Journal of Computational and Applied Mathematics Copyright © 2003 Elsevier B.V. All rights reserved

Volume 121, Issues 1-2, Pages 1-465 (1 September 2000)

North-Holland

		View: Citations	- Go
1. 🗖	Approximation in normed linear spaces, Pages 1-36 G. A. Watson SummaryPlus Full Text + Links PDF (210 K)		
2.	A tutorial history of least squares with applications to astronomy and Yves Nievergelt SummaryPlus Full Text + Links PDF (294 K)	geodesy, Pages 37-72	
3. 🗖	Shape-preserving approximation by polynomials , <i>Pages 73-94</i> D. Leviatan <u>SummaryPlus</u> <u>Full Text + Links</u> <u>PDF (183 K)</u>		
4.	Classical orthogonal polynomials: dependence of parameters, Pages 9. A. Ronveaux, A. Zarzo, I. Area and E. Godoy SummaryPlus Full Text + Links PDF (143 K)	5-112	
5.	Computing integral transforms and solving integral equations using C approximations, <i>Pages 113-124</i> Robert Piessens SummaryPlus Full Text + Links PDF (88 K)	hebyshev polynomial	
6. 🗖	Developments in bivariate spline interpolation , <i>Pages 125-152</i> G. Nürnberger and F. Zeilfelder <u>SummaryPlus</u> <u>Full Text + Links</u> <u>PDF (311 K)</u>		
7. 🗖	Multivariate spline and algebraic geometry, Pages 153-163 Ren-Hong Wang SummaryPlus Full Text + Links PDF (98 K)		
8. 🗖	A unified theory of radial basis functions: Native Hilbert spaces for rad Robert Schaback SummaryPlus Full Text + Links PDF (105 K)	dial basis functions II, Pages 165-177	
9.	Constraint-selected and search-optimized families of Daubechies wav factorization, Pages 179-195 Carl Taswell SummaryPlus Full Text + Links PDF (303 K)	elet filters computable by spectral	
10. 🗖	Multivariate Padé approximation, <i>Pages 197-219</i> Philippe Guillaume and Alain Huard SummaryPlus Full Text + Links PDF (175 K)		
11. 🗖	Numerical and asymptotic aspects of parabolic cylinder functions, Pa Nico M. Temme SummaryPlus Full Text + Links PDF (209 K)	ges 221-246	

12. 🗖	Computational strategies for the Riemann zeta function, Pages 247-296 Jonathan M. Borwein, David M. Bradley and Richard E. Crandall SummaryPlus Full Text + Links PDF (302 K)	
13. 🗖	Recursion formulae for basic hypergeometric functions, <i>Pages 297-312</i> Stanislaw Lewanowicz SummaryPlus Full Text + Links PDF (136 K)	
14. 🗖	The left-definite spectral theory for the classical Hermite differential equation, <i>Pages 313-330</i> W. N. Everitt, L. L. Littlejohn and R. Wellman SummaryPlus Full Text + Links PDF (149 K)	
15. 🗖	Minimal state-space realization in linear system theory: an overview, <i>Pages 331-354</i> B. De Schutter <u>SummaryPlus</u> <u>Full Text + Links</u> <u>PDF (150 K)</u>	
16. 🗖	Rational approximation in linear systems and control, <i>Pages 355-378</i> A. Bultheel and B. De Moor SummaryPlus Full Text + Links PDF (195 K)	
17. 🗖	Summary of sinc numerical methods, <i>Pages 379-420</i> Frank Stenger <u>SummaryPlus</u> <u>Full Text + Links</u> <u>PDF (276 K)</u>	
18. 🗖	Interval analysis: theory and applications, Pages 421-464 Götz Alefeld and Günter Mayer Abstract PDF (285 K)	
19. 🗖	Index, Page 465 PDF (30 K)	
20.	Numerical Analysis in the 20th Century Vol. I: Approximation Theory, <i>Pages ix-x</i> Luc Wuytack and Jet Wimp SummaryPlus Full Text + Links PDF (28 K)	



JOURNAL OF COMPUTATIONAL AND APPLIED MATHEMATICS

Journal of Computational and Applied Mathematics 121 (2000) ix-x

www.elsevier.nl/locate/cam

Preface

Numerical Analysis in the 20th Century Vol. I: Approximation Theory

The field of numerical analysis has witnessed many significant developments in the 20th century and will continue to enjoy major new advances in the years ahead. Therefore, it seems appropriate to compile a "state-of-the-art" volume devoted to numerical analysis in the 20th century. This volume on "Approximation Theory" is the first of seven volumes that will be published in this Journal. It brings together the papers dealing with historical developments, survey papers and papers on recent trends in selected areas.

In his paper, G.A. Watson gives an *historical survey* of methods for solving approximation problems in normed linear spaces. He considers approximation in L_p and Chebyshev norms of real functions and data. Y. Nievergelt describes the history of least-squares approximation. His paper surveys the development and applications of ordinary, constrained, weighted and total least-squares approximation. D. Leviatan discusses the degree of approximation of a function in the uniform or L_p -norm.

The development of numerical algorithms is strongly related to the type of approximating functions that are used, e.g. orthogonal polynomials, splines and wavelets, and several authors describe these different approaches.

E. Godoy, A. Ronveaux, A. Zarzo, and I. Area treat the topic of classical *orthogonal polynomials.* R. Piessens, in his paper, illustrates the use of Chebyshev polynomials in computing integral transforms and for solving integral equations.

Some developments in the use of *splines* are described by G. Nürnberger, F. Zeilfelder (for the bivariate case), and by R.-H. Wang in the multivariate case. For the numerical treatment of functions of several variables, radial basis functions are useful tools. R. Schaback treats this topic in his paper. Certain aspects of the computation of Daubechie *wavelets* are explained and illustrated in the paper by C. Taswell. P. Guillaume and A. Huard explore the case of multivariate Padé approximation.

Special functions have played a crucial role in approximating the solutions of certain scientific problems. N. Temme illustrates the usefulness of parabolic cylinder functions and J.M. Borwein, D.M. Bradley, R.E. Crandall provide a compendium of evaluation methods for the Riemann zeta function. S. Lewanowicz develops recursion formulae for basic hypergeometric functions. Aspects of the spectral theory for the classical Hermite differential equation appear in the paper by W.N. Everitt, L.L. Littlejohn and R. Wellman.

Many *applications* of approximation theory are to be found in linear system theory and model reduction. The paper of B. De Schutter gives an overview of minimal state space realization in

linear system theory and the paper by A. Bultheel and B. De Moor describes the use of rational approximation in linear systems and control.

For problems whose solutions may have singularities or infinite domains, *sinc approximation methods* are of value. F. Stenger summarizes the results in this field in his contribution.

G. Alefeld, and G. Mayer, provide a survey of the historical development of *interval analysis*, including several applications of interval mathematics to numerical computing.

These papers illustrate the profound impact that ideas of approximation theory have had in the creation of numerical algorithms for solving real-world scientific problems. Furthermore, approximation- theoretical concepts have proved to be basic tools in the analysis of the applicability of these algorithms.

We thank the authors of the above papers for their willingness to contribute to this volume. Also, we very much appreciate the referees for their role in making this volume a valuable source of information for the next millennium.

Luc Wuytack Departement Wiskunde Universiteit Antwerpen U.I.A. Universiteitsplein 1 2610 Wilrijk, Belgium E-mail address: wuytack@uia.ua.ac.be

Jet Wimp Dept. Mathematics & Comp. Drexel University Philadelphia, PA 19104 USA E-mail address: jwimp@msc.drexel.edu



Journal of Computational and Applied Mathematics 121 (2000) 1-36

JOURNAL OF COMPUTATIONAL AND APPLIED MATHEMATICS

www.elsevier.nl/locate/cam

Approximation in normed linear spaces

G.A. Watson

Department of Mathematics, University of Dundee, Dundee DD1 4HN, Scotland, UK

Received 10 October 1999; received in revised form 10 January 2000

Abstract

A historical account is given of the development of methods for solving approximation problems set in normed linear spaces. Approximation of both real functions and real data is considered, with particular reference to L_p (or l_p) and Chebyshev norms. As well as coverage of methods for the usual linear problems, an account is given of the development of methods for approximation by functions which are nonlinear in the free parameters, and special attention is paid to some particular nonlinear approximating families. © 2000 Elsevier Science B.V. All rights reserved.

1. Introduction

The purpose of this paper is to give a historical account of the development of numerical methods for a range of problems in best approximation, that is problems which involve the minimization of a norm. A treatment is given of approximation of both real functions and data. For the approximation of functions, the emphasis is on the use of the Chebyshev norm, while for data approximation, we consider a wider range of criteria, including the other l_p norms, $1 \le p < \infty$. As well as the usual linear problems, a general account is given of nonlinear best approximation, and we also consider some special cases. Only a passing mention is made of least-squares problems, as that is considered elsewhere. The focus is also entirely on the approximation of real quantities, and so best approximation of complex quantities is not covered. A partial justification of this is that dealing with problems in generality as complex ones would introduce additional complication not entirely justified by the additional algorithmic initiatives.

Since we are concerned here with historical development, technical details are not included for their own sake. The intention is, where appropriate, to be descriptive, rather to give a technically rigorous and detailed account of methods. However, it seemed necessary at times for the sake of comprehensibility, and in order to fully appreciate algorithmic developments, to include a reasonable amount of technical detail.

E-mail address: gawatson@maths.dundee.ac.uk (G.A. Watson).

Obviously a major factor in the development of methods has been the advent of powerful computing facilities, as this has opened up opportunities to tackle a wide range of practical problems. Whereas at one time, the main consideration may have been elegance and simplicity, with attention perhaps focussed on a set of problems satisfying "classical" assumptions, those considerations now usually have to take second place to the treatment of problems which are seen to be of practical importance, for which algorithms have to be robust and efficient.

The paper is effectively divided into two parts, the first (Section 2) being concerned with approximation by linear families, and the second (Section 3) being concerned with approximation by nonlinear families. These sections themselves further subdivide into two parts, where we consider separately approximation of data and of functions, and these are dealt with in that order within the two sections, with a further breakdown in what seems to be a reasonably natural way to take account of important special cases.

For the approximation of functions, we are primarily concerned with univariate functions on an interval [a, b], because that is where most effort has been concentrated. However, some relevant comments are made on the extent to which multivariate functions may also be treated, with a few references made to this.

2. Linear approximation

The approximation of a given function defined on an interval by a linear combination of given functions is the most fundamental problem in approximation theory. The functions involved are usually continuous, and this can be thought of as a continuous infinite dimensional approximation problem. If the functions are replaced by vectors in \mathbb{R}^m , then we have a class of finite dimensional or discrete problems, many of which have their origins in data fitting. That solutions to linear best approximation problems always exist is a result which goes back at least to Riesz in 1918 [174]. We will consider the finite dimensional problem first, and begin by making some general remarks, before looking at special cases.

2.1. Linear approximation in \mathbb{R}^m

Let $A \in \mathbb{R}^{m \times n}$ where $m \ge n$, and let $\boldsymbol{b} \in \mathbb{R}^m$. Then the statement of a linear best approximation problem in \mathbb{R}^m can be given as

find
$$\mathbf{x} \in \mathbb{R}^n$$
 to minimize $||\mathbf{r}||$, (1)

where

$$r = Ax - b$$

and ||.|| is a given norm on \mathbb{R}^m . The dependence of r on x will generally be suppressed, unless confusion is possible.

This particular problem has attracted enormous interest. It will be assumed throughout that rank(A) = n, and there is no x such that r = 0. These are not essential, neither in theory nor in practice; however, they are conditions that are normally satisfied in practice, and their assumption considerably simplifies the presentation. If the norm is a differentiable function of x, then we can easily characterize a minimum by zero derivative conditions: these are necessary, and, exploiting

convexity, also sufficient. The best known example is when the norm is the least-squares norm, when zero derivative conditions just give the usual normal equations

$$A^{\mathrm{T}}A\boldsymbol{x} = A^{\mathrm{T}}\boldsymbol{b}$$

The method of least squares is considered in detail elsewhere. But in a data fitting context, other l_p norms, particularly those for values of p satisfying $1 \le p < 2$ are also important. The reason for this is that it is common for the usual conditions justifying the use of the l_2 norm not to hold, for example there may be wild points or gross errors in the data, and these other norms give reduced weight to these wild points. This is considered in Sections 2.2 and 2.3. Of great interest also has been the use of the Chebyshev norm; this is perhaps of less value in a data fitting context, but problems arise for example in continuous function approximation when the region of approximation is discretized. The problem is rich in structure and the theory is a beautiful one; we consider this case in Section 2.4.

We will restrict attention here to the problem (1), although there are many modifications of that problem which are relevant in a data fitting context. Most modifications have only been given serious treatment comparatively recently, and so they are of lesser interest from a historical point of view.

2.2. Linear l_1 approximation in \mathbb{R}^m

Consider now the problem (1) with the l_1 norm

$$||\mathbf{r}||_{1} = \sum_{i=1}^{m} |r_{i}|.$$
(2)

This problem has a long history: its statement goes back well into the mid eighteenth century, and predates the introduction of least squares. Certainly, it was used in work of Laplace in 1786, in solving the overdetermined system of linear equations determining planetary orbits [110]. The first systematic methods for solving this problem seem due to Edgeworth [61]; in 1887 he gave a method based on tabulation, and in 1888 a method for the case when n = 2 which was essentially graphical and conceptual, but based on calculating descent directions. In 1930, Rhodes [167], motivated by the problem of fitting a parabola to data, tried Edgeworth's later method but found it "cumbrous". He gave a method where each iteration was calculated by solving 2 interpolation conditions for 2 of the parameters, and minimizing with respect to the remaining parameter. A proof that this kind of approach can give a solution was established by Singleton in 1940 [182]. A detailed historical account is given by Farebrother in a 1987 paper [63], covering the period 1793 to 1930.¹

The first modern systematic study of this problem appears to be by Motzkin and Walsh [131,132] in the late 1950s, and characterization results are given in the 1964 book by Rice [172]. A convenient form of these may be deduced from these results or as a simple consequence of applying to this special case known results in abstract approximation theory: we will not attempt to go down that historical route, since it is something of a diversion from the main theme. However, it is the case

¹ The 1999 book by Farebrother [64] is also relevant.

that a vector $x \in \mathbb{R}^n$ solves the l_1 problem if and only if there exists a vector $v \in \mathbb{R}^m$ satisfying

$$A^{\mathrm{T}}\boldsymbol{v}=0,$$

where $||v||_{\infty} \leq 1$, and $v_i = \operatorname{sign}(r_i)$ whenever $r_i \neq 0$. The first simple (direct) proof of this was probably given by Watson [199] in 1980. A formal treatment of the important result that when *A* has rank *n*, a solution will be such that *n* components of r_i are zero, was given by Motzkin and Walsh [131] in 1955. In the context of the l_1 problem, any point characterized in this way can be defined to be a *vertex*. The interpolation result (in special cases) appears to have been known to Gauss, and to have been used in early methods: for example, the methods of Rhodes and Singleton are essentially vertex to vertex descent methods.

The results of Motzkin and Walsh were arrived at by direct consideration of the problem. However, its relationship with a linear programming problem was recognized around the same time,² and linear programming theory provides a parallel route to the same properties. Around 1947, Dantzig did his pioneering work on the simplex method of linear programming, and over the next few years, duality theory was developed, largely by von Neumann, Gale, Kuhn and Tucker. The significance of these developments for numerical methods for the l_1 (and the l_{∞}) problem cannot be overemphasized.

The first representation of the l_1 problem as a tractable linear programming problem seems due to Charnes et al. [35] in 1955. The key observation is that if extra variables u and $v \in \mathbb{R}^m$ are introduced, then the problem can be posed as

minimize
$$\sum_{i=1}^{m} (u_i + v_i)$$
 subject to
$$[I : -I : A] \begin{bmatrix} u \\ v \\ x \end{bmatrix} = b$$
 $u \ge 0, \quad v \ge 0.$
(3)

Since in the simplex method, no columns of I and -I can simultaneously be basic, then

$$u_i v_i = 0, \quad i = 1, \ldots, m.$$

It follows that $u_i + v_i = |u_i - v_i|$ for all *i* and the equivalence of the simplex method applied to this problem with the minimization of (2) can readily be established.

Another version of the primal can be stated:

minimize $e^{T}s$ subject to

 $-s \leq Ax - b \leq s$.

This goes back at least to the 1964 Russian edition of the book by Zuhovitskii and Avdeyeva [211]. However, this form of the problem does not seem to have attracted as much attention as (3). The zero residuals will result in a form of degeneracy.

² Farebrother [63] in his 1987 paper interprets the work of Edgeworth in this context, and states that "...it must be conceded that Edgeworth had developed a fully operational, if somewhat complex, linear programming procedure for the L_1 estimation problem in 1888".

5

Fisher [66] in 1961 gave some publicity to (3) for the benefit of the statistical community, and this form was also used by Barrodale and Young [13] in 1966, who provided an Algol implementation and numerical results. The fact that the components of x may be non-negative is not a major problem in this context: for example, they can each be replaced by the difference of two non-negative variables. It was also noted that no first phase simplex calculation is required because an initial basic feasible solution can readily be obtained: if $b_i < 0$ then e_i can be present in the initial basis, if $b_i > 0$ then $-e_i$ can be, with either used if $b_i = 0$.

The linear programming connection is sometimes wrongly credited to Wagner [192] in 1959, who posed the problem as a bounded variable or interval programming problem. In fact the form of the problem considered by Wagner [192] can be interpreted as the dual of (3). This can be written as

maximize
$$\boldsymbol{b}^{\mathrm{T}}\boldsymbol{v}$$
 subject to (4)

$$A^{\mathrm{T}}\boldsymbol{v} = \boldsymbol{0}$$

$$-e \leqslant v \leqslant e$$
,

where e is a vector with every component equal to 1. Attention was re-focussed on (4) by Robers and Ben-Israel [175] in 1969, and Robers and Robers [176] in 1973, who argued the advantages of that approach, which included computational efficiency: the problem with the primal appeared to be the large number of extra variables required. However, an improved version of the primal linear programming method was given by Davies [53] in 1967 and Barrodale and Roberts [10] in 1970, where a special pivot column selection rule was employed, and in 1973, both Spyropoulos et al. [183] and Barrodale and Roberts [11] gave efficient implementations of the simplex method applied to the primal which fully exploited the structure. The Barrodale and Roberts method achieved efficiency by taking multiple pivot steps, exploiting the fact that descent can continue beyond the usual point when feasibility is lost, because feasibility can readily be recovered by swapping certain variables into and out of the basis. Further efficiency was achieved by imposing certain restrictions on the choice of variables to enter and leave the basis. A Fortran programme and numerical results were provided, together with favourable comparisons with some other primal and dual methods [12].

In 1975, Abdelmalik [2] developed a special purpose method for the dual, using the dual simplex method, and his method seemed comparable with that of Barrodale and Roberts [10]. This turned out not really to be surprising, because, as pointed out by Armstrong and Godfrey [6] in 1979, the application of the dual simplex method to the dual is equivalent to applying the primal simplex method to the primal. So apart from implementation aspects, the methods were the same.

A basic feasible solution to (3) in which all columns of A are present in the basis can readily be shown to correspond to a vertex as defined above. Therefore, once the columns of A are present in the basis, the simplex method is a vertex to vertex descent method. There are many other variants of these linear programming methods, but away from a linear programming context, *direct* descent methods were being considered. For given x, let

$$Z = \{i : r_i = 0\}.$$

Then since for full rank problems the solution occurs at a point x with Z containing n indices (a vertex), we want to systematically descend to such a point. Perhaps the first modern direct descent methods were given by Usow [189] in 1967, and Claerbout and Muir [43] in 1973. A natural way

to implement descent methods is by first finding a vertex, and then descending through a sequence of vertices. Thus there are two types of step depending on whether at the current point, Z contains (a) fewer than n indices (b) exactly n indices. (The possibility that Z contains more than n indices corresponds to a degenerate situation, and although there are ways round it, will for our purposes be ignored.) Then in case (a) movement as far as possible is made in the direction d in such a way that the number of indices in Z at the new point is increased, and in case (b) movement as far as possible is made in the direction d in such a way that the number of indices in Z is maintained. Effective methods of this type, therefore, have this strategy in common, and are distinguished by the way the descent direction is calculated. There are mainly two approaches, (i) *reduced gradient methods*, where the "active constraints" are used to express certain variables in terms of others, the objective function is expressed in terms of the latter group, and its gradient is obtained in terms of those, and (ii) *projected gradient methods*, where the gradient is obtained by projecting the gradient of the objective function onto the orthogonal complement of the span of the active constraints.

Bartels et al. [18] in 1978 gave a projected gradient method, and reduced gradient methods were given by Osborne [147,148] in 1985 and 1987. Both projected and reduced gradient methods were analyzed in detail by Osborne [147] in 1985, and he pointed out that although reduced gradient methods seem more suitable for implementation using a tableau format, with updating, in fact such organization is available for implementing both methods. On relationships with linear programming methods, he showed that there is an exact equivalence between the possible options available in implementing the simplex method and those available in the direct application of the reduced gradient. The usual simplex step corresponds to a particular option in the reduced gradient method, based on an unnormalized steepest edge test for determining the variable to leave the basis. A different way of choosing this variable (a normalized steepest edge test, which is scale invariant) was used by Bloomfield and Stieger [26] in 1983, and their evidence showed that this can lead to improvement.

Nearly all the good methods considered to the end of the 1980s were vertex to vertex methods, which exploit the polyhedral nature of the function to be minimized, and (in the absence of degeneracy) they are finite. There has been recent interest in interior point methods for linear programming problems, stimulated by the results of Karmarker [102] in 1984. In conjunction with a formal connection with classical barrier methods for constrained optimization problems, this has resulted in renewed interest in linear programming, and there has of course been an impact on special cases such as the l_1 problem.

The use of interior point methods for l_1 problems goes back at least as far as work of Meketon [127] in 1987, and methods have been given since then by Ruzinsky and Olsen [178] in 1989, Zhang in 1993 [209] and Duarte and Vanderbei [56] in 1994. Portnoy and Koenker [157] in 1997 make a case for the superiority of interior point methods over simplex-based methods for large problems. Based on comparisons of l_1 problems having *n* up to 16 and *m* from 1000 to 200 000, they conclude that there is "a compelling general case for the superiority of interior point methods for large linear programming problems". Their algorithm of choice for the l_1 problem is based on a primal–dual log barrier method due to Mehrotra [123] in 1992, and includes a statistical preprocessing approach which estimates whether a residual is zero or not. The opposition is represented by a variant of the Barrodale and Roberts method.

Meantime, two other types of smoothing method were being developed for the l_1 problem.³ The first of these is typified by an algorithm of Coleman and Li [46] in 1992, which is based on affine scaling: while not strictly speaking an interior point method, it is nevertheless in the spirit of such methods. Here, an attempt is made to satisfy the characterization conditions by an iterative descent method which has the following characteristics: (a) it generates a sequence of points which are such that Z is empty, so that derivatives exist, (b) it is globally convergent, (c) it ultimately takes damped Newton steps (damped to satisfy (a)), but with sufficiently accurate approximations to the full Newton step to permit quadratic convergence (under nondegeneracy conditions). Careful implementation of the method can avoid difficulties with near-zero components of r and the approach seems promising for large problems as it is insensitive to problem size. Some comparisons show that it is superior to Meketon's interior point method for problems with n up to 200, m up to 1000.

A second approach to smoothing the l_1 problem was developed by Madsen and Nielsen [116] in 1993. It is based on the use of the Huber M-estimator, defined by

$$\psi_{\gamma} \equiv \psi_{\gamma}(\mathbf{r}) = \sum_{i=1}^{m} \rho(r_i), \tag{5}$$

where

$$\rho(t) = \begin{cases} t^2/2, & |t| \leq \gamma, \\ \gamma(|t| - \gamma/2), & |t| > \gamma, \end{cases}$$
(6)

and γ is a scale factor or tuning constant. The function (5) is convex and once continuously differentiable, but has discontinuous second derivatives at points where $|r_i| = \gamma$. The mathematical structure of the Huber M-estimator seems first to have been considered in detail by Clark [44] in 1985. Clearly if γ is chosen large enough, then ψ_{γ} is just the least-squares function; in addition if γ tends to zero, then limit points of the set of solutions may be shown to minimize the l_1 norm. It is the latter property which concerns us here.

It has been suggested by Madsen and Nielsen [116] in 1993 and also by Li and Swetits [113] in 1998 that the preferred method for solving the l_1 problem is via a sequence of Huber problems for a sequence of scale values $\gamma \to 0$. This algorithmic development has lead to increased interest in the relationship between the Huber M-estimator and the l_1 problem; for example there is recent work of Madsen et al. [117] in 1994, and Li and Swetits [113] in 1998. The method of Madsen and Nielsen generates Huber solutions for a sequence of values of γ , tending to zero. The solutions are obtained by solving least-square problems, exploiting structure so that new solutions can be obtained using updating often in $O(n^2)$ operations. A key feature is that it is not necessary to let γ reach zero; once a sufficiently small value is identified, then the l_1 solution may be obtained by solving an $n \times n$ linear system. Madsen and Neilsen give some comparisons (for randomly generated problems, and with *m* mostly set to 2n for *m* up to 1620) with the method of Barrodale and Roberts [10] and claim superiority.

An important issue as far as the implementation of simplex type methods is the efficiency of the line search. The Barrodale and Roberts [10] method incorporates the equivalent of a comparison sort, and this leaves room for considerable improvement. Bloomfield and Stieger [26] considered

³ The observation that a best approximation can always be computed as the limit of a sequence of l_p approximations as $p \to 1$ is due to Fischer [65] in 1983 (an algorithm based on this was in fact given by Abdelmalik [1] in 1971), although this is not a very practical approach.

this aspect in their 1983 book, and suggested using a fast median method. An alternative based on the use of the secant algorithm was considered (in a related context) by George and Osborne [71] in 1990, and again by Osborne [147] in 1985. Numerical experiments were reported by Osborne and Watson [154] in 1996, where the secant-based method was seen to be as good as fast median methods on randomly generated problems, and to perform considerably better on problems with systematic data. Comparisons of other types of method with simplex methods really need to take this into account before definitive conclusions can be drawn.

2.3. Linear l_p approximation in \mathbb{R}^m , $1 , <math>p \neq 2$

For given $x \in \mathbb{R}^n$, let $D_{|r|}$ be defined by

$$D_{|\mathbf{r}|} = \operatorname{diag}\{|r_1|, \ldots, |r_m|\}$$

Then x minimizes

$$||\mathbf{r}||_{p}^{p} = \sum_{i=1}^{m} |r_{i}|^{p}$$

with 1 if and only if derivatives with respect to x are zero, that is if

$$A^{\mathrm{T}}D^{p-1}_{|r|}\theta = 0, (7)$$

where $\theta_i = \text{sign}(r_i)$, i = 1, ..., m. This is a nonlinear system of equations for x.

This criterion (for p even) was mentioned by Gauss as a generalization of his least-squares criterion. Apart from this special case, the more general l_p problem only seems to have attracted relatively recent computational attention. The range $1 is of particular interest computationally because there is potentially reduced smoothness: problems with <math>p \ge 2$ are twice differentiable, but problems with $1 may be only once differentiable. If <math>p \ge 2$ or if 1 and no component of <math>r is zero then twice differentiability is guaranteed and so (7) can be written as

$$A^{\mathrm{T}}D\mathbf{r} = \mathbf{0},\tag{8}$$

where

$$D = D_{|\mathbf{r}|}^{p-2},$$

and this is a particularly convenient form with which to work. It represents a generalized system of normal equations, effectively a least-squares problem apart from the "weighting" matrix D. Fixing x to an approximate value in D and solving this weighted system for a new approximation gives an example of the technique known as *iteratively reweighted least squares* or IRLS, which seems to have been introduced by Beaton and Tukey [20] in 1974. Since good software for (weighted) least-squares problems was then available, this seemed an attractive idea, additionally so since there are some apparently good theoretical properties: this simple iteration process will converge locally if p is close to 2, and if zero components of r are avoided, it is globally convergent (from any initial approximation) for 1 . The last result seems first to have been given by Dutter [60] in 1975. However, convergence can be slow, particularly as <math>p nears 1 (it is linear with convergence constant |p - 2|, as shown by Wolfe [206] in 1979), and there are potential numerical difficulties for reasons which will be clear from the previous section. The matrix D (which may not exist) can

be replaced by approximations (even by the unit matrix), and this gives rise to variants of the IRLS technique, but again convergence can be very slow.

Most recent algorithms for solving (8) are based on Newton's method, and many variants were proposed in the 1970s. It is interesting that the Newton step is just 1/(p-1) times the IRLS step (as measured by the difference between successive approximations), as pointed out by Watson [196] in 1977, and this gave an explanation of some success obtained by Merle and Späth [128] in 1974 in using a damped IRLS procedure with step length (p-1). Thus apart from differences due to the line search, IRLS and Newton's method with line searches *are essentially the same method*. It is easily seen that the region of convergence of Newton's method is proportional to |(p-1)/(p-2)|, so good line search procedures are needed even with the basic method, certainly far from p = 2. However, for p > 2, Newton's method with line search is usually perfectly satisfactory.

Since from a practical point of views the interesting cases are those when 1 , differentstrategies have been proposed for getting round the difficulties arising from zero (or near zero) components of r. These included the substitution of small nonzero values, solving a slightly perturbed problem, or identifying and so removing these components from the set. However, not just zero components but *nearly zero* components are potentially troublesome. There is some evidence, however, that these phenomena are not by themselves a major problem, but only if they are accompanied by p being close to 1. The main difficulty appears to be due the fact that as p approaches 1, we are coming closer to a discontinuous problem, effectively to a constrained problem. It seems necessary to recognize this in a satisfactory algorithm, and consider some of the elements of the l_1 problem in devising an approach which will deal in a satisfactory with small values of p. This is the philosophy in a recent method due to Li [114] in 1993, which is essentially equivalent to the method for the l_1 problem of Coleman and Li [46] referred to in the previous section. Numerical results show that the new method is clearly superior to IRLS (with the same line search) for values of p close to 1, with the gap between the two methods widening as p approaches 1. There is little difference for values of $p \ge 1.5$ or so. As with the l_1 case, the number of iterations appears to be independent of the problem size.

2.4. Linear Chebyshev approximation in \mathbb{R}^m

The use of the criterion now known as the Chebyshev norm

$$||\boldsymbol{r}||_{\infty} = \max |r_i|,\tag{9}$$

seems to go back to Laplace in 1786, who gave a solution procedure for n = 2. Cauchy in 1814 and Fourier in 1824 gave descent methods. A detailed historical account is given by Farebrother [63] in his 1987 paper, covering the period 1793 to 1824. The function space analogue was studied first by Chebyshev⁴ from the 1850s, arising from an analysis of a steam engine linking, and both continuous and discrete problems now carry his name.

For any $x \in \mathbb{R}^n$, let

 $\overline{I}(\boldsymbol{x}) = \{i: |r_i(\boldsymbol{x})| = ||\boldsymbol{r}||_{\infty}\}.$

⁴ The number of variants in the western literature which have been used for Chebyshev is legendary, but most people now seemed to have settled on this one. Butzer and Jongmans [33] in 1999 gave a detailed account of Chebyshev's life and work.

.

Then x is a solution if and only if there exists a subset $I \subset \overline{I}$ containing at most n + 1 indices, and a nontrivial vector $\lambda \in \mathbb{R}^m$ such that

$$\lambda_i = 0, \quad i \notin I,$$

 $A^{\mathrm{T}}\lambda = 0,$
 $\lambda_i \operatorname{sign}(r_i) \ge 0, \quad i \in I.$

This is an example of a "zero in the convex hull" type of characterization result, and is the discrete analogue of Kirchburger's 1903 result for the continuous problem [106]. A simple consequence is that for the full rank problem, there always exists a solution with \overline{I} containing n + 1 indices. Such a point can be thought of as a *vertex*.

An early method for the minimization of (9) was the Polya algorithm [156], which solves a sequence of l_p problems with $p \to \infty$: the assumption here is that the l_p problems are relatively easy to solve, being differentiable for large finite p. This method was given (in fact for continuous functions) in 1913, and convergence is readily established if the Chebyshev solution is unique. A proof of convergence to a particular Chebyshev approximation called the strict Chebyshev approximation (in the event of nonuniqueness of the Chebyshev solution) was given by Descloux [54] in 1963. Fletcher et al. [69] in 1974 used an extrapolation technique to accelerate convergence of the Polya algorithm, and in the same year Boggs [27] used a technique based on deriving a differential equation describing the l_p solution as a function of p. An algorithm due to Lawson [111] in 1961 was based on the solution of a sequence of weighted least-squares problems, but like the Polya algorithm, it can be very slowly convergent. Indeed none of these methods has been regarded as giving a particularly practical approach.

A fundamental assumption which was identified as important at an early stage was the *Haar* condition, that every $n \times n$ submatrix of A is nonsingular. This is sufficient for uniqueness of x minimizing (9) (and also necessary in the case when m = n + 1). This "classical" assumption played a major role in the minimization of (9) until the 1960s. It goes back to Haar [79] in 1918.

Before proceeding, it is helpful to point out an important property which is satisfied at a minimum of (9). The result, due to de la Vallée Poussin [190] in 1911, tells us that if J runs through all subsets of n + 1 indices from $\{1, ..., m\}$, then

$$\min_{x} \max_{i} |r_{i}| = \max_{J} \left\{ \min_{x} \max_{i \in J} |r_{i}| \right\}.$$
(10)

Therefore, if we can identify a set J where the maximum on the right-hand side is attained, solving a Chebyshev problem *on that subset* (and this is relatively easy) will give a solution to the original problem. For any J such that the corresponding problem matrix is full rank, the solution on J will occur at a vertex. Therefore, if A has full rank, so that the problem has a solution at a vertex, then it is sufficient to investigate all the vertices in a systematic way.

An exchange algorithm for finding an extremal subset or optimal vertex was given by Stiefel [184] in 1959. It assumed that A satisfied the Haar condition, and worked with a sequence $J_1, J_2, ...$ of subsets of n+1 components of r. The key aspect of the method was that J_{k+1} differed from J_k by one index, and a rule was given for exchanging one of the indices in J_k for another index outside it to give J_{k+1} in such a way that

$$h_{k+1} > h_k,$$

where

$$r_i(\mathbf{x}) = \theta_i h_k, \quad i \in J_k, \tag{11}$$

with $|\theta_i| = 1$, $i \in J_k$. Thus, we have an example of a vertex-to-vertex *ascent method*. Because there are only a finite number of selections of n + 1 indices from *m* the method must terminate, when from (10) it follows that a solution has been obtained. If the Haar condition is not satisfied, then strict inequality may hold for successive h_k values and the theory of the method is compromised.

Because the function of x given by (9) is piecewise linear, the Chebyshev problem may be posed as a linear programming problem, and as in the l_1 case, properties of the problem are again available through this route. Let $h = \max_i |r_i|$. Then the problem may be stated

minimize h subject to

 $-h \leq r_i \leq h$, $i = 1, \ldots, m$.

In terms of the variables h and x, this may be restated

minimize
$$z = e_{n+1}^{\mathsf{T}} \begin{bmatrix} \mathbf{x} \\ h \end{bmatrix}$$
 subject to
 $\begin{bmatrix} A & e \\ -A & e \end{bmatrix} \begin{bmatrix} \mathbf{x} \\ h \end{bmatrix} \ge \begin{bmatrix} \mathbf{b} \\ -\mathbf{b} \end{bmatrix}.$

One of the first to consider the linear programming formulation of the problem was Zuhovickii [211,212] in a series of papers originating in the Russian literature in the early 1950s. The above form is not particularly suitable for the application of standard techniques such as the simplex method because 2m slack variables are required, the basis matrices will be $2m \times 2m$, and although h is nonnegative, this is not true in general of the components of x.

All of these difficulties are overcome by turning to the dual problem, which is

maximize $z = [\boldsymbol{b}^{\mathrm{T}}, -\boldsymbol{b}^{\mathrm{T}}]\boldsymbol{w}$ subject to

$$[A^{\mathrm{T}} - A^{\mathrm{T}}]w = 0,$$
$$[e^{\mathrm{T}} e^{\mathrm{T}}]w \leq 1,$$
$$w \geq 0.$$

Only one slack variable is required (to make the inequality an equality), the basis matrices are only $(n + 1) \times (n + 1)$, and all the variables are nonnegative. The advantage in using the dual seems to have been first pointed out by Kelley in [105] 1958 in an application to curve fitting. Standard linear programming theory tells us that if a variable is dual basic, then the corresponding primal constraint holds with equality. Thus a basic feasible solution corresponds precisely to a solution to a set of equations having the form (11). It would appear therefore that there is a precise equivalence between a step of the simplex method applied to the dual, and a step of the Stiefel exchange method. This result was known to Stiefel [185] in 1960, who gave an indication of it by considering a small problem and using a geometric argument. He also (unnecessarily) eliminated the unconstrained variables from the primal before proceeding to the dual. The precise equivalence was first worked out in detail by Osborne and Watson [149] in 1967, although Bittner [24] in 1961 examined how linear programming could be used to relax the Haar condition. In order for the usual simplex method to be applied to the dual, the Haar condition is not required, only the nonsingularity of successive basis matrices: for this it is only necessary for A to have full rank n. The point here is that the simplex method does not permit nonsingular basis matrices. Note however that if the Haar condition does not hold, degeneracy may cause cycling in the simplex algorithm, although this can always be resolved.

A programme implementing the standard simplex method for the problem considered here seems first to have been given by Barrodale and Young [13] in 1966, who gave an Algol programme along with numerical results, and Bartels and Golub [19] gave a version in 1968 which used a numerically stable factorization procedure. In 1975, Barrodale and Phillips [8] used the special structure present in the dual formulation to greatly reduce the number of iterations required: conditions were imposed on variables entering and leaving the basis, and the fact exploited that some variables could easily be exchanged for others. The usual simplex rules were modified to permit ascent through a number of vertices, beyond the one which would usually be reached in a simplex step, by exploiting the fact that feasibility could easily be regained by such exchanges. Modifications of this basic technique to allow more than one index to be exchanged at each step were given by Hopper and Powell [90] in 1977 and by Armstrong and Kung [7] in 1979.

The Stiefel exchange method and variants which solve the dual formulation of the problem are examples of ascent methods, whose justification is based on (10). However, it is possible to solve the problem by a *descent* process. The primal linear programming problem is an example of a descent method, and although its direct solution is not recommended for the reasons already given, it is nevertheless possible to implement satisfactory descent methods.

As for the l_1 problem, good direct descent methods might be expected to follow the common strategy of having (in the absence of degeneracy) basically two types of step depending on whether the current point x is such that \overline{I} contains (a) fewer than n + 1 indices, or (b) exactly n + 1indices. In a manner precisely analogous to that considered for the l_1 problem, a strategy can be developed which ultimately gives a vertex-to-vertex descent process. Methods of reduced gradient type were given by Cheney and Goldstein [37] in 1959 and Cline [45] in 1976. A projected gradient method was given by Bartels et al. [16] in 1978. It appeared to be the case that such methods did not seriously compete with ascent methods. However, improvements in descent methods were considered by Bartels et al. [17] in 1989: they argued that the good performance of the Barrodale and Phillips method was due to the way the method chose a good starting point. By modifying the way in which a starting point is obtained for their descent method, they enhanced its performance and made a case for its superiority for data fitting problems.

All the approaches considered so far are essentially vertex-to-vertex methods. They exploit the polyhedral nature of the function to be minimized, and are of course (in the absence of degeneracy) finite. The recent interest in interior point methods for linear programming problems has, as in the l_1 case, extended to the special case of Chebyshev problems. Ruzinsky and Olsen [178] in 1989, Zhang [209] in 1993 and Duarte and Vanderbei [56] in 1994 all proposed interior point methods. An affine scaling algorithm analogous to that for the l_1 problem was given by Coleman and Li [47] in 1992. This is a descent method which involves a sequence of least-squares problems to define descent directions. It provides a smooth transition from guaranteed descent steps far from a solution, to steps close to a solution which are sufficiently accurate

approximations to the Newton step to permit quadratic convergence under suitable nondegeneracy assumptions.

In contrast to the l_1 situation, detailed comparisons of other methods with simplex type methods for large problems do not yet seem to be available. It should in any event not be assumed that conclusions can be drawn from the l_1 case, because large Chebyshev problems normally arise from discretizations of continuous Chebyshev approximation problems on intervals or multidimensional regions, and the data are highly systematic. Indeed, the solution is then normally part of a method for the continuous problem, or exploits the connection: we will defer further consideration of this until the following Section.

2.5. Linear Chebyshev approximation in C[a,b]

Let C[a, b] denote the set of continuous functions defined on the real interval [a, b], and let f(x), $\phi_1(x), \ldots, \phi_n(x)$ be in C[a, b]. Then the usual Chebyshev approximation problem in C[a, b] can be expressed as

find
$$\boldsymbol{a} \in \mathbb{R}^n$$
 to minimize $||f - \phi||_{\infty}$, (12)

where $\phi = \sum_{i=1}^{n} a_i \phi_i(x)$, and

$$||f||_{\infty} = \max_{a \leqslant x \leqslant b} |f(x)|.$$

This class of problems was systematically investigated by Chebyshev from the 1850s, although Chebyshev credits Poncelet with originating the problem. The "classical" case occurs when the set of functions forms a Chebyshev set (or is a Haar subspace) on [a, b], that is any nontrivial linear combination has at most (n - 1) zeros; the model problem here is approximation by polynomials of degree n - 1. The problem (12), with the interval [a, b] replaced by m points in [a, b], reduces to a problem of the form considered in the previous section. Indeed it is readily seen that the matrix A in this case satisfies the Haar condition if and only if the set of functions $\phi_1(x), \ldots, \phi_n(x)$ forms a Chebyshev set on [a, b]. Continuing this theme for a moment, arbitrarily good solutions to (12) can be obtained by choosing finer and finer discretizations; the main convergence results here are due to Motzkin and Walsh [132] in 1956. Although this observation by itself does not give practical algorithms, the use of a sequence of discretizations, where successive point sets are carefully chosen, is the key to the success of many good algorithms.

A general characterization result was obtained by Kirchberger [106] in 1903. Let

$$\bar{E} = \{x \in [a, b], |r(x, a)| = ||r(., a)||_{\infty}\}.$$

Then *a* is a solution if and only if there exists $E \subset \overline{E}$ containing $t \leq n+1$ points x_1, \ldots, x_t and a nontrivial vector $\lambda \in \mathbb{R}^t$ such that

$$\sum_{i=1}^{l} \lambda_i \phi_j(x_i) = 0, \quad j = 1, \dots, n,$$
$$\lambda_i \operatorname{sign}(r_i) \ge 0, \quad i = 1, \dots, t.$$

Borel [28] in 1905 established the well-known alternation result for approximation by degree (n-1) polynomials, that *a* is a solution if and only if there are n+1 points in [a,b] where the

norm is attained with alternating sign as we move from left to right through the points; we can state this concisely in the form

$$\mathscr{A}(f-p_{n-1})_{[a,b]} \ge n+1,$$

where p_n denotes the best degree *n* polynomial approximation. Uniqueness of solutions under these conditions is also due to Borel [28] in 1905. That this result extends to approximation by functions forming a Chebyshev set was shown by Young [208] in 1907, who also established uniqueness in this case. Haar [79] in 1918 showed that the solution is unique for all possible functions f(x) if and only if $\phi_1(x), \ldots, \phi_n(x)$ forms a Chebyshev set on [a, b].

Polya [156] in 1913 gave his algorithm for this problem, where a sequence of continuous L_p problems is solved with $p \to \infty$. A counterexample to a general convergence result for non-Chebyshev set problems was given by Descloux [54] in 1963. As in the discrete case this is anyway not a particularly effective approach.

Two important algorithms for solving the Chebyshev problem were given by Remes [165,166] in the 1930s. The method traditionally known as the "Second Algorithm" applies to Chebyshev set problems, exploiting the alternation property. It solves a sequence of discrete problems in \mathbb{R}^{n+1} defined by sets of n+1 points in [a, b]: each of these is just the solution of a system of n+1 equations for n+1 unknowns, using the fact the solutions have an alternation property. By exchanging the current set of n+1 points for n+1 local maxima of the modulus of the error function, subject to some simple rules, an ascent process is obtained. Under mild conditions this converges to the (unique) Chebyshev approximation, and at a second-order rate: the result, due to Veidinger [191] in 1960, is based on showing that the method is asymptotically Newton's method for solving the characterization conditions. A comparatively modern implementation of the method was given by Golub and Smith [73] in 1971. Note that if only one point is exchanged at each iteration (bringing in a point where the norm of the error is attained), then an equivalence can be drawn between a step of the method and a step of the Stiefel exchange method. An analysis of the one-point exchange method is given by Powell [158] in his 1980 book, where it is shown that this method also converges at a second-order rate.

The "First Algorithm of Remes" applies to general problems. Again it corresponds to the solution of a sequence of discrete problems, but of increasing size. Starting with a solution on $m_1 \ge (n + 1)$ discrete points in [a, b], a point where the error function attains the norm is added, and a new solution obtained on $m_1 + 1$ points. If the matrix A of the initial problem has rank n, then successive matrices also have rank n and so linear programming techniques, for example, can be used, and implemented efficiently using postoptimality theory. This is an "implicit" exchange method, since every solution corresponds to a vertex defined on the current set of points. In fact since much of the work in implementing such a method lies in finding a global maximum of the error function, and this would normally involve calculating all the local maxima, it is sensible to add in all such local maxima: the method is then an implicit multiple exchange method. For Chebyshev set problems this is equivalent to the second algorithm of Remes. Modifications of the first algorithm of Remes to allow multiple exchanges have been considered by Carasso and Laurent [34] in 1978, and Blatt [25] in 1984, based on constructing "chains of references".

Unfortunately there are examples where this kind of approach performs badly, when the solution to the continuous problem does not occur at a vertex, that is it attains the norm in fewer than n + 1 points: such problems were called singular problems by Osborne and Watson [151] in 1969. Note

that this phenomenon is specific to problems on a continuum, and has no analogue in the (full rank) discrete case. Therefore, because each discrete problem has a solution at a vertex, the limiting situation in this case is obtained by some of these points coalescing, slowing down convergence and giving ill-conditioned simplex bases.

For multivariate problems (where x is a vector in \mathbb{R}^s , s > 1), singularity is very common. A partial explanation for this is that Chebyshev sets of more than one function do not exist in continuums of dimension higher than one: this was first pointed out by Mairhuber [119] in 1956. Nevertheless, a method of this type can be developed for multivariate problems, as demonstrated by Watson [194] in 1975.

Therefore, there are two main difficulties with such methods: (a) the calculation of the local and global maxima of the error function, (b) the problem of singularity. It is perhaps only recently that close attention has been paid to efficient calculation in (a), for example by Reemtsen [163] in 1991, and Price and Coope [159] in 1996: it is usually assumed that all local maxima can be calculated to sufficient accuracy, and so the relevant algorithms are always implementable. But attempts to avoid (a) have been made, for example by Dunham [58] in 1981, Hettich [85] in 1986 and Reemtsen [162] in 1990. The main idea is to only require maxima of the error at each step on a grid, where the discretization error tends to zero as the method progresses. In particular, Reemtsen proved the convergence of a modified version of the first algorithm of Remes, in which the maximum of the *k*th error function was computed on a grid, with the grid density tending to zero. The method of Hettich is also based on successive grid refinement (and using a numerically stable simplex algorithm) and applies to one- and two-dimensional problems; solutions have been successfully obtained for problems with n up to 37.

An alternative approach which tries to avoid both (a) and (b) is through the use of *two-phase methods*. The first phase involves the solution of a single discretization of the problem, on a sufficiently dense set to enable identification of the number of points (with signs) where the norm is attained and good approximations of these. In the second phase, the characterization conditions, together with zero derivative conditions at points identified as extrema in (a, b), can then be solved (for example by Newton's method). This main idea for an approach of this type (in a more general context) is due to Gustafson [77] in 1970. Its application to Chebyshev approximation problems was considered by a number of people in the mid-1970s, among them Gustafson, Hettich, Andreassen and Watson [5,78,84]. The approach can be successful, but while the difficulty (a) above is essentially removed, (b) can still emerge in the first phase, and there is also the (new) difficulty of having to decide what level of discretization to use, or when to enter the second phase, and also when the information provided at that point is completely reliable. It may be necessary to permit re-entry to phase 1 with a more stringent exit criterion.

The second phase can be considered in two ways, depending on whether or not the local maxima are considered as differentiable functions of the unknown parameters, and whether or not this is exploited. If it is, then the zero derivative conditions can be used to eliminate these maxima in terms of the other unknowns, and there is a consequent reduction in the size of the linear system to be solved for the Newton step.

Of course the second phase applies equally to nonlinear problems, so we will return to some of these ideas in Section 3.4. Indeed, continuous Chebyshev approximation problems (both linear and nonlinear) are special cases of semi-infinite programming problems, that is problems with a finite number of variables and an infinite number of constraints, and many algorithms which have been developed for the more general problem class may be adapted for the Chebyshev approximation problem. Semi-infinite programming is an active research area – the recent survey paper of Reemtsen and Görner [164] in 1998 has 233 references, 96 of them dated 1990 or later. Algorithmic development has encompassed methods based on the ideas considered above, but also other approaches, for example the use of interior point methods. These are of comparatively recent origin, their usefulness (certainly as far as continuous Chebyshev approximation is concerned) does not appear to have been established, and we will not consider them further here.

2.6. Chebyshev approximation by splines with fixed knots

Approximation by splines is considered in some detail elsewhere, so we will not go into the history of the origins of this class of function. The main focus of approximation by splines has been on interpolation; however, Chebyshev approximation by splines has also attracted a lot of attention. Because we are concerned at present with linear problems, we assume in the present section that the knots are fixed a priori, and we will consider approximation from the space of spline functions defined as follows. Let integers *m* and *k* be given, and let $a = x_0 < x_1 < \cdots < x_{k+1} = b$. Then

$$S_m = \{s \in C^{m-1}[a,b] : s(x) \in \Pi_m \text{ on } [x_i, x_{i+1}], i = 0, \dots, k\},\$$

where Π_m denotes the space of polynomials of degree *m*, is the space of polynomial splines of degree *m* with *k* fixed knots. S_m is a linear space with dimension m + k + 1. The first results on Chebyshev approximation by splines seem to be due to Johnson [96] in 1960.

The theory of approximation by Chebyshev sets does not apply to approximation from S_m . However, S_m is an example of a family of functions forming a *weak Chebyshev set*: any linear combination of such a set of *n* functions has at most (n - 1) changes of sign. For such sets Jones and Karlowitz [100] showed in 1970 that there exists at least one best Chebyshev approximation ϕ to any continuous function *f* which has the classical alternation property

$$\mathscr{A}(f-\phi)_{[a,b]} \ge n+1,$$

(although there may be others which do not).

The theory of Chebyshev approximation by splines with fixed knots is fully developed, and a characterization of best approximation goes back to the Ph.D. dissertation of Schumaker in 1965, and his publications over the next few years, e.g. [180]. Results were also given by Rice [173] in 1967. What is required is the existence of an interval $[x_p, x_{p+q}] \subset [a, b]$, with $q \ge 1$ such that there are at least q + m + 1 alternating extrema on $[x_p, x_{p+q}]$, or in the notation previously introduced

$$\mathscr{A}(f-s)_{[x_{p},x_{p+q}]} \ge q+m+1,$$

where $s \in S_m$. In addition to characterization of solutions, there has been interest in conditions for uniqueness (and strong uniqueness) of best approximations. In general of course, best approximations are not unique. However, the uniqueness (and strong uniqueness) of best spline approximations is characterized by the fact that all knot intervals contain sufficiently many alternating extrema as shown by Nürnberger and Singer [143] in 1982.

The solution of a discretized problem by linear programming techniques was suggested by Barrodale and Young [14] in 1966 and also by Esch and Eastman in an 1967 technical report (see their 1969 paper [62]). These methods do not make explicit use of characterization results, in contrast to the (explicit) Remez exchange method of Schumaker presented again in technical reports about the same time (see his 1969 paper [181]). The latter method also solved the discretized problem, but had no convergence results.

Of course any methods for best Chebyshev approximation by linear functions may be used, but a special iterative algorithm for computing best Chebyshev approximations from spline spaces was given by Nürnberger and Sommer [144] in 1983. As in the classical Remes method, a substep at each iteration is the computation of a spline $s \in S_m$ such that

$$(-1)^{i}(f(\xi_{i}) - s(\xi_{i})) = h, \quad i = 1, \dots, m + k + 2,$$

for some real number h, and given points $\xi_1, \ldots, \xi_{m+k+2}$ in [a, b]. The number of equations reflects the fact that S_m has dimension m + k + 1. Then one of the points ξ_i is replaced by a point where ||f-s|| is attained in [a, b] to get a new set of points $\{\xi_i\}$. The usual Remes exchange rule can result in a singular system of equations, so a modified exchange rule is needed. Such a rule was given by Nürnberger and Sommer [144], which ensures that the new system has a unique solution. Because of possible nonuniqueness of best approximations, the proof of convergence is fairly complicated. However, convergence can be established.

A multiple exchange procedure can also be implemented, and quadratic convergence is possible. The above results can be extended to more general spline spaces, where the polynomials are replaced by linear combinations of functions forming Chebyshev sets: this was considered by Nürnberger et al. [141] in 1985.

To permit the full power of splines, one should allow the knots to vary, rather than be fixed in advance. The corresponding approximation problem is then a difficult nonlinear problem and we say more about this in Section 3.7.

2.7. Linear L_1 approximation in C[a,b]

Given the same setting as at the start of Section 2.5, we consider here the problem

find
$$\boldsymbol{a} \in \mathbb{R}^n$$
 to minimize $\int_a^b \left| f(x) - \sum_{i=1}^n a_i \phi_i(x) \right| \, \mathrm{d}x.$ (13)

This problem was apparently first considered by Chebyshev in 1889.

Characterization results go back to James [93] in 1947. A convenient form is the analogue of that available in the discrete case: a vector a solves the L_1 problem if and only if there exists a function v with $||v||_{\infty} \leq 1$ such that

$$\int_{a}^{b} v(x)\phi_{j}(x) \,\mathrm{d}x = 0, \quad j = 1, \dots, n,$$
$$v(x) = \operatorname{sign} r(x), \quad r(x) \neq 0.$$

If the set $\{\phi_1(x), \ldots, \phi_n(x)\}$ forms a Chebyshev set in [a, b], then Jackson [92] in 1921 showed that the solution is unique. For polynomial approximation, perhaps the first "algorithm" was given by Hoel [89] in 1935, who showed that the polynomials of best L_p approximation converge to the best L_1 approximation as $p \to 1$. This is the analogue of the Polya algorithm for Chebyshev approximation. A more general convergence result, and a characterization of the limiting element, was given by Landers and Rogge [109] in 1981. The L_1 problem is greatly simplified if it can be assumed that the zeros of $f(x) - \sum_{i=1}^{n} a_i \phi_i(x)$ form a set of measure zero in the interval [a, b] (for example the zeros just consist of a finite set of points). Then the function to be minimized in (13) is differentiable, and necessary and sufficient conditions for a solution are that

$$\int_a^b g(x, \boldsymbol{a})\phi_j(x) = 0, \quad j = 1, \dots, n,$$

where g(x, a) denotes the sign of $f(x) - \sum_{i=1}^{n} a_i \phi_i(x)$. This was known to Laasonen [107] in 1949. This means that great store is placed on the points where there are sign changes, or equivalently where the approximation interpolates f. If these points were known, and were exactly n in number, then we could compute the best approximation by interpolation, *provided that there were no other changes of sign in the error of the resulting approximation*. The points $x_1 < \cdots < x_t \in (a, b) = (x_0, x_{t+1})$, where $1 \le t \le n$, are called *canonical points* if

$$\sum_{i=0}^{t} (-1)^{i} \int_{x_{i}}^{x_{i+1}} \phi_{j}(x) \, \mathrm{d}x = 0, \quad j = 1, \dots, n.$$
(14)

For the Chebyshev set case, Laasonen [107] in 1949 showed that there is a unique sign function and further t = n. This was extended to weak Chebyshev sets by Micchelli [129] in 1977. Existence of a set of $t \le n$ canonical points for the general problem was shown by Hobby and Rice [87] in 1965.

For the special case when $\phi_i(x) = x^{i-1}$, i = 1, ..., n, then the location of the *n* canonical points is known – they lie at the zeroes of the Chebyshev polynomial of the second kind of degree *n* (shifted if necessary). This result is due to Bernstein [23] in 1926. Thus their location is independent of *f*. Interpolation at these points can quite frequently result in the best polynomial approximation, for example, if the set

$\{f(x),\phi_1(x),\ldots,\phi_n(x)\}$

forms a Chebyshev set in [a,b]. However, this is not usually the case, and so this is not a reliable method in general.

An algorithm of descent type seems first to have been given by Usow [188] in 1967, who gave an analysis applicable to problems with Chebyshev sets, and some numerical results for polynomial approximation. However, Marti [120] in 1975 gave an example where the method converges to a point which is not a solution. He gave an alternative descent method, valid when the functions $\{\phi_i\}$ form a Markov set (any rearrangement is a Chebyshev set).

The first general method seems due to Glashoff and Schultz [72] in 1979, based on using Newton's method to solve the characterization conditions (14) together with the corresponding interpolation conditions. A variant of this, which is globally convergent, was given by Watson [200] in 1981. It is essentially of exchange type, based on the calculation of the zeroes of the error at each iteration and the construction of descent directions. It is also of Newton type, since it constructs the Hessian matrix of the error when it exists, and therefore can have a second-order convergence rate. In a sense, it can be thought of as analogous to the second algorithm of Remes for Chebyshev problems, where here a sequence of sets of zeroes plays the role of a sequence of sets of extreme points in that problem; the connection with Newton's method under appropriate circumstances is also something the methods have in common. A method for L_1 problems based on Newton's method was also given by Blatt [25] in 1984.

3. Nonlinear approximation

There are two major differences which arise in moving from linear to nonlinear best approximation problems. Firstly, existence of solutions cannot generally be guaranteed. Secondly, there is normally a gap between conditions which are necessary and conditions which are sufficient for a best approximation. This reflects the loss of convexity. From an algorithmic pont of view, it is usual to seek to satisfy first-order conditions which are necessary for a solution to the best approximation problem, and such a point is conventionally referred to as a *stationary point*. At best this can be expected to be a local minimum of the norm. Assuming that the members of the approximating family are differentiable with respect to the free parameters at least in the region of interest, then a characterization of stationary points is straightforward: it is appropriate simply to replace in the linear case the basis elements (either vectors making up the columns of A or functions $\phi_i, i=1,...,n$) by the partial derivatives of the approximating function with respect to the free parameters at the relevant points.

3.1. Nonlinear approximation in \mathbb{R}^m

Consider now the discrete problem

find $x \in \mathbb{R}^n$ to minimize ||f(x)||,

where $f \in \mathbb{R}^m$ depends nonlinearly on the components of x, and where the norm is any norm on \mathbb{R}^m .

A general approach to this problem is through a sequence of linear subproblems. Assume that f is continuously differentiable in the region of interest, and at a given point x, let A denote the $m \times n$ matrix of partial derivatives of the components of f with respect to the components of x. Then consider the iterative method based on computing an updated x as follows:

(i) find $\boldsymbol{d} \in \mathbb{R}^n$ to minimize $||\boldsymbol{f} + A\boldsymbol{d}||$,

(ii) replace x by $x + \gamma d$, where $\gamma > 0$ is suitably chosen.

The problem in (i) is just a linear approximation problem in the given norm (a linear subproblem), and (ii) involves choosing γ so that

$$||f(x+\gamma d)|| < ||f(x)||, \tag{15}$$

if this is possible: for example we may try to minimize the expression on the left-hand side with respect to γ .

When the norm is the least-squares norm, this kind of method (with $\gamma = 1$) most probably dates back to Gauss and is now known as the Gauss-Newton method. For the Chebyshev problem, this kind of approach was suggested by Zuhovickii et al. [212] in 1963, by Ishizaki and Watanabe [91] in 1968, and by Osborne and Watson [150] in 1969. Unless x is a stationary point, improvement can always be obtained via step (ii) since (15) holds for $\gamma > 0$ small enough. The theory given in the Osborne and Watson paper required that successive matrices A satisfied the Haar condition, and in that case convergence to a stationary point was established. The method was extended to the l_1 norm by Osborne and Watson [152] in 1971. Also in 1971, Osborne [146] was able to relax the Haar condition assumption for the l_{∞} algorithm, and showed that the method was quadratically convergent if the maximum error at the limit point of the iteration was attained at n + 1 points. In that case, unit length steps were ultimately possible, and a ready connection could be drawn with Newton's method applied to the nonlinear equations satisfied at the stationary point. Osborne contrasted this with the behaviour of the method in the l_2 case, when good performance was dependent on the goodness of fit of the model, rather than on properties of the data.

The behaviour of the algorithm in a completely general setting was considered in 1978 by Osborne and Watson [153]; in particular, (15) was always shown to hold for $\gamma > 0$ small enough away from a stationary point. It was also pointed out that the above behaviour typified the situation for polyhedral norms on the one hand, and smooth strictly convex monotonic norms on the other.

A common basis for a convergence analysis which included this kind of algorithm was given by Cromme [51] in 1978: he showed that for second-order convergence, it was sufficient for the best approximation to be strongly unique. This criterion was also studied for the above algorithms in 1980 by Jittorntum and Osborne [95], who showed that strong uniqueness was not always necessary.

Meantime, (at least) two developments were taking place. The fact that the solution of the linear subproblem could be such that very small step lengths were sometimes required led to the idea of explicitly incorporating bounds. This Levenberg–Marquardt or trust region idea was finding favour in descent methods for more general optimization calculations. Another development was to do with the line search. Trying to find the value of γ to minimise ||f|| is clearly impractical, and the idea of inexact, but sufficiently good, line searches was again imported from contemporary optimization algorithms. These modifications were used by Madsen [115] in an algorithm for the Chebyshev problem, and by Anderson and Osborne [4] in 1977 for polyhedral norm problems (which include l_1 and l_{∞}). While this could improve things in certain cases, slow convergence could, however, still occur for many problems.

For fast local convergence in general, it was recognized that second derivative information had to be incorporated. Two stage methods for Chebyshev problems were given independently in 1979 by Watson [198] and by Hald and Madsen [80]. These methods solved a sequence of linear subproblems to build up information about the limit point (in particular, the number of points where the norm was attained, with signs). This information could then (if necessary) be used as input to a second (locally convergent) phase such as Newton's method applied to the nonlinear system of equations characterizing a stationary point. Thus they extended fast local convergence to a much wider range of problems.

It had long been recognized that the Chebyshev approximation problem could be posed as a nonlinearly constrained optimization problem, analogous to the way in which the linear problem could, although it seemed at one time that treating the problem in this way was likely to be less efficient than using linear subproblems. However, following advances in techniques for constrained optimization problems, and a recognition that there was much structure in the Chebyshev problem which could be exploited, Conn [48] in 1979, Murray and Overton [135] in 1980, Han [82] in 1981, and Womersley and Fletcher [207] in 1986 all proposed methods. These are all variants of a technique based on the solution of a sequence of quadratic programming problems, involving a Lagrangian function and linearizations of $r_i = h$, for *i* in a set which estimates the set of indices where the extrema are attained at a stationary point. They all incorporate second derivative information, and involve exploiting the structure and giving descent with respect to the norm. A line search descent method due to Conn and Li [49] in 1989 is claimed to make more explicit use of the structure: in addition to giving descent, it attempts to force satisfaction of the stationary point characterization conditions at the same time.

This general approach now seems the most effective for small problems with dense matrices A. However, for large problems with sparse structure in A, solving linear rather than quadratic programming problems is preferable, as the structure may be exploited. Therefore, for such problems, there has been some recent re-interest in methods of trust region type which use sequential linear programming. Some work of Jonasson and Madsen [97,98] from the mid-1990s is of relevance here.

As in the linear case, large problems may arise as discretizations of continuous problems; therefore we will return to this in Section 3.4.

There were analogous developments for the solution of the nonlinear l_1 problem. The first attempt to incorporate second derivative information into general classes of problems was probably by McLean and Watson [122] in 1980. This method was of two-phase type which used the solution of a sequence of bounded variable linear subproblems to provide information about Z at the desired stationary point, and then used Newton's method to get an accurate point. The exact Jacobian matrix was used for the Newton step. A similar method by Hald and Madsen [81] in 1985 used quasi-Newton approximations, and allowed several switches between phases. Meantime, (single phase) methods based on solving a sequence of quadratic programming problems were being developed, analogous to those mentioned before for Chebyshev approximation problems. In the main, these constructed the quadratic programming problems by defining a Lagrangian function, and by involving linear approximations to $r_i = 0$ for $i \in Z^k$, where Z^k was an estimate at iteration k to Z at the solution. Methods of this type which used line searches were proposed by Murray and Overton [136] in 1981 and Bartels and Conn [15] in 1982, and trust region methods were given by Fletcher [67,68] in 1981 and 1985.

Perhaps because there is no simple connection analogous to that between continuous and discrete Chebyshev approximation problems, the nonlinear l_1 problem has attracted much less recent interest.

3.2. Rational Chebyshev approximation in \mathbb{R}^{t}

Approximation by rational functions goes back to Chebyshev in 1859. The basic (discrete) problem is as follows. Let x_i , i = 1, ..., t be in [a, b]. Then a best approximation is sought from the set

$$\mathbb{R}_{nm}^{D} = \left\{ P(x)/Q(x): P(x) = \sum_{j=0}^{n} a_{j} p_{j}(x), \quad Q(x) = \sum_{j=0}^{m} b_{j} q_{j}(x), \quad Q(x_{i}) > 0, \ i = 1, \dots, t \right\},$$

to the set of values f_1, \ldots, f_t , in the sense that

 $\max_{1 \leq i \leq t} \left| \mathbb{R}(x_i) - f_i \right|$

is minimized over all $\mathbb{R} \in \mathbb{R}_{nm}^{D}$. For this problem, existence of best approximations is not guaranteed, even in the case of quotients of polynomials, and characterization and uniqueness results are not available, although of course necessary conditions for a solution may be obtained. In fact necessary conditions based on alternations may be derived analogous to the characterization conditions which are available in the case of approximation to a continuous function on an interval: see Section 3.6. Because of this it is possible to implement an algorithm equivalent to the second algorithm of Remes, although for the discrete problem there are better approaches which do not explicitly use alternations.

The quest for algorithms for rational Chebyshev approximation appears to go back at least as far as Wenzl [201] in 1954. In the late 1950s Loeb considered some approaches which formed the basis for what was perhaps the first really effective algorithm for the discrete problem, the differential correction algorithm, given by Cheney and Loeb [38] in 1961. At that time, the convergence properties were uncertain, and a modified version was subsequently considered by the same authors in 1962 [39], and also by Cheney and Southard [42] in 1963, which was shown to have sure convergence properties, and drew attention away from the original method. However, in 1972 Barrodale et al. [9] studied both approaches, and showed that the method in its original form had not only guaranteed convergence from any starting approximation in \mathbb{R}_{nm}^D , but usually had a second-order convergence rate. Further, their comparisons of the methods showed that the performance of the original method was better. Some further analysis was given by Cheney and Powell [41] in 1987.

The differential correction algorithm is an iterative method where successive approximations from \mathbb{R}_{nm}^D are computed by solving a linear programming subproblem, where one variable is minimized subject to 2t linear constraints involving also variables representing the coefficients of the new approximation, and bound constraints on the coefficients of the denominator. Each step of the method may be interpreted as working with an approximation to the original problem which is correct up to first order, and this "Newton method" connection gives a partial explanation of the second-order convergence rate. In fact, from the point of view of implementation, it is more efficient to solve the dual of the original linear programming subproblem.

A potentially unsatisfactory feature of approximation from \mathbb{R}_{nm}^{D} is that the denominator, although positive, can become arbitrarily close to zero at certain points. It is not sufficient simply to impose a lower bound on Q, because of the possibility of multiplying both numerator and denominator by an arbitrary constant. A modification of the differential correction algorithm which applies to problems with a lower bound on the denominator and upper bounds on the absolute values of the coefficients b_j was given by Kaufmann and Taylor [104] in 1981.⁵ It is more natural, however, to impose upper and lower bounds on the denominators themselves ("constrained denominators"). A modified differential correction algorithm for this problem has been given by Gugat [75] in 1996. This involves constraints of the form

$$\mu(x_i) \leqslant Q(x_i) \leqslant v(x_i), \quad i = 1, \dots, t, \tag{16}$$

where μ and v are continuous functions, which replace the constraints on $Q(x_i)$ in the definition of \mathbb{R}_{nm}^D .

The linear programming subproblem corresponding to (16) above differs in that the additional conditions are imposed on the denominators. However, Gugat's method differs also in that there is greater flexibility in choice of initial values, and this turns out to be important. The original algorithm starts with an approximation \mathbb{R}_1 in \mathbb{R}_{nm}^D and a value Δ_1 which is the maximum modulus error of this approximation on the discrete set. The method of Gugat starts with \mathbb{R}_1 as usual, but with an arbitrary number Δ_1 that is allowed to be smaller than the current maximum error. This flexibility turns out to be an important advantage: for example numerical results show that the choice $\Delta_1 = 0$ is a good one. It is shown by Gugat that convergence results for the original version carry over.

⁵ This is an example of a constrained problem, which arises in a natural way from the rational approximation problem: it is not our intention to consider constrained problems per se.

It has been pointed out that the quadratic convergence of the differential correction algorithm is a consequence of a connection which it may be shown to have with Newton's method. Methods which set out deliberately to use variants of Newton's method are given by Hettich and Zenke [86] in 1990 and Gugat [76] in 1996. However, in contrast to the methods based on the differential correction algorithm, these do not generate a monotonically decreasing sequence of maximum modulus errors on successive approximations.

3.3. Nonlinear approximation in C[a,b]

Consider now the problem

find
$$\boldsymbol{a} \in \mathbb{R}^n$$
 to minimize $||f(., \boldsymbol{a})||,$ (17)

where the norm is a given norm on C[a, b] and where *a* occurs nonlinearly in *f*. It was shown by Watson [197] in 1978 that, provided that *f* was differentiable in the region of interest, methods of Gauss–Newton type (the continuous analogues of the methods introduced in Section 3.1) can be applied to this class of problems. However, while this may be of some theoretical interest, it does not lead to practical algorithms. Indeed, such problems cannot really be considered in any generality, and we will in fact restrict attention to the Chebyshev norm, and some important special cases.

3.4. Nonlinear Chebyshev approximation in C[a,b]

Here we consider (17) when the norm is the Chebyshev norm

 $||f||_{\infty} = \max_{x \in [a,b]} |f(x)|.$

Some general problems of this type were considered by Chebyshev in 1859 [36], with particular reference to rational approximation.

Aside from some special cases (for example see below) it is not possible to say very much about the number of points where the norm is attained at a stationary point. In common with other general nonlinear problems, characterization results are not available, and numerical methods set out to find a stationary point.

The first practical numerical methods seem to have been of two-phase type (see Section 2.5), and these were proposed independently by Hettich [83] and Watson [195] in 1976. The basic idea is similar to that used for linear problems: a first phase is to solve a discretized problem, whose solution identifies the number and associated signs, along with good approximations, of the points where the norm is attained at a stationary point, and a second phase corresponding to the solution of a nonlinear system comprising the equations to be satisfied there. Only the first-phase calculation needs a method which is tailored to whether the problem is linear or not. If a single discretized problem is to be solved, then any of the methods for solving discrete Chebyshev problems can of course be used.

The second phase calculation is a Newton type method, whose steps may be interpreted as quadratic programming problems. The approach can be globalized, thus extending the domain of convergence. This idea was central to the single phase method given by Jonasson and Watson [99] in the mid-1980s, based on the use of a Lagrangian function, and solving a sequence of quadratic programming problems defined on the current set of local maxima of the modulus of the error

function. Descent directions were defined, and both line search and trust region algorithms were developed. Second-order convergence is normal, and there is a nice connection with the second algorithm of Remes; however, the requirement to calculate exact local extrema at each step is a major disadvantage, and there can be sometimes slow progress far from a stationary point. A similar method was given by Jing and Fam [94] in 1987.

The connection between continuous Chebyshev approximation problems and semi-infinite programming problems has already been drawn, and the earlier comments apply to nonlinear problems. It may be that more recent methods being devised for nonlinear semi-infinite programming problems may also improve on these earlier methods for Chebyshev approximation problems. For example, a method by Görner [74] in 1997 consists of the solution of a finite set of discretized problems by sequential quadratic programming methods, following on from similar ideas used by Zhou and Tits [210] in 1996. These methods can lead into a second phase for accurate solution of the continuous problem: a feature of the method of Görner is that the same superlinearly convergent sequential quadratic programming method is used in both phases.

In any event, it would appear that this much at least can be said: a two-phase method with a discretization technique as first phase, and a variant of Newton's method as second phase, seems to be the most reliable and efficient method for solving small to medium size continuous Chebyshev set problems. However, the difficulties referred to near the end of Section 2.5 are still relevant for larger problems.

3.5. Nonlinear Chebyshev approximation in C[a,b]-some special cases

In order to close the gap between conditions which are necessary and conditions which are sufficient, it is necessary to restrict the class of approximating functions, and the point at which this process converges may conveniently be described in terms of alternation conditions, analogous to those which apply in the (linear) Chebyshev set case. This clearly has implications for numerical methods, and so it is appropriate to look briefly at some of this theory. In the linear case, the Chebyshev set condition simultaneously implies the existence of an interpolation function with a certain (fixed) number of zeros. In nonlinear cases, these become two conditions which have to be assumed separately: the interpolation property is a *local* one (which depends on the approximation), but in addition we require a global property on the zeros.

The concept of unisolvency was introduced in 1949 by Motzkin [130]. Let $\phi(x, a) : \mathbb{R}^n \to C[a, b]$. Then given any $d \in \mathbb{R}^n$ and *n* distinct points x_i , i = 1, ..., n in [a, b], this family is *unisolvent* if there exists a unique vector $a \in \mathbb{R}^n$ such that

 $\phi(x_i, \boldsymbol{a}) = d_i, \quad i = 1, \dots, n.$

This particular generalization of the Chebyshev set property in the linear case leads to the existence of a unique best approximation ϕ which is characterized by

$$\mathscr{A}(f-\phi)_{[a,b]} \ge n+1,$$

as shown by Tornheim [187] in 1950. Unfortunately this is an extremely restrictive property, possessed by a small number of approximating functions, and Rice in his Ph.D. thesis in 1959, and in papers published in the next few years, suggested a more general property of *varisolvency*, which (provided the error is not constant) leads to the best approximation ϕ being characterized by

$$\mathscr{A}(f-\phi)_{[a,b]} \ge m(\phi)+1,$$

where $m(\phi)$ is the degree of local solvency [168,170]. Rice also showed that there is at most one best approximation. If ϕ is formed from a linear combination of *n* functions forming a Chebyshev set in [*a*, *b*], then this is in fact a varisolvent family of constant degree *n*.

A related theory for nonlinear Chebyshev approximation on an interval was established by Meinardus and Schwedt [126] in 1964, valid for approximating functions differentiable with respect to their parameters. It essentially replaces the local condition required in varisolvency by a local Chebyshev set condition on the tangent space. An alternation characterization condition was established, along with an uniqueness result. Braess [31] in 1974 demonstrated the precise relationship between these various results.

Attempting to define a general class of nonlinear approximating functions which would be varisolvent, and so satisfy this kind of characterization result, Hobby and Rice [88] in 1967 defined γ -polynomials,

$$\phi(x, \boldsymbol{a}) = \sum_{i=1}^{n} a_i \gamma(a_{i+n} x),$$

where γ is a continuous function of its parameters. This class is of interest because it includes some important special cases, for example exponentials and spline functions. Subject to an additional assumption (Descartes' rule of signs), Hobby and Rice [88] established that the theory of varisolvent families applied. This condition is satisfied if the set

$$\{\gamma(t_1,x),\ldots,\gamma(t_n,x)\}$$

forms a Chebyshev set in [a, b] for distinct t_i 's. A best approximation ϕ is then characterized by

$$\mathscr{A}(f-\phi)_{[a,b]} \ge n+l(\phi)+1$$

where $l(\phi)$ is the length of the γ -polynomial ϕ , defined by the restriction that ϕ cannot be expressed by a sum of fewer terms. The closure of the set of γ -polynomials is in fact required for existence of best approximations, but then the alternating characterization is lost.

An important special case is given by taking

$$\gamma(t,x)=\mathrm{e}^{tx},$$

when we have approximation by sums of exponentials. This was studied first by Rice [169] in 1960 (n = 1), and in 1962 (general *n*) [171]. Because the set $\{e^{t_1x}, \ldots, e^{t_nx}\}$ forms a Chebyshev set in [a, b] for distinct t_i 's, then a Descartes' rule of signs holds (this result seems to go back to Laguerre [108] in 1898), and it follows that the approximating family is varisolvent. This was shown by Rice [171] in 1962, who also showed that a best approximation ϕ is characterized by

$$\mathscr{A}(f-\phi)_{[a,b]} \ge n+k(\phi)+1,$$

where the gradient vector of ϕ with respect to a_i , i = 1, ..., 2n has $n + k(\phi)$ nonzero components. There is at most one best approximation. Existence of best approximations from the closure of the set of exponential functions was proved by Rice [171] in 1962 and Werner [204,205] in 1969. As Bellman [21] wrote in 1970, "exponential approximation is a notoriously delicate enterprise", mainly because widely varying parameter values can give nearly optimal results. Therefore, the calculation of best Chebyshev approximations (or indeed any approximations) by sums of exponentials can be difficult. If an assumption is made about the number of alternations (that $k(\phi) = n$), then a method of Remes type can be applied with a nonlinear system of equations to be solved for the new coefficients at each iteration. This is considered by Dunham [57] in 1979, and in subsequent work with Zhu: it was necessary to have very good starting approximations.

The fact that *n* of the parameters occur linearly means that if the parameters a_{n+1}, \ldots, a_{2n} (the frequencies) are fixed, then the remaining parameters can be obtained by applying a linear solution method; this gives a problem which is essentially in the frequencies alone, and which could be tackled by iteration on the frequencies to obtain optimal values. Local descent methods were suggested by Braess [29] and Werner [205] in the late 1960s, and related methods were implemented in the 1970s by others such as Cromme, Kammler, Robitzsch and Schaback [50,101,177]. A method due to Dunham [59] in 1988 worked well with one frequency, but had difficulties with two or more. Nearly equal frequencies, or coalescing frequencies, are generally a problem.

One feature is that good initial approximations are necessary: in particular it is important to estimate the positions of the frequencies, before applying an optimal method, and this has led to interest in "suboptimal approximations". Prony's method of "approximate interpolation" may be applied, although the method is not generally stable. An alternative is Bellman's 1970 [21] method of differential approximation. These methods were considered in detail by Robitzsch and Schaback [177] in 1978 and by Schaback [179] in 1979. Any suboptimal method may be considered as a first phase method which can lead into a second phase based on Newton's method to satisfy the nonlinear system characterizing the solution.

But it would seem that in practice additional constaints are both natural physically, and necessary mathematically and computationally – for example, to bound frequencies, or to prevent frequencies from crossing each other. The computational approach then depends on precisely what is being assumed, and we will not pursue this further.

3.6. Rational Chebyshev approximation in C[a,b]

The continuous analogue of the class of problems considered in Section 3.2 is based on the approximating set \mathbb{R}_{nm} defined by

$$\mathbb{R}_{nm} = \left\{ P(x)/Q(x) : P(x) = \sum_{j=0}^{n} a_j p_j(x), \\ Q(x) = \sum_{j=0}^{m} b_j q_j(x), \ Q(x) > 0 \text{ on } [a,b] \right\},$$

where the $p_j(x)$ and $q_j(x)$ are given sets of functions. Then given $f(x) \in C[a,b]$, we require to determine $\mathbb{R} \in \mathbb{R}_{nm}$ to minimize $||f - \mathbb{R}||$, where the norm is the Chebyshev norm on [a,b]. For the special case when P(x) and Q(x) are polynomials of degree n and m, respectively, existence of a best approximation is guaranteed, as shown by Walsh [193] in 1931. Achieser in 1947 (see his 1956 book [3]) showed that the best approximation is unique (up to a normalization), and earlier,

in 1930 (again see his 1956 book [3]), he showed that a best approximation $\mathbb{R} = P/Q \in \mathbb{R}_{nm}$ is characterized by

$$\mathscr{A}(f-\mathbb{R})_{[a,b]} \ge n+m+2-d(\mathbb{R}),$$

where $d(\mathbb{R})$ is the *defect* of the approximation: the defect is just the minimum difference between the *actual* degree of P(x) and Q(x) and *n* and *m* respectively. If $d(\mathbb{R}) > 0$, the best approximation is said to be degenerate. These results also follow from the fact that the approximating family is varisolvent. (The necessary conditions referred to in Section 3.2 correspond to this alternation result defined *on the points of the set* $x_1 < \cdots < x_t$ introduced there.)

For more general quotients (of linear combinations of functions), existence is no longer guaranteed, although characterization results are available (not necessarily of alternation type), and uniqueness results may be extended. The main contributions here are from Cheney and Loeb [39,40] in the mid-1960s.

For rational approximation by quotients of polynomials on an interval, the analogue of the Remes exchange method may be applied, using sets of m+n+2 points. It, therefore, requires nondegeneracy of the best approximation, and can converge at a second-order rate if started from close enough to the solution: the analysis is primarily due to Werner in a series of papers in the early 1960s [202,203]. The system of linear equations which needs to be solved in the linear problem is replaced by a nonlinear system in the rational problem, equivalent to an eigenvalue problem. Werner [203] in 1963 showed that the eigenvalues are always real, and there is at most one pole free solution, that is a rational approximation with Q(x) > 0 on [a, b]. Maehly in 1963 [118] gave an example which showed that in fact no pole free solution need exist; even if it is does exist, it need not be associated with the smallest eigenvalue. Despite these potential problems, the second algorithm of Remes has been successfully used for rational approximation. Fraser and Hart [70] in 1962, Werner [202] in 1962 and Stoer [186] in 1964 gave methods based on solving the system of nonlinear equations directly. In 1966, Curtis and Osborne [52] gave an algorithm which used the eigenvalue connection explicitly, solving the eigenvalue problem by inverse iteration with zero as an initial estimate for the eigenvalue; they also established quadratic convergence. Breuer [32] in 1987 suggested a different direct approach to this subproblem which used continued fraction interpolation, and which it was claimed can lead to a considerable increase in efficiency, and also accuracy and robustness.

Variants of the second algorithm of Remes apply to rational Chebyshev approximation problems which incorporate a generalized weight function. Important work involving rational approximation on an interval to provide optimal starting values for computing \sqrt{x} by the Newton Raphson method was done, for example, by Moursand [133] in the late 1960s.

The algorithms fail if the solution is degenerate, and indeed for problems which are nearly degenerate, extremely good starting approximations are required. Ralston [160,161] in a series of papers in the late 1960s and early 1970s considered degeneracy in detail. The computation of nearly degenerate approximations should if possible be avoided, as equally good results can be obtained through the use of smaller m and n.

It is possible to make the second algorithm of Remes more robust, by combining its merits with the differential correction algorithm. In particular the discrete subproblems can be solved by that method, and if no pole-free solution is obtained, additional points can be included. If sufficiently many points are taken in [a, b], and always assuming that the continuous problem is not degenerate, then a pole-free solution can be obtained so that the algorithm can be continued. Methods based

on this idea were given by Belogus and Liron [22] and also Kaufman et al. [103] both in 1978. Numerical evidence is that such an approach can be successful for problems which give difficulties with the traditional Remes method.

The differential correction algorithm may be applied to problems defined on an interval, although the subproblems are no longer finite. Dua and Loeb [55] in 1973 established a second order convergence rate if the best approximation is normal. The potentially unsatisfactory feature referred to in Section 3.2 where the deniminator, although positive, can become arbitrarily close to zero, also applies to \mathbb{R}_{nm} . The algorithm of Gugat referred to there also may be applied to intervals, although the numerical performance is unclear.

3.7. Chebyshev approximation by spline functions with free knots

To permit the full power of splines, one should allow the knots to vary, rather than be fixed in advance. The corresponding approximation problem is then a difficult nonlinear problem. This problem can be considered in terms of γ polynomials. However, the structure of the problem, and the way in which degeneracies can be introduced makes an attempt to make a straightforward application unhelpful.

To guarantee existence of best approximations, multiple knots have to be allowed. There may be local solutions; a characterization of best approximations is not known. For the case of k free knots, necessary and (different) sufficient conditions of the alternation kind given above may be proved. Let q' denote the sum of the knot multiplicities at the points $x_{p+1}, \ldots, x_{p+q-1}$. Then it is *necessary* for $s \in S_m$ to be a best Chebyshev approximation with k free knots to f in [a, b] that there exists an interval $[x_p, x_{p+q}] \subset [a, b]$ with $q \ge 1$ such that

$$\mathscr{A}(f-s)_{[x_p,x_{p+q}]} \ge m+q+q'+1,$$

as shown by Nürnberger et al. [142] in 1989; it is *sufficient* for $s \in S_m$ to be a best Chebyshev approximation with k free knots to f in [a,b] that there exists an interval $[x_p, x_{p+q}] \subset [a,b]$ with $q \ge 1$ such that

$$\mathscr{A}(f-s)_{[x_{p},x_{p+q}]} \ge m+k+q'+2,$$

as shown by Braess [30] in 1971. The necessary condition was strengthened to a possibly longer alternant by Mulansky [134] in 1992. Although a characterization of best spline approximations with free knots is not known, a characterization of strongly unique best spline approximations with free simple knots is available: what is required is that *all* knot intervals contain sufficiently many alternating extrema. The relevant work here is by Nürnberger [137,138] in 1987 and 1994.

Since approximation by splines with free knots is a nonlinear Chebyshev approximation problem, of course general methods can be used. However, the way in which the knots enter as free parameters makes this a particularly awkward problem and makes it important that the special structure be exploited.

For a discretization of the problem, a descent method based on Newton's method was given by Esch and Eastman [62] in 1969. Most algorithmic work has been concerned with uniform approximation from a space of piecewise polynomials where the continuity conditions at the knots are relaxed. A standard algorithmic approach is based on so-called segment approximation, originating from work of Lawson in 1964 [112], and methods were proposed by Pavlidis and Maika [155] in

1974, and McLaughlin and Zacharski [121] in 1984. Because pieces were fitted separately, continuity could be lost between segments. A recent method of this type is due to Nürnberger et al. [145] in 1986 (see also [124]). The algorithm converges through sequences of knot sets from an arbitrary set of knots. For each set of k knots, best Chebyshev degree m polynomial approximations to f are obtained on each subinterval using the classical Remes algorithm. The knots are then adjusted by a "levelling" process, so that the maximum errors of the polynomial best approximations are equalized. The result of this is a piecewise polynomial which is usually discontinuous. However, the procedure is augmented by the application of the method for fitting splines with fixed knots to the optimal knot positions obtained from the first part. The outcome of this is a differentiable spline approximation, which numerical results show to be a good one. Note that at present there is no algorithm for computing (global) best Chebyshev spline approximations with free knots. At best a local approximation can be expected, so producing a "good" spline approximation may be the most sensible strategy.

Generalizations to multivariate splines have mainly been concerned with interpolation problems. But consider bivariate splines on $[a_1, b_1] \times [a_2, b_2]$. This region can be divided into rectangles by knot lines $x = x_i$, $y = y_i$, i = 1, ..., s, and a tensor product spline space can be defined. As in the univariate problem, partitions can be defined and improved systematically in such a way that best Chebyshev approximations are obtained in the limit. Some recent work on this problem is given by Meinardus et al. [125] in 1996, and by Nürnberger [140] in 1997. However, there are many unsolved problems, as pointed out by Nürnberger [139] in 1996.

Acknowledgements

Charles Dunham attributes to P. Whippey, on the effort of writing a history: "It's a complex problem: the costs are real, the benefits imaginary". In any event, I am grateful to the many people who took the trouble to answer my questions about aspects of this work, or provided me with references. I am especially grateful to Mike Osborne, because he read through a draft and provided me with many helpful comments and suggestions. Of course, the responsibility for what is set down here is entirely mine.

Mike Osborne in fact has to shoulder quite a lot of responsibility for the existence of this paper, because he introduced me to the subject of approximation in 1964, he supervised my postgraduate study, and generally he has remained since that time a strong and guiding influence on my career. I dedicate this work to him, on the occasion of his 65th birthday, with my gratitude and my affection.

References

- [1] N.N. Abdelmalik, Linear L₁ approximation for a discrete point set and L₁ solutions of overdetermined linear equations, J. Assoc. Comput. Mach. 18 (1971) 41–47.
- [2] N.N. Abdelmalik, An efficient method for the discrete linear L_1 approximation problem, Math. Comp. 29 (1975) 844–850.
- [3] N.I. Achieser, Theory of Approximation, Ungar, New York, 1956.
- [4] D.H. Anderson, M.R. Osborne, Discrete, nonlinear approximation problems in polyhedral norms, Numer. Math. 28 (1977) 143-156.
- [5] D.O. Andreassen, G.A. Watson, Linear Chebyshev approximation without Chebyshev sets, BIT 16 (1976) 349-362.

- [6] R.D. Armstrong, J. Godfrey, Two linear programming algorithms for the linear discrete L_1 norm problem, Math. Comp. 33 (1979) 289–300.
- [7] R.D. Armstrong, D.S. Kung, A dual method for discrete Chebyshev curve fitting, Math. Programming 19 (1979) 186–199.
- [8] I. Barrodale, C. Phillips, Algorithm 495: solution of an overdetermined system of linear equations in the Chebyshev norm, ACM Trans. Math. Software 1 (1975) 264–270.
- [9] I. Barrodale, M.J.D. Powell, F.D.K. Roberts, The differential correction algorithm for rational l_{∞} approximation, SIAM J. Numer. Anal. 9 (1972) 493–504.
- [10] I. Barrodale, F.D.K. Roberts, Applications of mathematical programming to l_p approximation, in: J.B. Rosen, O.L. Mangasarian, K. Ritter (Eds.), Nonlinear Programming, Academic Press, New York, 1970, pp. 447–464.
- [11] I. Barrodale, F.D.K. Roberts, An improved algorithm for discrete l_1 linear approximation, SIAM J. Numer. Anal. 10 (1973) 839–848.
- [12] I. Barrodale, F.D.K. Roberts, Algorithm 478: solution of an overdetermined system of equations in the L_1 norm, Comm. ACM 17 (1974) 319–320.
- [13] I. Barrodale, A. Young, Algorithms for best L_1 and L_{∞} linear approximation on a discrete set, Numer. Math. 8 (1966) 295–306.
- [14] I. Barrodale, A. Young, A note on numerical procedures for approximation by spline functions, Comput. J. 9 (1966) 318–320.
- [15] R. Bartels, A.R. Conn, An approach to nonlinear l₁ data fitting, in: J.P. Hennart (Ed.), Numerical Analysis, Cocoyoc, 1981, Springer, Berlin, 1982, pp. 48–58.
- [16] R. Bartels, A.R. Conn, C. Charalambous, On Cline's direct method for solving overdetermined linear systems in the l_{∞} sense, SIAM J. Numer. Anal. 15 (1978) 255–270.
- [17] R. Bartels, A.R. Conn, Y. Li, Primal methods are better than dual methods for solving overdetermined linear systems in the l_{∞} sense, SIAM J. Numer. Anal. 26 (1989) 693–726.
- [18] R. Bartels, A.R. Conn, J.W. Sinclair, Minimization techniques for piecewise differentiable functions: the l_1 solution to an overdetermined linear system, SIAM J. Numer. Anal. 15 (1978) 224–241.
- [19] R. Bartels, G.H. Golub, Stable numerical methods for obtaining the Chebyshev solution to an overdetermined system of equations, Comm. ACM. 11 (1968) 401–406.
- [20] A.E. Beaton, J.W. Tukey, The fitting of power series, meaning polynomials, illustrated on band-spectrographic data, Technometrics 16 (1974) 147–185.
- [21] R. Bellmann, Methods of Nonlinear Analysis I, Academic Press, New York, 1970.
- [22] D. Belogus, N. Liron, DCR2: an improved algorithm for l_{∞} rational approximation, Numer. Math. 31 (1978) 17–29.
- [23] S.N. Bernstein, Lecons sur les Propriétés Extrémales et la Meillure Approximation des Fonctions Analytiques d'une Variable Réelle, Gauthier-Villars, Paris, 1926.
- [24] L. Bittner, Das Austauschverfahren der linearen Tschebyscheff-Approximation bei nicht erfullter Haarscher Bedingung, Z. Angew. Math. Mech. 41 (1961) 238–256.
- [25] H.-P. Blatt, Exchange algorithms, error estimates and strong unicity in convex programming and Chebyshev approximation, in: S.P. Singh, J.W.H. Burry, B. Watson (Eds.), Approximation Theory and Spline Functions, Reidel, Dodrecht, 1984, pp. 1–41.
- [26] P. Bloomfield, W.L. Steiger, Least Absolute Deviations, Birkhaüser, Boston, 1983.
- [27] P.T. Boggs, A new algorithm for the Chebyshev solution of overdetermined linear systems, Math. Comp. 28 (1974) 203–217.
- [28] E. Borel, Lecons sur les Fonctions de Variables Reélles, Gauthier-Villars, Paris, 1905.
- [29] D. Braess, Approximation mit Exponentialsummen, Computing 2 (1967) 309–321.
- [30] D. Braess, Chebyshev approximation by spline functions with free knots, Numer. Math. 17 (1971) 357–366.
- [31] D. Braess, Geometrical characterization for nonlinear uniform approximation, J. Approx. Theory 11 (1974) 260–274.
- [32] P.T. Breuer, A new method for real rational uniform approximation, in: J.C. Mason, M.G. Cox (Eds.), Algorithms for Approximation, Clarendon Press, Oxford, 1987, pp. 265–283.
- [33] P. Butzer, F. Jongmans, P.L. Chebyshev (1821–1894). A guide to his life and work, J. Approx. Theory 96 (1999) 11–138.

- [34] C. Carasso, P.J. Laurent, An algorithm of successive minimization in convex programming, R. A. I. R. O. Numer. Anal. 12 (1978) 377–400.
- [35] A. Charnes, W.W. Cooper, R.O. Ferguson, Optimal estimation of executive compensation by linear programming, Management Sci. 1 (1955) 138–151.
- [36] P.L. Chebyshev, Sur les questions de minima qui se rattachent à la representation approximative des fonctions, Oeuvres I (1859) 273–378.
- [37] E.W. Cheney, A.A. Goldstein, Newton's method for convex programming and Tchebycheff approximation, Numer. Math. 1 (1959) 253–268.
- [38] E.W. Cheney, H.L. Loeb, Two new algorithms for rational approximation, Numer. Math. 3 (1961) 72–75.
- [39] E.W. Cheney, H.L. Loeb, On rational Chebyshev approximation, Num. Math. 4 (1962) 124–127.
- [40] E.W. Cheney, H.L. Loeb, Generalized rational approximation, SIAM J. on Num. Anal. 1 (1964) 11-25.
- [41] E.W. Cheney, M.J.D. Powell, The differential correction algorithm for generalized rational functions, Constr. Approx. 3 (1987) 249–256.
- [42] E.W. Cheney, T.H. Southard, A survey of methods for rational approximation, with particular reference to a new method based on a formula of Darboux, SIAM Rev. 5 (1963) 219–231.
- [43] J.F. Claerbout, F. Muir, Robust modelling with erratic data, Geophysics 38 (1973) 826–844.
- [44] D.I. Clark, The mathematical structure of Huber's M-estimator, SIAM J. Sci. Statist. Comput. 6 (1985) 209-219.
- [45] A.K. Cline, A descent method for the uniform solution to overdetermined systems of linear equations, SIAM J. Numer. Anal. 13 (1976) 293–309.
- [46] T.F. Coleman, Y. Li, A globally and quadratically convergent affine scaling algorithm for l_1 problems, Math. Programming 56 (1992) 189–222.
- [47] T.F. Coleman, Y. Li, A globally and quadratically convergent method for linear l_{∞} problems, SIAM J. Numer. Anal. 29 (1992) 1166–1186.
- [48] A.R. Conn, An efficient second order method to solve the (constrained) minimax problem, Report CORR 79-5, University of Waterloo, 1979.
- [49] A.R. Conn, Y. Li, An efficient algorithm for nonlinear minimax problems, Research Report CS-88-41, University of Waterloo, 1989.
- [50] L. Cromme, Eine Klasse von Verfahren zur Ermittlung bester nichtlinearen Tschebyscheff-Approximationen, Numer. Math. 25 (1976) 447–459.
- [51] L. Cromme, Strong uniqueness: a far reaching criterion for the convergence of iterative processes, Numer. Math. 29 (1978) 179–193.
- [52] A.R. Curtis, M.R. Osborne, The construction of minimax rational approximations to functions, Comput. J. 9 (1966) 286–293.
- [53] M. Davies, Linear approximation using the criterion of least total deviations, J. Roy. Statist. Soc. Ser. B 29 (1967) 101–109.
- [54] J. Descloux, Approximations in L^p and Chebyshev approximations, SIAM J. 11 (1963) 1017–1026.
- [55] S.N. Dua, H.L. Loeb, Further remarks on the differential correction algorithm, SIAM J. Numer. Anal. 10 (1973) 123–126.
- [56] A.M. Duarte, R.J. Vanderbei, Interior point algorithms for lsad and lmad estimation, Technical Report SOR-94-07, Programs in Operations Research and Statistics, Princeton University, 1994.
- [57] C.B. Dunham, Chebyshev approximation by exponential-polynomial sums, J. Comput. Appl. Math. 5 (1979) 53-57.
- [58] C.B. Dunham, The weakened first algorithm of Remes, J. Approx. Theory 31 (1981) 97–98.
- [59] C.B. Dunham, Approximation with one (or few) parameters nonlinear, J. Comput. Appl. Math. 21 (1988) 115–118.
- [60] R. Dutter, Robust regression: different approaches to numerical solutions and algorithms, Research Report No 6, Technische Hochschule, Zurich, 1975.
- [61] F.Y. Edgeworth, On a new method of reducing observations relating to several quantities, Philos. Mag. 5 (1888) 185–191.
- [62] R.E. Esch, W.L. Eastman, Computational methods for best spline function approximation, J. Approx. Theory 2 (1969) 85–96.
- [63] R.W. Farebrother, The historical development of the L_1 and L_{∞} estimation procedures 1793–1930, in: Y. Dodge (Ed.), Statistical Data Analysis Based on the L_1 Norm and Related Methods, North-Holland, Amsterdam, 1987, pp. 37–63.

- [64] R.W. Farebrother, Fitting Linear Relationships: A History of the Calculus of Observations 1750–1900, Springer, Berlin, 1999.
- [65] J. Fischer, The convergence of the best discrete linear L_p approximation as $p \rightarrow 1$, J. Approx. Theory 39 (1983) 374–385.
- [66] W.D. Fisher, A note on curve fitting with minimum deviations by linear programming, J. Amer. Statist. Assoc. 56 (1961) 359–362.
- [67] R. Fletcher, Numerical experiments with an exact l₁ penalty function method, in: O.L. Mangasarian, R.R. Meyer, S.M. Robinson (Eds.), Nonlinear Programming 4, Academic Press, New York, 1981, pp. 99–129.
- [68] R. Fletcher, An l₁ penalty method for nonlinear constraints, in: P.T. Boggs, R.H. Byrd, R.B. Schnabel (Eds.), Numerical Optimization 1984, SIAM Publications, Philadelphia, 1985, pp. 26–40.
- [69] R. Fletcher, J.A. Grant, M.D. Hebden, Minimax approximation as the limit of best L_p approximation, SIAM J. Numer. Anal. 11 (1974) 123–136.
- [70] W. Fraser, J.F. Hart, On the computation of rational approximations to continuous functions, Comm. ACM 5 (1962) 401–403.
- [71] K. George, M.R. Osborne, The efficient computation of linear rank statistics, J. Comput. Simul. 35 (1990) 227–237.
- [72] K. Glashoff, R. Schultz, Uber die genaue Berechnung von besten L¹-Approximierenden, J. Approx. Theory 25 (1979) 280–293.
- [73] G.H. Golub, L.B. Smith, Algorithm 414: Chebyshev approximation of continuous functions by a Chebyshev system, Comm. ACM 14 (1971) 737–746.
- [74] S. Görner, Ein Hybridverfahren zur Losung nichtlinearer semi-infiniter Optimierungsprobleme, Ph.D. Thesis, Technical University of Berlin, 1997.
- [75] M. Gugat, An algorithm for Chebyshev approximation by rationals with constrained denominators, Constr. Approx. 12 (1996) 197–221.
- [76] M. Gugat, The Newton differential correction algorithm for uniform rational approximation with constrained denominators, Numer. Algorithms 13 (1996) 107–122.
- [77] S.-A. Gustafson, On the computational solution of a class of generalized moment problems, SIAM J. Numer. Anal. 7 (1970) 343–357.
- [78] S.-A. Gustafson, K. Kortanek, Numerical treatment of a class of semi-infinite programming problems, Naval. Res. Logist. Quart. 20 (1973) 477–504.
- [79] A. Haar, Die Minkowskische Geometrie und die Annaherung an stetige Funktionen, Math. Ann. 78 (1918) 294-311.
- [80] J. Hald, K. Madsen, A two-stage algorithm for minimax optimization, in: A. Bensoussan, J. Lions (Eds.), International Symposium on Systems Optimization and Analysis, Springer, Berlin, 1979, pp. 225–239.
- [81] J. Hald, K. Madsen, Combined LP and quasi-Newton methods for nonlinear l_1 optimization, SIAM J. Numer. Anal. 22 (1985) 68–80.
- [82] S.-P. Han, Variable metric methods for minimizing a class of nondifferentiable functions, Math. Prog. 20 (1981) 1–13.
- [83] R. Hettich, A Newton method for nonlinear Chebyshev approximation, in: R. Schaback, K. Scherer (Eds.), Approximation Theory, Springer, Berlin, 1976, pp. 222–236.
- [84] R. Hettich, Numerical methods for nonlinear Chebyshev approximation in: G. Meinardus (Ed.), Approximation in Theorie und Praxis, B.I.-Wissenschaftsverlag, Mannheim, 1979, pp. 139–156.
- [85] R. Hettich, An implementation of a discretization method for semi-infinite programming, Math. Programming 34 (1986) 354–361.
- [86] R. Hettich, P. Zenke, An algorithm for general restricted rational Chebyshev approximation, SIAM J. Numer. Anal. 27 (1990) 1024–1033.
- [87] C.R. Hobby, J.R. Rice, A moment problem in L_1 -approximation, Proc. Amer. Math. Soc. 65 (1965) 665–670.
- [88] C.R. Hobby, J.R. Rice, Approximation from a curve of functions, Arch. Rational Mech. Anal. 27 (1967) 91-106.
- [89] P.G. Hoel, Certain problems in the theory of closest approximation, Amer. J. Math. 57 (1935) 891–901.
- [90] M.J. Hopper, M.J.D. Powell, A technique that gains speed and accuracy in the minimax solution of overdetermined linear equations, in: J.R. Rice (Ed.), Mathematical Software III, Academic Press, New York, 1977.
- [91] Y. Ishizaki, H. Watanabe, An iterative Chebyshev approximation method for network design, IEEE Trans. Circuit Theory 15 (1968) 326–336.
- [92] D. Jackson, Note on a class of polynomials of approximation, Trans. Amer. Math. Soc. 13 (1921) 320-326.
- [93] R.C. James, Orthogonality and linear functionals in normed linear spaces, Trans. Amer. Math. Soc. 61 (1947) 265–292.
- [94] Z. Jing, A.T. Fam, An algorithm for computing continuous Chebyshev approximations, Math. Comp. 48 (1987) 691–710.
- [95] K. Jittorntrum, M.R. Osborne, Strong uniqueness and second order convergence in nonlinear discrete approximation, Numer. Math. 34 (1980) 439–455.
- [96] R.S. Johnson, On monosplines of least deviation, Trans. Amer. Math. Soc. 96 (1960) 458-477.
- [97] K. Jonasson, A projected conjugate gradient method for sparse minimax problems, Numer. Algorithms 5 (1993) 309–323.
- [98] K. Jonasson, K. Madsen, Corrected sequential linear programming for sparse minimax optimization, BIT 34 (1994) 372–387.
- [99] K. Jonasson, G.A. Watson, A Lagrangian method for multivariate continuous Chebyshev approximation problems, in: W. Schempp, K. Zeller (Eds.), Multivariate Approximation Theory 2, I.S.N.M. 61, Birkhaüser, Basel, 1982, pp. 211–221.
- [100] R.C. Jones, L.A. Karlovitz, Equioscillation under nonuniqueness in the approximation of continuous functions, J. Approx. Theory 3 (1970) 138–145.
- [101] D.W. Kammler, Chebyshev approximation of completely monotonic functions by sums of exponentials, SIAM J. Numer. Anal. 13 (1976) 761–774.
- [102] N. Karmarker, A new polynomial time algorithm for linear programming, Combinatorica 4 (1984) 373–395.
- [103] E.H. Kaufmann, D.J. Leeming, G.D. Taylor, A combined Remes-Differential Correction algorithm for rational approximation, Math. Comp. 32 (1978) 233–242.
- [104] E.H. Kaufmann, G.D. Taylor, Uniform approximation by rational functions having restricted denominators, J. Approx. Theory 32 (1981) 9–26.
- [105] J.E. Kelley Jr., An application of linear programming to curve fitting, SIAM J. 6 (1958) 15-22.
- [106] P. Kirchberger, Uber Tchebychefsche Annaherungsmethoden, Math. Ann. 57 (1903) 509–540.
- [107] P. Laasonen, Einige Satze uber Tschebyscheffsche Funktionensysteme, Ann. Acad. Sci. Fenn. Ser. AI 52 (1949) 3–24.
- [108] E. Laguerre, Ouvres I, Gauthier-Villars, Paris, 1898.
- [109] D. Landers, L. Rogge, Natural choice of L_1 -Approximants, J. Approx. Theory 33 (1981) 268–280.
- [110] P.S. Laplace, Mechanique Celeste, Tome 111, No 39, 1799.
- [111] C.L. Lawson, Contributions to the Theory of Linear Least Maximum Approximation, Ph.D. Dissertation, University of California, Los Angeles, 1961.
- [112] C.L. Lawson, Characteristic properties of the segmented rational minimax approximation problem, Num. Math. 6 (1964) 293–301.
- [113] W. Li, J.J. Swetits, The linear l_1 estimator and Huber M-estimator, SIAM J. on Optim. 8 (1998) 457–475.
- [114] Y. Li, A globally convergent method for L_p problems, SIAM J. Optim. 3 (1993) 609–629.
- [115] K. Madsen, An algorithm for minimax solution of over-determined systems of non-linear equations, J. Inst. Math. Appl. 16 (1975) 321–328.
- [116] K. Madsen, H.B. Neilsen, A finite smoothing algorithm for linear l_1 estimation, SIAM J. Optim. 3 (1993) 223–235.
- [117] K. Madsen, H.B. Nielson, M.C. Pinar, New characterizations of l_1 solutions of overdetermined linear systems, Operations Research Letters 16 (1994) 159–166.
- [118] H.J. Maehly, Methods of fitting rational approximations, Part II, J. Assoc. Comput. Mach. 10 (1963) 257-266.
- [119] J.C. Mairhuber, On Haar's theorem concerning Chebyshev approximation problems having unique solutions, Proc. Amer. Math. Soc. 7 (1956) 609–615.
- [120] J.T. Marti, A method for the numerical computation of best L_1 approximations of continuous functions, Proceedings of Oberwolfach Conference, ISNM, Vol. 26, Birkhauser, Basel, 1975.
- [121] H.W. McLaughlin, J.J. Zacharski, Segmented approximation, in: E.W. Cheney (Ed.), Approximation Theory III, Academic Press, New York, 1980, pp. 647–654.
- [122] R.A. McLean, G.A. Watson, Numerical methods for nonlinear discrete L_1 approximation problems, Proceedings of Oberwolfach Conference on Numerical Methods in Approximation Theory 1979, ISNM, Vol. 52, Birkhauser, Basel, 1980.
- [123] S. Mehrotra, On the implementation of a primal-dual interior point method, SIAM J. Optim. 2 (1992) 575-601.

- [124] G. Meinardus, G. Nürnberger, M. Sommer, H. Strauss, Algorithms for piecewise polynomials and splines with free knots, Math. Comp. 53 (1989) 235–247.
- [125] G. Meinardus, G. Nürnberger, G. Walz, Bivariate segment approximation and splines, Adv. Comput. Math. 6 (1996) 25–45.
- [126] G. Meinardus, D. Schwedt, Nicht-lineare Approximationen, Arch. Rational Mech. Anal. 17 (1964) 297-326.
- [127] M.S. Meketon, Least absolute value regression, Technical Report, AT&T Bell Laboratories, Murray Hill, New Jersey, 1987.
- [128] G. Merle, H. Späth, Computational experience with discrete L_p approximation, Computing 12 (1974) 315–321.
- [129] C.A. Micchelli, Best L^1 approximation by weak Chebyshev systems and the uniqueness of interpolating perfect splines, J. Approx. Theory 19 (1977) 1–14.
- [130] T.S. Motzkin, Approximation by curves of a unisolvent family, Bull. Amer. Math. Soc. 55 (1949) 789–793.
- [131] T.S. Motzkin, J.L. Walsh, Least pth power polynomials on a real finite point set, Trans. Amer. Math. Soc. 78 (1955) 67–81.
- [132] T.S. Motzkin, J.L. Walsh, Least *pth* power polynomials on a finite point set, Trans. Amer. Math. Soc. 83 (1956) 371–396.
- [133] D.G. Moursund, Optimal starting values for Newton-Raphson calculation of \sqrt{x} , Comm. ACM 10 (1967) 430–432.
- [134] B. Mulansky, Chebyshev approximation by spline functions with free knots, IMA J. Numer. Anal. 12 (1992) 95–105.
- [135] W. Murray, M.L. Overton, A projected Lagrangian algorithm for nonlinear minimax optimization, SIAM J. Sci. Statist. Comput. 1 (1980) 345–370.
- [136] W. Murray, M.L. Overton, A projected Lagrangian algorithm for nonlinear l₁ optimization, SIAM J. Sci. Statist. Comput. 2 (1981) 207–214.
- [137] G. Nürnberger, Strongly unique spline approximation with free knots, Constr. Approx. 3 (1987) 31-42.
- [138] G. Nürnberger, Approximation by univariate and bivariate splines, in: D. Bainov, V. Covachev (Eds.), Second International Colloquium on Numerical Analysis, VSP, Utrecht, 1994, pp. 143–153.
- [139] G. Nürnberger, Bivariate segment approximation and free knot splines: research problems 96-4, Constr. Approx. 12 (1996) 555–558.
- [140] G. Nürnberger, Optimal partitions in bivariate segment approximation, in: A. Le Méhauté, C. Rabut, L.L. Schumaker (Eds.), Surface Fitting and Multiresolution Methods, Vanderbilt University Press, Nashville, 1997, pp. 271–278.
- [141] G. Nürnberger, L.L. Schumaker, M. Sommer, H. Strauss, Approximation by generalized splines, J. Math. Anal. Appl. 108 (1985) 466–494.
- [142] G. Nürnberger, L.L. Schumaker, M. Sommer, H. Strauss, Uniform approximation by generalized splines with free knots, J. Approx. Theory 59 (1989) 150–169.
- [143] G. Nürnberger, I. Singer, Uniqueness and strong uniqueness of best approximations by spline subspaces and other spaces, J. Math. Anal. Appl. 90 (1982) 171–184.
- [144] G. Nürnberger, M. Sommer, A Remez type algorithm for spline functions, Numer. Math. 41 (1983) 117-146.
- [145] G. Nürnberger, M. Sommer, H. Strauss, An algorithm for segment approximation, Numer. Math. 48 (1986) 463-477.
- [146] M.R. Osborne, An algorithm for discrete, nonlinear best approximation problems, Proceedings of Oberwolfach Conference on ISNM, Vol. 16, Birkhauser, Basel, 1971.
- [147] M.R. Osborne, Finite Algorithms in Optimisation and Data Analysis, Wiley, Chichester, 1985.
- [148] M.R. Osborne, The reduced gradient algorithm, in: Y. Dodge (Ed.), Statistical Data Analysis Based on the L_1 Norm and Related Methods, North-Holland, Amsterdam, 1987, pp. 95–107.
- [149] M.R. Osborne, G.A. Watson, On the best linear Chebyshev approximation, Comput. J. 10 (1967) 172-177.
- [150] M.R. Osborne, G.A. Watson, An algorithm for minimax approximation in the nonlinear case, Comput. J. 12 (1969) 63–68.
- [151] M.R. Osborne, G.A. Watson, A note on singular minimax approximation problems, J. Math. Anal. Appl. 25 (1969) 692–700.
- [152] M.R. Osborne, G.A. Watson, On an algorithm for discrete nonlinear L_1 approximation, Comput. J. 14 (1971) 184–188.
- [153] M.R. Osborne, G.A. Watson, Nonlinear approximation problems in vector norms, in: G.A. Watson (Ed.), Proceedings of Dundee Numerical Analysis Conference, Springer, Berlin, 1978.

- [154] M.R. Osborne, G.A. Watson, Aspects of *M*-estimation and *l*₁ fitting problems, in: D.F. Griffiths, G.A. Watson (Eds.), Numerical Analysis: A R Mitchell 75th Birthday Volume, World Scientific Publishing Co, Singapore, 1996, pp. 247–261.
- [155] T. Pavlidis, A.P. Maika, Uniform piecewise polynomial approximation with variable joints, J. Approx. Theory 12 (1974) 61–69.
- [156] G. Polya, Sur une algorithme toujours convergent pour obtenir les polynomes de meillure approximation de Tchebysheff pour une fonction continue quelconque, Comptes Rendues 157 (1913) 840–843.
- [157] S. Portnoy, R. Koenker, The Gaussian hare and the Laplacian tortoise: computability of squared-error versus absolute-error estimators, Statist. Sci. 12 (1997) 279–296.
- [158] M.J.D. Powell, Approximation Theory and Methods, Cambridge University Press, Cambridge, 1980.
- [159] C.J. Price, I.D. Coope, Numerical experiments in semi-infinite programming problems, Comput. Optim. Appl. 6 (1996) 169–189.
- [160] A. Ralston, Rational Chebyshev approximation by Remes algorithms, Numer. Math. 7 (1965) 322-330.
- [161] A. Ralston, Some aspects of degeneracy in rational approximation, JIMA 11 (1973) 157-170.
- [162] R. Reemtsen, Modifications of the first Remez algorithm, SIAM J. Numer. Anal. 27 (1990) 507-518.
- [163] R. Reemtsen, Discretization methods for the solution of semi-infinite programming problems, J. Optim. Theory Appl. 71 (1991) 85–103.
- [164] R. Reemtsen, S. Görner, Numerical methods for semi-infinite programming: a survey, in: R. Reemtsen, J.-J. Ruckman (Eds.), Semi-Infinite Programming, Kluwer Academic Publishers, Boston, 1998, pp. 195–275.
- [165] E.YA. Remes, Sur un procédé convergent d'approximation successives pour déterminer les polynomes d'approximation, Comptes Rendues 198 (1934) 2063–2065.
- [166] E.YA. Remes, Sur le calcul effectif des polynomes d'approximation de Tchebichef, Comptes Rendues 199 (1934) 337–340.
- [167] E.C. Rhodes, Reducing observations by the method of minimum deviations, Philos. Mag. 9 (1930) 974–992.
- [168] J.R. Rice, The characterization of best nonlinear Tchebycheff approximation, Trans. Amer. Math. Soc. 96 (1960) 322–340.
- [169] J.R. Rice, Chebyshev approximation by $ab^x + c$, J. SIAM 10 (1960) 691–702.
- [170] J.R. Rice, Tchebycheff approximation by functions unisolvent of variable degree, Trans. Amer. Math. Soc. 99 (1961) 298–302.
- [171] J.R. Rice, Chebyshev approximation by exponentials, J. SIAM 10 (1962) 149-161.
- [172] J.R. Rice, The Approximation of Functions, Vol. I, Addison-Wesley, Reading, MA, 1964.
- [173] J.R. Rice, Characterization of Chebyshev approximation by splines, SIAM J. Math. Anal. 4 (1967) 557-567.
- [174] F. Riesz, Uber lineare Funktionalgleichungen, Acta. Math. 41 (1918) 71-98.
- [175] P.D. Robers, A. Ben-Israel, An interval programming algorithm for discrete linear L_1 approximation problems, J. Approx. Theory 2 (1969) 323–331.
- [176] P.D. Robers, S.S. Robers, Algorithm 458: discrete linear L_1 approximation by interval linear programming, Comm. ACM 16 (1973) 629–631.
- [177] H. Robitzsch, R. Schaback, Die numerische Berechnung von Startnaherungen bei der Exponentialapproximation, in: L. Collatz, G. Meinardus, H. Werner (Eds.), Numerical Methods of Approximation Theory, ISNM, Vol. 42, Birkhauser, Basel, 1978, pp. 260–280.
- [178] S.A. Ruzinsky, E.T. Olsen, l_1 and l_{∞} minimization via a variant of Karmarkar's algorithm, IEEE Trans. Acoust. Speech Signal Process. 37 (1989) 245–253.
- [179] R. Schaback, Suboptimal exponential approximation, Report Nr 23, University of Gottingen, 1979.
- [180] L.L. Schumaker, Uniform approximation by Tchebycheffian spline functions, J. Math. Mech. 18 (1968) 369-378.
- [181] L.L. Schumaker, Some algorithms for the computation of interpolating and approximating spline functions, in: T.N.E. Greville (Ed.), Theory and Applications of Spline Functions, Academic Press, New York, 1969, pp. 87–102.
- [182] R.R. Singleton, A method for minimizing the sum of absolute values of deviations, Ann. Math. Statist. 11 (1940) 301–310.
- [183] K. Spyropoulos, E. Kiountouzis, A. Young, Discrete approximation in the L_1 norm, Comput. J. 16 (1973) 180–186.
- [184] E.L. Stiefel, Uber diskrete und lineare Tschebyscheff- Approximationen, Numer. Math. 1 (1959) 1-28.
- [185] E.L. Stiefel, Note on Jordan elimination, linear programming and Tchebyscheff approximation, Numer. Math. 2 (1960) 1–17.

- [186] J. Stoer, A direct method for Chebyshev approximation by rational functions, J. Assoc. Comput. Mach. 1 (1964) 59–69.
- [187] L. Tornheim, On *n*-parameter families of functions and associated convex functions, Trans. Amer. Math. Soc. 69 (1950) 457–467.
- [188] K.N. Usow, On L₁ approximation I: computation for continuous functions and continuous dependence, SIAM J. Numer. Anal. 4 (1967) 70–88.
- [189] K.N. Usow, On L₁ approximation II: computation for discrete functions and discretization effects, SIAM J. Numer. Anal. 4 (1967) 233–244.
- [190] C.J. de la Vallée Poussin, Sur la methode de l'approximation minimum, Societe Scientifique de Bruxelles, Annales, Memoires, Vol. 35, 1911, pp. 1–16.
- [191] L. Veidinger, On the numerical determination of the best approximations in the Chebyshev sense, Num. Math. 2 (1960) 99–105.
- [192] H.M. Wagner, Linear programming techniques for regression analysis, J. Amer. Statist. Assoc. 54 (1959) 206-212.
- [193] J.L. Walsh, The existence of rational functions of best approximation, Trans. Amer. Math. Soc. 33 (1931) 668-689.
- [194] G.A. Watson, A multiple exchange algorithm for multivariate Chebyshev approximation, SIAM J. Numer. Anal. 12 (1975) 46–52.
- [195] G.A. Watson, A method for calculating best nonlinear Chebyshev approximations, JIMA 18 (1976) 351-360.
- [196] G.A. Watson, On two methods for discrete L_p approximation, Computing 18 (1977) 263–266.
- [197] G.A. Watson, On a class of methods for nonlinear approximation problems, in: D.C. Handscomb (Ed.), Multivariate Approximation, Academic Press, London, 1978, pp. 219–227.
- [198] G.A. Watson, The minimax solution of an overdetermined system of nonlinear equations, JIMA 23 (1979) 167–180.
- [199] G.A. Watson, Approximation Theory and Numerical Methods, Wiley, Chichester, 1980.
- [200] G.A. Watson, An algorithm for linear L_1 approximation of continuous functions, IMA J. Numer. Anal. 1 (1981) 157–167.
- [201] F. Wenzl, Uber Gleichungssysteme der Tschebysheffschen Approximation, Z. Angew. Math. Mech. 34 (1954) 385–391.
- [202] H. Werner, Die konstruktive Errmittlung der Tschebyscheff-Approximation in Bereich der rationalen Funktionen, Arch. Rational Mech. Anal. 11 (1962) 368–384.
- [203] H. Werner, Rationale-Tschebyscheff-Approximation, Eigenwerttheorie und Differenzenrechnung, Arch. Rational Mech. Anal. 13 (1963) 330–347.
- [204] H. Werner, Der Existenzsatz für das Tschebyscheffsdhe Approximationsproblem mit Exponenttialsummen, in: L. Collatz, H. Unger (Eds.), Funktionalanalytische Methoden der numerischen Mathematik, ISNM, Vol. 12, Birkhauser, Basel, 1969, pp. 133–143.
- [205] H. Werner, Tschebysheff-Approximation with sums of exponentials, in: A. Talbot (Ed.), Approximation Theory, Academic Press, London, 1969, pp. 109–134.
- [206] J.M. Wolfe, On the convergence of an algorithm for discrete L_p approximation, Numer. Math. 32 (1979) 439–459.
- [207] R.S. Womersley, R. Fletcher, An algorithm for composite nonsmooth optimization problems, J. Optim. Theory Appl. 48 (1986) 493–523.
- [208] J.W. Young, General theory of approximation by functions involving a given number of arbitrary parameters, Trans. Amer. Math. Soc. 8 (1907) 331–344.
- [209] Y. Zhang, A primal-dual interior point approach for computing the l_1 and l_{∞} solutions of overdetermined linear systems, J. Optim. Theory and Appl. 77 (1993) 323–341.
- [210] J.J. Zhou, A.L. Tits, An SQP algorithm for finitely discretized continuous minimax optimization problems and other minimax problems with many objective functions, SIAM J. Optim. 6 (1996) 461–487.
- [211] S.I. Zuhovickii, L.I. Avdeyeva, Linear and Convex Programming, (Translated from the 1964 Russian edition), Saunders Co., Philadelphia, 1966.
- [212] S.I. Zuhovickii, R.A. Poljak, M.E. Primak, An algorithm for the solution of the problem of Cebysev approximation, Soviet. Math. Dokl. 4 (1963) 901–904.



JOURNAL OF COMPUTATIONAL AND APPLIED MATHEMATICS

Journal of Computational and Applied Mathematics 121 (2000) 37-72

www.elsevier.nl/locate/cam

A tutorial history of least squares with applications to astronomy and geodesy $\stackrel{\text{tot}}{\approx}$

Yves Nievergelt*

Department of Mathematics, Eastern Washington University, MS-32, Cheney, WA 99004-2431, USA

Received 2 October 1999; received in revised form 20 December 1999

Abstract

This article surveys the history, development, and applications of least squares, including ordinary, constrained, weighted, and total least squares. The presentation includes proofs of the basic theory, in particular, unitary factorizations and singular-value decompositions of matrices. Numerical examples with real data demonstrate how to set up and solve several types of problems of least squares. The bibliography lists comprehensive sources for more specialized aspects of least squares. © 2000 Elsevier Science B.V. All rights reserved.

Keywords: History; Constrained; Ordinary; Total; Weighted; Least squares

1. Introduction

The mathematical concept of least squares is the basis for several methods to fit certain types of curves and surfaces to data. Problems of fitting curves and surfaces have a history spanning several millenia, which is outlined in Section 2 to set in perspective the contribution of least squares to their solution. The citations provided here include page numbers from Dreyer's book [13] to identify the original texts. Examples of such problems include the determination of the shape and size of celestial bodies and of their trajectories.

These problems were still without satisfactory solutions near the end of the eighteenth century A.D., at the time of the development of the concepts of problems of least squares and their solution with normal equations; see Section 3. (For greater detail, see Stewart's translation [16] of Gauss's work.)

[†] Work supported by a faculty research grant from Eastern Washington University.

^{*} Tel.: +1-509-359-4259; fax: +1-509-359-4700.

E-mail address: ynievergelt@ewu.edu (Y. Nievergelt).

For computations with floating-point or other approximate arithmetics, normal equations can exhibit a sensitivity to errors in data or in rounding larger than the sensitivity of methods with unitary factorizations. These factorizations also provide methods to solve problems of constrained and total least squares, as explained in Sections 4 and 5.

For the state-of-the-art in computing with least squares near the end of the second millenium A.D., Björk [1], Dennis Jr., and Schnabel [12], and Lawson and Hanson [32] present algorithms to solve least-squares problems, and Higham [22] also treats the analysis of sensitivity to errors. Van Huffel and Vandewalle [50] focus on total least-squares. These references also contain extensive bibliographies. To compute solutions of practically significant problems, the usual recommendation is to use one of the professionally maintained libraries of computer programs, for instance, netlib (http://www.netlib.org/lapack/).

2. An ancient history of curve and surface fitting

2.1. Fitting surfaces: the shapes of the earth and of the moon

One fitting problem consists in estimating the shape of the earth. Early in the first millenium B.C., several shapes were fitted to various combinations of religious canons, philosophical doctrines, and observations of the rôles of air, earth, fire, and water. Types of surfaces fitted to such ideas included a *circular disc* (Thales of Miletus, about 640–562 B.C. [13, p. 11]), an *infinite plane* (Xenophanes of Kolophon, about 570–475 B.C. [13, p. 18]), and a *sphere* (Parmenides of Elea, early in the fifth century B.C. [13, p. 20]). The type of surface was also fitted to observations of inequalities reported by travelers. For example, the star Canopus remained invisible to a traveler in Greece, became just visible above the horizon at Rhodes, and then appeared higher and higher above the horizon as the traveler went further and further south [13, p. 20]. Also while sailing toward the setting sun, mariners in the north saw the sun on their left, but mariners in the south saw the sun on their right [13, p. 39]. From the fifth century B.C., in Greece and India, the type of surface fitted to such observations was a sphere [13, pp. 39, 242].

Similarly, for the shape of the moon, a sphere fitted the observation that the lighted side of the moon always faces the sun (Parmenides [13, p. 21]; Anaxagoras of Klazomenæ, about 500–428 B.C. [13, p. 32]).

With the shape settled to be a sphere arises the problem of estimating its size.

To estimate the circumference of the earth, Posidonius of Apameia (about 135–50 B.C.) referred to a result attributed to Archimedes (287–212 B.C.) and Dikæarchus of Messana (about 285 B.C.), using two stars seen from two cities; see Fig. 1. The cities are Lysimachia in Thrace, and Syene in Upper Egypt, which lie 20 000 stadia apart from each other. The first star, γ Draconis, appears at the zenith (vertical direction) above Lysimachia. The second star, in the constellation Cancer, appears at the zenith above Syene. The difference between the declinations (angular elevations from the celestial equator) of the two stars is 1/15 of a full circle, which is thus the difference between the vertical directions at the two cities. Therefore, the circumference of the earth is 15 * 20000 = 300000stadia, corresponding to approximately 100 000 stadia for the earth's diameter [13, pp. 173–174]. (Though Archimedes and Apollonius already knew the approximations $\pi \approx 22/7$ and $\pi \approx 3.1416$



Fig. 1. Posidonius's estimate of the earth's circumference. Stars appear in the same direction from every point on earth. Two stars make an angle $2\pi/15$. One of them is at the zenith above Syene, the other is at the zenith at Lysimachia. Therefore, 15 times the distance from Syene to Lysimachia equals the earth's circumference.



Fig. 2. Eratosthenes's estimate of the earth's circumference. The sun rays appear parallel on earth. They are vertical at Syene. At Alexandria, with a vertical stick they make an angle 1/50 of a full circle. Therefore, 50 times the distance from Syene to Alexandria equals the earth's circumference.

[49, pp. 185–186], the approximation $\pi \approx 3$ was then common for practical purposes not only in Greece but also in Babylon, in Egypt [49, p. 173], and in China [49, p. 196].)

With a different procedure, Eratosthenes of Alexandria (276–194 B.C.) used the shadows of vertical rods in two cities; see Fig. 2. At the summer soltice, at Syene the rod casts no shadow, so that the sun rays fall vertically, while at Alexandria the sun rays and the vertical rod make and angle equal to 1/50 of a full circle. (According to van der Waerden, the computation of this angle from measurements of the lengths of the rod and of its shadow proceeded through the Theorem of Pythagoras and tables of sines [49, p. 214].) Because Syene lies 5000 stadia away from Alexandria, it follows that the circumference of the earth is about 50*5000=250 000 stadia. Kleomedes corroborated this results through the same procedure at the same locations but at the winter soltice. Table 1 shows comparisons with the World Geodetic System WGS-84 [23].

Source	Circumference	Radius
Archimedes, Dikæarchus,	300 000 stadia	50000 stadia
Posidonius, 3rd century B.C.	$(47250000\mathrm{m})$	(7875000m)
Eratosthenes, 2nd century B.C.	250 000 stadia	
$(1 \text{ stade} \approx 157.5 \text{ m})$	(39375000m)	(6562500m)
WGS-84, 1984 A.D.		$a = 6378137.00000 \mathrm{m}$
		$b = 6356752.31425\mathrm{m}$
		$e^2 = 0.00669437999013$
$Mathematica \ 4a \texttt{EllipticE}[e^2]$	40 007 862.917 27 m	

Table 1 Comparisons of estimates of the earth's polar circumference and radius



Fig. 3. Hipparchus's estimate of the radii of the moon and its orbit.

The estimate of the same circumference by different observers through different procedures or through repeated measurements hints at some attempts to detect errors, but no records of such attempts appear to remain [13, p. 177].

2.2. Fitting curves: the radii of the moon and its orbit

Another fitting problem consists in estimating the trajectories of celestial bodies. For example, rectilinear motions fitted the poetical ideas of Xenophanes in the sixth century B.C. [13, p. 18]. A century later, Philolaus of Thebes proposed circular orbits for the earth, the moon, the planets, and the sun, all around a "central fire" reflected by the sun toward the earth; such orbits fitted coarse observations of planetary motions [13, pp. 40–49]. In the third century B.C., Aristarchus of Samos outlined a heliocentric system with a circular orbit for the earth around the sun [13, p. 137].

With the orbits settled as circles arises the problem of estimating their size.

To estimate simultaneously the distance from the earth to the moon and the radius of the moon, Hipparchus of Nicæa (second century B.C.) used a full lunar eclipse [13, pp. 183–184]; see Fig. 3. Within the measurement accuracy available then, the sun's parallax p is nearly zero. Seen from the earth, the sun sustends an angle u=16'36''55''', and the path of the moon across the earth's shadow sustends and angle v = 41'32''17.5'''. The ratio $180^{\circ}/v \approx 260$ can also be calculated as the ratio t_1/t_2 of the time t_1 of a full revolution of the moon (29.5 days) and the time t_2 taken by the moon to cross the earth's shadow. Consequently, the parallax of the earth's shadow on the moon is nearly q = u + v = 58'09''12.5'''. Therefore, the ratio d/r_e of the distance d from the earth to

Source	Moon's radius	Orbit's radius
Hipparchus, 2nd century B.C.	$r_{\rm m} = r_{\rm e}/3.5$	$d = 59.1r_{\rm e}$
	(18/5000 m)	(38/843/50 m)
Hipparchus, 2nd century B.C.,	$r_{\rm m} = r_{\rm e}/3.4$	$d = 60\frac{2}{6}r_{\rm e}$
reported by Kleomedes.	(193014/m)	(399218/50 m)
[26, p. 476], 1984 A.D.	1738 000 m	384 400 000 m
	$(b/r_{ m m} \approx 3.658)$	$(d/b \approx 60.47)$

Table 2

Comparisons of estimates of the radii of the moon and its orbit

the moon and the earth's radius r_e is $1/\sin(q) = 1/\sin(u+v) = 59.1$. Moreover, the diameter of the earth's shadow at distance d from the earth equals about $d * v/180^\circ = d * t_2/t_1$. A measurement of the time t_3 from the moment the moon touches the earth's shadow to the moment it disappears in it then gives an estimate of the radius of the moon r_m in the form $2r_m/(d * t_2/t_1) = t_3/t_2$, whence $r_m = (d/2) * (t_3/t_1) = (59.1r_e/2) * (t_3/t_1) = r_e/3.5$.

According to Ptolemy's account, Hipparchus attempted to measure a lower bound and an upper bound for the sun's parallax p. The results just presented correspond to the lower bound 0. Kleomedes's report of another result from Hipparchus, $d = 60\frac{5}{6}r_e$ [13, pp. 183–184], corresponds to the upper bound 2'44". Such bounds hint at attempts to detect the maximum error.

With Eratosthenes's measure of the earth's radius, Hipparchus's results give 387 843 750 m for the distance of the moon, and 1 875 000 m for the radius of the moon. Table 2 shows comparisons with textbook values [26, p. 476].

2.3. Fitting curves and surfaces: planetary orbits and earth's geoid

It was also considerations of maximum errors, of the order of 8' between Tycho Brahe's observations of Mars and Copernicus's heliocentric model, which led Johann Kepler to abandon circles for the orbits, and finally (about 18 December 1604 A.D.) to substitute *ellipses* with a focus at the sun, along which planets sweep equal areas in equal times [13, pp. 389–392]. In 1687, Isaac Newton outlines in the *Philosophiae Naturalis Principia Mathematica* a proof that Kepler's laws are mathematically equivalent to the action of an attraction from the sun and inversely proportional to the square of the distance from the sun to the planet [39].

From Newton's law of gravitational attraction, it follows (from mathematical derivations by Newton, Ivory, Huygens, Clairaut, and Laplace [31, Book III, Section 18]) that a rotating mass of a homogenous and incompressible fluid can have the shape of an *ellipsoid* rotating around its shortest axis [20, pp. 172–175]. From 1700 through 1733, three surveys in France all suggested that the earth was an ellipsoid rotating around its *largest* axis [4, pp. 250–251]; such a surface failed to fit Newton's mathematical theory, based on Kepler's physics, itself based on Tycho Brahe's measurements. Ordered by Louis XV, a survey in Lapland and a survey in Peru in 1735 reversed the earlier results and confirmed that the earth was an ellipsoid rotating around it shortest axis [4, pp. 251–252].

The foregoing historical outline shows that for nearly three millenia, curves and surfaces were fitted to ideologies and theories. Yet errors — discrepancies between the fitted curve or surface and



Fig. 4. The geodetic latitude of a point (ρ, z) is the angle λ between the normal to the surface through (ρ, z) and the equatorial plane.

observations — drew attention through gross departures from the theory or through unacceptable *maximum* values.

3. Weighted ordinary least squares and geodesy

3.1. Precursors: minimax and minimum average modulus

By the end of the 18th century A.D., the Marquis Pierre Simon de Laplace (1749–1827) was using a sequence of several methods to fit curves and surfaces to measurements in geodesy and astronomy. Each of his methods minimizes either the maximum residual, the average absolute residual, or the average squared residual, of a linearized model.

For example, consider the problem of fitting an ellipse to a polar cross section of the earth, with principal semi-axes of lengths $a \ge b > 0$. Let $e^2 := 1 - (b/a)^2$ be its squared eccentricity, and set $\sigma^2 := 1 - e^2 = (b/a)^2$. For each point x := (x, y, z) on the earth's surface, the geodetic latitude of x is the angle λ between the normal to the surface at x and the equatorial plane, as in Fig. 4. With the cylindrical coordinate $\rho := \sqrt{x^2 + y^2}$, calculus gives

$$\rho = \frac{a\cos(\lambda)}{\sqrt{1 - e^2[\sin(\lambda)]^2}}, \qquad z = \frac{a\sigma^2\sin(\lambda)}{\sqrt{1 - e^2[\sin(\lambda)]^2}}$$

Hence, the differential of the arclength s along a meridian becomes

$$ds = \frac{a\sigma^2}{\{1 - e^2[\sin(\lambda)]^2\}^{3/2}} d\lambda$$

= $a\sigma^2\{1 + \frac{3}{2}e^2[\sin(\lambda)]^2 + \frac{3*5}{2*2*2!}e^4[\sin(\lambda)]^4 + \cdots\} d\lambda.$

Crude approximations indicate that $e^2 < 0.01$. Beyond the first two terms,

$$\sum_{k=2}^{\infty} \prod_{\ell=1}^{k} (2\ell+1) \frac{|e\sin(\lambda)|^{2k}}{2^{k} * (k!)} < \frac{15e^{4}}{8} \left(1 + \frac{7e^{2}}{6}\right) \sum_{k=0}^{\infty} e^{k} < 0.000\,25.$$

Thus, with a relative error less than 0.00025 uniformly over the earth's surface, the length Δs of an arc $\Delta \lambda$ of meridian at the geodetic latitude λ takes the following form, with $c_0 := a\sigma^2$ and $c_1 := \frac{3}{2}a\sigma^2 e^2$:

$$\frac{\Delta s}{\Delta \lambda} = c_0 + c_1 [\sin(\lambda)]^2.$$

Thus, measurements of the lengths of n arcs of a meridian produce n equations.

Example 1. With lengths in double toises (1/0.256537 m) and angles in grads $(2\pi = 400^{\circ})$, Laplace considered the following system [31, Book III, Section 41]:

$\Delta s/\Delta \lambda = c_0 + c_1 [\sin(\lambda)]^2;$	location;	latitude λ ;	arc $\Delta \lambda$;
$25538.85 = c_0 + c_1 * 0.00000;$	Peru;	$00.0000^{\circ};$	3.4633°;
$25666.65 = c_0 + c_1 * 0.30156;$	Good Hope;	37.0093°;	1.3572°;
$25599.60 = c_0 + c_1 * 0.39946;$	Pennsylvania;	43.5556°;	1.6435°;
$25640.55 = c_0 + c_1 * 0.46541;$	Italy;	47.7963°;	2.4034°;
$25658.28 = c_0 + c_1 * 0.52093;$	France;	51.3327°;	$10.7487^{\circ};$
$25683.30 = c_0 + c_1 * 0.54850;$	Austria;	53.0926°;	3.2734°;
$25832.25 = c_0 + c_1 * 0.83887;$	Lapland;	73.7037°;	1.0644° .

The problem then consisted in fitting c_0 and c_1 to this linear system.

Laplace's first method aimed at determining the ellipsoid that *minimizes the maximum error* between the fitted ellipsoid and the measurements [31, Book III, Section 39]. From this first method he concluded that the earth's surface was not exactly an ellipsoid but the maximum error was within the measurement accuracy, with a flattening f:=1-(b/a)=1/277 [31, Book III, Section 41], which corresponds to a squared eccentricity $e^2 < 0.007207 < 0.01$.

Laplace's second method aimed at determining the ellipsoid that *minimizes the average absolute* values of the errors subject to the constraint that the sum of the errors equal zero; the result yielded what he considered the most probable ellipsoid [31, Book III, Section 40].

The second method presented several difficulties. Firstly, the "most probable" estimate depends on the probability distribution of the errors and can fail to coincide with the minimum average absolute error [24, pp. 400-401]. Secondly, Laplace's method did not lend itself to the methods of power series, and no efficient algorithm existed to determine the solutions (until George B. Dantzig's simplex algorithm in the 1950s [6,10,11]). Finally, for an overdetermined system of linear equations with a matrix of any rank, Laplace's method can lead to multiple solutions filling an entire polytope [6, p. 219].

Example 2. Consider the following system Ax = b with maximal rank:

x + y = 4, x - y = 0, x - y = 2,x + y = 6. The residuals r = Ax - b add to zero at x = 3. Setting x = 3 gives $\{|(3 + y) - 4| + |(3 - y)| + |(3 - y) - 2| + |(3 + y) - 6|\}/4$ $= \{|y - 1| + |y - 3|\}/2$ $= \begin{cases} 2 - y > 1 \text{ if } y < 1, \\ 1 = 1 & \text{if } 1 \le y \le 3, \\ y - 2 > 1 \text{ if } 3 < y. \end{cases}$

The average reaches its minimum everywhere on the segment $\{3\} \times [1,3]$.

Each of Laplace's numerical examples of a minimization of the average absolute error consists of an *odd* number of equations [31, Book III, Sections 41–42]. In contrast, for the determination of orbits of celestial bodies, Laplace used an *even* number of linearized equations, corresponding to measurements at times scattered symmetrically about a central time t_0 :

$$t_0 - t_k, t_0 - t_{k-1}, \dots, t_0 - t_1, t_0 + t_1, \dots, t_0 + t_{k-1}, t_0 + t_k.$$

This produces a peculiar type of linear system, where the first column of coefficients A(;1) is perpendicular to the second column of coefficients A(;2), as in Example 2. For such systems, Laplace did not minimize the average absolute error. Instead, in effect, he computed the dot product of the system with the transposed column $A(;1)^*$ and solved for x, and then computed the dot product of the system with $A(;2)^*$ and solved for y [31, Book II, Section 37].

Example 3. Consider the system Ax = b from Example 2:

$$\begin{pmatrix} 1 & 1 \\ 1 & -1 \\ 1 & -1 \\ 1 & 1 \end{pmatrix} \begin{pmatrix} x \\ y \end{pmatrix} = \begin{pmatrix} 4 \\ 0 \\ 2 \\ 6 \end{pmatrix},$$

$$\begin{pmatrix} 4x \\ 4y \end{pmatrix} = \begin{pmatrix} 1 & 1 & 1 & 1 \\ 1 & -1 & -1 & 1 \end{pmatrix} \begin{pmatrix} 1 & 1 \\ 1 & -1 \\ 1 & -1 \\ 1 & 1 \end{pmatrix} \begin{pmatrix} x \\ y \end{pmatrix} = \begin{pmatrix} 1 & 1 & 1 & 1 \\ 1 & -1 & -1 & 1 \end{pmatrix} \begin{pmatrix} 4 \\ 0 \\ 2 \\ 6 \end{pmatrix} = \begin{pmatrix} 12 \\ 8 \end{pmatrix}.$$

Thus, x = 3 and y = 2.

For the peculiar type of linear system in Examples 2 and 3, Laplace's method amounts to solving the normal equations for the least squares solution.

3.2. Weighted ordinary least squares

Around 1800, Laplace, Legendre, and Gauss were fitting functional forms to data through various types of least squares. Laplace's method applied to systems with mutually orthogonal columns. Legendre (1752–1833) published the method of normal equations in 1805 [33]. In 1821–1823, Gauss published the method of weighted least squares to solve linear systems Ax = b with a matrix A with

n linearly independent columns and $m \ge n$ rows [16]. Though Gauss did not employ a notation with matrices, a derivation of weighted least squares with matrices can proceed as follows [47, p. 144].

The problem consists in determining a linear function u of n variables a_1, \ldots, a_n , which amounts to determining coefficients x_1, \ldots, x_n so that

$$u(a_1,\ldots,a_n)=a_1x_1+\cdots+a_nx_n.$$

The data consist of *m* points $A(i;) = (a_{i,1}, ..., a_{i,n})$, arranged as the rows of the matrix *A*, and of the value b_i if *u* at each point. The problem then amounts to fitting coefficients $x_1, ..., x_n$ to the linear system Ax = b:

$$a_{1,1}x_1 + \dots + a_{1,n}x_n = b_1,$$

$$\vdots$$

$$a_{m,1}x_1 + \dots + a_{m,n}x_n = b_m.$$

The data can also include estimates of the precision of the measurements, in the form of the reciprocal of the variance of each measurement, as investigated by Gauss, or, more generally, in the form of the inverse V^{-1} of the covariance matrix V of the measurements, as investigated by Aiken [2]. Specifically, if b_i represents the average $E(B_i)$ of a random variable B_i , estimated by the average of several observations, then $V_{i,j} = E[(B_i - b_i)(B_j - b_j)]$ is the covariance of B_i and B_j . The solution X of the linear system AX = B is then also a random variable. The problem solved by Gauss consists in finding a linear transformation L such that LA = I, to solve for $\tilde{x} = I\tilde{x} = LA\tilde{x} = Lb$, such that $\tilde{x} = Lb$ minimizes the covariance

$$U = E[(X - \tilde{x})(X - \tilde{x})^*].$$

Gauss showed that \tilde{x} is also the solution of the weighted least-squares system

$$WAx = Wb$$

with a matrix of weights W such that $W^*W = V^{-1}$, and then

$$L = (A^* V^{-1} A)^{-1} A^* V^{-1}.$$

Indeed, for every matrix K such that KA = I,

$$U = E[(X - \tilde{x})(X - \tilde{x})^*]$$

= $E[(X - KB)(X - KB)^*]$
= $E\{[X - KAX - K(B - AX)][X - KAX - K(B - AX)]^*\}$
= $E\{[K(B - AX)][K(B - AX)]^*\}$
= $KE\{(B - AX)(B - AX)^*\}K^*$
= KVK^*
= $LVL^* + (K - L)VL^* + LV(K - L)^* + (K - L)V(K - L)^*.$

The two middle terms equal zero, because of the condition KA = I and the definition of L. The last term, $(K - L)V(K - L)^*$, is hermitian positive semidefinite. Hence, for each vector z,

$$z^*Uz = z^*LVL^*z + z^*(K - L)V(K - L)^*z \ge z^*LVL^*z,$$

with z^*Uz minimum for K:=L. Moreover, the formula for U simplifies to

 $U = (A^* V^{-1} A)^{-1},$

which is thus the covariance matrix of the weighted least squares solution X.

Example 4. For the system in Example 1, Laplace weighted each equation by the number of degrees $\Delta \lambda$ in the corresponding measured arc. There was no known correlation between the measurements from the different teams assigned to measure different arcs. Thus the weight matrix W is diagonal with the corresponding values of $\Delta \lambda$ on its diagonal:

W =diagonal (3.4633, 1.3572, 1.6435, 2.4034, 10.7487, 3.2734, 1.0644).

The weighted least-squares solution (computed through the command LSQ on a Hewlett–Packard's HP48GX calculator [21, pp. 14,15]) is

$$c_0 = 25534.47,$$

 $c_1 = 242.81.$

Hence

 $e^{2} = \frac{2c_{1}}{3c_{0}} = 0.006\ 339\dots,$ $\sigma^{2} = 1 - e^{2} = 0.993\ 661\dots,$ $f = 1 - b/a = 1 - \sigma = 0.003\ 175\dots,$ $a = \frac{c_{0}}{\sigma^{2}} = 25697.38^{R}\dots = 100\ 170.25\ m.$

Laplace gives f = 1/277 = 0.003610, though Bowditch's calculations of Laplace's method lead to f = 1/250 = 0.004 [31, Book III, Section 41]. The values from WGS-84 are f = 0.00335281066474 and a = 6378137 m [23, p. xxiii]. Finally,

$$U = (A^* V^{-1} A)^{-1} = (A^* W^* W A)^{-1} = \begin{pmatrix} 0.007\ 701 & -0.147\ 686 \\ -0.147\ 686 & 0.309\ 706 \end{pmatrix}$$

The method of weighted least squares assigns weights only to the measured values b of the function u, but not to be coordinates of the points $(a_{i,1}, \ldots, a_{i,n})$. In Laplace's application, this would correspond to treating the measurements of the lengths of arcs of meridians as random variables, but considering the determinations of the geodetic latitudes as exact. Allowances for adjustments of all data require different methods, as explained below in Section 6.

4. Unitary factorizations and constrained least squares

4.1. Householder symmetries and unitary factorizations

To solve linear systems, Gaussian elimination performs a linear transformation known as a shear that maps a column of coefficients r = A(;j) to a multiple of a canonical basis vector e_j , which "eliminates" the coefficients below the *j*th row. Yet shears alter Euclidean distances, in particular,



Fig. 5. A Householder symmetry maps r to $-\operatorname{sign}(r_1) ||r||_2 e_1$.

they do not reveal which vector lies closest to the "right-hand side" of the system. In contrast, one of the strategies for solving least squares problems consists in replacing Gaussian elimination by a type of linear elimination that preserves Euclidean distances, for instance, Modified Gram-Schmidt (MGS) orthogonalization [22, Section 19.3, 43], Givens rotations, or Householder symmetries [1,32].

Householder symmetries involves the function sign: $\mathbb{C} \to \mathbb{C}$ defined by

sign(z):=
$$\begin{cases} z/|z| & \text{if } z \neq 0, \\ 1 & \text{if } z = 0. \end{cases}$$

For each non-zero vector $r \in \mathbb{C}^m \setminus \{0\}$, a Householder symmetry reflects r onto a multiple $-\operatorname{sign}(r_1)||r||_2 e_1$ of the basis vector e_1 across the hyperplane $H \subset \mathbb{C}^m$ that passes through the origin 0 perpendicularly to the bisectrix of the angle formed by r and $\operatorname{sign}(r_1)e_1$, which lies in the direction of $v:=r + \operatorname{sign}(r_1)||r||_2 \cdot e_1$, as in Fig. 5. The choice of $\operatorname{sign}(r_1)$ minimizes rounding inaccuracies, so that if $r \neq 0$ then $v:=r + \operatorname{sign}(r_1)||r||_2 e_1 \neq 0$, because $|v_1| = |r_1 + \operatorname{sign}(r_1)||r||_2 |\geqslant ||r||_2 > 0$. The hyperplane H is then perpendicular to the unit vector $u:=(1/||v||_2)v$.

A Householder symmetry S thus amounts to subtracting from r twice its projection along u, so that $S(r) = r - 2\langle r, u \rangle u$, which leads to Algorithm 1.

Algorithm 1. Data: any non-zero vector $r \in \mathbb{C}^m \setminus \{0\}$ (1) $s:=sign(r_1)$. (2) $v:=r+s||r||_2e_1$. (3) $v:=1/\{||r||_2(||r||_2+|r_1|)\}.$

Result. $S(Z) = Z - vv(v^*Z)$ for every $Z \in M_{m \times n}(\mathbb{C})$.

Proposition 5 verifies that Algorithm 1 produces a Householder symmetry.

Proposition 5. The transformation S defined by algorithm 1 reflects r onto $S(r) = -s \cdot ||r||_2 \cdot e_1$. Moreover, the matrix S of S is hermitian and unitary.

Proof. With $S(r) = r - vv(v^*r)$ defined as in Algorithm 1,

$$v^*r = \sum_{j=1}^m \overline{v_j}r_j = \overline{v_1}r_1 + \sum_{j=2}^m \overline{v_j}r_j = (\overline{r_1} + \overline{s}||r||_2)r_1 + \sum_{j=2}^m \overline{r_j}r_j$$
$$= ||r||_2^2 + \overline{s}||r||_2r_1 = ||r||_2(||r||_2 + |r_1|) = 1/v,$$

 $S(r) = r - v \cdot v \cdot (1/v) = r - v = r - (r + s \cdot ||r||_2 \cdot e_1) = -s||r||_2 e_1.$ Moreover, $||v||_2^2 = 2/v$:

$$||v||_{2}^{2} = \sum_{j=1}^{m} |v_{j}|^{2} = |r_{1} + s||r||_{2}|^{2} + \sum_{j=2}^{m} |r_{j}|^{2}$$
$$= |s|^{2} ||r||_{2}^{2} + (r_{1}\bar{s} + \bar{r_{1}}s)||r||_{2} + |r_{1}|^{2} + \sum_{j=2}^{m} |r_{j}|^{2}$$

$$= ||r||_{2}^{2} + 2|r_{1}| \cdot ||r||_{2} + ||r||_{2}^{2} = 2||r||_{2}(||r||_{2} + |r_{1}|) = 2v.$$

Consequently, the Householder symmetry S has a hermitian matrix, $S^* = S$:

$$S^* = (I - vvv^*)^* = I^* - v(v^*)^*v^* = S.$$

Finally, the Householder symmetry S is a unitary transformation, $S^*S = I$:

$$S^*S = (I - vvv^*) \cdot (I - vvv^*) = I - 2Ivvv^* + vvv^*vvv^*$$
$$= I - 2Ivvv^* + v^2v ||v||_2^2 v^* = I - 2vvv^* + 2vvv^* = I. \square$$

Applied to the first column r:=A(;1) of any rectangular matrix $A \in M_{m \times n}(\mathbb{C})$, the Householder symmetry S produces zeros under the first entry $(SA)_{1,1}=-s||A(;1)||_2$, and transforms the subsequent columns into $(SA(;2),\ldots,SA(;n))$. By induction, Householder symmetries S_1,\ldots,S_n (such that each S_k modifies only entries on or below the kth row) produce a unitary — but not necessarily hermitian — matrix $Q = S_n^* \cdots S_1^*$, and an upper-triangular matrix R, with

A = QR.

4.2. Solving least-squares problems with orthogonal factorizations

Consider a linear system Ax = b with $n \le m$ linearly independent columns in $A \in M_{m \times n}(\mathbb{C})$. If A = QR with Q unitary and R upper triangular, then

Ax = b,

 $Q^*Ax = Q^*b,$

$$Rx = Q^*b,$$

where multiplication Q preserves Euclidean distances, whence

$$||Ax - b||_2 = ||Rx - Q^*b||_2$$

Because *R* has *n* linearly independent columns and has only zeros below the *r*th row, $||Rx - Q^*b||_2$ reaches a minimum if and only if *x* is the unique solution \tilde{x} of the first *n* equations. Moreover,

$$||R\tilde{x} - Q^*b||_2 = ||((Q^*b)_{n+1}, \dots, (Q^*b)_m)||_2$$

For a matrix $A \in M_{m \times n}(\mathbb{C})$ with rank $r \leq \min\{m, n\}$ and columns that *need not* be linearly independent, there exists a unitary factorization

$$AP = QR.$$

The matrix $P \in M_{n \times n}(\mathbb{C})$ permutes the columns of A so that the first r columns of AP are linearly independent. Householder symmetries then yield a unitary factorization of the first r columns,

$$[AP(;1),\ldots,AP(;r)] = Q[R(;1),\ldots,R(;r)],$$

and $R = Q^*(AP)$ contains only zeros below the *r*th row. With the change of coordinates $z := P^{-1}x$, there is then an affine subspace of dimension n - r of least-squares solutions to the system

$$Ax = b,$$

$$(Q^*AP)(P^{-1}x) = Q^*b,$$

$$Rz = Q^*b.$$

One solution z results from setting $z_{r+1}:=\cdots:=z_n:=0$ and solving the z_1,\ldots,z_r . the shortest least-squares solution x is then the orthogonal projection of any solution z on the orthogonal complement of the null space of R, in other words, on the row space of R.

Such a projection can employ a unitary factorization of R^* ,

 $R^* = WT$

with W unitary and T upper triangular. Because $T = W^*R^*$ has only zeros below its rth row, it follows that the last n - r columns w_{r+1}, \ldots, w_n of W form an orthonormal basis of Kernel (R), while the first r columns w_1, \ldots, w_r form an orthonormal basis of its row space. Consequently,

$$\tilde{x} := (w_1 \dots w_r) \begin{pmatrix} w_1^* \\ \vdots \\ w_r^* \end{pmatrix} z$$

minimizes $||A\tilde{x} - b||_2$ with the smallest norm $||\tilde{x}||_2$; see also [32, Chapter 14].

In principle, the permutations P can be generated during the computation of each symmetry S_k , by swapping columns A(;k) and $A(;\ell)$ for some $\ell > k$ if A(;k) lies in the subspace spanned by $A(;1), \ldots, A(;k-1)$. However, detecting such linear dependencies and selecting a permutation amounts to computing the ranks of submatrices, which is not reliable with floating-point or other approximate arithmetics [12, p. 66]. The singular value decomposition will provide some information on the reliability of such computations.

4.3. Constrained least squares and geodesy

Such practical situations as geodesy lead to problems of least squares with linear constraints. The outline presented here expands on that of Lawson and Hanson [32, Chapter 20]. Specifically, for matrices

 $C \in \mathbb{M}_{k imes n}(\mathbb{C}),$ $E \in \mathbb{M}_{\ell imes n}(\mathbb{C}),$ $d \in \mathbb{C}^k,$ $f \in \mathbb{C}^\ell,$ the problem consists in determining a vector $x \in \mathbb{C}^n$ that minimizes

$$\|Ex-f\|_2$$

subject to the constraint

Cx = d.

The strategy for solving such a problem uses an orthonormal basis $(q_1, \ldots, q_k; q_{k+1}, \ldots, q_n)$, where (q_{k+1}, \ldots, q_n) is an orthonormal basis on the null space of *C*. The basis (q_{k+1}, \ldots, q_n) provides a parametrization of the solution space of the system Cx = d, which reduces the problem to an unconstrained least squares problem in the subspace of \mathbb{C}^n spanned by (q_1, \ldots, q_k) .

In the generic situation where C has k linearly independent rows and E has n linearly independent columns, C^* factors in the form

$$C^* = QR,$$

$$C = LQ^*,$$

where $Q \in M_{n \times n}(\mathbb{C})$ is unitary, $R \in M_{n \times k}(\mathbb{C})$ is upper triangular, and $L = R^*$ is lower triangular with linearly independent rows. Because $R = Q^*C^*$ has only zeros below the *k*th row, it follows that in Q^* all the rows q_{k+1}^*, \ldots, q_n^* are perpendicular to all the columns of C^* . Hence, the rows q_1^*, \ldots, q_k^* span the column space of C^* . Thus Q performs the required change of basis. With

 $w:=Q^*x$,

the system becomes

$$\begin{pmatrix} L \\ EQ \end{pmatrix} \begin{pmatrix} w_1 \\ w_2 \end{pmatrix} = \begin{pmatrix} C \\ E \end{pmatrix} Q(Q^*x) = \begin{pmatrix} d \\ f \end{pmatrix}.$$

Therefore, there exists exactly one solution $w_1 \in \mathbb{C}^k$ to the system

$$L\begin{pmatrix}w_1\\0\end{pmatrix}=d.$$

The initial problem thus reduces to determining $w_2 \in \mathbb{C}^\ell$ minimizing

$$\left\| \begin{pmatrix} L \\ EQ \end{pmatrix} \begin{pmatrix} w_1 \\ w_2 \end{pmatrix} - \begin{pmatrix} d \\ f \end{pmatrix} \right\|_2 = \left\| EQ \begin{pmatrix} 0 \\ w_2 \end{pmatrix} - \left[f - EQ \begin{pmatrix} w_1 \\ 0 \end{pmatrix} \right] \right\|_2$$

The following application uses the Gauss-Bonnet Theorem.

Theorem 6 (Gauss-Bonnet). Let D be a compact oriented domain with Euler characteristic χ on a Riemannian surface M in \mathbb{R}^3 . Let $C = \partial D$ be the boundary of D in M, and let $\alpha_1, \ldots, \alpha_L$ be the oriented internal angles at the vertices (if any) of C. Moreover, let K be the Gaussian curvature of M, and let k_g be the geodesic curvature of C. Then

$$\sum_{\ell=1}^{L} (\alpha_l - \pi) = \int \int_D K \, \mathrm{d}\sigma + \int_C k_g \, \mathrm{d}s - 2\pi \chi.$$

For a proof, see Chern's book [5, pp. 125-126].

For a triangle $D = \Delta$, with v = 3 vertices, s = 3 sides, and f = 1 facet, $\chi_{\Delta} = v - s + f = 1$. If each side lies on a geodesic on M, then $k_g = 0$, whence

$$\sum_{\ell=1}^{3} \alpha_{\ell} = \pi + \int \int_{D} K \, \mathrm{d}\sigma.$$

Example 7. With the geodetic latitude λ and the longitude φ , the parametrization of the spheroidal earth surface takes the form

$$\rho = \frac{a \cos(\lambda)}{\sqrt{1 - e^2 [\sin(\lambda)]^2}},$$
$$x = \rho \cos(\varphi),$$
$$y = \rho \sin(\varphi),$$
$$z = \frac{a\sigma^2 \sin(\lambda)}{\sqrt{1 - e^2 [\sin(\lambda)]^2}}.$$

Hence, calculus gives the surface area

$$\mathrm{d}\sigma = \frac{a^2 \sigma^2 \cos(\lambda)}{\{1 - e^2 [\sin(\lambda)]^2\}^2} \mathrm{d}\varphi \,\mathrm{d}\lambda$$

and the Gaussian curvature

$$K(\varphi,\lambda) = \frac{\{1 - e^2[\sin(\lambda)]^2\}^2}{a^2\sigma^2},$$

which is the reciprocal of the product of the radii of curvature R' in the plane of the meridian and N in the perpendicular plane [44, pp. 24, 25]:

$$R' = \frac{a(1-e^2)}{\{1-e^2[\sin(\lambda)]^2\}^{3/2}}, \quad N = \frac{a}{\{1-e^2[\sin(\lambda)]^2\}^{1/2}}$$

Thus, with Ω being the domain of the parametrization of Δ ,

$$\int \int_{\Delta} K \, \mathrm{d}\sigma = \int \int_{\Omega} \frac{\{1 - e^2 [\sin(\lambda)]^2\}^2}{a^2 \sigma^2} \frac{a^2 \sigma^2 \cos(\lambda)}{\{1 - e^2 [\sin(\lambda)]^2\}^2} \mathrm{d}\varphi \, \mathrm{d}\lambda$$
$$= \int \int_{\Omega} \cos(\lambda) \, \mathrm{d}\varphi \, \mathrm{d}\lambda.$$

Example 8. Gauss investigated triangulations measured by De Krayenhof, for instance, the following internal angles of a spheroidal triangle [16, Section 23, p. 149]:

$$\alpha = 50^{\circ}58'15.238''$$
 at Harlingen,
 $\beta = 82^{\circ}47'15.351''$ at Leeuwarden,
 $\gamma = 46^{\circ}14'27.202''$ at Ballum.

In the plane, no such triangle exists, because the sum of the three angles $\alpha + \beta + \gamma = 179^{\circ}59'57.791''$ fails to equal 180°. On an ellipsoid, the sum of the internal angles in a geodesic triangle Δ exceeds 180° by the integral of the Gaussian curvature K over the triangle, which Gauss computed to be 1.749'' for this example, so that $\alpha + \beta + \gamma = 180^{\circ}0'1.749''$. In either case, it is impossible to place the three cities on a map without altering the data. One strategy consists in making the "smallest" adjustment while preserving $\alpha + \beta + \gamma = 180^{\circ}$, in other words, minimizing

$$\begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix} \begin{pmatrix} \alpha \\ \beta \\ \gamma \end{pmatrix} - \begin{pmatrix} 50^{\circ}58'15.238'' \\ 82^{\circ}47'15.351'' \\ 46^{\circ}14'27.202'' \end{pmatrix}$$

subject to the linear constraint

(...)

$$(1 \quad 1 \quad 1) \begin{pmatrix} \alpha \\ \beta \\ \gamma \end{pmatrix} = 180^{\circ} + \frac{180^{\circ}}{\pi} \int \int_{\Delta} K \, \mathrm{d}\sigma = 180^{\circ} 0' 1.749''.$$

More generally, with *n* measurements f_1, \ldots, f_n of quantities $\alpha_1, \ldots, \alpha_n$ subject to a constraint $\alpha_1 + \cdots + \alpha_n = d$, the system becomes

$$\begin{pmatrix} 1^* \\ I \end{pmatrix} \begin{pmatrix} \alpha_1 \\ \vdots \\ \alpha_n \end{pmatrix} = \begin{pmatrix} d \\ f_1 \\ \vdots \\ f_n \end{pmatrix}.$$

For the unitary factorization of the constraint equation,

$$r = C^{*}(; 1) = 1^{*} = (1, \dots, 1)^{*},$$

$$v = r + ||r||_{2}e_{1} = \begin{pmatrix} \sqrt{n} + 1 \\ 1 \\ \vdots \\ 1 \end{pmatrix},$$

$$v = \frac{1}{||r||_{2}(||r||_{2} + |r_{1}|)} = \frac{1}{\sqrt{n}(\sqrt{n} + 1)},$$

$$S = I - vvv^{*},$$

$$L = CS^{*} = -\sqrt{3}1^{*} = (-\sqrt{n}, 0, 0) = -\sqrt{n}e_{1}.$$

Consequently,

$$Lw_1 = d,$$

$$-\sqrt{n}w_1 = d,$$

$$w_1 = \frac{d}{-\sqrt{n}} = \frac{180^{\circ}0'1.749''}{-\sqrt{3}} = -103^{\circ}55'33.072''.$$

For the least-squares system, E = I. Consequently,

$$EQ = IS = S = (q_1; q_2, q_3, \dots, q_n),$$

and the least-squares system takes the form

$$(q_1; q_2 q_3 \dots q_n) \begin{pmatrix} 0 \\ w_2 \end{pmatrix} = f - q_1 w_1$$

Hence,

$$\begin{pmatrix} 0\\ w_2 \end{pmatrix} = \begin{pmatrix} q_1^*;\\ q_2^*\\ q_3^*\\ \vdots\\ q_n^* \end{pmatrix} (f - q_1 w_1).$$

Because $q_j^*q_1 = 0$ for every j > 1, the least-squares solution is

$$w_{2} = \begin{pmatrix} q_{2}^{*} \\ q_{3}^{*} \\ \vdots \\ q_{n}^{*} \end{pmatrix} (f - q_{1}w_{1}) = \begin{pmatrix} q_{2}^{*} \\ q_{3}^{*} \\ \vdots \\ q_{n}^{*} \end{pmatrix} (f),$$

and the first coordinate (in this example) gives the least-squares error

$$|Ex - f||_2 = q_1^*(f - q_1w_1) = q_1^*f - w_1 = -(1/\sqrt{n})1^*f + d/\sqrt{n}$$
$$= \frac{1}{\sqrt{n}} \left[180^\circ + \frac{180^\circ}{\pi} \int \int_A K d\sigma - (f_1 + f_2 + \dots + f_n) \right].$$

Reverting to the canonical basis through the inverse change of basis gives the solution

$$x = Sw = (q_1; q_2, q_3, \dots, q_n) \begin{pmatrix} w_1 \\ w_2 \end{pmatrix}$$
$$= (w_1 - q_1^* f)q_1 + (q_1; q_2, q_3, \dots, q_n) \begin{pmatrix} q_1^* f \\ q_2^* \\ q_3^* \\ q_n^* \end{pmatrix} (f)$$
$$= (w_1 - q_1^* f)q_1 + f$$

$$= \begin{pmatrix} 50^{\circ}58'16.557\,333''\\82^{\circ}47'16.670\,333''\\46^{\circ}14'28.521\,333'' \end{pmatrix} = \begin{pmatrix} \tilde{\alpha}\\ \tilde{\beta}\\ \tilde{\gamma} \end{pmatrix}$$

which add to $180^{\circ}0'1.749''$. The formula

$$x = (w_1 - q_1^* f)q_1 + f = \left(\frac{d}{\sqrt{n}} - \frac{1^* f}{\sqrt{n}}\right)\frac{1}{\sqrt{n}}1 + f$$

shows that the measurements f are all adjusted by the average discrepancy

$$\frac{d-1^*f}{n} = \frac{180^\circ 0'1.749'' - 179^\circ 59'57.791''}{3} = \frac{0^\circ 0'3.958''}{3} = 1.319\,333\dots''\,.$$

5. The singular-value decomposition and error analysis

5.1. The singular-value decomposition

Ordinary least-squares problems consist in determining the shortest vector \tilde{x} that minimizes $||A\tilde{x} - b||_2$, perhaps subject to linear constraints. If $\tilde{b}:=A\tilde{x}$, then the solution minimizes the discrepancy in the right-hand side, $||\tilde{b}-b||_2$, but it does *not* adjust the matrix A. In other words, $\tilde{x}=(\tilde{x}_1,\ldots,\tilde{x}_n)$ is the gradient of the linear function $u:\mathbb{C}^n \to \mathbb{C}$ that minimizes the average squared discrepancy between the measurement b_j and the value u[A(j;)], but it does *not* minimize the Euclidean distance from the graph of u (a hyperplane) to the data $(a_{j,1},\ldots,a_{j,n};b_j)$ in \mathbb{C}^{n+1} . Such more general problems of least squares admit solutions in term of a matrix factorization called the "singular-value decomposition" that was published independently by Eugenio Beltrami in 1873 and Camille Jordan in 1874, and extended by Erhard Schmidt in 1907 and Hermann Weyl in 1912. (For greater detail on the history of the singular value decomposition consult Stewart's account [45].)

Theorem 9. For each matrix $A \in M_{m \times n}(\mathbb{C})$ of rank r, there exist unitary matrices U, V, and a diagonal matrix Σ , such that

$$A = U\Sigma V^*$$

= $u_1 \sigma_1 v_1^* + \dots + u_r \sigma_r v_r^*$
= $\tilde{U} \tilde{\Sigma} \tilde{V}^*$

with $\sigma_i := \Sigma_{i,i}$ and with the following features.

- (U) The matrix $U \in M_{m \times m}(\mathbb{C})$ is unitary. The first r columns (u_1, \ldots, u_r) of U form an orthonormal basis for the range (column space) of A. The last m r columns (u_{r+1}, \ldots, u_m) of U form an orthonormal basis for the null space (kernel) of A^* .
- (V) The matrix $V \in \mathbb{M}_{n \times n}(\mathbb{C})$ is unitary. The first r columns (v_1, \ldots, v_r) of V form an orthonormal basis for the row space of A ([Kernel(A)]^{\perp}). The last n r columns (v_{r+1}, \ldots, v_n) of V form an orthonormal basis for the null space (kernel) of A.
- (Σ) The matrix $\Sigma \in \mathbb{M}_{m \times n}(\mathbb{C})$ is diagonal: $\Sigma_{k,\ell} = 0$ for all $k \neq \ell$, with

$$\sigma_1 \ge \sigma_2 \ge \cdots \ge \sigma_r > 0 = \sigma_{r+1} = \cdots = \sigma_{\min\{m,n\}}.$$

Moreover,
 $Av_j = \sigma_j u_j,$
 $A^*u_j = \sigma_j v_j,$
for every $j \in \{1, \dots, r\},$ and $Av_j = 0$ for every $j \in \{r+1, \dots, n\}.$ Finally,
 $\tilde{U} = (u_1, \dots, u_r) \in \mathbb{M}_{m \times r}(\mathbb{C}),$
 $\tilde{V} = (v_1, \dots, v_r) \in \mathbb{M}_{n \times r}(\mathbb{C}),$
 $\tilde{\Sigma} = \text{diagonal}(\sigma_1, \dots, \sigma_r) \in \mathbb{M}_{r \times r}(\mathbb{C}).$

Proof. Let $V = (v_1, \ldots, v_r, v_{r+1}, \ldots, v_n)$ be an orthonormal basis of eigen vectors for the hermitian positive semi-definite matrix $A^*A \in \mathbb{M}_{n \times n}(\mathbb{C})$, corresponding to its eigenvalues in nondecreasing order $\lambda_1 \ge \lambda_2 \ge \cdots \ge \lambda_r > 0 = \lambda_{r+1} = \cdots = \lambda_n$. Define $\sigma_j := \sqrt{\lambda_j}$, and $u_j := (1/\sigma_j)Av_j$ for every $j \in \{1, \ldots, r\}$. The remainder of the proof consists of straightforward verifications [30, Section 5.4; 46, Section 6.4]. \Box

Definition 10. The factorization $A = U\Sigma V^*$ is the singular-value decomposition of A. The scalars $\sigma_1, \ldots, \sigma_r$ are the singular values of A. The vectors v_1, \ldots, v_n are the right singular vectors of A. The vectors u_1, \ldots, u_m are the left singular vectors of A.

The singular-value decomposition also provides a means to solve ordinary least-squares problems. Firstly, the product \tilde{U}^*b projects b orthogonally on the column space of A, whence

$$||Ax - U^{T}b||_{2} \leq ||Ax - b||_{2}$$

for every $x \in \mathbb{C}^n$. Because $\tilde{U}^* b$ lies in the column space of A, there exists a solution $x \in \mathbb{C}^n$ such that $Ax = \tilde{U}^* b$. Secondly, every solution to this system differs from x by a vector in the null space of A. Consequently, the shortest solution is the orthogonal projection $x^{\dagger} = \tilde{V}^* x$ of x on the orthogonal complement of the null space of A. A derivation of a formula for x^{\dagger} can proceed as follows:

$$Ax = b$$
,

 $(\tilde{U}\tilde{\Sigma}\tilde{V}^*)x = b,$ $\tilde{\Sigma}(\tilde{V}^*x) = \tilde{U}^*b,$ $\tilde{V}^*x = \tilde{\Sigma}^{-1}\tilde{U}^*b,$ $x^{\dagger} = (\tilde{V}\tilde{\Sigma}^{-1}\tilde{U}^*)b.$

Definition 11. The *pseudoinverse* of A is the matrix

$$A^{\dagger} := \tilde{V} \tilde{\Sigma}^{-1} \tilde{U}^*.$$

Thus, the shortest least-squares solution of Ax = b is $x^{\dagger} := A^{\dagger}b$.

5.2. Norms and condition numbers of matrices

5.2.1. Norms of matrices

The concepts of norms and condition numbers for matrices provide means to estimate the propagation of errors from the data and during computations through the solutions of linear systems, as developed by Gastinel [15].

Definition 12. For each norm $\| \|_p$ on \mathbb{C}^n and each norm $\| \|_q$ on \mathbb{C}^m , the subordinate matrix norm $\| \|_{p,q}$ on $\mathbb{M}_{m \times n}(\mathbb{C})$ is defined by

$$||A||_{p,q} := \max\{||Au||_q : u \in \mathbb{C}^n, ||u||_p = 1\}$$
$$= \max\{||Au||_q/||u||_p : u \in \mathbb{C}^n, u \neq 0\}.$$

Example 13. With $||x||_{\infty} := \max_{j} |x_{j}|$ on \mathbb{C}^{n} and $||Ax||_{\infty}$ on \mathbb{C}^{m} ,

$$||A||_{\infty,\infty} = \max_{1 \le i \le m} \sum_{j=1}^{n} |A_{i,j}|$$

With $||x||_1 := \sum_j |x_j|$ on \mathbb{C}^n and $||Ax||_1$ on \mathbb{C}^m ,

$$||A||_{1,1} = \max_{1 \le j \le n} \sum_{i=1}^{m} |A_{i,j}|.$$

With $||x||_1$ on \mathbb{C}^n , and $||Ax||_{\infty}$ on \mathbb{C}^m ,

$$||A||_{1,\infty} = \max_{1 \leq i \leq m} \max_{1 \leq j \leq n} |A_{i,j}|.$$

(For $p \in \{1, \infty\}$ the formulae for $||x||_p$ and $||A||_{p,p}$ coincide.)

The following considerations show that $||A||_{2,2} = \sigma_1$ and $\kappa_{2,2}(A) = \sigma_1/\sigma_n$ is the ratio of the largest to the smallest singular values of A.

Lemma 14. For all real numbers $\sigma_1 \ge \sigma_2 \ge \cdots \ge \sigma_{n-1} \ge \sigma_n \ge 0$,

$$\min_{\|x\|_{2}=1} \sum_{i=1}^{n} (\sigma_{i}x_{i})^{2} = \min_{i \in \{1,...,n\}} \sigma_{i}^{2} = \sigma_{n}^{2},$$
$$\max_{\|x\|_{2}=1} \sum_{i=1}^{n} (\sigma_{i}x_{i})^{2} = \max_{i \in \{1,...,n\}} \sigma_{i}^{2} = \sigma_{1}^{2}.$$

Proof. Solving $\sum_{i=1}^{n} x_i^2 = 1$ for x_n^2 gives $x_n^2 = 1 - \sum_{i=1}^{n-1} x_i^2$. Hence

$$\sum_{i=1}^{n} (\sigma_i x_i)^2 = \sigma_n^2 x_n^2 + \sum_{i=1}^{n-1} (\sigma_i x_i)^2$$

$$= \sigma_n^2 \left(1 - \sum_{i=1}^{n-1} x_i^2 \right) + \sum_{i=1}^{n-1} \sigma_i^2 x_i^2$$
$$= \sigma_n^2 + \sum_{i=1}^{n-1} (\sigma_i^2 - \sigma_n^2) x_i^2$$
$$\ge \sigma_n^2,$$

with equality if and only if $x_i = 0$ for $\sigma_i \neq \sigma_n$. Similarly, $\sum_{i=1}^n (\sigma_i x_i)^2 = \sigma_1^2 + \sum_{i=2}^n (\sigma_i^2 - \sigma_1^2) x_i^2 \leq \sigma_1^2$ with equality if and only if $x_i = 0$ for $\sigma_i \neq \sigma_1$.

Proposition 15. For each matrix $A \in M_{m \times n}(\mathbb{C})$, the subordinate Euclidean norm $||A||_{2,2} = \sigma_1$ is the largest singular value of A. Moreover, $\kappa_{2,2}(A) = \sigma_1/\sigma_n$ is the ratio of the largest to the smallest singular values of A.

Proof. Consider a singular-value decomposition $A = U\Sigma V^*$. For each vector $x \in \mathbb{C}^n$ with $||x||_2 = 1$, let $w := V^*x$. Then $||w||_2 = ||x||_2 = 1$. Hence,

$$\|Ax\|_{2}^{2} = x^{*}A^{*}Ax = x^{*}(V\Sigma^{*}U^{*})(U\Sigma V^{*})x$$
$$= x^{*}V\Sigma^{*}\Sigma V^{*}x = \|\Sigma w\|_{2}^{2} = \sum_{i=1}^{n} (\sigma_{i}w_{i})^{2} \leqslant \sigma_{1}^{2}$$

with the maximum value reached for $w = e_1$, or, equivalently, $x = v_1$. Hence, $\kappa_2(A) = ||A||_2 ||A^{-1}||_2 = (\max_i \sigma_i)(\max_i \sigma_i^{-1}) = \sigma_1/\sigma_n$. \Box

For norms of the type $||x||_p := (|x_1|^p + \cdots + |x_n|^p)^{1/p}$ with $p,q \notin \{1,2,\infty\}$, no formula for the subordinate norm $||A||_{p,q}$ seems to be known [22, p. 124].

5.2.2. Condition numbers of matrices

For a square and invertible matrix $A \in M_{n \times n}(\mathbb{C})$, the condition number provides lower and upper bounds on the discrepancy $\|\tilde{x} - x\|_p$ between the solution x of a linear system Ax = b and any vector \tilde{x} . Such a vector \tilde{x} can result, for instance, from an attempt at solving the system with floating-point or any other approximate arithmetic. To this end, let $\tilde{b}:=A\tilde{x}$.

Definition 16. For each norm $\| \|_p$ on \mathbb{C}^n and each norm $\| \|_q$ on \mathbb{C}^m , the *condition number* $\kappa_{p,q}$ is defined by

$$\kappa_{p,q}(A) := \|A\|_{p,q} \|A^{\dagger}\|_{q,p}$$

Proposition 17. For all b, \tilde{b} , x, \tilde{x} and A invertible with Ax = b and $A\tilde{x} = \tilde{b}$,

$$\frac{1}{\kappa_{p,q}(A)} \frac{\|\tilde{b} - b\|_q}{\|b\|_q} \leqslant \frac{\|\tilde{x} - x\|_p}{\|x\|_p} \leqslant \kappa_{p,q}(A) \frac{\|\tilde{b} - b\|_q}{\|b\|_q}$$

Proof. Use $||b||_q = ||Ax||_q \le ||A||_{p,q} \cdot ||x||_p$ and $||\tilde{x} - x||_p = ||A^{-1}(\tilde{b} - b)||_p \le ||A^{-1}||_{q,p} \cdot ||\tilde{b} - b||_q$ [30, Section 4.4; 46, Section 4.4]. \Box

Proposition 17 compares the solutions x and \tilde{x} of two systems with right-hand sides b and b but with the same matrix A. In contrast, with different matrices A and C the following result holds. For each invertible $A \in \mathbb{M}_{n \times n}(\mathbb{C})$, for each $C \in \mathbb{M}_{n \times n}(\mathbb{C})$. If $||A - C|| < 1/||A^{-1}||$, then for each nonzero vector $b \in \mathbb{C}^n$ and for the solutions $x \in \mathbb{C}^n$ of Ax = b and $w \in \mathbb{C}^n$ of Cw = b,

$$\frac{\|w - x\|}{\|x\|} \leq \frac{\kappa(A)}{1 - \kappa(A) \cdot \|A - C\|/\|A\|} \cdot \frac{\|A - C\|}{\|A\|}$$

For a proof see [46, pp. 188–198], and for other similar error bounds see [22, Chapter 7]. Yet more generally, a theorem of Wedin for all matrices $A \in \mathbb{M}_{m \times n}(\mathbb{C})$ and $C \in \mathbb{M}_{m \times n}(\mathbb{C})$, with rank $r = n \leq m$, and for all vectors $b \in \mathbb{C}^m$ and $d \in \mathbb{C}^m$, if there exists a positive real ε for which

 $\kappa_{2}(A)\varepsilon < 1,$ $\|A - C\|_{2} \leq \varepsilon \|A\|_{2},$ $\|b - d\|_{2} \leq \varepsilon \|b\|_{2},$

then the *least-squares* solutions $\tilde{x} \in \mathbb{C}^n$ and $\tilde{z} \in \mathbb{C}^n$ of the systems Ax = b and Cz = d satisfy the following inequalities [22, Chapter 19]:

$$\frac{\|x-z\|_2}{\|x\|_2} \leqslant \frac{\kappa_2(A)\varepsilon}{1-\kappa_2(A)\varepsilon} \left\{ 2 + [1+\kappa_2(A)] \frac{\|b-Ax\|_2}{\|A\|_2 \cdot \|x\|_2} \right\},\\ \frac{\|(b-Ax) - (d-Az)\|_2}{\|b\|_2} \leqslant [1+2\kappa_2(A)]\varepsilon.$$

The following theorem of Kahan [25, pp. 775,776], who credits Gastinel, shows that for each invertible matrix A the distance to the closest singular matrix is $1/||A^{-1}||$.

Theorem 18. For every invertible matrix A and every subordinate norm:

$$\min_{\det(S)=0} \|A - S\| = \frac{1}{\|A^{-1}\|}.$$

Proof. For each singular matrix S there exists a vector $z \neq 0$ with Sz = 0:

$$||A - S|| \ge \frac{||(A - S)z||}{||z||} = \frac{||Az||}{||z||} = \frac{||A^{-1}|| ||Az||}{||A^{-1}|| ||z||} \ge \frac{||A^{-1}Az||}{||A^{-1}|| ||z||} = \frac{1}{||A^{-1}||}.$$

There exists a vector $y \neq 0$ with $||A^{-1}y|| = ||A^{-1}|| ||y||$. As in the Hahn–Banach theorem [48], choose a linear functional w dual to $A^{-1}y$, so that

$$w(A^{-1}y) = ||w|| \cdot ||A^{-1}y|| = 1,$$

let w^* be the matrix of w relative to the canonical basis, so that $w(z) = w^*z$ for every vector z, and define

$$S:=A - yw^*$$
.

Then S is singular, because

$$S(A^{-1}y) = (A - yw^*) \cdot (A^{-1}y) = y - y \cdot 1 = 0$$

Moreover,

$$||A - S|| = \max\{||(yw^*)x||: ||x|| = 1\}$$

= max{||y(w^*x)||: ||x|| = 1}
= ||y|| \cdot max{w^*x: ||x|| = 1}
= ||y|| \cdot ||w^*||
= ||y|| \cdot \frac{1}{||A^{-1}y||}
= ||y|| \cdot \frac{1}{||A^{-1}|| \cdot ||y||}
= \frac{1}{||A^{-1}||}. \Box

6. Matrix approximation and total least squares

6.1. The approximation theorems of Schmidt, Mirsky, and Weyl

A theorem of Schmidt [43], with later versions by Mirsky [34] and Weyl [52], approximates a matrix $C \in \mathbb{M}_{m \times n}(\mathbb{C})$ of rank r by a singular matrix $S \in \mathbb{M}_{m \times n}(\mathbb{C})$ of rank s < r that minimizes the Frobenius norm $||C - S||_F$, defined by

$$\|A\|_{F}^{2} := \sum_{i=1}^{m} \sum_{j=1}^{n} |A_{i,j}|^{2} = \sum_{i=1}^{m} \|A(i;)\|_{2}^{2} = \sum_{j=1}^{n} \|A(;j)\|_{2}^{2}.$$

All unitary matrices U and V preserve Euclidean and Frobenius norms:

$$\|UA\|_F^2 = \sum_{j=1}^n \|UA(;j)\|_2^2 = \sum_{j=1}^n \|A(;j)\|_2^2 = \|A\|_F,$$
$$\|AV\|_F = \sum_{i=1}^m \|A(i;)V\|_2^2 = \sum_{i=1}^m \|A(i;)\|_2^2 = \|A\|_F.$$

In particular, with a singular-value decomposition $A = U\Sigma V^*$,

$$||A||_F^2 = ||U\Sigma V^*||_F^2 = ||\Sigma||_F^2 = \sum_{i=1}^r \sigma_r^2 \ge \sigma_1^2 = ||A||_2^2.$$

The following theorem follows Stewart's version [45, pp. 561, 562].

Theorem 19. For each matrix $C \in M_{m \times n}(\mathbb{C})$ with $\sigma_1 \ge \cdots \ge \sigma_r > 0$ and

$$C = \sum_{i=1}^{\prime} \sigma_i u_i v_i^*$$

and for each matrix $S \in M_{m \times n}(\mathbb{C})$ of rank $k \in \{0, ..., r\}$,

$$\|C-S\|_F^2 \ge \sum_{i=k+1}^r \sigma_i^2$$

with the minimum $\sigma_{k+1}^2 + \cdots + \sigma_r^2$ reached for

$$S = \sum_{i=1}^k \sigma_i u_i v_i^*.$$

Proof. If k = r then the theorem holds because S = C. Henceforth, assume that k < r. Also, for each matrix A, let $\sigma_i(A)$, $u_i(A)$, and $v_i(A)$ be the *i*th singular value and singular vectors of A, and define

$$A_k := \sum_{i=1}^k \sigma_i(A) u_i(A) v_i^*(A).$$

The following argument shows that $\sigma_1(C-S) \ge \sigma_{k+1}(C)$. If *S* has rank *k* then *S* has a singular value decomposition $S = \sum_{i=1}^{k} \tau_i w_i z_i^* = W \tau Z^*$. Moreover, the linear space Z^{\perp} perpendicular to z_1, \ldots, z_k has dimension $n - k > n - (k + 1) \ge n - r$. Because the column space *V* spanned by v_1, \ldots, v_{k+1} has dimension k + 1, it follows that $Z^{\perp} \cap V \neq \{0\}$. Thus, there exists a non-zero vector of coefficients $\gamma \in \mathbb{C}^{k+1}$, for instance with $\|\gamma\|_2 = 1$, such that $x := V_{\gamma} = \sum_{i=1}^{k+1} \gamma_i v_i \in Z^{\perp} \cap V$, whence 0 = Zx and hence Sx = 0. Let $\tilde{\gamma} := (\gamma^*, 0^*)^* \in \mathbb{C}^n$:

$$\sigma_1^2(C-S) \ge x^*(C-S)^*(C-S)x$$
$$= x^*C^*Cx$$
$$= \tilde{\gamma}^*V^*V\Sigma U^*U\Sigma V^*V\tilde{\gamma}$$
$$= \tilde{\gamma}^*\Sigma^2\tilde{\gamma} = \sum_{i=1}^{k+1} (\gamma_i\sigma_i)^2$$
$$\ge \sigma_{k+1}^2.$$

The next argument provides an upper bound on the change in the largest singular value caused by a change in a matrix. From the reverse triangle inequality for norms, it follows that

$$\sigma_1(A - B) = ||A - B||_2 \ge |||A||_2 - ||B||_2| = |\sigma_1(A) - \sigma_1(B)|$$

The following generalization provides inequalities for the other singular values. For each matrix G and each index ℓ , $\sigma_1(G - G_\ell) = \sigma_{\ell+1}(G)$. Consequently, for all matrices $G, H \in \mathbb{M}_{m \times n}(\mathbb{C})$, and for all indices k and ℓ , the foregoing result leads to

$$\sigma_{\ell+1}(G) + \sigma_{k+1}(H) = \sigma_1(G - G_\ell) + \sigma_1(H - H_k) \ge \sigma_1([G - G_\ell] + [H - H_k])$$

$$= \sigma_1([G+H] - [G_\ell + H_k])$$

$$\geq \sigma_{\ell+k+1}(G+H)$$

because the rank of $G_{\ell} + H_k$ cannot exceed $\ell + k$. Equivalently, if A := G + H and B := H, then

$$\sigma_{\ell+1}(A-B) \geq \sigma_{\ell+k+1}(A) - \sigma_{k+1}(B)$$

Finally, in the particular case where S has rank k, setting G := C - S and H := S gives

$$\sigma_{\ell+1}(C-S) + 0 = \sigma_{\ell+1}(C-S) + \sigma_{k+1}(S) \ge \sigma_{\ell+1+k}(C)$$

Finally,

$$\|C-S\|_F^2 = \sum_{i=1}^r \sigma_i^2 (C-S) \ge \sum_{i=1}^r \sigma_{i+k}^2 (C) = (\sigma_{k+1}^2 + \dots + \sigma_r^2) (C).$$

Equality holds with $S = \sum_{i=1}^{k} \sigma_i u_i v_i^*$, for which $||C - S||_F^2 = ||\sigma_{k+1}u_{k+1}v_{k+1}^* + \dots + \sigma_r u_r v_r^*||_F^2 = \sigma_{k+1}^2 + \dots + \sigma_r^2$.

The approximation theorem of Schmidt, Mirsky, and Weyl amounts to identifying a matrix S minimizing a rotationally invariant norm ||C-S||, for instance, the Euclidean norm, or the Frobenius norm, subject to the linear constraints $\sigma_{k+1}(S) = \cdots = \sigma_n(S) = 0$. There also exist other types of constraints, for example, with the vector of singular values $\sigma = (\sigma_1(S), \dots, \sigma_n(S))$ subject to a linear system of constraints $K\sigma = d$ [37].

6.2. Total least squares

For a linear system Ax = b, the problem of *ordinary* least squares consists in determining the shortest vector \tilde{x} that minimizes the Euclidean norm of the discrepancy between b and $\tilde{b}:=A\tilde{x}$, possibly subject to constraints. In other words, the *ordinary* least-squares solution \tilde{x} solves *exactly* a related linear system $A\tilde{x} = \tilde{b}$ with $\|\tilde{b} - b\|_2$ minimum. In contrast, the problem of *total* least squares allows for minimal adjustments not only of b but also of A, also possibly subject to constraints. The problem of total least squares admit several mutually equivalent mathematical formulations. Their solutions in terms of singular value decompositions was published in 1980 by Golub and Van Loan [18;19, pp. 576–581]. Van Huffel and Vandewalle's monograph [50] describes further extensions and applications.

6.2.1. Geometric formulations of total least squares

Geometrically, the problem of total least squares amounts to fitting a hyperplane H minimizing the average squared Euclidean distance (measured perpendicularly to the fitted hyperplane) to data points c_1, \ldots, c_m in \mathbb{C}^{n+1} . The problem then reduces to finding a point $c_0 \in H$ and a non-zero normal vector $x \perp H$ that minimize the sum D of the squared distances:

$$D(x,c_0;c_1,\ldots,c_m):=\sum_{i=1}^m\frac{|\langle c_i-c_0,x\rangle|^2}{\langle x,x\rangle}.$$

To simplify notation, for every point $c_0 \in \mathbb{C}^{n+1}$, and for all data c_1, \ldots, c_m in \mathbb{C}^{n+1} , define a matrix $C_{c_0} \in \mathbb{M}_{m \times (n+1)}(\mathbb{C})$ with *i*th row $c_i^* - c_0^*$:

$$C_{c_0} \! := \! egin{pmatrix} c_1^* - c_0^* \ dots \ c_m^* - c_0^* \end{pmatrix}.$$

Consequently,

$$D(x, c_0; c_1, \ldots, c_m) = \frac{\|C_{c_0}x\|_2^2}{\|x\|_2^2}.$$

The following lemma reveals that an optimal hyperplane must pass through the centroid of the data,

$$\bar{c} = \frac{1}{m} \sum_{i=1}^m c_i,$$

which can thus serve as the point $c_0 \in H$.

Lemma 20. For every normal vector $x \in \mathbb{C}^{n+1} \setminus \{0\}$, for every point $c_0 \in \mathbb{C}^{n+1}$, and for all data c_1, \ldots, c_m in \mathbb{C}^{n+1} ,

 $D(x,c_0;c_1,\ldots,c_m) \ge D(x,\bar{c};c_1,\ldots,c_m),$

with equality if and only if $\langle x, (r - c_0) \rangle = \langle x, (r - \bar{c}) \rangle$ for every r. Consequently, a hyperplane of total least squares must pass through the centroid \bar{c} .

Proof. Consider the vector $w := C_{c_0} x$, so that $w_i = \langle c_i - c_0, x \rangle$ and

$$D(x, c_0; c_1, \ldots, c_m) = \frac{\|w\|_2^2}{\|x\|_2^2}.$$

Also, consider the vector $z := C_{\bar{c}}x$, so that $z_i = \langle c_i - \bar{c}, x \rangle$ and

$$D(x,\bar{c};c_1,\ldots,c_m)=\frac{\|z\|_2^2}{\|x\|_2^2}.$$

Moreover, define $1:=(1,\ldots,1) \in \mathbb{C}^m$, and $h:=\langle x,(\bar{c}-c_0)\rangle$, so that

$$w = z + h1.$$

Then $z \perp 1$:

$$\langle z,1\rangle = 1^* (C_{\bar{c}}x) = (1^*C_{\bar{c}})x = \left(m\bar{c}^* - \sum_{j=1}^m c_j^*\right)x = 0^*x = 0.$$

Finally, the Pythagorean Theorem applied to $z \perp 1$ and w = z + h1 gives

$$D(x, c_0; c_1, \dots, c_m) = \|w\|_2^2 / \|x\|_2^2$$
$$= (\|z\|_2^2 + h^2 \|1\|_2^2) / \|x\|_2^2$$

$$= D(x, \bar{c}; c_1, \dots, c_m) + h^2 m / ||x||_2^2$$

$$\ge D(x, \bar{c}; c_1, \dots, c_m),$$

with equality if and only if $0 = h = \langle x, (\bar{c} - c_0) \rangle$, which means that c_0 also lies in the hyperplane passing through \bar{c} perpendicularly to x. \Box

The following lemma reveals that an optimal normal vector must be a right-singular vector corresponding to the smallest singular value of C_{c_0} .

Lemma 21. For every point $c_0 \in \mathbb{C}^{n+1}$ and all data c_1, \ldots, c_m in \mathbb{C}^{n+1} , let v be a right-singular vector corresponding to the smallest singular value σ of C_{c_0} . Then for every vector $x \in \mathbb{C}^{n+1} \setminus \{0\}$, the following inequality holds:

 $D(x,c_0;c_1,\ldots,c_m) \ge D(v,c_0;c_1,\ldots,c_m),$

with equality if, but only if, x is also a right-singular vector corresponding to the smallest singular value σ of C_{c_0} . Consequently, a hyperplane of total least squares must be perpendicular to such a singular vector. Moreover,

 $D(v,c_0;c_1,\ldots,c_m)=\sigma^2.$

Proof. From $D(x, c_0; c_1, ..., c_m) = ||C_{c_0}x||_2^2/||x||_2^2$ it follows that D reaches its minimum at a unit vector $v = x/||x||_2$ that minimizes $||C_{c_0}v||_2$. The theory of the SVD shows that v coincides with any singular vector v for to the smallest singular value σ of C_{c_0} , with $D(v, c_0; c_1, ..., c_m) = ||C_{c_0}v||_2^2 = \sigma^2$. \Box

Theorem 22. For every set of data points c_1, \ldots, c_m in \mathbb{C}^{n+1} , each hyperplane of total least-squares passes through the centroid of the data \bar{c} perpendicularly to a right-singular vector v corresponding the smallest singular value σ of the matrix $C_{\bar{c}}$ with ith row $c_i^* - \bar{c}^*$. Moreover, for such a hyperplane, the sum of the squared distances to the data is σ^2 :

Proof. Combine the proofs of Lemmas 20 and 21. \Box

The matrix *C* can have a multiple smallest singular value $\sigma = \sigma_{k+1} = \cdots = \sigma_{k+\ell}$, corresponding to a linear subspace $V_{\sigma} \subseteq \mathbb{C}^{n+1}$ spanned by multiple singular vectors $v_{k+1}, \ldots, v_{k+\ell}$. In this situation, there exists a set \mathscr{H} of hyperplanes of total least squares, with each hyperplane $H \in \mathscr{H}$ perpendicular to a vector $v \in V$ and containing the "axis" $\overline{c} + V^{\perp}$. In particular, if $\sigma = 0$, then the data lies at the intersection $\cap \mathscr{H}$ of all such hyperplanes, which is an affine subspace $\overline{c} + V^{\perp}$ of dimension $n+1-\ell$. For example, if n+1=3 and $\ell = 2$, then $n+1-\ell = 1$ and all the data points lie on a common straight line in space.

With x = v and σ computed, the vector

$$\hat{c}_i := c_i - \langle c_i - \bar{c}, v \rangle v$$

is the orthogonal projection of the data c_i onto H. Consequently,

$$\sum_{i=1}^{m} \|\hat{c}_i - c_i\|_2^2 = D(v, \bar{c}; c_1, \dots, c_m) = \sigma^2.$$



Fig. 6. The TLS plane minimizes the Euclidean distance to the data.

Example 23. Consider the four data points in space displayed in Fig. 6:

 $c_1 = (11 \ 45 \ 38),$ $c_2 = (47 \ 54 \ 38),$ $c_3 = (17 \ 12 \ 14),$ $c_4 = (21 \ 29 \ 58).$

For these data, $\bar{c} = (\frac{1}{4}) \sum_{i=1}^{4} c_i = (24, 35, 37)$, and

$$C_{\bar{c}} = \begin{pmatrix} c_1 - \bar{c} \\ c_2 - \bar{c} \\ c_3 - \bar{c} \\ c_4 - \bar{c} \end{pmatrix} = \begin{pmatrix} -13 & 10 & 1 \\ 23 & 19 & 1 \\ -7 & -23 & -23 \\ -3 & -6 & 21 \end{pmatrix}$$

The smallest singular value of $C_{\bar{c}}$ is $\sigma = \sigma_3 = 18$, and the corresponding singular vector is $v_3 = (\frac{2}{3}, -\frac{2}{3}, \frac{1}{3})^*$. Thus, the hyperplane *H* passes through $c_0 = \bar{c} = (24, 35, 37)$ and lies perpendicularly to the vector $x = v_3 = (\frac{2}{3}, -\frac{2}{3}, \frac{1}{3})^*$, so that *H* satisfies the equation

$$\frac{2}{3}(x-24) - \frac{2}{3}(y-35) + \frac{1}{3}(z-37) = 0.$$

Moreover, $C_{\bar{c}}v = (-15, 3, 3, 9)^*$ contains the signed distances $d(c_i, H)$ from the data points to the hyperplane H, here $\sum_{i=1}^4 d(c_i, H)^2 = \|C_{\bar{c}}v\|_2^2 = \sigma^2 = 18^2$, which gives the orthogonal projections

 $\hat{c}_1, \ldots, \hat{c}_4$ of the data on *H*:

$$\hat{C} = \begin{pmatrix} \hat{c}_1 \\ \vdots \\ \hat{c}_m \end{pmatrix} = \begin{pmatrix} c_1 \\ \vdots \\ c_m \end{pmatrix} - \begin{pmatrix} -15 \\ 3 \\ 3 \\ 9 \end{pmatrix} \begin{pmatrix} 2 \\ 3 \\ -\frac{2}{3} \\ -\frac{2}{3} \\ \frac{1}{3} \end{pmatrix} = \begin{pmatrix} 21 & 35 & 43 \\ 45 & 56 & 37 \\ 15 & 14 & 13 \\ 15 & 35 & 55 \end{pmatrix}$$

6.2.2. Algebraic formulations of total least squares

For a linear system Ax = b, the problem of *ordinary* least squares consists in determining a vector \tilde{b} that minimizes $\|\tilde{b} - b\|_2$ subject to the constraint that the system $A\tilde{x} = \tilde{b}$ have a solution. More generally, the algebraic formulation of the problem of *total* least squares consists in determining a *matrix* \hat{A} and a vector \hat{b} that minimize the Frobenius norm

 $\|[\hat{A};\hat{b}]-[A;b]\|_{\rm F}$

subject to the condition that the system

$$\hat{A}\hat{x} = \hat{b}$$

have a solution. Thus, with

$$C:=[A; b],$$

 $\hat{C}:=[\hat{A}; \hat{b}],$

the problem reduces to determining a matrix \hat{C} that minimizes $\|\hat{C} - C\|_{\rm F}$ subject to the condition that Kernel (\hat{C}) contains a vector of the form $(\hat{x}^*, -1)^*$.

If the matrix C = [A; b] has a singular-value decomposition

$$C = \sum_{i=1}^{n+1} \sigma_i u_i v_i^*$$

and the n + 1 columns of *C* are linearly independent, then $\hat{C} = [\hat{A}, \hat{b}]$ must have rank at most *n*, and the Schmidt–Mirsky approximation theorem states that the closest such matrix is

$$\hat{C} = \sum_{i=1}^{n} \sigma_{i} u_{i} v_{i}^{*} = [A; b] - \sigma_{n+1} u_{n+1} v_{n+1}^{*}.$$

Then v_{n+1} spans Kernel (\hat{C}) , and two cases arise. If $(v_{n+1})_{n+1} \neq 0$, then the problem admits the solution

$$\binom{\hat{x}}{-1} = \frac{-1}{(v_{n+1})_{n+1}} v_{n+1}.$$

If $(v_{n+1})_{n+1} = 0$, then the problem has no solution.

Some practical problems lead to problems of total least squares subject to the condition that the first k columns $A_1:=(A(;1),\ldots,A(;k))$ of the matrix $A \in M_{m \times n}(\mathbb{C})$ be kept exact (unadjusted) while the last n - k columns $A_2:=(A(;k+1),\ldots,A(;n))$ are subject to adjustments. Golub, Hoffman, and Stewart published the following method of solution [17]. If A_1 has rank $\ell \leq r$, if Q is the

orthogonal projection on the column space of A_1 and Q^{\perp} is the orthogonal projection on its orthogonal complement, then the matrix $[A_1; \hat{A}_2]$ of rank r that minimizes the Frobenius norm

 $\|[A_1; \hat{A}_2] - [A_1; A_2]\|_{\mathrm{F}}$

is defined in terms of a singular-value decomposition of $Q^{\perp}A_2$,

$$Q^{\perp}A_2 = \sum_{i=1}^{n-k} \tau_i w_i z_i^*,$$

by the formula

$$\hat{A}_2 := QA_2 + \sum_{i=1}^{r-\ell} \tau_i w_i z_i^*.$$

Example 24. To the data in Example 1, Laplace fitted an affine model for the length Δs of and arc of 1 grad along the meridian,

$$c_0 + c_1[\sin(\lambda)]^2 = \Delta s.$$

The coefficient 1 of c_0 is exact. Consequently, the first column $A_1:=1$ remains fixed. The orthogonal projection on the space spanned by $1 \in \mathbb{C}^m$ has matrix $Q:=1/m \, 11^*$. Thus, $Q^{\perp} = I - Q$ subtracts from each column the mean of that column. Here, the matrix $C:=Q^{\perp}A_2$ corresponds to a linear system for c_1 :

$$\begin{pmatrix} 0.00\ 000 - 0.43\ 925\\ 0.30\ 156 - 0.43\ 925\\ 0.39\ 946 - 0.43\ 925\\ 0.46\ 541 - 0.43\ 925\\ 0.52\ 093 - 0.43\ 925\\ 0.54\ 850 - 0.43\ 925\\ 0.83\ 887 - 0.43\ 925 \end{pmatrix} (c_1) = \begin{pmatrix} 25\ 538.85 - 25\ 659.93\\ 25\ 666.65 - 25\ 659.93\\ 25\ 599.60 - 25\ 659.93\\ 25\ 640.55 - 25\ 659.93\\ 25\ 683.30 - 25\ 659.93\\ 25\ 683.30 - 25\ 659.93\\ 25\ 832.25 - 25\ 659.93\\ 25\ 832.25 - 25\ 659.93 \end{pmatrix}$$

The singular-value decomposition of C (computed with the command SVD on the HP 48GX [21, pp. 14-22]) shows two singular values,

 $221.279 = \sigma_1 > \sigma_2 = 0.266\,719,$

the smallest of which corresponds to the right singular vector

$$\begin{pmatrix} \hat{x} \\ -1 \end{pmatrix} = \frac{-1}{(v_2)_2} v_2 = \frac{-1}{0.002\,562} \begin{pmatrix} -0.999\,997 \\ 0.002\,562 \end{pmatrix} = \begin{pmatrix} 390.356 \\ -1 \end{pmatrix}.$$

Adding QA_2 amounts to adding back the means, which yields

$$\hat{c}_1 = 390.356,$$

 $\hat{c}_0 = 25\,659.93 - 390.356 * 0.43\,925 = 25\,488.46.$

These values lead to the estimate of the squared eccentricity $\hat{e}^2 = (\frac{2}{3})\hat{c}_1/\hat{c}_0 = 0.010210$, which is farther from the current estimate $e^2 = 0.00669437999013$ than the value obtained by ordinary least squares $\tilde{e}^2 = 0.006339$. The same values then lead to the estimate of the equatorial radius $\hat{a} = \hat{c}_0/(1-\hat{e}^2) = 99352$ m, which is closer to the current estimate a = 6378137.00000 m than the value obtained by ordinary least squares $\tilde{a}=100170.25$ m, but still very inaccurate. The corresponding results for weighted total least squares, with C = W[A; b], are $\hat{e}^2 = 0.008153$ and $\hat{a} = 100307$ m. Unweighted least squares, ordinary or total, give yet worse results.

Because the estimate of the eccentricity closest to the current estimate comes from ordinary least squares, rather than total least squares, such results suggest than most of the errors lie in the measurement of lengths along the meridian, rather than in the geodetic latitudes of the locations. Bowditch's comments corroborate such suggestions [31, Book II, Section 41].

For a system AX = B with r right-hand sides, the problem of total least squares consists in determining matrices $\hat{A} \in \mathbb{M}_{m \times n}(\mathbb{C})$, $\hat{B} \in \mathbb{M}_{m \times r}(\mathbb{C})$, and $\hat{X} \in \mathbb{M}_{n \times r}(\mathbb{C})$, minimizing $||[A;B] - [\hat{A};\hat{B}]||_F$ subject to the constraint $\hat{A}\hat{X} = \hat{B}$. Equivalently, with $I \in \mathbb{M}_{r \times r}(\mathbb{C})$, there must exist a solution $[\hat{X}^*; -I]^*$ to the system $[\hat{A}; \hat{B}] [\hat{X}^*; -I]^* = O$. In particular, because the last r rows of $[\hat{X}^*; -I]^*$ are linearly independent, it follows that the rank of $S:=[\hat{A}; \hat{B}]$ cannot exceed (n+r) - r = n. Therefore, with

$$C:=[A;B]=\sum_{i=1}^{n+r}\sigma_i u_i v_i^*$$

the matrix S must be a matrix of rank at most n that minimizes $||C - S||_F$. By the approximation theorem of Schmidt, Mirsky, and Weyl, it follows that

$$S = [\hat{A}; \hat{B}] = \sum_{i=1}^{n} \sigma_i u_i v_i^*.$$

The problem then admits a solution in the form $[\hat{X}^*; -I]^*$ if and only if in the matrix $(v_{n+1}, \ldots, v_{n+r})$ the last n+r rows are linearly independent and hence form in invertible matrix $V_{n+1;n+r} \in \mathbb{M}_{r \times r}(\mathbb{C})$, so that

$$\begin{pmatrix} X\\ -I \end{pmatrix} = V_{n+1;n+r}^{-1}(v_{n+1},\ldots,v_{n+r}).$$

6.3. Relations between the algebraic and geometric formulations

The matrix $\hat{C} = [\hat{A}; \hat{b}]$ in the algebraic formulation corresponds to the matrix C_0 in the geometric formulation. In other words, the algebraic formulation corresponds to the problem of fitting to the rows of C = [A; b] a hyperplane constrained to pass through the origin $0 \in \mathbb{C}^{n+1}$ instead of through the centroid \bar{c} . Indeed, the system $\hat{A}\hat{x} = \hat{b}$ in the form

$$[\hat{A};\hat{b}]\begin{pmatrix}\hat{x}\\-1\end{pmatrix}=0$$

states that every row of $[\hat{A}; \hat{b}]$ lies on the hyperplane *H* passing through the origin perpendicularly to $(\hat{x}^*; -1)^*$. Moreover, the condition that \hat{C} minimizes the Frobenius norm, or, equivalently, its square,

$$\|\hat{C} - C\|_F^2 = \sum_{i=1}^m \sum_{j=1}^{n+1} (c_{i,j} - \hat{c}_{i,j})^2 = \sum_{i=1}^m \|c_i - \hat{c}_i\|_2^2,$$

shows that \hat{C} minimizes the sum of the squared distances from the rows of C in \mathbb{C}^{n+1} to the rows of \hat{C} on H. Therefore, the *i*th row $\hat{C}(i;)$ of \hat{C} is the orthogonal projection of the *i*th row C(i;) of C, for otherwise these orthogonal projections would lie on the same hyperplane H and would give a smaller total least squares, or squared Frobenius norm. However, Lemma 21 shows that v_{n+1} and hence $(\hat{x}^*; -1)^*$ is the normal direction of total least squares for all the hyperplanes passing through the origin.

7. Nonlinear least squares

7.1. Nonlinear least squares in astronomy and geodesy

The old problem of estimating the shape of the earth can be formulated as the total least-squares problem of fitting an ellipsoid by minimizing the sum of the squared distances to data points. The problem of reliably computing the distance from a point to an ellipse already causes difficulties, because it amounts to solving a quartic equation, and there does not seem to be any practical forward error bounds for the solutions by the quartic formulae. Nevertheless, for small eccentricities $(e^2 < 2 - \sqrt{2})$, there exists a provably reliable algorithm to compute the distance with a contracting map [28,38].

Similarly, the old problem of estimating the shape of the orbit of a celestial body can be formulated as the total least-squares problem of fitting a plane and a conic section in it by minimizing the sum of the squared distances to data points [41].

For both problems, the particular formulation depends on the type of data, for instance, azimuth and elevation only, or azimuth, elevation, and range (measured by radar, for instance) [41, Chapter 10; 51, pp. 302–305]. Despite the practicality of such problems, however, there does not yet seem to exist any theorem that guarantees the global convergence of any algorithm toward the globally optimum surface or orbit [41, p. 180].

7.2. Fitting circles by total least squares

Although the problem of fitting circles and spheres to observations can be traced to the first millenium B.C., the problem of designing an *algorithm* to calculate the center and the radius of a circle or a sphere fitted to a finite set of points can be traced to the 1970's A.D., through computer scientists' developments of algorithms [8] for medical devices [3] and typography [40], electrical engineers' adjustments of microwave calibrations [27], and particle physicists' writings on fitting circular trajectories to a large number of automated measurements of positions of electrically charged particles within uniform magnetic fields [9]. One method — called an *algebraic fit* — to
fit a circle or a sphere to data points $c_1, \ldots, c_m \in \mathbb{R}^n$ consists in computing the center $x \in \mathbb{R}^n$ and the radius $r \in \mathbb{R}$ that minimize the function $f : \mathbb{R}^{n+1} \to \mathbb{R}$ defined by

$$f(x,r;c_1,\ldots,c_m):=\sum_{j=1}^m (\|x-c_j\|_2^2-r^2)^2=\sum_{j=1}^m (x^*x-r^2-2x^*c_j+c_j^*c_j)^2.$$

For each center $x \in \mathbb{R}^n$ the radius r(x) that minimizes f is such that

$$[r(x)]^{2} = \frac{1}{m} \sum_{j=1}^{m} ||x - c_{j}||_{2}^{2}.$$

Substituting r(x) for r in f leads to a linear system Ax = b for the center x, where A is 8m times the covariance matrix of the data, and the vector b is defined by $b_i := \sum_{j=1}^m \{(c_j)_i - \bar{c}_i\} \|c_i\|_2^2$ [35]. With an approximate arithmetic, however, the computation of the entries of A and b can introduce errors.

An alternate method by Coope [7] performs the charge of coordinates

$$z:=2x, \quad z_{n+1}:=r^2-x^*x,$$

which leads to the following ordinary linear least-squares system Cz = d:

$$\begin{pmatrix} c_1^* & 1 \\ \vdots \\ c_m^* & 1 \end{pmatrix} \begin{pmatrix} z \\ z_{n+1} \end{pmatrix} = \begin{pmatrix} \|c_1\|_2^2 \\ \vdots \\ \|c_m\|_2^2 \end{pmatrix}.$$

Thus with Coope's method forming the matrix C does not involve any computations and hence does not introduce any computational approximation.

However, Gander, Golub, and Strebel have demonstrated with examples that curves fitted by minimizing such algebraic objectives as f can lie farther from the data than curves fitted by total least squares of the Euclidean distances, called *geometric fits* [14]. Therefore, several authors have suggested using algebraic fits as initial estimates to start iterative methods for the computation of geometric fits [1, p. 357; 7, 14]. Several problems remain unsolved yet.

Example 25. Consider the following data in the plane:

$$c_2:=(0,2),$$

 $c_4:=(0,0),$
 $c_3:=(-\sqrt{3},-1), \quad c_1:=(-\sqrt{3},-1),$

with the sum of the squared distances to the circle with center x and radius r:

1)

$$g(x,r;c_1,\ldots,c_m) := \sum_{j=1}^m (\|x-c_j\|_2 - r)^2.$$

Firstly, calculus yields three minima for g, corresponding to circles with centers x_k opposite to c_k , radius $r:=\frac{7}{4}$, and $g(x_k,r;c_1,\ldots,c_m)=2$ for each k:

$$x_1:=e^{4\pi i/3}x_2,$$
 $x_3:=e^{2\pi i/3}x_2,$
 $x_2:=(0,-\frac{3}{4}).$

Secondly, perturbations of any of the data can turn any of the local minima into a global minimum. Finally, the algebraic fit cannot serve as an initial estimate for Newton's methods to converge to any geometric fit. Indeed, the algebraically fitted circle has its center at the origin, which coincides with a data point, where the objective function is not differentiable.

7.3. Open problems in nonlinear least squares

Problems of fitting *affine* manifolds by minimizing a weighted sum of squared distances to data are extensively documented. Indeed, linear algebra yields affine parametrizations of their solutions and provides several methods of solution through orthogonal factorizations, for which there exist proven upper bounds on errors from the data or from approximate computations [1,19,22,32,50,51,53].

In contrast, problems of nonlinear least squares, for instance, problems of fitting nonaffine manifolds as simple as circles, remain mostly unsolved. There exist a substantial documentation of algorithms that converge globally (from every initial point) to a *local* minimum [12,29,42]. However, some of their shortcuts can succumb to rounding errors [36], and there does not yet seem to exist any theorem guaranteeing the convergence to a *global* minimum to fit curves as simple as conic sections and surfaces as simple as spheres.

References

- [1] Å. Björk, Numerical Methods for Least Squares Problems, Society for Industrial and Applied Mathematics, Philadelphia, PA, 1996.
- [2] A.C. Aiken, On least squares and linear combinations of observations, Proc. Roy. Soc. Edinburgh, Section A 55 (1934) 42–47.
- [3] F.L. Bookstein, Fitting conic sections to scattered data, Comput. Graphics Image Process. 9 (1979) 56-71.
- [4] L.A. Brown, The Story of Maps, Dover, New York, NY, 1979.
- [5] S.-s. Chern, Curves and surfaces in Euclidean space, in: S.-s. Chern (Ed.), Global Differential Geometry, MAA Studies in Mathematics, Vol. 27, Mathematical Association of America, Washington, D.C, 1989, pp. 99–139.
- [6] V. Chvátal, Linear Programming, W.H. Freeman, New York, NY, 1983.
- [7] I.D. Coope, Circle fitting by linear and nonlinear least squares, J. Optim. Theory Appl. 76 (2) (1993) 381-388.
- [8] M.G. Cox, H.M. Jones, An algorithm for least-squares circle fitting to data with specified uncertainty ellipses, IMA J. Numer. Anal. 9 (3) (1989) 285–298.
- [9] J.F. Crawford, A non-iterative method for fitting circular arcs to measured points, Nucl. Instr. and Meth. 211 (2) (1983) 223-225.
- [10] G.B. Dantzig, Inductive proof of the simplex method, IBM J. Res. Dev. 4 (1960) 505-506.
- [11] G.B. Dantzig, Linear Programming and Extensions, Princeton University Press, Princeton, NJ, 1963.
- [12] J.E. Dennis, Jr., R.B. Schnabel, Numerical Methods for Unconstrained Optimization and Nonlinear Equations, Classics In Applied Mathematics, Vol. 16, Society for Industrial and Applied Mathematics, Philadelphia, PA, 1996.
- [13] J.L.E. Dreyer, A History of Astronomy from Thales to Kepler, 2nd Edition, Dover, New York, NY, 1953.
- [14] W. Gander, G.H. Golub, R. Strebel, Least-squares fitting of circles and ellipses, BIT 34 (1994) 558-578.

- [15] N. Gastinel, Matrices du Second Degré et Normes Générales en Analyse Numérique Linéaire, Thèse d'Etat, Université de Grenoble, 1960.
- [16] C.F. Gauss, Theoria Combinationis Observationum Erroribus Minimis Obnoxiae, Society for Industrial and Applied Mathematics, Philadelphia, PA, 1995. (Theory of the Combination of Observations Least Subject to Errors, original with translation by Gilbert W. Stewart).
- [17] G.H. Golub, A. Hoffman, G.W. Stewart, A generalization of the Eckart-Young-Mirsky matrix approximation theorem, Linear Algebra Appl. 88/89 (1987) 317–327.
- [18] G.H. Golub, C.F.V. Loan, An analysis of the total least squares problem, SIAM J. Numer. Anal. 17 (6) (1980) 883–893.
- [19] G.H. Golub, C.F.V. Loan, Matrix Computations, 2nd Edition, Johns Hopkins University Press, Baltimore, MD, 1989.
- [20] N. Grossman, The Sheer Joy of Celestial Mechanics, Birkhäuser, Boston, Cambridge, MA, 1996.
- [21] Hewlett-Packard Co., Corvallis Division, 1000 NE Circle Blvd., Corvallis, OR 97330, USA. HP 48G Series User's Guide, 7th edition, March 1994. (Part Number 00048-90126.).
- [22] N.J. Higham, Accuracy and Stability of Numerical Algorithms, Society for Industrial and Applied Mathematics, Philadelphia, PA, 1996.
- [23] B. Hofmann-Wellenhof, H. Lichteneggar, J. Collins, Global Positioning System: Theory and Practice, 4th revised Edition, Springer-Verlag, Wien, 1997.
- [24] R.V. Hogg, A.T. Craig, Introduction to Mathematical Statistics, 4th Edition, Macmillan, New York, NY, 1978.
- [25] W.M. Kahan, Numerical linear algebra, Canad. Math. Bull. 9 (6) (1966) 757-801.
- [26] H. Karttunen, P. Kröger, H. Oja, M. Poutanen, K.J. Donner (Eds.), Fundamental Astronomy, 2nd enlarged Edition, Springer, New York, 1994.
- [27] I. Kasa, A circle fitting procedure and its error analysis, IEEE Trans. Instr. Meas. 25 (1) (1976) 8–14.
- [28] S.P. Keeler, Y. Nievergelt, Computing geodetic coordinates, SIAM Rev. 40 (2) (1998) 300-309.
- [29] C.T. Kelley, Iterative Methods for Linear and Nonlinear Equations, Frontiers in Applied Mathematics, Vol. 16, Society for Industrial and Applied Mathematics, Philadelphia, PA, 1995.
- [30] D. Kincaid, W. Cheney, Numerical Analysis: Mathematics of Scientific Computing, Brooks/Cole, Pacific Grove, CA, 1996.
- [31] P.S.d. Laplace, Traité de Mécanique Céleste, Durprat, Paris, 1798–1825. (Translated with commentary by Nathaniel Bowditch, Boston, MA, 1829; Chelsea, Bronx, NY, 1966).
- [32] C.L. Lawson, R.J. Hanson, Solving Least Squares Problems, Classics In Applied Mathematics, Vol. 15, Society for Industrial and Applied Mathematics, Philadelphia, PA, 1995.
- [33] A.M. Legendre, Nouvelle méthodes pour la déterminiation des orbites des comètes, Courcier, Paris, 1805.
- [34] L. Mirsky, Symmetric gauge functions and unitarily invariant norms, Quart. J. Math. Oxford, Ser. 2 11 (41) (1960) 50–59.
- [35] Y. Nievergelt, Computing circles and spheres of arithmetic least squares, Comput. Phys. Comm. 81 (3) (1994) 343–350.
- [36] Y. Nievergelt, The condition of Steffensen's acceleration in several variables, J. Comput. Appl. Math. 58 (1995) 291–305.
- [37] Y. Nievergelt, Schmidt-Mirsky matrix approximation with linearly constrained singular values, Linear Algebra Appl. 261 (1997) 207–219.
- [38] Y. Nievergelt, S.P. Keeler, Computing geodetic coordinates in space, J. Spacecraft Rockets 37 (2000) 293–296.
- [39] B. Pourciau, Reading the master: Newton and the birth of celestial mechanics, Amer. Math. Mon. 104 (1) (1997) 1–19.
- [40] V. Pratt, Direct least-squares fitting of algebraic surfaces, ACM Comput. Graphics 21 (4) (1987) 145–152.
- [41] J.E. Prussing, B.A. Conway, Orbital Mechanics, Oxford University Press, New York, NY, 1993.
- [42] W.C. Rheinboldt, Methods for Solving Systems of Nonlinear Equations, CBMS-NSF Regional Conference Series in Applied Mathematics, Vol. 14, Society for Industrial and Applied Mathematics, Philadelphia, PA, 1974.
- [43] E. Schmidt, Zur Theorie der linearen und nichtlinearen Integralgleichungen. 1. Teil: Entwicklung willkürlicher Funktionen nach Systemen vorgeschriebener, Math. Ann. 63 (1907) 433–476.
- [44] J.P. Snyder, Map Projections A Working Manual (U.S. Geological Survey Professional Paper 1395). United States Government Printing Office, Washington, D.C. 20402, 1987.
- [45] G.W. Stewart, On the early history of the singular value decomposition, SIAM Rev. 35 (4) (1993) 551-566.

- [46] J. Stoer, R. Bulirsch, Introduction to Numerical Analysis, 2nd Edition, Springer, New York, NY, 1993.
- [47] G. Strang, Introduction to Applied Mathematics, Wellesley-Cambridge Press, Wellesley, MA, 1986.
- [48] A.E. Taylor, Introduction to Functional Analysis, Wiley, New York, NY, 1958.
- [49] B.L. van der Waerden, Geometry and Algebra in Ancient Civilizations, Springer, New York, NY, 1983.
- [50] S. Van Huffel, J. Vandewalle, The Total Least Squares Problem: Computational Aspects and Analysis, Society for Industrial and Applied Mathematics, Philadelphia, PA, 1991.
- [51] C.F. Van Loan, Introduction to Scientific Computing, Prentice-Hall, Upper Saddle River, NJ, 1997.
- [52] H. Weyl, Das asymptotische Vertielungsgesetz der Eigenwert linearer partieller Differentialgleichungen (mit einer Anwendung auf der Theorie der Hohlraumstrahlung), Math. Ann. 71 (1912) 441–479.
- [53] J.H. Wilkinson, The Algebraic Eigenvalue Problem, Oxford Science Publication, Monographs on Numerical Analysis, Oxford University Press, Oxford, UK, paperback edition, 1988.



Journal of Computational and Applied Mathematics 121 (2000) 73-94

JOURNAL OF COMPUTATIONAL AND APPLIED MATHEMATICS

www.elsevier.nl/locate/cam

Shape-preserving approximation by polynomials

D. Leviatan

School of Mathematical Sciences, Sackler Faculty of Exact Sciences, Tel Aviv University, Tel Aviv 69978, Israel

Received 25 July 1999

Abstract

We are going to survey recent developments and achievements in shape-preserving approximation by polynomials. We wish to approximate a function f defined on a finite interval, say [-1, 1], while preserving certain intrinsic "shape" properties. To be specific we demand that the approximation process preserves properties of f, like its sign in all or part of the interval, its monotonicity, convexity, etc. We will refer to these properties as the shape of the function. © 2000 Elsevier Science B.V. All rights reserved.

Keywords: Degree of approximation; Jackson estimates; Polynomial approximation; Shape preserving approximation

1. Introduction

We are going to discuss the degree of constrained approximation of a function f in either the uniform norm or in the $\mathbb{L}_p[-1,1]$, norm $0 , and we will use the notation <math>\mathbb{L}_{\infty}[-1,1]$ for $\mathbb{C}[-1,1]$, whenever we state a result which is valid both for $\mathbb{C}[-1,1]$ as well as for $\mathbb{L}_p[-1,1]$, for a proper range of p's. The degree of approximation will be measured by the appropriate (quasi-)norm which we denote by $\|\cdot\|_p$. The approximation will be carried out by polynomials $p_n \in \Pi_n$, the space of polynomials of degree not exceeding n, which have the same shape in which we are interested, as f, namely, have the same sign as f does in various parts of [-1,1], or change their monotonicity or convexity exactly where f does in [-1,1]. Most of the proofs of the statements in this survey and especially those of the affirmative results, are technically involved and will be omitted. All we are going to say about the technique of proof is that we usually first approximate f well by splines or just continuous piecewise polynomials with the same shape as f, and then we replace the polynomial pieces by polynomials of the same shape. Thus, while this survey deals only with polynomial approximation, there are similar affirmative results for continuous piecewise polynomials

E-mail address: leviatan@math.tau.ac.il (D. Leviatan).

and in many cases for splines. We will sometimes indicate a proof or construct a counterexample which we consider illustrative while not too involved.

Interest in the subject began in the 1960s with work on monotone approximation by Shisha and by Lorentz and Zeller. It gained momentum in the 1970s and early 1980s with the work on monotone approximation of DeVore, and the work on comonotone approximation of Shvedov, of Newman and of Beatson and Leviatan. The last 15 years have seen extensive research and many new results, the most advanced of which are being summarized here. We are not going to give an elaborate historical account and we direct the interested reader to an earlier survey by the author [22] and to the references therein.

The theory we are going to develop is much richer when dealing with the uniform norm and much less is known for the L_p -(quasi-)norm, when 0 . We are not going to state specifically too many open problems; however, the reader will only have to compare the results for the former and for the latter norms for the questions to be apparent. Also comparison between the results in the various sections will show where work is still to be done.

To be specific, let $s \ge 0$ and let \mathbb{Y}_s be the set of all collections $Y_s := \{y_i\}_{i=1}^s$ of points, so that $y_{s+1} := -1 < y_s < \cdots < y_1 < 1 = : y_0$, where for s = 0, $Y_0 = \emptyset$. For $Y_s \in \mathbb{Y}_s$ we set

$$\Pi(x,Y_s):=\prod_{i=1}^s(x-y_i),$$

where the empty product =1.

We let $\Delta^0(Y_s)$ be the set of functions f which change their sign (in the weak sense) exactly at the points $y_i \in Y_s$, and which are nonnegative in $(y_1, 1)$ (if s = 0, then this means that $f \ge 0$ in [-1, 1], and we will write $f \in \Delta^0$, suppressing the $Y_0 = \emptyset$). Note that our assumption is equivalent to

$$f(x)\Pi(x,Y_s) \ge 0, \quad -1 \le x \le 1.$$

For $f \in \Delta^0(Y_s) \cap \mathbb{L}_p[-1, 1]$, we denote by

$$E_n^{(0)}(f, Y_s)_p := \inf_{p_n \in \Pi_n \cap A^0(Y_s)} \|f - p_n\|_p$$

the degree of copositive approximation of f by algebraic polynomials. If $Y_0 = \emptyset$, then we write $E_n^{(0)}(f)_p := E_n^{(0)}(f, \emptyset)_p$, which is usually referred to as the degree of positive approximation.

Also, we let $\Delta^1(Y_s)$, be the set of functions f which change monotonicity at the points $y_i \in Y_s$, and which are nondecreasing in $(y_1, 1)$, that is, f is nondecreasing in the intervals (y_{2j+1}, y_{2j}) and it is nonincreasing in (y_{2j}, y_{2j-1}) . In particular, if s = 0, then f is nondecreasing in [-1, 1], and we will write $f \in \Delta^1$. Moreover, if f is differentiable in (-1, 1), then

$$f \in \Delta^1(Y_s)$$
 iff $f'(x)\Pi(x, Y_s) \ge 0$, $-1 < x < 1$.

Now for $f \in \Delta^1(Y_s) \cap \mathbb{L}_p[-1, 1]$, we denote by

$$E_n^{(1)}(f, Y_s)_p := \inf_{p_n \in \Pi_n \cap A^1(Y_s)} \|f - p_n\|_p,$$

the degree of comonotone polynomial approximation. Again if $Y_0 = \emptyset$, then we write $E_n^{(1)}(f)_p := E_n^{(1)}(f, \emptyset)_p$, which is usually referred to as the degree of monotone approximation.

Finally, we let $\Delta^2(Y_s)$ be the set of functions f which change convexity at the points $y_i \in Y_s$, and which are convex in $(y_1, 1)$ (again Δ^2 if $Y_0 = \emptyset$), and for $f \in \Delta^2(Y_s) \cap \mathbb{L}_p[-1, 1]$, we denote by

$$E_n^{(2)}(f, Y_s)_p := \inf_{p_n \in \Pi_n \cap \varDelta^2(Y_s)} \|f - p_n\|_p,$$

the degree of coconvex approximation. Once again if $Y_0 = \emptyset$, then we write $E_n^{(2)}(f)_p := E_n^{(2)}(f, \emptyset)_p$, which is usually referred to as the degree of convex approximation.

Remark. While it is obvious that f may belong to $\Delta^0(Y_{s_0}^0) \cap \Delta^1(Y_{s_1}^1)$, say, where $Y_{s_0}^0 \neq Y_{s_1}^1$ and $s_0 \neq s_1$, it should be emphasized that f may belong $\Delta^{\nu}(Y_s)$, where $0 \leq \nu \leq 2$ is fixed, for many different sets Y_s and different s's, since we assumed weak changes in the sign, the monotonicity or the convexity. Thus, we find it useful for such a function to introduce the best degree of constrained approximation, namely, for the appropriate $0 \leq \nu \leq 2$ and a fixed s, we denote

$$e_n^{(v,s)}(f)_p := \inf E_n^{(v)}(f, Y_s)_p, \tag{1.1}$$

where the infimum is taken over all admissible sets Y_s of s points in which f changes its sign, monotonicity or convexity according to whether v = 0, 1 or 2, respectively. In this survey we make use of this notation only in negative results in comonotone approximation.

For comparison purposes we need the degree of unconstrained approximation, so for $f \in \mathbb{L}_p[-1, 1]$, let us write

$$E_n(f)_p := \inf_{p_n \in \Pi_n} \|f - p_n\|_p.$$

Suppose $f \in \mathbb{C}[-1, 1]$, $f \ge 0$. Then for $n \ge 0$, $P_n \in \Pi_n$ exists such that

$$||f-P_n||_{\infty}=E_n(f)_{\infty}.$$

Thus

$$P_n(x) - f(x) \ge -E_n(f)_{\infty},$$

so that

$$Q_n(x) := P_n(x) + E_n(f)_{\infty} \ge f(x) \ge 0.$$

Hence, Q_n is nonnegative and we have

$$||f-Q_n||_{\infty} \leq 2E_n(f)_{\infty},$$

which yields

$$E_n^{(0)}(f)_{\infty} \leqslant 2E_n(f)_{\infty}, \quad n \ge 0.$$

$$(1.2)$$

Thus, there is nothing to investigate in this case. However, the situation is completely different when asking for either pointwise estimates for the approximation of nonnegative functions by nonnegative polynomials, or for L_p estimates of positive polynomial approximation. We will discuss recent results on these subjects and in copositive polynomial approximation in Section 2.

Now suppose $f \in \mathbb{C}^1[-1,1]$ is monotone nondecreasing. Then of course $f' \ge 0$. By the above, for $n \ge 1$, a nonnegative $q_{n-1} \in \Pi_{n-1}$ exists such that

$$||f'-q_{n-1}||_{\infty} \leq 2E_{n-1}(f')_{\infty}.$$

Put $Q_n(x) := \int_0^x q_{n-1} + f(0)$. Then Q_n is nondecreasing and

$$\|f-Q_n\|_{\infty} \leqslant 2E_{n-1}(f')_{\infty}.$$

Hence,

$$E_n^{(1)}(f)_{\infty} \leq 2E_{n-1}(f')_{\infty}, \quad n \ge 1.$$
(1.3)

Similarly, if $f \in \mathbb{C}^2[-1,1]$ is convex, i.e., $f'' \ge 0$, then we get

$$E_n^{(2)}(f)_{\infty} \leqslant E_{n-2}(f'')_{\infty}, \quad n \ge 2.$$

$$(1.4)$$

Recall that if $f \in W_p^1[-1,1]$, $1 \le p \le \infty$, the Sobolev space of locally absolutely continuous functions on [-1,1], such that $f' \in \mathbb{L}_p[-1,1]$, then in unconstrained approximation by polynomials we have

$$E_n(f)_p \leqslant \frac{C}{n} E_{n-1}(f')_p, \quad n \ge 1,$$

$$(1.5)$$

where C = C(p) is an absolute constant. It should be emphasized that (1.5) is not valid for 0 . $Evidently, (1.5) in turn implies for <math>f \in W_p^r[-1, 1]$, $1 \le p \le \infty$,

$$E_n(f)_p \leqslant \frac{C}{n^r} E_{n-r}(f^{(r)})_p, \quad n \ge r,$$

where C = C(p) is an absolute constant.

Thus, in (1.3) we have lost an order of n and in (1.4) we have a loss of order of n^2 . We will try to retrieve some of these losses in the estimates we present in this paper. These will be Jackson type estimates which are analogous to those in unconstrained approximation, namely, on the right-hand side of (1.3) and (1.4), we will have various moduli of smoothness of different kinds which we are going to define below. However, we will also show that constrained approximation restricts the validity of these estimates. At this stage we just point out that Shevchuk separately and with the author [33,23] have proved,

Theorem 1.1. There exists a constant C > 0 such that for any $n \ge 1$, a function $f = f_n \in \mathbb{C}^1[-1,1] \cap \Delta^1$ exists, such that

$$E_n^{(1)}(f)_\infty \ge C E_{n-1}(f')_\infty > 0,$$

and for any $n \ge 2$, a function $f \in \mathbb{C}^2[-1,1] \cap \Delta^2$ exists, such that

$$E_n^{(2)}(f)_{\infty} \geq C E_{n-2}(f'')_{\infty} > 0.$$

Hence, it is clear that (1.3) and (1.4) by themselves, cannot be improved.

In fact, if 0 , the situation is even more pronounced. Indeed, Kopotun [18] proved that

Theorem 1.2. Let 0 and <math>v = 1, 2. Then for each $n \ge v$, and every constant A > 0, there exists an $f = f_{p,n,A} \in \mathbb{C}^{\infty}[-1,1] \cap \Delta^{v}$ for which

$$E_n^{(\nu)}(f)_p \ge A E_{n-\nu}(f^{(\nu)})_p$$

We will discuss monotone and comonotone approximation in Section 3 and convex and coconvex approximation in Section 4.

In order to use consistent notation we will use $r \ge 0$ for the number of derivatives that the function possesses, and $k \ge 0$ for the order of the moduli of smoothness. In addition to the ordinary moduli of smoothness $\omega_k(f,t)_p$, $0 (where <math>\omega_0(f,t)_p := ||f||_p$), defined for $f \in \mathbb{L}_p[-1,1]$, we recall the Ditzian–Totik moduli of smoothness which are defined for such an f, as follows. With $\varphi(x) := \sqrt{1 - x^2}$, we let

$$\Delta_{h\varphi}^{k}f(x) := \begin{cases} \sum_{i=0}^{k} (-1)^{i} \binom{k}{i} f(x+(i-\frac{k}{2})h\varphi(x)), & x \pm \frac{k}{2}h\varphi(x) \in [-1,1], \\ 0, & \text{otherwise,} \end{cases}$$

and we set $\Delta_{ho}^0 f := f$. Then we denote

$$\omega_k^{\varphi}(f,t)_p := \sup_{0 < h \leqslant t} \| \varDelta_{h\varphi}^k f \|_p, \qquad \omega^{\varphi}(f,t)_p := \omega_1^{\varphi}(f,t)_p.$$

When $p = \infty$ we will also use a modification of these moduli where we will take into account not only the position of x in the interval when setting $\Delta_{h\phi}^k f$, but also how far the endpoints of the interval $[x - \frac{1}{2}kh\phi(x), x + \frac{1}{2}kh\phi(x)]$ are from the endpoints of [-1, 1]. To this end we set

$$\varphi_{\delta}(x) := \sqrt{\left(1 - x - \frac{\delta}{2}\varphi(x)\right)\left(1 + x - \frac{\delta}{2}\varphi(x)\right)}, \qquad x \pm \frac{\delta}{2}\varphi(x) \in [-1, 1],$$

and we restrict $f \in \mathbb{C}[-1,1]$, to the space \mathbb{C}_{φ}^r , of functions $f \in \mathbb{C}^r(-1,1)$, such that $\lim_{x \to \pm 1} \varphi^r(x) f^{(r)}(x) = 0$. We denote

$$\omega_{k,r}^{\varphi}(f^{(r)},t) := \sup_{0 \leqslant h \leqslant t} \sup_{x} |\varphi_{kh}^{r}(x) \varDelta_{h\varphi(x)}^{k} f^{(r)}(x)|, \quad t \ge 0$$

where the inner supremum is taken over all x so that

$$x \pm \frac{k}{2}h\varphi(x) \in (-1,1).$$

Note that for k = 0, we have

$$\omega_{0,r}^{\varphi}(f^{(r)},t) = \|\varphi^r f^{(r)}\|_{\infty},\tag{1.6}$$

and that for r = 0,

$$\omega_{k,0}^{\varphi}(f^{(0)},t) := \omega_k^{\varphi}(f,t).$$

The above restriction guarantees that for $k \ge 1$, $\omega_{k,r}^{\varphi}(f^{(r)},t) \to 0$, as $t \to 0$. Also, it can be shown that if $f \in \mathbb{C}_{\varphi}^{r}$ and $0 \le m < r$, then

$$\omega_{k+r-m,m}^{\varphi}(f^{(m)},t) \leq C(k,r)t^{r-m}\omega_{k,r}^{\varphi}(f^{(r)},t), \quad t \ge 0$$

and conversely if $f \in \mathbb{C}[-1,1]$, $m < \alpha < k$ and $\omega_k^{\varphi}(f,t) \leq t^{\alpha}$, then $f \in \mathbb{C}_{\varphi}^m$ and

$$\omega_{k-m,m}^{\varphi}(f^{(m)},t) \leq C(\alpha,k)t^{\alpha-m}, \quad t \geq 0.$$

Finally, if $f \in \mathbb{C}_{\varphi}^{m}$ and $\omega_{r-m,m}^{\varphi}(f^{(m)},t) \leq t^{r-m}$, then

$$\|\varphi^r f^{(r)}\|_{\infty} \leqslant C(r).$$

We will denote the collection of functions satisfying the last inequality by \mathbb{B}^r , and the converse is also valid, namely, if $f \in \mathbb{B}^r$ and $0 \le m < r$, then $f \in \mathbb{C}^m_{\omega}$, and

 $\omega_{r-m,m}^{\varphi}(f^{(m)},t) \leq C(r)t^{r-m} \|\varphi^r f^{(r)}\|_{\infty}, \quad t \geq 0.$

2. Positive and copositive approximation

Pointwise estimates in the approximation of a nonnegative $f \in \mathbb{C}^r[-1,1] \cap \Delta^0$, there are of two types. The Timan–Brudnyi-type estimates of the form

$$|f(x) - p_n(x)| \leq C(r,k)\rho_n^r(x)\omega_k(f^{(r)},\rho_n(x))_{\infty}, \quad 0 \leq x \leq 1, \ n \geq N,$$
(2.1)

where $\rho_n(x) := 1/n^2 + (1/n)\varphi(x)$. Here C(r,k) is a constant which depends only on r and k, and which is independent of f and n, and the Telyakovskii–Gopengauz or interpolatory type estimates,

$$|f(x) - p_n(x)| \leq C(r,k)\delta_n^r(x)\omega_k(f^{(r)},\delta_n(x)), \quad 0 \leq x \leq 1, \ n \geq N,$$
(2.2)

where $\delta_n(x) := (1/n)\varphi(x)$.

Dzyubenko [5] has shown that estimates of the form (2.1) are valid for positive approximation for all $n \ge N := r + k - 1$. Namely, for each $n \ge N := r + k - 1$, there exists a polynomial $p_n \in \Pi_n \cap \Delta^0$, for which (2.1) holds. In contrast, in a recent paper, Gonska et al. [9] have shown that (2.2) is valid only when either r = 0 and k = 1, 2, or if $k \le r$. Specifically they proved the following two complementing results, namely,

Theorem 2.1. Let either r = 0 and k = 1, 2, or $1 \le k \le r$. If $f \in \mathbb{C}^r[-1, 1] \cap \Delta^0$, then for each $n \ge N := 2[(r + k + 1)/2]$, there is a polynomial $p_n \in \Pi_n \cap \Delta^0$, such that

$$|f(x) - p_n(x)| \leq C(r)\delta_n^r(x)\omega_k(f^{(r)}, \delta_n(x)), \quad 0 \leq x \leq 1.$$

$$(2.3)$$

(Note that the case $r + k \leq 2$ is due to DeVore and Yu [4]). And

Theorem 2.2. Let either r=0 and k > 2, or $k > r \ge 1$. Then for each $n \ge 1$ and any constant A > 0, a function $f = f_{k,r,n,A} \in \mathbb{C}^r[-1,1] \cap \Delta^0$ exists, such that for any polynomial $p_n \in \Pi_n \cap \Delta^0$, there is a point $x \in [0,1]$, for which

$$|f(x) - p_n(x)| > A \frac{(1-x)^{r/2}}{n^r} \omega_k \left(f^{(r)}, \frac{\sqrt{1-x}}{n} \right),$$
(2.4)

holds.

As was alluded to in the introduction the L_p -norm estimates for 0 , do not behave like $the case of the sup-norm. Denote by <math>W_p^r[-1, 1]$, 0 , the Sobolev space of functions <math>f such that $f^{(r-1)}$ is locally absolutely continuous in (-1, 1) and $f^{(r)} \in L_p[-1, 1]$. If $f \in W_p^1[-1, 1] \cap \Delta^0$, $1 \le p < \alpha$, then we come close to (1.2), with an estimate due to Stojanova (see [9]).

$$E_n^{(0)}(f)_p \leqslant \frac{C}{n} E_{n-1}(f')_p \leqslant \frac{C(k)}{n} \omega_k^{\varphi} \left(f', \frac{1}{n}\right)_p, \quad n \ge 1.$$

The constant C(k) depends only on k and on p, we are going to suppress indicating the dependence on the latter.

However, if we merely assume that $f \in \Delta^0 \cap \mathbb{L}_p[-1,1]$, then Hu et al. [10], proved that for 0 , there is a constant C such that

$$E_n^{(0)}(f)_p \leqslant C\omega\left(f,\frac{1}{n}\right)_p,$$

but on the other hand for each A > 0, every $n \ge 1$, and any $0 , there exists an <math>f := f_{A,n,p} \in \Delta^0 \cap \mathbb{L}_p[-1,1]$, such that

$$E_n^{(0)}(f)_p \ge A\omega_2(f,1)_p$$

Stojanova also proved that for $f \in \Delta^0 \cap \mathbb{L}_p[-1,1]$, we always have the estimate

$$E_n^{(0)}(f)_p \leqslant C(k)\tau_k\left(f,\frac{1}{n}\right)_p, \quad n \ge 1,$$

where $\tau_k(f, \cdot)_p$ are the averaged moduli of smoothness which were introduced by Sendov (see [10] for details and references).

We turn now to copositive approximation. Here we still have variations in the estimates for $p = \infty$ and for $1 \le p < \infty$, but in no case the behavior is as in unconstrained approximation. Recall that in this case we deal with a function which changes its sign at $Y_s \in \mathbb{V}_s$.

For the sup-norm the estimates are due to Hu and Yu [13], and Kopotun [17], and negative results are due to Zhou [37]. We summarize their results in

Theorem 2.3. Let $f \in \mathbb{C}[-1,1] \cap \Delta^0(Y_s)$. Then there exists a constant $C = C(Y_s)$ such that

$$E_n^{(0)}(f,Y_s)_{\infty} \leqslant C(Y_s)\omega_3^{\varphi}\left(f,\frac{1}{n}\right)_{\infty}, \quad n \ge 2,$$
(2.5)

and for each $n \ge 2$, there exists a polynomial $p_n \in \Pi_n \cap \Delta^0(Y_s)$, such that

 $|f(x)-p_n(x)| \leq C(Y_s)\omega_3(f,\rho_n(x))_{\infty}.$

Furthermore, if $f \in \mathbb{C}^1[-1,1] \cap \Delta^0(Y_s)$, then

$$E_n^{(0)}(f,Y_s)_{\infty} \leqslant \frac{C(k,Y_s)}{n} \omega_k \left(f',\frac{1}{n}\right)_{\infty}, \quad n \ge k$$

and for each $n \ge 2$, there exists a polynomial $p_n \in \Pi_n \cap \Delta^0(Y_s)$, such that

$$|f(x) - p_n(x)| \leq \frac{C(k, Y_s)}{n} \omega_k(f', \rho_n(x))_{\infty}, \quad n \geq k$$

Conversely, there is an $f \in \mathbb{C}^1[-1,1] \cap \Delta^0(\{0\})$, for which

$$\limsup_{n\to\infty}\frac{E_n^{(0)}(f,\{0\})_\infty}{\omega_4(f,1/n)_\infty}=\infty.$$

Since in Section 3 we are going to discuss the dependence of the constants on the collection Y_s , especially in contrast to dependence on s alone, we mention that if in (2.5), we replace the third modulus of smoothness of f by its modulus of continuity, then Leviatan has proved that the

inequality holds with a constant C = C(s). In view of (2.3) one may ask whether it is possible to give some interpolatory estimates for copositive approximation. This question is completely open. If $1 \le p < \infty$, then it was proved by Hu et al. [10] that

Theorem 2.4. If $f \in \mathbb{L}_p[-1,1] \cap \Delta^0(Y_s)$, then

$$E_n^{(0)}(f,Y_s)_p \leq C(Y_s)\omega^{\varphi}\left(f,\frac{1}{n}\right)_p, \quad n \geq 1,$$

and if $f \in W_p^1[-1,1] \cap \Delta^0(Y_s)$, then

$$E_n^{(0)}(f,Y_s)_p \leqslant \frac{C(Y_s)}{n} \omega^{\varphi} \left(f',\frac{1}{n}\right)_p, \quad n \ge 1.$$

Furthermore if $f \in W_p^2[-1,1] \cap \Delta^0(Y_s)$, then

$$E_n^{(0)}(f,Y_s)_p \leqslant \frac{C(k,Y_s)}{n^2} \omega_k^{\varphi} \left(f'', \frac{1}{n} \right)_p, \quad n \ge k+1.$$

Conversely, for every $n \ge 1$ and 0 , and for any constants <math>A > 0 and $0 < \varepsilon \le 1$, there is an $f = f_{n, p, \varepsilon, A} \in \mathbb{C}^{\infty}[-1, 1]$, satisfying $xf(x) \ge 0$, $-1 \le x \le 1$, such that for each $p_n \in \Pi_n$, for which $p_n(0) \ge 0$, we have

$$||f - p_n||_{L_p[0,\varepsilon]} > A\omega_2(f,1)_p$$

Also, there is a strictly increasing $f = f_{n, p, \varepsilon, A} \in \mathbb{C}^{\infty}[-1, 1]$, satisfying f(0) = 0 such that for each $p_n \in \Pi_n$, for which $p_n(0) = 0$, and $p_n(x) \ge 0$, $0 \le x \le \varepsilon$, we have

$$||f - p_n||_{\mathbb{L}_p[0,\varepsilon]} > A\omega_3(f',1)_p.$$

In [10] there are some estimates involving the "tau" modulus but we will not detail them here. It should also be pointed out that in [10] the authors introduce an interesting new concept of intertwining approximation which is related to both copositive approximation and one-sided approximation. We will not discuss this concept here and the interested reader should consult that paper.

3. Monotone and comonotone approximation

If $f \in \mathbb{L}_p[-1,1]$, 0 , is nondecreasing, then the following is known (see [3,34]).

Theorem 3.1. Let $f \in \mathbb{L}_p[-1,1] \cap \Delta^1$. Then for each $n \ge 1$, we have

$$E^{(1)}(f)_p \leqslant C\omega_2^{\varphi}\left(f, \frac{1}{n}\right)_p,\tag{3.1}$$

where C = C(p). (The dependence on p is crucial only when $p \to 0$.)

Conversely, if $k \ge 3$, then for any A > 0 and $n \ge 1$, there exists a function $f = f_{p,n,A} \in L_p[-1,1] \cap \Delta^1$, such that

 $E^{(1)}(f)_p \ge A\omega_k(f,1)_p > 0.$

If $1 \le p \le \infty$, then (3.1) readily implies that for $f \in W_p^1[-1,1] \cap \Delta^1$, we have

$$E^{(1)}(f)_p \leqslant \frac{C}{n} \omega^{\varphi} \left(f', \frac{1}{n} \right)_p.$$

Thus, one may hope that for smooth functions it would be possible to obtain estimates involving the moduli of smoothness of the appropriate derivatives and this way have better rates of monotone approximation. This is true for the sup-norm but it is not so for any of the L_p -norms. Specifically, Shevchuk [30,31] has proved that

Theorem 3.2. If $f \in \mathbb{C}^1[-1,1] \cap \Delta^1$, then for for each $k \ge 1$, there is a constant C = C(k) such that

$$E^{(1)}(f)_{\infty} \leqslant \frac{C}{n} \omega_k \left(f', \frac{1}{n} \right)_{\infty}.$$
(3.2)

However, Kopotun [18] has shown that

Theorem 3.3. Let $0 , <math>k \ge 0$, and $\max\{1, 3 - k\} \le v < 1 + 1/p$. Then for each $n \ge 1$ and $\varepsilon > 0$, and every constant A > 0, there exists an $f = f_{p,k,v,n,\varepsilon,A} \in \mathbb{C}^{\infty}[-1,1] \cap \Delta^1$, such that for any $p_n \in \Pi_n$ for which $p'_n(-1) \ge 0$, it follows that

$$||f - p_n||_{\mathbb{L}_p[-1,-1+\varepsilon]} > A\omega_k(f^{(v)},1)_p.$$

Note that in particular if $k \ge 2$, then one cannot replace in (3.2), the sup-norm by any of the L_v -norms, 0 .

One may ask whether we may have (3.1), if we relax the requirement on the constant by allowing such a constant to depend on the function f (but not on n). Wu and Zhou [35] have shown that this is impossible with $k = 4 + \lfloor 1/p \rfloor$. On the other hand, Shevchuk and Leviatan [26] have recently closed the gap when f is monotone and continuous, proving

Theorem 3.4. If $f \in \mathbb{C}[-1,1] \cap \Delta^1$, then there exists a constant C = C(f) such that

$$E^{(1)}(f)_{\infty} \leq C \omega_3^{\varphi} \left(f, \frac{1}{n} \right)_{\infty}, \quad n \geq 2.$$

In fact, when dealing with monotone continuous functions a lot more is known. For instance if $f \in \mathbb{C}^r[-1,1] \cap \Delta^1$, then for each $n \ge r + k - 1$, there exists a polynomial $p_n \in \Delta^1$, for which the pointwise estimates (2.1) are valid. This has recently been shown by Dzyubenko [6]. However, the interpolatory estimates (2.2) are valid in very few cases, namely, only when $r + k \le 2$, the affirmative result being due to DeVore and Yu [4]. We now know (see [9]) that

Theorem 3.5. If r > 2, then for each n there is a function $f = f_{r,n} \in W_{\infty}^{r}[-1,1] \cap \Delta^{1}$, such that for every polynomial $p_{n} \in \Pi_{n} \cap \Delta^{1}$, either

$$\limsup_{x \to -1} \frac{|f(x) - p_n(x)|}{\varphi^r(x)} = \infty,$$
(3.3)

or

$$\limsup_{x\to 1}\frac{|f(x)-p_n(x)|}{\varphi^r(x)}=\infty.$$

This readily implies

Corollary 3.6. Let r + k > 2. Then for each *n* there exists a function $f = f_{r,k,n} \in \mathbb{C}^r[-1,1] \cap \Delta^1$, such that for every polynomial $p_n \in \Pi_n \cap \Delta^1$, either

$$\limsup_{x \to -1} \frac{|f(x) - p_n(x)|}{\varphi^r(x)\omega_k(f^{(r)}, \varphi(x))} = \infty$$

or

$$\limsup_{x\to 1}\frac{|f(x)-p_n(x)|}{\varphi^r(x)\omega_k(f^{(r)},\varphi(x))}=\infty.$$

Here we have an opportunity to show the nature of the function yielding Theorem 3.5.

Proof of Theorem 3.5. Given *n*, set $b = n^{-2}$ and let

$$f(x) := \begin{cases} (b^r - (b - x - 1)^r)/r!, & -1 \le x \le -1 + b, \\ b^r/r!, & -1 + b < x \le 1. \end{cases}$$

Then $f \in W_{\infty}^{r}[-1,1] \cap \Delta^{1}$. Suppose that there is a nondecreasing polynomial p_{n} for which (3.3) fails. Then for that polynomial and some constant A, we have

$$|f(x) - p_n(x)| \leq A \varphi^r(x) \leq A (1+x)^{3/2}, \quad -1 \leq x \leq -1+b,$$

where the right-hand inequality follows since r > 2. Hence $p_n(-1) = f(-1) = 0$ and $p'_n(-1) = f'(-1) = b^{r-1}/(r-1)!$. Since $p_n \in \Delta^1$, we have $||p_n|| = p_n(1)$, so that applying Markov's inequality we conclude that

$$\frac{b^{r-1}}{(r-1)!} = p'_n(-1) \leqslant n^2 ||p_n|| = n^2 p_n(1)$$

or

$$p_n(1) \ge \frac{b^{r-1}}{(r-1)!n^2} = \frac{b^r}{(r-1)!}.$$

On the other hand,

$$f(1) = \frac{b^r}{r!} < \frac{b^r}{(r-1)!}.$$

Thus $f(1) \neq p_n(1)$, and (1.12) is satisfied. \Box

We still can salvage something if we are willing to settle for interpolation at only one of the endpoints while approximating well throughout the interval. Namely, it is proved in [9] that

82

Theorem 3.7. If $k \leq \max\{r, 2\}$, and $f \in \mathbb{C}^r[-1, 1] \cap \Delta^1$, then for each $n \geq N := k + r - 1$, there is a polynomial $p_n \in \Pi_n \cap \Delta^1$, such that

$$|f(x) - p_n(x)| \leq C(r)\rho_n^r(x)\omega_k(f^{(r)}, \rho_n(x)), \quad -1 \leq x \leq 1,$$

and

$$|f(x) - p_n(x)| \leq C(r) \frac{(1-x)^{r/2}}{n^r} \omega_k\left(f^{(r)}, \frac{\sqrt{1-x}}{n}\right), \quad -1 \leq x \leq 1.$$

For all other pairs (k, r), Theorem 3.7 does not hold. In fact we have

Theorem 3.8. If $k > \max\{r, 2\}$, then for each $n \ge 1$ and any constant A > 0, a function $f = f_{r,k,n,A} \in C^r[-1,1] \cap \Delta^1$ exists, such that for any polynomial $p_n \in \Pi_n \cap \Delta^1$, there is a point $x \in [-1,1]$ for which (2.4) holds.

We conclude the part on monotone approximation with a result on simultaneous pointwise estimates due to Kopotun [14].

Theorem 3.9. If $f \in C^1[-1,1] \cap \Delta^1$, then for every $n \ge 1$, a polynomial $p_n \in \Pi_n \cap \Delta^1$ exists, such that

$$|f^{(i)}(x) - p_n^{(i)}(x)| \leq C\omega_{2-i}(f^{(i)}, \rho_n(x)), \quad i = 0, 1, -1 \leq x \leq 1.$$

We now proceed to investigate the degree of comonotone approximation of a function $f \in \mathbb{L}_p[-1, 1]$, $0 , which changes its monotonicity at <math>Y_s \in \mathbb{V}_s$. Thus for the remainder of this section $s \ge 1$ unless we specifically say otherwise.

Again we have only some results for $p < \infty$, and most of the recent developments are for estimates in the max-norm. Denote

$$d(Y_s) := \min_{0 \le i \le s} (y_i - y_{i+1}).$$
(3.4)

Then the following general estimates have been obtained by Kopotun and Leviatan [19].

Theorem 3.10. Let $f \in \mathbb{L}_p[-1,1] \cap \Delta^1(Y_s)$, 0 . Then there exists a constant <math>C = C(s) such that for each $n \ge C/d(Y_s)$,

$$E_n^{(1)}(f,Y_s)_p \leq C\omega_2^{\varphi}\left(f,\frac{1}{n}\right)_p.$$

On the other hand, Zhou [38] has shown that for every $0 and each <math>s \ge 1$, there is a collection Y_s and a function $f \in \mathbb{L}_p[-1,1] \cap \Delta^1(Y_s)$, for which

$$\limsup_{n\to\infty}\frac{E_n^{(1,s)}(f)_p}{\omega_k(f,1/n)_p}=\infty,$$

with $k = 3 + \lfloor 1/p \rfloor$. Thus taking $p = \infty$, one sees that Theorem 3.10 is not valid with any $k \ge 3$, even with C = C(f) and N = N(f).

If $f \in \mathbb{C}[-1,1] \cap \Delta^1(Y_s)$, then we can say much more. We begin with the following results of Dzyubenko et al. [7] (see also [36]).

Theorem 3.11. Let $f \in \mathbb{C}^r[-1,1] \cap \Delta^1(Y_s)$. Then the estimates

$$E_n^{(1)}(f, Y_s)_{\infty} \leqslant \frac{C}{n^r} \omega_k \left(f^{(r)}, \frac{1}{n} \right)_{\infty}, \quad n \ge N,$$
(3.5)

is valid with C = C(k, r, s) and N = N(k, r, s), only when either k = 1, or r > s, or in the particular case k = 2 and r = s, moreover, in these cases one can always take N = k + r - 1. If k = 2 and $0 \le r < s$, or k = 3 and $1 \le r \le s$, or if k > 3 and $2 \le r \le s$, then the estimates hold either with $C = C(k, r, Y_s)$ and N = k + r, or with C = C(k, r, s) and $N = N(k, r, Y_s)$, and they fail to hold with C = C(k, r, s) and N = N(k, r, s).

On the other hand, if either r = 0 or r = 1, then for each $s \ge 1$, there is a collection $Y_s \in \mathbb{Y}_s$ and a function $f \in \mathbb{C}^r[-1, 1] \cap \Delta^1(Y_s)$, for which

$$\limsup_{n\to\infty}\frac{n^r E_n^{(1)}(f,Y_s)_\infty}{\omega_{3+r}(f^{(r)},1/n)}=\infty,$$

i.e., (3.5) is not valid even with constants which depend on f.

We found it easier to remember, especially when later on we compare with other types of estimates, to illustrate the above in an array in which + in the (k, r) entry means that both constants C and N depend only on k, r and s; \oplus means that one of the two constants depends on k and r and on the location of the points of change of monotonicity, namely on Y_s ; while - asserts that (3.5) is not valid at all (see Fig. 1).

In particular, the first column of the array implies that if $f \in W_{\infty}^{r}$, then

$$E_n^{(1)}(f, Y_s)_{\infty} \leqslant C(r, s) \frac{\|f^{(r)}\|_{\infty}}{n^r}, \quad n \ge r-1.$$
(3.6)

Pointwise estimates of the type (2.1), for comonotone approximation present new phenomena. If s = 1, then when either $r \ge 2$; or in three special cases, k = 1 and r = 0, 1; and k = 2 and r = 1; we have a polynomial $p_n \in \Pi_n \cap \Delta^1$ satisfying

$$|f(x) - p_n(x)| \leq C(r)\rho_n^r(x)\omega_k(f^{(r)}, \rho_n(x))_{\infty}, \quad 0 \leq x \leq 1, \quad n \geq k+r-1.$$

Two other pairs k = 2 and r = 0, and k = 3 and r = 1, yield (2.1) with $C = C(Y_1) = C(y_1)$, while for the remaining pairs, namely, r = 0 and $k \ge 3$, and r = 1 and $k \ge 4$, we have no estimate of the type (2.1). Thus the array is exactly the one we had in Fig. 1, for s = 1. If on the other hand, s > 1, then the array looks entirely different. To be specific, (2.1) holds with C = C(r, k, s), only for $n \ge N = N(r, k, Y_s)$ so that the array is as in Fig. 2.

Estimates involving the D–T moduli are similar to those of the ordinary moduli and yield the same array as Fig. 1. This raises the expectation of having an estimate analogous to (3.6) for functions in \mathbb{B}^r . However, this is not so except when f is monotone. Indeed, Leviatan and Shevchuk [27] have proved that



Theorem 3.12. Let $s \ge 0$ and assume that $f \in \mathbb{B}^r \cap \Delta^1(Y_s)$, $r \ge 1$. Then

$$E_n^{(1)}(f,Y_s)_{\infty} \leqslant C(r,Y_s) \frac{\|\varphi^r f^{(r)}\|_{\infty}}{n^r}, \quad n \ge r-1,$$

and

$$E_n^{(1)}(f,Y_s)_{\infty} \leqslant C(r,s) \frac{\|\varphi^r f^{(r)}\|_{\infty}}{n^r}, \quad n \ge N(r,Y_s).$$

Furthermore, if $f \in \mathbb{B}^r \cap \Delta^1(Y_s)$, with either s = 0 or r = 1, or r = 3 and s = 1, or r > 2s + 2, then

$$E_n^{(1)}(f, Y_s)_{\infty} \leqslant C(r) \frac{\|\varphi^r f^{(r)}\|_{\infty}}{n^r}, \quad n \ge r-1.$$
(3.7)

For all other cases (3.7) is not valid, that is, we have (see [24])

Theorem 3.13. Given $s \ge 1$. Let the constant A > 0 be arbitrary and let $2 \le r \le 2s + 2$, excluding the case r = 3 and s = 1. Then for any n, there exists a function $f = f_{r,s,n} \in \mathbb{B}^r$, which changes monotonicity s times in [-1, 1], for which

$$e_n^{(1,s)}(f)_{\infty} \geq A \|\varphi^r f^{(r)}\|_{\infty}$$

See (1.1) for the definition of $e_n^{(1,s)}(f)_{\infty}$.

It is in order to investigate this phenomenon that we introduced the modified moduli $\omega_{k,r}^{\varphi}$. In fact, we recall that in (1.6), we have noted that $\omega_{0,r}^{\varphi}(f^{(r)},t) = \|\varphi^r f^{(r)}\|_{\infty}$. Indeed Leviatan and Shevchuk [27] have obtained the following estimates for these moduli. (The case of monotone approximation, i.e., s = 0 had been treated earlier by Dzyubenko et al. [8].)

Theorem 3.14. Let $s \ge 0$ and assume that $f \in \mathbb{C}^r_{\omega} \cap \Delta^1(Y_s)$, with r > 2. Then

$$E_{n}^{(1)}(f, Y_{s})_{\infty} \leq \frac{C(k, r, Y_{s})}{n^{r}} \omega_{k, r}^{\varphi} \left(f^{(r)}, \frac{1}{n} \right), \quad n \geq k + r - 1,$$
(3.8)

and

$$E_n^{(1)}(f, Y_s)_{\infty} \leq \frac{C(k, r, s)}{n^r} \omega_{k, r}^{\varphi} \left(f^{(r)}, \frac{1}{n} \right), \quad n \geq N(k, r, Y_s).$$

$$(3.9)$$

Furthermore, if $f \in \mathbb{C}_{\varphi}^r \cap \Delta^1(Y_s)$, with r > 2s + 2, then

$$E_n^{(1)}(f, Y_s)_{\infty} \leq \frac{C(k, r, s)}{n^r} \omega_{k, r}^{\varphi} \left(f^{(r)}, \frac{1}{n} \right), \quad n \geq k + r - 1.$$
(3.10)

Remark. Obviously, when s = 0 there is no dependence on $Y_0 = \emptyset$ in (3.8) and (3.9) and hence the former is just (3.10). Also in this case, (3.10) is valid for $0 \le r + k \le 2$, as follows from (3.1).

To the contrary we have

Theorem 3.15. For $s \ge 1$, let $0 \le r \le 2s+2$, excluding the three cases $r+k \le 1$. Then for any constant A > 0 and every $n \ge 1$, there is function $f := f_{k,r,s,n,A} \in \mathbb{C}_{\varphi}^{r}$ which changes monotonicity s times in [-1,1], for which

$$e_n^{(1,s)}(f)_{\infty} > A\omega_{k,r}^{\varphi}(f^{(r)},1).$$

Finally, we have some cases where (3.8) is valid with a constant C = C(f), others when even so, it is not valid, and a few which are still open. We summarize what we know due to Leviatan and Shevchuk [25,28].

Theorem 3.16. If $f \in \Delta^1$, then there exist constants C = C(f) and N = N(f), and an absolute constant *c*, such that for all $0 \le k + r \le 3$,

$$E_n^{(1)}(f)_{\infty} \leq C\omega_{k,r}^{\varphi}\left(f,\frac{1}{n}\right), \quad n \geq 2,$$
(3.11)

and

$$E_n^{(1)}(f)_{\infty} \leq c \omega_{k,r}^{\varphi}\left(f,\frac{1}{n}\right), \quad n \geq N$$

Theorem 3.17. Let $s \ge 0$. Then there is a collection $Y_s \in \mathbb{Y}_s$ and a function $f \in \mathbb{C}^2_{\varphi} \cap \Delta^1(Y_s)$, satisfying

$$\limsup_{n\to\infty}\frac{n^2E_n^{(1)}(f,Y_s)_\infty}{\omega_{3,2}^\varphi(f'',1/n)}=\infty.$$

m	:	:	:	:	:	:	
T	•	•	•	•	•	•	•
4	+	+	+	+	+	+	•••
3	+	+	+	+	+	+	• • •
2	+	θ	?		—		•••
1	+	+	θ	?		-	•••
0		+	+	θ			•••
	0	1	2	3	4	5	${k}$

Fig. 3. s = 0 (the monotone case).

	•		•	•	•	•		
r	:	:	:	:	:	:	• •	
6	+	+	+	+	+	+	•••	
5	+	+	+	+	+	+	• • •	
4	\oplus	\oplus	\oplus	\oplus	\oplus	\oplus	•••	
3	+	\oplus	\oplus	\oplus	\oplus	\oplus	• • •	
2	\oplus	\oplus	?				•••	
1	+	\oplus	\oplus	?			•••	
0		+	\oplus				• • •	
	0	1	2	3	4	5	${k}$	
Fig. 4. $s = 1$.								

The reader may have noticed that there are (very few) cases in which we have not given a complete and clear answer as to whether (3.11) is valid with C = C(f), when nothing better is known. It is clear that it is not easy to differentiate between all cases without the assistance of arrays, so again we summarize the results in three arrays, one for the monotone case, one for one change of monotonicity which is special, and the third for s > 1. In addition to the symbols $+, \oplus$ and -, which have already been used in Figs. 1 and 2, here we also have the symbol \oplus which when appearing in entry (k, r) means that (3.8) and (3.9) do not hold but (3.10) holds with C = C(f). We have indicated the still open cases by ?.

Remark. Note that while in Fig. 3 the open cases ? are either \ominus or -, in Figs. 4 and 5 they may also be \oplus 's.

We conclude this section with a result on simultaneous approximation in comonotone approximation, due to Kopotun [16].

Theorem 3.18. If $f \in C^1[-1,1] \cap \Delta^1(Y_s)$, then there exists a constant C = C(s) such that for every $n \ge C/d(Y_s)$, a polynomial $p_n \in \Pi_n \cap \Delta^1(Y_s)$ exists, simultaneously yielding

$$\|f-p_n\| \leq \frac{C}{n} \omega^{\varphi} \left(f', \frac{1}{n}\right)_{\infty}$$

r	:	÷	:	:	:	:	· • •
2s+4	+	+	+	+	+	+	•••
2s + 3	+	+	+	+	+	+	•••
2s+2	\oplus	\oplus	\oplus	\oplus	\oplus	\oplus	•••
:	÷	÷	÷	÷	÷	:	÷
3	\oplus	\oplus	\oplus	\oplus	\oplus	\oplus	•••
2	\oplus	\oplus	?				•••
1	+	\oplus	\oplus	?			•••
0		+	\oplus				•••
	0	1	2	3	4	5	${m k}$
Fig. 5. $s > 1$.							

and

$$||f'-p'_n|| \leq \frac{C}{d_0}\omega^{\varphi}\left(f',\frac{1}{n}\right)_{\infty},$$

where $d_0 := \min\{\sqrt{1+y_s}, \sqrt{1-y_1}\}.$

4. Convex and coconvex approximation

We turn to convex approximation. Linear approximation methods similar to the ones for monotone approximation yielded estimates involving second moduli of smoothness of various types, while on the negative side, Shvedov [34] proved that it is impossible to get an estimate involving $\omega_4(f, 1)_p$ with an absolute constant. See also [36] for related results.

In 1994--1996 the gap between the affirmative estimates and the negative ones was closed in a series of papers by DeVore, Hu, Kopotun, Leviatan and Yu (see [12,14,1]) who proved using nonlinear methods,

Theorem 4.1. Let $f \in \mathbb{L}_p[-1,1] \cap \Delta^2$, 0 . Then there is an absolute constant <math>C = C(p), so that for each $n \ge 2$

$$E_n^{(2)}(f)_p \leqslant C\omega_3^{\varphi}\left(f,\frac{1}{n}\right)_p.$$

For convex approximation in the sup norm of convex functions we know a little more. Kopotun [14] has obtained some pointwise estimates. He has proved

Theorem 4.2. Let $f \in C^r[-1,1] \cap \Delta^2$, $0 \leq r \leq 2$. Then for each $n \geq 2$, a polynomial $p_n \in \Pi_n \cap \Delta^2$, exists such that

$$|f^{(i)}(x) - p_n^{(i)}(x)| \leq C\omega_{r-i}(f^{(i)}, \rho_n(x)), \quad 0 \leq i \leq r, \ -1 \leq x \leq 1.$$

In fact, for convex differentiable functions with at least two continuous derivatives, according to Shevchuk [32, p. 148, Theorem 17.2], Manya proved, but has never published, the following estimates.

Theorem 4.3. If $f \in C^r[-1,1] \cap \Delta^2$, $r \ge 2$, then for each $n \ge r + k - 1$, there exists a polynomial $p_n \in \Pi_n \cap \Delta^2$, such that

$$|f(x) - p_n(x)| \leq C\rho_n^r(x)\omega_k(f^{(r)}, \rho_n(x))_{\infty},$$

$$(4.1)$$

where C = C(r,k). In particular

$$E_n^{(2)}(f)_{\infty} \leq C n^{-r} \omega_k(f^{(r)}, 1/n)_{\infty}, \quad n \geq r+k-1.$$

Clearly, by virtue of Shvedov's result [34], for $f \in C[-1, 1] \cap \Delta^2$, one cannot, in general, achieve pointwise estimates of the type (4.1), where the right-hand side is $\omega_4(f, \rho_n(x))$. Very recently at a conference in Kiev, L.P. Yushenko, a student of Shevchuk announced proving that for $f \in C^1[-1, 1] \cap \Delta^2$, one cannot, in general, even have estimates of the type (4.1) where the right-hand side is $\rho_n(x)\omega_3(f', \rho_n(x))$.

Estimates involving the modified D–T moduli are due to Kopotun [15]. They can be summarized in the following result.

Theorem 4.4. Let $r,k \ge 0$. Then for every convex $f \in C_{\varphi}^{r}$

$$E_n^{(2)}(f)_{\infty} \leq C n^{-r} \omega_{k,r}^{\varphi} \left(f^{(r)}, \frac{1}{n} \right), \quad n \geq r+k-1,$$

with C = C(r,k), if and only if either $0 \le r + k \le 3$, or $r \ge 5$.

We know even less about coconvex approximation, and what we know is restricted to the sup-norm. Recent results of Kopotun et al. [21], yield

Theorem 4.5. Let $f \in \mathbb{C}[-1,1] \cap \Delta^2(Y_s)$. Then there exists a constant C = C(s) such that for each $n \ge C/d(Y_s)$,

$$E_n^{(2)}(f,Y_s)_{\infty} \leq C\omega_3^{\varphi}\left(f,\frac{1}{n}\right)_{\infty},$$

where $d(Y_s)$ was defined in (3.4).

It is also possible to obtain simultaneous approximation of f and its first and second derivatives when they exist while coconvexly approximating f (see [16,20]), namely,

Theorem 4.6. Let $f \in \mathbb{C}^r[-1,1] \cap \Delta^2(Y_s)$, $1 \leq r \leq 2$. Then there exists a constant C = C(s) such that for each $n \geq C/d(Y_s)$, there exist polynomials $p_n \in \Pi_n \cap \Delta^2(Y_s)$, such that if r = 1, we simultaneously have

$$\|f^{(i)} - p_n^{(i)}\|_{\infty} \leq \frac{C}{n^{1-i}} \omega_2^{\varphi} \left(f', \frac{1}{n}\right)_{\infty}, \quad 0 \leq i \leq 1;$$

and if r = 2, we simultaneously have

$$\|f^{(i)} - p_n^{(i)}\|_{\infty} \leq \frac{C}{n^{2-i}} \omega^{\varphi} \left(f^{\prime\prime}, \frac{1}{n}\right)_{\infty}, \quad 0 \leq i \leq 1,$$

and

$$\|f''-p_n''\|_{\infty} \leqslant \frac{C}{d_0} \omega^{\varphi} \left(f'', \frac{1}{n}\right)_{\infty},$$

where d_0 was defined in Theorem 3.18.

5. Relaxing the constraints

In an effort to improve the estimates beyond what we have seen, we [25,29], have recently attempted to approximate a function $f \in C^r[-1,1] \cap \Delta^1(Y_s)$ by polynomials which are comonotone with it in a major portion of the interval, but not necessarily in small neighborhoods of the points Y_s and the endpoints ± 1 , in other words relaxing a little the comonotonicity requirements. To be specific, given Y_s , $s \ge 0$, we set

$$O(n, Y_s) := [-1, 1] \cap \bigcup_{i=1}^{s} (y_i - \rho_n(y_i), y_i + \rho_n(y_i))$$

and

$$O^*(n, Y_s) := O(n, Y_s) \cup [-1, -1 + 1/n^2] \cup [1 - 1/n^2, 1].$$

Then we have the following results (compare with Theorem 3.11).

Theorem 5.1. For each natural number M, there is a constant C=C(s,M), so that if $f \in \mathbb{C}[-1,1] \cap \Delta^1(Y_s)$, then for every $n \ge 2$ a polynomial $P_n \in \Pi_n$ which is comonotone with f on $I \setminus O^*(Mn, Y_s)$ exists (i.e.,

$$p'_n(x)\prod_{i=1}^{3}(x-y_i) \ge 0, \quad x \in [-1,1] \setminus O^*(Mn,Y_s)),$$

such that

$$\|f-P_n\|_{\infty} \leqslant C\omega_3^{\varphi}\left(f,\frac{1}{n}\right).$$

If we assume that f is differentiable, then we do not need to relax the comonotonicity requirements near the endpoints. Namely, we have

Theorem 5.2. For each $k \ge 1$ and any natural number M, there is a constant C = C(k, s, M), for which if $f \in \Delta^1(Y_s) \cap \mathbb{C}^1[-1, 1]$, then for every $n \ge k$, a polynomial $P_n \in \Pi_n$, which is comonotone with f on $I \setminus O(Mn, Y)$ exists, such that

$$\|f-P_n\|_{\infty} \leq C \frac{1}{n} \omega_k^{\varphi} \left(f', \frac{1}{n}\right)_{\infty}.$$

Further, if we relax the requirements near the endpoints, then we can have

Theorem 5.3. For each $k \ge 1$ and any natural number M, there is a constant C = C(k, s, M), for which if $f \in \Delta^1(Y_s) \cap \mathbb{C}^1_{\varphi}$, then for every $n \ge k$, a polynomial $P_n \in \Pi_n$ which is comonotone with f on $I \setminus O^*(Mn, Y)$ exists, such that

$$\|f-P_n\|_{\infty} \leq C \frac{1}{n} \omega_{k,1}^{\varphi} \left(f', \frac{1}{n}\right).$$

We also have improved pointwise estimates (compare with Fig. 2 and the paragraph preceding it).

Theorem 5.4. There are a natural number M = M(s) and a constant C(s) such that if $f \in \mathbb{C}^1[-1,1] \cap \Delta^1(Y_s)$, then for every $n \ge 2$, a polynomial $p_n \in \Pi_n$, which is comonotone with f on $[-1,1] \setminus O^*(Mn, Y_s)$ exists, such that

$$|f(x) - p_n(x)| \leq C(s)\omega_3(f, \rho_n(x)), \quad -1 \leq x \leq 1.$$

Also

Theorem 5.5. There are a natural number M = M(s,k) and a constant C = C(s,k) for which, if $f \in \mathbb{C}^1[-1,1] \cap \Delta^1(Y_s)$, then for each $n \ge k$, a polynomial $p_n \in \Pi_n$ which is comonotone with f on $[-1,1] \setminus O(Mn,Y)$ exists such that

$$|f(x) - p_n(x)| \leq C(s,k)\rho_n(x)\omega_k(f',\rho_n(x)), \quad -1 \leq x \leq 1.$$

Remark. One should note one major difference between Theorems 5.1–5.3 which yield norm estimates, and Theorems 5.4 and 5.5 which yield pointwise estimates. The excluded neighborhoods in the former theorems may be taken proportionally as small as we wish (the number M may be arbitrarily big), while in the latter theorems the neighborhoods may not be too small (there is a number M = M(s) or M = M(s, k), as the case may be, and it may not be too big) as can be seen from the following (see [25]).

Theorem 5.6. For each $A \ge 1$ and any $n \ge 60A$, there exists a collection $Y_2^n := \{-1 < y_2^n < y_1^n < 1\}$, and a function $f_n \in \mathbb{C}[-1,1] \cap \Delta^1(Y_2^n)$, such that any polynomial $p_n \in \Pi_n$ which is comonotone with f_n on $[-1,1] \setminus O^*(27n, Y_2^n)$, necessarily satisfies

$$\left\|\frac{f_n-p_n}{\omega(f_n,\rho_n(\cdot))_{\infty}}\right\|>A.$$

Finally, we cannot push the estimates to ω_4 by relaxing the comonotonicity requirements on the *n*th polynomial, on any set of positive measure which tends to 0 when $n \to \infty$. In order to state the results we need some notation. Given an $\varepsilon > 0$ and a function $f \in \Delta^1(Y_s)$, we denote

$$E_n^{(1)}(f;\varepsilon;Y) := \inf_{p_n} \|f - p_n\|,$$

where the infimum is taken over all polynomials $p_n \in \Pi_n$ satisfying

$$\operatorname{meas}(\{x: P'_n(x)\Pi(x, Y_s) \ge 0\} \cap I) \ge 2 - \varepsilon.$$

The following was proved by DeVore et al. [2], for monotone approximation and by Leviatan and Shevchuk [29], when the function changes monotonicity.

Theorem 5.7. Given Y_s . For each sequence $\bar{\varepsilon} := \{\varepsilon_n\}_{n=1}^{\infty}$, of nonnegative numbers tending to 0, there exists a function $f := f_{\bar{\varepsilon}} \in \Delta^1(Y_s)$ such that

$$\limsup_{n\to\infty}\frac{E_n^{(1)}(f;\varepsilon_n;Y)}{\omega_4(f,1/n)_{\infty}}=\infty.$$

Following up on the above ideas Hu et al. [11] have investigated the analogous nearly positive and copositive approximation, and two variants of nearly intertwining approximation in $\mathbb{L}_p[-1, 1]$, $1 \le p \le \infty$. Again we will not discuss intertwining approximation here and the interested reader should consult that paper. The nearly copositive estimates they have obtained (compare with the statements following Theorems 2.2 and 2.4) are,

Theorem 5.8. If $f \in \mathbb{L}_p[-1,1] \cap \Delta^0(Y_s)$, $1 \le p < \infty$, then for each $n \ge 1$, there is a polynomial $p_n \in \Pi_n$ which is copositive with f in $[-1,1] \setminus O^*(n, Y_s)$, and such that

$$||f-p_n||_p \leq C\omega_2^{\varphi}\left(f,\frac{1}{n}\right)_p,$$

where $C = C(p, Y_s)$.

Furthermore, if $f \in W_p^1[-1,1] \cap \Delta^0(Y_s), \ 1 \leq p < \infty$, then

$$\|f-p_n\|_p \leq \frac{C}{n}\omega_k^{\varphi}\left(f',\frac{1}{n}\right)_p,$$

where $C = C(p, k, Y_s)$.

Conversely, for each $1 \le p < \infty$, any constant A > 0 and every $n \ge 1$, there exists a function $f := f_{p,n,A} \in \mathbb{L}_p[-1,1] \cap \Delta^0$, for which if a polynomial $p_n \in \Pi_n$ is nonnegative in $[-1+1/n^2, 1-1/n^2]$, then

 $||f - p_n||_p > A\omega_3(f, 1)_p.$

Also, for each 1 , any constant <math>A > 0 and every $n \ge 1$, there exists a function $f := f_{p,n,A} \in \mathbb{L}_p[-1,1] \cap \Delta^0(Y_s)$, for which if a polynomial $p_n \in \Pi_n$ is copositive with it in $[-1,1] \setminus O^*(n,Y_s)$, then

$$||f-p_n||_p > A\omega_3\left(f,\frac{1}{n}\right)_p,$$

and if p = 1 we can achieve

$$\|f-p_n\|_1 > A\omega_4\left(f,\frac{1}{n}\right)_1.$$

(Note the gap between the affirmative and negative estimates in the case p = 1.)

If $p = \infty$, we do not have to deal with nonnegative functions, whence we assume $s \ge 1$. What we have is

Theorem 5.9. If $f \in \mathbb{C}[-1,1] \cap \Delta^0(Y_s)$, $s \ge 1$, then for each $n \ge k-1$, there is a polynomial $p_n \in \Pi_n$ which is copositive with f in $[-1,1] \setminus O^*(n, Y_s)$, and such that

$$||f-p_n||_{\infty} \leq C \omega_k^{\varphi} \left(f, \frac{1}{n}\right)_{\infty},$$

where $C = C(k, Y_s)$

References

- [1] R.A. DeVore, Y.K. Hu, D. Leviatan, Convex polynomial and spline approximation in L_p , 0 , Constr. Approx. 12 (1996) 409–422.
- [2] R.A. DeVore, D. Leviatan, I.A. Shevchuk, Approximation of monotone functions: a counter example, in: A. Le Méhauté, C. Rabut, L.L. Schumaker (Eds.), Curves and Surfaces with Applications in CAGD, Proceedings of the Chamonix Conference, 1996, Vanderbilt University Press, Nashville, TN, 1997, pp. 95–102.
- [3] R.A. DeVore, D. Leviatan, X.M. Yu, Polynomial Approximation in L_p (0 < p < 1), Constr. Approx. 8 (1992) 187–201.
- [4] R.A. DeVore, X.M. Yu, Pointwise estimates for monotone polynomial approximation, Constr. Approx. 1 (1985) 323–331.
- [5] G.A. Dzyubenko, Pointwise estimates for comonotone approximation, Ukrain. Mat. Zh. 46 (1994) 1467–1472 (in Russian).
- [6] G.A. Dzyubenko, Copositive pointwise approximation, Ukrain. Mat. Zh. 48 (1996) 326–334 (in Russian); English transl. in Ukrainian Math. J. 48 (1996) 367–376.
- [7] G.A. Dzyubenko, J. Gilewicz, I.A. Shevchuk, Piecewise monotone pointwise approximation, Constr. Approx. 14 (1998) 311–348.
- [8] G.A. Dzyubenko, V.V. Listopad, I.A. Shevchuk, Uniform estimates of monotone polynomial approximation, Ukrain. Mat. Zh. 45 (1993) 38–43 (in Russian).
- [9] H.H. Gonska, D. Leviatan, I.A. Shevchuk, H.-J. Wenz, Interpolatory pointwise estimates for polynomial approximation, Constr. Approx., to appear.
- [10] Y.K. Hu, K. Kopotun, X.M. Yu, Constrained approximation in Sobolev spaces, Canad. Math. Bull. 49 (1997) 74-99.
- [11] Y.K. Hu, K. Kopotun, X.M. Yu, Weak copositive and intertwining approximation, J. Approx. Theory 96 (1999) 213–236.
- [12] Y.K. Hu, D. Leviatan, X.M. Yu, Convex polynomial and spline approximation in C[−1, 1], Constr. Approx. 10 (1994) 31–64.
- [13] Y.K. Hu, X.M. Yu, The degree of copositive approximation and a computer algorithm, SIAM J. Numer. Anal. 33 (1996) 388–398.
- [14] K.A. Kopotun, Pointwise and uniform estimates for convex approximation of functions by algebraic polynomials, Constr. Approx. 10.2 (1994) 153–178.
- [15] K.A. Kopotun, Uniform estimates of monotone and convex approximation of smooth functions, J. Approx. Theory 80 (1995) 76–107.
- [16] K.A. Kopotun, Coconvex polynomial approximation of twice differentiable functions, J. Approx. Theory 83 (1995) 141–156.
- [17] K.A. Kopotun, On copositive approximation by algebraic polynomials, Anal. Math. 21 (1995) 269–283.
- [18] K.A. Kopotun, On *k*-monotone polynomial and spline approximation in L_p , 0 (quasi)norm, in: C. Chui, L. Schumaker (Eds.), Approximation Theory VIII, Vol. 1: Approximation and Interpolation, World Scientific, Singapore, 1995, pp. 295–302.

- [19] K. Kopotun, D. Leviatan, Comonotone polynomial approximation in $L_p[-1,1]$, 0 , Acta Math. Hungar. 77 (1997) 301–310.
- [20] K. Kopotun, D. Leviatan, Degree of simultaneous coconvex polynomial approximation, Result Math. 34 (1998) 150–155.
- [21] K. Kopotun, D. Leviatan, I.A. Shevchuk, The degree of coconvex polynomial approximation, Proc. Amer. Math. Soc. 127 (1999) 409–415.
- [22] D. Leviatan, Shape preserving approximation by polynomial and splines, in: J. Szabados, K. Tandori (Eds.), Approximation and Function Series, Bolyai Society, Budapest, Hungary, 1996, pp. 63–84.
- [23] D. Leviatan, I.A. Shevchuk, Counter examples in convex and higher order constrained approximation, East J. Approx. 1 (1995) 391–398.
- [24] D. Leviatan, I.A. Shevchuk, Some positive results and counter examples in comonotone approximation, J. Approx. Theory 89 (1997) 195–206.
- [25] D. Leviatan, I.A. Shevchuk, Nearly comonotone approximation, J. Approx. Theory 95 (1998) 53-81.
- [26] D. Leviatan, I.A. Shevchuk, Monotone approximation estimates involving the third modulus of smoothness, in: Ch.K. Chui, L.L. Schumaker (Eds.), Approx. Theory IX, Vanderbilt University Press, Nashville, TN, 1998, pp. 223–230.
- [27] D. Leviatan, I.A. Shevchuk, Some positive results and counterexamples in comonotone approximation II, J. Approx. Theory 100 (1999) 113–143.
- [28] D. Leviatan, I.A. Shevchuk, More on comonotone polynomial approximation, Constr. Approx., to appear.
- [29] D. Leviatan, I.A. Shevchuk, Nearly comonotone approximation II, Acta Math. Szeged, to appear.
- [30] I.A. Shevchuk, On coapproximation of monotone functions, Dokl. Akad. Nauk SSSR 308 (1989) 537–541; English transl. in Soviet Math. Dokl. 40 (1990) 349–354.
- [31] I.A. Shevchuk, Approximation of monotone functions by monotone polynomials, Russ. Akad. Nauk Matem. Sbornik 183 (1992) 63–78; English transl. in Russ. Acad. Sci. Sbornik Math. 76 (1993) 51–64.
- [32] I.A. Shevchuk, Polynomial Approximation and Traces of Functions Continuous on a Segment, Naukova Dumka, Kiev, 1992 (in Russian).
- [33] I.A. Shevchuk, One example in monotone approximation, J. Approx. Theory 86 (1996) 270-277.
- [34] A.S. Shvedov, Orders of coapproximation of functions by algebraic polynomials, Mat. Zametki 29 (1981) 117–130; English transl. in Math. Notes 29 (1981) 63–70.
- [35] X. Wu, S.P. Zhou, A problem on coapproximation of functions by algebraic polynomials, in: A. Pinkus, P. Nevai (Eds.), Progress in Approximation Theory, Academic Press, New York, 1991, pp. 857–866.
- [36] X. Wu, S.P. Zhou, A conterexample in comonotone approximation in L^p space, Colloq. Math. 114 (1993) 265–274.
- [37] S.P. Zhou, A counterexample in copositive approximation, Israel J. Math. 78 (1992) 75-83.
- [38] S.P. Zhou, On comonotone approximation by polynomials in L^p space, Analysis 13 (1993) 363–376.



Journal of Computational and Applied Mathematics 121 (2000) 95-112

JOURNAL OF COMPUTATIONAL AND APPLIED MATHEMATICS

www.elsevier.nl/locate/cam

Classical orthogonal polynomials: dependence of parameters

A. Ronveaux^{a,1}, A. Zarzo^{b,2}, I. Area^c, E. Godoy^{c,*,3}

^a Mathematical Physics, Facultés Universitaires Notre-Dame de la Paix, B-5000 Namur, Belgium

^bInstituto Carlos I de Física Teórica y Computacional, Facultad de Ciencias, Universidad de Granada, 18071-Granada, Spain

^cDepartamento de Matemática Aplicada, Escuela Técnica Superior de Ingenieros Industriales y Minas, Universidad de Vigo, 36200-Vigo, Spain

Received 15 July 1999; received in revised form 20 February 2000

Abstract

Most of the classical orthogonal polynomials (continuous, discrete and their q-analogues) can be considered as functions of several parameters c_i . A systematic study of the variation, infinitesimal and finite, of these polynomials $P_n(x, c_i)$ with respect to the parameters c_i is proposed. A method to get recurrence relations for connection coefficients linking $(\partial^r/\partial c_i^r)P_n(x, c_i)$ to $P_n(x, c_i)$ is given and, in some situations, explicit expressions are obtained. This allows us to compute new integrals or sums of classical orthogonal polynomials using the digamma function. A basic theorem on the zeros of $(\partial/\partial c_i)P_n(x, c_i)$ is also proved. © 2000 Elsevier Science B.V. All rights reserved.

MSC: 33C25; 42C05; 33B15

Keywords: Classical orthogonal polynomials; Difference and q-derivative operators; Digamma function

1. Introduction. Notations. Motivations

Classical orthogonal polynomials (continuous, discrete or q-analogues) contain parameters c_i appearing in the orthogonality weight $\varrho(x, c_i)$. The dependence of $P_n(x, c_i)$ on the parameter c_i can be

³ The work of E.G. has partial financial support by DGES (Ministerio de Educación y Cultura of Spain) under contract DGES PB-96-0952.

0377-0427/00 - see front matter © 2000 Elsevier Science B.V. All rights reserved. PII: S0377-0427(00)00350-2

^{*} Corresponding author.

E-mail address: egodoy@dma.uvigo.es (E. Godoy).

¹This work was partially financed by the FNRS (Fonds National de la Recherche Scientifique (Belgium)) which supported the visit of A.R. in Vigo.

² A.Z. wishes to acknowledge partial financial support from DGES (Ministerio de Educación y Cultura of Spain) under contract PG95-1205 and also from Junta de Andalucía (Grant FQM-0207). Permanent address: Departamento de Matemática Aplicada, Escuela Técnica Superior de Ingenieros Industriales, Universidad Politécnica de Madrid, c/José Gutiérrez Abascal 2, 28006 Madrid, Spain.

analyzed in an infinitesimal way $\partial P_n(x,c_i)/\partial c_i$ or in a finite way, $\Delta_{c_i}P_n(x,c_i)$ or $D_{q(c_i)}P_n(x,c_i)$ for any parameter c_i denoted from now on by c. By Δ_c and $D_{q(c)}$ we denote the usual difference and q-derivative operators acting on the parameter c.

The parameter derivative does not change the polynomial character (in x) of the varied polynomials, denoted by $(\partial/\partial c)P_n(x,c) \equiv \dot{P}_n(x,c)$, and it is therefore natural to expand $\dot{P}_n(x,c)$ in the starting basis $\{P_n(x,c)\}_n$.

Explicit results in this direction were given by Froehlich [6], Koepf [14], Koepf and Schmersau [15], investigating the first parameter derivative of classical continuous and discrete families.

The first derivative of the hypergeometric function

$$_{p}F_{q}\left(\begin{array}{c}a_{1},\ldots,a_{p}\\b_{1},\ldots,b_{q}\end{array}\middle|x\right)$$

with respect to a_i or b_i appears in [19], but it is, in general, no more hypergeometric function.

On the other hand, integer change in a_i or b_j are controlled by contiguous relations, but it seems that the study is not entirely covered when the parameter's changes are not integer.

First derivatives of Jacobi polynomial $P_n^{(\alpha,\beta)}(x)$ with respect to α or β too appear in an unexpected way in spectral methods as Galerkin technique in order to approximate solution u(x,t) of evolutive partial differential equations [5,6].

Approximate solution $u^{N}(x,t)$ is developed in Jacobi polynomials in [6] as

$$u^{N}(x,t) = (1-x)^{\alpha(t)}(1+x)^{\beta(t)} \sum_{k=0}^{N-1} \hat{u}_{k}(t) P_{k}^{(\alpha(t),\beta(t))}(x)$$

and obviously partial derivative with respect to t implies first the knowledge of partial derivative with respect to α and β .

1.1. Notations

The Δ_{ω} and ∇_{ω} difference operators are defined by

$$(\varDelta_{\omega}f)(x) := f(x+\omega) - f(x), \quad (\nabla_{\omega}f)(x) := \varDelta_{\omega}f(x-\omega).$$

When $\omega = 1$ the operator Δ_{ω} becomes the forward difference operator Δ and ∇_{ω} is the backward difference operator ∇ .

The q-derivative operator D_q is defined [11] as

$$(D_q f)(x) := \frac{f(qx) - f(x)}{(q-1)x}, \quad q \neq 1, \ x \neq 0$$

and $(D_q f)(0) := f'(0)$, provided f'(0) exists.

Classical orthogonal polynomials $y(x) = P_n(x,c)$ are solutions of the second-order differential [17], difference [18] or *q*-difference equations [4,11,16]

$$\sigma(x)y''(x) + \tau(x)y'(x) + \lambda_n y(x) = 0, \qquad (1.1)$$

$$\sigma(x)(\Delta \nabla y)(x) + \tau(x)(\Delta y)(x) + \lambda_n y(x) = 0, \qquad (1.2)$$

$$\sigma(x)(D_q D_{q^{-1}} y)(x) + \tau(x)(D_q y)(x) + [\lambda_n] y(x) = 0,$$
(1.3)

$$\lambda_n = n(\frac{1}{2}(1-n)\sigma'' - \tau')$$

in (1.1) and (1.2), and

$$[\lambda_n] = [n]_q \left(\frac{1}{2}[1-n]_q \sigma'' - \tau'\right), \quad \text{with } [x]_q := \frac{q^x - 1}{q - 1}, \quad x \in \mathbb{C},$$

in the q-difference equation (1.3).

Thus, if c is any parameter, we can define operators acting on c (called **dot** operators) as follows:

$$(\dot{D}f)(x,c) := \frac{\partial f(x,c)}{\partial c}, \quad (\dot{\Delta}f)(x,c) := f(x,c+1) - f(x,c),$$
$$(\dot{\Delta}_{\omega}f)(x,c) := f(x,c+\omega) - f(x,c), \quad (\dot{D}_{q}f)(x,c) := \frac{f(x,qc) - f(x,c)}{(q-1)c}$$

Let $\mathscr{L} = \{D, \Delta, \nabla, \Delta_{\omega}, \nabla_{\omega}, D_q, D_{q^{-1}}\}$ and $\dot{\mathscr{L}} = \{\dot{D}, \dot{\Delta}, \dot{\nabla}, \dot{\Delta}_{\omega}, \dot{\nabla}_{\omega}, \dot{D}_q, \dot{D}_{q^{-1}}\}$. It is easy to check that any operator belonging to \mathscr{L} commutes with any operator belonging to $\dot{\mathscr{L}}$,

 $[\mathscr{T},\dot{\mathscr{T}}] = \mathscr{T}\dot{\mathscr{T}} - \dot{\mathscr{T}}\mathscr{T} = 0, \quad \mathscr{T} \in \mathscr{L}, \ \dot{\mathscr{T}} \in \dot{\mathscr{L}}.$

The following notations will be used in the case of higher (partial) derivatives and for each operator (D for example),

$$(DP_n)(x,c) = \frac{\partial}{\partial x} P_n(x,c) \equiv P'_n(x,c), \quad (D^2 P_n)(x,c) = P''_n(x,c), \tag{1.4}$$

$$(\dot{D}P_n)(x,c) = \frac{\partial}{\partial c} P_n(x,c) \equiv \dot{P}_n(x,c), \quad (\dot{D}^2 P_n)(x,c) = \ddot{P}_n(x,c), \tag{1.5}$$

and in general

$$P_n^{(r)}(x,c) := \frac{\partial^r}{\partial x^r} P_n(x,c) = (D^r P_n)(x,c),$$
(1.6)

$$P_{n}^{[r]}(x,c) \equiv (D^{[r]}P_{n})(x,c) := \frac{\partial^{r}}{\partial c^{r}} P_{n}(x,c) = (\dot{D}^{r}P_{n})(x,c),$$
(1.7)

where the upper script bracket stands for the dot derivative and the upper script parenthesis stands for the x derivative.

1.2. Motivations

The aims of this paper are:

(i) Generate, and solve in some cases, recurrence relations for the connection coefficients $C_m^{\{r\}}(n)$ in the expansion

$$P_n^{[r]}(x,c) := \frac{\partial^r}{\partial c^r} P_n(x,c) = \sum_{m=0}^n C_m^{\{r\}}(n) P_m(x,c)$$
(1.8)

taking into account that $P_n^{[r]}(x,c)$ is a polynomial in x of degree at most n, depending on the monic or nonmonic character of the $P_n(x,c)$ family.

Moreover, the connection coefficient $C_m^{\{1\}}(n)$ is strongly related to the integral

$$\int_a^b P_n(x) P_m(x) \varrho^{[r]}(x,c) \,\mathrm{d} x,$$

where $\varrho(x,c)$ is the orthogonality weight of the P_n family, and so, the knowledge of $C_m^{\{1\}}(n)$ provides a way of computing new integrals (and sums or q-sums) of classical orthogonal polynomials in terms of some special functions.

(ii) Classify in a systematic way the study of parameter dependence of classical continuous, discrete and *q*-orthogonal polynomials.

Following the restrictions on the degree of σ and τ , in order to stay inside classical families, the number of possible parameters c can be 3 and generic c stands for c_1 , c_2 , c_3 (without q). If we now mix the 3 equations, called D, Δ , D_q , with the dot operations \dot{D} , $\dot{\Delta}$, \dot{D}_q , we get for each r the 9 situations labelled by a couple of operators belonging to the set

 $\begin{array}{ll} (D,D^{[r]}), & (D,\Delta^{[r]}), & (D,D^{[r]}_q), \\ (\Delta,D^{[r]}), & (\Delta,\Delta^{[r]}), & (\Delta,D^{[r]}_q), \\ (D_q,D^{[r]}), & (D_q,\Delta^{[r]}), & (D_q,D^{[r]}_q), \end{array}$

where the notations are those of Eqs. (1.4)–(1.7), i.e. $D^{[r]}:=\dot{D}^r$, $\Delta^{[r]}:=\dot{\Delta}^r$ and $D^{[r]}_q:=\dot{D}^r_q$. In the case r=1, Froehlich [6], Koepf and Schmersau [15] already examined the two situations (D,\dot{D}) and (Δ,\dot{D}) and they gave the corresponding expansions for (1.8) using hypergeometric representation of the *D*-classical and Δ -classical orthogonal polynomials.

(iii) Prove interlacing properties of zeros of $\dot{P}_n(x,c)$ with respect to the zeros of $P_n(x,c)$ for arbitrary orthogonal family $P_n(x,c)$, with a rather weak assumption which is precised in classical discrete situations.

2. (D, \dot{D}) , (Δ, \dot{D}) , (D_q, \dot{D}) cases. Connection problems

Assume first that r = 1. Using the commutativity of the dot and no dot operators, and knowing the action of each operator on products, we first obtain from (1.1) to (1.3) the three relations for the respective (D, \dot{D}) , (Δ, \dot{D}) , (D_q, \dot{D}) situations:

$$\sigma \dot{P}_{n}^{\prime\prime} + \tau \dot{P}_{n}^{\prime} + \lambda_{n} \dot{P}_{n} = R_{1}^{(D,D)}[P_{n}], \qquad (2.1)$$

$$\sigma \Delta \nabla \dot{P}_n + \tau \Delta \dot{P}_n + \lambda_n \dot{P}_n = R_1^{(\Lambda, \dot{D})}[P_n], \qquad (2.2)$$

$$\sigma D_q D_{q^{-1}} \dot{P}_n + \tau D_q \dot{P}_n + [\lambda_n] \dot{P}_n = R_1^{(D_q, \dot{D})} [P_n]$$
(2.3)

with

$$R_1^{(D,D)}[P_n] = -\dot{\tau}P'_n - \dot{\lambda}_n P_n \quad (\dot{\sigma} = 0),$$
(2.4)

A. Ronveaux et al. | Journal of Computational and Applied Mathematics 121 (2000) 95-112

$$R_1^{(\Delta,\dot{D})}[P_n] = -\dot{\sigma}\Delta\nabla P_n - \dot{\tau}\Delta P_n - \dot{\lambda}_n P_n, \qquad (2.5)$$

$$R_1^{(D_q,\dot{D})}[P_n] = -\dot{\sigma}D_q D_{q^{-1}} P_n - \dot{\tau}D_q P_n - [\dot{\lambda}_n] P_n.$$
(2.6)

The connection problem

$$\frac{\partial}{\partial c}P_n(x,c) = \dot{P}_n(x,c) = \sum_{m=0}^{n-1} C_m^{\{1\}}(n)P_m(x,c)$$
(2.7)

for monic classical orthogonal polynomials $P_n(x,c)$ can be solved in these situations using the recently developed Navima algorithm for nonhomogeneous problems [9] as explained in the examples. The first step consists in introducing the above development in (2.1), (2.2) or (2.3) to obtain

$$\sum_{m=0}^{n-1} C_m^{\{1\}}(n) (\lambda_n - \lambda_m) P_m = -i P_n' - \dot{\lambda}_n P_n,$$
(2.8)

$$\sum_{m=0}^{n-1} C_m^{\{1\}}(n)(\lambda_n - \lambda_m)P_m = -\dot{\sigma} \varDelta \nabla P_n - \dot{\tau} \varDelta P_n - \dot{\lambda}_n P_n, \qquad (2.9)$$

$$\sum_{m=0}^{n-1} C_m^{\{1\}}(n)([\lambda_n] - [\lambda_m])P_m = -\dot{\sigma}D_q D_{q^{-1}} P_n - \dot{\tau}D_q P_n - [\dot{\lambda}_n]P_n, \qquad (2.10)$$

using (1.1), (1.2) or (1.3), respectively.

This last situation (2.10) will be treated separately, being new. The representation of $C_m^{\{1\}}(n)$, in the (D, \dot{D}) and (Δ, \dot{D}) cases, coincides with the results given in [6,14].

The generalized derivatives cases $(D, D^{[r]})$, $(\Delta, D^{[r]})$, $(D_a, D^{[r]})$ are treated in the following way:

Let us assume that $r \ge 2$ and let us consider first the (D, \dot{D}^2) case. The dot derivative of (2.1) gives, according to the notations of (1.4)-(1.5),

$$\sigma \ddot{P}_n'' + \tau \ddot{P}_n' + \lambda_n \ddot{P}_n = R_2^{(D,\dot{D})}[P_n]$$

with

$$R_{2}^{(D,\dot{D})}[P_{n}] = \dot{R}_{1}^{(D,D)}[P_{n}] - (\dot{\sigma}\dot{P}_{n}^{''} + \dot{\tau}\dot{P}_{n}^{'} + \dot{\lambda}_{n}\dot{P}_{n}) = -2(\dot{\tau}\dot{P}_{n}^{'} + \dot{\lambda}_{n}\dot{P}_{n})$$
$$= 2\dot{R}_{1}^{(D,\dot{D})}[P_{n}] \quad (\dot{\sigma} = 0, \ddot{\tau} = 0, \ddot{\lambda}_{n} = 0), \qquad (2.11)$$

where $\dot{R}_1^{(D,\dot{D})}[P_n]$ is given in (2.4).

Results on mixed dot derivative with respect to two parameters c_1 , c_2 need some more notations. For any polynomial $P_n = P_n(c_1, c_2)$ and for any function $f = f(c_1, c_2)$ depending on two parameters, we shall denote

$$\dot{f}_{c_i} := \frac{\partial f}{\partial c_i}, \quad \dot{P}_{n,c_i} := \frac{\partial P_n}{\partial c_i}, \quad \ddot{P}_{n,c_i,c_j} := \frac{\partial^2 P_n}{\partial c_i \partial c_j}, \quad (1 \le i, j \le 2).$$
(2.12)

With these notations, Eqs. (2.1) and (2.4) are rewritten as

$$\sigma \dot{P}_{n,c_1}^{\prime\prime} + au \dot{P}_{n,c_1}^{\prime} + \lambda_n \dot{P}_{n,c_1} = - \dot{ au}_{c_1} P_n^{\prime} - \dot{\lambda}_{n,c_1} P_n.$$

Table 1

σ	τ	$\dot{\lambda}_n$	$(\dot{ au})'$	Polynomials
0	0	0	0	Stieltjes–Wigert Discrete q-Hermite I Discrete q-Hermite II
0	eq 0	0	0	Little q-Laguerre
eq 0	eq 0	0	0	Al-Salam–Carlitz I Al-Salam–Carlitz II q-Meixner $(b,c)Big q-Laguerre (a,b)Big q-Jacobi (c)$
0	eq 0	eq 0	eq 0	<i>q</i> -Laguerre Alternative <i>q</i> -Charlier Little <i>q</i> -Jacobi (<i>a</i> , <i>b</i>) Big <i>q</i> -Jacobi (<i>b</i>)
$\neq 0$	eq 0	eq 0	eq 0	Big q-Jacobi (a)

Dependence of σ , τ , λ_n and τ' with respect to the parameters for each family of polynomials belonging to the q-Hahn tableau

The new dot derivative with respect to c_2 gives

$$\sigma \ddot{P}_{n,c_1,c_2}^{\prime\prime} + \tau \ddot{P}_{n,c_1,c_2}^{\prime} + \lambda_n \ddot{P}_{n,c_1,c_2} = -\dot{\tau}_{c_1} \dot{P}_{n,c_2}^{\prime} - \dot{\lambda}_{n,c_1} \dot{P}_{n,c_2} \quad (\ddot{\tau} = \ddot{\lambda}_n = 0)$$

and, expanding $\ddot{P}_{n,c_1,c_2}(x)$ as

$$\ddot{P}_{n,c_1,c_2}(x) = \sum_{m=0}^{n-1} C_{m,c_1,c_2}^{\{2\}}(n) P_m(x),$$

we can write

$$\sum_{n=0}^{n-1} C_{m,c_1,c_2}^{\{2\}}(n)(\lambda_n - \lambda_m)P_m(x) = -\dot{\tau}_{c_1}\dot{P}'_{n,c_2}(x) - \dot{\lambda}_{n,c_1}\dot{P}_{n,c_2}(x)$$
(2.13)

like in (2.8).

It is clear that inserting (2.7) in this later expression, a relation between $C_m^{\{2\}}(n)$ and $C_m^{\{1\}}(n)$ comes out.

The general situation $(D, D^{[r]})$ gives

$$\sigma(P_n^{[r]})'' + \tau(P_n^{[r]})' + \lambda_n(P_n^{[r]}) = R_r^{(D,D^{[r]})}[P_n] = -r[\dot{\tau}(P_n^{[r-1]})' + \dot{\lambda}_n P_n^{[r-1]}].$$

Investigation of this general situation proposed in (1.8) can also be done linking, as in the case r = 2, $C_m^{\{r\}}(n)$ to $C_m^{\{r-1\}}(n)$. In $(D_q, D^{[r]})$, from the presence of q in the q-derivative operator D_q , q is not an allowed parameter,

In $(D_q, D^{[r]})$, from the presence of q in the q-derivative operator D_q , q is not an allowed parameter, and therefore in all situations belonging to the q-Hahn tableau [16], $\dot{\sigma}''$ is also equal to zero, and $[\dot{\lambda}]_n$ depends only on $\dot{\tau}'$ which can be zero or not. The situation for each family of polynomials belonging to the q-Hahn tableau [16] is shown in Table 1.

From (2.3) we see immediately that the simplest cases for q-orthogonal polynomials with parameters appear when $\dot{\sigma}$ is zero. In the next section we give the three-term recurrence relation for

100

 $C_{m,a}^{\{1\}}(n)$ in the case of little q-Jacobi polynomials which is solved when b = 0 corresponding to the little q-Laguerre polynomials.

3. Recurrence relations for the connection coefficients $C_m^{\{r\}}(n)$ using the Navima algorithm

Recurrence relations for $C_m^{\{r\}}(n)$ in (2.8), (2.9) and (2.10) are easily obtained with the Navima⁴ algorithm [3,4,8,9,20,21,24].

Let us summarize the algorithm for r = 1 in the (D, \dot{D}) case. Extensions to other cases can be done *mutatis mutandis*.

(i) We expand first classical orthogonal polynomials $P_n(x,c)$ and $P_m(x,c)$ in (2.8) using the derivative representation [8]

$$P_n(x,c) = \frac{P'_{n+1}(x,c)}{n+1} + F_n P'_n(x,c) + G_n P'_{n-1}(x,c),$$
(3.1)

where F_n and G_n are independent of x.

- (ii) If $\dot{\tau}$ depends (at most linearly) on x, we expand $xP'_n(x,c)$ in three elements of $\{P'_n(x,c)\}$ basis using the recurrence relation for the $P'_n(x,c)$ family.
- (iii) The recurrence relation for $C_m^{\{1\}}(n)$ arises from a shift of indices in (2.8) written from (i) and (ii) as a constant coefficient combination of $P'_n(x,c)$.

3.1. Laguerre case

Let

$$L_n^{(\alpha)}(x):=(-1)^n(\alpha+1)_{n-1}F_1\left(\begin{array}{c}-n\\\alpha+1\end{array}\middle|x\right),\quad n\ge 0,$$

be the monic *n*th degree Laguerre polynomial. By using this procedure, we obtain that the coefficients $C_i^{\{2\}}(n)$ in the expansion

$$\frac{\partial^2}{\partial \alpha^2} L_n^{(\alpha)}(x) = \ddot{L}_n^{(\alpha)}(x) = \sum_{j=0}^{n-2} C_j^{\{2\}}(n) L_j^{(\alpha)}(x)$$

satisfy the following nonhomogeneous two-term recurrence relation:

$$(n-j)C_{j}^{\{2\}}(n) + \frac{n-j+1}{j}C_{j-1}^{\{2\}}(n) = -2\frac{(-n)_{n-j}}{n-j}$$

with the initial condition

$$C_{n-2}^{\{2\}}(n) = n(n-1),$$

which has the following solution:

$$C_{j}^{\{2\}}(n) = \frac{(-1)^{n-j}n!}{j!} \left(\sum_{s=j}^{n-2} \frac{1}{(s-j+1)(n-s-1)} \right), \quad 0 \leq j \leq n-2.$$

⁴ http://www.uvigo.es/webs/t10/navima.

By using Mathematica [23] symbolic language, we get

$$\sum_{s=j}^{n-2} \frac{1}{(s-j+1)(n-s-1)} = \frac{2(\gamma + \psi(n-j))}{n-j},$$

where

$$\gamma := \lim_{k \to \infty} \left[1 + \frac{1}{2} + \frac{1}{3} + \dots + \frac{1}{k} - \ln k \right],$$

is the Euler constant [1, p. 250] and

$$\psi(z) := \frac{\Gamma'(z)}{\Gamma(z)} \tag{3.2}$$

is the digamma function [1, p. 258] (logarithmic derivative of Gamma function). Thus,

$$\ddot{L}_{n}^{(\alpha)}(x) = 2n! \sum_{j=0}^{n-2} \frac{(-1)^{n-j}(\gamma + \psi(n-j))}{j!(n-j)} L_{j}^{(\alpha)}(x).$$
(3.3)

Remark 1. In this case, $C_n^{\{1\}}(n)$ being independent of the parameter $c = \alpha$, the following alternate approach can be used. In [15, p. 80] one finds the expression

$$\dot{L}_{n}^{(\alpha)}(x) = \sum_{m=0}^{n-1} \frac{(-n)_{n-m}}{n-m} L_{m}^{(\alpha)}(x),$$
(3.4)

which could be also deduced using our approach. The derivative with respect to α of the above equation gives

$$\ddot{L}_{n}^{(\alpha)}(x) = \frac{\partial}{\partial \alpha} \dot{L}_{n}^{(\alpha)}(x) = \sum_{m=0}^{n-1} \frac{(-n)_{n-m}}{n-m} \frac{\partial}{\partial \alpha} L_{m}^{(\alpha)}(x).$$

By using (3.4), relation (3.3) is again obtained.

3.2. Jacobi case

Let

be the monic *n*th degree Jacobi polynomial.

Taking into account (2.11), in the connection problem

$$\ddot{P}_{n,\alpha,\alpha}^{(\alpha,\beta)}(x) := \frac{\partial^2 P_n^{(\alpha,\beta)}(x)}{\partial \alpha^2} = \sum_{k=0}^{n-1} C_{k,\alpha,\alpha}^{\{2\}}(n) P_k^{(\alpha,\beta)}(x) \quad (\alpha \neq \beta),$$

102

we obtain the following nonhomogeneous three-term recurrence relation for the coefficients $C_{k,\alpha,\alpha}^{\{2\}}(n) \equiv C_k^{\{2\}}(n)$:

$$(n - k + 1)(n + \alpha + \beta + k)C_{k-1}^{\{2\}}(n) + k((n - k)(n + \alpha + \beta + k + 1)C_{k}^{\{2\}}(n)F_{k}$$

+ $(n - k - 1)(n + \alpha + \beta + k + 2)C_{k+1}^{\{2\}}(n))$
= $-2((n - k + 1)C_{k-1}^{\{1\}}(n) + k(C_{k}^{\{1\}}(n)((n + 1)F_{k} - \beta_{k} - 1))$
+ $C_{k+1}^{\{1\}}(n)((n + 1)G_{k+1} - \gamma_{k+1})))$

valid for $1 \le k \le n - 2$, with the initial conditions

$$C_{n-1}^{\{2\}}(n) = \frac{-4n(\beta+n)}{(\alpha+\beta+2n)^3},$$

$$C_{n-2}^{\{2\}}(n) = \frac{4(n-1)n(n+\beta-1)(\beta+n)}{(-2+\alpha+\beta+2n)(-1+\alpha+\beta+2n)^3(\alpha+\beta+2n)^2} \times (1+\alpha^2-3\beta-6n+(\beta+2n)^2+\alpha(-3+2\beta+4n)),$$

where β_n and γ_n are the coefficients of the three-term recurrence relation (3.1) satisfied by monic Jacobi polynomials

$$\beta_n = \frac{\beta^2 - \alpha^2}{(2n + \alpha + \beta)(2n + 2 + \alpha + \beta)},$$
$$\gamma_n = \frac{4n(n + \alpha)(n + \alpha + \beta)(n + \beta)}{(2n + \alpha + \beta - 1)(2n + \alpha + \beta)^2(2n + \alpha + \beta + 1)},$$

 F_n and G_n are the coefficients of the derivative representation of monic Jacobi polynomials

$$F_n = \frac{2(\alpha - \beta)}{(2n + \alpha + \beta)(2n + 2 + \alpha + \beta)},$$
$$G_n = \frac{-4n(n + \alpha)(n + \beta)}{(2n + \alpha + \beta - 1)(2n + \alpha + \beta)^2(2n + \alpha + \beta + 1)}$$

and $C_k^{\{1\}}(n)$ are the coefficients in the expansion

$$\dot{P}_{n,\alpha}^{(\alpha,\beta)}(x) := \frac{\partial P_n^{(\alpha,\beta)}(x)}{\partial \alpha} = \sum_{k=0}^{n-1} C_k^{\{1\}}(n) P_k^{(\alpha,\beta)}(x)$$

given by (see [6,15])

$$C_k^{\{1\}}(n) = \frac{2^{n-k}}{(n-k)} \frac{\binom{2k+\alpha+\beta}{k}}{\binom{2n+\alpha+\beta}{n}} \frac{(\alpha+\beta+2k+1)}{(\alpha+\beta+k+n+1)} \frac{(\beta+k+1)_{n-k}}{(\alpha+\beta+k+1)_{n-k}}$$

valid for $0 \leq k \leq n - 1$.

3.3. Little q-Jacobi case

For $x, q \in \mathbb{C}$ define the *q*-shifted factorial by [2]

$$(x;q)_n := \prod_{j=0}^{n-1} (1-q^j x) \quad (n=0,1,\ldots).$$
(3.5)

Let us define the monic little q-Jacobi polynomials [13],

$$p_n(x;a,b|q) := \frac{(-1)^n q^{n(n-1)/2} (aq;q)_n}{(abq^{n+1};q)_n} {}_2\phi_1 \begin{pmatrix} q^{-n}, abq^{n+1} \\ \\ aq \end{pmatrix} = q;qx$$

where the basic hypergeometric series is defined by [2]

$${}_{r}\phi_{s}\begin{pmatrix}a_{1},a_{2},\ldots,a_{r}\\b_{1},b_{2},\ldots,b_{s}\end{vmatrix} q;z := \sum_{k=0}^{\infty} \frac{(a_{1};q)_{k}\ldots(a_{r};q)_{k}}{(q;q)_{k}(b_{1};q)_{k}\ldots(b_{s};q)_{k}} ((-1)^{k}q^{\binom{k}{2}})^{1+s-r}z^{k}$$
(3.6)

for $r, s \in \mathbb{Z}_+$ and a_1, a_2, \ldots, a_r , b_1, b_2, \ldots, b_s , $z \in \mathbb{C}$. In order to have a well-defined series, we require that $b_1, b_2, \ldots, b_s \neq q^{-k}$ $(k = 0, 1, \ldots)$.

If we consider the expansion

$$\dot{p}_{n,a}(x;a,b|q) := \frac{\partial}{\partial a} p_n(x;a,b|q) = \sum_{m=0}^{n-1} C_{m,a}^{\{1\}}(n) p_m(x;a,b|q),$$

we obtain the following three-term recurrence relation for the connection coefficients $C_{m,a}^{\{1\}}(n) \equiv C_m^{\{1\}}(n)$,

$$C_{m-1}^{\{1\}}(n)(\lambda_n - \lambda_{m-1}) + (C_m^{\{1\}}(n)F_m(\lambda_n - \lambda_m) + C_{m+1}^{\{1\}}(n)G_{m+1}(\lambda_n - \lambda_{m+1}))[m]_q = 0,$$

with the initial conditions

$$C_{n-1}^{\{1\}}(n) = \frac{-2q^n(-1+q+(b-bq)\beta_n+b(q^{1+n}-1)F_n)[n]_q}{(q-1)(abq^{2n}-1)},$$

$$C_{n-2}^{\{1\}}(n) = \frac{1}{(-1+q)(1+q)(-q+abq^{2n})}(-2q^{1+n}[n-1]_q((b-bq)\gamma_n+b(q^{1+n}-1)G_n+F_{n-1}(1-q+b((q-1)\beta_n+F_n-q^{1+n}F_n))[n]_q)),$$

where β_n and γ_n are the coefficients of the three-term recurrence relation satisfied by monic little *q*-Jacobi polynomials

$$\beta_n = \frac{q^n (1 + a + a (1 + a) bq^{1+2n} - a (1 + b) q^n (1 + q))}{(abq^{2n} - 1) (abq^{2+2n} - 1)},$$

$$\gamma_n = \frac{aq^{2n} (q^n - 1) (aq^n - 1) (bq^n - 1) (abq^n - 1)}{(abq^{2n} - 1)^2 (abq^{2n} - q) (abq^{1+2n} - 1)}$$

104
and

$$F_{n} = -\frac{aq^{n} (q-1)(1+bq^{n} (q (aq^{n}-1)-1))}{(abq^{2n}-1)(abq^{2+2n}-1)},$$

$$G_{n} = -\frac{a^{2}bq^{3n} (q^{n}-1)(q-1)(aq^{n}-1)(bq^{n}-1)}{(abq^{2n}-1)^{2} (abq^{2n}-q)(abq^{1+2n}-1)}$$

are the coefficients of the q-derivative representation

$$p_n(x;a,b|q) = \frac{D_q p_{n+1}(x;a,b|q)}{[n+1]_q} + F_n D_q p_n(x;a,b|q) + G_n D_q p_{n-1}(x;a,b|q)$$

of monic little q-Jacobi polynomials.

If b = 0 the monic little q-Jacobi polynomials are the monic little q-Laguerre polynomials [13]. In this case, the above recurrence relation becomes a two-term recurrence relation which can be explicitly solved. We have obtained

$$\dot{p}_{n}(x;a|q) := \frac{\partial}{\partial a} p_{n}(x;a|q)$$

$$= \sum_{m=0}^{n-1} (-1)^{n-m} {n \brack m}_{q} \frac{a^{n-m-1}q^{(n-m)(n+m+1)/2}(q;q)_{n-m}}{q^{n-m}-1} p_{m}(x;a|q), \qquad (3.7)$$

where $p_n(x; a|q)$ are the monic little q-Laguerre polynomials, and

$$\begin{bmatrix}n\\m\end{bmatrix}_q := \frac{(q;q)_n}{(q;q)_m (q;q)_{n-m}}$$

is the q-binomial coefficient [2].

4. Remaining cases $(D, \dot{\Delta}), (D, \dot{D}_q), (\Delta, \dot{\Delta}), (\Delta, \dot{D}_q), (D_q, \dot{\Delta}), (D_q, \dot{D}_q)$

We are faced here with computing inside the Hahn class (Δ) and the q-Hahn class (D_q) the Δ and \dot{D}_q actions. These are pure connection problems, i.e.,

$$P_n(x, c+1) = \sum_{m=0}^{n} C_m(n) P_m(x, c),$$

$$P_n(x, qc) = \sum_{m=0}^{n} C_m(n) P_m(x, c),$$
(4.1)

which can be exactly solved as before by using the Navima algorithm. $P_n(x,c)$ being given by hypergeometric functions, integer change in parameters $c \rightarrow c+1$ gives simply contiguous hypergeometric functions. If the change of parameters is not integer, like with $\dot{\Delta}_{\omega}$, contiguous hypergeometric relations are lost, but the recurrence approach developed here still continues to work. A variant of these kinds of parameter's changes occurs for instance in [7].

For higher dot derivative $\Delta^{[r]}$ and $D_q^{[r]}$ the expansion of $P_n(x, c+r)$ and $P_n(x, q^r c)$ in terms of $P_n(x, c)$ are also pure connection problems.

5. New integrals and sums involving classical orthogonal polynomials

The coefficients $C_m^{\{r\}}(n)$ in (1.8) for monic $P_m(x,c)$, explicitly computed from the recurrence, allow us to compute new integrals, sums and q-integrals involving orthogonal polynomials. In the integral case, we have

(i) The dot derivative of the following expression:

$$\int_{a}^{b} P_{n}(x)P_{m}(x)\varrho(x,c)\,\mathrm{d}x = 0, \quad m < n,$$
(5.1)

gives

$$\int_a^b \dot{P}_n(x)P_m(x)\varrho(x,c)\,\mathrm{d}x + \int_a^b P_n(x)\dot{P}_m(x)\varrho(x,c)\,\mathrm{d}x + \int_a^b P_n(x)P_m(x)\dot{\varrho}(x,c)\,\mathrm{d}x = 0.$$

Since the second integral is equal to 0 for m < n, we obtain

$$\int_{a}^{b} P_{n}(x) P_{m}(x) \dot{\varrho}(x,c) \,\mathrm{d}x = -\int_{a}^{b} \dot{P}_{n}(x) P_{m}(x) \varrho(x,c) \,\mathrm{d}x = -C_{m}^{\{1\}}(n) d_{m}^{2}, \tag{5.2}$$

where d_m^2 is the squared norm of $P_m(x)$ given by

$$d_m^2 := \int_a^b P_m^2(x)\varrho(x,c)\,\mathrm{d}x.$$

Obviously, for n < m, we have

$$\int_{a}^{b} P_{n}(x) P_{m}(x) \dot{\varrho}(x,c) \, \mathrm{d}x = -C_{n}^{\{1\}}(m) d_{n}^{2}.$$

(ii) If n = m, the dot derivative of $\int_a^b P_n^2(x,c)\varrho(x,c) dx = d_n^2$, gives

$$\int_{a}^{b} P_{n}^{2}(x,c)\dot{\varrho}(x,c)\,\mathrm{d}x = \frac{\partial}{\partial c}\,d_{n}^{2}.$$

(iii) The integral

$$\int_a^b P_n(x) P_m(x) \ddot{\varrho}(x,c) \,\mathrm{d}x$$

will be computed after linearization

$$P_n(x,c) P_m(x,c) = \sum_{k=0}^{n+m} L_k(n,m) P_k(x,c),$$

obtaining results on

$$\int_{a}^{b} P_{k}(x)\ddot{\varrho}(x,c) \,\mathrm{d}x.$$
Using (5.2), we get
$$\int_{a}^{b} P_{k}(x)\dot{\varrho}(x,c) \,\mathrm{d}x = -C_{0}^{\{1\}}(k) \,d_{0}^{2} = -C_{0}^{\{1\}}(k) \int_{a}^{b} \varrho(x,c) \,\mathrm{d}x \tag{5.3}$$
and the dot derivative of the above expression gives

pression gi

$$\int_{a}^{b} P_{k}(x)\ddot{\varrho}(x,c) \,\mathrm{d}x = -\frac{\partial}{\partial c} [C_{0}^{\{1\}}(k) d_{0}^{2}] - \int_{a}^{b} \dot{P}_{k}(x)\dot{\varrho}(x,c) \,\mathrm{d}x$$
$$= -\frac{\partial}{\partial c} [C_{0}^{\{1\}}(k) d_{0}^{2}] - C_{0}^{\{1\}}(k) \int_{a}^{b} \dot{\varrho}(x,c) \,\mathrm{d}x + \sum_{m=1}^{k-1} C_{m}^{\{1\}}(k) C_{0}^{\{1\}}(m) d_{0}^{2}$$

using (2.7).

Of course, when the scalar product (5.1) is not given by an integral, but by a sum or a q-integral, the previous results are still valid.

5.1. Examples

5.1.1. Laguerre integrals

In the Laguerre case, $\varrho(x, \alpha) = e^{-x}x^{\alpha}$, $\alpha > -1$. Since $\dot{\varrho}(x, \alpha) = \varrho(x, \alpha)\ln(x)$ and using (3.4) we have,

$$\int_0^\infty \ln(x) L_n^{(\alpha)}(x) L_m^{(\alpha)}(x) e^{-x} x^{\alpha} \, \mathrm{d}x = \frac{m! \Gamma \left(1 + \alpha + m\right)(-n)_{n-m}}{m-n}, \quad m < n.$$

Moreover,

$$\int_0^\infty L_n^{(\alpha)}(x)\ddot{\varrho}(x,\alpha)\,\mathrm{d}x = \int_0^\infty L_n^{(\alpha)}(x)\varrho(x,\alpha)\ln^2(x)\,\mathrm{d}x$$
$$= 2(-1)^n\Gamma(1+\alpha)\Gamma(n)(\gamma-\psi(1+\alpha)+\psi(n)),$$

where $\psi(z)$ is the digamma function (3.2).

All these integrals can be computed for fixed n and m from [10, p. 576] expanding the Laguerre polynomials in terms of x^j .

5.1.2. Jacobi integrals

In the Jacobi case $\varrho(x, \alpha, \beta) = (1 - x)^{\alpha} (1 + x)^{\beta}$, $\alpha, \beta > -1$. From (5.2) we obtain, for m < n,

$$\int_{-1}^{1} P_n^{(\alpha,\beta)}(x) P_m^{(\alpha,\beta)}(x) (1-x)^{\alpha} (1+x)^{\beta} \ln(1-x) dx$$
$$= \frac{2^{1+\alpha+\beta+m+n}\Gamma(1+\alpha+m)\Gamma(1+\alpha+\beta+m)\Gamma(1+\alpha+\beta+m)}{(m-n)(1+\alpha+\beta+m+n)\Gamma(1+\alpha+\beta+2m)\Gamma(1+\alpha+\beta+2n)}$$

and

$$\int_{-1}^{1} P_{n}^{(\alpha,\beta)}(x) P_{m}^{(\alpha,\beta)}(x) (1-x)^{\alpha} (1+x)^{\beta} \ln(1+x) dx$$

= $\frac{(-1)^{n-m} 2^{1+\alpha+\beta+m+n} \Gamma (1+\beta+m)}{(m-n)(1+\alpha+\beta+m+n)} \frac{\Gamma (1+\alpha+\beta+m) \Gamma (1+n) \Gamma (1+\alpha+n)}{\Gamma (1+\alpha+\beta+2m) \Gamma (1+\alpha+\beta+2n)}.$

5.1.3. Meixner sum

In Meixner case, $\rho(x, \gamma, \mu) = \mu^x \Gamma(\gamma + x) / \Gamma(x + 1)$, $\gamma > 0$, $0 < \mu < 1$. Let us introduce monic Meixner polynomials

$$M_n^{(\gamma,\mu)}(x) := \left(\frac{\mu}{\mu-1}\right)^n (\gamma)_{n\,2} F_1\left(\begin{array}{c}-n,-x\\\gamma\end{array}\right) 1 - \frac{1}{\mu}.$$

From (5.2) and [15, p. 92], we obtain, for m < n,

$$\sum_{k=0}^{\infty} M_n^{(\gamma,\mu)}(k) M_m^{(\gamma,\mu)}(k) \frac{\mu^k \Gamma(\gamma+k)}{\Gamma(k+1)} \psi(\gamma+k)$$
$$= \frac{(1-\mu)^{-\gamma-m-n}(-\mu)^{-m+n} \mu^m \Gamma(\gamma+m) \Gamma(1+n)}{m-n},$$

where $\psi(z)$ is the digamma function (3.2).

5.1.4. Little q-Laguerre sum

In little q-Laguerre case, $\varrho(k, a, q) = (aq)^k/(q; q)_k$, 0 < aq < 1. In this situation, from (5.2) and (3.7), we obtain, for m < n,

$$\sum_{k=1}^{\infty} p_n(k;a|q) p_m(k;a|q) kq \frac{(aq)^{k-1}}{(q;q)_k} = \frac{(-1)^{n-m+1} a^{n-m-1} q^{(n-m)(1+m+n)/2} (aq^m)^m (q;q)_n (aq;q)_m}{(q^{n-m}-1)(aq;q)_\infty}$$

Many other results can be generated in the same way with other families and higher-order dot derivatives.

6. Interlacing properties of zeros of \dot{P}_n with respect to zeros of P_n

The following theorem is answer to a question indicated in [6] where the zeros of

$$\frac{\partial}{\partial \alpha} P_n^{(\alpha,\beta)}(x)$$

were investigated numerically, for $\alpha = \beta = -0.7$ and 10. With a weak assumption, it is proved here that the zeros of $\dot{P}_n(x,c)$ are simple and even more interlaced with the zeros of $P_n(x,c)$.

Theorem 6.1. Let $x_i = x_i(c)$ denote the zeros of the monic orthogonal polynomial $P_n(x,c)$ in decreasing order

$$x_1(c) > x_2(c) > \cdots > x_n(c).$$

Let us assume that the zeros $x_i(c)$ are all strictly increasing or all strictly decreasing functions of c. Then, the zeros of $\dot{P}_n(x,c) = (\partial/\partial c)P_n(x,c)$ and $P_n(x,c)$ interlace.

Proof. Let us write $P_n(x,c)$ in the factorized form

$$P_n(x,c) = \prod_{i=1}^n (x - x_i(c)).$$

The logarithmic dot derivative gives

$$\dot{P}_n(x,c) = -P_n(x,c) \sum_{i=1}^n \frac{\dot{x}_i(c)}{x - x_i(c)} = -\sum_{i=1}^n \left(\dot{x}_i(c) \prod_{j=1, j \neq i}^n (x - x_j(c)) \right)$$

and therefore

$$\dot{P}_n(x_i(c), c) = -\dot{x}_i(c) \prod_{j=1, j \neq i}^n (x_i(c) - x_j(c))$$

and $\operatorname{sgn}(\dot{P}_n(x_i(c), c)) = (-1)^{n+1-i} \operatorname{sgn}(\dot{x}_i(c)).$

If now for all $i, x_i(c)$ are monotonic function of c, we check easily the following ratio:

$$\frac{\dot{P}_n(x_{i+1}(c), c)}{\dot{P}_n(x_i(c), c)} < 0,$$

which proves the interlacing properties of the zeros of $\dot{P}_n(x,c)$ with respect to the zeros of $P_n(x,c)$ in the range of c corresponding to the orthogonality of $P_n(x,c)$. \Box

6.1. Discussion on the monotonicity assumptions of $x_i(c)$

In the Jacobi case with $\alpha > -1$, $\beta > -1$ we know from [22, p. 121] that for all *i*, $\dot{x}_i(c) < 0$ for $c = \alpha (\neq \beta)$ and $\dot{x}_i(c) > 0$ for $c = \beta (\neq \alpha)$; in the Laguerre case $\alpha > -1$ implies $\dot{x}_i(c) > 0$ for $c = \alpha$.

There are many other situations for which $\dot{x}_i(c)$ increases or decreases strictly for all *i* as a function of *c*. Let us mention [12] that the zeros of the *r*th associated Laguerre polynomial $L_n^{(\alpha)}(x;r)$ satisfy $\dot{x}_i(\alpha) > 0$ for $\alpha \ge 0$, r > -1.

For classical discrete orthogonal polynomials: Charlier, Meixner, Kravchuk and Hahn [18], the Markov theorem [22, p. 115] can be applied too and insures monotonicity of the zeros with respect to most of the parameters (we do not consider integer parameters fixing the length of the support of the measure).

The Markov theorem is based on the monotonicity in x of $\dot{\varrho}(x,c)/\varrho(x,c)$ which is

(i) For Charlier polynomials $C_n^{(\mu)}(x)$, $0 \le x < \infty$, $\varrho(x,\mu) = \mu^x / \Gamma(x+1), \ \mu > 0$,

$$\frac{\dot{\varrho}(x,c)}{\varrho(x,c)} = \frac{x}{\mu} \quad (c = \mu).$$

(ii) For Meixner polynomials $M_n^{(\gamma,\mu)}(x)$, $0 \le x < \infty$, $\varrho(x,\gamma,\mu) = \mu^x \Gamma(x+\gamma) / \Gamma(x+1)$, $0 < \mu < 1$, $\gamma > 0$,

$$\frac{\dot{\varrho}(x,c)}{\varrho(x,c)} = \begin{cases} \frac{x}{\mu} & \text{if } c = \mu, \\ \psi(x+\gamma) & \text{if } c = \gamma. \end{cases}$$

(iii) For Hahn polynomials $H_n^{(\alpha,\beta)}(x;N)$, $0 \le x < N$, $\varrho(x,\alpha,\beta) = \Gamma(\alpha+N-x)\Gamma(\beta+x+1)/(\Gamma(x+1)\Gamma(N-x))$, $\alpha > -1$, $\beta > -1$, $N \in \mathbb{N}$,

$$\frac{\dot{\varrho}(x,c)}{\varrho(x,c)} = \begin{cases} \psi(N+\alpha-x) & \text{if } c = \alpha, \\ \psi(\beta+1-x) & \text{if } c = \beta. \end{cases}$$

(iv) For Kravchuk polynomials $K_n^{(p)}(x;N)$, $0 \le x \le N$, $\varrho(x,p) = {N \choose x} p^x (1-p)^{N-x}$, $0 , <math>N \in \mathbb{N}$,

$$\frac{\dot{\varrho}(x,c)}{\varrho(x,c)} = \frac{x-pN}{p(1-p)} \quad (c=p).$$

The derivative

$$\frac{d}{dx} \left(\frac{\dot{\varrho}(x,c)}{\varrho(x,c)} \right) \tag{6.1}$$

is obviously positive in the case of Charlier ($c = \mu > 0$). For Meixner, with $c = \mu > 0$ the above derivative is again positive; for Meixner with $c = \gamma > 0$, $\psi'(z)$ is positive in the full range of x [1, (6.4.10)].

In the case of Hahn polynomials, derivative (6.1) is negative for $c = \alpha$ and again negative for $c = \beta$.

For Kravchuk polynomials (6.1) is always positive since 0 .

Remark 2. (i) (a) We emphasize the importance of normalization of the polynomial $P_n(x,c)$. If instead to use monic polynomials we introduce a normalization factor $N_n(c)$, the dot derivative of $\overline{P}_n(x,c) = N_n(c)P_n(x,c)$ gives

$$\bar{P}_n(x,c) = N_n(c)\dot{P}_n(x,c) + \dot{N}_n(c)P_n(x,c).$$

Of course, for each *c*-dependent normalization we get a new family of $\dot{P}_n(x,c)$ but each $\dot{P}_n(x,c)$ alternate again with the $\bar{P}_n(x,c)$ with the assumption that now $N_n(c)\dot{x}_i(c)$ is a strictly monotonic function of *c* for every *i*.

(b) For monic $P_n(x,c)$ or nonmonic $\overline{P}_n(x,c)$ with the normalization factor independent of c, $\dot{P}_n(x,c)$ is obviously of degree n-1. It stays of degree n in all other situations.

(ii) (a) For symmetric weight on symmetric interval the assumption $\dot{x}_i(c)$ monotonic for each *i* cannot be fulfilled, positive and negative roots moving in opposite direction.

(b) Parity argument in this symmetric case reduces the degree of $\dot{P}_n(x,c)$ for monic polynomial $P_n(x,c)$ to n-2.

111

(iii) If the zeros of monic $\dot{P}_n(x,c)$ are again strictly increasing or strictly decreasing functions of c, by using the same arguments as in the previous theorem, we obtain an interlacing property between zeros of monic $\dot{P}_n(x,c)$ and zeros of $\ddot{P}_n(x,c)$. It is an open problem to know what new hypothesis insure that the n-1 zeros of $\dot{P}_n(x,c)$ are also monotonic.

Final remarks. This elementary recursive approach gives a representation of infinitesimal or finite parameter's changes in classical orthogonal polynomials. Only a few results are given but more sophisticated situations with arbitrary D, Δ or D_q derivatives can be obtained in the same way.

In the case of semi-classical orthogonal polynomials like generalized Hermite or generalized Gegenbauer, the derivative representation of these polynomials does not exist, but a more general version of the Navima algorithm [20] allows again to build the recurrence relation for the connection coefficients between the $P_n^{[r]}(x,c)$ and $P_n(x,c)$ polynomials.

It is easy to generate a fourth-order D, Δ or D_q equation satisfied by $\dot{P}_n(x,c)$ building first the second-order corresponding operator annihilating $R_1^{(D,\dot{D})}$, $R_1^{(\Delta,\dot{D})}$, and $R_1^{(D_q,\dot{D})}$ in relations (2.4)–(2.6). This (factorized) fourth order equation could be used for connection problems between $\dot{P}_n(x,c)$ and an orthogonal family $\{Q_m(x)\}$, and for the linearization of $\dot{P}_i(x,c)\dot{P}_j(x,c)$ in terms of an orthogonal family $\{Q_m(x)\}$.

References

- [1] M. Abramowitz, I. Stegun, Handbook of Mathematical Functions, Dover, New York, 1965.
- [2] G.E. Andrews, R. Askey, R. Roy, Special Functions, Encyclopedia of Mathematics and its Applications, vol. 71, Cambridge University Press, Cambridge, 1999.
- [3] I. Area, E. Godoy, A. Ronveaux, A. Zarzo, Minimal recurrence relations for connection coefficients between classical orthogonal polynomials: discrete case, J. Comput. Appl. Math. 89 (1998) 309–325.
- [4] I. Area, E. Godoy, A. Ronveaux, A. Zarzo, Inversion problems in the q-Hahn tableau, J. Symbolic Comput. 28 (6) (1999) 767–776.
- [5] P. Deuflhard, M. Wulkow, Computational treatment of polyreaction kinetics by orthogonal polynomials of a discrete variable, Impact Comput. Sci. Eng. 1 (1989) 269–301.
- [6] J. Froehlich, Parameter derivatives of the Jacobi polynomials and Gaussian hypergeometric function, Integral Transforms Special Functions 2 (1994) 252–266.
- [7] B. Gerard, L. Roberts, Graphical discovery of a new identity for Jacobi polynomials, Amer. Math. Monthly (1998) 163–166.
- [8] E. Godoy, A. Ronveaux, A. Zarzo, I. Area, Minimal recurrence relations for connection coefficients between classical orthogonal polynomials: continuous case, J. Comput. Appl. Math. 84 (1997) 257–275.
- [9] E. Godoy, A. Ronveaux, A. Zarzo, I. Area, Connection problems in non-homogeneous differential and difference equations, J. Comput. Appl. Math. 99 (1–2) (1998) 177–187.
- [10] I.S. Gradshteyn, I.M. Ryzhik, Table of Integrals, Series and Products, Academic Press, London, 1980.
- [11] W. Hahn, Über Orthogonalpolynome, die q-differencengleichungen genügen, Math. Nachr. 2 (1949) 4–34.
- [12] M.E.H. Ismail, M.E. Muldoon, A discrete approach to monotonicity of zeros of orthogonal polynomials, Trans. Amer. Math. Soc. 323 (1) (1991) 65–78.
- [13] R. Koekoek, R.F. Swarttouw, The Askey-scheme of hypergeometric orthogonal polynomials and its q-analogue, Report 98-17, Fac. Techn. Math. Informatics, Delft University of Technology, Delft, 1998.
- [14] W. Koepf, Identities for families of orthogonal polynomials and special functions, Integral Transforms Special Functions 6 (1997) 69–102.
- [15] W. Koepf, D. Schmersau, Representations of orthogonal polynomials, J. Comput. Appl. Math. 90 (1998) 57-94.

- [16] T. Koornwinder, Compact quantum groups and q-special functions, in: V. Baldoni, M.A. Picardello (Eds.), Representations of Lie Groups and Quantum Groups, Pitman Research Notes in Mathematics Series, Vol. 311, Longman Scientific & Technical, 1994, pp. 46–128.
- [17] A. Nikiforov, V. Ouvarov, Éléments de la théorie des fonctions spéciales, Mir, Moscou, 1976.
- [18] A.F. Nikiforov, S.K. Suslov, V.B. Uvarov, Classical Orthogonal Polynomials of a Discrete Variable, Springer, Berlin, 1991.
- [19] A.P. Prudnikov, Yu.A. Brychkov, O.I. Marichev, Integrals and Series, 3 Volumes, Gordon and Breach, New York, 1986.
- [20] A. Ronveaux, A. Zarzo, E. Godoy, Recurrence relation for connection coefficients between two families of orthogonal polynomials, J. Comput. Appl. Math. 62 (1995) 67–73.
- [21] A. Ronveaux, S. Belmehdi, E. Godoy, A. Zarzo, Recurrence relation approach for connection coefficients. Application to classical discrete orthogonal polynomials, in: D. Levi, L. Vinet, P. Winternitz (Eds.), Proceedings of the International Workshop on Symmetries and Integrability of Difference Equations, CRM Proceedings and Lecture Note Series, Vol. 9, AMS, Providence, RI, 1996, pp. 321–337.
- [22] G. Szegö, Orthogonal Polynomials, A.M.S. Colloquium Publications, Vol. XXIII, AMS, Providence, RI, 1978.
- [23] Wolfram Research, Inc., Mathematica Version 3, Wolfram Research Inc., Champaign, Illinois, 1996.
- [24] A. Zarzo, I. Area, E. Godoy, A. Ronveaux, Results for some inversion problems for classical continuous and discrete orthogonal polynomials, J. Phys. A: Math. Gen. 30 (1997) L35–L40.



Journal of Computational and Applied Mathematics 121 (2000) 113-124

JOURNAL OF COMPUTATIONAL AND APPLIED MATHEMATICS

www.elsevier.nl/locate/cam

Computing integral transforms and solving integral equations using Chebyshev polynomial approximations

Robert Piessens

Dept. Computerwetenschappen, Katholieke Universiteit, Leuven, Celestijnenlaan 200 A, 3001 Leuven, Belgium

Received 10 February 2000

Abstract

This paper gives a survey of the use of Chebyshev polynomials in the computation and the inversion of integral transforms and in the solution of integral equations, especially with singular kernels. © 2000 Elsevier Science B.V. All rights reserved.

Keywords: Chebyshev polynomial; Integral equation; Integral transform; Singularity

1. Introduction

Pafnuty Chebyshev was born in 1821 in Russia. His early research was devoted to number theory. He defended his doctoral thesis 'Teoria sravneny' (Theory of congruences) in 1849. In 1850 he became extraordinary and in 1860 full professor of Mathematics at Petersburg University. This was the start of intensive research work in various fields. Besides research in probability which resulted in a generalization of the law of large numbers and a corresponding generalization of the central limit theorem of De Moivre and Laplace, he began his remarkable studies on the theory of mechanisms.

He studied the so-called Watt-parallelogram, a hinge mechanism employed in steam engines for transforming a rotating into a rectilinear movement. Since it is impossible to obtain strictly rectilinear movement by hinge mechanisms, Chebyshev elaborated a sound theory to reduce the deviation of

E-mail address: robert.piessens@cs.kuleuven.ac.be (R. Piessens).

^{0377-0427/00}/\$ - see front matter © 2000 Elsevier Science B.V. All rights reserved. PII: S0377-0427(00)00349-6

the resultant movement from the rectilinear. This problem is closely related to the theory of best approximations of functions.

The paper 'Théorie des mécanismes connus sous le nom de parallélogrammes' (1854) was first in a series of works in this area. In this paper, Chebyshev determined the polynomial of the *n*th degree with leading coefficient equal to unity which deviates least from zero on the interval [-1, 1]. This polynomial is

$$\frac{1}{2^{n-1}}\cos(n\arccos x) = \frac{1}{2^{n-1}}T_n(x).$$

The polynomials $T_n(x)$, named after Chebyshev, form an orthogonal system on [-1, 1] with respect to the weight function $(1 - x^2)^{-1/2}$.

The application of Chebyshev polynomials in numerical analysis starts with a paper of Lanczos [9] in 1938. The coming of the digital computer gave further emphasis to this development. From the middle of the 20th century, the numerical analysis literature abounds with papers on approximation of functions, computation of integrals and solution of differential equations, using Chebyshev polynomials. We mention especially the work of Lanczos [10,11], Clenshaw [1–3], Luke [12–15], and the handbooks of Fox and Parker [6] and Rivlin [25].

Chebyshev polynomials play also an important role in network synthesis, especially for the construction of equal ripple approximations to ideal low-pass filters.

In this paper we give a survey of the use of Chebyshev polynomials in the numerical computation of integral transforms and the solution of integral equations. We focus on problems showing singularity.

2. Properties of the Chebyshev polynomials

The popularity of Chebyshev polynomials in numerical analysis is due to a lot of important but simple properties:

- (i) The already mentioned property of least deviation from zero and the continuous and discrete orthogonality property.
- (ii) The recurrence relation

$$T_{n+1}(x) = 2xT_n(x) - T_{n-1}(x).$$
(1)

(iii) The differential equation

$$(1+x^2)T''(x) - xT'_n(x) + n^2T_n(x) = 0.$$
(2)

(iv) The difference-differential relation

$$(1 - x^{2})T'_{n}(x) = n(T_{n-1}(x) - xT_{n}(x))$$

= $\frac{n}{2}(T_{n-1}(x) - T_{n+1}(x)).$ (3)

(v) The expression for the indefinite integral

$$\int T_n(x) \, \mathrm{d}x = \frac{1}{2} \left(\frac{T_{n+1}(x)}{n+1} - \frac{T_{n-1}(x)}{n-1} \right), \quad n \ge 1.$$
(4)

A consequence is

$$\int_{-1}^{+1} T_n(x) dx = 0 \quad \text{when } n \text{ is odd}$$
$$= -\frac{1}{n^2 - 1} \quad \text{when } n \text{ is even,}$$

on which the Clenshaw-Curtis integration [4] method is based.

A very interesting property of the Chebyshev polynomials is that, in a relatively easy way, linear recurrence relations can be constructed for the computation of the so-called modified moments

$$I_n = \int_{-1}^{+1} \omega(x) T_n(x) \,\mathrm{d}x$$

for a lot of weight functions $\omega(x)$. In [18] the following weight functions are considered:

$$w_{1}(x) = (1 - x)^{\alpha}(1 + x)^{\beta},$$

$$w_{2}(x) = (1 - x)^{\alpha}(1 + x)^{\beta} \exp(-ax),$$

$$w_{3}(x) = (1 - x)^{\alpha}(1 + x)^{\beta} \ln((1 + x)/2) \exp(-ax),$$

$$w_{4}(x) = \exp(-ax^{2}),$$

$$w_{5}(x) = (1 - x)^{\alpha}(1 + x)^{\beta} \exp(-a(x + 1)^{2}),$$

$$w_{6}(x) = (1 - x)^{\alpha}(1 + x)^{\beta} \exp(-a/(x + 1)),$$

$$w_{7}(x) = (1 - x)^{\alpha}(1 + x)^{\beta} \exp(-a/(x + 1)^{2},$$

$$w_{8}(x) = (1 - x)^{\alpha}(1 + x)^{\beta} \ln((1 + x)/2),$$

$$w_{10}(x) = (1 - x)^{\alpha}(1 + x)^{\beta} \ln((1 + x)/2) \ln((1 - x)/2),$$

$$w_{11}(x) = |x - a|^{\alpha},$$

$$w_{12}(x) = |x - a|^{\alpha} \sin(x - a),$$

$$w_{13}(x) = |x - a|^{\alpha} \ln|x - a|,$$

$$w_{14}(x) = |x - a|^{\alpha} \ln|x - a|,$$

$$w_{15}(x) = (1 - x)^{\alpha}(1 + x)^{\beta}|x - a|^{\gamma},$$

$$w_{16}(x) = (1 - x)^{\alpha}(1 + x)^{\beta}|x - a|^{\gamma} \ln|x - a|,$$

$$w_{17}(x) = [(x - b)^{2} + a^{2}]^{\alpha},$$

$$w_{18}(x) = (1 + x)^{\alpha}J_{\nu}(a(x + 1)/2).$$

The modified moments have application in the construction of Gaussian quadrature formulas [7,21] and in the numerical computation of integrals with strongly oscillating, peaked or singular integrand [22].

In the next sections we will discuss the solution of integral equations and the computation of integral transforms using modified moments.

(vi) The expression as a hypergeometric function

$$T_n(x) = F\left(-n, n; \frac{1}{2}; \frac{1-x}{2}\right),$$
(5)

where F is the hypergeometric function. From this expression we derive the following results:

$$\mathscr{L}\{x^{\alpha}T_{n}(1-2x)\} = \frac{\Gamma(\alpha+1)}{p^{\alpha+1}} {}_{3}F_{1}\left(\frac{-n,n,\alpha+1}{1/2};\frac{1}{p}\right)$$
(6)

and

$$\mathscr{L}^{-1}\left\{p^{-\alpha}T_n\left(1-\frac{a}{p}\right)\right\} = \frac{t^{\alpha-1}}{\Gamma(\alpha)} {}_2F_2\left(\frac{-n,n}{1/2,\alpha};\frac{at}{2}\right),\tag{7}$$

where \mathscr{L} and \mathscr{L}^{-1} denote Laplace and inverse Laplace transform, α and *a* are real parameters and $_kF_l$ is the generalised hypergeometric function.

(vii) The relatively easy formulae for constructing near least-squares approximations of a function f(x):

$$f(x) \simeq \sum_{k=0}^{n} c_k T_k(x), \tag{8}$$

$$c_k = \frac{2}{N} \sum_{l=0}^{N} {}^{\prime\prime} f(x_l) T_l(x_k), \tag{9}$$

$$x_k = \cos\left(\frac{k\pi}{N}\right) \quad n \leqslant N,\tag{10}$$

where the prime denotes that the first term is taken with factor $\frac{1}{2}$, and where the double prime denotes that first and last term are taken with factor $\frac{1}{2}$.

The Chebyshev coefficients c_k can be evaluated using an efficient and numerically stable algorithm, based on FFT-techniques [8].

(viii) Chebyshev polynomials are members of larger families of orthogonal polynomials. (Jacobi polynomials and ultraspherical polynomials.) In many practical cases, the Chebyshev series expansion of a function is the best between all expansion into ultraspherical polynomials.

3. Solution of integral equations of the second kind using modified moments

We consider

$$\phi(x) = f(x) - \int_{-1}^{1} k(x, y)\phi(y) \,\mathrm{d}y, \tag{11}$$

where ϕ is the function to be determined. The kernel function k and the function f are given. We assume that $-1 \le x \le 1$. The use of Chebyshev polynomials for the numerical solution of such equations has been discussed in [5,6,26]. The use of modified moments is proposed in [19]. The solution $\phi(x)$ of (11) is approximated by

$$p(x) = \omega(x) \sum_{k=0}^{N} c_k T_k(x),$$
(12)

where the coefficients c_k are to be determined. If it is known that $\phi(x)$ shows a singular behaviour, the singularities can be catched in the function $\omega(x)$.

Substituting (12) into (11) we have

$$\sum_{k=0}^{N} c_k[\omega(x)T_k(x) + I_k(x)] = f(x),$$
(13)

where

$$I_k(x) = \int_{-1}^{+1} k(x, y)\omega(y)T_k(y) \,\mathrm{d}y.$$
(14)

Substituting at least N + 1 values of x into (13) yields a system of linear equations, the solution of which gives approximate values of the Chebyshev coefficients c_k .

In many practical cases, efficient evaluation of $I_k(x)$ is possible due to recurrence relations for modified moments. As an example, we consider Love's integral equation

$$\phi(x) = 1 \pm \frac{1}{\pi} \int_{-1}^{+1} \frac{a}{a^2 + (x - y)^2} \phi(y) \, \mathrm{d}y \tag{15}$$

the solution of which is the field of two equal circular coaxial conducting disks, separated by a distance *a* and on equal or opposite potential, with zero potential at infinity. We choose $\omega(x) \equiv 1$. The method of solution requires the evaluation of

$$I_k(x) = \int_{-1}^{+1} \frac{a}{a^2 + (x - y)^2} T_k(y) \,\mathrm{d}y.$$
(16)

When a is small, the kernel function $a/(a^2 + (x - y)^2)$ shows a strongly peaked behaviour, which is an handicap for numerical integration.

The recurrence relation, however,

$$I_{k+2}(x) - 4xI_{k+1}(x) + (2 + 4a^2 + 4x^2)I_k(x) - 4xI_{k-1}(x) + I_{k-2}$$

= $(4a/(1-k^2))[1 + (-1)^k]$ (17)

allows efficient computation.

Starting values are

$$I_0(x) = \arctan\left(\frac{1-x}{a}\right) + \arctan\left(\frac{1+x}{a}\right),$$

$$I_1(x) = xI_0(x) + \frac{a}{2}\ln\frac{(1-x)^2 + a^2}{(1+x)^2 + a^2},$$

$$I_2(x) = 4xI_1(x) - (2a^2 + 2x^2 + 1)I_0(x) + 4a,$$

$$I_3(3) = -(4a^2 - 12x^2 + 3)I_1(x) - 8x(a^2 + x^2)I_0(x) + 16xa.$$

Forward recursion of (17) is not completely numerically stable, but the stability is sufficient for practical purposes.

4. Solution of the Abel integral equation

The Abel integral equation

$$\int_0^x \phi(x)(x-y)^{-\alpha} \, \mathrm{d}y = f(x) \quad (0 < \alpha < 1),$$
(18)

occurs in a number of engineering problems.

If f(x) is differentiable, the solution of (18) is explicitly given by

$$\phi(x) = \frac{\sin(\alpha \pi)}{\pi} \left[\frac{f(0)}{x^{1-\alpha}} + \int_0^x \frac{f'(y)}{(x-y)^{1-\alpha}} \, \mathrm{d}y \right].$$
(19)

However, this formula is not of practical value in problems where no explicit mathematical expression for f(x) is known. In the case that f(x) is obtainable only from measured data, Chenshaw's curve fitting method [2] can be used to construct an approximation in the form

$$f(x) \simeq x^{\beta} \sum_{k=0}^{N} c_k T_k (1-2x),$$
 (20)

where $\beta > -\alpha$ is essentially a free parameter, which can be used to optimize the approximation, taking into account a singular behaviour for $x \to 0$.

The approximate solution of (18) is now [23]

$$\phi(x) \simeq \frac{x^{\alpha+\beta-1}}{\Gamma(\alpha+\beta)} \frac{\Gamma(1+\beta)}{\Gamma(1-\alpha)} \sum_{n=0}^{N} c_n f_n(x),$$
(21)

where

$$f_n(x) = {}_3F_2 \begin{pmatrix} -n, n, \beta+1 \\ 1/2, \alpha+\beta \end{pmatrix}.$$
(22)

Using Fasenmyer's technique [24], a recurrence formula for the computation of $f_n(x)$ can be derived, namely

$$f_n(x) + (A_n + B_n x) f_{n-1}(x) + (C_n + D_n x) f_{n-2}(x) + E_n f_{n-3}(x) = 0,$$
(23)

where

$$A_{n} = -\frac{1}{n-2} \left[n - 3 + \frac{(n-1)(2n-3)}{n+\alpha+\beta-1} \right],$$

$$B_{n} = 4 \frac{n+\beta}{n+\alpha+\beta-1},$$

,

$$C_{n} = \frac{1}{n-2} \left[-1 + \frac{n-1}{n+\alpha+\beta-1} (3n-\alpha-\beta-5) \right]$$
$$D_{n} = -4 \frac{(n-\beta-3)(n-1)}{(n+\alpha+\beta-1)(n-2)},$$
$$E_{n} = -\frac{(n-\alpha-\beta-2)(n-1)}{(n+\alpha+\beta-1)(n-2)}.$$

Starting values for (23) are

$$f_0(x) = 1,$$

$$f_1(x) = 1 - \frac{2(\beta + 1)}{\alpha + \beta} x,$$

$$f_2(x) = 1 - \frac{8(\beta + 1)}{\alpha + \beta} x + \frac{8(\beta + 1)(\beta + 2)}{(\alpha + \beta)(\alpha + \beta + 1)} x^2.$$

The recurrence formula (23) is a difference equation of Poincaré's type.

Forward recursion is numerically stable.

5. The computation of Laplace, Fourier and Hankel transforms

The Laplace transforms of f is defined as

$$\mathscr{L}\lbrace f\rbrace = F(s) = \int_0^\infty e^{-sx} f(x) \,\mathrm{d}x.$$
(24)

We approximate f(x) on $[0,\infty)$ by

$$f(x) \simeq (1+x)^{-\alpha} \sum_{k=0}^{N} a_k T_k^{\star} \left(\frac{1}{1+x}\right),$$
 (25)

where T_k^{\star} is the shifted Chebyshev polynomial of degree k and where $\alpha > 0$ is a real parameter, which can be choosen freely, although its value affects strongly the quality of the approximation. The coefficients a_k are computed as the Chebyshev series coefficients of

$$g(z) = \left(\frac{z+1}{2}\right)^{-\alpha} f\left(\frac{1-z}{1+z}\right).$$

An approximation of F(s) is now given by

$$F(s) \simeq \sum_{k=0}^{N} a_k I_k(\alpha, s), \tag{26}$$

where

$$I_k(\alpha, s) = 2^{1-\alpha} e^{-s} \int_{-1}^{+1} (x+1)^{\alpha-2} e^{2s/(x+1)} T_k(x) \, \mathrm{d}x.$$
(27)

Here again, the modified moments in (27) satisfy a linear recurrence relation [18]

$$-(k + \alpha + 1)I_{k+1} + 2(2s - k - 2)I_{k+1} + 2(\alpha - 3 - 4s)I_k + 2(2s + k - 2)I_{k-1} + (k - \alpha - 1)I_{k-2} = 0.$$
(28)

In (26) and (27) s may be replaced by $j\omega$, so that the formulae are applicable for the computation of Fourier integrals. Starting values for the recurrence relations and numerical stability are discussed in [20].

The Hankel transform of f(x) is defined as

$$\mathscr{H}_{v}\lbrace f\rbrace = F_{v}(s) = \int_{0}^{\infty} xf(x)J_{v}(sx)\,\mathrm{d}x,\tag{29}$$

where $J_{\nu}(x)$ is the Bessel function of the first kind and order ν . The inversion formula is, when $\nu > -\frac{1}{2}$:

$$f(x) = \mathscr{H}_{v}^{-1}\{F_{v}(s)\} = \int_{0}^{\infty} sF_{v}(s)J_{v}(sx) \,\mathrm{d}s.$$
(30)

Both direct and inverse transform are integrals of the form

$$I(s) = \int_0^\infty \varphi(x) J_\nu(sx) \,\mathrm{d}x,\tag{31}$$

which are difficult to compute numerically. However, if $\varphi(x)$ is rapidly decaying to zero, the infinite integration range may be truncated to a finite interval [0, A]. We have then

$$I(s) \simeq A \int_0^1 \varphi(Ax) J_{\nu}(\omega) \,\mathrm{d}x,\tag{32}$$

where $\omega = sA$.

Here the approximation on [0,1]

$$\varphi(Ax) \simeq x^{\alpha} \sum_{k=0}^{N} c_k T_k^{\star}(x)$$
(33)

yields

$$I(s) \simeq A \sum_{k=0}^{N} c_k M_k(\omega, \nu, \alpha), \tag{34}$$

where

$$M_k(\omega, \nu, \alpha) = \int_0^1 x^{\alpha} J_{\nu}(\omega x) T_k^{\star}(x) \,\mathrm{d}x.$$
(35)

These modified moments satisfy the following homogeneous, linear, nine-term recurrence relation:

$$\frac{\omega^2}{16}M_{k+4} + \left[(k+3)(k+3+2\alpha) + \alpha^2 - v^2 - \frac{\omega^2}{4} \right] M_{k+2} + \left[4(v^2 - \alpha^2) - 2(k+2)(2\alpha - 1) \right] M_{k+1}$$

$$-\left[2(k^{2}-4)+6(v^{2}-\alpha^{2})-2(2\alpha-1)-\frac{3\omega^{2}}{8}\right]M_{k}$$

$$+\left[4(v^{2}-\alpha^{2})+2(k-2)(2\alpha-1)\right]M_{k-1}$$

$$+\left[(k-3)(k-3-2\alpha)+\left(\alpha^{2}-v^{2}-\frac{\omega^{2}}{4}\right)\right]M_{k-2}+\frac{\omega^{2}}{16}M_{k-4}=0.$$
(36)

Because of the symmetry of the recurrence relation of the shifted Chebyshev polynomials, it is convenient to define

$$T^{\star}_{-k}(x) = T^{\star}_{k}(x), \quad k = 1, 2, 3, \dots$$

and consequently

$$M_{-k}(\omega, \nu, \alpha) = M_k(\omega, \nu, \alpha).$$

To start the recurrence relation with k = 0, 1, 2, 3, ... we need only M_0, M_1, M_2 , and M_3 . Using the explicit expressions of the shifted Chebyshev polynomials we obtain

$$\begin{split} M_0 &= G(\omega, \nu, \alpha), \\ M_1 &= 2G(\omega, \nu, \alpha + 1) - G(\omega, \nu, \alpha), \\ M_2 &= 8G(\omega, \nu, \alpha + 2) - 8G(\omega, \nu, \alpha + 1) + G(\omega, \nu, \alpha), \\ M_3 &= 32G(\omega, \nu, \alpha + 3) - 48G(\omega, \nu, \alpha + 2) + 18G(\omega, \nu, \alpha + 1) - G(\omega, \nu, \alpha), \end{split}$$

where

$$G(\omega, \nu, \alpha) = \int_0^1 x^{\alpha} J_{\nu}(\omega x) \, \mathrm{d}x.$$

Because

$$\omega^2 G(\omega, \nu, \alpha + 2) = [\nu^2 - (\alpha + 1)^2] G(\omega, \nu, \alpha) + (\alpha + \nu + 1) J_{\nu}(\omega) - \omega J_{\nu-1}(\omega),$$

we need only $G(\omega, v, \alpha)$ and $G(\omega, v, \alpha + 1)$. Luke [12] has given the following formulas:

1. a Neumann series expansion that is suitable for small ω

$$G(\omega, \nu, \alpha) = \frac{2}{\omega(\alpha + \nu + 1)} \sum_{k=0}^{\infty} \frac{(\nu + 2k + 1)((\nu - \alpha + 1)/2)}{((\nu + \alpha + 3)/2)_k} J_{\nu + 2k+1}(\omega),$$

2. an asymptotic expansion that is suitable for large ω

$$G(\omega, \nu, \alpha) = \frac{2^{\alpha}}{\omega^{\alpha+1}} \frac{\Gamma((\nu + \alpha + 1)/2)}{\Gamma((\nu - \alpha + 1)/2)} - \sqrt{\frac{2}{\pi\omega^3}} (g_1 \cos \theta + g_2 \sin \theta),$$

where

 $\theta = \omega - v\pi/2 + \pi/4$

and

$$g_{1} \sim \sum_{k=0}^{\infty} (-1)^{k} a_{2k} \omega^{-2k}, \quad \omega \to \infty,$$

$$g_{2} \sim \sum_{k=0}^{\infty} (-1)^{k} a_{2k+1} \omega^{-2k-1}, \quad \omega \to \infty,$$

$$a_{k} = \frac{(1/2 - \nu)_{k} (1/2 + \nu)_{k}}{2^{k} k!} b_{k},$$

$$b_{0} = 1,$$

$$b_{k+1} = 1 + \frac{2(k+1)(\alpha - k - 1/2)}{(\nu - k - 1/2)(\nu + k + 1/2)} b_{k}.$$

If α and v are integers, the following formulas are useful [12]:

$$\int_{0}^{1} J_{2\nu}(\omega x) \, \mathrm{d}x = \int_{0}^{1} J_{0}(\omega x) \, \mathrm{d}x - \frac{2}{\omega} \sum_{k=0}^{\nu-1} J_{2k+1}(\omega),$$
$$\int_{0}^{1} J_{2\nu+1}(\omega x) \, \mathrm{d}x = \frac{1 - J_{0}(\omega)}{\omega} - \frac{2}{\omega} \sum_{k=1}^{\nu} J_{2k}(\omega).$$

For the evaluation of

$$\int_0^1 J_0(\omega x) \,\mathrm{d}x$$

Chebyshev series approximations are given in [13].

Forward and backward recursion are asymptotically unstable. However, the instability of forward recursion is less pronounced if $k \leq \omega/2$. Indeed, practical experiments demonstrate that $M_k(\omega, \nu, \alpha)$ can be computed accurately using forward recursion for $k \leq \omega/2$. For $k > \omega/2$ the loss of significant figures increases and forward recursion is no longer applicable. In that case, Oliver's algorithm [16] has to be used. This means that (36) has to be solved as a boundary value problem with six initial values and two end values. The solution of this boundary value problem requires the solution of a linear system of equations having a band structure.

6. Inversion of the Laplace transform

The main difficulty in applying Laplace transform techniques is the determination of the original function f(t) from its transform F(p). In many cases, numerical methods must be used. The computation of f(t) from values of F(p) on the real axis is not well-posed, so that regularization is recommended. Inverting the approximation

$$F(p) \simeq p^{-\alpha} \sum_{n=0}^{N} c_n T_n \left(1 - \frac{b}{p} \right)$$
(37)

yields

$$f(t) \simeq \frac{t^{\alpha - 1}}{\Gamma(\alpha)} \sum_{n=0}^{N} c_n \varphi_n\left(\frac{bt}{2}\right),\tag{38}$$

where

$$\varphi_n(x) = {}_2F_2\left(\begin{array}{c} -n,n\\ 1/2,\alpha\end{array};x\right). \tag{39}$$

Here, $\varphi_n(x)$ is a polynomial of degree *n* which satisfies the recurrence formulae [17]

$$\varphi_n + (A + Bx)\varphi_{n-1} + (C + Dx)\varphi_{n-2} + E\varphi_{n-3} = 0, \tag{40}$$

n = 3, 4, ..., where

$$A = -\frac{\alpha n + 3n^2 - 9n - 3\alpha + 6}{(n-2)(\alpha + n - 1)},$$

$$B = \frac{4}{\alpha + n - 1},$$

$$C = \frac{n(3n - 9 - \alpha) + 6}{(n-2)(\alpha + n - 1)},$$

$$D = 4\frac{n - 1}{(n-2)(\alpha + n - 1)},$$

$$E = \frac{(n-1)(n - \alpha - 2)}{(n-2)(\alpha + n - 1)},$$

$$\varphi_0(x) = 1,$$

$$\varphi_0(x) = 1,$$

$$\varphi_1(x) = 1 - \frac{2x}{\alpha},$$

$$\varphi_2(x) = 1 - \frac{8x}{\alpha} + \frac{8x^2}{\alpha(\alpha + 1)}.$$

The polynomial $\varphi_n(x)$ has *n* real positive zeros. This means that the interval $[0, \infty)$ can be divided into an oscillation interval, in which lie the zeros and an interval in which the polynomial increases monotonically. In the oscillation interval, $\varphi_n(x)$ oscillates with strongly increasing amplitude. In evaluating expression (38), this fact produce some difficulty, because, for large values of *t*, the errors on the coefficients c_n are multiplied by a large number, especially for large *n*. Regularization consists in restricting the value of *N* in (37).

7. Conclusion

Chebyshev polynomials in numerical analysis are associated with the approximation of smooth functions. In this paper we have shown that the domain of applicability is much wider, going from the computation of integrals with singular integrand to the solution of integral equations and the inversion of integral transforms.

References

- [1] C.W. Clenshaw, A note on the summation of Chebyshev series, Math. Tab. Wash 9 (1955) 118-120.
- [2] C.W. Clenshaw, Curve fitting with a digital computer, Comput. J. 2 (1960) 170–173.
- [3] C.W. Clenshaw, Chebyshev Series for Mathematical Functions, Math. Tab. Nat. Phip. Lab., Vol. 5, H.M. Stationery Office, London, 1962.
- [4] C.W. Clenshaw, A.R. Curtis, A method for numerical integration on an automatic computer, Numer. Math. 2 (1960) 197–205.
- [5] D. Elliott, A Chebyshev series method for the numerical solution of Fredholm integral equations, Comput. J. 6 (1961) 102–111.
- [6] L. Fox, I.B. Parker, Chebyshev Polynomials in Numerical Analysis, Oxford University Press, London, 1968.
- [7] W. Gautschi, On the construction of Gaussian rules from modified moments, Math. Comput. 24 (1970) 245-260.
- [8] W.M. Gentleman, Algorithm 424, Clenshaw-Curtis quadrature, Comm. ACM 15 (1972) 353-355.
- [9] C. Lanczos, Trigonometric interpolation of empirical and analytic functions, J. Math. Phys. 17 (1938) 123-199.
- [10] C. Lanczos, Tables of Chebyshev Polynomials, Applicable Mathematics Series US Bureau Standards, Vol. 9, Government Printing Office, Washington, 1952.
- [11] C. Lanczos, Applied Analysis, Prentice-Hall, New York, 1957.
- [12] Y.L. Luke, Integrals of Bessel functions, McGraw-Hill Book Co, New York, 1962.
- [13] Y.L. Luke, The Special Functions and their Approximations, Vols. 1 & 2, Academic Press, New York, 1969.
- [14] Y.L. Luke, Mathematical Functions and their Approximations, Academic Press, New York, 1975.
- [15] Y.L. Luke, Algorithms for the Computation of Mathematical Functions, Academic Press, New York, 1977.
- [16] J. Oliver, The numerical solution of linear recurrence relations, Numer. Math. 11 (1968) 349-360.
- [17] R. Piessens, A new numerical method for the inversion of the Laplace transform, J. Inst. Math. Appl. 10 (1972) 185–192.
- [18] R. Piessens, Modified Chenshaw–Curtis integration and applications to numerical computation of integral transforms, in: P. Keast, G. Fairweather (Eds.), Numerical Integrations: Recent Developments, Software and Applications, Reidel, Dordrecht, 1987.
- [19] R. Piessens, M. Branders, Numerical solution of integral equations of mathematical physics, using Chebyshev polynomials, J. Comput. Phys. 21 (1976) 178–196.
- [20] R. Piessens, M. Branders, Computation of Fourier transform integrals using Chebyshev series expansions, Computing 32 (1984) 177–186.
- [21] R. Piessens, M. Chawla, N. Jayarajan, J. Comput. Phys. 21 (1976) 356-360.
- [22] R. Piesssens, E. de Doncker-Kapenga, C.W. Überhuber, D.K. Kahaner, Quadpack: A Subroutine Package for Automatic Integration, Springer, Berlin, 1983.
- [23] R. Piessens, P. Verbaeten, Numerical solution of the Abel integral equation, BIT 13 (1973) 451-457.
- [24] E.D. Rainville, Special Functions, Macmillan, New York, 1960.
- [25] T. Rivlin, Chebyshev Polynomials: From Approximation Theory to Algebra and Number Theory, Wiley, New York, 1990.
- [26] R.E. Scraton, Math. Comput. 23 (1969) 837-844.



Journal of Computational and Applied Mathematics 121 (2000) 125-152

JOURNAL OF COMPUTATIONAL AND APPLIED MATHEMATICS

www.elsevier.nl/locate/cam

Developments in bivariate spline interpolation

G. Nürnberger^{a,*}, F. Zeilfelder^b

^aFakultät für Mathematik und Informatik, Universität Mannheim, LS Mathematik IV, D-68131, Mannheim,

Germany

^bMax-Planck-Institut für Informatik, Im Stadtwald, D-66123 Saarbrücken, Germany

Received 22 January 2000

Abstract

The aim of this survey is to describe developments in the field of interpolation by bivariate splines. We summarize results on the dimension and the approximation order of bivariate spline spaces, and describe interpolation methods for these spaces. Moreover, numerical examples are given. © 2000 Elsevier Science B.V. All rights reserved.

MSC: 41A05; 41A15; 41A25; 65D05; 65D07; 65D17; 41A63

Keywords: Bivariate splines; Interpolation; Dimension; Approximation order; Finite elements; Macro elements; Bézier–Bernstein techniques; Triangulation

1. Introduction

Bivariate spline spaces have been studied intensively in the past 30 years. These spaces consist of piecewise polynomials defined on a triangulation of a polygonal domain. They are of considerable interest in approximation theory and numerical analysis, in particular, in scattered data fitting, the construction and reconstruction of surfaces in fields of application and, classically, in the numerical solution of boundary-value problems by finite-element-type methods.

The aim of this survey is to describe interpolation methods for bivariate splines (including numerical examples) and to summarize related results on the dimension and the approximation order of bivariate spline spaces. In contrast to the univariate case, even standard problems such as the dimension and the approximation order of bivariate spline spaces are difficult to solve. In particular, the construction of explicit interpolation schemes (especially Lagrange interpolation schemes) for spline spaces on given triangulations leads to complex problems.

* Corresponding author. Tel.: +49-621-181-25-00; fax: +49-621-181-24-98.

E-mail addresses: nuern@euklid.math.uni-mannheim.de (G. Nürnberger), zeilfeld@mpi-sb.mpg.de (F. Zeilfelder).

The paper is organized as follows. In Section 2, we give some results on the Bézier–Bernstein representation of bivariate polynomials and its relation to bivariate splines. Section 3 deals with the dimension of bivariate spline spaces. First, general lower and upper bounds for the dimension are known in the literature. Moreover, the dimension was determined for arbitrary triangulations if the degree is relatively large compared with the smoothness of the splines. In addition, dimension formulas were derived for general classes of triangulations in the case when the degree is near the smoothness.

In Section 4, we study the approximation order of bivariate spline spaces. For arbitrary triangulations, the approximation order was determined if the degree is sufficiently large compared with the smoothness. In the other case, the approximation order was given for general classes of triangulations. The results were proved by using interpolation and quasi-interpolation methods.

In Section 5, we discuss finite element and macro element methods. Classical finite element methods are based on Hermite interpolation by bivariate polynomials on each triangle of a given triangulation, where the triangles have to be subdivided in the case of low degree polynomials. The polynomials tied together lead to super splines. Macro element methods are generalizations of the finite element methods and also lead to supersplines. Unfortunately, the Hermite interpolation schemes cannot be transformed in a straightforward way into Lagrange interpolation schemes on the whole triangulation. On the other hand, this can be done when spline spaces are used.

Section 6 deals with Hermite and Lagrange interpolation methods for bivariate spline spaces. In contrast to the univariate case, Schoenberg–Whitney-type conditions do not characterize interpolation but the so-called almost interpolation by bivariate spline spaces. The construction of explicit interpolation schemes for bivariate spline spaces leads to complex problems in general. Concerning numerical purposes, it is desirable that the complexity of computing an interpolation spline is linear in the number of triangles. Interpolation methods of this type are known in the literature, where numerical examples with more than 100 000 interpolation points are given.

Such interpolation methods can be used for the construction and reconstruction of surfaces. Concerning scattered data fitting problems, the function values and derivatives which are needed for spline interpolation, respectively, for splines of finite element type can be computed approximately by using local methods. In this context, it is the advantage of Lagrange interpolation that only functional values (and no (orthogonal) derivatives) have to be determined approximately.

In Section 6, we summarize interpolation methods for bivariate spline space (which yield explicit interpolation schemes) of the following type: Interpolation by spline spaces of arbitrary degree and smoothness on uniform-type partitions; the construction of triangulations which are suitable for interpolation by spline spaces; interpolation by spline spaces (of higher degree) on arbitrary triangulations; interpolation by spline spaces (of low degree) on classes of triangulations, respectively quadrangulations; respectively quadrangulations. In addition, we discuss the approximation order of these methods.

We finally note that despite the great progress which has been made in the vast literature, several deep problems concerning spline spaces are still unsolved.

We also note that many papers on so-called *multivariate simplex splines* and *multivariate box splines* exist in the literature. Concerning these investigations, we refer to the survey of Dahmen and Michelli [48], the books by Bojanov et al. [20], by de Boor et al. [25] and by Michelli [111].

2. Spline spaces and Bézier–Bernstein techniques

Let Δ be a regular triangulation of a simply connected polygonal domain Ω in \mathbb{R}^2 , i.e., a set of closed triangles such that the intersection of any two triangles is empty, a common edge or a vertex. Following Alfeld, Piper and Schumaker [7], we set

 $V_{\rm I}$ = number of interior vertices of Δ ,

 $V_{\rm B}$ = number of boundary vertices of Δ ,

V =total number of vertices of Δ ,

 $E_{\rm I}$ = number of interior edges of Δ ,

 $E_{\rm B}$ = number of boundary edges of Δ ,

 $E = \text{total number of edges of } \Delta$,

N = number of triangles of Δ .

It is well known that the following Euler formulas hold:

$$E_{\rm B} = V_{\rm B},$$

$$E_{\rm I} = 3V_{\rm I} + V_{\rm B} - 3,$$

$$N = 2V_{\rm I} + V_{\rm B} - 2.$$

In the following, we define spline spaces which are natural generalizations of the classical *univariate spline spaces* (cf. the books by de Boor [21], by Nürnberger [116] and by Schumaker [144]) i.e., spaces of splines in one variable.

For given integers $r,q,0 \le r < q$, the space of *bivariate splines* of degree q and the smoothness r with respect to Δ is defined by

$$S_a^r(\varDelta) = \{ s \in C^r(\Omega) : s |_T \in \Pi_q, \ T \in \varDelta \},\$$

where

$$\Pi_q = \operatorname{span}\{x^i y^j : i, j \ge 0, i+j \le q\}$$

is the space of *bivariate polynomials* of total degree q. In addition, suppose ρ_i , i = 1, ..., V, are integers satisfying $r \leq \rho_i < q$, i = 1, ..., V, and let $\theta = (\rho_1, ..., \rho_V)$. The space of *bivariate super splines* with respect to Δ is defined by

$$S_q^{r,\theta}(\varDelta) = \{s \in S_q^r(\varDelta) \colon s \in C^{\rho_i}(v_i), \ i = 1, \dots, V\}.$$

Obviously, superspline spaces are subspaces of $S_q^r(\Delta)$.

In this survey, we consider the problem of constructing interpolation sets for bivariate spline spaces \mathscr{S} , where \mathscr{S} can be the space $S_a^r(\Delta)$ as well as a superspline space $S_a^{r,\theta}(\Delta)$.

A set $\{z_1, \ldots, z_d\}$ in Ω , where $d = \dim \mathcal{S}$, is called a *Lagrange interpolation set* for \mathcal{S} if for each function $f \in C(\Omega)$, a unique spline $s \in \mathcal{S}$ exists such that

$$s(z_i) = f(z_i), \quad i = 1, \dots, d.$$



Fig. 1. A degenerate edge e attached to the vertex v_1 .

If also partial derivatives of a sufficiently differentiable function f are involved and the total number of Hermite conditions is d, then we speak of a *Hermite interpolation set* for \mathcal{S} .

In investigating interpolation by bivariate splines, the following representation of the polynomial pieces of a spline is important. Given a triangle $T \in \Delta$ with vertices v_1, v_2, v_3 , the polynomial piece $p^T = s|_T \in \Pi_q$ of a given spline $s \in S_q^r(\Delta)$ can be written as

$$p^{T}(x,y) = \sum_{i+j+k=q} a^{T}_{i,j,k} \frac{q!}{i!j!k!} \Phi^{i}_{1}(x,y) \Phi^{j}_{2}(x,y) \Phi^{k}_{3}(x,y), \quad (x,y) \in T,$$
(1)

where the so-called *barycentric coordinates* $\Phi_{\mu} \in \Pi_1$, $\mu = 1, 2, 3$, are uniquely defined by $\Phi_{\mu}(v_{\nu}) = \delta_{\mu,\nu}$, $\nu = 1, 2, 3$. Representation (1) is called the *Bézier–Bernstein form* of p^T and the real numbers $a_{i,j,k}^T$ are called the *Bézier–Bernstein coefficients* of p^T .

The following result is important for investigating the structure of bivariate spline spaces. This theorem was given by Farin [69] and de Boor [22] (see also [29,47,80]) and characterizes smoothness conditions of polynomial pieces p^{T_l} , l = 1, 2, in representation (1) on adjacent triangles T_1, T_2 with vertices v_1, v_2, v_3 and v_1, v_2, v_4 , respectively.

Theorem 2.1. Let *s* be a piecewise polynomial function of degree *q* defined on $T_1 \cup T_2$. Then $s \in S_q^r(\{T_1, T_2\})$ iff for all $\rho \in \{0, ..., r\}$:

$$a_{i,j,\rho}^{T_2} = \sum_{i_1+j_1+k_1=\rho} a_{i+i_1,j+j_1,k_1}^{T_1} \frac{\rho!}{i_1!j_1!k_1!} \Phi_1^{i_1}(v_4) \Phi_2^{j_1}(v_4) \Phi_3^{k_1}(v_4), \quad i+j=q-\rho.$$

It is well known (cf. [29,69]) that for r = 1 the smoothness conditions of Theorem 2.1 have the geometric interpretation that the corresponding Bézier-Bernstein points lie in the same plane. Moreover, if the edge $e = [v_1, v_2]$ is *degenerate at* v_1 (i.e., the edges with vertex v_1 adjacent to e lie on a line, see Fig. 1), then for r = 1 the geometric interpretation of these smoothness conditions is that this plane degenerates to a line that contains three of the corresponding Bézier-Bernstein points. For $r \ge 2$, similar effects appear for degenerate edges. We note that degenerate edges can lead to complex problems in the investigation of bivariate spline spaces.

The following result, given in [22,69] (see also [35,36,125,159]), expresses the relation between the partial derivatives of a polynomial p^{T} in representation (1) at a vertex and its Bézier–Bernstein coefficients. This lemma plays an important role in the construction of interpolation sets for bivariate splines.

Lemma 2.2. Let $p^T \in \Pi_q$ be a polynomial on a triangle $T = \Delta(v_1, v_2, v_3)$ in representation (1) and d_j , j = 1, 2, be unit vectors in direction of the edge $[v_1, v_{j+1}]$, j = 1, 2. Then, for all $0 \le \alpha + \beta \le q$,

$$p_{d_1^{\alpha}d_2^{\beta}}^T(v_1) = \frac{q!}{(q-\alpha-\beta)!} \sum_{j=0}^{\alpha} \sum_{k=0}^{\beta} \binom{\alpha}{j} \binom{\beta}{k} (\Phi_1)_{d_1}^{\alpha-j} (\Phi_1)_{d_2}^{\beta-k} (\Phi_2)_{d_1}^j (\Phi_3)_{d_2}^k a_{q-j-k,j,k}.$$
 (2)

It easily follows from (2) and induction that if the Bézier-Bernstein coefficients

$$a_{q-j-k,j,k}, \quad j = 0, ..., \alpha, \ k = 0, ..., \beta$$

are uniquely determined, then all derivatives

$$p_{d_{j}^{j}d_{2}^{k}}^{T}(v_{1}), \quad j = 0, \dots, \alpha, \ k = 0, \dots, \beta$$

are given, and conversely. In particular, these relations show that Bézier–Bernstein coefficients can be considered as certain linear functionals (see Section 3).

The Bézier-Bernstein representation (1) of a polynomial $p^T \in \Pi_q$ has important applications in CAGD. We refer the reader to the surveys of Farin [69,70], Boehm et al. [19], the book by Chui [29] and the papers of de Boor [22] and Dahmen [47]. Moreover, concerning the so-called *blossoming approach*, we refer to the tutorial of de Rose et al. [135] and the survey of Seidel [149]. Triangulation methods were described in the survey of Schumaker [147]. For interpolation by bivariate polynomials, we refer to the survey of Gasca and Sauer [74].

Finally, we note that a characterization (different from Theorem 2.1) of the smoothness of polynomial pieces on adjacent triangles, without using Bézier–Bernstein techniques, was proved by Davydov et al. [59].

3. Dimension of spline spaces

In this section, we summarize results on the dimension of bivariate (super) spline spaces for arbitrary triangulations. Results on the dimension play a fundamental role for the construction of interpolation sets.

In contrast to univariate spline theory, it is a nontrivial problem to determine the dimension of bivariate spline spaces when $r \ge 1$. (For the case of continuous splines, i.e., r = 0, see [143].) In fact, this problem is not yet completely solved.

Historically, Strang [156] was the first in 1973 who posed the problem: What is the dimension of $S_a^r(\Delta)$? In the following we summarize results concerning this problem.

We begin with the following lower bound on the dimension, which was given in 1979 by Schumaker [143].

Theorem 3.1. Let Δ be an arbitrary triangulation and e_i be the number of edges with different slopes attached to the ith interior vertex of Δ . Set

$$\sigma_i = \sum_{j=1}^{q-r} (r+j+1-je_i)_+, \quad i=1,\ldots,V_{\rm I},$$



Fig. 2. The Morgan and Scott triangulation Δ_{MS} .

where $(x)_{+} = \max\{0, x\}$. Then

$$\dim S_q^r(\varDelta) \ge \binom{q+2}{2} + \binom{q-r+1}{2} E_I - \binom{q+2}{2} - \binom{r+2}{2} V_I + \sum_{i=1}^{V_I} \sigma_i.$$
(3)

In [145], it was shown that the lower bound in (3) also holds for spline spaces with respect to partitions more general than triangulations, the so-called *rectilinear partitions*. A lower bound for the dimension of spline spaces with respect to rectilinear partitions with holes was given by Jia [90].

A standard method to determine the exact dimension is to find an upper bound for the dimension of a given spline space that coincides with the lower bound in (3). In order to establish such an upper bound *n*, it suffices to construct linear functionals λ_i , i = 1, ..., n, such that

if
$$\lambda_i(s) = 0$$
, $s \in S_a^r(\Delta)$, $i = 1, \dots, n$, then $s \equiv 0$. (4)

In [145], an upper bound for the dimension of spline spaces was developed (see also [107]). This upper bound depends on the ordering of the interior vertices. An improved upper bound was given by Ripmeester [134], who used a special ordering of the vertices. We also mention that lower and upper bounds for the dimension hold for superspline spaces and that such bounds for spline spaces in several variables were developed by Alfeld [5]. In general, however, all these upper bounds do not coincide with the lower bound in (3).

The dimension of a bivariate spline space can be larger than the lower bound in (3). Examples concerning this fact were given in [145], where the first example constructed by Morgan and Scott [113] is also discussed (see also [4,76,152]). Morgan and Scott considered triangulations Δ_{MS} as in Fig. 2 and the space $S_2^1(\Delta_{MS})$. The dimension of this spline space is equal to 7 if Δ_{MS} has certain symmetry properties, and otherwise it is 6, which is equal to the lower bound in (3) for $S_2^1(\Delta_{MS})$. Hence, in the nonsymmetric case $S_2^1(\Delta_{MS})$ coincides with Π_2 . This example shows that the dimension can depend on the exact geometry of the whole triangulation. In general, such dependencies may appear if the degree q is small compared with the smoothness r. Diener [66] investigated the dimension of the space $S_{2r}^r(\Delta_{MS})$, $r \ge 2$, and found similar dependencies on the exact geometry of Δ_{MS} .

These results show that the structure of $S_q^r(\Delta)$ is becoming more complex when the degree q approaches the smoothness r. This is one of the fundamental phenomena in bivariate spline theory, in contrast to the univariate case.

We proceed by describing cases when the dimension of $S_q^r(\Delta)$ is known. In 1975, Morgan and Scott [112] determined the dimension of $S_q^1(\Delta)$, $q \ge 5$. Without using Bézier–Bernstein techniques,



Fig. 3. A singular vertex v.

they showed that the following formula holds for an arbitrary triangulation Δ :

$$\dim S_q^1(\varDelta) = \begin{pmatrix} q+2\\ 2 \end{pmatrix} N - (2q+1)E_1 + 3V_1 + \sigma, \quad q \ge 5.$$
(5)

Here and in the following, σ denotes the number of singular vertices of Δ . An interior vertex of Δ that has only two edges with different slopes attached to it is called a *singular vertex* (see Fig. 3). For C^1 -spline spaces, $\sigma_i = 1$ in (3) holds if the corresponding vertex is singular, and in all other cases $\sigma_i = 0$. By using this fact and Euler's formulas (see Section 2) it is easy to verify that the dimension formula (5) of Morgan and Scott is equal to the lower bound in (3). Moreover, in [112] a *nodal basis* for C^1 -spline spaces of degree at least 5 was constructed. This means that the splines in $S_q^1(\Delta)$, $q \ge 5$, are determined by their values and derivatives at d points in Ω , where $d = \dim S_q^1(\Delta)$. Davydov [53] showed that an alternative construction yields a basis for these spline spaces which is *locally linearly independent* (see Section 6).

The results of Morgan and Scott were extended to spline spaces $S_q^r(\Delta)$, $q \ge 4r+1$. These extensions are based on the results of Schumaker [146] for spline spaces on *cells* (see Fig. 4) coupled with the methods developed by Alfeld and Schumaker [8] and Alfeld et al. [6] (see also [27]). (For a generalization to trivariate splines of degree at least 8r + 1 on tetrahedral partitions, see [11].) In these papers, Bézier–Bernstein techniques were used and the concept of *minimal determining sets* was introduced. Roughly speaking, the Bézier–Bernstein coefficients $a_{i,j,k}^T$, i+j+k=q, $T \in \Delta$, of the polynomial pieces $p^T \in \Pi_q$ in representation (1) of a spline can be considered as linear functionals. A *determining set* is a subset $\{\lambda_i: i = 1, ..., n\}$ of these functionals such that (4) holds, and such a set is called *minimal* if there exists no determining set with fewer elements.

By using Bézier–Bernstein techniques Hong [80] determined the dimension of $S_q^r(\Delta)$ for arbitrary triangulations Δ in the case when $q \ge 3r + 2$.

Theorem 3.2. Let Δ be an arbitrary triangulation. If $q \ge 3r + 2$, then the dimension of $S_q^r(\Delta)$ is equal to the lower bound in (3).

This result was generalized by Ibrahim and Schumaker [82] to super spline spaces of degree at least 3r + 2 (see also [36]). The proof of Theorem 3.2 given by Hong [80] (respectively Ibrahim et al. [82]) is based on local arguments by considering vertices, edges and triangles separately. In particular, a basis of *star-supported* splines (i.e. splines whose supports are at most the set of triangles surrounding a vertex, i.e., a cell (for interior vertices)) of $S_q^r(\Delta)$, $q \ge 3r + 2$, is constructed (see also [59]). Recently, Alfeld and Schumaker [10] showed that such a basis does not exist in general if q < 3r + 2 and $r \ge 1$. Thus, local arguments as in [36,59,80,82] fail in these cases. The problem of finding an explicit formula for the dimension of $S_q^r(\Delta)$, q < 3r + 2, $r \ge 1$, remains open in general.

The only exception known in the literature is the space $S_4^1(\Delta)$. In 1987, Alfeld et al. [7] showed the following result.

Theorem 3.3. Let Δ be an arbitrary triangulation. Then

$$\dim S_4^1(\varDelta) = 6V + \sigma - 3. \tag{6}$$

Again, the number in (6) is equal to the lower bound in (3). The proof of Alfeld et al. [7] involves arguments from graph theory which are not purely local. As shown in [7], such complex arguments do not have to be used if Δ is a *nondegenerate triangulation* (i.e., a triangulation that contains no degenerate edges, see Section 2). For this class of triangulations Δ , the dimension of $S_{3r+1}^r(\Delta)$, $r \ge 2$, has been determined by Alfeld and Schumaker [9].

By Euler's formulas (see Section 2) the lower bound in (3) denoted by lb_3^1 for $S_3^1(\Delta)$ can be written as follows:

$$1b_3^1 = 3V_B + 2V_I + \sigma + 1.$$
(7)

Therefore, the dimension of $S_3^1(\Delta)$ is larger than the number of vertices of Δ . This is in contrast to the case of quadratic C^1 -splines (where the lower bound in (3) is equal to $V_B + \sigma + 3$, see also the general comment on these spaces in Section 5) and makes $S_3^1(\Delta)$ interesting for applications. On the other hand, the structure of the space $S_3^1(\Delta)$ is very complex. For instance, it is not known if the following conjecture is true for arbitrary triangulations Δ .

Conjecture. The dimension of $S_3^1(\Delta)$ is equal to lb_3^1 .

More general, the following problem for $r \ge 1$ is unsolved: What is the smallest integer $q \le 3r + 2$ (depending on r) such that the dimension of $S_q^r(\Delta)$ coincides with the lower bound in (3) for an arbitrary triangulation Δ ?

By using homological methods Billera [17] (see also [18,90,160]) showed that the above conjecture holds *generically*. Roughly speaking, this means that if the dimension of $S_3^1(\Delta)$ is not equal to lb_3^1 , then pertubations of the vertices exist such that equality holds.

In addition, the above conjecture holds for general classes of triangulations. In connection with their interpolation method, Davydov et al. [58] (see also [56,57]) proved that lb_3^1 is equal to the dimension of $S_3^1(\Delta)$, where Δ is contained in the general class of *nested polygon triangulations* (see Section 6).

Finally, Gmelig Meyling [75] discussed a numerical algorithm for determining the dimension of $S_3^1(\Delta)$. This is done by computing the rank of a global system associated with the smoothness constraints.

4. Approximation order of spline spaces

In this section, we summarize results on the *approximation order* of bivariate spline spaces. Such results are important for interpolation by these spaces. We say that $S_a^r(\Delta)$ has approximation order



Fig. 4. A cell.



Fig. 5. Δ^i , i = 1, 2, triangulations.

 ρ , if for any sufficiently differentiable function f

$$\operatorname{dist}(f, S_q^r(\Delta)) \leqslant Kh^{\rho}, \tag{8}$$

where $h = \max\{\text{diam}(T): T \in \Delta\}$, K is a constant depending on the smallest angle in Δ and f, and the distance is measured in the supremum norm. Moreover, we say that the approximation order is *optimal*, if $\rho = q + 1$.

It is clear that the approximation order of $S_q^r(\Delta)$ cannot be greater than q+1, and that the optimal approximation order of these spaces is obtained for r=0. In the context of the *finite element method*, it follows from a result of Ženišek [162] (see Section 5) that $S_q^r(\Delta)$ has optimal approximation order provided that $q \ge 4r + 1$, $r \ge 0$.

However, in contrast to the univariate case, the approximation order of a bivariate spline space is not always optimal. This was first proved in 1983 by de Boor and Höllig [23], who considered the space $S_3^1(\Delta^1)$. Here, Δ^1 is a triangulation obtained from a rectangular partition by adding the same diagonal to each rectangle (see Fig. 5). Extensions of this result, were given by Jia [87–89]. In 1993, de Boor and Jia [26] proved the following theorem.

Theorem 4.1. The approximation order of $S_q^r(\Delta^1)$ is at most q if q < 3r + 2, $r \ge 1$.

As shown in [26], this result holds for any L_p , $1 \le p \le \infty$, norm.

In general, for fixed q and r the approximation order of $S_q^r(\Delta)$ depends on the structure of the triangulation Δ . For instance, it was proved by Jia [91] that $S_q^r(\Delta^2)$, $q \ge 2r + 1$, $r \ge 0$, has optimal approximation order. Here, Δ^2 is a triangulation obtained from a rectangular partition by adding both diagonals to each quadrangle (see Fig. 5).

More general, *triangulated quadrangulations* were considered in the literature. These are triangulations obtained from a quadrangulation \Box as follows: for each convex quadrilateral \mathcal{Q} of \Box both diagonals are added and for each nonconvex quadrilateral \mathcal{Q} of \Box the diagonal is added and the center of this diagonal is connected with the remaining two vertices of \mathcal{Q} . A triangulated quadrangulation is called a *triangulated convex quadrangulation*, if every quadrilateral of the corresponding quadrangulation is convex.

Lai and Schumaker [101] proved optimal approximation order for $S_6^2(\Delta)$, where Δ is an arbitrary triangulated quadrangulation. For doing this, a *quasi-interpolation method* for $S_6^2(\Delta)$ was developed, which uses a superspline subspace $S_6^{2,\theta}(\Delta)$, with $\rho_i \in \{2,3\}$, in general. We note that by this method supersplines have only to be used if interior vertices of degree three or certain interior vertices of degree four appear in \Box . These are the only vertices, where smoothness three is needed.

Generally, given a basis $\{B_i, i = 1, ..., d\}$, of a spline space and a set of linear functionals $\{\lambda_i: i = 1, ..., d\}$, a quasi-interpolant s_f of a sufficiently smooth function f from this spline space can be written in the form

$$s_f = \sum_{i=1}^{a} (\lambda_i f) B_i.$$
⁽⁹⁾

These linear functionals λ_i typically consist of linear combinations of certain derivatives at points in Ω . In particular, Bézier–Bernstein coefficients (see Sections 2 and 3) can be considered as linear functionals of this type.

Moreover, Lai and Schumaker [103] proved optimal approximation order for $S_{3r}^r(\Delta)$, $r \ge 1$, where Δ is a triangulated convex quadrangulation. Again, this result was shown by using quasi-interpolation for a superspline subspace in general. We note that this quasi-interpolation method lead to an extension of Fraeijs de Veubeke's [71], and Sander's [140] finite element (see Section 5).

Concerning arbitrary triangulations Δ , the following theorem was first stated by de Boor and Höllig [24] in 1988, and later proved completely by Chui et al. [34].

Theorem 4.2. The approximation order of $S_q^r(\Delta)$ with respect to an arbitrary triangulation Δ is optimal when $q \ge 3r + 2$, $r \ge 1$.

While de Boor and Höllig [24] used abstract methods, Chui et al. [34] constructed a quasi-interpolant for a superspline subspace to prove this result.

By using a quasi-interpolation method different from Chui et al. [34], Lai and Schumaker [102] extended Theorem 4.2 to general L_p , $1 \le p \le \infty$, norms. These two quasi-interpolation methods have in common that both use an appropriate superspline subspace and Bézier–Bernstein methods. Recently, for $q \ge 3r + 2$, the first interpolation scheme which yields optimal approximation order (in the sense that the constant K only depends on the smallest angle of Δ and is independent of *near-degenerate* edges) such that the fundamental functions have minimal support was developed by Davydov et al. [59]. By using methods which are completely different from those for quasi-interpolation and by applying the concept of *weak interpolation* (introduced by Nürnberger [118]), the authors constructed a Hermite interpolating spline from the super spline space $S_q^{r,\theta^*}(\Delta)$ where $\theta^* = (r + [(r + 1)/2], \dots, r + [(r + 1)/2])$ and a nodal basis which yield these properties. The following theorem shows that the interpolating spline s_f of Davydov et al. [59] simultaneously



Fig. 6. The Argyis, Fried, and Scharpf element, $S_5^{1,\theta_1}(\Delta)$.

approximates the function f and its derivatives. We note that the quasi-interpolation methods of Lai and Schumaker [102,103] yield similar estimates.

Theorem 4.3. Let $q \ge 3r + 2$, $r \ge 1$, Δ an arbitrary triangulation, $f \in C^m(\Omega)$, $m \in \{2r, \ldots, q+1\}$ and $s_f \in S_q^{r,\theta^*}(\Delta)$ the interpolating spline of f. Then, for every triangle $T \in \Delta$,

$$\|D_x^{\alpha} D_y^{\beta} (f - s_f)\|_{L_{\infty}(T)} \leq K h_T^{m-\alpha-\beta} \max\{\|D_x^{\mu} D_y^{m-\mu} f\|_{C(T)} \colon \mu = 0, \dots, m\}$$
(10)

for all α , $\beta \ge 0$, $\alpha + \beta \le m$, where h_T is the diameter of T, and K is a constant which depends only on r, q and the smallest angle in Δ .

We close this section with the following problem.

Problem. Determine the approximation order of $S_q^r(\Delta)$, q < 3r + 2, $r \ge 1$, for general classes of triangulations Δ .

5. Finite elements and macro element methods

In this section, classical *finite elements* and *macro element* methods are considered. Classical finite elements are based either on interpolation by bivariate polynomials on every triangle or on interpolation by C^1 splines with respect to a triangulation that is obtained by applying a splitting procedure to *every* triangle or quadrilateral. Extensions of the latter case lead to macro elements.

We begin by describing finite elements, where each polynomial piece is determined separately. The idea of this classical method is to chose a suitable spline space such that interpolation by bivariate polynomials on every triangle of an arbitrary triangulation Δ automatically leads to a Hermite interpolation set for this spline space. Such spline spaces are superspline spaces of large degree.

As an example, we describe the well-known finite element of Argyis et al. [12], which yields Hermite interpolation by the super spline space $S_5^{1,\theta_1}(\Delta)$, where $\theta_1 = (2,...,2)$. This Hermite interpolation method is to interpolate function value, first and second-order derivatives at the vertices, and the normal derivative at the midpoint of each edge. (See Fig. 6, where the function value, vertex derivatives and normal derivatives are symbolized by circles and boxes, respectively). The corresponding *condensed scheme* is obtained by replacing the normal derivatives by other conditions, see [14,16]. This method was generalized by Ženišek [162,163] to Hermite interpolation by $S_a^{r,\theta_r}(\Delta), q \ge 4r + 1, r \ge 1$, where $\theta_r = (2r,...,2r)$. As mentioned in Section 4, the corresponding



Fig. 7. The Fraeijs de Veubeke, respectively Sander element, $S_3^1(\Delta)$.



Fig. 8. The Clough–Tocher element, $S_3^1(\Delta_{CT})$.

Hermite interpolating spline yields optimal approximation order. This follows from a result of Ciarlet and Raviart [43] concerning interpolation by bivariate polynomials. A link between the finite element method, super splines, and Bézier–Bernstein techniques was given by Schumaker [148] (see also [159], for the special case $S_9^{2,\theta_2}(\Delta)$). For further results on finite elements see [109,110] and the references therein.

It is desirable in general, however, to use low degree splines (in relation to the smoothness) to keep computational costs small. The following classical methods have been developed for this purpose. The idea of these methods is to modify the given partition, which can be a triangulation or a convex quadrangulation. This is done by applying a certain splitting procedure to each triangle or quadrilateral. In contrast to the finite element method described above, more than one polynomial piece is needed for each triangle or quadrilateral such that the method is local. These classical approaches lead to Hermite interpolation by cubic and quadratic C^1 splines.

We begin with the classical Hermite interpolation scheme for $S_3^1(\Delta)$ of Fraeijs de Veubeke [71] and Sander [140], where Δ is a triangulated convex quadrangulation (see Section 4). This classical Hermite interpolation set consists of the function and gradient value at the vertices of the underlying convex quadrangulation \Box and the orthogonal derivative at the midpoints of all edges of \Box . (See Fig. 7, where the function value, vertex derivatives and normal derivatives are symbolized by circles and boxes, respectively.) The approximation properties of this element were studied by Ciavaldini and Nèdélec [44]. It turns out that the corresponding Hermite interpolating spline yields optimal approximation order. We note that a modification of this Hermite interpolation set involving second-order derivatives (instead of the orthogonal derivatives) was given by Lai [99].

Another well-known element was given by Clough and Tocher [45] in 1966. These authors constructed a Hermite interpolation set for $S_3^1(\Delta_{\rm CT})$, where $\Delta_{\rm CT}$ is a triangulation obtained from an arbitrary triangulation Δ by splitting each triangle $T \in \Delta$ into three subtriangles (see Fig. 8). This classical Hermite interpolation set consists of function and gradient value at the vertices of Δ and



Fig. 9. The Powell-Sabin element, $S_2^1(\Delta_{PS})$.

the orthogonal derivative at the midpoints of all edges of Δ . (See Fig. 8, where the function value, the vertex derivatives and normal derivatives are symbolized by circles and boxes, respectively.) The approximation properties of this element and its condensed version were studied by Ciarlet [40,42] and Farin [68]. It turns out that the Hermite interpolation set described above yields optimal approximation order. Moreover, Percell [129] considered $S_4^1(\Delta_{\rm CT})$ and Alfeld [3] developed a trivariate Clough–Tocher-type method.

We turn now to the case of quadratic C^1 splines, $S_2^1(\Delta)$, and begin with a general comment on these spaces. As we mentioned in Section 3, the lower bound in (3) for these spline spaces, is equal to $V_B + \sigma + 3$. Since there are many cases, when the dimension of $S_2^1(\Delta)$ is close to this lower bound, it cannot be expected that these spaces have good approximation properties in general. In particular, for $S_2^1(\Delta^1)$. By considering this lower bound and having in mind the example of Morgan and Scott [113] (see Section 3), the only practical way to obtain a quadratic C^1 space in which interpolation yields good approximation order is to increase the number of singular and boundary vertices (cf. the method of Nürnberger and Zeilfelder [125] described in Section 6).

In 1977, Powell and Sabin [131] constructed a Hermite interpolation set for $S_2^1(\Delta_{PS})$. In that work, the triangulations Δ_{PS} are obtained by splitting every triangle *T* of a given triangulation Δ into six subtriangles (see Fig. 9). We note that the splitting points are chosen in such a way that each interior edge of Δ leads to a singular vertex of Δ_{PS} and that each boundary edge of Δ leads to an additional boundary vertex of Δ_{PS} . By using Euler's formulas (see Section 2) it can be easily seen that the dimension of the resulting space $S_2^1(\Delta_{PS})$ is 3V. The Hermite interpolation set constructed in [131] consists of function and gradient value for all vertices of Δ . (See Fig. 9 where the function values and vertex derivatives are symbolized by circles.) The corresponding Hermite interpolating spline of Powell and Sabin yields optimal approximation order. This was proved by Sablonnière [139] with special attention to the approximation constants. For further results using this space and a modification of this, see [51,67,79,130].

In the following, we describe extensions of these methods. These extensions are called macro element methods. Macro element methods use Bézier–Bernstein techniques and lead to Hermite interpolation by super spline spaces.

We start with the generalization of Fraeijs de Veubeke's and Sander's method, which was developed by Laghchim-Lahlou and Sablonnière [93,94,96]. In [93,94,96], the following cases were considered, where Δ is a triangulated convex quadrangulation: $S_{3r}^{r,\theta}(\Delta)$ if r is odd, $S_{3r+1}^{r,\theta}(\Delta)$ if ris even. Here, the components of θ concerning the vertices of the underlying quadrangulation are (3r-1)/2 if r is odd and 3r/2 if r is even. This Hermite interpolation method is to interpolate function value and derivatives up to order r + [r/2] at the vertices and suitable derivatives at interior points of each edge of the underlying quadrangulation. We note that if r is odd, the super spline spaces considered by Lai and Schumaker [103] (see Section 4) coincide with those of [96]. On the other hand, in [103] different supersplines with lower degree were used for even r.

Generalizations of the Clough–Tocher element were also given in the literature. In 1994, Laghchim-Lahlou et al. [95] (see also [92,94,137]) constructed Hermite interpolation sets for certain superspline space with respect to $\Delta_{\rm CT}$. Again, these sets consist of function value and derivatives up to a certain order at the vertices and suitable derivatives at interior points of each edge of the given (nonsplitted) triangulation. Lai and Schumaker [104] recently constructed a Hermite interpolation set for super spline spaces with respect to $\Delta_{\rm CT}$ that additionally contain function value and derivatives up to a certain order at each splitting point (except for the classical case $S_3^1(\Delta)$). These authors show that their construction provides a so-called *stable local basis*, which implies that the associated spline space has optimal approximation order. We mention that such bases have been constructed for $S_q^r(\Delta)$, $q \ge 3r+2$, by Chui et al. [34], by Lai and Schumaker [102] and by Davydov and Schumaker [61]. Finally, generalizations of the Powell–Sabin element were given in the literature. In 1996, Laghchim-Lahlou and Sablonnière [97] (see also [94]) considered the triangulation. There it is shown that the function value and derivatives up to order r + [r/2] at all vertices of Δ^1 yield Hermite interpolation by the superspline spaces $S_{2r}^{r,\theta}(\Delta_{\rm PS}^{\rm h})$ if r is even, where

 $\theta = (r + [r/2], ..., r + [r/2])$. Lai and Schumaker [105] recently constructed a Hermite interpolation set for super spline spaces with respect to Δ_{PS} that additionally contains the function values and derivatives of a certain order at points different from the vertices of the underlying given (nonsplitted) triangulation (except for the classical case $S_2^1(\Delta)$). This macro element method uses lower degree splines than earlier methods of this type. Again, in [105] a stable local basis was constructed. We note that the case $S_5^2(\Delta_{PS})$ has been considered earlier by Sablonnière [138] and Lai [100].

We close this section with two remarks. First, we note that all these Hermite interpolation methods cannot be transformed into Langrange interpolation on the *whole* triangulation straightforwardly. Second, all these methods lead to interpolation by superspline subspaces. With three exceptions: the classical methods of Fraeijs de Veubecke and Sander, Clough–Tocher and Powell–Sabin.

For results on interpolation by splines with constraints (e.g., shape preserving interpolation, convex interpolation) we refer to Dahmen and Micchelli [49,50], Schmidt and Walther [141,142], Lorente-Pardo et al. [106], Constantini and Manni [46], Chalmers et al. [28]. Concerning the finite element method, we refer to the books by Ciarlet [41], Prenter [132] and Strang and Fix [157].

6. Interpolation by spline spaces

If we consider the results discussed in the above sections, the natural problem of constructing interpolation sets for the *spline* space $S_q^r(\Delta)$ appears. In this section, we summarize results on interpolation by these spaces. Here, we do not consider interpolation by subspaces of $S_q^r(\Delta)$ such as super splines.

As mentioned above, results on interpolation by these spaces are strongly connected with the problem of determining the dimension of these spaces. Therefore, interpolation by $S_q^r(\Delta)$ leads to complex problems, in particular, for Lagrange interpolation.

It is well known that for univariate spline spaces, interpolation sets can be characterized through Schoenberg–Whitney-type conditions (see the books by de Boor [21], by Schumaker [144] and by Nürnberger [116]). In contrast to this, the Schoenberg–Whitney theorem cannot be extended to interpolation by bivariate splines, even for the simplest space $S_1^0(\Delta)$ (cf. [29, p. 136]).

It was shown by Sommer and Strauss [154,155] that the natural multivariate analogue of such Schoenberg–Whitney-type conditions characterizes almost interpolation sets, i.e., point sets that can be transformed into Lagrange interpolation sets for $S_q^r(\Delta)$ by arbitrary small perturbations of the points (see also [52,53,62–64]). In this context, locally linearly independent systems \mathscr{S} of splines play an important role. This means that each open subset in Ω contains a ball B such that the subsystem consisting of all elements of \mathscr{S} having a support with a nonempty intersection with B is linearly independent on B. In a general topological context, Davydov et al. [64] showed the relations of locally linearly independent systems with almost interpolation. In [64], a Schoenberg–Whitney-type characterization of almost interpolation sets for spline spaces that admit a locally linearly independent basis was developed (see also [53]). A locally linearly independent basis for $S_q^1(\Delta)$, $q \ge 5$, was constructed in [53]. For certain superspline spaces of higher smoothness such basis were given by Davydov et al. [59,63]. Recently, Davydov and Schumaker [60] constructed a locally linearly independent basis for $S_q^r(\Delta)$, $q \ge 3r + 2$. For further results on locally linearly independence, see the references in [60].

Concerning numerical purposes, it is desirable to construct explicit interpolation schemes for $S_q^r(\Delta)$, in particular, such that algorithmical complexity of computing a corresponding interpolating spline is $\mathcal{O}(\operatorname{card} \Delta)$. In the following, we describe explicit Lagrange and Hermite interpolation methods for $S_q^r(\Delta)$.

First, it is obvious that a Lagrange interpolation set for $S_q^0(\Delta)$, $q \ge 1$, is obtained by the union of all points $(i/q)v_1 + (j/q)v_2 + (k/q)v_3$, i+j+k=q, where v_1, v_2, v_3 are the vertices of a triangle in Δ . In particular, the set of vertices of Δ is a Lagrange interpolation set for $S_1^0(\Delta)$. In 1986, an algorithm for constructing more general Langrange interpolation sets for $S_1^0(\Delta)$ was given by Chui et al. [31]. Davydov et al. [65] recently gave a characterization of Lagrange interpolation sets for $S_1^0(\Delta)$.

The literature shows that for $q \ge 2$, it is complex problem to construct interpolation set for $S_q^r(\Delta)$, in particular concerning Lagrange interpolation. This was done for certain classes of triangulations, respectively, for splines of certain degree q and smoothness r. These methods are described in the following.

In the beginning 1990s Nürnberger and Rießinger [120,121] developed a general method for constructing Hermite and Lagrange interpolation sets for $S_q^r(\Delta^i)$, i=1,2. We note that the triangulations Δ^i , i=1,2 (see Fig. 5), have to be uniform if q,r arbitrary, whereas the triangulations may be nonuniform if i=1 and $r \in \{0,1\}$, respectively, i=2 and $r \in \{0,1,2\}$. The dimension for such type of spline spaces, and more generally for so-called *crosscut partitions*, was determined by Chui and Wang [37–39] and Schumaker [145] (for *quasi-crosscut partitions*, see [108]). For spline spaces on Δ^i , i=1,2, a basis consisting of the polynomials, *truncated power functions* and so-called *cone splines* exists in the above cases.

The method given in [120,121] is to construct line segments in Ω and to place points on these lines which satisfy the *interlacing condition* of Schoenberg and Whitney for certain univariate spline spaces such that the *principle of degree reduction* can be applied.

This construction of Lagrange interpolation sets uses certain basic steps. In this survey paper, we only describe these basic steps for $S_a^r(\Delta^1)$. For an arbitrary triangle $T \in \Delta^1$, one of the following

A B	B B	BB	B B
			D B
			D B
			D B

Fig. 10. Interpolation steps for $S_q^r(\Delta^1)$.

steps will be applied. (If the number of lines in Step C or D below is nonpositive, then no points are chosen.)

Step A (Starting step): Choose q + 1 disjoint line segments a_1, \ldots, a_{q+1} in T. For $j = 1, \ldots, q+1$, choose q + 2 - j distinct points on a_j .

Step B: Choose q-r disjoint line segments b_1, \ldots, b_{q-r} in T. For $j=1, \ldots, q-r$, choose q+1-r-j distinct points on b_j .

Step C: Choose q - 2r + [r/2] disjoint line segments $c_1, \ldots, c_{q-2r+[r/2]}$ in T. For $j = 1, \ldots, q - 2r$, choose q+1-r-j distinct points on c_j and for $j=q-2r+1, \ldots, q-2r+[r/2]$ choose 2(q-j)-3r+1 distinct points on c_j .

Step D: Choose q - 2r - 1 disjoint line segments d_1, \ldots, d_{q-2r-1} in T. For $j = 1, \ldots, q - 2r - 1$, choose q - 2r - j distinct points on d_j .

Given a Δ^1 triangulation, the above steps are applied to the triangles of Δ^1 as indicated in Fig. 10.

Hermite interpolation sets for $S_q^r(\Delta^i)$, i = 1, 2, are obtained by using the above Lagrange interpolation sets and by "taking limits". This means that Hermite interpolation sets are constructed by shifting the interpolation points to the vertices. More precisely, the Hermite interpolation conditions are obtained by considering a Lagrange interpolation set and letting certain points and line segments coincide. The corresponding Hermite interpolation conditions are as follows. If certain points on some line segment coincide, then directional derivatives along the line segment are considered, and if certain line segments coincide, then directional derivatives orthogonal to the line segment are considered.

For proving the approximation order of these interpolation methods, Nürnberger [118] introduced the principle of weak interpolation. The following results were proved by Nürnberger [118], respectively, Nürnberger and Walz [122] (see also Nürnberger [117]).

Theorem 6.1. Let f be a sufficiently differentiable function. The (Hermite) interpolating spline s_f of f yields (nearly) optimal approximation order for $S_a^1(\Delta^1)$, $q \ge 4$, and $S_a^1(\Delta^2)$, $q \ge 2$.

In particular, the approximation order of the interpolation methods is optimal for $S_q^1(\Delta^1)$, $q \ge 5$, and $S_q^1(\Delta^2)$, $q \ge 4$. In the remaining cases the interpolating spline yields approximation order q. Later, Davydov et al. [55] showed the following result by extending the weak interpolation principle and by using univariate Hermite–Birkhoff interpolation arguments.


Fig. 11. A polyhedron \mathcal{P} .



Fig. 12. Split of a triangle.

Theorem 6.2. Let f be a sufficiently differentiable function and $r \ge 1$. The (Hermite) interpolation spline $s_f \in S_a^r(\Delta^1)$, $q \ge 3.5r + 1$, of f yields optimal approximation order.

If we compare Theorem 6.2 with the general results on approximation order of these spaces in Section 4, we see that Theorem 6.2 is very close to what is possible, in general.

In 1996, the method of Nürnberger et al. [120,121] was extended by Adam [1] to cross-cuts partitions (see also [119]). The results in [1,120,121] were obtained by developing methods which were different from Bézier–Bernstein techniques.

In contrast to this, Bézier–Bernstein methods were used for constructing interpolation sets for spline spaces of degree less or equal to three with respect to Δ^i , i = 1, 2, and for proving their approximation order.

The following cases were considered between 1981 and 1994: $S_3^1(\Delta^1)$ by Sha [150], Bamberger [13], ter Morsche [114], $S_2^1(\Delta^2)$ by Sibson and Thompson [153], Sha [151], Sablonnière [136], Beatson and Ziegler [15], Chui and He [30], Rießinger [133], Zedek [161], Jeeawock-Zedek [84], Jeeawock-Zedek and Sablonnière [85], $S_3^1(\Delta^2)$ by Jeeawock-Zedek [83] and Lai [98].

We proceed by describing interpolation methods for more general classes than Δ^i , i = 1, 2. We start with methods, where triangulations suitable for interpolation by splines are constructed, then we describe interpolation methods for general classes of given triangulations.

In 1999, Nürnberger and Zeilfelder [125] developed an inductive method for constructing triangulations Δ that are suitable for Lagrange and Hermite interpolation by $S_q^r(\Delta)$, $q \ge 2r + 1$, r = 1, 2. Roughly speaking, the construction of these triangulations works as follows. By starting with one triangle, a polyhedron \mathscr{P} as in Fig. 11 is added in each step to obtain a larger triangulation. The polyhedra, which result from triangulations of locally chosen scattered points, have two common edges with the boundary of the subtriangulation constructed so far. This construction is such that the corresponding splines can be extended in each step. For doing this in the case of C^2 splines, it may be necessary to split *some* of the triangles of Δ (see Fig. 12).

The dimension of the resulting spline spaces was determined by using Bézier–Bernstein methods. Lagrange and Hermite interpolation sets were constructed simultaneously. For doing this, suitable



Fig. 13. Lagrange interpolation points in P.

Table 1

Interpolation of Franke's testfunction f(x, y) by splines s_f on a domain Ω containing the unit square

$f - s_f \ _{\infty}$
$.62 \cdot 10^{-5}$
$.80 \cdot 10^{-6}$
$.83 \cdot 10^{-6}$
$.30 \cdot 10^{-6}$

extensions of the above interpolation steps A, B, C, D were developed. Examples for Langrange interpolation points in \mathcal{P} are given in Fig. 13. Again, Hermite interpolation sets are obtained by "taking limits". We note that the corresponding interpolating splines can be computed step by step. In each step only small linear systems of equations have to be solved.

A variant of this method can be used to construct triangulations Δ for arbitrary scattered points in the plane which is similar to *Delaunay triangulations* and to construct interpolation points for $S_q^r(\Delta)$, r=1,2, as described above. Concerning this variant, in some steps it may be necessary, when small angles appear (in the subtriangulation constructed so far), to add instead of a polyhedron \mathscr{P} with at least two triangles (see Fig. 11) only one triangle which then has to be subdivided by a Clough–Tocher split.

Numerical tests with large numbers of interpolation conditions show that this interpolation method yields good approximations for $S_q^1(\Delta)$, $q \ge 4$, and $S_q^2(\Delta)$, $q \ge 7$. In order to obtain good approximations in the remaining cases (for nonuniform triangulations Δ) variants of this method were discussed in [125] (for some typical numerical results see Table 1). We note that in contrast to the macro element methods described in Section 5, by this method only some of the triangles have to be subdivided into three subtriangles, and Lagrange interpolation sets for spline spaces were constructed (see also Nürnberger and Zeilfelder [124]). Moreover, this method can be applied to certain classes of given triangulations Δ , in particular the class of triangulated quadrangulations which was investigated by Nürnberger and Zeilfelder [123].



Fig. 14. A double Clough–Tocher split Δ_{DCT} .



Fig. 15. Wang's special Morgan and Scott triangulation Δ_{MS} .

Earlier, Alfeld [2] and Wang [158] used completely different methods for Hermite interpolation by quintic C^2 splines. As in the finite element case (see Section 5), these approaches are based on a splitting procedure which is applied to *every* triangle of a given triangulation Δ . In [2], a double Clough–Tocher element Δ_{DCT} as in Fig. 14 was considered. Such an element consist of nine triangles. Alfeld [2] proposed to solve a linear system with approximately 100 equations for every triangle of Δ . This system is obtained by interpolating function values and derivatives up to order 2 at the vertices of Δ and certain additional conditions for the space $S_5^2(\Delta_{DCT})$.

Wang [158] proposed to split *every* triangle of Δ into seven subtriangles. The three corresponding splitting points have to be chosen as in Fig. 15. This is a special case of Morgan and Scott's triangulation Δ_{MS} (see Section 3). In this case, the Hermite interpolation set for $S_5^2(\Delta_{MS})$ consists of function values and derivatives up to order 2 at the vertices of Δ , of certain first and second order derivatives at interior points of every edge of Δ and function value and of one first derivative at each splitting point. Moreover, we mention that a Hermite interpolation set for $S_6^2(\Delta)$ was constructed by Gao [73], where Δ is a triangulated convex quadrangulation (see Section 4) that has to satisfy additional properties.

We proceed by describing interpolation methods for general classes of given triangulations.

Hermite interpolation sets for $S_q^1(\Delta)$, $q \ge 5$, where Δ is an arbitrary triangulation, were defined by Morgan and Scott [112] and Davydov [53]. The Hermite interpolation sets given in [53,112] cannot be transformed to Lagrange interpolation straightforwardly. In 1998, Davydov and Nürnberger [54] gave a different method for constructing explicit Hermite and Lagrange interpolation sets for $S_q^1(\Delta)$, $q \ge 5$. As shown in [54], this approach can also be applied to $S_4^1(\Delta)$, where Δ has to be slightly modified if exceptional constellations of triangles occur. Roughly speaking, the inductive method of constructing interpolation sets is as follows: In each step, a vertex of Δ and all triangles with this vertex having a common edge with the subtriangulation considered so far are added. Then the interpolation points are chosen locally on these triangles, where the number of interpolation points depends on so-called *semi-singular* vertices which may result from degenerate edges.

Earlier, Gao [72] defined a Hermite interpolation scheme for $S_4^1(\Delta)$ in the special case when Δ is an *odd degree triangulation*. These are triangulations, where every interior vertex has odd degree.



Fig. 16. A nested polygon triangulation.

Moreover, Chui and Hong [33] considered quasi-interpolation by $S_4^1(\Delta)$ (see also [32,81]). There it is shown that in order to achieve optimal approximation order of the quasi-interpolating spline certain edges of a given triangulation Δ have to be swapped.

We mention that interpolation by $S_4^1(\Delta)$ leads to an unsolved problem (see [7]; see also [115]) which we formulate as follows:

Problem. Given an arbitrary triangulation Δ . Does there exist a Hermite interpolation set for $S_4^1(\Delta)$ which includes function and gradient values for *all* vertices of Δ ?

The results on Hermite interpolation by $S_4^1(\Delta)$ described above show that for these classes of triangulations, the answer to this question is yes.

The case of cubic C^1 splines, $S_3^1(\Delta)$, is more complex since not even the dimension of these spaces is known for arbitrary triangulations Δ (see Section 3).

In 1987, Gmelig Meyling [75] considered Lagrange interpolation by these spaces. There, a global method for constructing Lagrange interpolation sets involving function values at all vertices of a given triangulation Δ is proposed. This approach requires to solve a large linear system of equations, where it is not guaranteed that this system is solvable. In [75] some numerical experiments were given.

Davydov et al. [58] (see also [56,57]) investigated interpolation by $S_3^1(\Delta)$, where Δ is contained in the general class of the so-called nested polygon triangulations. These are triangulations consisting of nested closed simple polygons $\mathcal{P}_0, \mathcal{P}_1, \ldots, \mathcal{P}_k$ whose vertices are connected by line segments (see Fig. 16).

The construction of interpolation sets for $S_3^1(\Delta)$ in [58] is inductive by passing through the points of the nested polygons $\mathscr{P}_0, \mathscr{P}_1, \ldots, \mathscr{P}_k$ in clockwise order: In each step, a point of a nested polygon and all triangles with this vertex having a common edge with the subtriangulation considered so far are added. Then the interpolation points are chosen locally on these triangles, where the number of interpolation points is different if semi-singular vertices exist or not. In addition, it was proved in [58] that the number of interpolation points coincides with Schumaker's lower bound lb_3^1 (see Section 3), and therefore the dimension of these spaces is equal to lb_3^1 .

The space $S_3^1(\Delta)$ is interesting for applications since the number of interpolation points is relatively small. It was remarked in [58] that numerical examples (with up to 100 000 interpolation conditions)



Fig. 17. Coloring of a given triangulation Δ .

show that in order to obtain good approximations, it is desirable to subdivide *some* of the triangles of Δ . The method of constructing interpolation points proposed in [58] also works for these modified triangulations.

We now discuss the problem of *local Langrange interpolation*. Concerning the construction and reconstruction of surfaces, it is sometimes desirable that only function values (and no (orthogonal) derivatives) are involved which for scattered data fitting can be computed approximately by using local methods. The classical method of Clough and Tocher [45] yields a local Hermite interpolation scheme for cubic C^1 splines on triangles which are splitted into three subtriangles (see Section 5). However, this Hermite interpolation scheme cannot be transformed straightforwardly into a local Lagrange interpolation scheme on a given triangulation.

The investigations of Nürnberger and Zeilfelder [127] show that local Lagrange interpolation schemes can be constructed by coloring the triangles of a given triangulation appropriately with two colors (see Fig. 17) and by subdividing the triangles of one color by a Clough–Tocher split. The authors developed an algorithm for constructing local Lagrange interpolation sets for spline spaces $S_3^1(\tilde{\Delta})$. This approach also works for higher degree C^1 spline spaces. Roughly speaking, the triangulations $\tilde{\Delta}$ are obtained by splitting half of the triangles of a given triangulation Δ into three subtriangles. Moreover, in this context "local" means that the corresponding fundamental functions $s_i \in S_3^1(\tilde{\Delta})$, i = 1, ..., d, determined by

 $s_i(z_j) = \delta_{i,j}, \quad j = 1, \dots, d$

have local support (here, $\delta_{i,j}$ denotes Kronecker's symbol).

The algorithm consists of two algorithmical steps. In the first step, Lagrange interpolation points are chosen on the edges of Δ such that interpolating spline is uniquely determined (only) on these edges. In the second step, the triangles are colored black and white (by a fast algorithm) such that at most two consecutive triangles (with a common edge) have the same color (see Fig. 17). Then the black triangles are subdivided by a Clough–Tocher split, and in the interior of the white triangles, additional Lagrange interpolation points are chosen. Then the interpolating spline is uniquely determined on the whole triangulation.

Since recently, Nürnberger et al. [128] are investigating the construction of local Lagrange interpolation schemes by cubic C^1 splines on convex quadrangulations. Since the classical Hermite

interpolation scheme of Fraeijs de Veubeke [71] and Sander [140] (see Section 5) cannot be transformed into a local Lagrange interpolation scheme on a given quadrangulation, in [128] a coloring method will be used for local Lagrange interpolation. It turned out that it is a much more complex problem to develop a fast algorithm for coloring quadrangulations in a desired way than for triangulations. The coloring methods known in graph theory (see the book by Jensen and Toft [86]) do not provide such an algorithm. In [128], the authors are investigating this problem for general classes of quadrangulations.

We now discuss an open problem concerning interpolation by cubic C^1 splines.

Problem. Given an arbitrary triangulation Δ . Does there exist an interpolation set for $S_3^1(\Delta)$ which includes function values for *all* vertices of Δ ?

Obviously, this question is strongly connected with the conjecture on the dimension of these spaces given in Section 3. We remark that for the interpolation sets constructed by Davydov et al. [58], it may happen that the function values at certain vertices are not included. On the other hand, for subclasses of nested polygon triangulations as described above function values at all vertices are included in the corresponding interpolation sets. This is also true for the local Lagrange interpolation sets constructed in [127].

By using Euler's formula (see Section 2), the lower bound in (3), lb_{2r+1}^r , for $S_{2r+1}^r(\Delta)$, $r \ge 1$, is

$$lb_{2r+1}^{r} = \binom{r+1}{2} + (r+1)V_{I} + \binom{r+2}{2}V_{B} + \sum_{i=1}^{V_{I}}\sigma_{i}.$$
(11)

Therefore, more general the question arises if an interpolation set for $S_{2r+1}^r(\Delta)$, $r \ge 1$, exists which includes function values at all vertices. Gmelig et al. [77] (see also [78]) proposed a method for constructing certain splines from $S_{2r+1}^r(\Delta)$, $r \ge 1$, that interpolate the function values at all vertices of a given triangulation Δ . This global method requires to solve a large linear system of equations, where it is not known if this system is solvable.

We turn now to the case $S_2^1(\Delta)$. As a consequence of their general method, Nürnberger and Zeilfelder [125] constructed Lagrange interpolation sets for $S_2^1(\Delta_Q)$, where Δ_Q is a triangulation of the following general type. By starting with one triangle, Δ_Q is described inductively as follows. Given a subtriangulation $\tilde{\Delta}_Q$, a triangle \tilde{T} which has one common edge with $\tilde{\Delta}_Q$ is added. Then in clockwise order, quadrangles (with two diagonals) having one common edge with $\tilde{\Delta}_Q$ and triangles having one common point with $\tilde{\Delta}_Q$ are added, where the last quadrangle also has one common edge with \tilde{T} (see Fig. 18). The resulting subtriangulation is again denoted by $\tilde{\Delta}_Q$. By proceeding with this method Δ_Q is finally obtained. The Lagrange interpolation set for $S_2^1(\Delta_Q)$ given in [125] consists of the vertices of Δ_Q (except the intersection points of the diagonals) together with three additional points in the starting triangle.

If in Δ_Q instead of quadrangles with two diagonals arbitrary quadrangles are considered, then for the quadrangles with only one diagonal no interpolation point can be chosen. In this case, no good approximations can be expected, in general (see the general comment on quadratic C^1 splines in Section 5). It was mentioned in [125] that in order to obtain good approximations, numerical tests (up to 40 000 Lagrange interpolation points) showed that the vertices of Δ_Q should be rather uniformly distributed.



Fig. 18. The triangulation Δ_Q .

We close this section with the following problem (see also Nürnberger and Zeilfelder [126]).

Problem. Construct interpolation sets for $S_q^1(\Delta)$, q = 2, 3, for general classes of triangulations Δ .

References

- [1] M.H. Adam, Bivariate spline-interpolation auf crosscut-partitionen, Dissertation, Mannheim, 1995.
- [2] P. Alfeld, A bivariate C^2 Clough-Tocher scheme, Comput. Aided Geom. Des. 1 (1984) 257–267.
- [3] P. Alfeld, A trivariate Clough-Tocher scheme for tetrahedral data, Comput. Aided Geom. Des. 1 (1984) 169-181.
- [4] P. Alfeld, On the dimension of piecewise polynomial functions, in: D.F. Griffiths, G.A. Watson (Eds.), Numerical Analysis, Longman Scientific & Technical, London, 1986, pp. 1–23.
- [5] P. Alfeld, Upper and lower bounds on the dimension of multivariate spline spaces, SIAM J. Numer. Anal. 33 (2) (1996) 571–588.
- [6] P. Alfeld, B. Piper, L.L. Schumaker, Minimally supported bases for spaces of bivariate piecewise polynomials of smoothness r and degree d≥4r + 1, Comput. Aided Geom. Des. 4 (1987) 105–123.
- [7] P. Alfeld, B. Piper, L.L. Schumaker, An explicit basis for C¹ quartic bivariate splines, SIAM J. Numer. Anal. 24 (1987) 891–911.
- [8] P. Alfeld, L.L. Schumaker, The dimension of bivariate spline spaces of smoothness r for degree $d \ge 4r + 1$, Constr. Approx. 3 (1987) 189–197.
- [9] P. Alfeld, L.L. Schumaker, On the dimension of bivariate spline spaces of smoothness r and degree d = 3r + 1, Numer. Math. 57 (1990) 651–661.
- [10] P. Alfeld, L.L. Schumaker, Non-existence of star-supported spline bases, SIAM J. Numer. Anal., to appear, preprint 1999.
- P. Alfeld, L.L. Schumaker, M. Sirvent, On dimension and existence of local bases for multivariate spline spaces, J. Approx. Theory 70 (1992) 243–264.
- [12] J.H. Argyis, I. Fried, D.W. Scharpf, The TUBA family of plate elements for the matrix displacement method, Aeronaut. J. Roy. Aeronaut. Soc. 72 (1968) 701–709.
- [13] L. Bamberger, Interpolation in bivariate spline spaces, in: W. Schempp, K. Zeller (Eds.), Multivariate Approximation Theory III, Proceedings of Conference, Oberwolfach, Germany, Birkhäuser, Basel, 1985, pp. 25–34.
- [14] R.E. Barnhill, G. Farin, C¹ quintic interpolation over triangles: two explicit representations, Int. J. Numer. Math. Eng. 17 (1981) 1763–1778.
- [15] R.K. Beatson, Z. Ziegler, Monotonicity preserving surface interpolation, SIAM J. Numer. Anal. 22 (2) (1985) 401–411.
- [16] K. Bell, A refined triangular plate bending element, J. Math. Eng. 1 (1969) 101-122.
- [17] L.J. Billera, Homology of smooth splines: generic triangulations and a conjecture of Strang, Trans. Amer. Math. Soc. 310 (2) (1988) 325–340.

- [18] L.J. Billera, R. Haas, The dimension and bases of divergence-free splines; a homological approach, Approx. Theory Appl. 7 (1) (1991) 91–99.
- [19] W. Boehm, G. Farin, J. Kahmann, A survey of curve and surface methods in CAGD, Comput. Aided Geom. Des. 1 (1984) 1–60.
- [20] B.D. Bojanov, H.A. Hakopian, A.A. Sahakian, Spline Functions and Multivariate Approximation, Kluwer, Dordrecht, 1993.
- [21] C. de Boor, A Practical Guide to Splines, Springer, New York, 1978.
- [22] C. de Boor, B-form basics, in: G. Farin (Ed.), Geometric Modeling, SIAM, Philadelphia, 1987, pp. 131-148.
- [23] C. de Boor, K. Höllig, Approximation order from bivariate C^1 cubics: a counterexample, Proc. AMS 87 (1983) 649–655.
- [24] C. de Boor, K. Höllig, Approximation power of smooth bivariate pp functions, Math. Z. 197 (1988) 343-363.
- [25] C. de Boor, K. Höllig, S. Riemenschneider, Box Splines, Springer, Berlin, 1993.
- [26] C. de Boor, Q. Jia, A sharp upper bound on the approximation order of smooth bivariate pp functions, J. Approx. Theory 72 (1993) 24–33.
- [27] J.M. Carncier, J.M. Peña, A Marsden type identity for periodic trigonometric splines, J. Approx. Theory 75 (1993) 248–255.
- [28] B.L. Chalmers, D. Leviatan, M.P. Prophet, Optimal interpolating spaces preserving shape, J. Approx. Theory 98 (1999) 354–373.
- [29] C.K. Chui, Multivariate Splines, CBMS, Vol. 54, SIAM, Philadelphia, 1988.
- [30] C.K. Chui, T.X. He, On location of sample points in C¹ quadratic bivariate spline interpolation, in: L. Collatz, G. Meinardus, G. Nürnberger (Eds.), Numerical Methods of Approximation Theory, International Series of Numerical Mathematics, Vol. 81, Birkhäuser, Basel, 1987, pp. 30–43.
- [31] C.K. Chui, T.X. He, R.H. Wang, Interpolation by bivariate linear splines, in: J. Szabados, J. Tandori (Eds.), Alfred Haar Memorial Conference, North-Holland, Amsterdam, 1986, pp. 247–255.
- [32] C.K. Chui, D. Hong, Construction of local C^1 quartic spline element for optimal-order approximation, Math. Comp. 65 (1996) 85–98.
- [33] C.K. Chui, D. Hong, Swapping edges of arbitrary triangulations to achieve the optimal order of approximation, SIAM J. Numer. Anal. 34 (1997) 1472–1482.
- [34] C.K. Chui, D. Hong, Q. Jia, Stability of optimal-order approximation by bivariate splines over arbitrary triangulations, Trans. Amer. Math. Soc. 347 (1995) 3301–3318.
- [35] C.K. Chui, M.-J. Lai, On multivariate vertex splines and applications, in: C. Chui, L.L. Schumaker, F. Utreras (Eds.), Topics in Multivariate Approximation, Academic Press, New York, 1987, pp. 19–36.
- [36] C.K. Chui, M.-J. Lai, On bivariate super vertex splines, Constr. Approx. 6 (1990) 399-419.
- [37] C.K. Chui, R.H. Wang, Multivariate spline spaces, J. Math. Anal. Appl. 94 (1983) 197-221.
- [38] C.K. Chui, R.H. Wang, On smooth multivariate spline functions, Math. Comp. 41 (1983) 131-142.
- [39] C.K. Chui, R.H. Wang, Spaces of bivariate cubic and quartic splines on type-1 triangulations, J. Math. Anal. Appl. 101 (1984) 540–554.
- [40] P.G. Ciarlet, Sur l'élément de Clough et Tocher, RAIRO Anal. Numér. 2 (1974) 19-27.
- [41] P.G. Ciarlet, The Finite Element Method for Elliptic Problems, North-Holland, Amsterdam, 1977.
- [42] P.G. Ciarlet, Interpolation error estimates for the reduced Hsieg–Clough–Tocher triangles, Math. Comp. 32 (1978) 335–344.
- [43] P.G. Ciarlet, P.A. Raviart, General Lagrange and Hermite interpolation in \mathbb{R}^n with applications to finite element methods, Arch. Rational Mech. Anal. 46 (1972) 177–199.
- [44] J.F. Ciavaldini, J.C. Nèdélec, Sur l' élément de Frajis de Veubecke et Sander, RAIRO Anal. Numér 2 (1974) 29–45.
- [45] R.W. Clough, J.L. Tocher, Finite element stiffness matries for analysis of plates in bending, Proceedings of Conference on Matrix Methods in Structural Mechanics, Wright Patterson A.F.B., OH, 1965.
- [46] P. Costantini, C. Manni, A local shape-preserving interpolation scheme for scattered data, Comput. Aided Geom. Des. 16 (1999) 385–405.
- [47] W. Dahmen, Bernstein-Bézier representation of polynomial surfaces, Proceedings of ACM SIGGRAPH, Dallas, 1986.

- [48] W. Dahmen, C.A. Michelli, Recent progress in multivariate splines, in: C.K. Chui, L.L. Schumaker, J.D. Ward (Eds.), Approximation Theory IV, Academic Press, New York, 1983, pp. 27–121.
- [49] W. Dahmen, C.A. Micchelli, Convexity of multivariate Bernstein polynomials and box spline surfaces, Stud. Sci. Math. Hung. 23 (1988) 265–287.
- [50] W. Dahmen, C.A. Micchelli, Convexity and Bernstein polynomials on *k*-simploids, Acta. Math. Appl. Sin. (1990) 50–66.
- [51] W. Dahmen, R.H.J. Gmelig Meyling, J.H.M. Ursem, Scattered data interpolation by bivariate C¹-piecewise quadratic functions, Approx. Theory Appl. 6 (1990) 6–29.
- [52] O. Davydov, On almost interpolation, J. Approx. Theory 91 (1997) 398-418.
- [53] O. Davydov, Locally linearly independent basis for C¹ bivariate splines of degree q≥5, in: M. Daehlen, T. Lyche, L.L. Schumaker (Eds.), Mathematical Methods for Curves and Surfaces II, Vanderbilt University Press, Nashville, 1998, pp. 71–77.
- [54] O. Davydov, G. Nürnberger, Interpolation by C^1 splines of degree $q \ge 4$ on triangulations, J. Comput. Appl. Math., to appear.
- [55] O. Davydov, G. Nürnberger, F. Zeilfelder, Approximation order of bivariate spline interpolation for arbitrary smoothness, J. Comput. Appl. Math. 90 (1998) 117–134.
- [56] O. Davydov, G. Nürnberger, F. Zeilfelder, Interpolation by cubic splines on triangulations, in: C.K. Chui, L.L. Schumaker (Eds.), Approximation Theory IX, Vanderbilt University Press, Nashville, 1998, pp. 17–24.
- [57] O. Davydov, G. Nürnberger, F. Zeilfelder, Interpolation by splines on triangulations, in: M.W. Muller, M.D. Buhmann, D.H. Mache, M. Felten (Eds.), New Developments in Approximation Theory, International Series of Numerical Mathematics, Vol. 132, Birkhäuser, Basel, 1999, pp. 49–70.
- [58] O. Davydov, G. Nürnberger, F. Zeilfelder, Cubic spline interpolation on nested polygon triangulations, in: A. Cohen, C. Rabut and L.L. Schumaker (Eds.), Curves and Surfaces, St. Malo, 1999. Vanderbilt University Press, to appear, Preprint 1999.
- [59] O. Davydov, G. Nürnberger, F. Zeilfelder, Bivariate spline interpolation with optimal approximation order, Constr. Approx., to appear, preprint 1998.
- [60] O. Davydov, L.L. Schumaker, Locally linearly independent bases for bivariate polynomial spline spaces, preprint 1999.
- [61] O. Davydov, L.L. Schumaker, On stable local bases for bivariate polynomial splines, preprint 1999.
- [62] O. Davydov, M. Sommer, H. Strauss, On almost interpolation by multivariate splines, in: G. Nürnberger, J.W. Schmidt, G. Walz (Eds.), Multivariate Approximation and Splines, International Series of Numerical Mathematics, Vol. 125, Birkhäuser, Basel, 1997, pp. 45–58.
- [63] O. Davydov, M. Sommer, H. Strauss, Locally linearly independent systems and almost interpolation, in: G. Nürnberger, J.W. Schmidt, G. Walz (Eds.), Multivariate Approximation and Splines, International Series of Numerical Mathematics, Vol. 125, Birkhäuser, Basel, 1997, pp. 59–72.
- [64] O. Davydov, M. Sommer, H. Strauss, On almost interpolation and locally linearly independent basis, East J. Approx. 5 (1999) 67–88.
- [65] O. Davydov, M. Sommer, H. Strauss, Interpolation by bivariate linear splines, J. Comput. Appl. Math. 119 (2000) 115–131.
- [66] D. Diener, Instability in the dimension of spaces of bivariate piecewise polynomials of degree 2r and smoothness order r, SIAM J. Numer. Anal. 27 (2) (1990) 543–551.
- [67] P. Diereckx, S. van Leemput, T. Vermeire, Algorithms for surface fitting using Powell-Sabin splines, IMA J. Numer. Anal. 12 (2) (1992) 271–299.
- [68] G. Farin, A modified Clough-Tocher interpolant, Comput. Aided Geom. Des. 2 (1985) 19-27.
- [69] G. Farin, Triangular Bernstein-Bézier patches, Comput. Aided Geom. Des. 3 (1986) 83-127.
- [70] G. Farin, Geometric Modeling, SIAM, Philadelphia, 1987.
- [71] G. Fraejis de Veubeke, Bending and stretching of plates, Proceedings of Conference on Matrix Methods in Structural Mechanics, Wright Patterson A.F.B., Ohio, 1965.
- [72] J. Gao, Interpolation by C^1 quartic bivariate splines, J. Math. Res. Expo. 11 (1991) 433–442.
- [73] J. Gao, A C^2 finite element and interpolation, Computing 50 (1993) 69–76.
- [74] M. Gasca, Th. Sauer, Polynomial interpolation in several variables, Adv. Comput. Math., to appear, preprint 1999.

- [75] R.H.J. Gmelig Meyling, Approximation by piecewise cubic C^1 -splines on arbitrary triangulations, Numer. Math. 51 (1987) 65–85.
- [76] R.H.J. Gmelig Meyling, P.R. Pfluger, On the dimension of the spline space $S_2^1(\Delta)$ in special cases, in: W. Schempp, K. Zeller (Eds.), Multivariate Approximation Theory III, Birkhäuser, Basel, 1985, pp. 180–190.
- [77] R.H.J. Gmelig Meyling, P.R. Pfluger, Smooth interpolation to scattered data by bivariate piecewise polynomials of odd degree, Comput. Aided Geom. Des. 7 (1990) 439–458.
- [78] T.A. Grandine, An iterative method for computing multivariate C^1 piecewise polynomial interpolants, Comput. Aided Geom. Des. 4 (1987) 307–320.
- [79] G. Heindl, Interpolation and approximation by piecewise quadratic C¹ functions of two variables, in: W. Schempp, K. Zeller (Eds.), Multivariate Approximation Theory, International Series of Numerical Mathematics, Vol. 51, Birkhäuser, Basel, 1979, pp. 146–161.
- [80] D. Hong, Spaces of bivariate spline functions over triangulation, Approx. Theory Appl. 7 (1991) 56-75.
- [81] D. Hong, H. Liu, R. Mohapatra, Optimal triangulations and smoothness conditions for bivariate splines, in: C.K. Chui, L.L. Schumaker (Eds.), Approximation Theory IX, Vanderbilt University Press, Nashville, 1998, pp. 129–136.
- [82] A. Ibrahim, L.L. Schumaker, Super spline spaces of smoothness r and degree $d \ge 3r + 2$, Constr. Approx. 7 (1991) 401–423.
- [83] F. Jeeawock-Zedek, Interpolation scheme by C¹ cubic splines on a non-uniform type-2 triangulation of a rectangular domain, C.R. Acad. Sci. Paris Sér. I Math. 314 (1992) 413–418.
- [84] F. Jeeawock-Zedek, Operator norm and error bounds for interpolating quadratic splines on a non-uniform type-2 triangulation of a rectangular domain, Approx. Theory Appl. 10 (2) (1994) 1–16.
- [85] F. Jeeawock-Zebek, P. Sablonnière, Hermite and Lagrange interpolation by quadratic splines on non-uniform criss-cross triangulations, in: P.J. Laurent, et al. (Eds.), Wavelets, Images and Surface Fitting, Addison-Wesley, Reading, MA, 1994, pp. 445–452.
- [86] T.R. Jensen, B. Toft, Graph Coloring Problems, Wiley, New York, 1995.
- [87] R.Q. Jia, Approximation by smooth bivariate splines on a three-direction mesh, in: C.K. Chui, L.L. Schumaker, J. Ward (Eds.), Approximation Theory IV, Academic Press, New York, 1983, pp. 539–545.
- [88] R.Q. Jia, Approximation order from certain spaces of smooth bivariate splines on a three directional mesh, Trans. Amer. Math. Soc. 295 (1986) 199–212.
- [89] R.Q. Jia, Local approximation order of box splines, Scentia Sinica 31 (1988) 274–285.
- [90] R.Q. Jia, Lower Bounds on the Dimension of Spaces of Bivariate Splines, International Series of Numerical Mathematics, Vol. 94, Birkhäuser, Basel, 1990, pp. 155–165.
- [91] R.Q. Jia, Lecture Notes on Multivariate Splines, University of Alberta, Edmonton, Canada, 1990.
- [92] M. Laghchim-Lahlou, Éléments finis composites de class C^k dans \mathbb{R}^2 , Dissertation, Rennes, 1991.
- [93] M. Laghchim-Lahlou, P. Sablonnière, Composite quadrilateral finite elements of class C^r, in: T. Lyche, L.L. Schumaker (Eds.), Mathematical Methods in CAGD, Academic Press, New York, 1989, pp. 413–418.
- [94] M. Laghchim-Lahlou, P. Sablonnière, C^r finite elements of HCT, PS and FVS types, in: J. Periaux, R.P. Shaw (Eds.), Proceedings of the Fifth International Symposium on Numerical Methods in Engineering, Vol. 2, Springer, Berlin, 1989, pp. 163–168.
- [95] M. Laghchim-Lahlou, P. Sablonnière, Triangular finite elements of HCT type and class C^{ρ} , Adv. Comput. Math. 2 (1994) 101–122.
- [96] M. Laghchim-Lahlou, P. Sablonnière, Quadrilateral finite elements of FVS type and class C^r , Numer. Math. 30 (1995) 229–243.
- [97] M. Laghchim-Lahlou, P. Sablonnière, C^r-finite elements of Powell–Sabin type on the three directional mesh, Adv. Comput. Math. 6 (1996) 191–206.
- [98] M.-J. Lai, Approximation order from bivariate C^1 -cubics on a four-directional mesh is full, Comput. Aided Geom. Des. 11 (1994) 215–223.
- [99] M.-J. Lai, Scattered data interpolation and approximation using bivariate C^1 piecewise cubic polynomials, Comput. Aided Geom. Des. 13 (1996) 81–88.
- [100] M.-J. Lai, On C² quintic spline functions over triangulations of Powell–Sabin's type, J. Comput. Appl. Math. 73 (1996) 135–155.
- [101] M.-J. Lai, L.L. Schumaker, Scattered data interpolation using C^2 supersplines of degree six, SIAM J. Numer. Anal. 34 (1997) 905–921.

- [102] M.-J. Lai, L.L. Schumaker, On the approximation power of bivariate splines, Adv. Comput. Math. 9 (1998) 251– 279.
- [103] M.-J. Lai, L.L. Schumaker, On the approximation power of splines on triangulated quadrangulations, SIAM J. Numer. Anal. 36 (1999) 143–159.
- [104] M.-J. Lai, L.L. Schumaker, Macro-elements and stable local bases for splines on Clough-Tocher triangulations, preprint 2000.
- [105] M.-J. Lai, L.L. Schumaker, Macro-elements and stable local bases for splines on Powell–Sabin triangulations, preprint 2000.
- [106] J. Lorente-Pardo, P. Sablonnière, M.C. Serrano-Perez, Subharmonicity and convexity properties of Bernstein polynomials and Bezier nets on triangles, Comput. Aided Geom. Des. 16 (1999) 287–300.
- [107] C. Manni, On the dimension of bivariate spline spaces over rectilinear partitions, Approx. Theory Appl. 7 (1) (1991) 23-34.
- [108] C. Manni, On the dimension of bivariate spline spaces over generalized quasi-cross-cut partitions, J. Approx. Theory 69 (1992) 141–155.
- [109] A. Le Méhauté, Construction of surfaces of class C^k on a domain $\Omega \subset \mathbb{R}^2$ after triangulation, in: W. Schempp, K. Zeller (Eds.), Multivariate Approximation Theory II, Birkhäuser, Basel, 1982, pp. 223–240.
- [110] A. Le Méhauté, Unisolvent interpolation in Rⁿ and the simplicial finite element method, in: C. Chui, L.L. Schumaker,
 F. Utreras (Eds.), Topics in Multivariate Approximation, Academic Press, New York, 1987, pp. 141–151.
- [111] C.A. Michelli, Mathematical Aspects of Geometric Modelling, CBMS, Vol. 65, SIAM, Philadelphia, 1995.
- [112] J. Morgan, R. Scott, A nodal basis for C^1 piecewise polynomials of degree $n \ge 5$, Math. Comp. 29 (1975) 736–740.
- [113] J. Morgan, R. Scott, The dimension of piecewise polynomials, 1977, unpublished manuscript.
- [114] H. ter Morsche, Bivariate cubic periodic spline interpolation on a three directional mesh, in: L.L. Schumaker et al. (Eds.), Approximation Theory V, Proceedings of Fifth International Symposium, College Station/Texas, 1986, pp. 487–490.
- [115] E. Nadler, Hermite interpolation by C¹ bivariate splines, in: W. Dahmen, M. Gasca, C. Michellli (Eds.), Contributions to the Computation of Curves and Surfaces, Vol. 2, Academia de Ciencias Exactas, Zaragoza, 1990, pp. 55–66.
- [116] G. Nürnberger, Approximation by Spline Functions, Springer, Berlin, 1989.
- [117] G. Nürnberger, Approximation by univariate and bivariate splines, in: D. Bainov, V. Covachev (Eds.), Second International Colloquium on Numerical Analysis, VSP, Utrecht, 1994, pp. 143–155.
- [118] G. Nürnberger, Approximation order of bivariate spline interpolation, J. Approx. Theory 87 (1996) 117–136.
- [119] G. Nürnberger, O. Davydov, G. Walz, F. Zeilfelder, Interpolation by bivariate splines on crosscut partitions, in: G. Nürnberger, J.W. Schmidt, G. Walz (Eds.), Multivariate Approximation and Splines, International Series of Numerical Mathematics, Vol. 125, Birkhäuser, Basel, 1997, pp. 189–204.
- [120] G. Nürnberger, T. Rießinger, Lagrange and Hermite interpolation by bivariate splines, Numer. Funct. Anal. Optim. 13 (1992) 75–96.
- [121] G. Nürnberger, T. Rießinger, Bivariate spline interpolation at grid points, Numer. Math. 71 (1995) 91-119.
- [122] G. Nürnberger, G. Walz, Error analysis in interpolation by bivariate C^1 -splines, IMA J. Numer. Anal. 18 (1998) 485–508.
- [123] G. Nürnberger, F. Zeilfelder, Spline interpolation on convex quadrangulations, in: C.K. Chui, L.L. Schumaker (Eds.), Approximation Theory IX, Vanderbilt University Press, Nashville, 1998, pp. 259–266.
- [124] G. Nürnberger, F. Zeilfelder, Lagrange interpolation by splines on triangulations, in: R.H. Wang (Ed.), Proceedings of the Morningside Institute, Peking, preprint 1999.
- [125] G. Nürnberger, F. Zeilfelder, Interpolation by spline spaces on classes of triangulations, J. Comput. Appl. Math. 119 (2000) 347–376.
- [126] G. Nürnberger, F. Zeilfelder, On bivariate spline spaces, in: W. Haußmann, K. Jetter and M. Reimer (Eds.), Advances in Multivariate Approximation, Wiley, 1998, pp. 227–231.
- [127] G. Nürnberger, F. Zeilfelder, in preparation.
- [128] G. Nürnberger, L.L. Schumaker, F. Zeilfelder, in preparation.
- [129] P. Percell, On cubic and quartic Clough-Tocher finite elements, SIAM J. Numer. Anal. 13 (1976) 100-103.
- [130] M.J.D. Powell, Piecewise quadratic surface fitting for contour plotting, in: D.J. Evans (Ed.), Software for Numerical Analysis, Academic Press, New York, 1974, pp. 253–277.

- [131] M.J.D. Powell, M.A. Sabin, Piecewise quadratic approximation on triangles, ACM Trans. Math. Software 4 (1977) 316–325.
- [132] P. Prenter, Splines and Variational Methods, Wiley, New York, 1975.
- [133] T. Rießinger, Interpolation by bivariate quadratic splines on a four-directional mesh, Computing 49 (1992) 129–137.
- [134] D.J. Ripmeester, Upper bounds on the dimension of bivariate spline spaces and duality in the plane, in: M. Daehlen, T. Lyche, L.L. Schumaker (Eds.), Mathematical Methods for Curves and Surfaces, Vanderbilt University Press, Nashville, 1995, pp. 455–466.
- [135] T. de Rose, R. Goldman, M. Lounsbery, A tutorial introduction to blossoming, in: H. Hagen, D. Roller (Eds.), Geometric Modelling, Methods and Applications, Springer, Berlin, 1991, pp. 267–286.
- [136] P. Sablonnière, Bernstein-Bézier methods for the construction of bivariate spline approximants, Comput. Aided Geom. Des. 2 (1985) 29–36.
- [137] P. Sablonnière, Composite finite elements of class C^k, J. Comput. Appl. Math. 12 (1985) 541-550.
- [138] P. Sablonnière, Composite finite elements of class C², in: C.K. Chui, L.L. Schumaker, F.I. Utreras (Eds.), Topics in Multivariate Approximation, Academic Press, New York, 1987, pp. 207–217.
- [139] P. Sablonnière, Error bounds for Hermite interpolation by quadratic splines on an α-Triangulation, IMA J. Numer. Anal. 7 (1987) 495–508.
- [140] G. Sander, Bornes supérieures et inférieures dans l'analyse matricielle des plaques en flexion-torsion, Bull. Soc. Roy. Sci. Liège 33 (1964) 456–494.
- [141] J.W. Schmidt, M. Walther, Gridded data interpolation with restrictions on the first order derivatives, in: G. Nürnberger, J.W. Schmidt, G. Walz (Eds.), Multivariate Approximation and Splines, International Series of Numerical Mathematics, Vol. 125, Birkhäuser, Basel, 1997, pp. 289–305.
- [142] J.W. Schmidt, M. Walther, Tensor product splines on refined grids in S-convex interpolation, in: W. Haußmann, K. Jetter, M. Riemer, et al. (Eds.), Multivariate Approximation, Mathematical Research, Vol. 101, Akademie Verlag, Berlin, 1997, pp. 189–202.
- [143] L.L. Schumaker, On the dimension of piecewise polynomials in two variables, in: W. Schempp, K. Zeller (Eds.), Multivariate Approximation Theory, Birkhäuser, Basel, 1979, pp. 396–412.
- [144] L.L. Schumaker, Spline Functions: Basic Theory, Wiley-Interscience, New York, 1980.
- [145] L.L. Schumaker, Bounds on the dimension of spaces of multivariate piecewise polynomials, Rocky Mountain J. Math. 14 (1984) 251–264.
- [146] L.L. Schumaker, Dual bases for spline spaces on a cell, Comput. Aided Geom. Des. 5 (1987) 277-284.
- [147] L.L. Schumaker, Triangulation methods, in: C.K. Chui, L.L. Schumaker, F.I. Uteras (Eds.), Topics in Multivariate Approximation, Academic Press, New York, 1987, pp. 219–232.
- [148] L.L. Schumaker, On super splines and finite elements, SIAM J. Numer. Anal. 4 (1989) 997-1005.
- [149] H.-P. Seidel, An introduction to polar forms, IEEE Comput. Graphics Appl. 13 (1) (1993) 38-46.
- [150] Z. Sha, On interpolation by $S_3^1(\Delta_{m,n}^1)$, Approx. Theory Appl. 1 (1985) 1–18.
- [151] Z. Sha, On interpolation by $S_2^1(\Delta_{m,n}^2)$, Approx. Theory Appl. 1 (1985) 71–82.
- [152] X.Q. Shi, The singularity of Morgan-Scott triangulation, Comput. Aided Geom. Des. 8 (1991) 201-206.
- [153] R. Sibson, G.D. Thomson, A seamed quadratic element for contouring, Comput. J. 24 (4) (1981) 378–382.
- [154] M. Sommer, H. Strauss, Interpolation by uni- and multivariate generalized splines, J. Approx. Theory 83 (1995) 423–447.
- [155] M. Sommer, H. Strauss, A condition of Schoenberg–Whitney type for multivariate spline interpolation, Adv. Comput. Math. 5 (1996) 381–397.
- [156] G. Strang, Piecewise polynomials and the finite element method, Bull. Amer. Math. Soc. 79 (1973) 1128–1137.
- [157] G. Strang, G. Fix, An analysis of the Finite Element Method, Prentice-Hall, New York, 1973.
- [158] T. Wang, A C^2 -quintic spline interpolation scheme on triangulation, Comput. Aided Geom. Des. 9 (1992) 379–386.
- [159] T. Whelan, A representation of a C^2 interpolant over triangles, Comput. Aided Geom. Des. 3 (1986) 53–66.
- [160] W. Whiteley, A matrix for splines, in: P. Nevai, A. Pinkus (Eds.), Progress in Approximation Theory, Academic Press, Boston, 1991, pp. 821–828.
- [161] F. Zedek, Interpolation de Lagrange par des splines quadratique sur un quadrilatere de ℝ², RAIRO Anal. Numér. 26 (1992) 575–593.
- [162] A. Ženišek, Interpolation polynomials on the triangle, Numer. Math. 15 (1970) 283–296.
- [163] A. Ženišek, A general theorem on triangular finite C^m -elements, RAIRO Anal. Numér. 2 (1974) 119–127.



JOURNAL OF COMPUTATIONAL AND APPLIED MATHEMATICS

Journal of Computational and Applied Mathematics 121 (2000) 153-163

www.elsevier.nl/locate/cam

Multivariate spline and algebraic geometry $\stackrel{\text{tr}}{\sim}$

Ren-Hong Wang

Institute of Mathematical Sciences, Dalian University of Technology, Dalian 116024, People's Republic of China

Received 2 November 1999; received in revised form 26 February 2000

Abstract

The purpose of this survey is to emphasize the special relationship between multivariate spline and algebraic geometry. We will not only point out the algebraic-geometric method of multivariate spline, but also the algebraic-geometric background and essence of multivariate spline. Especially, we have made an introduction to the so-called piecewise algebraic curve, piecewise algebraic variety, and some of their properties. © 2000 Elsevier Science B.V. All rights reserved.

MSC: 14C17; 14G40; 41A15; 41A46; 65D07; 65D10

Keywords: Multivariate (weak) spline; Algebraic geometry; Piecewise algebraic curve; Piecewise algebraic variety

1. On multivariate spline

It is well known that the polynomial is an important approximation tool of functions as well as curves and surfaces. In fact, according to the classical Weierstrass approximation theorem, any continuous function on a bounded closed domain can be uniformly approximated by polynomials on this domain.

Unfortunately, polynomials have global properties that are so strong that a polynomial can be determined solely by its properties on a neighbourhood of a given point in the domain. This is, however, not the case for most practical geometric objects such as the surfaces of aircrafts, cars, ships and satellites.

Splines as piecewise polynomials, instead, can be used to approximate any continuous, smooth, and even discontinuous function within any given tolerance. Moreover, a spline is easy to store, to evaluate, and to manipulate on a digital computer; a myriad of applications in scientific and engineering computation have been found. Spline has become a kind of fundamental tool for computational geometry, numerical analysis, approximation, and optimization, etc. [3,5].

^{*} Project supported by the National Natural Science Foundation of China, and 973 Foundation of China. *E-mail address:* renhong@dlut.edu.cn (R.-H. Wang).

Now let us turn to multivariate splines. Multivariate splines are piecewise polynomials defined on domains of more than one variable. In what follows, we consider bivariate splines, for it is easy to understand.

Let *D* be a domain in \mathbb{R}^2 and Δ a partition of *D* consisting of finite irreducible algebraic curves $\Gamma_i : \ell_i(x, y) = 0, i = 1, ..., N$, where the coefficients of $\ell_i(x, y)$ are real numbers.

Denote by D_i , i = 1, ..., T, all the cells of Δ . For integer $k > \mu \ge 0$, we say that

$$S_k^\mu(\varDelta) := \{s \in C^\mu(D) | s|_{D_i} \in \mathbf{P}_k, \ \forall i\}$$

is a multivariate spline space with degree k and smoothness μ , where \mathbf{P}_k denotes the collection of polynomials

$$\mathbf{P}_k := \left\{ \sum_{i=0}^k \sum_{j=0}^{k-i} c_{ij} x^i y^j | c_{ij} \text{ real} \right\}$$

if $k \ge 0$, and $\mathbf{P}_k = \{0\}$ if k < 0.

According to the above definition, multivariate splines will be very useful in solving problems in computational geometry and at the same time, we will encounter and have to deal with a number of problems in differential geometry, such as tangent, curvature, offset, and minimal surface problems. However, most people have not mentioned the relationship between algebraic geometry and multivariate splines.

First, by using Bezout's theorem in algebraic geometry, the author discovered the following fundamental theorem on multivariate splines [15].

Theorem 1. $s(x, y) \in S_k^{\mu}(\Delta)$ if and only if the following conditions are satisfied:

1. For each interior edge of Δ , which is defined by $\Gamma_i : \ell_i(x, y) = 0$, there exits the so-called smoothing cofactor $q_i(x, y)$ such that

$$p_{i1} - p_{i2} = \ell_i^{\mu+1} q_i, \tag{1}$$

where the polynomials p_{i1} and p_{i2} are determined by the restriction of s(x, y) on the two cells D_{i1} and D_{i2} with Γ_i as the common edge and $q_i \in \mathbf{P}_{\alpha-(\mu+1)n_i}$, $\alpha = \max(\text{degrees of } p_{i1} \text{ and } p_{i2})$, $n_i = \text{degree of } \ell_i$.

2. For any interior vertex v_j of Δ , the following conformality conditions are satisfied

$$\sum \left[\ell_i^{(j)}(x,y)\right]^{\mu+1} q_i^{(j)}(x,y) \equiv 0,$$
(2)

where the summation is taken over all the interior edges $\Gamma_i^{(j)}$ passing through v_j , and the signs of the smoothing cofactors $q_i^{(j)}$ are refixed in such a way that when a point crosses $\Gamma_i^{(j)}$ from D_{i2} to D_{i1} , it goes around v_j in a counter-clockwise manner.

Theorem 1 shows that the multivariate spline, in principle, is equivalent to an algebraic subject defined by Theorem 1. Furthermore, Theorem 1 has also shown a most general method for studying the multivariate splines over any given partition. It was called the smoothing cofactor-conformality method.

2. Dimension of multivariate spline space

Let the homogeneous system of linear algebraic equations corresponding to the global conformality condition be

$$BQ = 0, (3)$$

where Q is the column vector whose components are formed by coefficients of the smoothing cofactor of all the interior edges. The elements in matrix B are formed by coefficients of the expansion on $[l_i(x, y)]^{\mu+1}$.

Denote by N the total number of interior edges in Δ , and denote by n_i the degree of the *i*th interior edge, then the unknowns in the homogeneous system of linear equations BQ = 0 is

$$\sum_{i=1}^N \binom{k-n_i(\mu+1)+2}{2}.$$

If we denote $\sigma = \operatorname{rank} B$, according to the algebraic theory, then the dimension of solution space of BQ = 0 is

$$\sum_{i=1}^{N} \binom{k-n_i(\mu+1)+2}{2} - \sigma.$$

Adding the freedom of a polynomial of degree k defined in the "source cell" $\binom{k+2}{2}$, we have the following theorem:

Theorem 2 (Wang [17]).

$$\dim S_k^{\mu}(\Delta) = \binom{k+2}{2} + \sum_{i=1}^N \binom{k-n_i(\mu+1)+2}{2} - \sigma.$$
(4)

Especially, if every interior edge is a straight line, we have

Theorem 3 (Wang [15,17]).

$$\dim S_k^{\mu}(\varDelta) = \binom{k+2}{2} + N\binom{k-\mu+1}{2} - \sigma,$$
(5)

where N is the total number of the interior edges in Δ , and σ is the rank of matrix B whose elements are the coefficients of the homogeneous system of linear equations corresponding to the global conformality condition.

Although the theorems mentioned above have given the dimension formulae of multivariate spline spaces $S_k^{\mu}(\Delta)$, in principle, it is very complicated to calculate σ . Moreover, it is not only dependent on the topological property of Δ , but also sometimes dependent on the geometric property of Δ .

In fact, multivariate splines have a strong background of algebraic geometry.

Let us consider in detail the special partition Δ consisting of finite straight lines. By using the map $\varphi : (x, y) \to [x, y, 1]$, the partition Δ is embedded in CP^2 , and any edge Γ_i can be represented

$$\binom{k+2}{2} \times \binom{k-\mu+1}{2} N_v$$

matrix.

It is clear that the dimension of the solution space defined by $M_v \zeta = 0$ is

$$\binom{k-\mu+1}{2}N\binom{k+2}{2}+\tau,$$

where τ is the dimension of the solution space of the following linear system:

$$M_{v}^{\mathrm{T}}\xi = 0. \tag{6}$$

However, (6) shows that the points $\{(\alpha_i, \beta_i, \gamma_i)\}_{i=1}^{N_v}$ lie on algebraic curves as follows

$$\sum_{|\lambda|=\mu+1} C_\lambda lpha_i^{\lambda_1} eta_i^{\lambda_2} \gamma_i^{\lambda_3} = 0, \quad i=1,\ldots,N_v,$$

where N_v is the number of edges $\Gamma_i : \alpha_i x + \beta_i y + \gamma_i z = 0$ passing $v, \lambda = (\lambda_1, \lambda_2, \lambda_3), \lambda_1, \lambda_2, \lambda_3 \ge 0, |\lambda| = \lambda_1 + \lambda_2 + \lambda_3$, and the unknown $\binom{\mu+3}{2}$ -vectors C_{λ} are taken from a $\binom{k+2}{2}$ -vector in different ways.

Hence, the problems on dim $S_k^{\mu}(\Delta)$ should be also the problems in algebraic geometry.

They seem to be heavily dependent on the geometric properties of the partition Δ .

By using Pascal's theorem in the algebraic geometry, Du shows how dim $S_{\frac{1}{2}}(\Delta_{MS})$ depends on the geometric properties of the triangulation Δ_{MS} ([7,8,10]).

The close relationship between bivariate C^1 -quadratic spline based on Morgan–Scott partition and the classical Pascal's Theorem and some results in multivariate C^1 -quadratic splines on Morgan–Scott partition in \mathbb{R}^n [11] stimulate us to generalize Pascal's Theorem to *n*-dimensional cases.

To do this, one first takes Pascal's Theorem as the following problem: when does the hexagon obtained from a triangle by cutting out its three corners inscribe a conic? A natural problem is when the polyhedron obtained by cutting out the corners of an *n*-simplex (such a polyhedron is called Pascal's polyhedron) inscribes an *n*-dimensional quadratic hypersurface. We have [13].

Theorem 4. An *n*-Pascal's polyhedron inscribes an *n*-dimensional quadratic hypersurface if and only if each of its (n - 1)-faces inscribes an (n - 1)-dimensional quadratic hypersurface.

Billera [1] developed a homological approach to the problem of computing the dimension of $S_k(\Delta)$, where Δ is a finite *d*-dimensional simplicial complex embedded in \mathbb{R}^d . He also applied it specifically to the case of triangulated manifold Δ in the plane, getting lower bounds on dim $S_k^{\mu}(\Delta)$ for all μ . An algebraic criterion developed by Billera [1] is both necessary and sufficient for the piecewise polynomial on a *d*-dimensional complex Δ to be smooth of order μ . The basic idea of this criterion can be found in [15,6]. A *d*-complex Δ is called strongly connected if for any two *d*-simplices $\sigma, \sigma' \in \Delta$, there is a sequence of *d*-simplices

$$\sigma = \sigma_1, \sigma_2, \ldots, \sigma_r = \sigma'$$

such that for each i < r, $\sigma_i \cap \sigma_{i+1}$ has dimension d-1. Here σ_i and σ_{i+1} are called adjacent.

Theorem 5 (Billera, [1]). Suppose that Δ is a strongly connected d-simplex such that all links of simplices are also strongly connected complices. Let F be a piecewise polynomial which, restricted on each simplex of Δ is a polynomial of degree $\leq k$, and $\mu \geq 0$. Then $F \in S_k^{\mu}(\Delta)$ if and only if a relation similar to (1) in Theorem 1 holds for each pair σ_1, σ_2 of adjacent d-simplices in Δ .

The homology $H_*(C_k^{\mu})$ of the complex C_k^{μ} is useful for studying C^{μ} piecewise polynomials of degree at most k. Billera proved [1].

Theorem 6. Let Δ be a strongly connected d-complex with strongly connected links. Then for $k \ge 0$ and $\mu \ge 0$,

 $S_k^{\mu}(\varDelta) = H_d(C_k^{\mu}).$

A well-known Strang's conjecture has been proved by Billera [1].

Theorem 7. For generic embeddings of any triangulated 2-maniford Δ in \mathbb{R}^2 ,

$$\dim S_k^1(\Delta) = \binom{k+2}{2} f_2 - (2k+1)f_1^0 + 3f_0^0$$

where f_2 is the number of triangles in Δ , f_1^0 and f_0^0 the number of interior edges and vertices, respectively.

Billera and Rose [2] considered the formal power series

$$\sum_{k\geq 0} \dim_R S_k^{\mu}(\varDelta) \lambda^k,$$

and showed that the following form

$$P(\lambda)/(1-\lambda)^{d+1}$$

holds under mild conditions on Δ , where $d = \dim(\Delta)$, and $P(\lambda)$ is a polynomial in λ with integral coefficients and satisfies

P(0) = 1, $P(1) = f_d(\Delta)$, and $P'(1) - (\mu + 1)f_{d-1}^0(\Delta)$.

Moreover, Billera and Rose showed how the polynomial $P(\lambda)$ and bases of the spaces $S_k^{\mu}(\Delta)$ can be calculated by using Gröbner basis techniques of computational commutative algebra [4].

3. Multivariate weak spline

Let Δ be a partition of the domain D consisting of finite straight line segments. Denote by D_i , i = 1, ..., T, all the cells of Δ . Let Γ be an edge of Δ, S_{Γ} be a set of points on Γ , and $|S_{\Gamma}| := \operatorname{card}(S_{\Gamma})$ be limited. Suppose that S is a set of points. Denote by $C^{\mu}_{(0)}(S)$ the set of functions with μ -smoothness at each point of S.

$$W_k^{\mu}(\varDelta) := \{ w(x, y) \in C_{(0)}^{\mu}(S) | w(x, y) |_{D_i} \in P_k, \ \forall D_i, \ S = U_j S_{\Gamma_j} \}$$

is called the multivariate weak spline space.

Theorem 8 (Wang [16]). Suppose that $w(x, y) \in P_k$, and S is a set of points on the line y - ax - b = 0, then

$$D^j w(x_i, y_i) = 0, \quad j = 0, \dots, \mu, \quad \forall (x_i, y_i) \in S$$

if and only if there exist q(x, y) and $c_m(x)$ such that

$$w(x,y) = (y - ax - b)^{\mu+1}q(x,y) + \sum_{m=1}^{\mu+1} (y - ax - b)^{\mu+1-m} \left(\prod_{i=1}^{|S|} (x - x_i)\right)^m c_m(x),$$
(7)

where $q \in P_{k-\mu-1}$, $c_m(x) \in P_{k-\mu-(|S|-1)m-1}(x)$, and $c_m(x) = 0$ identically provided $k - \mu - (|S| - 1)m - 1 < 0$.

The polynomial q(x, y) shown above is still called the smoothing cofactor; however, $c_m(x)$ is called the weak smoothing cofactor of order m. The corresponding conformality condition, and global conformality condition of the multivariate weak spline are

$$\sum_{A} l_{ij}^{\mu+1} q_{ij}(x, y) + \sum_{m=1}^{\mu+1} l_{ij}^{\mu+1-m} \left(\prod_{x_t \in S_{l_{ij}}} (x - x_t) \right)^m c_{mij}(x) = 0$$
(8)

and

$$\sum_{A_r} l_i^{\mu+1} q_i(x, y) + \sum_{m=1}^{\mu+1} l_i^{\mu+1-m} \left(\prod_{x_i \in S_i} (x - x_i) \right)^m c_{mi}(x) = 0, \quad r = 1, \dots, M,$$
(9)

respectively, where M is the number of interior edges of Δ .

Theorem 9 (Xu & Wang [23]). For any given partition Δ , $w(x, y) \in W_k^{\mu}(\Delta)$, if and only if there exist a smoothing cofactor and the weak smoothing cofactors of order m, $m = 1, ..., \mu + 1$ on each interior edge, and the global conformality condition is satisfied.

4. Scattered data interpolation and piecewise algebraic curve

The interpolation of scattered data is an important topic in computational geometry. It is concerned with several practical areas such as CAD (computer-aided design), CAM (computer-aided manufacture), CAE (computer-aided engineering), and Image processing, etc. Let $\{(x_i, y_i; z_i)\}_{i=1}^N$ be a given scattered data $(N \gg 1)$. The problem of scattered data interpolation is how we can find a function (it should be simple) z = f(x, y) such that the following interpolation conditions:

$$z_i = f(x_i, y_i), \quad i = 1, \dots, N$$
 (10)

are satisfied

A system of N funcitons $\varphi_1, \ldots, \varphi_N$ defined on a point set S is called unisolvent on S if

$$|\varphi_i(x_j)|\neq 0$$

holds for every selection of distinct points x_1, \ldots, x_N in S.

It follows that $\varphi_1, \ldots, \varphi_N$ is unisolvent on S if and only if the linear combination of the $\varphi's$ that vanishes on N distinct points of S vanishes identically.

Haar's Theorem. Let S be a point set in n-dimensional Euclidean space \mathbb{R}^n , $n \ge 2$. Suppose that S contains at least an interior point, say p. Let $\varphi_1, \ldots, \varphi_N$ (N > 1) be defined on S and continuous in a neighborhood of p. Then this set of functions cannot be unisolvent on S.

Therefore, if the interpolation function f(x, y) shown in (10) is taken from a linear space, such as a polynomial space or a spline space, then to get a unique solution of (10), one has to choose carefully the knot-set $K := \{(x_i, y_i)\}_{i=1}^N$.

The knot-set $K := \{(x_i, y_i)\}_{i=1}^N$ is called a suitable knot-set for (10) if (10) has a unique solution for any given z_1, \ldots, z_N .

By using algebra, $K := \{(x_i, y_i)\}_{i=1}^N$ is suitable for (10), if and only if K does not lie on a curve defined by

$$\Gamma := \{(x, y) \mid f(x, y) = 0, f \in \mathscr{S}\}$$
(11)

where \mathscr{S} is the linear space spanned by $\varphi_1, \ldots, \varphi_N$ [20].

In principle, to solve an interpolation problem, one has to deal with the properties of curves.

When $\mathscr{S} := S_k^{\mu}(\varDelta)$ is a multivariate spline space, the curve

$$\Gamma := \{ (x, y) \, | \, s(x, y) = 0, s \in S_k^{\mu}(\Delta) \}$$

is called a piecewise algebraic curve. Therefore, a key problem on the interpolation by multivariate splines is to study the piecewise algebraic curve. It is obvious that the piecewise algebraic curve is a kind of generalization of the classical algebraic curve [9,18].

Because of the possibility $\{(x, y) \in D | s|_{D_i} = p(x, y) = 0\} \cap \overline{D}_i = \emptyset$ (empty), it is more difficult to study the piecewise algebraic curve. It is well known that Bezout's theorem is an important and classical result in algebraic geometry. Its weak form says that two algebraic curves will have infinitely many intersection points if they have more intersection points than the product of their degrees (it is called Bezout's number). For $S_m^r(\Delta)$ and $S_n^t(\Delta)$, denote by BN = BN($m, r; n, t; \Delta$) the Bezout's number. Then any two piecewise algebraic curves.

$$\Gamma: f(x, y) = 0$$
, and $\gamma: g(x, y) = 0$, $f \in S_m^r(\Delta)$, $g \in S_n^t(\Delta)$

must have infinitely many intersection points provided that they have more than BN intersection points.

A fundamental problem on the piecewise algebraic curve is how to find the Bezout's number. In general, this problem is very difficult. A triangulation Δ is called 2-triangle-signed, if each triangle in Δ can be marked by -1 or 1 such that any two adjacent triangles in Δ are marked by different signs. The following generalization of Bezout's theorem has been obtained.

Theorem 10 (Shi & Wang [12]). BN $(m, 0; n, 0; \Delta) = mnT$ if Δ is a 2-triangle-signed triangulation or if mn is even, and BN $(m, 0; n, 0; \Delta) \leq mnT - [(V_{odd} + 2)/3]$ in general, where T is the number of triangles in Δ and V_{odd} is the number of odd vertices in Δ .

By using the resultant on the polar coordinates, we have [19,22]

Theorem 11. Let Γ_i : $s_i(x, y) = 0$, $s_i \in S_{k_i}^1$, i = 1, 2 be piecewise algebraic curves. Then for any given interior vertex, say v, the Bezout number on R(v) satisfies the following inequality:

$$BN(k_1, 1; k_2, 1) \le n(k_1k_2 - 1) + 1, \tag{12}$$

where n is the number of edges passing through v.

An algebraic curve l(x, y) = 0 is called a branch of the algebraic curve Γ : L(x, y) = 0, if l(x, y) is a factor of L(x, y). Noting the feature if the piecewise algebraic curve, we define the so-called "local branch" of a piecewise algebraic curve [18]: A piecewise algebraic curve γ : t(x, y) = 0 is called a local branch of the piecewise algebraic curve Γ : s(x, y) = 0, if there exists a union U of cells of the partition Δ such that γ is a branch of Γ on U. It is important to discuss the real intersection points of piecewise algebraic curves. One of the basic problems is how one can determine the number of real intersection points of two piecewise algebraic curves on a certain cell of the partition. It is well known that any $p(x, y) \in P_n$ can also be represented in B-net form on a triangle Δ as follows

$$p(u_1, u_2, u_3) = \sum_{|\lambda|=n} b_{\lambda} u_1^{\lambda_1} u_2^{\lambda_2} u_3^{\lambda_3},$$
(13)

where

$$b_{\lambda} = p_{\lambda} \frac{n!}{\lambda_1! \lambda_2! \lambda_3},\tag{14}$$

and (u_1, u_2, u_3) is the barycentric coordinates of (x, y) on Δ , $\lambda = (\lambda_1, \lambda_2, \lambda_3)$, $|\lambda| = \lambda_1 + \lambda_2 + \lambda_3$, $i \in \{0, 1, ..., n\}$, p_{λ} are Bezier ordinates of p. By the change of variables as follows

$$u_{1} = \left(\frac{2t_{1}^{2}}{1+t_{1}^{2}+t_{2}^{2}}\right)^{2}, \quad u_{2} = \left(\frac{2t_{2}^{2}}{1+t_{1}^{2}+t_{2}^{2}}\right)^{2}, \quad u_{3} = \left(\frac{t_{1}^{2}+t_{2}^{2}-1}{1+t_{1}^{2}+t_{2}^{2}}\right)^{2}, \quad (15)$$

the polynomial $p(u_1, u_2, u_3)$ is represented by

$$p(u_1, u_2, u_3) = \frac{4}{(1 + t_1^2 + t_2^2)^{2n}} P^*(t_1, t_2),$$
(16)

where

$$P^{*}(t_{1},t_{2}) = \sum_{|\lambda|=n} b_{\lambda} t_{1}^{2\lambda_{1}} t_{2}^{2\lambda_{2}} (t_{1}^{2} + t_{2}^{2} - 1)^{2\lambda_{3}}$$
(17)

is defined on the whole R^2 . Therefore, one can estimate the number of real intersection points of the curves

$$\Gamma_i: p_i(x, y) = 0, \quad i = 1, 2$$

on Δ by computing the number of real intersection points of the curves Γ_1^* : $P_i^*(t_1, t_2) = 0$, i = 1, 2 on R^2 . In fact, the following Sturm-type theorem holds.

Theorem 12. The number of real intersection points of two algebraic curves

$$\Gamma_i: p_i(x, y) = 0, \quad i = 1, 2$$

on a triangle Δ is bounded by the number of real intersection points of the following two algebraic curves:

 $\Gamma_i^*: p_i^*(t_1, t_2) = 0, \quad i = 1, 2$

on the whole R^2 , where the corresponding polynomials p_i and P_i^* , i = 1, 2, are defined as above.

5. Piecewise algebraic variety

Let k be a fixed algebraically closed field. We define affine n-space over k, denoted by A_k^n , or simply A^n , to be the set of all n-tuples of elements of k. Denote by $k[x_1, \ldots, x_n]$ the polynomial ring in n variables over k. Let Δ be a partition of A^n , and $P(\Delta)$ be the set of piecewise polynomials defined on Δ . It is obvious that

$$S^{\mu}(\varDelta) = \{ f \mid f \in C^{\mu}(A^n) \cap p(\varDelta) \}$$

is also a ring on k, which is called C^{μ} spline ring. It is clear that $k[x_1, \ldots, x_n] \subset S^{\mu}(\Delta)$. A C^{μ} piecewise algebraic variety X is defined as the zero set of C^{μ} splines in $S^{\mu}(\Delta)$, i.e.,

$$X = \{ x \in A^n \, | \, s(x) = 0, s \in F \subset S^{\mu}(\Delta) \}.$$

Piecewise algebraic variety is a new topic in the algebraic geometry as well as the computational geometry. The degree of $f \in S^{\mu}(\Delta)$ is the maximal degree of polynomials corresponding to f on all cells of Δ , which is denoted by

 $\deg f = \max_{\sigma_i \in \varDelta} \deg(f|_{\sigma_i}).$

Proposition 13. The set

$$S_m^{\mu}(\varDelta) = \{ f \mid \deg f \leqslant m, \ f \in S^{\mu}(\varDelta) \}$$

is a finite-dimensional linear vector space on $k, m \ge 0$.

For *n*-dimensional space \mathbb{R}^n , n > 2, we always assume that the facets in partition Δ are hyperplanes. In this case, the conclusions in Theorem 1 corresponding to spline space $S_m^{\mu}(\Delta)$ also hold.

Definition 14 (Su, Wang et al. [14,21]). Let Δ be a partition of A^n , $X \subseteq A^n$. If there exist $f_1, \ldots, f_r \in S^{\mu}(\Delta)$, such that

$$X = z(f_1, \dots, f_r) = \bigcap_{i=1}^r z(f_i),$$
(18)

then X is called a C^{μ} piecewise algebraic variety in A^{n} with respect to Δ . If there exists $f \in S^{\mu}(\Delta)$, such that X = z(f), then X is called a C^{μ} piecewise algebraic hypersurface. A one-dimensional C^{μ} piecewise algebraic curve. A two-dimensional piecewise algebraic variety is called a C^{μ} piecewise algebraic surface.

Let every facet in Δ be a hyperplane, according to the conformality condition and the representation of multivariate spline, we have

Theorem 15 (Su, Wang et al. [14,21]). $S^{\mu}(\Delta)$ is a Nöther ring.

Proposition 16. Let X be a C^{μ} piecewise algebraic variety in A^n with respect to Δ , then X is irreducible if and only if I(X) is a prime ideal of $S^{\mu}(\Delta)$, where

$$I(X) := \{ f \mid f(x) = 0, \ x \in X, \ f \in S^{\mu}(\Delta) \}.$$

Theorem 17 (Su, Wang et al. [14,21]). Every C^{μ} piecewise algebraic variety in A^n with respect to Δ can be represented by the union of finite number of irreducible C^{μ} piecewise algebraic varieties X_1, \ldots, X_r , that is,

$$X = \bigcup_{i=1}^{n} X_i.$$
⁽¹⁹⁾

References

- [1] L. Billera, Homology of smooth splines: generic triangulations and a conjecture of Strang, Trans. Amer. Math. Soc. 310 (1988) 325–340.
- [2] L. Billera, L. Rose, A dimension series for multivariate splines, Discrete Comput. Geom. 6 (1991) 107-128.
- [3] C. de Boor, A Practical Guide to Splines, Springer, Berlin, 1978.
- [4] B. Buchberger, Gröbner bases an algorithmic method in polynomial ideal theory, in: N.K. Bose (Ed.), Recent Trends in multidimensional Systems Theory, Reidel, Dordrecht, 1985.
- [5] C.K. Chui, R.H. Wang, Spaces of bivariate cubic and quartic splines on type-1 triangulations, J. Math. Anal. Appl. 101 (1984) 540–554.
- [6] P. Ciarlet, Lectures on the Finite Element Method, Tata Institute of Fundamental Research, Bombay, 1975.
- [7] H. Du, A geometric approach to dim $S_2^1(\Delta_{MS})$, in: R.H. Wang (Ed.), Proceedings on Computational Geometry, AMS & Intern. Press, Boston, 2000, to appear.
- [8] J. Morgan, R. Scott, The dimension of piecewise polynomials, manuscript, 1975.
- [9] I.R. Shafarevich, Basic Algebraic Geometry Vols. 1 & 2, Springer, Berlin, 1977, 1974.
- [10] X.Q. Shi, High Dimensional Splines, Ph.D. Dissertation, Jilin University, 1988.
- [11] X.Q. Shi, The singularity of Morgan-Scott triangulation, CAGD 8 (1991) 201-206.
- [12] X.Q. Shi, R.H. Wang, Bezout's number for piecewise algebraic curves, BIT 2 (1999) 339-349.
- [13] X.Q. Shi, R.-H. Wang, The generalization of Pascal's Theorem and Morgan–Scott's partition, in: R.H. Wang (Ed.), Proceedings on Computational Geometry, AMS & Intern. Press, Boston, 2000, to appear.
- [14] Z.X. Su, Piecewise algebraic curves and surfaces and their applications in CAGD, Ph.D. Thesis, DUT, 1993.
- [15] R.H. Wang, The structural characterization and interpolation for multivariate splines, Acta Math. Sinica 18 (1975) 91–106 (English transl.ibid. 18 (1975) 10–39).
- [16] R.H. Wang, Multivariate Weak Splines, The Congress of Chinese Computation Mathematical Society, Guangzhou, 1979.
- [17] R.H. Wang, The dimension and basis of spaces of multivariate spline, J. Comput. Appl. Math. 12 & 13 (1985) 163–177.
- [18] R.H. Wang, Multivariate spline and its applications in science and technology, Rend. Sem. Mat. Fis. Milano, LXIII (1993) 213–229.
- [19] R.H. Wang, On piecewise algebraic curves, in: F. Fontanella, et al., (Eds.), Advanced Topics in Multivariate Approximation, World Scientific, Singapore, 1996, pp. 355–361.

- [20] R.H. Wang, X.Z. Liang, Multivariate Function Approximations, Scientific Press, Beijing, 1988.
- [21] R.H. Wang, X.Q. Shi, Z.X. Luo, Z.X. Su, Multivariate spline and its Applications, Science Press, Beijing, 1994, (English edition, Kluwer Press, Dordrecht to appear).
- [22] R.H. Wang, G.H. Zhao, An introduction to the piecewise algebraic curve, in: T. Mitsui (Ed.), Theory and Application of Scientific and Technical Computing, RIMS, Kyoto University, 1997, pp. 196–205.
- [23] Z.Q. Xu, R.H. Wang, Multivariate weak spline, submitted for publication.



Journal of Computational and Applied Mathematics 121 (2000) 165-177

JOURNAL OF COMPUTATIONAL AND APPLIED MATHEMATICS

www.elsevier.nl/locate/cam

A unified theory of radial basis functions Native Hilbert spaces for radial basis functions II

Robert Schaback

Universität Göttingen, Lotzestraße 16-18, D-37083 Göttingen, Germany

Received 4 January 2000

Abstract

This contribution provides a new formulation of the theory of radial basis functions in the context of integral operators. Instead of Fourier transforms, the most important tools now are expansions into eigenfunctions. This unifies the theory of radial basis functions in \mathbb{R}^d with the theory of zonal functions on the sphere S^{d-1} and the theory of kernel functions on Riemannian manifolds. New characterizations of native spaces and positive definite functions are included. The paper is a self-contained continuation of an earlier survey (R. Schaback, International Series of Numerical Mathematics, Vol. 132, Birkhäuser, Basel, 1999, pp. 255–282) over the native spaces associated to (not necessarily radial) basis functions. (C) 2000 Elsevier Science B.V. All rights reserved.

1. Introduction and overview

For the numerical treatment of functions of many variables, *radial basis functions* are useful tools. They have the form $\phi(||x-y||_2)$ for vectors $x, y \in \mathbb{R}^d$ with a *univariate* function ϕ defined on $[0, \infty)$ and the Euclidean norm $\|\cdot\|_2$ on \mathbb{R}^d . This allows to work efficiently for large dimensions d, because the function boils the multivariate setting down to a univariate setting. Usually, the multivariate context comes back into play by picking a large number M of points x_1, \ldots, x_M in \mathbb{R}^d and working with linear combinations

$$s(x) := \sum_{j=1}^{M} \lambda_j \phi(\|x_j - x\|_2).$$
(1.1)

In certain cases, functions from a space \mathscr{P} of low-degree polynomials have to be added, and these complications are dealt with in Section 5 of [8], while Section 6 shows how to get rid of these.

E-mail address: schaback@math.uni-goettingen.de (R. Schaback).

However, in this paper we completely ignore additional polynomials and the related notion of *conditional* positive definiteness.

Besides the classical *radial* basis functions on the whole space \mathbb{R}^d , there are *zonal* functions on the (d-1)-dimensional sphere $S^{d-1} \subset \mathbb{R}^d$. These have the form $\phi(x^T y) = \phi(\cos(\alpha(x, y)))$ for points x, y on the sphere spanning an angle of $\alpha(x, y) \in [0, \pi]$ at the origin. Here, the symbol ^T denotes vector transposition, and the function ϕ should be defined on [-1,1]. *Periodic* multivariate functions can also be treated, e.g. by reducing them to products of univariate periodic functions. Another very important case are basis functions on *Riemannian manifolds*, as introduced by Narcowich [4] and investigated by Dyn et al. [2]. In this paper we generally consider symmetric functions $\Phi : \Omega \times \Omega \to \mathbb{R}$ on some domain $\Omega \subseteq \mathbb{R}^d$, covering the four situations described above.

All of these cases of *basis functions* share a common Hilbert space foundation, but they differ when going over to more specific analytical techniques. The survey [8] deals with the Hilbert space basics, but it stops at the point where the four cases need different techniques. This paper goes a step further by providing analytical techniques that serve as a common ground for a simultaneous advanced treatment.

Let us give a somewhat more specific overview, adding some motivation from the standard theory of univariate spline functions. The crucial ingredient is the well-known representation of the "energy inner product" of classical splines in the form

$$(Lf, Lg)_{L_2(\Omega)} \rightleftharpoons (f, g)_L \tag{1.2}$$

with some linear differential operator *L*. Natural univariate splines of odd degree 2n - 1 are related to $L = d^n/dx^n$ on $\Omega = [a, b] \subset \mathbb{R}$. Furthermore, the fundamental work of Duchon [1] on thin-plate and polyharmonic splines is based strongly on the use of $L = \Delta^m$ on $\Omega = \mathbb{R}^d$. For general symmetric positive definite (not necessarily radial) basis functions $\Phi : \Omega \times \Omega \to \mathbb{R}$, there is no obvious analogue of such an operator.

However, we want to take advantage of (1.2) and thus proceed to work our way towards a proper definition of L, starting from the bilinear form

$$(\Phi(\cdot, x), \Phi(\cdot, y))_{\Phi} := \Phi(x, y), \quad x, y \in \Omega$$

on all linear combinations of functions $\Phi(\cdot, x)$. Then Hilbert space completion leads to the notion of the native space $\mathscr{F}_{\Phi}(\Omega)$ of a symmetric positive definite function $\Phi: \Omega \times \Omega \to \mathbb{R}$. Within functional analysis, the current theories of basis functions (radial, zonal, periodic, and Riemannian) have a common theory of their associated "native" Hilbert space of functions in which they act as a generalized reproducing kernel. The different special cases are naturally related to geometric invariants of the native spaces. Thus we recollect the basic facts on native spaces in a preliminary section, quoting proofs from [8].

The actual paper then starts in Section 3 by imbedding the native space $\mathscr{F}_{\Phi}(\Omega)$ into $L_2(\Omega)$ and studying (in Section 4) the adjoint *C* of the embedding, which turns out to be a convolution-type integral operator with kernel Φ . This operator *C* will finally be equal to $(L^*L)^{-1}$ when we write (1.2) later in the form

$$(Lf, Lg)_{L_2(\Omega)} = (f, g)_{\Phi} \tag{1.3}$$

with the inner product of the native space forming the right-hand side.

Since we have $C = (L^*L)^{-1}$ at hand and want to construct L, we have to form the "square root" of the operator C and invert it to get L. Taking the square root requires nonnegativity of C in the sense

of integral operators. This is a property that is intimately related to (strict) positive definiteness of the kernel Φ , and thus in Section 6 we take a closer look at the relation of these two notions.

This provides in Section 7 a very useful link to the theory of integral operators and their eigenfunction expansions. We use these expansions as a common replacement for different transforms or expansions in the classical variations of the theory. Consequently, we get new characterizations of native spaces via such expansions in Section 8, and we end up with new construction techniques for positive definite functions in Section 10.

In between, Section 5 will provide a first application of the technique we develop here: we can generalize a proof of an increased convergence order, replacing Fourier transforms by eigenfunction expansions. Section 9 contains some facts for understanding certain phenomena related to numerical stability and regularization.

2. Basic definitions

For the convenience of the reader, we collect some basics on symmetric positive definite basis functions and their native spaces from Section 4 of [8].

Definition 2.1. A function $\Phi: \Omega \times \Omega \to \mathbb{R}$ is symmetric and (strictly) positive definite (SPD), if for arbitrary finite sets $X = \{x_1, \ldots, x_M\} \subseteq \Omega$ of distinct points the matrix

 $A_{\Phi,X} = (\Phi(x_k, x_j))_{1 \le j,k \le M}$

is symmetric and positive definite.

Theorem 2.2. Every SPD function Φ on some domain Ω has a unique native Hilbert space $\mathscr{F}_{\Phi}(\Omega)$. It is the closure of the space

$$F_{\Phi}(\Omega) := \left\{ \sum_{j=1}^{M} \lambda_j \Phi(x_j, \cdot) : \lambda_j \in \mathbb{R}, \ M \in \mathbb{N}, \ x_j \in \Omega \right\}$$
(2.1)

of all functions of form (1.1) under the inner product

$$(\Phi(x,\cdot),\Phi(y,\cdot))_{\Phi} = \Phi(x,y) \quad for \ all \ x, y \in \Omega.$$

$$(2.2)$$

The elements of the native space can be interpreted as functions via the reproduction formula

$$f(x) = \delta_x(f) = (f, \Phi(x, \cdot))_{\phi} \quad \text{for all } x \in \Omega, f \in \mathscr{F}_{\phi}(\Omega).$$

$$(2.3)$$

Theorem 2.3. The dual $\mathscr{L}_{\Phi}(\Omega)$ of the native space is the closure of the space $L_{\Phi}(\Omega)$ spanned by all point evaluation functionals δ_x for all $x \in \Omega$ under the inner product $(.,.)_{\Phi}$ defined by

$$(\delta_x, \delta_y)_{\Phi} = \Phi(x, y) \quad \text{for all } x, y \in \Omega.$$
 (2.4)

From Section 8 of [8] we cite the following characterization of the native space due to Madych and Nelson [3].

Theorem 2.4. The space

$$\mathscr{M}_{\Phi}(\Omega) := \{ f : \Omega \to \mathbb{R} : |\lambda(f)| \leq C_f \|\lambda\|_{\Phi} \text{ for all } \lambda \in L_{\Phi}(\Omega) \}$$

coincides with the native space $\mathcal{N}_{\Phi}(\Omega)$. The norm

$$||f||_{\mathscr{M}} := \sup\{|\lambda(f)|: \lambda \in L_{\Phi}(\Omega), ||\lambda||_{\Phi} \leq 1\}$$

$$(2.5)$$

coincides with $||f||_{\Phi}$ defined via (2.2).

3. Embedding into L_2

There is an easy way to imbed a native space into an L_2 space.

Lemma 3.1. Let Φ be symmetric and positive definite (SPD) on Ω . Assume

$$C_2^2 := \int_{\Omega} \Phi(x, x) \, \mathrm{d}x < \infty. \tag{3.1}$$

Then the native Hilbert space $\mathscr{F}_{\Phi}(\Omega)$ for Φ has a continuous linear embedding into $L_2(\Omega)$ with norm at most C_2 .

Proof. For all $f \in \mathscr{F}_{\Phi}(\Omega)$ and all $x \in \Omega$ we can use (2.2) and the reproduction property (2.3) to get

$$f(x)^2 = (f, \Phi(x, \cdot))^2_{\phi}$$
$$\leqslant ||f||^2_{\phi} ||\Phi(x, \cdot)||^2_{\phi}$$
$$= ||f||^2_{\phi} \Phi(x, x).$$

This implies $\Phi(x,x) \ge 0$, and the assertion follows by integration over Ω . \Box

By the way, the above inequality shows in general how upper bounds for functions in the native space can be derived from the behaviour of Φ on the diagonal of $\Omega \times \Omega$. And, sometimes, the related geometric mean inequality

$$\Phi(x, y)^2 \leq \Phi(x, x)\Phi(y, y)$$

is useful, following directly from (2.2) or via $f_y(x) := \Phi(x, y)$ from the above argument.

4. The convolution mapping

We now go the other way round and map $L_2(\Omega)$ into the native space.

Theorem 4.1. Assume (3.1) to hold for an SPD function Φ on Ω . Then the integral operator

$$C(v)(x) := \int_{\Omega} v(t)\Phi(x,t) \,\mathrm{d}t \tag{4.1}$$

of generalized convolution type maps $L_2(\Omega)$ continuously into the native Hilbert space $\mathscr{F}_{\Phi}(\Omega)$. It has norm at most C_2 and satisfies

$$(f,v)_2 = (f,C(v))_{\phi} \quad \text{for all } f \in \mathscr{F}_{\phi}(\Omega), \ v \in L_2(\Omega),$$

$$(4.2)$$

i.e., it is the adjoint of the embedding of the native space $\mathscr{F}_{\Phi}(\Omega)$ into $L_2(\Omega)$.

Proof. We use the definition of $\mathcal{M}_{\phi}(\Omega)$ in Theorem 2.4 and pick some finitely supported functional $\lambda \in L_{\phi}(\Omega)$ to get

$$\lambda(C(v)) = \int_{\Omega} v(t) \lambda^{x} \Phi(x, t) dt$$
$$\leq \|v\|_{2} \|\lambda^{x} \Phi(x, \cdot)\|_{2}$$
$$\leq C_{2} \|v\|_{2} \|\lambda\|_{\Phi}$$

for all $v \in L_2(\Omega)$, where λ^x stands for the evaluation of λ with respect to x. In case of $f(t) := \Phi(x, t)$ with arbitrary $x \in \Omega$, Eq. (4.2) follows from the definition of the operator C and from the reproduction property. The general case is obtained by continuous extension to the full native space. \Box

We add two observations following from general properties of adjoint mappings:

Corollary 4.2. The range of the convolution map C is dense in the native Hilbert space $\mathscr{F}_{\Phi}(\Omega)$. The latter is dense in $L_2(\Omega)$ iff C is injective. \Box

To prove criteria for injectivity of *C* or, equivalently, for density of the Hilbert space $\mathscr{F}_{\Phi}(\Omega)$ in $L_2(\Omega)$, is an open problem, at least in the general situation. For SPD functions $\Phi(x, y) = \phi(x - y)$ on $\Omega = \mathbb{R}^d$ with a strictly positive *d*-variate Fourier transform $\hat{\phi}$ there is a neat argument due to Brown that does the job. In fact, if there is some $v \in L_2(\mathbb{R}^d)$ such that $(v, \Phi(x, \cdot))_{L_2(\mathbb{R}^d)} = 0$ for all $x \in \mathbb{R}^d$, then $\hat{v} \cdot \hat{\phi} = 0$ must hold on \mathbb{R}^d , and then v = 0 in $L_2(\mathbb{R}^d)$.

We finally remark that the above problem is related to the specific way of defining an SPD function via finitely supported functionals. Section 6 will shed some light on another feasible definition, and we can revisit the problem in Section 10 after we have replaced Fourier transforms by eigenfunction expansions.

5. Improved convergence results

The space $C(L_2(\Omega))$ allows an improvement of the standard error estimates for reconstruction processes of functions from native spaces. Roughly speaking, the error bound can be "squared". But we first want to describe the standard error estimate, based on material of Sections 10 and 11 of [8]. If Φ is an SPD function on Ω , one can interpolate any function f from the native space $\mathscr{F}_{\Phi}(\Omega)$ on any scattered set $\{x_1, \ldots, x_M\} \subset \Omega$ by a unique function s_f^* of form (1.1). The error functional

$$\varepsilon_x^* : f \mapsto f(x) - s_f^*(x)$$

is in the dual of the native space, and its norm

$$\|\varepsilon_x^*\|_{\Phi} =: P^*(x), \quad x \in \Omega$$

is called the Power Function. The standard error bound is

$$|f(x) - s_f^*(x)| \leq P^*(x) ||f - s_f^*||_{\Phi} \leq P^*(x) ||f||_{\Phi}$$
(5.1)

for all $f \in \mathscr{F}_{\Phi}(\Omega)$ and all $x \in \Omega$, and subsequent analysis (see e.g. [2,6] for an overview in the Riemannian and the radial case, respectively) usually proves uniform bounds on the power function in terms of the fill distance

 $h:=h_{X,\Omega}:=\sup_{y\in\Omega}\min_{x\in X} \|y-x\|_2.$

We now improve the error bound:

Theorem 5.1. For all $f = C(v) \in \mathscr{F}_{\Phi}(\Omega)$ with $v \in L_2(\Omega)$ we have

 $|f(x) - s_f^*(x)| \leq P^*(x) ||P^*||_2 ||v||_2$

for all $x \in \Omega$.

Proof. Taking the L_2 norm of the standard error bound (5.1) we get

 $||f - s_f^*||_2 \leq ||P^*||_2 ||f - s_f^*||_{\Phi}.$

Now we use (4.2) and the orthogonality relation from Theorem 11.3 of [8] to get

$$\begin{split} \|f - s_f^*\|_{\phi}^2 &= (f - s_f^*, f - s_f^*)_{\phi} \\ &= (f - s_f^*, f)_{\phi} \\ &= (f - s_f^*, C(v))_{\phi} \\ &= (f - s_f^*, v)_2 \\ &\leqslant \|f - s_f^*\|_2 \|v\|_2 \\ &\leqslant \|P^*\|_2 \|f - s_f^*\|_{\phi} \|v\|_2. \end{split}$$

Cancelling $||f - s_f^*||_{\Phi}$ and inserting the result into the error bound (5.1) proves the assertion. \Box

An earlier version of this result, based on Fourier transforms and restricted to functions on $\Omega = \mathbb{R}^d$ was given in [9]. Note that Theorem 5.1 holds only for functions in the range of the convolution map *C*, i.e., in a subspace of the native space. The study of the range of *C* is a challenging task, because there are numerical reasons to suggest that certain boundary effects are involved. We shall come back to this issue in Section 9.

6. Positive integral operators

We now look at the operator C from the point of view of integral equations. The compactness of C as an operator on $L_2(\Omega)$ will be delayed somewhat, because we first want to relate our definition

of a positive definite function to that of a positive integral operator. The latter property will be crucial in later sections.

Definition 6.1. An operator C of form (4.1) is *positive* (*nonnegative*), if the bilinear form

 $(w, C(v))_2, v, w \in L_2(\Omega)$

is symmetric and positive (nonnegative) definite on $L_2(\Omega)$.

In our special situation we can write

 $(w, C(v))_2 = (C(w), C(v))_{\Phi}, \quad v, w \in L_2(\Omega)$

and get

Theorem 6.2. If a symmetric and positive semidefinite function Φ on Ω satisfies (3.1), then the associated integral operator C is nonnegative. If this holds, positivity is equivalent to injectivity.

Theorem 6.3. Conversely, if C is a nonnegative integral operator of form (3.1) with a symmetric and continuous function $\Phi : \Omega \times \Omega \to \mathbb{R}$, then Φ is positive semidefinite on Ω .

Proof. We simply approximate point evaluation functionals δ_x by functionals on $L_2(\Omega)$ that take a local mean. Similarly, we approximate finitely supported functionals by linear combinations of the above form. The rest is standard, but requires continuity of Φ . \Box

Unfortunately, the above observations do not allow to conclude positive definiteness of Φ from positivity of the integral operator C. It seems to be an open problem to bridge this gap. However, due to the symmetry of Φ , the integral operator C is always self-adjoint.

7. Eigenfunction expansions

To apply strong results from the theory of integral equations, we still need that C is compact on $L_2(\Omega)$. This is implied (see, e.g., [5]) by the additional condition

$$\int_{\Omega} \int_{\Omega} \Phi(x, y)^2 \, \mathrm{d}x \, \mathrm{d}y < \infty, \tag{7.1}$$

which is automatically satisfied if our SPD function Φ is continuous and Ω is compact. Note the difference to (3.1), which is just enough to ensure embedding of the native space into $L_2(\Omega)$. Note further that (7.1) rules out certain familiar cases like the Gaussian on \mathbb{R}^d . It is an open problem to handle this situation, and here may be a subtle difference between working on bounded or unbounded domains.

From now on, we assume Φ to be an SPD kernel satisfying (3.1) and (7.1). Then C is a compact self-adjoint nonnegative integral operator. Now spectral theory and the theorem of Mercer [5] imply

the following facts:

1. There is a finite or countably infinite set of positive real eigenvalues $\rho_1 \ge \rho_2 \ge \cdots > 0$ and eigenfunctions $\varphi_1, \varphi_2, \ldots \in L_2(\Omega)$ such that

$$C(\varphi_n) = \rho_n \varphi_n, \quad n = 1, 2, \dots$$

- 2. The eigenvalues ρ_n converge to zero for $n \to \infty$, if there are infinitely many.
- 3. There is an absolutely and uniformly convergent representation

$$\Phi(x, y) = \sum_{n} \rho_n \varphi_n(x) \varphi_n(y), \quad x, y \in \Omega.$$
(7.2)

- 4. The functions φ_n are orthonormal in $L_2(\Omega)$.
- 5. Together with an orthonormal basis of the kernel of C, the functions φ_n form a complete orthonormal system in $L_2(\Omega)$.
- 6. There is a nonnegative self-adjoint operator $\sqrt[4]{C}$ such that $C = \sqrt[4]{C} \sqrt[4]{C}$ and with an absolutely and uniformly convergent kernel representation

$$\sqrt[*]{\Phi}(x,y) := \sum_{n} \sqrt{\rho_n} \varphi_n(x) \varphi_n(y), \quad x, y \in \Omega,$$
(7.3)

where

$$\sqrt[*]{C}(v)(x) := \int_{\Omega} v(t) \sqrt[*]{\Phi}(x,t) \,\mathrm{d}t, \quad x \in \Omega, \ v \in L_2(\Omega).$$

We use the symbol $\sqrt[3]{\Phi}$ to denote the "convolution square-root", because

$$\Phi(x,y) = \int_{\Omega} \sqrt[s]{\Phi}(x,t) \sqrt[s]{\Phi}(t,y) dt$$
(7.4)

is a generalized convolution. We remark that this equation can be used for construction of new positive-definite functions by convolution, and we provide details in Section 10.

The situation of finitely many eigenvalues cannot occur for the standard case of continuous SPD kernels on bounded domains with infinitely many points and linearly independent point evaluations. Otherwise, the rank of matrices of the form $(\Phi(x_j, x_k))_{1 \le j,k \le N}$ would have a global upper bound.

8. The native space revisited

The action of C on a general function $v \in L_2(\Omega)$ can now be rephrased as

$$C(v) = \sum_{n} \rho_n \varphi_n(v, \varphi_n)_2$$

and it is reasonable to define an operator L such that $(L^*L)^{-1} = C$ formally by

$$L(v) = \sum_{n} (\rho_n)^{-1/2} \varphi_n(v, \varphi_n)_2.$$
(8.1)

We want to show that this operator nicely maps the native space into $L_2(\Omega)$ as required for (1.3), but for this we first have to characterize functions from the native space in terms of expansions with respect to the functions φ_n . **Theorem 8.1.** The native space for an SPD function Φ which generates a nonnegative compact integral operator on $L_2(\Omega)$ can be characterized as the space of functions $f \in L_2(\Omega)$ with $L_2(\Omega)$ -expansions

$$f = \sum_{n} \varphi_{n}(f, \varphi_{n})_{2}$$

such that the additional summability condition

$$\sum_{n} \frac{(f,\varphi_n)_2^2}{\rho_n} < \infty$$

holds.

Proof. We first show that on the subspace $C(L_2(\Omega))$ of the native space $\mathscr{F}_{\Phi}(\Omega)$ we can rewrite the inner product as

$$(C(v), C(w))_{\phi} = (v, C(w))_{2}$$

= $\sum_{n} (v, \varphi_{n})_{2} (C(w), \varphi_{n})_{2}$
= $\sum_{n} \frac{(C(v), \varphi_{n})_{2} (C(w), \varphi_{n})_{2}}{\rho_{n}}$

But this follows from $(C(v), \varphi_n)_2 = \rho_n(v, \varphi_n)_2$ for all $v \in L_2(\Omega)$. Since $C(L_2(\Omega))$ is dense in $\mathscr{F}_{\phi}(\Omega)$ due to Corollary 4.2, and since $\mathscr{F}_{\phi}(\Omega)$ is embedded into $L_2(\Omega)$, we can rewrite the inner product on the whole native space as

$$(f,g)_{\Phi} = \sum_{n} \frac{(f,\varphi_{n})_{2}(g,\varphi_{n})_{2}}{\rho_{n}} \quad \text{for all } f,g \in \mathscr{F}_{\Phi}(\Omega).$$

$$(8.2)$$

The rest is standard. \Box

Corollary 8.2. The functions $\sqrt{\rho_n} \varphi_n$ are a complete orthonormal system in the native space $\mathscr{F}_{\Phi}(\Omega)$.

Proof. Orthonormality immediately follows from (8.2), and Theorem 8.1 allows to rewrite all functions from the native space in the form of an orthonormal expansion

$$f = \sum_{n} (f, \sqrt{\rho_n} \varphi_n)_{\Phi} \sqrt{\rho_n} \varphi_n$$

with respect to the inner product of the native space. \Box

Corollary 8.3. The operator L defined in (8.1) maps the native space $\mathscr{F}_{\Phi}(\Omega)$ into $L_2(\Omega)$ such that (1.2) holds. It is an isometry between its domain $\mathscr{F}_{\Phi}(\Omega)$ and its range $L_2(\Omega)/\ker C = \operatorname{clos}(\operatorname{span}\{\varphi_n\}_n)$.

Corollary 8.4. The operator $\sqrt[*]{C}$ defined in (7.3) maps $L_2(\Omega)$ onto the native space $\mathscr{F}_{\Phi}(\Omega)$. Its inverse on $\mathscr{F}_{\Phi}(\Omega)$ is L. Any function f in the native space has the integral representation

$$f = \int_{\Omega} v(t) \sqrt[4]{C}(\cdot, t) dt$$
(8.3)

with a function $v \in L_2(\Omega)$.

Corollary 8.5. The range of the mapping C consists of the functions f in $L_2(\Omega)$ such that the summability condition

$$\sum_{n} \frac{(f, \varphi_n)_2^2}{\rho_n^2} < \infty$$

holds.

Note that the above series of corollaries reach the goals we set ourselves in the introduction. It is an interesting open problem to generalize results for the radial case on $\Omega = \mathbb{R}^d$ to this setting, replacing Fourier transforms by eigenfunction expansions.

The operator *L* will in general not be a classical differential operator as in the spline case. For certain functions $\phi(x - y) = \Phi(x, y)$ on \mathbb{R}^d it is a pseudodifferential operator, e.g., for the Gaussian. Specific analytical arguments in the work of Duchon [1] yield optimal error bounds, but they seem to rely heavily on *L* being a classical differential operator. It is a difficult open problem to generalize those results.

9. Implications for numerical techniques

The reconstruction of a fairly general function f on Ω from function values $f(x_k)$ on centers $\{x_1, \ldots, x_M\}$ via a function

$$s_f^*(x) := \sum_{j=1}^M \lambda_j \Phi(x_j, x)$$

of form (1.1) is usually provided by interpolation, i.e., by a solution of the system

$$f(x_k) = \sum_{j=1}^M \lambda_j \Phi(x_j, x_k), \quad 1 \le k \le M$$
(9.4)

for the coefficients λ_j . We now look at this numerical problem from the viewpoint of integral operators, and our goal is to show that we get some new hints for further research.

In view of Corollary 8.4 and (8.3) we can write

$$\int_{\Omega} v(t) \sqrt[4]{C}(x_k, t) \, \mathrm{d}t = \int_{\Omega} \sqrt[4]{C}(x_k, t) \sum_{j=1}^{M} \lambda_j \sqrt[4]{C}(x_j, t) \, \mathrm{d}t,$$

to see that we are recovering v from the functions $\sqrt[n]{C}(x_j, t)$ via best approximation in $L_2(\Omega)$. The coefficients λ_j in system (9.4) have a natural interpretation via the approximation

$$v(t) \approx \sum_{j=1}^{M} \lambda_j \sqrt[s]{C}(x_j, t).$$

The above argument is a simple implication of the fact that all functions f from the native space are solutions of the operator equation

$$f = \sqrt[*]{C(v)}, \quad v \in L_2(\Omega).$$

Since this is (under certain assumptions) an integral equation of the first kind, numerical problems will automatically arise whenever the function f is not in the range of the operator $\sqrt[*]{C}$, i.e., if f is not in the native space. But we see what actually happens: the numerical process is a best approximation in $L_2(\Omega)$ with respect to the functions $\sqrt[*]{C}(x_j, t)$ and thus *always* numerically executable. The above argument also sheds some light on why in [7] the treatment of functions f outside the native space actually worked after truncation of the Fourier transform. The applied technique suitably regularizes the ill-posed integral equation problem, and it still guarantees optimal approximation orders for given smoothness of f.

We now make things worse and turn to the operator equation

$$f = C(v), \quad v \in L_2(\Omega).$$

Again, this is an integral equation of the first kind, and its solvability requires that f be in the range of C. This is precisely the situation of Theorem 5.1, and we get some explanation for the improved convergence rate. The interpretation of the coefficients λ_i in system (9.4) now is somewhat different:

$$f(x_k) = \int_{\Omega} v(t) \Phi(x_k, t) \, \mathrm{d}t = \sum_{j=1}^M \lambda_j \Phi(x_j, x_k),$$

makes it reasonable to compare with a quadrature formula

$$\int_{\Omega} g(t) \, \mathrm{d}t \approx \sum_{j=1}^{M} \beta_j g(x_j)$$

to arrive at

$$\lambda_i \approx \beta_i v(x_i).$$

This implies that for smooth f and fairly regular configurations the coefficients at nearby points should be similar, and it provides a first technique to prolong values of coarse approximations to coefficients regarding finer center distributions. This observation (in a somewhat different form) was made by Wenz [11].

Another possible progress from here is the investigation of multilevel techniques, taking the eigensystem of C into account. Research in this direction is currently going on.

10. Construction of positive-definite functions

We now know that many strictly positive definite functions Φ on a domain Ω induce positive integral operators in $L_2(\Omega)$ and have a representation (7.2). But we can turn things upside down and *define* Φ by (7.2), starting with a complete orthonormal system $\{\varphi_n\}_n$ in $L_2(\Omega)$ and a sequence $\{\rho_n\}_n$ of nonnegative numbers, converging to zero. In some sense, this approach is more general than the original one, because discontinuous or singular functions may result, depending on the decay of ρ_n for $n \to \infty$. Furthermore, the orthonormal systems arising from eigenfunction expansions are somewhat special, because they often are smoother than general L_2 functions. We thus have to expect a wider class of functions Φ when starting from (7.2). To actually carry out the construction, we first observe that Φ defined by (7.2) is a generalized positive semidefinite function in the sense that

$$(\lambda,\mu)_{\Phi} := \sum_{n} \rho_n \lambda(\varphi_n) \mu(\varphi_n) \tag{10.5}$$

is a continuous bilinear form on the dual of $L_2(\Omega)$. We cannot use the standard definition, because point evaluations are not continuous. Note here that for any functional λ in the dual of $L_2(\Omega)$ we have

$$\|\lambda\|_2^2 = \sum_n \lambda(\varphi_n)_2^2 < \infty$$

and thus can bound the bilinear form by

$$(\lambda,\mu)_{\Phi}^2 \leq \left(\sum_n \sqrt{\rho_n}\lambda(\varphi_n)^2\right) \left(\sum_n \sqrt{\rho_n}\mu(\varphi_n)^2\right).$$

The bilinear form is an inner product, if all ρ_n are positive. Now we can define the future native space via Theorem 8.1 and provide it with the bilinear form (8.2). The Riesz map $R_{\phi,\Omega}$ comes out to be

$$R_{\Phi,\Omega}(\lambda) = \lambda^{x} \Phi(x, \cdot) = \sum_{n} \rho_{n} \lambda(\varphi_{n}) \varphi_{n}$$

as expected, and the dual of the native space will be the closure of all functionals λ in the dual of $L_2(\Omega)$ under the inner product (10.5). Naturally, the dual of the native space will be larger than the dual of $L_2(\Omega)$, i.e. $L_2(\Omega)$ itself.

If some of the ρ_n are zero, we see that we get something like a generalized conditionally positive-definite case, and regularization of the kernel along the lines of Section 6 of [8] just does the right thing. Finally, it now is somewhat more clear why conditions for injectivity of *C* are nontrivial: one may be in a situation where some of the ρ_n are zero, and then everything has to be done modulo the kernel of *C* or, equivalently, the span of the φ_n with $\rho_n = 0$.

A look at (7.4) reveals another technique to construct positive semidefinite functions. In fact, if some function $P: \Omega \times \Omega \to \mathbb{R}$ has the property $P(x, \cdot) \in L_2(\Omega)$ for all $x \in \Omega$, we can form the generalized convolution

$$\Phi(x, y) := \int_{\Omega} P(x, t) P(y, t) \, \mathrm{d}t.$$

The two construction techniques of this section have not yet been exploited to generate new and interesting basis functions. For the radial case, a toolbox was provided by [10], but there is no generalization so far.

Acknowledgements

Special thanks go to Holger Wendland for careful proofreading and various improvements of a preliminary version of this paper.

References

- J. Duchon, Interpolation des fonctions de deux variables suivant le principe de la flexion des plaques minces, Rev. Française Automat. Informat. Rech. Opér. Anal. Numer. 10 (1976) 5–12.
- [2] N. Dyn, F. Narcowich, J. Ward, Variational principles and Sobolev-type estimates for generalized interpolation on a Riemannian manifold, Constr. Approx. 15 (1999) 175–208.
- [3] W.R. Madych, S.A. Nelson, Multivariate Interpolation and conditionally positive definite functions, Approx. Theory Appl. 4 (1988) 77–89.
- [4] F.J. Narcowich, Generalized Hermite interpolation by positive definite kernels on a Riemannian manifold, J. Math. Anal. Appl. 190 (1995) 165–193.
- [5] D. Porter, D.G. Stirling, Integral equations: a practical treatment from spectral theory to applications, Cambridge Texts in Applied Mathematics, Cambridge University Press, Cambridge, 1990.
- [6] R. Schaback, Error estimates and condition numbers for radial basis function interpolation, Adv. Comput. Math. 3 (1995) 251–264.
- [7] R. Schaback, Approximation by radial basis functions with finitely many centers, Constr. Approx. 12 (1996) 331–340.
- [8] R. Schaback, Native spaces for radial basis functions I, in: M.W. Müller, M.D. Buhmann, D.H. Mache, M. Felten (Eds.), New Developments in Approximation Theory, International Series of Numerical Mathematics, Vol. 132, Birkhäuser, Basel, 1999, pp. 255–282.
- [9] R. Schaback, Improved error bounds for radial basis function interpolation, Math. Comput. 68 (1999) 201-216.
- [10] R. Schaback, Z. Wu, Operators on radial basis functions, J. Comput. Appl. Math. 73 (1996) 257-270.
- [11] J. Wenz, private communication 1998.


Journal of Computational and Applied Mathematics 121 (2000) 179-195

JOURNAL OF COMPUTATIONAL AND APPLIED MATHEMATICS

www.elsevier.nl/locate/cam

Constraint-selected and search-optimized families of Daubechies wavelet filters computable by spectral factorization

Carl Taswell

Computational Toolsmiths, P.O. Box 18925, Stanford, CA 94309-8925, USA

Received 10 December 1999; received in revised form 12 February 2000

Abstract

A unifying algorithm has been developed to systematize the collection of compact Daubechies wavelets computable by spectral factorization of a symmetric positive polynomial. This collection comprises all classes of real and complex orthogonal and biorthogonal wavelet filters with maximal flatness for their minimal length. The main algorithm incorporates spectral factorization of the Daubechies product filter into analysis and synthesis filters. The spectral factors are found for search-optimized families by examining a desired criterion over combinatorial subsets of roots indexed by binary codes, and for constraint-selected families by imposing sufficient constraints on the roots without any optimizing search for an extremal property. Daubechies wavelet filter families have been systematized to include those constraint-selected by the principle of separably disjoint roots, and those search-optimized for time-domain regularity, frequency-domain selectivity, time-frequency uncertainty, and phase nonlinearity. The latter criterion permits construction of the least and most asymmetric real and complex orthogonal filters. Biorthogonal symmetric spline and balanced-length filters with linear phase are also computable by these methods. This systematized collection has been developed in the context of a general framework enabling evaluation of the equivalence of constraint-selected families have been demonstrated to be equivalent to some of the search-optimized families, thereby obviating the necessity for any search in their computation. (c) 2000 Elsevier Science B.V. All rights reserved.

Keywords: Multirate filter design; Wavelets and filter banks

1. Introduction

Since the discovery of compact orthogonal and biorthogonal wavelets by Daubechies, various discussions of the general theory and specific parameterizations of her wavelets have also been published (cf. [2,5,12,16] for literature reviews). These compact Daubechies wavelets, which have

E-mail address: ctaswell@toolsmiths.com (C. Taswell).

^{0377-0427/00/\$ -} see front matter \odot 2000 Elsevier Science B.V. All rights reserved. PII: \$0377-0427(00)00340-X\$

the maximal number of vanishing moments for their minimal length, can be implemented as discrete filters that are iterated or auto-convolved to generate approximations of the continuous functions.

The Daubechies wavelet filters can be readily computed via spectral factorization of a symmetric positive polynomial [1]. Significant advantages of the spectral factorization approach include its generalizability to many different classes and families of wavelets, its suitability for easily interpretable visual displays, and thus its practicality in pedagogy. All of the complex orthogonal, real orthogonal, and real biorthogonal families of the Daubechies class computable by spectral factorization and constructed with a single unifying computational algorithm have been studied experimentally in the systematized collection developed by Taswell [10-12,15-17] over a wide range of vanishing moment numbers and filter lengths.

In contrast, angular parameterization methods have usually been demonstrated for wavelets with only one vanishing moment (i.e., less than maximal flatness) and very short lengths [9] with the exception of [13]. But the latter only verified orthogonality and vanishing moment numbers for the filters and did not attempt any search through the angular parametrization space for filters with desirable properties.

These comments highlight one of the essential questions in the development of an algorithm for the design of wavelet filters: How much computational effort should be expended in the construction of a wavelet filter possessing which properties over which range of filter lengths? A basic assumption inherent in the systematized collection of Daubechies wavelets [11,15,17] hypothesizes that the spectral factorization approach affords the most economical generation of wavelet filters with the best variety and combination of properties over the widest range of filter lengths.

The economy of the spectral factorization method in comparison with the angular parameterization method is achieved by the reduced size of the search space for the filter root codes [16] relative to that for the filter coefficient angles [9]. In [16], conjectures were made regarding schemes to enhance the efficiency of the combinatorial search used in the design algorithm. In [17], a new design principle was introduced within a general framework to demonstrate that the search can be completely eliminated for those search-optimized filter families for which equivalence has been demonstrated with constraint-selected filter families. This survey reviews the development of the systematized collection of Daubechies wavelets and summarizes the essential computational methods.

2. General framework

Consider a filter expressed as the complex z-domain polynomial F(z) with corresponding vectors for the roots $z \equiv [z_j] \in \mathscr{Z}$ and the coefficients $f \equiv [f_n] \in \mathscr{F}$. Associated with F(z), assume there exist three parameters, vectors $\gamma \in \Gamma$, $\xi \in \Xi$, and scalar $\lambda \in \Lambda$, respectively, that index the filter within a set of such filters forming a defined family, specify each indexed filter of the family within a search space, and characterize its properties.

Applying this notation to the orthonormal Daubechies [1] and Rioul [7] wavelets, $\gamma \equiv [\gamma_1, \gamma_2] = [N, K]$ represents the number K of vanishing moments for wavelet filters of length N = 2K and N > 2K, respectively. For angle space methods [9] to generate orthonormal wavelets, ξ represents the set of angles that specifies f for F(z). For binomial space methods [16] to generate Daubechies wavelets, ξ represents the set of binary codes that specifies z for F(z). In both cases, λ represents a criterion obtained from an individual property or a weighted combination of properties computed

from z and/or f (such as the filter's time-domain regularity [14], phase nonlinearity [16], etc.) that characterizes F(z).

Thus, γ and ξ determine F(z) and then F(z) determines λ with the mapping of spaces $\Gamma \times \Xi \mapsto \mathscr{F} \times \mathscr{Z} \mapsto \Lambda$. The parameters γ and ξ that determine F(z) are called the *indexing parameter* and *specification parameter*, respectively. The parameter λ that is determined by F(z) is called the *characterization parameter*. If λ represents an individual property (rather than weighted combination of properties), then λ is also termed a *characteristic property* of F(z).

2.1. Existence and uniqueness

Given a defined filter family $\{F_{\gamma}(z)\}$ indexed by γ , assume for fixed γ that a finite sequence of filters $F_{\gamma,i}(z)$ indexed by *i* can be generated by and evaluated for corresponding sequences, respectively, of specification parameters ξ_i and characterization parameters λ_i . If Ξ is an unbounded or continuous space, then it can be appropriately bounded and discretized to permit a countably finite sequence ξ_i .

Assuming restriction to a countably finite space Ξ , then the corresponding spaces $\mathscr{F} \times \mathscr{Z}$ and Λ are also countably finite. Further assuming a one-to-one invertible mapping and uniqueness of the elements $\lambda_i \in \Lambda$ (achieved if necessary by the use of "tie-breaker" rules for the definition of the characterization parameter λ), then finite countability of unique elements for an invertible mapping implies that it is feasible to search for both elements $\underline{\lambda} \equiv \min_i \lambda_i$ and $\overline{\lambda} \equiv \max_i \lambda_i$ in the range and select the corresponding filters $F_{\gamma,i}(z)$ in the domain.

2.2. Definitions and inferences

A filter F(z) is called *extremal* if it can be shown to possess a characterization parameter attaining an extreme manifested by either $\underline{\lambda}$ or $\overline{\lambda}$. A filter F(z) is called *search optimized* if it is generated by an algorithm that optimizes $\lambda \in \Lambda$ with an exhaustive search to ensure identification of either $\underline{\lambda}$ or $\overline{\lambda}$. A filter F(z) is called *constraint selected* if it is generated by an algorithm that specifies sufficient constraints on $\boldsymbol{\xi}, \boldsymbol{f}$, or \boldsymbol{z} to ensure uniqueness of F(z) and selection of F(z) without a search. An indexed set of filters $\{F_{\gamma}(z)\} \equiv \{F(z;\gamma): \gamma \in \Gamma\}$ is called a *family* if all members of the set are generated by the *same* algorithm, a function $g(\boldsymbol{\xi};\gamma), g(\boldsymbol{f};\gamma),$ or $g(z;\gamma)$, subject to the control of the indexing parameter γ .

Two different filter families $\{F_{\gamma}(z)\}$ and $\{F'_{\gamma}(z)\}$ generated by two different algorithms $g(\cdot;\gamma)$ and $g'(\cdot;\gamma)$ are \mathscr{F} -equivalent, or equivalent with respect to (w.r.t.) the filter coefficient space \mathscr{F} , if $||f_{\gamma} - f'_{\gamma}|| < \tau$ for all $\gamma \in \Gamma$ with given error tolerance $\tau(\mathscr{F})$. Analogously, $\{F_{\gamma}(z)\}$ and $\{F'_{\gamma}(z)\}$ are \mathscr{L} -equivalent, or equivalent w.r.t. the filter root space \mathscr{L} , if $||z_{\gamma} - z'_{\gamma}|| < \tau$ for all $\gamma \in \Gamma$ with given error tolerance $\tau(\mathscr{L})$. Finally, they are Λ -equivalent, or equivalent w.r.t. the characterization parameter space Λ , if $|\lambda_{\gamma} - \lambda'_{\gamma}| < \tau$ for all $\gamma \in \Gamma$ with given error tolerance $\tau(\Lambda)$.

A search-optimized filter is necessarily an extremal filter, whereas a constraint-selected filter may or may not be an extremal filter. If a constraint-selected filter can be shown to be equivalent to a search-optimized filter, then the constraint-selected filter is also an extremal filter. Both \mathscr{F} -equivalence and \mathscr{Z} -equivalence of two different filter families imply Λ -equivalence, but the converse is not true.

3. Daubechies polynomials

The generation of Daubechies wavelet filter families computable by spectral factorization of the Daubechies polynomials requires a separate algorithm for computing the roots of the product filter

$$\mathscr{P}_{D}(z) = (z+1)^{2(D+1)} \mathscr{Q}_{D}(z)$$
(1)

or its related form the quotient filter

$$\mathcal{Q}_D(z) = (z+1)^{-2(D+1)} \mathscr{P}_D(z), \tag{2}$$

which is a Laurent polynomial of degree $d_2 = D$ with 2D roots. Both forms are indexed by the integer parameter $D \ge 0$.

Consider mappings $x \to y \to z$ between three planes in the complex variables x, y, and z. Use the x plane to find the roots of the conditioned polynomial $\mathscr{C}_D(x)$, map to the y plane for the roots of the binomial polynomial $\mathscr{B}_D(y)$, and map again to the z plane for the roots of the quotient polynomial $\mathscr{Q}_D(z)$. All three polynomials $\mathscr{C}_D(x)$, $\mathscr{B}_D(y)$, and $\mathscr{Q}_D(z)$ are considered related forms of $\mathscr{P}_D(z)$ called the conditioned, binomial, and quotient forms, respectively.

The quotient form $\mathcal{Q}_D(z)$ derives simply from division of the product form $\mathcal{P}_D(z)$ by all of its roots at z = -1. The binomial form [2, Eq. (6.1.12); 8, Eq. (1); 3, Eq. (1.7)]

$$\mathscr{B}_{D}(y) = \sum_{i=0}^{D} {D+i \choose i} y^{i}$$
(3)

derives from the binomial series for $(1 - y)^{-(D+1)}$ truncated at D + 1 terms. These forms can be related through conformal mappings (see below).

To improve the numerical conditioning of the root finding problem for the roots y_i of $\mathscr{B}_D(y)$, Shen and Strang [8] recommended the change of variables $x = \kappa y$ with $\kappa = 4$, while Goodman et al. [3] recommended the change of variables x=1/y. Incorporating both transformations with $x=1/(\kappa y)$, then

$$\mathcal{B}_{D}(y) = \sum_{i=0}^{D} {D+i \choose i} y^{i}$$
$$= (\kappa y)^{D} \sum_{i=0}^{D} \kappa^{-i} {D+i \choose i} (\kappa y)^{i-D}$$
$$= x^{-D} \mathscr{C}_{D}(x)$$

yields the conditioned form

$$\mathscr{C}_D(x) = \sum_{i=0}^{D} \kappa^{-i} \left(\frac{D+i}{i} \right) x^{D-i}.$$
(4)

Now obtain the *D* roots x_i of $\mathscr{C}_D(x)$ by computing the eigenvalues of the companion matrix. Then the *D* roots y_i of the binomial form $\mathscr{B}_D(y)$ can be calculated simply as $y_i = 1/(\kappa x_i)$.

With another change of variables $z + z^{-1} = 2 - 4y$ as described by Daubechies [1,2], map the binomial form $\mathscr{B}_D(y)$, a regular polynomial with D roots, to the quotient form $\mathscr{D}_D(z)$, a Laurent

polynomial with 2D roots. Given the Joukowski transformations [4, Vol. 1, pp. 197, 223]

$$w = f(z) = (z + z^{-1})/2,$$
(5)

$$z = f^{-1}(w) = w \pm \sqrt{w^2 - 1}$$
(6)

and the affine transformations

$$y = g(w) = (1 - w)/2,$$
 (7)

$$w = g^{-1}(y) = 1 - 2y, \tag{8}$$

then the composite mappings¹ yield the explicit solutions

$$y = g(f(z)) = (1 - (z + z^{-1})/2)/2,$$
(9)

$$z = f^{-1}(g^{-1}(y)) = 1 - 2y \pm \sqrt{(1 - 2y)^2 - 1}.$$
(10)

The latter equation yields a doubly valued solution with the reciprocal pair $\{z, z^{-1}\}$. When the pairs are regrouped as complex quadruplets $\{z, z^{-1}, \overline{z}, \overline{z}^{-1}\}$ and factors $\mathscr{U}(z; z_i) \equiv (z - z_i)(z - z_i^{-1})(z - \overline{z_i})(z - \overline{z_i})(z - \overline{z_i}^{-1})$ with any real duplets $\{r, r^{-1}\}$ and factors $\mathscr{V}(z; r_j) \equiv (z - r_j)(z - r_j^{-1})$, the Daubechies product polynomial $\mathscr{P}_D(z)$ expressed in regular form can be factored as

$$\mathscr{P}_{D}(z) = (z+1)^{2(D+1)} \prod_{i=1}^{n^{cq}} \mathscr{U}(z;z_{i}) \prod_{j=1}^{n^{d}} \mathscr{V}(z;r_{j}),$$
(11)

where $n^{cq} = \lfloor D/2 \rfloor$ and $n^{rd} = D \mod 2$. For further details on the numerical performance of these methods, refer to [12,16].

4. Spectral factorization rules

For an arbitrary polynomial $\mathscr{F}(z)$ with length N coefficients, there are N-1 roots of which $0 \le K \le N-1$ may be at z=-1. When considering spectral factorization, the product filter polynomial $\mathscr{P}_D(z)$ with $N_p = 4D + 3$ coefficients and $K_p = 2D + 2$ roots at z = -1 is factored into the analysis and synthesis filter polynomials $\mathscr{A}(z)$ and $\mathscr{P}(z)$ with N_a and N_s coefficients, and K_a and K_s roots at z = -1, respectively. This factorization yields the constraints

$$N_{\rm p} = N_{\rm a} + N_{\rm s} - 1, \tag{12}$$

$$K_{\rm p} = K_{\rm a} + K_{\rm s},\tag{13}$$

on the lengths of the three filters and their roots at z = -1. Each family of filters described in subsequent sections has been named with an identifying acronym followed by (N; K) in the orthogonal

¹ Unlike other sections where f and g may denote filters or arbitrary functions, here f and g denote functions that are conformal maps in the complex domain.

cases for which

$$N = N_{\rm a} = N_{\rm s},\tag{14}$$

$$K = K_a = K_s \tag{15}$$

is required, and by $(N_a, N_s; K_a, K_s)$ in the biorthogonal cases for which

$$N_{\rm a} = K_{\rm a} + 4n_{\rm a}^{\rm cq} + 2n_{\rm a}^{\rm rd} + 1, \tag{16}$$

$$N_{\rm s} = K_{\rm s} + 4n_{\rm s}^{\rm cq} + 2n_{\rm s}^{\rm rd} + 1,\tag{17}$$

$$N_{\rm p} = 2K_{\rm p} - 1 \tag{18}$$

is required. Here n_a^{cq} , n_s^{cq} , n_a^{rd} , and n_s^{rd} are the numbers of complex quadruplet factors $\mathscr{U}(z;z_i)$ and real duplet factors $\mathscr{V}(z;r_j)$ for each of $\mathscr{A}(z)$ and $\mathscr{G}(z)$. Both n^{cq} and n^{rd} may be whole or half-integer. In the latter case, half of a complex quadruplet and half of a complex duplet denote, respectively, a complex duplet and a real singlet.

For K_a and K_s necessarily both odd or both even, then K_p is always even and $K = K_p/2$ a whole integer determines $n_p^{cq} = n_a^{cq} + n_s^{cq}$ and $n_p^{rd} = n_a^{rd} + n_s^{rd}$ according to $n_p^{cq} = \lfloor (K - 1)/2 \rfloor$ and $n_p^{rd} = (K - 1) \mod 2$. If K_a and K_s are given, then K_p and K yield n_p^{cq} and n_p^{rd} split into $\{n_a^{cq}, n_a^{rd}\}$ and $\{n_s^{cq}, n_s^{rd}\}$ and the roots are factored accordingly. For real coefficients, a root z must be paired with its conjugate \bar{z} . For symmetric coefficients, a root z must be paired with its reciprocal z^{-1} . For 2-shift orthogonal coefficients, a root z must be *separated* from its conjugate reciprocal \bar{z}^{-1} .

Thus, in the real biorthogonal symmetric case, each complex quadruplet $\mathscr{U}(z;z_i)$ and real duplet $\mathscr{V}(z;r_j)$ must be assigned in its entirety to either $\mathscr{A}(z)$ or $\mathscr{S}(z)$. In the real orthogonal case, each complex quadruplet is split into two conjugate duplets $(z - z_i)(z - \bar{z}_i)$ and $(z - z_i^{-1})(z - \bar{z}_i^{-1})$, while each real duplet is split into two singlets $(z - r_j)$ and $(z - r_j^{-1})$, with one factor assigned to $\mathscr{A}(z)$ and the other to $\mathscr{S}(z)$. The complex orthogonal case is analogous to the real orthogonal case except that the complex quadruplets are split into reciprocal duplets $(z - z_i)(z - z_i^{-1})$ and $(z - \bar{z}_i)(z - \bar{z}_i^{-1})$ instead of conjugate duplets. The complex orthogonal symmetric case requires use of complex quadruplets without real duplets.

All orthogonal cases require $K = K_a = K_s = K_p/2$, $n_a^{cq} = n_s^{cq} = n_p^{cq}/2$, and $n_a^{rd} = n_s^{rd} = n_p^{rd}/2$ with $N = N_a = N_s = 2K$. Note that n_p^{rd} can only equal 0 or 1. Therefore, in biorthogonal cases, either $\{n_a^{rd} = 0, n_s^{rd} = 1\}$ or $\{n_a^{rd} = 1, n_s^{rd} = 0\}$. However, in orthogonal cases, either $\{n_a^{rd} = n_s^{rd} = n_s^{rd} = 0\}$ or $\{n_a^{rd} = n_s^{rd} = \frac{1}{2}\}$ with $\frac{1}{2}$ of a duplet denoting a singlet. For all real orthogonal cases as well as those complex orthogonal cases not involving symmetry criteria, K can be any positive integer. For the complex orthogonal least-asymmetric and most-asymmetric cases, K must be a positive even integer. For the complex orthogonal least-symmetric and most-symmetric cases, K must be a positive odd integer.

For the real biorthogonal symmetric cases, K_a and K_s must be both odd or both even. In the biorthogonal symmetric spline case, all additional roots (other than those at z = -1 with assignment determined by K_a and K_s) are assigned to the analysis filter leaving the synthesis filter as the spline filter. All other biorthogonal symmetric cases incorporate a root assignment constraint that balances the lengths of the analysis and synthesis filters such that $N_a \approx N_s$ as much as possible. For $K_a = 2i - 1$ and $K_s = 2j - 1$ both odd with $i, j \in \{1, 2, 3, ...\}$, balancing of equal filter lengths is possible. In fact,

requiring both $K_a = K_s$ and $N_a = N_s$ is also possible when $N = N_a = N_s = 2K$ with $K = K_a = K_s$ for $\{K = 1 + 4k | k = 1, 2, 3...\}$. However, for $K_a = 2i$ and $K_s = 2j$ both even, equal balancing of filter lengths N_a and N_s is not possible. The additional unbalanced roots are assigned to the analysis filter such that $N_a > N_s$ leaving the synthesis filter as the shorter filter.

5. Daubechies wavelet filter families

All filter families surveyed here are named, defined, and generated according to the conventions, notation, and methods established in [15,16] for the systematized collection of wavelet filters computable by spectral factorization of the Daubechies polynomial. However, one of the original families, named DROLD in [15], was renamed DROMD in [17] in order to achieve consistency with the more recent collection of families introduced in [17]. All of the acronyms used for the filter family names abbreviate 'D' for Daubechies as the first character, 'C' or 'R' for complex or real as the second character, 'O' or 'B' for orthogonal or biorthogonal as the third character, and then two additional characters denoting an additonal description to distinguish each family from the others.

5.1. Constraint-selected families

In addition to the spectral factorization rules (Section 4) imposing the necessary contraints for complex orthogonality, real orthogonality, and real biorthogonality, the least and most disjoint families are defined according to constraints derived from the principle of separably disjoint root sets in the complex z-domain. Consider only the roots of the quotient polynomial $\mathcal{Q}(z)$ (Eq. (2)) and split this set of roots into two sets of roots $\{z_k^a\}$ and $\{z_l^s\}$ for the analysis and synthesis filters $\mathcal{A}(z)$ and $\mathcal{S}(z)$.

These root sets from $\mathcal{Q}(z)$ must be disjoint with

$$\emptyset = \{z_k^{\mathbf{a}}\} \cap \{z_l^{\mathbf{s}}\} \tag{19}$$

(because common roots at z = -1 for both $\mathscr{A}(z)$ and $\mathscr{G}(z)$ from $\mathscr{P}(z)$ have been excluded from consideration). Now let $\{\mathscr{C}_i^a\}$ and $\{\mathscr{C}_j^s\}$ denote finite collections of open convex regions with the largest area domains that do not intersect yet still cover the sets $\{z_k^a\}$ and $\{z_l^s\}$, respectively. More precisely,

$$\bigcup_{k} z_{k}^{a} \subset \bigcup_{i} \mathscr{C}_{i}^{a}, \tag{20}$$

$$\bigcup_{l} z_{l}^{s} \subset \bigcup_{i} \mathscr{C}_{i}^{s}, \tag{21}$$

$$\emptyset = \bigcap_{i} \mathscr{C}_{i}^{a}, \tag{22}$$

$$\emptyset = \bigcap_{j} \mathscr{C}_{j}^{s}, \tag{23}$$

Acronym	$\mathscr{Q}(z) ightarrow \mathscr{A}(z)$	$\mathscr{Q}(z) ightarrow \mathscr{S}(z)$
DCOMD DROMD	$ \{ (z_j, z_j^{-1}): (r_j < 1) \land (\theta_j \ge 0) \} $ $ \{ (z_j, \bar{z}_j): r_j < 1 \} $	$ \{(z_j, z_j^{-1}): (r_j > 1) \land (\theta_j \leq 0)\} $ $ \{(z_j, \bar{z}_j): r_j > 1\} $
DRBMD DRBSS	$ \{ (z_j, \bar{z}_j, z_j^{-1}, \bar{z}_j^{-1}) \colon \theta_j < \theta^* \} $ $ \{ (z_j, \bar{z}_j, z_j^{-1}, \bar{z}_j^{-1}) \} $	$\begin{cases} (z_j, \bar{z}_j, z_j^{-1}, \bar{z}_j^{-1}): \theta_j > \theta^* \\ \emptyset \end{cases}$

Table 1 Filter designs for some constraint-selected families with roots $z_j = r_j e^{i\theta_j}$

$$\emptyset = \left(\bigcup_{i} \mathscr{C}_{i}^{a}\right) \cap \left(\bigcup_{j} \mathscr{C}_{j}^{s}\right).$$
(24)

Finally, let C denote the cardinality of the set

$$\{\mathscr{C}_{i}^{a}: i = 1, \dots, I; \ \mathscr{C}_{j}^{s}: j = 1, \dots, J\}$$
(25)

as measured by the number C = I + J of regions covering all the roots of $\mathcal{Q}(z)$. Then root sets $\{z_k^a\}$ and $\{z_l^s\}$ are called least and most disjoint if C is, respectively, the maximum or minimum possible subject to the constraints of the spectral factorization rules imposed.

Table 1 summarizes the spectral factorizations for the DCOMD, DROMD, and DRBMD filter families designed with most disjoint (MD) root sets. The factorizations for the DCOLD, DROLD, and DRBLD filters designed with least disjoint (LD) root sets cannot be summarized as concisely. However, the corresponding algorithms order the roots by angle and impose the maximum number of alternations for the assignments in the split to $\mathscr{A}(z)$ and $\mathscr{S}(z)$. The algorithm for DRBLD was also modified to devise another family called DRBRD with regular disjoint (RD) root sets. For comparison, Table 1 also includes the spectral factorization for the DRBSS family with symmetric spline (SS) root sets.

5.2. Search-optimized families

Numerical estimates of defined filter characterization parameters λ are used as selection criteria for all other families subjected to optimization in combinatorial searches of the root sets. These criteria [14] include the phase nonlinearity pnl(\mathscr{A}), time-domain regularity tdr(\mathscr{A}), frequency-domain selectivity fds(\mathscr{A}), and time-frequency uncertainty tfu(\mathscr{A}). Most of the orthogonal families are defined by pnl(\mathscr{A}) selecting for varying degrees of asymmetry or symmetry. Work reported in [11,12,15] was later revised in [16] by the shift of the integration interval for pnl(\mathscr{A}) from [0,2 π] to [$-\pi,\pi$] and by the use of pnl(\mathscr{A}) as a "tie-breaker" criterion for families selected by the other criteria. These revisions now insure unique criterion values for each root set examined in the combinatorial search (which can be performed ignoring binary complements for orthogonal families).

Minimizing or maximizing $pnl(\mathscr{A})$ for real filters defines DROLA and DROMA, respectively, the least asymmetric (LA) and most asymmetric (MA) families. If the parity of K is ignored, then minimizing or maximizing $pnl(\mathscr{A})$ for complex filters defines DCOLN and DCOMN, respectively, the least nonlinear (LN) and most nonlinear (MN) families. Phase nonlinearity does not exist and cannot be used for the real biorthogonal families all of which are symmetric. Therefore, one of the other characterization parameters must be used as an optimization criterion. Also, these biorthogonal

Real biorthogonal	Description	Index constraint	Optimization
DRBLU	Least uncertain	Even $(K_a + K_s)$	$\min tfu(\mathscr{A})$
DRBMS	Most selective	Even $(K_a + K_s)$	$\max f ds(\mathscr{A})$
DRBMR	Most regular	Even $(K_a + K_s)$	$\max t dr(\mathscr{A})$
DRBBR	Balanced regular	Even $(K_a + K_s)$	$\max B(\operatorname{t} \operatorname{dr}(\cdot), \mathscr{A}, \mathscr{S})$
Real orthogonal	Description	Constraint	Optimization
DROLU	Least uncertain	$K \ge 1$	min tfu(𝒴)
DROMR	Most regular	$K \ge 1$	$\max t dr(\mathscr{A})$
DROLA	Least asymmetric	$K \ge 1$	$\min pnl(\mathscr{A})$
DROMA	Most asymmetric	$K \ge 1$	$\max pnl(\mathscr{A})$
Complex orthogonal	Description	Constraint	Optimization
DCOLU	Least uncertain	<i>K</i> ≥3	$\min tfu(\mathscr{A})$
DCOMR	Most regular	<i>K</i> ≥3	$\max t dr(\mathscr{A})$
DCOLS	Least symmetric	Odd $K \ge 3$	$\max pnl(\mathscr{A})$
DCOMS	Most symmetric	Odd $K \ge 3$	$\min pnl(\mathscr{A})$
DCOLA	Least asymmetric	Even $K \ge 4$	$\min pnl(\mathscr{A})$
DCOMA	Most asymmetric	Even $K \ge 4$	$\max pnl(\mathscr{A})$
DCOLN	Least nonlinear	$K \ge 3$	$\min pnl(\mathscr{A})$
DCOMN	Most nonlinear	$K \ge 3$	$\max pnl(\mathscr{A})$

 Table 2

 Filter designs for some search-optimized families

families are subjected to the length constraints determined by the principle of maximally balancing the filter lengths for both $\mathcal{A}(z)$ and $\mathcal{G}(z)$.

For all but several of the search-optimized families, the selection criterion is optimized for $\mathscr{A}(z)$. The exceptions are the DRBBR, DRBBS, and DRBBU families with balanced regular (BR), balanced selective (BS), and balanced uncertain (BU) root sets. Instead, the selection criterion is optimized for both $\mathscr{A}(z)$ and $\mathscr{G}(z)$ by maximizing a balancing measure *B* defined as

$$B(\lambda(\cdot),\mathscr{A},\mathscr{S}) = \left| \frac{\lambda(\mathscr{A}) + \lambda(\mathscr{S})}{\lambda(\mathscr{A}) - \lambda(\mathscr{S})} \right|,\tag{26}$$

where $\lambda(\cdot)$ is either tdr(\cdot), fds(\cdot), or tfu(\cdot), respectively, for DRBBR, DRBBS, and DRBBU.

Table 2 summarizes filter designs for some of the search-optimized families. The index constraints tabulated are those required to generate the defined family. However, for purposes of comparison between families in tables and figures, the definitions for all orthogonal families have been extended to begin at K = 1. For example, DCOLN(6;3) is complex as expected, but DCOLN(4;2) and DCOLN(2;1) are real. Also, note that the DCOLN family is the union of the even-indexed DCOLA and odd-indexed DCOLS families, while the DCOMN family is the union of the even-indexed DCOMA and odd-indexed DCOLS families. Complete details for the algorithms to compute each of the various selection criteria can be found elsewhere [12,14].

6. Unifying algorithm

All filter families of the systematized collection of Daubechies wavelet filters [12,16] are generated by the spectral factorization and selection of root sets (with either the predetermined constraints or the optimizing combinatorial search) incorporated in the following algorithm:

- (1) Input the identifying name FiltName for the family of filters and the indexing design parameters K_a and K_s .
- (2) Compute the numbers $K_p = K_a + K_s$, $D = K_p/2 1$, $n_p^{cq} = \lfloor D/2 \rfloor$, and $n_p^{rd} = D \mod 2$.
- (3) Compute the n_p^{cq} sets of complex quadruplet roots and the n_p^{rd} sets of real duplet roots of the quotient filter $\mathcal{Q}_D(z)$.
- (4) Access the factorization and selection rules that define the family of filters named FiltName.
- (5) Apply the rules to $\{n_p^{cq}, n_p^{rd}\}$ for the FiltName filter pair indexed by $\{K_a, K_s\}$ and compute the splitting number pairs $\{n_a^{cq}, n_s^{cq}\}$ and $\{n_a^{rd}, n_s^{rd}\}$.
- (6) If FiltName is a constraint-selected family, apply the rules to select the $4n_a^{cq} + 2n_a^{rd}$ roots for $\mathscr{A}(z)$ and the $4n_s^{cq} + 2n_s^{rd}$ roots for $\mathscr{S}(z)$ and jump to Step 11.
- (7) Sort the roots in an order convenient for the class of splitting appropriate to the type of filter. All roots of a complex quadruplet should be adjacent with duplets of the quadruplet subsorted according to conjugates or reciprocals depending on the filter type. Assign binary coded labels 0 and 1 to the first and second duplet of each quadruplet. Analogously, assign binary codes to the first and second singlet of the real reciprocal duplet if present. If biorthogonal, assign binary coded labels 0 or 1 to each of the entire quadruplets and duplets.
- (8) Generate the possible binomial subsets for these binary codes [6] subject to the imposed factorization rules and splitting numbers. For orthogonal filters, there are a total of $n_a^{cq} + n_a^{rd}$ binary selections without constraint on the bit sum, and thus $2^{n_a^{cq}+n_a^{rd}-1}$ binomial subsets ignoring complements. For biorthogonal filters, there are a total of n_p^{cq} binary selections with bit sum constrained to n_a^{cq} , and thus $\binom{n_p^{cq}}{n_q^{cq}}$ binomial subsets.
- (9) For each root subset selected by the binomial subset codes, characterize the corresponding filter by the optimization criterion appropriate for the FiltName family. These optimization criteria may be any of the numerically estimated characterization parameters λ computed from the roots z or the coefficients f.
- (10) Search all root subsets to find the one with the optimal value of the desired criterion. If necessary, apply the "tie-breaker" criterion.
- (11) Include the $K_a + K_s$ required roots at z = -1 with K_a for the optimal subset of roots intended for the analysis factor $\mathscr{A}(z)$ and with K_s for the complementary subset intended for the synthesis factor $\mathscr{S}(z)$ and compute the filter coefficients.
- (12) If FiltName is an orthogonal search-optimized family, compare the selected (primary) subset of filter roots and coefficients with its complementary subset to choose the one with minimax group delay over the interval $\omega \in [0, \pi]$ as the subset for $\mathscr{A}(z)$. If FiltName is a biorthogonal search-optimized family, compare the primary and complementary subsets only if $K_a = K_s$, $n_a^{cq} = n_s^{cq}$, and $n_a^{rd} = 0 = n_s^{rd}$ in order to choose the one with the defining criterion optimized for $\mathscr{A}(z)$.
- (13) Output roots z and coefficients f for each of $\mathcal{A}(z)$ and $\mathcal{S}(z)$.



Fig. 1. Examples of disjoint sets of Daubechies polynomial roots.

For search-optimized families, full searches of all possible combinatorial subsets should be performed for a sufficient number of values of K indexing the filter family's members in order to infer the appropriate pattern of binary codes with bit sums characterizing the family. Using such a pattern permits successful *partial* rather than *full* combinatorial searches. These partial searches provide significant reduction in computational complexity convenient for larger values of K, for example, for searches with K > 30 computed on desktop workstations current in 1999.

7. Examples and comparisons

Fig. 1 displays spectral factorizations for each of the least and most disjoint filter families at $K_a = K_s = 16$ for D = 15. Roots for $\mathscr{A}(z)$ and $\mathscr{G}(z)$ are marked with "o" and "x", respectively. As an example of the principle of minimizing and maximizing the cardinality C, observe that C = 3 for DRBMD and C = 13 for DRBLD. Note that $C \neq 2$ for DRBMD because convexity is required for each of the non-intersecting covering regions, and $C \neq 26$ for DRBLD because the largest area possible is required for each of the regions. Fig. 2 displays the wavelets corresponding to $\mathscr{A}(z)$ for the six examples in Fig. 1. Both the real parts (solid lines) and imaginary parts (dotted lines) are shown for complex scalets and wavelets.



Fig. 2. Analysis wavelets for disjoint root set examples.

All filters of all families were demonstrated to meet or surpass requirements for orthogonality, biorthogonality, and reconstruction when tested [14] in 2-band wavelet filter banks. In general, reconstruction errors ranged from "perfect" at $\mathcal{O}(10^{-16})$ to "near-perfect" at $\mathcal{O}(10^{-8})$ as K ranged from K = 1 to 24 for both orthogonal and biorthogonal classes. All search-optimized filter families were observed to have the optimal values of their defining selection criterion when compared to the other families.

Figs. 3–6 display values of various characteristic properties for the filter families. The families are listed in the legends sorted in order of the properties' median values for $\mathcal{A}(z)$ over the range of the indexing parameter. These figures and the corresponding numerical values in tables can be examined to assess Λ -equivalence. Refer to [12,16] for a complete catalogue of all results for all of the filter families with both numerical tables of parameter estimates and graphical displays of the filters in the time, frequency, and z domains.

Although named distinctly because of their different computational algorithms, there are several pairs of filter families which should ideally be \mathcal{F} -, \mathcal{Z} - and Λ -equivalent. These pairs provide a test for verifying computational methods. The DROMD and DROMA families should be equivalent real families, while the DCOMD and DCOMN families should be equivalent complex families. Numerical experiments have confirmed these expected results. All constraint-selected families have been compared with the search-optimized families for $K_a = K_s = 1, ..., 24$. Each member of the



Orthogonal Filters Time-Domain Regularity





Biorthogonal Filters Time-Domain Regularity

Fig. 4. Time-domain regularity for biorthogonal filters.



Orthogonal Filters Time-Frequency Uncertainty

Fig. 5. Time-frequency uncertainty for orthogonal filters.



Orthogonal Filters Phase NonLinearity

Fig. 6. Phase nonlinearity for orthogonal filters.

following sets of filter families have been demonstrated to be \mathscr{F} -equivalent to the other members of the set with $\tau(\mathscr{F})$ at machine precision: {DRBMD, DRBMU, DRBLS, DRBLR}, {DRBRD, DRBMR}, {DROMD, DROMA}, and {DCOMD, DCOMN}.

Figs. 3 and 4 present visually dramatic contrasting examples of the presence and absence of Λ -equivalence, respectively, for the orthogonal and biorthogonal families with regard to the property of time-domain regularity. Examination of these figures reveals that of those displayed, all of the orthogonal families, but none of the biorthogonal families, are Λ -equivalent with $\tau(\Lambda) < 0.2$ for time-domain regularity. Figs. 5 and 6 demonstrate that {DROLD, DROLU} and {DROLD, DROLA} are each Λ -equivalent pairs of orthogonal families, respectively, with regard to time-frequency uncertainty and phase nonlinearity. Analogous results for biorthogonal families have shown that {DRBMR, DRBLU} is a Λ -equivalent pair with regard to time-frequency uncertainty for $\mathscr{A}(z)$, but there is no such pair with regard to frequency-domain selectivity. Note that since the pair {DRBRD, DRBMR} is \mathscr{F} -equivalent, then the pair {DRBRD, DRBMR} is Λ -equivalent with regard to time-frequency with regard to time-domain regularity and the pair {DRBRD, DRBMR} is Λ -equivalent with regard to time-frequency.

8. Discussion

An algorithm has been developed to unify all of the diverse families of real and complex orthogonal and biorthogonal Daubechies wavelets. This automated algorithm is valid for any order Kof wavelet and insures that the same consistent choice of roots is always made in the computation of the filter coefficients. It is also sufficiently flexible and extensible that it can be generalized to select roots for filters designed by criteria other than those that already comprise the systematized collection of Daubechies wavelets [11,15,17].

Systematizing a collection of filters with a mechanism both for generating and evaluating the filters enables the development of filter catalogues with tables of numerical parameter estimates characterizing their properties. Providing estimates for a variety of characteristics in both time and frequency domains, rather than just the optimized characteristic, constitutes an important aspect of these tables which enhances their utility. Use of these catalogues as a resource enables the investigator to choose an available filter with the desirable characteristics most appropriate to his research problem or development application.

The systematized collection of Daubechies wavelets has been developed within the context of a general filter design framework consisting of indexing parameters $\gamma \in \Gamma$, specification parameters $\xi \in \Xi$, filter coefficients $f \in \mathscr{F}$, filter roots $z \in \mathscr{Z}$, characterization parameters $\lambda \in \Lambda$, their corresponding spaces, and the mappings between the spaces. Within this framework, definitions have been introduced for filter families that are either search optimized or constraint selected, for the equivalence of families, and for new design principles based on disjoint root sets and filter characteristic properties.

Several pairs of both \mathscr{F} -equivalence and Λ -equivalence have been demonstrated for both orthogonal and biorthogonal classes of filter families. If Λ -equivalence exists between a constraint-selected family and a search-optimized family with respect to a particular characterization parameter λ as an extremal property, then the constraint-selected family can be used to replace the search-optimized family, and thus to obviate the necessity for a search in the computational algorithm. As an important example, the DROLD (least disjoint) family can be used as an effective substitute for the DROLA (least asymmetric) family.

The Λ -equivalent substitution of a constraint-selected family for a search-optimized family enables fast computation of those constraint-selected family members for which the corresponding search-optimized family members would require excessively slow computation. Because of the Λ -equivalence, this substitution can be performed without any loss greater than the tolerance $\tau(\Lambda)$ for the parameter λ representing the characteristic property of the filter. Sufficiently fast computation of filters within required error tolerances becomes critically important for real-time or on-line adaptive applications.

The spectral factorization approach advocated here for the systematized collection of Daubechies wavelets has been criticized [18,9] for the numerical instabilities associated with finding the roots of a symmetric positive polynomial at high orders. However, the angular parameterization methods, albeit avoiding the root-finding problem, do not guarantee that filters generated by lattices will have other desireable characteristics such as maximal frequency-domain selectivity or minimal time-frequency uncertainty. Although the parameter-space constraint on the angles for K = 1 vanishing moment on the wavelet [9] may insure some time-domain regularity and other desireable characteristics with relevance to low order filters with small N, it does not necessarily for high-order filters with large N. Searching a parameter space for the corresponding large K becomes increasingly computationally expensive. Thus, finding a filter with desireable characteristics becomes more difficult because of the unrestricted search space. Although the angular parameterization of Zou and Tewfik [18] does impose constraints for more than one vanishing moment, they did not present any filter examples for K > 2.

In contrast, Daubechies wavelets with a wide variety and combination of desireable filter characteristics can be readily computed via spectral factorization as demonstrated in the systematized collection developed in [11,15,17] and reviewed here. Thus, despite the criticism of other authors [18,9] regarding the numerical instabilities inherent in spectral factorization, so far the method remains more useful in generating higher order wavelets with more than one vanishing moment. Clearly, each of the different approaches has advantages and disadvantages. Therefore, the most prudent and practical position to adopt would be that of verifying for each algorithm its utility in terms of the class of filters and range of filter lengths N for which the algorithm is valid, the possible combinations of desired filter characteristics for which a search can be done, and the computational complexity of the search for filters with those characteristics. As reviewed here, this task has been completed for the Daubechies wavelets computed via spectral factorization.

References

- [1] I. Daubechies, Orthonormal bases of compactly supported wavelets, Comm. Pure Appl. Math. 41 (1988) 909-996.
- [2] I. Daubechies, Ten Lectures on Wavelets, in CBMS-NSF Series in Applied Mathematics, Vol. 61, Society for Industrial and Applied Mathematics, Philadelphia, PA, 1992.
- [3] T.N.T. Goodman, C.A. Micchelli, G. Rodriguez, S. Seatzu, Spectral factorization of Laurent polynomials, Advances in Computational Mathematics 7 (1997) 429–454.
- [4] A.I. Markushevich, Theory of Functions of a Complex Variable, 2nd Edition, Chelsea Publishing Company, New York, 1977 (translated by Richard A. Silverman).
- [5] Y. Meyer, Wavelets: Algorithms and Applications, Society for Industrial and Applied Mathematics, Philadelphia, PA, 1993 (translated by Robert D. Ryan).

- [6] E.M. Reingold, J. Nievergelt, N. Deo, Combinatorial Algorithms: Theory and Practice, Prentice-Hall, Englewood Cliffs, NJ, 1977.
- [7] O. Rioul, P. Duhamel, A Remez exchange algorithm for orthonormal wavelets, IEEE Trans. Circuits Systems II, Analog Digital Signal Process. 41 (8) (1994) 550–560.
- [8] J. Shen, G. Strang, The zeros of the daubechies polynomials, Proc. Amer. Math. Soc. 124 (12) (1996) 3819–3833.
- [9] B.G. Sherlock, D.M. Monro, On the space of orthonormal wavelets, IEEE Trans. Signal Process. 46 (6) (1998) 1716–1720.
- [10] C. Taswell, Algorithms for the generation of Daubechies orthogonal least asymmetric wavelets and the computation of their Holder regularity, Technical report, Scientific Computing and Computational Mathematics, Stanford University, August 1995.
- [11] C. Taswell, Computational algorithms for Daubechies least-asymmetric, symmetric, and most-symmetric wavelets, in: Proceedings of the International Conference on Signal Processing Applications and Technology, Miller Freeman, September 1997, pp. 1834–1838.
- [12] C. Taswell, The Systematized Collection of Wavelet Filters Computable by Spectral Factorization of the Daubechies Polynomial, Computational Toolsmiths, www.toolsmiths.com, 1997.
- [13] C. Taswell, Correction for the Sherlock-Monro algorithm for generating the space of real orthonormal wavelets, Technical Report CT-1998-05, Computational Toolsmiths, www.toolsmiths.com, June 1998.
- [14] C. Taswell, Empirical tests for the evaluation of multirate filter bank parameters, Technical Report CT-1998-02, Computational Toolsmiths, www.toolsmiths.com, March 1998.
- [15] C. Taswell, A spectral-factorization combinatorial-search algorithm unifying the systematized collection of Daubechies wavelets, in: V.B. Bajić (Ed.), Proceedings of the IAAMSAD International Conference SSCC'98: Advances in Systems, Signals, Control, and Computers, Vol. 1, September 1998, pp. 76–80 (invited paper and session keynote lecture).
- [16] C. Taswell, The systematized collection of Daubechies wavelets, Technical Report CT-1998-06, Computational Toolsmiths, www.toolsmiths.com, June 1998.
- [17] C. Taswell, Least and most disjoint root sets for Daubechies wavelets, in: Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing, 1999, Paper #1164.
- [18] H. Zou, A.H. Tewfik, Parametrization of compactly supported orthonormal wavelets, IEEE Trans. Signal Process. 41 (3) (1993) 1428–1431.



JOURNAL OF COMPUTATIONAL AND APPLIED MATHEMATICS

Journal of Computational and Applied Mathematics 121 (2000) 197-219

www.elsevier.nl/locate/cam

Multivariate Padé approximation

Philippe Guillaume*, Alain Huard

UMR MIP 5640, Département de Mathématiques, INSA, Complexe Scientifique de Rangueil, 31077 Toulouse Cedex, France

Received 11 November 1999; received in revised form 20 February 2000

Abstract

This paper is a survey on the multivariate Padé approximation. Two types of approximants are considered: those which can approximate general meromorphic functions f = h/g where both h and g are holomorphic, and those which are specialized to the approximation of functions of the same form where g is a polynomial. Algorithms are described, together with the different techniques used for proving convergence. © 2000 Elsevier Science B.V. All rights reserved.

MSC: 41A21; 41A63; 65D15

Keywords: Multivariate Padé approximants; Convergence of multivariate Padé approximants; Homogeneous approximants; Nonhomogeneous approximants; Padé o Padé approximants; Nested Padé approximants; Least-squares Padé approximants

1. Introduction: from univariate to multivariate Padé approximation

Let f(z) be a function defined on a subset of the complex plane. In many applications, the function is known through the first terms of its power series expansion. For example, in Electromagnetics or in Mechanics, the electric field (respectively the displacement) is the solution to a square linear system which depends on a parameter z (e.g. the frequency):

$$(A + Bz + Cz2) f(z) = b(z).$$

If the matrix $A \in \mathcal{M}_n(\mathbb{C})$ is invertible and if b(z) is holomorphic, then the solution f(z) is holomorphic around the origin, and has a power series expansion

$$f(z) = \sum_{k \ge 0} c_k z^k, \quad c_k \in \mathbb{C}^n.$$
(1)

* Corresponding author.

0377-0427/00/\$ - see front matter © 2000 Elsevier Science B.V. All rights reserved. PII: S0377-0427(00)00337-X

E-mail addresses: guillaum@gmm.insa-tlse.fr (P. Guillaume), huard@gmm.insa-tlse.fr (A. Huard).

The vectors c_k in this series can easily be computed by solving successively the systems

$$egin{aligned} &Ac_0 = b_0, \ &Ac_1 = -Bc_0 + b_1, \ &Ac_k = -Bc_{k-1} - Cc_{k-2} + b_k, \quad k \geqslant 2, \end{aligned}$$

where $b(z) = \sum_{k \ge 0} b_k z^k$ [13]. These systems are obtained by the identification of the coefficients of z^k in $(A+Bz+Cz^2) \sum_{k \ge 0} c_k z^k = \sum_{k \ge 0} b_k z^k$. Of course, only a finite number of coefficients c_k , $0 \le k \le N$, are computed and a good approximation of f(z) may be obtained by a Taylor polynomial

$$f(z) \simeq \sum_{k=0}^{N} c_k z^k.$$

However, such an approximation will be accurate if series (1) itself converges, that is, if $|z| < \rho$, where ρ is the convergence radius of the series. Unfortunately, ρ is often finite because there are complex numbers z_i , i = 1, 2, ..., such that $\det(A + Bz_i + Cz_i^2) = 0$, for which $f(z_i)$ is not defined. Hence, the function f(z) is usually a meromorphic function with poles z_i and a convergence radius $\rho = \min_i |z_i|$.

In such a case, it is well known that a Padé approximation can be far more accurate than a Taylor approximation. Essentially, it is a consequence of the famous Montessus de Ballore theorem, who established in 1902 the uniform convergence of Padé approximants on compact subsets excluding the poles. Particularly, a good approximation of f(z) can be obtained outside the disk of convergence of series (1), where the Taylor expansion fails to converge.

The above-mentioned example depended on a single variable z. Often, there are, in fact, some other parameters like shape variables, material properties (Hookes law, electromagnetic properties), boundary conditions, etc. For such cases, it would be desirable to construct a multivariate Padé approximation. However, the problem in several variables is much more difficult than in one variable, and many research has been done in the last 30 years in order to find a generalization with good convergence properties.

In order to understand these difficulties, and the different solutions which have been proposed to overcome them, first let us consider the univariate Padé approximant of a scalar and meromorphic function f defined on the complex plane,

$$f(z) = \frac{u(z)}{v(z)},\tag{2}$$

where u and v are both holomorphic functions on \mathbb{C} and $v(0) \neq 0$. For given integers m and n, let p(z)/q(z), deg $p \leq m$, deg $q \leq n$, be a nontrivial solution to the homogeneous and linear system

$$q(z)f(z) - p(z) = O(z^{m+n+1}).$$
 (3)

The proof of the Montessus de Ballore theorem is essentially based on the fact that the zeros of the function v form an at most countable set $Z = \{z_1, z_2, z_3, ...\}$ of isolated points. This property, which is particular to the univariate case, allows to rewrite the function f on a given disk D(0,r) as a fraction with a polynomial denominator:

$$f(z) = \frac{h(z)}{g(z)}, \qquad g(z) = \prod_{|z_i| < r} (z - z_i)^{\alpha_i}, \quad \forall z \in D(0, r) \backslash Z, \tag{4}$$

where α_i is the order of the pole z_i , and h(z) = u(z)g(z)/v(z) is holomorphic on D(0,r). Let *n* be the degree of the polynomial *g*. Using this form of the function *f*, the Montessus de Ballore theorem states that the Padé approximants p(z)/q(z) solution to (3) converge uniformly to *f* on all compact subset of $D(0,r)\backslash Z$ when $m \to \infty$. A natural generalization of this theorem would ask for multivariate rational approximants to converge to *f* uniformly on compact subsets excluding the zero set of the denominator.

There are mainly two difficulties in the generalization to several variables. The first one is that the substitution of a polynomial for v in Eq. (2) is no longer possible for a multivariate meromorphic function. The reason is that usually the zero set of an holomorphic function in \mathbb{C}^d , d > 1, does not coincide, even locally, with the zero set of a polynomial. This implies that for a general meromorphic function, one cannot hope from a sequence p/q of fractions where the degree of q remains bounded, to converge to f. The first way for overcoming this difficulty was proposed in 1988 by Chaffy who introduced the Padé \circ Padé approximants [3]. In the case of two complex variables x, y, they are obtained in two steps. The first one consists in the computation of the Padé approximant of the function $f_y: x \mapsto f(x, y)$ with respect to the variable x. The second step consists in the computation of the Padé approximant of the resulting function with respect to the variable y. Using a similar approach, one of the authors introduced the nested Padé approximants [11]. The difference lies in the second step, where the Padé approximants of the coefficients of the first step result are computed. These two approximants are rational approximants of f in the field $\mathbb{C}(x, y)$ of fractions in x and y with coefficients in \mathbb{C} , but they are computed in the field $\mathbb{C}(x)(y)$ (or $\mathbb{C}(y)(x)$ for the second one) of fractions in y with coefficients in $\mathbb{C}(x)$. Of course, these fields are the same, but the representation is changed, and particularly the number of coefficients is not the same for a given degree. It may seem that such a representation is not as elegant as the usual one. For example, the symmetry in x, y is lost, though this fact may be used when the two variables are of different nature. However, what is gained through this representation is that convergence can be obtained for a large class of meromorphic functions, even with non polynomial denominators. The proofs of convergence are based on the fact that under some suitable assumptions the zeros of a function v(x, y) can locally be identified with the zeros of a function g(x, y) which is a polynomial with respect to one variable, that is of the form $g(x, y) = \prod_{i=1}^{n} (x - x_i(y))$. Then univariate techniques can be applied to this form.

In order to get around the previous difficulty, many authors have concentrated their attention on the functions f which can be written in the form

$$f(z) = \frac{h(z)}{g(z)}, \quad z = (z_1, z_2, \dots, z_d) \in \mathbb{C}^d,$$
 (5)

where h is holomorphic and g is a polynomial of degree n.

The second difficulty, which appears when the field $\mathbb{C}(z_1, z_2, ..., z_d)$ is used for the approximation, is that "no formal equation analogous to (the univariate case) gives the correct number of linear equations to determine the coefficients" [4]. Thus, several choices have been made in order to define multivariate Padé approximants. One of the first definitions was proposed by Chisholm in 1973 [4]. A few years later, a most general definition was given by Levin [16], from which a particular case was studied by Cuyt [5], the so-called homogeneous Padé approximants (see also [15] for a definition based on orthogonal polynomials). They are closely related to the univariate approximation, and they allowed Cuyt to obtain in 1985 the first uniform convergence result for the multivariate case [6], recently improved in [7]. For numerical applications, these approximants suffer from a lack of convergence on the complex lines where $t \mapsto g(tz)$ has less than *n* roots. Particularly, they have a singularity at the origin, which has been carefully studied by Werner [19]. Up to now, no convergence has been obtained for the general definition proposed by Levin. The standard proofs of consistency and convergence break down for the same reason that there are not enough equations to uniquely determine a Padé approximant (see, for example, (11)). In the Padé approximation theory, consistency usually means that if f = h/g is a rational fraction, then its Padé approximant P/Q should be equal to f if the degrees of P and Q are correctly chosen. The least-squares Padé approximants introduced by the authors in [12] have allowed to obtain consistency and uniform convergence on compact subsets excluding the zero set of g. In the univariate case, the latter formulation provides an alternative to the classical Padé approximation, and coincides with it for a particular choice of the interpolation set.

This paper is organized as follows. The next section discusses the multivariate Padé approximation of functions f = h/g where h is holomorphic and g is a polynomial, and Section 3 considers the more general case where g is holomorphic. A simple algorithm is given for computing each kind of multivariate Padé approximant. We also describe their convergence properties and show how the proofs of convergence are closely related to the univariate case.

2. Multivariate Padé approximants of f/g with g polynomial

Many definitions have been proposed for the multivariate approximation of a function f(z) = h(z)/g(z), $z \in \mathbb{C}^d$, $d \ge 1$, where *h* is holomorphic and *g* is a polynomial of degree *n*. We will focus our attention to the two consistent approximations for which uniform convergence has been proved: the homogeneous (HPA) and the least-squares (LSPA) multivariate Padé approximants. It seems that although some other approximants may have some historical interest, their lack of convergence is a serious handicap for numerical applications.

In the HPA, the coefficients of the approximant P/Q are defined by a linear system which is over-determined for more than two variables. Due the particular choice of the degrees and the interpolation indices set, this system can be solved exactly. In the LSPA, the over-determined linear system defining the coefficients is solved in a weighted least-squares sense.

First some notation is introduced, and the consistency of a multivariate Padé approximation is discussed. Next, we describe the HPA and the LSPA. The proofs of convergence are very similar, and reported at the end of the section.

2.1. Notation

For a given finite subset $M \subset \mathbb{N}^d$, the set of polynomials $P \in \mathbb{C}[z]$ having the form $P(z) = \sum_{\alpha \in M} P_{\alpha} z^{\alpha}$ is denoted by \mathbb{P}_M . The standard notation $z^{\alpha} = \prod_{i=1}^d z_i^{\alpha_i}$ is used for $\alpha \in \mathbb{N}^d$ and $z \in \mathbb{C}^d$. A polynomial $P \in \mathbb{P}_M$ is said *M*-maximal if for all polynomial $Q \in \mathbb{C}[z]$, the condition $PQ \in \mathbb{P}_M$ implies $Q \in \mathbb{C}$. The degree of a polynomial $P = \sum_{\alpha} P_{\alpha} z^{\alpha}$ is $\max\{|\alpha|, P_{\alpha} \neq 0\}$ where $|\alpha| = \alpha_1 + \alpha_2 + \cdots + \alpha_d$, and the valuation of a series $S = \sum_{\alpha} S_{\alpha} z^{\alpha}$ is $\min\{|\alpha|, S_{\alpha} \neq 0\}$.

A subset $M \subset \mathbb{N}^d$ has the *rectangular inclusion property* if the conditions $\alpha \in \mathbb{N}^d$, $\beta \in M$ and $\alpha \leq \beta$ imply $\alpha \in M$. The standard partial order of \mathbb{N}^d is used, that is, $\alpha \leq \beta$ means $\alpha_i \leq \beta_i$ for $1 \leq i \leq d$. The number of elements of a finite subset M is denoted by |M|, and M + N denotes the set $\{\alpha + \beta; \alpha \in M, \beta \in N\}$. If $P \in \mathbb{P}_M$ and $Q \in \mathbb{P}_N$, then $PQ \in \mathbb{P}_{M+N}$.

For a sequence $(M_m)_{m\geq 0}$, $M_m \subset \mathbb{N}^d$, we say that $\lim_{m\to\infty} M_m = \infty$ if for all bounded subset B of \mathbb{N}^d , there exists an integer k such that $B \subset M_m$ for all $m \geq k$. For the sake of simplicity, we will omit the subscript m and write $M \to \infty$.

For a function f which is holomorphic around the origin, the coefficient of z^{α} in its power series expansion is denoted by f_{α} , that is, $f(z) = \sum_{\alpha \ge 0} f_{\alpha} z^{\alpha}$, and for $E \subset \mathbb{N}^d$, $f_E(z)$ denotes the partial sum $f_E(z) = \sum_{\alpha \in E} f_{\alpha} z^{\alpha}$.

2.2. Consistency of a rational approximation

Here we suppose that f is a fraction,

$$f(z) = \frac{h(z)}{g(z)}, \quad h \in \mathbb{P}_R, \ g \in \mathbb{P}_S,$$

where R and S are finite subsets of \mathbb{N}^d , and $g(0) \neq 0$. Consider three finite subsets $M, N, E \subset \mathbb{N}^d$, and a fraction P/Q, $P \in \mathbb{P}_M$, $Q \in \mathbb{P}_N$, such that

$$(f - P/Q)_E = 0.$$
 (6)

If the requirements $M \subset E$ and |E| = |M| + |N| - 1 are added, then this equation corresponds to the general definition given by Levin [16].

Like in other domains of numerical analysis, consistency is almost necessary for convergence. The question is under which conditions on the sets M, N, E, does Eq. (6) define a consistent approximation, that is, Eq. (6) implies P/Q = f.

In the univariate case, one has $M = R = \{0, 1, ..., m\}$, $N = S = \{0, 1, ..., n\}$, and

$$E = M + N, (7)$$

$$|E| = |M| + |N| - 1.$$
(8)

An extra condition such as Q(0) = 1 is usually added in order to avoid the zero solution. Then, Eq. (8) means that a square system is obtained for the free coefficients of P and Q. We will see that Eq. (7) is a consistency condition. When the sets M, N, E satisfy the rectangular inclusion property, the two conditions (7) and (8) are equivalent in the univariate case but incompatible in the multivariate case, because the identity |M + N| = |M| + |N| - 1 holds only for d = 1. Hence one of them must be abandoned.

Proposition 2.1. Let f = h/g be an irreducible fraction, where $h \in \mathbb{P}_R$, $g \in \mathbb{P}_S$, $g(0) \neq 0$. For a given $E \subset \mathbb{N}^d$ satisfying the inclusion property, let $(P,Q) \in \mathbb{P}_M \times \mathbb{P}_N$ be a nontrivial solution to the linear and homogeneous system

$$(Qf - P)_E = 0. (9)$$

If

$$N + R \subset E \quad \text{and} \quad M + S \subset E, \tag{10}$$

then

$$\frac{P}{Q} = \frac{h}{g}.$$

Moreover, if N = S and if g is N-maximal, then there exists a constant $c \in \mathbb{C}$ such that

$$P = ch, \quad Q = cg.$$

Proof. Due to $g(0) \neq 0$ and the rectangular inclusion property, system (9) is equivalent to $(Qh - Pg)_E = 0$. Due to $Qh \in \mathbb{P}_{N+R}$, $Pg \in \mathbb{P}_{M+S}$, $N + R \subset E$ and $M + S \subset E$, we have Qh = Pg. It follows from the Gauss lemma that there exists a polynomial c such that P = hc, Q = gc, hence P/Q = h/g. If $Q \in \mathbb{P}_N$ and g is N-maximal, then $c \in \mathbb{C}$. \Box

This proposition shows that condition (7) is sufficient for consistency when M = R and N = S. Without any special assumption on the function f, the latter condition is also necessary, as illustrated by the following example where trying to preserve (8) instead of (7) leads to divergence. Let

$$f(x,y) = \frac{1}{(1-x)(1-y)}.$$
(11)

The power series expansion of function f around the origin reads $f(x, y) = \sum_{ij} x^i y^j$. Let $M = R = \{(i, j) \in \mathbb{N}^2; i + j \leq 2\}, N = S = \{0, 1\}^2, E = \{0, 1, 2\}^2, P(x, y) = 1 + x + y + x^2 + y^2 \in \mathbb{P}_M$ and $Q(x, y) = 1 - xy \in \mathbb{P}_N$. Then P and Q are solution to Eq. (9), condition (8) is fulfilled, but $P/Q \neq f$. Such an indetermination of the denominator coefficients appears also for higher degrees of the numerator, making it impossible to obtain uniform convergence of the Padé approximants to the function f on compact subsets of $\{(x, y) \in \mathbb{C}^2; (1 - x)(1 - y) \neq 0\}$ when M increases.

Observe that system (9) has the nontrivial solution (h, g), although it is usually strongly overdetermined. When h is not a polynomial, depending on the choice of the sets M, N, E, this system will not necessarily have some nontrivial solutions. In the homogeneous approach, system (9) can always be solved exactly by letting $|N| \to \infty$ when $M \to \infty$. In the least-squares approach, the set N is kept constant with N = S, and system (9) is solved in a least-squares sense.

2.3. The homogeneous multivariate Padé approximants

The HPA were introduced by Cuyt [5]. Let f(z) = h(z)/g(z) be a meromorphic function where g is a polynomial of degree n with $g(0) \neq 0$.

2.3.1. Definition

The polynomials P and Q and the interpolation set E are chosen in the following way. For a given $m \ge 0$, consider the three sets

$$M = \{ \alpha \in \mathbb{N}^d, \, mn \leqslant |\alpha| \leqslant mn + m \},\tag{12}$$

$$N = \{ \alpha \in \mathbb{N}^d, \, mn \leq |\alpha| \leq mn+n \},\tag{13}$$

$$E = \{ \alpha \in \mathbb{N}^d, \ 0 \leqslant |\alpha| \leqslant mn + m + n \}.$$
(14)

We look for polynomials $P \in \mathbb{P}_M$ and $Q \in \mathbb{P}_N$ such that

$$(Qf - P)_E = 0.$$
 (15)

For $R = \{\alpha \in \mathbb{N}^d, 0 \leq |\alpha| \leq m\}$ and $S = \{\alpha \in \mathbb{N}^d, 0 \leq |\alpha| \leq n\}$, one has M + S = N + R = E, and it follows from Proposition 2.1 that this approximation is consistent.

The idea is to introduce the univariate polynomials p and q defined, for fixed $z \in \mathbb{C}^d$, by $p(t) = t^{-mn}P(tz)$ and $q(t) = t^{-mn}Q(tz)$, $t \in \mathbb{C}$, of respective degrees m and n (at most). The coefficient of t^k in P(tz) is an homogeneous polynomial in z of degree k, which explains the name given to these approximants. One can consider p and q as elements of the ring $\mathbb{C}[z][t]$, and compute a univariate Padé approximant solution to the linear system

$$q(t)f(tz) - p(t) = O(t^{m+n+1}).$$
(16)

This system with m + n + 2 unknown coefficients and m + n + 1 equations has always a nontrivial solution in $(\mathbb{C}[z])^{m+n+2}$. Moreover, if the denominator below is not zero, a solution is given explicitly by Jacobi's determinant representation

$$\frac{p(t)}{q(t)} = \frac{\begin{vmatrix} t^n F_{m-n}(t) & t^{n-1} F_{m-n+1}(t) & \cdots & F_m(t) \\ f_{m-n+1} & f_{m-n+2} & \cdots & f_{m+1} \\ \vdots & \vdots & & \vdots \\ f_m & f_{m+1} & \cdots & f_{m+n} \end{vmatrix}}{\begin{vmatrix} t^n & t^{n-1} & \cdots & 1 \\ f_{m-n+1} & f_{m-n+2} & \cdots & f_{m+1} \\ \vdots & \vdots & & \vdots \\ f_m & f_{m+1} & \cdots & f_{m+n} \end{vmatrix}}$$
(17)

with

$$f_k = \sum_{|\alpha|=k} f_{\alpha} z^{\alpha} \quad \text{if } k \ge 0, \qquad f_k = 0 \quad \text{if } k < 0, \qquad F_k(t) = \sum_{i=0}^k f_i t^i.$$

It follows from (16) that $t^{mn}q(t)f(tz) - t^{mn}p(t) = O(t^{mn+m+n+1})$ in $\mathbb{C}[z][[t]]$, the ring of power series in t with coefficients in $\mathbb{C}[z]$. By construction, the coefficient of t^k in $t^{mn}q(t)f(tz) - t^{mn}p(t)$ is homogeneous in z of degree k. This implies that $(q(1)f - p(1))_E = 0$. Moreover, P(z) = p(1) and Q(z) = q(1) are polynomials in z with valuations at least mn and respective degrees at most mn + mand mn + n, thus P and Q are a solution to Eq. (15).

Observe that as Qf - P has valuation at least mn, Eq. (15) can also be written

$$(Qf - P)_{E'} = 0, \quad E' = \{ \alpha \in \mathbb{N}^d, \, mn \leq |\alpha| \leq mn + m + n \}.$$

$$(18)$$

For the particular dimension d = 2, $(Qf - P)_{E'} = 0$ is an homogeneous and linear system with

$$s = \binom{mn+m+n+2}{2} - \binom{mn+1}{2}$$

equations and s + 1 unknowns. Hence it is not surprising to find a nontrivial solution. However, for d > 2, the latter system becomes over-determined, but still has some nontrivial solutions.

The solution P/Q found above is not necessarily an irreducible fraction. If P_1/Q_1 is another solution to Eq. (15), then $PQ_1 - P_1Q = 0$. Hence the irreducible form $P_{(m,n)}/Q_{(m,n)}$ of P/Q is unique.

Definition 2.1. The [m, n] homogeneous multivariate approximant of the function f is the irreducible form $P_{(m,n)}/Q_{(m,n)}$ of P/Q where P and Q satisfy (15).

When P/Q is not irreducible, it may happen that the valuation of the polynomial $Q/Q_{(m,n)}$ has a positive valuation *s*, and there is a "backward shift" *s* on the valuations of *P* and *Q*. In that case, the polynomials $P_{(m,n)}$ and $Q_{(m,n)}$ do not necessarily satisfy Eq. (15). For example, if the backward shift is *mn*, then one can only guaranty $(Q_{(m,n)}f - P_{(m,n)})_F = 0$ where $F = \{\alpha \in \mathbb{N}^d, 0 \leq |\alpha| \leq m+n\}$.

In a more algebraic presentation, and following Brezinski's univariate theory, Kida defines in [16] the same multivariate approximant as a particular case of a Padé-type approximant for which the generating polynomial is precisely q(t) in (17). However, one can observe that the substitution of tz for z, which allows to use the univariate construction, is made possible because of the particular choice of the sets E, M, N (12)–(14).

2.3.2. Computation

The [m, n] homogeneous multivariate approximant can be computed in several ways. One possibility consists in solving directly (18). Here the unknowns are complex numbers. Although over-determined if d > 2, this system has always some nontrivial solutions. The algorithm is the following.

Algorithm 1: HPA computation

- 1. Choose three enumerations $(\alpha_i)_{1 \leq i \leq |M|}$, $(\beta_i)_{1 \leq i \leq |N|}$, $(\gamma_i)_{1 \leq i \leq |E' \setminus M|}$, of the respective subsets $M = \{\alpha \in \mathbb{N}^d, mn \leq |\alpha| \leq mn + m\}$, $N = \{\beta \in \mathbb{N}^d, mn \leq |\beta| \leq mn + n\}$ and $E' \setminus M = \{\gamma \in \mathbb{N}^d, mn + m < |\gamma| \leq mn + m + n\}$.
- 2. Compute the coefficients of the $|E' \setminus M| \times |N|$ complex matrix A:

$$A_{ij} = f_{\gamma_i - \beta_j}, \quad 1 \leq i \leq |E' \setminus M|, \ 1 \leq j \leq |N|,$$

where $f_{\alpha} := 0$ if $0 \leq \alpha$.

3. Compute a nontrivial solution $v \in \mathbb{C}^{|N|}$ to the system Av = 0, and define the polynomial Q by

$$Q(z) = \sum_{i=1}^{|N|} v_i z^{\beta_i}.$$

4. Compute

$$w_i = \sum_{j=1}^{|N|} v_j f_{\alpha_i - \beta_j}, \quad 1 \leq i \leq |M|$$

and define the polynomial P by

$$P(z) = \sum_{i=1}^{|M|} w_i z^{\alpha_i}.$$

5. The fraction P/Q is the [m, n] HPA of the function f.

A second possibility is to use symbolic computation for solving (16) in the ring $\mathbb{C}[z]$. Here the unknowns are complex polynomials, and this system is always under-determined (size $(m + n + 1) \times (m + n + 2)$). Such an approach is also used for computing the Padé \circ Padé approximants (cf. Section 3.3).

Finally, one can also take advantage of the construction of the homogeneous approximant. For $z \in \mathbb{C}^d$ such that $t \mapsto g(tz)$, $t \in \mathbb{C}$, has exactly *n* roots, the rational fraction $t \mapsto P_{(m,n)}(tz)/Q_{(m,n)}(tz)$ is the [m,n] univariate Padé approximant of the function $f_z(t) := f(tz)$. Hence all the algorithms developed in the univariate case can be applied to compute $P_{(m,n)}(tz)/Q_{(m,n)}(tz) = [m,n]_{f_z}(t)$. Particularly, computation of staircase sequences like in the ε -algorithm or the qd-algorithm can be used. For a description of these two algorithms, we refer the reader to [8].

2.3.3. Convergence

Here we consider sequences of [m, n] HPA where *n* is fixed and $m \to \infty$. The degrees and the valuations of the numerator and the denominator in the nonreduced form P/Q increase when $m \to \infty$, and the convergence is obtained on compact subsets excluding the zero set of *g* if there exists a subsequence of approximants $P_{(m(k),n)}/Q_{(m(k),n)}$ such that $Q_{(m(k),n)} \neq 0$. This implies that the backward shift in the denominator *Q* of [m, n] must be at least *mn*. For a given function *f*, the existence of such a subsequence remains to our knowledge an open question. When it exists, it can be interpreted as a particular case of the LSPA for which the underlying system is solved exactly. In the general case, the HPA converge on a smaller subset obtained by removing also the complex cone formed by the vectors *z* for which $t \mapsto g(tz), t \in \mathbb{C}$, has less than *n* roots (Theorem 2.3).

Suppose that f = h/g is meromorphic on a neighborhood of a polydisc $\overline{D}(0,\rho) = \{z \in \mathbb{C}^d; |z_i| \leq \rho_i, i = 1, ..., d\}$, and h is holomorphic on a neighborhood of $\overline{D}(0,\rho)$. The polynomial g is normalized in such a way that $\sum_{\alpha \in N} |g_{\alpha}|^2 = 1$. Its decomposition into irreducible factors reads

$$g = \prod_{i=1}^{l} g_i^{\tau_i}$$

and the associated algebraic set G with its decomposition into irreducible components G_i are:

$$G = \{z \in \mathbb{C}^d; g(z) = 0\},\$$

$$G_i = \{z \in \mathbb{C}^d; g_i(z) = 0\}$$

Suppose also that $G_i \cap D(0, \rho) \neq \emptyset$ for $1 \leq i \leq l$, and that $h(z) \neq 0$ on a dense subset of $G \cap D(0, \rho)$. The following theorem was proved in [6].

Theorem 2.2 (Cuyt [6]). Let $(P_{(m(k),n)}/Q_{(m(k),n)})_{k\geq 0}$ be a subsequence of homogeneous multivariate approximants such that $Q_{(m(k),n)}(0) \neq 0$ for all $k \geq 0$. Then

$$\lim_{k \to \infty} P_{(m(k),n)} / Q_{(m(k),n)}(z) = f(z)$$

uniformly on all compact subsets of $\{z \in D(0, \rho); g(z) \neq 0\}$. Moreover, the subsequence $(Q_{(m(k),n)})_{k \ge 0}$ converges to g(z) uniformly on all compact subsets of $D(0, \rho)$.

The following result has been obtained in [7] where more general sets than the polydisc $D(0, \rho)$ are considered. Let Λ be the set of vectors $z \in \partial D(0, \rho)$ for which the polynomial $t \mapsto g(tz), t \in \mathbb{C}$, has less than *n* roots counted with multiplicity in $\overline{D}(0, 1)$, and denote by E_{Λ} the cone $\{tz; t \in \mathbb{C}, z \in \Lambda\}$.

Theorem 2.3 (Cuyt and Lubinsky [7]). If $h(z) \neq 0$ for $z \in G$, then

 $\lim_{m \to \infty} P_{(m,n)}/Q_{(m,n)}(z) = f(z)$

uniformly on all compact subsets of $\{z \in D(0, \rho); z \notin E_A, g(z) \neq 0\}$.

2.4. The least-squares multivariate Padé approximants

Studying least-squares orthogonal polynomials, Brezinski proposed recently a least-squares formulation for univariate Padé approximants [2]. This idea has been generalized to the multivariate case in [12]. The formulation [2] did not involve any particular weights in the least-squares approximation, whereas some weights were introduced in [12], which have an important role as it can be seen from the proof of Theorem 2.4.

2.4.1. Definition

The requirements on the function f are the same than in Section 2.3.3. The norm on \mathbb{P}_N is defined by $||Q|| = (\sum_{\alpha \in N} |Q_{\alpha}|^2)^{1/2}$. For $P \in \mathbb{P}_M$, $Q \in \mathbb{P}_N$, and a finite set $E \subset \mathbb{N}^d$, consider the function

$$j(P,Q) = \left(\sum_{\alpha \in E} \rho^{2\alpha} |(Qf - P)_{\alpha}|^2\right)^{1/2}.$$
(19)

Definition 2.2. Let $M, N, E \subset \mathbb{N}^d$ be three finite subsets such that $E \supset M + N$ and E satisfies the inclusion property. A least-squares multivariate Padé approximant of the function f is a fraction P/Q with $(P,Q) \in \mathbb{P}_M \times \mathbb{P}_N$, ||Q|| = 1, and

$$j(P,Q) \leq j(R,S), \quad \forall (R,S) \in \mathbb{P}_M \times \mathbb{P}_N, \quad ||S|| = 1.$$
 (20)

A solution to this problem is denoted by $[M,N]_f$.

For d = 1 and E = M + N, this definition coincides with the standard definition of the univariate Padé approximation. For d = 1 and $E \supseteq M + N$, the least-squares formulation provides an alternative to the exact Padé interpolation.

Observe that j(P,Q) = 0 if $(g,h) \in \mathbb{P}_M \times \mathbb{P}_N$, and it follows from Proposition 2.1 that this approximation is consistent. Although there may exist several LSPA for given M, N and E (even if one considers the irreducible form), the next theorem shows that it has no incidence on the convergence.

2.4.2. Computation

In order to solve (20), first the coefficients of Q are computed, then the coefficients of P are recovered by expanding and truncating the product Qf, that is, $P = (Qf)_M$. The coefficients of Q are solution to

$$\min_{||\mathcal{Q}||=1} \sum_{\alpha \in E \setminus M} \rho^{2\alpha} |(\mathcal{Q}f)_{\alpha}|^2,$$

which can be written in the form

$$\min_{||s||^2=1} s^* A^* A s,$$
(21)

where the vector $s \in \mathbb{C}^{|N|}$ contains the coefficients of Q, A is an $|E \setminus M| \times |N|$ complex matrix and A^* is the conjugate transpose of A. The optimality condition reads

$$A^*As = \lambda s, \quad \lambda \in \mathbb{R}.$$

Hence $s^*A^*As = \lambda \ge 0$, and a solution is given by any normalized eigenvector associated to the smallest eigenvalue value of A^*A . The algorithm is the following.

Algorithm 2: LSPA computation

- 1. Choose three enumerations $(\alpha_i)_{1 \leq i \leq |M|}$, $(\beta_i)_{1 \leq i \leq |N|}, (\gamma_i)_{1 \leq i \leq |E \setminus M|}$, of the respective subsets M, N and $E \setminus M$.
- 2. Choose $\rho \in \mathbb{R}^d_+$ and compute the coefficients of the matrix *A*:

$$A_{ij} = \rho^{\gamma_i} f_{\gamma_i - \beta_j}, \quad 1 \leq i \leq |E \setminus M|, \ 1 \leq j \leq |N|,$$

where $f_{\alpha} := 0$ if $0 \leq \alpha$.

3. Compute an eigenvector $v \in \mathbb{C}^{|N|}$ associated to the smallest eigenvalue of A^*A , and define the polynomial Q by

$$Q(z) = \sum_{i=1}^{|N|} v_i z^{\beta_i}$$

4. Compute

$$w_i = \sum_{j=1}^{|N|} v_j f_{\alpha_i - \beta_j}, \quad 1 \leq i \leq |M|$$

and define the polynomial P by

$$P(z) = \sum_{i=1}^{|M|} w_i z^{\alpha_i}.$$

5. The fraction P/Q is an [M,N] LSPA of the function f.

2.4.3. Convergence

In the following theorem [12], the set N is fixed, and can be any finite subset of \mathbb{N}^d such that $g \in \mathbb{P}_N$ and g is N-maximal (cf. Section 2.1).

Theorem 2.4 (Guillaume et al. [12]). Let $[M,N]_f = P^M/Q^M$ be a sequence of least-squares multivariate Padé approximants with $M \to \infty$. Then

$$\lim_{M \to \infty} [M, N]_f(z) = f(z)$$

uniformly on all compact subsets of $\{z \in D(0, \rho); g(z) \neq 0\}$. Moreover, the sequence $Q^{M}(z)$ converges to g(z) uniformly on all compact subsets of $D(0, \rho)$.

2.5. Proof of Theorems 2.2, 2.3 and 2.4

The proofs are an extension of the beautiful technique introduced by Karlsson and Wallin in the univariate case [14], which is based on the uniform convergence to zero of the function H^M in (22).

2.5.1. Proof of Theorems 2.2 and 2.4

Theorem 2.2 can be seen as a particular case of Theorem 2.4 by using the set $E' = \{\alpha \in \mathbb{N}^d, 0 \leq |\alpha| \leq m + n + mn - s\}$ instead of *E* (where *s*, $mn \leq s \leq mn + n$, is the backward shift on the valuation of *Q*, coming from the assumption $Q_{(m,n)}(0) \neq 0$), the notation $P^M = P_{(m,n)}, Q^M = Q_{(m,n)}$, and the fact that $(Q^M f - P^M)_{E'} = 0$, that is, the least-squares approximation is exact for the HPA. We outline the proof of Theorem 2.4.

Let $(P^M, Q^M) \in \mathbb{P}_M \times \mathbb{P}_N$ be a solution to problem (20) and consider the function

$$H^M = g(Q^M f - P^M), (22)$$

which is holomorphic on a neighborhood of $\overline{D}(0,\rho)$. The keystone of the proof is the following lemma, whose proof is given at the end of the section.

Lemma 2.5. One has

$$\lim_{M \to \infty} H^M(z) = 0 \tag{23}$$

uniformly on all compact subsets of $D(0, \rho)$.

According to Definition 2.2, the sequence $(Q^M)_M$ is bounded in \mathbb{P}_N . Consider an arbitrary subsequence, still denoted by $(Q^M)_M$ for simplicity, which converges to a polynomial $Q \in \mathbb{P}_N$ with ||Q|| = 1. The subsequence $(Q^M)_M$ converges also to Q, uniformly on all compact subsets of \mathbb{C}^d when $M \to \infty$.

The set $G_i \cap D(0,\rho)$ was supposed nonempty. For $z \in G_i \cap D(0,\rho)$, one has $H^M(z) = h(z)Q^M(z)$ and $(H^M)_M$ converges to 0 on $D(0,\rho)$, thus h(z)Q(z) = 0, and Q(z) = 0 because $h(z) \neq 0$ on a dense subset of $G \cap D(0,\rho)$. The set of regular points of G_i is open, connected and dense in G_i , thus Q = 0 on G_i and g_i divides Q [1,18]. Similarly $g_i^{\tau_i}$ divides Q (consider the partial derivatives of H^M), which implies that g divides Q. One has $Q \in \mathbb{P}_N$ and g is N-maximal with ||g|| = 1, thus

$$Q = cg, \quad |c| = 1.$$

Hence $\lim_{M\to\infty} Q^M = cg$ uniformly on all compact subsets of \mathbb{C}^d , and after division of (22) by gQ^M , one obtains with (23)

$$\lim_{M\to\infty}\left(f-\frac{P^M}{Q^M}\right)(z)=0$$

uniformly on all compact subsets of $\{z \in D(0, \rho); g(z) \neq 0\}$. As this holds for all convergent subsequences of the bounded sequence $(Q^M)_M$, the whole sequence $f - P^M/Q^M$ converges to zero in the same way. \Box

Proof of the Lemma. The main line is the following. One has $gP^M \in \mathbb{P}_{M+N}$ and $E \supset M + N$, thus

$$H^{M}_{\alpha} = \begin{cases} (hQ^{M})_{\alpha} & \text{if } \alpha \notin E, \\ (g(Q^{M}f - P^{M})_{E})_{\alpha} & \text{if } \alpha \in E \end{cases}$$

(and $H^M_{\alpha} = 0$ for the HPA if $\alpha \in E'$). The Cauchy integral yields

$$H^{M}_{\alpha} = \frac{1}{(2i\pi)^{d}} \int_{\Gamma_{+}} \frac{hQ^{M}}{z^{\alpha+1}} \,\mathrm{d}z \quad \text{if } \alpha \notin E,$$
(24)

$$H^M_{\alpha} = \frac{1}{(2\mathrm{i}\pi)^d} \int_{\Gamma_+} \frac{g(Q^M f - P^M)_E}{z^{\alpha+1}} \,\mathrm{d}z \quad \text{if } \alpha \in E,$$
(25)

where $\alpha + \mathbf{1} = (\alpha_1 + 1, \dots, \alpha_d + 1)$. The sequence $(Q^M)_M$ is bounded in \mathbb{P}_N and *h* is continuous on Γ_+ , thus

$$|H^M_{\alpha}| \leq \frac{c}{\rho^{\alpha}} \quad \text{if } \alpha \notin E.$$
(26)

The change of variable $z = (\rho_1 \exp(2i\pi\theta_1), \dots, \rho_d \exp(2i\pi\theta_d))$ in Eq. (25) yields

$$H^M_{\alpha} = \int_{[0,1]^d} \frac{g(Q^M f - P^M)_E}{z^{\alpha}} \,\mathrm{d}\theta.$$

The Cauchy-Schwarz inequality and Parseval's formula give

$$|H^M_{\alpha}| \leq \left(\int_{[0,1]^d} \frac{|g|^2}{\rho^{2\alpha}} \,\mathrm{d}\theta\right)^{1/2} \left(\sum_{\alpha \in E} \rho^{2\alpha} |(\mathcal{Q}^M f - P^M)_{\alpha}|^2\right)^{1/2}.$$

Thus, using definition (19) of the function j, one has (possibly with a different c)

$$|H^M_{\alpha}| \leqslant \frac{c}{\rho^{\alpha}} j(P^M, Q^M) \quad \text{if } \alpha \in E.$$
(27)

Due to the definition of P_M , Q_M and to $(h_M, g) \in \mathbb{P}_M \times \mathbb{P}_N$, one has $j(P^M, Q^M) \leq j(h_M, g)$, and gathering Eqs. (26) and (27), one obtains

$$|H^{M}(z)| \leq c \left(j(h_{M},g) \sum_{\alpha \in E} \left| \frac{z}{\rho} \right|^{\alpha} + \sum_{\alpha \notin E} \left| \frac{z}{\rho} \right|^{\alpha} \right),$$
(28)

where $|z/\rho|^{\alpha} = |z_1/\rho_1|^{\alpha_1} \cdots |z_d/\rho_d|^{\alpha_d}$. It follows from

$$j(h_M,g) = \left(rac{1}{(2\mathrm{i}\pi)^d}\int_{\Gamma_+}rac{|h_{E\setminus M}|^2}{z}\,\mathrm{d}z
ight)^{1/2},$$

that $\lim_{M\to\infty} j(h_M,g) = 0$, hence $\lim_{M\to\infty} |H^M(z)| = 0$, uniformly on all compact subsets of $D(0,\rho)$. \Box

2.5.2. Proof of Theorem 2.3

The proof is based on a univariate projection, which allows to shift the degrees of $P_{(m,n)}$ and $Q_{(m,n)}$, that is, for a given $z \in \partial D(0, \rho)$, the univariate Padé approximant of $t \mapsto f(tz)$ reads

$$P_{m,z}(t)/Q_{m,z}(t) = P_{(m,n)}(tz)/Q_{(m,n)}(tz),$$

where deg $P_{m,z} \leq m$, deg $Q_{m,z} \leq n$. Here again, the key idea is to show that the function

$$H_{m,z}(t) = g(tz)(Q_{m,z}(t)f(tz) - P_{m,z}(t))$$

$$=Q_{m,z}(t)h(tz)-P_{m,z}(t)g(tz),$$

converges uniformly to zero. A local extension of an estimation similar to (26) is obtained, which leads to the local convergence (in z) of $Q_{m,z}$ to a polynomial Q_z , deg $Q_z \leq n$. Particularly, if g(tz)=0, then $H_{m,z}(tz) = Q_{m,z}(t)h(tz)$. Taking the limit, it follows from $h(tz) \neq 0$ that $Q_z(t)=0$. Hence Q_z has exactly the same *n* roots than the polynomial $t \mapsto g(tz)$, which allows to complete the proof after division by gQ_z . \Box

3. Multivariate Padé approximants of f/g with g holomorphic

In this section are presented the Padé \circ Padé approximants (PRPA) and the nested Padé approximants (NPA) for a meromorphic function f(z) = h(z)/g(z). Both of them have a natural recursive structure and can be defined for $z \in \mathbb{C}^d$, d > 1. For the sake of simplicity they are here presented in the case of two complex variables x and y. For more variables z_1, z_2, \ldots, z_d , one substitutes z_1 for x and (z_2, \ldots, z_d) for y. Both PRPA and NPA are consistent and convergent, and start for fixed y with the univariate Padé approximant $[m, n]_{f_y}(x)$ of the function $f_y: x \mapsto f(x, y)$. The fraction $[m, n]_{f_y}(x)$ is an element of $\mathbb{C}[[y]](x)$, and is nothing else than a parameterized univariate Padé approximant.

In the PRPA, one computes in $\mathbb{C}(x)[[y]]$ the power series expansion of $[m, n]_{f_y}(x)$, and then, for fixed x, one computes the univariate Padé approximant of the function $x \mapsto [m, n]_{f_y}(x)$. The calculations are done in the field $\mathbb{C}(x)$, hence a good way of doing them is to use symbolic computation. In the NPA, one computes directly the univariate Padé approximants of the coefficients of $[m, n]_{f_y}(x)$, which belong to $\mathbb{C}[[y]]$. The computation does not need symbolic computation.

We point out the fact that in both cases only univariate Padé approximants are computed, for which much knowledge has been accumulated. Also noteworthy is that this kind of approximation can be applied to a larger class of functions than the approximations described in Section 2 because g needs not to be a polynomial.

Due to their construction, the convergence of PRPA or NPA cannot be obtained on all compact subset excluding the singularity of f. The complex lines (x, y) such that $[m, n]_{f_y}$ is not defined must also be removed from the convergence set. It is a sort of intermediate situation between the HPA and the LSPA, where the extra singularities of the HPA have been shifted away: instead of complex lines passing through the origin, these lines are here parallel to the x-axis.

First some notation is introduced. Then we describe the first step and give an intermediate convergence result which will be used for the convergence analysis of both PRPA and NPA.

3.1. Notation

Recall that in the univariate case, if the following linear system

$$q(x)u(x) - p(x) = O(x^{m+n+1}), \quad q(0) = 1$$
(29)

has a unique solution, then the fraction p/q is irreducible and is called the [m, n] Padé approximant of the function $u = \sum_{k \ge 0} u_k x^k$. This fraction is denoted by $[m, n]_u$. The Hankel matrix corresponding to this system is denoted by H(u, m, n), and the right member by C(u, m, n):

$$H(u,m,n) = \begin{pmatrix} u_{m-n+1} & \dots & u_m \\ \vdots & & \vdots \\ u_m & \dots & u_{m+n-1} \end{pmatrix}, \quad C(u,m,n) = - \begin{pmatrix} u_{m+1} \\ \vdots \\ u_{m+n} \end{pmatrix},$$

where $u_i := 0$ if i < 0. The coefficients $S = (q_n, \ldots, q_1)^T$ are solution to the system

$$H(u,m,n)S = C(u,m,n) \tag{30}$$

and the other coefficients p_i , $0 \le i \le m$, are recovered by expanding the product u(x)q(x).

3.2. First step (a parameterized Padé approximant)

Let f be a meromorphic function on a neighborhood of a polydisc $D(0, \rho_1, \rho_2)$,

$$f(x, y) = \frac{u(x, y)}{v(x, y)},$$

where the functions u and v are holomorphic on a neighborhood of $\overline{D}(0, \rho_1, \rho_2)$. For the sake of simplicity, we make the following assumption: $v(x, y) = \sum_{i=0}^{n} v_i(y)x^i$ is a polynomial in x such that $x \mapsto v(x, 0)$ has n simple roots with $v(0, 0) \neq 0$. A particular case is when v is a polynomial in the two variables x and y. In the general case, the set where v(x, y) vanishes is not necessarily algebraic.

Let $\mathscr{Y}_0 \subset D(0, \rho_2)$ be an open subset where the function $y \mapsto f(0, y)$ is holomorphic and the determinant of $H(f_y, m, n)$ is nonzero, and suppose that $0 \in \mathscr{Y}_0$. For a fixed $y \in \mathscr{Y}_0$, we can consider the [m, n] Padé approximant of the function $f_y : x \mapsto f(x, y)$,

$$[m,n]_{f_y}(x) = \frac{U^m(x,y)}{V^m(x,y)} = \frac{\sum_{i=0}^m s_i^m(y) x^i}{1 + \sum_{i=1}^n s_{m+i}^m(y) x^i}.$$
(31)

The subscript m indicates the dependence on m, whereas n is fixed once for all.

3.2.1. Computation

The coefficients of this parameterized Padé approximant can be computed in $\mathbb{C}[[y]]$ in the following way. For $y \in \mathscr{Y}_0$, the vector $S(y) = (s_{m+n}^m(y), \dots, s_{m+1}^m(y))^T$ is the unique solution to the linear system

$$H(f_{y}, m, n)S(y) = C(f_{y}, m, n).$$
 (32)

Due to the assumption det $H(f_0, m, n) \neq 0$, the vector-valued function S(y) is holomorphic around zero and has a power series expansion

$$S(y) = \sum_{j \ge 0} S_j y^j, \quad S_j \in \mathbb{C}^n.$$

The power series expansion of H and C read

$$H(f_{y}, m, n) = \sum_{j \ge 0} H_{j} y^{j}, \quad H_{j} \in \mathbb{C}^{n \times n},$$
$$C(f_{y}, m, n) = \sum_{j \ge 0} C_{j} y^{j}, \quad C_{j} \in \mathbb{C}^{n}.$$

It follows from (32) that the vectors S_i are solution to the systems

$$H_0 S_0 = C_0, ag{33}$$

$$H_0 S_j = -\sum_{k=1}^{j} H_k S_{j-k} + C_j, \quad j \ge 1,$$
(34)

which all have the same matrix. Like in the univariate case, the other coefficients $s_i^m(y)$, $0 \le i \le m$ are obtained by expanding in x the product $f_y(x)V^m(x, y)$. The pseudo-algorithm is the following (series are here considered, which will be later truncated in Algorithms 4 and 5).

Algorithm 3: Intermediate Padé approximant computation

1. Compute

$$H_{j} = \begin{pmatrix} f_{m-n+1,j} & \cdots & f_{m,j} \\ \vdots & & \vdots \\ f_{m,j} & \cdots & f_{m+n-1,j} \end{pmatrix}, \qquad C_{j} = -\begin{pmatrix} f_{m+1,j} \\ \vdots \\ f_{m+n,j} \end{pmatrix}, \quad j \ge 0,$$

where $f(x, y) = \sum_{i,j \ge 0} f_{i,j} x^i y^j$ and $f_{i,j} := 0$ if i < 0. 2. Solve (33) and (34) for $j \ge 1$. Using the numbering $S_j = (s_{m+n,j}^m, \dots, s_{m+1,i}^m)^T$, define

$$s_{m+i}^{m}(y) = \sum_{j \ge 0} s_{m+i,j}^{m} y^{j}, \quad 1 \le i \le n$$
$$V(x, y) = 1 + \sum_{i=1}^{n} s_{m+i}^{m}(y) x^{i}.$$

3. Compute

$$s_{i,j}^{m} = \sum_{k=0}^{n} s_{m+k,j}^{m} f_{i-k,j}, \quad 0 \leq i \leq m, \ j \geq 0,$$

where $s_{m,0}^m := 1$ and $s_{m,j}^m := 0$ for j > 0, and define

$$s_i^m(y) = \sum_{j \ge 0} s_{i,j}^m y^j, \quad 0 \le i \le m$$
$$U(x, y) = \sum_{i=0}^m s_i^m(y) x^i.$$

4. The function U(x, y)/V(x, y) is the [m, n] intermediate Padé approximant of f.

3.2.2. Intermediate convergence

The convergence is a direct consequence of the theory developed in the univariate case. Let $\mathcal{Y} \subset D(0, \rho_2)$ be an open subset with $0 \in \mathcal{Y}$ such that for all $y \in \mathcal{Y}$:

- $v_0(y) \neq 0, v_n(y) \neq 0$,
- the polynomial $x \mapsto v(x, y)$ has *n* simple roots $\alpha_i(y)$, $1 \le i \le n$, $|\alpha_i(y)| < \rho_1$, the functions α_i being holomorphic on \mathscr{Y} (simple roots can be replaced by roots of constant multiplicity),
- $u(x, y) \neq 0$ if v(x, y) = 0.

The following lemma was proved in [3].

Lemma 3.1 (Chaffy-Camus [3]). For all compact subsets $\mathscr{H}_y \subset \mathscr{Y}$, there is an integer m_0 such that for all $m \ge m_0$ and all $y \in \mathscr{H}_y$, there is a unique intermediate Padé approximant $[m,n]_{f_y} = U^m(x, y)/V^m(x, y)$. Let \mathscr{O} be the open subset

$$\mathcal{O} = \{(x, y) \in D(0, \rho_1, \rho_2), y \in \mathcal{Y}, v(x, y) \neq 0\}.$$

The sequence $([m,n]_{f_y})_{m \ge m_0}$ converges uniformly to f on all compact subsets of $(D(0,\rho_1) \times \mathscr{Y}) \cap \mathscr{O}$. Moreover, the sequence $(V^m(x,y))_{m \ge 0}$ converges to v(x,y) uniformly on all compact subsets of $\mathbb{C} \times \mathscr{Y}$. **Proof.** We give the main line of the proof, which is adapted from the very elegant technique used by Saff in the univariate case [17]. Define

$$\tilde{U}^{m}(x, y) = \frac{U^{m}(x, y)}{s_{m+n}^{m}(y)}, \quad \tilde{V}^{m}(x, y) = \frac{V^{m}(x, y)}{s_{m+n}^{m}(y)}$$

After dividing in f the numerator and the denominator by the function v_n (which does not vanish on \mathscr{Y}), the function f can be put in the following form which fits the form \tilde{U}^m/\tilde{V}^m of $[m,n]_{f_n}$:

$$f(x, y) = \frac{h(x, y)}{g(x, y)}, \quad g(x, y) = \sum_{i=0}^{n-1} g_i(y)x^i + x^n,$$

where the functions h and g are meromorphic on $D(0, \rho_1, \rho_2)$ and holomorphic on $D(0, \rho_1) \times \mathscr{Y}$. The idea is to search $\tilde{V}^m(x, y)$ under the form

$$\tilde{V}^{m}(x, y) = g(x, y) + \sum_{k=0}^{n-1} w_{k}^{m}(y)W_{k}(x, y),$$

where $W_0 \equiv 1$, $W_k(x, y) = (x - \alpha_1(y)) \cdots (x - \alpha_k(y))$ is a polynomial in x of degree k, holomorphic on $\mathbb{C} \times \mathscr{Y}$, and to reformulate the problem as follows.

For fixed $y \in \mathscr{Y}$, let $\pi_m(x, y)$ be the Taylor expansion of degree m + n at x = 0 of the function $x \mapsto \tilde{V}^m(x, y)h(x, y)$. The coefficients $w_k^m(y)$ are chosen in such a way that the polynomial in x, $\pi_m(x, y)$ vanishes at the *n* roots $\alpha_k(y)$ of g(., y). Hence, there exists a polynomial in x, $\tilde{U}^m(x, y)$ such that $\pi_m(x, y) = \tilde{U}^m(x, y)g(x, y)$, and it follows that $(h\tilde{V}^m - g\tilde{U}^m)(x, y) = O(x^{m+n+1})$. If $\tilde{V}^m(0, y) \neq 0$, these conditions coincide with the ones defining U^m and V^m .

Owing to Hermite's formula

$$\pi_m(x,y) = \frac{1}{2i\pi} \int_{|z|=\rho_1} \left(1 - \left(\frac{x}{z}\right)^{m+n+1} \right) \frac{\tilde{V}^m(z,y)h(z,y)}{z-x} \, \mathrm{d}z,\tag{35}$$

the coefficients $w_0^m(y), \ldots, w_{n-1}^m(y)$ are solution to the system

$$\sum_{k=0}^{n-1} A_{jk}^{m}(y) w_{k}^{m}(y) = B_{j}^{m}(y), \quad j = 1, 2, ..., n,$$

$$A_{jk}^{m}(y) = \frac{1}{2i\pi} \int_{|z|=\rho_{1}} \left(1 - \left(\frac{\alpha_{j}(y)}{z}\right)^{m+n+1} \right) \frac{W_{k}(z, y)h(z, y)}{z - \alpha_{j}(y)} \, \mathrm{d}z$$

$$B_{j}^{m}(y) = \frac{1}{2i\pi} \int_{|z|=\rho_{1}} \left(\frac{\alpha_{j}(y)}{z}\right)^{m+n+1} \frac{g(z, y)h(z, y)}{z - \alpha_{j}(y)} \, \mathrm{d}z,$$

which converges uniformly on \mathscr{K}_y to a triangular, homogeneous and invertible system. Thus, for $m \ge m_0$ sufficiently large and $y \in \mathscr{K}_y$, the coefficients $w_0^m(y), \ldots, w_{n-1}^m(y)$ are uniquely determined, holomorphic in y, they converge uniformly to zero, and \tilde{V}^m converges uniformly to g on all compact subsets of $\mathbb{C} \times \mathscr{K}_y$. The proof is completed as in Section 2.5. \Box

3.3. The PadéoPadé approximants

The Padé \circ Padé approximants were introduced by Chaffy [3]. In Section 3.2 were defined the intermediate Padé approximant $[m,n]_{f_y}(x) = U^m(x,y)/V^m(x,y)$ of the function f = u/v, where U^m and V^m belong to $\mathbb{C}[[y]][x]$, and have respective degrees m and n in x. The basic idea is now to compute an [m',n'] Padé approximant of $[m,n]_{f_y}(x)$ with respect to the variable y.

Some restrictions are made which are more technical than really necessary. For instance, a disk $D(0, \rho_3)$ is substituted for \mathscr{Y} because, in order to apply once more Lemma 3.1, one needs to be sure that the function $y \mapsto v(x_0, y)$ has exactly n' simple zeros in $D(0, \rho_3)$. The Padé \circ Padé approximant is defined locally in x in the following way.

Let $\rho_3 > 0$ be such that $D(0, \rho_3) \subset \mathscr{Y}$ and $v(0, y) \neq 0$ for all $y \in D(0, \rho_3)$. Let $x_0 \in D(0, \rho_1)$ be fixed such that $v(x_0, 0) \neq 0$, and suppose that the function $y \mapsto v(x_0, y)$ has exactly n' simple zeros in $D(0, \rho_3)$. The case of zeros of constant multiplicity could also be considered.

Recall that $V^m(x, y)$ converges to v(x, y) uniformly on all compact subsets of $\mathbb{C} \times \mathscr{Y}$. It follows from the implicit functions theorem, Lemma 3.1 and Rouché's theorem, that there exists an integer m_0 and a neighborhood $\mathscr{V}(x_0)$ of x_0 such that for $m \ge m_0$ and $(x, y) \in \mathscr{V}(x_0) \times D(0, \rho_3)$, v(x, y) and $V^m(x, y)$ can be written under the form

$$v(x, y) = c \prod_{i=1}^{n'} (y - \beta_i(x)) w(x, y),$$

$$V^m(x, y) = c_m \prod_{i=1}^{n'} (y - \beta_i^m(x)) W^m(x, y),$$

where the functions β_i and β_i^m are holomorphic and do not vanish on $\mathscr{V}(x_0)$, the functions w and W^m are holomorphic and do not vanish on $\mathscr{V}(x_0) \times \mathscr{Y}$. Moreover $\lim_{m\to\infty} \beta_i^m(x) = \beta_i(x)$ uniformly on all compact subsets of $\mathscr{V}(x_0)$.

Definition 3.1. For fixed $m \ge 0$, let $s(x, y) = [m, n]_{f_y}(x)$. The Padé \circ Padé approximant $[m', n']_y \circ [m, n]_x(f)$, if it exists, is defined on $\mathscr{V}(x_0)$ by

$$[m',n']_{v} \circ [m,n]_{x}(f)(x,y) = [m',n']_{s_{x}}(y)$$

Remark 3.1. The rational fraction $r(x, y) = [m', n']_y \circ [m, n]_x(f)$ has the following interpolation property:

$$\partial_x^k r(0,0) = \partial_x^k f(0,0), \quad 0 \le k \le m+n,$$

$$\partial_y^k r(x,0) = \partial_y^k s(x,0), \quad 0 \le k \le m'+n', \ \forall x \in \mathscr{V}(x_0).$$

If $0 \in \mathcal{V}(x_0)$ (for example if $x_0=0$), then the Padé \circ Padé satisfies at the origin the usual interpolation property. This follows from $s(0, y) = [m, n]_{f_y}(0) = f(0, y)$ for small |y|.

3.3.1. Convergence

We are now in the same position than the beginning of the first step if we exchange the variables x and y, and substitute $[m,n]_{f_y}(x)$ for the function f. The next theorem [3] follows from Lemma 3.1 applied first to the function $[m,n]_{f_y}(x)$, next to the function f_y .

Theorem 3.2 (Chaffy-Camus [3]). There exists an integer m'_0 and a neighborhood $\mathscr{V}(x_0)$ of x_0 such that for all $m' \ge m'_0$ and all $x \in \mathscr{V}(x_0)$, there is a unique Padé approximant $[m',n']_y \circ [m,n]_x(f)$. The Padé \circ Padé approximants converge in the following sense:

$$\lim_{m \to \infty} \left(\lim_{m' \to \infty} \left[m', n' \right]_{y} \circ [m, n]_{x}(f)(x, y) \right) = f(x, y)$$

uniformly on all compact subsets of $\{(x, y) \in \mathscr{V}(x_0) \times D(0, \rho_3), v(x, y) \neq 0\}$.

3.3.2. Computation

In order to compute $[m',n']_y \circ [m,n]_x(f)$, one first need to compute the double power series expansion of $[m,n]_{f_y}(x)$. Using symbolic computation, one expands fraction (31), $[m,n]_{f_y}(x) = (\sum_{i=0}^m s_i^m(y)x^i)/(1 + \sum_{i=1}^n s_{m+i}^m(y)x^i)$, in the form

$$[m,n]_{f_{y}}(x) = \sum_{j \ge 0} b_{j}(x)y^{j},$$
(36)

where the b_j 's are rational fractions, solution in $\mathbb{C}(x)$ to (37)-(38). Observe that $b_0(x)$ is a fraction with degrees at most [m,n], and more generally $b_j(x)$ is a fraction with degrees at most [m + jn, (j + 1)n]. Next the Padé \circ Padé approximant of f is obtained by computing the univariate Padé approximant of (36) with respect to y. Here again symbolic computation is used to solve the associated linear system in the field $\mathbb{C}(x)$, and the degrees in x will increase once more. The algorithm is the following.

Algorithm 4: PRPA computation

- 1. Use Algorithm 3 to compute the coefficients $s_{i,j}^m$, $0 \le i \le m + n$, $0 \le j \le m' + n'$ (in Algorithm 3, substitute $0 \le j \le m' + n'$ for $j \ge 0$).
- 2. Solve in $\mathbb{C}(x)$ the following triangular system (using symbolic computation):

$$\left(1 + \sum_{i=1}^{n} s_{m+i,0}^{m} x^{i}\right) b_{0}(x) = \sum_{i=0}^{m} s_{i,0}^{m} x^{i},$$
(37)

$$\sum_{k=0}^{j} \left(\sum_{i=1}^{n} s_{m+i,j-k}^{m} x^{i} \right) b_{k}(x) = \sum_{i=0}^{m} s_{i,j}^{m} x^{i}, \quad j = 1, 2, \dots, n' + m'.$$
(38)

3. Solve the system (using symbolic computation):

$$\begin{pmatrix} b_{m'-n'+1}(x) & \dots & b_{m'}(x) \\ \vdots & & \vdots \\ b_{m'}(x) & \dots & b_{m'+n'-1}(x) \end{pmatrix} \begin{pmatrix} q_{n'}(x) \\ \vdots \\ q_1(x) \end{pmatrix} = - \begin{pmatrix} b_{m'+1}(x) \\ \vdots \\ b_{m'+n'}(x) \end{pmatrix},$$

where $b_i(x) := 0$ if i < 0, and define

$$Q(x, y) = 1 + \sum_{i=0}^{n'} q_i(x) y^i$$
4. Compute

$$p_i(x) = \sum_{j=0}^{n'} q_j(x) b_{i-j}(x), \quad 0 \le i \le m',$$

where $q_0(x) := 1$, and define

$$P(x, y) = \sum_{i=0}^{m'} p_i(x) y^i.$$

5. The fraction P(x, y)/Q(x, y) is the Padé \circ Padé of the function f.

3.4. The nested multivariate approximants

The nested multivariate approximants were introduced in [10]. Consider a fraction $R \in \mathbb{C}(y)(x)$ of the form

$$R(x,y) = \frac{P(x,y)}{Q(x,y)} = \frac{\sum_{i=0}^{m} r_i(y) x^i}{1 + \sum_{i=1}^{n} r_{m+i}(y) x^i},$$
(39)

where the $r_i(y)$ are also fractions:

$$r_i(y) = \frac{p_i(y)}{q_i(y)} = \frac{\sum_{j=0}^{n_i} p_{ij} y^j}{1 + \sum_{j=1}^{n_i} q_{ij} y^j}, \quad 0 \le i \le n+m$$
(40)

with

$$m + n = d_1, \qquad m_i + n_i = d_2, \quad 0 \le i \le m + n,$$
(41)

$$\deg Q(x,0) = n, \qquad \deg q_i = n_i, \quad 0 \le i \le n + m.$$
(42)

Let $E(d_1, d_2) = \{0, 1, \dots, d_1\} \times \{0, 1, \dots, d_2\}$. For $\alpha = (\alpha_1, \alpha_2) \in \mathbb{N}^2$, we denote by ∂^{α} the usual differential operator $\partial^{\alpha} = \partial^{|\alpha|} / \partial_x^{\alpha_1} \partial_y^{\alpha_2}$.

Definition 3.2. Consider the following equation:

$$\partial^{\alpha} R(0,0) = \partial^{\alpha} f(0,0) \quad \forall \alpha \in E(d_1, d_2).$$

$$\tag{43}$$

If the fraction R (39) is the unique solution to this equation, it is called the $[m, n, (m_i), (n_i)]$ nested Padé approximant of the function f and denoted by $[m, n, (m_i), (n_i), x, y]_f$.

Similarly to the univariate case, existence occurs as soon as the degrees of the numerators are sufficiently large (cf. Theorem 3.4). A sufficient condition for uniqueness, which implies consistency, is the following [11]. The fraction R (39) is said irreducible if the fractions $x \mapsto R(x,0)$ and r_i , $0 \le i \le n + m$ are irreducible.

Proposition 3.3. If the fraction R (39) is a solution to (43) and is irreducible, then it is the unique solution to (43).

3.4.1. Computation

Eq. (43) is a nonlinear system of $(d_1+1)(d_2+1)$ equations, with the same number of unknowns. However the solution of this system can be obtained in two steps by solving small linear systems. The first step has been described in Section 3.2 and Algorithm 3, where the coefficients $s_i^m(y)$ of the intermediate Padé were obtained. The second step is the following.

For a given *m* and $0 \le i \le d_1$, degrees m_i and n_i are chosen in such a way that $m_i + n_i = d_2$ (see, e.g., [9] for the choice of the degrees). We suppose here that the following Padé approximants:

$$r_i(y) = [m_i, n_i]_{s^m}, \quad 0 \leq i \leq d_1,$$

exist in the sense of definition (29), that their denominators are of degree n_i , and that $r_{m+n}(0) \neq 0$. Let

$$R(x, y) = \frac{\sum_{i=0}^{m} r_i(y) x^i}{1 + \sum_{i=1}^{n} r_{m+i}(y) x^i}.$$
(44)

It can easily be proved that if this fraction R is irreducible, then it is the nested Padé approximant of order $[m, n, (m_i), (n_i)]$ of the function f [10]. The algorithm is the following, which needs no symbolic computation.

Algorithm 5: NPA computation

- 1. Use Algorithm 3 to compute the coefficients $s_{i,j}^m$, $0 \le i \le d_1$, $0 \le j \le d_2$ (in Algorithm 3, substitute $0 \le j \le d_2$ for $j \ge 0$).
- 2. For i = 1 to d_1 :
 - solve

$$egin{pmatrix} s^m_{i,m_i-n_i+1}&\ldots&s^m_{i,m_i}\dots&dots&dots\ s^m_{i,m_i}&dots&dots\ s^m_{i,m_i}&\ldots&s^m_{i,m_i+n_i-1} \end{pmatrix}egin{pmatrix} q_{n_i}\dots\ q_{n_i}\ q_{n_i}\dots\ q_{n_i}\dots\ q_{n_i}\dots\ q_{n_i}$$

where $s_{i,j} := 0$ if j < 0, and compute

$$p_j = \sum_{k=0}^{n_i} q_k s_{i,j-k}, \quad 0 \leqslant j \leqslant m_i,$$

• define

$$p_i(y) = \sum_{j=0}^{m_i} p_j y^j, \qquad q_i(y) = 1 + \sum_{j=1}^{n_i} q_j y^j, \qquad r_i(y) = \frac{p_i(y)}{q_i(y)}.$$

3. Define

$$P(x, y) = \sum_{i=0}^{m} r_i(y) x^i,$$
$$Q(x, y) = 1 + \sum_{i=1}^{n} r_{m+i}(y) x^i$$

4. The fraction P(x, y)/Q(x, y) is the nested Padé approximant of the function f.

3.4.2. Convergence

The next theorem was proved in [11].

Theorem 3.4 (Guillaume [11]). The sequence of nested Padé approximants converges uniformly to f on all compact subsets of $\mathcal{O} = \{(x, y) \in D(0, \rho_1, \rho_2), y \in \mathcal{Y}, v(x, y) \neq 0\}$ in the following sense: for all $\varepsilon > 0$ and all compact subsets $\mathcal{K} \subset \mathcal{O}$, there is an integer m_0 such that for all $m \ge m_0$, there exist integers d_0 and $n_i^m \le d_0$, $0 \le i \le m + n$, such that for all $d_2 \ge d_0$, the nested Padé approximant $[m, n, (d_2 - n_i^m), (n_i^m), x, y]_f$ of the function f is well defined and

 $\sup_{(x,y)\in\mathscr{K}}|f(x,y)-[m,n,(d_2-n_i^m),(n_i^m),x,y]_f(x,y)|<\varepsilon.$

Each n_i^m can be chosen equal to the number of poles (counted with multiplicity) within the disk $D(0, \rho_2)$ of the function s_i^m , $0 \le i \le m + n$.

Proof. The main line is the following. It follows from Lemma 3.1 that for $m \ge m_0$ sufficiently large, V^m is well defined and holomorphic around the origin. Hence the functions $s_i^m(y)$ are holomorphic around zero. Due to their construction (32), they are meromorphic on $D(0, \rho_2)$ and have a finite number n_i^m of poles (counted with multiplicity) within this disk. Owing to the Montessus de Ballore theorem, there is an integer d_0 such that the Padé approximants $[d_2 - n_i^m, n_i^m]_{s_i^m}$ are well defined for $d_2 \ge d_0$, and each sequence $([d_2 - n_i^m, n_i^m]_{s_i^m})_{d_2}$ converges to s_i^m uniformly on \mathcal{K} when $d_2 \to \infty$. \Box

Remark 3.2. Although the number of poles n_i^m of the functions s_i^m are not known, the technique described in [9] for counting the number of poles of meromorphic functions within a disk can be used here. Besides, the existence of an upper bound of the numbers n_i^m remains an open question, although numerical tests indicate they are bounded.

4. Final comments

An open question is whether convergence of more or less diagonal sequences in the LSPA table can be obtained for the approximation of general meromorphic functions. If such a result was obtained, it could be an improvement over the PRPA or NPA, where, due do their recursive construction, artificial singularities are present.

References

- [1] M. Bôcher, Introduction to Higher Algebra, The Macmillan Company, New York, 1952.
- [2] C. Brezinski, Least-squares orthogonal polynomials, J. Comput. Appl. Math. 46 (1993) 229-239.
- [3] C. Chaffy-Camus, Convergence uniforme d'une nouvelle classe d'approximants de Padé à plusieurs variables, C. R. Acad. Sci. Paris, Sér. I 306 (1988) 387–392.
- [4] J.S.R. Chisholm, Rational approximants defined from double power series, Math. Comp. 27 (1973) 841-848.
- [5] A. Cuyt, Multivariate Padé approximants, J. Math. Anal. Appl. 96 (1983) 283-293.
- [6] A. Cuyt, A Montessus de Ballore Theorem for Multivariate Padé Approximants, J. Approx. Theory 43 (1985) 43–52.
- [7] A. Cuyt, D.S. Lubinsky, A de Montessus theorem for multivariate homogeneous Padé approximants, Ann. Numer. Math. 4 (1997) 217–228.

- [8] A. Cuyt, L. Wuytack, in: Nonlinear Methods in Numerical Analysis, North-Holland Mathematics Studies, Vol. 136, Studies in Computational Mathematics, Vol. 1, North-Holland Publishing Co, Amsterdam, 1987.
- [9] B. Gleyse, V. Kaliaguine, On algebraic computation of number of poles of meromorphic functions in the unit disk, in: A. Cuyt (Ed.), Nonlinear Numerical Methods and Rational Approximation II, Kluwer Academic Publishers, Netherlands, 1994, pp. 241–246.
- [10] Ph. Guillaume, Nested Multivariate Padé Approximants, J. Comput. Appl. Math. 82 (1997) 149-158.
- [11] Ph. Guillaume, Convergence of the nested multivariate Padé approximants, J. Approx. Theory 94 (3) (1998) 455–466.
- [12] Ph. Guillaume, A. Huard, V. Robin, Generalized Multivariate Padé Approximants, J. Approx. Theory 95 (2) (1998) 203–214.
- [13] Ph. Guillaume, M. Masmoudi, Solution to the time-harmonic Maxwell's equations in a waveguide, use of higher order derivatives for solving the discrete problem, SIAM J. Numer. Anal. 34 (4) (1997) 1306–1330.
- [14] J. Karlsson, H. Wallin, Rational approximation by an interpolation procedure in several variables, in: E. Saaf, R. Varga (Eds.), Padé and Rational Approximation: Theory and Applications, Academic Press, New York, 1977, pp. 83–100.
- [15] S. Kida, Padé-type and Padé approximants in several variables, Appl. Numer. Math. 6 (1989/90) 371-391.
- [16] D. Levin, General order Padé-type rational approximants defined from double power series, J. Inst. Math. Appl. 18 (1976) 395–407.
- [17] E.B. Saff, An extension of Montessus de Ballore's theorem on the convergence of interpolating rational functions, J. Approx. Theory 6 (1972) 63–67.
- [18] B.L. Van der Waerden, Einführung in die algebraische Geometrie, Springer, Berlin, 1973.
- [19] H. Werner, Multivariate Padé Approximation, Numer. Math. 48 (1986) 429-440.



JOURNAL OF COMPUTATIONAL AND APPLIED MATHEMATICS

Journal of Computational and Applied Mathematics 121 (2000) 221-246

www.elsevier.nl/locate/cam

Numerical and asymptotic aspects of parabolic cylinder functions

Nico M. Temme*

CWI, P.O. Box 94079, 1090 GB Amsterdam, Netherlands

Received 19 August 1999

Abstract

Several uniform asymptotics expansions of the Weber parabolic cylinder functions are considered, one group in terms of elementary functions, another group in terms of Airy functions. Starting point for the discussion are asymptotic expansions given earlier by F.W.J. Olver. Some of his results are modified to improve the asymptotic properties and to enlarge the intervals for using the expansions in numerical algorithms. Olver's results are obtained from the differential equation of the parabolic cylinder functions; we mention how modified expansions can be obtained from integral representations. Numerical tests are given for three expansions in terms of elementary functions. In this paper only real values of the parameters will be considered. © 2000 Elsevier Science B.V. All rights reserved.

MSC: 33C15; 41A60; 65D20

Keywords: Parabolic cylinder functions; Uniform asymptotic expansion; Airy-type expansions; Numerical evaluation of special functions

1. Introduction

The solutions of the differential equation

$$\frac{d^2 y}{dz^2} - \left(\frac{1}{4}z^2 + a\right)y = 0,$$
(1.1)

are associated with the parabolic cylinder in harmonic analysis; see [20]. The solutions are called parabolic cylinder functions and are entire functions of z. Many properties in connection with physical applications are given in [4].

* Corresponding author.

E-mail address: nico.temme@cwi.nl (N.M. Temme).

As in [1, Chapter 19] and [17, Chapter 7], we denote two standard solutions of (1.1) by U(a,z), V(a,z). These solutions are given by the representations

$$U(a,z) = \sqrt{\pi} 2^{-(1/2)a} \left[\frac{2^{-1/4} y_1(a,z)}{\Gamma(3/4 + (1/2)a)} - \frac{2^{1/4} y_2(a,z)}{\Gamma(1/4 + (1/2)a)} \right],$$

$$V(a,z) = \frac{\sqrt{\pi} 2^{-(1/2)a}}{\Gamma((1/2) - a)} \left[\tan \pi \left(\frac{1}{2}a + \frac{1}{4} \right) \frac{2^{-1/4} y_1(a,z)}{\Gamma(3/4 + (1/2)a)} + \cot \pi \left(\frac{1}{2}a + \frac{1}{4} \right) \frac{2^{1/4} y_2(a,z)}{\Gamma(1/4 + (1/2)a)} \right],$$
(1.2)

where

$$y_{1}(a,z) = e^{-(1/4)z^{2}} {}_{1}F_{1}\left(-\frac{1}{2}a + \frac{1}{4}, \frac{1}{2}; -\frac{1}{2}z^{2}\right) = e^{(1/4)z^{2}} {}_{1}F_{1}\left(\frac{1}{2}a + \frac{1}{4}, \frac{1}{2}; \frac{1}{2}z^{2}\right),$$

$$y_{2}(a,z) = ze^{-(1/4)z^{2}} {}_{1}F_{1}\left(\frac{1}{2}a + \frac{3}{4}, \frac{3}{2}; \frac{1}{2}z^{2}\right) = ze^{(1/4)z^{2}} {}_{1}F_{1}\left(-\frac{1}{2}a + \frac{3}{4}, \frac{3}{2}; -\frac{1}{2}z^{2}\right)$$
(1.3)

and the confluent hypergeometric function is defined by

$${}_{1}F_{1}(a,c;z) = \sum_{n=0}^{\infty} \frac{(a)_{n}}{(c)_{n}} \frac{z^{n}}{n!}$$
(1.4)

with $(a)_n = \Gamma(a+n)/\Gamma(a)$, n = 0, 1, 2, ...Another notation found in the literature is

$$D_{\nu}(z) = U(-\nu - \frac{1}{2}, z).$$

There is a relation with the Hermite polynomials. We have

$$U(-n - \frac{1}{2}, z) = 2^{-n/2} e^{-(1/4)z^2} H_n(z/\sqrt{2}),$$

$$V(n + \frac{1}{2}, z) = 2^{-n/2} e^{(1/4)z^2} (-i)^n H_n(iz/\sqrt{2}).$$
(1.5)

Other special cases are error functions and Fresnel integrals.

The Wronskian relation between U(a,z) and V(a,z) reads

$$U(a,z)V'(a,z) - U'(a,z)V(a,z) = \sqrt{2/\pi},$$
(1.6)

which shows that U(a,z) and V(a,z) are independent solutions of (1.1) for all values of a. Other relations are

$$U(a,z) = \frac{\pi}{\cos^2 \pi a \Gamma((1/2) + a)} [V(a, -z) - \sin \pi a V(a, z)],$$

$$V(a,z) = \frac{\Gamma((1/2) + a)}{\pi} [\sin \pi a U(a, z) + U(a, -z)].$$
(1.7)

The functions $y_1(a,z)$ and $y_2(a,z)$ are the simplest even and odd solutions of (1.1) and the Wronskian of this pair equals 1. From a numerical point of view, the pair $\{y_1, y_2\}$ is not a satisfactory pair (see [8]), because they have almost the same asymptotic behavior at infinity. The behavior of U(a,z) and V(a,z) is, for large positive z and $z \gg |a|$

$$U(a,z) = e^{-(1/4)z^2} z^{-a-(1/2)} [1 + \mathcal{O}(z^{-2})],$$

$$V(a,z) = \sqrt{2/\pi} e^{(1/4)z^2} z^{a-(1/2)} [1 + \mathcal{O}(z^{-2})].$$
(1.8)

Clearly, numerical computations of U(a,z) that are based on the representations in (1.2) should be done with great care, because of the loss of accuracy if z becomes large.

Eq. (1.1) has two turning points at $\pm 2\sqrt{-a}$. For real parameters they become important if *a* is negative, and the asymptotic behavior of the solutions of (1.1) as $a \to -\infty$ changes significantly if *z* crosses the turning points. At these points Airy functions are needed. By changing the parameters it is not difficult to verify that $U(-\frac{1}{2}\mu^2, \mu t\sqrt{2})$ and $V(-\frac{1}{2}\mu^2, \mu t\sqrt{2})$ satisfy the simple equation

$$\frac{d^2 y}{dt^2} - \mu^4 (t^2 - 1)y = 0 \tag{1.9}$$

with turning points at $t = \pm 1$. For physical applications, negative *a*-values are most important (with special case the real Hermite polynomials, see (1.5)). For positive *a* we can use the notation $U(\frac{1}{2}\mu^2, \mu t \sqrt{2})$ and $V(\frac{1}{2}\mu^2, \mu t \sqrt{2})$, which satisfy the equation

$$\frac{d^2 y}{dt^2} - \mu^4 (t^2 + 1)y = 0.$$
(1.10)

The purpose of this paper is to give several asymptotic expansions of U(a,z) and V(a,z) that can be used for computing these functions for the case that at least one of the real parameters is large. In [10] an extensive collection of asymptotic expansions for the parabolic cylinder functions as $|a| \to \infty$ has been derived from the differential equation (1.1). The expansions are valid for complex values of the parameters and are given in terms of elementary functions and Airy functions. In Section 2 we mention several expansions in terms of elementary functions derived by Olver and modify some his results in order to improve the asymptotic properties of the expansions, to enlarge the intervals for using the expansions. In Section 3 we give similar results for expansions in terms of Airy functions. In Section 4 we give information on how to obtain the modified results by using integral representations of the parabolic cylinder functions. Finally we give numerical tests for three expansions in terms of elementary functions. Finally we give numerical tests for three expansions in terms of elementary functions. Finally we give numerical tests for three expansions in terms of elementary functions, with a few number of terms in the expansions. Only real parameters are considered in this paper.

1.1. Recent literature on numerical algorithms

Recent papers on numerical algorithms for the parabolic cylinder functions are given in [14] (Fortran; U(n,x) for natural *n* and positive *x*) and [13] (Fortran; U(a,x), V(a,x), *a* integer and half-integer and $x \ge 0$). The methods are based on backward and forward recursion.

Baker [2] gives programs in C for U(a,x), V(a,x), and uses representations in terms of the confluent hypergeometric functions and asymptotic expressions, including those involving Airy functions. Zhang and Jin [23] gives Fortran programs for computing U(a,z), V(a,z) with real orders and real argument, and for half-integer order and complex argument. The methods are based on

recursions, Maclaurin series and asymptotic expansions. They refer also to [3] for the evaluation of $U(-ia, ze^{(1/4)\pi i})$ for real *a* and *z* (this function is a solution of the differential equation $y'' + (\frac{1}{4}z^2 - a)y = 0$). Thompson [19] uses series expansions and numerical quadrature; Fortran and *C* programs are given, and *Mathematica* cells to make graphical and numerical objects.

Maple has algorithms for hypergeometric functions, which can be used in (1.2) and (1.3) [5]. *Mathematica* refers for the parabolic cylinder functions to their programs for the hypergeometric functions [21] and the same advice is given in [12]. For a survey on the numerical aspects of special functions we refer to [7].

2. Expansions in terms of elementary functions

2.1. The case $a \leq 0, z > 2\sqrt{-a}, -a + z \gg 0$

Olver's expansions in terms of elementary functions are all based on the expansion O-(4.3)¹

$$U(-\frac{1}{2}\mu^2, \mu t \sqrt{2}) \sim \frac{g(\mu) e^{-\mu^2 \xi}}{(t^2 - 1)^{1/4}} \sum_{s=0}^{\infty} \frac{\mathscr{A}_s(t)}{\mu^{2s}}$$
(2.1)

as $\mu \to \infty$, uniformly with respect to $t \in [1 + \varepsilon, \infty)$; ε is a small positive number and ξ is given by

$$\xi = \frac{1}{2}t\sqrt{t^2 - 1} - \frac{1}{2}\ln[t + \sqrt{t^2 - 1}].$$
(2.2)

The expansion is valid for complex parameters in large domains of the μ - and *t*-planes; details on these domains are not given here.

The coefficients $\mathscr{A}_s(t)$ are given by the recursion relation

$$\mathscr{A}_{s+1}(t) = \frac{1}{2} \frac{1}{\sqrt{t^2 - 1}} \frac{d\mathscr{A}_s(t)}{dt} + \frac{1}{8} \int_{c_{s+1}}^t \frac{3u^2 + 2}{(u^2 - 1)^{5/2}} \mathscr{A}_s(u) \, \mathrm{d}u, \quad \mathscr{A}_0(t) = 1,$$
(2.3)

where the constants c_s can be chosen in combination with the choice of $g(\mu)$. Olver chose the constants such that

$$\mathscr{A}_{s}(t) = \frac{u_{s}(t)}{(t^{2} - 1)^{3s/2}},$$
(2.4)

where the $u_s(t)$ are polynomials in t of degree 3s, (s odd), 3s - 2 (s even, $s \ge 2$). The first few are

$$u_0(t) = 1, \quad u_1(t) = \frac{t(t^2 - 6)}{24}, \quad u_2(t) = \frac{-9t^2 + 249t^2 + 145}{1152}$$

and they satisfy the recurrence relation

$$(t2 - 1)u'_{s}(t) - 3stu_{s}(t) = r_{s-1}(t),$$
(2.5)

where

$$8r_s(t) = (3t^2 + 2)u_s(t) - 12(s+1)tr_{s-1}(t) + 4(t^2 - 1)r'_{s-1}(t).$$

The quantity $g(\mu)$ in (2.1) is only available in the form of an asymptotic expansion

$$g(\mu) \sim h(\mu) \left(\sum_{s=0}^{\infty} \frac{g_s}{\mu^{2s}}\right)^{-1},$$
 (2.6)

¹ We refer to Olver's equations by writing O-(4.3), and so on.

where

$$h(\mu) = 2^{-(1/4)\mu^2 - (1/4)} e^{-(1/4)\mu^2} \mu^{(1/2)\mu^2 - (1/2)},$$
(2.7)

$$g_0 = 1$$
, $g_1 = \frac{1}{24}$, $g_3 = -\frac{2021}{207360}$, $g_{2s} = 0$ (s = 1, 2, ...),

and in general

$$g_s = \lim_{t \to \infty} \mathscr{A}_s(t).$$
(2.8)

2.1.1. Modified expansions

We modify the expansion in (2.1) by writing

$$U(-\frac{1}{2}\mu^{2},\mu t\sqrt{2}) = \frac{h(\mu)e^{-\mu^{2}\xi}}{(t^{2}-1)^{1/4}}F_{\mu}(t), \quad F_{\mu}(t) \sim \sum_{s=0}^{\infty} \frac{\phi_{s}(\tau)}{\mu^{2s}},$$
(2.9)

where $h(\mu)$ and ξ are as before, and

$$\tau = \frac{1}{2} \left[\frac{t}{\sqrt{t^2 - 1}} - 1 \right].$$
(2.10)

The coefficients $\phi_s(\tau)$ are polynomials in τ , $\phi_0(\tau) = 1$, and are given by the recursion

$$\phi_{s+1}(\tau) = -4\tau^2(\tau+1)^2 \frac{\mathrm{d}}{\mathrm{d}\tau} \phi_s(\tau) - \frac{1}{4} \int_0^\tau (20\tau'^2 + 20\tau' + 3)\phi_s(\tau') \,\mathrm{d}\tau'.$$
(2.11)

This recursion follows from (2.3) by substituting $t = (\tau + \frac{1}{2})/\sqrt{\tau(\tau + 1)}$, which is the inverse of the relation in (2.10). Explicitly,

$$\begin{split} \phi_0(\tau) &= 1, \\ \phi_1(\tau) &= -\frac{\tau}{12} (20\tau^2 + 30\tau + 9), \\ \phi_2(\tau) &= \frac{\tau^2}{288} (6160\tau^4 + 18\,480\tau^3 + 19\,404\tau^2 + 8028\tau + 945), \\ \phi_3(\tau) &= -\frac{\tau^3}{51\,840} (27\,227\,200\tau^6 + 122\,522\,400\tau^5 + 220\,540\,320\tau^4 \\ &+ 200\,166\,120\tau^3 + 94\,064\,328\tau^2 + 20\,545\,650\tau + 1\,403\,325), \end{split}$$
(2.12)

where τ is given in (2.10). Observe that $\lim_{t\to\infty} \tau(t) = 0$ and that all shown coefficients $\phi_s(\tau)$ vanish at infinity for s > 0. These properties of $\phi_s(\tau)$ follow by taking different constants c_s than Olver did in (2.3). In fact we have the relation

$$\sum_{s=0}^{\infty} \frac{g_s}{\mu^{2s}} \sum_{s=0}^{\infty} \frac{\phi_s(\tau)}{\mu^{2s}} \sim \sum_{s=0}^{\infty} \frac{u_s(t)}{(t^2-1)^{(3/2)s} \mu^{2s}},$$

where the first series appears in (2.6). Explicitly,

$$u_s(t) = (t^2 - 1)^{(3/2)s} \sum_{j=0}^{s} g_{s-j} \phi_j(\tau).$$
(2.13)

The relation (2.13) can easily be verified for the early coefficients, but it holds because of the unicity of Poincaré-type asymptotic expansions.

The expansion in (2.9) has several advantages compared with (2.1):

- (i) In the recursion relation (2.5), both u_s and u'_s occur in the left-hand side. By using computer algebra it is not difficult to compute any number of coefficients u_s , but the relation for the polynomials $\phi_s(\tau)$ is simpler than (2.5), with this respect.
- (ii) The quantity $h(\mu)$ in (2.9) is defined as an exact relation, and not, as $g(\mu)$ in (2.1), by an asymptotic expansion (cf. (2.6)).
- (iii) Most important, the expansion in (2.9) has a double asymptotic property: it holds if one or both parameters t and μ are large, and not only if μ is large.

For the function V(a,z) we have

$$V(-\frac{1}{2}\mu^2,\mu t\sqrt{2}) = \frac{e^{\mu^2\xi}}{\mu\sqrt{\pi}h(\mu)(t^2-1)^{(1/4)}}P_{\mu}(t), \quad P_{\mu}(t) \sim \sum_{s=0}^{\infty} (-1)^s \frac{\phi_s(\tau)}{\mu^{2s}}, \tag{2.14}$$

where the $\phi_s(\tau)$ are the same as in (2.9). This expansion is a modification of O-(11.19) (see also O-(2.12)).

For the derivatives we can use the identities

$$\frac{\mathrm{d}}{\mathrm{d}t} \frac{\mathrm{e}^{-\mu^{2}\xi}}{(t^{2}-1)^{1/4}} F_{\mu}(t) = -\mu^{2}(t^{2}-1)^{1/4} \mathrm{e}^{-\mu^{2}\xi} G_{\mu}(t), \quad G_{\mu}(t) \sim \sum_{s=0}^{\infty} \frac{\psi_{s}(\tau)}{\mu^{2s}},
\frac{\mathrm{d}}{\mathrm{d}t} \frac{\mathrm{e}^{+\mu^{2}\xi}}{(t^{2}-1)^{1/4}} P_{\mu}(t) = +\mu^{2}(t^{2}-1)^{1/4} \mathrm{e}^{+\mu^{2}\xi} Q_{\mu}(t), \quad Q_{\mu}(t) \sim \sum_{s=0}^{\infty} (-1)^{s} \frac{\psi_{s}(\tau)}{\mu^{2s}}.$$
(2.15)

The coefficients ψ_s can be obtained from the relation

$$\psi_s(t) = \phi_s(\tau) + 2\tau(\tau+1)(2\tau+1)\phi_{s-1}(\tau) + 8\tau^2(\tau+1)^2 \frac{\mathrm{d}\phi_{s-1}(\tau)}{\mathrm{d}\tau}, \qquad (2.16)$$

 $s = 0, 1, 2, \dots$ The first few are

$$\begin{split} \psi_0(t) &= 1, \\ \psi_1(t) &= \frac{\tau}{12} (28\tau^2 + 42\tau + 15), \\ \psi_2(t) &= -\frac{\tau^2}{288} (7280\tau^4 + 21\,840\tau^3 + 23\,028\tau^2 + 9684\tau + 1215), \\ \psi_3(t) &= \frac{\tau^3}{51\,840} (30\,430\,400\tau^6 + 136\,936\,800\tau^5 + 246\,708\,000\tau^4 \\ &+ 224\,494\,200\tau^3 + 106\,122\,312\tau^2 + 23\,489\,190\tau + 1\,658\,475). \end{split}$$

This gives the modifications (see O(4.13))

$$U'(-\frac{1}{2}\mu^2,\mu t\sqrt{2}) = -\frac{\mu}{\sqrt{2}}h(\mu)(t^2-1)^{1/4}e^{-\mu^2\xi}G_{\mu}(t), \quad G_{\mu}(t) \sim \sum_{s=0}^{\infty}\frac{\psi_s(\tau)}{\mu^{2s}}$$
(2.18)

and

$$V'(-\frac{1}{2}\mu^2,\mu t\sqrt{2}) = \frac{(t^2-1)^{1/4} e^{\mu^2 \xi}}{\sqrt{2\pi}h(\mu)} Q_{\mu}(t), \quad Q_{\mu}(t) \sim \sum_{s=0}^{\infty} (-1)^s \frac{\psi_s(\tau)}{\mu^{2s}}.$$
(2.19)

Remark 2.1. The functions $F_{\mu}(t)$, $G_{\mu}(t)$, $P_{\mu}(t)$ and $Q_{\mu}(t)$ introduced in the asymptotic representations satisfy the following exact relation:

$$F_{\mu}(t)Q_{\mu}(t) + G_{\mu}(t)P_{\mu}(t) = 2.$$
(2.20)

This follows from the Wronskian relation (1.6). The relation in (2.20) provides a convenient possibility for checking the accuracy in numerical algorithms that use the asymptotic expansions of $F_{\mu}(t), G_{\mu}(t), P_{\mu}(t)$ and $Q_{\mu}(t)$.

2.2. The case $a \leq 0, \ z < -2\sqrt{-a}, \ -a - z \gg 0$

For this case we mention the modification of O-(11.16). That is, for $t \ge 1 + \varepsilon$ we have the representations

$$U(-\frac{1}{2}\mu^{2},-\mu t\sqrt{2}) = \frac{h(\mu)}{(t^{2}-1)^{1/4}} \left[\sin(\frac{1}{2}\pi\mu^{2})e^{-\mu^{2}\xi}F_{\mu}(t) + \frac{\Gamma(1/2+(1/2)\mu^{2})\cos((1/2)\pi\mu^{2})}{\mu\sqrt{\pi}h^{2}(\mu)}e^{\mu^{2}\xi}P_{\mu}(t) \right], \qquad (2.21)$$

where $F_{\mu}(t)$ and $P_{\mu}(t)$ have the expansions given in (2.9) and (2.14), respectively. An expansion for $V(-\frac{1}{2}\mu^2, -\mu t\sqrt{2})$ follows from the second line in (1.7), (2.9) and (2.21). A few manipulations give

$$V(-\frac{1}{2}\mu^{2},-\mu t\sqrt{2}) = \frac{h(\mu)}{(t^{2}-1)^{1/4}\Gamma(1/2+(1/2)\mu^{2})} \left[\cos(\frac{1}{2}\pi\mu^{2})e^{-\mu^{2}\xi}F_{\mu}(t) -\frac{\Gamma(1/2+(1/2)\mu^{2})\sin((1/2)\pi\mu^{2})}{\mu\sqrt{\pi}h^{2}(\mu)}e^{\mu^{2}\xi}P_{\mu}(t)\right].$$
(2.22)

Expansions for the derivatives follow from the identities in (2.15). If $a = -\frac{1}{2}\mu^2 = -n - \frac{1}{2}$, n = 0, 1, 2, ..., the cosine in (2.21) vanishes, and, hence, the dominant part vanishes. This is the Hermite case, cf. (1.5).

2.3. The case $a \ll 0, -2\sqrt{-a} < z < 2\sqrt{-a}$

For negative a and -1 < t < 1 the expansions are essentially different, because now oscillations with respect to t occur. We have (O-(5.11) and O-(5.23))

$$U(-\frac{1}{2}\mu^{2},\mu t\sqrt{2}) \sim \frac{2g(\mu)}{(1-t^{2})^{1/4}} \left[\cos(\mu^{2}\eta - \frac{1}{4}\pi) \sum_{s=0}^{\infty} \frac{(-1)^{s} u_{2s}(t)}{(1-t^{2})^{3s} \mu^{4s}} - \sin(\mu^{2}\eta - \frac{1}{4}\pi) \sum_{s=0}^{\infty} \frac{(-1)^{s} u_{2s+1}(t)}{(1-t^{2})^{3s+(3/2)} \mu^{4s+2}} \right],$$
(2.23)

with $u_s(t)$ defined in (2.5) and $g(\mu)$ in (2.6), and

$$U'(-\frac{1}{2}\mu^{2},\mu t\sqrt{2}) \sim \mu \sqrt{2}g(\mu)(1-t^{2})^{1/4} \left[\sin(\mu^{2}\eta - \frac{1}{4}\pi) \sum_{s=0}^{\infty} \frac{(-1)^{s} v_{2s}(t)}{(1-t^{2})^{3s} \mu^{4s}} + \cos(\mu^{2}\eta - \frac{1}{4}\pi) \sum_{s=0}^{\infty} \frac{(-1)^{s} v_{2s+1}(t)}{(1-t^{2})^{3s+(3/2)} \mu^{4s+2}} \right],$$
(2.24)

as $\mu \to \infty$, uniformly with respect to $|t| \leq 1 - \varepsilon$, where the coefficients v_s are given by (see O-(4.15))

$$v_s(t) = u_s(t) + \frac{1}{2}tu_{s-1}(t) - r_{s-2}(t)$$
(2.25)

and

$$\eta = \frac{1}{2}\arccos t - \frac{1}{2}t\sqrt{1 - t^2}.$$
(2.26)

For the function V(a,z) we have (O-(11.20) and O-(2.12))

$$V(-\frac{1}{2}\mu^{2},\mu t\sqrt{2}) \sim \frac{2g(\mu)}{\Gamma(1/2+(1/2)\mu^{2})(1-t^{2})^{1/4}} \left[\cos(\mu^{2}\eta + \frac{1}{4}\pi) \sum_{s=0}^{\infty} \frac{(-1)^{s} u_{2s}(t)}{(1-t^{2})^{3s} \mu^{4s}} - \sin(\mu^{2}\eta + \frac{1}{4}\pi) \sum_{s=0}^{\infty} \frac{(-1)^{s} u_{2s+1}(t)}{(1-t^{2})^{3s+(3/2)} \mu^{4s+2}} \right],$$

$$V'(-\frac{1}{2}\mu^{2},\mu t\sqrt{2}) \sim \frac{\mu\sqrt{2}g(\mu)(1-t^{2})^{1/4}}{\Gamma(1/2+(1/2)\mu^{2})} \left[\sin(\mu^{2}\eta + \frac{1}{4}\pi) \sum_{s=0}^{\infty} \frac{(-1)^{s} v_{2s}(t)}{(1-t^{2})^{3s}\mu^{4s}} + \cos(\mu^{2}\eta + \frac{1}{4}\pi) \sum_{s=0}^{\infty} \frac{(-1)^{s} v_{2s+1}(t)}{(1-t^{2})^{3s+(3/2)}\mu^{4s+2}} \right].$$
(2.27)

By using the Wronskian relation (1.6) it follows that we have the following asymptotic identity

$$\sum_{s=0}^{\infty} \frac{(-1)^{s} u_{2s}(t)}{(1-t^{2})^{3s} \mu^{4s}} \sum_{s=0}^{\infty} \frac{(-1)^{s} v_{2s}(t)}{(1-t^{2})^{3s} \mu^{4s}} + \sum_{s=0}^{\infty} \frac{(-1)^{s} u_{2s+1}(t)}{(1-t^{2})^{3s+(3/2)} \mu^{4s+2}} \sum_{s=0}^{\infty} \frac{(-1)^{s} v_{2s+1}(t)}{(1-t^{2})^{3s+(3/2)} \mu^{4s+2}} \\ \sim \frac{\Gamma(1/2 + (1/2)\mu^{2})}{2\mu\sqrt{\pi}g^{2}(\mu)} \sim 1 - \frac{1}{576\mu^{4}} + \frac{2021}{2488320\mu^{8}} + \cdots$$
(2.28)

2.3.1. Modified expansions

We can give modified versions based on our earlier modifications, with $g(\mu)$ replaced with $h(\mu)$, and so on. Because in the present case t belongs to a finite domain, the modified expansions do not have the double asymptotic property. We prefer Olver's versions for this case.

This completes the description of U(a,z), U'(a,z), V(a,z), V'(a,z) in terms of elementary functions for negative values of a.

2.4. The case $a \ge 0$, $z \ge 0$, $a + z \gg 0$

For positive values of a the asymptotic behavior is rather simple because no oscillations occur now. The results follow from Olver's expansions O-(11.10) and O-(11.12). The modified forms are

$$U(\frac{1}{2}\mu^{2},\mu t\sqrt{2}) = \frac{\tilde{h}(\mu)e^{-\mu^{2}\tilde{\xi}}}{(t^{2}+1)^{1/4}}\tilde{F}_{\mu}(t), \quad \tilde{F}_{\mu}(t) \sim \sum_{s=0}^{\infty} (-1)^{s} \frac{\phi_{s}(\tilde{\tau})}{\mu^{2s}},$$
(2.29)

where

$$\tilde{\xi} = \frac{1}{2} [t\sqrt{1+t^2} + \ln(t+\sqrt{1+t^2})], \qquad (2.30)$$

$$\tilde{h}(\mu) = e^{(1/4)\mu^2} \mu^{-(1/2)\mu^2 - (1/2)} 2^{(1/4)\mu^2 - (1/4)}.$$
(2.31)

The coefficients ϕ_s in (2.29) are the same as in (2.9), with τ replaced by

$$\tilde{\tau} = \frac{1}{2} \left[\frac{t}{\sqrt{1+t^2}} - 1 \right].$$
 (2.32)

For the derivative we have

$$U'(\frac{1}{2}\mu^{2},\mu t\sqrt{2}) = -\frac{1}{\sqrt{2}}\mu\tilde{h}(\mu)(1+t^{2})^{1/4}e^{-\mu^{2}\tilde{\xi}}\tilde{G}_{a}(z), \quad \tilde{G}_{a}(z) \sim \sum_{s=0}^{\infty}(-1)^{s}\frac{\psi_{s}(\tilde{\tau})}{\mu^{2s}},$$
(2.33)

where $\psi_s(\tilde{\tau})$ is given in (2.16), with $\tilde{\tau}$ defined in (2.32).

2.5. The case $a \ge 0$, $z \le 0$, $a - z \gg 0$

Olver's expansion O-(11.10) and O-(11.12) cover both cases $z \ge 0$ and $z \le 0$. We have the modified expansions

$$U(\frac{1}{2}\mu^{2}, -\mu t\sqrt{2}) = \frac{\sqrt{2\pi}}{\Gamma(1/2 + (1/2)\mu^{2})} \frac{h(\mu)e^{\mu^{2}\tilde{\xi}}}{(1+t^{2})^{1/4}} \tilde{P}_{\mu}(t),$$

$$U'(\frac{1}{2}\mu^{2}, -\mu t\sqrt{2}) = -\frac{\mu}{\sqrt{2}} \frac{\sqrt{2\pi}}{\Gamma(1/2 + (1/2)\mu^{2})} h(\mu)e^{\mu^{2}\tilde{\xi}}(1+t^{2})^{1/4}\tilde{Q}_{\mu}(t),$$
(2.34)

where

$$ilde{P}_{\mu}(t)\sim\sum_{s=0}^{\infty}rac{\phi_{s}(ilde{ au})}{\mu^{2s}}, \quad ilde{Q}_{\mu}(t)\sim\sum_{s=0}^{\infty}rac{\psi_{s}(ilde{ au})}{\mu^{2s}}.$$

In Section 4.1.2 we give details on the derivation of these expansions.

Remark 2.2. By using the second relation in (1.7), the representations for V(a,z) and V'(a,z) for positive *a* can be obtained from the results for U(a,z) and U'(a,z) in (2.29), (2.33) and (2.34).

Remark 2.3. The functions $\tilde{F}_{\mu}(t)$, $\tilde{G}_{\mu}(t)$, $\tilde{P}_{\mu}(t)$ and $\tilde{Q}_{\mu}(t)$ introduced in (2.29), (2.32) and (2.34) satisfy the following exact relation:

$$\tilde{F}_{\mu}(t)\tilde{Q}_{\mu}(t) + \tilde{G}_{\mu}(t)\tilde{P}_{\mu}(t) = 2.$$
(2.35)



Fig. 1. Regions for the modified asymptotic expansions of U(a,z) given in Section 2 and the Airy-type expansions of Section 3 (which are valid in much larger domains than those indicated by the arrows).

This follows from the Wronskian relation

$$U(a,z)U'(a,-z) + U'(a,z)U(a,-z) = -\frac{\sqrt{2\pi}}{\Gamma(a+(1/2))}.$$

See also Remark 2.1.

Remark 2.4. The expansions of Sections 2.4 and 2.5 have the double asymptotic property: they are valid if the $a + |z| \to \infty$. In Sections 2.4 and 2.5 we consider the cases $z \ge 0$ and $\zeta \le 0$, respectively, as two separate cases. Olver's corresponding expansions O-(11.10) and O-(11.12) cover both cases and are valid for $-\infty < t < \infty$. As always, in Olver's expansions large values of μ are needed, whatever the size of t.

In Fig. 1 we show the domains in the t, a-plane where the various expansions of U(a, z) of this section are valid.

3. Expansions in terms of Airy functions

The Airy-type expansions are needed if z runs through an interval containing one of the turning points $\pm 2\sqrt{-a}$, that is, $t = \pm 1$.

3.1. The case $a \ll 0, z \ge 0$

We summarize the basic results O(8.11), O(8.15) and O(11.22) (see also O(2.12)):

$$U(-\frac{1}{2}\mu^{2},\mu t\sqrt{2}) = 2\pi^{1/2}\mu^{1/3}g(\mu)\phi(\zeta) \left[A_{i}(\mu^{4/3}\zeta)A_{\mu}(\zeta) + \frac{A_{i'}(\mu^{4/3}\zeta)}{\mu^{8/3}}B_{\mu}(\zeta)\right],$$
(3.1)

$$U'(-\frac{1}{2}\mu^{2},\mu t\sqrt{2}) = \frac{(2\pi)^{1/2}\mu^{2/3}g(\mu)}{\phi(\zeta)} \left[\frac{A_{i}(\mu^{4/3}\zeta)}{\mu^{4/3}}C_{\mu}(\zeta) + A_{i'}(\mu^{4/3}\zeta)D_{\mu}(\zeta)\right],$$
(3.2)

$$V(-\frac{1}{2}\mu^{2},\mu t\sqrt{2}) = \frac{2\pi^{1/2}\mu^{1/3}g(\mu)\phi(\zeta)}{\Gamma(1/2+(1/2)\mu^{2})} \left[B_{i}(\mu^{4/3}\zeta)A_{\mu}(\zeta) + \frac{B_{i'}(\mu^{4/3}\zeta)}{\mu^{8/3}}B_{\mu}(\zeta) \right],$$
(3.3)

$$V'(-\frac{1}{2}\mu^2,\mu t\sqrt{2}) = \frac{(2\pi)^{1/2}\mu^{2/3}g(\mu)}{\phi(\zeta)\Gamma(1/2+(1/2)\mu^2)} \left[\frac{B_i(\mu^{4/3}\zeta)}{\mu^{4/3}}C_\mu(\zeta) + B_{i'}(\mu^{4/3}\zeta)D_\mu(\zeta)\right].$$
(3.4)

The coefficient functions $A_{\mu}(\zeta), B_{\mu}(\zeta), C_{\mu}(\zeta)$ and $D_{\mu}(\zeta)$ have the following asymptotic expansions:

$$A_{\mu}(\zeta) \sim \sum_{s=0}^{\infty} \frac{a_s(\zeta)}{\mu^{4s}}, \quad B_{\mu}(\zeta) \sim \sum_{s=0}^{\infty} \frac{b_s(\zeta)}{\mu^{4s}},$$
 (3.5)

$$C_{\mu}(\zeta) \sim \sum_{s=0}^{\infty} \frac{c_s(\zeta)}{\mu^{4s}}, \quad D_{\mu}(\zeta) \sim \sum_{s=0}^{\infty} \frac{d_s(\zeta)}{\mu^{4s}},$$
 (3.6)

as $\mu \to \infty$, uniformly with respect to $t \ge -1 + \delta$, where δ is a small fixed positive number. The quantity ζ is defined by

$$\frac{2}{3}(-\zeta)^{3/2} = \eta(t), \quad -1 < t \le 1, \ (\zeta \le 0),$$

$$\frac{2}{3}\zeta^{3/2} = \xi(t), \quad 1 \le t, \ (\zeta \ge 0),$$

(3.7)

where η, ξ follow from (2.26), (2.2), respectively, and

$$\phi(\zeta) = \left(\frac{\zeta}{t^2 - 1}\right)^{1/4}.$$
(3.8)

The function $\zeta(t)$ is real for t > -1 and analytic at t = 1. We can invert $\zeta(t)$ into $t(\zeta)$, and obtain

$$t = 1 + 2^{-1/3}\zeta - \frac{1}{10}2^{-2/3}\zeta^2 + \frac{11}{700}\zeta^3 + \cdots$$

The function $g(\mu)$ has the expansion given in (2.6) and the coefficients $a_s(\zeta), b_s(\zeta)$ are given by

$$a_{s}(\zeta) = \sum_{m=0}^{2s} \beta_{m} \zeta^{-(3/2)m} \mathscr{A}_{2s-m}(t) \quad \sqrt{\zeta} b_{s}(\zeta) = -\sum_{m=0}^{2s+1} \alpha_{m} \zeta^{-(3/2)m} \mathscr{A}_{2s-m+1}(t),$$
(3.9)

where $\mathscr{A}_{s}(t)$ are used in (2.1), $\alpha_{0} = 1$ and

$$\alpha_m = \frac{(2m+1)(2m+3)\cdots(6m-1)}{m!(144)^m}, \quad \beta_m = -\frac{6m+1}{6m-1}a_m.$$
(3.10)

A recursion for α_m reads

$$\alpha_{m+1} = \alpha_m \frac{(6m+5)(6m+3)(6m+1)}{144(m+1)(2m+1)}, \quad m = 0, 1, 2, \dots$$

The numbers α_m , β_m occur in the asymptotic expansions of the Airy functions, and the relations in (3.9) follow from solving (3.1) and (3.3) for $A_{\mu}(\zeta)$ and $B_{\mu}(\zeta)$, expanding the Airy functions (assuming that ζ is bounded away from 0) and by using (2.1) and a similar result for V(a,z)(O-(11.16) and O-(2.12)).

231

For negative values of ζ (i.e., -1 < t < 1) we can use (O-(13.4))

$$a_{s}(\zeta) = (-1)^{s} \sum_{m=0}^{2s} \beta_{m}(-\zeta)^{-(3/2)m} \tilde{\mathscr{A}}_{2s-m}(t),$$

$$\sqrt{-\zeta} b_{s}(\zeta) = (-1)^{s-1} \sum_{m=0}^{2s+1} \alpha_{m}(-\zeta)^{-(3/2)m} \tilde{\mathscr{A}}_{2s-m+1}(t),$$
(3.11)

where

$$\tilde{\mathscr{A}}_{s}(t) = \frac{u_{s}(t)}{(1-t^{2})^{(3/2)s}}.$$

The functions $C_{\mu}(\zeta)$ and $D_{\mu}(\zeta)$ of (3.2) and (3.4) are given by

$$C_{\mu}(\zeta) = \chi(\zeta)A_{\mu}(\zeta) + A'_{\mu}(\zeta) + \zeta B_{\mu}(\zeta), \quad D_{\mu}(\zeta) = A_{\mu}(\zeta) + \frac{1}{\mu^{4}}[\chi(\zeta)B_{\mu}(\zeta) + B'_{\mu}(\zeta)].$$
(3.12)

The coefficients $c_s(\zeta)$ and $d_s(\zeta)$ in (3.6) are given by

$$c_{s}(\zeta) = \chi(\zeta)a_{s}(\zeta) + a'_{s}(\zeta) + \zeta b_{s}(\zeta), \quad d_{s}(\zeta) = a_{s}(\zeta) + \chi(\zeta)b_{s-1}(\zeta) + b'_{s-1}(\zeta), \quad (3.13)$$

where

$$\chi(\zeta) = \frac{\phi'(\zeta)}{\phi(\zeta)} = \frac{1 - 2t[\phi(\zeta)]^6}{4\zeta}$$
(3.14)

with $\phi(\zeta)$ given in (3.8). Explicitly,

$$\frac{1}{\sqrt{\zeta}}c_s(\zeta) = -\sum_{m=0}^{2s+1} \beta_m \zeta^{-(3/2)m} \mathscr{B}_{2s-m+1}(\tau) \quad d_s(\zeta) = -\sum_{m=0}^{2s} \alpha_m \zeta^{-(3/2)m} \mathscr{B}_{2s-m}(\tau), \tag{3.15}$$

where $\mathscr{B}_s(\tau) = v_s(t)/(t^2 - 1)^{(3/2)s}$, with $v_s(t)$ defined in (2.25). Other versions of (3.15) are needed for negative values of ζ , i.e., if -1 < t < 1; see (3.11).

3.2. The case $a \ll 0, z \leq 0$

Near the other turning point t = -1 we can use the representations (O-(9.7))

$$U(-\frac{1}{2}\mu^{2},-\mu t\sqrt{2}) = 2\pi^{1/2}\mu^{1/3}g(\mu)\phi(\zeta) \left[\sin(\frac{1}{2}\pi\mu^{2})\left\{A_{i}(\mu^{4/3}\zeta)A_{\mu}(\zeta) + \frac{A_{i'}(\mu^{4/3}\zeta)}{\mu^{8/3}}B_{\mu}(\zeta)\right\} + \cos(\frac{1}{2}\pi\mu^{2})\left\{B_{i}(\mu^{4/3}\zeta)\sum_{s=0}^{\infty}A_{\mu}(\zeta) + \frac{B_{i'}(\mu^{4/3}\zeta)}{\mu^{8/3}}B_{\mu}(\zeta)\right\}\right]$$
(3.16)

as $\mu \to \infty$, uniformly with respect to $t \ge -1 + \delta$, where δ is a small fixed positive number. Expansions for V(a,z) follow from (3.1) and (3.16) and the second relation in (1.7). Results for the derivatives of U(a,z) and V(a,z) follow easily from the earlier results.

3.3. Modified forms of Olver's Airy-type expansions

Modified versions of the Airy-type expansions (3.1)–(3.4) can also be given. In the case of the expansions in terms of elementary functions our main motivation for introducing modified expansions was the double asymptotic property of these expansions. In the case of the Airy-type expansions the interesting domains for the parameter t, from a numerical point of view, are finite domains that contain the turning points ± 1 . So, considering the expansions given so far, there is no need to have Airy-type expansions with the double asymptotic property; if μ remains finite and $|t| \gg 1$ we can use the expansions in terms of elementary functions. However, we have another interest in modified expansions in the case of Airy-type expansions. We explain this by first discussing a few properties of the coefficient functions $A_{\mu}(\zeta), B_{\mu}(\zeta), C_{\mu}(\zeta)$ and $D_{\mu}(\zeta)$.

By using the Wronskian relation (1.6) we can verify the relation

$$A_{\mu}(\zeta)D_{\mu}(\zeta) - \frac{1}{\mu^4}B_{\mu}(\zeta)C_{\mu}(\zeta) = \frac{\Gamma(1/2 + (1/2)\mu^2)}{2\mu\sqrt{\pi}g^2(\mu)},$$
(3.17)

where $g(\mu)$ is defined by means of an asymptotic expansion given in (2.6). By using the differential equation (O-(7.2))

$$\frac{\mathrm{d}^2 W}{\mathrm{d}\zeta^2} = \left[\mu^4 \zeta + \Psi(\zeta)\right] W,\tag{3.18}$$

where

$$\Psi(\zeta) = \frac{5}{16\zeta^2} - \frac{(3t^2 + 2)\zeta}{4(t^2 - 1)^3} = 2^{1/3} \left[-\frac{9}{280} + \frac{7}{150} 2^{-1/3}\zeta - \frac{1359}{26950} 2^{-2/3}\zeta^2 + \frac{196}{8125}\zeta^3 \dots \right], \quad (3.19)$$

we can derive the following system of equations for the functions $A_{\mu}(\zeta), B_{\mu}(\zeta)$:

$$A'' + 2\zeta B' + B - \Psi(\zeta)A = 0,$$

$$B'' + 2\mu^4 A' - \Psi(\zeta)B = 0,$$
(3.20)

where primes denote differentiation with respect to ζ . A Wronskian for this system follows by eliminating the terms with $\Psi(\zeta)$. This gives

$$2\mu^4 A'A + AB'' - A''B - 2\zeta B'B - B^2 = 0,$$

which can be integrated as

$$\mu^{4} A_{\mu}^{2}(\zeta) + A_{\mu}(\zeta) B_{\mu}'(\zeta) - A_{\mu}'(\zeta) B_{\mu}(\zeta) - \zeta B_{\mu}^{2}(\zeta) = \mu^{4} \frac{\Gamma(1/2 + (1/2)\mu^{2})}{2\mu\sqrt{\pi}g^{2}(\mu)},$$
(3.21)

where the quantity on the right-hand side follows from (3.17) and (3.12). It has the expansion

$$\mu^{4} \left[1 - \frac{1}{576\mu^{4}} + \frac{2021}{2\,488\,320\mu^{8}} + \cdots \right], \tag{3.22}$$

as follows from O-(2.22) and O-(5.21).

As mentioned before, the interesting domain of the Airy-type expansions given in this section is the domain that contains the turning point t = 1, or $\zeta = 0$. The representations of the coefficients of the expansions given in (3.9) cannot be used in numerical algorithms when $|\zeta|$ is small, unless we expand all relevant coefficients in powers of ζ . This is one way how to handle this problem numerically; see [18]. In that paper we have discussed another method that is based on a system like (3.20), with applications to Bessel functions. In that method the functions $A_{\mu}(\zeta)$ and $B_{\mu}(\zeta)$ are expanded in powers of ζ , for sufficiently small values of $|\zeta|$, say $|\zeta| \leq 1$, and the Maclaurin coefficients are computed from (3.20) by recursion. A normalizing relation (the analogue of (3.21)) plays a crucial role in that algorithm. The method works quite well for relatively small values of a parameter (the order of the Bessel functions) that is the analogue of μ .

When we want to use this algorithm for the present case only large values of μ are allowed because the function $g(\mu)$ that is used in (3.1)–(3.4) and (3.21) is only defined for large values of μ . For this reason we give the modified versions of Olver's Airy-type expansions. The modified versions are more complicated than the Olver's expansions, because the analogues of the series in (3.5) and (3.6) are in powers of μ^{-2} , and not in powers of μ^{-4} . Hence, when we use these series for numerical computations we need more coefficients in the modified expansions, which is certainly not desirable from a numerical point of view, given the complexity of the coefficients in Airy-type expansions. However, in the algorithm based on Maclaurin expansions of the analogues of the coefficient functions $A_{\mu}(\zeta), B_{\mu}(\zeta), C_{\mu}(\zeta)$ and $D_{\mu}(\zeta)$ this point is of minor concern.

The modified expansions are the following:

$$U(-\frac{1}{2}\mu^{2},\mu t\sqrt{2}) = \frac{\Gamma(1/2 + (1/2)\mu^{2})\phi(\zeta)}{\mu^{2/3}h(\mu)} \left[A_{i}(\mu^{4/3}\zeta)F_{\mu}(\zeta) + \frac{A_{i'}(\mu^{4/3}\zeta)}{\mu^{8/3}}G_{\mu}(\zeta) \right],$$
(3.23)

$$V(-\frac{1}{2}\mu^{2},\mu t\sqrt{2}) = \frac{\phi(\zeta)}{\mu^{2/3}h(\mu)} \left[B_{i}(\mu^{4/3}\zeta)F_{\mu}(\zeta) + \frac{B_{i'}(\mu^{4/3}\zeta)}{\mu^{8/3}}G_{\mu}(\zeta) \right].$$
(3.24)

The functions $F_{\mu}(\zeta)$ and $G_{\mu}(\zeta)$ have the following asymptotic expansions:

$$F_{\mu}(\zeta) \sim \sum_{s=0}^{\infty} \frac{f_s(\zeta)}{\mu^{2s}}, \quad G_{\mu}(\zeta) \sim \sum_{s=0}^{\infty} \frac{g_s(\zeta)}{\mu^{2s}}.$$
 (3.25)

The quantity ζ and the functions $\phi(\zeta)$ and $h(\mu)$ are as in Section 3.1. Comparing (3.23), (3.24) with (3.1), (3.3) we conclude that

$$F_{\mu}(\zeta) = H(\mu)A_{\mu}(\zeta), \quad G_{\mu}(\zeta) = H(\mu)B_{\mu}(\zeta), \quad H(\mu) = \frac{2\sqrt{\pi}\mu g(\mu)h(\mu)}{\Gamma(1/2 + (1/2)\mu^2)}.$$
(3.26)

The function $H(\mu)$ can be expanded (see O-(2.22), O-(2.27), O-(6.2) and (2.6))

$$H(\mu) \sim 1 + \frac{1}{2} \sum_{s=1}^{\infty} (-1)^s \frac{\gamma_s}{(\frac{1}{2}\mu^2)^s},$$
(3.27)

where γ_s are the coefficients in the gamma function expansions

$$\Gamma(\frac{1}{2}+z) \sim \sqrt{2\pi} e^{-z} z^z \sum_{s=0}^{\infty} \frac{\gamma_s}{z^s}, \quad \frac{1}{\Gamma(\frac{1}{2}+z)} \sim \frac{e^z z^{-z}}{\sqrt{2\pi}} \sum_{s=0}^{\infty} (-1)^s \frac{\gamma_s}{z^s}.$$
 (3.28)

The first few coefficients are

$$\gamma_0 = 1, \quad \gamma_1 = -\frac{1}{24}, \quad \gamma_2 = \frac{1}{1152}, \quad \gamma_3 = \frac{1003}{414720}.$$

The second expansion in (3.28) can be used in (3.26) to find relations between the coefficients $a_s(\zeta)$ and $b_s(\zeta)$ of (3.5) and of $f_s(\zeta)$ and $g_s(\zeta)$ of (3.25). That is

$$f_0(\zeta) = 1, \quad f_1(\zeta) = \frac{1}{24}, \quad f_2(\zeta) = a_1(\zeta) + \frac{1}{576}, \quad f_3(\zeta) = \frac{1}{24}a_1(\zeta) - \frac{1003}{103\,680},$$

$$g_0(\zeta) = b_0(\zeta), \quad g_1(\zeta) = \frac{1}{24}b_0(\zeta),$$

$$g_2(\zeta) = b_1(\zeta) + \frac{1}{576}b_0(\zeta), \quad g_3(\zeta) = \frac{1}{24}b_1(\zeta) - \frac{1003}{103\,680}b_0(\zeta).$$

The coefficients $f_s(\zeta)$, $g_s(\zeta)$ can also be expressed in terms of the coefficients $\phi_s(\tau)$ that are introduced in (2.9) by deriving the analogues of (3.9).

The system of equations (3.20) remains the same:

$$F'' + 2\zeta G' + G - \Psi(\zeta)F = 0,$$

$$G'' + 2\mu^4 F' - \Psi(\zeta)G = 0$$
(3.29)

and the Wronskian relation becomes

$$\mu^{4}F_{\mu}^{2}(\zeta) + F_{\mu}(\zeta)G_{\mu}'(\zeta) - F_{\mu}'(\zeta)G_{\mu}(\zeta) - \zeta G_{\mu}^{2}(\zeta) = \mu^{4}\frac{2\sqrt{\pi}\mu h^{2}(\mu)}{\Gamma(1/2 + (1/2)\mu^{2})}.$$
(3.30)

The right-hand side has the expansion (see (3.28) and (2.7)) $\mu^4 \sum_{s=0}^{\infty} (-1)^s \gamma_s / (\frac{1}{2}\mu^2)^s$. Observe that (3.30) is an exact relation, whereas (3.21) contains the function $g(\mu)$, of which only an asymptotic expansion is available.

3.4. Numerical aspects of the Airy-type expansions

In [18, Section 4], we solved the system (3.29) (for the case of Bessel functions) by substituting Maclaurin series of $F(\zeta), G(\zeta)$ and $\Psi(\zeta)$. That is, we wrote

$$F(\zeta) = \sum_{n=0}^{\infty} c_n(\mu)\zeta^n, \quad G(\zeta) = \sum_{n=0}^{\infty} d_n(\mu)\zeta^n, \quad \Psi(\zeta) = \sum_{n=0}^{\infty} \psi_n\zeta^n,$$

where the coefficients ψ_n can be considered as known (see (3.19)), and we substituted the expansions in (3.29). This gives for n = 0, 1, 2, ..., the recursion relations

$$(n+2)(n+1)c_{n+2} + (2n+1)d_n = \rho_n, \quad \rho_n = \sum_{k=0}^n \psi_k c_{n-k},$$

$$(n+2)(n+1)d_{n+2} + 2\mu^4(n+1)c_{n+1} = \sigma_n, \quad \sigma_n = \sum_{k=0}^n \psi_k d_{n-k}.$$
(3.31)

If μ is large, the recursion relations cannot be solved in forward direction, because of numerical instabilities. For the Bessel function case we have shown that we can solve the system by iteration and backward recursion. The relation in (3.30) can be used for normalization of the coefficients in the backward recursion scheme.

For details we refer to [18]. The present case is identical to the case of the Bessel functions; only the function $\Psi(\zeta)$ is different, and instead of μ^2 in (3.31) we had the order v of the Bessel functions.

4. Expansions from integral representations

The expansions developed by Olver, of which some are given in the previous sections, are all valid if |a| is large. For several cases we gave modified expansions that hold if at least one of the two parameters a, z is large and we have indicated the relations between Olver's expansions and the new expansions. The modified expansions have in fact a double asymptotic property. Initially, we derived these expansions by using integral representations of the parabolic cylinder functions, and later we found the relations with Olver's expansions. In this section we explain how some of the modified expansions can be obtained from the integrals that define U(a,z) and V(a,z). Again we only consider real values of the parameters.

4.1. Expansions in terms of elementary functions by using integrals

4.1.1. The case $a \ge 0, z \ge 0; a + z \gg 0$

We start with the well-known integral representation

$$U(a,z) = \frac{e^{-(1/4)z^2}}{\Gamma(a+\frac{1}{2})} \int_0^\infty w^{a-(1/2)} e^{-(1/2)w^2 - zw} \, \mathrm{d}w, \quad a > -\frac{1}{2}$$
(4.1)

which we write in the form

$$U(a,z) = \frac{z^{a+(1/2)} e^{-(1/4)z^2}}{\Gamma(a+(1/2))} \int_0^\infty w^{-1/2} e^{-z^2 \phi(w)} dw,$$
(4.2)

where

$$\phi(w) = w + \frac{1}{2}w^2 - \lambda \ln w, \quad \lambda = \frac{a}{z^2}.$$
 (4.3)

The positive saddle point w_0 of the integrand in (4.3) is computed from

$$\frac{\mathrm{d}\phi(w)}{\mathrm{d}w} = \frac{w^2 + w - \lambda}{w} = 0,\tag{4.4}$$

giving

$$w_0 = \frac{1}{2} [\sqrt{1 + 4\lambda} - 1]. \tag{4.5}$$

We consider z as the large parameter. When λ is bounded away from 0 we can use Laplace's method (see [11] or [22]). When a and z are such that $\lambda \to 0$ Laplace's method cannot be applied. However, we can use a method given in [15] that allows small values of λ .

To obtain a standard form for this Laplace-type integral, we transform $w \to t$ (see [16]) by writing

$$\phi(w) = t - \lambda \ln t + A, \tag{4.6}$$

where A does not depend on t or w, and we prescribe that w = 0 should correspond with t = 0 and $w = w_0$ with $t = \lambda$, the saddle point in the t-plane.

This gives

$$U(a,z) = \frac{z^{a+(1/2)} e^{-(1/4)z^2 - Az^2}}{(1+4\lambda)^{1/4} \Gamma(a+1/2)} \int_0^\infty t^{a-1/2} e^{-z^2 t} f(t) dt,$$
(4.7)

where

$$f(t) = (1+4\lambda)^{1/4} \sqrt{\frac{t}{w}} \frac{\mathrm{d}w}{\mathrm{d}t} = (1+4\lambda)^{1/4} \sqrt{\frac{w}{t}} \frac{t-\lambda}{w^2+w-\lambda}.$$
(4.8)

By normalizing with the quantity $(1 + 4\lambda)^{1/4}$ we obtain $f(\lambda) = 1$, as can be verified from (4.8) and a limiting process (using l'Hôpital's rule). The quantity A is given by

$$A = \frac{1}{2}w_0^2 + w_0 - \lambda \ln w_0 - \lambda + \lambda \ln \lambda.$$
(4.9)

A first uniform expansion can be obtained by writing

$$f(t) = \sum_{n=0}^{\infty} a_n(\lambda)(t-\lambda)^n.$$
(4.10)

Details on the computation of $a_n(\lambda)$ will be given in the appendix.

By substituting (4.10) into (4.7) we obtain

$$U(a,z) \sim \frac{e^{-(1/4)z^2 - Az^2}}{z^{a+(1/2)}(1+4\lambda)^{(1/4)}} \sum_{n=0}^{\infty} a_n(\lambda) P_n(a) z^{-2n},$$
(4.11)

where

$$P_n(a) = \frac{z^{2a+2n+1}}{\Gamma(a+1/2)} \int_0^\infty t^{a-1/2} e^{-z^2 t} (t-\lambda)^n \, \mathrm{d}t, \quad n = 0, 1, 2, \dots$$
 (4.12)

The $P_n(a)$ are polynomials in a. They follow the recursion relation

$$P_{n+1}(a) = (n + \frac{1}{2})P_n(a) + anP_{n-1}(a), \quad n = 0, 1, 2, \dots$$

with initial values

$$P_0(a) = 1, \quad P_1(a) = \frac{1}{2}.$$

We can obtain a second expansion

$$U(a,z) \sim \frac{e^{-(1/4)z^2 - Az^2}}{z^{a+(1/2)}(1+4\lambda)^{1/4}} \sum_{k=0}^{\infty} \frac{f_k(\lambda)}{z^{2k}}$$
(4.13)

with the property that in the series the parameters λ and z are separated, by introducing a sequence of functions $\{f_k\}$ with $f_0(t) = f(t)$ and by defining

$$f_{k+1}(t) = \sqrt{t} \frac{\mathrm{d}}{\mathrm{d}t} \left[\sqrt{t} \frac{f_k(t) - f_k(\lambda)}{t - \lambda} \right], \quad k = 0, 1, 2, \dots$$

$$(4.14)$$

The coefficients $f_k(\lambda)$ can be expressed in terms of the coefficients $a_n(\lambda)$ defined in (4.10). To verify this, we write

$$f_k(t) = \sum_{n=0}^{\infty} a_n^{(k)}(\lambda)(t-\lambda)^n$$
(4.15)

and by substituting this in (4.14) it follows that

$$a_n^{(k+1)}(\lambda) = \lambda(n+1)a_{n+2}^{(k)}(\lambda) + (n+\frac{1}{2})a_{n+1}^{(k)}(\lambda), \quad k \ge 0, \ n \ge 0.$$
(4.16)

Hence, the coefficients $f_k(\lambda)$ of (4.13) are given by

$$f_k(\lambda) = a_0^{(k)}(\lambda), \quad k \ge 0.$$
(4.17)

We have

$$f_{0}(\lambda) = 1,$$

$$f_{1}(\lambda) = \frac{1}{2}[a_{1}(\lambda) + 2\lambda a_{2}(\lambda)],$$

$$f_{2}(\lambda) = \frac{1}{4}[12\lambda^{2}a_{4}(\lambda) + 14\lambda a_{3}(\lambda) + 3a_{2}(\lambda)],$$

$$f_{3}(\lambda) = \frac{1}{8}[120\lambda^{3}a_{6}(\lambda) + 220\lambda^{2}a_{5}(\lambda) + 116\lambda a_{4}(\lambda) + 15a_{3}(\lambda)].$$
(4.18)

Explicitly,

$$f_{0}(\lambda) = 1,$$

$$f_{1}(\lambda) = -\frac{\rho}{24}(20\sigma^{2} - 10\sigma - 1),$$

$$f_{2}(\lambda) = \frac{\rho^{2}}{1152}(6160\sigma^{4} - 6160\sigma^{3} + 924\sigma^{2} + 20\sigma + 1),$$

$$f_{3}(\lambda) = -\frac{\rho^{3}}{414720}(27227200\sigma^{6} - 40840800\sigma^{5} + 16336320\sigma^{4} - 1315160\sigma^{3} - 8112\sigma^{2} + 2874\sigma + 1003),$$
(4.19)

where

$$\sigma = \frac{1}{2} \left[1 + \frac{z}{\sqrt{4a + z^2}} \right], \quad \rho = \frac{(2\sigma - 1)^2}{\sigma} = \frac{2z^2}{\sqrt{4a + z^2}(z + \sqrt{4a + z^2})}.$$
(4.20)

We observe that $f_k(\lambda)$ is a polynomial of degree 2k in σ multiplied with ρ^k .

If a and z are positive then $\sigma \in [0,1]$. Furthermore, the sequence $\{\rho^k/z^{2k}\}$ is an asymptotic scale when one or both parameters a and z are large. The expansion in (4.13) is valid for $z \to \infty$ and holds uniformly for $a \ge 0$. It has a double asymptotic property in the sense that it is also valid as $a \to \infty$, uniformly with respect to $z \ge 0$. As follows from the coefficients given in (4.19) and relations to be given later, we can indeed let $z \to 0$ in the expansion.

The expansion in (4.13) can be obtained by using an integration by parts procedure. We give a few steps in this method. Consider the integral

$$F_a(z) = \frac{1}{\Gamma(a+1/2)} \int_0^\infty t^{a-(1/2)} \mathrm{e}^{-z^2 t} f(t) \,\mathrm{d}t.$$
(4.21)

We have (with $\lambda = a/z^2$)

$$F_{a}(z) = \frac{f(\lambda)}{\Gamma(a+(1/2))} \int_{0}^{\infty} t^{a-(1/2)} e^{-z^{2}t} dt + \frac{1}{\Gamma(a+(1/2))} \int_{0}^{\infty} t^{a-(1/2)} e^{-z^{2}t} [f(t) - f(\lambda)] dt$$

$$= z^{-2a-1} f(\lambda) - \frac{1}{z^{2} \Gamma(a+(1/2))} \int_{0}^{\infty} t^{(1/2)} \frac{[f(t) - f(\lambda)]}{t - \lambda} de^{-z^{2}(t - \lambda \ln t)}$$

$$= z^{-2a-1} f(\lambda) + \frac{1}{z^{2} \Gamma(a+(1/2))} \int_{0}^{\infty} t^{a-(1/2)} e^{-z^{2}t} f_{1}(t) dt,$$

238

where f_1 is given in (4.14) with $f_0 = f$. Repeating this procedure we obtain (4.13). More details on this method and proofs of the asymptotic nature of the expansions (4.11) and (4.13) can be found in our earlier papers. We concentrate on expansion (4.13) because (4.11) cannot be compared with Olver's expansions.

To compare (4.13) with Olver's expansion (2.16), we write

$$a = \frac{1}{2}\mu^2, \quad z = \mu\sqrt{2}t.$$
 (4.22)

Then the parameters σ and ρ defined in (4.20) become

$$\sigma = \frac{1}{2} \left[1 + \frac{t}{\sqrt{1+t^2}} \right] = \tilde{\tau} + 1, \quad \rho = \frac{2t^2}{\sqrt{1+t^2}(t+\sqrt{1+t^2})}, \tag{4.23}$$

where $\tilde{\tau}$ is given in (2.32). After a few manipulations we write (4.13) in the form (cf. (2.29))

$$U(\frac{1}{2}\mu^{2},\mu t\sqrt{2}) = \frac{\tilde{h}(\mu)e^{-\mu^{2}\tilde{\xi}}}{(t^{2}+1)^{1/4}}\tilde{F}_{\mu}(z), \quad \tilde{F}_{\mu}(z) \sim \sum_{k=0}^{\infty} (-1)^{k}\frac{\tilde{\phi}_{k}(\sigma)}{\mu^{2k}},$$
(4.24)

where

$$\tilde{\xi} = \frac{1}{2} [t\sqrt{1+t^2} + \ln(t+\sqrt{1+t^2})], \tag{4.25}$$

$$h(\mu) = e^{(1/4)\mu^2} \mu^{-(1/2)\mu^2 - (1/2)} 2^{(1/4)\mu^2 - (1/4)}$$
(4.26)

and

$$\tilde{\phi}_k(\sigma) = \frac{(-1)^k}{(2t^2)^k} f_k(\lambda). \tag{4.27}$$

Explicitly,

$$\begin{split} \tilde{\phi}_{0}(\sigma) &= 1, \\ \tilde{\phi}_{1}(\sigma) &= \frac{1-\sigma}{12} (20\sigma^{2} - 10\sigma - 1), \\ \tilde{\phi}_{2}(\sigma) &= \frac{(1-\sigma)^{2}}{288} (6160\sigma^{4} - 6160\sigma^{3} + 924\sigma^{2} + 20\sigma + 1), \\ \tilde{\phi}_{3}(\sigma) &= \frac{(1-\sigma)^{3}}{51840} (27\,227\,200\sigma^{6} - 40\,840\,800\sigma^{5} + 16\,336\,320\sigma^{4} \\ &- 1\,315\,160\sigma^{3} - 8112\sigma^{2} + 2874\sigma + 1003), \end{split}$$
(4.28)

where σ is given in (4.23). Comparing (4.24) with (2.29) we obtain $\tilde{\phi}_k(\sigma) = \phi_k(\tilde{\tau}), \ k \ge 0$, because $\sigma = 1 + \tilde{\tau}$.

4.1.2. The case $a \ge 0$, $z \le 0$; $a - z \gg 0$

To derive the first expansion in (2.34) we use the contour integral

$$U(a,-z) = \frac{\sqrt{2\pi}e^{(1/4)z^2}}{\Gamma(a+(1/2))}H_a(z), \quad H_a(z) = \frac{\Gamma(a+(1/2))}{2\pi i}\int_{\mathscr{C}} e^{zs+(1/2)s^2}s^{-a-(1/2)}\,\mathrm{d}s, \tag{4.29}$$

where \mathscr{C} is a vertical line in the half-plane $\Re s > 0$. This integral can be transformed into a standard form that involves the same mapping as in the previous subsection. We first write (by transforming via s = zw)

$$H_{a}(z) = \frac{z^{(1/2)-a}\Gamma(a+(1/2))}{2\pi i} \int_{\mathscr{C}} e^{z^{2}(w+(1/2)w^{2})} w^{-a-(1/2)} dw$$

$$= \frac{z^{(1/2)-a}\Gamma(a+(1/2))}{2\pi i} \int_{\mathscr{C}} e^{z^{2}\phi(w)} \frac{dw}{\sqrt{w}},$$
(4.30)

where $\phi(w)$ is defined in (4.3). By using the transformation given in (4.6) it follows that

$$H_a(z) = \frac{z^{(1/2)-a}\Gamma(a+(1/2))e^{Az^2}}{2\pi i} \int_{\mathscr{C}} e^{z^2 t} t^{-a-(1/2)} f(t) dt.$$
(4.31)

The integration by parts method used for (4.21) gives the expansion (see [18])

$$H_a(z) \sim \frac{z^a e^{Az^2}}{(4a+z^2)^{1/4}} \sum_{k=0}^{\infty} (-1)^k \frac{f_k(\lambda)}{z^{2k}},$$
(4.32)

where the $f_k(\lambda)$ are the same as in (4.13); see also (4.18). This gives the first expansion of (2.34).

Remark 2.5. The first result in (2.34) can also be obtained by using (4.1) with z < 0. The integral for U(a, -z) can be written as in (4.2), now with $\phi(w) = \frac{1}{2}w^2 - w - \ln \lambda$, $\lambda = a/z^2$. In this case the relevant saddle point at $w_0 = (1 + \sqrt{1 + 4\lambda})/2$ is always inside the interval $[1, \infty)$ and the standard method of Laplace can be used. The same expansion will be obtained with the same structure and coefficients as in (2.34), because of the unicity of Poincaré-type asymptotic expansions. See also Section 4.1.4 where Laplace's method will be used for an integral that defines V(a,z).

4.1.3. The case $a \leq 0, z > 2\sqrt{-a}, -a + z \gg 0$

Olver's starting point (2.1) can also be obtained from an integral. Observe that (4.1) is not valid for $a \le -\frac{1}{2}$. We take as integral (see [1, p. 687, 19.5.1])

$$U(-a,z) = \frac{\Gamma(1/2+a)}{2\pi i} e^{-(1/4)z^2} \int_{\alpha} e^{zs - (1/2)s^2} s^{-a - (1/2)} \,\mathrm{d}s, \tag{4.33}$$

where α is a contour that encircles the negative s-axis in positive direction. Using a transformation we can write this in the form (cf. (4.2))

$$U(-a,z) = \frac{\Gamma(1/2+a)}{2\pi i} z^{(1/2)-a} e^{-(1/4)z^2} \int_{\alpha} e^{\phi(w)} w^{-1/2} \, \mathrm{d}w, \tag{4.34}$$

where

$$\phi(w) = w - \frac{1}{2}w^2 - \lambda \ln w, \quad \lambda = \frac{a}{z^2}.$$
 (4.35)

The relevant saddle point is now given by

$$w_0 = \frac{1}{2} [1 - \sqrt{1 - 4\lambda}], \quad 0 < \lambda < \frac{1}{4}.$$
(4.36)

When $\lambda \to 0$ the standard saddle point method is not applicable, and we can again use the methods of our earlier papers [15,16] and transform

$$\phi(w) = t - \lambda \ln t + A, \tag{4.37}$$

where the points at $-\infty$ in the w- and t-plane should correspond, and $w = w_0$ with $t = \lambda$. We obtain

$$U(-a,z) = \frac{\Gamma((1/2) + a)}{(1 - 4\lambda)^{1/4} 2\pi i} z^{(1/2) - a} e^{-(1/4)z^2 + z^2 A} \int_{\alpha} e^{z^2 t} t^{-a - (1/2)} f(t) dt,$$
(4.38)

where α is a contour that encircles the negative *t*-axis in positive direction and

$$f(t) = (1 - 4\lambda)^{1/4} \sqrt{\frac{t}{w}} \frac{\mathrm{d}w}{\mathrm{d}t} = (1 - 4\lambda)^{1/4} \sqrt{\frac{w}{t}} \frac{t - \lambda}{w - w^2 - \lambda}.$$
(4.39)

Expanding f(t) as in (4.10), and computing $f_k(\lambda)$ as in the procedure that yields the relations in (4.18), we find that the same values $f_k(\lambda)$ as in (4.19), up to a factor $(-1)^k$ and a different value of τ and ρ . By using the integration by parts method for contour integrals [15], that earlier produced (4.32), we obtain the result

$$U(-a,z) \sim \frac{z^a e^{Az^2 - (1/4)z^2}}{(z^2 - 4a)^{1/4}} \sum_{k=0}^{\infty} (-1)^k \frac{f_k(\lambda)}{z^{2k}},$$
(4.40)

where the first $f_k(\lambda)$ are given in (4.19) with

$$\sigma = \frac{1}{2} \left[1 + \frac{z}{\sqrt{z^2 - 4a}} \right], \quad \rho = \frac{(2\sigma - 1)^2}{\sigma} = \frac{2z^2}{\sqrt{z^2 - 4a + (z + \sqrt{z^2 - 4a})}}.$$
(4.41)

This expansion can be written in the form (2.9).

4.1.4. The case $a \leq 0$, $z < -2\sqrt{-a}$, $-a - z \gg 0$ We use the relation (see (1.7))

$$U(-a, -z) = \sin \pi a \, U(-a, z) + \frac{\pi}{\Gamma((1/2) - a)} \, V(-a, z), \tag{4.42}$$

and use the result of U(-a,z) given in (4.40) or the form (2.9). An expansion for V(-a,z) in (4.42) can be obtained from the integral (see [9])

$$V(a,z) = \frac{e^{-(1/4)z^2}}{2\pi} \int_{\gamma_1 \cup \gamma_2} e^{-(1/2)s^2 + zs} s^{a-(1/2)} \, \mathrm{d}s, \tag{4.43}$$

where γ_1 and γ_2 are two horizontal lines, γ_1 in the upper half plane $\Im s > 0$ and γ_2 in the lower half plane $\Im s < 0$; the integration is from $\Re s = -\infty$ to $\Re s = +\infty$. (Observe that when we integrate on γ_1 in the other direction (from $\Re s = +\infty$ to $\Re s = \infty$) the contour $\gamma_1 \cup \gamma_2$ can be deformed into α of (4.33), and the integral defines U(a,z), up to a factor.) We can apply Laplace's method to obtain the expansion given in (2.14) (see Remark 4.1).

4.2. The singular points of the mapping (4.6)

The mapping defined in (4.6) is singular at the saddle point

$$w_{-} = -\frac{1}{2}(\sqrt{1+4\lambda}+1). \tag{4.44}$$

If $\lambda = 0$ then $w_{-} = -1$ and the corresponding *t*-value is $-\frac{1}{2}$. For large values of λ we have the estimate:

$$t(w_{-}) \sim \lambda \left[-0.2785 - \frac{0.4356}{\sqrt{\lambda}} \right].$$

$$(4.45)$$

This estimate is obtained as follows. The value $t_{-} = t(w_{-})$ is implicitly defined by Eq. (4.6) with $w = w_{-}$. This gives

$$t_{-} - \lambda \ln t_{-} - \lambda + \lambda \ln \lambda = -\frac{1}{2}\sqrt{1+4\lambda} \pm \lambda \pi i + \lambda \ln \frac{4\lambda}{(1+\sqrt{1+4\lambda})^2}$$
$$= \pm \lambda \pi i - 2\sqrt{\lambda} \left[1 + \frac{1}{24\lambda} + \mathcal{O}(\lambda^{-2})\right], \qquad (4.46)$$

as $\lambda \to \infty$. The numerical solution of the equation $s - \ln s - 1 = \pm \pi i$ is given by $s_{\pm} = 0.2785 \cdots e^{\mp \pi i}$. This gives the leading term in (4.16). The other term follows by a further simple step.

4.3. Expansions in terms of Airy functions

All results for the modified Airy-type expansions given in Section 3.3 can be obtained by using certain loop integrals. The integrals in (4.33) and (4.43) can be used for obtaining (3.23) and (3.24), respectively. The method is based on replacing $\phi(w)$ in (4.34) by a cubic polynomial, in order to take into account the influence of both saddle points of $\phi(w)$. This method is first described in [6]; see also [11,22].

5. Numerical verifications

We verify several asymptotic expansions by computing the error in the Wronskian relation for the series in the asymptotic expansions. Consider Olver's expansions of Section 2.3 for the oscillatory region -1 < t < 1 with negative *a*. We verify the relation in (2.28). Denote the left-hand side of the first line in (2.28) by $W(\mu, t)$. Then we define as the error in the expansions

$$\Delta(\mu, t) := \left| \frac{W(\mu, t)}{1 - (1/576\mu^4) + (2021/2\,488\,320\mu^8)} - 1 \right|.$$
(5.1)

Taking three terms in the series of (2.23), (2.24) and (2.27), we obtain for several values of μ and t the results given in Table 1. We clearly see the loss of accuracy when t is close to 1. Exactly the same results are obtained for negative values of t in this interval.

Next, we consider the modified expansions of Section 2.1. Denote the left-hand side of (2.20) by $W(\mu, t)$. Then we define as the error in the expansions

$$\Delta(\mu, t) := \left| \frac{1}{2} W(\mu, t) - 1 \right|.$$
(5.2)

When we use the series in (2.9), (2.14), (2.18) and (2.19) with five terms, we obtain the results given in Table 2. We observe that the accuracy improves as μ or t increase. This shows the double asymptotic poperty of the modified expansions of Section 2.1.

Finally we consider the expansions of Sections 2.4 and 2.5. Let the left-hand side of (2.35) be denoted by $W(\mu, t)$. Then we define as the error in the expansions

$$\Delta(\mu, t) := \left| \frac{1}{2} W(\mu, t) - 1 \right|.$$
(5.3)

When we use the series in (2.29), (2.33) and (2.34) with five terms, we obtain the results of Table 3. We again observe that the accuracy improves as μ or *t* increase. This shows the double asymptotic property of the modified expansions of Sections 2.4 and 2.5.

μ	5	10	25	50	100
t					
0.00	0.32e - 09	0.78e - 13	0.13e - 17	0.32e - 21	0.78e - 25
0.10	0.26e - 09	0.63e - 13	0.11e - 17	0.26e - 21	0.63e - 25
0.20	0.81e - 10	0.20e - 13	0.33e - 18	0.82e - 22	0.20e - 25
0.30	0.16e - 08	0.39e - 12	0.65e - 17	0.16e - 20	0.39e - 24
0.40	0.88e - 08	0.22e - 11	0.36e - 16	0.89e - 20	0.22e - 23
0.50	0.51e - 07	0.13e - 10	0.21e - 15	0.52e - 19	0.13e - 22
0.60	0.40e - 06	0.99e - 10	0.17e - 14	0.40e - 18	0.99e - 22
0.70	0.53e - 05	0.13e - 08	0.22e - 13	0.54e - 17	0.13e - 20
0.80	0.20e - 03	0.50e - 07	0.84e - 12	0.20e - 15	0.50e - 19
0.90	0.35e - 00	0.24e - 04	0.41e - 09	0.10e - 12	0.25e - 16

Table 1 Relative accuracy $\Delta(\mu, t)$ defined in (5.1) for the asymptotic series of Section 2.3

Table 2											
Relative accuracy	$\varDelta(\mu,t)$	defined	in	(5.2)	for	the	asymptotic	series	of	Section	2.1

50	100
00	100
0 0.18e - 13	0.43e - 17
2 0.32e - 16	0.78e - 20
0.78e - 18	0.19e - 21
5 0.55e - 19	0.13e - 22
0.70e - 20	0.17e - 23
9 0.10e - 22	0.25e - 26
1 0.21e - 24	0.52e - 28
5 0.12e - 28	0.28e - 32
9 0.20e - 32	0.48e - 36
0.30e - 37	0.73e - 41
	$\begin{array}{c} & & & & \\ 0 & & & & 0.18e - 13 \\ 2 & & & & 0.32e - 16 \\ 4 & & & & 0.78e - 18 \\ 5 & & & & 0.55e - 19 \\ 5 & & & & 0.70e - 20 \\ 0 & & & & 0.10e - 22 \\ 1 & & & & 0.21e - 24 \\ 5 & & & & 0.12e - 28 \\ 0 & & & & 0.20e - 32 \\ 3 & & & & 0.30e - 37 \end{array}$

Table 3 Relative accuracy $\Delta(\mu, t)$ defined in (5.3) for the asymptotic series of Sections 2.4 and 2.5

μ t	5	10	25	50	100
0.00	0.32e - 09	0.78e - 13	0.13e - 17	0.32e - 21	0.78e - 25
0.25	0.12e - 09	0.28e - 13	0.47e - 18	0.12e - 21	0.28e - 25
0.50	0.45e - 11	0.11e - 14	0.19e - 19	0.46e - 23	0.11e - 26
0.75	0.57e - 11	0.14e - 14	0.24e - 19	0.58e - 23	0.14e - 26
1.0	0.27e - 11	0.65e - 15	0.11e - 19	0.27e - 23	0.65e - 27
1.5	0.29e - 13	0.70e - 17	0.12e - 21	0.29e - 25	0.70e - 29
2.0	0.20e - 13	0.48e - 17	0.81e - 22	0.20e - 25	0.48e - 29
2.5	0.43e - 14	0.11e - 17	0.18e - 22	0.43e - 26	0.11e - 29
5.0	0.45e - 17	0.11e - 20	0.18e - 25	0.45e - 29	0.11e - 32
10.0	0.16e - 20	0.38e - 24	0.64e - 29	0.16e - 32	0.38e - 36

6. Concluding remarks

As mentioned in Section 1.1, several sources for numerical algorithms for evaluating parabolic cylinder functions are available in the literature, but not so many algorithms make use of asymptotic expansions. The paper [10] is a rich source for asymptotic expansions, for all combinations of real and complex parameters, where always |a| has to be large. There are no published algorithms that make use of Olver's expansions, although very efficient algorithms can be designed by using the variety of these expansions; [3] is the only reference we found in which Olver's expansions are used for numerical computations.

We started our efforts in making algorithms for the case of real parameters. We selected appropriate expansions from Olver's paper and for some cases we modified Olver's expansions in order to get expansions having a double asymptotic property. A serious point is making efficient use of the powerful Airy-type expansions that are valid near the turning points of the differential equation (and in much larger intervals and domains of the complex plane). In particular, constructing reliable software for all possible combinations of the complex parameters a and z is a challenging problem.

A point of research interest is also the construction of error bounds for Olver's expansions and the modified expansions. Olver's paper is written before he developed the construction of bounds for the remainders, which he based on methods for differential equations, and which are available now in his book [11].

Appendix. Computing the coefficients $f_k(\lambda)$ of (4.13)

We give the details on the computation of the coefficients $f_k(\lambda)$ that are used in (4.13). The first step is to obtain coefficients d_k in the expansion

$$w = d_0 + d_1(t - \lambda) + d_2(t - \lambda)^2 + \cdots,$$
(A.1)

where $d_0 = w_0$. From (4.6) we obtain

$$\frac{\mathrm{d}w}{\mathrm{d}t} = \frac{w}{t} \frac{t-\lambda}{w^2 + w - \lambda}.\tag{A.2}$$

Substituting (A.1) we obtain

$$d_1^2 = \frac{w_0}{\lambda(1+2w_0)},\tag{A.3}$$

where the saddle point w_0 is defined in (4.5). From the conditions on the mapping (4.6) it follows that $d_1 > 0$. Higher order coefficients d_k can be obtained from the first ones by recursion.

When we have determined the coefficients in (A.1) we can use (4.8) to obtain the coefficients $a_n(\lambda)$ of (4.10).

For computing in this way a set of coefficients $f_k(\lambda)$, say $f_0(\lambda), \ldots, f_{15}(\lambda)$, we need more than 35 coefficients d_k in (A.1). Just taking the square root in (A.3) gives for higher coefficients d_k very complicated expressions, and even by using computer algebra programs, as Maple, we need suitable methods in computing the coefficients.

$$4\lambda = \tan^2 \theta. \tag{A.4}$$

We also write

$$\sigma = \cos^2 \frac{1}{2}\theta,\tag{A.5}$$

which is introduced earlier in (4.20) and (4.23). Then

$$w_0 = \frac{1-\sigma}{2\sigma-1}, \quad \lambda = \frac{\sigma(1-\sigma)}{(2\sigma-1)^2}, \quad d_1 = \frac{2\sigma-1}{\sigma}.$$
 (A.6)

In particular the expressions for w_0 and d_1 are quite convenient, because we can proceed without square roots in the computations. Higher coefficients d_k can be obtained by using (A.2).

The first relation $f_0(\lambda) = a_0^{(0)}(\lambda) = 1$ easily follows from (4.3), (4.8), (A.7) and (A.6):

$$f_0(\lambda) = (1+4\lambda)^{1/4} \sqrt{\frac{\lambda}{w_0}} d_1 = 1.$$

Then using (4.8) we obtain

$$a_0(\lambda) = 1, \quad a_1(\lambda) = -\frac{\cos^2\theta(1+2c)^2}{6(c+1)c^2}, \quad a_2(\lambda) = \frac{\cos^4\theta(20c^4 + 40c^3 + 30c^2 + 12c + 3)}{24(c+1)^2c^4}$$

where $c = \sqrt{\sigma} = \cos \frac{1}{2}\theta$. Using the scheme leading to (4.17) one obtains the coefficients $f_k(\lambda)$. The first few coefficients are given in (4.19).

We observe that $f_k(\lambda)$ is a polynomial of degree 2k in σ multiplied with ρ^k . If a and z are positive then $\sigma \in [0, 1]$. It follows that the sequence $\{\rho^k/z^{2k}\}$ is an asymptotic scale when one or both parameters a and z are large, and, hence, that $\{f_k(\lambda)/z^{2k}\}$ of (4.13) is an asymptotic scale when one or both parameters a and z are large.

Because of the relation in (4.27) and $\tilde{\phi}_k(\sigma) = \phi_k(\tilde{\tau})$, higher coefficients $f_k(\lambda)$ can also be obtained from the recursion relation (2.11), which is obtained by using the differential equation of the parabolic cylinder functions.

References

- M. Abramowitz, I.A. Stegun, Handbook of Mathematical Functions with Formulas, Graphs and Mathematical Tables, National Bureau of Standards Appl. Series, Vol. 55, U.S. Government Printing Office, Washington, DC, 1964 (paperback edition published by Dover, New York).
- [2] L. Baker, C Mathematical Function Handbook, McGraw-Hill, New York, 1992. Diskette with software included.
- [3] J.L. Blanchard, E.H. Newman, Numerical evaluation of parabolic cylinder functions, IEEE Trans. Antennas and Propagation 37 (1989) 519–523.
- [4] H. Buchholz, The Confluent Hypergeometric Function, Springer, Berlin, 1969.
- [5] B.W. Char, K.O. Geddes, G.H. Gonnet, B.L. Leong, M.B. Monagan, S.M. Watt, Maple V, Library Reference Manual, Corr. 2nd print, Springer, New York, 1991.
- [6] C. Chester, B. Friedman, F. Ursell, An extension of the method of steepest descent, Proc. Cambridge Philos. Soc. 53 (1957) 599–611.
- [7] D.W. Lozier, F.W.J. Olver, Numerical evaluation of special functions, in: W. Gautschi (Ed.), Mathematics of Computation 1943–1993: A Half-Century of Computational Mathematics, PSAM, AMS, Providence, 1994.

- [8] J.C.P. Miller, On the choice of standard solutions to Weber's equation, Proc. Cambridge Philos. Soc. 48 (1952) 428–435.
- [9] J.C.P. Miller, Tables of Weber Parabolic Cylinder Functions, H.M. Stationary Office, London, 1955.
- [10] F.W.J. Olver, Uniform asymptotic expansions for Weber parabolic cylinder functions of large order, J. Res. NBS 63B (1959) 131–169.
- [11] F.W.J. Olver, Asymptotics and Special Functions, Academic Press, New York, 1974 & 1997. Reprinted in 1997 by A.K. Peters.
- [12] W.H. Press, S.A. Teukolsky, W.T. Vetterling, B.P. Flannery, Numerical recipes. The art of scientific computing, 2nd Edition, Cambridge University Press, Cambridge, 1992, diskettes and example books available. Editions exist in Basic (1991), *C* (1992), Fortran (1992), Macintosh Fortran (1988) and Pascal (1989).
- [13] J. Segura, A. Gil, Parabolic cylinder functions of integer and half-integer orders for non-negative arguments, Comput. Phys. Comm. 115 (1998) 69–86.
- [14] G. Taubmann, Parabolic cylinder functions U(n,x) for natural *n* and positive *x*, Comput. Phys. Comm. 69 (1992) 415–419.
- [15] N.M. Temme, Uniform asymptotic expansions of Laplace integrals, Analysis 3 (1983) 221-249.
- [16] N.M. Temme, Laplace integrals: transformation to standard form and uniform asymptotic expansion, Quart. Appl. Math. XLIII (1985) 103–123.
- [17] N.M. Temme, Special Functions: An Introduction to the Classical Functions of Mathematical Physics, Wiley, New York, 1996.
- [18] N.M. Temme, Numerical Algorithms for uniform Airy-type asymptotic expansions, Numer. Algorithms 15 (1997) 207–225.
- [19] W.J. Thompson, Atlas for Computing Mathematical Functions: An Illustrated Guide for Practitioners, with Programs in Fortran 90 and Mathematica. There is also an edition with programs in *C*, Wiley, New York, 1997.
- [20] H.F. Weber, Über die Integration der partiellen Differential-gleichung: $\partial^2 u/\partial x^2 + \partial^2 u/\partial y^2 + k^2 u = 0$, Math. Anal. 1 (1869) 1–36.
- [21] S. Wolfram, The Mathematica Book, 3rd Edition, Wolfram Media, Cambridge University Press, Champaign, IL, 1996.
- [22] R. Wong, Asymptotic Approximations of Integrals, Academic Press, New York, 1989.
- [23] S. Zhang, J. Jin, Computation of Special Functions, Wiley, New York, 1996. Diskette with software included.



Journal of Computational and Applied Mathematics 121 (2000) 247-296

JOURNAL OF COMPUTATIONAL AND APPLIED MATHEMATICS

www.elsevier.nl/locate/cam

Computational strategies for the Riemann zeta function

Jonathan M. Borwein^{a,*}, David M. Bradley^b, Richard E. Crandall^c

^aCentre for Experimental and Constructive Mathematics, Simon Fraser University, Burnaby, BC, Canada V5A 1S6 ^bDepartment of Mathematics and Statistics, University of Maine, 5752 Neville Hall, Orono, ME 04469-5752, USA ^cCenter for Advanced Computation, Reed College, Portland, OR, 97202, USA

Received 30 December 1999; received in revised form 14 January 2000

Abstract

We provide a compendium of evaluation methods for the Riemann zeta function, presenting formulae ranging from historical attempts to recently found convergent series to curious oddities old and new. We concentrate primarily on practical computational issues, such issues depending on the domain of the argument, the desired speed of computation, and the incidence of what we call "value recycling". © 2000 Elsevier Science B.V. All rights reserved.

1. Motivation for efficient evaluation schemes

It was, of course, a profound discovery of Riemann that a function so superbly exploited by Euler, namely

$$\zeta(s) = \sum_{n=1}^{\infty} \frac{1}{n^s} = \prod_{p \text{ prime}} (1 - p^{-s})^{-1}$$
(1)

could be interpreted – to great advantage – for general complex *s*-values. Sum (1) defines the *Riemann zeta function* in the half-plane of absolute convergence $\Re(s) > 1$, and in the entire complex plane (except for the pole at s = 1) by analytic continuation. The purpose of the present treatise is to provide an overview of both old and new methods for evaluating $\zeta(s)$.

Starting with Riemann himself, algorithms for evaluating $\zeta(s)$ have been discovered over the ensuing century and a half, and are still being developed in earnest. But why concentrate at all on computational schemes? One reason, of course, is the intrinsic beauty of the subject; a beauty which cannot be denied. But another reason is that the Riemann zeta function appears – perhaps surprisingly – in many disparate domains of mathematics and science, well beyond its canonical

^{*} Corresponding author.

E-mail addresses: jborwein@cecm.sfu.ca (J.M. Borwein), bradley@gauss.umemat.maine.edu (D.M. Bradley), crandall@reed.edu (R.E. Crandall).

domain of analytic number theory. Accordingly, we shall provide next an overview of some such connections, with the intent to underscore the importance of efficient computational methods.

Typically, a particular method is geared to a specific domain, such as the critical strip $0 < \Re(s) < 1$, or the positive integers, or arguments lying in arithmetic progression, and so on. We shall honor this variety of purpose in presenting both old and new evaluation methods with a view to the specific domain in question. Just as the method of choice for evaluation tends to depend on the domain, the domain in turn typically depends on the theoretical or computational problem at hand. Though much of the present treatment involves new results for *s*-values in *integer* arithmetic progression, we shall digress presently to mention the primary historical motivation for ζ evaluation: analytic number theory applications.

There are well-known and utterly beautiful connections between number-theoretical facts and the behavior of the Riemann zeta function in certain complex regions. We shall summarize some basic connections with a brevity that belies the depth of the subject. First we state that ζ evaluations in certain complex regions of the *s*-plane have been used to establish theoretical bounds. Observe from definition (1) that, in some appropriate sense, full knowledge of ζ behavior should lead to full knowledge of the prime numbers. There is Euler's rigorous deduction of the infinitude of primes from the appearance of the pole at s=1; in fact, he deduced the stronger result that the sum of the reciprocals of the primes diverges. There is the known [60] equivalence of the *prime number theorem* [55]:

$$\pi(x) \sim \operatorname{li}(x) := \int_0^x \frac{\mathrm{d}u}{\log u} \sim \frac{x}{\log x}$$
(2)

with the nonvanishing of $\zeta(s)$ on the line $\Re(s)=1$. Here, the li integral assumes its Cauchy principal value. (Note that some authors define li in terms of an integral starting at u=2 and differing from our present integral by an absolute constant.)

Another way to witness a connection between prime numbers and the Riemann zeta function is the following. We observe that behavior of $\zeta(s)$ on a line such as $\Re(s) = 2$ in principle determines $\pi(x)$. In fact, for any noninteger x > 1,

$$\pi^*(x) := \pi(x) + \frac{1}{2}\pi(x^{1/2}) + \frac{1}{3}\pi(x^{1/3}) + \dots = \frac{1}{2\pi i} \int_{c-i\infty}^{c+i\infty} \frac{x^s}{s} \log \zeta(s) \, \mathrm{d}s, \tag{3}$$

for any real c > 1. If one can perform the contour integral to sufficient precision, then one has a value for π^* and may peel off the terms involving $\pi(x^{1/n})$ successively, for example by recursive appeal to the same integral formula with reduced x. This notion underlies the Lagarias–Odlyzko method for evaluation of $\pi(x)$ [76]. Those authors suggest clever modification, based on Mellin transforms, of the contour integrand. The idea is to transform x^s/s to a more convergent function of $\Im(s)$, with a relatively small penalty in necessary corrections to the π^* function. Experimental calculations using standard 64-bit floating point arithmetic for the ζ evaluations for quadrature of the contour integral – with, say, Gaussian decay specified for the integrand – can evidently reach up to $x \sim 10^{14}$ but not much further [58,48]. Still, it should eventually be possible via such analytic means to exceed current records such as:

 $\pi(10^{20}) = 2220819602560918840$

obtained by M. Deléglise, J. Rivat, and P. Zimmerman via nonanalytic (i.e. combinatorial) means. In fact, the Lagarias–Odlyzko remains the (asymptotically) fastest known $\pi(x)$ counting method, requiring only $O(x^{1/2+\varepsilon})$ bit complexity and $O(x^{1/4+\varepsilon})$ memory. The primary remaining obstacle to analytic

superiority is the sheer difficulty of high-precision ζ evaluations, especially in regard to rigorous error bounds, of which there is historically a definite paucity when one looks away from the critical line.

Then there are profound bounds on the fluctuations of prime densities – that is, error bounds on the prime number formula – depending on the celebrated *Riemann hypothesis*, that all the zeros in the critical strip $0 < \Re(s) < 1$ – call these the critical zeros – lie on the *critical line* $s = \frac{1}{2} + it$. In this regard, a different way of exploring the connection between ζ and prime numbers runs as follows. Riemann established the following relation, valid for noninteger x > 1:

$$\pi^*(x) = \mathrm{li}(x) - \sum_{\rho} \mathrm{li}(x^{\rho}) + \int_x^{\infty} \frac{\mathrm{d}u}{u(u^2 - 1)\log u} - \log 2,$$

where ρ runs over all zeros in the critical strip, that is $0 < \Re(\rho) < 1$, and counting multiplicity. Incidentally, the conditionally convergent sum over zeros $\rho = \sigma + it$ is to be interpreted as the limit of the sum over $|t| \leq T$ as $T \to \infty$ [62,55,63]. Arising from this kind of analysis is a highly refined prime-number estimate – due in essence to Riemann – involving not π^* but the elusive π function itself. Since one can write

$$\pi(x) = \sum_{m=1}^{\infty} \frac{\mu(m)}{m} \pi^*(x^{1/m}),$$

where μ denotes the *Möbius function*, it should be the case that, in some appropriate sense

$$\pi(x) \sim \operatorname{Ri}(x) - \sum_{\rho} \operatorname{Ri}(x^{\rho}) \tag{4}$$

with Ri denoting the Riemann function defined:

$$\operatorname{Ri}(x) = \sum_{m=1}^{\infty} \frac{\mu(m)}{m} \operatorname{li}(x^{1/m}).$$
(5)

This relation (4) has been called "exact" [94], yet we could not locate a proof in the literature; such a proof should be nontrivial, as the conditionally convergent series involved are problematic. In any case relation (4) is quite accurate (see below), and furthermore the Riemann function Ri can be calculated efficiently via evaluations of ζ at *integer* arguments in the Gram formula we encounter later (relation (70)).

The sum in (4) over critical zeros is not absolutely convergent, and furthermore the phases of the summands interact in a frightfully complicated way. Still, we see that the known equivalence of the Riemann hypothesis with the "best-possible" prime number theorem:

$$\pi(x) - \operatorname{li}(x) = \operatorname{O}(\sqrt{x}\log x)$$

makes heuristic sense, as under the celebrated hypothesis $|x^{\rho}|$ would be \sqrt{x} for every relevant zero in (4). Incidentally, as far as this equivalence goes, it is even possible to give explicit values for the implied big-O constant [10]. For example, for x > 2700 the magnitude of the left-hand side – under the Riemann hypothesis – never exceeds $(1/8\pi)\sqrt{x}\log x$. One way to find rigorous, explicit bounds on certain sums over critical zeros (on the Riemann hypothesis) is to use the known [10] exact relation

$$\sum |\rho|^{-2} = 1 + \frac{1}{2}\gamma - \frac{1}{2}\log(4\pi),$$

which incidentally is one possible overall check on any computational runs over many zeros. For example, the left-hand sum above, over the first 200 zeros (with t > 0) and their conjugate zeros, is ~ 0.021 while the right-hand constant is ~ 0.023.

Let us consider numerical experiments pertaining to $\pi(x)$ itself. If one uses the Ri formalism together with the first 200 critical zeros (with t > 0) and their conjugates, a numerical estimate from relation (4) is

 $\pi(10^{20}) \sim 2220819602591885820,$

evidently correct to about 1 part in 10¹¹. This is certainly more accurate than the direct, prime-number-theorem estimate:

 $li(10^{20}) \sim 2220819602783663484.$

It is in this way that Riemann critical zeros reveal, albeit somewhat unforgivingly, truths about prime numbers. Incidentally, as a computational matter, a convenient way to obtain numerical evaluations for li is to use the formal identity $li(z)=Ei(\log z)$, where Ei denotes the standard exponential integral, the latter standard function often having the superior machine implementation.

Because of such analytical connections, each of which underscoring the importance of the Riemann hypothesis, massive numerical calculations have been carried out over certain complex regions, such manipulations in turn depending on rapid evaluation of $\zeta(s)$. In 1979 Brent [32] showed that the first 81 million critical zeros lie on the critical line. In 1986 van de Lune et al. [82] showed that the first 1.5 billion critical zeros also lie on the critical line. The Odlyzko–Schönhage method for ζ evaluation in complex regions – which method we discuss in later sections – can be used to extend such massive calculations yet further. Indeed, Odlyzko showed efficacy by calculating $1.5 \cdot 10^8$ zeros near the 10^{20} th zero, and lately he has pressed such computations further, to the region of the 10^{22} nd zero. Then there is the *Mertens conjecture*, that

$$\left|\sum_{n \leq x} \mu(n)\right| < \sqrt{x} \quad \text{for all } x \geq 1,$$

where μ denotes the Möbius function, which conjecture was disproved by numerical efforts involving computation of the first 2000 critical zeros [88]. We note here that an exploratory discussion – from various vantage points – of the Riemann hypothesis appears in Section 8. In the earlier part of the 20th century Littlewood [81] performed a *tour de force* of analysis by establishing that $\pi(x)$ and li(x) trade dominance infinitely often, in fact

$$\pi(x) - \operatorname{li}(x) = \Omega_{\pm}\left(\frac{\sqrt{x}\log\log\log x}{\log x}\right),$$

although we know not a single explicit x > 2 such that $\pi(x)$ is the greater. After Littlewood's proof an upper bound on the first instance of integer x with $\pi(x) > li(x)$ was given, on the Riemann hypothesis, as a gargantuan, triply exponentiated "Skewes number":

 $10^{10^{10^{34}}}$.

Skewes later removed the dependency to give an even larger, unconditional bound [98,99]. Through the work of Lehman and te Riele the bound has been brought down to 10^{371} , again using numerical values of critical zeros [100]. Rosser and Schoenfeld have likewise analyzed complex zeros of related functions to establish interesting bounds on yet other number-theoretical conjectures. For example, they show that every integer greater than or equal to 2 is a sum of at most 7 primes [85]. More recently, Bays and Hudson [15] have shown how to use zeros of Dirichlet *L*-functions to quite

efficiently compute the difference $\pi_1(x) - \pi_3(x)$, for large $x \sim 10^{300}$ say, with $\pi_k(x)$ here being the number of primes $\equiv k \pmod{4}$ and not exceeding x. Because of the obvious relevance to number theory, we shall touch upon the problem of computational complexity for $\zeta(\frac{1}{2} + it)$ in Section 7.

But there are likewise beautiful, less familiar connections between the Riemann zeta function and number-theoretical conjectures. Consider, for example, as proposed by Bach [7,8] and analyzed in part also by Flajolet and Vardi [56] the following three constants: the *Artin constant A*, the *Mertens constant B*, and the *twin-prime constant C*:

$$A = \prod_{p} \left(1 - \frac{1}{p(p-1)} \right),\tag{6}$$

$$B = \gamma + \sum_{p} \left(\log(1 - p^{-1}) + p^{-1} \right), \tag{7}$$

$$C = \prod_{p>2} \left(1 - \frac{1}{(p-1)^2} \right),$$
(8)

in which product (6) and sum (7) run over all primes p, and product (8) runs over all odd primes. The constant A arises in the theory of primitive roots, B arises in the powerful asymptotic relation $\sum_{p \le x} 1/p \sim B + \log \log x$, and C arises in detailed conjectures regarding the density of twin prime pairs. Relevant series developments for these constants are:

$$-\log A = \sum_{n=2}^{\infty} \frac{\log \zeta(n)}{n} \sum_{m < n, m \mid n} \mu(m) a_{n/m-1},$$
(9)

$$B - \gamma = \sum_{n=2}^{\infty} \frac{\log \zeta(n)}{n} \mu(n), \tag{10}$$

$$-\log C = \sum_{n=2}^{\infty} \frac{\log((1-2^{-n})\zeta(n))}{n} \sum_{m < n, m \mid n} \mu(m)(2^{n/m}-2),$$
(11)

where $a_0 = 0$, $a_1 = 1$, otherwise $a_k = a_{k-1} + a_{k-2} + 1$. A fascinating aspect of these relations is this: whereas the original definitions (6)–(8), if used directly for computation, involve agonizingly slow convergence (not to mention determination of primes), the three series (9)–(11) each converge so rapidly that any of *A*, *B*, *C* may be determined to hundreds of digits in a convenient sitting. Incidentally, there are yet more interesting relations between number-theoretical constants and such entities as the logarithmic derivative $\zeta'(s)/\zeta(s)$ [9].

It is worthwhile to observe that the so-called "prime- ζ " function

$$P(s) = \sum_{p \text{ prime}} p^{-s}$$

can be evaluated to surprisingly high precision due to the identity

$$P(s) = \sum_{n=1}^{\infty} \frac{\mu(n)}{n} \log \zeta(ns).$$

For example, a certain problem in connection with the arrangement of pairs of coprime planar coordinates [107] amounts to analyzing the product

$$f(z) = \prod_{p \text{ prime}} \left(1 - \frac{z}{p^2}\right)$$

for some $z \neq -1, 0, 1$ (for each of which three values the product is well known). The problem can be solved in one sense by observing that

$$\log f(z) = -\sum_{m=1}^{\infty} P(2m) \frac{z^m}{m},$$

whence the Taylor coefficients of log f can be obtained to extreme precision without one having to know a vast collection of primes. Incidentally, one theoretically convenient aspect of the prime- ζ is that in the prime-counting relation (3), if one replaces $\zeta(s)$ with P(s), and x > 0 is again not an integer, then the left-hand side is just $\pi(x)$ rather than $\pi^*(x)$.

Still in connection with analysis, many interesting identities are manifestations of what we shall call *"rational \zeta-series"*, being explicit representations of some real number *x*, in the form

$$x = \sum_{n=2}^{\infty} q_n \zeta(n, m), \tag{12}$$

where each q_n is a rational number, *m* is fixed, and the $\zeta(n,m)$ are instances of the standard *Hurwitz* zeta function

$$\zeta(s,m) = \sum_{n=0}^{\infty} \frac{1}{(n+m)^s}.$$
(13)

Note that $\zeta(s, 1) = \zeta(s)$; the easy rule-of-thumb is that for integer *m* the Hurwitz $\zeta(s, m)$ is a zeta-like sum that starts with $1/m^s$. Thus for integer *m* the rational ζ -series (12) takes the form

$$x = \sum_{n=1}^{\infty} q_n \left(\zeta(n) - \sum_{j=1}^{m-1} j^{-n} \right),$$

in which the *n*th term decays roughly as q_n/m^n . We shall see in Section 4 that many fundamental constants enjoy convenient, rational ζ -series representation; and we shall be concentrating, then, on the variety involving $\zeta(n, 2)$.

Relations (9)–(11) involve collections of ζ -values and thus provide additional motive for what we call "value recycling" (Section 6). By this we refer to scenarios in which initial calculated values convey some information in regard to other values; so for instance some set of known ζ -values are used to get others, or many values interact symbiotically. (We had thought to call such approaches "parallel" schemes, but that is a slight misnomer because a single, scalar processor can benefit full well from most of the strategies we describe.) The motive for recycling ζ -values at integer arguments is especially strong when a rational ζ -series is essentially the only known recourse for numerical evaluation, for in such cases one desires large collections of ζ -values. In Section 4 we give examples to show that this last resort – when one is compelled to rely upon a rational ζ -series – does arise in practice.
2. Collected relations

We next list standard properties of the *Riemann zeta function*. For $\Re(s) > 1$, $\Re(\mu) > -1$ we have a Hurwitz zeta representation:

$$\zeta(s,\mu+1) = \frac{1}{\Gamma(s)} \int_0^\infty \frac{t^{s-1} e^{-\mu t}}{e^t - 1} dt, \quad \zeta(s) = \zeta(s,1),$$
(14)

whereas over the somewhat larger region $\Re(s) > 0$ the Riemann zeta function can be determined in proportion to the η function:

$$\eta(s) = (1 - 2^{1-s})\zeta(s) = \frac{1}{\Gamma(s)} \int_0^\infty \frac{t^{s-1}}{e^t + 1} dt.$$

As we shall see in Section 3, these integrals themselves already yield interesting, convergent expansions suitable for computation; not, however, always the fastest available. In Riemann's own works one finds integral representations that define $\zeta(s)$ for all complex s, for example,

$$\pi^{-s/2}\Gamma(\frac{1}{2}s)\zeta(s) = -\frac{1}{s} - \frac{1}{1-s} + \frac{1}{2}\int_{1}^{\infty} (t^{(1-s)/2} + t^{s/2})(\theta_3(e^{-\pi t}) - 1)\frac{\mathrm{d}t}{t},\tag{15}$$

in which the *Jacobi theta-function* [22] is $\theta_3(q) = \sum_{n=-\infty}^{\infty} q^{n^2}$. This representation will give rise to an (extremely) rapidly converging series (30), although the summands will be nonelementary. The collection of entire representations is by no means limited to (15). For example, there is the Jensen formula

$$\zeta(s) = \frac{1}{2} + \frac{1}{s-1} + 2\int_0^\infty \frac{\sin(s\tan^{-1}t)\,\mathrm{d}t}{(1+t^2)^{s/2}(\mathrm{e}^{2\pi t}-1)},$$

also valid for all $s \neq 1$, and useful in certain proofs of the prime number theorem [60].

From (15), there follows immediately the celebrated functional equation. If we define

$$\xi(s) = \frac{1}{2}s(s-1)\pi^{-s/2}\Gamma(\frac{1}{2}s)\zeta(s),$$
(16)

then the functional equation can be written elegantly [55] as

$$\xi(s) = \xi(1-s).$$
 (17)

Furthermore, by considering complex values $s = \frac{1}{2} + it$, one sees that the Riemann hypothesis is true if and only if all zeros of the function

$$\Xi(t) = -\frac{1}{2}(t^2 + \frac{1}{4})\xi(\frac{1}{2} + it)$$
(18)

are real [101]. The idea of forging a real-valued facsimile on the critical line is a good one, conducive to numerical analysis such as locating critical zeros. But the Ξ -function decays rapidly for large t, so in practice a more reasonable choice is the function (sometimes called the *Hardy function* [72]):

$$Z(t) = \exp(i\vartheta(t))\zeta(\frac{1}{2} + it),$$
(19)

where we define ϑ implicitly by

$$e^{i\vartheta(t)} = \chi(\frac{1}{2} + it)^{-1/2}$$
(20)

and the square root is defined continuously, with fixation $\sqrt{\chi(\frac{1}{2})} = 1$. In general, one may write

$$\chi(s) = \pi^{s-1/2} \frac{\Gamma((1-s)/2)}{\Gamma(s/2)}, \qquad \vartheta(t) = \Im(\ln\Gamma(\frac{1}{4} + it/2)) - \frac{1}{2}t\ln\pi,$$
(21)

of which the latter is computationally convenient (and circumvents the need for computing χ ; see [51]). Now for real *t*, the Hardy *Z*-function is real and the equality $|Z(t)| = |\zeta(\frac{1}{2} + it)|$ holds. These convenient properties render *Z* useful in modern searches for critical zeros [32,82]. In particular, simple zeros of $\zeta(\frac{1}{2} + it)$ on the critical line are signified by sign changes – as *t* increases – of the *Z*-function, and this notion can be made rigorous by careful constraint on numerical error, so that a machine can *prove* that all zeros in the critical strip interval $t \in [0, T]$ for some fixed *T* do, in fact, lie precisely on the critical line [85]. Later in Section 3 we shall describe the kinds of error contributions that appear in prevailing series developments of the *Z*-function.

It is well known that for positive even integer arguments we have exact evaluations

$$\zeta(2n) = -\frac{(2\pi i)^{2n} B_{2n}}{2(2n)!},\tag{22}$$

in terms of the *Bernoulli numbers* $1, -\frac{1}{2}, \frac{1}{6}, 0, -\frac{1}{30}, 0, \frac{1}{42}, \dots$ defined by the generating series

$$\frac{t}{e^t - 1} = \sum_{m=0}^{\infty} \frac{B_m}{m!} t^m,$$
(23)

in which $B_{2n+1} = 0$ for n > 0. For computational purposes it will turn out to be important that series (23) has radius of convergence 2π . Now from the functional equation (17) one may deduce the analytic continuation value $\zeta(0) = -\frac{1}{2}$ and the values at negative integer arguments

$$\zeta(-2n) = 0, \quad \zeta(1-2n) = -\frac{B_{2n}}{2n} \tag{24}$$

for positive integer *n*. An elegant and computationally lucrative representation for the even-argument ζ -values is

$$\pi t \cot \pi t = -2 \sum_{m=0}^{\infty} \zeta(2m) t^{2m}.$$
(25)

Series (25) converges for |t| < 1, and with this constraint in mind can be used in many different computational algorithms, including some recycling ones, as we shall discuss. On the issue of whether a convenient generating function can be obtained for *odd*-argument ζ -values, there is at least one candidate, namely the following relation involving the logarithmic derivative of the gamma function, i.e., the *digamma function* $\psi(z) = d \log \Gamma(z)/dz$:

$$\psi(1-t) = -\gamma - \sum_{n=2}^{\infty} \zeta(n) t^{n-1}, \quad |t| < 1,$$
(26)

which will be useful in the matter of value recycling.

Standard recurrence relations for Bernoulli numbers can be invoked to provide relations such as

$$\sum_{k=0}^{m} \frac{(\pi i)^{2k}}{(2k+1)!} (1 - 2^{2k-2m+1})\zeta(2m-2k) = 0$$
(27)

and for integer $k \ge 2$,

$$\sum_{j=1}^{k-1} \zeta(2j)\zeta(2k-2j) = (k+\frac{1}{2})\zeta(2k).$$

See [52] for generalizations to sums of products of $N \ge 2 \zeta$ -values, Bernoulli/Euler polynomials, and the like. Similar relations for odd-argument ζ -values are difficult if not fundamentally impossible to obtain. There are, however, some interesting relations between the values at odd integer arguments if we allow easily computed residual terms, which can be cast as rational ζ -series, as we shall see in Section 5.

Many interrelations between ζ values can be inferred from the following series development for the *complex Lerch*, or *periodic zeta function* [6,50]:

$$\sum_{n=1}^{\infty} \frac{e^{2\pi i n x}}{n^s} = -\sum_{j=0}^{\infty} \frac{(\pi i)^j}{j!} \eta(s-j)(2x-1)^j$$
(28)

valid for $\Re(s) > 0$ and real x with $|2x - 1| \le 1$. An immediate representation obtains on setting x = 0:

$$\zeta(s) = -\sum_{j=0}^{\infty} \frac{(-\pi i)^j}{j!} \eta(s-j)$$

valid for $\Re(s) > 0$. Note that if $\zeta(s)$ be real, then the imaginary part of the right-hand side vanishes, and this gives certain ζ -series representations. On the other hand, using just the real part of the right-hand side yields, for even integer *s*, the previous relation (27) for $\zeta(\text{even})$; while for odd *s* we obtain certain representations of $\zeta(\text{odd})$. The Lerch-series approach will be discussed later as a computational tool.

3. Evaluations for general complex arguments

Until the 1930s the workhorse of the evaluation art for the Riemann zeta function was Euler–Maclaurin expansion. The standard *Euler–Maclaurin formula* applied to $x \mapsto x^{-s}$ yields, for two cutoff integers M, N:

$$\zeta(s) = \sum_{n=1}^{N-1} \frac{1}{n^s} + \frac{1}{2N^s} + \frac{N^{1-s}}{s-1} + \sum_{k=1}^{M} T_{k,N}(s) + E(M,N,s),$$
(29)

where

$$T_{k,N}(s) = \frac{B_{2k}}{(2k)!} N^{1-s-2k} \prod_{j=0}^{2k-2} (s+j).$$

If $\sigma := \Re(s) > -2M - 1$ the error is rigorously bounded as [87,40]:

$$|E(M,N,s)| \leq \left|\frac{s+2M+1}{\sigma+2M+1}T_{M+1,N}(s)\right|.$$

One disadvantage of such expansions is universal, i.e., relegated not only to the Riemann zeta function. The problem is, one does not obtain a manifestly convergent expansion; rather, the expansion is of asymptotic character and one is compelled to rescale the cutoff parameters when attempting a new precision goal. With this in mind, we proceed for much of the rest of this treatment to focus on convergent series.

Since $(s-1)\zeta(s)$ is entire, we may write

$$\zeta(s) = \frac{1}{s-1} + \sum_{n=0}^{\infty} \frac{(-1)^n \gamma_n}{n!} (s-1)^n$$

The coefficients are generally referred to as the Stieltjes constants and are given by

$$\gamma_n = \lim_{m \to \infty} \left\{ \sum_{k=1}^m \frac{\log^n(k)}{k} - \frac{\log^{n+1}(m)}{n+1} \right\}.$$

Note that $\gamma_0 = 0.5772156649...$ is the *Euler constant* (which we heretofore call just γ). In principle, the Stieltjes expansion here gives a scheme for evaluation of Euler's constant, provided one has a sufficiently sharp scheme for $\zeta(1 + \varepsilon)$.

From (15), one has

$$\zeta(s)\Gamma(\frac{1}{2}s) = \frac{\pi^{s/2}}{s(s-1)} + \sum_{n=1}^{\infty} n^{-s}\Gamma(\frac{1}{2}s,\pi n^2) + \pi^{s-1/2} \sum_{n=1}^{\infty} n^{s-1}\Gamma(\frac{1}{2}(1-s),\pi n^2),$$
(30)

in principle, a consummately convergent expansion, the only obstacle to high efficiency being the evaluations of the *incomplete gamma function*, given (at least for $\Re(z) > 0$) by

$$\Gamma(a,z) = \int_{z}^{\infty} t^{a-1} e^{-t} dt = \frac{2z^{a} e^{-z}}{\Gamma(1-a)} \int_{0}^{\infty} \frac{t^{1-2a} e^{-t^{2}}}{t^{2}+z} dt$$

where the latter integral representation is valid for (an important region) $\Re(a) < 1$. But the evaluation of $\Gamma(a,z)$ is not as problematic as it may seem; many computer systems of today have suitable incomplete-gamma machinery. There are the special cases $\Gamma(s,0) = \Gamma(s)$ and $\Gamma(1,z) = e^{-z}$, with a recursion

$$a\Gamma(a,z) = \Gamma(a+1,z) - z^a e^{-z}$$
(31)

that proves useful, as we shall see, in the art of value recycling. The recursion also reveals that when a is a positive integer $\Gamma(a,z)$ is an elementary function of z. There is an at least threefold strategy for evaluating the incomplete gamma [44]. For $a \neq 0, -1, -2, \ldots$ one has an ascending hypergeometric series and transformed counterpart:

$$\Gamma(a,z) = \Gamma(a) - a^{-1} z_1^a F_1(a;a+1;-z),$$

= $\Gamma(a) \left(1 - z^a e^{-a} \sum_{m=0}^{\infty} \frac{z^m}{\Gamma(a+m+1)} \right),$

while for larger values of |z| one may use the continued fraction (when it exists – the convergence issues for general complex *a* are intricate and fascinating, see [60] or the recent treatment [3]):

$$\Gamma(a,z) = \frac{z^{a}e^{-z}}{z + \frac{1-a}{1 + \frac{1}{z + \frac{2-a}{1 + \frac{1}{z + \frac{2-a}{1 + \frac{1}{z + \frac{1$$

where pairs of consecutive numerators here take the form $\{n - a, n\}$ as *n* runs through the positive integers. For extremely large |z| values one has a standard asymptotic series:

$$\Gamma(a,z) \sim z^{a-1} e^{-z} \left(1 + \frac{a-1}{z} + \frac{(a-1)(a-2)}{z^2} + \cdots \right)$$

valid at least for $\Im(z) > 0$. Convergence and error-bounding issues can be resolved via proper analysis of appropriate Mellin–Barnes contour integrals, as discussed in [48]. For the moment we underscore the rapid convergence of series (30) by noting the behavior for positive real $z > \sigma$:

$$|\Gamma(\sigma + \mathrm{i}t, z)| < \max(1, 2^{\sigma}) z^{\sigma-1} \mathrm{e}^{-z}.$$

This bound is quite convenient in practice, and means that only $O(\sqrt{D})$ terms of a summand in series (30) are required to achieve *D* correct digits. A generalization of the incomplete-gamma series is useful in higher dimensions, specifically when Epstein zeta functions (generalizations of the Riemann zeta function) are to be evaluated [47].

Series (30) should always be considered as a possible expedient for evaluating ζ . We note that, especially for large $|\Im(s)|$, the Riemann–Siegel formula can be superior, easier to apply in practice, and also supports recycled evaluation of the Odlyzko–Schönhage type. But a recycling option also exists – albeit in a different sense and over different complex domains – for relation (30); for example, the recursion relation (31) allows recycling for certain arithmetic progressions of arguments, as we shall see later.

It is sometimes noted that a formula such as (30) suffers from precision loss when $|\Im(s)|$ is large, due to the factor $\Gamma(s/2)$ on the left, which factor in such instances being an exponentially small one, decaying as $\sim \exp(-\pi |\Im(s)|/4)$. But there is the notion of using a free parameter in formula (30), and furthermore allowing said parameter to attain complex values in order to reduce this precision loss. The interesting work of Rubinstein [96] on more general *L*-function evaluation contains analysis of this type, along with yet more incomplete-gamma representations. Other treatments of ζ on the critical line depend also on incomplete-gamma asymptotics, such as the Temme formulae [89]. In the same spirit there is ongoing research into the matter of casting series of type (30) in more elementary terms, with a view to practical computation, by using a combination of: complex free parameter, rigorous error bounds, and special expansions of the incomplete gamma at certain saddle points [49].

From the integral representation (14) together with the generating series (23), we can choose $|\lambda| < 2\pi$ and obtain

$$\zeta(s)\Gamma(s) = -\frac{\lambda^s}{2s} + \frac{\lambda^{s-1}}{s-1} + \sum_{n=0}^{\infty} n^{-s}\Gamma(s,\lambda n) - 2\lambda^{s-1}\sum_{n=1}^{\infty} \left(\frac{\lambda}{2\pi i}\right)^{2n} \frac{\zeta(2n)}{2n+s-1},\tag{32}$$

which is valid over the entire complex s-plane, provided we properly handle the limiting case $s \rightarrow n$ for a negative integer n. In fact, the pole in $\Gamma(s)$ on the left corresponds to the pole in the relevant summand in the second sum, and we derive all at once evaluations (24). Now (32) is an intriguing and sometimes useful expansion. The free parameter λ allows one to test quite stringently any numerical scheme: one must obtain invariant results for any λ chosen in the allowed domain. For positive integer arguments s, the incomplete gamma is *elementary*; furthermore, for such s and rational $\lambda/(2\pi)$, the second sum in (32) has all rational coefficients of the $\zeta(2n)$. We shall have more to say about this expansion in Section 7.

An interesting method for development of manifestly convergent series such as (30) and (32) starts with representation (28) for the Lerch function. If we set $x = \frac{1}{2} + i\lambda/(2\pi)$ then, formally at least

$$(1-2^{1-s})\zeta(s) = \eta(s) = \sum_{n=1}^{\infty} \frac{(-1)^{n-1}}{n^s} e^{-\lambda n} - \sum_{j=1}^{\infty} \frac{(-\lambda)^j}{j!} \eta(s-j).$$
(33)

It can be shown that this relation is valid for all complex s, with free parameter $\lambda \in (0, \pi]$. Later, in Section 5 we discuss specific applications for integer arguments s.

The *Stark formula*, also analyzed by Keiper [45,106], provides a different approach for general complex s and N a positive integer:

$$\zeta(s,N) = \frac{1}{s-1} \sum_{k=1}^{\infty} \left(N + \frac{s-1}{k+1} \right) (-1)^k \binom{s+k-1}{k} \zeta(s+k,N),$$

which by its very construction admits of interesting recursion schemes: one can write a Hurwitz- ζ function that calls itself. Reminiscent of the Stark–Keiper approach is the formula

$$\zeta(s) = \lim_{N \to \infty} \frac{1}{2^{N-s+1} - 2^N} \sum_{k=0}^{2N-1} \frac{(-1)^k}{(k+1)^s} \left(\sum_{m=0}^{k-N} \binom{N}{m} - 2^N \right)$$
(34)

for which it is possible to give a rigorous error bound as a function of the cutoff N and s itself [23,27]. Very recently there appeared the interesting Woon formula, which amounts to a relation involving Bernoulli numbers that generalizes the celebrated formula (24). We paraphrase the Woon formula thus: for free real parameter w > 0 and $\Re(s) > 1/w$, one has

$$\zeta(s) = -\pi (2\pi w)^{s-1} \sec(\frac{1}{2}\pi s) \sum_{n=0}^{\infty} \frac{(-1)^n b(w,n) \Gamma(s)}{n! \Gamma(s-n)},$$

where we define

$$b(w,n) = \frac{1}{2} + 2w \sum_{m=2}^{n+1} \left(\frac{i}{2\pi w}\right)^{2m} \binom{n}{2m-1} \zeta(2m).$$

Note that for positive even integer s, this whole scheme boils down to a tautology, because we have intentionally replaced (on the right-hand side of the b definition) the Bernoulli coefficients of Woon's original rendition with ζ (even) values. It is of interest that this formula becomes singular only at odd integer values of s (where the secant diverges), although Woon has specified a limiting process in such cases [108].

We end this section with a discussion of practical issues for the *Riemann–Siegel formula*. This formula and its variants amount to the most powerful evaluation scheme known for *s* possessed of large imaginary part – the quite elegant and profound developments are referenced in [32,101,55,18,87,19, 64,63]. Another unique aspect of the Riemann–Siegel formula is that it is relatively difficult to implement, having several terms each requiring its own special strategy. Yet another is the fact that different variants apply best in different complex regions, with different error-bounding formula applicable in problem-dependent fashion. Loosely speaking, the Riemann–Siegel formulae apply in two modes. Representatives of these modes are first, calculations on the critical line $s = \frac{1}{2} + it$ (for which the Z-function (19) is appropriate); and second, evaluations with $\Re(s) > 1$, as in the evaluation algorithms for integral (3) (for which $\log \zeta$ is desired). In all such instances any bounding formula must take into account the decay of error as a function of the imaginary part *t*.

The Riemann–Siegel formula for ζ itself – as opposed to variants attendant on the Hardy Z-function – can be written as an "approximate functional equation":

$$\zeta(s) = \sum_{n=1}^{M} \frac{1}{n^{s}} + \chi(s) \sum_{n=1}^{M} \frac{1}{n^{1-s}} + E_{M}(s),$$

where M is a certain cutoff value, the χ -function is from relation (21), and E_M is an error term that, although depending in a complicated way on the intended domain of s, can be bounded explicitly for computations in certain useful regions of the complex s-plane [87]. We note that the formula admits of more general rendition – in which the limits on the summands are unequal – and that an optimized inequality of said limits may be called for when one is working off the critical line. There is a long-studied theory for this kind of approximation, and there remain open questions on the precise asymptotic nature of the errors [101,63,55,18]. In particular, there is a distinct paucity of useful explicit bounds for s off the critical line, but research is ongoing into this dilemma [48].

The Riemann–Siegel formula above is certainly a streamlined rendition. The detailed error terms are complicated [87,101,63,57]; moreover, the level of asymptotic correction, the number of error terms to add in, and so on depend on the required precision and the complex domain of *s*. Accordingly, we shall give below just one practical variant and some explicit error bounds. As for the Hardy function (19), a similarly stripped-down rendition is [64]

$$Z(t) = 2 \sum_{1 \le n \le \tau} n^{-1/2} \cos(t \log(n^{-1}\tau) - \frac{1}{2}t - \frac{1}{8}\pi) + O(t^{-1/4}), \quad t > 0,$$

where $\tau = \sqrt{t/(2\pi)}$. It turns out the big-O error term here is best possible, because the indicated error is actually $\Omega_{\pm}(t^{-1/4})$ (not surprising – for one thing, the discontinuity implicit in the summation cutoff is of this magnitude). We now give just one reliable form of an expanded Riemann–Siegel formula for Z. In their numerical researches on the critical zeros, Brent et al. [32,33,82] used the following practical variant. To simplify notation, let $m = \lfloor \tau \rfloor$, $z = 2(\tau - m) - 1$. Then the variant involves the angle ϑ from definition (20), (21), which angle is relatively easy to calculate from gamma-function asymptotics, and reads

$$Z(t) = 2\sum_{n=1}^{m} n^{-1/2} \cos(t \log n - \vartheta(t)) + (-1)^{m+1} \tau^{-1/2} \sum_{j=0}^{M} (-1)^j \tau^{-j} \Phi_j(z) + R_M(t).$$
(35)

Here, M is a cutoff integer of choice, the Φ_j are entire functions defined for $j \ge 0$ in terms of a function Φ_0 and its derivatives, and $R_M(t)$ is the error. For computational rigor one needs to know an explicit big-O constant. A practical instance is Brent's choice M = 2, for which we need

$$\begin{split} \Phi_0(z) &= \frac{\cos(\frac{1}{2}\pi z^2 + \frac{3}{8}\pi)}{\cos(\pi z)}, \\ \Phi_1(z) &= \frac{1}{12\pi^2} \Phi_0^{(3)}(z), \\ \Phi_2(z) &= \frac{1}{16\pi^2} \Phi_0^{(2)}(z) + \frac{1}{288\pi^4} \Phi_0^{(6)}(z). \end{split}$$

All of this notational tangle may appear stultifying, but the marvelous benefit is this: the errors R_M have been rigorously bounded, in computationally convenient fashion by various investigators – notably Gabcke [57] – to achieve such as the following, for $t \ge 200$, $M \le 10$:

$$|R_M(t)| < B_M t^{-(2M+3)/4}$$

for a set of bounding numbers:

$${B_0, \ldots, B_{10}} = \{0.127, 0.053, 0.011, 0.031, 0.017, 0.061, 0.661, 9.2, 130, 1837, 25966\}.$$

Now the computationalist does not have to interpret big-O notation in numerical experiments. Perhaps surprisingly, regardless of these beautiful bounds the Riemann–Siegel formula with just M = 1 – so that R_1 is in force – was enough to resolve the first 1.5 billion zeros, in the following sense. The optimized strategy in [82] for finding and *proving* that zeros lie exactly on the critical line, which strategy stems from that used originally by Brent [32], was reported never to have failed with the R_1 bound in hand. Incidentally, the zero-location method is ingenious: one uses known rigorous bounds on the number of zeros in a vertical segment of the critical strip. For example the number of zeros having $t \in [0, T]$ can be obtained from [101]

$$N(T) = 1 + \pi^{-1}\vartheta(T) + \pi^{-1} \Delta \arg \zeta(s),$$

where ϑ is the angle from assignment (20) and Δ signifies the variation in the argument, defined to start from $\arg \zeta(2) = 0$ and varying continuously to s = 2 + iT, then to $s = \frac{1}{2} + iT$. If some number of sign changes of Z(t) has been counted, and this count saturates the theoretical bound (e.g., bound says N(t) < 15.6 zeros and we have found 15), then all the zeros in the segment must have been found: they must lie precisely on $\Re(s) = \frac{1}{2}$ and furthermore they must be simple zeros because Z sustained changes in sign.

It should be pointed out that most of the work in these hunts for critical zeros is in the evaluation of a finite sum:

$$\sum_{n=1}^{\infty} n^{-1/2} \cos(t \log n - \vartheta),$$
(36)

where we recall that $m = \lfloor \tau \rfloor$ is the greatest integer not exceeding $\sqrt{t/(2\pi)}$. The authors of [82] in fact vectorized this sum in supercomputer fashion. Computational issues aside, one can also envision – by pondering the phase of the cosine – how it is that zeros occur, and with what (approximate) frequency [101].

There is an interesting way to envision the delicate inner workings of the Riemann-Siegel formula (35). Note the implicit discontinuity of the *n*-summation; after all, the summation limit $m = \lfloor \tau \rfloor$ changes suddenly at certain *t*. The idea is, the *M* terms of the *j*-summation must cancel said discontinuity, up to some hopefully insignificant error. As Berry and Keating note [18], the summation limit *m* itself is a kind of critical-phase point during the analysis of those integral representations of ζ that underlie the Riemann-Siegel formalism. Berry and Keating gave, in fact, an alternative, free-parameter representation of the *Z*-function, which representation avoids discontinuities in summation. Though their leading sum is more complicated, it is also more accurate (instead of a discontinuous cutoff there is a smooth, error-function decay near the Riemann-Siegel critical-phase point *m*), and the same kind of accuracy-complexity tradeoff occurs for their ensuing correction terms. Thus the Berry-Keating form is perhaps a viable computational alternative; at the very least it has

261

theoretical importance in connection with semi-classical quantum theory and stationary states of operators (see Section 8).

Since the Riemann–Siegel formula can be derived by application of saddle point methods to integral representations, Galway [58] has noted such integrals themselves are well suited for numerical integration. This allows computation of ζ values to arbitrary accuracy while still retaining many of the advantages of the Riemann–Siegel formula. Another advantage of this method is that the analysis of the error terms is simplified.

We observe that the Riemann–Siegel formula exhibits properties in common with both the Euler– Maclaurin formula (29) and the incomplete-gamma expansion (30). As for the former similarity, the Riemann–Siegel form is asymptotic in nature, at least in the sense that one chooses a set of about M correction terms depending, in principle, on *both* the range of the argument and the required accuracy. As for the similarity with the incomplete-gamma formula, note that both formulae tend to require $O(t^{1/2})$ summands – the Riemann–Siegel by its very construction, and the incomplete-gamma by accuracy requirements. Of course, the Riemann–Siegel summands involve exclusively elementary functions, which is a strong advantage as we have intimated. We shall have more to say about such computational matters in Section 7.

4. Rational zeta series

Consider a natural specialization of the rational ζ -series (12), obtained by setting m = 1 in the Hurwitz zeta function (13). We shall discuss representations of real numbers x in the form

$$x = \sum_{n=2}^{\infty} q_n(\zeta(n) - 1),$$
(37)

where the rational coefficients q_n are, in some appropriate sense, well behaved. It is not hard to prove that *any* real x admits a rational ζ -series of the form (37) for unrestricted rational q_n ; but we are concerned with expansions for which the q_n are particularly simple in structure. One might demand the $|q_n|$ be bounded, or constrain the denominator of q_n to possess O(log n) bits, and so on. This kind of series for some desired number x tends to be computationally convenient because, of course, $\zeta(n) - 1$ decays like $(\frac{1}{2})^n$ for increasing n. It will turn out that many fundamental constants enjoy simple representations. To mention a few: π (in fact any positive integer power of π), log π , log rfor any rational r, the Euler constant γ , the Catalan constant G,¹ the Khintchine constant K_0 (actually (log K_0)(log 2)), and any quadratic surd ($A + \sqrt{B}$)/C (including, for example, the golden mean $(1 + \sqrt{5})/2$) are representable with relatively simple, explicit coefficients.

Let us consider some fundamental numbers from such disparate classes. First, there is a "representation of unity"

$$1 = \sum_{n=2}^{\infty} (\zeta(n) - 1),$$
(38)

 $^{{}^{1}}G$, γ and $\zeta(5)$ are quintessential examples of constants whose irrationality though suspected is unproven. Efficient high precision algorithms allow one to prove in these and many other cases that any rational representation must have an enormous denominator. See for example [34].

which has tremendous value in testing evaluation schemes – in particular the recycling schemes – for the $\zeta(n)$ themselves. Curiously, this representation can be partitioned into disjoint sums over even and odd *n*, respectively; the even-indexed sum having the value $\frac{3}{4}$, the odd-indexed sum having the value $\frac{1}{4}$. There are attractive representations for log 2 and the Euler constant γ :

$$\log 2 = \sum_{n=1}^{\infty} \frac{\zeta(2n) - 1}{n},$$

$$1 - \gamma = \sum_{n=2}^{\infty} \frac{\zeta(n) - 1}{n}.$$
(39)

As we shall see, the convergence of these and many related series can be duly accelerated. To give just one side example of the analytic depth of this subject, we note that Ramanujan once observed that a formula of Glaisher:

$$\gamma = 2 - 2\log 2 - 2\sum_{n=3, \text{ odd}} \frac{\zeta(n) - 1}{n(n+1)}$$

(as one can deduce from identities above and below this one) could be generalized to *infinitely* many different formulae for γ [93].

Many relations can be obtained upon manipulation of identities such as

$$\sum_{n=1}^{\infty} t^{2n}(\zeta(2n)-1) = \frac{1}{2} - \frac{1}{2}\pi t \cot \pi t - t^2(1-t^2)^{-1}, \quad |t| < 2,$$

$$\sum_{n=2}^{\infty} t^n(\zeta(n)-1) = -t(\gamma + \psi(1-t) - t(1-t)^{-1}), \quad |t| < 2,$$
(40)
(41)

which arise from expansions (25) and (26), respectively. Thus, for example, one may integrate (41) to achieve a formal expansion involving the Euler constant:

$$t(1-\gamma) + \log \Gamma(2-t) = \sum_{n=2}^{\infty} n^{-1} t^n (\zeta(n) - 1),$$
(42)

which expansion will have application later in Section 8. For $t = \frac{3}{2}$ we obtain a representation of $\log \pi$:

$$\log \pi = \sum_{n=2}^{\infty} n^{-1} (2(\frac{3}{2})^n - 3)(\zeta(n) - 1).$$

Evaluations of rational ζ -series with simple coefficients q_n can take attractive forms. For example, whereas (40) can be used to derive

$$\sum_{n=1}^{\infty} \frac{\zeta(2n) - 1}{2^{2n}} = \frac{1}{6}$$
(43)

and one of many ζ -series for π :

$$\sum_{n=1}^{\infty} \frac{\zeta(2n) - 1}{4^{2n}} = \frac{13}{30} - \frac{\pi}{8},\tag{44}$$

it also leads to

$$\sum_{n=1}^{\infty} \frac{\zeta(2n)-1}{8^{2n}} = \frac{61}{126} - \frac{\pi}{16} \sqrt{\frac{\sqrt{2}+1}{\sqrt{2}-1}}.$$

Not only can we establish such series for certain $\pi\alpha$, with α a nontrivial algebraic number; we may also insert appropriate roots of unity as *t*-parameters in (40) to obtain such as

$$\sum_{n=1}^{\infty} \left(\zeta(4n) - 1 \right) = \frac{7}{8} - \frac{1}{4} \pi \left(\frac{e^{2\pi} + 1}{e^{2\pi} - 1} \right).$$

We note that in (43), $q_n = (\frac{1}{2})^n$ for *n* even, else $q_n = 0$, provides an alternative representation of unity, to be contrasted with (38). In fact, there are infinitely many representations of unity. For example, the case $q_n = 1$ can be generalized to the following, valid for any nonnegative integer *k*.

$$1 = \sum_{n=k+2}^{\infty} \binom{n-1}{k} \left(\zeta(n) - 1\right)$$

Likewise (41) leads to interesting series, such as the following obtained by integration:

$$\sum_{n=2}^{\infty} \frac{\zeta(n) - 1}{(-1)^n (n+1)} = \frac{1}{2}\gamma + \frac{1}{2} - \log 2 - \int_1^2 \log \Gamma(z) \, \mathrm{d}z$$
$$= \frac{1}{2}(\gamma + 3 - \log 2\pi) - \log 2, \tag{45}$$

which result having a theoretical application we encounter in Section 8.

There are yet other rational ζ -series that interrelate various *Dirichlet series*. One way to derive such relations is to know, first, a cotangent integral such as

$$I_n := \int_0^{1/2} x^n \cot \pi x \, \mathrm{d}x,$$

then use expansion (25) within the integral. Actually, this integral I_n is known exactly for every positive integer *n* in terms of logarithms and values of $\zeta(\text{odd})$ [50]. One example provides a relation involving $\pi^{-2}\zeta(3)$:

$$\sum_{n=1}^{\infty} \frac{\zeta(2n) - 1}{4^n(n+1)} = \frac{3}{2} - 9\log 2 + 4\log 3 + \frac{7}{2}\pi^{-2}\zeta(3).$$
(46)

Consideration of integrals over $0 < x < \frac{1}{4}$ provide representations for $\pi^{-1}G$, where $G = 1 - \frac{1}{3}^2 + \frac{1}{5}^2 - \frac{1}{7}^2 + \cdots$ is the Catalan constant:

$$\sum_{n=1}^{\infty} \frac{\zeta(2n) - 1}{16^n(2n+1)} = \frac{3}{2} - \pi^{-1}G - \frac{1}{4}\log 2 - 2\log 5 + 2\log 3$$
(47)

and

$$\sum_{n=1}^{\infty} \frac{\zeta(2n) - 1}{16^n (2n+1)n} = 2\pi^{-1}G - 3 + 5\log 5 + \log \pi - 5\log 2 - 3\log 3.$$
(48)

Incidentally, not only rationals but logarithms of rationals as appear in (45)–(48) are easy to absorb, if necessary, into the ζ sum. We shall encounter a general logarithmic representation later in this section.

A rational ζ -series can often be accelerated for computational purposes, provided one can resolve the exact sum

$$\sum_{n=2}^{\infty} \frac{q_n}{a^n}$$

for some contiguous sequence a = 2, 3, 4, ..., A. One simply "peels off" these terms, leaving a series involving the Hurwitz terms $\zeta(n, A+1)$, i.e., ζ -like sums starting with $1/(A+1)^n$. For example, it turns out that one may peel off *any* number of terms from (46) [45]. The exact corrections for a = 2, 3, ... simply add to the detail of the logarithmic term. Perhaps the canonical example of "peeling" is the γ series (39) previously encountered. By peeling of N terms (including 1) from the ζ summand, one has

$$\gamma = \sum_{j=1}^{N} j^{-1} - \log N - \sum_{m=2}^{\infty} m^{-1} \zeta(m, N+1),$$

in which one witnesses the classical limit expression for γ plus an exact (always negative) correction. Computational complexity issues for this peeling – and other evaluation schemes – are discussed in Section 7. For the moment, we observe that if peeling be taken to its extreme limits, there may be no special advantage. For example, if we peel *all* summands in relation (46) for $\zeta(3)$, so that the whole rational ζ -series vanishes, we get the peculiar relation

$$\zeta(3) = \frac{5\pi^2}{36} - \frac{2\pi^2}{3} \sum_{n=1}^{\infty} \left\{ -\frac{5}{12} - 2n^2 + n(n+1)(2n+1)\log(1+1/2n) -n(n-1)(2n-1)\log(1-1/2n) \right\},$$
(49)

a slowly converging series indeed. Thus, the primary motivation for peeling is to optimize sums for actual computation – by peeling an optimal number of terms.

We next mention results of Flajolet and Vardi [56,102], who demonstrate that if $f(z) = \sum_{m \ge 2} f_m z^m$ is analytic on the closed unit disk, then

$$\sum_{n=1}^{\infty} f(1/n) = f(1) + \sum_{m=2}^{\infty} f_m(\zeta(m) - 1)$$

along with peeled such forms involving $\zeta(m, N)$ for N > 2. Some of the immediate results along these lines are for π :

$$\pi = \frac{8}{3} + \sum_{m=1}^{\infty} 4^{-m} (3^m - 1)(\zeta(m+1) - 1)$$
(50)

and for the Catalan constant:

$$G = \frac{8}{9} + \frac{1}{16} \sum_{m=1}^{\infty} (m+1)4^{-m}(3^m-1)(\zeta(m+2)-1).$$

The latter arises from the identity

$$(1-3z)^{-2} - (1-z)^{-2} = \sum_{m=1}^{\infty} (m+1)4^{-m}(3^m-1)z^m.$$

265

It is of interest that we thus know rational ζ -series for both G and, as formula (47) yields, G/π . One may also derive a series for π^3 , starting with the generating function $f(z) = z^3(1 - 3z/4)^{-3} - z^3(1 - z/4)^{-3}$. In fact, any odd power of π can be cast first as a Dirichlet series (actually, a rational multiple of the *beta function*, $\beta(d) = 1^{-d} - 2^{-d} + 3^{-d} - \cdots$), then one constructs f(z), quickly obtaining a series for π^d . Flajolet and Vardi [56] were able to augment the aforementioned number-theoretical representations described in Section 1 by casting such as the Landau–Ramanujan and Hafner–Sarnak–McCurley constants in terms of convergent ζ constructs.

These curious and attractive series aside, there can actually be practical import for rational ζ -series, thus motivating efficient schemes for evaluation of the relevant $\zeta(n)$. One of the most interesting applications is a result from the measure theory of continued fractions [109,11]. The celebrated Khintchine constant K_0 , defined as the limiting geometric mean of the elements of almost all simple continued fractions, can be bestowed with an efficient series development. The development is particularly compelling in that one of the standard definitions of K_0 is a cumbersome, slowly converging, infinite product. The rational ζ -series we have in mind is the Shanks–Wrench form [109] which for N > 2 can be peeled N - 2 times to yield [11]:

$$(\log K_0)(\log 2) = \sum_{n=1}^{\infty} \frac{\zeta(2n) - 1}{n} \left(1 - \frac{1}{2} + \frac{1}{3} - \dots + \frac{1}{2n - 1} \right)$$
$$= \sum_{k=3}^{N} \log \left(1 - \frac{1}{k} \right) \log \left(1 + \frac{1}{k} \right)$$
$$+ \sum_{n=1}^{\infty} \frac{\zeta(2n, N)}{n} \left(1 - \frac{1}{2} + \frac{1}{3} - \dots + \frac{1}{2n - 1} \right).$$
(51)

The peeled form has been used, together with recycling methods for evaluating ζ at the even positive integers, to obtain K_0 to thousands of digits. In like manner, for negative integers p the *p*-Hölder means (for almost all reals) denoted K_p , of which the harmonic mean K_{-1} is an example, can be given representations:

$$(K_p)^p \log 2 = \sum_{n=2}^{\infty} Q_{np}(\zeta(n+|p|)-1),$$

where all Q coefficients have been given explicit rational form [11]. Again there is a peeled form, and the harmonic mean K_{-1} in particular is now known, via such machinations, to more than 7000 decimal places [11].

Beyond the evident beauty of the world of ζ -expansions, there are important computational questions partially addressed by such high-precision efforts. For example, is the geometric mean of the partial quotients in the simple continued fraction for K_0 equal to K_0 ? The various formulae of [11] relevant to the Khintchine constant and its relatives depend in general on all integer arguments *n* for $\zeta(n)$, not just the even ones. For such reasons, rapid evaluation schemes – including recycling ones – for positive integer *n* are always of special interest.

Here is another example of the utility of the series forms of our present interest. The classical acceleration formula [79](4.28)

$$\frac{\operatorname{Cl}_2(\theta)}{\theta} = 1 - \log|\theta| + \sum_{n=1}^{\infty} \frac{\zeta(2n)}{n(2n+1)} \left(\frac{\theta}{2\pi}\right)^{2n}, \quad |\theta| < 2\pi$$

for the Clausen function

$$\operatorname{Cl}_2(\theta) = \sum_{n=1}^{\infty} \frac{\sin(n\theta)}{n^2}, \quad \theta \text{ real}$$

is useful for computing certain Dirichlet *L*-series values, e.g., $Cl_2(\pi/2) = G$, the Catalan constant. For actual computations an accelerated, "peeled" form

$$\frac{\operatorname{Cl}_{2}(\theta)}{\theta} = 3 - \log\left(\left|\theta\right| \left(1 - \frac{\theta^{2}}{4\pi^{2}}\right)\right) - \frac{2\pi}{\theta} \log\left(\frac{2\pi + \theta}{2\pi - \theta}\right) + \sum_{n=1}^{\infty} \frac{\zeta(2n) - 1}{n(2n+1)} \left(\frac{\theta}{2\pi}\right)^{2n}$$

could be used.

We next describe one way to generate a vast collection of examples of rational ζ -series, by establishing a certain connection with *Laplace transforms*. Observe the following formal manipulations, where we disregard for the moment issues of convergence and summation interchange. Let μ be a fixed complex number and let f be the exponential generating series of the (presumed rational) sequence f_0, f_1, \ldots

$$f(x) = \sum_{n=0}^{\infty} \frac{f_n}{n!} x^n.$$
 (52)

Proceeding formally, we derive

$$\int_{0}^{\infty} f(x/a) e^{-\mu x} dx = \int_{0}^{\infty} f(x/a) e^{-\mu x} (e^{x} - 1)^{-1} \sum_{k=1}^{\infty} x^{k}/k! dx$$
$$= \sum_{n=0}^{\infty} \sum_{k=1}^{\infty} \frac{a^{-n} f_{n}}{k! n!} \int_{0}^{\infty} \frac{e^{-\mu x} x^{n+k}}{e^{x} - 1} dx.$$
(53)

Now, we invoke the integral representation (14) for the Hurwitz zeta function to arrive at the formal Laplace transform

$$\int_0^\infty f(x/a) e^{-\mu x} dx = \sum_{n=2}^\infty \zeta(n, \mu+1) \sum_{k=0}^{n-2} \binom{n-1}{k} \frac{f_k}{a^k},$$

where a is so far arbitrary, but eventually to be constrained by convergence requirements. Up to this point μ is likewise unrestricted; if we specify $\mu = 1$ and assume the coefficients f_n be rational, then we have a formal relation

$$\int_0^\infty f(x) \operatorname{e}^{-ax} \mathrm{d}x = \sum_{n=2}^\infty q_n(\zeta(n) - 1),$$

where the q_n are explicit and rational:

$$q_n = \sum_{k=0}^{n-2} \binom{n-1}{k} \frac{f_k}{a^{k+1}}.$$

The recreational possibilities of the Laplace transform approach seem endless. One may use a *Bessel function* of the first kind, $f(x) = J_0(x) = 1 - (x^2/4)/(1!)^2 + (x^2/4)^2/(2!)^2 - \cdots$, whose Laplace transform is known

$$\int_0^\infty J_0(x) e^{-ax} \, \mathrm{d}x = (1+a^2)^{-1/2}$$

267

to obtain (again, merely formally as yet)

$$\frac{1}{\sqrt{1+b}} = \sum_{n=2}^{\infty} \left(\zeta(n) - 1\right) \sum_{k=0}^{n/2-1} \left(-b/4\right)^k \binom{2k}{k} \binom{n-1}{2k},\tag{54}$$

which already shows the importance of convergence considerations; evidently |b| must be sufficiently small; certainly |b| < 2 suffices. Now observe that for integers μ , ν a square root of μ/ν may be written

$$\sqrt{\frac{\mu}{\nu}} = \frac{1}{\sqrt{1 + (\nu/\mu - 1)}}$$

if $\mu > v$, otherwise we use $(\mu/v)\sqrt{v/\mu}$, and so the ζ -series (54) applies with

$$b = \min(\mu, \nu) / \max(\mu, \nu) - 1$$

to yield a series for \sqrt{q} for any rational q and therefore any quadratic surd. Along these lines one may establish infinitely many different rational ζ -series for the golden mean, $\tau = (1 + \sqrt{5})/2$. For example, setting b = 1/465124, for which $\sqrt{1+b} \in Q[\tau]$, results in just one explicit series.

To represent π as a rational ζ -series, one may use the integral

$$\int_0^\infty \frac{\mathrm{e}^{-x}\sin x}{x}\,\mathrm{d}x = \frac{\pi}{4}$$

to obtain the series

$$\frac{\pi}{4} = \sum_{n=2}^{\infty} (\zeta(n) - 1) \sum_{k=0}^{n/2-1} \frac{(-1)^k}{2k+1} \binom{n-1}{2k}$$
$$= \sum_{n=2}^{\infty} n^{-1} (\zeta(n) - 1) \mathfrak{F} ((1+i)^n - 1 - i^n), \tag{55}$$

where interestingly enough the coefficients q_n vanish for n = 4, 8, 12, ... This rational ζ -series for π , like the form (50) and the aforementioned scheme for π^{odd} , is nontrivial in the sense that, whereas π^{2n} , being a rational multiple of $\zeta(2n)$, is trivially representable, odd powers of π evidently require some nontrivial analysis.

We have intimated that logarithms of rationals can always be given an explicit ζ -series. One may show this by invoking the Laplace transform:

$$\int_0^\infty \frac{e^{-x}(1 - e^{-ax})}{x} \, \mathrm{d}x = \log(1 + a)$$

to infer

$$\log(1-a) = \sum_{n=2}^{\infty} n^{-1} (\zeta(n) - 1)(1 + a^n - (1+a)^n).$$

Though this series has a finite domain of convergence, one may forge a series for $\log N$ for any integer $N \ge 2$ by using $\log N = -\log(1 + (1/N - 1))$. Thus $\log M/N$ for any integers M, N can be cast as a rational ζ -series. And the story by no means ends here. One may take

$$f(x) = \frac{\sinh\sqrt{x}}{\sqrt{x}}$$

to obtain a series for the error function at rational points z. (More precisely, one obtains a series for $\sqrt{\pi} \exp(z^2) \operatorname{erf}(z)$.) As the error function is essentially an incomplete gamma function, there is the possibility of casting more general incomplete gammas in rational ζ -series.

There is the intriguing possibility that one may effect numerical integration for some Laplacetransform integrands by way of appropriate rational ζ -series. There is also the possibility of discovering new identities by inversion; that is, one may work the Laplace transform technique backwards, to observe (let us say formally, as before):

$$\sum_{n=2}^{\infty} q_n(\zeta(n)-1) = \sum_{k=0}^{\infty} f_k,$$

where the f_k are defined via the recurrence

$$(k+1)f_k = q_{k+2} - \sum_{j=0}^{k-1} {\binom{k+1}{j}} f_j.$$

A different – and elegant – integral transform technique was enunciated by Adamchik and Srivastava [1], in the following form to which our Laplace-transform method stands as a kind of complement. Working formally as before, one can quickly derive from representation (14) a general relation

$$\sum_{n=2}^{\infty} q_n(\zeta(n) - 1) = \int_0^\infty \frac{F(t)e^{-t}}{e^t - 1} dt,$$
(56)

where

$$F(t) = \sum_{n=1}^{\infty} q_{n+1} \frac{t^n}{n!}.$$

As with our Laplace-transform technique, when one can do the integral one obtains a rational ζ -series. Adamchik and Srivastava went on to derive in this fashion such attractive series as

$$\sum_{n=1}^{\infty} n^{-1} t^n (\zeta(2n) - 1) = \log ((1 - t) \pi \sqrt{t} \csc(\pi \sqrt{t}))$$

which can also be derived by integration of relation (40); and the following curiosity which involves a derivative of ζ :

$$\sum_{n=2}^{\infty} \frac{\zeta(n)-1}{(n+1)(n+2)} = -\frac{1}{6}(1+\gamma) - 2\zeta'(-1).$$

Adamchik and Srivastava also employed their ζ -summation methods together with a certain polylogarithm series from [11] to derive an alternative representation for the Khintchine constant:

$$(\log K_0)(\log 2) = \frac{1}{12}\pi^2 + \frac{1}{2}\log^2 2 + \int_0^{\pi} t^{-1}\log(t|\cot t|)\,\mathrm{d}t$$

This kind of analysis shows again that a rational ζ -series can enjoy, quite beyond its natural allure, some theoretical importance. Incidentally, when a successful Laplace-transform kernel is used in the Adamchik–Srivastava formalism, the effects can be appealing. As just one example, if we use a Bessel kernel not as in the previous quadratic-surd analysis, but for *F* in relation (56), the result is

a convergent scheme for certain sums $\sum_{m} (x^2 + m^2)^{-s}$, which can in turn be summed over x to yield such as

$$\sum_{N=1}^{\infty} \frac{d(N)}{(1+N^2)^{3/2}} = \sum_{n=1}^{\infty} 4^{-n} (\zeta(2n+1)-1) 2n \zeta(2n+1) (-1)^{n-1} {\binom{2n}{n}} = 0.197785480715675063088236301582 \dots,$$

where d(N) is the number of divisors of N. Though the relevant coefficients this time are not rational (as they involve the $\zeta(2n+1)$ themselves), the indicated numerical value would evidently be difficult to achieve without the aid of such manifest convergence.

Because the series of choice for practical calculation of some constants (such as the Khintchine constant as just one example) is some form of rational ζ -series, we are interested in ζ -evaluations for integer arguments, to which subject we next turn.

5. Integer arguments

Because of existing fast algorithms for computation of π and its powers in (22), not to mention finite recurrences between the ζ -values at even positive integer arguments, computations for positive odd integer arguments are relatively more difficult.

Our first observation is that various of the formulae of previous sections may be applied directly when s is a positive odd integer. As just one example, the free-parameter choice $\lambda = i\pi$ in (32), together with recursion relation (31), gives rise to an interrelation between the ζ -values at odd positive integer arguments in the following form. Let m be a positive integer. With s = 2m + 1, we obtain

$$-\frac{(1-2^{-2m-1})2\zeta(2m+1)}{(\pi i)^{2m}} = \sum_{k=1}^{m-1} \frac{(1-4^{-k})\zeta(2k+1)}{(\pi i)^{2k}(2m-2k)!} + \frac{1}{(2m)!} \left\{ \log 2 - \frac{1}{2m} + \sum_{n=1}^{\infty} \frac{\zeta(2n)}{4^n(n+m)} \right\}.$$
(57)

When m = 1, the formula yields

$$\zeta(3) = \frac{2\pi^2}{7} \left\{ \log 2 - \frac{1}{2} + \sum_{n=1}^{\infty} \frac{\zeta(2n)}{4^n(n+1)} \right\},\,$$

which can be peeled once to give relation (46). Such as $\zeta(5)$ could be obtained in terms of $\zeta(3)$ and a convergent series, and so on. It is interesting that the weight factor 1/(2m)! of the troublesome series part decays so rapidly; that is, we have for large *m* an "almost exact" interrelation between the relevant $\zeta(\text{odd})$, in the spirit of, say, the even-argument relation (27).

From the Lerch expansion (33) one can derive other interrelations amongst ζ evaluations. Using the functional equation (17) we can write, for example:

$$\frac{\frac{3}{4}\zeta(3)}{=} \frac{1}{12}(1+\pi^2) - \frac{1}{2}\log 2 + \sum_{n=1}^{\infty} \frac{(-1)^{n-1}}{n^3 \mathrm{e}^n} - 2\sum_{j=1}^{\infty} \left(\frac{\mathrm{i}}{2\pi}\right)^{2j} \frac{(1-4^j)\zeta(2j)}{2j(2j+1)(2j+2)},$$

where the last sum on the right has purely rational summands decaying as π^{-2j} .

There are other similar series for $\zeta(\text{odd})$, for example that of Boo [21]:

$$\zeta(3) = -\frac{4\pi^2}{7} \sum_{n=0}^{\infty} \frac{\zeta(2n)}{(2n+1)(2n+2)4^n}$$

and of Williams [110]:

$$\zeta(3) = -2\pi^2 \sum_{n=0}^{\infty} \frac{\zeta(2n)}{(2n+2)(2n+3)4^n}.$$

Specific hyperbolic series, to be chosen as odd positive integer s is $1, -1 \pmod{4}$, respectively, are essentially due to Ramanujan and ala Zagier run as follows:

$$p\zeta(4p+1) = \frac{1}{\pi} \sum_{n=0}^{2p+1} (-1)^n (n - \frac{1}{2}) \zeta(2n) \zeta(4p+2-2n) -2 \sum_{n>0} \frac{n^{-4p-1}}{\exp(2\pi n) - 1} \left(p + \frac{\pi n}{1 - \exp(-2\pi n)} \right),$$
(58)

$$\zeta(4p-1) = -\frac{1}{\pi} \sum_{n=0}^{2p} (-1)^n \zeta(2n) \zeta(4p-2n) - 2 \sum_{n>0} \frac{n^{-4p+1}}{\exp(2\pi n) - 1}.$$
(59)

For p = 0, (58) evaluates to $\frac{1}{4}$ and (59) to $-\frac{1}{12} = \zeta(-1)$, as might be hoped. Note that there is no longer an infinite set of ζ -values required; the sums involving ζ are finite in (58) and (59). Moreover, while these require evaluation of $e^{2\pi k}$, the number e^{π} can be computed once and recycled.

Recently, similar but demonstrably different series have been found (the first few cases empirically by Simon Plouffe). A most striking example – which can be obtained, ex post facto, from [16, Chapter 14, Entry 21(i)] – in implicit form is

$$(2 - (-4)^{-n}) \left(2 \sum_{k=1}^{\infty} \frac{1}{(e^{2k\pi} - 1)k^{4n+1}} + \zeta(4n+1) \right)$$
$$-(-4)^{-2n} \left(-2 \sum_{k=1}^{\infty} \frac{1}{(e^{2k\pi} + 1)k^{4n+1}} + \zeta(4n+1) \right)$$
$$= \pi^{4n+1} \sum_{k=0}^{2n+1} (-1)^{k+1} (4^k + (-1)^{k(k-1)/2} (-4)^n 2^k) \frac{B_{4n+2-2k}}{(4n+2-2k)!} \frac{B_{2k}}{(2k)!}, \tag{60}$$

in which Bernoulli numbers can be replaced by even ζ -values using (24); and whose first case yields:

$$\zeta(5) = -\frac{72}{35} \sum_{k=1}^{\infty} \frac{1}{(e^{2k\pi} - 1)k^5} - \frac{2}{35} \sum_{k=1}^{\infty} \frac{1}{(e^{2k\pi} + 1)k^5} + \frac{\pi^5}{294}$$

A classical formula

$$\zeta(2) = 3\sum_{k=1}^{\infty} \frac{1}{k^2 \binom{2k}{k}}$$

270

has analog

$$\zeta(4) = \frac{36}{17} \sum_{k=1}^{\infty} \frac{1}{k^4 \binom{2k}{k}}$$

given in [41, p. 90], or see [22]. The next formula – which has no known single-term analogue yielding $\zeta(5)$ – played a signal role in Apéry's proof of the irrationality of $\zeta(3)$. (These matters are discussed further in [25,24].) The precise formula, due to Hjortnaes [61] is

$$\zeta(3) = \frac{5}{2} \sum_{k=1}^{\infty} \frac{(-1)^{k+1}}{k^3 \binom{2k}{k}}.$$
(61)

There is however a two-term analogue yielding $\zeta(5)$, namely this due to Koecher [74,73]

$$\zeta(5) = 2\sum_{k=1}^{\infty} \frac{(-1)^{k+1}}{k^5 \binom{2k}{k}} - \frac{5}{2} \sum_{k=1}^{\infty} \frac{(-1)^{k+1}}{k^3 \binom{2k}{k}} \sum_{j=1}^{k-1} \frac{1}{j^2}$$
(62)

and, more generally we have the following formal expansion in powers of z:

$$\sum_{k=1}^{\infty} \frac{1}{k^3(1-z^2/k^2)} = \sum_{k=1}^{\infty} \frac{(-1)^{k+1}}{k^3\binom{2k}{k}} \left(\frac{1}{2} + \frac{2}{1-z^2/k^2}\right) \prod_{j=1}^{k-1} (1-z^2/j^2).$$

Borwein-Bradley [25,24,4] established

$$\zeta(7) = \frac{5}{2} \sum_{k=1}^{\infty} \frac{(-1)^{k+1}}{k^7 \binom{2k}{k}} + \frac{25}{2} \sum_{k=1}^{\infty} \frac{(-1)^{k+1}}{k^3 \binom{2k}{k}} \sum_{j=1}^{k-1} \frac{1}{j^4}$$
(63)

and more generally the power series in z:

$$\sum_{k=1}^{\infty} \frac{1}{k^3 (1 - z^4/k^4)} = \frac{5}{2} \sum_{k=1}^{\infty} \frac{(-1)^{k+1}}{k^3 \binom{2k}{k}} \frac{1}{1 - z^4/k^4} \prod_{j=1}^{k-1} \frac{j^4 + 4z^4}{j^4 - z^4}.$$
(64)

Note that (64) contains (61) and (63) as its constant term and next term (coefficient of z^4), respectively. Formula (64) was discovered empirically and reduced in [25] to an equivalent finite form by a process of "creative telescoping" and analytic continuation. This finite form was subsequently proven by Almkvist and Granville. Formulae (61)–(63) are well suited for numerical computation due to the fact that the series terms decay roughly geometrically with ratio $\frac{1}{4}$. Algorithms 1, 2 and 3 below are based on the Hjortnaes formula (61), the Koecher formula (62), and the Borwein–Bradley formula (63), respectively [25].

Algorithm 1. Given D, compute $\zeta(3)$ to D digits using (61). Computations are performed to D digits.

```
N = 1 + \lfloor 5D/3 \rfloor; c = 2; s = 0;
for n = 1 to N do begin
s = s + (-1)^{n+1}/(n^3c);
c = c(4n + 2)/(n + 1);
end;
return 5s/2;
```

Note that this kind of algorithm can be naturally extended to yet more efficient $\zeta(3)$ series, such as the accelerated formula (68) appearing later in this paper.

Algorithm 2. Given D, compute $\zeta(5)$ to D digits using (63). Computations are performed to D digits.

```
N = 1 + \lfloor 5D/3 \rfloor; a = 0; c = 2; s = 0;
for n = 1 to N do begin
g = 1/n^2; s = s + (-1)^{n+1}(4n - 5a)/(n^3c);
c = c(4n + 2)/(n + 1); a = a + g;
end;
return s/2;
```

Algorithm 3. Given D, compute $\zeta(7)$ to D digits using (63). Computations are performed to D digits.

 $N = 1 + \lfloor 5D/3 \rfloor; a = 0; c = 2; s = 0;$ for n = 1 to N do begin $g = 1/n^2; s = s + (-1)^{n+1}(5a + g)/(n^3c);$ c = c(4n + 2)/(n + 1); a = a + g;end; return 5s/2;

The operational complexity of Algorithms 1–3 will be discussed in Section 7. Generally speaking, for fixed precision (say D digits) these are the fastest schemes available for the indicated ζ (integer) values. One should keep in mind that there are asymptotically (very large D) even faster ways of handling the relevant summations, using a so-called FEE method also discussed in Section 7.

6. Value recycling

We have mentioned the multivalue computations of Odlyzko and Schönhage [87], such an approach being of interest for complex *s* lying, say, in some (complex) arithmetic progression. It turns out that for certain sets of arguments with integer differences (the arguments not necessarily in arithmetic progression) one can invoke alternative value-recycling schemes. The basic notion of recycling here

273

is that previously calculated ζ -values – or initialization tables of those calculations – are re-used to aid in the extraction of other ζ -values, or many ζ -values at once are somehow simultaneously determined, and so on. So by value recycling we mean that somehow the computation of a collection of ζ -values is more efficient than would be the establishment of independent values.

First, one can use either of (30) or (32) to efficiently evaluate ζ at each of N arguments $\{s, s+2, s+4, \ldots, s+2(N-1)\}$ for any complex s. This approach might be fruitful for obtaining a collection of ζ -values at odd positive integers, for example. The idea is to exploit the recursion relation (31) for the incomplete gamma function and thereby, when N is sufficiently large, effectively unburden ourselves of the incomplete gamma evaluations. One may evaluate such as $\Gamma(\{s/2\}, x), \Gamma(\{(1-s)/2\}, x)$ where $\{z\}$ denotes generally the fractional part of z, over a collection of x-values, then use the above recursion either backward or forward to rapidly evaluate series terms for the whole set of desired ζ -values. Given the initial $\Gamma(\{s/2\}, x)$ evaluations, the rest of the calculation to get all the ζ -values is sharply reduced. In the case that the $\{s + 2k\}$ are odd integers, the precomputations involve only $\Gamma(0,x)$ and $\Gamma(1/2,x)$ values; known classically as exponential-integral and error-function values. Ref. [45] contains explicit pseudocode for a recycling evaluation of $\zeta(3), \zeta(5), \ldots, \zeta(L)$ via series (30), in which evaluation one initializes error function and exponential-integral values, respectively:

$$\{\Gamma(\frac{1}{2},\pi n^2): n \in [1,\lfloor D \rfloor]\},\tag{65}$$

$$\{\Gamma(0,\pi n^2): n \in [1,\lfloor D \rfloor]\},$$

where D decimal digits of precision are ultimately desired for each ζ value. The notion of "recycling" takes its purest form in this method, for the incomplete-gamma evaluations above are reused for every ζ (odd).

A second recycling approach, relevant for *even* integer arguments, involves a method of series inversion pioneered by J. P. Buhler for numerical analyses on Fermat's "Last Theorem" and on the Vandiver conjecture [36–38]. One uses a generating function for Bernoulli numbers, and invokes the Newton method for series inversion of the key elementary function. To get values at even positive integers, one may use an expansion related to (25). One has

$$\frac{\sinh(2\pi\sqrt{t})}{4\pi\sqrt{t}} \, \frac{2\pi^2 t}{\cosh(2\pi\sqrt{t}) - 1} = -\sum_{n=0}^{\infty} (-1)^n \zeta(2n) t^n,$$

which we have derived and written in this particular form to allow the algorithm following. Note that we have separated the left-hand side into two series-dependent factors, each in the *t* variable: one series being essentially of the form $(\sinh\sqrt{z})/\sqrt{z}$ and the other being $(\cosh\sqrt{z}-1)/z$. The idea, then, is to invert the latter series via a fast polynomial inversion algorithm (Newton method). Using *t* as a place-holder throughout, one then reads off the ζ -values as coefficients in a final polynomial. In the algorithm display following, we assume that $\zeta(2), \zeta(4), \ldots, \zeta(2N-2)$ are desired. The polynomial arithmetic is most efficient when truncation of large polynomials occurs at the right junctures. For a polynomial q(t), we denote by $q(t) \mod t^k$ the truncation of *q* through the power t^{k-1} inclusive; that is, terms t^k and beyond are dropped. Also in what follows, a polynomial multiplication operation is signified by "*".

Algorithm 4. Recycling scheme for a sequence $\zeta(0), \zeta(2), \zeta(4), \dots, \zeta(2(N-1))$. [Denominator setup] (1)Create the polynomial $f(t) = (\cosh(2\pi\sqrt{t}) - 1)/(2\pi^2 t)$, through degree N (i.e., through power t^N inclusive); [Newton polynomial inversion, to obtain $q := f^{-1}$] (2)p = q = 1;while $(p < \deg(f))$ do begin $p = \max(2p, \deg(f));$ $h = f \mod t^p$; $q = (q + q * (1 - h * q)) \mod t^{p}$; end: [Numerator setup] (3) Create the polynomial $k(t) = \sinh(2\pi\sqrt{t})/(4\pi\sqrt{t})$, through degree N; $g = g * k \mod t^{2N-1};$ For $n \in [0, 2N - 2]$, read off $\zeta(2n)$ as $-(-1)^n$ times the coefficient of t^n in polynomial q(t).

It is important to note that this algorithm can be effected in *either* numerical or symbolic mode. That is, in step (1) the polynomial in question can have floating point coefficients, or symbolic ones with their respective powers of π and so on. If symbolic mode is in force, the ζ values of the indicated finite set are all exact, through $\zeta(2N-2)$ inclusive. The method has actually been used – in numerical mode so that fast Fourier transform methods may also be applied to the numerical multiplications – to calculate the relevant ζ -values for high-precision values of the Khintchine constant [11]. Incidentally, if one worries about memory storage in such a Buhler inversion, there is a powerful technique called "multisectioning", whereby one calculates all the $\zeta(2k)$ for k lying in some congruence class (mod 4, 8 or 16 say), using limited memory for that calculation, then moving on to the next congruence class, and so on. Observe first that, by looking only at even-indexed Bernoulli numbers in the previous algorithm, we have effectively multisectioned by 2 already. To go further and multisection by 4, one may observe

$$\frac{x\cosh x \sin x \pm x \cos x \sinh x}{\sinh x \sin x} = 2\sum_{n \in S^{\pm}} \frac{B_n}{n!} (2x)^n,$$

where the sectioned sets are $S^+ = \{0, 4, 8, 12, ...\}$ and $S^- = \{2, 6, 10, 14, ...\}$. The key is that the denominator $(\sinh x \sin x)$ is, perhaps surprisingly, x^2 times a series in x^4 , namely we have the attractive series

$$\sinh x \sin x = \sum_{n \in S^{-}} (-1)^{(n-2)/4} 2^{n/2} \frac{x^n}{n!},$$
(66)

so that the key Newton inversion of a polynomial approximant to said denominator only has *one-fourth* the terms that would accrue with the standard Bernoulli denominator $(e^x - 1)$ (and one-half as many terms as required in Algorithm 5). Thus, reduced memory is used to establish a congruence class of Bernoulli indices, then that memory is reused for the next congruence class, and so on. Thus, these methods function well in either parallel or serial environments.

Multisectioning was used by Buhler and colleagues – as high as level-16 sections – to verify Fermat's "Last Theorem" to exponent 8 million [37]. They desired Bernoulli numbers modulo primes, and so employed integer arithmetic, but the basic Newton iteration is the same for either symbolic (rational multiples of powers of π) or numerical (floating-point) ζ -values.

A third approach is to contemplate continued fraction representations that yield ζ -values. For example, the well-known fraction for $\sqrt{z} \coth \sqrt{z}$ gives

$$\frac{\pi^2 z}{3 + \frac{\pi^2 z}{5 + \frac{\pi^2 z}{7 + \frac{\pi$$

The computational advantage here would obtain if one already had in hand an efficient, continued fraction engine. There is also the possibility of fast evaluation of the convergents, although it is unclear whether this technique could be brought to the efficiency of the Buhler approach above. Incidentally, if one desires not values at the even positive integers but the actual Bernoulli numbers as exact rational numbers, there is an alternative fraction due to Bender:

$$\frac{1}{1 + \frac{b(1)z}{1 + \frac{b(2)z}{1 + \frac{b(3)z}{1 + \cdots}}}} = 1 + 6\sum_{n=2}^{\infty} B_{2n} z^{n-1}$$

with

$$b(n) = \frac{n(n+1)^2(n+2)}{4(2n+1)(2n+3)}.$$

Note that the series does not converge in any obvious sense; it is a symbolic series. Again, this form might be recommended if a good continued fraction calculator were in place. As a last alternative for fast evaluation at even positive integer arguments, there is an interesting approach of Plouffe and Fee [90], in which the Von-Staudt–Clausen formula for the fractional part of B_n is invoked, then asymptotic techniques are used to ascertain the integer part. In this way the number B_{200000} has been calculated in exact, rational form. Yet another strategy for Bernoulli numbers – untested as far as we know – is to resolve B_n via Chinese remainder methods, where one would establish via Voronoi formulae the values $B_n \pmod{p_i}$ for sufficiently many small primes p_i .

A fourth approach stands as a kind of complement to the previous, even-argument method. There is actually a way to calculate ζ -values at consecutive positive integers in recycled fashion. Now, the generating function will not be a cotangent function but the ψ function defined in (26). Previous implementations of a ψ -based recycling algorithm, as in [45], do work but are not of the fast algorithm class. More recently [48], there has appeared an asymptotically "fast" rendition of the idea, which method we now briefly describe.

Since the standard gamma function can be estimated via such approximations as [22]

$$\left|\Gamma(z)-N^{z}\sum_{k=0}^{6N}\frac{(-1)^{k}N^{k}}{k!(k+z)}\right| \leq 2N\mathrm{e}^{-N},$$

valid for real $z \in [1, 2]$, one might expect that the kind of sum appearing would, if the series inversions of the (k+z) were carried out as polynomials in z, provide a reasonable series for the ψ function (the logarithmic derivative of Γ). Indeed, it turns out [48] that the logarithmic derivative of a function with summation limit 4N, namely

$$g(z) = \sum_{k=0}^{4N-1} \frac{(-1)^k N^k}{k!(k+1-z)}$$

is an appropriate power series in z, in fact

$$\frac{\mathrm{d}}{\mathrm{d}z}\log g(z)\sim (\log N+\gamma)+\zeta(2)z+\zeta(3)z^2+\zeta(4)z^3+\cdots,$$

in some appropriate asymptotic sense [48]. Thus, the same polynomial arithmetic ideas as for the Buhler method previous can be used in principle to evaluate ζ at consecutive positive integer arguments. The following algorithm display follows the treatment in [48]:

Algorithm 5. Recycling scheme for a collection of the L values: $\zeta(2), \zeta(3), \zeta(4), \dots, \zeta(L+1)$.

```
[Set precision]
(1)
         Choose a power-of-two N, such that 2^{-N} is less than
            the required precision (i.e., N is greater than the required bit-precision).
            and also N \ge L (a common condition in numerical settings).
(2)
      [Quotient array]
         Create g[k] = P[k]/Q[k], for k \in [0, 4N - 1], where
            P[k] = (-N)^k, \ Q[k] = k!(k+1-z),
            with z being a place-holder as in standard polynomial computations.
(3)
      [Resolve g function]
         p = 1;
         while p \leq 2n do begin
           for q = 0 to 4N - 1 - p step p do begin
               g[q] = g[q] + g[q + p];
               Resolve the new g[q] into numerator/denominator,
                 each clipped mod z^{L+1});
            end;
            p = 2p;
          end;
```

- (4) [Monic reversion] Now g[0] = P[0]/Q[0], each of P,Q being of degree at most L, so force a reverse-monic property, by dividing each of P,Q through by its constant coefficient;
 - [Inversion] Perform Newton inversions as in step (2) of Algorithm 4, to create the reciprocal polynomials P^{-1} and Q^{-1} ;
- (6) [Coefficient computation] Compute the coefficients R_k in the polynomial

$$R(z) = \sum_{k=0}^{2} R_k z^k = ((dP/dz)P^{-1} - (dQ/dz)Q^{-1}) \mod z^{L+1};$$

(7) [*Read off the* ζ values] *Optionally read off* $\gamma \sim R_0 - \log N$ and in any case read off, for $k \in [2, L+1]$, the desired ζ approximations as

$$\zeta(k) \sim R_{k-1}$$
.

(5)

A typical experiment with Algorithm 5 works out as follows. Take N=L=16, meaning that degree-16 polynomials will be used and we shall obtain in recycling fashion a set of 16 separate ζ values, together with an approximation to γ :

$$R(x) \sim \log N + 0.57721 + 1.64493x + 1.20205x^{2} + 1.08232x^{3} + 1.03692x^{4} + \dots + 1.000122713347x^{12} + \dots,$$

where we indicate good digits by virtue of their appearance. Note that $\zeta(13)$ as the coefficient of x^{12} is more accurate than the low-lying coefficients. This trend is universal to the algorithm, and in some ways is a good thing because if the values $\zeta(n) - 1$ are employed, we enjoy relative precision after the 1 is subtracted. Note also that even the low-lying coefficients have errors of order 2^{-16} as expected. Of course, the algorithm can be modified to yield only values at odd positive integers, for example by subtracting off at a key juncture a truncated cotangent series. Detailed error analysis and asymptotics are described in [48], though we do touch upon complexity issues for Algorithm 5 in the next section. It should also be observed that fast, single-argument evaluation of the gamma function and functions such as our g(z) were worked out by Karatsuba [66–69], about which we have more to say in the next section; so perhaps her methods may be used to accelerate even further the series computations of Algorithm 5.

7. Computational complexity

Herein we focus on evaluations of ζ -values for integer arguments and arguments in certain arithmetic progressions. However, in a spirit of completeness, let us first comment on the complexity issue for those analytic number theory computations briefly reviewed in Section 1. Consider first the highly important evaluation of $\zeta(1/2 + it)$ where t is positive but otherwise unrestricted; and say we desire the evaluation to have a fixed precision (one only needs enough precision actually to locate

zeros, say) but that t is unrestricted. It should be stated right off that for this problem there is no known polynomial-time algorithm, say an algorithm of $O(\log^k t)$ operation complexity to perform a *single* ζ evaluation. We note the interesting remarks in [19], where the author suggests outright that the calculation of $\zeta(\frac{1}{2} + it)$ is fundamentally of exponential operation complexity $O_{\varepsilon}(t^{1/2-o(1)})$ to achieve errors bounded by a fixed ε and furthermore that this is a special property of the critical line (indeed, off the critical line the complexity is reduced). Whereas it is known that the classical Euler–Maclaurin approach has operation complexity O(t), the Riemann–Siegel formula allows $O(t^{1/2+\varepsilon})$. Indeed, we recall that most of the work for the latter method is a sum over $O(\sqrt{t})$ elementary summands. Furthermore, the Odlyzko–Schönhage approach allows the (approximately $T^{1/2}$) critical zeros of the interval $t \in [T, T + T^{1/2})$ to be found in $O(T^{1/2+\varepsilon})$ operations [85–87]. So the average operation complexity per critical zero works out to be impressive: $O(T^{\varepsilon})$.

Riemann–Siegel formula (35), $\Re(s = \sigma + it) > 0$ fixed, t > 0 arbitrary, and precision fixed:

Operation complexity $O(t^{1/2+\varepsilon})$.

Odlyzko–Schönhage enhancement, for $t \in [T, T + T^{1/2}]$:

Operation complexity $O(T^{\varepsilon})$ per each of $O(T^{1/2+\varepsilon}) \zeta$ values.

Note that the Odlyzko–Schönhage method enjoys its tremendous efficiency because it is, in our present sense of the word, a recycling scheme. As Ref. [87] describes, the evaluation of multiple ordinates *t* simultaneously can be done via FFT-like methods, in particular rational-complex function evaluation which can also be considered as fast interpolation along the lines of the works of Dutt et al. [53] and Dutt and Rokhlin [54]. The essential idea is to attempt to perform sums of the form (36) for a set of *t* values (which may or may not be equispaced). Sometimes, depending on the problem at hand, a simple FFT approach with the Euler–Maclaurin formula (29) is a good option. For example, $\pi(x)$ calculations, for moderate *x*, carried out in the style described after relation (3) may benefit from such a simplified approach [47].

Incidentally, the Euler–Maclaurin series (29) for fixed precision and arbitrary t is not as good as the Riemann–Siegel series, in fact, Euler–Maclaurin formula (29), $\Re(s = \sigma + it) > 0$ fixed, t > 0 arbitrary, and precision fixed:

Operation complexity $O(t^{1+\varepsilon})$.

Incidentally, because the Euler–Maclaurin method also starts out with a sum of terms $n^{-\sigma-it}$, the Odlyzko–Schönhage acceleration applies equally well, with the ultimate complexity being reduced accordingly to $O(T^{1/2+\varepsilon})$ per zero for resolution of $O(T^{1/2+\varepsilon})$ zeros in $[T, T + T^{1/2}]$. Note also that the Bernoulli components of the Euler–Maclaurin sum can be obtained in recycled fashion, as we discuss below. Such methods can sometimes pull computationally (perhaps not always theoretically) important logarithmic factors off complexity bounds. There is a moral here: regardless of superior asymptotic behavior, the Riemann–Siegel formulae may sometimes involve too many practical details when Euler–Maclaurin, far simpler to implement, and susceptible to some interesting optimizations, would suffice. The Euler–Maclaurin scheme can be used, for example, in serious practical evaluations of ζ (see, for example, [40], where careful Euler–Maclaurin error bounds are developed).

Unlike the analysis for fixed σ and large t, most every other aspect of the present treatment involves the following scenario: the argument s, or arguments $\{s_1, s_2, \ldots\}$ (and their population) are fixed, and we consider varying the precision, measured say as D digits.

Let us start with the incomplete-gamma series (30). Because an incomplete gamma function can be evaluated via fast Fourier transform acceleration in $O(D^{1/2} \log^2 D)$ operations [22], and because we require $O(D^{1/2})$ summands of either sum, and because elementary functions (e.g. arbitrary powers) require $O(\log^k D)$ operations, for some k [30,31], we conclude: Incomplete-gamma formula (30), for fixed complex s, to D-digit precision:

Operation complexity $O(D^{1+\varepsilon})$.

Recycling enhancement to incomplete-gamma formula (based on precomputations (65)), for set of arguments $\{s, s+2, s+4, \dots, s+2(N-1)\}$:

Operation complexity $O(D^{1/2+\varepsilon})$ per ζ value.

This average complexity in recycling mode is impressive; we know of no simple schemes for say $\zeta(\text{odd})$ that run faster than O(D); however see the material later in this section for low bit-complexity schemes that exploit dynamically changing precision, such as Karatsuba's FEE method and possible hybrid alternatives that might stem from it.

Because the values at even positive integers appear in so many studies, we next discuss the Buhler recycling scheme, Algorithm 4. It is evident that the even-argument values $\zeta(2), \ldots, \zeta(2N)$ can all be obtained in $O(\log N)$ Newton iterations. However, these iterations can be done with dynamically increasing precision, so that the asymptotic complexity is dominated by that for the last Newton step: a single polynomial multiply for polynomials of degree O(N). One can achieve such by using a fast convolution algorithm for the polynomial multiplication, such as the Nussbaumer method [45], thus obtaining all the indicated ζ -values in $O(N \log N)$ operations. To summarize

Buhler recycling scheme, Algorithm 4, for $\zeta(0), \zeta(2), \zeta(4), \dots, \zeta(2N-2)$ each to D-digit precision:

Operation complexity $O(\log N)$ per ζ value.

This estimate now has implications for various formulae, such as the Bernoulli-based series (32) and the Euler-Maclaurin method (29), as both depend on the values at even positive integers.

As for the more general recycling scheme of Algorithm 5, the complexity analysis can be found in [48], the essential idea being that the recombination of polynomials in step (3) involves $N/2, N/4, N/8, \ldots$ pairwise polynomial-ratio combinations, respectively, on successive loop passes, and these are of growing degree, yet fast polynomial multiplication can be used, with the result that the complexity is $O(N\log^2 L)$ operations for the very construction of the g function as the ratio of two polynomials each of degree L. We conclude:

Psi-function recycling scheme, Algorithm 5, for $\gamma, \zeta(2), \zeta(3), \dots, \zeta(L+1)$ each to *D*-digit precision (with $L \sim D$ also):

Operation complexity $O(L^{-1}N \log^2 L)$ per each of the L evaluations of ζ .

Note that for $L \sim N$, equivalently: one desires about D different ζ values each to D digits, the average cost is $O(\log^2 D)$ per value. This is somewhat worse than the cost of Algorithm 4, but certainly falls into the "fast algorithm" category: for both algorithms we could say that "polynomial rate" is achieved, meaning polynomial time complexity $O(\log^k D)$ as a per-evaluation average.

Next, we look at the Euler-Maclaurin scheme. For precision 10^{-D} we can take $M = O(D/\log N)$ in the Bernoulli summation of series (29). But we have just estimated the operation complexity as $O(M \log M)$ for the generation of the relevant Bernoulli numbers. As general exponentiation

is $O(\log^k D)$ operations for some k [30,31], the work for first summation in the Euler-Maclaurin formula requires $O(N \log^k D)$ operations. Thus for any (fixed) complex s, we end up with operation complexity $O(N \log^k D) + O(D \log(D/\log N)/\log N)$, and we conclude:

Euler-Maclaurin formula (29), for s fixed, D-digit precision:

Operation complexity $O(D^{1+\varepsilon})$.

Of course, for integer s the Euler-Maclaurin method will – as with most other schemes – be somewhat more efficient.

For the Bernoulli series (32) to *D*-digit precision, we again apply the recycling of Buhler for $O(D/\log(1/\lambda))$ summands in the second sum, with $O(D/\lambda)$ summands in the first. This means we optimize the free parameter as: $\lambda \sim (\log \log D)/\log D$ and conclude:

Free-parameter formula (32), for s fixed, D-digit precision:

Operation complexity $O(D \log D / \log \log D)$.

This is of course also $O(D^{1+\epsilon})$, but the analysis is particularly straightforward for the free-parameter formula, so we exhibit the detailed complexity. Note that the asymptotic decay of the free parameter λ is consistent with the hard constraint on the allowed range $0 \le \lambda < 2\pi$. Incidentally the "peeled series" approach, whereby one peels terms from a rational- ζ series, is in complexity terms very similar to the free-parameter series. Writing

$$\sum_{n=2}^{\infty} q_n(\zeta(n)-1) = \sum_{m=2}^{M} \sum_{n=2}^{\infty} \frac{q_n}{m^s} + \sum_{n=2}^{\infty} q_n\zeta(n,M+1),$$
(67)

we see that if the last summation above is over $n \in [2, N]$ then for *D*-digit precision we require $N = O(D/\log M)$. If the (we presume closed-form) peeled terms are each of polynomial operation complexity, and we use recycling, we have overall cost $O(M \log^k D) + O(D \log D/\log M)$. If we set $M \sim D/\log^k D$ and $N \sim D/\log M$ we obtain:

General peeled-series form (67), for s fixed, D-digit precision:

Operation complexity $O(D^{1+\varepsilon})$.

Heretofore in this section we have concentrated on *operation* counts, whereby one takes each summand of a series to full precision. Also, *s* arguments have heretofore been general. But for certain series of our immediate interest, notably some old and new series for $\zeta(\text{odd})$, one can adroitly adjust precision so that very low *bit* complexity is achieved. Our first observation is that a modern series having rational summands, and exhibiting linear convergence can be evaluated to *D* good digits, for fixed integer argument *s*, in N = O(D) summands. Thus the operation complexity is simply:

Rational-summand series, such as (61) and many others, as in Algorithms 1–3, for D-digit precision:

Operation complexity O(D).

This is as good as any of the previous complexity estimates, except for the recycling cases (when the average, per-value complexity may be genuinely less than O(D)); furthermore the terms in the various series are generally simple in structure.

But now we wish momentarily to drop the notion of "operation complexity for D digits" and concentrate instead on *bit* complexity for, let us say, N-bit precision. In modern times there has been

281

a revolution of sorts in the matter of bit-complexity estimates for ζ evaluation, or for that matter the evaluation of more general series. The idea is to combine subseries of a given, well-convergent series in certain, efficient ways, employing recursion relations and other algebraic expedients cleverly. We shall refer to this as the fast *E*-function evaluation (FEE) method of Karatsuba. The algorithm has sometimes been called "binary splitting," which was foreshadowed in the works of Schönhage and Brent [30,31,97,22] for decimal-base conversion, calculation of fundamental constants and some elementary functions; yet was brought into powerful, general, and rigorous form by Karatsuba, resulting in unprecedented low bit complexity for hypergeometric series of algebraic parameters and argument (see [65–70], Ref. [71] being especially informative).²

One way to think of the FEE method is to imagine, in the words of [59], pushing "as much multiplication work as possible to the region where multiplication becomes efficient". The complexity of the FEE method, when said method applies, turns out to be

$$O(M(N)\log^2 N),$$

where M(N) is either the bit complexity of multiplying two integers each of N bits by grammar-school (naive, $O(N^2)$ means), or the bit complexity that is the lowest known. As for minimal-complexity multiplication, the celebrated Schönhage–Strassen bit-complexity bound, namely [97]

 $M(N) = \mathcal{O}(N \log N \log \log N),$

thus yields a bit complexity for the FEE method in the form

$$O(N \log^3 N \log \log N)$$

for evaluation of appropriate series to *N*-bit precision, which bound can be thought of as $O(N^{1+\varepsilon})$ and thus "near-optimal"; and we remind ourselves that this bound thus applies to a very wide class of series.³ In this class are computations of certain constants such as ζ -values at odd positive integers, Euler's constant γ , powers e^x for bounded *x*, and generally to series whose *k*th terms are rational, possessed of $O(\log k)$ bits in numerator and denominator; and yet more generally to hypergeometric series with suitably bounded algebraic argument and parameters [71].

It should be remarked right off that the FEE method gives no gain whatsoever – over direct summation – if standard, grammar-school multiplication (of bit-complexity O(NN')) for two respective N, N'-bit operands) be used. To see this, consider a typical series to which the FEE method applies:

$$S = \sum_{n=0}^{\infty} \frac{a(n)}{b(n)} \prod_{j=0}^{n} \frac{p(j)}{q(j)},$$

where each of a, b, p, q is an integer-valued function of $O(\log n)$ bits, and assume (as is typically required for the FEE method) that a truncation error bound of 2^{-N} , for N-bit precision, obtains after O(N) terms of the series. It is not hard to see that if each term be evaluated to N bits, we require under grammar-school multiplication $O(N \log j)$ bit operations per term, so that the summation of the required N terms has bit complexity $O(N^2 \log N)$. Thus if the grammar-school bound is used

² There is also a succinct and accessible modern treatment of such technique, by Haible and Papanikolaou [59], yet those authors unfortunately were unaware of the original works of Karatsuba. For reasons of scholarship therefore, we choose to refer to the general series-manipulation paradigm in question as the FEE method.

³ Incidentally, there is another multiplication algorithm enjoying the same bit complexity as Schönhage–Strassen; we speak of Nussbaumer convolution which is at least as easy to implement, as described in say [45,51].

with FEE, the bit complexity is $O(M(N)\log^2 N) = O(N^2 \log^2 N)$ which amounts to no gain over conventional summation.

For the present compendium we have carefully chosen an illustrative FEE example. It is neither the simplest (perhaps the calculation of e or some such constant would qualify for that), nor is it the most recondite (one can even apply FEE to special functions of applied science, such as Bessel functions and so on). But the example shows the essential ingredients of FEE, and intentionally moves a little away from the above S form to underscore the possibility of algebraic-irrational arguments. Consider the polylogarithm evaluation

$$L = \text{Li}_{3}(\tau^{-2}) = \sum_{n=1}^{\infty} \frac{\tau^{-2n}}{n^{3}},$$

where $\tau = (1 + \sqrt{5})/2$ is the (patently algebraic) golden mean. This *L* constant is especially interesting because knowing it is essentially to know $\zeta(3)$, as we see discuss in Section 8. Now if we truncate the *L* series through the (n = N)th term inclusive, we have at least *N*-bit precision, so let us for algorithmic convenience choose some suitable $N = 2^k$, and note first that

$$\sum_{n=1}^{N} \frac{\tau^{-2n}}{n^3} = \sum_{m=1}^{N/2} \tau^{2m-1} \frac{8m^3 + (8m^3 - 12m^2 + 6m - 1)\tau}{8m^3(2m-1)^3},$$

where we have pairwise combined terms from the left-hand sum, to forge a half-length sum with, in Karatsuba's words, "obvious denominators". Likewise, the right-hand sum can be pairwise processed to forge a yet shorter sum:

$$\sum_{p=1}^{N/4} \tau^{4p-3} \frac{A+B\tau}{C},$$

where A, B, C are more complicated polynomials in the index p, e.g., A, B now have degree 6. In general one obtains, as Karatsuba showed, a recurrence relation for ever more complicated numerator terms. In our case, one must use the quadratic reduction $\tau^2 = 1 + \tau$ to keep all numerators in $Z[\tau]$. Upon detailed analysis of the work to perform the pairwise combinations and so on, one finds that the bit complexity to perform $k = \lg N$ such series contractions – which work yields just *one* final term, a singleton summation – is a sum:

$$\sum_{j=1}^{k} \frac{N}{2^{j}} M(\mathcal{O}(2^{j} \log N)) = \mathcal{O}(M(N) \log^{2} N)$$

for either grammar-school or minimal-complexity multiply, as claimed.

Let us perform an example of evaluation of the L constant through N = 16 summands. This means that we carry out four recursion levels, obtaining on the first level eight terms:

$$\left\{\tau \ \frac{8+\tau}{8}, \dots, \tau^{15} \ \frac{4096+3375\tau}{13824000}\right\},\,$$

the sum over which eight being exactly the original 16-fold sum intended. At recursion bottom we end up with a solitary term, namely

$$L \sim \frac{842439095385706230219 - 376615379847138777145\sqrt{5}}{748737728234496000} \sim 0.4026839629\ldots,$$

where we have taken the liberty to cast the result in the form of a surd $(a + b\sqrt{5})/c$. The numerical value is incidentally correct to the 10 places shown. To convey an idea of the efficiency of the method, we note that for N = 32 summands and so five recursion levels, the numerical value of the solitary surd is correct to 17 decimals, which makes sense because to jump from 16 to 32 summands we only have to do a little more than twice the multiplication work.

It is especially intriguing that the final result of such FEE processing is not only a single term, but an *exact* term in the sense that it could be used later in a truncated series of twice the length, i.e., the single term in hand can act as the left-hand term of a one-higher recursion level. Likewise, FEE is a *parallel* method in that separate processors can, in principle, handle separate pairings of terms at any recursion level.

We have merely sketched the technique in brief fashion. For the rigorous details of such applications of FEE, a good reference is [68], where the celebrated formula (61) for $\zeta(3)$ is used to establish the $O(M(N)\log^2 N)$ bit complexity for *N*-bit precision; and therein of course the numerator recursions are of pure-integer form.

As an example application of such techniques for very-high-precision work, in [59] the identity of Amdeberhan and Zeilberger [5]:

$$\zeta(3) = \frac{1}{2} \sum_{m=1}^{\infty} \frac{(-1)^{m-1} (205m^2 - 160m + 32)}{m^5 {\binom{2m}{m}}^5}$$
(68)

is noted, together with the (S-series) assignments: $a(n) = 205n^2 + 250n + 77$, b(n) = 1, p(0) = 1, $p(n) = -n^5$ for positive n, and $q(n) = 32(2n + 1)^5$.

In spite of Karatsuba's FEE and its wide applicability, there remain some interesting open questions. For example, note that one can, in principle, use FEE recursion, but *symbolically*, in the following sense. One recurses down only "half way", to render an original sum of N terms as a new sum of $O(\sqrt{N})$ terms, each new term now being rational polynomial with each numerator and denominator having say $O(\sqrt{N})$ degree with integer coefficients. (In our above example for the L constant, just one level yields a degree-3 numerator, and we are saying one would continue the construction of higher-degree numerators but only to a certain depth.) Now it is known that a degree-d polynomial can be evaluated at O(d) points in $O(d \log^2 d)$ operations with fast algorithms [45], so perhaps there is a compromise in some cases, between full FEE recursion and a mixed, symbolic-FEE-polynomial scheme. At the very least, these considerations lead out of the bit-complexity paradigm into a world in which $O(D^{1/2+\epsilon})$ operation complexity – meaning full-precision operations for every term – suffices for D good digits.

8. Curiosities and open questions

We end this treatise with a tour of some attractive curiosities from the annals of ζ function studies. We do this not only because of the allure of such oddities, but also because there may well be algorithmic consequences in the verification or application of various of our recollections and observations.

Let us first focus on the special case $\zeta(3)$, which number being for many reasons a kind of celebrity in the world of ζ evaluation. The Apéry proof of the irrationality of $\zeta(3)$ which invokes

formula (61), is by now legendary [104]. But within what we might call the Apéry formalism, there are yet more interesting relations. If, like the present authors, one believes that polylogarithms of algebraic arguments are fundamental constants, then there is a "closed-form" expression for $\zeta(3)$ due to Landen [78,79] (6.13) namely

$$\zeta(3) = \frac{5}{4} \operatorname{Li}_3(\tau^{-2}) + \frac{1}{6} \pi^2 \log \tau - \frac{5}{6} \log^3 \tau,$$

where τ is as before the golden mean $(1 + \sqrt{5})/2$, and the polylogarithm is standardly defined:

$$\mathrm{Li}_{s}(z) = \sum_{n=1}^{\infty} \frac{z^{n}}{n^{s}}.$$

An equivalent form is the integral [79] (6.14)

$$\zeta(3) = 10 \int_0^{\log \tau} t^2 \coth t \, \mathrm{d}t$$

with equivalence following from known polylogarithm relations [79]. An open question is whether, in view of the fact that there are coth expansions available, the Li_3 form above can be computationally accelerated. Another byproduct of the Apéry formalism is the remarkable continued fraction:

$$\zeta(3) = \frac{6}{d(0) - \frac{1^6}{d(1) - \frac{2^6}{d(2) - \frac{3^6}{d(3) - \frac{3^6}{2}}}}$$

in which $d(n) = 34n^3 + 51n^2 + 27n + 5$. Such continued fractions can be used to prove irrationality, in yielding theoretical bounds on rational approximations of $\zeta(3)$, although Apéry's original proof and the accepted variants of same do not really concentrate on the fraction per se [104]. Complementary to the theoretical value of the continued fraction, there are intriguing computational questions. One should not rule out the continued fraction as a computational expedient. For one thing, the usual recurrence relations for the convergents p_n/q_n of the fraction need not consume O(n) operations. Because the fraction above has polynomial forms for the elements, one may consider the application of fast polynomial evaluation methods. An open question is, just how efficient can such an evaluation approach be made?

Still on the topic of the illustrious $\zeta(3)$, Broadhurst [35] gave a remarkable formula, amounting to a generalized polylogarithm expansion:

$$\begin{aligned} \zeta(3) &= \frac{8}{7} \sum_{k=1}^{\infty} \frac{1}{k^3} \left(\frac{6a_k}{2^{\lfloor (k+1)/2 \rfloor}} + \frac{4b_k}{2^{\lfloor 3(k+1)/2 \rfloor}} \right), \\ \{a_k\} &= \{1, -7, -1, 10, -1, -7, 1, 0, \ldots\}, \\ \{b_k\} &= \{1, 1, -1, -2, -1, 1, 1, 0\}. \end{aligned}$$

The Broadhurst formula is an extension of the discovery of Bailey et al. [12], that numbers such as π and other constants can be cast in such periodic forms. The forms permit the determination of isolated digits – albeit in restricted bases. In this way, Broadhurst gives the hexadecimal digits, starting from the 10 millionth place (inclusive) of $\zeta(3)$, as: CDA01... It should be remarked that

Broadhurst was also able to determine isolated digits of $\zeta(5)$ using a more complicated summation involving three periodic coefficient sequences. Bailey and Crandall have used such expansions to establish, under a general dynamical hypothesis, random properties of the binary bits in various ζ values [13]. Open questions include this one: as all summands are rational and the terms decay geometrically in k, how best to adapt the Broadhurst series to the FEE method of Karatsuba, for example what should be the "obvious denominators" during series contractions?

It seems as if research on $\zeta(3)$ will never end. As just one example of new directions in this regard, Lan [77] recently proposed a possible avenue for proving $\zeta(3)$, or in fact any $\zeta(\text{odd})$ transcendental. His method involves the theory of certain cyclic fields, to arrive at a formula

$$\zeta(2k+1) = \zeta(2k) \frac{2q^{2k+1} - q^{2k} - q}{q^{2k+1} - 1} A_k(q),$$

where q is a prime and the A_k coefficient can be approximated via calculations in "tamely ramified cyclic fields". The point is, if an A_k could be shown to be algebraic, then $\zeta(2k+1)$ is automatically shown transcendental.

Interdisciplinary appearances of $\zeta(\text{integer})$ can be amusing, attractive. In physics, because the so-called Planck radiation law has terms of the form $(e^x - 1)^{-1}$, the theory of "blackbody radiation" involves the integral (14) and perforce a ζ value. For example $\zeta(3), \zeta(4)$ thus become embedded in certain physical constants involving the theoretical rate at which a hot body radiates energy (in two, three dimensions respectively). Another amusing – and quite different – connection is in number theory, where asymptotic relations can involve $\zeta(\text{integer})$. Here is a well known such relation: the probability that two random integers be relatively prime is $1/\zeta(2)$. But sometimes one encounters a more obscure relation. For example, one has the result of [29] that, if *n* be a power of two, the number #(n) of solutions to n = p + xy with *p* prime and *x*, *y* positive integers enjoys the asymptotic relation

$$\frac{\#(n)}{n} \sim \frac{105\zeta(3)}{2\pi^4}.$$

It is unclear how to attempt high-precision numerical verification of this peculiar result. One may calculate for example that $\#(2^{29}) = 382203245$, giving the poor approximation $\zeta(3) \sim 1.320...$ which is off the mark by 10%.

Next, we mention a computational connection between ζ -values and the gamma function. One can derive intriguing limit relations for values at the odd positive integers, such as

$$\zeta(3) = \lim_{\varepsilon \to 0} \frac{1}{2\varepsilon^3} \log \frac{\Gamma^3(1+\varepsilon)\Gamma(1-\varepsilon)}{\Gamma(1+2\varepsilon)}$$

which shows that a fast algorithm for general Γ evaluation implies an analogous algorithm for $\zeta(3)$. This limiting Γ -formula can be derived from the aforementioned expansion (26) for the ψ function. Incidentally, in practice the actual error in this approximation to $\zeta(3)$ is just about 2 ϵ . Conversely, the functional equation for the Riemann zeta function and duplication formula for the gamma function allow one to compute Γ as efficiently as ζ . We mention in passing that $\Gamma(n/24)$ for positive integers n may be computed with the same reduced complexity as $\pi = \Gamma(1/2)^2$ (see [26]), via elliptic integral evaluations.

One may use the recycling ideas of Algorithm 5 to deduce evaluations for specific arguments, for example:

$$\zeta(3) = -G(1,N)^3 - 3G(1,N)G(2,N) - 3G(3,N) + O(e^{-N}),$$

where we define the G-function as a finite sum:

$$G(s,N) = \sum_{k=1}^{4N} \frac{(-N)^k}{k!k^s}.$$

In fact, $\zeta(2n + 1)$ for any positive integer *n* can be expressed in terms of similar series [68,67]. It is intriguing that this approach yields so simply to the FEE method of Karatsuba: the rational *G* coefficients are so very simple, the summation limit on *G* can conveniently be made a power of two, and so on.

As for interesting interrelations involving general s we note the formulae of Landau:

$$\frac{1}{s-1} = \sum_{n=0}^{\infty} {\binom{s+n-1}{n-1}} \frac{\zeta(s+n) - 1}{n}$$

and of Ramaswami:

$$(1-2^{1-s})\zeta(s) = \sum_{n=1}^{\infty} \binom{s+n-1}{n} \zeta(s+n)$$

Remarkably, either formula is valid for all complex s; either one may be used to define the complete analytic continuation of ζ [101]. We present them here on the idea that perhaps they have computational value. The Landau formula may possibly be used to accelerate other rational ζ -series we have encountered.

An intriguing formula of quite a different character is the following remarkable, van der Pol integral representation, valid on the (open) critical strip, which representation amounts to a complete Fourier decomposition of $\zeta(s)/s$:

$$\zeta(s) = s \int_{-\infty}^{\infty} e^{-\sigma \omega} (\lfloor e^{\omega} \rfloor - e^{\omega}) e^{-i\omega t} d\omega$$

where $s = \sigma + it$, and $\Re(s) \in (0, 1)$. (Actually, the representation can be extended to the half-plane $\Re(s) > 0$ by integrating only over $(0, \infty)$ and adding back a pole term s/(s - 1) on the right-hand side.) The representation is especially intriguing for the Riemann critical line, that is for $\sigma = \frac{1}{2}$. This Fourier approach was actually used fifty years ago by van der Pol, who went so far as to construct an electronic circuit to estimate – in what is called analog fashion – the critical behavior of the Riemann zeta function [10,103]. An open computational question is: can discrete fast Fourier transform methods be efficiently used to estimate the Fourier integral? Of course, one cannot rule out possible convergent schemes arising from theoretical manipulations *per se* of the van der Pol integral representation. One connection between the van der Pol representation and our rational ζ -series runs as follows. One of the known asymptotic relations for ζ on the Riemann critical line is [101]

$$\int_{-T}^{T} |\zeta(1/2 + it)|^2 dt \sim 2T \log T.$$

But we can say something similar by appeal to the Fourier integral above. What might be called the "signal power" relevant to the van der Pol decomposition is

$$P = \int_{-\infty}^{\infty} \left| \frac{\zeta(1/2 + it)}{1/2 + it} \right|^2 dt = 2\pi \int_{-\infty}^{\infty} e^{-\omega} (\lfloor e^{\omega} \rfloor - e^{\omega})^2 d\omega$$
$$= 4\pi \left\{ \frac{3}{2} - \log 2 - \sum_{m=2}^{\infty} \frac{\zeta(m) - 1}{(-1)^m (m+1)} \right\}.$$
(69)

This last relation can be shown via the substitution $\omega \mapsto \log R$ in the power integral, then partitioning the *R* domain into intervals [n, n + 1). At any rate, we have come full circle back to a ζ -series, and provided at least one means for numerical evaluation of the power integral on the critical line. Indeed the ζ -series (69) admits of exact evaluation, as in relation (45), yielding the exact signal power value:

 $P = 2\pi(\log 2\pi - \gamma) = 7.920969195282313657947\dots$

It is likewise intriguing that the Riemann hypothesis can be formulated in terms of the collection of ζ -values at the even positive integers. There is the theorem of Riesz, that the Riemann hypothesis is equivalent to the following big-O behavior of a certain, peculiar *Riesz function R* [101]:

$$R(x) = \sum_{n=1}^{\infty} \frac{(-x)^n}{\zeta(2n)(n-1)!} = \mathcal{O}(x^{1/4+\varepsilon}).$$

Alternatively the Riemann hypothesis is equivalent to a different big-O condition of Hardy and Littlewood [101]:

$$\sum_{n=1}^{\infty} \frac{(-x)^n}{\zeta(2n+1)n!} = \mathcal{O}(x^{-1/4}),$$

It is unclear whether there be any computational value whatsoever to these equivalencies, especially as the big-O statement is involved and therefore infinite computational complexity is implicit, at least on the face of it. Still, if there be any reason to evaluate such sums numerically, the aforementioned methods for recycling of ζ (even) or ζ (odd) values would come into play.

Predating the Riesz function is the Riemann function defined by (5), together with its fascinating connection with the distribution of prime numbers. What makes such connections yet more compelling from a practical viewpoint is that various computational expedients exist for accelerating certain evaluations. For example we have the *Gram formula* (see [95] for a derivation) as

$$\operatorname{Ri}(x) = 1 + \sum_{n=1}^{\infty} \frac{(\log x)^n}{n\zeta(n+1)n!},$$
(70)

whose very form may provide additional motivation for performing recycled computations of ζ -values at positive integer arguments.

We should mention an interesting new foray into the world of asymptotic equivalencies for the Riemann hypothesis: an application of the so-called Carleman continuation problem, described in a treatment due to Aizenberg et al. [2]. Let us paraphrase here the authors' proposition, that the Riemann hypothesis is true if and only if we have the (large-n) behavior:

$$\limsup_{n} |a_{n}^{-}|^{1/n} = \limsup_{n} |a_{n}^{+}|^{1/n} = 1,$$

where

288

$$a_n^{\pm} = \int_{(1\pm 1)/4}^{(3\pm 1)/4} \left(\frac{(z^2 - 1)\mathbf{i} \pm 2z\cos 2\pi x}{1 + z^2 \pm 2z\sin 2\pi x} \right)^n \frac{\mathrm{d}x}{\zeta(x)}$$

where 0 < z < 1 is otherwise unrestricted. It is possible to take the power *n* as high as $N = 10^{20}$, for which the authors find

$$|a_N^-|^{1/N} \sim 0.999999999999999999999956...,$$

 $|a_N^+|^{1/N} \sim 0.9999999999999999999982.....$

It is not yet known what is a proper scale in this asymptotic behavior; that is whether such numerical results imply compelling bounds on locations of critical zeros.

More recent, but in the same general vein of integral equivalencies, is a theorem [14], to the effect that the Riemann hypothesis is true if and only if the integral

$$I = \int \frac{\log|\zeta(s)|}{|s|^2} \,\mathrm{d}t,$$

taken along the critical line $s = \frac{1}{2} + it$, vanishes. This would perhaps not be so compelling if it were not for the exact expression those authors derived for the above integral, namely a sum formula:

$$I = 2\pi \sum_{\Re(\rho) > 1/2} \log \left| \frac{\rho}{1 - \rho} \right|,$$

where ρ denotes any zero in the critical strip, but to the *right* of the critical line as indicated, counting multiplicity. It is interesting to plot the defining *I* integral for ever-increasing integration limits, say, and witness a slow but chaotic tendency toward I = 0. For example, the approximation

$$I(T) = 2 \int_{t=0}^{T} \frac{\log|\zeta(1/2 + it)|}{1/4 + t^2} dt$$

appears to oscillate between about 10^{-9} and 10^{-6} in the vicinity of $T \sim 1000$. One interesting question is: even if the Riemann hypothesis be true, what is a valid positive α such that

$$I(T) = \mathcal{O}(T^{-\alpha}) ?$$

On the basis of preliminary numerical evidence (the aforementioned $T \sim 1000$ data) we are moved to conjecture that $\alpha = 2$ is admissible. It is intriguing that such a numerically motivated statement about a positive α is *stronger* than the Riemann hypothesis. Moreover, the sum formula for *I* could conceivably be used to infer bounds on possible violations of the Riemann hypothesis. For example, here is another interesting question: what could be inferred from sheer computation and the sum formula if one assumed the existence of a *single* errant zero $(\sigma_1 > \frac{1}{2}) + i(t_1 > 0)$ and its redundant reflections?

Also recent is the tantalizing result of [92], to the effect that the Riemann hypothesis is equivalent to a positivity condition on the ξ function defined in (16), which condition applies at a *single* point $s = \frac{1}{2}$ as

$$\frac{\mathrm{d}^n\xi}{\mathrm{d}s^n}\left(\frac{1}{2}\right)>0$$
for every n = 2, 4, 6, ... This brings up an interesting computational exercise; namely, to provide numerical values for a great number of such derivatives. It is nontrivial even to produce the first few, which we list here (to unguaranteed, but suspected implied precision):

$$\frac{\mathrm{d}^{2}\xi}{\mathrm{d}s^{2}}\left(\frac{1}{2}\right) = 0.022971944315145437535249\ldots,$$
$$\frac{\mathrm{d}^{4}\xi}{\mathrm{d}s^{4}}\left(\frac{1}{2}\right) = 0.002962848433687632165368\ldots,$$
$$\frac{\mathrm{d}^{6}\xi}{\mathrm{d}s^{6}}\left(\frac{1}{2}\right) = 0.000599295946597579491843\ldots$$

with the 18th derivative being of order $2 \cdot 10^{-6}$, and so on. Some possible, numerically motivated conjectures are that the sequence of such derivatives is monotone decreasing, but that the successive ratios of the (2m+1)th over the (2m)th are monotone increasing. Note that various of our convergent series for ζ admit of internal differentiation. For example, one might invoke either series (30) or (32) and differentiate with respect to *s* inside the summations. This will entail derivatives of the incomplete gamma function; thus if one uses the integral representation following series (30), powers of logarithms of the integration variable will appear in the formalism, yet we know from the works of Karatsuba (see [69] for example) how to calculate such log-power integrals rapidly from series. What may also work is the differentiation of a sufficiently deep rational polynomial expression as such arises from the continued fraction formalism for incomplete gamma. It goes without saying that if a single negative (2m)th derivative could be found – say to within rigorously bounded numerical error – then the Riemann hypothesis would perforce be broken.

Seemingly reminiscent results in recent times are that of Li [80,20] to the effect that the Riemann hypothesis is equivalent to the positivity property:

$$\lambda_n = \sum_{\rho} \left(1 - \left(1 - \frac{1}{\rho} \right)^n \right) > 0$$

holding for each n = 1, 2, 3, ..., with the sum over critical-strip zeros being interpreted in the usual limit sense. Interestingly, the λ_n constants can be cast in terms of derivatives of log $\xi(s)$, but this time all such evaluated at s = 1. Yet another criterion equivalent to the Riemann hypothesis is that of Lagarias [75]:

$$\Re \left(\frac{\xi'(s)}{\xi(s)}\right) > 0$$

whenever $\Re(s) > \frac{1}{2}$. Furthermore it may well be that the infimum of the real part always occurs for a given $\Re(s)$ at $\Re(s) + 0i$, that is on the real axis.

We close this somewhat recreational section with "interdisciplinary" observations, some highly speculative but some revealing connections between ζ -function theory and other scientific fields.

Let us briefly touch upon experiments that have been performed in the matter of "listening" to the Riemann ζ function, by which we mean hearing a sound signal created as the real part of $\zeta(\sigma + it)$, with imaginary part t taken to be time. One can easily hear qualitative differences between sounds for say $\sigma = 0, \frac{1}{2}, 1$ and so on. We expect this on the basis of differing growth behavior of ζ along

these lines of the *s*-plane. An heuristic can be forwarded [45], to the effect that along the critical line $\sigma = \frac{1}{2}$ the resulting sound is "whitest" in the sense of white (flat) spectrum. One can argue that, in view of the formal sum:

$$\zeta(\sigma+\mathrm{i}t)=\sum_{n=1}^{\infty} \frac{\mathrm{e}^{-\mathrm{i}t\log n}}{n^{\sigma}},$$

the ζ function is a certain superposition of oscillators, with a scaling law that comes down to the estimate

$$P(\omega) \sim e^{-\omega(2\sigma-1)}$$

for the power present at frequency ω . Indeed if this formula be continued – shall we say heuristically – over to the critical strip, the power spectrum would be white on the critical line. Actually, when one "hears" the critical-line signal, it is not random noise as we know it, but the signal does sound like a roughly equal-strength mix of many oscillators. To achieve rigor in these heuristics, one would have to analyze integrals such as (for large *T*):

$$\frac{1}{T} \int_{-T/2}^{T/2} \zeta(1/2 + it) e^{-i\omega t} dt,$$

whose absolute square is essentially the power $P(\omega)$. Due to the existence of such as the van der Pol integral representation earlier in this section, such delicate spectral analysis may well be possible (and may have been performed elsewhere, unbeknownst to the present authors).

More serious (less recreational) is the Hilbert–Pólya conjecture, saying in essence that the behavior of the Riemann zeta function on the critical line $\Re(s) = \frac{1}{2}$ depends somehow on a mysterious (complex) Hermitian operator, of which the critical zeros would be eigenvalues. There is interesting literature, of both theoretical and computational flavors, in this regard. In particular, the pioneering work of Montgomery and Dyson [84] on the statistical correlations amongst consecutive critical zeros now has numerical supporting evidence; and it is widely conjectured that the mysterious Hilbert-Pólya operator is of the Gaussian unitary ensemble (GUE) class. A relevant $n \times n$ matrix *G* in such a theory has $G_{aa} = x_{aa}\sqrt{2}$ and for a > b, $G_{ab} = x_{ab} + iy_{ab}$, together with the Hermiticity condition $G_{ab} = G_{ba}^*$; where every x_{ab}, y_{ab} is a Gaussian random variable with unit variance, mean zero. The computations of Odlyzko [85,86] show that the statistics of consecutive critical zeros are in many ways equivalent – experimentally speaking – to the theoretical distribution of eigenvalues of a large such matrix G. (Actually, there is evidence that a more refined class, namely that of unitary symplectic operators, may be more reasonable as the basis of such conjectures [96].) In these connections, a great deal of fascinating work – by M.V. Berry and colleagues – under the rubric of "quantum chaology" has arisen [17,18]. In some of this work [18], there even appears an asymptotic expansion, reminiscent of the Riemann-Siegel expansion, motivated by semiclassical ideas yet suitable perhaps for high-accuracy calculations on the critical line.

In another connection with quantum chaology, Connes [42] has recently given a spectral interpretation of the critical zeros, as comprising an "absorption spectrum", with noncritical zeros appearing as "resonances". His work connects quantum chaology, algebraic geometry and field theory, yielding interesting equivalent forms of the Riemann hypothesis. There has also appeared an actual, claimed proof of the Riemann hypotheses by de Branges [28], although the present authors are at the time of this writing unaware of any confirmation of that proof. It is intriguing that many of the various new expansion and associated observations relevant to the critical zeros arise from the field of quantum theory, feeding back, as it were, into the study of the Riemann zeta function. But the feedback of which we speak can move in the other direction, as techniques attendant on the Riemann ζ apply to quantum studies. There is the so-called "quantum zeta function", which is a sum (when it exists)

$$z(s) = \sum_{n} \frac{1}{E_n^s}$$

over eigenvalues $\{E_0, E_1, ...\}$ of a specified quantum system. For such as the quantum oscillator with potential x^2 , so that energies are evenly spaced, the quantum z is essentially a scaled form of the Riemann ζ . But – and this is quite the fascinating thing – it turns out that for some quantum systems and certain s, we can evaluate z(s) exactly, even when not a *single* eigenvalue E_n be known. Voros [105] showed that for the so-called power-law oscillator, in which the system potential is x^m for an integer m > 0, one has the exact evaluation:

$$z(1) = \left(\frac{2}{(m+2)^2}\right)^{m/(m+2)} \frac{\Gamma^2(2/(m+2))\Gamma(3/(m+2))}{\Gamma(4/(m+2))\Gamma((m+1)/(m+2))} \left(1 + \sec\frac{2\pi}{m+2}\right)$$

Later, Crandall [46] showed that this relation holds for an arbitrary power-law (i.e., m > 0 need only be real), and conjectured that this relation for z(1) is correct as an analytic continuation in some scenarios for which the literal sum $\sum 1/E_n$ diverges. This is very much like the fact of $\zeta(0) = -\frac{1}{2}$ even though the literal Riemann sum is, of course, divergent at s = 0. The point is, machinery developed over the years on behalf of the Riemann ζ may well apply to the problem of evaluating the quantum z. What is more, the zeros of z(s) may signal, by way of their distribution, the level of quantum chaos inherent to the system. For this intriguing connection, see [46] and references therein.

But in a somewhat different vein there is a precise – in nature neither statistical nor asymptotic – connection between quantum-theoretical operators and the critical zeros. In 1991 it was observed by Crandall [43] that, in the standard formulation of quantum theory there exists a wave function (smooth, devoid of zeros) which, after a finite evolution time under an harmonic-oscillator Schroedinger equation, possesses infinitely many zeros; furthermore these zeros coincide precisely with the Riemann critical zeros. Specifically, define an initial wave function

$$\psi(x,0) = 2\pi \sum_{n=1}^{\infty} n^2 \exp(-\pi n^2 e^{2|x|}) (2\pi n^2 e^{9|x|/2} - 3e^{5|x|/2}),$$

which appears in the standard theory of the critical zeros [101,85], and amounts to the Fourier transform of the Ξ function defined in (18). When plotted graphically this initial wave function looks essentially like a "bell curve", certainly innocent, if you will, on casual inspection. However, evolution of a wave function $\psi(x, \tau)$ under a Schroedinger oscillator equation (where *a* is any positive real constant):

$$i\frac{\partial\psi}{\partial\tau} = -\frac{1}{a^2} \frac{\partial^2\psi}{\partial x^2} + a^2x^2\psi$$

for a precise time interval $0 \le \tau \le \pi/4$ yields a very complicated wave function $\psi(x, \pi/4)$ whose zeros on the x-axis are the zeros of $\zeta(\frac{1}{2} + ix)$, said zeros being therefore infinite in number. All of this

is not hard to show from standard ζ function theory [101] and the theory of quantum harmonic oscillators. For within the latter formalism one can show that after one-quarter of a classical period of the oscillator evolution, a wave packet becomes essentially its own Fourier transform. However, one also knows that basis expansions of wave functions can be useful, so we might contemplate an eigenfunction expansion:

$$\psi(x/a,0) = \sum_{n=2}^{\infty} c_n H_{2n}(x) \exp(-x^2/2),$$

where H_k denotes the Hermite polynomial of degree k, with the coefficients c_n computable in terms of the initial wave packet, via

$$c_n = \frac{\sqrt{\pi}}{2^{2m-1}(2m)!} \int_0^\infty \psi(x/a, 0) H_{2m}(x) \exp(-x^2/2) \, \mathrm{d}x$$

with the parameter a free to be chosen for computational efficiency (a = 4 is a good choice in practice, as below). The result of quantum evolution of a Hermite-basis expansion is particularly simple, and we obtain

$$\Xi(x) = f(x)\zeta(\frac{1}{2} + ix) = a^{-1}\sqrt{2\pi}\exp(-x^2/(2a^2))\sum_{n=0}^{\infty} c_n(-1)^n H_{2n}(x/a),$$
(71)

where we recall, as in definition (18), that the function $\Xi(x) = f(x)\zeta(\frac{1}{2} + ix)$ where f has no real zeros. It is a fascinating thing that the Hermite expansion of the initial wave function only needs these alternating $(-1)^n$ factors to change from a simple-looking wave packet to one with all the complications relevant to the critical line. These observations, albeit recreational, are not entirely specious. For one can perform an actual experiment, taking a=4 and the sum in (71) to say n=N=27 inclusive. In this way there will be 28 of the c coefficients – obtained via numerical integration of the initial packet – and we end up with a degree-54 polynomial in x as an approximation to $\Xi(x)$. This stated experiment yields the specific approximation:

$$\Xi(x) \sim \exp(-x^2/32)(0.497120778225837245 + 0.00404905216049614136x^2 + 0.00000725014346774865092x^4 \cdots - 1.39799726436057536 \cdot 10^{-71}x^{54}),$$

and real zeros of this degree-54 polynomial are located at 14.13472514, 21.022039, 25.01086, 30.4248, 32.93, 37.6, 40.9, and their negatives, where we have indicated the good digits in comparison with established critical zeros – i.e., only good digits have been provided. Incidentally, one does not forget that the degree-54 polynomial must have 54 complex zeros. It turns out that the 40 zeros remaining all have significant imaginary argument. The general picture seems to be this: if one adopts a large-degree-N polynomial, and plots its zeros on the complex ($s = \frac{1}{2} + ix$)-plane, then some number – increasing somehow with N itself – of said zeros lie on the critical line, the rest forming a kind of oval that circumscribes the collection of these real zeros. If the Riemann hypothesis were to be cast in terms of the asymptotic behavior of the zeros of the polynomial

$$\sum_{n=0}^{N} c_n (-1)^n H_{2n}(x/a),$$

3.7

the relevant statement would have to involve the eventual expulsion of all the nonreal zeros, away from, in some appropriate asymptotic sense, the Riemann critical strip. It is likewise intriguing that,

as with any polynomial-root problem, the relevant zeros can, in principle, be described as eigenvalues of a Hessenberg matrix involving the polynomial coefficients.

Incidentally, Hermite polynomials figure into the theory of the Riemann zeta function in at least three other ways. They figure into the Berry–Keating expansion, which we have said is an alternative to the Riemann–Siegel formula [18]. The polynomials have also been used in Motohashi's spectral theory pertinent to ζ [83]. Recently, Bump et al. have analyzed a "local Riemann hyopthesis" into which theory the zeros of Mellin transforms of orthogonal polynomials – including the Hermite variety – figure strongly [39].

Recreational aspects aside, an open issue is whether there be any computational benefit to this quantum connection. We observe that even though a differential equation would be solved numerically, there exist a great many techniques for value recycling – including fast Fourier transform analysis of the Schroedinger equation – in this case meaning simultaneous computation of many wave function values at once. And there is yet another intriguing, interdisciplinary connection. There has been some research on whether solutions to differential equations need be computable. Indeed in [91] it is shown that one can have computable boundary conditions and yet suffer from incomputable solutions. In turn, one recalls Bombieri's suggestion that the Riemann ζ on the critical line is not computable in polynomial (in log t) time. This is all speculative, indeed, but speculation has been a common activity over the long history of the Riemann zeta function.

Acknowledgements

We heartily thank J.P. Buhler, E. Bach, S. Plouffe, S. Wolfram, M. Trott, V. Adamchik, A. Odlyzko, C. Pomerance, P. Flajolet, A. Karatsuba for their ideas and support during the development of this compendium. We are especially grateful to W. Galway for his various theoretical ideas and critical reading, which contributions substantially improved the overall exposition.

References

- [1] V.S. Adamchik, H.M. Srivastava, Some series of the zeta and related functions, Analysis 18 (1998) 131-144.
- [2] L. Aizenberg, V. Adamchik, V. Levit, Approaching the Riemann hypothesis with Mathematica, Math. J. 7 (1997) 54–57.
- [3] S. Akiyama, Y. Tanigawa, Calculation of values of *L*-functions associated to elliptic curves, Math. Comp. 68 (1999) 1201–1231.
- [4] G. Almkvist, A. Granville, Borwein and Bradley's Apéry-like formulae for $\zeta(4n+3)$, Exp. Math. 8 (1999) 197–203.
- [5] T. Amdeberhan, D. Zeilberger, Hypergeometric series acceleration via the WZ method, Electron. J. Combin. 4 (3) (1996) #R3.
- [6] T. Apostol, Introduction to Analytic Number Theory, Springer, Berlin, 1976.
- [7] E. Bach, Comments on the complexity of some number-theoretic constants, manuscript, 28 July 1992.
- [8] E. Bach, The complexity of number-theoretic constants, Inform. Process. Lett. 62 (1997) 145-152.
- [9] E. Bach, R. Lukes, J. Shallit, H. Williams, Results and estimates on pseudopowers, Math. Comp. 65 (1996) 1737–1747.
- [10] E. Bach, J. Shallit, Algorithmic Number Theory, Vol. I, MIT Press, Cambridge, MA, 1996.
- [11] D. Bailey, J. Borwein, R. Crandall, On the Khintchine constant, Math. Comp. 66 (1997) 417-431.
- [12] D. Bailey, P. Borwein, S. Plouffe, On the rapid computation of various polylogarithmic constants, Math. Comp. 66 (1997) 903–913.

- [13] D. Bailey, R. Crandall, On the random character of fundamental constant expansions, manuscript, http://www.perfsci.com/free/techpapers.
- [14] M. Balazard, E. Saias, M. Yor, Notes sur la fonction ζ de Riemann, 2, Adv. = 80 Math 143 (1999) 284–287.
- [15] C. Bays, R. Hudson, Zeroes of Dirichlet L-functions and irregularities in the distribution of primes, Math. Comp. S 0025-5718(99)01105-9 (1999), elec. pub. 10 Mar.
- [16] B.C. Berndt, Ramanujan's Notebooks: Part II, Springer, Berlin, 1985, p. 359.
- [17] M.V. Berry, Quantum chaology, Proc. Roy. Soc. London A 413 (1987) 183-198.
- [18] M.V. Berry, J.P. Keating, A new asymptotic representation for $\zeta(1/2+it)$ and quantum spectral determinants, Proc. Roy. Soc. London A 437 (1992) 151–173.
- [19] E. Bombieri, Remarks on the analytic complexity of Zeta functions, in: Y. Motohashi (Ed.), Analytic Number Theory, London Mathematical Society Lecture Note Series, Vol. 247, Cambridge University Press, Cambridge, 1998, pp. 21–30.
- [20] E. Bombieri, J. Lagarias, Complements to Li's criterion for the Riemann hypothesis, J. Number Theory 77 (1999) 274–287.
- [21] R.C. Boo, An elementary proof of $\sum_{n \ge 1} 1/n^2 = 3D\pi^2/6$, Amer. Math. Mon. 94 (1987) 662–663.
- [22] J.M. Borwein, P.B. Borwein, Pi and the AGM, Wiley, New York, 1987.
- [23] D. Borwein, P. Borwein, A. Jakinovski, An efficient algorithm for the Riemann Zeta function, available from the CECM preprint server, URL = 3D http://www.cecm.sfu.ca/preprints/1995pp.html, CECM-95-043.
- [24] J. Borwein, D. Bradley, Searching symbolically for Apéry-like formulae for values of the Riemann Zeta function, SIGSAM Bull. Symbolic Algebraic Manipulation 116 (1996) 2–7. Also available from the CECM preprint server, URL = 3D http://www.cecm.sfu.ca/preprints/1996pp.html, CECM-96-076.
- [25] J. Borwein, D. Bradley, Empirically determined Apéry-like formulae for $\zeta(4n + 3)$, Exp. Math. 6(3) (1997) 181– 194. Also available from the CECM preprint server, URL = 3D http://www.cecm.sfu.ca/preprints /1996pp.html, CECM-96-069.
- [26] J.M. Borwein, I.J. Zucker, Elliptic integral evaluation of the Gamma function at rational values of small denominator, IMA J. Numer. Anal. 12 (1992) 519–526.
- [27] P. Borwein, An efficient algorithm for the Riemann Zeta function, manuscript, 1995.
- [28] L. de Branges, Proof of the Riemann hypothesis, 1999. ftp://ftp.math.purdue.edu/pub/branges/proof.ps
- [29] B. Bredihin, Applications of the dispersion method in binary additive problems, Dokl. Akad. Nauk. SSSR 149 (1963) 9–11.
- [30] R.P. Brent, The complexity of multi-precision arithmetic. Complexity of Computational Problem Solving, manuscript, 1976.
- [31] R.P. Brent, Fast multiple-precision evaluation of elementary functions, J. ACM 23 (1976) 242–251.
- [32] R.P. Brent, On the zeros of the Riemann zeta function in the critical strip, Math. Comp. 33 (1979) 1361–1372.
- [33] R.P. Brent, J. van de Lune, H.J.J. te Riele, D.T. Winter, On the zeros of the Riemann zeta function in the critical strip. II, Math. Comp. 39 (1982) 681–688.
- [34] R.P. Brent, E.M. McMillan, Some new algorithms for high-precision calculation of Euler's constant, Math. Comp. 34 (1980) 305–312.
- [35] D.J. Broadhurst, Polylogarithmic ladders, hypergeometric series and the ten millionth digits of $\zeta(3)$ and $\zeta(5)$, manuscript: OUT-4102-71, math.CA/9803067 (16 March 1998).
- [36] J. Buhler, R. Crandall, R. Ernvall, T. Metsänkylä, Irregular primes to four million, Math. Comp. 61 (1993) 151–153.
- [37] J. Buhler, R. Crandall, R. Ernvall, T. Metsankyla, A. Shokrollahi, Irregular primes and cyclotomic invariants to eight million, manuscript, 1996.
- [38] J. Buhler, R. Crandall, R.W. Sompolski, Irregular primes to one million, Math. Comp. 59 (1992) 717-722.
- [39] D. Bump, K.K. Choi, P. Kurlberg, J. Vaaler, A local Riemann hypothesis, I, Math. Z. 233 (2000) 1–19.
- [40] H. Cohen, M. Olivier, Calcul des valeurs de la fonction zêta de Riemann en multiprécision, C.R. Acad. Sci. Sér. I Math. 314 (6) (1992) 427–430.
- [41] L. Comtet, Advanced Combinatorics, Reidel, Dordrecht, 1974.
- [42] A. Connes, Trace formula in noncommutative Geometry and the zeros of the Riemann zeta function, manuscript, http://xxx.lanl.gov/ps/math/9811068, 1988.
- [43] R.E. Crandall, Mathematica for the Sciences, Addison-Wesley, Reading MA, 1991.
- [44] R.E. Crandall, Pojects in Scientific Computation, Springer, New York, 1994.

- [45] R.E. Crandall, Topics in Advanced Scientific Computation, Springer, New York, 1996.
- [46] R.E. Crandall, On the quantum zeta function, J. Phys. A. Math. Gen. 29 (1996) 6795-6816.
- [47] R.E. Crandall, Fast evaluation of Epstein zeta functions, manuscript, at http://www.perfsci.com, October 1998.
- [48] R.E. Crandall, Recycled (simultaneous) evaluations of the Riemann zeta function, manuscript, 1999.
- [49] R.E. Crandall, Alternatives to the Riemann-Siegel formula, manuscript, 1999.
- [50] R.E. Crandall, J.P. Buhler, On the evaluation of Euler sums, Exp. Math. 3 (4) (1995) 275-285.
- [51] R. Crandall, C. Pomerance, Prime Numbers: A Computational Perspective, Springer, Berlin, manuscript.
- [52] K. Dilcher, Sums of products of Bernoulli numbers, J. Number Theory 60 (1) (1996) 23-41.
- [53] A. Dutt, V. Rokhlin, Fast Fourier transforms for nonequispaced data, SIAM J. Comput. 14 (6) (1993) 1368–1393.
- [54] A. Dutt, M. Gu, V. Rokhlin, Fast algorithms for polynomial interpolation, integration, and differentiation, SIAM J. Numer. Anal. 33 (5) (1996) 1689–1711.
- [55] H.M. Edwards, Riemann's Zeta Function, Academic Press, New York, 1974.
- [56] P. Flajolet, I. Vardi, Zeta function expansions of classical constants, 1996 24 Feb, manuscript. http://pauillac.inria.fr/ algo/flajolet/Publications/publist.html
- [57] W. Gabcke, Neue Herleitung und Explizite Restabschätzung der Riemann-Siegel-Formel, Thesis, Georg-August-Universität zu Göttingen, 1979.
- [58] W. Galway, Fast computation of the Riemann Zeta function to abitrary accuracy, manuscript, 1999.
- [59] B. Haible, T. Papanikolaou, Fast multiprecision evaluation of series of rational numbers, in: J.P. Buhler (Ed.), Algorithmic Number Theory, Third International Symposium, ANTS-III, Lecture Notes in Computer Science, Vol. 1423, Springer, Berlin, 1998.
- [60] P. Henrici, Applied and Computational Complex Analysis, Vol. 2, Wiley, New York, 1977.
- [61] M.M. Hjortnaes, Overf ϕ ring av rekken $\sum_{k=3D1}^{\infty} (1/k^3)$ till et bestemt integral, Proceedings of 12th Congress on Scand. Maths, Lund 1953, (Lund 1954).
- [62] A.E. Ingham, The Distribution of Prime Numbers 2nd edition Cambridge Tracts in Mathematics, Vol. 30, Cambridge University Press, Cambridge 1932, 1990, with a Foreword by R.C. Vaughan.
- [63] A. Ivic, The Riemann Zeta-Function, Wiley, New York, 1985.
- [64] A. Ivic, On some results concerning the Riemann hypothesis, in: Y. Motohashi (Ed.), Analytic Number Theory London Mathematical Society, Lecture Note Series, Vol. 247, Cambridge University Press, Cambridge, 1998, pp. 139–167.
- [65] E.A. Karatsuba, On new method of fast computation of transcendental functions, Uspechi Mat. Nauk (Successes of Mathematical Sciences) 46 (2) (1991) 219–220.
- [66] E.A. Karatsuba, On fast computation of transcendental functions, Soviet Math. Dokl. 43 (1991) 693-694.
- [67] E.A. Karatsuba, Fast evaluation of transcendental functions, Probl. Peredachi Inform. 27 (1991) 76–99.
- [68] E.A. Karatsuba, Fast evaluation of $\zeta(3)$, Probl. Peredachi Inform. 29 (1993) 68–73.
- [69] E.A. Karatsuba, Fast Calculation of the Riemann Zeta function $\zeta(s)$ for integer values of the argument *s*, Probl. Inform. Transm. 31 (1995) 353–362.
- [70] E.A. Karatsuba, On fast computation of Riemann zeta function for integer values of argument, Dokl.RAN 349 (4) (1996) 463.
- [71] E.A. Karatsuba, Fast evaluation of hypergeometric function by FEE, in: N. Papamichael, St. Ruscheweyh, E.B. Saff (Eds.), Computational Methods and Function Theory 1997, Proceedings of the Third CMFT Conference, Series in Approximation and Decomposition, World Scientific, Singapore, 1999, pp. 303–314 = 20.
- [72] A.A. Karatsuba, S.M. Voronin, The Riemann Zeta-Function, Walter de Gruyter, Berlin, 1992.
- [73] M. Koecher, Klassische Elementare Analysis, Birkhäuser, Basel, 1987.
- [74] M. Koecher, Lett. Math. Intelligencer 2 (1980) 62–64.
- [75] J. Lagarias, On a positivity property of the Riemann ξ -function, Acta Arithmetica 89 (1999) 217–234.
- [76] J. Lagarias, A. Odlyzko, Computing $\pi(x)$: an analytic method, J. Algorithms 8 (1987) 173–191.
- [77] Y. Lan, A limit formula for $\zeta(2k+1)$, J. Number Theory = 20 78 (1999) 271–286.
- [78] J. Landen, Math. Memoirs 1 (1780) 118.
- [79] L. Lewin, Polylogarithms and Associated Functions, Elsevier, North-Holland Inc, New York, 1981.
- [80] X.-J. Li, The positivity of a sequence of numbers and the Riemann hypothesis, J. Number Theory 65 (1997) 325–333.
- [81] J.E. Littlewood, Sur la distribution des nombres premiers, C.R. Acad. Sci. (Paris) 158 (1914) 1869–1872.

- [82] J. van de Lune, H.J.J. te Riele, D.T. Winter, On the zeros of the Riemann zeta function in the critical strip. IV, Math. Comp. 46 (1986) 667–681.
- [83] Y. Motohashi, Spectral Theory of the Riemann Zeta-Function, Cambridge Tracts in Mathematics, Vol. 127, Cambridge University Press, Cambridge, 1997.
- [84] A. Odlyzko, On the distribution of spacings between zeros of the Zeta function, Math. Comp. 48 (1991) 273–308.
- [85] A. Odlyzko, Analytic computations in number theory, Proc. Symp. Appl. Math. 48 (1994) 441-463.
- [86] A. Odlyzko, The 10²⁰th zero of the Riemann zeta function and 175 million of its neighbors, 1998 http://www.research.att.com/ amo/unpublished/index.html.
- [87] A. Odlyzko, A. Schönhage, Fast algorithms for multiple evaluations of the Riemann zeta-function, Trans. Amer. Math. Soc. 309 (1988) 797–809.
- [88] A. Odlyzko, H.J.J. te Riele, Disproof of the Mertens conjecture, J. Reine Angew. Math. 357 (1985) 138-160.
- [89] R.B. Paris, An asymptotic representation for the Riemann zeta function on the critical line, Proc. Roy. Soc. London Ser. A 446 (1994) 565–587.
- [90] S. Plouffe, http://www.lacim.uqam.ca/pi/index.html (1998).
- [91] M.B. Pour-El, I. Richards, The wave equation and computable initial data such that its unique solution is not computable, Adv. Math. 39 (1981) 215–239.
- [92] L.D. Pustyl'nikov, On a property of the classical zeta-function associated with the Riemann conjecture on zeros, Russian Math. Surveys 54 (1) (1999) 262–263.
- [93] R. Ramanujan, in: G. Hardy, P. Aiyar, B. Wilson (Eds.), Collected Papers of Srinivasa Ramanujan, Cambridge University Press, Cambridge, 1927, Section 19.
- [94] P. Ribenboim, The New Book of Prime Number Records, Springer, New York, 1995.
- [95] H. Riesel, Prime Numbers and Computer Methods for Factorization, Birkhauser, Boston, 1994, p. 126.
- [96] M.O. Rubinstein, Evidence for a spectral interpretation of the zeros of *L*-functions, Dissertation, Princeton University, 1998.
- [97] A. Schönhage, Schnelle Multiplikation grosser Zahlen, Computing 7 (1971) 281–292.
- [98] S. Skewes, On the Difference $\pi(x) li(x)$, J. London Math. Soc. 8 (1933) 277–283.
- [99] S. Skewes, On the Difference $\pi(x) \text{li}(x)$, II, J. London Math. Soc. 5 (1955) 48–70.
- [100] H.J.J. te Riele, On the Sign of the difference of $\pi(x) \text{li}(x)$, Math. Comp. 48 (1987) 323–328.
- [101] E.C. Titchmarsh, The Theory of the Riemann Zeta-Function, University Press, Oxford, 1967.
- [102] I. Vardi, Computational Recreations in Mathematica, Addison-Wesley, Reading, MA, 1991.
- [103] B. van der Pol, Bull. AMS 53 (1947) 976 ff.
- [104] A. van der Poorten, A proof that Euler missed: Apéry's proof of the irrationality of $\zeta(3)$, Math. Intelligencer 1 (1970) 195–203.
- [105] A. Voros, Nucl. Phys. B165 (1980) 209-236.
- [106] E.T. Whittaker, G.N. Watson, A Course of Modern Analysis, 4th Edition, Cambridge University Press, London, 1969.
- [107] S. Wolfram, D. Hillman, private communication, 1998.
- [108] S.C. Woon, Generalization of a relation between the Riemann zeta function and the Bernoulli numbers, 1998 December, manuscript.
- [109] J.W. Wrench, D. Shanks, Questions concerning Khintchine's constant and the efficient computation of regular continued fractions, Math. Comp. 20 (1966) 444–448.
- [110] N.Y. Zhang, K.S. Williams, Some Series Representations of $\zeta(2n + 1)$, Rocky Mountain J. Math. 23 (4) (1993) 1581–1592.



Journal of Computational and Applied Mathematics 121 (2000) 297-312

JOURNAL OF COMPUTATIONAL AND APPLIED MATHEMATICS

www.elsevier.nl/locate/cam

Recursion formulae for basic hypergeometric functions¹

Stanislaw Lewanowicz*

Institute of Computer Science, University of Wroclaw, ul. Przesmyckiego 20, 51-151 Wroclaw, Poland

Received 16 June 1999

Abstract

We show that the basic hypergeometric functions

$$F_{k}(\omega) := \frac{\prod_{i=1}^{r} (a_{i};q)_{k}}{(q;q)_{k} \prod_{j=1}^{s} (b_{j};q)_{k}} \omega^{k} \left[(-1)^{k} q^{\binom{k}{2}} \right]^{s+1-r} r_{r+1} \phi_{s+1} \begin{pmatrix} a_{1}q^{k}, \dots, a_{r}q^{k}, \alpha q^{k+1} \\ b_{1}q^{k}, \dots, b_{s}q^{k}, \alpha \beta q^{2k+2} \end{pmatrix} q; \omega q^{k(s+1-r)} \end{pmatrix}$$

satisfy a recurrence relation of the form

$$\sum_{i=0}^{\vartheta} \left[A_i(k) + \frac{1}{\omega} B_i(k) \right] F_{k+i}(\omega) = 0, \qquad \vartheta = \max(r+1, s+2),$$

where $A_i(k)$, $B_i(k)$ are rational functions of q^k , and $B_0(k) = B_{\vartheta}(k) \equiv 0$.

When r = s + 1 and $\omega = q$, this result can be refined. Namely, we show that the functions

$$F_{k}(q) := \frac{\prod_{i=1}^{s+1} (a_{i}; q)_{k}}{(q; q)_{k} \prod_{j=1}^{s} (b_{j}; q)_{k}} q^{k}{}_{s+2} \phi_{s+1} \begin{pmatrix} a_{1}q^{k}, \dots, a_{s+1}q^{k}, \alpha q^{k+1} \\ b_{1}q^{k}, \dots, b_{s}q^{k}, \alpha \beta q^{2k+2} \end{pmatrix} q; q \end{pmatrix},$$

satisfy a recurrence relation of order s + 1,

$$\sum_{i=0}^{s+1} C_i(k) F_{k+i}(q) = 0$$

with rational coefficients in q^k . © 2000 Elsevier Science B.V. All rights reserved.

MSC: 33C25; 33D45

Keywords: Basic hypergeometric functions; Recurrence relations; q-difference equations; Little q-Jacobi polynomials

¹ Supported by Komitet Badań Navkowych (Poland) under the grant 2P03A02717.

^{*} Fax: +71-251-271.

E-mail address: stanislaw.lewanowicz@ii.uni.wroc.pl (S. Lewanowicz).

1. Notation

The notation used in this section is that of [1, Chapter 1] (see also [4,8]). From now on we shall always assume that 0 < q < 1.

For $x, q \in \mathbb{C}$ define the *q*-shifted factorial by

$$(x;q)_n := \prod_{j=0}^{n-1} (1-q^j x), \quad n=0,1,\dots$$
 (1.1)

The basic hypergeometric series is defined by

$${}_{r}\phi_{s}\left(\begin{array}{c}a_{1},a_{2},\ldots,a_{r}\\b_{1},b_{2},\ldots,b_{s}\end{array}\middle|q;z\right) := \sum_{k=0}^{\infty}\frac{\prod_{i=1}^{r}(a_{i};q)_{k}}{\prod_{j=1}^{s}(b_{j};q)_{k}}\left[(-1)^{k}q^{\binom{k}{2}}\right]^{1+s-r}z^{k}.$$
(1.2)

Here $r, s \in \mathbb{Z}_+$ and parameters $a_1, a_2, \ldots, a_r, b_1, b_2, \ldots, b_s$, and z are in \mathbb{C} . In order to have a well-defined series, we require that $b_1, b_2, \ldots, b_s \neq q^{-k}$, $k = 0, 1, \ldots$.

The *q*-difference operator D_q is given by

$$\boldsymbol{D}_{q}f(x) := \frac{(\boldsymbol{E}_{q} - \boldsymbol{I})f(x)}{(\boldsymbol{E}_{q} - \boldsymbol{I})x} = \frac{f(qx) - f(x)}{(q - 1)x}, \quad x \neq 0$$
(1.3)

with $D_q f(0) := f'(0)$, provided f'(0) exists. Here E_q is the q-shift operator, $E_q f(x) = f(qx)$, and I is the identity operator, If(x) = f(x). (By convention, all the bold letter operators act on the variable x.)

The *q*-integral is defined by

$$\int_0^z f(x) \, \mathrm{d}_q x := z(1-q) \sum_{k=0}^\infty f(zq^k) q^k, \quad z > 0.$$
(1.4)

Given a set c of constants c_1, c_2, \ldots, c_u , we define the symmetric polynomials $S_i(c)$ by

$$\begin{cases} S_0(\mathbf{c}_{-}) := 1, \\ S_i(\mathbf{c}_{-}) := \sum_{\substack{t_1, t_2, \dots, t_i \in \{1, 2, \dots, u\} \\ t_1 < t_2 < \dots < t_i}} c_{t_1} c_{t_2} \cdots c_{t_i}, \quad i = 1, 2, \dots, u. \end{cases}$$
(1.5)

2. Introduction

The monic little q-Jacobi polynomials $\{p_k(x; \alpha, \beta | q)\}$ are defined by

$$p_k(x;\alpha,\beta|q) = (-1)^k q^{\binom{k}{2}} \frac{(\alpha q;q)_k}{(\alpha\beta q^{k+1};q)_k} {}_2\phi_1 \left(\begin{array}{c} q^{-k}, \alpha\beta q^{k+1} \\ \alpha q \end{array} \middle| q;qx, \right), \quad k \ge 0$$

$$(2.1)$$

(see, e.g., [3, Section 3.12]). The following expansion can be deduced from a general result given in [1, Eq. (3.7.9)] (see also [9]):

$${}_{r}\phi_{s}\left(\begin{array}{c}a_{1},a_{2},\ldots,a_{r}\\b_{1},b_{2},\ldots,b_{s}\end{array}\middle|q;\omega x\right)=\sum_{k=0}^{\infty}F_{k}(\omega)\ p_{k}(x;\alpha,\beta|q),$$
(2.2)

where $0 < \alpha q < 1$, $\beta q < 1$, and

$$F_{k}(\omega) := \frac{\prod_{i=1}^{r} (a_{i};q)_{k}}{(q;q)_{k} \prod_{j=1}^{s} (b_{j};q)_{k}} \omega^{k} [(-1)^{k} q^{\binom{k}{2}}]^{s+1-r} \times_{r+1} \phi_{s+1} \left(\begin{array}{c} a_{1}q^{k}, \dots, a_{r}q^{k}, \alpha q^{k+1} \\ b_{1}q^{k}, \dots, b_{s}q^{k}, \alpha \beta q^{2k+2} \end{array} \middle| q; \omega q^{k(s+1-r)} \right).$$

$$(2.3)$$

In the present paper, we show that the functions $F_k(\omega)$ satisfy a recurrence relation of the form

$$\sum_{i=0}^{\vartheta} \left[A_i(k) + \frac{1}{\omega} B_i(k) \right] F_{k+i}(\omega) = 0, \qquad \vartheta = \max(r+1, s+2),$$

where $A_i(k)$, $B_i(k)$ are rational functions of q^k , and $B_0(k) = B_{\vartheta}(k) \equiv 0$.

In the special case when r = s + 1 and $\omega = q$, this result can be refined. Namely, we show that the functions

$$F_{k}(q) := \frac{\prod_{i=1}^{s+1} (a_{i};q)_{k}}{(q;q)_{k} \prod_{j=1}^{s} (b_{j};q)_{k}} q^{k}{}_{s+2} \phi_{s+1} \begin{pmatrix} a_{1}q^{k}, \ldots, a_{s+1}q^{k}, \alpha q^{k+1} \\ b_{1}q^{k}, \ldots, b_{s}q^{k}, \alpha \beta q^{2k+2} \end{pmatrix} q;q \end{pmatrix},$$

satisfy a recurrence relation of order s + 1,

$$\sum_{i=0}^{s+1} C_i(k) F_{k+i}(q) = 0$$

with rational coefficients in q^k .

This may be considered as a partial analogue of Wimp's result [10] (see also [7, vol. 2, pp. 135 ff]) for the generalized hypergeometric functions

$$U_{k}(\omega) := \frac{\prod_{i=1}^{r} (a_{i})_{k}}{(k+\gamma)_{k} \prod_{j=1}^{s} (b_{j})_{k}} \omega^{k}{}_{r+1}F_{s+1} \left(\begin{array}{c} a_{1}+k, \dots, a_{r}+k, \alpha+k \\ b_{1}+k, \dots, b_{s}+k, 2k+\gamma+1 \end{array} \right| \omega \right)$$

and its refinment given in [5] in case when $\omega = 1$. Here $(c)_k := \Gamma(c+k)/\Gamma(c)$.

The proposed approach is a specialization of a general method (see [6]) for obtaining a recurrence relation for the coefficients of expansion with respect to q-classical orthogonal polynomials of a function satisfying a linear q-difference equation with polynomial coefficients. Notice that the function

$$f(x) := {}_r \phi_s \left(\begin{array}{c} a_1, a_2, \dots, a_r \\ b_1, b_2, \dots, b_s \end{array} \middle| q; \omega x \right),$$

satisfies the equation

$$\boldsymbol{P}_{\theta}f(x) \equiv \left\{ \omega x \left(\prod_{i=1}^{r} \left(a_{i}\boldsymbol{E}_{q} - \boldsymbol{I} \right) \right) \boldsymbol{E}_{q}^{s+1-r} - \left(\boldsymbol{E}_{q} - \boldsymbol{I} \right) \prod_{j=1}^{s} \left(q^{-1}b_{j}\boldsymbol{E}_{q} - \boldsymbol{I} \right) \right\} f(x) = 0,$$
(2.4)

of the order $\theta := \max(r, s+1)$ (see [1, p. 27]).

In Section 3 we list certain properties of the little q-Jacobi polynomials, as well as of the related Fourier coefficients. The main results of the paper are given in Section 4 (see Theorems 4.2 and 4.5). Some illustrative examples are given in Section 5.

3. Properties of the little q-Jacobi polynomials

3.1. Basic properties

The little q-Jacobi polynomials $\{p_k(x; \alpha, \beta | q)\}$ (cf. (2.1)) belong to a large q-Hahn class of basic hypergeometric orthogonal polynomials [2,4]. The orthogonality relation for these polynomials reads

$$\int_0^1 \varrho(x) p_k(x) p_l(x) \operatorname{d}_q x = h_k \delta_{kl}, \qquad (3.1)$$

where $p_m(x) \equiv p_m(x; \alpha, \beta | q), m = 0, 1, \dots, 0 < \alpha q < 1, \beta q < 1,$

$$\varrho(x) := x^a \frac{(qx;q)_{\infty}}{(\beta qx;q)_{\infty}} \quad (\alpha = q^a),$$
(3.2)

$$h_{k} := q^{k(k+a)} \frac{(q;q)_{k}(\alpha q;q)_{k}(\beta q;q)_{k}(\alpha \beta q;q)_{k}}{(\alpha \beta q;q)_{2k}(\alpha \beta q^{2};q)_{2k}}.$$
(3.3)

They satisfy the three-term recurrence relation

$$x p_k(x) = \xi_0(k) p_{k-1}(x) + \xi_1(k) p_k(x) + \xi_2(k) p_{k+1}(x),$$

$$k = 0, 1, \dots, p_{-1}(x) \equiv 0, \quad p_0(x) \equiv 1,$$
(3.4)

where the coefficients $\xi_i(k)$ are given by (see, e.g. [3, Section 3.12] or [8, Section 3.1])

$$\xi_0(k) := A(k-1)C(k), \quad \xi_1(k) := A(k) + C(k), \quad \xi_2(k) := 1.$$
(3.5)

Here

$$A(k) := q^{k} \frac{(\alpha q^{k+1} - 1)(\alpha \beta q^{k+1} - 1)}{(\alpha \beta q^{2k+1} - 1)(\alpha \beta q^{2k+2} - 1)}, \qquad C(k) := \alpha q^{k} \frac{(q^{k} - 1)(\beta q^{k} - 1)}{(\alpha \beta q^{2k} - 1)(\alpha \beta q^{2k+1} - 1)}.$$
(3.6)

We shall need some further properties enjoyed by $\{p_k(x)\}$ (see, e.g. [3, Section 3.12] or [8, Section 2]).

First, for any $k \in \mathbb{Z}_+$, the polynomial $p_k(x)$ satisfies a second-order q-difference equation

$$\boldsymbol{L}_{k} p_{k}(x) \equiv \left\{ \sigma(x) \boldsymbol{D}_{q} \boldsymbol{D}_{q^{-1}} + \tau(x) \boldsymbol{D}_{q} + \lambda_{k} \boldsymbol{I} \right\} p_{k}(x) = 0,$$
(3.7)

where

$$\sigma(x) = x(x-1), \qquad \tau(x) = \frac{1 - \alpha q - (1 - \alpha \beta q^2)x}{q-1}, \tag{3.8}$$

$$\lambda_k := q(q^{-k} - 1)(\alpha\beta q^{k+1} - 1)/(q - 1)^2.$$
(3.9)

Second, the weight function ρ satisfies the q-difference equation

$$\boldsymbol{D}_q(\sigma\varrho) = \tau\varrho. \tag{3.10}$$

Third, we have the D_q -structure relation [8, Section 3.2]

$$\sigma^{+}(x)\boldsymbol{D}_{q}p_{k}(x) = \delta_{0}(k)p_{k-1}(x) + \delta_{1}(k)p_{k}(x) + \delta_{2}(k)p_{k+1}(x), \qquad (3.11)$$

where

$$\sigma^{+}(x) := \sigma(x) + (q-1)x\tau(x) = \alpha q x (\beta q x - 1)$$
(3.12)

and

$$\delta_{0}(k) := q^{1-k} \frac{1 - \alpha \beta q^{k+1}}{q - 1} \xi_{0}(k),$$

$$\delta_{1}(k) := \frac{q}{q^{2} - 1} \{ \alpha q - 1 - [(q^{-k} - 1)(\alpha \beta q^{k+1} - 1) + \alpha \beta q^{2} - 1] \xi_{1}(k) \},$$

$$\delta_{2}(k) := \frac{\alpha \beta q^{2}(q^{k} - 1)}{q - 1}.$$
(3.13)

Fourth, we have the following q-difference-recurrence identities [6]:

$$\frac{\sigma^+(x)}{x-\zeta} \boldsymbol{D}_q \{ \omega_1(k) p_k(x) + \omega_2 p_{k+1}(x) \} = \pi_1(k) p_k(x) + \pi_2(k) p_{k+1}(x),$$
(3.14)

where $x - \zeta$ is any linear divisor of σ^+ (thus $\zeta = 0$ or $\zeta = 1/(\beta q)$), and

$$\omega_{1}(k) := \xi_{0}(k+1)\delta_{2}(k+1)/\xi_{2}(k+1) - \delta_{0}(k+1),$$

$$\omega_{2}(k) := \delta_{1}(k) - \delta_{0}(k)(\xi_{1}(k) - \zeta)/\xi_{0}(k),$$

$$\pi_{1}(k) := \delta_{0}(k)\omega_{1}(k)/\xi_{0}(k),$$

$$\pi_{2}(k) := \delta_{2}(k+1)\omega_{2}(k)/\xi_{2}(k+1).$$

(3.15)

3.2. Identities for the Fourier coefficients

Given a function f(x), the Fourier coefficients of f are defined by

$$a_k[f] := h_k^{-1} b_k[f], \quad k = 0, 1, \dots$$
 (3.16)

where

$$b_k[f] := \int_0^1 \varrho(x) p_k(x) f(x) \, \mathrm{d}_q x.$$

Let $\mathscr{X}, \mathscr{D}, \mathscr{P}$, and \mathscr{D} be the difference operators (acting on k) defined by

$$\mathscr{X} := \xi_0(k)\mathscr{E}^{-1} + \xi_1(k)\mathscr{I} + \xi_2(k)\mathscr{E}, \tag{3.17}$$

$$\mathscr{D} := \delta_0(k) \mathscr{E}^{-1} + \delta_1(k) \mathscr{I} + \delta_2(k) \mathscr{E}, \tag{3.18}$$

$$\mathscr{P} := \pi_1(k)\mathscr{I} + \pi_2(k)\mathscr{E}, \tag{3.19}$$

$$\mathcal{Q} := \omega_1(k)\mathcal{I} + \omega_2(k)\mathcal{E},\tag{3.20}$$

the notation used being that of (3.4), (3.11), and (3.14). Here \mathscr{I} is the identity operator, $\mathscr{I}b_k[f] =$ $b_k[f]$, and \mathscr{E}^m is the *m*th shift operator, $\mathscr{E}^m b_k[f] = b_{k+m}[f]$ ($m \in \mathbb{Z}$). For the sake of simplicity, we write \mathscr{E} in place of \mathscr{E}^1 . (We adopt the convention that all the script letter operators act on the variable k.)

Further, let us define the difference operators L and Z_{ζ} (acting on x) by

$$\boldsymbol{L} := \sigma(\boldsymbol{x})\boldsymbol{D}_{q}\boldsymbol{D}_{q^{-1}} + \tau(\boldsymbol{x})\boldsymbol{D}_{q}, \tag{3.21}$$

$$Z_{\zeta} := (x - \zeta)D_q$$
, where $(x - \zeta)$ is a linear divisor of $\sigma^+(x)$. (3.22)

Using (3.4)–(3.14), the following lemma can be proved.

Lemma 3.1 (Lewanowicz et al. [6]). The coefficients $b_k[f]$ satisfy the identities:

.

$$b_{k}[pf] = p(\mathcal{X})b_{k}[f], \quad p\text{-arbitrary polynomial},$$

$$\mathcal{D}b_{k}[\boldsymbol{D}_{q}f] = \lambda_{k}b_{k}[f],$$

$$\mathcal{P}b_{k}[\boldsymbol{Z}_{\zeta}f] = \mathcal{Q}(\lambda_{k}b_{k}[f]),$$

$$b_{k}[\boldsymbol{L}f] = -\lambda_{k}b_{k}[f].$$
(3.23)

In the remaining part of the paper, we use the notation Z for the operator defined in (3.22) with $\zeta = 0$, thus

$$\boldsymbol{Z} \equiv \boldsymbol{Z}_0 := \boldsymbol{x} \boldsymbol{D}_q. \tag{3.24}$$

Let us define for i = 0, 1, ..., the operators \mathcal{P}_i and \mathcal{Q}_i by

$$\mathcal{P}_{i} := \pi_{i1}(k)\mathcal{I} + \pi_{i2}(k)\mathcal{E},$$

$$\mathcal{Q}_{i} := \omega_{i1}(k)\mathcal{I} + \omega_{i2}(k)\mathcal{E},$$

(3.25)

where

$$\omega_{i1}(k) := \lambda_k \omega_1(k+i) \prod_{j=1}^i \left[\vartheta(k+j) - \eta(k+i) \right],$$

$$\omega_{i2}(k) := \lambda_{k+1} \omega_2(k) \prod_{j=0}^{i-1} \left[\vartheta(k) - \eta(k+j) \right],$$

$$\pi_{i1}(k) := \vartheta(k) \omega_{i1}(k),$$

$$\pi_{i2}(k) := \eta(k+i) \omega_{i2}(k)$$
(3.26)

and

$$\eta(k) := \delta_2(k+1)/\xi_2(k+1), \qquad \vartheta(k) := \delta_0(k)/\xi_0(k).$$

(Hence $\mathscr{P}_0 = \mathscr{P}, \ \mathscr{Q}_0 = \mathscr{Q}(\lambda_k \mathscr{I})$, where \mathscr{P} and \mathscr{Q} are defined according to (3.19), (3.20), respectively, with $\zeta = 0.$)

Further, let

$$\mathcal{R}_{ij} := \begin{cases} \mathscr{I}, & i < j, \\ & i, j \ge 0, \\ \mathscr{P}_i \mathscr{P}_{i-1} \cdots \mathscr{P}_j, & i \ge j, \end{cases}$$
$$\mathcal{F}_i := \mathscr{R}_{i-1,0}, \quad i \ge 0, \\ \mathscr{V}_0 := \mathscr{I}, \\ \mathscr{V}_i := \mathscr{Q}_{i-1} \dots \mathscr{Q}_1 \mathscr{Q}_0, \quad i \ge 1. \end{cases}$$
(3.27)

Lemma 3.2. The identity

$$\mathscr{T}_i b_k[\mathbf{Z}^i f] = \mathscr{V}_i b_k[f] \tag{3.28}$$

holds for i = 0, 1, ...

Proof. We use induction on *i*. For i = 0, (3.28) is obviously true, and for i = 1 it takes the form $\mathscr{P}_0 b_k[\mathbb{Z}f] = \mathscr{Q}_0 b_k[f]$, that is a disguised form of (3.23).

Now, assume that (3.28) holds for a certain i ($i \ge 1$). We have

$$\mathscr{T}_{i+1}b_k[\mathbf{Z}^{i+1}f] = \mathscr{P}_i\mathscr{T}_ib_k[\mathbf{Z}^i(\mathbf{Z}f)] = \mathscr{P}_i\mathscr{V}_i(b_k[\mathbf{Z}f]).$$

It can be checked that

 $\mathscr{P}_i \mathscr{Q}_{i-1} = \mathscr{Q}_i \mathscr{P}_{i-1}, \quad i = 1, 2, \dots,$

which implies

$$\mathscr{P}_i \mathscr{V}_i = \mathscr{Q}_i \dots \mathscr{Q}_2 \mathscr{Q}_1 \mathscr{P}_0.$$

Hence, by virtue of the first part of the proof,

$$\mathscr{T}_{i+1}b_k[\mathbf{Z}^{i+1}f] = \mathscr{V}_{i+1}b_k[f]. \qquad \Box$$

4. Results

4.1. General case

We shall need the following result.

Lemma 4.1. Difference equation (2.4), satisfied by the function

$$f(x) := {}_{r}\phi_{s} \left(\begin{array}{c} a_{1}, a_{2}, \dots, a_{r} \\ b_{1}, b_{2}, \dots, b_{s} \end{array} \middle| q; \omega x \right), \tag{4.1}$$

can be written in the equivalent form

$$(\boldsymbol{Q}_{\theta}f)(x) \equiv \boldsymbol{D}_{q} \left\{ \sum_{i=0}^{\theta-1} \boldsymbol{Z}^{i} \left(\left(u_{i}x - \frac{v_{i}}{\omega} \right) f(x) \right) + \sum_{j=0}^{d-1} \bar{u}_{j} \boldsymbol{Z}^{j}(xf(x)) \right\} + vf(x) = 0,$$
(4.2)

where $Z := xD_q$, $\theta := \max(r, s + 1)$, d := s + 1 - r, and

$$u_{i} := q^{-d} (q-1)^{i} \sum_{h=0}^{m_{i}} (-1)^{h} \left[\begin{pmatrix} \theta - h \\ i+1 \end{pmatrix} - \begin{pmatrix} \theta - m_{i} \\ i+1 \end{pmatrix} \right] S_{r-h}(\mathfrak{a}/q),$$

$$m_{i} := \min(\theta - i, r), \quad i = 0, 1, \dots, \theta - 1,$$
(4.3)

$$v_{i} := (q-1)^{i} \sum_{h=0}^{\min(\theta-1-i,s)} (-1)^{h} \begin{pmatrix} \theta-1-h\\ i \end{pmatrix} S_{s-h}(\mathfrak{b}/q), \quad i = 0, 1, \dots, \theta-1,$$
(4.4)

$$\bar{u}_j := u(q-1)^{1+j} \binom{d}{j+1}, \quad j = 0, 1, \dots, d-1,$$
(4.5)

$$v := q^{-s-1}(q-1)^{-1} \prod_{i=1}^{r} (a_i - q).$$
(4.6)

Here the symbols $S_l(\mathfrak{a}/q)$, $S_n(\mathfrak{b}/q)$ have the meaning given in the definition (1.5).

Proof. Let $s + 1 \ge r$. Let us write the operator P_{θ} given by (2.4) as

$$\boldsymbol{P}_{\theta} = \boldsymbol{S}_{r} \boldsymbol{E}_{q}^{s+1-r} - (\boldsymbol{E}_{q} - \boldsymbol{I}) \boldsymbol{V}_{s}, \tag{4.7}$$

where

$$S_r := \omega x \prod_{i=1}^r (a_i E_q - I) = \omega x \sum_{j=0}^r (-1)^{r-j} S_j(\mathfrak{a}) E_q^j,$$
(4.8)

$$V_{s} := \prod_{i=1}^{s} (q^{-1}b_{i}\boldsymbol{E}_{q} - \boldsymbol{I}) = \sum_{j=0}^{s} (-1)^{s-j} S_{j}(\boldsymbol{b}/q) \boldsymbol{E}_{q}^{j}.$$
(4.9)

We have

$$S_r = \omega(E_q - I)T_{r-1} + \omega x t I,$$

where

$$\boldsymbol{T}_{r-1} := x \sum_{i=0}^{r-1} t_i q^i \boldsymbol{E}_q^i,$$

and

$$t_i := \sum_{j=i+1}^r (-1)^{r-j} S_j(\mathfrak{a}/q), \quad i = 0, 1, \dots, r-1,$$
$$t = \sum_{j=0}^r (-1)^{r-j} S_j(\mathfrak{a}/q) = \prod_{i=1}^r (a_i/q - 1).$$

Also, we have

$$\boldsymbol{E}_q^{s+1-r} = (\boldsymbol{E}_q - \boldsymbol{I}) \sum_{i=0}^{s-r} \boldsymbol{E}_q^i + \boldsymbol{I}.$$

Thus,

$$S_{r}E_{q}^{s+1-r} = \omega(E_{q} - I)\left\{T_{r-1}E_{q}^{s+1-r} + q^{r-s-1}t\sum_{i=0}^{s-r}E_{q}^{i}(xI)\right\} + q^{r-s-1}\omega xtI.$$
Using $Z = (q-1)^{-1}(E_{q} - I)$ and
$$E_{q}^{i} = \sum_{m=0}^{i} {i \choose m}(q-1)^{m}Z^{m},$$
(4.10)

we obtain

$$\boldsymbol{U}_{s} := \boldsymbol{T}_{r-1} \boldsymbol{E}_{q}^{s+1-r} = q^{r-s-1} \sum_{j=0}^{r-1} t_{j} \boldsymbol{E}_{q}^{j+s+1-r}(\boldsymbol{x}\boldsymbol{I}) = \sum_{m=0}^{s} u_{j} \boldsymbol{Z}^{j}(\boldsymbol{x}\boldsymbol{I}),$$

where

$$u_m := q^{r-s-1}(q-1)^m \sum_{h=\max\{0,m-s-1+r\}}^s \binom{h+s+1-r}{m} t_h$$

and

$$\bar{\boldsymbol{U}}_{s-r} := q^{r-s-1}t\sum_{i=0}^{s-r}\boldsymbol{E}_q^i(\boldsymbol{x}\boldsymbol{I}) = \sum_{m=0}^{s-r}\bar{\boldsymbol{u}}_m\boldsymbol{Z}^m(\boldsymbol{x}\boldsymbol{I}),$$

where

$$\bar{u}_m := tq^{r-s-1}(q-1)^m {s+1-r \choose m+1}, \quad m=0,1,\ldots,s-r.$$

Similarly, we write

$$\boldsymbol{V}_s = \sum_{m=0}^s \boldsymbol{v}_m \boldsymbol{Z}^m,$$

where

$$v_m := (q-1)^m \sum_{h=m}^s (-1)^{s-m} \binom{h}{m} S_h(\mathfrak{b}/q).$$

Summing up, we have

$$\begin{aligned} \boldsymbol{Q}_{\theta} &:= \frac{1}{(q-1)\omega x} \boldsymbol{P}_{\theta} \\ &= \boldsymbol{D}_{q} \left\{ \sum_{j=0}^{s} \boldsymbol{Z}^{j} \left(\left(u_{j} x - \frac{v_{j}}{\omega} \right) \boldsymbol{I} \right) + \sum_{m=0}^{s-r} \bar{u}_{m} \boldsymbol{Z}^{m}(x \boldsymbol{I}) \right\} + q^{r-s-1} \frac{t}{q-1} \boldsymbol{I}. \end{aligned}$$

Hence the result.

In the case r > s + 1, we start with the operator

$$\bar{\boldsymbol{P}}_{\theta} := \boldsymbol{E}_{q}^{r-s-1} \boldsymbol{P}_{\theta},$$

 P_{θ} being the operator given in (4.7), and proceed in a manner similar to that of the first part of the proof. \Box

Now we can prove the following result.

Theorem 4.2. The functions $F_k(\omega)$ (cf. (2.3)), satisfy the recurrence relation

$$\{\mathcal{M} + \omega^{-1}\mathcal{N}\}(h_k F_k(\omega)) = 0 \tag{4.11}$$

of order $\theta + 1$, where $\theta := \max(r, s + 1)$, the difference operators \mathcal{M} and \mathcal{N} are given by

$$\mathscr{M} := \left(\sum_{j=0}^{\theta-1} \hat{u}_j \mathscr{R}_{\theta-2,j} \mathscr{V}_j\right) \mathscr{X} + v \mathscr{T}_{\theta-1} \lambda_k^{-1} \mathscr{D}, \tag{4.12}$$

$$\mathscr{N} := \sum_{j=0}^{\theta-1} v_j \mathscr{R}_{\theta-2,j} \mathscr{V}_j \tag{4.13}$$

with

$$\hat{u}_j := \begin{cases} u_j + \bar{u}_j, & j = 0, 1, \dots, d - 1, \\ u_j, & j = \max(d, 0), \dots, \theta - 1, \end{cases} \qquad d = s + 1 - r$$

and the notation used is that of (3.17), (3.18), (3.27), and (4.3)–(4.6).

Proof. Eq. (4.2) implies

$$b_k[\boldsymbol{Q}_{\boldsymbol{\theta}}f] = 0.$$

Applying the operator $\mathscr{T}_{\theta-1}\lambda_k^{-1}\mathscr{D}$ to both sides of the above equation, and using Lemmata 3.1 and 3.2, we obtain

$$\left\{\sum_{j=0}^{\theta-1} \mathscr{R}_{\theta-2,j} \mathscr{V}_j(u_j \mathscr{X} - \omega^{-1} v_j \mathscr{I}) + \sum_{m=0}^{d-1} \bar{u}_m \mathscr{R}_{\theta-2,m} \mathscr{V}_m \mathscr{X} + u \mathscr{T}_{\theta-1} \lambda_k^{-1} \mathscr{D} \right\} b_k[f] = 0.$$

Hence follows (4.11) with \mathcal{M} and \mathcal{N} given by (4.12) and (4.13), respectively. \Box

Corollary 4.3. The recurrence (4.11) can be written in the form

$$\sum_{i=0}^{\theta+1} \left[A_i(k) + \frac{1}{\omega} B_i(k) \right] F_{k+i}(\omega) = 0,$$
(4.14)

where $A_i(k)$, $B_i(k)$ are rational functions of q^k , and $B_0(k) = B_{\theta+1}(k) \equiv 0$.

4.2. Special case

In the special case where r = s + 1 and $\omega = q$, the results given in the preceding section can be refined. Namely, we show that the functions $F_k(q)$ satisfy a recurrence in k of order s + 1 (see Theorem 4.5). We shall need the following result.

Lemma 4.4. Function

$$f(x) := {}_{s+1}\phi_s \left(\begin{array}{c} a_1, a_2, \dots, a_{s+1} \\ b_1, b_2, \dots, b_s \end{array} \middle| q; qx \right), \quad s \ge 1,$$
(4.15)

satisfies the difference equation

$$(\boldsymbol{Q}_{s+1}^*f)(x) \equiv \frac{1-q}{q} \boldsymbol{L}f(x) + \boldsymbol{D}_q \sum_{m=0}^{s-1} \boldsymbol{Z}^m[(\boldsymbol{u}_m^* x - \boldsymbol{v}_m^*)f(x)] + \boldsymbol{v}^*f(x) = 0,$$
(4.16)

where $\boldsymbol{Z} := x \boldsymbol{D}_q$, and

$$u_m^* := q(q-1)^m \sum_{j=m+2}^{s+1} (-1)^{j-1} {j-1 \choose m+1} S_j(\mathfrak{a}/q) + \delta_{m0} \,\alpha\beta,$$
(4.17)

$$v_m^* := (q-1)^m \sum_{j=m+1}^s (-1)^j \binom{j-1}{m} S_j(\mathfrak{b}/q) + \delta_{m0} \alpha, \quad m = 0, 1, \dots, s-1,$$
(4.18)

$$v^* := (-1)^s \frac{q}{q-1} \prod_{j=1}^{s+1} (a_j/q - 1) - \alpha\beta + 1.$$
(4.19)

Proof. First observe that the operator L introduced in (3.21) can be written as

$$(q-1)\boldsymbol{L}f(x) = \boldsymbol{D}_{q}(\sigma_{0}^{+}(x/q)f(x)) - \boldsymbol{D}_{q^{-1}}(\sigma_{0}(qx)f(x)) - q(\alpha\beta - 1)f(x),$$
(4.20)

where

$$\sigma_0(x) := \sigma(x)/x = x - 1, \qquad \sigma_0^+(x) := \sigma^+(x)/x = \alpha q(\beta q x - 1).$$
(4.21)

Let us define

$$\boldsymbol{Q}_{s+1}^* := (-1)^s \frac{1}{(q-1)x} \boldsymbol{E}_{q^{-1}} \boldsymbol{P}_{s+1}, \tag{4.22}$$

the operator P_{s+1} being defined as in (2.4) for r=s+1 and $\omega=q$. Then, by (4.8), (4.9) and (4.21), we have

$$\boldsymbol{Q}_{s+1}^{*}f(x) = \frac{1}{q}\boldsymbol{D}_{q^{-1}}(\sigma_{0}(qx)f(x)) + \frac{1}{(q-1)x}\sum_{j=0}^{s}(-1)^{j}x[S_{j+1}(\mathfrak{a}) - q\delta_{j0}]\boldsymbol{E}_{q}^{j}f(x)$$
$$-\boldsymbol{D}_{q}\sum_{j=0}^{s}(-1)^{j+1}S_{j+1}(\mathfrak{b}/q)\boldsymbol{E}_{q}^{j}f(x).$$

Hence,

$$\left\{\boldsymbol{Q}_{s+1}^{*} + \frac{q-1}{q}\boldsymbol{L}\right\}f(x) = \boldsymbol{T}_{s}f(x) - \boldsymbol{D}_{q}\boldsymbol{V}_{s}f(x) + \frac{1}{q}\boldsymbol{D}_{q}(\sigma_{0}^{+}(x/q)f(x)) - (\alpha\beta - 1)f(x), \quad (4.23)$$

where

$$T_{s}^{*} := \frac{1}{(q-1)x} \sum_{j=0}^{s} (-1)^{j} x [S_{j+1}(\mathfrak{a}) - q \delta_{j0}] E_{q}^{j},$$

$$V_{s}^{*} := \sum_{j=0}^{s-1} (-1)^{j+1} S_{j+1}(\mathfrak{b}/q) E_{q}^{j}.$$
(4.24)
(4.25)

Now, we can write

$$\boldsymbol{T}_{s}^{*}f(x) = \frac{1}{(q-1)x} \sum_{j=0}^{s} (-1)^{j} q^{-j} [S_{j+1}(\mathfrak{a}) - q\delta_{j0}] \boldsymbol{E}_{q}^{j}(xf(x))$$
$$= \boldsymbol{D}_{q} \boldsymbol{U}_{s}^{*}(xf(x)) + v^{*}f(x), \qquad (4.26)$$

where

$$\boldsymbol{U}^*_s := \sum_{i=0}^{s-1} t^*_i \boldsymbol{E}^i_q$$

with

$$t_i^* := q \sum_{j=i+2}^{s+1} (-1)^{j-1} S_j(\mathfrak{a}/q), \quad i = 0, 1, \dots, s-1,$$
(4.27)

and

$$v^* := (-1)^s \frac{q}{q-1} \prod_{i=1}^{s+1} (a_i/q-1).$$

Now, on using (4.10), we obtain

$$\boldsymbol{U}_{s}^{*} = \sum_{m=0}^{s-1} u_{m}^{*} \boldsymbol{Z}^{m}, \qquad \boldsymbol{V}_{s}^{*} = \sum_{m=0}^{s-1} v_{m}^{*} \boldsymbol{Z}^{m}$$
(4.28)

for

$$u_m^* := q(q-1)^m \sum_{i=m+1}^s \binom{i-1}{m} \sum_{j=i+1}^{s+1} (-1)^{j-1} S_j(\mathfrak{a}/q)$$

$$= q(q-1)^{m} \sum_{j=m+2}^{s+1} (-1)^{j-1} {j-1 \choose m+1} S_{j}(\mathfrak{a}/q), \qquad (4.29)$$

$$v_m^* := (q-1)^m \sum_{j=m+1}^s (-1)^j \binom{j-1}{m} S_j(\mathfrak{b}/q), \quad m = 0, 1, \dots, s-1.$$
(4.30)

Now, using (4.24)-(4.26), (4.28)-(4.30), in (4.23), the result follows. \Box

Theorem 4.5. Let $s \ge 1$. The functions

$$F_{k}(q) := \frac{\prod_{i=1}^{s+1} (a_{i};q)_{k}}{(q;q)_{k} \prod_{j=1}^{s} (b_{j};q)_{k}} q^{k}{}_{s+2} \phi_{s+1} \begin{pmatrix} a_{1}q^{k}, \dots, a_{s+1}q^{k}, \alpha q^{k+1} \\ b_{1}q^{k}, \dots, b_{s}q^{k}, \alpha \beta q^{2k+2} \end{pmatrix} q;q \end{pmatrix}$$

(cf. (2.3)) satisfy the recurrence relation

$$\mathscr{L}^*(h_k F_k(q)) = 0 \tag{4.31}$$

of order s + 1, where the difference operator \mathscr{L}^* is given by

$$\mathscr{L}^* := \sum_{m=0}^{s-1} \mathscr{R}_{s-2,m} \mathscr{V}_m(u_m^* \mathscr{X} - v_m^* \mathscr{I}) + \mathscr{T}_{s-1} \lambda_k^{-1} \mathscr{D}(\mu_k \mathscr{I})$$
(4.32)

with $\mu_k := v^* - \lambda_k (q-1)/q$, and the notation used is that of (3.17), (3.18), (3.27), and (4.17)–(4.19).

Proof. On applying the operator $\mathscr{T}_{s-1}\lambda_k^{-1}\mathscr{D}$ to both sides of

 $b_k[\boldsymbol{Q}_{s+1}^*f] = 0$

and using Lemmata 3.1 and 3.2, we obtain $\mathscr{L}^* b_k[f] = 0$ with \mathscr{L}^* given by (4.32). Hence the result. \Box

Corollary 4.6. The recurrence (4.31) can be written in the form

$$\sum_{i=0}^{s+1} C_i(k) F_{k+i}(q) = 0$$
(4.33)

with rational coefficients in q^k .

5. Examples

5.1. First example

For r = s + 1 = 2, Theorem 4.2. implies the *third-order* recurrence relation $\{\mathscr{G} + \omega^{-1}\mathscr{H}\}(h_k F_k(\omega)) = 0$

for the quantities

$$F_{k}(\omega) := \frac{(a_{1};q)_{k}(a_{2};q)_{k}}{(q;q)_{k}(b;q)_{k}} \omega^{k}{}_{3}\phi_{2} \left(\begin{array}{c} a_{1}q^{k}, a_{2}q^{k}, \alpha q^{k+1} \\ bq^{k}, \alpha \beta q^{2k+2} \end{array} \middle| q; \omega, \right),$$

where

$$\begin{split} \mathscr{G} &:= \{ \gamma_0 \mathscr{P}_0 + \gamma_1 \mathscr{Q}_0 \} \mathscr{X} + \gamma \mathscr{P}_0 \lambda_k^{-1} \mathscr{D}, \\ \mathscr{H} &:= \eta_0 \mathscr{P}_0 + \eta_1 \mathscr{Q}_0 \end{split}$$

and

$$\gamma_0 := 2a_1a_2 - q(a_1 + a_2), \quad \gamma_1 := (q - 1)a_1a_2, \quad \gamma := (a_1 - q)(a_2 - q)/(q - 1),$$

$$\eta_0 := q(q-b), \quad \eta_1 := -bq(q-1).$$

Here the difference operators \mathscr{X} , \mathscr{D} are defined by (3.17), (3.28), respectively, and

$$\mathscr{P}_0 := \pi_{01}(k)\mathscr{I} + \pi_{02}(k)\mathscr{E}, \tag{5.1}$$

$$\mathcal{Q}_0 := \omega_{01}(k)\mathscr{I} + \omega_{02}(k)\mathscr{E}$$
(5.2)

with

$$\pi_{01}(k) := \frac{\alpha(q^{k}-1)(\beta q^{k+1}-1)(\alpha \beta q^{k+1}-1)^{2}}{(q-1)(\alpha \beta q^{2k+1}-1)(\alpha \beta q^{2k+2}-1)}, \quad \pi_{02}(k) := -\frac{\alpha \beta(q^{k+1}-1)(\alpha \beta q^{k+2}-1)}{q-1},$$
$$\omega_{01}(k) := -\frac{\alpha q^{k}(q^{k}-1)(\beta q^{k+1}-1)(\alpha \beta q^{2k+2}-1)}{(\alpha \beta q^{2k+1}-1)(\alpha \beta q^{2k+2}-1)}, \quad \omega_{02}(k) := -\frac{\alpha \beta q^{k+2}-1}{q}$$

(cf. (3.25); notice cancellation of a common factor).

Note that

$$\begin{split} \mathscr{G} &= c_{10}(k)\mathscr{E}^{-1} + (c_{11} + c_{20})(k)\mathscr{I} + (c_{12} + c_{21})(k)\mathscr{E} + c_{22}(k)\mathscr{E}^2, \\ \mathscr{H} &= d_1(k)\mathscr{I} + d_2(k)\mathscr{E}, \end{split}$$

where

$$\begin{aligned} c_{ij}(k) &:= (\gamma_0 \pi_{0i}(k) + \gamma_1 \omega_{0i}(k)) \xi_j(k+i-1) + \gamma \pi_{0i}(k) \delta_j(k+i-1) / \lambda_{k+i-1}, \\ i &= 1, 2; \ j = 0, 1, 2, \\ d_m(k) &:= \eta_0 \pi_{0m}(k) + \eta_1 \omega_{0m}(k), \quad m = 1, 2. \end{aligned}$$

5.2. Second example

Theorem 4.5. implies that the functions

$$F_{k}(q) := \frac{(a_{1};q)_{k}(a_{2};q)_{k}}{(q;q)_{k}(b;q)_{k}} q^{k}{}_{3}\phi_{2} \left(\begin{array}{c} a_{1}q^{k}, a_{2}q^{k}, \alpha q^{k+1} \\ bq^{k}, \alpha \beta q^{2k+2} \end{array} \middle| q;q \right)$$
(5.3)

satisfy the second-order recurrence relation

$$\{\mathscr{D}(\mu_k\mathscr{I}) + \lambda_k(u_0^*\mathscr{X} - v_0^*\mathscr{I})\}(h_kF_k(q)) = 0$$

with $\mu_k := v^* - (1-q)\lambda_k/q$, and

$$u_0^*(x) := -a_1 a_2/q + \alpha \beta,$$

$$v_0^*(x) := \alpha - b/q,$$

 $v^* := -(a_1 - q)(a_2 - q)/[q(q - 1)] - \alpha\beta + 1$

or, in the scalar form,

$$A_0^*(k)F_{k-1}(q) + A_1^*(k)F_k(q) + A_2^*(k)F_{k+1}(q) = 0,$$
(5.4)

where

$$\begin{split} A_0^*(k) &:= (a_1q^k - q)(a_2q^k - q)(\alpha\beta q^{2k} - 1) \\ &\times (\alpha\beta q^{2k+1} - 1)(\alpha\beta q^{2k+2} - 1)^2(\alpha\beta q^{2k+3} - 1), \\ A_1^*(k) &:= q^k(q^k - 1)(\alpha\beta q^{2k+1} - 1)(\alpha\beta q^{2k+2} - 1)(\alpha\beta q^{2k+3} - 1) \\ &\times \{q^{k+1}\alpha(a_1 + a_2)[a_2q^k(\alpha q^{k+1} - 1) - (\beta q^{k+1} - 1)] \\ &+ q^k(a_1a_2 + \alpha\beta q^2)[a_1q^k(\beta q^{k+1} - 1) - (\alpha q^{k+1} - 1)] \\ &- b(\alpha\beta q^{2k} - 1)(\alpha\beta q^{2k+2} - 1)\}, \\ A_2^*(k) &:= -\alpha q^{3k+1}(q^k - 1)(q^{k+1} - 1)(\alpha\beta q^{k+1} - 1)(\beta q^{k+1} - 1) \\ &\times (\alpha\beta q^{2k} - 1)(\alpha\beta q^{k+2} - a_1)(\alpha\beta q^{k+2} - a_2). \end{split}$$

Now, replace α by q^{α} , β by q^{β} , a_i by q_i^a , b by q^b , and let $q \to 1$. The limit form of (5.3) is

$$\bar{U}_k := \frac{(a_1)_k(a_2)_k}{k!(b)_k} {}_3F_2 \left(\begin{array}{c} a_1 + k, a_2 + k, \alpha + k + 1 \\ b + k, \alpha + \beta + 2 + 2k \end{array} \right| 1 \right)$$

and the limit form of Eq. (5.4) agrees with the equation predicted for \bar{U}_k by Wimp's theorem (cf. [10]).

5.3. Third example

By virtue of Theorem 4.2, the functions

$$F_{k}(\omega) := \frac{(a;q)_{k}}{(q;q)_{k}(b_{1};q)_{k}(b_{2};q)_{k}} \omega^{k} q^{k(k-1)}{}_{2} \phi_{3} \left(\begin{array}{c} aq^{k}, \alpha q^{k+1} \\ b_{1}q^{k}, b_{2}q^{k}, \alpha \beta q^{2k+2} \end{array} \middle| q; \omega q^{2k} \right)$$

satisfy the *fourth-order* recursion

$$\{\mathcal{M}+\omega^{-1}\mathcal{N}\}(h_kF_k(\omega))=0,$$

where

$$\mathcal{M} := \{\mu_0 \mathcal{P}_1 \mathcal{P}_0 + \mu_1 \mathcal{P}_1 \mathcal{Q}_0 + \mu_2 \mathcal{Q}_1 \mathcal{Q}_0\} \mathcal{X} + \mu \mathcal{P}_1 \mathcal{P}_0 \lambda_k^{-1} \mathcal{Q},$$
$$\mathcal{N} := \mathcal{P}_1 \{v_0 \mathcal{P}_0 + v_1 \mathcal{Q}_0\} + v_2 \mathcal{Q}_1 \mathcal{Q}_0$$

and

$$\mu_0 := 3a - 2q, \quad \mu_1 := (q - 1)(3a - q), \quad \mu_2 := (q - 1)^2 a,$$

$$v_0 := q(b_1 - q)(b_2 - q), \quad v_1 := q(q - 1)[b_1(b_2 - q) + b_2(b_1 - q)],$$

$$w_2 := q(q-1)^2 b_1 b_2, \quad \mu := (a-q)/(q-1).$$

The difference operators \mathscr{X} , \mathscr{D} , \mathscr{P}_0 , and \mathscr{Q}_0 are defined by (3.17), (3.18), (5.1) and (5.2), respectively, and the operators \mathscr{P}_1 and \mathscr{Q}_1 are given by

$$\begin{split} \mathscr{P}_{1} &= -\frac{\alpha(q^{k+1}-1)(\beta q^{k+2}-1)(\alpha \beta q^{k+1}-1)}{q^{k-1}(q-1)(\alpha \beta q^{2k+3}-1)}\mathscr{I} + \frac{\alpha \beta(q^{k+2}-1)(\alpha \beta q^{2k+2}-1)}{q^{k-1}(q-1)}\mathscr{E}, \\ \mathscr{Q}_{1} &= \frac{\alpha q(q^{k+1}-1)(\beta q^{k+2}-1)}{\alpha \beta q^{2k+3}-1}\mathscr{I} + \frac{\alpha \beta q^{2k+2}-1}{q^{k}}\mathscr{E} \end{split}$$

(cf. (3.25); note cancellation of common factors which simplifies the formulae).

References

- [1] G. Gasper, M. Rahman, Basic Hypergeometric Series, Cambridge University Press, Cambridge, 1990.
- [2] W. Hahn, Über Orthogonalpolynome, die q-Differenzengleichungen genügen, Math. Nachr. 2 (1949) 4-34.
- [3] R. Koekoek, R.F. Swarttouw, The Askey scheme of hypergeometric orthogonal polynomials and its q-analogue, Report 98-17, Fac. Techn. Math. Informatics, Delft Univ. of Technology, Delft, 1998.
- [4] T. Koornwinder, Compact quantum groups and q-special functions, in: V. Baldoni, M.A. Picardello (Eds.), Representations of Lie groups and quantum groups, Pitman Research Notes in Mathematics Series, Vol. 311, Longman Scientific & Technical, New York, 1994, pp. 46–128.
- [5] S. Lewanowicz, Recurrence relations for hypergeometric functions of unit argument, Math. Comput. 45 (1985) 521–535, Errata 48 (1987) 853.
- [6] S. Lewanowicz, E. Godoy, I. Area, A. Ronveaux, A. Zarzo, Recurrence relations for the coefficients of the Fourier series expansions with respect to *q*-classical orthogonal polynomials, Numer. Algorithms 23 (2000) 31–50.
- [7] Y.L. Luke, The Special Functions and Their Approximations, Academic Press, New York, 1969.
- [8] J.C. Medem, Polinomios ortogonales q-semiclásicos, Ph.D. Dissertation, Universidad Politécnica de Madrid, Madrid, 1996.
- [9] A. Verma, Certain expansions of the basic hypergeometric functions, Math. Comput. 20 (1966) 151–157.
- [10] J. Wimp, Recursion formulae for hypergeometric functions, Math. Comput. 22 (1968) 363-373.



Journal of Computational and Applied Mathematics 121 (2000) 313-330

JOURNAL OF COMPUTATIONAL AND APPLIED MATHEMATICS

www.elsevier.nl/locate/cam

The left-definite spectral theory for the classical Hermite differential equation $\stackrel{\text{\tiny{\scale}}}{\rightarrow}$

W.N. Everitt^a, L.L. Littlejohn^{b,*}, R. Wellman^b

^aSchool of Mathematics and Statistics, University of Birmingham, Edgbaston, Birmingham B15 2TT, England, UK ^bDepartment of Mathematics and Statistics, Utah State University, Logan, Utah, 84322-3900, USA

Received 4 July 1999

Dedicated to Professor A.M. Krall upon his retirement

Abstract

In this paper, we develop the left-definite spectral theory associated with the self-adjoint operator A in $L^2((-\infty,\infty); \exp(-t^2))$, generated from the classic second-order Hermite differential equation

$$\ell_H[y](t) = -y'' + 2ty' + ky = \lambda y \quad (t \in (-\infty, \infty))$$

that has the Hermite polynomials $\{H_m(t)\}_{m=0}^{\infty}$ as eigenfunctions. More specifically, for each $n \in \mathbb{N}$, we explicitly determine the unique left-definite Hilbert–Sobolev space W_n and associated inner product $(\cdot, \cdot)_n$, which is generated from the *n*th integral power $\ell_H^n[\cdot]$ of $\ell_H[\cdot]$. Moreover, for each $n \in \mathbb{N}$, we determine the corresponding unique left-definite self-adjoint operator A_n in W_n and characterize its domain in terms of another left-definite space. As a consequence of this, we explicitly determine the domain of each integral power of A and, in particular, we obtain a new characterization of the domain of the classical right-definite operator A. © 2000 Elsevier Science B.V. All rights reserved.

MSC: 33C65; 34B20; 34B30; 47B25; 47B65

Keywords: Spectral theorem; Left-definite Sobolev space; Left-definite self-adjoint operator; Lagrangian symmetric; Hermite polynomials; Stirling numbers of the second kind

1. Introduction

When A is an unbounded self-adjoint operator in a Hilbert space $(H, (\cdot, \cdot))$ that is bounded below by a positive multiple of the identity operator, the authors in [4] show that there is a continuum of

 $^{^{\}pm}$ Research is partially supported from the National Science Foundation under grant numbers DMS-9970478 (W.N.E. and L.L.L.) and IMD-9919107 (R.W.).

^{*} Corresponding author.

E-mail address: lance@math.usu.edu (L.L. Littlejohn).

unique Hilbert spaces $\{(W_r, (\cdot, \cdot)_r)\}_{r>0}$ and, for each r > 0, a unique self-adjoint restriction A_r of A in W_r . The Hilbert space W_r is called the *r*th *left-definite Hilbert space* associated with the pair (H, A) and the operator A_r is called the *r*th *left-definite operator* associated with (H, A).

Left-definite spectral theory has its roots in the classic treatise of Weyl [12] on the theory of formally symmetric second-order differential expressions. We remark, however, that even though our motivation for the general left-definite theory developed in [4] arose through our interest in certain self-adjoint differential operators, the theory developed in [4] can be applied to an arbitrary self-adjoint operator that is bounded below.

The terminology *left-definite* is due to Schäfke and Schneider (who used the German *Links-definit*) [9] in 1965 and describes one of the Hilbert space settings in which certain formally symmetric differential expressions can be analyzed. For example, consider the differential equation

$$S[y](t) = \lambda w(t)y(t) \quad (t \in I; \ \lambda \in \mathbb{C}),$$
(1.1)

where I = (a, b) is an open interval of the real line \mathbb{R} , w is Lebesgue measurable, locally integrable and positive almost everywhere on I, and where $S[\cdot]$ is the formally symmetric differential expression

$$S[y](t) = \sum_{j=0}^{n} (-1)^{j} (b_{j}(t) y^{(j)}(t))^{(j)} \quad (t \in I),$$

with *nonnegative*, infinitely differentiable coefficients $b_j(t)$ (j = 0, 1, ..., n) on I (as in all known cases when (1.1) has a sequence of orthogonal polynomial solutions). The classical Glazman–Krein–Naimark theory (see [5]) applies to (1.1) and describes all self-adjoint extensions of the minimal operator T_{\min} , generated by $w^{-1}S[\cdot]$, in the weighted Hilbert space $L^2_w(I)$ of all Lebesgue measurable functions $f: I \to \mathbb{C}$ with inner product

$$(f,f) = \int_I |f(t)|^2 w(t) \,\mathrm{d}t < \infty.$$

Due to the appearance of w on the right-hand side of (1.1), the space $L^2_w(I)$ is called the *right-definite Hilbert space* for $w^{-1}S[\cdot]$. On the other hand, spectral properties of the differential expression $w^{-1}S[\cdot]$ can also be studied in a Hilbert space W generated by the Sobolev inner product

$$(f,g)_W = \sum_{j=0}^n b_j(t) f^{(j)}(t) \bar{g}^{(j)}(t) \quad (f,g \in W),$$

called the *Dirichlet inner product*, which arises naturally in connection to Green's formula for the expression $S[\cdot]$. Since this inner product is generated by the left-hand side of (1.1), we call the spectral study of $w^{-1}S[\cdot]$ in W a *left-definite spectral setting* and call W a *left-definite Hilbert space*.

In this paper, we apply this left-definite theory to the self-adjoint Hermite differential operator A, generated by the classical second-order formally symmetric Hermite differential expression

$$\ell_H[y](t) := -y''(t) + 2ty'(t) + ky(t)$$

= $\exp(t^2)(-(\exp(-t^2)y'(t))' + k\exp(-t^2)y(t)) \quad (t \in \mathbb{R} = (-\infty, \infty))$ (1.2)

and having the Hermite polynomials as eigenfunctions. Here, k is a fixed, positive constant. The right-definite setting in this case is the Hilbert space $L^2((-\infty, \infty); \exp(-t^2))$ of Lebesgue measurable

functions $f: (-\infty, \infty) \to \mathbb{C}$ satisfying $||f|| < \infty$, where $|| \cdot ||$ is the norm generated by the inner product

$$(f,g) := \int_{-\infty}^{\infty} f(t)\overline{g}(t)\exp(-t^2) dt \quad (f,g \in H).$$

$$(1.3)$$

Even though the theory developed in [4] guarantees the existence of a continuum of spaces $\{W_r\}_{r>0}$ and left-definite operators $\{A_r\}_{r>0}$ (they are all *differential* operators), we can only effectively determine the left-definite spaces, their inner products, and the domains of the left-definite operators when r is a positive integer; reasons for this will be made clear in the analysis below.

The contents of this paper are as follows. In Section 2, we state some of the main results developed in [4]. In Section 3, we review some of the properties of the Hermite differential equation, the Hermite polynomials and the right-definite self-adjoint operator A, generated by the second-order Hermite expression (1.2), having the Hermite polynomials as eigenfunctions. Also in this section, we obtain the formally symmetric form of each integral power of the second-order Hermite expression; as we shall see, these higher-order expressions are key to determining the various left-definite inner products. Interestingly, these powers involve the Stirling numbers of the second kind. Lastly, in Section 4, we establish the left-definite theory for the Hermite expression (1.2). Specifically, we determine explicitly

- (a) the sequence $\{W_n\}_{n=1}^{\infty}$ of left-definite spaces associated with the pair $(L^2((-\infty,\infty); \exp(-t^2)), A),$
- (b) the sequence of left-definite self-adjoint operators $\{A_n\}_{n=1}^{\infty}$, and their domains $\mathscr{D}(A_n)\}_{n=1}^{\infty}$, associated with $(L^2((-\infty,\infty); \exp(-t^2)), A)$, and
- (c) the domains $\mathscr{D}(A^n)$ of each integral power A^n of A.

These results culminate in Theorem 4.5. An application of this theorem yields a new result (see Corollary 4.5) concerning the characterization of functions in the domain of the right-definite operator A.

2. Left-definite Hilbert spaces and left-definite operators

Let V denote a vector space (over the complex field \mathbb{C}) and suppose that (\cdot, \cdot) is an inner product with norm $|| \cdot ||$ generated from (\cdot, \cdot) such that $H = (V, (\cdot, \cdot))$ is a Hilbert space. Suppose V_r (the subscripts will be made clear shortly) is a linear manifold of the vector space V and let $(\cdot, \cdot)_r$ and $|| \cdot ||_r$ denote an inner product (quite possibly different from (\cdot, \cdot)) and associated norm, respectively, over V_r . We denote the resulting inner product space by $W_r = (V_r, (\cdot, \cdot)_r)$.

Throughout this paper, we assume that $A: \mathscr{D}(A) \subset H \to H$ is a self-adjoint operator that is bounded below by kI, for some k > 0, that is

$$(Ax,x) \ge k(x,x) \quad (x \in \mathcal{D}(A)),$$

that is to say, A is bounded below in H by kI, where I is the identity operator. It follows that A^r , for each r > 0, is a self-adjoint operator that is bounded below in H by $k^r I$.

We now make the definitions of left-definite spaces and left-definite operators.

Definition 2.1. Let r > 0 and suppose V_r is a linear manifold of the Hilbert space $H = (H, (\cdot, \cdot))$ and $(\cdot, \cdot)_r$ is an inner product on V_r . Let $W_r = (V_r, (\cdot, \cdot)_r)$. We say that W_r is an *r*th *left-definite space* associated with the pair (H, A) if each of the following conditions hold:

- (1) W_r is a Hilbert space,
- (2) $\mathscr{D}(A^r)$ is a linear manifold of V_r ,
- (3) $\mathscr{D}(A^r)$ is dense in W_r ,
- (4) $(x,x)_r \ge k^r(x,x)$ $(x \in V_r)$, and
- (5) $(x, y)_r = (A^r x, y) \quad (x \in \mathscr{D}(A^r), y \in V_r).$

It is not clear, from the definition, if such a self-adjoint operator A generates a left-definite space for a given r > 0. However, in [4], the authors prove the following theorem; the Hilbert space spectral theorem plays a prominent role in establishing this result.

Theorem 2.2 (see Littlejohn and Wellman [4, Theorem 3.1]). Suppose $A : \mathscr{D}(A) \subset H \to H$ is a selfadjoint operator that is bounded below by kI, for some k > 0. Let r > 0. Define $W_r = (V_r, (\cdot, \cdot)_r)$ by

$$V_r = \mathscr{D}(A^{r/2}) \tag{2.1}$$

and

$$(x, y)_r = (A^{r/2}x, A^{r/2}y) \quad (x, y \in V_r).$$

Then W_r is a left-definite space associated with the pair (H, A^r) . Moreover, suppose $W_r := (V_r, (\cdot, \cdot)_r)$ and $W'_r := (V'_r, (\cdot, \cdot)'_r)$ are rth left-definite spaces associated with the pair (H, A). Then $V_r = V'_r$ and $(x, y)_r = (x, y)'_r$ for all $x, y \in V_r = V'_r$, i.e., $W_r = W'_r$. That is to say, $W_r = (V_r, (\cdot, \cdot)_r)$ is the unique left-definite space associated with (H, A).

Definition 2.3. For r > 0, let $W_r = (V_r, (\cdot, \cdot)_r)$ denote the *r*th left-definite space associated with (H, A). If there exists a self-adjoint operator $A_r : \mathscr{D}(A_r) \subset W_r \to W_r$ that is a restriction of A; that is to say

 $A_r f = A f \quad (f \in \mathscr{D}(A_r) \subset \mathscr{D}(A)),$

we call such an operator a rth left-definite operator associated with (H,A).

Again, it is not immediately clear that such an A_r exists; in fact, however, A_r exists and is unique.

Theorem 2.4 (see Littlejohn and Wellman [4, Theorem 3.2]). Suppose A is a self-adjoint operator in a Hilbert space H that is bounded below by kI, for some k > 0. For any r > 0, let $W_r = (V_r, (\cdot, \cdot)_r)$ be the rth left-definite space associated with (H, A). Then there exists a unique left-definite operator A_r in W_r associated with (H, A). Moreover,

$$\mathscr{D}(A_r) = V_{r+2}.$$

The last theorem that we state in this section shows that the point spectrum, continuous spectrum, and resolvent set of a self-adjoint operator A and each of its associated left-definite operators A_r (r > 0) are identical. We recall (see [3, Chapter 7]) that:

- (i) the *point spectrum* $\sigma_p(A)$ of A consists of all $\lambda \in \mathbb{C}$ such that $R_{\lambda}(A) := (A \lambda I)^{-1}$ does not exist;
- (ii) the *continuous spectrum* $\sigma_c(A)$ of A consists of all $\lambda \in \mathbb{C}$ such that $R_{\lambda}(A)$ exists with a dense domain but is an unbounded operator;
- (iii) the *resolvent set* $\rho(A)$ of A consists of all $\lambda \in \mathbb{C}$ such that $R_{\lambda}(A)$ exists with a dense domain and is a bounded operator;

Moreover, for a self-adjoint operator A, we remark that \mathbb{C} is the disjoint union of $\sigma_p(A)$, $\sigma_c(A)$, and $\rho(A)$.

Theorem 2.5 (see Littlejohn and Wellman [4, Theorem 3.6]). For each r > 0, let A_r denote the rth left-definite operator associated with the self-adjoint operator A that is bounded below by kI, where k > 0. Then

- (a) the point spectra of A and A_r coincide, i.e., $\sigma_p(A_r) = \sigma_p(A)$;
- (b) the continuous spectra of A and A_r coincide, i.e., $\sigma_c(A_r) = \sigma_c(A)$;
- (c) the resolvent sets of A and A_r are equal, i.e., $\rho(A_r) = \rho(A)$.

We refer the reader to [4] for other results established on left-definite theory for self-adjoint operators A that are bounded below.

3. Preliminary results on the Hermite differential equation

When $\lambda = 2m + k$, where $m \in \mathbb{N}_0$, the Hermite equation $\ell_H[y](t) = (\lambda + k)y(t)$, where $\ell_H[\cdot]$ is defined in (1.2), has a polynomial solution $H_m(t)$ of degree *m*; the sequence of polynomials $\{H_m(t)\}_{m=0}^{\infty}$ is called the *Hermite* polynomials. These polynomials form a complete orthogonal set in the Hilbert space $L^2((-\infty, \infty); \exp(-t^2))$ of Lebesgue measurable functions $f: (-\infty, \infty) \to \mathbb{C}$ satisfying $||f|| < \infty$, where $||\cdot||$ is the norm generated from the inner product (\cdot, \cdot) , defined by

$$(f,g) := \int_{-\infty}^{\infty} f(t)\overline{g}(t)\exp(-t^2) dt \quad (f,g \in L^2_H(-\infty,\infty)).$$
(3.1)

In fact, with the *m*th Hermite polynomial defined by

$$H_m(t) = \frac{(m!)^{1/2} 2^{m/2}}{\pi^{1/4}} \sum_{j=0}^{\lfloor m/2 \rfloor} \frac{(-1)^j}{2^{2j} (m-2j)! j!} t^{m-2j} \quad (m \in \mathbb{N}_0),$$

it is the case that the sequence $\{H_m(t)\}_{m=0}^{\infty}$ is orthonormal in $L^2((-\infty,\infty); \exp(-t^2))$; that is,

$$(H_m, H_r) = \delta_{m,r} \quad (m, r \in \mathbb{N}_0), \tag{3.2}$$

where $\delta_{m,r}$ is the Kronecker delta function. We refer the reader to [7, Chapter 12] or [10, Chapter V] for various properties of the Hermite polynomials. The derivatives of these polynomials satisfy the

identity

$$\frac{d^{j}(H_{m}(t))}{dt^{j}} = 2^{j/2} (P(m,j))^{1/2} H_{m-j}(t) \quad (m,j \in \mathbb{N}_{0}),$$
(3.3)

where

$$P(m,j) := m(m-1)\dots(m-j+1) \quad (m,j \in \mathbb{N}_0; \ j \le m).$$
(3.4)

From (3.2) and (3.3), we see that

$$\int_{-\infty}^{\infty} \frac{d^{j}(H_{m}(t))}{dt^{j}} \frac{d^{j}(H_{r}(t))}{dt^{j}} \exp(-t^{2}) dt = 2^{j} P(m, j) \delta_{m, r} \quad (m, r, j \in \mathbb{N}_{0}).$$
(3.5)

The maximal domain Δ of $\exp(-t^2)\ell_H[\cdot]$ in $L^2((-\infty,\infty);\exp(-t^2))$ is defined to be

$$\Delta = \{ f \in L^2((-\infty,\infty); \exp(-t^2)) | f, f' \in AC_{\text{loc}}(-\infty,\infty);
\ell_H[f] \in L^2((-\infty,\infty); \exp(-t^2)) \}.$$
(3.6)

For functions $f, g \in \Delta$ and $[a, b] \subset \mathbb{R}$, we have Dirichlet's formula

$$\int_{a}^{b} \ell_{H}[f](t)\bar{g}(t)\exp(-t^{2}) dt = -\exp(-t^{2})f'(t)\bar{g}(t)|_{a}^{b} + \int_{a}^{b} [f'(t)\bar{g}'(t)\exp(-t^{2}) + kf(t)\bar{g}(t)\exp(-t^{2})] dt.$$
(3.7)

It is well known (for example, see, [6,11]) that $\exp(-t^2)\ell_H[\cdot]$ is strong limit point at $t = \pm \infty$:

$$\lim_{t \to \pm \infty} \exp(-t^2) f'(t) \bar{g}(t) = 0 \quad (f, g \in \Delta)$$
(3.8)

and Dirichlet at $t = \pm \infty$:

$$\int_{0}^{\infty} |f'(t)|^{2} \exp(-t^{2}) dt, \qquad \int_{-\infty}^{0} |f'(t)|^{2} \exp(-t^{2}) dt < \infty \quad (f \in \Delta).$$
(3.9)

From the Glazman-Krein-Naimark theory (see [5, Theorem 4, pp. 79-80]), it follows that the operator (in the terminology of this paper, the right-definite operator)

$$A: L^2((-\infty,\infty); \exp(-t^2)) \to L^2((-\infty,\infty); \exp(-t^2)),$$

defined by

$$(Af)(t) = \ell_H[f](t) \quad (f \in \mathscr{D}(A); \text{ a.e. } t \in (-\infty, \infty))$$
(3.10)

with domain

$$\mathscr{D}(A) = \Delta \tag{3.11}$$

is self-adjoint and has the Hermite polynomials as a complete set of eigenfunctions (see also [2, Appendix II, pp. 210-211;11]); moreover, the spectrum of A is given by

$$\sigma(A) = \{2m + k \mid m \in \mathbb{N}_0\}. \tag{3.12}$$

319

From (3.7)-(3.9), it follows that

$$(Af, f) = \int_{-\infty}^{\infty} \left[|f'(t)|^2 \exp(-t^2) + k |f(t)|^2 \exp(-t^2) \right] dt \ge k(f, f) \quad (f \in \mathcal{D}(A)),$$
(3.13)

that is, A is bounded below in $L^2((-\infty,\infty); \exp(-t^2))$ by kI. It is this inequality that explains the importance of the term ky(t) in (1.2). Consequently, we can apply Theorems 2.1–2.3. Notice that $(\cdot, \cdot)_1$, defined by

$$(f,g)_1 = \int_{-\infty}^{\infty} \left[f'(t)\overline{g}'(t)\exp(-t^2) + kf(t)\overline{g}(t)\exp(-t^2) \right] \mathrm{d}t \quad (f,g \in \mathcal{D}(A))$$

is an inner product; in fact, it is the inner product for the first left-definite space associated with the pair $(L^2((-\infty,\infty); \exp(-t^2)), A)$. Moreover, the closure of $\mathcal{D}(A)$ in the topology generated from this inner product is the first left-definite space W_1 associated with the pair $(L^2((-\infty,\infty); \exp(-t^2)), A)$.

We now turn our attention to the explicit construction of the sequence of left-definite inner products $(\cdot, \cdot)_n$ $(n \in \mathbb{N})$ associated with $(L^2((-\infty, \infty); \exp(-t^2)), A)$. As we shall see, these are generated from the integral powers $\ell_H^n[\cdot]$ $(n \in \mathbb{N})$ of the Hermite expression $\ell_H[\cdot]$, inductively given by

$$\ell_{H}^{1}[y] = \ell_{H}[y], \quad \ell_{H}^{2}[y] = \ell_{H}(\ell_{H}[y]), \dots, \ell_{H}^{n}[y] = \ell_{H}(\ell_{H}^{n-1}[y]) \quad (n \in \mathbb{N}).$$

A key to the explicit determination of these powers are certain numbers $\{b_j(n,k)\}_{j=0}^n$ which we now define.

Definition 3.1. For $n \in \mathbb{N}$ and $j \in \{0, 1, \dots, n\}$, let

$$b_j(n,k) := \sum_{i=0}^{j} \frac{(-1)^{i+j}}{j!} {j \choose i} (k+i)^n.$$
(3.14)

If we expand the term $(k+i)^n$ in (3.14) and switch the order of summation, we find that

$$b_{j}(n,k) = \sum_{m=0}^{n} \left(\sum_{i=0}^{j} \frac{(-1)^{i+j}}{j!} {j \choose i} i^{n-m} \right) {n \choose m} k^{m}$$

= $\sum_{m=0}^{n} {n \choose m} S_{n-m}^{(j)} k^{m},$ (3.15)

where

$$S_n^{(j)} = \sum_{i=0}^j \frac{(-1)^{i+j}}{j!} \binom{j}{i} i^n \quad (n, j \in \mathbb{N}_0)$$
(3.16)

is the Stirling number of the second kind. By definition, $S_n^{(j)}$ is the number of ways of partitioning *n* elements into *j* nonempty subsets (in particular, $S_0^j = 0$ for any $j \in \mathbb{N}$); we refer the reader to [1, pp. 824–825] for various properties of these numbers. Consequently, we see that

$$b_0(n,k) = \begin{cases} 0 & \text{if } k = 0, \\ k^n & \text{if } k > 0 \end{cases}$$
(3.17)

and, for $j \in \{1, 2, ..., n\}$,

$$b_{j}(n,k) = \begin{cases} S_{n}^{(j)} & \text{if } k = 0, \\ \sum_{m=0}^{n-1} {n \choose m} S_{n-m}^{(j)} k^{m} & \text{if } k > 0. \end{cases}$$
(3.18)

In particular, observe that when k > 0, each $b_j(n,k)$ is positive. In [4], Littlejohn and Wellman show that these numbers $\{b_j(n,k)\}_{j=0}^n$ appear in the symmetric form of the *n*th power of the classical Laguerre differential expression $\ell_{\alpha}[\cdot]$, defined by

$$\ell_{\alpha}[y](t) = -ty'' + (t - 1 - \alpha)y' + ky \quad (t \in (0, \infty))$$

= $t^{-\alpha} \exp(t)[-(t^{\alpha+1}\exp(-t)y'(t))' + kt^{\alpha}\exp(-t)y(t)]$

indeed, they prove that

$$t^{\alpha} e^{-t} \ell_{\alpha}^{n}[y](t) = \sum_{j=0}^{n} (-1)^{j} (b_{j}(n,k) t^{\alpha+j} e^{-t} y^{(j)}(t))^{(j)} \quad (n \in \mathbb{N}).$$

Moreover, in [4], the authors prove the following result concerning the numbers $\{b_j(n,k)\}_{j=0}^n$, which is important in our discussion on the Hermite differential expression.

Lemma 3.2. For each $n \in \mathbb{N}$, the numbers $b_j = b_j(n,k)$, defined in (3.14), are the unique solutions to the equations

$$(m+k)^n = \sum_{j=0}^n P(m,j)b_j \quad (m \in \mathbb{N}_0),$$
 (3.19)

where P(m, j) is defined in (3.4)

With \mathscr{P} denoting the space of all (possibly complex-valued) polynomials, we are now in position to prove the following theorem.

Theorem 3.3. Let $n \in \mathbb{N}$ and let $\ell_H[\cdot]$ denote the Hermite differential expression defined in (1.2). *Then*

$$\int_{-\infty}^{\infty} \ell_{H}^{n}[p](t)\bar{q}(t)\exp(-t^{2}) dt = \sum_{j=0}^{n} c_{j}(n,k) \int_{-\infty}^{\infty} p^{(j)}(t)\bar{q}^{(j)}(t)\exp(-t^{2}) dt \quad (p,q \in \mathscr{P}), \quad (3.20)$$

where

$$c_0(n,k) = \begin{cases} 0 & \text{if } k = 0, \\ k^n & \text{if } k > 0 \end{cases}$$
(3.21)

and, for
$$j = 1, 2, ..., n$$
,

$$c_j(n,k) = \begin{cases} 2^{n-j} S_n^{(j)} & \text{if } k = 0, \\ 2^{n-j} \sum_{m=0}^{n-1} {n \choose m} S_{n-m}^{(j)} \left(\frac{k}{2}\right)^m & \text{if } k > 0. \end{cases}$$
(3.22)

321

- (b) For k > 0, each $c_i(n,k)$ is positive (j = 0, 1, ..., n).
- (c) For each $n \in N$, the nth power $\ell_{H}^{n}[\cdot]$ of the Hermite expression $\ell_{H}[\cdot]$ is Lagrangian symmetrizable with symmetry factor $w(t) = \exp(-t^{2})$ and the Lagrangian symmetric form of $\exp(-t^{2})\ell_{H}^{n}[\cdot]$ is given by

$$\exp(-t^2)\ell_H^n[y](t) = \sum_{j=0}^n (-1)^j \left(c_j(n,k)\exp(-t^2)y^{(j)}(t)\right)^{(j)}.$$
(3.23)

Proof. The fact that the numbers $\{c_j(n,k)\}_{j=0}^n$ are positive follows from the positivity of the numbers $\{b_j(n,k)\}_{j=0}^n$ for k > 0. Since the Hermite polynomials $\{H_m(t)\}_{m=0}^\infty$ form a basis for \mathcal{P} , it suffices to show (3.20) is valid for $p = H_m(t)$ and $q = H_r(t)$, where $m, r \in \mathbb{N}_0$ are arbitrary. From the identity

$$\ell_H^n[H_m](t) = (2m+k)^n H_m(t) \quad (m \in \mathbb{N}_0),$$

it follows, with this particular choice of p and q, that the left-hand side of (3.20) reduces to $(2m+k)^n \delta_{m,r}$. On the other hand, from (3.3), the right-hand side of (3.20) yields

$$\sum_{j=0}^{n} c_{j}(n,k) \int_{-\infty}^{\infty} \frac{\mathrm{d}^{j}(H_{m}(t))}{\mathrm{d}t^{j}} \frac{\mathrm{d}^{j}(H_{r}(t))}{\mathrm{d}t^{j}} \exp(-t^{2}) \,\mathrm{d}t$$
$$= \sum_{j=0}^{n} 2^{j} P(m,j) c_{j}(n,k) \delta_{m,r} \quad \text{by (3.3)}.$$
(3.24)

Consequently, the identity in (3.20) holds if and only if

$$(2m+k)^n = \sum_{j=0}^n 2^j P(m,j) c_j(n,k),$$
(3.25)

or, after rearranging terms,

$$\left(m+\frac{k}{2}\right)^n = \sum_{j=0}^n b_j(n,k)P(m,j)$$

where

$$b_i(n,k) = 2^{j-n}c_i(n,k).$$

From Lemma 3.2, (3.17) and (3.18), it follows that the numbers $\{c_j(n,k)\}_{j=0}^n$ are given as in (3.21) and (3.22), establishing (3.20).

To prove (3.23), define the differential expression

$$m_H[y](t) := e^{t^2} \sum_{j=0}^n (-1)^j (c_j(n,k) \exp(-t^2) y^{(j)}(t))^{(j)}, \qquad (3.26)$$

where the numbers $c_j(n,k)$ (j = 0, 1, ..., n) are as above. For $p, q \in \mathscr{P}$ and $[a,b] \subset (-\infty, \infty)$, we apply integration by parts to obtain

$$\int_{a}^{b} m_{H}[p](t)\bar{q}(t)\exp(-t^{2}) dt$$

= $\sum_{j=0}^{n} (-1)^{j} c_{j}(n,k) \sum_{r=1}^{j} (-1)^{r+1} (p^{(j)}(t)\exp(-t^{2}))^{(j-r)} \bar{q}^{(r-1)}(t)|_{a}^{b}$

+
$$\sum_{j=0}^{n} c_j(n,k) \int_a^b p^{(j)}(t) \bar{q}^{(j)}(t) \exp(-t^2) dt$$

Now, for any $p \in \mathscr{P}$, $(p^{(j)}(t) \exp(-t^2))^{(j-r)} = p_{j,r}(t) \exp(-t^2)$ for some $p_{j,r} \in \mathscr{P}$; in particular, $\lim_{t \to +\infty} (p^{(j)}(t) \exp(-t^2))^{(j-r)} \bar{q}^{(r-1)}(t) = 0 \quad (p,q \in \mathscr{P}; r,j \in \mathbb{N}, r \leq j).$

Consequently, as $a \to -\infty$ and $b \to \infty$, we see that

$$\int_{-\infty}^{\infty} m_H[p](t)\bar{q}(t)\exp(-t^2)\,\mathrm{d}t = \sum_{j=0}^n c_j(n,k)\int_{-\infty}^{\infty} p^{(j)}(t)\bar{q}^{(j)}(t)\exp(-t^2)\,\mathrm{d}t \quad (p,q\in\mathscr{P}).$$
(3.27)

Consequently, from (3.27) and (3.20), we see that for all polynomials p and q, we have

$$(\ell_H^n[p] - m_H[p], q) = 0.$$

From the density of polynomials in $L^2((-\infty,\infty); \exp(-t^2))$, it follows that

$$\ell_{H}^{n}[p](t) = m_{H}[p](t) \quad (t \in (-\infty, \infty))$$
(3.28)

for all polynomials p. This latter identity implies that the expression $\ell_H^n[\cdot]$ has the form given in (3.23). \Box

For example, we see from this theorem that

$$\exp(-t^2)\ell_H^2[y](t) = (\exp(-t^2)y'')'' - ((2k+2)\exp(-t^2)y')' + k^2\exp(-t^2)y$$

and

$$\exp(-t^2)\ell_H^3[y](t) = -(\exp(-t^2)y''')''' + ((3k+6)\exp(-t^2)y'')'' - ((3k^2+6k+4)\exp(-t^2)y')' + k^3\exp(-t^2)y.$$

The following corollary lists some additional properties of $\ell_H^n[\cdot]$.

Corollary 3.4. Let $n \in \mathbb{N}$. Then

(a) the nth power of the classical Hermite differential expression

 $\mathscr{L}_{H}[y](t) := -y''(t) + 2ty'(t)$

is symmetrizable with symmetry factor $w(t) = \exp(-t^2)$ and has the Lagrangian symmetric form

$$\exp(-t^2)\mathscr{L}_{H}^{n}[y](t) := \sum_{j=1}^{n} (-1)^{j} \left(S_{n}^{(j)} 2^{n-j} \exp(-t^2) y^{(j)}(t) \right)^{(j)},$$

where $S_n^{(j)}$ is the Stirling number of the second kind defined in (3.16);

(b) the bilinear form $(\cdot, \cdot)_n$ defined on $\mathscr{P} \times \mathscr{P}$ by

$$(p,q)_n := \sum_{j=0}^n c_j(n,k) \int_{-\infty}^{\infty} p^{(j)}(t)\bar{q}^{(j)}(t) \exp(-t^2) dt \quad (p,q \in \mathscr{P})$$
(3.29)

is an inner product when k > 0 and satisfies

$$(\ell_H^n[p],q) = (p,q)_n \quad (p,q \in \mathscr{P}); \tag{3.30}$$

(c) the Hermite polynomials $\{H_m(t)\}_{m=0}^{\infty}$ are orthogonal with respect to the inner product $(\cdot, \cdot)_n$; in fact,

$$(H_m, H_r)_n = \sum_{j=0}^n c_j(n, k) \int_{-\infty}^{\infty} \frac{\mathrm{d}^j(H_m(t))}{\mathrm{d}t^j} \frac{\mathrm{d}^j(H_r(t))}{\mathrm{d}t^j} \exp(-t^2) \,\mathrm{d}t = (2m+k)^n \delta_{m,r}.$$
 (3.31)

Proof. The proof of (i) follows immediately from Theorem 3.1 and identities (3.21) and (3.22). The proof of (ii) is clear since all the numbers $\{c_j(n,k)\}_{j=0}^n$ are positive when k > 0. The identity in (3.30) follows from (3.27) and (3.28). Lastly, (3.31) is a restatement of (3.24), using (3.25).

4. The left-definite theory for the Hermite equation

For results that follow in this section, it is convenient to use the following notation. For $n \in \mathbb{N}$, let

$$AC_{\rm loc}^{(n-1)}(-\infty,\infty) := \{ f : (-\infty,\infty) \to \mathbb{C} \mid f, f', \dots, f^{(n-1)} \in AC_{\rm loc}(-\infty,\infty) \}.$$

For the rest of this section, we assume that k > 0.

Definition 4.1. For each $n \in \mathbb{N}$, define

$$V_{n} := \{ f: (-\infty, \infty) \to \mathbb{C} \mid f \in AC_{\text{loc}}^{(n-1)}(-\infty, \infty); \ f^{(j)} \in L^{2}((-\infty, \infty); \ \exp(-t^{2})) \\ (j = 0, 1, \dots, n) \}$$
(4.1)

and let $(\cdot, \cdot)_n$ and $|| \cdot ||_n$ denote, respectively, the inner product

$$(f,g)_n = \sum_{j=0}^n c_j(n,k) \int_{-\infty}^{\infty} f^{(j)}(t)\bar{g}^{(j)}(t) \exp(-t^2) dt \quad (f,g \in V_n)$$
(4.2)

(see (3.29) and (3.30)) and the norm $||f||_n = (f, f)_n^{1/2}$, where the numbers $c_j(n,k)$ are defined in (3.21) and (3.22).

The inner product $(\cdot, \cdot)_n$, defined in (4.2), is a Sobolev inner product and is more commonly called the *Dirichlet inner product* associated with the symmetric differential expression $\exp(-t^2)\ell_H^n[\cdot]$.

We remark that, for each r > 0, the *r*th left-definite inner product $(\cdot, \cdot)_r$ is abstractly given by

$$(f,g)_r = \int_{\mathbb{R}} \lambda^r \, \mathrm{d}E_{f,g} \quad (f,g \in V_r := \mathscr{D}(A^{r/2}))$$

where *E* is the spectral resolution of the identity for *A*; see [4]. However, we are able to determine this inner product in terms of the differential expression $\ell_H^r[\cdot]$ only when $r \in \mathbb{N}$.

We aim to show (see Theorem 4.4) that

$$W_n := (V_n, (\cdot, \cdot)_n)$$

is the *n*th left-definite space associated with the pair $(L^2((-\infty,\infty); \exp(-t^2)), A)$, where A is defined in (3.10) and (3.11). We begin by showing that W_n is a complete inner product space.

Theorem 4.2. For each $n \in \mathbb{N}$, W_n is a Hilbert space.

Proof. Let $n \in \mathbb{N}$. Suppose $\{f_m\}_{m=1}^{\infty}$ is Cauchy in W_n . Since each of the numbers $c_j(n,k)$ is positive, we see that $\{f_m^{(n)}\}_{m=1}^{\infty}$ is Cauchy in $L^2((-\infty,\infty); \exp(-t^2))$ and hence there exists $g_{n+1} \in L^2((-\infty,\infty); \exp(-t^2))$ such that

 $f_m^{(n)} \to g_{n+1}$ in $L^2((-\infty,\infty); \exp(-t^2))$.

Fix $t, t_0 \in \mathbb{R}$ (t_0 will be chosen shortly) and assume $t_0 \leq t$. From Hölder's inequality,

$$\int_{t_0}^t |f_m^{(n)}(t) - g_{n+1}(t)| \, \mathrm{d}t = \int_{t_0}^t |f_m^{(n)}(t) - g_{n+1}(t)| \exp(-t^2/2) \exp(t^2/2) \, \mathrm{d}t$$

$$\leq \left(\int_{t_0}^t |f_m^{(n)}(t) - g_{n+1}(t)|^2 \exp(-t^2) \, \mathrm{d}t \right)^{1/2} \left(\int_{t_0}^t \exp(t^2) \, \mathrm{d}t \right)^{1/2}$$

$$= M(t_0, t) \left(\int_{t_0}^t |f_m^{(n)}(t) - g_{n+1}(t)|^2 \exp(-t^2) \, \mathrm{d}t \right)^{1/2} \to 0 \quad \text{as } m \to \infty.$$

Moreover, since $f_m^{(n-1)} \in AC_{loc}(-\infty,\infty)$, we see that

$$f_m^{(n-1)}(t) - f_m^{(n-1)}(t_0) = \int_{t_0}^t f_m^{(n)}(t) \, \mathrm{d}t \to \int_{t_0}^t g_{n+1}(t) \, \mathrm{d}t$$
(4.3)

and, in particular, $g_{n+1} \in L^1_{loc}(-\infty,\infty)$. Furthermore, from the definition of $(\cdot,\cdot)_n$, we see that $\{f_m^{(n-1)}\}_{m=0}^{\infty}$ is Cauchy in $L^2((-\infty,\infty); \exp(-t^2))$; hence, there exists $g_n \in L^2((-\infty,\infty); \exp(-t^2))$ such that

$$f_m^{(n-1)} \to g_n$$
 in $L^2((-\infty,\infty); \exp(-t^2))$.

Repeating the above argument, we see that $g_n \in L^1_{loc}(-\infty,\infty)$ and, for any $t, t_1 \in \mathbb{R}$,

$$f_m^{(n-2)}(t) - f_m^{(n-2)}(t_1) = \int_{t_1}^t f_m^{(n-1)}(t) \, \mathrm{d}t \to \int_{t_1}^t g_n(t) \, \mathrm{d}t.$$
(4.4)

Moreover, from [8, Theorem 3.12], there exists a subsequence $\{f_{m_{k,n-1}}^{(n-1)}\}$ of $\{f_m^{(n-1)}\}_{m=1}^{\infty}$ such that

$$f_{m_{k,n-1}}^{(n-1)}(t) \rightarrow g_n(t)$$
 a.e. $t \in \mathbb{R}$.

Choose $t_0 \in \mathbb{R}$ in (4.3) such that $f_{m_{k,n-1}}^{(n-1)}(t_0) \to g_n(t_0)$ and then pass through this subsequence in (4.3) to obtain

$$g_n(t) - g_n(t_0) = \int_{t_0}^t g_{n+1}(t) \, \mathrm{d}t \quad (\text{a.e. } t \in \mathbb{R})$$

That is to say,

$$g_n \in AC_{\text{loc}}(-\infty,\infty)$$
 and $g'_n(t) = g_{n+1}(t)$ a.e. $t \in \mathbb{R}$. (4.5)

Again, from the definition of $(\cdot, \cdot)_n$, we see that $\{f_m^{(n-2)}\}_{m=1}^{\infty}$ is Cauchy in $L^2((-\infty, \infty); \exp(-t^2))$; consequently, there exists $g_{n-1} \in L^2((-\infty, \infty); \exp(-t^2))$ such that

$$f_m^{(n-2)} \to g_{n-1}$$
 in $L^2((-\infty,\infty); \exp(-t^2))$.
As above, we find that $g_{n-1} \in L^1_{loc}(-\infty,\infty)$; moreover, for any $t, t_2 \in \mathbb{R}$

$$f_m^{(n-3)}(t) - f_m^{(n-3)}(t_2) = \int_{t_2}^t f_m^{(n-2)}(t) \, \mathrm{d}t \to \int_{t_2}^t g_{n-1}(t) \, \mathrm{d}t$$

and there exists a subsequence $\{f_{m_{k,n-2}}^{(n-2)}\}$ of $\{f_m^{(n-2)}\}$ such that

$$f_{m_{k,n-2}}^{(n-2)}(t) \rightarrow g_{n-1}(t)$$
 a.e. $t \in \mathbb{R}$.

In (4.4), choose $t_1 \in \mathbb{R}$ such that $f_{m_{k,n-2}}^{(n-2)}(t_1) \to g_{n-1}(t_1)$ and pass through the subsequence $\{f_{m_{k,n-2}}^{(n-2)}\}$ in (4.4) to obtain

$$g_{n-1}(t) - g_{n-1}(t_1) = \int_{t_1}^t g_n(t) dt$$
 (a.e. $t \in \mathbb{R}$)

Consequently, $g_{n-1} \in AC_{loc}^{(1)}(-\infty,\infty)$ and $g_{n-1}''(t) = g_n'(t) = g_{n+1}(t)$ a.e. $t \in \mathbb{R}$. Continuing in this fashion, we obtain n+1 functions $g_{n-j+1} \in L^2((-\infty,\infty); \exp(-t^2)) \cap L_{loc}^1(-\infty,\infty)$ $(j=0,1,\ldots,n)$ such that

(i) $f_m^{(n-j)} \to g_{n-j+1}$ in $L^2((-\infty,\infty); \exp(-t^2))$ (j = 0, 1, ..., n), (ii) $g_1 \in AC_{\text{loc}}^{(n-1)}(-\infty,\infty); g_2 \in AC_{\text{loc}}^{(n-2)}(-\infty,\infty), ..., g_n \in AC_{\text{loc}}(-\infty,\infty)$, (iii) $g'_{n-j}(t) = g_{n-j+1}(t)$ a.e. $t \in \mathbb{R}$ (j = 0, 1, ..., n-1), (iv) $g_1^{(j)} = g_{j+1}$ (j = 0, 1, ..., n).

In particular, we see that $f_m^{(j)} \to g_1^{(j)}$ in $L^2((-\infty,\infty); \exp(-t^2))$ for j = 0, 1, ..., n and $g_1 \in V_n$. Hence, we see that

$$||f_m - g_1||_n^2 = \sum_{j=0}^n c_j(n,k) \int_{-\infty}^\infty |f_m^{(j)}(t) - g_1^{(j)}(t)|^2 \exp(-t^2) dt$$

\$\to 0\$ as \$m \to \infty\$.

Hence W_n is complete. \Box

We now show that \mathscr{P} is dense in W_n ; consequently, $\{H_m(t)\}_{m=0}^{\infty}$ is a complete orthogonal set in W_n .

Theorem 4.3. The Hermite polynomials $\{H_m(t)\}_{m=0}^{\infty}$ form a complete orthogonal set in the space \mathcal{P} of polynomials is dense in W_n .

Proof. Let $f \in W_n$; in particular, $f^{(n)} \in L^2((-\infty, \infty); \exp(-t^2))$. Consequently, from the completeness and orthonormality of $\{H_m(t)\}_{m=0}^{\infty}$ in $L^2((-\infty, \infty); \exp(-t^2))$, it follows that

$$\sum_{m=0}^{r} c_{m,n} H_m \to f^{(n)} \quad \text{as } r \to \infty \quad \text{in } L^2((-\infty,\infty); \exp(-t^2)),$$

where the numbers $\{c_{m,n}\}_{m=0}^{\infty} \subset \ell^2$ are the Fourier coefficients of $f^{(n)}$ defined by

$$c_{m,n} = \int_{-\infty}^{\infty} f^{(n)}(t) H_m(t) \exp(-t^2) dt \quad (m \in \mathbb{N}_0).$$

For $r \ge n$, define the polynomials

$$p_r(t) = \sum_{m=n}^{\prime} \frac{c_{m-n,n}}{2^{n/2} (P(m,n))^{1/2}} H_m(t)$$
(4.6)

(see (3.3)). Then, using the derivative formula (3.3) for the Hermite polynomials, we see that

$$p_r^{(j)}(t) = \sum_{m=n}^r \frac{c_{m-n,n} 2^{j/2} (P(m,j))^{1/2}}{2^{n/2} (P(m,n))^{1/2}} H_{m-j}(t) \quad (j = 1, 2, ...),$$
(4.7)

and, in particular, as $r \to \infty$,

$$p_r^{(n)} = \sum_{m=n}^r c_{m-n,n} H_{m-n} \to f^{(n)}$$
 in $L^2((-\infty,\infty); \exp(-t^2)).$

Furthermore, from [8, Theorem 3.12], there exists a subsequence $\{p_{r_i}^{(n)}\}$ of $\{p_r^{(n)}\}$ such that

$$p_{r_j}^{(n)}(t) \to f^{(n)}(t) \quad \text{a.e. } t \in \mathbb{R}.$$

$$(4.8)$$

Returning to (4.7), observe that since $2^{j/2}(P(m,j))^{1/2}/2^{n/2}(P(m,n))^{1/2} \rightarrow 0$ as $m \rightarrow \infty$ for j = 0, 1, ..., n-1, we see that

$$\left\{\frac{c_{m-n,n}2^{j/2}(P(m,j))^{1/2}}{2^{n/2}(P(m,n))^{1/2}}\right\}_{m=n}^{\infty}$$

is a square-summable sequence. Thus, from the completeness of the Hermite polynomials $\{H_m(t)\}_{m=0}^{\infty}$ in $L^2((-\infty,\infty); \exp(-t^2))$ and the Riesz–Fischer theorem (see [8, Chapter 4, Theorem 4.17]), there exists $g_i \in L^2((-\infty,\infty); \exp(-t^2))$ such that

$$p_r^{(j)} \to g_j \quad \text{in } L^2((-\infty,\infty); \exp(-t^2)) \quad \text{as } r \to \infty \quad (j=0,1,\ldots,n-1).$$
 (4.9)

Since, for a.e. $a, t \in (-\infty, \infty)$,

$$p_{r_j}^{(n-1)}(t) - p_{r_j}^{(n-1)}(a) = \int_a^t p_{r_j}^{(n)}(u) \, \mathrm{d}u \to \int_a^t f^{(n)}(u) \, \mathrm{d}u = f^{(n-1)}(t) - f^{(n-1)}(a) \quad (j \to \infty),$$

we see that, as $j \to \infty$,

$$p_{r_j}^{(n-1)}(t) \to f^{(n-1)}(t) + c_1 \quad (\text{a.e. } t \in (-\infty, \infty)),$$
(4.10)

where c_1 is some constant. From (4.9), with j = n - 1, we deduce that

$$g_{n-1}(t) = f^{(n-1)}(t) + c_1$$
 (a.e. $t \in (-\infty, \infty)$).

Next, from (4.10) and one integration, we obtain

$$p_{r_j}^{(n-2)}(t) \to f^{(n-2)}(t) + c_1 t + c_2 \quad (j \to \infty)$$

for some constant c_2 and hence, from (4.9),

$$g_{n-2}(t) = f^{(n-2)}(t) + c_1 t + c_2$$
 (a.e. $t \in (-\infty, \infty)$).

We continue this process to see that, for j = 0, 1, ..., n - 1,

$$g_j(t) = f^{(j)}(t) + q_{n-j-1}(t)$$
 (a.e. $t \in (-\infty, \infty)$),

where q_{n-j-1} is a polynomial of degree $\leq n-j-1$ satisfying

$$q'_{n-j-1}(t) = q_{n-j-2}(t).$$

Combined with (4.9), we see that, as $r \to \infty$,

$$p_r^{(j)} \to f^{(j)} + q_{n-j-1}$$
 in $L^2((-\infty,\infty); \exp(-t^2))$ $(j = 1, 2, ..., n).$

For each $r \ge n$, define the polynomials

$$\pi_r(t) := p_r(t) - q_{n-1}(t)$$

and observe that

$$\pi_r^{(j)} = p_r^{(j)} - q_{n-1}^{(j)}$$

= $p_r^{(j)} - q_{n-j-1}$
 $\rightarrow f^{(j)}$ in $L^2((-\infty,\infty); \exp(-t^2))$

Hence, as $r \to \infty$,

$$||f - \pi_r||_n^2 = \sum_{j=0}^n c_j(n,k) \int_{-\infty}^{\infty} |f^{(j)}(t) - \pi_r^{(j)}|^2 \exp(-t^2) dt \to 0.$$

The next result, which gives a simpler characterization of the function space V_n , follows from ideas in the above proof of Theorem 4.3. Due to the importance of this theorem (which can be seen in the statement of Corollary 4.5), we sketch the proof; specific details are given in Theorem 4.3.

Theorem 4.4. For each $n \in \mathbb{N}$,

$$V_n = \{ f : (-\infty, \infty) \to \mathbb{C} \mid f \in AC_{\text{loc}}^{(n-1)}(-\infty, \infty); f^{(n)} \in L^2((-\infty, \infty); \exp(-t^2)) \}.$$
(4.11)

Proof. Let $n \in \mathbb{N}$ and recall the definition of V_n in (4.1). Define

$$V'_n = \{ f : (-\infty, \infty) \to \mathbb{C} \mid f \in AC^{(n-1)}_{\text{loc}}(-\infty, \infty); f^{(n)} \in L^2((-\infty, \infty); \exp(-t^2)) \}$$

It is clear that $V_n \subset V'_n$. Conversely, suppose $f \in V'_n$ so $f^{(n)} \in L^2((-\infty,\infty); \exp(-t^2))$ and $f \in AC_{loc}^{(n-1)}(-\infty,\infty)$. As shown in Theorem 4.3, as $r \to \infty$,

$$\sum_{m=0}^{\prime} c_{m,n} H_m \to f^{(n)} \quad \text{in } L^2((-\infty,\infty); \exp(-t^2)),$$

where

$$c_{m,n} = \int_{-\infty}^{\infty} f^{(n)}(t) H_m(t) \exp(-t^2) dt \quad (m \in \mathbb{N}_0).$$

For $r \ge n$, let $p_r(t)$ be the polynomial that is defined in (4.6). Then, for any $j \in \mathbb{N}_0$, the *j*th derivative of p_r is given in (4.7) and, as in Theorem 4.3,

$$p_r^{(n)} \to f^{(n)}$$
 as $r \to \infty$ in $L^2((-\infty,\infty); \exp(-t^2))$

and, for $j=0,1,\ldots,n-1$, there exists polynomials q_{n-j-1} of degree $\leq n-j-1$ satisfying $q'_{n-j-1}(t) = q_{n-j-2}(t)$, such that

$$p_r^{(j)} \to f^{(j)} + q_{n-j-1}$$
 as $r \to \infty$ in $L^2((-\infty, \infty); \exp(-t^2))$,
= $f^{(j)} + q_{n-1}^{(j)}$.

Consequently, for each $j=0,1,\ldots,n-1$, $\{p_r^{(j)}-q_{n-1}^{(j)}\}_{r=n}^{\infty}$ converges in $L^2((-\infty,\infty); \exp(-t^2))$ to $f^{(j)}$. From the completeness of $L^2((-\infty,\infty); \exp(-t^2))$, we conclude that $f^{(j)} \in L^2((-\infty,\infty); \exp(-t^2))$ for $j=0,1,\ldots,n-1$. That is to say, $f \in V_n$. This completes the proof. \Box

We are now in position to prove the main result of this section.

Theorem 4.5. For k > 0, let $A : \mathcal{D}(A) \subset L^2((-\infty, \infty); \exp(-t^2)) \to L^2((-\infty, \infty); \exp(-t^2))$ denote the self-adjoint operator, defined in (3.6), (3.10), and (3.11), having the Hermite polynomials $\{H_m(t)\}_{m=0}^{\infty}$ as eigenfunctions. For each $n \in \mathbb{N}$, let V_n be given as in (4.1) or (4.11) and let $(\cdot, \cdot)_n$ denote the inner product defined in (3.29). Then $W_n = (V_n, (\cdot, \cdot)_n)$ is the nth left-definite space for the pair $(L^2((-\infty, \infty); \exp(-t^2)), A)$. Moreover, the Hermite polynomials $\{H_m(t)\}_{m=0}^{\infty}$ form a complete orthogonal set in W_n satisfying the orthogonality relation (3.31). Furthermore, define

$$A_n:\mathscr{D}(A_n)\subset W_n\to W_n$$

by

$$A_n f = \ell_H[f] \quad (f \in \mathscr{D}(A_n) := V_{n+2}),$$

where $\ell_H[\cdot]$ is the Hermite differential expression defined in (1.2). Then A_n is a self-adjoint differential operator in W_n ; more specifically, A_n is the nth left-definite operator associated with the pair $(L^2((-\infty,\infty); \exp(-t^2)), A)$. Furthermore, the Hermite polynomials $\{H_m(t)\}_{m=0}^{\infty}$ are eigenfunctions of A_n and the spectrum of A_n is given by

$$\sigma(A_n) = \{2m + k \mid m \in \mathbb{N}_0\}.$$

Proof. To show that W_n is the *n*th left-definite space for the pair $(L^2((-\infty,\infty); \exp(-t^2)), A)$, we must show that the five conditions in Definition 2.1 are satisfied.

- (i) W_n is complete: The proof of (i) is given in Theorems 4.2 and 4.4.
- (ii) $\mathscr{D}(A^n) \subset W_n \subset L^2((-\infty,\infty); \exp(-t^2))$: Let $f \in \mathscr{D}(A^n)$. Since the Hermite polynomials $\{H_m(t)\}_{m=0}^{\infty}$ form a complete orthonormal set in $L^2((-\infty,\infty); \exp(-t^2))$, we see that

$$p_j \to f \quad \text{in } L^2((-\infty,\infty); \exp(-t^2)) \quad (j \to \infty),$$

$$(4.12)$$

where

$$p_j(t) := \sum_{m=0}^j c_m H_m(t)$$

and $\{c_m\}_{m=0}^{\infty}$ are the Fourier coefficients of f in $L^2((-\infty,\infty); \exp(-t^2))$ defined by

$$c_m = (f, H_m) = \int_{-\infty}^{\infty} f(t) H_m(t) t^a \mathrm{e}^{-t} \, \mathrm{d}t \quad (m \in \mathbb{N}_0)$$

Since $A^n f \in L^2((-\infty,\infty); \exp(-t^2))$, we see that

$$\sum_{m=0}^{j} \alpha_m H_m \to A^n f \quad \text{in } L^2((-\infty,\infty); \exp(-t^2)) \quad (j \to \infty),$$

where

$$\alpha_m = (A^n f, H_m) = (f, A^n H_m) = (2m)^n (f, H_m) = (2m)^n c_m,$$

that is to say,

$$A^n p_j \to A^n f$$
 in $L^2((-\infty,\infty); \exp(-t^2))$ $(j \to \infty).$

Moreover, from (3.30), we see that

$$||p_j - p_r||_n^2 = (A^n[p_j - p_r], p_j - p_r)$$

$$\rightarrow 0 \quad \text{as } j, r \rightarrow \infty,$$

that is to say, $\{p_j\}_{j=0}^{\infty}$ is Cauchy in W_n . From Theorem 4.2, we see that there exists $g \in W_n \subset L^2((-\infty,\infty); \exp(-t^2))$ such that

 $p_j \to g$ in W_n $(j \to \infty)$.

Furthermore, by definition of $(\cdot, \cdot)_n$ and the fact that $c_0(n, k) = k^n$ for k > 0, we see that

$$(p_j-g, p_j-g)_n \geq k^n (p_j-g, p_j-g),$$

hence

$$p_j \to g \quad \text{in } L^2((-\infty,\infty); \exp(-t^2)).$$
 (4.13)

Comparing (4.12) and (4.13), we see that $f = g \in W_n$; this completes the proof of (ii).

(iii) $\mathscr{D}(A^n)$ is dense in W_n : Since polynomials are contained in $\mathscr{D}(A^n)$ and are dense in W_n (see Theorem 4.3), it is clear that (iii) is valid. Furthermore, from Theorem 4.3, we see that $\{H_m(t)\}_{m=0}^{\infty}$ forms a complete orthogonal set in W_n ; see also (3.31).

(iv) $(f, f)_n \ge k^n(f, f)$ for all $f \in V_n$: This is clear from the definition of $(\cdot, \cdot)_n$, the positivity of the coefficients $c_i(n,k)$, and the fact that $c_0(n,k) = k^n$.

(v) $(f,g)_n = (A^n f,g)$ for $f \in \mathcal{D}(A^n)$ and $g \in V_n$: Observe that this identity is true for any $f,g \in \mathscr{P}$; indeed, this is seen in (3.30). Let $f \in \mathscr{D}(A^n) \subset W_n$ and $g \in W_n$; since polynomials are dense in both W_n and $L^2((-\infty,\infty); \exp(-t^2))$ and convergence in W_n implies convergence in $L^2((-\infty,\infty); \exp(-t^2))$, there exists sequences of polynomials $\{p_j\}_{j=0}^{\infty}$ and $\{q_j\}_{j=0}^{\infty}$ such that, as $j \to \infty$,

$$p_j \to f$$
 in $W_n, A^n p_j \to A^n f$ in $L^2((-\infty, \infty); \exp(-t^2))$ (see the proof of part (ii))

and

$$q_j \to g$$
 in W_n and $L^2((-\infty,\infty); \exp(-t^2))$.

Hence, from (3.30),

$$(A^n[f],g) = \lim_{j \to \infty} (A^n[p_j],q_j) = \lim_{j \to \infty} (p_j,q_j)_n = (f,g)_n.$$

This proves (v). The rest of the proof follows immediately from Theorems 2.4 and 2.5. \Box

The following corollary follows immediately from Theorems 4.5 and 4.4, as well as (2.1). Remarkably, it characterizes the domain of each of the integral powers of A. In particular, the characterization given below of the domain $\mathcal{D}(A)$ of the classical Hermite differential operator A having the Hermite polynomials as eigenfunctions seems to be new.

Corollary 4.6. For each $n \in \mathbb{N}$, the domain $\mathcal{D}(A^n)$ of the nth power A^n of the classical self-adjoint operator A, defined in (3.10), (3.6), and (3.11), is given by

$$\mathscr{D}(A^{n}) = V_{2n} = \{ f : (-\infty, \infty) \to \mathbb{C} \mid f \in AC_{loc}^{(2n-1)}(-\infty, \infty); f^{(2n)} \in L^{2}((-\infty, \infty); \exp(-t^{2})) \}.$$

In particular,

$$\mathscr{D}(A) = V_2 = \{ f : (-\infty, \infty) \to \mathbb{C} \mid f \in AC^{(1)}_{\text{loc}}(-\infty, \infty); f'' \in L^2((-\infty, \infty); \exp(-t^2)) \}.$$

References

- [1] M. Abramowitz, I. Stegun (Eds.), Handbook of Mathematical Functions, Dover Publications, New York, NY, 1972.
- [2] N.I. Akhiezer, I.M. Glazman, Theory of Linear Operators in Hilbert space, Dover Publications, New York, NY, 1993.
- [3] E. Kreyszig, Introductory functional analysis with applications, Wiley Classics Library, Wiley, New York, 1989.
- [4] L.L. Littlejohn, R. Wellman, Left-definite spectral theory with applications to differential equations, submitted for publication.
- [5] M.A. Naimark, Linear Differential Operators II, Frederick Ungar, New York, 1968.
- [6] V. P. Onyango-Otieno, The application of ordinary differential operators to the study of classical orthogonal polynomials, Ph.D. Thesis, University of Dundee, Dundee, Scotland, 1980.
- [7] E.D. Rainville, Special Functions, Chelsea Publishing, New York, 1960.
- [8] W. Rudin, Real and Complex Analysis, 3rd Edition, McGraw-Hill Publishers, New York, 1987.
- [9] F.W. Schäfke, A. Schneider, SS-Hermitesche Randeigenwertprobleme I, Math. Ann. 162 (1965) 9-26.
- [10] G. Szegö, Orthogonal polynomials, American Mathematical Society Colloquium Publications, Vol. 23, Providence, RI, 1978.
- [11] E.C. Titchmarsh, Eigenfunction Expansions Associated With Second-Order Differential Equations, Clarendon Press, Oxford, England, 1946.
- [12] H. Weyl, Über gewöhnliche Differentialgleichungen mit Singularitäten und die zugehörigen Entwicklungen willkürlicher Funktionen, Mathematicsche Annalen, Vol. 68, 1910, pp. 220–269.



Journal of Computational and Applied Mathematics 121 (2000) 331-354

JOURNAL OF COMPUTATIONAL AND APPLIED MATHEMATICS

www.elsevier.nl/locate/cam

Minimal state-space realization in linear system theory: an overview

B. De Schutter

Control Laboratory, Faculty of Information Technology and Systems, Delft University of Technology, P.O. Box 5031, 2600 GA Delft, Netherlands

Received 2 July 1999; received in revised form 30 January 2000

Abstract

We give a survey of the results in connection with the minimal state-space realization problem for linear time-invariant systems. We start with a brief historical overview and a short introduction to linear system theory. Next we present some of the basic algorithms for the reduction of nonminimal state-space realizations and for the minimal state-space realization of infinite or finite sequences of Markov parameters of linear time-invariant systems. Finally, we discuss some extensions of this problem to other classes of systems and point out some related problems. © 2000 Elsevier Science B.V. All rights reserved.

Keywords: Minimal realization; Linear system theory; State-space models

1. Introduction

1.1. Overview

In this paper we give an overview of the results in connection with the minimal state-space realization problem for linear time-invariant (LTI) systems. The reason for focusing on LTI systems is that on the one hand, they form a very simple class of systems that can be analyzed rather easily and for which many analytic and numerical results are available, but that on the other, they have been used to solve many problems that appear in practice in a very satisfactory way. For sake of simplicity and conciseness, we will limit ourselves mainly to finite-dimensional discrete-time systems with real inputs and outputs in this paper. This choice is also motivated by the fact that most physical systems have real inputs and by the fact that some concepts (especially the Markov parameters) have

E-mail address: b.deschutter@its.tudelft.nl (B. De Schutter).

a more natural physical interpretation for discrete-time systems than for continuous-time systems. Furthermore, most of the techniques for discrete-time systems with real-valued inputs and outputs are also valid for systems with complex inputs and outputs and for continuous-time systems.

In general the minimal state-space realization problem for LTI systems can be formulated as follows: "Given some data about an LTI system, find a state-space description of minimal size that explains the given data". The data are typically the impulse response of the system, the step response, input-output measurements, frequency response data, or more general frequency measurements. The minimal state-space realization problem starting from impulse responses (or more general: sequences of Markov parameters) has been studied since the early 1960s and many algorithms have been developed to solve the problem. In this paper we will give an overview of some of these algorithms. At the end of the paper we will also briefly discuss the minimal state-space realization problem for some other classes of dynamical systems. Furthermore, we will also point out the relation between the minimal state-space realization problem and more involved problems such as model reduction and identification.

This paper is organized as follows. In Sections 1.2 and 1.3 we give a brief overview of the history of linear system theory and we discuss the main differences between the state-space representation and the transfer function representation of linear systems. In Section 2 we give a short and informal introduction to some of the basic concepts of linear system theory that are used in this paper. In Section 3 we treat various aspects of the minimal state-space realization problem for LTI systems and discuss some algorithms for solving this problem. Finally, we consider some related problems and extensions of the basic minimal state-space realization problem for LTI systems.

In order to limit the already large number of references in the bibliography of this paper we have selected a small subset of possible references, thereby aiming at historical papers, seminal papers, survey papers and reference works. Whenever we refer to a general book or paper, the reference is also intended to encompass the references included in that work.

1.2. Some historic notes on linear system theory and state-space models¹

Linear systems have already been studied for a long time and from many different points of view: in physics, mathematics, engineering, and so on. In an engineering context linear systems have been extensively studied since the 1930s. In those early days most researchers used frequency-domain techniques (i.e. input–output or transfer function descriptions). Moreover, most of this work was done for single-input–single-output (SISO) systems. At first sight the frequency-domain techniques did not seem to extend satisfactorily to the multi-input–multi-output (MIMO) systems that became increasingly important in aerospace, process control, and econometric applications in the late 1950s. This fact, and the importance of time-varying systems and time-domain characteristics in aerospace problems, led to a renewed interest in the state-space description of linear systems, triggered by the work of Bellman and Kalman. The papers [18,32] give a good idea of the situation around 1960. The state-space formulation led to many new ideas for systems design and feedback control. In the early 1970s Popov and Rosenbrock [43] have shown that many of the scalar transfer function concepts developed for SISO systems could also be extended to matrix transfer functions for MIMO systems. Now we could say that transfer functions descriptions (which are basically frequency domain

¹This section is based on [31].

methods) and state-space descriptions (which are more oriented towards the time domain) are only two extremes of a whole spectrum of possible descriptions of finite-dimensional LTI systems. We can work exclusively with one description or the other, but we can also easily translate results from one framework to another, and it really depends on the application we have in mind which method best suits our needs.

In this paper we will only consider state-space descriptions. The minimal realization problem for transfer functions is related to Padé approximation of rational functions, a topic that will be discussed in the contributions in this volume by Bultheel and De Moor, and Guillaume and Huard [9,23] (see also Section 4.1).

In the next section we will briefly discuss some differences between the state-space description and the transfer function description of a linear system.

1.3. State-space models versus transfer functions

The most important differences between the state-space representation and the transfer function representation of a given system are [12,48].

- The transfer function of an LTI system describes the relation between the input and the output of the system under the assumption that the system is initially relaxed (i.e., the initial state is zero). Hence, if this assumption does not hold, the description is not applicable.² In contrast to the state-space description, the transfer function representation does not reveal what will happen if the system is not initially relaxed (e.g., observable modes can be excited due to a nonzero initial state but may not appear in the transfer function due to pole-zero cancellation).
- The transfer function formulation does not reveal the behavior inside the system, such as unobservable unstable modes. Therefore, the transfer function matrix cannot always be used to study the stability properties of an LTI system. This problem of hidden pole-zero cancellation was not really understood prior to the work of Gilbert [18] and Kalman [32], who proved that the input–output description reveals only the controllable and observable part of a dynamical system.
- Although most results that are available for MIMO state-space descriptions can now also be obtained in the transfer function approach, the state-space formulation stays the most elegant way of dealing with generalizations like MIMO systems or nonlinear systems. Moreover, in practice the state-space formulation is very important for numerical computations and controller design.
- The state-space formulation can easily be extended to the time-varying case (see also Section 4.7). The extension of the transfer function to the time-varying case has not been very successful.

 $^{^{2}}$ Note that this assumption does hold for the minimal state-space realization problem based on the sequence of Markov parameters of an LTI system, which is the main topic of this paper.

2. Linear system theory

In this section we give an informal introduction to some concepts of linear system theory that will be used in the subsequent sections. The notation used in this section and the following sections is mainly based on [31]. Unless explicitly indicated otherwise, the proofs of the theorems and properties given below can be found in [31]. Other introductions to linear system theory can be found in [12,50].

2.1. Notation

The set of the real numbers is denoted by \mathbb{R} . All the vectors that appear in this paper are assumed to be column vectors, i.e., matrices with one column. If *a* is a vector then a_i represents the *i*th component of *a*. If *A* is a matrix then a_{ij} and $(A)_{ij}$ represent the entry on the *i*th row and the *j*th column of *A*. To select rows, columns and submatrices of a given matrix *A* we use the following Matlab-like notation. The *i*th row of *A* is denoted by A(i, :), and the *j*th column by A(:, j). Let *i*, *j* with i < j be two row indices of *A*, and let *k*, *l* with k < l be two column indices of *A*. The submatrix of *A* consisting of the entries on rows $i, i + 1, \ldots, j$ and columns $k, k + 1, \ldots, l$ is denoted by A(i : j, k : l). The submatrix of *A* consisting of columns $k, k + 1, \ldots, j$ is denoted by A(:, k : l). The $n \times n$ identity matrix is denoted by I_n and the $m \times n$ zero matrix by $0_{m,n}$. If the dimensions of the identity matrix or the zero matrix are not indicated, they should be clear from the context.

2.2. Linear time-invariant systems

A system or model is said to be time invariant if its response to any arbitrary input signal does not depend on absolute time. Consider a time-invariant system and let $\mathscr{S}(u)$ be the output of the system if the input signal u is applied to the system. Then we say that the system is linear if for every input signal u_1, u_2 and for every $c_1, c_2 \in \mathbb{R}$ we have $\mathscr{S}(c_1u_1 + c_2u_2) = c_1\mathscr{S}(u_1) + c_2\mathscr{S}(u_2)$. If we know and are interested in the inputs and outputs of the system at each time instant, then we will use a continuous-time model. On the other hand, in sampled or digital systems we often only know the signals of the system at certain discrete-time instants (e.g. at integer multiples of the sampling period). This leads to discrete-time models.

The behavior of a continuous-time LTI system with m inputs and l outputs can be described by a model of the form

$$\frac{\mathrm{d}x(t)}{\mathrm{d}t} = Ax(t) + Bu(t),\tag{1}$$

$$y(t) = Cx(t) + Du(t)$$
⁽²⁾

with $A \in \mathbb{R}^{n \times n}$, $B \in \mathbb{R}^{n \times m}$, $C \in \mathbb{R}^{l \times n}$ and $D \in \mathbb{R}^{l \times m}$, and where *u* is the input of the system, *y* the output and *x* the state. Similarly, the evolution of a discrete-time LTI system can be described by a model of the form

$$x(k+1) = Ax(k) + Bu(k), \tag{3}$$

$$y(k) = Cx(k) + Du(k).$$
(4)

Models (1)-(4) are called state-space models. The number of components of the state vector x is called the order of the model. A state-space model will be represented by the 4-tuple (A, B, C, D) of system matrices.

The Markov parameters G_k of an LTI system are defined by

$$G_0 = D$$
 and $G_k = CA^{k-1}B$ for $k = 1, 2, ...$ (5)

We say that (A, B, C, D) is a realization of the sequence $\{G_k\}_{k=0}^{\infty}$ if (5) holds. The realization is minimal if the model order is minimal. The model order of a minimal realization is called the minimal system order or sometimes also the McMillan degree of the system.

Consider a discrete-time LTI system and assume that x(0) = 0. If we apply a unit impulse $e(\cdot)$ defined by

$$e(k) = \begin{cases} 1 & \text{if } k = 0, \\ 0 & \text{otherwise,} \end{cases}$$

to the *i*th input of the system and if we apply a zero signal to the other inputs, then the output of the system is given by

$$y(0) = D(:,i)$$
 and $y(k) = CA^{k-1}B(:,i)$ for $k = 1, 2, ...$

This output is called the impulse³ response due to an impulse at the *i*th input. Note that y(k) corresponds to the *i*th column of the matrix $CA^{k-1}B$ for k = 1, 2, 3, Therefore, the sequence D, CB, CAB, CA^2B ,... is called the *impulse response* of the system. Note that the terms of this sequence corresponds to the Markov parameters of the system. So for a discrete-time LTI system the sequence $\{G_k\}_{k=0}^{\infty}$ of Markov parameters corresponds to the impulse response of the system.

Remark 2.1. For a continuous-time LTI system the situation is a little bit more complicated: let $y^i(t)$ be the output of a continuous-time LTI system with models (1) and (2) if we apply a Dirac impulse to the *i*th input and a zero signal to the other inputs. The matrix-valued function $Y(\cdot) = [y^1(\cdot) \ y^2(\cdot) \ \dots \ y^m(\cdot)]$ is called the impulse response of the system. It can be shown that the Taylor series expansion of $Y(\cdot)$ around the point t = 0 is given by

$$Y(t) = \sum_{k=0}^{\infty} G_k \frac{t^k}{k!}.$$

So for a continuous-time LTI system the relation between the Markov parameters and the impulse response is given by

$$G_k = \frac{d^{k-1}Y(t)}{dt^{k-1}}\bigg|_{t=0}.$$
 (6)

³Note that some authors prefer to use the term "pulse response" for the discrete-time case and reserve the term "impulse response" for the continuous-time case. However, in this paper we follow the terminology of [31] in which the term "impulse response" is used for both the discrete- and the continuous-time case.

2.3. Controllability and observability

Consider a 4-tuple (A, B, C, D) of system matrices of an LTI system and let N be a positive integer. We define

 $\mathcal{O}_{N}(C,A) = egin{bmatrix} C \ CA \ dots \ CA^{N-1} \end{bmatrix},$

$$\mathscr{C}_N(B,A) = [B \quad AB \quad \dots \quad A^{N-1}B].$$

If *n* is the model order of the realization (A, B, C, D) then $\mathcal{O}_n(C, A)$ is called the observability matrix of the realization and $\mathcal{C}_n(A, B)$ is called the controllability matrix.

A realization (A, B, C, D) is called observable if the observability matrix $\mathcal{O}_n(C, A)$ has full rank. If a realization is observable, then we can always reconstruct the initial state x(0) from observing the output evolution for $k \ge 0$ or $t \ge 0$ provided that we also know the input evolution for $k \ge 0$ or $t \ge 0$.

A realization is (A, B, C, D) is called controllable if the controllability matrix $\mathscr{C}_n(A, B)$ has full rank. If a realization is controllable, then for any initial state it is always possible to design an input sequence that steers the system to a desired final state.

The concepts observability and controllability are dual in the sense that a realization (A, B, C, D) is observable if and only if the dual realization (A^{T}, C^{T}, B^{T}, D) is controllable, and vice versa.

The following theorem which is due to Kalman gives a characterization of minimal state-space realizations.

Theorem 2.2. A realization (A, B, C, D) is minimal if and only if it is controllable and observable.

In general, a state-space realization of a given LTI system is not unique. Nevertheless, minimal state-space representations are unique up to a change of basis of the state space, or equivalently, any two minimal state-space realizations are connected by a unique similarity transformation [18,32].

Proposition 2.3. If (A, B, C, D) and $(\tilde{A}, \tilde{B}, \tilde{C}, \tilde{D})$ are two minimal state-space realizations of a given LTI system, there exists a unique invertible matrix T such that

$$\tilde{A} = T^{-1}AT, \quad \tilde{B} = T^{-1}B, \quad \tilde{C} = CT \quad and \quad \tilde{D} = D.$$
 (7)

Furthermore, the matrix T can be specified as $T = \mathscr{C}\widetilde{\mathscr{C}}^{\mathsf{T}}(\widetilde{\mathscr{C}}\widetilde{\mathscr{C}}^{\mathsf{T}})^{-1} = ((\widetilde{\mathscr{O}}^{\mathsf{T}}\widetilde{\mathscr{O}})^{-1}\widetilde{\mathscr{O}}^{\mathsf{T}}\mathscr{O})^{-1}$ with $\mathscr{C} = \mathscr{C}_{\rho}(A,B), \ \widetilde{\mathscr{C}} = \mathscr{C}_{\rho}(\widetilde{A},\widetilde{B}), \ \mathcal{O} = \mathscr{O}_{\rho}(C,A) \text{ and } \ \widetilde{\mathscr{O}} = \mathscr{O}_{\rho}(\widetilde{C},\widetilde{A}) \text{ where } \rho \text{ is the minimal system order.}$

The similarity transformation (7) corresponds to a transformation of the state $\tilde{x}(\cdot)=Tx(\cdot)$ where $x(\cdot)$ and $\tilde{x}(\cdot)$ are the state vectors of the realizations (A, B, C, D) and $(\tilde{A}, \tilde{B}, \tilde{C}, \tilde{D})$ respectively. Each choice of basis for the state-space will lead to another state-space representation (i.e., other system matrices). This results in several possible canonical forms such as the observer canonical form, the observability canonical form, the controller canonical form, etc. [31]. Different properties stand out more clearly in different realizations, and some canonical forms may have advantages in some applications. Note however that the input–output properties of the system such as the transfer function, the Markov parameters, the impulse response, and so on are not changed by similarity transformations.

In the next section we turn to the main topic of this paper: the minimal state-space realization problem for LTI systems.

3. The minimal state-space realization problem for LTI systems

3.1. Overview

The origins of the minimal state-space realization problem lie in the early 1960s. The minimal state-space realization problem for (continuous) LTI systems was first stated by Gilbert [18], who gave an algorithm for transforming a transfer function into a system of differential equations (i.e., a state-space description). A second algorithm for the problem was given around the same time by Kalman [32]. The approach of Gilbert was based on partial-fraction expansions and worked under the assumption that each entry of the transfer function matrix has distinct poles. Kalman's algorithm was based on the theory of controllability and observability and reduced a nonminimal state-space realization until it became minimal (cf. Theorem 2.2). Ho and Kalman [26,27] approached the minimal realization problem from an entirely new point of view: they solved the problem starting from the sequence of Markov parameters of the system. Their algorithm will be discussed extensively below. All these algorithms assume that the entire sequence of Markov parameters is available. However, many times only a limited number of Markov parameters is available. The corresponding minimal *partial* state-space realization problem for MIMO systems was first explored by Kalman [34] and Tether [54]. Later, Rissanen [42] gave a recursive solution of the SISO version of this problem (which he claims can easily be extended to the MIMO case).

Most of the early work on the minimal state-space realization problem dealt with the realization given the sequence of Markov parameters of the system. From a system-theoretical point of view this problem is often regarded as being somewhat academic. Nevertheless, there are several reasons why the minimal state-space realization problem for LTI systems deserves to be studied:

- This problem is one of the most fundamental problems in system theory and can be considered as a simplified version of problems with noisy data, nonlinear models, etc. that occur frequently in practice. Before we deal with these more complex problems, it is useful to study the simplified version, which might lead to additional insight in the original problems. As such the solution of the minimal state-space realization problem can also be seen as the first step towards problems such as model reduction and identification, which are of important practical interest.
- In order to analyze systems it is advantageous to have a compact description of the system. The aim of the minimal state-space realization problem is to find a state-space model of minimal size of the given system. Moreover, minimal realization techniques can also be used to reduce the order of existing state-space models.
- Since the minimal realization is both controllable and observable, it is a good basis for designing an observer to estimate the states of the system from measurements of the outputs, and also for subsequently designing a state feedback controller (using e.g. pole placement).

• Furthermore, the minimal state-space realization problem can be solved very elegantly using linear matrix algebra methods, that can be implemented in a numerically stable way.

The minimal state-space realization problem has attracted much attention since the early 1960s, which has resulted in a wide variety of algorithms to solve the problem. In the next sections we will discuss some of these minimal state-space realization algorithms.

In the remainder of the paper we will only consider discrete-time systems since for these systems the Markov parameters coincide with the terms of the impulse response, whereas for continuous-time systems the relation between the Markov parameters and the impulse response is more complicated (see Remark 2.1). Nevertheless, if we have in some way obtained the Markov parameters of a continuous-time LTI system then the techniques discussed below can also be used to obtain a minimal state-space realization of that system. Note however that (6) implies that matching an increasing number of Markov parameters of a continuous-time system means placing increasing emphasis on the high-frequency behavior of the system, which is more susceptible to noise.

In general, the basic minimal state-space realization methods can be classified into two main groups:

- The first group consists of methods that start with a nonminimal realization which could be obtained fairly easily and then reduce it to get a realization that is both controllable and observable and therefore also minimal. These methods will be discussed in Section 3.2.
- The second group consists of those methods that start with the impulse response (or Markov parameters) of the system and obtain the minimal realization directly by suitable transformations of the resulting Hankel matrix. These methods are treated in Section 3.3.

Afterwards we will also consider the minimal partial realization problem in Section 3.4, and the realization or approximation of noisy measurements of the impulse response (in Section 3.5) and the step response (in Section 3.6).

3.2. Minimal realization based on reduction of nonminimal realizations

Suppose that we have a (not necessarily minimal) *n*th-order state-space realization (A, B, C, D) of a given LTI system. Rosenbrock [43] has developed a procedure to transform this realization into a minimal realization in two steps. In fact, this algorithm is merely a small modification to the standard algorithm for reducing matrices to echelon form [37]. Rosenbrock's method works as follows. The matrices A, B and C are put in a matrix

$$P = \left[\frac{A \mid B}{C \mid 0}\right].$$

By applying a similarity transformation on P that consists of a sequence of elementary row operations (such as interchanging two rows or adding the multiple of a row to another row) on the first n rows of P and the corresponding column operations on the first n columns of P, the matrix P can be

transformed into a matrix of the form

$$\tilde{P} = \begin{bmatrix} A_{11} & 0 & 0 \\ A_{21} & A_{22} & B_2 \\ \hline C_1 & C_2 & 0 \end{bmatrix} \stackrel{\text{def}}{=} \begin{bmatrix} \tilde{A} & \tilde{B} \\ \tilde{C} & 0 \end{bmatrix},$$

where (A_{22}, B_2, C_2, D) is controllable. Since $(\tilde{A}, \tilde{B}, \tilde{C}, D)$ is connected to (A, B, C, D) by a similarity transformation, it is also a realization of the given system. Furthermore, since $\tilde{C} \tilde{A}^k \tilde{B} = C_2 A_{22}^k B_2$ for k = 0, 1, 2, ..., the 4-tuple (C_2, A_{22}, B_2, D) is a (controllable) state-space realization of the given system. By an analogous procedure on the matrix

$$Q = \begin{bmatrix} A_{22}^{\mathrm{T}} | C_2^{\mathrm{T}} \\ B_2^{\mathrm{T}} | 0 \end{bmatrix},$$

we obtain an observable realization. The resulting realization is then both controllable and observable and therefore also minimal (cf. Theorem 2.2).

A variant of Rosenbrock's method is implemented in the minreal command of Matlab. A stabilized version of Rosenbrock's algorithm is given in [56]. This algorithm is implemented in the freeware subroutine library SLICOT [7], which provides Fortran implementations of numerical algorithms for computations in systems and control theory.

3.3. Minimal realization of impulse responses

In this section we consider the problem of constructing a minimal realization starting from the impulse response $\{G_k\}_{k=0}^{\infty}$ of the system. Note that we always have $D = G_0$. Therefore, the problem of reconstructing D can be separated from the construction of A, B and C.

Many algorithms for minimal state-space realization of impulse responses use the following block Hankel matrix:

$$H_{r,r'}(\mathscr{G}) = \begin{bmatrix} G_1 & G_2 & G_3 & \dots & G_{r'} \\ G_2 & G_3 & G_4 & \dots & G_{r'+1} \\ G_3 & G_4 & G_5 & \dots & G_{r'+2} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ G_r & G_{r+1} & G_{r+2} & \dots & G_{r+r'-1} \end{bmatrix}.$$

Note that if (A, B, C, D) is a realization of the impulse response \mathcal{G} then we have

$$H_{r,r'}(\mathscr{G}) = \mathcal{O}_r(C,A)\mathscr{C}_{r'}(A,B).$$

We also define the shifted block Hankel matrix $\bar{H}_N(\mathscr{G})$ as

$$\bar{H}_{r,r'}(\mathscr{G}) = \begin{bmatrix} G_2 & G_3 & G_4 & \dots & G_{r'+1} \\ G_3 & G_4 & G_5 & \dots & G_{r'+2} \\ G_4 & G_5 & G_6 & \dots & G_{r'+3} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ G_{r+1} & G_{r+2} & G_{r+3} & \dots & G_{r+r'} \end{bmatrix}$$

The order of any minimal state-space realization of the sequence $\mathscr{G} = \{G_k\}_{k=0}^{\infty}$ is given by

$$\rho = \operatorname{rank} H_{\infty,\infty}(\mathscr{G}).$$

This result was discovered independently by Ho [25-27], Silverman [47], and Youla and Tissi [60].

Note that it is not always necessary to build the semi-infinite Hankel matrix $H_{\infty,\infty}(\mathscr{G})$ to determine the minimal system order. Indeed, if there is a linear relation between the Markov parameters of the form

$$G_{r+j} = \sum_{k=0}^{r-1} \alpha_k G_{k+j} \quad \text{for } j = 0, 1, 2, \dots$$
(8)

with $\alpha_0, \alpha_1, \ldots, \alpha_{r-1} \in \mathbb{R}$, then we have $\rho = \operatorname{rank} H_{r,r}(\mathscr{G})$ [26,27]. If the system matrix A of a (possible nonminimal) realization of the system is available, then a linear relation of the form (8) can easily be derived from the characteristic equation of the matrix A in combination with the Cayley–Hamilton theorem.

The use of Hankel matrices in realization theory was developed independently in the work of Ho and Kalman [25–27], Silverman [47], and Youla and Tissi [60]. These minimal realization algorithms can be divided into two groups:

- Some algorithms first determine the observable part of a system, and then the controllable part of the resulting system (or vice versa). Since the observability and controllability are dual concepts (see Section 2.3), the basic requirement is an algorithm for determining the controllable part. Most algorithms achieve this by selecting a largest set of linearly independent columns from the controllability matrix and use this set to construct a suitable transformation matrix (which removes the uncontrollable part). The resulting algorithms are quite complex. The algorithm of Silverman, which will be discussed more extensively below, belongs to this group.
- Another group of algorithms is based on a decomposition of the Hankel matrix. Both the algorithm of Ho and the algorithm of Youla and Tissi belong to this group.

3.3.1. Silverman's algorithm

The following theorem characterizes the sequences of Markov parameters that can be realized by an LTI system [48]:

Theorem 3.1. An infinite sequence of Markov parameters $\mathscr{G} = \{G_k\}_{k=0}^{\infty}$ is realizable by an LTI state-space model if and only if there exist positive integers r, r' and ρ such that

$$\operatorname{rank} H_{r,r'}(\mathscr{G}) = \operatorname{rank} H_{r+1,r'+j}(\mathscr{G}) = \rho \tag{9}$$

for j = 1, 2, ... The integer ρ then is the minimal system order.

In theory, the entire infinite sequence \mathscr{G} is needed to determine realizability since in general it is not true that rank $H_{r,r'+1} = \operatorname{rank} H_{r,r'}$ implies that (9) holds for all positive integers j [48]. However, for r large enough the rank of the Hankel matrix satisfies rank $H_{r,r}(\mathscr{G}) = \rho$ where ρ is the minimal system order.

Let r, r' and ρ be determined as in Theorem 3.1. The method of Silverman [48,49] is based on finding linearly independent rows in $H_{r,r'}(\mathscr{G})$. Let G be the submatrix of $H_{r,r'}(\mathscr{G})$ formed by the first ρ linearly independent rows of $H_{r,r'}(\mathscr{G})$, and let \tilde{G} be the submatrix of $H_{r+1,r'}$ positioned l rows below G. Let F be the nonsingular $\rho \times \rho$ matrix formed by the first ρ linearly independent columns of G, and let \tilde{F} be the $\rho \times \rho$ matrix occupying the same column positions in \tilde{G} as does F in G. Let F_1 be the $l \times \rho$ matrix occupying the same column positions in $H_{1,r'}(\mathscr{G})$ as does F in G. If we define $A = \tilde{F}F^{-1}$, B = G(:, 1:m), $C = F_1F^{-1}$, and $D = G_0$ then (A, B, C, D) is a minimal state-space realization of \mathscr{G} .

3.3.2. Ho's algorithm

The celebrated algorithm of Ho [26,27] can be stated as follows:

- 1. Determine a linear relation of the form (8) or select r large enough (e.g., larger than or equal to the order of another possibly nonminimal realization if that is available) and define $\rho = \operatorname{rank} H_{r,r}(\mathcal{G})$.
- 2. Find nonsingular matrices P and Q such that⁴

$$PH_{r,r}(\mathscr{G})Q = \begin{bmatrix} I_{\rho} & 0\\ 0 & 0 \end{bmatrix}.$$
 (10)

3. Now define

$$A = E_{\rho,rl} P \bar{H}_{r,r}(\mathscr{G}) Q E_{\rho,rm}^{\mathrm{T}}$$
$$B = E_{\rho,rl} P H_{r,r}(\mathscr{G}) E_{m,rm}^{\mathrm{T}},$$
$$C = E_{l,rl} H_{r,r}(\mathscr{G}) Q E_{\rho,rm}^{\mathrm{T}},$$
$$D = G_{0},$$

where $E_{p,q}$ is the $p \times q$ block matrix $[I_p \quad 0_{p,q-p}]$.

This yields a minimal state-space realization (A, B, C, D) of the sequence \mathscr{G} . Related algorithms using a reduced (i.e. smaller) Hankel matrix are described in [13,44].

Note that (10) corresponds to a decomposition of the matrix $H_{r,r}(\mathscr{G})$ as $H_{r,r}(\mathscr{G}) = H_0H_c$ with $H_0 \in \mathbb{R}^{rl \times \rho}$ and $H_c \in \mathbb{R}^{\rho \times rm}$ full rank matrices (with rank ρ). The algorithm of Youla and Tissi [60]

⁴ This is a standard problem in linear algebra. Apart from noting that P and Q may be taken to be lower and upper triangular, Ho and Kalman did not specify a particular matrix decomposition to be used in [26,27].

is also based on such a decomposition of $H_{r,r}(\mathscr{G})$. It can be shown that for any full rank matrix decomposition $H_{r,r}(\mathscr{G}) = H_0 H_c$ with $H_0 \in \mathbb{R}^{r l \times \rho}$ and $H_c^{\rho \times rm}$ satisfying

rank $H_{\rm o}$ = rank $H_{\rm c}$ = rank $H_{r,r}(\mathscr{G}) = \rho$,

there exist matrices A, B, C from a ρ -dimensional state-space model such that

$$H_{o} = \mathcal{O}_{r}(C, A)$$
 and $H_{c} = \mathscr{C}_{r}(A, B)$.

Furthermore, $\bar{H}_{r,r}(\mathcal{G}) = H_0 A H_c$. The matrices A, B and C can then be constructed as follows: $A = H_0^+ \bar{H}_N(\mathcal{G}) H_c^+$ where M^+ is the pseudo-inverse of the matrix M, $B = H_c(:, 1:m)$, and $C = H_0(1:l,:)$.

A numerically very reliable procedure for both the full rank decomposition of $H_{r,r}(\mathscr{G})$ and for the construction of the pseudo-inverses H_0^+ and H_c^+ is the singular value decomposition (SVD) [22,28]. The SVD also yields the most reliable numerical calculation of the rank of a matrix. The SVD of a matrix $M \in \mathbb{R}^{m \times n}$ is a decomposition of the form $M = U\Sigma V^T$ with $U \in \mathbb{R}^{m \times m}$ and $V \in \mathbb{R}^{n \times n}$ orthogonal matrices and $\Sigma \in \mathbb{R}^{m \times n}$ a diagonal matrix with $(\Sigma)_{11} \ge (\Sigma)_{22} \ge \cdots \ge 0$. The number of nonzero diagonal entries is equal to the rank of M.

The SVD can be used for the decomposition of the Hankel matrix $H_{r,r}(\mathscr{G})$ in the second step of Ho's algorithm as follows. Compute the SVD of $H_{r,r}(\mathscr{G})$: $H_{r,r}(\mathscr{G}) = U\Sigma V$ and define $H_o = U\Sigma^{1/2}$ and $H_c = \Sigma^{1/2} V^T$. This yields a decomposition that is equivalent to (10). The use of the SVD for the decomposition of the Hankel matrix was introduced by Zeiger and McEwen in their paper [61] in which they considered the problem of determining approximate state-space realizations of noisy data (see also Section 3.5).

Remark 3.2. In general, the system matrices A, B, C and D that result from the minimal realization algorithms discussed above do not exhibit a specific structure, i.e., all the system matrices are filled with nonzero coefficients. This implies that in general all $\rho(\rho + l + m) + lm$ entries have to be computed where ρ is the minimal system order. This has motivated work on algorithms that provide state-space models with specific canonical structures such as, e.g., the method of Ackerman and Bucy [1]. This method also consists in determining a set of linearly independent rows in the matrix $H_{r,r}(\mathscr{G})$. The resulting realization is in the canonical form of Bucy and has at most $\rho(l+m)$ parameters (the other entries are fixed at either 0 or 1).

3.4. The minimal partial realization problem

Now we assume that only a finite number of Markov parameters is available. So given a finite sequence $\mathscr{G}_N = \{G_k\}_{k=0}^N$ we want to find a 4-tuple (A, B, C, D) such that $D = G_0$ and $CA^{k-1}B = G_k$ for k = 1, 2, ..., N. In that case we say that (A, B, C, D) is a *partial realization* of \mathscr{G}_N . Note that trivially we have $D = G_0$. The 4-tuple (A, B, C, D) is said to be a minimal partial realization of \mathscr{G}_N if and only if the size of A is minimal among all other partial realizations of \mathscr{G}_N .

Clearly, a minimal partial realization always exists. However, uniqueness (even up to a similarity transformation) is only guaranteed under certain conditions [54]:

Proposition 3.3. Given a finite sequence $\mathscr{G}_N = \{G_k\}_{k=0}^N$ such that

rank $H_{r,r'}(\mathscr{G}_N) = \operatorname{rank} H_{r+1,r'}(\mathscr{G}_N) = \operatorname{rank} H_{r,r'+1}(\mathscr{G}_N)$

for some positive integers r, r' with r+r'=N, then the extension of the sequence \mathscr{G}_N to the infinite sequence $\mathscr{G}_{\infty} = \{G_k\}_{k=N+1}^{\infty}$ for which

$$\operatorname{rank} H_{p',p}(\mathscr{G}_{\infty}) = \operatorname{rank} H_{r',r}(\mathscr{G}_{\infty}) = \operatorname{rank} H_{r',r}(\mathscr{G}_{N})$$

with p' + p = N + k for k = 1, 2, ..., is unique.

If the conditions of this proposition hold, we can still apply the algorithms that are developed for the full minimal realization problem [34,54].

Proposition 3.4. The minimal partial realization problem of the sequence $\mathscr{G}_N = \{G_k\}_{k=0}^N$ may be solved by Ho's algorithm if and only if there exist positive integers r and r' with r + r' = N such that

$$\operatorname{rank} H_{r',r}(\mathscr{G}_N) = \operatorname{rank} H_{r',r}(\mathscr{G}_N) = \operatorname{rank} H_{r'+1,r}(\mathscr{G}_N).$$
(11)

The dimension of the minimal partial realization is equal to rank $H_{r',r}$.

If the rank condition (11) is satisfied then any pair of two different minimal partial realizations of the sequence $\mathscr{G} = \{G_k\}_{k=0}^N$ are connected by a similarity transformation.

Note that if the rank condition (11) is satisfied and if we have a partial realization of $\{G_k\}_{k=0}^N$, then we cannot be sure that this minimal partial realization is also a realization of the entire sequence $\mathscr{G}_{\infty} = \{G_k\}_{k=0}^{\infty}$ since rank $H_{r',r}(\mathscr{G}_{\infty})$ may increase if we increase r or r'.

If we have a finite sequence $\mathscr{G}_N = \{G_k\}_{k=0}^N$ for which the rank condition (11) does not hold for any positive integer r, then the only possibility for utilizing Proposition 3.4 is to try to extend \mathscr{G}_N to a longer sequence until (11) is satisfied. There could exist many extensions that satisfy the rank condition and each extension might yield a different minimal system order. Therefore, we now look for the extension that yields that smallest minimal system order among all possible extensions of \mathscr{G}_N that satisfy the rank condition (11). A characterization of the resulting minimal system order is too complex to state here, but can be found in [34]. A similar result was discovered simultaneously and independently by Tether [54].

The procedure of reduction of a nonminimal state-space representation of a finite sequence \mathscr{G}_N of Markov parameters to a controllable and observable one does not necessarily lead to a minimal realization of \mathscr{G}_N . A compression algorithm to reduce an arbitrary finite realization of \mathscr{G}_N to a minimal realization is given in [21]. This paper also provides a criterion for the minimality of a partial realization and an expression for the minimal system order.

Rissanen [42] has developed a recursive algorithm for the minimal partial state-space realization problem. His algorithm is based on a decomposition of $p \times q$ submatrices $H_{p,q}$ of the Hankel matrix $H_{\infty,\infty}(\mathscr{G})$ as PQ with $P \in \mathbb{R}^{p \times p}$ a lower triangular matrix with 1s on the diagonal and with certain entries of the matrix $Q \in \mathbb{R}^{p \times q}$ set to 0 so that entries in the lower triangular part of P can be computed recursively one by one and such that the numbers already calculated do not change if extra rows or columns are added to $H_{p,q}$. This yields an efficient algorithm for subsequently computing minimal partial state-space realizations of the finite sequences $\{G_k\}_{k=0}^N$, $\{G_k\}_{k=0}^{N+1}, \ldots$ where more data are taken into account as new measurements become available. In contrast to the other minimal realization algorithms discussed above, which require a complete recalculation of all parameters each time a new measurement becomes available, this algorithm has the advantage that only a few new parameters need to be calculated to extend a partial realization.

3.5. Minimal realization of noisy measurements of the impulse response

In practice, we will never have the exact Markov parameters of an LTI system at our disposition, but we will have measured data which are disturbed by noise. Furthermore, in practice, we will also only have a finite number of terms. Now, we ask ourselves how we can extract the underlying LTI state-space model from these noisy measurements.

If the minimal system order of the underlying "real" LTI system is ρ , then the measured sequence $\mathscr{G}_N = \{G_k\}_{k=0}^N$ can in general not be generated exactly by a ρ th-order state-space model. Furthermore, the Hankel matrix $H_{r,r}(\mathscr{G})$ will generically be of full rank, which implies that it is not possible to construct a low-order state-space realization that exactly matches the given sequence \mathscr{G}_N . Therefore, it may be better to make a good low-order approximation of the sequence \mathscr{G}_N rather than to try to match it exactly. Here we already enter the field of identification which will be discussed more extensively in Section 4.3. However, since Ho's algorithm can easily be extended to the special case of this section, we already treat it here. The method presented here is due to Kung [35] and is based on the SVD.

- 1. Given the sequence $\mathscr{G}_N = \{G_k\}_{k=0}^N$, construct a Hankel matrix $H_{r,r'}(\mathscr{G})$ with r + r' = N.
- 2. Compute the SVD of $H_{r,r'}(\mathcal{G})$: $H_{r,r'}(\mathcal{G}) = U\Sigma V^{T}$. Look how the singular values $(\Sigma)_{ii}$ decrease as a function of the index *i*, and decide how many singular values are significant. The remaining singular values will be neglected. Let ρ be the number of singular values that are retained.
- 3. Construct $U_{\rho} = U(:, 1 : \rho), V_{\rho} = V(:, 1 : \rho)$ and $\Sigma_{\rho} = \Sigma(1 : \rho, 1 : \rho).$
- 4. Now apply Ho's algorithm to the matrix $H_{\text{red}}(\mathscr{G}) = U_{\rho} \Sigma_{\rho} V_{\rho}^{\text{T}}$. Since $H_{\text{red}}(\mathscr{G})$ has rank ρ , the order of the resulting minimal state-space realization will be equal to ρ .

A related algorithm is given in [61] in which the SVD was also used, but no method for determining the resulting system order was specified.

Since in general the matrix $H_{red}(\mathscr{G})$ will not have a block Hankel structure, the Markov parameters of the resulting realization (A, B, C, D) will not exactly match the blocks of $H_{red}(\mathscr{G})$.

3.6. Minimal realization based on step response data

In many industrial processes we have step response measurements available instead of impulse response data. A straightforward way to do the realization then is to construct impulse response data by differencing or differentiating the step response data. However, this operation is not attractive since it will introduce an amplification of high-frequency noise in the data. As an alternative approach for discrete-time LTI systems, it is possible to use the step response data directly in a realization method that is a modified version of the Kung method. This modification is due to

van Helmont et al. [57], and consists in applying similar operations as the Kung algorithm of Section 3.5 but this time on the matrix

$$T_{r,r'} = \begin{bmatrix} S_1 & S_2 & S_3 & \cdots & S_{r'} \\ S_2 & S_3 & S_4 & \cdots & S_{r'+1} \\ S_3 & S_4 & S_5 & \cdots & S_{r'+2} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ S_r & S_{r+1} & S_{r+2} & \cdots & S_{r+r'-1} \end{bmatrix} - \begin{bmatrix} S_0 & S_0 & S_0 & \cdots & S_0 \\ S_1 & S_1 & S_1 & \cdots & S_1 \\ S_2 & S_2 & S_2 & \cdots & S_2 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ S_{r-1} & S_{r-1} & S_{r-1} & \cdots & S_{r-1} \end{bmatrix}$$

with r + r' = N + 1 where $\{S_k\}_{k=0}^N$ is the measured step response.

In practice, the measurements that are available will not necessarily be impulse response or step response data, but general input-output data. Since these data will in general always contain noise, an exact realization of the data by an LTI model (of low order) will not be possible. This brings us to the topic of identification, which will be discussed in the next section together with other related problems and extensions of the minimal state-space realization problem for LTI systems.

4. Related problems and extensions

4.1. Rational approximation

If we apply the z-transform to the discrete-time LTI state-space model (3) and (4) and if we assume that the initial condition of the system is x(0) = 0, then we obtain the following relation between the input and the output of the system:

$$Y(z) = H(z)U(z)$$

with the transfer function $H(\cdot)$ of the system given by

$$H(z) = C(zI - A)^{-1}B + D = \sum_{k=0}^{\infty} G_k z^{-k}.$$
(12)

Since

$$H(z) = \frac{1}{\det(zI - A)}C\operatorname{adj}(zI - A)B + D,$$

where adj(M) represents the adjoint matrix of M, the transfer function will always be a rational (matrix) function.

If we have a state-space representation of a system, then the transfer function can be computed using (12). On the other hand, if we have a SISO transfer function

$$H(z) = \frac{\sum_{i=0}^{n} a_{n-i} z^{i}}{\sum_{i=0}^{n} b_{n-i} z^{i}}$$

of a discrete-time LTI system with b_0 normalized to 1, then a possible state-space representation is given by the 4-tuple (A, B, C, D) with

$$A = \begin{bmatrix} -b_1 & -b_2 & \cdots & -b_{n-1} & -b_n \\ 1 & 0 & \cdots & 0 & \\ 0 & 1 & \cdots & 0 & \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & 0 & \cdots & 1 & 0 \end{bmatrix}, \quad B = \begin{bmatrix} 1 \\ 0 \\ \vdots \\ 0 \end{bmatrix},$$
$$C = \begin{bmatrix} a_1 - b_1 a_0 & a_2 - b_2 a_0 & \cdots & a_n - b_n a_0 \end{bmatrix} \text{ and } D = a_0.$$

A similar result holds for SISO continuous-time LTI models. For the MIMO case, the SISO state-space models that correspond to the individual transfer functions from each input to each output, could be stacked into one large MIMO state-space model. However, the resulting state-space models will in general certainly not be minimal. Therefore, several authors have developed methods to transform transfer function matrices into a minimal state-space realization (see, e.g., [33,41]).

Since the state-space representation can be converted into a transfer function and vice versa, we can also rephrase the minimal realization problem of Section 3.3 as follows: "Given the sequence of Markov parameters of an LTI system, determine the transfer function of the system with minimal McMillan degree". Since this transfer function is a rational function, this leads to the problem of approximation a given series by a rational function. This problem is related to the Padé approximation problem. For more information on this topic the reader is referred to [10,11] and the contributions of Bultheel and De Moor, and Guillaume and Huard in this volume [9,23].

4.2. Model reduction

In many practical applications high-order LTI state-space models are obtained (e.g. by combining models of separate components to build the model of a large plant, as the result of a filter or controller design, and so on). It is often desirable to replace them by lower-order models without introducing too much errors. Consequently, a wide variety of model reduction methods have been proposed. We shall concentrate on one method since it is connected to Ho's algorithm. It can be shown that the state-space model obtained using Ho's algorithm with SVD will be "balanced". The idea of balanced realizations of systems has first been introduced to the control area by Moore [39] and uses similarity transformations to put the system in a form from which reduced models can be obtained. Loosely speaking, in a balanced realization every state is as controllable as it is observable. As a consequence, the states can be ordered in terms of their contribution to the input–output properties of the system. In order to model reduction the states with the least contribution can be removed.

More information on this topic can be found in [19].

4.3. Identification

In practice the input–output measurements of a system will be disturbed by sensor and process noise. Furthermore, there will be nonlinear effects, modeling errors and so on, which makes that the given data can almost never be explained by a linear model. This brings us to the topic of identification, where we want to determine a linear model that explains the given data as well as possible (and that has also good generalization properties).

There are several approaches to generate a linear model of a system. We could, e.g., start from first principles and write down the basic physical laws that govern the behavior of the system. If the resulting model is nonlinear, we could linearize it in the operating point of the system in order to obtain a linear model. This "white-box" approach works for simple examples, but its complexity increases rapidly for real-world systems. An alternative approach is system identification, which is also called the "black-box" approach.⁵ In system identification we first collect measurements of the input–output behavior of the system and afterwards we compute a model that explains the measured data. The field of identification has developed rapidly during the past decades. We can now distinguish two main groups of algorithms to identify linear LTI models on the basis of measured data: prediction error methods and subspace methods. Let us now briefly discuss these two main groups of techniques.

The prediction error methods were developed by Ljung and his co-workers [36]. In prediction error methods the model of the system is first parameterized in some canonical way, and then the model parameters are determined such that the measurements are explained as accurately as possible by the model. This is done by formulating a constrained optimization problem with the unknown parameters of the model as variables, with a measure of the deviation between the measured data and the predictions obtained from the model as the objective function, and the model equations as the constraints.

In the beginning of the 1990s a new type of linear system identification algorithms, called subspace methods, emerged. Subspace identification algorithms yield state-space models and consist of two steps [14]. Most subspace methods first estimate the states of the system explicitly or implicitly using a projection of certain subspaces generated from the data. Next, they determine the state-space model by a linear least squares method.

So in subspace methods the identification problem is reduced to a simple least squares problem, whereas in prediction error methods generally nonlinear optimization problems have to be solved. Since subspace identification methods do not involve nonlinear optimization techniques (which are in general iterative), they are faster than prediction error methods. Another advantage is that subspace methods – provided they are implemented correctly – have better numerical properties than prediction error methods. Furthermore, they do not suffer from problems with local minima. The price to be paid is that subspace methods are suboptimal.

Since giving an overview of this domain is beyond a scope of this paper, we refer the interested reader to the following papers and books for more information on this topic. An excellent recent overview of subspace identification methods can be found in [14]. Prediction error methods are described in [36]. Some other key references for the field are [5,6,8,51].

⁵ Note that there also exists a "grey-box" approach that is used when the state-space equations of the system are known up to some unknown parameters, which are estimated using a parameter estimation method.

In the next sections we will discuss the minimal realization problem for state-space models that are not linear time-invariant. Although for most of these cases there exist theoretical characterizations of the minimal state-space realization, for almost all of the cases there are currently no efficient algorithms to compute minimal realizations (except for the linear time-varying case).

4.4. Positive linear systems

Positive linear systems are LTI systems for which the state and the output are always nonnegative for any nonnegative input signal. Positive linear models appear when we have a system in which the variables must take nonnegative value due to nature of the underlying physical system. Typical examples of positive linear systems are networks of reservoirs, industrial processes involving chemical reactors, heat exchangers and distillation columns, age-structure population models, compartmental systems (which are frequently used for modeling transport and accumulation phenomena of substances in human body), water and atmospheric pollution models, stochastic models with probabilities as state variables, and many other models commonly used in economy and sociology.

So a discrete-time positive LTI system (or positive linear system for short) is a system that can be described by a model of the form

$$x(k+1) = Ax(k) + Bu(k),$$
(13)

$$y(k) = Cx(k) + Du(k), \tag{14}$$

in which the components of the input, the state and the output are always nonnegative. This implies that the entries of the system matrices A, B, C and D are also nonnegative [55].

Now we consider the minimal state-space realization problem for positive linear systems: "Given the impulse response $\mathscr{G} = \{G_k\}_{k=0}^{\infty}$ of a positive linear system, determine a positive state-space realization (A, B, C, D) of \mathscr{G} with the dimension of A as small as possible". Although the problem of finding a finite-dimensional positive state-space realization for positive systems has been solved, the minimal positive state-space realization problem has not been solved completely yet [2]. If $\mathscr{G} = \{G_k\}_{k=0}^{\infty}$ is the impulse response of the system, then in contrast to general discrete-time LTI systems, the rank of the Hankel matrix $H_{\infty,\infty}(\mathscr{G})$ is only a lower bound for the minimal positive system order, and there are systems for which the actual minimal positive system order is larger than the rank of the Hankel matrix. In general the minimal positive system order can be characterized as follows [55].

Proposition 4.1. Given the impulse response $\mathscr{G} = \{G_k\}_{k=0}^{\infty}$ of a positive linear system with l inputs, the minimal positive system order is equal to the smallest integer ρ for which there exist matrices $H_o \in \mathbb{R}^{\rho \times \rho}$, $H_c \in \mathbb{R}^{\rho \times \rho}$ and $A \in \mathbb{R}^{\rho \times \rho}$ such that

$$H_{\infty,\infty}(\mathscr{G}) = H_0 H_c, \tag{15}$$

$$H_{\rm o}A = \hat{H}_{\rm o},\tag{16}$$

where \hat{H}_{o} is the matrix obtained by removing the first l rows of H_{o} .

However, there exist no efficient algorithms to compute a minimal decomposition of form (15) and (16). It is easy to verify that if we have a minimal decomposition of form (15) and (16) of

 $H_{\infty}(\mathscr{G})$ then the 4-tuple $(A, H_{c}(:, 1 : m), H_{0}(1 : l, :), G_{0})$ is a minimal state-space realization of the given impulse response. More information on this problem can be found in [17,55].

4.5. Max-plus-algebraic models

In this section we focus on state-space models for a class of discrete-event systems. Typical examples of discrete-event systems are manufacturing systems, telecommunication networks, railway traffic networks, and multi-processor computers. One of the characteristic features of discrete-event systems, as opposed to the continuous-variable systems⁶ considered above, is that their dynamics are *event-driven* as opposed to time-driven. An event corresponds to the start or the end of an activity. For a manufacturing system possible events are: the completion of a part on a machine, a machine breakdown, or a buffer becoming empty.

In general, models that describe the behavior of discrete-event systems are nonlinear, but there exists a class of discrete-event systems for which the model becomes "linear" when formulated in the max-plus algebra, which has maximization (represented by \oplus) and addition (represented as \otimes) as its basic operations. Loosely speaking, this class of discrete-event systems can be characterized as the class of deterministic time-invariant discrete-event systems in which only synchronization and no concurrency occurs. If we write down a model for the behavior of such a system, then the operations maximization and addition arise as follows. Synchronization corresponds to maximization (a new activity can only start when all the preceding activities have been finished, i.e., after the maximum of the finishing times of the preceding activities), whereas the duration of activities corresponds to addition (the finishing time of an activity is the starting time plus the duration of the activity). This leads to a model of the following form:⁷

$$x(k+1) = A \otimes x(k) \oplus B \otimes u(k), \tag{17}$$

$$y(k) = C \otimes x(k). \tag{18}$$

For a manufacturing system, u(k) would typically represent the time instants at which raw material is fed to the system for the (k + 1)th time, x(k) the time instants at which the machines start processing the *k*th batch of intermediate products, and y(k) the time instants at which the *k*th batch of finished products leaves the system.

Note that description (17) and (18) closely resembles the state-space description (3) and (4) for discrete-time LTI systems, but with + replaced by \oplus and \times by \otimes . Therefore, we say that (17) and (18) is a max-plus-linear model, i.e., a model that is linear in the max-plus algebra.

The reason for using the symbols \oplus and \otimes to denote maximization and addition is that there is a remarkable analogy between \oplus and addition, and between \otimes and multiplication: many concepts and properties from conventional linear algebra and linear system theory (such as the Cayley–Hamilton theorem, eigenvectors and eigenvalues, Cramer's rule, ...) can be translated to the max-plus algebra and max-plus-algebraic system theory by replacing + by \oplus and \times by \otimes . However, since there does not exist a max-plus-algebraic equivalent of the minus operator, we cannot straightforwardly transfer all the techniques from linear system theory to the max-plus-algebraic system theory.

⁶ That is, systems the behavior of which can be described by difference or differential equations.

⁷ The max-plus-algebraic matrix sum and product are defined in the same way as in linear algebra but with + replaced by \oplus and \times by \otimes . So $(A \oplus B)_{ij} = a_{ij} \oplus b_{ij} = \max(a_{ij}, b_{ij})$ and $(A \otimes B)_{ij} = \bigoplus_k a_{ik} \otimes b_{kj} = \max(a_{ik} + b_{kj})$.

We can also define the minimal state-space realization problem for max-plus-linear time-invariant discrete-event systems. This problem is strongly related to the minimal realization problem for positive linear systems that was considered in the previous section (e.g., with the proper change of notation, Proposition 4.1 also holds for max-plus-linear time-invariant systems). Just as for positive linear systems, there are currently no efficient, i.e., polynomial-time, algorithms to solve the general max-plus-algebraic minimal state-space realization problem, and there are strong indications that the problem is at least NP-hard. Nevertheless, there are also some special cases for which efficient algorithms exist. An recent overview of the current status of research and the open questions in connection with this problem is given in [15,40].

4.6. Multi-dimensional minimal state-space realization

In recent years there has been an increasing interest in the study of multi-dimensional systems, due to a wide range of applications in image processing, seismological data, geophysics, computer tomography, control of multi-pass processes, and so on. An *n*-dimensional state-space model has the following form:

$$\hat{x} = Ax + Bu(i_1, i_2, \dots, i_n),$$

 $y(i_1, i_2, \dots, i_n) = Cx + Du(i_1, i_2, \dots, i_n)$

with

$$\mathbf{\dot{x}} = \begin{bmatrix} x_{11}(i_1 + 1, i_2, \dots, i_n) \\ x_{12}(i_1, i_2 + 1, \dots, i_n) \\ \vdots \\ x_{1n}(i_1, i_2, \dots, i_n + 1) \\ \hline x_{21}(i_1 + 1, i_2, \dots, i_n) \\ x_{22}(i_1, i_2 + 1, \dots, i_n) \\ \vdots \\ x_{2n}(i_1, i_2, \dots, i_n + 1) \\ \hline \vdots \\ x_{nn}(i_1, i_2, \dots, i_n + 1) \end{bmatrix}$$
 and
$$\mathbf{x} = \begin{bmatrix} x_{11}(i_1, i_2, \dots, i_n) \\ x_{12}(i_1, i_2, \dots, i_n) \\ \vdots \\ x_{21}(i_1, i_2, \dots, i_n + 1) \\ \hline \vdots \\ x_{2n}(i_1, i_2, \dots, i_n + 1) \end{bmatrix}$$

The minimal state-space realization problem and the model reduction problem play an important role in the analysis and design of multi-dimensional systems because of the large amount of data involved in multi-dimensional signal processing. However, the general problem of minimal state-space realization of multidimensional systems has not been solved even for two-dimensional systems. Nevertheless, for some special cases minimal state-space realization methods have been derived. For more information the interested reader is referred to [3,38] and the references therein.

4.7. Linear time-varying models

The system matrices in the state-space models of the previous sections were constant over time. However, we can also consider time-varying linear systems in which the system matrices also depend on time

$$x_{k+1} = A_k x_k + B_k u_k$$

 $y_k = C_k x_k + D_k.$

Some authors even consider models in which the dimensions of the system matrices may change over time. Minimal state-space realizations for linear time-varying systems can also be characterized as being both controllable and observable [16]. Furthermore, the algorithm of Youla and Tissi can be extended to yield minimal state-space realizations for time-varying linear systems. We refer to [4,16,20,45] for more information on this topic.

4.8. Nonlinear models

When we use linear models to model physical systems, we are making some assumptions that correspond to an idealization of the real world, which is in fact nonlinear. Although LTI models turn out to be able to approximate many real-world systems and processes very well in practice, sometimes nonlinear models are required. In general, a discrete-time nonlinear time-invariant model has the following form:

$$x_{k+1} = f(x_k, u_k)$$

$$y_k = g(x_k, u_k).$$

We can also define a state-space realization and a minimal state-space realization for nonlinear systems. In analogy with linear systems, some authors define a minimal realization of a nonlinear system as a realization that is both controllable and observable [53]. However, where for a linear systems the dimension of the minimal realization can easily be determined from the impulse response or input–output data of the system, the situation is far more complicated for nonlinear systems. For more information in this context, the reader is referred to [24,29,30,46,53].

There are many other classes of linear and nonlinear time-invariant or time-varying systems (such as linear systems that operate on finite fields or integers (instead of real numbers), descriptor systems, periodic systems, ...) for which minimal state-space realization results exist, but it would be beyond the scope of this paper to discuss them all. More information on this topic can be found in [31,52,58,59] and the references therein.

5. Conclusion

In this paper we have given an overview of the minimal state-space realization problem for linear time-invariant systems and discussed some related problems and extensions. The basic problem has been solved satisfactorily since the mid-1960s and has led to a renewed research in various fields

such as model reduction, approximation and identification. Especially, for general nonlinear systems and special classes of nonlinear systems there still is much active research going on.

References

352

- J.E. Ackermann, R.S. Bucy, Canonical minimal realization of a matrix of impulse response sequences, Inform. and Control 19 (3) (1971) 224–231.
- [2] B.D.O. Anderson, Positive system realizations, in: V.D. Blondel, E.D. Sontag, M. Vidyasagar, J.C. Willems (Eds.), Open Problems in Mathematical Systems and Control Theory. Springer, London, 1999 (Chapter 2).
- [3] G.E. Antoniou, P.N. Paraskevopoulous, S.J. Varoufakis, Minimal state-space realization of factorable 2-D transfer functions, IEEE Trans. Circuits Systems 35 (8) (1988) 1055–1058.
- [4] D.Z. Arov, M.A. Kaashoek, D.R. Pik, Optimal time-variant systems and factorization of operators, I: Minimal and optimal systems, Integral Equations Oper. Theory 31 (4) (1998) 389–420.
- [5] K.J. Åström, P. Eykhoff, System identification a survey, Automatica 7 (2) (1971) 123–162.
- [6] K.J. Åström, T. Söderström (Eds.), Automatica (Special Issue on Trends in System Identification) 31 (12) (1995) 1691–1892.
- [7] P. Benner, V. Sima, V. Mehrmann, S. Van Huffel, A. Varga, SLICOT a subroutine library in systems and control theory, in: B. Datta (Ed.), Applied and Computational Control, Signals, and Circuits, Vol. 1, Birkhäuser, Boston, 1999, pp. 504–546 (Chapter 10), see also http://www.win.tue.nl/niconet/NIC2/slicot.html.
- [8] G.E.P. Box, G.M. Jenkins, G.C. Reinsel, Time Series Analysis: Forecasting and Control, 3rd Edition, Prentice-Hall, Englewood Cliffs, NJ, 1994.
- [9] A. Bultheel, B. De Moor, Rational approximation in linear systems and control, this volume (Numerical Analysis in the 20th Century Vol. I: Approximation Theory), J. Comput. Appl. Math. (2000) 355–378.
- [10] A. Bultheel, M. Van Barel, Padé techniques for model reduction in linear system theory: a survey, J. Comput. Appl. Math. 14 (3) (1986) 401–438.
- [11] A. Bultheel, M. Van Barel, Linear Algebra, Rational Approximation and Orthogonal Polynomials, Studies in Computational Mathematics, Vol. 6, Elsevier, Amsterdam, 1997.
- [12] C.T. Chen, Linear System Theory and Design, Holt, Rinehart and Winston, New York, 1984.
- [13] C.T. Chen, D.P. Mital, A simplified irreducible realization algorithm, IEEE Trans. Automat. Control 17 (1972) 535–537.
- [14] B. De Moor, P. Van Overschee, W. Favoreel, Algorithms for subspace state-space system identification: an overview, in: B. Datta (Ed.), Applied and Computational Control, Signals, and Circuits, Vol. 1, Birkhäuser, Boston, 1999, pp. 271–335 (Chapter 6).
- [15] B. De Schutter, G.J. Olsder, The minimal state-space realization problem in the max-plus algebra: an overview, Technical Report, Control Laboratory, Faculty of Information Technology and Systems, Delft University of Technology, Delft, The Netherlands, May 1999, submitted for publication.
- [16] P. Dewilde, A.J. van der Veen, Time-Varying Systems and Computations, Kluwer Academic Publishers, Boston, 1998.
- [17] L. Farina, L. Benvenuti, Positive realizations of linear systems, Systems Control Lett. 26 (1995) 1-9.
- [18] E.G. Gilbert, Controllability and observability in multi-variable control systems, SIAM J. Control 1 (2) (1963) 128–151.
- [19] K. Glover, All optimal Hankel-norm approximations of linear multivariable systems and their L^{∞} -error bounds, Internat. J. Control 39 (6) (1984) 1115–1193.
- [20] I. Gohberg, M.A. Kaashoek, On minimality and stable minimality of time-varying linear systems with well-posed boundary conditions, Internat. J. Control 43 (5) (1986) 1401–1411.
- [21] I. Gohberg, M.A. Kaashoek, L. Lerer, On minimality in the partial realization problem, Systems Control Lett. 9 (1987) 97-104.
- [22] G.H. Golub, C.F. Van Loan, Matrix Computations, The John Hopkins University Press, Baltimore, MD, 1989.
- [23] P. Guillaume, A. Huard, Multivariate Padé approximation, this volume (Numerical Analysis in the 20th Century Vol. I: Approximation Theory), J. Comput. Appl. Math. (2000) 197–219.

- [24] R. Hermann, A. Krener, Nonlinear controllability and observability, IEEE Trans. Automat. Control 22 (1977) 728– 740.
- [25] B.L. Ho, An effective construction of realizations from input/output descriptions, Ph.D. Thesis, Stanford University, Stanford, CA, 1966.
- [26] B.L. Ho, R.E. Kalman, Effective construction of linear state-variable models from input/output functions, in: M.E. Van Valkenburg (Ed.), Proceedings of the Third Annual Allerton Conference on Circuit and System Theory, Monticello, IL, October 1965, pp. 449–459.
- [27] B.L. Ho, R.E. Kalman, Effective construction of linear, state-variable models from input/output functions, Regelungstechnik 14 (12) (1966) 545–548.
- [28] R.A. Horn, C.R. Johnson, Matrix Analysis, Cambridge University Press, Cambridge, UK, 1985.
- [29] A. Isidori, Nonlinear Control Systems, 3rd Edition, Springer, Berlin, 1995.
- [30] B. Jakubczyk, Existence and uniqueness of realizations of nonlinear systems, SIAM J. Control Optim. 18 (4) (1980) 455–471.
- [31] T. Kailath, Linear Systems, Prentice-Hall, Englewood Cliffs, NJ, 1980.
- [32] R.E. Kalman, Mathematical description of linear dynamical systems, SIAM J. Control 1 (2) (1963) 152–192.
- [33] R.E. Kalman, Irreducible realizations and the degree of a rational matrix, J. Soc. Ind. Appl. Math. 13 (2) (1965) 520-544.
- [34] R.E. Kalman, On minimal partial realizations of a linear input/output map, in: R.E. Kalman, N. DeClaris (Eds.), Aspects of Network and System Theory, Holt, Rinehart and Winston, New York, 1971, pp. 385–407.
- [35] S.Y. Kung, A new identification and model reduction algorithm via singular value decomposition, in: Proceedings of the 12th Asilomar Conference on Circuits, Systems and Computers, Pacific Grove, CA, 1978, pp. 705–714.
- [36] L. Ljung, System Identification: Theory for the User, 2nd Edition, Prentice-Hall, Upper Saddle River, NJ, 1999.
- [37] D.Q. Mayne, An elementary derivation of Rosenbrock's minimal realization algorithm, IEEE Trans. Automat. Control 18 (3) (1973) 306–307.
- [38] S.H. Mentzelopoulou, N.J. Theodorou, n-dimensional minimal state-space realization, IEEE Trans. Circuits Systems 38 (3) (1991) 340–343.
- [39] B.C. Moore, Principal component analysis in linear systems: controllability, observability and model reduction, IEEE Trans. Automat. Control 26 (1) (1981) 17–31.
- [40] G.J. Olsder, B. De Schutter, The minimal realization problem in the max-plus algebra, in: V.D. Blondel, E.D. Sontag, M. Vidyasagar, J.C. Willems (Eds.), Open Problems in Mathematical Systems and Control Theory, Springer, London, 1999, pp. 157–162 (Chapter 32).
- [41] I.S. Pace, S. Barnett, Efficient algorithms for linear system calculations Part II Minimal realization, Internat. J. Systems Sci. 5 (5) (1974) 413–424.
- [42] J. Rissanen, Recursive identification of linear systems, SIAM J. Control Optim. 9 (3) (1971) 420-430.
- [43] H.H. Rosenbrock, State-Space and Multivariable Theory, Thomas Nelson and Sons, London, 1970.
- [44] P. Rózsa, N.K. Sinha, Efficient algorithm for irreducible realization of a rational matrix, Internat. J. Control 20 (5) (1974) 739–751.
- [45] W.J. Rugh, Linear System Theory, 2nd Edition, Prentice-Hall, Upper Saddle River, NJ, 1996.
- [46] J.M.A. Scherpen, W.S. Gray, Minimality and similarity invariants of a nonlinear state-space realization, IEEE Trans. Automat. Control (1999), in press.
- [47] L.M. Silverman, Structural properties of time-variable linear systems, Ph.D. Thesis, Department of Electrical Engineering, Columbia University, New York, 1966.
- [48] L.M. Silverman, Realization of linear dynamical systems, IEEE Trans. Automat. Control 16 (6) (1971).
- [49] L.M. Silverman, H.E. Meadows, Equivalence and synthesis of time-variable linear systems, in: W.R. Perkins, J.B. Cruz Jr. (Eds.), Proceedings of the Fourth Annual Allerton Conference on Circuit and System Theory, Monticello, IL, October 1966, 776–784.
- [50] R.E. Skelton, Dynamic Systems Control, Wiley, New York, 1988.
- [51] T. Söderström, P. Stoica, System Identification, Prentice-Hall, London, 1989.
- [52] J. Sreedhar, P. Van Dooren, P. Misra, Minimal order time invariant representation of periodic descriptor systems, in: Proceedings of the 1999 American Control Conference, San Diego, CA, June 1999, pp. 1309–1313.
- [53] H.J. Sussmann, Existence and uniqueness of minimal realizations of nonlinear systems, Math. Systems Theory 10 (3) (1977) 263–284.

354

- [54] A.J. Tether, Construction of minimal linear state-variable models from finite input-output data, IEEE Trans. Automat. Control 17 (4) (1970) 427–436.
- [55] J.M. van den Hof, System Theory and System Identification of Compartmental Systems, Ph.D. Thesis, Faculty of Mathematics and Natural Sciences, University of Groningen, Groningen, The Netherlands, November 1996.
- [56] P.M. Van Dooren, The generalized eigenstructure on linear system theory, IEEE Trans. Automat. Control 26 (1) (1981) 111–129.
- [57] J.B. van Helmont, A.J.J. van der Weiden, H. Anneveld, Design of optimal controllers for a coal fired Benson boiler based on a modified approximate realization algorithm, in: R. Whalley (Ed.), Application of Multivariable System Techniques, Elsevier, London, 1990, pp. 313–320.
- [58] A. Varga, Computation of minimal realizations of periodic systems, in: Proceedings of the 37th IEEE Conference on Decision and Control, Tampa, FL, December 1998, pp. 3825–3830.
- [59] Y. Wang, E.D. Sontag, Orders of input/output differential equations and state-space dimensions, SIAM J. Control Optim. 33 (4) (1995) 1102–1127.
- [60] D.C. Youla, P. Tissi, *n*-port synthesis via reactance extraction Part I, IEEE Internat. Convention Record, part 7 14 (1966) 183–205.
- [61] H.P. Zeiger, A.J. McEwen, Approximate linear realizations of given dimension via Ho's algorithm, IEEE Trans. Automat. Control 19 (2) (1974) 153.



Journal of Computational and Applied Mathematics 121 (2000) 355-378

JOURNAL OF COMPUTATIONAL AND APPLIED MATHEMATICS

www.elsevier.nl/locate/cam

Rational approximation in linear systems and control $\stackrel{\text{tr}}{\sim}$

A. Bultheel^{a, *,1}, B. De Moor^{b, 2}

^aDepartment of Computer Science (NALAG), K.U.Leuven, Celestijnenlaan 200 A, B-3001 Leuven, Belgium ^bDepartment of Electrical Engineering (ESAT-SISTA), K.U.Leuven, Kardinaal Mercierlaan 94, B-3001 Leuven, Belgium

Received 7 July 1999; received in revised form 27 January 2000

Abstract

In this paper we want to describe some examples of the active interaction that takes place at the border of rational approximation theory and linear system theory. These examples are mainly taken from the period 1950–1999 and are described only at a skindeep level in the simplest possible (scalar) case. We give comments on generalizations of these problems and how they opened up new ranges of research that after a while lived their own lives. We also describe some open problems and future work that will probably continue for some years after 2000. © 2000 Elsevier Science B.V. All rights reserved.

Keywords: Rational approximation; Linear system theory; Model reduction; Identification

1. Introduction

In linear systems, control, and signal processing, rational approximation has always been an important issue and it has given rise to specific problems and insights in approximation theory, it has

 $^{^{\}ddagger}$ This work is partially supported by 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 authors.

^{*} Corresponding author. Tel.: +32-16-327540; fax: +32-16-327996.

E-mail addresses: adhemar.bultheel@cs.kuleuven.ac.be (A. Bultheel), bart.demoor@esat.kuleuven.ac.be (B. De Moor)

¹Partially supported by the Fund for Scientific Research (FWO), project "Orthogonal systems and their applications", grant #G.0278.97.

² Research associate of the Fonds Wetenschappetyk Onderzoek Vlaanderen and Associate Professor at the K.U. Leuven. Also supported by grants from the Flemish Government (Research council KUL, GOA-MIPS, FWO-Vlaanderen (grants, research projects and communities), AWI (Bil. Int. Coll.), IWT (Eureka SINOPSYS, Eureka Impact, STWW)), the Belgian Federal Government (IUAP-IV/02 and IUAP-IV/24, Program Dur. Dev. (Mobility)), the European Commission (TMR Networks (Alapedes, Niconet), Science (ERNSI)) and several industrial grants (Electrabel, Laborelec, Verhaert).

revived forgotten methods and initiated new directions of research. It is the intention of this paper to illustrate some of these innovating ideas that were born from this interaction of system theory, linear algebra and approximation theory and formulate some open problems or aspects that are still under development.

First we recall some mathematical notation and concepts. Next, we shall narrow the focus of our system theoretic problems by introducing some concepts from system theory and selecting the subjects we shall discuss from those we shall not consider in this paper.

By \mathbb{Z} we denote the set of integers, by \mathbb{N} the positive integers, and $\ell_p = \ell_p(\mathbb{Z})$, $(1 \leq p < \infty)$ is the Banach space of complex sequences $x = (x_k)$ with $||x||_p := [\sum_{k \in \mathbb{Z}} |x_k|^p]^{1/p} < \infty$ while ℓ_{∞} is the space for which $||x||_{\infty} := \sup_{k \in \mathbb{Z}} |x_k| < \infty$. The set of real and complex numbers is \mathbb{R} and \mathbb{C} , respectively, and the notation \mathbb{T} , \mathbb{D} and \mathbb{E} are reserved for the unit circle, its interior and its exterior: $\mathbb{T} := \{z \in \mathbb{C} : |z| = 1\}, \ \mathbb{D} := \{z \in \mathbb{C} : |z| < 1\}, \ \mathbb{E} := \{z \in \mathbb{C} : |z| > 1\}.$ The spaces $L_p = L_p(\mathbb{T}), \ \mathbb{C} := \{z \in \mathbb{C} : |z| > 1\}$. $(1 \leq p < \infty)$ are defined as the function spaces for which $||F||_p := [(1/2\pi) \int_{-\pi}^{\pi} |F(e^{i\omega})|^p d\omega]^{1/p} < \infty$, and $||F||_{\infty} := \sup_{t \in \mathbb{T}} |F(t)|$. L_2 is a Hilbert space with inner product $\langle F, G \rangle = (1/2\pi) \int_{-\pi}^{\pi} F(e^{i\omega}) \overline{G(e^{i\omega})} d\omega$. The \mathscr{Z} -transform of a sequence $a = (a_k)_{k \in \mathbb{Z}}$ is $A(z) = \mathscr{Z}(a) = \sum_{k \in \mathbb{Z}} a_k z^{-k}$. We will use the convention that \mathscr{Z} -transforms are indicated by capital letters: $\mathscr{Z}(a) = A(z)$. Note that we use here the system engineering convention that the \mathscr{Z} -transform is defined as above, while the mathematical convention is that z is replaced by z^{-1} . We shall define the Fourier transform correspondingly. Thus $\mathscr{F}(a) = A(e^{i\omega}) = \sum_{k \in \mathbb{Z}} a_k e^{-ik\omega}$. The Fourier transform is an isometric isomorphism be-tween ℓ_2 and L_2 . Although an integrable function $F \in L_1$ has a Fourier series $\sum_{k \in \mathbb{Z}} f_k e^{-ik\omega}$ with $f_k = (1/2\pi) \int_{-\pi}^{\pi} F(e^{i\omega}) e^{ik\omega} d\omega$, the partial sums need not converge in norm or pointwise. The Cesàro sums are summation techniques, for example the Fejér sums that take the average over the first npartial sums, and these have better convergence properties than the ordinary partial sums. The Hardy space $H_p = H_p(\mathbb{D})$ $(1 \le p \le \infty)$ is the subspace of L_p of the functions F whose *positive* Fourier coefficients f_k with $k \ge 1$ vanish. Because the series $\sum_{k=0}^{\infty} f_{-k} z^k$ converge for $z \in \mathbb{D}$, these functions have an analytic extension in D. Similar to Hardy spaces, one can define a closed subset of the continuous functions on \mathbb{T} namely $\mathscr{A} = \mathscr{A}(\mathbb{D}) = \{F \in C(\mathbb{T}): f_{-n} = \int_0^{2\pi} F(e^{i\omega}) e^{in\omega} d\omega = 0, n = 1, 2, ...\}$ which is the *disk algebra*. Again such functions can be extended analytically to \mathbb{D} . The closure of the polynomials in L_p give H_p for $1 \le p < \infty$ but it gives \mathscr{A} for $p = \infty$. The inclusion $\mathscr{A} \subset H_{\infty}$ is proper.

To describe the system theoretic topics of this paper, let us now introduce some terminology from system theory. Mathematically, a *linear system* is a linear operator transforming an input into an output. Depending on the spaces where the input and the output live, there are several possibilities. A system has an input signal u_t say, and an output signal y_t . The variable t stands for "time". If t ranges over a continuous set (like \mathbb{R}) we have a *continuous time* system, otherwise (e.g., $t \in \mathbb{Z}$) we have a *digital* or *discrete time* system. Here we shall restrict ourselves to *discrete time* systems.

The system can be *single input-single output* (SISO) if the input and output are scalar. Otherwise, if they are in higher dimensional vector spaces, it is a *multi-input-multi-output* system. We restrict ourselves mainly to the simple SISO case.

There are two settings for these systems: the I/O signals can be stochastic or they can be purely deterministic, or even a mixture, like a deterministic signal with stochastic noise. In this paper, we look mainly at the purely deterministic case.

So we have excluded a lot of possible problems, but the choice that remains is still overwhelming. For the areas that are not treated, a specialist would know other problems analogous to those discussed here. The nonspecialist will find enough material to start with.

In the simplest possible form, a system can be described as a convolution: $y = T_h(u) = h * u$, thus $y_k = \sum_i h_{k-i}u_i$, $k \in \mathbb{Z}$, where $u = (u_k)$, $y = (y_k)$ are the input and output signals, and $h = (h_k)$ is called the *impulse response* of the system. The h_k are also known as *Markov parameters*. Taking \mathscr{Z} -transforms, we can write the I/O relation as Y(z) = H(z)U(z). If the series H(z) converges, then H(z) represents a function that is called the *transfer function* of the system.

Unless otherwise implied, we shall assume that we are dealing with *causal* systems, which means that the impulse response $h_k = 0$ for k < 0. If a sequence (s_k) has finite energy $\sum_k |s_k|^2 < \infty$, then it means that $s \in \ell_2$. A system is called *stable* if the I/O operator $T_h: u \mapsto y = T_h(u)$ is bounded. If $u, y \in \ell_2$ then the system is *stable* if $h \in \ell_\infty$, in other words, according to our convention, a stable causal ℓ_2 system will have a transfer function H such that H(1/z) is a bounded analytic function in the open unit disk: $H \in H_\infty(\mathbb{D})$.

The I/O signals (sequences) live in the *time domain* while the \mathscr{Z} -transforms live in the \mathscr{Z} domain or frequency domain. Indeed, if $z = e^{i\omega}$ in the \mathscr{Z} -transform, then H(z) is just the Fourier transform of $h = (h_k)$. The function $H(e^{i\omega})$ is called the frequency response and $|H(e^{i\omega})|^2$ is the *power spectrum*.

In many cases, it is very convenient to write a state-space formulation for the system. This means to define an intermediate vector x_k called the *state* such that

$$\begin{aligned} x_{k+1} &= Ax_k + Bu_k, \quad u_k, y_k \in \mathbb{C}, \ x_k, x_{k+1} \in \mathbb{C}^{d \times 1}, \\ y_k &= Cx_k + Du_k, \quad A \in \mathbb{C}^{d \times d}, \ B \in \mathbb{C}^{d \times 1}, \ C \in \mathbb{C}^{1 \times d}, \ D \in \mathbb{C}, \end{aligned}$$

where we shall assume that the dimension d of the *state space* is finite. The relation between state-space, time-domain, and frequency-domain formulations are (under appropriate conditions) given by

$$H(z) = C(zI - A)^{-1}B + D = \sum_{k} h_{k} z^{-k},$$

$$h_{0} = D, \quad h_{k} = CA^{k-1}B, \quad k = 1, 2, \dots$$

The state space is a space intermediate between input and output space. The previous equations for x_{k+1} , and y_{k+1} show one way in which the input is mapped to the state and the state is mapped to the output. In control problems it is of great importance to know how much of the state space can be reached from the input side and how much from the state space can be read off from the output. This is characterized to some extent by the controllability and observability Gramians, respectively. The controllability matrix C is the array whose kth column is $A^{k-1}B$, k = 1, 2, ... and the observability matrix O is the array whose kth row is CA^{k-1} , k = 1, 2, The corresponding Gramians are $P = CC^*$ and $Q = O^*O$. They solve the Lyapunov equations $APA^* - P = -BB^*$ and $A^*QA - Q = -C^*C$, respectively. (The superscript * means conjugate transpose.) Another important tool is the Hankel operator Γ_H of the system (see Section 3.2) which maps the past input onto the future output. With respect to the standard basis, it has a Hankel matrix representation $[h_{i+k-1}]_{i,j=1}^{\infty}$, which we can also write as OC.

The state-space formulation is very important for numerical computations and it is the most elegant way of dealing with generalizations like block Hankel matrices or MIMO systems or time-varying systems. State-space methods involve basically linear algebra techniques. Many of the most practical algorithms are based on state-space descriptions and linear algebra techniques. However, conceptually it is sometimes easier to work in function spaces. So, in order not to develop the ideas twice, we have made a choice in the different sections of this paper: sometimes we give a state-space description, sometimes the linear algebra aspects are deliberately neglected.

The general problem that is addressed here is:

Problem 1.1. (General problem) *Given some data about the system, find (a rational) approximation of the system.*

Of course, this problem has several "names". For example, in *model reduction* the aim is to find a simple (i.e., a low degree) approximation of a high-order system. In *realization* problems the objective is to find an explicit, implementable form of the system. Most often this is meant to be a state-space realization. *Identification* of the system means again that the system is pinned down in some way, and this is usually meant to be in the time domain or in the frequency domain. For instance, in the time domain this means that a sequence of (possibly perturbed) inputs u_k and/or outputs y_k is known. *Prediction* theory (or *forecasting*) is vaguely described as trying to identify or model or approximate a system for which we know only the output (like a seismic or a speech signal). In *control* theory one wants a system to generate a certain output, and if the actual output deviates from the desired one, then the input or the system itself is modified (controlled). To do this in an effective way, it is of course necessary to have at least an approximate model of the system, and thus the previous problems reappear.

The data that are given for the system can also have many different forms. Sometimes data are given for the behaviour of H(z) for z near ∞ (e.g., the first few Markov parameters $h_k = CA^{k-1}B$ which are the coefficients in the expansion of H(z) at $z = \infty$, which describe best the steady state behavior of the system, i.e., the behavior of the system for large t), or at the origin (e.g., the so-called time moments, or equivalently the coefficients of H(z) expanded at z = 0 which describe best the transient behavior of the system, i.e., the behavior of the system for small t), or on the unit circle (e.g., the power spectrum $|H(z)|^2$ for |z|=1). These data could be assumed exact, but they are in all practical situations contaminated by measurement error or model errors (for example nonlinear effects, while the model is linear, or an underestimation of the model order). Not only the data and the desired result but also the *approximation criterions* can be formulated in the time or frequency domain. That may be some kind of norm (like L_p norms) or some other criterion like interpolation with minimal degree or minimal norm, etc.

We want to give some samples from the last 50 years that we consider as stepping stones in the fertile interaction between system theory and rational approximation in the complex plane. We emphasize that it is never our ambition to be complete, and we do not want to deprecate any work we do not mention. We give some examples that are obviously colored by our personal interest. It should be sufficient though to illustrate our main point, namely the rich soil between system theory, rational approximation, linear algebra, numerical analysis, and operator theory that has been an incubation place for many new ideas and methods and will continue to be so in many years to come.

2. Realization theory

One of the main achievements in system theory of the 1960s was the solution of the so-called *minimal partial realization* problem.

Problem 2.1. (Minimal partial realization) Given the expansion $H(z) = h_0 + h_1 z^{-1} + \dots + h_N z^{-N} + \dots$ for $z \to \infty$. Find a rational function \hat{H} of minimal degree satisfying $H(z) - \hat{H}(z) = O(z^{-(N+1)})$ as $z \to \infty$.

Writing $\hat{H}(z) = P_n(z)/Q_n(z)$ with P_n and Q_n polynomials, one can solve the linearized problem

$$H(z)Q_n(z) - P_n(z) = O(z^{-(N+1)}), \quad z \to \infty$$

instead of the proper problem. If this is written down coefficientwise in descending order, we see that the first n+1 equations give P_n in terms of Q_n and h_0, \ldots, h_n . The most obvious choice is N = n, so that generically the next n equations define the coefficients of Q_n , up to a multiplicative factor, as the solution of a homogeneous Hankel system of linear equations. If this Hankel matrix is not of full rank there can be more solutions. Then the ratio \hat{H} is still unique, but it may or may not be a solution of the proper problem.

This is equivalent with the construction of a diagonal Padé approximation at infinity [2]. In Padé approximation one considers rational approximants of general type (m, n) [that is (numerator, denominator)-degree] whose expansion matches a given series in its first m + n + 1 terms. The (m, n)-approximants are arranged in a table called Padé table.

The problem can be solved recursively, i.e., it is computationally cheap to update a solution for a given order N to a solution for order N + 1. This kind of algorithms corresponds to fast algorithms for the solution of Hankel matrices that were recognized as variants of the Euclidean algorithm [14]. These algorithms solve a Hankel system by solving all the systems corresponding to the leading principal submatrices of the Hankel matrix. In this sequence there may appear singular submatrices according to a certain "pattern" [29]. This phenomenon is well known in Padé theory and corresponds to singular blocks in the Padé table. Such singular blocks are always square and that explains the "pattern".

These algorithms are fast because they exploit the structure of the Hankel matrix. However, the linear algebra operations do not use orthogonal matrices and to maintain the structure of the matrix, pivoting is not allowed. Therefore, these algorithms potentially suffer from numerical instabilities. The leading principal submatrices may be nearly singular, which causes large rounding errors. This problem has initiated an intensive research in the 1990s about look-ahead techniques for such structured linear algebra problems. These algorithms test if some quantity is below a certain threshold, and in that case the matrix is considered as being singular, and an update in the recursion is postponed until the quantity raises above the threshold [25].

The coefficients that are computed by such an algorithm correspond to recurrence relations for formal orthogonal polynomials [6] and to the numerators and denominators in a continued fraction expansion of the transfer function, which are known in realization theory as *Cauer* fractions.

Note, however, that this technique does not guarantee stability of the approximant and that can be considered as a major drawback of this technique. The coefficients of the recurrence relation for the orthogonal polynomials, or equivalently of the continued fractions, can be used in a stability test. This is the so-called Routh algorithm: a method to test whether a certain polynomial is stable (in the sense of continuous time systems, i.e., having zeros in the left half-plane) or not. The parameters in the Routh algorithm can be restricted, such that the rational approximant that results is guaranteed to be stable. Another possibility is to start from a high-order denominator polynomial of a stable system. The Routh algorithm computes the Routh parameters and by truncation, one obtains a low degree denominator polynomial for the approximation. The free coefficients in the numerator are used to match as many Markov parameters as possible. This is an example of a Padé-type approximant: the denominator is given and the numerator is determined in Padé sense.

More on the topic of minimal partial realization and on Padé and multivariate Padé approximation can be found in the extended contributions in this volume by De Schutter, Wuytack, and Guillaume, respectively.

3. Model reduction techniques

In the model reduction problem one wants to approximate a linear system by one of lower McMillan degree. The methods used here are diverse and rely on different mathematical techniques, on different presentations of the system and on different objectives that are optimized. We give some examples.

3.1. Balanced truncation and projection of dynamics

The idea is to perform a similarity transformation on the state space, so that the observability and controllability are "in balance". This means that the controllability Gramian P and the observability Gramian Q are equal and diagonal (see [40]). In this balanced realization, a simple truncation (keeping the most important eigenvalues) does the actual model reduction. The balanced realization is obtained from an eigenvalue decomposition of the product $PQ = TAT^{-1}$. Assume that the d eigenvalues in Λ are ordered in decreasing order: $\Lambda = \text{diag}(\Lambda_1, \Lambda_2)$ with Λ_1 containing the $n \leq d$ largest eigenvalues. After the similarity transformation $(A, B, C, D) \rightarrow (TAT^{-1}, TB, CT^{-1}, D)$, we then keep only the the first n rows and/or columns to isolate the reduced system. The reduced system is stable if the original system is, but the reduced system is not balanced in the discrete case (it is balanced for continuous time systems though). We note that the eigenvalues of PQ are related to the Hankel singular values. These are the singular values of the Hankel matrix of the system $\Gamma_H = OC$ which maps past input into future output (see Section 3.2). Indeed one has

$$\lambda_i(PQ) = \lambda_i(CC^*O^*O) = \lambda_i(C^*O^*OC) = \sigma_i^2(\Gamma_H).$$

Thus, this technique of model reduction typically throws away the smaller Hankel singular values and keeps the most important ones.

There are also several extensions. Actually, whenever one has a theory in which two matrix equations (Lyapunov, Riccati) result in two positive-definite solutions P and Q, one can define a balancing transformation T, a so-called contragredient transformation, that transforms P and Q so that they are equal and diagonal. Examples are stochastic balancing (in which case the solutions to the so-called forward and backward Riccati equations of a stochastic system are used to obtain the balancing transformation, LQG-balancing (where one starts from the solutions of the Kalman filter
and LQR Riccati equations, balancing on the two solutions of the H_{∞} Riccati filtering and control equations, relative error balancing (see, e.g., [44,47] for details and references).

Balanced model reduction can be considered as one particular case of the so-called *projection of dynamics*. Hereto define the projection $\Pi = RL$, where $R, L^* \in \mathbb{C}^{d \times n}$ and $LR = I_n$. Π is the projection onto the range space of R along the null space of L. Now, the idea of projection of dynamics is that at each time instant, the state x(t) is approximated by $\Pi x(t) = RLx(t)$ so that

$$RLx_{k+1} \approx ARLx_k + Bu_k, \quad y_k \approx CRLx_k + Du_k,$$

or

 $z_{k+1} = (LAR)z_k + (LB)u_k, \quad y_k = (CR)z_k + Du_k$

with $z_k := Lx_k$.

Since $LR = I_n$, it is always possible to find a similarity transformation T such that L consists of the first n rows of T and R of the first n columns of T^{-1} . Hence the approximating system (LAR, LB, CR, D) is obtained by first transforming the model by choosing a particular basis transformation matrix T, giving the realization $(TAT^{-1}, TB, CT^{-1}, D)$, and then truncating the transformed state-space model by restriction to the first n rows and/or columns. The oldest methods that can be interpreted in terms of *projection of dynamics* are *modal decomposition and reduction methods*, in which case the similarity transformation T diagonalizes the matrix A. The eigenvalues of the matrix A, which are in case of a minimal realization also the poles of the system, are thus revealed. This allows us to easily choose the poles of the reduced order system as a subset of the poles of the original system, or sometimes poles of the reduced system are simply fixed in advance. This is similar to the Padé-type approximants mentioned at the end of Section 2. The advantage of these method is their simplicity, typically only requiring the solution of one or more least-squares problems.

3.2. Hankel norm approximation

The Hankel norm approximation problem was inspired by a paper of Adamjan, Arov and Krein that appeared in 1971 [1] and it is therefore also known as AAK approximation. It was worked out in the context of system theory in the late 1970s and early 1980s.

If the system is stable and causal, then the I/O operator T_h defined in Section 1 will map the past onto the future. If we assume that the input and output have finite energy, then T_h is a Hankel operator in ℓ_2 . In the \mathscr{Z} or Fourier domain, this corresponds to a Hankel operator Γ_H mapping $H_2 = H_2(\mathbb{D})$ (the Hardy space of the unit disk) to its orthogonal complement H_2^{\perp} . It is defined by $\Gamma_H = P_{H_2^{\perp}}M_H$, where $H(z) = \sum_{k=1}^{\infty} h_k z^{-k}$ for $z \in \mathbb{T}$ is the Fourier transform of the impulse response h, and M_H is the operator representing the multiplication with H, and $P_{H_2^{\perp}}$ is the orthogonal projection onto H_2^{\perp} . If the system is stable, then Γ_H is a bounded operator and hence $H \in L_{\infty}$. The function $H \in L_{\infty}$ is called the *symbol* of the Hankel operator Γ_H . Note that only the Fourier coefficients h_k with k > 0 are relevant to define the operator. Given a Hankel operator, then its symbol is only determined up to an additive arbitrary H_{∞} function.

The representation of Γ_H with respect to the standard bases $\{z^k\}$ is a Hankel matrix whose entries are the Markov parameters.

Problem 3.1. (Hankel norm/AAK approximation) Given Γ_H , find an approximation $\Gamma_{\hat{H}}$ such that we solve one of the following two problems:

- minimum norm problem: $\|\Gamma_{H-\hat{H}}\|_2$ is minimal with rank $\Gamma_{\hat{H}} \leq n$, or the
- minimum rank problem: rank $\Gamma_{\hat{H}}$ is minimal with $\|\Gamma_{H-\hat{H}}\|_2 \leq \varepsilon$ for given ε .

Of course, the minimal norm problem has a solution for general matrices that is given by a singular value decomposition (SVD), truncated after the first (i.e., the largest) *m* singular values. Thus if $\Gamma = \sum_k \sigma_k \langle \cdot, v_k \rangle w_k$ is the SVD of Γ with singular values $\sigma_1 \ge \sigma_2 \ge \cdots$, and corresponding Schmidt pairs (v_k, w_k) , then the best rank *m* approximant is given by keeping only the first *n* terms in this sum. The remarkable fact here is that for a Hankel matrix Γ_H , where this truncated approximation is in general not a Hankel matrix, we can find a Hankel approximant $\Gamma_{\hat{H}}$ that approximates equally well. Thus $\inf \{ \|\Gamma_H - \Gamma_{\hat{H}}\|_2 : \operatorname{rank} \Gamma_{\hat{H}} \le \sigma_{n+1}$.

The rank of any Hankel operator Γ_H is related to the degree of its symbol H by Kronecker's theorem (1890) which says that rank $\Gamma_H \leq n$ iff $H \in \mathcal{R}_n + H_\infty$ where \mathcal{R}_n is the subset of all rational L_∞ functions with at most n poles inside the open unit disk \mathbb{D} .

It is clear that if $\Gamma_{\hat{H}}$ approximates Γ_{H} , then \hat{H} should approximate H in some sense. However, the symbol of a Hankel operator can be in L_{∞} in general and therefore we have to say that the "Hankel norm" defined as $||H||_{\Gamma} = ||\Gamma_{H}||_{2}$ is not really a norm for H(z), unless we know that H has no component in H_{∞} . This Hankel norm is however closely related to the L_{∞} norm. A theorem by Nehari (1957) says, for example, that $||H||_{\Gamma} = ||\Gamma_{H}||_{2} = \inf\{||H - F||_{\infty}: F \in H_{\infty}\} = \operatorname{dist}(H, H_{\infty})$. So we arrive at the AAK theorem in its simplest form, which solves the minimal norm problem and gives an approximation result for the symbols of Hankel matrices.

Theorem 3.1. (Adamjan et al. [1]) Let Γ_H be a compact Hankel operator with Schmidt pairs (v_k, w_k) for the singular values σ_k . Then, with the notation introduced above

 $\inf\{\|\Gamma_H - \Gamma_{\hat{H}}\|_2: \operatorname{rank} \Gamma_{\hat{H}} \leq n\} = \inf\{\|H - \hat{H}\|_{\infty}: \hat{H} \in \mathcal{R}_n + H_{\infty}\} = \sigma_{n+1}.$

Let us introduce the \mathscr{Z} transforms $V_k = \mathscr{Z}(v_k)$ and $W_k = \mathscr{Z}(w_k)$. Then if $\sigma_n > \sigma_{n+1}$, there is a unique solution \hat{H} that is defined by

$$H - \hat{H} = \frac{\sigma_{n+1}W_{n+1}}{V_{n+1}} = \frac{\Gamma_H V_{n+1}}{V_{n+1}}.$$

The error function $E = H - \hat{H}$ satisfies $|E| = \sigma_{n+1}$ a.e. on \mathbb{T} .

Note that the solution gives a best approximation \hat{H} that is in L_{∞} , even if the given H has no H_{∞} part. So, to have a causal stable approximant, one should get rid of the H_{∞} part of \hat{H} . Since \hat{H} is rational, this could, in principle, be done by partial fraction decomposition, although this is not the most advisable way to be followed for numerical computations.

The trouble is that for the solution of the AAK problem, one needs to solve an SVD problem for an infinite Hankel matrix. However, when the rank of that matrix is finite, then the computations can be done on a finite number of data like, for example, the state-space description of a matrix. The most elegant solution of the problem came therefore from the state-space approach by Glover [26], which is a benchmark paper in this theory. The operation of balancing the state-space representation of the linear system, as briefly explained previously, is crucial in the state-space approach by Glover. In this paper it can also be found that generically (when all the Hankel singular values σ_i are distinct) one has for a system of degree *d* that is approximated by a system of degree *n* that $||H - \hat{H}||_{\infty} \leq 2 \sum_{k=n+1}^{d} \sigma_i(\Gamma_H)$.

3.3. H_2 -model reduction

The Hankel norm approximant discussed in the previous section models an approximating system that minimized the H_2 -deviation of the future outputs, given that the original and reduced system had the same past inputs. The H_2 -norm of the system is the H_2 -norm of the transfer function $||H||_2$ which equals the ℓ_2 -norm of the impulse response $(\sum_k |h_k|^2)^{1/2} = ||h||_2$. For a finite-dimensional state space, one can derive that this norm can also be expressed in terms of state-space quantities as $||H||_2^2 = B^*QB + D^*D = CPC^* + DD^*$. The physical interpretation of the H_2 -norm is that its square is the expected value of the power in the output signal, when the input is white Gaussian zero mean noise with unit variance, or in a deterministic model, a unit impulse. Approximation in the H_2 -norm means finding a least-squares approximation of the impulse response. It is also known that if \hat{H}_n is an optimal H_2 approximant with a pole p, then $(d^m/dz^m)(H(z) - \hat{H}_n(z))_{z=p^{-1}} = 0$, m = 0, 1. This necessary condition can and has been be exploited in algorithms. Compare with Section 4 on linear prediction where interpolation problems also satisfy least-squares optimality conditions. We shall stick here to the straightforward characterization, and so our problem is

Problem 3.2. (h_2 norm approximation) Given a system with transfer function H or with state space (A, B, C, D). Find a system with transfer function \hat{H} or state space $(\hat{A}, \hat{B}, \hat{C}, \hat{D})$ such that $||H - \hat{H}||_2$ is minimal where the approximating system can have degree n at most.

Finding $\hat{H}_n = \arg \min ||H - H_n||_2$ where H_n ranges over all systems of degree *n* at most is a classical least-squares problem that is typically solved using an orthogonal basis and orthogonal projection techniques. It can be proved that the solution will generically have degree *n* since the minimum strictly decreases as *n* increases. This result is valid even for local minimizers. The H_2 -norm squared is clearly differentiable everywhere and therefore necessary conditions for minimality may easily be derived by calculating the gradient of the square of the H_2 -norm with respect to the state-space matrices of the candidate approximant $(\hat{A}, \hat{B}, \hat{C}, \hat{D})$. It can also be shown that the state-space model of an H_2 -optimal approximant can be characterized in terms of *projection of dynamics*. This characterization is useful in deriving gradient-based optimization algorithms (none of which can, of course, guarantee that a global minimum will be found) or in deriving *homotopy*-based optimization methods. For an excellent survey and many references we refer to [44].

A quite interesting derivation of properties, characterization of optimality, and even algorithms goes via linear algebra. Here we reconsider the Markov parameters $h_k = CA^{k-1}B$, k = 1, 2, Let $p(\lambda)$ be the characteristic polynomial³ of A, i.e., $p(\lambda) = \det(\lambda I - A) = \lambda^n + \alpha_1 \lambda^{n-1} + \cdots + \alpha_{n-1} \lambda + \alpha_n$. Then, it follows from the Cayley–Hamilton theorem that the h_k satisfy a *n*th-order recurrence relation, and thus the $p \times q$ Hankel matrix $\mathscr{H}_{p,q} = [h_{i+j-1}]_{i=1,...,p}^{j=1,...,p}$ is rank deficient, when both p > n and q > n.

³We do not make the distinction here between the characteristic and the minimal polynomial. 'Generically', they coincide. But if not, the presentation here can be refined.

Now, let $h_1, h_2, h_3, \ldots, h_N$ be N given data. We would like to model these datapoints by the impulse response of an *n*th-order linear system, where n is (a typically small integer) specified by the user. Using the rank deficiency result of the Hankel matrix, we can formulate this least-squares approximation problem as $\min \sum_{k=1}^{N} (h_k - \hat{h}_k)^2$, where the minimum is taken over all $\hat{h}_1, \ldots, \hat{h}_N$, subject to the condition that $\hat{\Gamma}$, the Hankel matrix of size $(N - n) \times (n + 1)$ with symbol \hat{H} , is rank deficient, i.e., $\hat{\Gamma}v = 0$, for some v, with $||v||_2 = 1$. When $N \to \infty$, this problem becomes the optimal H_2 -approximation problem we considered above. Let us however first treat the case where $N < \infty$. Then the problem is a so-called *structured total least-squares* problem, the structure being that $\hat{\Gamma}$ is required to be a Hankel matrix. The solution to this follows from the smallest singular value and corresponding singular vector of a nonlinear generalized SVD of the Hankel matrix $H \in \mathbb{R}^{(N-n)\times(n+1)}$ with the data:

$$Hv = D_v u\tau, \quad u^* D_v u = 1,$$

$$H^*u = D_u v\tau, \quad v^*D_u v = 1$$

and $v^*v = 1$, which is called the Riemannian SVD [19]. Specifically for the Hankel case, we refer to [18,20]. In these expressions, D_u and D_v are symmetric weighting matrices of appropriate dimensions and with a particular structure, the elements of which are quadratic functions of the elements of u, resp. v. For all values of u and v they are positive definite. More specifically, in this case D_u and D_v are symmetric, positive-definite banded Toeplitz matrices, the elements of which are quadratic in the components of u and v. The least-squares approximant \hat{H} of McMillan degree n now follows from the 'smallest' singular triplet (u, τ, v) that satisfies the Riemannian SVD equations, as $\hat{h} = h - (u\tau) * v$, where a * denotes the convolution of the vector sequences. Heuristic algorithms, remeniscent of and inspired by the power method to find the smallest singular value of a matrix, are described in [44,17]. For a translation to the \mathscr{Z} -domain of the H_2 problem and the Riemannian SVD problem as $N \to \infty$ we also refer to [44].

3.4. The Hilbert-Schmidt-Hankel norm

The Hilbert–Schmidt–Hankel norm is yet another performance criterion. It is defined as the Frobenius norm of its Hankel operator: $||H||_{\text{HSH}}^2 = ||\Gamma_H||_F^2$. It immediately follows from results in the previous sections that for a system of degree d

$$|H||_{\text{HSH}}^2 = \sum_{i=1}^n \sigma_i^2(\Gamma_H) = \text{Trace}(PQ) = \sum_{i=1}^\infty \sum_{j=1}^\infty h_i \bar{h}_j = \sum_{i=1}^\infty i |h_i|^2.$$

The last equality implies the interpretation of the HSH-norm of a system as a time-weighted H_2 -norm: It can be considered as the energy storage capacity, i.e., the expected energy stored in the system at a certain time, when the system has been driven by white zero mean unit variance Gaussian noise up to that time [44, p. 28].

A relatively little known interpretation of the HSH-norm is that it is the area enclosed by the oriented Nyquist plot of the linear system in the complex plane (see [32] for an elaboration and a proof). Hence, when doing model reduction in the HSH-norm, one tries to minimize the area between the Nyquist plot of the given system H(z) and its HSH-optimal approximant of lower McMillan degree.

From what has been discussed about model reduction so far, it should be clear that the essence of model reduction in the HSH norm is again a problem of finding a least-squares rank deficient approximation of a Hankel matrix, that is still a Hankel matrix. It thus should come as no surprise that this problem is again a so-called structured total least-squares problem. For further details, see [44].

Finally, we mention that for SISO systems in discrete time, we have $||H||_2 \leq ||H||_{\infty}$, but there is no upper bound for the H_{∞} -norm in terms of the H_2 -norm. This means that these two norms are not equivalent. However the HSH-norm, the Hankel norm and the H_{∞} -norm are equivalent:

$$||H||_{\Gamma} \leq ||H||_{\mathrm{HSH}} \leq \sqrt{n} ||H||_{\Gamma} \quad \text{and} \quad \frac{1}{2n} ||H||_{\infty} \leq ||H||_{\mathrm{HSH}} \leq \sqrt{n} ||H||_{\infty},$$

where n is the degree of H (see, e.g., [44]).

4. Linear prediction

The aspect of prediction theory is even older and dates back to the theory of Wiener [48], Wiener-Masani [49], Grenander and Szegő [30], and it was revived in the 1980s by Dewilde and Dym [21]. The problem can be described as follows [12]. Suppose that for an arbitrary n we observe a stationary signal $\{h_k\}_{k=-\infty}^{n-1}$ up to time moment n-1 and we want to predict the next value h_n . In other words, we should find a model for the signal h. Assume that h_n is predicted as $\hat{h}_n = -\sum_{k=1}^{\infty} a_k h_{n-k}$, then the prediction error $e_n = h_n - \hat{h}_n$ should be minimized in ℓ_2 sense. Taking \mathscr{Z} transforms, we get E(z) = A(z)H(z), where $A(z) = 1 + a_1z^{-1} + a_2z^{-1} + \cdots$. Minimizing the L_2 -norm $||E|| = ||AH|| = ||A||_{|H|^2}$ leads to a weighted least-squares problem formulated and solved by Szegő. To give an easy formulation, we introduce $F = A_*$ where the substar is used to denote the parahermitian conjugate $A_*(z) = \overline{A(1/\overline{z})}$. Note that F(0) = 1.

Problem 4.1. (Szegő) Given some positive measure μ on \mathbb{T} , find a function $F \in H_2^{\mu}$ that solves $\inf\{\|F\|_{\mu}^2: F \in H_2^{\mu}, F(0) = 1\}.$

The norm is taken in the Hilbert space L_2^{μ} with inner product

$$\langle f,g\rangle_{\mu} = \int_{\mathbb{T}} f(t)\overline{g(t)} \,\mathrm{d}\mu(t)$$

and H_2^{μ} is the subspace of functions analytic in \mathbb{D} . In the prediction problem, the measure μ is called spectral measure since the weight is the power spectrum: $d\mu(e^{i\omega}) = (|H(e^{i\omega})|^2/2\pi) d\omega$. The Fourier expansion for the power spectrum $W = |H|^2$ is $\sum_k c_k t^{-k}$, $t \in \mathbb{T}$ where the c_k are the autocorrelation coefficients of the signal h. These coefficients satisfy $c_{-k} = \bar{c}_k$. Since $|H|^2 \ge 0$, the Toplitz matrix $T_W = [c_{i-j}]$ is positive definite. Note also that $||A_*||_{\mu} = ||A||_{\mu}$ for any A and any positive measure on \mathbb{T} .

Once the predictor $A = F_*$ is known, we can invert the whitening filter relation E = AH and model H as E/A. However, E is not known, but if the prediction is good, then we can assume that it has a relatively flat spectrum. If the infimum in the Szegő problem is $G^2 \ge 0$, then we can approximate H by $\hat{H} = G/A$. Thus, the signal h is approximated as the impulse response of a system with transfer function \hat{H} which is called the modeling filter.

Szegő's theorem says that if $W \in L_1(\mathbb{T})$, then the optimal predictor is \hat{H} with $\hat{H}_*(z) = F(z) := S(0)/S(z)$ where for $c \in \mathbb{T}$ and $z \in \mathbb{D}$

$$S(z) = c \exp\left\{\frac{1}{4\pi} \int_{\mathbb{T}} D(t,z) \log W(t) \,\mathrm{d}\omega\right\}, \quad D(t,z) = \frac{t+z}{t-z}$$

is an outer spectral factor of W. If W is rational, then this means that S has no zeros and no poles in \mathbb{D} and $|S(t)|^2 = W(t)$ for $t \in \mathbb{T}$.

The practical computation of this infinite-dimensional problem is by computing the solution of the prediction (alias Szegő) problem in finite-dimensional subspaces of H_2^{μ} . Let \mathscr{L}_n be (n+1)-dimensional subspaces that are nested $\cdots \subset \mathscr{L}_n \subset \mathscr{L}_{n+1} \subset \cdots$ and such that the L_2^{μ} -closure of $\bigcup \mathscr{L}_n$ is H_2^{μ} . We then try to solve the partial Szegő problem

Problem 4.2. (Partial Szegő) Find a function $F_n \in \mathscr{L}_n \subset H_2^{\mu}$, with $F_n(0) = 1$ such that we solve one of the following problems:

- minimum norm problem: given n, find F_n that minimizes $||F_n||_{\mu}$, or
- minimum degree problem: given ε , find F_n with smallest n such that $||F_n||_{\mu} \leq \varepsilon$.

If $\{\phi_k\}_{k=0}^n$ is an orthonormal basis for \mathscr{L}_n , $n=0,1,\ldots$, then $k_n(z,w) = \sum_{k=0}^n \phi_k(z)\overline{\phi_k(w)}$ is a reproducing kernel for \mathscr{L}_n . This means that $\langle f, k_n(\cdot, w) \rangle_{\mu} = f(w)$ for all $f \in \mathscr{L}_n$. The solution F_n of the minimum norm partial Szegő problem in \mathscr{L}_n is then given by $k_n(z,0)/k_n(0,0)$ and the infimum is $1/k_n(0,0)$.

In the original theory of Grenander, Wiener and Szegő, the subspaces were the polynomial subspaces: $\mathscr{L}_n = \Pi_n$, the polynomials of degree at most *n*. Then, it can be shown that $k_n(z,0) = \kappa_n \phi_n^{\#}(z) = \kappa_n z^n \phi_{n*}(z)$ with $\{\phi_n\}$ the orthonormal polynomials and $\kappa_n > 0$ the leading coefficient of ϕ_n . Thus $\kappa_n = \phi_n^{\#}(0)$ and so $k_n(0,0) = \kappa_n^2$, giving $\hat{H}_n(z) = z^n/\phi_n(z)$. Note that the zeros of ϕ_n are in \mathbb{D} so that this is a stable and minimal phase transfer function, or in mathematical terms a conjugate outer function since all its poles and zeros are in \mathbb{D} . This solution also results in a *Chebyshev–Padé* approximation of W(z) since indeed the Fourier series of $W_n = |\hat{H}_n|^2$ is $\sum_k \hat{c}_k e^{-ik\omega}$ where $\hat{c}_k = c_k$ for all |k| < n + 1. Moreover, it can be shown that $d\mu_n = (W_n/2\pi) d\omega$, solves the partial trigonometric moment problem as formulated below.

Problem 4.3. (Trigonometric moments) Given c_k , $k \in \mathbb{Z}$, find a positive measure on \mathbb{T} such that $c_k = \int_{\mathbb{T}} t^k d\mu(t)$, $k \in \mathbb{Z}$. The partial problem is: given c_0, \ldots, c_n , find a positive measure μ_n such that it has exactly these moments or equivalently such that $\langle f, g \rangle_{\mu} = \langle f, g \rangle_{\mu_n}$ for all $f, g \in \Pi_n$.

Thus $\int_{\mathbb{T}} t^k d\mu_n(t) = \int_{\mathbb{T}} t^k d\mu(t)$ for |k| < n + 1. The Riesz-Herglotz transform of μ , given by

$$C(z) = \int_{\mathbb{T}} D(t,z) \,\mathrm{d}\mu(t), \quad D(t,z) = \frac{t+z}{t-z}$$

is in the Carathéodory class $\mathscr{C} = \{C \in H_{\infty}: \text{Re } C(z) \ge 0, z \in \mathbb{D}\}$, and it has the expansion $C(z) = c_0/2 + \sum_{k=1}^{\infty} c_k z^k, z \in \mathbb{D}$. The c_k are the trigonometric moments. If C_n is the Riesz–Herglotz transform of μ_n , then $C(z) - C_n(z) = O(z^{n+1})$ for $z \to 0$, so that we have also solved a Carathéodory–Fejér interpolation problem (see below).

When the construction of the orthogonal polynomials are formulated in terms of linear algebra, then the coefficients of the orthogonal polynomials are obtained as the solution of the Yule–Walker equations, which are in fact the normal equations for the least-squares problem posed by the linear prediction. The matrix of the system is a (positive-definite Hermitian) Toeplitz matrix, and again, like in the Hankel case, fast algorithms exist that solve subsystems by considering the leading principal submatrices of the Toeplitz matrix. For the duality between Hankel and Toeplitz systems and the associated (formal) orthogonal polynomials, see [13]. However, here the Toeplitz matrices are positive definite, unlike the Hankel systems in the partial realization problem. Therefore, in the linear prediction problem we are not confronted with a numerical and system theoretical instability problem as in the partial realization problem.

The well-known Levinson algorithm is a fast (i.e. $O(n^2)$) algorithm to solve Toeplitz systems. It is a version of the Gram–Schmidt orthogonalization that is fast because it exploits the structure of the Toeplitz matrix. Since the solution of the system is computed recursively, the algorithm computes as a bonus the so-called *reflection coefficients* that are related to the recurrence coefficients that occurred in the recurrence relations for the orthogonal polynomials as derived by Szegő. They are also called Schur coefficients because they also occur in a continued fraction like algorithm that was designed by Schur to see whether a given function is in the Schur class, that is the class $\mathscr{S} = \{f \in H_{\infty} : ||f||_{\infty} \leq 1\}$. The Schur algorithm is based on the simple lemma that $f_k \in \mathscr{S}$ iff $\rho_k = f_k(0) \in \mathbb{D}$ and $f_{k+1} = (1/z)[f_k - \rho_k]/[1 - \bar{\rho}_k f_k] \in \mathscr{S}$. These ρ_k are the Schur parameters. When translated in terms of linear algebra we can say that the Levinson algorithm gives a L^*DL factorization of T_W^{-1} while the Schur algorithm gives an LDL^* factorization of T_W where T_W is the Toeplitz matrix with symbol W. The Jury test to check the discrete stability of a polynomial (i.e., to see whether all its zeros are inside the unit circle) can also be seen as a variant of the Schur algorithm.

The mathematical ideas that we have just described were developed around the beginning of the 20th century. The multipoint generalization by Nevanlinna and Pick was published around 1920. The breakthrough of this multipoint generalization in signal processing, system theory, inverse scattering, transmission lines, etc., came not before the 1980s and was related to the introduction of the AAK ideas.

Let us reconsider the linear prediction problem, but now we take for \mathscr{L}_n subspaces of rational functions to be defined as follows. Let $\{z_k\}_{k=1}^{\infty}$ be a sequence of not necessarily different points in \mathbb{D} and set $\pi_0 = 1$ and $\pi_n(z) = \prod_{k=1}^n (1 - \overline{z}_k z)$. The spaces \mathscr{L}_n are then defined as the rational functions of degree at most *n* whose denominator is π_n : $\mathscr{L}_n = \{f = p_n/\pi_n: p_n \in \Pi_n\}$. Defining the Blaschke factors $\zeta_k(z) = \eta_k(z - z_k)/(1 - \overline{z}_k z)$, k = 1, 2, ... where $\eta_k = 1$ if $z_k = 0$ and $\eta_k = -\overline{z}_k/|z_k|$ otherwise, then it is obvious that \mathscr{L}_n is spanned by the Blaschke products $B_0 = 1$ and $B_k = \prod_{i=1}^k \zeta_i, \ k = 1, ..., n$. Note that if we choose all $z_k = 0$, then we are back in the polynomial case.

Following the same lines as above, we can construct an orthogonal basis by Gram–Schmidt orthogonalization of the $\{B_k\}$. Let us denote the orthonormal basis as $\phi_k = a_{k0}B_0 + a_{k1}B_1 + \cdots + a_{kk}B_k$, with $\kappa_k = a_{kk} > 0$. To solve the partial linear prediction problem in \mathcal{L}_n , we then construct $k_n(z, 0)/k_n(0, 0)$ where $k_n(z, w) = \sum_{k=0}^n \phi_k(z)\overline{\phi_k(w)}$. However, in the general case $k_n(z, 0)$ will not simplify as in the polynomial case so that we are stuck with the expression $\hat{H} = 1/K_{n*}$ with $K_n(z) = k_n(z, 0)/\sqrt{k_n(0, 0)}$ but this is again a minimal phase and stable transfer function. Indeed, \hat{H}_n is of the form $\prod_{k=1}^n (z-z_k)/P_n(z)$ where P_n is a polynomial with all its zeros in \mathbb{D} .

All the approximation properties that we had before are transformed into multipoint versions. For example, if $W_n = |\hat{H}_n|^2$ and $d\mu_n = (W_n/2\pi) d\omega$, then $\langle f, g \rangle_{\mu} = \langle f, g \rangle_{\mu_n}$ for all $f, g \in \mathscr{L}_n$. This means that we solve a partial moment problem in $\mathscr{L}_n \cdot \mathscr{L}_{n*}$ where the moments are given by $c_k = \int_{\mathbb{T}} B_k(t) d\mu(t), k \in \mathbb{Z}$ with $B_{-k} = B_{k*} = 1/B_k$. The Schur interpolation algorithm is replaced by the Nevanlinna–Pick (NP) interpolation algorithm, which solves the NP interpolation problem.

Problem 4.4. (Nevanlinna–Pick interpolation) Given $z_0 = 0, z_1, z_2, ..., z_n \in \mathbb{D}$, and $w_0 = 0, w_1, w_2, ..., w_n \in \mathbb{C}$, find a function $F \in H_{\infty}$ such that $F(z_k) = w_k$, k = 0, 1, 2, ..., n. For the partial problem, n is finite, for the full problem, n is infinite. If there are more solutions, one can characterize all the solutions and solve one of the following two problems:

- minimum norm problem: find a solution with minimal norm $||F||_{\infty}$ or the
- minimum degree problem: among the solutions with $||F||_{\infty} < \varepsilon$, find one of minimal degree.

This formulation corresponds to mutually different points z_i , but it is not difficult to imagine what the confluent case involves. If several of the z_k points coincide, then it is, in fact, a reformulation of an Hermite–Fejér problem and if all the points coincide at $z_k = 0$, then the NP algorithm becomes the Schur algorithm and the above problem becomes a Schur or Carathéodory–Fejér problem, although the latter is usually formulated by conformally mapping the function values from the unit circle to the right half-plane, so that the Schur class \mathscr{S} is replaced by the Carathéodory class \mathscr{C} . The original NP and Schur algorithms just checked whether some F was a Schur function, thus whether there is a solution with $||F||_{\infty} \leq 1$. Like the Schur algorithm, the NP algorithm is based on a simple lemma that is a slight generalization of the Schur lemma: $f_k \in \mathscr{S}$ iff for some $z_k \in \mathbb{D}$, $\rho_k = f_k(z_k) \in \mathbb{D}$ and $f_{k+1} = (1/\zeta_k)[f_k - \rho_k]/[1 - \bar{\rho}_k f_k] \in \mathscr{S}$.

This is a good place to introduce the *Nehari problem* since it can be seen as a generalization of the NP problem and hence also of the Schur problem.

Problem 4.5. (Nehari) Given $\{h_k\}_{k=1}^{\infty}$, find the function $H \in L_{\infty}$ such that $||H||_{\infty}$ is minimal and $h_k = (1/2\pi) \int_0^{2\pi} H(e^{i\omega}) e^{ik\omega} d\omega, \ k = 1, 2, \dots$.

If we define B_n as the Blaschke product with zeros z_1, \ldots, z_n and $G \in H_\infty$ as a function that satisfies the partial NP interpolation conditions, then the set of all functions in H_∞ satisfying the partial NP interpolation conditions is given by $G + B_n H_\infty$, and a minimal norm solution F is given by solving the Nehari problem $\inf\{||B_n^{-1}G - H||_\infty : H \in H_\infty\}$ and setting $F = G - B_n H$. This minimal norm solution is rational. As in previous problems, one can require that (after rescaling) $||F||_\infty \le 1$, and then find the rational solution with minimal degree.

The Nehari problem is particularly important for applications in control theory, where it usually appears under a slightly modified form which is a minimum degree version: find a solution satisfying $||H||_{\infty} \leq \varepsilon$ for a given ε . By appropriate rescaling, one can reduce this problem to the standard form where $\varepsilon = 1$. If there is more than one solution then one could select the one with the minimal degree.

The relation between the Nehari problem and the minimal norm AAK problem should be clear. Given a Hankel operator Γ_H , i.e., the numbers $\{h_k\}_{k=1}^{\infty}$, find a function $E \in L_{\infty}$ such that $h_k =$ $(1/2\pi) \int_0^{2\pi} E(e^{i\omega}) e^{ik\omega} d\omega$ and such that it has minimal norm $||E||_{\infty} = \sigma_{n+1}$. The latter means that *E* is equal to σ_{n+1} times an all pass function, or in mathematical terms, a function that is unimodular on \mathbb{T} almost everywhere, and the solution of the AAK problem is then $\hat{H} = H - E$. Note that if *H* is rational, then *E* is rational as well. Thus the all pass function is a Blaschke product and thus is *E* equal to σ_{n+1} times a Blaschke product.

5. Chain scattering matrices and H_{∞} control

Consider a 2 × 2 matrix valued functions M(z), and suppose J is a constant diagonal matrix with diagonal entries +1 and -1. Then M is called J-unitary if $M(z)^*JM(z)=J$. The matrix is called J-contractive if $J - M(z)^*JM(z) \ge 0$ where the inequality is to be understood in the sense of positive-definite matrices.

In a more general context, *J*-unitary matrices were studied by Potapov [43]. It was only since 1978 [24,21,22] that Dewilde and coworkers used the full power of *J*-unitary matrices in prediction theory and related matters like transmission lines, and inverse scattering. In fact, the recurrence relation for the orthogonal rational functions ϕ_n and for the reproducing kernels $k_n(z, w)$, of the previous section can be written in the form of a matrix relation. For example with $K_n(z) = k_n(z,0)/\sqrt{k_n(0,0)}$ and $K_n^{\#}(z) = B_n(z)K_{n*}(z)$, there exist *J*-unitary matrices $\theta_n(z)$ such that

$$\begin{bmatrix} K_{n+1}^{\#}(z) \\ K_{n+1}(z) \end{bmatrix} = \theta_n(z) \begin{bmatrix} K_n^{\#}(z) \\ K_n(z) \end{bmatrix} = \theta_n(z) \cdots \theta_0(z) \begin{bmatrix} K_0^{\#}(z) \\ K_0(z) \end{bmatrix} = \Theta_n(z) \begin{bmatrix} 1 \\ 1 \end{bmatrix},$$

where we assumed an appropriate normalization: $\int_{\mathbb{T}} d\mu(t) = 1$, so that $\phi_0 = 1$ and hence $K_0 = K_0^{\#} = 1$. Since the product of *J*-unitary matrices is a *J*-unitary matrix, the matrix Θ_n represents a scattering medium consisting of *n* layers. At the boundary of layers *n* and *n*+1, K_n and $K_n^{\#}$ can be considered as incident and reflected wave on the *n* side of the boundary, while at side *n* + 1 of the boundary, we have K_{n+1} and $K_{n+1}^{\#}$. The θ_n are called *chain scattering matrices* (CSM) because if the medium consists of several layers, then the CSM for the whole medium is the product of the CSMs of each layer. Adding one extra layer just requires an extra θ factor. Since the matrix θ_n will depend on the part of the energy that is reflected and the part that is transmitted, it will depend on the reflection coefficients. In fact, this is the origin of the name reflection coefficient. The variable *z* enters as a delay operator representing the time needed for the wave to pass through and back an homogeneous layer. Physically, if the system is passive, i.e., if it does not add or absorb energy, then the CSM is *J*-unitary in T and *J*-contractive in D. It also explains why the reflection coefficients are bounded by 1 in modulus: they represent the fraction that is reflected.

In terms of electrical circuits [5], the θ matrices represent a 2-port (two I/O pairs) mapping one I/O pair into another I/O pair. A CSM is equivalent to a scattering matrix mapping inputs into outputs. A scattering matrix of a passive network is a unitary matrix on \mathbb{T} and contractive in \mathbb{D} , but the concatenation of 2-ports gives rise to a complicated star product for the scattering matrices, replacing the ordinary product of the CSMs.

The special structure of the θ matrices does not only give a direct lattice realization of the whitening filter (analysis) or modeling filter (synthesis), but they can even be used for the design of dedicated hardware implementation with systolic arrays [36].

The CSM can also play a prominent role in H_{∞} control. Let us consider a simple standard control problem

$$\begin{bmatrix} Z \\ Y \end{bmatrix} = \begin{bmatrix} P_{11} & P_{12} \\ P_{21} & P_{22} \end{bmatrix} \begin{bmatrix} W \\ U \end{bmatrix},$$
$$U = KY,$$

where Z is the errors to be controlled, W the exogenous input, Y the observed output, U the control input, P the plant matrix, K the controller.

The closed-loop transfer function from W to Z is

$$H = P_{11} + P_{12}K(I - P_{22}K)^{-1}P_{21}.$$

Thus if W is some disturbing noise, then the controller should ensure that the observed error Z is only influenced marginally.

Problem 5.1. (H^{∞} control) Find the controller K such that $||H||_{\infty} < \gamma$ and such that the system is internally stable.

Internal stability means that no fuses in the plant may burn, i.e., no state variable in the plant should become infinite. This problem can be reformulated in terms of CSMs. Indeed, the natural way to look at this problem is to consider the plant P as a 2-port (inputs W and U, outputs Z and Y) that is loaded by the controller: a 1-port (input Y, output U). The above description is a typical scattering matrix formulation, mapping inputs into outputs. However a CSM approach is much more interesting. Then the I/O pair (Y, U) is mapped into the I/O pair (W, Z). This gives

$$\begin{bmatrix} Z \\ W \end{bmatrix} = \Theta \begin{bmatrix} U \\ Y \end{bmatrix} \quad \text{with } \begin{bmatrix} Z \\ W \end{bmatrix} = \begin{bmatrix} H \\ I \end{bmatrix} W \quad \text{and} \quad \begin{bmatrix} U \\ Y \end{bmatrix} = \begin{bmatrix} K \\ I \end{bmatrix} Y.$$

Note that this relation expresses that *H* is a linear fractional transform of *K*, characterized by $\Theta: H = \text{LFT}(\Theta; K)$. If Θ is *J*-unitary, then it maps $K \in \mathscr{S}$ into $H \in \mathscr{S}$. Thus $||H||_{\infty} < \gamma$ if $||K||_{\infty} < \gamma$. The solvability of the control problem reduces to the existence of a *J*-lossless factorization. This is an indefinite matrix version of the classical inner-outer factorization of complex functions. Assume that Θ is the CSM representation of a plant and assume it can be factorized as $\Theta = \Delta \Pi$ with Π an invertible *J*-unitary matrix and Δ is the CSM of a stable plant, which means that it is *J*-contractive in \mathbb{D} . Thus Π absorbs the instabilities of the plant. Since Δ now has the property that any controller \tilde{K} , with $||\tilde{K}||_{\infty} < \gamma$ will solve the problem, we have an infinite set of controllers for the plant Θ given by $K = \text{LFT}(\Pi^{-1}; \tilde{K})$ since indeed $H = \text{LFT}(\Theta; K) = \text{LFT}(\Theta \Pi^{-1}; \text{LFT}(\Pi; K)) = \text{LFT}(\Delta; \tilde{K})$. This *J*-lossless factorization can be obtained by a Nevanlinna–Pick type of algorithm. For more details we refer to [38] where it is also shown that many other related control problems can be reduced to the present one or a vector generalization thereof.

6. Identification

We consider an identification problem in the frequency domain. Suppose we know the frequency response *H* at some specific points $\{t_i: i = 1, ..., N\} \subset \mathbb{T}$. Depending on what norm or what other objectives one wants to achieve, there are several different approximation problems to solve.

6.1. Discrete linearized least squares

As a first example, one can try to find a rational function (of smallest possible degree) that interpolates the given data, or if a restriction is given on the degree, find the one that interpolates in a discrete least-squares sense. The problem remains linear if we multiply out the denominator.

Problem 6.1. (Discrete linearized least squares) Given $\{t_i\}_{i=1}^N \subset \mathbb{T}$ and couples of numbers $\{(f_i, e_i)\}_{i=1}^N$ such that $H(t_i) = f_i/e_i$, i = 1, ..., N, find an approximant $\hat{H}(z) = N(z)/D(z)$ of degree at most n such that $\sum_{i=1}^N w_i |R_i|^2$ is minimal where $R_i = N(t_i)e_i - D(t_i)f_i$ and $\{w_i\}_{i=1}^N$ are positive weights.

The solution of this problem is again related to the Schur algorithm, but now with a discrete measure that has some mass concentrated only at the points t_i . For such a measure the Schur algorithm will break down after a finite number of steps because a reflection coefficient will become one in modulus. If the degree *n* is reached before all the interpolation conditions are satisfied, then the algorithm is artificially forced to an end, pushing some masses to the points t_i on \mathbb{T} . Linear least-squares problems are typically solved using orthogonal polynomials and can therefore be formulated solely in terms of linear algebra. A fast algorithm for discrete polynomial least-squares approximation on the real line is given by Forsythe (1957). It performs a Gram–Schmidt orthogonalization and stores the polynomials by their three-term recurrence relation. The analog for the unit circle uses the Szegő recurrence and stores the polynomials via their reflection coefficients [46]. More precisely, we write the solution as a polynomial vector

$$S(z) = \begin{bmatrix} D(z) \\ N(z) \end{bmatrix} = \sum_{i=0}^{n} \Phi_i(z) C_i, \quad C_i = \begin{bmatrix} \rho_i \\ \tau_i \end{bmatrix},$$

where the C_i are constant vectors and Φ_i are 2×2 polynomial matrices that are orthogonal in the following sense:

$$\langle \Phi_k, \Phi_l \rangle_W = \sum_{i=1}^N \Phi_k^*(t_i) W_i \Phi_l(t_i) = \delta_{k,l} I_2, \quad W_i = E_i^* v_i E_i, \quad E_i = [f_i - e_i].$$

The superscript * denotes the adjoint. We want to minimize $\langle S(z), S(z) \rangle_W$. The minimum is obtained for all $C_i = 0$, except for C_n , because the degree should be *n*. Choosing the two variables ρ_n and τ_n is a very simple problem because $\langle S(z), S(z) \rangle_W = |\rho_n|^2 + |\tau_n|^2$. The algorithm to generate the orthogonal polynomials Φ_i is a block version of a Szegő-like recurrence relation. It results in the following recursive matrix interpretation. Define the matrix *E* whose *i*th row is E_i , i = 1, ..., N and $Z = \text{diag}(t_1, ..., t_N)$. Then define

$$M = \begin{bmatrix} 0 | E^* \\ \overline{E} | Z \end{bmatrix} \text{ and } \tilde{M} = Q^* M Q = \begin{bmatrix} 0 | \sigma_0^* | 0 \\ \overline{\sigma_0} \\ 0 | H \end{bmatrix},$$

where Q is a unitary matrix, σ_0 is a 1×2 vector and H is a unitary upper Hessenberg matrix. It is unitary because it is a unitary similarity transformation of the original matrix Z, which is unitary. If \tilde{M} is extended with the new data E_{N+1} and t_{N+1} , then the same Hessenberg structure is restored by a unitary similarity transformation. This updating is fast because the Hessenberg matrix is stored by its block Schur parameters, and this update is performed very efficiently.

6.2. Robust identification

Robustness is an important issue in systems and control. It essentially means that it is not only sufficient to solve the problem within some tolerance, but the performance should remain within this tolerance in the worst-case situation when certain perturbations are allowed. For example, in the control problem we had to design a controller that generated a transfer H with $||H||_{\infty} < \gamma$. In robust control, this should be satisfied for all possible perturbations of the plant P that remain within certain bounds. Sometimes this robustness is obtained by formulating a weighted problem where the weight is chosen so as to emphasize the sensitive parts of the objective function. There is also a technique of structured singular values or μ -analysis where some singular value is monitored within some bounded perturbations of variables that cause a structured perturbation of the matrix. Because this is essentially a linear algebra technique, we will not go deeper into this matter. As an example of a robustness problem, we discuss here the robust identification problem.

Suppose that we know the frequency response *H* of a stable system in *N* points $t_i \in \mathbb{T}$, i = 1, ..., N. We can, of course, find an approximant \hat{H}_N and we want the algorithm to be such that $\hat{H}_N \to H$ for $N \to \infty$ for all stable H. By a mapping $z \mapsto z^{-1}$, we can reformulate the problem in the disk algebra \mathscr{A} . Thus setting $F = H_*$, we have $F \in \mathscr{A}$. Robustness means that we now allow the observed data to be contaminated by noise. Thus, we are given $F(t_i) + \eta_i$ with $|\eta_i| < \varepsilon$ (i.e., $||\eta||_{\infty} < \varepsilon$). We still want the algorithm to be such that in the worst case situation, the convergence still holds for $\varepsilon \rightarrow 0$. The problem is thus described as follows.

Problem 6.2. (Robust identification) Given is a function $F \in \mathcal{A}$. This function can be computed in points $t_i \in \mathbb{T}$ with some error η_i bounded by ε , so that we can compute $\tilde{F}_i := F(t_i) + \eta_i$, i = 1, ..., Nwith $t_i \in \mathbb{T}$ and $\|\eta\|_{\infty} < \varepsilon$. Design an algorithm $A_N : \mathcal{A} \to \mathbb{C}^N \to \mathcal{A} : F \mapsto \hat{F}_N$ that constructs an approximant \hat{F}_N using the values $\{\tilde{F}_i\}_{i=1}^N$ such that

$$\lim_{N\to\infty,\varepsilon\to 0} \sup_{\|\eta\|_{\infty}<\varepsilon} \sup_{F\in\mathscr{A}} \|\hat{F}_N-F\|_{\infty}=0.$$

It was shown by Partington in 1992 that there is no linear algorithm that solves this problem. So the problem is usually solved in two steps.

- (1) Find a *linear* algorithm V_N that constructs an approximant G_N in L_{∞} . (2) Find a rational approximant $\hat{F}_N \in \mathscr{A}$ of $G_N \in L_{\infty}$ that minimizes $||G_N \hat{F}_N||_{\infty}$.

The first problem is a problem of sampling theory: how much (i.e., what samples) do we need to know about a function to be able to recover it, if we know that the function is in a certain class. The second problem is in fact a Nehari problem, that has been discussed before (Problem 4.5). Suppose the approximant of F generated by V_N , given $\tilde{F}_i = F(t_i) + \eta_i$, i = 1, ..., N, is denoted as $V_N(F + \eta)$. Because V_N is linear and $\|\eta\|_{\infty} < \varepsilon$, we have

$$\begin{aligned} \|V_N(F+\eta) - F\|_{\infty} &\leq \|V_N(F) - F\|_{\infty} + \|V_N(\eta)\|_{\infty} \\ &\leq \|V_N(F) - F\|_{\infty} + \|V_N\|_{\infty} \varepsilon. \end{aligned}$$

Thus $V_N(F + \eta)$ will converge to F in the sense of robust identification in the set L_∞ if $\lim_{N \to \infty} ||V_N(F) - F||_\infty = 0$ and V_N is bounded in \mathscr{A} .

A simple example for the algorithm V_N is to choose t_l as the *N*th roots of unity $t_l = e^{2\pi i l/N}$ and to compute the discrete Fourier transform (DFT) for the given samples: $c_N(k) = (1/N) \sum_{l=1}^{N} f(t_l) t_l^{-k}$, $k \in \mathbb{Z}$, which defines an approximant $\tilde{G}_{n,N}(e^{i\omega}) = \sum_{k=-n}^{n} c_N(k)e^{ik\omega}$. However, this approximant does not in general converge to *F* when *n* and *N* go to infinity. We need a special summation technique, for example the Fejér sum, which takes the average of the $\tilde{G}_{k,N}$, $k = 1, \ldots, n$: $G_{n,N} = (1/n) \sum_{k=1}^{n} \tilde{G}_{k,N}$. This corresponds to the introduction of weight coefficients in the DFT, thus we use a windowed DFT: $G_{n,N}(e^{i\omega}) = \sum_{k=-n}^{n} w_k c_N(k) e^{ik\omega}$ with $w_k = 1 - |k|/n$. Other summation techniques (e.g., de la Vallée-Poussin) exist that correspond to other weights. Finally, the approximant G_N is then given as $G_N = \lim_{n \to \infty} G_{n,N}$.

This construction of trigonometric approximants should be generalized so that we need not necessarily take equidistant points on \mathbb{T} . However, it is intuitively clear (and a hard proof exists) that the points t_i should be eventually dense in \mathbb{T} if we want to recover F exactly. Of course, modifications can be made if we know that the signal is band limited, or if only approximants in a certain band are important, while outside that interval, the approximant may even diverge.

Anyway, when using trigonometric approximants, the convergence may be very slow, especially when there are poles close to \mathbb{T} . In that case it might be wise to use rational basis functions instead of (trigonometric) polynomials. In fact, it was for a similar reason that in prediction theory, the Nevanlinna–Pick algorithm replaced the Schur algorithm, so that AR models could be replaced by ARMA models.

Assume that we have some estimates of the poles. So we are given a sequence of (not necessarily distinct) $z_k \in \mathbb{D}$. The orthogonal rational functions ϕ_n with poles $\{1/\bar{z}_k\}_{k=1}^n$ as discussed in Section 4, but now made orthogonal with respect to the Lebesgue measure on \mathbb{T} are known explicitly (Walsh attributes them to Malmquist) and they are given by

$$\phi_n(z) = \frac{\sqrt{1-|z_n|^2}}{1-\overline{z_n}z} z \prod_{k=1}^{n-1} \frac{z-z_k}{1-\overline{z_k}z}, \quad n \ge 0.$$

It is known that span{ ϕ_k : k = 0, 1, ...} is dense in H_2 and in the disk algebra \mathscr{A} iff $\sum_{k=0}^{\infty} (1 - |z_k|) = \infty$. Some special cases became rather popular in identification: when all the z_k are equal to some fixed $a \in [-1, 1]$, then this system is known as the Laguerre system because it is related to a transformation of the Laguerre functions. This system was used by Wahlberg in 1991. In 1994 he also introduced the Kautz system based an a complex conjugate pair: $z_k = \overline{z}_{k+1} = a$. For a survey of all kinds of orthogonal (rational) bases and their use in system identification see [41]. Schipp and Bokor proposed yet another system which corresponds to a cyclic repetition of the poles $\{1/\overline{z}_1, \ldots, 1/\overline{z}_d\}$. It is constructed as follows. Let B_d be the Blaschke product with zeros z_1, \ldots, z_d and consider an orthonormal basis $\{\phi_1, \ldots, \phi_d\}$ of $B_d H_1(\mathbb{D})$. Then $\{\phi_l B_d^k: 1 \le l \le d; k = 0, 1, \ldots\}$ is an orthonormal basis for $H_2(\mathbb{D})$.

More recently, also rational wavelet-like bases were used to represent functions in the disk algebra. For example, the Franklin system is a piecewise linear L_2 -orthogonal system in $C(\mathbb{T})$. First let $\psi(\omega)$ be the hat function in $[0,\pi]$ (zero in 0 and π , one in $\pi/2$ and linear in between). Then define $\psi_{nk}(e^{i\omega}) = \psi(2^n\omega - k\pi)$ for $k = 0, ..., 2^n$ and $n \in \mathbb{N}$. These functions are orthogonalized to give a Faber–Schauder basis ϕ_{nk} for functions in $C(\mathbb{T}^+)$ where \mathbb{T}^+ means the upper half of \mathbb{T} . Because the trigonometric conjugate system $\tilde{\phi}_{nk}$ is also continuous, the functions $\Phi_{nk} := \phi_{nk} + i\tilde{\phi}_{nk}$, when extended to the lower half of \mathbb{T} by $\Phi_{nk*} = \Phi_{nk}$ and analytically extended to \mathbb{D} , form a basis for the disk algebra \mathscr{A} .

Another rational wavelet-like basis can be obtained by using the the Cauchy kernel $C(z, w) = 1/(1 - \bar{w}z)$. Therefore define the set $W = \{w_{nk} = \rho_n e^{i\omega_{nk}}: k = 0, ..., 2^n - 1; n = 1, 2, ...\}$ where $\rho_n = 1 - 2^{-n}$ and $\omega_{nk} = 2k\pi 2^{-n}$. The system $\{C(z, w): w \in W\}$ is dense in the disk algebra \mathscr{A} . In fact, this is an overcomplete system because it is sufficient that $\sum_{k=n} (1 - |w_{nk}|) = \infty$ for the system to be dense.

7. Generalizations: past and future

The basic idea relating *Padé-like approximations* and *partial realization* has been extended in many directions, e.g., to MIMO systems (giving rise to block Hankel matrices) and to two-point Padé approximations, (using information at 0 and ∞) and multipoint Padé approximation and general rational interpolation. Sometimes a combination of Markov parameters and correlation coefficients are fitted like in q-COVER [44]. Many generalizations of the Padé approximation and interpolation problem lead to new developments Laurent–Padé [7], vector-Padé [2]. Many *matrix valued rational interpolation* problems are for example discussed in [4] and related problems are discussed in several volumes of the same series, edited by Gohberg. It is also interesting to see how many techniques like state-space descriptions, controllability and observability matrices, are used in the analysis and solution of these problems. See [11] for a bibliography on Padé techniques in systems theory. The original matrix minimal partial realization problem got a first satisfactory solution in state space from, described by Ho and Kalman [35].

In many applications, notably in controller design applications, the objective of the model reduction problem can be weighted in the frequency domain. The reason is that often (especially in control system design) one is interested in a good match between the reduced model and the original one, at a certain frequency or in the neigborhood of a certain frequency (e.g. the so-called cross-over frequency in control system design). Therefore typically frequency-domain weighting matrices are included in the model reduction framework, so as to minimize the input-output weighted error $W_0(z)[H(z) - \hat{H}(z)]W_0(z)$. Here $W_0(z)$ and $W_0(z)$ are the output, resp. input weighting functions that emphasize certain frequency regions in which the approximation error should preferably be small. References to extensions in this direction to H_2 -optimal model reduction, Hankel norm approximation and balanced truncation can be found in [44, p. 33]. An important special case occurs when one of the weights in the weighted error is the inverse of the original system, in which case one is minimizing the so-called *relative error*. Balanced stochastic truncation is one such method that achieves balanced truncation in a relative error framework (see [44, Chapter 4] for a survey and some error bounds upper bounding the H_{∞} -norm of the 'relative' error to the so-called balanced stochastic singular values). For references on frequency weighted open- and closed-loop balanced truncation, we refer to [44, Chapter 3].

The generalization of the realization problem to the situation where only input-output data of the system are available, u_k , y_k , k = 0, 1, 2, ... (e.g., via measurements obtained from sensor devices), can be solved via so-called *prediction error methods* [39] or *subspace system identification methods* [47] (also check these references for literature surveys). Explaining this in detail would lead us too far. Suffice it to say that the identification problem is very important in many mathematical engineering problems in the process industry and that it leads to model-based control system design and optimization, softsensors, observers, etc.

Stimulated by the *AAK problem* from system theory, several related research projects were started in approximation theory. The fact that a Hankel norm approximation is, under certain conditions, nearly a best approximant in ∞ -norm, was exploited by Trefethen and Gutknecht to construct near best polynomial and rational approximants. They called it CF approximation because they started from the classical *Carathéodory–Fejér* theorem that deals with the polynomial case (i.e., n=0). The system theoretic approximant is again diagonal in the sense that the degree of numerator and denominator are equal. This has been generalized in many directions including an analysis of the structure of a CF table in analogy with the Padé table [33]. The equi-oscillation condition of the error curve in real Chebyshev approximation on an interval is replaced by the circularity of the error curve and the winding number is large enough. As it is usually assumed in computations, *H* is a rational of high degree and \hat{H} is a low degree approximant. Then the error $H - \hat{H}$ is σ_{n+1} times a Blaschke product. The winding number is then associated with the degree of this Blaschke product and this is related to the multiplicity of σ_{n+1} as a singular value [45].

There is a tendency to generalize one or two point interpolation (as in Padé approximation or in the Levinson algorithm) to more general *multipoint* (Hermite) *interpolation* problems (like in the Nevanlinna–Pick interpolation). This implies a shift from polynomial bases to rational bases. We gave some examples of this idea before. Here is another one. Recall that in the Nevanlinna–Pick problem, some points z_k were chosen which were used as interpolation points, but at the same time they featured as transmission zeros in the model. Recently, the problem was raised whether it is possible to keep all the nice properties of this approach but separate the role of interpolation points and transmission zeros [15]. The problem reduces to a constrained optimization problem to guarantee stability and yet obtain an interpolant that is of minimal degree. The search space is the set of all stable solutions to the interpolation problem which can, for example, be parametrized by the reflection coefficients.

The H_{∞} control problem has many facets and many different kind of subproblems. The discussion given in Section 5 is just a start. It should be obvious that the analysis of *J*-unitary matrices is essential. The fine structure for matrix and operator valued functions that are unitary and/or contractive with respect to an indefinite matrix was initiated by Potapov, but because it is so essential in all kind of generalizations of the Nevanlinna–Pick theory, of moment problems, and all the engineering applications, the study of these matrices and all the related problems has grown out into an independent discipline generally known as *Schur analysis*. Some generalizations of Potapov's work are found in [27]. The importance of *J*-unitary matrices is surveyed in [10] which contains many references, especially to the Russian literature. For all kind of generalizations of the Nevanlinna–Pick algorithm see also [16].

The linear algebra problems (Hankel matrices in realization and Toeplitz matrices in linear prediction) were solved by fast algorithms because the structure of these matrices could be exploited. When considering signals which are not stationary, then the covariance matrix is not Toeplitz, but if the nonstationarity is only mild, then the matrices do not deviate too much from a Toeplitz matrices. The structural deviation from a Toeplitz matrix could be measured by the *displacement rank*, introduced around 1979 and studied by Kailath and many others [34,37]. Designing stable and fast or superfast algorithms for all sorts of structured matrices is still an active field of research.

The *nonstationary prediction* problem has been generalized to the time-varying case. The easiest way to see what this means is to consider the state-space description with the matrices (A, B, C, D). The *time-varying* case is obtained when we let these matrices depend on the (time) variable k

[3,23,31]. Many of the concepts, from Hankel matrix to orthogonal polynomials have been generalized. There is however much more work to be done before these techniques will be widely accepted.

The algorithm of Section 6.1 is a unit circle analogue for discrete time systems of an algorithm that starts from data given on the real line and that is used for continuous time systems. In the latter case the problem is related to a Hankel matrix and because of the symmetry, the (block) Hessenberg matrix will be (block) tridiagonal. This idea was first elaborated by Ammar and Gragg in 1984. Further generalizations are available and many variants of the algorithm exist and several applications are found in different domains. Also the so-called UHQR (unitary Hessenberg QR) and the TQR (tridiagonal QR) algorithms are closely related. They solve not the least squares, but the eigenvalue problems for unitary Hessenberg or tridiagonal matrices. Many references related to this circle of ideas are found in [28].

There is much research still going on about the generalization of classical *sampling theorems*. This is also stimulated by the interest in wavelets. The ultimate question is: how much function values does one have to know of a function to be able to reconstruct the function perfectly. The use of other bases than the usual complex exponentials is thereby an essential element. Mathematically, this is related to density problems: will the span of a certain set of basis functions or a frame be dense in the space where we want to approximate? There are many recent results, but there are still a lot of questions to be answered.

As for the use of *orthogonal rational bases*, there is a lot of theory accumulated on orthogonal rational functions with respect to a general positive measure on the unit circle (and on \mathbb{R}) in [9]. This monograph gives convergence and density results, interpolation properties and generalized moment problems in the scalar case. It is very useful if the ideas of identification are used with weighted approximation conditions, thus using another basis than the one orthogonal with respect to the Lebesgue measure. Moreover, this theory can be generalized to the matrix case and although there are many partial results available in the literature, a systematic study is not published. Another way of potential application of the monograph lies in the fact that also orthogonal rational functions are discussed whose poles are on the unit circle (and not inside the open disk \mathbb{D}). This has many potential applications in systems and identification that has not been explored so far.

The use of *wavelets* in identification is just starting up and a lot of work has to be done here. For example, the orthogonal rational wavelets based on reproducing kernels [8] may open a new horizon. Among the problems that are only partially explored, one can count the problem of selecting a best basis in a dictionary of possible bases.

An extremely useful survey from an approximation theoretical point of view about rational approximation and system related problems is given in [42].

References

- V.M. Adamjan, D.Z. Arov, M.G. Krein, Analytic properties of Schmidt pairs for a Hankel operator and the generalized Schur-Takagi problem, Math. USSR-Sb. 15 (1971) 31–73.
- [2] G.A. Baker Jr., P.R. Graves-Morris, Padé Approximants, 2nd Edition, Encyclopedia of Mathematics and its Applications, Vol. 59, Cambridge University Press, Cambridge, 1996.
- [3] J.A. Ball, I. Gohberg, M.A. Kaashoek, Nevanlinna–Pick interpolation for time-varying input-output maps, in: I. Gohberg (Ed.), Time-Variant Systems and Interpolation, Vol. 56, Birkhäuser, Basel, 1992, pp. 1–51.

- [4] J.A. Ball, I. Gohberg, L. Rodman, Interpolation of rational matrix functions, Operator Theory: Advances in Applications, Vol. 45, Birkhäuser, Basel, 1990.
- [5] V. Belevitch, Classical Network Theory, Holden-Day, San Francisco, 1968, pp. 93,136,141. (Zbl 172.204).
- [6] C. Brezinski, Padé-type approximation and general orthogonal polynomials, Operator Theory: Advances in Applications, International Series of Numerical Mathematics, Vol. 50, Birkhäuser, Basel, 1980.
- [7] A. Bultheel, Laurent series and their Padé approximations, Operator Theory: Advances in Applications, Vol. OT-27, Birkhäuser, Basel, 1987.
- [8] 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 of the American Mathematical Society 236, (1999) 101–123.
- [9] A. Bultheel, P. González-Vera, E. Hendriksen, O. Njåstad, in: Orthogonal Rational Functions, Cambridge Monographs on Applied and Computational Mathematics, Vol. 5, Cambridge University Press, Cambridge, 1999.
- [10] A. Bultheel, K. Müller, On several aspects of *J*-inner functions in Schur analysis, Bull. Belgian Math. Soc. Simon Stevin 5 (1998) 603–648.
- [11] A. Bultheel, M. Van Barel, Padé techniques for model reduction in linear system theory, J. Comput. Appl. Math. 14 (1986) 401–438.
- [12] A. Bultheel, M. Van Barel, Linear prediction: mathematics and engineering, Bull. Belgian Math. Soc. Simon Stevin 1 (1994) 1–58.
- [13] A. Bultheel, M. Van Barel, Formal orthogonal polynomials and Hankel/Toeplitz duality, Numer. Algorithms 10 (1995) 289–335.
- [14] A. Bultheel, M. Van Barel, Linear Algebra, Rational Approximation and Orthogonal Polynomials, Studies in Computational Mathematics, Vol. 6, Elsevier, Amsterdam, 1997, 446 pp.
- [15] C.I. Byrnes, S.V. Gusev, A. Lindquist, A convex optimization approach to the rational covariance extension problem, SIAM J. Control Optim. 37 (1) (1999) 211–229.
- [16] T. Contantinescu, Schur Analysis, Factorization and Dilation Problems, Operator Theory: Advances and Applications, Vol. 82, Birkhäuser, Basel, 1996.
- [17] B. De Moor, Structured total least squares and L_2 approximation problems, Linear Algebra Appl. 188–189 (1993) 63–207.
- [18] B. De Moor, Total least squares for affinely structured matrices and the noisy realization problem, IEEE Trans. Signal Process. 42 (11) (1994) 3104–3113.
- [19] B. De Moor, The Riemannian singular value decomposition, in: SVD and Signal Processing, III: Algorithms, Architectures and Applications, Elsevier, Amsterdam, 1995, pp. 61–78.
- [20] B. De Moor, Structured total least squares for Hankel matrices, in: A. Paulraj, V. Roychowdhury, C. Schaper (Eds.), Communications, Computation, Control and Signal Processing, Kluwer, Dordrecht, 1997, pp. 243–258 (At the occasion of Thomas Kailath's 60th birthday, Stanford, USA).
- [21] P. Dewilde, H. Dym, Schur recursions, error formulas, and convergence of rational estimators for stationary stochastic sequences, IEEE Trans. Inform. Theory IT-27 (1981) 446–461.
- [22] P. Dewilde, H. Dym, Lossless inverse scattering, digital filters, and estimation theory, IEEE Trans. Inform. Theory IT-30 (1984) 644–662.
- [23] P. Dewilde, A.J. Van Der Veen, Time-Varying Systems and Computations, Kluwer, Dordrecht, 1998.
- [24] P. Dewilde, A. Viera, T. Kailath, On a generalized Szegő–Levinson realization algorithm for optimal linear predictors based on a network synthesis approach, IEEE Trans. Circuits Systems CAS-25 (1978) 663–675.
- [25] R.W. Freund, H. Zha, A look-ahead strategy for the solution of general Hankel systems, Numer. Math. 64 (1993) 295–322.
- [26] K. Glover, All optimal Hankel norm appoximations of linear multivariable systems and their L_{∞} error bounds, Int. J. Control 39 (6) (1984) 1115–1193.
- [27] I. Gohberg, L.A. Sakhnovich (Eds), Matrix and Operator Valued Functions, Operator Theory: Advances in Applications, Vol. 72, Birkhäuser, Basel, 1994.
- [28] W.B. Gragg, Stabilization of the UHQR-algorithm, in: Z. Chen, Y. Li, C.A. Michelli, Y. Xu (Eds.), Advances in Computational Mathematics, Lecture Notes in Pure and Applied Mathematics, Vol. 202, Marcel Dekker, New York, 1999, pp. 139–154.
- [29] W.B. Gragg, A. Lindquist, On the partial realization problem, Linear Algebra Appl. 50 (1983) 277-319.

- [30] U. Grenander, G. Szegő, Toeplitz Forms and Their Applications, University of California Press, Berkley, 1958.
- [31] A. Halanay, V. Ionescu, Time-Varying Discrete Linear Systems, Operator Theory: Advances and Applications, Vol. 68, Birkhäuser, Basel, 1994.
- [32] B. Hanzon, The area enclosed by the oriented Nyquist diagram and the Hilbert-Schmidt-Hankel norm of a linear system, IEEE Trans. Automat. Control 37 (6) (1992) 835–839.
- [33] E. Hayashi, L.N. Trefethen, M.H. Gutknecht, The CF table, Constr. Approx. 6 (1990) 195-223.
- [34] G. Heinig, K. Rost, Algebraic Methods for Toeplitz-like Matrices and Operators, Akademie, Berlin, also Birkhäuser, Basel, 1984.
- [35] B.L. Ho, R.E. Kalman, Effective construction of linear state-variable models from input/output functions, Regelungstechnik 14 (12) (1996) 545–548.
- [36] I.C.F. Ipsen, Systolic algorithms for the parallel solution of dense symmetric positive definite Toeplitz systems, in: M. Schultz (Ed.), Numerical Algorithms for Modern Parallel Computer Architectures, Springer, New York, 1988, pp. 85–108.
- [37] T. Kailath, A.H. Sayed, Displacement structure: theory and applications, SIAM Rev. 37 (1995) 297-386.
- [38] M. Kimura, Chain Scattering Approach to H-Infinity-Control, Birkhäuser, Basel, 1997.
- [39] L. Ljung, in: System Identification: Theory for the User, 2nd Edition, Information and System Sciences Series, Prentice-Hall, Englewood Cliffs, NJ, 1999.
- [40] B.C. Moore, Principal component analysis in linear systems: controllability, observability and model reduction, IEEE Trans. Automat. Control 26 (1981) 17–32.
- [41] B. Ninness, F. Gustafsson, A unified construction of orthogonal bases for system identification, IEEE Trans. Automat. Control 42 (1997) 515–522.
- [42] J. Partington, Interpolation, Identification and Sampling, London Mathematical Society Monographs, Vol. 17, Oxford University Press, Oxford, 1997.
- [43] V.P. Potapov, The Multiplicative Structure of J-Contractive Matrix Functions, American Mathematical Society Translation Series 2, Vol. 15, American Mathematics Society, Providence, RI, 1960, pp. 131–243.
- [44] G. Schelfhout, Model reduction for control design, Ph.D. Thesis, Electrotechnical Engineering ESAT, K.U. Leuven, January 1996.
- [45] L.N. Trefethen, Near circularity of the error curve in complex Chebyshev approximation, J. Approx. Theory 31 (1981) 344–367.
- [46] M. Van Barel, A. Bultheel, Orthogonal polynomial vectors and least squares approximation for a discrete inner product, Electron. Trans. Numer. Anal. 3 (1995) 1–23.
- [47] P. Van Overschee, B. De Moor, Subspace Identification for Linear Systems: Theory, Implementation, Applications, Kluwer, Dordrecht, 1996.
- [48] N. Wiener, Extrapolation, Interpolation and Smoothing of Time Series, Wiley, New York, 1949.
- [49] N. Wiener, P. Masani, The prediction theory of multivariate stochastic processes, I, The regularity condition, Acta Math. 98 (1957) 111–150.



JOURNAL OF COMPUTATIONAL AND APPLIED MATHEMATICS

Journal of Computational and Applied Mathematics 121 (2000) 379-420

www.elsevier.nl/locate/cam

Summary of Sinc numerical methods

Frank Stenger

Department of Computer Science, University of Utah, Salt Lake City, UT 84112, USA

Received 10 August 1999

Abstract

Sinc approximation methods excel for problems whose solutions may have singularities, or infinite domains, or boundary layers. This article summarizes results obtained to date, on Sinc numerical methods of computation. Sinc methods provide procedures for function approximation over bounded or unbounded regions, encompassing interpolation, approximation of derivatives, approximate definite and indefinite integration, solving initial value ordinary differential equation problems, approximation and inversion of Fourier and Laplace transforms, approximation of Hilbert transforms, and approximation of indefinite convolutions, the approximate solution of partial differential equations, and the approximate solution of integral equations, methods for constructing conformal maps, and methods for analytic continuation. Indeed, Sinc are ubiquitous for approximating every operation of calculus. © 2000 Elsevier Science B.V. All rights reserved.

Keywords: Sinc methods; Numerical methods

1. Introduction and summary

This article attempts to summarize the existing numerical methods based on *Sinc approximation*. Due to limited space, we have followed the instructions of the editors, and have attempted to restrict referencing to the original articles. Our manner of description of the methods is in symbolic form. We include methods for collocation, function interpolation and approximation, for approximation of derivatives, for approximate definite and indefinite integration, for solving initial and boundary value ordinary differential equation problems, for approximation and inversion of Fourier and Laplace transforms, for the approximation of Hilbert transforms, for the approximation of definite and indefinite solution of partial differential equations, and for the approximate solution of partial differential equations, and for methods for analytic continuation. Special effort is made to give credit to the original discoverers of the methods.

E-mail address: stenger@hsr.no (F. Stenger).

To date there are three textbooks on Sinc numerical methods; two of these [18,44] are solely related to this subject, while the third [15] contains two expository chapters, as well as theorems and proofs about approximation via Sinc methods.

Sinc methods are based on the use of the *Cardinal function*, C(f,h), which is a sinc expansion of f, defined by

$$C(f,h)(x) = \sum_{k \in \mathbb{Z}} f(kh) \operatorname{sinc}\left\{\frac{x}{h} - k\right\}, \quad x \in \mathbb{R},$$
(1.1)

and where the step size h > 0, and where the function sinc is defined by

$$\operatorname{sinc}(x) = \frac{\sin\left(\pi x\right)}{\pi x}.$$
(1.2)

This function dates back to the works of Borel [6], and Whittaker [53], although related work on the trapezoidal rule (obtained by termwise integration of C(f,h)) dates back earlier, to Plana [29]. The *Cardinal function* occupies an important place in the theory of analytic functions. My mentor, *J.J. McNamee*, rightly called this beautiful function "a function of royal blood, whose distinguished properties separate it from its bourgeois brethren". Whittaker was the first to make a connection with analytic functions, where the Sinc expansion has its natural "home". C(f,h) is replete with identities within a *Wiener class of functions*, $W(\pi/h)$ of all entire functions of order 1 and type π/h that are also square integrable over the real line \mathbb{R} – identities which yield accurate approximation formulas when applied to functions in classes other that than $W(\pi/h)$.

The term "sinc" originated in engineering literature [33]; this function was introduced to the world of communication - see [14,33] and the historical accounts in [12].

The study during the last century, of numerical methods by use of complex variables dates back to the work of Davis [9].

It is even more convenient, for purposes of Sinc computation to use the more powerful notation introduced in [19],

$$S(k,h) \circ (u) = \operatorname{sinc} \left\{ \frac{u}{h} - k \right\}$$
(1.3)

where u may be a function of x. This type of substitution, or transformation, enables use of (a finite number of terms of) C(f,h) to approximate functions over intervals other than the real line \mathbb{R} . Throughout this paper we have written Sinc rather than sinc to emphasize this feature of approximation.

Presently, there exist close to 200 publications on Sinc methods. Due to limited space, we have attempted to present only those papers in which the original works appeared. A complete listing of all of the publications can be obtained from the author.

Although studies of quadrature via the trapezoidal rule and Sinc approximation have enjoyed independent developments, they are in fact, intimately related. Indeed, the Sinc expansion can be used to derive not only the trapezoidal rule, but also the DFT (discrete Fourier transform), and many other well known expansions [43, Chapters 2, 3, 5].

It is perhaps also not well known, that whereas the Sinc methods of this paper were developed before the emergence of *wavelets*, they do, in fact satisfy all of the relationships of wavelets, and moreover, because Sinc methods are an optimal basis for approximation in spaces of functions that

are analytic on an open interval and of class Lip_{α} on the closed interval, any n-basis wavelet approximation can never be appreciably more accurate than an *n*-basis Sinc approximation. Furthermore, whereas it is frequently impossible in the absence of Sinc theory, to determine a priori the error of a wavelet approximation, Sinc methods make it possible to estimate the error of wavelet approximation.

The layout of the paper is as follows:

- 1. In the remainder of this section, Section 1, we draw some distinctions between polynomial and Sinc approximation.
- 2. In Section 2, we present explicit spaces of analytic functions for one-dimensional Sinc approximation.
- 3. In Section 3 we present the basic formulas for one-dimensional Sinc approximation.
- 4. In Section 4 we summarize applications of Sinc indefinite integration and collocation to the solution of initial and boundary value ordinary differential equation problems.
- 5. In Section 5 we summarize results obtained for solution of partial differential equations, via Sinc approximation of derivatives. Although this is an important area of applications, our coverage here is brief, since the subject matter has already been abmply covered in textbooks [18,44], and via program packages [28].
- 6. In Section 6 we summarize some results obtained on the solution of integral equations, including results for solving Cauchy singular integral equatins, boundary integral equations, and the construction of conformal maps.
- 7. In Section 7 we illustrate the use of Sinc convolution, a technique for evaluating one and multidimensional convolution-type integrals, and for the application of this procedure to the solution of differential and integral equations. This novel technique of approximation has to date been used only with Sinc approximation, where it has had incredibly surprising consequences, such as enabling a unified approach to the solution of elliptic, parabolic, and hyperbolic differential equations, and for incredible speed-up in efficiency of solving differential and integral equations.
- 8. In Section 8 we list some existing computer algorithms based on Sinc methods.

Most numerical approximation techniques are based on the use of polynomials. The main reasons for this are that polynomials readily yield easy to compute approximations to all of the operations of calculus. It is perhaps less well known that the identities of the function C(f,h) also enable approximation of all of the operations of calculus [22,41,43, Section 1.10, Chapters 3,4]. Initially, this approach yielded approximation only on the whole real line, although suitable transformations introduced in [39] enabled approximation over arbitrary intervals, or even over contours.

We end this section with a summary of the main differences between Sinc and polynomial approximation.

• Approximation formulas

Polynomial. Explicit approximation based on polynomials can be obtained either from Taylor, or else from Lagrange interpolations formulas. Both of these procedures yield splines which are very popular.

Sinc. Sinc approximation over (a,b) is based on the use of a truncated Sinc series of the form (1.1), i.e.,

$$g(x) \approx \sum_{k} g(z_k) \operatorname{sinc} \left\{ \frac{u(x)}{h} - k \right\},$$
(1.4)

where *u* is a one-to-one transformation of (a,b) onto \mathbb{R} , and with *h* as defined in (1.1), and the *Sinc points*, given by $z_k = u^{-1}(kh)$ [43, Section 4.1].

• Domain of approximation

Polynomial. Polynomial approximation formulas are applicable for approximation over *finite* intervals, or finite contours. In more than one dimension, polynomials are the basis for finite element, or finite difference approximation. The domains can have curvilinear boundaries, although they need to be finite.

Sinc. Sinc approximation formulas are applicable over finite, semi-infinite, infinite intervals, or over contours. In more than one dimension, curvilinear regions are also readily dealt with. In addition, Sinc methods readily handle infinite regions.

• Spaces of approximation and rates of convergence. This is the interesting realm of greatest difference, both in philosophy, and in rates of convergence. On the one hand, polynomials and splines seem to have 'found their home' in Sobolev spaces defined over bounded domains, whereas Sinc functions 'live more naturally' in spaces of analytic functions defined over bounded or unbounded domains. Finite element methods were developed first, and this is perhaps why they are by far the most popular, whereas Sinc methods came later, and are much less widely in use. Sinc and finite element methods both enjoy simplicity and flexibility, although this is less well known for Sinc methods. From the point of view of computation, our philosophy is that when a scientist models a problem via a differential or integral equation, he/she invariably uses calculus, and the solution to the equation is then always piecewise analytic, giving Sinc methods an advantage in the approximation of such problems. Convergence proofs of finite difference and finite element procedures are usually accomplished via the mathematics of Sobolev spaces, whereas Sinc methods, while also easy to use, require some understanding of analytic functions - an area of mathematics that is well understood by a relatively few solvers of differential equations - for proofs of convergence. Also, the derived system of algebraic equations is usually more sparse for finite difference and finite element methods than for Sinc methods. In spite of this, if accuracy is desired in the solution of a differential or integral equation, the complexity of solving a differential equation problem via a finite difference or finite element method is usually far larger than the corresponding complexity for Sinc methods. Finally, while the h-p finite element method also enabled Sinc-like exponential convergence, these methods do not converge as fast as Sinc methods (see [13]). Let us make some more specific comparisons.

Polynomial. The existence of a finite number of derivatives is assumed on the function to be approximated (call it f). For example, the assumption that $f \in \mathbb{C}^p$ enables a convergence rate of $\mathcal{O}(n^{-p})$ when approximating f with a polynomial of degree n - 1. When solving differential equations, the space of approximation is a *Sobolev space*, p is usually 2, so that it is possible to achieve a convergence rate of $\mathcal{O}(n^{-2})$ via use of either polynomials of degree n, or degree ≥ 1 splines with maximum spacing of the order of 1/n. If f is analytic in a domain containing a finite interval [a, b], then it is possible to approximate f to within an error that is $\mathcal{O}(\exp(-\gamma n))$ by a polynomial of degree n, with γ some positive constant that is independent of n. On the other hand, if f has an unknown type singularity of the form $(x - c)^{\alpha}$, with, e.g., c either a or b, and $\alpha > 0$, then this rate of convergence drops to $\mathcal{O}(n^{-\alpha})$. If the value of α is known, then this rate can be improved via the use of n - h - p type spline basis functions to $\mathcal{O}(\exp(-1.7627(\alpha n)^{1/2}))$, but this is not as fast as the rate of convergence of Sinc methods (see the above cited references).

383

Sinc. For Sinc approximation, the space of functions typically consists of functions that are analytic in a domain containing the (open) interval or contour of approximation, and of class Lip_{α} on the interval. It is possible to achieve an error if approximation that is $\mathcal{O}(\exp(-\gamma n^{1/2}))$, with γ considerably larger than the corresponding constant using splines. For example, for the case $f(x) = (x - c)^{\alpha}$, ten ordinary Sinc approximation that uses the transformation in Example 2.2 of this paper together with the bases as given in Eq. (3.2) already converges faster to f on [0, 1] than then h-p finite element approximation. Furthermore, if α and c are explicitly known then it is possible via use of a *double exponential transformation* [13] to achieve an *n*-basis Sinc approximation of this function to within a uniform error that is $\mathcal{O}(\exp(-n/\log(n)))$.

• Collocating convolutions. This novel approach [44] has to date been applied only for approximation via Sinc methods; it makes possible the accurate and efficient approximation of one and multidimensional, definite and indefinite convolution-type integrals, and this in turn yields a novel procedure for, e.g., solving elliptic, hyperbolic, and parabolic differential equations by essentially the same technique [46], as well as for solving many different types of convolution-type integral equations.

2. Sinc spaces of approximation

As mentioned above, the function C(f,h) exactly represents functions in the space, W(pi/h). Such exact representations have been studied by Rahman and Schmeisser (see, e.g., [30]), who have published many mathematically beautiful articles based on interpolation at the zeros of other functions, such as Bessel functions, etc. Such formulas undoubtedly yield novel numerical techniques yet to be derived, and are currently outside of the scope of this article.

While no longer being an exact representation, the function C(f, h) provides an incredibly accurate approximation on \mathbb{R} to functions f that are analytic and uniformly bounded on the strip

$$D_d = \{ z \in \mathbb{C} \colon |\mathscr{T}z| < d \} \quad d > 0, \tag{2.1}$$

where [22]

$$\sup_{x \in \mathbb{R}} |f(x) - \mathcal{C}(f,h)(x)| = \mathcal{O}(e^{-\pi d/h}) \quad h \to 0.$$
(2.2)

This fact guides us in getting accurate approximations to a function F over other intervals, or even contours, Γ , by selecting functions φ which provide a *one-to-one* transformations of Γ onto \mathbb{R} , and which also provide a conformal map of a region \mathcal{D} on which F is analytic and bounded onto D_d . In that way, the problem of approximation of F on Γ is transformed into the problem of approximation of f on \mathbb{R} , with $f = F \circ \varphi^{-1}$ now an analytic and bounded function in D_d . We have tacitly assumed here that \mathcal{D} is a simply connected domain in the complex plane \mathbb{C} . For practical usefulness, it is preferable to select functions φ which can be explicitly expressed, and for which it is also possible to explicitly express the inverse functions, φ^{-1} .

Let us also denote by $H^{\infty}(\mathcal{D})$ the family of all functions f that are analytic and uniformly bounded in \mathcal{D} .

For purposes of Sinc approximation, consider first the case of a finite interval, (a,b). Define φ by $w = \varphi(z) = \log[(z-a)/(b-z)]$; this function φ provides a conformal transformation of the

"eye-shaped" region $\mathscr{D} = \{z \in \mathbb{C}: |\arg[(z-a)/(b-z)| < d\}$ onto the strip D_d defined above. The same function φ also provided a one-to-one transformation of (a, b) onto the real line \mathbb{R} . The Sinc points are defined for h > 0 and $k \in \mathbb{Z} = 0, \pm 1, \pm 2, ...,$ by $z_k = \varphi^{-1}(kh) = (a + be^{kh})/(1 + e^{kh})$.

There are two important spaces of functions, $L_{\alpha,\beta}(\mathscr{D})$ and $M_{\alpha,\beta}(\mathscr{D})$ associated with Sinc approximation on the finite interval (a, b). For the case of $L_{\alpha,\beta}(\mathscr{D})$, we assume that α, β , and d are arbitrary fixed positive numbers. The space $L_{\alpha,\beta}(\mathscr{D})$ consists of the family of all functions f that are analytic and uniformly bounded in the domain \mathscr{D} defined above, such that,

$$f(z) = \begin{cases} \mathcal{O}(|z-a|^{\alpha}), \text{ uniformly as } z \to a \text{ from within } \mathcal{D}, \\ \mathcal{O}(|z-b|^{\beta}), \text{ uniformly as } z \to b \text{ from within } \mathcal{D}. \end{cases}$$
(2.3)

In order to define the second space, $M_{\alpha,\beta}(\mathcal{D})$, it is convenient to assume that α , β , and d are restricted such that $0 < \alpha \leq 1$, $0 < \beta \leq 1$, and $d \in (0, \pi)$. Then, $M_{\alpha,\beta}(\mathcal{D})$ denotes the family of all functions g that are analytic and uniformly bounded in \mathcal{D} , such that $f \in L_{\alpha,\beta}(\mathcal{D})$, where f is defined by

$$f = g - \mathscr{L}g,\tag{2.4}$$

and where

$$\mathscr{L}g(z) = \frac{(b-z)g(a) + (z-a)g(b)}{b-a}.$$
(2.5)

For example, $M_{\alpha,\beta}(\mathscr{D})$ includes all those functions $g \in Hol(\mathscr{D})$ which are of class Lip_{α} in that part of \mathscr{D} within a distance $R \leq (b-a)/2$ from a, and which are of class Lip_{β} in that part of \mathscr{D} within a distance R from b. The class $M_{\alpha,\beta}(\mathscr{D})$ thus includes functions that are analytic in \mathscr{D} , but which may have singularities at the end points of (a, b).

The $\mathcal{O}(e^{-cn^{\nabla^2}})$ convergence rate for quadratures in the Hardy space $H^p(\mathcal{U})$ with \mathcal{U} denoting the unit disc was probably due to Bojanov [5]. The general approach via the use of a mapping φ was originally carried out for quadrature, in [38,36], and later, for more other types of Sinc approximation in [39]. The spaces $L_{\alpha,\beta}(\mathcal{D})$ and $M_{\alpha,\beta}(\mathcal{D})$ were originally explicitly defined in [43], although they were in use well before then [36,38–42].

Thus, if (a, b) is a contour Γ , such as, e.g., the interval $(0, \infty)$, or the real line \mathbb{R} (or even an analytic arc in the complex plane), the mapping φ is selected to be a conformal mapping of a domain \mathcal{D} onto D_d , with D_d defined as above, such that φ is also a one-to-one map of Γ onto \mathbb{R} . The Sinc points are defined for h > 0 and $k \in \mathbb{Z}$ by $z_k = p^{-1}(kh)$. We define ρ by $\rho = e^{\varphi}$. Note that $\rho(z)$ increases from 0 to ∞ as z traverses Γ from a to b.

Let α , β and d denote arbitrary, fixed positive numbers. We denote by $L_{\alpha,\beta}(\mathcal{D})$ the family of all functions that are analytic and uniformly bounded in \mathcal{D} , such that

$$f(z) = \begin{cases} \mathcal{O}(|\rho(z)|^{\alpha}), & \text{uniformly as } z \to a \text{ from within } \bar{\mathcal{D}}, \\ \mathcal{O}(|\rho(z)|^{-\beta}), & \text{uniformly as } z \to b \text{ from within } \bar{\mathcal{D}}. \end{cases}$$
(2.6)

We next define the class of functions $M_{\alpha,\beta}(\mathcal{D})$, but this time restricting α , β and d such that $\alpha \in (0,1], \beta \in (0,1]$ and $d \in (0,\pi)$. This class consists of all those functions $g \in Hol(\mathcal{D})$, that have

finite limits at a and b, so that the function $\mathcal{L}g$ is well defined, where

$$\mathscr{L}g(z) = \frac{f(a) + \rho(z)f(b)}{1 + \rho(z)}, \quad \rho = e^{\rho},$$
(2.7)

and such that if f is defined by

$$f = g - \mathscr{L}g \tag{2.8}$$

then $f \in L_{\alpha,\beta}(\mathcal{D})$.

Note that if $0 < d < \pi$, then $\mathscr{L}(g)$ is uniformly bounded in $\overline{\mathscr{D}}$, the closure of \mathscr{D} , and moreover, $\mathscr{L}(g)(z) - f(a) = \mathscr{O}(|\rho(z)|)$ as $z \to a$, and $\mathscr{L}(g)(z) - f(b) = \mathscr{O}(1/|\rho(z)|)$ as $z \to b$, i.e., $\mathscr{L}(g) \in M_{1,1}(\mathscr{D})$. Furthermore, $M_{1,1}(\mathscr{D}) \subseteq M_{\alpha,\beta}(\mathscr{D})$ for any $\alpha \in (0,1]$, $\beta \in (0,1]$, and $d \in (0,\pi)$. The class $L_{\alpha,\beta}(\mathscr{D})$ is contained in the class $M_{\alpha,\beta}(\mathscr{D})$.

It is at times convenient to work with the somewhat simpler to describe spaces, $L_{\alpha}(\mathcal{D})$ and $M_{\alpha}(\mathcal{D})$, where

$$L_{\alpha}(\mathscr{D}) = L_{\alpha,\alpha}(\mathscr{D}), \qquad M_{\alpha}(\mathscr{D}) = M_{\alpha,\alpha}(\mathscr{D}).$$
(2.9)

These spaces are readily connected with a Hardy space of functions $H^p(\mathcal{U})$, with \mathcal{U} the unit disc, for which the best approximation rate of the form $\mathcal{O}(\exp[-cn^{1/2}])$ have been extensively studied (see [1,7,15,49,54]). The constant *c* in our own estimates of the rate of convergence $\mathcal{O}(\exp[-cn^{1/2}])$ which we obtained in this chapter is not as large as that of the optimal rate possible, as predicted in [1,54]. While excellent upper bounds on the error of Sinc approximation have been obtained, best possible bounds are still missing, and it is not known exactly how Sinc methods compare with the optimal rates of convergence established in the above cited papers.

The spaces $L_{\alpha,\beta}(\mathcal{D})$ and $M_{\alpha,\beta}(\mathcal{D})$ are motivated by the premise that most scientists and engineers use calculus to model differential and integral equation problems, and under this premise the solution to these problems are (at least piecewise) analytic. The spaces $L_{\alpha,\beta}(\mathcal{D})$ and $M_{\alpha,\beta}(\mathcal{D})$ house nearly all solutions to such problems, including solutions with singularities at end points of (finite or infinite) intervals (or at boundaries of finite or infinite domains in more than one dimension). Although these spaces also house singularities, they are not as large as Sobolev spaces which assume the existence of only a finite number of derivatives in a solution, and consequently (see below) when Sinc methods are used to approximate solutions of differential or integral equations, they are usually more efficient than finite difference or finite element methods. In addition, Sinc methods are replete with interconnecting simple identities, including the *Discrete Fourier Transform* (DFT), which is one of the sinc identities, enabling the use of the *Fast Fourier Transform* (FFT), making it possible to use a Sinc approximation for nearly every type of operation arising in the solution of differential and integral equations.

The spaces $L_{\alpha,\beta}(\mathscr{D})$ and $M_{\alpha,\beta}(\mathscr{D})$ are invariant, in the sense that if for j = 1, 2 we have conformal mappings $\varphi_j : \mathscr{D}_j \to D_d$, and if $f \in L_{\alpha,\beta}(\mathscr{D}_1)$ (resp., $f \in M_{\alpha,\beta}(\mathscr{D}_1)$), then $f \circ \varphi_1^{-1} \circ \varphi_2 \in L_{\alpha,\beta}(\mathscr{D}_2)$ (resp., $f \circ \varphi_1^{-1} \circ \varphi_2 \in M_{\alpha,\beta}(\mathscr{D}_2)$). Let us note that if the same function φ provides the conformal mappings $\varphi : \mathscr{D}' \to \mathscr{D}_{d'}, \varphi : \mathscr{D} \to D_d$, with 0 < d < d', then $\mathscr{D} \subset \mathscr{D}'$.

The results of the following theorem and its proof are originally given in [43, Section 4.1]; see also [15, pp. 119–121]. This theorem summarizes some important properties about the spaces $L_{\alpha,\beta}(\mathcal{D})$ and $M_{\alpha,\beta}(\mathcal{D})$, that are useful for proving convergence of Sinc approximations.

Theorem 2.1. Let $\alpha \in (0, 1]$, $\beta \in (0, 1]$, $d' \in (0, \pi)$, let $\mathcal{D}_{d'}$ be defined as above, let $\mathcal{D}' = \varphi^{-1}(\mathcal{D}_{d'})$ and for some fixed $d \in (0, d')$, let $\mathcal{D} = \varphi^{-1}(D_d)$. Let $f \in Hol(\mathcal{D})$, and let $\mathscr{I}f$ denote the indefinite integral of f.

- 1. If $f \in H^{\infty}(\mathscr{D}')$, then $f'/\varphi' \in H^{\infty}(\mathscr{D})$.
- 2. If $f \in H^{\infty}(\mathcal{D}')$, and if $(1/\varphi')'$ is uniformly bounded in \mathcal{D}' , then $f^{(n)}/(\varphi')^n \in H^{\infty}(\mathcal{D})$, n = 1, 2, 3, ...
- 3. If $f \in M_{\alpha,\beta}(\mathscr{D}')$, then $f'/\varphi' \in L_{\alpha,\beta}(\mathscr{D})$.
- 4. If $f \in M_{\alpha,\beta}(\mathscr{D}')$, and if $(1/\varphi')'$ is uniformly bounded in \mathscr{D}' then $f^{(n)}/(\varphi')^n \in L_{\alpha,\beta}(\mathscr{D})$, n = 1, 2, 3, ...
- 5. If $f \in H^1(\mathcal{D})$, then $\mathcal{I}f \in H^{\infty}(\mathcal{D})$.
- 6. If $f'/\varphi' \in L_{\alpha,\beta}(\mathcal{D})$, then $f \in M_{\alpha,\beta}(\mathcal{D})$.
- 7. If $f \in L_{\alpha,\beta}(\mathscr{D})$, then $\varphi' f \in H^1(\mathscr{D})$.

Let us describe some specific spaces for Sinc approximation. Of these, the transformations over (0,1) and $(0,\infty)$ were originally used by Moran [23,35], and later, by Schwartz. The log(sinh) map was originally due to Lund [17]. The remaining ones, as well as the general definition of such transformations [38,39,36] were discovered by the author.

Example 2.2. If $\Gamma = (0, 1)$, and if \mathscr{D} is the "eye-shaped" region, $\mathscr{D} = \{z \in \mathbb{C}: |\arg[z/(1-z)]| < d\}$, then $\varphi(z) = \log[z/(1-z)]$, relation (2.7) reduces to f = g - (1-x)g(0) - xg(1), and $L_{\alpha,\beta}(\mathscr{D})$ is the class of all functions $f \in Hol(\mathscr{D})$, such that for all $z \in \mathscr{D}$, $|f(z)| < c|z|^{\alpha}|1-z|^{\beta}$. In this case, if, e.g., $\delta = \max\{\alpha, \beta\}$, and a function w is such that $w \in Hol(\mathscr{D})$, and $w \in Lip_{\delta}(\mathscr{D})$, then $w \in M_{\alpha,\beta}(\mathscr{D})$. The Sinc points z_j are $z_j = e^{jh}/(1 + e^{jh})$, and $1/\varphi'(z_j) = e^{jh}/(1 + e^{jh})^2$.

Example 2.3. If Γ is the arc $\{z \in \mathbb{C} : z = e^{i\theta}, u < \theta < v\}$, where $0 < v - u < 2\pi$, and \mathscr{D} is the "eye-shaped" region $\mathscr{D} = \{z \in \mathbb{C} : |(v - u)/2 + \arg[(z - e^{iu})/(e^{iv} - z)] < d\}$, then $\varphi(z) = i(v - u)/2 + \log[(z - e^{iu})/(e^{iv} - z)]$, relation (2.7) reduces to $f = g - [(e^{iv} - z)f(e^{iu}) + (z - e^{iu})f(e^{iv})]/(e^{iv} - e^{iu})$, and $L_{\alpha,\beta}(\mathscr{D})$ is the class of all functions $f \in Hol(\mathscr{D})$ such that for all $z \in \mathscr{D}$, $|f(z)| < c|z - e^{iu}|^{\alpha}|e^{iv} - z|^{\beta}$. In this case, if, e.g., $\delta = \max\{\alpha, \beta\}$, and a function w is such that $w \in Hol(\mathscr{D})$, and $w \in Lip_{\delta}(\mathscr{D})$, then $w \in M_{\alpha,\beta}(\mathscr{D})$. The Sinc points z_j are $z_j = [e^{jh+iv} + e^{i(u+v)/2}]/[e^{jh} + e^{i(v-u)/2}]$, and $1/\varphi'(z_j) = e^{jh+i(3v-u)/2}/[e^{jh} + e^{i(v-u)/2}]^2$.

Example 2.4. For problems requiring approximation of functions with a singularity at the origin, and/or having algebraic decay at infinity. If $\Gamma = (0, \infty)$, and if \mathscr{D} is the "sector" $\mathscr{D} = \{z \in \mathbb{C} : |\arg(z)| < d\}$, then $\varphi(z) = \log(z)$, relation (2.7) reduces to $f(z) = g(z) - [g(0) + zg(\infty)]/(1+z)$, and the class $L_{\alpha,\beta}(\mathscr{D})$ is the class of all functions $f \in Hol(\mathscr{D})$ such that if $z \in \mathscr{D}$ and $|z| \leq 1$ then $|f(z)| \leq c|z|^{\alpha}$, while if $z \in \mathscr{D}$ and $|z| \geq 1$, then $|f(z)| \leq c|z|^{-\beta}$. This map thus allows for algebraic decay at both x = 0 and $x = \infty$. The Sinc points z_j are defined by $z_j = e^{jh}$, and $1/\varphi'(z_j) = e^{jh}$.

Example 2.5. For problems requiring approximation of functions with a singularity at the origin, and/or oscillatory, exponential decay at infinity. If $\Gamma = (0, \infty)$, and if \mathcal{D} is the "bullet-shaped" region $\mathcal{D} = \{z \in \mathbb{C} : |\arg(\sinh(z))| < d\}$, then $\varphi(z) = \log(\sinh(z))$. The relation (2.7) then reduces

to $f(z) = g(z) - [g(0) + \sinh(z)g(\infty)]/(1 + \sinh(z))$, and $L_{\alpha,\beta}(\mathcal{D})$ is the class of all functions $f \in Hol(\mathcal{D})$ such that if $z \in \mathcal{D}$ and $|z| \leq 1$ then $|f(z)| \leq c|z|^{\alpha}$, while if $z \in \mathcal{D}$ and $|z| \geq 1$, then $|f(z)| \leq c \exp\{-\beta |z|\}$. This map thus allows for algebraic decay at x = 0 and exponential decay at $x = \infty$. The Sinc points z_j are defined by $z_j = \log[e^{jh} + (1 + e^{2jh})^{1/2}]$, and $1/\varphi'(z_j) = (1 + e^{-2jh})^{-1/2}$.

Example 2.6. For problems requiring approximation of functions with exponential decay at at least one of the points $\pm \infty$. If $\Gamma = \mathbb{R}$, and if \mathscr{D} is the above defined "strip", $\mathscr{D} = D_d$, take $\varphi(z) = z$. The relation (2.7) then reduces to $f(z) = g(z) - [g(-\infty) + e^z g(\infty)]/(1 + e^z)$. The class $L_{\alpha,\beta}(\mathscr{D})$ is the class of all functions $f \in Hol(\mathscr{D})$ such that if $z \in \mathscr{D}$ and $\Re z \leq 0$ then $|f(z)| \leq c e^{-\alpha |z|}$, while if $z \in \mathscr{D}$ and $\Re z \geq 0$, then $|f(z)| \leq c e^{-\beta |z|}$. Thus this map allows for exponential decay at both $x = -\infty$ and $x = \infty$. The Sinc points z_i are defined by $z_i = jh$, and $1/\varphi'(z_i) = 1$.

Example 2.7. For problems requiring approximation of algebraic decaying functions at at least one of the points $\pm\infty$. If $\Gamma = \mathbb{R}$, and if \mathscr{D} is the "hour glass-shaped" region, $\mathscr{D} = \{z \in \mathbb{C}: |\arg[z + (1 + z^2)^{1/2}]| < d\}$, take $\varphi(z) = \log[z + (1 + z^2)^{1/2}]$. The relation (2.7) reduces to $f(z) = g(z) - [g(-\infty) + (z + (1 + z^2)^{1/2})g(\infty)]/[1 + z + (1 + z^2)^{1/2}]$, and the class $L_{\alpha,\beta}(\mathscr{D})$ is the class of all functions $f \in Hol(\mathscr{D})$ such that if $z \in \mathscr{D}$ and $\Re z \leq 0$, then $|f(z)| \leq c(1 + |z|)^{-\alpha}$, while if $z \in \mathscr{D}$ and $\Re z \geq 0$, then $|f(z)| \leq c(1 + |z|)^{-\alpha}$, while if $z \in \mathscr{D}$ and $\Re z \geq 0$. The Sinc points z_i are defined by $z_i = \sinh(jh)$, and $1/\varphi'(z_i) = \cosh(jh)$.

Example 2.8. For problems requiring approximation of functions on \mathbb{R} that are oscillatory and have exponential decay at $-\infty$, and that have algebraic decay at ∞ . If $\Gamma = \mathbb{R}$, and if \mathcal{D} is the "funnel-shaped" region, $\mathcal{D} = \{z \in \mathbb{C} : |\arg\{\sinh[z + (1 + z^2)^{1/2}]\}| < d\}$, take $\varphi(z) = \log\{\sinh[z + (1 + z^2)^{1/2}]\}$. The relation (2.7) then reduces to $f(z) = g(z) - [g(-\infty) + \sinh(z + (1 + z^2)^{1/2})g(\infty)]/[1 + \sinh(z + (1 + z^2)^{1/2})]$, and $L_{\alpha,\beta}(\mathcal{D})$ is the class of all functions $f \in Hol(\mathcal{D})$ such that if $z \in \mathcal{D}$ and $\Re z \leq 0$, then $|f(z)| \leq c(1 + |z|)^{-\alpha}$, while if $z \in \mathcal{D}$ and $\Re z \geq 0$, then $|f(z)| \leq ce^{-\beta|z|}$. This map thus allows for algebraic decay at $x = -\infty$ and exponential decay at $x = \infty$. The Sinc points z_j are defined by $z_j = (1/2)[t_j - 1/t_j]$, where $t_j = \log[e^{jh} + (1 + e^{2jh})^{1/2}]$, and $1/\varphi'(z_j) = (1/2)(1 + 1/t_j^2)$ $(1 + e^{-2jh})^{-1/2}$.

Example 2.9. Double exponential transformations. These were introduced originally by Takahasi and Mori [52], who advocate, when possible, to use more than one of the above transformations in succession, in order to achieve more rapid convergence – the second transformation usually being that of Example 2.7 above, although the transformation ϕ of Example 2.8 and other transformations that transform \mathbb{R} onto \mathbb{R} in a one-to-one manner will at times apply. The double exponential procedure has been extensively studied by Mori, Sugihara and others. A complete set of references may be found in [24]. We provide an explicit example in Section 3.5 below.

3. Sinc approximation

In this section we illustrate the various processes of one-dimensional Sinc approximation.

3.1. Notation

Sinc approximation in $M_{\alpha,\beta}(\mathcal{D})$ is defined as follows. Let N denote a positive integer, and let integers M, and m, a diagonal matrix D(u) and an operator V_m be defined as follows.

$$N = \text{positive integer},$$

$$M = [\beta N/\alpha],$$

$$m = M + N + 1,$$

$$D(u) = \text{diag}[u(z_{-M}), \dots, u(z_{N})],$$

$$V_{m}(u) = (u(z_{-M}), \dots, u(z_{N}))^{\mathrm{T}},$$
(3.1)

where $[\cdot]$ denotes the greatest integer function, where *u* is an arbitrary function defined on Γ , and where "T" denotes the transpose. Letting \mathbb{Z} denote the set of all integers, set [43]

$$\operatorname{sinc}(z) = \frac{\sin(\pi z)}{\pi z},$$

$$h = \left(\frac{\pi d}{\beta N}\right)^{1/2},$$

$$z_j = \varphi^{-1}(jh), \quad j \in \mathbb{Z}$$

$$\gamma_j = \operatorname{sinc}\left\{[\varphi - jh]/h\right\}, \quad j = -M, \dots, N,$$

$$\omega_j = \gamma_j, \quad j = -M + 1, \dots, N - 1,$$

$$\omega_{-M} = \frac{1}{1+\rho} - \sum_{j=-M+1}^{N} \frac{1}{1+e^{jh}} \gamma_j,$$

$$\omega_N = \frac{\rho}{1+\rho} - \sum_{j=-M}^{N-1} \frac{e^{jh}}{1+e^{jh}} \gamma_j,$$

$$\varepsilon_N = N^{1/2} e^{-(\pi d\beta N)^{1/2}},$$

$$w_m = (\omega_{-M}, \dots, \omega_N).$$
(3.2)

For given vector $\boldsymbol{c} = (c_{-M}, \dots, c_N)^{\mathrm{T}}$, set

$$\boldsymbol{w}_{m}\boldsymbol{c} = \sum_{j=-M}^{N} c_{j} \,\omega_{j}, \tag{3.3}$$

with ω_j defined as in (3.2). This operation $w_m c$ can thus be interpreted as vector dot product multiplication. We shall also define a norm by

$$||f|| = \sup_{x \in \Gamma} |f(x)|,$$

and throughout this section C will denote a generic constant, independent of N.

3.2. Sinc interpolation and approximation

A proof of the following result originally derived in [39] may be found in [41,43] (see e.g., [15, pp. 126–132].

Theorem 3.1. If $f \in M_{\alpha,\beta}(\mathcal{D})$, then

$$|f - \mathbf{w}_m \, V_m \, f|| \leqslant C \varepsilon_N. \tag{3.4}$$

The constants in the exponent in the definition of ε_N are the best constants for approximation in $M_{\alpha,\beta}(\mathscr{D})$ (see [7]). Hence accurate Sinc approximation of f is based on our being able to make good estimates on α , β , and d. If these constants cannot be accurately estimated, e.g., if instead of as in (3.2) above, we define h by $h = \gamma/N^{1/2}$, with γ a constant independent of N, then the right-hand side of (3.4) is replaced by $Ce^{-\delta N^{1/2}}$, where C and δ are some positive constants independent of N. Henceforth, we shall take h as defined in (3.2).

Remark. We remark, that if $f \in L_{\alpha,\beta}(\mathcal{D})$, then it is convenient to take $\omega_j = \operatorname{sinc}\{[\varphi - jh]/h\}$, $j = -M, \ldots, N$, instead of as defined in (3.2), since the corresponding approximation of f then also vanishes at the end points of Γ , just as f then vanishes at the end points of Γ .

3.3. Sinc approximation of derivatives

A proof of the following result is originally in [20]; it may be found on pp. 135–136 of [15]. This result forms the basis for Sinc approximation of ordinary and partial differential equation boundary (and even initial) value problems.

Theorem 3.2. Let $f \in M_{\alpha,\beta}(\mathcal{D}')$, let μ be any non-negative integer, and if $\mu > 1$, let $(1/\varphi')'$ be uniformly bounded in \mathcal{D}' . Then there exists a constant C_{μ} which is independent of N, such that

$$\left\| \left(\frac{h}{\varphi'}\right)^k \left[f - \mathbf{w}_m V_m f \right]^{(k)} \right\| \leq C_\mu \varepsilon_N, \tag{3.5}$$

for $k = 0, 1, ..., \mu$.

We remark that if $\omega_j = \text{sinc}\{[\varphi - jh]/h\}$, j = -M, ..., N, then the coefficient matrix with (i, j)th element $(h/\varphi'(z_i))^k \omega_j^{(k)}(z_i)$ (e.g., the matrix obtained when using the approximation in (3.4) to express the column vector

$$(f^{(k)}(z_{-M}),\ldots,f^{(k)}(z_N))^{\mathrm{T}}$$

in terms of the column vector

$$(f(z_{-M}),...,f(z_N))^{\mathrm{T}}$$

(with the error term on the right-hand side of (3.4) replaced by 0) is a well conditioned matrix, with condition number of the order of N^k , for the case when k is even. This result makes Sinc collocation of differential equations possible.

3.4. Sinc collocation

The following result is originally stated and proved in [43, Section 7.3]; a proof of it may also be found in [15, p. 132]. It guarantees an accurate final approximation of f on Γ , provided that we know a good approximation to f at the Sinc points, i.e., Sinc collocation is justified.

Theorem 3.3. Let $f \in M_{\alpha,\beta}(\mathcal{D})$, and let the conditions of Theorem 3.1 be satisfied. Let $c = (c_{-M}, \ldots, c_N)^T$ be a complex vector of order m, such that

$$\left(\sum_{j=-M}^{N} |f(z_j) - c_j|^2\right)^{1/2} < \delta,$$
(3.6)

where δ is a positive number. If C and ε_N are defined as in (3.4), and if ω_j is defined as in (3.2), then

$$\|f - \mathbf{w}_m \mathbf{c}\| < C\varepsilon_N + \delta. \tag{3.7}$$

3.5. Sinc quadrature

Quadratures based on transformations were originally carried out by Moran [23,35], and later by Schwartz, in [36,38,39] and in [50,51].

Theorem 3.4. If $f/\varphi' \in L_{\alpha,\beta}(\mathcal{D})$, then

$$\left|\int_{a}^{b} f(x) \,\mathrm{d}x - h\{V_{m}(1/\varphi')\}^{\mathrm{T}}(V_{m}f)\right| \leq C\varepsilon_{N}.$$
(3.8)

Example 3.5. Let us provide the following example, which illustrates both the application of this theorem, as well as the use of the *double exponential transformation*. We wish to approximate the integral

$$I=\int_{\mathbb{R}}\frac{\mathrm{d}z}{(1+z^2)^{\gamma}},$$

in which $\gamma = 0.577...$ denotes Euler's constant. We may note that the integrand is analytic and uniformly bounded in \mathcal{D}_d , for any $d \in (0, 1)$, although it does not have exponential decay at $\pm \infty$, i.e., the decay conditions of Example 2.6 are not satisfied. However, the integrand is, in fact, analytic and bounded in the region \mathcal{D} of Example 2.7 above, for any $d \in (0, 1)$, and we can therefore set $t = \varphi(z)$, with φ defined as in Example 2.7, to get

$$I = \int_{\mathbb{R}} (\cosh(t))^{1-2\gamma} \, \mathrm{d}t.$$

We may note at this point that the newly obtained integrand in this integral is analytic and bounded in D_d , $d \in (0, \pi/2)$, and moreover, it has requisite exponential decay there, with $\alpha = \beta = 2\gamma - 1$. All of the conditions of Example 2.7 as well as of Theorem 3.4 above are therefore satisfied, and we may apply the trapezoidal formula, to approximate this integral using a relatively small number of points.

We may also note, at this point, that in making the transformation φ of Example 2.7, we have tacitly assumed that the original integrand in z is analytic *only* in the region \mathscr{D} of Example 2.7, so that this newly obtained integrand in t is analytic *only* in $\mathscr{D}_{\pi/2}$, i.e., we did not take advantage of the fact that the above integrand in z is analytic on an infinitely sheeted Riemann surface (call it \mathscr{S}) with branch points at $z = \pm i$. This analyticity property is exemplified by the fact that the new integrand in t analytic not only in the region $\mathscr{D}_{\pi/2}$, but in the larger region \mathscr{D} of Example 2.7, which is mapped onto $D_{\pi/2}$ by the transformation of Example 2.7. This function φ maps the complete region of analyticity of the above integrand in z onto $\mathscr{D}_{\pi/2}$, and no improvement on the convergence rate is possible upon further exponential transformation [49].

At this point, the double exponential advocate will therefore recommend repeating the transformation of Example 2.7, to get

$$I = \int_{\mathbb{R}} \left(\cosh(\sinh(x))\right)^{1-2\gamma} \cosh(x) \, \mathrm{d}x.$$

The new integrand is now analytic and uniformly bounded in D_d , for any $d \in (0, \pi/2)$, but it now has a much faster, "double exponential" rate of decay on \mathbb{R} , yielding a more rapidly convergent approximation upon application of the trapezoidal rule. We may note that the inverse of the *composite* transformation is given explicitly by $z = \varphi^{-1}(w) = \sinh(\sinh(w))$, and while this function is a one-to-one function on \mathbb{R} , the function φ itself is a conformal map of the infinitely sheeted *Riemann surface* \mathscr{S} onto $\mathscr{D}_{\pi/2}$.

3.6. Sinc indefinite integration

The use of indefinite integration was originally stated without proof in [43]; proofs on the convergence of this formula were later given in [42, Section 3.6].

At the outset, we define numbers σ_k and e_k , by

$$\sigma_k = \int_0^k \operatorname{sinc}(x) \, \mathrm{d}x, \quad k \in \mathbb{Z},$$

$$e_k = 1/2 + \sigma_k.$$
(3.9)

We use the notation of (3.2), and we define a Toeplitz matrix $I^{(-1)}$ of order *m* by $I^{(-1)} = [e_{i-j}]$, with e_{i-j} denoting the (i, j)th element of $I^{(-1)}$. We then define operators \mathscr{J} and \mathscr{J}' , and matrices A_m and B_m by

$$(\mathscr{J}f)(x) = \int_{a}^{x} f(t) dt, \qquad (\mathscr{J}'f)(x) = \int_{x}^{b} f(t) dt,$$

$$A_{m} = hI^{(-1)}D(1/\varphi'), \qquad B_{m} = h(I^{(-1)})^{\mathrm{T}}D(1/\varphi'),$$

$$\mathscr{J}_{m} = w_{m}A_{m}V_{m} \qquad \mathscr{J}'_{m} = w_{m}B_{m}V_{m}.$$
(3.10)

with $(I^{(-1)})^{T}$ denoting the transpose of $I^{(-1)}$. (A still unsolved problem is to prove or disprove that if λ is an eigenvalue of $I^{(-1)}$ then $\Re \lambda > 0$.) We can thus state the following theorem, the result of

which enables us to collocate (linear or nonlinear, nonstiff or stiff) initial value problems over an interval or a contour.

Theorem 3.6. If
$$f/\varphi' \in L_{\alpha,\beta}(\mathcal{D})$$
, then
 $\|\mathscr{J}f - \mathscr{J}_m f\| \leq C\varepsilon_N, \quad \|\mathscr{J}'f - \mathscr{J}'_m f\| \leq C\varepsilon_N.$ (3.11)

3.7. Sinc indefinite convolution

This unusual approximation scheme was originally described in [44]. It is adaptable to any basis, although it is described here for use via the Sinc basis. It is especially useful for solving ordinary and partial differential equations that can be conveniently transformed into one or multidimensional convolution-type integrals, or equations. Such integral equation representations can be then collocated via methods based on Sinc formulas for collocating the indefinite convolution integrals

$$p(x) = \int_{a}^{x} f(x-t)g(t) dt,$$
 (3.12a)

$$q(x) = \int_{x}^{b} f(t-x)g(t) dt,$$
 (3.12b)

where $x \in \Gamma$. In presenting these convolution results, we shall assume that $\Gamma = (a, b) \subseteq \mathbb{R}$, unless otherwise indicated. Note also, that being able to collocate p and q enables us to collocate both definite convolutions

$$\int_{a}^{b} f(x-t)g(t) dt, \quad \int_{a}^{b} f(|x-t|)g(t) dt.$$
(3.13)

Other applications include the solution of convolution type integral equations which may consider to be difficult, such as Abel's integral equation, or integral equations to which the classical Wiener–Hopf method is applicable, or inversion of the Laplace transform.

Sinc collocation of p and q is possible under the following

Assumption 3.7. We assume that the "Laplace transform",

$$F(s) = \int_{E} f(t) \mathrm{e}^{-t/s} \,\mathrm{d}t \tag{3.14}$$

with $E \supseteq (0, b - a)$, exists for all $s \in \Omega^+ \equiv \{s \in \mathbb{C} : \Re s > 0\}$. Let P(r, x) be defined by

$$P(r,x) = \int_{a}^{x} f(r+x-t)g(t) dt.$$
(3.15)

We assume that

(i) $P(r, \cdot) \in M_{\alpha,\beta}(\mathscr{D}')$, uniformly for $r \in [0, b - a]$; and that

(ii) $P(\cdot,x)$ is of bounded variation on (0, [b-a]), uniformly for $x \in [a, b]$.

Under these assumptions, we get the following result (see [43, Section 4.6], or [44] for a proof), for which we may note the convenient to evaluate expressions

$$F(\mathscr{J}_m)g = \mathbf{w}_m F(A_m)V_m g, \qquad F(\mathscr{J}'_m)g = \mathbf{w}_m F(B_m)V_m g. \tag{3.16}$$

$$F(A_m) = X \operatorname{diag}[F(s_{-M}), \dots, F(s_N)]X^{-1},$$

and similarly for $F(B_m)$. Moreover, B_m need not be diagonalized, since its eigenvalues are the same as those of A_m , and the eigenvectors of B_m are easily expressed in terms of those of A_m .

Theorem 3.8. If the above assumptions are satisfied, then

$$p = F(\mathcal{J})g, \qquad q = F(\mathcal{J}')g,$$

Moreover, if A_m and B_m are defined as in (3.10), then

$$|p - F(\mathscr{J}_m)g|| \leq C\varepsilon_N, \qquad ||q - F(\mathscr{J}'_m)V_mg|| \leq C\varepsilon_N.$$

$$(3.17)$$

Remark. We remark here that it may be shown [44] that every eigenvalue of the matrices $I^{(-1)}$ lies in the closed region $\overline{\Omega^+}$, where Ω^+ denotes the right half (complex) plane. A still unsolved problem is to prove or disprove that if λ is any eigenvalue of $I^{(-1)}$ then $\Re \lambda > 0$, although this has been shown to be the case via direct computation by Naghsh–Nilchi, for m = 1, 2, ..., 513. It follows, thus, at least for the case when (a, b) is a subinterval of \mathbb{R} and 0 < m < 514, that the matrices $F(A_m)$ and $F(B_m)$ are well defined, and may be evaluated in the usual way, via diagonalization of A_m and B_m . (We have also tacitly assumed here that A_m and B_m can be diagonalized, which has not been proved or disproved to date, although this has so far always been the case for the problems that we have attempted.)

The above theorem has an important extension to application of Sinc convolution to the approximate evaluation of multidimensional convolution type integral expressions over curvilinear regions or surfaces, based on the known expression

$$\int_{0}^{\infty} \left\{ \int_{0}^{x} k(x, x - t) g(t) dt \right\} e^{-x/s} dx = \hat{K}(s) \hat{g}(s),$$
(3.18)

where \hat{g} denotes the usual "Laplace transform" of g, (i.e., with s replaced by 1/s) and

$$\hat{K}(s) = \int_0^\infty \int_0^\infty k(u, v) e^{-(u+v)/s} du dv.$$
(3.19)

It thus follows that

$$\int_0^x k(x,x-t)g(t) \,\mathrm{d}t \approx \hat{K}(\mathscr{J}_m)g. \tag{3.20}$$

3.8. Harmonic-Sinc approximation

We now give a harmonic extension of the basis defined in (3.2), enabling effective procedures for solving Poisson's equation, or for computing conformal maps of bounded or unbounded regions whose boundary is the union of a finite number of analytic arcs, or for analytic continuation.

Let Γ , \mathcal{D} , φ , ρ , N, M, m and h be defined as in (3.2), and let \mathcal{D}^+ denote the part of \mathcal{D} to the left of Γ as one traverses Γ from a to b. We take $z \in \mathcal{D}^+$, and we introduce the definitions

$$\sigma_j(z) = \mathscr{T}\left\{\frac{\mathrm{e}^{\mathrm{i}\pi[\varphi(z)-jh]/h}-1}{\pi[\varphi(z)-jh]/h}\right\}, \quad j \in \mathbb{Z},$$

$$\tau_{a}(z) = \left[1 - \frac{\mathscr{T}\varphi(z)}{\pi}\right] \Re \left\{\frac{1}{1+\rho(z)}\right\} - \frac{\Re\varphi(z)}{\pi} \mathscr{T} \left\{\frac{1}{1+\rho(z)}\right\},$$

$$\tau_{b}(z) = \left[1 - \frac{\mathscr{T}\varphi(z)}{\pi}\right] \Re \left\{\frac{\rho(z)}{1+\rho(z)}\right\} - \frac{\Re\varphi(z)}{\pi} \mathscr{T} \left\{\frac{\rho(z)}{1+\rho(z)}\right\},$$

$$\delta_{j} = \sigma_{j}, \quad -M < j < N,$$

$$\delta_{-M} = \tau_{a} - \sum_{j=-M+1}^{N} \frac{1}{1+e^{jh}} \sigma_{j},$$

$$\delta_{N} = \tau_{b} - \sum_{j=-M}^{N-1} \frac{1}{1+e^{-jh}} \sigma_{j}.$$
(3.21)

We then arrive at the following theorem, whose proof is a matter of inspection.

Theorem 3.9. Let δ_j be defined as in (3.21). Given any *m* complex numbers c_{-M}, \ldots, c_N , the expression

$$u_m = \sum_{j=-M}^{N} c_j \delta_j \tag{3.22}$$

is harmonic and uniformly bounded on \mathcal{D}^+ . Moreover, if $x \in \Gamma$,

$$\lim_{z \to x, z \in \mathscr{D}^+} u_m(z) = \sum_{j=-M}^N c_j \omega_j(x), \tag{3.23}$$

where the ω_i are defined as in (3.2).

3.9. Hilbert and Cauchy transforms

The Hilbert transform of a function f may be defined, e.g., if $f/\phi' \in L_{\alpha,\beta}(\mathcal{D})$, by the integral

$$(\mathscr{H}f)(x) = \frac{\mathrm{PV}}{\pi} \int_{\Gamma} \frac{f(t)}{t-x} \,\mathrm{d}t.$$
(3.24)

The main reason for our being able to accurately approximate Hilbert transforms is that [11] if $f \in L_{\alpha,\beta}(\mathcal{D})$, then $\mathscr{H}f \in M_{\alpha,\beta}(\mathcal{D})$.

We take $x \in \Gamma$, and $z \in \mathscr{D}^+$, the part of \mathscr{D} to the right of Γ , and define the functions

$$s_k(x) = \frac{h}{\pi} \frac{\sin\{\pi[\varphi(x) - kh]/h\}}{\varphi'(z_k)(x - z_k)}$$
(3.25)

$$t_k(x) = \frac{h}{2\pi i} \frac{\cos\left\{\frac{\pi i}{h} \left(\varphi(z) - kh\right)\right\} - 1}{\varphi'(z_k)(z - z_k)}$$
(3.26)

$$c_k(z) = \frac{h}{2\pi i} \frac{\exp\left\{\frac{\pi i}{h} \left(\varphi(z) - kh\right)\right\} - 1}{\varphi'(z_k)(z - z_k)}.$$
(3.27)

Since $c_k(x) = \frac{1}{2}(s_k(x) + t_k(x))$, we have, for k and ℓ integers, that

$$s_{k}(z_{\ell}) = \begin{cases} 0 & \text{if } k \neq \ell, \\ 1 & \text{if } k = \ell, \end{cases}$$

$$t_{k}(z_{\ell}) = \begin{cases} \frac{h}{\pi i} \frac{1 - (-1)^{k-\ell}}{\varphi'(z_{k})(z_{k} - z_{\ell})} & \text{if } k \neq \ell, \\ 0 & \text{if } k = \ell, \end{cases}$$

$$c_{k}(z_{\ell}) = \begin{cases} \frac{h}{2\pi i} \frac{1 - (-1)^{k-\ell}}{\varphi'(z_{k})(z_{k} - z_{\ell})} & \text{if } k \neq \ell \\ \frac{1}{2} & \text{if } k = \ell. \end{cases}$$
(3.28)

We then get the following result, the proof of which follows by combining results of [43, Sections 3.4 and 5.2]:

Theorem 3.10. Let α and β denote positive numbers such that $0 < \alpha \leq 1$, $0 < \beta \leq 1$, let $g \in L_{\alpha,\beta}(\mathcal{D})$, let N be a positive integer, let us select M as in (3.1), h, z_j and ε_N as in (3.2). Let ∂D^+ denote the part of \mathcal{D} to the right of Γ as we traverse Γ from a to b. Then there exist positive constants C_i , i = 1, 2, 3, that are independent of N, such that

$$\sup_{x\in\Gamma} \left| g(x) - \sum_{j=-M}^{N} g(z_j) s_j(x) \right| < C_1 \varepsilon_N, \quad x\in\Gamma,$$
(3.29)

$$\sup_{x\in\Gamma} \left| (\mathscr{H}g)(x) - \sum_{j=-M}^{N} g(z_j) t_j(x) \right| < C_2 \varepsilon_N, \quad x\in\Gamma$$
(3.30)

$$\sup_{z\in\mathscr{D}^+} \left| \frac{1}{2\pi i} \int_{\Gamma} \frac{g(t)}{t-z} dt - \sum_{j=-M}^{N} g(z_j) c_j(z) \right| < C_3 \varepsilon_N, \quad z\in\mathscr{D}^+.$$
(3.31)

4. Solution of ordinary differential equations

In this section we present the basic results obtained on the application of Sinc procedures for approximating solutions of ordinary differential equation initial and boundary value problems.

4.1. Initial value problems

We summarize here the approach in [45] to solve ordinary differential equation initial value problems.

Consider first, the simple (system of n) first-order equation(s),

$$y' = K(t)y + g(t), \quad t \in \Gamma,$$

$$y(a) = y_a.$$
 (4.1)

Assuming (7.1) to be a scalar equation, we have the following result, the proof of which may be found in [43, Section 7.1].

Theorem 4.1. Let both K/φ' and g/φ' belong to $L_{\alpha}(\mathcal{D})$. Then the initial value problem (4.1) has a unique solution $y \in M_{\alpha}(\mathcal{D})$.

Similar analyticity properties may be established for nonlinear ordinary differential equations [43, Section 7.1].

Let us consider here, the nonlinear Volterra integral equation,

$$\tau f(x) \equiv f(x) - g(x) - \int_{a}^{x} k(t, f(t)) dt = 0.$$
(4.2)

We seek to approximate f, the solution to this equation, by combining Sinc approximation and *Newton's* method, under the assumption that k and g are given. To this end, we let $X = Hol(\mathcal{D}) \cap C(\overline{\mathcal{D}})$, and we define a ball with center $f^{(0)}$ and radius r, i.e., $\mathcal{B}(f^{(0)};r)$ for $f^{(0)} \in X$ with r > 0, and with ||f|| defined as above. We assume that the partial derivatives $k_f(\cdot, f)$ and $k_{ff}(\cdot, f)$ exist, and corresponding to fixed $f^{(0)} \in X$, we assume the following:

- $g \in M_{\alpha}(\mathscr{D});$
- $k(\cdot, f)/\varphi'$, $k_f(\cdot, f)/\varphi'$, and $k_{ff}(\cdot, f)/\varphi'$ all belong to $L_{\alpha}(\mathcal{D})$ for all $f \in \mathcal{B}(f^{(0)}, r)$;
- With the operator τ defined as in (4.2) above, we may define the linear operator $\tau'_f : X \to M_{\alpha}(\mathcal{D})$, where $f \in X$ by

$$(\tau'_f)w(x) = w(x) - \int_a^x k_f(t, f(t)) w(t) dt.$$
(4.3)

The inverse $(\tau'_f)^{-1}: M_{\alpha}(\mathscr{D}) \to X$ of τ'_f is given by

$$(\tau_f')^{-1}v(x) = v(x) + \int_a^x k_f(t, f(t)) \exp\left[\int_a^t k_f(\tau, f(\tau)) \,\mathrm{d}\tau\right] v(t) \,\mathrm{d}t.$$
(4.4)

Let us now introduce the "usual" constants β , η , and K, that are required for the convergence of Newton's method, as follows:

$$\beta = \exp\left[\int_{\Gamma} |k_{f}(t, f^{(0)}(t) dt|\right];$$

$$\eta = \sup_{x \in \Gamma} |(\tau'_{f^{(0)}})^{-1} \tau(f^{(0)})|;$$

$$K = \sup_{f \in \mathscr{B}(f^{(0)};r)} \int_{\Gamma} |k_{ff}(t, f(t)) dt|,$$
(4.5)

and let us assume that

$$\gamma \equiv \beta \eta K < 1/2,$$

$$\frac{1 - \sqrt{1 - 2\gamma}}{\gamma} < r.$$
(4.6)

Then we have the following variant of the well known Newton-Kantorovich theorem.
Theorem 4.2. Starting with $f^{(0)} \in X$, the sequence $\{f^{(m)}\}$, such that

$$f^{(m+1)} = f^{(m)} - (\tau'_{f^{(m)}})^{-1} \tau(f^{(m)}), \quad m = 0, 1, \dots,$$
(4.7)

is well defined, $f^{(m)} \in M_{\alpha}(\mathcal{D}) \cap \mathcal{B}(f^{(0)}; r)$ for every positive integer m, and $f^{(m)} \to f^*$, as $m \to \infty$, where $\tau(f^*) = 0$. Moreover,

$$||f^{(m)} - f^*|| \leq (2\gamma)^{2^m} \frac{\eta}{\gamma}.$$
 (4.8)

Let us next turn to the discrete form of Newton's method, which is made possible via Sinc methods. To this end, we first carry out Sinc collocation of the expression $\tau(f)$ given in (4.2), yielding

$$\boldsymbol{t}(\boldsymbol{f}) \equiv \boldsymbol{f} - \boldsymbol{g} - \boldsymbol{h} \boldsymbol{I}^{(-1)} \boldsymbol{D}(1/\varphi') \boldsymbol{k}.$$
(4.9)

In (4.9), $f = (f_{-N}, ..., f_N)^T$, h, $I^{(-1)}$, and $D(\cdot)$ are defined as in Section 4.1, and where $k = (k(z_{-N}, f_{-N}), ..., k(z_N, f_N))^T$.

Sinc discretization provides explicit methods for solving the equation t(f) = 0, via Newton's method, starting with an initial vector $f^{(0)}$.

For the case of a system of n equations, we first write (4.2) in the form

$$F(x) - G(x) - \int_{a}^{x} K(t, F(t)) dt = \mathbf{0},$$
(4.10)

where

$$F(x) = (f_1(x), \dots, f_n(x))^{\mathrm{T}},$$

$$G(x) = (g_1(x), \dots, g_n(x))^{\mathrm{T}},$$

$$K(t, F(t)) = (k_1(t, F(t)), \dots, k_n(t, F(t)))^{\mathrm{T}}.$$

Application of Sinc indefinite integration and collocation now enables us to write this equation in the discrete, *Kronecker product* form

$$\boldsymbol{T}(\boldsymbol{F}) \equiv \boldsymbol{F} - \boldsymbol{G} - \boldsymbol{I} \otimes \{h \, \boldsymbol{I}^{(-1)} \boldsymbol{D}(1/\varphi')\} \boldsymbol{K},\$$

where I is the unit matrix of order n,

$$F = (f_{1,-N}, \dots, f_{1,N}, f_{2,-N}, \dots, f_{2,N}, \dots, f_{n,N})^{\mathrm{T}},$$

$$G = (g_{1}(z_{-N}), \dots, g_{1}(z_{N}), g_{2}(z_{-N}), \dots, g_{2}(z_{N}), \dots, g_{n}(z_{N}))^{\mathrm{T}},$$

$$K = (k_{1}(z_{-N}, F_{-N}), \dots, k_{1}(z_{N}, F_{N}), k_{2}(z_{-N}, F_{-N}), \dots, k_{2}(z_{N}, F_{N}), \dots, k_{n}(z_{N}, F_{N}))^{\mathrm{T}},$$

$$F_{j} = (f_{1,j}, \dots, f_{n,j})^{\mathrm{T}}$$

and where for a given rectangular $p \times q$ matrix $A = [a_{ij}]$ and an $r \times s$ matrix $B = [b_{ij}]$, we denote $A \otimes B$ to be the huge Kronecker product matrix $[a_{ij}B]$.

The Fréchet derivative equation corresponding to (4.10) is

$$(t'_F)W(x) \equiv W(x) - \int_a^x J(t, F(t)) W(t) dt$$
 (4.11)

where, upon denoting w_i to be the Fréchet derivative of f_i , we set $W = (w_1, \ldots, w_n)^T$, and we define J to be the $n \times n$ matrix J_{ii} , where

$$J_{ij}(x,F(x)) = \frac{\partial k_i(x,F(x))}{\partial f_j}.$$

Once again applying collocation to Eq. (4.11), we may define the Jacobian matrix for a discrete Newton method by

$$\boldsymbol{T}_{\boldsymbol{F}}' = \boldsymbol{I}_1 \otimes \boldsymbol{I}_2 - \boldsymbol{I}_1 \otimes \{h\boldsymbol{I}^{(-1)}\boldsymbol{D}(1/\varphi')\}[\boldsymbol{B}_{ij}],$$

where I_1 is the unit matrix of order n, I_2 is the unit matrix of order 2N+1, $hI^{(-1)}D(1/\varphi')$ is defined as in (3.9) above, and where for i, j = 1, ..., n, B_{ij} is the diagonal matrix of order 2N + 1,

 $\boldsymbol{B}_{ij} = \operatorname{diag}(J_{ij}(z_{-N}, F_{-N}), \dots, J_{ij}(z_N, F_N)).$

Newton's method for the vector

$$\boldsymbol{F}^{m+1} = (f_{1,-N}^{m+1}, \dots, f_{1,N}^{m+1}, f_{2,-N}, \dots, f_{2,N}^{m+1}, \dots, f_{n,N}^{m+1})^{\mathrm{T}}$$

thus takes the form

 $T'_{F^m}(F^{m+1}-F^m)=-T(F^m).$

4.2. ODE – boundary value problems

For sake of simplicity of presentation, we restrict ourselves to the model problem,

$$(Ly)(x) = y''(x) + \mu(x)y'(x) + \nu(x)y(x) - \sigma(x) = 0, \quad x \in \Gamma,$$

$$y(a) = y(b) = 0,$$
 (4.12)

There are two methods, based on Sinc approximation, to solve this problem: by a Galerkin-type scheme, and by collocation. The Galerkin scheme was developed first [40]; both procedures are described in [43]. Indeed, it is shown in [43, Chapter 7] that both of these procedures are, in effect, equivalent, in that they converge at the same rate. We only describe Sinc collocation here, since it is easier to apply, and since the following assumptions that yield convergence are also simpler for this case.

Assumption 4.3. We assume for the differential equation in (4.12) that μ/φ' , $(1/\varphi')'$, and $\nu/[\varphi']^2$ belong to $H^{\infty}(\mathscr{D})$, that $\sigma/[\varphi']^2 \in L_{\alpha}(\mathscr{D})$, and that the Problem (4.12) has a unique solution $y \in L_{\alpha}(\mathscr{D})$.

In order to arrive at a system of equations which approximates a solution to (4.12), we set

$$\delta_{i-j}^{(m)} = \left(\left(\frac{\mathrm{d}}{\mathrm{d}x} \right)^m \operatorname{sinc}(x-i) \right) \Big|_{x=j},$$
(4.13)

and define the matrix

$$I^{(m)} = [\delta_{i-j}^{(m)}], \tag{4.14}$$

$$A = I^{(2)} + hD\left(\left(\frac{1}{\varphi'}\right)' - \frac{\mu}{\varphi'}\right)I^{(1)} + h^2D\left(\frac{\nu}{[\varphi']^2}\right),$$

$$\boldsymbol{p} = h^2D\left(\frac{1}{[\varphi']^2}\right)\boldsymbol{s},$$

$$\boldsymbol{s} = (\sigma(z_{-N}), \dots, \sigma(z_N))^{\mathrm{T}}.$$
(4.15)

We then obtain the system of algebraic equations

$$A\boldsymbol{w} = \boldsymbol{p}.\tag{4.16}$$

We also mention here, that the matrix A in (4.16) is not symmetric, in the case when the differential equation (4.12) is self-adjoint, and to this end, yet another Galerkin scheme has been developed by Lund [17] which may be more suitable for solving (4.12) for self-adjoint problems, since the method of Lund yields a symmetric matrix for this case.

The solution $\mathbf{w} = (w_{-M}, \dots, w_N)^T$ of the system (4.16) yields an approximation

$$y_N(x) = \sum_{k=-N}^{N} w_k S(k,h) \circ \varphi(x)$$
(4.17)

to the solution of problem (4.12). The following result is established in [43, Section 7.2]:

Theorem 4.4. Let Assumption 4.3 be satisfied. Let y denote the exact solution of (4.12), and let y_N , defined as in (4.17), denote the approximate solution, where the vector $\mathbf{w} = (w_{-N}, \dots, w_N)^T$ denotes the exact solution of the system of equations (4.16). Then there exists a constant c which is independent of N, such that

$$\sup_{x \in \Gamma} |y(x) - y_N(x)| \le c N^{5/2} e^{-(\pi d\alpha N)^{1/2}}.$$
(4.18)

Nonlinear equations can of course also be effectively dealt with via Sinc collocation, as has been demonstrated in [19, p. 2], and discussed in [43, Section 7.2]. The method can also be used to solve eigenvalue problems [11], and even inverse problems that can be modeled via second order ODE [43, Section 7.2].

5. Partial differential equations

In this section we briefly discuss the solution of elliptic, and parabolic partial differential equations (PDE) based on Sinc-Galerkin, or Sinc collocation methods. At the outset, we briefly discuss the type of situations that we desire, in order to achieve analyticity in the solutions, enabling the high accuracy that we may expect with Sinc methods. As mentioned above, Burke [8] was the first to develop algorithms to solve PDE over rectangular regions. This undertaking was a noble one,

although too ambitious for a masters thesis, and his approach was not the best for all the model problems that he attempted. Later, a number of the faculty at Montana State University developed better algorithms for solving such problems [18]. Still later, Parker [28] developed Sinc algorithms based on Maple, for nonrectangular regions.

5.1. The regions, and the spaces of functions

The class of regions V_n which we shall consider are a union of regions of the form

$$K_{n} = \{ (\xi^{1}, \dots, \xi^{n}) \in \mathbb{R}^{n} \colon u^{1} \leqslant \xi^{1} \leqslant v^{1}, \quad u^{2}(\xi^{1}) \leqslant \xi^{2} \leqslant v^{2}(\xi^{1}), \\ \dots, u^{n}(\xi^{1}, \dots, \xi^{n-1}) \leqslant \xi^{n} \leqslant v^{n}(\xi^{1}, \dots, \xi^{n-1}) \}$$
(5.1)

where it is assumed that the u^i and v^i are either identically infinite, or else they are bounded a.e. on their domain of definition.

The Sinc approximations of the previous section readily lend themselves to approximations over rectangular regions of the form

$$V_n = \{ (x^1, \dots, x^n) \in \mathbb{R}^n : u^i \leq x^i \leq v^i, \ i = 1, \dots, n \},$$
(5.2)

with (u^i, v^i) any one of either a finite interval, $(0, \infty)$, $(-\infty, 0)$, or $(-\infty, \infty)$, or for that matter, even for the case when x^i ranges over a contour Γ_i .

We would like to consider functions $F(\xi^1, \ldots, \xi^n)$ which belong to the space $M_{\alpha}(\mathcal{D}_i)$ with respect to the variable ξ^i , when the other variables are fixed. More precisely, let α and d be positive constants, such that $0 < \alpha \leq 1$, and $0 < d < \pi$. Let φ_i be a conformal map of the region $\mathcal{D}_i \subset \mathbb{C}$ onto D_d , with D_d defined as in (2.1), let $\psi_i = \varphi_i^{-1}$ denote the inverse map, such that $\Gamma_i = [u_i, v_i] = \{\psi_i(w): w \in \mathbb{R}\}$. Set $\rho^i = \exp(\varphi^i)$, $h = [\pi d/(\alpha N)]^{1/2}$, and $z_k^i = \psi^i(kh)$. Let $(x^1, \ldots, x^{i-1}, x^{i+1}, \ldots, x^n)$ be a given point in the region $\prod_{j \neq i} \Gamma_j$, and denote by $F^i = F^i(z)$ the

Let $(x^1, \ldots, x^{i-1}, x^{i+1}, \ldots, x^n)$ be a given point in the region $\prod_{j \neq i} \Gamma_j$, and denote by $F^i = F^i(z)$ the function $F(x^1, \ldots, x^{i-1}, z, x^{i+1}, \ldots, x^n)$. Let $X_{\alpha}(V_n)$ denote the family of all functions F defined on V_n such that $F^i \in M_{\alpha}(\mathscr{D}_i)$ for $i = 1, \ldots, n$. Define basis functions ω_k^i by means of the equations

$$\omega_{k}^{i} = S(k,h) \circ \varphi^{i}, \quad k = -N + 1, \dots, N - 1,$$

$$\omega_{-N}^{i} = \frac{1}{1+\rho^{i}} - \sum_{k=-N+1}^{N} \frac{1}{1+e^{kh}} S(k,h) \circ \varphi^{i},$$

$$\omega_{N}^{i} = \frac{\rho^{i}}{1+\rho^{i}} - \sum_{k=-N}^{N-1} \frac{1}{1+e^{-kh}} S(k,h) \circ \varphi^{i}.$$
(5.3)

If $F \in X_{\alpha}(V_n)$, then it is readily shown [43, Section 6.5], that for N > 1, there exists a constant, C, independent of N, such that

$$\sup_{\substack{(x^{1},...,x^{n})\in V_{n}}} \left| F(x^{1},...,x^{n}) - \sum_{k_{1}=-N}^{N} \dots \sum_{k_{n}=-N}^{N} F(z_{k_{1}}^{1},...,z_{k_{n}}^{n}) \omega_{k_{1}}^{1}(x^{1}) \dots \omega_{k_{n}}^{n}(x^{n}) \right| \\ \leqslant CN^{1/2} \{\log(N)\}^{n-1} \mathrm{e}^{-(\pi d\alpha N)^{1/2}}.$$
(5.4)

Given a problem over the region K_n defined as in (5.1), we first transform the region K_n into the region V_n of the form (5.2). This can always be accomplished via simple "quasi-linear" transformations. For example, if u^i and v^i defined as in (5.1) are finite everywhere, we set

$$\xi^{i} = u^{i} + (v^{i} - u^{i})x^{i}$$
(5.5)

so that if, for example, this is the case for i = 1, ..., n, the new region of the variables x^i is now the cube $\prod_{i=1}^{n} [0, 1]$. If $u^i = -\infty$, $v^i = \infty$, we usually do not need to apply any transformation, unless some other advantages may be gained, depending on the problem. If u^i is a function, and $v^i = \infty$, we set

$$\xi^i = u^i + x^i. \tag{5.6}$$

It may be readily shown that if the functions u^i and v^i belong to $M_{\gamma}(\mathcal{D}^i)$ in each of the variables x^{i-k} , k = 1, ..., i-1, with γ a positive constant, then the transformed function under (5.5) or (5.6) in the variables x^i belongs to the space $X_{\alpha\gamma}(V_n)$.

We recommend decomposing the original domain (say \mathscr{B}) of the solution of the partial differential equation into a finite number of subdomains, V_n , such that the solution of the partial differential equation belongs to the space $X_{\alpha'}(V_n)$, for each V_n . We will then be able to construct a solution to within an accuracy given on the right-hand side of (5.4) (in which α is replaced by α') throughout the region \mathscr{B} . This was the approach used by Parker, in his *Ptolemy* package to solve PDE [28]. Furthermore, this process of subdivision also provides the key to parallel computation via Sinc methods.

5.2. Elliptic, parabolic and hyperbolic PDE

Let us consider, as a representative example, the Sinc approximate solution of Laplace's equation,

$$U_{\xi\xi} + U_{\eta\eta} = 0, \quad (\xi, \eta) \in K_2,$$
 (5.7)

over the region K_2 . If K_2 is a rectangular region, then we can solve the problem directly, via e.g., Sinc collocation in each variable, or equivalently, by substitution of a sum of the form in (5.4) into Eq. (5.7), and evaluating the result at all pairs of Sinc points $(z_i^{(1)}, z_j^{(2)})$ to obtain a system of algebraic equations of the form

$$AU + UB = C \tag{5.8}$$

where A, and B are square matrices (see (4.15) above for the definition of A and B), with A, B and C known, and with $U = [u_{i,j}]$, a rectangular matrices, where $u_{i,j}$ approximates the solution U at the point $(z_i^{(1)}, z_j^{(2)})$.

Suppose, now, that K_2 is given by

$$K_2 = \{ (\xi, \eta) \in \mathbb{R}^2 \colon 0 < \xi < 1, \ u(\xi) < \eta < v(\xi) \},$$
(5.9)

with the functions u and v belong to the class $M_{\alpha}(\mathcal{D})$, with \mathcal{D} defined as in Example 2.2, and (a,b) = (0,1). The substitution

$$\xi = x, \quad \eta = u + (v - u)y, \tag{5.10}$$

then transforms the differential equation (5.7) into the equation

$$U_{xx} - 2\left(\frac{u' + (v' - u')y}{v - u}\right)U_{xy} + \left(\frac{[u' + (v' - u')y]^2 + 1}{(v - u)^2}\right)U_{yy} + \left\{\frac{u' + (v' - u')y}{v - u}\left(\frac{v' - u'}{v - u}\right) - \frac{\partial}{\partial x}\left(\frac{u' + (v' - u')y}{v - u}\right)\right\}U_y = 0$$
(5.11)

which can now be solved, e.g., via Sinc collocation over $V_2 = (0, 1) \times (0, 1)$. Once an approximate solution has been obtained over V_2 , we can readily obtain an approximate solution to the original problem over K_2 by use of the transformation (5.10).

This rectangular domain approach also works for parabolic [18,16], and even Hyperbolic PDE [21]. Sinc collocation has also been used by Morley [25] to approximate a surface given its Gaussian curvature, via approximate solution of the nonlinear Monge–Ampère partial differential equations.

6. Integral equations

Integral equation approaches are becoming more popular, even for solution of ordinary and partial differential equations. Such approaches have been somewhat slow in gaining popularity, mainly because the kernels of such equations have "moving singularities" which are not easy to deal with via standard quadrature schemes. On the other hand, these approaches at times offer great savings in size of the system of algebraic equations that are required to be solved compared to sizes of systems obtained by more classical methods. In addition, such "moving singularities" can be readily dealt with via Sinc methods. In this section we briefly summarize results obtained for Volterra equations, for Cauchy singular integral equations, for conformal maps, and for solving the Lippmann–Schwinger integral equations [43, Section 6.6], since the solution of such equations can now be more efficiently obtained via Sinc convolution, a procedure which we present in Section 7.

6.1. Volterra equations

In essence, these have already been dealt with in Section 4.1 of this paper, and we shall say no more about them here.

6.2. Fredholm equations

One-dimensional linear Fredholm equations of the second kind take the form

$$f(x) - \int_{\Gamma} K(x,t) f(t) dt = g(x).$$
(6.1)

Many conditions may be cited, for accurate Sinc approximation of this equation, such as, e.g., those in the following theorem:

Theorem 6.1. Let $K(x, \cdot) \in L^1(\Gamma)$ for all $x \in \Gamma$, let $K(\cdot, t) \in M_{\alpha,\beta}(\mathcal{D})$ for all $t \in \Gamma$, and let $g \in M_{\alpha,\beta}(\mathcal{D})$. If Eq. (6.1) has a solution, f, then $f \in M_{\alpha,\beta}(\mathcal{D})$.

$$f(x) = g(x) + \int_{\Gamma} K(x,t) f(t) dt$$
(6.2)

clearly belongs to $M_{\alpha,\beta}(\mathscr{D})$. \Box

In applications of solution of integral equations it is convenient to replace the condition of Theorem 6.1, that $K(x, \cdot) \in L^1(\Gamma)$ by $K(x, \cdot)/\varphi' \in L_{\alpha,\beta}(\mathcal{D})$, in order to be able to accurately replace the integral equation (6.1) by a system of algebraic equations. In that way (see (3.1), (3.2), (3.8)) we get a *collectively compact* [3] sequence of functions which can also be used to deduce the existence of a solution to (6.1).

6.3. Cauchy singular integral equations

Cauchy singular integral equations (CSIE) over Γ are of the form

$$u\omega + v\mathcal{H}\omega + \mathbf{k}_1\omega = g_1, \tag{6.3}$$

for ω , with $\Gamma = \Gamma(a, b)$ an open arc of finite length, having endpoints at a and b, with $t \in \Gamma$, $\mathscr{H}\omega(t)$ defined in (3.24), and

$$(\mathbf{k}_{1}\omega)(t) = \int_{\Gamma} \mathbf{k}_{1}(t,\tau)\,\omega(\tau)\,\mathrm{d}\tau.$$
(6.4)

Furthermore, we assume that each of the functions $u, v, k_1(t, \cdot)$ $(t \in \Gamma)$ and $k_1(\cdot, \tau)$ $(\tau \in \Gamma)$ belong to $Hol(\mathcal{D}) \cap Lip_{\alpha'}(\bar{\mathcal{D}})$, with $\alpha' \in (0, 1]$ a positive constant. Moreover, we assume that

$$r(z) \equiv u^2(z) + v^2(z) \neq 0, \quad z \in \mathcal{D}.$$
(6.5)

We seek a solution ω to Problem (6.3), $\omega \in H^1(\mathscr{D})$.

The solution to this problem is usually carried out via *regularization* of Eq. (6.3), a procedure that enables replacement of the CSIE (6.3) by a Fredholm equation of the form

$$w - \mathbf{k}_2 w = g_2, \tag{6.6}$$

in which w and ω are simply related, and for which the conditions of the previous subsection are satisfied. The details for carrying this out were developed in [10,4], and can be found in [43, Section 6.7].

6.4. Construction of conformal maps

In [47] a Sinc procedure was described for constructing a conformal map of a region *B* whose boundary, ∂B consists of a finite number of smooth arcs Γ_j , j = 1, ..., n, onto the unit disc, under the assumption that the origin is in *B*, and that "the solution" *f* satisfies f(0) = 0. Defining \mathscr{G} by the equation

$$f(z) = z \exp\{\mathscr{G}(z)\},\tag{6.7}$$

and for $z = \zeta \in \partial B$, we set

$$\mathscr{G} = g + \mathrm{i}h,\tag{6.8}$$

with g and h real-valued functions, and with $g(\zeta) = -\log|\zeta|$. Assuming that \mathscr{G} can be expressed in the form

$$\mathscr{G}(z) = \frac{1}{\pi i} \int_{\partial B} \frac{u(t)}{t-z} dt + i\delta, \quad z \in \mathbb{C},$$
(6.9)

where u is an unknown real-valued function and δ is a real-valued constant, we can deduce the following boundary integral equation for u:

$$u(\zeta) + \Re\left(\mathscr{H}u\right)(\zeta) = g(\zeta). \tag{6.10}$$

Under suitable assumptions on the contours Γ_j that are analogous to those in Section 2 of this paper, the function g is the real part of a function belonging to a space of the form $M_{\alpha_j,\beta_j}(D_j)$, where each α_j ($\beta_j = \alpha_{j+1}$) may be simply expressed in terms of the angle of intersection of the arcs Γ_j . Moreover, (6.10) always has a unique solution u which also belongs to this space. Using standard Sinc approximation as well as the results of (3.24)-(3.27) on each Γ_j , we arrive at a system of equations for the values of u at the Sinc points of each Γ_j . Furthermore, the sequence of approximations thus obtained converges to the exact solution of (6.10). The values of u thus obtained can then be used in conjunction with (6.9) and (3.24)-(3.27) to accurately approximate $\mathscr{G}(z)$ and hence f(z) at any point $z \in B$.

6.5. Analytic continuation

The results of Theorem 3.9 may be used to carry out analytic continuation.

The results of Theorem 3.9 can also be used to solve Dirichlet's problem over a region in $B \subset \mathbb{R}^2$. Assume, for example, that *B* is a region belonging to \mathbb{R}^2 such that ∂B consists of a finite number, *n*, arcs Γ_j , and such that each arc Γ_j satisfies the conditions on the contour Γ as given at the outset of Section 2. A sum of *n* representations of the form (3.22) also satisfies Laplace's equation, and forcing such a sum to satisfy Dirichlet boundary conditions (which we allow to be discontinuous at junctions of the arcs) yields a block system of algebraic equations, for which all of the diagonal matrices are unit matrices. The solution of this system then yields a solution to the Dirichlet problem. By the maximum principle, the error of the approximate solution within *B* is bounded by the error on ∂B . This form of a boundary element method to solve partial differential equations has been successfully implemented in [27].

6.6. Boundary integral equations

Boundary integral equations provide concise representations of solutions of many problems from applications, such as three-dimensional stress-strain and potential problems, and also to the solution of such problems in two dimensions, as well as to the solution of conformal mapping problems (Section 6.4). Several papers have been written about the effectiveness of Sinc methods for the solution of boundary integral equations [43, Section 6.5], as well as for solution of two-dimensional problems already discussed in the previous subsections. Boundary integral equation methods are also becoming popular using bases other than Sinc, although Sinc methods yield algorithms of relatively small complexity for solution of such problems. In this subsection we briefly discuss the solution of such three-dimensional problems via Sinc methods.

Let us consider, for sake of illustration, the approximate solution of a class of Fredholm integral equations of the second kind having the form

$$f(\bar{r}) - \lambda \int \int_{S} K(\bar{r}, \bar{r}') f(\bar{r}') \,\mathrm{d}S(\bar{r}') = g(\bar{r}), \quad \bar{r} \in S.$$

$$(6.11)$$

Here \bar{r} and \bar{r}' are points of a two-dimensional surface S embedded in \mathbb{R}^3 , $dS(\bar{r}')$ is the element of surface area at $\bar{r}' \in S$, λ is a complex number which may or may not be a characteristic value, and g is a given function defined on S. The kernel $K(\bar{r}, \bar{r}')$ will usually have a *weak singularity* at $\bar{r}' = \bar{r}$, i.e., $K(\bar{r}, \bar{r})$ may be infinite at all $\bar{r} \in S$, although $\int \int_S |K(\bar{r}, \bar{r}')| dS(\bar{r}') < \infty$.

Problems from applications typically have the property that the surface S consists of a relatively small number of "analytic" patches on each of which the function g is also "analytic". For suitable K, it may then be shown that the solution f is also "analytic" on the interior of every such patch. We shall give a precise definition of "analyticity" in what follows.

Let us assume for now, that (6.11) is given over the square $Q = [-1,1] \times [-1,1]$. Let \mathscr{D} be defined in Eq. (6.11), with a = -1, b = 1, and $0 < d \le \pi/2$, and let $\overline{\mathscr{D}}$ denote the closure of \mathscr{D} . Let Ω be defined by

$$\Omega = \{ [-1,1] \times \bar{\mathscr{D}} \} \cup \{ \bar{\mathscr{D}} \times [-1,1] \}.$$
(6.12)

Let A denote the family of all functions $f \in C(\Omega)$ such that

• $f(\cdot, y) \in H^{\infty}(\mathcal{D})$ for all $y \in [-1, 1]$, and $f(x, \cdot) \in H^{\infty}(\mathcal{D})$ for all $x \in [-1, 1]$.

Let A_{α} denote the family of all functions $f \in C(\Omega)$, such that

• $f(\cdot, y) \in M_{\alpha}(\mathcal{D})$ for all $y \in [-1, 1]$; and $f(x, \cdot) \in M_{\alpha}(\mathcal{D})$ for all $x \in [-1, 1]$.

We may define a norm on these spaces, by

$$\|f\|_{\Omega} = \sup_{(x,y)\in\Omega} |f(x,y)|.$$
(6.13)

Corresponding to some positive integer N, let us now take $h = (\pi d/(\alpha N)^{1/2})$, let us define Sinc points z_k and basis functions ω_k as in (3.2), and corresponding to any $f \in C(\Omega)$, let us define a projection,

$$P_n(f)(x,y) = \sum_{j=-N}^N \sum_{k=-N}^N f(z_j, z_k) \,\omega_j(x) \,\omega_k(y).$$
(6.14)

The following result is then established in [43, Section 6.5]:

Theorem 6.2. If $f \in A_{\alpha}$, then there exist constants C_1 and C_2 that are independent of N, such that

$$\|f - P_N(f)\|_{\Omega} \leq (C_1 + C_2 \log(N)) N^{1/2} \exp(-(\pi d\alpha N)^{1/2}).$$
(6.15)

With reference to (6.11), let us now set

$$\kappa f(x,y) = \int \int_{S} K(\vec{r},\vec{r}') f(\vec{r}') \,\mathrm{d}S(\vec{r}'),\tag{6.16}$$

with $\bar{r} = (x, y)$, and $\bar{r}' = (\xi, \eta)$, and let us assume that K has the particular form

$$K(\bar{r},\bar{r}') = \frac{F(\bar{r},\bar{r}')}{|\bar{r}-\bar{r}'|}.$$
(6.17)

We then let A_F denote the family of all functions F, such that

- $F(\cdot, y; \xi, \eta) \in H^{\infty} \cap M_{\alpha}(\mathscr{D})$ for all fixed $(y, \xi, \eta) \in [-1, 1] \times \overline{\mathscr{D}} \times [-1, 1];$
- $F(x, \cdot; \xi, \eta) \in H^{\infty} \cap M_{\alpha}(\mathcal{D})$ for all fixed $(x, \xi, \eta) \in [-1, 1] \times [-1, 1] \times \overline{\mathcal{D}};$
- $F(x, y; \cdot, \eta) \in H^{\infty} \cap M_{\alpha}(\mathcal{D})$ for all fixed $(x, y, \eta) \in \overline{\mathcal{D}} \times [-1, 1] \times [-1, 1];$
- $F(x, y; \xi, \cdot) \in H^{\infty} \cap M_{\alpha}(\mathcal{D})$ for all fixed $(x, y, \xi) \in \overline{\mathcal{D}} \times [-1, 1] \times [-1, 1]$.

We then have the following [52, Section 6.5]:

Theorem 6.3. If $F \in A_F$, and if $f \in A$, then $\kappa f \in A_{\alpha}$.

Since $P_N f \in A$, this theorem enables us to prove that the solution to (6.11) may be obtained by Galerkin method, using $P_N f$ as an approximation, and moreover, in the case when the solution of (6.11) is either unique, or if it is not unique, but the eigenvalue λ in (6.11) is simple, then the error in the resulting Galerkin approximation may also be bounded via the right-hand side of (6.15). It has furthermore been shown [48,43, Section 6.5], that if (6.11) has a nonunique solution, for some λ , with multiplicity $p \ge 1$, then the error in the resulting approximation λ (and similarly for the error in the corresponding eigensolution) obtained, e.g., via *singular value decomposition* again has the form given in (6.15) above, but with the exponent $(\pi d\alpha N)^{1/2}$ replaced by $(\pi d\alpha N)^{1/2}/p$.

These results extend readily to the case when S is represented as a union of "analytic" patches S_1, \ldots, S_m , whose interiors are disjoint. Indeed, situations of this type generally occur in applications. Furthermore, if these surface patches can be expressed, e.g., explicitly via functions from calculus, then these functional forms will enable explicitly transformation of Q onto each of these surface patches, such that Eq. (6.11), with K as in (6.17) taken over S, will transform into a system of m equations, each over Q. By taking care to identify function values on boundaries of adjacent patches, we can again achieve a rapidly convergent scheme, with error of the form given in (6.15) [43, Section 6.5].

7. Sinc convolution

The Sinc convolution procedure involves a new process of approximation that enables surprisingly simple solutions to problems which were hitherto difficult. In this section we illustrate the solution to such problems, including the time domain conversion of control problems, the inversion of Laplace transforms, the solution of Wiener–Hopf integral equation, a unified approach to the solution of elliptic, parabolic, and hyperbolic equations, the solution of Burgers' equation, and the solution of the electric field integral equations.

7.1. Feedback control

Given functions k and g defined on $(a,b) \subseteq \mathbb{R}$, the solution we can express the time domain formulation of a feedback control problem in the form

$$f(t) - \int_a^t k(t-\tau)f(\tau)\,\mathrm{d}\tau = g(t), \quad t \in (a,b).$$
(7.1)

Letting K denote the "Laplace transform" of k, as defined in (3.14), the results of Section 3.7 immediately yield

$$f \approx w_m (I_m - K(A_m))^{-1} V_m g, \tag{7.2}$$

where I_m denotes the unit matrix of order *m*. We remark here that this approach broadens the classical solution method of feedback control, in that the equations in the classical schemes stem from ordinary differential equations with constant coefficients, since the solution method is based on finding roots of polynomials, whereas the present method not only works in such cases, but is also not restricted in this manner.

7.2. Fourier and Laplace transform inversion

The Fourier transform of the Cardinal expansion (1.1) is the DFT which is the basis of FFT. The use of this formula for the approximation and inversion of Fourier transforms is discussed at length in [43, Sections 2.2 and 3.3], and we will not repeat that discussion here. Similarly, effective methods of getting Laplace transforms based on Sinc methods are given in [2] and in [43, Section 6.9].

We illustrate here an alternate, yet unpublished method, based on Sinc convolution, for inversion of the Laplace transform, and while we have had relatively little practical experience in using this method, we know from the convergence proof in [43] that it works, and moreover, the sheer simplicity of the procedure should impress anyone who has previously attempted to invert Laplace transforms.

Assume that f satisfies the properties, f(0) = 0, so that $f(x) = \int_0^x f'(t) dt = \int_0^x f'(x-t) 1 dt$, i.e., we have a convolution of the form (3.12a), in which g(t) = 1, and assume, furthermore, that the integral defining the Laplace transform exists for all s on the open right-half plane. The (ordinary) Laplace transform equivalent of this last equation in f is $\hat{f}(s) = \{s \hat{f}(s)\}\{1/s\}$, or, to get the "Laplace transform" via replacement of s by 1/s, we get $G(s) = \hat{f}(1/s)$, and setting F(s) = G(s)/s, then in the notation of (3.12a), we get

$$f(x) = (F(\mathscr{J})1)(x). \tag{7.3}$$

Hence, using (3.10) and (3.16), we get the accurate approximation

$$f(x) \approx (\mathbf{w}_m F(A_m) \mathbf{1})(x), \tag{7.4}$$

where $\mathbf{1} = (1, 1, \dots, 1)^{\mathrm{T}}$.

7.3. Solution of Wiener–Hopf equations

Wiener-Hopf equations arising in applications are either explicitly solvable, or extremely difficult to solve, even computationally. The theoretical Wiener-Hopf procedure, while mathematically sophisticated and very beautiful, does not lend itself to a computationally efficient method of solution of this equation [37]. However, Sinc convolution offers a general, yet incredibly simple solution of Wiener–Hopf equations. These equations take the form

$$f(x) - \int_0^\infty k(x-t)f(t) \, \mathrm{d}t = g(x), \quad x \in (0,\infty),$$
(7.5)

in which we seek f, given k and g. Noting that $\int_0^\infty k(x-t) f(t) dt = (\int_0^x + \int_x^\infty) k(x-t) f(t) dt$, we can approximate each of these indefinite integrals via Sinc convolution. Upon setting

$$F_{\pm}(s) = \int_0^\infty f(\pm t) \exp\left(-\frac{t}{s}\right) \, \mathrm{d}t, \quad \Re s > 0, \tag{7.6}$$

we thus immediately get, in the notation of (3.9), the approximating system of linear equations

$$(I_m - F_+(A_m) - F_-(B_m))f = Vg, (7.7)$$

whose solution is $\mathbf{f} = (f_{-M}, \dots, f_N)^T$ where f_j approximates the solution f of (7.5) at the Sinc point z_j , i.e., by Sinc collocation, we have $f \approx w_m f$.

We remark that the same technique applies to enable the solution of convolution – type integral equations over a finite interval, and also, over \mathbb{R} .

7.4. Multidimensional Sinc convolution

As already noted above, a definite integral convolution in one dimension can be split into a sum of two indefinite ones, and each of these indefinite convolution integrals can then be approximated as illustrated in the previous subsection. Similarly a *v*-dimensional definite convolution integral can be split into 2^{v} indefinite convolution integrals, and moreover, the method of approximation extends readily from 1 to *v* dimensions. The resulting method yields an efficient "separation of variables" scheme, which readily lends itself to parallel computation.

The Sinc convolution technique enables novel and highly efficient methods of solving PDE, in the cases when the solution to such equations can be represented as integrals, or integral equations, in terms of Green's functions. The procedure has already been successfully applied to the solution of elliptic, parabolic, and hyperbolic differential equations [46], and indeed, it has been shown that:

- (i) The same procedure works for all three types of equations, elliptic, parabolic, and hyperbolic.
- (ii) The procedure is more efficient than current methods of solution of partial differential equations. (iii) While yet unpublished, it also works for curvilinear regions, essentially, by use of (3.18)–
- (iii) while yet unpublished, it also works for curvinnear regions, essentially, by use of (5.18)–(3.20).
- (iv) Although the procedure enables collocation of very large, multidimensional systems, the resulting matrices in these systems (e.g., the coefficient matrix of the vector g in (7.13) below) need never be stored, since it can be represented via Kronecker products in terms of the sum of a small number of relatively small one-dimensional matrices.

The Sinc convolution procedure has also been successfully applied to solve the (vector) electric field integral equations in three space and one time dimension, where it has been shown to be orders of magnitude faster than classical methods of solution of this equation.

It suffices to illustrate the procedure for the case of two dimensions.

We take as our "model integral", the expression

$$p(x, y) = \int_{a_2}^{y} \int_{x}^{b_1} f(x - \xi, \eta - y) g(\xi, \eta) \,\mathrm{d}\xi \,\mathrm{d}\eta, \tag{7.8}$$

where the approximation is sought over the region $\mathscr{B} = \prod_{i=1}^2 \otimes (a_i, b_i)$, and with $(a_i, b_i) \subseteq \mathbb{R}$. We assume that the mappings $\varphi_j: \mathscr{D}'_j \to \mathscr{D}_{d'}$ have been determined. We furthermore assume that positive integers N_i and M_j as well as positive numbers h_i (j = 1, 2) have been selected, we set $m_i = M_i + N_i + 1$, and we define the Sinc points by $z_{\ell}^{(j)} = \varphi_i^{-1}(\ell h_i)$, for $\ell = -M_i, \ldots, N_i$; j = 1, 2. Next, we determine matrices A_i , X_i , and S_i , such that

$$A_{1} = h_{1}(I_{m_{1}}^{(-1)})^{\mathrm{T}} D(1/\varphi_{1}') = X_{1} S_{1} X_{1}^{-1},$$

$$A_{2} = h_{2} I_{m_{2}}^{(-1)} D(1/\varphi_{2}') = X_{2} S_{2} X_{2}^{-1}.$$
(7.9)

In (7.9), $I_{m_i}^{(-1)}$ is defined as in (3.9) above, and the S_i are diagonal matrices,

$$S_j = \text{diag}[s_{-M_j}^{(j)}, \dots, s_{N_j}^{(j)}].$$
(7.10)

We require the two dimensional "Laplace transform"

$$F(s^{(1)}, s^{(2)}) = \int_0^\infty \int_0^\infty f(x, y) e^{-x/s^{(1)} - y/s^{(2)}} dx dy,$$
(7.11)

which we assume to exist for all $s^{(j)} \in \Omega^+$, with Ω^+ denoting the right-half plane. It can then be shown (see [43, Section 4.6]), that the values $p_{i,j}$ which approximate $p(z_i^{(1)}, z_i^{(2)})$ can be computed via the following succinct algorithm. In this algorithm the we use the notation, e.g., $h_{i,\cdot}$ = $(h_{i_1-M_2},\ldots,h_{i_1,N_2})^{\mathrm{T}}$. We again emphasize the obvious ease of adaptation of this algorithm to parallel computation.

Algorithm 7.1. 1. Form the arrays $z_i^{(j)}$, and $(d/dx)\varphi^{(j)}(x)$ at $x = z_i^{(j)}$ for j = 1, 2, and $i = -M_i, \dots, N_i$, and then form the block of numbers $[g_{i,j}] = [g(z_i^{(1)}, z_j^{(2)})].$

2. Determine A_i , S_j , X_j , and X_i^{-1} for j = 1, 2, as defined in (4.3).

- 3. Form $\boldsymbol{h}_{\cdot,j} = X_1^{-1} \boldsymbol{g}_{\cdot,j}, \ j = -M_2, \dots, N_2;$ 4. Form $\boldsymbol{k}_{i,\cdot} = X_2^{-1} \boldsymbol{h}_{i,\cdot}, \ i = -M_1, \dots, N_1;$
- 5. Form $\vec{r}_{i,j} = \vec{F(s_i^{(1)}, s_j^{(2)})} k_{i,j}, \ i = -M_1, \dots, N_1, \ j = -M_2, \dots, N_2;$
- 6. Form $\boldsymbol{q}_{i,\cdot} = X_2 \boldsymbol{r}_{i,\cdot}, \ i = -M_1, \dots, N_1;$
- 7. Form $p_{i,j} = X_1 q_{i,j}, j = -M_2, \dots, N_2$.

Remark. We remark here that it is unnecessary to compute the matrices X_1^{-1} and X_2^{-1} in Steps 3 and 4 of this algorithm, since the vectors $\mathbf{h}_{i,i}$ and $\mathbf{k}_{i,j}$ can be found via the LU factorization of the matrices X_1 and X_2 .

Thus starting with the rectangular array $[g_{i,j}]$, Algorithm 7.1 transforms this into the rectangular array $[p_{i,j}]$. By stacking the array of numbers $g_{i,j}$ as a vector g of length $m_1 \times m_2$, and similarly, forming a diagonal matrix **F**, in the form

$$\boldsymbol{g} = (g_{-M_1,-M_2}, g_{-M_1,-M_2+1}, \dots, g_{-M_1,N_2}, g_{-M_1+1,-M_2}, \dots, g_{N_1,N_2})^{\mathrm{T}},$$

$$\boldsymbol{F} = \mathrm{diag}[F_{-M_1,-M_2}, F_{-M_1,-M_2+1}, \dots, F_{-M_1,N_2}, F_{-M_1+1,-M_2}, \dots, F_{N_1,N_2}],$$
(7.12)

with $F_{i,j} = F(s_i^{(1)}, s_j^{(2)})$, and similarly, forming a vector **p** with the resulting array of numbers $p_{i,j}$, it is easily seen that **p** can be represented in terms of **g** via the Kronecker product matrix expression

$$\boldsymbol{p} = (X_2 \otimes X_1) \boldsymbol{F} (X_2^{-1} \otimes X_1^{-1}) \boldsymbol{g}.$$
(7.13)

The numbers $p_{i,j}$ approximate p(x, y) at $(x, y) = (z_i^{(1)}, z_j^{(2)})$; once they have been computed, we can then use the procedure of (5.3)–(5.4) to approximate p in \mathcal{B} .

To get an idea of the complexity of the above procedure, we make the simplifying assumption that $M_j = N_j = N$, for j = 1, 2. We may readily deduce that if the above two-dimensional "Laplace transform" F is either known explicitly, or if the evaluation of this transform can be reduced to the evaluation of a one-dimensional integral, then the complexity, i.e., the total amount of work required to achieve an error ε when carrying out the computations of the above algorithm (to approximate p(x, y) at $(2N + 1)^2$ points) on a sequential machine, is $\mathcal{O}([\log(\varepsilon)]^6)$.

As already mentioned, the above algorithm extends readily to v dimensions, in which case the complexity for evaluating a v-dimensional convolution integral (at $(2N + 1)^{v}$ points) to within an error of ve is of the order of $[\log(\varepsilon)]^{2v+2}$.

7.5. Sinc convolution solution of Burgers' equation

The integral equation formulation of Burgers equation takes the form

$$u(x,t) = \frac{1}{(4\pi\varepsilon t)^{1/2}} \int_{\mathbb{R}} \exp\left\{-\frac{(x-\xi)^2}{4\varepsilon t}\right\} u_0(\xi) d\xi$$
$$+\pi \int_0^t \int_{\mathbb{R}} \frac{x-\xi}{\{4\pi\varepsilon(t-\tau)\}^{3/2}} \exp\left\{-\frac{(x-\xi)^2}{4\varepsilon(t-\tau)}\right\} u^2(\xi,\tau) d\xi d\tau,$$
(7.14)

where u_0 is a given function defined on \mathbb{R} , which we define, for sake of illustration by

$$u_0(x) = a \exp\{-b(x-c)^2\}.$$
(7.15)

This choice of u_0 enables an explicit expression for the first term on the right-hand side of (7.14), which, while unnecessary for purposes of solution via Sinc convolution, nevertheless simplifies our illustration. The equivalent integral equation formulation is

$$u(x,t) = v(x,t) + \pi \int_0^t \left[\int_{-\infty}^x \frac{x-\xi}{\{4\pi\varepsilon(t-\tau)\}^{3/2}} \exp\left\{ -\frac{(x-\xi)^2}{4\varepsilon(t-\tau)} \right\} u^2(\xi,\tau) \,\mathrm{d}\xi - \int_x^\infty \frac{\xi-x}{\{4\pi\varepsilon(t-\tau)\}^{3/2}} \exp\left\{ -\frac{(x-\xi)^2}{4\varepsilon(t-\tau)} \right\} u^2(\xi,\tau) \,\mathrm{d}\xi \right] \,\mathrm{d}\tau,$$
(7.16)

where

$$v(x,t) = \frac{a}{\{1+4b\varepsilon t\}^{1/2}} \exp\left\{-\frac{b(x-c)^2}{1+4b\varepsilon t}\right\}.$$
(7.17)

We now proceed to discretize Eq. (7.14) as outlined above. To this end we may note that it is possible to explicitly evaluate the "Laplace transform" of the Green's function convolution kernel in

(7.14), i.e.,

$$F(s,\sigma) = \int_0^\infty \int_0^\infty \exp\left\{-\frac{x}{s} - \frac{t}{\sigma}\right\} \frac{x}{\{4\pi\varepsilon t\}^{3/2}} \exp\left\{-\frac{x^2}{4\varepsilon t}\right\} \,\mathrm{d}x \,\mathrm{d}t$$
$$= \frac{1}{4\varepsilon^{1/2}} \frac{s \,\sigma^{1/2}}{s + \varepsilon^{1/2} \sigma^{1/2}}.$$
(7.18)

We now select $\varepsilon = 1/2$, b = 1, c = 0, $\varphi_t(t) = \log\{\sinh(t)\}$, $\varphi_x(x) = x$, $d_t = \pi/2$, $\alpha_t = \beta_t = 1/2$, $d_x = \pi/4$, $\alpha_x = \beta_x = 1$, and in this case it is convenient to take $M_t = N_t = M_x = N_x = N$. We thus form matrices

$$A_{x} = h_{x}I^{(-1)} = X_{x}S_{x}X_{x}^{-1}, \quad A_{x}' = h_{x}(I^{(-1)})^{\mathrm{T}} = (X_{x}^{-1})^{\mathrm{T}}S_{x}X_{x}^{\mathrm{T}},$$

$$B_{t} = h_{t}I^{(-1)}D(1/\varphi_{t}') = X_{t}S_{t}X_{t}^{-1},$$
(7.19)

where the superscript "T" denotes the transpose, and where S_x and S_t are diagonal matrices, and then proceed as in the previous subsection, to reduce the integral equation problem (7.14) to the nonlinear matrix problem

$$[u_{ij}] = F(A_x, B_t, [u_{ij}^2]) - F(A'_x, B_t, [u_{ij}^2]) + [v_{ij}],$$
(7.20)

where the notation $F(\cdot, \cdot, \cdot)$ signifies the result of applying Algorithm 7.1 above. Note also, that the function v_{ij} may be evaluated a priori, via the formula $v_{ij} = v(ih_x, z_j)$, with v(x, t) defined as in (7.17), and with $z_j = \log[e^{jh_t} + (1 + e^{2jh_t})^{1/2}]$.

The system (7.20) may be solved by Neumann iteration, for *a* (defined as in (7.17)) sufficiently small. Neumann iteration takes the form

$$[u_{ij}^{(k+1)}] = F(A_x, B_t, [(u_{ij}^{(k)})^2]) - F(A'_x, B_t, [(u_{ij}^{(k)})^2]) + [v_{ij}],$$
(7.21)

for k = 0, 1, 2, ..., starting with $[u_{ij}^{(0)}] = [v_{ij}]$. For example, with a = 1/2, and using the map $\varphi_t(t) = \log[\sinh(t)]$ we achieved convergence in 4 iterations, for all values of N (between 10 and 30). We can also solve the above equation via Neumann iteration for larger values of a, if we restrict the time t to a finite interval, (0, T), via the map $\varphi_t(t) = \log\{t/(T - t)\}$.

7.6. A unified approach to the solution of PDE

The Sinc convolution technique enables novel and highly efficient methods of solving PDE, in the cases when the solution to such equations can be represented as integrals, or integral equations, in terms of Green's functions. The procedure has already been successfully applied to the solution of elliptic, parabolic, and hyperbolic differential equations [46], and indeed, it has been shown that

- (i) The same procedure works for all three types of equations;
- (ii) The procedure is more efficient than current methods of solution of partial differential equations, and
- (iii) While yet unpublished, it also works for curvilinear regions, essentially, via combination of Theorem 3.7, Eqs. (3.18)–(3.20), and Algorithm 7.1.

We illustrate here that we can obtain the "Laplace transforms" of the Green's functions for all such problems, and while we restrict ourselves to 3 space (and one time, when appropriate) dimensions, the procedure works similarly for 2 space (and one time) dimensions.

Example 7.1 (A Poisson problem). Suppose, for sake of illustration, that we wish to evaluate the following three-dimensional convolution integral expression of the solution Ψ to a Poisson problem of the form $u_{xx} + u_{yy} + u_{zz} = -g$ in $V = (a_1, b_1) \times (a_2, b_2) \times (a_3, b_3)$, given by

$$\Psi(x, y, z) = \int_{a_1}^{b_1} \int_{a_2}^{b_2} \int_{a_3}^{b_3} \frac{g(\xi, \eta, \zeta)}{4\pi\sqrt{(x-\xi)^2 + (y-\eta)^2 + (z-\zeta)^2}} \,\mathrm{d}\xi \,\mathrm{d}\eta \,\mathrm{d}\zeta. \tag{7.22}$$

for all $(x, y, z) \in V$. In order to solve this problem via Sinc convolution, we require the threedimensional "Laplace transform" $\tilde{G}(u, v, w)$ of the convolution kernel, which can, in fact, be explicitly obtained, based on the approach developed by Naghsh–Nilchi, in his thesis – see [46]. The presentations given here illustrate his procedure.

Lemma 7.2. Let $\tilde{G}(u, v, w)$ be defined by

$$\tilde{G}(u,v,w) = \int_0^\infty \int_0^\infty \int_0^\infty \frac{\exp\{-\frac{x}{u} - \frac{y}{v} - \frac{z}{w}\}}{4\pi\sqrt{x^2 + y^2 + z^2}} \,\mathrm{d}x \,\mathrm{d}y \,\mathrm{d}z.$$
(7.23)

Then

$$\tilde{G}(u,v,w) = \left(\frac{1}{u^2} + \frac{1}{v^2} + \frac{1}{w^2}\right)^{-1} \cdot \left(-\frac{1}{8} + H(u,v,w) + H(v,w,u) + H(w,u,v)\right),$$
(7.24)

where, setting

$$\lambda = \sqrt{\frac{1}{v^2} + \frac{1}{w^2}},\tag{7.25}$$

we have

$$H(u,v,w) = \frac{1}{8\pi u\lambda} \ln\left(\frac{(\lambda+\frac{1}{v})(\lambda+\frac{1}{w})}{(\lambda-\frac{1}{v})(\lambda-\frac{1}{w})}\right).$$
(7.26)

Example 7.3 (A heat problem). Consider obtaining an accurate approximation to the integral expression

$$U(\bar{r},t) \equiv \int_{0}^{t} \int \int \int_{\mathbb{R}^{3}} G(|\bar{r}-\bar{r}'|,t-t') g(\bar{r}',t') \,\mathrm{d}\bar{r}' \,\mathrm{d}t',$$
(7.27)

where we use the notation $\bar{r} = (x, y, z)$, $\bar{r}' = (x', y', z')$, where $r = |\bar{r}| = \sqrt{x^2 + y^2 + z^2}$, and where

$$G(r,t) = \frac{1}{(4\pi\varepsilon t)^{3/2}} \exp\left(-\frac{r^2}{4\varepsilon t}\right).$$
(7.28)

In this case, the Green's function (7.28) satisfies the equation

$$G_t - \varepsilon \nabla^2 G = \delta(t) \,\delta^3(\bar{r}),\tag{7.29}$$

Taking the "Laplace transform" of (7.28) with respect to t, we find, using the initial condition $G(r, 0^+) = 0$, that

$$\varepsilon \nabla^2 \tilde{G} - \frac{1}{\tau} \tilde{G} = -\delta^3(\bar{r}).$$
(7.30)

The solution to this equation is well known; it is

$$\tilde{G}(\bar{r},\tau) = \frac{1}{\varepsilon} \frac{\exp(-pr)}{4\pi r}.$$
(7.31)

where $p = 1/(\varepsilon \tau)$. Next, we want to evaluate the "Laplace transform" of \hat{G}_{xx} with respect to x as above. Since $\tilde{G}_x(0, y, z, \tau) = 0$ for $(y, z) \neq (0, 0)$, we have, upon integration by parts,

$$\int_0^\infty e^{-x/u} \tilde{G}_{xx}(x, y, z, \tau) \, \mathrm{d}x = -\frac{1}{u} \, \tilde{G}(0, y, z, \tau) + \frac{1}{u^2} \int_0^\infty e^{-x/u} \tilde{G}(x, y, z, \tau) \, \mathrm{d}x.$$
(7.32)

Hence, setting

$$H(u, v, w, \tau) \equiv \frac{1}{u} \int_0^\infty \int_0^\infty e^{-y/v - z/w} \tilde{G}(0, y, z, \tau) \, \mathrm{d}y \, \mathrm{d}z,$$
(7.33)

we find, after converting to polar coordinates, $y + jz = =\rho e^{j\theta}$, where $\rho = \sqrt{y^2 + z^2}$, $1/v + 1/w = \lambda e^{j\phi}$, with $\lambda = \sqrt{1/v^2 + 1/w^2}$, and then setting $z = e^{j\theta}$, and $\zeta = e^{j\phi}$, that

$$H(u, v, w, \tau) = \frac{j\zeta}{2\pi u \varepsilon \lambda} \int_C \frac{\mathrm{d}z}{z^2 + \frac{2\zeta z}{\varepsilon \tau \lambda} + \zeta^2}.$$
(7.34)

where

$$C = \{ z \in \mathbb{C} \colon z = e^{j\theta}, \ 0 \le \theta \le \pi/2 \}.$$
(7.35)

Upon denoting the roots of the quadratic in the denominator of the integrand by z_1 and z_2 , we find that

$$z_{1,2} = -\frac{\zeta}{\lambda} \left(\frac{1}{\varepsilon \tau} \pm \sqrt{\frac{1}{\varepsilon^2 \tau^2} - \lambda^2} \right), \tag{7.36}$$

so that

$$H(u, v, w, \tau) = -\frac{j}{4\pi u \varepsilon \sqrt{\frac{1}{\varepsilon^2 \tau^2} - \lambda^2}} \{ Q(z_1) - Q(z_2) \},$$
(7.37)

and where, with ζ denoting an arbitrary complex number, we define $Q(\zeta)$ by

$$Q(\zeta) = \int_C \frac{\mathrm{d}z}{z - \zeta}.$$
(7.38)

The numerical evaluation of $Q(\zeta)$ needs to be carried out with care. Setting $\zeta = \zeta + j\eta$, with ζ and η real, we define a linear form $L(\zeta)$ by

$$L(\zeta) = \xi + \eta - 1 \tag{7.39}$$

and we define a region A (with closure \overline{A}) in the complex plane by

$$A = \{ \zeta = \xi + j\eta \in \mathbb{C} \colon |\zeta| < 1, \ \mathscr{L}(\zeta) > 0 \}.$$

$$(7.40)$$

This area A in the complex plane plays an important role for the correct evaluation of $Q(\zeta)$ in (7.38).

As conventional, let $\ln(\zeta)$ denote the principal value of the logarithm, i.e. if ζ is a complex number, then $\ln(\zeta) = \ln|\zeta| + j \arg(\zeta)$, with $\arg(\zeta)$ taking its principal value in the range $-\pi < \arg(\zeta) \le \pi$. We then get the following result.

Lemma 7.4. Let A, ζ , $L(\zeta)$, and $Q(\zeta)$ be defined as above.

• If $\zeta \in \mathbb{C} \setminus \overline{A}$, then

$$Q(\zeta) = \ln\left(\frac{j-\zeta}{1-\zeta}\right); \tag{7.41}$$

- If $\zeta = 1$ or $\zeta = j$, then $Q(\zeta)$ is infinite. (This case cannot occur if the time interval [0,T] is chosen sufficiently small.)
- If $\zeta \in A$,

$$Q(\zeta) = \ln \left| \frac{j - \zeta}{1 - \zeta} \right| + j \left(2\pi - \left| \arg \left(\frac{j - \zeta}{1 - \zeta} \right) \right| \right);$$

$$(7.42)$$

• If $\zeta = \xi + j\eta$ with $L(\zeta) = 0$, and with $|\zeta| < 1$,

$$Q(\zeta) = \ln \left| \frac{j - \zeta}{1 - \zeta} \right| + j\pi; \tag{7.43}$$

and

• If $\zeta = \xi + j\eta$ with $L(\zeta) > 0$, and with $|\zeta| = 1$,

$$Q(\zeta) = \ln \left| \frac{j - \zeta}{1 - \zeta} \right| + i \left(\pi - \left| \arg \left(\frac{j - \zeta}{1 - \zeta} \right) \right| \right).$$
(7.44)

The four dimensional "Fourier transform" as defined in (7.28) is thus given by

$$\hat{G}(u, v, w, \tau) = \left(\frac{1}{u^2} + \frac{1}{v^2} + \frac{1}{w^2} - \frac{1}{\epsilon\tau}\right)^{-1} \times \left(-\frac{1}{8\epsilon} + H(u, v, w, \tau) + H(v, w, u, \tau) + H(w, u, v, \tau)\right)$$
(7.45)

with $H(u, v, w, \tau)$ defined as in (7.37), (7.38), and Lemma 7.2.

Example 7.5 (The wave equation). Our starting point is the equation

$$\frac{1}{c^2}\frac{\partial^2 u(\bar{r},t)}{\partial t^2} - \nabla^2 u(\bar{r},t) = g(\bar{r},t), \quad \bar{r} \in V, \ t \in (0,T)$$

$$(7.46)$$

$$u(\bar{r},0^{+}) = \frac{\partial u}{\partial t}(\bar{r},0^{+}) = 0, \qquad (7.47)$$

where $V = \prod_{i=1}^{3} (a_i, b_i)$, and with $-\infty \leq a_i < b_i \leq \infty$.

The Green's function $G(\bar{r} - \bar{r}', t - t')$ of this problem satisfies (7.46), with $g(\bar{r}, t)$ replaced by $\delta^3(\bar{r} - \bar{r}')\delta(t - t')$. The four dimensional "Laplace transform" of the function $G(\bar{r}, t)$ is defined by

$$\hat{G}(u, v, w, \tau) = \int_0^\infty \int_0^\infty \int_0^\infty \int_0^\infty G(x, y, z, t) \exp\left\{-\frac{x}{u} - \frac{y}{v} - \frac{z}{w} - \frac{t}{\tau}\right\} dx \, dy \, dz \, dt.$$
(7.48)

By first taking the "Laplace transform" of each side of (7.46) with respect to t and applying the initial conditions specified in (7.47), we find, upon denoting this transform by $\tilde{G}(\bar{r}, \tau)$, that

$$\nabla^2 \tilde{G}(\tau, \bar{r}) - \frac{1}{c^2 \tau^2} \tilde{G}(\bar{r}, \tau) = -\delta^3(\bar{r}), \tag{7.49}$$

an equation for which the solution is well known, i.e.,

$$\tilde{G}(\bar{r},\tau) = \frac{\exp\left\{-\frac{r}{c\tau}\right\}}{4\pi r}.$$
(7.50)

Moreover, by proceeding as in the previous subsection, we can now write down an explicit expression for the three-dimensional "Laplace transform" $\hat{G}(u, v, w, \tau)$ of $\tilde{G}(\bar{r}, \tau)$ taken with respect to \bar{r} , namely,

$$\hat{G}(u, v, w, \tau) = \left(\frac{1}{u^2} + \frac{1}{v^2} + \frac{1}{w^2} - \frac{1}{c^2 \tau^2}\right)^{-1} \times \left\{-\frac{1}{8} + H(u, v, w, \tau) + H(v, w, u, \tau) + H(w, u, v, \tau)\right\}$$
(7.51)

where

$$H(u, v, w, \tau) = \frac{v + wj}{4\pi u} \left\{ \frac{1}{v^2} + \frac{1}{w^2} - \frac{1}{c^2 \tau^2} \right\}^{1/2} \left\{ Q(z_1) - Q(z_2) \right\},\tag{7.52}$$

where

$$z_{1,2} = \frac{vw}{v + wj} \left\{ \pm \sqrt{\frac{1}{w^2} + \frac{1}{v^2} - \frac{1}{c^2 \tau^2}} - \frac{j}{c\tau} \right\}$$

$$\zeta = \frac{\frac{1}{v} + \frac{j}{w}}{\sqrt{\frac{1}{v^2} + \frac{1}{w^2}}},$$
(7.53)

and where $Q(\zeta)$ is defined as in (7.38) and Lemma 7.2.

We remark that the approxumate solution in all three of the above examples can be written in the form

$$\boldsymbol{u} = \boldsymbol{C}\boldsymbol{g},$$

$$\boldsymbol{C} = \sum_{\nu=1}^{8} C_{\nu},$$
(7.54)

with u and g denoting suitably defined vectors of values at Sinc point coordinates, and each of the A_v being a square matrix. For example, for the cases of the heat and wave equations above, each matrix C_v has the *Kronecker product* form

$$C_{\nu} = X_4 \otimes X_3 \otimes X_2 \otimes X_1 G X_4^{-1} \otimes X_3^{-1} \otimes X_2^{-1} \otimes X_1^{-1},$$
(7.55)

where G is a diagonal matrix, with entries $\hat{G}(s_i^{(1)}, s_j^{(2)}, s_k^{(3)}, s_\ell^{(4)})$, the $s_i^{(\sigma)}$ being eigenvalues of matrices A_{σ} defined as in (7.9), while the matrices X_{σ} are the corresponding eigenvectors of these matrices, for $\sigma = 1, 2, 3$ corresponding to space variables, and $\sigma = 4$ to time. Furthermore, the diagonal matrix G is the same for each C_{ν} , and only 4 of the matrices X_k need to be stored, since they are simply related.

Suppose, for example, that the spatial matrices A_i , i = 1, 2, 3, as well as the time interval, [0, T], and a corresponding "time matrix", A_4 , have been determined to yield a given accuracy in the solution. It may then be shown [26] that if we take $\varphi_4(t) = \log(t/(T-t))$, then each $s_{\ell}^{(4)}$ is of the form $s_{\ell}^{(4)} = T\lambda_{\ell}$, with the λ_{ℓ} independent of T, and moreover, the matrices X_4 are independent of T. Furthermore, the eigenvalues $s_{\ell}^{(4)}$ replace the variable τ in the transforms \tilde{G} of (7.45) and (7.51), whereas $s_i^{(1)}$ replaces u, $s_j^{(2)}$ replaces v, and $s_k^{(3)}$ replaces w. It readily deduced that for fixed (u, v, w), $\tilde{G}(u, v, w, \tau) = \mathcal{O}(\tau)$, as $\tau \to 0$. That is, if A is defined as in (7.54) above, then the norm of A is $\mathcal{O}(T)$, as $T \to 0$. Moreover, it is readily seen that accuracy will not diminish with decreasing T.

It thus follows, that the norm of the matrix A in (7.54) is less than 1 for all sufficiently small T.

This important feature of the method enables us to prove convergence of Neumann iteration, for T sufficiently small, of the corresponding integral equations for the cases of the heat and wave equations, when either the Green's functions (7.28) or that implicitly defined in (7.48) are used on forming kernels of integral equations.

We wish to add some additional observations. Suppose that each matrix A_i is an $m \times m$ matrix, so that each X_i in (7.55) is an $m \times m$ matrix. Then each C_v , as well as C is an $m^4 \times m^4$ matrix. For example, m = 40 should usually suffice to get 6 places of accuracy. Then C is a matrix with more than 6.5×10^{12} entries, which is too large to be stored. On the other hand it does not have to be stored; by the Sinc convolution procedure, we only need to store 4 matrices X_i , each of order m, as well as the m^4 values of the 4-dimensional array $g_{i,j,k,\ell}$. Even the diagonal matrix G need not be stored.

Imagine, now, solving the integral equation analogue of either the heat or the wave equation via a Galerkin scheme that uses m^4 basis functions of the form $\omega_i^{(1)}\omega_j^{(2)}\omega_k^{(3)}\omega_\ell^{(4)}$. To construct the matrix C will then require the evaluation of m^8 entries, with each entry requiring the approximation of a 4 dimensional integral! Sinc convolution yields this matrix for us so much more easily.

Lastly, by premultiplication of the matrix C by a Kronecker product of a row vectors of bases of the form of

 $w_m(t) \otimes w_m(z) \otimes w_m(y) \otimes w_m(x)$

in (3.2), and post-multiplying by a similar column vector, with the variables replace by primes, we get a Sinc function representation of the Green's function.

Example 7.6 (Electric field integral equation). This has been solved by Naghsh–Nilchi [26], for the (vector) electric field in three space and one time dimension, where it has been shown to be orders of magnitude faster than classical methods of solution of this equation. Table 1 compares computation times of the method with a finite difference method [55]. The first two entries in the finite difference column, and all but the last entry in the Sinc convolution column are actual times of computation. The remaining entries are estimates, based on known rates of convergence of the procedures.

It should perhaps also be mentioned here that in order to achieve an error less than 10^{-5} , the coefficient matrix of the resulting system of equations is a square matrix with over 6.5×10^{12} entries, which is too large to be stored. On the other hand, via the Sinc convolution procedure, we only need to store four square matrices of order 40. We should also mention that the resulting system

Table 1

Precision	Finite difference run-time	Sinc-convolution run-time
10^{-1}	< 1 s	< 1 s
10^{-2}	000:00:00:27	000:00:00:06
10^{-3}	003:00:41:40	000:00:02:26
10^{-4}	> 82 yr	000:00:43:12
10^{-5}	> 800000 yr	000:06:42:20
10^{-6}	> 8.2 billion yr	001:17:31:11*

The IBM RISC/560 workstation run-time for computations required by Finite difference and Sinc-convolution methods vs. desired precision

*Note: Computer run-time is shown as Days:Hours:Minutes:Seconds.

of equations was solved by successive approximation, a procedure that always converges if the time interval is taken sufficiently small.

8. Computer packages

A number of computer program packages have already been developed, including:

- 1. a package for one-dimensional quadrature [32];
- 2. a mouse-automated package that runs on Unix operated machines, for all of the one-dimensional Sinc operations, including interpolation, quadrature, indefinite integral equations, Hilbert transforms, and indefinite convolution [56];
- 3. a Fortran package for solving ODE initial value problems [45];
- 4. a Least squares package, in "C" for Sinc approximation [34];
- 5. a "Sinc-Maple package for solving PDE [28]; and
- 6. a Fortran package for conformal mapping [31].

We are currently in the process of developing a Sinc computer program package for solution of stress–strain problems based on solving boundary integral equations via Sinc approximation. Our plans are to combine these packages.

Acknowledgements

Part of this research was carried out while the author was visiting the University of Aizu, during the summer of 1999. The hospitality and support of the University of Aizu is gratefully acknowledged. I have written this paper in fond memory of J.J. McNamee (deceased), E.L. Whitney (deceased), and Philip Davis, 3 people, whose contributions were necessary to make this work possible.

References

- [1] J.-E. Andersson, B.D. Bojanov, A note on the optimal quadrature in H^p , Numer. Math. 44 (1984) 301–308.
- [2] D.D. Ang, J.R. Lund, F. Stenger, Complex variable and regularization methods of inversion of the Laplace transform, Math. Comp. 51 (1989) 589–608.
- [3] P. Anselone, Collectively Compact Operator Approximation Theory and Applications to Integral Equations, Prentice-Hall, Englewood Cliffs, NJ, 1971.
- [4] B. Bialecki, F. Stenger, Sinc-Nyström method for numerical solution of one-dimensional Cauchy singular integral equations given on a smooth arc in the complex plane, Math. Comp. 51 (1988) 133–165.
- [5] B. Bojanov, Best Quadrature Formula for a Certain Class of Analytic Functions, Zastosowania Matematyki Appl. Mat. XIV (1974) 441–447.
- [6] E. Borel, Sur l'interpolation, C.R. Acad. Sci. Paris 124 (1897) 673-676.
- [7] H.G. Burchard, K. Höllig, N-width and entropy of H^p classes in $L^q(-1,1)$, SIAM J. Math. Anal. 16 (1985) 405-421.
- [8] R. Burke, Application of the Sinc-Galerkin method to the solution of boundary value problems, M. Sc. Thesis, University of Utah, 1977.
- [9] P.J. Davis, Errors of numerical approximation for analytic functions, J. Rational Mech. Anal. 2 (1953) 303-313.
- [10] D. Elliott, F. Stenger, Sinc method of solution of singular integral equations, IMACS Conference on CSIE, Philadelphia, PA, 1984, pp. 155–166.
- [11] M. Jarratt, J. Lund, K.L. Bowers, Galerkin schemes and the Sinc-Galerkin method for singular Sturm-Liouville problems, J. Comput. Phys. 89 (1990) 41–62.
- [12] A. Jerri, The Shannon sampling theorem its various extensions and applications: a tutorial review, Proc. IEEE 65 (1977) 1565–1596.
- [13] F. Keinert, Uniform approximation to $|x|^{\beta}$ by Sinc functions, J. Approx. Theory 66 (1991) 44–52.
- [14] V.A. Kotel'nikov, On the carrying capacity of the 'Ether' and wire in telecommunications, Material for the First All-Union Conference on Questions of Communication, Izd. Red. Upr. Svyazi RKKA, Moscow, 1933.
- [15] M. Kowalski, K. Sikorski, F. Stenger, Selected Topics in Approximation and Computation, Oxford University Press, Oxford, 1993.
- [16] A. Lippke, Analytic solution and Sinc function approximation in thermal conduction with nonlinear heat generation, J. Heat Transfer (Trans. ASME) 113 (1991) 5–11.
- [17] J. Lund, Symmetrization of the Sinc-Galerkin method for boundary value problems, Math. Comput. 47 (1986) 571–588.
- [18] J. Lund, K.L. Bowers, Sinc Methods for Quadrature and Differential Equations, SIAM, Philadelphia, PA, 1992.
- [19] L. Lundin, A Cardinal Function Method of Solution of the Equation $\Delta u = u u^3$, Math. Comput. 35 (1980) 747–756.
- [20] L. Lundin, F. Stenger, Cardinal-type approximation of a function and its derivatives, SIAM J. Numer. Anal. 10 (1979) 139–160.
- [21] K. McArthur, K. Bowers, J. Lund, Numerical implementation of the Sinc-Galerkin method for second-order hyperbolic equations, Numer. Methods Partial Differential Equations 3 (1987) 169–185.
- [22] J. McNamee, F. Stenger, E.L. Whitney, Whittaker's Cardinal function in retrospect, Math. Comp. 25 (1971) 141–154.
- [23] P.A.P. Moran, Numerical integration in the presence of singularities, Acta Math. Sci. 1 (1981) 83-85.
- [24] M. Mori, M. Sugihara, The double exponential transformation in numerical analysis, in: Numerical Analysis in the 20th Century Vol. V: Quadrature and Orthogonal Polynomials, J. Comput. Appl. Math. 127 (2001), to appear.
- [25] D. Morley, On the convergence of a Newton-like method for closed surfaces, Ph.D. Thesis, University of Utah, 1979.
- [26] A. Naghsh-Nilchi, F. Stenger, Sinc-convolution method of solving the electric field integral equations over $\mathbb{R}^2 \times \mathbb{R}^+ \times [0, T]$, submitted for publication.
- [27] S. Narashimhan, K. Chen, F. Stenger, The Harmonic-Sinc solution of the Laplace equation for problems with singularities and semi-infinite domains, Numer. Heat Transfer, to appear.
- [28] K. Parker, PTOLEMY, A Sinc collocation sub-system, Ph.D. Thesis, University of Utah, 1999.
- [29] G. Plana, Sur une nouvelle expression analytique des nombres Bernoulliens, Academia di Torino 25 (1820) 403–418.
- [30] Q.I. Rahman, G. Schmeisser, The summation formulæ of Poisson, Plana, Euler-Maclaurin and their relationship, J. Math. Sci. 28 (1994) 151–171.

- [31] R. Schmidtlein, F. Stenger, A Sinc-Fortran package for conformal mapping, to appear.
- [32] J. Schwing, K. Sikorski, F. Stenger, ALGORITHM 614, A FORTRAN subroutine for numerical integration in H^p , ACM TOMS 10 (1984) 152–160.
- [33] C.E. Shannon, A mathematical theory of communication, Bell Systems Tech. J. 27 (1948) 379-423,623-656.
- [34] J. Soller, Global optimization for neural network problems, Ph.D. Thesis, University of Utah, 1994.
- [35] W. Squire, Numerical evaluation of integrals using Moran transformations, Aerospace Eng. Report No. Tr-14, 1969.
- [36] F. Stenger, An analytic function which is an approximate characteristic function, SIAM J. Numer. Anal. 12 (1975) 239–254.
- [37] F. Stenger, The approximate solution of convolution-type integral equations, SIAM J. Math. Anal. 4 (1973) 536–555.
- [38] F. Stenger, Integration formulas based on the trapezoidal formula, J. Inst. Math. Appl. 12 (1973) 103-114.
- [39] F. Stenger, Approximations via the Whittaker Cardinal Function, J. Approx. Theory 17 (1976) 222-240.
- [40] F. Stenger, A Sinc-Galerkin method of solution of boundary value problems, Math. Comp. 33 (1979) 85-109.
- [41] F. Stenger, Numerical methods based on the Whittaker Cardinal, or Sinc functions, SIAM Rev. 23 (1981) 165-224.
- [42] F. Stenger, Sinc methods for solving partial differential equations, in: A. Miller (Ed.), Contributions of Mathematical Analysis to the Approximate Solution of Partial Differential Equations, Proceedings of the Centre for Mathematical Analysis, Australian National University, Vol. 7, 1984, pp. 40–64.
- [43] F. Stenger, Numerical Methods Based on Sinc and Analytic Functions, Springer, NewYork, 1993.
- [44] F. Stenger, Collocating convolutions, Math. Comp. 64 (1995) 211-235.
- [45] F. Stenger, S.A. Gustafson, B. Keyes, M. O'Reilly, K. Parker, ODE-IVP-PACK, via Sinc indefinite integration and Newton iteration, Numer. Algorithms, to appear.
- [46] F. Stenger, A. Naghsh-Nilchi, J. Niebsch, R. Ramlau, A unified approach to the approximate solution of PDE, Comm. Appl. Anal., to appear.
- [47] F. Stenger, R. Schmidtlein, Conformal maps via Sinc methods, in: N. Papamichael, St. Ruscheweyh, E.B. Saff (Eds.), Computational Methods in Function Theory (CMFT '97), World Scientific, Singapore, pp. 505–549.
- [48] M. Stromberg, Solution of shock problems by methods using Sinc functions, Ph.D. Thesis, University of Utah, 1988.
- [49] 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).
- [50] H. Takahasi, M. Mori, Error estimation in the numerical integration of analytic functions, Report Comput. Centre, Univ. Tokyo 3 (1970) 41–108.
- [51] H. Takahasi, M. Mori, Quadrature formulas obtained by variable transformation, Numer. Math. 21 (1973) 206–219.
- [52] H. Takahasi, M. Mori, Double exponential formulas for numerical integration, Publ. RIMS Kyoto Univ. 9 (1974) 721–741.
- [53] E.T. Whittaker, On the functions which are represented by expansion of the interpolation theory, Proc. Roy. Soc. Edinburgh 35 (1915) 181–194.
- [54] K. Wilderotter, *n*-Widths of H^p -spaces in $L_q(-1,1)$, J. Complexity 8 (1992) 324–335.
- [55] K. Yee, Numerical solution of boundary value problems involving Maxwell's equations in isotropic media, IEEE Trans. Antennas and Propagation AP-16 (1966) 302–307.
- [56] T. Zhang, F. Stenger, SINC PACK A Package of Sinc Algorithms, in preparation.

Further reading

- [1] K. Al-Khaled, Theory and computation in Hyperbolic model problems, Ph.D. Thesis, University of Nebraska, 1996.
- [2] W. Faltenbacher, Sinc method of solution of the Helmholtz Equation on a Half Space, University of Utah, Utah, 1984.
- [3] Y. Ikebe, T.Y. Li, F. Stenger, Numerical solution of Hilbert's problem, in: A.G. Law, B. Sahney (Eds.), Theory of Approximation with Applications, Academic Press, NewYork, 1976 pp. 338–358.
- [4] S.A. Johnson, Y. Zhou, M.L. Tracy, M.J. Berggren, F. Stenger, Inverse scattering solutions by Sinc basis, multiple source moment method, – Part III: fast algorithms, Ultrasonic Imaging 6 (1984) 103–116.
- [5] D.L. Lewis, A fully Galerkin method for parabolic problems, Ph.D. Thesis, Montana State University, 1989.
- [6] J. Mueller, Inverse problems for singular differential equations, Ph.D. Thesis, University of Nebraska, 1997.

- [7] M. O'Reilly, F. Stenger, Computing solutions to medical problems, IEEE Trans. Automat. Control 43 (1998) 843-846.
- [8] K. Pfabe, A problem in nonlinear ion transport, Ph.D. Thesis, University of Nebraska, 1995.
- [9] J. Schwing, Numerical solution of integral equations in potential theory problems, Ph.D. Thesis, University of Utah, 1976.
- [10] F. Stenger, B. Barkey, R. Vakili, Sinc convolution method of solution of Burgers' equation, in: K. Bowers, J. Lund (Eds.), Proceedings of Computation and Control III, Birkhäuser, Basel, 1993.



Journal of Computational and Applied Mathematics 121 (2000) 421-464

JOURNAL OF COMPUTATIONAL AND APPLIED MATHEMATICS

www.elsevier.nl/locate/cam

Interval analysis: theory and applications

Götz Alefeld^{a, *}, Günter Mayer^b

^aInstitut für Angewandte Mathematik, Universität Karlsruhe, D-76128 Karlsruhe, Germany ^bFachbereich Mathematik, Universität Rostock, D-18051 Rostock, Germany

Received 13 August 1999

Abstract

We give an overview on applications of interval arithmetic. Among others we discuss verification methods for linear systems of equations, nonlinear systems, the algebraic eigenvalue problem, initial value problems for ODEs and boundary value problems for elliptic PDEs of second order. We also consider the item software in this field and give some historical remarks. © 2000 Elsevier Science B.V. All rights reserved.

Contents

- 1. Historical remarks and introduction
- 2. Definitions, notations and basic facts
- 3. Computing the range of real functions by interval arithmetic tools
- 4. Systems of nonlinear equations
- 5. Systems of linear equations
- 6. The algebraic eigenvalue problem and related topics
- 7. Ordinary differential equations
- 8. Partial differential equations
- 9. Software for interval arithmetic

1. Historical remarks and introduction

First, we try to give a survey on how and where interval analysis was developed. Of course, we cannot give a report which covers all single steps of this development. We simply try to list some

* Corresponding author.

E-mail addresses: goetz.alefeld@math.uni-karlsruhe.de (G. Alefeld), guenter.mayer@mathematik.uni-rostock.de (G. Mayer).

important steps and published papers which have contributed to it. This survey is, of course, strongly influenced by the special experience and taste of the authors.

A famous and very old example of an interval enclosure is given by the method due to Archimedes. He considered inscribed polygons and circumscribing polygons of a circle with radius 1 and obtained an increasing sequence of lower bounds and at the same time a decreasing sequence of upper bounds for the aera of the corresponding disc. Thus stopping this process with a circumscribing and an inscribed polygon, each of n sides, he obtained an interval containing the number π . By choosing n large enough, an interval of arbitrary small width can be found in this way containing π .

One of the first references to interval arithmetic as a tool in numerical computing can already be found in [35, p. 346 ff] (originally published in Russian in 1951) where the rules for the arithmetic of intervals (in the case that both operands contain only positive numbers) are explicitly stated and applied to what is called today interval arithmetic evaluation of rational expressions (see Section 2 of the present paper). For example, the following problem is discussed: What is the range of the expression

$$x = \frac{a+b}{(a-b)c}$$

if the exact values of a, b and c are known to lie in certain given intervals. By plugging in the given intervals the expression for x delivers a superset of the range of x.

According to Moore [64] P.S. Dwyer has discussed matrix computations using interval arithmetic already in his book [29] in 1951.

Probably the most important paper for the development of interval arithmetic has been published by the Japanese scientist Teruo Sunaga [88]. In this publication not only the algebraic rules for the basic operations with intervals can be found but also a systematic investigation of the rules which they fulfill. The general principle of bounding the range of a rational function over an interval by using only the endpoints via interval arithmetic evaluation is already discussed. Furthermore, interval vectors are introduced (as multidimensional intervals) and the corresponding operations are discussed. The idea of computing an improved enclosure for the zero of a real function by what is today called interval Newton method is already presented in Sunaga's paper (Example 9.1). Finally, bounding the value of a definite integral by bounding the remainder term using interval arithmetic tools and computing a pointwise enclosure for the solution of an initial value problem by remainder term enclosing have already been discussed there. Although written in English these results did not find much attention until the first book on interval analysis appeared which was written by Moore [64].

Moore's book was the outgrowth of his Ph.D. thesis [63] and therefore was mainly concentrated on bounding solutions of initial value problems for ordinary differential equations although it contained also a whole bunch of general ideas.

After the appearance of Moore's book groups from different countries started to investigate the theory and application of interval arithmetic systematically. One of the first survey articles following Moore's book was written by Kulisch [49]. Based on this article the book [12] was written which was translated to English in 1983 as [13].

The interplay between algorithms and the realization on digital computers was thoroughfully investigated by U. Kulisch and his group. Already in the 1960s, an ALGOL extension was created and

implemented which had a type for real intervals including provision of the corresponding arithmetic and related operators.

During the last three decades the role of compact intervals as independent objects has continuously increased in numerical analysis when verifying or enclosing solutions of various mathematical problems or when proving that such problems cannot have a solution in a particular given domain. This was possible by viewing intervals as extensions of real or complex numbers, by introducing interval functions and interval arithmetics and by applying appropriate fixed point theorems. In addition thoroughful and sophisticated implementations of these arithmetics on a computer together with – partly new – concepts such as controlled roundings, variable precision, operator overloading or epsilon–inflation made the theory fruitful in practice and effected that in many fields solutions could be automatically verified and (mostly tightly) enclosed by the computer.

In this survey article we report on some interval arithmetic tools. In particular, we present various crucial theorems which form the starting point for efficient interval algorithms. In Section 2 we introduce the basic facts of the 'standard' interval arithmetic: We define the arithmetic operations, list some of its properties and present a first way how the range of a given function can be included. We continue this latter topic in Section 3 where we also discuss the problem of overestimation of the range. Finally, we demonstrate how range inclusion (of the first derivative of a given function) can be used to compute zeros by a so-called enclosure method.

An enclosure method usually starts with an interval vector which contains a solution and improves this inclusion iteratively. The question which has to be discussed is under what conditions is the sequence of including interval vectors convergent to the solution. This will be discussed in Section 4 for selected enclosure methods of nonlinear systems. An interesting feature of such methods is that they can also be used to prove that there exists no solution in an interval vector. It will be shown that this proof needs only few steps if the test vector has already a small enough diameter. We also demonstrate how for a given nonlinear system a test vector can be constructed which will very likely contain a solution.

In Section 5 we address to systems of linear equations Ax = b, where we allow A and b to vary within given matrix and vector bounds, respectively. The ideas of Section 4 are refined and yield to interval enclosures of the corresponding set of solutions. As a particularity we restrict A within its bounds to be a symmetric matrix and provide methods for enclosing the associated smaller symmetric solution set. In both cases we show how the amount of overestimation by an interval vector can be measured without knowing the exact solution set.

Section 6 is devoted to mildly nonlinear topics such as the algebraic eigenvalue problem, the generalized algebraic eigenvalue problem, the singular value problem, and - as an application - a particular class of inverse eigenvalue problems.

In Section 7 we present crucial ideas for verifying and enclosing solutions of initial value problems for ordinary differential equations. For shortness, however, we must confine to the popular class of interval Taylor series methods.

Section 8 contains some remarks concerning selected classes of partial differential equations of the second order. We mainly consider elliptic boundary value problems and present an access which leads to a powerful verification method in this field.

The practical importance of interval analysis depends heavily on its realization on a computer. Combining the existing machine arithmetic with direct roundings it is possible to implement an interval arithmetic in such a way that all interval algorithms keep their – theoretically proved –

properties on existence, uniqueness and enclosure of a solution when they are performed on a computer. Based on such a machine interval arithmetic, software is available which delivers verified solutions and bounds for them in various fields of mathematics. We will shortly consider this topic in Section 9.

In the last 20 years both the algorithmic components of interval arithmetic and their realization on computers (including software packages for different problems) were further developed. Today the understanding of the theory and the use of adapted programming languages are indispensible tools for reliable advanced scientific computing.

2. Definitions, notations and basic facts

Let $[a] = [\underline{a}, \overline{a}]$, $b = [\underline{b}, \overline{b}]$ be real compact intervals and \circ one of the basic operations 'addition', 'subtraction', 'multiplication' and 'division', respectively, for real numbers, that is $\circ \in \{+, -, \cdot, /\}$. Then we define the corresponding operations for intervals [a] and [b] by

$$[a] \circ [b] = \{a \circ b | a \in [a], b \in [b]\},\tag{1}$$

where we assume $0 \notin [b]$ in case of division.

It is easy to prove that the set $I(\mathbb{R})$ of real compact intervals is closed with respect to these operations. What is even more important is the fact that $[a] \circ [b]$ can be represented by using only the bounds of [a] and [b]. The following rules hold:

$$[a] + [b] = [\underline{a} + \underline{b}, \overline{a} + b],$$

$$[a] - [b] = [\underline{a} - \overline{b}, \overline{a} - \underline{b}],$$

$$[a] \cdot [b] = [\min\{\underline{ab}, \underline{a}\overline{b}, \overline{a}\underline{b}, \overline{a}\overline{b}\}, \max\{\underline{ab}, \underline{a}\overline{b}, \overline{a}\underline{b}, \overline{a}\overline{b}\}].$$

If we define

$$\frac{1}{[b]} = \left\{ \frac{1}{b} \middle| b \in [b] \right\} \quad \text{if } 0 \notin [b],$$

then

$$[a]/[b] = [a] \cdot \frac{1}{[b]}.$$

If $\underline{a} = \overline{a} = a$, i.e., if [a] consists only of the element a, then we identify the real number a with the degenerate interval [a,a] keeping the real notation, i.e., $a \equiv [a,a]$. In this way one recovers at once the real numbers \mathbb{R} and the corresponding real arithmetic when restricting $I(\mathbb{R})$ to the set of degenerate real intervals equipped with the arithmetic defined in (1). Unfortunately, $(I(\mathbb{R}), +, \cdot)$ is neither a field nor a ring. The structures $(I(\mathbb{R}), +)$ and $(I(\mathbb{R})/\{0\}, \cdot)$ are commutative semigroups with the neutral elements 0 and 1, respectively, but they are not groups. A nondegenerate interval [a] has no inverse with respect to addition or multiplication. Even the distributive law has to be replaced by the so-called subdistributivity

$$[a]([b] + [c]) \subseteq [a][b] + [a][c].$$
⁽²⁾

The simple example $[-1,1](1+(-1))=0 \subset [-1,1] \cdot 1+[-1,1] \cdot (-1)=[-2,2]$ illustrates (2) and shows that -[-1,1] is certainly not the inverse of [-1,1] with respect to +. It is worth noticing

that equality holds in (2) in some important particular cases, for instance if [a] is degenerate or if [b] and [c] lie on the same side with respect to 0.

From (1) it follows immediately that the introduced operations for intervals are inclusion monotone in the following sense:

$$[a] \subseteq [c], [b] \subseteq [d] \Rightarrow [a] \circ [b] \subseteq [c] \circ [d].$$

$$(3)$$

Standard interval functions $\varphi \in F = \{ sin, cos, tan, arctan, exp, ln, abs, sqr, sqrt \}$ are defined via their range, i.e.,

$$\varphi([x]) = \{\varphi(x) | x \in [x]\}.$$

$$\tag{4}$$

Apparently, they are extensions of the corresponding real functions. These real functions are continuous and piecewise monotone on any compact subinterval of their domain of definition. Therefore, the values $\varphi([x])$ can be computed directly from the values at the bounds of [x] and from selected constants such as 0 in the case of the square, or -1, 1 in the case of sine and cosine. It is obvious that the standard interval functions are inclusion monotone, i.e., they satisfy

$$[x] \subseteq [y] \Rightarrow \varphi([x]) \subseteq \varphi([y]). \tag{5}$$

Let $f: D \subseteq \mathbb{R} \to \mathbb{R}$ be given by a mathematical expression f(x) which is composed by finitely many elementary operations $+, -, \cdot, /$ and standard functions $\varphi \in F$. If one replaces the variable x by an interval $[x] \subseteq D$ and if one can evaluate the resulting interval expression following the rules in (1) and (4) then one gets again an interval. It is denoted by f([x]) and is usually called (an) interval arithmetic evaluation of f over [x]. For simplicity and without mentioning it separately we assume that f([x]) exists whenever it occurs in the paper.

From (3) and (5) the interval arithmetic evaluation turns out to be inclusion monotone, i.e.,

$$[x] \subseteq [y] \Rightarrow f([x]) \subseteq f([y]) \tag{6}$$

holds. In particular, f([x]) exists whenever f([y]) does for $[y] \supseteq [x]$. From (6) we obtain

$$x \in [x] \Rightarrow f(x) \in f([x]), \tag{7}$$

whence

$$R(f;[x]) \subseteq f([x]). \tag{8}$$

Here R(f; [x]) denotes the range of f over [x].

Relation (8) is the fundamental property on which nearly all applications of interval arithmetic are based. It is important to stress what (8) really is delivering: Without any further assumptions is it possible to compute lower and upper bounds for the range over an interval by using only the bounds of the given interval.

Example 1. Consider the rational function

$$f(x) = \frac{x}{1-x}, \qquad x \neq 1,$$

and the interval [x] = [2,3]. It is easy to see that

 $R(f; [x]) = [-2, -\frac{3}{2}],$

f([x]) = [-3, -1],

which confirms (8).

For $x \neq 0$ we can rewrite f(x) as

$$f(x) = \frac{1}{1/x - 1}, \quad x \neq 0, \ x \neq 1$$

and replacing x by the interval [2,3] we get

$$\frac{1}{1/[2,3]-1} = [-2, -\frac{3}{2}] = R(f; [x]).$$

From this example it is clear that the quality of the interval arithmetic evaluation as an enclosure of the range of f over an interval [x] is strongly dependent on how the expression for f(x) is written. In order to measure this quality we introduce the so-called Hausdorff distance $q(\cdot, \cdot)$ between intervals with which $I(\mathbb{R})$ is a complete metric space:

Let $[a] = [\underline{a}, \overline{a}], [b] = [\underline{b}, \overline{b}]$, then

$$q([a],[b]) = \max\{|\underline{a} - \underline{b}|, |\bar{a} - \bar{b}|\}.$$
(9)

Furthermore, we use

$$\begin{split} \check{a} &= \frac{1}{2}(\underline{a} + \bar{a}), \\ d[a] &= \bar{a} - \underline{a}, \\ |[a]| &= \max\{|a| \, | a \in [a]\} = \max\{|\underline{a}|, |\bar{a}|\}, \\ \langle [a] \rangle &= \min\{|a| | a \in [a]\} = \begin{cases} 0, & \text{if } 0 \in [a], \\ \min\{|\underline{a}|, |\bar{a}|\} & \text{if } 0 \notin [a] \end{cases} \end{split}$$
(10)

and call \check{a} center, d[a] diameter and |[a]| absolute value of [a].

In order to consider multidimensional problems we introduce $m \times n$ interval matrices $[A] = ([a_{ij}])$ with entries $[a_{ij}]$, i = 1, ..., m, j = 1, ..., n, and interval vectors $[x] = ([x_i])$ with n components $[x_i]$, i = 1, ..., n. We denote the corresponding sets by $I(\mathbb{R}^{m \times n})$ and $I(\mathbb{R}^n)$, respectively. Trivially, [A] coincides with the matrix interval $[\underline{A}, \overline{A}] = \{B \in \mathbb{R}^{m \times n} | \underline{A} \leq B \leq \overline{A}\}$ if $\underline{A} = (\underline{a}_{ij}), \overline{A} = (\overline{a}_{ij}) \in \mathbb{R}^{m \times n}$ and if $A = (a_{ij}) \leq B = (b_{ij})$ means $a_{ij} \leq b_{ij}$ for all i, j. Since interval vectors can be identified with $n \times 1$ matrices, a similar property holds for them. The null matrix O and the identity matrix I have the usual meaning, e denotes the vector $e = (1, 1, ..., 1)^T \in \mathbb{R}^n$. Operations between interval matrices and between interval vectors are defined in the usual manner. They satisfy an analogue of (6)-(8). For example,

$$\{Ax \mid A \in [A], x \in [x]\} \subseteq [A][x] = \left(\sum_{j=1}^{n} [a_{ij}][x_j]\right) \in I(\mathbb{R}^m)$$

$$(11)$$

if $[A] \in I(\mathbb{R}^{m \times n})$ and $[x] \in I(\mathbb{R}^n)$. It is easily seen that [A][x] is the smallest interval vector which contains the left set in (11), but normally it does not coincide with it. An interval item which encloses some set *S* as tight as possible is called (interval) hull of *S*. The above-mentioned operations with two interval operands always yield to the hull of the corresponding underlying sets.

An interval matrix $[A] \in I(\mathbb{R}^{n \times n})$ is called nonsingular if it contains no singular real $n \times n$ matrix.

The Hausdorff distance, the center, the diameter and the absolute value in (9), (10) can be generalized to interval matrices and interval vectors, respectively, by applying them entrywise. Note that the results are real matrices and vectors, respectively, as can be seen, e.g., for

$$q([A], [B]) = (q([a_{ij}], [b_{ij}])) \in \mathbb{R}^{m \times n}$$

if [A], [B] $\in I(\mathbb{R}^{m \times n})$. We also use the comparison matrix $\langle [A] \rangle = (c_{ij}) \in \mathbb{R}^{n \times n}$ which is defined for $[A] \in I(\mathbb{R}^{n \times n})$ by

$$c_{ij} = \begin{cases} \langle [a_{ij}] \rangle & \text{ if } i = j, \\ -|[a_{ij}]| & \text{ if } i \neq j. \end{cases}$$

By int([x]) we denote the interior of an interval vector [x], by $\rho(A)$ the spectral radius of $A \in \mathbb{R}^{n \times n}$ and by $|| \cdot ||_{\infty}$ the usual maximum norm for vectors from \mathbb{R}^n or the row sum norm for matrices from $\mathbb{R}^{n \times n}$. In addition, the Euclidean norm $|| \cdot ||_2$ in \mathbb{R}^n will be used. We recall that $A \in \mathbb{R}^{n \times n}$ is an Mmatrix if $a_{ij} \leq 0$ for $i \neq j$ and if A^{-1} exists and is nonnegative, i.e., $A^{-1} \geq O$. If each matrix A from a given interval matrix [A] is an M matrix then we call [A] an M matrix, too.

Let each component f_i of $f: D \subseteq \mathbb{R}^m \to \mathbb{R}^n$ be given by an expression $f_i(x)$, i = 1, ..., n, and let $[x] \subseteq D$. Then the interval arithmetic evaluation f([x]) is defined analogously to the one-dimensional case.

In this paper we restrict ourselves to real compact intervals. However, complex intervals of the form [z] = [a] + i[b] ($[a], [b] \in I(\mathbb{R})$) and $[z] = \langle \tilde{z}, r \rangle$ ($\tilde{z}, r \in \mathbb{R}, r \ge 0$) are also used in practice. In the first form [z] is a rectangle in the complex plane, in the second form it means a disc with midpoint \tilde{z} and radius r. In both cases a complex arithmetic can be defined and complex interval functions can be considered which extend the presented ones. See [3,13] or [73], e.g., for details.

3. Computing the range of real functions by interval arithmetic tools

Enclosing the range R(f; [x]) of a function $f: D \subseteq \mathbb{R}^n \to \mathbb{R}^m$ with $[x] \subseteq D$ is an important task in interval analysis. It can be used, e.g., for

- localizing and enclosing global minimizers and global minima of f on [x] if m = 1,
- verifying $R(f; [x]) \subseteq [x]$ which is needed in certain fixed point theorems for f if m = n,
- enclosing R(f'; [x]), i.e., the range of the Jacobians of f if m = n,
- enclosing $R(f^{(k)}; [x])$, i.e., the range of the kth derivative of f which is needed when verifying and enclosing solutions of initial value problems,
- verifying the nonexistence of a zero of f in [x].

According to Section 2 an interval arithmetic evaluation f([x]) is automatically an enclosure of R(f, [x]). As Example 1 illustrates f([x]) may overestimate this range. The following theorem shows how large this overestimation may be.

Theorem 1 (Moore [64]). Let $f: D \subset \mathbb{R}^n \to \mathbb{R}$ be continuous and let $[x] \subseteq [x]^0 \subseteq D$. Then (under mild additional assumptions)

 $q(R(f;[x]), f([x])) \leq \gamma ||d[x]||_{\infty}, \quad \gamma \geq 0,$

$$df([x]) \leq \delta ||d[x]||_{\infty}, \quad \delta \geq 0,$$

where the constants γ and δ depend on $[x]^0$ but not on [x].

Theorem 1 states that if the interval arithmetic evaluation exists then the Hausdorff distance between R(f; [x]) and f([x]) goes linearly to zero with the diameter d[x]. Similarly the diameter of the interval arithmetic evaluation goes linearly to zero if d[x] is approaching zero.

On the other hand, we have seen in the second part of Example 1 that f([x]) may be dependent on the expression which is used for computing f([x]). Therefore the following question is natural:

Is it possible to rearrange the variables of the given function expression in such a manner that the interval arithmetic evaluation gives higher than linear order of convergence to the range of values?

A first result in this respect shows why the interval arithmetic evaluation of the second expression in Example 1 is optimal:

Theorem 2 (Moore [64]). Let a continuous function $f:D \subset \mathbb{R}^n \to \mathbb{R}$ be given by an expression f(x) in which each variable x_i , i = 1, ..., n, occurs at most once. Then

$$f([x]) = R(f; [x]) \quad for \ all \ [x] \subseteq D.$$

Unfortunately, not many expressions f(x) can be rearranged such that the assumptions of Theorem 2 are fulfilled. In order to propose an alternative we consider first a simple example.

Example 2. Let $f(x) = x - x^2$, $x \in [0, 1] = [x]^0$. It is easy to see that for $0 \le r \le \frac{1}{2}$ and $[x] = [\frac{1}{2} - r, \frac{1}{2} + r]$ we have

$$R(f; [x]) = \left[\frac{1}{4} - r^2, \frac{1}{4}\right]$$

and

$$f([x]) = [\frac{1}{4} - 2r - r^2, \frac{1}{4} + 2r - r^2].$$

From this it follows

$$q(R(f; [x]), (f([x])) \leq \gamma d[x] \text{ with } \gamma = 1$$

and

 $df([x]) \leq \delta d[x]$ with $\delta = 2$

in agreement with Theorem 1.

If we rewrite f(x) as

$$x - x^2 = \frac{1}{4} - (x - \frac{1}{2})(x - \frac{1}{2})$$

and plug in the interval $[x] = [\frac{1}{2} - r, \frac{1}{2} + r]$ on the right-hand side then we get the interval $[\frac{1}{4} - r^2, \frac{1}{4} + r^2]$ which, of course, includes R(f; [x]) again, and

$$q(R(f; [x]), [\frac{1}{4} - r^2, \frac{1}{4} + r^2]) = r^2 = \frac{1}{4}(d[x])^2$$

Hence the distance between R(f; [x]) and the enclosure interval $[\frac{1}{4} - r^2, \frac{1}{4} + r^2]$ goes quadratically to zero with the diameter of [x].

The preceding example is an illustration for the following general result.

Theorem 3 (The centered form). Let the function $f:D \subseteq \mathbb{R}^n \to \mathbb{R}$ be represented in the 'centered form'

$$f(x) = f(z) + h(x)^{T}(x - z)$$
(12)

for some $z \in [x] \subseteq [x]^0 \subseteq D$ and $h(x) \in \mathbb{R}^n$. If

$$f([x]) = f(z) + h([x])^{\mathrm{T}}([x] - z),$$
(13)

then

$$R(f;[x]) \subseteq f([x]) \tag{14}$$

and (under some additional assumptions)

$$q(R(f;[x]), f([x])) \leq \kappa ||d[x]||_{\infty}^{2}, \quad \kappa \geq 0,$$
(15)

where the constant κ depends on $[x]^0$ but not on [x] and z.

Relation (15) is called 'quadratic approximation property' of the centered form. For rational functions it is not difficult to find a centered form, see for example [77].

After having introduced the centered form it is natural to ask if there are forms which deliver higher than quadratic order of approximation of the range. Unfortunately, this is not the case as has been shown recently by Hertling [39]; see also [70].

Nevertheless, in special cases one can use the so-called generalized centered forms to get higherorder approximations of the range; see, e.g., [18]. Another interesting idea which uses a so-called 'remainder form of f' was introduced by Cornelius and Lohner [27].

Finally, we can apply the subdivision principle in order to improve the enclosure of the range. To this end we represent $[x] \in I(\mathbb{R}^n)$ as the union of k^n interval vectors $[x]^l$, $l = 1, ..., k^n$, such that $d[x_i]^l = d[x_i]/k$ for i = 1, ..., n and $l = 1, ..., k^n$. Defining

$$f([x];k) = \bigcup_{l=1}^{k^n} f([x]^l),$$
(16)

the following result holds:

Theorem 4. Let $f:D \subseteq \mathbb{R}^n \to \mathbb{R}$. (a) With the notations and assumptions of Theorem 1 and with (16) we get

$$q(R(f; [x]), f([x]; k)) \leq \frac{\hat{\gamma}}{k},$$

where $\hat{\gamma} = \gamma ||d[x]^0||_{\infty}.$

(b) Let the notations and assumptions of Theorem 3 hold. Then using in (16) for $f([x]^{l})$ the expression (13) with $z = z^{l} \in [x]^{l}$, l = 1, ..., k, it follows that

$$q(R(f; [x]), f([x]; k)) \leq \frac{\hat{\kappa}}{k^2},$$

where $\hat{\kappa} = \kappa ||d[x]^0||_{\infty}^2.$

Theorem 4 shows that the range can be enclosed arbitrarily close if k tends to infinity, i.e., if the subdivision of $[x] \subseteq [x]^0$ is sufficiently fine, for details see, e.g., [78].

In passing we note that the principal results presented up to this point provide the basis for enclosing minimizers and minima in global optimization. Necessary refinements for practical algorithms in this respect can be found in, e.g., [36,37,38,42,44] or [79].

As a simple example for the demonstration how the ideas of interval arithmetic can be applied we consider the following problem:

Let there be given a continuously differentiable function $f: D \subset \mathbb{R} \to \mathbb{R}$ and an interval $[x]^0 \subseteq D$ for which the interval arithmetic evaluation of the derivative exists and does not contain zero: $0 \notin f'([x]^0)$. We want to check whether there exists a zero x^* in $[x]^0$, and if it exists we want to compute it by producing a sequence of intervals containing x^* with the property that the lower and upper bounds are converging to x^* . (Of course, checking the existence is easy in this case by evaluating the function at the endpoints of $[x]^0$. However, the idea following works also for systems of equations. This will be shown in the next section.)

For $[x] \subseteq [x]^0$ we introduce the so-called interval Newton operator

$$N[x] = m[x] - \frac{f(m[x])}{f'([x])}, \quad m[x] \in [x]$$
(17)

and consider the following iteration method:

$$[x]^{k+1} = N[x]^k \cap [x]^k, \quad k = 0, 1, 2, \dots,$$
(18)

which is called interval Newton method.

Properties of operator (17) and method (18) are described in the following result.

Theorem 5. Under the above assumptions the following holds for (17) and (18): (a) *If*

$$N[x] \subseteq [x] \subseteq [x]^0, \tag{19}$$

then f has a zero $x^* \in [x]$ which is unique in $[x]^0$. (b) If f has a zero $x^* \in [x]^0$ then $\{[x]^k\}_{k=0}^{\infty}$ is well defined, $x^* \in [x]^k$ and $\lim_{k\to\infty} [x]^k = x^*$. If $df'([x]) \leq cd[x], \ [x] \subseteq [x]^0$, then $d[x]^{k+1} \leq \gamma (d[x]^k)^2$. . ^

(c)
$$N[x]^{\kappa_0} \cap [x]^{\kappa_0} = \emptyset$$
 (= empty set) for some $k_0 \ge 0$ if and only if $f(x) \ne 0$ for all $x \in [x]^{\circ}$.

Theorem 5 delivers two strategies to study zeros in $[x]^0$. By the first it is proved that f has a unique zero x^* in $[x]^0$. It is based on (a) and can be realized by performing (18) and checking (19) with $[x] = [x]^k$. By the second – based on (c) – it is proved that f has no zero x^* in $[x]^0$. While the second strategy is always successful if $[x]^0$ contains no zero of f the first one can fail as the simple example $f(x) = x^2 - 4$, $[x]^0 = [2, 4]$ shows when choosing $m[x]^k > \underline{x}^k$. Here the iterates have the form $[x]^{k} = [2, a_{k}]$ with appropriate $a_{k} > 2$ while $N[x]^{k} < 2$. Hence (19) can never be fulfilled.

In case (b), the diameters are converging quadratically to zero. On the other hand, if method (18) breaks down because of empty intersection after a finite number of steps then from a practical point of view it would be interesting to have qualitative knowledge about the size of k_0 in this case. This will be discussed in the next section in a more general setting.

4. Systems of nonlinear equations

In the present section we consider systems of nonlinear equations in the form

$$f(x) = 0 \tag{20}$$

and

$$f(x) = x, \tag{21}$$

respectively, i.e., we look for zeros and for fixed points of f, respectively. (It is well known that problems (20) and (21) are equivalent when choosing f in (21) appropriately.) Using interval arithmetic we want to derive simple criteria which guarantee that a given interval [x] contains at least one zero x^* of f or a corresponding fixed point. We also list conditions for x^* to be unique within [x], and we show how [x] can be improved iteratively to some vector $[x]^*$ which contains x^* and has a smaller diameter.

In the whole section we assume that $f:D \subseteq \mathbb{R}^n \to \mathbb{R}^n$ is at least continuous in D, and often we assume that it is at least once continuously (Fréchet-) differentiable.

We first consider fixed points x^* of f in $[x] \subseteq D$. A simple method for verifying such a point is based on (6)-(8) and Brouwer's fixed point theorem and reads as follows.

Theorem 6. Let $f:D \subseteq \mathbb{R}^n \to \mathbb{R}^n$ be continuous and let

$$f([x]) \subseteq [x] \subseteq D. \tag{22}$$

Then f has at least one fixed point in [x] and the iteration

$$[x]^{0} = [x],$$

$$[x]^{k+1} = f([x]^{k}), \quad k = 0, 1, \dots$$
 (23)

converges to some $[x]^*$ such that

$$[x]^* \subseteq [x]^{k+1} \subseteq [x]^k \subseteq \dots \subseteq [x]^0 = [x].$$

$$(24)$$

The limit $[x]^*$ contains all fixed points of f in [x].

We call an interval sequence $\{[x]^k\}_{k=0}^{\infty}$ monotonically decreasing if it fulfills (24). Theorem 6 says nothing on the uniqueness of $x^* \in [x]$ nor on the width of $[x]^*$. In fact, the simple example f(x) = -x, [x] = [-1, 1] with $[x]^k = [x]^* = [x]$ shows that $d[x]^* > 0$ can occur although $x^* = 0$ is the only fixed point of f in \mathbb{R} . For P contractions, however, sharper results can be proved by a direct application of Banach's fixed point theorem. Note that $f:D \subseteq \mathbb{R}^n \to \mathbb{R}^n$ is a *P* contraction on the set I([x]) of all compact intervals contained in $[x] \subseteq D$ if there is a matrix $P \ge O \in \mathbb{R}^{n \times n}$ with spectral radius $\rho(P) < 1$ and

 $q(f([y]), f([z])) \leq Pq([y], [z])$ for all $[y], [z] \subseteq [x]$.

Trivial examples are linear functions f(x) = Ax - b with $D = \mathbb{R}^n$, $A \in \mathbb{R}^{n \times n}$, $\rho(|A|) < 1$, $b \in \mathbb{R}^n$ and P = |A|.

Theorem 7. Let $f:D \subseteq \mathbb{R}^n \to \mathbb{R}^n$ be a P contraction on I([x]), $[x] \subseteq D$, and let (22) hold. Then f has exactly one fixed point $x^* \in [x]$ and iteration (23) converges to x^* for all starting vectors $[x]^0 \subseteq [x]$. Moreover, $x^* \in [x]^k$, k = 1, 2, ..., if $x^* \in [x]^0$ which holds, in particular, if $[x]^0 = [x]$.

Remark 1. Condition (22) can be omitted in Theorem 7 if f is a P contraction on the whole space $I(\mathbb{R}^n)$ (cf. [13]). For any $[x]^0 \in I(\mathbb{R}^n)$ the unique fixed point x^* is then contained in $[-\underline{x}^0 - \Delta, \overline{x}^0 + \Delta]$, $\Delta = (I - P)^{-1}q([x]^1, [x]^0)$.

Remark 1 is interesting since it is not always an easy task to find an [x] such that (22) holds. There is, however, a method of trial and error which goes back to Rump [81] and which, in practice, mostly ends up with such an [x] in a few steps. The technique is called epsilon inflation and is a quite general interval arithmetic tool. It consists in replacing the current interval iterate by an interval vector which is a proper superset of the iterate and which differs from it by a small parameter ε . This can be done, e.g., in the following way: first compute an approximation \tilde{x} of x^* by applying any appropriate standard method in numerical analysis. Then iterate according to

$$[x]^{0} = \tilde{x},$$

$$[x]^{k+1} = f([x]^{k} + d[x]^{k} [-\varepsilon, \varepsilon] + [-\eta, \eta]e), \quad k = 0, 1, ...,$$
 (25)

where ε , η are some small positive real numbers. If f is a P contraction on $I(\mathbb{R}^n)$ then (25) ends up after finitely many steps with an iterate which fulfills (22). This is stated in our next theorem.

Theorem 8. Let $f:D = \mathbb{R}^n \to \mathbb{R}^n$ be a *P* contraction on $I(\mathbb{R}^n)$. With $[x]^0_{\varepsilon}$ being given, iterate by inflation according to

$$[x]_{\varepsilon}^{k+1} = f([x]_{\varepsilon}^{k}]) + [\delta]^{k}, \quad k = 0, 1, \dots$$

where $[\delta]^k \in I(\mathbb{R}^n)$ are given vectors which converge to some limit $[\delta]$. If $0 \in int([\delta])$ then there is an integer $k_0 = k_0([x]_{\varepsilon}^0)$ such that

$$f([x]_{\varepsilon}^{k_0}) \subseteq \operatorname{int}([x]_{\varepsilon}^{k_0}).$$

In view of (25) we can try to apply Theorem 8 with $[\delta]^k = (df[x]^k_{\varepsilon})[-\varepsilon,\varepsilon] + [-\eta,\eta]e$ and $[x]^0_{\varepsilon} = [x]^0 + (d[x]^0)[-\varepsilon,\varepsilon] + [-\eta,\eta]e$. If $[\delta] = \lim_{k\to\infty} [\delta]^k$ exists then $0 \in int([\delta])$ since $0 \in [-\eta,\eta]e \subseteq [\delta]^k$ for $k = 0, 1, \ldots$

Theorem 8 was originally stated and proved by Rump [83] for linear functions f. It was generalized to P contractions and contractive interval functions in [58,59] where also the case $D \neq \mathbb{R}^n$
is considered and where various examples for epsilon inflations are presented. Unfortunately, Theorem 8 says nothing on the number of steps which are needed to succeed with (22). Therefore, other possibilities become interesting which we are going to present in the second part of this section and in Section 6.

We consider now zeros of a given function f.

A first method is based on a result of C. Miranda (see [62] or Corollary 5.3.8 in [69]) which is equivalent to Brouwer's fixed point theorem. We use it in the following modified interval version.

Theorem 9. Let
$$f:D \subseteq \mathbb{R}^n \to \mathbb{R}^n$$
 be continuous and let $[x] \subseteq D$,
 $[l]^i = ([x_1], ..., [x_{i-1}], \underline{x}_i, [x_{i+1}], ..., [x_n])^T$,
 $[u]^i = ([x_1], ..., [x_{i-1}], \overline{x}_i, [x_{i+1}], ..., [x_n])^T$.
If $\overline{f_i([l]^i)} \leq 0$, $\underline{f_i([u]^i)} \geq 0$ or $\underline{f_i([l]^i)} \geq 0$, $\overline{f_i([u]^i)} \leq 0$ holds for each $i = 1, ..., n$ then f has at least one zero in $[x]$.

Combined with subdivisions, lists and exclusion techniques Theorem 9 forms the basis of a simple but efficient verification and enclosure method for zeros of functions $f:D \subseteq \mathbb{R}^n \to \mathbb{R}^m$ even if m < n. Curves and surfaces can thus be tightly enclosed and problems in CAGD like ray tracing can be handled. We refer to [31,52,68].

Another method for verifying zeros consists in generalizing the interval Newton method of Section 3 to the multidimensional case. To this end we denote by

IGA([A], [b]),

the result of the Gaussian algorithm applied formally to a nonsingular interval matrix $[A] \in I(\mathbb{R}^{n \times n})$ and an interval vector $[b] \in I(\mathbb{R}^n)$, see, for example, [13, Section 15]. Here we assumed that no division by an interval which contains zero occurs in the elimination process. It is easy to see that

$$S = \{x = A^{-1}b \mid A \in [A], b \in [b]\} \subseteq \text{IGA}([A], [b])$$
(26)

holds. By

IGA([A])

we denote the interval matrix whose *i*th column is obtained as $IGA([A], e^i)$ where e^i is the *i*th unit vector. In other words, IGA([A]) is an enclosure for the inverses of all matrices $A \in [A]$.

Now assume that

$$f:D \subset \mathbb{R}^n \to \mathbb{R}^n \tag{27}$$

is continuously differentiable. If $x, y \in [x] \subseteq D$ then

$$f(x) - f(y) = J(y,x)(x - y),$$
 (28)

where

$$J(y,x) = \int_0^1 f'(y + t(x - y)) dt.$$
 (29)

Note that J is a continuous mapping of x and y which satisfies J(y,x) = J(x, y). Since $t \in [0,1]$ we have $y + t(x - y) \in [x]$ and therefore

$$J(y,x) \in f'([x]), \tag{30}$$

where f'([x]) denotes the interval arithmetic evaluation of the Jacobian of f. For fixed $y \in [x]$ we obtain from (28) and (30)

$$p(x) = x - J^{-1}(y, x)f(x) = y - J^{-1}(y, x)f(y) \in y - \text{IGA}(f'([x]), f(y)).$$
(31)

If $x \in [x]$ is a zero of f then (31) implies $x \in y - \text{IGA}(f'([x]), f(y))$. This leads to the following definition of the interval Newton operator N[x] which we introduce in analogy to (18): suppose that $m[x] \in [x]$ is a real vector. Then

$$N[x] = m[x] - \text{IGA}(f'([x]), f(m[x])).$$
(32)

The interval Newton method is defined by

$$[x]^{k+1} = N[x]^k \cap [x]^k, \quad k = 0, 1, 2, \dots$$
(33)

Analogously to Theorem 5 we have the following result.

Theorem 10. Let $f: D \subseteq \mathbb{R}^n \to \mathbb{R}^n$ be continuously differentiable and assume that $IGA(f'([x]^0))$ exists for some interval vector $[x]^0 \subseteq D$: (This is identical to assuming that the Gaussian algorithm is feasible for $f'([x]^0)$. In particular, $f'([x]^0)$ is nonsingular in this case.) (a) If

 $N[x] \subseteq [x]$

for some $[x] \subseteq [x]^0$ then f has a zero x^* in [x] which is unique even in $[x]^0$. Assume that

$$\rho(A) < 1, \quad \text{where } A = |I - \mathrm{IGA}(f'([x]^0))f'([x]^0)|.$$
(34)

(b) If f has a zero x^* in $[x]^0$ then the sequence $\{[x]^k\}_{k=0}^{\infty}$ defined by (33) is well defined, $x^* \in [x]^k$ and $\lim_{k\to\infty} [x]^k = x^*$. In particular, $\{[x]^k\}_{k=0}^{\infty}$ is monotonically decreasing and x^* is unique in $[x]^0$. Moreover, if

Moreover, if

$$df'([x])_{ij} \leq \alpha ||d[x]||_{\infty}, \quad \alpha \geq 0, \quad 1 \leq i, j \leq n$$
for all $[x] \subseteq [x]^0$ then
$$(35)$$

$$||d[x]^{k+1}||_{\infty} \leq \gamma ||d[x]^{k}||_{\infty}^{2}, \quad \gamma \geq 0.$$
(36)
(c) $N[x]^{k_{0}} \cap [x]^{k_{0}} = \emptyset$ for some $k_{0} \geq 0$ if and only if $f(x) \neq 0$ for all $x \in [x]^{0}$.

The proof of (a) can be quickly done by applying Brouwer's fixed point theorem to p of (31) The results of (b) and (c) can be found in [9].

Note that in contrast to the onedimensional case we need condition (34) in cases (b) and (c).

Because of continuity reasons this condition always holds if the diameter $d[x]^0$ of the given interval vector ('starting interval') is componentwise small enough (and if $f'([x]^0)$ contains no singular matrix) since because of Theorem 1 we have A = O in the limit case $d[x]^0 = 0$. Schwandt

[86] has discussed a simple example in the case $\rho(A) \ge 1$ which shows that for a certain interval vector (33) is feasible, $x^* \in [x]^k$, but $\lim_{k\to\infty} [x]^k \neq x^*$.

In case (a) of the preceding theorem we have by (36) quadratic convergence of the diameters of the enclosing intervals to the zero vector. This is the same favorable behavior as it is well known for the usual Newton method. If there is no solution x^* of f(x) = 0 in $[x]^0$ this can be detected by applying (33) until the intersection becomes empty for some k_0 . From a practical point of view it is important that k_0 is not big in general. Under natural conditions it can really be proved that k_0 is small if the diameter of $[x]^0$ is small:

Let $N[x] = [\underline{n}, \overline{n}]$ for the interval Newton operator (32). It is easy to prove that

$$N[x] \cap [x] = \emptyset$$

if and only if for at least one component i_0 either

$$(\bar{n} - \underline{x})_{i_0} < 0 \tag{37}$$

$$(\bar{x}-\underline{n})_{i_0} < 0 \tag{38}$$

holds. Furthermore, it can be shown that

$$\bar{x} - \underline{n} \leq O(||d[x]||_{\infty}^2)e + A^2 f(\bar{x})$$
(39)

and

$$\bar{n} - \underline{x} \leq \mathcal{O}(||d[x]||_{\infty}^2)e - A^1 f(\underline{x})$$

$$\tag{40}$$

provided (35) holds. Here A^1 and A^2 are two real matrices contained in IGA($f'([x]^0)$). Furthermore, if $f(x) \neq 0$, $x \in [x]$, then for sufficiently small diameter d[x] there is at least one $i_0 \in \{1, 2, ..., n\}$ such that

$$(A^1 f(\underline{x}))_{i_0} \neq 0 \tag{41}$$

and

$$\operatorname{sign}(A^1 f(\underline{x}))_{i_0} = \operatorname{sign}(A^2 f(\overline{x}))_{i_0}.$$
(42)

Assume now that $\operatorname{sign}(A^1 f(\underline{x}))_{i_0} = 1$. Then for sufficiently small diameter d[x] we have $(\overline{n} - \underline{x})_{i_0} < 0$ by (40) and by (37) the intersection becomes empty. If $\operatorname{sign}(A^1 f(\underline{x}))_{i_0} = -1$ then by (39) we obtain $(\overline{x} - \underline{n})_{i_0} < 0$ for sufficiently small d[x] and by (38) the intersection becomes again empty. If $N[x]^{k_0} \cap [x]^{k_0} = \emptyset$ for some k_0 then the interval Newton method breaks down and we speak of

If $N[x]^{k_0} \cap [x]^{k_0} = \emptyset$ for some k_0 then the interval Newton method breaks down and we speak of divergence of this method. Because of the terms $O(||d[x]||_{\infty}^2)$ in (39) and (40) we can say that in the case $f(x) \neq 0$, $x \in [x]^0$, the interval Newton method is quadratically divergent.

We demonstrate this behavior by a simple one-dimensional example.

Example 3. Consider the polynomial

$$f(x) = x^5 + x^4 - 11x^3 - 3x^2 + 18x$$

which has only simple real zeros contained in the interval $[x]^0 = [-5, 6]$. Unfortunately, (18) cannot be performed since $0 \in f'([x]^0)$. Using a modification of the interval Newton method described already in [3] one can compute disjoint subintervals of $[x]^0$ for which the interval arithmetic evaluation does

not contain zero. Hence (18) can be performed for each of these intervals. If such a subinterval contains a zero then (a) of Theorem 5 holds, otherwise (b) is true. Table 1 contains the intervals which were obtained by applying the above-mentioned modification of the interval Newton method until $0 \notin f'([x])$ for all computed subintervals of $[x]^0$ (for simplicity we only give three digits in the mantissa).

The subintervals which do not contain a zero of f are marked by a star in Table 2. The number in the second line exhibits the number of steps until the intersection becomes empty. For n = 9 we have a diameter of approximately 2.75, which is not small, and after only 3 steps the intersection becomes empty. The intervals with the numbers n=1,2,3,6,8 each contain a zero of f. In the second line the number of steps are given which have to be performed until the lower and upper bound can be no longer improved on the computer. These numbers confirm the quadratic convergence of the diameters of the enclosing intervals. (For n = 3 the enclosed zero is $x^* = 0$ and we are in the underflow range.)

For more details concerning the speed of divergence see [8].

6

7

8 9

The interval Newton method has the big disadvantage that even if the interval arithmetic evaluation $f'([x]^0)$ of the Jacobian contains no singular matrix its feasibility is not guaranteed, IGA($f'([x]^0)$), $f(m[x]^0)$ can in general only be computed if $d[x]^0$ is sufficiently small. For this reason Krawczyk [48] had the idea to introduce a mapping which today is called the Krawczyk operator:

Assume again that a mapping (27) with the corresponding properties is given. Then analogously to (32) we consider the so-called Krawczyk operator

$$K[x] = m[x] - Cf(m[x]) + (I - Cf'([x]))([x] - m[x]),$$
(43)

The mo	odified interval Newto f from Example 3	on method ap-
п		
1	$[-0.356 \cdot 10^1,$	$-0.293 \cdot 10^{1}$]
2	$[-0.141 \cdot 10^1,$	$-0.870 \cdot 10^{0}$
3	$[-0.977 \cdot 10^{0},$	$0.499 \cdot 10^{0}$]
4	$[0.501 \cdot 10^0,$	$0.633 \cdot 10^{0}$]
5	$[0.140 \cdot 10^1]$	$0.185 \cdot 10^{1}$

 $[0.188 \cdot 10^1]$

 $[0.265 \cdot 10^1]$ $[0.297 \cdot 10^1]$

 $[0.327 \cdot 10^1,$

 $0.212 \cdot 10^{1}$] $0.269 \cdot 10^{1}$]

 $0.325 \cdot 10^{1}$]

 $0.600 \cdot 10^{1}$]

Table 1	
The modified interval Newton method	ap-
blied to f from Example 3	

Tabl	le 2							
The	interval	Newton	method	applied	to f	from	Example	3

n	1	2	3	4*	5*	6	7*	8	9*
	5	6	9	1	2	6	1	5	3

where C is a nonsingular real matrix and where $m[x] \in [x]$. For fixed C we define the so-called Krawczyk method by

$$[x]^{k+1} = K[x]^k \cap [x]^k, \quad k = 0, 1, 2, \dots$$
(44)

For this method an analogous result holds as was formulated for the interval Newton method in Theorem 10:

Theorem 11. Let $f: D \subseteq \mathbb{R}^n \to \mathbb{R}^n$ be continuously differentiable and assume that the interval arithmetic evaluation $f'([x]^0)$ of the Jacobian exists for some interval vector $[x]^0 \subseteq D^0$. (a) If

$$K[x] \subseteq [x]$$
for some $[x] \subseteq [x]^0$ then f has a zero x^* in $[x]$.
$$(45)$$

If (45) is slightly sharpened to

$$(K[x])_i \subset [x_i] \subseteq [x_i]^0$$
 for $i = 1, ..., n,$
(46)

then $\rho(|I - Cf'([x])|) < 1$ holds, f'([x]) is nonsingular and x^* is unique in [x]. Let m[x] be the center of [x] and assume that

 $\rho(B) < 1$ where $B = |I - Cf'([x]^0)|$.

(b) If f has a zero x^* in $[x]^0$ then the sequence $\{[x]^k\}_{k=0}^{\infty}$ defined by (44) is well defined, $x^* \in [x]^k$ and $\lim_{k\to\infty} [x]^k = x^*$. In particular, $\{[x]^k\}_{k=0}^{\infty}$ is monotonically decreasing and x^* is unique in $[x]^0$. Moreover, if $C = C_k$ varies with k such that it is the inverse of some matrix from $f'([x]^k)$, and if

$$df'([x])_{ij} \leq \alpha ||d[x]||_{\infty}, \quad \alpha \geq 0, \quad 1 \leq i, j \leq n$$
(48)

(47)

for all
$$[x] \subseteq [x]^0$$
 then
 $||d[x]^{k+1}||_{\infty} \leq \gamma ||d[x]^k||_{\infty}^2, \quad \gamma \geq 0.$
(49)

(c) $K[x]^{k_0} \cap [x]^{k_0} = \emptyset$ for some $k_0 \ge 0$ if and only if $f(x) \ne 0$ for all $x \in [x]^0$.

Proof. (a) Consider for the nonsingular matrix C in K[x] the continuous mapping

 $g:D\subseteq\mathbb{R}^n\to\mathbb{R}^n$

defined by

$$g(x) = x - Cf(x).$$

It follows, using (28) and the assumption,

$$g(x) = x - Cf(x)$$

= $x - C(f(x) - f(m[x])) - Cf(m[x])$
= $m[x] + (x - m[x]) - CJ(m[x], x)(x - m[x]) - Cf(m[x])$
 $\in m[x] - Cf(m[x]) + (I - Cf'([x]))([x] - m[x])$
= $K[x] \subseteq [x], \quad x \in [x].$

By Brouwer's fixed point theorem g has a fixed point $x^* \in [x]$. This fixed point is a zero of f.

If (45) is replaced by (46) then $|I - Cf'([x])|d[x] \leq dK[x] < d[x]$. Therefore,

$$\max_{1 \le i \le n} \frac{\sum_{j=1}^{n} |I - Cf'([x])|_{ij} d[x_j]}{d[x_i]} < 1$$

which is equivalent to

$$||\hat{D}^{-1}|I - Cf'([x])|\hat{D}||_{\infty} < 1.$$

Here, \hat{D} is the diagonal matrix with $\hat{d}_{ii} = d[x_i], i = 1, ..., n$. Therefore,

$$\rho(|I - Cf'([x])|) = \rho(\hat{D}^{-1}|I - Cf'([x])|\hat{D}) \leq ||\hat{D}^{-1}|I - Cf'([x])|\hat{D}||_{\infty} < 1.$$

If f'([x]) contained a singular matrix A then I - CA would have the eigenvalue 1 and we would get the contradiction

$$1 \le \rho(I - CA) \le \rho(|I - CA|) \le \rho(|I - Cf'([x])|) < 1.$$
(50)

Therefore, f'([x]) is nonsingular. If f had two zeros x^* , $y^* \in [x]$ then (28) and (30) would imply $x^* = y^*$.

(b) By (28) we have

$$f(x^*) - f(m[x]) = J(m[x], x^*)(x^* - m[x])$$

and since $f(x^*) = 0$ it follows

$$x^* = m[x] - Cf(m[x]) + (I - CJ(m[x], x^*))(x^* - m[x])$$

$$\in m[x] - Cf(m[x]) + (I - Cf'([x]))([x] - m[x])$$

$$= K[x].$$

Hence if $x^* \in [x]^0$ then $x^* \in K[x]^0$ and therefore $x^* \in K[x]^0 \cap [x]^0 = [x]^1$. Mathematical induction proves $x^* \in [x]^k$, $k \ge 0$.

For the diameters of the sequence $\{[x]^k\}_{k=0}^{\infty}$ we have $d[x]^{k+1} \leq dK[x]^k \leq Bd[x]^k$, where the last inequality holds because we assumed that $m[x]^k$ is the center of $[x]^k$. Since $\rho(B) < 1$ we have $\lim_{k\to\infty} d[x]^k = 0$, and from $x^* \in [x]^k$ it follows $\lim_{k\to\infty} [x]^k = x^*$. In particular, x^* is unique within $[x]^0$.

Analogously to (a) assumption (47) implies that $f'([x^0])$ is nonsingular. Since it is compact and since the inverse of a matrix $M \in \mathbb{R}^{n \times n}$ depends continuously on the entries of M the set $\{|M^{-1}| | M \in f'([x]^0)\}$ is bounded by some matrix \hat{C} . The quadratic convergence behavior (49) follows now from

$$d[x]^{k+1} \leq |I - C_k f'([x]^k)| d[x]^k$$

$$\leq |C_k| |C_k^{-1} - f'([x]^k)| d[x]^k$$

$$\leq \hat{C} |f'([x]^k) - f'([x]^k)| d[x]^k$$

$$= \hat{C} df'([x]^k) d[x]^k$$

by using (48).

(c) Assume now that $K[x]^{k_0} \cap [x]^{k_0} = \emptyset$ for some $k_0 \ge 0$. Then $f(x) \ne 0$ for $x \in [x]^0$ since if $f(x^*) = 0$ for some $x^* \in [x]^{\bar{0}}$ then Krawczyk's method is well defined and $x^* \in [x]^k$, $k \ge 0$.

If on the other hand $f(x) \neq 0$ and $K[x]^k \cap [x]^k \neq \emptyset$ then $\{[x]^k\}$ is well defined. Because of $\rho(B) < 1$ we have $d[x]^k \to 0$ and since we have a nested sequence it follows $\lim_{k\to\infty} [x]^k = \hat{x} \in \mathbb{R}^n$. Since the Krawczyk operator is continuous and since the same holds for forming intersections we obtain by passing to infinity in (44)

$$\hat{x} = K\hat{x} \cap \hat{x} = K\hat{x} = \hat{x} - Cf(\hat{x}).$$

From this it follows that $f(\hat{x}) = 0$ in contrast to the assumption that $f(x) \neq 0$ for $x \in [x]^0$. This completes the proof of Theorem 11. \Box

Remark 2. (a) When we defined the Krawczyk operator in (43) we required C to be nonsingular. We need not know this in advance if (45) or (47) holds since either of these two conditions implies the nonsingularity by an analogous argument as in the proof for (a).

(b) It is easy to see that in case (a) of the preceding theorem all the zeros x^* of f in [x] are even in K[x].

(c) If m[x] is not the center of [x] but still an element of it the assertions in (b), (c) remain true if (47) is replaced by $\rho(B) < \frac{1}{2}$.

(d) Assertion (47) certainly holds if (34) is true with $C \in IGA(f'([x]^0))$.

In case (c) of the Theorem 11, that is if $K[x]^{k_0} \cap [x]^{k_0} = \emptyset$ for some k_0 , we speak again of divergence (of the Krawczyk method). Similar as for the interval Newton method k_0 is small if the diameter of $[x]^0$ is small. This will be demonstrated subsequently under the following assumptions:

(i) $f'([x]^0)$ is nonsingular,

(ii) (48) holds,

(iii) $C = C_k$ varies with k such that it is the inverse of some matrix from $f'([x]^k)$.

Note that these assumptions certainly hold if the assumptions for (49) are fulfilled.

As for the interval Newton operator we write $K[x] = [k, \bar{k}]$. Now $K[x] \cap [x] = \emptyset$ if and only if

$$(\bar{x} - \underline{k})_{i_0} < 0 \tag{51}$$

or

$$(\bar{k} - \underline{x})_{i_0} < 0 \tag{52}$$

for at least one $i_0 \in \{1, 2, \dots, n\}$. (Compare with (37) and (38).)

We first prove that for K[x] defined by (43) we have the vector inequalities

$$\bar{x} - \underline{k} \leqslant \mathcal{O}(||d[x]||_{\infty}^2)e + Cf(\bar{x})$$
(53)

and

$$\bar{k} - \underline{x} \leq \mathcal{O}(||d[x]||_{\infty}^2)e - Cf(\underline{x}), \tag{54}$$

where again $e = (1, 1, \dots, 1)^{\mathrm{T}} \in \mathbb{R}^{n}$.

We prove (54). For $[x] \subseteq [x]^0$ let $f'([x]) = [\underline{F}', \overline{F}']$ and set $C = \hat{M}^{-1}$ with some matrix $\hat{M} \in f'([x])$. An easy computation shows that

$$I - Cf'([x]) = C[\hat{M} - \bar{F}', \hat{M} - \underline{F}'] \subseteq |C|[\underline{F}' - \bar{F}', \bar{F}' - \underline{F}'] \subseteq [-1, 1]\hat{C} df'([x]),$$

where \hat{C} is any upper bound for the set $\{|M^{-1}| | M \in f'([x]^0)\}$. Therefore

$$K[x] \subseteq m[x] - Cf(m[x]) + [-1,1]\hat{C} df'([x]) \cdot |[x] - m[x]|.$$

Hence,

$$\bar{k} - \underline{x} \le m[x] - \underline{x} - Cf(m[x]) + \hat{C}df'([x])d[x]$$

$$\leq \frac{1}{2}d[x] - Cf(m[x]) + O(||d[x]||_{\infty}^{2})e,$$

where we have used (48) and $m[x] \in [x]$.

Choosing x = m[x], $y = \underline{x}$ in (28) we obtain

$$f(m[x]) - f(\underline{x}) = J(\underline{x}, m[x])(m[x] - \underline{x})$$

It follows that

$$\bar{k} - \underline{x} \leq \frac{1}{2}d[x] - Cf(\underline{x}) - \frac{1}{2}CJ(\underline{x}, m[x])d[x] + O(||d[x]||_{\infty}^{2})e$$
$$= \frac{1}{2}(I - CJ(\underline{x}, m[x]))d[x] - Cf(\underline{x}) + O(||d[x]||_{\infty}^{2})e.$$

Since

$$I - CJ(\underline{x}, m[x]) = C(C^{-1} - J(\underline{x}, m[x])) \in \hat{C}(f'([x]) - f'([x])) = \hat{C}df'([x])$$

the assertion follows by applying (48).

The second inequality can be shown in the same manner, hence (53) and (54) are proved.

If $f(x) \neq 0$, $x \in [x]$ and d[x] is sufficiently small, then there exists an $i_0 \in \{1, 2, ..., n\}$ such that $(C f(x)) \neq 0$ (55)

$$(Cf(\underline{x}))_{i_0} \neq 0 \tag{55}$$

and

$$\operatorname{sign}\left(Cf(\bar{x})\right)_{i_0} = \operatorname{sign}\left(Cf(\underline{x})\right)_{i_0}.$$
(56)

This can be seen as follows: Since $\underline{x} \in [x]$ we have $f(\underline{x}) \neq 0$ and since C is nonsingular it follows that $Cf(\underline{x}) \neq 0$ and therefore $(Cf(\underline{x}))_{i_0} \neq 0$ for at least one $i_0 \in \{1, 2, ..., n\}$ which proves (55). Using again (28) with $x = \overline{x}$, $y = \underline{x}$ we get

$$f(\bar{x}) - f(\underline{x}) = J(\underline{x}, \bar{x})(\bar{x} - \underline{x}).$$

It follows

$$Cf(\bar{x}) = Cf(\underline{x}) + CJ(\underline{x},\bar{x})(\bar{x}-\underline{x}).$$

Since the second term on the right-hand side approaches zero if $d[x] \rightarrow 0$ we have (56) for sufficiently small diameter d[x].

Using (53), (54) together with (55) and (56) we can now show that for sufficiently small diameters of [x] the intersection $K[x] \cap [x]$ becomes empty. See the analogous conclusions for the interval Newton method using (41), (42) together with (39) and (40). By the same motivation as for the interval Newton method we denote this behavior as 'quadratic divergence' of the Krawczyk method.

Part (a) of the two preceding theorems can be used in a systematic manner for verifying the existence of a solution of a nonlinear system in an interval vector. Besides of the existence of a solution also componentwise errorbounds are delivered by such an interval vector. We are now going to discuss how such an interval vector can be constructed.

For a nonlinear mapping $f: D \subset \mathbb{R}^n \to \mathbb{R}^n$ we consider Newton's method

$$x^{k+1} = x^k - f'(x^k)^{-1} f(x^k), \quad k = 0, 1, \dots$$
(57)

The Newton–Kantorovich theorem gives sufficient conditions for the convergence of Newton's method starting at x^0 . Furthermore, it contains an error estimation. A simple discussion of this estimation in conjunction with the quadratic convergence property (36) which we have also proved (under mild additional assumptions) for the Krawczyk method will lead us to a test interval which can be computed using only iterates of Newton's method.

Theorem 12 (See Ortega and Rheinboldt, [71, Theorem 12.6.2]). Assume that $f:D \subseteq \mathbb{R}^n \to \mathbb{R}^n$ is differentiable in the ball $\{x \mid ||x - x^0||_{\infty} \leq r\}$ and that

$$||f'(x) - f'(y)||_{\infty} \leq L||x - y||_{\infty}$$

for all x, y from this ball. Suppose that $f'(x^0)^{-1}$ exists and that $||f'(x^0)^{-1}||_{\infty} \leq B_0$. Let

$$||x^{1} - x^{0}||_{\infty} = ||f'(x^{0})^{-1} \cdot f(x^{0})||_{\infty} = \eta_{0}$$

and assume that

$$h_0 = B_0 \eta_0 L \leqslant \frac{1}{2}, \quad r_0 = \frac{1 - \sqrt{1 - 2h_0}}{h_0} \eta_0 \leqslant r.$$

Then the Newton iterates are well defined, remain in the ball $\{x \mid ||x - x^0||_{\infty} \leq r_0\}$ and converge to a solution x^* of f(x) = 0 which is unique in $D \cap \{x \mid ||x - x^0||_{\infty} < r_1\}$ where

$$r_1 = \frac{1 + \sqrt{1 - 2h_0}}{h_0} \eta_0$$

provided $r \ge r_1$. Moreover the error estimate

$$||x^* - x^k||_{\infty} \leq \frac{1}{2^{k-1}} (2h_0)^{2^k - 1} \eta_0, \quad k \ge 0$$
(58)

holds.

Since $h_0 \leq \frac{1}{2}$, the error estimate (58) (for k = 0, 1 and the ∞ -norm) leads to

$$\begin{aligned} ||x^* - x^0||_{\infty} &\leq 2\eta_0 = 2||x^1 - x^0||_{\infty}, \\ ||x^* - x^1||_{\infty} &\leq 2h_0\eta_0 \leq \eta_0 = ||x^1 - x^0||_{\infty}. \end{aligned}$$

This suggests a simple construction of an interval vector containing the solution x^* . If x^0 is close enough to the solution x^* then x^1 is much closer to x^* than x^0 since Newton's method is quadratically convergent. The same holds if we choose any vector ($\neq x^*$) from the ball $\{x \mid ||x - x^1||_{\infty} \leq \eta_0\}$ as starting vector for Newton's method. Because of (36) and since $x^* \in K[x]$ it is reasonable to assume that

$$K[x] = x^{1} - f'(x^{0})^{-1}f(x^{1}) + (I - f'(x^{0})^{-1}f'([x]))([x] - x^{1}) \subseteq [x]$$

for

$$[x] = \{x \mid ||x - x^{1}||_{\infty} \leq \eta_{0}\}.$$
(59)

The important point is that this test interval [x] can be computed without knowing B_0 and L. Of course all the preceding arguments are based on the assumption that the hypothesis of the Newton–Kantorovich theorem is satisfied, which may not be the case if x^0 is far away from x^* .

We try to overcome this difficulty by performing first a certain number of Newton steps until we are close enough to a solution x^* of f(x) = 0. Then we compute the interval (59) with x^{k+1} instead of x^1 . Using the Krawczyk operator we test whether this interval contains a solution. The question of when to terminate the Newton iteration is answered by the following considerations.

Our general assumption is that the Newton iterates are convergent to x^* . For ease of notation we set

$$[y] = x^{k+1} - f'(x^k)^{-1} f(x^{k+1}) + (I - f'(x^k)^{-1} f([x]))([x] - x^{k+1})$$

where

$$[x] = \{x \in \mathbb{R}^n \mid ||x^{k+1} - x||_{\infty} \le \eta_k\},\$$

$$\eta_k = ||x^{k+1} - x^k||_{\infty}$$
(60)

for some fixed k. Our goal is to terminate Newton's method as soon as

$$\frac{||d[y]||_{\infty}}{||x^{k+1}||_{\infty}} \le \text{eps}$$
(61)

holds where eps is the machine precision of the floating point system. If $x^* \in [x]$ then $x^* \in [y]$ so that for any $y \in [y]$ we have

$$\frac{||x^*-y||_{\infty}}{||x^*||_{\infty}} \leqslant \frac{||d[y]||_{\infty}}{||x^*||_{\infty}}.$$

Since $||x^*||_{\infty}$ differs only slightly from $||x^{k+1}||_{\infty}$ if x^{k+1} is near x^* , condition (61) guarantees that the relative error with which any $y \in [y]$ approximates x^* is close to machine precision. Using (35) it can be shown that

$$||df'([x])||_{\infty} \leq \hat{L}||d[x]||_{\infty}$$

and

$$||d[y]||_{\infty} \leq ||f'(x^k)^{-1}||_{\infty} \tilde{L}||d[x]||_{\infty}^2,$$

where $\tilde{L} = \max{\{\hat{L}, L\}}$, and since $||d[x]||_{\infty} = 2\eta_k$ the inequality (61) holds if

$$4\frac{||f'(x^k)^{-1}||_{\infty}\tilde{L}\eta_k^2}{||x^{k+1}||_{\infty}} \leq \text{eps}$$
(62)

is true.

From Newton's method we have

$$x^{k+1} - x^k = f'(x^k)^{-1} \{ f(x^k) - f(x^{k-1}) - f'(x^{k-1})(x^k - x^{k-1}) \}$$

and by 3.2.12 in [71] it follows that

$$\eta_k \leq \frac{1}{2} ||f'(x^k)^{-1}||_{\infty} \tilde{L} \eta_{k-1}^2.$$

Replacing the inequality sign by equality in this relation and eliminating $||f'(x^k)^{-1}||_{\infty}\tilde{L}$ in (62) we get the following stopping criterion for Newton's method:

$$\frac{8\eta_k^3}{||x^{k+1}||_{\infty}\eta_{k-1}^2} \leqslant \text{eps.}$$
(63)

Of course, this is not a mathematical proof that if (63) is satisfied then the interval [y] constructed as above will contain x^* and that the vectors in [y] will approximate x^* with a relative error close to eps. However as has been shown in [11] the test based on the stopping criterion (63) works extremely well in practice.

Some of the ideas of this section have been generalized to nonsmooth mappings by Chen [24].

Nonlinear interval systems, i.e., systems of nonlinear equations with parameter-dependent input data, have been considered, e.g., in [58].

A very important point is also the fact that for the verification of solutions of nonlinear systems one can often replace the interval arithmetic evaluation of the Jacobian by an interval arithmetic enclosure of the slope-matrix of f. In this connection slopes have first been considered in [5], see also [75].

5. Systems of linear equations

Given $[A] \in I(\mathbb{R}^{n \times n})$, $[b] \in I(\mathbb{R}^n)$ we want to characterize and to enclose the solution set

$$S = \{ x \in \mathbb{R}^n | \ Ax = b, \ A \in [A], \ b \in [b] \}$$
(64)

and the symmetric solution set

$$S_{\text{sym}} = \{ x \in \mathbb{R}^n | Ax = b, A = A^{\mathsf{T}} \in [A] = [A]^{\mathsf{T}}, b \in [b] \}.$$
(65)

These sets occur when dealing with systems of linear equations whose input data are afflicted with tolerances (cf., e.g. [13,69] or [84]). This is the case when data $\check{A} \in \mathbb{R}^{n \times n}$, $\check{b} \in \mathbb{R}^n$ are perturbed by errors caused, e.g., by measurements or by a conversion from decimal to binary digits on a computer. Assume that these errors are known to be bounded by some quantities $\Delta A \in \mathbb{R}^{n \times n}$ and $\Delta b \in \mathbb{R}^n$ with nonnegative entries. Then it seems reasonable to accept a vector \tilde{x} as the 'correct' solution of $\check{A}x = \check{b}$ if it is in fact the solution of a perturbed system $\tilde{A}x = \tilde{b}$ with

$$\tilde{A} \in [A] = [\check{A} - \Delta A, \check{A} + \Delta A], \qquad \tilde{b} \in [b] = [\check{b} - \Delta b, \check{b} + \Delta b].$$

The characterization of all such \tilde{x} led Oettli and Prager [72] to statements (a) and (b) of the following theorem.

Theorem 13. For $[A] \in I(\mathbb{R}^{n \times n})$, $[b] \in I(\mathbb{R}^{n})$ the following properties are equivalent: (a) $x \in S$; (b) $|\check{A}x - \check{b}| \leq \frac{1}{2}(d([A])|x| + d([b]))$; (c) $[A]x \cap [b] \neq \emptyset$; (d)

$$\left\{\begin{array}{l} \underline{b}_i - \sum_{j=1}^n a_{ij}^+ x_j \leq 0\\ -\overline{b}_i + \sum_{j=1}^n a_{ij}^- x_j \leq 0\end{array}\right\}, \quad i = 1, \dots, n,$$

where a_{ij}^- and a_{ij}^+ are determined by the equality

$$[\underline{a}_{ij}, \bar{a}_{ij}] = \begin{cases} [a_{ij}^-, a_{ij}^+] & \text{if } x_j \ge 0, \\ [a_{ij}^+, a_{ij}^-] & \text{if } x_j < 0. \end{cases}$$

The inequality in (b) relates the midpoint residual to the diameters of [A] and [b], (c) is a short interval version of (b) due to Beeck [22] and (d) characterizes S in each orthant as intersection of finitely many half spaces. This last property shows, in particular, that S cannot easily be described. Therefore, one often encloses S by an interval vector [x]. According to (26) such a vector can be computed, e.g., by the Gaussian algorithm performed with the interval data as in Section 4. It is an open question to find necessary and sufficient conditions for the feasibility of the Gaussian elimination process if [A] contains nondegenerate entries. For instance, IGA([A], [b]) exists if $\langle [A] \rangle$ is an M matrix as was shown in [4]. Other sufficient conditions can be found in [13,55,60]. See also the references there.

Iterative methods can also be used for enclosing S. Two simple ones are the interval Jacobi method

$$[x_i]^{k+1} = \left([b_i] - \sum_{\substack{j=1\\j\neq i}}^n [a_{ij}] [x_j]^k \right) \middle/ [a_{ii}], \quad i = 1, \dots, n$$
(66)

and the interval Gauss-Seidel method

$$[x_i]^{k+1} = \left([b_i] - \sum_{j=1}^{i-1} [a_{ij}] [x_j]^{k+1} - \sum_{j=i+1}^n [a_{ij}] [x_j]^k \right) / [a_{ii}], \quad i = 1, \dots, n$$
(67)

with $0 \notin [a_{ii}]$ for i = 1, ..., n. They can be modified by intersecting the right-hand sides of (66) and (67) with $[x_i]^k$ before assigning it to $[x_i]^{k+1}$.

Denote by [D], -[L] and -[U], respectively, the diagonal part, the strictly lower triangular part and the strictly upper triangular part of [A], respectively. Then [A] = [D] - [L] - [U], and the unmodified methods can be written in the form

$$[x]^{k+1} = f([x]^k) \quad \text{with } f([x]) = \text{IGA}([M], [N][x] + [b]), \tag{68}$$

where [A] = [M] - [N] and where we assume that IGA([M]) exists. For [M] = [D] we recover the Jacobi method (66) and for [M] = [D] - [L] the Gauss–Seidel method (67). The following result holds for these two cases and for a slight generalization concerning the shape of [M]:

444

Theorem 14. Let $[A] = [M] - [N] \in I(\mathbb{R}^{n \times n})$, $[b] \in I(\mathbb{R}^n)$ with [M] being a nonsingular lower triangular interval matrix:

(a) Iteration (68) is equivalent to the iteration

$$[x_i]^{k+1} = \left([b_i] - \sum_{j=1}^{i-1} [m_{ij}] [x_j]^{k+1} + \sum_{j=1}^n [n_{ij}] [x_j]^k \right) / [m_{ii}], \quad i = 1, \dots, n.$$
(69)

- (b) Iteration (68) is convergent to some limit [x]* ∈ I(ℝⁿ) (i.e., each sequence {[x]^k}_{k=0}[∞] of iterates defined by (68) is convergent to [x]*) if and only if ρ(⟨[M]⟩⁻¹|[N]|) < 1. In this case S ⊆ [x]*.
- (c) If [A] and [M] are M matrices and if $\underline{N} \ge O$ then $\rho(\langle [M] \rangle^{-1} | [N] |) = \rho(\underline{M}^{-1} \overline{N}) < 1$ and $[x]^*$ from (b) is the hull of S.
- (d) Let $[x] \in I(\mathbb{R}^n)$. If f([x]) from (68) satisfies $(f([x]))_i \subset [x_i]$ for i = 1, ..., n, then $\rho(\langle [M] \rangle^{-1} |[N]|) < 1$.

Proof. (a) follows by induction with respect to *i* taking into account that for lower triangular matrices the *i*th elimination step of the Gaussian algorithm changes only the *i*th column of [A].

(b) Let $P = \langle [M] \rangle^{-1} | [N] |$. Since [M] is triangular, $\langle [M] \rangle$ is an M matrix, hence $P \ge O$. ' \Rightarrow ': From (69) we get

$$d[x_i]^{k+1} \ge \left(\sum_{j=1}^{i-1} |m_{ij}| d[x_j]^{k+1} + \sum_{j=1}^{n} |n_{ij}| d[x_j]^k\right) / \langle [m_{ii}] \rangle, \quad i = 1, \dots, n,$$
(70)

which is equivalent to $\langle [M] \rangle d[x]^{k+1} \ge |[N]| d[x]^k$. From this, $d[x]^{k+1} \ge Pd[x]^k$, and, by induction, $d[x]^k \ge P^k d[x]^0$ follow. Choose $[x]^0$ such that $d[x]^0$ is a Perron vector for P with $d[x_{i_0}]^* < d[x_{i_0}]^0$ for some index i_0 . If $\rho(P) \ge 1$ then

$$d[x_{i_0}]^k \ge \rho(P)^k d[x_{i_0}]^0 \ge d[x_{i_0}]^0 > d[x_{i_0}]^*$$

and $k \to \infty$ yields to a contradiction.

' \Leftarrow ': Let f([x]) = IGA([M], [N][x] + [b]). From (69) we get

$$q(f([x]), f([y]))_i \leq \frac{1}{\langle [m_{ii}] \rangle} \left(\sum_{j=1}^{i-1} |[m_{ij}]| q(f([x_j]), f([y_j])) + \sum_{j=1}^{n} |[n_{ij}]| q([x_j], [y_j]) \right),$$

$$i = 1, \dots, n,$$

whence $\langle [M] \rangle q(f([x]), f([y])) \leq |[N]| q([x], [y])$ and $q(f([x]), f([y])) \leq Pq([x], [y])$. Hence f is a P contraction, and Theorem 7 together with Remark 1 proves the convergence.

Let now (68) be convergent for all $[x]^0$ and choose $\tilde{x} \in S$. There are $\tilde{A} \in [A]$, $\tilde{b} \in [b]$, $\tilde{M} \in [M]$, $\tilde{N} \in [N]$ such that $\tilde{A}\tilde{x} = \tilde{b}$, $\tilde{A} = \tilde{M} - \tilde{N}$ and $\tilde{x} = \tilde{M}^{-1}(\tilde{N}\tilde{x} + \tilde{b})$. Then $\tilde{x} \in \text{IGA}([M], [N]\tilde{x} + [b])$. Start (68) with $[x]^0 = \tilde{x}$. Then $\tilde{x} \in [x]^k$ for k = 0, 1, ..., hence $\tilde{x} \in [x]^*$. This proves $S \subseteq [x]^*$.

(c) The assumptions imply that $\underline{A} = \underline{M} - \overline{N}$ is a regular splitting of \underline{A} and that $\underline{A}^{-1} \ge O$. Therefore, 2.4.17 in [71] guarantees $\rho(\langle [M] \rangle^{-1} | [N] |) = \rho(\underline{M}^{-1} \overline{N}) < 1$.

In order to prove the hull property let $[x]^*$ be the limit of (68), define

$$m_{ij}^* = \begin{cases} \underline{m}_{ij} & \text{if } \underline{x}_j^* \leq 0, \\ \overline{m}_{ij} & \text{if } \underline{x}_j^* > 0, \end{cases} \qquad n_{ij}^* = \begin{cases} \overline{n}_{ij} & \text{if } \underline{x}_j^* \leq 0, \\ \underline{n}_{ij} & \text{if } \underline{x}_j^* > 0 \end{cases}$$

and let $A^* = M^* - N^*$. Then $A^* \in [A]$, and from (69) with $k \to \infty$ we get $A^* \underline{x}^* = \underline{b}$, hence $\underline{x}^* \in S$. Analogously one can show that $\overline{x}^* \in S$.

(d) Replace $[x_j]^k$ by $[x_j]$ and $[x_i]^{k+1}$ by $f([x])_i$ in (70). Together with the assumption this yields to $Pd[x] \leq d f([x]) < d[x]$, and analogously to the proof of Theorem 11(a) we get $\rho(P) < 1$. \Box

For the Richardson splitting [A] = I - (I - [A]) parts of Theorem 14 were already stated and proved in [61]. Most of its present form can be found in [69, Chapters 4.4 and 4.5].

We now apply the Krawczyk operator (43) to the function Ax - b and replace $A \in \mathbb{R}^{n \times n}$, $b \in \mathbb{R}^n$ by $[A] \in I(\mathbb{R}^{n \times n})$, $[b] \in I(\mathbb{R}^n)$. Then we get the modified Krawczyk operator

$$K_{\text{mod}}[x] = m[x] + C([b] - [A]m[x]) + (I - C[A])([x] - m[x])$$
(71)

with some nonsingular matrix $C \in \mathbb{R}^{n \times n}$ and any vector m[x] from \mathbb{R}^n . For $K_{\text{mod}}[x]$ and for the iteration

$$[x]^{k+1} = K_{\text{mod}}[x]^k \cap [x]^k$$
(72)

with fixed C the following analogue of Theorem 11 holds.

Theorem 15. Let
$$[A] \in I(\mathbb{R}^{n \times n}), [b] \in I(\mathbb{R}^{n})$$
:
(a) If
 $\rho(|I - C[A]|) < 1,$ (73)

then [A] is nonsingular, i.e., each linear system Ax = b with $A \in [A]$ and $b \in [b]$ is uniquely solvable. If, in addition, $S \subseteq [x]^0$ then the sequence $\{[x]^k\}_{k=0}^{\infty}$ defined by (72) is well defined, $S \subseteq [x]^k$ and $\lim_{k\to\infty} [x]^k = [x]^* \supseteq S$. In particular, $\{[x]^k\}_{k=0}^{\infty}$ is monotonically decreasing. (b) If

$$K_{\text{mod}}[x] \subseteq [x] \tag{74}$$

for some $[x] \in I(\mathbb{R}^n)$ then each linear system Ax = b with $A \in [A]$ and $b \in [b]$ has a solution $x^* \in [x]$.

If (74) is slightly sharpened to

$$(K_{\text{mod}}[x])_i \subset [x_i] \quad for \ i = 1, \dots, n,$$

$$(75)$$

then $\rho(|I - C[A]|) < 1$, i.e., the properties in (a) hold with $S \subset [x]$. (c) If

$$|||I - C[A]|||_{\infty} < 1, \tag{76}$$

then the properties in (a) hold. In addition,

$$S \subseteq [\tilde{x}] = [\tilde{x} - \alpha e, \tilde{x} + \alpha e], \tag{77}$$

where

$$\alpha = \frac{|||C([b] - [A]\tilde{x})|||_{\infty}}{1 - |||I - C[A]|||_{\infty}}$$

Therefore, the second part of (a) holds for any $[x]^0 \supseteq [\tilde{x}]$.

Proof. (a) Can be proved via an analog of (50) and by using the representation

$$x^* = m[x] + C(b - Am[x]) + (I - CA)(x^* - m[x]) \in K_{\text{mod}}[x]$$
(78)

for $x^* = A^{-1}b, A \in [A], b \in [b]$.

(b) Is proved analogously to part (a) of Theorem 11.

(c) Since the assertion implies $\rho(|I - C[A]|) < 1$ all properties of (a) hold. Let $x^* \in S$. Then there are $A \in [A]$, $b \in [b]$ such that $Ax^* = b$. Hence

$$||x^* - \tilde{x}||_{\infty} = ||A^{-1}(b - A\tilde{x})||_{\infty} \leq ||\{I - (I - CA)\}^{-1}||_{\infty}||C(b - A\tilde{x})||_{\infty} \leq \alpha,$$

where we used the Neumann series for the last inequality. \Box

Remark 3. (a) As in Remark 2 it is not necessary to know whether C is nonsingular if (73), (75) or (76) hold. Either of these assumptions guarantees the nonsingularity of C.

(b) If (74) or (75) holds then $S \subseteq K_{\text{mod}}[x]$.

(c) If [A] and [b] are degenerate, i.e., $[A] \equiv A$, $[b] \equiv b$ then the assumption $\rho(|I - CA|) < 1$ in Theorem 15 implies

$$\lim_{k \to \infty} [x]^k = x^*,$$

where $Ax^* = b$.

Remark 3(b) leads to the question how good the enclosures are which one gets as iterates obtained by (72). The following result is due to Rump [82] and answers this question if (75) holds. To this end we define S_i as the projection of S to the *i*th coordinate axis, i.e.,

$$S_i = \{x_i \mid x \in S\} \subseteq \mathbb{R}.$$
(79)

For nonsingular [A] Cramer's rule shows that x_i depends continuously on $A \in [A]$ and $b \in [b]$. Since [A] and [b] are connected and compact, the sets S_i are compact intervals.

Theorem 16. Let $[A] \in I(\mathbb{R}^{n \times n})$, $[b] \in I(\mathbb{R}^n)$, S_i as in (79). Compute $K_{\text{mod}}[x]$ from (71) with any $m[x] = \tilde{x} \in \mathbb{R}^n$ and any nonsingular $C \in \mathbb{R}^{n \times n}$, and let

$$[z] = C([b] - [A]\tilde{x}), \quad [\delta] = (I - C[A])([x] - \tilde{x}).$$

If $(K_{\text{mod}}[x])_i \subset [x_i]$ for $i = 1, ..., n$ then
 $\tilde{x}_i + \underline{z}_i + \underline{\delta}_i \leqslant \min S_i \leqslant \tilde{x}_i + \underline{z}_i + \overline{\delta}_i,$

$$(80)$$
 $\tilde{x}_i + \overline{z}_i + \underline{\delta}_i \leqslant \max S_i \leqslant \tilde{x}_i + \overline{z}_i + \overline{\delta}_i,$

$$(81)$$

i.e., $d[\delta]$ is a measure for the overestimation of S by $K_{mod}[x]$.

Proof. The left inequality of (80) and the right inequality of (81) follow directly from Remark 3(b). In order to prove the two remaining inequalities note that the interval $[z_i]$ is the interval arithmetic evaluation of the function $f : \mathbb{R}^{n^2+n} \to \mathbb{R}$ which is defined by $f(A,b) = (C(b-A\tilde{x}))_i$. In f(A,b) each variable occurs only once. Therefore, Theorem 2 implies

$$f([A], [b]) = R(f; [A], [b]),$$
(82)

i.e., there are some $A^* \in [A]$, $b^* \in [b]$ such that $\underline{z}_i = f(A^*, b^*)$. From (78) for $x^* = (A^*)^{-1}b^* \in S$ and with $\delta^* = (I - CA^*)(x^* - \tilde{x})$ we get

 $\min S_i \leqslant x_i^* = \tilde{x}_i + \underline{z}_i + \delta_i^* \leqslant \tilde{x}_i + \underline{z}_i + \bar{\delta}_i,$

which shows the right inequality of (80). The left inequality of (81) is proved analogously. \Box

Remark 4. Let (75) holds with *C* being the inverse of the center of [*A*] and let \tilde{x} be a good approximation of some element of *S*. Assume that d[A], d[b] are small and that (75) holds for some [*x*] with $m[x] = \tilde{x} \in [x]$. Then $d[z] = |C|(d[b] + d[A]\tilde{x})$ can be expected to be small and from

$$[\delta] = |C|[-\frac{1}{2}d[A], \frac{1}{2}d[A]]([x] - \tilde{x}) = |C|[-\frac{1}{2}d[A], \frac{1}{2}d[A]]|[x] - \tilde{x}|,$$

we get $d[\delta] \leq |C|d[A]d[x]$. Hence if d[x] is also small (which can be expected if some $A \in [A]$ is not ill-conditioned) then $d[\delta]$ is quadratically small, i.e., $d[\delta] \leq d[z]$. This indicates a small overestimation of *S* by $K_{\text{mod}}[x]$.

If, in fact, at least $d[\delta] \leq d[z]$ holds then $\underline{z} + \overline{\delta} \leq \overline{z} + \underline{\delta}$ and $[x]^{\text{int}} = [\underline{x}^{\text{int}}, \overline{x}^{\text{int}}] = \tilde{x} + [\underline{z} + \overline{\delta}, \overline{z} + \underline{\delta}]$ is an interval vector which satisfies min $S_i \leq \underline{x}_i^{\text{int}} \leq \overline{x}_i^{\text{int}} \leq \max S_i$ for i = 1, ..., n. Such a vector is called an inner enclosure of S by Rump [84]. If an inner enclosure of S is known one can estimate the quality of an enclosure (in the set-theoretical sense) of S in a straightforward way. Inner enclosures and related topics are considered for instance in [84,87].

Now we address to the symmetric solution set S_{sym} from (65), i.e., we are interested in linear systems Ax = b with symmetric matrices $A \in [A] \in I(\mathbb{R}^{n \times n})$. For simplicity, we assume

$$[A] = [A]^{\mathrm{T}}.$$
(83)

Otherwise the subsequent results hold for the largest interval matrix which is contained in [A] and which has property (83).

Trivially, S_{sym} is a subset of S. Its shape is even more complicated than that of S: Curved boundaries can occur as the following theorem indicates.

Theorem 17. Let S_{sym} be defined for a given nonsingular interval matrix $[A] = [A]^T \in I(\mathbb{R}^{n \times n})$ and a given interval vector $[b] \in I(\mathbb{R}^n)$. Then for any closed orthant $O \subseteq \mathbb{R}^n$ the set $S_{sym} \cap O$ can be represented as the intersection of finitely many closed sets whose boundaries are quadrics or hyperplanes. These sets can be described by inequalities which result, e.g., from a Fourier–Motzkin elimination process.

The proof of this theorem can be found in [15], corresponding properties on classes of matrices with more general dependencies in [16,17]. For the Fourier–Motzkin elimination see, for instance, [85].

We want to enclose S_{sym} by an interval vector. Trivially, each method for enclosing S delivers such a vector. But the symmetric solution set often contains much less elements than S. Therefore, it is useful to look for methods which enclose S_{sym} but not necessarily S. Such a method is the interval Cholesky method which is defined by applying formally the formulas of the Cholesky method to the interval data $[A] = [A]^T$ and [b]. It produces an interval vector which we denote by ICh([A], [b]). In the algorithm the squares and the square roots are defined via (4). We assume that no division by an interval occurs which contains zero. If $\langle [A] \rangle$ is an *M* matrix with $\underline{a}_{ii} > 0$ for i = 1, ..., n then ICh([*A*],[*b*]) exists. This was shown in [19] where the interval version of the Cholesky method was introduced and studied in detail. See also [21].

Another method to enclose S_{sym} was considered by Jansson in [41]. He starts with a modification of $K_{mod}[x]$ from (71): Let

$$K_{\text{mod}}^{\text{sym}}[x] = m[x] + [z]^{\text{sym}} + (I - C[A])([x] - m[x]),$$
(84)

where $[z]^{\text{sym}} = ([z_i]^{\text{sym}}) \in I(\mathbb{R}^n)$ is defined by

$$[z_i]^{\text{sym}} = \sum_{j=1}^n c_{ij}([b_j] - [a_{jj}](m[x])_j) - \sum_{j=1}^n \sum_{l=1}^{j-1} (c_{ij}(m[x])_l + c_{il}(m[x])_j)[a]_{jl}.$$

Iterate analogously to (72) with $K_{\text{mod}}^{\text{sym}}[x]$ replacing $K_{\text{mod}}[x]$. Since by the same reasoning as above

$$[z_i]^{\text{sym}} = \{ (C(b - Am[x]))_i | A = A^{\text{T}} \in [A], b \in [b] \},\$$

Theorems 15 and 16 hold with S, [z] being replaced by S_{sym} , $[z]^{sym}$.

6. The algebraic eigenvalue problem and related topics

In this section we look for intervals $[\lambda] \in I(\mathbb{R})$ and interval vectors $[x] \in I(\mathbb{R}^n)$ such that $[\lambda]$ contains an eigenvalue $\lambda^* \in \mathbb{R}$ and [x] contains an associated eigenvector $x^* \in \mathbb{R}^n \setminus \{0\}$ for a given matrix $A \in \mathbb{R}^{n \times n}$. We restrict ourselves only to real eigenpairs. Complex ones have also been studied; cf. [56,57], e.g., for an overview.

We start with the mild nonlinear equation

$$f(x,\lambda) = \begin{pmatrix} Ax - \lambda x \\ x_{i_0} - \alpha \end{pmatrix} = 0,$$
(85)

where i_0 is a fixed index from $\{1, ..., n\}$ and $\alpha \neq 0$ is a constant. It is obvious that (x^*, λ^*) is a solution of (85) if and only if (x^*, λ^*) is an eigenpair of A with the normalization $x_{i_0}^* = \alpha$ of the eigenvector x^* . Expanding f into a Taylor series at an approximation $(\tilde{x}, \tilde{\lambda})$ of (x^*, λ^*) yields to

$$f(x,\lambda) = f(\tilde{x},\tilde{\lambda}) + \begin{pmatrix} A - \tilde{\lambda}I_n & -\tilde{x} \\ (e^{(i_0)})^{\mathrm{T}} & 0 \end{pmatrix} \begin{pmatrix} \Delta x \\ \Delta \lambda \end{pmatrix} - \begin{pmatrix} \Delta \lambda & \Delta x \\ 0 \end{pmatrix},$$
(86)

where $\Delta x = x - \tilde{x}$, $\Delta \lambda = \lambda - \tilde{\lambda}$, I_k is the $k \times k$ identity matrix and $e^{(i_0)}$ is the i_0 th column of I_n . Multiplying (86) by a preconditioning matrix $-C \in \mathbb{R}^{(n+1)\times(n+1)}$ and adding $((\Delta x)^T, \Delta \lambda)^T$ on both sides results in the fixed point equation

$$\begin{pmatrix} \Delta x \\ \Delta \lambda \end{pmatrix} = g(\Delta x, \Delta \lambda) = -Cf(\tilde{x}, \tilde{\lambda}) + \left\{ I_{n+1} - C \begin{pmatrix} A - \tilde{\lambda}I_n & -\tilde{x} - \Delta x \\ (e^{(i_0)})^{\mathrm{T}} & 0 \end{pmatrix} \right\} \begin{pmatrix} \Delta x \\ \Delta \lambda \end{pmatrix},$$
(87)

for the error $(\Delta x, \Delta \lambda) = (\Delta x^*, \Delta \lambda^*) = (x^* - \tilde{x}, \lambda^* - \tilde{\lambda})$ of an eigenpair (x^*, λ^*) . The following theorem is due to Rump [81].

Theorem 18. Let $A \in \mathbb{R}^{n \times n}$, $\tilde{\lambda} \in \mathbb{R}$, $\tilde{x} \in \mathbb{R}^n$, $C \in \mathbb{R}^{(n+1) \times (n+1)}$, and define g by (87). Let \tilde{x} be normalized by $\tilde{x}_{i_0} = \alpha \neq 0$. If g fulfills the inclusion

$$g([\Delta x], [\Delta \lambda]) \subseteq \operatorname{int}([\Delta x]^{\mathrm{T}}, [\Delta \lambda])^{\mathrm{T}}$$
(88)

then the following assertions hold:

(a) C is nonsingular.

- (b) There exists exactly one eigenvector $x^* \in \tilde{x} + [\Delta x]$ of A which is normalized by $x_{i_0}^* = \alpha$.
- (c) There exists exactly one eigenvalue $\lambda^* \in \tilde{\lambda} + [\Delta \lambda]$ of A.
- (d) $Ax^* = \lambda^* x^*$ with x^* from (b) and λ^* from (c).
- (e) The eigenvalue λ^* from (d) is geometric simple.
- (f) If $(\tilde{x}, \tilde{\lambda})$ is a sufficiently good approximation of the eigenpair (x^*, λ^*) from (d) then it can be guaranteed that λ^* is algebraic simple.

(g) If one starts the iteration

$$\begin{pmatrix} [\Delta x]^{k+1} \\ [\Delta \lambda]^{k+1} \end{pmatrix} = g([\Delta x]^k, [\Delta \lambda]^k), \quad k = 0, 1, \dots,$$
(89)

with

$$([\Delta x]^0, [\Delta \lambda]^0) = ([\Delta x], [\Delta \lambda])$$

from (88) then the iterates converge satisfying

$$([\Delta x]^{k+1}, [\Delta \lambda]^{k+1}) \subseteq ([\Delta x]^k, [\Delta \lambda]^k), \quad k = 0, 1, \dots$$

and

$$(x^*, \lambda^*) \in (\tilde{x}, \tilde{\lambda}) + ([\Delta x]^k, [\Delta \lambda]^k), \quad k = 0, 1, \dots$$

for the eigenpair (x^*, λ^*) from (d).

Interval quantities [x], $[\lambda]$ with (88) can be found, e.g., via ε -inflation; cf. [58] or [59]. Another way was indicated in [6] by the following theorem.

Theorem 19. With the notations of Theorem 18 define

$$\rho = \left\| C \begin{pmatrix} A\tilde{x} - \tilde{\lambda}\tilde{x} \\ 0 \end{pmatrix} \right\|_{\infty}, \quad \sigma = \left\| I_{n+1} - C \begin{pmatrix} A - \tilde{\lambda}I_n & -\tilde{x} \\ (e^{(i_0)})^{\mathrm{T}} & 0 \end{pmatrix} \right\|_{\infty}, \quad \tau = ||C||_{\infty}$$
(90)

and assume

$$\sigma < 1, \quad \Delta = (1 - \sigma)^2 - 4\rho\tau \ge 0. \tag{91}$$

Then the numbers

$$\beta^{-} = (1 - \sigma - \sqrt{\Delta})/(2\tau) = \frac{2\rho}{1 - \sigma + \sqrt{\Delta}},$$
$$\beta^{+} = (1 - \sigma + \sqrt{\Delta})/(2\tau)$$

are nonnegative, and the condition (88) of Theorem 18 is fulfilled for $([\Delta x]^T, [\Delta \lambda])^T = [-\beta, \beta]e \in I(\mathbb{R})^{(n+1)\times(n+1)}$ with arbitrary $\beta \in (\beta^-, \beta^+)$. In particular, all the assertions of that theorem hold.

If β is restricted to $[\beta^-, (\beta^- + \beta^+)/2)$ then the iterates of (89) converge to the error $(\frac{\Delta x^*}{\Delta x^*})$.

In [58] it is shown how (87) can be reduced to an *n*-dimensional problem which, originally, formed the starting point in [6]. It is also indicated there how (87) has to be modified if the normalization $x_{i_0}^* = \alpha$ is replaced by $||x^*||_2 = 1$.

A second method for enclosing eigenpairs starts with the centered form

$$f(x,\lambda) = f(\tilde{x},\tilde{\lambda}) + \begin{pmatrix} A - \tilde{\lambda}I_n & -\tilde{x} - \Delta x \\ (e^{(i_0)})^{\mathrm{T}} & 0 \end{pmatrix} \begin{pmatrix} \Delta x \\ \Delta \lambda \end{pmatrix}.$$

It is obvious that the subdivision principle discussed in Section 3 can be applied to any initial domain $([x]^0, [\lambda]^0)$ chosen by the user. The crucial problem remains to verify that $0 \in f([\hat{x}], [\hat{\lambda}])$ yields to $f(x^*, \lambda^*) = 0$ in a subdomain $([\hat{x}], [\hat{\lambda}]) \subseteq ([x]^0, [\lambda]^0)$.

A third method is due to H. Behnke and F. Goerisch. It assumes A to be symmetric and is based on a complementary variational principle. For details see, e.g., [23, Section 6], and the references there.

Symmetric matrices can also be handled by an access due to Lohner [54]. First A is reduced to nearly diagonal form using Jacobi rotations and a sort of staggered correction. Finally Gershgorin's theorem is applied in order to obtain bounds for the eigenvalues. A theorem due to Wilkinson allows the enclosure of eigenvectors.

There is no problem to generalize the ideas above to the generalized eigenvalue problem $Ax = \lambda Bx$, $x \neq 0$, $B \in \mathbb{R}^{n \times n}$ nonsingular. The analogue of (85) reads

$$f(x,\lambda) = \begin{pmatrix} Ax - \lambda Bx \\ x_{i_0} - \alpha \end{pmatrix} = 0.$$

In a similar way one can treat the singular value problem for a given $m \times n$ matrix A with $m \ge n$. Here, we look for orthogonal matrices $U \in \mathbb{R}^{n \times n}$, $V \in \mathbb{R}^{m \times m}$ and for a diagonal matrix $\Sigma = \text{diag}(\sigma_1, \ldots, \sigma_r, \ldots, \sigma_n) \in \mathbb{R}^{m \times n}$ with the singular values $\sigma_1 \ge \sigma_2 \ge \cdots \ge \sigma_r > \sigma_{r+1} = 0 = \cdots = \sigma_n$, r = rank(A), such that $A = V\Sigma U^T$. One starts with

$$f(u,v,\sigma) = \begin{pmatrix} Au - \sigma v \\ A^{\mathrm{T}}v - \sigma u \\ u^{\mathrm{T}}u - 1 \end{pmatrix} \quad \text{or with} \quad f(u,v,\sigma,\sigma') = \begin{pmatrix} Au - \sigma v \\ A^{\mathrm{T}}v - \sigma' u \\ u^{\mathrm{T}}u - 1 \\ v^{\mathrm{T}}v - 1 \end{pmatrix}.$$

In the first case a zero of f satisfies $v^{T}v = 1$, in the second one gets $\sigma = \sigma'$. In either of the cases u is a column of U, v a corresponding column of V and σ a singular value of A associated with u and v. For details, additional remarks and references to further methods for verifying and enclosing singular values see [7,57].

We also mention verification methods in [14] for generalized singular values (c^*, s^*) of a given matrix pair (A, B), $A \in \mathbb{R}^{p \times n}$, $B \in \mathbb{R}^{q \times n}$, which are defined as the zeros of the function $f(c, s) = \det(s^2 A^T A - c^2 B^T B)$ restricted to $c, s \ge 0, c^2 + s^2 = 1$. For applications of generalized singular values see [33].

The methods and results of the Sections 4-6 can be combined in order to study the following inverse eigenvalue problem:

Given n+1 symmetric matrices $A_i \in \mathbb{R}^{n \times n}$, $i=0,1,\ldots,n$. Find *n* real numbers c_i^* , $i=1,\ldots,n$, such that the matrix $A(c) = A_0 + \sum_{i=1}^n c_i A_i$, $c = (c_i) \in \mathbb{R}^n$, has for $c = c^* = (c_i^*)$ prescribed eigenvalues $\lambda_1^* < \lambda_2^* < \cdots < \lambda_n^*$. (92)

Here one starts with the function $f(c) = \lambda(c) - \lambda^* \in \mathbb{R}^n$, *c* sufficiently close to c^* , where the components $\lambda_i(c)$ of $\lambda(c)$ are the eigenvalues of A(c) ordered increasingly, and where $\lambda^* = (\lambda_i^*)$ is defined with (92). One can show that the equation for Newton's method reads

$$(x^{i}(c^{k}))^{T}A_{j}(x^{i}(c^{k}))(c^{k+1}-c^{k}) = -(\lambda(c^{k})-\lambda^{*});$$
(93)

 $x^{i}(c^{k})$ are the eigenvectors of $A(c^{k})$ associated with the eigenvalues $\lambda_{i}(c^{k})$ and normalized by $x^{i}(c^{k})^{T}x^{i}(c^{k}) = 1$, $\operatorname{sign}(x_{i_{0}}^{i}(c^{k})) = 1$ for some fixed index $i_{0} \in \{1, \ldots, n\}$.

In a first step approximations of $x^i(c^k)$, $\lambda_i(c^k)$ are computed for i = 1, ..., n. With these values Eq. (93) is formed and solved. This is done for k = 0, 1, ... up to some k_0 . In a second step the verification process is performed using the interval Newton method and results from Section 6 which are generalized from point matrices to interval matrices. For details see [10,20] or [57].

7. Ordinary differential equations

Many contributions to verification numerics refer to initial value problems for ordinary differential equations

$$y' = f(y), \tag{94}$$

$$y(x_0) = y^0,$$
 (95)

where we assume that $f:D \subseteq \mathbb{R}^n \to \mathbb{R}^n$ is sufficiently smooth and that (94) has a unique solution in some given interval $[x_0, x_0 + T]$ for any initial value $y^0 \in [y^0] \subseteq D$. For ease of presentation we choose (94) to be autonomous. This is not a severe restriction since any nonautonomous initial value problem can be reduced to an autonomous one by introducing the additional component $y_{n+1}=x$, the additional differential equation $y'_{n+1}=1$ and the additional initial value $y_{n+1}(x_0)=y^0_{n+1}=x_0$. We shall use a grid $x_0 < x_1 < \cdots < x_k < \cdots < x_K = x_0 + T$ with grid points x_k and stepsizes $h_k = x_{k+1} - x_k$ to be determined later on, and we shall consider (94) with initial values $y(x_k)$ from some intermediate interval vectors $[y^k]$. To this end we introduce the set

$$y(x;x_k,[y^k]) = \{y(x) \mid y' = f(y), \ y(x_k) \in [y^k]\}$$
(96)

of all solutions of (94) with initial values in $[y^k]$. In the sequel, we shall need the following auxiliary result.

Theorem 20. If $[\tilde{y}] + [0,h]f([\hat{y}]) \subseteq [\hat{y}]$ for f from (94) and some h > 0, $[\tilde{y}] \subseteq [\hat{y}] \subseteq D$, then $y(x;\tilde{x}, [\tilde{y}]) \subseteq [\hat{y}]$ for all $x \in [\tilde{x}, \tilde{x} + h]$.

Proof. For fixed $\tilde{y}^0 \in [\tilde{y}]$ apply Banach's fixed point theorem to the Picard–Lindelöf operator $(Tu)(x) = \tilde{y}^0 + \int_{\tilde{x}}^x f(u(t)) dt$, to the set $U = \{u \mid u \in C^0[\tilde{x}, \tilde{x} + h] \text{ and } u(x) \in [\hat{y}] \text{ for } x \in [\tilde{x}, \tilde{x} + h] \}$ and to the metric $||u||_{\alpha} = \max_{\tilde{x} \leq x \leq \tilde{x} + h} \{e^{-\alpha(x-\tilde{x})} \mid ||u(x)||_{\infty}\}$ with any $\alpha > |||\partial f([\hat{y}])/\partial y|||_{\infty}$. \Box

One of the most popular methods for verifying and enclosing solutions of initial value problems is known as interval Taylor series method. It goes back to R.E. Moore and was modified in various ways – cf., for instance, [30,53], and overviews in [26,66,80]. In order to describe this method we assume that we know the grid point $x_k < x_K$ and an enclosure $[y^k]$ of $y(x_k; x_0, [y^0])$. Such an enclosure is given for k = 0. The method consists of two major steps:

In the first step a new stepsize $h_k > 0$ and a rough a priori enclosure $[\hat{y}^k]$ is computed such that

$$y(x;x_k,[y^k]) \subseteq [\hat{y}^k] \quad \text{for all } x \in [x_k, x_k + h_k].$$

$$(97)$$

To this end let $[\hat{y}^k]$ be any vector which contains $[y^k]$ in its interior and choose $h_k > 0$ so small that $[y^k] + [0, h_k] f([\hat{y}^k]) \subseteq [\hat{y}^k]$. Then (97) is guaranteed by Theorem 20. With h_k we know $x_{k+1} = x_k + h_k$, and from (97) with $x = x_{k+1}$ we see that $[\hat{y}^k]$ is a candidate for $[y^{k+1}]$.

In the second step of the method this candidate is improved in the following way: consider any particular solution y^* of (94) with $y^*(x_k) \in [y^k]$. Using (94) and the Taylor expansion of y^* at x_k we get for a fixed $p \in \mathbb{N}$ and $h = x - x_k$

$$y^{*}(x) = \psi(h, y^{*}(x_{k})) + r_{p}(h, y^{*})$$
(98)

with

$$\psi(h, y) = y + \sum_{j=1}^{p} h^{j} f^{[j]}(y), \quad f^{[1]} = f, \quad f^{[j]} = \frac{1}{j} (f^{[j-1]})' = \frac{1}{j} \frac{\partial f^{[j-1]}}{\partial y} f \text{ for } j \ge 2$$

and with the remainder term $r_p(h, y^*) \in h^{p+1}f^{[p+1]}([\hat{y}^k])$. Throughout this section we assume that the Taylor coefficients $f^{[j]}(y^*(x_k))$ exist. They can be computed recursively by means of automatic differentiation which is described, e.g., in [34] or [76]. Obviously,

$$y(x;x_0,[y^0]) \subseteq y(x;x_k,[y^k]) \subseteq \psi(h,[y^k]) + h^{p+1} f^{[p+1]}([\hat{y}^k]) \quad \text{for } x_k \leqslant x \leqslant x_{k+1}.$$
(99)

By virtue of $d\psi(h_k, [y^k]) \ge d[y^k]$ the right expression in (99) with $h = h_k$ seems not yet to be suited as a good candidate for $[y^{k+1}]$ since its diameter dominates $d[y^k]$. Therefore, we represent $\psi(h, y)$ as centered form

$$\psi(h, y) = \psi(h, \tilde{y}^k) + \left\{ I + \sum_{j=1}^p h^j J(y, \tilde{y}^k; f^{[j]}) \right\} (y - \tilde{y}^k)$$
(100)

$$\in \psi(h, \tilde{y}^k) + \left\{ I + \sum_{j=1}^p h^j \frac{\partial f^{[j]}([y^k])}{\partial y} \right\} ([y^k] - \tilde{y}^k), \tag{101}$$

where $y, \tilde{y}^k \in [y^k]$ and where J(y,z; f) is defined as J(y,z) in (29) using the third argument as underlying function. With y^* as in (98) and

$$S_k^* = I + \sum_{j=1}^p h_k^j J(y^*, \tilde{y}^k; f^{[j]}),$$
(102)

$$[S_k] = I + \sum_{j=1}^p h_k^j \frac{\partial f^{[j]}([y^k])}{\partial y},\tag{103}$$

$$[\tilde{y}^{k+1}] = \psi(h_k, \tilde{y}^k) + h_k^{p+1} f^{[p+1]}([\hat{y}^k])$$
(104)

for $k = 0, 1, \dots, K - 1$ we therefore get

$$y^{*}(x_{k+1}) = \psi(h_{k}, \tilde{y}^{k}) + r_{p}(h_{k}, y^{*}) + S_{k}^{*}(y^{*}(x_{k}) - \tilde{y}^{k})$$
(105)

$$\in [\tilde{y}^{k+1}] + [S_k]([y^k] - \tilde{y}^k).$$
(106)

The partial derivatives in (101) and (103) can again be computed using automatic differentiation or by differentiating the code list of $f^{[j]}$. Formula (105) represents the basis for most variants of the interval Taylor series method as long as they differ in their second step. Obviously,

$$y(x_{k+1}; x_0, [y^0]) \subseteq y(x_{k+1}; x_k, [y^k]) \subseteq [\tilde{y}^{k+1}] + [S_k]([y^k] - \tilde{y}^k),$$
(107)

so that the right expression is a candidate for $[y^{k+1}]$, this time with $d[y^{k+1}] \leq d[y^k]$ being possible. The successive construction of $[y^{k+1}]$ via (106) is called mean value method. Since $0 \in [S_k]([y^k] \tilde{y}^{k}$), we get $[\tilde{y}^{k+1}] \subseteq [y^{k+1}]$. Therefore, we can assume for the succeeding interval $[x_{k+1}, x_{k+2}]$ that $\tilde{y}^{k+1} \in [y^{k+1}]$ in (100) is chosen from $[\tilde{y}^{k+1}]$ – preferably its midpoint – which justifies our notation.

Unfortunately, $y(x_{k+1}; x_k, [y^k])$ is not necessarily an interval vector. Therefore, $[y^{k+1}]$ can overestimate this set and, consequently, $y(x_{k+1}; x_0, \lfloor y^0 \rfloor)$. This phenomenon which occurs at each grid point $x_k, k > 0$, is called wrapping effect. Its existence is an intrinsic feature of interval arithmetic and does not depend on the particular method. Its size, however, is strongly influenced by the choice of the method. In order to reduce this size the original mean value method often has to be modified. If $h_k > 0$ is small and p is large one can expect that the second summand $[S_k]([y^k] - \tilde{y}^k)$ in (106) contributes most to the wrapping effect. It can be influenced by preconditioning with a regular matrix $A_k \in \mathbb{R}^{n \times n}$ which yields to the following variant of the mean value method:

• Choose $\tilde{v}^0 \in [v^0]$ and let $[r^0] = [v^0] - \tilde{v}^0$, $A_0 = I \in \mathbb{R}^{n \times n}$.

For $k = 0, 1, \dots, K - 1$ do the following steps:

- Compute [S_k], [ỹ^{k+1}] as in (103), (104).
 Choose ỹ^{k+1} ∈ [ỹ^{k+1}].
- Choose $A_{k+1} \in \mathbb{R}^{n \times n}$ (regular) as described below.
- Compute

$$[r^{k+1}] = \{A_{k+1}^{-1}([S_k]A_k)\}[r^k] + A_{k+1}^{-1}([\tilde{y}^{k+1}] - \tilde{y}^{k+1}),$$
(108)

$$[y^{k+1}] = [\tilde{y}^{k+1}] + ([S_k]A_k)[r^k].$$
(109)

Before we consider particular choices of matrices A_k we prove an analogue of (107).

Theorem 21. Let \tilde{y}^k , $[\tilde{y}^k]$, $[r^k]$, A_k be defined for $k = 0, 1, \dots, K$ as in the preceding variant of the mean value method and let, formally, $x_{-1} = x_0$, $[y^{-1}] = [y^0]$. Then for $k = 0, 1, \dots, K$ we get

$$y(x_k; x_{k-1}, [y^{k-1}]) \subseteq [y^k],$$
(110)

$$A_{k}^{-1}(y^{*}(x_{k}) - \tilde{y}^{k}) \in [r^{k}] \quad for \ any \ solution \ y^{*} \ of \ (94) \ with \ y^{*}(x_{k-1}) \in [y^{k-1}].$$
(111)

Proof. The assertion is true for k = 0 by the definition of x_{-1} , $[y^{-1}]$ and by $A_0 = I$. Let it hold for some k < K and let y^* be a solution of (94) with $y^*(x_k) \in [y^k]$. From (105), (111) and (109) we get

$$y^{*}(x_{k+1}) \in [\tilde{y}^{k+1}] + S_{k}^{*}(y^{*}(x_{k}) - \tilde{y}^{k}) = [\tilde{y}^{k+1}] + (S_{k}^{*}A_{k})\{A_{k}^{-1}(y^{*}(x_{k}) - \tilde{y}^{k})\}$$
(112)

$$\subseteq [\tilde{y}^{k+1}] + ([S_k]A_k)[r^k] = [y^{k+1}],$$
(113)

hence (110) follows for k + 1. Since (112) implies $y^*(x_{k+1}) - \tilde{y}^{k+1} \in [\tilde{y}^{k+1}] - \tilde{y}^{k+1} + S_k^*(y^*(x_k) - \tilde{y}^k)$ we obtain

$$\begin{aligned} A_{k+1}^{-1}(y^*(x_{k+1}) - \tilde{y}^{k+1}) &\in A_{k+1}^{-1}([\tilde{y}^{k+1}] - \tilde{y}^{k+1}) + (A_{k+1}^{-1}S_k^*A_k)\{A_k^{-1}(y^*(x_k) - \tilde{y}^k)\} \\ &\subseteq A_{k+1}^{-1}([\tilde{y}^{k+1}] - \tilde{y}^{k+1}) + (A_{k+1}^{-1}[S_k]A_k)[r^k] = [r^{k+1}], \end{aligned}$$

where we used (111) and (108). \Box

An easy induction shows that one can retrieve the mean value method from its variant above if $A_k = I$ for k = 0, 1, ..., K.

If $A_{k+1} \in [S_k]A_k$ then $I \in A_{k+1}^{-1}([S_k]A_k)$, and $(A_{k+1}^{-1}[S_k]A_k)[r^k] \approx [r^k]$ can be expected if A_k is not ill-conditioned (cf. [66, p. 32]). Therefore, the wrapping effect should not lead to large overestimations in this case. Unfortunately, A_k is not always well-conditioned. So, other choices for A_k become important. R. Lohner starts in [53] with $\tilde{A}_{k+1} \in [S_k]A_k$ and performs a *QR*-decomposition of \tilde{A}_{k+1} (eventually after having permuted the columns of this matrix), i.e., $\tilde{A}_{k+1} = Q_{k+1}R_{k+1}$. Then he chooses $A_{k+1} = Q_{k+1}$ which effects a rotation of the coordinate system. For details cf. [53] or [66].

We also mention variants due to Eijgenraam [30] and Rihm [80] and Lohner's implementation AWA. For further reading we recommend [66] in which an interval Hermite–Obreschkoff method is considered, and [67] in which an enclosure method for the solution of linear ODEs with polynomial coefficients is given.

Based on the preceding ideas boundary value problems can be handled via the well-known shooting method as it was done in [53].

The stability of the Orr–Sommerfeld equation for different parameters was investigated in [51] by enclosure methods.

ODEs are closely related to integral equations. Therefore, it is interesting to ask for verified enclosures of such equations and of definite integrals. Due to space limit, however, we must refer the reader to the literature, for instance to [25,32,43] and to various contributions in [1].

8. Partial differential equations

Like the theory of partial differential equations the verification methods in this field are very heterogeneous. As in many cases in the previous sections they are mostly based on fixed point theorems and on particular function spaces. In order to give a taste of some ideas we outline a method due to Plum [74] which applies for second order elliptic boundary value problems of the form

$$-\Delta u + F(x, u, \nabla u) = 0 \quad \text{in } \Omega, \tag{114}$$

$$B[u] = 0 \qquad \text{on } \partial\Omega, \tag{115}$$

where $\Omega \subseteq \mathbb{R}^n$, $n \in \{2,3\}$, is a bounded domain whose boundary $\partial \Omega$ is at least Lipschitz continuous. The boundary operator *B* is defined by

$$B[u] = \begin{cases} u & \text{on } \Gamma_0, \\ \frac{\partial u}{\partial v} = v \cdot \nabla u & \text{on } \partial \Omega \setminus \Gamma_0 \end{cases}$$

with $\Gamma_0 \subseteq \partial \Omega$ being closed and with v denoting the unit outward normal vector. The function F is given by $F:\overline{\Omega}\times\mathbb{R}\times\mathbb{R}^n\to\mathbb{R}$ with $|F(x,y,z)|\leqslant C(1+||z||_2^2)$ for some $C\geqslant 0$ and all $x\in\overline{\Omega}$, $y \in \mathbb{R}, |y| \leq \alpha, z \in \mathbb{R}^n$. We assume that F and its derivatives $F_y = \partial F/\partial y, F_z = (\partial F/\partial z_1, \dots, \partial F/\partial z_n)^T$, are continuous.

In view of the theory for (114) we assume that for some $\sigma \in \mathbb{R}$ and each $r \in L^2(\Omega)$ (= set of square integrable functions) the boundary value problem $-\Delta u + \sigma u = r$ in Ω is uniquely solvable in $H_B^2 = \operatorname{cl}\{u \in C^2(\overline{\Omega}) \mid B[u] = 0 \text{ on } \partial\Omega\}$ where 'cl' means the closure in the Sobolev space $H^2(\Omega)$.

We start with a function $\omega \in H^2_B(\Omega)$ which can be thought to be an approximation of a solution u^* of (114), (115), although – at the moment – we do not know whether such a solution exists.

We will apply the operator $L: H^2_B(\Omega) \to L^2(\Omega)$ given by

$$L[u] = -\Delta u + b \cdot \nabla u + cu, \quad b = F_z(\cdot, \omega, \nabla \omega), \ c = F_y(\cdot, \omega, \nabla \omega).$$
(116)

In order to guarantee the invertibility of L needed later on we assume $\nabla \omega \in (L^{\infty}(\Omega))^n$ and we have to check numerically that all eigenvalues of L on $H^2_B(\Omega)$ are nonzero. In addition, we suppose that, for some Banach space $X \supseteq H^2_R(\Omega)$ with some norm $|| \cdot ||_X$:

(a) the function

$$\Phi:\begin{cases} X & \to & L^2(\Omega), \\ u & \mapsto & b \cdot \nabla u + cu - F(\cdot, u, \nabla u) \end{cases}$$
(117)

is continuous, bounded on bounded sets, and Fréchet differentiable at ω with $\Phi'(\omega) = 0$, (b) the imbedding $H^2_{\mathcal{B}}(\Omega) \hookrightarrow X$ is compact.

As fixed point operator we choose the simplified Newton operator

$$Tu = u - \mathscr{F}'(\omega)^{-1} \mathscr{F}(u) \tag{118}$$

with $\mathscr{F}(u) = -\Delta u + F(\cdot, u, \nabla u)$, with the Fréchet derivative \mathscr{F}' of \mathscr{F} and with ω as above. Since $\mathcal{F}'(\omega) = L$ and $-\Delta u = L[u] - b \cdot \nabla u - cu$ we obtain

$$Tu = u - L^{-1}[-\Delta u + F(\cdot, u, \nabla u)] = L^{-1}[b \cdot \nabla u + cu - F(\cdot, u, \nabla u)] = L^{-1}[\Phi(u)].$$
(119)

Due to our assumptions it can be shown that $T: X \to X$ is continuous, compact and Fréchet differentiable at ω with $T'(\omega) = 0$. If we can find some closed, bounded, convex function set $U \subseteq X$ such that

$$TU \subseteq U, \tag{120}$$

then Schauder's fixed point theorem guarantees the existence of some fixed point $u^* \in U$ of T which, by virtue of (119), is a solution of (114), (115). In order to construct U we first apply a shift $u \mapsto v = u - \omega$ which yields to a set $V = U - \omega$ and which emphasizes the approximative character of ω . Moreover, it follows the lines of centered forms which we exploited successfully already several times. From $u^* = Tu^*$ and $v^* = u^* - \omega \in X$ we get

$$v^* = T\omega - \omega + \{T(\omega + v^*) - T\omega\} = L^{-1}[-\delta[\omega] + \varphi(v^*)]$$
(121)

with

$$\delta[\omega] = -\Delta\omega + F(\cdot, \omega, \nabla\omega),$$

$$\varphi(v) = -\{F(\cdot, \omega + v, \nabla\omega + \nabla v) - F(\cdot, \omega, \nabla\omega) - b \cdot \nabla v - cv\}.$$
(122)

If we replace (120) by

$$L^{-1}[-\delta[\omega] + \varphi(V)] \subseteq V, \tag{123}$$

then Schauder's fixed point theorem applies again yielding to a fixed point v^* such that $u^* = \omega + v^*$ is a solution of (114), (115). We now construct a closed, bounded, convex set V which satisfies (123). Since $T'(\omega)=0$ by definition of ω , we have $T(\omega+v)-T(\omega)=T'(\omega)[v]+o(||v||_X)=o(||v||_X)$, hence, by virtue of (121), v^* can be expected to be small if ω is a good approximation of a solution u^* of (114), (115). Therefore, we assume V to be some small ball

$$V = \{ v \in X \mid ||v||_X \leqslant \alpha \}$$

$$(124)$$

with some $\alpha > 0$. In [74] X is suggested to be the space $H^{1,4}(\Omega)$ with the norm

$$||u||_{X} = \max\{||u||_{\infty}, \gamma||\nabla u||_{4}\}$$
(125)

and with

$$||u||_p = \left\{\frac{1}{\operatorname{meas}(\Omega)} \int_{\Omega} |v(x)|^p \, \mathrm{d}x\right\}^{1/p} \quad \text{for } p \in \{2, 4\}$$

here and in the remaining part of this section. The constant $\gamma > 0$ is adapted such that

$$||L^{-1}[r]||_{X} \leqslant K ||r||_{2} \quad \text{for all } r \in L^{2}(\Omega)$$
(126)

with a computable constant K > 0. Due to $\Phi'(\omega) = 0$ we have

$$||\varphi(v)||_2 = ||\Phi(\omega + v) - \Phi(\omega)||_2 = o(||v||_X) \text{ for } ||v||_X \to 0$$

Let $G:[0,\infty) \to [0,\infty)$ be a majorizing monotonically nondecreasing function such that

$$||\varphi(v)||_2 \leqslant G(||v||_X) \quad \text{for all } v \in X$$
(127)

and

$$G(t) = o(t) \quad \text{for } t \to +0. \tag{128}$$

Such a function can be found explicitly via an ansatz according to the lines in [74]. The following theorem is then crucial in view of (123).

Theorem 22. With the notation and the assumptions above let $||\delta[\omega]||_2 \leq \beta$ for some $\beta > 0$. If

$$\beta \leqslant \frac{\alpha}{K} - G(\alpha),\tag{129}$$

then V from (124) satisfies (123), i.e., there exists a solution $u^* \in H^2_B(\Omega)$ of (114), (115) with $||u^* - \omega||_X \leq \alpha$.

The proof follows immediately from

$$||L^{-1}[-\delta[\omega] + \varphi(v)]||_{X} \leqslant K(||\delta[\omega]||_{2} + ||\varphi(v)||_{2}) \leqslant K(\beta + G(||v||_{X})) \leqslant K(\beta + G(\alpha)) \leqslant \alpha$$

for each $v \in V$. Note that the right-hand side of (129) is positive for small α , hence (129) can be fulfilled if ω is a sufficiently good approximation of u^* which makes the defect $\delta[\omega]$ small. Some care has to be taken when computing the constants for the inequalities. It is here, among others, where interval arithmetic comes into play. For instance, in order to obtain the constant K in (126) and to check the invertibility of L (on $H_B^2(\Omega)$) one has to verify $\lambda_1 > 0$ for the smallest eigenvalue λ_1 of the eigenvalue problem (in weak formulation)

$$u \in H^2_B(\Omega), \quad \langle L[u], L[\psi] \rangle = \lambda \langle u, \psi \rangle \quad \text{for all } \psi \in H^2_B(\Omega)$$

with $\langle \cdot, \cdot \rangle$ denoting the canonical inner product in $L^2(\Omega)$. By means of interval arithmetic one is able to provide verified bounds for λ_1 and K. Details on the method including the computation of the approximation ω via finite elements can be found in [74] and in papers cited there.

While Plum's method can be characterized as an analytic one there are other methods for elliptic differential equations which use intervals in a more direct way. Thus for the Dirichlet problem

$$-\Delta u = f(u) \quad \text{in } \Omega,$$
$$u = 0 \qquad \text{on } \partial \Omega,$$

Nakao [65] works with some set U which has the form

$$U = \omega + \sum_{j=1}^{m} [a_j]\phi_j + \{\phi \in S^{\perp} \mid ||\phi||_{H^1_0} \leq \alpha\},$$

where $S \subseteq H_0^1(\Omega)$ is a finite-dimensional (finite element) subspace, S^{\perp} is its orthogonal complement in H_0^1 , $\{\phi_1, \ldots, \phi_m\}$ forms a basis of S and α is some constant which has to be determined numerically.

We also mention verification methods for hyperbolic equations - cf. for instance [28,47] and the literature there.

The investigation of old and the introduction of new ideas for the enclosure of solutions of differential equations is still a very active part of research.

9. Software for interval arithmetic

Interval arithmetic has been implemented on many platforms and is supported by several programming languages. The extended scientific computation (XSC) languages provide powerful tools necessary for achieving high accuracy and reliability. They provide a large number of predefined numerical data types and operations to deal with uncertain data.

PASCAL-XSC [46] is a general purpose programming language. Compared with PASCAL it provides an extended set of mathematical functions that are available for the types real, complex, interval and cinterval (complex interval) and delivers a result of maximum accuracy. Routines

for solving numerical problems have been implemented in PASCAL-XSC. PASCAL-XSC systems are available for personal computers, workstations, mainframes and supercomputers.

Similar remarks hold for the languages C-XSC [45] and FORTRAN-XSC [89].

ACRITH-XSC [40] is an extension of FORTRAN 77. It was developed in a joint project between IBM/Germany and the Institute of Applied Mathematics of the University of Karlsruhe (U. Kulisch). Unfortunately, it can be used only on machines with IBM/370 architecture that operates under the VMCMS operating system. It is a FORTRAN like programing library. Its features are dynamic arrays, subarrays, interval and vector arithmetic and problem solving routines for mathematical problems with verified results.

In the last section of the paper [50] one can find a general discussion of the availability of the necessary arithmetic for automatic result verification in hardware and suitable programming support. A detailed information of latest developments in the group of U. Kulisch can be found under http://www.uni-karlsruhe.de/~iam.

Via http://interval.usl.edu/kearfott one can get an overview on software written in the Computer Science Department of the University of South Louisiana, Lafayette, under the guidance of R. Baker Kearfott. Here is a short outline of available software:

- INTBIS (FORTRAN 77 code to find all solutions to polynomial systems of equations),
- INTLIB (ACM TOMS Algorithm 737 A FORTRAN 77 library for interval arithmetic and for rigorous bounds on the ranges of standard functions),
- INTERVAL ARITHMETIC (A FORTRAN 77 module that uses INTLIB to define an interval data type).

Programmer's Runtime Optimized Fast Library (PROFIL) developed at the Technical University of Hamburg-Harburg (S.M. Rump) is a C++ class library which has available usual real operations and the corresponding ones for intervals. Presently, the following data types are supported: int, real, interval, vectors and matrices for these types and complex numbers. For more details see http://www.ti3.tu-harburg.de/Software/PROFIL.html.

Recently, Rump announced the availability of an interval arithmetic package for MATLAB, called "INTLAB – A MATLAB library for interval arithmetic routines". Elements (toolboxes) of INTLAB are

- arithmetic operations for real and complex intervals, vectors and matrices over those, including sparse matrices,
- rigorous (real) standard functions,
- automatic differentiation including interval data,
- automatic slopes including interval data,
- multiple precision including interval data,
- rigorous input and output,
- some sample verification routines.

All INTLAB code is written in MATLAB for best portability. There is exactly one exception to that statement, that is one assembly language routine for switching the rounding mode of the processor (provided for some hardware platform).

Major objective of INTLAB is speed and ease of use. The first is achieved by a special concept for arithmetic routines, the second by the operator concept in MATLAB.

INTLAB code is easy to read and to write, almost as a specification. INTLAB is available for WINDOWS and UNIX systems, prerequisite is MATLAB Version 5. For more details and down-loading see http://www.ti3.tu-harburg.de/rump/intlab/.

References

- E. Adams, U. Kulisch (Eds.), Scientific Computing with Automatic Result Verification, Academic Press, Boston, 1993.
- [2] R. Albrecht, G. Alefeld, H.J. Stetter (Eds.), Validation Numerics, Theory and Applications, Springer, Wien, 1993.
- [3] G. Alefeld, Intervallrechnung über den komplexen Zahlen und einige Anwendungen, Ph.D. Thesis, Universität Karlsruhe, Karlsruhe, 1968.
- [4] G. Alefeld, Über die Durchführbarkeit des Gaußschen Algorithmus bei Gleichungen mit Intervallen als Koeffizienten, Comput. Suppl. 1 (1977) 15–19.
- [5] G. Alefeld, Bounding the slope of polynomials and some applications, Computing 26 (1981) 227–237.
- [6] G. Alefeld, Berechenbare Fehlerschranken f
 ür ein Eigenpaar unter Einschlu
 ß von Rundungsfehlern bei Verwendung des genauen Skalarprodukts, Z. Angew. Math. Mech. 67 (1987) 145–152.
- [7] G. Alefeld, Rigorous error bounds for singular values of a matrix using the precise scalar product, in: E. Kaucher, U. Kulisch, C. Ullrich (Eds.), Computerarithmetic, B.G. Teubner, Stuttgart, 1987, pp. 9–30.
- [8] G. Alefeld, Über das Divergenzverhalten des Intervall-Newton-Verfahrens, Computing 46 (1991) 289-294.
- [9] G. Alefeld, Inclusion methods for systems of nonlinear equations the interval Newton method and modifications, in: J. Herzberger (Ed.), Topics in Validated Computations, Elsevier, Amsterdam, 1994, pp. 7–26.
- [10] G. Alefeld, A. Gienger, G. Mayer, Numerical validation for an inverse matrix eigenvalue problem, Computing 53 (1994) 311–322.
- [11] G. Alefeld, A. Gienger, F. Potra, Efficient numerical validation of solutions of nonlinear systems, SIAM J. Numer. Anal. 31 (1994) 252–260.
- [12] G. Alefeld, J. Herzberger, Einführung in die Intervallrechnung, Bibliographisches Institut, Mannheim, 1974.
- [13] G. Alefeld, J. Herzberger, Introduction to Interval Computations, Academic Press, New York, 1983.
- [14] G. Alefeld, R. Hoffmann, G. Mayer, Verification algorithms for generalized singular values, Math. Nachr. 208 (1999) 5–29.
- [15] G. Alefeld, V. Kreinovich, G. Mayer, On the shape of the symmetric, persymmetric, and skew-symmetric solution set, SIAM J. Matrix Anal. Appl. 18 (1997) 693–705.
- [16] G. Alefeld, V. Kreinovich, G. Mayer, The shape of the solution set of linear interval equations with dependent coefficients, Math. Nachr. 192 (1998) 23–36.
- [17] G. Alefeld, V. Kreinovich, G. Mayer, On the solution sets of particular classes of linear systems, submitted for publication.
- [18] G. Alefeld, R. Lohner, On higher order centered forms, Computing 35 (1985) 177-184.
- [19] G. Alefeld, G. Mayer, The Cholesky method for interval data, Linear Algebra Appl. 194 (1993) 161–182.
- [20] G. Alefeld, G. Mayer, A computer-aided existence and uniqueness proof for an inverse matrix eigenvalue problem, Int. J. Interval Comput. 1994 (1) (1994) 4–27.
- [21] G. Alefeld, G. Mayer, On the symmetric and unsymmetric solution set of interval systems, SIAM J. Matrix Anal. Appl. 16 (1995) 1223–1240.
- [22] H. Beeck, Über Struktur und Abschätzungen der Lösungsmenge von linearen Gleichungssystemen mit Intervallkoeffizienten, Computing 10 (1972) 231–244.
- [23] H. Behnke, F. Goerisch, Inclusions for eigenvalues of selfadjoint problems, in: J. Herzberger (Ed.), Topics in Validated Computations, Elsevier, Amsterdam, 1994, pp. 277–322.
- [24] X. Chen, A verification method for solutions of nonsmooth equations, Computing 58 (1997) 281–294.
- [25] G.F. Corliss, Computing narrow inclusions for definite integrals, MRC Technical Summary Report # 2913, University of Madison, Madison, Wisconsin, February 1986.
- [26] G.F. Corliss, Introduction to validated ODE solving, Technical Report No. 416. Marquette University, Milwaukee, Wisconsin, March 1995.

- [27] H. Cornelius, R. Lohner, Computing the range of values of real functions with accuracy higher than second order, Computing 33 (1984) 331–347.
- [29] P.S. Dwyer, Linear Computations, Wiley, New York, 1951.
- [30] P. Eijgenraam, The solution of initial value problems using interval arithmetic. Formulation and analysis of an algorithm, Thesis, Mathematisch Centrum, Amsterdam, 1981.
- [31] W. Enger, Intervall Ray Tracing Ein Divide-and-Conquer Verfahren f
 ür photorealistische Computergrafik, Thesis, Universität Freiburg, Freiburg, 1990.
- [32] A. Gienger, Zur Lösungsverifikation bei Fredholmschen Integralgleichungen zweiter Art, Thesis, Universität Karlsruhe, Karlsruhe, 1997.
- [33] G.H. Golub, C.F. van Loan, Matrix Computations, 3rd Edition, John Hopkins, Baltimore, 1995.
- [34] A. Griewank, G.F. Corliss (Eds.), Automatic Differentiation of Algorithms, SIAM, Philadelphia, PA, 1992.
- [35] H. Grell, K. Maruhn, W. Rinow (Eds.), Enzyklopädie der Elementarmathematik, Band I Arithmetik, Dritte Auflage, VEB Deutscher Verlag der Wissenschaften, Berlin, 1966.
- [36] R. Hammer, M. Hocks, U. Kulisch, D. Ratz, Numerical Toolbox for Verified Computing I, Springer, Berlin, 1993.
- [37] E. Hansen, An overview of global optimization using interval analysis, in: R.E. Moore (Ed.), Reliability in Computing, The Role of Interval Methods in Scientific Computing, Perspectives in Computing, Vol. 19, Academic Press, Boston, 1988, pp. 289–307.
- [38] E. Hansen, Global Optimization Using Interval Analysis, Dekker, New York, 1992.
- [39] P. Hertling, A Lower Bound for Range Enclosure in Interval Arithmetic, Centre for Discrete Mathematics and Theoretical Computer Science Research Report Series, Department of Computer Science, University of Auckland, January 1998.
- [40] IBM High Accuracy Arithmetic-Extended Scientific Computation (ACRITH-XSC), General Information GC 33-646-01, IBM Corp., 1990.
- [41] C. Jansson, Rigorous sensitivity analysis for real symmetric matrices with interval data, in: E. Kaucher, S.M. Markov, G. Mayer (Eds.), Computer Arithmetic, Scientific Computation and Mathematical Modelling, IMACS Annals on Computing and Applied Mathematics, Vol. 12, J.C. Baltzer AG, Scientific Publishing, Basel, 1994, pp. 293–316.
- [42] C. Jansson, On self-validating methods for optimization problems, in: J. Herzberger (Ed.), Topics in Validated Computations, Elsevier, Amsterdam, 1994, pp. 381–438.
- [43] E.W. Kaucher, W.L. Miranker, in: Self-Validating Numerics for Function Space Problems, Computations with Guarantees for Differential and Integral Equations, Academic Press, Orlando, 1984.
- [44] R.B. Kearfott, A review of techniques in the verified solution of constrained global optimization problems, in: R.B. Kearfott, V. Kreinovich (Eds.), Applications of Interval Computations, Kluwer, Dordrecht, 1996, pp. 23–59.
- [45] R. Klatte, U. Kulisch, C. Lawo, M. Rauch, A. Wiethoff, C-XSC. A C⁺⁺ Class Library for Extended Scientific Computing, Springer, Berlin, 1993.
- [46] R. Klatte, U. Kulisch, M. Neaga, D. Ratz, C. Ullrich, PASCAL-XSC, Language Reference with Examples, Springer, Berlin, 1992.
- [47] M. Koeber, Lösungseinschlie
 ßung bei Anfangswertproblemen f
 ür quasilineare hyperbolische Differentialgleichungen, Thesis, Universit
 ät Karlsruhe, Karlsruhe, 1997.
- [48] R. Krawczyk, Newton-Algorithmen zur Bestimmung von Nullstellen mit Fehlerschranken, Computing 4 (1969) 187–201.
- [49] U. Kulisch, Grundzüge der Intervallrechnung, in: Jahrbuch Überblicke Mathematik, Vol. 2, Bibliographisches Institut, Mannheim, 1969.
- [50] U. Kulisch, Numerical algorithms with automatic result verification, Lectures in Applied Mathematics, Vol. 32, 1996, pp. 471–502.
- [51] J.-R. Lahmann, Eine Methode zur Einschließung von Eigenpaaren nichtselbstadjungierter Eigenwertprobleme und ihre Anwendung auf die Orr-Sommerfeld-Gleichung, Thesis, Universität Karlsruhe, Karlsruhe, 1999.
- [52] B. Lang, Lokalisierung und Darstellung von Nullstellenmengen einer Funktion, Diploma Thesis, Universität Karlsruhe, Karlsruhe, 1989.

- [53] R. Lohner, Einschließung der Lösung gewöhnlicher Anfangs- und Randwertaufgaben und Anwendungen, Thesis, Universität Karlsruhe, Karlsruhe, 1988.
- [54] R. Lohner, Enclosing all eigenvalues of symmetric matrices, in: C. Ullrich, J. Wolff von Gudenberg (Eds.), Accurate Numerical Algorithms. A Collection of Research Papers, Research Reports, ESPRIT, Project 1072, Diamond, Vol. 1, Springer, Berlin, 1989, pp. 87–103.
- [55] G. Mayer, Old and new aspects for the interval Gaussian algorithm, in: E. Kaucher, S.M. Markov, G. Mayer (Eds.), Computer Arithmetic, Scientific Computation and Mathematical Modelling, IMACS Annals on Computing and Applied Mathematics, Vol. 12, J.C. Baltzer AG, Scientific Publishing, Basel, 1991, pp. 329–349.
- [56] G. Mayer, Enclosures for eigenvalues and eigenvectors, in: L. Atanassova, J. Herzberger (Eds.), Computer Arithmetic and Enclosure Methods, Elsevier, Amsterdam, 1992, pp. 49–67.
- [57] G. Mayer, Result verification for eigenvectors and eigenvalues, in: J. Herzberger (Ed.), Topics in Validated Computations, Elsevier, Amsterdam, 1994, pp. 209–276.
- [58] G. Mayer, Epsilon-inflation in verification algorithms, J. Comput. Appl. Math. 60 (1995) 147-169.
- [59] G. Mayer, Epsilon-inflation with contractive interval functions, Appl. Math. 43 (1998) 241-254.
- [60] G. Mayer, J. Rohn, On the applicability of the interval Gaussian algorithm, Reliable Comput. 4 (1998) 205-222.
- [61] O. Mayer, Über die in der Intervallrechnung auftretenden Räume und einige Anwendungen, Thesis, Universität Karlsuhe, Karlsruhe, 1968.
- [62] C. Miranda, Un' osservazione su un teorema di Brouwer, Bol. Un. Mat. Ital. Ser. II 3 (1941) 5-7.
- [63] R.E. Moore, Interval Arithmetic and Automatic Error Analysis in Digital Computing, Thesis, Stanford University, October 1962.
- [64] R.E. Moore, Interval Analysis, Prentice-Hall, Englewood Cliffs, NJ, 1966.
- [65] M.T. Nakao, State of the art for numerical computations with guaranteed accuracy, Math. Japanese 48 (1998) 323–338.
- [66] N.S. Nedialkov, Computing rigorous bounds on the solution of an initial value problem for an ordinary differential equation, Thesis, University of Toronto, Toronto, 1999.
- [67] M. Neher, An enclosure method for the solution of linear ODEs with polynomial coefficients, Numer. Funct. Anal. Optim. 20 (1999) 779–803.
- [68] A. Neumaier, The enclosure of solutions of parameter-dependent systems of equations, in: R.E. Moore (Ed.), Reliability in Computing. The Role of Interval Methods in Scientific Computing, Perspectives in Computing, Vol. 19, Academic Press, Boston, 1988, pp. 269–286.
- [69] A. Neumaier, Interval Methods for Systems of Equations, University Press, Cambridge, 1990.
- [70] H.T. Nguyen, V. Kreinovich, V. Nesterov, M. Nakumura, On hardware support for interval computations and for soft computing: theorems, IEEE Trans. Fuzzy Systems 5 (1) (1997) 108–127.
- [71] J.M. Ortega, W.C. Rheinboldt, Iterative Solution of Nonlinear Equations in Several Variables, Academic Press, New York, 1970.
- [72] W. Oettli, W. Prager, Compatibility of approximate solution of linear equations with given error bounds for coefficients and right-hand sides, Numer. Math. 6 (1964) 405–409.
- [73] M. Petković, L.D. Petković, Complex Interval Arithmetic and Its Applications, Wiley, New York, 1998.
- [74] M. Plum, Inclusion methods for elliptic boundary value problems, in: J. Herzberger (Ed.), Topics in Validated Computations, Elsevier, Amsterdam, 1994, pp. 323–379.
- [75] L.B. Rall, Computational Solution of Nonlinear Operator Equations, Wiley, New York, 1969.
- [76] L.B. Rall, Automatic Differentiation: Techniques and Applications, Lecture Notes in Computer Science, Vol. 120, Springer, Berlin, 1981.
- [77] H. Ratschek, Centered forms, SIAM J. Numer. Anal. 17 (1980) 656-662.
- [78] H. Ratschek, J. Rokne, Computer Methods for the Range of Functions, Ellis Horwood, Chichester, 1984.
- [79] H. Ratschek, J. Rokne, New Computer Methods for Global Optimization, Ellis Horwood, Chichester, UK, 1988.
- [80] R. Rihm, Interval methods for initial value problems in ODE's, in: J. Herzberger (Ed.), Topics in Validated Computations, Elsevier, Amsterdam, 1994, pp. 173–207.
- [81] S.M. Rump, Solving algebraic problems with high accuracy, in: U.W. Kulisch, W.L. Miranker (Eds.), A New Approach to Scientific Computation, Academic Press, New York, 1983, pp. 53–120.
- [82] S.M. Rump, Rigorous sensitivity analysis for systems of linear and nonlinear equations, Math. Comp. 54 (1990) 721–736.

- [83] S.M. Rump, On the solution of interval linear systems, Computing 47 (1992) 337–353.
- [84] S.M. Rump, Verification methods for dense and sparse systems of equations, in: J. Herzberger (Ed.), Topics in Validated Computations, Elsevier, Amsterdam, 1994, pp. 63–135.
- [85] A. Schrijver, Theory of Linear and Integer Programming, Wiley, New York, 1986.
- [86] H. Schwandt, Schnelle fast global konvergente Verfahren für die Fünf-Punkte-Diskretisierung der Poissongleichung mit Dirichletschen Randbedingungen auf Rechteckgebieten, Thesis, Fachbereich Mathematik der TU Berlin, Berlin, 1981.
- [87] P.S. Shary, Solving the linear interval tolerance problem, Math. Comput. Simulation 39 (1995) 53–85.
- [88] T. Sunaga, Theory of an interval algebra and its application to numerical analysis, RAAG Memoirs 2 (1958) 29-46.
- [89] W.V. Walter, FORTRAN-XSC: a portable FORTRAN 90 module library for accurate and reliable scientific computing, in: R. Albrecht, G. Alefeld, H.J. Stetter (Eds.), Validation Numerics, Theory and Applications, Springer, Wien, 1993, pp. 265–285.

Further reading

- G. Alefeld, A. Frommer, B. Lang (Eds.), Scientific Computing and Validated Numerics, Mathematical Research, Vol. 90, Akademie Verlag, Berlin, 1996.
- [2] G. Alefeld, R.D. Grigorieff (Eds.), Fundamentals of Numerical Computation, Computer-Oriented Numerical Analysis, Computing Supplementum, Vol. 2, Springer, Wien, 1980.
- [3] G. Alefeld, J. Herzberger (Eds.), Numerical Methods and Error Bounds, Mathematical Research, Vol. 89, Akademie Verlag, Berlin, 1996.
- [4] L. Atanassova, J. Herzberger (Eds.), Computer Arithmetic and Enclosure Methods, Elsevier, Amsterdam, 1992.
- [5] H. Bauch, K.-U. Jahn, D. Oelschlägel, H. Süße, V. Wiebigke, Intervallmathematik, Theorie und Anwendungen, Mathematisch-Naturwissenschaftliche Bibliothek, Vol. 72, BSB B.G. Teubner, Leipzig, 1987.
- [6] E. Hansen (Ed.), Topics in Interval Analysis, Oxford University Press, Oxford, 1969.
- [7] J. Herzberger (Ed.), Topics in Validated Computations, Elsevier, Amsterdam, 1994.
- [8] J. Herzberger (Ed.), Wissenschaftliches Rechnen, Eine Einführung in das Scientific Computing, Akademie Verlag, Berlin, 1995.
- [9] S.A. Kalmykov, J.I. Shokin, S.C. Yuldashev, Methods of Interval Analysis, Novosibirsk, 1986 (in Russian).
- [10] E. Kaucher, U. Kulisch, C. Ullrich (Eds.), Computerarithmetic, B.G. Teubner, Stuttgart, 1987.
- [11] E. Kaucher, S.M. Markov, G. Mayer (Eds.), in: Computer Arithmetic, Scientific Computation and Mathematical Modelling, IMACS Annals on Computing and Applied Mathematics, Vol. 12, J.C. Baltzer AG, Scientific Publishing, Basel, 1991.
- [12] R.B. Kearfott, V. Kreinovich (Eds.), Applications of Interval Computations, Kluwer, Dordrecht, 1996.
- [13] V. Kreinovich, A. Lakeyev, J. Rohn, P. Kahl, Computational Complexity and Feasibility of Data Processing and Interval Computations, Kluwer, Dordrecht, 1998.
- [14] U. Kulisch (Ed.), Wissenschaftliches Rechnen mit Ergebnisverifikation, Vieweg, Braunschweig, 1989.
- [15] U.W. Kulisch, W.L. Miranker, Computer Arithmetic in Theory and Practice, Academic Press, New York, 1981.
- [16] U.W. Kulisch, W.L. Miranker (Eds.), A New Approach to Scientific Computation, Academic Press, New York, 1983.
- [17] U. Kulisch, W. Miranker, The arithmetic of the digital computer: a new approach, SIAM Rev. 28 (1986) 1-40.
- [18] U. Kulisch, H.J. Stetter (Eds.), Scientific Computation with Automatic Result Verification, Computing Supplementum, Vol. 6, Springer, Wien, 1988.
- [19] S.M. Markov (Ed.), Scientific Computation and Mathematical Modelling, DATECS Publishing, Sofia, 1993.
- [20] W.L. Miranker, R.A. Toupin (Eds.), Accurate Scientific Computations, Lecture Notes in Computer Science, Vol. 235, Springer, Berlin, 1986.
- [21] R.E. Moore, Methods and Applications of Interval Analysis, SIAM, Philadelphia, 1979.
- [22] R.E. Moore, Computational Functional Analysis, Ellis Horwood, Chichester, 1985.
- [23] R.E. Moore (Ed.), Reliability in Computing. The Role of Interval Methods in Scientific Computing, Perspectives in Computing, Vol. 19, Academic Press, Boston, 1988.
- [24] K. Nickel (Ed.), Interval Mathematics, Lecture Notes in Computer Science, Vol. 29, Springer, Berlin, 1975.

- [25] K.L.E. Nickel (Ed.), Interval Mathematics 1980, Academic Press, New York, 1980.
- [26] K. Nickel (Ed.), Interval Mathematics 1985, Lecture Notes in Computer Science, Vol. 212, Springer, Berlin, 1985.
- [27] M. Petković, Iterative Methods for Simultaneous Inclusion of Polynomial Zeros, Lecture Notes in Mathematics, Vol. 1387, Springer, Berlin, 1989.
- [28] C. Ullrich (Ed.), Computer Arithmetic and Self-Validating Numerical Methods, Academic Press, Boston, 1990.
- [29] C. Ullrich (Ed.), Contributions to Computer Arithmetic and Self-Validating Numerical Methods, IMACS Annals on Computing and Applied Mathematics, Vol. 7, J.C. Baltzer AG, Scientific Publishing, Basel, 1990.
- [30] C. Ullrich, J. Wolff von Gudenberg (Eds.), Accurate Numerical Algorithms. A Collection of Research Papers, Research Reports, ESPRIT, Project 1072, Diamond, Vol. 1, Springer, Berlin, 1989.
- [31] T. Csendes (Ed.), Developments in Reliable Computing, Kluwer, Dordrecht, 1999.



Journal of Computational and Applied Mathematics 121 (2000) 465

JOURNAL OF COMPUTATIONAL AND APPLIED MATHEMATICS

www.elsevier.nl/locate/cam

421-464

Author Index Volume 121 (2000)

Mayer, G., see Alefeld, G.

Alefeld, G. and G. Mayer, Interval analysis: theory and applications	421-464
Area, I., see Ronveaux, A.	95-112
Borwein, J.M., D.M. Bradley and R.E. Crandall, Computational strategies for the Riemann zeta	
function	247-296
Bradley, D.M., see Borwein, J.M.	247-296
Bultheel, A. and B. De Moor, Rational approxima-	
tion in linear systems and control	355-378
Crandall, R.E., see Borwein, J.M.	247-296
De Moor, B., see Bultheel, A.	355-378
linear system theory: an overview	331-354
Everitt, W.N., L.L. Littlejohn and R. Wellman, The	
mite differential equation	313-330
Godoy, E., see Ronveaux, A.	95-112
proximation	197–219
Huard, A., see Guillaume, P.	197–219
Leviatan, D., Shape-preserving approximation by	
polynomials	73- 94
Lewanowicz, S., Recursion formulae for basic hyper-	
geometric functions	297-312
Littlejohn, L.L., see Everitt, W.N.	313-330

Nievergelt, Y., A tutorial history of least squares with applications to astronomy and geodesy Nürnberger, G. and F. Zeilfelder, Developments in bivariate spline interpolation	37- 72 125-152
orvariate spinie interpolation	120 102
Piessens, R., Computing integral transforms and solving integral equations using Chebyshev polynomial approximations	113-124
Ronveaux, A., A. Zarzo, I. Area and E. Godoy, Classi- cal orthogonal polynomials: dependence of para- meters	95-112
 Schaback, R., A unified theory of radial basis functions. Native Hilbert spaces for radial basis functions II Stenger, F., Summary of Sinc numerical methods 	165–177 379–420
 Taswell, C., Constraint-selected and search-optimized families of Daubechies wavelet filters computable by spectral factorization Temme, N.M., Numerical and asymptotic aspects of parabolic cylinder functions 	179–195 221–246
 Wang, RH., Multivariate spline and algebraic geometry Watson, G.A., Approximation in normed linear spaces Wellman, R., see Everitt, W.N. 	153–163 1– 36 313–330
Zarzo, A., see Ronveaux, A.	95-112
Zeilfelder, F., see Nürnberger, G.	125-152