BIRKHÄUSER

Operator Theory
Advances and Applications
Vol. 199

Numerical Methods for Structured Matrices and Applications

The Georg Heinig Memorial Volume

Dario Andrea Bini
Volker Mehrmann
Vadim Olshevsky
Eugene E. Tyrtyshnikov
Marc van Barel
Editors

BIRKHÄUSER

Operator Theory: Advances and Applications

Vol. 199

Founded in 1979 by Israel Gohberg

Numerical Methods for Structured Matrices and Applications

The Georg Heinig Memorial Volume

Dario Andrea Bini
Volker Mehrmann
Vadim Olshevsky
Eugene E. Tyrtyshnikov
Marc van Barel
Editors

Birkhäuser

Editors:

Dario Andrea Bini
Dipartimento di Matematica
Università di Pisa
Largo Bruno Pontecorvo, 5
56127 Pisa
Italy
e-mail: bini@dm.unipi.it

Eugene E. Tyrtyshnikov
Institute of Numerical Mathematics
Russian Academy of Sciences
Gubkina Street, 8
Moscow, 119991
Russia
e-mail: tee@inm.ras.ru

Volker Mehrmann
Institut für Mathematik
Technische Universität Berlin
Straße des 17. Juni 136
10623 Berlin
Germany
e-mail: mehrmann@math.tu-berlin.de

Marc van Barel
Katholieke Universiteit Leuven
Department of Computer Science
Celestijnenlaan 200A
3001 Leuven (Heverlee)
Belgium
e-mail: marc.vanbarel@cs.kuleuven.be

Vadim Olshevsky
Department of Mathematics
University of Connecticut
196 Auditorium Road, U-9
Storrs, CT 06269
USA
e-mail: olshevsky@uconn.edu

2010 Mathematics Subject Classification: 15

Library of Congress Control Number: 2010920068

Bibliographic information published by Die Deutsche Bibliothek.
Die Deutsche Bibliothek lists this publication in the Deutsche Nationalbibliografie;
detailed bibliographic data is available in the Internet at http://dnb.ddb.de

ISBN 978-3-0346-8995-6

P.O. Box 133, CH-4010 Basel, Switzerland
Part of Springer Science+Business Media
Printed on acid-free paper produced from chlorine-free pulp. TCF∞
Printed in Germany

ISBN 978-3-7643-8995-6 e-ISBN 978-3-7643-8996-3

9 8 7 6 5 4 3 2 1 www.birkhauser.ch

Contents

Foreword

Georg Heing, a charming, erudite man, and a first rate mathematician died unexpectedly of a heart attack on May 10, 2005. Georg is survived by his wife Gerti, his daughter Susanne, and his son Peter.

We have lost one the leading experts in the field of structured matrices, a wonderful colleague, and a terrific friend.

Georg Heinig's results, approaches, and his scientific taste influenced our community of researchers working on structured matrices. In fact, the community's focus grew to reflect his interdisciplinary vision ranging from applications (e.g., in systems and control theory and signal processing) through fundamental mathematics (structured matrices, periodic Jacobi, Toeplitz, and Wiener-Hopf operators, classes of singular integral operators, resultants and Bezoutians for operator-valued polynomials and continual analogs thereof) to numerical analysis and fast algorithms. The broad spectrum of Georg Heinig's interests are represented in this collection.

Georg served as an Associate Editor of two top journals: Integral Equations and Operator Theory and Linear Algebra and Its Applications. This volume starts with two eulogies published earlier by IEOT and LAA. The first one, published in IEOT is by Albrecht Böttcher, Israel Gohberg (who was Georg's advisor during his Ph.D. studies), and Bernd Silbermann. The second one, published in LAA is by Karla Rost who collaborated with Georg during last three decades until day of his death. They have produced together more than 30 papers and a monograph.

We refer to these two eulogies for the details of Georg's career, and here we would like to emphasize only one point, namely the influence of his work in the area of structured matrices. Matrices with structure (e.g., Toeplitz matrices) are encountered in a surprising variety of areas in sciences and mathematics. There were many approaches to study Toeplitz structure and its generalizations, one of them was known under the name "displacement structure method." In their 1984 monograph *"Algebraic Methods for Toeplitz-like Matrices and perators"* G.Heinig and K.Rost demonstrated that this method (they called it the UV-reduction method) can be successfully used not only for Toeplitz structure and its derivatives, but also for many other patterns of structure, e.g., Hankel, Vandermonde, Cauchy matrices, Bezoutians and their generalizations. This breakthrough discovery facilitated a lot of interest in the community. Moreover, the new technique was immediately picked up and it was heavily used in the work of a number of research groups in Germany, USA, Israel, Leuven, Moscow, Hong Kong.

As Georg mentioned many times, about 20 years ago he was virtually alone delivering talks on structured matrices at such conferences as IWOTA and ILAS meetings. Nowadays special sessions and minisymposia on structured matrices are routinely included in programs of a number of conferences such as IWOTA, ILAS, SIAM annual meetings, SPIE, MTNS. Moreover, a number of conferences dedicated exclusively to structured matrices has been held (two AMS meetings in the USA, four conferences in Italy, three in Moscow, three in Hong Kong). Needless to say, Georg's results, ideas, his energy, and service to the community facilitated this development and strongly influenced the research efforts of structured matrices community.

We are happy to include in this volume a joint paper of Georg Heinig and Karla Rost on Bezoutians. This is a subject Georg worked on since the very beginning of his career, and to which he made a number of significant contributions. The paper blends an wonderful exposition of classical results with a survey recent development in the field.

It was a great honor and a privilege to edit this volume of papers dedicated in Georg's memory.

The Editors

Part I

Georg Heinig

Operator Theory:
Advances and Applications, Vol. 199, 3–6

Georg Heinig (1947–2005)
In Memoriam

On May 10, 2005, Georg Heinig died unexpectedly of a heart attack in his apartment in Kuwait. We have lost one of the top experts in the field of structured matrices, an irreplaceable colleague, and a good friend. He was an active member of the editorial boards of the journal *Integral Equations and Operator Theory* and the book series *Operator Theory: Advances and Applications* since 1993. Our heartfelt condolences go out to his wife and his family.

Originally published in *Integr. equ. oper. theory* **53** (2005), 297–300.

Georg Heinig was born on November 24, 1947 in the small town of Zschopau in the Ore Mountains (Erzgebirge) in East Germany. From 1954 to 1964 he attended the school in Zschopau and from 1964 to 1966 the elite class for mathematics at Chemnitz University of Technology. Such elite classes were established to provide especially gifted pupils with an extraordinary education in mathematics (but also in the natural sciences and in languages) under the guidance of experienced university teachers. The careers of many successful East German scientists started at elite classes. None of these classes has survived the German reunification.

He studied mathematics at Chemnitz University of Technology from 1966 to 1970 and graduated with the diploma degree in 1970. His diploma thesis was written under the supervision of Siegfried Prössdorf and was devoted to certain properties of normally solvable operators in Banach spaces.

After defending his diploma thesis with the best possible grade, Georg Heinig was given the opportunity of entering a PhD program abroad. He decided to continue his studies at Kishinev (now Chisinau) University under the supervision of the second of us. His wife Gerti accompanied him in Kishinev and also completed a dissertation during that period. Georg Heinig was a very talented and dedicated researcher. In Kishinev he embarked on research into the theory of Toeplitz, Wiener-Hopf, and singular integral operators with scalar and matrix-valued symbols, and it was during those wonderful years that he has fallen in love with all the exciting mathematics of structured matrices. His deep results in this area formed the basis of his excellent PhD thesis, which he defended in Spring of 1974. Many other mathematical insights gained by Georg during the years in Kishinev went into his habilitation thesis, which he completed in Chemnitz. The early paper Gohberg/Heinig, Inversion of finite Toeplitz matrices consisting of elements of a non-commutative algebra (Russian), Rev. Roumaine Math. Phys. Appl. 19, 623–663 (1974) became one of his most frequently cited works.

Georg Heinig returned to Chemnitz in 1974. In the following five years the first of us had the pleasure of attending his classes as a student, the third of us received an outstanding member of his research group, and the second of us was proud of Georg's outstanding mathematical achievements. Georg Heinig integrated several young people into his research, Karla Rost being the most prominent figure of them. In 1979 he defended his habilitation thesis, which was on the spectral theory of operator bundles and the algebraic theory of finite Toeplitz matrices. His two children Peter and Susanne were born in 1974 and 1977.

The scientific outcome of the research directed by Georg Heinig in the 1970s and early 1980s is summarized in his and Karla Rost's book *Algebraic Methods for Toeplitz-like Matrices and Operators,* which was originally published by Akademie-Verlag, Berlin in 1984 and was republished by Birkhäuser Verlag, Basel in the same year. This book has found a warm reception and perpetual interest by a large community for now about twenty years. Some of its basic ideas, such as the so-called UV reduction (which later received more popularity under the name displacement operation), have become important tools for workers in the field of

structured matrices. Moreover, the scientific collaboration of Georg Heinig with Karla Rost lasted three decades until the day of Georg's death. Their joint research resulted in more than 30 papers. The results and methods of these papers are an essential ingredient to the present-day mathematical high-technology one is encountering in connection with structured matrices.

In 1982, Georg Heinig was a guest professor at Aleppo University in Syria, and from 1987 to 1989, he held a guest professorship at Addis Ababa University in Ethiopia. In the late 1980s he was appointed full professor at Leipzig University.

After the political events in Germany at the turn to the 1990s the life for Georg changed dramatically. All people working at East German universities were formally dismissed and had to apply for a position anew. Those who had shown a certain extent of political proximity to the former socialist system had no chance of receiving a new position at a German university, neither in East Germany nor in the subsequently reunified Germany. The situation was extremely difficult, and the efforts of Georg's friends to help him did not bring any positive results. Certainly Georg was very disappointed and despaired. Some time he planned to take over his father's store for vegetables, but eventually he looked for a job at a foreign university.

In 1993, Georg Heing went to Kuwait University, where he worked as a professor until his tragic death. The scientific conditions at Kuwait University were excellent and Georg has always thankfully acknowledged the recognition and friendship he received from his Kuwaiti colleagues. In 2003, he was awarded as the Scientist of the Year by the Amir of Kuwait. Despite all these successes, his and his wife's dream was to endure the university job only until the age of 60 years and then simply to relish life together, including travelling around the world. His unexpected death at the age of 57 abruptly dispersed this dream.

Georg Heinig's scientific legacy is immense. In more than 100 publications he made outstanding contributions to a variety of fields, including

- theory and fast algorithms for several classes of structured matrices,
- periodic Jacobi, Toeplitz, and Wiener-Hopf operators,
- classes of singular integral operators,
- resultants and Bezoutians for operator-valued polynomials,
- continual analogs of resultants and Bezoutians,
- numerical methods for convolution equations,
- applications in systems and control theory and signal processing.

Discoveries by Georg and his co-workers, such as the structure of the kernel and of the pseudoinverse for certain classes of structured matrices, significantly shaped the development of numerical algorithms. He also remarkably enriched various areas of operator theory, for example by deep results on the spectral theory of Jacobi matrices and of Toeplitz and Wiener-Hopf operators. He supervised 6 dissertations.

Georg Heing was a very pleasant person and an inspiring colleague. His sense of humor and his characteristic bright laughing will be missed by everyone who was lucky enough to meet him. His permanent endeavor for disclosing the absolute essence of a matter and his untiring aspiration for clearness and brevity were challenges for his co-workers on the one hand and have resulted in grateful appreciation by his students and the readers of his publications on the other.

Another dream of Georg Heinig was a joint textbook with Karla Rost on structured matrices, ranging from the basics for beginners up to recent developments. About one year ago they started writing this book and three chapters are already more or less complete. It is unimaginable that he will never have this book in his hands some day. This tragedy bitterly reveals the gap that Georg has left and painfully reminds us of the projects and ideas that passed away with him. However, his work will endure and we will always remember this outstanding mathematician, excellent colleague, and wonderful friend.

Albrecht Böttcher, Israel Gohberg, Bernd Silbermann

Operator Theory:
Advances and Applications, Vol. 199, 7–19

Georg Heinig
November 24, 1947 – May 10, 2005
A Personal Memoir and Appreciation

Karla Rost

Originally published in *Linear Algebra and its Applications* **413** (2006), 1–12.

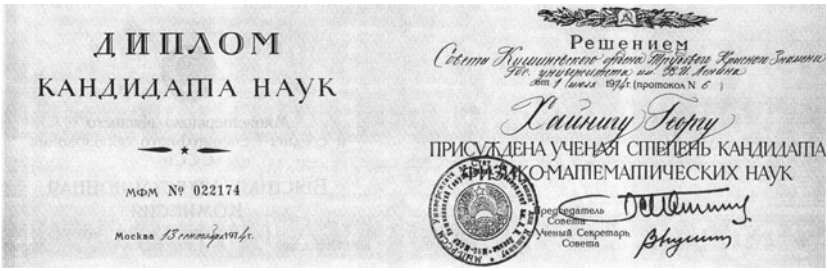
On May 10, 2005, Georg Heinig, an excellent mathematician died unexpectedly at the age of 57. He was a world leader in the field of structured matrices. As associate editor of the journal Linear Algebra and its Applications since his appointment in 1991 he contributed much to the journal's success by his valuable and extensive work. In what follows I want to try to capture some aspects of this mathematical life and his personality.

Georg was born on November 24, 1947, in the small town of Zschopau in the Ore Mountains (Erzgebirge) in East Germany. From 1954 to 1964 he attended the elementary school there. Because of his good performance he was admitted to the elite school of Karl-Marx-Stadt (now Chemnitz) University of Technology, where he received his graduation diploma with the grade "very good". During his time at this school he already showed extraordinary talent for mathematics and natural sciences, and his passion and skills for solving mathematical problems grew.

Subsequently he studied mathematics at Karl-Marx-Stadt (now Chemnitz) University of Technology. He wrote his diploma thesis under the supervision of Siegfried Prössdorf [SP] on some properties of normally solvable operators in Banach spaces. Some years ago S. Prössdorf told me that he liked to recall the time he spent with the gifted and creative student Georg. In the summer of 1970 Georg received the best possible grade for the defense of his diploma thesis.

He received a scholarship to study for the Ph.D. abroad and he decided to go to the State University of Moldavia at Kishinev from 1971 to 1974. There he worked on his PhD thesis on the subject of Wiener-Hopf block operators and singular integral operators under the supervision of Israel Gohberg [IG] who even then was internationally well known and respected.

Here is a copy of the authenticated Russian document certifying Georg's degree as "candidate of sciences", which is the equivalent to a PhD:


ДИПЛОМ
КАНДИДАТА НАУК
МФМ № 022174
Решением
Хайнигу Георгу
ПРИСУЖДЕНА УЧЕНАЯ СТЕПЕНЬ КАНДИДАТА
ФИЗИКО-МАТЕМАТИЧЕСКИХ НАУК
Председатель Совета
Ученый Секретарь Совета

A well-known, important and often cited work with I. Gohberg from this time is the paper [111]. With great admiration and deep gratitude Georg always considered I. Gohberg as his scientific father ([IG], page 63).

By this time Georg was cast irretrievably into the realm of matrix theory, in particular the theory of structured matrices. His commitment to this field over three decades has benefited several scientific grandchildren of Israel Gohberg's of whom I am one. Georg's connection to I. Gohberg has never ceased. From 1993

on, he was a member of the editorial board of the journal Integral Equations and Operator Theory.

I became acquainted with Georg when he returned to Karl-Marx-Stadt in late 1974, where he worked at the Department of Mathematics, first in the group chaired by S. Prössdorf and then (after Prössdorf's leave to Berlin 1975) in B. Silbermann's group [BS]. There he found extremely good conditions. Prössdorf and Silbermann considered him as an equal partner, and hence he could pursue his inclinations in research unhampered and even build a small research group.

In all honesty, I have to admit that I was not euphoric at my first encounters with Georg. He was very young, with no experience as a supervisor, and in addition, he appeared to me as too self-oriented. It was his ability to awake my interest in the topics he proposed, which in 1975 led me to decide to write my diploma thesis under his supervision despite my initial hesitations. In fact he turned out to be an extraordinary supervisor, and I soon became aware that starting my scientific career with him was a lucky decision. In later years we managed better and better to get attuned to each other, and consequently I wrote a large part of my dissertation on the method of UV-reduction for inverting structured matrices under his supervision in 1980. Meanwhile 30 years of fruitful and intense joint work have passed. One joint monograph and almost 40 papers in journals testify to this.

In 1979 Georg defended his habilitation thesis (of an imposing length of 287 pages) on the spectral theory of operator bundles and the algebraic theory of finite Toeplitz matrices with excellence.

Georg was very optimistic and in love with life. I very much miss his cheerful and bright laughter. Certainly his stay in Kishinev intensified his wanderlust and his curiosity for other countries. Despite the travel restrictions for citizens of the G.D.R., the former socialist part of Germany, there was scientific cooperation with Syria and Ethiopia, and Georg was offered a research and working visit at Aleppo University in Syria in 1982. During a longer stay from 1985–87 at Addis Ababa University in Ethiopia, he was accompanied by his wife Gerti and by his two children Peter (born 1974) and Susanne (born 1977). Later on, both these stays certainly helped him to settle down in Kuwait.

Georg was well established at Karl-Marx-Stadt (now Chemnitz) University of Technology. He was a respected and highly recognized colleague with outstanding achievements in research and teaching. Thus Georg was appointed as a full professor for numerical mathematics at Karl Marx University of Leipzig. Since the late seventies his international recognition has grown enormously, which is, for example, reflected by the interest of the Birkhäuser publishing house in the joint monograph [85, 88], which was originally intended to be published by the Akademie-Verlag only.

The "Wende" in the fall of 1989 was an incisive break and turning point in the life of many people in East Germany, and thus also for Georg. All scientists working at universities were formally dismissed and had to apply for a position anew. An important criterion for a refusal of such an application was the political

proximity to the old socialist system. Due to this, in 1993 Georg went to Kuwait University, where he worked as a professor for more than 10 years. He died of a heart attack on May 10, 2005, in his apartment in Kuwait.

During this long period in Kuwait he continuously maintained scientific and personal contacts with his friends and former colleagues from Chemnitz, including Albrecht Böttcher, Bernd Silbermann, Steffen Roch, and myself. In May 1998 we all had the opportunity to participate in the International Conference on Fourier Analysis and Applications in Kuwait. Georg had an especially high admiration for Albrecht Böttcher and was therefore very glad that Albrecht agreed to enter the scientific committee and the editorial board of the proceedings of this conference [33]. In the course of that conference we convinced ourselves with great pleasure of the respect in which Georg was held by his colleagues and students in Kuwait. Thus he found very good friends and supporters in his Kuwaiti colleagues Fadhel Al-Musallam and Mansour Al-Zanaidi as well as in his colleague Christian Grossmann from Dresden University of Technology, who stayed in Kuwait from 1992 to 1998. One of the highlights of his life occurred in 2002, when the Amir of Kuwait distinguished him as the Researcher of the Year. Since 2004 he was also a member of the editorial board of the Kuwait Journal of Science and Engineering.

My mathematical knowledge and my ability to tackle problems have benefited immensely from Georg. He had an extraordinary gift to explain complicated things in simple terms. This was also appreciated by his students. His lectures and scientific talks were very sought after and well attended. The aesthetic component is well to the fore in his work. He mastered with equal facility problems of extreme generality and abstraction as well as down-to-earth questions.

Georg is the author and coauthor of more than 100 scientific publications. He always made high demands on himself and on his coauthors regarding not only mathematical originality and exactness but also regarding clear and short exposition.

His main research interests are

- structured matrices: algebraic theory and fast algorithms,
- interpolation problems,
- operator theory and integral equations,
- numerical methods for convolution equations,
- applications in systems and control theory and signal processing.

In each of these topics he achieved essential contributions which is impressively shown by his list of publications. In my opinion, especially the importance of his results concerning the algebraic theory of structured matrices are striking and imposing. In particular, in our joint paper [74] we show that inverses of matrices which are the sum of a Toeplitz and a Hankel matrix possess a Bezoutian structure as inverses of Hankel or Toeplitz matrices do separately. On the basis of this structure, for example, matrix representations can be found and fast algorithms can be designed. Thus a breakthrough for the class of Toeplitz-plus-Hankel matrices was achieved.

Moreover, Georg's observation of the kernel structure of (block) Toeplitz and Toeplitz-plus-Hankel matrices turns out to be a suitable key to develop algorithms without additional assumptions. His ideas how to connect the structure of a matrix with its additional symmetries lead to more efficient inversion and solution algorithms as well as a new kind of factorization. But also his contributions to Toeplitz least square problems, to transformation techniques for Toeplitz and Toeplitz-plus-Hankel matrices leveraged the research in these fields.

I am proud that he has completed (and still wanted to complete) many mathematical projects with me. For many years we dreamt about writing a neat textbook on structured matrices for graduate students, ranging from the basics for beginners up to recent developments. One year ago we really started writing the first chapters of this book. Except for some interruptions, when we were very busy with teaching, the collaboration was therefore especially intense, partly culminating in a dozen of emails per day. I received two emails from him on May 10, 2005. I then did not know that they were his last!

Georg's work leaves behind a trail that points to directions for future research. His early death leaves a loss from which we cannot recover, for it is tragic how many plans and original ideas have passed away with him. Colleagues like me are left behind in shock and ask themselves how we can at least partially close the gap that he has left.

In such situations the persistent optimist Georg used to say:

"Lamenting does not help. Things are as they are.
Let us confidently continue to work. This helps!"

References

[BS] Toeplitz Matrices and Singular Integral Equations: The Bernd Silbermann Anniversary Volume (Pobershau, 2001), eds. A. Böttcher, I. Gohberg, P. Junghanns, Operator Theory: Advances and Applications, Vol. 135, Birkhäuser, Basel, 2002.

[SP] Problems and Methods in Mathematical Physics: The Siegfried Prössdorf Memorial Vol. (Proceed. of the 11 TMP Chemnitz, 1999), eds. J. Elschner, I. Gohberg, B. Silbermann, Operator Theory: Advances and Applications, Vol. 121, Birkhäuser, Basel, 2001.

[IG] The Gohberg Anniversary Collection, Vol. I (Calgary, 1988), eds. H. Dym, S. Goldberg, M.A. Kaashoek, P. Lancaster, Operator Theory: Advances and Applications, Vol. 40, Birkhäuser, Basel, 1989.

List of Georg Heinig's (refereed) publications

(chronologically ordered, including one monograph [85, 88], two edited proceedings [14, 33], and one book translation [30])

[1] G. Heinig, K. Rost, Split algorithms for centrosymmetric Toeplitz-plus-Hankel matrices with arbitrary rank profile, 129–146, Oper. Theory Adv. Appl., 171, Birkhäuser, Basel, 2007.

[2] G. Heinig, K. Rost, Schur-type algorithms for the solution of Hermitian Toeplitz systems via factorization, 233–252, Oper. Theory Adv. Appl., 160, Birkhäuser, Basel, 2005.

[3] G. Codevico, G. Heinig, M. Van Barel, A superfast solver for real symmetric Toeplitz systems using real trigonometric transformations. Numer. Linear Algebra Appl. 12 (2005), 699–713.

[4] G. Heinig, K. Rost, Fast "split" algorithms for Toeplitz and Toeplitz-plus-Hankel matrices with arbitrary rank profile. Proceedings of the International Conference on Mathematics and its Applications (ICMA 2004), 285–312, Kuwait, 2005.

[5] G. Heinig, K. Rost, Split algorithms for symmetric Toeplitz matrices with arbitrary rank profile. Numer. Linear Algebra Appl. 12 (2005), no. 2-3, 141–151.

[6] G. Heinig, K. Rost, Split algorithms for Hermitian Toeplitz matrices with arbitrary rank profile. Linear Algebra Appl. 392 (2004), 235–253.

[7] G. Heinig, Fast algorithms for Toeplitz least squares problems. Current trends in operator theory and its applications. 167–197, Oper. Theory Adv. Appl., 149, Birkhäuser, Basel, 2004.

[8] G. Heinig, K. Rost, Split algorithms for skewsymmetric Toeplitz matrices with arbitrary rank profile. Theoret. Comput. Sci. 315 (2004), no. 2-3, 453–468.

[9] G. Heinig, K. Rost, New fast algorithms for Toeplitz-plus-Hankel matrices. SIAM J. Matrix Anal. Appl. 25 (2003), no. 3, 842–857.

[10] G. Heinig, K. Rost, Fast algorithms for centrosymmetric and centro-skewsymmetric Toeplitz-plus-Hankel matrices. International Conference on Numerical Algorithms, Vol. I (Marrakesh, 2001), Numer. Algorithms 33 (2003), no. 1-4, 305–317.

[11] G. Heinig, Inversion of Toeplitz-plus-Hankel matrices with arbitrary rank profile. Fast algorithms for structured matrices: theory and applications (South Hadley, MA, 2001), 75–89, Contemp. Math., 323, Amer. Math. Soc., Providence, RI, 2003.

[12] M. Van Barel, G. Heinig, P. Kravanja, A superfast method for solving Toeplitz linear least squares problems. Special issue on structured matrices: analysis, algorithms and applications (Cortona, 2000), Linear Algebra Appl. 366 (2003), 441–457.

[13] G. Heinig, K. Rost, Centrosymmetric and centro-skewsymmetric Toeplitz-plus-Hankel matrices and Bezoutians. Special issue on structured matrices: analysis, algorithms and applications (Cortona, 2000), Linear Algebra Appl. 366 (2003), 257–281.

[14] Special issue on structured matrices: analysis, algorithms and applications. Papers from the workshop held in Cortona, September 21–28, 2000. eds. D. Bini, G. Heinig, E. Tyrtyshnikov. Linear Algebra Appl. 366 (2003). Elsevier Science B.V., Amsterdam, 2003.

[15] G. Heinig, K. Rost, Fast algorithms for skewsymmetric Toeplitz matrices. Toeplitz matrices and singular integral equations (Pobershau, 2001), 193–208, Oper. Theory Adv. Appl., 135, Birkhäuser, Basel, 2002.

[16] G. Heinig, On the reconstruction of Toeplitz matrix inverses from columns. Linear Algebra Appl. 350 (2002), 199–212.

[17] G. Heinig, K. Rost, Centro-symmetric and centro-skewsymmetric Toeplitz matrices and Bezoutians. Special issue on structured and infinite systems of linear equations. Linear Algebra Appl. 343/344 (2002), 195–209.

[18] G. Heinig, Kernel structure of Toeplitz-plus-Hankel matrices. Linear Algebra Appl. 340 (2002), 1–13.

[19] G. Heinig, Fast and superfast algorithms for Hankel-like matrices related to orthogonal polynomials. Numerical analysis and its applications (Rousse, 2000), 385–392, Lecture Notes in Comput. Sci., 1988, Springer, Berlin, 2001.

[20] M. Van Barel, G. Heinig, P. Kravanja, An algorithm based on orthogonal polynomial vectors for Toeplitz least squares problems. Numerical analysis and its applications (Rousse, 2000), 27–34, Lecture Notes in Comput. Sci., 1988, Springer, Berlin, 2001.

[21] M. Van Barel, G. Heinig, P. Kravanja, A stabilized superfast solver for nonsymmetric Toeplitz systems. SIAM J. Matrix Anal. Appl. 23 (2001), no. 2, 494–510.

[22] G. Heinig, K. Rost, Efficient inversion formulas for Toeplitz-plus-Hankel matrices using trigonometric transformations. Structured matrices in mathematics, computer science, and engineering, II (Boulder, CO, 1999), 247–264, Contemp. Math., 281, Amer. Math. Soc., Providence, RI, 2001.

[23] G. Heinig, Stability of Toeplitz matrix inversion formulas. Structured matrices in mathematics, computer science, and engineering, II (Boulder, CO, 1999), 101–116, Contemp. Math., 281, Amer. Math. Soc., Providence, RI, 2001.

[24] G. Heinig, V. Olshevsky, The Schur algorithm for matrices with Hessenberg displacement structure. Structured matrices in mathematics, computer science, and engineering, II (Boulder, CO, 1999), 3–15, Contemp. Math., 281, Amer. Math. Soc., Providence, RI, 2001.

[25] G. Heinig, Not every matrix is similar to a Toeplitz matrix. Proceedings of the Eighth Conference of the International Linear Algebra Society (Barcelona, 1999). Linear Algebra Appl. 332/334 (2001), 519–531.

[26] G. Heinig, Chebyshev-Hankel matrices and the splitting approach for centrosymmetric Toeplitz-plus-Hankel matrices. Linear Algebra Appl. 327 (2001), no. 1-3, 181–196.

[27] S. Feldmann, G. Heinig, Partial realization for singular systems in standard form. Linear Algebra Appl. 318 (2000), no. 1-3, 127–144.

[28] G. Heinig, K. Rost, Representations of inverses of real Toeplitz-plus-Hankel matrices using trigonometric transformations. Large-scale scientific computations of engineering and environmental problems, II (Sozopol, 1999), 80–86, Notes Numer. Fluid Mech., 73, Vieweg, Braunschweig, 2000.

[29] M. Van Barel, G. Heinig, P. Kravanja, Least squares solution of Toeplitz systems based on orthogonal polynomial vectors. Advanced Signal Processing Algorithms, Architectures, and Implementations X. ed. F.T. Luk, Vol 4116, Proceedings of SPIE (2000), 167-172.

[30] V. Maz'ya, S. Nazarov, B. Plamenevskij, Asymptotic theory of elliptic boundary value problems in singularly perturbed domains. Vol. I. Translated from the German by Georg Heinig and Christian Posthoff. Operator Theory: Advances and Applications, 111. Birkhäuser, Basel, 2000.

[31] G. Heinig, K. Rost, Hartley transform representations of symmetric Toeplitz matrix inverses with application to fast matrix-vector multiplication. SIAM J. Matrix Anal. Appl. 22 (2000), no. 1, 86–105.

[32] G. Heinig, K. Rost, Hartley transform representations of inverses of real Toeplitz-plus-Hankel matrices. Proceedings of the International Conference on Fourier Analysis and Applications (Kuwait, 1998). Numer. Funct. Anal. Optim. 21 (2000), no. 1-2, 175–189.

[33] Proceedings of the International Conference on Fourier Analysis and Applications. Held at Kuwait University, Kuwait, May 3–6, 1998, Eds. F. Al-Musallam, A. Böttcher, P. Butzer, G. Heinig, Vu Kim Tuan. Numer. Funct. Anal. Optim. 21 (2000), no. 1-2. Marcel Dekker, Inc., Monticello, NY, 2000.

[34] G. Heinig, F. Al-Musallam, Hermite's formula for vector polynomial interpolation with applications to structured matrices. Appl. Anal. 70 (1999), no. 3-4, 331–345.

[35] S. Feldmann, G. Heinig, Parametrization of minimal rank block Hankel matrix extensions and minimal partial realizations. Integral Equations Operator Theory 33 (1999), no. 2, 153–171.

[36] G. Heinig, K. Rost, DFT representations of Toeplitz-plus-Hankel Bezoutians with application to fast matrix-vector multiplication. ILAS Symposium on Fast Algorithms for Control, Signals and Image Processing (Winnipeg, MB, 1997). Linear Algebra Appl. 284 (1998), no. 1-3, 157–175.

[37] G. Heinig, Matrices with higher-order displacement structure. Linear Algebra Appl. 278 (1998), no. 1-3, 295–301.

[38] G. Heinig, A. Bojanczyk, Transformation techniques for Toeplitz and Toeplitz-plus-Hankel matrices. II. Algorithms. Linear Algebra Appl. 278 (1998), no. 1-3, 11–36.

[39] G. Heinig, Properties of "derived" Hankel matrices. Recent progress in operator theory (Regensburg, 1995), 155–170, Oper. Theory Adv. Appl., 103, Birkhäuser, Basel, 1998.

[40] G. Heinig, K. Rost, Representations of Toeplitz-plus-Hankel matrices using trigonometric transformations with application to fast matrix-vector multiplication. Proceedings of the Sixth Conference of the International Linear Algebra Society (Chemnitz, 1996). Linear Algebra Appl. 275/276 (1998), 225–248.

[41] G. Heinig, F. Al-Musallam, Lagrange's formula for tangential interpolation with application to structured matrices. Integral Equations Operator Theory 30 (1998), no. 1, 83–100.

[42] G. Heinig, Generalized Cauchy-Vandermonde matrices. Linear Algebra Appl. 270 (1998), 45–77.

[43] G. Heinig, The group inverse of the transformation $S(X) = 3DAX - XB$. Linear Algebra Appl. 257 (1997), 321–342.

[44] G. Heinig, A. Bojanczyk, Transformation techniques for Toeplitz and Toeplitz-plus-Hankel matrices. I. Transformations. Proceedings of the Fifth Conference of the International Linear Algebra Society (Atlanta, GA, 1995). Linear Algebra Appl. 254 (1997), 193–226.

[45] G. Heinig, L.A. Sakhnovich, I.F. Tidniuk, Paired Cauchy matrices. Linear Algebra Appl. 251 (1997), 189–214.

[46] G. Heinig, Solving Toeplitz systems after extension and transformation. Toeplitz matrices: structures, algorithms and applications (Cortona, 1996). Calcolo 33 (1998), no. 1-2, 115–129.

[47] S. Feldmann, G. Heinig, On the partial realization problem for singular systems. Proceedings of the 27th Annual Iranian Mathematics Conference (Shiraz, 1996), 79–100, Shiraz Univ., Shiraz, 1996.

[48] S. Feldmann, G. Heinig, Vandermonde factorization and canonical representations of block Hankel matrices. Proceedings of the Fourth Conference of the International Linear Algebra Society (ILAS) (Rotterdam, 1994). Linear Algebra Appl. 241/243 (1996), 247–278.

[49] G. Heinig, Inversion of generalized Cauchy matrices and other classes of structured matrices. Linear algebra for signal processing (Minneapolis, MN, 1992), 63–81, IMA Vol. Math. Appl., 69, Springer, New York, 1995.

[50] G. Heinig, Matrix representations of Bezoutians. Special issue honoring Miroslav Fiedler and Vlastimil Pták. Linear Algebra Appl. 223/224 (1995), 337–354.

[51] G. Heinig, K. Rost, Recursive solution of Cauchy-Vandermonde systems of equations. Linear Algebra Appl. 218 (1995), 59–72.

[52] G. Heinig, Generalized inverses of Hankel and Toeplitz mosaic matrices. Linear Algebra Appl. 216 (1995), 43–59.

[53] G. Heinig, F. Hellinger, Displacement structure of generalized inverse matrices. Generalized inverses (1993). Linear Algebra Appl. 211 (1994), 67–83.

[54] G. Heinig, F. Hellinger, The finite section method for Moore-Penrose inversion of Toeplitz operators. Integral Equations Operator Theory 19 (1994), no. 4, 419–446.

[55] G. Heinig, F. Hellinger, Displacement structure of pseudoinverses. Second Conference of the International Linear Algebra Society (ILAS) (Lisbon, 1992). Linear Algebra Appl. 197/198 (1994), 623–649.

[56] S. Feldmann, G. Heinig, Uniqueness properties of minimal partial realizations. Linear Algebra Appl. 203/204 (1994), 401–427.

[57] G. Heinig, F. Hellinger, Moore-Penrose inversion of square Toeplitz matrices. SIAM J. Matrix Anal. Appl. 15 (1994), no. 2, 418–450.

[58] A.W. Bojańczyk, G. Heinig, A multi-step algorithm for Hankel matrices. J. Complexity 10 (1994), no. 1, 142–164.

[59] G. Heinig, F. Hellinger, On the Bezoutian structure of the Moore-Penrose inverses of Hankel matrices. SIAM J. Matrix Anal. Appl. 14 (1993), no. 3, 629–645.

[60] T. Finck, G. Heinig, K. Rost, An inversion formula and fast algorithms for Cauchy-Vandermonde matrices. Linear Algebra Appl. 183 (1993), 179–191.

[61] G. Heinig, P. Jankowski, Kernel structure of block Hankel and Toeplitz matrices and partial realization. Linear Algebra Appl. 175 (1992), 1–30.

[62] G. Heinig, Inverse problems for Hankel and Toeplitz matrices. Linear Algebra Appl. 165 (1992), 1–23.

[63] G. Heinig, Fast algorithms for structured matrices and interpolation problems. Algebraic computing in control (Paris, 1991), 200–211, Lecture Notes in Control and Inform. Sci., 165, Springer, Berlin, 1991.

[64] G. Heinig, On structured matrices, generalized Bezoutians and generalized Christoffel-Darboux formulas. Topics in matrix and operator theory (Rotterdam, 1989), 267–281, Oper. Theory Adv. Appl., 50, Birkhäuser, Basel, 1991.

[65] G. Heinig, Formulas and algorithms for block Hankel matrix inversion and partial realization. Signal processing, scattering and operator theory, and numerical methods (Amsterdam, 1989), 79–90, Progr. Systems Control Theory, 5, Birkhäuser Boston, Boston, MA, 1990.

[66] G. Heinig, P. Jankowski, Parallel and superfast algorithms for Hankel systems of equations. Numer. Math. 58 (1990), no. 1, 109–127.

[67] G. Heinig, P. Jankowski, Fast algorithms for the solution of general Toeplitz systems. Wiss. Z. Tech. Univ. Karl-Marx-Stadt 32 (1990), no. 1, 12–17.

[68] G. Heinig, K. Rost, Matrices with displacement structure, generalized Bezoutians, and Moebius transformations. The Gohberg anniversary collection, Vol. I (Calgary, AB, 1988), 203–230, Oper. Theory Adv. Appl., 40, Birkhäuser, Basel, 1989.

[69] G. Heinig, K. Rost, Inversion of matrices with displacement structure. Integral Equations Operator Theory 12 (1989), no. 6, 813–834.

[70] G. Heinig, W. Hoppe, K. Rost, Structured matrices in interpolation and approximation problems. Wiss. Z. Tech. Univ. Karl-Marx-Stadt 31 (1989), no. 2, 196–202.

[71] G. Heinig, K. Rost, Matrix representations of Toeplitz-plus-Hankel matrix inverses. Linear Algebra Appl. 113 (1989), 65–78.

[72] G. Heinig, T. Amdeberhan, On the inverses of Hankel and Toeplitz mosaic matrices. Seminar Analysis (Berlin, 1987/1988), 53–65, Akademie-Verlag, Berlin, 1988.

[73] G. Heinig, P. Jankowski, K. Rost, Tikhonov regularisation for block Toeplitz matrices. Wiss. Z. Tech. Univ. Karl-Marx-Stadt 30 (1988), no. 1, 41–45.

[74] G. Heinig, K. Rost, On the inverses of Toeplitz-plus-Hankel matrices. Linear Algebra Appl. 106 (1988), 39–52.

[75] G. Heinig, P. Jankowski, K. Rost, Fast inversion algorithms of Toeplitz-plus-Hankel matrices. Numer. Math. 52 (1988), no. 6, 665–682.

[76] G. Heinig, Structure theory and fast inversion of Hankel striped matrices. I. Integral Equations Operator Theory 11 (1988), no. 2, 205–229.

[77] G. Heinig, K. Rost, Inversion of generalized Toeplitz-plus-Hankel matrices. Wiss. Z. Tech. Univ. Karl-Marx-Stadt 29 (1987), no. 2, 209–211.

[78] G. Heinig, U. Jungnickel, Hankel matrices generated by Markov parameters, Hankel matrix extension, partial realization, and Padé-approximation. Operator theory and systems (Amsterdam, 1985), 231–253, Oper. Theory Adv. Appl., 19, Birkhäuser, Basel, 1986.

[79] G. Heinig, U. Jungnickel, Lyapunov equations for companion matrices. Linear Algebra Appl. 76 (1986), 137–147.

[80] G. Heinig, U. Jungnickel, Hankel matrices generated by the Markov parameters of rational functions. Linear Algebra Appl. 76 (1986), 121–135.

[81] G. Heinig, Partial indices for Toeplitz-like operators. Integral Equations Operator Theory 8 (1985), no. 6, 805–824.

[82] G. Heinig, K. Rost, Fast inversion of Toeplitz-plus-Hankel matrices. Wiss. Z. Tech. Hochsch. Karl-Marx-Stadt 27 (1985), no. 1, 66–71.

[83] G. Heinig, U. Jungnickel, On the Bezoutian and root localization for polynomials. Wiss. Z. Tech. Hochsch. Karl-Marx-Stadt 27 (1985), no. 1, 62–65.

[84] G. Heinig, B. Silbermann, Factorization of matrix functions in algebras of bounded functions. Spectral theory of linear operators and related topics (Timişoara/Herculane, 1983), 157–177, Oper. Theory Adv. Appl., 14, Birkhäuser, Basel, 1984.

[85] G. Heinig, K. Rost, Algebraic methods for Toeplitz-like matrices and operators. Oper. Theory Adv. Appl., 13, Birkhäuser, Basel, 1984.

[86] G. Heinig, U. Jungnickel, On the Routh-Hurwitz and Schur-Cohn problems for matrix polynomials and generalized Bezoutians. Math. Nachr. 116 (1984), 185–196.

[87] G. Heinig, K. Rost, Schnelle Invertierungsalgorithmen für einige Klassen von Matrizen. (German) [Fast inversion algorithms for some classes of matrices] Wiss. Z. Tech. Hochsch. Karl-Marx-Stadt 26 (1984), no. 2, 235–241.

[88] G. Heinig, K. Rost, Algebraic methods for Toeplitz-like matrices and operators. Mathematical Research, 19, Akademie-Verlag, Berlin, 1984.

[89] G. Heinig, K. Rost, Invertierung von Toeplitzmatrizen und ihren Verallgemeinerungen. I. Die Methode der UV-Reduktion. (German) [Inversion of Toeplitz matrices and their generalizations. I. The method of UV-reduction] Beiträge Numer. Math. No. 12 (1984), 55–73.

[90] G. Heinig, Generalized resultant operators and classification of linear operator pencils up to strong equivalence. Functions, series, operators, Vol. I, II (Budapest, 1980), 611–620, Colloq. Math. Soc. János Bolyai, 35, North-Holland, Amsterdam, 1983.

[91] G. Heinig, Inversion of Toeplitz and Hankel matrices with singular sections. Wiss. Z. Tech. Hochsch. Karl-Marx-Stadt 25 (1983), no. 3, 326–333.

[92] G. Heinig, U. Jungnickel, Zur Lösung von Matrixgleichungen der Form $AX - XB = 3DC$. (German) [On the solution of matrix equations of the form $AX - XB = 3DC$] Wiss. Z. Tech. Hochsch. Karl-Marx-Stadt 23 (1981), no. 4, 387–393.

[93] G. Heinig, Linearisierung und Realisierung holomorpher Operatorfunktionen. (German) Wiss. Z. Tech. Hochsch. Karl-Marx-Stadt 22 (1980), no. 5, 453–459.

[94] G. Heinig, K. Rost, Invertierung einiger Klassen von Matrizen und Operatoren. I. Endliche Toeplitzmatrizen und ihre Verallgemeinerungen. (German) [Inversion of some classes of matrices and operators. I. Finite Toeplitz matrices and their generalizations] Wissenschaftliche Informationen [Scientific Information], 12, Technische Hochschule Karl-Marx-Stadt, Sektion Mathematik, Karl-Marx-Stadt, 1979.

[95] G. Heinig, Bezoutiante, Resultante und Spektralverteilungsprobleme für Operatorpolynome. (German) Math. Nachr. 91 (1979), 23–43.

[96] G. Heinig, Transformationen von Toeplitz- und Hankelmatrizen. (German) Wiss. Z. Tech. Hochsch. Karl-Marx-Stadt 21 (1979), no. 7, 859–864.

[97] G. Heinig, Invertibility of singular integral operators. (Russian) Soobshch. Akad. Nauk Gruzin. SSR 96 (1979), no. 1, 29–32.

[98] G. Heinig, Über ein kontinuierliches Analogon der Begleitmatrix eines Polynoms und die Linearisierung einiger Klassen holomorpher Operatorfunktionen. (German) Beiträge Anal. 13 (1979), 111–126.

[99] G. Heinig, Verallgemeinerte Resultantenbegriffe bei beliebigen Matrixbüscheln. II. Gemischter Resultantenoperator. (German) Wiss. Z. Tech. Hochsch. Karl-Marx-Stadt 20 (1978), no. 6, 701–703.

[100] G. Heinig, Verallgemeinerte Resultantenbegriffe bei beliebigen Matrixbüscheln. I. Einseitiger Resultantenoperator. (German) Wiss. Z. Tech. Hochsch. Karl-Marx-Stadt 20 (1978), no. 6, 693–700.

[101] G. Heinig, Endliche Toeplitzmatrizen und zweidimensionale Wiener-Hopf-Operatoren mit homogenem Symbol. II. Über die normale Auflösbarkeit einer Klasse zweidimensionaler Wiener-Hopf Operatoren. (German) Math. Nachr. 82 (1978), 53–68.

[102] G. Heinig, Endliche Toeplitzmatrizen und zweidimensionale Wiener-Hopf-Operatoren mit homogenem Symbol. I. Eigenschaften endlicher Toeplitzmatrizen. (German) Math. Nachr. 82 (1978), 29–52.

[103] G. Heinig, K. Rost, Über homogene Gleichungen vom Faltungstyp auf einem endlichen Intervall. (German) Demonstratio Math. 10 (1977), no. 3-4, 791–806.

[104] G. Heinig, Über Block-Hankelmatrizen und den Begriff der Resultante für Matrixpolynome. (German) Wiss. Z. Techn. Hochsch. Karl-Marx-Stadt 19 (1977), no. 4, 513–519.

[105] G. Heinig, The notion of Bezoutian and of resultant for operator pencils. (Russian) Funkcional. Anal. i Priložen. 11 (1977), no. 3, 94–95.

[106] I.C. Gohberg, G. Heinig, The resultant matrix and its generalizations. II. The continual analogue of the resultant operator. (Russian) Acta Math. Acad. Sci. Hungar. 28 (1976), no. 3-4, 189–209.

[107] G. Heinig, Periodische Jacobimatrizen im entarteten Fall. (German) Wiss. Z. Techn. Hochsch. Karl-Marx-Stadt 18 (1976), no. 4, 419–423.

[108] G. Heinig, Über die Invertierung und das Spektrum von singulären Integraloperatoren mit Matrixkoeffizienten. (German) 5. Tagung über Probleme und Methoden der Mathematischen Physik (Techn. Hochschule Karl-Marx-Stadt, Karl-Marx-Stadt, 1975), Heft 1, 52–59. Wiss. Schr. Techn. Hochsch. Karl-Marx-Stadt, Techn. Hochsch., Karl-Marx-Stadt, 1975.

[109] I.C. Gohberg, G. Heinig, On matrix integral operators on a finite interval with kernels depending on the difference of the arguments. (Russian) Rev. Roumaine Math. Pures Appl. 20 (1975), 55–73.

[110] I.C. Gohberg, G. Heinig, The resultant matrix and its generalizations. I. The resultant operator for matrix polynomials. (Russian) Acta Sci. Math. (Szeged) 37 (1975), 41–61.

[111] I.C. Gohberg, G. Heinig, Inversion of finite Toeplitz matrices consisting of elements of a noncommutative algebra. (Russian) Rev. Roumaine Math. Pures Appl. 19 (1974), 623–663.

[112] G. Heinig, The inversion and the spectrum of matrix Wiener-Hopf operators. (Russian) Mat. Sb. (N.S.) 91(133) (1973), 253–266.

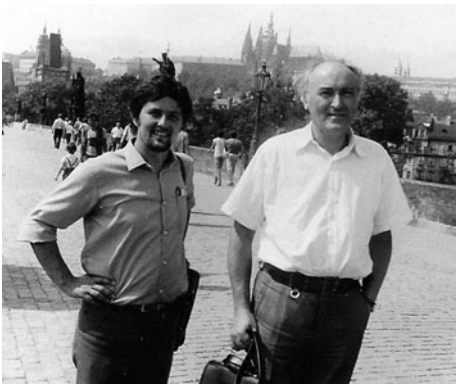
[113] G. Heinig, The inversion and the spectrum of matrix-valued singular integral operators. (Russian) Mat. Issled. 8 (1973), no. 3(29), 106–121.

[114] I.C. Gohberg, G. Heinig, The inversion of finite Toeplitz matrices. (Russian) Mat. Issled. 8 (1973), no. 3(29), 151–156.

[115] G. Heinig, Inversion of periodic Jacobi matrices. (Russian) Mat. Issled. 8 (1973), no. 1(27), 180–200.

Karla Rost
Department of Mathematics
Chemnitz University of Technology
D-09107 Chemnitz, Germany
krost@mathematik.tu-chemnitz.de

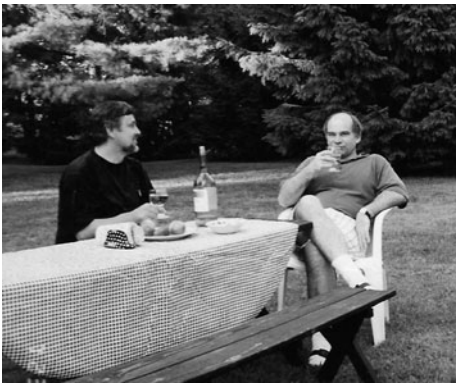
With Israel Gohberg



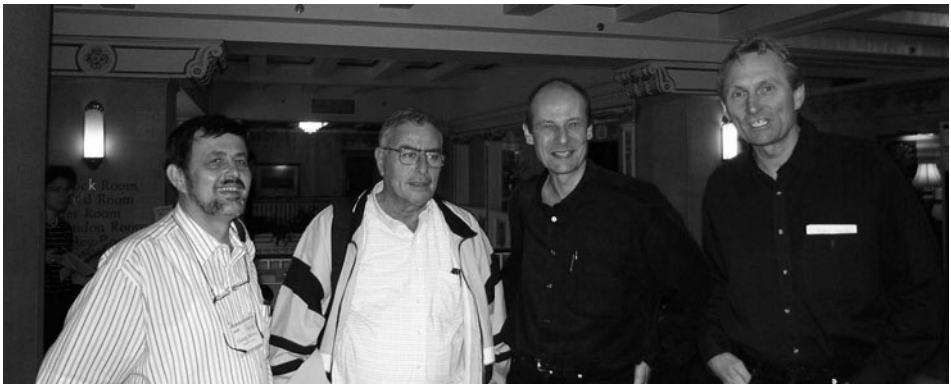
With Vlastimil Ptak

With Eric Kaltofen, Lothar Reichel and Dario Bini

With Israel Gohberg and Karla Rost



With Adam Bojanzcyk



With Paul Furmann, Uwe Helmke and Volker Mehrmann

With Dario Bini

With Paul Van Doren

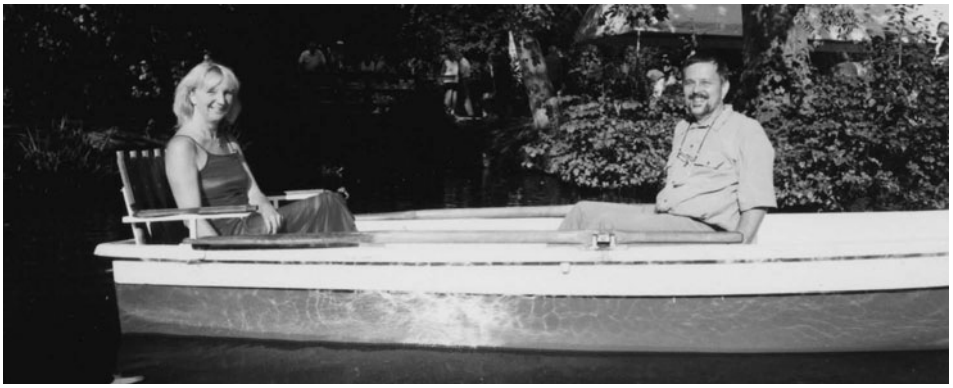
With Victor Pan and Vadim Olshevsky

In Kuwait

Researcher of the Year award, 2002

With Gerti

With Children and Grandchildren



With Gerti

In Ethiopia

Operator Theory:
Advances and Applications, Vol. 199, 25–118

Introduction to Bezoutians

Georg Heinig and Karla Rost

Mathematics Subject Classification (2000). Primary 15-01; Secondary 15A09, 15A23, 65F05.

Keywords. Bezoutian, Toeplitz matrix inverses, Hankel matrix inverses, resultant matrix, Toeplitz-plus-Hankel matrix inverses.

Foreword

In the present paper we consider classes of matrices the entries of which are in a given field $\mathbb{F}$. These matrices have a special structure, they are *Bezoutians*. Historically, Bezoutians were at first introduced in connection with the elimination for the solution of systems of nonlinear algebraic equations and in connection with root localization problems. Only much later their importance for Hankel and Toeplitz matrix inversion became clear.

We will introduce three kinds of Bezoutians: Toeplitz Bezoutians, Hankel Bezoutians, and Toeplitz-plus-Hankel Bezoutians. The classes of Toeplitz and Hankel Bezoutians are related to Toeplitz and Hankel matrices in two ways. First, the inverses of Toeplitz and Hankel matrices are Toeplitz and Hankel Bezoutians, respectively. Furthermore, in case where $\mathbb{F} = \mathbb{C}$, Hermitian Toeplitz and Hankel Bezoutians are congruent to Toeplitz and Hankel matrices. The class of Toeplitz-plus-Hankel Bezoutians includes the inverses of Toeplitz-plus-Hankel matrices. Instead of a summary of the content we will offer the table of contents at the end of this foreword.

The present paper is not a usual paper. It originated from the draft of one chapter of a text-book on structured matrices planned by both authors. This textbook for graduate students was intended to range from the basics for beginners up to recent investigations. At the beginning of 2005 the outlines for the first three chapters were ready and parts of the text were in an acceptable form when Georg Heinig, the head of this project, unexpectedly died of a heart attack on May 10, 2005. We have lost one of the top experts in the field of structured matrices. His death reveals a gap we cannot overcome. This is the tragedy of the planned book and also of the present paper.

In the last period of our cooperation that had lasted 30 years we mainly worked on the third chapter of the textbook, which was dedicated to Bezoutians, so that I think that this part of the book was perhaps the favorite "child" of Georg.

Thus I felt obliged to continue and complete this text to achieve a self-contained, improved version which can be published separately. I started with a preliminary section to make the presentation more selfcontained. Then I corrected and completed the other sections. Since the Toeplitz-plus-Hankel case was not included, I added main results concerning this case in Sections 11 and 12. Moreover, I finished, as planned, with exercises – part of which were already discussed with Georg – and then I make some short historical notes and provide hints to literature pursuing and accentuating the topic in different directions.

I hope I was able to do all these things in such a way that Georg would not be ashamed. In fact it is a hard burden of responsibility for me, in particular, since Georg was an outstanding mathematician with excellent abilities in teaching and in writing papers.

A further reason for this paper is that the topic of Bezoutians is very nice, interesting and important with a lot of connections and applications. In the last few years, one can even observe a revival of the interest in Bezoutians, mainly motivated by their importance in many modern fields such as numerical computing and control theory. Thus it would be very useful to have an introductional paper into this topic, where a lot of properties and relations are systematically collected and explained.

I neither intend to quote a huge number of relevant papers nor to mention all corresponding generalizations and applications. What I try and do is to appreciate Georg's contributions, since his legacy concerning this topic is enormously.

In 1971 Georg started his PhD studies at the State University of Moldavia in Kishinev under the supervision of Israel Gohberg. By this time he was irretrievably cast into the realm of structured matrices, in particular of Toeplitz and Hankel matrices as well as Bezoutians. His very early joint papers with I. Gohberg [12], [13] dealt with the inversion of finite Toeplitz matrices, the papers [14], [15], [16], [17], [18], [19], [20], [21] were dedicated to Bezoutians and resultant matrices mainly for operator-valued polynomials or to continual analogs of resultants and Bezoutians. The main results of these papers were milestones of the research in this field.

In 1975 Georg waked up my interests in the topic of Toeplitz and Hankel matrices and their inverses. Thus, in 1981, I wrote a large part of my PhD thesis on the method of UV-reduction for inverting structured matrices under his excellent supervision. When we started our work on the book "Algebraic Methods for Toeplitz-like Matrices and Operators" [32], [33] he introduced me, in particular, into the wonderful world of Bezoutians. In this book, Section 2 of Part I is dedicated to Bezoutians and resultant matrices. Some of the results presented there are, of course, also offered here but, as the result of new thoughts about the matter, from another point of view.

Moreover, in Subsection 2.2, Part II of [33] we present first ideas and results concerning matrices which are the sum of a Toeplitz and a Hankel matrix (briefly

T+H matrices). In my opinion, one of our most important joint result is that in 1986 we discovered a Bezoutian structure also for inverses of T+H matrices (see [34]). This was the starting point of a long interesting and fruitful joint work on these special cases of structured matrices In fact, until now I feel a motivation given by Georg to deal with the T+H case (see [48], [64]).

Beginning with the joint paper [37] we wrote a number of papers on matrix representations for T+H matrices and their inverses which allow fast matrix-vector multiplication (see, e.g., [39], [38], [40], [42], [43]).

Then we dealt with the problem how to connect the Toeplitz or T+H structure of matrices with possibly additional symmetries in order to reduce the number of parameters involved in these formulas or in the corresponding algorithms. Georg's paper [22] showed that splitting ideas in the spirit of Delsarte and Genin were very promising. The splitting approaches of our joint papers [44], [45], [46] differ from those of [22].

(Note that in [34] the concept of ω-structured matrices was introduced as a generalization of matrices possessing a Toeplitz, Hankel, or T+H structure. This class of and further investigated in [35], [36]. But these considerations are not included in this paper.)

I was only one of a large number of Georg's coauthors and pupils. In particular, Uwe Jungnickel wrote his PhD thesis under Georg's supervision in 1986. In their joint papers [27], [28] they considered Routh-Hurwitz or Schur-Cohn problems of counting the roots of a given polynomial in a half-plane or in a circle. They investigated Hankel matrices generated by the Markov parameters of rational functions and their importance for partial realization and Pade approximation in [30], [29]. They investigated the connection of Bezoutians and resultant matrices for the solution of matrix equations in [26], [31].

In Section 7 of Part I of [33] some first results concerning the Bezoutian structure of generalized inverses are presented. Georg continued this investigation together with his student Frank Hellinger. Their results published in [23], [25], [24] have found perpetual interest by a large community. Since they go beyond the scope of the present paper they are not included.

It is not possible to recognize the full extent and importance of Georg's work concerning Bezoutians. I beg your pardon for all I will forget to mention or I will not appreciate to the due extend.

Karla Rost

Contents

1. Preliminaries

1. Notation. Throughout the paper, $\mathbb{F}$ will denote an arbitrary field. In some sections we restrict ourselves to the case that $\mathbb{F} = \mathbb{C}$ or $\mathbb{R}$, the fields of complex or real numbers, respectively. By $\{ \mathbf{e}_1, \dots, \mathbf{e}_n \}$ the standard basis of $\mathbb{F}^n$ is denoted. Furthermore, $\mathbf{0}_k$ will stand for a zero vector of length k. If there is no danger of misunderstanding we will omit the subscript k.

As usual, an element of the vector space $\mathbb{F}^n$ will be identified with the corresponding $n \times 1$ (column) matrix. That means

$$(x_i)_{i=1}^n = (x_1, \dots, x_n) = \begin{bmatrix} x_1 \\ \vdots \\ x_n \end{bmatrix} .$$

In all what follows we denote by $\ell_n(t)$, $t \in \mathbb{F}$, the vector

$$\ell_n(t) = (1, t, t^2, \dots, t^{n-1}) . \tag{1.1}$$

The Bezoutian concept is convenient introduced in polynomial language. First we introduce "polynomial language" for vectors. For $\mathbf{x} = (x_i)_{i=1}^n \in \mathbb{F}^n$, we consider the polynomial

$$\mathbf{x}(t) = \ell_n(t)^T \mathbf{x} = \sum_{k=1}^{n} x_k t^{k-1} \in \mathbb{F}^n(t)$$

and call it *generating polynomial* of $\mathbf{x}$. Polynomial language for matrices means that we introduce the *generating polynomial of an $m \times n$ matrix* $A = [\, a_{ij} \,]_{i=1,j=1}^{m \quad n} \in \mathbb{F}^{m \times n}$ as the bivariate polynomial

$$A(t, s) = \ell_m(t)^T A \ell_n(s) = \sum_{i=1}^{m} \sum_{j=1}^{n} a_{ij}\, t^{i-1} s^{j-1}.$$

At several places in this paper we will exploit symmetry properties of matrices. Besides symmetry, skewsymmetry and Hermitian symmetry in the usual sense we also deal with persymmetry and centrosymmetry. To be more precise we introduce some notations. Let J_n be the matrix of the flip operator in $\mathbb{F}^n$ mapping $(x_1, x_2, \dots, x_n)$ to $(x_n, x_{n-1}, \dots, x_1)$,

$$J_n = \begin{bmatrix} 0 & & 1 \\ & \iddots & \\ 1 & & 0 \end{bmatrix} . \tag{1.2}$$

For a vector $\mathbf{x} \in \mathbb{F}^n$ we denote by $\mathbf{x}^J$ the vector $J_n \mathbf{x}$ and, in case $\mathbb{F} = \mathbb{C}$, by $\mathbf{x}^\#$ the vector $J_n \overline{\mathbf{x}}$, where $\overline{\mathbf{x}}$ is the vector with the conjugate complex entries,

$$\mathbf{x}^J = J_n \mathbf{x} \quad \text{and} \quad \mathbf{x}^\# = J_n \overline{\mathbf{x}} .$$

In polynomial language the latter looks like

$$\mathbf{x}^J(t) = \mathbf{x}(t^{-1}) t^{n-1}, \quad \mathbf{x}^\#(t) = \overline{\mathbf{x}}(t^{-1}) t^{n-1} .$$

A vector is called *symmetric* if $\mathbf{x}^J = \mathbf{x}$, *skewsymmetric* if $\mathbf{x}^J = -\mathbf{x}$, and *conjugate symmetric* if $\mathbf{x}^\# = \mathbf{x}$. Let $\mathbb{F}^n_+(\mathbb{F}^n_-)$ denote the subspace of all symmetric (skewsymmetric) vectors of $\mathbb{F}^n$, and let $P_\pm$ be the matrices

$$P_\pm = \frac{1}{2}(I_n \pm J_n)\,. \tag{1.3}$$

These matrices are projections onto $\mathbb{F}^n_\pm$ and

$$P_+ + P_- = I_n, \quad P_+ - P_- = J_n\,.$$

For an $n \times n$ matrix A, we denote

$$A^J = J_n A J_n \quad \text{and} \quad A^\# = J_n \overline{A} J_n\,,$$

where $\overline{A}$ is the matrix with the conjugate complex entries. An $n \times n$ matrix A is called *persymmetric* if $A^J = A^T$. The matrix A is called *centrosymmetric* if $A^J = A$. It is called *centro-skewsymmetric* if $A^J = -A$ and *centro-Hermitian* if $A^\# = A$.

2. Sylvester's inertia law. Assume that $\mathbb{F} = \mathbb{C}$. Let A be an Hermitian $n \times n$ matrix. The triple of integers

$$\operatorname{In} A = (p_+, p_-, p_0)$$

in which p_+ is the number of positive, p_- the number of negative, and p_0 the number of zero eigenvalues, counting multiplicities, is called the *inertia of* A. Clearly $p_+ + p_- + p_0 = n$. The integer

$$\operatorname{sgn} A = p_+ - p_-$$

is called the *signature of* A. Note that $p_- + p_+$ is the rank of A, so that rank and signature of an Hermitian matrix determine its inertia.

Two Hermitian $n \times n$ matrices A and B are called *congruent* if there is a nonsingular matrix C such that $B = C^* A C$, where C^* denotes the conjugate transpose of C. The following is *Sylvester's inertia law*, which will frequently be applied in this paper.

Theorem 1.1. *Congruent matrices have the same inertia.*

We will often apply the following version of Sylvester's inertia law.

Corollary 1.2. *Let A be an Hermitian $m \times m$ matrix and C an $m \times n$ matrix with $m \le n$ and* $\operatorname{rank} C = m$. *Then the signatures of A and $C^* A C$ coincide.*

To see that Corollary 1.2 follows from Theorem 1.1 we extend C to a nonsingular $n \times n$ matrix $\widetilde{C}$ by adding rows at the bottom. Then $C^* A C = \widetilde{C}^* \widetilde{A} \widetilde{C}$, where $\widetilde{A}$ is the extension of A by $n - m$ zero columns and zero rows on the right and at the bottom, respectively. This means that $C^* A C$ is congruent to $\widetilde{A}$, and thus $\operatorname{sgn} C^* A C = \operatorname{sgn} \widetilde{A} = \operatorname{sgn} A$.

3. Toeplitz, Hankel, and Toeplitz-plus-Hankel matrices. Let $\mathcal{T}_{mn}$ be the subspace of $\mathbb{F}^{m \times n}$ consisting of all $m \times n$ Toeplitz matrices

$$T_{mn}(\mathbf{a}) = [\, a_{i-j} \,]_{i=1,j=1}^{m \quad n}, \quad \mathbf{a} = (a_i)_{i=1-n}^{m-1} \in \mathbb{F}^{m+n-1}.$$

The subspace of all $m \times n$ Hankel matrices

$$H_{mn}(\mathbf{s}) = [\, s_{i+j-1} \,]_{i=1,j=1}^{m \quad n}, \quad \mathbf{s} = (s_i)_{i=1}^{m+n-1} \in \mathbb{F}^{m+n-1}$$

is denoted by $\mathcal{H}_{mn}$. The dimension of these subspaces is $m+n-1$. The intersection $\mathcal{T}_{mn} \cap \mathcal{H}_{mn}$ consists of all *chess-board matrices*,

$$B = \begin{bmatrix} c & b & c & \cdots \\ b & c & b & \cdots \\ c & b & c & \cdots \\ \vdots & \vdots & \vdots & \end{bmatrix} \qquad (c, b \in \mathbb{F}) \tag{1.4}$$

which form a two-dimensional subspace of $\mathbb{F}^{m\times n}$. The subspace of all $m\times n$ matrices R_{mn} which are the sum of a Toeplitz and a Hankel matrix (briefly *T+H matrices*)

$$R_{mn} = T_{mn}(\mathbf{a}) + H_{mn}(\mathbf{s})$$

is $2(m + n - 2)$ dimensional. Since for an $m \times n$ Hankel matrix H_{mn} the matrix $H_{mn}J_n$ is Toeplitz any T+H matrix can be represented in the form

$$R_{mn} = T_{mn}(\mathbf{a}) + T_{mn}(\mathbf{b})J_n \qquad (\mathbf{a}, \mathbf{b} \in \mathbb{F}^{m+n-1}). \tag{1.5}$$

From this another representation is derived involving the projections $P_\pm$ introduced in (1.3),

$$R_{mn} = T_{mn}(\mathbf{c})P_+ + T_{mn}(\mathbf{d})P_- \tag{1.6}$$

with $\mathbf{c} = \mathbf{a}+\mathbf{b}, \mathbf{d} = \mathbf{a}-\mathbf{b}$. Obviously, all these representations are not unique (see Exercises 15 and 16).

4. Quasi-Toeplitz matrices, quasi-Hankel matrices, and quasi-T+H matrices. We consider the transformation ∇_+ in the space of $n \times n$ matrices defined by

$$\nabla_+(A) = A - S_n A S_n^T, \tag{1.7}$$

where S_n is the forward shift in $\mathbb{F}^n$ mapping $(x_1, x_2, \ldots, x_n)$ to $(0, x_1, \ldots, x_{n-1})$,

$$S_n = \begin{bmatrix} 0 & 0 & & 0 \\ 1 & 0 & & 0 \\ & \ddots & \ddots & \\ 0 & & 1 & 0 \end{bmatrix}. \tag{1.8}$$

It can easily be checked that this transformation is one-to-one. The transformation ∇_+ is called *shift displacement operator*. For a Toeplitz matrix $T_n = [\, a_{i-j} \,]_{i,j=1}^n$ we have, obviously,

$$\nabla_+(T_n) = \begin{bmatrix} a_0' & 1 \\ a_1 & 0 \\ \vdots & \vdots \\ a_{n-1} & 0 \end{bmatrix} \begin{bmatrix} a_0' & a_{-1} & \cdots & a_{1-n} \\ 1 & 0 & \cdots & 0 \end{bmatrix},$$

where $a_0' = \frac{1}{2}\, a_0$. In particular, the rank of $\nabla_+(T_n)$ equals 2, unless T_n is triangular. In the latter case the rank of $\nabla_+(T_n)$ equals 1, unless $T_n = O$.

Notice that if T_n is Hermitian, then $\nabla_+(T_n)$ is also Hermitian, and the signature of $\nabla_+(T_n)$ equals zero, unless T_n is diagonal. (Obviously, T_n diagonal means $T_n = a_0 I_n$ and $\operatorname{sgn}(\nabla_+(T_n))$ equals the signum of a_0.)

Moreover, a matrix A is Toeplitz if and only if the $(n-1)\times(n-1)$ submatrix in the lower right corner of $\nabla_+(A)$ is the zero matrix. An $n \times n$ matrix A is called *quasi-Toeplitz* if $\operatorname{rank} \nabla_+(A) \le 2$.

Clearly, Toeplitz matrices are also quasi-Toeplitz, but not vice versa. The following proposition gives a complete description of quasi-Toeplitz matrices. Since the proof is an elementary calculation, we leave it to the reader.

Proposition 1.3. *Suppose that* $\nabla_+(A) = \mathbf{g}_+\mathbf{g}_-^T - \mathbf{h}_+\mathbf{h}_-^T$, $\mathbf{g}_\pm = (g_i^\pm)_{i=1}^n$, $\mathbf{h}_\pm = (h_i^\pm)_{i=1}^n$. *Then* A *can be represented as the sum of* 2 *products of triangular Toeplitz matrices,*

$$A = \begin{bmatrix} g_1^+ & & 0 \\ \vdots & \ddots & \\ g_n^+ & \dots & g_1^+ \end{bmatrix} \begin{bmatrix} g_1^- & \dots & g_n^- \\ & \ddots & \vdots \\ 0 & & g_1^- \end{bmatrix} - \begin{bmatrix} h_1^+ & & 0 \\ \vdots & \ddots & \\ h_n^+ & \dots & h_1^+ \end{bmatrix} \begin{bmatrix} h_1^- & \dots & h_n^- \\ & \ddots & \vdots \\ 0 & & h_1^- \end{bmatrix}. \tag{1.9}$$

Conversely, if A *is given by* (1.9), *then* $\nabla_+(A) = \mathbf{g}_+\mathbf{g}_-^T - \mathbf{h}_+\mathbf{h}_-^T$.

Analogously, we consider the transformation $\nabla^+ : \mathbb{F}^n \longrightarrow \mathbb{F}^n$ defined by

$$\nabla^+(A) = S_n A - A S_n^T. \tag{1.10}$$

A matrix A is Hankel if and only if the $(n-1)\times(n-1)$ submatrix in the lower right corner of $\nabla^+(A)$ is the zero matrix. We call a matrix A *quasi-Hankel* if $\operatorname{rank} \nabla^+(A) \le 2$. A similar representation to (1.9) can be obtained.

Let W_n be the matrix $W_n = S_n + S_n^T$, and let $\nabla : \mathbb{F}^n \longrightarrow \mathbb{F}^n$ be defined by

$$\nabla(A) = AW_n - W_n A. \tag{1.11}$$

Proposition 1.4. *A matrix* A *is a T+H matrix if and only if the* $(n-2)\times(n-2)$ *submatrix in the center of* $\nabla(A)$ *is the zero matrix.*

We call a matrix A *quasi-T+H* if $\operatorname{rank} \nabla(A) \le 4$. T+H matrices are also quasi-T+H, but not vice versa.

5. Möbius transformations. The flip operator J_n introduced in (1.2) is a special case of a class of operators which will be described in this subsection. Let $\phi = \begin{bmatrix} a & c \\ b & d \end{bmatrix}$ be a nonsingular 2×2 matrix with entries from $\mathbb{F}$. We associate ϕ with the linear fractional function

$$\phi(t) = \frac{at+b}{ct+d}.$$

Despite we use the name "function", $\phi(t)$ is understood here in a formal sense, i.e., t is considered as an abstract variable. In the case where $\mathbb{F} = \mathbb{C}$, $\phi(t)$ can be seen as a function mapping the Riemann sphere onto itself. These linear fractional functions form a group $\mathcal{M}$ with respect to composition. This group is *isomorphic*, modulo multiples of I_2, to the group $\mathrm{GL}(\mathbb{F}^2)$ of nonsingular 2×2 matrices. The

latter means that if $\phi = \phi_1\phi_2$, then $\phi(t) = \phi_2(\phi_1(t))$, and $\phi(t) = t$ if and only if $\phi = \alpha I_2$ for some $\alpha \in \mathbb{F}$.

We will make use of the fact that the group $\mathrm{GL}(\mathbb{F}^2)$ is generated by matrices of the form

$$\text{(a)} \quad \begin{bmatrix} a & 0 \\ 0 & 1 \end{bmatrix} \ (a \neq 0), \quad \text{(b)} \quad \begin{bmatrix} 1 & 0 \\ b & 1 \end{bmatrix}, \quad \text{(c)} \quad \begin{bmatrix} 0 & 1 \\ 1 & 0 \end{bmatrix}. \tag{1.12}$$

For $\phi \in \mathrm{GL}(\mathbb{F}^2)$ and a natural number n, let $K_n(\phi)$ denote the operator defined by

$$K_n(\phi)\mathbf{x}(t) = \mathbf{x}(\phi(t))(ct + d)^{n-1}$$

for $\mathbf{x}(t) \in \mathbb{F}^n(t)$. An operator of this form will be called *Möbius transformation*. It is easily checked that $K_n(\phi)$ maps $\mathbb{F}^n(t)$ into itself and is linear. In the special cases (1.12) we have

$$\text{(a) } K_n(\phi)\mathbf{x}(t) = \mathbf{x}(at), \ \text{(b) } K_n(\phi)\mathbf{x}(t) = \mathbf{x}(t + b), \ \text{(c) } K_n(\phi)\mathbf{x}(t) = \mathbf{x}(t^{-1})t^{n-1}. \tag{1.13}$$

The matrix representations of these transformations (called Möbius matrices) with respect to the standard basis in $\mathbb{F}^n(t)$ are

$$\text{(a) } K_n(\phi) = \mathrm{diag}(a^j)_{j=0}^{n-1}, \ \text{(b) } K_n(\phi) = \left[\binom{k}{j} b^{k-j} \right]_{j,k=0}^{n-1}, \ \text{(c) } K_n(\phi) = J_n. \tag{1.14}$$

Furthermore, the following is true.

Proposition 1.5. *If $\phi_1, \phi_2 \in \mathrm{GL}(\mathbb{F}^2)$ and $\phi = \phi_1\phi_2$, then $K_n(\phi) = K_n(\phi_1)K_n(\phi_2)$.*

It is sufficient to prove the proposition for the special matrices (1.12). We leave this to the reader.

According to Proposition 1.5 the Möbius transformations are all invertible and form a subgroup of the group of invertible linear operators on $\mathbb{F}^n$. Furthermore, $K_n(\phi)^{-1} = K_n(\phi^{-1})$. Möbius matrices are mostly used in connection with matrix transformations of the form

$$A \mapsto K_n(\psi)^T A K_n(\phi) \tag{1.15}$$

for fixed $\phi, \psi \in \mathrm{GL}(\mathbb{F}^2)$. In the literature special cases of such transformations are called *Frobenius-Fischer transformations*. We will use this name for all transformations of this form. Some Frobenius-Fischer transformations are mappings inside a class of structured matrices, other build a bridge between different classes. We discuss here the situation with Hankel and Toeplitz matrices.

Let $\mathcal{H}_n$ denote the class of $n \times n$ Hankel matrices $H_n(\mathbf{s}) = [\, s_{i+j-1} \,]_{i,j=1}^n$, where $\mathbf{s} = (s_i)_{i=1}^{2n-1} \in \mathbb{F}^{2n-1}$. The following proposition describes Frobenius-Fischer transformations that map $\mathcal{H}_n$ into itself.

Proposition 1.6. *For $\phi \in \mathrm{GL}(\mathbb{F}^2)$ and $\mathbf{s} \in \mathbb{F}^{2n-1}$, the equality*

$$K_n(\phi)^T H_n(\mathbf{s}) K_n(\phi) = H_n(\widetilde{\mathbf{s}})$$

with $\widetilde{\mathbf{s}} = K_{2n-1}(\phi)^T \mathbf{s}$ is satisfied.

Proof. It suffices to prove the proposition for the special cases (1.12). The cases (a) and (c) are obvious. Let ϕ be now of the form (b), $K_n(\phi)^T H_n(\mathbf{s}) K_n(\phi) = [\, g_{ij} \,]_{i,j=0}^{n-1}$. Then

$$g_{ij} = \sum_{k=0}^{i} \sum_{l=0}^{j} \binom{i}{k}\binom{j}{l} b^{i+j-k-l} s_{k+l+1} = \sum_{r=0}^{i+j} \binom{i+j}{r} b^{i+j-r} s_{r+1} \ .$$

This implies the assertion. □

Now we consider besides Hankel also Toeplitz matrices $T_n = [\, a_{i-j} \,]_{i,j=1}^{n}$. The class of $n \times n$ Toeplitz matrices will be denoted by $\mathcal{T}_n$. Obviously, T_n is Toeplitz if and only if $J_n T_n$ is Hankel. Remember that J_n is the special Möbius matrix $K_n(J_2)$. Thus modifications of Propositions 1.6 can be stated about Frobenius-Fischer transformations transforming Toeplitz into Hankel, Hankel into Toeplitz and Toeplitz into Toeplitz matrices. In particular, we have the following.

Corollary 1.7. *For $\phi \in \mathrm{GL}(\mathbb{F}^2)$, the transformation $A \mapsto K_n(\psi)^T A\, K_n(\phi)$ maps*

1. *$\mathcal{H}_n$ into $\mathcal{H}_n$ if $\psi = \phi$,* 2. *$\mathcal{T}_n$ into $\mathcal{H}_n$ if $\psi = J_2\phi$,*
3. *$\mathcal{H}_n$ into $\mathcal{T}_n$ if $\psi = \phi J_2$,* 4. *$\mathcal{T}_n$ into $\mathcal{T}_n$ if $\psi = J_2\phi J_2$.*

In the case $\mathbb{F} = \mathbb{C}$ we are in particular interested in congruence transformations, i.e., transformations that preserve Hermitian symmetry. For this we have to check under which condition $K_n(\psi)^T = K_n^*(\phi)$. In terms of the matrix $\phi = \begin{bmatrix} a & c \\ b & d \end{bmatrix}$ this condition is equivalent to

1. a, b, c, d real, 2. $a = \overline{b}$, $c = \overline{d}$, 3. $a = \overline{c}$, $b = \overline{d}$, 4. $a = \overline{d}$, $b = \overline{c}$

in the cases of Corollary 1.7. In terms of the linear fractional function $\phi(t) = \dfrac{at+b}{ct+d}$ this means that

1. $\phi(t)$ maps $\mathbb{R}$ to $\mathbb{R}$, 2. $\phi(t)$ maps $\mathbb{T}$ to $\mathbb{R}$, 3. $\phi(t)$ maps $\mathbb{R}$ to $\mathbb{T}$, 4. $\phi(t)$ maps $\mathbb{T}$ to $\mathbb{T}$,

where $\mathbb{T}$ denotes the unit circle. For transformations with this property the inertia of the matrix remains invariant by Sylvester's inertia law.

2. Definitions and properties for the Hankel and Toeplitz case

1. Hankel Bezoutians. Let $\mathbf{u}(t), \mathbf{v}(t) \in \mathbb{F}^{n+1}(t)$ be two polynomials. The *Hankel Bezoutian* or briefly *H-Bezoutian* of $\mathbf{u}(t)$ and $\mathbf{v}(t)$ is, by definition, the $n \times n$ matrix $B = \mathrm{Bez}_H(\mathbf{u}, \mathbf{v})$ with the generating polynomial

$$B(t,s) = \frac{\mathbf{u}(t)\mathbf{v}(s) - \mathbf{v}(t)\mathbf{u}(s)}{t-s} \ .$$

We will also say that B is the H-Bezoutian of the vectors $\mathbf{u}$ and $\mathbf{v}$. It is easily seen that $B(t,s)$ is really a polynomial in t and s. A simple argumentation for this is as follows. We fix $s = s_0$. Then the numerator is a polynomial in t vanishing at $t = s_0$. Hence we obtain a polynomial after dividing the numerator by $t - s_0$. Thus

$B(t, s_0)$ is a polynomial in t. Analogously $B(t_0, s)$ is a polynomial in s for fixed $t = t_0$, and so the claim is proved.

For example, if $\mathbf{u}(t) = t - a$ and $\mathbf{v}(t) = t - b$, then we have

$$B(t,s) = \frac{(t-a)(s-b) - (t-b)(s-a)}{t-s} = a - b\,.$$

Thus $\mathrm{Bez}_H(\mathbf{u}, \mathbf{v}) = a - b$.

As a second example we consider the flip matrix J_n introduced in (1.2). The generating function of J_n is

$$J_n(t) = \frac{t^n - s^n}{t-s}\,.$$

Thus J_n is the H-Bezoutian of the polynomials t^n and 1 or, in other words, of the last and first unit vectors $\mathbf{e}_{n+1}$ and $\mathbf{e}_1$.

As a more general example, let us compute the H-Bezoutian B_k of a general polynomial $\mathbf{u}(t) \in \mathbb{F}^{n+1}(t)$ and $\mathbf{e}_k(t) = t^{k-1}$, $k = 1, \ldots, n+1$. Suppose that $\mathbf{u}(t) = \sum_{i=1}^{n+1} u_i t^{i-1}$. Then we have

$$\begin{aligned} B_k(t,s) &= \sum_{i=1}^{n+1} u_i \frac{t^{i-1}s^{k-1} - t^{k-1}s^{i-1}}{t-s} \\ &= \sum_{i=k+1}^{n+1} u_i \frac{t^{i-k} - s^{i-k}}{t-s}\, t^{k-1}s^{k-1} - \sum_{i=1}^{k-1} u_i \frac{t^{k-i} - s^{k-i}}{t-s}\, t^{i-1}s^{i-1}\,. \end{aligned}$$

In matrix language this means that

$$\mathrm{Bez}_H(\mathbf{u}, \mathbf{e}_k) = \begin{bmatrix} & & -u_1 & & & \\ & \iddots & \vdots & & O & \\ -u_1 & \ldots & -u_{k-1} & & & \\ & & & u_{k+1} & \ldots & u_{n+1} \\ & O & & \vdots & \iddots & \\ & & & u_{n+1} & & \end{bmatrix}. \tag{2.1}$$

The case $k = 1$ is of particular importance. For this reason we introduce the notation

$$B(\mathbf{u}) = \mathrm{Bez}_H(\mathbf{u}, \mathbf{e}_1) = \begin{bmatrix} u_2 & \ldots & u_{n+1} \\ \vdots & \iddots & \\ u_{n+1} & & \end{bmatrix}. \tag{2.2}$$

H-Bezoutians are obviously symmetric matrices. They are skewsymmetric with respect to the arguments, i.e.,

$$\mathrm{Bez}_H(\mathbf{u}, \mathbf{v}) = -\mathrm{Bez}_H(\mathbf{v}, \mathbf{u})\,.$$

Furthermore, $\mathrm{Bez}_H(\mathbf{u},\mathbf{v})$ is linear in each argument. That means that, for c_1, $c_2 \in \mathbb{F}$,

$$\mathrm{Bez}_H(c_1\mathbf{u}_1 + c_2\mathbf{u}_2, \mathbf{v}) = c_1\mathrm{Bez}_H(\mathbf{u}_1,\mathbf{v}) + c_2\mathrm{Bez}_H(\mathbf{u}_2,\mathbf{v})\,.$$

To present a product rule for H-Bezoutians we need the matrix of the multiplication operator which is introduced as follows. Let $\mathbf{a}(t) = \sum_{k=0}^{m} a_k t^k \in \mathbb{F}^{m+1}(t)$. For $n = 1, 2, \ldots$, define a linear operator $M_n(\mathbf{a}) : \mathbb{F}^n \longrightarrow \mathbb{F}^{m+n}$ by

$$(M_n(\mathbf{a})\mathbf{x})(t) = \mathbf{a}(t)\mathbf{x}(t)\,.$$

The matrix of this operator with respect to the standard bases is the $(m+n)\times n$ matrix

$$M_n(\mathbf{a}) = \left.\begin{bmatrix} a_0 & & & \\ a_1 & a_0 & & \\ \vdots & a_1 & \ddots & \\ a_m & \vdots & \ddots & a_0 \\ & a_m & & a_1 \\ & & \ddots & \vdots \\ & & & a_m \end{bmatrix}\right\} m+n\,. \tag{2.3}$$

Moreover, we need the following matrix. Let $\mathbf{a}(t) = \sum_{k=0}^{m} a_k t^k$ and $\mathbf{b}(t) = \sum_{k=0}^{n} b_k t^k$ be two given polynomials. Then

$$(\mathbf{x}(t),\mathbf{y}(t)) \mapsto \mathbf{a}(t)\mathbf{x}(t) + \mathbf{b}(t)\mathbf{y}(t), \quad \mathbf{x}(t) \in \mathbb{F}^n(t),\ \mathbf{y}(t) \in \mathbb{F}^m(t)$$

is a linear operator from the direct product $\mathbb{F}^n(t) \otimes \mathbb{F}^m(t)$ to $\mathbb{F}^{m+n}(t)$. The matrix of this operator with respect to the standard bases is given by $[\,M_n(\mathbf{a})\ \ M_m(\mathbf{b})\,]$. (Here we identify $(\mathbf{x}(t),\mathbf{y}(t))$ with $\mathbf{x}(t) + t^n\mathbf{y}(t)$.) The transpose of this matrix is called the *resultant matrix* (or *Sylvester matrix*) of $\mathbf{a}(t)$ and $\mathbf{b}(t)$ (or of the vectors $\mathbf{a}$ and $\mathbf{b}$) and is denoted by $\mathrm{Res}(\mathbf{a},\mathbf{b})$,

$$\mathrm{Res}(\mathbf{a},\mathbf{b}) = \begin{bmatrix} M_n(\mathbf{a})^T \\ M_m(\mathbf{b})^T \end{bmatrix}. \tag{2.4}$$

If we assume that $a_m \neq 0$ or $b_n \neq 0$ then $\mathrm{Res}(\mathbf{a},\mathbf{b})$ is nonsingular if and only if $\mathbf{a}(t)$ and $\mathbf{b}(t)$ are coprime (cf. Exercise 3).

Proposition 2.1. *Let* $\mathbf{u},\mathbf{v} \in \mathbb{F}^{n+1}$, $\mathbf{u}(t) = \mathbf{u}_1(t)\mathbf{u}_2(t)$, $\mathbf{v}(t) = \mathbf{v}_1(t)\mathbf{v}_2(t)$, *where* $\mathbf{u}_i, \mathbf{v}_i \in \mathbb{F}^{n_i+1}$ $(i = 1, 2)$ *and* $n_1 + n_2 = n - 1$. *Then*

$$\mathrm{Bez}_H(\mathbf{u},\mathbf{v}) = \mathrm{Res}\,(\mathbf{u}_2,\mathbf{v}_1)^T \begin{bmatrix} \mathrm{Bez}_H(\mathbf{u}_1,\mathbf{v}_1) & O \\ O & \mathrm{Bez}_H(\mathbf{u}_2,\mathbf{v}_2) \end{bmatrix} \mathrm{Res}\,(\mathbf{v}_2,\mathbf{u}_1)\,. \tag{2.5}$$

Proof. Let $B = \mathrm{Bez}_H(\mathbf{u},\mathbf{v})$. Then, $B(t,s)$ has the representation

$$\mathbf{u}_2(t)\,\frac{\mathbf{u}_1(t)\mathbf{v}_1(s) - \mathbf{v}_1(t)\mathbf{u}_1(s)}{t-s}\,\mathbf{v}_2(s) + \mathbf{v}_1(t)\,\frac{\mathbf{u}_2(t)\mathbf{v}_2(s) - \mathbf{v}_2(t)\mathbf{u}_2(s)}{t-s}\,\mathbf{u}_1(s)\,.$$

In matrix language this means

$$B = M_{n_1}(\mathbf{u}_2)\mathrm{Bez}_H(\mathbf{u}_1,\mathbf{v}_1)M_{n_1}(\mathbf{v}_2)^T + M_{n_2}(\mathbf{v}_1)\mathrm{Bez}_H(\mathbf{u}_2,\mathbf{v}_2)M_{n_2}(\mathbf{u}_1)^T\,.$$

From this relation the assertion is immediate. □

2. The transformation ∇_H. Next we clarify what means for a matrix to be an H-Bezoutian in matrix language. For this we introduce the transformation ∇_H transforming an $n\times n$ matrix $A = [\,a_{ij}\,]_{i,j=1}^{n}$ into a $(n+1)\times(n+1)$ matrix according to

$$\nabla_H A = [\,a_{i-1,j} - a_{i,j-1}\,]_{i,j=1}^{n+1}\,.$$

Here we set $a_{ij}=0$ if one of the integers i or j is not in the set $\{1,2,\dots,n\}$. We have

$$\nabla_H A = \begin{bmatrix} S_nA - AS_n^T & * \\ * & * \end{bmatrix} = \begin{bmatrix} * & * \\ * & AS_n - S_n^TA \end{bmatrix}, \tag{2.6}$$

where S_n is defined in (1.8). Comparing the coefficients it is easy to verify that

$$(\nabla_H A)(t,s) = (t-s)A(t,s)\,.$$

Hence the Bezoutians $B = \mathrm{Bez}_H(\mathbf{u},\mathbf{v})$ can be characterized with the help of ∇_H by

$$\nabla_H B = [\,\mathbf{u}\ \ \mathbf{v}\,]\begin{bmatrix} 0 & 1 \\ -1 & 0 \end{bmatrix}[\,\mathbf{u}\ \ \mathbf{v}\,]^T\,. \tag{2.7}$$

In particular, the rank of ∇B_H is equal to 2, unless $\mathbf{u}$ and $\mathbf{v}$ are linearly dependent. In the latter case the H-Bezoutian is the zero matrix. The representation (2.6) shows that the transformation ∇^+ introduced in (1.10) is a restriction of ∇_H. Thus, H-Bezoutians are quasi-Hankel matrices.

3. Uniqueness. Different pairs of polynomials may produce the same H-Bezoutian. However, from (2.7) one can conclude that if $\mathrm{Bez}_H(\mathbf{u},\mathbf{v}) = \mathrm{Bez}_H(\mathbf{u}_1,\mathbf{v}_1) \neq O$, then $\mathrm{span}\{\mathbf{u},\mathbf{v}\} = \mathrm{span}\{\mathbf{u}_1,\mathbf{v}_1\}$. In the latter case there is a nonsingular 2×2 matrix φ such that

$$[\,\mathbf{u}_1\ \ \mathbf{v}_1\,] = [\,\mathbf{u}\ \ \mathbf{v}\,]\,\varphi\,. \tag{2.8}$$

Lemma 2.2. *Let $\mathbf{u},\mathbf{v},\mathbf{u}_1,\mathbf{v}_1 \in \mathbb{F}^{n+1}$ related via (2.8). Then*

$$\mathrm{Bez}_H(\mathbf{u}_1,\mathbf{v}_1) = (\det\varphi)\,\mathrm{Bez}_H(\mathbf{u},\mathbf{v})\,.$$

Proof. Suppose that $\varphi = \begin{bmatrix} a & c \\ b & d \end{bmatrix}$, $B_1 = \mathrm{Bez}_H(\mathbf{u}_1,\mathbf{v}_1)$. Then

$$\begin{aligned}(t-s)B_1(t,s) &= (a\mathbf{u}(t)+b\mathbf{v}(t))(c\mathbf{u}(s)+d\mathbf{v}(s)) - (c\mathbf{u}(t)+d\mathbf{v}(t))(a\mathbf{u}(s)+b\mathbf{v}(s)) \\ &= (ad-bc)(\mathbf{u}(t)\mathbf{v}(s)-\mathbf{v}(t)\mathbf{u}(s))\,,\end{aligned}$$

which proves the lemma. □

Corollary 2.3. *The H-Bezoutians $\mathrm{Bez}_H(\mathbf{u},\mathbf{v}) \neq O$ and $\mathrm{Bez}_H(\mathbf{u}_1,\mathbf{v}_1)$ coincide if and only if the vectors $\mathbf{u},\mathbf{v}$ and $\mathbf{u}_1$, $\mathbf{v}_1$ are related via (2.8) with $\det\varphi = 1$.*

From Corollary 2.3 we can conclude that the H-Bezoutian $\mathrm{Bez}_H(\mathbf{u},\mathbf{v})$ is equal to a H-Bezoutian $\mathrm{Bez}_H(\widetilde{\mathbf{u}},\widetilde{\mathbf{v}})$ in which the last coefficient of $\widetilde{\mathbf{v}}$ vanishes, i.e., $\widetilde{\mathbf{v}}(t) \in \mathbb{F}^n(t)$.

4. Quasi-H-Bezoutians. A matrix B is called *quasi-H-Bezoutian* if $\operatorname{rank} \nabla_H B \le 2$. We give a general representation of quasi-H-Bezoutians that is also important for H-Bezoutians.

Proposition 2.4. *A quasi-H-Bezoutian $B \neq O$ of order n admits a representation*

$$B = M_r(\mathbf{p})\,\mathrm{Bez}_H(\mathbf{u},\mathbf{v})\,M_r(\mathbf{q})^T\,, \tag{2.9}$$

where $\mathbf{u}(t),\mathbf{v}(t) \in \mathbb{F}^{r+1}(t)$ are coprime and $r \le n$. Here $M_r(\,\cdot\,)$ is defined in (2.3).

Proof. For B is a quasi-H-Bezoutian, there exist $\mathbf{a},\mathbf{b},\mathbf{c},\mathbf{d} \in \mathbb{F}^{n+1}$ such that

$$(t-s)B(t,s) = \mathbf{a}(t)\mathbf{d}(s) - \mathbf{b}(t)\mathbf{c}(s)\,.$$

Since for $t = s$ the left-hand side vanishes, we have $\mathbf{a}(t)\mathbf{d}(t) = \mathbf{b}(t)\mathbf{c}(t)$. Let $\mathbf{p}(t)$ be the greatest common divisor of $\mathbf{a}(t)$ and $\mathbf{b}(t)$ and $\mathbf{q}(t)$ the greatest common divisor of $\mathbf{c}(t)$ and $\mathbf{d}(t)$. Then $\mathbf{a}(t) = \mathbf{p}(t)\mathbf{u}(t)$ and $\mathbf{b}(t) = \mathbf{p}(t)\mathbf{v}(t)$ for some coprime $\mathbf{u}(t),\mathbf{v}(t) \in \mathbb{F}^{r+1}(t)$ $(r \le n)$. Furthermore, $\mathbf{c}(t) = \mathbf{q}(t)\mathbf{u}_1(t)$ and $\mathbf{d}(t) = \mathbf{q}(t)\mathbf{v}_1(t)$ for some coprime $\mathbf{u}_1(t)$, $\mathbf{v}_1(t) \in \mathbb{F}^{r_1+1}(t)$ $(r_1 \le n)$. Since

$$\frac{\mathbf{a}(t)}{\mathbf{b}(t)} = \frac{\mathbf{u}(t)}{\mathbf{v}(t)} = \frac{\mathbf{c}(t)}{\mathbf{d}(t)} = \frac{\mathbf{u}_1(t)}{\mathbf{v}_1(t)}\,,$$

we conclude that, for some $\gamma \neq 0$, $\mathbf{u}_1 = \gamma\mathbf{u}$, $\mathbf{v}_1 = \gamma\mathbf{v}$, and $r = r_1$. Now we have

$$\mathbf{a}(t)\mathbf{d}(s) - \mathbf{b}(t)\mathbf{c}(s) = \gamma\,\mathbf{p}(t)(\mathbf{u}(t)\mathbf{v}(s) - \mathbf{v}(t)\mathbf{u}(s))\mathbf{q}(s)\,.$$

We can replace $\gamma\,\mathbf{p}$ by $\mathbf{p}$. Now it remains to translate this into matrix language to obtain (2.9). □

The matrix on the right-hand side of (2.9) has rank r at most. Hence if $r < n$, then B is singular. This leads to the following somehow surprising conclusion.

Corollary 2.5. *Any nonsingular quasi-H-Bezoutian is an H-Bezoutian of two coprime polynomials.*

Later we will show that, vice versa, the H-Bezoutian of two coprime polynomials is nonsingular (cf. Corollary 3.4).

If the quasi-H-Bezoutian is symmetric, then in (2.9) we must have $\mathbf{q} = \mathbf{p}$, since the middle factor is symmetric. This implies the following.

Corollary 2.6. *Any symmetric quasi-H-Bezoutian is an H-Bezoutian*

$$B = \mathrm{Bez}_H(\mathbf{a},\mathbf{b})\,.$$

In particular, (2.9) *can be written in the form*

$$B = M_r(\mathbf{p})\,\mathrm{Bez}_H(\mathbf{u},\mathbf{v})\,M_r(\mathbf{p})^T\,, \tag{2.10}$$

where $\mathbf{p}(t)$ is the greatest common divisor of $\mathbf{a}(t)$ and $\mathbf{b}(t)$.

5. Frobenius-Fischer transformations. We show now that Frobenius-Fischer transformations introduced in Section 1 transform the class of H-Bezoutians into itself. In particular, the following result is the Bezoutian counterpart of Proposition 1.6.

Theorem 2.7. *For any $\varphi \in \mathrm{GL}(\mathbb{F}^2)$, the transformation*

$$B \mapsto K_n(\varphi) B K_n(\varphi)^T$$

maps H-Bezoutians into H-Bezoutians. Moreover

$$K_n(\varphi)\mathrm{Bez}_H(\mathbf{u},\mathbf{v})K_n(\varphi)^T = \frac{1}{\det\varphi}\,\mathrm{Bez}_H(\widetilde{\mathbf{u}},\widetilde{\mathbf{v}})\,, \tag{2.11}$$

where $\widetilde{\mathbf{u}} = K_{n+1}(\varphi)\mathbf{u}$ and $\widetilde{\mathbf{v}} = K_{n+1}(\varphi)\mathbf{v}$.

Proof. Let $B = \mathrm{Bez}_H(\mathbf{u},\mathbf{v})$ and $\widetilde{B} = \mathrm{Bez}_H(\widetilde{\mathbf{u}},\widetilde{\mathbf{v}})$. It is sufficient to prove the theorem for the matrices (1.12) that generate $\mathrm{GL}(\mathbb{F}^2)$. If $\varphi = \begin{bmatrix} a & 0 \\ 0 & 1 \end{bmatrix}$, then

$$\widetilde{B}(t,s) = a\,\frac{\mathbf{u}(at)\mathbf{v}(as) - \mathbf{v}(at)\mathbf{u}(as)}{at - as} = a\,B(at,as) = a\,(K_n(\varphi)BK_n(\varphi)^T)(t,s)\,,$$

which is equivalent to (2.11). If $\varphi = \begin{bmatrix} 1 & 0 \\ b & 1 \end{bmatrix}$, then

$$\begin{aligned}\widetilde{B}(t,s) &= \frac{\mathbf{u}(t+b)\mathbf{v}(s+b) - \mathbf{v}(t+b)\mathbf{u}(s+b)}{t-s} = B(t+b,s+b)\\ &= (K_n(\varphi)BK_n(\varphi)^T)(t,s)\,,\end{aligned}$$

which is equivalent to (2.11). Finally, let $\varphi = J_2$. Then

$$\begin{aligned}\widetilde{B}(t,s) &= \frac{\mathbf{u}(t^{-1})t^n\mathbf{v}(s^{-1})s^n - \mathbf{v}(t^{-1})t^n\mathbf{u}(s^{-1})s^n}{t-s} = -B(t^{-1},s^{-1})(ts)^{n-1}\\ &= -B^J(t,s)\,,\end{aligned}$$

which is again equivalent to (2.11). □

6. Splitting of H-Bezoutians. In some applications, like stability tests for real polynomials, $\mathbf{u}(t)$ and $\mathbf{v}(t)$ have the special form

$$\mathbf{u}(t) = \mathbf{a}(t^2) \quad \text{and} \quad \mathbf{v}(t) = t\,\mathbf{b}(t^2). \tag{2.12}$$

That means $\mathbf{u}(t)$ has only even powers and $\mathbf{v}(t)$ only odd powers. In this case we have for the generating polynomial of $B = \mathrm{Bez}_H(\mathbf{u},\mathbf{v})$ after multiplying numerator and denominator by $(t+s)$

$$\begin{aligned}B(t,s) &= \frac{ts(\mathbf{a}(t^2)\mathbf{b}(s^2) - \mathbf{b}(t^2)\mathbf{a}(s^2)) + \mathbf{a}(t^2)s^2\mathbf{b}(s^2) - t^2\mathbf{b}(t^2)\mathbf{a}(s^2)}{t^2 - s^2}\\ &= tsB_1(t^2,s^2) + B_0(t^2,s^2)\,,\end{aligned}$$

where $B_1 = \mathrm{Bez}_H(\mathbf{a},\mathbf{b})$ and $B_0 = \mathrm{Bez}_H(\mathbf{a},t\mathbf{b})$. To translate this into matrix language we introduce the matrix Σ_n of the even-odd shuffle operator:

$$\Sigma_n(x_i)_{i=1}^n = (x_1, x_3, \ldots, x_2, x_4, \ldots)\,.$$

Proposition 2.8. *Let* $\mathbf{u}(t) \in \mathbb{F}^n(t)$ *and* $\mathbf{v}(t) \in \mathbb{F}^n(t)$ *be given by (*2.12*). Then*

$$\Sigma_n^T \mathrm{Bez}_H(\mathbf{u}, \mathbf{v}) \Sigma_n = \begin{bmatrix} \mathrm{Bez}_H(\mathbf{a}, t\,\mathbf{b}) & O \\ O & \mathrm{Bez}_H(\mathbf{a}, \mathbf{b}) \end{bmatrix}.$$

7. Toeplitz Bezoutians. We introduce the Toeplitz analogue of the H-Bezoutian. The *Toeplitz Bezoutian* or briefly *T-Bezoutian* of the two polynomials $\mathbf{u}(t) \in \mathbb{F}^{n+1}(t)$ and $\mathbf{v}(t) \in \mathbb{F}^{n+1}(t)$ is, by definition, the matrix $B = \mathrm{Bez}_T(\mathbf{u}, \mathbf{v})$ with the generating polynomial

$$B(t,s) = \frac{\mathbf{u}(t)\mathbf{v}^J(s) - \mathbf{v}(t)\mathbf{u}^J(s)}{1 - ts}.$$

Like for H-Bezoutians, it is easily checked that $B(t,s)$ is really a polynomial in t and s. If, for example, the polynomials $\mathbf{u}(t) = t - a$ and $\mathbf{v}(t) = t - b$ of $\mathbb{F}^{n+1}(t)$ are given, then for $n = 1$

$$B(t,s) = \frac{(t-a)(1-bs) - (t-b)(1-as)}{1-ts} = b - a\,.$$

Hence $\mathrm{Bez}_T(\mathbf{u}, \mathbf{v}) = b - a$. But in case $n > 1$ we have

$$\mathrm{Bez}_T(\mathbf{u}, \mathbf{v}) = \begin{bmatrix} O & b-a \\ O & O \end{bmatrix}.$$

We state that the definition of a T-Bezoutian of two polynomials depends, in contrast to the H-Bezoutian, essentially on the integer n. That means if we consider $\mathbf{u}(t)$ and $\mathbf{v}(t)$ as elements of $\mathbb{F}^{N+1}(t)$ for $N > n$, then we will have a different T-Bezoutian. Indeed, let B_N denote the T-Bezoutian in this sense. Then we obtain

$$B_N(t,s) = \frac{\mathbf{u}(t)\mathbf{v}(s^{-1})s^N - \mathbf{v}(t)\mathbf{u}(s^{-1})s^N}{1-ts} = B(t,s)s^{N-n}\,,$$

where B is the T-Bezoutian of $\mathbf{u}$ and $\mathbf{v}$ in the original sense. Thus, B_N is of the form

$$B_N = \begin{bmatrix} O & B \\ O & O \end{bmatrix}.$$

If $t = 0$ is a common zero of $\mathbf{u}(t)$ and $\mathbf{v}(t)$, $\mathbf{u}(t) = t^r\mathbf{u}_0(t)$, $\mathbf{v}(t) = t^r\mathbf{v}_0(t)$ $(r > 0)$, then B is of the form

$$B = \begin{bmatrix} O & O \\ B_0 & O \end{bmatrix}$$

where B_0 is the $(n-r) \times (n-r)$ T-Bezoutian of $\mathbf{u}_0$ and $\mathbf{v}_0$.

As an example, we compute the T-Bezoutian of a polynomial and a power of t. Let $B_{(k)} = \mathrm{Bez}_T(\mathbf{u}, \mathbf{e}_k)$ and $\mathbf{u} = (u_i)_{i=1}^{n+1}$. Then

$$\begin{aligned} B_{(k)}(t,s) &= \sum_{i=1}^{n+1} u_i \frac{t^{i-1}s^{n-k+1} - t^{k-1}s^{n-i+1}}{1-ts} \\ &= \sum_{i=1}^{k-1} u_i\, t^{i-1}s^{n-k+1} \frac{1-(ts)^{k-i}}{1-ts} + \sum_{i=k+1}^{n+1} u_i\, t^{k-1}s^{n-i+1} \frac{(ts)^{i-k}-1}{1-ts}\,. \end{aligned}$$

We take into account that $\dfrac{1-(ts)^k}{1-ts}$ is the generating polynomial of I_k and obtain in matrix form

$$\mathrm{Bez}_T(\mathbf{u},\mathbf{e}_k) = \begin{bmatrix} & & & u_1 & & \\ & O & & \vdots & \ddots & \\ & & & u_{k-1} & \dots & u_1 \\ -u_{n+1} & \dots & -u_{k+1} & & & \\ & \ddots & \vdots & & O & \\ & & -u_{n+1} & & & \end{bmatrix}. \tag{2.13}$$

For the special cases $k=1$ and $k=n+1$ we introduce the notations

$$B_+(\mathbf{u}) = -\begin{bmatrix} u_{n+1} & \dots & u_2 \\ & \ddots & \vdots \\ & & u_{n+1} \end{bmatrix}, \quad B_-(\mathbf{u}) = \begin{bmatrix} u_1 & & \\ \vdots & \ddots & \\ u_n & \dots & u_1 \end{bmatrix}. \tag{2.14}$$

Obviously, the T-Bezoutian is, like the H-Bezoutian, linear in its arguments. Furthermore, it is skewsymmetric with respect to the arguments, i.e.,

$$\mathrm{Bez}_T(\mathbf{u},\mathbf{v}) = -\mathrm{Bez}_T(\mathbf{v},\mathbf{u})\,.$$

Moreover, we have

$$\mathrm{Bez}_T(\mathbf{u},\mathbf{v})^T = -\mathrm{Bez}_T(\mathbf{u}^J,\mathbf{v}^J) \tag{2.15}$$

and

$$\begin{aligned} \mathrm{Bez}_T(\mathbf{u},\mathbf{v})^J(t,s) &= \frac{\mathbf{u}(t^{-1})\mathbf{v}(s)s^{-n} - \mathbf{v}(t^{-1})\mathbf{u}(s)s^{-n}}{1-t^{-1}s^{-1}}\, t^{n-1}s^{n-1} \\ &= \frac{\mathbf{u}(t^{-1})t^n\mathbf{v}(s) - \mathbf{v}(t^{-1})t^n\mathbf{u}(s)}{ts-1}\,, \end{aligned}$$

which means that

$$\mathrm{Bez}_T(\mathbf{u},\mathbf{v})^J = -\mathrm{Bez}_T(\mathbf{u}^J,\mathbf{v}^J)\,. \tag{2.16}$$

From (2.15) and (2.16) we conclude that

$$\mathrm{Bez}_T(\mathbf{u},\mathbf{v})^T = \mathrm{Bez}_T(\mathbf{u},\mathbf{v})^J\,.$$

Hence T-Bezoutians are, like Toeplitz matrices, persymmetric.

The discussion about uniqueness of H-Bezoutians in Section 2.1 can be immediately transferred to T-Bezoutians. In fact, there is a Toeplitz analogue of Lemma 2.2.

Lemma 2.9. *If* $[\,\mathbf{u}_1\ \ \mathbf{v}_1\,] = [\,\mathbf{u}\ \ \mathbf{v}\,]\,\varphi$ *for some* 2×2 *matrix* φ*, then*

$$\mathrm{Bez}_T(\mathbf{u}_1,\mathbf{v}_1) = (\det\varphi)\,\mathrm{Bez}_T(\mathbf{u},\mathbf{v})\,.$$

Hence the following is true.

Corollary 2.10. *The T-Bezoutians* $\mathrm{Bez}_T(\mathbf{u},\mathbf{v}) \neq O$ *and* $\mathrm{Bez}_T(\mathbf{u}_1,\mathbf{v}_1)$ *coincide if and only if*

$$[\,\mathbf{u}_1 \;\; \mathbf{v}_1\,] = [\,\mathbf{u} \;\; \mathbf{v}\,]\,\varphi$$

for some matrix φ *with* $\det\varphi = 1$.

8. The transformation ∇_T. The Toeplitz analogue of the transformation ∇_H is the transformation ∇_T transforming an $n\times n$ matrix $A = [\,a_{ij}\,]_{i,j=1}^{n}$ into a $(n+1)\times(n+1)$ matrix according to

$$\nabla_T A = [\,a_{ij} - a_{i-1,j-1}\,]_{i,j=1}^{n+1}\,.$$

Here we set $a_{ij} = 0$ if one of the integers i or j is not in the set $\{1,2,\dots,n\}$. Obviously,

$$\nabla_T A = \begin{bmatrix} A - S_n A S_n^T & * \\ * & * \end{bmatrix} = \begin{bmatrix} * & * \\ * & S_n^T A S_n - A \end{bmatrix}. \tag{2.17}$$

In polynomial language the transformation ∇_T is given by

$$(\nabla_T A)(t,s) = (1-ts)A(t,s)\,.$$

That means the T-Bezoutian $B = \mathrm{Bez}_T(\mathbf{u},\mathbf{v})$ is characterized by

$$\nabla_T B = [\,\mathbf{u} \;\; \mathbf{v}\,]\begin{bmatrix} 0 & 1 \\ -1 & 0 \end{bmatrix}[\,\mathbf{u}^J \;\; \mathbf{v}^J\,]^T\,.$$

Taking into account (2.16) we observe that $(\nabla_T B)^J = -\nabla_T B^J$.

The representation (2.17) shows that the transformation ∇_+ introduced in (1.7) is a restriction of ∇_T. In particular, we conclude that T-Bezoutians are quasi-Toeplitz matrices. Furthermore, if B is a T-Bezoutian, then B^J is also a quasi-Toeplitz matrix.

9. Symmetric and skewsymmetric T-Bezoutians. We discuss now how symmetric and skewsymmetric T-Bezoutians can be characterized. First we observe that (2.16) implies that $\mathrm{Bez}_T(\mathbf{u},\mathbf{v})$ is symmetric if one of the vectors $\mathbf{u}$ or $\mathbf{v}$ is symmetric and the other one is skewsymmetric. Furthermore, $\mathrm{Bez}_T(\mathbf{u},\mathbf{v})$ is skewsymmetric if both vectors $\mathbf{u}$ and $\mathbf{v}$ are symmetric or both are skewsymmetric. We show that the converse is also true. For simplicity of notation we write $B(\mathbf{u},\mathbf{v})$ instead of $\mathrm{Bez}_T(\mathbf{u},\mathbf{v})$.

Let $\mathbf{u},\mathbf{v}\in\mathbb{F}^{n+1}$ be any vectors, $\mathbf{u} = \mathbf{u}_+ + \mathbf{u}_-$ and $\mathbf{v} = \mathbf{v}_+ + \mathbf{v}_-$, where $\mathbf{u}_+,\mathbf{v}_+$ are symmetric and $\mathbf{u}_-,\mathbf{v}_-$ are skewsymmetric. Then $B(\mathbf{u},\mathbf{v}) = B_+ + B_-$, where

$$B_+ = B(\mathbf{u}_+,\mathbf{v}_-) + B(\mathbf{u}_-,\mathbf{v}_+)\,, \quad B_- = B(\mathbf{u}_+,\mathbf{v}_+) + B(\mathbf{u}_-,\mathbf{v}_-)\,,$$

B_+ is symmetric, and B_- is skewsymmetric. Suppose that $B = B(\mathbf{u},\mathbf{v})$ is symmetric. Then $B_- = O$. Hence

$$B(\mathbf{u}_+,\mathbf{v}_+) = B(\mathbf{v}_-,\mathbf{u}_-)\,.$$

Since the vectors $\mathbf{u}_+$ and $\mathbf{v}_+$ cannot be linear combinations of $\mathbf{u}_-$ and $\mathbf{v}_-$ from Corollary 2.10 it becomes clear that

$$B(\mathbf{u}_+,\mathbf{v}_+) = B(\mathbf{v}_-,\mathbf{u}_-) = O\,.$$

Thus $\mathbf{v}_\pm = \alpha_\pm \mathbf{u}_\pm$ for some $\alpha_\pm \in \mathbb{F}$ or $B(\mathbf{u}, \mathbf{v}) = O$. We conclude that

$$B = B_+ = B((\alpha_- - \alpha_+)\mathbf{u}_+, \mathbf{u}_-)\,.$$

That means that B is the T-Bezoutian of a symmetric and a skewsymmetric vector.

Suppose now that $B = B(\mathbf{u}, \mathbf{v}) \neq O$ is skewsymmetric. Then $B_+ = O$. Hence

$$B(\mathbf{u}_+, \mathbf{v}_-) = B(\mathbf{v}_+, \mathbf{u}_-)\,.$$

From Corollary 2.10 and the symmetry properties of the vectors we conclude that either $\{\mathbf{u}_+, \mathbf{v}_+\}$ as well as $\{\mathbf{u}_-, \mathbf{v}_-\}$ are linearly dependent or

$$B(\mathbf{u}_+, \mathbf{v}_-) = B(\mathbf{v}_+, \mathbf{u}_-) = O\,.$$

In the former case we would have $B_- = O$, so we have the latter case. Using again the symmetry properties of the vectors we find that either $\mathbf{u}_- = \mathbf{v}_- = \mathbf{0}$ or $\mathbf{u}_+ = \mathbf{v}_+ = \mathbf{0}$. That means that B is the Bezoutian of two symmetric or two skewsymmetric vectors. Let us summarize.

Proposition 2.11. *A T-Bezoutian is symmetric if and only if it is the T-Bezoutian of a symmetric and a skewsymmetric vector. A T-Bezoutian is skewsymmetric if and only if it is the T-Bezoutian of two symmetric vectors or two skewsymmetric vectors.*

Note that the T-Bezoutian $B(\mathbf{u}, \mathbf{v})$ of two skewsymmetric vectors cannot be nonsingular. In fact, in this case we have $\mathbf{u}(1) = \mathbf{v}(1) = 0$ such that $\mathbf{u}(t) = (t-1)\mathbf{u}_1(t)$ and $\mathbf{v}(t) = (t-1)\mathbf{v}_1(t)$. Then $\mathbf{u}_1$ and $\mathbf{v}_1$ are symmetric, and as in Proposition 2.4 we obtain

$$B(\mathbf{u}, \mathbf{v}) = M_{n-1}(t-1)\mathrm{Bez}_T(\mathbf{u}_1, \mathbf{v}_1)M_{n-1}(t-1)^T\,.$$

Thus $B(\mathbf{u}, \mathbf{v})$ has rank $n-1$ at most.

There is an alternative representation for symmetric T-Bezoutians, which has no skewsymmetric counterpart. Suppose that $B = B(\mathbf{u}_+, \mathbf{u}_-)$. We set $\mathbf{v} = -\frac{1}{2}\,\mathbf{u}_+ + \mathbf{u}_-$. Then

$$B(\mathbf{v}, \mathbf{v}^J) = B\,.$$

On the other hand, $B(\mathbf{v}, \mathbf{v}^J)$ is symmetric for any vector $\mathbf{v} \in \mathbb{F}^{n+1}$. Thus the following is true.

Corollary 2.12. *A T-Bezoutian B is symmetric if and only if it can be represented in the form $B = \mathrm{Bez}_T(\mathbf{v}, \mathbf{v}^J)$ for some $\mathbf{v} \in \mathbb{F}^{n+1}$.*

10. Hermitian T-Bezoutians. Now we characterize Hermitian T-Bezoutians. Suppose that $\mathbf{u}_+, \mathbf{u}_- \in \mathbb{C}^{n+1}$ are conjugate-symmetric. Then we conclude from (2.15) that the matrix $\mathrm{i}B(\mathbf{u}_+, \mathbf{u}_-)$ is Hermitian. Conversely, let $B = B(\mathbf{u}, \mathbf{v})$ be Hermitian, $\mathbf{u} = \mathbf{u}_+ + \mathrm{i}\mathbf{u}_-$ and $\mathbf{v} = \mathbf{v}_+ + \mathrm{i}\mathbf{v}_-$, where $\mathbf{u}_\pm, \mathbf{v}_\pm$ are conjugate-symmetric. Then $B(\mathbf{u}, \mathbf{v}) = B_+ + \mathrm{i}B_-$, where

$$B_+ = \mathrm{i}(B(\mathbf{u}_+, \mathbf{v}_-) + B(\mathbf{u}_-, \mathbf{v}_+))\,, \quad B_- = \mathrm{i}(B(\mathbf{u}_-, \mathbf{v}_-) - B(\mathbf{u}_+, \mathbf{v}_+))\,.$$

The matrices B_+ and $\mathrm{i}B_-$ are Hermitian. Since $B = B(\mathbf{u}, \mathbf{v})$ is assumed to be Hermitian, we have $B_- = O$, which means

$$B(\mathbf{u}_+, \mathbf{v}_+) = B(\mathbf{u}_-, \mathbf{v}_-)\,.$$

Using Corollary 2.10 we conclude that

$$[\,\mathbf{u}_- \;\; \mathbf{v}_-\,] = [\,\mathbf{u}_+ \;\; \mathbf{v}_+\,] \begin{bmatrix} a & c \\ b & d \end{bmatrix}$$

for some a, b, c, d with $ad - bc = 1$. Since all vectors under consideration are conjugate-symmetric, these numbers must be real. We obtain after elementary calculations

$$B(\mathbf{u}, \mathbf{v}) = \mathrm{i}B(\mathbf{u}_+, -(a+d)\mathbf{v}_+) = B(\mathbf{u}_+, -(a+d)\mathrm{i}\mathbf{v}_+)\,.$$

Thus, B is the Bezoutian of a conjugate-symmetric and a conjugate-skewsymmetric vector. Let us summarize.

Proposition 2.13. *A T-Bezoutian B is Hermitian if and only if it is of the form*

$$B = \mathrm{i}\mathrm{Bez}_T(\mathbf{u}_+, \mathbf{u}_-)$$

for conjugate-symmetric vectors $\mathbf{u}_+$ and $\mathbf{u}_-$.

As for symmetric T-Bezoutians, we have an alternative form. Suppose that $B = \mathrm{i}B(\mathbf{u}_+, \mathbf{u}_-)$ and set $\mathbf{v} = -\frac{1}{2}\,\mathbf{u}_+ + \mathrm{i}\mathbf{u}_-$. Then $B(\mathbf{v}, \mathbf{v}^{\#}) = B$. Since, on the other hand, the matrix $B(\mathbf{v}, \mathbf{v}^{\#})$ is Hermitian for any vector $\mathbf{v} \in \mathbb{F}^{n+1}$, which is easily checked, the following is true.

Corollary 2.14. *A T-Bezoutian B is Hermitian if and only if it can be represented in the form $B = \mathrm{Bez}_T(\mathbf{v}, \mathbf{v}^{\#})$ for some $\mathbf{v} \in \mathbb{C}^{n+1}$.*

11. Splitting of symmetric T-Bezoutians. It was mentioned in Section 2.7 that T-Bezoutians are persymmetric. Hence a symmetric T-Bezoutian B is also centrosymmetric. That means that the subspaces of symmetric or skewsymmetric vectors $\mathbb{F}^n_\pm$ are invariant under B. We show that the restrictions of a symmetric T-Bezoutian to $\mathbb{F}^n_\pm$ can be characterized by another kind of Bezoutians which is introduced next.

Let $\mathbf{p}, \mathbf{q} \in \mathbb{F}^{n+2}$ be either both symmetric or both skewsymmetric. Then

$$B_{\mathrm{split}}(t, s) = \frac{\mathbf{p}(t)\mathbf{q}(s) - \mathbf{q}(t)\mathbf{p}(s)}{(t-s)(ts-1)}$$

is a polynomial in t and s. The $n \times n$ matrix with the generating polynomial $B_{\mathrm{split}}(t, s)$ will be called *split Bezoutian* of $\mathbf{p}(t)$ and $\mathbf{q}(t)$ and denoted by

$$\mathrm{Bez}_{\mathrm{split}}(\mathbf{p}, \mathbf{q})\,.$$

Obviously, $\mathrm{Bez}_{\mathrm{split}}(\mathbf{p}, \mathbf{q})$ is a symmetric and centrosymmetric matrix. If $\mathbf{p}$ and $\mathbf{q}$ are symmetric, then we will speak about a *split Bezoutian of (+)-type* and if these vectors are skewsymmetric about a *split Bezoutian of (−)-type*. Instead of B_{split} we write B_+ or B_-, respectively.

The columns and rows of a split Bezoutian of (+)-type are all symmmetric and of a split Bezoutian of (−)-type are all skewsymmetric, so that its rank is at most $\frac{1}{2}(n+1)$ in the (+) case and $\frac{1}{2}n$ in the (−) case. As an example we consider the case $\mathbf{p}(t) = t^{2k} + 1 \in \mathbb{F}^{2k+1}(t)$ and $\mathbf{q}(t) = t^k \in \mathbb{F}^{2k+1}(t)$. In this case

$$\begin{aligned} B_+(t,s) &= \frac{(t^{2k}+1)s^k - t^k(s^{2k}+1)}{(t-s)(ts-1)} = \frac{t^k - s^k}{t-s}\,\frac{(ts)^k - 1}{ts-1} \\ &= (t^{k-1} + t^{k-2}s + \cdots + s^{k-1})(1 + ts + \cdots + t^{k-1}s^{k-1})\,. \end{aligned}$$

For $k = 3$, the matrix with this generating polynomial is

$$B_+ = \begin{bmatrix} & & 1 & & \\ & 1 & & 1 & \\ 1 & & 1 & & 1 \\ & 1 & & 1 & \\ & & 1 & & \end{bmatrix}.$$

For a general symmetric $\mathbf{p} = (p_i)_{i=1}^7 \in \mathbb{F}^7$ and $\mathbf{q}$ as before, $\mathbf{q} = \mathbf{e}_4$, the split Bezoutian of $\mathbf{p}$ and $\mathbf{q}$ is given by

$$B_+ = \begin{bmatrix} & & p_1 & & \\ & p_1 & p_2 & p_1 & \\ p_1 & p_2 & p_1 + p_3 & p_2 & p_1 \\ & p_1 & p_2 & p_1 & \\ & & p_1 & & \end{bmatrix}.$$

Moreover, from this special case it is clear how the split Bezoutian of a general $\mathbf{p} \in \mathbb{F}_+^{2k+1}$ and $\mathbf{q} = \mathbf{e}_{k+1}$ looks like.

Recall that $P_\pm = \frac{1}{2}(I_n \pm J_n)$ are the projections from $\mathbb{F}^n$ onto $\mathbb{F}^n_\pm$ along $\mathbb{F}^n_\mp$. Our aim is to describe $BP_\pm = P_\pm BP_\pm$ for a symmetric T-Bezoutian B. As we know from Proposition 2.11, a symmetric T-Bezoutian B is the T-Bezoutian of a symmetric vector $\mathbf{u}_+ \in \mathbb{F}_+^{n+1}$ and a skewsymmetric vector $\mathbf{v}_- \in \mathbb{F}_-^{n+1}$. From these vectors we form the polynomials

$$\mathbf{p}_\pm(t) = (t \pm 1)\mathbf{u}_+(t) \quad \text{and} \quad \mathbf{q}_\pm(t) = (t \mp 1)\mathbf{v}_-(t)\,.$$

Clearly, $\mathbf{p}_+$ and $\mathbf{q}_+$ are symmetric, and $\mathbf{p}_-$ and $\mathbf{q}_-$ are skewsymmetric.

Proposition 2.15. *The symmetric T-Bezoutian $B = \mathrm{Bez}_T(\mathbf{u}_+, \mathbf{v}_-)$ can be represented as $B = B_+ + B_-$, where $B_\pm = BP_\pm$ and*

$$B_\pm = \mp\frac{1}{2}\,\mathrm{Bez}_{\mathrm{split}}(\mathbf{p}_\pm, \mathbf{q}_\pm)\,.$$

Proof. We compute the generating polynomial $B_+(t,s)$ of $B_+ = BP_+$. By definition we have

$$\begin{aligned} B_+(t,s) &= \frac{1}{2}\left(B(t,s) + B(t,s^{-1})s^{n-1}\right) \\ &= -\frac{1}{2}\left(\frac{\mathbf{u}_+(t)\mathbf{v}_-(s) + \mathbf{v}_-(t)\mathbf{u}_+(s)}{1-ts} + \frac{\mathbf{u}_+(t)\mathbf{v}_-(s) - \mathbf{v}_-(t)\mathbf{u}_+(s)}{t-s}\right) \\ &= -\frac{1}{2}\,\frac{(t+1)\mathbf{u}_+(t)(s-1)\mathbf{v}_-(s) - (t-1)\mathbf{v}_-(t)(s+1)\mathbf{u}_+(s)}{(t-s)(ts-1)}\,. \end{aligned}$$

This is just the generating polynomial of the matrix $-\frac{1}{2}\operatorname{Bez}_{\text{split}}(\mathbf{p}_+, \mathbf{q}_+)$. The other case is proved analogously. □

12. Relations between H- and T-Bezoutians. There is a simple relation between H- and T-Bezoutians, namely

$$\operatorname{Bez}_T(\mathbf{u},\mathbf{v}) = -\operatorname{Bez}_H(\mathbf{u},\mathbf{v})J_n\,.$$

More general relations can be described with the help of Frobenius-Fischer transformations. Analogously to Theorem 2.7 we obtain the following.

Theorem 2.16. *For any $\varphi \in \mathrm{GL}(\mathbb{F}^2)$, the transformation*

$$\Phi\,:\,B \mapsto K_n(\varphi)BK_n(J_2\varphi)^T$$

maps T-Bezoutians into H-Bezoutians. Moreover,

$$K_n(\varphi)\operatorname{Bez}_T(\mathbf{u},\mathbf{v})K_n(J_2\varphi)^T = \frac{1}{\det(\varphi J_2)}\operatorname{Bez}_H(\widetilde{\mathbf{u}},\widetilde{\mathbf{v}}) = \frac{1}{\det\varphi}\operatorname{Bez}_H(\widetilde{\mathbf{v}},\widetilde{\mathbf{u}})\,, \quad (2.18)$$

where $\widetilde{\mathbf{u}} = K_{n+1}(\varphi)\mathbf{u}$ and $\widetilde{\mathbf{v}} = K_{n+1}(\varphi)\mathbf{v}$.

In the case $\mathbb{F} = \mathbb{C}$ it is of particular interest to describe congruence transformations that transform Hermitian T-Bezoutians into real symmetric H-Bezoutians, in other words to describe a coordinate transformation that transforms Hermitian T-Bezoutian forms into real quadratic H-Bezoutian forms. The transformation Φ has this property if and only if $K_n(J_2\varphi)^T = K_n(\varphi)^*$ (up to multiples of I_2), which is equivalent to $J_2\varphi = \overline{\varphi}$. Suppose that $\varphi = \begin{bmatrix} a & c \\ b & d \end{bmatrix}$, then this is equivalent to $a = \overline{b}$ and $c = \overline{d}$. It can be easily checked that φ has this property if and only if the linear fractional function $\varphi(t) = \dfrac{at+b}{ct+d}$ maps the unit circle onto the real line (compare Section 1.5). Hence we have the following.

Corollary 2.17. *If $\varphi(t)$ maps the unit circle onto the real line, then the transformation $\Phi\,:\,B \mapsto K_n(\varphi)BK_n(\varphi)^*$ maps Hermitian T-Bezoutians into real symmetric H-Bezoutians. In particular, the signatures of B and $\Phi(B)$ coincide.*

3. Resultant matrices and matrix representations of Bezoutians

In this section we show that Bezoutians are closely related to resultant matrices and that the relations between these two classes can be used to derive important matrix representations of Bezoutians. We present two kinds of relations between resultant matrices and Bezoutians. The first is due to Kravitsky and Russakovsky, the second an interpretation of Bezoutians as Schur complements in resultant matrices.

The resultant matrix $\mathrm{Res}(\mathbf{u},\mathbf{v})$ of two polynomials $\mathbf{u}(t) \in \mathbb{F}^{m+1}, \mathbf{v}(t) \in \mathbb{F}^{n+1}(t)$ was introduced in (2.4) as the $(m+n)\times(m+n)$ matrix

$$\mathrm{Res}\,(\mathbf{u},\mathbf{v}) = \begin{bmatrix} M_n(\mathbf{u})^T \\ M_m(\mathbf{v})^T \end{bmatrix}.$$

In this section we restrict ourselves to the case $m = n$, which is no restriction of generality when speaking about nonsingularity, rank and related quantities. Recall that $\mathrm{Res}\,(\mathbf{u},\mathbf{v})$ is nonsingular if and only if the polynomials $\mathbf{u}(t)$ and $\mathbf{v}(t)$ are coprime and at least one of the leading coefficients of $\mathbf{u}(t)$ or $\mathbf{v}(t)$ is not zero.

1. Kravitsky-Russakovsky formulas. To begin with we generalize the resultant concept. Let $\mathbf{u}(t)$ and $\mathbf{v}(t)$ be polynomials of degree n. The *p-resultant matrix* $(p = 0, 1, \dots)$ of $\mathbf{u}(t)$ and $\mathbf{v}(t)$ is, by definition, the $(2n+2p)\times(2n+p)$ matrix

$$\mathrm{Res}_p(\mathbf{u},\mathbf{v}) = \begin{bmatrix} M_{n+p}(\mathbf{u})^T \\ M_{n+p}(\mathbf{v})^T \end{bmatrix}.$$

In the case $p = 0$ we have the resultant matrix in the former sense. For the sequel it is important to observe that

$$\mathrm{Res}_p(\mathbf{u},\mathbf{v})\ell_{2n+p}(t) = \begin{bmatrix} \mathbf{u}(t)\ell_{n+p}(t) \\ \mathbf{v}(t)\ell_{n+p}(t) \end{bmatrix}, \tag{3.1}$$

where $\ell_m(t) = (t^{i-1})_{i=1}^m$.

Theorem 3.1. *Let* $\mathbf{u}(t)$ *and* $\mathbf{v}(t)$ *be polynomials of degree* n. *Then*

1.
$$\mathrm{Res}_p(\mathbf{u},\mathbf{v})^T \begin{bmatrix} O & J_{n+p} \\ -J_{n+p} & O \end{bmatrix} \mathrm{Res}_p(\mathbf{u},\mathbf{v}) = \begin{bmatrix} O & O & -B_H \\ O & O & O \\ B_H & O & O \end{bmatrix}, \tag{3.2}$$

where $B_H = \mathrm{Bez}_H(\mathbf{u},\mathbf{v})$, *and*

2.
$$\mathrm{Res}_p(\mathbf{u},\mathbf{v})^T \begin{bmatrix} I_{n+p} & O \\ O & -I_{n+p} \end{bmatrix} \mathrm{Res}_p(\mathbf{v}^J,\mathbf{u}^J) = \begin{bmatrix} B_T & O & O \\ O & O & O \\ O & O & -B_T \end{bmatrix}, \tag{3.3}$$

where $B_T = \mathrm{Bez}_T(\mathbf{u},\mathbf{v})$.

Proof. We compare the generating polynomials of the right-hand and of the left-hand sides. According to (3.1) we have

$$\begin{aligned} & \ell_{2n+p}(t)^T \mathrm{Res}_p(\mathbf{u},\mathbf{v})^T \begin{bmatrix} O & J_{n+p} \\ -J_{n+p} & O \end{bmatrix} \mathrm{Res}_p(\mathbf{v},\mathbf{u})\, \ell_{2n+p}(s) \\ = \;& (\mathbf{u}(t)\mathbf{v}(s) - \mathbf{v}(t)\mathbf{u}(s))\ell_{n+p}(t)^T J_{n+p}\, \ell_{n+p}(s) \\ = \;& (\mathbf{u}(t)\mathbf{v}(s) - \mathbf{v}(t)\mathbf{u}(s))\, \frac{t^{n+p} - s^{n+p}}{t-s} \\ = \;& (t^{n+p} - s^{n+p})\mathrm{Bez}_H(\mathbf{u},\mathbf{v})(t,s)\;, \end{aligned}$$

which is the polynomial form of the first assertion.

To prove the second relation we observe that (3.1) implies

$$\begin{aligned} & \ell_{2n+p}(t)^T \mathrm{Res}_p(\mathbf{u},\mathbf{v})^T \begin{bmatrix} I_{n+p} & O \\ O & -I_{n+p} \end{bmatrix} \mathrm{Res}_p(\mathbf{v}^J,\mathbf{u}^J)\ell_{2n+p}(s) \\ = \;& \frac{1-(ts)^{n+p}}{1-ts}\,(\mathbf{u}(t)\mathbf{v}^J(s) - \mathbf{v}(t)\mathbf{u}^J(s))\,. \end{aligned}$$

This leads to the second assertion. □

2. Matrix representations of Bezoutians. The Kravitsky-Russakovsky formulas (3.2) and (3.3) provide an elegant way to obtain matrix representations of Bezoutians in terms of triangular Toeplitz matrices. These formulas are very important in connection with inversion of Toeplitz and Hankel matrices. They represent so-called "inversion formulas". Note that from computational point of view the formulas presented here are not the most efficient ones. Other, more efficient formulas for the cases $\mathbb{F} = \mathbb{C}$ or $\mathbb{F} = \mathbb{R}$ can be found in [38], [40], [41], [42], [43].

We define, for $\mathbf{u} = (u_i)_{i=1}^{n+1}$, the lower triangular $n \times n$ Toeplitz matrix

$$T(\mathbf{u}) = \begin{bmatrix} u_1 & & \\ \vdots & \ddots & \\ u_n & \dots & u_1 \end{bmatrix}.$$

Note that $T(\mathbf{u})$ is the T-Bezoutian $B_-(\mathbf{u})$ of $\mathbf{u}(t)$ and t^n, which was introduced in (2.14). Note also that the matrix $T(\mathbf{u})$ is related to the H-Bezoutian $B(\mathbf{u})$ defined by (2.2) and the T-Bezoutian $B_+(\mathbf{u})$ defined by (2.14) via

$$B(\mathbf{u}) = J_n T(\mathbf{u}^J)\;, \quad B_+(\mathbf{u}) = -T(\mathbf{u}^J)^T\,.$$

Furthermore, let us mention that we have commutativity

$$T(\mathbf{u}_1)T(\mathbf{u}_2) = T(\mathbf{u}_2)T(\mathbf{u}_1)$$

and the relation $T(\mathbf{u})^T = T(\mathbf{u})^J$. The nonsingular matrices $T(\mathbf{u})$ form a commutative subgroup of $GL(\mathbb{F}^n)$. With this notation the resultant matrix $\mathrm{Res}(\mathbf{u},\mathbf{v})$ for $\mathbf{u},\mathbf{v} \in \mathbb{F}^{n+1}$ can be written in the form

$$\mathrm{Res}(\mathbf{u},\mathbf{v}) = \begin{bmatrix} T(\mathbf{u})^T & T(\mathbf{u}^J) \\ T(\mathbf{v})^T & T(\mathbf{v}^J) \end{bmatrix}.$$

The application of Theorem 3.1 for $p = 0$ leads now to the following.

Theorem 3.2. *The H-Bezoutian of two polynomials* $\mathbf{u}(t), \mathbf{v}(t) \in \mathbb{F}^{n+1}$ *admits*

1. *the representations*

$$\mathrm{Bez}_H(\mathbf{u}, \mathbf{v}) = T(\mathbf{v})J_nT(\mathbf{u}^J) - T(\mathbf{u})J_nT(\mathbf{v}^J)$$

and

$$\mathrm{Bez}_H(\mathbf{u}, \mathbf{v}) = T(\mathbf{u}^J)^TJ_nT(\mathbf{v})^T - T(\mathbf{v}^J)^TJ_nT(\mathbf{u})^T\ .$$

2. *the representations*

$$\mathrm{Bez}_T(\mathbf{u}, \mathbf{v}) = T(\mathbf{u})T(\mathbf{v}^J)^T - T(\mathbf{v})T(\mathbf{u}^J)^T$$

and

$$\mathrm{Bez}_T(\mathbf{u}, \mathbf{v}) = T(\mathbf{v}^J)T(\mathbf{u})^T - T(\mathbf{u}^J)T(\mathbf{v})^T\ .$$

3. Bezoutians as Schur complements. We assume that the polynomial $\mathbf{u}(t)$ has degree n. Then the matrix $T(\mathbf{u}^J)$ is nonsingular. Now the second expression for $\mathrm{Bez}_H(\mathbf{u}, \mathbf{v})$ in Theorem 3.2,1 can be written in the form

$$C = T(\mathbf{u}^J)^{-1}J_n\mathrm{Bez}_H(\mathbf{u}, \mathbf{v}) = T(\mathbf{v})^T - T(\mathbf{v}^J)T(\mathbf{u}^J)^{-1}T(\mathbf{u})^T\ .$$

We see that C is the Schur complement of the left upper block in

$$\widetilde{R} = \mathrm{Res}\,(\mathbf{u}, \mathbf{v}) \begin{bmatrix} O & I_n \\ I_n & O \end{bmatrix} = \begin{bmatrix} T(\mathbf{u}^J) & T(\mathbf{u})^T \\ T(\mathbf{v}^J) & T(\mathbf{v})^T \end{bmatrix}\ .$$

Recall that the concept of Schur complement is defined in connection with the factorization of a block matrix

$$G = \begin{bmatrix} A & B \\ C & D \end{bmatrix} = \begin{bmatrix} I_n & O \\ CA^{-1} & I_n \end{bmatrix} \begin{bmatrix} A & O \\ O & D - CA^{-1}B \end{bmatrix} \begin{bmatrix} I_n & A^{-1}B \\ O & I_n \end{bmatrix}\ ,$$

where A is assumed to be invertible. Here $D - CA^{-1}B$ is said to be the *Schur complement of* A *in* G. Applying this factorization to our case we obtain the following.

Proposition 3.3. *Let* $\mathbf{u}(t) \in \mathbb{F}^{n+1}(t)$ *be a polynomial of degree* n, $\mathbf{v}(t) \in \mathbb{F}^{n+1}(t)$. *Then the resultant of* $\mathbf{u}(t)$ *and* $\mathbf{v}(t)$ *can be represented in the form*

$$\mathrm{Res}\,(\mathbf{u}, \mathbf{v}) = \begin{bmatrix} T(\mathbf{u}^J) & O \\ T(\mathbf{v}^J) & T(\mathbf{u}^J)^{-1}J_n \end{bmatrix} \begin{bmatrix} I_n & O \\ O & \mathrm{Bez}_H(\mathbf{u}, \mathbf{v}) \end{bmatrix} \begin{bmatrix} T(\mathbf{u}^J)^{-1}T(\mathbf{u})^T & I_n \\ I_n & O \end{bmatrix}.$$

From this proposition we see that $\mathrm{Res}(\mathbf{u}, \mathbf{v})$ is nonsingular if and only if $\mathrm{Bez}_H(\mathbf{u}, \mathbf{v})$ has this property. Hence $\mathrm{Bez}_H(\mathbf{u}, \mathbf{v})$ is nonsingular if and only if the polynomials $\mathbf{u}(t)$ and $\mathbf{v}(t)$ are coprime. Taking (2.10) into account we conclude the following.

Corollary 3.4. *The nullity of* $\mathrm{Bez}_H(\mathbf{u}, \mathbf{v})$ *is equal to the degree of the greatest common divisor of* $\mathbf{u}(t)$ *and* $\mathbf{v}(t)$.

Clearly, the same is also true for T-Bezoutians.

4. Inverses of Hankel and Toeplitz matrices

The most striking property of H- and T-Bezoutians is that inverses of Hankel and Toeplitz matrices belong to these classes. In view of Theorem 3.2 a consequence of this fact is that inverses of Toeplitz and Hankel matrices can be represented as product sum of triangular Toeplitz matrices, which is important for fast matrix-vector multiplication. Later, in Section 7.6 and Section 8.7 we will see that, vice versa, inverses of H- and T-Bezoutians are Hankel or Toeplitz matrices, respectively. Let us start with the Hankel case.

1. Inverses of Hankel matrices. Let $H_n = [\, s_{i+j-1} \,]_{i,j=1}^{n}$ be a nonsingular Hankel matrix. Besides H_n we consider the $(n-1) \times (n+1)$ Hankel matrix ∂H_n which is obtained from H_n after deleting the last row and adding another column on the right so that the Hankel structure is preserved. That means

$$\partial H_n = \begin{bmatrix} s_1 & \dots & s_{n+1} \\ \vdots & \iddots & \vdots \\ s_{n-1} & \dots & s_{2n-1} \end{bmatrix}. \tag{4.1}$$

For H_n is nonsingular, ∂H_n has a two-dimensional nullspace. A basis $\{\, \mathbf{u}, \mathbf{v} \,\}$ of the nullspace of ∂H_n will be called *fundamental system for* H_n. We consider for fixed $s \in \mathbb{F}$ the linear system of equations

$$H_n \mathbf{x}_s = \ell_n(s)\,, \tag{4.2}$$

where $\ell_n(s)$ is introduced in (1.1). It can be checked that

$$\partial H_n \begin{bmatrix} \mathbf{x}_s \\ 0 \end{bmatrix} = \ell_{n-1}(s) \quad \text{and} \quad \partial H_n \begin{bmatrix} 0 \\ \mathbf{x}_s \end{bmatrix} = s\ell_{n-1}(s)\,.$$

Hence $\begin{bmatrix} 0 \\ \mathbf{x}_s \end{bmatrix} - s \begin{bmatrix} \mathbf{x}_s \\ 0 \end{bmatrix}$ belongs to the kernel of ∂H_n. In polynomial language, this means that there are constants a_s and b_s such that

$$(t-s)\mathbf{x}_s(t) = a_s \mathbf{u}(t) - b_s \mathbf{v}(t)\,.$$

Now we consider s as a variable. From (4.2) it is clear that $\mathbf{x}_s(t) = \ell_n(t)^T H_n^{-1} \ell_n(s)$ is a polynomial in s of degree $n-1$. (It is just the generating polynomial of the matrix H_n^{-1}.) We conclude that $a_s = \mathbf{a}(s)$ and $b_s = \mathbf{b}(s) \in \mathbb{F}^{n+1}(s)$. Thus, H_n^{-1} is a quasi-H-Bezoutian. According to Corollary 2.5, this implies that H_n^{-1} is an H-Bezoutian, which means that $\mathbf{a}(t) = \gamma \mathbf{v}(t)$ and $\mathbf{b}(t) = \gamma \mathbf{u}(t)$, and $H_n^{-1} = \gamma \operatorname{Bez}_H(\mathbf{u}, \mathbf{v})$ for some nonzero constant γ. It remains to compute γ. For this we introduce the $2 \times (n+1)$ matrix

$$F = \begin{bmatrix} s_n & \dots & s_{2n-1} & s_{2n} \\ 0 & \dots & 0 & 1 \end{bmatrix}.$$

Here $s_{2n} \in \mathbb{F}$ is arbitrary. A fundamental system $\{\mathbf{u}, \mathbf{v}\}$ will be called *canonical* if

$$F[\, \mathbf{u} \;\; \mathbf{v} \,] = I_2\,.$$

Let $\{\mathbf{u}, \mathbf{v}\}$ be canonical. Then, in particular, $u := \mathbf{e}_{n+1}^T \mathbf{u} = 0$ and $v = \mathbf{e}_{n+1}^T \mathbf{v} = 1$. Furthermore, if we consider $\mathbf{u}$ as a vector in $\mathbb{F}^n$, then it is just the last column of H_n^{-1}, i.e.,

$$H_n \mathbf{u} = \mathbf{e}_n. \tag{4.3}$$

We compare $\mathbf{u}$ with the last column of $\mathrm{Bez}_H(\mathbf{u}, \mathbf{v})$, which is equal to $\mathbf{v}u - \mathbf{u}v = -\mathbf{u}$ (cf. Theorem 3.2). Thus, $\gamma = -1$. Note that $\mathbf{v}$ is of the form $\mathbf{v} = \begin{bmatrix} -\mathbf{z} \\ 1 \end{bmatrix}$, where $\mathbf{z}$ is the solution of the system

$$H_n \mathbf{z} = \mathbf{g} \quad \text{with} \quad \mathbf{g} = (s_{n+i})_{i=1}^n . \tag{4.4}$$

Hereafter we need the following fact.

Proposition 4.1. *Let the equations* (4.3) *and* (4.4) *be solvable. Then H_n is nonsingular.*

Proof. Assume that H_n is singular, and let $\mathbf{v} = (v_j)_{j=1}^n$ be a nontrivial vector such that $H_n \mathbf{v} = \mathbf{0}$. Then applying $\mathbf{v}^T$ from the left side to the equations (4.3) and (4.4) leads to

$$\mathbf{v}^T H_n \mathbf{u} = \mathbf{v}^T \mathbf{e}_n = \mathbf{0} \text{ and } \mathbf{v}^T H_n \mathbf{z} = \mathbf{v}^T \mathbf{g} = \mathbf{0},$$

which means, in particular, that $v_n = 0$. Taking into account

$$H_n S_n - S_n^T H_n = \mathbf{e}_n \mathbf{g}^T - \mathbf{g} \mathbf{e}_n^T$$

we conclude $(S_n \mathbf{v})^T H_n = \mathbf{0}$. Repeating the above arguments for the $S_n \mathbf{v}$ instead of $\mathbf{v}$ shows that $v_{n-1} = 0$, and so on. Finally we have $\mathbf{v} = \mathbf{0}$ which is a contradiction. Thus, the nonsingularity of H_n is proved. □

Now we consider a general fundamental system $\{\mathbf{u}, \mathbf{v}\}$. The matrix $\varphi = F[\,\mathbf{u} \;\; \mathbf{v}\,]$ is nonsingular. In fact, suppose it is singular. Then there is a nontrivial linear combination $\mathbf{w}(t)$ of $\mathbf{u}(t)$ and $\mathbf{v}(t)$ such that $F\mathbf{w} = \mathbf{0}$. In particular the highest-order coefficient vanishes, i.e., $\mathbf{w} \in \mathbb{F}^n$. Since $\mathbf{w} \in \ker \partial H_n$ we conclude that $H_n \mathbf{w} = \mathbf{0}$, which means that H_n is singular. The columns of $[\,\mathbf{u} \;\; \mathbf{v}\,]\varphi^{-1}$ form now a canonical fundamental system. It remains to apply Lemma 2.2 to obtain the following.

Theorem 4.2. *Let $\{\mathbf{u}, \mathbf{v}\}$ be a fundamental system for H_n. Then*

$$H_n^{-1} = \frac{1}{\det \varphi} \mathrm{Bez}_H(\mathbf{v}, \mathbf{u}) , \tag{4.5}$$

where $\varphi = F[\,\mathbf{u} \;\; \mathbf{v}\,]$.

Since $\mathrm{Bez}_H(\mathbf{u}, \mathbf{v})$ is nonsingular, the polynomials $\mathbf{u}(t)$ and $\mathbf{v}(t)$ must be coprime (cf. Corollary 3.4). Hence the following is true.

Corollary 4.3. *If $\{\mathbf{u}, \mathbf{v}\}$ is a fundamental system for a nonsingular Hankel matrix, then the polynomials $\mathbf{u}(t)$ and $\mathbf{v}(t)$ are coprime.*

2. Characterization of fundamental systems. There are several possibilities to characterize fundamental systems via solutions of special linear systems. We are mainly interested in characterizations by vectors that will be computed recursively using Levinson algorithms. In the Hankel case these vectors are the last columns $\mathbf{x}_k$ of H_k^{-1} or alternatively the monic solutions $\mathbf{u}_k$ of the Yule-Walker equations $H_k\mathbf{u}_k = \rho_k\mathbf{e}_k$, where ρ_k is so that $\mathbf{e}_k^T\mathbf{u}_k = 1$.

It is convenient to consider an $(n+1)\times(n+1)$ extension $H_{n+1} = [\, s_{i+j-1}\,]_{i,j=1}^{n+1}$. The matrix H_{n+1} is for almost all choices of s_{2n} and s_{2n+1} nonsingular. In fact, H_{n+1} is nonsingular if the Schur complement of the leading principal submatrix H_n in H_{n+1} is nonsingular. This Schur complement is equal to

$$s_{2n+1} - \mathbf{g}^T H_n^{-1}\mathbf{g}.$$

That means, for any s_{2n} there is only one value of s_{2n+1} for which H_{n+1} is singular. Now, since the vector $\begin{bmatrix} \mathbf{u}_n \\ 0 \end{bmatrix}$, which will be also denoted by $\mathbf{u}_n$, and the vector $\mathbf{u}_{n+1}$ are linearly independent and belong both to the kernel of ∂H_n they form a fundamental system for H_n. To compute the factor $\dfrac{1}{\det\varphi}$ in (4.5) we observe that

$$F\begin{bmatrix} \mathbf{u}_n & \mathbf{u}_{n+1} \end{bmatrix} = \begin{bmatrix} \rho_n & 0 \\ 0 & 1 \end{bmatrix}.$$

For the corresponding vectors $\mathbf{x}_n$ and $\mathbf{x}_{n+1}$ we find that

$$F\begin{bmatrix} \mathbf{x}_n & \mathbf{x}_{n+1} \end{bmatrix} = \begin{bmatrix} 1 & 0 \\ 0 & \xi_{n+1} \end{bmatrix},$$

where ξ_{n+1} is the last component of $\mathbf{x}_{n+1}$.

Corollary 4.4. *The inverse of the Hankel matrix H_n is given by*

$$H_n^{-1} = \frac{1}{\rho_n}\operatorname{Bez}_H(\mathbf{u}_{n+1},\mathbf{u}_n) = \frac{1}{\xi_{n+1}}\operatorname{Bez}_H(\mathbf{x}_{n+1},\mathbf{x}_n)\,.$$

3. Christoffel-Darboux formula. We compare the first Bezoutian formula for Hankel matrix inversion of Corollary 4.4 with the UL-factorization of H_n^{-1} (see, e.g., [33]) which can be written in polynomial language as

$$H_n^{-1}(t,s) = \sum_{k=1}^{n}\frac{1}{\rho_k}\mathbf{u}_k(t)\mathbf{u}_k(s)\,.$$

We conclude

$$\sum_{k=1}^{n}\frac{1}{\rho_k}\mathbf{u}_k(t)\mathbf{u}_k(s) = \frac{1}{\rho_n}\,\frac{\mathbf{u}_{n+1}(t)\mathbf{u}_n(s) - \mathbf{u}_n(t)\mathbf{u}_{n+1}(s)}{t-s}\,. \tag{4.6}$$

This relation is called *Christoffel-Darboux formula*. It is important in the theory of orthogonal polynomials.

4. Inverses of Toeplitz matrices. The proof of the fact that inverses of Toeplitz matrices are T-Bezoutians follows the same lines as that for Hankel matrices. We

introduce the $(n-1)\times(n+1)$ Toeplitz matrix ∂T_n obtained from $T_n = [\, a_{i-j} \,]_{i,j=1}^{n}$ after deleting the first row and adding another column to the right by preserving the Toeplitz structure,

$$\partial T_n = \begin{bmatrix} a_1 & a_0 & \dots & a_{2-n} & a_{1-n} \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ a_{n-1} & a_{n-2} & \dots & a_0 & a_{-1} \end{bmatrix}. \tag{4.7}$$

If T_n is nonsingular, then ∂T_n has a two-dimensional nullspace. Each basis of this subspace is called *fundamental system for* T_n. The role of the matrix F is taken by

$$F = \begin{bmatrix} a_0 & \dots & a_{1-n} & a_{-n} \\ 0 & \dots & 0 & 1 \end{bmatrix},$$

where a_{-n} is arbitrary.

Theorem 4.5. *Let* $\{\mathbf{u}, \mathbf{v}\}$ *be a fundamental system for* T_n. *Then*

$$T_n^{-1} = \frac{1}{\det \varphi}\, \mathrm{Bez}_T(\mathbf{u}, \mathbf{v}) \,,$$

where $\varphi = F[\, \mathbf{u} \;\; \mathbf{v} \,]$.

The Toeplitz analogue of Proposition 4.1 is now as follows.

Proposition 4.6. *Let the equations*

$$T_n \mathbf{y} = \mathbf{e}_1 \quad \text{and} \quad T_n \mathbf{z} = \mathbf{f}^J$$

with $\mathbf{f} = (a_{-i})_{i=1}^{n}$ *be solvable. Then* T_n *is nonsingular.*

Taking into account that

$$T_n S_n - S_n T_n = \mathbf{e}_1 \mathbf{f}^T - \mathbf{f}^J \mathbf{e}_n^T$$

the proof of this proposition is analogous to that one of Proposition 4.1.

5. Characterization of fundamental systems. In the Toeplitz case the Levinson algorithm computes recursively the first and last columns $\mathbf{x}_k^-$ and $\mathbf{x}_k^+$ of T_k^{-1} or alternatively the solutions $\mathbf{u}_k^\pm$ of the Yule-Walker equations

$$T_k \mathbf{u}_k^- = \rho_k^- \mathbf{e}_1, \qquad \text{and} \qquad T_k \mathbf{u}_k^+ = \rho_k^+ \mathbf{e}_k \,, \tag{4.8}$$

where $\rho_k^\pm \in \mathbb{F}$ are so that

$$\mathbf{e}_1^T \mathbf{u}_k^- = 1 \qquad \text{and} \qquad \mathbf{e}_k^T \mathbf{u}_k^+ = 1 \,.$$

(In other words $\mathbf{u}_k^+(t)$ is assumed to be monic and $\mathbf{u}_k^-(t)$ *comonic*, which means that $(\mathbf{u}_k^-)^J(t)$ is monic.) So it is reasonable to describe the fundamental system with these vectors.

It can easily be seen that $\begin{bmatrix} \mathbf{x}_n^- \\ 0 \end{bmatrix}$ and $\begin{bmatrix} 0 \\ \mathbf{x}_n^+ \end{bmatrix}$ belong to the nullspace of ∂T_n and in the case where T_{n-1} is nonsingular they are linearly independent.

Thus, they form a fundamental system. Likewise $\begin{bmatrix} \mathbf{u}_n^- \\ 0 \end{bmatrix}$ and $\begin{bmatrix} 0 \\ \mathbf{u}_n^+ \end{bmatrix}$ form a fundamental system. We find that

$$F \begin{bmatrix} \mathbf{u}_n^- & 0 \\ 0 & \mathbf{u}_n^+ \end{bmatrix} = \begin{bmatrix} \rho_n & * \\ 0 & 1 \end{bmatrix}, \quad F \begin{bmatrix} \mathbf{x}_n^- & 0 \\ 0 & \mathbf{x}_n^+ \end{bmatrix} = \begin{bmatrix} 1 & * \\ 0 & \xi_n \end{bmatrix},$$

where ξ_n is the first component of $\mathbf{x}_n^-$ which equals the last component of $\mathbf{x}_n^+$. Consequently, $\rho_n^+ = \rho_n^- = \rho_n$.

We will have a problem with these systems, if the submatrix T_{n-1} is singular. For example, in this case the solution $\mathbf{u}_n^+$ does not exist and $\xi_n = 0$. For this reason we also consider, like in the Hankel case, an $(n+1) \times (n+1)$ Toeplitz extension $T_{n+1} = [\, a_{i-j} \,]_{i,j=1}^{n+1}$. This extension is nonsingular for almost all choices of $a_{\pm n}$. The proof of this fact is, however, less trivial than in the Hankel case.

The Schur complement of T_n in T_{n+1} is given by

$$\sigma = a_0 - (\mathbf{g}_+ + a_n \mathbf{e}_1)^T T_n^{-1} (\mathbf{g}_- + a_{-n} \mathbf{e}_1),$$

where $\mathbf{g}_\pm = [\, 0 \; a_{\pm(n-1)} \; \dots \; a_{\pm 1} \,]^T$. Hence

$$-\sigma = \xi a_n a_{-n} + \eta_- a_n + \eta_+ a_{-n} + \zeta, \tag{4.9}$$

where $\xi = \mathbf{e}_1^T T_n^{-1} \mathbf{e}_1 = \mathbf{e}_1^T \mathbf{x}_n^-$, $\zeta = \mathbf{g}_+^T T_n^{-1} \mathbf{g}_- - a_0$, and

$$\eta_- = \mathbf{e}_1^T T_n^{-1} \mathbf{g}_- = \mathbf{g}_-^T \mathbf{x}_n^+, \quad \eta_+ = \mathbf{g}_+^T T_n^{-1} \mathbf{e}_1 = \mathbf{g}_+^T \mathbf{x}_n^-.$$

If $\xi \neq 0$, which is equivalent to the nonsingularity of T_{n-1}, the set of pairs (a_n, a_{-n}) for which T_{n+1} is singular is a quadratic curve in $\mathbb{F}^2$. (Choosing, for example, $a_n = a_{-n}$ there are at most 2 values of a_n for which T_{n+1} is singular.)

We show that if $\xi = 0$, then $\eta_\pm \neq 0$. In fact, in the case where $\xi = 0$ we have $T_n S_n^T \mathbf{x}_n^- = \eta_+ \mathbf{e}_n$. Since T_n is assumed to be nonsingular, we have $\eta_+ \neq 0$. Analogously, $\eta_- \neq 0$. That means in the case $\xi = 0$ the pairs (a_n, a_{-n}) for which T_{n+1} is singular are on the graph of a polynomial of first degree. (Choose, for example, $a_{-n} = 0$, then T_{n+1} is nonsingular with the exception of one value of a_n.)

Let now T_{n+1} be a nonsingular Toeplitz extension of T_n, $\mathbf{x}_{n+1}^-$ the first and $\mathbf{x}_{n+1}^+$ the last column of T_{n+1}^{-1}. Furthermore, let $\mathbf{u}_{n+1}^\pm$ be the solutions of the corresponding Yule-Walker equations(4.8) for $k = n+1$. Then $\{\mathbf{x}_{n+1}^-, \mathbf{x}_{n+1}^+\}$ and $\{\mathbf{u}_{n+1}^-, \mathbf{u}_{n+1}^+\}$ are fundamental systems for T_n and

$$F \begin{bmatrix} \mathbf{x}_{n+1}^- & \mathbf{x}_{n+1}^+ \end{bmatrix} = \begin{bmatrix} 1 & 0 \\ * & \xi_{n+1} \end{bmatrix}, \quad F \begin{bmatrix} \mathbf{u}_{n+1}^- & \mathbf{u}_{n+1}^+ \end{bmatrix} = \begin{bmatrix} \rho_{n+1} & 0 \\ * & 1 \end{bmatrix}. \tag{4.10}$$

Corollary 4.7. *The inverse of the Toeplitz matrix T_n is given by*

$$T_n^{-1} = \frac{1}{\xi_{n+1}} \operatorname{Bez}_T(\mathbf{x}_{n+1}^-, \mathbf{x}_{n+1}^+) = \frac{1}{\rho_{n+1}} \operatorname{Bez}_T(\mathbf{u}_{n+1}^-, \mathbf{u}_{n+1}^+).$$

6. Inverses of symmetric Toeplitz matrices. We discuss now the case of a symmetric Toeplitz matrix T_n. Let T_{n+1} be a symmetric Toeplitz extension of T_n. Since in this case $\mathbf{g}_+ = \mathbf{g}_-$ we have in (4.9) $\eta_+ = \eta_-$. From this we conclude that T_{n+1} is nonsingular with the exception of at most two values of a_n. Thus, we may assume that T_{n+1} is nonsingular.

Since we have $\mathbf{x}_{n+1} := \mathbf{x}^+_{n+1} = (\mathbf{x}^-_{n+1})^J$, the vectors $\mathbf{w}^+_{n+1} = \mathbf{x}_{n+1} + \mathbf{x}^J_{n+1}$ and $\mathbf{w}^-_{n+1} = \mathbf{x}_{n+1} - \mathbf{x}^J_{n+1}$ form a fundamental system consisting of a symmetric and a skewsymmetric vector. The vectors $\mathbf{w}^\pm_{n+1}$ are the solutions of $T_{n+1}\mathbf{w}^\pm_{n+1} = \mathbf{e}_{n+1} \pm \mathbf{e}_1$ and

$$F\begin{bmatrix} \mathbf{w}^-_{n+1} & \mathbf{w}^+_{n+1} \end{bmatrix} = \begin{bmatrix} -1 & 1 \\ -\mathbf{w}^-_{n+1}(0) & \mathbf{w}^+_{n+1}(0) \end{bmatrix}.$$

Corollary 4.8. *The inverse of a nonsingular symmetric Toeplitz matrix T_n is given by*

$$T_n^{-1} = \frac{1}{\gamma}\operatorname{Bez}_T(\mathbf{w}^-_{n+1}, \mathbf{w}^+_{n+1}),$$

where $\gamma = \mathbf{w}^-_{n+1}(0) - \mathbf{w}^+_{n+1}(0)$.

One can show that for solving a system $T_n\mathbf{z} = \mathbf{b}$ it is sufficient to compute the vectors $\mathbf{w}^+_k$. So it is reasonable to ask whether it is possible to describe $\mathbf{w}^-_{n+1}$ in terms of $\mathbf{w}^+_k$. The following proposition gives an answer to this question. Let T_{n+2} be a nonsingular $(n+2)\times(n+2)$ symmetric Toeplitz extension of T_{n+1} and $\mathbf{w}^\pm_{n+2}$ the solutions of $T_{n+2}\mathbf{w}^\pm_{n+2} = \mathbf{e}_{n+2} \pm \mathbf{e}_1$.

Proposition 4.9. *The polynomials $\mathbf{w}^\pm_{n+1}$ are given by*

$$\mathbf{w}^\pm_{n+1}(t) = \frac{t\mathbf{w}^+_n(t) - c_\pm\mathbf{w}^+_{n+2}(t)}{1 \pm t}, \tag{4.11}$$

where $\mathbf{w}^+_{n+2}(1) \neq 0$ and $c_- = \mathbf{w}^+_n(1)/\mathbf{w}^+_{n+2}(1)$. If n is odd, then $\mathbf{w}^+_{n+2}(-1) \neq 0$ and $c_+ = -\mathbf{w}^+_n(-1)/\mathbf{w}^+_{n+2}(-1)$. If n is even, then $\mathbf{w}^+_{n+2}(-1) = 0$ and c_+ is not determined by $\mathbf{w}^+_n$ and $\mathbf{w}^+_{n+2}$ alone.

Proof. We have

$$T_{n+2}\begin{bmatrix} \mathbf{w}^\pm_{n+1} & 0 \\ 0 & \mathbf{w}^\pm_{n+1} \end{bmatrix} = \begin{bmatrix} \pm 1 & \pm a_\pm \\ 0 & \pm 1 \\ \mathbf{0} & \mathbf{0} \\ 1 & 0 \\ a_\pm & 1 \end{bmatrix}, \quad T_{n+2}\begin{bmatrix} 0 & \\ \mathbf{w}^+_n & \mathbf{w}^+_{n+2} \\ 0 & \end{bmatrix} = \begin{bmatrix} b & 1 \\ 1 & 0 \\ \mathbf{0} & \mathbf{0} \\ 1 & 0 \\ b & 1 \end{bmatrix}$$

for some $a_\pm, b \in \mathbb{F}$. Consequently,

$$\begin{bmatrix} \mathbf{w}^\pm_{n+1} \\ 0 \end{bmatrix} \pm \begin{bmatrix} 0 \\ \mathbf{w}^\pm_{n+1} \end{bmatrix} = \begin{bmatrix} 0 \\ \mathbf{w}^+_n \\ 0 \end{bmatrix} - c_\pm w^+_{n+2}$$

for some $c_\pm \in \mathbb{F}$. Writing this in polynomial language, we see that $\mathbf{w}^\pm_{n+1}(t) \pm t\mathbf{w}^\pm_{n+1}(t) = t\mathbf{w}^+_n(t) - c_\pm\mathbf{w}^+_{n+2}(t)$ and obtain (4.11).

To prove the rest of the proposition we recall that the polynomials $\mathbf{w}^+_{n+1}(t)$ and $\mathbf{w}^-_{n+1}(t)$ form a fundamental system. Therefore, they are coprime. Suppose that $\mathbf{w}^+_{n+2}(1) = 0$. Then (4.11) implies $\mathbf{w}^+_{n+1}(1) = 0$. But we have also $\mathbf{w}^-_{n+1}(1) = 0$, since $\mathbf{w}^-_{n+1}$ is skewsymmetric. This contradicts the coprimeness of $\mathbf{w}^+_{n+1}(t)$ and $\mathbf{w}^-_{n+1}(t)$. Consequently, $\mathbf{w}^+_{n+2}(1) \neq 0$. Analogously, if n is odd and $\mathbf{w}^+_{n+2}(-1) = 0$, then (4.11) implies $\mathbf{w}^-_{n+1}(-1) = 0$. But we have also $\mathbf{w}^+_{n+1}(-1) = 0$, since $\mathbf{w}^+_{n+1}$ is symmetric and has an even length. This contradiction shows that $\mathbf{w}^+_{n+2}(-1) \neq 0$. If n is even, then T_n is not completely determined by its restriction to symmetric vectors. That means $\mathbf{w}^+_{n+1}$ is not completely given by $\mathbf{w}^+_n$ and $\mathbf{w}^+_{n+2}$. □

If n is even, then the constant c_+ can be obtained by applying a test functional, which could be the multiplication by any row of T_{n+1}.

7. Inverses of skewsymmetric Toeplitz matrices. In the case of a nonsingular skewsymmetric Toeplitz matrix T_n, $n = 2m$, the Levinson-type algorithm can be used to compute vectors spanning the nullspace of T_{2k-1} for $k = 1, \dots, m$. So it is reasonable to ask for a fundamental system $\{\mathbf{u}, \mathbf{v}\}$ consisting of vectors of this kind.

Let $\mathbf{x}$ be any vector spanning the nullspace of T_{n-1}. From the relation $T^J_{n-1} = -T_{n-1}$ follows that also the vector $\mathbf{x}^J$ belongs to the nullspace of T_{n-1}. Thus $\mathbf{x}$ is either symmetric or skewsymmetric. We show that the latter is not possible.

Lemma 4.10. *The vector* $\mathbf{x}$ *is symmetric.*

Proof. Let $\mathbf{f}_j$ denote the jth row of T_{n-1}, $n = 2m$. State that the row $\mathbf{f}_m$ in the middle of T_{n-1} is skewsymmetric. We introduce vectors $\mathbf{f}^\pm_j = \mathbf{f}_j \mp \mathbf{f}_{n-j}$ for $j = 1, \dots, m-1$. Then the $\mathbf{f}^+_j$ are symmetric, the $\mathbf{f}^-_j$ are skewsymmetric, $\mathbf{f}^\pm_j \in \mathbb{F}^{n-1}_\pm$, and the system $T_{n-1}\mathbf{v} = O$ is equivalent to $\mathbf{f}^\pm_j \mathbf{v} = 0$ for $j = 1, \dots, m-1$ and $\mathbf{f}_m \mathbf{v} = 0$. Since $\dim \mathbb{F}^{n-1}_+ = \frac{n}{2}$, there exists a symmetric vector $\mathbf{v} \neq O$ such that $\mathbf{f}^+_j \mathbf{v} = 0$ for $j = 1, \dots, m-1$. Since, obviously, $\mathbf{f}^-_j \mathbf{v} = O$ and $\mathbf{f}_m \mathbf{v} = 0$ we have $T_{n-1}\mathbf{v} = O$. Taking into account that $\dim \ker T_{n-1} = 1$ we conclude that $\mathbf{x} = c\mathbf{v}$ for some $c \in \mathbb{F}$. Thus, $\mathbf{x} \in \mathbb{F}^{n-1}_+$. □

Now, by Lemma 4.10, $\mathbf{x}$ is symmetric and

$$\mathbf{u} = \begin{bmatrix} 0 \\ \mathbf{x} \\ 0 \end{bmatrix} \in \ker \partial T_n \,.$$

(Since we do not want to assume that T_{n-2} is nonsingular, we cannot assume that $\mathbf{x}$ is monic.) Furthermore, let T_{n+1} be any $(n+1)\times(n+1)$ skewsymmetric Toeplitz extension of T_n and $\mathbf{v}$ a (symmetric) vector spanning the nullspace of T_{n+1}. Since T_n is nonsingular, we may assume that $\mathbf{v}$ is monic. Now $\{\mathbf{u}, \mathbf{v}\}$ is a fundamental system, and

$$F \begin{bmatrix} \mathbf{u} & \mathbf{v} \end{bmatrix} = \begin{bmatrix} \gamma & 0 \\ 0 & 1 \end{bmatrix}, \quad \gamma = [\, a_1 \dots a_{n-1} \,]\mathbf{x} \,.$$

(Here $\gamma \neq 0$ since otherwise $\begin{bmatrix} 0 \\ \mathbf{x} \end{bmatrix}$ belongs to the kernel of T_n.) Thus we obtain the following.

Corollary 4.11. *The inverse of the nonsingular skewsymmetric Toeplitz matrix T_n is given by*

$$T_n^{-1} = \frac{1}{\gamma} \operatorname{Bez}_T(\mathbf{u}, \mathbf{v}) .$$

8. Inverses of Hermitian Toeplitz matrices. Finally we discuss the case of a nonsingular Hermitian Toeplitz matrix T_n. Besides T_n we consider an $(n+1)\times(n+1)$ Hermitian Toeplitz extension T_{n+1} of T_n. With similar arguments as above one can show that for almost all values of a_n the matrix T_{n+1} is nonsingular, so we may assume this. In the Hermitian case we have for the first and last columns $\mathbf{x}_{n+1}^-, \mathbf{x}_{n+1}^+$ of T_{n+1}^{-1} that

$$\mathbf{x}_{n+1} := \mathbf{x}_{n+1}^- = (\mathbf{x}_{n+1}^+)^\#$$

and for the solutions $\mathbf{u}_{n+1}^\pm$ of the Yule-Walker equations

$$\mathbf{u}_{n+1} := \mathbf{u}_{n+1}^- = (\mathbf{u}_{n+1}^+)^\# .$$

Taking Corollary 4.7 into account we obtain

$$T_n^{-1} = \frac{1}{\xi_{n+1}} \operatorname{Bez}_T(\mathbf{x}_{n+1}, \mathbf{x}_{n+1}^\#) = \frac{1}{\rho_{n+1}} \operatorname{Bez}_T(\mathbf{u}_{n+1}, \mathbf{u}_{n+1}^\#) , \tag{4.12}$$

where ξ_{n+1} is the first component of $\mathbf{x}_{n+1}$ and ρ_{n+1} is so that $\mathbf{u}_{n+1}(t)$ is comonic.

In the Levinson-type algorithm described in [53], [47] not the vectors $\mathbf{x}_k$ are computed but the solutions of the equations $T_k\mathbf{q}_k = \mathbf{e}$, where $\mathbf{e}$ is the vector all components of which are equal to 1. For an inversion formula we need the vectors $\mathbf{q}_n$ and $\mathbf{q}_{n+1}$. Since $\mathbf{q}_{n+1}$ and $\mathbf{q}_n$ are conjugate-symmetric, $\mathbf{q}_{n+1}(t) - t\mathbf{q}_n(t)$ is not identically equal to zero. Hence

$$\mathbf{x}_{n+1}(t) = b(\mathbf{q}_{n+1}(t) - t\mathbf{q}_n(t)) \tag{4.13}$$

for some nonzero $b \in \mathbb{C}$.

Besides $\mathbf{q}_n$ we consider the coefficient vector $\mathbf{w}$ of $\mathbf{w}(t) = \mathrm{i}\,(t-1)\mathbf{q}_n(t)$, which is obviously conjugate-symmetric.

Proposition 4.12. *The inverse of a nonsingular Hermitian Toeplitz matrix T_n is given by*

$$T_n^{-1} = \frac{\mathrm{i}}{c} \operatorname{Bez}_T(\mathbf{w}, \mathbf{q}_{n+1}) - \frac{1}{c}\, \mathbf{q}_n \overline{\mathbf{q}}_n^T , \tag{4.14}$$

where c is the real constant $\mathbf{q}_{n+1}(1) - \mathbf{q}_n(1)$.

Proof. We insert (4.13) into (4.12) and obtain, after an elementary calculation, formula (4.14) with $c = \dfrac{\xi_{n+1}}{|b|^2} \neq 0$. Taking into account that $\mathbf{q}_n(t) = (T_n^{-1}\mathbf{e})(t) = T_n^{-1}(t,1)$ and that, due to (4.14), $T_n^{-1}(t,1) = \dfrac{1}{c}\,\mathbf{q}_n(t)(\overline{\mathbf{q}}_{n+1}(1) - \overline{\mathbf{q}}_n(1))$ we find that $c = \overline{c} = \mathbf{q}_{n+1}(1) - \mathbf{q}_n(1)$. □

9. Solution of systems. The formulas for the inverses of Toeplitz and Hankel matrices presented in this section can be used in combination with the matrix representations of Bezoutians to solve Toeplitz and Hankel systems. This is in particular convenient if systems have to be solved with different right-hand sides and one and the same coefficient matrix. The advantage compared with factorization methods is that only $O(n)$ parameters have to be stored.

The application of the formulas requires 4 matrix-vector multiplications by triangular Toeplitz matrices. If these multiplications are carried out in the classical way, then $2\,n^2$ multiplications and $2\,n^2$ additions are needed, which is more than, for example, if back substitution in the LU-factorization is applied. However, due to the Toeplitz structure of the matrices there are faster methods, actually methods with a complexity less than $O(n^2)$, to do this. In the cases $\mathbb{F} = \mathbb{C}$ and $\mathbb{F} = \mathbb{R}$ the Fast Fourier and related real trigonometric transformations with a computational complexity of $O(n \log n)$ can be applied.

5. Generalized triangular factorizations of Bezoutians

In this section we describe algorithms that lead to a generalized UL-factorization of Bezoutians. In the case of H-Bezoutians the algorithm is just the Euclidian algorithm.

1. Division with remainder. Suppose that $\mathbf{u} = (u_i)_{i=1}^{n+1} \in \mathbb{F}^{n+1}$,$\mathbf{v} = (v_i)_{i=1}^{m+1} \in \mathbb{F}^{m+1}$, $m \leq n$, and that the last components of $\mathbf{u}$ and $\mathbf{v}$ are not zero. Division with remainder means to find polynomials $\mathbf{q}(t) \in \mathbb{F}^{n-m+1}(t)$ and $\mathbf{r}(t) \in \mathbb{F}^{m}(t)$ such

$$\mathbf{u}(t) = \mathbf{q}(t)\mathbf{v}(t) + \mathbf{r}(t)\,. \tag{5.1}$$

In matrix language this means that we first solve the $(n-m+1) \times (n-m+1)$ triangular Toeplitz system

$$\begin{bmatrix} v_{m+1} & \dots & v_{n-2m+1} \\ & \ddots & \vdots \\ & & v_{m+1} \end{bmatrix} \mathbf{q} = \begin{bmatrix} u_{m+1} \\ \vdots \\ u_{n+1} \end{bmatrix},$$

where we put $v_i = 0$ for $i \notin \{1, \dots, m+1\}$. With the notation (2.3) we find $\mathbf{r}$ via

$$\begin{bmatrix} \mathbf{r} \\ \mathbf{0} \end{bmatrix} = \mathbf{u} - M_{n-m+1}(\mathbf{v})\mathbf{q}\,.$$

2. Factorization step for H-Bezoutians. We clarify what means division with remainder in terms of the H-Bezoutian. From (5.1) we obtain for $B = \mathrm{Bez}_H(\mathbf{u}, \mathbf{v})$

$$B(t,s) = \mathbf{v}(t)\,\frac{\mathbf{q}(t) - \mathbf{q}(s)}{t-s}\,\mathbf{v}(s) + \frac{\mathbf{r}(t)\mathbf{v}(s) - \mathbf{v}(t)\mathbf{r}(s)}{t-s}\,,$$

which can be written in the form

$$\mathrm{Bez}_H(\mathbf{u}, \mathbf{v}) = M_{n-m}(\mathbf{v})B(\mathbf{q})M_{n-m}(\mathbf{v})^T + \mathrm{Bez}_H(\mathbf{r}, \mathbf{v})\,,$$

where $B(\mathbf{q})$ is defined in (2.2). This is equivalent to

$$\mathrm{Bez}_H(\mathbf{u},\mathbf{v}) = \begin{bmatrix} I_m & M_{n-m}(\mathbf{v}) \\ O & \end{bmatrix} \begin{bmatrix} \mathrm{Bez}_H(\mathbf{r},\mathbf{v}) & O \\ O & B(\mathbf{q}) \end{bmatrix} \begin{bmatrix} I_m & M_{n-m}(\mathbf{v}) \\ O & \end{bmatrix}^T . \tag{5.2}$$

Note that in this equation the left factor is upper triangular and the right factor is lower triangular.

3. Euclidian algorithm. The Euclidian algorithm is the successive application of division with remainder. We set $\mathbf{u}_0(t) = \mathbf{u}(t)$ and $\mathbf{u}_1(t) = \mathbf{v}(t)$. Then, for $i = 1, 2, \dots$ we find $-\mathbf{u}_{i+1}(t)$ as remainder of the division of $\mathbf{u}_{i-1}(t)$ by $\mathbf{u}_i(t)$, i.e.,

$$\mathbf{u}_{i+1}(t) = \mathbf{q}_i(t)\mathbf{u}_i(t) - \mathbf{u}_{i-1}(t)\,. \tag{5.3}$$

We take the minus sign with $\mathbf{u}_{i+1}(t)$ for convenience. With this definition one factorization step (5.2) reduces $\mathrm{Bez}_H(\mathbf{u}_{i-1},\mathbf{u}_i)$ to $\mathrm{Bez}_H(\mathbf{u}_i,\mathbf{u}_{i+1})$, without the minus to $\mathrm{Bez}_H(\mathbf{u}_{i+1},\mathbf{u}_i)$. If for some $i = l$ we have $\mathbf{u}_{l+1}(t) = 0$, then the algorithm is terminated and $\mathbf{u}_l(t)$ is the greatest common divisor of $\mathbf{u}(t)$ and $\mathbf{v}(t)$.

There are modifications of the Euclidian algorithm with different normalizations. One could assume, for example, that the polynomials $\mathbf{u}_i(t)$ are monic or that the polynomials $\mathbf{q}_i(t)$ are monic. For these two possibilities one has to admit a constant factor at $\mathbf{u}_{i-1}$.

4. Generalized UL-factorization of H-Bezoutians. Applying (5.2) successively to the polynomials $\mathbf{u}_i(t)$ appearing in the Euclidian algorithm (5.3) we arrive at the following.

Theorem 5.1. *Let* $\mathbf{u}(t), \mathbf{v}(t) \in \mathbb{F}^{n+1}(t)$ *where* $\mathbf{u}(t)$ *has a nonzero leading coefficient. Then* $\mathrm{Bez}_H(\mathbf{u},\mathbf{v})$ *admits the factorization*

$$\mathrm{Bez}_H(\mathbf{u},\mathbf{v}) = UDU^T \tag{5.4}$$

where

$$U = \begin{bmatrix} I_d & M_{n_l}(\mathbf{u}_l) & \dots & M_{n_1-n_2}(\mathbf{u}_2) & M_{n-n_1}(\mathbf{u}_1) \\ O & O & & O & \end{bmatrix},$$

$n_i = \deg \mathbf{u}_i(t)$, *$d$ is the degree of the greatest common divisor of* $\mathbf{u}(t)$ *and* $\mathbf{v}(t)$, *and*

$$D = \mathrm{diag}\,(\,O_d,\, B(\mathbf{q}_l),\, \dots,\, B(\mathbf{q}_1)\,)\,. \tag{5.5}$$

Note that U is a nonsingular upper triangular matrix, $B(\mathbf{q}_i)$ are nonsingular upper triangular Hankel matrices defined by (2.2). From the theorem we can again conclude that the nullity of $\mathrm{Bez}_H(\mathbf{u},\mathbf{v})$ is equal to d (cf. Corollary 3.4).

Let us discuss the case, in which $\mathbf{u}(t)$ and $\mathbf{v}(t)$ are coprime and all polynomials $\mathbf{q}_i(t)$ have degree 1. This is called the *generic case*. Suppose that $\mathbf{q}_i(t) = \rho_i t + \delta_i$. Then (5.4) turns into an UL-factorization of the H-Bezoutian in which, U is the upper triangular matrix the kth row of which is equal to the transpose of $\begin{bmatrix} \mathbf{u}_{n+1-k} \\ \mathbf{0}_{k-1} \end{bmatrix}$ and $D = \mathrm{diag}\,(\rho_{n+1-i})_{i=1}^n$. The UL-factorization of the H-Bezoutian exists if and only if the matrix $\mathrm{Bez}_H(\mathbf{u},\mathbf{v})^J$ is strongly nonsingular. That means

in the generic case the matrix has this property. The converse is also true, since in the non-generic case the matrix $\mathrm{Bez}_H(\mathbf{u},\mathbf{v})^J$ has singular leading principal submatrices.

5. Inertia computation. It is an important consequence of Theorem 5.1 that the signature of a real H-Bezoutian can be computed via running the Euclidian algorithm. In fact, in the case of real polynomials $\mathbf{u}(t)$ and $\mathbf{v}(t)$ the matrix $\mathrm{Bez}_H(\mathbf{u},\mathbf{v})$ is congruent to the block diagonal matrix D given by (5.5). It remains to compute the signature of $B(\mathbf{q}_i)$.

Let ρ_i denote the leading coefficient of $\mathbf{q}_i(t)$. Then the signature of $B(\mathbf{q}_i)$ is equal to the signature of $\rho_i J_{m_i}$, $m_i = n_{i-1} - n_i$. This can be shown using a homotopy argument. Let $H(t) = t\rho_i J_{m_i} + (1-t)B(\mathbf{q}_i)$ for $0 \le t \le 1$. Then $H(0) = B(\mathbf{q}_i)$ and $H(1) = \rho_i J_{m_i}$. Furthermore, $H(t)$ is nonsingular for all t and depends continuously on t. Hence $\operatorname{sgn} H(t)$ is constant for $0 \le t \le 1$. The signature of $\rho_i J_{m_i}$ is obviously equal to zero if m_i is even and is equal to the sign of ρ_i if m_i is odd. Applying Sylvester's inertia law we conclude the following.

Corollary 5.2. *The signature of the real H-Bezoutian* $\mathrm{Bez}_H(\mathbf{u},\mathbf{v})$ *is given by*

$$\operatorname{sgn} \mathrm{Bez}_H(\mathbf{u},\mathbf{v}) = \sum_{n_{i-1}-n_i \text{ odd}} \operatorname{sgn} \rho_i \,,$$

where ρ_i *are the antidiagonal entries of* $B(\mathbf{q}_i)$.

Since the Euclidian algorithm computes besides the signature s also the rank r of the H-Bezoutian, it gives a complete picture about the inertia of $\mathrm{Bez}_H(\mathbf{u},\mathbf{v})$,

$$\operatorname{In} \mathrm{Bez}_H(\mathbf{u},\mathbf{v}) = (s_+, s_-, d),$$

where $s_\pm = \frac{n-d\pm s}{2}$ and $d = n - r$.

6. Factorization step for T-Bezoutians in the generic case. We consider the problem of triangular factorization of a T-Bezoutian $B = \mathrm{Bez}_T(\mathbf{u},\mathbf{v})$, where $\mathbf{u},\mathbf{v} \in \mathbb{F}^{n+1}$. This problem is more complicated than for H-Bezoutians, unless the matrix is strongly nonsingular. We introduce the 2×2 matrix

$$\Gamma = \Gamma(\mathbf{u},\mathbf{v}) = \begin{bmatrix} \mathbf{e}_1^T\mathbf{u} & \mathbf{e}_1^T\mathbf{v} \\ \mathbf{e}_{n+1}^T\mathbf{u} & \mathbf{e}_{n+1}^T\mathbf{v} \end{bmatrix}.$$

The case of nonsingular Γ is referred to as *generic case*, the case of singular Γ as *non-generic case*. In this subsection we consider the generic case. Observe that $\gamma := B(0,0) = \det \Gamma$. That means that γ is the entry in the left upper corner of B. Thus, we have the generic case if B is strongly nonsingular. Note that γ is also the entry in the right lower corner of B, due to the persymmetry of B. In the generic case, $\begin{bmatrix} \widetilde{\mathbf{u}} & \widetilde{\mathbf{v}} \end{bmatrix} = \begin{bmatrix} \mathbf{u} & \mathbf{v} \end{bmatrix} \Gamma^{-1}$ is of the form

$$\begin{bmatrix} \widetilde{\mathbf{u}} & \widetilde{\mathbf{v}} \end{bmatrix} = \begin{bmatrix} \mathbf{u}_1 & 0 \\ 0 & \mathbf{v}_1 \end{bmatrix}, \quad \mathbf{u}_1, \mathbf{v}_1 \in \mathbb{F}^n \,.$$

According to Lemma 2.9 we have

$$\mathrm{Bez}_T(\widetilde{\mathbf{u}},\widetilde{\mathbf{v}}) = \frac{1}{\gamma}\,\mathrm{Bez}_T(\mathbf{u},\mathbf{v})\,.$$

Furthermore, for $\widetilde{B} = \mathrm{Bez}_T(\widetilde{\mathbf{u}},\widetilde{\mathbf{v}})$ we obtain

$$\begin{aligned}\widetilde{B}(t,s) &= \frac{\mathbf{u}_1(t)\mathbf{v}_1(s^{-1})s^{n-1} - ts\mathbf{v}_1(t)\mathbf{u}_1(s^{-1})s^{n-1}}{1-ts}\\ &= B_1(t,s) + \mathbf{v}_1(t)\mathbf{u}_1(s^{-1})s^{n-1}\,,\end{aligned}$$

where $B_1 = \mathrm{Bez}_T(\mathbf{u}_1,\mathbf{v}_1)$. We also have

$$\widetilde{B}(t,s) = ts\,B_1(t,s) + \mathbf{u}_1(t)\mathbf{v}_1(s^{-1})s^{n-1}\,.$$

In matrix language this can be written as

$$\mathrm{Bez}_T(\mathbf{u},\mathbf{v}) = \begin{bmatrix} I_{n-1} & \mathbf{v}_1 \\ \mathbf{0}^T & \end{bmatrix}\begin{bmatrix} \gamma\,\mathrm{Bez}_T(\mathbf{u}_1,\mathbf{v}_1) & \mathbf{0} \\ \mathbf{0}^T & \gamma \end{bmatrix}\begin{bmatrix} I_{n-1} & \mathbf{u}_1^J \\ \mathbf{0}^T & \end{bmatrix}^T$$

or

$$\mathrm{Bez}_T(\mathbf{u},\mathbf{v}) = \begin{bmatrix} \mathbf{u}_1 & I_{n-1} \\ & \mathbf{0}^T \end{bmatrix}\begin{bmatrix} \gamma & \mathbf{0}^T \\ \mathbf{0} & \gamma\,\mathrm{Bez}_T(\mathbf{u}_1,\mathbf{v}_1) \end{bmatrix}\begin{bmatrix} \mathbf{v}_1^J & I_{n-1} \\ & \mathbf{0}^T \end{bmatrix}^T. \quad (5.6)$$

7. LU-factorization of T-Bezoutians. Let $B = \mathrm{Bez}_T(\mathbf{u},\mathbf{v})$ be strongly nonsingular, which is equivalent to the strongly nonsingularity of B^J, due to persymmetry. We can apply now the factorization step of the previous subsection, since the property of strongly nonsingularity is inherited after a factorization step. If we carry out the factorization step successively, then we obtain the following algorithm. We set $\mathbf{u}_1 = \mathbf{u}$ and $\mathbf{v}_1 = \mathbf{v}$ and find recursively polynomials $\mathbf{u}_k(t)$ and $\mathbf{v}_k(t)$ via

$$\begin{bmatrix} \mathbf{u}_{k+1}(t) & \mathbf{v}_{k+1}(t) \end{bmatrix} = \begin{bmatrix} \mathbf{u}_k(t) & \mathbf{v}_k(t) \end{bmatrix}\Gamma_k^{-1}\begin{bmatrix} 1 & 0 \\ 0 & t^{-1} \end{bmatrix}, \quad (5.7)$$

where $\Gamma_k = \Gamma(\mathbf{u}_k,\mathbf{v}_k)$. This algorithm has the same structure as the Schur algorithm for Toeplitz matrices. We call it also *Schur algorithm*. Like for Toeplitz matrices, it can be slightly modified by replacing the matrix Γ_k^{-1} by a matrix of the form $\begin{bmatrix} 1 & * \\ * & 1 \end{bmatrix}$. This will reduce the number of operations.

To simplify the notation we agree upon the following. For a sequence $(\mathbf{w}_j)_{j=1}^n$ with $\mathbf{w}_j \in \mathbb{F}^{n+1-j}$, by $L(\mathbf{w}_j)_{j=1}^n$ will be denoted the lower triangular matrix the kth column of which is equal to

$$L(\mathbf{w}_j)_{j=1}^n\mathbf{e}_k = \begin{bmatrix} \mathbf{0}_{k-1} \\ \mathbf{w}_k \end{bmatrix}.$$

Now we conclude the following from (5.6).

Theorem 5.3. *Let $B = \mathrm{Bez}_T(\mathbf{u},\mathbf{v})$ be strongly nonsingular, and let $\mathbf{u}_k(t)$ and $\mathbf{v}_k(t)$ be the polynomials obtained by the Schur algorithm* (5.7)*. Then B admits an LU-factorization*

$$B = LDU\,,$$

where

$$L = L(\mathbf{u}_i)_{i=1}^n\,, \quad U = (L(\mathbf{v}_i)_{i=1}^n)^T$$

and

$$D = \operatorname{diag}(\widetilde{\gamma}_i^{-1})_{i=1}^n\,, \quad \widetilde{\gamma}_i = \prod_{j=1}^{i} \gamma_j\,, \quad \gamma_j = \det \Gamma_j\,.$$

8. Non-generic case for T-Bezoutians. Now we consider the case where the matrix $\Gamma = \Gamma(\mathbf{u}, \mathbf{v})$ is singular. If Γ has a zero row, then $\mathbf{u}(t)$ and $\mathbf{v}(t)$ or $\mathbf{u}^J(t)$ and $\mathbf{v}^J(t)$ have a common factor t. Suppose that $\mathbf{u}(t) = t^{\mu_-}\mathbf{u}_0(t)$, $\mathbf{v}(t) = t^{\mu_-}\mathbf{v}_0(t)$, $\mathbf{u}^J(t) = t^{\mu_+}\mathbf{u}_0^J(t)$, and $\mathbf{v}^J(t) = t^{\mu_+}\mathbf{v}_0^J(t)$ such that $\Gamma(\mathbf{u}_0, \mathbf{v}_0)$ has no zero row. Then $B(t,s) = t^{\mu_-}s^{\mu_+}B_0(t,s)$, where $B_0 = \operatorname{Bez}_T(\mathbf{u}_0, \mathbf{v}_0)$ or, in matrix language

$$B = \begin{bmatrix} O & O & O \\ O & B_0 & O \\ O & O & O \end{bmatrix},$$

where the zero matrix in the left upper corner is $\mu_- \times \mu_+$.

Now we assume that Γ is singular but has no zero row. Then there is a 2×2 matrix Φ with $\det \Phi = 1$ such that the last column of $\Gamma\Phi$ is zero, but the first column consists of nonzero elements. We set

$$\begin{bmatrix} \widetilde{\mathbf{u}}(t) & \widetilde{\mathbf{v}}(t) \end{bmatrix} = \begin{bmatrix} \mathbf{u}(t) & \mathbf{v}(t) \end{bmatrix} \Phi.$$

Then, according to Lemma 2.9, we have $B = \operatorname{Bez}_T(\widetilde{\mathbf{u}}, \widetilde{\mathbf{v}}) = \operatorname{Bez}_T(\mathbf{u}, \mathbf{v})$. Furthermore, let us write $\widetilde{\mathbf{v}}$ in the form

$$\widetilde{\mathbf{v}} = \begin{bmatrix} \mathbf{0}_{\nu_-} \\ \mathbf{w} \\ \mathbf{0}_{\nu_+} \end{bmatrix}$$

with some vector $\mathbf{w} = (w_i)_{i=1}^{m+1} \in \mathbb{F}^{m+1}$, $m + \nu_+ + \nu_- = n$, with nonzero first and last components. We apply now a two-sided division with remainder to find polynomials $\mathbf{q}_-(t) \in \mathbb{F}^{\nu_-}(t)$, $\mathbf{q}_+(t) \in \mathbb{F}^{\nu_+ + 1}(t)$, and $\mathbf{r}(t) \in \mathbb{F}^m(t)$ such that

$$\widetilde{\mathbf{u}}(t) = (t^{\nu_-}\mathbf{q}_+(t) + \mathbf{q}_-(t))\mathbf{w}(t) + t^{\nu_-}\mathbf{r}(t)\,.$$

The vectors $\mathbf{q}_\pm$ can be found by solving the triangular Toeplitz systems

$$\begin{bmatrix} w_1 & & \\ \vdots & \ddots & \\ w_{\nu_-} & \dots & w_1 \end{bmatrix} \mathbf{q}_- = \begin{bmatrix} \widetilde{u}_1 \\ \vdots \\ \widetilde{u}_{\nu_-} \end{bmatrix},$$

$$\begin{bmatrix} w_{m+1} & \dots & w_{m-\nu_++1} \\ & \ddots & \vdots \\ & & w_{m+1} \end{bmatrix} \mathbf{q}_+ = \begin{bmatrix} \widetilde{u}_{n-\nu_++1} \\ \vdots \\ \widetilde{u}_{n+1} \end{bmatrix}.$$

Then we have

$$\mathbf{u}^J(t) = (t^{\nu_+ + 1}\mathbf{q}_-^J(t) + \mathbf{q}_+^J(t))\mathbf{w}^J(t) + t^{\nu_+ + 1}\mathbf{r}^J(t)$$

and

$$\begin{aligned} B(t,s) &= \mathbf{w}(t)\left(\frac{\mathbf{q}_-(t)-t^{\nu_-}\mathbf{q}_-^J(s)s}{1-ts}\,s^{\nu_+}+t^{\nu_-}\,\frac{\mathbf{q}_+(t)s^{\nu_+}-\mathbf{q}_+^J(s)}{1-ts}\right)\mathbf{w}^J(s)\\ &\quad +t^{\nu_-}\,\frac{\mathbf{r}(t)\mathbf{w}^J(s)-\mathbf{w}(t)\mathbf{r}^J(s)s}{1-ts}\,s^{\nu_+}\\ &= \mathbf{w}(t)\left(\mathrm{Bez}_T(\mathbf{q}_-,\mathbf{e}_{\nu_-+1})(t,s)s^{\nu_+}+t^{\nu_-}\mathrm{Bez}_T(\mathbf{q}_+,\mathbf{e}_1)(t,s)\right)\mathbf{w}^J(s)\\ &\quad +t^{\nu_-}\mathrm{Bez}_T(\mathbf{r},\mathbf{w})(t,s)s^{\nu_+}\,. \end{aligned}$$

In matrix form this can be written as

$$B=M_{\nu_++\nu_-}(\mathbf{w})\begin{bmatrix} O & B_-(\mathbf{q}_-)\\ B_+(\mathbf{q}_+) & O\end{bmatrix}M_{\nu_++\nu_-}(\mathbf{w}^J)^T+\begin{bmatrix} O & O & O\\ O & B_1 & O\\ O & O & O\end{bmatrix},$$

where $B_1=\mathrm{Bez}_T(\mathbf{r},\mathbf{w})$ is of order m, $B_+(\mathbf{q}_+)=\mathrm{Bez}_T(\mathbf{q}_+,\mathbf{e}_1)$ and $B_-(\mathbf{q}_-)=\mathrm{Bez}_T(\mathbf{q},\mathbf{e}_{\nu_-+1})$ are of order $\nu_\pm$ (cf. (2.14)), and the zero matrix in the left upper corner of the last term has size $\nu_-\times\nu_+$.

9. Hermitian T-Bezoutians. We discuss now the specifics of the case of an Hermitian T-Bezoutian B. Our main attention is dedicated to the question how to compute the signature, because this is the most important application of the procedure. First we remember that there are two possibilities to represent Hermitian T-Bezoutian. The first is $B=\mathrm{Bez}_T(\mathbf{u},\mathbf{u}^\#)$ for a general vector $\mathbf{u}\in\mathbb{C}^{n+1}$, the second is $B=\mathrm{i}\,\mathrm{Bez}_T(\mathbf{u}_+,\mathbf{u}_-)$ for two conjugate-symmetric vectors $\mathbf{u}_\pm$ (see Section 2.10).

In the generic case we use the first representation. In the representation (5.6) we have $\mathbf{v}_1^J=\overline{\mathbf{u}}_1$, and γ is real. Thus, for a strongly nonsingular B, Theorem 5.3 provides a factorization $B=LDL^*$. Consequently,

$$\operatorname{sgn}B=\sum_{i=1}^{n}\operatorname{sgn}\widetilde{\gamma}_i\,.$$

That means that the signature of B can be computed via the Schur algorithm in $O(n^2)$ operations.

In the non-generic case, i.e., in the case where Γ is singular, we switch from the first representation of B to the second one. This is done as follows. Suppose B is given as $B=\mathrm{Bez}_T(\mathbf{u},\mathbf{u}^\#)$. Then Γ is centro-Hermitian. Hence the homogeneous equation

$$\Gamma\begin{bmatrix}\alpha\\ \beta\end{bmatrix}=\begin{bmatrix}0\\ 0\end{bmatrix}$$

has a nontrivial conjugate-symmetric solution, which is a solution with $\beta=\overline{\alpha}$. We can assume that $|\alpha|=1$. We set

$$\Phi=\begin{bmatrix}\alpha\mathrm{i} & \alpha\\ -\overline{\alpha}\mathrm{i} & \overline{\alpha}\end{bmatrix}$$

and

$$\left[\begin{array}{cc} \mathbf{u}_+(t) & \mathbf{u}_-(t) \end{array}\right] = \left[\begin{array}{cc} \mathbf{u}(t) & \mathbf{u}^{\#}(t)) \end{array}\right] \Phi\,.$$

Then $\mathbf{u}_\pm$ are conjugate-symmetric and according to Lemma 2.9 we obtain

$$B = \frac{1}{2\mathrm{i}}\,\mathrm{Bez}_T(\mathbf{u}_+, \mathbf{u}_-)\,.$$

We can now apply the reduction step described in Section 5.8 for $\widetilde{\mathbf{u}} = \mathbf{u}_+$ and $\widetilde{\mathbf{v}} = \mathbf{u}_-$. Due to the Hermitian symmetry we have $\nu_- = \nu_+ =: \nu$. The vector $\mathbf{q}_-^{\#}$ is just the vector $\mathbf{q}_+$ after cancelling its first component. The vectors $\mathbf{w}$ and $\mathbf{r}$, both considered as elements of $\mathbb{F}^{m+1}$, are conjugate-symmetric. This leads to

$$2\mathrm{i}\,B = M_{2\nu}(\mathbf{w}) \left[\begin{array}{cc} O & B_-(\mathbf{q}_-) \\ (B_-(\mathbf{q}_-))^* & O \end{array}\right] M_{2\nu}(\mathbf{w})^* + \left[\begin{array}{ccc} O & O & O \\ O & B_1 & O \\ O & O & O \end{array}\right],$$

where $B_1 = \mathrm{Bez}_T(\mathbf{r}, \mathbf{w})$ is $m \times m$. Taking into account that $B_-(\mathbf{q})$ and the zero matrices in the corners of the last term are $\nu \times \nu$ matrices, the sum of the ranks of the two terms on the right-hand side is equal to the rank of B. This is also true for the signature. The signature of the first term is equal to zero. Hence

$$\operatorname{sgn} B = \operatorname{sgn} B_1\,.$$

If $\mathrm{Bez}_T(\mathbf{r}, \mathbf{w})$ is singular, then we carry out another non-generic step. If $\mathrm{Bez}_T(\mathbf{r}, \mathbf{w})$ is nonsingular we go over to the first Bezoutian representation by introducing $\mathbf{v} = \frac{1}{2}(\mathbf{r} - \mathrm{i}\mathbf{w})$ and obtain $B_1 = \mathrm{Bez}_T(\mathbf{v}, \mathbf{v}^{\#})$. Now we can apply a generic step. Summing up, we have described a procedure that computes the signature of an arbitrary Hermitian T-Bezoutian in $O(n^2)$ operations.

6. Bezoutians and companion matrices

In this section we show that Bezoutians are related to functions of companion matrices.

1. Factorization of the companion. The companion matrix of the monic polynomial $\mathbf{u}(t) = \sum_{k=1}^{n+1} u_k t^{k-1} \in \mathbb{F}^{n+1}$ is, by definition, the $n \times n$ matrix

$$C(\mathbf{u}) = \left[\begin{array}{cccc} 0 & 1 & & \\ & & \ddots & \\ & & & 1 \\ -u_1 & -u_2 & \dots & -u_n \end{array}\right]$$

It is easy to show that the characteristic polynomial of $C(\mathbf{u})$, $\det(tI_n - C(\mathbf{u}))$, is equal to $\mathbf{u}(t)$. This is also a consequence of the following useful relation.

Lemma 6.1. *We have*

$$tI_n - C(\mathbf{u}) = \begin{bmatrix} 0 & -1 & & \\ & & \ddots & \\ & & & -1 \\ 1 & \mathbf{u}_1(t) & \dots & \mathbf{u}_{n-1}(t) \end{bmatrix} \begin{bmatrix} \mathbf{u}(t) & \\ & I_{n-1} \end{bmatrix} \begin{bmatrix} 1 & & & \\ -t & 1 & & \\ & \ddots & \ddots & \\ & & -t & 1 \end{bmatrix}, \tag{6.1}$$

where $\mathbf{u}_k(t) = u_{k+1} + u_{k+2}t + \cdots + u_{n+1}t^{n-k}$.

Proof. It is immediately checked that

$$(tI_n - C(\mathbf{u})) \begin{bmatrix} 1 & & & & \\ t & 1 & & & \\ t^2 & t & 1 & & \\ \vdots & \vdots & \ddots & \ddots & \\ t^{n-1} & t^{n-2} & \dots & t & 1 \end{bmatrix} = \begin{bmatrix} 0 & -1 & & & \\ & & \ddots & & \\ & & & & -1 \\ \mathbf{u}(t) & \mathbf{u}_1(t) & \dots & \mathbf{u}_{n-1}(t) \end{bmatrix}.$$

This equality can be rearranged to (6.1). □

The polynomials $\mathbf{u}_k(t)$ appearing in the first factor of the right side of (6.1) are the *Horner polynomials* of $\mathbf{u}^J(t)$. They satisfy the recursion

$$\mathbf{u}_k(t) = t\mathbf{u}_{k+1}(t) + u_{k+1}$$

and can be represented as

$$[\,\mathbf{u}_1(t)\ \dots\ \mathbf{u}_n(t)\,] = \ell_n(t)^T B(\mathbf{u})\,.$$

where $B(\mathbf{u})$ is introduced in (2.2). For $t_0 \in \mathbb{F}$, the matrix $t_0 I_n - C(\mathbf{u})$ is a special case of a resultant matrix (cf. (2.4)). In fact,

$$t_0 I_n - C(\mathbf{u}) = \mathrm{Res}(t_0 - t, \mathbf{u}(t) + t^{n-1}(t_0 - t))\,.$$

Since the resultant matrix is nonsingular if and only if the polynomials are coprime, we conclude again that $t_0 I_n - C(\mathbf{u})$ is singular if and only if $t - t_0$ is a divisor of $\mathbf{u}(t)$, i.e., $\mathbf{u}(t_0) = 0$.

2. Functional calculus. Before we continue with companions and Bezoutians we recall some general definitions and facts concerning functions of a matrix. Let A be an $n \times n$ matrix and $\mathbf{u}(t) = \sum_{k=1}^{m} u_k t^{k-1}$ a polynomial. Then $\mathbf{u}(A)$ denotes the matrix

$$\mathbf{u}(A) = \sum_{k=1}^{m} u_k A^{k-1}$$

in which we set $A^0 = I_n$. The matrices of the form $\mathbf{u}(A)$ form a commutative matrix algebra and the transformation $\mathbf{u}(t) \mapsto \mathbf{u}(A)$ is a linear operator and a ring homomorphism. In particular, if $\mathbf{u}(t) = \mathbf{u}_1(t)\mathbf{u}_2(t)$, then $\mathbf{u}(A) = \mathbf{u}_1(A)\mathbf{u}_2(A)$. If $\mathbf{u}(t)$ is the characteristic polynomial of A, then, according to the Cayley-Hamilton

theorem, $\mathbf{u}(A) = O$. Let a polynomial $\mathbf{v}(t)$ and the characteristic polynomial $\mathbf{u}(t)$ of A be coprime. Then the Bezout equation

$$\mathbf{v}(t)\mathbf{x}(t) + \mathbf{u}(t)\mathbf{y}(t) = 1 \tag{6.2}$$

has a solution $(\mathbf{x}(t), \mathbf{y}(t))$. Replacing t by A we obtain that $\mathbf{v}(A)\mathbf{x}(A) = I_n$. That means $\mathbf{v}(A)$ is nonsingular and $\mathbf{x}(A)$ is its inverse.

3. Barnett's formula. The following remarkable formula is due to *S. Barnett.*

Theorem 6.2. *Let* $\mathbf{u}(t), \mathbf{v}(t) \in \mathbb{F}^{n+1}(t)$ *and* $\mathbf{u}(t)$ *be monic. Then*

$$\mathrm{Bez}_H(\mathbf{u}, \mathbf{v}) = B(\mathbf{u})\mathbf{v}(C(\mathbf{u}))\,, \tag{6.3}$$

where $B(\mathbf{u})$ *is introduced in* (2.2).

Proof. Due to linearity, it is sufficient to prove the formula for $\mathbf{v}(t) = \mathbf{e}_k(t) = t^{k-1}$. We set $B_k = \mathrm{Bez}_H(\mathbf{u}, \mathbf{e}_k)$. Since $B_k = B(\mathbf{u})C(\mathbf{u})^{k-1}$ is true for $k = 1$ we still have to show that $B_{k+1} = B_k C(\mathbf{u})$. Taking into account that

$$C(\mathbf{u})\ell_n(s) = s\ell_n(s) - \mathbf{u}(s)\mathbf{e}_n\,,$$

we obtain

$$\ell_n(t)^T B_k C(\mathbf{u})\ell_n(s) = sB_k(t,s) - \mathbf{u}(s)\ell_n(t)^T B_k \mathbf{e}_n = sB_k(t,s) - t^{k-1}\mathbf{u}(s)\,. \tag{6.4}$$

On the other hand,

$$B_{k+1}(t,s) = \frac{(\mathbf{u}(t)s^{k-1} - t^{k-1}\mathbf{u}(s))s}{t-s} - t^{k-1}\mathbf{u}(s) = sB_k(t,s) - t^{k-1}\mathbf{u}(s)\,. \tag{6.5}$$

Comparing (6.4) and (6.5) we obtain the recursion $B_{k+1} = B_k C(\mathbf{u})$. □

From this theorem we can conclude again (cf. Corollary 3.4) that the H-Bezoutian of $\mathbf{u}(t)$ and $\mathbf{v}(t)$ is nonsingular if $\mathbf{u}(t)$ and $\mathbf{v}(t)$ ($\mathbf{u}(t) \in \mathbb{F}^{n+1}(t)$ monic) are coprime and that its inverse is given by

$$\mathrm{Bez}_H(\mathbf{u}, \mathbf{v})^{-1} = \mathbf{x}(C(\mathbf{u}))B(\mathbf{u})^{-1}\,, \tag{6.6}$$

where $\mathbf{x}(t)$ is from the solution of the Bezout equation (6.2). In the next subsection we will show that $\mathbf{x}(C(\mathbf{u}))$ is actually a Hankel matrix.

4. Barnett's formula for T-Bezoutians. Now we consider T-Bezoutians. Let $\mathbf{u}(t)$ be a comonic polynomial of degree $\le n$ and $B_k = \mathrm{Bez}_T(\mathbf{u}, \mathbf{e}_k)$. Then

$$B_k(t,s) = B_{k+1}(t,s)s - t^{k-1}\mathbf{u}^J(s)\,.$$

This can be written as

$$B_k = B_{k+1}C(\mathbf{u}^J)\,.$$

With the notation of (2.14) we obtain the *Toeplitz analogue of Barnett's formula.*

Theorem 6.3. *Let* $\mathbf{u}(t), \mathbf{v}(t) \in \mathbb{F}^{n+1}(t)$, *where* $\mathbf{u}(t)$ *is comonic. Then*

$$\mathrm{Bez}_T(\mathbf{u}, \mathbf{v}) = B_-(\mathbf{u})\mathbf{v}^J(C(\mathbf{u}^J))\,.$$

In particular, for $\mathbf{v}(t) = 1$ we obtain the equality

$$B_+(\mathbf{u}) = B_-(\mathbf{u})C(\mathbf{u}^J)^n\,, \tag{6.7}$$

which yields an LU-factorization of $C(\mathbf{u}^J)^n$ and will be applied below to prove an inversion formula for T-Bezoutions.

7. Hankel matrices generated by rational functions

In this section we consider Hankel matrices generated by rational functions and show that they are closely related to H-Bezoutians. We understand "rational functions" in an abstract sense as elements of the quotient field of the ring of polynomials. But occasionally, in particular if we restrict ourselves to the case $\mathbb{F} = \mathbb{C}$, we interpret them in the analytic sense as functions defined in $\mathbb{F}$.

By a *proper* rational function we mean a rational function for which the degree of the numerator polynomial is not greater than the degree of the denominator polynomial. We say that the representation $\mathbf{f}(t) = \frac{\mathbf{p}(t)}{\mathbf{u}(t)}$ is in *reduced form* if $\mathbf{u}(t)$ and $\mathbf{p}(t)$ are coprime and $\mathbf{u}(t)$ is monic. This representation is unique. The *degree of a proper rational function* is the degree of the denominator polynomial in the reduced representation.

1. Generating functions of Hankel matrices. A proper rational function $\mathbf{f}(t)$ can be represented as

$$\mathbf{f}(t) = h_0 + h_1 t^{-1} + h_2 t^{-2} + \cdots\,. \tag{7.1}$$

If $\mathbb{F} = \mathbb{C}$, then (7.1) can be interpreted as the Laurent series expansion of $\mathbf{f}(t)$ at infinity converging outside a disk with center 0. For a general field $\mathbb{F}$ (7.1) has a meaning as quotient of two formal power series. The coefficients h_i can be obtained recursively by an obvious formula. For $\mathbf{f}(t)$ having a Laurent expansion (7.1), we set $\mathbf{f}(\infty) = h_0$ and write $\mathbf{f}(t) = O(t^{-m})$ if $h_0 = \cdots = h_{m-1} = 0$.

Note that if $\mathbf{f}(t)$ is given by (7.1), then we have

$$\frac{\mathbf{f}(t) - \mathbf{f}(s)}{t - s} = \sum_{k=1}^{\infty} h_k\, \frac{t^{-k} - s^{-k}}{t - s} = -\sum_{i,j=1}^{\infty} h_{i+j-1}\, t^{-i} s^{-j}\,. \tag{7.2}$$

This relation suggests the following definition. For $n = 1, 2, \ldots$, the $n \times n$ *Hankel matrix generated by* $\mathbf{f}(t)$ is, by definition, the matrix

$$H_n(\mathbf{f}) = [\, h_{i+j-1} \,]_{i,j=1}^{n}.$$

Let us point out that the entry h_0 does not enter the definition of $H_n(\mathbf{f})$. If for some $n \times n$ Hankel matrix H_n there is a function $\mathbf{f}$ so that $H_n = H_n(\mathbf{f})$, then $\mathbf{f}(t)$ will be called *generating function* of H_n.

Example 7.1. *As an example, let us compute the Hankel matrices generated by a partial fraction* $\dfrac{1}{(t-c)^m}$ $(c \in \mathbb{F},\ m = 1, 2, \dots, 2n-1)$. *We denote*

$$L_n(c,m) = \frac{1}{(m-1)!}\, H_n\left(\frac{1}{(t-c)^m}\right)$$

and write $L_n(c)$ *instead of* $L_n(c,1)$. *In view of*

$$\frac{1}{t-c} = t^{-1} + c\,t^{-2} + c^2\,t^{-3} + \cdots \tag{7.3}$$

we have

$$L_n(c) = [\, c^{i+j-2} \,]_{i,j=1}^{n}\,. \tag{7.4}$$

Differentiating the equality (7.3) *we obtain*

$$L_n(c,2) = [\, (i+j-2)c^{i+j-3} \,]_{i,j=1}^{n}$$

and in general, for $m = 1, \dots, 2n-1$,

$$L_n(c,m) = \left[\begin{pmatrix} i+j-2 \\ m-1 \end{pmatrix} c^{i+j-1-m} \right]_{i,j=1}^{n}\,.$$

It is obvious that the rank of $L_n(c,m)$ *is equal to* m.

The matrices $L_n(c,m)$ are called *elementary Hankel matrices*.

Example 7.2. *For our second example we assume that* $\mathbb{F}$ *is algebraically closed. Let* $\mathbf{u}(t)$ *be a polynomial of degree* n *and let* $t_1, \dots, t_n$ *be the zeros of* $\mathbf{u}(t)$. *The* Newton sums s_i $(i = 1, 2, \dots)$ *are given by*

$$s_i = \sum_{k=1}^{n} t_k^{i-1}\,.$$

We form the Hankel matrix $H_n = [\, s_{i+j-1} \,]_{i,j=1}^{n}$. *Then we have*

$$H_n = H_n\left(\frac{\mathbf{u}'(t)}{\mathbf{u}(t)}\right).$$

This follows from the obvious relation

$$\frac{\mathbf{u}'(t)}{\mathbf{u}(t)} = \sum_{k=1}^{n} \frac{1}{t-t_k}\,.$$

The transformation

$$\mathcal{H} \,:\, \mathbf{f}(t) \longrightarrow H_n(\mathbf{f})$$

is clearly a linear operator from the vector space of all proper rational functions to the space of $n \times n$ Hankel matrices. The kernel of this operator consists of all proper rational function $\mathbf{f}(t)$ for which $\mathbf{f}(t) - \mathbf{f}(\infty) = O(t^{-2n})$. We show that this transformation is onto. That means any $k \times k$ Hankel matrix can be regarded as generated by a proper rational function.

Proposition 7.3. *Let $\mathbf{u}(t)$ be a fixed monic polynomial of degree $2n-1$. Then any $n \times n$ Hankel matrix H_n can be represented uniquely in the form*

$$H_n = H_n\left(\frac{\mathbf{p}}{\mathbf{u}}\right) \tag{7.5}$$

for some $\mathbf{p} \in \mathbb{F}^{2n-1}$.

Proof. Clearly, a matrix of the form (7.5) with $\mathbf{p} \in \mathbb{F}^{2n-1}$ does not belong to the kernel of the transformation $\mathcal{H}$. That means that the mapping of the vector $\mathbf{p} \in \mathbb{F}^{2n-1}$ to the Hankel matrix $H_n\left(\frac{\mathbf{p}}{\mathbf{u}}\right)$ is one-to-one. Hence the dimension of its range equals $2n-1$. This is just the dimension of the space of $n \times n$ Hankel matrices. Thus the mapping is onto. □

Since in an algebraically closed field $\mathbb{F}$ any proper rational function has a partial fraction decomposition we conclude that in this case any Hankel matrix can be represented as a linear combination of elementary Hankel matrices. The reader may observe that the problem to find the generating function of a Hankel matrix is closely related to the Padé approximation problem at infinity and the partial realization problem.

In connection with these and other problems the question about a generating function of minimal degree arises. We will see later that the degree of the generating function is at least equal to the rank of H_n. But it can be bigger. For example, the rank-one Hankel matrix $H_n = \mathbf{e}_n \mathbf{e}_n^T$ has no generating function of degree less than $2n-1$. Here we restrict ourselves to the nonsingular case. For a nonsingular $n \times n$ Hankel matrix, a generating function of degree n always exists, as the next proposition shows. Let us use the notation of Section 4.1.

Theorem 7.4. *Let $H_n = [\, s_{i+j-1} \,]_{i,j=1}^n$ be a nonsingular Hankel matrix, $\{\mathbf{u}(t), \mathbf{v}(t)\}$ be a fundamental system of H_n, where $\mathbf{u}(t)$ is monic and $\deg \mathbf{v}(t) < n$. Then, for any $\alpha \in \mathbb{F}$, there is a vector $\mathbf{p} \in \mathbb{F}^n$ such that*

$$H_n = H_n\left(\frac{\mathbf{p}(t)}{\mathbf{u}(t) - \alpha \mathbf{v}(t)}\right) .$$

Proof. We consider the $(n-1) \times (n+1)$ matrix ∂H_n, which was introduced in (4.1). The vector $\mathbf{w} = \mathbf{u} - \alpha \mathbf{v}$ is a monic vector belonging to the nullspace of ∂H_n. We set

$$\mathbf{p} = \begin{bmatrix} 0 & s_1 & \dots & s_n \\ & \ddots & \ddots & \vdots \\ & & 0 & s_1 \end{bmatrix} \mathbf{w} .$$

From this definition and from $\mathbf{w} \in \ker \partial H_n$ we can see that in the expansion

$$\frac{\mathbf{p}(t)}{\mathbf{w}(t)} = h_1 t^{-1} + h_2 t^{-2} + \cdots$$

we have $h_i = s_i$, $i = 1, \dots, 2n-1$. Hence, $H_n = H_n\left(\frac{\mathbf{p}(t)}{\mathbf{w}(t)}\right)$. □

Example 7.5. *Let us find generating functions of degree n of the matrix $H_n = J_n$. For this matrix $\{\mathbf{e}_{n+1}, \mathbf{e}_1\}$ is a fundamental system. Furthermore, $\mathbf{p} = \mathbf{e}_1$. Thus, for any $\alpha \in \mathbb{F}$,*

$$J_n = H_n \left(\frac{1}{t^n - \alpha} \right) .$$

Let us present a special case of Theorem 7.4 involving the solutions of the equations

$$H_k \mathbf{u}_k = \rho_k \mathbf{e}_k , \quad \mathbf{e}_k^T \mathbf{u}_k = 1 \tag{7.6}$$

for $k = n, n+1$. Here H_{n+1} is a nonsingular extension of H_n . As we already know from Section 4.2 these monic solutions form a fundamental system for H_n. Thus, the following is immediately clear.

Corollary 7.6. *Let H_n and H_{n+1} be as above and $\mathbf{u}_n, \mathbf{u}_{n+1}$ be the solutions of* (7.6) *for $k = n, n+1$. Then, for an $\alpha \in \mathbb{F}$, there are vectors $\mathbf{p}_{n+1} \in \mathbb{F}^n$ and $\mathbf{p}_n \in \mathbb{F}^{n-1}$ such that*

$$H_n = H_n \left(\frac{\mathbf{p}_{n+1}(t) - \alpha \mathbf{p}_n(t)}{\mathbf{u}_{n+1}(t) - \alpha \mathbf{u}_n(t)} \right) .$$

2. Vandermonde factorization of Hankel matrices. In this subsection we assume that $\mathbb{F} = \mathbb{C}$. Let H_n be an $n \times n$ nonsingular Hankel matrix. Then, by Theorem 7.4, it has a generating function of degree n. We can assume that the denominator polynomial of this function has only simple roots, which follows from the following lemma. Recall from Corollary 4.3 that the polynomials forming a fundamental system of a nonsingular Hankel matrix are coprime.

Lemma 7.7. *Let $\mathbf{u}(t)$ and $\mathbf{v}(t)$ be coprime. Then for all $\alpha \in \mathbb{C}$, with the exception of a finite number of points, the polynomial $\mathbf{w}(t) = \mathbf{u}(t) - \alpha \mathbf{v}(t)$ has only simple roots.*

Proof. Suppose that τ_0 is a multiple root of $\mathbf{w}(t)$. Then $\mathbf{u}(\tau_0) = \alpha \mathbf{v}(\tau_0)$ and $\mathbf{u}'(\tau_0) = \alpha \mathbf{v}'(\tau_0)$. Since $\mathbf{u}(t)$ and $\mathbf{v}(t)$ are coprime, $\mathbf{v}(\tau_0) \neq 0$. Hence τ_0 is a root of the nonzero polynomial $\mathbf{z}(t) = \mathbf{u}(t)\mathbf{v}'(t) - \mathbf{v}(t)\mathbf{u}'(t)$. Now we choose α such that $\mathbf{u}(\tau) \neq \alpha \mathbf{v}(\tau)$ for all roots τ of $\mathbf{z}(t)$. Then none of the τ is a root of $\mathbf{w}(t)$, so $\mathbf{w}(t)$ has no multiple roots. □

Let $\mathbf{f}(t) = \frac{\mathbf{p}(t)}{\mathbf{w}(t)}$ with $(\infty) = \mathbf{0}$ be a proper rational function in reduced form such that $\mathbf{w}(t)$ has simple roots $t_1, \ldots, t_n$. Then $\mathbf{f}(t)$ has a partial fraction decomposition

$$\mathbf{f}(t) = \sum_{i=1}^{n} \frac{\delta_i}{t - t_i} ,$$

where

$$\delta_i = \frac{\mathbf{p}(t_i)}{\mathbf{w}'(t_i)} = \left(\left(\frac{1}{\mathbf{f}} \right)' (t_i) \right)^{-1} . \tag{7.7}$$

Hence $H_k(\mathbf{f})$ can be represented as a linear combination of elementary Hankel matrices $L_k(t_i)$ defined by (7.4)

$$H_k(\mathbf{f}) = \sum_{i=1}^{n} \delta_i L_k(t_i)\,. \tag{7.8}$$

This relation can be stated as a matrix factorization. In fact, observe that $L_k(t_i)$ equals $\ell_k(t_i)\ell_k(t_i)^T$, where $\ell_k(t_i) = (\,t_i^{j-1})_{j=1}^{k}$. We form the $n \times k$ Vandermonde matrix $V_k(\mathbf{t})$, $\mathbf{t} = (t_i)_{i=1}^{n}$, with the rows $\ell_k(t_i)^T$ $(i = 1, \ldots, n,\ )$

$$V_k(\mathbf{t}) = [\,t_i^{j-1}\,]_{i=1,j=1}^{n\quad k}\,.$$

Now (7.8) is equivalent to the following.

Proposition 7.8. *Let $\mathbf{f}(t)$ and $\mathbf{t}$ be as above. Then for $k \geq n$, the Hankel matrix $H_k = H_k(\mathbf{f})$ admits a representation*

$$H_k = V_k(\mathbf{t})^T D V_k(\mathbf{t})\,, \tag{7.9}$$

where $D = \operatorname{diag}(\delta_i)_{i=1}^{n}$ and the δ_i are given by (7.7).

Example 7.9. *For the Hankel matrix of Example* 7.2 *we obtain the following factorization,*

$$H_n\left(\frac{\mathbf{u}'}{\mathbf{u}}\right) = V_n(\widetilde{\mathbf{t}})^T \operatorname{diag}(\nu_i)_{i=1}^{r} V_n(\widetilde{\mathbf{t}})\,, \tag{7.10}$$

where $\widetilde{\mathbf{t}}$ is the vector of different zeros $\widetilde{t}_i$ $(i = 1, \ldots, r)$ of $\mathbf{u}(t)$, and ν_i are their multiplicities.

A consequence of this Proposition 7.8 is that $\operatorname{rank} H_k(\mathbf{f}) = n$ for all $k \geq n$. In Section 7.5 we will show that this is true for a general field $\mathbb{F}$. Combining Proposition 7.8 with Proposition 7.4 we obtain the following.

Corollary 7.10. *Let H_n be a nonsingular $n \times n$ Hankel matrix, $\{\mathbf{u}, \mathbf{v}\}$ a fundamental system of H_n, and $\alpha \in \mathbb{C}$ such that $\mathbf{w}(t) = \mathbf{u}(t) - \alpha\mathbf{v}(t)$ has simple roots $t_1, \ldots, t_n$. Then H_n admits a representation*

$$H_n = V_n(\mathbf{t})^T D V_n(\mathbf{t})$$

with a diagonal matrix D.

Let us find the Vandermonde factorization for Example 7.5, i.e., for $H_n = J_n$. The polynomial $t^n - \alpha$ has simple roots for all $\alpha \neq 0$, and these roots t_i are the nth complex roots of α. The diagonal matrix is given by

$$D = \frac{1}{n} \operatorname{diag}(t_1^{1-n}, \ldots, t_n^{1-n})\,.$$

3. Real Hankel matrices. We consider the special case of a real, nonsingular Hankel matrix H_n. In this case the fundamental system of H_n is also real. We choose $\alpha \in \mathbb{R}$, since then the non-real roots of $\mathbf{w}(t) = \mathbf{u}(t) - \alpha\mathbf{v}(t)$ appear in conjugate complex

pairs. Let $t_1, \dots, t_r$ be the real roots and $t_{r+1}, t_{r+2} = \overline{t}_{r+1}, \dots, t_{n-1}, t_n = \overline{t}_{n-1}$ be the non-real roots of $\mathbf{w}(t)$. Then

$$V_n(\mathbf{t})^T = V_n(\mathbf{t})^* \operatorname{diag}(I_r, \underbrace{J_2, \dots, J_2}_{l}),$$

where $r + 2l = n$. Thus, we obtain from Proposition 7.8 the following.

Corollary 7.11. *If the Hankel matrix H_k $(k \geq n)$ in Proposition* 7.8 *is real, then it admits a representation*

$$H_k = V_k(\mathbf{t})^* D_1 V_k(\mathbf{t}),$$

where

$$D_1 = \operatorname{diag}\left(\delta_1, \dots, \delta_r, \begin{bmatrix} 0 & \overline{\delta}_{r+1} \\ \delta_{r+1} & 0 \end{bmatrix}, \dots, \begin{bmatrix} 0 & \overline{\delta}_{n-1} \\ \delta_{n-1} & 0 \end{bmatrix}\right).$$

In particular, the matrices H_n and D_1 are congruent.

Combining this with Sylvester's inertia law, with (7.9), and with the fact that the signature of a matrix $\begin{bmatrix} 0 & \overline{\delta} \\ \delta & 0 \end{bmatrix}$ equals 0, we conclude the following.

Corollary 7.12. *Let $\mathbf{f}(t)$ be a real rational function of degree n with simple poles $t_1, \dots, t_n$, where $t_1, \dots, t_r$ are real and the other poles non-real. Then for $k \geq n$ the signature of the Hankel matrix $H_k = H_k(\mathbf{f})$ is given by*

$$\operatorname{sgn} H_k = \sum_{i=1}^{r} \operatorname{sgn} \delta_i ,$$

where δ_i is defined by (7.7). *In particular, H_n is positive definite if and only if all roots of $\mathbf{w}(t)$ are real and $\delta_i > 0$ for all i.*

Let us specify the criterion of positive definiteness further.

4. The Cauchy index. Let $\mathcal{C}$ be an oriented closed curve in the extended complex plane $\mathbb{C} \cup \{\infty\}$ and $\mathbf{f}(t)$ a rational function with real values on $\mathcal{C}$ with the exception of poles. A pole c of $\mathbf{f}(t)$ on $\mathcal{C}$ is said to be of *positive type* if

$$\lim_{\substack{t \to c- \\ t \in \mathcal{C}}} \mathbf{f}(t) = -\infty \quad \text{and} \quad \lim_{\substack{t \to c+ \\ t \in \mathcal{C}}} \mathbf{f}(t) = \infty .$$

It is said to be of *negative type* if c is of positive type for $-\mathbf{f}(t)$. If a pole is not of positive or negative type, then it is called *neutral*. The *Cauchy index of* $\mathbf{f}(t)$ *along* $\mathcal{C}$ is, by definition, the integer

$$\operatorname{ind}_{\mathcal{C}} \mathbf{f}(t) = p_+ - p_-$$

where p_+ is the number of poles of positive and p_- the number of poles of negative type. The pole c is of positive (negative) type if and only if the function $\frac{1}{\mathbf{f}(t)}$ is increasing (decreasing) in a neighborhood of c.

It is clear that if c is a pole of positive or negative type on $\mathcal{C}$, then a small perturbation of the coefficients of $\mathbf{f}(t)$ leads only to a small change of the pole on

$\mathcal{C}$ by preserving its type (which is not true for neutral poles). Now we are in the position to relate the signature of Hankel matrices generated by a rational function with the Cauchy index of this function along $\mathbb{R}$.

Proposition 7.13. *Let* $\mathbf{f}(t)$ *be a real proper rational function of degree* n. *Then*

$$\operatorname{sgn} H_n(\mathbf{f}) = \operatorname{ind}_{\mathbb{R}} \mathbf{f}(t)\,.$$

Proof. Suppose that $\mathbf{f}(t) = \frac{\mathbf{p}(t)}{\mathbf{u}(t)}$ is the reduced representation as quotient of polynomials. Let us first assume that $\mathbf{u}(t)$ has simple roots $t_1, \dots, t_n$, i.e., $\mathbf{f}(t)$ has simple poles. Simple poles cannot be neutral. If t_i is a simple pole of positive type, then $\left(\frac{1}{\mathbf{f}}\right)'(t_i) > 0$. Comparing this with (7.7) we conclude that $\delta_i > 0$. Analogously, we have $\delta_i < 0$ for a pole t_i of negative type. Now it remains to apply Corollary 7.12.

Now let $\mathbf{u}(t)$ have multiple roots. Neutral poles of $\mathbf{f}(t)$ correspond to roots of $\mathbf{u}(t)$ of even order and do not contribute to the Cauchy index of $\mathbf{f}(t)$. It is easy to see that we can disturb $\mathbf{u}(t)$ additively with an $\alpha \in \mathbb{R}$ as small as we want such that the disturbed roots of even order disappear or become simple roots of $\mathbf{u}(t) + \alpha$, so that the respective pairs of poles do not contribute to the Cauchy index of $\mathbf{f}_\alpha(t) = \frac{\mathbf{p}(t)}{\mathbf{u}(t)+\alpha}$. The other poles remain simple and of the same type. So the first part of the proof applies to $\mathbf{f}_\alpha(t)$. Due to Proposition 7.17 below $H_n(\mathbf{f})$ is nonsingular. Taking into account that the signature of a nonsingular Hankel matrix is invariant with respect to small perturbations the assertion follows. □

Some readers might be unsatisfied with the analytic argument in the proof of the algebraic Proposition 7.13. For those readers we note that a purely algebraic proof of this proposition is possible if more general Vandermonde representations of Hankel matrices are considered.

Let us discuss the question how positive definiteness of $H_n(\mathbf{f})$ can be characterized in terms of $\mathbf{f}(t)$. According to Proposition 7.13, $H_n(\mathbf{f})$ is positive definite if and only if the Cauchy index of $\mathbf{f}(t)$ along $\mathbb{R}$ is equal to n. That means that $\mathbf{f}(t)$ must have n poles of positive type. Between two poles of positive type there must be a zero of $\mathbf{f}(t)$, i.e., a root of the numerator polynomial. In other words, the poles and zeros of $\mathbf{f}(t)$ must interlace.

We say that the *real roots of two polynomials* $\mathbf{u}(t)$ *and* $\mathbf{p}(t)$ *interlace* if between two roots of $\mathbf{u}(t)$ there is exactly one root of $\mathbf{p}(t)$. Polynomials with roots that interlace are coprime. Let us summarize our discussion.

Corollary 7.14. *Let* $\mathbf{f}(t)$ *be a real rational function of degree* n, *and let* $\mathbf{f}(t) = \frac{\mathbf{p}(t)}{\mathbf{u}(t)}$ *be its reduced representation. Then* $H_n(\mathbf{f})$ *is positive definite if and only if the polynomials* $\mathbf{u}(t)$ *and* $\mathbf{p}(t)$ *have only real simple roots that interlace.*

5. Congruence to H-Bezoutians. In Section 4 we showed that inverses of Hankel matrices are H-Bezoutians. Now we are going to explain another relation between Hankel matrices and H-Bezoutians. In the real case this relation just means that Hankel matrices and H-Bezoutians are congruent.

Throughout this subsection, let $\mathbf{u}(t)$ be a polynomial of degree n, $\mathbf{v}(t) \in \mathbb{F}^{n+1}(t)$, and $\mathbf{f}(t) = \frac{\mathbf{v}(t)}{\mathbf{u}(t)}$. Then $B = \mathrm{Bez}_H(\mathbf{u}, \mathbf{v})$ is an $n \times n$ matrix. For $k > n$, we identify this matrix with the $k \times k$ matrix obtained from B by adding $k - n$ zero rows and zero columns at the bottom and on the right. The same we do for $B(\mathbf{u}) = \mathrm{Bez}_H(\mathbf{u}, \mathbf{e}_1)$ introduced in (2.2).

Proposition 7.15. *For $k \geq n$, the $k \times k$ Hankel matrix generated by $\mathbf{f}(t) = \frac{\mathbf{v}(t)}{\mathbf{u}(t)}$ is related to the H-Bezoutian of $\mathbf{u}(t)$ and $\mathbf{v}(t)$ via*

$$\mathrm{Bez}_H(\mathbf{u}, \mathbf{v}) = B(\mathbf{u})\, H_k\,(\mathbf{f})\, B(\mathbf{u})\,. \tag{7.11}$$

Proof. For $B = \mathrm{Bez}_H(\mathbf{u}, \mathbf{v})$ and $\mathbf{u} = (u_i)_{i=1}^{n+1}$ we have in view of (7.2)

$$\begin{aligned} B(t,s) &= -\mathbf{u}(t)\,\frac{\mathbf{f}(t) - \mathbf{f}(s)}{t-s}\,\mathbf{u}(s) \\ &= \sum_{i,j=1}^{n+1} \sum_{p,q=1}^{\infty} u_i\, h_{p+q-1}\, u_j\, t^{i-p-1} s^{j-q-1} \\ &= \sum_{m,l=1}^{n} \sum_{p,q=1}^{\infty} u_{m+p}\, h_{p+q-1}\, u_{q+l}\, t^{m-1} s^{l-1}\,, \end{aligned}$$

where we set $u_j = 0$ for $j > n+1$. The coefficient matrix of this polynomial can be written as a product of three matrices, as it is claimed. □

Recall from Corollary 3.4 that the nullity of $\mathrm{Bez}_H(\mathbf{u}, \mathbf{v})$ is equal to the degree of the greatest common divisor of $\mathbf{u}(t)$ and $\mathbf{v}(t)$. This implies that the rank of $\mathrm{Bez}_H(\mathbf{u}, \mathbf{v})$ is equal to the degree of the rational function $\mathbf{f}(t)$.

Corollary 7.16. *In the real case the matrices $\mathrm{Bez}_H(\mathbf{u}, \mathbf{v})$ and $H_n\left(\frac{\mathbf{u}}{\mathbf{v}}\right)$ are congruent.*

Proposition 7.17. *Let $\mathbf{f}(t)$ be a rational function of degree n. Then, for $k \geq n$, the rank of $H_k(\mathbf{f})$ is equal to n. In particular, $H_n(\mathbf{f})$ is nonsingular.*

Proof. Let $\mathbf{f}(t) = \frac{\mathbf{v}(t)}{\mathbf{u}(t)}$ be in reduced form and $\deg \mathbf{u}(t) = n$. Since the matrix $B(\mathbf{u})$ in (7.11) as an $k \times k$ matrix is singular for $k > n$, the assertion cannot be concluded directly from Proposition 7.15. Thus, for $k > n$, define $\mathbf{u}_k$ and $\mathbf{v}_k$ by $\mathbf{u}_k(t) = t^k \mathbf{u}(t)$ and $\mathbf{v}_k(t) = t^k \mathbf{v}(t)$, respectively. The $\mathbf{f}(t) = \frac{\mathbf{v}_k(t)}{\mathbf{u}_k(t)}$ and, due to Proposition 7.15,

$$\mathrm{Bez}_H(\mathbf{u}_k, \mathbf{v}_k) = B(\mathbf{u}_k)\, H_k(\mathbf{f})\, B(\mathbf{u}_k)\,, \tag{7.12}$$

where $B(\mathbf{u}_k)$ is nonsingular. Since the nullity of $\mathrm{Bez}_H(\mathbf{u}_k, \mathbf{v}_k)$ equals k, we have that $\operatorname{rank} \mathrm{Bez}_H(\mathbf{u}_k, \mathbf{v}_k) = n = \operatorname{rank} H_k(\mathbf{f})$. □

Note that under the assumptions of Proposition 7.8 the assertion of Proposition 7.17 follows already from the Vandermonde factorization of $H_k(\mathbf{f})$ given in (7.9).

In the real case, (7.12) is a congruence relation. Applying Sylvester's inertia law we can conclude the following for the case $\mathbb{F} = \mathbb{R}$.

Proposition 7.18. *If* $\mathbf{u}(t) \in \mathbb{R}^{n+1}(t)$ *with* $\deg \mathbf{u}(t) = n$ *and* $\mathbf{v}(t) \in \mathbb{R}^{n+1}(t)$, *then the matrices* $\mathrm{Bez}_H(\mathbf{u}, \mathbf{v})$ (*considered as* $k \times k$ *matrix*) *and* $H_k\left(\frac{\mathbf{v}}{\mathbf{u}}\right)$ *have the same inertia.*

Proof. Due to formula (7.12), which is also true if $\frac{\mathbf{v}(t)}{\mathbf{u}(t)}$ is not in reduced form, and since $\mathrm{Bez}_H(\mathbf{u}, \mathbf{v})$ and $\mathrm{Bez}_H(\mathbf{u}_k, \mathbf{v}_k)$ have the same rank it remains to show that these two Bezoutians have the same signature. But, this follows from

$$\mathrm{Bez}_H(\mathbf{u}_k, \mathbf{v}_k) = M_n(\mathbf{e}_{k+1})\mathrm{Bez}_H(\mathbf{u}, \mathbf{v})M_n(\mathbf{e}_{k+1})^T$$

and from Corollary 1.2. □

6. Inverses of H-Bezoutians. Comparing (7.11) with Barnett's formula (6.3) we obtain that the Hankel matrix $H_n\left(\frac{\mathbf{p}}{\mathbf{u}}\right)$ admits a representation

$$H_n\left(\frac{\mathbf{p}}{\mathbf{u}}\right) = \mathbf{p}(C(\mathbf{u}))B(\mathbf{u})^{-1}\,. \tag{7.13}$$

Together with (6.6) this immediately leads to the following.

Theorem 7.19. *Let* $\mathbf{u}(t)$ *and* $\mathbf{v}(t)$ *be coprime and* $(\mathbf{q}(t), \mathbf{p}(t))$, *be the* (*unique*) *solution of the Bezout equation*

$$\mathbf{u}(t)\mathbf{q}(t) + \mathbf{v}(t)\mathbf{p}(t) = 1 \tag{7.14}$$

with $\mathbf{q}(t), \mathbf{p}(t) \in \mathbb{F}^n(t)$. *Then*

$$\mathrm{Bez}_H(\mathbf{u}, \mathbf{v})^{-1} = H_n\left(\frac{\mathbf{p}}{\mathbf{u}}\right)\,.$$

An immediate consequence of this theorem is the converse of Theorem 4.2.

Corollary 7.20. *The inverse of a nonsingular H-Bezoutian is a Hankel matrix.*

Furthermore, Theorem 7.19 tells us that the inverse of a Bezoutian or the inverse of a Hankel matrix generated by a rational function can be computed with an algorithm which solves the Bezout equation (7.14). We show next that this equation can be solved with the help of the Euclidian algorithm.

7. Solving the Bezout equation. Let $\mathbf{u}_i(t)$ be the polynomials computed by the Euclidian algorithm, as described in Section 5.3, then with the help of the data of the Euclidian algorithm we can recursively solve the Bezout equations

$$\mathbf{u}(t)\mathbf{x}_i(t) + \mathbf{v}(t)\mathbf{y}_i(t) = \mathbf{u}_i(t) \quad (i = 0, 1, \dots), \tag{7.15}$$

where, for initialization, we have

$$\begin{array}{rclcrcl} \mathbf{x}_0(t) & = & 1 & & \mathbf{x}_1(t) & = & 0 \\ \mathbf{y}_0(t) & = & 0 & , & \mathbf{y}_1(t) & = & 1 \end{array}\,.$$

The recursion is given by

$$\begin{array}{rcl} \mathbf{x}_{i+1}(t) & = & \mathbf{x}_{i-1}(t) - \mathbf{q}_i(t)\mathbf{x}_i(t) \\ \mathbf{y}_{i+1}(t) & = & \mathbf{y}_{i-1}(t) - \mathbf{q}_i(t)\mathbf{y}_i(t)\,. \end{array} \tag{7.16}$$

In fact,

$$\begin{aligned}\mathbf{u}(t)\mathbf{x}_{i+1}(t)+\mathbf{v}(t)\mathbf{y}_{i+1}(t) &= \mathbf{u}(t)(\mathbf{x}_{i-1}-\mathbf{q}_i(t)\mathbf{x}_i(t))+\mathbf{v}(t)(\mathbf{y}_{i-1}(t)-\mathbf{q}_i(t)\mathbf{y}_i(t))\\ &= \mathbf{u}_{i-1}(t)-\mathbf{q}_i(t)\mathbf{u}_i(t)=\mathbf{u}_{i+1}(t)\ .\end{aligned}$$

Introducing the 2×2 matrix polynomials

$$X_i(t)=\begin{bmatrix}\mathbf{x}_{i-1}(t) & \mathbf{x}_i(t)\\ \mathbf{y}_{i-1}(t) & \mathbf{y}_i(t)\end{bmatrix} \quad\text{and}\quad \Phi_i(t)=\begin{bmatrix}0 & 1\\ 1 & -\mathbf{q}_i(t)\end{bmatrix}$$

we can write the recursion (7.16) as

$$X_{i+1}(t)=X_i(t)\Phi_i(t)\ , \qquad X_1(t)=I_2\ .$$

If $\mathbf{u}(t)$ and $\mathbf{v}(t)$ are coprime, then we have for some $i=l$ that $\mathbf{u}_l(t)=c=\text{const}$. Thus, the solution of (7.14) is obtained from $(\mathbf{x}_l(t),\mathbf{y}_l(t))$ by dividing by c.

As mentioned in Section 5.3 the Euclidian algorithm for finding the greatest common divisor of two polynomials is closely related to a Schur-type algorithm for Hankel matrices. The Euclidian algorithm for solving the Bezout equation is related to a mixed Levinson-Schur-type algorithm for Hankel matrices. It is also possible to design a pure Levinson-type algorithm for the solution of the Bezout equation.

8. Toeplitz matrices generated by rational functions

This section is the Toeplitz counterpart of the previous one. We define and study Toeplitz matrices generated by rational functions and show that they are closely related to T-Bezoutians. Some features are completely analogous to the Hankel case, but there are also some significant differences.

1. Generating functions of Toeplitz matrices. Let $\mathbf{f}(t)=\frac{\mathbf{p}(t)}{\mathbf{u}(t)}$ be a proper rational function with $\mathbf{u}(0)\neq 0$. Then $\mathbf{f}(t)$ admits series expansions in powers of t and as well as in powers of t^{-1},

$$\mathbf{f}(t)=a_0^+ + a_1 t + a_2 t^2 + \cdots\ , \quad \mathbf{f}(t)=-a_0^- - a_{-1}t^{-1} - a_{-2}t^{-2} - \cdots\ . \tag{8.1}$$

If $\mathbb{F}=\mathbb{C}$, then (8.1) can be understood as the Laurent series expansion at $t=0$ and at $t=\infty$, respectively. For a general $\mathbb{F}$, (8.1) makes sense as the quotient of two formal Laurent series. The coefficients can be obtained recursively by obvious relations. Note that, in a formal sense,

$$\frac{\mathbf{f}(t)-\mathbf{f}(s^{-1})}{1-ts}=\sum_{i,j=0}^{\infty} a_{i-j}t^i s^j\ , \tag{8.2}$$

where $a_0=a_0^+ + a_0^-$. The latter follows from the obvious relation

$$(1-ts)T(t,s)=a_0+\sum_{i=1}^{\infty}a_i t^i+\sum_{j=1}^{\infty}a_{-j}s^j\ ,$$

where $T=[\,a_{i-j}\,]_{i,j=1}^{\infty}$. This relation makes the following definition natural.

For $n = 1, 2, \dots$ the $n \times n$ *Toeplitz matrix generated by* $\mathbf{f}(t)$ with the expansions (8.1) is, by definition, the matrix $T_n(\mathbf{f}) = [\, a_{i-j} \,]_{i,j=1}^n$, where $a_0 = a_0^+ + a_0^-$. If $T_n = T_n(\mathbf{f})$, then the function $\mathbf{f}(t)$ is called *generating function* of T_n. Obviously, $T_n(\mathbf{f})$ is the zero matrix if $\mathbf{f}$ is a constant function. Note that finding the generating function of a given Toeplitz matrix is a two-point Padé approximation problem for the points 0 and ∞.

Example 8.1. *Since for* $c \neq 0$

$$\frac{1}{1-ct} = \sum_{k=0}^{\infty} c^k t^k \quad \text{and} \quad \frac{1}{1-ct} = -\sum_{k=1}^{\infty} c^{-k} t^{-k}\,,$$

we have

$$T_n\left(\frac{1}{1-ct}\right) = [\, c^{i-j} \,]_{i,j=1}^n = \ell_n(c)\ell_n\left(c^{-1}\right)^T\,, \tag{8.3}$$

where $\ell_n(c)$ *is defined in* (1.1).

Example 8.2. *Our second example is the Toeplitz analogue of Example* 7.2. *Let* $\mathbb{F}$ *be algebraically closed and* $\mathbf{u}(t)$ *a polynomial of degree* n *with the roots* $t_1, \dots, t_n$ *and* $\mathbf{u}(0) \neq 0$. *We define*

$$c_i = \sum_{k=1}^{n} t_k^i \quad (i = 0, \pm 1, \pm 2, \dots)$$

and form the Toeplitz matrix $T_n = [\, c_{i-j} \,]_{i,j=1}^n$. *Then* $T_n = T_n(\mathbf{f})$ *for*

$$\mathbf{f}(t) = \sum_{k=1}^{n} \frac{1}{1-t_k t}\,.$$

Taking $\mathbf{u}^J(t) = \prod_{k=1}^n (1 - t_k t)$ *into account we find that*

$$\mathbf{f}(t) = \frac{(\mathbf{u}')^J(t)}{\mathbf{u}^J(t)}\,.$$

Here $\mathbf{u}'$ *has to be considered as a vector in* $\mathbb{F}^n$, *that means* $(\mathbf{u}')^J(t) = t^{n-1}\mathbf{u}'(t^{-1})$.

Like for Hankel matrices, it can be shown that, for any given polynomial $\mathbf{u}(t)$ of degree $2n-2$ with $\mathbf{u}(0) \neq 0$, any $n \times n$ Toeplitz matrix has a generating function with denominator polynomial $\mathbf{u}(t)$. More important is the following Toeplitz analogue of Theorem 7.4 about generating functions of nonsingular Hankel matrices. Note that its proof is somehow different to the Hankel case.

Theorem 8.3. *Let* $T_n = [\, a_{i-j} \,]_{i,j=1}^n$ *be a nonsingular Toeplitz matrix,* $\{\mathbf{u}(t), \mathbf{v}(t)\}$ *be a fundamental system of* T_n. *Furthermore, let* $\alpha, \beta \in \mathbb{F}$ *be such that* $\mathbf{w}(t) = \alpha\mathbf{u}(t) + \beta\mathbf{v}(t)$ *is of degree* n *and* $\mathbf{w}(0) \neq 0$. *Then there is a* $\mathbf{p} \in \mathbb{F}^{n+1}$ *such that*

$$T_n = T_n\left(\frac{\mathbf{p}}{\mathbf{w}}\right)\,.$$

Proof. Let ∂T_n be the matrix defined in (4.7). Then $\mathbf{w}$ belongs to the nullspace of ∂T_n. We find a_n and a_{-n} via the equations

$$\begin{bmatrix} a_n & a_{n-1} & \dots & a_0 \end{bmatrix} \mathbf{w} = \begin{bmatrix} a_0 & \dots & a_{1-n} & a_{-n} \end{bmatrix} \mathbf{w} = 0$$

and form the $(n+1)\times(n+1)$ Toeplitz matrix $T_{n+1} = [\, a_{i-j} \,]_{i,j=1}^{n+1}$. Then we have $T_{n+1}\mathbf{w} = \mathbf{0}$. Now we represent T_{n+1} as $T_{n+1} = T_{n+1}^{+} + T_{n+1}^{-}$, where T_{n+1}^{+} is a lower triangular and T_{n+1}^{-} is an upper triangular Toeplitz matrix and define

$$\mathbf{p} = T_{n+1}^{+}\mathbf{w}\,. \tag{8.4}$$

We have also $\mathbf{p} = -T_{n+1}^{-}\mathbf{w}$. A comparison of coefficients reveals that

$$\frac{\mathbf{p}(t)}{\mathbf{w}(t)} = a_0^{+} + a_1 t + \dots + a_n t^n + \dots \quad \text{and} \quad \frac{\mathbf{p}(t)}{\mathbf{w}(t)} = -a_0^{-} - a_{-1}t^{-1} - \dots - a_{-n}t^{-n} - \dots\,,$$

where $a_0^{\pm}$ are the diagonal entries of $T_{n+1}^{\pm}$, respectively. Thus $T_n = T_n(\mathbf{f})$ for $\mathbf{f}(t) = \frac{\mathbf{p}(t)}{\mathbf{w}(t)}$. □

Example 8.4. *Let us compute generating functions for the identity matrix I_n. We observe first that $\{\mathbf{e}_1, \mathbf{e}_{n+1}\}$ is a fundamental system. To meet the conditions of the theorem we choose $\alpha \neq 0, \beta \neq 0$. Then we find that $a_n = -\frac{\beta}{\alpha}$ and $a_{-n} = -\frac{\alpha}{\beta}$. We can choose now*

$$T_{n+1}^{+} = \gamma I_{n+1} - \frac{\beta}{\alpha}\,\mathbf{e}_{n+1}\mathbf{e}_1^T\,,$$

where γ is arbitrary. For $\gamma = 0$ this leads to $\mathbf{p} = -\beta\mathbf{e}_{n+1}$. Thus generating functions for I_n are given by

$$\mathbf{f}(t) = \frac{-\beta t^n}{\alpha + \beta t^n}\,.$$

If we choose a different γ, then the resulting function differs from that function only by a constant.

2. Matrices with symmetry properties. It is a little bit surprising that if $\mathbf{f}(t)$ is symmetric in the sense that

$$\mathbf{f}(t^{-1}) = \mathbf{f}(t)\,, \tag{8.5}$$

then the matrix $T_n(\mathbf{f})$ becomes skewsymmetric. Symmetric matrices $T_n(\mathbf{f})$ are obtained if $\mathbf{f}(t)$ satisfies

$$\mathbf{f}(t^{-1}) = -\mathbf{f}(t)\,. \tag{8.6}$$

In the case $\mathbb{F} = \mathbb{C}$ the matrices $T_n(\mathbf{f})$ are Hermitian if

$$\mathbf{f}(\overline{t}^{\,-1}) = -\overline{\mathbf{f}(t)}\,. \tag{8.7}$$

This is equivalent to saying that $\mathbf{f}(t)$ takes purely imaginary values on the unit circle. We show that the converse is, in a sense, also true.

Proposition 8.5. *If T_n is a nonsingular symmetric, skewsymmetric or Hermitian Toeplitz matrix, then there exists a generating function $\mathbf{f}(t)$ for T_n of degree n that satisfies the conditions* (8.6), (8.5), (8.7), *respectively.*

Proof. If T_n is symmetric, then the fundamental system consists of a symmetric vector $\mathbf{u}$ and a skewsymmetric vector $\mathbf{v}$. If the last component of $\mathbf{u}$ does not vanish, then we can choose $\mathbf{w} = \mathbf{u}$ to satisfy the conditions of Theorem 8.3. Further, we obtain $a_{-n} = a_n$, thus T_{n+1} is symmetric, and the choice $T_{n+1}^- = (T_{n+1}^+)^T$ is possible. Hence we have

$$\mathbf{p} = T_{n+1}^+ \mathbf{u} = -(T_{n+1}^+)^T \mathbf{u} = -T_{n+1}^+ \mathbf{u}^J = -\mathbf{p}^J \, .$$

That means that $\mathbf{p}$ is skewsymmetric, which implies that $\mathbf{f}(t) = \frac{\mathbf{p}(t)}{\mathbf{u}(t)}$ satisfies (8.6).

If the last component of $\mathbf{u}$ vanishes, then the last component of $\mathbf{v}$ must be nonzero and we can choose $\mathbf{w} = \mathbf{v}$. Again we obtain $a_{-n} = a_n$, so that T_{n+1} is symmetric. With the choice $T_{n+1}^- = (T_{n+1}^+)^T$ we have

$$\mathbf{p} = T_{n+1}^+ \mathbf{v} = -(T_{n+1}^+)^T \mathbf{v} = -T_{n+1}^+ \mathbf{v}^J = \mathbf{p}^J \, .$$

Thus, $\mathbf{p}$ is symmetric and $\mathbf{f}(t) = \frac{\mathbf{p}(t)}{\mathbf{v}(t)}$ satisfies (8.6). The proof of the other cases is analogous. We have to take into account that a fundamental system of a nonsingular skewsymmetric Toeplitz matrix consists of two symmetric vectors and the fundamental system of a nonsingular Hermitian Toeplitz matrix of two conjugate-symmetric vectors. □

To discuss this proposition we consider the Examples 8.1, 8.4. The Toeplitz matrix $T_n = [\, c^{i-j} \,]_{i,j=1}^n$ in Example 8.1 is Hermitian if c is on the unit circle, but its generating function $\frac{1}{1-ct}$ does not satisfy (8.7). However, since $\frac{1}{1-ct} = \frac{1}{2}\left(\frac{1+ct}{1-ct} + 1\right)$ we have also

$$T_n = T_n\left(\frac{1}{2}\,\frac{1+ct}{1-ct}\right) \, ,$$

and this generating function satisfies (8.7). Generating functions for the identity matrix I_n satisfying (8.6) and so reflecting its symmetry are

$$\mathbf{f}(t) = \frac{1}{2}\,\frac{1-t^n}{1+t^n} \quad \text{and} \quad \mathbf{f}(t) = \frac{1}{2}\,\frac{1+t^n}{1-t^n} \, .$$

3. Vandermonde factorization of nonsingular Toeplitz matrices. Let T_n be a nonsingular $n \times n$ Toeplitz matrix with complex entries and $\mathbf{f}(t) = \frac{\mathbf{p}(t)}{\mathbf{w}(t)}$ be a generating function of degree n with $\mathbf{f}(\infty) = 0$. According to Theorem 8.3 and Lemma 7.7 such a function exists and, due to the freedom in the choice of $\mathbf{w}(t)$ we can assume that $\mathbf{w}(t)$ has only simple roots $t_1, \ldots, t_n$. Using the partial fraction decomposition of $\mathbf{f}(t)$ in the form

$$\mathbf{f}(t) = \sum_{i=1}^{n} -\frac{1}{t_i}\,\frac{\gamma_i}{1 - t_i^{-1} t}$$

as well as (8.3) we can conclude, in analogy to the Hankel case, the following.

Proposition 8.6. *Let T_n be a nonsingular $n \times n$ Toeplitz matrix, $\{\mathbf{u}, \mathbf{v}\}$ a fundamental system of T_n and $\alpha, \beta \in \mathbb{C}$ such that $\mathbf{w}(t) = \alpha \mathbf{u}(t) + \beta \mathbf{v}(t)$ is of degree n,*

satisfies $\mathbf{w}(0) \neq 0$*, and has simple roots* $t_1, \ldots, t_n$*. Then* T_n *admits a representation*

$$T_n = V_n(\mathbf{t}^{-1})^T D V_n(\mathbf{t}) \tag{8.8}$$

where $\mathbf{t} = (t_i)_{i=1}^n$ *and* $\mathbf{t}^{-1} = (t_i^{-1})_{i=1}^n$*, and* D *is diagonal,* $D = \operatorname{diag}\left(-t_k^{-1}\gamma_k\right)_{k=1}^n$.

The diagonal matrix can be expressed in terms of the generating function. If $D = \operatorname{diag}(\delta_i)_{i=1}^n$, then

$$\delta_i = -\frac{1}{t_i}\,\frac{\mathbf{p}(t_i)}{\mathbf{w}'(t_i)} = -\frac{1}{t_i}\left(\left(\frac{1}{\mathbf{f}}\right)'(t_i)\right)^{-1} . \tag{8.9}$$

Note that, like for Hankel matrices, the Vandermonde factorization for a nonsingular Toeplitz matrix $T_n(\mathbf{f})$ extends to all Toeplitz matrices $T_k(\mathbf{f})$ with $k \geq n$ as

$$T_k(\mathbf{f}) = V_k(\mathbf{t}^{-1})^T D V_k(\mathbf{t})\,. \tag{8.10}$$

4. Hermitian Toeplitz matrices. Let the assumptions of Proposition 8.6 be satisfied. We consider the special case of an Hermitian Toeplitz matrix T_n. In this case T_n has a fundamental system consisting of two conjugate-symmetric vectors. If we choose α and β as reals, then the vector $\mathbf{w}$ in Proposition 8.6 is also conjugate-symmetric. For a polynomial with a conjugate-symmetric coefficient vector the roots are symmetric with respect to the unit circle $\mathbb{T}$. That means if t_0 is a root, then $\overline{t}_0^{-1}$ is also a root. In particular, the roots not on $\mathbb{T}$ appear in pairs that are symmetric with respect to $\mathbb{T}$.

Let $t_1, \ldots, t_r$ be the roots of $\mathbf{w}(t)$ on $\mathbb{T}$ and $t_{r+1}, t_{r+2} = \overline{t}_{r+1}^{-1}, \ldots, t_{n-1}, t_n = \overline{t}_{n-1}^{-1}$ be the roots of $\mathbf{w}(t)$ which are not on $\mathbb{T}$. Note that for the coefficients γ_i in the partial fraction decomposition of $\mathbf{f}(t)$, which are the residuals at t_i, we have $\delta_{r+2} = \overline{\delta}_{r+1}, \ldots, \delta_n = \overline{\delta}_{n-1}$. Furthermore,

$$V_n(\mathbf{t}^{-1})^T = V_n(\mathbf{t})^* \operatorname{diag}(I_r, \underbrace{J_2, \ldots, J_2}_{l})\,,$$

where $r + 2l = n$. Now Proposition 8.6 leads to the following.

Corollary 8.7. *If the Toeplitz matrix in Proposition* 8.6 *is Hermitian, then it admits a representation*

$$T_n = V_n(\mathbf{t})^* D_1 V_n(\mathbf{t})\,,$$

where

$$D_1 = \operatorname{diag}\left(\delta_1, \ldots, \delta_r, \begin{bmatrix} 0 & \overline{\delta}_{r+1} \\ \delta_{r+1} & 0 \end{bmatrix}, \ldots, \begin{bmatrix} 0 & \overline{\delta}_{n-1} \\ \delta_{n-1} & 0 \end{bmatrix}\right)$$

In particular, the matrices T_n *and* D_1 *are congruent.*

5. Signature and Cauchy index. Corollary 8.7 allows us to express the signature of T_n in terms of the signs of the diagonal elements $\delta_i\,(i = 1, \ldots, r)$ of D_1 (cf. Corollary 7.12). Our aim is now to express it in terms of the Cauchy index.

Let $\mathbf{f}(t)$ be a rational function of degree n satisfying (8.7). Then the function $\frac{1}{\mathrm{i}}\mathbf{f}(t)$ takes real values on the unit circle, thus the Cauchy index (see Section 7.4) $\mathrm{ind}_{\mathbb{T}}\frac{1}{\mathrm{i}}\mathbf{f}(t)$ is well defined. It is the difference of the number of poles on $\mathbb{T}$ of positive type and the number of poles of negative type. Let $t_j = e^{\mathrm{i}\theta_j}$ be a pole on $\mathbb{T}$. Then t_j is of positive (negative) type if and only if the (real-valued) function

$$\varphi(\theta) = \frac{\mathrm{i}}{\mathbf{f}(e^{\mathrm{i}\theta})}$$

is increasing (decreasing) in a neighborhood of θ_j. For a simple pole this is equivalent to $\varphi'(\theta_j) > 0$ ($\varphi'(\theta_j) < 0$). We have

$$\varphi'(\theta_j) = -t_j \left(\frac{1}{\mathbf{f}}\right)'(t_j)\,.$$

Comparing this with (8.9) we conclude that $\varphi(\theta_j) = \delta_j^{-1}$. We arrive at the following statement for the case of simple poles. It can be generalized to multiple poles by using the continuity arguments from the proof of Proposition 7.13.

Proposition 8.8. *Let $\mathbf{f}(t)$ be a proper rational function with degree n that takes imaginary values on the unit circle. Then*

$$\operatorname{sgn} T_n(\mathbf{f}) = \mathrm{ind}_{\mathbb{T}}\,\frac{1}{\mathrm{i}}\mathbf{f}(t)\,.$$

We also obtain a criterion of positive definiteness.

Corollary 8.9. *Let $\mathbf{f}(t)$ be a proper rational function of degree n, and let*

$$\mathbf{f}(t) = \frac{\mathrm{i}\mathbf{p}(t)}{\mathbf{u}(t)}$$

be its reduced representation in which $\mathbf{u}$ and $\mathbf{p}$ are conjugate-symmetric. Then $T_n(\mathbf{f})$ is positive definite if and only if the polynomials $\mathbf{u}(t)$ and $\mathbf{p}(t)$ have only roots on the unit circle, these roots are simple and interlaced.

6. Congruence to T-Bezoutians. The Toeplitz analogue of Proposition 7.15 is as follows. (Concerning the order of $\mathrm{Bez}_T(\mathbf{u},\mathbf{v})$ and of $B_\pm(\mathbf{u})$ compare the remarks before Proposition 7.15.)

Proposition 8.10. *Let $\mathbf{u},\mathbf{v} \in \mathbb{F}^{n+1}$, where $\mathbf{u}$ has nonvanishing first and last components. Then, for $k \geq n$, the T-Bezoutian of $\mathbf{u}(t)$ and $\mathbf{v}(t)$ is related to the $k\times k$ Toeplitz matrix $T_k(\mathbf{f})$ generated by $\mathbf{f}(t) = \frac{\mathbf{v}(t)}{\mathbf{u}(t)}$ via*

$$\mathrm{Bez}_T(\mathbf{u},\mathbf{v}) = B_-(\mathbf{u})\,T_k(\mathbf{f})\,B_+(\mathbf{u})\,,$$

where $B_\pm(\mathbf{u})$ are introduced in (2.14).

Proof. For $B = \mathrm{Bez}_T(\mathbf{u},\mathbf{v})$ we have

$$B(t,s) = \mathbf{u}(t)\,\frac{\mathbf{f}(t) - \mathbf{f}(s^{-1})}{1-ts}\,(-\mathbf{u}^J(s))\,.$$

Applying (8.2) we obtain

$$B = \begin{bmatrix} u_1 & & & \\ \vdots & \ddots & & O \\ u_n & \dots & u_1 & \end{bmatrix} T \begin{bmatrix} -u_{n+1} & \dots & -u_2 \\ & \ddots & \vdots \\ & & -u_{n+1} \\ & O & \end{bmatrix},$$

where $\mathbf{u} = (u_i)_{i=1}^{n+1}$ and $T = [\, a_{i-j} \,]_{i,j=1}^{\infty}$. Hence $B = B_-(\mathbf{u})T_k(\mathbf{f})B_+(\mathbf{u})$. □

Corollary 8.11. *If* $T_n\left(\frac{\mathbf{v}}{\mathbf{u}}\right)$ *is Hermitian and if* $\mathbf{u}$ *is a conjugate-symmetric* (*or conjugate-skewsymmetric*) *polynomial of degree* $n+1$ *then* $\mathrm{Bez}_T(\mathbf{v},\mathbf{u})$ (*or* $\mathrm{Bez}_T(\mathbf{u},\mathbf{v})$) *and* $T_n(\mathbf{f})$ *are congruent.*

Corollary 8.12. *For* $k \geq n$ *the rank of* $T_k(\mathbf{f})$ *is equal to the degree of* $\mathbf{f}(t)$. *In particular, if* $\mathbf{f}(t)$ *has degree* n, *then* $T_n(\mathbf{f})$ *is nonsingular.*

If we combine Proposition 8.10 with Barnett's formula in Theorem 6.3, then we obtain the representation

$$T_n\left(\frac{\mathbf{v}}{\mathbf{u}}\right) = \mathbf{v}^J(C(\mathbf{u}^J))B_+(\mathbf{u})^{-1}, \tag{8.11}$$

where we assume that $\mathbf{u}(t)$ is comonic.

7. Inverses of T-Bezoutians. Now we show that relation (8.11) leads to an inversion formula for T-Bezoutians.

Theorem 8.13. *Let* $\mathbf{u}, \mathbf{v} \in \mathbb{F}^{n+1}$ *be such that* $\mathbf{u}(t)$ *and* $\mathbf{v}(t)$ *are coprime and the first and last components of* $\mathbf{u}$ *do not vanish. If* $(\mathbf{q}(t), \mathbf{p}(t))$, $\mathbf{q}, \mathbf{p} \in \mathbb{F}^{n+1}$, *is the solution of the Diophantine equation*

$$\mathbf{u}(t)\mathbf{q}(t) + \mathbf{v}(t)\mathbf{p}(t) = t^n, \tag{8.12}$$

then $\mathrm{Bez}_T(\mathbf{u},\mathbf{v})$ *is invertible and the inverse is given by*

$$\mathrm{Bez}_T(\mathbf{u},\mathbf{v})^{-1} = T_n\left(\frac{\mathbf{p}}{\mathbf{u}}\right).$$

Proof. First we note that (8.12) is equivalent to

$$\mathbf{u}^J(t)\mathbf{q}^J(t) + \mathbf{v}^J(t)\mathbf{p}^J(t) = t^n.$$

Thus, $\mathbf{v}^J(t)\mathbf{p}^J(t) \equiv t^n$ modulo $\mathbf{u}^J(t)$. From Theorem 6.3, the representation (8.11), and the Cayley-Hamilton theorem, now we obtain

$$\begin{aligned} \mathrm{Bez}_T(\mathbf{u},\mathbf{v})T_n\left(\frac{\mathbf{p}}{\mathbf{u}}\right) &= B_-(\mathbf{u})\mathbf{v}^J(C(\mathbf{u}^J))p^J(C(\mathbf{u}^J))B_+(\mathbf{u})^{-1} \\ &= B_-(\mathbf{u})C(\mathbf{u}^J)^n B_+(\mathbf{u})^{-1}. \end{aligned}$$

Taking (6.7) into account this leads to

$$\mathrm{Bez}_T(\mathbf{u},\mathbf{v})T_n\left(\frac{\mathbf{p}}{\mathbf{u}}\right) = B_+(\mathbf{u})B_+(\mathbf{u})^{-1} = I_n,$$

and the theorem is proved. □

8. Relations between Toeplitz and Hankel matrices. We know that if T_k is a $k \times k$ Toeplitz matrix, then $J_k T_k$ is Hankel, and vice versa. We show how the generating functions are related.

Proposition 8.14. *Let* $\mathbf{u}, \mathbf{q} \in \mathbb{F}^{n+1}$, *where* $\mathbf{u}$ *has nonvanishing first and last components, and let* $\frac{\mathbf{q}(t)}{\mathbf{u}(t)}$ *be a generating function of* T_k. *For* $k \geq n$, *let* $\mathbf{p} \in \mathbb{F}^n$ *be such that* $-\mathbf{p}(t) \in \mathbb{F}^n(t)$ *is the remainder polynomial of* $t^k \mathbf{q}^J(t)$ *divided by* $\mathbf{u}^J(t)$. *Then*

$$J_k T_k \left(\frac{\mathbf{q}}{\mathbf{u}}\right) = H_k \left(\frac{\mathbf{p}}{\mathbf{u}}\right) .$$

Proof. According to the definition of $\mathbf{p}$ we have

$$t^k \mathbf{q}^J(t) = -\mathbf{p}^J(t) + \mathbf{r}(t)\mathbf{u}^J(t)$$

for some $\mathbf{r}(t) \in \mathbb{F}^{k+1}(t)$. This is equivalent to $\mathbf{q}(t) = -t^k \mathbf{p}(t) + t^k \mathbf{r}(t^{-1})\mathbf{u}(t)$, and we obtain

$$\frac{\mathbf{q}(t)}{\mathbf{u}(t)} = -t^k \frac{\mathbf{p}(t)}{\mathbf{u}(t)} + t^k \mathbf{r}(t^{-1}).$$

Thus

$$t^k \mathbf{r}(t^{-1}) = a_0^+ + a_1 t + \cdots + a_k t^k .$$

On the other hand,

$$\frac{\mathbf{p}(t)}{\mathbf{u}(t)} = \mathbf{r}(t^{-1}) - t^{-k} \frac{\mathbf{q}(t)}{\mathbf{u}(t)} .$$

Consequently,

$$\frac{\mathbf{p}(t)}{\mathbf{u}(t)} = a_k + a_{k-1} t^{-1} + \cdots + a_0 t^{-k} a_{-1} t^{-k-1} + \cdots ,$$

where $a_0 = a_0^+ + a_0^-$. This means that

$$H_k = H_k \left(\frac{\mathbf{p}}{\mathbf{u}}\right) = \begin{bmatrix} a_{k-1} & a_{k-2} & \cdots & a_0 \\ a_{k-2} & & \iddots & \\ \vdots & \iddots & & \vdots \\ a_0 & \cdots & & a_{1-k} \end{bmatrix} .$$

Thus $H_k = J_k T_k \left(\frac{\mathbf{q}}{\mathbf{u}}\right)$. □

9. Vandermonde reduction of Bezoutians

In this section the underlying field is the field of complex numbers, $\mathbb{F} = \mathbb{C}$. Some of the results can be extended to general algebraically closed fields.

In Sections 7 and 8 we showed that nonsingular Hankel and Toeplitz matrices can be represented as a product of the transpose of a Vandermonde matrix V_n^T, a diagonal matrix, and the Vandermonde matrix V_n, and we called this Vandermonde factorization. Since inverses of Hankel and Toeplitz matrices are H- and T-Bezoutians, respectively, this is equivalent to the fact that nonsingular Bezoutians can be reduced to diagonal form by multiplying them by V_n from the left and by

V_n^T from the right. We call this kind of factorization *Vandermonde reduction of Bezoutians*. In this section we give a direct derivation of Vandermonde reduction of Bezoutians and generalize it to general, not necessarily nonsingular Bezoutians.

Hereafter, the notations *confluent* and *non-confluent* are used in connection with confluent and non-confluent Vandermonde matrices, respectively.

1. Non-confluent Hankel case. To begin with, let us recall that, for $\mathbf{t} = (t_i)_{i=1}^n$, the (non-confluent) Vandermonde matrix $V_m(\mathbf{t})$ is defined by

$$V_m(\mathbf{t}) = \begin{bmatrix} 1 & t_1 & \dots & t_1^{m-1} \\ 1 & t_2 & \dots & t_2^{m-1} \\ \vdots & \vdots & & \vdots \\ 1 & t_n & \dots & t_n^{m-1} \end{bmatrix}.$$

If $m = n$ and the t_i are distinct, then $V_n(\mathbf{t})$ is nonsingular.

Obviously, for $\mathbf{x} \in \mathbb{C}^m$, $V_m(\mathbf{t})\mathbf{x} = (\mathbf{x}(t_i))_{i=1}^n$. Furthermore, for an $n \times n$ matrix B and $\mathbf{s} = (s_j)_{j=1}^n$,

$$V_n(\mathbf{t})BV_n(\mathbf{s})^T = [\, B(t_i, s_j)\,]_{i,j=1}^n . \tag{9.1}$$

We specify this for $\mathbf{t} = \mathbf{s}$ with pairwise distinct components and an H-Bezoutian $B = \mathrm{Bez}_H(\mathbf{u}, \mathbf{v})$, where $\mathbf{u}(t), \mathbf{v}(t) \in \mathbb{C}^{n+1}(t)$ and $\mathbf{u}(t)$ has degree n. In this case the off diagonal entries c_{ij} of $C = V_n(\mathbf{t})BV_n(\mathbf{t})^T$ are given by

$$c_{ij} = \frac{\mathbf{u}(t_i)\mathbf{v}(t_j) - \mathbf{v}(t_i)\mathbf{u}(t_j)}{t_i - t_j} . \tag{9.2}$$

Our aim is to find $\mathbf{t}$ such that C is diagonal. One possibility is to choose the zeros of $\mathbf{u}(t)$. This is possible if $\mathbf{u}(t)$ has only simple zeros. A more general case is presented next.

Proposition 9.1. *Let $\alpha \in \mathbb{C}$ be such that $\mathbf{w}(t) = \mathbf{u}(t) - \alpha\mathbf{v}(t)$ has simple roots $t_1, \dots, t_n$, and let $\mathbf{t} = (t_i)_{i=1}^n$. Then*

$$V_n(\mathbf{t})\mathrm{Bez}_H(\mathbf{u}, \mathbf{v})V_n(\mathbf{t})^T = \mathrm{diag}(\gamma_i)_{i=1}^n , \tag{9.3}$$

where

$$\gamma_i = \mathbf{u}'(t_i)\mathbf{v}(t_i) - \mathbf{u}(t_i)\mathbf{v}'(t_i) . \tag{9.4}$$

Proof. According to Lemma 2.2 we have $\mathrm{Bez}_H(\mathbf{u}, \mathbf{v}) = \mathrm{Bez}_H(\mathbf{w}, \mathbf{v})$. From (9.2) we see that the off diagonal elements of the matrix $V_n(\mathbf{t})\mathrm{Bez}_H(\mathbf{u}, \mathbf{v})V_n(\mathbf{t})^T$ vanish. The expression (9.4) follows from (9.2), $\gamma_i = \lim\limits_{t \to t_i} \frac{\mathbf{u}(t)\mathbf{v}(t_i) - \mathbf{v}(t)\mathbf{u}(t_i)}{t - t_i}$. □

Remark 9.2. If $\mathbf{u}(t)$ and $\mathbf{v}(t)$ are coprime, which is the same as saying that $\mathrm{Bez}_H(\mathbf{u}, \mathbf{v})$ is nonsingular, then according to Lemma 7.7 for almost all values of α the polynomial $\mathbf{w}(t)$ has simple roots.

Remarkably the special case $\mathbf{v}(t) = 1$ of (9.3) leads to a conclusion concerning the inverse of a Vandermonde matrix.

Corollary 9.3. *Let* $\mathbf{t} = (t_i)_{i=1}^n$ *have pairwise distinct components, and let* $\mathbf{u}(t)$ *be defined by* $\mathbf{u}(t) = \prod_{i=1}^n (t - t_i)$. *Then the inverse of* $V_n(\mathbf{t})$ *is given by*

$$V_n(\mathbf{t})^{-1} = B(\mathbf{u}) V_n(\mathbf{t})^T \operatorname{diag}\left(\frac{1}{\mathbf{u}'(t_i)}\right)_{i=1}^n , \tag{9.5}$$

where $B(\mathbf{u})$ *is the upper triangular Hankel matrix introduced in* (2.2).

Now we consider the H-Bezoutian of real polynomials $\mathbf{u}(t)$ and $\mathbf{v}(t)$, which is an Hermitian matrix. Similarly to Corollary 7.11, Proposition 9.1 leads to a matrix congruence.

Corollary 9.4. *Let in Proposition* 9.1 *the polynomials* $\mathbf{u}(t)$ *and* $\mathbf{v}(t)$ *be real and* $\alpha \in \mathbb{R}$. *Furthermore, let* $t_1, \dots, t_r$ *be the (simple) real and* $(t_{r+1}, \overline{t}_{r+1}), \dots, (t_{n-1}, \overline{t}_{n-1})$ *the (simple) non-real roots of* $\mathbf{w}(t)$. *Then*

$$V_n(\mathbf{t}) \operatorname{Bez}_H(\mathbf{u}, \mathbf{v}) V_n(\mathbf{t})^* = \operatorname{diag}\left(\gamma_1, \dots, \gamma_r, \begin{bmatrix} 0 & \overline{\gamma}_{r+1} \\ \gamma_{r+1} & 0 \end{bmatrix}, \dots, \begin{bmatrix} 0 & \overline{\gamma}_{n-1} \\ \gamma_{n-1} & 0 \end{bmatrix}\right).$$

In particular,

$$\operatorname{sgn} \operatorname{Bez}_H(\mathbf{u}, \mathbf{v}) = \sum_{i=1}^r \operatorname{sgn} \gamma_i .$$

Note that it follows from the Hermitian symmetry of the matrix that the numbers $\gamma_1, \dots, \gamma_r$ are real.

2. Non-confluent Toeplitz case. Let $\mathbf{t} = (t_i)_{i=1}^n$ have distinct nonzero components, $\mathbf{t}^{-1} = \left(\frac{1}{t_i}\right)_{i=1}^n$. If we specify (9.1) for a T-Bezoutian $\operatorname{Bez}_T(\mathbf{u}, \mathbf{v})$, then we obtain for the off diagonal entries c_{ij} of $C = V_n(\mathbf{t}) \operatorname{Bez}_T(\mathbf{u}, \mathbf{v}) V_n(\mathbf{t}^{-1})^T$ the relation

$$c_{ij} = -\frac{\mathbf{u}(t_i)\mathbf{v}(t_j) - \mathbf{v}(t_i)\mathbf{u}(t_j)}{t_i - t_j}\, t_j^{1-n} .$$

For the diagonal entries we obtain using l'Hospital's rule

$$c_{ii} = (\mathbf{v}'(t_i)\mathbf{u}(t_i) - \mathbf{u}'(t_i)\mathbf{v}(t_i)) t_i^{1-n} .$$

In the same way as in Proposition 9.1 we derive the following.

Proposition 9.5. *Let* $\alpha \in \mathbb{C}$ *be such that* $\mathbf{w}(t) = \mathbf{u}(t) - \alpha \mathbf{v}(t)$ *has simple nonzero roots* $t_1, \dots, t_n$, *and let* $\mathbf{t} = (t_i)_{i=1}^n$. *Then*

$$V_n(\mathbf{t}) \operatorname{Bez}_T(\mathbf{u}, \mathbf{v}) V_n(\mathbf{t}^{-1})^T = \operatorname{diag}(\gamma_i)_{i=1}^n ,$$

where

$$\gamma_i = (\mathbf{v}'(t_i)\mathbf{u}(t_i) - \mathbf{u}'(t_i)\mathbf{v}(t_i)) t_i^{1-n} .$$

Let $\mathbf{u}$ be conjugate-symmetric and $\mathbf{v}$ conjugate-skewsymmetric. Then the matrix $\operatorname{Bez}_T(\mathbf{u}, \mathbf{v})$ is Hermitian. For purely imaginary α, the roots of $\mathbf{w}(t) = \mathbf{u}(t) - \alpha \mathbf{v}(t)$ are located symmetrically with respect to the unit circle $\mathbb{T}$.

Corollary 9.6. *Let in Proposition* 9.5 *the polynomial* $\mathbf{u}$ *be conjugate-symmetric and* $\mathbf{v}(t)$ *be conjugate-skewsymmetric, and let* $\alpha \in \mathrm{i}\mathbb{R}$. *Furthermore, let* $t_1, \dots, t_r$ *be the* (*simple*) *roots of* $\mathbf{w}(t)$ *on* $\mathbb{T}$ *and* $t_{r+1}, t_{r+2} = \overline{t}_{r+1}^{-1}, \dots, t_{n-1}, t_n = \overline{t}_{n-1}^{-1}$ *the* (*simple*) *roots of* $\mathbf{w}(t)$ *outside* $\mathbb{T}$. *Then*

$$V_n(\mathbf{t})\mathrm{Bez}_T(\mathbf{u},\mathbf{v})V_n(\mathbf{t})^* = \mathrm{diag}\left(\gamma_1, \dots, \gamma_r, \begin{bmatrix} 0 & \overline{\gamma}_{r+1} \\ \gamma_{r+1} & 0 \end{bmatrix}, \dots, \begin{bmatrix} 0 & \overline{\gamma}_{n-1} \\ \gamma_{n-1} & 0 \end{bmatrix}\right).$$

In particular,

$$\operatorname{sgn}\mathrm{Bez}_T(\mathbf{u},\mathbf{v}) = \sum_{i=1}^{r} \operatorname{sgn}\gamma_i\,.$$

3. Confluent case. Here we need the following generalization of a Vandermonde matrix. Let $\mathbf{t} = (t_i)_{i=1}^m$ and $\mathbf{r} = (r_i)_{i=1}^m \in \mathbb{N}^m$. We denote by $V_n(t_i, r_i)$ the $r_i \times n$ matrix

$$V_n(t_i, r_i) = \begin{bmatrix} 1 & t_i & t_i^2 & & \dots & t_i^{n-1} \\ 0 & 1 & 2t_i & & \dots & (n-1)t_i^{n-2} \\ \vdots & & \ddots & & & \vdots \\ 0 & \dots & 0 & 1 & \dots & \binom{n-1}{r_i-1} t_i^{n-r_i} \end{bmatrix}$$

and introduce the matrix

$$V_n(\mathbf{t},\mathbf{r}) = \begin{bmatrix} V_n(t_1, r_1) \\ \vdots \\ V_n(t_m, r_m) \end{bmatrix},$$

which is called *confluent Vandermonde matrix*.

Now we show that in case that $\mathbf{u}(t)$ has multiple roots Bezoutians can be reduced to block diagonal form with the help of confluent Vandermonde matrices $V_n(\mathbf{t},\mathbf{r})$. We restrict ourselves to the case of H-Bezoutians. The case of T-Bezoutians is analogous.

First we consider the special single node case $\mathbf{t} = 0$, $\mathbf{r} = r$. Suppose that $\mathbf{u} = (u_i)_{i=1}^{n+1}$ and $\mathbf{u}(t)$ has the root $t = 0$ with multiplicity r, i.e., $u_1 = u_2 = \dots = u_r = 0$. Obviously, $V_n(0, r) = \begin{bmatrix} I_r & O \end{bmatrix}$. Hence $V_n(0,r)BV_n(0,r)^T$ is the $r \times r$ leading principal submatrix of $B = \mathrm{Bez}_H(\mathbf{u},\mathbf{v})$. We denote this matrix by $\Gamma(0)$ and observe, taking Theorem 3.2 into account, that

$$\Gamma(0) = \begin{bmatrix} v_1 & & \\ \vdots & \ddots & \\ v_r & \dots & v_1 \end{bmatrix} \begin{bmatrix} & & u_{r+1} \\ & \iddots & \vdots \\ u_{r+1} & \dots & u_{2r+1} \end{bmatrix} = \begin{bmatrix} & & w_1 \\ & \iddots & \vdots \\ w_1 & \dots & w_r \end{bmatrix},$$

where

$$\sum_{i=1}^{2n-r+1} w_i t^{i-1} = \mathbf{u}(t)\mathbf{v}(t)t^{-r}\,.$$

Taylor expansion gives us

$$u_i = \frac{1}{(i-1)!}\,\mathbf{u}^{(i-1)}(0)$$

and an analogous expression for v_i.

Suppose now that t_0 is a root of $\mathbf{u}(t)$ with multiplicity r. We consider the polynomials $\widetilde{\mathbf{u}}(t) = \mathbf{u}(t-t_0)$ and $\widetilde{\mathbf{v}}(t) = \mathbf{v}(t-t_0)$ and $\widetilde{B} = [\,\widetilde{b}_{ij}\,]_{i,j=1}^{n} = \mathrm{Bez}_H(\widetilde{\mathbf{u}}, \widetilde{\mathbf{v}})$. Then, for $k = 0, 1, \ldots,$

$$\widetilde{\mathbf{u}}^{(k)}(0) = \mathbf{u}^{(k)}(t_0)\,, \quad \widetilde{\mathbf{v}}^{(k)}(0) = \mathbf{v}^{(k)}(t_0)$$

and

$$\begin{aligned}\widetilde{b}_{ij} &= \frac{1}{(i-1)!(j-1)!}\,\frac{\partial^{i+j-2}}{\partial t^{i-1}\partial s^{j-1}}\,\widetilde{B}(t,s)\,|_{(0,0)} \\ &= \frac{1}{(i-1)!(j-1)!}\,\frac{\partial^{i+j-2}}{\partial t^{i-1}\partial s^{j-1}}\,B(t,s)\,|_{(t_0,t_0)}\,.\end{aligned}$$

Hence

$$\Gamma(t_0) = \begin{bmatrix} \widetilde{v}_1 & & \\ \vdots & \ddots & \\ \widetilde{v}_r & \ldots & \widetilde{v}_1 \end{bmatrix} \begin{bmatrix} & & \widetilde{u}_{r+1} \\ & \iddots & \vdots \\ \widetilde{u}_{r+1} & \ldots & \widetilde{u}_{2r+1} \end{bmatrix} = \begin{bmatrix} & & w_1 \\ & \iddots & \vdots \\ w_1 & \ldots & w_r \end{bmatrix},$$

where $\widetilde{u}_i = \dfrac{1}{(i-1)!}\,\mathbf{u}^{(i-1)}(t_0)$, analogously for $\widetilde{v}_i$, and $\sum_{i=1}^{2n-r+1} w_i(t-t_0)^{i-1} = \mathbf{u}(t)\mathbf{v}(t)(t-t_0)^{-r}$. We arrive at the following.

Proposition 9.7. *Let $t_1, \ldots, t_m$ be the (different) roots of $\mathbf{u}(t)$ and $r_1, \ldots, r_m$ the corresponding multiplicities, $\mathbf{t} = (t_i)_{i=1}^m$ and $\mathbf{r} = (r_i)_{i=1}^m$. Then*

$$V_n(\mathbf{t},\mathbf{r})\mathrm{Bez}_H(\mathbf{u},\mathbf{v})V_n(\mathbf{t},\mathbf{r})^T = \mathrm{diag}(\Gamma_i)_{i=1}^m\,,$$

where

$$\Gamma_i = \begin{bmatrix} & & w_{i1} \\ & \iddots & \vdots \\ w_{i1} & \ldots & w_{ir_i} \end{bmatrix} \quad \textit{and} \quad \sum_{j=1}^{2n-r_i+1} w_{ij}(t-t_i)^{j-1} = \mathbf{u}(t)\mathbf{v}(t)(t-t_i)^{-r_i}\,.$$

The case $\mathbf{v}(t) = 1$ provides a formula for the inverse of confluent Vandermonde matrices.

Corollary 9.8. *Let $\mathbf{t} = (t_1, \ldots, t_q)$ have distinct components, and let*

$$\mathbf{u}(t) = \prod_{i=1}^{q}(t-t_i)^{r_i},$$

where $n = r_1 + \cdots + r_q$. Then the inverse of the confluent Vandermonde matrix $V_n(\mathbf{t},\mathbf{r})$, $\mathbf{r} = (r_, \ldots, r_q)$, is given by

$$V_n(\mathbf{t},\mathbf{r})^{-1} = B(\mathbf{u})V_n(\mathbf{t},\mathbf{r})^T\mathrm{diag}\,(\Gamma_i^{-1})_{i=1}^m\,,$$

where $\Gamma_i = \begin{bmatrix} & & w_{i1} \\ & \iddots & \vdots \\ w_{i1} & \dots & w_{ir_i} \end{bmatrix}$ *with* $w_{ij} = \frac{\mathbf{u}^{(j+r_i)}(t_i)}{(j+r_i-1)!}$.

We leave it to the reader to state the analogous properties of T-Bezoutians.

10. Root localization problems

In this section we show the importance of Bezoutians, Hankel and Toeplitz matrices for root localization problems. Throughout the section, let $\mathbb{F} = \mathbb{C}$.

1. Inertia of polynomials. Let $\mathcal{C}$ be a simple oriented closed curve in the extended complex plane $\mathbb{C}^\infty = \mathbb{C} \cup \{\infty\}$ dividing it into an "interior" part Ω_+ and an "exterior" part Ω_-. We assume that the domain Ω_+ is situated left from $\mathcal{C}$ if a point moves along $\mathcal{C}$ in positive direction. The *inertia of the polynomial* $\mathbf{u}(t) \in \mathbb{C}^{n+1}(t)$ *with respect to* $\mathcal{C}$ is, by definition, a triple of nonnegative integers

$$\mathrm{in}_{\mathcal{C}}(\mathbf{u}) = (\pi_+(\mathbf{u}), \pi_-(\mathbf{u}), \pi_0(\mathbf{u}))\,,$$

where $\pi_\pm(\mathbf{u})$ is the number of zeros of $\mathbf{u}(t)$ in $\Omega_\pm$, respectively, and $\pi_0(\mathbf{u})$ is the number of zeros on $\mathcal{C}$. In all cases multiplicities are counted. We say that $\mathbf{u}(t) \in \mathbb{C}^{n+1}(t)$ has a *zero at* ∞ *with multiplicity* r if the r leading coefficients of $\mathbf{u}(t)$ are zero. By a *root localization problem* we mean the problem to find the inertia of a given polynomial with respect to a curve $\mathcal{C}$.

In the sequel we deal only with the cases that $\mathcal{C}$ is the real line $\mathbb{R}$, the imaginary line $\mathrm{i}\,\mathbb{R}$, or the unit circle $\mathbb{T}$. We relate the inertia of polynomials to inertias of Hermitian matrices, namely to Bezoutians. Recall that the inertia of an Hermitian matrix A is the triple

$$\mathrm{In}\, A = (p_+(A), p_-(A), p_0(A))\,,$$

where $p_+(A)$ is the number of positive, $p_-(A)$ the number of negative eigenvalues (counting multiplicities), and $p_0(A)$ the nullity of A. Clearly, the inertia of A is completely determined by the rank and the signature of A. The importance of the relation consists in the fact that the inertia of Bezoutians can be computed via recursive algorithms for triangular factorization, which were described in Section 5.

2. Inertia with respect to the real line. Let $\mathbf{u}(t)$ be a given monic polynomial of degree n, $\mathbf{q}(t)$ its real and $\mathbf{p}(t)$ its imaginary part, i.e., $\mathbf{u}(t) = \mathbf{q}(t) + \mathrm{i}\,\mathbf{p}(t)$. We consider the matrix

$$B = \frac{1}{2\mathrm{i}}\,\mathrm{Bez}_H(\mathbf{u}, \overline{\mathbf{u}}). \tag{10.1}$$

For example, if $\mathbf{u}(t) = t - c$, then $B = \frac{1}{2\mathrm{i}}(c - \overline{c}) = \mathrm{Im}\, c$. Hence c is in the upper half-plane if and only if $B > 0$. In general, we have

$$\begin{aligned}(t-s)B(t,s) &= \left[\frac{1}{2\mathrm{i}}(\mathbf{q}(t)+\mathrm{i}\mathbf{p}(t))(\mathbf{q}(s)-\mathrm{i}\mathbf{p}(s)) - (\mathbf{q}(t)-\mathrm{i}\mathbf{p}(t))(\mathbf{q}(s)+\mathrm{i}\mathbf{p}(s))\right] \\ &= \mathbf{p}(t)\mathbf{q}(s) - \mathbf{q}(t)\mathbf{p}(s)\,.\end{aligned}$$

Hence,

$$B = \mathrm{Bez}_H(\mathbf{p}, \mathbf{q})\,. \tag{10.2}$$

In particular, we see that B is a real symmetric matrix. The following is usually referred to as *Hermite's theorem*.

Theorem 10.1. *Let* $\mathbf{u}(t)$ *be a monic polynomial of degree* n,

$$\mathrm{in}_{\mathbb{R}}(\mathbf{u}) = (\pi_+(\mathbf{u}), \pi_-(\mathbf{u}), \pi_0(\mathbf{u})),$$

and B *be defined by* (10.1) *or* (10.2). *Then the signature of* B *is given by*

$$\operatorname{sgn} B = \pi_+(\mathbf{u}) - \pi_-(\mathbf{u})\,.$$

In particular, B *is positive definite if and only if* $\mathbf{u}(t)$ *has all its roots in the upper half-plane. Furthermore, if* $\mathbf{u}(t)$ *and* $\overline{\mathbf{u}}(t)$ *are coprime, then* $\operatorname{In} B = \mathrm{in}_{\mathbb{R}}(\mathbf{u})$.

Proof. Let $\mathbf{d}(t)$ be the greatest common divisor of $\mathbf{u}(t)$ and $\overline{\mathbf{u}}(t)$ and δ its degree. Then $\mathbf{d}(t)$ is also the greatest common divisor of $\mathbf{p}(t)$ and $\mathbf{q}(t)$, and let $\mathbf{u}(t) = \mathbf{d}(t)\mathbf{u}_0(t)$. Then $\overline{\mathbf{u}}(t) = \mathbf{d}(t)\overline{\mathbf{u}}_0(t)$, since $\mathbf{d}(t)$ is real. According to (2.5) we have

$$B = \operatorname{Res}(\mathbf{d}, \mathbf{u}_0)^* \begin{bmatrix} B_0 & O \\ O & O \end{bmatrix} \operatorname{Res}(\mathbf{d}, \mathbf{u}_0)\,,$$

where $B_0 = \frac{1}{2\mathrm{i}} \mathrm{Bez}_H(\mathbf{u}_0, \overline{\mathbf{u}}_0)$. Since $\mathbf{d}(t)$ and $\mathbf{u}_0(t)$ are coprime, $\operatorname{Res}(\mathbf{d}, \mathbf{u}_0)$ is non-singular. By Sylvester's inertia law we have $\operatorname{sgn} B = \operatorname{sgn} B_0$. We find now $\operatorname{sgn} B_0$.

Let z_1 be a root of $\mathbf{u}_0(t)$ and $\mathbf{u}_0(t) = (t - z_1)\mathbf{u}_1(t)$. Then $\overline{\mathbf{u}}_0(t) = (t - \overline{z}_1)\overline{\mathbf{u}}_1(t)$. Taking into account that

$$\frac{1}{2\mathrm{i}} \frac{(t - z_1)(s - \overline{z}_1) - (t - \overline{z}_1)(s - z_1)}{t - s} = \operatorname{Im} z_1$$

we obtain using (2.5)

$$B_0 = \operatorname{Res}(t - \overline{z}_1, \mathbf{u}_1)^* \begin{bmatrix} B_1 & \mathbf{0} \\ \mathbf{0}^T & \operatorname{Im} z_1 \end{bmatrix} \operatorname{Res}(t - \overline{z}_1, \mathbf{u}_1)\,,$$

where $B_1 = \frac{1}{2\mathrm{i}} \mathrm{Bez}_H(\mathbf{u}_1, \overline{\mathbf{u}}_1)$. Repeating these arguments for the other roots z_k of $\mathbf{u}_0(t)$ $(k = 2, \ldots, n - \delta)$ we conclude that there is a matrix R such that

$$B_0 = R^* \operatorname{diag}(\operatorname{Im} z_1, \ldots, \operatorname{Im} z_{n-\delta})\, R\,.$$

Thus, B_0 is congruent to the diagonal matrix of the $\operatorname{Im} z_i$. Applying Sylvester's inertia law we obtain

$$\operatorname{sgn} B = \operatorname{sgn} B_0 = \sum_{i=1}^{n-\delta} \operatorname{sgn}(\operatorname{Im} z_i) = \pi_+(\mathbf{u}) - \pi_-(\mathbf{u})\,,$$

which proves the main part of the theorem.

If $\mathbf{u}(t)$ and $\overline{\mathbf{u}}(t)$ are coprime, then B is nonsingular, thus $p_0(B) = 0$, and $\pi_0(\mathbf{u}) = 0$. Hence $\pi_+(\mathbf{u}) + \pi_-(\mathbf{u}) = n$. Consequently $\pi_\pm(\mathbf{u}) = p_\pm(B)$, and the theorem is proved. □

If $\mathbf{u}(t)$ is a real polynomial, then the Bezoutian B is zero, so all information about the polynomial is lost. It is remarkable that in the other cases information about the polynomial is still contained in B.

Example 10.2. *Let* $\mathbf{u}(t) = (t - z_0)\mathbf{d}(t)$*, where* $\mathbf{d}(t)$ *is real and* z_0 *is in the upper half-plane. Then* $B = \operatorname{Im} z_0\, \mathbf{d}\mathbf{d}^*$*. This matrix has signature 1 saying that* $\pi_+(\mathbf{u}) - \pi_-(\mathbf{u}) = 1$ *but not specifying the location of the roots of* $\mathbf{d}(t)$*. Nevertheless, the polynomial* $\mathbf{d}(t)$ *and so information about* $\mathbf{u}(t)$ *can be recovered from* B*.*

Recall that according to the results in Section 5 the Euclidian algorithm applied to the polynomials $\mathbf{p}(t)$ and $\mathbf{q}(t)$ provides a method to compute the signature of B in $O(n^2)$ operations.

We know from Proposition 7.18 that the Bezoutian $B = \operatorname{Bez}_H(\mathbf{p}, \mathbf{q})$ and the Hankel matrix $H_n = H_n\left(-\frac{\mathbf{p}}{\mathbf{q}}\right)$ have the same inertia. Hence we can conclude the following.

Corollary 10.3. *Let* $\mathbf{p}(t)\,,\ \mathbf{q}(t) \in \mathbb{R}^{n+1}(t)$ *be two coprime polynomials, where* $\mathbf{q}(t)$ *is monic with degree* n *and* $\mathbf{u}(t) = \mathbf{q}(t) + \mathrm{i}\mathbf{p}(t)$*. Then*

$$\operatorname{in}_{\mathbb{R}}(\mathbf{u}) = \operatorname{In} H_n\left(-\frac{\mathbf{p}}{\mathbf{q}}\right).$$

Theorem 10.1, gives a complete answer to the problem to find the inertia of a polynomial only if $\mathbf{u}(t)$ has no real and conjugate complex roots. In the other cases we have only partial information. More precisely, if δ denotes the number of real roots of $\mathbf{u}(t)$, then

$$\pi_\pm(\mathbf{u}) = p_\pm(B) + \frac{1}{2}(p_0(B) - \delta).$$

Note that $\frac{1}{2}(p_0(B) - \delta)$ is the number of conjugate complex pairs of roots of $\mathbf{u}(t)$. The number δ is also the number of real roots of the greatest common divisor $\mathbf{d}(t)$ of $\mathbf{u}(t)$ and $\overline{\mathbf{u}}(t)$. Clearly, $\mathbf{d}(t)$ is a real polynomial. Thus we are led to the problem to count the number of real roots of a real polynomial. This problem will be discussed next.

3. Real roots of real polynomials. Let $\mathbf{p}(t)$ be a real polynomial of degree n. We consider the rational function $\frac{\mathbf{p}'(t)}{\mathbf{p}(t)}$ and the Hankel matrix $H_n\left(\frac{\mathbf{p}'(t)}{\mathbf{p}(t)}\right)$. This is just our Example 7.2. According to Proposition 7.17 the rank of this matrix is equal to the number of *different* roots of $\mathbf{p}(t)$ and, due to Corollaries 7.11 and 1.2, the signature is the number of *different real* roots of $\mathbf{p}(t)$. Let $\pi_0'(\mathbf{p})$ denote the number of different real roots of $\mathbf{p}(t)$. Taking also into account Proposition 7.18, we have now the following *Theorem of Jacobi-Borchardt.*

Theorem 10.4. *The number of different real roots $\pi_0'(\mathbf{p})$ of the real polynomial $\mathbf{p}$ is given by*

$$\pi_0'(\mathbf{p}) = \operatorname{sgn} \operatorname{Bez}_H(\mathbf{p}, \mathbf{p}') = \operatorname{sgn} H_n\left(\frac{\mathbf{p}'(t)}{\mathbf{p}(t)}\right).$$

Example 10.5. *Let* $\mathbf{p}(t) = t^4 - 1$. *Then*

$$H_4\left(\frac{\mathbf{p}'(t)}{\mathbf{p}(t)}\right) = 4\begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 \\ 0 & 0 & 1 & 0 \\ 0 & 1 & 0 & 0 \end{bmatrix}, \quad \operatorname{Bez}_H(\mathbf{p}, \mathbf{p}') = 4\begin{bmatrix} 0 & 0 & 1 & 0 \\ 0 & 1 & 0 & 0 \\ 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix}.$$

Both matrices have signature 2, *which confirms Theorem* 10.4.

Theorem 10.4 completely solves the root localization problem for real polynomials in case of simple roots. The multiple roots are just the roots of the greatest common divisor $\mathbf{p}_1(t)$ of $\mathbf{p}(t)$ and $\mathbf{p}'(t)$. If we find the numbers of different real roots of $\mathbf{p}_1(t)$ by Theorem 10.4, then we can obtain the number of real roots with multiplicity of at least two. If we continue we obtain the number of different real roots with multiplicity of at least 3, and so on.

From the algorithmic view point we have to apply the Euclidian algorithm again and again. If it terminates at a certain polynomial $\mathbf{d}(t)$, then we continue with $\mathbf{d}(t)$ and $\mathbf{d}'(t)$ until the remainder is a constant. It is remarkable that all together there are not more than n steps.

This also concerns the general root localization problem for a complex polynomial $\mathbf{u}(t)$ which was discussed above in Section 10.2. First we apply the Euclidian algorithm to the real and imaginary parts of $\mathbf{u}(t)$, and when it terminates at a non-constant $\mathbf{d}(t)$, then we continue with the Euclidian algorithm for $\mathbf{d}(t)$ and $\mathbf{d}'(t)$. Again we have at most n steps if the degree of the original polynomial is n.

4. Inertia with respect to the imaginary axis. To find the inertia of a *real* polynomial with respect to the imaginary axis $\mathrm{i}\mathbb{R}$ is a very important task in many applications, because it is related to the question of stability of systems. In order to avoid confusion, let us point out that if $\operatorname{in}_{\mathrm{i}\mathbb{R}}(\mathbf{u}) = (\pi_+(\mathbf{u}), \pi_-(\mathbf{u}), \pi_0(\mathbf{u}))$, then according to the definition above $\pi_+(\mathbf{u})$ is the number of roots of $\mathbf{u}(t)$ with *negative* real part and $\pi_-(\mathbf{u})$ those with *positive* real part.

Clearly, the imaginary axis case can easily be transformed into the real line case by a transformation of the variable. It remains to study the specifics that arises from the fact that the polynomial under investigation is real. Suppose that $\mathbf{p}(t)$ is a real polynomial of degree n. We set $\mathbf{u}(t) = \mathbf{p}(\mathrm{i}t)$. Then

$$\operatorname{in}_{\mathrm{i}\mathbb{R}}(\mathbf{p}) = \operatorname{in}_{\mathbb{R}}(\mathbf{u}).$$

Furthermore, if $\mathbf{p}(t) = \sum_{j=1}^{n+1} p_j t^{j-1}$, then $\mathbf{u}(t)$ admits a representation

$$\mathbf{u}(t) = \mathbf{a}(t^2) + \mathrm{i}\, t\, \mathbf{b}(t^2),$$

where, for odd $n = 2m + 1$,

$$\mathbf{a}(t) = p_1 - p_3 t + \cdots + (-1)^m p_{2m+1} t^m, \quad \mathbf{b}(t) = p_2 - p_4 t + \cdots + (-1)^m p_{2m+2} t^m, \tag{10.3}$$

and for even $n = 2m$,

$$\mathbf{a}(t) = p_1 - p_3 t + \cdots + (-1)^m p_{2m+1} t^m, \quad \mathbf{b}(t) = p_2 - p_4 t + \cdots + (-1)^{m-1} p_{2m} t^{m-1}. \tag{10.4}$$

From (10.2) and Proposition 2.8 we conclude that the matrix $B = \frac{1}{2\mathrm{i}} \operatorname{Bez}_H(\mathbf{u}, \overline{\mathbf{u}})$ is congruent to the direct sum of the matrices

$$B_0 = \operatorname{Bez}_H(t\mathbf{b}, \mathbf{a}) \quad \text{and} \quad B_1 = \operatorname{Bez}_H(\mathbf{b}, \mathbf{a}). \tag{10.5}$$

Using Theorem 10.1 we arrive at the following.

Theorem 10.6. *Let $\mathbf{p}(t)$ be a real polynomial of degree n and let $\mathbf{a}(t)$ and $\mathbf{b}(t)$ be defined by* (10.3) *or* (10.4)*, depending on whether n is odd or even, and let B_0 and B_1 be given by* (10.5)*. Then*

$$\operatorname{sgn} B_0 + \operatorname{sgn} B_1 = \pi_+(\mathbf{p}) - \pi_-(\mathbf{p}),$$

where $\pi_+(\mathbf{p})$ denotes the number of roots of $\mathbf{p}(t)$ in the left and $\pi_-(\mathbf{p})$ the number of roots in the right half-plane. In particular, the roots of $\mathbf{p}(t)$ lie entirely in the left half-plane if and only both matrices B_0 and B_1 are positive definite. Furthermore, if $\mathbf{a}(t)$ and $\mathbf{b}(t)$ are coprime, then

$$\operatorname{in}_{\mathrm{i}\mathbb{R}}(\mathbf{p}) = \operatorname{In} B_0 + \operatorname{In} B_1.$$

5. Roots on the imaginary axis and positive real roots of real polynomials. In order to get a full picture about the location of the roots of the real polynomial $\mathbf{p}(t)$ with respect to the imaginary axis we have to find the number π_0 of all roots on the imaginary axis, counting multiplicities. As a first step we find the number π_0' of different roots on $\mathrm{i}\mathbb{R}$. This number is equal to the number of different real roots of the greatest common divisor of $\mathbf{a}(t^2)$ and $t\mathbf{b}(t^2)$, where $\mathbf{a}(t)$ and $\mathbf{b}(t)$ are defined by (10.3) or (10.4). Let $\mathbf{p}(0) \neq 0$ and $\mathbf{d}(t)$ be the greatest common divisor of $\mathbf{a}(t)$ and $\mathbf{b}(t)$. Then the greatest common divisor of $\mathbf{a}(t^2)$ and $t\mathbf{b}(t^2)$ equals $\mathbf{d}(t^2)$. The number of real roots of the polynomial $\mathbf{d}(t^2)$ equals twice the number of positive real roots of $\mathbf{d}(t)$. Thus, we are led to the problem to count the roots of a real polynomial on the positive real half-axis.

Let $\mathbf{p}(t)$ be a real polynomial of degree n and $\mathbf{p}(0) \neq 0$. We consider the function $\mathbf{f}_1(t) = \frac{t\mathbf{p}'(t)}{\mathbf{p}(t)}$. This function has a partial fraction decomposition

$$\mathbf{f}_1(t) = \sum_{i=1}^{n} \frac{t}{t - t_i} = n + \sum_{i=1}^{n} \frac{t_i}{t - t_i},$$

where $t_1, \ldots, t_n$ are the roots of $\mathbf{p}(t)$. This leads to a Vandermonde factorization similar to the factorization (7.10) in Example 7.9. From this factorization we conclude that

$$\operatorname{sgn} H_n(\mathbf{f}_1) = \delta_+ - \delta_-,$$

where δ_+ denotes the number of different positive and δ_- the number of different negative real roots of $\mathbf{p}(t)$. Using Proposition 7.18 and the result of Theorem 10.4, which is $\operatorname{sgn}\operatorname{Bez}_H(\mathbf{p},\mathbf{p}') = \delta_+ + \delta_-$, we conclude the following.

Theorem 10.7. *Let* $\mathbf{p}(t)$ *be as in Theorem* 10.6 *with* $\mathbf{p}(0) \neq 0$. *The number of positive real roots* δ_+ *of* $\mathbf{p}(t)$ *is given by*

$$\delta_+ = \frac{1}{2}(\operatorname{sgn}\operatorname{Bez}_H(\mathbf{p},\mathbf{p}') + \operatorname{sgn}\operatorname{Bez}_H(\mathbf{p},t\mathbf{p}'))\,.$$

Furthermore, all roots of $\mathbf{p}(t)$ *are real and positive if and only if* $\operatorname{Bez}_H(\mathbf{p},t\mathbf{p}')$ *is positive definite.*

Example 10.8. *For* $\mathbf{p}(t) = t^4 - 1$ *we obtain*

$$\operatorname{Bez}_H(\mathbf{p},t\mathbf{p}') = 4\,J_4\,.$$

This matrix has a signature equal to zero. Thus (*cf. Example* 10.5),

$$\frac{1}{2}(\operatorname{sgn}\operatorname{Bez}_H(\mathbf{p},\mathbf{p}') + \operatorname{sgn}\operatorname{Bez}_H(\mathbf{p},t\mathbf{p}')) = 1$$

which confirms Theorem 10.7.

6. Inertia with respect to the unit circle. Now we discuss the problem how to find the inertia $\operatorname{in}_{\mathbb{T}}(\mathbf{u}) = (\pi_+(\mathbf{u}),\pi_-(\mathbf{u}),\pi_0(\mathbf{u}))$ of a complex monic polynomial $\mathbf{u}(t)$ of degree n with respect to the unit circle $\mathbb{T}$. According to the definition in Section 10.1, $\pi_+(\mathbf{u})$ is the number of roots inside the unit circle, $\pi_-(\mathbf{u})$ is the number of roots outside the unit circle, and $\pi_0(\mathbf{u})$ the number of roots on the unit circle. We consider the matrix

$$B = \operatorname{Bez}_T(\mathbf{u}^{\#},\mathbf{u})\,. \tag{10.6}$$

For example, if $n = 1$ and $\mathbf{u}(t) = t - c$, then $B = 1 - |c|^2$. Thus c belongs to the open unit disk if and only if $B > 0$. A general $\mathbf{u}(t)$ can be represented as $\mathbf{u}(t) = \mathbf{u}_+(t) + \mathrm{i}\mathbf{u}_-(t)$, where $\mathbf{u}_\pm$ are conjugate symmetric. We have now that the polynomial $(1-ts)B(t,s)$ is equal to

$$(\mathbf{u}_+(t) - \mathrm{i}\mathbf{u}_-(t))(\overline{\mathbf{u}}_+(s) + \mathrm{i}\overline{\mathbf{u}}_-(s)) - ((\mathbf{u}_+(t) + \mathrm{i}\mathbf{u}_-(t))(\overline{\mathbf{u}}_+(s) - \mathrm{i}\overline{\mathbf{u}}_-(s))).$$

Thus $(1-ts)B(t,s) = 2\mathrm{i}(\mathbf{u}_+(t)\overline{\mathbf{u}}_-(s) - \mathbf{u}_-(t)\overline{\mathbf{u}}_+(s))$, which means

$$B = 2\mathrm{i}\operatorname{Bez}_T(\mathbf{u}_+,\mathbf{u}_-)\,. \tag{10.7}$$

The following is usually referred to as the *Schur-Cohn theorem*. It can be proved with the same arguments as Hermite's theorem (Theorem 10.1).

Theorem 10.9. *Let* $\mathbf{u}(t)$ *be a monic polynomial of degree* n *and* B *be defined by* (10.6) *or* (10.7). *Then the signature of* B *is given by*

$$\operatorname{sgn} B = \pi_+(\mathbf{u}) - \pi_-(\mathbf{u})\,.$$

In particular, B *is positive definite if and only if* $\mathbf{u}(t)$ *has all its roots in the open unit disk. Furthermore, if* $\mathbf{u}(t)$ *and* $\mathbf{u}^{\#}(t)$ *are coprime, then* $\operatorname{In} B = \operatorname{in}_{\mathbb{T}}(\mathbf{u})$.

Theorem 10.9 provides full information about the inertia of $\mathbf{u}(t)$ only if $\mathbf{u}(t)$ has no roots on the unit circle and symmetric with respect to the unit circle or, what is the same if $\mathbf{u}(t)$ and $\mathbf{u}^{\#}(t)$ are coprime. To complete the picture we still have to find the inertia of the greatest common divisor of $\mathbf{u}(t)$ and $\mathbf{u}^{\#}(t)$, which is a conjugate-symmetric polynomial.

7. Roots of conjugate-symmetric polynomials. Let $\mathbf{w}(t)$ be a monic conjugate-symmetric polynomial of degree n and $t_1, \dots, t_n$ its roots. Then

$$\mathbf{w}(t) = \prod_{k=1}^{n}(t - t_k) = \mathbf{w}^{\#}(t) = \prod_{k=1}^{n}(1 - \overline{t}_k t)\,,$$

which implies $\overline{t}_k^{-1} = t_k\,,\ k = 1, \dots, n\,.$ We consider the function $\mathbf{f}(t) = \frac{(\mathbf{w}')^{\#}(t)}{\mathbf{w}^{\#}(t)}$, where $(\mathbf{w}')^{\#} = J_n\overline{\mathbf{w}'}$ This function has a partial fraction decomposition (cf. Example 8.2)

$$\mathbf{f}(t) = \sum_{k=1}^{n}\frac{1}{1 - \overline{t}_k t}\,.$$

From this representation we can see that the Toeplitz matrix $T_n(\mathbf{f})$ is Hermitian and its signature is equal to the number of different roots of $\mathbf{w}(t)$ on $\mathbb{T}$ (cf. Corollary 8.7, (8.10), and Corollary 1.2).

Theorem 10.10. *For a conjugate-symmetric polynomial* $\mathbf{w}(t)$, $\mathrm{Bez}_T((\mathbf{w}')^{\#}, \mathbf{w}^{\#})$ *is Hermitian, and its signature is equal to the number of different roots of* $\mathbf{w}(t)$ *on the unit circle.*

11. Toeplitz-plus-Hankel Bezoutians

Some important results for the Toeplitz and Hankel case can be generalized to matrices which are the sum of such structured matrices. In particular, we will show that the inverse of a (nonsingular) matrix which is the sum of a Toeplitz plus a Hankel matrix possesses again a (generalized) Bezoutian structure. To be more precise we define the following.

1. Definition. An $n \times n$ matrix B is called *Toeplitz-plus-Hankel Bezoutian*, briefly *T+H-Bezoutian,* if there are eight polynomials $\mathbf{g}_i(t)$, $\mathbf{f}_i(t)$ $(i = 1, 2, 3, 4)$ of $\mathbb{F}^{n+2}(t)$ such that

$$B(t, s) = \frac{\sum_{i=1}^{4}\mathbf{g}_i(t)\mathbf{f}_i(s)}{(t - s)(1 - ts)}\,. \tag{11.1}$$

In analogy to the Hankel or Toeplitz case we use here the notation

$$B = \mathrm{Bez}_{T+H}((\mathbf{g}_i, \mathbf{f}_i)_1^4).$$

H-Bezoutians or T-Bezoutians are also T+H-Bezoutians. For example, the flip matrix J_n introduced in (1.2) is an H-Bezoutian, $J_n(t,s)$ can be written as

$$J_n(t,s) = \frac{t^n - s^n}{t-s} = \frac{t^n - s^n - t^{n+1}s + ts^{n+1}}{(t-s)(1-ts)}\,,$$

which shows that J_n is the T+H-Bezoutian (11.1), where

$$\mathbf{g}_1 = -\mathbf{f}_2 = \mathbf{e}_{n+1},\ \mathbf{g}_2 = \mathbf{f}_1 = 1,\ \mathbf{g}_3 = \mathbf{f}_4 = \mathbf{e}_{n+2},\ \mathbf{g}_4 = -\mathbf{f}_3 = \mathbf{e}_2\,.$$

The shift matrix (1.8) is a T-Bezoutian and a T+H-Bezoutian,

$$S_n(t,s) = \frac{t - t^n s^{n-1}}{1-ts} = \frac{t^2 - t^{n+1}s^{n-1} - ts + t^n s^n}{(t-s)(1-ts)}\,.$$

For these examples the sum $S_n + J_n$ is also a T+H-Bezoutian,

$$(S_n + J_n)(t,s) = \frac{(t^n + t^2) - t^{n+1}(s + s^{n-1}) + (t^n - 1)s^n + t(s^{n+1} - s)}{(t-s)(1-ts)}\,.$$

But for any vectors $\mathbf{u}, \mathbf{v}, \mathbf{g}, \mathbf{h} \in \mathbb{F}^{n+1}, n > 3$, the rank of the matrix with the generating polynomial

$$(1-ts)(\mathbf{u}(t)\mathbf{v}(s) - \mathbf{v}(t)\mathbf{u}(s)) + (t-s)(\mathbf{g}(t)\mathbf{h}^J(s) - \mathbf{h}(t)\mathbf{g}^J(s))$$

is not expected to be less or equal to 4. This means that the sum of a T- and an H-Bezoutian $\mathrm{Bez}_H(\mathbf{u},\mathbf{v}) + \mathrm{Bez}_T(\mathbf{g},\mathbf{f})$ is, in general, not a T+H-Bezoutian.

2. The transformation $\boldsymbol{\nabla_{T+H}}$. The T+H analogue of the transformations ∇_H or ∇_T (introduced in Section 2.2 and in Section 2.8) is the transformation ∇_{T+H} mapping a matrix $A = [\,a_{ij}\,]_{i,j=1}^n$ of order n onto a matrix of order $n+2$ according to

$$A = [\,a_{i-1,j} - a_{i,j-1} + a_{i-1,j-2} - a_{i-2,j-1}\,]_{i,j=1}^{n+2}\,.$$

Here we put $a_{ij} = 0$ if $i \notin \{1,2,\dots,n\}$ or $j \notin \{1,2,\dots,n\}$. Denoting $W_n = S_n + S_n^T$ we have

$$\nabla_{T+H}A = \begin{bmatrix} 0 & -\mathbf{e}_1^T A & 0 \\ A\mathbf{e}_1 & AW_n - W_nA & A\mathbf{e}_n \\ 0 & -\mathbf{e}_n^T A & 0 \end{bmatrix}\,. \tag{11.2}$$

The generating polynomial of $\nabla_{T+H}A$ is

$$(\nabla_{T+H}A)(t,s) = (t-s)(1-ts)A(t,s)\,. \tag{11.3}$$

Hence a matrix B is a T+H-Bezoutian if and only if

$$\mathrm{rank}\nabla_{T+H}B \le 4\,.$$

Recall that the $n \times n$ matrix in the center of (11.2) is the matrix $\nabla(A)$ introduced in (1.11). In other words, the transformation ∇ is a restriction of $\nabla_{\mathrm{T+H}}$, and it is clear that T+H-Bezoutians are quasi-T+H matrices, but not vice versa.

3. Uniqueness. Different vector systems $\{\mathbf{g}_i, \mathbf{f}_i\}_{i=1}^4$, $\{\widetilde{\mathbf{g}}_i, \widetilde{\mathbf{f}}_i\}_{i=1}^4$ may produce the same T+H-Bezoutian.

Note that $B = \mathrm{Bez}_{T+H}((\mathbf{g}_i, \mathbf{f}_i)_1^4)$ is equal to $\widetilde{B} = \mathrm{Bez}_{T+H}((\widetilde{\mathbf{g}}_i, \widetilde{\mathbf{f}}_i)_1^4)$ if and only if $\nabla_{T+H} B = \nabla_{T+H} \widetilde{B}$. To answer the questions under which conditions this happens we use the following lemma.

Lemma 11.1. *Let $G_j, F_j\,(j = 1, 2)$ be full rank matrices of order $m \times r, n \times r$, respectively, $r = \operatorname{rank} G_j = \operatorname{rank} F_j$. Then*

$$G_1 F_1^T = G_2 F_2^T \tag{11.4}$$

if and only if there is a nonsingular $r \times r$ matrix φ such that

$$G_2 = G_1 \varphi\,,\ F_1 = F_2 \varphi^T\,. \tag{11.5}$$

Proof. Assume there is a nonsingular φ so that $G_2 = G_1\varphi$ and $F_2^T = \varphi^{-1} F_1^T$, then $G_1 F_1^T = G_2 F_2^T$. Now let (11.4) be satisfied and $A = G_1 F_1^T$. The image of A is spanned by the columns of G_1 as well as of G_2. Thus there exists a nonsingular matrix φ so that $G_2 = G_1\varphi$. With the same arguments for A^T we obtain that there is a nonsingular matrix ψ so that $F_2 = F_1\psi$. Hence

$$G_1 F_1^T = G_2 F_2^T = G_1 \varphi \psi^T F_1^T\,. \tag{11.6}$$

Since G_1, F_1 are of full rank they are one-sided invertible, and we conclude from (11.6) that $\varphi \cdot \psi^T = I_r$. □

Let $B, \widetilde{B}$ be $n \times n$ T+H-Bezoutians and $G, \widetilde{G}, F, \widetilde{F}$ be full rank matrices with

$$r = \operatorname{rank} G = \operatorname{rank} F \le 4\,,\ \widetilde{r} = \operatorname{rank} \widetilde{G} = \operatorname{rank} \widetilde{F} \le 4$$

such that the matrices $\nabla_{\mathrm{T+H}} B$ and $\nabla_{\mathrm{T+H}} \widetilde{B}$ allow the following rank decompositions

$$\nabla_{\mathrm{T+H}} B = G F^T\,,\ \nabla_{\mathrm{T+H}} \widetilde{B} = \widetilde{G} \widetilde{F}^T\,.$$

Proposition 11.2. *The T+H-Bezoutians B and $\widetilde{B}$ coincide if and only if $r = \widetilde{r}$, and there is a nonsingular $r \times r$ matrix φ so that*

$$\widetilde{G} = G\varphi\,,\ F = \widetilde{F}\varphi^T\,.$$

To specify this for nonsingular Bezoutians we make the following observation.

Proposition 11.3. *Let B be an $n \times n$ matrix ($n \ge 2$) with rank $\nabla_{\mathrm{T+H}} B < 4$. Then B is a singular matrix.*

Proof. Let us prove this by contradiction. Assume B is nonsingular and $\nabla_{\mathrm{T+H}} B < 4$. Taking (11.2) into account elementary considerations show that $\nabla_{\mathrm{T+H}} B$ allows the following decomposition

$$\nabla_{\mathrm{T+H}} B = \begin{bmatrix} 0 \\ B\mathbf{e}_1 \\ 0 \end{bmatrix} [\,1\ *\ 0\,] + \begin{bmatrix} 0 \\ B\mathbf{e}_n \\ 0 \end{bmatrix} [\,0\ *\ 1\,] - \begin{bmatrix} 1 \\ * \\ 0 \end{bmatrix} [\,0\ \mathbf{e}_1^T B\ 0\,] - \begin{bmatrix} 0 \\ * \\ 1 \end{bmatrix} [\,0\ \mathbf{e}_n^T B\ 0\,], \tag{11.7}$$

where $*$ stands for some vector of $\mathbb{F}^n$. Due to the nonsingularity of B its first and last rows as well as its first and last columns are linearly independent. Thus,

$$\operatorname{rank}\begin{bmatrix} 0 & 0 & 1 & 0 \\ B\mathbf{e}_1 & B\mathbf{e}_n & * & * \\ 0 & 0 & 0 & 1 \end{bmatrix} = \operatorname{rank}\begin{bmatrix} 1 & * & 0 \\ 0 & * & 1 \\ 0 & \mathbf{e}_1^T B & 0 \\ 0 & \mathbf{e}_n^T B & 0 \end{bmatrix} = 4\,,$$

which contradicts rank $\nabla_{\mathrm{T+H}} B < 4$. □

Corollary 11.4. *If rank $\nabla_{\mathrm{T+H}} B < 4$ then the first and the last rows (or the first and the last columns) of B are linearly dependent.*

For T-(or H-)Bezoutians B, the condition rank $\nabla_T B < 2$ (or rank $\nabla_H B < 2$) leads to $B \equiv 0$. But in the T+H case nontrivial T+H-Bezoutians B with rank $\nabla_{\mathrm{T+H}} B < 4$ exist. Examples are $B = I_n + J_n$ $(n \geq 2)$ and split Bezoutians introduced in Section 2.11. In these cases rank $\nabla B \leq 2$. Now we present the result for the nonsingular case.

Proposition 11.5. *The nonsingular T+H-Bezoutians*

$$B = \mathrm{Bez}_{\mathrm{T+H}}((\mathbf{g}_i, \mathbf{f}_i)_1^4) \text{ and } \widetilde{B} = \mathrm{Bez}_{\mathrm{T+H}}((\widetilde{\mathbf{g}}_i, \widetilde{\mathbf{f}}_i)_1^4)$$

coincide if and only if there is a nonsingular 4×4 matrix φ such that

$$[\,\mathbf{g}_1\ \mathbf{g}_2\ \mathbf{g}_3\ \mathbf{g}_4\,]\,\varphi = [\,\widetilde{\mathbf{g}}_1\ \widetilde{\mathbf{g}}_2\ \widetilde{\mathbf{g}}_3\ \widetilde{\mathbf{g}}_4\,]$$

and

$$[\,\widetilde{\mathbf{f}}_1\ \widetilde{\mathbf{f}}_2\ \widetilde{\mathbf{f}}_3\ \widetilde{\mathbf{f}}_4\,]\,\varphi^T = [\,\mathbf{f}_1\ \mathbf{f}_2\ \mathbf{f}_3\ \mathbf{f}_4\,]\,.$$

4. Inverses of T+H-Bezoutians. Recall that in the Hankel and Toeplitz cases we proved that a nonsingular matrix is an H- or a T-Bezoutian if and only if it is the inverse of a Hankel or of a Toeplitz matrix, respectively (see Sections 4.1, 4.4, 7.6, 8.7). Such an assertion is also true in the T+H case. We start with proving the following part of it.

Theorem 11.6. *Let B be a nonsingular T+H-Bezoutian. Then B^{-1} is a T+H matrix.*

Proof. Taking Proposition 11.3 into account we have rank $\nabla_{T+H} B = 4$, and a rank decomposition of $\nabla_{T+H} B$ is of the form (11.7). In particular, this means that there are vectors $\mathbf{z}_i \in \mathbb{F}^n, i = 1, 2, 3, 4$, such that

$$BW_n - W_n B = B\mathbf{e}_1\,\mathbf{z}_1^T + B\mathbf{e}_n\,\mathbf{z}_2^T + \mathbf{z}_3\,\mathbf{e}_1^T B + \mathbf{z}_4\,\mathbf{e}_n^T B\,.$$

Applying B^{-1} from both sides this equality leads to

$$B^{-1}W_n - W_n B^{-1} = -(\mathbf{e}_1\,\mathbf{z}_1^T B^{-1} + \mathbf{e}_n\,\mathbf{z}_2^T B^{-1} + B^{-1}\mathbf{z}_3\,\mathbf{e}_1^T + B^{-1}\mathbf{z}_4\,\mathbf{e}_n^T)\,.$$

Thus, the matrix of order $n-2$ in the center of $\nabla(B^{-1})$ is the zero matrix. By Proposition 1.4 this proves that B^{-1} is a T+H matrix. □

In the next section we will show that the converse is also true, i.e., the inverse of a (nonsingular) T+H matrix is a T+H-Bezoutian.

12. Inverses of T+H-matrices

We consider now $n \times n$ matrices R_n which are the sum of a Toeplitz matrix T_n and a Hankel matrix H_n. For our purposes it is convenient to use a representation (1.5) for $m = n$,

$$T_n = T_n(\mathbf{a})\,,\ \mathbf{a} = (a_i)_{i=1-n}^{n-1}\,, \quad H_n = T_n(\mathbf{b})J_n\,,\ \mathbf{b} = (b_i)_{i=1-n}^{n-1}\,,$$

$$R_n = T_n(\mathbf{a}) + T_n(\mathbf{b})J_n = \begin{bmatrix} a_0 & \dots & a_{1-n} \\ \vdots & \ddots & \vdots \\ a_{n-1} & \dots & a_0 \end{bmatrix} + \begin{bmatrix} b_{1-n} & \dots & b_0 \\ \vdots & \iddots & \vdots \\ b_0 & \dots & b_{n-1} \end{bmatrix}. \quad (12.1)$$

We want to prove that the inverse of a T+H matrix R_n is a T+H-Bezoutian and even more, we want to present inversion formulas

$$R_n^{-1} = \mathrm{Bez}_{T+H}((\mathbf{g}_i, \mathbf{f}_i)_1^4)\,.$$

Thus, we have to answer the question how to obtain the vectors $\mathbf{g}_i, \mathbf{f}_i\,,\ i = 1, 2, 3, 4$. Note that representations of inverses of T+H matrices as T+H-Bezoutians allow fast matrix-vector multiplication by these matrices (in case $\mathbb{F} = \mathbb{C}$ see [38], in case $\mathbb{F} = \mathbb{R}$ [40], [42]).

1. Fundamental systems. Besides the nonsingular T+H matrix R_n of (12.1) we consider the $(n-2) \times (n+2)$ T+H matrices $\partial R_n, \partial R_n^T$ obtained from R_n, R_n^T after deleting the first and last rows and adding one column to the right and to the left by preserving the T+H structure,

$$\partial R_n = \begin{bmatrix} a_2 & a_1 & \dots & a_{2-n} & a_{1-n} \\ a_3 & a_2 & \dots & a_{3-n} & a_{2-n} \\ \vdots & \vdots & & \vdots & \vdots \\ a_{n-1} & a_{n-2} & \dots & a_{-1} & a_{-2} \end{bmatrix} + \begin{bmatrix} b_{1-n} & b_{2-n} & \dots & b_1 & b_2 \\ b_{2-n} & b_{3-n} & \dots & b_2 & b_3 \\ \vdots & \vdots & & \vdots & \vdots \\ b_{-2} & b_{-1} & \dots & b_{n-2} & b_{n-1} \end{bmatrix}, \quad (12.2)$$

$$\partial R_n^T = \begin{bmatrix} a_{-2} & a_{-1} & \dots & a_{n-2} & a_{n-1} \\ a_{-3} & a_{-2} & \dots & a_{n-3} & a_{n-2} \\ \vdots & \vdots & & \vdots & \vdots \\ a_{1-n} & a_{2-n} & \dots & a_1 & a_2 \end{bmatrix} + \begin{bmatrix} b_{1-n} & b_{2-n} & \dots & b_1 & b_2 \\ b_{2-n} & b_{3-n} & \dots & b_2 & b_3 \\ \vdots & \vdots & & \vdots & \vdots \\ b_{-2} & b_{-1} & \dots & b_{n-2} & b_{n-1} \end{bmatrix}. \quad (12.3)$$

Since R_n is nonsingular both matrices ∂R_n and ∂R_n^T are of full rank, which means

$$\dim \ker \partial R_n = \dim \ker \partial R_n^T = 4\,.$$

Any system of eight vectors $\{\mathbf{u}_i\}_{i=1}^4, \{\mathbf{v}_i\}_{i=1}^4$, where $\{\mathbf{u}_i\}_{i=1}^4$ is a basis of $\ker \partial R_n$ and $\{\mathbf{v}_i\}_{i=1}^4$ is a basis of $\ker \partial R_n^T$, is called *fundamental system for* R_n. The reason for this notation is that these vectors completely determine the inverse R_n^{-1}. In order to understand this we consider first a special fundamental system.

Hereafter we use the following notation. For a given vector $\mathbf{a} = (a_j)_{j=1-n}^{n-1}$ we define

$$\mathbf{a}_\pm = (a_{\pm j})_{j=1}^{n}\,, \tag{12.4}$$

where $a_{\pm n}$ can be arbitrarily chosen. The matrix $\nabla(R_n) = R_nW_n - W_nR_n$ allows a rank decomposition of the form,

$$\nabla(R_n) = -(\mathbf{a}_+ + \mathbf{b}_-^J)\mathbf{e}_1^T - (\mathbf{a}_-^J + \mathbf{b}_+)\mathbf{e}_n^T + \mathbf{e}_1(\mathbf{a}_- + \mathbf{b}_-^J)^T + \mathbf{e}_n(\mathbf{a}_+^J + \mathbf{b}_+)^T\,. \tag{12.5}$$

Multiplying (12.5) from both sides by R_n^{-1} we obtain a rank decomposition of $\nabla(R_n^{-1})$.

Proposition 12.1. *We have*

$$\nabla\left(R_n^{-1}\right) = \mathbf{x}_1\mathbf{y}_1^T + \mathbf{x}_2\mathbf{y}_2^T - \mathbf{x}_3\mathbf{y}_3^T - \mathbf{x}_4\mathbf{y}_4^T\,, \tag{12.6}$$

where $\mathbf{x}_i$ $(i = 1, 2, 3, 4)$ *are the solutions of*

$$R_n\mathbf{x}_1 = \mathbf{a}_+ + \mathbf{b}_-^J\,,\ R_n\mathbf{x}_2 = \mathbf{a}_-^J + \mathbf{b}_+\,,\ R_n\mathbf{x}_3 = \mathbf{e}_1\,,\ R_n\mathbf{x}_4 = \mathbf{e}_n\,, \tag{12.7}$$

and $\mathbf{y}_i$ $(i = 1, 2, 3, 4)$ *are the solutions of*

$$R_n^T\mathbf{y}_1 = \mathbf{e}_1\,,\ R_n^T\mathbf{y}_2 = \mathbf{e}_n\,,\ R_n^T\mathbf{y}_3 = \mathbf{a}_- + \mathbf{b}_-^J\,,\ R_n^T\mathbf{y}_4 = \mathbf{a}_+^J + \mathbf{b}_+\,. \tag{12.8}$$

According to (12.2), (12.3) we obtain the following fundamental system for R_n.

Proposition 12.2. *Let* $\mathbf{x}_i, \mathbf{y}_i \in \mathbb{F}^n$ *be defined by* (12.7), (12.8). *The vector system*

$$\left\{\mathbf{u}_1 = \begin{bmatrix} 1 \\ -\mathbf{x}_1 \\ 0 \end{bmatrix},\ \mathbf{u}_2 = \begin{bmatrix} 0 \\ -\mathbf{x}_2 \\ 1 \end{bmatrix},\ \mathbf{u}_3 = \begin{bmatrix} 0 \\ \mathbf{x}_3 \\ 0 \end{bmatrix},\ \mathbf{u}_4 = \begin{bmatrix} 0 \\ \mathbf{x}_4 \\ 0 \end{bmatrix}\right\} \tag{12.9}$$

is a basis of $\ker \partial R_n$, *the vector system*

$$\left\{\mathbf{v}_1 = \begin{bmatrix} 0 \\ \mathbf{y}_1 \\ 0 \end{bmatrix},\ \mathbf{v}_2 = \begin{bmatrix} 0 \\ \mathbf{y}_2 \\ 0 \end{bmatrix},\ \mathbf{v}_3 = \begin{bmatrix} 1 \\ -\mathbf{y}_3 \\ 0 \end{bmatrix},\ \mathbf{v}_4 = \begin{bmatrix} 0 \\ -\mathbf{y}_4 \\ 1 \end{bmatrix}\right\} \tag{12.10}$$

is a basis of $\ker \partial R_n^T$.

2. Inversion of T+H matrices. The special fundamental system of Proposition 12.2 deliver the parameters needed in a Bezoutian formula for R_n^{-1}. This is the initial point for our further considerations.

Theorem 12.3. *Let* R_n *be the nonsingular T+H matrix* (12.1) *and* $\{\mathbf{u}_i\}_{i=1}^4$, $\{\mathbf{v}_i\}_{i=1}^4$ *be the fundamental system for* R_n *given by* (12.9), (12.7), (12.10), (12.8). *Then* R_n^{-1} *is the T+H-Bezoutian defined by its generating polynomial as follows,*

$$R_n^{-1}(t, s) = \frac{\mathbf{u}_3(t)\mathbf{v}_3(s) + \mathbf{u}_4(t)\mathbf{v}_4(s) - \mathbf{u}_1(t)\mathbf{v}_1(s) - \mathbf{u}_2(t)\mathbf{v}_2(s)}{(t-s)(1-ts)}\,. \tag{12.11}$$

Proof. Since $\mathbf{x}_3$ is the first, $\mathbf{x}_4$ the last column, $\mathbf{y}_1^T$ is the first, $\mathbf{y}_2^T$ the last row of R_n^{-1} we conclude from (11.2)

$$\nabla_{T+H} R_n^{-1} = \begin{bmatrix} 0 & -\mathbf{y}_1^T & 0 \\ \mathbf{x}_3 & \nabla(R_n^{-1}) & \mathbf{x}_4 \\ 0 & -\mathbf{y}_2^T & 0 \end{bmatrix}.$$

Taking (12.6) into account this leads to

$$\nabla_{T+H} R_n^{-1} = [\,-\mathbf{u}_1 \;\; -\mathbf{u}_2 \;\; \mathbf{u}_3 \;\; \mathbf{u}_4\,]\,[\,\mathbf{v}_1 \;\mathbf{v}_2 \;\mathbf{v}_3 \;\mathbf{v}_4\,]^T,$$

where the vectors $\mathbf{u}_i$ and $\mathbf{v}_i$ are defined in (12.9), (12.10). The inversion formula follows now from (11.3). □

In particular, this theorem shows that if we want to use the vectors of any fundamental system in a Bezoutian formula for R_n^{-1} a "normalization" of them is necessary. For this purpose we introduce the following $(n+2)\times 4$ matrices

$$F = [\,\mathbf{e}_1 \;\mathbf{e}_{n+2} \;\mathbf{f}_1 \;\mathbf{f}_2\,], \quad G = [\,\mathbf{g}_1 \;\mathbf{g}_2 \;\mathbf{e}_1 \;\mathbf{e}_{n+2}\,],$$

where

$$\mathbf{f}_1 = (a_{1-i} + b_{i-n})_{i=0}^{n+1}\,,\;\; \mathbf{f}_2 = (a_{n-i} + b_{i-1})_{i=0}^{n+1}\,,$$
$$\mathbf{g}_1 = (a_{i-1} + b_{i-n})_{i=0}^{n+1}\,,\;\; \mathbf{g}_2 = (a_{i-n} + b_{i-1})_{i=0}^{n+1}\,,$$

with $a_{\pm n}, b_{\pm n}$ arbitrarily chosen. We call a fundamental system $\{\mathbf{u}_i\}_{i=1}^4, \{\mathbf{v}_i\}_{i=1}^4$ for R_n *canonical* if

$$F^T\,[\,\mathbf{u}_1 \;\mathbf{u}_2 \;\mathbf{u}_3 \;\mathbf{u}_4\,] = G^T\,[\,\mathbf{v}_1 \;\mathbf{v}_2 \;\mathbf{v}_3 \;\mathbf{v}_4\,] = I_4\,. \tag{12.12}$$

Proposition 12.4. *A fundamental system $\{\mathbf{u}_i\}_{i=1}^4, \{\mathbf{v}_i\}_{i=1}^4$ for R_n is canonical if and only if $\mathbf{u}_i$ is of the form* (12.9), (12.7) *and $\mathbf{v}_i$ is of the form* (12.10), (12.8) *for $i = 1, 2, 3, 4$.*

Proof. If $\{\mathbf{u}_i\}_{i=1}^4$ and $\{\mathbf{v}_i\}_{i=1}^4$ are canonical then (12.12) means, in particular, that the first component of $\mathbf{u}_1$ and $\mathbf{v}_3$ as well as the last components of $\mathbf{u}_2$ and $\mathbf{v}_4$ are one. The first and last components of the other vectors are zero. Hence there are vectors $\mathbf{x}_i, \mathbf{y}_i \in \mathbb{F}^n$ such that $\mathbf{u}_i, \mathbf{v}_i$ are of the form (12.9), (12.10). Now by (12.12) we have

$$[\,I_{+-}\mathbf{f}_1 \;\; I_{+-}\mathbf{f}_2\,]^T[\,\mathbf{x}_3 \;\mathbf{x}_4\,] = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}. \tag{12.13}$$

Here, for a given vector $\mathbf{h} = (h_i)_{i=0}^{n+1} \in \mathbb{F}^{n+2}$ the vector $I_{+-}\mathbf{h} \in \mathbb{F}^n$ is defined by

$$I_{+-}\mathbf{h} = (h_i)_{i=1}^n. \tag{12.14}$$

Since

$$(I_{+-}\mathbf{f}_1)^T = e_1^T R_n\,,\;\; (I_{+-}\mathbf{f}_2)^T = e_n^T R_n$$

and since $\begin{bmatrix} 0 \\ \mathbf{x}_3 \\ 0 \end{bmatrix}, \begin{bmatrix} 0 \\ \mathbf{x}_4 \\ 0 \end{bmatrix}$ are in $\ker \partial R_n$ equality (12.13) leads to

$$R_n\mathbf{x}_3 = e_1\,,\; R_n\mathbf{x}_4 = e_n\,.$$

Moreover, $\begin{bmatrix} 1 \\ -\mathbf{x}_1 \\ 0 \end{bmatrix} \in \ker \partial R_n$ means that $R_n\mathbf{x}_1 = \mathbf{a}_+ + \mathbf{b}_-^J$ and $\begin{bmatrix} 0 \\ -\mathbf{x}_2 \\ 1 \end{bmatrix} \in \ker \partial R_n$ means that $R_n\mathbf{x}_2 = \mathbf{a}_-^J + \mathbf{b}_+$. Similar arguments show that $\mathbf{y}_i$, $i = 1,2,3,4$, are the solutions of (12.8), and the necessity part of the proof is complete.

If $\{\mathbf{u}_i\}_{i=1}^4, \{\mathbf{v}_i\}_{i=1}^4$ are of the form (12.9), (12.7), and (12.10), (12.8) then, obviously, (12.12) is satisfied. □

Given an arbitrary fundamental system $\{\widetilde{\mathbf{u}}_i\}_{i=1}^4, \{\widetilde{\mathbf{v}}_i\}_{i=1}^4$ we define two 4×4 nonsingular matrices Γ_F, Γ_G,

$$F^T\,[\,\widetilde{\mathbf{u}}_1\ \widetilde{\mathbf{u}}_2\ \widetilde{\mathbf{u}}_3\ \widetilde{\mathbf{u}}_4\,] = \Gamma_F\,, \quad G^T\,[\,\widetilde{\mathbf{v}}_1\ \widetilde{\mathbf{v}}_2\ \widetilde{\mathbf{v}}_3\ \widetilde{\mathbf{v}}_4\,] = \Gamma_G\,.$$

We conclude that by

$$[\,\mathbf{u}_1\ \mathbf{u}_2\ \mathbf{u}_3\ \mathbf{u}_4\,] = [\,\widetilde{\mathbf{u}}_1\ \widetilde{\mathbf{u}}_2\ \widetilde{\mathbf{u}}_3\ \widetilde{\mathbf{u}}_4\,]\,\Gamma_F^{-1} \tag{12.15}$$

and

$$[\,\mathbf{v}_1\ \mathbf{v}_2\ \mathbf{v}_3\ \mathbf{v}_4\,] = [\,\widetilde{\mathbf{v}}_1\ \widetilde{\mathbf{v}}_2\ \widetilde{\mathbf{v}}_3\ \widetilde{\mathbf{v}}_4\,]\,\Gamma_G^{-1} \tag{12.16}$$

a canonical fundamental system $\{\mathbf{u}_i\}_{i=1}^4, \{\mathbf{v}_i\}_{i=1}^4$ is given. Note that for fixed $a_{\pm n}, b_{\pm n}$ the canonical fundamental system is unique. The following becomes clear.

Theorem 12.5. *Let R_n be the nonsingular T+H matrix* (12.1) *and $\{\widetilde{\mathbf{u}}_i\}_{i=1}^4, \{\widetilde{\mathbf{v}}_i\}_{i=1}^4$ be a fundamental system of R_n. Then the inverse R_n^{-1} is the T+H-Bezoutian* (12.11)*, where $\{\mathbf{u}_i\}_{i=1}^4, \{\mathbf{v}_i\}_{i=1}^4$ are given by* (12.15)*,* (12.16)*.*

Let R_n be given by (12.1). Hereafter we use also a representation of R_n which involves the projections $P_\pm = \frac{1}{2}(I_n \pm J_n)$ onto $\mathbb{F}_\pm^n$ introduced in (1.3) and the vectors

$$\mathbf{c} = (c_j)_{j=1-n}^{n-1} = \mathbf{a} + \mathbf{b}\,,\ \mathbf{d} = (d_j)_{j=1-n}^{n-1} = \mathbf{a} - \mathbf{b}\,,$$

namely

$$R_n = T_n(\mathbf{c})P_+ + T_n(\mathbf{d})P_-\,. \tag{12.17}$$

Instead of the solutions $\mathbf{x}_i$ of (12.7) and the solutions $\mathbf{y}_i$ of (12.8) we consider now the solutions of the following equations the right-hand sides of which depend on $\mathbf{c}, \mathbf{d}$ and $\widetilde{\mathbf{c}} = \mathbf{a}^J + \mathbf{b}$, $\widetilde{\mathbf{d}} = \mathbf{a}^J - \mathbf{b}$,

$$R_n\mathbf{w}_1 = \frac{1}{2}(\mathbf{c}_+ + \mathbf{c}_-^J)\,,\ R_n\mathbf{w}_2 = \frac{1}{2}(\mathbf{d}_+ - \mathbf{d}_-^J)\,,\ R_n\mathbf{w}_3 = P_+\mathbf{e}_1\,,\ R_n\mathbf{w}_4 = P_-\mathbf{e}_1 \tag{12.18}$$

and

$$R_n^T\mathbf{z}_1 = P_+\mathbf{e}_1\,,\ R_n^T\mathbf{z}_2 = P_-\mathbf{e}_1\,,\ R_n^T\mathbf{z}_3 = \frac{1}{2}(\widetilde{\mathbf{c}}_+ + \widetilde{\mathbf{c}}_-^J)\,,\ R_n^T\mathbf{z}_4 = \frac{1}{2}(\widetilde{\mathbf{d}}_+ - \widetilde{\mathbf{d}}_-^J)\,. \tag{12.19}$$

Here we use the notation (12.4). We introduce the vectors

$$\breve{\mathbf{u}}_1 = \begin{bmatrix} 1 \\ -2\mathbf{w}_1 \\ 1 \end{bmatrix}, \quad \breve{\mathbf{u}}_2 = \begin{bmatrix} 1 \\ -2\mathbf{w}_2 \\ -1 \end{bmatrix}, \quad \breve{\mathbf{u}}_3 = \begin{bmatrix} 0 \\ \mathbf{w}_3 \\ 0 \end{bmatrix}, \quad \breve{\mathbf{u}}_4 = \begin{bmatrix} 0 \\ \mathbf{w}_4 \\ 0 \end{bmatrix},$$

$$\breve{\mathbf{v}}_1 = \begin{bmatrix} 0 \\ \mathbf{z}_1 \\ 0 \end{bmatrix}, \quad \breve{\mathbf{v}}_2 = \begin{bmatrix} 0 \\ \mathbf{z}_2 \\ 0 \end{bmatrix}, \quad \breve{\mathbf{v}}_3 = \begin{bmatrix} 1 \\ -2\mathbf{z}_3 \\ 1 \end{bmatrix}, \quad \breve{\mathbf{v}}_4 = \begin{bmatrix} 1 \\ -2\mathbf{z}_4 \\ -1 \end{bmatrix}. \tag{12.20}$$

Now an inversion formula which involves these vectors follows from formula (12.11).

Proposition 12.6. *Let R_n be the nonsingular T+H matrix* (12.17)*. Then the inverse R_n^{-1} is given by*

$$R_n^{-1}(t,s) = \frac{\breve{\mathbf{u}}_3(t)\breve{\mathbf{v}}_3(s) + \breve{\mathbf{u}}_4(t)\breve{\mathbf{v}}_4(s) - \breve{\mathbf{u}}_1(t)\breve{\mathbf{v}}_1(s) - \breve{\mathbf{u}}_2(t)\breve{\mathbf{v}}_2(s)}{(t-s)(1-ts)}, \tag{12.21}$$

where $\{\breve{\mathbf{u}}_i\}_{i=1}^4, \{\breve{\mathbf{v}}_i\}_{i=1}^4$ are defined in (12.20)*.*

Proof. Since

$$[\,\breve{\mathbf{u}}_1\ \breve{\mathbf{u}}_2\ \breve{\mathbf{u}}_3\ \breve{\mathbf{u}}_4\,] = [\,\mathbf{u}_1\ \mathbf{u}_2\ \mathbf{u}_3\ \mathbf{u}_4\,]\,\varphi$$

and

$$[\,\breve{\mathbf{v}}_1\ \breve{\mathbf{v}}_2\ \breve{\mathbf{v}}_3\ \breve{\mathbf{v}}_4\,] = [\,\mathbf{v}_1\ \mathbf{v}_2\ \mathbf{v}_3\ \mathbf{v}_4\,]\,\varphi^{-1},$$

where φ is the block diagonal matrix

$$\varphi = \operatorname{diag}\left(\begin{bmatrix} 1 & 1 \\ 1 & -1 \end{bmatrix}, \frac{1}{2}\begin{bmatrix} 1 & 1 \\ 1 & -1 \end{bmatrix}\right),$$

the proposition follows from Proposition 11.5 and (12.11). □

3. Inversion of symmetric T+H matrices. It is easy to see that a T+H matrix is symmetric if and only if the Toeplitz part has this property. Let R_n be a nonsingular, symmetric T+H matrix (12.1). Then the solutions of (12.7) and (12.8) coincide,

$$\mathbf{y}_1 = \mathbf{x}_3\,,\ \mathbf{y}_2 = \mathbf{x}_4\,,\ \mathbf{y}_3 = \mathbf{x}_1\,,\ \mathbf{y}_4 = \mathbf{x}_2\,.$$

Using the inversion formula (12.11) R_n^{-1} is given by the vectors $\{\mathbf{u}_i\}_{i=1}^4$ of (12.9),

$$R_n^{-1}(t,s) = \frac{\mathbf{u}_3(t)\mathbf{u}_1(s) - \mathbf{u}_1(t)\mathbf{u}_3(s) + \mathbf{u}_4(t)\mathbf{u}_2(s) - \mathbf{u}_2(t)\mathbf{u}_4(s)}{(t-s)(1-ts)}. \tag{12.22}$$

Since $\mathbf{a} = \mathbf{a}^J$ we have $\mathbf{c} = \widetilde{\mathbf{c}}$, $\mathbf{d} = \widetilde{\mathbf{d}}$, and the inversion formula (12.21) can be simplified as well,

$$R_n^{-1}(t,s) = \frac{\breve{\mathbf{u}}_3(t)\breve{\mathbf{u}}_1(s) - \breve{\mathbf{u}}_1(t)\breve{\mathbf{u}}_3(s) + \breve{\mathbf{u}}_4(t)\breve{\mathbf{u}}_2(s) - \breve{\mathbf{u}}_2(t)\breve{\mathbf{u}}_4(s)}{(t-s)(1-ts)}. \tag{12.23}$$

If we have any basis $\{\widetilde{\mathbf{u}}_i\}_{i-1}^4$ of $\ker \partial R_n$, it remains to compute Γ_F, and $\{\mathbf{u}_i\}_{i=1}^4$ is given by (12.15).

We will not consider the skewsymmetric case, since a skewsymmetric T+H matrix is always a pure Toeplitz matrix. (For the skewsymmetric Toeplitz case see Section 4.7.)

4. Inversion of centrosymmetric T+H matrices. If R_n from (12.1) is centrosymmetric, i.e., $R_n^J = R_n$, then in view of $J_n T_n(\mathbf{a}) J_n = T_n(\mathbf{a}^J)$,

$$R_n = \frac{1}{2}(R_n + R_n^J) = T_n\left(\frac{1}{2}(\mathbf{a} + \mathbf{a}^J)\right) + T_n\left(\frac{1}{2}(\mathbf{b} + \mathbf{b}^J)\right) J_n\,.$$

Together with Exercises 15, 16 we conclude the following.

Proposition 12.7. *Let R_n be an $n \times n$ T+H matrix. Then the following assertions are equivalent.*

1. *R_n is centrosymmetric.*
2. *In the representation* (12.1) (*resp.* (12.17)) *the Toeplitz matrices $T_n(\mathbf{a})$ and $T_n(\mathbf{b})$* (*resp. $T_n(\mathbf{c})$ and $T_n(\mathbf{d})$*) *are symmetric.*
3. *In the representation* (12.1) (*resp.* 12.17)) $\mathbf{a}$ *and* $\mathbf{b}$ (*resp.* $\mathbf{c}$ *and* $\mathbf{d}$) *are symmetric vectors.*

Corollary 12.8. *A centrosymmetric T+H matrix R_n is also symmetric.*

Moreover, in the centrosymmetric case the representation (12.17) can be written in the form

$$R_n = P_+ T_n(\mathbf{c}) P_+ + P_- T_n(\mathbf{d}) P_-\,. \tag{12.24}$$

Now we specify the results for general T+H matrices to centrosymmetric T+H matrices R_n. Since R_n is symmetric we can use the simplifications of the previous subsection. To begin with we observe that the right-hand sides of the first and the third equations of (12.18) are symmetric and of the second and the fourth equations are skewsymmetric if we choose

$$c_n = c_{-n},\ d_n = d_{-n}\,.$$

Since centrosymmetric matrices map symmetric (skewsymmetric) vectors into symmetric (skewsymmetric) vectors, we conclude that the solutions $\mathbf{w}_1, \mathbf{w}_3$ of (12.18) as well as their extensions $\check{\mathbf{u}}_1, \check{\mathbf{u}}_3$ of (12.20) are symmetric, whereas $\mathbf{w}_2, \mathbf{w}_4$ and $\check{\mathbf{u}}_2, \check{\mathbf{u}}_4$ are skewsymmetric vectors. This leads to further simplifications of the inversion formula (12.23). But before presenting this formula let us introduce a more unified notation, where the subscript + designates symmetric, − skewsymmetric vectors in the fundamental system,

$$\mathbf{u}_+ = \begin{bmatrix} 0 \\ \mathbf{w}_3 \\ 0 \end{bmatrix},\ \mathbf{u}_- = \begin{bmatrix} 0 \\ \mathbf{w}_4 \\ 0 \end{bmatrix},\ \mathbf{v}_+ = \begin{bmatrix} 1 \\ -2\mathbf{w}_1 \\ 1 \end{bmatrix},\ \mathbf{v}_- = \begin{bmatrix} 1 \\ -2\mathbf{w}_2 \\ -1 \end{bmatrix}. \tag{12.25}$$

Here $\mathbf{w}_i$ are the solutions of (12.18) which turn obviously into pure Toeplitz equations,

$$T_n(\mathbf{c})\mathbf{w}_1 = P_+\mathbf{c}_+\,,\ T_n(\mathbf{d})\mathbf{w}_2 = P_-\mathbf{d}_+\,,\ T_n(\mathbf{c})\mathbf{w}_3 = P_+\mathbf{e}_1\,,\ T_n(\mathbf{d})\mathbf{w}_4 = P_-\mathbf{e}_1\,. \tag{12.26}$$

Note that these equations have unique symmetric or skewsymmetric solutions. Thus, the inversion formula (12.23) can be rewritten as a sum of a split Bezoutian of (+)-type and a split Bezoutian of (−)-type. These special Bezoutians were introduced in Section 2.11. Let us use the notations adopted there.

Theorem 12.9. *Let R_n be a nonsingular, centrosymmetric T+H matrix given by* (12.17) *and $\mathbf{u}_\pm, \mathbf{v}_\pm$ be the vectors of $\mathbb{F}_\pm^{n+2}$ defined in* (12.25)*, where the $\mathbf{w}_i$ are the unique symmetric or skewsymmetric solutions of* (12.26)*. Then*

$$R_n^{-1} = B_+ + B_- \,,$$

where $B_\pm$ are the split Bezoutians of $(\pm)$-type

$$B_\pm = \mathrm{Bez}_{\mathrm{split}}(\mathbf{v}_\pm, \mathbf{u}_\pm)\,.$$

Similar ideas as those of Section 4.5 lead to a slight modification of the last theorem. We extend the nonsingular centrosymmetric T+H matrix R_n given by (12.17) to a nonsingular centrosymmetric T+H matrix R_{n+2}, such that R_n is its central submatrix of order n.

$$R_{n+2} = T_{n+2}(\mathbf{c})P_+ + T_{n+2}(\mathbf{d})P_- \,. \tag{12.27}$$

Here $\mathbf{c}$ and $\mathbf{d}$ are extensions of the original vectors $\mathbf{c}$ and $\mathbf{d}$ by corresponding components $c_{-n} = c_n$, $d_{-n} = d_n$, $c_{-n-1} = c_{n+1}$, $d_{-n-1} = d_{n+1}$. Let $\mathbf{x}_{n+2}^\pm, \mathbf{x}_n^\pm$ be the unique symmetric or skewsymmetric solutions of

$$\begin{array}{ll} T_{n+2}(\mathbf{c})\mathbf{x}_{n+2}^+ = P_+\mathbf{e}_1\,, & T_n(\mathbf{c})\mathbf{x}_n^+ = P_+\mathbf{e}_1\,, \\ T_{n+2}(\mathbf{d})\mathbf{x}_{n+2}^- = P_-\mathbf{e}_1\,, & T_n(\mathbf{d})\mathbf{x}_n^- = P_-\mathbf{e}_1\,. \end{array} \tag{12.28}$$

(Note that $\mathbf{x}_n^+ = \mathbf{w}_3\,, \mathbf{x}_n^- = \mathbf{w}_4$. The solutions $\mathbf{x}_{n+2}^\pm$ are up to a constant factor equal to the vectors $\mathbf{v}_\pm$.)

Corollary 12.10. *Let R_{n+2} be a nonsingular, centrosymmetric extension* (12.27) *of R_n. Then the equations* (12.28) *have unique symmetric or skewsymmetric solutions and*

$$R_n^{-1} = \frac{1}{r_+}\mathrm{Bez}_{\mathrm{split}}(\mathbf{x}_{n+2}^+, \mathbf{u}_+) + \frac{1}{r_-}\mathrm{Bez}_{\mathrm{split}}(\mathbf{x}_{n+2}^-, \mathbf{u}_-)\,,$$

where $r_\pm$ are the first components of $\mathbf{x}_{n+2}^\pm$ and $\mathbf{u}_\pm = \begin{bmatrix} 0 \\ \mathbf{x}_n^\pm \\ 0 \end{bmatrix}$.

If $T_n(\mathbf{c})$ and $T_n(\mathbf{d})$ are nonsingular then R_n is nonsingular. Indeed, taking (12.24) into account $R_n\mathbf{u} = 0$ leads to

$$P_+T_n(\mathbf{c})P_+\mathbf{u} = -P_-T_n(\mathbf{d})P_-\mathbf{u}\,.$$

Hence $P_+\mathbf{u} = \mathbf{0}$ and $P_-\mathbf{u} = \mathbf{0}$ which means $\mathbf{u} = \mathbf{0}$. The converse is not true. Take, for example, $\mathbf{c} = (1,\ 1,\ 1)$ and $\mathbf{d} = (-1,\ 1, -1)$, then $T_2(\mathbf{c})$ and $T_2(\mathbf{d})$ are singular, whereas $R_2 = 2I_2$ is nonsingular. One might conjecture that for a nonsingular R_n

there is always a representation (12.17) with nonsingular $T_n(\mathbf{c})$ and $T_n(\mathbf{d})$. For $n = 2$ this is true. But this fails to be true for $n = 3$. Consider, for example,

$$\mathbf{c} = (1, 0, 1, 0, 1) \text{ and } \mathbf{d} = (0, 0, 1, 0, 0)\,.$$

Then

$$R_3 = T_3(\mathbf{c})P_+ + T_3(\mathbf{d})P_- = \frac{1}{2}\begin{bmatrix} 3 & 0 & 1 \\ 0 & 2 & 0 \\ 1 & 0 & 3 \end{bmatrix}$$

is nonsingular. But $T_3(\mathbf{c})$ is a chess-board matrix (1.4) with $c = 1, b = 0$ which is singular and uniquely determined in the representation of R_3 (cf. Exercise 16).

Let us consider besides $R_n = T_n(\mathbf{a}) + T_n(\mathbf{b})J_n$ the matrix $R_n^- = T(\mathbf{a}) - T(\mathbf{b})J_n$. If R_n is represented in the form (12.24) then the corresponding representation of R_n^- is

$$R_n^- = P_+T_n(\mathbf{d})P_+ + P_-T_n(\mathbf{c})P_-\,,$$

which means that the roles of $\mathbf{c}$ and $\mathbf{d}$ are interchanged. We conclude the following

Proposition 12.11. *The (symmetric) Toeplitz matrices $T_n(\mathbf{c})$ and $T_n(\mathbf{d})$ are nonsingular if and only if both R_n and R_n^- are nonsingular.*

Proof. We have already shown that the nonsingularity of $T_n(\mathbf{c})$ and $T_n(\mathbf{d})$ implies the nonsingularity of R_n. The nonsingularity of R_n^- follows with the same arguments. It remains to show that the singularity of $T_n(\mathbf{c})$ (or $T_n(\mathbf{d})$) leads to the singularity of R_n or R_n^-. Let $\mathbf{u}$ be a nontrivial vector such that $T_n(\mathbf{c})\mathbf{u} = 0$. We split $\mathbf{u}$ into its symmetric and skewsymmetric parts

$$\mathbf{u} = \mathbf{u}_+ + \mathbf{u}_- \quad (\mathbf{u}_\pm \in \mathbb{F}_\pm^n).$$

Clearly, at least one of the vectors $\mathbf{u}_+$ or $\mathbf{u}_-$ is nonzero, and $T_n(\mathbf{c})\mathbf{u}_+ = T_n(\mathbf{c})\mathbf{u}_- = 0$. Since

$$R_n\mathbf{u}_+ = T_n(\mathbf{c})\mathbf{u}_+\,, \quad R_n^-\mathbf{u}_- = T_n(\mathbf{c})\mathbf{u}_-$$

we obtain that R_n or R_n^- is singular. This is also obtained if we assume that $T_n(\mathbf{d})$ is singular. □

5. Inversion of centro-skewsymmetric T+H matrices. In this subsection let us consider T+H matrices R_n which are centro-skewsymmetric, $R_n = -R_n^J$. Since for an $n \times n$ centro-skewsymmetric matrix A, $\det A = (-1)^n \det A$, all centro-skewsymmetric matrices of odd order are singular. Hence we consider here mainly matrices of even order. The centro-skewsymmetric counterpart of Proposition 12.7 is as follows.

Proposition 12.12. *Let R_n be an $n \times n$ T+H matrix. Then the following assertions are equivalent.*

1. *R_n is centro-skewsymmetric.*
2. *There is a representation* (12.1) (*resp.* (12.17)) *such that the Toeplitz matrices $T_n(\mathbf{a})$ and $T_n(\mathbf{b})$ (resp. $T_n(\mathbf{c})$ and $T_n(\mathbf{d})$) are skewsymmetric.*
3. *There is a representation* (12.1) (*resp.* (12.17)) *such that* $\mathbf{a}$ *and* $\mathbf{b}$ (*resp.* $\mathbf{c}$ *and* $\mathbf{d}$) *are skewsymmetric vectors.*

In the remaining part of this subsection we only use such representations. In this case (12.17) can be rewritten as

$$R_n = P_- T_n(\mathbf{c}) P_+ + P_+ T_n(\mathbf{d}) P_- \,.$$

Its transposed matrix is given by

$$R_n^T = -(P_- T_n(\mathbf{d}) P_+ + P_+ T_n(\mathbf{c}) P_-) \,.$$

In the equations (12.19) we have $\widetilde{\mathbf{c}} = -\mathbf{d}$ and $\widetilde{\mathbf{d}} = -\mathbf{c}$.

In general, R_n is neither symmetric nor skewsymmetric, thus a connection between the solutions of (12.18) and (12.19) is not obvious. If we choose $c_n = -c_{-n}$ and $d_n = -d_{-n}$ than $\mathbf{c}_- = -\mathbf{c}_+$, $\mathbf{d}_- = -\mathbf{d}_+$. Hence the right-hand sides of the equations (12.18), (12.19) are either symmetric or skewsymmetric. Since R_n as a centro-skewsymmetric matrix maps $\mathbb{F}_\pm^n$ to $\mathbb{F}_\mp^n$, we obtain that the solutions are also either symmetric or skewsymmetric. Let us indicate these symmetry properties again by denoting

$$\begin{gathered} \mathbf{w}_+ = \mathbf{w}_1 \,, \mathbf{w}_- = \mathbf{w}_2 \,, \mathbf{x}_- = \mathbf{w}_3 \,, \mathbf{x}_+ = \mathbf{w}_4, \\ \widetilde{\mathbf{x}}_- = \mathbf{z}_1 \,, \ \widetilde{\mathbf{x}}_+ = \mathbf{z}_2 \,, \ \widetilde{\mathbf{w}}_+ = \mathbf{z}_3 \,, \ \widetilde{\mathbf{w}}_- = \mathbf{z}_4. \end{gathered}$$

Since these symmetries pass to the augmented vectors $\breve{\mathbf{u}}_j$, $\breve{\mathbf{v}}_j$ of (12.20) we set

$$\begin{gathered} \mathbf{v}_+ = \breve{\mathbf{u}}_1 \,, \ \mathbf{v}_- = \breve{\mathbf{u}}_2 \,, \ \mathbf{u}_- = \breve{\mathbf{u}}_3 \,, \ \mathbf{u}_+ = \breve{\mathbf{u}}_4 \,, \\ \widetilde{\mathbf{v}}_+ = \breve{\mathbf{v}}_3 \,, \ \widetilde{\mathbf{v}}_- = \breve{\mathbf{v}}_4 \,, \ \widetilde{\mathbf{u}}_- = \breve{\mathbf{v}}_1 \,, \ \widetilde{\mathbf{u}}_+ = \breve{\mathbf{v}}_2 \,. \end{gathered} \tag{12.29}$$

The equations (12.18), (12.19) turn into Toeplitz equations,

$$T_n(\mathbf{c})\mathbf{x}_+ = P_-\mathbf{e}_1 \,, \ T_n(\mathbf{c})\mathbf{w}_+ = P_-\mathbf{c}_+ \,, \ T_n(\mathbf{d})\mathbf{x}_- = P_+\mathbf{e}_1 \,, \ T_n(\mathbf{d})\mathbf{w}_- = P_+\mathbf{d}_+ \tag{12.30}$$

and

$$T_n(\mathbf{c})\widetilde{\mathbf{x}}_- = -P_+\mathbf{e}_1 \,, \ T_n(\mathbf{c})\widetilde{\mathbf{w}}_- = P_+\mathbf{c}_+ \,, \ T_n(\mathbf{d})\widetilde{\mathbf{x}}_+ = -P_-\mathbf{e}_1 \,, \ T_n(\mathbf{d})\widetilde{\mathbf{w}}_+ = P_-\mathbf{d}_+ \,. \tag{12.31}$$

According to Proposition 12.6 and (11.3) R_n^{-1} given by the augmented vectors (12.29) of these solutions via

$$\nabla_{T+H} R_n^{-1} = \mathbf{u}_- \widetilde{\mathbf{v}}_+^T - \mathbf{v}_+ \widetilde{\mathbf{u}}_-^T - \mathbf{v}_- \widetilde{\mathbf{u}}_+^T + \mathbf{u}_+ \widetilde{\mathbf{v}}_-^T \,. \tag{12.32}$$

Now we show how the solutions of (12.30) and (12.31) are related. First we compare the equations $T_n(\mathbf{c})\widetilde{\mathbf{x}}_- = P_+\mathbf{e}_1$ and $T_n(\mathbf{c})\mathbf{x}_+ = P_-\mathbf{e}_1$ for any $\mathbf{c} \in \mathbb{F}_-^{2n-1}$. The following lemma shows that there is an essential difference between the centrosymmetric and centro-skewsymmetric cases.

Lemma 12.13. *Let $T_n(\mathbf{c})$ be skewsymmetric. If the equation $T_n(\mathbf{c})\widetilde{\mathbf{x}}_- = -P_+\mathbf{e}_1$ is solvable, then equation $T_n(\mathbf{c})\mathbf{x}_+ = P_-\mathbf{e}_1$ is also solvable, and if n is even, then the converse is also true. If $\widetilde{\mathbf{x}}_-$ is a skewsymmetric solution of the first equation, then a solution of the second equation is given by*

$$\mathbf{x}_+(t) = \frac{1+t}{1-t}\,\widetilde{\mathbf{x}}_-(t). \tag{12.33}$$

Proof. If $T_n(\mathbf{c})\widetilde{\mathbf{x}}_- = -P_+\mathbf{e}_1$ is solvable, then there exists a skewsymmetric solution $\widetilde{\mathbf{x}}_-$. Since $\widetilde{\mathbf{x}}_-$ is skewsymmetric, we have $\widetilde{\mathbf{x}}_-(1) = 0$. Hence (12.33) defines a polynomial $\mathbf{x}_+(t)$. Moreover, the coefficient vector $\mathbf{x}_+$ is symmetric.

Let $\mathbf{z} \in \mathbb{F}^n$ be defined by $\mathbf{z}(t) = \frac{1}{t-1}\,\widetilde{\mathbf{x}}_-(t)$ and $\mathbf{z}^1(t) = t\mathbf{z}(t)$. If now $T_n(\mathbf{c})\mathbf{z} = (r_k)_{k=1}^n$, then $T_n(\mathbf{c})\mathbf{z}^1 = (r_{k-1})_{k=1}^n$, where r_0 is some number. In view of $T_n(\mathbf{c})(\mathbf{z}^1 - \mathbf{z}) = -P_+\mathbf{e}_1$, we have

$$r_0 - r_1 = -\frac{1}{2},\ r_1 = r_2 = \cdots = r_{n-1},\ r_{n-1} - r_n = -\frac{1}{2}.$$

Since the $(n-1) \times (n-1)$ principal submatrix T_{n-1} of $T_n(\mathbf{c})$ is skewsymmetric and the vector $\mathbf{z}' \in \mathbb{F}^{n-1}$ obtained from $\mathbf{z}$ by deleting the last (zero) component is symmetric, the vector $T_{n-1}\mathbf{z}' = (r_k)_{k=1}^{n-1}$ is skewsymmetric. Hence

$$r_0 = -\frac{1}{2},\ r_1 = r_2 = \cdots = r_{n-1} = 0,\ r_n = \frac{1}{2}\,.$$

We conclude that $T_n(\mathbf{c})(\mathbf{z} + \mathbf{z}^1) = -P_-\mathbf{e}_1$. This means that $\mathbf{x}_+ = \mathbf{z} + \mathbf{z}^1$.

The proof of the converse direction follows the same lines. One has to take into account that if n is even and $\mathbf{x}_+$ is symmetric, then $\mathbf{x}_+(-1) = 0$. Hence $\mathbf{z}(t) = \frac{1}{t+1}\mathbf{x}_+(t)$ is a polynomial. □

Note that the converse direction of Lemma 12.13 is not true if n is odd. If, for example,

$$T_3(\mathbf{c}) = \begin{bmatrix} 0 & -1 & 0 \\ 1 & 0 & -1 \\ 0 & 1 & 0 \end{bmatrix}$$

then $T_3(\mathbf{c})\mathbf{x}_+ = P_-\mathbf{e}_1$ is solvable but $T_n(\mathbf{c})\widetilde{\mathbf{x}}_- = P_+\mathbf{e}_1$ is not.

The relation between the solutions $\mathbf{x}_+$ and $\widetilde{\mathbf{x}}_-$ extends to the augmented vectors $\mathbf{u}_+$ and $\widetilde{\mathbf{u}}_-$. We have

$$\mathbf{u}_+(t) = \frac{1+t}{1-t}\,\widetilde{\mathbf{u}}_-(t).$$

Replacing $\mathbf{c}$ by $\mathbf{d}$ we obtain

$$\widetilde{\mathbf{u}}_+(t) = \frac{1+t}{1-t}\,\mathbf{u}_-(t).$$

Now we compare the equations $T_n(\mathbf{c})\mathbf{w}_+ = P_-\mathbf{c}_+$ and $T_n(\mathbf{c})\widetilde{\mathbf{w}}_- = P_+\mathbf{c}_+$. More precisely, we compare the augmented vectors $\mathbf{v}_+$ and $\widetilde{\mathbf{v}}_-$.

Lemma 12.14. *Let $T_n(\mathbf{c})$ be skewsymmetric. If the equation $T_n(\mathbf{c})\widetilde{\mathbf{w}}_- = P_+\mathbf{c}_+$ is solvable, then the equation $T_n(\mathbf{c})\mathbf{w}_+ = P_-\mathbf{c}_+$ is also solvable, and the augmented vectors of these solutions are related via*

$$\mathbf{v}_+(t) = \frac{1+t}{1-t}\,\widetilde{\mathbf{v}}_-(t). \tag{12.34}$$

If n is even, then the solvability of $T_n(\mathbf{c})\mathbf{w}_+ = P_-\mathbf{c}_+$ implies the solvability of $T_n(\mathbf{c})\widetilde{\mathbf{w}}_- = P_+\mathbf{c}_+$.

Proof. Let $\widetilde{T}$ denote the $n \times (n+2)$ matrix $\widetilde{T} = [c_{i-j+1}]_{i=0\ j=0}^{n-1\ n+1}$ with $c_{-n} = c_n$. If $T_n(\mathbf{c})\widetilde{\mathbf{w}}_- = P_+\mathbf{c}_+$, then $\widetilde{T}\widetilde{\mathbf{v}}_- = \mathbf{0}$. Furthermore, if $\widetilde{\mathbf{w}}_-$ is skewsymmetric, then $\widetilde{\mathbf{v}}_-$ is skewsymmetric. Hence $\widetilde{\mathbf{v}}_-(1) = 0$ and $\mathbf{z}(t) = \frac{1}{1-t}\widetilde{\mathbf{v}}_-(t)$ is a polynomial. We consider the coefficient vector $\mathbf{z}$ of $\mathbf{z}(t)$ as a vector in $\mathbb{F}^{n+2}$ and denote the coefficient vector of $t\mathbf{z}(t)$ by $\mathbf{z}^1$.

Suppose that $\widetilde{T}\mathbf{z} = (r_k)_1^n$, then $\widetilde{T}\mathbf{z}^1 = (r_{k-1})_1^n$, where r_0 is some number. Since $\mathbf{z} - \mathbf{z}^1 = \widetilde{\mathbf{v}}_-$ and $\widetilde{T}\widetilde{\mathbf{v}}_- = 0$, we conclude that $r_0 = \cdots = r_n$.

Let $T_{n+1}(\mathbf{c})$ denote the $(n+1)\times(n+1)$ matrix $[c_{i-j}]_{i,j=0}^n$ and $\mathbf{z}' \in \mathbb{F}^{n+1}$ the vector obtained from $\mathbf{z}$ deleting the last (zero) component. Then $T_{n+1}(c)\mathbf{z}' = (r_k)_{k=0}^n$. Here $T_{n+1}(\mathbf{c})$ is skewsymmetric and $\mathbf{z}'$ is symmetric, thus the vector $(r_k)_{k=0}^n$ is skewsymmetric. Since all components are equal, it must be the zero vector. We obtain $\widetilde{T}(\mathbf{z} + \mathbf{z}^1) = \mathbf{0}$. Observe that $\mathbf{z} + \mathbf{z}^1$ is symmetric and that its first component is equal to 1. Therefore, $\mathbf{z} + \mathbf{z}^1 = \mathbf{v}_+ = [\,1\ \ -2\mathbf{w}_+^T\ 1\,]^T$ for some symmetric vector $\mathbf{w}_+ \in \mathbb{F}^n$. This vector is now a solution of the equation $T_n(\mathbf{c})\mathbf{w}_+ = P_-\mathbf{c}_+$.

The converse direction is proved analogously taking into account that if n is even, then the length of $\mathbf{v}_+$, which is $n+2$, is even. Hence $\mathbf{v}_+(-1) = 0$ and $\mathbf{z}(t) = \frac{1}{1+t}\mathbf{v}_+(t)$ is well defined. □

Replacing $\mathbf{c}$ by $\mathbf{d}$ we obtain

$$\widetilde{\mathbf{v}}_+(t) = \frac{1+t}{1-t}\,\mathbf{v}_-(t).$$

Taking (12.32), Lemma 12.13 and Lemma 12.14 together we arrive at

$$\begin{aligned}\nabla_{T+H}(R_n^{-1})(t,s) \;&=\; \mathbf{u}_-(t)\frac{1+s}{1-s}\mathbf{v}_-(s) - \mathbf{v}_-(t)\frac{1+s}{1-s}\mathbf{u}_-(s)\\ &\quad -\mathbf{v}_+(t)\frac{1-s}{1+s}\mathbf{u}_+(s) + \mathbf{u}_+(t)\frac{1-s}{1+s}\mathbf{v}_+(s)\,,\end{aligned} \tag{12.35}$$

which finally leads to the following theorem taking (11.3) into account.

Theorem 12.15. *Let the centro-skewsymmetric T+H matrix R_n be nonsingular and given by* (12.17)*. Then the equations* (12.30) *are solvable and the generating function of the inverse matrix is given by the augmented vectors of the solutions of these equations via*

$$R_n^{-1}(t,s) = B_+(t,s)\,\frac{s-1}{s+1} + B_-(t,s)\,\frac{s+1}{s-1}$$

and

$$B_\pm = \mathrm{Bez}_{\mathrm{split}}(\mathbf{u}_\pm, \mathbf{v}_\pm)\,.$$

Note that for a nonsingular matrix R_n all equations (12.30) and (12.31) are uniquely solvable. Moreover, we observe that $\mathbf{x} = \mathbf{x}_+ - \widetilde{\mathbf{x}}_-$ is a solution of $T_n(\mathbf{c})\mathbf{x} = \mathbf{e}_1$ and $\mathbf{w} = \mathbf{w}_+ - \widetilde{\mathbf{w}}_-$ is a solution of $T_n(\mathbf{c})\mathbf{w} = \mathbf{c}_-^J$. Taking Proposition

4.6 into account we obtain the nonsingularity of $T_n(\mathbf{c})$. Analogously, $T_n(\mathbf{d})$ is nonsingular. This leads to the following surprising conclusion.

Corollary 12.16. *For a centro-skewsymmetric T+H matrix*

$$R_n = T(\mathbf{a}) + T(\mathbf{b})J_n = T(\mathbf{c})P_+ + T(\mathbf{d})P_-$$

with skewsymmetric vectors $\mathbf{a}, \mathbf{b}, \mathbf{c}, \mathbf{d}$, *the following assertions are equivalent:*

1. R_n *is nonsingular.*
2. $R_n^- = T(\mathbf{a}) - T(\mathbf{b})J_n$ *is nonsingular.*
3. $T(\mathbf{c})$ *and* $T(\mathbf{d})$ *are nonsingular.*

To present the counterpart of Corollary 12.10 let us extend the nonsingular centro-skewsymmetric T+H matrix R_n given by (12.17) to a nonsingular centro-skewsymmetric T+H matrix R_{n+2}, such that R_n is its central submatrix of order n.

$$R_{n+2} = T_{n+2}(\mathbf{c})P_+ + T_{n+2}(\mathbf{d})P_- \,, \tag{12.36}$$

where $\mathbf{c}$ and $\mathbf{d}$ are extensions of the original vectors $\mathbf{c}$ and $\mathbf{d}$ by corresponding components $c_{-n} = -c_n$, $d_{-n} = -d_n$, $c_{-n-1} = -c_{n+1}$, $d_{-n-1} = -d_{n+1}$. Let $\mathbf{x}_{n+2}^{\pm}, \mathbf{x}_n^{\pm}$ be the unique symmetric or skewsymmetric solutions of

$$\begin{array}{ll} T_{n+2}(\mathbf{c})\mathbf{x}_{n+2}^- = P_+\mathbf{e}_1 \,, & T_n(\mathbf{c})\mathbf{x}_n^- = P_+\mathbf{e}_1 \,, \\ T_{n+2}(\mathbf{d})\mathbf{x}_{n+2}^+ = P_-\mathbf{e}_1 \,, & T_n(\mathbf{d})\mathbf{x}_n^+ = P_-\mathbf{e}_1 \,. \end{array} \tag{12.37}$$

Note that $\mathbf{x}_n^{\pm} = -\widetilde{\mathbf{x}}_{\pm}$, (solutions of (12.31)), thus $-\mathbf{u}_{\pm}$ are the augmented vectors defined by $\mathbf{u}_{\pm}(t) = t\mathbf{x}_n^{\pm}(t)$. The solutions $\mathbf{x}_{n+2}^{\pm}$ are up to a constant factor equal to the vectors $\mathbf{v}_{\pm}$.

Corollary 12.17. *Let* R_{n+2} *be a nonsingular and centro-skewsymmetric extension* (12.36) *of* R_n. *Then the equations* (12.37) *have unique symmetric or skewsymmetric solutions and*

$$R_n^{-1} = \frac{1}{r_+}\mathrm{Bez}_{\mathrm{split}}(\mathbf{x}_{n+2}^+, \mathbf{u}_+)\frac{s-1}{s+1} + \frac{1}{r_-}\mathrm{Bez}_{\mathrm{split}}(\mathbf{x}_{n+2}^-, \mathbf{u}_-)\frac{s+1}{s-1} \,,$$

where $r_{\pm}$ *are the first components of* $\mathbf{x}_{n+2}^{\pm}$.

13. Exercises

1. An $n \times n$ matrix B is called *quasi-T-Bezoutian* if $\nabla_T B$ introduced in (2.17) has rank 2 at most.
   (a) Show that B is a quasi-T-Bezoutian if and only if BJ_n is a quasi-H-Bezoutian (introduced in Section 2.4).
   (b) State and prove a proposition about the representation of a quasi-T-Bezoutian that is analogous to Proposition 2.4.

2. The special Toeplitz matrix

$$Z_n^{\alpha}(\mathbf{a}) = \begin{bmatrix} a_0 & \alpha a_{n-1} & \dots & \alpha a_1 \\ a_1 & a_0 & \ddots & \vdots \\ \vdots & \ddots & \ddots & \alpha a_{n-1} \\ a_{n-1} & \dots & a_1 & a_0 \end{bmatrix} \qquad (\alpha \in \mathbb{F})$$

is called *α-circulant*.

   (a) Show that the T-Bezoutian of a polynomial $\mathbf{u}(t) \in \mathbb{F}^{n+1}$ and $t^n - \alpha$ is an α-circulant and each α-circulant is of this form.

   (b) Show that a T-Bezoutian is a Toeplitz matrix if and only if it is an α-circulant for some α or an upper triangular Toeplitz matrix.

3. Let $\mathbf{u}(t)$ be a polynomial of degree n and $\mathbf{v}(t)$ a polynomial of degree $\leq n$. Describe the nullspace of the transpose of $\mathrm{Res}_p(\mathbf{u}, \mathbf{v})$ (introduced in Section 3.1) in terms of the greatest common divisor of $\mathbf{u}(t)$ and $\mathbf{v}(t)$. Use this to show that the nullity of $\mathrm{Res}_p(\mathbf{u}, \mathbf{v})$ is, independently of p, equal to the degree of the greatest common divisor of $\mathbf{u}(t)$ and $\mathbf{v}(t)$.

4. (a) Show that an $n \times n$ matrix A is Toeplitz if and only if $I_{+-} \nabla_T(A) I_{+-}^T$ is the zero matrix, where I_{+-} is introduced in (12.14).

   (b) Show that the product of two nonzero Toeplitz matrices is Toeplitz again if and only if both factors are α-circulants for the same α or both are upper (or lower) triangular Toeplitz matrices.

5. Prove that the nonsingularity of a Toeplitz matrix $T_n = [\, a_{i-j} \,]_{i,j=1}^n$ follows from the solvability of the equations

$$T_n \mathbf{y} = \mathbf{e}_1 \quad \text{and} \quad T_n \mathbf{z} = (a_{-n+j-1})_{j=1}^n \,,$$

where a_{-n} is arbitrarily chosen. Construct a fundamental system from these solutions.

6. Design a Levinson-type algorithm for the solution of the Bezout equation (7.14), i.e., an algorithm that does not rely on successive polynomial division. Compare the complexity of this algorithm with the complexity of the algorithm described in Section 7.7.

   Hint. Consider first the "regular" case in which the degrees of all quotients $\mathbf{q}_i(t)$ are equal to 1.

7. Let $\mathbf{p}(t) = p_1 + p_2 t + p_3 t^2 + t^3$ be a monic real polynomial. Show the following *theorem of Vyshnegradsky*. The polynomial $\mathbf{p}(t)$ has all its roots in the left half-plane if and only if all coefficients are positive and $p_2 p_3 > p_1$.

8. The factorizations presented in this paper can be used to derive formulas for the determinants of Bezoutians, Hankel, Toeplitz and resultant matrices. To solve the following problems one can use Vandermonde factorization or reduction and take into account that

$$\det V_n(\mathbf{t}) = \prod_{i>j} (t_i - t_j) \,,$$

where $\mathbf{t} = (t_1, \ldots, t_n)$ or apply Barnett's formula and

$$\det \mathbf{p}(C(\mathbf{u})) = \prod_{i=1}^{n} \mathbf{p}(t_i)$$

if $\mathbf{u}(t) = \prod_{i=1}^{n}(t - t_i)$. Suppose that $\mathbf{v}(t) = \prod_{i=1}^{m}(t - s_i)$ and $\mathbf{u}(t) = \prod_{i=1}^{n}(t - t_i)$ are complex polynomials and $m \le n$.

(a) Show that

$$\det H_n\left(\frac{\mathbf{v}}{\mathbf{u}}\right) = (-1)^{\frac{n(n-1)}{2}} \prod_{i=1}^{n}\prod_{j=1}^{m}(t_i - s_j)\,,$$

and find an analogous formula for $\det T_n\left(\frac{\mathbf{v}}{\mathbf{u}}\right)$.

(b) Derive from (a) formulas for the determinants of $\mathrm{Bez}_H(\mathbf{u}, \mathbf{v})$, $\mathrm{Bez}_T(\mathbf{u}, \mathbf{v})$, and $\mathrm{Res}\,(\mathbf{u}, \mathbf{v})$.

9. Find the Toeplitz matrices

$$T_n\left(\frac{1}{(1-ct)^m}\right) \quad \text{and} \quad T_n\left(\left(\frac{1+ct}{1-ct}\right)^m\right)$$

10. Let $\mathbf{u}(t)$ and $\mathbf{v}(t) = \mathbf{v}_1(t)\mathbf{v}_2(t)$ be polynomials of degree n.

(a) Show that

$$\mathrm{Bez}_H(\mathbf{u}, \mathbf{v}) = \mathrm{Bez}_H(\mathbf{u}, \mathbf{v}_1)B(\mathbf{u})^{-1}\mathrm{Bez}_H(\mathbf{u}, \mathbf{v}_2)$$

(b) If $\mathbf{u}(t)$ and $\mathbf{v}(t)$ are monic, show that

$$\mathrm{Bez}_H(\mathbf{u}, \mathbf{v}) = B(\mathbf{u})J_nB(\mathbf{v})(C(\mathbf{u})^n - C(\mathbf{v})^n)\,.$$

11. Let $\mathbf{u}(t)$ and $\mathbf{v}(t)$ be complex polynomials of degree n and m, respectively, where $m \le n$, and let $t_1, \ldots, t_r$ be the different roots of the greatest common divisor of $\mathbf{u}(t)$ and $\mathbf{v}(t)$ and $\nu_1, \ldots, \nu_r$ their multiplicities. Let vectors $\ell_k(c, \nu)$ be defined by

$$\ell_k(c, \nu) = \left(\binom{i-1}{\nu-1}c^{i-k}\right)_{i=1}^{k}\,.$$

Show that the vectors $\ell_n(t_i, k)$, where $k = 1, \ldots, \nu_i$, $i = 1, \ldots, r$ form a basis of the nullspace of $\mathrm{Bez}_H(\mathbf{u}, \mathbf{v})$ and the corresponding vectors $\ell_{m+n+p}(t_i, k)$ a basis of the nullspace of $\mathrm{Res}_p(\mathbf{u}, \mathbf{v})$ introduced in Section 3.1.

12. Let $\mathbf{u}(t)$ and $\mathbf{p}(t)$ be coprime polynomials and $\deg \mathbf{p}(t) \le \deg \mathbf{u}(t) = n$. Show that for $k > n$ the coefficient vectors of $t^j\mathbf{u}(t)$, $j = 1, \ldots, k - n$ form a basis of the nullspace of $H_k\left(\frac{\mathbf{p}}{\mathbf{u}}\right)$.

13. Let $\mathbf{u}(t)$ be a polynomial with real coefficients of degree n.

(a) Describe the number of different positive real roots in terms of the signatures of the matrices $H_n\left(\frac{t\mathbf{u}'(t)}{\mathbf{u}(t)}\right)$ and $H_n\left(\frac{\mathbf{u}'}{\mathbf{u}}\right)$.

(b) Let a and b be real numbers, $a < b$. Describe the number of different real roots of $\mathbf{u}(t)$ in the interval $[\,a, b\,]$ in terms of the signatures of the

matrices $H_n(\mathbf{g})$ and $H_n\left(\frac{\mathbf{u}'}{\mathbf{u}}\right)$, where

$$\mathbf{g}(t) = \frac{(t-a)(b-t)\mathbf{u}'(t) + nt^2\mathbf{u}(t)}{\mathbf{u}(t)}\,.$$

(c) Prove a representation of $\mathrm{Res}(\mathbf{u}, \mathbf{v})$ which is analogous to that one of Proposition 3.3 but involves $\mathrm{Bez}_T(\mathbf{u}, \mathbf{v})$ instead of $\mathrm{Bez}_H(\mathbf{u}, \mathbf{v})$.

14. Prove that a matrix A is a T+H matrix if and only if the matrix $I_{+-}\nabla(A)I_{+-}^T$ is the zero matrix, where $\nabla(A)$ is introduced in (1.11) and I_{+-} in (12.14).
15. Let $\mathbf{e}$ and $\mathbf{e}_\sigma$ denote the vectors of $\mathbb{F}^{2n-1}$

$$\mathbf{e} = (\,1,\,1\ldots,\,1\,) \quad \text{and} \quad \mathbf{e}_\sigma = (\,(-1)^i\,)_{i=1}^{2n-1}\,.$$

Show that a T+H matrix $R_n = T_n(\mathbf{a}) + T_n(\mathbf{b})J_n$ is equal to $R'_n = T_n(\mathbf{a}') + T_n(\mathbf{b}')J_n$ if and only if, for some $\alpha, \beta \in \mathbb{F}$, $\mathbf{a}' = \mathbf{a} + \alpha\mathbf{e} + \beta\mathbf{e}_\sigma$ and $\mathbf{b}' = \mathbf{b} - \alpha\mathbf{e} - \beta(-1)^{n-1}\mathbf{e}_\sigma$.

16. Let R_n be an $n \times n$ T+H matrix given by (12.17) and by $R_n = T_n(\mathbf{c}')P_+ + T_n(\mathbf{d}')P_-$. Show that
    (a) If n is odd, then $\mathbf{c}' = \mathbf{c}$, i.e., $\mathbf{c}$ is unique, and $\mathbf{d}'$ is of the form $\mathbf{d}' = \mathbf{d} + \alpha\mathbf{e} + \beta\mathbf{e}_\sigma$ for $\alpha, \beta \in \mathbb{F}$.
    (b) If n is even, then $\mathbf{c}'$ is of the form $\mathbf{c}' = \mathbf{c} + \alpha\mathbf{e}_\sigma$ and $\mathbf{d}'$ of the form $\mathbf{d}' = \mathbf{d} + \beta\mathbf{e}$ for $\alpha, \beta \in \mathbb{F}$.

    Here $\mathbf{e}$ and $\mathbf{e}_\sigma$ are as above.
17. Let R_n be an $n \times n$ nonsingular, centro-skewsymmetric T+H matrix given by (12.17). Show that

$$R_n^{-1} = T_n(\mathbf{c})^{-1}P_- + T_n(\mathbf{d})^{-1}P_+\,.$$

14. Notes

1. The Bezoutian and the resultant matrix have a long history, which goes back to the 18th century. Both concepts grew up from the work of Euler [5] in connection with the elimination of variables for the solution of systems of nonlinear algebraic equations. In 1764, Bezout generalized a result of Euler [1]. In his solution the determinant of a matrix occurred which was only in 1857 shown by Cayley [3] to be the same as that being today called the (Hankel) Bezoutian. For more detailed information see [66].
2. The classical studies of Jacobi [51] and Sylvester [65] utilized Bezoutians in the theory of separation of polynomial roots. Hermite [49] studied the problem of counting the roots of a polynomial in the upper half-plane. Clearly, this is equivalent to finding the number of roots in the left half-plane, which is important for stability considerations. A nice review of classical results concerning root localization problems is given in the survey paper [52]; see also [50], [8], [63], [61], [60].
3. The importance of Bezoutians for the inversion of Hankel and Toeplitz matrices became clear much later. Only in 1974, Lander [55] established the

fundamental result that the inverse of a (nonsingular) Hankel matrix can be represented as a Bezoutian of two polynomials and that, conversely, any nonsingular Bezoutian is the inverse of a Hankel matrix. Similar results are true for Toeplitz matrices. In [55] also a Vandermonde factorization of Bezoutians was presented.

4. There is a huge number of papers and books dedicated to Bezoutians, resultant matrices and connected problems. Let me recommend some books and survey papers (see also the references therein) to light the younger history and recent developments, to pursue and to accentuate the topic in different directions. (This list is far away from being complete!)

   Books: Gohberg, Lancaster, and Rodman [10], Heinig and Rost [33], Lancaster and Tismenetsky [54], Bini and Pan [2], Fuhrmann [6], Pan [62], Lascoux [56].

   Papers: Gohberg, Kaashoek, Lerer, and Rodman [9], Lerer and Tismenetsky [58], Lerer and Rodman [57], Fuhrmann and Datta [7], Gohberg and Shalom [11], Emiris and Mourrain [4], Mourrain and Pan [59].

References

[1] E. Bezout. Recherches sur le degré des équations résultants de l'évanouissement des inconnues, et sur les moyens qu'il convenient d'employer pour trouver ces equations. *Mem. Acad. Roy. Sci. Paris*, pages 288–338, 1764.

[2] D. Bini and V.Y. Pan. *Polynomial and matrix computations. Vol.* 1. Progress in Theoretical Computer Science. Birkhäuser Boston Inc., Boston, MA, 1994. Fundamental algorithms.

[3] A. Cayley. Note sur la méthode d'élimination de Bezout. *J. Reine Angew. Math.*, 53:366–367, 1857.

[4] I.Z. Emiris and B. Mourrain. Matrices in elimination theory. *J. Symbolic Comput.*, 28(1-2):3–44, 1999. Polynomial elimination – algorithms and applications.

[5] L. Euler. *Introductio in analysin infinitorum. Tomus primus*. Sociedad Andaluza de Educación Matemática "Thales", Seville, 2000. Reprint of the 1748 original.

[6] P.A. Fuhrmann. *A polynomial approach to linear algebra*. Universitext. Springer-Verlag, New York, 1996.

[7] P.A. Fuhrmann and B.N. Datta. On Bezoutians, Vandermonde matrices, and the Liénard-Chipart stability criterion. In *Proceedings of the Fourth Haifa Matrix Theory Conference (Haifa, 1988)*, volume 120, pages 23–37, 1989.

[8] F.R. Gantmacher. *Matrizentheorie*. Hochschulbücher für Mathematik [University Books for Mathematics], 86. VEB Deutscher Verlag der Wissenschaften, Berlin, 1986. With a foreword by D.P. Želobenko, Translated from the Russian by Helmut Boseck, Dietmar Soyka and Klaus Stengert.

[9] I. Gohberg, M.A. Kaashoek, L. Lerer, and L. Rodman. Common multiples and common divisors of matrix polynomials. II. Vandermonde and resultant matrices. *Linear and Multilinear Algebra*, 12(3):159–203, 1982/83.

[10] I. Gohberg, P. Lancaster, and L. Rodman. *Matrix polynomials*. Academic Press Inc. [Harcourt Brace Jovanovich Publishers], New York, 1982. Computer Science and Applied Mathematics.

[11] I. Gohberg and T. Shalom. On Bezoutians of nonsquare matrix polynomials and inversion of matrices with nonsquare blocks. *Linear Algebra Appl.*, 137/138:249–323, 1990.

[12] I.C. Gohberg and G. Heinig. The inversion of finite Toeplitz matrices. *Mat. Issled.*, 8(3(29)):151–156, 183, 1973.

[13] I.C. Gohberg and G. Heinig. Inversion of finite Toeplitz matrices consisting of elements of a noncommutative algebra. *Rev. Roumaine Math. Pures Appl.*, 19:623–663, 1974.

[14] I.C. Gohberg and G. Heinig. The resultant matrix and its generalizations. I. The resultant operator for matrix polynomials. *Acta Sci. Math.* (*Szeged*), 37:41–61, 1975.

[15] I.C. Gohberg and G. Heinig. The resultant matrix and its generalizations. II. The continual analogue of the resultant operator. *Acta Math. Acad. Sci. Hungar.*, 28(3-4):189–209, 1976.

[16] G. Heinig. The notion of Bezoutian and of resultant for operator pencils. *Funkcional. Anal. i Priložen.*, 11(3):94–95, 1977.

[17] G. Heinig. Über Block-Hankelmatrizen und den Begriff der Resultante für Matrixpolynome. *Wiss. Z. Techn. Hochsch. Karl-Marx-Stadt*, 19(4):513–519, 1977.

[18] G. Heinig. Verallgemeinerte Resultantenbegriffe bei beliebigen Matrixbüscheln. I. Einseitiger Resultantenoperator. *Wiss. Z. Tech. Hochsch. Karl-Marx-Stadt*, 20(6): 693–700, 1978.

[19] G. Heinig. Verallgemeinerte Resultantenbegriffe bei beliebigen Matrixbüscheln. II. Gemischter Resultantenoperator. *Wiss. Z. Tech. Hochsch. Karl-Marx-Stadt*, 20(6):701–703, 1978.

[20] G. Heinig. Bezoutiante, Resultante und Spektralverteilungsprobleme für Operatorpolynome. *Math. Nachr.*, 91:23–43, 1979.

[21] G. Heinig. Generalized resultant operators and classification of linear operator pencils up to strong equivalence. In *Functions, series, operators, Vol. I, II* (*Budapest,* 1980), volume 35 of *Colloq. Math. Soc. János Bolyai*, pages 611–620. North-Holland, Amsterdam, 1983.

[22] G. Heinig. Chebyshev-Hankel matrices and the splitting approach for centrosymmetric Toeplitz-plus-Hankel matrices. *Linear Algebra Appl.*, 327(1-3):181–196, 2001.

[23] G. Heinig and F. Hellinger. On the Bezoutian structure of the Moore-Penrose inverses of Hankel matrices. *SIAM J. Matrix Anal. Appl.*, 14(3):629–645, 1993.

[24] G. Heinig and F. Hellinger. Displacement structure of generalized inverse matrices. *Linear Algebra Appl.*, 211:67–83, 1994. Generalized inverses (1993).

[25] G. Heinig and F. Hellinger. Moore-Penrose inversion of square Toeplitz matrices. *SIAM J. Matrix Anal. Appl.*, 15(2):418–450, 1994.

[26] G. Heinig and U. Jungnickel. Zur Lösung von Matrixgleichungen der Form $AX - XB = C$. *Wiss. Z. Tech. Hochsch. Karl-Marx-Stadt*, 23(4):387–393, 1981.

[27] G. Heinig and U. Jungnickel. On the Routh-Hurwitz and Schur-Cohn problems for matrix polynomials and generalized Bezoutians. *Math. Nachr.*, 116:185–196, 1984.

[28] G. Heinig and U. Jungnickel. On the Bezoutian and root localization for polynomials. *Wiss. Z. Tech. Hochsch. Karl-Marx-Stadt*, 27(1):62–65, 1985.

[29] G. Heinig and U. Jungnickel. Hankel matrices generated by Markov parameters, Hankel matrix extension, partial realization, and Padé-approximation. In *Operator theory and systems* (*Amsterdam,* 1985), volume 19 of *Oper. Theory Adv. Appl.*, pages 231–253. Birkhäuser, Basel, 1986.

[30] G. Heinig and U. Jungnickel. Hankel matrices generated by the Markov parameters of rational functions. *Linear Algebra Appl.*, 76:121–135, 1986.

[31] G. Heinig and U. Jungnickel. Lyapunov equations for companion matrices. *Linear Algebra Appl.*, 76:137–147, 1986.

[32] G. Heinig and K. Rost. *Algebraic methods for Toeplitz-like matrices and operators*, volume 19 of *Mathematical Research*. Akademie-Verlag, Berlin, 1984.

[33] G. Heinig and K. Rost. *Algebraic methods for Toeplitz-like matrices and operators*, volume 13 of *Operator Theory: Advances and Applications*. Birkhäuser Verlag, Basel, 1984.

[34] G. Heinig and K. Rost. On the inverses of Toeplitz-plus-Hankel matrices. *Linear Algebra Appl.*, 106:39–52, 1988.

[35] G. Heinig and K. Rost. Inversion of matrices with displacement structure. *Integral Equations Operator Theory*, 12(6):813–834, 1989.

[36] G. Heinig and K. Rost. Matrices with displacement structure, generalized Bezoutians, and Moebius transformations. In *The Gohberg anniversary collection, Vol. I* (*Calgary, AB,* 1988), volume 40 of *Oper. Theory Adv. Appl.*, pages 203–230. Birkhäuser, Basel, 1989.

[37] G. Heinig and K. Rost. Matrix representations of Toeplitz-plus-Hankel matrix inverses. *Linear Algebra Appl.*, 113:65–78, 1989.

[38] G. Heinig and K. Rost. DFT representations of Toeplitz-plus-Hankel Bezoutians with application to fast matrix-vector multiplication. *Linear Algebra Appl.*, 284(1-3):157–175, 1998. ILAS Symposium on Fast Algorithms for Control, Signals and Image Processing (Winnipeg, MB, 1997).

[39] G. Heinig and K. Rost. Representations of Toeplitz-plus-Hankel matrices using trigonometric transformations with application to fast matrix-vector multiplication. In *Proceedings of the Sixth Conference of the International Linear Algebra Society* (*Chemnitz,* 1996), volume 275/276, pages 225–248, 1998.

[40] G. Heinig and K. Rost. Hartley transform representations of inverses of real Toeplitz-plus-Hankel matrices. In *Proceedings of the International Conference on Fourier Analysis and Applications* (*Kuwait,* 1998), volume 21, pages 175–189, 2000.

[41] G. Heinig and K. Rost. Hartley transform representations of symmetric Toeplitz matrix inverses with application to fast matrix-vector multiplication. *SIAM J. Matrix Anal. Appl.*, 22(1):86–105 (electronic), 2000.

[42] G. Heinig and K. Rost. Representations of inverses of real Toeplitz-plus-Hankel matrices using trigonometric transformations. In *Large-scale scientific computations of engineering and environmental problems, II* (*Sozopol,* 1999), volume 73 of *Notes Numer. Fluid Mech.*, pages 80–86. Vieweg, Braunschweig, 2000.

[43] G. Heinig and K. Rost. Efficient inversion formulas for Toeplitz-plus-Hankel matrices using trigonometric transformations. In *Structured matrices in mathematics,*

computer science, and engineering, II (*Boulder, CO,* 1999), volume 281 of *Contemp. Math.*, pages 247–264. Amer. Math. Soc., Providence, RI, 2001.

[44] G. Heinig and K. Rost. Centro-symmetric and centro-skewsymmetric Toeplitz matrices and Bezoutians. *Linear Algebra Appl.*, 343/344:195–209, 2002. Special issue on structured and infinite systems of linear equations.

[45] G. Heinig and K. Rost. Centrosymmetric and centro-skewsymmetric Toeplitz-plus-Hankel matrices and Bezoutians. *Linear Algebra Appl.*, 366:257–281, 2003. Special issue on structured matrices: analysis, algorithms and applications (Cortona, 2000).

[46] G. Heinig and K. Rost. Fast algorithms for centro-symmetric and centro-skewsymmetric Toeplitz-plus-Hankel matrices. *Numer. Algorithms*, 33(1-4):305–317, 2003. International Conference on Numerical Algorithms, Vol. I (Marrakesh, 2001).

[47] G. Heinig and K. Rost. Split algorithms for Hermitian Toeplitz matrices with arbitrary rank profile. *Linear Algebra Appl.*, 392:235–253, 2004.

[48] G. Heinig and K. Rost. Split algorithms for centrosymmetric Toeplitz-plus-Hankel matrices with arbitrary rank profile. In *The extended field of operator theory*, volume 171 of *Oper. Theory Adv. Appl.*, pages 129–146. Birkhäuser, Basel, 2007.

[49] C. Hermite. Extrait d'une lettre de Mr. Ch. Hermite de Paris à Mr. Borchard de Berlin, sur le nombre des racines d'une équation algébrique comprises entre des limites données. *J. Reine Angew. Math.*, 52:39–51, 1856.

[50] A.S. Householder. Bezoutiants, elimination and localization. *SIAM Rev.*, 12:73–78, 1970.

[51] C. Jacobi. De eliminatione variablis e duabus aequatione algebraicis. *J. Reine Angew. Math.*, 15:101–124, 1836.

[52] M.G. Krein and M.A. Naimark. The method of symmetric and Hermitian forms in the theory of the separation of the roots of algebraic equations. *Linear and Multilinear Algebra*, 10(4):265–308, 1981. Translated from the Russian by O. Boshko and J.L. Howland.

[53] B. Krishna and H. Krishna. Computationally efficient reduced polynomial based algorithms for Hermitian Toeplitz matrices. *SIAM J. Appl. Math.*, 49(4):1275–1282, 1989.

[54] P. Lancaster and M. Tismenetsky. *The theory of matrices*. Computer Science and Applied Mathematics. Academic Press Inc., Orlando, FL, second edition, 1985.

[55] F.I. Lander. The Bezoutian and the inversion of Hankel and Toeplitz matrices (in Russian). *Mat. Issled.*, 9(2 (32)):69–87, 249–250, 1974.

[56] A. Lascoux. *Symmetric functions and combinatorial operators on polynomials*, volume 99 of *CBMS Regional Conference Series in Mathematics*. Published for the Conference Board of the Mathematical Sciences, Washington, DC, 2003.

[57] L. Lerer and L. Rodman. Bezoutians of rational matrix functions. *J. Funct. Anal.*, 141(1):1–36, 1996.

[58] L. Lerer and M. Tismenetsky. The Bezoutian and the eigenvalue-separation problem for matrix polynomials. *Integral Equations Operator Theory*, 5(3):386–445, 1982.

[59] B. Mourrain and V.Y. Pan. Multivariate polynomials, duality, and structured matrices. *J. Complexity*, 16(1):110–180, 2000. Real computation and complexity (Schloss Dagstuhl, 1998).

[60] A. Olshevsky and V. Olshevsky. Kharitonov's theorem and Bezoutians. *Linear Algebra Appl.*, 399:285–297, 2005.

[61] V. Olshevsky and L. Sakhnovich. An operator identities approach to bezoutians. A general scheme and examples. In *Proc.of the MTNS' 04 Conference*. 2004.

[62] V.Y. Pan. *Structured matrices and polynomials*. Birkhäuser Boston Inc., Boston, MA, 2001. Unified superfast algorithms.

[63] M.M. Postnikov. *Ustoichivye mnogochleny* (*Stable polynomials*). "Nauka", Moscow, 1981.

[64] K. Rost. Toeplitz-plus-Hankel Bezoutians and inverses of Toeplitz and Toeplitz-plus-Hankel matrices. *Oper. Matrices*, 2(3):385–406, 2008.

[65] I. Sylvester. On a theory of syzygetic relations of two rational integral functions, comprising an application to the theory of Sturm's functions, and that of the greatest algebraical common measure. *Phylos. Trans. Roy. Soc. London*, 143:407–548, 1853.

[66] H.K. Wimmer. On the history of the Bezoutian and the resultant matrix. *Linear Algebra Appl.*, 128:27–34, 1990.

Karla Rost
Dept. of Mathematics
Chemnitz University of Technology
Reichenhainer Straße 39
D-09126 Chemnitz, Germany
e-mail: karla.rost@mathematik.tu-chemnitz.de

Operator Theory:
Advances and Applications, Vol. 199, 119–123

On Matrices that are not Similar to a Toeplitz Matrix and a Family of Polynomials

Tewodros Amdeberhan and Georg Heinig

Abstract. A conjecture from the second author's paper [Linear Algebra Appl., 332–334 (2001) 519–531] concerning a family of polynomials is proved and strengthened. A consequence of this is that for any $n > 4$ there is an $n \times n$ matrix that is not similar to a Toeplitz matrix, which was proved before for odd n and $n = 6, 8, 10$.

Mathematics Subject Classification (2000). Primary 15A21; Secondary 15A18.

Keywords. Toeplitz matrix; Jordan normal form; Inverse eigenvalue problem.

1. Introduction

In the paper [4] D.S. Mackey, N. Mackey and S. Petrovic posed and studied the inverse Jordan structure problem for complex Toeplitz matrices. They showed, in particular, that every $n \times n$ complex nonderogatory matrix is similar to an upper Hessenberg Toeplitz matrix, with ones on the subdiagonal. Such a choice guarantees uniqueness of the unit upper Hessenberg Toeplitz matrix. This result was recently extended by Willmer [6], who showed that a block companion matrix is similar to a unique block unit Hessenberg matrix.

The authors [4] also investigated the problem of what happens if the nonderogatority condition is dropped and asked the question, *"Is every complex matrix similar to a Toeplitz matrix?"* This poses the inverse Jordan structure problem for Toeplitz matrices – which Jordan forms are achievable by Toeplitz matrices. Then, [4] gave an affirmative answer to this question for matrices of order $n \leq 4$ and conjectured that this might be true for all n. It is worth noting that the inverse eigenvalue question for *real* symmetric $n \times n$ Toeplitz matrices was posed in 1983 by Delsarte and Genin [1] and resolved by them for $n \leq 4$; the general case was settled only recently by Landau [3]. Landau's non-constructive proof uses topological degree theory to show that any list of n real numbers can be realized as the spectrum of an $n \times n$ real symmetric Toeplitz matrix.

In [2] the second author of the present note showed that there are matrices that are not similar to a Toeplitz matrix. Examples for such matrices are

$$\bigoplus_{j=1}^{m}(S_2 \oplus c) \qquad \text{and} \qquad \bigoplus_{j=1}^{m-2}(S_2 \oplus S_3)$$

for all $m > 1$ and $c \neq 0$. Here S_k denotes the $k \times k$ matrix of the forward shift, i.e.,

$$S_2 = \begin{bmatrix} 0 & 0 \\ 1 & 0 \end{bmatrix}, \quad S_3 = \begin{bmatrix} 0 & 0 & 0 \\ 1 & 0 & 0 \\ 0 & 1 & 0 \end{bmatrix},$$

and $\oplus$ stands for the direct sum. Note that the order of the first set of these matrices is $2m + 1$ and the second matrix is nilpotent. That means that for any odd integer $n > 4$ there is an $n \times n$ matrix that is not similar to a Toeplitz matrix.

For even n the problem is more complicated. Candidates for matrices that are not similar to a Toeplitz matrix are

$$\bigoplus_{j=1}^{m-1}(S_2 \oplus 0 \oplus c) \qquad \text{and} \qquad \bigoplus_{j=1}^{m-2}(S_2 \oplus S_3 \oplus 0)\,, \tag{1.1}$$

where $c \neq 0$ and $m > 2$. It was proved in [2] that these matrices are really not similar to a Toeplitz for $m = 3, 4, 5$, that means for matrices of order 6, 8 and 10. For the general case the problem was reduced to the property of a class of polynomials defined as follows:

$$p_0(t) = p_1(t) = 1\,, \quad p_2(t) = t\,, \quad p_j(t) = -\frac{1}{2}\sum_{k=1}^{j-1} p_k(t)p_{j-k}(t) \quad (j > 2)\,. \tag{1.2}$$

It was shown that the matrices (1.1) are not similar to a Toeplitz matrix if the following is true.

Condition 1.1 ([2], p. 528). *For $m > 3$, the system of $m - 2$ equations*

$$p_{m+2}(t) = p_{m+3}(t) = \cdots = p_{2m-1}(t) = 0$$

has only the trivial solution $t = 0$.

In the present note we show that this condition is always satisfied. Even more, the following is shown, which is the main result of the paper.

Theorem 1.2. *For $m > 1$, $p_{m+1}(t) = p_m(t) = 0$ has only the trivial solution $t = 0$.*

A consequence of this theorem is the following.

Corollary 1.3. *For any $m > 4$ there is an $m \times m$ matrix that is not similar to a Toeplitz matrix.*

2. On a family of polynomials

First we compute the generating function of the family of polynomials $\{p_j(t)\}$ defined by (1.2), which is

$$p(z,t) = \sum_{j=0}^{\infty} p_j(t) z^j .$$

Lemma 2.1. *The generating function $p(z,t)$ is given by*

$$p(z,t) = \big(1 + 2z + z^2(2t+1)\big)^{1/2} . \tag{2.1}$$

Proof. According to the definition of $p_j(t)$ we have

$$\sum_{i+k=j} p_i(t) p_k(t) = 0$$

for $j > 2$. That means that the coefficients of z^j in the expansion of $(p(z,t))^2$ in powers of z vanish if $j > 2$. Hence $p(z,t)^2$ is a quadratic polynomial in z, i.e., $p(z,t) = A(t) + B(t)z + C(t)z^2$. Taking the definition of $p_j(t)$ for $j = 0, 1, 2$ into account we obtain

$$A(t) = 1, \quad B(t) = 2, \quad C(t) = 2t + 1 ,$$

which completes the proof. □

Expanding $p(z,t)$ in powers of z we obtain the following explicit representation of $p_j(t)$[1]:

$$p_j(t) = \sum_{k=0}^{\lfloor j/2 \rfloor} 2^{j-2k} \binom{1/2}{j-k} \binom{j-k}{k} (2t+1)^k , \tag{2.2}$$

where $\lfloor j/2 \rfloor$ is the integer part of $j/2$.

The key for proving Theorem 1.2 is the following lemma.

Lemma 2.2. *The polynomials $p_j(t)$ $(j = 0, 1, \dots)$ satisfy the 3-term recursion*

$$(j+2)p_{j+2}(t) + (2j+1)p_{j+1}(t) + (j-1)(2t+1)p_j(t) = 0. \tag{2.3}$$

Proof. Let $h(z,t)$ denote the generating function of the polynomial family $\{p_j(t)\}$ defined by (2.3) with initial conditions $p_0(t) = p_1(t) = 1$. We show that $h(z,t) = p(z,t)$. Let h' denote the partial derivative of $h(z,t)$ by z and $h = h(z,t)$.

We have

$$\sum_{j=0}^{\infty} (j+2) p_{j+2} z^{j+1} = h' - 1 , \qquad \sum_{j=0}^{\infty} (2j+1) p_{j+1} z^{j+1} = 2zh' - h + 1 ,$$

$$\sum_{j=0}^{\infty} (j-1) p_j z^{j+1} = z^2 h' - zh .$$

[1] A typo in [2] p. 528 is corrected here. The expression is never used to affect the results of [2].

Summing up we obtain the ordinary differential equation

$$(1+2z+(2t+1)z^2)h' - (1+(2t+1)z)h = 0\ .$$

As it is easily checked, the generating function $p(z,t)$ also satisfies this equation. Since $p(0,t) = h(0,t)$, we conclude that $p(z,t) = h(z,t)$. □

An alternative way to prove the lemma is to employ the explicit expression (2.2) for $p_j(t)$. This appears in the Appendix section.

Proof of Theorem 1.2. The theorem can be proved now by induction in a standard fashion. The base case, $m = 2$, is evident since $p_2 = t = p_3 = 0$ iff $t = 0$. Assume the theorem is valid for $m > 1$, then we claim the same is true for $m+1$. Suppose not! i.e., $p_{m+2}(\tau) = p_{m+1}(\tau) = 0$ for some $\tau \neq 0$. Then Lemma 2.2 implies that $\tau = -\frac{1}{2}$. Once again, make application of the recurrence (2.3) but this time re-index m by $m-1$ to get

$$(m+1)p_{m+1}(\tau) + (2m-1)p_m(\tau) + (m-2)(2\tau+1)p_{m-2}(\tau) = 0. \tag{2.4}$$

So, $p_m(-\frac{1}{2}) = 0$. Hence both p_{m+1} and p_m vanish at $-\frac{1}{2}$. This contradiction to the induction step proves the theorem. □

Let us finally mention two consequences of our result. The following is immediate from Theorem 1.2 where variables are switched $w = \frac{b}{2a}z$ and the value $t = \frac{4ac}{b^2} - 1$ is selected. The case $b = 0$ is treated separately. It is important that $t \neq 0$.

Corollary 2.3. *Let $f(w) = (a + bw + cw^2)^{\frac{1}{2}}$, where $a \neq 0$ and $b^2 - 4ac \neq 0$, and $f(w) = \sum_{k=0}^{\infty} f_k z^k$ be its Maclaurin expansion. Then for all j, f_j and f_{j+1} cannot both vanish.*

The following is an equivalent formulation of Condition 1.1.

Corollary 2.4. *For $n > 4$ there is no polynomial $P(t)$ of degree n such that $P(t)^2 = q(t) + t^{2n-1}r(t)$ for quadratic polynomials $q(t)$ and $r(t)$, except for the trivial cases $P(t) = a + bt$ and $P(t) = at^{n-1} + bt^n$.*

Proof. Compare proof of Lemma 6.1 in [2] where the polynomials $p_j(t)$ take the place of u_k. Then, convert u_k via u_k/u_1^k. □

3. Appendix

We show a scheme on how to arrive at the recursion

$$(j+2)p_{j+2}(t) + (2j+1)p_{j+1}(t) + (j-1)(2t+1)p_j(t) = 0 \tag{3.1}$$

for the explicit expression

$$p_j(t) = \sum_{k=0}^{\lfloor j/2 \rfloor} 2^{j-2k} \binom{1/2}{j-k} \binom{j-k}{k} (2t+1)^k$$

of the sequence $\{p_j(t)\}_j$. The idea utilizes the so-called *Wilf-Zeilberger* (WZ) method of proof [5].

Let $F(j,k) := 2^j \binom{1/2}{j-k}\binom{j-k}{k}(2t+1)^k$, and $G(j,k) := -2\frac{(j-1)(2j-2k-1)k}{(j+1-2k)(j+2-2k)}F(j,k)$.

Then one can check, preferably using a symbolic software, that

$$(j+2)F(j+2,k)+(2j+1)F(j+1,k)+(j-1)(2t+1)F(j,k) = G(j,k+1)-G(j,k).$$

Telescoping: Sum over all $-\infty < k < \infty$ and observe that

$$\sum_{k=-\infty}^{\infty} F(j,k) = \sum_{k=0}^{\lfloor j/2 \rfloor} F(j,k) = p_j(t) \quad \text{while} \quad \sum_{k=-\infty}^{\infty} G(j,k+1) = \sum_{k=-\infty}^{\infty} G(j,k),$$

since $G(j,k)$ has compact support. Then assertion (3.1) follows.

Acknowledgment

The first author gratefully acknowledges the wonderful support rendered, at onset of this project, by the DIMACS center at Rutgers University. He also takes this opportunity to commemorate the second author as a great *Mensch*.

References

[1] Philippe Delsarte, Yves V. Genin. Spectral properties of finite Toeplitz matrices. *in Mathematical Theory of networks and Systems: Proc. MTNS-83 Int. Symp., Beer Sheva, Israel, June* 1983*, P.A. Fhrmann, ed., Lecture Notes in Control and Information Sciences, New York, Springer-Verlag*, 58:194–213, 1984.

[2] Georg Heinig. Not every matrix is similar to a Toeplitz matrix. *Linear Algebra Appl.*, 332–334:519–531, 2001.

[3] Henry J Landau. The inverse eigenvalue problem for real symmetric Toeplitz matrices. *J. Amer. Math. Soc.*, 7:749–767, 1994.

[4] D. Steven Mackey, Niloufer Mackey, and Srdjan Petrovic. Is every matrix similar to a Toeplitz matrix? *Linear Algebra Appl.* 297:87–105, 1999.

[5] Marko Petkovsek, Herbert Wilf, Doron Zeilberger. $A = B$ Toeplitz matrices. *A.K. Peters Ltd., USA*, 1996.

[6] Harald K. Wimmer. Similarity of block companion and block Toeplitz matrices. *Linear Algebra Appl.* 343–344:381–387, 2002.

Tewodros Amdeberhan
Mathematics
Tulane University
New Orleans, LA 70118, USA
e-mail: tamdeber@tulane.edu

Part II

Research Contributions

Operator Theory:
Advances and Applications, Vol. 199, 127–154

A Traub-like Algorithm for Hessenberg-quasiseparable-Vandermonde Matrices of Arbitrary Order

T. Bella, Y. Eidelman, I. Gohberg, V. Olshevsky,
E. Tyrtyshnikov and P. Zhlobich

Abstract. Although Gaussian elimination uses $\mathcal{O}(n^3)$ operations to invert an arbitrary matrix, matrices with a special Vandermonde structure can be inverted in only $\mathcal{O}(n^2)$ operations by the *fast* Traub algorithm. The original version of Traub algorithm was numerically unstable although only a minor modification of it yields a high accuracy in practice. The Traub algorithm has been extended from Vandermonde matrices involving monomials to polynomial-Vandermonde matrices involving real orthogonal polynomials, and the Szegö polynomials.

In this paper we consider a new more general class of polynomials that we suggest to call Hessenberg order m quasiseparable polynomials, or (H, m)-quasiseparable polynomials. The new class is wide enough to include all of the above important special cases, e.g., monomials, real orthogonal polynomials and the Szegö polynomials, as well as new subclasses. We derive a fast $\mathcal{O}(n^2)$ Traub-like algorithm to invert the associated (H, m)-*quasiseparable-Vandermonde matrices*.

The class of *quasiseparable matrices* is garnering a lot of attention recently; it has been found to be useful in designing a number of fast algorithms. The derivation of our new Traub-like algorithm is also based on exploiting quasiseparable structure of the corresponding Hessenberg matrices. Preliminary numerical experiments are presented comparing the algorithm to standard structure ignoring methods.

This paper extends our recent results in [6] from the $(H, 0)$- and $(H, 1)$-quasiseparable cases to the more general (H, m)-quasiseparable case.

Mathematics Subject Classification (2000). 15A09, 15-04, 15B05.

Keywords. Orthogonal polynomials, Szego polynomials, quasiseparable matrices, Vandermonde matrices, Hessenberg matrices, inversion, fast algorithm.

1. Introduction. Polynomial-Vandermonde matrices and quasiseparable matrices

1.1. Inversion of polynomial-Vandermonde matrices

In this paper we consider the problem of inverting the class of polynomial-Vandermonde matrices. For a set of n distinct nodes $\{x_k\}_{k=1}^n$, the classical Vandermonde matrix $V(x) = \left[\ x_i^{j-1}\ \right]$ is known to be invertible (provided the nodes are distinct). One can generalize this structure by evaluating a different basis (other than the monomials) at the nodes in the following way. That is, for a set of n polynomials $R = \{r_0(x), r_1(x), \ldots, r_{n-1}(x)\}$ satisfying $\deg r_k(x) = k$, the matrix of the form

$$V_R(x) = \begin{bmatrix} r_0(x_1) & r_1(x_1) & \cdots & r_{n-1}(x_1) \\ r_0(x_2) & r_1(x_2) & \cdots & r_{n-1}(x_2) \\ \vdots & \vdots & & \vdots \\ r_0(x_n) & r_1(x_n) & \cdots & r_{n-1}(x_n) \end{bmatrix} \tag{1.1}$$

is called a polynomial-Vandermonde matrix. It is clear that such a matrix is invertible if and only if the chosen nodes are distinct. Indeed, let T be an invertible, upper triangular matrix, and consider the product $V(x) \cdot T$. The effect of post-multiplication by T is that the entries of the product are polynomials determined by the columns of T, evaluated at the given nodes. Hence this is an alternate definition of a polynomial-Vandermonde matrix as the product of invertible matrices (provided the nodes are distinct).

In the simplest case where $R = \{1, x, x^2, \ldots, x^{n-1}\}$ (i.e., when $T = I$), the matrix $V_R(x)$ reduces to a classical Vandermonde matrix and the inversion algorithm is due to Traub [28]. It was observed in [16] that a minor modification of the original Traub algorithm results in very good accuracy.

The structure-ignoring approach of Gaussian elimination for inversion of $V_R(x)$ requires $\mathcal{O}(n^3)$ operations, and for a general matrix $V_R(x)$ (i.e., no special recurrence relations satisfied by the polynomial system R involved), the algorithm derived in this paper also requires $\mathcal{O}(n^3)$ operations. However, in several special cases, the structure has been exploited, resulting in fast algorithms that can compute the n^2 entries of the inverse in only $\mathcal{O}(n^2)$ operations. It also allows the construction of fast system solvers; one of the pioneering works in this area belongs to Björck and Pereyra [2]. Table 1 lists the previous work in deriving fast inversion algorithms and fast system solvers for various special cases of the polynomial system R.

1.2. Capturing recurrence relations via confederate matrices

To generalize the inversion algorithms of Table 3 we will use the concept of a *confederate matrix* introduced in [21]. Let polynomials $R = \{r_0(x),\ r_1(x)\ , \ldots,\ r_n(x)\}$ be specified by the general n-term recurrence relations[1]

$$r_k(x) = (\alpha_k x - a_{k-1,k}) \cdot r_{k-1}(x) - a_{k-2,k} \cdot r_{k-2}(x) - \cdots - a_{0,k} \cdot r_0(x), \quad \alpha_k \neq 0 \tag{1.2}$$

[1]It is easy to see that any polynomial system $\{r_k(x)\}$ satisfying $\deg r_k(x) = k$ obeys (1.2).

TABLE 1. Fast $\mathcal{O}(n^2)$ algorithms for polynomial-Vandermonde matrices.

Matrix $V_R(x)$	Polynomial System R	$\mathcal{O}(n^2)$ inversion	$\mathcal{O}(n^2)$ system solver
Classical Vdm	monomials	Traub [28]	Björck–Pereyra [2]
Chebyshev–Vdm	Chebyshev plns	Gohberg–Ols [14]	Reichel–Opfer [26]
Three-term Vdm	Real orthogonal plns	Calvetti–Reichel [7]	Higham [19]
Szegö–Vdm	Szegö plns	Olshevsky [23]	BEGKO [3]

Vdm = Vandermonde; Ols = Olshevsky; plns = polynomials

for $k > 0$, and r_0 is a constant. Define for the polynomial

$$P(x) = P_0 \cdot r_0(x) + P_1 \cdot r_1(x) + \cdots + P_{n-1} \cdot r_{n-1}(x) + P_n \cdot r_n(x) \tag{1.3}$$

its *confederate matrix* (with respect to the polynomial system R) by

$$C_R(P) = \underbrace{\begin{bmatrix} \frac{a_{01}}{\alpha_1} & \frac{a_{02}}{\alpha_2} & \frac{a_{03}}{\alpha_3} & \cdots & \frac{a_{0,k}}{\alpha_k} & \cdots & \cdots & \frac{a_{0,n}}{\alpha_n} \\ \frac{1}{\alpha_1} & \frac{a_{12}}{\alpha_2} & \frac{a_{13}}{\alpha_3} & \cdots & \frac{a_{1,k}}{\alpha_k} & \cdots & \cdots & \frac{a_{1,n}}{\alpha_n} \\ 0 & \frac{1}{\alpha_2} & \frac{a_{23}}{\alpha_3} & \cdots & \vdots & \cdots & \cdots & \frac{a_{2,n}}{\alpha_n} \\ 0 & 0 & \frac{1}{\alpha_3} & \ddots & \frac{a_{k-2,k}}{\alpha_k} & \ddots & & \vdots \\ \vdots & \vdots & \ddots & \ddots & \frac{a_{k-1,k}}{\alpha_k} & \ddots & \ddots & \vdots \\ \vdots & \vdots & \ddots & \ddots & \frac{1}{\alpha_k} & \ddots & \ddots & \vdots \\ \vdots & \vdots & \ddots & \ddots & \ddots & \ddots & \ddots & \vdots \\ 0 & 0 & \cdots & \cdots & \cdots & 0 & \frac{1}{\alpha_{n-1}} & \frac{a_{n-1,n}}{\alpha_n} \end{bmatrix}}_{C_R(r_n)} - \begin{bmatrix} P_0 \\ P_1 \\ P_2 \\ \vdots \\ \vdots \\ P_{n-1} \end{bmatrix} \begin{bmatrix} 0 & \cdots & 0 & \frac{1}{\alpha_n P_n} \end{bmatrix} \tag{1.4}$$

In the special case where $P(x) = r_n(x)$, we have $P_0 = P_1 = \cdots = P_{n-1} = 0$, and hence the last term on the right-hand side of (1.4) vanishes.

Notice that the coefficients of the recurrence relations for the k^{th} polynomial $r_k(x)$ from (1.2) are contained in the k^{th} column of $C_R(r_n)$, as the highlighted column shows. We refer to [21] for many useful properties of the confederate matrix and only recall here that

$$r_k(x) = \alpha_0 \cdot \alpha_1 \cdot \ldots \cdot \alpha_k \cdot \det(xI - [C_R(P)]_{k\times k}),$$

and

$$P(x) = \alpha_0 \cdot \alpha_1 \cdots \alpha_n \cdot P_n \cdot \det(xI - C_R(P)),$$

where $[C_R(P)]_{k\times k}$ denotes the $k \times k$ leading submatrix of $C_R(P)$ in the special case where $P(x) = r_n(x)$.

Next in Table 2 we list confederate matrices for the polynomial systems[2] of Table 3.

TABLE 2. Systems of polynomials and corresponding recurrence relations.

Polynomial System Recurrence relations	Corresponding confederate matrix $C_R(r_n)$
monomials $r_k(x) = x \cdot r_{k-1}(x)$	$\begin{bmatrix} 0 & 0 & \cdots & \cdots & 0 \\ 1 & \ddots & \ddots & & 0 \\ 0 & \ddots & \ddots & \ddots & \vdots \\ \vdots & \ddots & \ddots & \ddots & \vdots \\ 0 & \cdots & 0 & 1 & 0 \end{bmatrix}$ lower shift matrix
Chebyshev polynomials $r_k(x) = 2x \cdot r_{k-1}(x) - r_{k-2}(x)$	$\begin{bmatrix} 0 & \frac{1}{2} & \cdots & \cdots & 0 \\ \frac{1}{2} & \ddots & \ddots & & 0 \\ 0 & \ddots & \ddots & \ddots & \vdots \\ \vdots & \ddots & \ddots & \ddots & \frac{1}{2} \\ 0 & \cdots & 0 & \frac{1}{2} & 0 \end{bmatrix}$ tridiagonal matrix
real orthogonal polynomials $r_k(x) = (\alpha_k x - \delta_k) r_{k-1}(x)$ $-\gamma_k \cdot r_{k-2}(x)$	$\begin{bmatrix} \frac{\delta_1}{\alpha_1} & \frac{\gamma_2}{\alpha_2} & 0 & \cdots & 0 \\ \frac{1}{\alpha_1} & \frac{\delta_2}{\alpha_2} & \ddots & \ddots & \vdots \\ 0 & \frac{1}{\alpha_2} & \ddots & \frac{\gamma_{n-1}}{\alpha_{n-1}} & 0 \\ \vdots & \ddots & \ddots & \frac{\delta_{n-1}}{\alpha_{n-1}} & \frac{\gamma_n}{\alpha_n} \\ 0 & \cdots & 0 & \frac{1}{\alpha_{n-1}} & \frac{\delta_n}{\alpha_n} \end{bmatrix}$ tridiagonal matrix
Szegö polynomials[3] $\begin{bmatrix} \phi_k(x) \\ \phi_k^{\#}(x) \end{bmatrix} = \frac{1}{\mu_k} \begin{bmatrix} 1 & -\rho_k^* \\ -\rho_k & 1 \end{bmatrix} \times \begin{bmatrix} \phi_{k-1}(x) \\ x\phi_{k-1}^{\#}(x) \end{bmatrix}$	$\begin{bmatrix} -\rho_1\rho_0^* & \cdots & -\rho_{n-1}\mu_{n-2}\dots\mu_1\rho_0^* & -\rho_n\mu_{n-1}\dots\mu_1\rho_0^* \\ \mu_1 & \ddots & -\rho_{n-1}\mu_{n-2}\dots\mu_2\rho_1^* & -\rho_n\mu_{n-1}\dots\mu_2\rho_1^* \\ 0 & \ddots & \vdots & \vdots \\ \vdots & & -\rho_{n-1}\rho_{n-2}^* & -\rho_n\mu_{n-1}\rho_{n-2}^* \\ 0 & \cdots & \mu_{n-1} & -\rho_n\rho_{n-1}^* \end{bmatrix}$ unitary Hessenberg matrix matrix

[2]For the monomials, Chebyshev polynomials and real orthogonal polynomials the structure of the confederate matrices can be immediately deduced from their recurrence relations. For Szegö polynomials it is also well known, see, e.g., [23] and the references therein.

It turns out that all matrices of Table 2 are special cases of the more general class of matrices defined next. It is this larger class of matrices, and the class of polynomials related to them via (1.2) that we consider in this paper.

1.3. Main tool: quasiseparable matrices and polynomials

Definition 1.1. (Quasiseparable matrices and polynomials)

- *A matrix A is called (H, m)-quasiseparable (i.e., Hessenberg lower part and order m upper part) if* **(i)** *it is strongly upper Hessenberg (i.e., nonzero first subdiagonal, $a_{i+1,i} \neq 0$), and* **(ii)** $\max(\operatorname{rank} A_{12}) = m$, *where the maximum is taken over all symmetric partitions of the form*

$$A = \left[\begin{array}{c|c} * & A_{12} \\ \hline * & * \end{array}\right]$$

- *Let $A = [a_{ij}]$ be a (H, m)-quasiseparable matrix. For $\alpha_i = 1/a_{i+1,i}$, then the system of polynomials related to A via*

$$r_k(x) = \alpha_1 \cdots \alpha_k \det (xI - A)_{(k\times k)} .$$

is called a system of (H, m)-quasiseparable polynomials.

Remark 1.2. The class of (H, m)-quasiseparable polynomials is wide enough to include monomials, Chebyshev polynomials, real orthogonal and Szegö polynomials (i.e., all polynomials of Tables 3 and 2) as special cases. This can be seen by inspecting, for each confederate matrix, its typical submatrix A_{12} from the partition described in Definition 1.1.

- **The lower shift matrix is $(H, 0)$-quasiseparable** Indeed, if A is such a matrix, then any submatrix A_{12} is simply a zero matrix.
- **Tridiagonal matrices are also $(H, 1)$-quasiseparable**. Indeed, if A is tridiagonal, then the submatrix A_{12} has the form $(\gamma_j/\alpha_j)e_k e_1^T$, which can easily be observed to have rank one.
- **Unitary Hessenberg matrices are $(H, 1)$-quasiseparable**. Indeed, if A corresponds to the Szegö polynomials, then the corresponding $3\times(n-1)$ submatrix A_{12} has the form

$$\begin{bmatrix} -\rho_k\mu_{k-1}\cdots\mu_3\mu_2\mu_1\rho_0^* & -\rho_{k-1}\mu_{k-2}\cdots\mu_3\mu_2\mu_1\rho_0^* & \cdots & -\rho_n\mu_{n-1}\cdots\mu_3\mu_2\mu_1\rho_0^* \\ -\rho_k\mu_{k-1}\cdots\mu_3\mu_2\rho_1^* & -\rho_{k-1}\mu_{k-2}\cdots\mu_3\mu_2\rho_1^* & \cdots & -\rho_n\mu_{n-1}\cdots\mu_3\mu_2\rho_1^* \\ -\rho_k\mu_{k-1}\cdots\mu_3\rho_2^* & -\rho_{k-1}\mu_{k-2}\cdots\mu_3\rho_2^* & \cdots & -\rho_n\mu_{n-1}\cdots\mu_3\rho_2^* \end{bmatrix},$$

which is also rank 1 since the rows are scalar multiples of each other. The same is true for all other symmetric partitions of A.

[3]It is known that, under the additional restriction of $\rho_k \neq 0$ for each k, the corresponding Szegö polynomials satisfy the three-term recurrence relations

$$\phi_0^{\#}(x) = \tfrac{1}{\mu_0}, \quad \phi_1^{\#}(x) = \tfrac{1}{\mu_1}(x \cdot \phi_0^{\#}(x) + \rho_1\rho_0^* \cdot \phi_0^{\#}(x))$$
$$\phi_k^{\#}(x) = \left[\tfrac{1}{\mu_k} \cdot x + \tfrac{\rho_k}{\rho_{k-1}}\tfrac{1}{\mu_k}\right] \phi_{k-1}^{\#}(x) - \tfrac{\rho_k}{\rho_{k-1}}\tfrac{\mu_{k-1}}{\mu_k} \cdot x \cdot \phi_{k-2}^{\#}(x).$$

Hence all of the polynomials corresponding to the confederate matrices listed above are $(H,1)$-quasiseparable polynomials.

1.4. Main problem: Inversion of (H,m)-quasiseparable-Vandermonde matrices

As shown in the previous remark, $(H,0)$- and $(H,1)$-quasiseparable polynomials generalize the previous cases of monomials, real orthogonal polynomials, and Szegö polynomials. In the paper [6], an algorithm for inversion of $(H,0)$- and $(H,1)$-quasiseparable-Vandermonde matrices is derived, and hence that algorithm is applicable to these special cases. However there are important cases not covered by either $(H,0)$- or $(H,1)$-quasiseparable polynomials. An example of such a system of polynomials is given next.

Example 1.3 (l-recurrent polynomials). From (1.4) it follows that if polynomials satisfy l-term recurrence relations

$$r_k(x) = (\alpha_k x - a_{k-1,k})\cdot r_{k-1}(x) - a_{k-2,k}\cdot r_{k-2}(x) - \cdots - a_{k-(l-1),k}\cdot r_{k-(l-1)}(x) \quad (1.5)$$

then their confederate matrices

$$A = \begin{bmatrix} \frac{a_{0,1}}{\alpha_1} & \cdots & \frac{a_{0,l-1}}{\alpha_{l-1}} & 0 & \cdots & 0 \\ \frac{1}{\alpha_1} & \frac{a_{1,2}}{\alpha_2} & \cdots & \frac{a_{1,l}}{\alpha_l} & \ddots & \vdots \\ 0 & \frac{1}{\alpha_2} & & & \ddots & 0 \\ \vdots & \ddots & \ddots & \ddots & & \frac{a_{n-(l-1),n}}{\alpha_n} \\ \vdots & & \ddots & \frac{1}{\alpha_{n-2}} & & \vdots \\ 0 & \cdots & \cdots & 0 & \frac{1}{\alpha_{n-1}} & \frac{a_{n-1,n}}{\alpha_n} \end{bmatrix} \quad (1.6)$$

are $(1,l-2)$-banded, i.e., they have only one nonzero subdiagonal and $l-2$ nonzero superdiagonals. Considering a typical element A_{12} of the partition of Definition 1.1, in this case for a 5×5, $(1,2)$-banded example, we have

$$A = \left[\begin{array}{c|c} * & A_{12} \\ \hline * & * \end{array}\right] = \left[\begin{array}{cc|ccc} \frac{a_{0,1}}{\alpha_1} & \frac{a_{0,2}}{\alpha_2} & \frac{a_{0,3}}{\alpha_3} & 0 & 0 \\ \frac{1}{\alpha_1} & \frac{a_{1,2}}{\alpha_2} & \frac{a_{1,3}}{\alpha_3} & \frac{a_{1,4}}{\alpha_4} & 0 \\ \hline 0 & \frac{1}{\alpha_2} & \frac{a_{2,3}}{\alpha_3} & \frac{a_{2,4}}{\alpha_4} & \frac{a_{2,5}}{\alpha_5} \\ 0 & 0 & \frac{1}{\alpha_3} & \frac{a_{3,4}}{\alpha_4} & \frac{a_{3,5}}{\alpha_5} \\ 0 & 0 & 0 & \frac{1}{\alpha_4} & \frac{a_{4,5}}{\alpha_5} \end{array}\right]$$

One can see that any A_{12} of the partition of Definition 1.1 has rank at most 2, implying that A is a $(H,2)$-quasiseparable matrix by definition. More generally, a system of l-recurrent polynomials are $(H,l-2)$-quasiseparable (and so polynomials satisfying three-term recurrence relations are $(H,1)$-quasiseparable).

A Björck–Pereyra-like algorithm for solving linear systems with Hessenberg-quasiseparable-Vandermonde coefficient matrices was proposed in [4], and this algorithm is applicable for any order of quasiseparability (i.e., for any $m \geqslant 1$ of (H,m)-quasiseparable matrix). A Traub-like inversion algorithm is derived in [6], but it is valid only for $(H,0)$- and $(H,1)$-quasiseparable-Vandermonde matrices.

In this paper we extend the Traub-like algorithm to a corresponding algorithm for an arbitrary order of quasiseparability. Previous work in this area, including that using quasiseparable matrices, is given next in Table 3.

TABLE 3. Fast $\mathcal{O}(n^2)$ algorithms for polynomial-Vandermonde matrices.

Matrix $V_R(x)$	Polynomial System R	$\mathcal{O}(n^2)$ inversion	$\mathcal{O}(n^2)$ system solver
Classical Vdm	monomials	Traub [28]	Björck–Pereyra [2]
Chebyshev–Vdm	Chebyshev plns	Gohberg–Ols [14]	Reichel–Opfer [26]
Three-term Vdm	Real orthogonal plns	Calvetti–Reichel [7]	Higham [19]
Szegö–Vdm	Szegö plns	Ols [23]	BEGKO [3]
$(H,1)$-qsep-Vdm	$(H,1)$-qsep	BEGOT [6]	BEGKO [4]
(H,m)-qsep-Vdm	(H,m)-qsep	this paper	

Vdm = Vandermonde; Ols = Olshevsky;
plns = polynomials; qsep = quasiseparable

This new inversion algorithm is applicable to the special cases of polynomial-Vandermonde matrices for monomials, real orthogonal polynomials, and Szegö polynomials, which are themselves special cases of (H,m)-quasiseparable polynomials. In addition, it is also applicable to new classes of polynomials for which no Traub-like algorithm is currently available. One such class of polynomials are those satisfying the motivating recurrence relations (1.5).

As was the case for the Traub-like algorithm for $(H,0)$- and $(H,1)$-quasiseparable-Vandermonde matrices of [6], the proposed Traub-like algorithm for (H,m)-quasiseparable-Vandermonde matrices is fast, requiring only $\mathcal{O}(n^2)$ operations by exploiting the sparse recurrence relations (1.5).

2. Inversion formula

In this section we recall the formula that will be used to invert a polynomial-Vandermonde matrix as in (1.1). Such a matrix is completely determined by n polynomials $R = \{r_0(x), \ldots, r_{n-1}(x)\}$ and n nodes $x = (x_1, \ldots, x_n)$. The desired inverse $V_R(x)^{-1}$ is given by the formula

$$V_R(x)^{-1} = \tilde{I} \cdot V_{\tilde{R}}^T(x) \cdot \operatorname{diag}(c_1, \ldots, c_n), \tag{2.1}$$

(see [22], [23]) where

$$c_i = \prod_{\substack{k=1 \\ k \neq i}}^{n} (x_k - x_i)^{-1}, \tag{2.2}$$

$\tilde{I}$ is the antidiagonal matrix

$$\tilde{I} = \begin{bmatrix} 0 & \cdots & 0 & 1 \\ \vdots & \iddots & 1 & 0 \\ 0 & \iddots & \iddots & \vdots \\ 1 & 0 & \cdots & 0 \end{bmatrix}, \tag{2.3}$$

and $\widehat{R}$ is the system of *associated (generalized Horner) polynomials*, defined as follows: if we define the *master polynomial* $P(x)$ by $P(x) = (x - x_1) \cdots (x - x_n)$, then for the polynomial system $R = \{r_0(x), \ldots, r_{n-1}(x), P(x)\}$, the associated polynomials $\widehat{R} = \{\widehat{r}_0(x), \ldots, \widehat{r}_{n-1}(x), P(x)\}$ are those satisfying the relations

$$\frac{P(x) - P(y)}{x - y} = \sum_{k=0}^{n-1} r_k(x) \cdot \widehat{r}_{n-k-1}(y), \tag{2.4}$$

see [20]. A discussion showing the existence of polynomials satisfying these relations (2.4) for any polynomial system R is given in [3]. This definition can be seen as a generalization of the Horner polynomials associated with the monomials, cf. with the discussion in Section 2.1 below.

This discussion gives a relation between the inverse $V_R(x)^{-1}$ and the polynomial-Vandermonde matrix $V_{\widehat{R}}(x)$, where $\widehat{R}$ is the system of polynomials associated with R. To use this in order to invert $V_R(x)$, one needs to evaluate the polynomials $\widehat{R}$ at the nodes x to form $V^T_{\widehat{R}}(x)$. Such evaluation can be done by using confederate matrices (defined in Section 1.2) associated with system of polynomials, which will be discussed in the next section, but at this point the formula (2.1) allows us to present a sketch of the Traub-like inversion algorithm next. The detailed algorithm will be provided in Section 6 below after deriving next several formulas that will be required to implement its steps 2 and 3.

Algorithm 2.1 (A sketch of the Traub-like inversion algorithm).

1. *Compute the entries of diag*$(c_1, \ldots, c_n)$ *via* (2.2).
2. *Compute the coefficients* $\{P_0, P_1, \ldots, P_n\}$ *of the master polynomial* $P(x)$ *as in* (1.3).
3. *Evaluate the* n *polynomials of* $\widehat{R}$ *with confederate matrix specified via* (2.8) *at the* n *nodes* x_k *to form* $V_{\widehat{R}}(x)$.
4. *Compute the inverse* $V_R(x)^{-1}$ *via* (2.1).

2.1. The key property of all Traub-like algorithms: pertransposition

In this section we use the classical Traub algorithm to explain the key property used in deriving a Traub-like algorithms in terms of quasiseparable ranks of confederate matrices of systems of polynomials. According to (1.3), let

$$P(x) = P_0 + P_1 \cdot x + \cdots + P_{n-1} \cdot x^{n-1} + x^n$$

be a polynomial in the monomial base. The monomials $R = \{1, x, x^2, \ldots, x^{n-1}\}$ satisfy the obvious recurrence relations $x^k = x \cdot x^{k-1}$ and hence the confederate

matrix (1.4) of $P(x)$ with respect to R becomes

$$C_R(P) = \begin{bmatrix} 0 & 0 & \cdots & 0 & -P_0 \\ 1 & 0 & \cdots & 0 & -P_1 \\ 0 & 1 & \ddots & \vdots & \vdots \\ \vdots & \ddots & \ddots & 0 & \vdots \\ 0 & \cdots & 0 & 1 & -P_{n-1} \end{bmatrix} \tag{2.5}$$

which is the well-known companion matrix. By Definition 1.1 its leading submatrices $[C_R(P)]_{k\times k}$ are $(H,0)$-quasiseparable for $k = 1 \ldots n-1$. Hence monomials are $(H,0)$-quasiseparable polynomials.

From the well-known recurrence relations for the Horner polynomials (which invert the classical Vandermonde matrix, see [28])

$$\widehat{r}_0(x) = 1, \quad \widehat{r}_k(x) = x \cdot \widehat{r}_{k-1}(x) + P_{n-k}, \tag{2.6}$$

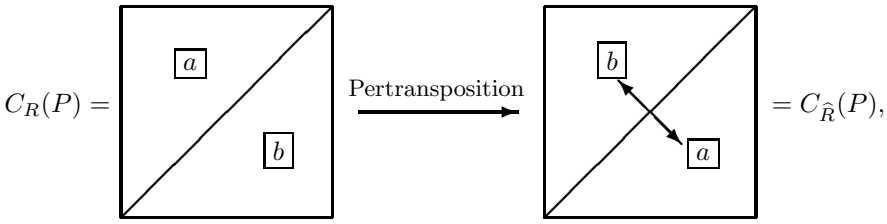
we obtain the confederate matrix

$$C_{\widehat{R}}(P) = \begin{bmatrix} -P_{n-1} & -P_{n-2} & \cdots & -P_1 & -P_0 \\ 1 & 0 & \cdots & 0 & 0 \\ 0 & 1 & \ddots & \vdots & 0 \\ \vdots & \ddots & \ddots & 0 & \vdots \\ 0 & \cdots & 0 & 1 & 0 \end{bmatrix}. \tag{2.7}$$

for the Horner polynomials. This relation between the confederate matrices of R and $\widehat{R}$ can be seen as

$$C_{\widehat{R}}(P) = \tilde{I} \cdot C_R(P)^T \cdot \tilde{I}, \tag{2.8}$$

or more visually,



and holds in the general case (see [22], [23]). The passage from $C_R(P)$ to $C_{\widehat{R}}(P)$ in (2.8) is called a *pertransposition*, or reflection across the antidiagonal. We will show later that recurrence relations for a given system of polynomials together with (2.8) allow fast evaluation of the polynomials $\widehat{R}$ at the nodes x, a required step for a fast Traub-like algorithm.

The leading submatrices of the matrix (2.7) are easily seen to be Hessenberg and quasiseparable[4] but due to the perturbation $\{P_0, P_1, \ldots, P_{n-1}\}$ their quasiseparable ranks are 1. Hence the Horner polynomials are $(H, 1)$-quasiseparable; that is, the quasiseparable rank increases by one due to the inclusion of the perturbation terms in each principal submatrix. Analogously, for an arbitrary system of (H, m)-quasiseparable polynomials the order of quasiseparability increases by at most one, which we state as follows.

Remark 2.2. For R a system of (H, m)-quasiseparable polynomials, when passing to the system $\widehat{R}$ of polynomials associated with R, the order of quasiseparability may increase by one. That is, the system $\widehat{R}$ is either (H, m)-quasiseparable or $(H, m + 1)$-quasiseparable.

This property is used by all of the previous Traub-like algorithms given in Table 3. The quasiseparable rank of a confederate matrix after pertransposition is increased only by at most one. This allows derivation of cheap recurrence relations for the system of polynomials $\widehat{R}$ and use them in computing $V_{\widehat{R}}$.

In summary, the classical Traub algorithm was based on deriving formulas for systems of polynomials as the confederate matrices changed from $(H, 0)$-quasiseparable to $(H, 1)$-quasiseparable, and subsequent Traub-like algorithms were based on changes from $(H, 1)$-quasiseparable to $(H, 2)$-quasiseparable. Such derivations were already very involved.

In order to derive a Traub-like algorithm in the more general case considered in this paper, we need to **(i)** attain the formulas for the original (H, m)-quasiseparable polynomials which, unlike previous cases, are not readily available[5], and **(ii)** have a derivation allowing us to pass from (H, m)-quasiseparable to $(H, m + 1)$-quasiseparable confederate matrices. Hence our plan for the next two sections is to solve the problems (i) and (ii) listed above, respectively.

3. Recurrence relations for (H, m)-quasiseparable polynomials

Real orthogonal polynomials are typically defined in terms of the recurrence relations they satisfy, and then these recurrence relations are used to give equivalent definitions in terms of Hessenberg matrices, as in Table 2. Currently we have only the latter definition for (H, m)-quasiseparable polynomials, i.e., in terms of the related (H, m)-quasiseparable matrix. Since the main tool is designing fast algorithms is the recurrence relations, as in (2.6), it is the goal of this section to derive a set of sparse recurrence relations satisfied by (H, m)-quasiseparable polynomials.

[4]Pertransposition changes the order in which the submatrices A_{12} of Definition 1.1 appear and transposes them, but does not change their ranks, hence the quasiseparability is preserved.
[5]In the $(H, 1)$-quasiseparable case, these formulas were derived in [6].

3.1. Generators of quasiseparable matrices

We begin with an equivalent definition of quasiseparability in terms of *generators*. Such generators are the compressed representation of a quasiseparable matrix; that is, the $\mathcal{O}(n)$ entries of the generators define the $\mathcal{O}(n^2)$ entries of the matrix. Operations with generators are the key in designing various types of fast algorithms. In this case, the sparse recurrence relations for (H, m)-quasiseparable polynomials will be given in terms of these generators.

Definition 3.1 (Generator definition for (H, m)-quasiseparable matrices). *A matrix A is called (H, m)-quasiseparable if* **(i)** *it is strongly upper Hessenberg (i.e., nonzero first subdiagonal, $a_{i+1,i} \neq 0$), and* **(ii)** *it can be represented in the form*

$$A = \begin{array}{|ccccc|}\hline d_1 & & & & g_i b_{ij}^{\times} h_j \\ & \ddots & & & \\ p_2 q_1 & & \ddots & & \\ & & & \ddots & \\ & 0 & & p_n q_{n-1} & d_n \\ \hline \end{array} \tag{3.1}$$

where $b_{ij}^{\times} = b_{i+1} \cdots b_{j-1}$ for $j > i+1$ and $b_{ij}^{\times} = 1$ for $j = i+1$. The elements

$$\{p_k, q_k, d_k, g_k, b_k, h_k\},$$

called the generators of the matrix A, are matrices of sizes

	p_k	q_k	d_k	g_k	b_k	h_k
sizes	1×1	1×1	1×1	$1 \times u_k$	$u_{k-1} \times u_k$	$u_{k-1} \times 1$
range	$k \in [2, n]$	$k \in [1, n-1]$	$k \in [1, n]$	$k \in [1, n-1]$	$k \in [2, n-1]$	$k \in [2, n]$

subject to $\max_k u_k = m$.

Remark 3.2. For a given (H, m)-quasiseparable matrix the set of generators of Definition 3.1 is not unique. There is a freedom in choosing generators without changing the matrix.

Remark 3.3. It is useful to note that Definition 3.1 together with (1.2) imply that (H, m)-quasiseparable polynomials satisfy n-term recurrence relations

$$r_k(x) = \frac{1}{p_{k+1} q_k} \left[(x - d_k) r_{k-1}(x) - \sum_{j=0}^{k-2} \left(g_{j+1} b_{j+1,k}^{\times} h_k r_j(x) \right) \right]. \tag{3.2}$$

This formula is not sparse and hence expensive. In order to design a fast algorithm, sparse recurrence relations are required, and such are stated and proved in the next section.

3.2. Sparse recurrence relations for (H, m)-quasiseparable polynomials

The next theorem gives, for any (H, m)-quasiseparable matrix, a set of sparse recurrence relations satisfied by the corresponding (H, m)-quasiseparable polynomials. These recurrence relations are given in terms of the generators of the (H, m)-quasiseparable matrix.

Theorem 3.4. *Let A be a (H, m)-quasiseparable matrix specified by the generators $\{p_k, q_k, d_k, g_k, b_k, h_k\}$. Then the polynomial system $R = \{r_k(x)\}_{k=0}^{n}$ corresponding to A (such that $A = C_R(r_n)$) satisfies*

$$\left[\begin{array}{c} F_k(x) \\ \hline r_k(x) \end{array}\right] = \frac{1}{p_{k+1}q_k} \left[\begin{array}{cc|c} p_kq_k & b_k^T & -q_k\, g_k^T \\ \hline p_k\, h_k^T & & x - d_k \end{array}\right] \left[\begin{array}{c} F_{k-1}(x) \\ \hline r_{k-1}(x) \end{array}\right] \tag{3.3}$$

The proof is given at the end of this section, however first some special cases are given in detail.

Example 3.5 ($(H, 0)$- and $(H, 1)$-quasiseparable case). For the case where $m \leqslant 1$, the recurrence relations (3.3) reduce to those derived in [11], which were used in [6] to derive the Traub-like algorithm for $(H, 0)$- and $(H, 1)$-quasiseparable-Vandermonde matrices; that is, of the form

$$\begin{bmatrix} F_k(x) \\ r_k(x) \end{bmatrix} = \begin{bmatrix} \alpha_k & \beta_k \\ \gamma_k & \delta_k x + \theta_k \end{bmatrix} \begin{bmatrix} F_{k-1}(x) \\ r_{k-1}(x) \end{bmatrix}. \tag{3.4}$$

These were referred to as [EGO05]-type recurrence relations in [6], and as the recurrence relations (3.3) are a generalization of these, we refer to (3.3) as [EGO05]-type recurrence relations as well.

A motivating example for considering the larger class of (H, m)-quasiseparable polynomials was their inclusion of the class of l-recurrent polynomials, a class not contained by previous cases.

Example 3.6 (l-recurrent polynomials). By introducing the auxiliary polynomials $f_k^{(1)}(x), \ldots, f_k^{(l-2)}(x)$, the relation (1.5) can be rewritten as

$$\begin{bmatrix} f_k^{(1)}(x) \\ \vdots \\ \vdots \\ \vdots \\ f_k^{(l-2)} \\ r_k(x) \end{bmatrix} = \begin{bmatrix} 0 & 1 & 0 & \ldots & 0 & -a_{k-2,k} \\ \vdots & \ddots & \ddots & \ddots & \vdots & \vdots \\ \vdots & & \ddots & \ddots & 0 & \vdots \\ \vdots & & & \ddots & 1 & \vdots \\ 0 & \ldots & \ldots & \ldots & 0 & -a_{k-(l-1),k} \\ 1 & 0 & \ldots & \ldots & 0 & \alpha_k x - a_{k-1,k} \end{bmatrix} \begin{bmatrix} f_{k-1}^{(1)}(x) \\ \vdots \\ \vdots \\ \vdots \\ f_{k-1}^{(l-2)} \\ r_{k-1}(x) \end{bmatrix},$$

which is the reduction of (3.3) in this special case.

Proof of Theorem 3.4. The recurrence relations (3.3) define a system of polynomials which satisfy the n-term recurrence relations

$$r_k(x) = (\alpha_k x - a_{k-1,k}) \cdot r_{k-1}(x) - a_{k-2,k} \cdot r_{k-2}(x) - \cdots - a_{0,k} \cdot r_0(x) \tag{3.5}$$

for some coefficients $\alpha_k, a_{k-1,k}, \ldots, a_{0,k}$. The proof is presented by showing that these n-term recurrence relations in fact coincide exactly with (3.2), so these coefficients coincide with those of the n-term recurrence relations of the polynomials R. Using relations for $r_k(x)$ and $F_{k-1}(x)$ from (3.3), we have

$$r_k(x) = \frac{1}{p_{k+1}q_k}\left[(x-d_k)r_{k-1}(x) - g_{k-1}h_k r_{k-2}(x) + p_{k-1}h_k^T b_{k-1}^T F_{k-2}(x)\right]. \tag{3.6}$$

Notice that again using (3.3) to eliminate $F_{k-2}(x)$ from the equation (3.6) will result in an expression for $r_k(x)$ in terms of $r_{k-1}(x)$, $r_{k-2}(x)$, $r_{k-3}(x)$, $F_{k-3}(x)$, and $r_0(x)$ without modifying the coefficients of $r_{k-1}(x)$, $r_{k-2}(x)$, or $r_0(x)$. Again applying (3.3) to eliminate $F_{k-3}(x)$ results in an expression in terms of $r_{k-1}(x)$, $r_{k-2}(x)$, $r_{k-3}(x)$, $r_{k-4}(x)$, $F_{k-4}(x)$, and $r_0(x)$ without modifying the coefficients of $r_{k-1}(x)$, $r_{k-2}(x)$, $r_{k-3}(x)$, or $r_0(x)$. Continuing in this way, the n-term recurrence relations of the form (3.5) are obtained without modifying the coefficients of the previous ones.

Suppose that for some $0 < j < k-1$ the expression for $r_k(x)$ is of the form

$$\begin{aligned} r_k(x) = \frac{1}{p_{k+1}q_k}\Big[&(x-d_k)r_{k-1}(x) - g_{k-1}h_k r_{k-2}(x) - \cdots \\ &\cdots - g_{j+1}b_{j+1,k}^{\times} h_k r_j(x) + p_{j+1}h_k^T \left(b_{j,k}^{\times}\right)^T F_j(x)\Big]. \end{aligned} \tag{3.7}$$

Using (3.3) for $F_j(x)$ gives the relation

$$F_j(x) = \frac{1}{p_{j+1}q_j}\left(p_j q_j b_j^T F_{j-1}(x) - q_j g_j^T r_{j-1}(x)\right) \tag{3.8}$$

Inserting (3.8) into (3.7) gives

$$\begin{aligned} r_k(x) = \frac{1}{p_{k+1}q_k}\Big[&(x-d_k)r_{k-1}(x) - g_{k-1}h_k r_{k-2}(x) - \cdots \\ &\cdots - g_j b_{j,k}^{\times} h_k r_{j-1}(x) + p_j h_k^T \left(b_{j-1,k}^{\times}\right)^T F_{j-1}(x)\Big]. \end{aligned}$$

Therefore since (3.6) is the case of (3.7) for $j = k-2$, (3.7) is true for each $j = k-2, k-3, \ldots, 0$, and for $j = 0$, using the fact that $F_0 = 0$ we have

$$r_k(x) = \frac{1}{p_{k+1}q_k}\left[(x-d_k)r_{k-1}(x) - g_{k-1}h_k r_{k-2}(x) - \cdots - g_1 b_{1,k}^{\times} h_k r_0(x)\right]$$

Since these coefficients coincide with (3.2) that are satisfied by the polynomial system R, the polynomials given by (3.3) must coincide with these polynomials. This proves the theorem. □

4. Recurrence relations for polynomials associated with (H, m)-quasiseparable polynomials

In the previous section, sparse recurrence relations for (H, m)-quasiseparable polynomials were derived. However, the classical Traub algorithm is based on the recurrence relations for the Horner polynomials, not the original monomial system. In this section, sparse recurrence relations for the system of polynomials associated with a system of (H, m)-quasiseparable polynomials (i.e., the generalized Horner polynomials) are derived. It is these recurrence relations that form the basis of the Traub-like algorithm.

4.1. Introduction of a perturbation term via pertransposition of confederate matrices

Let $R = \{r_k(x)\}_{k=0}^n$ be a system of (H, m)-quasiseparable polynomials, and $\{x_k\}_{k=1}^n$ a set of distinct nodes. Decomposing the master polynomial into the R basis as

$$\prod_{k=1}^{n}(x - x_k) =: P(x) = P_0 \cdot r_0(x) + P_1 \cdot r_1(x) + \cdots + P_{n-1} \cdot r_{n-1}(x) + P_n \cdot r_n(x)$$

yields the coefficients $P_0, P_1, \ldots, P_n$, and we have

$$C_R(P) = \begin{bmatrix} d_1 & & g_i b_{ij}^{\times} h_j \\ p_2 q_1 & \ddots & \\ & \ddots & \ddots \\ 0 & & p_n q_{n-1} \quad d_n \end{bmatrix} - \frac{1}{P_n} \left[\begin{array}{cc} & P_0 \\ 0 & \vdots \\ & P_{n-1} \end{array}\right] \tag{4.1}$$

Applying (2.8) gives us the confederate matrix for the associated polynomials as

$$C_{\widehat{R}}(P) = \begin{bmatrix} d_n & & g_{n-j} b_{n-j,n-i}^{\times} h_{n-i} \\ p_n q_{n-1} & \ddots & \\ & \ddots & \ddots \\ 0 & & p_2 q_1 \quad d_1 \end{bmatrix} - \frac{1}{P_n} \left[\begin{array}{ccc} P_{n-1} & \cdots & P_0 \\ \hline & 0 & \end{array}\right] \tag{4.2}$$

From this last equation we can see that the n-term recurrence relations satisfied by the associated polynomials $\widehat{R}$ are given by

$$\widehat{r}_k(x) = \frac{1}{\widehat{p}_{k+1}\widehat{q}_k}\Bigg[\underbrace{(x-\widehat{d}_k)\widehat{r}_{k-1}(x) - \sum_{j=0}^{k-2}\left(\widehat{g}_{j+1}\widehat{b}^{\times}_{j+1,k}\widehat{h}_k\widehat{r}_j(x)\right)}_{\text{typical term as in (3.2)}} - \underbrace{\frac{P_{n-k}}{P_n}\widehat{r}_0(x)}_{\text{perturbation term}}\Bigg] \tag{4.3}$$

where, in order to simplify the formulas, we introduce the notation

$$\begin{aligned} \widehat{p}_k &= q_{n-k+1}, \quad \widehat{q}_k = p_{n-k+1}, \\ \widehat{d}_k &= d_{n-k+1}, \quad \widehat{g}_k = h^T_{n-k+1}, \\ \widehat{b}_k &= b^T_{n-k+1}, \quad \widehat{h}_k = g^T_{n-k+1}. \end{aligned} \tag{4.4}$$

We will see in a moment that the *nonzero top row* of the second matrix in (4.2) introduces perturbation terms into all formulas that we derive. These perturbation terms are the cause of the increase in quasiseparable rank by at most one.

4.2. Perturbed [EGO05]-type recurrence relations

The previous section showed how n-term recurrence relations for a system of (H, m)-quasiseparable polynomials change after a rank one perturbation of the first row of the corresponding confederate matrix. This small increase in quasiseparable rank allows construction of the desired sparse recurrence relations for associated polynomials, as presented in the next theorem.

Theorem 4.1 (Perturbed [EGO05]-type recurrence relations). *Let*

$$R = \{r_0(x), \ldots, r_{n-1}(x), P(x)\}$$

be a system of (H, m)-quasiseparable polynomials corresponding to a (H, m)-quasiseparable matrix of size $n \times n$ with generators $\{p_k, q_k, d_k, g_k, b_k, h_k\}$ as in Definition 3.1, with the convention that $q_n = 0, b_n = 0$. Then the system of polynomials $\widehat{R}$ associated with R satisfy the recurrence relations

$$\left[\begin{array}{c|c} F_0(x) & \\ \hline & r_0(x) \end{array}\right] = \left[\begin{array}{c|c} 0 & \\ \hline & P_n \end{array}\right] \tag{4.5}$$

$$\begin{bmatrix} \widehat{F}_k(x) \\ \hline \widehat{r}_k(x) \end{bmatrix} = \underbrace{\frac{1}{\widehat{p}_{k+1}\widehat{q}_k} \left[\begin{array}{c|c} \widehat{p}_k\widehat{q}_k \; \widehat{b}_k^T & -\widehat{q}_k \; \widehat{g}_k^T \\ \hline \widehat{p}_k \; \widehat{h}_k^T & x - \widehat{d}_k \end{array}\right] \begin{bmatrix} \widehat{F}_{k-1}(x) \\ \hline \widehat{r}_{k-1}(x) \end{bmatrix}}_{\text{typical terms}} + \underbrace{\frac{1}{\widehat{p}_{k+1}\widehat{q}_k} \begin{bmatrix} 0 \\ \hline P_{n-k} \end{bmatrix}}_{\text{perturbation term}} \tag{4.6}$$

with the vector of auxiliary polynomials $\widehat{F}_k(x)$*, and the coefficients* $P_k, k = 0, \ldots, n$ *are as defined in* (1.3).

Proof. The recurrence relations (4.6) define a system of polynomials which satisfy the n-term recurrence relations

$$\widehat{r}_k(x) = (\alpha_k x - a_{k-1,k}) \cdot \widehat{r}_{k-1}(x) - a_{k-2,k} \cdot \widehat{r}_{k-2}(x) - \cdots - a_{0,k} \cdot \widehat{r}_0(x) \tag{4.7}$$

for some coefficients $\alpha_k, a_{k-1,k}, \ldots, a_{0,k}$. The proof is presented by showing that these n-term recurrence relations in fact coincide exactly with (4.3), so these coefficients coincide with those of the n-term recurrence relations of the associated polynomials $\widehat{R}$; that is,

$$\begin{aligned} \alpha_k = \frac{1}{\widehat{p}_{k+1}\widehat{q}_k}, \quad a_{k-1,k} = \frac{1}{\widehat{p}_{k+1}\widehat{q}_k}\widehat{d}_k, \quad a_{0,k} = \frac{1}{\widehat{p}_{k+1}\widehat{q}_k}\left(\widehat{g}_1\widehat{b}_{1,k}^{\times}\widehat{h}_k - \frac{P_{n-k}}{P_n}\right) \\ a_{j,k} = \frac{1}{\widehat{p}_{k+1}\widehat{q}_k}\widehat{g}_{j+1}\widehat{b}_{j+1,k}^{\times}\widehat{h}_k, \quad j = 1, \ldots, k-2 \end{aligned} \tag{4.8}$$

Using relations for $\widehat{r}_k(x)$ and $\widehat{F}_{k-1}(x)$ from (4.6), we have

$$\begin{aligned} \widehat{r}_k(x) = \frac{1}{\widehat{p}_{k+1}\widehat{q}_k}\Big[&(x - \widehat{d}_k)\widehat{r}_{k-1}(x) - \widehat{g}_{k-1}\widehat{h}_k\widehat{r}_{k-2}(x) \\ &+ \widehat{p}_{k-1}\widehat{h}_k^T\widehat{b}_{k-1}^T\widehat{F}_{k-2}(x) + \frac{P_{n-k}}{P_n}\widehat{r}_0(x)\Big]. \end{aligned} \tag{4.9}$$

Notice that again using (4.6) to eliminate $\widehat{F}_{k-2}(x)$ from the equation (4.9) will result in an expression for $\widehat{r}_k(x)$ in terms of $\widehat{r}_{k-1}(x)$, $\widehat{r}_{k-2}(x)$, $\widehat{r}_{k-3}(x)$, $\widehat{F}_{k-3}(x)$,

and $\widehat{r}_0(x)$ without modifying the coefficients of $\widehat{r}_{k-1}(x)$, $\widehat{r}_{k-2}(x)$, or $\widehat{r}_0(x)$. Again applying (4.6) to eliminate $\widehat{F}_{k-3}(x)$ results in an expression in terms of $\widehat{r}_{k-1}(x)$, $\widehat{r}_{k-2}(x)$, $\widehat{r}_{k-3}(x)$, $\widehat{r}_{k-4}(x)$, $\widehat{F}_{k-4}(x)$, and $\widehat{r}_0(x)$ without modifying the coefficients of $\widehat{r}_{k-1}(x)$, $\widehat{r}_{k-2}(x)$, $\widehat{r}_{k-3}(x)$, or $\widehat{r}_0(x)$. Continuing in this way, the n-term recurrence relations of the form (4.7) are obtained without modifying the coefficients of the previous ones.

Suppose that for some $0 < j < k-1$ the expression for $\widehat{r}_k(x)$ is of the form

$$\widehat{r}_k(x) = \frac{1}{\widehat{p}_{k+1}\widehat{q}_k}\Bigg[(x-\widehat{d}_k)\widehat{r}_{k-1}(x) - \widehat{g}_{k-1}\widehat{h}_k\widehat{r}_{k-2}(x) - \cdots \\ \cdots - \widehat{g}_{j+1}\widehat{b}^{\times}_{j+1,k}\widehat{h}_k\widehat{r}_j(x) + \widehat{p}_{j+1}\widehat{h}_k^T(\widehat{b}^{\times}_{j,k})^T\widehat{F}_j(x) + \frac{P_{n-k}}{P_n}\widehat{r}_0(x)\Bigg]. \tag{4.10}$$

Using (4.6) for $\widehat{F}_j(x)$ gives the relation

$$\widehat{F}_j(x) = \frac{1}{\widehat{p}_{j+1}\widehat{q}_j}\left(\widehat{p}_j\widehat{q}_j\widehat{b}_j^T\widehat{F}_{j-1}(x) - \widehat{q}_j\widehat{g}_j^T\widehat{r}_{j-1}(x)\right) \tag{4.11}$$

Inserting (4.11) into (4.10) gives

$$\widehat{r}_k(x) = \frac{1}{\widehat{p}_{k+1}\widehat{q}_k}\Bigg[(x-\widehat{d}_k)\widehat{r}_{k-1}(x) - \widehat{g}_{k-1}\widehat{h}_k\widehat{r}_{k-2}(x) - \cdots \\ \cdots - \widehat{g}_j\widehat{b}^{\times}_{j,k}\widehat{h}_k\widehat{r}_{j-1}(x) + \widehat{p}_j\widehat{h}_k^T(\widehat{b}^{\times}_{j-1,k})^T\widehat{F}_{j-1}(x) + \frac{P_{n-k}}{P_n}\widehat{r}_0(x)\Bigg]. \tag{4.12}$$

Therefore since (4.9) is the case of (4.10) for $j = k-2$, (4.10) is true for each $j = k-2, k-3, \ldots, 0$, and for $j = 0$, using the fact that $\widehat{F}_0 = 0$ we have

$$\widehat{r}_k(x) = \frac{1}{\widehat{p}_{k+1}\widehat{q}_k}\Bigg[(x-\widehat{d}_k)\widehat{r}_{k-1}(x) - \widehat{g}_{k-1}\widehat{h}_k\widehat{r}_{k-2}(x) - \cdots \\ \cdots - \widehat{g}_1\widehat{b}^{\times}_{1,k}\widehat{h}_k\widehat{r}_0(x) + \frac{P_{n-k}}{P_n}\widehat{r}_0(x)\Bigg] \tag{4.13}$$

Since these coefficients coincide with those in (4.8) that are satisfied by the associated polynomials, the polynomials given by (4.6) must coincide with the associated polynomials. This proves the theorem. □

4.3. Known special cases of these more general recurrence relations

In this section, recurrence relations valid for the class of polynomials associated with (H, m)-quasiseparable polynomials were derived. Such are needed to provide a Traub-like algorithm. We next demonstrate that these more general recurrence relations reduce as expected in the classical cases. That is, since monomials and real orthogonal polynomials are themselves (H, m)-quasiseparable, the above formulas are valid for those classes as well, and furthermore, the special cases of these formulas are, in fact, the classical formulas for these cases.

Example 4.2 (Classical Traub case: monomials and the Horner polynomials). As shown earlier, the well-known companion matrix (2.5) results when the polynomial

system R is simply a system of monomials. By choosing the generators $p_k = 1, q_k = 1, d_k = 0, g_k = 1, b_k = 1$, and $h_k = 0$, the matrix (4.1) reduces to (2.5), and also (4.2) reduces to the confederate matrix for the Horner polynomials (2.7). In this special case, the perturbed three-term recurrence relations of Theorem 4.1 become

$$\widehat{r}_0(x) = P_n, \quad \widehat{r}_k(x) = x\widehat{r}_{k-1}(x) + P_{n-k}, \tag{4.14}$$

after eliminating the auxiliary polynomials present, coinciding with the known recurrence relations for the Horner polynomials, used in the evaluation of the polynomial

$$P(x) = P_0 + P_1 x + \cdots + P_{n-1}x^{n-1} + P_n x^n. \tag{4.15}$$

Example 4.3 (Calvetti–Reichel case: Real orthogonal polynomials and the Clenshaw rule). Consider the almost tridiagonal confederate matrix

$$C_R(P) = \begin{bmatrix} d_1 & h_2 & 0 & \cdots & 0 & -P_0/P_n \\ q_1 & d_2 & h_3 & \ddots & \vdots & -P_1/P_n \\ 0 & q_2 & d_3 & h_4 & 0 & \vdots \\ 0 & 0 & q_3 & d_4 & \ddots & -P_{n-3}/P_n \\ \vdots & \ddots & \ddots & \ddots & \ddots & h_n - P_{n-2}/P_n \\ 0 & \cdots & 0 & 0 & q_{n-1} & d_n - P_{n-1}/P_n \end{bmatrix}. \tag{4.16}$$

The corresponding system of polynomials R satisfy three-term recurrence relations; for instance, the highlighted column implies

$$r_3(x) = \frac{1}{q_3}(x - d_3)r_2(x) - \frac{h_3}{q_3}r_1(x) \tag{4.17}$$

by the definition of the confederate matrix. Thus, confederate matrices of this form correspond to systems of polynomials satisfying three-term recurrence relations, or systems of polynomials orthogonal on a real interval, and the polynomial $P(x)$. Such confederate matrices can be seen as special cases of our general class by choosing scalar generators, with $p_k = 1, b_k = 0$, and $g_k = 1$, and in this case the matrix (4.1) reduces to (4.16).

With these choices of generators, applying Theorem 4.1 and eliminating the auxiliary polynomials yields the recurrence relations

$$\widehat{r}_k(x) = \frac{1}{q_{n-k}}(x - d_{n-k})\widehat{r}_{k-1}(x) - \frac{q_{n-k+1}}{q_{n-k}}h_{n-k+1}\widehat{r}_{k-2}(x) + \frac{1}{q_{n-k}}P_{n-k} \tag{4.18}$$

which coincides with the Clenshaw rule for evaluating a polynomial in a real orthogonal basis, i.e., of the form

$$P(x) = P_0 r_0(x) + P_1 r_1(x) + \cdots + P_{n-1}r_{n-1}(x) + P_n r_n(x). \tag{4.19}$$

By the discussion of pertransposition in Section 2, recurrence relations for the system of polynomials associated with real orthogonal polynomials can be found

by considering the confederate matrix

$$C_R(P) = \begin{bmatrix} d_n - P_{n-1}/P_n & h_n - P_{n-2}/P_n & -P_{n-3}/P_n & \cdots & -P_1/P_n & -P_0/P_n \\ q_{n-1} & d_{n-1} & h_{n-1} & \ddots & \vdots & \\ 0 & q_{n-2} & d_{n-2} & h_{n-2} & 0 & \vdots \\ 0 & 0 & q_{n-3} & d_{n-3} & \ddots & \\ \vdots & \ddots & \ddots & \ddots & \ddots & h_2 \\ 0 & \cdots & 0 & 0 & q_1 & d_1 \end{bmatrix}. \tag{4.20}$$

obtained by pertransposition of (4.16). Note that the highlighted column corresponds to the full recurrence relation

$$\widehat{r}_3(x) = \frac{1}{q_{n-3}}(x - d_{n-2})\widehat{r}_2(x) - \frac{h_{n-1}}{q_{n-3}}\widehat{r}_1(x) + \frac{1}{q_{n-3}}\frac{P_{n-3}}{P_n}\widehat{r}_0(x) \tag{4.21}$$

Thus our formula generalizes both the Clenshaw rule and the algorithms designed for inversion of three-term-Vandermonde matrices in [7] and [14].

5. Computing the coefficients of the master polynomial

Note that in order to use the recurrence relations of the previous section it is necessary to decompose the master polynomial $P(x)$ into the R basis; that is, the coefficients P_k as in (1.3) must be computed. To this end, an efficient method of calculating these coefficients follows.

It is easily seen that the last polynomial $r_n(x)$ in the system R does not affect the resulting confederate matrix $C_R(P)$. Thus, if

$$\bar{R} = \{r_0(x), \ldots, r_{n-1}(x), xr_{n-1}(x)\},$$

we have $C_R(P) = C_{\bar{R}}(P)$. Decomposing the polynomial $P(x)$ into the $\bar{R}$ basis can be done recursively by setting $r_n^{(0)}(x) = 1$ and then for $k = 0, \ldots, n-1$ updating $r_n^{(k+1)}(x) = (x - x_{k+1}) \cdot r_n^{(k)}(x)$.

The following lemma gives this procedure, and is from [23].

Lemma 5.1 ([23]). *Let $R = \{r_0(x), \ldots, r_n(x)\}$ be given by (1.2), and $f(x) = \sum_{i=1}^{k} a_i \cdot r_i(x)$, where $k < n-1$. Then the coefficients of $x \cdot f(x) = \sum_{i=1}^{k+1} b_i \cdot r_i(x)$ can be computed by*

$$\begin{bmatrix} b_0 \\ \vdots \\ b_k \\ b_{k+1} \\ 0 \\ \vdots \\ 0 \end{bmatrix} = \left[\begin{array}{ccccc|c} & & C_R(r_n) & & & 0 \\ \hline 0 & \cdots & 0 & \frac{1}{\alpha_n} & & 0 \end{array}\right] \begin{bmatrix} a_0 \\ \vdots \\ a_k \\ 0 \\ 0 \\ \vdots \\ 0 \end{bmatrix} \tag{5.1}$$

Proof. It can be easily checked that

$$\begin{aligned} & x \cdot \begin{bmatrix} r_0(x) & r_1(x) & \cdots & r_n(x) \end{bmatrix} \\ & \quad - \begin{bmatrix} r_0(x) & r_1(x) & \cdots & r_n(x) \end{bmatrix} \cdot \left[\begin{array}{ccccc|c} & & C_R(r_n) & & & 0 \\ \hline 0 & \cdots & 0 & \frac{1}{\alpha_n} & & 0 \end{array}\right] \\ & = \begin{bmatrix} 0 & \cdots & 0 & x \cdot r_n(x) \end{bmatrix}. \end{aligned}$$

Multiplying the latter equation by the column of the coefficients we obtain (5.1). □

This lemma suggests the following algorithm for computing coefficients $\{P_0, P_1, \ldots, P_{n-1}, P_n\}$ in

$$\prod_{k=1}^{n} (x - x_k) = P_0 r_0(x) + P_1 r_1(x) + \cdots + P_{n-1} r_{n-1}(x) + P_n r_n(x). \tag{5.2}$$

of the master polynomial.

Algorithm 5.2 (Coefficients of the master polynomial in the R basis).
Cost: *$\mathcal{O}(n \times m(n))$, where $m(n)$ is the cost of multiplication of an $n \times n$ quasiseparable matrix by a vector.*
Input: *A quasiseparable confederate matrix $C_R(r_n)$ and n nodes $x = (x_1, x_2, \ldots, x_n)$.*

1. *Set* $\begin{bmatrix} P_0^{(0)} & \cdots & P_{n-1}^{(0)} & P_n^{(0)} \end{bmatrix} = \begin{bmatrix} 1 & 0 & \cdots & 0 \end{bmatrix}$
2. *For* $k = 1 : n$,

$$\begin{bmatrix} P_0^{(k)} \\ \vdots \\ P_{n-1}^{(k)} \\ P_n^{(k)} \end{bmatrix} = \left(\left[\begin{array}{ccccc|c} & & C_{\bar{R}}(x \cdot r_{n-1}(x)) & & & 0 \\ \hline 0 & \cdots & 0 & 1 & & 0 \end{array}\right] - x_k \cdot I \right) \cdot \begin{bmatrix} P_0^{(k-1)} \\ \vdots \\ P_{n-1}^{(k-1)} \\ P_n^{(k-1)} \end{bmatrix}$$

   where $\bar{R} = \{r_0(x), \ldots, r_{n-1}(x), x r_{n-1}(x)\}$.
3. *Take* $\begin{bmatrix} P_0 & \cdots & P_{n-1} & P_n \end{bmatrix} = \begin{bmatrix} P_0^{(n)} & \cdots & P_{n-1}^{(n)} & P_n^{(n)} \end{bmatrix}$

Output: *Coefficients $\{P_0, P_1, \ldots, P_{n-1}, P_n\}$ such that (5.2) is satisfied.*

It is clear that the computational burden in implementing this algorithm is in multiplication of the matrix $C_{\bar{R}}(r_n)$ by the vector of coefficients. The cost of each such step is $\mathcal{O}(m(n))$, where $m(n)$ is the cost of multiplication of an $n \times n$ quasiseparable matrix by a vector, thus the cost of computing the n coefficients is $\mathcal{O}(n \times m(n))$. Using a fast $\mathcal{O}(n)$ algorithm for multiplication of a quasiseparable matrix by a vector first derived in [9] (or its matrix interpretation of [4]), the cost of this algorithm is $\mathcal{O}(n^2)$.

6. The overall Traub-like algorithm

6.1. Quasiseparable generator input

The main algorithm of this section is the Traub-like algorithm that outputs the inverse of a (H, m)-quasiseparable-Vandermonde matrix. It takes as input the generators $\{p_k, q_k, d_k, g_k, b_k, h_k\}$ of the (H, m)-quasiseparable confederate matrix corresponding to the system of polynomials R.

In this algorithm we will make use of MATLAB notations; for instance $V_{\widehat{R}}(i : j, k : l)$ will refer to the block of $V_{\widehat{R}}(x)$ consisting of rows i through j and columns k through l. For each node x_k we have a vector of auxiliary polynomials $F_{\widehat{R}}(x_k)$. Let us compose a matrix of such vectors $[F_{\widehat{R}}(x_1)|\cdots|F_{\widehat{R}}(x_n)]$ and denote it as $\widehat{F_k}$ on each step.

Algorithm 6.1 (Traub-like inversion algorithm).
Cost: $\mathcal{O}(n^2)$ *operations.*
Input: *Generators* $\{p_k, q_k, d_k, g_k, b_k, h_k\}$ *of a quasiseparable confederate matrix corresponding to a system of polynomials* R *and* n *nodes* $x = (x_1, x_2, \ldots, x_n)$.

1. *Compute the entries of* $diag(c_1, \ldots, c_n)$ *via* (2.2)*:* $c_i = \prod_{\substack{k=1 \\ k\neq i}}^{n} (x_k - x_i)^{-1}$.
2. *Compute the coefficients* $\{P_0, \ldots, P_n\}$ *of the master polynomial* $P(x)$ *as in* (1.3) *via Algorithm* 5.2.
3. *Evaluate the* n *polynomials of* $\widehat{R}$ *specified via* (2.8) *at the* n *nodes* x_k *to form* $V_{\widehat{R}}(x)$*. Theorems* 4.1 *provides an algorithm for this.*
   (a) *Set* $V_{\widehat{R}}(:, 1) = P_n \begin{bmatrix} 1 \\ 1 \\ \vdots \\ 1 \end{bmatrix}$, $\widehat{F_1} = \begin{bmatrix} 0 & \cdots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \cdots & 0 \end{bmatrix}$.
   (b) *For* $k = 1 : n-1$*, compute*

$$V_{\widehat{R}}(:, k+1) = \frac{1}{\widehat{p}_{k+1}\widehat{q}_k} \left(\widehat{p}_k \widehat{F_k} \widehat{h}_k + \left(\begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{bmatrix} - \widehat{d}_k \begin{bmatrix} 1 \\ 1 \\ \vdots \\ 1 \end{bmatrix} \right) V_{\widehat{R}}(:, k) + P_{n-k} \begin{bmatrix} 1 \\ 1 \\ \vdots \\ 1 \end{bmatrix} \right)$$

and

$$F_{\widehat{R}}(:, k+1) = \frac{1}{\widehat{p}_{k+1}\widehat{q}_k} \left(\widehat{p}_k \widehat{q}_k \widehat{b}_k^T \widehat{F_k} - \widehat{q}_k \widehat{g}_k^T V_{\widehat{R}}(:, k)^T \right)$$

Note: The product of two column vectors is understood to be componentwise.

4. *Compute the inverse* $V_R(x)^{-1}$ *via* (2.1)*:*

$$V_R(x)^{-1} = \tilde{I} \cdot V_{\widehat{R}}^T(x) \cdot \mathrm{diag}(c_1, \ldots, c_n)$$

Output: *Entries of* $V_R(x)^{-1}$*, the inverse of the polynomial-Vandermonde matrix.*

6.2. Recurrence relation coefficient input

The previous section provides the Traub-like algorithm, which takes as input the generators of the (H, m)-quasiseparable polynomials involved in forming the quasiseparable-Vandermonde matrix. However, as in the motivating cases of real orthogonal polynomials and Szegö polynomials, problems may be stated in terms of the coefficients of the involved recurrence relations instead of in terms of generators.

In this section, we present a result allowing conversion from the language of recurrence relation coefficients to that of quasiseparable generators. Applying this conversion as a preprocessor, the algorithm of the previous section can then be used for problems stated in terms of recurrence relation coefficients.

Theorem 6.2 (Recurrence relations coefficients ⇒ quasiseparable generators). *Let $R = \{r_k(x)\}_{k=0}^{n}$ be a system of polynomials satisfying the [EGO05]-type two-term recurrence relations* (3.4)*:*

$$\begin{bmatrix} F_k(x) \\ r_k(x) \end{bmatrix} = \begin{bmatrix} \alpha_k & \beta_k \\ \gamma_k & \delta_k x + \theta_k \end{bmatrix} \begin{bmatrix} F_{k-1}(x) \\ r_{k-1}(x) \end{bmatrix}.$$

Then the (H, m)-quasiseparable matrix

$$C_R(r_n) = \tag{6.1}$$

$$\begin{bmatrix}
-\frac{\theta_1}{\delta_1} & -(\frac{1}{\delta_2})\gamma_2\beta_1 & -\frac{1}{\delta_3}\gamma_3\alpha_2\beta_1 & -\frac{1}{\delta_4}\gamma_4\alpha_3\alpha_2\beta_1 & \cdots & -\frac{1}{\delta_n}\gamma_n\alpha_{n-1}\alpha_{n-2}\cdots\alpha_3\alpha_2\beta_1 \\
\frac{1}{\delta_1} & -\frac{\theta_2}{\delta_2} & -\frac{1}{\delta_3}\gamma_3\beta_2 & -\frac{1}{\delta_4}\gamma_4\alpha_3\beta_2 & \cdots & -\frac{1}{\delta_n}\gamma_n\alpha_{n-1}\alpha_{n-2}\cdots\alpha_3\beta_2 \\
0 & \frac{1}{\delta_2} & -\frac{\theta_3}{\delta_3} & -\frac{1}{\delta_4}\gamma_4\beta_3 & \ddots & -\frac{1}{\delta_n}\gamma_n\alpha_{n-1}\cdots\alpha_4\beta_3 \\
0 & 0 & \frac{1}{\delta_3} & -\frac{\theta_4}{\delta_4} & \ddots & \vdots \\
\vdots & \ddots & \ddots & \ddots & \ddots & -\frac{1}{\delta_n}\gamma_n\beta_{n-1} \\
0 & \cdots & 0 & 0 & \frac{1}{\delta_{n-1}} & -\frac{\theta_n}{\delta_n}
\end{bmatrix}$$

with generators

$$d_k = -(\theta_k/\delta_k), \quad (k = 1, \ldots, n), \qquad p_{k+1}q_k = (1/\delta_k), \quad (k = 1, \ldots, n-1),$$

$$g_k = \beta_k^T, \quad (k = 1, \ldots, n-1),$$

$$b_k = \alpha_k^T, \quad (k = 2, \ldots, n-1),$$

$$h_k = -\frac{1}{\delta_k}\gamma_k^T, \quad (k = 2, \ldots, n)$$

is a confederate matrix for the system of polynomials R.

Proof. Inserting the specified choice of generators into the general n-term recurrence relations (3.2), we arrive at

$$\begin{aligned} r_k(x) = (\delta_k x + \theta_k) r_{k-1}(x) + \gamma_k \beta_{k-1} r_{k-2}(x) + \gamma_k \alpha_{k-1} \beta_{k-2} r_{k-3}(x) \\ + \gamma_k \alpha_{k-1} \alpha_{k-2} \beta_{k-3} r_{k-4}(x) + \cdots + \gamma_k \alpha_{k-1} \cdots \alpha_2 \beta_1 r_0(x) \end{aligned} \tag{6.2}$$

It suffices to show that the polynomial system satisfying the two-term recurrence relations also satisfies these n-term recurrence relations. Beginning with

$$r_k(x) = (\delta_k x + \theta_k) r_{k-1}(x) + \gamma_k F_{k-1}(x) \tag{6.3}$$

and using the relation $F_{k-1}(x) = \alpha_{k-1} F_{k-2}(x) + \beta_{k-1} r_{k-2}(x)$, (6.3) becomes

$$r_k(x) = (\delta_k x + \theta_k) r_{k-1}(x) + \gamma_k \beta_{k-1} r_{k-2}(x) + \gamma_k \alpha_{k-1} F_{k-2}(x)$$

and continuing this procedure to obtain n-term recurrence relations. It can easily be checked that this procedure yields exactly (6.2). □

7. Numerical Experiments

The numerical properties of the Traub algorithm and its generalizations (that are the special cases of the algorithm proposed in this paper) were studied by different authors. It was noticed in [16] that a modification of the classical Traub algorithm of [28] can yield high accuracy in certain cases if the algorithm is preceded with the *Leja ordering* of the nodes; that is, ordering such that

$$|x_1| = \max_{1 \leqslant i \leqslant n} |x_i|, \qquad \prod_{j=1}^{k-1} |x_k - x_j| = \max_{k \leqslant i \leqslant n} \prod_{j=1}^{k-1} |x_i - x_j|, \quad k = 2, \ldots, n-1$$

(see [26], [19], [24]) It was noticed in [16] that the same is true for Chebyshev-Vandermonde matrices.

No error analysis was done, but the conclusions of the above authors was that in many cases the Traub algorithm and its extensions can yield much better accuracy than Gaussian elimination, even for very ill-conditioned matrices.

We made our preliminary experiments with the proposed Traub-like algorithm, and our conclusions for the most general case are consistent with the experience of our colleagues made for special cases. The results of experiments with the proposed algorithm yields better accuracy than Gaussian elimination. However, our experiments need to be done for different special cases of (H, m)-quasiseparable polynomials.

The algorithm was implemented in *C* using *Lapack* for all supplementary matrix computations (such as matrix multiplication GEMM). For Gaussian elimination we used the *Lapack* subroutine GESV. To estimate the accuracy of all of the above algorithms we took the output of new Traub-like algorithm in double precision $V_R(x)^{-1}$ as the exact solution.

We compare the forward accuracy of the inverse computed by the algorithm in single precision $\widehat{V_R(x)^{-1}}$ with respect to the inverse computed in double precision, defined by

$$e = \frac{\|\widehat{V_R(x)^{-1}} - V_R(x)^{-1}\|_2}{\|V_R(x)^{-1}\|_2} \tag{7.1}$$

where $V_R^s(x)^{-1}$ is the solution computed by each algorithm in single precision. In the tables, New Algorithm denotes the proposed Traub-like algorithm with Leja ordering, and GESV indicates *Lapack*'s inversion subroutine. Finally, cond(V) denotes the condition number of the matrix V computed via the *Lapack* subroutine GESVD.

Experiment 1. (Random choice of generators) In this experiment, the generators we chosen randomly in $(-1, 1)$, and the nodes x_k were selected equidistant on $(-1, 1)$ via the formula

$$x_k = -1 + 2\left(\frac{k}{n-1}\right), \quad k = 0, 1, \ldots, n-1$$

We test the accuracy of the inversion algorithm for various sizes n and quasiseparable ranks m of matrices generated in this way. Some results are tabulated in Table 4.

Notice that the performance of the proposed inversion algorithm is an improvement over that of *Lapack*'s standard inversion subroutine GESV in this specific case. And in almost all cases relative errors are around e-7, which means that all digits of the errors in single precision coincide with corresponding digits in double precision. There are occasional examples in which the proposed algorithm can lose several decimal digits, but it still outperforms Gaussian elimination.

Experiment 2. (l-recurrent polynomials) In this experiment we consider l-recurrent polynomials

$$r_k(x) = (\alpha_k x - a_{k-1,k}) \cdot r_{k-1}(x) - a_{k-2,k} \cdot r_{k-2}(x) - \cdots - a_{k-(l-1),k} \cdot r_{k-(l-1)}(x) \tag{7.2}$$

by choosing coefficients of (7.2) randomly in $(-1, 1)$, and the nodes x_k equidistant on $(-1, 1)$.

We test the accuracy of the inversion algorithm for various sizes n and number of terms l. We remind the reader that quasiseparable rank of polynomials given by (7.2) is $l - 2$. Table 4 presents some results generated in this way.

8. Conclusions

In this paper we extend the previous work in the area of fast Traub-like inversion algorithms to the general class of (H, m)-quasiseparable-Vandermonde matrices. This generalizes results for Vandermonde, three-term-Vandermonde, Szegö–Vandermonde, and $(H, 1)$-quasiseparable-Vandermonde matrices. Exploiting the quasiseparable structure yields sparse recurrence relations which allow the desired computational speedup, resulting in a fast $\mathcal{O}(n^2)$ algorithm as opposed to Gaussian

TABLE 4. Random generators on $(-1, 1)$. Equidistant nodes on $(-1, 1)$.

n	m	cond(V)	GESV	New Algorithm
10	1	1.8e+007	1.5e-006	1.4e-006
	2	5.6e+007	4.6e-005	5.4e-007
20	1	2.2e+020	5.9e-001	5.0e-007
	2	1.6e+019	2.6e+000	1.9e-007
	3	1.0e+021	5.6e-002	2.3e-006
	4	6.1e+020	2.8e+000	5.5e-006
30	1	7.2e+029	1.2e+000	2.6e-006
	2	3.4e+025	9.2e-001	2.7e-006
	3	2.9e+029	1.0e+000	2.0e-006
	4	7.5e+026	1.0e+000	1.5e-006
	5	5.0e+024	1.0e+000	1.4e-006
	6	2.5e+026	1.0e+000	5.1e-007
40	1	2.1e+034	1.0e+000	1.6e-005
	2	3.3e+033	1.0e+000	1.9e-006
	3	4.1e+029	1.0e+000	1.1e-003
	4	4.5e+028	1.0e+000	3.5e-007
	5	1.2e+031	1.0e+000	7.2e-007
	6	3.5e+032	1.0e+000	3.3e-006
	7	6.0e+027	1.0e+000	1.7e-004
	8	7.8e+031	1.0e+000	5.0e-007
50	1	1.5e+039	1.0e+000	3.9e-007
	2	3.7e+038	1.0e+000	7.6e-001
	3	2.6e+041	1.0e+000	4.5e-004
	4	2.0e+037	1.0e+000	3.8e-001
	5	1.7e+037	1.0e+000	5.7e-007
	6	8.0e+038	1.0e+000	9.2e-005
	7	1.7e+038	1.0e+000	9.3e-007
	8	7.5e+036	1.0e+000	4.7e-007

elimination, which requires $\mathcal{O}(n^3)$ operations. Finally, some numerical experiments were presented that indicate that, under some circumstances, the resulting algorithm can give better performance than Gaussian elimination.

TABLE 5. l-recurrent polynomials. Random coefficients on $(-1, 1)$.

n	l	cond(V)	inv()	TraubQS(Leja)
10	3	9.5e+004	1.5e-004	1.9e-007
	4	1.2e+006	7.1e-004	3.3e-007
20	3	4.3e+013	1.0e+000	2.8e-007
	4	2.2e+013	1.0e+000	3.4e-007
	5	1.5e+012	1.0e+000	5.1e-007
	6	4.1e+011	1.0e+000	2.2e-007
30	3	4.7e+016	1.0e+000	4.2e-007
	4	1.8e+016	1.0e+000	3.2e-007
	5	3.0e+018	1.0e+000	4.4e-007
	6	7.3e+016	1.0e+000	4.1e-007
	7	1.2e+017	1.0e+000	5.1e-007
	8	3.6e+017	1.0e+000	2.8e-007
40	3	8.9e+017	1.0e+000	4.8e-007
	4	1.2e+020	1.0e+000	6.5e-007
	5	2.3e+018	1.0e+000	8.3e-007
	6	2.2e+021	1.0e+000	4.5e-007
	7	2.4e+020	1.0e+000	6.8e-007
	8	1.8e+018	1.0e+000	9.7e-007
	9	8.9e+019	1.0e+000	1.1e-006
	10	8.6e+020	1.0e+000	6.3e-007
50	3	3.6e+018	1.0e+000	2.8e-007
	4	3.1e+019	1.0e+000	2.0e-007
	5	5.3e+019	1.0e+000	6.1e-007
	6	7.8e+019	1.0e+000	4.8e-007
	7	1.8e+020	1.0e+000	1.8e-007
	8	3.6e+019	1.0e+000	4.2e-007
	9	7.0e+019	1.0e+000	5.5e-007
	10	2.2e+020	1.0e+000	8.5e-007
	11	3.9e+021	1.0e+000	1.8e-007
	12	5.3e+020	1.0e+000	5.3e-007

References

[1] M.Bakonyi and T.Constantinescu, *Schur's algorithm and several applications*, in Pitman Research Notes in Mathematics Series, vol. 61, Longman Scientific and Technical, Harlow, 1992.

[2] A. Björck and V. Pereyra, *Solution of Vandermonde Systems of Equations*, Math. Comp., **24** (1970), 893–903.

[3] T. Bella, Y. Eidelman, I. Gohberg, I. Koltracht and V. Olshevsky, *A Björck–Pereyra-type algorithm for Szegö-Vandermonde matrices based on properties of unitary Hessenberg matrices*, Linear Algebra and Applications, Volume 420, Issues 2-3 pp. 634–647, 2007.

[4] T. Bella, Y. Eidelman, I. Gohberg, I. Koltracht and V. Olshevsky *A fast Bjorck–Pereyra like algorithm for solving Hessenberg-quasiseparable-Vandermonde systems*, submitted to SIAM Journal of Matrix Analysis (SIMAX), 2007.

[5] T. Bella, Y. Eidelman, I. Gohberg, V. Olshevsky, *Classifications of three-term and two-term recurrence relations and digital filter structures via subclasses of quasiseparable matrices*, submitted to SIAM Journal of Matrix Analysis (SIMAX), 2007.

[6] T. Bella, Y. Eidelman, I. Gohberg, V. Olshevsky, E. Tyrtyshnikov *Fast Traub-like inversion algorithm for Hessenberg order one quasiseparable Vandermonde matrices*, submitted to Journal of Complexity, 2007.

[7] Calvetti, D. and Reichel, L., *Fast inversion of Vandermonde-like matrices involving orthogonal polynomials*, BIT, 1993.

[8] Y. Eidelman and I. Gohberg, *On a new class of structured matrices*, Integral Equations and Operator Theory, **34** (1999), 293–324.

[9] Y. Eidelman and I. Gohberg, *Linear complexity inversion algorithms for a class of structured matrices*, Integral Equations and Operator Theory, **35** (1999), 28–52.

[10] Y. Eidelman and I. Gohberg, *A modification of the Dewilde-van der Veen method for inversion of finite-structured matrices*, Linear Algebra Appl., **343–344** (2002), 419–450.

[11] Y. Eidelman, I. Gohberg and V. Olshevsky, *Eigenstructure of Order-One-Quasiseparable Matrices. Three-term and Two-term Recurrence Relations*, Linear Algebra and its Applications, Volume 405, 1 August 2005, Pages 1–40.

[12] G. Forney, *Concatenated codes*, The M.I.T. Press, 1966, Cambridge.

[13] L.Y. Geronimus, *Polynomials orthogonal on a circle and their applications*, Amer. Math. Translations, **3** p. 1–78, 1954 (Russian original 1948).

[14] I. Gohberg and V. Olshevsky, *Fast inversion of Chebyshev-Vandermonde matrices*, Numerische Mathematik, **67, No. 1** (1994), 71–92.

[15] I. Gohberg and V. Olshevsky, *A fast generalized Parker-Traub algorithm for inversion of Vandermonde and related matrices*, Journal of Complexity, **13(2)** (1997), 208–234. A short version in Proceedings in *Communications, Computation, Control and Signal Processing: A tribute to Thomas Kailath*, Eds. A. Paulraj, V. Roychowdhury and C. Shaper, Kluwer Academic Publishing, 1996, 205–221.

[16] I. Gohberg and V. Olshevsky, *The fast generalized Parker-Traub algorithm for inversion of Vandermonde and related matrices*, J. of Complexity, 13(2) (1997), 208–234.

[17] U. Grenader and G. Szegö, *Toeplitz forms and Applications*, University of California Press, 1958.

[18] W.G. Horner, *A new method of solving numerical equations of all orders by continuous approximation*, Philos. Trans. Roy. Soc. London, (1819), 308–335.

[19] N.J. Higham, *Stability analysis of algorithms for solving confluent Vandermonde-like systems*, SIAM J. Matrix Anal. Appl., **11**(1) (1990), 23–41.

[20] T. Kailath and V. Olshevsky, *Displacement structure approach to polynomial Vandermonde and related matrices*, Linear Algebra and Its Applications, **261** (1997), 49–90.

[21] J. Maroulas and S. Barnett, Polynomials with respect to a general basis. I. Theory, J. of Math. Analysis and Appl., **72** (1979), 177–194.

[22] V. Olshevsky, *Eigenvector computation for almost unitary Hessenberg matrices and inversion of Szegö-Vandermonde matrices via Discrete Transmission lines*. Linear Algebra and Its Applications, 285 (1998), 37–67.

[23] V. Olshevsky, *Associated polynomials, unitary Hessenberg matrices and fast generalized Parker–Traub and Bjorck–Pereyra algorithms for Szegö-Vandermonde matrices* invited chapter in the book "Structured Matrices: Recent Developments in Theory and Computation," 67–78, (D. Bini, E. Tyrtyshnikov, P. Yalamov., Eds.), 2001, NOVA Science Publ., USA.

[24] V. Olshevsky, *Pivoting for structured matrices and rational tangential interpolation*, in Fast Algorithms for Structured Matrices: Theory and Applications, CONM/323, 1–75, AMS publications, May 2003.

[25] F. Parker, *Inverses of Vandermonde matrices*, Amer. Math. Monthly, **71** (1964), 410–411.

[26] L. Reichel and G. Opfer, *Chebyshev-Vandermonde systems*, Math. of Comp., **57** (1991), 703–721.

[27] J. Stoer, R. Bulirsch, *Introduction to Numerical Analysis*, Springer-Verlag, 1992. 277–301.

[28] J. Traub, *Associated polynomials and uniform methods for the solution of linear problems*, SIAM Review, **8**, No. 3 (1966), 277–301.

T. Bella, V. Olshevsky and P. Zhlobich
Department of Mathematics
University of Connecticut
Storrs CT 06269-3009, USA
e-mail: bella@math.uconn.edu (contact author)
olshevsky@math.uconn.edu

Y. Eidelman and I. Gohberg
School of Mathematical Sciences
Raymond and Beverly Sackler
Faculty of Exact Sciences
Tel Aviv University
Ramat-Aviv 69978, Israel
e-mail: eideyu@post.tau.ac.il
gohberg@post.tau.ac.il

E. Tyrtyshnikov
Institute of Numerical Mathematics
Russian Academy of Sciences
Gubkina Street, 8
Moscow, 119991, Russia
e-mail: tee@inm.ras.ru

Operator Theory:
Advances and Applications, Vol. 199, 155–173

A Fast Algorithm for Approximate Polynomial GCD Based on Structured Matrix Computations

Dario A. Bini and Paola Boito

In Memory of Georg Heinig

Abstract. An $O(n^2)$ complexity algorithm for computing an ϵ-greatest common divisor (gcd) of two polynomials of degree at most n is presented. The algorithm is based on the formulation of polynomial gcd given in terms of resultant (Bézout, Sylvester) matrices, on their displacement structure and on the reduction of displacement structured matrices to Cauchy-like form originally pointed out by Georg Heinig. A Matlab implementation is provided. Numerical experiments performed with a wide variety of test problems, show the effectiveness of this algorithm in terms of speed, stability and robustness, together with its better reliability with respect to the available software.

Mathematics Subject Classification (2000). 68W30, 65F05, 15A23.

Keywords. Cauchy matrices, polynomial gcd, displacement structure, Sylvester matrix, Bézout matrix.

1. Introduction

A basic problem in algebraic computing is the evaluation of polynomial gcd: given the coefficients of two univariate polynomials

$$u(x) = \sum_{i=0}^{n} u_i x^i, \quad v(x) = \sum_{i=0}^{m} v_i x^i,$$

compute the coefficients of their greatest common divisor $g(x)$.

In many applications, input data are represented as floating point numbers or derive from the results of physical experiments or previous computations, so

Supported by MIUR grant 2006017542.

that they are generally affected by errors. If $u(x)$ and $v(x)$ have a nontrivial gcd, it turns out that arbitrarily small perturbations in the coefficients of $u(x)$ and $v(x)$ may transform $u(x)$ and $v(x)$ into relatively prime polynomials. Therefore, it is clear that the concept of gcd is not well suited to deal with applications where data are approximatively known. This is why the notion of approximate gcd, or ϵ-gcd, has been introduced. For more details on this topic we refer the reader to [19] [7],[18], [23] and to the references therein.

We use the following definition where $\|\cdot\|$ denotes the Euclidean norm.

Definition 1.1. *A polynomial $g(x)$ is said to be an* ϵ-divisor *of $u(x)$ and $v(x)$ if there exist polynomials $\hat{u}(x)$ and $\hat{v}(x)$ of degree n and m, respectively, such that $\|u(x)-\hat{u}(x)\| \le \epsilon\|u(x)\|$, $\|v(x)-\hat{v}(x)\| \le \epsilon\|v(x)\|$ and $g(x)$ divides $\hat{u}(x)$ and $\hat{v}(x)$. If $g(x)$ is an ϵ-divisor of maximum degree of $u(x)$ and $v(x)$, then it is called* ϵ-gcd *of $u(x)$ and $v(x)$. The polynomials $p(x) = \hat{u}(x)/g(x)$ and $q(x) = \hat{v}(x)/g(x)$ are called ϵ-cofactors.*

Several algorithms for the computation of an approximate polynomial gcd can be found in the literature; they rely on different techniques, such as the Euclidean algorithm [1], [2], [12], [17], optimization methods [15], SVD and factorization of resultant matrices [5], [4], [23], Padé approximation [3], [18], root grouping [18]. Some of them have been implemented inside numerical/symbolic packages like the algorithm of Zeng [23] in Matlab$^{\mathrm{TM}}$ and the algorithms of Kaltofen [14], of Corless et al. [4], of Labahn and Beckermann [13] in Maple$^{\mathrm{TM}}$. These algorithms have a computational cost of $O(n^3)$ which makes them expensive for moderately large values of n.

Algorithms based on the Euclidean scheme have a typical cost of $O(n^2)$ but they are prone to numerical instabilities; look-ahead strategies can improve the numerical stability with an increase of the complexity to $O(n^3)$. More recently, $O(n^2)$ algorithms have been proposed in [24] and [16]. They are based on the QR factorization of a displacement structured matrix obtained by means of the normal equations. The use of the normal equations generally squares the condition number of the original problem with a consequent deterioration of the stability.

In this paper we present an algorithm for (approximate) gcd computation which has a cost of $O(n^2)$ arithmetic operations and, from the several numerical experiments performed so far, results robust and numerically stable. The algorithm relies on the formulation of the gcd problem given in terms of the Bézout matrix $B(u,v)$ or of the Sylvester matrix $S(u,v)$ associated with the pair of polynomials (u,v), and on their reduction to Cauchy-like matrices by means of unitary transforms. This kind of reduction, which is fundamental for our algorithm, was discovered and analyzed by Georg Heinig in [10] in the case of general Toeplitz-like matrices.

For exact gcd, where $\epsilon = 0$, the degree k_ϵ of the ϵ-gcd coincides with the nullity (i.e., the dimension of the kernel) of $B(u,v)$ and of $S(u,v)$, or equivalently, with the nullity of the Cauchy matrices obtained through the reduction.

Our algorithm can be divided into two stages. In the first stage, from the coefficients of the input polynomials a resultant matrix (Sylvester or Bézout) is computed and reduced to Cauchy-like form. The GKO algorithm of Gohberg, Kailath and Olshevsky [8] for the PLU factorization is applied to the Cauchy-like matrix obtained in this way. The algorithm relies on the pivoting strategy and on a suitable technique used to control the growth of the generators. This algorithm is rank-revealing since in exact arithmetic it provides a matrix U with the last k rows equal to zero, where k is the nullity of the matrix. In our case, where the computation is performed in floating point arithmetic with precision μ and where $\epsilon > \mu$, the algorithm is halted if the last computed pivot a is such that $|a| \leq \epsilon\sqrt{m+n}$. This provides an estimate of the value k_ϵ and a candidate $g_\epsilon(x)$ to an ϵ-divisor of $u(x)$ and $v(x)$.

In the second stage, the tentative ϵ-divisor $g_\epsilon(x)$ is refined by means of Newton's iteration and a test is applied to check that $g_\epsilon(x)$ is an ϵ-common divisor. In this part, the value of k_ϵ can be adaptively modified in the case where $g_\epsilon(x)$ has not the maximum degree, or if $g_\epsilon(x)$ is not an ϵ-divisor of $u(x)$ and $v(x)$.

It is important to point out that the Jacobian system, which has to be solved at each Newton's iteration step, is still a Toeplitz-like linear system which can be reduced once again to Cauchy-like form and solved by means of the pivoted GKO algorithm.

In the refinement stage we have complemented Newton's iteration with a line search step in order to guarantee the monotonic behavior of the residual.

The algorithm has been implemented in Matlab, tested with a wide set of polynomials and compared with the currently available software, in particular the Matlab and Maple packages UVGCD by Zeng [23], STLN by Kaltofen et al. [14] and QRGCD by Corless et al. [4]. We did not compare our algorithm to the ones of [16], [24] since the software of the latter algorithms is not available.

We have considered the test polynomials of [23] and some new additional tests which are representative of difficult situations. In all the problems tested so far our algorithm has shown a high reliability and effectiveness, moreover, its $O(n^2)$ complexity makes it much faster than the currently available algorithms already for moderately large values of the degree. Our Matlab code is available upon request.

The paper is organized as follows. In Section 2 we recall the main tools used in the paper, among which, the properties of Sylvester and Bézout matrices, their interplay with gcd, the reduction to Cauchy-like matrices and a modified version of the GKO algorithm. In Section 3 we present the algorithms for estimating the degree and the coefficients of the ϵ-gcd together with the refinement stage based on Newton's iteration. Section 4 reports the results of the numerical experiments together with the comparison of our Matlab implementation with the currently available software.

2. Resultant matrices and ϵ-gcd

We recall the definitions of Bézout and Sylvester matrices and their interplay with gcd.

2.1. Sylvester and Bézout matrices

The Sylvester matrix of $u(x)$ and $v(x)$ is the $(m+n)\times(m+n)$ matrix

$$S(u,v)=\begin{pmatrix} u_n & u_{n-1} & \dots & u_0 & & 0\\ & \ddots & & & \ddots & \\ 0 & & u_n & u_{n-1} & \dots & u_0\\ v_m & v_{m-1} & \dots & v_0 & & 0\\ & \ddots & & & \ddots & \\ 0 & & v_m & v_{m-1} & \dots & v_0 \end{pmatrix}. \tag{1}$$

where the coefficients of $u(x)$ appear in the first m rows.

Assume that $n\ge m$ and observe that the rational function

$$b(x,y)=\frac{u(x)v(y)-u(y)v(x)}{x-y}$$

is actually a polynomial $\sum_{i,j=1}^{n} x^{i-1}y^{j-1}b_{i,j}$ in the variables x, y. The coefficient matrix $B(u,v)=(b_{i,j})$ is called the Bézout matrix of $u(x)$ and $v(x)$.

The following property is well known:

Lemma 2.1. *The nullities of $S(u,v)$ and of $B(u,v)$ coincide with deg(g).*

The next two results show how the gcd of $u(x)$ and $v(x)$ and the corresponding cofactors are related to Sylvester and Bézout submatrices. Recall that, for an integer $\nu\ge 2$ and a polynomial $a(x)=\sum_{j=0}^{\mu}a_jx^j$, the ν-th *convolution matrix* associated with $a(x)$ is the Toeplitz matrix having $[a_0,\dots,a_\mu,\underbrace{0\dots0}_{\nu-1}]^T$ as its first column and $[a_0,\underbrace{0,\dots0}_{\nu-1}]$ as its first row.

Lemma 2.2. *Let $u(x)=g(x)p(x)$, $v(x)=g(x)q(x)$, then the vector $[q_0,\dots,q_{m-k},-p_0,\dots,-p_{n-k}]^T$ belongs to the null space of the matrix $S_k=[\mathcal{C}_u\ \ \mathcal{C}_v]$, where $\mathcal{C}_u$ is the $(m-k+1)$-st convolution matrix associated with $u(x)$ and $\mathcal{C}_v$ is the $(n-k+1)$-st convolution matrix associated with $v(x)$.*

Theorem 2.3. [6] *Assume that $B(u,v)$ has rank $n-k$ and denote by $c_1,\dots,c_n$ its columns. Then $c_{k+1},\dots,c_n$ are linearly independent. Moreover writing each c_i $(1\le i\le k)$ as a linear combination of $c_{k+1},\dots,c_n$*

$$c_{k-i}=h_{k-i}^{k+1}c_{k+1}+\sum_{j=k+2}^{n}h_{k-i}^{j}c_j,\qquad i=0,\dots,k-1,$$

one finds that $D(x)=d_0x^k+d_1x^{k-1}+\dots+d_{k-1}x+d_k$ is a gcd for $u(x)$ and $v(x)$, where $d_1,\dots,d_k$ are given by $d_j=d_0h_{k-j+1}^{k+1}$, with $d_0\in\mathbb{R}$ or $\mathbb{C}$.

Moreover, we have:

Remark 2.4. *Let $g(x) = \sum_{i=0}^{k} g_i x^i$ be the gcd of $u(x)$ and $v(x)$, and let $\hat{u}(x)$ and $\hat{v}(x)$ be such that $u(x) = \hat{u}(x)g(x)$, $v(x) = \hat{v}(x)g(x)$. Then we have $B(u, v) = GB(\hat{u}, \hat{v})G^T$, where G is the $(n-k)$th convolution matrix associated with $g(x)$.*

2.2. Cauchy-like matrices

An $n \times n$ matrix C is called *Cauchy-like* of rank r if it has the form

$$C = \Big[\frac{\mathbf{u}_i \mathbf{v}_j^H}{f_i - \bar{a}_j}\Big]_{i,j=0}^{n-1}, \tag{2}$$

with $\mathbf{u}_i$ and $\mathbf{v}_j$ row vectors of length $r \leq n$, and f_i and a_j complex scalars such that $f_i - \bar{a}_j \neq 0$ for all i, j. The matrix G whose rows are given by the $\mathbf{u}_i$'s and the matrix B whose columns are given by the $\mathbf{v}_i$'s are called the *generators* of C.

Equivalently, C is Cauchy-like of rank r if the matrix

$$\nabla_C C = FC - CA^H, \tag{3}$$

where $F = \mathrm{diag}(f_0, \ldots, f_{n-1})$ and $A = \mathrm{diag}(a_0, \ldots, a_{n-1})$, has rank r. The operator ∇_C defined in (3) is a displacement operator associated with the Cauchy-like structure, and C is said to have *displacement rank* equal to r.

The algorithm that we now present is due to Gohberg, Kailath and Olshevsky [8], and is therefore known as *GKO algorithm*; it computes the Gaussian elimination with partial pivoting (GEPP) of a Cauchy-like matrix and can be extended to other classes of displacement structured matrices. The algorithm relies on the following

Fact 2.5. *Performing Gaussian elimination on an arbitrary matrix is equivalent to applying recursive Schur complementation; Schur complementation preserves the displacement structure; permutations of rows and columns preserve the Cauchy-like structure.*

It is therefore possible to directly apply Gaussian elimination with partial pivoting to the generators rather than to the whole matrix C, resulting in increased computational speed and less storage requirements.

So, a step of the fast GEPP algorithm for a Cauchy-like matrix $C = C_1$ can be summarized as follows (we assume that generators (G_1, B_1) of the matrix are given):

(i) Use (2) to recover the first column $\begin{bmatrix} d_1 \\ l_1 \end{bmatrix}$ of C_1 from the generators.

(ii) Determine the position (say, $(k, 1)$) of the entry of maximum magnitude in the first column.

(iii) Let P_1 be the permutation matrix that interchanges the first and kth rows. Interchange the first and kth diagonal entries of F_1; interchange the first and kth rows of G_1.

(iv) Recover from the generators the first row $\begin{bmatrix} \tilde{d}_1 & u_1 \end{bmatrix}$ of P_1C_1. Now one has the first column $\begin{bmatrix} 1 \\ \frac{1}{\tilde{d}_1}\tilde{l}_1 \end{bmatrix}$ of L and the first row $\begin{bmatrix} \tilde{d}_1 & u_1 \end{bmatrix}$ of U in the LU factorization of P_1C_1.

(v) Compute generators (G_2, B_2) of the Schur complement C_2 of $P_1 \cdot C_1$ as follows:

$$\begin{bmatrix} 0 \\ G_2 \end{bmatrix} = G_1 - \begin{bmatrix} 1 \\ \frac{1}{\tilde{d}_1}\tilde{l}_1 \end{bmatrix} g_1, \qquad \begin{bmatrix} 0 & B_2 \end{bmatrix} = B_1 - b_1 \begin{bmatrix} 1 & \frac{1}{\tilde{d}_1}u_1 \end{bmatrix}, \tag{4}$$

where g_1 is the first row of G_1 and b_1 is the first column of B_1.

Proceeding recursively, one obtains the factorization $C_1 = P \cdot L \cdot U$, where P is the product of the permutation matrices used in the process.

Now, let

$$Z_\phi = \begin{pmatrix} 0 & \dots & \dots & 0 & \phi \\ 1 & 0 & \dots & \dots & 0 \\ 0 & 1 & \ddots & & \vdots \\ \vdots & & \ddots & \ddots & \vdots \\ 0 & \dots & 0 & 1 & 0 \end{pmatrix} \tag{5}$$

and define the matrix operator

$$\nabla_T T = Z_1 T - T Z_{-1}. \tag{6}$$

An $n \times n$ matrix T having low displacement rank with respect to the operator ∇_T (i.e., such that $\nabla_T = GB$, with $G \in \mathbb{C}^{n\times r}$ and $B \in \mathbb{C}^{r\times n}$) is called Toeplitz-like. Sylvester and Bézout matrices are Toeplitz-like, with displacement rank 2.

Toeplitz-like matrices can be transformed into Cauchy-like as follows [10]. Here and hereafter $\hat{\imath}$ denotes the imaginary unit such that $\hat{\imath}^2 = -1$.

Theorem 2.6. *Let T be an $n \times n$ Toeplitz-like matrix. Then $C = \mathcal{F}TD_0^{-1}\mathcal{F}^H$ is a Cauchy-like matrix, i.e.,*

$$\nabla_{D_1,D_{-1}}(C) = D_1C - CD_{-1} = \hat{G}\hat{B}, \tag{7}$$

where

$$\mathcal{F} = \frac{1}{\sqrt{n}}[e^{\frac{2\pi\hat{\imath}}{n}(k-1)(j-1)}]_{k,j}$$

is the normalized $n \times n$ Discrete Fourier Transform matrix

$$D_1 = \mathrm{diag}(1, e^{\frac{2\pi\hat{\imath}}{n}}, \dots, e^{\frac{2\pi\hat{\imath}}{n}(n-1)}), \quad D_{-1} = \mathrm{diag}(e^{\frac{\pi\hat{\imath}}{n}}, e^{\frac{3\pi\hat{\imath}}{n}}, \dots, e^{\frac{(2n-1)\pi\hat{\imath}}{n}}),$$

$$D_0 = \mathrm{diag}(1, e^{\frac{\pi\hat{\imath}}{n}}, \dots, e^{\frac{(n-1)\pi\hat{\imath}}{n}}),$$

and

$$\hat{G} = \mathcal{F}G, \qquad \hat{B}^H = \mathcal{F}D_0B^H. \tag{8}$$

Therefore the GKO algorithm can be also applied to Toeplitz-like matrices, provided that reduction to Cauchy-like form is applied beforehand.

In particular, the generators (G, B) of the matrix $S(u, v)$ with respect to the Toeplitz-like structure can be chosen as follows. Let $N = n + m$; then G is the $N \times 2$ matrix having all zero entries except the entries $(1, 1)$ and $(m + 1, 2)$ which are equal to 1; the matrix B is $2 \times N$, its first and second rows are

$$[-u_{n-1}, \ldots, -u_1, v_m - u_0, v_{m-1}, \ldots, v_1, v_0 + u_n],$$
$$[-v_{m-1}, \ldots, -v_1, u_n - v_0, u_{n-1}, \ldots, u_1, u_0 + v_m],$$

respectively. Generators for $B(u, v)$ can be similarly recovered from the representation of the Bézout matrix as sum of products of Toeplitz/Hankel triangular matrices. Generators for the associated Cauchy-like matrix are computed from (G, B) by using (8).

2.3. Modified GKO algorithm

Gaussian elimination with partial pivoting (GEPP) is usually regarded as a fairly reliable method for solving linear systems. Its fast version, though, raises more stability issues.

Sweet and Brent [22] have done an error analysis of the GKO algorithm applied to a Cauchy-like matrix C. They point out that the error propagation depends not only on the magnitude of the triangular factors in the LU factorization of C (as is expected for ordinary Gaussian elimination), but also on the magnitude of the generators. In some cases, the generators can suffer large internal growth, even if the triangular factors do not grow too large, and therefore cause a corresponding growth in the backward and forward error. Experimental evidence shows that this is the case for Cauchy-like matrices derived from Sylvester and Bézout matrices.

However, it is possible to modify the GKO algorithm so as to prevent generator growth, as suggested for example in [21] and [9]. In particular, the latter paper proposes to orthogonalize the first generator before each elimination step; this guarantees that the first generator is well conditioned and allows a good choice of a pivot. In order to orthogonalize G, we need to:

- QR-factorize G, obtaining $G = \mathcal{G}R$, where $\mathcal{G}$ is an $n \times r$ column orthogonal matrix and R is upper triangular;
- define new generators $\tilde{G} = \mathcal{G}$ and $\tilde{B} = RB$.

This method performs partial pivoting on the column of C corresponding to the column of B with maximum norm. This technique is not equivalent to complete pivoting, but nevertheless allows a good choice of pivots and effectively reduces element growth in the generators, as well as in the triangular factors.

3. Fast ϵ-gcd computation

3.1. Estimating degree and coefficients of the ϵ-gcd

We first examine the following problem: find a fast method to determine whether two given polynomials $u(x)$ and $v(x)$ have an ϵ-divisor of given degree k. Throughout we assume that the input polynomials have unitary Euclidean norm.

The coefficients of the cofactors $p(x)$ and $q(x)$ can be obtained by applying Lemma 2.2. Once the cofactors are known, a tentative gcd can be computed as $g(x) = u(x)/p(x)$ or $g(x) = v(x)/q(x)$. Exact or nearly exact polynomial division (i.e., with a remainder of small norm) can be performed in a fast and stable way via evaluation/interpolation techniques (see [3]), which exploit the properties of the discrete Fourier transform.

Alternatively, Theorem 2.3 can be employed to determine the coefficients of a gcd; the cofactors, if required, are computed as $p(x) = u(x)/g(x)$ and $q(x) = v(x)/g(x)$.

The matrix in Lemma 2.2 is formed by two Toeplitz blocks and has displacement rank 2 with respect to the straightforward generalization of the operator ∇_T defined in (6) to the case of rectangular matrices. We seek to employ the modified GKO algorithm to solve the system that arises when applying Lemma 2.2, or the linear system that yields the coefficients of a gcd as suggested by Theorem 2.3.

In order to ensure that the matrices F and A defining the displacement operator ∇_C associated with the reduced matrix have well-separated spectra, a modified version of Theorem 2.6 is needed. Observe that a Toeplitz-like matrix T also has low displacement rank with respect to the operator $\nabla_{Z_1,Z_\theta}(T) = Z_1T - T \cdot Z_\theta$, for any $\theta \in \mathbb{C}$, $|\theta| = 1$. Then we have:

Theorem 3.1. *Let $T \in \mathbb{C}^{n\times m}$ be a Toeplitz-like matrix, satisfying*

$$\nabla_{Z_1,Z_\theta}(T) = Z_1T - TZ_\theta = GB,$$

where $G \in \mathbb{C}^{n\times\alpha}$, $B \in \mathbb{C}^{\alpha\times m}$ and Z_1, Z_θ are as in (5). Let $N = lcm\ (n, m)$. Then $C = \mathcal{F}_nTD_\theta\mathcal{F}_m$ is a Cauchy-like matrix, i.e.,

$$\nabla_{D_1,D_\theta}(C) = D_1C - CD_\theta = \hat{G}\hat{B}, \tag{9}$$

where $\mathcal{F}_n$ and $\mathcal{F}_m$ are the normalized Discrete Fourier Transform matrices of order n and m respectively,

$$D_\theta = \theta \cdot D_1,$$
$$D = \mathrm{diag}(1, e^{\frac{\pi\hat{\imath}}{Nm}}, e^{\frac{2\pi\hat{\imath}}{Nm}}, \dots)$$
$$D_1 = \mathrm{diag}(1,\ e^{\frac{2\pi\hat{\imath}}{n}}, \dots, e^{\frac{2\pi\hat{\imath}}{n}(n-1)})$$

and $\hat{G} = \mathcal{F}_nG, \qquad \hat{B}^H = \mathcal{F}_mDB^H.$

The optimal choice for θ is then $\theta = e^{\frac{\pi\hat{\imath}}{N}}$.

The gcd and cofactors obtained from Lemma 2.2 or Theorem 2.3 can be subsequently refined as described in the next section. After the refining step, it is easy to check whether an ϵ-divisor has actually been computed.

We are left with the problem of choosing a tentative gcd degree k_ϵ. A possibility is to employ a bisection technique, which requires to test the existence of an approximate divisor $\log_2 n$ times and therefore preserves the overall quadratic cost of the method.

Alternatively, a heuristic method of choosing a tentative value for k_ϵ can be designed by observing that, as a consequence of the properties of resultant matrices presented in Section 2.1, the choice of a suitable k_ϵ is mainly a matter of approximate rank determination, and the fast LU factorization of the Sylvester or Bézout matrix might provide reasonably useful values for k_ϵ.

Observe that the incomplete fast LU factorization computes a Cauchy-like perturbation matrix ΔC such that $C - \Delta C$ has rank $n - k$. If a is the last pivot computed in the incomplete factorization, then as a consequence of Lemma 2.2 in [9], $|a| \leq \|\Delta C\|_2$.

Now, let $u_\epsilon(x)$ and $v_\epsilon(x)$ be polynomials of minimum norm and same degrees as $u(x)$ and $v(x)$, such that $u+u_\epsilon$ and $v+v_\epsilon$ have an exact gcd of degree k. Assume $\|u_\epsilon\|_2 \leq \epsilon$ and $\|v_\epsilon\|_2 \leq \epsilon$. Let C_ϵ be the Cauchy-like matrix obtained via Theorem 2.6 from the Sylvester matrix $S_\epsilon = S(u_\epsilon, v_\epsilon)$. Then $C + C_\epsilon$ has rank $n - k$, too.

If we assume that $\|\Delta C\|_2$ is very close to the minimum norm of a Cauchy-like perturbation that decreases the rank of C to $n - k$, then we have

$$|a| \leq \|\Delta C\|_2 \leq \|C_\epsilon\|_2 = \|S_\epsilon\|_2 \leq \epsilon\sqrt{n+m}, \tag{10}$$

where the last inequality follows from the structure of the Sylvester matrix. Therefore, if $|a| > \epsilon/\sqrt{n+m}$, then $u(x)$ and $v(x)$ cannot have an ϵ-divisor of degree k. This gives an upper bound on the ϵ-gcd degree based on the absolute values of the pivots found while applying the fast Gaussian elimination to C. The same idea can be applied to the Bézout matrix.

This is clearly a heuristic criterion since it assumes that some uncheckable condition on $||\Delta C||_2$ is satisfied. However, this criterion seems to work quite well in practice. When it is applied, the gcd algorithm should check whether it actually provides an upper bound on the gcd degree. We use this criterion for the determination of a tentative gcd degree in our implementation of the algorithm. In fact, experimental evidence shows that this criterion is usually more efficient in practice than the bisection strategy, though in principle it does not guarantee that the quadratic cost of the overall algorithm is preserved.

3.2. Refinement

Since the computed value of k_ϵ is the result of a tentative guess, it might happen in principle that the output provided by the algorithm of Section 3.1 is not an ϵ-divisor, is an ϵ-divisor of lower degree, or is a poor approximation of the sought divisor. In order to get rid of this uncertainty, it is suitable to refine this output by means of an *ad hoc* iterative technique followed by a test on the correctness

of the ϵ-degree. For this purpose we apply Newton's iteration to the least squares problem defined by

$$F(\mathbf{z}) = \begin{bmatrix} \mathcal{C}_p\mathbf{g} - \mathbf{u} \\ \mathcal{C}_q\mathbf{g} - \mathbf{v} \end{bmatrix}, \qquad \mathbf{z} = \begin{bmatrix} \mathbf{g} \\ \mathbf{p} \\ \mathbf{q} \end{bmatrix}, \tag{11}$$

where the Euclidean norm of the function $F(\mathbf{z})$ is to be minimized. Here, in boldface we denote the coefficient vectors of the associated polynomials. The matrices $\mathcal{C}_p$ and $\mathcal{C}_q$ are convolution matrices of suitable size associated with the polynomials $p(x)$ and $q(x)$ respectively.

The Jacobian matrix J associated with the problem (11) has the form

$$J = \begin{pmatrix} \mathcal{C}_p & \mathcal{C}_g & 0 \\ \mathcal{C}_q & 0 & \mathcal{C}_g \end{pmatrix}, \tag{12}$$

where each block is a convolution matrix associated with a polynomial; $\mathcal{C}_p$ is of size $(n+1)\times(k+1)$, $\mathcal{C}_q$ is $(m+1)\times(k+1)$, $\mathcal{C}_g$ in the first block row is $(n+1)\times(n-k+1)$ and $\mathcal{C}_g$ in the second block row is $(m+1)\times(m-k+1)$. This Jacobian matrix, however, is always rank deficient, because of the lack of a normalization for the gcd.

Remark 3.2. *Under the hypotheses stated above, the Jacobian matrix* (12) *computed at any point* $\mathbf{z} = [\mathbf{g}^T \quad -\mathbf{p}^T \quad -\mathbf{q}^T]^T$ *is singular. Moreover, the nullity of* J *is* 1 *if and only if* $p(x)$, $q(x)$ *and* $g(x)$ *have no common factors. In particular, if* $\mathbf{z}$ *is a solution of* $F(\mathbf{z}) = 0$ *and* $g(x)$ *has maximum degree, i.e., it is a gcd, then* J *has nullity one and any vector in the null space of* J *is a multiple of* $\mathbf{w} = [\mathbf{g}^T \quad -\mathbf{p}^T \quad -\mathbf{q}^T]^T$, *where* $p(x)$ *and* $q(x)$ *are cofactors.*

In order to achieve better stability and convergence properties, we force the Jacobian to have full rank by adding a row, given by $\mathbf{w}^T$. Nevertheless, it can be proved, by relying on the results of [20], that the quadratic convergence of Newton's method in the case of zero residual also holds, in this case, with a rank deficient Jacobian. This property is useful when the initial guess for k_ϵ is too small, since in this case the rank deficiency of the Jacobian is unavoidable.

The new Jacobian $\tilde{J} = \begin{bmatrix} J \\ \mathbf{w}^T \end{bmatrix}$ is associated with the least squares problem that minimizes $\tilde{F}(\mathbf{z}) = \begin{bmatrix} F(\mathbf{z}) \\ \|\mathbf{g}\|^2 - \|\mathbf{p}\|^2 - \|\mathbf{q}\|^2 - K \end{bmatrix}$, where K is a constant. The choice of $\mathbf{w}^T$ as an additional row helps to ensure that the solution of each Newton's step

$$\mathbf{z}_{j+1} = \mathbf{z}_j - \tilde{J}(\mathbf{z}_j)^\dagger \tilde{F}(\mathbf{z}_j) \tag{13}$$

is nearly orthogonal to $\ker J$. Here $\tilde{J}(\mathbf{z}_j)^\dagger$ is the Moore-Penrose pseudoinverse of the matrix $\tilde{J}(\mathbf{z}_j)$. For ease of notation, the new Jacobian will be denoted simply as J in the following.

The matrix J has a Toeplitz-like structure, with displacement rank 5. We propose to exploit this property by approximating the solution of each linear least squares problem

$$J\eta_j = \tilde{F}(\mathbf{z}_j), \quad \eta_j = \mathbf{z}_j - \mathbf{z}_{j+1}$$

via fast LU factorization still preserving the quadratic convergence of the modified Newton's iteration obtained in this way.

We proceed as follows:

- Compute the factorization $J = LU$, where $J \in \mathbb{C}^{N\times M}$, $L \in \mathbb{C}^{N\times N}$ and $U \in \mathbb{C}^{N\times M}$. For the sake of simplicity, we are overlooking here the presence of permutation matrices due to the pivoting procedure; we can assume that either J or the vectors η_j and $\mathbf{x}_j = \tilde{F}(\mathbf{z}_j)$ have already undergone appropriate permutations.
  Consider the following block subdivision of the matrices L e U, where the left upper block has size $M \times M$:
  $$L = \left[\begin{array}{c|c} L_1 & 0 \\ \hline L_2 & I \end{array}\right], \qquad U = \left[\begin{array}{c} U_1 \\ \hline 0 \end{array}\right].$$
  Analogously, let $\mathbf{x}_j = \left[\begin{array}{c} \mathbf{x}_j^{(1)} \\ \hline \mathbf{x}_j^{(2)} \end{array}\right]$ and observe that $L^{-1} = \left[\begin{array}{c|c} L_1^{-1} & 0 \\ \hline -L_2L_1^{-1} & I \end{array}\right]$.
- Let $\mathbf{y}_j = L_1^{-1}\mathbf{x}_j^{(1)}$. If U_1 is nonsingular, then compute $\mathbf{w}_j$ as solution of $U_1\mathbf{w}_j = \mathbf{y}_j$. Else, consider the block subdivision
  $$U_1 = \left[\begin{array}{c|c} U_{11} & U_{12} \\ \hline 0 & 0 \end{array}\right], \quad \mathbf{w}_j = \left[\begin{array}{c} \mathbf{w}_j^{(1)} \\ \hline \mathbf{w}_j^{(2)} \end{array}\right], \quad \mathbf{y}_j = \left[\begin{array}{c} \mathbf{y}_j^{(1)} \\ \hline \mathbf{y}_j^{(2)} \end{array}\right],$$
  such that U_{11} is nonsingular; set all the entries of $\mathbf{w}_j^{(2)}$ equal to zero, and compute $\mathbf{w}_j^{(1)}$ as solution of $U_{11}\mathbf{w}_j^{(1)} = \mathbf{y}_j^{(1)}$
- If J is rank deficient, find a basis for $\mathcal{K} = \ker J$.
- Subtract from $\mathbf{w}_j$ its projection on $\mathcal{K}$, thus obtaining a vector χ_j. This is the vector that will be used as approximation of a solution of the linear least squares system in the iterative refinement process.

  Let $\mathcal{R}$ be the subspace of $\mathbb{C}^N$ spanned by the columns of J. We have
  $$\mathbb{C}^N = \mathcal{R} \oplus \mathcal{R}^{\perp}. \tag{14}$$

Let $\mathbf{x}_j = \alpha_j + \beta_j$ be the decomposition of $\mathbf{x}_j$ with respect to (14), i.e., we have $\alpha_j \in \mathcal{R}$ and $\beta_j \in \mathcal{R}^{\perp}$.

The Moore-Penrose pseudoinverse of J acts on $\mathbf{x}_j$ as follows: $J^{\dagger}\alpha_j$ is the preimage of α_j with respect to J and it is orthogonal to $\mathcal{K} = \ker J$, whereas $J^{\dagger}\beta_j$ is equal to zero.

The LU-based procedure, on the other hand, acts exactly like $J^{\dagger}$ on α_j, whereas the component β_j is not necessarily sent to 0. Therefore, χ_j is the sum of η_j and of the preimage of β_j with respect to the LU decomposition.

In a general linear least squares problem, there is no reason for $\|\beta_j\|_2$ to be significantly smaller than $\|\mathbf{x}_j\|_2$. In our case, though, the Taylor expansion of $F(\mathbf{z})$ yields:
$$0 = F(\mathbf{z}^*) = F(\mathbf{z}_j) - J(\mathbf{z}_j)\epsilon_j + \mathcal{O}(\|\epsilon_j\|_2^2), \tag{15}$$

where $\epsilon_j = \mathbf{z}_j - \mathbf{z}^*$ and $\mathbf{z}^*$ is such that $F(\mathbf{z}^*) = 0$. It follows from (15) that $\mathbf{x}_j = J(\mathbf{z}_j)\epsilon_j + \mathcal{O}(\|\epsilon_j\|_2^2)$. Since $J(\mathbf{z}_j)\epsilon_j \in \mathcal{R}$, we conclude that $\|\beta_j\|_2 = \mathcal{O}(\|\epsilon_j\|_2^2)$. Therefore, Newton's method applied to the iterative refinement of the polynomial gcd preserves its quadratic convergence rate, even though the linear least squares problems (13) are treated using via the LU factorization of the Jacobian.

The iterative process ends when at least one of the following criteria is satisfied:

1. the residual (that is, the Euclidean norm of the function $F(\mathbf{z})$) becomes smaller than a fixed threshold,
2. the number of iteration reaches a fixed maximum,
3. the residual given by the last iteration is greater than the residual given by the previous iteration.

The purpose of the third criterion is to avoid spending computational effort on tentative gcds that are not in fact suitable candidates. However, its use with Newton's method may pose some difficulties, because it is generally difficult to predict the global behaviour of this method; in particular, it might happen that the residual does not decrease monotonically. The usual way to overcome this obstacle is to use instead a relaxed version of Newton that includes a line search. More precisely, instead of the iteration (13) one computes

$$\mathbf{z}_{j+1} = \mathbf{z}_j - \alpha_j \tilde{J}(\mathbf{z}_j)^\dagger \tilde{F}(\mathbf{z}_j), \tag{16}$$

where α_j is chosen – using a one-dimensional minimization method – so as to approximately minimize the norm of $\tilde{F}(\mathbf{z}_j)$.

The drawback of this technique is that it slows down convergence: the quadratic convergence that was one of the main interesting points of Newton's method is lost if one consistently performs iterations of the type (16). For this reason we employ here a hybrid method: At each step, the algorithm evaluates the descent direction $\tilde{J}(\mathbf{z}_j)^\dagger \tilde{F}(\mathbf{z}_j)$ and checks if a pure Newton step (that is, (16) with $\alpha_j = 1$) decreases the residual. If this is the case, then the pure Newton step is actually performed; otherwise, α_j and subsequently $\mathbf{z}_{j+1}$ are computed by calling a line search routine. In this way, most of the optimization work is still performed by pure Newton iterations, so that the overall method remains computationally cheap; the line search, called only when necessary, is helpful in some difficult cases and ensures that the method has a sound theoretical basis.

3.3. The overall algorithm

Algorithm Fastgcd

Input: the coefficients of polynomials $u(x)$ and $v(x)$ and a tolerance ϵ.

Output: an ϵ-gcd $g(x)$; a backward error (residual of the gcd system); possibly perturbed polynomials $\hat{u}(x)$ and $\hat{v}(x)$ and cofactors $p(x)$ and $q(x)$.

Computation:

- Compute the Sylvester matrix S associated with $u(x)$ and $v(x)$;
- Use Lemma 2.6 to turn S into a Cauchy-like matrix C;

- Perform fast Gaussian elimination with almost complete pivoting on C; stop when a pivot a such that $|a| < \epsilon/\sqrt{n+m}$ is found; let k_0 be the order of the not-yet-factored submatrix $\tilde{U}$ that has a as upper left entry;
- Choose $k = k_0$ as tentative gcd degree;
- Is there an ϵ-divisor of degree k? The answer is found as follows:
  - find tentative cofactors by applying the modified GKO algorithm to the system given by Lemma 2.2,
  - compute a tentative gcd by performing polynomial division via evaluation/interpolation,
  - perform iterative refinement and check whether the backward error is smaller than ϵ;
- If yes, check for $k+1$; if there is also an ϵ-divisor of degree $k+1$, keep checking for increasing values of the degree until a maximum is reached (i.e., a degree is found for which there is no ϵ-divisor);
- If not, keep checking for decreasing values of the degree, until an ϵ-divisor (and gcd) is found.

Observe that a slightly different version of the above algorithm is still valid by replacing the Sylvester matrix with the Bézout matrix. With this replacement the size of the problem is roughly reduced by a factor of 2 with clear computational advantage.

It should also be pointed out that the algorithm generally outputs an approximate gcd with complex coefficients, even if $u(x)$ and $v(x)$ are real polynomials. This usually allows for a higher gcd degree or a smaller backward error.

4. Numerical experiments

The algorithm Fastgcd has been implemented in Matlab and tested on many polynomials, with satisfactory results. Some of these results are shown in this section and compared to the performance of other implemented methods that are found in the literature, namely UVGCD by Zeng [23], STLN by Kaltofen et al. [14] and QRGCD by Corless et al. [4]. Matlab experiments with Fastgcd and UVGCD are performed using version 7.5.0 running under Windows; we use here the P-code for UVGCD contained in the Apalab toolbox.

It must be pointed out that comparison with the STLN method is not straightforward, since this methods follows an optimization approach, i.e., it takes two (or more) polynomials and the desired gcd degree k as input, and seeks a perturbation of minimum norm such that the perturbed polynomials have an exact gcd of degree k. Moreover, the algorithms UVGCD and STLN do not normalize the input polynomials, whereas QRGCD and Fastgcd do; therefore all test polynomials are normalized (with unitary Euclidean norm) beforehand.

In the following tests, we generally display the residual (denoted as "res") associated with the gcd system (recall that the residual is defined here as the Euclidean norm of the function $F(\mathbf{z})$ and it may slightly differ from the residual as

defined by other authors). In some examples, where a nearly exact gcd is sought, we report the coefficient-wise error on the computed gcd (denoted as "cwe"), since the "correct" gcd is known.

4.1. Badly conditioned polynomials

The test polynomials in this section are taken from [23]. The polynomials in the first example are specifically chosen so that the gcd problem is badly conditioned.

Example 4.1. *Let n be an even positive integer and $k = n/2$; define $p_n = u_n v_n$ and $q_n = u_n w_n$, where*

$$u_n = \prod_{j=1}^{k}[(x - r_1\alpha_j)^2 + r_1^2\beta_j^2], \quad v_n = \prod_{j=1}^{k}[(x - r_2\alpha_j)^2 + r_2^2\beta_j^2],$$

$$w_n = \prod_{j=k+1}^{n}[(x - r_1\alpha_j)^2 + r_1^2\beta_j^2], \quad \alpha_j = \cos\frac{j\pi}{n}, \quad \beta_j = \sin\frac{j\pi}{n},$$

for $r_1 = 0.5$ and $r_2 = 1.5$. The roots of p_n and q_n lie on the circles of radius r_1 and r_2.

The following table shows the coefficient-wise errors given by the examined gcd methods as n increases.

n	Fastgcd	UVGCD	QRGCD
10	6.44×10^{-13}	3.24×10^{-13}	1.57×10^{-12}
12	5.23×10^{-12}	1.40×10^{-12}	3.28×10^{-4}
14	1.79×10^{-11}	2.27×10^{-11}	$(*)$
16	5.27×10^{-10}	4.41×10^{-11}	$(*)$
18	6.11×10^{-9}	3.63×10^{-10}	$(*)$

$(*)$ Here QRGCD fails to find a gcd of correct degree.

In this case, there are no substantial differences between the (good) results provided by Fastgcd and by UVGCD, while QRGCD outputs failure for very ill-conditioned cases. It should be pointed out, however, that the results given by UVGCD vary between trials, which makes comparisons more difficult.

In the following test, the gcd degree is very sensitive to the choice of the tolerance ϵ.

Example 4.2. *Let*

$$p(x) = \prod_{1}^{10}(x - x_j), \qquad q(x) = \prod_{1}^{10}(x - x_j + 10^{-j}),$$

with $x_j = (-1)^j(j/2)$. The roots of p and q have decreasing distances 0.1, 0.01, 0.001, *etc.*

The table shows, for several values of the tolerance, the corresponding gcd degree and residual found by Fastgcd and UVGCD. Fastgcd gives better results, since it generally finds gcds of higher degree. The algorithm QRGCD, on the contrary, outputs failure for all values of ϵ smaller than 10^{-2}.

ϵ	Fastgcd		UVGCD	
	deg	res	deg	res
10^{-2}	9	0.0045	9	0.0040
10^{-3}	8	2.63×10^{-4}	8	1.72×10^{-4}
10^{-4}	7	9.73×10^{-6}	$(*)$	
10^{-6}	6	2.78×10^{-7}	1	3.34×10^{-16}
10^{-7}	5	8.59×10^{-9}	1	3.34×10^{-16}

$(*)$ Here UVGCD outputs the same result as above due to a different definition of residual.

It is interesting to observe that for $\epsilon \leq 10^{-5}$, UVGCD computes a common ϵ-divisor which does not have the maximum degree, while Fastgcd always provides an ϵ-divisor of higher degree.

We have also studied this example using the STLN method, though the employed approach is entirely different. The following table shows the residuals computed by STLN for several values of the degree.

deg gcd	res	deg gcd	res
9	5.65×10^{-3}	6	2.58×10^{-7}
8	2.44×10^{-4}	5	6.34×10^{-9}
7	1.00×10^{-5}	4	1.20×10^{-10}

4.2. High gcd degree

In this example, also taken from [23], the gcd has a large degree.

Example 4.3. *Let* $p_n = u_n v$ *and* $q_n = u_n w$, *where* $v(x) = \sum_{j=0}^{3} x^j$ *and* $w(x) = \sum_{j=0}^{4} (-x)^j$ *are fixed polynomials and* u_n *is a polynomial of degree* n *whose coefficients are random integer numbers in the range* $[-5, 5]$.

The following table shows the residuals and the coefficient-wise errors on the computed gcd for large values of n. Here, Fastgcd and UVGCD perform similarly while QRGCD provides a worse coefficient-wise error.

n	Fastgcd		UVGCD		QRGCD
	res	cwe	res	cwe	cwe
50	2.97×10^{-16}	5.04×10^{-16}	2.43×10^{-16}	8.32×10^{-16}	1.72×10^{-12}
100	2.91×10^{-16}	1.41×10^{-15}	1.83×10^{-16}	7.77×10^{-16}	4.80×10^{-8}
200	5.08×10^{-16}	7.29×10^{-15}	1.72×10^{-16}	9.99×10^{-16}	2.39×10^{-11}
500	4.04×10^{-16}	3.12×10^{-15}	2.10×10^{-15}	1.35×10^{-14}	
1000	3.98×10^{-16}	3.28×10^{-15}	2.26×10^{-16}	1.67×10^{-15}	

4.3. Unbalanced coefficients

This is another example taken from [23].

Example 4.4. *Let $p = uv$ and $q = uw$, where $v(x)$ and $w(x)$ are as in Example* 4.3 *and*

$$u(x) = \sum_{j=0}^{15} c_j 10^{e_j} x^j,$$

where c_j and e_j are random integers in $[-5, 5]$ and $[0, 6]$ respectively.

In this example $u(x)$ is the gcd of $p(x)$ and $q(x)$ and the magnitude of its coefficients varies between 0 and 5×10^6. If an approximate gcd algorithm is applied and the coefficient-wise relative error θ is calculated, then $N = \log_{10} \theta$ is roughly the minimum number of correct digits for the coefficients of $u(x)$ given by the chosen method. 100 repetitions of this test are performed. The average number of correct digits found in an experiment of this type is 10.63 for Fastgcd and 10.83 for UVGCD. Therefore the two algorithms give comparable results. Residuals are always about 10^{-16}. QRGCD, on the contrary, achieves an average of 7.46 correct digits.

4.4. Multiple roots

Example 4.5. *Let $u(x) = (x^3 + 3x - 1)(x - 1)^k$ for a positive integer k, and let $v(x) = u'(x)$. The gcd of $u(x)$ and $v(x)$ is $g(x) = (x - 1)^{k-1}$.*

The coefficient-wise errors computed by Fastgcd, UVGCD and QRGCD for several values of k and for $\epsilon = 10^{-6}$ are shown in the following table. Unless otherwise specified, the computed gcd degrees are understood to be correct.

k	Fastgcd	UVGCD	QRGCD
15	5.18×10^{-13}	4.27×10^{-13}	7.04×10^{-7}
25	9.31×10^{-11}	1.99×10^{-11}	$(*)$
35	1.53×10^{-8}	4.44×10^{-9}	$(*)$
45	6.61×10^{-6}	4.04×10^{-8}	$(*)$

$(*)$ Here QRGCD does not detect a gcd of correct degree.

The algorithm UVGCD has been specifically designed for polynomials with multiple roots and is therefore very efficient. Fastgcd also provides good results, with backward errors (residuals) always of the order of the machine epsilon, whereas QRGCD fails to find a gcd of correct degree as soon as the root multiplicity is larger than Donnerstag, Oktober 8, 2009 at 3:44 pm15.

4.5. Small leading coefficient

A gcd with a small leading coefficient may represent in many cases a source of instability.

Example 4.6. *For a given (small) parameter $\alpha \in \mathbb{R}$, let $g(x) = \alpha x^3 + 2x^2 - x + 5$, $p(x) = x^4 + 7x^2 - x + 1$ and $q(x) = x^3 - x^2 + 4x - 2$ and set $u(x) = g(x)p(x)$, $v(x) = g(x)q(x)$.*

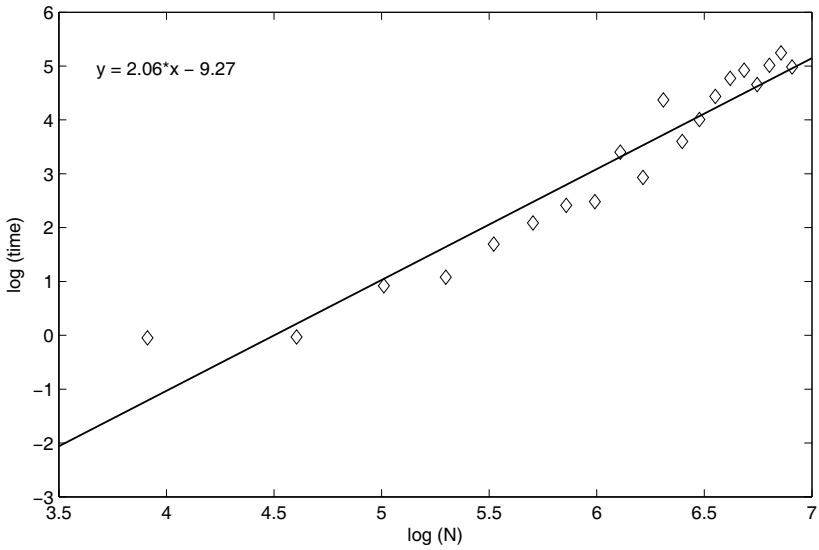


FIGURE 1. Running time of the algorithm Fastgcd

We applied Fastgcd and QRGCD to this example, with α ranging between 10^{-5} and 10^{-15}. It turns out that, for $\alpha < 10^{-5}$, QRGCD fails to recognize the correct gcd degree and outputs a gcd of degree 2. Fastgcd, on the contrary, always recognizes the correct gcd degree, with a residual of the order of the machine epsilon.

4.6. Running time

We have checked the growth rate of the running time of the algorithm Fastgcd on pairs of polynomials whose GCD and cofactors are defined like the polynomials $u_n(x)$ introduced in Section 4.2. Polynomials of degree $N = 2n$ ranging between 50 and 1000 have been used. Figure 1 shows the running time (in seconds) versus the degree in log-log scale, with a linear fit and its equation. Roughly speaking, the running time grows as $\mathcal{O}(N^\alpha)$, where α is the coefficient of the linear term in the equation, i.e., 2.06 in our case.

We next show a comparison between the running times of Fastgcd and UVGCD. In order to avoid randomly chosen coefficients, we define a family of test polynomials as follows. Let k be a positive integer and let $n_1 = 25k$, $n_2 = 15k$ and $n_3 = 10k$. For each value of k define the cofactors $p_k(x) = (x^{n_1}-1)(x^{n_2}-2)(x^{n_3}-3)$ and $q_k(x) = (x^{n_1}+1)(x^{n_2}+5)(x^{n_3}+\hat{\imath})$. The test polynomials are $u_k(x) = g(x)p_k(x)$ and $v_k(x) = g(x)q_k(x)$, where the gcd $g(x) = x^4+10x^3+x-1$ is a fixed polynomial.

Figure 2 shows the computing times required by Fastgcd and UVGCD on $u_k(x)$ and $v_k(x)$ for $k = 1, \ldots, 8$. The plot clearly shows that the time growth for Fastgcd is much slower than for UVGCD.

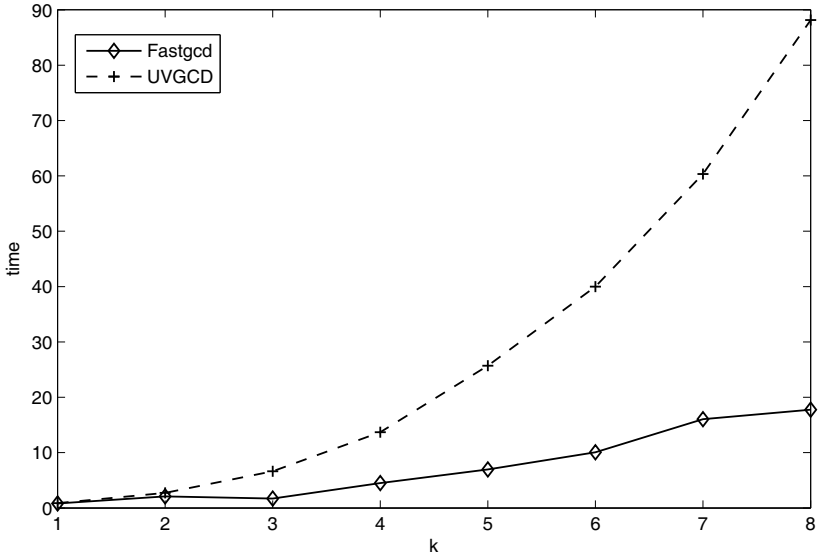


FIGURE 2. Comparison between the running times of Fastgcd and UVGCD.

References

[1] B. Beckermann, G. Labahn, *When are two numerical polynomials relatively prime?*, J. Symbolic Comput. **26**, 677–689 (1998).

[2] B. Beckermann, G. Labahn, *A fast and numerically stable Euclidean-like algorithm for detecting relatively prime numerical polynomials*, J. Symb. Comp. **26**, 691–714 (1998).

[3] D.A. Bini, V.Y. Pan, *Polynomial and Matrix Computations*, vol. I, Birkhäuser, 1994.

[4] R.M. Corless, S.M. Watt, L. Zhi, *QR Factoring to Compute the GCD of Univariate Approximate Polynomials*, IEEE Trans. Signal Processing **52**, 3394–3402 (2004).

[5] R.M. Corless, P.M. Gianni, B.M. Trager, S.M. Watt, *The Singular Value Decomposition for Approximate Polynomial Systems*, Proc. International Symposium on Symbolic and Algebraic Computation, July 10–12 1995, Montreal, Canada, ACM Press 1995, pp. 195–207.

[6] G.M. Diaz-Toca, L. Gonzalez-Vega, *Computing greatest common divisors and squarefree decompositions through matrix methods: The parametric and approximate cases*, Linear Algebra Appl. **412**, 222–246 (2006).

[7] I.Z. Emiris, A. Galligo, H. Lombardi, *Certified approximate univariate GCDs*, J. Pure Appl. Algebra **117/118**, 229–251 (1997).

[8] I. Gohberg, T. Kailath, V. Olshevsky, *Fast Gaussian elimination with partial pivoting for matrices with displacement structure*, Math. Comp. **64**, 1557–1576 (1995).

[9] M. Gu, *Stable and Efficient Algorithms for Structured Systems of Linear Equations*, SIAM J. Matrix Anal. Appl. **19**, 279–306 (1998).

[10] G. Heinig, *Inversion of generalized Cauchy matrices and other classes of structured matrices*, Linear Algebra in Signal Processing, IMA volumes in Mathematics and its Applications **69**, 95–114 (1994).

[11] G. Heinig, P. Jankowsky, K. Rost, *Fast inversion of Toeplitz-plus-Hankel matrices*, Numer. Math. **52**, 665–682 (1988).

[12] V. Hribernig, H.J. Stetter, *Detection and validation of clusters of polynomial zeros*, J. Symb. Comp. **24**, 667–681 (1997).

[13] C.-P. Jeannerod, G. Labahn, *SNAP User's Guide*, UW Technical Report no. CS-2002-22 (2002).

[14] E. Kaltofen, Z. Yang, L. Zhi, *Approximate Greatest Common Divisors of Several Polynomials with Linearly Constrained Coefficients and Singular Polynomials*, Proc. International Symposium on Symbolic and Algebraic Computations, 2006.

[15] N.K. Karmarkar, Y.N. Lakshman, *On Approximate GCDs of Univariate Polynomials*, J. Symbolic Comp. **26**, 653–666 (1998).

[16] B. Li, Z. Yang, L. Zhi, *Fast Low Rank Approximation of a Sylvester Matrix by Structure Total Least Norm*, Journal of Japan Society for Symbolic and Algebraic Computation **11**, 165–174 (2005).

[17] M.-T. Noda, T. Sasaki, *Approximate GCD and its application to ill-conditioned algebraic equations*, J. Comput. Appl. Math. **38**, 335–351 (1991).

[18] V.Y. Pan, *Numerical computation of a polynomial GCD and extensions*, Information and Computation **167**, 71–85 (2001).

[19] A. Schönhage, *Quasi-GCD Computations*, J. Complexity, **1**, 118–137 (1985).

[20] L.B. Rall, *Convergence of the Newton Process to Multiple Solutions*, Num. Math. **9**, 23–37 (1966).

[21] M. Stewart, *Stable Pivoting for the Fast Factorization of Cauchy-Like Matrices*, preprint (1997).

[22] D.R. Sweet, R.P. Brent, *Error analysis of a fast partial pivoting method for structured matrices*, in Adv. Signal Proc. Algorithms, Proc. of SPIE, T. Luk, ed., 266–280 (1995).

[23] Z. Zeng, *The approximate GCD of inexact polynomials Part I: a univariate algorithm*, to appear

[24] L. Zhi, *Displacement Structure in computing the Approximate GCD of Univariate Polynomials*, Mathematics, W. Sit and Z. Li eds., World Scientific (Lecture Notes Series on Computing), 288–298 (2003).

Dario A. Bini and Paola Boito
Dipartimento di Matematica
Università di Pisa
Largo Bruno Pontecorvo 5
I-56127 Pisa, Italy
e-mail: bini@dm.unipi.it
boito@mail.dm.unipi.it

Operator Theory:
Advances and Applications, Vol. 199, 175–189

On Inertia of Some Structured Hermitian Matrices

Vladimir Bolotnikov

Dedicated to the memory of Georg Heinig

Abstract. Two classes of structured Hermitian matrices are considered with the additional property that certain principal submatrices are all singular. Such matrices can be considered as the Pick matrices of certain (interior and boundary) norm constrained interpolation problems for functions meromorphic on the unit disk which the iterative Schur algorithm does not apply to. We characterize these matrices in terms of the parameters determining their structure and present formulas for their inertia.

Mathematics Subject Classification (2000). 15A57, 11C20.

Keywords. Structured Hermitian matrices, inertia.

1. Introduction

In this note we discuss two related classes of structured matrices whose structure is determined by Stein equations and the additional property that the principal minors of certain type are all zeros. Let $J_n(z)$ denote the $n \times n$ Jordan block with $z \in \mathbb{C}$ on the main diagonal and let E_n stand for the vector of the length n with the first coordinate equals one and other coordinates equal zero:

$$J_n(z) = \begin{bmatrix} z & 0 & \dots & 0 \\ 1 & z & \ddots & \vdots \\ & \ddots & \ddots & 0 \\ 0 & & 1 & z \end{bmatrix}, \quad E_n = \begin{bmatrix} 1 \\ 0 \\ \vdots \\ 0 \end{bmatrix}. \tag{1.1}$$

Given a tuple $\mathbf{z} = (z_1, \dots, z_k)$ of k distinct points in the open unit disk $\mathbb{D}$ and a tuple $\mathbf{n} = (n_1, \dots, n_k)$ of natural numbers, we let $|\mathbf{n}| := n_1 + \cdots + n_k$ and define

$$T_{\mathbf{n}}(\mathbf{z}) = \begin{bmatrix} J_{n_1}(z_1) & & \\ & \ddots & \\ & & J_{n_k}(z_k) \end{bmatrix} \quad \text{and} \quad E_{\mathbf{n}} = \begin{bmatrix} E_{n_1} \\ \vdots \\ E_{n_k} \end{bmatrix}. \tag{1.2}$$

Definition 1.1. *We say that a matrix $P \in \mathbb{C}^{|\mathbf{n}|\times|\mathbf{n}|}$ belongs to the class $\mathcal{D}_{\mathbf{n}}$ if it satisfies the Stein identity*

$$P - T_{\mathbf{n}}(\mathbf{z}) P T_{\mathbf{n}}(\mathbf{z})^* = E_{\mathbf{n}} E_{\mathbf{n}}^* - C_{\mathbf{n}} C_{\mathbf{n}}^* \tag{1.3}$$

for some vector

$$C_{\mathbf{n}} = \begin{bmatrix} C_{1,n_1} \\ \vdots \\ C_{k,n_k} \end{bmatrix}, \quad \textit{where} \quad C_{i,n_i} = \begin{bmatrix} c_{i,0} \\ \vdots \\ c_{i,n_i-1} \end{bmatrix}, \tag{1.4}$$

and if its compression to any $T_{\mathbf{n}}(\mathbf{z})^*$-invariant subspace of $\mathbb{C}^{|\mathbf{n}|}$ is singular.

Since all the eigenvalues of $T_{\mathbf{n}}(\mathbf{z})$ fall inside the unit disk, the Stein equation (1.3) has a unique solution P which is Hermitian. Solving (1.3) gives the following explicit formulas for the entries of P:

$$P = [P_{ij}]_{i,j=1}^{k}, \quad P_{ij} \in \mathbb{C}^{n_i \times n_j}, \tag{1.5}$$

where

$$\begin{aligned} [P_{ij}]_{\ell,r} &= \sum_{s=0}^{\min\{\ell,r\}} \frac{(\ell+r-s)!}{(\ell-s)!s!(r-s)!} \frac{z_i^{r-s}\bar{z}_j^{\ell-s}}{(1-z_i\bar{z}_j)^{\ell+r-s+1}} \\ &- \sum_{\alpha=0}^{\ell}\sum_{\beta=0}^{r}\sum_{s=0}^{\min\{\alpha,\beta\}} \frac{(\alpha+\beta-s)!}{(\alpha-s)!s!(\beta-s)!} \frac{z_i^{\beta-s}\bar{z}_j^{\alpha-s} c_{i,\ell-\alpha} c_{j,r-\beta}^*}{(1-z_i\bar{z}_j)^{\alpha+\beta-s+1}}. \end{aligned} \tag{1.6}$$

In (1.6) and in what follows we will use the symbol c^* for the complex conjugate of $c \in \mathbb{C}$. Also we will denote by $\pi(P)$, $\nu(P)$ and $\delta(P)$ respectively the numbers of positive, negative and zero eigenvalues, counted with multiplicities, of a Hermitian matrix P. For two tuples $\mathbf{n}$ and $\mathbf{m}$ in $\mathbb{Z}_+^k$, we will say that

$$\mathbf{m} = (m_1, \dots, m_k) \preceq (n_1, \dots, n_k) = \mathbf{n} \quad \text{if} \quad m_i \le n_i \quad \text{for} \quad i = 1, \dots, k. \tag{1.7}$$

For a matrix $P = P_{\mathbf{n}}$ decomposed in blocks as in (1.5) and a tuple $\mathbf{m} \preceq \mathbf{n}$ as in (1.7), define the principal submatrix $P_{\mathbf{m}} = [(P_{\mathbf{m}})_{ij}]_{i,j=1}^{k}$ of P whose block entries $(P_{\mathbf{m}})_{ij}$'s are equal to the leading $m_i \times m_j$ submatrices of the corresponding blocks in P:

$$P_{\mathbf{m}} = [(P_{\mathbf{m}})_{ij}]_{i,j=1}^{k} \quad \text{where} \quad (P_{\mathbf{m}})_{ij} = \begin{bmatrix} I_{m_i} & 0 \end{bmatrix} P_{ij} \begin{bmatrix} I_{m_j} \\ 0 \end{bmatrix}. \tag{1.8}$$

It is easily seen that any compression of P to a $T_{\mathbf{n}}(\mathbf{z})^*$-invariant subspace of $\mathbb{C}^{|\mathbf{n}|}$ is of the form $P_{\mathbf{m}}$ for some $\mathbf{m} \preceq \mathbf{n}$. Thus the class $\mathcal{D}_{\mathbf{n}}$ can be characterized as follows.

Definition 1.1′. $\mathcal{D}_{\mathbf{n}}$ consists of all matrices P of the form (1.5), (1.6) (for some $c_{i,j} \in \mathbb{C}$) and such that *for every* $\mathbf{m} \preceq \mathbf{n}$, *the submatrix* $P_{\mathbf{m}}$ *of* P *is singular.*

Classes $\mathcal{D}_{\mathbf{n}}$ admit the following functional-model interpretation. Let H^2 be the Hardy space of the unit disk, let $\theta(z)$ be the finite Blaschke product

$$\theta(z) = \prod_{i=1}^{k} \left(\frac{z - z_i}{1 - z\bar{z}_i} \right)^{n_i}$$

associated with the tuples $\mathbf{z}$ and $\mathbf{n}$, and let $K_\theta := H^2 \ominus \theta H^2$ be the model space. The functions

$$e_{i,j}(z) = \frac{1}{j!} \frac{d^j}{d\bar{z}_i} \left(\frac{1 - \theta(z)\theta(z_i)^*}{1 - z\bar{z}_i} \right) \quad (i = 1, \dots, k;\ j = 0, \dots, n_i - 1), \tag{1.9}$$

form a basis for K_θ and therefore, $\dim K_\theta = |\mathbf{n}|$. The space K_θ is invariant with respect to the backward shift operator $R: \ f \to \frac{f(z) - f(0)}{z}$ and the matrix of R with respect to the basis (1.9) is equal to $T_{\mathbf{n}}(\mathbf{z})^*$ defined in (1.2). Let T_f be the Toeplitz operator with symbol $f \in H^\infty$ and let $P_{\theta,f}$ be the compression of the operator $I - T_f T_f^*$ to the model space K_θ. Using the reproducing property of the kernel $\dfrac{1 - \theta(z)\theta(\zeta)^*}{1 - z\bar{\zeta}}$ for the space K_θ, it is not hard to show that the matrix of the operator $P_{\theta,f}$ with respect to the basis (1.9) equals

$$P_{\mathbf{n}}^f(\mathbf{z}) = \left[\left[\frac{1}{\ell! r!} \frac{\partial^{\ell + r}}{\partial z^\ell \partial \bar{\zeta}^r} \frac{1 - f(z)\overline{f(\zeta)}}{1 - z\bar{\zeta}} \Big|_{\substack{z = z_i \\ \zeta = z_j}} \right]_{\ell = 0, \dots, n_i - 1}^{r = 0, \dots, n_j - 1} \right]_{i,j=1}^{k}. \tag{1.10}$$

If in addition,

$$f^{(j)}(z_i) = j!\, c_{i,j} \qquad (i = 1, \dots, k;\ j = 0, \dots, n_i - 1), \tag{1.11}$$

then differentiation in (1.10) shows that $P_{\mathbf{n}}^f(\mathbf{z})$ is equal to the matrix P defined in (1.5), (1.6). Furthermore, $P \in \mathcal{D}_{\mathbf{n}}$ means that the compression of $P_{\theta,f}$ to any backward shift invariant subspace of K_θ is singular.

Now we will explain why the class $\mathcal{D}_{\mathbf{n}}$ (or equivalently, the class of operators on K_θ of the form $P_{K_\theta}(I - T_f T_f^*)|_{K_\theta}$ with all backward shift invariant compressions singular) are of some interest. Note that the matrix $P_{\mathbf{n}}^f(\mathbf{z})$ can be defined for every function f analytic at $z_1, \dots, z_k$ (not necessarily analytic on all of $\mathbb{D}$). If we will think of (1.11) as of interpolation conditions for an unknown function f (say, rational and with $\max_{z \in \mathbb{T}} |f(z)| \le 1$), then $P_{\mathbf{n}}^f(\mathbf{z}) = P$ for every solution f of the problem (1.11) with interpolation data $\{z_i, c_{i,j}\}$. The matrix P is then called the *Pick matrix* of the problem (1.11).

In the nondegenerate case (where P is invertible), the solution set of the problem (1.11) can be parametrized in terms of a linear fractional transformation

(see, e.g., [4]). If P is singular, one may start with a subproblem of (1.11) of the form

$$f^{(j)}(z_i) = j!\, c_{i,j} \qquad (i = 1, \dots, k;\ j = 0, \dots, m_i - 1), \tag{1.12}$$

where $m_i \le n_i$ for $i = 1, \dots, k$. It is readily seen that the Pick matrix of this subproblem is the principal submatrix $P_{\mathbf{m}}$ of P defined as in (1.8), and if this matrix is invertible, one can apply the Schur algorithm to reduce the original problem (1.11) to a problem with $|\mathbf{m}|$ fewer interpolation conditions. We now see that the class $\mathcal{D}_{\mathbf{n}}$ consists of Pick matrices corresponding to interpolation problems of the form (1.11) which the Schur algorithm (even the first step of this algorithm) does not apply to. On the other hand, even if we start with a degenerate interpolation problem (1.11) containing nondegenerate subproblems, then after a number of steps we still come up with a reduced problem whose Pick matrix belongs to the class $\mathcal{D}_{\mathbf{m}}$ for some $\mathbf{m} \preceq \mathbf{n}$. In other words, the Schur algorithm reduces any degenerate problem (1.11) to a problem with the Pick matrix from the class $\mathcal{D}_{\mathbf{m}}$. In Section 2 we characterize the matrices of the class $\mathcal{D}_{\mathbf{n}}$ in terms of the parameters $\{c_{i,j}\}$ and establish a simple formula for the inertia of a $P \in \mathcal{D}_{\mathbf{n}}$ (the results of this sort go back to [8]; see also [1], [2], [3] for further developments). The formula for inertia (see Theorem 2.2) is of certain interest for interpolation problems by rational functions unimodular on the unit circle $\mathbb{T}$ (i.e., ratios of finite Blaschke products: $f = b_1/b_2$), since the inertia of the Pick matrix controls the minimally possible degrees of b_1 and b_2. In Section 3 we consider the class $\mathcal{B}_{\mathbf{n}}$ (a "boundary" analog of the class $\mathcal{D}_{\mathbf{n}}$), the class of matrices with leading (with respect to a designated block decomposition) minors equal zero and with the structure determined by the Stein identity (1.3) where the (distinct) points $t_1, \dots, t_k$ fall on the unit circle $\mathbb{T}$ rather than inside the unit disk. In this case, for the Stein equation (1.3) to have a solution, the vector $C_{\mathbf{n}}$ must satisfy certain additional a'priori conditions. If these conditions are satisfied, the equation has infinitely many Hermitian solutions, and all of them are of the same special structure. We will see how this structure simplifies under the singularity assumption about leading submatrices. The matrices of the class $\mathcal{B}_{\mathbf{n}}$ serve as Pick matrices of boundary L^∞-norm constrained interpolation problems which do not contain nondegenerate subproblems. Applications of the results presented here to boundary interpolation problems will be demonstrated elsewhere.

2. The interior case

Since the unique solution P of the Stein equation (1.3) is determined by $z_1, \dots, z_k$ and the entries $c_{i,j}$ of the vector $C_{\mathbf{n}}$, the characterization of matrices from the class $\mathcal{D}_{\mathbf{n}}$ can be given in terms of this data. It turns out that (distinct) points $z_1, \dots, z_k$ do not play any role in this characterization.

Proposition 2.1. *A matrix P of the form* (1.5), (1.6) *belongs to $\mathcal{D}_{\mathbf{n}}$ if and only if*

$$c_{1,0} = c_{2,0} = \cdots = c_{k,0} = \gamma, \quad |\gamma| = 1 \tag{2.1}$$

and

$$c_{i,j} = 0 \quad \textit{for every} \quad i = 1, \ldots, k \ \textit{ and } \ 1 \le j \le \left[\frac{n_i}{2}\right], \tag{2.2}$$

where $[x]$ *stands for the greatest integer less than or equal to* x.

Proof. Let for short $\ell_i = \left[\frac{n_i}{2}\right]$ and let us assume that $P \in \mathcal{D}_{\mathbf{n}}$. Then in particular, the leading submatrices of diagonal blocks P_{ii} are singular. Since P_{ii} satisfies the Stein identity

$$P_{ii} - J_{n_i}(z_i) P_{ii} J_{n_i}(z_i)^* = E_{n_i} E_{n_i}^* - C_{i,n_i} C_{i,n_i}^* \tag{2.3}$$

(which is just the equality between the ith diagonal blocks in (1.7)), Theorem 5.1 in [1] applies and says that the $\ell_i \times \ell_i$ leading submatrix $\widetilde{P}_{ii}$ of P_{ii} is the zero matrix. Now we compare the $\ell_i \times \ell_i$ leading submatrices in (2.3) to conclude that $0 = E_{\ell_i} E_{\ell_i}^* - C_{i,\ell_i} C_{i,\ell_i}^*$ where E_{ℓ_i} and C_{i,ℓ_i} are defined via formulas (1.1) and (1.4). Equating the top rows in the latter equality gives

$$0 = \begin{bmatrix} 1 - |c_{i,0}|^2 & -c_{i,0}c_{i,1}^* & -c_{i,0}c_{i,2}^* & \ldots & c_{i,0}c_{i,\ell}^* \end{bmatrix},$$

which implies

$$|c_{i,0}| = 1 \quad \text{for} \quad i = 1, \ldots, k \tag{2.4}$$

and $c_{i,1} = c_{i,2} = \cdots = c_{i,\ell_i} = 0$ which is the same as (2.2). Now we take the tuple $\mathbf{m} = (m_1, \ldots, m_k)$ with $m_i = m_j = 1$ and $m_\ell = 0$ for $\ell \ne i, j$. Since $\mathbf{m} \preceq \mathbf{n}$, the submatrix

$$P_{\mathbf{m}} = \begin{bmatrix} \dfrac{1 - |c_{i,0}|^2}{1 - |z_i|^2} & \dfrac{1 - c_{i,0}c_{j,0}^*}{1 - z_i \bar{z}_j} \\ \dfrac{1 - c_{j,0}c_{i,0}^*}{1 - z_j \bar{z}_i} & \dfrac{1 - |c_{j,0}|^2}{1 - |z_j|^2} \end{bmatrix} = \begin{bmatrix} 0 & \dfrac{1 - c_{i,0}c_{j,0}^*}{1 - z_i \bar{z}_j} \\ \dfrac{1 - c_{j,0}c_{i,0}^*}{1 - z_j \bar{z}_i} & 0 \end{bmatrix}$$

of P is singular, which implies

$$c_{i,0} = c_{j,0} \quad \text{for} \quad i, j = 1, \ldots, k. \tag{2.5}$$

Equalities (2.1) follow from (2.4) and (2.5).

For the sufficiency part, observe that conditions (2.1) and (2.2) guarantee that all the entries in the $(\left[\frac{n_i}{2}\right] + 1) \times (\left[\frac{n_j}{2}\right] + 1)$ leading submatrix of the block P_{ij} of P are zeros (this is readily seen from the explicit formula (1.6) for the entries of P_{ij}):

$$P_{ij} = \begin{bmatrix} 0_{(\left[\frac{n_i}{2}\right]+1) \times (\left[\frac{n_j}{2}\right]+1)} & * \\ * & * \end{bmatrix} \quad (i, j = 1, \ldots, k). \tag{2.6}$$

Since $P_{ij} \in \mathbb{C}^{n_i \times n_j}$, it follows that the dimensions of the zero block in (2.6) are greater than halves of the corresponding dimensions of P_{ij}. Therefore, for every $\mathbf{m} \preceq \mathbf{n}$, the matrix $P_{\mathbf{m}}$ contains a zero principal submatrix whose dimensions are greater than $\frac{|\mathbf{m}|}{2}$. Therefore, $P_{\mathbf{m}}$ is singular which completes the proof of the theorem. □

The next theorem describes the inertia of a matrix $P \in \mathcal{D}_{\mathbf{n}}$. It shows in particular, that a matrix $P \in \mathcal{D}_{\mathbf{n}}$ has equally many positive and negative eigenvalues.

Note that in the single-block case (i.e., if $k = 1$) this theorem is a particular case of Theorem 5.1 in [1].

Theorem 2.2. *Let* $\mathbf{z} = (z_1, \ldots, z_k) \in \mathbb{D}^k$ *and let* $\mathbf{n} = (n_1, \ldots, n_k) \in \mathbb{N}^k$ *and* $\mathbf{d} = (d_1, \ldots, d_k) \in \mathbb{Z}_+^k$ *be two tuples such that*

$$\frac{n_i}{2} \le d_i \le n_i \quad \textit{for} \quad i = 1, \ldots, k. \tag{2.7}$$

Let P *be defined as in* (1.5), (1.6) *and let us assume that the numbers* $c_{i,j}$ *satisfy* (2.1) *and*

$$c_{i,d_i} \neq 0 \quad \textit{and} \quad c_{i,j} = 0 \quad \textit{for every} \quad i = 1, \ldots, k \ \textit{and} \ 1 \le j \le d_i - 1 \tag{2.8}$$

(*the first inequality in* (2.8) *is relevant only if* $d_i < n_i$). *Then*

$$\pi(P) = \nu(P) = |\mathbf{n}| - |\mathbf{d}| \quad \textit{and} \quad \delta(P) = 2|\mathbf{d}| - |\mathbf{n}|. \tag{2.9}$$

Proof. Note that under assumptions (2.8), the matrix P belongs to $\mathcal{D}_{\mathbf{n}}$ if and only if $\frac{n_i}{2} < d_i \le n_i$ for $i = 1, \ldots, k$ (by Proposition 2.1). However, the conclusion (2.9) of the theorem holds true even if $d_i = \frac{n_i}{2}$ for some i's. In particular, it follows that if $d_i = \frac{n_i}{2}$ for every $i = 1, \ldots, k$, then the matrix P is invertible and has equally many positive and negative eigenvalues.

For the proof, we first observe that conditions (2.1) and (2.8) guarantee that $P_{\mathbf{d}}$, the principal submatrix of P defined via (1.8), is equal to the zero matrix

$$P_{\mathbf{d}} = 0. \tag{2.10}$$

The latter can be seen directly from the explicit formula (1.6) for the entries of P. Since $P \in \mathbb{C}^{|\mathbf{n}| \times |\mathbf{n}|}$ and $P_{\mathbf{d}} \in \mathbb{C}^{|\mathbf{d}| \times |\mathbf{d}|}$, it follows from (2.10), that $\operatorname{rank}(P) \le 2(|\mathbf{n}| - |\mathbf{d}|)$ and therefore, that

$$\delta(P) \ge |\mathbf{n}| - 2(|\mathbf{n}| - |\mathbf{d}|) = 2|\mathbf{d}| - |\mathbf{n}| \ge 0, \tag{2.11}$$

where the last inequality follows from the assumption (2.7).

Let us extend the sequence $\{c_{i,j}\}_{i=1,\ldots,k}^{j=0,\ldots,n_i-1}$ to $\{c_{i,j}\}_{i=1,\ldots,k}^{j=0,\ldots,2d_i-1}$; the extending terms c_{i,n_i+r} $(r = 0, \ldots, 2d_i - n_i - 1)$ are arbitrary nonzero complex numbers (in fact we need only c_{i,n_i} to be nonzero in case where $n_i = d_i$). Let H be the solution of the Stein equation

$$H - T_{2\mathbf{d}} H T_{2\mathbf{d}}^* = E_{2\mathbf{d}} E_{2\mathbf{d}}^* - C_{2\mathbf{d}} C_{2\mathbf{d}}^* \tag{2.12}$$

where $2\mathbf{d} := (2d_1, \ldots, 2d_k)$ and where the matrices $T_{2\mathbf{d}} := T_{2\mathbf{d}}(\mathbf{z})$, $E_{2\mathbf{d}}$ and $C_{2\mathbf{d}}$ are defined via formulas (1.2) and (1.4). It is clear that the principal submatrix $H_{\mathbf{n}}$ of H is equal to the original matrix P. We will show that

$$\pi(H) = \nu(H) = |\mathbf{d}|. \tag{2.13}$$

Assuming that (2.13) is already proved, we complete the proof of the theorem as follows. Since P is an $|\mathbf{n}| \times |\mathbf{n}|$ principal submatrix of the $2|\mathbf{d}| \times 2|\mathbf{d}|$ Hermitian matrix H, it follows by the Cauchy's interlacing theorem that

$$\pi(P) \ge \pi(H) - (2|\mathbf{d}| - |\mathbf{n}|) \quad \text{and} \quad \nu(P) \ge \nu(H) - (2|\mathbf{d}| - |\mathbf{n}|)$$

which together with (2.13) imply

$$\pi(P) \geq |\mathbf{n}| - |\mathbf{d}| \geq 0 \quad \text{and} \quad \nu(P) \geq |\mathbf{n}| - |\mathbf{d}| \geq 0. \tag{2.14}$$

Since $\pi(P)+\nu(P)+\delta(P) = |\mathbf{n}|$, equalities (2.9) follow from inequalities (2.11) and (2.14). It remains to verify (2.13). To this end, let

$$F_{i,d_i} := \begin{bmatrix} c_{i,d_i} \\ c_{i,d_i+1} \\ \vdots \\ c_{i,2d_i-1} \end{bmatrix} \quad \text{and} \quad \mathbf{F}_{i,d_i} := \begin{bmatrix} c_{i,d_i} & 0 & \dots & 0 \\ c_{i,d_i+1} & c_{i,d_i} & \ddots & \vdots \\ \vdots & \ddots & \ddots & 0 \\ c_{i,2d_i-1} & \dots & c_{i,d_i+1} & c_{i,d_i} \end{bmatrix}, \tag{2.15}$$

and let

$$F_{\mathbf{d}} = \begin{bmatrix} F_{1,d_1} \\ \vdots \\ F_{k,d_k} \end{bmatrix} \quad \text{and} \quad \mathbf{F}_{\mathbf{d}} = \begin{bmatrix} \mathbf{F}_{1,d_1} & & 0 \\ & \ddots & \\ 0 & & \mathbf{F}_{k,d_k} \end{bmatrix}. \tag{2.16}$$

Since $C_{2\mathbf{d}} = \mathrm{Col}_{1\leq i\leq k} C_{i,2d_i}$, since $C_{i,2d_i} = \begin{bmatrix} C_{i,d_i} \\ F_{i,d_i} \end{bmatrix}$ and since $C_{i,d_i} = \gamma E_{d_i}$ (by assumptions (2.1) and (2.8)), we have eventually

$$C_{2\mathbf{d}} = \mathrm{Col}_{1\leq i\leq k} \begin{bmatrix} \gamma E_{d_i} \\ F_{i,d_i} \end{bmatrix}. \tag{2.17}$$

Let U be the $2|\mathbf{d}| \times |2\mathbf{d}|$ permutation matrix defined by $U = \begin{bmatrix} U_1 \\ U_2 \end{bmatrix}$ where U_1 and U_2 are diagonal block matrices with the ith diagonal blocks

$$U_{1,i} = \begin{bmatrix} I_{d_i} & 0_{d_i} \end{bmatrix} \quad \text{and} \quad U_{2,i} = \begin{bmatrix} 0_{d_i} & I_{d_i} \end{bmatrix}.$$

Then

$$UT_{2\mathbf{d}}U^* = \begin{bmatrix} T_{\mathbf{d}} & 0 \\ * & T_{\mathbf{d}} \end{bmatrix}, \quad UE_{2\mathbf{d}} = \begin{bmatrix} E_{\mathbf{d}} \\ 0 \end{bmatrix}, \quad UC_{2\mathbf{d}} = \begin{bmatrix} \gamma E_{\mathbf{d}} \\ F_{\mathbf{d}} \end{bmatrix}, \tag{2.18}$$

where the two first equalities follow from definitions (1.2) and the third is a consequence of (2.17). We also have (by (2.10))

$$UHU^* = \begin{bmatrix} P_{\mathbf{d}} & B^* \\ B & D \end{bmatrix} = \begin{bmatrix} 0 & B^* \\ B & D \end{bmatrix} \qquad (B,\, D \in \mathbb{C}^{|\mathbf{d}|\times|\mathbf{d}|}). \tag{2.19}$$

Multiplying both parts of (2.12) by U on the left and by U^* on the right and making use of (2.18) and (2.19), we get

$$\begin{bmatrix} 0 & B \\ B^* & D \end{bmatrix} - \begin{bmatrix} T_{\mathbf{d}} & 0 \\ * & T_{\mathbf{d}} \end{bmatrix} \begin{bmatrix} 0 & B \\ B^* & D \end{bmatrix} \begin{bmatrix} T_{\mathbf{d}}^* & * \\ 0 & T_{\mathbf{d}}^* \end{bmatrix} = \begin{bmatrix} 0 & -\gamma E_{\mathbf{d}} F_{\mathbf{d}}^* \\ -\gamma^* F_{\mathbf{d}} E_{\mathbf{d}}^* & -F_{\mathbf{d}} F_{\mathbf{d}}^* \end{bmatrix}.$$

Comparison the 21-blocks in the latter matrix equality gives

$$B - T_{\mathbf{d}} B T_{\mathbf{d}}^* = -\gamma^* F_{\mathbf{d}} E_{\mathbf{d}}^*. \tag{2.20}$$

Let $Q \in \mathbb{C}^{|\mathbf{d}|\times|\mathbf{d}|}$ be the unique solution of the Stein equation

$$Q - T_{\mathbf{d}} Q T_{\mathbf{d}}^* = E_{\mathbf{d}} E_{\mathbf{d}}^*. \tag{2.21}$$

Since the pair $(E_{\mathbf{d}}, T_{\mathbf{d}})$ is observable, the matrix Q is positive definite. Multiplying both parts of (2.21) on the left by the matrix $\mathbf{F_d}$ given in (2.16) and taking into account that $T_{\mathbf{d}}\mathbf{F_d} = \mathbf{F_d}T_{\mathbf{d}}$ and $\mathbf{F_d}E_{\mathbf{d}} = F_{\mathbf{d}}$, we arrive at

$$\mathbf{F_d}Q - T_{\mathbf{d}}\mathbf{F_d}QT_{\mathbf{d}}^* = E_{\mathbf{d}}E_{\mathbf{d}}^*.$$

Comparing the latter equality with (2.20) we conclude that the matrix $-\gamma\mathbf{F_d}Q$ and B solve the same Stein equation. Since this equation has a unique solution, it follows that $B = -\gamma\mathbf{F_d}Q$. Since $\gamma \neq 0$ and since the matrices Q and $\mathbf{F_d}$ are not singular (recall that the diagonal entries of the triangular matrix $\mathbf{F_d}$ are all nonzero either by assumption (2.8) or by construction), it follows that B is not singular. Then the matrix UHU^* in (2.19) is invertible and therefore, it has equally many positive and negative eigenvalues (this is a direct consequence of the Cauchy's interlacing theorem). Thus, $\pi(UHU^*) = \nu(UHU^*) = |\mathbf{d}|$, and (2.13) follows, which completes the proof of the theorem. □

Corollary 2.3. *Let P belong to $\mathcal{D}_{\mathbf{n}}$. Then*

$$\pi(P) = \sum_{i=1}^{k} \pi(P_{ii}), \quad \nu(P) = \sum_{i=1}^{k} \nu(P_{ii}), \quad \delta(P) = \sum_{i=1}^{k} \delta(P_{ii}). \tag{2.22}$$

Proof. Applying (2.9) separately to each diagonal block P_{ii} we get

$$\pi(P_{ii}) = \nu(P_{ii}) = n_i - d_i \quad \text{and} \quad \delta(P_{ii}) = 2d_i - n_i$$

for $i = 1, \ldots, k$ which together with (2.9) (applied to the whole P) imply the statement.

Corollary 2.4. *If $P \geq 0$ ($P \leq 0$) belongs to $\mathcal{D}_{\mathbf{n}}$, then $P = 0$.*

Proof. The statement follows from (2.9), since in this case $\nu(P) = 0$ ($\pi(P) = 0$).

Corollary 2.5. *Let $\mathbf{n} = (n_1, \ldots, n_k) \in \mathbb{N}^k$ with $n_i \leq 2$ for $i = 1, \ldots, k$. If $P \in \mathcal{D}_{\mathbf{n}}$, then $P = 0$.*

Proof. By Theorem 2.2, equalities (2.1) hold and moreover, $c_{i,1} = 0$ whenever $n_i = 2$. Then the right-hand side matrix in the Stein equation (1.3) is the zero matrix and the result follows since the homogeneous Stein equation $P - \mathbf{T_n}(\mathbf{z})P\mathbf{T_n}(\mathbf{z})^* = 0$ has only trivial solution.

3. The boundary case

In this section we consider the Stein equation (1.3) where $T_{\mathbf{n}}(\mathbf{z})$, $E_{\mathbf{n}}$ and $C_{\mathbf{n}}$ are still given by formulas (1.2), (1.4), but the (distinct) points $t_1, \ldots, t_k$ fall on the unit circle $\mathbb{T}$ rather than inside the unit disk:

$$P - T_{\mathbf{n}}(\mathbf{t})PT_{\mathbf{n}}(\mathbf{t})^* = E_{\mathbf{n}}E_{\mathbf{n}}^* - C_{\mathbf{n}}C_{\mathbf{n}}^*, \tag{3.1}$$

where $\mathbf{t} = (t_1, \ldots, t_k) \in \mathbb{T}^k$. In this "boundary" situation, the Stein equation (3.1) either has no solutions or it has infinitely many of them. In the latter case, the structure of every solution P is determined by $|\mathbf{n}|$ additional parameters which

will be now explained (we refer to [5, Section 10], [6, Section 3] and [7, Section 2] for proofs and some more detail). Given $t \in \mathbb{T}$ and $n \in \mathbb{N}$, let

$$\mathbf{\Psi}_n(t) = \begin{bmatrix} t & -t^2 & t^3 & \cdots & (-1)^{n-1}\binom{n-1}{0} t^n \\ 0 & -t^3 & 2t^4 & \cdots & (-1)^{n-1}\binom{n-1}{1} t^{n+1} \\ \vdots & & t^5 & \cdots & (-1)^{n-1}\binom{n-1}{2} t^{n+2} \\ \vdots & & & \ddots & \vdots \\ 0 & \cdots & \cdots & 0 & (-1)^{n-1}\binom{n-1}{n-1} t^{2n-1} \end{bmatrix}, \tag{3.2}$$

be the upper triangular matrix with the entries $\Psi_{j\ell} = (-1)^{\ell}\binom{\ell}{j} t^{\ell+j+1}$ for $0 \le j \le \ell \le n-1$. Let $\mathbf{C}_{i,n_i}$ be the lower triangular Toeplitz matrix

$$\mathbf{C}_{i,n_i} = \begin{bmatrix} c_{i,0} & 0 & \dots & 0 \\ c_{i,1} & c_{i,0} & \dots & 0 \\ \vdots & \ddots & \ddots & \vdots \\ c_{i,n_i-1} & \dots & c_{i,1} & c_{i,0} \end{bmatrix}, \qquad i = 1, \dots, k, \tag{3.3}$$

so that $\mathbf{C}_{i,n_i} E_{n_i} = C_{i,n_i}$. A necessary and sufficient condition for the Stein equation (3.1) to have a solution is that

$$\mathbf{C}_{i,n_i}^{\top} \mathbf{\Psi}_{n_i}(t_i) \mathbf{C}_{i,n_i}^{*} = \mathbf{\Psi}_{n_i}(t_i) \quad \text{for} \quad i = 1, \dots, k, \tag{3.4}$$

where $^{\top}$ stands for the transpose. Equating the left upper corner entries in (3.4) gives

$$|c_{i,0}| = 1 \quad \text{for} \quad i = 1, \dots, k. \tag{3.5}$$

Sequences $\{c_{i,j}\}_{j=0}^{n_i-1}$ satisfying (3.7) can be extended (not uniquely) to $\{c_{i,j}\}_{j=0}^{2n_i-1}$ so that

$$\mathbf{C}_{i,2n_i}^{\top} \mathbf{\Psi}_{2n_i}(t_i) \mathbf{C}_{i,2n_i}^{*} = \mathbf{\Psi}_{2n_i}(t_i) \qquad (i = 1, \dots, k) \tag{3.6}$$

and in general, (3.4) and (3.5) follow from (3.6) due to upper triangular structure of matrices in (3.6). The next theorem (in a slightly different formulation) can be found in [6].

Theorem 3.1. *Given* $\mathbf{t} = (t_1, \dots, t_k) \in \mathbb{T}^k$ *and* $\{c_{i,j}\}_{i=1,\dots,k}^{j=0,\dots,2n_i-1}$, *let*

$$P = [P_{ij}]_{i,j=1}^{k} \tag{3.7}$$

be the block matrix with the $n_i \times n_j$ *blocks* P_{ij} *given by*

$$\begin{aligned} [P_{ij}]_{\ell,r} \;=\; & \sum_{s=0}^{\min\{\ell,r\}} \frac{(\ell+r-s)!}{(\ell-s)!s!(r-s)!} \frac{t_i^{r-s}\bar{t}_j^{\ell-s}}{(1-t_i\bar{t}_j)^{\ell+r-s+1}} \\ & - \sum_{\alpha=0}^{\ell}\sum_{\beta=0}^{r} \sum_{s=0}^{\min\{\alpha,\beta\}} \frac{(\alpha+\beta-s)!}{(\alpha-s)!s!(\beta-s)!} \frac{t_i^{\beta-s}\bar{t}_j^{\alpha-s} c_{i,\ell-\alpha} c_{j,r-\beta}^{*}}{(1-t_i\bar{t}_j)^{\alpha+\beta-s+1}} \quad \text{if} \quad i \ne j \end{aligned} \tag{3.8}$$

and

$$P_{ii} := H_{i,n_i} \mathbf{\Psi}_{n_i}(t_i) \mathbf{C}_{i,n_i}^{*} \quad \text{where} \quad H_{i,n_i} = [c_{i,\ell+j+1}]_{\ell,j=0}^{n_i-1} \quad (i = 1, \dots, k). \tag{3.9}$$

The following are equivalent

1. *Conditions* (3.6) *hold.*
2. *P is Hermitian and satisfies the Stein identity* (3.1).
3. *P is Hermitian and $|c_{i,0}| = 1$ for $i = 1, \ldots, k$.*

Moreover, if equation (3.1) *has a solution* (*i.e., if conditions* (3.4) *are in force*), *then every Hermitian solution P is necessarily of the form* (3.7)–(3.9) *for some $c_{i,j}$ $(i = 1, \ldots, k;\ j = n_i, \ldots, 2n_i - 1)$ satisfying extended conditions* (3.6).

Hermitian matrices of the form (3.7)–(3.9) can be generated by rational functions unimodular on $\mathbb{T}$. Given such a function f and given a tuple $\mathbf{t} = (t_1, \ldots, t_k) \in \mathbb{T}$, let us define the *boundary Schwarz-Pick matrix* by

$$P_{\mathbf{n}}^f(\mathbf{t}) = \lim_{\mathbf{z} \to \mathbf{t}} P_{\mathbf{n}}^f(\mathbf{z}) \tag{3.10}$$

where $\mathbf{z} \in \mathbb{D}^k$ and $P_{\mathbf{n}}^f(\mathbf{z})$ is defined as in (1.10). The proof of the next theorem can be found in [7].

Theorem 3.2. *Let $\mathbf{t} = (t_1, \ldots, t_k) \in \mathbb{T}^k$ and let f be a rational function unimodular on $\mathbb{T}$. Then the numbers*

$$c_{i,j} = \frac{f^{(j)}(t_i)}{j!} \qquad (i = 1, \ldots, k;\ j = 0, \ldots, 2n_i - 1)$$

satisfy conditions (3.6). *Furthermore, the limit in* (3.10) *exists and is equal to the matrix P defined by formulas* (3.7)–(3.9).

Now we address the questions from Section 2 to the present boundary setting.

Definition 3.3. We will say that P of the form (3.7)–(3.9) and with parameters $\{c_{i,j}\}$ satisfying conditions (3.6) belongs to the class $\mathcal{B}_{\mathbf{n}}$ if for every $\mathbf{m} \preceq \mathbf{n}$, the principal submatrix $P_{\mathbf{m}}$ of P defined as in (1.8) is singular.

The next theorem is an analog of Proposition 2.1. In contrast to the interior case, the matrices of the class $\mathcal{B}_{\mathbf{n}}$ are block diagonal.

Proposition 3.4. *Let $\mathbf{t} = (t_1, \ldots, t_k) \in \mathbb{T}^k$ and $\mathbf{n} = (n_1, \ldots, n_k) \in \mathbb{N}^k$, let P be defined as in* (3.7)–(3.9) *and let us assume that conditions* (3.6) *are satisfied. Then $P \in \mathcal{B}_{\mathbf{n}}$ if and only if* (2.1) *holds and*

$$c_{i,j} = 0 \quad \text{for every} \quad i = 1, \ldots, k \ \text{ and } \ 1 \le j \le n_i. \tag{3.11}$$

In this case, the block entries of P simplify to

$$P_{ii} = \gamma H_{i,n_i} \mathbf{\Psi}_{n_i}(t_i) \quad \text{and} \quad P_{ij} = 0 \ \ (i \ne j). \tag{3.12}$$

Proof. Since conditions (3.6) are satisfied, then in particular, we have (3.5) and therefore, the matrices $\mathbf{C}_{i,j}$ defined via formula (3.3) are invertible for every $i = 1, \ldots, k$ and $j = 0, \ldots, n_i - 1$. Since $t_i \in \mathbb{T}$, the triangular matrices $\mathbf{\Psi}_j(t_i)$ are also invertible.

Let $P \in \mathcal{B}_{\mathbf{n}}$. Then $P_{ii} \in \mathcal{B}_{n_i}$ (i.e., all leading submatrices of P_{ii} are singular) for every $i = 1, \ldots, k$. It follows from the "Hankel-$\mathbf{\Psi}$-Toeplitz" structure (3.9) of P_{ii} that the $j \times j$ leading submatrix of P_{ii} equals $H_{i,j}\mathbf{\Psi}_j(t_i)\mathbf{C}_{i,j}^*$ and therefore, the membership of P_{ii} in $\mathcal{B}_{n_i}$ implies that

$$\det H_{i,j} = 0 \quad \text{for every} \quad j = 0, \ldots, n_i - 1. \tag{3.13}$$

Now we recursively get (3.11) from (3.13). Indeed, letting $j = 0$ in (3.13) we get $c_{i,1} = 0$; assuming that $c_{i,j} \neq 0$ for some $j \in \{1, \ldots, n_i\}$ and $c_{i,r} = 0$ for $r = 1, \ldots, j-1$ we get

$$\det H_{i,j} = \det \begin{bmatrix} 0 & \ldots & 0 & c_{i,j} \\ \vdots & \iddots & \iddots & c_{i,j+1} \\ 0 & \iddots & \iddots & \vdots \\ c_{i,j} & c_{i,j+1} & \ldots & c_{i,2j-1} \end{bmatrix} \neq 0,$$

which contradicts (3.13) and proves (3.11).

Equalities (2.1) are obtained in much the same way as in the proof of Theorem 2.1. We take the tuple $\mathbf{m} = (m_1, \ldots, m_k)$ with $m_i = m_j = 1$ and $m_\ell = 0$ for $\ell \neq i, j$. Since the leading entries in the blocks P_{ii} and P_{jj} are equal to zeros, the submatrix $P_{\mathbf{m}}$ takes the form

$$P_{\mathbf{m}} = \begin{bmatrix} 0 & \dfrac{1 - c_{i,0}c_{j,0}^*}{1 - t_i\bar{t}_j} \\ \dfrac{1 - c_{j,0}c_{i,0}^*}{1 - t_j\bar{t}_i} & 0 \end{bmatrix}.$$

Since $P \in \mathcal{B}_{\mathbf{n}}$, the matrix $P_{\mathbf{m}}$ must be singular which implies that $c_{i,0} = c_{j,0}$ for $i, j = 1, \ldots, k$. The latter implies (2.1) due to (3.5).

On account of (2.1) and (3.11), $\mathbf{C}_{i,n_i} = \gamma I_{n_i}$ and therefore, formula (3.9) for P_{ii} collapses to the first formula in (3.12). On the other hand, substituting (2.1) and (3.11) into (3.8) gives

$$\begin{aligned} [P_{ij}]_{\ell,r} &= \sum_{s=0}^{\min\{\ell,r\}} \frac{(\ell + r - s)!}{(\ell - s)!s!(r - s)!} \frac{t_i^{r-s}\bar{t}_j^{\ell-s}(1 - c_{i,0}c_{j,0}^*)}{(1 - t_i\bar{t}_j)^{\ell+r-s+1}} \\ &= \sum_{s=0}^{\min\{\ell,r\}} \frac{(\ell + r - s)!}{(\ell - s)!s!(r - s)!} \frac{t_i^{r-s}\bar{t}_j^{\ell-s}(1 - |\gamma|^2)}{(1 - t_i\bar{t}_j)^{\ell+r-s+1}} = 0 \end{aligned}$$

for every $\ell \leq n_i$ and $r \leq n_j$ which proves the second equality in (3.12).

Finally, let us assume that conditions (2.1) and (3.11) hold. Then representations (3.12) are in force. Furthermore, since $c_{i,1} = \cdots = c_{i,n_i} = 0$ and since the matrix $\mathbf{\Psi}_{n_i}(t_i)$ is upper triangular, it follows from (3.12) that all the entries of P_{ii} on and above the main "southwest-northeast" diagonal are zeroes. Therefore every leading submatrix of P_{ii} is singular and thus, $P_{ii} \in \mathcal{B}_{n_i}$ for every $i = 1, \ldots, k$. Since all non-diagonal blocks in P are zero matrices, $P \in \mathcal{B}_{\mathbf{n}}$. □

The next theorem is the "boundary" analog of Theorem 2.2. It establishes explicit formulas for inertia of a matrix $P \in \mathcal{B}_{\mathbf{n}}$ of the form (3.7)–(3.9). Since every such P is necessarily block diagonal (by (3.12)), it suffices to consider the one-block case. We will use the previous notation with the subscript i dropped.

Theorem 3.5. *Let $t \in \mathbb{T}$, let $c_0, \dots, c_{2n-1}$ be the complex numbers such that*

$$\mathbf{C}_{2n}^{\top} \mathbf{\Psi}_{2n}(t) \mathbf{C}_{2n}^{*} = \mathbf{\Psi}_{2n}(t) \tag{3.14}$$

and let

$$P_n := H_n \mathbf{\Psi}_n(t) \mathbf{C}_n^{*}, \tag{3.15}$$

where the matrices H_n and $\mathbf{C}_n$ are constructed via formulas (3.9) and (3.3). Let

$$c_d \neq 0 \quad \text{and} \quad c_1 = c_2 = \cdots = c_{d-1} = 0 \quad \text{for some } d\ (n < d \leq 2n-1). \tag{3.16}$$

Then $\delta(P_n) = d - n$. Furthermore,

1. *If $d = 2k$, then $\pi(P) = \nu(P) = n - k$.*
2. *If $d = 2k+1$, then $t^d c_d c_0^{*} \in \mathbb{R} \setminus \{0\}$ and*
   (a) *$\nu(P_n) = \pi(P_n) + 1 = n - k$ if $(-1)^k \mathbf{sgn}\,(t^d c_d c_0^{*}) < 0$.*
   (b) *$\pi(P_n) = \nu(P_n) + 1 = n - k$ if $(-1)^k \mathbf{sgn}\,(t^d c_d c_0^{*}) > 0$.*

Proof. Condition (3.14) implies $|c_0| = 1$ and therefore, $\mathbf{C}_n = c_0 I_n$ so that formula (3.15) reads

$$P_n = c_0^{*} H_n \mathbf{\Psi}_n(t). \tag{3.17}$$

Let $\mathbf{D}_j(A)$ $(j = 1, \dots, 2n-1)$ denote the jth "southwest-northeast" diagonal of a matrix $A \in \mathbb{C}^{n \times n}$. With a self-evident interpretation,

$$\mathbf{D}_j(A) = \begin{cases} (a_{j,1},\ a_{j-1,2},\ \dots,\ a_{1,j}), & \text{if} \quad 1 \leq j \leq n, \\ (a_{n,j-n+1},\ a_{n-1,j-n+2},\ \dots,\ a_{j-n+1,n}), & \text{if} \quad n+1 \leq j \leq 2n-1. \end{cases}$$

Since $\mathbf{\Psi}_n(t)$ is invertible for every $t \neq 0$ and $n \in \mathbb{N}$, it follows from (3.17) that $\operatorname{rank}(P_n) = \operatorname{rank}(H_n)$. Assumptions (3.16) and the Hankel structure of H_n imply that

$$\mathbf{D}_j(H_n) = 0 \ \ (j = 0, \dots, d-1) \quad \text{and} \quad \mathbf{D}_d(H_n) = (c_d, c_d, \dots, c_d) \neq 0.$$

Since $d > n$, it now follows that $\operatorname{rank}(H_n) = 2n - d$ and therefore,

$$\delta(P_n) = \delta(H_n) = n - (2n - d) = d - n. \tag{3.18}$$

Let us extend the sequence $\{c_i\}_{i=0}^{2n-1}$ to $\{c_i\}_{i=0}^{2d-1}$ in such a way that

$$\mathbf{C}_{2d}^{\top} \mathbf{\Psi}_{2d}(t) \mathbf{C}_{2d}^{*} = \mathbf{\Psi}_{2d}(t), \tag{3.19}$$

where $\mathbf{C}_{2d}$ and $\mathbf{\Psi}_{2d}$ are defined via formulas (3.3) and (3.2). Such an extension is always possible (see [7, Section 2], where these extensions were called t-isometric). Now we can introduce the Hankel matrix $H_d := [c_{i+j+1}]_{i,j=0}^{d-1}$ and the matrix

$$P_d := H_d \mathbf{\Psi}_d(t) \mathbf{C}_d^{*} \tag{3.20}$$

which is Hermitian due to (3.19) by Theorem 3.1. It is readily checked that

$$H_d = \begin{bmatrix} H_n & * \\ * & * \end{bmatrix}, \quad \mathbf{C}_d = \begin{bmatrix} \mathbf{C}_n & * \\ 0 & * \end{bmatrix}, \quad \mathbf{\Psi}_d(t) = \begin{bmatrix} \mathbf{\Psi}_n(t) & * \\ 0 & * \end{bmatrix}, \quad P_d = \begin{bmatrix} P_n & * \\ * & * \end{bmatrix}.$$

Since all the entries above the main diagonal $\mathbf{D}_d(H_d)$ of H_d are zeros, it follows that

$$\det H_d = (-1)^{d-1} c_d^d. \tag{3.21}$$

Since the matrices $\mathbf{\Psi}_d(t)$ and $\mathbf{C}_d^*$ are upper triangular, their determinants are equal to products of diagonal entries, so that

$$\det \mathbf{\Psi}_d(t) = \prod_{j=0}^{d-1} (-1)^j t^{2j+1} = (-1)^{\frac{d(d-1)}{2}} t^{d^2} \quad \text{and} \quad \det \mathbf{C}_d^* = (c_0^*)^d,$$

which together with (3.21) and (3.20) gives

$$\det P_d = (-1)^{\frac{(d+2)(d-1)}{2}} (t^d c_0^* c_d)^d. \tag{3.22}$$

By triangular structure of H_d, $\mathbf{\Psi}_d(t)$ and $\mathbf{C}_d^*$, we conclude from (3.20) that $\mathbf{D}_j(P_d) = 0$ for $j = 1, \ldots, d-1$. Note also that P_d is Hermitian and invertible: $\det H_d \neq 0$, by (3.21). Therefore, if $d = 2k$, then P_d is of the form

$$P_d = \begin{bmatrix} 0 & B \\ B^* & D \end{bmatrix}, \quad B \in \mathbb{C}^{k \times k}, \ \det B \neq 0.$$

Therefore

$$\pi(P_d) = \nu(P_d) = k. \tag{3.23}$$

Since $P_n \in \mathbb{C}^{n \times n}$ is a principal submatrix of $P_d \in \mathbb{C}^{d \times d}$, it follows by the Cauchy's interlacing theorem that

$$\pi(P_n) \geq \pi(P_d) - (d-n) \quad \text{and} \quad \nu(P_n) \geq \nu(P_d) - (d-n). \tag{3.24}$$

Since $d = 2k < 2n$, the latter inequalities and equalities (3.23) imply

$$\pi(P_n) \geq k - (2k-n) = n - k > 0 \quad \text{and} \quad \nu(P_n) \geq n - k > 0. \tag{3.25}$$

Since $\pi(P_n) + \nu(P_n) + \delta(P_n) = n$, it follows from (3.18) and (3.25) that in fact, $\pi(P_n) = n - k$ and $\nu(P_n) = n - k$.

Finally, let us assume that $d = 2k + 1$ and let p_{kk} be the "central" entry of P_d. Since P_d is Hermitian, $p_{kk} \in \mathbb{R}$ and P_d takes the form

$$P_d = \begin{bmatrix} 0 & 0 & B \\ 0 & p_{kk} & X \\ B^* & X^* & D \end{bmatrix}, \quad \text{where} \quad B \in \mathbb{C}^{k \times k} \quad \text{and} \quad X \in \mathbb{C}^{1 \times k}. \tag{3.26}$$

Since P_d is invertible, we have $\det B \neq 0$ and $p_{kk} \neq 0$. Now we will show that

$$\mathbf{sgn}\,(p_{kk}) = \mathbf{sgn}\,(\det P_d) = (-1)^k \mathbf{sgn}\,(t^k c_d c_0^*). \tag{3.27}$$

Indeed, since the matrix B in (3.26) is upper triangular with respect to its main "southwest-northeast" diagonal $\mathbf{D}_k(B)$, we have $\det P_d = p_{kk} \cdot |\det B|^2$ and the first equality in (3.27) follows. To verify the second, let us equate the entries on

the intersection of the first row and of the $(d+1)$-st column in matrix equality (3.19); by definitions (3.2) and (3.3) of matrices $\mathbf{C}_{2d}$ and $\mathbf{\Psi}_{2d}(t)$ we have

$$\begin{bmatrix} c_0 & c_1 & \dots & c_{2d} \end{bmatrix} \mathbf{\Psi}_{2d}(t) \begin{bmatrix} c_0^* & c_1^* & \dots & c_d^* & 0 & \dots & 0 \end{bmatrix}^\top = (-1)^{d-1} t^{d+2},$$

which in view of (3.2), (3.16) collapses to

$$tc_0c_d^* + (-1)^d t^{2d+1} c_d c_0^* + (-1)^{d-1} t^{d+2} |c_0|^2 = (-1)^{d-1} t^{d+2}.$$

Since d is odd and since $|c_0| = 1$, we get $tc_0c_d^* - t^{2d+1}c_dc_0^* = 0$. Multiplying the latter equality by $\bar{t}^{d+1}$ and taking into account that $|t| = 1$, we get $\bar{t}^d c_0 c_d^* - t^d c_d c_0^* = 0$. Therefore, $t^d c_d c_0^* \in \mathbb{R}$ and since d is odd,

$$\mathbf{sgn}\,(t^k c_d c_0^*) = \mathbf{sgn}\,(t^k c_d c_0^*)^d. \tag{3.28}$$

Substituting $d = 2k+1$ into (3.22) gives

$$\det P_d = (-1)^{k(2k+3)} (t^d c_0^* c_d)^d = (-1)^k (t^d c_0^* c_d)^d$$

which together with (3.28) implies the second equality in (3.27). To complete the proof of the theorem, observe that the Schur complement $\mathbf{S}$ of p_{kk} in P_d equals

$$\mathbf{S} = \begin{bmatrix} 0 & B \\ B^* & D \end{bmatrix} - \begin{bmatrix} 0 \\ X^* \end{bmatrix} p_{kk}^{-1} \begin{bmatrix} 0 & X \end{bmatrix} = \begin{bmatrix} 0 & B \\ B^* & D - X^* p_{kk}^{-1} X \end{bmatrix}$$

and has equally many positive and negative eigenvalues by the preceding analysis. Since $\pi(P_d) = \pi(p_{kk}) + \pi(\mathbf{S})$ and $\nu(P_d) = \nu(p_{kk}) + \nu(\mathbf{S})$, it follows that P_d has k positive eigenvalues, k negative eigenvalues and one more eigenvalue whose sign coincides with that of p_{kk}. According to (3.28),

$$\pi(P_d) = \nu(P_d) - 1 = k \quad \text{if} \quad (-1)^k \mathbf{sgn}\,(t^d c_d c_0^*) < 0 \tag{3.29}$$

and

$$\nu(P_d) = \pi(P_d) - 1 = k \quad \text{if} \quad (-1)^k \mathbf{sgn}\,(t^d c_d c_0^*) > 0. \tag{3.30}$$

Now we again apply the interlace theorem to get inequalities (3.24) which then turn out to be equalities:

$$\pi(P_n) = \pi(P_d) - (d-n) \quad \text{and} \quad \nu(P_n) = \nu(P_d) - (d-n),$$

and then the second statement of the theorem follows from (3.29) and (3.30). Note that statement (1) also covers the trivial case $d = 2n$ (which was not included in (3.16)). In this case $P_n = 0$ which agrees with statement (1) asserting that $\pi(P_n) = \nu(P_n) = n - n = 0$. □

In conclusion we note that Corollary 2.5 does not hold for $P \in \mathcal{B}_{\mathbf{n}}$, whereas Corollary 2.3 does (which is obvious due to the block diagonal structure of $P \in \mathcal{B}_{\mathbf{n}}$). As for Corollary 2.4, we have the following boundary analog which we formulate for the single-block case and which follows immediately from Theorem 3.5: *If* $P \geq 0$ *(*$P \leq 0$*) belongs* $\mathcal{B}_n$*, then* $\operatorname{rank}(P) \leq 1$.

References

[1] D. Alpay and H. Dym, *Structured invariant spaces of vector-valued rational functions, Hermitian matrices, and a generalization of the Iohvidov laws*, Linear Algebra Appl. **137/138** (1990), 137–181.

[2] D. Alpay and H. Dym, *Structured invariant spaces of vector-valued functions, sesquilinear forms, and a generalization of the Iohvidov laws*, Linear Algebra Appl. **137/138** (1990), 413–451

[3] D. Alpay and H. Dym, *On a new class of reproducing kernel spaces and a new generalization of the Iohvidov laws*, Linear Algebra Appl. **178** (1993), 109–183.

[4] J.A. Ball, I. Gohberg, and L. Rodman, *Interpolation of rational matrix functions*, OT45, Birkhäuser Verlag, 1990.

[5] V. Bolotnikov and H. Dym, *On boundary interpolation for matrix Schur functions*, Mem. Amer. Math. Soc. 181 (2006), no. 856.

[6] V. Bolotnikov and A. Kheifets, *The higher-order Carathéodory–Julia theorem and related boundary interpolation problems*, Operator Theory: Advances and Applications **OT 179** (2007), 63–102.

[7] V. Bolotnikov and A. Kheifets, *Carathéodory-Julia type conditions and symmetries of boundary asymptotics for analytic functions on the unit disk*, Math. Nachr. **282** (2009), no. 11, 1513–1536.

[8] I.S. Iohvidov, *Hankel and Toeplitz matrices and forms. Algebraic theory.* Birkhäuser, Boston, Mass., 1982.

Vladimir Bolotnikov
Department of Mathematics
The College of William and Mary
Williamsburg, VA 23187-8795, USA
e-mail: vladi@math.wm.edu

Operator Theory:
Advances and Applications, Vol. 199, 191–202

Variable-coefficient Toeplitz Matrices with Symbols beyond the Wiener Algebra

Albrecht Böttcher and Sergei Grudsky

In Memory of Georg Heinig

Abstract. Sequences of so-called variable-coefficient Toeplitz matrices arise in many problems, including the discretization of ordinary differential equations with variable coefficients. Such sequences are known to be bounded if the generating function satisfies a condition of the Wiener type, which is far away from the minimal requirement in the case of constant coefficients. The purpose of this paper is to uncover some phenomena beyond the Wiener condition. We provide counterexamples on the one hand and prove easy-to-check sufficient conditions for boundedness on the other.

Mathematics Subject Classification (2000). Primary 47B35; Secondary 15A60, 65F35.

Keywords. Toeplitz matrix, variable coefficients, matrix norm.

1. Introduction

Let a be a complex-valued continuous function on $[0,1] \times [0,1] \times \mathbb{T}$, where $\mathbb{T}$ is the complex unit circle, $a : [0,1] \times [0,1] \times \mathbb{T} \to \mathbb{C}$, $(x,y,t) \mapsto a(x,y,t)$. For $n \in \mathbb{Z}$, we put $\hat{a}_n(x,y) = \int_{\mathbb{T}} a(x,y,t)\, t^{-n}\, |dt|/(2\pi)$ and so have the Fourier series

$$a(x,y,t) = \sum_{n=-\infty}^{\infty} \hat{a}_n(x,y) t^n, \tag{1}$$

where equality holds at least in the L^2 sense. Let $A_N(a)$ be the matrix

$$A_N(a) = \left(\hat{a}_{j-k} \left(\frac{j}{N}, \frac{k}{N} \right) \right)_{j,k=0}^{N}. \tag{2}$$

Occasionally we allow us to write $a(x,y,t)$ and $A_N(a(x,y,t))$ for a and $A_N(a)$.

This work was partially supported by CONACYT project U46936-F, Mexico.

We refer to $A_N(a)$ as a Toeplitz matrix with variable coefficients. Clearly, if a does not depend on x and y, then $A_N(a)$ is a pure Toeplitz matrix. It is easily seen that every $(N+1)\times(N+1)$ matrix may be written as $A_N(a)$ with some a. Thus, the notion of a Toeplitz matrix with variable coefficients is rather an asymptotic notion which makes sense for the entire sequences $\{A_N(a)\}_{N=0}^{\infty}$ but not for a single $(N+1)\times(N+1)$ matrix.

Variable-coefficient Toeplitz matrices and their modifications and generalizations are currently emerging in many applications (see, for example, [1], [4], [5], [6], [7], [8], [11], [12]) and go under various names, such as generalized Toeplitz [9], locally Toeplitz [11] or generalized locally Toeplitz matrices [8], Berezin-Toeplitz matrices [1], twisted Toeplitz matrices [12], or generalized discrete convolutions [10], [13]. In our opinion, variable-coefficient Toeplitz matrices [3] is the perhaps best name, at least when considering matrices of the form (2). The problem of primary interest is the understanding of the spectral and pseudospectral properties of $A_N(a)$ as $N\to\infty$. Accordingly, the works cited above and also [3], [9], [14] deal with extensions of the Szegö and Avram-Parter theorems, that is, with asymptotic formulas for $\operatorname{tr} f(A_N(a))$, and with asymptotic inverses and pseudospectra of $A_N(a)$.

This paper addresses the problem of the uniform boundedness of $A_N(a)$, that is, the question whether $\|A_N(a)\|_\infty$ remains bounded as $N\to\infty$. Here and in what follows, $\|A\|_\infty$ is the spectral norm of A.

If a is independent of x and y, $a(t)=\sum_{n=-\infty}^{\infty}\hat{a}_n t^n$, we denote $A_N(a)$ by $T_N(a)$. Thus, $T_N(a)=(\hat{a}_{j-k})_{j,k=0}^{N}$. It is well known that $\|T_N(a)\|_\infty\le\|T_{N+1}(a)\|_\infty$ for all N and that

$$\lim_{N\to\infty}\|T_N(a)\|_\infty = M_\infty(a) := \sup_{t\in\mathbb{T}}|a(t)|. \tag{3}$$

If a is of the form $a(x,y,t)=b(x,y)t^n$, then

$$A_N(a) = T_N(t^n)\operatorname{diag}\left(b\left(\frac{j+n}{N},\frac{j}{N}\right)\right)_{j=0}^{N}$$

(where $j+n$ is taken modulo $N+1$) and hence

$$\|A_N(b(x,y)t^n)\|_\infty \le M_{\infty,\infty}(b) := \sup_{x\in[0,1]}\sup_{y\in[0,1]}|b(x,y)|.$$

Consequently, for a given by (1) we have

$$\|A_N(a)\|_\infty \le \sum_{n=-\infty}^{\infty} M_{\infty,\infty}(\hat{a}_n). \tag{4}$$

Thus, if

$$\sum_{n=-\infty}^{\infty} M_{\infty,\infty}(\hat{a}_n) < \infty, \tag{5}$$

then $\{\|A_N(a)\|_\infty\}_{N=0}^\infty$ is a bounded sequence. Condition (5) is a condition of the Wiener type. In the case where a does not depend on x and y, it amounts to saying that $\|T_N(a)\|_\infty$ remains bounded as $N \to \infty$ if a belongs to the Wiener algebra, that is, if $\sum_{n=-\infty}^\infty |\hat{a}_n| < \infty$. This is clearly far away from (3) and is therefore a source of motivation for looking whether the uniform boundedness of $\|A_N(a)\|_\infty$ can be guaranteed under weaker assumptions.

The first question one might ask is whether the sole continuity of the generating function a on $[0,1]^2 \times \mathbb{T} := [0,1] \times [0,1] \times \mathbb{T}$ ensures the uniform boundedness of $\|A_N(a)\|_\infty$. We show that, surprisingly, the answer to this question is no.

Theorem 1.1. *There exist* $a \in C([0,1]^2 \times \mathbb{T})$ *such that* $\sup \|A_N(a)\|_\infty = \infty$.

On the other hand, we will prove that $\|A_N(a)\|_\infty$ remains bounded if $a(x,y,t)$ has certain smoothness in x and y. (Notice that (5) is a requirement on the smoothness in t.) Our results will imply the following.

Theorem 1.2. *If* $a \in C^{4,0}([0,1]^2 \times \mathbb{T})$, *which means that the function* $a(x,y,t)$ *has continuous partial derivatives with respect to* x *and* y *up to the order* 4, *then* $\sup \|A_N(a)\|_\infty < \infty$.

For $a(x,y,t)$ independent of x and y, this is equivalent to the statement that $T_N(a)$ is uniformly bounded if $a \in C(\mathbb{T})$. In fact, we can sharpen Theorem 1.2 as follows.

Theorem 1.3. *If* $a \in L^\infty(\mathbb{T}, C^4([0,1]^2))$, *that is, if* a *is an* L^∞ *function on* $\mathbb{T}$ *with values in the Banach space of all functions on* $[0,1]^2$ *that have continuous partial derivatives up to the order* 4, *then* $\sup \|A_N(a)\|_\infty < \infty$.

In the case of constant coefficients, this theorem is best possible: it says that $\sup \|T_N(a)\|_\infty < \infty$ if $a \in L^\infty(\mathbb{T})$.

The next question is whether the exponent 4 is close to a kind of a minimum. Let $0 < \alpha \le 1$. For a continuous function $a(x,y,t)$ on $[0,1]^2 \times \mathbb{T}$ we define

$$M_{\alpha,\infty,\infty}(a) = \sup_{t\in\mathbb{T}} \sup_{y\in[0,1]} \sup_{x_1,x_2} \frac{|a(x_2,y,t) - a(x_1,y,t)|}{|x_2 - x_1|^\alpha},$$

$$M_{\infty,\alpha,\infty}(a) = \sup_{t\in\mathbb{T}} \sup_{x\in[0,1]} \sup_{y_1,y_2} \frac{|a(x,y_2,t) - a(x,y_1,t)|}{|y_2 - y_1|^\alpha},$$

and

$$M_{\alpha,\alpha,\infty}(a) = \sup_{t\in\mathbb{T}} \sup_{x_1,x_2} \sup_{y_1,y_2} \frac{|\Delta_2 a(x_1,x_2,y_1,y_2,t)|}{|x_2 - x_1|^\alpha |y_2 - y_1|^\alpha}$$

where $\Delta_2 a(x_1,x_2,y_1,y_2,t)$ is the second difference

$$\Delta_2 a(x_1,x_2,y_1,y_2,t) = a(x_2,y_2,t) - a(x_2,y_1,t) - (a(x_1,y_2,t) - a(x_1,y_1,t)).$$

Here $\sup_{z_1,z_2}$ means the supremum over all $z_1, z_2 \in [0,1]$ such that $z_1 \ne z_2$. We say that the function $a(x,y,t)$ belongs to $H_{\alpha,\alpha,\infty}$ if the three numbers $M_{\alpha,\alpha,\infty}(a)$,

$M_{\alpha,\infty,\infty}(a)$, $M_{\infty,\alpha,\infty}(a)$ are finite and we denote by $H_{1+\alpha,1+\alpha,\infty}$ the set of all functions $a(x,y,t)$ that have continuous partial derivatives up to the order 2 in x and y and for which the three numbers

$$M_{1+\alpha,\infty,\infty}(a) := M_{\alpha,\infty,\infty}(\partial_x a), \quad M_{\infty,1+\alpha,\infty}(a) := M_{\infty,\alpha,\infty}(\partial_y a),$$
$$M_{1+\alpha,1+\alpha,\infty}(a) := M_{\alpha,\alpha,\infty}(\partial_x \partial_y a)$$

are finite. Notice that $C^{4,0}([0,1]^2 \times \mathbb{T}) \subset H_{2,2,\infty}$. We here prove the following.

Theorem 1.4. *If $\beta < 1/2$, there exist functions a in the space $H_{\beta,\beta,\infty}$ such that $\sup \|A_N(a)\|_\infty = \infty$. If $\beta > 1$, then $\sup \|A_N(a)\|_\infty < \infty$ for every function a in $H_{\beta,\beta,\infty}$.*

The theorem leaves a gap. We can actually remove this gap and prove that $\sup \|A_N(a)\|_\infty < \infty$ whenever $a \in H_{\beta,\beta,\infty}$ and $\beta > 1/2$, which is guaranteed if $a \in C^{2,0}([0,1]^2 \times \mathbb{T})$. However, the proof of this result is very sophisticated. We see the purpose of this paper in revealing the delicacy of the problem of the uniform boundedness of $\|A_N(a)\|_\infty$ and in providing results that might be sufficient for applications. Drawing down things to $\beta > 1/2$ is a matter of mathematical ambition and will be the topic of our subsequent paper [2].

2. Hölder continuity

We already defined the space $H_{\beta,\beta,\infty}$ and the quantities $M_{\beta,\beta,\infty}(a)$, $M_{\beta,\infty,\infty}(a)$, $M_{\infty,\beta,\infty}(a)$ for a continuous function $a(x,y,t)$ on $[0,1]^2 \times \mathbb{T}$ and for $0 < \beta \le 2$. In addition, we put

$$M_{\infty,\infty,\infty}(a) = \sup_{t\in\mathbb{T}} \sup_{x\in[0,1]} \sup_{y\in[0,1]} |a(x,y,t)|.$$

Note that if $a(x,y,t) = x^\gamma + y^\gamma$ with $0 < \gamma < 1$, then

$$M_{\beta,\infty,\infty}(a) = M_{\infty,\beta,\infty}(a) = \infty, \quad M_{\beta,\beta,\infty}(a) = 0$$

for $\gamma < \beta < 1$, which shows that the assumption $M_{\beta,\beta,\infty}(a) < \infty$ does not imply that $M_{\beta,\infty,\infty}(a)$ and $M_{\infty,\beta,\infty}(a)$ are finite. In the introduction we also mentioned that $C^{2,0}([0,1]^2 \times \mathbb{T})$ is contained in $H_{1,1,0}$ and thus in $H_{\beta,\beta,0}$ for $0 < \beta < 1$. This follows from the representations

$$a(x_2,y,t) - a(x_1,y,t) = \int_{x_1}^{x_2} \partial_x a(\xi,y,t)\,d\xi,$$
$$a(x,y_2,t) - a(x,y_1,t) = \int_{y_1}^{y_2} \partial_y a(x,\eta,t)\,d\eta,$$
$$\Delta_2 a(x_1,x_2,y_1,y_2,t) = \int_{x_1}^{x_2}\int_{y_1}^{y_2} \partial_x\partial_y a(\xi,\eta,t)\,d\eta\,d\xi,$$

which show that in fact the continuity of the partial derivatives $\partial_x a$, $\partial_y a$, $\partial_x\partial_y a$ would suffice. Since $C^{2,0}([0,1]^2 \times \mathbb{T}) \subset H_{1,1,0}$, it follows that $C^{4,0}([0,1]^2 \times \mathbb{T})$ is a subset of $H_{2,2,0}$.

In what follows we work with functions $a(x, y, t)$ that are independent of one of the variables x and y and therefore need the following modifications. Let $0 < \alpha \leq 1$. We say that a continuous functions $a(x, t)$ on $[0, 1] \times \mathbb{T}$ is in $H_{\alpha,\infty}$ if

$$M_{\alpha,\infty}(a) := \sup_{t\in\mathbb{T}} \sup_{x_1,x_2} \frac{|a(x_2, t) - a(x_1, t)|}{|x_2 - x_1|^\alpha} < \infty,$$

and we denote by $M_{\infty,\infty}(a)$ the maximum of $|a(x, t)|$ on $[0, 1] \times \mathbb{T}$. The function $a(x, t)$ is said to be in $H_{1+\alpha,\infty}$ if it is continuously differentiable in x and $\partial_x a$ is in $H_{\alpha,\infty}$. In that case $M_{1+\alpha,\infty}(a)$ is defined as $M_{\alpha,\infty}(\partial_x a)$. Analogously, we say that a continuous functions $a(y, t)$ on $[0, 1] \times \mathbb{T}$ is in $H_{\alpha,\infty}$ if

$$M_{\alpha,\infty}(a) := \sup_{t\in\mathbb{T}} \sup_{y_1,y_2} \frac{|a(y_2, t) - a(y_1, t)|}{|y_2 - y_1|^\alpha} < \infty,$$

we denote by $M_{\infty,\infty}(a)$ the maximum of $|a(y, t)|$ on $[0, 1] \times \mathbb{T}$, and say that $a(y, t)$ belongs to $H_{1+\alpha,\infty}$ if $\partial_y a \in H_{\alpha,\infty}$, in which case $M_{1+\alpha,\infty}(a)$ is defined as $M_{\alpha,\infty}(\partial_y a)$. Finally, H_α is the set of all continuous functions $f(x)$ on $[0, 1]$ for which

$$M_\alpha(f) := \sup_{x_1,x_2} \frac{|f(x_2) - f(x_1)|}{|x_2 - x_1|^\alpha} < \infty,$$

and $H_{1+\alpha}$ is the space of all continuously differentiable functions $f(x)$ on $[0, 1]$ with $M_{1+\alpha}(f) := M_\alpha(f') < \infty$; we put $M_\infty(f) = \sup_{x\in[0,1]} |f(x)|$.

3. Counterexamples

In this section we prove Theorem 1.1 and the first statement of Theorem 1.4. We show that counterexamples can even be found within the functions $a(x, y, t)$ that are independent of one of the variables x and y.

Theorem 3.1. *There exist functions $a(x, t)$ in $C([0, 1] \times \mathbb{T})$ such that*

$$\sup_{N\geq 0} \|A_N(a)\|_\infty = \infty.$$

Proof. Assume the contrary, that is, $\sup \|A_N(a)\|_\infty < \infty$ for every function a in $C([0, 1] \times \mathbb{T})$. Let $\mathcal{S}$ denote the Banach space of all sequences $\{B_N\}_{N=0}^\infty$ of matrices $B_N \in \mathbb{C}^{(N+1)\times(N+1)}$ such that

$$\|\{B_N\}_{N=0}^\infty\| := \sup_{N\geq 0} \|B_N\|_\infty < \infty.$$

By our assumption, the map

$$T : C([0, 1] \times \mathbb{T}) \to \mathcal{S}, \quad a \mapsto \{A_N(a)\}_{N=0}^\infty$$

is a linear operator defined on all of $C([0, 1] \times \mathbb{T})$. To show that T is bounded, we employ the closed graph theorem. Thus, let $a_n, a \in C([0, 1] \times \mathbb{T})$ and suppose $a_n \to a$ in $C([0, 1] \times \mathbb{T})$ and $Ta_n \to b = \{B_N\}_{N=0}^\infty$ in $\mathcal{S}$. Then, for fixed $N \geq 0$, $\|A_N(a_n) - B_N\|_\infty \to 0$ as $n \to \infty$ and hence the jk entry of $A_N(a_n)$ converges to

the jk entry of B_N as $n \to \infty$, that is, $[A_N(a_n)]_{jk} \to [B_N]_{jk}$ for $0 \le j,k \le N$. On the other hand,

$$|\,[A_N(a_n)]_{jk} - [A_N(a)]_{jk}| = \left| (\hat{a}_n)_{j-k}\left(\frac{j}{N}\right) - \hat{a}_{j-k}\left(\frac{j}{N}\right)\right|$$
$$= \left|\int_{\mathbb{T}} \left(a_n\left(\frac{j}{N},t\right) - a\left(\frac{j}{N},t\right)\right) t^{-(j-k)}\,\frac{|dt|}{2\pi}\right| \le M_{\infty,\infty}(a_n - a) = o(1),$$

which yields the equality $[B_N]_{jk} = [A_N(a)]_{jk}$. Consequently, $Ta = b$. The closed graph theorem therefore implies that T is bounded.

We have shown that there is a constant $C < \infty$ such that

$$\|A_N(a)\|_\infty \le C M_{\infty,\infty}(a) \tag{6}$$

for all $a \in C([0,1] \times \mathbb{T})$. Fix $N \ge 2$ and for $j = 1, \dots, N-1$, denote by I_j the segment

$$I_j = \left[\frac{j}{N} - \frac{1}{2N}, \frac{j}{N} + \frac{1}{2N}\right].$$

Let a_j be the function that is identically zero on $[0,1] \setminus I_j$, increases linearly from 0 to 1 on the left half of I_j, and decreases linearly from 1 to 0 on the right half of I_j. Put

$$a(x,t) = a_1(x)t^1 + a_2(x)t^2 + \cdots + a_{N-1}(x)t^{N-1}.$$

As the spectral norm of a matrix is greater than or equal to the ℓ^2 norm of its first column and as $\hat{a}_j(x) = a_j(x)$ for $1 \le j \le N-1$, it follows that

$$\|A_N(a)\|_\infty^2 \ge \sum_{j=1}^{N-1} \left|a_j\left(\frac{j}{N}\right)\right|^2 = \sum_{j=1}^{N-1} 1^2 = N-1.$$

Since $a(x,t) = 0$ for $x \notin \cup I_j$ and $|a(x,t)| = |a_j(x)t^j| \le 1$ for $x \in I_j$, we obtain that $M_{\infty,\infty}^2(a) = 1$. Consequently, (6) gives $N - 1 \le C^2 \cdot 1$ for all $N \ge 2$, which is impossible. □

Theorem 3.2. *If $0 < \alpha < 1/2$, there exist functions $a(x,t)$ in $H_{\alpha,\infty}$ such that*

$$\sup_{N \ge 0} \|A_N(a)\|_\infty = \infty.$$

Proof. Assume that $\sup \|A_N(a)\|_\infty < \infty$ for every $a \in H_{\alpha,\infty}$. The space $H_{\alpha,\infty}$ is a Banach space under the norm $\|a\| := M_{\infty,\infty}(a) + M_{\alpha,\infty}(a)$ and hence the same argument as in the proof of Theorem 3.1 gives

$$\|A_N(a)\|_\infty \le C(M_{\infty,\infty}(a) + M_{\alpha,\infty}(a)) \tag{7}$$

for all $a \in H_{\alpha,\infty}$. Let $a(x,t)$ be exactly as the proof of Theorem 3.1. We then have $\|A_N(a)\|_\infty \ge \sqrt{N-1}$, $M_{\infty,\infty}(a) = 1$, and it is easily seen that $M_{\alpha,\infty}(a) = O(N^\alpha)$. Thus, (7) delivers $\sqrt{N-1} = O(N^\alpha)$, which is impossible for $\alpha < 1/2$. □

Since $A_N(a(y,t))$ is the transpose of $A_N(a(x,1/t))$, the above two theorems also deliver counterexamples with functions of the form $a(y,t)$.

4. Sufficient conditions

In this section we prove the second half of Theorem 1.4 and thus also Theorem 1.2. The following result is well known; see, for example, [15, Chap. 2, Sec. 4]. We cite it with a full proof for the reader's convenience.

Lemma 4.1. *If $f(x)$ is a function in $H_{1+\alpha}$ and $f(0) = f(1)$, then*

$$f(x) = \sum_{n=-\infty}^{\infty} f_n e^{2\pi i n x}$$

with

$$|f_n| \le \frac{M_\alpha(f')}{2^{2+\alpha}\pi|n|^{1+\alpha}} \quad \text{for} \quad |n| \ge 1.$$

Proof. Let $|n| \ge 1$. Then

$$\begin{aligned} f_n &= \int_0^1 f(x)e^{-2\pi i n x}dx = \int_0^1 f(x)d\frac{e^{-2\pi i n x}}{-2\pi i n} \\ &= f(x)\ \frac{e^{-2\pi i n x}}{-2\pi i n}\bigg|_0^1 + \frac{1}{2\pi i n}\int_0^1 f'(x)e^{-2\pi i n x}dx = \frac{1}{2\pi i n}\int_0^1 f'(x)e^{-2\pi i n x}dx, \end{aligned}$$

the last equality resulting from the requirement that $f(0) = f(1)$. The substitution $x \to x - 1/(2n)$ yields

$$\int_0^1 f'(x)e^{-2\pi i n x}dx = -\int_0^1 f'\left(x - \frac{1}{2n}\right)e^{-2\pi i n x}dx,$$

whence

$$f_n = \frac{1}{4\pi i n}\int_0^1 \left(f'(x) - f'\left(x - \frac{1}{2n}\right)\right)e^{-2\pi i n x}dx.$$

Taking into account that $|f'(x) - f'(x - 1/(2n))| \le M_\alpha(f')/(2n)^\alpha$, we arrive at the asserted inequality. □

We first consider functions $a(x, y, t)$ that are independent of either x or y.

Theorem 4.2. *Let $\alpha > 0$. There exists a constant $C(\alpha)$ depending only on α such that*

$$\|A_N(a)\|_\infty \le C(\alpha)(M_{\infty,\infty}(a) + M_{1+\alpha,\infty}(a))$$

for all functions $a(x, t)$ in $H_{1+\alpha,\infty}$.

Proof. We write $a = a_0 + a_1$ with

$$a_1(x, t) = (a(1, t) - a(0, t))x + a(0, t), \quad a_0(x, t) = a(x, t) - a_1(x, t).$$

Then $A_N(a) = A_N(a_0) + A_N(a_1)$. Obviously,

$$A_N(b(x)c(t)) = \left(b\left(\frac{j}{N}\right)\hat{c}_{j-k}\right)_{j,k=0}^{N} = D_N(b)T_N(c), \tag{8}$$

where $D_N(b) = \mathrm{diag}(b(j/N))_{j=0}^N$ and $T_N(c) = (\hat{c}_{j-k})_{j,k=0}^N$. Taking into account that $\|D_N(b)\|_\infty \le M_\infty(b)$ and $\|T_N(c)\|_\infty \le M_\infty(c)$, we obtain that

$$\|A_N(a_1)\|_\infty \le M_\infty(x)M_\infty(a(1,t) - a(0,t)) + M_\infty(a(0,t)) \le 3M_{\infty,\infty}(a).$$

As $a_0(0,t) = a_0(1,t)$ $(= 0)$, Lemma 4.1 gives

$$a_0(x,t) = \sum_{n=-\infty}^{\infty} a_n^0(t)e^{2\pi inx}$$

with

$$|a_n^0(t)| \le \frac{M_\alpha(\partial_x a_0(x,t))}{2^{2+\alpha}\pi|n|^{1+\alpha}} \tag{9}$$

for $|n| \ge 1$. From (8) we infer that

$$\|A_N(a_n^0(t)e^{2\pi inx})\|_\infty \le M_\infty(e^{2\pi inx})M_\infty(a_n^0(t)) = M_\infty(a_n^0).$$

Thus, by (9),

$$\begin{aligned}\|A_N(a_0)\|_\infty &\le M_\infty(a_0^0) + \sum_{|n|\ge 1} M_\infty(a_n^0)\\ &\le M_\infty(a_0^0) + \frac{1}{2^{2+\alpha}\pi}\sum_{|n|\ge 1}\frac{M_{\alpha,\infty}(\partial_x a_0(x,t))}{|n|^{1+\alpha}}.\end{aligned}$$

Since $a_0(x,t) = a(x,t) - a_1(x,t)$ and $\partial_x a_1(x,t)$ is independent of x, we get

$$M_{\alpha,\infty}(\partial_x a_0(x,t)) = M_{\alpha,\infty}(\partial_x a(x,t)) = M_{1+\alpha,\infty}(a).$$

Furthermore,

$$\begin{aligned}M_\infty(a_0^0) &= \sup_{t\in\mathbb{T}}\left|\int_0^1 a_0(x,t)dx\right| \le M_{\infty,\infty}(a_0) = M_{\infty,\infty}(a - a_1)\\ &\le M_{\infty,\infty}(a) + M_{\infty,\infty}(a_1) \le 4\,M_{\infty,\infty}(a).\end{aligned}$$

In summary,

$$\|A_N(a)\|_\infty \le 7\,M_{\infty,\infty}(a) + \left(\frac{1}{2^{2+\alpha}\pi}\sum_{|n|\ge 1}\frac{1}{|n|^{1+\alpha}}\right) M_{1+\alpha,\infty}(a),$$

which implies the assertion at once. □

Theorem 4.3. *Let $\alpha > 0$. There exists a constant $C(\alpha)$ that depends only on α such that*

$$\|A_N(a)\|_\infty \le C(\alpha)(M_{\infty,\infty}(a) + M_{1+\alpha,\infty}(a))$$

for all functions $a(y,t)$ in $H_{1+\alpha,\infty}$.

Proof. This follows from Theorem 4.2 by taking transposed matrices. □

We now turn to the case where $a(x,y,t)$ depends on all of the three variables.

Lemma 4.4. *Let $f(x,y,t)$ be a function in $H_{1+\alpha,1+\alpha,\infty}$ $(\alpha > 0)$ and suppose $f(0,y,t) = f(1,y,t)$ for all y and t. Then*

$$f(x,y,t) = \sum_{n=-\infty}^{\infty} f_n(y,t)e^{2\pi inx}$$

with

$$M_{\infty,\infty}(f_n) \le \frac{1}{2^{2+\alpha}\pi|n|^{1+\alpha}}\, M_{1+\alpha,\infty,\infty}(f), \tag{10}$$

$$M_{1+\alpha,\infty}(f_n) \le \frac{1}{2^{2+\alpha}\pi|n|^{1+\alpha}}\, M_{1+\alpha,1+\alpha,\infty}(f) \tag{11}$$

for $|n| \ge 1$.

Proof. Estimate (10) is immediate from Lemma 4.1 and the definition of the number $M_{1+\alpha,\infty,\infty}(f)$. To prove (11), we first note that

$$\partial_y f_n(y_2,t) - \partial_y f_n(y_1,t) = \int_0^1 (\partial_y f(x,y_2,t) - \partial_y f(x,y_1,t))e^{-2\pi inx}dx \tag{12}$$

and $\partial_y f(0,y,t) = \partial_y(1,y,t)$. Integrating by parts we therefore see that (12) equals

$$\frac{1}{2\pi in}\int_0^1 (\partial_x\partial_y f(x,y_2,t) - \partial_x\partial_y f(x,y_1,t))e^{-2\pi inx}dx,$$

and the substitution $x \to x - 1/(2n)$ shows that this is

$$\frac{1}{4\pi in}\int_0^1 \Delta_2\partial_x\partial_y f\left(x+\frac{1}{2n},x,y_1,y_2,t\right)e^{-2\pi inx}dx.$$

Consequently,

$$|\partial_y f_n(y_2,t) - \partial_y f_n(y_1,t)| \le \frac{1}{4\pi|n|}\int_0^1 M_{1+\alpha,1+\alpha,\infty}(f)\,\frac{1}{|2n|^\alpha}\,|y_2-y_1|^\alpha dx,$$

which gives (11). □

Theorem 4.5. *Let $\alpha > 0$. Then there exists a constant $D(\alpha) < \infty$ depending only on α such that*

$$\|A_N(a)\|_\infty \le D(\alpha)(M_{\infty,\infty,\infty}(a)+M_{1+\alpha,\infty,\infty}(a)+M_{\infty,1+\alpha,\infty}(a)+M_{1+\alpha,1+\alpha,\infty}(a))$$

for all $N \ge 0$ and all functions $a(x,y,t)$ in $H_{1+\alpha,1+\alpha,\infty}$.

Proof. We write $a = a_0 + a_1$ and accordingly $A_N(a) = A_N(a_0) + A_N(a_1)$ with

$$a_1(x,y,t) = (a(1,y,t) - a(0,y,t))x + a(0,y,t),$$
$$a_0(x,y,t) = a(x,y,t) - a_1(x,y,t).$$

For every function $c(y,t)$ we have

$$A_N(c(y,t)x) = \left(\hat{c}_{j-k}\left(\frac{k}{N}\right)\frac{j}{N}\right)_{j,k=0}^{N} = \operatorname{diag}\left(\frac{j}{N}\right)_{j=0}^{N} A_N(c(y,t))$$

and the spectral norm of the diagonal matrix is 1. Hence

$$\begin{aligned}\|A_N(a_1)\|_\infty &\le \|A_N(a(1,y,t)-a(0,y,t))\|_\infty + \|A_N(a(0,y,t))\|_\infty \\ &\le \|A_N(a(1,y,t))\|_\infty + 2\,\|A_N(a(0,y,t))\|_\infty.\end{aligned}$$

Theorem 4.3 gives

$$\begin{aligned}\|A_N(a(1,y,t))\|_\infty &\le C(\alpha)(M_{\infty,\infty}(a(1,y,t)) + M_{1+\alpha,\infty}(a(1,y,t))) \\ &\le C(\alpha)(M_{\infty,\infty,\infty}(a) + M_{\infty,1+\alpha,\infty}(a)), \\ \|A_N(a(0,y,t))\|_\infty &\le C(\alpha)(M_{\infty,\infty}(a(0,y,t)) + M_{1+\alpha,\infty}(a(0,y,t))) \\ &\le C(\alpha)(M_{\infty,\infty,\infty}(a) + M_{\infty,1+\alpha,\infty}(a)).\end{aligned}$$

Thus,

$$\|A_N(a_1)\|_\infty \le 3\,C(\alpha)(M_{\infty,\infty,\infty}(a) + M_{\infty,1+\alpha,\infty}(a)).$$

On the other hand, since $a_0(0,y,t) = a_0(1,y,t)$, Lemma 4.4 implies that

$$a_0(x,y,t) = \sum_{n=-\infty}^{\infty} a_n^0(y,t)e^{2\pi inx}$$

where the functions $a_n^0(y,t)$ satisfy estimates like (10) and (11). This time

$$\begin{aligned}A_N(a_n^0(y,t)e^{2\pi inx}) &= \left((\hat{a}_n^0)_{j-k}\left(\frac{k}{N}\right)e^{2\pi inj/N}\right)_{j,k=0}^{N} \\ &= \operatorname{diag}\left(e^{2\pi ijn/N}\right)_{j=0}^{N} A_N(a_n^0(y,t))\end{aligned}$$

and the spectral norm of the diagonal matrix is again 1. It follows from Theorem 4.3 and Lemma 4.4 that

$$\begin{aligned}&\|A_N(a_0)\|_\infty \le \|A_N(a_0^0)\|_\infty + \sum_{|n|\ge 1}\|A_N(a_n^0)\|_\infty \\ &\le \|A_N(a_0^0)\|_\infty + \sum_{|n|\ge 1}\frac{1}{2^{2+\alpha}\pi|n|^{1+\alpha}}(M_{1+\alpha,\infty,\infty}(a_0) + M_{1+\alpha,1+\alpha,\infty}(a_0)) \\ &=: \|A_N(a_0^0)\|_\infty + D_0(\alpha)(M_{1+\alpha,\infty,\infty}(a_0) + M_{1+\alpha,1+\alpha,\infty}(a_0))\end{aligned}$$

We have $M_*(a_0) = M_*(a-a_1) \le M_*(a) + M_*(a_1)$. Since $a_1(x,y,t)$ depends on x linearly, we get $M_{1+\alpha,\infty,\infty}(a_1) = M_{1+\alpha,1+\alpha,\infty}(a_1) = 0$. Thus,

$$M_{1+\alpha,\infty,\infty}(a_0) + M_{1+\alpha,1+\alpha,\infty}(a_0) \le M_{1+\alpha,\infty,\infty}(a) + M_{1+\alpha,1+\alpha,\infty}(a)$$

Finally, Theorem 4.3 shows that

$$\|A_N(a_0^0)\|_\infty \le C(\alpha)(M_{\infty,\infty}(a_0^0) + M_{1+\alpha,\infty}(a_0^0)).$$

As $a_0^0(t) = \int_0^1 a_0(x,y,t)dx$, it results that

$$\begin{aligned}M_{\infty,\infty}(a_0^0) &\le M_{\infty,\infty,\infty}(a_0) = M_{\infty,\infty,\infty}(a-a_1) \\ &\le M_{\infty,\infty,\infty}(a) + M_{\infty,\infty,\infty}(a_1) \le 4\,M_{\infty,\infty,\infty}(a_0)\end{aligned}$$

and

$$\begin{aligned} M_{1+\alpha,\infty}(a_0^0) &\le \sup_{t\in\mathbb{T}} \sup_{y_1,y_2} \int_0^1 \frac{|\partial_y a_0(x,y_2,t) - \partial_y a_0(x,y_1,t)|}{|y_2-y_1|^\alpha}\,dx \\ &\le M_{\infty,1+\alpha,\infty}(a_0) \le M_{\infty,1+\alpha,\infty}(a) + M_{\infty,1+\alpha,\infty}(a_1) \\ &= M_{\infty,1+\alpha,\infty}(a) + 3\,M_{\infty,1+\alpha,\infty}(a). \end{aligned}$$

Putting things together we arrive at the theorem with $D(\alpha) = 7\,C(\alpha) + D_0(\alpha)$. □

Clearly, Theorem 4.5 implies the second half of Theorem 1.4 and thus also Theorem 1.2.

5. Discontinuous generating functions

So far we have assumed that $a \in C([0,1]^2 \times \mathbb{T})$. Inequality (4) implies that $\sup \|A_N(a)\|_\infty < \infty$ if a is given by (1) and $\hat{a}_n$ are any bounded functions on $[0,1]^2$ such that $\sum_{n=-\infty}^{\infty} M_{\infty,\infty}(\hat{a}_n) < \infty$. Thus, sufficient smoothness in t allows us to admit arbitrary bounded coefficient $\hat{a}_n(x,y)$.

For $0 < \alpha \le 1$, we denote by $H_{1+\alpha,1+\alpha}$ the Banach space of all continuous functions $f : [0,1]^2 \to \mathbb{C}$ which have continuous partial derivatives up to the order 2 and for which

$$\|f\|_{1+\alpha} := M_{\infty,\infty}(f) + M_{\alpha,\infty}(\partial_x f) + M_{\infty,\alpha}(\partial_y f) + M_{\alpha,\alpha}(\partial_x \partial_y f) < \infty,$$

where $M_{\infty,\infty}(f)$ is the maximum modulus of $f(x,y)$ on $[0,1]^2$ and

$$M_{\alpha,\infty}(g) = \sup_{y\in[0,1]} \sup_{x_1,x_2} \frac{|g(x_2,y) - g(x_1,y)|}{|x_2-x_1|^\alpha},$$

$$M_{\infty,\alpha}(g) = \sup_{x\in[0,1]} \sup_{y_1,y_2} \frac{|g(x,y_2) - g(x,y_1)|}{|y_2-y_1|^\alpha},$$

$$M_{\alpha,\alpha}(g) = \sup_{x_1,x_2} \sup_{y_1,y_2} \frac{|\Delta_2 g(x_1,x_2,y_1,y_2)|}{|x_2-x_1|^\alpha |y_2-y_1|^\alpha}.$$

Let $L^\infty(\mathbb{T}, H_{1+\alpha,1+\alpha})$ be the set of all measurable and essentially bounded functions $a : \mathbb{T} \to H_{1+\alpha,1+\alpha}$. A check of the proofs shows that these work literally also for functions a in $L^\infty(\mathbb{T}, H_{1+\alpha,1+\alpha})$. Thus, if a is in $L^\infty(\mathbb{T}, H_{1+\alpha,1+\alpha})$ then $\sup \|A_N(a)\|_\infty < \infty$. Since $C^4([0,1]^2) \subset H_{2,2}$, we obtain in particular Theorem 1.3.

References

[1] D. Borthwick and A. Uribe, *On the pseudospectra of Berezin-Toeplitz operators,* Methods Appl. Anal. **10** (2003), 31–65.

[2] A. Böttcher and S. Grudsky, *Uniform boundedness of Toeplitz matrices with variable coefficients,* Integral Equations Operator Theory **60** (2008), 313–328.

[3] T. Ehrhardt and B. Shao, *Asymptotic behavior of variable-coefficient Toeplitz determinants,* J. Fourier Anal. Appl. **7** (2001), 71–92.

[4] D. Fasino and S. Serra Capizzano, *From Toeplitz matrix sequences to zero distribution of orthogonal polynomials,* Fast Algorithms for Structured Matrices: Theory and Applications (South Hadley, MA, 2001), 329–339, Contemp. Math., **323**, Amer. Math. Soc., Providence, RI 2003.

[5] M. Kac, W.L. Murdock, and G. Szegö, *On the eigenvalues of certain Hermitian forms,* J. Rational Mech. Anal. **2** (1953), 767–800.

[6] A.B. Kuijlaars and S. Serra Capizzano, *Asymptotic zero distribution of orthogonal polynomials with discontinuously varying recurrence coefficients,* J. Approx. Theory **113** (2001), 142–155.

[7] V. Rabinovich, S. Roch, and B. Silbermann, *Limit Operators and Their Applications in Operator Theory,* Operator Theory: Advances and Applications, **150**, Birkhäuser Verlag, Basel 2004.

[8] S. Serra Capizzano, *Generalized locally Toeplitz sequences: spectral analysis and applications to discretized partial differential equations,* Linear Algebra Appl. **366** (2003), 371–402.

[9] B. Shao, *On the singular values of generalized Toeplitz matrices,* Integral Equations Operator Theory **49** (2004), 239–254

[10] I.B. Simonenko, *Szegö-type limit theorems for generalized discrete convolution operators* (Russian), Mat. Zametki **78** (2005), 265–277.

[11] P. Tilli, *Locally Toeplitz sequences: spectral properties and applications,* Linear Algebra Appl. **278** (1998), 91–120.

[12] L.N. Trefethen and S.J. Chapman, *Wave packet pseudomodes of twisted Toeplitz matrices,* Comm. Pure Appl. Math. **57** (2004), 1233–1264.

[13] O.N. Zabroda, *Generalized Convolution Operators and Asymptotic Spectral Theory,* Dissertation, TU Chemnitz 2006.

[14] O.N. Zabroda and I.B. Simonenko, *Asymptotic invertibility and the collective asymptotic behavior of the spectrum of generalized one-dimensional discrete convolutions,* Funct. Anal. Appl. **38** (2004), 65–66.

[15] A. Zygmund, *Trigonometric series,* 2nd ed., Vols. I, II, Cambridge University Press, New York 1959.

Albrecht Böttcher
Fakultät für Mathematik
TU Chemnitz
D-09107 Chemnitz, Germany
e-mail: aboettch@mathematik.tu-chemnitz.de

Sergei Grudsky
Departamento de Matemáticas
CINVESTAV del I.P.N.
Apartado Postal 14-740
07000 México, D.F., México
e-mail: grudsky@math.cinvestav.mx

Operator Theory:
Advances and Applications, Vol. 199, 203–220

A Priori Estimates on the Structured Conditioning of Cauchy and Vandermonde Matrices

Enrico Bozzo and Dario Fasino

Abstract. We analyze the componentwise and normwise sensitivity of inverses of Cauchy, Vandermonde, and Cauchy-Vandermonde matrices, with respect to relative componentwise perturbations in the nodes defining these matrices. We obtain *a priori*, easily computable upper bounds for these condition numbers. In particular, we improve known estimates for Vandermonde matrices with generic real nodes; we consider in detail Vandermonde matrices with nonnegative or symmetric nodes; and we extend the analysis to the class of complex Cauchy-Vandermonde matrices.

Mathematics Subject Classification (2000). Primary 15A12; Secondary 15A57, 65F35.

Keywords. Condition number, displacement structure, Cauchy matrix, Vandermonde matrix.

1. Introduction

A structured matrix is, in some sense, a matrix whose entries depend on a small set of parameters. Clearly, this dependence is somehow preserved by matrix inversion. The main question addressed here is: How sensitive is the inverse of a structured matrix to perturbations in its parameters?

The answer to the above question is what we call *structured conditioning*. This paper focuses on the structured conditioning of Cauchy, Vandermonde and Cauchy-Vandermonde matrices, which have been one of the main research topics of Georg Heinig, see, e.g., [9, 12, 13]. The main reasons motivating this study are the analysis of the influence of data errors in the solution of linear systems with structured matrices, and the assessment of stability properties of fast algorithms for solving such linear systems. These algorithms act directly on the set of parameters defining a structured matrix rather than on its entries, see, e.g., [3, 4, 5, 9, 10, 12, 15], hence

their stability characteristics, which are sometimes surprising, should be examined in the light of suitable refinements of classical condition numbers, taking in due consideration both more accurate measures of perturbations and the structure of the problem considered, as observed, e.g., in [1, 10, 11, 14, 15, 22]. Indeed, suppose that the solution computed by such an algorithm in finite precision is the exact solution of a similarly structured system defined by slightly perturbed parameters. To assess the quality of the computed solution with no further assumptions on the right-hand side, we should make use of some measure of the sensitivity of matrix inversion with respect to perturbation in its parameters. Theoretical bases of this argument can be found in [14], where such a *backward structured error analysis* in introduced rather formally for structured matrices whose dependence on the parameters is linear, and in [1, 15, 22], where a detailed analysis has been carried out for Vandermonde systems, mainly motivated by the stability analysis of the Björk-Pereyra algorithm. Other results in the same streamline can be found in [7, 25], concerning possibly rectangular Cauchy, Vandermonde, Toeplitz and Hankel matrices.

By the way, the results in the above-mentioned papers are mainly aimed at characterizing the structured conditioning in terms of exact (possibly generalized) inverse matrices and solutions of suitable linear systems, hence they are *a posteriori* results, and the resulting expressions may be hard to compute.

The goal of this paper is to obtain *a priori* bounds for the structured conditioning of Cauchy, Vandermonde and other related matrices, that are easily computable right from their parameters, with no involvement of exact (generalized) inverse matrices. Besides to their pervasive occurrence in computations with polynomial and rational functions, Cauchy and Vandermonde matrices play an important role in deriving structural and computational properties of many relevant matrix classes with displacement structure, see, e.g., [9, 13, 17, 18, 19]. For example, they occur as fundamental blocks (together with trigonometric transforms) in decomposition formulas for Toeplitz, Hankel, and related matrices. In this paper, we will pay particular attention to the structured conditioning of the individual columns of their inverses, as they have a special relevance in polynomial and rational interpolation problems. Indeed, let $x_1 \dots x_n$ be pairwise distinct points in the complex plane, considered as parameters, and let $\phi_1(x) \dots \phi_n(x)$ be a fixed set of functions such that the collocation matrix $X \equiv (\phi_j(x_i))$ is nonsingular. In fact, Vandermonde, Cauchy, and Cauchy-Vandermonde matrices arise as collocation matrices when the functions $\phi_j(x)$ are monomials or particular rational functions. Let $X^{-1} \equiv (v_{i,j})$. Then, the functions $\ell_k(x) = \sum_j v_{j,k}\phi_j(x)$ for $1 \le k \le n$ are the Lagrange functions for the interpolation problem defined by the points x_i and the functions $\phi_j(x)$. Hence, a further reason for investigating the structured conditioning of X^{-1} is to give a measure of the sensitivity of the functions $\ell_k(x)$ with respect to perturbations in the interpolation points x_i.

After giving in Section 2 a quick look at Cauchy matrices, we consider Vandermonde matrices in Section 3. There, we will improve the result in [11, Thm. 1] on the mixed structured conditioning of Vandermonde matrices (see Corollary 2),

and consider in detail Vandermonde matrices with nonnegative nodes and symmetric nodes. In Section 4 we will extend our analysis to Cauchy-Vandermonde matrices with complex nodes.

1.1. Main definitions and notations

We borrow from [10, 11] the following definitions and notations. Let p and q be positive integers, and let F be a (densely defined) continuous function $F : \mathbb{C}^p \mapsto \mathbb{C}^q$. One usually defines the *normwise condition number* of F in a given point $x \in \mathbb{C}^p$ as

$$\kappa(F, x) = \lim_{\varepsilon \to 0} \sup_{\|x - \tilde{x}\| \le \varepsilon} \frac{\|F(x) - F(\tilde{x})\|}{\|F(x)\|} \frac{\|x\|}{\|x - \tilde{x}\|}.$$

In the limit as the perturbation size tends to zero, this number gives the worst possible magnification of the quantity $\|x - \tilde{x}\|/\|x\|$ in the computation of F. Here and in what follows, $\|\cdot\| = \|\cdot\|_\infty$, unless otherwise noted.

The number $\|x - \tilde{x}\|/\|x\|$ is the *relative normwise distance* between x and $\tilde{x}$. In what follows, we also consider the *relative componentwise distance* between two points $x, \tilde{x} \in \mathbb{C}^p$, defined as

$$\delta(x, \tilde{x}) = \min\{\varepsilon : |x_i - \tilde{x}_i| \le \varepsilon |x_i|, 1 \le i \le p\}. \tag{1}$$

The above definition is of interest in numerical analysis, because it is the most appropriate way to measure errors induced by the finite precision representation of machine numbers. Observe that if an entry of the vector x is zero, the corresponding entry in $\tilde{x}$ must be zero for $\delta(x, \tilde{x})$ be definite: Componentwise relative perturbations do not affect null entries.

Accordingly, we consider the *componentwise condition number* of F in $x \neq 0$ as

$$c(F, x) = \lim_{\varepsilon \to 0} \sup_{\delta(x, \tilde{x}) \le \varepsilon} \frac{\delta(F(x), F(\tilde{x}))}{\delta(x, \tilde{x})}. \tag{2}$$

Throughout this paper, we consider matrix and vector moduli and inequalities as applied componentwise. Hence, we observe that $c(F, x)$ is characterized by the following inequality:

$$|F(x) - F(\tilde{x})| \le |F(x)| c(F, x) \delta(x, \tilde{x}) + o(\delta(x, \tilde{x})). \tag{3}$$

In some cases, it may be also of interest to consider the *mixed condition number*

$$m(F, x) = \lim_{\varepsilon \to 0} \sup_{\delta(x, \tilde{x}) \le \varepsilon} \frac{\|F(x) - F(\tilde{x})\|}{\|F(x)\|} \frac{1}{\delta(x, \tilde{x})}. \tag{4}$$

Since $\|x - \tilde{x}\|/\|x\| \le \delta(x, \tilde{x})$, we have both $m(F, x) \le c(F, x)$ and $m(F, x) \le \kappa(F, x)$, but in general $c(F, x)$ and $\kappa(F, x)$ are unrelated. Moreover, the equivalent definition

$$\delta(x, \tilde{x}) = \sup_{D \text{ diagonal}} \frac{\|D(x - \tilde{x})\|}{\|Dx\|}$$

implies that $m(F, x)$ results by minimizing $\kappa(F, x)$ with respect to all argument normalizations, while $c(F, x) = \sup_D m(DF, x)$ is a "worst case" measure of the

sensitivity of F in presence of diagonal scalings in both the parameters and the function values.

If the map F is differentiable, the condition numbers introduced above can be related to its differential as follows: For any n-vector $a = (a_1, \ldots, a_n)^T$, let D_a denote the $n \times n$ diagonal matrix whose ith diagonal entry is a_i. Then, if F' denotes the differential of F, we have

$$\kappa(F, x) = \|F'(x)\|\|x\|/\|F(x)\| \tag{5}$$

$$c(F, x) = \|D^{-1}_{F(x)} F'(x) D_x\| \tag{6}$$

$$m(F, x) = \|F'(x) D_x\|/\|F(x)\|. \tag{7}$$

Some further notations are used throughout this paper: Let e_i be the ith standard basis vector, whose order will be made clear from the context, and $\mathbf{1} = (1, \ldots, 1)^T$. Let $\mathrm{Vec} : \mathbb{C}^{n\times n} \mapsto \mathbb{C}^{n^2}$ be the operator such that $\mathrm{Vec}(X)$ is the n^2-order vector obtained by stacking downward the columns of X, into one long column. Moreover, for $x = (x_1, \ldots, x_p)^T \in \mathbb{C}^p$ and $y = (y_1, \ldots, y_q)^T \in \mathbb{C}^q$, let

$$\Delta(x) = \max_{i\neq j} \frac{|x_i|}{|x_i - x_j|}, \qquad \Delta(x, y) = \max_{i,j} \frac{|x_i|}{|x_i - y_j|}. \tag{8}$$

In the above equations, fractions with vanishing denominators assume the value $+\infty$, whatever the numerators are. In particular, $\Delta(x) < +\infty$ if and only if the points $x_1 \ldots x_n$ are pairwise distinct, and similarly for $\Delta(x, y)$.

1.2. Basic displacement structured matrices

Cauchy, Vandermonde and Cauchy-Vandermonde matrices are among the best known matrices with a *displacement structure*. Such kind of structure is defined in terms of a *displacement operator*

$$\mathcal{L}_{M,N}(X) = MX - XN,$$

where M, N are two fixed $n \times n$ matrices with disjoint spectra, so that the operator $\mathcal{L}_{M,N}$ is invertible. Throughout this paper, we will deal with matrix spaces having the following form:

$$\mathcal{D}_{M,N} = \{X : \mathrm{rank}(\mathcal{L}_{M,N}(X)) = 1\}.$$

A basic fact that will be used here to reduce the analysis of Vandermonde and Cauchy-Vandemonde matrices to the simpler case of Cauchy matrices is this: If $N = SDS^{-1}$ and $X \in \mathcal{D}_{M,N}$, then $XS \in \mathcal{D}_{M,D}$.

Regarding the connection between displacement structure and conditioning, we recall that in the paper [20] an exponentially growing lower bound is derived for the spectral conditioning of matrices X such that $AX + XA^T = -BB^T$, where B has low rank and all eigenvalues of A have negative real part. By Lyapunov theorem, any such matrix X is symmetric and positive definite; remarkably, the displacement structure induced by the operator $\mathcal{L}_{A,-A^T}$ forces X to be very badly conditioned.

2. Cauchy matrices

Let $x = (x_1, \ldots, x_n)^T \in \mathbb{C}^n$ and $y = (y_1, \ldots, y_n)^T \in \mathbb{C}^n$ have pairwise distinct entries, with $x_i \neq y_j$ for $1 \leq i, j \leq n$. The *Cauchy matrix* associated with x and y is defined by $C_{x,y} \equiv (1/(x_i - y_j))$. Since the displacement operator $\mathcal{L}_{D_x, D_y}$ is nonsingular, $C_{x,y}$ is the unique solution of the displacement equation $D_x C_{x,y} - C_{x,y} D_y = \mathbf{1}\mathbf{1}^T$. The following explicit formula for the entries of the inverse of $C_{x,y}$ is well known, see, e.g., [10]: For $C_{x,y}^{-1} \equiv (v_{i,j})$ we have

$$v_{i,j} = \frac{\prod_l (y_i - x_l)}{\prod_{l \neq i} (y_i - y_l)} \frac{1}{y_i - x_j} \frac{\prod_l (x_j - y_l)}{\prod_{l \neq j} (x_j - x_l)}. \tag{9}$$

Remark that $v_{i,j} \neq 0$. In [10] the above expression is differentiated with respect to x_i and y_i in order to bound the componentwise conditioning of the inversion of $C_{x,y}$, via (6). In the next result we show a simple bound for the componentwise conditioning of the columns of $C_{x,y}^{-1}$, when only the x node vector is subject to perturbations. This result plays a fundamental role in the subsequent sections. Moreover, as mentioned in the Introduction, it is useful to estimate the sensitivity of a set of rational Lagrange functions with respect to the interpolation nodes. It is apparent from the definition of $C_{x,y}$ that we obtain the corresponding result for y by simply considering the transpose matrix.

Theorem 1. *Let $1 \leq i \leq n$ and $y \in \mathbb{C}^n$ be fixed, with pairwise distinct entries. Let $F_i : \mathbb{C}^n \mapsto \mathbb{C}^n$ be defined as $F_i(x) = C_{x,y}^{-1} e_i$. Then,*

$$c(F_i, x) \leq 2(n-1)\Delta(x) + (2n-1)\Delta(x, y).$$

Proof. From (6) we have

$$c(F_i, x) = \max_{1 \leq j \leq n} \sum_k \left| \frac{x_k}{v_{j,i}} \frac{\partial v_{j,i}}{\partial x_k} \right| = \max_{1 \leq j \leq n} \sum_k \left| x_k \frac{\partial}{\partial x_k} \log |v_{j,i}| \right|, \tag{10}$$

where $v_{j,i}$ is given in (9). For $k \neq i$ we have:

$$\left| x_k \frac{\partial \log |v_{j,i}|}{\partial x_k} \right| = \left| -\frac{x_k}{x_k - x_i} \frac{x_k}{x_k - y_j} \right| \leq \Delta(x) + \Delta(x, y).$$

On the other hand,

$$\left| x_i \frac{\partial \log |v_{j,i}|}{\partial x_i} \right| = \left| \sum_{l \neq i} \frac{x_i}{x_i - x_l} - \sum_l \frac{x_i}{x_i - y_l} \right| \leq (n-1)\Delta(x) + n\Delta(x, y).$$

Plugging the latter inequalities into (10) we arrive at the claim. □

For notational simplicity, in the sequel we use the shorthand

$$\Delta(n, x, y) = 2(n-1)\Delta(x) + (2n-1)\Delta(x, y). \tag{11}$$

In the next corollary we consider the structured conditioning of the matrices in $\mathcal{D}_{D_x, D_y}$. Observe that any matrix in this set can be expressed as $D_1 C_{x,y} D_2$ for some diagonal matrices D_1 and D_2.

Corollary 1. *Let $F : \mathbb{C}^n \mapsto \mathbb{C}^{n^2}$ be defined as $F(x) = \mathrm{Vec}(D_1 C_{x,y}^{-1} D_2)$, where D_1, D_2 are arbitrary diagonal matrices and y is as in the preceding theorem. Then,*

$$c(F, x) \leq \Delta(n, x, y),$$

where $\Delta(n, x, y)$ is defined in (11).

Proof. Scaling the entries of $C_{x,y}^{-1}$ by constant factors does not affect their conditioning, as it should be clear from the definitions (1) and (2). Hence we can assume $D_c = D_d = I$. In this case, we observe that

$$c(F, x) = \max_{1 \leq i \leq n} c(F_i, x) \leq \Delta(n, x, y),$$

where F_i is as in the preceding theorem. □

If F is as in the preceding corollary and $D_c = D_d = I$, then $F'(x)$ is the map $z \mapsto \mathrm{Vec}(C_{x,y}^{-1} E(z) C_{x,y}^{-1})$, where $E(z) \equiv (-z_i/(x_i - y_j)^2)$. Hence, in view of (5), the normwise conditioning $\kappa(F, x)$ is essentially driven by the (unstructured) conditioning of $C_{x,y}$, that is, $\|C_{x,y}\| \|C_{x,y}^{-1}\|$. Quite few estimates are currently available for the latter, and only for special vectors x and y. Here we only mention the paper [23], dealing with the spectral conditioning of the Cauchy-Toeplitz matrix $C_T \equiv (1/(a+i-j))$, and [8], for Cauchy matrices $C_{-y,y}$ with positive vector y (note that these two examples fall into the generic case considered in [20] and mentioned in Subsection 1.2). Generally, all these matrices are very ill conditioned, and this fact should be contrasted with the slowly-growing estimate for the componentwise conditioning shown in the above corollary; for example, if y has positive entries we have (see [8])

$$\|C_{-y,y}\|_2 \|C_{-y,y}^{-1}\|_2 > \left(\frac{y_n + y_1}{y_n - y_1} \right)^{2n-2}.$$

Note that also the exponentially ill-conditioned Hilbert matrix $C_H \equiv (1/(i+j-1))$ is a special Cauchy matrix. For that matrix, an $O(n^2)$ bound for its structured conditioning is derived in [10], while its spectral conditioning grows roughly like 34^n [2].

3. Vandermonde matrices

Given a vector $x = (x_1, \ldots, x_n)^T \in \mathbb{C}^n$ with pairwise distinct entries, for any fixed $0 \leq \theta < 2\pi$ the Vandermonde matrix $V_x \equiv (x_i^{j-1})$ fulfills the displacement equation

$$\mathcal{L}_{D_x, P_\theta}(V_x) = D_x V_x - V_x P_\theta = \begin{pmatrix} x_1^n - \mathrm{e}^{\mathrm{i}n\theta} \\ \vdots \\ x_n^n - \mathrm{e}^{\mathrm{i}n\theta} \end{pmatrix} e_n^T,$$

where i is the imaginary unit, and

$$P_\theta = \begin{pmatrix} 0 & \cdots & 0 & \mathrm{e}^{\mathrm{i}n\theta} \\ 1 & \ddots & 0 & 0 \\ & \ddots & \ddots & \vdots \\ O & & 1 & 0 \end{pmatrix}.$$

Recall that the spectral decomposition of P_θ is explicitly computable,

$$P_\theta = \mathrm{e}^{\mathrm{i}\theta} D(\theta) \Phi D_\xi \Phi^{-1} D(\theta)^{-1}, \tag{12}$$

where $\xi = (1, \mathrm{e}^{\mathrm{i}2\pi/n}, \ldots, \mathrm{e}^{\mathrm{i}(n-1)2\pi/n})^T$, $\Phi = \sqrt{1/n}\, V_\xi$ is the unitary Fourier matrix of order n, and $D(\theta) = \mathrm{Diag}(1, \mathrm{e}^{-\mathrm{i}\theta}, \ldots, \mathrm{e}^{-\mathrm{i}(n-1)\theta})$.

Only for notational simplicity, in what follows we suppose that the operator $\mathcal{L}_{D_x,P_0}$ is nonsingular, that is, $\Delta(x,\xi) < +\infty$. This hypothesis is of no restriction, and can always be fulfilled by a suitable rotation of the complex plane. Indeed, let $\hat{x} = \mathrm{e}^{-\mathrm{i}\theta} x$; we have $V_x D(\theta) = V_{\hat{x}}$ and, from (12),

$$\begin{aligned} \mathcal{L}_{D_x,P_\theta}(V_x) &= [D_x V_x D(\theta) - \mathrm{e}^{\mathrm{i}\theta} V_x D(\theta) P_0] D(\theta)^{-1} \\ &= \mathrm{e}^{\mathrm{i}\theta} [D_{\hat{x}} V_{\hat{x}} - V_{\hat{x}} P_0] D(\theta)^{-1} \\ &= \mathrm{e}^{\mathrm{i}n\theta} \mathcal{L}_{D_{\hat{x}},P_0}(V_{\hat{x}}). \end{aligned}$$

The last equation uses the fact that $e_n^T D(\theta)^{-1} = \mathrm{e}^{\mathrm{i}(n-1)\theta} e_n^T$. Hence, the discussion of the Vandermonde matrix with node vector x, for a chosen matrix P_θ, can be carried out equivalently by considering the Vandermonde matrix defined by $\hat{x}$ and the displacement equation with the matrix P_0. Since the entries of $V_{\hat{x}}$ and $V_{\hat{x}}^{-1}$ have the same modulus of the entries of V_x and V_x^{-1}, respectively, the rotation $x \mapsto \hat{x}$ leaves unaltered the conditioning properties we are investigating.

In contrast to what happens with Cauchy matrices, inverse Vandermonde matrices can have zero entries, hence the componentwise conditioning of Vandermonde matrices cannot be bounded in general. For this reason, we will consider the mixed conditioning, in the generic case, as in [11, 15, 22]. Furthermore, we will obtain estimates for the componentwise conditioning for particular configurations of the nodes, namely, nonnegative or symmetric nodes.

We will use the following lemma to bridge Cauchy and Vandermode matrices. The statement can be obtained as a consequence of Proposition 3.2 in [9]. For convenience, we provide here a short and self-contained proof.

Lemma 1. *Let $x = (x_1, \ldots, x_n)^T \in \mathbb{C}^n$ and $a = (x_1^n - 1, \ldots, x_n^n - 1)^T$. If D_a is nonsingular, the matrix V_x can be factorized as follows:*

$$V_x = \frac{1}{\sqrt{n}} D_a C_{x,\xi} D_\xi^{-1} \Phi^{-1}.$$

Proof. From (12) we have

$$D_x V_x \Phi - V_x \Phi D_\xi = (D_x V_x - V_x P_0)\Phi = a e_n^T \Phi = \frac{1}{\sqrt{n}} a \mathbf{1}^T D_\xi^{-1}.$$

Since diagonal matrices commute, we have

$$\begin{aligned}\mathcal{L}_{D_x,D_\xi}(D_a^{-1}V_x\Phi D_\xi) &= D_a^{-1}[D_xV_x\Phi - V_x\Phi D_\xi]D_\xi \\ &= \sqrt{1/n}\, D_a^{-1}a\mathbf{1}^TD_\xi^{-1}D_\xi \\ &= \sqrt{1/n}\,\mathbf{1}\mathbf{1}^T.\end{aligned}$$

By hypothesis, the nodes x_i are not roots of unit, hence the operator $\mathcal{L}_{D_x,D_\xi}$ is invertible. Thus $D_a^{-1}V_x\Phi D_\xi = \sqrt{1/n}\,C_{x,\xi}$, whence we obtain the thesis. □

Observe that, if the hypothesis on D_a is false, then the matrix $C_{x,\xi}$ is not even defined. Indeed, the same hypothesis can be restated as $\Delta(x,\xi) < +\infty$.

Lemma 2. *Let $F : \mathbb{C}^n \mapsto \mathbb{C}^n$ be the map $F(v) = \Phi v$. Then, $m(F,v) = \sqrt{n}$.*

Proof. See [6, Corollary 1]. □

Theorem 2. *Let $1 \le i \le n$ be fixed, and let $F_i : \mathbb{C}^n \mapsto \mathbb{C}^n$ be the function defined as $F_i(x) = V_x^{-1}e_i$. Furthermore, let $\xi = (1, \mathrm{e}^{\mathrm{i}2\pi/n}, \ldots, \mathrm{e}^{\mathrm{i}(n-1)2\pi/n})^T$. Then, for any $0 \le \theta < 2\pi$ we have*

$$m(F_i,x) \le \sqrt{n}\left(\Delta(n,\mathrm{e}^{-\mathrm{i}\theta}x,\xi) + \left|\frac{nx_i^n}{x_i^n - \mathrm{e}^{\mathrm{i}n\theta}}\right|\right),$$

where $\Delta(n,\mathrm{e}^{-\mathrm{i}\theta}x,\xi)$ can be obtained from (11).

Proof. In the light of the argument outlined at the beginning of this section, we can restrict the proof to the case $\theta = 0$, since the general case follows by considering the matrix $V_{\hat{x}}$, with $\hat{x} = \mathrm{e}^{-\mathrm{i}\theta}x$. From Lemma 1 we obtain $V_x^{-1} = \sqrt{n}\Phi D_\xi C_{x,\xi}^{-1}D_a^{-1}$, where $a = (x_1^n - 1, \ldots, x_n^n - 1)^T$. Hence, for $1 \le i \le n$,

$$F_i(x) = \frac{\sqrt{n}}{x_i^n - 1}\Phi D_\xi C_{x,\xi}^{-1}e_i.$$

Consider the decomposition $F_i(x) = G(H_i(x))$, where $G(x) = \Phi x$ and $H_i(x) = \sqrt{n}(x_i^n - 1)^{-1}D_\xi C_{x,\xi}^{-1}e_i$. It can be shown that $m(F_i,x) \le m(G,H_i(x))c(H_i,x)$, see [11, p. 692]. Then, from Lemma 2 we obtain $m(F_i,x) \le \sqrt{n}c(H_i,x)$.

Furthermore, for x and $\tilde{x}$ such that $\delta(x,\tilde{x}) = \epsilon$, we have from Theorem 1

$$\begin{aligned}|H_i(x) - H_i(\tilde{x})| &\le \sqrt{n}|x_i^n - 1|^{-1}|C_{x,\xi}^{-1}e_i - C_{\tilde{x},\xi}^{-1}e_i| \\ &\quad + \sqrt{n}|(x_i^n - 1)^{-1} - (\tilde{x}_i^n - 1)^{-1}||C_{\tilde{x},\xi}^{-1}e_i| \\ &\le \epsilon\Delta(n,x,\xi)|H_i(x)| + \epsilon\left|\frac{nx_i^n}{x_i^n - 1}\right||H_i(x)| + o(\epsilon).\end{aligned}$$

From the equivalence of (3) and (2) we have

$$c(H_i,x) \le \Delta(n,x,\xi) + \left|\frac{nx_i^n}{x_i^n - 1}\right|,$$

completing the proof. □

Note that the factor $\sqrt{n}$ in the right-hand side of the thesis of the preceding theorem can be dropped off, if the 2-norm is used instead of the ∞-norm in the definition of $m(F_i, x)$. Indeed, one proves easily that $m_2(G, x) = \|G'(x)D_x\|_2/\|G(x)\|_2 = 1$, and we have the attainable upper bound $m_2(F_i, x) \leq c(H_i, x)$, in the notations of the preceding proof. Hence, in some sense, the Euclidean norm is more appropriate than the ∞-norm to analyze the structured conditioning of V_x.

In the case of real nodes, Gohberg and Koltracht proved in [11] the upper bound

$$m(F_i, x) \leq n^2 \max(n\Delta(x), n + \Delta(x)).$$

We improve this bound in the following corollary.

Corollary 2. *In the notations of Theorem 2, if $x \in \mathbb{R}^n$ we have*

$$m(F_i, x) < \sqrt{n}(2(n-1)\Delta(x) + 2n^2).$$

Proof. Let $\theta = \pi/(2n)$. We have:

$$\Delta(x, \mathrm{e}^{\mathrm{i}\theta}\xi) \leq \frac{1}{\sin(\theta)} < \frac{\pi}{2\theta} = n.$$

Hence from (11)

$$\Delta(n, \mathrm{e}^{-\mathrm{i}\theta}x, \xi) = \Delta(n, x, \mathrm{e}^{\mathrm{i}\theta}\xi) < 2(n-1)\Delta(x) + (2n-1)n.$$

Moreover,

$$\left|\frac{nx_i^n}{x_i^n - \mathrm{e}^{\mathrm{i}n\theta}}\right| = n\left|\frac{x_i^n}{x_i^n - \mathrm{i}}\right| < n,$$

and the claim follows from Theorem 2. □

We consider the complete inverse of V_x in the next corollary, which follows from an argument analogous to that used in Corollary 1.

Corollary 3. *Let $F : \mathbb{C}^n \mapsto \mathbb{C}^{n^2}$ be defined as $F(x) = \mathrm{Vec}(V_x^{-1}D_d)$, where D_d is an arbitrary diagonal matrix whose entries do not depend on x. Then, for any $0 \leq \theta < 2\pi$,*

$$m(F, x) \leq \sqrt{n}\left(\Delta(n, \mathrm{e}^{-\mathrm{i}\theta}x, \xi) + n \max_{1\leq i\leq n}\left|\frac{x_i^n}{x_i^n - \mathrm{e}^{\mathrm{i}n\theta}}\right|\right),$$

where $\Delta(n, \mathrm{e}^{-\mathrm{i}\theta}x, \xi)$ can be obtained from (11).

Proof. Since a constant scaling of the columns of V_x^{-1} does not affect their conditioning, we suppose $D_d = I$. Let F_i be the map introduced in Theorem 2 for $1 \leq i \leq n$. Then, $F(x) = (F_1(x), \ldots, F_n(x))^T$. For arbitrary $x, \tilde{x}$ we have

$$\frac{\|F(x) - F(\tilde{x})\|}{\|F(x)\|} = \frac{\max_i \|F_i(x) - F_i(x)\|}{\max_i \|F_i(x)\|} \leq \max_{1\leq i\leq n} \frac{\|F_i(x) - F_i(x)\|}{\|F_i(x)\|}.$$

Hence, from the definition (4) we obtain $m(F, x) \leq \max_i m(F_i, x)$, and the claim follows from Theorem 2. □

3.1. Vandermonde matrices with nonnegative nodes

If $x_k > 0$ for $1 \le k \le n$ then V_x^{-1} has no zero entries, and we can obtain precise upper bounds for its structured componentwise conditioning. Indeed, let $V_x^{-1} \equiv (v_{i,j})$. Then, it is well known that

$$v_{j,i} = (-1)^{n-j}\sigma_{n-j}^{(n-1)}(x_1, \ldots, x_{i-1}, x_{i+1}, \ldots, x_n)\prod_{l\neq i}(x_i - x_l)^{-1},$$

see, e.g., [11], where $\sigma_i^{(n)}(a_1, \ldots, a_n)$ is the ith *elementary symmetric function* on n variables,

$$\sigma_i^{(n)}(a_1, \ldots, a_n) = \sum_{1\le j_1<\cdots<j_i\le n} a_{j_1}a_{j_2}\cdots a_{j_i}. \tag{13}$$

By definition (13), for any $1 \le k \le n$ we have

$$\begin{aligned}\sigma_i^{(n)}(a_1, \ldots, a_n) &= a_k\sigma_{i-1}^{(n-1)}(a_1, \ldots, a_{k-1}, a_{k+1}, \ldots, a_n)\\ &\quad + \sigma_i^{(n-1)}(a_1, \ldots, a_{k-1}, a_{k+1}, \ldots, a_n).\end{aligned}$$

Recall that $\sigma_0^{(n)} \equiv 1$ and $\sigma_i^{(n)} \equiv 0$ for $i > n$ or $i < 0$. Furthermore, we have

$$\sum_{k=1}^{n} a_k\sigma_i^{(n-1)}(a_1, \ldots, a_{k-1}, a_{k+1}, \ldots, a_n) = (i+1)\sigma_{i+1}^{(n)}(a_1, \ldots, a_n). \tag{14}$$

Actually, we can allow one entry in the vector x, say x_1, to be zero. In this case, the first row of V_x^{-1} is parallel to e_1^T, but no other zeros are introduced in V_x^{-1}; moreover, owing to the definition (1), x_1 is untouched by relative perturbations, hence the zero entries in V_x^{-1} don't vary. In the light of the preceding facts, we obtain the following bound for the componentwise conditioning of the columns of V_x^{-1} with nodes in $\mathbb{R}_+ = \{x \ge 0\}$:

Theorem 3. *Let $F_i : \mathbb{R}_+^n \mapsto \mathbb{C}^n$ be the function defined as $F_i(x) = V_x^{-1}e_i$, for any fixed $1 \le i \le n$. Then*

$$c(F_i, x) \le (n-1)(2\Delta(x) + 1).$$

Proof. From (6) we have

$$c(F_i, x) = \max_{1\le j\le n}\sum_k \left|\frac{x_k}{v_{j,i}}\frac{\partial v_{j,i}}{\partial x_k}\right|.$$

For $k \neq i$ we have:

$$\frac{x_k}{v_{j,i}}\frac{\partial v_{j,i}}{\partial x_k} = \frac{x_k\sigma_{n-j-1}^{(n-2)}(x_1, \ldots, x_{k-1}, x_{k+1}, \ldots, x_{i-1}, x_{i+1}, \ldots, x_n)}{\sigma_{n-j}^{(n-1)}(x_1, \ldots, x_{i-1}, x_{i+1}, \ldots, x_n)} - \frac{x_k}{x_i - x_k}.$$

Moreover,

$$\frac{x_i}{v_{j,i}}\frac{\partial v_{j,i}}{\partial x_i} = x_i\frac{\partial}{\partial x_i}\log|v_{j,i}| = -\sum_{l\neq i}\frac{x_i}{x_i - x_l}.$$

Owing to the positivity of $x_1 \dots x_n$, from (14) we obtain

$$\sum_{k \neq i} \left| \frac{x_k \sigma_{n-j-1}^{(n-1)}(x_1, \dots, x_{k-1}, x_{k+1}, \dots, x_{i-1}, x_{i+1}, \dots, x_n)}{\sigma_{n-j}^{(n-1)}(x_1, \dots, x_{i-1}, x_{i+1}, \dots, x_n)} \right| = n - j.$$

Hence, we obtain

$$c(F_i, x) \leq \max_{1 \leq j \leq n} |n - j| + \sum_{k \neq i} \left| \frac{x_k}{x_i - x_k} \right| + \left| \sum_{l \neq i} \frac{x_i}{x_i - x_l} \right|,$$

and the claim follows from the triangle inequality and the definitions (8). □

The next corollary follows by an argument analogous to the ones exploited in the proof of Corollary 1 and 3, hence we omit it.

Corollary 4. *Let $F : \mathbb{R}_+^n \mapsto \mathbb{C}^{n^2}$ be defined as $F(x) = \mathrm{Vec}(D_c V_x^{-1} D_d)$, where D_c, D_d are arbitrary diagonal matrices whose entries do not depend on x. Then, $c(F, x) \leq (n-1)(2\Delta(x) + 1)$.*

The preceding results should be contrasted with well-known lower bounds on the conditioning of Vandermonde matrices with real or positive nodes [2, 16, 21, 24], which are exponentially growing functions in the order n. Moreover, we observe that the results in this subsections are trivially extended to the case where the points $x_1 \dots x_n$ belong to a ray in the complex plane, $x_i = |x_i|\omega$, where $\omega = \mathrm{e}^{\mathrm{i}\theta}$, as it is apparent from the factorization $V_x = V_{\omega^{-1}x} \mathrm{Diag}(1, \omega \dots \omega^{n-1})$.

3.2. Vandermonde matrices with symmetric nodes

When the nodes are restricted to be real, in many circumstances they are also symmetrically located with respect to zero. Indeed, symmetric configurations arise naturally when the nodes are Fekete points or zeros of special polynomial sequences (e.g., orthogonal polynomials from symmetric weights), or when one attempts to minimize the (classical) conditioning of Vandermonde matrices, see [16].

For the sake of simplicity, we consider the Vandermonde matrix $V_{(x,-x)}$, of order $2n$, whose nodes are $x_1, \dots, x_n, -x_1, \dots, -x_n$ with $x_i > 0$ (however, consider that the structured conditioning is invariant under permutation of the nodes). Introducing the vector $\hat{x} = (x_1^2, \dots, x_n^2)^T$, we have

$$V_{(x,-x)} = \begin{pmatrix} V_x & D_x^n V_n \\ V_{-x} & D_{-x}^n V_{-n} \end{pmatrix} = \begin{pmatrix} I & I \\ I & -I \end{pmatrix} \begin{pmatrix} V_{\hat{x}} & O \\ O & D_x V_{\hat{x}} \end{pmatrix} \Pi^T,$$

where Π is the perfect shuffle permutation matrix. We obtain

$$\begin{aligned} V_{(x,-x)}^{-1} &= \frac{1}{2} \Pi \begin{pmatrix} V_{\hat{x}}^{-1} & O \\ O & (D_x V_{\hat{x}})^{-1} \end{pmatrix} \begin{pmatrix} I & I \\ I & -I \end{pmatrix} \\ &= \frac{1}{2} \Pi \begin{pmatrix} V_{\hat{x}}^{-1} & (D_x V_{\hat{x}})^{-1} \\ V_{\hat{x}}^{-1} & -(D_x V_{\hat{x}})^{-1} \end{pmatrix}, \end{aligned}$$

and we see from the above decomposition that every entry of the matrix $V_{(x,-x)}^{-1}$ coincides, apart of a constant, with one entry of either $V_{\hat{x}}^{-1}$ or $(D_x V_{\hat{x}})^{-1}$. On the basis of this argument we obtain the following result:

Corollary 5. *Let $F : \mathbb{R}_+^n \mapsto \mathbb{C}^{4n^2}$ be defined as $F(x) = \mathrm{Vec}(D_c V_{(x,-x)}^{-1} D_d)$, where D_c, D_d are diagonal matrices of order $2n$, whose entries do not depend on x. Moreover, for $x = (x_1, \ldots, x_n)^T$, let $\hat{x} = (x_1^2, \ldots, x_n^2)^T$. Then,*

$$c(F, x) \leq 2(n-1)(2\Delta(\hat{x}) + 1) + 1.$$

Proof. We can set $D_c = D_d = I$, without loss in generality. Let $F^{(1)}, F^{(2)} : \mathbb{R}_+^n \mapsto \mathbb{C}^{n^2}$ be defined as $F^{(1)}(x) = \mathrm{Vec}(V_{\hat{x}}^{-1})$, and $F^{(2)}(x) = \mathrm{Vec}((D_x V_{\hat{x}})^{-1})$. By virtue of the preceding argument we have

$$c(F, x) \leq \max\{c(F^{(1)}, x), c(F^{(2)}, x)\}.$$

Our goal reduces to obtain upper bounds for $c(F^{(1)}, x)$ and $c(F^{(2)}, x)$.

First, observe that $F^{(1)}(x) = G(H(x))$, where $G(x) = \mathrm{Vec}(V_x^{-1})$ and $H(x) = \hat{x}$. Hence $c(F^{(1)}, x) \leq c(G, H(x))c(H, x)$, see [11, p. 692]. From Corollary 4 we have $c(G, H(x)) \leq (n-1)(2\Delta(\hat{x})+1)$. Moreover, from (6) we obtain $c(H, x) = 2$. Hence $c(F^{(1)}, x) \leq 2(n-1)(2\Delta(\hat{x}) + 1)$.

In order to estimate the componentwise conditioning of $F^{(2)}$, denote $V_{\hat{x}}^{-1} \equiv (\hat{v}_{i,j})$. Then $(D_x V_{\hat{x}})^{-1} \equiv (\hat{v}_{i,j}/x_j)$. Let fix one particular pair (i, j), and consider the scalar function $F_{i,j}^{(2)}(x) = \hat{v}_{i,j}/x_j$, considering $\hat{v}_{i,j}$ a function of x. Clearly we have $c(F^{(2)}, x) = \max_{i,j} c(F_{i,j}^{(2)}, x)$. Again using (6) we obtain

$$\begin{aligned} c(F_{i,j}^{(2)}, x) &= \sum_k \left| \frac{x_k}{F_{i,j}^{(2)}(x)} \frac{\partial F_{i,j}^{(2)}(x)}{\partial x_k} \right| \\ &\leq \sum_{k \neq j} \left| \frac{x_k}{\hat{v}_{i,j}} \frac{\partial \hat{v}_{i,j}}{\partial x_k} \right| + \left| \frac{x_j^2}{\hat{v}_{i,j}} \frac{\partial}{\partial x_j} \left(\frac{\hat{v}_{i,j}}{x_j} \right) \right| \\ &\leq c(F^{(1)}, x) + 1. \end{aligned}$$

By Corollary 4, we have $c(F^{(2)}, x) \leq 2(n-1)(2\Delta(\hat{x}) + 1) + 1$ and the proof is complete. □

Observe that in the above corollary we have

$$\Delta(\hat{x}) = \max_{i \neq j} \frac{x_i^2}{|x_i^2 - x_j^2|} = \max_{i \neq j} \frac{x_i}{|x_i - x_j|} \frac{x_i}{x_i + x_j} < \Delta(x).$$

4. Cauchy-Vandermonde matrices

Let $x = (x_1, \ldots, x_n)^T \in \mathbb{C}^n$ and $y = (y_1, \ldots, y_k)^T \in \mathbb{C}^k$, where $0 < k < n$, have pairwise distinct entries and $x_i \neq y_j$ for all i, j. The $n \times n$ matrix

$$K_{x,y} = \left(\begin{array}{ccc|ccc} \frac{1}{x_1-y_1} & \cdots & \frac{1}{x_1-y_k} & 1 & \cdots & x_1^{n-k-1} \\ \vdots & & \vdots & \vdots & & \vdots \\ \frac{1}{x_n-y_1} & \cdots & \frac{1}{x_n-y_k} & 1 & \cdots & x_n^{n-k-1} \end{array} \right)$$

is called *Cauchy-Vandermonde matrix*. Matrices with the above structure appear in connection with rational interpolation problems for functions with prescribed poles, see [9, 12, 18, 19]. In particular, in [9, 18, 19] special factorizations of the inverse of $K_{x,y}$ are deduced from representation formulas of the rational function solving a particular interpolation problem; such factorizations allows to compute the solution of linear systems with a Cauchy-Vandermonde matrix at the cost of $O(n^2)$ or even $O(n \log^2 n)$ arithmetic operations. Other low-cost algorithms for such systems, based on recursions of Levinson and Schur type, are presented in [12].

For any $0 \leq \theta < 2\pi$, the matrix $K_{x,y}$ fulfills the displacement equation

$$D_x K_{x,y} - K_{x,y} N_\theta = \begin{pmatrix} x_1^{n-k} - \mathrm{e}^{\mathrm{i}(n-k)\theta} \\ \vdots \\ x_n^{n-k} - \mathrm{e}^{\mathrm{i}(n-k)\theta} \end{pmatrix} e_n^T, \qquad N_\theta = \begin{pmatrix} D_y & O \\ e_1 \mathbf{1}^T & P_\theta \end{pmatrix},$$

where the matrix N_θ has square diagonal blocks and P_θ is as in (12). Analogously to the Vandermonde case, we can limit ourselves to consider the case $\theta = 0$. Indeed, if we let $\omega = \mathrm{e}^{-\mathrm{i}\theta}$, we have

$$K_{\omega x, \omega y} = K_{x,y} \Omega, \qquad \Omega = \begin{pmatrix} \omega^{-1} I & O \\ O & \mathrm{Diag}(1, \omega, \ldots, \omega^{n-k-1}) \end{pmatrix}. \tag{15}$$

Since Ω is diagonal and unitary, the entries of $K_{\omega x, \omega y}$ and its inverse have the same modulus of the corresponding entries of $K_{x,y}$ and its inverse, respectively, hence their structured conditioning is the same.

We stress the fact that, differently to the case of Cauchy and Vandermonde matrices, no closed-form formulas are known for the entries of $K_{x,y}^{-1}$, for generic k. Indeed, the inversion formulas presented in [9] involve a polynomial division, and the inversion algorithms introduced in [12, 18, 19] have a recursive character, namely, the entries of $K_{x,y}^{-1}$ are computed according to suitable orderings. Hence, it is not obvious how to study the structured conditioning of $K_{x,y}$ by a straightforward use of the definitions. In the sequel we will exploit a factorization approach analogous to the one already introduced in the preceding section.

Lemma 3. *Let* $\xi = (\xi_1, \ldots, \xi_{n-k})^T$, $\xi_j = \mathrm{e}^{\mathrm{i}(j-1)2\pi/(n-k)}$. *Furthermore, let* $a = (x_1^{n-k} - 1, \ldots, x_n^{n-k} - 1)^T$ *and* $\hat{y} \in \mathbb{C}^n$, $\hat{y} = (y_1, \ldots, y_k, \xi_1, \ldots, \xi_{n-k})^T$. *Let*

$$T = \begin{pmatrix} I & O \\ (n-k)^{-1/2} C_{\xi,y} & \Phi^{-1} \end{pmatrix}, \tag{16}$$

partitioned as N_0, i.e., the Cauchy matrix $C_{\xi,y}$ has order $(n-k)\times k$ and $\Phi = (n-k)^{-1/2}V_\xi$ is the Fourier matrix of order $n-k$. Furthermore, let f be the solution of the linear system $T^T f = e_n$. If $\Delta(x,\hat{y}) < +\infty$ then

$$K_{x,y} = D_a C_{x,\hat{y}} D_f T.$$

Proof. It is a simple task to verify that

$$T^{-1} = \begin{pmatrix} I & O \\ -(n-k)^{-1/2}\Phi C_{\xi,y} & \Phi \end{pmatrix}. \tag{17}$$

Moreover, since $\Phi^{-1}e_1 = (n-k)^{-1/2}\mathbf{1}$ and $\Phi^{-1}P_0\Phi = D_\xi$, we have

$$\begin{aligned} TN_0T^{-1} &= \begin{pmatrix} I & O \\ (n-k)^{-1/2}C_{\xi,y} & \sqrt{n-k}\Phi^{-1} \end{pmatrix}\begin{pmatrix} D_y & O \\ e_1\mathbf{1}^T & P_0 \end{pmatrix} \\ &\quad \times \begin{pmatrix} I & O \\ -(n-k)^{-1/2}\Phi C_{\xi,y} & \Phi \end{pmatrix} \\ &= \begin{pmatrix} D_y & O \\ (n-k)^{-1/2}[C_{\xi,y}D_y - D_\xi C_{\xi,y} + \mathbf{1}\mathbf{1}^T] & \Phi^{-1}P_0\Phi \end{pmatrix} \\ &= D_{\hat{y}}. \end{aligned}$$

Hence $T^{-1}D_{\hat{y}}T = N_0$. Since diagonal matrices commute, we obtain:

$$\begin{aligned} \mathcal{L}_{D_x,D_{\hat{y}}}(D_a^{-1}K_{x,y}T^{-1}) &= D_a^{-1}[D_xK_{x,y} - K_{x,y}N_0]T^{-1} \\ &= D_a^{-1}ae_n^TT^{-1} \\ &= \mathbf{1}\mathbf{1}^TD_f. \end{aligned}$$

The hypothesis stated on x implies both the invertibility of D_a and that of the operator $\mathcal{L}_{D_x,D_{\hat{y}}}$. We obtain $D_a^{-1}K_{x,y}T^{-1} = C_{x,\hat{y}}D_f$, whence thesis follows. □

We remark that $\det(K_{x,y}) \neq 0$ if all entries from x and y are distinct, see, e.g., [9, Thm. 3.1] or [19]. As a consequence, by Binet-Cauchy Theorem we see that, in the hypotheses of the preceding lemma, we have $\det(D_f) \neq 0$, that is, f has no zero entries.

We consider in the following theorem the structured conditioning of the columns of $K_{x,y}^{-1}$ with respect to perturbations in the vector x.

We will use some further notations: For any $x = (x_1,\ldots,x_n)^T \in \mathbb{C}^n$, consider the two subvectors

$$x^{(1)} = (x_1,\ldots,x_k)^T \in \mathbb{C}^k, \qquad x^{(2)} = (x_{k+1},\ldots,x_n)^T \in \mathbb{C}^{n-k},$$

where k is the integer appearing in the definition of the Cauchy-Vandermonde matrix under consideration. Furthermore, introduce the relative distance

$$\hat{\delta}(x,\tilde{x}) = \max\{\delta(x^{(1)},\tilde{x}^{(1)}), \|x^{(2)} - \tilde{x}^{(2)}\|/\|x\|\},$$

and consider the following condition measure, naturally induced by it:

$$\widehat{m}(F,x) = \lim_{\epsilon\to 0}\sup_{\delta(x,\tilde{x})\le\epsilon} \frac{\hat{\delta}(F(x),F(\tilde{x}))}{\delta(x,\tilde{x})}. \tag{18}$$

Remark that $\|x-\tilde{x}\|/\|x\| \leq \hat{\delta}(x,\tilde{x}) \leq \delta(x,\tilde{x})$, and

$$m(F,x) \leq \widehat{m}(F,x). \tag{19}$$

Theorem 4. *Let $1 \leq i \leq n$ be fixed, and let $F_i : \mathbb{C}^n \mapsto \mathbb{C}^n$ be the function defined as $F_i(x) = K_{x,y}^{-1} e_i$. Moreover, let $0 \leq \theta < 2\pi$ be arbitrary and let $\omega = \mathrm{e}^{-\mathrm{i}\theta}$. Then,*

$$\begin{aligned} \widehat{m}(F_i,x) &\leq \left(\frac{\gamma}{\sqrt{n-k}} + \sqrt{n-k}\right)(\gamma + \sqrt{n-k}) \\ &\quad \times \left(\Delta(n,\omega x,\xi) + \left|\frac{(n-k)x_i^{n-k}}{x_i^{n-k} - \mathrm{e}^{\mathrm{i}(n-k)\theta}}\right|\right), \end{aligned}$$

where $\gamma = \|C_{\xi,\omega y}\|$, $\xi = (\xi_1,\ldots,\xi_{n-k})$, $\xi_j = \mathrm{e}^{\mathrm{i}(j-1)2\pi/(n-k)}$ and $\Delta(n,\omega x,\xi)$ can be derived from (11).

Proof. As in Theorem 2, we can reduce the analysis to the case $\theta = 0$ by simply considering the rotated vectors $\hat{x} = \omega x$ and $\hat{y} = \omega y$, by virtue of (15). In the notations of Lemma 3 we have

$$K_{x,y}^{-1} = T^{-1} D_f^{-1} C_{x,\hat{y}}^{-1} D_a^{-1}.$$

Recall that f has no zero entries, hence D_f is invertible. We obtain

$$F_i(x) = \frac{1}{x_i^{n-k} - 1} T^{-1} D_f^{-1} C_{x,\hat{y}}^{-1} e_i.$$

Then, $F_i(x) = G(H_i(x))$, where $G(x) = T^{-1}x$ and

$$H_i(x) = (x_i^{n-k} - 1)^{-1} D_f^{-1} C_{x,\xi}^{-1} e_i.$$

Using (18), it is straightforward to check that $\widehat{m}(F_i,x) \leq \widehat{m}(G,H_i(x))c(H_i,x)$.

Let $u = G(x)$ and $\tilde{u} = G(\tilde{x})$, for arbitrary $x, \tilde{x}$. In order to bound the quantity $\hat{\delta}(u,\tilde{u})$, observe firstly that $\delta(u^{(1)},\tilde{u}^{(1)}) = \delta(x^{(1)},\tilde{x}^{(1)}) \leq \delta(x,\tilde{x})$. Using $\|u\| \geq \|x\|/\|T\|$, we have:

$$\begin{aligned} \frac{\|u^{(2)} - \tilde{u}^{(2)}\|}{\|u\|} &\leq \|T\| \frac{\|(n-k)^{-1/2}\Phi C_{\xi,y}(x^{(1)} - \tilde{x}^{(1)})\| + \|\Phi(x^{(2)} - \tilde{x}^{(2)})\|}{\|x\|} \\ &\leq \|T\| \left(\gamma \frac{\|x^{(1)} - \tilde{x}^{(1)}\|}{\|x\|} + \sqrt{n-k}\frac{\|x^{(2)} - \tilde{x}^{(2)}\|}{\|x\|}\right) \\ &\leq \|T\|(\gamma + \sqrt{n-k})\delta(x,\tilde{x}). \end{aligned}$$

Since $\|T\| \leq \gamma/\sqrt{n-k} + \sqrt{n-k}$ we have

$$\hat{\delta}(u,\tilde{u}) \leq (\gamma/\sqrt{n-k} + \sqrt{n-k})(\gamma + \sqrt{n-k})\delta(x,\tilde{x}).$$

From (18) we obtain

$$\widehat{m}(G,x) \leq (\gamma/\sqrt{n-k} + \sqrt{n-k})(\gamma + \sqrt{n-k}).$$

Furthermore, if we let $\delta(x,\tilde{x}) = \epsilon$, we have from Theorem 1

$$\begin{aligned} |H_i(x) - H_i(\tilde{x})| &\leq |x_i^{n-k} - 1|^{-1} |D_f^{-1}| |C_{x,\xi}^{-1} e_i - C_{\tilde{x},\xi}^{-1} e_i| \\ &\quad + \left| \frac{1}{x_i^{n-k} - 1} - \frac{1}{\tilde{x}_i^{n-k} - 1} \right| |D_f^{-1} C_{\tilde{x},\xi}^{-1} e_i| \\ &\leq \epsilon |x_i^{n-k} - 1|^{-1} |D_f^{-1} C_{x,\xi}^{-1} e_i| \Delta(n,x,\xi) \\ &\quad + \epsilon \frac{(n-k)|x_i^{n-k}|}{|x_i^{n-k} - 1|^2} |D_f^{-1} C_{x,\xi}^{-1} e_i| + o(\epsilon) \\ &\leq \epsilon \Delta(n,x,\xi) |H_i(x)| + \epsilon \left| \frac{(n-k) x_i^{n-k}}{x_i^{n-k} - 1} \right| |H_i(x)| + o(\epsilon), \end{aligned}$$

whence we obtain

$$c(H_i, x) \leq \Delta(n,x,\xi) + \left| \frac{(n-k) x_i^{n-k}}{x_i^{n-k} - 1} \right|,$$

and the proof is complete. □

Remark that an upper bound for the constant γ appearing in the preceding theorem is obtained as follows:

$$\gamma = \max_{1 \leq i \leq n-k} \sum_{j=1}^{k} \frac{1}{|\xi_i - \omega y_j|} \leq \frac{k}{\min_{i,j} |\xi_i - \omega y_j|} = k \Delta(\xi, \omega y).$$

We omit the proof of the following corollary, as it is essentially the same as that of Corollary 3, in the light of (19):

Corollary 6. *Let $F : \mathbb{C}^n \mapsto \mathbb{C}^{n^2}$ be defined as $F(x) = \mathrm{Vec}(K_{x,y}^{-1} D_d)$, where D_d is an arbitrary diagonal matrix whose entries do not depend on x. Furthermore, let $\xi = (\xi_1, \ldots, \xi_{n-k})$, $\xi_j = \mathrm{e}^{\mathrm{i}(j-1)2\pi/(n-k)}$. For any $0 \leq \theta < 2\pi$, let $\omega = \mathrm{e}^{-\mathrm{i}\theta}$. Then,*

$$\begin{aligned} m(F,x) &\leq (\gamma/\sqrt{n-k} + \sqrt{n-k})(\gamma + \sqrt{n-k}) \\ &\quad \times \left(\Delta(n, \omega x, \xi) + (n-k) \max_i \left| \frac{x_i^{n-k}}{x_i^{n-k} - \mathrm{e}^{\mathrm{i}(n-k)\theta}} \right| \right), \end{aligned}$$

where $\gamma = \|C_{\xi,\omega y}\|$ and $\Delta(n, \omega x, \xi)$ can be obtained from (11).

References

[1] S.G. Bartels and D.J. Higham; The structured sensitivity of Vandermonde-like systems. *Numer. Math.* 62 (1992), 17–33.

[2] B. Beckermann; The condition number of real Vandermonde, Krylov and positive definite Hankel matrices. *Numer. Math.* 85 (2000), 553–577.

[3] T. Bella, Y. Eidelman, I. Gohberg, I. Koltracht and V. Olshevsky; A Björck-Pereyra-type algorithm for Szegö-Vandermonde matrices based on properties of unitary Hessenberg matrices. *Linear Algebra Appl.* 420 (2007), 634–647.

[4] T. Boros, T. Kailath and V. Olshevsky; A fast parallel Björck-Pereyra-type algorithm for solving Cauchy linear equations. *Linear Algebra Appl.* 302/303 (1999), 265–293.

[5] T. Boros, T. Kailath and V. Olshevsky; Pivoting and backward stability of fast algorithms for solving Cauchy linear equations. *Linear Algebra Appl.* 343/344 (2002), 63–99.

[6] E. Bozzo, D. Fasino and O. Menchi; The componentwise conditioning of the DFT. *Calcolo* 39 (2002), 181–187.

[7] F. Cucker and H. Diao; Mixed and componentwise condition numbers for rectangular structured matrices. *Calcolo* 44 (2007), 89–115.

[8] D. Fasino and V. Olshevsky; How bad are symmetric Pick matrices?, in: *Structured Matrices in Operator Theory, Numerical Analysis, Control, Signal and Image Processing* (V. Olshevsky, Ed.) AMS Series on Contemporary Mathematics, 280 (2001), 301–311.

[9] T. Finck, G. Heinig and K. Rost; An inversion formula and fast algorithms for Cauchy-Vandermonde matrices. *Linear Algebra Appl.* 183 (1993), 179–191.

[10] I. Gohberg and I. Koltracht; On the inversion of Cauchy matrices, in: M.A. Kaashoek, J.H. van Schuppen, A.C.M. Ran (Eds.), *Signal processing, scattering and operator theory, and numerical methods* (Proceedings of MTNS-98), pp. 381–392; Birkhäuser, 1990.

[11] I. Gohberg and I. Koltracht; Mixed, componentwise, and structured condition numbers. *SIAM J. Matrix Anal. Appl.* 14 (1993), 688–704.

[12] G. Heinig and K. Rost; Recursive solution of Cauchy-Vandermonde systems of equations. *Linear Algebra Appl.* 218 (1995), 59–72.

[13] G. Heinig and K. Rost; *Representations of inverses of real Toeplitz-plus-Hankel matrices using trigonometric transformations*, in: Large-Scale Scientific Computation of Engineering and Environmental Problems II (M. Griebel, S. Margenov and P. Yalamov, Eds.), Vieweg, 73 (2000) 80–86.

[14] D.J. Higham and N.J. Higham; Backward error and condition of structured linear systems. *SIAM J. Matrix Anal. Appl.* 13 (1992), 162–175.

[15] N.J. Higham; Error analysis of the Björck-Pereyra algorithms for solving Vandermonde systems. *Numer. Math.* 50 (1987), 613–632.

[16] R.-C. Li; Asymptotically optimal lower bounds for the condition number of a real Vandermonde matrix. *SIAM J. Matrix Anal. Appl.* 28 (2006), 829–844.

[17] V. Olshevsky and V. Pan; *Polynomial and rational evaluation and interpolation* (*with structured matrices*). ICALP99 Proceedings, Springer, LNCS 1644 (1999), 585–594.

[18] G. Mühlbach; Interpolation by Cauchy-Vandermonde systems and applications. *J. Comput. Appl. Math.* 122 (2000), 203–222.

[19] G. Mühlbach; On Hermite interpolation by Cauchy-Vandermonde systems: the Lagrange formula, the adjoint and the inverse of a Cauchy-Vandermonde matrix. *J. Comput. Appl. Math.* 67 (1996), 147–159.

[20] T. Penzl; Eigenvalue decay bounds for solutions of Lyapunov equations: the symmetric case. *Systems Control Lett.* 40 (2000), 139–144.

[21] S. Serra Capizzano; An elementary proof of the exponential conditioning of real Vandermonde matrices. *Boll. Unione Mat. Ital. Sez. B Artic. Ric. Mat.* 10 (2007), 761–768.

[22] J.-G. Sun; Bounds for the structured backward errors of Vandermonde systems. *SIAM J. Matrix Anal. Appl.* 20 (1999), 45–59.

[23] E.E. Tyrtyshnikov; Singular values of Cauchy-Toeplitz matrices. *Linear Algebra Appl.* 161 (1992), 99–116.

[24] E.E. Tyrtyshnikov; How bad are Hankel matrices? *Numer. Math.* 67 (1994), 261–269.

[25] H. Xiang and Y. Wei; Structured mixed and componentwise condition numbers of some structured matrices. *J. Comput. Appl. Math.* 202 (2007), 217–229.

Enrico Bozzo and Dario Fasino
Dipartimento di Matematica e Informatica
Università di Udine
Via delle Scienze, 208
I-33100 Udine, Italy
e-mail: dario.fasino@dimi.uniud.it
enrico.bozzo@dimi.uniud.it

Operator Theory:
Advances and Applications, Vol. 199, 221–227

Factorizations of Totally Negative Matrices

V. Cortés and J.M. Peña

In memory of Georg Heinig

Abstract. A matrix is called totally negative if all its minors are negative. In this paper we characterize two decompositions of totally negative matrices: the QR decomposition and the symmetric-triangular decomposition.

Mathematics Subject Classification (2000). 15A23; 65F25; 15A48; 15A15.

Keywords. Totally negative matrix, QR decomposition, symmetric-triangular decomposition.

1. Introduction and background

This paper deals with two factorizations of totally negative matrices relevant in numerical analysis: QR decomposition and symmetric-triangular decomposition (see [9]). A matrix is totally negative (TN) if all its minors are negative. Totally negative matrices belong to the class of N-matrices (matrices with all principal minors negative), which play an important role in Economy (see [3], [10] and [13]). Others aspects of totally negative matrices have been considered in [2], [5] and [7]. Besides, they belong to the class of strictly sign regular matrices, which will be defined below and which are very important in many applications due to their property known as variation diminishing property (see [1], [11] and [12]). Let us start with basic notations.

Following [1] and [6], for $k, n \in N$, $1 \leq k \leq n$, $Q_{k,n}$ will denote the set of all increasing sequences of k natural numbers not greater than n. For each $\alpha \in Q_{k,n}$, its dispersion number $d(\alpha)$ is defined by $d(\alpha) := \sum_{i=1}^{k-1}(\alpha_{i+1} - \alpha_i - 1) = \alpha_k - \alpha_1 - (k-1)$, with the convention $d(\alpha) = 0$ for $\alpha \in Q_{1,n}$. Let us observe that $d(\alpha) = 0$ means that α consists of k consecutive integers.

For $\alpha = (\alpha_1, \alpha_2, \ldots, \alpha_k)$, $\beta = (\beta_1, \beta_2, \ldots, \beta_k) \in Q_{k,n}$ and A an $n \times n$ matrix, we denote by $A[\alpha|\beta]$ the $k \times k$ submatrix of A containing rows $\alpha_1, \alpha_2, \ldots, \alpha_k$ and columns $\beta_1, \beta_2, \ldots, \beta_k$ of A. If $\alpha = \beta$, we denote by $A[\alpha] := A[\alpha|\alpha]$ the corresponding principal minor. A *column-initial minor* of A is a minor of the form

$\det A[\alpha|1,\dots,k]$, where $\alpha \in Q_{k,n}$ with $d(\alpha) = 0$ and $1 \leq k \leq n$. Analogously, a *row-initial minor* of A is a minor of the form $\det A[1,\dots,k|\alpha]$, where α and k are given as above. Let A be an $n \times n$ lower (resp., upper) triangular matrix. Following [4], the minors $\det A[\alpha|\beta]$ with $\alpha_k \geq \beta_k$ (resp., $\alpha_k \leq \beta_k\ \forall k$) are called nontrivial minors of A because all the remaining minors are obviously equal to zero. A triangular matrix A is called ΔTP if its nontrivial minors are all positive. We remark that these matrices are also called ΔSTP in other papers (see [8]).

By a *signature sequence* we mean an (infinite) real sequence $\varepsilon = (\varepsilon_i)$ with $|\varepsilon_i| = 1$, $i = 1, 2, \dots$. An $n \times n$ matrix A verifying $\varepsilon_k \det A[\alpha|\beta] > 0$ for all $\alpha, \beta \in Q_{k,n}$ and $k = 1, \dots, n$ is called *strictly sign regular* with signature $\varepsilon_1, \dots, \varepsilon_n$, and will be denoted by SSR.

Two important subclasses of the strictly sign regular matrices are the totally positive matrices (with all its minors positive) and the totally negative matrices (defined above). This terminology is more frequent nowadays, although these matrices also have been called in the literature as strictly totally positive and strictly totally negative, respectively.

The first part of the following result comes from Theorem 3.1 of [4] and the second part is provided by Remark 3.6 of [7].

Theorem 1.1.

(i) *A lower (resp., upper) triangular matrix M is ΔTP if and only if all column-initial (resp., row-initial) minors of M are positive.*

(ii) *Let $A = (a_{ij})_{1 \leq i,j \leq n}$ be a nonsingular matrix and $a_{nn} < 0$. Then A is TN if and only if all its initial minors are negative.*

An LDU-factorization of a matrix A is the decomposition $A = LDU$ where L (resp., U) is a lower (resp., upper) triangular, unit diagonal matrix (i.e., with all diagonal entries equal to 1), and D is a diagonal matrix. From now on, $A = LDU$ will refer to this decomposition.

Let us recall the following result, which will be used in the following sections. The proof can be seen in Proposition 2.1 of [7].

Theorem 1.2. *If A is SSR, then $A = LDU$ with L (resp., U) a ΔTP and lower (resp., upper) triangular, unit diagonal matrix and D a diagonal nonsingular matrix.*

Finally, let us recall the well-known Cauchy-Binet formula. If A, B are $n \times n$ matrices, then we have the following determinantal identity:

$$\det(AB)[\alpha|\beta] = \sum_{w \in Q_{k,n}} \det A[\alpha|w] \det B[w|\beta], \quad \alpha, \beta \in Q_{k,n}.$$

Section 2 contains some auxiliary results. In Section 3 we characterize the QR factorization of totally negative matrices. Section 4 contains the characterization of the symmetric-triangular factorization of totally negative matrices and includes an example showing that a condition used in our characterizations (the fact that $a_{nn} < 0$) cannot be suppressed.

2. Auxiliary results

We are going to define some special classes of matrices which will play a key role in the characterizations of this paper.

Definition 2.1. A nonsingular matrix A is said to be *lowerly TN* if it can be decomposed in the form $A = LDU$ and $LD\Sigma$ is ΔTP, where Σ is a diagonal matrix with diagonal entries $-1, +1, \ldots, +1$. If, in addition, U^{-1} satisfies that $\Sigma U^{-1}\Sigma$ is ΔTP, then the matrix is called *strict γ^--matrix*.

In the previous definitions, since L is unit diagonal, the fact that $LD\Sigma$ is ΔTP implies that L and $D\Sigma$ are ΔTP.

The following result relates the two concepts introduced in the previous definition.

Proposition 2.2. *If $A = LDU$ and $(A^T)^{-1}$ are lowerly TN, then A is a strict γ^--matrix. Therefore, an orthogonal matrix is lowerly TN if and only if it is strict γ^--matrix.*

Proof. We have to see that, under the conditions of the proposition, $\Sigma U^{-1}\Sigma$ is ΔTP. From the factorizations $A = LDU$ and $(A^T)^{-1} = \tilde{L}\tilde{D}\tilde{U}$ we get

$$(U^T)^{-1} = DL^T(A^T)^{-1} = DL^T(\tilde{L}\tilde{D}\tilde{U}). \tag{2.1}$$

Since A and $(A^T)^{-1}$ are lowerly TN, we have from (2.1) that

$$\Sigma(U^T)^{-1}\Sigma = (\Sigma DL^T)(\tilde{L}\tilde{D}\Sigma)(\Sigma\tilde{U}\Sigma), \tag{2.2}$$

where the upper triangular matrix ΣDL^T and the lower triangular matrix $\tilde{L}\tilde{D}\Sigma$ are ΔTP. Moreover, $\Sigma\tilde{U}\Sigma$ is an upper triangular, unit diagonal matrix.

By the Cauchy-Binet formula, for all $\alpha \in Q_{k,n}$ with $1 \le k \le n$ and $d(\alpha) = 0$ we can derive from (2.2)

$$\det B[\alpha|1,\ldots,k] = \det C[\alpha|1,\ldots,k]\det F[1,\ldots,k] = \det C[\alpha|1,\ldots,k], \tag{2.3}$$

where $B := \Sigma(U^T)^{-1}\Sigma$, $C := (\Sigma DL^T)(\tilde{L}\tilde{D}\Sigma)$ and $F := \Sigma\tilde{U}\Sigma$.

Again by the Cauchy-Binet formula, for all $\alpha \in Q_{k,n}$ with $d(\alpha) = 0$, we have

$$\det C[\alpha|1,\ldots,k] = \sum_{\beta\in Q_{k,n}} \det(\Sigma DL^T)[\alpha|\beta]\det(\tilde{L}\tilde{D}\Sigma)[\beta|1,\ldots,k]. \tag{2.4}$$

From (2.4) and taking into account that ΣDL^T and $\tilde{L}\tilde{D}\Sigma$ are ΔTP, we obtain $\det C[\alpha|1,\ldots,k] > 0$ for all $\alpha \in Q_{k,n}$ with $d(\alpha) = 0$. Then, from (2.3) $\det B[\alpha|1,\ldots,k] > 0$ for all $\alpha \in Q_{k,n}$ with $d(\alpha) = 0$. Therefore, all column-initial minors of the lower triangular matrix B are positive and so, by Theorem 1.1 (i), $B = \Sigma(U^T)^{-1}\Sigma$ is ΔTP, which implies $\Sigma U^{-1}\Sigma$ is ΔTP. □

The following result for lowerly TN matrices will be very useful in this paper.

Proposition 2.3. *Let A be an $n \times n$ matrix. If $A = CV$ with C lowerly TN and V upper triangular with positive diagonal, then all column-initial minors of A are negative.*

Proof. Since C is lowerly TN, we have that $C = L_C D_C U_C$ and the matrix $L_C D_C \Sigma$ is ΔTP, where Σ is a diagonal matrix with diagonal entries $-1, +1, \ldots, +1$. Then by the Cauchy-Binet formula, we get for $k = 1, \ldots, n$

$$0 < \det(L_C D_C \Sigma)[\alpha|1, \ldots, k] = \det(L_C D_C)[\alpha|1, \ldots, k] \det \Sigma[1, \ldots, k] = \\ - \det(L_C D_C)[\alpha|1, \ldots, k]$$

for all $\alpha \in Q_{k,n}$ and $d(\alpha) = 0$. So, the sign of $\det(L_C D_C)[\alpha|1, \ldots, k]$ is -1.

Again by the Cauchy-Binet formula, we get

$$\det C[\alpha|1, \ldots, k] = \det(L_C D_C)[\alpha|1, \ldots, k] \det U_C[1, \ldots, k] = \\ \det(L_C D_C)[\alpha|1, \ldots, k]$$

for all $\alpha \in Q_{k,n}$ and $d(\alpha) = 0$. So, all column-initial minors of order k of C are negative. Applying again the Cauchy-Binet formula to $A = CV$, we obtain

$$\det A[\alpha|1, \ldots, k] = \det C[\alpha|1, \ldots, k] \det V[1, \ldots, k] \tag{2.5}$$

for all $\alpha \in Q_{k,n}$ and $d(\alpha) = 0$. Since V has positive diagonal entries, we deduce from (2.5) that all column-initial minors of A are negative. □

3. QR decomposition of TN matrices

The following theorem gives a characterization of TN matrices by their QR factorization.

Theorem 3.1. *Let A be an $n \times n$ matrix. A is TN if and only if $a_{nn} < 0$ and*

$$A = QR, \quad A^T = \tilde{Q}\tilde{R}, \tag{3.1}$$

where Q and $\tilde{Q}$ are orthogonal and strict γ^--matrices and R and $\tilde{R}$ are upper triangular and ΔTP matrices.

Proof. If B is a TN matrix, then $b_{nn} < 0$. Since B is nonsingular, $B = QR$ with Q orthogonal and R upper triangular with positive diagonal entries. If $L_Q D_Q U_Q$ is the LDU-factorization of Q, then we have

$$B = L_Q D_Q (U_Q R). \tag{3.2}$$

Since R is upper triangular with positive diagonal, we can write $U_Q R = \bar{D}\bar{U}$, where $\bar{U}$ is upper triangular, unit diagonal matrix and $\bar{D}$ is a diagonal matrix with positive diagonal. So, from (3.2) we get

$$B = L_Q (D_Q \bar{D}) \bar{U}. \tag{3.3}$$

Since B is TN (so that, B is SSR with signature $\varepsilon_i = -1$ for all i, by Theorem 1.2, $B = LDU$ with D a diagonal nonsingular matrix and L (resp., U) ΔTP and lower (resp., upper) triangular with unit diagonal. The uniqueness of the LDU-factorization implies by (3.3) that $L_Q = L$, $D = D_Q \bar{D}$, $U = \bar{U}$, and so L_Q is ΔTP. The diagonal entries d_i of D satisfy

$$d_i = \frac{\det B[1, \ldots, i]}{\det B[1, \ldots, i-1]} \tag{3.4}$$

and so $\text{sign}(d_i) = \varepsilon_i/\varepsilon_{i-1} = +1$ for $i = 2, \ldots, n$, and $\text{sign}(d_1) = \varepsilon_1/\varepsilon_0 = -1$, taking $\varepsilon_0 := +1$. Hence, if Σ is a diagonal matrix with diagonal entries $-1, +1, \ldots, +1$, $D\Sigma = D_Q\bar{D}\Sigma$ has positive diagonal entries and so $D_Q\Sigma$ is also a diagonal matrix with positive diagonal. Therefore, Q is lowerly TN and by Proposition 2.2 it is a strict γ^--matrix.

Since $B = QR$,

$$B^TB = R^TR \tag{3.5}$$

and R^TR can be computed from R by multiplying each row by a positive number and adding a linear combination of the previous rows. Therefore, the row-initial minors of R^TR have the same sign as the corresponding minors of R. Since the product of two TN matrices is TP by Theorem 3.1 of [1], we deduce from (3.5) that R^TR is TP. Then the row-initial minors of R^TR and R are positive, and by Theorem 1.1 (i), R is ΔTP.

Now, since A and A^T are TN, applying the previous reasoning to A and A^T, (3.1) follows with the stated properties.

Conversely, applying Proposition 2.3 to the matrices A and A^T with $C = Q$, $\tilde{Q}$ and $V = R$, $\tilde{R}$ (respectively), we conclude that all the initial minors of A are negative. Then, taking into account that $a_{nn} < 0$, by Proposition 1.1 (ii) A is TN. □

4. Symmetric-triangular decomposition of TN matrices

Recently, Golub and Yuan have introduced in [9] a symmetric-triangular decomposition of a nonsingular matrix. We now characterize TN matrices in terms of this decomposition.

Theorem 4.1. *Let A be an $n \times n$ matrix. A is TN if and only if $a_{nn} < 0$ and*

$$A = SR, \quad A^T = \tilde{S}\tilde{R}, \tag{4.1}$$

where S and $\tilde{S}$ are symmetric and lowerly TN matrices and R and $\tilde{R}$ are upper triangular matrices with unit diagonal.

Proof. Let B be a TN matrix. By Theorem 1.2, $B = LDU$ with D a diagonal nonsingular matrix and L (resp., U) ΔTP and lower (upper) triangular with unit diagonal. If we write $B = (LDL^T)(L^T)^{-1}U$, then we can consider

$$S := LDL^T,\ R := (L^T)^{-1}U, \tag{4.2}$$

where S is a nonsingular and symmetric matrix and R is an upper triangular matrix with unit diagonal.

Now, let us see S is lowerly TN. So, we have to prove that $LD\Sigma$ is ΔTP, where Σ is a diagonal matrix with diagonal entries $-1, +1, \ldots, +1$. Reasoning through (3.4) as in the proof of Theorem 3.1, it can be deduced that $D\Sigma$ is a diagonal matrix with positive diagonal. So, $LD\Sigma$ is ΔTP because L is ΔTP. Then, from (4.2), we conclude that S is lowerly TN.

Since the matrices A and A^T are TN, applying the previous reasoning to A and A^T, the necessary condition holds.

Conversely, if we apply Proposition 2.3 to the matrices A and A^T with $C = S$, $\tilde{S}$ and $V = R$, $\tilde{R}$ (respectively), then we conclude that all the initial minors of A are negative. Then, taking into account that $a_{nn} < 0$, by Proposition 1.1 (ii) A is TN. □

The following example shows that the condition $a_{nn} < 0$ cannot be suppressed in theorems 3.1 and 4.1.

Example. Let A be a nonsingular matrix given by

$$A = \begin{pmatrix} -1 & -1 \\ -1 & 0 \end{pmatrix}.$$

Then $A = QR$ with Q orthogonal and R upper triangular with positive diagonal entries given by

$$Q = \begin{pmatrix} -1/\sqrt{2} & -1/\sqrt{2} \\ -1/\sqrt{2} & 1/\sqrt{2} \end{pmatrix}, \quad R = \begin{pmatrix} \sqrt{2} & 1/\sqrt{2} \\ 0 & 1/\sqrt{2} \end{pmatrix}.$$

We also have $A^T = A = QR$. The matrix R is ΔTP. Besides, $Q = L_Q D_Q U_Q$ where

$$L_Q = \begin{pmatrix} 1 & 0 \\ 1 & 1 \end{pmatrix}, \quad D_Q = \begin{pmatrix} -1/\sqrt{2} & 0 \\ 0 & \sqrt{2} \end{pmatrix}, \quad U_Q = \begin{pmatrix} 1 & 1 \\ 0 & 1 \end{pmatrix}.$$

Clearly the matrices

$$L_Q D_Q \Sigma = \begin{pmatrix} 1/\sqrt{2} & 0 \\ 1/\sqrt{2} & \sqrt{2} \end{pmatrix}, \quad \Sigma U^{-1} \Sigma = \begin{pmatrix} 1 & 1 \\ 0 & 1 \end{pmatrix},$$

with $\Sigma = \mathrm{diag}\{-1, +1\}$, are ΔTP, and so Q is strict γ^--matrix. However, A is not TN because $\det A[2] = 0$.

As for the symmetric-triangular decomposition, we have $A = SR$ with S symmetric and R upper triangular with unit diagonal given by

$$S = \begin{pmatrix} -1 & -1 \\ -1 & 0 \end{pmatrix}, \quad R = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}.$$

We also have $A^T = A = SR$. The matrix $S = LDL^T$ where

$$L = \begin{pmatrix} 1 & 0 \\ 1 & 1 \end{pmatrix}, \quad D = \begin{pmatrix} -1 & 0 \\ 0 & 1 \end{pmatrix}.$$

Since the matrix

$$LD\Sigma = \begin{pmatrix} 1 & 0 \\ 1 & 1 \end{pmatrix}$$

with $\Sigma = \mathrm{diag}\{-1, +1\}$, is ΔTP, S is lowerly TN. However, A is not TN.

Acknowledgment

This work is partially supported by the Spanish Research Grant MTM2009-07315 and by Gobierno de Aragón and Fondo Social Europeo.

References

[1] T. Ando, *Totally positive matrices*, Linear Algebra Appl. **90** (1987) 165–219.

[2] C.M. Araújo, J.R. Torregrosa, A.M. Urbano, *Totally nonpositive completions on partial matrices*, Linear Algebra Appl. **413** (2006), 403–424.

[3] R.B. Bapat, T.E.S. Raghavan, *Nonnegative Matrices and Applications*, Cambridge University Press, New York, 1997.

[4] C.W. Cryer, *LU-factorization of totally positive matrices*, Linear Algebra Appl. **7** (1973) 83–92.

[5] S.M. Fallat, P. Van Den Driessche, *On matrices with all minors negative*, Electron. J. Linear Algebra **7** (2000), 92–99.

[6] J. Garloff, *Intervals of almost totally positive matrices*, Linear Algebra Appl. **363** (2003) 103–108.

[7] M. Gasca and J.M. Peña, *A test for strict sign-regularity*, Linear Algebra Appl. **197** (1994) 133–142.

[8] M. Gasca and J.M. Peña, *A matricial description of Neville elimination with application to total positivity*, Linear Algebra Appl. **202** (1994) 33–54.

[9] G.H. Golub and J.Y. Yuan, *Symmetric-triangular decomposition and its applications. I. Theorems and algorithms*, BIT **42** (2002) 814–822.

[10] T. Parthasarathy and G. Ravindran, *N-matrices*, Linear Algebra Appl. **139** (1990), 89–102.

[11] J. M. Peña (Ed.), *Shape preserving representations in Computer-Aided Geometric Design*, Nova Science Publishers, Commack (New York), 1999.

[12] J.M. Peña, *On nonsingular sign regular matrices*, Linear Algebra Appl. **359** (2003), 91–100.

[13] R. Saigal, *On the class of complementary cones and Lemke's algorithm*, SIAM J. Appl. Math. **23** (1972), 46–60.

V. Cortés and J.M. Peña
Depto. Matemática Aplicada
Universidad de Zaragoza
E-50009 Zaragoza, Spain
e-mail: vcortes@unizar.es
jmpena@unizar.es

Operator Theory:
Advances and Applications, Vol. 199, 229–254

QR-factorization of Displacement Structured Matrices Using a Rank Structured Matrix Approach

Steven Delvaux, Luca Gemignani and Marc Van Barel

Abstract. A general scheme is proposed for computing the QR-factorization of certain displacement structured matrices, including Cauchy-like, Vandermonde-like, Toeplitz-like and Hankel-like matrices, hereby extending some earlier work for the QR-factorization of the Cauchy matrix. The algorithm employs a chasing scheme for the recursive construction of a diagonal plus semiseparable matrix of semiseparability rank r, where r is equal to the given displacement rank. The complexity is $O(r^2n^2)$ operations in the general case, and $O(rn^2)$ operations in the Toeplitz- and Hankel-like case, where n denotes the matrix size. Numerical experiments are provided.

Mathematics Subject Classification (2000). 65F18, 15A23, 15A03.

Keywords. Displacement structures, QR-factorization, lower semiseparable plus diagonal matrices, chasing procedure.

1. Introduction

In this paper we present a novel unified approach for computing the QR-factorization of displacement structured matrices, including some cases of main interest, namely Vandermonde-like, Toeplitz-like and Hankel-like matrices. Cauchy-like ma-

The research of the first and third author was partially supported by the Research Council K.U. Leuven, project OT/05/40 (Large rank structured matrix computations), Center of Excellence: Optimization in Engineering, by the Fund for Scientific Research – Flanders (Belgium), G.0455.0 (RHPH: Riemann-Hilbert problems, random matrices and Padé-Hermite approximation), G.0423.05 (RAM: Rational modelling: optimal conditioning and stable algorithms), and by the Belgian Programme on Interuniversity Poles of Attraction, initiated by the Belgian State, Prime Minister's Office for Science, Technology and Culture, project IUAP V-22 (Dynamical Systems and Control: Computation, Identification & Modelling). The scientific responsibility rests with the authors.
The first author is a Postdoctoral Fellow of the Fund for Scientific Research – Flanders (Belgium).
The work of the second author has been supported by MIUR under project 2004015437.

trices can also be treated in this framework, provided they have at least one set of data points lying on the unit circle of the complex plane $\mathbb{C}$. Generalizations of the algorithm to deal with Toeplitz-plus-Hankel-like matrices and Cauchy-like matrices with data points on the real axis $\mathbb{R}$ are also possible and will be discussed elsewhere.

In the literature, some algorithms for the QR-factorization of Toeplitz-like matrices are already known, see [1, 12] and the references therein. These methods typically compute the upper triangular factor by factoring the normal equation matrix via the generalized Schur algorithm. If the input matrix is ill-conditioned or exactly rank deficient, then the computed factorization of these algorithms can be very poorly accurate or even not an exact QR-factorization [12].

In contrast to these existing methods, the algorithms of the present paper proceed by translating the QR-factorization problem for a given displacement structured matrix into an equivalent chasing problem for an associated unitary rank structured matrix.

The approach followed in the present paper is motivated by the result in [7], where the QR-factorization of the classical Cauchy matrix with real data points was translated into a chasing problem for an associated Hermitian diagonal plus semiseparable matrix of semiseparability rank one. However, differently from [7] here we will focus on those displacement structures leading to *unitary* rank structured matrices, since this is the case for some of the more important classes such as Vandermonde-like, Toeplitz-like matrices and so on. Incidentally, let us point out that also transforms of a Toeplitz matrix into a Cauchy-like matrix with data points lying on the real line $\mathbb{R}$ are possible [9, 10, 11], but these will not be covered in the present paper.

The unitary rank structured matrices in this paper will be represented and manipulated as a product of Givens transformations [5, 8]. This representation leads to an asymptotically optimal number of $O(rn)$ parameters. Moreover, in contrast to the so-called sequentially semiseparable, quasiseparable, uv and Givens-weight representations used to represent rank structured matrices in the literature, the Givens product representation has the advantage that the unitarity of the matrix is an explicit part of the representation.

Chasing problems for tridiagonal and banded matrices are generally solved by using *bulge-chasing* techniques. The rationale for employing these techniques is that band matrices have a specified *sparsity pattern* and we can annihilate a bulge by moving it along this pattern downwards or upwards until it disappears by using a sequence of Givens rotations applied as a similarity transform. Rank structured matrices are generally dense so that the occurrence of a bulge in the rank structure is not visible. Our major contribution is to show that appropriate generalizations of bulge-chasing techniques still work for unitary rank structures represented in factored form thus leading to an efficient solution of the associated chasing problem. The cumulative similarity transformation returned as output by the chasing procedure allows to determine the Q-factor of the QR-factorization of the given displacement structured matrix.

The total complexity of this algorithm for QR-factorization is $O(r^2n^2)$ operations in the general case, and $O(rn^2)$ operations in the Toeplitz-like case, where r denotes the displacement rank, and where n denotes the matrix size. It is worth pointing out that the algorithm computes the QR-factorization of the given displacement structured matrix $A = QR$ in *full* form, i.e., the R-factor is computed as a full upper triangular matrix having $\frac{(n+1)n}{2}$ entries, while the Q-factor is computed to consist of a full product of $\frac{n(n-1)}{2}$ Givens transformations. Hence, the proposed scheme does *not* lead to a sparsity in the representation of the QR-factorization, but rather to an efficient way for computing these parameters.

The methods introduced in this paper are also strongly related to the chasing methods for diagonal plus semiseparable matrices in solving inverse eigenvalue problems [15]. More precisely,

- we present methods which can be literally used for solving some of the inverse eigenvalue problems too,
- we explain the mechanism behind a source of numerical instability, which may illuminate the numerical experiments for inverse eigenvalue problems as well.

The remainder of this paper is organized as follows. In Section 2 we establish basic notation and briefly recall some rank structure concepts. In Section 3 we describe the general scheme for the QR-factorization of a displacement structured matrix and its relation to a chasing problem for an associated lower semiseparable plus diagonal matrix. Sections 4, 5 and 6 specify this scheme to the case of Cauchy-like, Vandermonde-like and Toeplitz-like matrices. Finally, Section 7 reports on the results of some numerical experiments, and it discusses some interpretations of the observed numerical stability issues.

2. Rank Structure Preliminaries

Let us first define in more detail the rank structures arising from the methods of this paper.

Definition 1. *We define a* rank structure $\mathcal{R}$ *on* $\mathbb{C}^{n\times n}$ *as a collection of so-called structure blocks* $\mathcal{R} = \{\mathcal{B}_k\}_k$. *Each structure block* $\mathcal{B}_k$ *is characterized as a 4-tuple*

$$\mathcal{B}_k = (i_k, j_k, r_k, \lambda_k),$$

where i_k *is the row index,* j_k *the column index,* r_k *the rank upper bound and* $\lambda_k \in \mathbb{C}$ *is called the shift element. We say a matrix* $A \in \mathbb{C}^{n\times n}$ *to satisfy the rank structure* $\mathcal{R}$ *if for each* k,

$$\mathrm{Rank} A_k(i_k : n, 1 : j_k) \le r_k, \quad \text{where } A_k = A - \lambda_k I.$$

Thus after subtracting the shift element λ_k *from the diagonal entries, we must get a low rank block.*

As a special case, when a structure block $\mathcal{B}_k$ has shift element equal to zero, or when it is situated strictly below the main diagonal, then we call it pure. *We sometimes denote such a structure block by $\mathcal{B}_{\mathrm{pure},k}$.*

Figure 1 shows an example with two structure blocks.

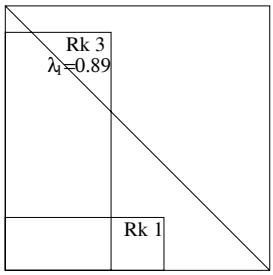


FIGURE 1. Example of a rank structure with two structure blocks. The left structure block $\mathcal{B}_1$ intersects the diagonal and has shift element $\lambda_1 = 0.89$, while the second structure block $\mathcal{B}_2$ is 'pure'. The notation 'Rk r' denotes that the structure block is of rank at most r.

As a special case, a matrix is called *lower semiseparable plus diagonal* of semiseparability rank r if it satisfies a set of structure blocks $\mathcal{B}_k : (i_k, j_k, r_k, \lambda_k) = (k, k, r, \lambda_k)$, $k = 1, \ldots, n$. Here the $\lambda_k \in \mathbb{C}$ are an integral part of the rank structure. We will sometimes refer to them as the *diagonal elements* of the rank structure. Another remarkable special structure is defined by $\mathcal{B}_{\mathrm{pure},k} : (i_k, j_k, r_k, \lambda_k) = (s + k, k, 0, 0)$, $k = 1, \ldots, n - s$. If $A \in \mathbb{C}^{n \times n}$ satisfies all the $\mathcal{B}_{\mathrm{pure},k}$ then A has lower bandwidth $s-1$. Therefore, rank structures are a generalization of *band structures*.

A unitary rank structured matrix admits a condensed representation as the product of Givens rotations. For computational purposes it is useful to describe pictorially this product by means of an associated *Givens product representation graph* [5]. The following example clarifies these concepts. Let $A \in \mathbb{C}^{6 \times 6}$ be a unitary lower semiseparable matrix of semiseparability rank 1, i.e., A satisfies a set of structure blocks $\mathcal{B}_k : (i_k, j_k, r_k, \lambda_k) = (k, k, 1, 0)$, $k = 1, \ldots, 6$. The QR-decomposition of A yields a condensed parametrization of the matrix as a product of Givens transformations $A = G_{5,6} G_{4,5} \cdots G_{1,2}$; here the notation $G_{i,i+1}$ denotes a Givens transformation, i.e., a unitary transformation which equals the identity matrix except for its entries in rows and columns $i, i + 1$. We represent the above matrix product by the following graph

Here each little line segment represents a Givens transformation $G_{i,i+1}$ 'acting' on the rows of an (invisible) matrix standing on the right of it. The row index i of the Givens transformation $G_{i,i+1}$ can be derived from the height at which the Givens transformation is standing in the figure, e.g., the top rightmost line segment in the above figure corresponds to $G_{1,2}$, and so on.

Now let us suppose that A is modified both in the third and in the fourth column in such a way that the perturbed matrix $\widetilde{A}$ is still unitary and, moreover, $G_{1,2}^H \cdots G_{5,6}^H \widetilde{A} = I_2 \oplus G'_{3,4} \oplus I_2$. The 'bulge 'in the rank structure of $\widetilde{A}$ is revealed by the associated graph

At the core of our proposed chasing methods for unitary rank structured matrices is the basic observation that, similarly as in the banded case, this bulge can be moved along the graph by a sequence of unitary similarity transformations determined by swapping techniques applied to short sequences of consecutive Givens rotations.

3. QR-factorization of displacement structured matrices

In this section we start with some general theory for the QR-factorization of displacement structured matrices. We will assume that $A \in \mathbb{C}^{n\times n}$ is an invertible *displacement structured matrix*, i.e., a matrix satisfying a *displacement equation* of the form

$$YA - AZ = \text{Rk } r, \tag{1}$$

where $Y, Z \in \mathbb{C}^{n\times n}$ are fixed coefficient matrices, and where the right-hand side Rk r has to satisfy the constraint that it is of rank at most r, where $r \in \mathbb{N}$ is the *displacement rank*. Examples of such displacement equations could be those defining Cauchy-like, Vandermonde-like, Toeplitz-like matrices and so on, each of them having a specific choice for the coefficient matrices Y and Z as well as for the displacement rank r. The treatment of these specific classes is deferred to the next sections.

Remark 2. *Some comments are in order concerning the assumption that A is invertible. This condition seems to preclude our method to be applied in many interesting cases where the matrix is rank-deficient or it is ill conditioned. However, in* [2] *it has been proved that the assumption can be removed by using a continuity argument and, indeed, our approach still works in the singular case. More precisely, since the QR decomposition of a singular matrix is not essentially unique, we can show that a QR decomposition of A exists such that the Q-factor is the solution of a chasing problem for an associated unitary rank structured matrix.*

Let us now assume a general displacement equation (1). Inserting the QR-factorization $A = QR$ into this equation leads to

$$Y(QR) - (QR)Z = \mathrm{Rk}\ r \tag{2}$$

$$\Leftrightarrow \quad (Q^H Y Q)R - RZ = \widetilde{\mathrm{Rk}}\ r \tag{3}$$

$$\Leftrightarrow \quad (Q^H Y Q) = RZR^{-1} + \widetilde{\widetilde{\mathrm{Rk}}}\ r, \tag{4}$$

where $\widetilde{\mathrm{Rk}}\ r$ and $\widetilde{\widetilde{\mathrm{Rk}}}\ r$ are new matrices of rank at most r. Note that we assumed here R^{-1} to exist.

Now we would like to use the above relations to obtain an alternative way of computing the QR-factorization $A = QR$. We will have to assume to this end that

- both Y and Z are upper triangular matrices. (5)

By the above assumption, the matrix RZR^{-1} occurring in (4) must also be an upper triangular matrix. A closer look reveals that it has diagonal elements precisely equal to those of Z: $z_{i,i}$, $i = 1, \ldots, n$.

Hence, the right-hand side of (4) will be a *lower semiseparable plus diagonal matrix* of semiseparability rank r, with diagonal elements of the structure precisely equal to the $z_{i,i}$.

In order to exploit this observation, let us partition

$$Y = \begin{bmatrix} y_{1,1} & Y_{1,2} \\ 0 & Y_{2,2} \end{bmatrix}, \quad Z = \begin{bmatrix} Z_{1,1} & Z_{1,2} \\ 0 & z_{2,2} \end{bmatrix}, \tag{6}$$

where $y_{1,1}$ and $z_{2,2}$ denote the top left and bottom right elements of Y and Z, respectively. We also partition

$$A = \begin{bmatrix} A_{1,1} & a_{1,2} \\ A_{2,1} & A_{2,2} \end{bmatrix}, \tag{7}$$

where $a_{1,2}$ is the top right element of A.

From the assumptions on the upper triangularity of Y and Z, it follows that the submatrix $A_{2,1}$ *inherits* the displacement structure from A. Indeed, this follows by evaluating the last block row and first block column of the displacement equation (1), and using (6), from which $Y_{2,2}A_{2,1} - A_{2,1}Z_{1,1} = (\mathrm{Rk}\ r)_{2,1}$. Thus the matrix $A_{2,1}$ is displacement structured too.

This suggests a recursive procedure: Assume that a QR-factorization has been computed for the displacement structured matrix $A_{2,1}$:

$$A_{2,1} = QR. \tag{8}$$

Clearly, given the knowledge of the QR-factorization (8), the *full* matrix A in (7) can be transformed into a Hessenberg matrix H, and hence its QR-factorization can be completed by applying a single downgoing sequence of Givens transformations to the rows.

On the other hand, we know that the QR-factorization (8) leads to an associated lower semiseparable plus diagonal equation (4), i.e.,

$$Q^H Y_{2,2} Q = S, \tag{9}$$

where

$$S := RZ_{1,1}R^{-1} + \mathrm{Rk}\ r \tag{10}$$

is already lower semiseparable plus diagonal of semiseparability rank r, with diagonal elements of the structure equal to $z_{i,i}$, $i = 1, \ldots, n-1$. By embedding this relation in a full n by n form, it follows that

$$\begin{bmatrix} 1 & 0 \\ 0 & Q^H \end{bmatrix} \begin{bmatrix} y_{1,1} & Y_{1,2} \\ 0 & Y_{2,2} \end{bmatrix} \begin{bmatrix} 1 & 0 \\ 0 & Q \end{bmatrix} = \begin{bmatrix} y_{1,1} & Y_{1,2}Q \\ 0 & S \end{bmatrix}. \tag{11}$$

We should stress that (11) has been obtained from the QR-factorization (8) of size $n-1$. But we would like now to bring it to the required form (4), induced by the QR-factorization of the *full* n by n matrix A. It was already mentioned before that this aim can be achieved by incorporating a downgoing sequence of Givens transformations $G_{n-1,n}^H, \ldots, G_{1,2}^H$ into the matrix Q^H. Applying these transformations to both unitary factors in the left-hand side of (11), and applying them also to the right-hand side of (11) transforms the latter to

$$\tilde{S} := G_{n-1,n}^H \cdots G_{1,2}^H \begin{bmatrix} y_{1,1} & Y_{1,2}Q \\ 0 & S \end{bmatrix} G_{1,2} \cdots G_{n-1,n}. \tag{12}$$

This relation (12) is the basis of the chasing procedure of this paper. It means that given the lower semiseparable plus diagonal matrix S of size $n-1$, we must apply a similarity operation with a downgoing sequence of Givens transformations, such that the resulting matrix $\tilde{S}$ in (12) is lower semiseparable plus diagonal of size n.

Under quite mild assumptions on the rank structure of S it can be proved [2] that the unitary Hessenberg matrix $G_{1,2} \cdots G_{n-1,n}$ satisfying (12) is essentially unique. Roughly speaking, this means that the Q-factor in the QR decomposition (8) of $A_{2,1}$ can be updated in an essentially unique way to obtain a solution Q of (4). Whence, this latter Q must be the Q-factor of a certain QR factorization of the full matrix A in (7).

The above description, and in particular (12), suggests that we could use a *structure-preserving chasing* procedure to determine the subsequent Givens transformations $G_{i,i+1}$, $i = 1, \ldots, n-1$.

In fact, one could retrieve more details about the intended chasing procedure, by comparing the position of the diagonal elements of the structure $z_{i,i}$ before and after the chasing, cf. (10). This reveals that each of the original diagonal elements of the structure $z_{i,i}$, $i = 1, \ldots, n-1$ should be chased one position upwards, while a new diagonal element $z_{n,n}$ is 'installed' at the bottom of the matrix (12).

To manipulate the lower diagonal plus semiseparable matrix (12) efficiently during the chasing scheme, we would like that not only the lower, but also the

upper triangular part of this matrix has bounded ranks. We will realize this by assuming that

- Y is a unitary/Hermitian plus low rank matrix. (13)

Indeed, it is known that this type of structure is preserved under unitary similarity transformations, and moreover, that the rank structure in the lower triangular part of such matrices, induces rank structure in the upper triangular part as well. Applying this to the lower semiseparable plus diagonal structure of the matrix S in (9) shows that this matrix must be *upper* semiseparable plus diagonal as well, or *semiseparable plus diagonal* for short. This will allow the use of efficient representations to represent the matrix S during the chasing scheme, such as sequentially semiseparable, quasiseparable, uv, or Givens-weight representations, see, e.g., [6, 3]. The best choice of representation depends on many factors. For example, for each of the practical examples of the next sections we will have that the matrix Y, and hence the matrix S in (9), are purely unitary matrices, and consequently we will make use there of Givens product representations [5].

Incidentally, note that for the special case where Y in (13) is *exactly* unitary or Hermitian, combining this with the upper triangularity of Y reveals that Y must be diagonal. However, note that the diagonality of Y does not imply the diagonality of the matrix S in (9), since the latter is semiseparable plus diagonal with semiseparability rank determined by the displacement rank, cf. (10).

One element is still missing in the above scheme. Namely, there is the fact that the structure-chasing procedure will unavoidably break down for the determination of the first r, as well as for the final r Givens transformations $G_{i,i+1}$ in (12), since the corresponding Rk r structure blocks are 'trivially satisfied', and hence preserved by *any* Givens transformation $G_{i,i+1}$. Hence, we should be able to find these Givens transformations in a direct way, using the fact that they belong to the QR-factorization of A.

To this end, recall the Hessenberg matrix

$$H = \begin{bmatrix} A_{1,1} & a_{1,2} \\ R & Q^H A_{2,2} \end{bmatrix} \tag{14}$$

into which the matrix A has already been brought by means of the unitary matrix Q^H of size $n-1$, where $R := Q^H A_{2,1}$ is already upper triangular. We should then be able to derive the first and last r columns of this Hessenberg matrix, since the action of making these columns upper triangular will reveal the required Givens transformations $G_{i,i+1}$ needed for doing this.

Let us assume by induction that we have already knowledge of the first and last r columns of the *submatrix* R in (14). It will clearly suffice if we can determine the required elements of the first row and last column of the matrix H in (14).

Let us begin with the first row of H. It is clear from (14) that this row is equal to that of the original displacement structured matrix A. Due to the supposed upper triangularity of Y and Z, we will be able to derive these entries by means of the original displacement equation (1) which the matrix A has to satisfy, which

we restate here for convenience:

$$YA - AZ = \text{Rk } r. \tag{15}$$

Indeed, evaluating the last row and first column of (15) reveals that $(y_{n,n} - z_{1,1})a_{n,1} = (\text{Rk } r)_{n,1}$, from which $a_{n,1}$ can be derived. One can then gradually proceed towards the top right direction of the matrix A, using a recursive scheme to determine the subsequent entries of this matrix. Note that we have to assume here that the displacement operator is invertible, i.e., that $y_{i,i} \neq z_{j,j}$ for all i and j.

It is clear that the complexity of determining all entries of the matrix A in this way will require $O(n^3)$ operations in the general case, which we consider to be unacceptably expensive. In order to obtain a better complexity, we assume from now on that, in addition to being upper triangular, we have that

- both Y and Z have upper bandwidth at most b, with $b \leq r$.

Using this assumption, the complexity of determining all entries of A from its displacement equation is easily seen to reduce to $O(b^2n^2)$ operations. We remark that we need each time only those entries in the first and last r columns of the top row of (14), but in the process of determining the latter, the computation of the elements in the other columns seems to be an unavoidable step. determined one by one in a recursive way. (Exception to this is when the bandwidth $b = 0$, thus in case where Y and Z are diagonal matrices, in which case each entry of the displacement structured matrix A can be determined independently.) We mention also that the condition that Y and Z have limited bandwidth can be replaced by the more general condition that they are *rank structured*, but we will not go further into this.

Now we consider the computation of the last column of H in (14). To this end, let us characterize this Hessenberg matrix in an alternative way. From the decomposition $A = QH$ (we write here Q instead of $1 \oplus Q$ by abuse of notation), one can derive in an analogous way to (2), (3) that

$$\begin{aligned} & Y(QH) - (QH)Z = \text{Rk } r \\ \Leftrightarrow \quad & (Q^H Y Q)H - HZ = \widetilde{\text{Rk}}\, r, \end{aligned}$$

where $\widetilde{\text{Rk}}\, r$ is a new matrix of rank at most r. Evaluating the last column of this last equation leads to

$$[(Q^H Y Q) - z_{n,n}I]H_{\text{col } n} = \sum_{k=n-b}^{n-1} z_{k,n} H_{\text{col } k} + (\widetilde{\text{Rk}}\, r)_{\text{col } n}, \tag{16}$$

where the subscript col k is used to denote the kth column of a matrix, and where b denotes again the supposed bandwidth of the coefficient matrix Z.

This equation can be used to solve for the last column $H_{\text{col } n}$ in terms of the previous columns of H. Note that the coefficient matrix $Q^H Y Q$ occurring in (16) is nothing but the lower diagonal plus semiseparable matrix S used during the chasing procedure, cf. (9). Since this matrix is rank structured, it will be possible to solve (16) in an efficient way; see further.

Summarized, we have now specified how to determine the entries of the first and last r columns of the matrix H in (14), whose information is important for determining the first r and last r Givens transformations $G_{i,i+1}$ in the chasing scheme of (12). This ends the description of one inductive step $n-1 \mapsto n$ in the QR-factorization of the displacement structured matrix A.

At the end of performing all these inductive steps $k-1 \mapsto k$, $k = 2, \ldots, n$, we will finally have obtained the Givens transformations constituting the Q-factor of the QR-factorization $A = QR$.

To end the description of the method, we should still specify how to compute the subsequent columns of the R-factor of the QR-factorization. But this can be easily done by solving again the subsequent columns of the relation (3). Note that doing this for *all* columns of R will lead to the same order of complexity as computing *all* the column vectors $H_{\text{col } n}$, during each of the inductive steps $n-1 \mapsto n$, as described above.

We point out that the complexity of computing all the columns of the R-factor of the QR-factorization, and hence the complexity of computing each of the column vectors $H_{\text{col } n}$ during the algorithm, typically requires $O(r^2n^2)$ operations, and $O(rn^2)$ operations in the Toeplitz-like case. Hence, these computations typically form the *dominant* term in the total complexity of the QR-factorization, since the chasing steps usually require only $O(rn^2)$ operations.

In the next 3 sections we will show how the general scheme for the QR-factorization of displacement structured matrices described in the current section can be specified to the displacement structures of most interest.

4. The Cauchy-like case

In this section we describe the QR-factorization in case where $A \in \mathbb{C}^{n\times n}$ is a Cauchy-like matrix:

$$D_y A - A D_z = \text{Rk } r, \tag{17}$$

where the coefficient matrices D_y, $D_z \in \mathbb{C}^{n\times n}$ are diagonal matrices whose diagonal entries $y_{i,i}$, $z_{j,j}$ are called *data points*, with $y_{i,i} \neq z_{j,j}$ for all i, j, and where the displacement rank r is supposed to be known; note that $r = 1$ in case of a classical *Cauchy matrix* $A = \left[\dfrac{1}{y_{i,i} - z_{j,j}}\right]_{i,j=1,\ldots,n}$.

It follows from the general scheme of Section 3 that the matrices S in (10) will now be lower diagonal plus semiseparable matrices, with diagonal elements of the structure given by the $z_{i,i}$. In fact, for reasons that will become clear soon, we will prefer to keep these diagonal elements out of the representation, and to work instead with the induced *pure* rank structure of S, which is situated just below the main diagonal: see Figure 2.

For the scheme to be described in this section to be of practical interest, we will need the following specification of (13):

- D_y is unitary.

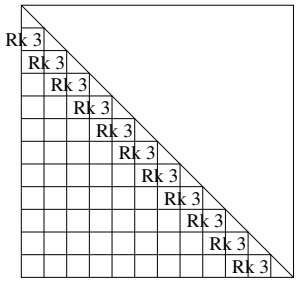


FIGURE 2. Induced pure rank structure for a lower diagonal plus semiseparable matrix S of semiseparability rank $r = 3$.

This unitarity will imply the matrices S in (9) to be unitary as well. Thus, the chasing techniques of this section will be expressed in terms of unitary diagonal plus semiseparable matrices.

We turn now to the practical representation and manipulation of the unitary diagonal plus semiseparable matrix S of semiseparability rank r. From Section 2 this can be achieved by means of a suitable *Givens product representation graph* [5]. In the present case, the representation takes the form of Figure 3[1].

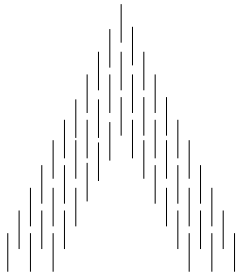


FIGURE 3. Givens product representation graph for a unitary diagonal plus semiseparable matrix S of semiseparability rank $r = 3$.

To explain the specific pattern formed by the Givens transformations in Figure 3, it is useful to think in terms of the QR-factorization $S = QR$ of the given unitary diagonal plus semiseparable matrix S of semiseparability rank r. Due to the unitarity of S, the R-factor of its QR-factorization can be chosen to be the identity matrix. The above QR-factorization is then equivalent to the equation

$$Q^H S = I,$$

[1]The specific pattern formed by the topmost Givens transformations in Figure 3 corresponds to the so-called *zero-creating* variant of the Givens product representation, as introduced in [5].

the identity matrix. Thus we should apply to the rows of S a set of unitary operations Q^H in order to eliminate all entries in the lower triangular part of S. Due to the rank structure of S, this process can be applied efficiently via a two step procedure. The first step in this process is to *peel off* all Givens transformations in the left branch of Figure 3, i.e., to multiply S with the Hermitian transposes of these Givens transformations to the rows, going from bottom to top, and hereby compressing the structure blocks of the given rank structured matrix S into blocks of zeros. Incidentally, this reveals that the width of the left branch of the Givens product representation directly reflects the underlying ranks of the rank structure, which is $r = 3$ in the case of Figure 3.

What remains after peeling off the Givens transformations of the left branch of the representation will be a Hessenberg matrix with three subdiagonals, which is nothing but the right branch of the representation of Figure 3. The fact that the right branch has the same width as the left branch, indicates that the structure is situated here just below the main diagonal. For more details about the representation indicated in Figure 3 we refer to [5].

With the aim of manipulating the Givens product representation, we recall the following result.

Lemma 3 (Pull-through lemma). *Given a unitary* 3 *by* 3 *matrix* Q *which is factorized as*

$$Q = G'_{1,2} G_{2,3} G_{1,2},$$

then there exists a refactorization

$$Q = \tilde{G}'_{2,3} \tilde{G}_{1,2} \tilde{G}_{2,3}.$$

See Figure 4.

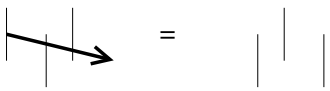


FIGURE 4. Pull-through lemma applied in the downward direction. One could imagine that the leftmost Givens transformation is really 'pulled through' the two rightmost Givens transformations.

We will now use the above representation for the unitary diagonal plus semiseparable matrix S to determine the subsequent Givens transformations $G_{i,i+1}$ used in the chasing procedure (12). First we will show how to determine the Givens transformations $G_{i,i+1}$ with $i = r+1, \ldots, n-r-1$; the determination of the first and last r Givens transformations $G_{i,i+1}$ will be discussed later in this section. From the general scheme of Section 3, we know that the similarity transform with $G_{i,i+1}$, $G_{i,i+1}^H$ has to preserve the diagonal plus semiseparable structure, while each of the diagonal elements of the structure $z_{i,i}$ is chased one position upwards by the chasing procedure.

We will neglect here the information about the diagonal elements $z_{i,i}$, and put our focus only on the induced *pure* structure, situated just below the main diagonal, since the latter structure can be read off directly from the sparsity pattern of the representation in Figure 3, as explained above. Note that this induced pure structure must be preserved during *each* step of the algorithm.

We will use this latter observation to derive the required Givens transformations. Suppose that we are in the ith step of the chasing process, $i = r+1, \ldots, n-r-1$, and that we have to determine a Givens transformation $G := G_{i,i+1}$ such that a similarity transformation with G preserves the pure induced rank structure of the unitary diagonal plus semiseparable matrix S.

This condition can be ensured by requiring that, after pulling G downwards through the left branch, and G^H downwards through the right branch of the representation, by using each time r subsequent applications of the pull-through lemma, then the pulled through Givens transformations must precisely annihilate each other. Stated another way, the pulled through versions of G and G^H must be related to each other as H and H^H, for a suitable Givens transformation H: see Figure 5.

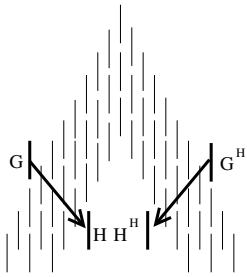


FIGURE 5. Elementary chasing step for the unitary diagonal plus semiseparable matrix S of semiseparability rank $r = 3$. The preservation of structure is ensured when the pulled through versions of G and G^H are related to each other as H and H^H, hence cancelling each other out in the middle of the figure.

We want now to exploit this observation to obtain a practical scheme for determining G. To this end, we will first investigate more closely the leftmost of the two pull-through relations between G and H in Figure 5: see Figure 6.

By expanding the middle Givens product on the left-hand side of Figure 6 into its full matrix form, it turns out that this relation can be restated in terms of the lower triangularity of a certain 2 by 2 matrix A being preserved by multiplying with G on the left and with H^H on the right: See Figure 7.

In a similar way, one can consider also the *rightmost* of the two pull-through relations between G and H in Figure 5. Taking the Hermitian transpose of this relation will lead in exactly the same way to a 2 by 2 lower triangular matrix B,

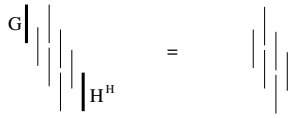


FIGURE 6. The figure expresses the fact that the pulled-through version of G is precisely H, and hence can be removed by multiplying with H^H on the right. This guarantees the existence of a refactorization having the shape shown in the right part of the figure.

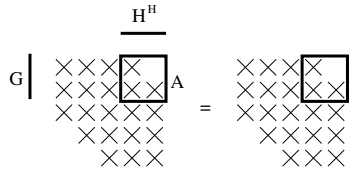


FIGURE 7. Equivalent formulation of Figure 6 in its full matrix form.

whose lower triangularity must be preserved by multiplying with G on the left and with H^H on the right.

The determination of the required Givens transformations G and H reduces then to the following problem.

Problem 4. *Given two lower triangular matrices*

$$A = \begin{bmatrix} \times & 0 \\ \times & \times \end{bmatrix}, \quad B = \begin{bmatrix} \times & 0 \\ \times & \times \end{bmatrix}.$$

We are interested in finding two non-trivial Givens transformations G and H such that the lower triangularity of A, B is preserved in

$$GAH^H = \begin{bmatrix} \times & 0 \\ \times & \times \end{bmatrix}, \quad GBH^H = \begin{bmatrix} \times & 0 \\ \times & \times \end{bmatrix},$$

where the generically nonzero values are denoted as $\times$.

It is clear that there always exists a trivial solution $(G, H) = (I_2, I_2)$ to this problem, where I_2 is the identity matrix of size 2. But we will be interested in the non-trivial solution.

It turns out that this non-trivial solution can be easily obtained. First we are going to determine G. We will search it in the form

$$G = \begin{bmatrix} c & s \\ \times & \times \end{bmatrix},$$

where c and s are unknown complex numbers such that $|c|^2 + |s|^2 = 1$, and where the values that will be irrelevant in the computation are denoted as $\times$. We compute

$$GA = \begin{bmatrix} ca_{1,1} + sa_{2,1} & sa_{2,2} \\ \times & \times \end{bmatrix}, \quad GB = \begin{bmatrix} cb_{1,1} + sb_{2,1} & sb_{2,2} \\ \times & \times \end{bmatrix}. \tag{18}$$

Now it is clear that the existence of the column operation H^H occurring in Problem 4, is equivalent to the condition

$$\begin{vmatrix} ca_{1,1} + sa_{2,1} & sa_{2,2} \\ cb_{1,1} + sb_{2,1} & sb_{2,2} \end{vmatrix} = 0.$$

A trivial case where this equation is satisfied is when $s = 0$. This corresponds to the trivial solution $(G, H) = (I_2, I_2)$ mentioned before.

But recall that we were interested in the non-trivial solution. This non-trivial solution can be obtained by skipping the factor s in the second column of the above determinant, and expanding the remaining part of this determinant as

$$(a_{1,1}b_{2,2} - a_{2,2}b_{1,1})c = (a_{2,2}b_{2,1} - a_{2,1}b_{2,2})s,$$

or alternatively in matrix-vector form as

$$G \begin{bmatrix} a_{1,1}b_{2,2} - a_{2,2}b_{1,1} \\ a_{2,2}b_{2,1} - a_{2,1}b_{2,2} \end{bmatrix} = \begin{bmatrix} 0 \\ \times \end{bmatrix}. \tag{19}$$

It is clear that, provided at least one of the two components of the vector on the left-hand side of (19) is non-zero, then the required Givens transformation G is essentially uniquely determined from this equation.

Given the knowledge of G, one can now determine H as well. Indeed, we know that H should be chosen to restore the lower triangularity of each of the two matrices in (18), i.e., to create a zero in the $(1, 2)$ position of *both* matrices. We point out that in principle, it does not matter which of these two matrices is taken to determine H, but for stability reasons, one should use the one whose top row has the largest norm, or alternatively a least squares variant containing information of both top rows.

We should still clarify one point here. Suppose that we insert the computed value of G in the left part of Figure 6, and subsequently pull it through, so that this Givens transformation is chased to the bottom right. Will it then be guaranteed that we end up with (essentially) the computed value of H at the end of this pull-through process, so that the factor H^H on the left of Figure 6 can be cancelled, ending up with the Givens pattern on the right of Figure 6?

Maybe surprisingly, the answer to this question is *no*, since the pull-through operation may be badly determined. This means that in certain cases, the pull-through of two Givens transformations which are equal up to the machine precision $\epsilon \approx 10^{-16}$ may lead to relatively large differences in the pulled through Givens transformations. Stated in another way, inserting G does not always guarantee that (essentially) the computed H will come out at the other side.

The solution to this problem is to use the information of *both* Givens transformation G and H, and to perform an explicit refactorization procedure of the

unitary matrix Q in the left-hand side of Figure 6. In other words, we are given the matrix Q and our aim is to refactorize this matrix into a new Givens product having the form of the (unknown) right-hand side of Figure 6.

Let us assume that $r = 3$, and let us denote the required refactorization as

$$Q = \tilde{G}_{2,3}\tilde{\tilde{G}}_{1,2}\tilde{G}_{3,4}\tilde{\tilde{G}}_{2,3}\tilde{G}_{4,5}\tilde{\tilde{G}}_{3,4}, \tag{20}$$

where we have to search each of the Givens transformations $\tilde{G}_{i+1,i+2}$ and $\tilde{\tilde{G}}_{i,i+1}$, $i = 1, \ldots, 3$; cf. Figure 6.

To obtain the refactorization (20) in a stable way, we will determine the subsequent Givens transformations $\tilde{G}_{i+1,i+2}$ and $\tilde{\tilde{G}}_{i,i+1}$, $i = 1, \ldots, 3$ as illustrated in Figures 8 and 9.

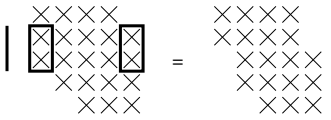


FIGURE 8. Determine the first Givens transformation $\tilde{G}^H_{2,3}$ from a least squares condition.

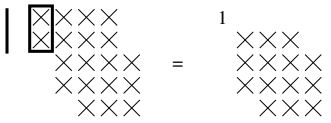


FIGURE 9. Determine the second Givens transformation $\tilde{\tilde{G}}^H_{1,2}$.

Let us comment on these figures. The first step is to note that the multiplication of the matrix Q on the left with the Givens transformation $\tilde{G}^H_{2,3}$ should create a zero in the $(3, 1)$ element as well as in the $(2, 5)$ element of Q, i.e.,

$$\tilde{G}^H_{2,3}\begin{bmatrix} q_{2,1} & q_{2,5} \\ q_{3,1} & q_{3,5} \end{bmatrix} = \begin{bmatrix} \times & 0 \\ 0 & \times \end{bmatrix}, \tag{21}$$

see Figure 8. In practice, it is advisable to determine $\tilde{G}^H_{2,3}$ by using a pivoting or least-squares strategy based on these two requirements. Let us then update $Q := \tilde{G}^H_{2,3}Q$.

The second step is to note that the multiplication of Q on the left with the Givens transformation $\tilde{\tilde{G}}^H_{1,2}$ should create a zero in the $(2, 1)$ element of Q, i.e.,

$$\tilde{\tilde{G}}^H_{1,2}\begin{bmatrix} q_{1,1} \\ q_{2,1} \end{bmatrix} = \begin{bmatrix} \times \\ 0 \end{bmatrix}, \tag{22}$$

see Figure 9. Note that we can obviously determine $\tilde{\tilde{G}}_{1,2}^H$ from this equation. We can then update $Q := \tilde{\tilde{G}}_{1,2}^H Q$.

The first column of Q has then been brought completely into upper triangular form and hence, due to the unitarity, also the first row of this matrix will vanish: see the right-hand side of Figure 9. Since the situation at the end of Figure 9 is similar to the one we started from in Figure 8, but with smaller dimension of the problem, the determination of the next Givens transformations of the refactorization (20) is not shown anymore.

A suited use of the pull-through lemma enables the above refactorization scheme to be carried out in $O(r)$ operations, rather than the $O(r^2)$ suggested in the above figures. The clue to this speed-up is to represent the matrix Q by its Givens product representation rather than expanding it in full form. The initial Givens product representation of Q is provided by the left-hand side of Figure 6. For each of the updates $Q := \tilde{G}_{i+1,i+2}^H Q$ and $Q := \tilde{\tilde{G}}_{i,i+1}^H Q$, we then update the Givens product representation of Q by applying the pull-through lemma maximally two times in the upward direction; note that this requires only $O(1)$ operations. Moreover, it turns out that in this way, the required elements $q_{2,1}, q_{3,1}$ in (21) and $q_{1,1}, q_{2,1}$ in (22) can be computed from the Givens product representation of Q using only $O(1)$ operations. Applying this method for all the Givens transformations $\tilde{G}_{i+1,i+2}$ and $\tilde{\tilde{G}}_{i,i+1}$, $i = 1, \ldots, r$, reveals that the total complexity will be $O(r)$ operations.

To complete the above description we should then show that also the elements $q_{2,r+2}, q_{3,r+2}, \ldots$ in (21) during this and the next steps of the refactorization process can be computed in $O(r)$ operations. But since these elements all belong to the rightmost column of Q, this can be realized by just precomputing the latter column in its full form, and next updating it during the refactorization process. It is easy to see that the total cost of this process is $O(r)$ operations as well.

Summarized, we have described now how to implement the basic chasing step in Figure 5 in a practical way, using not more than $O(r)$ operations. Applying this scheme for $i = r + 1, \ldots, n - r - 1$, we can obtain the subsequent Givens transformations $G := G_{i,i+1}$ of (12) using a total number of $O(rn)$ operations.

To complete the description of the QR-factorization in the Cauchy-like case, we should still explain how to determine the first and last r Givens transformations $G_{i,i+1}$ in (12), and how to update the representation under the action of these operations.

First, let us consider the computation of the first r Givens transformations $G_{i,i+1}$, $i = 1, \ldots, r$. This can be done by just specializing the general scheme of Section 3 in a straightforward way. Hence, it will suffice if we can show how the representation of the matrix S should be updated under the influence of a similarity transform with these operations. This is done in Figure 10.

Let us comment on this figure. First, we embed the given Givens product representation of size $n-1$, in a larger n by n representation. This can be achieved by pulling the two original branches apart, as done in Figure 10(a), and inserting

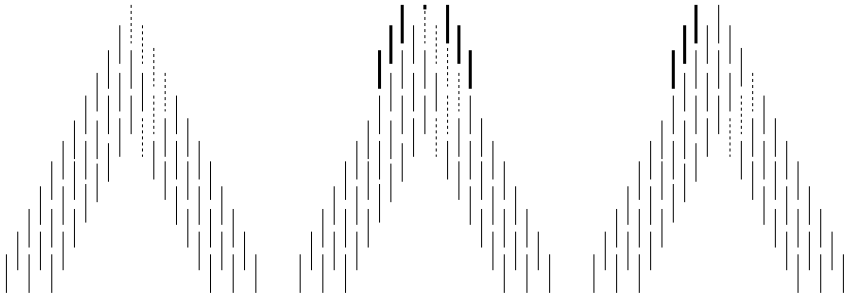


FIGURE 10. Absorbing the first Givens transformations into the representation.

some identity Givens transformations between them, as indicated by the dashed line segments in Figure 10(a).

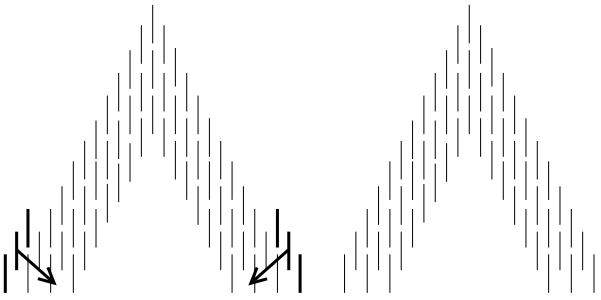
Next, we multiply with the next unitary diagonal element $y_{1,1}$, indicated by the fat dot, and with the first r Givens transformations $G_{i,i+1}$ on the left, $G_{i,i+1}^H$ on the right, indicated by the fat line segments in Figure 10(b).

One can now immediately absorb the rightmost added Givens transformations into the representation, and place them on the position of the dashed lines, without any actual computation: see Figure 10(c). Having done this, one should apply the pull-through lemma $r \times r = r^2$ in the downward direction to bring the leftmost added Givens transformations downwards, so that they ultimately appear on the positions of the remaining set of dashed lines in Figure 10(c). Clearly, the complexity of this entire process is only $O(r^2)$ operations, which can be considered as negligible w.r.t. the other parts of the algorithm. Having done these operations, the Givens product representation will be brought back into its usual form, as in Figure 3.

We move now to the absorption of the last r Givens transformations $G_{i,i+1}$, $G_{i,i+1}^H$ into the representation, $i = n-r, \ldots, n-1$. This can be done even simpler, using straightforward pull-through operations as indicated in Figure 11.

We wish to point out here the striking fact that absorbing the first and last r Givens transformations into the representation both requires about $O(r^2)$ operations, and essentially the same number of pull-through applications.

Finally, to conclude the description of the QR-factorization algorithm in the Cauchy-like case, we should still explain how each time the linear system is solved to compute the last column of the Hessenberg matrix H, as in (16). Since this linear system contains as coefficient matrix $S - z_{n,n}I = Q^H D_y Q - z_{n,n}I$, with in general $z_{n,n} \neq 0$, one can *not* proceed here with the Givens product representation anymore. Instead, we suggest to transform the Givens product representation first into a *Givens-weight representation*, using the $O(r^2 n)$ transformation algorithm



(a) (b)

FIGURE 11. Absorbing the final Givens transformations into the representation.

described in [5]. Once the Givens-weight representation has been computed, the diagonal correction term $-z_n I$ is easily added to the matrix, and one can then proceed by the standard Givens-weight solvers, including also $O(r^2 n)$ operations [4].

5. The Vandermonde-like case

In this section we briefly consider the QR-factorization in the case of a Vandermonde-like matrix A. Denoting the *upshift matrix*

$$Z_\alpha = \begin{bmatrix} 0 & I_{n-1} \\ \alpha & 0 \end{bmatrix}, \tag{23}$$

for any $\alpha \in \mathbb{C}$, and with I_{n-1} the identity matrix of size $n-1$, we will work with the Vandermonde-like matrices described by the displacement equation

$$Z_1 A - A D_z = \text{Rk } r, \tag{24}$$

where Z_1 is the *circulant* upshift matrix, having $\alpha = 1$, and where $D_z \in \mathbb{C}^{n \times n}$ is a diagonal matrix. We also assume that the displacement operator (24) is invertible and, therefore, that the spectra of Z_1 and D_z do not intersect each other. The diagonal elements z_i are sometimes called the *interpolation nodes* of the Vandermonde-like matrix. The displacement rank r is assumed to be known; note that $r = 1$ in case of a classical (transposed) *Vandermonde matrix* $A = [z_j^i]_{i,j=0,\ldots,n-1}$.

It should be observed that the displacement equation (24) does not allow the general scheme for QR-factorization in Section 3 to be applied, since the first coefficient matrix of the displacement equation violates the requirement of being upper triangular.

The solution consists in working instead with the matrix $\tilde{A} := F^H A$, where F is the Fourier matrix. From the well-known spectral decomposition $Z_1 = F D_y F^H$,

see, e.g., [13], the equation (24) transforms then to

$$D_y\tilde{A} - \tilde{A}D_z = \widetilde{\mathrm{Rk}}\ r, \tag{25}$$

where D_y is now a unitary diagonal matrix containing roots of unity. Clearly, the problem of QR-factorization has been reduced now to the Cauchy-like case described in the previous section. The total complexity of the QR-factorization remains $O(r^2n)$ operations.

6. The Toeplitz-like case

In this section we consider the case where A has Toeplitz-like structure. A first idea could be to use the displacement equation

$$Z_1A - AZ_0 = \mathrm{Rk}\ r,$$

where we use the notations of (23) above. The displacement rank r is again assumed to be known; note that $r = 2$ in case of a classical *Toeplitz matrix* $A = [t_{i-j}]_{i,j=1,\ldots,n}$.

Using again a premultiplication with the Fourier matrix F, the above problem of QR-factorization can be reduced immediately to the form

$$D_y\tilde{A} - \tilde{A}Z_0 = \widetilde{\mathrm{Rk}}\ r,$$

where D_y is again a unitary diagonal matrix containing roots of unity, and where we defined the updated matrix $\tilde{A} := F^HA$.

It can be observed now that in this case, the coefficient matrix $Z := Z_0$ is *strictly* upper triangular, i.e., $z_{i,i} = 0$ for each i. This has two important consequences:

- also the main diagonal is involved into the lower semiseparable structure;
- the coefficient matrix in (16) is the matrix Q^HYQ itself, i.e., there is no diagonal correction which has to be added to this coefficient matrix.

This last observation will allow for an $O(rn^2)$ solution of the QR-factorization problem; this should be compared to the $O(r^2n^2)$ solution in case of Cauchy-like matrices, see further.

Let us illustrate here the corresponding unitary rank structured matrix S: see Figure 12.

The difference for the corresponding unitary representation is now that the left branch of the representation of the unitary rank structured matrix S is slightly thicker than the right one, by the fact that the diagonal is part of the rank structure. This has the important consequence that the left and right Givens transformations G and G^H with $G := G_{i,i+1}$ do not influence the same structure block anymore, i.e., the algorithm has obtained a well-determined *flow direction*. In practice, this amounts to the fact that the Givens transformation $G := G_{i,i+1}$, after pulling it through the representation, arrives on the right of the right branch of the representation, one position lower than it was originally standing. This pulled-through

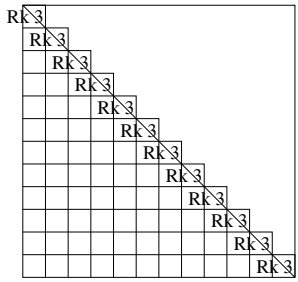


FIGURE 12. Rank structure in the Toeplitz-like case; we have here $r = 3$.

Givens transformation determines then immediately the Givens transformation $G_{i+1,i+2}$ to be applied during the next step, and so on. See Figure 13.

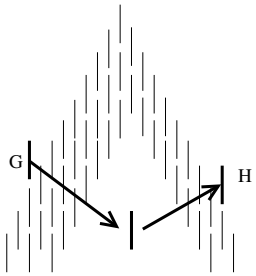


FIGURE 13. The figure shows the ith chasing step in the Toeplitz-like case, $i \in \{3, \ldots, n-4\}$. Note that the original Givens transformation $G = G_{i,i+1}$ is pulled through to the right, leading to a new Givens transformation $H = G_{i+1,i+2}$. The latter serves then as input for the $(i+1)$th chasing step. In the figure we assume that $r = 3$.

Applying the above scheme for $i = r, \ldots, n-r-1$, one can obtain the subsequent Givens transformations $G_{i,i+1}$ in (12).

We should then show how to obtain the first r and last $r-1$ Givens transformations $G_{i,i+1}$. This can be done by just specializing the general scheme of Section 3 in a straightforward way. Hence, it will suffice if we can show how the representation of the matrix S should be updated under the action of the first $r-1$ operations on the left and the first r operations on the right. But this is completely similar to the Cauchy-like case: see Figure 14.

Concerning this figure, note that the only difference w.r.t. the Cauchy-like case is that the symmetry between the number of Givens transformations on the rows and columns has somewhat changed, due to the asymmetry of the two branches of the representation. In case of Figure 14, this implies the absorption

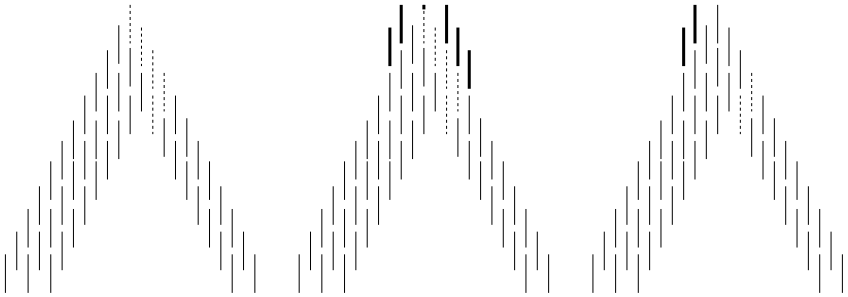


FIGURE 14. Absorbing the first Givens transformations into the representation in the Toeplitz-like case.

of the first three Givens transformations on the right, but only of the first *two* Givens transformations on the left. Indeed, this reflects the fact that starting from the third Givens transformation $G_{3,4}$ on the left, one can start with the chasing scheme of Figure 13.

In a similar way, one can treat the absorption of the last r Givens transformations on the rows and the last $r-1$ Givens transformations on the columns, where now also the symmetry between the number of Givens transformations on the rows and columns has changed, in an obvious way. The details are straightforward.

7. Numerical experiments

In this section we discuss the practical behavior of the proposed algorithms for the recursive QR-factorization of displacement structured matrices and report on the results of some numerical experiments. The algorithm for the Cauchy-like case described in Section 4 has been implemented in Matlab and then tested on both random and specific input matrices. In each experiment we measured the accuracy of the computed QR factorization of the input matrix C by evaluating the errors $||Q \cdot Q^H - I||_F$, $||C - Q \cdot R||_F$ and $||\mathrm{tril}(Q^H \cdot C, -1)||_F$, where $\mathrm{tril}(A, -1)$ denotes the strictly lower triangular part of the matrix A and $||\cdot||_F$ is the Frobenius matrix norm. Our program also returns as output the estimates of the conditioning of C and of the maximal conditioning of the lower-left submatrices of C employed in the recursive factorization process.

In our first test session the input matrices were generated as follows. We started from a Toeplitz matrix $T \in \mathbb{C}^{n\times n}$ determined by the entries a_j, $j = 1, \ldots, 2n-1$, with a_j, $j = 1, \ldots, n$ corresponding to the first row and a_j, $j = n, \ldots, 2n-1$ corresponding to the first column of T. Applying a similarity transform $C := FTF^H$, with F the Fourier matrix of size n, transforms the Toeplitz matrix T into a Cauchy-like matrix C satisfying $D_y C - C D_z = \mathrm{Rk}\ 2$, where D_y and D_z contain interlaced roots of unity; this is the same construction as we used in

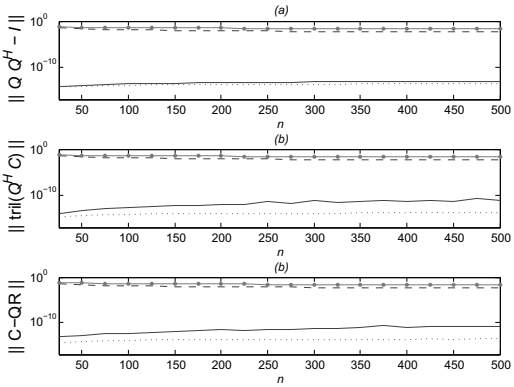


FIGURE 15. Average accuracy for subsequent values of n in the Cauchy-like case. The used norm is the Frobenius norm.

[2]. This Cauchy-like matrix C was then used as input for the QR-factorization algorithm. Recall that our scheme computes the Q-factor of the QR-factorization as a product of $\frac{n(n-1)}{2}$ Givens transformations $G_{i,i+1}$, while the R-factor is computed column by column by means of (3). The results are shown in Figure 15.

Let us comment on these figures. The figures were constructed from a Cauchy-like matrix $C := FTF^H$ where T is a Toeplitz matrix with uniformly randomly generated entries a_j on the interval $[0, 1]$, $j = 1, \ldots, 2n-1$. The lower lines show the accuracy of our method (continuous line) compared with the accuracy of the standard QR Matlab routine (dot line). The upper lines show the reciprocal of the condition number and the reciprocal of the maximal condition number of the lower-left submatrices. It can be noticed that both C and its relevant submatrices are fairly well conditioned and the fast algorithm behaves quite well even if it is probably not backward stable.

To investigate better the role of the conditioning of the matrices involved for the accuracy of computed results, in the second test session we considered the case where T equals the prolate matrix with parameter $\alpha \in \{0.45, 0.48, 0.6\}$, i.e., the matrix having $a_j = \sin(2\pi\alpha(n-j))/(\pi(n-j))$ and $a_{n+j} = \sin(-2\pi\alpha j)/(-\pi j)$ for $j = 1, \ldots, n-1$, and $a_n = 2 \times \alpha$.

It turned out that now the accuracy of our algorithm was substantially worse, with even not a single significant digit left for $\alpha = 0.6$ and $n = 400$. The reason was a catastrophic rank-one (instead of rank-two) submatrix detected in the very first step of the chasing procedure making the corresponding Givens transformation very badly determined. To remedy this, we applied a *randomization* of the last two rows and first two columns of the obtained Cauchy-like matrix C. We realized this by means of a random update of the last two rows and first two columns of the generators of the low displacement rank matrix Rk 2 in the right-hand side of (1). The fact that we have applied a low rank perturbation to the matrix C can then be

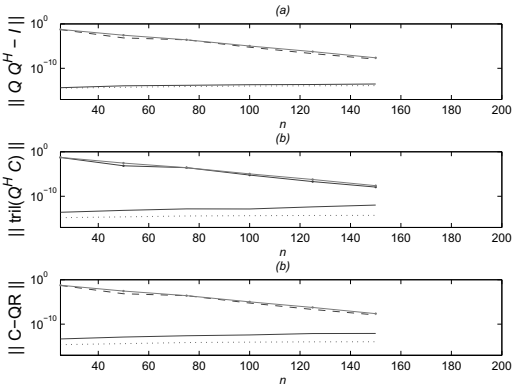


FIGURE 16. Accuracy for the prolate matrix of size n with parameter $\alpha = 0.48$.

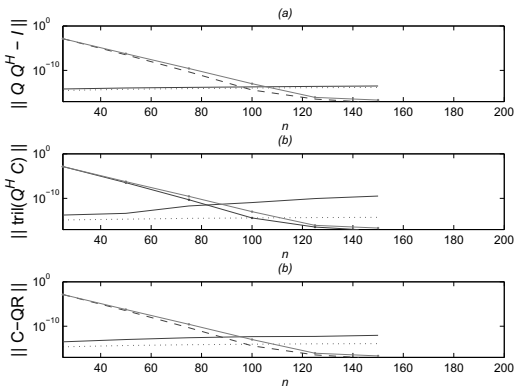


FIGURE 17. Accuracy for the prolate matrix of size n with parameter $\alpha = 0.45$.

dealt with by using any standard stable routine for updating a QR-factorization under the influence of a low rank correction term.

Figures 16 and 17 report the results for $\alpha = 0.48$ and $\alpha = 0.45$, respectively. It is seen that the initial randomization improves substantially the accuracy of the method. Moreover, Figure 17 seems to indicate that the accuracy is influenced by the conditioning of the matrices involved in the recursive factorization process. This claim sounds reasonable since from the classical perturbation analysis for the QR factorization of a nonsingular matrix C it follows that the magnitude of the perturbation ΔQ of the Q-factor of C under a small perturbation ΔC of C can be amplified by a factor depending on the conditioning of C [14].

However, we want to conclude by stressing that the mechanism of error amplification in the chasing-based procedures for the QR factorization of displacement structured matrices in not yet completely clear. This is specifically true for not diagonal displacement operators as those occurring with Toeplitz-like structures. Some very preliminary numerical experiments with Toeplitz matrices suggest that in that case the accuracy of computed results can be seriously worse than it is predicted by the conditioning of the problem. At this time we have no analysis or informal explanation for this phenomena and more research is still required for the design of numerically robust algorithms.

Acknowledgment

The first author would like to thank Prof. Gemignani for inviting him in September 2005 at the University of Pisa, Italy, where this joint work has started.

References

[1] A.W. Bojanczyk, R.P. Brent, and F.R. de Hoog. QR factorization of Toeplitz matrices. *Numerische Mathematik*, 49:81–94, 1986.

[2] S. Delvaux, L. Gemignani, and M. Van Barel. Fast QR-factorization of Cauchy-like matrices. *Linear algebra and its Applications*, 428, 2008.

[3] S. Delvaux and M. Van Barel. A Givens-weight representation for rank structured matrices. *SIAM Journal on Matrix Analysis and its Applications*, 29(4):1147–1170, 2007.

[4] S. Delvaux and M. Van Barel. A QR-based solver for rank structured matrices. *SIAM Journal on Matrix Analysis and its Applications*, 2008. To appear.

[5] S. Delvaux and M. Van Barel. Unitary rank structured matrices. *Journal of Computational and Applied Mathematics*, 215(1):49–78, 2008.

[6] P. Dewilde and A.-J. van der Veen. *Time-varying systems and computations*. Kluwer Academic Publishers, Boston, June 1998.

[7] D. Fasino and L. Gemignani. A Lanczos type algorithm for the QR-factorization of regular Cauchy matrices. *Numerical Linear Algebra with Applications*, 9:305–319, 2002.

[8] L. Gemignani, D.A. Bini, Y. Eidelman, and I.C. Gohberg. Fast QR eigenvalue algorithms for Hessenberg matrices which are rank-one perturbations of unitary matrices. *SIAM Journal on Matrix Analysis and its Applications*, 29, 2007.

[9] G. Heinig and A.W. Bojanczyk. Transformation techniques for Toeplitz and Toeplitz-plus-Hankel matrices I. Transformations. *Linear Algebra and its Applications*, 254:193–226, 1997.

[10] G. Heinig and A.W. Bojanczyk. Transformation techniques for Toeplitz and Toeplitz-plus-Hankel matrices II. Algorithms. *Linear Algebra and its Applications*, 278:11–36, 1998.

[11] G. Heinig and K. Rost. Representations of Toeplitz-plus-Hankel matrices using the trigonometric transformations with application to fast matrix-vector multiplication. *Linear Algebra and its Applications*, 275-276:225–248, 1998.

[12] T. Kailath and A.H. Sayed, editors. *Fast reliable algorithms for matrices with structure*. SIAM, Philadelphia, PA, USA, May 1999.

[13] V.Y. Pan. *Structured matrices and polynomials*. Birkhäuser Springer, 2001.

[14] X.W. Chang and C.C Paige and G.W. Stewart. Perturbation analyses for the QR factorization. *SIAM Journal on Matrix Analysis and its Applications*, 18(3):775–791, 1997.

[15] M. Van Barel, D. Fasino, L. Gemignani, and N. Mastronardi. Orthogonal rational functions and structured matrices. *SIAM Journal on Matrix Analysis and its Applications*, 26(3):810–829, 2005.

[16] C.F. Van Loan. *Computational Frameworks for the Fast Fourier Transform*. Frontiers in Applied Mathematics. SIAM, 1992.

Steven Delvaux
Department of Mathematics
Katholieke Universiteit Leuven
Celestijnenlaan 200B
B-3001 Leuven (Heverlee), Belgium
e-mail: steven.delvaux@wis.kuleuven.be

Luca Gemignani
Dipartimento di Matematica
Università di Pisa
Largo Bruno Pontecorvo 5
I-56127 Pisa, Italy
e-mail: gemignan@dm.unipi.it

Marc Van Barel
Department of Computer Science
Katholieke Universiteit Leuven
Celestijnenlaan 200A
B-3001 Leuven (Heverlee), Belgium
e-mail: marc.vanbarel@cs.kuleuven.be

Operator Theory:
Advances and Applications, Vol. 199, 255–288

Bezoutians Applied to Least Squares Approximation of Rational Functions

Sven Feldmann

To Georg Heinig

Abstract. A projection method to reduce large scale discrete systems which has been introduced in [12, 21] will be generalized to continues systems without to transform it bilinear. To achieve that goal depending on an algebraic curve $\gamma \subset \mathbb{C}$ and a rational function $h \in \mathbb{C}(z)$ a non negative function $F : \mathbb{C}^m \to \mathbb{R}$ is introduced whose minimizer provides an approximant of degree m. Special cases are obtained via specification of γ and h.

Mathematics Subject Classification (2000). 15A24; 41A20.

Keywords. Bezoutian, matrix equation, model reduction, stability.

1. Introduction

The paper concerns an approximation problem. Let $\gamma \subset \mathbb{C}$ be a curve which divides the complex plane $\mathbb{C}$ in two parts $\mathbb{C}_\ell$ and $\mathbb{C}_r$, and $h \in \mathbb{C}(z)$ be a stable strictly proper rational function, that means all it poles, denoted with $\sigma(h)$, are located in $\mathbb{C}_\ell$ and $\lim_{z\to\infty} h(z) = 0$. We are interested in the approximation of h by a low (Mc-Millan) degree strictly proper rational function h_{red} such that along γ the least squares distance between h and h_{red} is satisfying and h_{red} is again stable. As usual the degree of h is defined by the degree of $q \in \mathbb{C}[z]$ where p/q is a coprime fraction representation of h. That approximation problem admits a system theoretical interpretation via definition of h by the transfer function $h_\Sigma(z) := C(zI - A)^{-1}B$ of the systems

$$\begin{array}{rl} x_{k+1} & = Ax_k + Bu_k \\ y_k & = Cx_k \end{array}, \; x_0 = 0 \qquad \begin{array}{rl} \dot{x}(t) & = Ax(t) + Bu(t) \\ y(t) & = Cx(t) \end{array}, \; x(0) = 0$$

$$\Sigma := (A, B, C) \in \mathbb{R}^{n\times n} \times \mathbb{R}^{n\times 1} \times \mathbb{R}^{1\times n}$$

which transform the inputs

$$\begin{aligned} u_k &:= C_u A_u^k B_u, \quad u(t) := C_u \exp(A_u t) B_u, \\ \Sigma_u &:= (A_u, B_u, C_u) \in \mathbb{R}^{\mathfrak{n}\times\mathfrak{n}} \times \mathbb{R}^{\mathfrak{n}\times 1} \times \mathbb{R}^{1\times\mathfrak{n}}, \end{aligned} \quad \sigma(A_u) \cap \sigma(A) = \emptyset$$

to the outputs

$$y_k = C_u A_u^k \widehat{B}_u + C A^k \widehat{B}, \quad y(t) = C_u \exp(A_u t)\widehat{B}_u + C \exp(At)\widehat{B}$$
$$\widehat{B}_u := h_\Sigma(A_u) B_u, \quad \widehat{B} := h_{\Sigma_u}(A) B.$$

As usual $\exp(A)$ and $h(A)$ are defined, respectively, by $\sum_{k=0}^{\infty} \frac{A^k}{k!}$ and $p(A)q(A)^{-1}$. Among other things, large n makes the computation of the outputs infeasible in acceptable time. That is one reason why one is interested in the replacement of Σ by a smaller triple $\Sigma_{\rm red}$ which generates systems that have similar transfer behavior. To explain the term 'similar transfer behavior' we restrict ourself to

$$\gamma \in \{\mathbb{T}, \mathbf{i}\mathbb{R}\}, \quad \mathbb{T} := \{z \in \mathbb{C} : |z| = 1\}, \quad \mathbf{i}\mathbb{R} := \{z \in \mathbb{C} : \Re(z) = 0\}.$$

Here, $\Re(z)$ designates the real part of the complex number z. For such γ the left-hand parts $\mathbb{C}_\ell$ are the unit disc and the left half-plane

$$\mathbb{D} := \{z \in \mathbb{C} : |z| < 1\}, \quad \mathbb{H}_- := \{z \in \mathbb{C} : \Re(z) < 0\}.$$

Thus, the matrices A^k and $\exp(At)$ tend to zero as $k, t \to \infty$ such that for large enough k and t the quantities y_k and $y(t)$ are in essential equal to $C_u A_u^k \widehat{B}_u$ and $C_u \exp(A_u t)\widehat{B}_u$. Consequently, two stable systems Σ and $\Sigma_{\rm red}$ possess similar transfer behavior relative to the input signal class

$$\mathscr{S}_\gamma := \{\Sigma_u : \sigma(A_u) \subset \gamma\}$$

if for all $\Sigma_u \in \mathscr{S}_\gamma$ the number $d(A_u) := \|h_\Sigma(A_u) - h_{\Sigma_{\rm red}}(A_u)\|_{\rm F}$ is small. For $\Sigma_u \in \mathscr{S}_\gamma$ the smallness of $\|h_\Sigma - h_{\Sigma_{\rm red}}\|_2^\gamma$ leads to the smallness of $d(A_u)$. Here, $\|.\|_{\rm F}$ denotes the Frobenius matrix norm $\|A\|_{\rm F} := \sqrt{\sum_{i,j=1}^n |a_{ij}|^2}$, [[18], p. 56] and $\|.\|_2^\gamma$ the L_2-norm along γ, defined by

$$\|f\|_2^\gamma := \sqrt{\langle f|f\rangle_2^\gamma}, \quad \langle f|g\rangle_2^\gamma := \tfrac{1}{2\pi}\int_a^b \overline{f(w(t))} g(w(t)) |\dot{w}(t)| \mathrm{d}t$$
$$\gamma := \{w(t) : t \in [a, b]\}, \quad w : [a, b] \subseteq \mathbb{R} \to \mathbb{C}.$$

Hence, for a stable Σ, stimulated by elements of $\mathscr{S}_\gamma$, the reduction problem is equivalent to approximate h_Σ along γ by a stable strictly proper low degree rational function $h_{\rm red}$. The desired reduced system $\Sigma_{\rm red}$ is provided by a realization of $h_{\rm red}$. For $\gamma \in \{\mathbb{T}, \mathbf{i}\mathbb{R}\}$ the balanced truncation method [[35], Section 21.8, Chapter 7] represents a prominent reduction procedure. An overview with respect to the huge number of papers which deal with approximation problems of this nature offer [4, 5]. The optimal γ-approximation problem of order m consists in finding

$$h_{\rm opt} \in R_m^\gamma := \{g \in \mathbb{C}(z) : \sigma(g) \in \mathbb{C}_\ell, \deg g = m\}$$

such that

$$\|h - h_{\rm opt}\|_2^\gamma = \min\{\|h - g\|_2^\gamma : g \in R_m^\gamma\}. \tag{1.1}$$

Since R_m^γ does not not form a subspace of $\{g \in \mathbb{C}(z) : \|g\|_2^\gamma < \infty\}$ the determination of an h_{opt} becomes difficult. Via the specification $\gamma = \mathbf{i}\mathbb{R}$ the so-called optimal H_2-model reduction problem is obtained. There exists a lot of corresponding investigations, consult for instance [30, 32, 34].

We do not determine h_{opt} but we try to provide an h_{red} such that

$$\|h - h_{\text{red}}\|_2^\gamma \approx \|h - h_{\text{opt}}\|_2^\gamma .$$

To achieve that goal we associate h with two ordered polynomial sets

$$\mathscr{A} := \{a_0, \ldots, a_{n-1}\},\ \mathscr{R} := \{r_0, \ldots, r_{n-1}\}$$

and consider as approximation candidates the rational functions

$$h_x(z) := \frac{a_m(z) - \sum_{k=0}^{m-1} x_k a_k(z)}{z^m - \sum_{k=0}^{m-1} x_k z^k}, \quad x \in \mathbb{C}^m, \quad a_k \in \mathscr{A}, \quad 0 < m < n.$$

To find appropriate weight vectors x, depending on $\mathscr{R}$ and γ we define a non negative function $F : \mathbb{C}^m \to \mathbb{R}$ and derive the representation

$$F(x)^2 = \text{row}(x^*, -1) G_{m+1} \text{col}(x, -1), \quad G_m \in \mathbb{C}^{m \times m}. \tag{1.2}$$

It turns out that the minimizer ξ of F generates an h_ξ having all desired properties of h_{red} and that $G := G_n$ admits the factorization

$$G = R^* B_\gamma^{-1}(q) R. \tag{1.3}$$

We call G the Gramian of h with respect to γ. Due to $B_\gamma(q)$ is representing the Bezoutian of q with respect to γ, that factorization brings together the Bezout concept with rational approximation problems, admits a natural generalization of the Bezout concept and justifies paper's title. The context to [12, 21] is established by the possibility to compute ξ via a projection method which has been introduced there.

To analyze the properties of h_ξ the class of admissible curves will be restricted to the class of algebraic curves

$$\gamma := \{z \in \mathbb{C} : p_\gamma(z, \overline{z}) = 0\}. \tag{1.4}$$

Here, p_γ is a bivariate polynomial with Hermitian coefficient matrix Γ:

$$p_\gamma(x, y) := \psi_{\nu+1}(x) \Gamma \psi_{\nu+1}^{\mathrm{T}}(y),\ 0 \le \gamma_{\nu\nu},\ \Gamma := [\gamma_{ij}]_{i,j=0}^{\nu},\ \psi_n(x) := \text{row}(x^j)_{j=0}^{n-1}.$$

That restriction allows us to associate h and γ with the Hermitian matrix L defined by

$$\begin{gathered} L := L_\gamma(C_q, M) \in \mathbb{C}^{n \times n}, \quad M := G^{-1} \\ L_\gamma(A, X) : \mathbb{C}^{n \times n} \to \mathbb{C}^{n \times n},\ L_\gamma(A, X) := \textstyle\sum_{i,j=0}^{\nu} \gamma_{ij} A^i X A^{*j},\ A \in \mathbb{C}^{n \times n}. \end{gathered} \tag{1.5}$$

Remember, that for $A, B \in \mathbb{C}^{n \times n}$, $\gamma = \mathbb{T}$ and $\nu = 1$ the matrix equation

$$L_\gamma(A, X) = B$$

is said to be a Stein and for $\gamma = \mathbf{i}\mathbb{R}$ a Lyapunov equation. We pay our main attention to the case $\nu = 1$. Then the curvature of γ is constant such that $\mathbb{C}_\ell$ is a disc or a half-plane and non-positive L implies

$$\sigma(h_\xi) \subset \overline{\mathbb{C}_\ell} := \mathbb{C}_\ell \cup \gamma.$$

Furthermore, for positive

$$\Delta := \sqrt{-\det \Gamma}$$

the desired non-positivity is guaranteed by the existence of $\mathfrak{w} \in \mathbb{C}^n$ satisfying

$$L = -\Delta \mathfrak{w}\mathfrak{w}^*. \tag{1.6}$$

Our construction of $\mathfrak{w}$ is derived from the special cases $\gamma \in \{\mathbb{T}, \mathbf{i}\mathbb{R}\}$. In particular, in the unit circle case Parseval's identity states the equivalence between the definition of the denominator polynomial Q of h_ξ provided here, definition (10) of Q_n in [12] and definition (3.4) of χ_{LS} in [21]. Numerical experiences show that a small distance between $\sigma(h)$ and γ is preserved by $\sigma(h_\xi)$ and that $\sigma(h_\xi)$ possesses an element nearby to $\alpha \in \sigma(h)$ supposed the associated partial fraction coefficient is large. A heuristical discussion of that observations took place in [13]. The theoretical part of the present paper finishes with a mathematical foundation.

Their dependencies determine the paper's structure. In Section 2 we define F and G precisely and deduce representation (1.2). Afterwards, an estimate of the distance between h and h_x directly proportional to $F(x)$ is derived such that the definition of x via its minimizer ξ becomes natural. Due to factorization (1.3) Section 3 is dedicated to Bezoutians. In Section 4 we prove $\sigma(h_\xi) \subset \overline{\mathbb{C}_\ell}$ by exploitation of the matrix equation (1.6) and give an explicit construction of $\mathfrak{w}$. The paper's main theorem (Theorem 4.4) combines the pole separation result with the distance estimate along γ such that we arrive at a generalization of [[12], Theorem 2] and [[21], Theorem 3.1.]. As mentioned above, the construction receipt for $\mathfrak{w}$ is derived from the special cases $\gamma \in \{\mathbb{T}, \mathbf{i}\mathbb{R}\}$. Therefore, Section 5 and 6 deal with that cases. Section 7 focuses on the computation of h_ξ starting from the knowledge of a minimal realization of h. That happens by identification of G as the observability Gramian of the controllability realization of h with respect to γ. In the extreme situation $h = 1/q$ the set $\mathscr{A}$ becomes a singleton consisting only of the zero polynomial such that as approximant the zero function is only in question. To overcome that collapse, in Section 8 a second minimization step is concatenated to the minimization of F which defines the numerator of h_ξ more sensitive such that for the resulting rational function h_ξ^η the relation

$$\|h - h_\xi^\eta\|_2^\gamma \approx \|h - h_{\mathrm{opt}}\|_2^\gamma$$

is achieved. In particular by introduction of the mirror image of $\sigma(Q)$ with respect to γ the numerator polynomial $\mathscr{P}$ of h_ξ^η can be obtained as solution to an interpolation problem. In Section 9 the limit behavior of $\sigma(h_\xi)$ is studied when some pole of h moves to the border γ or a partial fraction coefficient tends to infinity. The main tool to analyze such limit processes are the generating functions of Bezoutians. Since there exist bilinear transformations between $\mathbb{D}$ and $\mathbb{H}_-$ the

statements with respect to $\mathbb{D}$ can be translated in corresponding statements with respect to $\mathbb{H}_{\_}$. In Section 10 we sketch that approach. Finally, in Section 11 we illustrate our theoretical results by means of numerical examples. The both last examples touch the concept of positive realness. It turns out that the direct as well as the transformation approach do not preserve the underlying mapping property. In a forthcoming paper that problem will be studied.

2. The construction of h_ξ

Depending on a differentiable curve γ, a positive measure μ acting on γ and polynomials $r_j \in \mathscr{R} \subset \mathbb{C}_n[z] := \{p \in \mathbb{C}[z] : \deg p < n\}$ we define $F : \mathbb{C}^m \to \mathbb{R}$, $m < n$, according to

$$\begin{array}{rlrl} F(x) & := \|p_x\|^\gamma_{\mathrm{d}\mu}\,, & p_x(z) & := r_m(z) - \sum_{j=0}^{m-1} x_j r_j(z), \\ \|f\|^\gamma_{\mathrm{d}\mu} & := \langle f \mid f\rangle^\gamma_{\mathrm{d}\mu}{}^{\frac{1}{2}}, & \langle f \mid g\rangle^\gamma_{\mathrm{d}\mu} & := \int_\gamma \overline{f} g \,\mathrm{d}\mu. \end{array}$$

To adapt F to our approximation problem relative to $p/q := h$, $p, q \in \mathbb{C}[z]$, the sets $\mathscr{A}$ and $\mathscr{R}$ are specified by the solutions to

$$z^j p = a_j q + r_j,\ r_j \in \mathbb{C}_n[z],\ a_j \in \mathbb{C}_j[z],\ j = 0, \ldots, n-1,\ \mathbb{C}_0[z] := \{0\}.$$

We call r_j the $(j+1)^{\text{th}}$ residue of p with respect to q. Cayley-Hamilton's Theorem [[35], p. 21] shows the coincidence of the coefficient vector $\mathfrak{r}_j$ of r_j regarded as an element of $\mathbb{C}^n$ with the j^{th} column of $p(C_q)$. Thus, the definition

$$R := \left[\begin{array}{ccc} \mathfrak{r}_0 & \ldots & \mathfrak{r}_{n-1} \end{array}\right] \in \mathbb{C}^{n\times n}$$

leads to $R = p(C_q)$. Due to $r_0 = p$ the first column of R coincides with the coefficient vector $\mathfrak{p}$ of p. Moreover,

$$p(C_q) = \mathscr{K}(C_q, \mathfrak{p}), \quad \mathscr{K}(A, B) := \mathscr{K}_n(A, B), \quad \mathscr{K}_\ell(A, B) := \mathrm{row}(A^j B)_{j=0}^{\ell-1}.$$

The measure μ is specified by

$$\mathrm{d}\mu(t) := \frac{|\dot{w}(t)|}{2\pi|q(w(t))|^2}\,\mathrm{d}t, \quad \begin{array}{c}\gamma := \{w(t) : t \in [a, b]\},\\ w : [a, b] \subseteq \mathbb{R} \to \mathbb{C},\end{array} \quad \dot{w} := \frac{\mathrm{d}}{\mathrm{d}t}\Re(w) + \mathbf{i}\frac{\mathrm{d}}{\mathrm{d}t}\Im(w).$$

As approximation candidates now the elements h_x of the rational function set

$$R_m(h) := \left\{ \frac{P}{Q} : \begin{array}{l} P(z) := a_m(z) - \sum_{k=0}^{m-1} x_k a_k(z), \\ Q(z) := z^m - \sum_{k=0}^{m-1} x_k z^k, \end{array} \ x \in \mathbb{C}^m \right\}$$

are in question. Note that for $f_k := r_k/q$ we have

$$\begin{array}{c} f_k(z) = z^k h(z) - a_k(z) = CA^k(zI - A)^{-1}B, \\ a_0(z) \equiv 0, \quad a_{k+1}(z) = \sum_{i=0}^{k} h_{k-i} z^i, \quad h_k := CA^k B. \end{array}$$

Here, (A, B, C) represents a realization of h as for example the controllability realization $\Sigma_{\mathrm{co}} := (C_q, e_1, \mathfrak{h}^*)$ [[14], p. 288] or the controller realization $\left(C_q^{\mathrm{T}}, e_n, \mathfrak{q}_n^{-1}\mathfrak{p}^{\mathrm{T}}\right)$

where (regard blank spaces as filled with zeros)

$$C_q := \left[\begin{array}{ccc|c} & & & -\mathfrak{q}_0/\mathfrak{q}_n \\ \hline 1 & & & -\mathfrak{q}_1/\mathfrak{q}_n \\ & \ddots & & \vdots \\ & & 1 & -\mathfrak{q}_{n-1}/\mathfrak{q}_n \end{array}\right], \quad \begin{array}{l} p = \psi_n \mathfrak{p} \\ q = \psi_{n+1}\mathfrak{q} \\ \mathfrak{h} := \operatorname{col}(\overline{h_i})_{i=0}^{n-1} \end{array} \tag{2.1}$$

$$e_1 := \left[\begin{array}{cccc} 1 & 0 & \dots & 0 \end{array}\right]^{\mathrm{T}}, \quad e_n := \left[\begin{array}{cccc} 0 & \dots & 0 & 1 \end{array}\right]^{\mathrm{T}}.$$

The equation $a_{k+1}(z) = \sum_{i=0}^{k} h_{k-i} z^i$ and the definition of P admit to express the coefficient vector ω of P according to

$$\omega = \left[\begin{array}{ccc} h_0 & \dots & h_{m-1} \\ & \ddots & \vdots \\ & & h_0 \end{array}\right] \left[\begin{array}{c} -\operatorname{col}(x_i)_{i=1}^{m-1} \\ 1 \end{array}\right]. \tag{2.2}$$

Therefore, $R_m(h)$ forms the set of all strictly proper rational functions of degree m whose first m Laurent coefficients coincide with $(h_k)_{k=0}^{m-1}$. For discrete systems that property means the coincidence of the unit impulse responses up to the length m. If one renounces to preserve that system property, the gained freedom to choose P can be used to proceed with a second minimization step after minimization of F. Section 8 focuses on the corresponding details. To estimate $\|h - h_x\|_2^\gamma$ we introduce

$$\min_\gamma(f) := \min\{|f(w)| : w \in \gamma\}, \quad \max_\gamma(f) := \max\{|f(w)| : w \in \gamma\}.$$

Proposition 2.1. *Supposed $\sigma(Q) \cap \gamma = \emptyset$ the estimate $\|h - h_x\|_2^\gamma \le \frac{F(x)}{\min_\gamma(Q)}$ holds.*

Proof. Replacement of r_j by $z^j p - a_j q$ yields

$$\begin{aligned} p_x = \;& r_m - \textstyle\sum_{j=0}^{m-1} x_j r_j = (z^m p - a_m q) - \sum_{j=0}^{m-1} x_j \left(z^j p - a_j q\right) \\ = \;& \left(z^m - \textstyle\sum_{j=0}^{m-1} x_j z^j\right) p - \left(a_m - \textstyle\sum_{j=0}^{m-1} x_j a_j\right) q = Qp - Pq. \end{aligned}$$

Thus $\|h - h_x\|_2^\gamma = \|p_x/(Qq)\|_2^\gamma \le \max_\gamma(Q^{-1}) \; \|p_x/q\|_2^\gamma = \frac{1}{\min_\gamma(Q)} \|p_x\|_{\mathrm{d}\mu}^\gamma$. □

Consequently, to get along γ a small distance between h and h_x, the choice of x via the minimizer ξ of F is natural. To determine ξ note that

$$\begin{aligned} F(x)^2 \;&=\; \left\langle r_m - \textstyle\sum_{j=0}^{m-1} x_j r_j \;\middle|\; r_m - \sum_{j=0}^{m-1} x_j r_j \right\rangle_{\mathrm{d}\mu}^\gamma \\ &= \left[\begin{array}{cc} x^* & -1 \end{array}\right] \left[\langle r_i \,|\, r_j \rangle_{\mathrm{d}\mu}^\gamma\right]_{i,j=0}^{m} \left[\begin{array}{c} x \\ -1 \end{array}\right]. \end{aligned}$$

Using the abbreviations

$$G := G_n,\; G_m := [g_{ij}]_{i,j=0}^{m-1},\; g_{ij} := \langle r_i \,|\, r_j \rangle_{\mathrm{d}\mu}^\gamma$$

the factorization (1.2) is obtained. We call G the Gramian of h with respect to γ. In addition we set

$$g_m := \operatorname{col}(g_{im})_{i=0}^{m-1} \in \mathbb{C}^m.$$

If one defines $a_n \in \mathbb{C}_n[z]$ and $r_n \in \mathbb{C}_{n+1}[z]$ via $z^n p = a_n q + r_n$ then the vector g_n is defined and q admits the representation

$$q(z) = \mathfrak{q}_n(z^n - \psi_n(z)G^{-1}g_n).$$

Proposition 2.2. *Supposed p and q are coprime F achieves its minimum in $G_m^{-1}g_m$. Thus, we set*

$$Q(z) := z^m - \psi_m(z)\xi, \quad \xi := G_m^{-1}g_m. \tag{2.3}$$

Proof. Obviously,

$$G = R^* \left[\langle z^i \mid z^j\rangle_{\mathrm{d}\mu}^{\gamma}\right]_{i,j=0}^{n-1} R \tag{2.4}$$

where the positivity of the middle factor is a consequence of the positivity of μ. Since p and q are coprime, together with $R = p(C_q)$ the invertibility of R follows. Thus G is positive. We proceed with the principal axis transformation of G_m. Let

$$G_m = U\mathrm{diag}(d_k)_{k=1}^m U^*$$

where U is unitary and all d_k are positive. With $\hat{x} := Ux$ and $\hat{g}_m := Ug_m$ we get

$$F(x)^2 = g_{mm} + \textstyle\sum_{k=1}^m f(\hat{x}_k, d_k, \hat{g}_{km}), \quad f(z,d,g) := d|z|^2 - 2\Re(g\overline{z}).$$

Therefore, the minimum of F^2 is equal to

$$g_{mm} + \textstyle\sum_{k=1}^m \min_{z\in\mathbb{C}} f(z, d_k, \hat{g}_{km}).$$

The condition $f_x = f_y = 0$, $z = x + \mathbf{i}y$, leads to

$$2dx - 2\Re(g) = 0, \quad 2dy - 2\Im(g) = 0.$$

Hence, $\hat{x}_k = \hat{g}_{km}/d_k$ that means $G_m^{-1}g_m$ represents the only one stationary point of F^2. Since the (2×2)-matrix of the second derivations of f is equal to $\mathrm{diag}(d_k, d_k)$ and d_k is positive, F attains its minimum there. □

3. Bezoutians

We introduce now the Bezout concept for q with respect to γ. The classical Bezoutians are special Hermitian matrices that were introduced to study the root location problem of polynomials. The naming of Bezoutians goes back to J. Sylvester [31]. Historical remarks can be found in [25, 26, 33]. In view of (1.3) and (2.4) we set

$$B_\gamma(q) := \left(\left[\ \langle z^i \ \big|\ z^j\rangle_{\mathrm{d}\mu}^{\gamma}\ \right]_{i,j=0}^{n-1}\right)^{-1}.$$

In Section 5 and 6 for $\gamma \in \{\mathbb{T}, \mathbf{i}\mathbb{R}\}$ and $\sigma(q) \subset \mathbb{C}_\ell$ the coincidence of $B_\gamma(q)$ with the classical Bezoutians

$$\begin{aligned} B_d(q) &:= H_{\triangleright}(\mathfrak{q}_1, \ldots, \mathfrak{q}_n)H_{\triangleright}(\overline{\mathfrak{q}_1}, \ldots, \overline{\mathfrak{q}_n}) - H_{\triangleright}(\overline{\mathfrak{q}_{n-1}}, \ldots, \overline{\mathfrak{q}_0})H_{\triangleright}(\mathfrak{q}_{n-1}, \ldots, \mathfrak{q}_0) \\ B_c(q) &:= I_{\pm}(T_{\triangleright}(\widetilde{\mathfrak{q}}_0, \ldots, \widetilde{\mathfrak{q}}_{n-1})H_{\triangleright}(\overline{\mathfrak{q}_1}, \ldots, \overline{\mathfrak{q}_n}) - T_{\triangleright}(\overline{\mathfrak{q}_0}, \ldots, \overline{\mathfrak{q}_{n-1}})H_{\triangleright}(\widetilde{\mathfrak{q}}_1, \ldots, \widetilde{\mathfrak{q}}_n)) \end{aligned} \tag{3.1}$$

will be proved, a fact that justifies to refer $B_\gamma(q)$ as Bezoutian. Here, $\widetilde{\mathfrak{q}} := I_\pm \mathfrak{q}$ and

$$I_\pm := \begin{bmatrix} 1 & & \\ & -1 & \\ & & \ddots \end{bmatrix}, H_\triangleright(x_1,\dots,x_n) := \begin{bmatrix} x_1 & \dots & x_n \\ \vdots & \iddots & \\ x_n & & \end{bmatrix}, T_\triangleright(x_1,\dots,x_n) := \begin{bmatrix} x_1 & & \\ \vdots & \ddots & \\ x_n & \dots & x_1 \end{bmatrix}.$$

The right-hand side of (3.1) is said to be the Gohberg-Semencul representation of $B_\gamma(q)$ [[14], p. 206]. In view of [[26], Theorem XVa], [[24], Theorem 4.4], [[15], Theorem B2], [[14], Theorem 9.1.5.] the relations

$$\begin{aligned} \sigma(q) \subset \mathbb{D} &\Leftrightarrow 0 < B_d(q), \quad \sigma(q) \subset \mathbb{C} \setminus \overline{\mathbb{D}} \Leftrightarrow B_d(q) < 0 \\ \sigma(q) \subset \mathbb{H}_- &\Leftrightarrow 0 < B_c(q), \quad \sigma(q) \subset \mathbb{C} \setminus \overline{\mathbb{H}}_- \Leftrightarrow B_c(q) < 0 \end{aligned} \tag{3.2}$$

hold. To find the vector $\mathfrak{w}$ satisfying $L = -\Delta \mathfrak{w}\mathfrak{w}^*$ it is worth to consider appropriate representations of the bivariate polynomial

$$B_\gamma(q,x,y) := \psi_n(x) B_\gamma(q) \psi_n^{\mathrm{T}}(y).$$

According to [[14], Theorem 9.1.5, p. 256] we have

$$\begin{gathered} B_d(q,x,y) = \frac{\widehat{\overline{q}}(x)\widehat{q}(y) - q(x)\overline{q}(y)}{1-xy}, \quad B_c(q,x,y) = \frac{q(x)\overline{q}(y) - \widetilde{\overline{q}}(x)\widetilde{q}(y)}{x+y} \\ \widehat{q}(z) := q(z^{-1})z^n, \quad \widetilde{q}(z) := q(-z), \quad \overline{q}(z) := \mathfrak{q}^* \psi_{n+1}^{\mathrm{T}}(z). \end{gathered} \tag{3.3}$$

The first factors of the products arising in the numerators provide now $\mathfrak{w}$. Indeed, $\mathfrak{w} = R^{-1}\mathfrak{v}$ where for $\gamma = \mathbb{T}$ the polynomial $v =: \psi_n \mathfrak{v}$ is the first residue of $\widehat{\overline{q}}$ with respect to q and for $\gamma = \mathbf{i}\mathbb{R}$ the first residue of q with respect to $\widetilde{\overline{q}}$. The representation (3.3) of $B_\gamma(q,x,y)$ is said to be the generating function of $B_\gamma(q)$.

4. The matrix L for $\nu = 1$ and indefinite Γ

We specify γ by (1.4) and set

$$\mathbb{C}_\ell := \{z \in \mathbb{C} : p_\gamma(z,\overline{z}) < 0\}.$$

Our proof of $\sigma(h_\xi) \subset \overline{\mathbb{C}_\ell}$ needs the following representation of Q. Remember that M is defined by G^{-1}.

Proposition 4.1. *Let M be partitioned according to $[M_{ij}]_{i,j=1}^2$, $M_{11} \in \mathbb{C}^{m\times m}$. Then M_{22} is invertible and Q admits the representation*

$$Q(z) = \textstyle\sum_{j=m+1}^n \zeta_j m_j(z),\ \zeta := M_{22}^{-1} e_1 \in \mathbb{C}^{n-m},\ m_j := \psi_n M e_j \in \mathbb{C}_n[z].$$

Proof. Since G is positive, the same holds for M. Consequently, M_{22} is invertible. Remembering the equation

$$G_{11}^{-1} G_{12} = -M_{12} M_{22}^{-1}, \quad [G_{ij}]_{i,j=1}^2 := G, \quad M_{11}, G_{11} \in \mathbb{C}^{m\times m}$$

with $G_{11} = G_m$, $g_m = G_{12} e_1$ and $\xi = G_m^{-1} g_m$ one concludes

$$\begin{aligned} Q(z) &= z^m - \psi_m(z)\xi = z^m + \psi_m(z) M_{12} M_{22}^{-1} e_1 \\ &= z^m + \psi_m(z) M_{12} \zeta = \psi_n(z) \mathrm{col}(M_{i2})_{i=1}^2 \zeta = \textstyle\sum_{j=m+1}^n \zeta_j m_j(z). \end{aligned}$$

□

We state and prove now our stability result assuming $\nu = 1$ and the non-positivity of L. As mentioned above, $\nu = 1$ implies the coincidence of $\mathbb{C}_\ell$ with an open disc or an open half-plane.

Proposition 4.2. *For $\nu = 1$ and $L \le 0$ we have $\sigma(h_\xi) \subset \overline{\mathbb{C}_\ell}$. In the circle case h_ξ is stable.*

Proof. A not vanishing $q(\alpha)$ provides the inversion formula

$$(\alpha I - C_q)^{-1} = \begin{bmatrix} \mathfrak{q}_n & \dots & \mathfrak{q}_1 \\ & \ddots & \vdots \\ & & \mathfrak{q}_n \end{bmatrix} \begin{bmatrix} \alpha^{n-1} \\ \vdots \\ \alpha^0 \end{bmatrix} \frac{\psi_n(\alpha)}{q(\alpha)} - \begin{bmatrix} 0 & \alpha^0 & \dots & \alpha^{n-2} \\ & 0 & \ddots & \vdots \\ & & \ddots & \alpha^0 \\ & & & 0 \end{bmatrix}.$$

According to Proposition 4.1 the polynomial Q admits the representation

$$Q = \psi_n \tau, \quad \tau := M\varrho, \quad \varrho := \operatorname{col}(0, \zeta).$$

Since $\deg Q = m$ the vector τ is of the form $\operatorname{col}(\mathfrak{Q}, 0)$, $\mathfrak{Q} \in \mathbb{C}^{m+1}$. Assume $Q(\alpha) = 0$, $q(\alpha) \ne 0$ and let $\epsilon \in \mathbb{C}^n$ be defined by $\epsilon := (\alpha I - C_q)^{-*} \varrho$. Then ϵ and τ are orthogonal:

$$\epsilon^* \tau = \varrho^* (\alpha I - C_q)^{-1} \tau$$

$$= \varrho^* \begin{bmatrix} \mathfrak{q}_n & \dots & \mathfrak{q}_1 \\ & \ddots & \vdots \\ & & \mathfrak{q}_n \end{bmatrix} \begin{bmatrix} \alpha^{n-1} \\ \vdots \\ \alpha^0 \end{bmatrix} \frac{Q(\alpha)}{q(\alpha)} - \begin{bmatrix} 0 \\ \zeta \end{bmatrix}^* \begin{bmatrix} 0 & \alpha^0 & \dots & \alpha^{n-2} \\ & 0 & \ddots & \vdots \\ & & \ddots & \alpha^0 \\ & & & 0 \end{bmatrix} \begin{bmatrix} \mathfrak{Q} \\ 0 \end{bmatrix} = 0.$$

Together with the definition of the real numbers

$$\vartheta := \epsilon^* L \epsilon \le 0, \quad \kappa := \epsilon^* M \epsilon > 0, \quad \lambda := \varrho^* M \varrho > 0$$

and the equations $\epsilon^* C_q = \alpha \epsilon^* - \varrho^*$, $\tau = M\varrho$ we get

$$\begin{aligned} \vartheta &= \gamma_{11} \epsilon^* C_q M C_q^* \epsilon + \gamma_{10} \epsilon^* C_q M \epsilon + \gamma_{01} \epsilon^* M C_q^* \epsilon + \gamma_{00} \epsilon^* M \epsilon \\ &= \gamma_{11} (\alpha \epsilon^* - \varrho^*) M (\epsilon \overline{\alpha} - \varrho) + \gamma_{10} (\alpha \epsilon^* - \varrho^*) M \epsilon + \gamma_{01} \epsilon^* M (\epsilon \overline{\alpha} - \varrho) + \gamma_{00} \kappa \\ &= \gamma_{11} (\alpha \overline{\alpha} \kappa - \tau^* \epsilon \overline{\alpha} - \alpha \epsilon^* \tau + \lambda) + \gamma_{10} (\alpha \kappa - \tau^* \epsilon) + \gamma_{01} (\overline{\alpha} \kappa - \epsilon^* \tau) + \gamma_{00} \kappa \\ &= (\gamma_{11} \alpha \overline{\alpha} + \gamma_{10} \alpha + \gamma_{01} \overline{\alpha} + \gamma_{00}) \kappa + \gamma_{11} \lambda = p_\gamma(\alpha, \overline{\alpha}) \kappa + \gamma_{11} \lambda. \end{aligned}$$

Thus

$$p_\gamma(\alpha, \overline{\alpha}) = (\vartheta - \gamma_{11} \lambda)/\kappa \le 0$$

that means $\alpha \in \overline{\mathbb{C}_\ell}$. For positive γ_{11} the inequality $p_\gamma(\alpha, \overline{\alpha}) < 0$ holds that means $\alpha \in \mathbb{C}_\ell$. □

In Example 2 we meet the situation $\sigma(h_\xi) \cap \gamma \ne \emptyset$ actually. Thus, without additional assumptions the statement of Proposition 4.2 cannot be strengthened. In particular, for the case where γ is a line and $h := 1/q$ the localization of $\sigma(Q)$ is a consequence of Fejér's convex hull theorem [[9], Thm. 10.2.2], [10]. In Section 6 for

$h = p/q$ and $m := \deg q - \deg p$ we prove that Q becomes stable also for the half-plane case, and in [11] it will be shown that for $m \in \{\deg q - \deg p + 1, \ldots, n-1\}$ the polynomial Q is generic stable.

The proofs of Proposition 4.3 and 6.1 apply the bilinear transformation

$$\phi : \mathbb{C} \to \mathbb{C}, \quad \phi(z) := \frac{az+b}{cz+d}, \quad ad - bc \neq 0$$

with appropriate parameters $a, b, c, d \in \mathbb{C}$. The associated n-dimensional Möbius matrix M_ϕ is formed by the coefficient vectors of

$$\{(az+b)^{j-1}(cz+d)^{n-j} : j = 1, \ldots, n\}.$$

For vanishing c we extent the coefficient vector of $(az+b)^j d^{n-j}$ to an element of $\mathbb{C}^n$ via a corresponding number of zeros such that M_ϕ becomes a right upper triangular matrix. If one sets

$$p_\phi(z) := p(\phi(z))(cz+d)^{n-1}, \quad \deg p < n$$

then the associated coefficient vectors $\mathfrak{p}$ and $\mathfrak{p}_\phi$ are related by $M_\phi \mathfrak{p} = \mathfrak{p}_\phi$. For the transformations z^{-1} and $-z$ the Möbius matrices read as $J := \begin{bmatrix} & & 1 \\ & \iddots & \\ 1 & & \end{bmatrix}$ and $I_\pm$. Thus

$$\widehat{\mathfrak{q}} = J\mathfrak{q}, \quad \widehat{\overline{\mathfrak{q}}} = J\overline{\mathfrak{q}}, \quad \widetilde{\mathfrak{q}} = I_\pm \mathfrak{q}, \quad \widetilde{\overline{\mathfrak{q}}} = I_\pm \overline{\mathfrak{q}}.$$

For the parameters $a = -1$ and $b = c = d = 1$ the properties of M_ϕ are described in [[27], p. 37]. Note, that for $\nu = 1$ the quantity Δ is equal to $\sqrt{|\gamma_{01}|^2 - \gamma_{00}\gamma_{11}}$ and is assumed to be positive.

Proposition 4.3. *Stable h implies the existence of $\mathfrak{w} \in \mathbb{C}^n$ with $L = -\Delta \mathfrak{w}\mathfrak{w}^*$.*

Proof. By assumption Δ is positive. Thus for positive γ_{11}, the curve γ represents a circle and for vanishing γ_{11} a line. Let $w(t) := \phi(z(t))$ where $\phi(z) := az + b$ $(c = 0, d = 1)$ and $t \in \mathscr{I}$. Then $w(t)$ is a parameterization of γ where $a, b, z(t), \mathscr{I}$ are defined by

	a	b	$z(t)$	$\mathscr{I}$
$0 < \gamma_{11}$	Δ/γ_{11}	$-\gamma_{01}/\gamma_{11}$	$\exp(\mathbf{i}t)$	$[0, 2\pi]$
$0 = \gamma_{11}$	$1/\gamma_{10}$	$-\gamma_{00}/(2\gamma_{10})$	$\mathbf{i}t$	$\mathbb{R}$

Using the abbreviations $\gamma' := \phi^{-1}(\gamma)$, $B := B_\gamma(q)$, $B' := B_{\gamma'}(q_\phi)$ immediately by definition the identities

$$B' = |a| M_\phi B M_\phi^*, \quad a C_{q_\phi} M_\phi = M_\phi (C_q - bI)$$

follow. Combination of (1.3) and (1.5) with the commutator equation $RC_q = C_q R$ yields

$$L = R^{-1} L_\gamma(C_q, B) R^{-*}.$$

Hence, setting $\mathfrak{w} := R^{-1} M_\phi^{-1} \mathfrak{v}$ it remains to compute the appropriate $\mathfrak{v}$.

In the circle case we have $\gamma_{11}a = \Delta$. Thus

$$\begin{aligned} L_\gamma(C_q, B) &= \gamma_{11}((C_q - bI)B(C_q - bI)^* - a^2 B) \\ &= \Delta a(M_\phi^{-1} C_{q_\phi} M_\phi B M_\phi^* C_{q_\phi}^* M_\phi^{-*} - B) \\ &= \Delta M_\phi^{-1}(C_{q_\phi} B' C_{q_\phi}^* - B') M_\phi^{-*}. \end{aligned}$$

Due to $\gamma' = \mathbb{T}$ and $\sigma(q) \subset \mathbb{C}_\ell \Leftrightarrow \sigma(q_\phi) \subset \mathbb{D}$ Proposition 5.2 states the existence of $\mathfrak{v}$ such that $C_{q_\phi} B' C_{q_\phi}^* - B' = -\mathfrak{v}\mathfrak{v}^*$.

In the line case we have $|a|^{-1} = \Delta$. Thus

$$\begin{aligned} L_\gamma(C_q, B) &= \gamma_{10}(C_q - bI)B + B(C_q - bI)^* \gamma_{01} \\ &= \underbrace{\gamma_{10} a}_{=1} M_\phi^{-1} C_{q_\phi} M_\phi B + B M_\phi^* C_{q_\phi} M_\phi^{-*} \underbrace{\overline{a}\gamma_{01}}_{=1} \\ &= \Delta M_\phi^{-1}(C_{q_\phi} B' + B' C_{q_\phi}^*) M_\phi^{-*}. \end{aligned}$$

Due to $\gamma' = \mathbf{i}\mathbb{R}$ and $\sigma(q) \subset \mathbb{C}_\ell \Leftrightarrow \sigma(q_\phi) \subset \mathbb{H}_-$ Proposition 6.2 states the existence of $\mathfrak{v}$ such that $C_{q_\phi} B' + B' C_{q_\phi}^* = -\mathfrak{v}\mathfrak{v}^*$. □

The combination of the Propositions 2.1, 4.2 and 4.3 generalizes [[12], Theorem 2] and [[21], Theorem 3.1].

Theorem 4.4. *In general we have $\sigma(h_\xi) \subset \overline{\mathbb{C}_\ell}$. Moreover, for the circle case h_ξ is stable. Supposed $\sigma(h_\xi) \cap \gamma = \emptyset$ the approximant h_ξ satisfies*

$$\|h - h_\xi\|_2^\gamma \leq F(\xi)/\min_\gamma(Q).$$

Proof. Proposition 2.1 provides the estimate and Proposition 4.3 states the assumption of Proposition 4.2 which provides the location of $\sigma(h_\xi)$. □

5. The unit circle case $\gamma = \mathbb{T}$

The entries of G read as

$$g_{ij} = \langle r_i | r_j \rangle_{\mathrm{d}\mu}^{\mathbb{T}}, \quad \langle f | g \rangle_{\mathrm{d}\mu}^{\mathbb{T}} = \tfrac{1}{2\pi} \int_0^{2\pi} \overline{f(e^{\mathbf{i}t})} g(e^{\mathbf{i}t}) |q(e^{\mathbf{i}t})|^{-2} \, \mathrm{d}t.$$

Moreover, for $\sigma(q) \subset \mathbb{D}$ the coincidence of $B_{\mathbb{T}}(q)$ with the Toeplitz Bezoutian $B_d(q)$ can be proved. The name Toeplitz Bezoutian is justified by the fact that $B_d^{-1}(q)$ is a Toeplitz matrix. Matrices whose entries are constant along each diagonal are called Toeplitz matrices. To prove the asserted coincidence we use the semi-infinite Hankel matrix H generated by $1/q$:

$$H := [h_{i+j}]_{i,j=0}^{\infty, n-1}, \quad h_0 z^{-1} + h_1 z^{-2} + h_2 z^{-3} + \cdots := q(z)^{-1}.$$

Matrices whose entries are constant along each anti-diagonal are called Hankel matrices.

Proposition 5.1. $\sigma(q) \subset \mathbb{D} \Leftrightarrow B_d(q) = B_{\mathbb{T}}(q)$.

Proof. Due to $\sigma(q) \subset \mathbb{D}$ the complex number sequence $(h_k)_{k=0}^{\infty}$ belongs to ℓ_2 where as usual ℓ_2 denotes the normed vector space of all square-summable complex number sequences:

$$\varepsilon \in \ell_2 \Leftrightarrow \|\varepsilon\|_2 := \left(|\varepsilon_0|^2 + |\varepsilon_1|^2 + \cdots\right)^{1/2} < \infty.$$

The Schwarz inequality [[29], p. 77] provides the existence of H^*H. The Hankel structure of H implies the solution property of H^*H with respect to

$$C_q^* X C_q - X = -e_n |\mathfrak{q}_n|^{-2} e_n^*. \tag{5.1}$$

By exploitation of the structure of C_q one shows that the product

$$\widehat{q}(C_q^*)^{-1} \overline{\mathfrak{q}_n^{-1}} \mathscr{K}(C_q^*, e_n) J$$

solves that equation as well. Since the solution is unique we get

$$K\widehat{q}(C_q^*) = (H^*H)^{-1}, \quad K := J\mathscr{K}(C_q^*, e_n)^{-1} \overline{\mathfrak{q}_n}.$$

The left-hand product

$$K\widehat{q}(C_q^*) \tag{5.2}$$

is known as Barnett's factorization of $B_d(q)$ [[14], p. 204], hence $B_d^{-1}(q) = H^*H$. Finally, Parseval's identity [[29], p. 85] provides

$$B_d^{-1}(q) = H^*H = \tfrac{1}{2\pi} \textstyle\int_0^{2\pi} \psi_n^*(e^{\mathbf{i}t}) \psi_n(e^{\mathbf{i}t}) |q(e^{\mathbf{i}t})|^{-2} \mathrm{d}t = B_{\mathbb{T}}^{-1}(q).$$

Suppose now $B_d(q) = B_{\mathbb{T}}(q)$. Because $B_{\mathbb{T}}(q)$ is positive, the assumption yields the positivity of $B_d(q)$. With (3.2) relation $\sigma(q) \subset \mathbb{D}$ follows. □

As we have seen $B_{\mathbb{T}}^{-1}(q)$ solves the Stein equation (5.1). To create for G and $B_{\mathbb{T}}(q)$ corresponding equations, we generate their right-hand sides, respectively, by $\mathfrak{h}$ defined as in (2.1) and by the coefficient vector $\mathfrak{v}$ of the first residue v of $\widehat{\overline{q}}$ with respect to q:

$$v(z) = \psi_n(z)\mathfrak{v}, \quad v(z) := \widehat{\overline{q}}(z) - a_0 q(z), \quad a_0 := \overline{\mathfrak{q}_0}/\mathfrak{q}_n.$$

Proposition 5.2. *Supposed $\sigma(q) \subset \mathbb{D}$ the Gramian G and the Bezoutian $B_{\mathbb{T}}(q)$ solve, respectively,*

$$C_q^* X C_q - X = -\mathfrak{h}\mathfrak{h}^*, \quad C_q X C_q^* - X = -\mathfrak{v}\mathfrak{v}^*. \tag{5.3}$$

Proof. In view of (1.3), Proposition 5.1, $R = p(C_q)$ and $e_n^* R = \mathfrak{q}_n \mathfrak{h}^*$ with the abbreviation $B := B_{\mathbb{T}}(q)$ we obtain

$$C_q^* G C_q - G = R^*(C_q^* B^{-1} C_q - B^{-1})R = -R^* e_n |\mathfrak{q}_n|^{-2} e_n^* R = -\mathfrak{h}\mathfrak{h}^*.$$

According to the proof of Proposition 5.1 the matrix B can be replaced by $K\widehat{q}(C_q^*)$. Thus

$$C_q B C_q^* - B = C_q K \widehat{q}(C_q^*) C_q^* - K\widehat{q}(C_q^*) = (C_q K C_q^* - K)\widehat{q}(C_q^*).$$

The definition of v implies $\mathfrak{v} = \widehat{\overline{q}}(C_q)e_1$. The identity $C_q K C_q^* - K = -\mathfrak{v}e_1^*$ completes the proof. □

Let $\mathscr{Q} \in \mathbb{C}^{n\times n}$ be the observability Gramian of $\Sigma := (A, B, C)$ with respect to γ meaning that $\mathscr{Q}$ solves the matrix equation

$$L_{\overline{\gamma}}(A^*, X) = -\Delta C^* C, \quad L_{\overline{\gamma}}(A, X) := \gamma_{00}X + \gamma_{01}AX + \gamma_{10}XA^* + \gamma_{11}AXA^*.$$

Here, Σ is a minimal realization of h meaning that the size of A is small as possible. For stable h the matrix $\mathscr{Q}$ exists and is unique. Consideration of the controllability realization Σ_{co} of h reveals G as the observability Gramian of Σ_{co} with respect to the unit circle.

For $\sigma(q) \subset \mathbb{D}$ Proposition 5.3 states the equivalence of definition (2.3) of Q, definition (10) of Q_n in [12] and definition (3.4) of χ_{LS} in [21]. As usual, let the Moore-Penrose pseudo inverse of A be denoted by $A^\dagger$ [[18], p. 243].

Proposition 5.3. *Supposed $\sigma(q) \subset \mathbb{D}$ the minimizer ξ of F satisfies $\xi = H_m^\dagger V$ where*

$$H_m := [h_{i+j}]_{i,j=0}^{\infty,m-1}, \quad V := \operatorname{col}(h_i)_{i=m}^{\infty}, \quad \textstyle\sum_{k=0}^{\infty} h_k z^{-(k+1)} := \frac{p(z)}{q(z)}. \tag{5.4}$$

Proof. Parseval's identity states $G_m = H_m^* H_m$ and $g_m = H_m^* V$. Thus,

$$H_m^\dagger V = (H_m^* H_m)^{-1} H_m^* V = \xi. \qquad \square$$

Hence, the vector $H_m\xi$ represents the projection of V on the linear subspace

$$\mathscr{H}_m := \operatorname{span}(H_m) := \{H_m x : x \in \mathbb{C}^m\} \subset \ell_2$$

meaning that the statement of Proposition 5.3 can be interpreted as projection method, moreover, since H_m possesses Hankel structure, as a semi-infinite structured least squares problem. Together with formula (2.2) where x is to replace by ξ, the computation receipt proposed as in [12] is obtained. The projection point of view allows us to express the approximation error via the ℓ_2-distance between V and its projection on $\mathscr{H}_m$.

Proposition 5.4. *Let for $\sigma(q) \subset \mathbb{D}$ and $x \in \mathbb{C}^m$ the ℓ_2-sequence ε be defined by the components of the infinite-dimensional vector $H_m x - V$. Supposed $\sigma(Q) \cap \mathbb{T} = \emptyset$ we have $\|h - h_x\|_2^{\mathbb{T}} \le \|\varepsilon\|_2 / \min_{\mathbb{T}}(Q)$.*

Proof. In virtue of Proposition 2.1 it suffices $F(x) = \|\varepsilon\|_2$ to show. Obviously, the j^{th} column of $[H_m, V]$ coincide with the Laurent coefficient vector generated by f_j. Thus, the definition of ε and the equation $f_k(z) = z^k h(z) - a_k(z)$ lead to

$$E(z) := \textstyle\sum_{i=0}^{\infty} \varepsilon_i z^{-(i+1)} = \sum_{k=0}^{m-1} f_k(z)x_k - f_m(z) = P(z) - h(z)Q(z).$$

Together with $p_x = Qp - Pq$ we conclude

$$E/Q = h_x - h = -p_x/(Qq) \;\Rightarrow\; E = -p_x/q \;\Rightarrow\; \|E\|_2^{\mathbb{T}} = \|p_x\|_{\mathrm{d}\mu}^{\mathbb{T}} = F(x).$$

Parseval's identity yields $\|E\|_2^{\mathbb{T}} = \|\varepsilon\|_2$. $\square$

6. The imaginary axis case $\gamma = \mathbf{i}\mathbb{R}$

The entries of G read as $g_{ij} = \langle r_i | r_j \rangle_{\mathrm{d}\mu}^{\mathbf{i}\mathbb{R}}$ where

$$\langle f|g\rangle_{\mathrm{d}\mu}^{\mathbf{i}\mathbb{R}} = \tfrac{1}{2\pi}\int_{\mathbb{R}} \overline{f(\mathbf{i}t)}g(\mathbf{i}t)|q(\mathbf{i}t)|^{-2}\,\mathrm{d}t.$$

To represent G as for the unit circle case the Hankel Bezoutian $B_c(q)$ is in question. The name Hankel Bezoutian is justified by the fact that $B_c^{-1}(q)I_{\pm}$ is a Hankel matrix.

Proposition 6.1. $\sigma(q) \subset \mathbb{H}_- \Leftrightarrow B_c(q) = B_{\mathbf{i}\mathbb{R}}(q)$.

Proof. Using the bilinear transformation $\phi(z) := \frac{z+1}{z-1} : \mathbb{D} \to \mathbb{H}_-$, the generating functions (3.3) imply $B_d(q_\phi) = 2M_\phi B_c(q) M_\phi^*$. The ϕ-transformation of q and r_j leads to

$$\sigma(q_\phi) \subset \mathbb{D}, \quad \left\langle\ r_i\ \middle|\ r_j\ \right\rangle_{\frac{\mathrm{d}}{|q|^2}}^{\mathbf{i}\mathbb{R}} = 2\left\langle\ r_{i\phi}\ \middle|\ r_{j\phi}\ \right\rangle_{\frac{\mathrm{d}}{|q_\phi|^2}}^{\mathbb{T}}. \tag{6.1}$$

Thus, with $r_j(z) = z^j$, $R_\phi := \mathrm{row}(\mathfrak{r}_{j\phi})_{j=0}^{n-1}$ and Proposition 5.1 the equations

$$\begin{aligned} B_{\mathbf{i}\mathbb{R}}^{-1}(q) &= \left[\left\langle\ r_i\ \middle|\ r_j\ \right\rangle_{\frac{\mathrm{d}}{2\pi|q|^2}}^{\mathbf{i}\mathbb{R}}\right]_{i,j=0}^{n-1} = 2\left[\left\langle\ r_{i\phi}\ \middle|\ r_{j\phi}\ \right\rangle_{\frac{\mathrm{d}}{2\pi|q_\phi|^2}}^{\mathbb{T}}\right]_{i,j=0}^{n-1} \\ &= 2R_\phi^* B_{\mathbb{T}}^{-1}(q_\phi) R_\phi = 2R_\phi^* B_d^{-1}(q_\phi) R_\phi \\ &= R_\phi^* M_\phi^{-*} B_c^{-1}(q) M_\phi^{-1} R_\phi = R^* B_c^{-1}(q) R \end{aligned}$$

are obtained. Finally, $R = I$ completes the proof of "⇒". Supposed $B_c(q) = B_{\mathbf{i}\mathbb{R}}(q)$ the Bezoutian $B_c(q)$ is positive such that (3.2) implies $\sigma(q) \subset \mathbb{H}_-$. □

As we have just stated, for stable q the (i,j)-entry of $B_d^{-1}(q)$ and $B_c^{-1}(q)$ admits the representation

$$\tfrac{1}{2\pi}\int_\gamma \overline{z}^i z^j |q(z)|^{-2}\mathrm{d}s.$$

Supposing $\sigma(q) \subset \{z \in \mathbb{C} : 0 < \Im(z)\}$ such a representation

$$\tfrac{1}{2\pi\mathbf{i}}\int_{\mathbb{R}} t^{i+j}|q(t)|^{-2}\mathrm{d}t$$

of the (i,j)-entry of $B(q,\overline{q})^{-1}$ have been proved already in [[22], Corollary 6.1]. Here, $\Im(z)$ designates the imaginary part of the complex number z.

To relate G and $B_{\mathbf{i}\mathbb{R}}(q)$ to Lyapunov equations, we generate the corresponding right-hand sides by $\mathfrak{h}$ and by the coefficient vector $\mathfrak{v}$ of the first residue v of q with respect to $\widetilde{\overline{q}}$:

$$v(z) = \psi_n(z)\mathfrak{v}, \quad v(z) := q(z) - a_0\widetilde{\overline{q}}(z), \quad a_0 := (-1)^n \mathfrak{q}_n/\overline{\mathfrak{q}_n}.$$

Note, that $\mathfrak{v}$ coincides with the last column of $B_{\mathbf{i}\mathbb{R}}(q)$.

Proposition 6.2. *Supposed $\sigma(q) \subset \mathbb{H}_-$ the Gramian G and the Bezoutian $B_{\mathbf{i}\mathbb{R}}(q)$ solve, respectively,*

$$C_q^* X + X C_q = -\mathfrak{h}\mathfrak{h}^*, \quad C_q X + X C_q^* = -\mathfrak{v}\mathfrak{v}^*. \tag{6.2}$$

Proof. Using the abbreviations $B := B_{\mathbf{i}\mathbb{R}}(q)$ and $K := H_{\triangleright}(\mathfrak{q}_1, \ldots, \mathfrak{q}_n)$ Barnett's factorization of B reads as

$$I_{\pm} K^* \widetilde{q}(C_q^*).$$

The definition of v implies $\mathfrak{v} = -\widetilde{\overline{q}}(C_q)e_1 a_0$. In addition one concludes

$$K^* C_q^* = \overline{C_q} K^*, \quad (C_q I_{\pm} + I_{\pm}\overline{C_q})\mathfrak{q}_n = \mathfrak{v}(-1)^n e_n^*, \quad K e_n = e_1 \mathfrak{q}_n.$$

Using that identities one gets

$$\begin{aligned} C_q B + B C_q^* &= C_q I_{\pm} K^* \widetilde{q}(C_q^*) + I_{\pm} K^* \widetilde{q}(C_q^*) C_q^* = (C_q I_{\pm} K^* + I_{\pm} K^* C_q^*)\widetilde{q}(C_q^*) \\ &= (C_q I_{\pm} + I_{\pm}\overline{C_q}) K^* \widetilde{q}(C_q^*) = \mathfrak{v}\mathfrak{q}_n^{-1}(-1)^n e_n^* K^* \widetilde{q}(C_q^*) \\ &= \mathfrak{v} \underbrace{\mathfrak{q}_n^{-1}(-1)^n \overline{\mathfrak{q}_n}}_{=\overline{a_0}} e_1^* \widetilde{q}(C_q^*) = -\mathfrak{v}\mathfrak{v}^*. \end{aligned}$$

With $G = R^* B^{-1} R$, $C_q R = R C_q$ and $e_n^* R = \mathfrak{q}_n \mathfrak{h}^*$ one obtains

$$\begin{aligned} C_q^* G + G C_q &= R^*(C_q^* B^{-1} + B^{-1} C_q) R = R^* B^{-1}(B C_q^* + C_q B) B^{-1} R \\ &= -R^* B^{-1} \mathfrak{v}\mathfrak{v}^* B^{-1} R = -R^* e_n \overline{\mathfrak{q}_n^{-1}} \mathfrak{q}_n^{-1} e_n^* R = -\mathfrak{h}\mathfrak{h}^*. \end{aligned}$$

□

Note, that the solution property of $B_c^{-1}(q)$ with respect to

$$C_q^* X + X C_q = -e_n e_n^*$$

is a special case of [[23]. Theorem 4.1].

Consideration of the controllability realization Σ_{co} of h reveals G as the observability Gramian of Σ_{co} with respect to $\mathbf{i}\mathbb{R}$. As a consequence of Theorem 4.4 for real h all coefficients of Q are non-negative. Thus, the relation $G_m^{-1} g_m \in \mathbb{R}_-^m$ can be stated where $\mathbb{R}_- := \{r \in \mathbb{R} : r \leq 0\}$.

As in the unit circle case, ξ can be obtained as solution to a least squares problem in ℓ_2.

Proposition 6.3. *Supposed $\sigma(q) \subset \mathbb{H}_-$ the minimizer ξ of F satisfies $\xi = M_m^{\dagger} V_m$ where*

$$M_m := [h_{ij}]_{i,j=0}^{\infty,m-1}, \quad V_m := \mathrm{col}(h_{im})_{i=0}^{\infty}, \quad \textstyle\sum_{i=0}^{\infty} h_{ij} z^{-(i+1)} := \frac{r_{j\phi}(z)}{q_\phi(z)}.$$

Proof. Parseval's identity states $B_d^{-1}(q_\phi) = H^* H$ where H is generated by $1/q_\phi$. From above we know that $G = 2R_\phi^* B_d^{-1}(q_\phi) R_\phi$. Thus $M_n = H R_\phi$, $2M_n^* M_n = G$ and

$$M_m^{\dagger} V_m = (M_m^* M_m)^{-1} M_m^* V_m = G_m^{-1} g_m = \xi.$$

□

As counter part of Proposition 5.4 we have Proposition 6.4.

Proposition 6.4. *Let for $\sigma(q) \subset \mathbb{H}_-$ and $x \in \mathbb{C}^m$ the ℓ_2-sequence ε be defined by the components of the infinite-dimensional vector $M_m x - V_m$. Supposed $\sigma(Q) \cap \mathbf{i}\mathbb{R} = \emptyset$ we have $\|h - h_x\|_2^{\mathbf{i}\mathbb{R}} \leq \sqrt{2}\|\varepsilon\|_2 / \min_{\mathbf{i}\mathbb{R}}(Q)$.*

Proof. According to Proposition 2.1 it suffices to show $F(x) = \sqrt{2}\|\varepsilon\|_2$. Obviously,

$$\begin{aligned}\frac{p_{x\phi}(z)}{q_\phi(z)} &= \frac{r_{m\phi}(z)}{q_\phi(z)} - \sum_{j=0}^{m-1} \frac{r_{j\phi}(z)}{q_\phi(z)} x_j \\ &= \sum_{i=0}^{\infty} h_{im} z^{-(i+1)} - \sum_{j=0}^{m-1} x_j \sum_{i=0}^{\infty} h_{ij} z^{-(i+1)} = \sum_{i=0}^{\infty} \varepsilon_i z^{-(i+1)}.\end{aligned}$$

Thus, transformation formula (6.1) and Parseval's identity lead to

$$F(x)^2 = \tfrac{1}{2\pi} \int_{\mathbb{R}} \left| \tfrac{p_x(\mathbf{i}t)}{q(\mathbf{i}t)} \right|^2 \mathrm{d}t = \tfrac{1}{\pi} \int_0^{2\pi} \left| \tfrac{p_{x\phi}(e^{\mathbf{i}t})}{q_\phi(e^{\mathbf{i}t})} \right|^2 \mathrm{d}t = 2 \textstyle\sum_{i=0}^{\infty} |\varepsilon_i|^2. \qquad \square$$

For the line case only $\sigma(h_\xi) \subset \overline{\mathbb{C}_\ell}$ is stated. Numerical experiences collected as in Example 2 show that actually the relation $\sigma(Q) \subset \gamma$ holds, supposed the degree m of Q is less than δ where

$$\delta := n - \kappa, \quad n := \deg q, \quad \kappa := \deg p, \quad h = p/q.$$

In addition, for $m = \delta$ the polynomial Q becomes stable. To suppress technical difficulties, we prove these observations only for real h and $\gamma = \mathbf{i}\mathbb{R}$. Moreover, for $m \in \{\delta + 1, \ldots, n-1\}$ and accidently generated p and stable q the polynomial Q becomes always stable. In [11] it is proved that this situation is generic. For acquainting the reader with the proof idea for the case $m \le \delta$ we consider the Lyapunov equations (6.2). As a consequence the entries g_{ij} of G satisfy

$$g_{ij} + g_{i-1,j+1} = -h_{i-1}\overline{h_j}, \quad i = 1, \ldots, n, \quad j = 0, \ldots, n-1. \tag{6.3}$$

Thus, setting $\mathfrak{g}_i := g_{ii}$ and $h_{ij} := h_i h_j$ for real h the coincidence of $[G_4|g_4]$ with

$$\left[\begin{array}{cccc|c} \mathfrak{g}_0 & -\frac{h_{00}}{2} & -\mathfrak{g}_1 - h_{01} & \frac{h_{11}}{2} - h_{02} & \mathfrak{g}_2 - h_{03} + h_{12} \\ -\frac{h_{00}}{2} & \mathfrak{g}_1 & -\frac{h_{11}}{2} & -\mathfrak{g}_2 - h_{12} & \frac{h_{22}}{2} - h_{13} \\ -\mathfrak{g}_1 - h_{01} & -\frac{h_{11}}{2} & \mathfrak{g}_2 & -\frac{h_{22}}{2} & -\mathfrak{g}_3 - h_{23} \\ \frac{h_{11}}{2} - h_{02} & -\mathfrak{g}_2 - h_{12} & -\frac{h_{22}}{2} & \mathfrak{g}_3 & -\frac{h_{33}}{2} \end{array}\right] \tag{6.4}$$

is obtained. Consequently, for $h_0 = h_1 = h_2 = 0$ the product $G_4 I_\pm$ becomes a Hankel matrix generated by the vector $[\mathfrak{g}_0, 0, -\mathfrak{g}_1, 0, \mathfrak{g}_2, 0, -\mathfrak{g}_3]^{\mathrm{T}}$ and g_4 is equal to $[\mathfrak{g}_2, 0, -\mathfrak{g}_3, \beta]^{\mathrm{T}}$, $\beta := -h_{33}/2$. Here, depending on the vector $\mathrm{col}(h_i)_{i=0}^{2m-2}$ we set

$$H(h_0, \ldots, h_{2m-2}) := [h_{i+j}]_{i,j=0}^{m-1}.$$

The zero distribution in the coefficient matrix and the right-hand side of the equation system $G_m x = g_m$ can be used now to give a mathematical foundation for our observations. But at first for $m = \delta$ we relate the Gramian G_δ to the Bezoutian $B_c(Q)$.

Proposition 6.5. *Let* $\xi \in \mathbb{R}^m$ *and* $\operatorname{col}(\mathfrak{g}_i)_{i=0}^{m-1} \in \mathbb{R}^m$ *be related by* $G_m\xi = g_m(\beta)$ *where*

$$G_m := HI_{\pm},\ H := H(f_0, 0, f_1, 0, f_2, \ldots, f_{m-2}, 0, f_{m-1}),\ f_i := (-1)^i\mathfrak{g}_i,\ \beta \in \mathbb{R}$$

and

$$g_m(\beta) := \begin{cases} \left[0, (-1)^{\frac{m-1}{2}}\mathfrak{g}_{\frac{m+1}{2}}, 0, \ldots, 0, \mathfrak{g}_{m-2}, 0, -\mathfrak{g}_{m-1}, \beta\right]^{\mathrm{T}}, & \text{if } m \text{ is odd,} \\ \left[(-1)^{\frac{m}{2}}\mathfrak{g}_{\frac{m}{2}}, 0, \ldots, 0, \mathfrak{g}_{m-2}, 0, -\mathfrak{g}_{m-1}, \beta\right]^{\mathrm{T}}, & \text{if } m \text{ is even.} \end{cases}$$

Then for invertible G_m *the polynomial* $Q(z) := z^m - \psi_m(z)\xi$ *satisfies*

$$G_m B_c(Q) = -2\beta I.$$

Proof. According to [[24], p. 19, Theorem 1.1′] we have

$$\psi_m(x)H^{-1}\psi_m^{\mathrm{T}}(y)(x-y) = T_\alpha(x)S(y) - S(x)T_\alpha(y)$$

where the polynomials T_α and S are defined by

$$T_\alpha(z) := z^m - \psi_m(z)H^{-1}f(\alpha), \quad S(z) := \psi_m(z)H^{-1}e_m,$$
$$f(\alpha) := \begin{cases} \left[0, f_{(m+1)/2}, \ldots, 0, f_{m-2}, 0, f_{m-1}, \alpha\right]^{\mathrm{T}}, & \text{if } m \text{ is odd,} \\ \left[f_{m/2}, 0, \ldots, 0, f_{m-2}, 0, f_{m-1}, \alpha\right]^{\mathrm{T}}, & \text{if } m \text{ is even.} \end{cases}$$

Here, α represents an arbitrary real number. The substitution $x := -x$ and the equation $G_m = HI_{\pm}$ lead to

$$B(x,y) := \psi_m(x)G_m^{-1}\psi_m^{\mathrm{T}}(y)(x+y) = \widetilde{S}(x)T_\alpha(y) - \widetilde{T}_\alpha(x)S(y), \quad \widetilde{S}(z) := S(-z).$$

For odd m we have $I_{\pm}g_m(\beta) = f(\beta)$. Thus $Q(z) = z^m - \psi_m(z)H^{-1}f(\beta) = T_\beta(z)$. Since the distribution of the vanishing entries in H will be preserved by its inverse, together with the zero distribution in $f(\beta)$ it follows that the product $(I_{\pm} + I)H^{-1}(f(\beta) - e_m\beta)$ is equal to the zero vector. Thus

$$\begin{aligned} \widetilde{Q}(z) + Q(z) &= T_\beta(-z) + T_\beta(z) = -\psi_m(z)(I_{\pm} + I)H^{-1}f(\beta) \\ &= -\psi_m(z)(I_{\pm} + I)H^{-1}(f(\beta) - e_m\beta + e_m\beta) \\ &= -2\beta\psi_m(z)H^{-1}e_m = -2\beta S(z). \end{aligned}$$

On the same way for even m we get $Q = T_{-\beta}$ and $\widetilde{Q} - Q = -2\beta S$. Thus

$$\begin{aligned} -2\beta B(x,y) &= -2\beta\widetilde{S}(x)T_{\pm\beta}(y) - \widetilde{T}_{\pm\beta}(x)(-2\beta S(y)) \\ &= (Q(x) \mp \widetilde{Q}(x))Q(y) - \widetilde{Q}(x)(\widetilde{Q}(y) \mp Q(y)) \\ &= Q(x)Q(y) - \widetilde{Q}(x)\widetilde{Q}(y). \end{aligned}$$

Finally, in view of the generating function (3.3) the statement follows. □

Proposition 6.6. *Let* $h \in \mathbb{R}(z)$ *be stable. Then for* $m < \delta$ *we have* $\sigma(Q) \subset \mathbf{i}\mathbb{R}$ *and for* $m = \delta$ *the polynomial* Q *becomes stable.*

Proof. Due to $\delta = \deg q - \deg p$ the Laurent coefficients $h_0, \dots, h_{\delta-2}$ of h vanish. Hence, in view of (6.3) the entries of G_δ satisfy

$$g_{ij} + g_{i-1,j+1} = 0, \quad i = 1, \dots, \delta - 1, \quad j = 0, \dots, \delta - 2. \tag{6.5}$$

That equation establishes a recurrence relation within the entries along every anti-diagonal of G_δ. Here, the k^{th} anti-diagonal A_k is formed by all entries g_{ij} for which $i + j = k$. For $k = 2i - 1$ equation (6.5) passes over to $g_{i,i-1} + g_{i-1,i} = 0$. Since h is real, we have $g_{ij} = g_{ji}$. Thus, $g_{i,i-1} = 0$ holds. Consequently, all entries of A_{2i-1} vanish. For $k = 2i$ equation (6.5) passes over to $g_{ii} + g_{i-1,i+1} = 0$. Thus, all entries of A_{2i} are equal to $\mathfrak{g}_i$ or $-\mathfrak{g}_i$. Collecting these observations we get

$$G_\delta = H(\mathfrak{g}_0, 0, -\mathfrak{g}_1, 0, \dots, 0, (-1)^{\delta-1}\mathfrak{g}_{\delta-1})I_\pm.$$

For $m < \delta$ the structure of G_δ leads to a polynomial $Q(z) := z^m - \psi_m(z)G_m^{-1}g_m$ whose coefficients of all even or of all odd powers vanish. In addition, Theorem 4.4 states $\sigma(Q) \subset \mathbb{H}_- \cup \mathbf{i}\mathbb{R}$. Since all zeros of such a polynomial are located on $\mathbf{i}\mathbb{R}$, the first statement is proved.

For $m = \delta$ the vector ξ satisfies $G_\delta \xi = g_\delta(-h_{\delta-1}^2/2)$. Thus, ξ and $\operatorname{col}(\mathfrak{g}_i)_{i=0}^{\delta-1}$ are related as in Proposition 6.5. In combination with $h_{\delta-1} = p_\kappa$ the matrix equation

$$G_\delta B_c(Q) = p_\kappa^2 I$$

is obtained. Since h is stable, the Gramian G_δ is positive. In combination with $p_\kappa \neq 0$ the same holds for $B_c(Q)$. Finally, (3.2) yields the stability of Q. □

As mentioned above, the statement of Proposition 6.6 for $m < \delta$ is a simple consequence of Fejér's convex hull theorem, the case $m = \delta$ have been treated here in detail and the case $m \in \{\delta + 1, \dots, n - 1\}$ is subject of [11].

7. Computation of h_ξ via its minimal realizations

There is a third approach to get h_ξ. Depending on a minimal realization $\Sigma := (A, B, C)$ of h we set $K := \mathscr{K}(A, B)$ and consider the observability Gramian $\mathscr{Q}$ of Σ with respect to γ.

Proposition 7.1. *For stable h the Gramians G and $\mathscr{Q}$ are related by $G = K^*\mathscr{Q}K$.*

Proof. The minimality of Σ provides the invertibility of K and the stability of h the existence of $\mathscr{Q}$. Immediately by definition the equations

$$CK = \mathfrak{h}^*, \quad AK = KC_q$$

follow. It turns out that (5.3) and (6.2) can be generalized to $L_{\overline{\gamma}}(C_q^*, G) = -\Delta\mathfrak{h}\mathfrak{h}^*$ that means G represents the observability Gramian of the controllability realization Σ_{co} with respect to γ. Thus, we have

$$L_{\overline{\gamma}}(A^*, K^{-*}GK) = K^{-*}L_{\overline{\gamma}}(C_q^*, G)K^{-1} = -\Delta K^{-*}\mathfrak{h}\mathfrak{h}^*K^{-1} = -\Delta C^*C.$$

Finally, the uniqueness of $\mathscr{Q}$ provides the statement. □

Assuming that $A = \mathrm{diag}(\alpha_i)_{i=0}^{n-1}$, $B = \mathrm{col}(1)_{i=1}^{n}$, and $C = \mathrm{row}(c_j)_{j=0}^{n-1}$, we define the Vandermonde matrix V and the Cauchy matrix Ω by

$$V := [\alpha_i^j]_{i,j=0}^{n-1}, \quad \Omega := \left[\frac{-\Delta\overline{c_i}c_j}{p_\gamma(\alpha_j, \overline{\alpha_i})}\right]_{i,j=0}^{n-1}. \tag{7.1}$$

Then $K = V$ and $\mathscr{Q} = \Omega$ such that supposed $\alpha_i \in \mathbb{C}_\ell$ the Gramian G admits the factorization $G = V^*\Omega V$. Supposed that all poles of h are simple, we call Ω the Cauchy matrix generated by h and γ. For numerical examples which are generated by prescribed $\alpha_i \in \mathbb{C}_\ell$ and $c_i \in \mathbb{C}$ the matrix G can be obtained by the product $V^*\Omega V$. In particular, for $\alpha, \beta \in \mathbb{C}_\ell$ the identity

$$\int_{\mathscr{I}} \frac{|\dot{w}(t)|\mathrm{d}t}{\left(\overline{w(t)-\alpha}\right)(w(t)-\beta)} = \frac{-2\pi\Delta}{p_\gamma(\beta,\overline{\alpha})}, \quad \mathscr{I} \in \{[0, 2\pi], \mathbb{R}\}, \quad \gamma \in \{\mathbb{T}, \mathbf{i}\mathbb{R}\}$$

holds where $w(t)$ is defined as in the proof of Proposition 4.3.

With the abbreviation $V := \mathscr{K}_m(A, B)$ we get $V^*\mathscr{Q}V = G_m$ meaning that $V^*\mathscr{Q}V$ is invertible and appropriate to define $Z := (V^*\mathscr{Q}V)^{-1}V^*\mathscr{Q}$.

Proposition 7.2. *The triple* (ZAV, ZB, CV) *realizes* h_ξ.

Proof. We have

$$ZAV = (V^*\mathscr{Q}V)^{-1}V^*\mathscr{Q}[AB, \ldots, A^mB] = [e_2, \ldots, e_m, (V^*\mathscr{Q}V)^{-1}V^*\mathscr{Q}A^mB].$$

Together with $G_m = V^*\mathscr{Q}V$ and $g_m = V^*\mathscr{Q}A^mB$ one gets $ZAV = C_Q$. Finally,

$$ZB = (V^*\mathscr{Q}V)^{-1}V^*\mathscr{Q}B = e_1, \quad CV = [h_0, \ldots, h_{m-1}]$$

thus, above triple represents the controllability realization of h_ξ. □

8. The construction of h_ξ^η

The renouncement on the disappearance of the first m Laurent coefficients of $h - h_\xi$ admits to proceed with a second optimization step resulting in a rational function h_ξ^η with

$$\|h - h_\xi^\eta\|_2^\gamma \approx \|h - h_{\mathrm{opt}}\|_2^\gamma.$$

Instead to apply (2.2), the numerator polynomial $\mathscr{P}$ of h_ξ^x will be defined now via the minimizer $\eta \in \mathbb{C}^m$ of

$$F : \mathbb{C}^m \to \mathbb{R}, \quad F(x) := \|h_\xi^x - h\|_2^\gamma, \quad h_\xi^x(z) := \tfrac{1}{Q(z)}\textstyle\sum_{k=0}^{m-1} x_k z^k$$

such that

$$h_\xi^\eta := \mathscr{P}/Q, \quad \mathscr{P}(z) := \psi_m(z)\eta.$$

As in Section 2 the minimizer ξ, the minimizer η turns out to be the solution to

$$\left[\left\langle \frac{z^i}{Q} \,\middle|\, \frac{z^j}{Q} \right\rangle_2^\gamma\right]_{i,j=0}^{m-1} x = W, \quad W := \mathrm{col}\left(\left\langle \frac{z^i}{Q} \,\middle|\, h \right\rangle_2^\gamma\right)_{i=0}^{m-1}. \tag{8.1}$$

Therefore, for the determination of η it suffices to find the stable polynomial Q and to apply (8.1). In addition, the definition of $B_\gamma(Q)$ implies $\eta = B_\gamma(Q)W$. That

approach becomes essential for the case where the first m Laurent coefficients of h are small in comparison with the subsequent coefficients. Then according to (2.2) the numerator of h_ξ becomes approximately the zero polynomial, a circumstance which does not lead to satisfying approximation results. Here, we say that a complex number is small, if its magnitude is small. For $h := 1/q$, $\gamma := \mathbb{T}$ and $m := n-1$ the approximant h_ξ^η can be expressed explicitly in terms of q:

$$h_\xi^\eta(z) = \frac{|\mathfrak{q}_n|^2 - |\mathfrak{q}_0|^2 + q(z)\overline{\mathfrak{q}_0} - \widehat{\overline{q}}(z)\mathfrak{q}_n}{\mathfrak{q}_n(q(z)\overline{\mathfrak{q}_n} - \widehat{\overline{q}}(z)\mathfrak{q}_0)}.$$

Note, that z represents a common divisor of the numerator and the denominator such that the degree of h_ξ^η is actually equal to $n-1$. Moreover, that formula reveals an interesting interpolation property of h_ξ^η. Let β be a pole of h_ξ^η. Then its denominator vanishes in β such that $q(\beta)\overline{\mathfrak{q}_n} = \widehat{\overline{q}}(\beta)\mathfrak{q}_0$. Supposed $\beta \neq 0$ together with $\widehat{q}(z) = z^n q(1/z)$ that equation implies

$$\mathfrak{q}_n(q(\widehat{\beta})\overline{\mathfrak{q}_n} - \widehat{\overline{q}}(\widehat{\beta})\mathfrak{q}_0) = q(\widehat{\beta})(|\mathfrak{q}_n|^2 - |\mathfrak{q}_0|^2),\ q(\widehat{\beta})\overline{\mathfrak{q}_0} - \widehat{\overline{q}}(\widehat{\beta})\mathfrak{q}_n = 0,\ \widehat{\beta} := \overline{\beta^{-1}}.$$

Thus, $h_\xi^\eta(\widehat{\beta}) = q(\widehat{\beta})^{-1} = h(\widehat{\beta})$ holds. The coincidence of h_ξ^η and h on the mirror image set $\phi_\gamma(\sigma(h_\xi^\eta))$ along γ will be maintained, if one sets

$$\phi_\gamma(z) := -\frac{\gamma_{00} + \gamma_{01}\overline{z}}{\gamma_{10} + \gamma_{11}\overline{z}}.$$

Proposition 8.1. *Let h be stable and $\beta \in \sigma(h_\xi^\eta)$. Then $h(\phi_\gamma(\beta)) = h_\xi^\eta(\phi_\gamma(\beta))$.*

Proof. Since h and h_ξ^η are strictly proper for infinite $\phi_\gamma(\beta)$ the statement holds. Suppose $|\phi_\gamma(\beta)| < \infty$. To avoid technical details, we prove the implication only for real h and $\gamma = \mathbb{T}$ meaning that $\phi_\gamma(\beta) = \overline{\beta^{-1}}$. Let $Q(z) := \sum_{k=0}^m \mathfrak{Q}_k z^k$ be an arbitrary real stable polynomial of degree m. Combination of Barnett's factorization (5.2) with $JB_d(Q)J = B_d(Q)$ leads to

$$B_d(Q) = J\widehat{Q}(C_Q)H_\triangleright, \quad H_\triangleright := H_\triangleright(\mathfrak{Q}_1, \ldots, \mathfrak{Q}_m).$$

The Hankel matrix Θ generated by $1/Q$ fulfills $\Theta^* = H_\triangleright^{-1}\mathrm{row}(C_Q^i e_1)_{i=0}^\infty$, thus the vector W defined as in (8.1) satisfies

$$W = \Theta^*\mathrm{col}(h_i)_{i=0}^\infty = H_\triangleright^{-1}\mathrm{row}(C_Q^i e_1)_{i=0}^\infty \mathrm{col}(h_i)_{i=0}^\infty = H_\triangleright^{-1}\sum_{i=0}^\infty C_Q^i e_1 h_i.$$

First, we consider for $k > 0$ the special case $h(z) := z^{-k}$. Then all Laurent coefficients h_i of $h(z)$ vanish excepted h_{k-1} which is equal to 1. Consequently, W simplifies to

$$W = H_\triangleright^{-1} C_Q^{k-1} e_1$$

such that the coefficient vector η_k of the numerator polynomial $\mathscr{P}_k$ of $h_\xi^{\eta_k}$ generated by Q and z^{-k} admits the representation

$$\eta_k = B_d(Q)W = J\widehat{Q}(C_Q)C_Q^{k-1}e_1.$$

The assumption $Q(\beta)=0$ implies now $\psi_m(\beta)C_Q^k=\beta^k\psi_m(\beta)$. Thus,

$$\begin{aligned}\mathscr{P}_k\left(\beta^{-1}\right) &= \beta^{1-m}\psi_m(\beta)J\eta_k=\beta^{1-m}\psi_m(\beta)\widehat{Q}(C_Q)C_Q^{k-1}e_1\\ &= \beta Q(\beta^{-1})\psi_m(\beta)C_Q^{k-1}e_1=\beta^kQ(\beta^{-1})\end{aligned}$$

proving the statement for $h(z):=z^{-k}$. Finally, let $h(z):=\sum_{k=0}^{\infty}h_k/z^{k+1}$ be real and stable. Then the coefficient vector η of $\mathscr{P}$ admits the representation

$$\eta=\textstyle\sum_{k=1}^{\infty}\eta_kh_{k-1}$$

such that

$$h(\beta^{-1})Q(\beta^{-1})=\textstyle\sum_{k=1}^{\infty}h_{k-1}\beta^kQ(\beta^{-1})=\sum_{k=1}^{\infty}\mathscr{P}_k(\beta^{-1})h_{k-1}=\mathscr{P}(\beta^{-1}).$$

Because $h,Q,\mathscr{P}$ are real the equation $h(\overline{\beta^{-1}})Q(\overline{\beta^{-1}})=\mathscr{P}(\overline{\beta^{-1}})$ holds. □

In the case where all zeros of Q are simple, the computation of η via $B_\gamma(Q)W$ can be replaced by solving the interpolation problem

$$\mathscr{P}(\gamma_i)=h(\gamma_i)Q(\gamma_i),\quad \gamma_i:=\phi_\gamma(\beta_i),\quad i=0,\dots,m-1. \tag{8.2}$$

Example 1 and 4 show the improved performance of h_ξ^η in comparison with h_ξ.

9. The limit behavior of Q

For simplification let $\gamma\in\{\mathbb{T},\mathbf{i}\mathbb{R}\}$, $h\in\mathbb{R}(s)$ and $q(z)=\prod_{i=0}^{n-1}(z-\alpha_i)$. To study the limit behavior of Q, remember its sum representation

$$Q(z)=\textstyle\sum_{j=m+1}^{n}\zeta_jm_j(z)$$

provided as in Proposition 4.1, the equation $R=p(C_q)$ and the definition of the polynomial v and the vector $\mathfrak{w}$ as in the end of Section 3. As usual we set $q':=\frac{\mathrm{d}}{\mathrm{d}z}q$ and assume the stability of h. Vanishing $q(\alpha)$ provides $\psi_n(\alpha)p(C_q)=p(\alpha)\psi_n(\alpha)$. Therefore, $w(z):=\psi_n(z)\mathfrak{w}$ satisfies the interpolation condition $w(\alpha)=v(\alpha)/p(\alpha)$. We establish now a recurrence relation within the numbers $m_j(\alpha)$ and express $m_n(\alpha)$ explicitly. With the definition

$$q_{\phi_\gamma}(z):=q(\phi_\gamma(\overline{z}))(\gamma_{10}+\gamma_{11}z)^n$$

we get $q_{\phi_{\mathbb{T}}}=\widehat{q}$ and $q_{\phi_{\mathbf{i}\mathbb{R}}}=\widetilde{q}$. In addition, we set

$$K(x,y):=-\frac{r(x)r(y)}{p_\gamma(x,y)},\quad r:=\frac{q_{\phi_\gamma}}{q'p}.$$

Proposition 9.1. *For $\alpha\in\sigma(q)$ the numbers $(m_j(\alpha))_{j=0}^n$ satisfy the recurrence relation*

$$\begin{aligned}\gamma=\mathbb{T}:\quad m_{j+1}(\alpha) &= \alpha m_j(\alpha)-\alpha\mathfrak{q}_jm_n(\alpha)+\mathfrak{w}_j\widehat{q}(\alpha)/p(\alpha)\\ \gamma=\mathbf{i}\mathbb{R}:\quad m_j(\alpha) &= m_n(\alpha)\mathfrak{q}_j-\alpha m_{j+1}(\alpha)+(-1)^{n+1}\mathfrak{w}_j\widetilde{q}(\alpha)/p(\alpha)\end{aligned} \tag{9.1}$$

where $m_0(z)\equiv 0$. In the case where all zeros α_i of q are simple, the polynomial $m_n(z)$ satisfies the interpolation conditions

$$m_n(\alpha_i)=q'(\alpha_i)\textstyle\sum_{j=0}^{n-1}K(\alpha_i,\overline{\alpha_j}),\quad i=0,\dots,n-1. \tag{9.2}$$

Proof. In the circle case the equation $C_qMC_q^* - M = -\mathfrak{w}\mathfrak{w}^*$ holds. Thus the companion structure of C_q provides for the columns of M the equation

$$\mathfrak{m}_{j+1} = C_q(\mathfrak{m}_j - \mathfrak{m}_n\mathfrak{q}_j) + \mathfrak{w}\mathfrak{w}_j.$$

Left multiplication with $\psi_n(\alpha)$ yields

$$m_{j+1}(\alpha) = \alpha(m_j(\alpha) - m_n(\alpha)\mathfrak{q}_j) + w(\alpha)\mathfrak{w}_j.$$

Since $w(\alpha) = v(\alpha)/p(\alpha)$ and $v(\alpha) = \widehat{q}(\alpha)$ we have $w(\alpha) = \widehat{q}(\alpha)/p(\alpha)$.

In the line case M satisfies $C_qM + MC_q^* = -\mathfrak{w}\mathfrak{w}^*$. Thus the companion structure of C_q provides for the columns of M the equation

$$\mathfrak{m}_j = \mathfrak{m}_n\mathfrak{q}_j - C_q\mathfrak{m}_{j+1} - \mathfrak{w}\mathfrak{w}_j.$$

Left multiplication with $\psi_n(\alpha)$ yields

$$m_j(\alpha) = m_n(\alpha)\mathfrak{q}_j - \alpha m_{j+1}(\alpha) - w(\alpha)\mathfrak{w}_j.$$

Since $w(\alpha) = v(\alpha)/p(\alpha)$ and $v(\alpha) = (-1)^{n+1}\widetilde{q}(\alpha)$ we have $w(\alpha) = (-1)^{n+1}\frac{\widetilde{q}(\alpha)}{p(\alpha)}$.

To prove in the circle case the validity of the asserted interpolation conditions, we introduce the diagonal matrix

$$\mathscr{D} := \operatorname{diag}\left(\widehat{q}(\alpha_i)\right)_{i=0}^{n-1}.$$

Then the generating function of $B_d(q)$ yields

$$VB_d(q)V^* = \mathscr{D}\Omega\mathscr{D}^*, \quad V := [\alpha_i^j]_{i,j=0}^{n-1}, \quad \Omega := [(1-\alpha_i\overline{\alpha_j})^{-1}]_{i,j=0}^{n-1}.$$

Furthermore, R admits the factorization $V^{-1}\operatorname{diag}(p(\alpha_i))_{i=0}^{n-1}V$ such that

$$M = G^{-1} = R^{-1}B_d(q)R^{-*} = V^{-1}D\Omega D^*V^{-*}, \quad D := \operatorname{diag}\left(\frac{\widehat{q}(\alpha_i)}{p(\alpha_i)}\right)_{i=0}^{n-1}. \tag{9.3}$$

Since the last column of V^{-*} is equal to $\operatorname{col}(q'(\overline{\alpha_i})^{-1})_{i=0}^{n-1}$ and $\psi_n(\alpha_i) = e_i^*V$ we have

$$\begin{aligned} m_n(\alpha_i) = \;& \psi_n(\alpha_i)Me_n = e_i^*D\Omega D^*\operatorname{col}(q'(\overline{\alpha_i})^{-1})_{i=0}^{n-1} \\ = \;& \frac{\widehat{q}(\alpha_i)}{p(\alpha_i)}e_i^*\Omega\operatorname{col}\left(\frac{\widehat{q}(\overline{\alpha_j})}{p(\overline{\alpha_j})q'(\overline{\alpha_j})}\right)_{j=0}^{n-1} = \frac{\widehat{q}(\alpha_i)}{p(\alpha_i)}\sum_{j=0}^{n-1}\frac{\widehat{q}(\overline{\alpha_j})}{(1-\alpha_i\overline{\alpha_j})p(\overline{\alpha_j})q'(\overline{\alpha_j})}. \end{aligned}$$

The proof of the statement with respect to $\mathbf{i}\mathbb{R}$ is left to the reader. □

We describe now the influence of p on $\sigma(h_\xi)$. The statements are formulated and proved only for the unit circle, but to ensure their validity for the half-plane Example 3 concerns that case. Instead of considering the variation of $\mathfrak{p}$, we consider the variation of $\varphi := p(\alpha) \in \mathbb{C}$, $\alpha \in \sigma(q)$. Since the associated partial fraction coefficient reads as $\varphi/q'(\alpha)$, the case where that coefficient tends to infinity is in question. The dependency of a function f and a matrix M on φ will be designated by $f(.,\varphi)$ and M_φ, respectively. In addition, we use the abbreviation $M(k,\ell) := [m_{ij}]_{i,j=k}^{\ell}$. For simplification of the proofs we assume $\alpha_i \neq \alpha_j$.

Proposition 9.2. *Let $\alpha \in \sigma(q) \subset \mathbb{D}$. Then α is a zero of $\lim_{\varphi\to\infty} Q(.,\varphi)$.*

Proof. W.l.o.g. let $\alpha := \alpha_0$. It suffices to show the existence of the limit vector $\lim_{\varphi\to\infty}\zeta_\varphi$ arising in the sum representation of $Q(.,\varphi)$ and that α becomes a zero of all polynomials $m_j(.,\varphi)$ as $\varphi\to\infty$. According to Proposition 9.1 we have

$$m_0(\alpha,\varphi)=0,\quad \lim_{\varphi\to\infty} m_n(\alpha,\varphi)=q'(\alpha)\sum_{j=0}^{n-1}\underbrace{\lim_{\varphi\to\infty}K(\alpha,\overline{\alpha_j},\varphi)}_{=0}=0.$$

The recurrence relation (9.1) implies $\lim_{\varphi\to\infty} m_j(\alpha,\varphi)=0$. In factorization (9.3) only D depends on φ. Therefore,

$$M_\varphi = V^{-1}D_\varphi\Omega D_\varphi^* V^{-*}.$$

Since $\alpha_i\neq\alpha_j$ and $\sigma(q)\subset\mathbb{D}$ the Vandermonde matrix V is regular and the Cauchy matrix Ω is positive. Let

$$\left[\ \mathfrak{c}_1\ \middle|\ \mathrm{col}(\mathfrak{z}_i)_{i=1}^n\ \right] := \mathrm{row}(\mathfrak{c}_j)_{j=1}^n := V^{-1},\quad \mathfrak{z}_i^*\in\mathbb{C}^{n-1},\ \mathfrak{c}_j\in\mathbb{C}^n.$$

Since the $(1,1)$-entry of D_φ tends to zero as $\varphi\to\infty$ the limit $M_\infty := \lim_{\varphi\to\infty} M_\varphi$ satisfies

$$M_\infty = \mathrm{col}(\mathfrak{z}_i)_{i=1}^n D(2,n)\Omega(2,n)D(2,n)^*\mathrm{row}(\mathfrak{z}_i^*)_{i=1}^n.$$

Since $e_1^*Ve_1=1$ Cramer's rule implies the independence of $\{\mathfrak{z}_2^*,\dots,\mathfrak{z}_n^*\}$. Thus, for $k>1$ the sub-matrices

$$M_\infty(k,n) = \mathrm{col}(\mathfrak{z}_i)_{i=k}^n D(2,n)\Omega(2,n)D(2,n)^*\mathrm{row}(\mathfrak{z}_i^*)_{i=k}^n$$

are invertible. Finally, $\lim_{\varphi\to\infty}\zeta_\varphi = M_\infty(m+1,n)^{-1}e_1$. □

We observe now the behavior of $\sigma(Q(.,\alpha))$ if the conjugate complex pair $\{\alpha,\overline{\alpha}\}\subset\sigma(q)\subset\mathbb{D}$ tends to the dotted boundary $\mathbb{T}_p := \mathbb{T}\setminus\{\sigma(p)\cup\{\pm1\}\}$.

Proposition 9.3. *For $m>1$ and $\alpha\in\sigma(q)\subset\mathbb{D}$ the relation $\alpha\in\lim_{\alpha\to\mathbb{T}_p}\sigma(Q(.,\alpha))$ holds.*

Proof. It suffices to show that $\lim_{\alpha\to\mathbb{T}_p} m_j(\alpha)=0$ and that for $k>2$ the limit $\lim_{\alpha\to\mathbb{T}_p} M_\alpha(k,n)$ is invertible. W.l.o.g. let $\{\alpha_0,\alpha_1\}=\{\alpha,\overline{\alpha}\}$. According to Proposition 9.1 we have

$$\lim_{\alpha_0\to\mathbb{T}_p} m_n(\alpha_0) = q'(\alpha_0)\textstyle\sum_{j=0}^{n-1}\lim_{\alpha_0\to\mathbb{T}_p}K(\alpha_0,\overline{\alpha_j})$$

$$K(\alpha_0,\overline{\alpha_j}) = \frac{\widehat{q}(\overline{\alpha_j})\prod_{i=0,i\neq j}^{n-1}(1-\alpha_0\overline{\alpha_i})}{q'(\alpha_0)p(\alpha_0)q'(\overline{\alpha_j})p(\overline{\alpha_j})}.$$

Because $\widehat{q}(z)=\prod_{i=0}^{n-1}(1-z\overline{\alpha_i})$ and $\overline{\alpha_1}=\alpha_0$, for $j=0$ the second factor of $\widehat{q}(\overline{\alpha_j})$, and for $j\neq 0$ the first factor of $\prod_{i=0,i\neq j}^{n-1}(1-\alpha_0\overline{\alpha_i})$ is equal to $1-|\alpha_0|^2$. Thus $K(\alpha_0,\overline{\alpha_j})\to 0$ as $\alpha_0\to\mathbb{T}_p$. Regarding (9.1) by induction $\lim_{\alpha\to\mathbb{T}_p} m_j(\alpha)=0$ follows.

To show the invertibility of $\lim_{\alpha\to\mathbb{T}_p} M_\alpha(k,n)$ observe that all factors in (9.3) depend on α:

$$M_\alpha = V_\alpha^{-1}D_\alpha\Omega_\alpha D_\alpha^* V_\alpha^{-*}.$$

Obviously, for $\alpha \to \mathbb{T}_p$ the both first main diagonal entries of D_α vanish. The $(1,1)$- and the $(2,2)$-entry of Ω_α are equal to $1/(1-|\alpha|^2)$. Thus, the corresponding entries of $D_\alpha \Omega_\alpha D_\alpha^*$ coincide with

$$\frac{\widehat{q}(\alpha)\widehat{q}(\overline{\alpha})}{p(\alpha)(1-|\alpha|^2)p(\overline{\alpha})}.$$

Since $\widehat{q}(\alpha)\widehat{q}(\overline{\alpha})$ possesses the factor $(1-|\alpha|^2)^2$ the limit $\lim_{\alpha\to\mathbb{T}} \frac{\widehat{q}(\alpha)\widehat{q}(\overline{\alpha})}{1-|\alpha|^2}$ vanishes. Consequently, $M_{\mathbb{T}} := \lim_{\alpha\to\mathbb{T}_p} M_\alpha$ satisfies

$$M_{\mathbb{T}} = \mathrm{row}(\mathfrak{c}_j)_{j=3}^n D(3,n)\Omega(3,n)D(3,n)^* \mathrm{row}(\mathfrak{c}_j)_{j=3}^{n*},\ \mathrm{row}(\mathfrak{c}_j)_{j=1}^n := (\lim_{\alpha\to\mathbb{T}_p} V_\alpha)^{-1}.$$

Note that $\pm 1 \notin \mathbb{T}_p$ ensures the invertibility of $\lim_{\alpha\to\mathbb{T}_p} V_\alpha$. Finally, the invertibility of the main sections $M_{\mathbb{T}}(k,n)$, $k > 2$, is obtained as in the proof of Proposition 9.2. □

For $\gamma = \mathbf{i}\mathbb{R}$ Example 3 illustrates the validity of Propositions 9.2 and 9.3.

10. Bilinear transformation between $\mathbb{D}$ and $\mathbb{H}_-$

The least squares method for the discrete case can be applied to the continues case as follows. Let $\phi(z) := \frac{z-1}{z+1}$ and h_ϕ be defined by

$$h_\phi(z) := h(\phi(z)).$$

For $1 \notin \sigma(A)$ the identity

$$(\phi(z)I - A)^{-1} = (I-A)^{-1} + 2(I-A)^{-2}(zI - (I+A)(I-A)^{-1})^{-1}$$

holds. Due to $\sigma(A_c) \subset \mathbb{H}_-$ we have $1 \notin \sigma(A_c)$. Thus, supposed (A_c, B_c, C_c) realizes h, the rational function h_ϕ admits the representation

$$h_\phi(z) = \underbrace{C_c(I-A_c)^{-1}B_c}_{=h(1)} + \underbrace{2C_c(I-A_c)^{-2}}_{=:C_d}(zI - \underbrace{(I+A_c)(I-A_c)^{-1}}_{=:A_d})^{-1}B_c \quad (10.1)$$

meaning that (A_d, B_c, C_d) realizes $h_\phi - h(1)$. Consequently, the Laurent coefficients h_k^ϕ of h_ϕ read as $(C_d A_d^k B_c)_{k=0}^\infty$, the vector ω_ϕ can be computed according to (2.2) and due to $\sigma(A_d) \subset \mathbb{D}$ the vector ξ_ϕ can be computed as the vector ξ in Proposition 5.3. Finally, as approximant of h the rational function

$$h_{\mathrm{red}}^{\mathrm{moeb}}(z) := h_{\xi_\phi}\left(\frac{1+z}{1-z}\right), \quad h_{\xi_\phi} := h(1) + \frac{P_\phi}{Q_\phi}$$

is in question. Example 5 compares the approximant h_ξ obtained directly with the approximant obtained as just carried out.

11. Examples

According to (2.3) for $\gamma \in \{\mathbb{T}, \mathbf{i}\mathbb{R}\}$ the minimizer $\xi := \mathrm{col}(\xi_i)_{i=0}^{m-1} \in \mathbb{C}^m$ of F can be computed as solution to $G_m x = g_m$ where $G_m := [g_{ij}]_{i,j=0}^{m-1}$, $g_m := \mathrm{col}(g_{im})_{i=0}^{m-1}$ and

$$\begin{aligned} \gamma = \mathbb{T}: \quad & g_{ij} := \tfrac{1}{2\pi}\int_0^{2\pi} \overline{f_i(e^{\mathbf{i}t})} f_j(e^{\mathbf{i}t})\mathrm{d}t \\ \gamma = \mathbf{i}\mathbb{R}: \quad & g_{ij} := \lim_{N\to\infty} \tfrac{1}{2\pi}\int_{-N}^{+N} \overline{f_i(\mathbf{i}t)} f_j(\mathbf{i}t)\mathrm{d}t \\ & f_k(z) := CA^k(zI-A)^{-1}B. \end{aligned} \tag{11.1}$$

By approximation of h via h_ξ the coefficient vector ω of its numerator polynomial P has to be computed according to (2.2) where x has to be replaced by ξ. By approximation of h via h_ξ^η the coefficient vector η of the numerator polynomial $\mathscr{P}$ of h_ξ^η can be obtained by $B_\gamma(Q)W$ where W is defined as in (8.1), or equivalently as solution to the interpolation problem (8.2) supposed that the zeros β_i of Q are known and simple. Thus,

$$\begin{aligned} h_\xi := \frac{P}{Q}, \quad & P := \psi_m \omega, \\ & Q := z^m - \psi_m \xi, \\ h_\xi^\eta := \frac{\mathscr{P}}{Q}, \quad & \mathscr{P} := \psi_m \eta, \end{aligned} \qquad \begin{aligned} \omega &:= \begin{bmatrix} h_0 & \dots & h_{m-1} \\ & \ddots & \vdots \\ & & h_0 \end{bmatrix} \begin{bmatrix} -\mathrm{col}(\xi_i)_{i=1}^{m-1} \\ 1 \end{bmatrix} \\ \eta &:= B_\gamma(Q)W, \; W := \mathrm{col}\left(\left\langle \frac{z^i}{Q} \,\middle|\, h \right\rangle_2^\gamma \right)_{i=0}^{m-1} \\ & \text{or equivalently} \\ \eta &:= \left([\gamma_i^j]_{i,j=0}^{m-1} \right)^{-1} \mathrm{col}\left(h(\gamma_i) Q(\gamma_i) \right)_{i=0}^{m-1} \\ \gamma_i &:= \phi_\gamma(\beta_i). \end{aligned} \tag{11.2}$$

Examples 1, 2, 5 use that computation receipt. If one refuses the computation of the integrals (11.1) and if large enough subsections of $(h_k)_{k=0}^\infty$, $h(z) = \sum_{k=0}^\infty h_k z^{-(k+1)}$, and $(\theta_k)_{k=0}^\infty$, $Q(z)^{-1} = \sum_{k=0}^\infty \theta_k z^{-(k+1)}$, are known, then in the unit circle case the vectors ξ and η can be obtained as limits of the truncated computation procedure

$$\begin{aligned} \xi_N &:= {H_m^N}^\dagger \mathrm{col}(h_i)_{i=m}^{N+m}, \quad H_m^N := [h_{i+j}]_{i,j=0}^{N,m-1}, \quad \lim_{N\to\infty} \xi_N \to \xi \\ \eta_N &:= {\Theta_m^N}^\dagger \mathrm{col}(h_i)_{i=0}^{N}, \quad \Theta_m^N := [\theta_{i+j}]_{i,j=0}^{N,m-1}, \quad \lim_{N\to\infty} \eta_N \to \eta \end{aligned} \tag{11.3}$$

which has been used in Example 4. If one knows q and the residues r_j, and is able to compute q_ϕ and $r_{j\phi}$, then according to Proposition 6.3 for the continues case the minimizer ξ can also be obtained via a projection method applied to elements of ℓ_2. But when $\sigma(h)$ is close to γ, only the application of solution algorithms, which address the Hankel structure of H_m^N and Θ_m^N, lead to satisfying numerical results. Thus, it is natural to ask for algorithms which compute $H^\dagger$ by exploitation of its structure. In the literature the Toeplitz structure is exploited in two different directions: one approach leads to algorithms of low complexity (fast and super

fast algorithms) and a second approach to algorithms which take care for accuracy and stability (high performance algorithms). For fast and super fast algorithms consult among others [1, 2, 3, 6, 7] and for high performance algorithms consult [16, 17, 20]. The paper [19] goes in both directions. Numerical problems arising in connection with the computation of the integrals (11.1), of the sequences $(h_k)_{k=0}^{\infty}$ and $(\theta_k)_{k=0}^{\infty}$, or of $H_m^{N\dagger}$ and $\Theta_m^{N\dagger}$ are not addressed in this paper. Supposed a fraction representation p/q of h is known and $\gamma \in \{\mathbb{T}, \mathbf{i}\mathbb{R}\}$, the Bezoutian $B_\gamma(q)$ can be computed via (3.1) and G via the product $p(C_q)^* B_\gamma^{-1}(q) p(C_q)$. In Example 3 and 6 we use that formula, but for large n the inversion of $B_\gamma(q)$ becomes very expensive which reflects the disadvantage of that approach. Finally, in the case where a minimal realization of h is available Proposition 7.2 can be applied. A disadvantage of that computation approach is the necessity to know the corresponding observability Gramian $\mathscr{Q}$ completely. To overcome associated storage and condition problems, in [28] a low rank approximation of $\mathscr{Q}$ is recommended which is found by a low-rank Smith type method. Observe that by exploitation of $G_m x = g_m$ the knowledge only of the $m \times (m+1)$-dimensional left upper corner of G is necessary, and that according to (6.4) the matrix $[G_m, g_m]$ is completely determined by the numbers g_{ii}, h_i, $i = 0, \ldots, m-1$.

The approximation procedures will be illustrated now by means of some examples. Here, we pay our main attention to the line case, the special case $h = 1/q$ and the situation where $\sigma(h)$ is close to γ or the partial fraction coefficient of a pole is large in comparison with the coefficients of the other poles. For the unit circle and general p a lot of examples are available in [12, 13, 21].

Example 1. We consider the m-degree approximation of

$$h(z) := \frac{1}{2c} \sum_{k=1}^{\kappa} \frac{c_k}{z - \alpha_k} + \frac{c_k}{z - \overline{\alpha_k}},\ \alpha_k := \frac{k}{8}\left(\mathbf{i} - \frac{1}{2\kappa}\right) \in \mathbb{H}_-,\ c_k := k^2,\ c := \sum_{k=1}^{\kappa} c_k$$

with respect to $\gamma := \mathbf{i}\mathbb{R}$. Due to $h_k = c^{-1} \Re \sum_{i=1}^{\kappa} c_i \alpha_i^k$ for $\kappa = 10$ we have

$$\begin{bmatrix} h_0 & h_1 & h_2 & h_3 \end{bmatrix} = \begin{bmatrix} 1 & -0.049 & -1.026 & 0.1680 \end{bmatrix}.$$

For $N = 5000$ the computation of the truncated integrals (11.1) yields

$$G_4 = [g_{ij}]_{i,j=0}^{3} = \begin{bmatrix} 1.550 & -0.500 & -1.671 & 1.027 \\ -0.500 & 1.720 & -0.001 & -2.113 \\ -1.671 & -0.001 & 2.063 & -0.526 \\ 1.027 & -2.113 & -0.526 & 2.807 \end{bmatrix}$$

$$g_4 = \begin{bmatrix} g_{04} & \ldots & g_{34} \end{bmatrix}^{\mathrm{T}} = \begin{bmatrix} 1.945 & 0.534 & -2.635 & -0.014 \end{bmatrix}^{\mathrm{T}}$$

what can be verified by utilization of (6.4) or of $G = V^* \Omega V$ where V and Ω are defined as in (7.1). The solution ξ to $G_4 x = g_4$ and the vector ω computed via (11.2) read as

$$\xi := -\begin{bmatrix} 0.865 & 0.652 & 2.126 & 0.578 \end{bmatrix}^{\mathrm{T}},\ \omega := \begin{bmatrix} 0.123 & 1.072 & 0.529 & 1 \end{bmatrix}^{\mathrm{T}}.$$

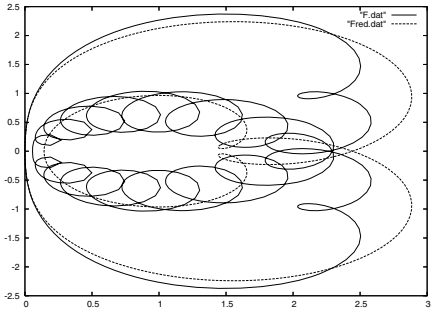


FIGURE 1. $n = 2\kappa = 20$, $m = 4$, solid line $\widehat{=}\ h(\mathbf{i}\mathbb{R})$, dashed line $\widehat{=}\ h_\xi(\mathbf{i}\mathbb{R})$

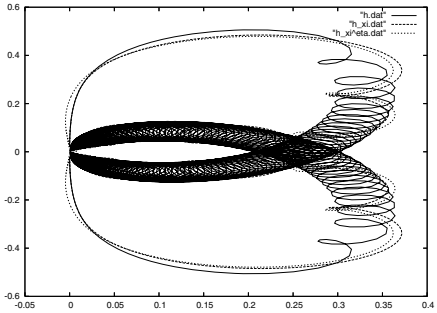


FIGURE 2. $n = 2\kappa = 200$, $m = 14$, solid line $\widehat{=}\ h(\mathbf{i}\mathbb{R})$, dashed line $\widehat{=}\ h_\xi(\mathbf{i}\mathbb{R})$, dashed line $\widehat{=}\ h_\xi^\eta(\mathbf{i}\mathbb{R})$

In Figure 1 the Nyquist plots of h and h_ξ are compared to each other. According to Theorem 4.4 the zeros of Q belong to $\overline{\mathbb{H}}_-$. Actually,

$$\sigma(Q) = \{-0.18 \pm 0.751\mathbf{i}, -0.109 \pm 1.2\mathbf{i}\} =: \{\beta_0, \ldots, \beta_3\}$$

holds. With $\phi_\gamma(z) = -\overline{z}$ the solution to the interpolation problem (8.2) reads as

$$\eta := \left[\ 0.106 \quad 1.012 \quad 0.515 \quad 0.949\ \right]^{\mathrm{T}}.$$

Since $\eta \approx \omega$ there is no significant difference between $h_\xi(\mathbf{i}\mathbb{R})$ and $h_\xi^\eta(\mathbf{i}\mathbb{R})$. Enlargement of κ and m shows the improved performance of h_ξ^η more better. Figure 2 plots for $\kappa = 100$ and $m = 14$ the sets $h(\mathbf{i}\mathbb{R})$, $h_\xi(\mathbf{i}\mathbb{R})$, $h_\xi^\eta(\mathbf{i}\mathbb{R})$ where G_{14} and g_{14} have been generated by the factorization $G = V^*\Omega V$. In particular

$$\|h - h_\xi\|_2^{\mathbf{i}\mathbb{R}} = 0.093, \quad \|h - h_\xi^\eta\|_2^{\mathbf{i}\mathbb{R}} = 0.087, \quad \|h\|_2^{\mathbf{i}\mathbb{R}} = 0.463.$$

Example 2. To illustrate the situation $\sigma(Q) \subset \gamma$ we consider the line

$$\gamma := \{w(t) : t \in \mathbb{R}\},\ w(t) := 1 + (2 + \mathbf{i})t$$

and the monic polynomial q generated by the zeros

$$\alpha_0 := -1 + 0.5\mathbf{i}, \quad \alpha_1 := -0.1 + 1\mathbf{i}, \quad \alpha_2 := 0.3 + 1.5\mathbf{i}, \quad \alpha_3 := 1.5 + 1.1\mathbf{i}.$$

It turns out, that $\sigma(q)$ is located on the left-hand side of γ, if t runs over $(-\infty, +\infty)$. The setting $h := 1/q$ simplifies the integral (11.1) to

$$g_{ij} = \lim_{N\to\infty} \tfrac{1}{2\pi} \int_{-N}^{N} \overline{w(t)}^i w(t)^j |q(w(t))|^{-2} |\dot{w}(t)| \,\mathrm{d}t.$$

For $N = 400$ and $m = 3$ we get

$$\xi = \begin{bmatrix} -0.972 - 1.819\mathbf{i} & -0.32 + 2.503\mathbf{i} & 2.358 - 0.321\mathbf{i} \end{bmatrix}^{\mathrm{T}}.$$

Actually, the parameters

$$t_1 := -0.8325, \quad t_2 := -0.0684, \quad t_3 := 0.5795$$

satisfy $\sigma(Q) = \{w(t_1), w(t_2), w(t_3)\}$.

Example 3. We give an example for the limit situations as described in Proposition 9.2 and 9.3. We have to consider a rational function h with the property, that some of its partial fraction coefficients are large in comparison with the others. Let $\sigma(q)$ be given by

$$\alpha_{0,1} := -1 \pm \mathbf{i}, \quad \alpha_{2,3} := -2 \pm \mathbf{i}$$

and p be defined via the interpolation conditions

$$p(\alpha_0) = p(\alpha_1) = 50, \quad p(\alpha_2) = p(\alpha_3) = 1.$$

Then $h := p/q$ where

$$p(z) = -19.6z^3 - 88.2z^2 - 137.2z - 48, \quad q(z) = z^4 + 6z^3 + 15z^2 + 18z + 10.$$

For $\gamma := \mathbf{i}\mathbb{R}$ and $m := 2$ the relation $\sigma(Q) \approx \{\alpha_0, \alpha_1\}$ is to be expected. To be exact we use $G = p(C_q^{\mathrm{T}}) B_c^{-1}(q) p(C_q)$. Then we have

$$\xi = -\begin{bmatrix} 1.984 \\ 1.986 \end{bmatrix} = G_2^{-1} g_2, \; G_2 = \begin{bmatrix} 108.237 & -192.08 \\ -192.08 & 409.555 \end{bmatrix}, g_2 = \begin{bmatrix} 166.685 \\ -432.18 \end{bmatrix}.$$

Thus, actually $\sigma(Q) = \{-0.993 \pm 0.999\mathbf{i}\}$ holds. To illustrate Proposition 9.3 we move $\alpha_{0,1}$ close to γ: $\alpha_{0,1} := -0.01 \pm \mathbf{i}$, let α_2, α_3 be unchanged and choose p according to

$$p(\alpha_0) = p(\alpha_1) = 1, \quad p(\alpha_2) = p(\alpha_3) = -1.$$

As above $\xi = -[0.99, 0.075]^{\mathrm{T}}$ is obtained, thus $\sigma(Q) = \{-0.037 \pm 0.994\mathbf{i}\} \approx \{\alpha_0, \alpha_1\}$ as expected. It turns out that for $N = 10^3$ the utilization of the truncated integrals (11.1) yields the same result.

Example 4. To illustrate for $n := 20$ and large $p_0 \in \mathbb{R}$ the advantage of a second optimization step as proposed in Section 8, we approximate $h := p/q$ along $\mathbb{T}$ by rational functions h_ξ and h_ξ^η of degree $m := 7$ where

$$p(z) := \sum_{k=1}^{n-1} z^k + p_0, \; q(z) := \prod_{k=0}^{\frac{n}{2}-1} (z - \alpha_k)(z - \overline{\alpha_k}), \; \alpha_k := \frac{8}{10} - \frac{3k}{n+1} - \frac{8k^2}{(n+10)^2}\mathbf{i}.$$

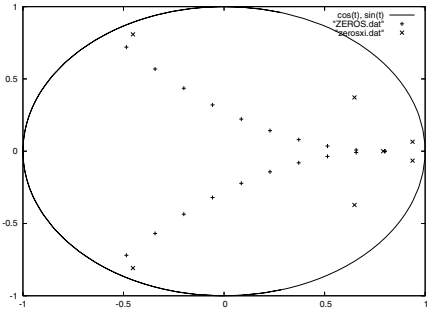


FIGURE 3. $n = 20, m = 7, + := \sigma(q), \times := \sigma(Q)$

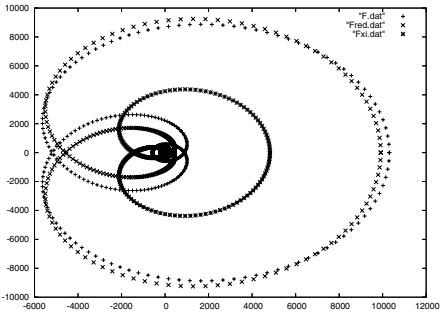


FIGURE 4. $n = 20, m = 7, + := h(\mathbb{U}), \times := h_\xi^\eta(\mathbb{U}), * := h_\xi(\mathbb{U})$

Figure 3 shows the distribution of the α_k marked by "+". We choose $p_0 := 10$, $N := 200$, and form the sequences $(h_k)_{k=0}^{N+m}$ and $(\theta_k)_{k=0}^{N}$ to find h_ξ and h_ξ^η. Then (11.2) and (11.3) provide

$$\xi = [0.336, -1.562, 2.889, -2.948, 2.879, -3.652, 3.057]^{\mathrm{T}},$$
$$\omega = [0.677, 0.752, 0.769, 0.866, 0.786, 1.085, 1.000]^{\mathrm{T}},$$
$$\eta = [120.391, -226.509, 170.689, -143.321, 183.663, -126.761, 34.052]^{\mathrm{T}}.$$

For sufficiently large N Theorem 4.4 states the relation $\sigma(Q) \subset \mathbb{D}$. Figure 3 confirms that localization where $\{\beta_0, \ldots, \beta_6\} := \sigma(Q)$ is marked with "×". It is easy to check, that $\mathscr{P}, Q, h$ satisfy the interpolation conditions

$$\mathscr{P}(\gamma_i) = h(\gamma_i)Q(\gamma_i),\ \gamma_i := 1/\beta_i.$$

In Figure 4 the Nyquist plots of h, h_ξ, h_ξ^η along the unit roots $\mathbb{U} := \sigma(z^{3000} - 1)$ are compared to each other. Obviously, h_ξ^η fits more better than h_ξ, precisely

$$\sqrt{\frac{\sum_{w\in\mathbb{U}} |h(w)-h_\xi(w)|^2}{\sum_{w\in\mathbb{U}} |h(w)|^2}} = 0.571, \quad \sqrt{\frac{\sum_{w\in\mathbb{U}} |h(w)-h_\xi^\eta(w)|^2}{\sum_{w\in\mathbb{U}} |h(w)|^2}} = 0.183.$$

Example 5. We approximate

$$h(z) := \sum_{k=0}^{2} \frac{1}{z-\alpha_k} + \frac{1}{z-\overline{\alpha_k}},\ \alpha_0 := -2+7\mathbf{i},\ \alpha_1 := -2+2\mathbf{i},\ \alpha_2 := -2+1\mathbf{i}$$

along $\gamma := \mathbf{i}\mathbb{R}$ by h_ξ and $h_{\mathrm{red}}^{\mathrm{moeb}}$ of degree $m := 3$, and compare the corresponding approximation quality. Obviously, h admits the fraction representation

$$h(z) = \frac{6z^5+60z^4+456z^3+1776z^2+3570z+2916}{z^6+12z^5+114z^4+592z^3+1785z^2+2916z+2120}.$$

Since $h_k := 2\Re \sum_{i=0}^{2} \alpha_i^k$ the first 3 Laurent coefficients of h read as

$$\left[\begin{array}{ccc} h_0 & h_1 & h_2 \end{array}\right] = \left[\begin{array}{ccc} 6 & -12 & -84 \end{array}\right].$$

The computation of g_{ii}, $i = 0, 1, 2$, via the integrals (11.1) and application of (6.4) yield the system

$$\left[\begin{array}{rrr} 4.832 & -18 & -8.934 \\ -18 & 80.934 & -72 \\ -8.934 & -72 & 1295.242 \end{array}\right] x = \left[\begin{array}{r} 576 \\ -2303.242 \\ -3528 \end{array}\right].$$

Its solution ξ and the resulting ω read as

$$\xi = -\left[\begin{array}{ccc} 152.464 & 69.145 & 7.619 \end{array}\right]^{\mathrm{T}},\quad \omega := \left[\begin{array}{ccc} 239.439 & 33.714 & 6 \end{array}\right]^{\mathrm{T}}.$$

The Laurent coefficients h_k^ϕ of h_ϕ are obtained via $C_d A_d^k B_c$ where (A_d, B_c, C_d) is defined as in (10.1) and

$$A_c := \mathrm{diag}(\alpha_0, \alpha_1, \alpha_2, \overline{\alpha_0}, \overline{\alpha_1}, \overline{\alpha_2}) \in \mathbb{C}^{6\times 6},\ B_c^{\mathrm{T}} := C_c := \left[\begin{array}{ccc} 1 & \dots & 1 \end{array}\right] \in \mathbb{R}^6.$$

For $N = 190$ the truncated computation procedure (11.3) provides as approximation of h_ϕ along $\mathbb{T}$ the rational function $h_{\xi_\phi} := h(1) + P_\phi/Q_\phi$ where

$$Q_\phi(z) := z^3 - \psi_3(z)\xi_\phi,\ \xi_\phi := -\left[\begin{array}{ccc} 0.303 & 1.238 & 1.904 \end{array}\right]^{\mathrm{T}} = [h_{i+j}^\phi]_{i,j=0}^{190,2\,\dagger} \mathrm{col}(h_i^\phi)_{i=3}^{193}$$

$$\begin{array}{l} P_\phi := \psi_3\omega_\phi \\ h(1) = 1.165 \end{array},\quad \omega_\phi := \left[\begin{array}{c} 0.099 \\ 0.448 \\ 0.391 \end{array}\right] = \left[\begin{array}{ccc} h_0^\phi & h_1^\phi & h_2^\phi \\ & h_0^\phi & h_1^\phi \\ & & h_0^\phi \end{array}\right] \left[\begin{array}{c} -\xi_1^\phi \\ -\xi_2^\phi \\ 1 \end{array}\right]$$

that means

$$h_{\xi_\phi}(z) = h(1) + \frac{P_\phi(z)}{Q_\phi(z)} = \frac{1.165z^3+2.609z^2+1.889z+0.453}{z^3+1.904z^2+1.238z+0.303}.$$

Finally,

$$h_{\mathrm{red}}^{\mathrm{moeb}}(z) := h_{\xi_\phi}\left(\frac{1+z}{1-z}\right) = \frac{-0.243z^3+11.764z^2+94.772z+202.879}{z^3+25.496z^2+91.439z+147.452}.$$

In Figure 5 the Nyquist plots of h, h_ξ and $h_{\mathrm{red}}^{\mathrm{moeb}}$ with respect to $\mathbf{i}\mathbb{R}$ are compared to each other. Obviously, for some intervals of $\mathbf{i}\mathbb{R}$ the approximation quality differs significantly. Along $\mathbf{i}[-5,5]$ the approximant $h_{\mathrm{red}}^{\mathrm{moeb}}$ fits better than h_ξ, and along $\mathbf{i}(\mathbb{R}\setminus[-5,5])$ the approximant h_ξ is to be preferred. Remember, a positive (real)

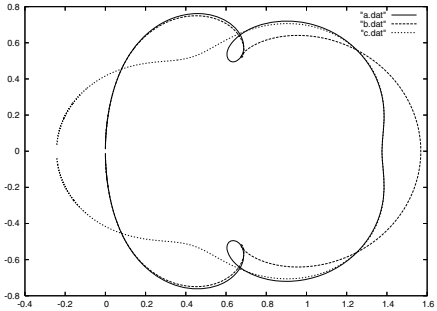


FIGURE 5. a$\widehat{=} h(\mathbf{i}\mathbb{R})$, b$\widehat{=} h_\xi(\mathbf{i}\mathbb{R})$, c$\widehat{=} h_{\mathrm{red}}^{\mathrm{moeb}}(\mathbf{i}\mathbb{R})$

function $h(z)$ is a (real) function whose real part is positive when the real part of z is positive [[8], p. 197]. According to [[8], Theorem V] a real rational function h is positive, if $\sigma(h) \subset \mathbb{H}_-$ and $\Re(h(\mathbf{i}\mathbb{R})) \subset \mathbb{R}_+$. The definition of the numbers α_i and Theorem 4.4 ensures the stability of h and h_ξ. Thus, Figure 5 shows the positivity of h and h_ξ, and the non-positivity of $h_{\mathrm{red}}^{\mathrm{moeb}}$. Consequently, if one like to preserve positivity, then h_ξ is to be preferred. Unfortunately, also the direct approach does not preserve positivity in general. We finish the paper with a counter example.

Example 6. Let $h(z) := \sum_{k=0}^{2} c_k (z - \alpha_k)^{-1}$ where

$$\alpha_0 := -1 + \mathbf{i},\ \alpha_1 := \overline{\alpha_0},\ \alpha_2 = -3,\ c_0 := c_1 := 1,\ c_2 := 2.$$

Then with $p(z) := 4z^2 + 12z + 10$ and $q(z) := z^3 + 5z^2 + 8z + 6$ the equation $h = p/q$ follows. Because all c_k are positive, and h is stable with respect to $\mathbf{i}\mathbb{R}$, we are dealing with a positive real function. To avoid the approximation of the integrals (11.1) we use factorization (1.3). The representation (3.1) of $B_c(q)$ and $R = p(C_q)$ provide

$$B_c^{-1}(q) = \tfrac{1}{|B_c(q)|} \begin{bmatrix} 680 & 0 & -816 \\ 0 & 816 & 0 \\ -816 & 0 & 6528 \end{bmatrix},\ R = \begin{bmatrix} 10 & -24 & 48 \\ 12 & -12 & 40 \\ 4 & -8 & 18 \end{bmatrix}$$

$$|B_c(q)| = 55488.$$

By exploitation of $B_c(q) = B_{\mathbf{i}\mathbb{R}}(q)$ and $G = RB_{\mathbf{i}\mathbb{R}}^{-1}(q)R^*$ one gets the equation system

$$\underbrace{\begin{bmatrix} 224672 & -443904 \\ -443904 & 891072 \end{bmatrix}}_{= G_2|B_c(q)|} x = \underbrace{\begin{bmatrix} 884544 \\ -1775616 \end{bmatrix}}_{= g_2|B_c(q)|}$$

to determine $\xi \in \mathbb{R}^2$. Thus, $\xi = -[0.0034, 1.9944]^{\mathrm{T}}$. Since $[h_0, h_1, h_2] = e_3^{\mathrm{T}} R$ we have $h_0 = 4$ and $h_1 = -8$ such that in view of (11.2) the coefficient vector ω of the numerator polynomial of h_ξ becomes $[-0.023, 4]$. Consequently, $h_\xi(0) = -\omega_0/\xi_0 < 0$ in contradiction to the stipulation $h_\xi : \mathbb{H}_+ \to \mathbb{H}_+ := \mathbb{C} \setminus (\mathbb{H}_- \cup \mathbf{i}\mathbb{R})$.

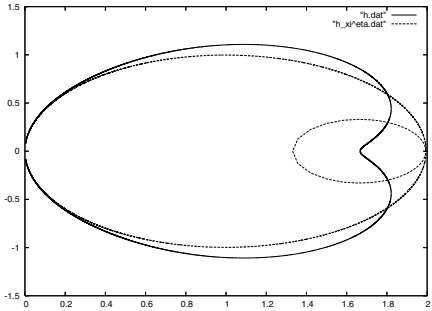


FIGURE 6. solid line := $h(\mathbf{i}\mathbb{R})$, dashed line := $h_\xi^\eta(\mathbf{i}\mathbb{R})$

But, if we replace ω by η, then a positive real approximant is obtained. To get η it is sufficient to compute the product $B_c(Q)W$. The generating function of $B_c(Q)$ yields $B_c(Q) = 2\xi_1 \mathrm{diag}(\xi_0, -1) = \mathrm{diag}(0.013, 3.989)$. In addition, for $N = 300$ as approximation of

$$W = \lim_{N\to\infty} \frac{1}{2\pi} \int_{-N}^{+N} \overline{\begin{bmatrix} 1 \\ \mathbf{i}t \end{bmatrix}} \frac{h(\mathbf{i}t)}{\overline{Q(\mathbf{i}t)}} \mathrm{d}t,$$

we get $[0.334, 0.997]^{\mathrm{T}}$ such that $\eta := B_c(Q)W$ becomes $[0.005, 3.977]^{\mathrm{T}}$. To check the positivity of h_ξ^η we represent it as continued fraction:

$$h_\xi^\eta(z) = \frac{\eta_1 z + \eta_0}{z^2 - \xi_1 z - \xi_0} = \cfrac{1}{(az+b) + \cfrac{r}{(\eta_1 z + \eta_0)}}.$$

Then the positivity of all coefficients a, b, r, η_0, η_1 implies the desired mapping property $h_\xi^\eta : \mathbb{H}_+ \to \mathbb{H}_+$. From above we know that η_0 and η_1 are positive. Comparison of the coefficients yields

$$a = 1/\eta_1, \quad b = -(a\eta_0 + \xi_1)/\eta_1, \quad r = -\xi_0 - b\eta_0$$

thus $a = 0.251$, $b = 0.501$, and $r = 0.001$. Figure 6 confirms the positivity of h and h_ξ^η. A mathematical exploration with respect to the preservation of positivity will be provided in a forthcoming paper.

References

[1] G.S. Ammar, W.B. Gragg, Numerical experience with a superfast real Toeplitz solver, LAA 34 (1980), 103–116

[2] G.S. Ammar, W.B. Gragg, The generalized Schur algorithm for the superfast solution of Toeplitz systems, in: J. Gilewicz, M. Pidor, W. Siemaszko (Eds.), Rational Approximation and it Applications in Mathematics and Physics, Lecture Notes in Mathematics, vol. 1237, 1987, pp. 315

[3] G.S. Ammar, W.B. Gragg, Superfast solution of real positive definite Toeplitz systems, SIAM J. Matrix Anal. Appl.9 (1) (1988), 61–76

[4] A.C. Antoulas, D. Sorenson, S. Gugercin, A survey of model reduction methods for large scale systems, Contemporary Mathematics, AMS Publications (2001), 193–221

[5] M. Van Barel, A. Bultheel, Padé techniques for model reduction in linear system theory: a survey, J. Comput. Appl. Math. 14 (1986), 401–438

[6] M. Van Barel, G. Heinig, P. Kravanja, A superfast method for solving Toeplitz linear least squares problems, LAA 366 (2003), 441–457

[7] M. Van Barel, G. Heinig, P. Kravanja, A stabilized superfast solver for nonsymmetric Toeplitz systems, SIAM J. Matrix Anal. Appl.23 (2) (2001), 494–510

[8] O. Brune, Synthesis of a finite two-terminal network whose driving-point impedance is a prescribed function of frequency, J. of Math. Phys. 10 (1931), 191–236

[9] P.J. Davis, *Interpolation & Approximation*, Dover, New York, 1975

[10] L. Fejér, Über die Lage der Nullstellen von Polynomen, die aus Minimumforderungen gewisser Art entspringen, Math. Annalen 85 (1922), 41–48

[11] S. Feldmann, Fejér's convex hull theorem related to least squares approximation of rational functions, in preparation

[12] S. Feldmann, P. Lang, A least squares approach to reduce stable discrete linear systems preserving their stability, LAA 381 (2004), 141–163

[13] S. Feldmann, P. Lang, D. Prätzel-Wolters, A unified least squares approach to identify and to reduce continuous asymptotically stable systems, LAA 426, Issues 2-3 (2007), 674–689

[14] P.A. Fuhrmann, *A Polynomial Approach to Linear Algebra*, Springer, New York, Berlin, Heidelberg, 1996

[15] M. Fujiwara, Über die algebraischen Gleichungen, deren Wurzeln in einem Kreise oder in einer Halbebene liegen, Math. Z. 24 (1926), 161–169

[16] K.A. Gallivan, S. Thirumalai, P. Van Dooren, V. Vermaut, High performance algorithms for Toeplitz and block Toeplitz matrices, LAA 241-243 (1996), 343–388

[17] I. Gohberg, T. Kailath, V. Olshevsky, Fast Gaussian elimination with partial pivoting for matrices with displacement structure, Math. Comp. 64 (212) (1995), 1557–1576

[18] G.H. Golub, Ch.F. van Loan, *Matrix Computations*, The Johns Hopkins University Press, Baltimore, 1991

[19] M. Gu, New fast algorithms for structured least squares problems, SIAM J. Matrix Anal. Appl. 20 (1) (1998), 244–269

[20] M. Gu, Stable and efficient algorithm for structured systems of equations, SIAM J. Matrix Anal. Appl. 19 (2) (1997), 279–306

[21] S. Gugercin, A.C. Antoulas, Model reduction of large-scale systems by least squares, LAA 415, Issue 2-3 (2006), 290–321

[22] G. Heinig, Bezoutiante, Resultante und Spektralverteilungsprobleme für Operatorpolynome, Math. Nachr. 91 (1979), 23–43

[23] G. Heinig, U. Jungnickel, Zur Lösung von Matrixgleichungen der Form $AX - XB = C$, Wiss. Zeitschrift der TH Karl-Marx-Stadt 23 (1981), 387–393

[24] G. Heinig, K. Rost, *Algebraic Methods for Toeplitz-like Matrices and Operators*, Akademie-Verlag, Berlin, 1984

[25] U. Helmke, P.A. Fuhrmann, Bezoutians, LAA 122-124 (1989), 1039–1097

[26] M.G. Krein, M.A. Naimark, The Method of Symmetric and Hermitian Forms in the Theory of the Separation of the Roots of Algebraic Equations, English translation in Linear and Multilinear Algebra 10 (1981), 265–308

[27] C. Lanczos, *Applied Analysis*, Prentice-Hall, Inc., Englewood Cliffs, New Jersey, 1956

[28] T. Penzl, A cyclic low-rank Smith method for large sparse Lyapunov equations, SIAM J. Sci. Comput., Vol. 21, No. 4 (2000), 1401–1418

[29] W. Rudin, *Real and Complex Analysis*, Mc Graw-Hill, New York, 1987

[30] J.T. Spanos, M.H. Milman, D.L. Mingori, A new algorithm for L_2-optimal model reduction, Automatica 28 (1992), 897–909

[31] J. Sylvester, On a Theory of the Syzygetic relations of two rational integral functions, comprising an application to the theory of Sturm's Functions, and that of the greatest Algebraic Common Measure, Philos. Trans. Roy. Soc. London 143 (1853), 407–548

[32] D.A. Wilson, Optimum solution of model reduction problem, Proc. Inst. Elec. Eng. (1970) 1161–1165

[33] H.K. Wimmer, On the history of the Bezoutian and the resultant matrix, LAA 128 (1990) 27–34

[34] W.Y. Yan, J. Lam, An approximative approach to H_2-optimal model reduction, IEEE Trans. Automatic Control, AC-44 (1999), 1341–1358

[35] K. Zhou, J.C. Doyle, K. Glover, *Robust and Optimal Control*, Prentice Hall, Simon & Schuster, New Jersey, 1995

Sven Feldmann
Thiemstrasse 21
D-04299 Leipzig, Germany
e-mail: svendrfeldmann@yahoo.de

Operator Theory:
Advances and Applications, Vol. 199, 289–332

On the Weyl Matrix Balls Corresponding to the Matricial Carathéodory Problem in Both Nondegenerate and Degenerate Cases

Bernd Fritzsche, Bernd Kirstein and Andreas Lasarow

Dedicated to the memory of Georg Heinig

Abstract. The main goal of the paper is to determine the Weyl matrix balls associated with an arbitrary matricial Carathéodory problem. For the special case of a nondegenerate matricial Carathéodory problem the corresponding Weyl matrix balls were computed by I.V. Kovalishina [Ko] and alternatively by the first and the second authors in [FK1, Parts IV and V].

Mathematics Subject Classification (2000). Primary: 44A60, 47A57, 30E05
Secondary: 47A56.

Keywords. Weyl matrix balls, matricial Carathéodory problem, matrix polynomials, matricial Carathéodory functions, matricial Schur functions.

0. Introduction

This paper is closely related to the authors' recent investigations [FKL1] on the matricial Carathéodory problem in both nondegenerate and degenerate cases. In [FKL1] the authors obtained a parametrization of the solution set of a general (possibly degenerate) matricial Carathéodory problem in terms of a linear fractional transformation. Using this parametrization we study the set of values of matrices which will be attained by the solutions of the matricial Carathéodory problem in a prescribed fixed point of the open unit disk. We show that this set of values fills a (closed) matrix ball and present explicit expressions for the center and the semi-radii of this matrix ball which is also called the Weyl matrix ball associated with the concrete matricial Carathéodory problem under consideration. Following the classical monograph [Akh] the terminology "Weyl circles" or later

The work of the third author of the present paper was supported by the German Research Foundation (Deutsche Forschungsgemeinschaft) on badge LA 1386/2–1.

"Weyl matrix balls" was consequently used in the Soviet literature (see, e.g., [Or], [KP], [Ko], [Du], and [Mi]). As explained in [Akh, Chapter 1] or in [Ev], the history of the scalar case is intimately related to the classical papers [We], [He], and [Ne].

In the case of a nondegenerate matricial Carathéodory problem the corresponding Weyl matrix balls were computed by Kovalishina [Ko] and alternatively by the first and the second authors in [FK1, Parts IV and V]. In view of the general theory of matrix balls due to Šmuljan [Sm], the semi-radii of a matrix ball are not uniquely determined. In his investigations [Du] on the matricial Schur problem, V.K. Dubovoj has constructed a clever normalization of the left semi-radius. This normalization was adopted to the case of the matricial Carathéodory problem in [FK1, Parts IV and V]. In the case of the nondegenerate matricial Carathéodory problem the normalized semi-radii of the Weyl matrix balls can be nicely rewritten with the aid of the Gohberg-Heinig formula (see [GH, Theorem 1.1], for the inverses of block Toeplitz matrices. This will be described at the end of Section 4.

In the scalar case the Weyl disks associated with the Carathéodory problem were already handled by Geronimus (see, e.g., [Ge1], [Ge2], and [Ge3]). A modern view on various aspects of Geronimus' work was presented in the paper [CG] of Chang and Georgiou who particularly worked out the role of the centers of the Weyl disks in the context of maximum entropy extension. In the paper [FKL2] the authors discussed similar questions and closely related matters for the matrix case of the nondegenerate matricial Carathéodory problem. The limit behaviour of the normalized semi-radii of the Weyl matrix balls associated with a nondegenerate matrix-valued Carathéodory function was studied in [FK2].

This paper is organized as follows: In Section 1, after presenting some notations and preliminaries we formulate the main result of this paper (see Theorem 1.1). In Section 2, we recall the parametrization of the solution set of a matricial Carathéodory problem which was obtained in [FKL1] and we give some additional comments concerning the uniqueness of the functions which appear as parameters in this description. It turns out that much information on the Weyl matrix balls associated with a matricial Carathéodory problem is contained in a distinguished rational matrix-valued function Θ_n built from the data. For this reason, the study of this function Θ_n is one of the central themes of the paper. In Section 3, we introduce the function Θ_n and show that the restriction of this function onto $\mathbb{D}$ is a matrix-valued Schur function. Hereby this Schur function is inner if and only if the considered Carathéodory problem is nondegenerate. In the particular case that the given sequence $(\Gamma_j)_{j=0}^n$ of data satisfies $\det \Gamma_0 \neq 0$ we associate with $(\Gamma_j)_{j=0}^n$ a second rational matrix-valued function $\Theta_n^\diamond$. The interplay between both functions Θ_n and $\Theta_n^\diamond$ is described in Proposition 3.11. In Section 4, we will prove the main result of this paper. More precisely, we will compute the Weyl matrix ball associated with an arbitrary matricial Carathéodory problem. In Section 5, we consider in detail the case that the given $q \times q$ Carathéodory sequence $(\Gamma_j)_{j=0}^n$ satisfies $\det \Gamma_0 \neq 0$. We indicate that the above-mentioned rational matrix-valued function $\Theta_n^\diamond$ is intimately connected with the so-called reciprocal Carathéodory sequence $(\Gamma_j^\sharp)_{j=0}^n$ corresponding to $(\Gamma_j)_{j=0}^n$. This enables us to describe the Weyl

matrix ball associated with $(\Gamma_j^\sharp)_{j=0}^n$ with the aid of $\Theta_n^\diamond$ (see Theorem 5.2). The main results of Section 6 are Theorem 6.8 and Theorem 6.11 which contain useful recurrence formulas for the functions Θ_n and $\Theta_n^\diamond$, respectively. Finally, in Section 7 we discuss the central $q \times q$ Carathéodory function $\Omega_{c,n}$ corresponding to $(\Gamma_j)_{j=0}^n$ from the view of Weyl matrix balls.

1. Preliminaries

Throughout this paper, let p and q be positive integers. We will use $\mathbb{C}$, $\mathbb{N}_0$, and $\mathbb{N}$ to denote the set of all complex numbers, the set of all nonnegative integers, and the set of all positive integers, respectively. If $m \in \mathbb{N}_0$ and if $\kappa \in \mathbb{N}_0$ or $\kappa = \infty$, then we will write $\mathbb{N}_{m,\kappa}$ for the set of all integers k satisfying $m \le k \le \kappa$. The set of all complex $p \times q$ matrices will be designated by $\mathbb{C}^{p\times q}$. For each $A \in \mathbb{C}^{p\times q}$, let A^+ be the Moore-Penrose inverse of A. If $A \in \mathbb{C}^{q\times q}$, then $\det A$ stands for the determinant of A and $\operatorname{tr} A$ denotes the trace of A. Further, for each $A \in \mathbb{C}^{q\times q}$, let $\operatorname{Re} A$ and $\operatorname{Im} A$ be the real part of A and the imaginary part of A, respectively, i.e., $\operatorname{Re} A := \frac{1}{2}(A + A^*)$ and $\operatorname{Im} A := \frac{1}{2i}(A - A^*)$. The zero matrix which belongs to $\mathbb{C}^{p\times q}$ will be denoted by $0_{p\times q}$ and the identity matrix which belongs to $\mathbb{C}^{q\times q}$ will be designated by I_q. If the size of a zero matrix or a identity matrix is obvious, we will omit the indices. A complex $p \times q$ matrix A is said to be contractive if the operator norm of A is not greater than 1. Obviously, a complex $p \times q$ matrix A is contractive if and only if the matrix $I - A^*A$ is nonnegative Hermitian.

Let $n \in \mathbb{N}_0$. If $(\Gamma_j)_{j=0}^n$ is a sequence of complex $q \times q$ matrices, then we associate with $(\Gamma_j)_{j=0}^n$ the block Toeplitz matrices S_n and T_n given by

$$S_n := \begin{pmatrix} \Gamma_0 & 0 & 0 & \dots & 0 \\ \Gamma_1 & \Gamma_0 & 0 & \dots & 0 \\ \Gamma_2 & \Gamma_1 & \Gamma_0 & \ddots & \vdots \\ \vdots & \vdots & \vdots & \ddots & 0 \\ \Gamma_n & \Gamma_{n-1} & \Gamma_{n-2} & \dots & \Gamma_0 \end{pmatrix} \tag{1.1}$$

and

$$T_n := \operatorname{Re} S_n.$$

A sequence $(\Gamma_j)_{j=0}^n$ of complex $q \times q$ matrices is called $q \times q$ *Carathéodory sequence* (respectively, *nondegenerate* $q \times q$ *Carathéodory sequence*) if the matrix T_n is nonnegative Hermitian (respectively, positive Hermitian). Obviously, if $(\Gamma_j)_{j=0}^n$ is a $q \times q$ Carathéodory sequence (respectively, a nondegenerate $q \times q$ Carathéodory sequence) and if $m \in \mathbb{N}_{0,n}$, then $(\Gamma_j)_{j=0}^m$ is also a $q \times q$ Carathéodory sequence (respectively, a nondegenerate $q \times q$ Carathéodory sequence). In addition, a sequence $(\Gamma_k)_{k=0}^\infty$ of complex $q \times q$ matrices is said to be a $q \times q$ *Carathéodory sequence* (respectively, a *nondegenerate* $q \times q$ *Carathéodory sequence*) if for every choice of a nonnegative integer m the sequence $(\Gamma_j)_{j=0}^m$ is a $q \times q$ Carathéodory sequence (respectively, a nondegenerate $q \times q$ Carathéodory sequence).

Let $\mathbb{D} := \{z \in \mathbb{C} : |z| < 1\}$ and $\mathbb{T} := \{z \in \mathbb{C} : |z| = 1\}$ be the unit disk and the unit circle of the complex plane, respectively. A complex $q \times q$ matrix-valued function $\Omega : \mathbb{D} \to \mathbb{C}^{q\times q}$ which is holomorphic in $\mathbb{D}$ and for which the real part $\operatorname{Re}\Omega(z)$ of $\Omega(z)$ is nonnegative Hermitian for each $z \in \mathbb{D}$ is called $q \times q$ *Carathéodory function* (*in* $\mathbb{D}$). The set of all $q \times q$ Carathéodory functions (in $\mathbb{D}$) will be denoted by $\mathcal{C}_q(\mathbb{D})$. It is a well-known fact that a matrix-valued function $\Omega : \mathbb{D} \to \mathbb{C}^{q\times q}$ which is holomorphic in $\mathbb{D}$ with Taylor series representation

$$\Omega(z) = \sum_{k=0}^{\infty} \Gamma_k z^k, \quad z \in \mathbb{D},$$

belongs to $\mathcal{C}_q(\mathbb{D})$ if and only if $(\Gamma_k)_{k=0}^{\infty}$ is a $q \times q$ Carathéodory sequence (see, e.g., [FK1, Section 4 in Part I]).

The matricial version of the classical Carathéodory interpolation problem consists of the following:

Let $n \in \mathbb{N}_0$ and let $(\Gamma_j)_{j=0}^n$ be a sequence of complex $q \times q$ matrices. Describe the set $\mathcal{C}_q[\mathbb{D}, (\Gamma_j)_{j=0}^n]$ of all $q \times q$ Carathéodory functions Ω (in $\mathbb{D}$) such that

$$\frac{1}{j!}\Omega^{(j)}(0) = \Gamma_j \tag{1.2}$$

holds for each $j \in \mathbb{N}_{0,n}$, where $\Omega^{(j)}(0)$ is the jth derivative of Ω at the point $z = 0$.

If $n \in \mathbb{N}_0$ and if $(\Gamma_j)_{j=0}^n$ is a sequence of complex $q \times q$ matrices, then the set $\mathcal{C}_q[\mathbb{D}, (\Gamma_j)_{j=0}^n]$ is nonempty if and only if $(\Gamma_j)_{j=0}^n$ is a $q \times q$ Carathéodory sequence (see, e.g., [FK1, Section 4 in Part I]). Several approaches to parametrize the solution set $\mathcal{C}_q[\mathbb{D}, (\Gamma_j)_{j=0}^n]$ of the matricial Carathéodory problem can be found in the literature (see, e.g., [AK], [Ko], [Dy], [FK1], [BGR], [FF], [Sa], and [FFGK]). An essential common feature is that the discussions are mainly concentrated on the so-called nondegenerate case which is connected with nondegenerate $q \times q$ Carathéodory sequences built from the interpolation data. Nowadays, quite different approaches to handle also degenerate cases of matrix interpolation were used (see, e.g., [BH], [BD], [Br], [CH1], [CH2], [DGK2], [Dy, Chapter 7], and [Sa, Chapter 5]). The starting point of the present paper are the descriptions of $\mathcal{C}_q[\mathbb{D}, (\Gamma_j)_{j=0}^n]$ via linear fractional transformations in the general case, i.e., without additional assumptions, given in [FKL1], where the parameters of the linear fractional transformations are expressed explicitly by the given data of the problem. For the convenience of the reader, we recall the main idea of these parametrizations in the following section (see Theorem 2.1). For this and our forthcoming considerations we need some further notation.

If $m \in \mathbb{N}_0$, let $e_{m,q}$ and $\varepsilon_{m,q}$ be the matrix polynomials defined by

$$e_{m,q}(z) := \Big(I_q, zI_q, z^2I_q, \ldots, z^mI_q\Big) \quad \text{and} \quad \varepsilon_{m,q}(z) := \begin{pmatrix} z^mI_q \\ z^{m-1}I_q \\ \vdots \\ zI_q \\ I_q \end{pmatrix} \tag{1.3}$$

for all $z \in \mathbb{C}$. Let e be a $q \times q$ matrix polynomial of degree not greater than m, i.e., there is a complex $(m+1)q \times q$ matrix E such that $e(z) = e_{m,q}(z)E$ for each $z \in \mathbb{C}$. Then the *reciprocal matrix polynomial* $\tilde{e}^{[m]}$ *of* e with respect to the unit circle $\mathbb{T}$ and the formal degree m is given, for all $z \in \mathbb{C}$, by $\tilde{e}^{[m]}(z) := E^* \varepsilon_{m,q}(z)$.

If $n \in \mathbb{N}_0$ and if $(\Gamma_j)_{j=0}^n$ is a sequence of complex $q \times q$ matrices, then let

$$L_1 := \operatorname{Re} \Gamma_0, \quad R_1 := \operatorname{Re} \Gamma_0, \tag{1.4}$$

and moreover (in the case $n \geq 1$) we will use for each $k \in \mathbb{N}_{1,n}$ the notations

$$Z_k := \frac{1}{2}\Big(\Gamma_k, \Gamma_{k-1}, \ldots, \Gamma_1\Big), \quad Y_k := \frac{1}{2}\begin{pmatrix} \Gamma_1 \\ \Gamma_2 \\ \vdots \\ \Gamma_k \end{pmatrix}, \tag{1.5}$$

and

$$L_{k+1} := \operatorname{Re} \Gamma_0 - Z_k T_{k-1}^+ Z_k^*, \quad R_{k+1} := \operatorname{Re} \Gamma_0 - Y_k^* T_{k-1}^+ Y_k. \tag{1.6}$$

For each $k \in \mathbb{N}_{0,n}$, the matrices L_{k+1} and R_{k+1} are nonnegative Hermitian if $(\Gamma_j)_{j=0}^n$ is a $q \times q$ Carathéodory sequence (see, e.g., [DFK, Lemma 1.1.9]).

In the sequel, let $n \in \mathbb{N}_0$ and let $(\Gamma_j)_{j=0}^n$ be a $q \times q$ Carathéodory sequence. In the case $n \geq 1$, the sets

$$\mathcal{Y}_n := \{V \in \mathbb{C}^{nq \times q} : T_{n-1} V = Y_n\} \quad \text{and} \quad \mathcal{Z}_n := \{W \in \mathbb{C}^{q \times nq} : W T_{n-1} = Z_n\}$$

will be linking elements in the studies below, since the matrix polynomials a_n, b_n, c_n, and d_n defined by

$$a_n(z) := \begin{cases} \Gamma_0 & \text{if } n = 0 \\ \Gamma_0 + z e_{n-1,q}(z) S_{n-1}^* V_n & \text{if } n \geq 1\,, \end{cases} \tag{1.7}$$

$$b_n(z) := \begin{cases} I_q & \text{if } n = 0 \\ I_q - z e_{n-1,q}(z) V_n & \text{if } n \geq 1\,, \end{cases} \tag{1.8}$$

$$c_n(z) := \begin{cases} \Gamma_0 & \text{if } n = 0 \\ W_n S_{n-1}^* z \varepsilon_{n-1,q}(z) + \Gamma_0 & \text{if } n \geq 1\,, \end{cases} \tag{1.9}$$

and

$$d_n(z) := \begin{cases} I_q & \text{if } n = 0 \\ -W_n z \varepsilon_{n-1,q}(z) + I_q & \text{if } n \geq 1 \end{cases} \tag{1.10}$$

with some $V_n \in \mathcal{Y}_n$ and $W_n \in \mathcal{Z}_n$ if $n \geq 1$ play an essential role in [FKL1]. Note that, if $n \geq 1$, the matrix $T_{n-1}^+ Y_n$ belongs to $\mathcal{Y}_n$ and the matrix $Z_n T_{n-1}^+$ belongs to $\mathcal{Z}_n$ (cf. [FK3, Remark 1.4]). Moreover, in the case $n \geq 1$, [FK3, Proposition 2.2] implies that $T_{n-1}^+ Y_n$ actually belongs to the set $\widetilde{\mathcal{Y}}_n$ of all $V_n \in \mathcal{Y}_n$ such that $\det b_n$ vanishes nowhere in $\mathbb{D}$ and from [FK3, Theorem 2.3] one can see that $Z_n T_{n-1}^+$ actually belongs to the set $\widetilde{\mathcal{Z}}_n$ of all $W_n \in \mathcal{Z}_n$ such that $\det d_n$ vanishes nowhere in $\mathbb{D}$, where b_n and d_n are the matrix polynomials defined by (1.8) and (1.10).

A main goal of this paper is to present a parametrization, for every choice of w in $\mathbb{D}$, of the set $\big\{\Omega(w) : \Omega \in \mathcal{C}_q[\mathbb{D}, (\Gamma_j)_{j=0}^n]\big\}$ in terms of the matrix polynomials a_n, b_n, c_n, and d_n defined by (1.7), (1.8), (1.9), and (1.10) with some $V_n \in \widetilde{\mathcal{Y}}_n$ and

$W_n \in \widetilde{\mathcal{Z}}_n$ if $n \geq 1$. We will prove that these sets are matrix balls. Recall that, for each $M \in \mathbb{C}^{q\times q}$, each $A \in \mathbb{C}^{q\times q}$, and each $B \in \mathbb{C}^{q\times q}$, the set $\mathfrak{K}(M; A, B)$ of all $X \in \mathbb{C}^{q\times q}$ which admit a representation $X = M + AKB$ with some contractive complex $q \times q$ matrix K is called the *matrix ball with center* M, *left semi-radius* A, *and right semi-radius* B. In order to describe the parameters of the concrete matrix balls in question we introduce the rational matrix-valued function

$$\Theta_n := \sqrt{R_{n+1}}\, b_n^{-1}\, \tilde{d}_n^{[n]} \sqrt{L_{n+1}}^{+}.$$

It will turn out that the function Θ_n is holomorphic in $\mathbb{D}$ and that the restriction of Θ_n onto $\mathbb{D}$ belongs to the Schur class $\mathcal{S}_{q\times q}(\mathbb{D})$ of all $q \times q$ Schur functions (in $\mathbb{D}$). Recall that a matrix-valued function $S : \mathbb{D} \to \mathbb{C}^{q\times q}$ which is holomorphic in $\mathbb{D}$ is called $q \times q$ *Schur function* (*in* $\mathbb{D}$) if at each point $z \in \mathbb{D}$ the value $S(z)$ of S is a contractive matrix. We will prove that the following theorem holds.

Theorem 1.1. *Let $n \in \mathbb{N}_0$ and let $(\Gamma_j)_{j=0}^n$ be a $q\times q$ Carathéodory sequence. If $n \geq 1$, then let $V_n \in \widetilde{\mathcal{Y}}_n$ and let $W_n \in \widetilde{\mathcal{Z}}_n$. Let the matrix polynomials a_n, b_n, c_n, and d_n be defined by* (1.7), (1.8), (1.9), *and* (1.10). *Furthermore, let the complex matrices L_{n+1} and R_{n+1} be given by* (1.4) *and* (1.6). *For each $w \in \mathbb{D}$, then by setting $\Theta_n(w) := \sqrt{R_{n+1}}\big(b_n(w)\big)^{-1}\tilde{d}_n^{[n]}(w)\sqrt{L_{n+1}}^{+}$, $F_n^{\blacksquare}(w) := \sqrt{L_{n+1}}^{+}\Theta_n^*(w)\sqrt{R_{n+1}}$, and $G_n^{\blacksquare}(w) := \sqrt{L_{n+1}}\Theta_n^*(w)\sqrt{R_{n+1}}^{+}$ the identity*

$$\Big\{\Omega(w) : \Omega \in \mathcal{C}_q[\mathbb{D}, (\Gamma_j)_{j=0}^n]\Big\} = \mathfrak{K}\Big(\mathcal{M}_{n+1}(w); 2|w|^{n+1}\mathcal{A}_{n+1}(w), \mathcal{B}_{n+1}(w)\Big)$$

is fulfilled, where

$$\mathcal{M}_{n+1}(w) := \Big(d_n(w) - |w|^2 G_n^{\blacksquare}(w)\tilde{b}_n^{[n]}(w)\Big)^{-1}\Big(c_n(w) + |w|^2 G_n^{\blacksquare}(w)\tilde{a}_n^{[n]}(w)\Big), \quad (1.11)$$

$$\mathcal{A}_{n+1}(w) := \big(d_n(w)\big)^{-1}\sqrt{L_{n+1}}\sqrt{I_q - |w|^2\Theta_n^*(w)\Theta_n(w)}^{-1}, \quad (1.12)$$

and

$$\mathcal{B}_{n+1}(w) := \sqrt{I_q - |w|^2\Theta_n(w)\Theta_n^*(w)}^{-1}\sqrt{R_{n+1}}\,\big(b_n(w)\big)^{-1} \quad (1.13)$$

and where the matrix $\mathcal{M}_{n+1}(w)$ also admits the representation

$$\mathcal{M}_{n+1}(w) = \Big(a_n(w) + |w|^2\tilde{c}_n^{[n]}(w)F_n^{\blacksquare}(w)\Big)\Big(b_n(w) - |w|^2\tilde{d}_n^{[n]}(w)F_n^{\blacksquare}(w)\Big)^{-1}. \quad (1.14)$$

2. On a parametrization of the solution set $\mathcal{C}_q[\mathbb{D}, (\Gamma_j)_{j=0}^n]$

In the present section we recall at first a description of the set $\mathcal{C}_q[\mathbb{D}, (\Gamma_j)_{j=0}^n]$ of all solutions of the matricial Carathéodory problem via linear fractional transformations which is derived in [FKL1]. Moreover, we give some additional comments concerning the uniqueness of the function $f \in \mathcal{S}_{q\times q}(\mathbb{D})$ appearing as parameter in these linear fractional transformations.

Theorem 2.1. *Let $n \in \mathbb{N}_0$ and let $(\Gamma_j)_{j=0}^n$ be a $q \times q$ Carathéodory sequence. If $n \geq 1$, then let $V_n \in \widetilde{\mathcal{Y}}_n$ and $W_n \in \widetilde{\mathcal{Z}}_n$. Let the matrix polynomials a_n, b_n, c_n, and d_n be given by* (1.7), (1.8), (1.9), *and* (1.10). *Furthermore, let the matrices L_{n+1} and R_{n+1} be defined by* (1.4) *and* (1.6).

(a) *If $f \in \mathcal{S}_{q\times q}(\mathbb{D})$, then the matrices $z\tilde{d}_n^{[n]}(z)\sqrt{L_{n+1}}^{+}f(z)\sqrt{R_{n+1}} + b_n(z)$ and $z\sqrt{L_{n+1}}f(z)\sqrt{R_{n+1}}^{+}\tilde{b}_n^{[n]}(z) + d_n(z)$ are nonsingular for each $z \in \mathbb{D}$ and the function $\Omega : \mathbb{D} \to \mathbb{C}^{q\times q}$ given by*

$$\Omega(z) := \Big(-z\tilde{c}_n^{[n]}(z)F(z) + a_n(z)\Big)\Big(z\tilde{d}_n^{[n]}(z)F(z) + b_n(z)\Big)^{-1} \tag{2.1}$$

belongs to $\mathcal{C}_q[\mathbb{D}, (\Gamma_j)_{j=0}^n]$ and satisfies, for each $z \in \mathbb{D}$, the representation

$$\Omega(z) = \Big(zG(z)\tilde{b}_n^{[n]}(z) + d_n(z)\Big)^{-1}\Big(-zG(z)\tilde{a}_n^{[n]}(z) + c_n(z)\Big), \tag{2.2}$$

where $F := \sqrt{L_{n+1}}^{+}f\sqrt{R_{n+1}}$ and $G := \sqrt{L_{n+1}}f\sqrt{R_{n+1}}^{+}$.

(b) *If $\Omega \in \mathcal{C}_q[\mathbb{D}, (\Gamma_j)_{j=0}^n]$, then there is an $f \in \mathcal{S}_{q\times q}(\mathbb{D})$ so that Ω admits, for each $z \in \mathbb{D}$, the representations*

$$\Omega(z) = \Big(-z\tilde{c}_n^{[n]}(z)F(z) + a_n(z)\Big)\Big(z\tilde{d}_n^{[n]}(z)F(z) + b_n(z)\Big)^{-1}$$

and (2.2), *where $F := \sqrt{L_{n+1}}^{+}f\sqrt{R_{n+1}}$ and $G := \sqrt{L_{n+1}}f\sqrt{R_{n+1}}^{+}$.*

A proof of Theorem 2.1 is given by [FKL1, Theorems 3.2 and 3.7]. Contrary to the well-known circumstance in the nondegenerate case (see, e.g., [AK], [Ko], [Dy], [BGR], [FF], and [FK1]), the underlying function $f \in \mathcal{S}_{q\times q}(\mathbb{D})$ by the linear fractional transformations stated in (2.1) and (2.2), respectively, is not uniquely determined via f in general. The following results clarify this fact.

Proposition 2.2. *Let $n \in \mathbb{N}_0$ and let $(\Gamma_j)_{j=0}^n$ be a $q \times q$ Carathéodory sequence. If $n \geq 1$, then let $V_n \in \widetilde{\mathcal{Y}}_n$ and let $W_n \in \widetilde{\mathcal{Z}}_n$. Let the matrix polynomials a_n, b_n, c_n, and d_n be defined by* (1.7), (1.8), (1.9), *and* (1.10). *Let the matrices L_{n+1} and R_{n+1} be given by* (1.4) *and* (1.6). *Further, for $j \in \{1, 2\}$, let f_j be a $q \times q$ Schur function (in $\mathbb{D}$), let $F_j := \sqrt{L_{n+1}}^{+}f_j\sqrt{R_{n+1}}$, and let $\Omega_j : \mathbb{D} \to \mathbb{C}^{q\times q}$ be defined by*

$$\Omega_j(z) := \Big(-z\tilde{c}_n^{[n]}(z)F_j(z) + a_n(z)\Big)\Big(z\tilde{d}_n^{[n]}(z)F_j(z) + b_n(z)\Big)^{-1}.$$

(a) *Let $w \in \mathbb{D} \setminus \{0\}$. Then the following statements are equivalent:*
   (i) $\Omega_1(w) = \Omega_2(w)$.
   (ii) $L_{n+1}L_{n+1}^{+}f_1(w)R_{n+1}R_{n+1}^{+} = L_{n+1}L_{n+1}^{+}f_2(w)R_{n+1}R_{n+1}^{+}$.

(b) *The following statements are equivalent:*
   (iii) $\Omega_1 = \Omega_2$.
   (iv) $L_{n+1}L_{n+1}^{+}f_1R_{n+1}R_{n+1}^{+} = L_{n+1}L_{n+1}^{+}f_2R_{n+1}R_{n+1}^{+}$.

Proof. For $j \in \{1, 2\}$, we also use the setting

$$G_j := \sqrt{L_{n+1}}f_j\sqrt{R_{n+1}}^{+}.$$

First we prove that (a) holds. Let $w \in \mathbb{D} \setminus \{0\}$.
(i) $\Rightarrow$ (ii): Suppose that (i) holds. Thus, part (a) of Theorem 2.1 yields

$$\begin{aligned}
&\Big(wG_1(w)\tilde{b}_n^{[n]}(w) + d_n(w)\Big)^{-1}\Big(-wG_1(w)\tilde{a}_n^{[n]}(w) + c_n(w)\Big) \\
&\quad = \Big(-w\tilde{c}_n^{[n]}(w)F_1(w) + a_n(w)\Big)\Big(w\tilde{d}_n^{[n]}(w)F_1(w) + b_n(w)\Big)^{-1} \\
&\quad = \Big(-w\tilde{c}_n^{[n]}(w)F_2(w) + a_n(w)\Big)\Big(w\tilde{d}_n^{[n]}(w)F_2(w) + b_n(w)\Big)^{-1}
\end{aligned}$$

and consequently

$$\begin{aligned}
&- w^2G_1(w)\tilde{a}_n^{[n]}(w)\tilde{d}_n^{[n]}(w)F_2(w) + c_n(w)b_n(w) - wG_1(w)\tilde{a}_n^{[n]}(w)b_n(w) \\
&+ wc_n(w)\tilde{d}_n^{[n]}(w)F_2(w) \\
&\quad = \Big(-wG_1(w)\tilde{a}_n^{[n]}(w) + c_n(w)\Big)\Big(w\tilde{d}_n^{[n]}(w)F_2(w) + b_n(w)\Big) \\
&\quad = \Big(wG_1(w)\tilde{b}_n^{[n]}(w) + d_n(w)\Big)\Big(-w\tilde{c}_n^{[n]}(w)F_2(w) + a_n(w)\Big) \\
&\quad = - w^2G_1(w)\tilde{b}_n^{[n]}(w)\tilde{c}_n^{[n]}(w)F_2(w) + d_n(w)a_n(w) + wG_1(w)\tilde{b}_n^{[n]}(w)a_n(w) \\
&\qquad - wd_n(w)\tilde{c}_n^{[n]}(w)F_2(w). \qquad (2.3)
\end{aligned}$$

Moreover, from [FK3, Theorems 1.7 and 2.3] and [DFK, Lemma 1.2.2] we get

$$d_n(w)a_n(w) = c_n(w)b_n(w) \quad \text{and} \quad \tilde{a}_n^{[n]}(w)\tilde{d}_n^{[n]}(w) = \tilde{b}_n^{[n]}(w)\tilde{c}_n^{[n]}(w). \qquad (2.4)$$

By virtue of [FKL1, part (b) of Proposition 2.4] we have

$$\tilde{a}_n^{[n]}(w)b_n(w) + \tilde{b}_n^{[n]}(w)a_n(w) = 2w^nR_{n+1}, \qquad (2.5)$$

$$c_n(w)\tilde{d}_n^{[n]}(w) + d_n(w)\tilde{c}_n^{[n]}(w) = 2w^nL_{n+1}. \qquad (2.6)$$

Using (2.4), (2.5), and (2.6) from (2.3) we can conclude

$$2w^{n+1}G_1(w)R_{n+1} = 2w^{n+1}L_{n+1}F_2(w).$$

Multiplying the last equation from the left by $\frac{1}{2w^{n+1}}\sqrt{L_{n+1}}^{+}$ and from the right by $\sqrt{R_{n+1}}^{+}$ we get the identity stated in (ii).
(ii) $\Rightarrow$ (i): From (ii) it follows $F_1(w) = F_2(w)$ and hence (i).
Therefore, part (a) is verified. Part (b) is then an easy consequence of (a) and a continuity argument (note part (a) of Theorem 2.1). □

Remark 2.3. Let the assumptions of Proposition 2.2 be fulfilled. Then a combination of Theorem 2.1 and part (b) of Proposition 2.2 shows that there is a bijective correspondence between the solution set $\mathcal{C}_q[\mathbb{D}, (\Gamma_j)_{j=0}^n]$ and the set of all matrix-valued functions $f \in \mathcal{S}_{q\times q}(\mathbb{D})$ satisfying $L_{n+1}L_{n+1}^{+}fR_{n+1}R_{n+1}^{+} = f$.

Remark 2.4. Let the assumptions of Proposition 2.2 be fulfilled. Furthermore, for $j \in \{1, 2\}$, let $G_j := \sqrt{L_{n+1}}f_j\sqrt{R_{n+1}}^{+}$. In view of Theorem 2.1 we mention that from Proposition 2.2 one can see that the following statements are equivalent:

(i) $\Omega_1 = \Omega_2$.
(ii) $F_1 = F_2$.
(iii) $G_1 = G_2$.
(iv) $L_{n+1}f_1R_{n+1} = L_{n+1}f_2R_{n+1}$.

3. Particular matrix-valued Schur functions associated with given $q \times q$ Carathéodory sequences

In this section, we will discuss particular rational matrix-valued functions the restriction of which onto $\mathbb{D}$ are $q\times q$ Schur functions (in $\mathbb{D}$) and which are constructed from given finite $q \times q$ Carathéodory sequences based on the complex $q \times q$ matrix polynomials a_n, b_n, c_n, and d_n defined by (1.7), (1.8), (1.9), and (1.10). For this reason, we firstly prove certain identities for these matrix polynomials.

If $\mathcal{G}$ is a nonempty subset of $\mathbb{C}$ and if $e : \mathcal{G} \to \mathbb{C}^{q\times q}$ is a matrix-valued function, then let

$$\mathcal{N}_e := \{w \in \mathcal{G} : \det e(w) = 0\}.$$

Remark 3.1. Let $n \in \mathbb{N}_0$ and let $(\Gamma_j)_{j=0}^n$ be a $q\times q$ Carathéodory sequence. If $n \geq 1$, then let $V_n \in \mathcal{Y}_n$ and let $W_n \in \mathcal{Z}_n$. Furthermore, let the matrix polynomials a_n, b_n, c_n, and d_n be defined by (1.7), (1.8), (1.9), and (1.10). In view of $b_n(0) = I$, $d_n(0) = I$, and [DFK, Lemma 1.2.2] one can immediately conclude that the set $\mathfrak{N} := \mathcal{N}_{b_n} \cup \mathcal{N}_{\tilde{b}_n^{[n]}} \cup \mathcal{N}_{d_n} \cup \mathcal{N}_{\tilde{d}_n^{[n]}}$ consists of at most $4nq$ complex numbers. Similarly, if the matrix Γ_0 is nonsingular, then from $a_n(0) = \Gamma_0$, $c_n(0) = \Gamma_0$, and [DFK, Lemma 1.2.2] one can see that the set $\mathfrak{M} := \mathcal{N}_{a_n} \cup \mathcal{N}_{\tilde{a}_n^{[n]}} \cup \mathcal{N}_{c_n} \cup \mathcal{N}_{\tilde{c}_n^{[n]}}$ consists of at most $4nq$ complex numbers.

Lemma 3.2. *Let $n \in \mathbb{N}_0$ and let $(\Gamma_j)_{j=0}^n$ be a $q \times q$ Carathéodory sequence. Let the matrices L_{n+1} and R_{n+1} be given by* (1.4) *and* (1.6). *Furthermore, if $n \geq 1$, then let $V_n \in \mathcal{Y}_n$ and let $W_n \in \mathcal{Z}_n$.*

(a) *The matrix polynomials b_n and d_n defined by* (1.8) *and* (1.10) *satisfy for each $w \in \mathbb{C} \setminus \mathfrak{N}$ the identity*

$$\big(\tilde{b}_n^{[n]}(w)\big)^{-1} R_{n+1}\big(b_n(w)\big)^{-1} = \big(d_n(w)\big)^{-1} L_{n+1}\big(\tilde{d}_n^{[n]}(w)\big)^{-1}, \tag{3.1}$$

where $\mathfrak{N} := \mathcal{N}_{b_n} \cup \mathcal{N}_{\tilde{b}_n^{[n]}} \cup \mathcal{N}_{d_n} \cup \mathcal{N}_{\tilde{d}_n^{[n]}}$.

(b) *Let the matrix polynomials a_n and c_n be defined by* (1.7) *and* (1.9). *Suppose that the set $\mathfrak{M} := \mathcal{N}_{a_n} \cup \mathcal{N}_{\tilde{a}_n^{[n]}} \cup \mathcal{N}_{c_n} \cup \mathcal{N}_{\tilde{c}_n^{[n]}}$ does not coincide with $\mathbb{C}$. Then*

$$\big(\tilde{a}_n^{[n]}(w)\big)^{-1} R_{n+1}\big(a_n(w)\big)^{-1} = \big(c_n(w)\big)^{-1} L_{n+1}\big(\tilde{c}_n^{[n]}(w)\big)^{-1} \tag{3.2}$$

is satisfied for each $\mathbb{C} \setminus \mathfrak{M}$.

Proof. We know from Remark 3.1 that $\mathfrak{N}$ consists of at most $4nq$ complex numbers. Applying [FK3, Theorems 1.7 and 2.3] and [DFK, Lemma 1.2.2] we get (2.4) for each $w \in \mathbb{C}$. From [FKL1, part (b) of Proposition 2.4] we see that the identities (2.5) and (2.6) are satisfied for each $w \in \mathbb{C}$. Since (2.4), (2.5), and (2.6) imply

$$\begin{aligned}
2w^n R_{n+1} &= \tilde{a}_n^{[n]}(w)b_n(w) + \tilde{b}_n^{[n]}(w)a_n(w)\\
&= \tilde{b}_n^{[n]}(w)\tilde{c}_n^{[n]}(w)\big(\tilde{d}_n^{[n]}(w)\big)^{-1} b_n(w) + \tilde{b}_n^{[n]}(w)\big(d_n(w)\big)^{-1} c_n(w)b_n(w)\\
&= \tilde{b}_n^{[n]}(w)\big(d_n(w)\big)^{-1}\Big(d_n(w)\tilde{c}_n^{[n]}(w) + c_n(w)\tilde{d}_n^{[n]}(w)\Big)\big(\tilde{d}_n^{[n]}(w)\big)^{-1} b_n(w)\\
&= 2w^n\tilde{b}_n^{[n]}(w)\big(d_n(w)\big)^{-1} L_{n+1}\big(\tilde{d}_n^{[n]}(w)\big)^{-1} b_n(w)
\end{aligned} \tag{3.3}$$

for each $w \in \mathbb{C} \setminus (\mathcal{N}_{d_n} \cup \mathcal{N}_{\tilde{d}_n^{[n]}})$, we can infer that the equality (3.1) is satisfied for each $w \in \mathbb{C} \setminus (\mathfrak{N} \cup \{0\})$. Finally, a continuity argument yields that (3.1) holds actually for every choice of w in $\mathbb{C} \setminus \mathfrak{N}$. Now suppose that $\mathfrak{M} \neq \mathbb{C}$. From the first equation in (3.3), (2.4), and (2.6), we obtain then

$$\begin{aligned} 2w^n R_{n+1} &= \tilde{a}_n^{[n]}(w)\big(c_n(w)\big)^{-1} d_n(w) a_n(w) + \tilde{a}_n^{[n]}(w)\tilde{d}_n^{[n]}(w)\big(\tilde{c}_n^{[n]}(w)\big)^{-1} a_n(w) \\ &= \tilde{a}_n^{[n]}(w)\big(c_n(w)\big)^{-1} \Big(d_n(w)\tilde{c}_n^{[n]}(w) + c_n(w)\tilde{d}_n^{[n]}(w)\Big)\big(\tilde{c}_n^{[n]}(w)\big)^{-1} a_n(w) \\ &= 2w^n \tilde{a}_n^{[n]}(w)\big(c_n(w)\big)^{-1} L_{n+1}\big(\tilde{c}_n^{[n]}(w)\big)^{-1} a_n(w) \end{aligned}$$

for each $w \in \mathbb{C} \setminus \mathfrak{M}$. By continuity, (3.2) follows for every choice of w in $\mathbb{C} \setminus \mathfrak{M}$. □

Remark 3.3. Let $n \in \mathbb{N}_0$ and let $(\Gamma_j)_{j=0}^n$ be a $q \times q$ Carathéodory sequence. If $n \geq 1$, then let $V_n, \mathbf{V}_n \in \mathcal{Y}_n$ and let $W_n, \mathbf{W}_n \in \mathcal{Z}_n$. Let a_n, b_n, c_n and d_n be the matrix polynomials which are defined by (1.7), (1.8), (1.9), and (1.10). Furthermore, let $\mathbf{a}_n$, $\mathbf{b}_n$, $\mathbf{c}_n$, and $\mathbf{d}_n$ be the matrix polynomials which are defined analogously as a_n, b_n, c_n, and d_n using the matrices $\mathbf{V}_n$ and $\mathbf{W}_n$ instead of V_n and W_n, respectively. From the definition of these matrix polynomials and Lemma 3.2 one can see that

$$\big(\tilde{b}_n^{[n]}(w)\big)^{-1} R_{n+1}\big(b_n(w)\big)^{-1} = \big(\tilde{\mathbf{b}}_n^{[n]}(w)\big)^{-1} R_{n+1}\big(\mathbf{b}_n(w)\big)^{-1}$$

for all $w \in \mathbb{C} \setminus (\mathcal{N}_{b_n} \cup \mathcal{N}_{\tilde{b}_n^{[n]}} \cup \mathcal{N}_{\mathbf{b}_n} \cup \mathcal{N}_{\tilde{\mathbf{b}}_n^{[n]}})$ and that

$$\big(d_n(w)\big)^{-1} L_{n+1}\big(\tilde{d}_n^{[n]}(w)\big)^{-1} = \big(\mathbf{d}_n(w)\big)^{-1} L_{n+1}\big(\tilde{\mathbf{d}}_n^{[n]}(w)\big)^{-1}$$

for all $w \in \mathbb{C} \setminus (\mathcal{N}_{d_n} \cup \mathcal{N}_{\tilde{d}_n^{[n]}} \cup \mathcal{N}_{\mathbf{d}_n} \cup \mathcal{N}_{\tilde{\mathbf{d}}_n^{[n]}})$ hold, where the matrices L_{n+1} and R_{n+1} are given by (1.4) and (1.6). Similarly, if $\mathfrak{N}_1 := \mathcal{N}_{a_n} \cup \mathcal{N}_{\tilde{a}_n^{[n]}} \cup \mathcal{N}_{\mathbf{a}_n} \cup \mathcal{N}_{\tilde{\mathbf{a}}_n^{[n]}}$ does not coincide with $\mathbb{C}$ and if $w \in \mathbb{C} \setminus \mathfrak{N}_1$, then

$$\big(\tilde{a}_n^{[n]}(w)\big)^{-1} R_{n+1}\big(a_n(w)\big)^{-1} = \big(\tilde{\mathbf{a}}_n^{[n]}(w)\big)^{-1} R_{n+1}\big(\mathbf{a}_n(w)\big)^{-1}$$

and if $\mathfrak{N}_2 := \mathcal{N}_{c_n} \cup \mathcal{N}_{\tilde{c}_n^{[n]}} \cup \mathcal{N}_{\mathbf{c}_n} \cup \mathcal{N}_{\tilde{\mathbf{c}}_n^{[n]}}$ does not coincide with $\mathbb{C}$ and if $w \in \mathbb{C} \setminus \mathfrak{N}_2$, then

$$\big(c_n(w)\big)^{-1} L_{n+1}\big(\tilde{c}_n^{[n]}(w)\big)^{-1} = \big(\mathbf{c}_n(w)\big)^{-1} L_{n+1}\big(\tilde{\mathbf{c}}_n^{[n]}(w)\big)^{-1}.$$

For every positive real number ρ let

$$K(0;\rho) := \{w \in \mathbb{C} : |w| < \rho\}.$$

Proposition 3.4. *Let $n \in \mathbb{N}_0$ and let $(\Gamma_j)_{j=0}^n$ be a $q \times q$ Carathéodory sequence. If $n \geq 1$, then let $V_n \in \widetilde{\mathcal{Y}}_n$ and let $W_n \in \widetilde{\mathcal{Z}}_n$. Let the $q \times q$ matrix polynomials b_n and d_n be given by* (1.8) *and* (1.10)*. Furthermore, let the matrices L_{n+1} and R_{n+1} be defined by* (1.4) *and* (1.6)*. Then there is a real number $\rho > 1$ such that the rational matrix-valued function*

$$\Theta_n := \sqrt{R_{n+1}}\, b_n^{-1} \tilde{d}_n^{[n]} \sqrt{L_{n+1}}^{+} \tag{3.4}$$

is holomorphic in $K(0;\rho)$. Moreover, Θ_n admits the representation

$$\Theta_n = \sqrt{R_{n+1}}^{+} \tilde{b}_n^{[n]} d_n^{-1} \sqrt{L_{n+1}}, \tag{3.5}$$

for each $w \in \mathbb{D}$ the inequalities

$$\Theta_n(w)\Theta_n^*(w) \le R_{n+1}R_{n+1}^+ \quad \textit{and} \quad \Theta_n^*(w)\Theta_n(w) \le L_{n+1}L_{n+1}^+ \tag{3.6}$$

are fulfilled, and for each $z \in \mathbb{T}$ the equations

$$\Theta_n(z)\Theta_n^*(z) = R_{n+1}R_{n+1}^+ \quad \textit{and} \quad \Theta_n^*(z)\Theta_n(z) = L_{n+1}L_{n+1}^+ \tag{3.7}$$

hold. In particular, the restriction of Θ_n onto $\mathbb{D}$ is a $q \times q$ Schur function (in $\mathbb{D}$).

Proof. Because of the choice of V_n and W_n (in the case $n \ge 1$), for each $w \in \mathbb{D}$ the matrices $b_n(w)$ and $d_n(w)$ are both nonsingular and

$$\Theta_n(w) = \sqrt{R_{n+1}}\big(b_n(w)\big)^{-1}\tilde{d}_n^{[n]}(w)\sqrt{L_{n+1}}^{+}$$

holds. Using Lemma 3.2 we get then

$$\begin{aligned}
\Theta_n(w) &= \sqrt{R_{n+1}}^{+}\tilde{b}_n^{[n]}(w)\big(\tilde{b}_n^{[n]}(w)\big)^{-1}R_{n+1}\big(b_n(w)\big)^{-1}\tilde{d}_n^{[n]}(w)\sqrt{L_{n+1}}^{+}\\
&= \sqrt{R_{n+1}}^{+}\tilde{b}_n^{[n]}(w)\big(d_n(w)\big)^{-1}L_{n+1}\big(\tilde{d}_n^{[n]}(w)\big)^{-1}\tilde{d}_n^{[n]}(w)\sqrt{L_{n+1}}^{+}\\
&= \sqrt{R_{n+1}}^{+}\tilde{b}_n^{[n]}(w)\big(d_n(w)\big)^{-1}\sqrt{L_{n+1}}
\end{aligned}$$

for every choice of w in $\mathbb{D}$. Hence (3.5) follows by a continuity argument. According to Remark 3.1 the set $\mathfrak{N} := \mathcal{N}_{b_n} \cup \mathcal{N}_{\tilde{b}_n^{[n]}} \cup \mathcal{N}_{d_n} \cup \mathcal{N}_{\tilde{d}_n^{[n]}}$ consists of at most $4nq$ complex numbers. Because of (3.5), [DFK, Lemma 1.2.2], and Lemma 3.2 we obtain

$$\begin{aligned}
\Theta_n(z)\Theta_n^*(z) &= \sqrt{R_{n+1}}^{+}\tilde{b}_n^{[n]}(z)\big(d_n(z)\big)^{-1}L_{n+1}\big(d_n(z)\big)^{-*}\big(\tilde{b}_n^{[n]}(z)\big)^{*}\sqrt{R_{n+1}}^{+}\\
&= \sqrt{R_{n+1}}^{+}\tilde{b}_n^{[n]}(z)\big(d_n(z)\big)^{-1}L_{n+1}\big(\tilde{d}_n^{[n]}(z)\big)^{-1}b_n(z)\sqrt{R_{n+1}}^{+}\\
&= \sqrt{R_{n+1}}^{+}R_{n+1}\sqrt{R_{n+1}}^{+} = R_{n+1}R_{n+1}^+
\end{aligned}$$

for each $z \in \mathbb{T} \setminus \mathfrak{N}$. Based on (3.4) one gets analogously

$$\Theta_n^*(z)\Theta_n(z) = L_{n+1}L_{n+1}^+$$

for every choice of z in $\mathbb{T} \setminus \mathfrak{N}$. Hence the rational matrix-valued function Θ_n fulfills the identities in (3.7) actually for each $z \in \mathbb{T}$ and there is a real number $\rho > 1$ such that Θ_n is holomorphic in $K(0;\rho)$. In particular, from $L_{n+1}L_{n+1}^+ \le I$ we see that for each $z \in \mathbb{T}$ the matrix $\Theta_n(z)$ is contractive. Thus the maximum modulus principle for holomorphic functions yields that the restriction of Θ_n onto $\mathbb{D}$ belongs to $\mathcal{S}_{q\times q}(\mathbb{D})$. Let $w \in \mathbb{D}$. Multiplying the inequality $\Theta_n^*(w)\Theta_n(w) \le I$ from the left and from the right by $L_{n+1}L_{n+1}^+$ and using (3.5) we get the second inequality in (3.6). The first one follows analogously. □

Corollary 3.5. *Let the assumptions of* Proposition 3.4 *be fulfilled. Then the following statements are equivalent:*

(i) *There is a unique $\Omega \in \mathcal{C}_q(\mathbb{D})$ such that* (1.2) *holds for each $j \in \mathbb{N}_{0,n}$.*
(ii) *$\Theta_n(w) = 0$ for each $w \in \mathbb{C}$.*
(iii) *There is some $z_0 \in \mathbb{T}$ such that $\Theta_n(z_0)$ is strictly contractive.*

Proof. The matrix $L_{n+1}L_{n+1}^+$ is strictly contractive if and only if $L_{n+1} = 0$. Thus Proposition 3.4 and [FKL1, Lemma 6.1] yield the asserted equivalences. □

In the next result, we will analyze the construction of a $q \times q$ Schur function given by Proposition 3.4 for the case that the underlying $q \times q$ Carathéodory sequence $(\Gamma_j)_{j=0}^n$ is nondegenerate. Recall firstly that, by a $q \times q$ *Blaschke-Potapov elementary factor* (with respect to the signature matrix I_q) with zero at $w_0 \in \mathbb{D}$ we mean a rational complex $q \times q$ matrix-valued function B_{w_0} which is given via

$$B_{w_0} = V_1\big(I_q + (b_{w_0} - 1)P\big)V_2, \tag{3.8}$$

where V_1 and V_2 are some unitary $q \times q$ matrices, where P is some non-zero idempotent and Hermitian $q \times q$ matrix, and where b_{w_0} is the *elementary Blaschke factor* corresponding to w_0, i.e., b_{w_0} is the rational function defined by

$$b_{w_0}(w) := \begin{cases} w & \text{for each } w \in \mathbb{C} \text{ if } w_0 = 0 \\ \dfrac{\overline{w_0}}{|w_0|}\dfrac{w_0 - w}{1 - \overline{w_0}w} & \text{for each } w \in \mathbb{C} \setminus \left\{\frac{1}{\overline{w_0}}\right\} \text{ if } w_0 \neq 0\,. \end{cases}$$

The notion *finite $q \times q$ Blaschke-Potapov product* (with respect to I_q) will be used to denote a constant unitary $q \times q$ matrix-valued function or a finite product of $q \times q$ Blaschke-Potapov elementary factors defined as in (3.8). Moreover, we note that each $q \times q$ Schur function f (in $\mathbb{D}$) has radial boundary values $\underline{\lambda}$-almost everywhere on $\mathbb{T}$, where $\underline{\lambda}$ stands for the linear Lebesgue-Borel measure on $\mathbb{T}$. If f is a $q \times q$ Schur function in $\mathbb{D}$ such that its radial boundary values are unitary $\underline{\lambda}$-almost everywhere on $\mathbb{T}$, then f is called an *inner $q \times q$ Schur function* (*in* $\mathbb{D}$).

Corollary 3.6. *Let the assumptions of* Proposition 3.4 *be fulfilled. Then the following statements are equivalent:*

(i) *The $q \times q$ Carathéodory sequence $(\Gamma_j)_{j=0}^n$ is nondegenerate.*
(ii) *For each $z \in \mathbb{T}$ the matrix $\Theta_n(z)$ is unitary.*
(iii) *There is some $w \in \mathbb{D} \cup \mathbb{T}$ such that the matrix $\Theta_n(w)$ is nonsingular.*
(iv) *The restriction of Θ_n onto $\mathbb{D}$ is an inner $q \times q$ Schur function.*
(v) *Θ_n is a finite Blaschke-Potapov product* (*with respect to I_q*).

Proof. Proposition 3.4 yields that the identities in (3.7) hold for each $z \in \mathbb{T}$, that the inequalities in (3.6) are satisfied for each $w \in \mathbb{D}$, and that the restriction of Θ_n onto $\mathbb{D}$ belongs to $\mathcal{S}_{q\times q}(\mathbb{D})$. Thus the equivalence of (i), (ii), (iii), and (iv) follows from [FKL1, Lemma 5.1], whereas the equivalence of (iv) and (v) is a consequence of [FFK, Corollary 13 and Proposition 31]. □

In the second part of this section we present now a similar construction of a $q \times q$ Schur function as in Proposition 3.4, where the matrix polynomials a_n and c_n defined by (1.7) and (1.9) are involved instead of the matrix polynomials b_n and d_n defined by (1.8) and (1.10). Note that, in view of (1.7) and (1.9), it is not hard to accept that one has to take for granted then a given $q \times q$ Carathéodory sequence $(\Gamma_j)_{j=0}^n$ with the additional condition that the matrix Γ_0 is nonsingular.

Remark 3.7. Let $n \in \mathbb{N}_0$ and let $(\Gamma_j)_{j=0}^n$ be a $q \times q$ Carathéodory sequence with nonsingular matrix Γ_0. If $n \geq 1$, then let $V_n \in \widetilde{\mathcal{Y}}_n$ and let $W_n \in \widetilde{\mathcal{Z}}_n$. From [FK1,

Remark 30 in Part V] and [FK3, Remark 1.1, Theorem 1.7, and Theorem 2.3] one can see that the functions $\det a_n$ and $\det c_n$ vanish nowhere in $\mathbb{D}$.

Proposition 3.8. *Let $n \in \mathbb{N}_0$ and let $(\Gamma_j)_{j=0}^n$ be a $q \times q$ Carathéodory sequence with nonsingular matrix Γ_0. If $n \geq 1$, then let $V_n \in \widetilde{\mathcal{Y}}_n$ and let $W_n \in \widetilde{\mathcal{Z}}_n$. Let the matrix polynomials a_n and c_n be defined by* (1.7) *and* (1.9)*. Furthermore, let the matrices L_{n+1} and R_{n+1} be given by* (1.4) *and* (1.6)*. Then there is a real number $\rho > 1$ such that the matrix-valued function*

$$\Theta_n^\diamond := \sqrt{R_{n+1}}\, a_n^{-1} \tilde{c}_n^{[n]} \sqrt{L_{n+1}}^{+} \tag{3.9}$$

is holomorphic in $K(0;\rho)$. Moreover, $\Theta_n^\diamond$ admits the representation

$$\Theta_n^\diamond = \sqrt{R_{n+1}}^{+} \tilde{a}_n^{[n]} c_n^{-1} \sqrt{L_{n+1}}, \tag{3.10}$$

for each $w \in \mathbb{D}$ the inequalities

$$\Theta_n^\diamond(w)\big(\Theta_n^\diamond(w)\big)^* \leq R_{n+1}R_{n+1}^+ \quad \text{and} \quad \big(\Theta_n^\diamond(w)\big)^*\Theta_n^\diamond(w) \leq L_{n+1}L_{n+1}^+$$

are fulfilled, and for each $z \in \mathbb{T}$ the equations

$$\Theta_n^\diamond(z)\big(\Theta_n^\diamond(z)\big)^* = R_{n+1}R_{n+1}^+ \quad \text{and} \quad \big(\Theta_n^\diamond(z)\big)^*\Theta_n^\diamond(z) = L_{n+1}L_{n+1}^+$$

hold. In particular, the restriction of $\Theta_n^\diamond$ onto $\mathbb{D}$ is a $q \times q$ Schur function (in $\mathbb{D}$).

Proof. Because of Remark 3.7 the functions $\det a_n$ and $\det c_n$ vanish nowhere in $\mathbb{D}$. Consequently, the relation

$$\Theta_n^\diamond(w) = \sqrt{R_{n+1}}\big(a_n(w)\big)^{-1} \tilde{c}_n^{[n]}(w) \sqrt{L_{n+1}}^{+}$$

is satisfied for each $w \in \mathbb{D}$. Using Remark 3.1 and part (b) of Lemma 3.2 we get then (3.10) and that the set $\mathfrak{M} := \mathcal{N}_{a_n} \cup \mathcal{N}_{\tilde{a}_n^{[n]}} \cup \mathcal{N}_{c_n} \cup \mathcal{N}_{c_n^{[n]}}$ consists of at most $4nq$ complex numbers. The rest of the assertion can be verified analogously to the given proof of Proposition 3.4. □

Corollary 3.9. *Let the assumptions of* Proposition 3.8 *be fulfilled. Then the following statements are equivalent:*

(i) *There is a unique $\Omega \in \mathcal{C}_q(\mathbb{D})$ such that* (1.2) *holds for each $j \in \mathbb{N}_{0,n}$.*
(ii) *$\Theta_n^\diamond(w) = 0$ for each $w \in \mathbb{C}$.*
(iii) *There is some $z_0 \in \mathbb{T}$ such that the matrix $\Theta_n^\diamond(z_0)$ is strictly contractive.*

Proof. Use Proposition 3.8, [FKL1, Lemma 6.1], and the fact that the matrix $L_{n+1}L_{n+1}^+$ is strictly contractive if and only if $L_{n+1} = 0$. □

Corollary 3.10. *Let the assumptions of* Proposition 3.8 *be fulfilled. Then the following statements are equivalent:*

(i) *The $q \times q$ Carathéodory sequence $(\Gamma_j)_{j=0}^n$ is nondegenerate.*
(ii) *For each $z \in \mathbb{T}$ the matrix $\Theta_n^\diamond(z)$ is unitary.*
(iii) *There is some $w \in \mathbb{D} \cup \mathbb{T}$ such that the matrix $\Theta_n^\diamond(w)$ is nonsingular.*
(iv) *The restriction of $\Theta_n^\diamond$ onto $\mathbb{D}$ is an inner $q \times q$ Schur function.*
(v) *$\Theta_n^\diamond$ is a finite Blaschke-Potapov product (with respect to I_q).*

Proof. Using Proposition 3.8, [FKL1, Lemma 5.1], and [FFK, Corollary 13 and Proposition 31] the proof is analogous to the proof of Corollary 3.6. □

Note that from Corollary 3.5 and Corollary 3.9 we see particularly that, under the assumptions of Proposition 3.8, if there is a unique $\Omega \in \mathcal{C}_q(\mathbb{D})$ such that (1.2) is satisfied for each $j \in \mathbb{N}_{0,n}$, then the rational matrix-valued functions Θ_n and $\Theta_n^\diamond$ defined by (3.4) and (3.9) coincide. Generally, the following connection between both functions is available.

Proposition 3.11. *Let $n \in \mathbb{N}_0$ and let $(\Gamma_j)_{j=0}^n$ be a $q \times q$ Carathéodory sequence with nonsingular matrix Γ_0. If $n \geq 1$, then let $V_n \in \widetilde{\mathcal{Y}}_n$ and let $W_n \in \widetilde{\mathcal{Z}}_n$. Let the matrix polynomials a_n, b_n, c_n, and d_n be defined by* (1.7), (1.8), (1.9), *and* (1.10). *Furthermore, let the matrices L_{n+1} and R_{n+1} be given by* (1.4) *and* (1.6) *and let the matrix-valued functions Θ_n and $\Theta_n^\diamond$ be given by* (3.4) *and* (3.9). *For every choice of w in $\mathbb{D}$, then the identities*

$$\Theta_n(w) + \Theta_n^\diamond(w) = 2w^n \sqrt{R_{n+1}}\big(a_n(w)\big)^{-1}\big(d_n(w)\big)^{-1}\sqrt{L_{n+1}} \tag{3.11}$$

and

$$\Theta_n(w) + \Theta_n^\diamond(w) = 2w^n \sqrt{R_{n+1}}\big(b_n(w)\big)^{-1}\big(c_n(w)\big)^{-1}\sqrt{L_{n+1}} \tag{3.12}$$

are fulfilled.

Proof. Obviously, the complex-valued functions $\det b_n$ and $\det d_n$ vanish nowhere in $\mathbb{D}$. Remark 3.7 shows that $\det a_n$ and $\det c_n$ vanish nowhere in $\mathbb{D}$ as well. Let $w \in \mathbb{D}$. Using [FK3, Remark 1.1, Theorem 1.7, and Theorem 2.3] we get

$$\begin{aligned}\Theta_n^\diamond(w) &= \sqrt{R_{n+1}}^{\,+}\tilde{a}_n^{[n]}(w)\big(c_n(w)\big)^{-1}d_n(w)\big(d_n(w)\big)^{-1}\sqrt{L_{n+1}} \\ &= \sqrt{R_{n+1}}^{\,+}\tilde{a}_n^{[n]}(w)b_n(w)\big(a_n(w)\big)^{-1}\big(d_n(w)\big)^{-1}\sqrt{L_{n+1}}\end{aligned} \tag{3.13}$$

and, in view of (3.9), similarly

$$\Theta_n^\diamond(w) = \sqrt{R_{n+1}}\big(b_n(w)\big)^{-1}\big(c_n(w)\big)^{-1}d_n(w)\tilde{c}_n^{[n]}(w)\sqrt{L_{n+1}}^{\,+}. \tag{3.14}$$

Moreover, from [FKL1, part (b) of Proposition 2.4] we obtain

$$\tilde{a}_n^{[n]}(w)b_n(w)\big(a_n(w)\big)^{-1} = 2w^n R_{n+1}\big(a_n(w)\big)^{-1} - \tilde{b}_n^{[n]}(w) \tag{3.15}$$

and

$$\big(c_n(w)\big)^{-1}d_n(w)\tilde{c}_n^{[n]}(w) = 2w^n\big(c_n(w)\big)^{-1}L_{n+1} - \tilde{d}_n^{[n]}(w). \tag{3.16}$$

Because of (3.4), (3.13), and (3.15) it follows (3.11). Analogously, (3.12) is a consequence of (3.5), (3.14), and (3.16). □

Corollary 3.12. *Let the assumptions of* Proposition 3.11 *be fulfilled. Then there is a real number $\rho > 1$ such that the matrix-valued function*

$$\Theta_n^\triangleright := \sqrt{R_{n+1}}a_n^{-1}d_n^{-1}\sqrt{L_{n+1}} \tag{3.17}$$

is holomorphic in $K(0;\rho)$. Moreover, $\Theta_n^\triangleright$ admits the representation

$$\Theta_n^\triangleright = \sqrt{R_{n+1}}b_n^{-1}c_n^{-1}\sqrt{L_{n+1}}$$

and the restriction of $\Theta_n^\triangleright$ onto $\mathbb{D}$ is a $q \times q$ Schur function (in $\mathbb{D}$).

Proof. Use Proposition 3.11 in combination with Proposition 3.4, Proposition 3.8, and a matricial version of Schwarz lemma (see, e.g., [DFK, Lemma 2.3.1]). □

4. On the Weyl matrix balls associated with matricial Carathéodory sequences

In this section, we will prove Theorem 1.1 which describes the Weyl matrix balls of the solutions of the matricial Carathéodory problem in both nondegenerate and degenerate cases. Moreover, we discuss the particular choice of the semi-radii of the matrix balls in question and we compare Theorem 1.1 with the corresponding result for the nondegenerate case. First of all, we remark the following.

Remark 4.1. Let $n \in \mathbb{N}_0$ and let $(\Gamma_j)_{j=0}^n$ be a $q \times q$ Carathéodory sequence. If $n \geq 1$, then let $V_n \in \widetilde{\mathcal{Y}}_n$ and let $W_n \in \widetilde{\mathcal{Z}}_n$. Let the matrix polynomials a_n, b_n, c_n, and d_n be defined by (1.7), (1.8), (1.9), and (1.10). Furthermore, let the matrix-valued function Θ_n be defined by (3.4). For each $w \in \mathbb{D}$, from Proposition 3.4 and [FKL1, Lemma 3.1] one can see that the matrices $I_q - |w|^2\Theta_n^*(w)\Theta_n(w)$ and $I_q - |w|^2\Theta_n(w)\Theta_n^*(w)$ are both positive Hermitian and that the matrices

$$d_n(w) - |w|^2\sqrt{L_{n+1}}\Theta_n^*(w)\sqrt{R_{n+1}}^{+}\tilde{b}_n^{[n]}(w)$$

and

$$b_n(w) - |w|^2\tilde{d}_n^{[n]}(w)\sqrt{L_{n+1}}^{+}\Theta_n^*(w)\sqrt{R_{n+1}}$$

are both nonsingular.

We start now with our proof of Theorem 1.1.

Proof of Theorem 1.1. Let $w \in \mathbb{D}$. Subject to Remark 4.1 the matrices $\mathcal{M}_{n+1}(w)$, $\mathcal{A}_{n+1}(w)$, and $\mathcal{B}_{n+1}(w)$ are well defined and representation (1.14) of $\mathcal{M}_{n+1}(w)$ is an immediate consequence of [FKL1, Lemma 3.1]. In view of Theorem 2.1, if $f \in \mathcal{S}_{q\times q}(\mathbb{D})$, then we use in the following the setting $\Omega_f : \mathbb{D} \to \mathbb{C}^{q\times q}$ defined by

$$\Omega_f(z) := \Big(-z\tilde{c}_n^{[n]}(z)F(z) + a_n(z)\Big)\Big(z\tilde{d}_n^{[n]}(z)F(z) + b_n(z)\Big)^{-1}, \tag{4.1}$$

where $F := \sqrt{L_{n+1}}^{+} f \sqrt{R_{n+1}}$. Therefore, Theorem 2.1 yields

$$\Big\{\Omega(w) : \Omega \in \mathcal{C}_q[\mathbb{D}, (\Gamma_j)_{j=0}^n]\Big\} = \Big\{\Omega_f(w) : f \in \mathcal{S}_{q\times q}(\mathbb{D})\Big\}. \tag{4.2}$$

Let $u \in \mathbb{T}$ be such that the relation $w = |w|u$ is satisfied, let $E := \overline{w}\Theta_n^*(w)$, and let $\Phi := d_n(w) - |w|^2\sqrt{L_{n+1}}\Theta_n^*(w)\sqrt{R_{n+1}}^{+}\tilde{b}_n^{[n]}(w)$, where Θ_n is the rational matrix-valued function defined by (3.4). Remark 4.1 shows that the matrix E is strictly contractive and that the matrix Φ is nonsingular. Moreover, Proposition 3.4 and the choice of V_n and W_n (in the case $n \geq 1$) provide us that

$$\Theta_n(w) = \sqrt{R_{n+1}}\big(b_n(w)\big)^{-1}\tilde{d}_n^{[n]}(w)\sqrt{L_{n+1}}^{+} = \sqrt{R_{n+1}}^{+}\tilde{b}_n^{[n]}(w)\big(d_n(w)\big)^{-1}\sqrt{L_{n+1}}.$$

Therefore, a straightforward calculation yields

$$\sqrt{L_{n+1}}\,(I-EE^*)=\Phi\left(d_n(w)\right)^{-1}\sqrt{L_{n+1}}$$

and hence

$$\Phi^{-1}\sqrt{L_{n+1}}=\left(d_n(w)\right)^{-1}\sqrt{L_{n+1}}\,(I-EE^*)^{-1}. \tag{4.3}$$

Furthermore, from [FKL1, Proposition 2.4] we get (2.5) and (2.6). Because of [FK3, Theorems 1.7 and 2.3] and [DFK, Lemma 1.2.2] we have (2.4). The following considerations of the proof are divided into three steps.

Step A. In this first step we consider an arbitrary $f\in\mathcal{S}_{q\times q}(\mathbb{D})$. Using (4.1) with $F:=\sqrt{L_{n+1}}^{\,+}f\sqrt{R_{n+1}}$, we are going to prove that

$$\begin{aligned}&\Omega_f(w)-\mathcal{M}_{n+1}(w)\\&\quad=-2w^{n+1}\mathcal{A}_{n+1}(w)\sqrt{I-EE^*}^{\,-1}\big(f(w)+E\big)\big(I+E^*f(w)\big)^{-1}\sqrt{I-E^*E}\,\mathcal{B}_{n+1}(w)\end{aligned} \tag{4.4}$$

holds. For this reason, first we note that Theorem 2.1 yields that the matrix

$$\Psi_f:=w\tilde{d}_n^{[n]}(w)\sqrt{L_{n+1}}^{\,+}f(w)\sqrt{R_{n+1}}+b_n(w) \tag{4.5}$$

is nonsingular. Because of (1.11), (4.1), (2.4), (2.5), and (2.6) we obtain then

$$\begin{aligned}&\Phi\big(\Omega_f(w)-\mathcal{M}_{n+1}(w)\big)\Psi_f\\&\quad=\Big(d_n(w)-|w|^2\sqrt{L_{n+1}}\Theta_n^*(w)\sqrt{R_{n+1}}^{\,+}\tilde{b}_n^{[n]}(w)\Big)\Big(-w\tilde{c}_n^{[n]}(w)F(w)+a_n(w)\Big)\\&\qquad-\Big(c_n(w)+|w|^2\sqrt{L_{n+1}}\Theta_n^*(w)\sqrt{R_{n+1}}^{\,+}\tilde{a}_n^{[n]}(w)\Big)\Big(w\tilde{d}_n^{[n]}(w)F(w)+b_n(w)\Big)\\&\quad=-wd_n(w)\tilde{c}_n^{[n]}(w)F(w)-|w|^2\sqrt{L_{n+1}}\Theta_n^*(w)\sqrt{R_{n+1}}^{\,+}\tilde{b}_n^{[n]}(w)a_n(w)\\&\qquad+d_n(w)a_n(w)+w|w|^2\sqrt{L_{n+1}}\Theta_n^*(w)\sqrt{R_{n+1}}^{\,+}\tilde{b}_n^{[n]}(w)\tilde{c}_n^{[n]}(w)F(w)\\&\qquad-wc_n(w)\tilde{d}_n^{[n]}(w)F(w)-|w|^2\sqrt{L_{n+1}}\Theta_n^*(w)\sqrt{R_{n+1}}^{\,+}\tilde{a}_n^{[n]}(w)b_n(w)\\&\qquad-c_n(w)b_n(w)-w|w|^2\sqrt{L_{n+1}}\Theta_n^*(w)\sqrt{R_{n+1}}^{\,+}\tilde{a}_n^{[n]}(w)\tilde{d}_n^{[n]}(w)F(w)\\&\quad=-2w^{n+1}L_{n+1}\sqrt{L_{n+1}}^{\,+}f(w)\sqrt{R_{n+1}}-2w^n|w|^2\sqrt{L_{n+1}}\Theta_n^*(w)\sqrt{R_{n+1}}^{\,+}R_{n+1}\\&\quad=-2w^{n+1}\sqrt{L_{n+1}}\big(f(w)+E\big)\sqrt{R_{n+1}}\end{aligned}$$

and consequently

$$\Omega_f(w)-\mathcal{M}_{n+1}(w)=-2w^{n+1}\Phi^{-1}\sqrt{L_{n+1}}\big(f(w)+E\big)\sqrt{R_{n+1}}\Psi_f^{-1}. \tag{4.6}$$

Moreover, from (3.4), $V_n\in\widetilde{\mathcal{Y}}_n$, $W_n\in\widetilde{\mathcal{Z}}_n$, and (4.5) we can conclude that

$$\begin{aligned}&\big(I+E^*f(w)\big)\sqrt{R_{n+1}}\\&\quad=\sqrt{R_{n+1}}+w\Theta_n(w)f(w)\sqrt{R_{n+1}}\\&\quad=\sqrt{R_{n+1}}\big(b_n(w)\big)^{-1}b_n(w)+w\sqrt{R_{n+1}}\big(b_n(w)\big)^{-1}\tilde{d}_n^{[n]}(w)\sqrt{L_{n+1}}^{\,+}f(w)\sqrt{R_{n+1}}\\&\quad=\sqrt{R_{n+1}}\big(b_n(w)\big)^{-1}\Psi_f.\end{aligned} \tag{4.7}$$

Since the matrix E is strictly contractive, the matrix $I+E^*f(w)$ is nonsingular (see, e.g., [DFK, Lemma 1.1.12, Lemma 1.1.13, and Remark 1.1.2]). Therefore, from (4.7) we get the identity

$$\sqrt{R_{n+1}}\Psi_f^{-1}=\big(I+E^*f(w)\big)^{-1}\sqrt{R_{n+1}}\big(b_n(w)\big)^{-1}. \tag{4.8}$$

Combining (4.6), (4.3), (4.8), (1.12), and (1.13) we obtain

$$\begin{aligned}
&\Omega_f(w) - \mathcal{M}_{n+1}(w)\\
&\quad = -2w^{n+1}\big(d_n(w)\big)^{-1}\sqrt{L_{n+1}}\big(I - EE^*\big)^{-1}\big(f(w) + E\big)\\
&\qquad \cdot \big(I + E^* f(w)\big)^{-1}\sqrt{R_{n+1}}\big(b_n(w)\big)^{-1}\\
&\quad = -2w^{n+1}\big(d_n(w)\big)^{-1}\sqrt{L_{n+1}}\sqrt{I - EE^*}^{-1}\sqrt{I - EE^*}^{-1}\big(f(w) + E\big)\\
&\qquad \cdot \big(I + E^* f(w)\big)^{-1}\sqrt{I - E^*E}\sqrt{I - E^*E}^{-1}\sqrt{R_{n+1}}\big(b_n(w)\big)^{-1}\\
&\quad = -2w^{n+1}\mathcal{A}_{n+1}(w)\sqrt{I-EE^*}^{-1}\big(f(w) + E\big)\big(I + E^* f(w)\big)^{-1}\sqrt{I-E^*E}\,\mathcal{B}_{n+1}(w).
\end{aligned}$$

Thus (4.4) is proved.

Step B. We are going to check that

$$\Big\{\Omega(w) : \Omega \in \mathcal{C}_q[\mathbb{D}, (\Gamma_j)_{j=0}^n]\Big\} \subseteq \mathfrak{K}\Big(\mathcal{M}_{n+1}(w); 2|w|^{n+1}\mathcal{A}_{n+1}(w), \mathcal{B}_{n+1}(w)\Big) \quad (4.9)$$

is satisfied. Let $X \in \big\{\Omega(w) : \Omega \in \mathcal{C}_q[\mathbb{D}, (\Gamma_j)_{j=0}^n]\big\}$. According to (4.2) there is an $f \in \mathcal{S}_{q\times q}(\mathbb{D})$ such that $X = \Omega_f(w)$, where $\Omega_f : \mathbb{D} \to \mathbb{C}^{q\times q}$ is given by (4.1) with $F := \sqrt{L_{n+1}}^{+} f \sqrt{R_{n+1}}$. By virtue of Step A we have then (4.4). Since the matrix

$$H(E) := \begin{pmatrix} \sqrt{I - EE^*}^{-1} & \sqrt{I - EE^*}^{-1}E \\ \sqrt{I - E^*E}^{-1}E^* & \sqrt{I - E^*E}^{-1} \end{pmatrix}$$

fulfills the identity

$$\big(H(E)\big)^* \begin{pmatrix} I_q & 0 \\ 0 & -I_q \end{pmatrix} H(E) = \begin{pmatrix} I_q & 0 \\ 0 & -I_q \end{pmatrix} \quad (4.10)$$

(see, e.g., [DFK, Lemma 3.6.32 and Lemma 1.1.12]), an application of [DFK, part (a) of Theorem 1.6.1] provides us that the complex $q \times q$ matrix

$$\sqrt{I - EE^*}^{-1}\big(f(w) + E\big)\big(I + E^* f(w)\big)^{-1}\sqrt{I - E^*E}$$

is contractive. Since u belongs to $\mathbb{T}$, the matrix

$$K := -u^{n+1}\sqrt{I - EE^*}^{-1}\big(f(w) + E\big)\big(I + E^* f(w)\big)^{-1}\sqrt{I - E^*E}$$

is also contractive and because of (4.4) we have finally

$$X - \mathcal{M}_{n+1}(w) = \Omega_f(w) - \mathcal{M}_{n+1}(w) = 2|w|^{n+1}\mathcal{A}_{n+1}(w)K\mathcal{B}_{n+1}(w).$$

Consequently, (4.9) is proved.

Step C. We are going to check that

$$\mathfrak{K}\Big(\mathcal{M}_{n+1}(w); 2|w|^{n+1}\mathcal{A}_{n+1}(w), \mathcal{B}_{n+1}(w)\Big) \subseteq \Big\{\Omega(w) : \Omega \in \mathcal{C}_q[\mathbb{D}, (\Gamma_j)_{j=0}^n]\Big\} \quad (4.11)$$

is satisfied as well. For this reason, we consider an arbitrary element X of the matrix ball $\mathfrak{K}\big(\mathcal{M}_{n+1}(w); 2|w|^{n+1}\mathcal{A}_{n+1}(w), \mathcal{B}_{n+1}(w)\big)$. Then there is a contractive $q \times q$ matrix C such that

$$X - \mathcal{M}_{n+1}(w) = -2w^{n+1}\mathcal{A}_{n+1}(w)\big(-\overline{u}^{n+1}C\big)\mathcal{B}_{n+1}(w)$$

holds. According to (4.10) and [DFK, part (a) of Theorem 1.6.2] there is a contractive $q \times q$ matrix D such that

$$-\overline{u}^{n+1}C = \sqrt{I - EE^*}^{\,-1}(D+E)(I+E^*D)^{-1}\sqrt{I-E^*E}.$$

Let f be the constant matrix-valued function defined on $\mathbb{D}$ with value D. Then f belongs to $\mathcal{S}_{q\times q}(\mathbb{D})$, where on the one hand the equality

$$\begin{aligned} &X - \mathcal{M}_{n+1}(w) \\ &\quad = -2w^{n+1}\mathcal{A}_{n+1}(w)\sqrt{I-EE^*}^{\,-1}\big(f(w)+E\big)\big(I+E^*f(w)\big)^{-1}\sqrt{I-E^*E}\,\mathcal{B}_{n+1}(w) \end{aligned}$$

holds and where Step A shows on the other hand that the function Ω_f defined by (4.1) with $F := \sqrt{L_{n+1}}^{\,+} f \sqrt{R_{n+1}}$ fulfills (4.4). Comparing both relations we get $\Omega_f(w) = X$. Applying Theorem 2.1 we see that Ω_f belongs to $\mathcal{C}_q[\mathbb{D}, (\Gamma_j)_{j=0}^n]$. Thus X belongs to $\big\{\Omega(w) : \Omega \in \mathcal{C}_q[\mathbb{D}, (\Gamma_j)_{j=0}^n]\big\}$ which implies finally (4.11).
In view of (4.9) and (4.11) the proof is complete. □

Remark 4.2. Let the assumptions of Theorem 1.1 be fulfilled. From Theorem 1.1 and [Sm, Theorem 1.3] (see also [DFK, Theorem 1.5.2]) one can see that

$$\Big\{\Omega(w) : \Omega \in \mathcal{C}_q[\mathbb{D}, (\Gamma_j)_{j=0}^n]\Big\} = \mathfrak{K}\left(\mathcal{M}_{n+1}(w); |w|^{n+1}\sqrt{2\,\mathcal{L}_{n+1}(w)}, \sqrt{2\,\mathcal{R}_{n+1}(w)}\,\right)$$

holds for each $w \in \mathbb{D}$, where $\mathcal{M}_{n+1}(w)$, $\mathcal{L}_{n+1}(w)$, and $\mathcal{R}_{n+1}(w)$ are given by (1.11),

$$\mathcal{L}_{n+1}(w) := \big(d_n(w)\big)^{-1}\sqrt{L_{n+1}}\,\big(I_q - |w|^2\Theta_n^*(w)\Theta_n(w)\big)^{-1}\sqrt{L_{n+1}}\,\big(d_n(w)\big)^{-*},$$

and

$$\mathcal{R}_{n+1}(w) := \big(b_n(w)\big)^{-*}\sqrt{R_{n+1}}\,\big(I_q - |w|^2\Theta_n(w)\Theta_n^*(w)\big)^{-1}\sqrt{R_{n+1}}\,\big(b_n(w)\big)^{-1}.$$

In the special case $n = 0$, for each $w \in \mathbb{D}$, straightforward calculations yield that

$$\mathcal{M}_1(w) = \frac{1+|w|^2}{1-|w|^2}\operatorname{Re}\Gamma_0 + i\operatorname{Im}\Gamma_0, \quad \mathcal{L}_1(w) = \frac{1}{1-|w|^2}\operatorname{Re}\Gamma_0, \quad \text{and} \quad \mathcal{R}_1(w) = \mathcal{L}_1(w).$$

Remark 4.3. Let the assumptions of Theorem 1.1 be fulfilled. By virtue of [FKL1, Lemma 6.1] one can get that the following statements are equivalent:

(i) There is one and only one $\Omega \in \mathcal{C}_q(\mathbb{D})$ such that (1.2) holds for each $j \in \mathbb{N}_{0,n}$.
(ii) For each $w \in \mathbb{D}$, the identities $\mathcal{A}_{n+1}(w) = 0$ and $\mathcal{B}_{n+1}(w) = 0$ hold.
(iii) There is some $w \in \mathbb{D}$ such that $\mathcal{A}_{n+1}(w) = 0$ or $\mathcal{B}_{n+1}(w) = 0$.

Remark 4.4. Let the assumptions of Theorem 1.1 be fulfilled. Taking into account [FK1, Remark 2 in Part I], for each $w \in \mathbb{D}$, one can immediately see that

$$\operatorname{rank}\mathcal{A}_{n+1}(w) = \operatorname{rank}L_{n+1} = \operatorname{rank}R_{n+1} = \operatorname{rank}\mathcal{B}_{n+1}(w).$$

Remark 4.5. Let the assumptions of Theorem 1.1 be fulfilled. From Remark 4.4 and [FKL1, Lemma 5.1] it follows that the following statements are equivalent:

(i) $(\Gamma_j)_{j=0}^n$ is a nondegenerate $q \times q$ Carathéodory sequence.
(ii) For each $w \in \mathbb{D}$, both matrices $\mathcal{A}_{n+1}(w)$ and $\mathcal{B}_{n+1}(w)$ are nonsingular.
(iii) There is some $w \in \mathbb{D}$ such that $\mathcal{A}_{n+1}(w)$ or $\mathcal{B}_{n+1}(w)$ is nonsingular.

Remark 4.6. Let the assumptions of Theorem 1.1 be fulfilled. Then

$$\det \mathcal{A}_{n+1}(w) = \det \mathcal{B}_{n+1}(w)$$

for each $w \in \mathbb{D}$. Indeed, if $(\Gamma_j)_{j=0}^n$ is a nondegenerate $q \times q$ Carathéodory sequence, then this follows from the equation $\det L_{n+1} = \det R_{n+1}$ (see [FK1, Remark 2 in Part I]), from the identity $\det d_n(w) = \det b_n(w)$ which holds for each $w \in \mathbb{C}$ (see [FKL1, Lemma 5.1, Remark 5.2, and Lemma 5.5] and [DGK1, equation (72)]), and from [DFK, Lemma 1.1.8]. In the degenerate case it suffices to apply Remark 4.5.

In the rest of this section let us consider the so-called nondegenerate case which was discussed in [FK1, Part IV and Part V] in order to understand how one can guess that the Weyl matrix ball representation stated in Theorem 1.1 holds. Let $n \in \mathbb{N}_0$. Firstly, we assume a given sequence $(\Gamma_j)_{j=0}^n$ of complex $q \times q$ matrices such that the matrix Γ_0 is nonsingular. Then the matrix S_n defined by (1.1) is nonsingular as well and there is a unique sequence $(\Gamma_j^\sharp)_{j=0}^n$ of complex $q \times q$ matrices such that the block Toeplitz matrix

$$S_n^\sharp := \begin{pmatrix} \Gamma_0^\sharp & 0 & 0 & \dots & 0 \\ \Gamma_1^\sharp & \Gamma_0^\sharp & 0 & \dots & 0 \\ \Gamma_2^\sharp & \Gamma_1^\sharp & \Gamma_0^\sharp & \ddots & \vdots \\ \vdots & \vdots & \vdots & \ddots & 0 \\ \Gamma_n^\sharp & \Gamma_{n-1}^\sharp & \Gamma_{n-2}^\sharp & \dots & \Gamma_0^\sharp \end{pmatrix} \tag{4.12}$$

coincides with S_n^{-1}. This sequence $(\Gamma_j^\sharp)_{j=0}^n$ fulfills the identity $S_k^\sharp = S_k^{-1}$ for each $k \in \mathbb{N}_{0,n}$. Moreover, by setting

$$T_n^\sharp := 2 \operatorname{Re} S_n^\sharp \tag{4.13}$$

we obtain

$$T_n^\sharp = S_n^{-*} T_n S_n^{-1} \quad \text{and} \quad T_n^\sharp = S_n^{-1} T_n S_n^{-*}. \tag{4.14}$$

Therefore, it is readily checked that $(\Gamma_j)_{j=0}^n$ is a $q \times q$ Carathéodory sequence if and only if $(\Gamma_j^\sharp)_{j=0}^n$ is a $q \times q$ Carathéodory sequence. If $(\Gamma_j)_{j=0}^n$ is a $q \times q$ Carathéodory sequence with nonsingular matrix Γ_0, then $(\Gamma_j^\sharp)_{j=0}^n$ is called the *reciprocal* $q \times q$ *Carathéodory sequence corresponding to* $(\Gamma_j)_{j=0}^n$. Now let $(\Gamma_j)_{j=0}^n$ be a nondegenerate $q \times q$ Carathéodory sequence. Then the matrix Γ_0 is necessarily nonsingular. Thus the matrix S_n is nonsingular and the reciprocal $q \times q$ Carathéodory sequence $(\Gamma_j^\sharp)_{j=0}^n$ corresponding to $(\Gamma_j)_{j=0}^n$ is well defined. Furthermore, in view of (4.14), one can see that this $q \times q$ Carathéodory sequence $(\Gamma_j^\sharp)_{j=0}^n$ is also nondegenerate, i.e., the block Toeplitz matrix $T_n^\sharp$ defined by (4.13) and (4.12) is positive Hermitian. Consequently, the matrix polynomials η_n, ζ_n, $\eta_n^\sharp$, and $\zeta_n^\sharp$ given by

$$\eta_n := e_{n,q} T_n^{-1} e_{n,q}(0), \quad \zeta_n := \varepsilon_{n,q}(0) T_n^{-1} \varepsilon_{n,q},$$

$$\eta_n^\sharp := e_{n,q} (T_n^\sharp)^{-1} e_{n,q}(0), \quad \zeta_n^\sharp := \varepsilon_{n,q}(0) (T_n^\sharp)^{-1} \varepsilon_{n,q}$$

are well defined, where $e_{n,q}$ and $\varepsilon_{n,q}$ are the matrix polynomials given by (1.3). Moreover, the matrices T_{n-1}, L_{n+1}, and R_{n+1} are positive Hermitian (see, e.g.,

[FKL1, Lemma 5.1]). In view of [FK1, Theorems 28 and 29 in Part V] one can see that, if $w \in \mathbb{D}$, then the set

$$\Big\{\Omega(w) : \Omega \in \mathcal{C}_q[\mathbb{D}, (\Gamma_j)_{j=0}^n]\Big\}$$

coincides with the matrix ball

$$\mathfrak{K}\Big(\mathbf{M}_{n+1}(w); |w|^{n+1}\sqrt{2\,\mathbf{L}_{n+1}(w)}, \sqrt{2\,\mathbf{R}_{n+1}(w)}\Big),$$

where

$$\mathbf{L}_{n+1}(w) := \Big(\zeta_n^*(w)L_{n+1}\zeta_n(w) - |w|^2\big(\tilde{\eta}_n^{[n]}(w)\big)^* R_{n+1}\tilde{\eta}_n^{[n]}(w)\Big)^{-1},$$

$$\mathbf{R}_{n+1}(w) := \Big(\eta_n(w)R_{n+1}\eta_n^*(w) - |w|^2\tilde{\zeta}_n^{[n]}(w)L_{n+1}\big(\tilde{\zeta}_n^{[n]}(w)\big)^*\Big)^{-1},$$

and

$$\mathbf{M}_{n+1}(w) := \mathbf{L}_{n+1}(w)\Big(\zeta_n^*(w)L_{n+1}\Gamma_0^{-*}\zeta_n^\sharp(w) + |w|^2\big(\tilde{\eta}_n^{[n]}(w)\big)^* R_{n+1}\Gamma_0^{-1}\widetilde{(\eta_n^\sharp)}^{[n]}(w)\Big).$$

Comparing this result for the nondegenerate case with the general one stated in Theorem 1.1, for each $w \in \mathbb{D}$, from [DFK, Corollary 1.5.1 and Theorem 1.5.2] we know that the identity $\mathbf{M}_{n+1}(w) = \mathcal{M}_{n+1}(w)$ holds and that there is a positive real number $\rho_n(w)$ such that both equalities

$$\mathbf{L}_{n+1}(w) = \rho_n(w)\mathcal{L}_{n+1}(w) \quad \text{and} \quad \mathbf{R}_{n+1}(w) = \frac{1}{\rho_n(w)}\mathcal{R}_{n+1}(w)$$

are fulfilled, where $\mathcal{L}_{n+1}(w)$ and $\mathcal{R}_{n+1}(w)$ are the matrices defined in Remark 4.2. We are going to show that actually $\rho_n(w) = 1$ holds for each $w \in \mathbb{D}$. For this reason, first we observe that, in the nondegenerate case we have

$$\mathcal{Y}_n = \widetilde{\mathcal{Y}}_n = \big\{T_{n-1}^{-1}Y_n\big\}, \quad \mathcal{Z}_n = \widetilde{\mathcal{Z}}_n = \big\{Z_nT_{n-1}^{-1}\big\},$$

and that the formulas in (4.14) and the equations

$$2Y_n - S_{n-1}T_{n-1}^{-1}Y_n = S_{n-1}^*T_{n-1}^{-1}, \quad 2Z_n - Z_nT_{n-1}^{-1}S_{n-1} = Z_nT_{n-1}^{-1}S_{n-1}^*,$$

$$T_n^{-1}\begin{pmatrix} I_q \\ 0 \end{pmatrix} = \begin{pmatrix} I_q \\ -T_{n-1}^{-1}Y_n \end{pmatrix}R_{n+1}^{-1},$$

$$\big(T_n^\sharp\big)^{-1}\begin{pmatrix} I_q \\ 0 \end{pmatrix}\Gamma_0^{-*} = S_n\begin{pmatrix} I_q \\ -T_{n-1}^{-1}Y_n \end{pmatrix}R_{n+1}^{-1} = \begin{pmatrix} \Gamma_0 \\ S_{n-1}^*T_{n-1}^{-1}Y_n \end{pmatrix}R_{n+1}^{-1},$$

$$\Big(0,\ I_q\Big)T_n^{-1} = L_{n+1}^{-1}\Big(-Z_nT_{n-1}^{-1},\ I_q\Big),$$

and

$$\Gamma_0^{-*}\Big(0,\ I_q\Big)(T_n^\sharp)^{-1} = L_{n+1}^{-1}\Big(-Z_nT_{n-1}^{-1},\ I_q\Big)S_n = L_{n+1}^{-1}\Big(Z_nT_{n-1}^{-1}S_{n-1}^*,\ \Gamma_0\Big)$$

are valid. Thus we can conclude that the identities

$$\eta_n = b_nR_{n+1}^{-1}, \quad \eta_n^\sharp = a_nR_{n+1}^{-1}\Gamma_0^*, \quad \zeta_n = L_{n+1}^{-1}d_n, \quad \text{and} \quad \zeta_n^\sharp = \Gamma_0^*L_{n+1}^{-1}c_n \quad (4.15)$$

are satisfied (see also [FK3, Section 1]). Let $w \in \mathbb{D}$. In view of the definition of $\mathcal{L}_{n+1}(w)$, (3.5), and (4.15) we get

$$\begin{aligned}\mathcal{L}_{n+1}(w) &= \big(d_n(w)\big)^{-1}\sqrt{L_{n+1}}\,\big(I - |w|^2\Theta_n^*(w)\Theta_n(w)\big)^{-1}\sqrt{L_{n+1}}\,\big(d_n(w)\big)^{-*}\\ &= \Big(d_n^*(w)L_{n+1}^{-1}d_n(w) - |w|^2\big(\tilde{b}_n^{[n]}(w)\big)^*R_{n+1}^{-1}\tilde{b}_n^{[n]}(w)\Big)^{-1}\\ &= \Big(\zeta_n^*(w)L_{n+1}\zeta_n(w) - |w|^2\big(\tilde{\eta}_n^{[n]}(w)\big)^*R_{n+1}\tilde{\eta}_n^{[n]}(w)\Big)^{-1} = \mathbf{L}_{n+1}(w)\end{aligned}$$

and by virtue of the definition of $\mathcal{R}_{n+1}(w)$, (3.4), and (4.15) analogously

$$\mathcal{R}_{n+1}(w) = \mathbf{R}_{n+1}(w).$$

Note that, based on (4.15), one can also check explicitly that $\mathcal{M}_{n+1}(w) = \mathbf{M}_{n+1}(w)$ is satisfied. Moreover, taking into account the definitions of $\mathbf{R}_{n+1}(w)$, $\mathbf{L}_{n+1}(w)$, η_n, and ζ_n as well as (1.3), (1.4), (1.5), and (1.6), an application of the Gohberg-Heinig formula (see [GH, Theorem 1.1]) for the inverse of a nonsingular block Toeplitz matrix shows again that the identities

$$\mathbf{L}_{n+1}(w) = \frac{1}{1-|w|^2}\Big(\varepsilon_{n,q}^*(w)T_n^{-1}\varepsilon_{n,q}(w)\Big)^{-1}$$

and

$$\mathbf{R}_{n+1}(w) = \frac{1}{1-|w|^2}\Big(e_{n,q}(w)T_n^{-1}e_{n,q}^*(w)\Big)^{-1}$$

are fulfilled as well (see also [FKL2, Section 6]).

5. On the Weyl matrix balls associated with reciprocal matrix-valued Carathéodory sequences

It is a well-known fact that if $\Omega \in \mathcal{C}_q(\mathbb{D})$ such that $\det\Omega(0) \neq 0$, then $\det\Omega(w) \neq 0$ for each $w \in \mathbb{D}$ (see, e.g., [FK1, Remark 30 in Part V]). Taking into account this, for a fixed point $w \in \mathbb{D}$, in the following section we give a matrix ball description of the set $\big\{\big(\Omega(w)\big)^{-1} : \Omega \in \mathcal{C}_q[\mathbb{D}, (\Gamma_j)_{j=0}^n]\big\}$ for the case that the matrix Γ_0 of the underlying $q\times q$ Carathéodory sequence $(\Gamma_j)_{j=0}^n$ is nonsingular. In fact, we present a similar description as given by Theorem 1.1 for $\big\{\Omega(w) : \Omega \in \mathcal{C}_q[\mathbb{D}, (\Gamma_j)_{j=0}^n]\big\}$, where the $q \times q$ Schur function $\Theta_n^\diamond$ defined by (3.9) is involved instead of Θ_n.

Let $n \in \mathbb{N}_0$ and let $(\Gamma_j)_{j=0}^n$ be a $q \times q$ Carathéodory sequence such that the matrix Γ_0 is nonsingular. Then the matrix S_n given by (1.1) is nonsingular and hence the reciprocal $q\times q$ Carathéodory sequence $(\Gamma_j^\sharp)_{j=0}^n$ corresponding to $(\Gamma_j)_{j=0}^n$ is well defined. Note that, if $n \geq 1$, if $V_n \in \mathcal{Y}_n$ and $W_n \in \mathcal{Z}_n$, and if we set

$$V_n^\sharp := -S_{n-1}^* V_n \Gamma_0^{-1} \quad \text{and} \quad W_n^\sharp := -\Gamma_0^{-1} W_n S_{n-1}^*, \tag{5.1}$$

then $V_n = -(S_{n-1}^\sharp)^* V_n^\sharp (\Gamma_0^\sharp)^{-1}$ and $W_n = -(\Gamma_0^\sharp)^{-1} W_n^\sharp (S_n^\sharp)^*$ as well as

$$V_n^\sharp \in \mathcal{Y}_n^\sharp \quad \text{and} \quad W_n^\sharp \in \mathcal{Z}_n^\sharp, \tag{5.2}$$

where $\mathcal{Y}_n^\sharp := \{V \in \mathbb{C}^{nq\times q} : T_{n-1}^\sharp V = Y_n^\sharp\}$ and $\mathcal{Z}_n^\sharp := \{W \in \mathbb{C}^{n\times nq} : WT_{n-1}^\sharp = Z_n^\sharp\}$ with

$$Y_n^\sharp := \frac{1}{2}\begin{pmatrix} \Gamma_1^\sharp \\ \Gamma_2^\sharp \\ \vdots \\ \Gamma_n^\sharp \end{pmatrix} \quad \text{and} \quad Z_n^\sharp := \frac{1}{2}\left(\Gamma_n^\sharp, \Gamma_{n-1}^\sharp, \ldots, \Gamma_1^\sharp\right)$$

(cf. [FK3, Lemma 3.2]). Moreover, if the matrix polynomials a_n, b_n, c_n, and d_n are defined as in (1.7), (1.8), (1.9), and (1.8) and if the matrix polynomials $a_n^\sharp$, $b_n^\sharp$, $c_n^\sharp$, and $d_n^\sharp$ are (based on (5.1)) given by

$$a_n^\sharp(z) := \begin{cases} \Gamma_0^\sharp & \text{if } n = 0 \\ \Gamma_0^\sharp + ze_{n-1,q}(z)(S_{n-1}^\sharp)^* V_n^\sharp & \text{if } n \geq 1\,, \end{cases} \tag{5.3}$$

$$b_n^\sharp(z) := \begin{cases} I_q & \text{if } n = 0 \\ I_q - ze_{n-1,q}(z)V_n^\sharp & \text{if } n \geq 1\,, \end{cases} \tag{5.4}$$

$$c_n^\sharp(z) := \begin{cases} \Gamma_0^\sharp & \text{if } n = 0 \\ W_n^\sharp(S_{n-1}^\sharp)^* z\varepsilon_{n-1,q}(z) + \Gamma_0^\sharp & \text{if } n \geq 1\,, \end{cases} \tag{5.5}$$

and

$$d_n^\sharp(z) := \begin{cases} I_q & \text{if } n = 0 \\ -W_n^\sharp z\varepsilon_{n-1,q}(z) + I & \text{if } n \geq 1 \end{cases} \tag{5.6}$$

with some $V_n \in \mathcal{Y}_n$ and $W_n \in \mathcal{Z}_n$ in the case of $n \geq 1$, then it is readily checked that the equalities

$$a_n^\sharp = b_n\Gamma_0^{-1}, \quad b_n^\sharp = a_n\Gamma_0^{-1}, \quad c_n^\sharp = \Gamma_0^{-1}d_n, \quad \text{and} \quad d_n^\sharp = \Gamma_0^{-1}c_n \tag{5.7}$$

hold. Using these identities, we obtain the following interrelation between the construction of $q \times q$ Schur functions stated in Propositions 3.4 and 3.8.

Proposition 5.1. *Let $n \in \mathbb{N}_0$, let $(\Gamma_j)_{j=0}^n$ be a $q\times q$ Carathéodory sequence such that the matrix Γ_0 is nonsingular, and let $(\Gamma_j^\sharp)_{j=0}^n$ be the reciprocal $q \times q$ Carathéodory sequence corresponding to $(\Gamma_j)_{j=0}^n$. If $n \geq 1$, then let $V_n \in \widetilde{\mathcal{Y}}_n$, let $W_n \in \widetilde{\mathcal{Z}}_n$, let $V_n^\sharp := -S_{n-1}^* V_n \Gamma_0^{-1}$, and let $W_n^\sharp := -\Gamma_0^{-1}W_n S_{n-1}^*$. Furthermore, let the matrix polynomials a_n, b_n, c_n, d_n, $a_n^\sharp$, $b_n^\sharp$, $c_n^\sharp$, and $d_n^\sharp$ be defined by* (1.7), (1.8), (1.9), (1.10), (5.3), (5.4), (5.5), *and* (5.6) *and let the matrix-valued functions Θ_n, $\Theta_n^\diamond$, $\Theta_n^\sharp$, and Ξ_n be given by* (3.4), (3.9),

$$\Theta_n^\sharp := \sqrt{R_{n+1}^\sharp}(b_n^\sharp)^{-1}\widetilde{(d_n^\sharp)}^{[n]}\sqrt{L_{n+1}^\sharp}^{+}, \tag{5.8}$$

and

$$\Xi_n := \sqrt{R_{n+1}^\sharp}(a_n^\sharp)^{-1}\widetilde{(c_n^\sharp)}^{[n]}\sqrt{L_{n+1}^\sharp}^{+},$$

where

$$L_1^\sharp := \operatorname{Re}\Gamma_0^\sharp, \quad R_1^\sharp := \operatorname{Re}\Gamma_0^\sharp \tag{5.9}$$

in the case $n = 0$ and

$$L_{n+1}^\sharp := \operatorname{Re}\Gamma_0^\sharp - Z_n^\sharp(T_{n-1}^\sharp)^+(Z_n^\sharp)^*, \quad R_{n+1}^\sharp := \operatorname{Re}\Gamma_0^\sharp - (Y_n^\sharp)^*(T_{n-1}^\sharp)^+ Y_n^\sharp \tag{5.10}$$

if $n \geq 1$. There are unitary $q \times q$ matrices U and V fulfilling

$$\Gamma_0^{-1}\sqrt{L_{n+1}} = \sqrt{L_{n+1}^\sharp}U^* \quad \text{and} \quad \sqrt{R_{n+1}}\Gamma_0^{-1} = V^*\sqrt{R_{n+1}^\sharp} \tag{5.11}$$

and, if U and V are such unitary $q \times q$ matrices, then

$$\Theta_n^\sharp = R_{n+1}^\sharp(R_{n+1}^\sharp)^+ V\,\Theta_n^\diamond U, \quad \Theta_n^\sharp = V\Theta_n^\diamond U L_{n+1}^\sharp(L_{n+1}^\sharp)^+ \tag{5.12}$$

and

$$\Xi_n = R_{n+1}^\sharp(R_{n+1}^\sharp)^+ V\,\Theta_n U, \quad \Xi_n = V\Theta_n U L_{n+1}(L_{n+1}^\sharp)^+. \tag{5.13}$$

In particular, if $\Gamma_0 = I$, then $\Theta_n^\sharp = \Theta_n^\diamond$ and $\Xi_n = \Theta_n$.

Proof. Using [FK3, the formulas by (3.15)] and the polar decomposition of matrices we get that there are unitary $q\times q$ matrices U and V such that the identities stated in (5.11) are satisfied. Lemma 3.2 yields

$$R_{n+1}b_n^{-1}\tilde{d}_n^{[n]} = \tilde{b}_n^{[n]}d_n^{-1}L_{n+1} \quad \text{and} \quad R_{n+1}a_n^{-1}\tilde{c}_n^{[n]} = \tilde{a}_n^{[n]}c_n^{-1}L_{n+1}. \tag{5.14}$$

Because of (5.8), (5.7), (5.14), (5.11), and (3.10) we can conclude

$$\begin{aligned}
\Theta_n^\sharp &= \sqrt{R_{n+1}^\sharp}(b_n^\sharp)^{-1}\widetilde{(d_n^\sharp)}^{[n]}\sqrt{L_{n+1}^\sharp}^+ = V\sqrt{R_{n+1}}a_n^{-1}\tilde{c}_n^{[n]}\Gamma_0^{-*}\sqrt{L_{n+1}^\sharp}^+ \\
&= V\sqrt{R_{n+1}}^+\tilde{a}_n^{[n]}c_n^{-1}L_{n+1}\Gamma_0^{-*}\sqrt{L_{n+1}^\sharp}^+ = V\Theta_n^\diamond\sqrt{L_{n+1}}\Gamma_0^{-*}\sqrt{L_{n+1}^\sharp}^+ \\
&= V\Theta_n^\diamond\big(\sqrt{L_{n+1}^\sharp}U^*\big)^*\sqrt{L_{n+1}^\sharp}^+ = V\Theta_n^\diamond U L_{n+1}^\sharp(L_{n+1}^\sharp)^+.
\end{aligned}$$

Thus the second equation in (5.12) is checked. Similarly, based on (5.7), (5.14), and (5.11) the first equation in (5.12) and the identities in (5.13) can be proved. In the particular case $\Gamma_0 = I_q$ one can choose $U = V = I_q$ (note [FK3, the formulas in (3.15)] and (5.11)) so that $\Theta_n^\sharp = \Theta_n^\diamond$ and $\Xi_n = \Theta_n$ follow from (5.7). □

Theorem 5.2. *Let $n \in \mathbb{N}_0$ and let $(\Gamma_j)_{j=0}^n$ be a $q \times q$ Carathéodory sequence such that the matrix Γ_0 is nonsingular. If $n \geq 1$, then let $V_n \in \widetilde{\mathcal{Y}}_n$ and let $W_n \in \widetilde{\mathcal{Z}}_n$. Let the matrix polynomials a_n, b_n, c_n, and d_n be defined by* (1.7), (1.8), (1.9), *and* (1.10). *Let the matrices L_{n+1}, R_{n+1}, $L_{n+1}^\sharp$, and $R_{n+1}^\sharp$ be given by* (1.4), (1.6), (5.9), *and* (5.10), *where $(\Gamma_j^\sharp)_{j=0}^n$ stands for the reciprocal $q\times q$ Carathéodory sequence corresponding to $(\Gamma_j)_{j=0}^n$. If $\Omega \in \mathcal{C}_q[\mathbb{D}, (\Gamma_j)_{j=0}^n]$, then the function $\det\Omega$ does not vanish in $\mathbb{D}$ and*

$$\Big\{\big(\Omega(w)\big)^{-1} : \Omega \in \mathcal{C}_q[\mathbb{D}, (\Gamma_j)_{j=0}^n]\Big\} = \mathfrak{K}\Big(\mathcal{M}'_{n+1}(w); 2|w|^{n+1}\mathcal{A}'_{n+1}(w), \mathcal{B}'_{n+1}(w)\Big)$$

for each $w \in \mathbb{D}$ with

$$\mathcal{M}'_{n+1}(w) := \Big(c_n(w) - |w|^2 G_n^\blacklozenge(w)\tilde{a}_n^{[n]}(w)\Big)^{-1}\Big(d_n(w) + |w|^2 G_n^\blacklozenge(w)\tilde{b}_n^{[n]}(w)\Big), \tag{5.15}$$

$$\mathcal{A}'_{n+1}(w) := \big(c_n(w)\big)^{-1}\sqrt{L_{n+1}}\sqrt{I_q - |w|^2\big(\Theta_n^\diamond(w)\big)^*\Theta_n^\diamond(w)}^{-1}, \tag{5.16}$$

and

$$\mathcal{B}'_{n+1}(w) := \sqrt{I_q - |w|^2\Theta_n^\diamond(w)\big(\Theta_n^\diamond(w)\big)^*}^{-1}\sqrt{R_{n+1}}\,\big(a_n(w)\big)^{-1}, \tag{5.17}$$

where $G_n^{\blacklozenge}(w) := \sqrt{L_{n+1}}\big(\Theta_n^{\diamond}(w)\big)^* \sqrt{R_{n+1}}^{+} \Gamma_0^* R_{n+1}^{\sharp} (R_{n+1}^{\sharp})^{+} \Gamma_0^{-*}$ *and where* $\Theta_n^{\diamond}$ *stands for the rational matrix-valued function defined by* (3.9)*. Moreover, for each* $w \in \mathbb{D}$*, the matrix* $\mathcal{M}'_{n+1}(w)$ *can be represented via*

$$\mathcal{M}'_{n+1}(w) = \Big(b_n(w) + |w|^2 \tilde{d}_n^{[n]}(w) F_n^{\blacklozenge}(w)\Big)\Big(a_n(w) - |w|^2 \tilde{c}_n^{[n]}(w) F_n^{\blacklozenge}(w)\Big)^{-1},$$

where $F_n^{\blacklozenge}(w) := \Gamma_0^{-*} L_{n+1}^{\sharp} (L_{n+1}^{\sharp})^{+} \Gamma_0^* \sqrt{L_{n+1}}^{+} \big(\Theta_n^{\diamond}(w)\big)^* \sqrt{R_{n+1}}$.

Proof. For each $\Omega \in \mathcal{C}_q[\mathbb{D}, (\Gamma_j)_{j=0}^n]$ we have $\det \Omega(0) = \det \Gamma_0 \neq 0$. Consequently, the inequality $\det \Omega(w) \neq 0$ holds for each $w \in \mathbb{D}$ and

$$\Big\{\Omega^{-1} : \Omega \in \mathcal{C}_q[\mathbb{D}, (\Gamma_j)_{j=0}^n]\Big\} = \mathcal{C}_q[\mathbb{D}, (\Gamma_j^{\sharp})_{j=0}^n] \tag{5.18}$$

is valid (see, e.g., [FK1, Remark 30 in Part V]). Let the matrix polynomials $a_n^{\sharp}$, $b_n^{\sharp}$, $c_n^{\sharp}$, and $d_n^{\sharp}$ be defined by (5.3), (5.4), (5.5), and (5.6) based on (5.1). Let $w \in \mathbb{D}$. Due to (5.18), (5.2), (5.7), and Remark 3.7 an application of Theorem 1.1 yields

$$\Big\{\big(\Omega(w)\big)^{-1} : \Omega \in \mathcal{C}_q[\mathbb{D}, (\Gamma_j)_{j=0}^n]\Big\} = \mathfrak{K}\Big(\mathcal{M}_{n+1}^{\sharp}(w); 2|w|^{n+1} \mathcal{A}_{n+1}^{\sharp}(w), \mathcal{B}_{n+1}^{\sharp}(w)\Big)$$

with the settings

$$\mathcal{M}_{n+1}^{\sharp}(w) := \Big(d_n^{\sharp}(w) - |w|^2 G_n^{\bullet}(w) \widetilde{(b_n^{\sharp})}^{[n]}(w)\Big)^{-1}\Big(c_n^{\sharp}(w) + |w|^2 G_n^{\bullet}(w) \widetilde{(a_n^{\sharp})}^{[n]}(w)\Big),$$

$$\mathcal{A}_{n+1}^{\sharp}(w) := \big(d_n^{\sharp}(w)\big)^{-1} \sqrt{L_{n+1}^{\sharp}} \sqrt{I - |w|^2 \big(\Theta_n^{\sharp}(w)\big)^* \Theta_n^{\sharp}(w)}^{-1},$$

and

$$\mathcal{B}_{n+1}^{\sharp}(w) := \sqrt{I - |w|^2 \Theta_n^{\sharp}(w) \big(\Theta_n^{\sharp}(w)\big)^*}^{-1} \sqrt{R_{n+1}^{\sharp}}\, \big(b_n^{\sharp}(w)\big)^{-1},$$

where $G_n^{\bullet}(w) := \sqrt{L_{n+1}^{\sharp}} \big(\Theta_n^{\sharp}(w)\big)^* \sqrt{R_{n+1}^{\sharp}}^{+}$ and where $\Theta_n^{\sharp}$ is the matrix-valued function defined by (5.8). From Theorem 1.1 we also see that

$$\mathcal{M}_{n+1}^{\sharp}(w) = \Big(a_n^{\sharp}(w) + |w|^2 \widetilde{(c_n^{\sharp})}^{[n]}(w) F_n^{\bullet}(w)\Big)\Big(b_n^{\sharp}(w) - |w|^2 \widetilde{(d_n^{\sharp})}^{[n]}(w) F_n^{\bullet}(w)\Big)^{-1} \tag{5.19}$$

holds, where $F_n^{\bullet}(w) := \sqrt{L_{n+1}^{\sharp}}^{+} \big(\Theta_n^{\sharp}(w)\big)^* \sqrt{R_{n+1}^{\sharp}}$. According to Proposition 5.1 there are unitary $q \times q$ matrices U and V such that (5.11), (5.12), and (5.13) are valid. Using (5.11), (5.12), and (3.9) we obtain

$$\begin{aligned}
G_n^{\bullet}(w) &= \Gamma_0^{-1} \sqrt{L_{n+1}} U \Big(R_{n+1}^{\sharp} (R_{n+1}^{\sharp})^{+} V \Theta_n^{\diamond}(w) U\Big)^* \sqrt{R_{n+1}^{\sharp}}^{+} \\
&= \Gamma_0^{-1} \sqrt{L_{n+1}} \big(\Theta_n^{\diamond}(w)\big)^* V^* \sqrt{R_{n+1}^{\sharp}}^{+} \\
&= \Gamma_0^{-1} \sqrt{L_{n+1}} \big(\Theta_n^{\diamond}(w)\big)^* \sqrt{R_{n+1}}^{+} \sqrt{R_{n+1}} V^* \sqrt{R_{n+1}}^{+} \\
&= \Gamma_0^{-1} \sqrt{L_{n+1}} \big(\Theta_n^{\diamond}(w)\big)^* \sqrt{R_{n+1}}^{+} \Gamma_0^* \sqrt{R_{n+1}^{\sharp}} \sqrt{R_{n+1}^{\sharp}}^{+} \\
&= \Gamma_0^{-1} \sqrt{L_{n+1}} \big(\Theta_n^{\diamond}(w)\big)^* \sqrt{R_{n+1}}^{+} \Gamma_0^* R_{n+1}^{\sharp} (R_{n+1}^{\sharp})^{+}
\end{aligned} \tag{5.20}$$

and analogously

$$F_n^{\bullet}(w) = L_{n+1}^{\sharp}(L_{n+1}^{\sharp})^{+}\Gamma_0^{*}\sqrt{L_{n+1}}^{+}\big(\Theta_n^{\diamond}(w)\big)^{*}\sqrt{R_{n+1}}\,\Gamma_0^{-1}. \tag{5.21}$$

In addition, from (5.12), (3.10), and (5.11) we also get

$$\begin{aligned}
&\big(\Theta_n^{\sharp}(w)\big)^{*}\Theta_n^{\sharp}(w)\\
&\quad= U^{*}\big(\Theta_n^{\diamond}(w)\big)^{*}V^{*}(R_{n+1}^{\sharp})^{+}R_{n+1}^{\sharp}V\Theta_n^{\diamond}(w)U\\
&\quad= U^{*}\big(\Theta_n^{\diamond}(w)\big)^{*}\sqrt{R_{n+1}}^{+}\sqrt{R_{n+1}}V^{*}(R_{n+1}^{\sharp})^{+}R_{n+1}^{\sharp}\big(\sqrt{R_{n+1}}V^{*}\big)^{*}\sqrt{R_{n+1}}^{+}\Theta_n^{\diamond}(w)U\\
&\quad= U^{*}\big(\Theta_n^{\diamond}(w)\big)^{*}\sqrt{R_{n+1}}^{+}\Gamma_0^{*}\sqrt{R_{n+1}^{\sharp}}\sqrt{R_{n+1}^{\sharp}}\Gamma_0\sqrt{R_{n+1}}^{+}\Theta_n^{\diamond}(w)U\\
&\quad= U^{*}\big(\Theta_n^{\diamond}(w)\big)^{*}\sqrt{R_{n+1}}^{+}R_{n+1}\sqrt{R_{n+1}}^{+}\Theta_n^{\diamond}(w)U \;=\; U^{*}\big(\Theta_n^{\diamond}(w)\big)^{*}\Theta_n^{\diamond}(w)U
\end{aligned} \tag{5.22}$$

and analogously

$$\Theta_n^{\sharp}(w)\big(\Theta_n^{\sharp}(w)\big)^{*} = V\Theta_n^{\diamond}(w)\big(\Theta_n^{\diamond}(w)\big)^{*}V^{*}. \tag{5.23}$$

Because of (5.7) and (5.20) we have

$$d_n^{\sharp}(w) - |w|^2G_n^{\bullet}(w)\widetilde{(b_n^{\sharp})}^{[n]}(w) = \Gamma_0^{-1}\left(c_n(w) - |w|^2G_n^{\blacklozenge}(w)\tilde{a}_n^{[n]}(w)\right)$$

and

$$c_n^{\sharp}(w) + |w|^2G_n^{\bullet}(w)\widetilde{(a_n^{\sharp})}^{[n]}(w) = \Gamma_0^{-1}\left(d_n(w) + |w|^2G_n^{\blacklozenge}(w)\tilde{b}_n^{[n]}(w)\right).$$

Thus we see that $\mathcal{M}'_{n+1}(w)$ is a well-defined matrix and that the identity

$$\mathcal{M}'_{n+1}(w) = \mathcal{M}_{n+1}^{\sharp}(w) \tag{5.24}$$

holds. Furthermore, from (5.19), (5.7), and (5.21) it follows the second representation of $\mathcal{M}'_{n+1}(w)$. By virtue of (5.7), (5.11), and (5.22) we can conclude

$$\begin{aligned}
\mathcal{A}_{n+1}^{\sharp}(w) \;&=\; \big(c_n(w)\big)^{-1}\sqrt{L_{n+1}}U\sqrt{I - |w|^2U^{*}\big(\Theta_n^{\diamond}(w)\big)^{*}\Theta_n^{\diamond}(w)U}^{-1}\\
&=\; \big(c_n(w)\big)^{-1}\sqrt{L_{n+1}}\sqrt{I - |w|^2\big(\Theta_n^{\diamond}(w)\big)^{*}\Theta_n^{\diamond}(w)}^{-1}U \;=\; \mathcal{A}'_{n+1}(w)U
\end{aligned}$$

and because of (5.23) analogously

$$\mathcal{B}_{n+1}^{\sharp}(w) = V\mathcal{B}'_{n+1}(w).$$

Application of [Sm, Theorem 1.3] completes the proof. □

Remark 5.3. Let the assumptions of Theorem 5.2 be fulfilled. Then Theorem 5.2 and [Sm, Theorem 1.3] (see also [DFK, Theorem 1.5.2]) imply the identity

$$\left\{\big(\Omega(w)\big)^{-1} \colon \Omega \in \mathcal{C}_q[\mathbb{D}, (\Gamma_j)_{j=0}^{n}]\right\} = \mathfrak{K}\left(\mathcal{M}'_{n+1}(w); |w|^{n+1}\sqrt{2\mathcal{L}'_{n+1}(w)}, \sqrt{2\mathcal{R}'_{n+1}(w)}\right)$$

for each $w \in \mathbb{D}$, where $\mathcal{M}'_{n+1}(w)$, $\mathcal{L}'_{n+1}(w)$, and $\mathcal{R}'_{n+1}(w)$ are given by (5.15),

$$\mathcal{L}'_{n+1}(w) := \big(c_n(w)\big)^{-1}\sqrt{L_{n+1}}\left(I_q - |w|^2\big(\Theta_n^{\diamond}(w)\big)^{*}\Theta_n^{\diamond}(w)\right)^{-1}\sqrt{L_{n+1}}\,\big(c_n(w)\big)^{-*},$$

and

$$\mathcal{R}'_{n+1}(w) := \big(a_n(w)\big)^{-*}\sqrt{R_{n+1}}\,\Big(I_q - |w|^2\Theta_n^{\diamond}(w)\big(\Theta_n^{\diamond}(w)\big)^*\Big)^{-1}\sqrt{R_{n+1}}\,\big(a_n(w)\big)^{-1}.$$

Remark 5.4. Let the assumptions of Theorem 5.2 be fulfilled and let $\Gamma_0 = I_q$. Then (5.11) shows that $L_{n+1}^{\sharp} = L_{n+1}$ and $R_{n+1}^{\sharp} = R_{n+1}$ hold. Thus it is readily checked that, for each $w \in \mathbb{D}$, the matrices $G_n^{\blacklozenge}(w)$ and $F_n^{\blacklozenge}(w)$, which lead to representations of $\mathcal{M}'_{n+1}(w)$ according to Theorem 5.2, have the simpler form

$$G_n^{\blacklozenge}(w) = \sqrt{L_{n+1}}\big(\Theta_n^{\diamond}(w)\big)^*\sqrt{R_{n+1}}^{+} \quad \text{and} \quad F_n^{\blacklozenge}(w) = \sqrt{L_{n+1}}^{+}\big(\Theta_n^{\diamond}(w)\big)^*\sqrt{R_{n+1}}$$

in that case. (These formulas are already satisfied if Γ_0 is a unitary $q \times q$ matrix.)

Remark 5.5. Let the assumptions of Theorem 5.2 be fulfilled. In view of [FK1, Remark 2 in Part I], (5.11), and Theorem 5.2 one can see that

$$\operatorname{rank}\mathcal{A}'_{n+1}(w) = \operatorname{rank}L_{n+1} = \operatorname{rank}R_{n+1} = \operatorname{rank}\mathcal{B}'_{n+1}(w)$$

for every choice of w in $\mathbb{D}$.

Let the assumptions of Theorem 5.2 be fulfilled. Based on Remarks 4.4, 5.5, and 4.3 one can immediately get further necessary and sufficient conditions for the fact that the solution set $\mathcal{C}_q[\mathbb{D}, (\Gamma_j)_{j=0}^n]$ of the matricial Carathéodory problem contains exactly one element. Moreover, application of Remarks 4.4, 5.5, and 4.5 yields further necessary and and sufficient conditions for the nondegeneracy of a given $q \times q$ Carathéodory sequence $(\Gamma_j)_{j=0}^n$. We omit the details.

Remark 5.6. Let the assumptions of Theorem 5.2 be fulfilled. From Remark 4.6 and (5.7) one can see that the identity $\det a_n(w) = \det c_n(w)$ holds for each $w \in \mathbb{D}$. Thus taking into account [FK1, Remark 2 in Part I], [DFK, Lemma 1.1.8], (5.16), and (5.17) for each $w \in \mathbb{D}$ it follows

$$\det\mathcal{A}'_{n+1}(w) = \det\mathcal{B}'_{n+1}(w).$$

6. Further observations on the matrix-valued functions Θ_n and $\Theta_n^{\diamond}$

Theorem 1.1 shows that the rational matrix-valued function Θ_n defined by (3.4) plays a key role in the description of the Weyl matrix balls of the solutions of the matricial Carathéodory problem. In this section, we check that this function Θ_n (and hence the parameters of the Weyl matrix balls defined by Theorem 1.1) does not depend on the concrete choice of the underlying matrix V_n in the set $\widetilde{\mathcal{Y}}_n$ and the underlying matrix W_n in the set $\widetilde{\mathcal{Z}}_n$. Furthermore, we will state a possibility to construct recursively the function Θ_n. Since there are matrix polynomials a_n, b_n, c_n, and d_n of the types which are considered in Theorem 1.1 which can be constructed recursively as well (see [FK3, Proposition 4.4, Remark 4.5, and Lemma 4.6]) one gets a particular possibility to calculate the parameters of the Weyl matrix balls of the solutions of matricial Carathéodory problem. Moreover, we

present analogous results corresponding to the rational matrix-valued function $\Theta_n^\diamond$ defined by (3.9), which appears in the Weyl matrix balls studied in Theorem 5.2.

At first, we show that the functions Θ_n and $\Theta_n^\diamond$ are independent of the concrete choice of V_n in $\widetilde{\mathcal{Y}}_n$ and W_n in $\widetilde{\mathcal{Z}}_n$. In other words, the functions Θ_n and $\Theta_n^\diamond$ depend only on the given $q \times q$ Carathéodory sequence $(\Gamma_j)_{j=0}^n$.

Proposition 6.1. *Let $n \in \mathbb{N}_0$ and let $(\Gamma_j)_{j=0}^n$ be a $q \times q$ Carathéodory sequence. If $n \geq 1$, then let $V_n, \mathbf{V}_n \in \widetilde{\mathcal{Y}}_n$ and let $W_n, \mathbf{W}_n \in \widetilde{\mathcal{Z}}_n$. Let a_n, b_n, c_n, and d_n be the matrix polynomials given by* (1.7), (1.8), (1.9), *and* (1.10), *and let $\mathbf{a}_n$, $\mathbf{b}_n$, $\mathbf{c}_n$, and $\mathbf{d}_n$ be the matrix polynomials which are defined analogously to a_n, b_n, c_n, and d_n using (if $n \geq 1$) the matrices $\mathbf{V}_n$ and $\mathbf{W}_n$ instead of V_n and W_n, respectively. Furthermore, let the matrices L_{n+1} and R_{n+1} be given by* (1.4) *and* (1.6). *Then:*

(a) *The matrix-valued functions*

$$\Theta_n := \sqrt{R_{n+1}}\, b_n^{-1} \tilde{d}_n^{[n]} \sqrt{L_{n+1}}^{+} \quad \text{and} \quad \mathbf{\Theta}_n := \sqrt{R_{n+1}}\, \mathbf{b}_n^{-1} \tilde{\mathbf{d}}_n^{[n]} \sqrt{L_{n+1}}^{+}$$

coincide.

(b) *If the matrix Γ_0 is nonsingular, then the matrix-valued functions*

$$\Theta_n^\diamond := \sqrt{R_{n+1}}\, a_n^{-1} \tilde{c}_n^{[n]} \sqrt{L_{n+1}}^{+} \quad \text{and} \quad \mathbf{\Theta}_n^\diamond := \sqrt{R_{n+1}}\, \mathbf{a}_n^{-1} \tilde{\mathbf{c}}_n^{[n]} \sqrt{L_{n+1}}^{+}$$

coincide.

Proof. (a) Let $w \in \mathbb{D} \setminus \{0\}$. Using Theorem 1.1 in combination with [DFK, Corollary 1.5.1] we can conclude

$$\begin{aligned} &\Big(a_n(w) + |w|^2 \tilde{c}_n^{[n]}(w) F_n^\blacksquare(w)\Big)\Big(b_n(w) - |w|^2 \tilde{d}_n^{[n]}(w) F_n^\blacksquare(w)\Big)^{-1} \\ &\quad = \Big(\mathbf{a}_n(w) + |w|^2 \tilde{\mathbf{c}}_n^{[n]}(w) \mathbf{F}_n^\blacksquare(w)\Big)\Big(\mathbf{b}_n(w) - |w|^2 \tilde{\mathbf{d}}_n^{[n]}(w) \mathbf{F}_n^\blacksquare(w)\Big)^{-1}, \quad (6.1) \end{aligned}$$

where $F_n^\blacksquare(w) := \sqrt{L_{n+1}}^{+} \Theta_n^*(w) \sqrt{R_{n+1}}$ and $\mathbf{F}_n^\blacksquare(w) := \sqrt{L_{n+1}}^{+} \mathbf{\Theta}_n^*(w) \sqrt{R_{n+1}}$. According to Proposition 3.4 the matrix $-\overline{w}\mathbf{\Theta}_n^*(w)$ is strictly contractive. Thus [FKL1, Lemma 3.1] yields

$$\begin{aligned} &\Big(\mathbf{a}_n(w) + |w|^2 \tilde{\mathbf{c}}_n^{[n]}(w) \mathbf{F}_n^\blacksquare(w)\Big)\Big(\mathbf{b}_n(w) - |w|^2 \tilde{\mathbf{d}}_n^{[n]}(w) \mathbf{F}_n^\blacksquare(w)\Big)^{-1} \\ &\quad = \Big(a_n(w) + |w|^2 \tilde{c}_n^{[n]}(w) \mathbf{F}_n^\blacksquare(w)\Big)\Big(b_n(w) - |w|^2 \tilde{d}_n^{[n]}(w) \mathbf{F}_n^\blacksquare(w)\Big)^{-1}. \end{aligned}$$

Combining this with (6.1) we obtain

$$\begin{aligned} &\Big(a_n(w) + |w|^2 \tilde{c}_n^{[n]}(w) F_n^\blacksquare(w)\Big)\Big(b_n(w) - |w|^2 \tilde{d}_n^{[n]}(w) F_n^\blacksquare(w)\Big)^{-1} \\ &\quad = \Big(a_n(w) + |w|^2 \tilde{c}_n^{[n]}(w) \mathbf{F}_n^\blacksquare(w)\Big)\Big(b_n(w) - |w|^2 \tilde{d}_n^{[n]}(w) \mathbf{F}_n^\blacksquare(w)\Big)^{-1}. \end{aligned}$$

Since we see from Proposition 3.4 that the matrices $-\overline{w}\Theta_n^*(w)$ and $-\overline{w}\mathbf{\Theta}_n^*(w)$ are both contractive, Proposition 2.2 provides us

$$-\overline{w} L_{n+1} L_{n+1}^{+} \Theta_n^*(w) R_{n+1} R_{n+1}^{+} = -\overline{w} L_{n+1} L_{n+1}^{+} \mathbf{\Theta}_n^*(w) R_{n+1} R_{n+1}^{+}.$$

Therefore, in view of $w \neq 0$ and the definitions of Θ_n and $\mathbf{\Theta}_n$, we get the identity $\Theta_n^*(w) = \mathbf{\Theta}_n^*(w)$, i.e., $\Theta_n(w) = \mathbf{\Theta}_n(w)$. Since Θ_n and $\mathbf{\Theta}_n$ are rational matrix-valued functions, we can finally conclude $\Theta_n = \mathbf{\Theta}_n$.
(b) Based on the reciprocal $q \times q$ Carathéodory sequence $(\Gamma_j^\sharp)_{j=0}^n$ to $(\Gamma_j)_{j=0}^n$, the assertion of part (b) follows from (a) in combination with Proposition 5.1. □

Corollary 6.2. *Let the assumptions of* Proposition 6.1 *be fulfilled. Then:*
(a) *The identities* $d_n^{-1}\sqrt{L_{n+1}} = \mathbf{d}_n^{-1}\sqrt{L_{n+1}}$ *and* $\sqrt{R_{n+1}}b_n^{-1} = \sqrt{R_{n+1}}\mathbf{b}_n^{-1}$ *hold.*
(b) *If the matrix* Γ_0 *is nonsingular, then the equalities* $c_n^{-1}\sqrt{L_{n+1}} = \mathbf{c}_n^{-1}\sqrt{L_{n+1}}$ *and* $\sqrt{R_{n+1}}a_n^{-1} = \sqrt{R_{n+1}}\mathbf{a}_n^{-1}$ *are satisfied.*

Proof. First we observe that one can easily see that each of the sets $\mathcal{N}_{b_n}$, $\mathcal{N}_{\tilde{b}_n^{[n]}}$, $\mathcal{N}_{d_n}$, $\mathcal{N}_{\tilde{d}_n^{[n]}}$, $\mathcal{N}_{\mathbf{b}_n}$, $\mathcal{N}_{\tilde{\mathbf{b}}_n^{[n]}}$, $\mathcal{N}_{\mathbf{d}_n}$, and $\mathcal{N}_{\tilde{\mathbf{d}}_n^{[n]}}$ consists of at most nq elements. Taking into account Lemma 3.2 we obtain

$$R_{n+1}b_n^{-1}\tilde{d}_n^{[n]} = \tilde{b}_n^{[n]}d_n^{-1}L_{n+1} \quad \text{and} \quad R_{n+1}b_n^{-1}\tilde{\mathbf{d}}_n^{[n]} = \tilde{b}_n^{[n]}\mathbf{d}_n^{-1}L_{n+1}. \tag{6.2}$$

Moreover, according to part (a) of Proposition 6.1, the rational matrix functions Θ_n and $\sqrt{R_{n+1}}b_n^{-1}\tilde{\mathbf{d}}_n^{[n]}\sqrt{L_{n+1}}^+$ coincide. Thus from (6.2) it follows

$$\begin{aligned}\tilde{b}_n^{[n]}d_n^{-1}\sqrt{L_{n+1}} &= \tilde{b}_n^{[n]}d_n^{-1}L_{n+1}\sqrt{L_{n+1}}^+ = R_{n+1}b_n^{-1}\tilde{d}_n^{[n]}\sqrt{L_{n+1}}^+ = \sqrt{R_{n+1}}\Theta_n \\ &= R_{n+1}b_n^{-1}\tilde{\mathbf{d}}_n^{[n]}\sqrt{L_{n+1}}^+ = \tilde{b}_n^{[n]}\mathbf{d}_n^{-1}L_{n+1}\sqrt{L_{n+1}}^+ = \tilde{b}_n^{[n]}\mathbf{d}_n^{-1}\sqrt{L_{n+1}}\end{aligned}$$

and similarly $\sqrt{R_{n+1}}b_n^{-1}\tilde{d}_n^{[n]} = \sqrt{R_{n+1}}\mathbf{b}_n^{-1}\tilde{d}_n^{[n]}$. Using continuity arguments we get

$$d_n^{-1}\sqrt{L_{n+1}} = \mathbf{d}_n^{-1}\sqrt{L_{n+1}} \quad \text{and} \quad \sqrt{R_{n+1}}b_n^{-1} = \sqrt{R_{n+1}}\mathbf{b}_n^{-1}.$$

Hence part (a) is proved. Applying Lemma 3.2, part (b) of Proposition 6.1, and Proposition 3.8, part (b) can be verified analogously. □

Remark 6.3. Let the assumptions of Proposition 6.1 be fulfilled. Further, suppose that the matrix Γ_0 is nonsingular. Because of Corollary 6.2 one can immediately conclude that the matrix-valued function $\Theta_n^{\triangleright}$ given by (3.17) coincides with

$$\mathbf{\Theta}_n^{\triangleright} := \sqrt{R_{n+1}}\mathbf{a}_n^{-1}\mathbf{d}_n^{-1}\sqrt{L_{n+1}}.$$

Based on part (a) of Proposition 6.1 we get that, for each $w \in \mathbb{D}$, the semiradii of the matrix ball described in Theorem 1.1 is independent of the concrete choice of the matrix V_n in $\widetilde{\mathcal{Y}}_n$ and the matrix W_n in $\widetilde{\mathcal{Z}}_n$.

Remark 6.4. Let the assumptions of Proposition 6.1 be fulfilled. From Proposition 6.1 and Corollary 6.2 one can see that the identities $\mathcal{A}_{n+1}(w) = \mathbf{A}_{n+1}(w)$ and $\mathcal{B}_{n+1}(w) = \mathbf{B}_{n+1}(w)$ hold for each $w \in \mathbb{D}$, where $\mathcal{A}_{n+1}(w)$ and $\mathcal{B}_{n+1}(w)$ are defined by (1.12) and (1.13) and where similarly

$$\mathbf{A}_{n+1}(w) := \big(\mathbf{d}_n(w)\big)^{-1}\sqrt{L_{n+1}}\sqrt{I - |w|^2\mathbf{\Theta}_n^*(w)\mathbf{\Theta}_n(w)}^{-1}$$

and

$$\mathbf{B}_{n+1}(w) := \sqrt{I - |w|^2\mathbf{\Theta}_n(w)\mathbf{\Theta}_n^*(w)}^{-1}\sqrt{R_{n+1}}\,\big(\mathbf{b}_n(w)\big)^{-1}.$$

The semi-radii of the matrix ball described in Theorem 5.2 do not depend on the concrete choice of V_n in $\widetilde{\mathcal{Y}}_n$ and W_n in $\widetilde{\mathcal{Z}}_n$ as well.

Remark 6.5. Let the assumptions of Proposition 6.1 be fulfilled. Further, suppose that the matrix Γ_0 is nonsingular. From Proposition 6.1 and Corollary 6.2 it follows that $\mathcal{A}'_{n+1}(w) = \mathbf{A}'_{n+1}(w)$ and $\mathcal{B}'_{n+1}(w) = \mathbf{B}'_{n+1}(w)$ hold for each $w \in \mathbb{D}$, where $\mathcal{A}'_{n+1}(w)$ and $\mathcal{B}'_{n+1}(w)$ are defined by (5.16) and (5.17) and where similarly

$$\mathbf{A}'_{n+1}(w) := \big(\mathbf{c}_n(w)\big)^{-1} \sqrt{L_{n+1}} \sqrt{I - |w|^2 \big(\mathbf{\Theta}^{\diamond}_n(w)\big)^* \mathbf{\Theta}^{\diamond}_n(w)}^{-1}$$

and

$$\mathbf{B}'_{n+1}(w) := \sqrt{I - |w|^2 \mathbf{\Theta}^{\diamond}_n(w) \big(\mathbf{\Theta}^{\diamond}_n(w)\big)^*}^{-1} \sqrt{R_{n+1}}\, \big(\mathbf{a}_n(w)\big)^{-1}.$$

Now we are going to verify recurrence relations for the rational matrix-valued functions Θ_n and $\Theta^{\diamond}_n$ defined by (3.4) and (3.9), respectively. Before, we state some technical results which are useful in view of the proof of these formulas.

Remark 6.6. Let $A \in \mathbb{C}^{p\times q}$ and let $B \in \mathbb{C}^{q\times p}$. Then $\det(I - BA) = \det(I - AB)$ holds (see, e.g., [DFK, Lemma 1.1.8]). Moreover, if $\det(I - AB) \neq 0$, then the identity $(I - AB)^{-1}A = A(I - BA)^{-1}$ is fulfilled.

Remark 6.7. Let $n \in \mathbb{N}$ and let $(\Gamma_j)_{j=0}^n$ be a $q \times q$ Carathéodory sequence. Let $\mathbf{a}_0$, $\mathbf{b}_0$, $\mathbf{c}_0$, and $\mathbf{d}_0$ be the constant matrix-valued functions given, for each $z \in \mathbb{C}$, by

$$\mathbf{a}_0(z) := \Gamma_0, \quad \mathbf{b}_0(z) := I_q, \quad \mathbf{c}_0(z) := \Gamma_0, \quad \mathbf{d}_0(z) := I_q,$$

and for each $m \in \mathbb{N}_{0,n-1}$ let the matrix polynomials $\mathbf{a}_{m+1}$, $\mathbf{b}_{m+1}$, $\mathbf{c}_{m+1}$, and $\mathbf{d}_{m+1}$ be recursively defined, for each $z \in \mathbb{C}$, by

$$\mathbf{a}_{m+1}(z) := \mathbf{a}_m(z) + z\tilde{\mathbf{c}}^{[m]}_m(z)t_{m+1}, \quad \mathbf{b}_{m+1}(z) := \mathbf{b}_m(z) - z\tilde{\mathbf{d}}^{[m]}_m(z)t_{m+1},$$
$$\mathbf{c}_{m+1}(z) := \mathbf{c}_m(z) + zu_{m+1}\tilde{\mathbf{a}}^{[m]}_m(z), \quad \mathbf{d}_{m+1}(z) := \mathbf{d}_m(z) - zu_{m+1}\tilde{\mathbf{b}}^{[m]}_m(z),$$

where

$$t_{m+1} := L^+_{m+1}\big(\tfrac{1}{2}\Gamma_{m+1} - M_{m+1}\big) \quad \text{and} \quad u_{m+1} := \big(\tfrac{1}{2}\Gamma_{m+1} - M_{m+1}\big)R^+_{m+1}. \tag{6.3}$$

In view of [FK3, Remark 4.5, Proposition 4.4, and Lemma 4.6] one can see that, for each $m \in \mathbb{N}_{1,n}$, there are matrices $\mathbf{V}_m \in \widetilde{\mathcal{Y}}_m$ and $\mathbf{W}_m \in \widetilde{\mathcal{Z}}_m$ such that

$$\mathbf{a}_m(z) = \Gamma_0 + ze_{m-1,q}(z)S^*_{m-1}\mathbf{V}_m, \quad \mathbf{b}_m(z) = I_q - ze_{m-1,q}(z)\mathbf{V}_m,$$
$$\mathbf{c}_m(z) = \mathbf{W}_m S^*_{m-1} z\varepsilon_{m-1,q}(z) + \Gamma_0, \quad \mathbf{d}_m(z) = -\mathbf{W}_m z\varepsilon_{m-1,q}(z) + I_q$$

hold. In particular, for each $m \in \mathbb{N}_{0,n}$ and each $w \in \mathbb{D}$, the complex $q \times q$ matrices $\mathbf{b}_m(w)$ and $\mathbf{d}_m(w)$ are nonsingular.

If $(\Gamma_j)_{j=0}^n$ is a $q \times q$ Carathéodory sequence, then we will use the notation

$$M_{k+1} := \begin{cases} 0_{q\times q} & \text{if } k = 0 \\ Z_k T^+_{k-1} Y_k & \text{if } k \in \mathbb{N}_{1,n}\,, \end{cases} \tag{6.4}$$

where for each $k \in \mathbb{N}_{1,n}$ the matrices Z_k and Y_k are given by (1.5). Furthermore, let Θ_n be the matrix-valued function defined by (3.4).

Theorem 6.8. *Let $n \in \mathbb{N}_0$ and let $(\Gamma_j)_{j=0}^n$ be a $q \times q$ Carathéodory sequence. Then:*

(a) *Θ_0 is the constant function with value $(\operatorname{Re}\Gamma_0)(\operatorname{Re}\Gamma_0)^+$.*

(b) *If $n \geq 1$, then for each $m \in \mathbb{N}_{1,n}$ the matrix*

$$K_m := \sqrt{L_m}^+ \left(\tfrac{1}{2}\Gamma_m - M_m\right) \sqrt{R_m}^+ \tag{6.5}$$

is contractive and the recurrence formulas

$$\Theta_m(w) = \sqrt{R_{m+1}}^+ \sqrt{R_m}\big(w\Theta_{m-1}(w) - K_m^*\big)\big(I - wK_m\Theta_{m-1}(w)\big)^{-1}\sqrt{L_m}^+\sqrt{L_{m+1}}$$

and

$$\Theta_m(w) = \sqrt{R_{m+1}}\sqrt{R_m}^+\big(I - w\Theta_{m-1}(w)K_m\big)^{-1}\big(w\Theta_{m-1}(w) - K_m^*\big)\sqrt{L_m}\sqrt{L_{m+1}}^+$$

hold for every choice of w in $\mathbb{D}$.

Proof. Since $(\Gamma_j)_{j=0}^n$ is a $q \times q$ Carathéodory sequence, for each $m \in \mathbb{N}_{0,n}$ the sequence $(\Gamma_j)_{j=0}^m$ is a $q \times q$ Carathéodory sequence as well. According to Proposition 3.4, for each $m \in \mathbb{N}_{0,n}$ the rational matrix-valued function Θ_m is well defined and its restriction onto $\mathbb{D}$ belongs to $\mathcal{S}_{q\times q}(\mathbb{D})$. Moreover, part (a) follows immediately from (1.4), (1.8), (1.10), and a well-known statement on Moore-Penrose inverses (see, e.g., [DFK, Lemma 1.1.6]). We consider now the case $n \geq 1$. For each $m \in \mathbb{N}_{0,n}$, let the matrix polynomials $\mathbf{a}_m$, $\mathbf{b}_m$, $\mathbf{c}_m$, and $\mathbf{d}_m$ be defined as in Remark 6.7. For each $m \in \mathbb{N}_{1,n}$, from [FKL1, Remark 2.1] we get that the matrix K_m is contractive and consequently that for each $w \in \mathbb{D}$ the matrices $I - wK_m\Theta_{m-1}(w)$ and $I - w\Theta_{m-1}(w)K_m$ are nonsingular. From Remark 6.7, Proposition 3.4, and Proposition 6.1 we can conclude that, for each $j \in \mathbb{N}_{0,n}$ and each $w \in \mathbb{D}$, the rational matrix-valued function Θ_j can be represented via

$$\Theta_j(w) = \sqrt{R_{j+1}}^+ \tilde{\mathbf{b}}_j^{[j]}(w)\big(\mathbf{d}_j(w)\big)^{-1}\sqrt{L_{j+1}} \tag{6.6}$$

and

$$\Theta_j(w) = \sqrt{R_{j+1}}\big(\mathbf{b}_j(w)\big)^{-1}\tilde{\mathbf{d}}_j^{[j]}(w)\sqrt{L_{j+1}}^+. \tag{6.7}$$

Now let $m \in \mathbb{N}_{1,n}$. Using Remark 6.7 we get

$$\begin{aligned}\Theta_m(w) &= \sqrt{R_{m+1}}^+\big(w\tilde{\mathbf{b}}_{m-1}^{[m-1]}(w) - t_m^*\mathbf{d}_{m-1}(w)\big)\\ &\qquad\cdot\big(\mathbf{d}_{m-1}(w) - wu_m\tilde{\mathbf{b}}_{m-1}^{[m-1]}(w)\big)^{-1}\sqrt{L_{m+1}}\\ &= \sqrt{R_{m+1}}^+\Big(w\tilde{\mathbf{b}}_{m-1}^{[m-1]}(w)\big(\mathbf{d}_{m-1}(w)\big)^{-1} - t_m^*\Big)\\ &\qquad\cdot\Big(I - wu_m\tilde{\mathbf{b}}_{m-1}^{[m-1]}(w)\big(\mathbf{d}_{m-1}(w)\big)^{-1}\Big)\sqrt{L_{m+1}}\end{aligned}$$

for each $w \in \mathbb{D}$. Because of (6.5), (6.3), and [FKL1, Remark 2.1] we have

$$t_m = \sqrt{L_m}^+ K_m \sqrt{R_m} \quad \text{and} \quad u_m = \sqrt{L_m}K_m\sqrt{R_m}^+.$$

Furthermore, the identities

$$\sqrt{L_m}\sqrt{L_m}^+\sqrt{L_{m+1}} = \sqrt{L_{m+1}} \quad \text{and} \quad \sqrt{R_{m+1}}^+\sqrt{R_m}\sqrt{R_m}^+ = \sqrt{R_{m+1}}^+$$

hold (see [DFK, Remark 3.4.3]). Thus it follows

$$\Theta_m(w) = \sqrt{R_{m+1}}^{+}\sqrt{R_m}\left(w\sqrt{R_m}^{+}\tilde{\mathbf{b}}_{m-1}^{[m-1]}(w)\big(\mathbf{d}_{m-1}(w)\big)^{-1} - K_m^*\sqrt{L_m}^{+}\right)$$
$$\cdot\left(I - w\sqrt{L_m}K_m\sqrt{R_m}^{+}\tilde{\mathbf{b}}_{m-1}^{[m-1]}(w)\big(\mathbf{d}_{m-1}(w)\big)^{-1}\right)^{-1}\sqrt{L_m}\sqrt{L_m}^{+}\sqrt{L_{m+1}}$$

and consequently by virtue of Remark 6.6 then

$$\Theta_m(w) = \sqrt{R_{m+1}}^{+}\sqrt{R_m}\left(w\sqrt{R_m}^{+}\tilde{\mathbf{b}}_{m-1}^{[m-1]}(w)\big(\mathbf{d}_{m-1}(w)\big)^{-1} - K_m^*\sqrt{L_m}^{+}\right)$$
$$\cdot\sqrt{L_m}\left(I - wK_m\sqrt{R_m}^{+}\tilde{\mathbf{b}}_{m-1}^{[m-1]}(w)\big(\mathbf{d}_{m-1}(w)\big)^{-1}\sqrt{L_m}\right)^{-1}\sqrt{L_m}^{+}\sqrt{L_{m+1}}$$

for each $w \in \mathbb{D}$. Thus, in view of (6.6) and $K_m^*\sqrt{L_m}^{+}\sqrt{L_m} = K_m^*$ we obtain that

$$\Theta_m(w) = \sqrt{R_{m+1}}^{+}\sqrt{R_m}\big(w\Theta_{m-1}(w) - K_m^*\big)\big(I - wK_m\Theta_{m-1}(w)\big)^{-1}\sqrt{L_m}^{+}\sqrt{L_{m+1}}$$

holds for each $w \in \mathbb{D}$. Using (6.7), the other recurrence formula of part (b) can be proved analogously. □

Corollary 6.9. *Let $n \in \mathbb{N}$ and let $(\Gamma_j)_{j=0}^n$ be a $q \times q$ Carathéodory sequence. For each $m \in \mathbb{N}_{1,n}$, then*

$$\Theta_m(0) = -\sqrt{R_{m+1}}^{+}\sqrt{R_m}\,K_m^*\sqrt{L_m}^{+}\sqrt{L_{m+1}}$$

and

$$\Theta_m(0) = -\sqrt{R_{m+1}}\sqrt{R_m}^{+}K_m^*\sqrt{L_m}\sqrt{L_{m+1}}^{+},$$

where K_m is the matrix given by (6.5)*. Moreover, if $m \in \mathbb{N}_{1,n}$ and if $w \in \mathbb{D}$ are such that $\Theta_{m-1}(w) = 0$, then $\Theta_m(w) = \Theta_m(0)$.*

Proof. Apply Theorem 6.8. □

Corollary 6.10. *Let $n \in \mathbb{N}$ and let $(\Gamma_j)_{j=0}^n$ be a $q \times q$ Carathéodory sequence which satisfies $\Gamma_n = 2M_n$, where M_n is defined by* (6.4)*. For each $w \in \mathbb{D}$, then*

$$\Theta_n(w) = w\Theta_{n-1}(w).$$

Proof. In view of $\Gamma_n = 2M_n$, an application of [DFK, Remark 3.4.3] implies

$$L_{n+1} = L_n \quad \text{and} \quad R_{n+1} = R_n$$

as well as (6.5) yields

$$K_n = 0_{q\times q}.$$

Thus part (b) of Theorem 6.8 and (3.4) lead to

$$\Theta_n(w) = \sqrt{R_{n+1}}^{+}\sqrt{R_n}w\Theta_{n-1}(w)\sqrt{L_n}^{+}\sqrt{L_{n+1}} = w\Theta_{n-1}(w)$$

for each $w \in \mathbb{D}$. □

Let $n \in \mathbb{N}$ and let $(\Gamma_j)_{j=0}^n$ be a $q \times q$ Carathéodory sequence. Since the matrix-valued functions $\Theta_0, \Theta_1, \ldots, \Theta_n$ given by (3.4) are rational, there is a finite subset $\mathcal{N}$ of $\mathbb{C} \setminus \mathbb{D}$ such that the recurrence formulas stated in part (b) of Theorem 6.8 hold actually for each $m \in \mathbb{N}_{1,n}$ and each $w \in \mathbb{C} \setminus \mathcal{N}$.

Now we are going to show that if a $q \times q$ Carathéodory sequence with nonsingular matrix Γ_0 is given, then the matrix-valued functions $\Theta_0^\diamond, \Theta_1^\diamond, \ldots, \Theta_n^\diamond$ defined by (3.9) fulfill some recurrence relations as well.

Theorem 6.11. *Let $n \in \mathbb{N}$ and let $(\Gamma_j)_{j=0}^n$ be a $q \times q$ Carathéodory sequence such that the matrix Γ_0 is nonsingular. Then:*

(a) *$\Theta_0^\diamond$ is the constant function with value $\sqrt{\operatorname{Re}\Gamma_0}^{+}\Gamma_0^* \Gamma_0^{-1}\sqrt{\operatorname{Re}\Gamma_0}$.*
(b) *If $n \geq 1$, then for each $m \in \mathbb{N}_{1,n}$ and each $w \in \mathbb{D}$ the recurrence formulas*

$$\Theta_m^\diamond(w) = \sqrt{R_{m+1}}^{+}\sqrt{R_m}\big(w\Theta_{m-1}^\diamond(w) + K_m^*\big)\big(I + wK_m\Theta_{m-1}^\diamond(w)\big)^{-1}\sqrt{L_m}^{+}\sqrt{L_{m+1}}$$

and

$$\Theta_m^\diamond(w) = \sqrt{R_{m+1}}\sqrt{R_m}^{+}\big(I + w\Theta_{m-1}^\diamond(w)K_m\big)^{-1}\big(w\Theta_{m-1}^\diamond(w) + K_m^*\big)\sqrt{L_m}\sqrt{L_{m+1}}^{+}$$

hold, where K_m is the contractive matrix given by (6.5).

Proof. Using Remark 3.7, Proposition 3.8, part (b) of Proposition 6.1, Remark 6.6, and Remark 6.7 one can prove Theorem 6.11 analogously to Theorem 6.8. We omit the details. □

Corollary 6.12. *Let $n \in \mathbb{N}$ and let $(\Gamma_j)_{j=0}^n$ be a $q \times q$ Carathéodory sequence with nonsingular matrix Γ_0. For each $m \in \mathbb{N}_{1,n}$, then $\Theta_m^\diamond(0) = -\Theta_m(0)$. Moreover, if $m \in \mathbb{N}_{1,n}$ and $w \in \mathbb{D}$ are such that $\Theta_{m-1}^\diamond(w) = 0$, then $\Theta_m^\diamond(w) = \Theta_m^\diamond(0)$.*

Proof. Use Theorem 6.11 and Corollary 6.9. □

Corollary 6.13. *Let $n \in \mathbb{N}$ and let $(\Gamma_j)_{j=0}^n$ be a $q \times q$ Carathéodory sequence such that the matrix Γ_0 is nonsingular and that the identity $\Gamma_n = 2M_n$ is fulfilled, where M_n is defined by* (6.4). *Then $\Theta_n^\diamond(w) = w\Theta_{n-1}^\diamond(w)$ for each $w \in \mathbb{D}$.*

Proof. Using the same argumentation as in the proof of Corollary 6.10, the assertion is an easy consequence of part (b) of Theorem 6.11. □

Let $n \in \mathbb{N}$ and let $(\Gamma_j)_{j=0}^n$ be a $q \times q$ Carathéodory sequence with nonsingular matrix Γ_0. Since $\Theta_0^\diamond, \Theta_1^\diamond, \ldots, \Theta_n^\diamond$ are rational matrix-valued functions, there is a finite subset $\mathcal{M}$ of $\mathbb{C} \setminus \mathbb{D}$ such that the recurrence formulas in Theorem 6.11 hold actually for each $m \in \mathbb{N}_{1,n}$ and each $w \in \mathbb{C} \setminus \mathcal{M}$. Furthermore, we note marginally that one can also verify recurrence relations for the rational matrix-valued functions $\Theta_0^\triangleright, \Theta_1^\triangleright, \ldots, \Theta_n^\triangleright$ given by (3.17) using the same argumentation as in the proof of Theorem 6.8 (respectively, Theorem 6.11), where the function $\Theta_m^\triangleright$ can be calculated based on $\Theta_{m-1}^\triangleright$, $\Theta_{m-1}^\diamond$, Θ_{m-1}, and K_m for each $m \in \mathbb{N}_{1,n}$.

7. Some remarks on central $q \times q$ Carathéodory functions

In the present section, we are going to describe the situation that, starting from a $q \times q$ Carathéodory sequence $(\Gamma_j)_{j=0}^n$ and a point w belonging to the unit disk $\mathbb{D}$, the value $\Omega_{c,n}(w)$ of the central $q \times q$ Carathéodory function $\Omega_{c,n}$ corresponding to $(\Gamma_j)_{j=0}^n$ coincides with the center $\mathcal{M}_{n+1}(w)$ of the Weyl matrix ball which is given by the value set $\{\Omega(w) : \Omega \in \mathcal{C}_q[\mathbb{D}, (\Gamma_j)_{j=0}^n]\}$ according to Theorem 1.1.

Let us consider an arbitrary nonnegative integer n and an arbitrary $q \times q$ Carathéodory sequence $(\Gamma_j)_{j=0}^n$. Then, in view of [DFK, Theorem 3.4.1], the set of all complex $q \times q$ matrices Γ_{n+1} such that $(\Gamma_j)_{j=0}^{n+1}$ is a $q \times q$ Carathéodory sequence coincides with the matrix ball $\mathfrak{K}\big(2M_{n+1}; \sqrt{2L_{n+1}}, \sqrt{2R_{n+1}}\big)$, where the matrices M_{n+1}, L_{n+1}, and R_{n+1} are defined by (6.4), (1.4), and (1.6). Thus the choice $\Gamma_{n+1} := 2M_{n+1}$ yields a particular $q \times q$ Carathéodory sequence $(\Gamma_j)_{j=0}^{n+1}$ and the set of all complex $q \times q$ matrices Γ_{n+2} such that $(\Gamma_j)_{j=0}^{n+2}$ is a $q \times q$ Carathéodory sequence coincides with $\mathfrak{K}\big(2M_{n+2}; \sqrt{2L_{n+2}}, \sqrt{2R_{n+2}}\big)$. In this way, choosing

$$\Gamma_{n+1+k} := 2M_{n+1+k} \tag{7.1}$$

for each $k \in \mathbb{N}_0$ one obtains a particular $q \times q$ Carathéodory sequence $(\Gamma_j)_{j=0}^{\infty}$ and hence a particular function which belongs to $\mathcal{C}_q[\mathbb{D}, (\Gamma_j)_{j=0}^n]$, the so-called *central $q \times q$ Carathéodory function $\Omega_{c,n}$ corresponding to $(\Gamma_j)_{j=0}^n$*. Clearly, it admits the Taylor series representation

$$\Omega_{c,n}(z) = \sum_{k=0}^{\infty} \Gamma_k z^k, \quad z \in \mathbb{D},$$

where for each $k \in \mathbb{N}_0$ the matrices Γ_{n+1+k} are defined by (7.1). Moreover, for each $z \in \mathbb{D}$, it admits the representations

$$\Omega_{c,n}(z) = a_n(z)\big(b_n(z)\big)^{-1} \quad \text{and} \quad \Omega_{c,n}(z) = \big(d_n(z)\big)^{-1} c_n(z), \tag{7.2}$$

where V_n and W_n are arbitrary matrices which belong to $\widetilde{\mathcal{Y}}_n$ and $\widetilde{\mathcal{Z}}_n$, respectively, and where a_n, b_n, c_n, and d_n are the matrix polynomials defined by (1.7), (1.8), (1.9), and (1.10) (see [FK3, Remark 1.1 and Theorems 1.2., 1.7, and 2.3]).

In the following, we use again the notations of Theorem 1.1 and Remark 4.2.

Theorem 7.1. *Let $n \in \mathbb{N}_0$, let $(\Gamma_j)_{j=0}^n$ be a $q \times q$ Carathéodory sequence, let $\Omega_{c,n}$ be the central $q \times q$ Carathéodory function corresponding to $(\Gamma_j)_{j=0}^n$, and let $w \in \mathbb{D}$. Furthermore, let the matrices L_{n+1} and R_{n+1} be defined by* (1.4) *and* (1.6)*. Then the following statements are equivalent:*

(i) $\Omega_{c,n}(w) = \mathcal{M}_{n+1}(w)$.
(ii) $w\Theta_n(w) = 0$.
(iii) $\mathcal{A}_{n+1}(w) = \big(d_n(w)\big)^{-1}\sqrt{L_{n+1}}$.
(iv) $\mathcal{B}_{n+1}(w) = \sqrt{R_{n+1}}\,\big(b_n(w)\big)^{-1}$.
(v) $\mathcal{L}_{n+1}(w) = \big(d_n(w)\big)^{-1} L_{n+1}\big(d_n(w)\big)^{-*}$.
(vi) $\mathcal{R}_{n+1}(w) = \big(b_n(w)\big)^{-*} R_{n+1}\big(b_n(w)\big)^{-1}$.

Proof. (i) $\Leftrightarrow$ (ii): If $w = 0$, the equivalence of (i) and (ii) is an easy consequence of (7.2) and (1.11). Now suppose $w \neq 0$. In view of Proposition 3.4 we see that $\big(-w\Theta_n(w)\big)^*$ is contractive. Because of (7.2) we have on the one hand

$$\begin{aligned}\Omega_{c,n}(w) &= a_n(w)(b_n(w))^{-1} \\ &= \Big(-w\tilde{c}_n^{[n]}(w)F_1(w) + a_n(w)\Big)\Big(w\tilde{d}_n^{[n]}(w)F_1(w) + b_n(w)\Big)^{-1}\end{aligned}$$

with $F_1(w) := \sqrt{L_{n+1}}^{+}0\sqrt{R_{n+1}}$, where on the other hand (1.14) implies

$$\mathcal{M}_{n+1}(w) = \Big(-w\tilde{c}_n^{[n]}(w)F_2(w) + a_n(w)\Big)\Big(w\tilde{d}_n^{[n]}(w)F_2(w) + b_n(w)\Big)^{-1}$$

with $F_2(w) := \sqrt{L_{n+1}}^{+}\big(-w\Theta_n(w)\big)^*\sqrt{R_{n+1}}$. Therefore, in view of $w \neq 0$ and part (a) of Proposition 2.2 we get that (i) is necessary and sufficient for

$$L_{n+1}L_{n+1}^{+}\big(-w\Theta_n(w)\big)^*R_{n+1}R_{n+1}^{+} = L_{n+1}L_{n+1}^{+}0R_{n+1}R_{n+1}^{+},$$

i.e., (i) is equivalent to

$$R_{n+1}R_{n+1}^{+}\big(w\Theta_n(w)\big)L_{n+1}L_{n+1}^{+} = 0.$$

Thus, because of (3.4), statement (i) is also tantamount to (ii) in this case.
(ii) $\Rightarrow$ (iii): This implication follows immediately from (1.12).
(iii) $\Rightarrow$ (v): Since the definitions of $\mathcal{A}_{n+1}(w)$ and $\mathcal{L}_{n+1}(w)$ imply the relation

$$\mathcal{A}_{n+1}(w)\mathcal{A}_{n+1}^*(w) = \mathcal{L}_{n+1}(w),$$

we can directly see that (iii) leads to (v).
(v) $\Rightarrow$ (ii): Because of (v) we obtain

$$\begin{aligned}&\big(d_n(w)\big)^{-1}\sqrt{L_{n+1}}\,\big(I - |w|^2\Theta_n^*(w)\Theta_n(w)\big)^{-1}\sqrt{L_{n+1}}\,\big(d_n(w)\big)^{-*} \\ &\quad = \mathcal{L}_{n+1}(w) \;=\; \big(d_n(w)\big)^{-1}L_{n+1}\big(d_n(w)\big)^{-*}\end{aligned}$$

and hence

$$\sqrt{L_{n+1}}\,\big(I - |w|^2\Theta_n^*(w)\Theta_n(w)\big)^{-1}\sqrt{L_{n+1}} = L_{n+1}.$$

Therefore, taking into account (3.5) and Remark 6.6 we can conclude that

$$\begin{aligned}L_{n+1} &= \sqrt{L_{n+1}}\left(I - |w|^2\Theta_n^*(w)\sqrt{R_{n+1}}^{+}\tilde{b}_n^{[n]}(w)\big(d_n(w)\big)^{-1}\sqrt{L_{n+1}}\right)^{-1}\sqrt{L_{n+1}} \\ &= \left(I - |w|^2\sqrt{L_{n+1}}\Theta_n^*(w)\sqrt{R_{n+1}}^{+}\tilde{b}_n^{[n]}(w)\big(d_n(w)\big)^{-1}\right)^{-1}L_{n+1}.\end{aligned}$$

This yields

$$\begin{aligned}L_{n+1} &= \left(I - |w|^2\sqrt{L_{n+1}}\Theta_n^*(w)\sqrt{R_{n+1}}^{+}\tilde{b}_n^{[n]}(w)\big(d_n(w)\big)^{-1}\right)L_{n+1} \\ &= L_{n+1} - |w|^2\sqrt{L_{n+1}}\Theta_n^*(w)\Theta_n(w)\sqrt{L_{n+1}}\end{aligned}$$

and consequently

$$|w|^2\sqrt{L_{n+1}}\Theta_n^*(w)\Theta_n(w)\sqrt{L_{n+1}} = 0.$$

Thus, in view of (3.5), we get

$$\big(w\Theta_n(w)\big)^*\big(w\Theta_n(w)\big) = |w|^2\sqrt{L_{n+1}}^{+}\sqrt{L_{n+1}}\Theta_n^*(w)\Theta_n(w)\sqrt{L_{n+1}}\sqrt{L_{n+1}}^{+} = 0.$$

Consequently, (ii) holds.
(ii) $\Rightarrow$ (iv): This follows immediately from (1.13).
(iv) $\Rightarrow$ (vi): By definition of $\mathcal{B}_{n+1}(w)$ and $\mathcal{R}_{n+1}(w)$ we see that (iv) leads to (vi).
(vi) $\Rightarrow$ (ii): This can be proved analogously to the implication "(v) $\Rightarrow$ (ii)". We omit the details. □

Now we consider a particular situation, where one of the equivalent conditions stated in Theorem 7.1 is satisfied.

Proposition 7.2. *Let $n \in \mathbb{N}$ and let $(\Gamma_j)_{j=0}^{n-1}$ be a $q \times q$ Carathéodory sequence. Let the matrices L_n, R_n, and M_n be defined by* (1.4), (1.6), *and* (6.4). *Furthermore, let $w \in \mathbb{D}$ and let*

$$\Gamma_n := 2M_n + \sqrt{2L_n}\left(w\Theta_{n-1}(w)\right)^* \sqrt{2R_n}. \tag{7.3}$$

Then $(\Gamma_j)_{j=0}^{n}$ is $q \times q$ Carathéodory sequence, $\Theta_n(w) = 0$ holds, and the central $q \times q$ Carathéodory function $\Omega_{c,n}$ corresponding to $(\Gamma_j)_{j=0}^{n}$ fulfills

$$\Omega_{c,n}(w) = \mathcal{M}_{n+1}(w). \tag{7.4}$$

Moreover, if $(\Gamma_j)_{j=0}^{n-1}$ is a nondegenerate $q \times q$ Carathéodory sequence, then $(\Gamma_j)_{j=0}^{n}$ is a nondegenerate $q \times q$ Carathéodory sequence as well.

Proof. Proposition 3.4 yields that matrix $\left(w\Theta_{n-1}(w)\right)^*$ is strictly contractive. Consequently, from [DFK, Theorem 3.4.1] we get that $(\Gamma_j)_{j=0}^{n}$ is a $q \times q$ Carathéodory sequence and, moreover, that $(\Gamma_j)_{j=0}^{n}$ is a nondegenerate $q \times q$ Carathéodory sequence if $(\Gamma_j)_{j=0}^{n-1}$ is a nondegenerate $q \times q$ Carathéodory sequence. Because of (3.4) and (7.3) we have

$$\left(w\Theta_{n-1}(w)\right)^* = \sqrt{L_n}^{\,+}\left(\tfrac{1}{2}\Gamma_n - M_n\right)\sqrt{R_n}^{\,+}.$$

Thus application of Theorem 6.8 provides us $\Theta_n(w) = 0$. Hence using Theorem 7.1 we get (7.4). □

In the nondegenerate case, the statement of Proposition 7.2 (respectively, Theorem 7.1) can be extended as follows.

Proposition 7.3. *Let $n \in \mathbb{N}$, let $(\Gamma_j)_{j=0}^{n}$ be a nondegenerate $q \times q$ Carathéodory sequence, let $\Omega_{c,n}$ be the central $q \times q$ Carathéodory function corresponding to $(\Gamma_j)_{j=0}^{n}$, and let $w \in \mathbb{D} \setminus \{0\}$. Then* (7.4) *is equivalent to*

$$\Gamma_n = 2M_n + \sqrt{2L_n}\left(w\Theta_{n-1}(w)\right)^* \sqrt{2R_n}. \tag{7.5}$$

Proof. By virtue of Proposition 7.2 we see that (7.5) implies (7.4). We suppose now (7.4). Because of Theorem 7.1 and $w \neq 0$ it follows $\Theta_n(w) = 0$. Since the $q \times q$ Carathéodory sequence $(\Gamma_j)_{j=0}^{n}$ is nondegenerate, from [FKL1, Lemma 5.1] we know that the matrices L_{n+1}, L_n, R_{n+1}, and R_n are positive Hermitian. Using this in combination with $\Theta_n(w) = 0$ and Theorem 6.8 we get then the identity

$$0 = w\Theta_{n-1}(w) - K_n^*$$

and, in view of (6.5) and $\Gamma_n \in \mathfrak{K}\left(2M_n; \sqrt{2L_n}, \sqrt{2R_n}\right)$, consequently (7.5). □

Let $n \in \mathbb{N}$, let $(\Gamma_j)_{j=0}^n$ be a $q \times q$ Carathéodory sequence with nonsingular matrix Γ_0, and let $\Omega_{c,n}$ be the central $q \times q$ Carathéodory function corresponding to $(\Gamma_j)_{j=0}^n$. Then $\det \Omega_{c,n}$ does not vanish in $\mathbb{D}$ (see [FK1, Remark 30 in Part I]) and the central $q \times q$ Carathéodory function $\Omega_{c,n}^\sharp$ corresponding to the reciprocal $q \times q$ Carathéodory sequence $(\Gamma_j^\sharp)_{j=0}^n$ corresponding to $(\Gamma_j)_{j=0}^n$ coincides with the matrix-valued function $(\Omega_{c,n})^{-1}$ (see [FK3, Theorem 3.3]). Taking this into account we get the following results which are analogues of Theorem 7.1, Proposition 7.2, and Proposition 7.3 with respect to the Weyl matrix balls stated in Theorem 5.2. Here we use again the notations of Theorem 5.2 and Remark 5.3.

Theorem 7.4. *Let $n \in \mathbb{N}_0$, let $(\Gamma_j)_{j=0}^n$ be a $q \times q$ Carathéodory sequence with nonsingular matrix Γ_0, let $\Omega_{c,n}$ be the central $q \times q$ Carathéodory function corresponding to $(\Gamma_j)_{j=0}^n$, and let $w \in \mathbb{D}$. Furthermore, let the matrices L_{n+1} and R_{n+1} be defined by* (1.4) *and* (1.6)*. Then the following statements are equivalent:*

(i) $\big(\Omega_{c,n}(w)\big)^{-1} = \mathcal{M}'_{n+1}(w)$.
(ii) $w\Theta_n^\diamond(w) = 0$.
(iii) $\mathcal{A}'_{n+1}(w) = \big(c_n(w)\big)^{-1}\sqrt{L_{n+1}}$.
(iv) $\mathcal{B}'_{n+1}(w) = \sqrt{R_{n+1}}\,\big(a_n(w)\big)^{-1}$.
(v) $\mathcal{L}'_{n+1}(w) = \big(c_n(w)\big)^{-1}L_{n+1}\big(c_n(w)\big)^{-*}$.
(vi) $\mathcal{R}'_{n+1}(w) = \big(a_n(w)\big)^{-*}R_{n+1}\big(a_n(w)\big)^{-1}$.

Proof. (i) $\Leftrightarrow$ (ii): If $w = 0$, the equivalence of (i) and (ii) is an easy consequence of (7.2) and (5.15). Now suppose $w \neq 0$. Moreover, let $(\Gamma_j^\sharp)_{j=0}^n$ be the reciprocal $q \times q$ Carathéodory sequence corresponding to $(\Gamma_j)_{j=0}^n$, let $a_n^\sharp$, $b_n^\sharp$, $c_n^\sharp$, and $d_n^\sharp$ be the matrix polynomials given by (5.3), (5.4), (5.5), and (5.6), and let $\Theta_n^\sharp$ be the rational matrix-valued function defined by (5.8). In view of Proposition 3.4 (with respect to the $q \times q$ Carathéodory sequence $(\Gamma_j^\sharp)_{j=0}^n$) we see that $\big(-w\Theta_n^\sharp(w)\big)^*$ is contractive. Because of (7.2), Remark 3.7, and (5.7), on the one hand we have

$$\begin{aligned}\big(\Omega_{c,n}(w)\big)^{-1} &= b_n(w)\big(a_n(w)\big)^{-1} = a_n^\sharp(w)\big(b_n^\sharp(w)\big)^{-1}\\ &= \Big(-w\widetilde{(c_n^\sharp)}^{[n]}(w)F_1^\bullet(w) + a_n^\sharp(w)\Big)\Big(w\widetilde{(d_n^\sharp)}^{[n]}(w)F_1^\bullet(w) + b_n^\sharp(w)\Big)^{-1}\end{aligned}$$

with $F_1^\bullet(w) := \sqrt{L_{n+1}^\sharp}^{+}\,0\sqrt{R_{n+1}^\sharp}$, where on the other hand (5.24) and (5.19) imply

$$\mathcal{M}'_{n+1}(w) = \Big(-w\widetilde{(c_n^\sharp)}^{[n]}(w)F_2^\bullet(w) + a_n^\sharp(w)\Big)\Big(w\widetilde{(d_n^\sharp)}^{[n]}(w)F_2^\bullet(w) + b_n^\sharp(w)\Big)^{-1}$$

with $F_2^\bullet(w) := \sqrt{L_{n+1}^\sharp}^{+}\big(-w\Theta_n^\sharp(w)\big)^*\sqrt{R_{n+1}^\sharp}$. Therefore, in view of $w \neq 0$ and part (a) of Proposition 2.2, we get that (i) is necessary and sufficient for

$$L_{n+1}^\sharp(L_{n+1}^\sharp)^+\big(-w\Theta_n^\sharp(w)\big)^*R_{n+1}^\sharp(R_{n+1}^\sharp)^+ = L_{n+1}^\sharp(L_{n+1}^\sharp)^+0R_{n+1}^\sharp(R_{n+1}^\sharp)^+,$$

i.e., (i) is equivalent to

$$R_{n+1}^\sharp(R_{n+1}^\sharp)^+\big(w\Theta_n^\sharp(w)\big)L_{n+1}^\sharp(L_{n+1}^\sharp)^+ = 0.$$

Thus, because of (5.8), statement (i) is tantamount to

$$w\Theta_n^\sharp(w) = 0. \tag{7.6}$$

Since $w\Theta_n^\diamond(w) = 0$ leads immediately to (7.6) in view of (5.12), we obtain that (ii) gives rise to (i). Conversely, if (i) is satisfied, then (7.6) holds and hence

$$w\Theta_n^\sharp(w)\sqrt{L_{n+1}^\sharp} = 0.$$

By virtue of (5.11) and (5.12) it follows then

$$0 = wV\Theta_n^\diamond(w)U\sqrt{L_{n+1}^\sharp} = wV\Theta_n^\diamond(w)\sqrt{L_{n+1}}\Gamma_0^{-*},$$

i.e., we get $w\Theta_n^\diamond(w)\sqrt{L_{n+1}} = 0$, which results in $w\Theta_n^\diamond(w) = 0$ due to (3.9). Consequently, (i) also implies (ii) in that case.
The remaining part of the assertion can be analogously proved as Theorem 7.1 by a straightforward calculation. We omit the details. □

Now we consider a particular situation, where one of the equivalent conditions stated in Theorem 7.4 is satisfied.

Proposition 7.5. *Let $n \in \mathbb{N}$ and let $(\Gamma_j)_{j=0}^{n-1}$ be a $q \times q$ Carathéodory sequence with nonsingular matrix Γ_0. Furthermore, let $w \in \mathbb{D}$ and let*

$$\Gamma_n := 2M_n + \sqrt{2L_n}\left(-w\Theta_{n-1}^\diamond(w)\right)^*\sqrt{2R_n}.$$

Then $(\Gamma_j)_{j=0}^n$ is a $q \times q$ Carathéodory sequence, $\Theta_n^\diamond(w) = 0$ holds, and the central $q \times q$ Carathéodory function $\Omega_{c,n}$ corresponding to $(\Gamma_j)_{j=0}^n$ fulfills

$$\left(\Omega_{c,n}(w)\right)^{-1} = \mathcal{M}'_{n+1}(w).$$

Moreover, if $(\Gamma_j)_{j=0}^{n-1}$ is a nondegenerate $q \times q$ Carathéodory sequence, then $(\Gamma_j)_{j=0}^n$ is a nondegenerate $q \times q$ Carathéodory sequence as well.

Proof. Applying Proposition 3.8, [DFK, Theorem 3.4.1], Theorem 6.11, and Theorem 7.4 one can prove Proposition 7.5 analogously to Proposition 7.2. □

In the nondegenerate case, the statement of Proposition 7.5 (respectively, Theorem 7.4) can be extended as follows.

Proposition 7.6. *Let $n \in \mathbb{N}$ and let $(\Gamma_j)_{j=0}^n$ be a $q \times q$ nondegenerate Carathéodory sequence, let $\Omega_{c,n}$ be the central $q \times q$ Carathéodory function corresponding to $(\Gamma_j)_{j=0}^n$, and let $w \in \mathbb{D} \setminus \{0\}$. Then $\left(\Omega_{c,n}(w)\right)^{-1} = \mathcal{M}'_{n+1}(w)$ is equivalent to*

$$\Gamma_n = 2M_n + \sqrt{2L_n}\left(-w\Theta_{n-1}^\diamond(w)\right)^*\sqrt{2R_n}. \tag{7.7}$$

Proof. Using Proposition 7.5, Theorem 7.4, [FKL1, Lemma 5.1], and Theorem 6.11 one can prove Proposition 7.6 analogously to Proposition 7.3. □

The following example illustrates that, in the case that the $q \times q$ Carathéodory sequence $(\Gamma_j)_{j=0}^n$ is degenerate, equation (7.5) (respectively, (7.7)) is not necessary for the fact that $\Theta_n(w) = 0$ (respectively, $\Theta_n^\diamond(w) = 0$) holds for some $w \in \mathbb{D}$.

Example 7.7. Let U be a unitary $q\times q$ matrix, let $\Gamma_0 := I_q$, and let $\Gamma_1 := 2U$. Then $(\Gamma_j)_{j=0}^1$ is a $q \times q$ Carathéodory sequence with $L_1 = R_1 = I_q$, $L_2 = R_2 = 0_{q\times q}$, and $\det \Gamma_0 \neq 0$. In view of (3.4) and (3.10) we have $\Theta_1(w) = 0$ and $\Theta_1^\diamond(w) = 0$ for each $w \in \mathbb{D}$. On the other hand,

$$\Gamma_1 = 2U \neq 2\overline{w}I_q = 2M_1 + \sqrt{2L_1}\big(w\Theta_0(w)\big)^*\sqrt{2R_1}$$

and

$$\Gamma_1 = 2U \neq -2\overline{w}I_q = 2M_1 + \sqrt{2L_1}\big(-w\Theta_0(w)\big)^*\sqrt{2R_1}.$$

We finish this paper with some comments on central $q \times q$ Carathéodory functions, which are conclusions of certain considerations carried out above. Firstly, we will point out that Proposition 2.2 leads to a characterization of the fact that given finite $q \times q$ Carathéodory sequences coincide in terms of equality of some value of the corresponding central $q \times q$ Carathéodory functions. The following result paves the way for a proof of this characterization.

Lemma 7.8. *Let $n \in \mathbb{N}_0$ and let $(\Gamma_j)_{j=0}^n$ be a $q\times q$ Carathéodory sequence. If $n \geq 1$, then let $V_n \in \widetilde{\mathcal{Y}}_n$ and let $W_n \in \widetilde{\mathcal{Z}}_n$. Let the matrix polynomials a_n, b_n, c_n, and d_n be defined by* (1.7), (1.8), (1.9), *and* (1.10). *Let the matrices L_{n+1} and R_{n+1} be given by* (1.4) *and* (1.6). *Furthermore, for $j \in \{1, 2\}$, let K_j be a contractive $q \times q$ matrix and let $F_j := \sqrt{L_{n+1}}^+ K_j \sqrt{R_{n+1}}$. Then the complex $q \times q$ matrices $z\tilde{d}_n^{[n]}(z)F_1 + b_n(z)$ and $z\tilde{d}_n^{[n]}(z)F_2 + b_n(z)$ are both nonsingular for each $z \in \mathbb{D}$ and, moreover, the following statements are equivalent:*

(i) $L_{n+1}L_{n+1}^+K_1R_{n+1}R_{n+1}^+ = L_{n+1}L_{n+1}^+K_2R_{n+1}R_{n+1}^+$.

(ii) *For each $z \in \mathbb{D}$,*

$$\begin{aligned}&\Big(-z\tilde{c}_n^{[n]}(z)F_1 + a_n(z)\Big)\Big(z\tilde{d}_n^{[n]}(z)F_1 + b_n(z)\Big)^{-1}\\ &\quad= \Big(-z\tilde{c}_n^{[n]}(z)F_2 + a_n(z)\Big)\Big(z\tilde{d}_n^{[n]}(z)F_2 + b_n(z)\Big)^{-1}.\end{aligned} \tag{7.8}$$

(iii) *There is some $z \in \mathbb{D} \setminus \{0\}$ such that* (7.8) *is satisfied.*

Proof. Apply part (a) of Proposition 2.2. □

Proposition 7.9. *Let $n \in \mathbb{N}_0$, let $(\Gamma_j)_{j=0}^{n+1}$ and $(\widetilde{\Gamma}_j)_{j=0}^{n+1}$ be $q \times q$ Carathéodory sequences such that $\Gamma_k = \widetilde{\Gamma}_k$ for each $k \in \mathbb{N}_{0,n}$, and let $\Omega_{c,n+1}$ and $\widetilde{\Omega}_{c,n+1}$ be the central $q \times q$ Carathéodory function corresponding to $(\Gamma_j)_{j=0}^{n+1}$ and $(\widetilde{\Gamma}_j)_{j=0}^{n+1}$, respectively. Then $\Gamma_{n+1} = \widetilde{\Gamma}_{n+1}$ if and only if there is some $z \in \mathbb{D} \setminus \{0\}$ such that*

$$\Omega_{c,n+1}(z) = \widetilde{\Omega}_{c,n+1}(z). \tag{7.9}$$

Proof. If the equality $\Gamma_{n+1} = \widetilde{\Gamma}_{n+1}$ is fulfilled, then the $q \times q$ Carathéodory sequences $(\Gamma_j)_{j=0}^{n+1}$ and $(\widetilde{\Gamma}_j)_{j=0}^{n+1}$ coincide. Thus, by definition we have $f_{c,n+1} = \widetilde{f}_{c,n+1}$. In particular, there is some $z \in \mathbb{D} \setminus \{0\}$ such that the identity (7.9) holds. Conversely, we assume now that (7.9) is satisfied for some $z \in \mathbb{D} \setminus \{0\}$. Because of the choice of $(\Gamma_j)_{j=0}^{n+1}$ and $(\widetilde{\Gamma}_j)_{j=0}^{n+1}$, it follows that $(\Gamma_j)_{j=0}^n$ is a $q \times q$ Carathéodory

sequence, where $\Gamma_k = \widetilde{\Gamma}_k$ for each $k \in \mathbb{N}_{0,n}$. Hence, from [DFK, Theorem 3.4.1] we obtain that there are some contractive $q \times q$ matrices K_1 and K_2 fulfilling

$$\Gamma_{n+1} = 2M_{n+1} + \sqrt{2L_{n+1}}K_1\sqrt{2R_{n+1}} \tag{7.10}$$

and

$$\widetilde{\Gamma}_{n+1} = 2M_{n+1} + \sqrt{2L_{n+1}}K_2\sqrt{2R_{n+1}}, \tag{7.11}$$

where L_{n+1}, R_{n+1}, and M_{n+1} are defined as in (1.4), (1.6), and (6.4). Therefore, if the matrix polynomials a_n, b_n, c_n, and d_n are given by (1.7), (1.8), (1.9), and (1.10) with some $V_n \in \widetilde{\mathcal{Y}}_n$ and $W_n \in \widetilde{\mathcal{Z}}_n$ if $n \geq 1$, then (7.9) and [FKL1, Corollary 2.7] yield (7.8) for some $z \in \mathbb{D} \setminus \{0\}$, where $F_j := \sqrt{L_{n+1}}^{+} K_j \sqrt{R_{n+1}}$ for $j \in \{1, 2\}$. By virtue of Lemma 7.8 we get

$$L_{n+1}L_{n+1}^{+}K_1R_{n+1}R_{n+1}^{+} = L_{n+1}L_{n+1}^{+}K_2R_{n+1}R_{n+1}^{+}$$

which implies finally $\Gamma_{n+1} = \widetilde{\Gamma}_{n+1}$ in view of (7.10) and (7.11). □

We note that, since the central $q \times q$ Carathéodory function corresponding to a $q \times q$ Carathéodory sequence $(\tilde{\Gamma}_j)_{j=0}^0$ is the constant function (defined on $\mathbb{D}$) with value $\tilde{\Gamma}_0$ (see, e.g., [FK3, Remark 1.1]), the assertion of Proposition 7.9 is obvious in the case of given $q \times q$ Carathéodory sequences $(\Gamma_j)_{j=0}^0$ and $(\tilde{\Gamma}_j)_{j=0}^0$.

Let $\Omega \in \mathcal{C}_q(\mathbb{D})$. Then the matricial version of the F. Riesz-Herglotz Representation Theorem (see, e.g., [DFK, Theorem 2.2.2]) shows that there is a unique nonnegative Hermitian $q \times q$ Borel measure F defined on the σ-algebra $\mathfrak{B}_{\mathbb{T}}$ of all Borel subsets of the unit circle $\mathbb{T}$ such that

$$\Omega(w) = \int_{\mathbb{T}} \frac{z+w}{z-w}\, F(dz) + i \operatorname{Im}\Omega(0) \tag{7.12}$$

for each $w \in \mathbb{D}$. This nonnegative Hermitian measure F is called the *Riesz-Herglotz measure of* Ω.

Let $\underline{\lambda}$ be denote the linear Lebesgue measure defined on $\mathfrak{B}_{\mathbb{T}}$. If Γ_0 is a complex $q \times q$ matrix with nonnegative Hermitian real part $\operatorname{Re}\Gamma_0$, then from [FK3, Remark 1.1] and the matricial versions of the F. Riesz-Herglotz Representation Theorem and the Herglotz-Bochner Theorem (see, e.g., [DFK, Theorem 2.2.1]) we know that the Riesz-Herglotz measure $F_{c,0}$ of the central $q \times q$ Carathéodory function $\Omega_{c,0}$ corresponding to $(\Gamma_j)_{j=0}^0$ is absolutely continuous with respect to $\frac{1}{2\pi}\underline{\lambda}$ and that $f_0 : \mathbb{T} \to \mathbb{C}^{q\times q}$ defined by $f_0(z) := \operatorname{Re}\Gamma_0$ is a version of the corresponding Radon-Nikodym derivative. In the case that a positive integer n and a nondegenerate $q \times q$ Carathéodory sequence $(\Gamma_j)_{j=0}^n$ are given the Riesz-Herglotz measure $F_{c,n}$ of the central $q \times q$ Carathéodory function $\Omega_{c,n}$ corresponding to $(\Gamma_j)_{j=0}^n$ is also absolutely continuous with respect to $\frac{1}{2\pi}\underline{\lambda}$ and the corresponding Radon-Nikodym derivative can be constructed explicitly from the given sequence $(\Gamma_j)_{j=0}^n$ as well (see, e.g., [FK1, Theorem 16 and Remark 18 in Part III]).

Let $n \in \mathbb{N}$, let $(\Gamma_j)_{j=0}^n$ be a nondegenerate $q \times q$ Carathéodory sequence, let $V_n := T_{n-1}^{+}Y_n$, let $W_n := Z_nT_{n-1}^{+}$, and let the $q \times q$ matrix polynomials b_n and d_n be defined by (1.8) and (1.10), respectively. From [FK3, Proposition 2.2

and Theorem 2.3] we know that the functions $\det b_n$ and $\det d_n$ vanish nowhere in $\mathbb{D} \cup \mathbb{T}$. Thus the following proposition is a generalization of the corresponding result for the nondegenerate situation.

Proposition 7.10. *Let $n \in \mathbb{N}$ and let $(\Gamma_j)_{j=0}^n$ be a $q \times q$ Carathéodory sequence. Let $V_n \in \widetilde{\mathcal{Y}}_n$, let $W_n \in \widetilde{\mathcal{Z}}_n$, and let the matrix polynomials b_n and d_n be defined by (1.8) and (1.10). Furthermore, let the matrices L_{n+1} and R_{n+1} be given by (1.6).*

(a) *If $\det b_n(z) \neq 0$ for each $z \in \mathbb{T}$, then the Riesz-Herglotz measure $F_{c,n}$ of the central $q \times q$ Carathéodory function $\Omega_{c,n}$ corresponding to $(\Gamma_j)_{j=0}^n$ is absolutely continuous with respect to $\frac{1}{2\pi}\underline{\lambda}$ and $f : \mathbb{T} \to \mathbb{C}^{q\times q}$ given by*

$$f(z) := \big(b_n(z)\big)^{-*} R_{n+1} \big(b_n(z)\big)^{-1}$$

is the corresponding Radon-Nikodym derivative.

(b) *If $\det d_n(z) \neq 0$ for each $z \in \mathbb{T}$, then $F_{c,n}$ is absolutely continuous with respect to $\frac{1}{2\pi}\underline{\lambda}$ and $g : \mathbb{T} \to \mathbb{C}^{q\times q}$ given by*

$$g(z) := \big(d_n(z)\big)^{-1} L_{n+1} \big(d_n(z)\big)^{-*}$$

is the corresponding Radon-Nikodym derivative.

Proof. (a) Suppose that $\det b_n$ does not vanish in $\mathbb{T}$. Since V_n belongs to $\widetilde{\mathcal{Y}}_n$ and since $\det b_n$ is a polynomial, we get that there is a real number $\rho > 1$ such that $\det b_n$ has no zero in $K(0;\rho)$. Hence $\Omega_{c,n}^{\square} := a_n b_n^{-1}$ is a rational matrix function which is holomorphic in $K(0;\rho)$. Using the matricial version of the H.A. Schwarz Formula, for each $w \in \mathbb{D}$, we get

$$\Omega_{c,n}^{\square}(w) = \frac{1}{2\pi} \int_{\mathbb{T}} \frac{z+w}{z-w} \mathrm{Re}\big(\Omega_{c,n}(z)\big)\, \underline{\lambda}(dz) + i \,\mathrm{Im}\, \Omega_{c,n}^{\square}(0),$$

i.e., in view of (7.2) and $\Omega_{c,n} \in \mathcal{C}_q[\mathbb{D}, (\Gamma_j)_{j=0}^n]$,

$$\Omega_{c,n}(w) = \frac{1}{2\pi} \int_{\mathbb{T}} \frac{z+w}{z-w} \left(\frac{1}{2} \Big(a_n(z) \big(b_n(z)\big)^{-1} + \big(b_n(z)\big)^{-*} a_n^*(z) \Big) \right) \underline{\lambda}(dz) + i \,\mathrm{Im}\, \Gamma_0.$$

Furthermore, from [FKL1, part (a) of Proposition 2.4] we obtain that

$$\mathrm{Re}\,\big(a_n^*(z) b_n(z)\big) = R_{n+1}$$

holds for each $z \in \mathbb{T}$. Thus we can conclude that

$$\Omega_{c,n}(w) = \frac{1}{2\pi} \int_{\mathbb{T}} \frac{z+w}{z-w} \big(b_n(z)\big)^{-*} R_{n+1} \big(b_n(z)\big)^{-1}\, \underline{\lambda}(dz) + i \,\mathrm{Im}\, \Gamma_0 \tag{7.13}$$

is satisfied for each $w \in \mathbb{D}$. Taking into account that the complex $q \times q$ matrix R_{n+1} is nonnegative Hermitian (see, e.g., [DFK, Lemma 1.1.9]) a comparison of (7.12) and (7.13) completes the proof of part (a).

(b) This can be proved analogously to (a). □

In view of Proposition 7.10, in the following result we use the notation $\widehat{\mathcal{Y}}_n$ for the set of all $V_n \in \mathcal{Y}_n$ such that $\det b_n$ vanishes nowhere in $\mathbb{D} \cup \mathbb{T}$ and the notation

$\widehat{\mathcal{Z}}_n$ for the set of all $W_n \in \mathcal{Z}_n$ such that $\det d_n$ vanishes nowhere in $\mathbb{D} \cup \mathbb{T}$, where b_n and d_n are the matrix polynomials defined by (1.8) and (1.10).

Corollary 7.11. *Let $n \in \mathbb{N}$ and let $(\Gamma_j)_{j=0}^n$ be a $q \times q$ Carathéodory sequence. Let $V_n \in \mathcal{Y}_n$, let $W_n \in \mathcal{Z}_n$, and let the $q \times q$ matrix polynomials b_n and d_n be defined by* (1.8) *and* (1.10). *Furthermore, let the matrices L_{n+1} and R_{n+1} be given by* (1.6). *If at least one of the sets $\widehat{\mathcal{Y}}_n$ and $\widehat{\mathcal{Z}}_n$ is nonempty, then the Riesz-Herglotz measure $F_{c,n}$ of the central $q \times q$ Carathéodory function $\Omega_{c,n}$ corresponding to $(\Gamma_j)_{j=0}^n$ is absolutely continuous with respect to $\frac{1}{2\pi}\underline{\lambda}$ and the corresponding Radon-Nikodym derivative $f : \mathbb{T} \to \mathbb{C}^{q\times q}$ satisfies $\underline{\lambda}$-a.e. on $\mathbb{T}$ the identities*

$$f = b_n^{-*} R_{n+1} b_n^{-1} \quad \text{and} \quad f = d_n^{-1} L_{n+1} d_n^{-*}.$$

Proof. An application of Proposition 7.10 in combination with Remark 3.1 and part (a) of Lemma 3.2 yields the assertion. □

Proposition 7.12. *Let $n \in \mathbb{N}$, let $(\Gamma_j)_{j=0}^n$ be a $q \times q$ Carathéodory sequence with nonsingular matrix Γ_0, and let $\Omega_{c,n}$ the central $q \times q$ Carathéodory function corresponding to $(\Gamma_j)_{j=0}^n$. Let $V_n \in \widetilde{\mathcal{Y}}_n$, let $W_n \in \widetilde{\mathcal{Z}}_n$, and let the matrix polynomials a_n and c_n be defined by* (1.7) *and* (1.9). *Furthermore, let the matrices L_{n+1} and R_{n+1} be given by* (1.6).

(a) *If $\det a_n(z) \neq 0$ for each $z \in \mathbb{T}$, then the Riesz-Herglotz measure $F_{c,n}^\sharp$ of $\Omega_{c,n}^{-1}$ is absolutely continuous with respect to $\frac{1}{2\pi}\underline{\lambda}$ and $f^\sharp : \mathbb{T} \to \mathbb{C}^{q\times q}$ given by*

$$f^\sharp(z) := \big(a_n(z)\big)^{-*} R_{n+1} \big(a_n(z)\big)^{-1}$$

is the corresponding Radon-Nikodym derivative.

(b) *If $\det c_n(z) \neq 0$ for each $z \in \mathbb{T}$, then $F_{c,n}^\sharp$ is absolutely continuous with respect to $\frac{1}{2\pi}\underline{\lambda}$ and $g : \mathbb{T} \to \mathbb{C}^{q\times q}$ given by*

$$g^\sharp(z) := \big(c_n(z)\big)^{-1} L_{n+1} \big(c_n(z)\big)^{-*}$$

is the corresponding Radon-Nikodym derivative.

Proof. Based on Remark 3.7, the matricial version of the H.A. Schwarz Formula, (7.2), and [FKL1, part (a) of Proposition 2.4] one can prove Proposition 7.12 analogously to Proposition 7.10. We omit the details. □

In view of Proposition 7.12 and Remark 3.7 we use in the following result the notation $\check{\mathcal{Y}}_n$ for the set of all $V_n \in \mathcal{Y}_n$ such that $\det a_n$ vanishes nowhere in $\mathbb{D} \cup \mathbb{T}$ and the notation $\check{\mathcal{Z}}_n$ for the set of all $W_n \in \mathcal{Z}_n$ such that $\det c_n$ vanishes nowhere in $\mathbb{D} \cup \mathbb{T}$, where a_n and c_n are the matrix polynomials defined by (1.7) and (1.9).

Corollary 7.13. *Let $n \in \mathbb{N}$, let $(\Gamma_j)_{j=0}^n$ be a $q \times q$ Carathéodory sequence with nonsingular matrix Γ_0, and let the matrices L_{n+1} and R_{n+1} be given by* (1.6). *Furthermore, let $V_n \in \mathcal{Y}_n$, let $W_n \in \mathcal{Z}_n$, and let the $q \times q$ matrix polynomials a_n and c_n be defined by* (1.7) *and* (1.9). *If at least one of the sets $\check{\mathcal{Y}}_n$ and $\check{\mathcal{Z}}_n$ is nonempty, then the Riesz-Herglotz measure $F_{c,n}^\sharp$ of $\Omega_{c,n}^{-1}$ is absolutely continuous with respect*

to $\frac{1}{2\pi}\underline{\lambda}$ *and the corresponding Radon-Nikodym derivative* $f^\sharp : \mathbb{T} \to \mathbb{C}^{q\times q}$ *satisfies* $\underline{\lambda}$*-a.e. on* $\mathbb{T}$ *the identities*

$$f^\sharp = a_n^{-*} R_{n+1} a_n^{-1} \quad \text{and} \quad f^\sharp = c_n^{-1} L_{n+1} c_n^{-*}.$$

Proof. An application of Proposition 7.12 in combination with Remark 3.1 and part (b) of Lemma 3.2 yields the assertion. □

References

[Akh] Akhiezer, N.I.: *The Classical Moment Problem* (Russian), Fizmatgiz, Moskva 1961; English translation: Oliver and Boyd, Edinburgh-London 1965.

[AK] Arov, D.Z.; Kreĭn, M.G.: *Problems of the search of the minimum of entropy in indeterminate extension problems* (Russian), Funkcional. Anal. Prilozen. **15** (1981), No. 2, 61–64; English translation: Func. Anal. Appl. **15** (1981), 123–126.

[BGR] Ball, J.A.; Gohberg, I.; Rodman, L.: *Interpolation of Rational Matrix Functions*, Operator Theory: Adv. Appl. 45, Birkhäuser, Basel 1990.

[BH] Ball, J.A.; Helton, J.W.: *Interpolation problems of Pick-Nevanlinna and Loewner types for meromorphic matrix functions: parametrization of the set of all solutions*, Integral Equations Operator Theory **9** (1986), 155–203.

[BD] Bolotnikov, V.; Dym, H.: *On degenerate interpolation, entropy and extremal problems for matrix Schur functions*, Integral Equations Operator Theory **32** (1998), 367–435.

[Br] Bruinsma, P.: *Degenerate interpolation problems for Nevanlinna pairs*, Indag. Math., N.S. **2** (1991), 179–200.

[CG] Chang, C.; Georgiou, T.: *Geometric aspects of the Carathéodory extension problem*, Linear Algebra Appl. **203/204** (1994), 209–251.

[CH1] Chen, G.N.; Hu, Y.J.: *The truncated Hamburger matrix moment problems in the nondegenerate and degenerate cases, and matrix continued fractions*, Linear Algebra Appl. **277** (1998), 199–236.

[CH2] Chen, G.N.; Hu, Y.J.: *On the multiple Nevanlinna-Pick matrix interpolation in the class* C_p *and the Carathéodory matrix coefficient problem*, Linear Algebra Appl. **283** (1998), 179–203.

[DGK1] Delsarte, P.; Genin, Y.; Kamp, Y.: *Orthogonal polynomial matrices on the unit circle*, IEEE Trans. Circuits and Systems CAS-**25** (1978), 149–160.

[DGK2] Delsarte, P.; Genin, Y.; Kamp, Y.: *The Nevanlinna-Pick problem for matrix-valued functions*, SIAM J. Appl. Math. **36** (1979), 47–61.

[Du] Dubovoj, V.K.: *Indefinite metric in the interpolation problem of Schur for analytic matrix functions IV* (Russian), Teor. Funktsii, Funkts. Anal. i Prilozen. **42** (1984), 46–57; English translation in: Topics in Interpolation Theory (Eds.: H. Dym, B. Fritzsche, V.E. Katsnelson, B. Kirstein), Operator Theory: Adv. Appl. 95, Birkhäuser, Basel 1997, pp. 93–104.

[DFK] Dubovoj, V.K.; Fritzsche, B.; Kirstein, B.: *Matricial Version of the Classical Schur Problem*, Teubner-Texte zur Mathematik 129, B.G. Teubner, Stuttgart-Leipzig 1992.

[Dy] Dym, H.: *J Contractive Matrix Functions, Reproducing Kernel Spaces and Interpolation,* CBMS Regional Conference Series in Mathematics 71, Amer. Math. Soc., Providence, R.I. 1989.

[Ev] Everitt, W.N.: *A personal history of the m-coefficient,* J. Comp. Appl. Math. **171** (2004), 185–197.

[FF] Foiaş, C.; Frazho, A.E.: *The Commutant Lifting Approach to Interpolation Problems*, Operator Theory: Adv. Appl. 44, Birkhäuser, Basel 1990.

[FFGK] Foiaş, C.; Frazho, A.E.; Gohberg, I.; Kaashoek, M.A.: *Metric Constrained Interpolation, Commutant Lifting and Systems*, Operator Theory: Adv. Appl. 100, Birkhäuser, Basel 1998.

[FFK] Fritzsche, B.; Fuchs, S.; Kirstein, B.: *A Schur type matrix extension problem*, Part V, Math. Nachr. **158** (1992), 133–159.

[FK1] Fritzsche, B.; Kirstein, B.: *An extension problem for non-negative hermitian block Toeplitz matrices*, Math. Nachr., Part I: **130** (1987), 121–135; Part II: **131** (1987), 287–297; Part III: **135** (1988), 319–341; Part IV: **143** (1989), 329–354; Part V: **144** (1989), 283–308.

[FK2] Fritzsche, B.; Kirstein, B.: *On the Weyl matrix balls associated with nondegenerate matrix-valued Carathéodory functions,* Zeitschr. für Analysis und ihre Anwendungen **12** (1993), 239–261.

[FK3] Fritzsche, B.; Kirstein, B.: *Representations of central matrix-valued Carathéodory functions in both nondegenerate and degenerate cases*, Integral Equations Operator Theory **50** (2004), 333–361.

[FKL1] Fritzsche, B.; Kirstein, B.; Lasarow, A.: *The matricial Carathéodory problem in both nondegenerate and degenerate case*, in: Operator Theory: Adv. Appl. (Subseries: Linear Operators and Linear Systems) Vol. 165 (Eds.: D. Alpay, I. Gohberg), Birkhäuser, Basel 2006, pp. 251–290.

[FKL2] Fritzsche, B.; Kirstein, B.; Lasarow, A.: *On a class of extremal solutions of the nondegenerate matricial Carathéodory problem*, Analysis **27** (2007), 109–164.

[Ge1] Geronimus, Ja.L.: *On polynomials orthogonal on the unit circle, on the trigonometric moment problem and on associated functions of classes of Carathéodory-Schur* (Russian), Mat. USSR-Sb. **15** (1944), 99–130.

[Ge2] Geronimus, Ja.L.: *On the trigonometric moment problem*, Ann. Math. **47** (1946), 742–761.

[Ge3] Geronimus, Ja.L.: *Polynomials orthogonal on a circle and their applications* (Russian), Zapiski Naučno-Issled. Mat. Meh. Har'kov. Mat. Obšč. **19** (1948), 35–120; English translation in: AMS Translations, Series 1, Volume 3, AMS, Providence, R.I. 1962, pp. 1–78.

[GH] Gohberg, I.Ts.; Heinig, G.: *Inversion of finite Toeplitz matrices consisting of elements of a non-commutative algebra* (Russian), Rev. Roum. Math. Pures et. Appl. **19** (1974), 623–663.

[He] Hellinger, E.: *Zur Stieltjesschen Kettenbruchtheorie*, Math. Ann. **86** (1922), 18–29.

[Ko] Kovalishina, I.V.: *Analytic theory of a class of interpolation problems* (Russian), Izv. Akad. Nauk SSSR, Ser. Mat. **47** (1983), 455–497; English translation: Math. USSR Izvestija **22** (1984), 419–463.

[KP] Kovalishina, I.V.; Potapov, V.P.: *The radii of a Weyl disk in the matricial Nevanlinna-Pick problem* (Russian), in: Operator Theory in Function Spaces and Their Application (Ed.: V.A. Marčenko), Naukova Dumka, Kiev 1981, pp. 25–49; English translation in: Seven Papers Translated from the Russian, AMS Translations, Series 2, Volume 138, AMS, Providence, R.I. 1988, pp. 37–54.

[Mi] Mikhailova, I.V.: *Weyl matrix circles as a tool for uniqueness in the theory of multiplicative representations of J-inner matrix functions* (Russian), in: Analysis in Infinite-dimensional Spaces and Operator Theory (Ed.: V.A. Marčenko), Naukova Dumka, Kiev 1983, pp. 101–117; English translation in: Topics in Interpolation Theory (Eds.: H. Dym, B. Fritzsche, V.E. Katsnelson, B. Kirstein), Operator Theory: Adv. Appl. 95, Birkhäuser, Basel 1997, pp. 397–417.

[Ne] Nevanlinna, R.: *Asymptotische Entwicklungen beschränkter Funktionen und das Stieltjessche Momentenproblem*, Ann. Acad. Sci. Fenn. **A18** (1922), 5, 1–53.

[Or] Orlov, S.A.: *Nested matrix balls, analytically depending on a parameter, and theorems on invariance of ranks of radii of limit matrix balls* (Russian), Izv. Akad. Nauk SSSR, Ser. Mat. **40** (1976), 593–644; English translation: Math. USSR Izvestija **10** (1976), 565–613.

[Sa] Sakhnovich, L.A.: *Interpolation Theory and its Applications*, Mathematics and its Applications 428, Kluwer, Dordrecht 1997.

[Sm] Šmuljan, Ju.L.: *Operator balls* (Russian), Teor. Funkcii, Funktional. Anal. i Prilozen. **6** (1968), 68–81; reprinted in: Integral Equations and Operator Theory **13** (1990), 864–882.

[We] Weyl, H.: *Über gewöhnliche Differentialgleichungen mit Singularitäten und die zugehörigen Entwicklungen willkürlicher Funktionen,* Math. Ann. **68** (1910), 220–269.

Bernd Fritzsche and Bernd Kirstein
Fakultät für Mathematik und Informatik
Universität Leipzig
Postfach: 10 09 20
D-04009 Leipzig, Germany
e-mail: fritzsche@mathematik.uni-leipzig.de
kirstein@mathematik.uni-leipzig.de

Andreas Lasarow
Departement Computerwetenschappen
Katholieke Universiteit Leuven
Celestijnenlaan 200A – postbus: 02402
B-3001 Leuven, Belgium
e-mail: Andreas.Lasarow@cs.kuleuven.be

Operator Theory:
Advances and Applications, Vol. 199, 333–346

On Extremal Problems of Interpolation Theory with Unique Solution

Bernd Fritzsche, Bernd Kirstein and Lev A. Sakhnovich

Dedicated to the memory of Georg Heinig

Abstract. The main goal of this paper is to investigate the matrix extremal interpolation problem formulated in Chapter 7 of the monograph [7]. We give natural conditions under which the problem has one and only one solution. The basic idea of the proof is to use the matrix Riccati equation deduced in [7, Chapter 7].

Mathematics Subject Classification (2000). Primary: 47A57.

Keywords. Matricial interpolation problems, minimal solutions, nonlinear matrix equation, regular interpolation problems.

0. Introduction

In this paper we consider a particular matrix extremal interpolation problem. More precisely, we try to find amongst the solutions $w(z)$ of the corresponding interpolation problem the solution which satisfies the additional extremal condition

$$w^*(z)w(z) \leq \rho_{\min}^2, \quad |z| < 1, \tag{0.1}$$

where $\rho_{\min}$ is a positive Hermitian $m \times m$ matrix. What concerns the statement of the problem we follow the book [7, Chapter 7] where some Riccati type equation for the matrix $\rho_{\min}$ was deduced. It was proved by Ran and Reurings (see [6]) that for the case of the Schur extremal interpolation problem this equation has one and only one solution $\rho_{\min}$.

The main result of this paper is to show that the above-mentioned result of Ran and Reurings can be extended to a broad class of interpolation problems. In particular, it is true for the Nevanlinna-Pick problem. In this paper we will develop a method for the computation of $\rho_{\min}$. We illustrate our general result with some concrete examples (Schur problem, Nevanlinna-Pick problem, Jordan blocks).

Remark 0.1. A well-posed physical problem should have one and only one solution. The extremal interpolation problem under consideration satisfies this requirement.

Remark 0.2. It is essential both from the applied and theoretical points of view that the solution of the extremal problem turns out to be a rational matrix function (see [7, Chapter 7], [4]).

Remark 0.3. It is possible that in the classical case

$$w^*(z)w(z) \leq I$$

the interpolation problem has not any solution whereas problem (0.1) has a solution.

Remark 0.4. The scalar case ($m = 1$) of the extremal interpolation problem was studied by N.I. Akhiezer [1]. It found its application in control theory (see Kimura [5]). The transition to the matrix case allows to enlarge considerably the class of extremal problems which have effective solutions.

1. Extremal interpolation problems

Let the matrices A, S_k and $\Psi_k, k = 1, 2$, have the sizes $mN \times mN$ and $mN \times m$, respectively, where S_k is nonnegative Hermitian. We suppose that these matrices are connected by the relations

$$S_k - AS_kA^* = \Psi_k\Psi_k^*, \quad k = 1, 2. \tag{1.1}$$

Setting

$$S := S_2 - S_1$$

we deduce from (1.1) the equality

$$S - ASA^* = \Psi_2\Psi_2^* - \Psi_1\Psi_1^*. \tag{1.2}$$

We introduce the block-diagonal matrix

$$R := \operatorname{diag}\ \underbrace{(\rho, \ldots, \rho)}_{N}$$

where ρ is a positive Hermitian matrix of size $m \times m$. In addition we shall assume the equality

$$AR = RA. \tag{1.3}$$

This is justified, since it was shown in [4] that condition (1.3) is true in a number of concrete examples.
From equations (1.1) and (1.3) it follows that

$$S_\rho - AS_\rho A^* = \Psi_2\Psi_2^* - \Psi_{1,\rho}\Psi_{1,\rho}^*. \tag{1.4}$$

where

$$S_\rho := S_2 - R^{-1}S_1R^{-1} \tag{1.5}$$

$$\Psi_{1,\rho} := R^{-1}\Psi_1. \tag{1.6}$$

Thus we have constructed a set of operator identities (1.5), where the positive Hermitian matrix ρ plays the role of a parameter. A set of interpolation problems, see

[7, Chapter 6] corresponds to this set of operator identities. A necessary condition for the solvability of these problems is the inequality

$$RS_2R - S_1 \geq 0. \tag{1.7}$$

Now we turn to extremal interpolation.

Definition 1.1. We shall call the matrix $\rho = \rho_{\min} > 0$ a minimal solution of inequality (1.7) if the following two requirements are fulfilled:

1. The inequality

$$R_{\min}S_2R_{\min} - S_1 \geq 0 \tag{1.8}$$

holds where

$$R_{\min} = \operatorname{diag} \underbrace{(\rho_{\min}, \ldots, \rho_{\min})}_{N}$$

is valid.

2. If $\rho > 0$ satisfies inequality (1.7), then

$$\operatorname{rank}\,(R_{\min}S_2R_{\min} - S_1) \leq \operatorname{rank}\,(RS_2R - S_1). \tag{1.9}$$

(In other words, $R_{\min}$ minimizes the rank of $RS_2R - S_1 \geq 0$.)

Remark 1.2. The existence of $\rho_{\min}$ follows directly from Definition 1.1

We shall write the nonnegative Hermitian matrices S_1, S_2 and R in the following block forms

$$S_k = \begin{pmatrix} S_{11}^{(k)} & S_{12}^{(k)} \\ S_{21}^{(k)} & S_{22}^{(k)} \end{pmatrix}, \quad k = 1, 2 \tag{1.10}$$

$$R = \begin{pmatrix} R_1 & 0 \\ 0 & \rho \end{pmatrix}, \quad R_1 = \operatorname{diag} \underbrace{(\rho, \ldots, \rho)}_{N-1} \tag{1.11}$$

where $S_{22}^{(k)}$ are blocks of size $m \times m$, $S_{11}^{(k)}$ has the size $(N-1)m \times (N-1)m$ and $S_{12}^{(k)}$ has the size $(N-1)m \times m$. The following result is proved in [7, Proposition 7.1.1].

Proposition 1.3. *Suppose that for all $\rho > 0$ satisfying inequality* (1.7) *the upper diagonal block is positive Hermitian, i.e., that*

$$R_1S_{11}^{(2)}R_1 - S_{11}^{(1)} > 0$$

holds. If $\rho = q > 0$ satisfies inequality (1.7) *and the relation*

$$qS_{22}^{(2)}q = S_{22}^{(1)} + C_1^* \left(Q_1S_{11}^{(2)}Q_1 - S_{11}^{(1)}\right)^{-1} C_1 \tag{1.12}$$

where

$$Q_1 := \operatorname{diag} \underbrace{(q, q, \ldots, q)}_{N-1}, \qquad C_1 := Q_1S_{12}^{(2)}q - S_{12}^{(1)}, \tag{1.13}$$

then

$$\rho_{\min} = q.$$

2. On a nonlinear matrix equation

In this section we consider the equation

$$qS_{22}^{(2)}q = S_{22}^{(1)} + S_{12}^{*}\left(Q_1 S_{11}^{(2)} Q_1 - S_{11}^{(1)}\right)^{-1} S_{12} \tag{2.1}$$

where

$$S_{12} := Q_1 S_{12}^{(2)} q - S_{12}^{(1)}. \tag{2.2}$$

We make the following two assumptions.

Condition 1. The matrix S_2 has the block structure

$$S_2 = (C_{jk})_{j,k=1}^{n}$$

where all $m \times m$ blocks C_{jk} have the shape

$$C_{jk} = \alpha_{jk} I_m$$

with some complex number α_{jk}.

Condition 2. The matrix S_2 is positive Hermitian ($S_2 > 0$).

In view of Condition 1 we obtain

$$\begin{aligned} S_{12} &= Q_1^2 S_{12}^{(2)} - S_{12}^{(1)} \\ &= Q_1^2 S_{11}^{(2)} \left(S_{11}^{(2)}\right)^{-1} S_{12}^{(2)} - S_{12}^{(1)} \\ &= \left[Q_1^2 S_{11}^{(2)} - S_{11}^{(1)}\right] \left(S_{11}^{(2)}\right)^{-1} S_{12}^{(2)} + S_{11}^{(1)} \left(S_{11}^{(2)}\right)^{-1} S_{12}^{(2)} - S_{21}^{(1)} \end{aligned}$$

and

$$Q_1 S_{11}^{(2)} Q_1 = Q_1^2 S_{11}^{(2)}.$$

We introduce the following notations

$$A := Q_1^2 S_{11}^{(2)} - S_{11}^{(1)}, \quad B := \left(S_{11}^{(2)}\right)^{-1} S_{12}^{(2)}, \tag{2.3}$$

and

$$C := S_{11}^{(1)} \left(S_{11}^{(2)}\right)^{-1} S_{12}^{(2)} - S_{12}^{(1)}. \tag{2.4}$$

Then obviously $A^* = A$ and the equation (2.1) can be written in the form

$$\alpha_{nn} q^2 = S_{22}^{(1)} + (B^* A + C^*) A^{-1} (AB + C)$$

or

$$\alpha_{nn} q^2 = S_{22}^{(1)} + B^* AB + B^* C + C^* B + C^* A^{-1} C. \tag{2.5}$$

Using (2.3) and (2.5) we infer

$$q^2 T = U + C^* A^{-1} C \tag{2.6}$$

where

$$T := \alpha_{nn} I_m - \left(S_{12}^{(2)}\right)^{*} \left(S_{11}^{(2)}\right)^{-1} S_{12}^{(2)} \tag{2.7}$$

and

$$\begin{aligned} U &:= B^*C + C^*B - B^*S_{11}^{(1)}B + S_{22}^{(1)} \\ &= \left(S_{12}^{(2)}\right)^* \left(S_{11}^{(2)}\right)^{-1} S_{11}^{(1)} \left(S_{11}^{(2)}\right)^{-1} S_{12}^{(2)} - \left(S_{12}^{(2)}\right)^* \left(S_{11}^{(2)}\right)^{-1} S_{12}^{(1)} \\ &\quad - \left(S_{12}^{(1)}\right)^* \left(S_{11}^{(2)}\right)^{-1} S_{12}^{(2)} + S_{22}^{(1)} \end{aligned} \tag{2.8}$$

In view of Condition 2 the relation

$$T > 0 \tag{2.9}$$

is true.

According to Condition 1 and (2.7) the matrix T has scalar type, i.e.,

$$T = \beta I_m. \tag{2.10}$$

Because of (2.9) and (2.10) it follows then

$$\beta > 0.$$

Hence the equation (2.6) takes the form

$$q^2 = \frac{1}{\beta}(U + C^*A^{-1}C). \tag{2.11}$$

If we compare the equations (2.1) and (2.11) we see that S_{12} depends on q, but C does not depend on q. Taking into account this fact we can apply Theorem 3.3 of the paper [6] to equation (2.11). Moreover, we observe that the matrix A can be represented in the form

$$A = DQ_1^2D - S_{11}^{(1)}$$

where

$$D := \sqrt{S_{11}^{(2)}} > 0.$$

Now we rewrite the equation (2.11) in the form

$$q^2 = \frac{1}{\beta}\left[U + C_1^*\left(Q_1^2 - D^{-1}S_{11}^{(1)}D^{-1}\right)^{-1}C_1\right] \tag{2.12}$$

where

$$C_1 := D^{-1}C.$$

We introduce the notation

$$\tilde{U} := \operatorname{diag}\underbrace{(U, U, \ldots, U)}_{n-1}.$$

Definition 2.1. We call an interpolation problem regular if the condition

$$\frac{1}{\beta}\tilde{U} > D^{-1}S_{11}^{(1)}D^{-1}$$

is satisfied.

Using Theorem 3.3 in [6] now we obtain our main result.

Theorem 2.2. *Let the conditions* 1 *and* 2 *be fulfilled and let the interpolation problem be regular. If* $S_{11}^{(1)} \geq 0$ *then equation* (2.1) *has a unique solution* q *such that* $q > 0$ *and* $Q_1 S_{11}^{(2)} Q_1 > S_{11}^{(1)}$.

Corollary 2.3. *Under the assumptions of Theorem* 2.2 *the relation* $\rho_{\min}^2 = q^2$ *holds.*

Remark 2.4. Under the assumption that $\rho_{\min}$ is known an explicit representation of the solution of the corresponding extremal interpolation problem is given in monograph [7, Chapter 7]. In the special case of the extremal interpolation problems named after Schur and Nevanlinna-Pick, respectively, this solution is written in a simpler form in the paper [4].

We have shown that there is one and only one positive Hermitian solution of equation (2.1) which satisfies condition (1.7). In the case $N = 2$ Ferrante/Levy [3] proved without taking into account condition (1.7) that equation (2.1) has one and only one positive Hermitian solution.

3. Methods of computation

1. We apply the method of successive approximation to the study of equation (2.1). We set

$$q_0^2 := \frac{1}{\beta} U$$

and for $p \in \{0, 1, 2, \ldots\}$ then

$$q_{p+1}^2 := \frac{1}{\beta} \left[U + C_1^* (Q_p^2 - D^{-1} S_{11}^{-1} D^{-1}) C_1 \right],$$

where

$$Q_p := \operatorname{diag} \underbrace{(q_p, q_p, \ldots, q_p)}_{N-1}.$$

The following assertion is proved in [7, Lemma 7.1.1].

Lemma 3.1. *Let the conditions of Theorem* 2.2 *be fulfilled. Then:*

a) *The sequence* $q_0^2, q_2^2, q_4^2, \ldots$ *is monotonously increasing and has the limit* q^2.
b) *The sequence* $q_1^2, q_3^2, q_5^2, \ldots$ *is monotonously decreasing and has the limit* $\overline{q}^2$.

The combination of Theorem 2.2 and Lemma 3.1 yields the following result.

Corollary 3.2. *The relations*

$$\underline{q}^2 = \overline{q}^2 = \rho_{\min}^2$$

and

$$q_{2p}^2 \leq \rho_{\min}^2 \leq q_{2p+1}^2, \; p \in \{0, 1, 2, \ldots\},$$

hold.

2. Now we introduce the matrix

$$S_{\min} := R_{\min} S_2 R_{\min} - S_1. \tag{3.1}$$

Setting

$$R_1 := \operatorname{diag} \underbrace{(\rho_{\min}, \ldots, \rho_{\min})}_{N-1}. \tag{3.2}$$

and, using the block partitions (1.10), moreover

$$S_{12} := R_1 S_{12}^{(2)} \rho_{\min} - S_{12}^{(1)} \tag{3.3}$$

from (3.1), (3.2), and (3.3) we obtain the block decomposition

$$S_{\min} = \begin{pmatrix} R_1 S_{11}^{(2)} R_1 - S_{11}^{(1)} & S_{12} \\ S_{12}^* & \rho_{\min} S_{22}^{(2)} \rho_{\min} - S_{22}^{(1)} \end{pmatrix}. \tag{3.4}$$

Taking into account (2.1), (2.2), (3.2), and (3.3) from Theorem 2.2 we get

$$R_1 S_{11}^{(2)} R_1 S_{11}^{(1)} > 0 \tag{3.5}$$

and

$$\rho_{\min} S_{22}^{(2)} \rho_{\min} = S_{22}^{(1)} + S_{12}^* (R_1 S_{11}^{(2)} R_1 - S_{11}^{(1)})^{-1} S_{12}. \tag{3.6}$$

In view of $S_{\min} \geq 0$, (3.4), (3.5) and (3.6) we infer

$$\operatorname{rank} S_{\min} = (N-1)m. \tag{3.7}$$

Defining the sequence $(Y_j)_{j=1}^n$ of complex $m \times m$ matrices via

$$\begin{pmatrix} Y_1 \\ Y_2 \\ \vdots \\ Y_{N-1} \end{pmatrix} := -(R_1 S_{11}^{(2)} R_1 - S_{11}^{(1)})^{-1} S_{12} \tag{3.8}$$

and

$$Y_N := I_m \tag{3.9}$$

we obtain from (3.4) and (3.6) the relation

$$S_{\min} Y = 0 \tag{3.10}$$

where

$$Y := \operatorname{col}(Y_j)_{j=1}^N. \tag{3.11}$$

3. Let us calculate the number β introduced in (2.10). According to Conditions 1 and 2 the matrix $S_{22}^{(2)}$ is nonsingular and the matrix $(S_{22}^{(2)})^{-1}$ has the block structure

$$(S_{22}^{(2)})^{-1} = (D_{jk})_{j,k=1}^{N-1} \tag{3.12}$$

where all $m \times m$ blocks D_{jk} have the shape

$$D_{jk} = \gamma_{jk} I_m \tag{3.13}$$

with some complex numbers γ_{jk}. Condition 1 implies moreover that

$$S_{12}^{(2)} = \operatorname{col}\left((\alpha_{jN} I_m)_{j=1}^{N-1}\right). \tag{3.14}$$

Combining (2.10), (2.7), (3.12), (3.13), and (3.14) we obtain

$$\beta = \alpha_{NN} - \alpha^* \Gamma \alpha$$

where

$$\alpha := \operatorname{col}(\alpha_{jN})_{j=1}^{N-1}$$

and

$$\Gamma := (\gamma_{jk})_{j,k=1}^{N-1}.$$

4. Schur extremal problem

Let us consider the following version of the Schur problem.

Problem 4.1. Let the complex $m \times m$ matrices $a_0, a_1, \dots, a_p$ and the positive Hermitian $m \times m$ matrix ρ be given. We wish to describe the set of $m \times m$ matrix functions $w(z)$ holomorphic in the circle $|z| < 1$ satisfying

$$w(z) = a_0 + a_1 z + \cdots + a_p z^p + \cdots \tag{4.1}$$

and

$$w^*(z) w(z) \le \rho^2, \quad |z| < 1. \tag{4.2}$$

The relations (4.1) and (4.2) can be written in the form

$$w_1(z) = b_0 + b_1 z + \cdots + b_p z^p + \cdots, \tag{4.3}$$

$$w_1^*(z) w_1(z) \le I_m, \quad |z| < 1, \tag{4.4}$$

where

$$w_1^*(z) := w(z) \cdot \rho^{-1} \tag{4.5}$$

$$b_0 := a_0 \rho^{-1}, \dots, b_p := a_p \rho^{-1}. \tag{4.6}$$

Thus, the modified Problem 4.1 is reduced to the matrix version of the classical Schur problem.

Using [7, Proposition 7.2.1] we deduce the following assertion.

Proposition 4.2. *Let the matrix C_p be defined by*

$$C_p := \begin{pmatrix} a_0 & a_1 & \dots & a_p \\ 0 & a_0 & \dots & a_{p-1} \\ \vdots & \vdots & & \vdots \\ 0 & 0 & \dots & a_0 \end{pmatrix}. \tag{4.7}$$

Then Problem 4.1 *has a solution if and only if the inequality*

$$C_p^* C_p \le R^2 \tag{4.8}$$

is valid, where

$$R := \operatorname{diag} \underbrace{(\rho, \rho, \dots, \rho)}_{p+1}. \tag{4.9}$$

Remark 4.3. If the strict inequality

$$C_p^* C_p < R^2 \tag{4.10}$$

holds then the set of solutions of the modified Schur problem 4.1 can be written with the aid of a linear fractional transformation (see [7, Chapter 6]).

Let us formulate now the extremal Schur problem.

Problem 4.4. Let the complex $m \times m$ matrices $a_0, a_1, \ldots, a_p$ be given. We seek an $m \times m$ matrix-valued function $w(z)$, holomorphic in the circle $|z| < 1$, satisfying

$$w(z) = a_0 + a_1 z + \cdots + a_p z^p + \cdots \tag{4.11}$$

and

$$w^*(z) w(z) \leq \rho_{\min}^2, \quad |z| < 1. \tag{4.12}$$

Here $\rho_{\min}$ will be defined by a minimal rank condition. More precisely, we indicate that the Schur extremal problem fits into the general scheme of interpolation problems studied in this paper and seek a minimal rank solution in the sense of Definition 1.1.

We present now the operator reformulation of the Schur problem (see [7, Chapter 7]). It is well known that in this case

$$S_2 = I_{(p+1)m}, \quad S_1 = C_p^* C_p \tag{4.13}$$

where C_p is given by (4.7). Moreover, the matrix A has in the case of the Schur problem the form

$$A = \underbrace{\begin{pmatrix} 0 & 0 & \ldots & 0 & 0 \\ I_m & 0 & \ldots & 0 & 0 \\ 0 & I_m & \ldots & 0 & 0 \\ \vdots & \vdots & & \vdots & \vdots \\ 0 & 0 & \ldots & I_m & 0 \end{pmatrix}}_{p+1}, \tag{4.14}$$

whereas the matrices Ψ_1 and Ψ_2 are defined by formulas

$$\Psi_1 = \operatorname{col}\left[a_0^*, a_1^*, \ldots, a_p^*\right] \tag{4.15}$$

and

$$\Psi_2 = \operatorname{col}\left[I_m, 0, \ldots, 0\right]. \tag{4.16}$$

Taking into account formulas (4.13)–(4.16) it is immediately checked that

$$S_k - A S_k A^* = \Psi_k \Psi_k^*, \quad k \in \{1, 2\}. \tag{4.17}$$

Thus, in view of (4.17) the results obtained above can be applied to the Schur extremal problem.

In this case we have

$$T = I_m, \qquad \beta = 1,$$
$$S_{12}^{(2)} = 0, \qquad S_{12} = -S_{12}^{(1)}, \qquad S_{11}^{(1)} = C_{p-1}^* C_{p-1},$$
$$U = S_{22}^{(1)} = a_0^* a_0 + a_1^* a_1 + \cdots + a_p^* a_p.$$

The regularity condition from Definition 2.1 has now the form

$$\operatorname{diag} \underbrace{(U, U, \ldots, U)}_{p} > S_{11}^{(1)} \tag{4.18}$$

Proposition 4.5. *If the condition (4.18) is fulfilled, then*

$$a_p^* a_p > 0.$$

The converse statement is true in the case $p = 1$.

Proposition 4.6. *If $p = 1$ and $a_1^* a_1 > 0$ then*

$$U = a_0^* a_0 + a_1^* a_1 > a_0^* a_0 = S_{11}^{(1)}.$$

5. Nevanlinna-Pick extremal problem

Let us consider the following version of the Nevanlinna-Pick problem.

Problem 5.1. Let the complex $m \times m$ matrices $\eta_1, \eta_2, \ldots, \eta_n$, the points $z_1, \ldots, z_n$ satisfying $|z_k| < 1$, $k \in \{1, \ldots, n\}$ and the positive Hermitian $m \times m$ matrix ρ be given. We wish to describe the set of $m \times m$ matrix functions $w(z)$ holomorphic in the circle $|z| < 1$ such that

$$w(-\overline{z_k}) = \eta_k^*, \ k \in \{1, \ldots, n\}, \tag{5.1}$$

and

$$w^*(z) \cdot w(z) \le \rho^2, \quad |z| < 1. \tag{5.2}$$

The relations (5.1) and (5.2) can be written in the form

$$w_1(-\overline{z_k}) = \tilde{\eta}_k, \ k \in \{1, \ldots, n\}, \tag{5.3}$$

$$w_1^*(z) \cdot w_1(z) \le I_m, \quad |z| < 1, \tag{5.4}$$

where

$$\tilde{\eta}_k := \eta_k \cdot \rho^{-1}, \ k \in \{1, \ldots, n\}, \tag{5.5}$$

$$w_1(z) := w(z) \cdot \rho^{-1}. \tag{5.6}$$

Using [7, Proposition 7.3.1] we deduce the following assertion.

Proposition 5.2. *Let the matrices S_1 and S_2 be defined by*

$$S_1 := \left(\frac{\eta_k \eta_l^*}{1 - z_k \overline{z}_l} \right)_{k,l=1}^n, \qquad S_2 := \left(\frac{I_m}{1 - z_k \overline{z}_l} \right)_{k,l=1}^n. \tag{5.7}$$

Then Problem 5.1 *has a solution if and only if the inequality*

$$S_2 - R^{-1}S_1R^{-1} \geq 0 \tag{5.8}$$

is valid, where

$$R := \operatorname{diag}\underbrace{(\rho, \ldots, \rho)}_{p}. \tag{5.9}$$

Remark 5.3. If the strict inequality

$$S_2 - R^{-1}S_1R^{-1} > 0$$

holds, then the set of solutions of the modified Nevanlinna-Pick Problem 5.1 can be written with the aid of a linear fractional transformation (see [7, Chapter 6]).

Let us formulate now the Nevanlinna-Pick extremal problem.

Problem 5.4. Let complex $m \times m$ matrices $\eta_1, \eta_2, \ldots, \eta_n$ and points $z_1, z_2, \ldots, z_n$ satisfying $|z_k| < 1,\ k \in \{1, \ldots, n\}$, be given. We seek an $m \times m$ matrix-valued function $w(z)$ which is holomorphic in the circle $|z| < 1$, such that

$$w(-\overline{z_k}) = \eta_k^*,\ k \in \{1, \ldots, n\}, \tag{5.10}$$

and

$$w^*(z)w(z) \leq \rho_{\min}^2, \quad |z| < 1. \tag{5.11}$$

Here $\rho_{\min}$ will be defined by a minimal rank condition. More precisely, we indicate that the Nevanlinna-Pick extremal problem fits into the general scheme of interpolation problems studied in this paper and seek a minimal rank solution in the sense of Definition 1.1.

We present now the operator reformulation of the Nevanlinna-Pick problem (see [7, Chapter 7]). In this case, the matrices S_1 and S_2 are given by (5.7) whereas the matrices A, Ψ_1, and Ψ_2 are defined by

$$A := \operatorname{diag}(z_1 I_m, z_2 I_m, \ldots, z_n I_m) \tag{5.12}$$

$$\Psi_1 := \operatorname{col}(\eta_1, \ldots, \eta_m) \tag{5.13}$$

$$\Psi_2 := \operatorname{col}(I_m, \ldots, I_m). \tag{5.14}$$

Taking into account formulas (5.7), (5.12), (5.13) and (5.14) it is immediately checked that

$$S_k - AS_kA^* = \Psi_k\Psi_k^*, \quad k \in \{1, 2\}. \tag{5.15}$$

Thus, in view of (5.15) the results obtained above can be applied to the Nevanlinna-Pick extremal problem. In this case we have

$$S_{11}^{(1)} = \left(\frac{\eta_k\eta_l^*}{1 - z_k\overline{z}_l}\right)_{k,l=1}^{n-1}, \qquad S_{11}^{(2)} = \left(\frac{I_m}{1 - z_k\overline{z}_l}\right)_{k,l=1}^{n-1}, \tag{5.16}$$

$$S_{22}^{(1)} = \frac{\eta_n\eta_n^*}{1 - |z_n|^2}, \qquad S_{22}^{(2)} = \frac{I_m}{1 - |z_n|^2}, \tag{5.17}$$

$$S_{12}^{(1)} = \left(\frac{\eta_k\eta_n^*}{1 - z_k\overline{z}_n}\right)_{k=1}^{n-1}, \qquad S_{12}^{(2)} = \left(\frac{I_m}{1 - z_k\overline{z}_n}\right)_{k=1}^{n-1}. \tag{5.18}$$

Example. We consider the case $n = 2$. Then

$$S_{11}^{(1)} = \frac{\eta_1 \eta_1^*}{1 - |z_1|^2}, \qquad S_{11}^{(2)} = \frac{I_m}{1 - |z_1|^2} \tag{5.19}$$

$$S_{22}^{(1)} = \frac{\eta_2 \eta_2^*}{1 - |z_2|^2}, \qquad S_{22}^{(2)} = \frac{I_m}{1 - |z_2|^2} \tag{5.20}$$

$$S_{12}^{(1)} = \frac{\eta_1 \eta_2^*}{1 - z_1 \overline{z_2}}, \qquad S_{12}^{(2)} = \frac{I_m}{1 - z_1 \overline{z_2}}. \tag{5.21}$$

Thus, using formulas (5.16)–(5.21) we obtain

$$\begin{aligned} \beta &= \frac{1}{1 - |z_2|^2} - \frac{1 - |z_1|^2}{(1 - z_1 \overline{z_2})(1 - \overline{z_1} z_2)} \\ &= \frac{|z_1 - z_2|^2}{[1 - |z_2|^2] \cdot |1 - z_1 \overline{z_2}|^2}. \end{aligned} \tag{5.22}$$

Using (2.8) and (5.22) we get

$$\begin{aligned} U &= \left(S_{12}^{(2)}\right)^* \left(S_{11}^{(2)}\right)^{-1} \left(S_{11}^{(1)}\right) \left(S_{11}^{(2)}\right)^{-1} S_{12}^{(2)} \\ &\quad - \left(S_{12}^{(2)}\right)^* \left(S_{11}^{(2)}\right)^{-1} S_{12}^{(1)} - \left(S_{12}^{(1)}\right)^* \left(S_{11}^{(2)}\right)^{-1} S_{12}^{(2)} + S_{22}^{(1)} \\ &= \frac{1 - |z_1|^2}{|1 - z_1 \overline{z_2}|^2} (\eta_1 - \eta_2)(\eta_1 - \eta_2)^* + \left[\frac{1}{1 - |z_2|^2} - \frac{1 - |z_1|^2}{|1 - z_1 \overline{z_2}|} \right] \eta_2 \eta_2^* \\ &= \beta \left[\frac{(1 - |z_1|^2)(1 - |z_2|^2)}{|z_1 - z_2|^2} (\eta_1 - \eta_2)(\eta_1 - \eta_2)^* + \eta_2 \eta_2^* \right] \end{aligned}$$

Hence,

$$\frac{1}{\beta} U = \frac{(1 - |z_1|^2)(1 - |z_2|^2)}{|z_1 - z_2|^2} (\eta_1 - \eta_2)(\eta_1 - \eta_2)^* + \eta_2 \eta_2^*.$$

In view of

$$D = \sqrt{S_{11}^{(2)}} = \frac{I_m}{\sqrt{1 - |z_1|^2}}$$

the regularity condition $\frac{1}{\beta} U > D^{-1} S_{11}^{(1)} D^{-1}$ has now the form

$$\frac{(1 - |z_1|^2)(1 - |z_2|^2)}{|z_1 - z_2|^2} (\eta_1 - \eta_2)(\eta_1 - \eta_2)^* + \eta_2 \eta_2^* > \eta_1 \eta_1^*.$$

In particular, if $\eta_2 \eta_2^* > \eta_1 \eta_1^*$, then this condition is satisfied.

6. Jordan block diagonal structure

In this section, we consider the case when the matrix A has the block diagonal form

$$A = \operatorname{diag}(A_1, A_2, \ldots, A_N),$$

where for each $k \in \{1, \dots, N\}$ the complex $mn_k \times mn_k$ matrix A_k has the shape

$$A_k = \begin{pmatrix} \lambda_k I_m & 0 & \dots & 0 & 0 \\ I_m & \lambda_k I_m & \dots & 0 & 0 \\ 0 & I_m & \dots & 0 & 0 \\ \vdots & \vdots & & \vdots & \vdots \\ 0 & 0 & \dots & I_m & \lambda_k I_m \end{pmatrix} \tag{6.1}$$

and where

$$|\lambda_k| < 1. \tag{6.2}$$

Let $n := \sum_{k=1}^{N} n_k$. Then we choose the complex $mn \times m$ matrix Ψ_2 via

$$\Psi_2 := \operatorname{col}(\Psi_{21}, \Psi_{22}, \dots, \Psi_{2,N}) \tag{6.3}$$

where

$$\Psi_{2,k} := \operatorname{col}(\underbrace{I_m, 0, \dots, 0}_{n_k}), \quad k \in \{1, \dots, N\} \tag{6.4}$$

Now we are looking for a complex $mn \times mn$ matrix S_2 satisfying the equation

$$S_2 - AS_2A^* = \Psi_2\Psi_2^*. \tag{6.5}$$

Proposition 6.1. *There is one and only complex $mn \times mn$ matrix S_2 satisfying* (6.5)*, namely*

$$S_2 = \sum_{j=0}^{\infty} A^j \Psi_2 \Psi_2^* (A^j)^*. \tag{6.6}$$

Moreover, the matrix S_2 is positive Hermitian.

Proof. In view of (6.1) and (6.2) we see that the spectrum of the matrix A is contained in the open unit disk. Thus, we obtain from [2, p. 578] that the equation (6.5) has a unique solution S_2 which is given by formula (6.6). This matrix S_2 is positive Hermitian. □

We introduce the block diagonal matrix

$$R := \operatorname{diag}(\underbrace{\rho, \rho, \dots, \rho}_{n})$$

where ρ is a positive Hermitian $m \times m$ matrix. Let us consider the equation

$$S_1 - AS_1A^* = \Psi_1\Psi_1^* \tag{6.7}$$

where the complex $mn \times mn$ matrix Ψ_1 is defined by the relations

$$\Psi_1 := R^{-1} \cdot \operatorname{col}(a_1, a_2, \dots, a_N)$$

and

$$a_k = \operatorname{col}(a_{1,k}, a_{2,k}, \dots, a_{n_k,k}), \quad k \in \{1, \dots, N\}.$$

Here the matrices a_k and $a_{j,k}$ have the sizes $mn_k \times m$ and $m \times m$, respectively. Using [2, Theorem A3.4, part (a)] we obtain the following result.

Proposition 6.2. *There is one and only complex $mn \times mn$ matrix S_1 satisfying* (6.7)*, namely*

$$S_1 = \sum_{j=0}^{\infty} A^j \Psi_1 \Psi_1^* (A^j)^*.$$

Moreover, the matrix S_1 is nonnegative Hermitian.

The identities (6.5) and (6.7) generate an extremal interpolation problem to which the results of Section 1–3 can be applied. We omit here its exact formulation.This will be done in detail in a forthcoming paper.

References

[1] Akhiezer, N.I.: *On a minimum problem in function theory and the number of roots of an algebraic equation inside the unit disc* (Russian), Izv. Akad. Nauk SSR, Otdel. Mat. i Est. Nauk **9** (1930), 1169–1189, English Translation in: Topics in Interpolation Theory (Eds.: H. Dym, B. Fritzsche, V.E. Katsnelson, B. Kirstein), OT Series, Volume 95, Birkhäuser, Basel-Boston-Berlin 1997, pp. 19–35.

[2] Ball, J.A.; Gohberg, I.; Rodman, L.: *Interpolation of Rational Matrix Functions*, OT Series, Volume 45, Birkhäuser, Basel-Boston-Berlin 1990.

[3] Ferrante, A.; Levy, B.C.: *Hermitian solutions of the equation $X = Q + NX^{-1}N^*$*, Linear Alg. Appl. **247** (1996), 359–373.

[4] Helton, J.W.; Sakhnovich, L.A.: *Extremal problems of interpolation theory*, Rocky Mount. J. Math. **35** (2005), 819–841.

[5] Kimura, H.: *State space approach to the classical interpolation problem and its applications*, in: Three Decades of Mathematical System Theory (Eds.: H. Nijmeijer, J.M. Schumacher), Lect. Notes Contr. Int. Sci., Volume 135, Springer, Berlin 1989, pp. 243–275.

[6] Ran, A.C.M.; Reurings, M.C.B.: *A nonlinear matrix equation connected to interpolation theory*, Linear Alg. Appl. **379** (2004), 289–302.

[7] Sakhnovich, L.A.: *Interpolation Theory and Its Applications*, Kluwer, Dordrecht 1997.

Bernd Fritzsche and Bernd Kirstein
Fakultät für Mathematik und Informatik
Universität Leipzig
Postfach: 10 09 20
D-04009 Leipzig, Germany
e-mail: fritzsche@mathematik.uni-leipzig.de
kirstein@mathematik.uni-leipzig.de

Lev A. Sakhnovich
99 Cove Avenue
Milford, CT 06461, USA
e-mail: lsakhnovich@gmail.com

Operator Theory:
Advances and Applications, Vol. 199, 347–358

$O(n)$ Algorithms for Banded Plus Semiseparable Matrices

Jitesh Jain, Hong Li, Cheng-Kok Koh and
Venkataramanan Balakrishnan

Abstract. We present a new representation for the inverse of a matrix that is a sum of a banded matrix and a semiseparable matrix. In particular, we show that under certain conditions, the inverse of a banded plus semiseparable matrix can also be expressed as a banded plus semiseparable matrix. Using this result, we devise a fast algorithm for the solution of linear systems of equations involving such matrices. Numerical results show that the new algorithm competes favorably with existing techniques in terms of computational time.

Mathematics Subject Classification (2000). 15A09, 15A23, 65F05, 65L10, 65R20.

Keywords. Semiseperable matrix, fast algorithms, linear solver, inverse, structured matrices.

1. Introduction

Understanding of structured matrices and computation with them have long been problems of theoretical and practical interest [1]. Recently a class of matrices called semiseparable matrices has received considerable attention [2, 3, 4, 5, 6, 7, 8, 9]. Perhaps the simplest example of a semiseparable matrix is given by the inverse of a symmetric tridiagonal matrix: If $A = A^T$ is irreducible tridiagonal and nonsingular, then it is well known that A^{-1} can be written as:

$$A_{ij}^{-1} = \begin{cases} u_i v_j & \text{if } i \leq j, \\ u_j v_i & \text{if } i > j. \end{cases} \tag{1}$$

Matrices such as the one in (1) and its generalizations arise in a number of practical applications like integral equations [10, 11], statistics [12], and vibrational analysis [13]. Modeling with a semiseparable matrix evidently offers the potential of reducing the number of parameters describing a matrix by up to an order of magnitude (from $O(n^2)$ to $O(n)$ in the example in (1)). Moreover, it is known that

This work was done when the authors were at Purdue University.

computation with semiseparable matrices requires significantly reduced effort; for example, for a semiseparable matrix of the form in (1), matrix-vector multiplies can be performed in $O(n)$ (as compared to $O(n^2)$ in general).

In several practical situations, semiseparable matrices do not arise alone; instead, matrices that are encountered, are a sum of diagonal and a semiseparable matrix or a banded and a semiseparable matrix. Examples where such matrices arise, are in boundary value problems [14, 15], and integral equations [16]. The computation with such matrices has been a subject of considerable interest. Several algorithms have been developed to deal with matrix inversion and linear equation solution with such matrices. Formulae for inversion of diagonal plus semiseparable matrices were first developed in [6]. However, these formulae were valid under the assumption that the matrix is strongly regular, i.e., it has non-vanishing principal minors. These restrictions were later removed in [4]. Recently several algorithms have been developed for solving linear systems of equation

$$Ax = c, \tag{2}$$

where the coefficient matrix A is a sum of diagonal and semiseparable matrix [3, 4, 5, 7]. Fast and numerically stable algorithms for banded plus semiseparable linear system of equations were proposed in [2].

We present two main results in this paper. First, we provide an explicit representation for inverses of banded plus semiseparable matrices. In particular, we show that under certain conditions, the inverse of a banded plus semiseparable matrix is again a banded plus semiseparable matrix. Our second contribution is to provide fast solutions of linear systems of equations with these matrices. A comparison with the state of the art shows that our method is about two times faster than existing solutions of linear system with diagonal plus semiseparable matrices. When banded plus semiseparable matrices are considered, our method is up to twenty times faster than existing solutions.

The remainder of the paper is organized as follows. In §2, we establish mathematical preliminaries and the notation used in the paper, as well as a brief review of the state of the art. In §3 we present formulae for the inverse of banded plus semiseparable matrices. We exploit this result in §4 to provide a fast algorithm for solving linear systems of equations. In §5, we establish the effectiveness of the new algorithm via numerical results. The extension of this algorithm to handle some special cases is presented in an appendix.

2. Preliminaries

For $k = 1, \ldots, a$, and $r = 1, \ldots, b$, let $u_k = \{u_k(i)\}_{i=1}^n$, $v_k = \{v_k(i)\}_{i=1}^n$, $p_r = \{p_r(i)\}_{i=1}^n$, and $q_r = \{q_r(i)\}_{i=1}^n$, be specified vectors. Then S_a^b, an $n \times n$ semiseparable matrix of order (a, b), is characterized as follows:

$$S_{ij} = \begin{cases} \sum_{k=1}^{a} u_k(i)v_k(j) & \text{if } i \leq j, \\ \sum_{r=1}^{b} p_r(j)q_r(i) & \text{if } i > j. \end{cases} \tag{3}$$

We use $\mathbb{S}_n$ as the generic notation for the class of semiseparable matrices of size n.

We use $\mathbb{B}_n$ to denote the class of banded matrices of size n. $B_l^m = \{B_{ij}\}_{i,j=1}^n$ is used to denote a banded matrix with l non-zero diagonals strictly above the main diagonal and m non-zero diagonals strictly below the main diagonal, i.e., if $B_l^m \in \mathbb{B}_n$ then $B_{ij} = 0$ if $i - j > m$ or $j - i > l$. The numbers l and m are called respectively the upper and lower bandwidths of a banded matrix B_l^m. $\mathbb{D}_n$ is used to denote the class of diagonal matrices. $D = \mathrm{diag}(d)$ is used to denote a diagonal matrix with d as the main diagonal. We now define a proper banded matrix.

Definition 2.1. A nonsingular banded matrix B_l^m is said to be proper if any submatrix obtained by deleting $r(= \max(l, m))$ consecutive rows and r consecutive columns is nonsingular.

It is well known that the inverses of banded matrices are semiseparable matrices.

Theorem 2.2 ([17]**).** *Let B_l^m be a $n \times n$ proper banded matrix. Then its inverse can be written as*

$$(B_l^m)^{-1} = S_l^m, S_l^m \in \mathbb{S}_n.$$

Remark 2.3. The above semiseparable representation, though elegant and compact, suffers from numerical instabilities [18], making it of limited practical use. Hence, the above representation will be only used as a theoretical tool, and not in any numerical implementations.

We next present a brief review of the state of the art for solving linear system of equation in (2), where A is a sum of banded and semiseparable matrices. We first consider the case when A is a sum of a diagonal and a semiseparable matrix. Two algorithms for solving such systems were developed in [3]. The first step is the same with both algorithms, where $A \in n \times n$ is reduced to a upper Hessenberg matrix H via $n - 1$ Givens rotations:

$$A = \underbrace{G_2^T G_3^T \cdots G_n^T}_{G_{2,\ldots,n}^T} H.$$

It was shown that $G_{2,\ldots,n}^T$ is a lower Hessenberg matrix, whose upper triangular part is the upper triangular part of a unit-rank matrix. The upper triangular part of H was shown to be the upper triangular part of a matrix of rank two. The second step of both algorithms is to reduce the upper Hessenberg matrix H into an upper triangular matrix via $n - 1$ Givens rotations. Two different algorithms were obtained by either applying the Givens rotations on the left of H, leading to QR algorithm, or applying Givens rotations to the right of H, obtaining the URV algorithm. Exploiting the low rank structure of $G_{2,\ldots,n}^T$ and H, both algorithms were shown to require $54n - 44$ flops as compared to $58n$ flops for the algorithm in [5] and $59n$ flops for the one in [2].

The authors in [2] proposed a fast and numerical stable algorithm for the more general case of the solution of (2) when A is a sum of banded and semiseparable matrix. The basic idea is to compute a two-sided decomposition of the matrix A such that $A = WLH$. Here L is a lower triangular matrix, and both W and H can be written as a product of elementary matrices. An efficient algorithm for solving (2) is obtained by inverting W, L, and H on the fly. The matrices W and H can be either obtained via standard Gaussian elimination, or by using Givens rotations and Householder reflection matrices. The algorithm based on Gaussian elimination was shown to be marginally better in computational time than the one based on Givens rotations and Householder reflections, with the latter algorithm performing better with respect to accuracy. For a banded matrix with upper and lower bandwidth l and m and a semiseparable matrix of size n and order (a, b), the algorithm based on Givens rotation and Householder reflection was shown to have an operation count of $\left(11a^2 + 2\left(2l + 2m + 3b + 5a\right)\left(l + a\right)\right) n$ flops, while the algorithm based on Gaussian elimination requires $\left(9a^2 + 2\left(l + 2m + 2b + 2a\right)\left(l + a\right)\right) n$ flops.

The approach we take in this paper is to first provide an explicit representation for inverses of banded plus semiseparable matrices. We then exploit these results to come up with a fast algorithm for solving linear systems of equations.

3. Structure for inverses of banded plus semiseparable matrices

We begin this section by considering semiseparable matrices of order (1,1). We first present a theorem on multiplicative structure of inverses of banded plus semiseparable matrices of order $(1, 1)$.

Theorem 3.1. *Let B_l^m be a $n \times n$ banded matrix and S_1^1 be a $n \times n$ semiseparable matrix of order $(1, 1)$. Then the inverse of their sum has the following multiplicative structure:*

$$\left(B_l^m + S_1^1\right)^{-1} = DL^T \left(B_{l+1}^{m+1}\right)^{-1} L\tilde{D},$$

where D, $\tilde{D} \in \mathbb{D}_n$, $B_{l+1}^{m+1} \in \mathbb{B}_n$, and L is a lower bidiagonal matrix.

Proof: Let

$$\left(S_1^1\right)_{ij} = \begin{cases} u_i v_j & \text{if } i \leq j, \\ p_j q_i & \text{if } i > j. \end{cases}$$

Assume $v_1(i) \neq 0$, $q_1(i) \neq 0$ for this and the next section. The results have been extended for the more general case in the appendix. Let $D_v = \text{diag}(1/v)$ and $D_q = \text{diag}(1/q)$ be diagonal matrices. Let L be a lower bidiagonal matrix, which is defined as follows:

$$L = \begin{pmatrix} 1 & & & & \\ -1 & 1 & & & \\ & -1 & 1 & & \\ & & \ddots & \ddots & \\ & & & -1 & 1 \end{pmatrix}. \tag{4}$$

It can be easily verified that

$$LD_q S_1^1 D_v L^T = \hat{L}, \tag{5}$$

$$D_v L^T \hat{S}_1^1 L D_q = \tilde{S}_1^1 + D, \tag{6}$$

$$LD_q \hat{S}_1^1 D_v L^T = \tilde{S}_1^1 + \tilde{L}, \tag{7}$$

where $\hat{L}$, $\tilde{L}$ are lower bidiagonal matrices, $D \in \mathbb{D}_n$, and $\hat{S}_1^1$, $\tilde{S}_1^1 \in \mathbb{S}_n$. Now, from (5)

$$LD_q \left(B_l^m + S_1^1\right) D_v L^T = LD_q B_l^m D_v L^T + \hat{L}.$$

It can be readily verified that matrix $LD_q B_l^m D_v L^T$ is banded with lower and upper bandwidth $l+1$ and $m+1$ respectively, i.e., $LD_v B_l^m D_q^T L^T = \tilde{B}_{l+1}^{m+1} \in \mathbb{B}_n$. Moreover, as $\hat{L}$ is lower bidiagonal, we have

$$LD_q \left(B_l^m + S_1^1\right) D_v L^T = B_{l+1}^{m+1}$$

Thus

$$\left(B_l^m + S_1^1\right)^{-1} = D_v L^T \left(B_{l+1}^{m+1}\right)^{-1} LD_q. \qquad \square$$

Remark 3.2. Suppose the banded matrix B_{l+1}^{m+1} in Theorem 3.1 is proper. Then using Theorem 2.2 and (6), we have the following elegant additive structure for the inverse of a banded and semiseparable matrix of order $(1,1)$:

$$\begin{aligned}\left(B_l^m + S_1^1\right)^{-1} &= D_v L^T \tilde{S}_{l+1}^{m+1} L D_q \\ &= S_{l+1}^{m+1} + D,\end{aligned}$$

where $S_{l+1}^{m+1} \in \mathbb{S}_n$, and $D \in \mathbb{D}_n$.

More generally, the following theorem characterizes the inverses of general banded plus semiseparable matrices.

Theorem 3.3. *Let B_l^m be a $n \times n$ banded matrix and S_a^b be a $n \times n$ semiseparable matrix. Then the inverse of their sum has the following multiplicative structure:*

$$\left(B_l^m + S_a^b\right)^{-1} = D_1' L^T \cdots D_b' L^T \left(B_{l+a}^{m+b}\right)^{-1} LD_a \cdots LD_1,$$

where $D_1, D_1', D_2, D_2', \ldots, D_a, D_a', D_b, D_b' \in \mathbb{D}_n$, $B_{l+a}^{m+b} \in \mathbb{B}_n$, and L is a lower bidiagonal matrix.

Proof: Without loss of generality, assume $b \geq a$. Now, from (5) and (7)

$$\begin{aligned}LD_1 \left(B_l^m + S_a^b\right) D_1' L^T &= B_{l+1}^{m+1} + S_{a-1}^{b-1} \\ LD_a \cdots LD_1 \left(B_l^m + S_a^b\right) D_1' L^T \cdots D_a' L^T &= B_{l+a}^{m+a} + S_0^{b-a} \\ LD_a \cdots LD_1 \left(B_l^m + S_a^b\right) D_1' L^T \cdots D_a' L^T \cdots D_b' L^T &= B_{l+a}^{m+b},\end{aligned}$$

where $D_1, D_1', D_2, D_2', \ldots, D_b, D_b' \in \mathbb{D}_n$. Thus

$$\left(B_l^m + S_a^b\right)^{-1} = D_1' L^T \cdots D_a{}' L^T \cdots D_b' L^T \left(B_{l+a}^{m+b}\right)^{-1} LD_a \cdots LD_1. \qquad \square$$

Remark 3.4. Suppose the banded matrix B_{l+a}^{m+b} in Theorem 3.3 is proper. Then using Theorem 2.2 and (6), we have the following elegant additive structure for the inverse of a banded and semiseparable matrix:

$$\begin{aligned}\left(B_l^m + S_a^b\right)^{-1} &= D_1' L^T \cdots D_a' L^T \cdots D_b' L^T \tilde{S}_{l+a}^{m+b} L D_a \cdots L D_1 \\ &= B_{a-1}^{b-1} + S_{l+a}^{m+b},\end{aligned}$$

where $B_{a-1}^{b-1} \in \mathbb{B}_n$, and $S_{l+a}^{m+b} \in \mathbb{S}_n$.

Corollary 3.5. *The inverse of a diagonal plus semiseparable matrix is again a diagonal plus semiseparable matrix.*

4. Fast solution of $Ax = c$

In this section, we consider the problem of finding the solution of linear systems of equation, $Ax = c$, where the coefficient matrix, A is a sum of banded and semiseparable matrix, i.e.,

$$A = B_l^m + S_a^b,\ B_l^m \in \mathbb{B}_n,\ S_a^b \in \mathbb{S}_n.$$

Without loss of generality assume $b \geq a$.

$$x = A^{-1}c = \left(B_l^m + S_a^b\right)^{-1} c = D_1' L^T \cdots D_a' L^T \cdots D_b' L^T \left(B_{l+a}^{m+b}\right)^{-1} L D_a \cdots L D_1 c. \tag{8}$$

`function` $x = \text{Ainvc}(S_a^b, B_l^m, c)$

1. Calculate $D_1, D_1', \ldots, D_a, D_a', \ldots, D_b, D_b',$ and B_{l+a}^{m+b} as described in proof of Theorem 3.3;
2. $z = L D_a \cdots L D_1 c$;
3. $y = B_{l+a}^{m+b} \setminus z$ (Solve $B_{l+a}^{m+b} y = z$);
4. $x = D_1' L^T \cdots D_a' L^T \cdots D_b' L^T y$;

`return` x;

Note that for solving $Ax = c$, we do not need the condition of the banded matrices being proper.

We now present a complexity analysis of the above-mentioned procedure of solving $Ax = c$. We will assume $1 \ll a, b, l, m \ll n$ to make the complexity analysis simpler. The flop count of the overall algorithm is dominated by the cost of steps 1 and 3. Total cost of step-1, i.e., calculating $D_1, D_1', \ldots, D_a, D_a', \ldots, D_b, D_b'$, and B_{l+a}^{m+b} is

$$\left(2\,(l+m)\max(a,b) + 8\left(a^2+b^2\right)\right) n$$

flops. In step-3 the cost of solving a banded system of equations is $2(l+a)(m+b)n$ flops. Hence the total complexity of the proposed algorithm is

$$\left(2(l+a)(m+b) + 2\,(l+m)\max(a,b) + 8\left(a^2+b^2\right)\right) n$$

operations. For diagonal plus semiseparable systems, the complexity reduces to $30n$ operations.

5. Numerical results

In this section, we compare the results for computational time and accuracy for solving a linear system of equations of form $Ax = b$, for several algorithms. We compare the results for following four algorithms:

- Algorithm I: QR algorithm for solving a diagonal plus semiseparable system, as given in [3].
- Algorithm II: URV algorithm for solving a diagonal plus semiseparable system, as given in [3].
- Algorithm III: Chandrasekaran-Gu algorithm for banded plus semiseparable system of equations, as given in [2].
- Algorithm IV: The new algorithm as described in §4.

All numerical experiments were performed in MATLAB running on a 4 CPU 1.5GHz Intel® Pentium® machine. For Algorithm I and II, we used the author's implementation, taken directly from [19]. All the matrix entries are randomly generated, drawn from a Gaussian distribution with zero mean and unit variance.

5.1. A is a sum of diagonal and semiseparable matrix

We first present the computational requirements and the accuracy of our approach against Algorithm I and II from [3]. The matrix A comprises of a diagonal plus a semiseparable matrix. Table 1 summarizes the results. Error in solving x in $Ax = b$ is defined as $\frac{\|(Ax-b)\|_\infty}{\|A\|_\infty \|x\|_\infty}$. As expected all three algorithms are linear in computational time. Algorithm IV is faster than and comparable in accuracy to Algorithms I and II. For a system with size 320000, Algorithm IV is 1.9× faster than Algorithm I and 2.4× faster than Algorithm II. This supports the theoretical complexities mentioned previously, where Algorithm IV takes $30n$ operations as compared to $54n - 44$ operations taken by Algorithm I and II.

Size	Error = $\frac{\|(Ax-b)\|_\infty}{\|A\|_\infty \|x\|_\infty}$			Time (in sec)		
	Alg I [3]	Alg II [3]	Alg IV	Alg I [3]	Alg II [3]	Alg IV
10000	3.33×10^{-19}	5.32×10^{-19}	3.27×10^{-18}	.34	.43	.14
20000	8.01×10^{-19}	4.66×10^{-19}	1.40×10^{-17}	.68	.85	.27
40000	9.24×10^{-19}	1.80×10^{-18}	5.63×10^{-18}	1.39	1.74	.65
80000	1.47×10^{-18}	8.27×10^{-19}	2.15×10^{-17}	2.77	3.50	1.41
160000	1.29×10^{-19}	5.49×10^{-19}	5.65×10^{-19}	5.59	7.00	3.05
320000	2.34×10^{-19}	5.24×10^{-19}	2.30×10^{-18}	11.06	13.92	5.86
640000	6.53×10^{-20}	1.71×10^{-19}	8.91×10^{-18}	22.02	27.79	11.86
1280000	8.30×10^{-20}	8.40×10^{-19}	2.88×10^{-18}	44.17	55.89	24.39

TABLE 1. Error values and Computational time as compared with Algorithms I and II.

5.2. A is a sum of banded and semiseparable matrix

We now present the computational requirements and the accuracy of our approach against Algorithm III from [2]. The matrix A comprises of a banded plus semiseparable matrix. The three variables of interest of the coefficient matrix are the bandwidth of the banded matrix, order of the semiseparable matrix and size of the system. We first give results by varying one of the quantities at a time, keeping other two constant. Table 2 shows the results for increasing sizes of the linear

Size	Error $= \frac{\|(Ax-b)\|_\infty}{\|A\|_\infty \|x\|_\infty}$		Time (in sec)	
	Alg III [2]	Alg IV	Alg III [2]	Alg IV
1000	8.58×10^{-17}	3.07×10^{-13}	1.33	0.13
2000	6.41×10^{-17}	8.42×10^{-13}	2.65	0.26
3000	1.78×10^{-16}	2.68×10^{-13}	3.96	0.39
4000	7.14×10^{-17}	1.57×10^{-12}	5.29	0.53
5000	1.84×10^{-16}	1.87×10^{-12}	6.57	0.65
6000	1.20×10^{-16}	1.25×10^{-12}	7.98	0.80
7000	9.05×10^{-17}	8.48×10^{-13}	9.27	0.96
8000	3.73×10^{-17}	1.03×10^{-12}	10.62	1.12

TABLE 2. Error values and Computational time as compared with Algorithm III; $l = m = a = b = 5$.

systems. The upper and lower bandwidth of the banded matrix in all cases is 5. The order of the semiseparable matrix is also kept constant at $(5, 5)$. Algorithm IV performs favorably in terms of computational time, being $10\times$ faster than Algorithm IV for the size of 8000. Similar results are seen in Table 3 and 4, where we are varying the bandwidth and the order respectively. Algorithm III exhibits better accuracy than Algorithm IV, as it relies heavily on Givens rotations.

We now give results when all three variables are varied at the same time. Table 5 shows the results for increasing sizes of the linear systems. The upper and lower bandwidth of banded matrices as well as the order of semiseparable matrices are varied as $\frac{n}{250}$, where n denotes the size of the system. Algorithm IV is faster, but at the expense of numerical accuracy. The computational times are consistent with the theoretical flop count. For $a = b = l = m = r$, flop count of Algorithm III reduces to $59r^2n$. For the same case, flop count of the proposed Algorithm is $28r^2n$.

$l = m$	Error = $\frac{\|(Ax-b)\|_\infty}{\|A\|_\infty \|x\|_\infty}$		Time (in sec)	
	Alg III [2]	Alg IV	Alg III [2]	Alg IV
11	7.51×10^{-17}	3.34×10^{-13}	2.24	0.16
21	2.33×10^{-16}	2.40×10^{-13}	3.57	0.24
31	8.64×10^{-16}	4.82×10^{-13}	4.90	0.35
41	6.88×10^{-16}	1.22×10^{-12}	5.54	.46
51	4.32×10^{-16}	5.61×10^{-13}	5.60	.55
61	1.33×10^{-15}	2.33×10^{-12}	7.24	.67
71	4.95×10^{-16}	1.26×10^{-12}	8.19	.71
81	9.71×10^{-16}	2.87×10^{-12}	9.61	.81

TABLE 3. Error values and Computational time as compared with Algorithm III; $n = 4000$, $a = b = 1$.

$a = b$	Error = $\frac{\|(Ax-b)\|_\infty}{\|A\|_\infty \|x\|_\infty}$		Time (in sec)	
	Alg III [2]	Alg IV	Alg III [2]	Alg IV
11	2.41×10^{-16}	2.70×10^{-10}	8.96	1.09
21	3.52×10^{-16}	2.88×10^{-10}	15.47	3.20
31	9.79×10^{-16}	1.82×10^{-8}	25.29	6.04
41	6.37×10^{-16}	2.48×10^{-9}	43.19	9.40
51	1.16×10^{-15}	3.75×10^{-9}	61.07	14.83
61	1.01×10^{-15}	6.06×10^{-9}	117.37	21.32
71	1.68×10^{-15}	1.62×10^{-7}	188.40	27.26
81	1.70×10^{-15}	1.03×10^{-7}	288.77	34.94

TABLE 4. Error values and Computational time as compared with Algorithm III; $n = 4000$, $l = m = 1$.

6. Conclusions

We have presented a representation for inverse of banded plus semiseparable matrices. We have also presented fast algorithms for solving linear system of equations with these matrices. Numerical results show that the proposed approach competes favorably with the state of the art algorithms in terms of computational efficiency.

Appendix

In the discussion till now, we have assumed we are given a semiseparable matrix S_a^b, as defined in (3), such that for all i, k $v_k(i) \neq 0, q_k(i) \neq 0$. Now, we will give procedure to modify the proposed methods when such assumptions do not hold

Size	Error = $\frac{\|(Ax-b)\|_\infty}{\|A\|_\infty \|x\|_\infty}$		Time (in sec)	
	Alg III [2]	Alg IV	Alg III [2]	Alg IV
1000	7.00×10^{-17}	1.21×10^{-12}	1.31	0.13
2000	6.68×10^{-17}	7.69×10^{-12}	4.49	0.90
3000	2.69×10^{-16}	1.69×10^{-11}	8.84	2.71
4000	2.51×10^{-16}	1.46×10^{-9}	16.46	6.53
5000	5.14×10^{-16}	6.25×10^{-10}	26.36	11.62
6000	2.32×10^{-16}	1.11×10^{-10}	41.16	19.50
7000	2.55×10^{-16}	2.67×10^{-9}	63.70	30.93
8000	6.37×10^{-16}	3.39×10^{-10}	96.16	45.39

TABLE 5. Error values and Computational time as compared with Algorithm III; $l = m = a = b = \frac{n}{250}$.

true. We will only consider the symmetric case, i.e., $p = u,\ q = v$. In addition, we assume that the order of semiseparable matrices in consideration is $(1, 1)$, i.e., $l = m = 1$. Hence

$$S_{ij} = \begin{cases} u_i v_j & \text{if } i \le j, \\ u_j v_i & \text{if } i > j. \end{cases}$$

The general case can be handled in a similar fashion.

Let us assume $v_k = 0$, and $v_i \neq 0$ if $i \neq k$. The technique that we propose next can be easily extended to handle the case when for more than one i, v_i is zero. Consider the matrices L, and D_v as defined in §3. In addition, let $L(k, k-1) = 0$, and $D_v(k, k) = 1$. Then it can be easily verified that

$$L D_v S_1^1 D_v L^T = M_{u,v},$$

where M is defined as follows:

$$M_{u,v} = \begin{pmatrix} \alpha_1 & & & & \alpha_1 & & \\ & \alpha_2 & & & \alpha_2 & & \\ & & \ddots & & \vdots & & \\ & & & 0 & \alpha_k & & \\ \alpha_1 & \alpha_2 & \dots & \alpha_k & \alpha_{k+1} & & \\ & & & & & \ddots & \\ & & & & & & \alpha_n \end{pmatrix},$$

where $\alpha_1 = u_1 v_1,\ \alpha_k = u_k,\ \alpha_{k+1} = u_{k+1} - u_k,\ \alpha_i = u_i v_i - u_{i-1} v_{i-1}$ for all $i \notin \{1, k, k+1\}$. We can now do a tridiagonal decomposition of $M_{u,v}$ as

$$P M P^T = T,$$

where T is a tridiagonal matrix, and P is given by

$$P = \begin{pmatrix} 1 & -\frac{\alpha_1}{\alpha_2} & & & \\ & 1 & -\frac{\alpha_2}{\alpha_3} & & \\ & & \ddots & \ddots & \\ & & & 1 & -\frac{\alpha_{n-1}}{\alpha_n} \\ & & & & 1 \end{pmatrix}$$

Proceeding as with the proof of Theorem 1, we can show that

$$\left(B_l^m + S_1^1\right)^{-1} = DL^T \left(B_{l+2}^{m+2}\right)^{-1} L\tilde{D},$$

where $D, \tilde{D} \in \mathbb{D}_n, B_{l+2}^{m+2} \in \mathbb{B}_n$, and L is a lower bidiagonal matrix.

References

[1] T. Kailath and A.H. Sayed. *Fast reliable algorithms for matrices with structure.* Society for Industrial and Applied Mathematics, Philadelphia, PA, USA, 1999.

[2] S. Chandrasekaran and M. Gu. Fast and Stable algorithms for banded plus semiseparable systems of linear equations. *SIAM Journal on Matrix Analysis and Applications*, 25(2):373–384, 2003.

[3] E. Van Camp, N. Mastronardi and M. Van Barel. Two fast algorithms for solving diagonal-plus-semiseparable linear systems. *Journal of Computational and Applied Mathematics*, 164-165:731–747, 2004.

[4] Y. Eidelman and I. Gohberg. Inversion formulas and linear complexity algorithm for diagonal plus semiseparable matrices. *Computer and Mathematics with Applications*, 33(4):69–79, 1996.

[5] Y. Eidelman and I. Gohberg. A modification of Dewilde-van der Veen method for inversion of finite structured matrices. *Linear Algebra Appl.*, 343-344:419–450, 2002.

[6] I. Gohberg and T. Kailath and I. Koltracht. Linear Complexity algorithms for semiseparable matrices. *Integral Equations and Operator Theory*, 8:780–804, 1985.

[7] N. Mastronardi, S. Chandrasekaran and S.V. Huffel. Fast and stable two-way algorithm for diagonal plus semi-separable system of linear equations. *Numerical Linear Algebra Appl.*, 8(1):7–12, 2001.

[8] P. Rozsa. Band Matrices and semi-separable matrices. *Colloquia Mathematica Societatis Janos Bolyai*, 50:229–237, 1986.

[9] P. Rozsa, R. Bevilacqua, F. Romani and P. Favati. On Band Matrices and their inverses. *Linear Algebra Appl.*, 150:287–295, 1991.

[10] E. Asplund. Inverses of Matrices a_{ij} which satisfy $a_{ij} = 0$ for $j > i+p$. *Mathematica Scandinavica*, 7:57–60, 1959.

[11] S.O. Asplund. Finite boundary value problems solved by Green's Matrix. *Mathematica Scandinavica*, 7:49–56, 1959.

[12] S.N. Roy and A.E Sarhan. On inverting a class of patterned matrices. *Biometrika*, 43:227–231, 1956.

[13] F.R. Gantmacher and M.G. Krein. *Oscillation matrices and kernels and small vibrations of mechanical systems.* AMS Chelsea publishing, 2002.

[14] L. Greengard and V. Rokhlin. On the numerical solution of two-point boundary value problems. *Communications on Pure and Applied Mathematics*, 44:419–452, 1991.

[15] J. Lee and L. Greengard. A fast adaptive numerical method for stiff two-point boundary value problems. *SIAM Journal on Scientific Computing*, 18:403–429, 1997.

[16] Starr, Jr., Harold Page. On the Numerical Solution of One-Dimensional Integral and Differential Equations. *Department of Computer Science, Yale University*, New Haven, CT, 1992.

[17] F. Romani. On the additive structure of inverses of banded matrices. *Linear Algebra Appl.*, 80:131–140, 1986.

[18] P. Concus and G. Meurant. On Computing INV block preconditionings for the conjugate gradient method. *BIT*, 26:493–504, 1986.

[19] http://www.cs.kuleuven.ac.be/∼marc/software/index.html.

Jitesh Jain
Intel Corporation
Hillsboro, OR 97124, USA
e-mail: jitesh.jain@intel.com

Hong Li
Synopsys Inc.
Mountain View, CA 94043, USA
e-mail: lhong@synopsys.com

Cheng-Kok Koh and Venkataramanan Balakrishnan
School of Electrical and Computer Engineering
Purdue University
West Lafayette, IN 47907-1285, USA
e-mail: chengkok@purdue.edu
ragu@purdue.edu

Operator Theory:
Advances and Applications, Vol. 199, 359–375

Unified Nearly Optimal Algorithms for Structured Integer Matrices

Victor Y. Pan, Brian J. Murphy and Rhys Eric Rosholt

Abstract. Our subject is the solution of a structured linear system of equations, which is closely linked to computing a shortest displacement generator for the inverse of its structured coefficient matrix. We consider integer matrices with the displacement structure of Toeplitz, Hankel, Vandermonde, and Cauchy types and combine the unified divide-and-conquer MBA algorithm (due to Morf 1974, 1980 and Bitmead and Anderson 1980) with the Chinese remainder algorithm to solve both computational problems within nearly optimal randomized Boolean and word time bounds. The bounds cover the cost of both solution and its correctness verification. The algorithms and nearly optimal time bounds are extended to the computation of the determinant of a structured integer matrix, its rank and a basis for its null space and further to some fundamental computations with univariate polynomials that have integer coefficients.

Mathematics Subject Classification (2000). 68W30, 68W20, 68Q25, 68W40.

Keywords. Structured matrices, the MBA divide-and-conquer algorithm.

1. Introduction

Linear systems of equations with displacement structure of Toeplitz, Hankel, Vandermonde and Cauchy types are omnipresent in scientific and engineering computations and signal and image processing. Due to the structure they can be solved fast, in quadratic rather than cubic arithmetic time [L47], [D59], [T64] or even

Some results of this paper have been presented at the Annual International Conference on Application of Computer Algebra, Volos, Greece, June 2002; ACM International Symposium on Symbolic and Algebraic Computation, Lille, France, July 2002; and the 5th Annual Conference on Computer Algebra in Scientific Computing, Yalta, Crimea, Ukraine, September 2002.

Supported by NSF Grant CCR 9732206 and PSC CUNY Awards 67297–0036, 68291–0037, 69330–0038, and 69350–0038.

superfast, in nearly linear time [BGY80], [M80], [BA80], [OP98], [P01]. Quite typically, however, application of the superfast algorithms leads to the problems of numerical stability [B85]. Moreover, structured linear systems of some important classes are ill conditioned [GI88], [T94]. This suggests devising fast and superfast symbolic algorithms, applied to the case of integer input coefficients, to which one can reduce the rational inputs by scaling. The complexity of such algorithms is usually analyzed under the Boolean (bit-operation) and word operation models [GG03]. Our main goal in this paper is to present the algorithm in a unified way for the general class of integer input matrices with displacement structure and to estimate the complexity of this algorithm and the probability of degeneration (cf. [OP98], [P01], [PW08]). Besides the complexity and degeneration issues, we elaborate upon recursive compression of the displacement generators in the MBA process, displacement transformation of the input matrices, and reconstruction of the rational solution from its representation modulo selected distinct random primes. We also cover the extension to computations with structured singular matrices. We arrive at the complexity bounds that cover both randomized solution and its correction verification and that are nearly optimal (up to a polylogarithmic factor) under both Boolean and word operation models. The bounds cover the cost of both solution and its correctness verification. Versus the information lower bound of $n^2 \log n$, the algorithm takes the order of $n^2 \log^3 n$ bit-operations for a structured linear system of n equations with n unknowns where all coefficients have absolute values in $n^{O(1)}$. The same cost bound covers the computation of the determinant, rank, and a basis for the null space. The algorithm can be called superfast because its Boolean cost bound is nearly quadratic versus the orders of n^4 and n^3 bit-operations required for the solution by Gaussian elimination and by the fast algorithms such as Levinson–Durbin's and Trench's, respectively. Our nearly optimal complexity estimates can be extended to Berlekamp–Massey's reconstruction of a linear recurrence coefficients from its values and to computing the greatest common divisor, least common multiples, and Padé approximation for univariate polynomials. We organize our presentation as follows. In the next section we state some definitions and basic results. We describe the unified MBA divide-and-conquer algorithm in Section 3. In Section 4 we cover the reconstruction of the rational solution from the solution modulo sufficiently many distinct primes and estimate the overall Boolean and word complexity of our computations. In Section 5 we recall various related works. The presented algorithms have been implemented by the third and mostly the second authors. Otherwise the paper is due to the first author.

2. Definitions and basic facts

2.1. Integers

We write log for $\log_2$ (unless we specify otherwise), $\mathbb{Z}$ for the ring of integers, $\mathbb{Z}_q$ for the ring of integers modulo an integer q, and $\mathbb{Q}$ for the field of rational numbers.

"Ops" stand for "arithmetic operations". $\tilde{O}(f(n))$ denotes $O(f(n)(\log\log n)^c)$ for a constant c. $f(s) = O(1)$ as well as "$f(s)$ is in $O(1)$" means that $f(s)$ is bounded by a constant independent of s. We write $a = z \mod q$, for three integers $q > 1$, a, and z, either to denote a unique integer a such that q divides $z - a$ and $0 \le a < q$ or, wherever due to the context this causes no confusion, just to show that q divides $z - a$.

Fact 2.1. *Assume two integers a and m such that $m > 1$ and $m > 2|a|$. Then the following expressions hold,*

$$a = a \mod m \text{ if } 2|a \mod m| < m,$$

$$a = a \mod m - m \text{ otherwise.}$$

2.2. Polynomial and integer multiplication

Definition 2.1. *Let $\mu(d))$ denote the minimum number of bit operations sufficient to perform an arithmetic operation modulo a prime $q < 2^d$, including division by an integer coprime with q, and let $m(n)$ denote the minimum number of field operations sufficient to multiply two polynomials of degree $n-1$ or less over any field, ring with unity, or algebra.*

Fact 2.2. *We have $\mu(d) = O((d\log d)\log\log d)$ and $m(n) = O((n\log n)\log\log n)$* [CK91], [K98], [B03], [GG03], [F07].

2.3. General matrices

Definition 2.2. *$M = (m_{i,j})_{i,j=1}^{k,l} \in \mathbb{R}^{k\times l}$ is a $k\times l$ matrix with entries $m_{i,j}$ in a ring $\mathbb{R}$. $\mathbf{v} = (v_i)_{i=1}^{k} \in \mathbb{R}^{k\times 1}$ is a column vector. I is the identity matrix of a proper size. I_l is the $l\times l$ identity matrix. (K, L) is a 1×2 block matrix with the blocks K and L. $D(\mathbf{v}) = \mathrm{diag}(\mathbf{v}) = \mathrm{diag}(v_i)_i$ is the diagonal matrix with the diagonal entries $d_{ii} = v_i$ given by the coordinates of the vector $\mathbf{v} = (v_i)_i$. M^T is the transpose of M. $M^{(h)}$ is the $h\times h$ leading principal (that is northwestern) submatrix of M. A matrix M of rank ρ has* generic rank profile *if its submatrices $M^{(k)}$ are nonsingular for $k = 1, \dots, \rho$, that is up to the rank size $\rho\times\rho$. M is* strongly nonsingular *if it is nonsingular and has generic rank profile. A block of a matrix is its submatrix in the intersection of a set of its contiguous rows and a set of its contiguous columns.*

Definition 2.3. *The symmetric matrix M^TM for a nonsingular matrix M is called positive definite. (In $\mathbb{Q}^{n\times n}$ such a matrix is strongly nonsingular.)*

Definition 2.4. *$\det M$ and $\mathrm{adj}\, M$ are the determinant and the adjoint of a matrix M, respectively.* ($\mathrm{adj}\, M = M^{-1}\det M$ *if M is nonsingular.*)

Definition 2.5. *$|M| = ||M||_\infty = \max_i \sum_j |m_{i,j}|$ is the row norm of a matrix $M = (m_{i,j})_{i,j}$; $\alpha(M) = \max_{i,j} |m_{i,j}|$; $|M|/n \le \alpha(M) \le |M|$; $|\mathbf{v}| = \beta(\mathbf{v}) = \max_i |v_i|$ is the maximum norm of a vector $\mathbf{v} = (v_i)_i$.*

Definition 2.6. *$m_S \le 2n^2 - n$ is the minimum number of arithmetic operations in an algorithm that multiplies an $n\times n$ matrix S by a vector.*

Clearly, we can multiply a pair of $n \times n$ matrices by using $2n^3 - n^2$ arithmetic operations. We refer the reader to [P84], [CW90], [K04], and the bibliography therein on theoretical and practical speed up.

Fact 2.3. $|\det M| \le \prod_j (\Sigma_i m_{i,j}^2)^{1/2} \le (\alpha(M)\sqrt{n})^n$, *and so* (*since the entries of the matrix* adj M *are the determinants of* $(n-1) \times (n-1)$ *submatrices of* M), *we have* $\alpha(\text{adj}\, M) \le (\alpha(M)\sqrt{n-1})^{n-1}$ *for an* $n \times n$ *matrix* $M = (m_{i,j})_{i,j}$.

Hadamard's bound $|\det M| \le (\alpha(M)\sqrt{n})^n$ above is known to be sharp in the worst case, but is an over-estimate on the average according to [ABM99].

Remark 2.1. *We can apply Fact* 2.1 *and readily recover the integer* $\det M$ *as soon as we have its value computed modulo* p_+ *for some* $p_+ > 2(\alpha(M)\sqrt{n})^n$ *because* $(\alpha(M)\sqrt{n})^n \ge |\det M|$ *in virtue of Fact* 2.3. *If* M *is a nonsingular matrix in* $\mathbb{Z}_{p_+}$ *and if its determinant* $\det M$ *and inverse* M^{-1} *have been computed in* $\mathbb{Z}_{p_+}$, *then we can immediately compute in* $\mathbb{Z}_{p_+}$ *the matrix* $\text{adj}\, M = M^{-1} \det M$. *In virtue of Fact* 2.3, *its integer entries lie in the range* $(-p_+/2, p_+/2)$, *and we can recover them from their values modulo* p_+ *by applying Fact* 2.1 *again.*

2.4. Matrices with displacement structure: general properties

In this subsection we define matrices with displacement structure and recall their basic properties.

Definition 2.7. $\Delta_{A,B}(M) = M - AMB = GH^T$ (*resp.* $\nabla_{A,B}(M) = AM - MB = GH^T$) *is the Stein* (*resp. Sylvester*) displacement *of an* $n \times n$ *matrix* M *where* $n \times n$ *matrices* A *and* B *are* operator matrices *and a pair of* $n \times l$ *matrices* G *and* H *form a* displacement generator *of length* l *for the matrix* M. *The rank of the displacement is called the* displacement rank *of the matrix* M. (*It minimizes the length* l *of displacement generators* $\Delta_{A,B}(M)$ (*resp.* $\nabla_{A,B}(M)$) *for a fixed triple* (A, B, M).)

The simple basic results below are from [P01, Theorems 1.3.1, 1.5.1–1.5.6].

Theorem 2.1. *If the matrix* A (*resp.* B) *is nonsingular, then we have* $\Delta_{A,B}(M) = A^{-1}\nabla_{A^{-1},B}(M)$ (*resp.* $\Delta_{A,B}(M) = -\nabla_{A,B^{-1}}(M)B^{-1}$).

Theorem 2.2. *For matrices* A, B, M, *and* N *of compatible sizes, displacement operators* $L = \Delta_{A,B}$ *and* $L = \nabla_{A,B}$, *and a scalar* a, *we have* $L(M + aN) = L(M) + aL(N)$, $\Delta_{A,B}(M^T) = (\Delta_{B^T,A^T}(M))^T$, $\nabla_{A,B}(M^T) = -(\nabla_{B^T,A^T}(M))^T$. *Furthermore* $\nabla_{B,A}(M^{-1}) = -M^{-1}\nabla_{A,B}(M)M^{-1}$ *if* M *is a nonsingular matrix,* $\Delta_{B,A}(M^{-1}) = BM^{-1}\Delta_{A,B}(M)B^{-1}M^{-1}$ *if the matrices* B *and* M *are nonsingular, and* $\Delta_{B,A}(M^{-1}) = M^{-1}A^{-1}\Delta_{A,B}(M)M^{-1}A$ *if the matrices* A *and* M *are nonsingular.*

Theorem 2.3. *For any* 5*-tuple* $\{A, B, C, M, N\}$ *of matrices of compatible sizes we have* $\nabla_{A,C}(MN) = \nabla_{A,B}(M)N + M\nabla_{B,C}(N)$, $\Delta_{A,C}(MN) = \Delta_{A,B}(M)N + AM\nabla_{B,C}(N)$. *Furthermore* $\Delta_{A,C}(MN) = \Delta_{A,B}(M)N + AMB\Delta_{B^{-1},C}(N)$ *if* B *is a nonsingular matrix and* $\Delta_{A,C}(MN) = \Delta_{A,B}(M)N - AM\Delta_{B,C^{-1}}(N)C$ *if* C *is a nonsingular matrix.*

Theorem 2.4. *Represent the matrices A, B, M, $\nabla_{A,B}(M)$, and $\Delta_{A,B}(M)$ as 2×2 block matrices with blocks $A_{i,j}, B_{i,j}, M_{i,j}, \nabla_{i,j}$, and $\Delta_{i,j}$, respectively, having compatible sizes (for $i, j \in \{0, 1\}$). Then*

$$\nabla_{A_{ii},B_{jj}}(M_{ij}) = \nabla_{ij} - R_{i,j},$$
$$\Delta_{A_{ii},B_{jj}}(M_{ij}) = \Delta_{ij} + S_{i,j},$$

where

$$R_{i,j} = M_{i,1-j}B_{1-j,j} - A_{i,1-i}M_{1-i,j},$$
$$S_{i,j} = A_{i,i}M_{i,1-j}B_{1-j,j} + A_{i,1-i}M_{1-i,j}B_{j,j} + A_{i,1-i}M_{1-i,1-j}B_{1-j,j},$$

for $i, j \in \{0, 1\}$.

Remark 2.2. *The expressions of Theorem* 2.4 *project the displacement generator of a matrix into those of its blocks so that the projection increases the length of the generator by at most* $\mathrm{rank}(R_{i,j})$ *or* $\mathrm{rank}(S_{i,j})$*, that is, in both cases at most* $\mathrm{rank}(A_{i,1-i}) + \mathrm{rank}(B_{1-j,j})$*. Hereafter (see Definitions* 2.8–2.11*) we only deal with diagonal and unit f-circulant operator matrices A and B; in both cases their blocks $A_{1-i,i}$ and $B_{1-j,j}$ have ranks zero or one (cf. Remark* 2.3*).*

In Section 3 for a nonsingular structured matrix M with $dr(M) = r$ we perform operations with short displacement generators to obtain a displacement generator of length r for its inverse. In the process of computing, the length of the generators can grow above the displacement rank, but then we compress the generators to the rank level based on the following results, valid in any field.

Theorem 2.5. *Given a pair of $n \times l$ matrices G and H, it is sufficient to perform $O(l^2 n)$ ops to compute a pair of $n \times r$ matrices $\tilde{G}$ and $\tilde{H}$ such that $\tilde{G}\tilde{H}^T = GH^T$ where $r = \mathrm{rank}(GH^T) \le l$.*

Proof. See [P01, Theorem 4.6.4]). □

Corollary 2.1. *Given a displacement generator of length l for a displacement operator L and an $n \times n$ matrix M with $dr_L(M) = r$, it is sufficient to use $O(l^2 n)$ ops to compute a displacement generator of length r for the same pair of L and M.*

2.5. Most popular matrices with displacement structure

Toeplitz, Hankel, Vandermonde, and Cauchy matrices have displacement ranks one or two for appropriate operator matrices. These are most used matrices with displacement structure. Next we specify their natural extensions (see some other important classes in [P01, Examples 4.4.8 and 4.4.9]).

Definition 2.8. *$T = (t_{i,j})_{i,j=1}^{n}$ is a Toeplitz matrix if $t_{i,j} = t_{i+1,j+1}$ for every pair of its entries $t_{i,j}$ and $t_{i+1,j+1}$. Such matrix T is f-circulant for a scalar f if $t_{i,j} = f t_{k,l}$ wherever $l - k + n = j - i > 0$. In this case we write $T = Z_f(\mathbf{t}) = \sum_{h=1}^{n} t_h Z_f^{h-1}$ where $\mathbf{t} = (t_h)_{h=1}^{n}$ is the first column of the matrix, $t_h = t_{h,1}$, $h = 1, \dots, n$, and Z_f is the unit f-circulant matrix with the first column $(0, 1, 0, \dots, 0)^T$ and the first row $(0, \dots, 0, f)$. $Z_0(\mathbf{t})$ is the lower triangular Toeplitz matrix with the first column $\mathbf{t}$.*

Definition 2.9. $J = (j_{g,h})_{g,h=0}^{n-1,n-1}$ *is the reflection* (*or the unit Hankel*) *matrix if* $j_{g,n-1-g} = 1$ *for* $g = 0, \dots, n-1$, $j_{g,h} = 0$ *for* $h + g \neq n-1$. (*J* $(v_i)_{i=0}^{n-1} = (v_{n-i-1})_{i=0}^{n-1}$, $J^2 = I$.) $H = (h_{i,j})_{i,j}$ *is a Hankel matrix if* $h_{i,j} = h_{i-1,j+1}$ *for every pair of its entries* $h_{i,j}$ *and* $h_{i-1,j+1}$ *or equivalently if* $H = TJ$ *for a Toeplitz matrix* T.

Definition 2.10. *For a positive integer* n *and a vector* $\mathbf{t} = (t_i)_{i=1}^n$, *define the* $n \times n$ *Vandermonde matrix* $V(\mathbf{t}) = (t_i^{j-1})_{i,j=1}^n$.

Definition 2.11. *For a positive integer* n *and two vectors* $\mathbf{s} = (s_i)_{i=1}^n$ *and* $\mathbf{t} = (t_i)_{i=1}^n$ *such that the* $2n$ *scalars* $\{s_i, t_j\}_{i,j}$ *are distinct, define the* $n \times n$ *Cauchy matrix* $C(\mathbf{s}, \mathbf{t}) = (\frac{1}{s_i - t_j})_{i,j}$.

Fact 2.4. (*See* [P01, Chapters 2 and 3].) *We have* $m_S = O(m(n))$ *if* S *is an* $n \times n$ *Toeplitz or Hankel matrix for* m_S *in Definition* 2.6 *and* $m(n)$ *in Definition* 2.1. $m_S = m_{S^T} = O(m(n) \log n)$ *if* S *is an* $n \times n$ *Vandermonde or Cauchy matrix.*

One can easily verify that Toeplitz, Hankel, Vandermonde, and Cauchy matrices have displacement ranks one or two for appropriate operator matrices.

Fact 2.5. *We have* a) $dr_{Z_e, Z_f^T}(T) \leq 2$ *and* $dr_{Z_e, Z_f}(H) \leq 2$ *for a Toeplitz matrix* T, *a Hankel matrix* H, *and a pair of scalars* e *and* f, $ef \neq 1$, b) $dr_{D(\mathbf{t}), Z_f}(V(\mathbf{t})) = 1$ *for a scalar* f *and a vector* $\mathbf{t} = (t_i)_{i=1}^n$ *such that* $t_i^n f \neq 1$, $i = 1, \dots, n$ *and* c) $dr_{D(\mathbf{s}), D(\mathbf{t})}(C(\mathbf{s}, \mathbf{t})) = 1$ *for a pair of vectors* $\mathbf{s}$ *and* $\mathbf{t}$ *with* $2n$ *distinct entries.*

We use the common nomenclatures of Toeplitz-like, Hankel-like, Vandermonde-like, and Cauchy-like matrices (or the matrices that have the structures of Toeplitz, Hankel, Vandermonde, and Cauchy types, respectively) to define the classes of matrices that have smaller displacement ranks under the same operators whose images have ranks one or two for Toeplitz, Hankel, Vandermonde, and Cauchy matrices, respectively. In view of our results in the previous section, this includes the transposes, blocks and inverses of the latter matrices, as well as the products, sums, and linear combinations of pairs of such matrices for the same or properly reconciled operators (see Theorems 2.2–2.4 and [P01, Chapter 4]).

Next we equivalently define the matrices of these classes (in the memory efficient way) as bilinear expressions via the entries of their displacements generators $G = (\mathbf{g}_i)_{i=1}^r$ and $H = (\mathbf{h}_i)_{i=1}^r$ (cf. [GO94] and [P01, Chapter 4]).

Theorem 2.6. (*Cf.* [KKM79], [P01, Example 4.4.1].) *A matrix* T *has displacement generator* $((\mathbf{g}_i)_i, (\mathbf{h}_i)_i)$ *under the operator* Δ_{Z_e, Z_f^T}, *that is* $T - Z_e T Z_f^T = \sum_{i=1}^r \mathbf{g}_i \mathbf{h}_i^T$ (*for* $2r$ *vectors* $\mathbf{g}_i$ *and* $\mathbf{h}_i$, $i = 1, \dots, r$, *and for a pair of scalars* e *and* f *such that* $ef \neq 1$) *if and only if* $T = \sum_{i=1}^r Z_e(\mathbf{g}_i) Z_f^T(\mathbf{h}_i)$.

Theorem 2.7. *A matrix* H *has displacement generator* $((\mathbf{g}_i)_i, (\mathbf{h}_i)_i)$ *under the operator* Δ_{Z_e, Z_f}, *that is* $H - Z_e H Z_f = \sum_{i=1}^r \mathbf{g}_i \mathbf{h}_i^T$ (*for* $2r$ *vectors* $\mathbf{g}_i$ *and* $\mathbf{h}_i$, $i = 1, \dots, r$, *and for a pair of scalars* e *and* f *such that* $ef \neq 1$) *if and only if* $H = \sum_{i=1}^r Z_e(\mathbf{g}_i) Z_f^T(J\mathbf{h}_i) J$.

Proof. Apply Theorem 2.6 for $T = HJ$ and observe that $JZ_f^T J = Z_f$. □

Theorem 2.8. (*Cf.* [P01, Example 4.4.6b].) *A matrix V has displacement generator $((\mathbf{g}_i)_i, (\mathbf{h}_i)_i)$ under the operator $\Delta_{D(\mathbf{t}),Z_f}$, that is, $V - \operatorname{diag}(\mathbf{t})VZ_f = \sum_{i=1}^r \mathbf{g}_i\mathbf{h}_i^T$) (for a vector $\mathbf{t} = (t_j)_{j=1}^n$ and a scalar f such that $t_i^n f \neq 1$ for $i = 1, \dots, n$) if and only if $V = \sum_{i=1}^r \operatorname{diag}(\frac{1}{1-t_i^n})_{i=1}^n \operatorname{diag}(\mathbf{g}_i)V(\mathbf{t})JZ_f(J\mathbf{h}_i)$.*

Theorem 2.9. (*Cf.* [P01, Example 4.4.1].) *A matrix C has displacement generator $((\mathbf{g}_i)_i, (\mathbf{h}_i)_i)$ under the operator $\nabla_{D(\mathbf{s}),D(\mathbf{t})}$, that is* $\operatorname{diag}(\mathbf{s})C - C\operatorname{diag}(\mathbf{t}) = \sum_{i=1}^r \mathbf{g}_i\mathbf{h}_i^T$ *(for $2r$ vectors $\mathbf{g}_i$ and $\mathbf{h}_i$, $i = 1, \dots, r$ and two vectors $\mathbf{s} = (s_i)_{i=1}^n$ and $\mathbf{t} = (t_i)_{i=1}^n$ such that the $2n$ scalars $\{s_i, t_j\}_{i,j}$ are distinct) if and only if $C = \sum_{i=1}^r \operatorname{diag}(\mathbf{g}_i)C(\mathbf{s}, \mathbf{t}) \operatorname{diag}(\mathbf{h}_i)$.*

Now let us assess our power in dealing with matrices having displacement structure. Theorems 2.2–2.4 enable us to express such matrices and the results of some operations with them in terms of their displacement generators (by using the order of rn parameters versus n^2 entries). Theorems 2.6–2.9 enable us to go back from the displacement to matrices. The bilinear expressions in Theorems 2.6–2.9 reduce multiplication by a vector of a matrix with Toeplitz-like, Hankel--like, Vandermonde-like or Cauchy-like structures essentially to $2r$ multiplications of Toeplitz, Hankel, Vandermonde or Cauchy matrices by $2r$ vectors. Theorems 2.2–2.4 extend this property to transposes, inverses, sums, and products of such matrices as well as their blocks. In particular we specify some respective estimates for the arithmetic complexity in Corollary 2.2 below.

Definition 2.12. *Suppose we have a nonsingular $n \times n$ matrix M preprocessed with a procedure P that outputs ν parameters $p_1, \dots, p_\nu$. Let $i_M(P)$ be the minimum number of ops required to solve a linear system $M\mathbf{x} = \mathbf{f}$ for any vector $\mathbf{f}$ provided a matrix M, its preprocessing P, and the parameters $p_1, \dots, p_\nu$ are fixed. Write $i_M = \min_P\{i_M(P)\}$ where the minimum is over all preprocessings P and write $i_{M,h} = \min_{P(h)}\{i_M(P)\}$ where the minimum is over all preprocessings $P = P(h)$ that amount to solving at most h linear systems of equations with the matrices M, M^T and M^TM.*

Corollary 2.2. *We have $m_T = m_H = O(lm(n))$, $m_V = O(lm(n)\log n)$, $m_C = O(lm(n)\log n)$, $i_{T,2l} = i_{H,2l} = O(lm(n))$, $i_{V,2l} = O(lm(n)\log n)$, and $i_{C,2l} = O(lm(n)\log n)$ where T, H, V, and C stand for $n \times n$ matrices given with their displacement generators of lengths at most l and having structures of Toeplitz, Hankel, Vandermonde, and Cauchy types, respectively.*

The following result enables adjustment of the input displacement structure towards subsequent acceleration of the solution algorithms. In Section 3 we apply this *method of displacement transformation* (proposed in [P89/90] (cf. [P01, Sections 1.7, 4.8, and 4.9])) to accelerate the solution of linear systems with the structures of Vandermonde and Cauchy types by reducing them to linear systems with Toeplitz-like structures.

Corollary 2.3. (*Cf.* [P89/90].) *Given a positive integer r and a displacement generator of a length l for an $n \times n$ matrix M with the structure of Vandermonde (resp. Cauchy) type, one can generate matrices V_1 and V_2 that have structure of Vandermonde type and that are defined with their displacement generators of lengths at most r and then apply $O((l+r)m(n)\log n)$ ops to compute a displacement generator of a length at most $l+r$ (resp. $l+2r$) for a Toeplitz-like matrix V_1M or MV_2 (resp. V_1MV_2).*

Proof. The corollary follows from Theorem 2.3 and Corollary 2.2. □

Recall that in virtue of Theorem 2.1 the transition $\Delta \leftrightarrow \nabla$ between the operators Δ and ∇ cannot seriously affect the displacement rank of a matrix. Likewise the simple estimates below imply that the modification of the parameter f in the operator matrices Z_f little affects the displacement rank of a matrix.

Fact 2.6.

$$|dr_{Z_e,Z_f^T}(T) - dr_{Z_e,Z_g^T}(T)| \le 1,\ |dr_{Z_g,Z_f^T}(T) - dr_{Z_e,Z_f^T}(T)| \le 1, \tag{2.1}$$

$$|dr_{Z_e,Z_f}(H) - dr_{Z_e,Z_g}(H)| \le 1,\ |dr_{Z_g,Z_f}(H) - dr_{Z_e,Z_f}(H)| \le 1 \tag{2.2}$$

for a 5-tuple (T, H, e, f, g).

Remark 2.3. (*Cf. Remark* 2.2.) *The rank of any off-diagonal block is zero for a diagonal matrix and is at most one for any matrix Z_f.*

2.6. Randomization

Unlike *deterministic algorithms*, which always produce correct output, *randomized algorithms* produce correct output with a probability of at least $1-\epsilon$ for any fixed positive tolerance ϵ. The randomized complexity estimates differ depending whether they cover the cost of verification of the correctness of computed solution (e.g., verification that $M\mathbf{x} = \mathbf{f}$ for the computed randomized solution $\mathbf{x}$). If they cover this cost, they are of the *Las Vegas* type. (In this case at the estimated cost one either fails with a low probability or outputs the correct solution.) The other randomized complexity estimates are of the *Monte Carlo* type. (They cover algorithms whose output can be erroneous, but with a bounded low probability.)

Given a nonsingular matrix in $\mathbb{Z}$, what is the probability that it stays such in $\mathbb{Z}_p$ for a random prime p in a fixed large range, e.g., in $(y/20, y]$ for a large integer y? Here is an estimate from [PMRW05], [PW08].

Theorem 2.10. *Suppose that ϵ is a positive number, the matrix $M \in \mathbb{Z}^{n\times n}$ is nonsingular, and a prime p is randomly sampled from the range $(y/20, y]$ under the uniform probability distribution in this range where $y = \frac{n\xi \ln |M|}{\epsilon} \ge 114$, $\xi = \frac{16\ln 114}{16\ln 5.7 - \ln 114} = 16\nu/(1-\nu) = 3.278885\ldots$, and $\nu = \frac{\ln 114}{16\ln 5.7} = 0.17007650\ldots$. Then $P = \text{Probability}((\det M) \bmod p = 0) < \epsilon$.*

3. Computations in $\mathbb{Z}_p$ with matrices having displacement structure

3.1. Inversion of strongly nonsingular structured matrices

Theorem 3.1. *Assume that a strongly nonsingular $n \times n$ matrix M in a field $\mathbb{F}$ has structure of Toeplitz, Hankel, Vandermonde or Cauchy type (cf. Section 2.5), has a displacement rank r, and is given with its displacement generator of a length l. Then a displacement generator of the minimum length r for the matrix M^{-1} as well as the scalar $\det M$ can be computed by using $O(l^2 n + m_M r \log n)$ field operations.*

Proof. The MBA divide-and-conquer algorithm by Morf 1974 and 1980 and Bitmead and Anderson 1980 [M74], [M80], and [BA80] was proposed for Toeplitz--like matrices. We adapt it to a more general class of matrices with displacement structure (cf. [OP98], [P01, Chapter 5]). Recall the well-known block triangular factorizations

$$M = \begin{pmatrix} M_{00} & M_{01} \\ M_{10} & M_{11} \end{pmatrix} = \begin{pmatrix} I & 0 \\ M_{10}M_{00}^{-1} & I \end{pmatrix} \begin{pmatrix} M_{00} & 0 \\ 0 & S \end{pmatrix} \begin{pmatrix} I & M_{00}^{-1}M_{01} \\ 0 & I \end{pmatrix}, \tag{3.1}$$

$$M^{-1} = \begin{pmatrix} \tilde{M}_{00} & \tilde{M}_{01} \\ \tilde{M}_{10} & \tilde{M}_{11} \end{pmatrix} = \begin{pmatrix} I & -M_{00}^{-1}M_{01} \\ 0 & I \end{pmatrix} \begin{pmatrix} M_{00}^{-1} & 0 \\ 0 & S^{-1} \end{pmatrix} \begin{pmatrix} I & 0 \\ -M_{10}M_{00}^{-1} & I \end{pmatrix}. \tag{3.2}$$

Here

$$\tilde{M}_{10} = (-S^{-1})(M_{10}M_{00}^{-1}),\ \tilde{M}_{01} = -(M_{00}^{-1}M_{01})S^{-1}, \tag{3.3}$$

$$\tilde{M}_{00} = M_{00}^{-1} - (M_{00}^{-1}M_{01})\tilde{M}_{10},\ \tilde{M}_{11} = S^{-1}, \tag{3.4}$$

the $k \times k$ block matrix M_{00} and the $(n-k) \times (n-k)$ Schur complement

$$S = S(M, M_{00}) = M_{11} - (M_{10}M_{00}^{-1})M_{01} \tag{3.5}$$

are strongly nonsingular if so is the matrix M, and the sizes $n_i \times n_i$ of the block matrices M_{ii}, $i = 0, 1$ are assumed to be balanced (say, $n_0 = \lceil n/2 \rceil$, $n_1 = n - n_0$). We obtain and recursively extend factorization (3.1) by applying the block Gauss–Jordan elimination to the matrix M and then recursively to the matrices M_{00} and S until the inversion problem is reduced to the case of one-by-one matrices. Actual computation goes back from the inverses of such one-by-one matrices to the matrix M^{-1}. In this recursive process we can also recursively factorize the scalar $\det M = (\det M_{00}) \det S$ and then can compute it in $n-1$ multiplications.

To yield the claimed complexity bound, we maintain and exploit the structure of the input matrix M. In particular, we recursively compress the displacement generator of the matrix M and of all computed auxiliary matrices to the level of their displacement ranks and perform all computations with these matrices by operating with their displacement generators (cf. Theorems 2.2–2.5, Corollary 2.1, and Remark 2.2). Let us examine the associated operator matrices. Assume the pairs of operator matrices (A,B) for an input matrix M and (A_{ii},B_{jj}) for its submatrices M_{ij} for $i, j \in \{0, 1\}$. Combine equations (3.3) and (3.4) with Theorems

2.2–2.4 and obtain the pairs of operator matrices (B_{ii},A_{jj}) for the submatrices $\tilde{M}_{ij}$ of M^{-1} for $i,j \in \{0,1\}$ and hence obtain the pair of operator matrices (B,A) for the matrix M^{-1} such that $dr_{A,B}(M) = dr_{B,A}(M^{-1})$ (cf. Theorem 2.2). Likewise we arrive at consistent pairs of operator matrices for every matrix computed in the forward recursive process and for its inverse computed in the respective backward step. Now we can deduce the complexity bound claimed in Theorem 3.1 in the case of Sylvester displacements by combining Theorems 2.2 and 2.3 and our next result, in which we use Sylvester displacement $\nabla_{A,B}(M)$ (this allows singular operator matrices when we apply Theorem 2.2 for the inverses).

Theorem 3.2. (*Cf. Remark* 3.1.) *Assume Sylvester displacement* $\nabla_{A,B}(M)$. *Then* a) *all trailing principal blocks* (*that is Schur complements*) *computed in the forward recursive process of the adapted MBA algorithm have displacement ranks at most* $r+4$, b) *all leading principal blocks processed in the forward recursive process of the adapted MBA algorithm have displacement ranks at most* $r+6$, *and* c) *all other matrices computed in the forward and backward recursive processes of the adapted MBA algorithm have displacement ranks at most* $2r+12$.

Proof. In the proof we will write $S^{(k)}(M)$ to denote the $(n-k)\times(n-k)$ Schur complement $S(M,M_{00})$ of the $k\times k$ block M_{00} in the $n\times n$ matrix M. We observe that $S^{(g)}(S^{(h)}(M)) = S^{(g+h)}(M)$ because the MBA algorithm is a (structure preserving) variant of the block Gauss–Jordan elimination. Likewise $S^{(h)}(M^{(g+h)}) = (S^{(h)}(M))^{(g)}$. Therefore all trailing principal blocks computed in the MBA forward recursive process are Schur complements in the respective submatrices $M^{(k)}$. Now part a) follows from Theorem 2.4 (applied in the case $i=j$) and Remark 2.2 because the inverse of every Schur complement is a trailing principal (that is, southeastern) block of the inverse of the matrix itself (cf. equation (3.4)) and because $dr(N) = dr(N^{-1})$ (cf. Theorem 2.2). Part b) follows from part a) and Theorem 2.4. Let us prove part c).

In the first step of the forward recursive MBA process we compute the off-diagonal blocks $M_{00}^{-1}M_{01}$ and $M_{10}M_{00}^{-1}$. In the next steps we compute similar products $\widehat{M}_{00}^{-1}\widehat{M}_{01}$ and $\widehat{M}_{01}\widehat{M}_{00}^{-1}$ where the blocks $\widehat{M}_{ij}$ denote the (i,j)th blocks of the respective principal block computed in the previous step of the forward recursive process. As we have observed in the proof of part a), such a principal block is the matrix $S^{(k)}(M^{(h)})^{(g)}$ for some integers g, h and k. Any of its blocks is also a block of the matrices M or $S^{(k)}(M^{(h)})$ for some pair of h and k. Now combining part a) with Theorems 2.2–2.4 implies the bound $dr(B) \le 2r+12$ claimed in part c). Indeed this bound surely covers the factors $M_{10}M_{00}^{-1}$ and $M_{00}^{-1}M_{01}$ of the blocks $\tilde{M}_{10}$ and $\tilde{M}_{01}$ of the inverse M^{-1}, respectively (cf. equations (3.3)), but the same bound is extended to the operands involved in the computation of the northwestern blocks computed in the backward process. Equations (3.4) and the inequalities $dr((M_{00}^{-1}M_{01})\tilde{M}_{00}) \le dr(M_{00}^{-1}M_{01}) + dr(\tilde{M}_{00}))$ support this extension at its final step, and similar relationships support it at the other steps. The stronger upper bound $r+4$ holds for the southeastern blocks computed in the backward

process because they are the inverses of the Schur complements $S^{(k)}(M^{(h)})$ for some integers h and k, and so we can apply Theorems 2.2 and 2.4. □

According to Theorem 3.2, displacement generators of all matrices involved into the MBA process have lengths in $O(r)$. For the final transition from M_{00}^{-1} and S^{-1} to M^{-1} (performed in terms of generators) we use $A = O(rm_M + r^2m(n)) = O(rm_M)$ ops. At the $(i-1)$st preceding level of the recursion we perform similar operations with 2^i matrices of sizes roughly $(n/2^i) \times (n/2^i)$. For a positive constant c and $m_M \geq cn$ this means $O(rm_M)$ ops at each of the $\lceil \log n \rceil$ levels of the MBA backward recursion and thus $O(rm_M \log n)$ ops overall. This completes the proof of Theorem 3.1 under Sylvester displacements. Theorem 2.1 enables extension to Stein displacements $\Delta_{A,B}$. We recall bounds (2.1) and (2.2) to treat the cases where $A, B \in \{Z_0, Z_0^T\}$. □

Remark 3.1. *In the case of the Cauchy-like structure, the MBA recursive process involves only diagonal operator matrices, and so the bounds in Theorem* 3.2 *decrease to* r *in parts* a) *and* b) *and to* $2r$ *in part* c). *We have a smaller decrease in the case of the Vandermonde-like structure, where one half of the operator matrices in the MBA recursive process are diagonal. In the latter case and in the case of Toeplitz-like structure, we can choose the operator matrix* $B{=}Z_0^T$, *to decrease the bounds in Theorem* 3.2 *because the* $(1,0)$*th block of this matrix is filled with zeros and thus has rank zero* (*cf. Remarks* 2.2 *and* 2.3).

Corollary 2.3 reduces the inversion of matrices with the structures of Vandermonde and Cauchy types to the inversion of Toeplitz-like matrices because $M^{-1} = V_2(V_1MV_2)^{-1}V_1$. This implies the following result.

Corollary 3.1. *The upper estimates of Theorem* 3.1 *can be decreased to* $O(l^2n + m(n)r^2 \log n)$ *field operations for* $m(n)$ *in Definition* 2.1, *even for matrices with the structures of Vandermonde and Cauchy types.*

3.2. Inversion of nonsingular structured matrices in $\mathbb{Z}_p$

Corollary 3.2. *Assume a random prime* p *in the range* $(y/20, y]$ *for a sufficiently large integer* y *and a nonsingular matrix* $M \in \mathbb{Z}^{n \times n}$ *having structure of Toeplitz, Hankel, Vandermonde or Cauchy type, given with its displacement generator of a length* l, *and having a displacement rank* r. *Then* a) *the matrix* M^TM *is expected to be strongly nonsingular in* $\mathbb{Z}_p$, *and* b) *if it is strongly nonsingular, then a displacement generator of length* r *for the matrix* $M^{-1} \mod p$ *can be computed by using* $O(l^2n + r^2m(n) \log n)$ *operations in* $\mathbb{Z}_p$.

Proof. The matrix M^TM is strongly nonsingular in $\mathbb{Z}$ (cf. Definition 2.3) and is expected to stay such in $\mathbb{Z}_p$ due to Theorem 2.10. This proves part a). Part b) follows from Corollary 3.1 for $\mathbb{F} = \mathbb{Z}_p$ and from the equation $M^{-1} = (M^TM)^{-1}M^T$. □

3.3. The case of singular matrices with displacement structure

Let us extend our study to the case of singular input matrices M having a rank ρ and a displacement rank r. In this case we seek the inverse of a $\rho \times \rho$ nonsingular submatrix of the matrix M.

Theorem 3.3. *Assume that in a field* $\mathbb{F}$ *an* $n \times n$ *matrix* M *of a rank* $\rho < n$ *has generic rank profile, has structure of* (a) *the Toeplitz or Hankel types or* (b) *Vandermonde or Cauchy types, has a displacement rank* r*, and is given with a displacement generator of a length* $l = O(r)$*. Then* (i) *the rank* ρ *and* (ii) *a displacement generator of length* r *for the matrix* $(M^{(\rho)})^{-1}$ *can be computed by using* $O(rm_{M^{(\rho)}} \log \rho)$ *ops in* $\mathbb{F}$*, that is* $O(m(\rho)r^2 \log^{1+\delta} \rho)$ *ops for* $\delta = 0$ *in case* (a) *and* $\delta = 1$ *in case* (b)*.* (iii) *Within the same cost bound one can compute a solution* $\mathbf{x}$ *to a consistent linear system* $M\mathbf{x} = \mathbf{f}$*, in* $O(m_M)$ *additional ops one can verify consistency of the system, and in* $O(rm_M)$ *additional ops one can compute a shortest displacement generator for a matrix whose columns define a basis for the null space of the matrix* M*.*

Proof. Apply the adapted MBA algorithm as in the case of strongly nonsingular input matrices until it factorizes the submatrix $M^{(\rho)}$ and computes a shortest displacement generator (of length $r + O(1)$) for the matrix $(M^{(\rho)})^{-1}$. This takes $O(rm_{M^{(\rho)}} \log \rho)$ ops overall. Then the algorithm stops because it is prompted to invert the one-by-one matrix filled with the zero. To solve a consistent nonhomogeneous linear system $M\mathbf{x} = \mathbf{f}$, multiply the matrix $(M^{(\rho)})^{-1}$ by the subvector made up of the first ρ coordinates of the vector $\mathbf{f}$ and append $n - \rho$ zero coordinates to the product to obtain a solution vector $\mathbf{x}$. This stage involves $O(m_{M^{(\rho)}})$ ops. To verify consistency of the nonhomogeneous linear system, multiply the matrix M by the vector $\mathbf{x}$ and compare the product with the vector $\mathbf{f}$. In fact one only needs to multiply the $(n - \rho) \times \rho$ southwestern submatrix by the leading subvector of the dimension ρ in the vector $\mathbf{x}$. If $\mathbf{f} = \mathbf{0}$ and if we seek a solution $\mathbf{x} = (x_i)_{i=1}^n$ to the system $M\mathbf{x} = \mathbf{0}$, then we substitute $x_n = 1$ into this system and arrive at a nonhomogeneous linear system with $n - 1$ unknowns and equations. Finally substitute $M_{00} = M^{(\rho)}$ into (3.1) and observe that the columns of the matrix $\begin{pmatrix} (M^{(\rho)})^{-1} M_{01} \\ -I_{n-\rho} \end{pmatrix}$ form a basis for the null space of the matrix M. One can compute a shortest displacement generator for the matrix $(M^{(\rho)})^{-1} M_{01}$ in $O(rm_M)$ additional ops (cf. Theorem 2.3). □

Theorem 3.4. I) *To extend Theorem* 3.3 *to the case of input matrices not having generic rank profile it is sufficient to perform* $O(m_M)$ *additional ops, to generate* $2n - 2$ *random parameters in the field* $\mathbb{F}$*, and to allow Monte Carlo randomization, that is, to allow erroneous output with a low probability.* II) *By performing additional* $O(rm_M)$ *ops one yields Las Vegas randomization, that is, either fails with a low probability or arrives at the correct output.*

Proof. Part I) is implied by the following theorem.

Theorem 3.5. *Let a finite set S of a sufficiently large cardinality $|S|$ lie in a field $\mathbb{F}$ and let a matrix M of a rank $\rho \le n$ lie in $\mathbb{F}^{n\times n}$. Let $\{s_i, t_j,\ i,j = 1,\dots,n\}$ and $\{u_i, v_j,\ i,j = 1,\dots,n\}$ be two sets of $2n$ distinct scalars each. Define randomized preprocessing of the matrix M overwriting it with the matrix XMY (we write $M \leftarrow XMY$) where $X = X_g$, $Y = Y_h$ for $g,h \in \{1,2\}$, $X_1 = (\frac{x_j}{s_i - t_j})_{i=1}^n$, $Y_1 = (\frac{y_j}{u_i - v_j})_{i=1}^n$, $X_2 = Z_0^T(\mathbf{x})$, $Y_2 = Z_0(\mathbf{y})$, $x_1 = y_1 = 1$, and the other $2n-2$ coordinates of the vectors $\mathbf{x} = (x_i)_{i=1}^n$ and $\mathbf{y} = (y_i)_{i=1}^n$ are randomly sampled from the set S. Then both matrices X_2 and Y_2 are nonsingular and with a probability of at least $(1 - n/|S|)^2$ both matrices X_1 and Y_1 are nonsingular. If the matrices X and Y are nonsingular, then with a probability of at least $1 - (\rho+1)\rho/|S|$ matrix XMY has generic rank profile (and therefore is strongly nonsingular if the matrix M is nonsingular).*

Proof. See [KS91] on the case $X = X_2$, $Y = Y_2$ and [P01, Corollary 5.6.3] on the case $X = X_1$, $Y = Y_1$. □

We can extend the displacement structure of the matrix M to the matrix XMY by choosing appropriate matrices $X = X_i$ and $Y = Y_i$ for $i = 1, 2$ to match the operator matrices associated with the matrix M. Then $dr(XMY) \le dr(M)+2$, a shortest displacement generator of the matrix XMY is computed at a low cost (see Theorem 2.3), and so its recursive factorization and a shortest displacement generator for the matrix $M^{-1} = Y(XMY)^{-1}X$ are computed within the cost bounds of Corollary 3.1. This proves part I) of Theorem 3.4. To prove part II), verify correctness of the rank computation as follows: first compute a displacement generator of length $O(r)$ for the Schur complement $S(XMY, (XMY)^{(\rho)})$ of the nonsingular block $(XMY)^{(\rho)}$ in the matrix XMY (cf. equation (3.5) and Theorems 2.2, 2.3, and 3.3), then compress this generator to the minimum length (cf. Corollary 2.1), and finally verify that this length is zero. □

Similarly to Corollary 3.1, we refine the estimates of Theorems 3.3 and 3.4 in the case (b).

Corollary 3.3. *In the upper estimates of Theorems* 3.3 *and* 3.4 *one can replace the bounds $m_{M(\rho)}$ ops with $rm(\rho)$ and m_M with $rm(n)$, even in the case of input matrices with the structures of Vandermonde and Cauchy types, provided that the estimates are increased by $rm(n)\log n$ ops required for computing a shortest displacement generator for the matrix $V_1 M V_2$ where V_1 and V_2 are nonsingular Vandermonde or Vandermonde-like matrices (one of them is replaced with the identity matrix if the input matrix has structure of Vandermonde type).*

Clearly, the results of this subsection can be applied to matrices M in the field $\mathbb{F} = \mathbb{Z}_p^{n\times n}$ for any prime p. Furthermore for a large integer y and a random prime p chosen in the range $(y/20, y]$, a matrix $M \in \mathbb{Z}^{n\times n}$ is likely to keep its rank, displacement rank, and displacement generator in the transition from the ring $\mathbb{Z}$ to the field $\mathbb{Z}_p$ (cf. Theorem 2.10). Therefore the results of this subsection can be extended to structured integer matrices, as we specify next.

4. Computations with structured integer matrices

The algorithms in the previous section compute in $\mathbb{Z}_p$ the determinant and a shortest generator for the inverse of a nonsingular $n \times n$ matrix M with displacement structure. We can repeat this computations for k distinct primes $p_1, \dots, p_k$. The probability of failure and the overall number of ops increase at most by the factor of k. Then the Chinese Remainder Algorithm can produce the determinant and a shortest generator for the inverse modulo $\prod_{i=1}^{k} p_i$. Suppose we let $k = n$ and choose the primes $p_1, \dots, p_n$ at random in the range $(y/20, y]$ for a value

$$y > y_0 = 20(2n)^{1/n} n \max\{\alpha(M)\sqrt{n}, \frac{\xi n}{\epsilon} \ln(n\alpha(M))\} \tag{4.1}$$

for a fixed positive ϵ and $\xi < 4$ in Theorem 2.10. Then we keep the probability of failure within ϵ and obtain $\det M$ and M^{-1} modulo $p_+ = \prod_{i=1}^{n} p_i > 2n(\alpha(M)\sqrt{n})^n$. Fact 2.3 implies that $p_+ > 2|\det M|$, and so we can apply the recipe in Remark 2.1 to yield $\det M$ in $\mathbb{Z}$ and $\operatorname{adj} M$ in $\mathbb{Z}^{n\times n}$. The overall computational cost is dominated at the stage of n applications of our algorithms of the previous section (see [GG03, Theorem 10.25] on the complexity of the Chinese Remainder Algorithm). In particular we arrive at the following result.

Theorem 4.1. *The asymptotic estimates of the previous section can be extended to computing the determinant and a shortest generator for the inverse of a strongly nonsingular $n \times n$ integer matrix M having displacement rank r and given with its displacement generator of a length l. The estimated numbers of random primes and ops increase by the factor n versus the estimates in the previous section. The ops are performed in the fields $\mathbb{Z}_{p_1}, \dots, \mathbb{Z}_{p_n}$ for n distinct primes $p_1, \dots, p_n$ chosen at random independently of each other in the range $(y/20, y]$ where y satisfies* (4.1). *If $l = O(r)$, then the overall cost bound turns into $O(r^2 nm(n) \log n)$ ops performed with the precision of $\log y_0$ bits. This implies the overall Boolean cost bound of $O((r^2 nm(n) \log n)\mu(\log y_0))$ ops for y_0 in* (4.1) *and $m(n)$ and $\mu(d)$ in Definition* 2.1. *The ops are word operations if $\log_2 y_0$ is within the word length. If $\log(1/\epsilon) = O(\log(n\alpha(M)))$, then the upper bound ϵ on the failure probability can be supported for $\log y_0 = O(\log(n\alpha(M)))$. If in addition $\log \alpha(M) = O(\log n)$, then we can choose $\log y_0$ of the order of $O(\log n)$. The techniques in Sections* 3.2 *and* 3.3 *enable extension of all these estimates to computations with any nonsingular and singular integer matrices M having displacement structure at the additional randomized cost specified in Sections* 3.2 *and* 3.3.

5. Related works

Computations with matrices having displacement structure are closely linked to various fundamental polynomial computations [BGY80], [P01], [PMRa]. These links enable extension of our nearly optimal cost bounds from matrix to polynomial computations where the input values are integers. In particular links to

Toeplitz computations enable extension of our cost bounds to the computation of the greatest common divisors, least common multiples, and Padé approximations for a pair of univariate polynomials and further to the Berlekamp–Massey problem of recovering the coefficients of a linear recurrence from its values (see [PMRa]). On the extension to theoretical and practical acceleration of Wiedemann and block Wiedemann algorithms for determinants and Smith's factors of a general matrix, see [P04a] and [PMRa]. For structured matrices M with $\log \alpha(m) = O(\log n)$ and $r = O(1)$, our upper estimates on the Boolean cost are within the factor $\tilde{O}(r^2 \log^2 n)$ from the information lower bound of $n^2 \log n$. The upper bound has been further decreased in [P02], [PMRa] by the factor of $r \log n$ based on combining Hensel's symbolic lifting with numerical iterative refinement and rational number reconstruction from their numerical approximation and from their values modulo a larger integers (cf. [PW02] and [WP03] on the latter subjects). Instead of k random primes for k of the order n, the lifting approach involves just a single prime. Moreover the paper [PMRa] elaborates upon lifting initialized with a power of two instead of customary basic primes p, so that the computations use the binary representation of all operands. The reader is referred to the paper [PMRa] on further details.

Acknowledgement

Criticism by Jesse Wolf helped us to detect and to fix a number of defects in the original draft of this paper.

References

[ABM99] J. Abbott, M. Bronstein, T. Mulders. Fast Deterministic Computation of the Determinants of Dense Matrices, *Proc. Intern. Symp. Symbolic and Algebraic Comput. (ISSAC'99)*, 197–204, ACM Press, New York, 1999.

[B85] J.R. Bunch, Stability of Methods for Solving Toeplitz Systems of Equations, *SIAM J. Sci. Stat. Comput.*, **6(2)**, 349–364, 1985.

[B03] D.J. Bernstein, Fast Multiplication and Its Applications, to be printed in *Algorithmic Number Theory* (edited by Joe Buhler, Peter Stevenhagen), **44**, Mathematical Sciences Research Institute Publications, Cambridge University Press. Available from http://cr.yp.to/papers.html

[BA80] R.R. Bitmead, B.D.O. Anderson, Asymptotically Fast Solution of Toeplitz and Related Systems of Linear Equations, *Linear Algebra and Its Applications*, **34**, 103–116, 1980.

[BGY80] R.P. Brent, F.G. Gustavson, D.Y.Y. Yun, Fast Solution of Toeplitz Systems of Equations and Computation of Padé Approximations, *J. Algorithms*, **1**, 259–295, 1980.

[CK91] D.G. Cantor, E. Kaltofen, On Fast Multiplication of Polynomials over Arbitrary Rings, *Acta Informatica*, **28(7)**, 697–701, 1991.

[CW90] D. Coppersmith, S. Winograd, Matrix Multiplication via Arithmetic Progressions. *J. Symbolic Comput.*, **9(3)**, 251–280, 1990.

[D59] J. Durbin, The Fitting of Time-Series Models, *Review of International Statistical Institute,* **28**, 229–249, 1959.

[F07] M. Fürer, Faster Integer Multiplication, *Proceedings of* 39*th Annual Symposium on Theory of Computing* (*STOC* 2007), 57–66, ACM Press, New York, 2007.

[GG03] J. von zur Gathen, J. Gerhard, *Modern Computer Algebra*, Cambridge University Press, Cambridge, UK, 2003 (second edition).

[GI88] W. Gautschi, G. Inglese, Lower Bounds for the Condition Number of Vandermonde Matrices, *Numerische Math.*, **52**, 241–250, 1988.

[GO94] I. Gohberg, V. Olshevsky, Complexity of Multiplication with Vectors for Structured Matrices, *Linear Algebra and Its Applications*, **202**, 163–192, 1994.

[H79] G. Heinig, Beiträge zur Spektraltheorie von Operatorbüscheln und zur algebraischen Theorie von Toeplitzmatrizen, Dissertation **B**, *TH Karl-Marx-Stadt*, 1979.

[HR84] G. Heinig, K. Rost, *Algebraic Methods for Toeplitz-like Matrices and Operators, Operator Theory*, **13**, Birkhäuser, 1984.

[K98] D.E. Knuth, *The Art of Computer Programming*, **2**: *Seminumerical Algorithms*, Addison-Wesley, Reading, Massachusetts, 1998.

[K04] I. Kaporin, The Aggregation and Cancellation Techniques As a Practical Tool for Faster Matrix Multiplication, *Theoretical Computer Science*, **315**, 469–510, 2004.

[KKM79] T. Kailath, S.Y. Kung, M. Morf, Displacement Ranks of Matrices and Linear Equations, *Journal Math. Analysis and Appls.*, **68(2)**, 395–407, 1979.

[KS91] E. Kaltofen, B.D. Saunders, On Wiedemann's Method for Solving Sparse Linear Systems, *Proceedings of AAECC–*5*, Lecture Notes in Computer Science*, **536**, 29–38, Springer, Berlin, 1991.

[L47] N. Levinson, The Wiener RMS (Root-Mean-Square) Error Criterion in the Filter Design and Prediction, *Journal of Mathematical Physics*, **25**, 261–278, 1947.

[M74] M. Morf, Fast Algorithms for Multivariable Systems, Ph.D. Thesis, *Dept. Electrical Engineering, Stanford Univ.,* CA, 1974.

[M80] M. Morf, Doubling Algorithms for Toeplitz and Related Equations, *Proceedings of IEEE International Conference on ASSP*, 954–959, IEEE Press, Piscataway, New Jersey, 1980.

[OP98] V. Olshevsky, V.Y. Pan, A Unified Superfast Algorithm for Boundary Rational Tangential Interpolation Problem and for Inversion and Factorization of Dense Structured Matrices, *Proc. 39th Ann. IEEE Symp. Foundation of Comp. Science* (*FOCS* 98), 192–201, IEEE Computer Soc. Press, Los Alamitos, CA, 1998.

[P84] V.Y. Pan, How Can We Speed up Matrix Multiplication?, *SIAM Review*, **26(3)**, 393–415, 1984.

[P89/90] V.Y. Pan, On Computations with Dense Structured Matrices, *Math. of Computation*, **55(191)**, 179–190, 1990. Proceedings version in *Proc. ISSAC'*89, 34-42, ACM Press, New York, 1989.

[P01] V.Y. Pan, *Structured Matrices and Polynomials: Unified Superfast Algorithms*, Birkhäuser/Springer, Boston/New York, 2001.

[P02] V.Y. Pan, Can We Optimize Toeplitz/Hankel Computations? *Proc. of the Fifth International Workshop on Computer Algebra in Scientific Computing* (*CASC*'02), Yalta, Crimea, Sept. 2002 (E.W. Mayr, V.G. Ganzha, E.V. Vorozhtzov, Editors), 253–264, *Technische Universität München*, Germany, 2002.

[P04a] V.Y. Pan, On Theoretical and Practical Acceleration of Randomized Computation of the Determinant of an Integer Matrix, *Zapiski Nauchnykh Seminarov POMI* (in English), **316**, 163–187, St. Petersburg, Russia, 2004. Also available at http://comet.lehman.cuny.edu/vpan/

[PMRa] V.Y. Pan, B. Murphy, R.E. Rosholt, Unified Nearly Optimal Algorithms for Structured Integer Matrices and Polynomials, Technical Report 2008003, *PhD Program in Computer Science, The Graduate Center, CUNY*, New York, 2008. Available at http://www.cs.gc.cuny.edu/tr/techreport.php?id=352

V.Y. Pan, B. Murphy, R.E. Rosholt, Nearly Optimal Symbolic-Numerical Algorithms for Structured Integer Matrices and Polynomials, *Proceedings of International Symposium on Symbolic-Numerical Computations* (Kyoto, Japan, August 2009), (edited by Hiroshi Kai and Hiroshi Sekigawa), pp. 105–113, ACM Press, New York, 2009

[PMRW05] V.Y. Pan, B. Murphy, R.E. Rosholt, X. Wang, Toeplitz and Hankel Meet Hensel and Newton Modulo a Power of Two, Technical Report 2005008, *PhD Program in Computer Science, The Graduate Center, CUNY*, New York, 2005. Available at http://www.cs.gc.cuny.edu/tr/techreport.php?id=352

[PW02] V.Y. Pan, X. Wang, Acceleration of Euclidean Algorithm and Extensions, *Proc. Intern. Symp. Symbolic and Algebraic Computation* (*ISSAC*'02), 207–213, ACM Press, New York, 2002.

[PW08] V.Y. Pan, X. Wang, Degeneration of Integer Matrices Modulo an Integer, *Linear Algebra and Its Application*, **429**, 2113–2130, 2008.

[T64] W.F. Trench, An Algorithm for Inversion of Finite Toeplitz Matrices, *J. of SIAM*, **12**, 515–4522, 1964.

[T94] E.E. Tyrtyshnikov, How Bad Are Hankel Matrices? *Numerische Mathematik*, **67(2)**, 261–269, 1994.

[WP03] X. Wang, V.Y. Pan, Acceleration of Euclidean Algorithm and Rational Number Reconstruction, *SIAM J. on Computing*, **32(2)**, 548–556, 2003.

Victor Y. Pan
Department of Mathematics and Computer Science
Lehman College of the City University of New York
Bronx, NY 10468, USA

http://comet.lehman.cuny.edu/vpan/
e-mail: victor.pan@lehman.cuny.edu

Brian J. Murphy and Rhys Eric Rosholt
Department of Mathematics and Computer Science
Lehman College of the City University of New York
Bronx, NY 10468, USA
e-mail: brian.murphy@lehman.cuny.edu
rhys.rosholt@lehman.cuny.edu

Operator Theory:
Advances and Applications, Vol. 199, 377–396

V-cycle Optimal Convergence for DCT-III Matrices

C. Tablino Possio

Dedicated to Georg Heinig

Abstract. The paper analyzes a two-grid and a multigrid method for matrices belonging to the DCT-III algebra and generated by a polynomial symbol. The aim is to prove that the convergence rate of the considered multigrid method (V-cycle) is constant independent of the size of the given matrix. Numerical examples from differential and integral equations are considered to illustrate the claimed convergence properties.

Mathematics Subject Classification (2000). Primary 65F10, 65F15, 15A12.

Keywords. DCT-III algebra, two-grid and multigrid iterations, multi-iterative methods.

1. Introduction

In the last two decades, an intensive work has concerned the numerical solution of structured linear systems of large dimensions [6, 14, 16]. Many problems have been solved mainly by the use of (preconditioned) iterative solvers. However, in the multilevel setting, it has been proved that the most popular matrix algebra preconditioners are not effective in general (see [23, 26, 20] and references therein). On the other hand, the multilevel structures often are the most interesting in practical applications. Therefore, quite recently, more attention has been focused (see [2, 1, 7, 5, 28, 9, 12, 10, 13, 22, 25, 19]) on the multigrid solution of multilevel structured (Toeplitz, circulants, Hartley, sine (τ class) and cosine algebras) linear systems in which the coefficient matrix is banded in a multilevel sense and positive definite. The reason is due to the fact that these techniques are very efficient, the total cost for reaching the solution within a preassigned accuracy being linear in the dimensions of the involved linear systems.

The work of the author was partially supported by MIUR, grant number 2006017542.

In this paper we deal with the case of matrices generated by a polynomial symbol and belonging to the DCT-III algebra. This kind of matrices appears in the solution of differential equations and integral equations, see for instance [4, 18, 24]. In particular, they directly arise in certain image restoration problems or can be used as preconditioners for more complicated problems in the same field of application [17, 18].

We consider the Two-Grid (TGM)/Multi-Grid (MGM) Method proposed and analyzed in [7] in terms of the algebraic multigrid theory developed by Ruge and Stüben [21] (for the foundation of the theory see [11] and [29]). Our aim is to provide general conditions under which the proposed MGM results to be optimally convergent with a convergence rate independent of the matrix size for a large class of matrices in the DCT-III algebra. We prove that for this class of matrices, the MGM results to be optimal in the sense of Definition 1.1 below, i.e., the problem of solving a linear system with coefficient matrix A is asymptotically of the same cost as the direct problem of multiplying A by a vector.

Definition 1.1. [3] Let $\{A_m x_m = b_m\}$ be a given sequence of linear systems with A_m of size m. An iterative method for solving the systems $A_m x_m = b_m$, $m \in \mathbb{N}^+$, is *optimal* if

1. the arithmetic cost of each iteration is at most proportional to the complexity of a matrix vector product with matrix A_m,
2. the number of iterations for reaching the solution within a fixed accuracy can be bounded from above by a constant independent of m.

A crucial role to prove our optimality result is played by the choice of the projection operator P_{s+1}^s which is used in the MGM to project from the level s to the level $s+1$. In fact, the total cost of the proposed MGM will be of $O(m)$ operations since for any coarse level s we can find a projection operator P_{s+1}^s such that

- the matrix vector product involving P_{s+1}^s costs $O(m_s)$ operations where $m_s = m/2^s$;
- the coarse grid matrix $A_{m_{s+1}} = P_{s+1}^s A_{m_s} (P_{s+1}^s)^T$ is also a matrix in the DCT-III algebra generated by a polynomial symbol and can be formed within $O(m_s)$ operations;
- the convergence rate of the MGM is independent of m.

The paper is organized as follows. In Section 2 we briefly recall TGM and MGM (standard V-cycle) and report the main tools regarding the convergence theory of algebraic multigrid methods [21]. In Section 3 we consider the TGM for matrices belonging to DCT-III algebra with reference to some optimal convergence properties, while Section 4 is devoted to the convergence analysis of its natural extension as V-cycle. Numerical evidences of the theoretical results are reported and discussed in Section 5, while Section 6 concerns complexity issues and conclusions.

2. Two-grid and multi-grid methods

In this section we briefly report the main results pertaining to the convergence theory of algebraic multigrid methods.

Let us consider the generic linear system $A_m x_m = b_m$ with large dimension m, where $A_m \in \mathbb{C}^{m\times m}$ is a Hermitian positive definite matrix and $x_m, b_m \in \mathbb{C}^m$. Let

$$m_0 = m > m_1 > \cdots > m_s > \cdots > m_{s_{\min}}$$

and let

$$P_{s+1}^s \in \mathbb{C}^{m_{s+1}\times m_s}$$

be a given full-rank matrix for any s. Lastly, let us denote by $\mathcal{V}_s$ a class of stationary iterative methods for linear systems of dimension m_s.

According to [11], the algebraic Two-Grid Method (TGM) is an iterative method whose generic step is defined as follow.

$$x_s^{\text{out}} = \mathcal{TGM}(s, x_s^{\text{in}}, b_s)$$

$x_s^{\text{pre}} = \mathcal{V}_{s,\text{pre}}^{\nu_{\text{pre}}}(x_s^{\text{in}}, b_s)$	Pre-smoothing iterations
$r_s = A_s x_s^{\text{pre}} - b_s$ $r_{s+1} = P_{s+1}^s r_s$ $A_{s+1} = P_{s+1}^s A_s (P_{s+1}^s)^H$ Solve $A_{s+1} y_{s+1} = r_{s+1}$ $\hat{x}_s = x_s^{\text{pre}} - (P_{s+1}^s)^H y_{s+1}$	Exact Coarse Grid Correction
$x_s^{\text{out}} = \mathcal{V}_{s,\text{post}}^{\nu_{\text{post}}}(\hat{x}_s, b_s)$	Post-smoothing iterations

where we refer to the dimension m_s by means of its subscript s.

In the first and last steps a *pre-smoothing iteration* and a *post-smoothing iteration* are applied ν_{pre} times and ν_{post} times, respectively, according to the chosen stationary iterative method in the class $\mathcal{V}_s$.

Moreover, the intermediate steps define the so-called *exact coarse grid correction operator*, that depends on the considered projector operator P_{s+1}^s.

The global iteration matrix of the TGM is then given by

$$TGM_s = V_{s,\text{post}}^{\nu_{\text{post}}} CGC_s V_{s,\text{pre}}^{\nu_{\text{pre}}}, \tag{2.1}$$

$$CGC_s = I_s - (P_{s+1}^s)^H A_{s+1}^{-1} P_{s+1}^s A_s \quad A_{s+1} = P_{s+1}^s A_s (P_{s+1}^s)^H, \tag{2.2}$$

where $V_{s,\text{pre}}$ and $V_{s,\text{post}}$ denote the pre-smoothing and post-smoothing iteration matrices, respectively.

By means of a recursive procedure, the TGM gives rise to a Multi-Grid Method (MGM): the standard V-cycle is defined as follows.

$$x_s^{\text{out}} = \mathcal{MGM}(s, x_s^{\text{in}}, b_s)$$

`if` $s \le s_{\min}$ `then`

$\text{Solve } A_s x_s^{\text{out}} = b_s$	Exact solution

`else`

$x_s^{\text{pre}} = \mathcal{V}_{s,\text{pre}}^{\nu_{\text{pre}}}(x_s^{\text{in}}, b_s)$	Pre-smoothing iterations
$r_s = A_s x_s^{\text{pre}} - b_s$ $r_{s+1} = P_{s+1}^s r_s$ $y_{s+1} = \mathcal{MGM}(s+1, \mathbf{0}_{s+1}, r_{s+1})$ $\hat{x}_s = x_s^{\text{pre}} - (P_{s+1}^s)^H y_{s+1}$	Coarse Grid Correction
$x_s^{\text{out}} = \mathcal{V}_{s,\text{post}}^{\nu_{\text{post}}}(\hat{x}_s, b_s)$	Post-smoothing iterations

Notice that in MGM the matrices $A_{s+1} = P_{s+1}^s A_s (P_{s+1}^s)^H$ are more profitably formed in the so-called *setup phase* in order to reduce the computational costs.

The global iteration matrix of the MGM can be recursively defined as

$$\begin{aligned} MGM_{s_{\min}} &= O \in \mathbb{C}^{s_{\min} \times s_{\min}}, \\ MGM_s &= V_{s,\text{post}}^{\nu_{\text{post}}} \left[I_s - (P_{s+1}^s)^H \left(I_{s+1} - MGM_{s+1} \right) A_{s+1}^{-1} P_{s+1}^s A_s \right] V_{s,\text{pre}}^{\nu_{\text{pre}}}, \\ & \qquad\qquad s = s_{\min} - 1, \ldots, 0. \end{aligned}$$

Hereafter, by $\|\cdot\|_2$ we denote the Euclidean norm on $\mathbb{C}^m$ and the associated induced matrix norm over $\mathbb{C}^{m \times m}$. If X is positive definite, $\|\cdot\|_X = \|X^{1/2} \cdot\|_2$ denotes the Euclidean norm weighted by X on $\mathbb{C}^m$ and the associated induced matrix norm. Finally, if X and Y are Hermitian matrices, then the notation $X \le Y$ means that $Y - X$ is nonnegative definite.

Some general conditions that ensure the convergence of an algebraic TGM and MGM are due to Ruge and Stüben [21].

Theorem 2.1 (TGM convergence [21]**).** *Let m_0, m_1 be integers such that $m_0 > m_1 > 0$, let $A \in \mathbb{C}^{m_0 \times m_0}$ be a positive definite matrix. Let $\mathcal{V}_0$ be a class of iterative methods for linear systems of dimension m_0 and let $P_1^0 \in \mathbb{C}^{m_1 \times m_0}$ be a given full-rank matrix. Suppose that there exist $\alpha_{\text{post}} > 0$ independent of m_0 such that*

$$\|V_{0,\text{post}}\, x\|_A^2 \le \|x\|_A^2 - \alpha_{\text{post}} \|x\|_{AD^{-1}A}^2 \quad \textit{for any } x \in \mathbb{C}^{m_0} \tag{2.3}$$

(*where D is the diagonal matrix formed by the diagonal entries of A*) *and that there exists $\gamma > 0$ independent of m_0 such that*

$$\min_{y \in \mathbb{C}^{m_1}} \|x - (P_1^0)^H y\|_D^2 \leq \gamma \|x\|_A^2 \quad \text{for any } x \in \mathbb{C}^{m_0}. \tag{2.4}$$

Then, $\gamma \geq \alpha_{\text{post}}$ and

$$\|TGM_0\|_A \leq \sqrt{1 - \alpha_{\text{post}}/\gamma}. \tag{2.5}$$

It is worth stressing that in Theorem 2.1 the matrix $D \in \mathbb{C}^{m_0 \times m_0}$ can be substituted by any Hermitian positive definite matrix X: clearly the choice $X = I$ can give rise to valuable simplifications [2].

At first sight, the MGM convergence requirements are more severe since the smoothing and CGC iteration matrices are linked in the same inequalities as stated below.

Theorem 2.2 (MGM convergence [21]**).** *Let $m_0 = m > m_1 > m_2 > \cdots > m_s > \cdots > m_{s_{\min}}$ and let $A \in \mathbb{C}^{m \times m}$ be a positive definite matrix. Let $P_{s+1}^s \in \mathbb{C}^{m_{s+1} \times m_s}$ be full-rank matrices for any level s. Suppose that there exist $\delta_{\text{pre}} > 0$ and $\delta_{\text{post}} > 0$ such that*

$$\|V_{s,\text{pre}}^{\nu_{\text{pre}}} x\|_{A_s}^2 \leq \|x\|_{A_s}^2 - \delta_{\text{pre}} \|CGC_s V_{s,\text{pre}}^{\nu_{\text{pre}}} x\|_{A_s}^2 \quad \text{for any } x \in \mathbb{C}^{m_s} \tag{2.6a}$$

$$\|V_{s,\text{posr}}^{\nu_{\text{post}}} x\|_{A_s}^2 \leq \|x\|_{A_s}^2 - \delta_{\text{post}} \|CGC_s x\|_{A_s}^2 \quad \text{for any } x \in \mathbb{C}^{m_s} \tag{2.6b}$$

both for each $s = 0, \ldots, s_{\min} - 1$, then $\delta_{\text{post}} \leq 1$ and

$$\|MGM_0\|_A \leqslant \sqrt{\frac{1 - \delta_{\text{post}}}{1 + \delta_{\text{pre}}}} < 1. \tag{2.7}$$

By virtue of Theorem 2.2, the sequence $\{x_m^{(k)}\}_{k \in \mathbb{N}}$ will converge to the solution of the linear system $A_m x_m = b_m$ and within a constant error reduction not depending on m and $s_{\min}$ if at least one between δ_{pre} and δ_{post} is independent of m and $s_{\min}$.

Nevertheless, as also suggested in [21], the inequalities (2.6a) and (2.6b) can be split as

$$\begin{cases} \|V_{s,\text{pre}}^{\nu_{\text{pre}}} x\|_{A_s}^2 & \leq \|x\|_{A_s}^2 - \alpha \|V_{s,\text{pre}}^{\nu_{\text{pre}}} x\|_{A_s D_s^{-1} A_s} \\ \|CGC_s x\|_{A_s}^2 & \leq \gamma \|x\|_{A_s D_s^{-1} A_s}^2 \\ \delta_{\text{pre}} = \alpha/\gamma \end{cases} \tag{2.8}$$

and

$$\begin{cases} \|V_{s,\text{post}}^{\nu_{\text{post}}} x\|_{A_s}^2 & \leq \|x\|_{A_s}^2 - \beta \|x\|_{A_s D_s^{-1} A_s}^2 \\ \|CGC_s\, x\|_{A_s}^2 & \leq \gamma \|x\|_{A_s D_s^{-1} A_s}^2 \\ \delta_{\text{post}} = \beta/\gamma \end{cases} \tag{2.9}$$

where D_s is the diagonal matrix formed by the diagonal entries of A_s (again, the $AD^{-1}A$-norm is not compulsory [2] and the A^2-norm will be considered in the following) and where, more importantly, the coefficients α, β and γ can differ in each recursion level s since the step from (2.8) to (2.6a) and from (2.9) to (2.6b) are purely algebraic and does not affect the proof of Theorem 2.2.

Therefore, in order to prove the V-cycle optimal convergence, it is possible to consider the inequalities

$$\|V_{s,\mathrm{pre}}^{\nu_{\mathrm{pre}}} x\|_{A_s}^2 \leq \|x\|_{A_s}^2 - \alpha_s \|V_{s,\mathrm{pre}}^{\nu_{\mathrm{pre}}} x\|_{A_s^2}^2 \qquad \text{for any } x \in \mathbb{C}^{m_s} \tag{2.10a}$$

$$\|V_{s,\mathrm{post}}^{\nu_{\mathrm{post}}} x\|_{A_s}^2 \leq \|x\|_{A_s}^2 - \beta_s \|x\|_{A_s^2}^2 \qquad \text{for any } x \in \mathbb{C}^{m_s} \tag{2.10b}$$

$$\|CGC_s x\|_{A_s}^2 \leq \gamma_s \|x\|_{A_s^2}^2 \qquad \text{for any } x \in \mathbb{C}^{m_s}. \tag{2.10c}$$

where it is required that α_s, β_s, $\gamma_s \geq 0$ for each $s = 0, \ldots, s_{\min} - 1$ and

$$\delta_{\mathrm{pre}} = \min_{0 \leq s < s_{\min}} \frac{\alpha_s}{\gamma_s}, \qquad \delta_{\mathrm{post}} = \min_{0 \leq s < s_{\min}} \frac{\beta_s}{\gamma_s}. \tag{2.11}$$

We refer to (2.10a) as the *pre-smoothing property*, (2.10b) as the *post-smoothing property* and (2.10c) as the *approximation property* (see [21]).

An evident benefit in considering the inequalities (2.10a)–(2.10c) relies on the fact that the analysis of the smoothing iterations is distinguished from the more difficult analysis of the projector operator. Moreover, the MGM smoothing properties (2.10a) and (2.10b) represent a generalization of the TGM smoothing property (2.3) with D substituted by I, in accordance with the previous reasoning.

3. Two-grid and multi-grid methods for DCT-III matrices

Let $\mathcal{C}_m = \{C_m \in \mathbb{R}^{m \times m} | C_m = Q_m D_m Q_m^T\}$ the one-level DCT-III cosine matrix algebra, i.e., the algebra of matrices that are simultaneously diagonalized by the orthogonal transform

$$Q_m = \left[\sqrt{\frac{2 - \delta_{j,1}}{m}} \cos \left\{ \frac{(i-1)(j-1/2)\pi}{m} \right\} \right]_{i,j=1}^{m} \tag{3.1}$$

with $\delta_{i,j}$ denoting the Kronecker symbol.

Let f be a real-valued even trigonometric polynomial of degree k and period 2π. Then, the DCT-III matrix of order m generated by f is defined as

$$C_m(f) = Q_m D_m(f) Q_m^T, \quad D_m(f) = \mathrm{diag}_{1 \leq j \leq m} f\left(x_j^{[m]}\right), \quad x_j^{[m]} = \frac{(j-1)\pi}{m}.$$

Clearly, $C_m(f)$ is a symmetric band matrix of bandwidth $2k+1$. In the following, we denote in short with $C_s = C_{m_s}(g_s)$ the DCT-III matrix of size m_s generated by the function g_s.

An algebraic TGM/MGM method for (multilevel) DCT-III matrices generated by a real-valued even trigonometric polynomial has been proposed in [7]. Here, we briefly report the relevant results with respect to TGM convergence analysis, the aim being to prove in Section 4 the V-cycle optimal convergence under suitable conditions.

Indeed, the projector operator P_{s+1}^s is chosen as

$$P_{s+1}^s = T_{s+1}^s C_s(p_s)$$

where $T_{s+1}^s \in \mathbb{R}^{m_{s+1}\times m_s}$, $m_{s+1} = m_s/2$, is the cutting operator defined as

$$\left[T_{s+1}^s\right]_{i,j} = \begin{cases} 1/\sqrt{2} & \text{for } \ j \in \{2i-1, 2i\},\ i = 1, \ldots, m_{s+1}, \\ 0 & \text{otherwise,} \end{cases} \tag{3.2}$$

and $C_s(p_s)$ is the DCT-III cosine matrix of size m_s generated by a suitable even trigonometric polynomial p_s. Here, the scaling by a factor $1/\sqrt{2}$ is introduced in order to normalize the matrix T_{s+1}^s with respect to the Euclidean norm. From the point of view of an algebraic multigrid this is a natural choice, while in a geometric multigrid it is more natural to consider just a scaling by $1/2$ in the projector, to obtain an average value.

The cutting operator plays a leading role in preserving both the structural and spectral properties of the projected matrix C_{s+1}: in fact, it ensures a spectral link between the space of the frequencies of size m_s and the corresponding space of frequencies of size m_{s+1}, according to the following Lemma.

Lemma 3.1. [7] *Let $Q_s \in \mathbb{R}^{m_s\times m_s}$ and $T_{s+1}^s \in \mathbb{R}^{m_{s+1}\times m_s}$ be given as in* (3.1) *and* (3.2)*, respectively. Then*

$$T_{s+1}^s Q_s = Q_{s+1}[\Phi_{s+1}, \Theta_{s+1}\Pi_{s+1}], \tag{3.3}$$

where

$$\Phi_{s+1} = \mathrm{diag}_{j=1,\ldots,m_{s+1}}\left[\cos\left(\frac{1}{2}\left(\frac{x_j^{[m_s]}}{2}\right)\right)\right], \quad x_j^{[m_s]} = \frac{(j-1)\pi}{m_s}, \tag{3.4a}$$

$$\Theta_{s+1} = \mathrm{diag}_{j=1,\ldots,m_{s+1}}\left[-\cos\left(\frac{1}{2}\left(\frac{x_j^{[m_s]}}{2} + \frac{\pi}{2}\right)\right)\right], \tag{3.4b}$$

and $\Pi_{s+1} \in \mathbb{R}^{m_{s+1}\times m_{s+1}}$ is the permutation matrix given by

$$\begin{bmatrix} 1 & 0 & \ldots & 0 \\ 0 & 0 & & 1 \\ \vdots & & \iddots & \\ 0 & 1 & & 0 \end{bmatrix}.$$

As a consequence, let $A_s = C_s(f_s)$ be the DCT-III matrix generated by f_s, then

$$A_{s+1} = P_{s+1}^s A_s (P_{s+1}^s)^T = C_{s+1}(f_{s+1})$$

where

$$\begin{aligned} f_{s+1}(x) \ = \ & \cos^2\left(\frac{x/2}{2}\right) f_s\left(\frac{x}{2}\right) p_s^2\left(\frac{x}{2}\right) \\ & + \cos^2\left(\frac{\pi - x/2}{2}\right) f_s\left(\pi - \frac{x}{2}\right) p_s^2\left(\pi - \frac{x}{2}\right), \quad x \in [0, \pi]. \end{aligned} \tag{3.5}$$

On the other side, the convergence of the proposed TGM at size m_s is ensured by choosing the polynomial as follows. Let $x^0 \in [0, \pi)$ be a zero of the generating function f_s. The polynomial p_s is chosen so that

$$\lim_{x \to x^0} \frac{p_s^2(\pi - x)}{f_s(x)} < +\infty, \tag{3.6a}$$

$$p_s^2(x) + p_s^2(\pi - x) > 0. \tag{3.6b}$$

In the special case $x^0 = \pi$, the requirement (3.6a) is replaced by

$$\lim_{x \to x^0 = \pi} \frac{p_s^2(\pi - x)}{\cos^2\left(\frac{x}{2}\right)\, f_s(x)} < +\infty. \tag{3.7a}$$

If f_s has more than one zero in $[0, \pi]$, then p_s will be the product of the polynomials satisfying the condition (3.6a) (or (3.7a)) for every single zero and globally the condition (3.6b). It is evident that the polynomial p_s must have zeros of proper order in any mirror point $\hat{x}^0 = \pi - x^0$, where x^0 is a zero of f_s.

It is worth stressing that conditions (3.6a) and (3.6b) are in perfect agreement with the case of other structures such as τ, symmetric Toeplitz and circulant matrices (see, e.g., [22, 25]), while the condition (3.7a) is proper of the DCT-III algebra and it corresponds to a worsening of the convergence requirements.

Moreover, as just suggested in [7], in the case $x^0 = 0$ the condition (3.6a) can also be weakened as

$$\lim_{x \to x^0 = 0} \frac{\cos^2\left(\frac{\pi - x}{2}\right)\, p_s^2(\pi - x)}{f_s(x)} < +\infty. \tag{3.8a}$$

We note that if f_s is a trigonometric polynomial of degree k, then f_s can have a zero of order at most $2k$. If $f_s(\pi) \neq 0$, then by (3.6a) the degree of p_s has to be less than or equal to $\lceil k/2 \rceil$. If $f_s(\pi) = 0$, then the degree of p_s is less than or equal to $\lceil (k+1)/2 \rceil$.

Notice also that from (3.5) it is easy to obtain the Fourier coefficients of f_{s+1} and hence the nonzero entries of $A_{s+1} = C_{s+1}(f_{s+1})$. In addition, we can obtain the zeros of f_{s+1} and their orders by knowing the zeros of f_s and their orders.

Lemma 3.2. [7] *If $0 \leq x^0 \leq \pi/2$ is a zero of f_s, then by* (3.6a), *$p_s(\pi - x^0) = 0$ and hence by* (3.5), *$f_{s+1}(2x^0) = 0$, i.e., $y^0 = 2x^0$ is a zero of f_{s+1}. Furthermore, because $p_s(\pi - x^0) = 0$, by* (3.6b), *$p_s(x^0) > 0$ and hence the orders of x^0 and y^0 are the same. Similarly, if $\pi/2 \leq x^0 < \pi$, then $y^0 = 2(\pi - x^0)$ is a zero of f_{s+1} with the same order as x^0. Finally, if $x^0 = \pi$, then $y^0 = 0$ with order equal to the order of x^0 plus two.*

In [7] the Richardson method has be considered as the most natural choice for the smoothing iteration, since the corresponding iteration matrix $V_m := I_m - \omega A_m \in \mathbb{C}^{m \times m}$ belongs to the DCT-III algebra, too. Further remarks about such a type of smoothing iterations and the tuning of the parameter ω are reported in [25, 1].

Theorem 3.3. [7] *Let $A_{m_0} = C_{m_0}(f_0)$ with f_0 being a nonnegative trigonometric polynomial and let $V_{m_0} = I_{m_0} - \omega A_{m_0}$ with $\omega = 1/\|f_0\|_\infty$ for the post-smoothing iteration. Then, under the quoted assumptions and definitions the inequalities* (2.3) *and* (2.4) *hold true, and the proposed TGM converges linearly.*

Here, it could be interesting to come back to some key steps in the proof of the quoted Theorem 3.3 in order to highlight the structure with respect to any point and its mirror point according to the considered notations.

By referring to a proof technique developed in [22], the claimed thesis is obtained by proving that the right-hand sides in the inequalities

$$\gamma \geq \frac{1}{d_s(x)} \left[\cos^2\left(\frac{\pi - x}{2}\right) \frac{p_s^2(\pi - x)}{f_s(x)} \right], \tag{3.9a}$$

$$\gamma \geq \frac{1}{d_s(x)} \left[\cos^2\left(\frac{\pi - x}{2}\right) \frac{p_s^2(\pi - x)}{f_s(x)} + \cos^2\left(\frac{x}{2}\right) \frac{p_s^2(x)}{f_s(\pi - x)} \right], \tag{3.9b}$$

$$d_s(x) = \cos^2\left(\frac{x}{2}\right) p_s^2(x) + \cos^2\left(\frac{\pi - x}{2}\right) p_s^2(\pi - x) \tag{3.9c}$$

are uniformly bounded on the whole domain so that γ is a universal constant.

It is evident that (3.9a) is implied by (3.9b). Moreover, both the two terms in (3.9b) and in $d_s(x)$ can be exchanged each other, up to the change of variable $y = \pi - x$.

Therefore, if $x^0 \neq \pi$ it is evident that (3.6a) and (3.6b) ensure the required uniform boundedness since the condition $p_s^2(x) + p_s^2(\pi - x) > 0$ implies $d_s(x) > 0$.

In the case $x^0 = \pi$, the inequality (3.9b) can be rewritten as

$$\gamma \geq \frac{1}{\dfrac{p_s^2(x)}{\cos^2\left(\frac{\pi - x}{2}\right)} + \dfrac{p_s^2(\pi - x)}{\cos^2\left(\frac{x}{2}\right)}} \left[\frac{p_s^2(\pi - x)}{\cos^2\left(\frac{x}{2}\right)\ f_s(x)} + \frac{p_s^2(x)}{\cos^2\left(\frac{\pi - x}{2}\right)\ f_s(\pi - x)} \right] \tag{3.10}$$

so motivating the special case reported in (3.7a).

4. V-cycle optimal convergence

In this section we propose a suitable modification of (3.6a), or (3.7a), with respect to the choice of the polynomial involved into the projector, that allows us to prove the V-cycle optimal convergence according to the verification of the inequalities (2.10a)–(2.10c) and the requirement (2.11).

Thanks to the choice $D = I$, it is worth stressing that the MGM smoothing properties do not require a true verification, since the proofs of (2.10a) and (2.10b) are identical in any matrix algebra with unitary transforms.

Proposition 4.1. *Let $A_s = C_{m_s}(f_s)$ for any $s = 0, \ldots, s_{\min}$, with $f_s \geq 0$, and let ω_s be such that $0 < \omega_s \leq 2/\|f_s\|_\infty$. If we choose α_s and β_s such that $\alpha_s \leq \omega_s \min\left\{2, (2 - \omega_s\|f_s\|_\infty)/(1 - \omega_s\|f_s\|_\infty)^2\right\}$ and $\beta_s \leqslant \omega_s(2 - \omega_s\|f_s\|_\infty)$ then for*

any $x \in \mathbb{C}^m$ *the inequalities*

$$\|V_{s,\mathrm{pre}}\, x\|_{A_s}^2 \leq \|x\|_{A_s}^2 - \alpha_s \|V_{s,\mathrm{pre}}\, x\|_{A_s^2}^2 \tag{4.1}$$

$$\|V_{s,\mathrm{post}}\, x\|_{A_s}^2 \leq \|x\|_{A_s}^2 - \beta_s \|x\|_{A_s^2}^2 \tag{4.2}$$

hold true.

Notice, for instance, that the best bound to β_s is given by $1/\|f_s\|_\infty$ and it is obtained by taking $\omega_s = 1/\|f_s\|_\infty$ [25, 1].

Concerning the analysis of the approximation condition (2.10c) we consider here the case of a generating function f_0 with a single zero at x^0. In such a case, the choice of the polynomial in the projector is more severe with respect to the case of TGM. Let $x^0 \in [0, \pi)$ a zero of the generating function f_s. The polynomial p_s is chosen in such a way that

$$\lim_{x \to x^0} \frac{p_s(\pi - x)}{f_s(x)} < +\infty, \tag{4.3a}$$

$$p_s^2(x) + p_s^2(\pi - x) > 0. \tag{4.3b}$$

In the special case $x^0 = \pi$, the requirement (4.3a) is replaced by

$$\lim_{x \to x^0 = \pi} \frac{p_s(\pi - x)}{\cos\left(\frac{x}{2}\right)\, f_s(x)} < +\infty. \tag{4.4a}$$

Notice also that in the special case $x^0 = 0$ the requirement (4.3a) can be weakened as

$$\lim_{x \to x^0 = 0} \frac{\cos\left(\frac{\pi - x}{2}\right)\, p_s(\pi - x)}{f_s(x)} < +\infty. \tag{4.5a}$$

Proposition 4.2. *Let* $A_s = C_{m_s}(f_s)$ *for any* $s = 0, \ldots, s_{\min}$, *with* $f_s \geq 0$. *Let* $P_{s+1}^s = T_{s+1}^s C_s(p_s)$, *where* $p_s(x)$ *is fulfilling* (4.3a) (*or* (4.4a)) *and* (4.3b). *Then, for any* $s = 0, \ldots, s_{\min} - 1$, *there exists* $\gamma_s > 0$ *independent of* m_s *such that*

$$\|CGC_s x\|_{A_s}^2 \leq \gamma_s \|x\|_{A_s^2}^2 \quad \text{for any } x \in \mathbb{C}^{m_s}, \tag{4.6}$$

where CGC_s *is defined as in* (2.2).

Proof. Since

$$CGC_s = I_s - (P_{s+1}^s)^T (P_{s+1}^s A_s (P_{s+1}^s)^T)^{-1} P_{s+1}^s A_s$$

is a unitary projector, it holds that $CGC_s^T\, A_s\, CGC_s = A_s\, CGC_s$. Therefore, the target inequality (4.6) can be simplified and symmetrized, giving rise to the equivalent matrix inequality

$$\widetilde{CGC_s} = I_s - A_s^{1/2} (P_{s+1}^s)^T (P_{s+1}^s A_s (P_{s+1}^s)^T)^{-1} P_{s+1}^s A_s^{1/2} \leq \gamma_s A_s. \tag{4.7}$$

Hence, by invoking Lemma 3.1, $Q_s^T \widetilde{CGC_s} Q_s$ can be permuted into a 2×2 block diagonal matrix whose jth block, $j = 1, \ldots, m_{s+1}$, is given by the rank-1 matrix

(see [8] for the analogous τ case)

$$I_2 - \frac{1}{c_j^2 + s_j^2} \begin{bmatrix} c_j^2 & c_j s_j \\ c_j s_j & s_j^2 \end{bmatrix},$$

where

$$c_j = \cos\left(\frac{x_j^{[m_s]}}{2}\right) p^2 f(x_j^{[m_s]}) \quad s_j = -\cos\left(\frac{\pi - x_j^{[m_s]}}{2}\right) p^2 f(\pi - x_j^{[m_s]}).$$

As in the proof of the TGM convergence, due to the continuity of f_s and p_s, (4.7) is proven if the right-hand sides in the inequalities

$$\gamma_s \geq \frac{1}{\widetilde{d}_s(x)} \left[\cos^2\left(\frac{\pi - x}{2}\right) \frac{p_s^2 f_s(\pi - x)}{f_s(x)}\right] \tag{4.8a}$$

$$\gamma_s \geq \frac{1}{\widetilde{d}_s(x)} \left[\cos^2\left(\frac{\pi - x}{2}\right) \frac{p_s^2 f_s(\pi - x)}{f_s(x)} + \cos^2\left(\frac{x}{2}\right) \frac{p_s^2 f_s(x)}{f_s(\pi - x)}\right] \tag{4.8b}$$

$$\widetilde{d}_s(x) = \cos^2\left(\frac{x}{2}\right) p_s^2 f_s(x) + \cos^2\left(\frac{\pi - x}{2}\right) p_s^2 f(\pi - x) \tag{4.8c}$$

are uniformly bounded on the whole domain so that γ_s are universal constants.

Once again, it is evident that (4.8a) is implied by (4.8b). Moreover, both the terms in (4.8b) and in $\widetilde{d}_s(x)$ can be exchanged each other, up to the change of variable $y = \pi - x$.

Therefore, if $x^0 \neq \pi$, (4.8b) can be rewritten as

$$\gamma_s \geq \frac{1}{\hat{d}_s(x)} \left[\cos^2\left(\frac{\pi - x}{2}\right) \frac{p_s^2(\pi - x)}{f_s^2(x)} + \cos^2\left(\frac{x}{2}\right) \frac{p_s^2(x)}{f_s^2(\pi - x)}\right] \tag{4.9}$$

where

$$\hat{d}_s(x) = \cos^2\left(\frac{x}{2}\right) \frac{p_s^2(x)}{f_s(\pi - x)} + \cos^2\left(\frac{\pi - x}{2}\right) \frac{p_s^2(\pi - x)}{f_s(x)},$$

so that (4.3a) and (4.3b) ensure the required uniform boundedness.

In the case $x^0 = \pi$, the inequality (4.8b) can be rewritten as

$$\gamma_s \geq \frac{1}{\dfrac{p_s^2(x)}{\cos^2\left(\frac{\pi - x}{2}\right) f_s(\pi - x)} + \dfrac{p_s^2(\pi - x)}{\cos^2\left(\frac{x}{2}\right) f_s(x)}} \left[\frac{p_s^2(\pi - x)}{\cos^2\left(\frac{x}{2}\right) f_s^2(x)} + \frac{p_s^2(x)}{\cos^2\left(\frac{\pi - x}{2}\right) f_s^2(\pi - x)}\right] \tag{4.10}$$

so motivating the special case reported in (4.4a). □

Remark 4.3. Notice that in the case of pre-smoothing iterations and under the assumption $V_{s,\text{pre}}$ nonsingular, the approximation condition

$$\|CGC_s V_{s,\text{pre}}^{\nu_{\text{pre}}} x\|_{A_s}^2 \leq \gamma_s \|V_{s,\text{pre}}^{\nu_{\text{pre}}} x\|_{A_s^2}^2 \text{ for any } x \in \mathbb{C}^{m_s}, \tag{4.11}$$

is equivalent to the condition, in matrix form, $\widetilde{CGC}_s \leq \gamma_s A_s$ obtained in Proposition 4.2.

In Propositions 4.1 and 4.2 we have obtained that for every s (independent of $m = m_0$) the constants α_s, β_s, and γ_s are absolute values not depending on $m = m_0$, but only depending on the functions f_s and p_s. Nevertheless, in order to prove the MGM optimal convergence according to Theorem 2.2, we should verify at least one between the following inf–min conditions [2]:

$$\delta_{\text{pre}} = \inf_{m_0} \min_{0 \leq s \leq \log_2(m_0)} \frac{\alpha_s}{\gamma_s} > 0, \quad \delta_{\text{post}} = \inf_{m_0} \min_{0 \leq s \leq \log_2(m_0)} \frac{\beta_s}{\gamma_s} > 0. \tag{4.12}$$

First, we consider the inf-min requirement (4.12) by analyzing the case of a generating function $\tilde{f}_0$ with a single zero at $x^0 = 0$.

It is worth stressing that in such a case the DCT-III matrix $\tilde{A}_{m_0} = C_{m_0}(\tilde{f}_0)$ is singular since 0 belongs to the set of grid points $x_j^{[m_0]} = (j-1)\pi/m_0$, $j = 1, \ldots, m_0$. Thus, the matrix $\tilde{A}_{m_0}$ is replaced by

$$A_{m_0} = C_{m_0}(f_0) = C_{m_0}(\tilde{f}_0) + \tilde{f}_0\left(x_2^{[m_0]}\right) \cdot \frac{ee^T}{m_0}$$

with $e = [1, \ldots, 1]^T \in \mathbb{R}^{m_0}$ and where the rank-1 additional term is known as Strang correction [30]. Equivalently, $\tilde{f}_0 \geq 0$ is replaced by the generating function

$$f_0 = \tilde{f}_0 + \tilde{f}_0\left(x_2^{[m_0]}\right) \chi_{w_1^{[m_s]} + 2\pi\mathbb{Z}} > 0, \tag{4.13}$$

where χ_X is the characteristic function of the set X and $w_1^{[m_0]} = x^0 = 0$.

In Lemma 4.4 below, it is reported the law to which the generating functions are subjected at the coarser levels. With respect to this target, it is useful to consider the following factorization result: let $f \geq 0$ be a trigonometric polynomial with a single zero at x^0 of order $2q$. Then, there exists a positive trigonometric polynomial ψ such that

$$f(x) = [1 - \cos(x - x_0)]^q \ \psi(x). \tag{4.14}$$

Notice also that, according to Lemma 3.2, the location of the zero is never shifted at the subsequent levels.

Lemma 4.4. *Let $f_0(x) = \tilde{f}_0(x) + c_0\chi_{2\pi\mathbb{Z}}(x)$, with $\tilde{f}_0(x) = [1 - \cos(x)]^q\psi_0(x)$, q being a positive integer and ψ_0 being a positive trigonometric polynomial and with $c_0 = \tilde{f}_0\left(x_2^{[m_0]}\right)$. Let $p_s(x) = [1 + \cos(x)]^q$ for any $s = 0, \ldots, s_{\min} - 1$. Then, under the same assumptions of Lemma* 3.1*, each generating function f_s is given by*

$$f_s(x) = \tilde{f}_s(x) + c_s\chi_{2\pi\mathbb{Z}}(x), \quad \tilde{f}_s(x) = [1 - \cos(x)]^q\psi_s(x).$$

The sequences $\{\psi_s\}$ and $\{c_s\}$ are defined as

$$\psi_{s+1} = \Phi_{q,p_s}(\psi_s), \qquad c_{s+1} = c_s p_s^2(0), \qquad s = 0, \ldots, s_{\min} - 1,$$

where $\Phi_{q,p}$ is an operator such that

$$[\Phi_{q,p}(\psi)](x) = \frac{1}{2^{q+1}}\left[(\varphi\, p\, \psi)\left(\frac{x}{2}\right) + (\varphi\, p\, \psi)\left(\pi - \frac{x}{2}\right)\right], \tag{4.15}$$

with $\varphi(x) = 1 + \cos(x)$. Moreover, each $\tilde{f}_s$ is a trigonometric polynomial that vanishes only at $2\pi\mathbb{Z}$ with the same order $2q$ as $\tilde{f}_0$.

Proof. The claim is a direct consequence of Lemma 3.1. Moreover, since the function ψ_0 is positive by assumption, the same holds true for each function ψ_s. □

Hereafter, we make use of the following notations: for a given function f, we will write $M_f = \sup_x |f|$, $m_f = \inf_x |f|$ and $\mu_\infty(f) = M_f/m_f$.

Now, if $x \in (0, 2\pi)$ we can give an upper bound for the left-hand side $R(x)$ in (4.9), since it holds that

$$\begin{aligned}
R(x) &= \frac{\dfrac{\cos^2\left(\frac{x}{2}\right) p_s^2(x)}{f_s^2(\pi-x)} + \dfrac{\cos^2\left(\frac{\pi-x}{2}\right) p_s^2(\pi-x)}{f_s^2(x)}}{\dfrac{\cos^2\left(\frac{x}{2}\right) p_s^2(x)}{f_s(\pi-x)} + \dfrac{\cos^2\left(\frac{\pi-x}{2}\right) p_s^2(\pi-x)}{f_s(x)}} \\
&= \frac{\dfrac{\cos^2\left(\frac{x}{2}\right)}{\psi_s^2(\pi-x)} + \dfrac{\cos^2\left(\frac{\pi-x}{2}\right)}{\psi_s^2(x)}}{\dfrac{\cos^2\left(\frac{\pi-x}{2}\right) p_s(x)}{\psi_s(\pi-x)} + \dfrac{\cos^2\left(\frac{\pi-x}{2}\right) p_s(\pi-x)}{\psi_s(x)}} \\
&\le \frac{M_{\psi_s}}{m_{\psi_s}^2} \frac{1}{\cos^2\left(\frac{x}{2}\right) p_s(x) + \cos^2\left(\frac{\pi-x}{2}\right) p_s(\pi-x)} \le \frac{M_{\psi_s}}{m_{\psi_s}^2},
\end{aligned}$$

we can consider $\gamma_s = M_{\psi_s}/m_{\psi_s}^2$. In the case $x = 0$, since $p_s(0) = 0$, it holds $R(0) = 1/f_s(\pi)$, so that we have also to require $1/f_s(\pi) \le \gamma_s$. However, since $1/f_s(\pi) \le M_{\psi_s}/m_{\psi_s}^2$, we take $\gamma_s^* = M_{\psi_s}/m_{\psi_s}^2$ as the best value.

In (2.9), by choosing $\omega_s^* = \|f_s\|_\infty^{-1}$, we simply find $\beta_s^* = \|f_s\|_\infty^{-1} \ge 1/(2^q M_{\psi_s})$ and as a consequence, we obtain

$$\frac{\beta_s^*}{\gamma_s^*} \ge \frac{1}{2^q M_{\psi_s}} \cdot \frac{m_{\psi_s}^2}{M_{\psi_s}} = \frac{1}{2^q \mu_\infty^2(\psi_s)}. \tag{4.16}$$

A similar relation can be found in the case of a pre-smoothing iteration. Nevertheless, since it is enough to prove one between the inf-min conditions, we focus our attention on condition (4.16). So, to enforce the inf-min condition (4.12), it is enough to prove the existence of an absolute constant L such that $\mu_\infty(\psi_s) \leqslant L < +\infty$ uniformly in order to deduce that $\|MGM_0\|_{A_0} \leqslant \sqrt{1 - (2^q L^2)^{-1}} < 1$.

Proposition 4.5. *Under the same assumptions of Lemma 4.4, let us define $\psi_s = [\Phi_{p_s,q}]^s(\psi)$ for every $s \in \mathbb{N}$, where $\Phi_{p,q}$ is the linear operator defined as in (4.15). Then, there exists a positive polynomial ψ_∞ of degree q such that ψ_s uniformly converges to ψ_∞, and moreover there exists a positive real number L such that $\mu_\infty(\psi_s) \leqslant L$ for any $s \in \mathbb{N}$.*

Proof. Due to the periodicity and to the cosine expansions of all the involved functions, the operator $\Phi_{q,p}$ in (4.15) can be rewritten as

$$[\Phi_{q,p}(\psi)](x) = \frac{1}{2^{q+1}} \left[(\varphi\, p\, \psi)\left(\frac{x}{2}\right) + (\varphi\, p\, \psi)\left(\pi + \frac{x}{2}\right) \right]. \tag{4.17}$$

The representation of $\Phi_{q,p}$ in the Fourier basis (see Proposition 4.8 in [2]) leads to an operator from $\mathbb{R}^{m(q)}$ to $\mathbb{R}^{m(q)}$, $m(q)$ proper constant depending only on q, which is identical to the irreducible nonnegative matrix $\bar{\Phi}_q$ in equation (4.14) of [2], with $q+1$ in place of q.

As a consequence, the claimed thesis follows by referring to the Perron–Frobenius theorem [15, 31] according to the very same proof technique considered in [2]. □

Lastly, by taking into account all the previous results, we can claim the optimality of the proposed MGM.

Theorem 4.6. *Let $\tilde{f}_0$ be an even nonnegative trigonometric polynomial vanishing at 0 with order $2q$. Let $m_0 = m > m_1 > \cdots > m_s > \cdots > m_{s_{\min}}$, $m_{s+1} = m_s/2$. For any $s = 0, \ldots, s_{\min}-1$, let P^s_{s+1} be as in Proposition 4.2 with $p_s(x) = [1+\cos(x)]^q$, and let $V_{s,\text{post}} = I_{m_s} - A_{m_s}/\|f_s\|_\infty$. If we set $A_{m_0} = C_{m_0}(\tilde{f}_0 + c_0 \chi_{2\pi\mathbb{Z}})$ with $c_0 = \tilde{f}_0(w_2^{[m_0]})$ and we consider $b \in \mathbb{C}^{m_0}$, then the MGM (standard V-cycle) converges to the solution of $A_{m_0}x = b$ and is optimal (in the sense of Definition 1.1).*

Proof. Under the quoted assumptions it holds that $\tilde{f}_0(x) = [1-\cos(x)]^q\, \psi_0(x)$ for some positive polynomial $\psi_0(x)$. Therefore, it is enough to observe that the optimal convergence of MGM as stated in Theorem 2.2 is implied by the inf-min condition (4.12). Thanks to (4.16), the latter is guaranteed if the quantities $\mu_\infty(\psi_s)$ are uniformly bounded and this holds true according to Proposition 4.5. □

Now, we consider the case of a generating function f_0 with a unique zero at $x^0 = \pi$, this being particularly important in applications since the discretization of certain integral equations leads to matrices belonging to this class. For instance, the signal restoration leads to the case of $f_0(\pi) = 0$, while for the super-resolution problem and image restoration $f_0(\pi, \pi) = 0$ is found [5].

By virtue of Lemma 3.2 we simply have that the generating function f_1 related to the first projected matrix uniquely vanishes at 0, i.e., at the first level the MGM projects a discretized integral problem, into another which is spectrally and structurally equivalent to a discretized differential problem.

With respect to the optimal convergence, we have that Theorem 2.2 holds true with $\delta = \min\{\delta_0, \bar{\delta}\}$ since δ results to be a constant and independent of m_0. More precisely, δ_0 is directly related to the finest level and $\bar{\delta}$ is given by the inf-min condition of the differential problem obtained at the coarser levels. The latter constant value has been previously shown, while the former can be proven as follows. We are dealing with $f_0(x) = (1+\cos(x))^q \psi_0(x)$ and according to (4.3a) we

choose $\tilde{p}_0(x) = p_0(x) + d_0\chi_{2\pi\mathbb{Z}}$ with $p_0(x) = (1+\cos(x))^{q+1}$ and $d_0 = p_0(w_2^{[m_0]})$. Therefore, an upper bound for the left-hand side $\tilde{R}(x)$ in (4.10) is obtained as

$$\tilde{R}(x) \leq \frac{M_{\psi_0}}{m_{\psi_0}^2},$$

i.e., we can consider $\gamma_0 = M_{\psi_0}/m_{\psi_0}^2$ and so that a value δ_0 independent of m_0 is found.

5. Numerical experiments

Hereafter, we give numerical evidence of the convergence properties claimed in the previous sections, both in the case of proposed TGM and MGM (standard V-cycle), for two types of DCT-III systems with generating functions having zero at 0 (differential-like problems) and at π (integral-like problems).

The projectors P_{s+1}^s are chosen as described in Section 3 and in Section 4. The Richardson smoothing iterations are used twice in each iteration with $\omega = 2/\|f\|_\infty$ and $\omega = 1/\|f\|_\infty$, respectively. The iteration is stopped when the Euclidean norm of the relative residual at dimension m_0 is less than 10^{-7}. Moreover, the exact solution of the system is found by a direct solver when the coarse grid dimension equals to 16 (16^2 in the additional two-level tests).

5.1. Case $x^0 = 0$ (differential-like problems)

First, we consider the case $A_m = C_m(f_0)$ with $f_0(x) = [2 - 2\cos(x)]^q$, i.e., with a unique zero at $x^0 = 0$ of order $2q$.

As previously outlined, the matrix $C_m(f_0)$ is singular, so that the solution of the rank-1 corrected system is considered, whose matrix is given by $C_m(f_0) + (f_0(\pi/m)/m)ee^T$, with $e = [1, \ldots, 1]^T$. Since the position of the zero $x^0 = 0$ at the coarser levels is never shifted, then the function $p_s(x) = [2 - 2\cos(\pi - x)]^r$ in the projectors is the same at all the subsequent levels s.

To test TGM/MGM linear convergence with rate independent of the size m_0 we tried for different r: according to (3.6a), we must choose r at least equal to 1 if $q = 1$ and at least equal to 2 if $q = 2, 3$, while according to (4.3a) we must always choose r equal to q. In Table 1, we report the number of iterations required for convergence. As expected, it results to be bounded by a constant irrespective of the problem size. Notice also that, in general, the MGM requires the same number of iterations than the TGM. In other words, the approximation due to the Coarse Grid Correction in the V-cycle does not introduce any significant loss of accuracy.

By using tensor arguments, the previous results plainly extend to the multilevel case. In Table 2 we consider the case of generating function $f_0(x, y) = f_0(x) + f_0(y)$, that arises in the uniform finite difference discretization of elliptic constant coefficient differential equations on a square with Neumann boundary conditions, see, e.g., [24].

TABLE 1. Number of required iterations – $1D$ Case: $f_0(x) = [2 - 2\cos(x)]^q$ and $p(x) = [2 - 2\cos(\pi - x)]^r$.

TGM						MGM					
	$q=1$	$q=2$		$q=3$			$q=1$	$q=2$		$q=3$	
m_0	$r=1$	$r=1$	$r=2$	$r=2$	$r=3$	m_0	$r=1$	$r=1$	$r=2$	$r=2$	$r=3$
16	7	15	13	28	24	16	1	1	1	1	1
32	7	16	15	34	32	32	7	16	15	34	32
64	7	16	16	35	34	64	7	17	16	35	34
128	7	16	16	35	35	128	7	18	16	35	35
256	7	16	16	35	35	256	7	18	16	35	35
512	7	16	16	35	35	512	7	18	16	35	35

TABLE 2. Number of required iterations – $2D$ Case: $f_0(x,y) = [2 - 2\cos(x)]^q + [2 - 2\cos(y)]^q$ and $p(x,y) = [2 - 2\cos(\pi - x)]^r + [2 - 2\cos(\pi - y)]^r$.

TGM						MGM					
	$q=1$	$q=2$		$q=3$			$q=1$	$q=2$		$q=3$	
m_0	$r=1$	$r=1$	$r=2$	$r=2$	$r=3$	m_0	$r=1$	$r=1$	$r=2$	$r=2$	$r=3$
16^2	15	34	30	–	–	16^2	1	1	1	1	1
32^2	16	36	35	71	67	32^2	16	36	35	71	67
64^2	16	36	36	74	73	64^2	16	36	36	74	73
128^2	16	36	36	74	73	128^2	16	36	36	74	73
256^2	16	36	36	74	73	256^2	16	37	36	74	73
512^2	16	36	36	74	73	512^2	16	37	36	74	73

5.2. Case $\mathbf{x^0 = \pi}$ (integral-like problems)

DCT-III matrices $A_{m_0} = C_{m_0}(f_0)$ whose generating function shows a unique zero at $x^0 = \pi$ are encountered in solving integral equations, for instance in image restoration problems with Neumann (reflecting) boundary conditions [18].

According to Lemma 3.2, if $x^0 = \pi$, then the generating function f_1 of the coarser matrix $A_{m_1} = C_{m_1}(f_1)$, $m_1 = m_0/2$, has a unique zero at 0, whose order equals the order of $x^0 = \pi$ with respect to f_0 plus two.

It is worth stressing that in such a case the projector at the first level is singular so that its rank-1 Strang correction is considered. This choice gives rise in a natural way to the rank-1 correction considered in Section 5.1. Moreover, starting from the second coarser level, the new location of the zero is never shifted from 0. In Table 3 the number of iterations required for convergence, both in the one-level and two-level case, is reported.

6. Computational costs and conclusions

Some remarks about the computational costs are required in order to highlight the optimality of the proposed procedure.

Since the matrix $C_{m_s}(p)$ appearing in the definition of P^s_{s+1} is banded, the cost of a matrix vector product involving P^s_{s+1} is $O(m_s)$. Therefore, the first con-

TABLE 3. Number of required iterations – $1D$ Case: $f_0(x) = 2 + 2\cos(x)$ and $p_0(x) = 2 - 2\cos(\pi - x)$ and $2D$ Case: $f_0(x, y) = 4 + 2\cos(x) + 2\cos(y)$ and $p_0(x, y) = 4 - 2\cos(\pi - x) - 2\cos(\pi - y)$.

m_0	TGM	MGM
16	15	1
32	14	14
64	12	13
128	11	13
256	10	12
512	8	10

m_0	TGM	MGM
16^2	7	1
32^2	7	7
64^2	7	7
128^2	7	6
256^2	7	6
512^2	7	6

dition in Definition 1.1 is satisfied. In addition, notice that the matrices at every level (except for the coarsest) are never formed since we need only to store the $O(1)$ nonzero Fourier coefficients of the related generating function at each level for matrix-vector multiplications. Thus, the memory requirements are also very low.

With respect to the second condition in Definition 1.1 we stress that the representation of $A_{m_{s+1}} = C_{m_{s+1}}(f_{s+1})$ can be obtained formally in $O(1)$ operations by virtue of (3.5). In addition, the zeros of f_{s+1} and their orders are obtained according to Lemma 3.2 by knowing the zeros of f_s and their orders. Furthermore, each iteration of TGM costs $O(m_0)$ operations as A_{m_0} is banded. In conclusion, each iteration of the proposed TGM requires $O(m_0)$ operations.

With regard to MGM, optimality is reached since we have proven that there exists δ independent from both m and $s_{\min}$ so that the number of required iterations results uniformly bounded by a constant irrespective of the problem size. In addition, since each iteration has a computational cost proportional to matrix-vector product, Definition 1.1 states that such a kind of MGM is *optimal*.

Finally, we report some remarks about the performances of the proposed method with respect to those achieved by considering Fast Cosine Transform (FCT) standard resolutions [27]. Let us consider the linear system $A_m x_m = b_m$, where A_m belongs to the DCT-III matrix algebra. Then, the solution is given by

$$x_m = Q_m \Lambda_m^{-1} Q_m^T b_m, \tag{6.1}$$

where Λ_m is the diagonal matrix holding the eigenvalues of A_m. Since, these eigenvalues can be evaluates as

$$[\Lambda_m]_{(i,i)} = \frac{[Q_m^T A_m e_1]_i}{[Q_m^T e_1]_i}, \quad i = 1, \ldots, m,$$

with $e_1 = [1, 0, \ldots, 0]^T$, the solution (6.1) can be obtained in three FCT, i.e., within $O(m \log m)$ operations. Thus, such a computational cost is not substantially higher than the cost of the proposed MGM. In addition, the MGM implementation is a delicate task since it clearly involves an efficient memory allocation, both with respect to the recursion itself and the data stored in the setup phase.

Nevertheless, even in the one-level case, for increasing dimensions, the solution computed by means of FCT can show a worsening in the accuracy greater

TABLE 4. Relative error in Euclidean norm e_r in the case of our MGM Matlab code and FCT built-in Matlab function based code – $1D$ case: $f_0(x) = [2 - 2\cos(x)]^2$ ($p(x) = [2 - 2\cos(\pi - x)]^2$ and stopping criterion equal to 10^{-7} for the MGM).

m_0	MGM e_r	FCT e_r
16	2.23e-011	8.43e-014
32	3.99e-010	1.87e-013
64	3.20e-010	1.16e-011
128	2.43e-008	2.68e-010
256	2.68e-007	3.54e-009
512	4.46e-006	1.04e-007
1024	5.47e-005	1.61e-006
2048	5.67e-004	1.72e-002
4096	9.22e-003	3.93e-001

than the solution computed by MGM, as, for instance, in the case of the generating function $f_0(x) = 2 - 2\cos(x)$. The loss of accuracy appears even more suddenly in the case of $f_0(x) = [2 - 2\cos(x)]^2$: for a size larger than 2000, the results delivered by our MGM implementation are definitely more accurate than those obtained by using the Matlab FCT code and show a more predictable pattern. More precisely, for increasing dimensions, Table 4 compares the relative error in Euclidean norm obtained by considering our MGM Matlab implementation and by using the built-in Matlab dct function. For the sake of completeness, we have to mention that our MGM CPU times are larger than those of the FCT code. Nevertheless, we think that the comparison is unfair since the former refer to a non-optimized Matlab code (the matrix-vector product key operation is implemented in $O(m)$ operations by using loops, but it is known that loops slow down the Matlab performances), while the latter make use of a built-in executable function.

Moreover, the MGM, due to its iterative nature, has also a greater flexibility related to the choice of threshold in the stopping criterion, whenever the considered application does not require a high accuracy. We recall that this situation is not academic and indeed it does occur when considering image restoration problems [18, 17], where we cannot expect an error less than the noise level.

As a conclusion, we observe that the reported numerical tests in Section 5 show that the requirements on the order of zero in the projector could be weakened. Future works will deal with this topic and with the extension of the convergence analysis in the case of a general location of the zeros of the generating function.

Acknowledgment

Warm thanks to the referee for very pertinent and useful remarks.

References

[1] A. Aricò, M. Donatelli, *A V-cycle multigrid for multilevel matrix algebras: proof of optimality.* Numer. Math. **105** (2007), no. 4, 511–547 (DOI 10.1007/s00211-006-0049-7).

[2] A. Aricò, M. Donatelli, S. Serra-Capizzano, *V-cycle optimal convergence for certain (multilevel) structured linear systems.* SIAM J. Matrix Anal. Appl. **26** (2004), no. 1, 186–214.

[3] O. Axelsson, M. Neytcheva, *The algebraic multilevel iteration methods – theory and applications.* In *Proceedings of the Second International Colloquium on Numerical Analysis* (Plovdiv, 1993), 13–23, VSP, 1994.

[4] R.H. Chan, T.F. Chan, C. Wong, *Cosine transform based preconditioners for total variation minimization problems in image processing.* In *Iterative Methods in Linear Algebra, II, V3, IMACS Series in Computational and Applied Mathematics, Proceedings of the Second IMACS International Symposium on Iterative Methods in Linear Algebra*, Bulgaria, 1995, 311–329.

[5] R.H. Chan, M. Donatelli, S. Serra-Capizzano, C. Tablino-Possio, *Application of multigrid techniques to image restoration problems.* In Proceedings of *SPIE-Session: Advanced Signal Processing: Algorithms, Architectures, and Implementations XII*, Vol. 4791 (2002), F. Luk Ed., 210–221.

[6] R.H. Chan, M.K. Ng, *Conjugate gradient methods for Toeplitz systems.* SIAM Rev. **38** (1996), no. 3, 427–482.

[7] R.H. Chan, S. Serra-Capizzano, C. Tablino-Possio, *Two-grid methods for banded linear systems from DCT III algebra.* Numer. Linear Algebra Appl. **12** (2005), no. 2-3, 241–249.

[8] G. Fiorentino, S. Serra-Capizzano, *Multigrid methods for Toeplitz matrices.* Calcolo **28** (1991), no. 3-4, 283–305.

[9] G. Fiorentino, S. Serra-Capizzano, *Multigrid methods for symmetric positive definite block Toeplitz matrices with nonnegative generating functions.* SIAM J. Sci. Comput. **17** (1996), no. 5, 1068–1081.

[10] R. Fischer, T. Huckle, *Multigrid methods for anisotropic BTTB systems.* Linear Algebra Appl. **417** (2006), no. 2-3, 314–334.

[11] W. Hackbusch, *Multigrid methods and applications.* Springer Series in Computational Mathematics, 4. Springer-Verlag, 1985.

[12] T. Huckle, J. Staudacher, *Multigrid preconditioning and Toeplitz matrices.* Electron. Trans. Numer. Anal. **13** (2002), 81–105.

[13] T. Huckle, J. Staudacher, *Multigrid methods for block Toeplitz matrices with small size blocks.* BIT **46** (2006), no. 1, 61–83.

[14] X.Q. Jin, *Developments and applications of block Toeplitz iterative solvers.* Combinatorics and Computer Science, 2. Kluwer Academic Publishers Group, Dordrecht; Science Press, Beijing, 2002.

[15] D.G. Luenberger, *Introduction to Dynamic Systems: Theory, Models, and Applications*, John Wiley & Sons Inc., 1979.

[16] M.K. Ng, *Iterative methods for Toeplitz systems.* Numerical Mathematics and Scientific Computation. Oxford University Press, 2004.

[17] M.K. Ng, R.H. Chan, T.F. Chan, A.M. Yip, *Cosine transform preconditioners for high resolution image reconstruction. Conference Celebrating the 60th Birthday of Robert J. Plemmons* (*Winston-Salem, NC,* 1999). Linear Algebra Appl. **316** (2000), no. 1-3, 89–104.

[18] M.K. Ng, R.H. Chan, W.C. Tang, *A fast algorithm for deblurring models with Neumann boundary conditions.* SIAM J. Sci. Comput. **21** (1999), no. 3, 851–866.

[19] M.K. Ng, S. Serra-Capizzano, C. Tablino-Possio. *Numerical behaviour of multigrid methods for symmetric sinc-Galerkin systems.* Numer. Linear Algebra Appl. **12** (2005), no. 2-3, 261–269.

[20] D. Noutsos, S. Serra-Capizzano, P. Vassalos, *Matrix algebra preconditioners for multilevel Toeplitz systems do not insure optimal convergence rate.* Theoret. Comput. Sci. **315** (2004), no. 2-3, 557–579.

[21] J. Ruge, K. Stüben, *Algebraic multigrid.* In *Frontiers in Applied Mathematics: Multigrid Methods. SIAM*, 1987, 73–130.

[22] S. Serra Capizzano, *Convergence analysis of two-grid methods for elliptic Toeplitz and PDEs matrix-sequences.* Numer. Math. **92** (2002), no. 3, 433–465.

[23] S. Serra-Capizzano, *Matrix algebra preconditioners for multilevel Toeplitz matrices are not superlinear. Special issue on structured and infinite systems of linear equations.* Linear Algebra Appl. **343-344** (2002), 303–319.

[24] S. Serra Capizzano, C. Tablino Possio, *Spectral and structural analysis of high precision finite difference matrices for elliptic operators.* Linear Algebra Appl. **293** (1999), no. 1-3, 85–131.

[25] S. Serra-Capizzano, C. Tablino-Possio, *Multigrid methods for multilevel circulant matrices.* SIAM J. Sci. Comput. **26** (2004), no. 1, 55–85.

[26] S. Serra-Capizzano and E. Tyrtyshnikov, *How to prove that a preconditioner cannot be superlinear.* Math. Comp. **72** (2003), no. 243, 1305–1316.

[27] G. Strang, *The discrete cosine transform*, SIAM Review, **41**, n. 1, pp. 135–147, 1999.

[28] H. Sun, X. Jin, Q. Chang, *Convergence of the multigrid method for ill-conditioned block Toeplitz systems.* BIT **41** (2001), no. 1, 179–190.

[29] U. Trottenberg, C.W. Oosterlee and A. Schüller, *Multigrid. With contributions by A. Brandt, P. Oswald and K. Stüben.* Academic Press, Inc., 2001.

[30] E. Tyrtyshnikov, *Circulant preconditioners with unbounded inverses.* Linear Algebra Appl. **216** (1995), 1–23.

[31] R.S. Varga, *Matrix Iterative Analysis.* Prentice-Hall, Inc., Englewood Cliffs, 1962.

C. Tablino Possio
Dipartimento di Matematica e Applicazioni,
Università di Milano Bicocca,
via Cozzi 53
I-20125 Milano, Italy
e-mail: cristina.tablinopossio@unimib.it

Operator Theory:
Advances and Applications, Vol. 199, 397–419

The Ratio Between the Toeplitz and the Unstructured Condition Number

Siegfried M. Rump and Hiroshi Sekigawa

Abstract. Recently it was shown that the ratio between the normwise Toeplitz structured condition number of a linear system and the general unstructured condition number has a finite lower bound. However, the bound was not explicit, and nothing was known about the quality of the bound. In this note we derive an explicit lower bound only depending on the dimension n, and we show that this bound is almost sharp for all n.

Mathematics Subject Classification (2000). 15A12, 26D05.

Keywords. Structured condition number, Toeplitz matrix, Mahler measure, polynomial norms.

1. Notation and problem formulation

For a system of linear equations $Ax = b$ with $A \in \mathbb{R}^{n\times n}$, $x, b \in \mathbb{R}^n$, the condition number characterizes the sensitivity of the solution x with respect to infinitely small perturbations of the matrix A. For $\varepsilon > 0$, denote

$$M_\varepsilon := M_\varepsilon(A) := \{\Delta A \in \mathbb{R}^{n\times n} : \|\Delta A\| \le \varepsilon\|A\|\}, \tag{1.1}$$

where throughout the paper $\|\cdot\|$ denotes the spectral norm for matrices and for vectors. Denote by $P_\varepsilon := P_\varepsilon(A, x)$ the set of all vectors $\Delta x \in \mathbb{R}^n$ for which there exists $\Delta A \in M_\varepsilon$ with $(A+\Delta A)(x+\Delta x) = Ax$. Then the (unstructured) normwise condition number is defined by

$$\kappa(A, x) := \lim_{\varepsilon\to 0} \sup_{\Delta x\in P_\varepsilon} \frac{\|\Delta x\|}{\varepsilon\|x\|}. \tag{1.2}$$

It is well known that $\kappa(A, x) = \|A^{-1}\|\,\|A\|$, such that the (unstructured) condition number does not depend on x.

If the matrix A has some structure, it seems reasonable to restrict the set M_ε to matrices with similar structure. For $a = (a_{-(n-1)}, \ldots, a_{-1}, a_0, a_1, \ldots, a_{n-1})$, the

$n \times n$ Toeplitz matrix $T_n(a)$ is of the form

$$T := T_n(a) := \begin{pmatrix} a_0 & a_1 & \dots & a_{n-1} \\ a_{-1} & a_0 & \ddots & \vdots \\ \vdots & \ddots & \ddots & a_1 \\ a_{-(n-1)} & \dots & a_{-1} & a_0 \end{pmatrix} \in \mathbb{R}^{n \times n}. \tag{1.3}$$

For given (nonsingular) Toeplitz matrix T, restricting M_ε to Toeplitz matrices changes (1.2) into the Toeplitz condition number $\kappa^{\mathrm{Toep}}(T, x)$ [10], [14], [2, Section 13.3]. Since the set of perturbations ΔA is restricted, it follows $\kappa^{\mathrm{Toep}}(T, x) \leq \kappa(T, x) = \|T^{-1}\| \, \|T\|$. Note that in contrast to the general condition number, the Toeplitz condition number depends on x. However, there exists always a worst case x such that both condition numbers coincide [14, Theorem 4.1]:

$$\sup_{x \neq 0} \left\{ \kappa^{\mathrm{Toep}}(T, x) \right\} = \|T^{-1}\| \, \|T\| \, .$$

In [14, Theorem 10.2] it was shown that $\kappa^{\mathrm{Toep}}(T, x) \geq 2^{-1/2} \sqrt{\kappa(T, x)}$ (see also [2, Theorem 13.14]), hence the ratio $\kappa^{\mathrm{Toep}}/\kappa$ is bounded below by $[2\|T^{-1}\| \, \|T\|]^{-1/2}$. The question arises, how small can the Toeplitz condition number actually be compared to the general condition number?

In a recent survey paper on Toeplitz and Hankel matrices [4], Böttcher and Rost note "One expects that $\kappa^{\mathrm{Toep}}(T, x)$ is in general significantly smaller than $\kappa(T, x)$, but, curiously up to now no convincing example in this direction is known." Furthermore, Böttcher and Rost continue to note that, as proved in [3] (submitted in 2002 but appeared in 2005), it seems rather hopeless to find examples numerically (see also [2, Theorem 13.20]):

Theorem 1.1 (Böttcher, Grudsky, 2002). *Let $x_0, x_1, \dots, x_{n-1} \in \mathbb{C}$ be independent random variables whose real and imaginary parts are subject to the standard normal distribution and put $x = (x_j)_{j=0}^{n-1}$. There are universal constants $\delta \in (0, \infty)$ and $n_0 \in \mathbb{N}$ such that*

$$\text{Probability} \left(\frac{\kappa^{\mathrm{Toep}}(T_n(a), x)}{\kappa(T_n(a), x)} \geq \frac{\delta}{n^{3/2}} \right) > \frac{99}{100}$$

for all finitely supported sequences a and all $n \geq n_0$.

Notice that generically $\kappa(T_n(a), x)$ remains bounded or increases exponentially fast as n goes to infinity. Since in the case of exponential growth the factor $\delta/n^{3/2}$ is harmless, it follows that with high probability that $\kappa^{\mathrm{Toep}}(T_n(a), x)$ increases exponentially fast together with $\kappa(T_n(a), x)$.

In [14] the first author showed a lower bound on the ratio $\kappa^{\mathrm{Toep}}/\kappa$ which surprisingly depends only on the solution x, not on A (see also [2, Theorem 13.16]). However, despite some examples of small dimension (inspired by Heinig [9]) no general examples could be derived.

In this note we

1. derive a general lower bound on $\kappa^{\text{Toep}}(T,x)/\kappa(T,x)$ only depending on the dimension n, and
2. show that this lower bound is almost sharp for all n.

The solution of both problems is based on the minimization of the smallest singular value of a class of Toeplitz matrices (2.2) and its surprising connection to a lower bound on the coefficients of the product of two polynomials. We will prove in Corollary 2.11 that

$$\frac{2n}{\Delta^{n-1}} \geq \inf\{\frac{\kappa^{\text{Toep}}(A,x)}{\kappa(A,x)} : \ A \in \mathbb{R}^{n\times n} \text{ Toeplitz}, \ 0 \neq x \in \mathbb{R}^n\} > \frac{\sqrt{2}}{n\Delta^{n-1}},$$

where $\Delta = 3.209912\ldots$.

We denote by $\sigma_{\min}(A)$ the smallest singular value of the matrix A, and by J the permutation matrix ("flip matrix") mapping $(1,\ldots,n)$ into $(n,\ldots,1)$.

2. Main results

Let a linear system $Ax = b$ with Toeplitz matrix A be given. The defining equation $(A+\Delta A)(x+\Delta x) = Ax$ with $\|\Delta A\| \leq \varepsilon\|A\|$ implies

$$\Delta x = -A^{-1}\Delta A x + \mathcal{O}(\varepsilon). \tag{2.1}$$

For Toeplitz perturbations, we have $\Delta A = T(\Delta a)$ with $\Delta a \in \mathbb{R}^{2n-1}$ according to (1.3), and using ideas from [10] a computation shows [14]

$$\Delta A x = J\Psi_x \Delta a \quad \text{with} \quad \Psi_x := \begin{pmatrix} x_1 & x_2 & \ldots & x_n & & \\ & x_1 & x_2 & \ldots & x_n & \\ & & & \ldots & & \\ & & x_1 & x_2 & \ldots & x_n \end{pmatrix} \in \mathbb{R}^{n\times(2n-1)}. \tag{2.2}$$

In [14, Lemma 6.3] it was shown that the spectral matrix norm of ΔA and Euclidean norm of Δa are related by

$$\frac{1}{\sqrt{n}}\|\Delta A\| \leq \|\Delta a\| \leq \sqrt{2}\|\Delta A\|. \tag{2.3}$$

Combining this with the definition of $\kappa^{\text{Toep}}(A,x)$ and (2.1) yields [14, Theorem 6.5]

$$\kappa^{\text{Toep}}(A,x) = \gamma\frac{\|A^{-1}J\Psi_x\|\,\|A\|}{\|x\|} \quad \text{with} \quad \frac{1}{\sqrt{n}} \leq \gamma \leq \sqrt{2}, \tag{2.4}$$

so that $\|A^{-1}J\Psi_x\| \geq \|A^{-1}\|\sigma_{\min}(\Psi_x)$ implies [14, Corollary 6.6]

$$\frac{\kappa^{\text{Toep}}(A,x)}{\kappa(A,x)} \geq \frac{1}{\sqrt{n}}\frac{\|A^{-1}J\Psi_x\|}{\|A^{-1}\|\,\|x\|} \geq \frac{1}{\sqrt{n}}\frac{\sigma_{\min}(\Psi_x)}{\|x\|}. \tag{2.5}$$

Surprisingly, this lower bound depends only on the solution x. That means, a given solution x implies a lower bound for $\kappa^{\text{Toep}}(A,x)/\kappa(A,x)$ for any Toeplitz matrix A.

We will show that the lower bound in (2.5) is achievable up to a small factor. For this we first construct for given x a Toeplitz matrix A with ratio $\kappa^{\rm Toep}/\kappa$ near $\sigma_{\min}(\Psi_x)/\|x\|$.

Let fixed but arbitrary $x \in \mathbb{R}^n$ be given. For simplicity assume $\|x\| = 1$. First we will show that for $\delta > 0$ there exists a Toeplitz matrix A with $\|A^{-1}J\Psi_x\| < \|A^{-1}\|\sigma_{\min}(\Psi_x) + \delta$.

Denote by $y \in \mathbb{R}^n$, $\|y\| = 1$, a left singular vector of Ψ_x to $\sigma_{\min}(\Psi_x)$, so that $\|y^T\Psi_x\| = \sigma_{\min}(\Psi_x)$. By $\Psi_x\Psi_x^T = J\Psi_x\Psi_x^TJ^T$ we may assume either $y = Jy$ or $y = -Jy$. Define by

$$L(p_1, \ldots, p_n) := \begin{pmatrix} p_1 & & & \\ p_2 & \ddots & & \\ \vdots & \ddots & \ddots & \\ p_n & \ldots & p_2 & p_1 \end{pmatrix} \in \mathbb{R}^{n\times n},$$

a lower triangular Toeplitz matrix depending on $p \in \mathbb{R}^n$. Define

$$B := L(y_1, y_2, \ldots, y_n) \quad \text{and} \quad C := L(0, y_n, y_{n-1}, \ldots, y_2)$$

and

$$R_\varepsilon := (B + \varepsilon I)(B + \varepsilon I)^T - CC^T. \tag{2.6}$$

If R_ε is invertible, then the Gohberg-Semencul formula ([8], see also [4, Th. 3.3]) implies that $A_\varepsilon := R_\varepsilon^{-1}$ is a (symmetric) Toeplitz matrix. Furthermore, a direct computation using $y = \pm Jy$ yields

$$R_0 = yy^T \tag{2.7}$$

which implies $\det(R_0) = 0$ for $n \geq 2$. The determinant of R_ε is a monic polynomial of degree $2n$ in ε, thus R_ε is nonsingular for all $0 \neq \varepsilon < \varepsilon_0$ for small enough ε_0. Hence there is a constant α, independent of ε, with

$$\|R_\varepsilon\Psi_x\| \leq \|yy^T\Psi_x\| + \alpha\varepsilon = \sigma_{\min}(\Psi_x) + \alpha\varepsilon,$$

the latter equality because $\sigma_{\min}^2(\Psi_x)$ is the only nonzero eigenvalue of

$$yy^T\Psi_x(yy^T\Psi_x)^T.$$

Since $A_\varepsilon = R_\varepsilon^{-1}$ is a Toeplitz matrix and $y = \pm Jy$, (2.4) implies the following result, which is trivially also true for $n = 1$.

Theorem 2.1. *Let $0 \neq x \in \mathbb{R}^n$ be given. Then for all $\delta > 0$ there exists a Toeplitz matrix $A \in \mathbb{R}^{n\times n}$ with*

$$\|A^{-1}\|\sigma_{\min}(\Psi_x) \leq \|A^{-1}J\Psi_x\| < \|A^{-1}\|\sigma_{\min}(\Psi_x) + \delta$$

and

$$\kappa^{\rm Toep}(A, x) = \gamma \cdot \kappa(A, x)\frac{\sigma_{\min}(\Psi_x)}{\|x\|} + \delta' \quad \textit{for} \quad \frac{1}{\sqrt{n}} \leq \gamma \leq \sqrt{2} \quad \textit{and} \quad 0 \leq \delta' < \sqrt{2}\delta.$$

For $x \neq 0$, the matrix Ψ_x has full rank because otherwise each $n \times n$ submatrix of Ψ_x would be singular, taking the leftmost submatrix in Ψ_x would imply $x_1 = 0$, the second leftmost would imply $x_2 = 0$ and so forth. Thus

$$\mu_n := \min_{0 \neq x \in \mathbb{R}} \frac{\sigma_{\min}(\Psi_x)}{\|x\|} = \min_{\|x\|=1} \sigma_{\min}(\Psi_x) > 0 \tag{2.8}$$

for all n, and Theorem 2.1 yields

Corollary 2.2. *For all n,*

$$\sqrt{2}\mu_n \geq \inf\{\frac{\kappa^{\text{Toep}}(A,x)}{\kappa(A,x)} : \ A \in \mathbb{R}^{n\times n} \text{ Toeplitz}, \ 0 \neq x \in \mathbb{R}^n\} \geq \frac{1}{\sqrt{n}}\mu_n.$$

In the remaining of the paper we will estimate μ_n to characterize the infimum of $\kappa^{\text{Toep}}/\kappa$. The matrix Ψ_x is also known as "polynomial matrix" [1]. Identifying a vector $x = (x_1, \ldots, x_n) \in \mathbb{R}^n$ with the polynomial $x(t) := \sum_{\nu=0}^{n-1} x_{n-\nu} t^\nu \in \mathbb{R}[t]$, a little computation yields

$$z = y^T \Psi_x \quad \Leftrightarrow \quad z(t) = y(t)x(t), \tag{2.9}$$

and therefore

$$y^T \Psi_x = x^T \Psi_y. \tag{2.10}$$

This, of course, can also be verified by direct computation. We define the norm $\|x(t)\|$ of a polynomial by the norm $\|x\|$ of its coefficient vector. Since $\|y^T \Psi_x\| = \sigma_{\min}(\Psi_x)$, we can characterize μ_n by

$$\mu_n = \min\{\|PQ\| : \ P, Q \in \mathbb{R}[t], \ \deg(P) = \deg(Q) = n-1, \ \|P\| = \|Q\| = 1\}. \tag{2.11}$$

To give lower and upper bounds for μ_n, we first describe some related results for polynomials. The supremum norm $\|P\|_E$ of a complex univariate polynomial P on a compact set $E \subset \mathbb{C}$ is defined as

$$\|P\|_E := \sup_{z \in E} |P(z)|. \tag{2.12}$$

In [12] Kneser gave the exact lower bound for the supremum norm on the interval $[-1, 1]$ of the product of two polynomials.

Theorem 2.3 (Kneser, 1934). *Suppose that $PQ = R$, where P, Q and R are complex polynomials of degree m, $n - m$ and n, respectively. Then for all m and n*

$$\|P\|_{[-1,1]} \|Q\|_{[-1,1]} \leq K_{m,n} \|R\|_{[-1,1]},$$

where

$$K_{m,n} := 2^{n-1} \prod_{k=1}^{m} \left(1 + \cos\frac{(2k-1)\pi}{2n}\right) \prod_{k=1}^{n-m} \left(1 + \cos\frac{(2k-1)\pi}{2n}\right).$$

This bound is exactly attained by the Chebyshev polynomial of degree n.

[1] Many thanks to Ludwig Elsner, Bielefeld, for pointing to this connection.

To estimate μ_n, we need similar results for the unit disc D. Boyd's result in [5, 6] gives a sharp inequality for this case. To describe Boyd's results, we define the Mahler measure. For a complex polynomial F in k variables the Mahler measure of F is defined as

$$M(F) := \exp\left(\int_0^1 \cdots \int_0^1 \log|F(e^{2\pi\sqrt{-1}t_1}, \ldots, e^{2\pi\sqrt{-1}t_k})|dt_1 \cdots dt_k\right).$$

Theorem 2.4 (Boyd, 1992/94). *Let R be a polynomial of degree n with complex coefficients and suppose that $PQ = R$. Then for the norm $\|\cdot\|_D$ as in* (2.12) *on the unit disc D*

$$\|P\|_D\|Q\|_D \le \delta^n\|R\|_D,$$

where $\delta = M(1 + x + y - xy) = 1.7916228\ldots$. The constant is best possible.

As written in section 3 of [5], the constant δ can be expressed in terms of Clausen's integral

$$Cl_2(\theta) = -\int_0^\theta \log\left(2\sin\frac{t}{2}\right)dt = \sum_{k=1}^{\infty}\frac{\sin k\theta}{k^2},$$

or, in terms of $I(\theta)$, where

$$I(\theta) = \int_0^\theta \log\left(2\cos\frac{t}{2}\right)dt = Cl_2(\pi - \theta).$$

Using Catalan's constant $G = I(\pi/2) = Cl_2(\pi/2) \approx 0.9160$, we can write $\delta = e^{2G/\pi}$.

Theorem 2.4 implies a lower bound for μ_n. To obtain an upper bound for μ_n, we estimate the supremum norms of the following polynomials. Define $F_n(t)$ as $t^{2n} + (-1)^n$. Let $\hat{P}_n(t)$ be the monic polynomial of degree n with the zeros of $F_n(t)$ in the right half-plane, and $\hat{Q}_n(t)$ be the monic polynomial of degree n with the zeros of $F_n(t)$ in the left half-plane. It follows $\hat{P}_n\hat{Q}_n = F_n$ and $\hat{Q}_n(t) = (-1)^n\hat{P}_n(-t)$.

Lemma 2.5. *For the norm $\|\cdot\|_D$ as in* (2.12) *on the unit disc D, the following inequalities hold true.*

$$e^{\frac{\pi}{8n}}\delta^n > \|\hat{P}_n\|_D = (-1)^n\hat{P}_n(-1) = \|\hat{Q}_n\|_D = \hat{Q}_n(1) > \delta^n.$$

Remark 2.6. When n is even, $2K_{n/2,n} = \hat{Q}_n(1)^2$, where $K_{p,q}$ is the constant in Theorem 2.3.

Combining Theorem 2.4 and Lemma 2.5, where the proof of the latter is deferred to the appendix, with (2.11), we obtain an upper and a lower bound for μ_n. Before we state our final result, we prove that we may assume without loss of generality that polynomials P and Q minimizing μ_n as in (2.11) must both have all their roots on the unit circle. This is also useful to identify such polynomials P and Q numerically for small n. In fact, the following Theorem 2.7 shows more, namely that for fixed (normed) Q there is a (normed) polynomial P with only roots on the unit circle and minimizing $\|PQ\|$.

Theorem 2.7. *For two nonzero real univariate polynomials P and Q with $\|P\| = \|Q\| = 1$, there exists a real univariate polynomial P' such that $\deg(P') = \deg(P)$, $\|P'\| = 1$, all zeros of P' lie on the unit circle and $\|P'Q\| \le \|PQ\|$.*

The proof of Theorem 2.7 is rather involved, and thus deferred to the appendix. An immediate consequence is the following corollary.

Corollary 2.8.

$$\mu_n = \min\{\|PQ\| : P, Q \in \mathbb{R}[t],\ \deg(P) = \deg(Q) = n-1,\ \|P\| = \|Q\| = 1, \text{ and } P,\ Q \text{ have all zeros on the unit circle}\}.$$

Now we can prove the following upper and lower bounds for μ_n.

Theorem 2.9.

$$\frac{\sqrt{2}(n+1)}{\Delta^n} > \mu_{n+1} \ge \frac{2}{\sqrt{2n+1}\Delta^n},$$

where $\Delta := e^{4G/\pi}$ for Catalan's constant G. It is $\Delta = \delta^2$, where δ is the constant in Theorem 2.4. Note that $\Delta = 3.209912\ldots$.

Remark 2.10. Using Proposition 2.12 at the end of this section, we can improve the upper bound to

$$\frac{C\sqrt{n+1}}{\Delta^n},$$

where C is a constant independent of n.

Proof. Let F be a complex polynomial $\sum_{\nu=0}^{n} a_\nu t^\nu$. Then, the following inequalities among norms of F hold.

$$\sqrt{n+1}\|F\| \ge |F|_1 \ge \|F\|_D \ge \|F\|. \tag{2.13}$$

Here, $|F|_1$ is defined as $\sum_{\nu=0}^{n} |a_\nu|$. Real polynomials P and Q minimizing μ_n have all their roots on the unit circle. For this case the right-most inequality in (2.13) improves into

$$\|F\|_D \ge \sqrt{2}\|F\| \tag{2.14}$$

which follows from a much more general result[2] in [16], see also [17, (7.71)]. From Theorem 2.4, for *real* polynomials P and Q of degree n, we have

$$\frac{\|PQ\|_D}{\|P\|_D\|Q\|_D} \ge \frac{1}{\delta^{2n}} = \frac{1}{\Delta^n}.$$

Therefore, for polynomials P and Q with $\|P\| = \|Q\| = 1$, the inequalities

$$\|PQ\| \ge \frac{2\|PQ\|_D}{\sqrt{2n+1}\|P\|_D\|Q\|_D} \ge \frac{2}{\sqrt{2n+1}\Delta^n}$$

follow from (2.13) and (2.14). This proves the lower bound for μ_{n+1}.

[2]Thanks to P. Batra, Hamburg, for pointing to this reference.

Let $\hat{P}_n$ and $\hat{Q}_n$ be as in Lemma 2.5. An upper bound for $\|\hat{P}_n\hat{Q}_n\|/(\|\hat{P}_n\|\|\hat{Q}_n\|)$ is an upper bound for μ_{n+1}. Since $\|\hat{P}_n\hat{Q}_n\| = \sqrt{2}$, the inequalities

$$\frac{\|\hat{P}_n\hat{Q}_n\|}{\|\hat{P}_n\|\|\hat{Q}_n\|} \leq \frac{\sqrt{2}}{(\|\hat{P}_n\|_D/\sqrt{n+1})(\|\hat{Q}_n\|_D/\sqrt{n+1})} < \frac{\sqrt{2}(n+1)}{\Delta^n}$$

follow from (2.13) and Lemma 2.5. □

Inserting this into Corollary 2.2 characterizes the asymptotic behavior of the worst ratio between the unstructured and structured condition number for Toeplitz matrices.

Corollary 2.11. *For all n,*

$$\frac{2n}{\Delta^{n-1}} > \inf\left\{\frac{\kappa^{\mathrm{Toep}}(A,x)}{\kappa(A,x)} : \ A \in \mathbb{R}^{n\times n} \text{ Toeplitz}, \ 0 \neq x \in \mathbb{R}^n\right\} > \frac{\sqrt{2}}{n\Delta^{n-1}}, \quad (2.15)$$

where $\Delta = 3.209912\ldots$ is the constant in Theorem 2.9.

We can improve the upper bound using the following proposition, the proof of which is given in the Appendix.

Proposition 2.12.

$$\lim_{n\to\infty} \frac{||\hat{P}_n||n^{1/4}}{||\hat{P}_n||_D} = \lim_{n\to\infty} \frac{||\hat{Q}_n||n^{1/4}}{||\hat{Q}_n||_D} = \frac{1}{\sqrt{2}}.$$

By similar arguments of the proof for Theorem 2.9, we obtain the following improved upper bound.

Corollary 2.13. *There exists a constant $C > 0$ such that for all n,*

$$\frac{C\sqrt{n}}{\Delta^{n-1}} > \inf\left\{\frac{\kappa^{\mathrm{Toep}}(A,x)}{\kappa(A,x)} : \ A \in \mathbb{R}^{n\times n} \text{ Toeplitz}, \ 0 \neq x \in \mathbb{R}^n\right\},$$

where $\Delta = 3.209912\ldots$ is the constant in Theorem 2.9.

3. Approximation of μ_n

Next we show how to approximate $\Psi_x \in \mathbb{R}^{n\times(2n-1)}$ minimizing $\sigma_{\min}(\Psi_x)$. Using Ψ_x, a Toeplitz matrix with small ratio $\kappa^{\mathrm{Toep}}/\kappa$ can be constructed following the discussion preceding Theorem 2.1. For given unit vector $x \in \mathbb{R}^n$ and $x(t) := \sum_{\nu=0}^{n-1} x_{n-\nu}t^\nu$ define Ψ_x as in (2.2), and let $y \in \mathbb{R}^n$ be a unit left singular vector to $\sigma_{\min}(\Psi_x)$ of Ψ_x. With $y(t) := \sum_{\nu=0}^{n-1} y_{n-\nu}t^\nu$ as in the discussion following Corollary 2.2 we have

$$\|x\| = \|y\| = \|x(t)\| = \|y(t)\| = 1 \quad \text{and} \quad \|x(t)y(t)\| = \|y^T\Psi_x\| = \sigma_{\min}(\Psi_x).$$

n	approximate μ_n	rigorous bounds of μ_n	$\hat{\mu}_n = \frac{\|\|\hat{P}_n\hat{Q}_n\|\|}{\|\|\hat{P}_n\|\|\|\|\hat{Q}_n\|\|}$	$\hat{\mu}_n/\mu_n$
2	0.70710678118655	[0.70710678118 , 0.70710678119]	0.707107	1.0000
3	0.33333333333333	[0.33333333333 , 0.33333333334]	0.353553	1.0607
4	0.13201959446019	[0.13201959446 , 0.13201959447]	0.141421	1.0712
5	0.04836936580270	[0.04836936580 , 0.04836936581]	0.051777	1.0705
6	0.01702151213258	[0.01702151213 , 0.01702151214]	0.018183	1.0682
7	0.00584679996238	[0.00584679996 , 0.00584679997]	0.006234	1.0662
8	0.00197621751074	[0.00197621751 , 0.00197621752]	0.002104	1.0647

TABLE 1

For fixed $x(t)$, the polynomial $y(t)$ minimizes $\|x(t)y(t)\|$ subject to $\|y(t)\| = 1$. Now (2.10) implies $\|x^T\Psi_y\| = \sigma_{\min}(\Psi_x)$ and therefore $\sigma_{\min}(\Psi_y) \leq \sigma_{\min}(\Psi_x)$. Iterating the process, that is replacing x by y and computing y as a left singular vector to $\sigma_{\min}(\Psi_x)$, generates a monotonically decreasing and therefore convergent sequence. Practical experience suggests that for generic starting vector x this sequence converges mostly to the same limit, presumably μ_n. In any case this limit is an upper bound for μ_n. Table 1 displays this limit for some values of n.

To ensure that the limit is not a local but the global minimum $\min_x \sigma_{\min}(\Psi_x)$, a verified global optimization method was used [13] for computing rigorous lower and upper bounds for μ_n. Such methods take all procedural, approximation and rounding errors into account and are, provided the computer system works to its specifications, rigorous (see, for example, [7]). For given n and using (2.11) this means $2n$ variables. This was possible up to $n = 5$ with reasonable effort. The rightmost column in Table 1 displays the computed bounds for μ_n. For larger values of n, the number of variables was significantly reduced using Theorem 2.7. Since minimizers P, Q have only roots on the unit circle it follows $P(z) = \pm z^n P(1/z)$ and similarly for Q, i.e., the coefficient vectors are (skew-)symmetric to reflection. Using this allows the computation of rigorous bounds for μ_n up to $n = 8$ with moderate effort.[3]

The best-known lower and upper bounds for μ_n are by Kaltofen et al. [11]. They compute verified lower bounds until $n = 18$. For bounding μ_{18} from below they need 25 days of computing time. Results and more background are summarized in [15]. Computational evidence supports the following conjecture.

Conjecture 3.1. *There are polynomials $P, Q \in \mathbb{R}[t]$ with $\deg P = \deg Q = n - 1$ and $\|P\| = \|Q\| = 1$ with $\mu_n = \|PQ\|$ such that all coefficients of P are positive, $Q(t) = P(-t)$ and all roots of P and Q lie on the unit circle. The roots $a_\nu \pm ib_\nu$ of P have all positive real parts a_ν, and the roots of Q are $-a_\nu \pm ib_\nu$.*

[3]Thanks to Kyoko Makino for performing the verified global optimization using the COSY-package [1].

Finally, the values $\hat{\mu}_n = \frac{||\hat{P}_n \hat{Q}_n||}{||\hat{P}_n|| ||\hat{Q}_n||}$ for the polynomials $\hat{P}_n, \hat{Q}_n$ as in Lemma 2.5 and the ratio $\hat{\mu}_n/\mu_n$ is displayed as well. It seems that $\hat{P}_n, \hat{Q}_n$ are not far from the optimum. This is supported by Proposition 2.12.

4. Appendix

Proof of Lemma 2.5. Since we can write

$$\hat{Q}_n(t) = \begin{cases} (t+1) \displaystyle\prod_{k=1}^{\frac{n-1}{2}} \left(t^2 + 2t \cos \frac{k\pi}{n} + 1 \right), & \text{if } n \text{ is odd,} \\ \displaystyle\prod_{k=1}^{\frac{n}{2}} \left(t^2 + 2t \cos \frac{(2k-1)\pi}{2n} + 1 \right), & \text{if } n \text{ is even,} \end{cases} \tag{4.1}$$

and $\cos \frac{k\pi}{n}, \cos \frac{(2k-1)\pi}{2n} > 0$ for the values of k in question, we have $\|\hat{Q}_n\|_D = \hat{Q}_n(1)$. From the definition of $\hat{Q}_n$, we have $\|\hat{Q}_n\|_D = \|\hat{P}_n\|_D$ and $\hat{Q}_n(1) = (-1)^n \hat{P}_n(-1)$.

First we prove the inequalities in Lemma 2.5 when n is odd. From (4.1), we have

$$\hat{Q}_n(1) = 2^{\frac{n+1}{2}} \prod_{k=1}^{\frac{n-1}{2}} \left(1 + \cos \frac{k\pi}{n} \right).$$

Therefore,

$$\log \hat{Q}_n(1) = \frac{(n+1)\log 2}{2} + \sum_{k=1}^{\frac{n-1}{2}} \log \left(1 + \cos \frac{k\pi}{n} \right).$$

Let a and b be real numbers with $a < b$. For a real function f such that $f'' \leq 0$ on the interval $[a, b]$, we have

$$(b-a) f\left(\frac{a+b}{2} \right) \geq \int_a^b f(x)dx \geq (b-a) \frac{f(a) + f(b)}{2}. \tag{4.2}$$

Applying (4.2) to $f = \log(1 + \cos \pi x)$ on intervals $[0, \frac{1}{2n}]$, $[\frac{2k-1}{2n}, \frac{2k+1}{2n}]$ ($k = 1, 2, \ldots, \frac{n-1}{2}$) for an upper estimation, and on intervals $[\frac{k}{n}, \frac{k+1}{n}]$ ($k = 0, 1, \ldots, \frac{n-3}{2}$), $[\frac{n-1}{2n}, \frac{1}{2}]$ for a lower estimation, we have

$$\begin{aligned} &\frac{\log \hat{Q}_n(1)}{n} - \frac{(n+1)\log 2}{2n} + \frac{1}{2n} \log \left(1 + \cos \frac{\pi}{4n} \right) \\ &\quad \geq \int_0^{\frac{1}{2}} \log(1 + \cos \pi x) dx \\ &\quad \geq \frac{\log \hat{Q}_n(1)}{n} - \frac{\log 2}{2} - \frac{1}{4n} \log \left(1 + \cos \frac{(n-1)\pi}{2n} \right). \end{aligned}$$

Since $1+\cos\pi x = 2\cos^2\frac{\pi x}{2}$, it follows

$$\begin{aligned}\int_0^{\frac{1}{2}} \log(1+\cos\pi x)dx &= \int_0^{\frac{1}{2}} \log\left(2\cos^2\frac{\pi x}{2}\right)dx \\ &= \frac{2}{\pi}\int_0^{\frac{\pi}{2}} \log\left(2\cos\frac{t}{2}\right)dt - \frac{\log 2}{2} = \log\delta - \frac{\log 2}{2}.\end{aligned} \tag{4.3}$$

From (4.3), we have

$$\begin{aligned}&\frac{\log\hat{Q}_n(1)}{n} - \frac{\log 2}{2n} + \frac{1}{2n}\log\left(1+\cos\frac{\pi}{4n}\right) \\ &\quad \geq \log\delta \geq \frac{\log\hat{Q}_n(1)}{n} - \frac{1}{4n}\log\left(1+\cos\frac{(n-1)\pi}{2n}\right).\end{aligned}$$

Therefore, the following inequalities hold.

$$\begin{aligned}n\log\delta + \frac{1}{4}\log\left(1+\cos\frac{(n-1)\pi}{2n}\right) &\geq \log\hat{Q}_n(1) \\ &\geq n\log\delta + \frac{\log 2}{2} - \frac{1}{2}\log\left(1+\cos\frac{\pi}{4n}\right).\end{aligned} \tag{4.4}$$

Since

$$\log\left(1+\cos\frac{(n-1)\pi}{2n}\right) = \log\left(1+\sin\frac{\pi}{2n}\right) < \log\left(1+\frac{\pi}{2n}\right) < \frac{\pi}{2n},$$

we can estimate the left-hand side of (4.4) by

$$n\log\delta + \frac{1}{4}\log\left(1+\cos\frac{(n-1)\pi}{2n}\right) < n\log\delta + \frac{\pi}{8n}.$$

An estimation for the right-hand side of (4.4) is as follows.

Since $\log 2 > \log\left(1+\cos\frac{\pi}{4n}\right)$, we have

$$n\log\delta + \frac{\log 2}{2} - \frac{1}{2}\log\left(1+\cos\frac{\pi}{4n}\right) > n\log\delta,$$

and therefore

$$n\log\delta + \frac{\pi}{8n} > \log\hat{Q}_n(1) > n\log\delta$$

proves Lemma 2.5 for odd n. When n is even, we have

$$\log\hat{Q}_n(1) = \frac{n\log 2}{2} + \sum_{k=1}^{\frac{n}{2}} \log\left(1+\cos\frac{(2k-1)\pi}{2n}\right).$$

Applying (4.2) to $f = \log(1+\cos\pi x)$ on intervals $[\frac{k}{n}, \frac{k+1}{n}]$ $(k = 0, 1, \ldots, \frac{n}{2}-1)$, $[\frac{n-1}{2n}, \frac{1}{2}]$ for an upper estimation, and on intervals $[0, \frac{1}{2n}]$, $[\frac{2k-1}{2n}, \frac{2k+1}{2n}]$ $(k = 1, 2,$

$\ldots, \frac{n}{2}-1), [\frac{n-1}{2n}, \frac{1}{2}]$ for a lower estimation, we have

$$\begin{aligned}\frac{\log \hat{Q}_n(1)}{n} - \frac{\log 2}{2} &\geq \int_0^{\frac{1}{2}} \log\left(1+\cos \pi x\right) dx \\ &\geq \frac{\log \hat{Q}_n(1)}{n} - \frac{\log 2}{2} - \frac{1}{2n}\log\left(1+\cos\frac{\pi}{2n}\right) \\ &\quad + \frac{\log 2}{2n} - \frac{1}{4n}\log\left(1+\cos\frac{(n-1)\pi}{2n}\right).\end{aligned}$$

Therefore, the inequalities

$$\begin{aligned}n\log\delta + \frac{1}{2}\log\left(1+\cos\frac{\pi}{2n}\right) - \frac{\log 2}{2} + \frac{1}{4}\log\left(1+\cos\frac{(n-1)\pi}{2n}\right) \\ \geq \log \hat{Q}_n(1) \geq n\log\delta\end{aligned}$$

hold, and from similar arguments for odd n, the inequalities

$$n\log\delta + \frac{\pi}{8n} > \log \hat{Q}_n(1) > n\log\delta$$

prove Lemma 2.5 for even n. □

To prove Theorem 2.7, we need the following lemmas, corollaries and algorithm.

Lemma 4.1. *Let F and G be nonzero complex univariate polynomials, and ζ be a fixed complex number on the unit circle. Define $\nu : \mathbb{R} \to \mathbb{R}$ by*

$$\nu(r) := \frac{\|(t-r\zeta)F\|}{\|(t-r\zeta)G\|}.$$

Then, the following statements hold.

1. *$\nu(r)$ has a minimum at either $r = 1$ or -1.*
2. *If $\nu(1)$ is not a minimum, then $\nu(r) > \nu(0)$ for any $r > 0$.*

Proof. Since $\nu(r)$ is nonnegative, it is sufficient to prove that $N(r) = \nu(r)^2$ has the above properties. For $P(t) = \sum_{k=0}^{n} a_k t^k$, we have

$$\|(t-r\zeta)P\|^2 = \|P\|^2 r^2 - \left(\zeta\sum_{k=1}^{n}\overline{a}_{k-1}a_k + \overline{\zeta}\sum_{k=1}^{n}a_{k-1}\overline{a}_k\right)r + \|P\|^2. \tag{4.5}$$

Therefore, we can write

$$\begin{aligned}\|(t-r\zeta)F\|^2 &= f_1 r^2 + f_2 r + f_1, \\ \|(t-r\zeta)G\|^2 &= g_1 r^2 + g_2 r + g_1,\end{aligned}$$

where $f_1 = \|F\|^2$, $g_1 = \|G\|^2$, and f_2, g_2 are real numbers. Therefore, we have

$$N'(r) = \frac{(f_1 g_2 - f_2 g_1)(r^2-1)}{\|(t-r\zeta)G\|^2}.$$

If $f_1 g_2 - f_2 g_1 = 0$, then $N(r)$ is constant and the statements are clear.

If $f_1g_2 - f_2g_1 > 0$, then $N(r)$ tends to $f_1/g_1 = N(0)$, as r tends to $\pm\infty$. Furthermore, $N(r)$ is monotonically increasing on $(-\infty, -1]$, monotonically decreasing on $[-1, 1]$ and monotonically increasing on $[1, \infty)$. Therefore, $N(r)$ has a minimum at $r = 1$.

If $f_1g_2 - f_2g_1 < 0$, then similar arguments hold. We have $N(r) > N(0)$ for any $r > 0$ and $N(-1)$ is a minimum. □

The following corollary immediately follows from Lemma 4.1.

Corollary 4.2. *Let F and G be nonzero complex univariate polynomials in t, and α be a nonzero complex number. Put $\zeta = \alpha/|\alpha|$. (That is, $|\zeta| = 1$.) Then, the following inequality holds.*

$$\frac{\|(t-\alpha)F\|}{\|(t-\alpha)G\|} \geq \min\left\{\frac{\|(t-\zeta)F\|}{\|(t-\zeta)G\|}, \frac{\|tF\|}{\|tG\|}\right\}.$$

When polynomials F and G are real, the following lemma holds.

Lemma 4.3. *Let F and G be nonzero real univariate polynomials in t, and r be a fixed nonzero real number. Define $\nu : \mathbb{C} \to \mathbb{R}$ by*

$$\nu(\zeta) := \frac{\|(t-r\zeta)F\|}{\|(t-r\zeta)G\|}.$$

We consider $\nu(\zeta)$ a function on the unit circle in $\mathbb{C}$. Then, $\nu(\zeta)$ has a minimum at $\zeta = -1$ or 1.

Proof. Since $\nu(\zeta)$ is nonnegative, it is sufficient to prove that $N(\zeta) = \nu(\zeta)^2$ has a minimum at $\zeta = -1$ or 1. From Equation (4.5) and considering F, $G \in \mathbb{R}[t]$, we have

$$\begin{aligned}\|(t-r\zeta)F\|^2 &= f_1(\zeta + \overline{\zeta}) + f_2,\\ \|(t-r\zeta)G\|^2 &= g_1(\zeta + \overline{\zeta}) + g_2,\end{aligned}$$

where $f_2 = (r^2+1)\|F\|^2$, $g_2 = (r^2+1)\|G\|^2$ and f_1, $g_1 \in \mathbb{R}$. Put $s = \zeta + \overline{\zeta}$ $(\in \mathbb{R})$. We can write $N(\zeta)$ as $\tilde{N}(s)$, which is a function of s $(-2 \leq s \leq 2)$. Then we have

$$\tilde{N}'(s) = \frac{f_1g_2 - f_2g_1}{(g_1s + g_2)^2},$$

That is, $\tilde{N}'(s)$ is monotonic on $[-2, 2]$. Therefore, it has a minimum at $s = -2$ or 2, which corresponds to $\zeta = -1$ or 1, respectively. □

Combining Lemmas 4.1 and 4.3, we can easily see that the following corollary holds.

Corollary 4.4. *Let F and G be nonzero real univariate polynomials, and α be a complex number. Then, the following inequality holds.*

$$\frac{\|(t-\alpha)F\|}{\|(t-\alpha)G\|} \geq \min\left\{\frac{\|(t-1)F\|}{\|(t-1)G\|}, \frac{\|(t+1)F\|}{\|(t+1)G\|}\right\}.$$

Finally, we describe the following algorithm.

Algorithm 4.5. Given a real polynomial $P(t) = (t-\alpha)(t-\overline{\alpha})P_0(t)$, where $\alpha \in \mathbb{C}$, $\notin \mathbb{R}$, this algorithm constructs $F \in \mathbb{R}[t]$ satisfying the following conditions.

1. The degree of F is two.
2. Both zeros of F lie on the unit circle.
3. F satisfies the following inequality.

$$\frac{\|PQ\|}{\|P\|} \geq \frac{\|FP_0Q\|}{\|FP_0\|}.$$

Step 1. Put $\zeta = \alpha/|\alpha|$. If

$$\frac{\|(t-\alpha)(t-\overline{\alpha})P_0Q\|}{\|(t-\alpha)(t-\overline{\alpha})P_0\|} \geq \frac{\|(t-\zeta)(t-\overline{\alpha})P_0Q\|}{\|(t-\zeta)(t-\overline{\alpha})P_0\|},$$

then go to Step 2. Otherwise, go to Step 3.

Step 2.

Step 2.1. If

$$\frac{\|(t-\zeta)(t-\overline{\alpha})P_0Q\|}{\|(t-\zeta)(t-\overline{\alpha})P_0\|} \geq \frac{\|(t-\zeta)(t-\overline{\zeta})P_0Q\|}{\|(t-\zeta)(t-\overline{\zeta})P_0\|},$$

then terminate with the output $(t-\zeta)(t-\overline{\zeta})$. Otherwise, go to Step 2.2.

Step 2.2. If

$$\frac{\|t(t-\zeta)P_0Q\|}{\|t(t-\zeta)P_0\|} \geq \frac{\|t(t-1)P_0Q\|}{\|t(t-1)P_0\|},$$

then put $b_1 = 1$. Otherwise put $b_1 = -1$. If

$$\frac{\|t(t-b_1)P_0Q\|}{\|t(t-b_1)P_0\|} \geq \frac{\|(t-1)(t-b_1)P_0Q\|}{\|(t-1)(t-b_1)P_0\|},$$

then put $b_2 = 1$. Otherwise put $b_2 = -1$.
Terminate with the output $(t-b_1)(t-b_2)$.

Step 3. If

$$\frac{\|t(t-\overline{\alpha})P_0Q\|}{\|t(t-\overline{\alpha})P_0\|} \geq \frac{\|t(t-1)P_0Q\|}{\|t(t-1)P_0\|},$$

then put $b_3 = 1$. Otherwise, put $b_3 = -1$. If

$$\frac{\|t(t-b_3)P_0Q\|}{\|t(t-b_3)P_0\|} \geq \frac{\|(t-1)(t-b_3)P_0Q\|}{\|(t-1)(t-b_3)P_0\|},$$

then put $b_4 = 1$. Otherwise put $b_4 = -1$.
Terminate with the output $(t-b_3)(t-b_4)$.

The validity of the algorithm is as follows. In Step 2.1, if the inequality does not hold, then we have

$$\frac{\|(t-\zeta)(t-\overline{\alpha})P_0Q\|}{\|(t-\zeta)(t-\overline{\alpha})P_0\|} \geq \frac{\|t(t-\zeta)P_0Q\|}{\|t(t-\zeta)P_0\|}$$

from Corollary 4.2.

In Step 2.2, the following inequalities hold from Corollary 4.4.

$$\frac{\|t(t-\zeta)P_0Q\|}{\|t(t-\zeta)P_0\|} \geq \frac{\|t(t-b_1)P_0Q\|}{\|t(t-b_1)P_0\|} \geq \frac{\|(t-b_1)(t-b_2)P_0Q\|}{\|(t-b_1)(t-b_2)P_0\|}.$$

In Step 3, the inequality

$$\frac{\|(t-\alpha)(t-\overline{\alpha})P_0Q\|}{\|(t-\alpha)(t-\overline{\alpha})P_0\|} \geq \frac{\|t(t-\overline{\alpha})P_0Q\|}{\|t(t-\overline{\alpha})P_0\|}$$

holds from Corollary 4.2. Furthermore, the inequalities

$$\frac{\|t(t-\overline{\alpha})P_0Q\|}{\|t(t-\overline{\alpha})P_0\|} \geq \frac{\|t(t-b_3)P_0Q\|}{\|t(t-b_3)P_0\|} \geq \frac{\|(t-b_3)(t-b_4)P_0Q\|}{\|(t-b_3)(t-b_4)P_0\|}$$

hold from Corollary 4.4.

Proof of Theorem 2.7. It is sufficient to show that the following two statements hold.

1. Given $P = (t-a)P_0$, where $a \in \mathbb{R}$, we can construct a real polynomial $R = (t-b)P_0$ ($b = 1$ or -1) satisfying the following inequality.
$$\frac{\|PQ\|}{\|P\|} \geq \frac{\|RQ\|}{\|R\|}.$$
2. Given $P = (t-\alpha)(t-\overline{\alpha})P_0$, where $\alpha \in \mathbb{C}$, $\notin \mathbb{R}$, we can construct a real polynomial $R = FP_0$ with the inequality
$$\frac{\|PQ\|}{\|P\|} \geq \frac{\|RQ\|}{\|R\|},$$
where F is a univariate real polynomial of degree two with both zeros on the unit circle.

The first statement and the second statement follow from Corollary 4.4 and Algorithm 4.5, respectively. □

To prove Proposition 2.12, we need some lemmas.

Lemma 4.6. *Let $P(t)$ be a real univariate polynomial of degree n. For an integer $m > n$, the equality*

$$\|P\|^2 = \frac{1}{m}\sum_{k=1}^{m} |P(\omega\zeta^k)|^2$$

holds, where $\omega \in \mathbb{C}$ lies on the unit circle and ζ is a primitive mth root of unity.

Lemma 4.7. *For arbitrary $\epsilon > 0$, there exists $\theta > 0$ such that the inequality*

$$1 - 2(1-\epsilon)x \geq \frac{1-\sin x}{1+\sin x}$$

holds for $\theta \geq x \geq 0$.

Proof. Since

$$\frac{1-\sin x}{1+\sin x} = 1 - \frac{2\sin x}{1+\sin x},$$

the inequality is equivalent to

$$\frac{\sin x}{1+\sin x} \geq (1-\epsilon)x. \tag{4.6}$$

As x tends to 0,

$$\frac{\sin x}{x} \to 1, \qquad \frac{1}{1+\sin x} \to 1,$$

hold. Therefore, for given $\epsilon > 0$, there exists $\theta > 0$ such that the inequality (4.6) holds for $\theta \geq x \geq 0$. □

Lemma 4.8. *For arbitrary $\epsilon > 0$, there exists $\theta > 0$ such that the inequalities*

$$\exp(-x) \geq 1 - x \geq \exp(-(1+\epsilon)x)$$

hold for $\theta \geq x \geq 0$.

Lemma 4.9 (Jordan's Inequality). *For $\pi/2 \geq x \geq 0$,*

$$x \geq \sin x \geq \frac{2x}{\pi}.$$

Proof of Proposition 2.12. First we prove the proposition when n is odd. It is sufficient to show that for any $\epsilon > 0$, there exists an integer N such that the inequalities

$$\begin{aligned} &\frac{1}{2\sqrt{1-\epsilon}} + \frac{1}{2\sqrt{n}} - \frac{\sqrt{n}}{2}\exp\left(-\frac{\lfloor n^{2/3}\rfloor^2}{n}\right) \\ &\quad > \frac{\|\hat{Q}_n\|^2\sqrt{n}}{\|\hat{Q}_n\|_D^2} > \frac{1}{2\sqrt{1+\epsilon}} - \frac{1}{\sqrt{n}} - \frac{\sqrt{n}\exp\left(-(1+\epsilon)\pi n^{1/3}\right)}{2(1+\epsilon)\pi} \end{aligned} \tag{4.7}$$

hold for any odd integer $n \geq N$.

Let ζ be $\exp(\pi\sqrt{-1}/n)$. Then we have

$$\|\hat{Q}_n\|^2 = \frac{1}{2n}\sum_{k=1}^{2n} |\hat{Q}(\zeta^k)|^2 = \frac{\hat{Q}_n(1)^2}{2n} + \frac{1}{n}\sum_{k=1}^{(n-1)/2} |\hat{Q}_n(\zeta^k)|^2.$$

The relation between $|\hat{Q}_n(\zeta^k)|^2$ and $|\hat{Q}_n(\zeta^{k-1})|^2$ is as follows.

$$|\hat{Q}_n(\zeta^k)|^2 = |\hat{Q}_n(\zeta^{k-1})|^2 \frac{\left|\zeta^{k-1} + \zeta^{-\frac{n+1}{2}}\right|^2}{\left|\zeta^{k-1} + \zeta^{\frac{n-1}{2}}\right|^2} = |\hat{Q}_n(\zeta^{k-1})|^2 \frac{\left|1 + \zeta^{-\frac{n-1}{2}-k}\right|^2}{\left|1 + \zeta^{\frac{n+1}{2}-k}\right|^2}.$$

Since the equalities

$$|1+\zeta^j|^2 = (1+\zeta^j)(1+\zeta^{-j}) = 2\left(1 + \cos\frac{j\pi}{n}\right)$$

hold for $j \in \mathbb{N}$, we have

$$|\hat{Q}_n(\zeta^k)|^2 = |\hat{Q}_n(\zeta^{k-1})|^2 \frac{1+\cos\left(\frac{\pi}{2}+\frac{(2k-1)\pi}{2n}\right)}{1+\cos\left(\frac{\pi}{2}-\frac{(2k-1)\pi}{2n}\right)} = |\hat{Q}_n(\zeta^{k-1})|^2 \frac{1-\sin\frac{(2k-1)\pi}{2n}}{1+\sin\frac{(2k-1)\pi}{2n}}. \tag{4.8}$$

First we show the upper bound. Take any $\epsilon > 0$. Then, there exists an integer L such that the above lemma holds for $\theta = L^{-1/3}\pi$. Take any $n \geq L$. Since we have

$$\frac{\pi}{L^{1/3}} \geq \frac{\pi}{n^{1/3}} \geq \frac{(2n^{2/3}-1)\pi}{2n} \geq \frac{(2k-1)\pi}{2n}$$

for $\lfloor n^{2/3} \rfloor \geq k \geq 1$, the following inequalities follow from Lemmas 4.7 and 4.8.

$$\frac{1-\sin\frac{(2k-1)\pi}{2n}}{1+\sin\frac{(2k-1)\pi}{2n}} \leq 1 - \frac{(1-\epsilon)(2k-1)\pi}{n} \leq \exp\left(-\frac{(1-\epsilon)(2k-1)\pi}{n}\right).$$

Therefore, for $\lfloor n^{2/3} \rfloor \geq k \geq 1$ we have

$$\begin{aligned}
|\hat{Q}_n(\zeta^k)|^2 &\leq |\hat{Q}_n(\zeta^{k-1})|^2 \exp\left(-\frac{(1-\epsilon)(2k-1)\pi}{n}\right) \\
&\leq \hat{Q}_n(1)^2 \prod_{j=1}^{k} \exp\left(-\frac{(1-\epsilon)(2j-1)\pi}{n}\right) \\
&= \hat{Q}_n(1)^2 \exp\left(-\frac{(1-\epsilon)\pi}{n}\sum_{j=1}^{k}(2j-1)\right) = \hat{Q}_n(1)^2 \exp\left(-\frac{(1-\epsilon)\pi}{n}k^2\right).
\end{aligned}$$

Since the inequality

$$\frac{1-\sin\frac{(2k-1)\pi}{2n}}{1+\sin\frac{(2k-1)\pi}{2n}} \leq 1 - \sin\frac{(2k-1)\pi}{2n}$$

holds for $(n-1)/2 \geq k > \lfloor n^{2/3} \rfloor$, the following inequalities follow from Lemmas 4.8 and 4.9.

$$\frac{1-\sin\frac{(2k-1)\pi}{2n}}{1+\sin\frac{(2k-1)\pi}{2n}} \leq \exp\left(-\sin\frac{(2k-1)\pi}{2n}\right) \leq \exp\left(-\frac{2k-1}{n}\right).$$

Hence, for $(n-1)/2 \geq k > \lfloor n^{2/3} \rfloor$, we have

$$\begin{aligned}
|\hat{Q}_n(\zeta^k)|^2 &\leq |\hat{Q}_n(\zeta^{k-1})|^2 \exp\left(-\frac{2k-1}{n}\right) \leq \hat{Q}_n(1)^2 \prod_{j=1}^{k} \exp\left(-\frac{2j-1}{n}\right) \\
&= \hat{Q}_n(1)^2 \exp\left(-\frac{1}{n}\sum_{j=1}^{k}(2j-1)\right) = \hat{Q}_n(1)^2 \exp\left(-\frac{k^2}{n}\right).
\end{aligned}$$

Therefore, the following inequality holds.

$$\frac{\hat{Q}_n(1)^2}{2n} + \frac{\hat{Q}_n(1)^2}{n}\sum_{k=1}^{\lfloor n^{2/3}\rfloor}\exp\left(-\frac{(1-\epsilon)\pi}{n}k^2\right) + \frac{\hat{Q}_n(1)^2}{n}\sum_{k=\lfloor n^{2/3}\rfloor+1}^{(n-1)/2}\exp\left(-\frac{k^2}{n}\right) \geq \|\hat{Q}_n\|^2.$$

Here,

$$\sum_{k=1}^{\lfloor n^{2/3}\rfloor}\exp\left(-\frac{(1-\epsilon)\pi}{n}k^2\right) < \int_0^\infty \exp\left(-\frac{(1-\epsilon)\pi}{n}x^2\right)dx = \frac{1}{2}\sqrt{\frac{n}{1-\epsilon}}$$

holds since

$$\int_0^\infty \exp(-cx^2)dx = \frac{1}{\sqrt{c}}\int_0^\infty \exp(-x^2)dx = \frac{1}{2}\sqrt{\frac{\pi}{c}}$$

holds for $c > 0$. Then we have

$$\begin{aligned}\sum_{k=\lfloor n^{2/3}\rfloor+1}^{(n-1)/2}\exp\left(-\frac{k^2}{n}\right) &< \int_{\lfloor n^{2/3}\rfloor}^\infty \exp\left(-\frac{x^2}{n}\right)dx < \int_{\lfloor n^{2/3}\rfloor}^\infty x\exp\left(-\frac{x^2}{n}\right)dx \\ &= \left[-\frac{n}{2}\exp\left(-\frac{x^2}{n}\right)\right]_{\lfloor n^{2/3}\rfloor}^\infty = -\frac{n}{2}\exp\left(-\frac{\lfloor n^{2/3}\rfloor^2}{n}\right).\end{aligned}$$

Hence, the following inequality holds.

$$\hat{Q}(1)^2\left(\frac{1}{2n} + \frac{1}{2\sqrt{(1-\epsilon)n}} - \frac{1}{2}\exp\left(-\frac{\lfloor n^{2/3}\rfloor^2}{n}\right)\right) > \|\hat{Q}_n\|^2.$$

Therefore, we obtain the upper bound. That is, the inequality

$$\frac{1}{2\sqrt{n}} + \frac{1}{2\sqrt{1-\epsilon}} - \frac{\sqrt{n}}{2}\exp\left(-\frac{\lfloor n^{2/3}\rfloor^2}{n}\right) > \frac{\|\hat{Q}_n\|^2\sqrt{n}}{\hat{Q}_n(1)^2} \tag{4.9}$$

holds for $n \geq L$.

Next, we show the lower bound. From (4.8) we have

$$|\hat{Q}_n(\zeta^j)|^2 > \hat{Q}_n(1)^2\prod_{k=1}^{j}\left(1-\sin\frac{(2k-1)\pi}{2n}\right)^2.$$

Take any $\epsilon > 0$. Then, there exists an integer M such that the above lemma holds for $\theta = M^{-1/3}\pi$. Take any $n \geq M$. Since for $\lfloor n^{2/3}\rfloor \geq k \geq 1$ we have

$$\frac{\pi}{M^{1/3}} \geq \frac{\pi}{n^{1/3}} \geq \frac{(2n^{2/3}-1)\pi}{2n} \geq \frac{(2k-1)\pi}{2n},$$

the following inequalities follow from Lemma 4.8.

$$1-\sin\frac{(2k-1)\pi}{2n} \geq 1-\frac{(2k-1)\pi}{2n} \geq \exp\left(\frac{-(1+\epsilon)(2k-1)\pi}{2n}\right).$$

Hence, for $\lfloor n^{2/3} \rfloor \geq k \geq 1$ we have

$$|\hat{Q}_n(\zeta^k)|^2 > \hat{Q}_n(1)^2 \exp\left(\sum_{j=1}^{k} \frac{-(1+\epsilon)(2k-1)\pi}{n}\right) = \hat{Q}_n(1)^2 \exp\left(\frac{-(1+\epsilon)\pi j^2}{n}\right).$$

Therefore, the following inequalities hold.

$$\begin{aligned} \|\hat{Q}_n\|^2 &> \frac{1}{n} \sum_{k=1}^{\lfloor n^{2/3} \rfloor} |\hat{Q}_n(\zeta^k)|^2 > \frac{1}{n} \sum_{k=1}^{\lfloor n^{2/3} \rfloor} \left(\hat{Q}_n(1)^2 \exp\left(\frac{-(1+\epsilon)\pi k^2}{n}\right)\right) \\ &= \frac{\hat{Q}_n(1)^2}{n} \sum_{k=1}^{\lfloor n^{2/3} \rfloor} \exp\left(\frac{-(1+\epsilon)\pi k^2}{n}\right). \end{aligned}$$

The following estimation holds.

$$\sum_{k=1}^{\lfloor n^{2/3} \rfloor} \exp\left(\frac{-(1+\epsilon)\pi k^2}{n}\right) > \int_1^{\lfloor n^{2/3} \rfloor + 1} \exp\left(\frac{-(1+\epsilon)\pi x^2}{n}\right) dx.$$

For $a > 0$ and $c \geq 1$ we have

$$\begin{aligned} \int_1^a \exp(-cx^2)dx &= \int_0^\infty \exp(-cx^2)dx - \int_0^1 \exp(-cx^2)dx - \int_a^\infty \exp(-cx^2)dx \\ &> \frac{\sqrt{\pi}}{2\sqrt{c}} - 1 - \int_a^\infty x \exp(-cx^2)dx \end{aligned}$$

and

$$\int_a^\infty x \exp(-cx^2)dx = \left[-\frac{\exp(-cx^2)}{2c}\right]_a^\infty = \frac{\exp(-ca^2)}{2c}.$$

Hence, the inequalities

$$\begin{aligned} \sum_{k=1}^{\lfloor n^{2/3} \rfloor} \exp\left(\frac{-(1+\epsilon)\pi k^2}{n}\right) &> \frac{\sqrt{n}}{2\sqrt{1+\epsilon}} - 1 - \frac{n \exp\left(-\frac{(1+\epsilon)\pi}{n}(\lfloor n^{2/3} \rfloor + 1)^2\right)}{2(1+\epsilon)\pi} \\ &> \frac{\sqrt{n}}{2\sqrt{1+\epsilon}} - 1 - \frac{n \exp\left(-(1+\epsilon)\pi n^{1/3}\right)}{2(1+\epsilon)\pi} \end{aligned}$$

hold. Therefore, we have

$$\begin{aligned} \|\hat{Q}_n\|^2 &> \frac{1}{n} \sum_{k=1}^{\lfloor n^{2/3} \rfloor} |\hat{Q}_n(\zeta^k)|^2 > \frac{\hat{Q}_n(1)^2}{n} \sum_{k=1}^{\lfloor n^{2/3} \rfloor} \exp\left(\frac{-(1+\epsilon)\pi j^2}{n}\right) \\ &> \frac{\hat{Q}_n(1)^2}{n} \int_1^{\lfloor n^{2/3} \rfloor + 1} \exp\left(\frac{-(1+\epsilon)\pi x^2}{n}\right) dx \\ &> \hat{Q}_n(1)^2 \left(\frac{1}{2\sqrt{n(1+\epsilon)}} - \frac{1}{n} - \frac{\exp\left(-(1+\epsilon)\pi n^{1/3}\right)}{2(1+\epsilon)\pi}\right). \end{aligned}$$

Then, we obtain the lower bound. That is, the inequality

$$\frac{\|\hat{Q}_n\|^2\sqrt{n}}{\hat{Q}_n(1)^2} > \frac{1}{2\sqrt{1+\epsilon}} - \frac{1}{\sqrt{n}} - \frac{\sqrt{n}\exp\left(-(1+\epsilon)\pi n^{1/3}\right)}{2(1+\epsilon)\pi} \tag{4.10}$$

holds for $n \geq M$. Combining (4.9) and (4.10), we have the statement (4.7) for $N = \max\{L, M\}$ when n is odd.

Next we prove the statement when n is even. Let ζ_{4n} be $\exp(\pi\sqrt{-1}/2n)$. Note that ζ_{4n}^2 is a primitive $2n$th root of unity. Then, we have

$$\|\hat{Q}_n\|^2 = \frac{1}{2n}\sum_{k=1}^{2n} |\hat{Q}(\zeta_{4n}^{2k-1})|^2 = \frac{1}{n}\sum_{k=1}^{n/2} |\hat{Q}_n(\zeta_{4n}^{2k-1})|^2.$$

The relation between $|\hat{Q}_n(\zeta_{4n}^{2k+1})|^2$ and $|\hat{Q}_n(\zeta_{4n}^{2k-1})|^2$ is as follows.

$$|\hat{Q}_n(\zeta_{4n}^{2k+1})|^2 = |\hat{Q}_n(\zeta_{4n}^{2k-1})|^2 \frac{|\zeta_{4n}^{2k-1} + \zeta_{4n}^{-n-1}|^2}{|\zeta_{4n}^{2k-1} + \zeta_{4n}^{n-1}|^2} = |\hat{Q}_n(\zeta_{4n}^{2k-1})|^2 \frac{|1 + \zeta_{4n}^{-n-2k}|^2}{|1 + \zeta_{4n}^{n-2k}|^2}.$$

Since

$$|1 + \zeta_{4n}^{j}|^2 = (1 + \zeta_{4n}^{j})(1 + \zeta_{4n}^{-j}) = 2\left(1 + \cos\frac{j\pi}{2n}\right),$$

we have

$$|\hat{Q}_n(\zeta_{4n}^{2k+1})|^2 = |\hat{Q}_n(\zeta_{4n}^{2k-1})|^2 \frac{1 + \cos(\frac{\pi}{2} + \frac{k\pi}{n})}{1 + \cos(\frac{\pi}{2} - \frac{k\pi}{n})} = |\hat{Q}_n(\zeta_{4n}^{2k-1})|^2 \frac{1 - \sin\frac{k\pi}{n}}{1 + \sin\frac{k\pi}{n}}.$$

From similar arguments for odd n, given $\epsilon > 0$ there exists an integer N such that the following inequalities hold for any even integer $n \geq N$.

$$\frac{1}{2\sqrt{1-\epsilon}} - \frac{\sqrt{n}}{2}\exp\left(-\frac{\lfloor n^{2/3}\rfloor^2}{n}\right) > \frac{\|\hat{Q}_n\|^2\sqrt{n}}{|\hat{Q}_n(\zeta_{4n})|^2}$$
$$> \frac{1}{2\sqrt{1+\epsilon}} - \frac{1}{\sqrt{n}} - \frac{\sqrt{n}\exp\left(-(1+\epsilon)\pi n^{1/3}\right)}{2(1+\epsilon)\pi}.$$

That is, we have

$$\frac{\|\hat{Q}_n\|^2\sqrt{n}}{|\hat{Q}_n(\zeta_{4n})|^2} \to \frac{1}{2} \tag{4.11}$$

as n tends to infinity.

According to the following Lemma, we have

$$\lim_{n\to\infty} \frac{|\hat{Q}_n(\zeta_{4n})|^2}{\|\hat{Q}_n\|_D^2} = 1,$$

and combining with (4.11), we have the statement. □

Lemma 4.10.

$$\lim_{n\to\infty} \frac{|\hat{Q}_n(\zeta_{4n})|^2}{\|\hat{Q}_n\|_D^2} = 1.$$

Proof. Since the following inequalities

$$\hat{Q}_n(1) = \prod_{k=-n/2+1}^{n/2} (1+\zeta_{4n}^{2k-1}), \qquad \hat{Q}_n(\zeta_{4n}) = \prod_{k=-n/2+1}^{n/2} (\zeta_{4n}+\zeta_{4n}^{2k-1})$$

hold, we have

$$\begin{aligned}
\hat{Q}_n(1)^2 &= \prod_{k=-n/2+1}^{n/2} (1+\zeta_{4n}^{2k-1})^2 = \prod_{k=1}^{n/2} (1+\zeta_{4n}^{2k-1})^2 (1+\zeta_{4n}^{-2k+1})^2 \\
&= \prod_{k=1}^{n/2} \left(2+2\cos\frac{(2k-1)\pi}{2n}\right)^2, \\
|\hat{Q}_n(\zeta_{4n})|^2 &= \prod_{k=-n/2+1}^{n/2} |1+\zeta_{4n}^{2k-2}|^2 = \prod_{k=1}^{n/2} |1+\zeta_{4n}^{2k-2}|^2 \cdot |1+\zeta_{4n}^{-2k}|^2 \\
&= \prod_{k=1}^{n/2} \left(2\cos\frac{\pi}{2n}+2\cos\frac{(2k-1)\pi}{2n}\right)^2.
\end{aligned}$$

Therefore, the following inequalities hold.

$$\begin{aligned}
\frac{|\hat{Q}_n(\zeta_{4n})|^2}{\hat{Q}_n(1)^2} &= \prod_{k=1}^{n/2} \frac{\left(\cos\frac{\pi}{2n}+\cos\frac{(2k-1)\pi}{2n}\right)^2}{\left(1+\cos\frac{(2k-1)\pi}{2n}\right)^2} \geq \prod_{k=1}^{n/2} \frac{\left(1-\frac{\pi^2}{8n^2}+\cos\frac{(2k-1)\pi}{2n}\right)^2}{\left(1+\cos\frac{(2k-1)\pi}{2n}\right)^2} \\
&= \prod_{k=1}^{n/2} \left(1-\frac{\pi^2}{8n^2\left(1+\cos\frac{(2k-1)\pi}{2n}\right)^2}\right)^2 > \left(1-\frac{\pi^2}{8n^2}\right)^n.
\end{aligned}$$

That is, we have

$$1 \geq \frac{|\hat{Q}_n(\zeta_{4n})|^2}{\hat{Q}_n(1)^2} > \left(1-\frac{\pi^2}{8n^2}\right)^n.$$

Therefore, we have

$$\frac{|\hat{Q}_n(\zeta_{4n})|^2}{\hat{Q}_n(1)^2} \to 1$$

as n tends to infinity. □

Acknowledgement

The first author wishes to thank Prashant Batra, Ludwig Elsner and Arnold Schönhage for their valuable hints and fruitful discussion. The authors thank Kyoko Makino for computing the rigorous lower and upper bounds in Table 1. Furthermore our thanks to two unknown referees for their constructive comments.

References

[1] M. Berz. From Taylor series to Taylor models. In *Nonlinear problems in accelerator physics, AIP Conference proceedings*, number CP405, pages 1–27, 1997.

[2] A. Böttcher and S. Grudsky. *Spectral properties of banded Toeplitz matrices*. SIAM, Philadelphia, 2005.

[3] A. Böttcher and S. Grudsky. Structured condition numbers of large Toeplitz matrices are rarely better than usual condition numbers. *Numerical Linear Algebra and its Applications*, 12:95–102, 2005.

[4] A. Böttcher and K. Rost. Topics in the numerical linear algebra of Toeplitz and Hankel matrices. *Mitt. Ges. Angew. Math. Mech.*, 27(2):174–188, 2004.

[5] D.W. Boyd. Two sharp inequalities for the norm of a factor of a polynomial. *Mathematika*, 39:341–349, 1992.

[6] D.W. Boyd. Sharp inequalities for the product of polynomials. *Bull. London Math. Soc.*, 26:449–454, 1994.

[7] S. Dallwig, A. Neumaier, and H. Schichl. GLOPT – a program for constrained global optimization. In I.M. Bomze et al., editor, *Developments in global optimization*, pages 19–36. Kluwer Academic Publishers, 1997.

[8] I. Gohberg and A.A. Semencul. The inversion of finite Toeplitz matrices and their continual analogues. *Matem. Issled*, 7:201–223, 1972.

[9] G. Heinig. private communication.

[10] D.J. Higham and N.J. Higham. Backward error and condition of structured linear systems. *SIAM J. Matrix Anal. Appl.*, 13(1):162–175, 1992.

[11] E. Kaltofen, L. Bin, Y. Zhengfeng, and Z. Lihong. Exact certification in global polynomial optimization via sums-of-squares of rational functions with rational coefficients. submitted for publication, 2009.

[12] H. Kneser. Das Maximum des Produkts zweier Polynome. In *Sitz. Preuss. Akad. Wiss., Phys.-Math. Kl.*, pages 426–431, 1934.

[13] A. Neumaier. Complete search in continuous global optimization and constraint satisfaction. *Acta Numerica*, 13:271–369, 2004.

[14] S.M. Rump. Structured perturbations Part I: Normwise distances. *SIAM J. Matrix Anal. Appl. (SIMAX)*, 25(1):1–30, 2003.

[15] S.M. Rump. A Model Problem for Global optimization. submitted for publication, 2009.

[16] E.B. Saff and T. Sheil-Small. Coefficient and integral mean estimates for algebraic and trigonometric polynomials with restricted zeros. *J. London Math. Soc.*, 9:16–22, 1974.

[17] T. Sheil-Small. *Complex polynomials*, volume 75 of *Cambridge Studies in Advanced Mathematics*. Cambridge University Press, 2002.

Siegfried M. Rump
Institute for Reliable Computing
Hamburg University of Technology
Schwarzenbergstraße 95
D-21071 Hamburg, Germany

and

Waseda University
Faculty of Science and Engineering
3–4–1 Okubo, Shinjuku-ku
Tokyo 169–8555, Japan
e-mail: rump@tu-harburg.de

Hiroshi Sekigawa
NTT Communication Science Laboratories
Nippon Telegraph and Telephone Corporation
3–1 Morinosato-Wakamiya, Atsugi-shi
Kanagawa, 243-0198, Japan
e-mail: sekigawa@theory.brl.ntt.co.jp

Operator Theory:
Advances and Applications, Vol. 199, 421–430

A New Algorithm for Finding Positive Eigenvectors for a Class of Nonlinear Operators Associated with M-matrices

Yuriy V. Shlapak

Abstract. In this paper we state the sufficient conditions for the existence and uniqueness of positive eigenvectors for a class of nonlinear operators associated with M-matrices. We also show how to construct a convergent iterative process for finding these eigenvectors. The details of numerical implementation of this algorithm for some spectral methods of discretization of elliptic partial differential equations are also discussed. Some results of numerical experiments for the Gross-Pitaevskii Equation with non-separable potentials in a rectangular domain are given in the end of the paper.

Mathematics Subject Classification (2000). 47J10.

Keywords. Positive eigenvectors, Nonlinear operators, M-matrices, Monotone fixed point theorem, Gross-Pitaevskii equation.

1. Introduction

Many problems of the modern physics require finding positive eigenvectors of some nonlinear elliptic operators. After the discretization these problems can be reduced to problems of finding positive eigenvectors of nonlinear operators acting from $\mathbb{R}^n$ to $\mathbb{R}^n$.

We consider the following problem: find a column vector $X \in \mathbb{R}^n$ that satisfies the equation

$$AX + V \circ X + F(X) = \lambda X \tag{1.1}$$

where λ is a real positive constant (eigenvalue), A is an M-matrix of size n, symbol $\circ$ is used to denote the Hadamard product of two matrices, $V = [v_1, \ldots, v_n]^T$ is a column vector that is called a potential, and $F(X)$ is a nonlinear vector function from $\mathbb{R}^n$ to $\mathbb{R}^n$ that depends only on the vector X. In addition to finding the vector X, we would like to establish conditions that are sufficient for the existence and uniqueness of a vector X in this problem.

To prove the main result of this paper, we use techniques that are similar to ones given in [1]. We consider a more complex equation (1.1), which has a linear term $V \circ X$ in it. This is a generalization of the equation considered in [1]. The discretizations of many important equations of physics have a linear term in them. For example, the Gross-Pitaevskii Equation from quantum physics

$$-\Delta u + V(x,y,z)u + ku^3 = \lambda u \tag{1.2}$$

in one-dimensional case can be discretized by using the 3-point central finite differences and a uniform mesh with the step h into the equation (1.1) with the matrix A of the form:

$$A = \frac{1}{h^2} \cdot \begin{bmatrix} 2 & -1 & & & 0 \\ -1 & 2 & -1 & & \\ & -1 & \ddots & \ddots & \\ & & \ddots & 2 & -1 \\ 0 & & & -1 & 2 \end{bmatrix}$$

We should point out that the algorithm described in [1] and [2] allows practical computation of positive eigenvectors for some special types of equation (1.1). In this paper we provide both the theoretical justification and practical algorithm for finding positive eigenvectors of equation (1.1) in its most general form.

We can reformulate the problem of finding a solution of equation (1.1) as a problem of finding a fixed point of some transformation, namely we need to find $X \in \mathbb{R}^n$ such that $X = S(X)$. We define $S(X)$ by the formula

$$S(X) = (cI + A)^{-1}((c+\lambda)X - V \circ X - F(X)) \tag{1.3}$$

where $c > 0$ is some positive constant (usually we will choose it to be a large number). In order to prove the main theorem of this article, we will also need a Monotone Fixed Point Theorem [3] applied in the context of our problem.

Theorem 1.1 (Monotone Fixed Point Theorem). *Consider a space $\mathbb{R}^n$ with the partial order relation $<$ defined in the following way: for any two vectors $X = [x_1, \ldots, x_n]^T \in \mathbb{R}^n$ and $Y = [y_1, \ldots, y_n]^T \in \mathbb{R}^n$ we will say that X is smaller than Y (denoted as $X < Y$) if $x_i \le y_i$ for all $i = 1, \ldots, n$ and $x_i < y_i$ for at least one i.*

If $X \in \mathbb{R}^n$, $Y \in \mathbb{R}^n$ and $X < Y$ we can define the interval $[X, Y] \subset \mathbb{R}^n$ in the following way: we will say that $T \in [X, Y]$ if $X \le T \le Y$.

Let $Y \in \mathbb{R}^n$ and $Z \in \mathbb{R}^n$ are such that $Y < Z$ and let S: $\mathbb{R}^n \to \mathbb{R}^n$ be defined and continuous on the interval $[Y, Z]$ and suppose the following properties are satisfied

1) $Y < S(Y) < Z$
2) $Y < S(Z) < Z$
3) $Y \le X_1 < X_2 \le Z$ *implies* $Y < S(X_1) < S(X_2) < Z$

Then

a) *the fixed point iteration $X_k = S(X_{k-1})$ with $X_0 = Y$ converges: $X_k \to X_*$, $S(X_*) = X_*$, $Y < X_* < Z$*

b) *the fixed point iteration* $X_k = S(X_{k-1})$ *with* $X_0 = Z$ *converges:* $X_k \to X^*$, $S(X^*) = X^*$, $Y < X^* < Z$
c) *if* X *is a fixed point of* S *in* $[Y, Z]$ *then* $X_* \le X \le X^*$
d) S *has a unique fixed point in* $[Y, Z]$ *if and only if* $X_* = X^*$

2. Monotone iteration for finding positive eigenvector

Now we are ready to prove our main result of this article. It will state the existence and uniqueness of the positive solution of equation (1.1) and the convergence of the iteration sequence $X_{n+1} = S(X_n)$, where $S(X)$ is defined by (1.3), to this solution. As was mentioned above, this theorem is a generalization of a theorem from [1], which is a particular case of our theorem when $V = \mathbf{0}$.

Theorem 2.1 (Monotone iteration for finding positive eigenvector). *Suppose in the equation* (1.1) A *is an M-matrix,* μ *is the smallest positive eigenvalue of the matrix* A. *Now we will use the notation* $X = [x_1, \dots, x_n]^T$ *and* $V = [v_1, \dots, v_n]^T$. *Let* $\lambda > \mu + \max\limits_{1 \le i \le n} v_i$, *and let*

$$F(X) = \begin{bmatrix} f_1(x_1) \\ f_2(x_2) \\ \vdots \\ f_n(x_n) \end{bmatrix} \tag{2.1}$$

be such a vector function that for $i = 1, \dots, n$ *the components* $f_i : (0, \infty) \to (0, \infty)$ *are* C^1*-functions satisfying conditions*

$$\lim_{t \to 0} \frac{f_i(t)}{t} = 0, \quad \lim_{t \to \infty} \frac{f_i(t)}{t} = \infty \tag{2.2}$$

Then (1.1) *has a positive solution.*

If, in addition to the conditions given above, for $i = 1, \dots, n$ *we have*

$$\frac{f_i(s)}{s} < \frac{f_i(t)}{t} \quad \text{whenever } 0 < s < t \tag{2.3}$$

then a positive solution of (1.1) *is unique and there exists a vector* X_0 *and a positive constant* c *such that the sequence* $X_{n+1} = S(X_n)$, *where*

$$S(X) = (cI + A)^{-1}((c + \lambda)X - V \circ X - F(X)),$$

converges to the unique positive solution of (1.1).

Proof. The proof is based on showing that $S(X)$ satisfies the conditions of the Monotone Fixed Point Theorem (Theorem 1.1). It will guarantee the existence of the positive solution and its uniqueness.

First of all, we choose a real number β_1 small enough so that $(\lambda - v_i - \mu)(\beta_1 p_i) > f_i(\beta_1 p_i)$ for all $i = 1, \dots, n$ and we choose $\beta_2 > \beta_1$ large enough so

that $(\lambda - v_i - \mu)(\beta_2 p_i) < f_i(\beta_2 p_i)$ for all $i = 1, \ldots, n$. This is always possible in view of limits (2.2). We also choose a positive number c such that

$$c > \max_{1 \le i \le n} \left(\sup_{\beta_1 p_i \le t \le \beta_2 p_i} |f_i'(t)| \right) - \lambda + \max_{1 \le i \le n} v_i \tag{2.4}$$

and then reformulate the problem of finding a solution of equation (1.1) into a problem of finding a fixed point of the transformation $S : \mathbb{R}^n \to \mathbb{R}^n$, where $S(X)$ is defined as in (1.3), i.e.,

$$S(X) = (cI + A)^{-1}((c + \lambda)X - V \circ X - F(X)) \tag{2.5}$$

To prove the existence of a fixed point of (1.3) it is enough to show that $S(X)$ satisfies the conditions of Monotone Fixed Point Theorem given above with $Y = \beta_1 p$ and $Z = \beta_2 p$.

First of all, let us show that the condition 1) of Monotone Fixed Point Theorem is satisfied, i.e., $Y < S(Y) < Z$. In the proof below we will use the obvious fact that $p = (A + cI)^{-1}(c + \mu)p$.

Since $(A + cI)^{-1}u > 0$ whenever $u > 0$ (i.e., $(A + cI)^{-1}$ is a positive matrix) it suffices to show that

$$(c + \mu)(\beta_1 p) < (c + \lambda)(\beta_1 p) - V \circ (\beta_1 p) - F(\beta_1 p) < (c + \mu)(\beta_2 p) \tag{2.6}$$

We start from proving the left part of this double inequality. The argument below is valid for any $i = 1, \ldots, n$. From $(\lambda - v_i - \mu)(\beta_1 p_i) > f_i(\beta_1 p_i)$ we have $(\lambda - v_i - \mu)(\beta_1 p_i) - f_i(\beta_1 p_i) > 0$ and so $(c + \mu)(\beta_1 p_i) + (\lambda - v_i - \mu)(\beta_1 p_i) - f_i(\beta_1 p_i) > (c + \mu)(\beta_1 p_i)$. After cancelling out the term $\mu\beta_1 p_i$ in the left part we get $(c + \lambda)(\beta_1 p_i) - v_i(\beta_1 p_i) - f_i(\beta_1 p_i) > (c + \mu)(\beta_1 p_i)$ which is exactly the componentwise notation of the left part of the double inequality (2.6).

Now we will prove the right part of the double inequality (2.6). By our choice of β_2 for any $i = 1, \ldots, n$ we have $f_i(\beta_2 p_i) > (\lambda - v_i - \mu)(\beta_2 p_i)$. It can be rewritten as $f_i(\beta_1 p_i) + (f_i(\beta_2 p_i) - f_i(\beta_1 p_i)) > (\lambda - v_i - \mu)(\beta_2 p_i)$ or, if we estimate the change in f by its derivative multiplied by the change in argument of f, and use (2.4), we will obtain that $f_i(\beta_1 p_i) + (c + \lambda - v_i)((\beta_2 p_i) - (\beta_1 p_i)) > (\lambda - v_i - \mu)(\beta_2 p_i)$ and after moving some terms into the left part we get $f_i(\beta_1 p_i) - (c + \lambda - v_i)(\beta_1 p_i) > (\lambda - v_i - \mu)(\beta_2 p_i) - (c + \lambda - v_i)(\beta_2 p_i)$. After simplification it becomes $f_i(\beta_1 p_i) - (c + \lambda - v_i)(\beta_1 p_i) > -(\mu + c)(\beta_2 p_i)$ or, if we multiply it by -1, it becomes $(\mu + c)(\beta_2 p_i) > (c + \lambda)(\beta_1 p_i) - v_i(\beta_1 p_i) - f_i(\beta_1 p_i)$, which is exactly the right part of equality (2.6).

We also need to prove that the second condition of Monotone Fixed Point Theorem is satisfied, namely, $Y < S(Z) < Z$. Due to the fact that $(A + cI)^{-1}$ is a positive matrix it suffices to show only that

$$(c + \mu)(\beta_1 p) < (c + \lambda)(\beta_2 p) - V \circ (\beta_2 p) - F(\beta_2 p) < (c + \mu)(\beta_2 p) \tag{2.7}$$

Let us show now that the right part of this double inequality holds. The argument below is valid for any $i = 1, \ldots, n$. From $(\lambda - v_i - \mu)(\beta_2 p_i) < f_i(\beta_2 p_i)$ we have $(\lambda - v_i - \mu)(\beta_2 p_i) - f_i(\beta_2 p_i) < 0$ and so we can write $(c + \mu)(\beta_2 p_i) + (\lambda - v_i -$

$\mu)(\beta_2 p_i) - f_i(\beta_2 p_i) < (c+\mu)(\beta_2 p_i)$ and then after cancelling out the term $\mu\beta_2 p_i$ in the left part we get $(c+\lambda)(\beta_2 p_i) - v_i(\beta_2 p_i) - f_i(\beta_2 p_i) < (c+\mu)(\beta_2 p_i)$, which is exactly the right part of the double inequality (2.7).

Now we will prove that the left part of the double inequality (2.7) holds. From our choice of β_1 we have $f_i(\beta_1 p_i) < (\lambda - v_i - \mu)(\beta_1 p_i)$, which can be written as $f_i(\beta_2 p_i) + (f_i(\beta_1 p_i) - f_i(\beta_2 p_i)) < (\lambda - v_i - \mu)(\beta_1 p_i)$ or, if we estimate the change in f by its derivative multiplied by the change in argument of f, and use (2.4), we will get $f_i(\beta_2 p_i) + (c + \lambda - v_i)((\beta_1 p_i) - (\beta_2 p_i)) < (\lambda - v_i - \mu)(\beta_1 p_i)$. After moving some terms into the left part, it becomes $f_i(\beta_2 p_i) - (c + \lambda - v_i)(\beta_2 p_i) < (\lambda - v_i - \mu)(\beta_1 p_i) - (c + \lambda - v_i)(\beta_1 p_i)$ and after some simplification it becomes $f_i(\beta_2 p_i) - (c + \lambda - v_i)(\beta_2 p_i) < -(\mu + c)(\beta_1 p_i)$. Finally, if we multiply it by -1 we will get $(\mu + c)(\beta_1 p_i) < (c + \lambda)(\beta_2 p_i) - v_i(\beta_2 p_i) - f_i(\beta_2 p_i)$ which is exactly the componentwise notation of the left part of double inequality (2.7).

Now we have to show that if $\beta_1 p \le X_1 < X_2 \le \beta_2 p$ then $S(X_1) < S(X_2)$. It will guarantee that the condition 3) of Monotone Fixed Point Theorem is satisfied. If we denote ith components of vectors X_1 and X_2 as x_{1i} and x_{2i} correspondingly then we can write the ith component of $S(X_2) - S(X_1)$ as $(A+cI)^{-1}((c+\lambda-v_i)x_{2i} - f_i(x_{2i}) - (c+\lambda-v_i)x_{1i} + f_i(x_{1i}))$. Due to the fact that $(A+cI)^{-1}$ is a positive matrix it suffices to show only that $(c+\lambda-v_i)x_{2i} - f_i(x_{2i}) - (c+\lambda-v_i)x_{1i} + f_i(x_{1i}) > 0$ for each $i = 1, \ldots, n.$. It can be shown by using some simple algebraic transformations and estimating the change in f by the maximum of its derivative multiplied by the change in argument of f and then using (2.4). For each i we have $(c+\lambda-v_i)x_{2i} - f_i(x_{2i}) - (c+\lambda-v_i)x_{1i} + f_i(x_{1i}) = (c+\lambda-v_i)(x_{2i} - x_{1i}) - (f_i(x_{2i} - f_i(x_{1i})) > (c+\lambda-v_i)(x_{2i} - x_{1i}) - (c+\lambda-v_i)(x_{2i} - x_{1i}) = 0$. So $\beta_1 p \le X_1 < X_2 \le \beta_2 p$ implies $S(X_1) < S(X_2)$.

We have checked that the conditions 1)-3) of Monotone Fixed Point Theorem are satisfied. It guarantees that there exists at least one fixed point of the transformation defined by (1.3) that will also be a solution of the equation (1.1). The Monotone Fixed Point Theorem also implies that if we choose $X_0 = \beta_1 p$ or $X_0 = \beta_2 p$ then the sequence $X_{n+1} = S(X_n)$ will converge to a fixed point of the transformation $S(X)$ (which will also be a solution of equation (1.1)). And finally, it guarantees that such fixed point(s) of $S(X)$ will lie between $\beta_1 p$ and $\beta_2 p$, which guarantees the positivity of all components of the solution.

Now we will prove the uniqueness of the positive solution under the conditions of Theorem 2.1. Suppose now that the condition (2.3), i.e.,

$$\frac{f_i(s)}{s} < \frac{f_i(t)}{t} \quad \text{whenever } 0 < s < t$$

is satisfied. We want to show that in this case for any two positive solutions X^* and X_* it must be $X^* = X_*$.

Since both X^* and X_* are solutions of the equation (1.1), we have

$$AX^* + V \circ X^* + F(X^*) = \lambda X^*$$
$$AX_* + V \circ X_* + F(X_*) = \lambda X_* \ .$$

Pre-multiplying the first equation by X_*^T and the second equation by X^{*T} and subtracting the second equation from the first one we will get (using an obvious identity $X^{*T} \cdot V \circ X_* = X_*^T \cdot V \circ X^*$) that $X_*^T F(X^*) = X^{*T} F(X_*)$ or, equivalently, that using the componentwise notation we can write

$$0 = \sum_{i=1}^{n} (f_i(x_i^*) x_{*i} - f_i(x_{*i}) x_i^*) = \sum_{i=1}^{n} \left(x_{*i} x_i^* \left(\frac{f_i(x_i^*)}{x_i^*} - \frac{f_i(x_{i*})}{x_{i*}} \right) \right)$$

Since x_i^* and x_{i*} are always positive, the only case when this sum can be zero is when for each $i = 1, \ldots, n$

$$\frac{f_i(x_i^*)}{x_i^*} = \frac{f_i(x_{i*})}{x_{i*}} \tag{2.8}$$

which, by condition (2.3), can only be true when $X_* = X^*$. So the uniqueness of the positive solution of (1.1) under the condition (2.3) is also proven.

The proof of Theorem 2.1 is completed. □

3. The numerical implementation in two-dimensional case

As an application of the algorithm described above, we consider solving a discretized boundary value problem for a nonlinear elliptic PDE in two dimensions. Suppose our domain is a square $\Omega = \{(x, y) | -T \leq x \leq T, -T \leq y \leq T\}$ and we need to find a function $u(x, y)$ such that for every $(x, y) \in \Omega$ it satisfies the equation

$$-\Delta u(x, y) + V(x, y) u(x, y) + f(u(x, y)) = \lambda u(x, y) \tag{3.1}$$

where Δ is Laplace operator, $V(x, y)$ is a function that is called a potential, f is a nonlinear function and λ is a positive constant. The function u must also satisfy the following boundary condition:

$$u|_\Gamma = 0 \tag{3.2}$$

where Γ is the boundary of a square Ω.

We approximate the function $u(x, y)$ by a linear combination of normalized Legendre polynomials:

$$u(x, y) = \sum_{i,j=1}^{N} \alpha_{ij} p_i \left(\frac{x}{T} \right) p_j \left(\frac{y}{T} \right), \tag{3.3}$$

where $p_i(x)$ is the normalized Legendre polynomial of degree i. We use the set of collocation points $\{(x_k, y_m) | \, k, m = 1, \ldots, N+1\}$, where x_k and y_k are zeros of the $N+1$-st normalized Legendre polynomial $p_{N+1}(x)$ that are multiplied by T. Naturally, we have

$$u(x_k, y_m) = \sum_{i,j=1}^{N} \alpha_{ij} p_i \left(\frac{x_k}{T} \right) p_j \left(\frac{y_m}{T} \right) \tag{3.4}$$

and now we define U to be the matrix of values of function u at the collocation points, i.e., $U_{km} = u(x_k, y_m)$. We similarly define matrix V as the matrix of values of potential V at the collocation points, i.e., $V_{km} = V(x_k, y_m)$.

It is widely known (see [4] or the Appendix of [2]) that if u is a linear combination of normalized Legendre polynomials as given in (3.4), and U is the matrix of values of u at the collocation points given above, then the matrices of values of partial derivatives $\frac{\partial^2 u}{\partial x^2}$ and $\frac{\partial^2 u}{\partial y^2}$ at the collocation points can be found correspondingly as DU and UD, where D is a known square matrix that depends only on N and T. The matrix D plays the role of the double differentiation operator for our functions at the collocation points. The boundary value problem (3.1)–(3.2) at the collocation points can be written as the following matrix equation:

$$-DU - UD + V \circ U + f(U) = \lambda U, \tag{3.5}$$

where $\circ$ is a symbol of Hadamard product of two matrices. The matrix D is not an M-matrix, but our numerical experiments showed that the inverse of the matrix $cI + D$ was a positive matrix in all our experiments. So the statement and the results of the Theorem 2.1 are also valid for the matrix equations that involve D instead of A. The iteration sequence generated by transformation (1.3) for finding the approximate solution of the equation (3.5) is defined by relation

$$cU_{n+1} - DU_{n+1} - U_{n+1}D = cU_n + \lambda U_n - V \circ U_n - f(U_n) \tag{3.6}$$

In order to find U_{n+1} (when U_n is known) we use a technique described in [2], [4]. This technique makes extensive use of properties of orthogonal polynomials at certain points. The numerical implementation of this technique is relatively straightforward and gives excellent results in terms of computational speed and precision.

To illustrate the feasibility and effectiveness of the method described in this paper, we found positive solutions of the Gross-Pitaevskii Equation (1.2) with non-separable potentials (i.e., potentials that cannot be expressed as sums of two functions that depend only on one spacial variable each). This type of the Gross-Pitaevskii Equation is more difficult to work with than the one with a separable potential. For example, the method described in [1], [2] cannot be used for such equations.

In our experiments, the domain was a square $[-10, 10] \times [-10, 10]$. It took about 31 seconds for our personal computer to do 1000 iterations on a 128 by 128 mesh of collocation points. Some graphs and tables related to our numerical experiments are given next.

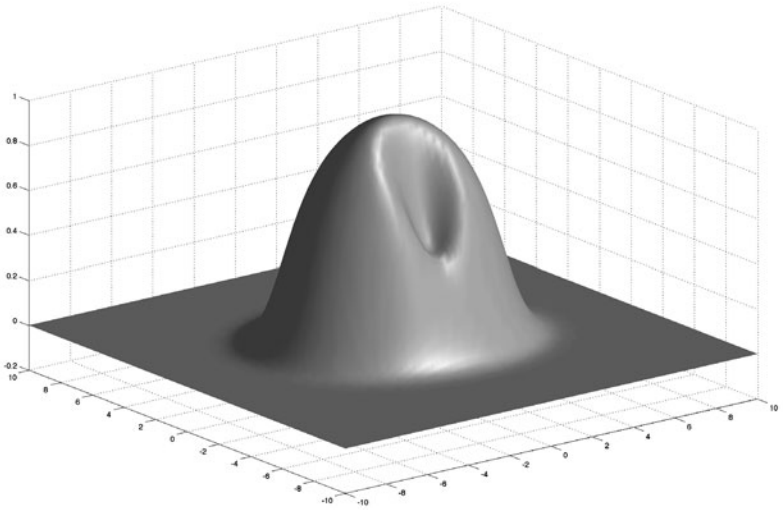


FIGURE 1. The solution for

$$V(x,y) = \ 0.6x^2 \ + \ 0.4y^2 \ + \ 7e^{-(x+2)^2-y^2}, \quad k = 10, \quad \lambda = 10$$

Table 1. Convergence results for

$$V(x,y) = 0.6x^2 + 0.4y^2 + 7e^{-(x+2)^2-y^2}, \quad k = 10, \quad \lambda = 10, \quad c = 100$$

Number of iterations	2-norm of the residue
2	14.4357
5	4.4489
10	3.7863
20	7.3529
50	46.2945
100	14.0350
200	0.6203
500	$1.6815 \cdot 10^{-5}$
1000	$2.4052 \cdot 10^{-9}$
2000	$2.4054 \cdot 10^{-9}$
5000	$2.4054 \cdot 10^{-9}$

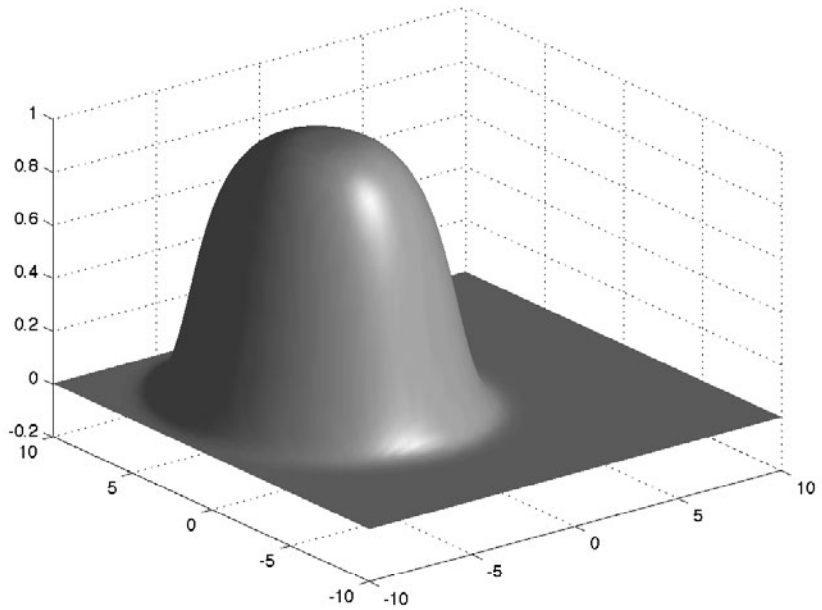


FIGURE 2. The solution for

$$V(x,y) = \ 0.1\left(\sqrt{0.6(x-2)^2+0.8(y+3)^2}\right)^3, \quad k=10, \quad \lambda=10$$

Table 2. Convergence results for

$$V(x,y) = 0.1\left(\sqrt{0.6(x-2)^2+0.8(y+3)^2}\right)^3, \quad k=10, \quad \lambda=10, \quad c=200$$

Number of iterations	2-norm of the residue
2	28.1881
5	9.8392
10	5.2898
20	6.2834
50	21.7926
100	83.8192
200	15.5093
500	0.0366
1000	$1.0538 \cdot 10^{-6}$
2000	$5.2324 \cdot 10^{-9}$
5000	$5.2321 \cdot 10^{-9}$

We should point out that increasing c in (3.6) slows down the convergence of our algorithm. However, as the condition (2.4) in Theorem 2.1 indicates, increasing c may be necessary to guarantee the convergence of the algorithm and our numerical experiments confirmed the necessity of this requirement.

References

[1] Choi Y.S., Koltracht I., McKenna P.J. A generalization of the Perron-Frobenius theorem for non-linear perturbations of Stiltjes Matrices. *Contemporary Mathematics*, Volume 281, 2001, pp. 325–330

[2] Choi Y.S., Javanainen J., Koltracht I., Kostrun M., McKenna P. J., Savytska N., A fast algorithm for the solution of the time-independent Gross-Pitaevskii equation. *Journal of Computational Physics*, 190 (2003), pp. 1–21

[3] Kantorovich L.V., Vulikh B.Z., Pinsker A.G., Functional Analysis in Semi-ordered Spaces. (in Russian Language). Moscow, GosIzdat Technico-Teoreticheskoi Literatury, 1950.

[4] Gottlieb D., Orszag S.A., Numerical Analysis of Spectral Methods: Theory and Applications. Philadelphia, Society for Industrial and Applied Mathematics, 1977.

Yuriy V. Shlapak
Department of Mathematics
University of Connecticut
196 Auditorium Road, Unit 3009
Storrs, CT 06269-3009, USA
e-mail: shlapak@math.uconn.edu

Operator Theory:
Advances and Applications, Vol. 199, 431–439

Hankel Minors and Pade Approximations

Eugene Tyrtyshnikov

To the memory of Georg Heinig

Abstract. Algebraic Pade theory is presented in a complete, brief and clear way as a corollary of one property of a sequence of nonzero leading minors in a semi-infinite Hankel matrix associated with a formal series.

Mathematics Subject Classification (2000). 15A12; 65F10; 65F15.

Keywords. Pade approximants; Hankel matrices; Toeplitz matrices; rational approximations.

1. Introduction

Algebraic theory of Pade approximations is in fact a theory of submatrices in a semi-infinite Hankel matrix. Basic assertions of this theory in [1] are formulated in an elegant and simple way. However, one could be disappointed by some proofs looking too tangled and even somewhat misleading. At the same time, algebraic facts are not in the focus of the extensive literature on rational approximations (see [7, 2, 3]) devoted chiefly to analytical questions and applications.

The purpose of this paper is a complete, brief and transparent presentation of the algebraic Pade theory as a corollary of a nice property of leading minors of a Hankel matrix, which is probably best exposed in the enlightening book by G. Heinig and K. Rost [5]. This property is related to a problem of rank-preserving augmentations [6] and was expounded in [10] as a base of the "method of jumps" for the inversion of Hankel matrices with some of leading minors equal to zero. Besides a certain development of the "method of jumps", in this paper we contribute with new proofs for a systematic presentation of the known statements of the algebraic Pade theory.

Supported by RFBR grants 08-01-00115, 09-01-91332, 09-01-112058, the State Contract П940 and Priority Research Programme of Department of Mathematical Sciences of Russian Academy of Sciences.

2. Series and matrices

Consider a formal series

$$f(x) = \sum_{i=0}^{\infty} a_i x^i.$$

We define a *Pade approximation of type* (m, n) for $f(x)$ as a pair of polynomials

$$u(x) = \sum_{i=0}^{m} u_i x^i, \quad v(x) = \sum_{i=0}^{n} v_i x^i$$

such that

$$f(x)v(x) - u(x) = O(x^{m+n+1}) \tag{2.1}$$

under an additional restriction

$$v(0) = 1. \tag{2.2}$$

It follows immediately that

$$f(x) - \frac{u(x)}{v(x)} = O(x^{m+n+1}).$$

Condition (2.1) is equivalent to the system of linear equations

$$\sum_{j=0}^{n} a_{i-j} v_j = 0, \quad m+1 \le i \le m+n,$$

or, in matrix notation,

$$\begin{bmatrix} a_{m+1} & a_m & \dots & a_{m-n+1} \\ a_{m+2} & a_{m+1} & \dots & a_{m-n} \\ \dots & \dots & \dots & \dots \\ a_{m+n} & a_{m+n-1} & \dots & a_m \end{bmatrix} \begin{bmatrix} v_0 \\ v_1 \\ \dots \\ v_n \end{bmatrix} = 0.$$

Taking into account the equation $v_0 = 1$, we obtain

$$\begin{bmatrix} a_m & \dots & a_{m-n+1} \\ \dots & \dots & \dots \\ a_{m+n-1} & \dots & a_m \end{bmatrix} \begin{bmatrix} v_1 \\ \dots \\ v_n \end{bmatrix} = - \begin{bmatrix} a_{m+1} \\ \dots \\ a_{m+n} \end{bmatrix}.$$

To clarify things, take $m = n = 3$, then

$$\begin{bmatrix} a_3 & a_2 & a_1 \\ a_4 & a_3 & a_2 \\ a_5 & a_4 & a_3 \end{bmatrix} \begin{bmatrix} v_1 \\ v_2 \\ v_3 \end{bmatrix} = - \begin{bmatrix} a_4 \\ a_5 \\ a_6 \end{bmatrix}.$$

Each entry in the coefficient matrix of this system is defined by the difference of the row and column indices – such matrices are called *Toeplitz matrices*. By taking the columns in the reverse order we obtain a matrix whose entries are defined by

the sum of row and column indices – such matrices are called *Hankel matrices*. By reversion of columns the above system reduces to the following one:

$$\begin{bmatrix} a_1 & a_2 & a_3 \\ a_2 & a_3 & a_4 \\ a_3 & a_4 & a_5 \end{bmatrix} \begin{bmatrix} v_3 \\ v_2 \\ v_1 \end{bmatrix} = - \begin{bmatrix} a_4 \\ a_5 \\ a_6 \end{bmatrix}.$$

In the general case we obtain a system of the form

$$\begin{bmatrix} a_{m-n+1} & \dots & a_m \\ \dots & \dots & \dots \\ a_m & \dots & a_{m+n-1} \end{bmatrix} \begin{bmatrix} v_n \\ \dots \\ v_1 \end{bmatrix} = - \begin{bmatrix} a_{m+1} \\ \dots \\ a_{m+n} \end{bmatrix}, \tag{2.3}$$

or, in short notation,

$$A_{mn} v^n = -a^{mn}. \tag{2.4}$$

The Hankel matrix A_{mn} is composed of the coefficients of the formal series $f(x)$ and defined by m and n as follows: a_m is located in its lower left corner and n is the matrix order. If $i < 0$, then $a_i = 0$ by definition. The right-hand side vector a^{mn} is the last column of the extended rectangular Hankel matrix $[A_{mn}, a^{mn}]$.

Thus, *the existence of a Pade approximation of type* (m, n) *is equivalent to the consistency of the linear system* (2.4). The latter means that

$$a^{mn} \in \mathrm{im} A_{mn},$$

and, by the Kronecker–Capelli theorem, is equivalent to the condition

$$\mathrm{rank} A_{mn} = \mathrm{rank}[A_{mn}, a^{mn}]. \tag{2.5}$$

3. Jumps over zero minors

Given a semi-infinite Hankel matrix

$$A = [a_{i+j-1}], \quad 1 \le i, j < \infty,$$

let us consider its leading submatrices. Let A_k be a leading submatrix of order k, and let a sequence of natural numbers

$$n_1 < n_2 < \cdots$$

determine the orders of those and only those of them which are nonsingular. The method of jumps suggested in [10] is a scheme of the transition from some compact representation (suggested in [4] for Toeplitz matrices and then generalized to many classes of structured matrices, cf. [5]) for $A_{n_k}^{-1}$ to a similar-style representation for $A_{n_{k+1}}^{-1}$. It can be thought of as a jump over all intermediate singular leading submatrices, which explains the name.

General questions of the design of fast algorithms for Hankel and Toeplitz matrices are considered, for instance, in [8, 9]. Algebraic properties of Hankel matrices needed to perform the above-discussed jump are presented in [5] together with a structure of null-spaces of Hankel matrices. They are closely connected with infinite rank-preserving augmentations studied in [6].

The method of jumps stems from the following observation. Let $p = n_k$ and $q = n_{k+1}$. Since A_p is nonsingular, the system

$$\begin{bmatrix} a_1 & \dots & a_p \\ \dots & \dots & \dots \\ a_p & \dots & a_{2p-1} \end{bmatrix} \begin{bmatrix} s_1 \\ \dots \\ s_p \end{bmatrix} = \begin{bmatrix} a_{p+1} \\ \dots \\ a_{2p} \end{bmatrix}$$

has a unique solution. In other words, the columns from the 1st to pth truncated up to p elements are linearly independent, and the column $p+1$ truncated in the same way can be written as their linear combination with the coefficients $s_1, \dots, s_p$. It may happen that the same coefficients can be used for the column $p+1$ to be a linear combination of the preceding columns when truncated up to $p+1$ or even larger number of elements.

Theorem 3.1. *Let $r(p) \geq p$ be the minimal size for truncation, in which the column $p+1$ is not in the linear span of preceding columns. Then $n_{k+1} = r(n_k)$.*

Proof. Let $p = n_k$ and $r = r(p)$. Then

$$\begin{bmatrix} a_1 & \dots & a_p & a_{p+1} \\ \dots & \dots & \dots & \dots \\ a_p & \dots & a_{2p-1} & a_{2p} \\ \dots & \dots & \dots & \dots \\ a_{r-1} & \dots & a_{r+p-2} & a_{r+p-1} \\ a_r & \dots & a_{r+p-1} & a_{r+p} \end{bmatrix} \begin{bmatrix} -s_1 \\ \dots \\ -s_p \\ 1 \end{bmatrix} = \begin{bmatrix} 0 \\ \dots \\ 0 \\ \dots \\ 0 \\ \gamma \end{bmatrix}, \quad \gamma \neq 0.$$

This implies the equation

$$A_r \left[\begin{array}{cccc|cccc} -s_1 & & & & 1 & & & \\ \dots & -s_1 & & & & 1 & & \\ \dots & \dots & \dots & & & & \dots & \\ -s_p & \dots & \dots & & & & & 1 \\ \hline 1 & -s_p & \dots & & & & & \\ & 1 & \dots & \dots & & & & \\ & & \dots & \dots & & & & \\ & & \dots & -s_p & & & & \\ & & & 1 & & & & \end{array}\right] = \left[\begin{array}{cccc|cccc} & & & & a_1 & a_2 & \dots & a_p \\ & & & & a_2 & a_3 & \dots & a_{p+1} \\ & & & & \dots & \dots & \dots & \dots \\ & & & & a_p & a_{p+1} & \dots & a_{2p-1} \\ \hline & & & \gamma & a_{p+1} & a_{p+2} & \dots & a_{2p} \\ & & \dots & \dots & \dots & \dots & \dots & \dots \\ & \gamma & \dots & \dots & \dots & \dots & \dots & \dots \\ \gamma & \dots & \dots & \dots & a_r & a_{r+1} & \dots & a_{r+p-1} \end{array}\right]. \tag{3.1}$$

If $p = n_k$ then the matrix A_p is nonsingular. Therefore, its augmentation A_r is nonsingular if and only if $\gamma \neq 0$. □

Corollary 3.2. *If $n_k \leq n \leq n_{k+1}$ then*

$$\dim \ker A_n = \min\{n - n_k, n_{k+1} - n\}.$$

Proof. We derive directly from (3.1) that A_n times a nonsingular matrix is a matrix having $\min\{n - n_k, n_{k+1} - n\}$ zero columns and all nonzero columns producing a linearly independent system. □

Denote by $\hat{A}_n$ the augmentation of A_n of the following form:

$$\hat{A}_n = \begin{bmatrix} a_1 & \dots & a_n & a_{n+1} \\ \dots & \dots & \dots & \dots \\ a_n & \dots & a_{2n-1} & a_{2n} \end{bmatrix}.$$

Corollary 3.3. *If $n_k \le n < n_{k+1}$, then the equality of ranks*

$$\mathrm{rank}A_n = \mathrm{rank}\hat{A}_n \tag{3.2}$$

takes place when and only when

$$n - n_k < n_{k+1} - n. \tag{3.3}$$

Proof. According to the Kronecker–Capelli theorem, the equality (3.2) is equivalent to the claim for the last column of the extended matrix $\hat{A}_n$ to be a linear combination of the preceding columns. Let $p = n_k$ and $q = n_{k+1}$.

Let $p \le n = p + i < q$, and note that the column $p + 1$ of A_{q-1} is a linear combination of the preceding columns with the coefficients $s_1, \dots, s_p$. The same is valid for the column $p + 1$ of all those of its leading submatrices that contain A_{p+1}. If

$$p + 2i < q,$$

then this holds true for the matrix A_{p+2i}. Allowing for the Hankel structure of matrices, we easily deduce that the column $p + 1 + i$ – (the last column of the extended matrix $\hat{A}_{p+i}$) is in the linear span of the preceding p columns (with the same coefficients $s_1, \dots, s_p$). The inequality $p + 2i < q$ is equivalently to $n - p < q - n$.

It remains to prove that (3.2) implies (3.3). In accordance with (3.2) the last column of the extended matrix $\hat{A}_n$ is in the linear span of the columns of A_n. In this case, when passing from A_n to A_{n+1} the rank may increase by at most 1. By contradiction, let us admit that

$$n - p \ge q - p.$$

On the base of Corollary 3.2,

$$\mathrm{rank}A_n = n - \min\{n - p, q - n\} = 2n - q,$$

$$\mathrm{rank}A_{n+1} = n + 1 - \min\{n + 1 - p, q - n - 1\} = 2n - q + 2.$$

Hence,

$$\mathrm{rank}A_{n+1} = \mathrm{rank}A_n + 2,$$

which is impossible since the rank cannot inflate greater than by 1. □

Corollary 3.4. *If $n_k \le n < n_{k+1}$, then the inequality* (3.3) *is fulfilled when and only when*

$$\mathrm{rank}A_{n+1} - \mathrm{rank}A_n \le 1.$$

4. An identity for determinants

Let A be a matrix of order n and A_{ij} be its submatrix of order $n-1$ obtained by ruling out the row i and column j. Let $A_{ik;jl}$ denote a submatrix of order $n-2$ appeared from A be deleting a pair rows with the indices i and k and a pair of columns with the indices j and l. The next result is known as the Sylvester identity.

Theorem 4.1. *Let $i < k$ and $j < l$. Then*

$$\det A \, \det A_{ik;jl} = \det A_{ij} \, \det A_{kl} - \det A_{il} \, \det A_{kj}.$$

Proof. Without loss of generality we may assume that $i = j = n-1$ and $k = l = n$. Let $B = A_{ik;jl}$. Then the matrix A is of the form

$$A = \begin{bmatrix} B & v & q \\ u & c & d \\ p & g & h \end{bmatrix}.$$

Suppose first that B is nonsingular. By block Gaussian elimination of u and p with the pivot B we find

$$\begin{bmatrix} I & 0 & 0 \\ -uB^{-1} & 1 & 0 \\ -pB^{-1} & 0 & 1 \end{bmatrix} \begin{bmatrix} B & v & q \\ u & c & d \\ p & g & h \end{bmatrix} = \begin{bmatrix} B & v & q \\ 0 & c_1 & d_1 \\ 0 & g_1 & h_1 \end{bmatrix},$$

where

$$h_1 = h - pB^{-1}q, \quad c_1 = c - uB^{-1}v, \quad g_1 = g - pB^{-1}v, \quad d_1 = d - uB^{-1}q.$$

Consequently,

$$\det A_{ij} \, \det A_{kl} - \det A_{il} \, \det A_{kj} = (\det B)^2 (h_1 c_1 - g_1 d_1) = \det B \, \det A.$$

If B is singular then the target identity is valid as soon as we replace B with any nonsingular block B_ε. Since B_ε can be chosen arbitrarily close to B, we complete the proof by transition to the limit. □

5. Table of minors

Proceed with the study of an infinite Hankel matrix $A = [a_{i+j}]$ composed of coefficients of the formal series $f(x) = \sum_{i=0}^{\infty} a_i x^i$ (if $i < 0$ then $a_i = 0$). Recall that A_{mn} is its Hankel submatrix of order n with the entry a_m in the lower left corner. We are interested to examine a semi-infinite matrix $C = [c_{mn}]$ collecting the minors of A: $\quad c_{mn} = \det A_{mn}, \quad 0 \le m, n < \infty. \quad$ Set $c_{m0} = 1$.

Lemma 5.1. $c_{m,n+1} c_{m,n-1} = c_{m+1,n} c_{m-1,n} - c_{mn}^2$.

Proof. This equality is nothing else than the Sylvester identity for determinants (Theorem 4.1) applied to the Hankel matrix $A_{m,n+1}$ and its submatrices after expunging the first and last rows and columns, the Hankel property being kept with this special choice. □

The table of minors C for the Hankel matrix A is sometimes referred to as *C-table* [1]. The basic property of the C-table consists in a special structure of zeroes (zero minors of A). Let us call a *window* any finite or infinite submatrix made up from the entries of contiguous rows and columns. A window is called a *square window* if it corresponds to a finite square submatrix or an infinite one with infinitely many both rows and columns. All the entries adjacent to a given window will be called a *frame* – these entries belong to a wider window with the rows and columns augmenting the given window. Above all, we need to investigate *zero windows* and *nonzero frames* – in the first case all the entries in the window are zeroes, in the second case all the entries in the frame differ from zero.

Here and throughout, assume that $a_0 \neq 0$. Then $c_{0n} \neq 0$ for $n \geq 1$ (the determinants of Hankel triangular matrices with a nonzero along the anti-diagonal). Moreover, let us agree that $c_{m0} = 1$ whenever $m \geq 0$.

Theorem 5.2. *Any zero entry in the table of minors C belongs to a square window with a nonzero frame.*

Proof. Assume that $c_{m,n-1} = c_{m+1,n} = 0$ and $c_{m,n+1} = c_{m+1,n} = 0$. From Lemma 5.1 we find $c_{mn} = 0$. As a corollary, submatrices of the form

$$\begin{bmatrix} 0 & * \\ * & 0 \end{bmatrix}, \qquad \begin{bmatrix} * & 0 \\ 0 & * \end{bmatrix}$$

must be zero. Combining this with the inequalities $c_{m0} \neq 0$ and $c_{0n} \neq 0$ we conclude that any zero entry of C belongs to a rectangular window with a nonzero frame. It remains to prove that this window is in fact a square one.

Let $c_{mn} \neq 0$ is an entry of the frame located at the upper left corner. Assume that $c_{m+r,n+r} \neq 0$ is one more entry of the same window's frame and prove that it lies at the lower right corner. If it were not so, then we would have two options:

(1) $c_{m+r,n+r-1} = 0$, or
(2) $c_{m+r-1,n+r} = 0$.

Option (1). Note that A_{mn} and $A_{m+r,n+r}$ are nonsingular leading submatrices in the Hankel matrix $A_{m+r,n+r}$ and the intermediate leading submatrices $A_{m+i,n+i}$ for $0 < i < r$ are singular. According to the method of jumps (Theorem 3.1), the column $n+1$ of $A_{m+r,n+r}$ without the last row is a linear combination of the preceding columns.

In this case $c_{m+1,n} \neq 0$, $c_{m+r+1,n+r} \neq 0$. This validates application of the method of jumps to nonsingular Hankel matrices $A_{m+1,n}$ and $A_{m+1+r,n+r}$, and so we come to the conclusion that the column $n+2$ of $A_{m+r,n+r}$ without the last row is a linear combination of the preceding columns starting from the 2nd one. By subtraction of these linear combinations from the columns $n+1$ and $n+2$ we do not change the determinant of $A_{m+r,n+r}$ and acquire two columns with zeroes except for the last row's entries. Consequently, $c_{m+r,n+r} = \det A_{m+r,n+r} = 0$, which contradicts to the initial assumption. Thus, $c_{m+r,n+r-1} \neq 0$.

Option (2). Observe that $c_{m+r,n+r+1} \neq 0$ (otherwise $c_{m+r,n+r} = 0$ by Lemma 5.1). Thus, $A_{m,n+1}$ are $A_{m+r,n+r+1}$ are nonsingular leading submatrices with singular

intermediate submatrices. In line with the method of jumps, the column $n+2$ in $A_{m+r,n+r+1}$ without the last row is a linear combination of the preceding columns. Hence, the column $n+2$ of $A_{m+r,n+r}$ without the last row is a linear combination of the preceding columns. As previously, the same holds true for the column $n+1$ of the same matrix. Again we come to contradiction with the nonsingularity of $A_{m+r,n+r}$. Therefore, $c_{m+r-1,n+r} \neq 0$.

All in all, both options (1) and (2) lead to contradiction. It means that we have simultaneously

$$c_{m+r,n+r} \neq 0, \quad c_{m+r,n+r-1} \neq 0, \quad c_{m+r-1,n+r} \neq 0.$$

This proves that $c_{m+r,n+r}$ is located at the lower right corner of the zero window's frame. Since the upper left corner is occupied by c_{mn}, this is a square window. The case when $c_{m+r,n+r} \neq 0$ for all $r > 0$ is merely simpler: a recursive use of Lemma 5.1 makes it clear that we deal with an infinite square window. □

6. Pade theory

Pade theory provides us with the necessary and sufficient condition for the existence of a Pade approximation of type (m, n) in the terms of zero structure in the table of minors associated with the formal series $f(x)$. As we already know, a condition of this kind is the consistency of the linear system (2.4). Obviously, the existence is guaranteed whenever $c_{mn} = \det A_{mn} \neq 0$. The main result for the case $c_{mn} = 0$ is formulated as follows.

Theorem 6.1. *Let $c_{mn} = 0$ belong to a zero window with a nonzero frame whose upper left and lower right corners keep $c_{kl} \neq 0$ and $c_{k+r,l+r} \neq 0$, respectively. Then, Pade approximation of type (m, n) exists if and only if $k + l < m + n < k + l + r$.*

Proof. Assume that c_{st} and $c_{s+p,t+p}$ belong to the nonzero frame of the given zero window and for some $h > 0$ we have $m = s + h$ and $n = t + h$. The matrix A_{st} is a nonsingular leading submatrix in the nonsingular Hankel matrix $A_{s+p,t+p}$, and the submatrices $A_{s+i,t+i}$ are all singular for $0 < i < p$. From the method of jumps (see Corollary 3.3) it emanates that the compatibility condition (2.5) is equivalent to the inequality

$$(t + h) - t < (t + p) - (t + h) \quad \Leftrightarrow \quad h < p/2.$$

It is fulfilled if and only if

$$m + n < k + l + r. \qquad \square$$

Theorem 6.2. *Let c_{kl} and $c_{k+r,l+r}$ be the corner entries of a nonzero frame of some zero window of the table of minors, and assume that $m \geq k$ and $n \geq l$ satisfy the inequalities $k + l \leq m + n < k + l + r$. Then, a Pade approximation of type (k, l) is also a Pade approximation of type (m, n).*

If c_{kl} is the corner nonzero entry of the infinite zero window's frame, then the same is valid for all $m \geq k$ and $n \geq l$ provided that $k + l \leq m + n$.

Proof. Let $k \le m$ and $l \le n$. The claim reduces to the following check: if for v^l we have

$$A_{kl} v^l = -a^{kl},$$

then, so long as $k + l \le m + n < k + l + r$, we also obtain

$$A_{mn} \begin{bmatrix} 0 \\ v^l \end{bmatrix} = -a^{mn}.$$

In the case of infinite window, the latter holds true for all $r > 0$. □

References

[1] G.A. Baker, Jr. and P. Graves-Morris, *Pade Approximants*, Addison-Wesley Publishing Co., 1981.

[2] A. Bultheel, M. Van Barel, Linear Algebra, Rational Approximation and Orthogonal Polynomials, *Studies in Computational Mathematics*, vol. 6, North-Holland, Elsevier Science, Amsterdam, 1997.

[3] A. Bultheel, M. Van Barel, Pade techniques for model reduction in linear system theory: a survey, *J. Comput. Appl. Math.*, 14, pp. 401–438 (1986).

[4] I. Gohberg, A.A. Semencul, On inversion of finite-section Toeplitz matrices and their continuous analogues, *Matem. Issled.* (Russian), Vol.7, No. 2, pp. 201–224 (1972).

[5] G. Heinig, K. Rost, *Algebraic Methods for Toeplitz-like Matrices and Operators*, Academie-Verlag, Berlin, 1984.

[6] I.C. Iohvidov, *Hankel and Toeplitz Matrices and Forms*, Nauka, 1984.

[7] E.M. Nikishin and V.N. Sorokin, *Rational approximations and orthogonality*, Nauka, 1988.

[8] E.E. Tyrtyshnikov, *Toeplitz Matrices, Some Analogs and Applications*, OVM RAN, 1989.

[9] E.E. Tyrtyshnikov, Euclidean Algorithm and Hankel Matrices, *Numerical Analysis and Applied Mathematics*, AIP Conference Proceedings, vol. 936, Melville, New York, pp. 27–30 (2007).

[10] V.V. Voevodin and E.E. Tyrtyshnikov, *Computational Processes with Toeplitz Matrices*, Nauka, 1987.

Eugene Tyrtyshnikov
Institute of Numerical Mathematics
Russian Academy of Sciences
Gubkin Street, 8
Moscow 119333, Russia
e-mail: `tee@inm.ras.ru`