Operator Theory: Advances and Applications
Vol. 179

Recent Advances in Matrix and Operator Theory

Joseph A. Ball
Yuli Eidelman
J. William Helton
Vadim Olshevsky
James Rovnyak
Editors

Birkhäuser
Basel · Boston · Berlin

Editors:

Joseph A. Ball
Department of Mathematics
Virginia Tech
Blacksburg, VA 24061, USA
e-mail: ball@math.vt.edu

Yuli Eidelman
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

J. William Helton
Department of Mathematics
University of California San Diego
9500 Gilman Drive
La Jolla, California 92093-0112, USA
e-mail: helton@math.ucsd.edu

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

James Rovnyak
Department of Mathematics
University of Virginia
P. O. Box 400137
Charlottesville, VA 22904-4137, USA
e-mail: rovnyak@virginia.edu

2000 Mathematics Subject Classification: 15Axx, 28C20, 30A38, 30E05, 34A55, 34D20, 34D45, 46E22, 46J20, 46L40, 47Axx, 47B35, 47J10, 49N45, 60G20, 60H40, 65F05, 65F10, 65F15, 65H17, 70G30, 90C22

Library of Congress Control Number: 2007937322

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 978-3-7643-8538-5 Birkhäuser Verlag AG, Basel - Boston - Berlin

ISBN 978-3-7643-8538-5 e-ISBN 978-3-7643-8539-2

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

Contents

Editorial Introduction

This volume contains the proceedings of the International Workshop on Operator Theory and Applications (IWOTA) which was held at the University of Connecticut, Storrs, USA, July 24–27, 2005. This was the sixteenth IWOTA; in fact, the workshop was held biannually since 1981, and annually in recent years (starting in 2002) rotating among ten countries on three continents. Here is the list of the fifteen workshops:

IWOTA'1981: Santa Monica, California, USA (J.W. Helton, Chair)
IWOTA'1983: Rehovot, Israel (H. Dym, Chair)
IWOTA'1985: Amsterdam, The Netherlands (M.A. Kaashoek, Chair)
IWOTA'1987: Mesa, Arizona, USA (L. Rodman, Chair)
IWOTA'1989: Rotterdam, The Netherlands (H. Bart, Chair)
IWOTA'1991: Sapporo, Hokkaido, Japan (T. Ando, Chair)
IWOTA'1993: Vienna, Austria (H. Langer, Chair)
IWOTA'1995: Regensburg, Germany (R. Mennicken, Chair)
IWOTA'1996: Bloomington, Indiana, USA (H. Bercovici, C. Foias, Co-chairs)
IWOTA'1998: Groningen, The Netherlands (A. Dijksma, Chair)
IWOTA'2000: Faro, Portugal (A.F. dos Santos, Chair)
IWOTA'2002: Blacksburg, Virginia, USA (J. Ball, Chair)
IWOTA'2003: Cagliari, Italy (S. Seatzu, C. van der Mee, Co-Chairs)
IWOTA'2004: Newcastle upon Tyne, UK (M.A. Dritschel, Chair)
IWOTA'2005: Storrs, Connecticut, USA (V. Olshevsky, Chair)

The aim of the 2005 IWOTA was to review recent advances in operator theory and its applications to several areas including mathematical systems theory and control theory.

Among the main topics of the workshop was the study of structured matrices, their applications, and their role in the design of fast and numerically reliable algorithms. This topic had already received a considerable attention at IWOTA'2002 and IWOTA'2003 when the main focus was mostly on the structures of Toeplitz, Hankel and Pick types. In the year 2005 the interest shifted towards matrices with quasiseparable structure.

The IWOTA'2005 was made possible through the generous financial support of National Science Foundation (award : 0536873) as well as thanks to the funds of the College of Arts and Sciences and of the Research Foundation of the University of Connecticut. All this support is acknowledged with a gratitude.

Joseph Ball, Yuli Eidelman, William Helton,
Vadim Olshevsky, and James Rovnyak (Editors)

Operator Theory:
Advances and Applications, Vol. 179, 1–24

Inverse Problems for First-Order Discrete Systems

Daniel Alpay and Israel Gohberg

Abstract. We study inverse problems associated to first-order discrete systems in the rational case. We show in particular that every rational function strictly positive on the unit circle is the spectral function of such a system. Formulas for the coefficients of the system are given in terms of realizations of the spectral function or in terms of a realization of a spectral factor. The inverse problems associated to the scattering function and to the reflection coefficient function are also studied. An important role in the arguments is played by the state space method. We obtain formulas which are very similar to the formulas we have obtained earlier in the continuous case in our study of inverse problems associated to canonical differential expressions.

Mathematics Subject Classification (2000). Primary: 34A55, 49N45, 70G30; Secondary: 93B15, 47B35.

Keywords. Inverse problems, spectral function, scattering function, Schur parameters, state space method.

1. Introduction

Here we continue to study first-order discrete systems. We defined the characteristic spectral functions associated to a first-order discrete in [7] and studied the corresponding inverse problems in [8] for scalar systems. In the matrix-valued case, see [3], a system of equations of the form

$$X_n(z) = \begin{pmatrix} I_p & \alpha_n \\ \beta_n & I_p \end{pmatrix}^* \begin{pmatrix} zI_p & 0 \\ 0 & I_p \end{pmatrix} X_{n-1}(z), \quad n = 1, 2, \ldots, \tag{1.1}$$

is called a canonical discrete first-order one-sided system. The sequence of matrices (α_n, β_n) is not arbitrary, but has the following property: there exists a sequence Δ

Daniel Alpay wishes to thank the Earl Katz family for endowing the chair which supported his research.

of strictly positive block diagonal matrices in $\mathbb{C}^{2p\times 2p}$ such that

$$\begin{pmatrix} I_p & \alpha_n \\ \beta_n & I_p \end{pmatrix} J\Delta_n \begin{pmatrix} I_p & \alpha_n \\ \beta_n & I_p \end{pmatrix}^* = J\Delta_{n-1}, \quad n = 1, 2, \ldots, \tag{1.2}$$

where

$$J = \begin{pmatrix} I_p & 0 \\ 0 & -I_p \end{pmatrix}.$$

The sequence is then called Δ-admissible. In the scalar case (that is, when $p = 1$) condition (1.2) forces $\alpha_n = \beta_n^*$ (see [3]). Still for $p = 1$ these systems arise as the discretization of the telegrapher equation; see [7] for a discussion and references. An *a posteriori* motivation for the study of such systems is the fact that we obtain formulas very close to the ones we proved in the continuous case in our study of inverse problems associated to canonical differential expressions. To be more precise we need to present our setting in greater details. We first gather the main results from [3] needed in the sequel. Let

$$Z = \begin{pmatrix} zI_p & 0 \\ 0 & I_p \end{pmatrix} \quad \text{and} \quad F_n = \begin{pmatrix} 0 & \beta_n^* \\ \alpha_n^* & 0 \end{pmatrix}, \quad n = 1, 2, \ldots.$$

Under the hypothesis

$$\sum_{n=1}^{\infty} (\|\alpha_n\| + \|\beta_n\|) < \infty, \tag{1.3}$$

the infinite product

$$Y(z) = \left(\prod_{n=1}^{\overset{\curvearrowleft}{\infty}} (I_{2p} + Z^{-n} F_n Z^n) \right) \tag{1.4}$$

converges absolutely and uniformly on the unit circle, and the functions

$$X_n(z) = Z^n \left((I_{2p} + Z^{-n} F_n Z^n) \cdots (I_{2p} + Z^{-1} F_1 Z) \right) Y(z)^{-1}, \quad n = 1, 2, \ldots,$$

define the unique $\mathbb{C}^{2p\times 2p}$-valued solution to the system (1.1) with the property that

$$\lim_{n\to\infty} \begin{pmatrix} z^{-n} I_p & 0 \\ 0 & I_p \end{pmatrix} X_n(z) = I_{2p}, \quad |z| = 1. \tag{1.5}$$

See [3, Section 2.1]. This solution is called the *fundamental solution* of the first-order discrete system (1.1). The function $Y(z)^{-1}$ is called the *asymptotic equivalence matrix function*; see [3, Section 2.2]. Under the supplementary hypothesis

$$\lim_{n\longrightarrow\infty} \Delta_n > 0 \tag{1.6}$$

the function $Y(z)$ allows to define the characteristic spectral functions of the system (1.1). We note that when (1.6) is not in force the situation seems to be much more involved, and leads to degenerate cases. Furthermore, conditions such as (1.2) and (1.6) seem to be specific of the discrete case; no counterpart of these conditions is needed in the continuous case.

Let

$$\lim_{n\to\infty} \Delta_n = \begin{pmatrix} \delta_1 & 0 \\ 0 & \delta_2 \end{pmatrix}. \tag{1.7}$$

The function

$$\begin{aligned} W(z) &= ((Y_{21}+Y_{22})(1/z))^{-1}\,\delta_2^{-1}\,((Y_{21}+Y_{22})(1/z))^{-*} \\ &= ((Y_{11}+Y_{12})(1/z))^{-1}\,\delta_1^{-1}\,((Y_{11}+Y_{12})(1/z))^{-*} \end{aligned}$$

is called the *spectral function*. The Weyl function is the uniquely defined function $N(z)$ analytic in the closed unit disk such that $N(0) = iI_p$ and

$$W(z) = \operatorname{Im} N(z), \quad |z| = 1.$$

Associated to N is the reproducing kernel space of functions with reproducing kernel $\frac{N(z)-N(w)^*}{z-w^*}$ and denoted by $\mathcal{L}(N)$. The function $W(z)$ is the spectral function of the unitary operator U defined in $\mathcal{L}(N)$ by

$$(U-\alpha I)^{-1} f(z) = \frac{f(z)-f(\alpha)}{z-\alpha}, \quad |\alpha| \neq 1.$$

See [11].

From (1.5) follows that there exists a $\mathbb{C}^{2p\times p}$-valued solution $B_n(z)$ to (1.1) with the following properties:

(a) $\begin{pmatrix} I_p & -I_p \end{pmatrix} B_0(z) = 0$, and
(b) $\begin{pmatrix} 0 & I_p \end{pmatrix} B_n(z) = I_p + o(n), \quad |z| = 1.$

It then holds that

$$\begin{pmatrix} I_p & 0 \end{pmatrix} B_n(z) = z^n S(z) + o(n)$$

where

$$S(z) = (Y_{11}(z)+Y_{12}(z))(Y_{21}(z)+Y_{22}(z))^{-1}. \tag{1.8}$$

The function (1.8) is called the *scattering matrix function* associated to the discrete system. The scattering matrix function has the following properties: it is in the Wiener algebra $\mathcal{W}^{p\times p}$ (see the end of the section for the definition), admits a Wiener–Hopf factorization and is such that

$$S(z)^*\delta_1 S(z) = \delta_2, \quad |z| = 1. \tag{1.9}$$

See [3, Section 2.3]. The inverse scattering problem considered in this paper is defined as follows: given a function $S(z)$ which admits a Wiener–Hopf factorization and satisfies moreover the condition (1.9) for some matrices δ_1 and δ_2, is $S(z)$ the scattering function of a first-order discrete system?

Some preliminary notation and remarks are needed to define the *reflection coefficient function*. First, for

$$M = \begin{pmatrix} M_{11} & M_{12} \\ M_{21} & M_{22} \end{pmatrix} \in \mathbb{C}^{2p\times 2p} \quad \text{and} \quad X \in \mathbb{C}^{p\times p}$$

we define the linear fractional transformation $T_M(X)$:

$$T_M(X) = (M_{11}X + M_{12})(M_{21}X + M_{22})^{-1}.$$

Recall that the semi-group property

$$T_{\Theta_1\Theta_2}(X) = T_{\Theta_1}(T_{\Theta_2}(X))$$

holds when the three matrices $T_{\Theta_2}(X)$, $T_{\Theta_1}(T_{\Theta_2})(X)$ and $T_{\Theta_1\Theta_2}(X)$ are well defined. Next, it follows from (1.2) that the matrices

$$C_n = \Delta_n^{1/2} \begin{pmatrix} I_p & \alpha_n \\ \beta_n & I_p \end{pmatrix}^* \Delta_{n-1}^{-1/2} \tag{1.10}$$

are J-unitary: $C_n^* J C_n = J$. Moreover, for every $n \in \mathbb{N}$ the solution $\Psi_n(z)$ of the system

$$\Psi_n(z) = \Psi_{n-1}(z) C_n^* \begin{pmatrix} zI_p & 0 \\ 0 & I_p \end{pmatrix}, \quad n = 1, 2, \dots \quad \text{and} \quad \Psi_0(z) = I_{2p} \tag{1.11}$$

is a matrix-valued function whose entries are polynomials of degree at most n and which is J-inner:

$$J - \Theta(z) J \Theta(z)^* \begin{cases} \leq 0, & |z| < 1, \\ = 0, & |z| = 1. \end{cases} \tag{1.12}$$

The *reflection coefficient function* is defined to be

$$R(z) = \lim_{n \to \infty} T_{\Psi_n(z)}(0).$$

We proved in [3, Section 2.4] that $R(z)$ belongs to the Wiener algebra $\mathcal{W}_+^{p \times p}$ and takes strictly contractive values on the unit circle. We also proved in [3, Section 2.4] that

$$R(z) = \frac{1}{z} Y_{21}(\overline{z})^* (Y_{22}(\overline{z}))^{-*} = \frac{1}{z} (Y_{11}(1/z))^{-1} Y_{12}(1/z), \quad |z| = 1,$$

and that the reflection coefficient function and the Weyl function are related by the formula

$$N(z) = i(I_p - zR(z))(I_p + zR(z))^{-1}. \tag{1.13}$$

This paper presents the solution of the inverse spectral problem in the rational case. We also briefly discuss how to recover the system using the scattering function or the reflection coefficient function. In the paper [8], where we considered the scalar case, a key role was played by the description of the solutions of an underlying Nehari problem which are unitary and admit a Wiener–Hopf factorization. The point of view in the present paper is different. A key tool is a certain uniqueness result in the factorization of J-inner polynomial functions (see Theorem 2.4).

We would like to mention that the formulas we obtain in Theorems 4.2 and 4.3 (that is, when one is given a minimal realization of the spectral function or a minimal realization of a spectral factor, respectively) are very similar to the formulas which we obtained earlier in the continuous case, in our study of inverse problems associated to canonical differential expressions with rational spectral data; see in particular formulas (4.7) and (4.12), which are the counterparts of [6, (3.1) p. 9] and [6, Theorem 3.5 p. 9], respectively.

The paper consists of five sections besides the introduction and its outline is as follows. In the second section we review part of the theory of certain finite dimensional reproducing kernel Hilbert spaces (called $\mathcal{H}(\Theta)$ spaces) which will be needed in the sequel. The inverse spectral problem is studied in Section 3 and the inverse scattering problem in Section 4. In the fifth and last section we consider the inverse problem associated to the reflection coefficient function.

We note that another kind of discrete systems have been studied in [18].

We will denote by $\mathbb{D}$ the open unit disk and by $\mathbb{T}$ the unit circle. The Wiener algebra of Fourier series $\sum_\ell z^\ell w_\ell$ with absolutely summable coefficients:

$$\sum_\ell |w_\ell| < \infty$$

will be denoted by $\mathcal{W}$. By $\mathcal{W}_+$ (resp. $\mathcal{W}_-$) we denote the sub-algebra of elements of $\mathcal{W}$ for which $w_\ell = 0$ for $\ell < 0$ (resp. $\ell > 0$). We denote by $\mathcal{W}^{p\times p}$ (resp. $\mathcal{W}_+^{p\times p}$, resp. $\mathcal{W}_-^{p\times p}$) the algebra of matrices with entries in $\mathcal{W}$ (resp. in $\mathcal{W}_+$, resp. in $\mathcal{W}_-$).

Finally, we denote by $\mathbb{C}_J$ the space $\mathbb{C}^{2p}$ endowed with the indefinite inner product

$$\langle f, g\rangle_{\mathbb{C}_J} = g^* J f, \quad f, g \in \mathbb{C}^{2p}. \tag{1.14}$$

2. Reproducing kernel Hilbert spaces

First recall that a Hilbert space $\mathcal{H}$ of $\mathbb{C}^k$-valued functions defined on a set Ω is called a *reproducing kernel Hilbert space* if there is a $\mathbb{C}^{k\times k}$-valued function $K(z, w)$ defined on $\Omega \times \Omega$ and with the following properties:

(i) For every $w \in \Omega$ and every $c \in \mathbb{C}^k$ the function $z \mapsto K(z, w)c$ belongs to $\mathcal{H}$.
(ii) It holds that

$$\langle f(z), K(z, w)c\rangle_{\mathcal{H}} = c^* f(w).$$

The function $K(z, w)$ is called the reproducing kernel of the space; it is positive in the sense that for every $\ell \in \mathbb{N}^*$ and every $w_1, \ldots, w_\ell \in \Omega$ the block matrix with ij block entry $K(w_i, w_j)$ is non-negative. Conversely, to any positive function corresponds a uniquely defined reproducing kernel Hilbert space with reproducing kernel the given positive function; see [9], [19], [1].

Finite dimensional reproducing kernel spaces with reproducing kernel of the form

$$K_\Theta(z, w) = \frac{J - \Theta(z)J\Theta(w)^*}{1 - zw^*}$$

have been studied in [2] and [4]. They correspond to rational functions which are J-unitary on the unit circle (but they may have singularities on the unit circle). In this work, a special role is played by the class $P(J)$ of $\mathbb{C}^{2p\times 2p}$-valued polynomial functions Θ which are J-inner (see (1.12) for the definition).

For $\Theta \in P(J)$ the function $K_\Theta(z, w)$ defined above is positive (in the sense of reproducing kernels) in $\mathbb{C}$. We denote by $\mathcal{H}(\Theta)$ the associated reproducing kernel Hilbert space and gather in the next theorem the main features of these spaces

which will be used in the sequel. In the statement, deg Θ denotes the McMillan degree of Θ and $\mathbf{H}_{2,J}$ denotes the Kreĭn space of pairs of functions $\binom{f(z)}{g(z)}$ with f and g in the Hardy space $\mathbf{H}_2^p$ and indefinite inner product

$$\left[\begin{pmatrix} f(z) \\ g(z) \end{pmatrix}, \begin{pmatrix} f(z) \\ g(z) \end{pmatrix}\right]_{\mathbf{H}_{2,J}} = \left\langle \begin{pmatrix} f(z) \\ g(z) \end{pmatrix}, J \begin{pmatrix} f(z) \\ g(z) \end{pmatrix} \right\rangle_{\mathbf{H}_2^{2p}}.$$

Furthermore, R_0 denotes the backward shift operator

$$R_0 f(z) = \frac{f(z) - f(0)}{z}.$$

Theorem 2.1. *Let $\Theta \in P(J)$.*

(i) *We have that $R_0\mathcal{H}(\Theta) \subset \mathcal{H}(\Theta)$.*

(ii) $\dim \mathcal{H}(\Theta) = \deg \Theta$.

(iii) $\det \Theta(z) = c_\Theta z^{\deg \Theta}$ for some $c_\Theta \in \mathbb{T}$.

(iv) *The space $\mathcal{H}(\Theta)$ is spanned by the columns of the matrix functions*

$$R_0^\ell \Theta(z), \quad \ell = 1, 2, \ldots,$$

and in particular the elements of $\mathcal{H}(\Theta)$ are $\mathbb{C}^{2p}$-valued polynomials.

(v)

$$\mathcal{H}(\Theta) = \mathbf{H}_{2,J} \ominus \Theta \mathbf{H}_{2,J}. \tag{2.1}$$

(vi) *The product of any two elements in $P(J)$ is always minimal, and for Θ_1 and Θ_2 in $P(J)$ it holds that*

$$\mathcal{H}(\Theta_1\Theta_2) = \mathcal{H}(\Theta_1) \oplus \Theta_1 \mathcal{H}(\Theta_2).$$

Proof. For the proofs of items (i), (ii) and (iv) and further references and information we refer to the papers [2] and [4]. These papers deal with the more general case of rational functions J-unitary on the unit circle (or the real line). To prove (iii) we note (see [2]) that Θ is a minimal product of degree one factors in $P(J)$ and that each one of these elementary factors has determinant equal to z. To prove (2.1) one checks that the space $\mathbf{H}_{2,J} \ominus \Theta \mathbf{H}_{2,J}$ has reproducing kernel $K_\Theta(z, w)$. By uniqueness of the reproducing kernel we have the desired equality. Since (property (iii))

$$\begin{aligned} \det \Theta_1\Theta_2(z) &= c_{\Theta_1\Theta_2} z^{\deg \Theta_1\Theta_2} \\ &= (\det \Theta_1)(\det \Theta_2) \\ &= c_{\Theta_1} z^{\deg \Theta_1} c_{\Theta_2} z^{\deg \Theta_2} \\ &= c_{\Theta_1} c_{\Theta_2} z^{\deg \Theta_1 + \deg \Theta_2} \end{aligned}$$

we have that

$$\deg \Theta_1\Theta_2 = \deg \Theta_1 + \deg \Theta_2.$$

Thus the product $\Theta_1\Theta_2$ is minimal. Finally from the equality

$$K_{\Theta_1\Theta_2}(z, w) = K_{\Theta_1}(z, w) + \Theta_1(z) K_{\Theta_2}(z, w) \Theta_1(w)^*$$

we see that

$$\mathcal{H}(\Theta_1\Theta_2) = \mathcal{H}(\Theta_1) + \Theta_1\mathcal{H}(\Theta_2).$$

The sum is direct and orthogonal since the product $\Theta_1\Theta_2$ is minimal, and this proves (vi). □

In the next theorem we precise the structure of $\mathcal{H}(\Theta)$ spaces.

Theorem 2.2. *Let $\Theta \in P(J)$. The space $\mathcal{H}(\Theta)$ has a basis which consists of $k \leq p$ chains of the form*

$$\begin{aligned} f_1(z) &= u_1, \\ f_2(z) &= zu_1 + u_2, \\ &\vdots \\ f_m(z) &= z^m u_1 + z^{m-1}u_2 + \cdots + u_m, \end{aligned} \tag{2.2}$$

where $u_1, \ldots, u_m \in \mathbb{C}^{2p}$.

Proof. The elements of $\mathcal{H}(\Theta)$ are polynomials (see (iv) of Theorem 2.1) and therefore the only eigenvalue of R_0 is 0, and the corresponding eigenvectors are vectors in $\mathbb{C}^{2p}$. Let $f_1, \ldots, f_k$ be the linear independent elements of $\mathbb{C}^{2p}$ in $\mathcal{H}(\Theta)$. The space spanned by the f_j is a strictly positive subspace of $\mathbf{H}_{2,J}$. On constant vectors the inner product of $\mathbf{H}_{2,J}$ coincides with the inner product of $\mathbb{C}_J$ (see Definition (1.14)) and so $k \leq p$. To conclude we note that each Jordan chain corresponding to an eigenvector is of the form (2.2). □

In general we can only state that $m \leq \deg \Theta$. Here we are in a more special situation. The $\Psi_n(z)$ defined by (1.11) have moreover the following property, which is important here: $\deg \Psi_n = np$ *and the entries of $\Psi_n(z)$ are scalar polynomials of degree less or equal to n*. Therefore, by Theorem 2.1 the components of the elements of $\mathcal{H}(\Psi_n)$ are polynomials of degree less or equal to $n-1$ and the following theorem shows that the space $\mathcal{H}(\Psi_n)$ is spanned by p chains of length n.

Theorem 2.3. *There exist matrices $S_0, S_1, \ldots, S_{n-1}$ such that a basis of $\mathcal{H}(\Psi_n)$ is given by the columns of $F_0(z), \ldots, F_{n-1}(z)$ where*

$$\begin{aligned} F_0(z) &= \begin{pmatrix} I_p \\ S_0^* \end{pmatrix}, \\ F_1(z) &= z\begin{pmatrix} I_p \\ S_0^* \end{pmatrix} + \begin{pmatrix} 0 \\ S_1^* \end{pmatrix}, \\ &\vdots \\ F_{n-1}(z) &= z^{n-1}\begin{pmatrix} I_p \\ S_0^* \end{pmatrix} + z^{n-2}\begin{pmatrix} 0 \\ S_1^* \end{pmatrix} + \cdots + \begin{pmatrix} 0 \\ S_{n-1}^* \end{pmatrix}. \end{aligned} \tag{2.3}$$

Proof. By Theorem 2.2, a basis of $\mathcal{H}(\Psi_n)$ is made of $k \leq p$ chains of the form (2.2). Since the components of the elements of $\mathcal{H}(\Psi_n)$ are polynomials of degree less or equal to $n-1$, these chains generate a space of dimension less or equal to kn. On the other hand,

$$\deg \Psi_n = np = \dim \mathcal{H}(\Psi_n).$$

Therefore, $k = p$ and each chain has length n. The space $\mathcal{H}(\Psi_n)$ contains therefore p linearly independent vectors $f_1, f_2, \ldots, f_p \in \mathbb{C}^{2p}$. Set

$$\begin{pmatrix} f_1 & f_2 & \cdots & f_p \end{pmatrix} \stackrel{\text{def.}}{=} \begin{pmatrix} X_1 \\ X_2 \end{pmatrix}$$

where X_1 and X_2 are in $\mathbb{C}^{p\times p}$. Since the f_j span a strictly positive subspace of $\mathbf{H}_{2,J}$ we have $X_1^* X_1 > X_2^* X_2$. Thus X_1 is invertible, and we can chose:

$$F_0(z) = \begin{pmatrix} I_p \\ X_2 X_1^{-1} \end{pmatrix} \in \mathcal{H}(\Psi_n).$$

We set $S_0^* = X_2 X_1^{-1}$. The next p elements in a basis of $\mathcal{H}(\Theta)$ form the columns of a matrix-function of the form

$$zF_0(z) + V = z \begin{pmatrix} I_p \\ S_0^* \end{pmatrix} + V, \quad V \in \mathbb{C}^{2p\times p}.$$

By subtracting a multiple of $F_0(z)$ to this function we obtain $F_1(z)$. The rest of the argument is proved by induction in the same way: if we know at rank ℓ that $F_\ell(z)$ is of the asserted form, then the next p elements in a basis of $\mathcal{H}(\Theta)$ form a matrix-function of the form $zF_\ell(z) + V$. Removing a multiple of $F_0(z)$ from this function we obtain $F_{\ell+1}(z)$. □

The following uniqueness theorem will be used in the solution of the inverse spectral problem; see the proof of Theorem 4.1:

Theorem 2.4. *Let (α_n, β_n) and (α'_n, β'_n) be two admissible sequences with associated sequences of diagonal matrices Δ_n and Δ'_n respectively, normalized by $\Delta_0 = \Delta'_0 = I_{2p}$. Let C_n be given by* (1.10) *and let C'_n be defined in a similar way, with (α'_n, β'_n) and Δ'_n. Assume that*

$$C_1^* \begin{pmatrix} zI_p & 0 \\ 0 & I_p \end{pmatrix} \cdots C_m^* \begin{pmatrix} zI_p & 0 \\ 0 & I_p \end{pmatrix} U = (C'_1)^* \begin{pmatrix} zI_p & 0 \\ 0 & I_p \end{pmatrix} \cdots (C'_m)^* \begin{pmatrix} zI_p & 0 \\ 0 & I_p \end{pmatrix} U' \stackrel{\text{def.}}{=} \Theta(z),$$

where U and U' are J-unitary constants. Then $U = U'$ and $C_\ell = C'_\ell$ for $\ell = 1, \ldots m$.

Proof. We denote by the superscript $'$ all the quantities related to the C'_n and we set $\Delta_n = \text{diag}\,(d_{1,n}, d_{2,n})$. Equation (1.2) can be rewritten as:

$$\begin{aligned} d_{1,n} - \alpha_n d_{2,n} \alpha_n^* &= d_{1,n-1}, && (2.4)\\ d_{1,n}\beta_n^* &= \alpha_n d_{2,n} && (2.5)\\ d_{2,n} - \beta_n d_{1,n} \beta_n^* &= d_{2,n-1}. && (2.6) \end{aligned}$$

We set

$$\theta_n(z) = C_n^* \begin{pmatrix} zI_p & 0 \\ 0 & I_p \end{pmatrix}, \tag{2.7}$$

so that $\Theta(z) = \theta_1(z) \cdots \theta_m(z)$.

By Theorem 2.1 (item (vi)) we have:

$$\begin{aligned} \mathcal{H}(\Theta) &= \mathcal{H}(\theta_1) \oplus \theta_1 \mathcal{H}(\theta_2) \oplus \theta_1\theta_2 \mathcal{H}(\theta_3) \oplus \cdots \\ &= \mathcal{H}(\theta'_1) \oplus \theta'_1 \mathcal{H}(\theta'_2) \oplus \theta'_1\theta'_2 \mathcal{H}(\theta'_3) \oplus \cdots . \end{aligned}$$

By Theorem 2.3, the constant functions of $\mathcal{H}(\Theta)$ span both the spaces $\mathcal{H}(\theta_1)$ and $\mathcal{H}(\theta'_1)$. Thus,

$$\mathcal{H}(\theta_1) = \mathcal{H}(\theta'_1).$$

These two spaces have the same reproducing kernel and we get

$$K_{\theta_1}(z,w) = K_{\theta'_1}(z,w).$$

Since

$$\begin{aligned} K_{\theta_1}(z,w) &= \frac{J - C_1^* \begin{pmatrix} zw^* I_p & 0 \\ 0 & -I_p \end{pmatrix} C_1}{1 - zw^*} \\ &= \frac{J - C_1^* \begin{pmatrix} (zw^* - 1 + 1) I_p & 0 \\ 0 & -I_p \end{pmatrix} C_1}{1 - zw^*} \\ &= \frac{J - C_1^* J C_1}{1 - zw^*} + C_1^* \begin{pmatrix} I_p & 0 \\ 0 & 0 \end{pmatrix} C_1 \\ &= \begin{pmatrix} I_p \\ \beta_1 \end{pmatrix} d_{1,1} \begin{pmatrix} I_p & \beta_1^* \end{pmatrix}, \end{aligned} \tag{2.8}$$

we get

$$\begin{pmatrix} I_p \\ \beta_1 \end{pmatrix} d_{1,1} \begin{pmatrix} I_p & \beta_1^* \end{pmatrix} = \begin{pmatrix} I_p \\ \beta'_1 \end{pmatrix} d'_{1,1} \begin{pmatrix} I_p & (\beta'_1)^* \end{pmatrix}.$$

It follows that $d_{1,1} = d'_{1,1}$ and $\beta_1 = \beta'_1$. From the normalization $\Delta_0 = \Delta'_0 = I_{2p}$ and equations (2.4)–(2.6) it follows that $d_{2,1} = d'_{2,1}$ and $\alpha_1 = \alpha'_1$.

By induction we see that

$$\mathcal{H}(\theta_n) = \mathcal{H}(\theta'_n), \quad n = 2, 3, \ldots.$$

But, in a way similar to (2.8),

$$K_{\theta_n}(z,w)=\frac{J-C_n^*\begin{pmatrix} zw^*I_p & 0\\ 0 & -I_p\end{pmatrix}C_n}{1-zw^*}=\Delta_{n-1}^{-1/2}\begin{pmatrix} I_p\\ \beta_n\end{pmatrix}d_{1,n}\begin{pmatrix} I_p & \beta_n^*\end{pmatrix}\Delta_{n-1}^{-1/2},$$

and it follows from $\Delta_{n-1}=\Delta'_{n-1}$ (induction hypothesis at rank $n-1$) that $\beta_n=\beta'_n$ and $d_{n,1}=d'_{n,1}$. Equations (2.4)–(2.6) imply then that $\alpha_n=\alpha'_n$ and $d_{n,2}=d'_{n,2}$, and finally that $U=U'$. □

Theorem 2.5. *Let $X(z)$ be analytic and contractive in the open unit disk and let $R(z)=\lim_{n\to\infty}T_{\Psi_n(z)}(X(z))$. Let $R(z)=R_0+R_1z+\cdots$ be the Taylor expansion of $R(z)$ at the origin. Then, the space $\mathcal{H}(\Psi_n)$ is spanned by the functions* (2.3) *with the coefficients $R_0,R_1,\dots,R_{n-1}$.*

Proof. Let $A_0,A_1,\dots$ be matrices such that $\mathcal{H}(\Psi_n)$ is spanned by the columns of the functions

$$\begin{aligned}F_0(z)&=\begin{pmatrix} I_p\\ A_0^*\end{pmatrix},\\ F_1(z)&=z\begin{pmatrix} I_p\\ A_0^*\end{pmatrix}+\begin{pmatrix} 0\\ A_1^*\end{pmatrix},\\ &\vdots\\ F_{n-1}(z)&=z^{n-1}\begin{pmatrix} I_p\\ A_0^*\end{pmatrix}+z^{n-2}\begin{pmatrix} 0\\ A_1^*\end{pmatrix}+\cdots+\begin{pmatrix} 0\\ A_{n-1}^*\end{pmatrix}.\end{aligned}$$

Since $\mathcal{H}(\Psi_n)=\mathbf{H}_{2,J}\ominus\Psi_n\mathbf{H}_{2,J}$ (see Theorem 2.1) we have that

$$\begin{aligned}\begin{pmatrix} I_p & -A_0\end{pmatrix}\Psi_n(0)&=0\\ \begin{pmatrix} I_p & -A_0\end{pmatrix}\Psi'_n(0)+\begin{pmatrix} 0 & -A_1\end{pmatrix}\Psi_n(0)&=0\\ &\vdots\end{aligned}\tag{2.9}$$

The first equation leads to $T_{\Psi_n(z)}(0)=A_0$. Letting $n\to\infty$ we have

$$A_0=R(0)=R_0.$$

The second equation will lead in a similar way to $R'(0)=A_1$. More generally, equations (2.9) lead to

$$\begin{pmatrix} I_p & -(A_0+A_1+\cdots+A_{n-1}z^{n-1})\end{pmatrix}\Psi_n(z)=O(z^n).\tag{2.10}$$

Set

$$\Psi_n(z)=\begin{pmatrix}\alpha_n(z) & \beta_n(z)\\ \gamma_n(z) & \delta_n(z)\end{pmatrix}.$$

Equation (2.10) implies that

$$\beta_n(z)-(A_0+A_1+\cdots+A_{n-1}z^{n-1})\delta_n(z)=O(z^n).$$

From the J-innerness of $\Psi_n(z)$ the matrix-function $\delta_n(z)$ is analytic and invertible in $\mathbb{D}$, with $\|\delta_n(z)^{-1}\| \leq 1$; see [13]. Hence,

$$T_{\Psi_n(z)}(0) = (A_0 + A_1 + \cdots + A_{n-1}z^{n-1}) + O(z^n)$$

and hence the result. □

3. Realization theory

As is well known a rational function $W(z)$ analytic at the origin can be written in the form

$$W(z) = D + zC(I - zA)^{-1}B$$

where $D = W(0)$ and where A, B and C are matrices of appropriate sizes. The realization is called minimal when the size of A is minimal; see [10]. Assume moreover that $W(z)$ is analytic on the unit circle. Then A has no spectrum on the unit circle and the entries of $W(z)$ are in the Wiener algebra $\mathcal{W}$; indeed, let P_0 denote the Riesz projection corresponding to the spectrum of A outside the closed unit disk:

$$P_0 = I - \frac{1}{2\pi i}\int_{\mathbb{T}} (\zeta I - A)^{-1} d\zeta.$$

Then,

$$\begin{aligned} W(z) &= D + zC(I - zA)^{-1}B \\ &= D + zCP_0(I - zA)^{-1}P_0B + zC(I - P_0)(I - zA)^{-1}(I - P_0)B \\ &= D - zCP_0A^{-1}z^{-1}(I - z^{-1}A^{-1})^{-1}P_0B \\ &\quad + zC(I - P_0)(I - zA)^{-1}(I - P_0)B \\ &= D - \sum_{k=0}^{\infty} z^{-k}CP_0A^{-k-1}P_0B \\ &\quad + \sum_{k=0}^{\infty} z^{k+1}C(I - P_0)A^k(I - P_0)B. \end{aligned}$$

and thus the coefficients r_k in the representation $W(z) = \sum_{\mathbb{Z}} z^k r_k$ (with $|z| = 1$) can be written as

$$r_k = \begin{cases} CA^{k-1}(I - P_0)B, & k > 0, \\ D\delta_{k0} - CA^{k-1}P_0B, & k \leq 0, \end{cases} \tag{3.1}$$

so that

$$\sum_{\mathbb{Z}} \|r_k\| < \infty.$$

The hypotheses of analyticity at the origin and at infinity are restrictive. In fact any rational function analytic on the unit circle belongs to the Wiener algebra.

We now review the relevant theory and follow the analysis in [14]. First recall that any rational function $W(z)$ analytic on the unit circle can be represented as

$$W(z) = I + C(zG - A)^{-1}B,$$

where $zG - A$ is invertible on $\mathbb{T}$; see [14, Theorem 3.1 p. 395]. The separating projection is defined by

$$P = \frac{1}{2\pi i}\int_{\mathbb{T}} G(\zeta G - A)^{-1}d\zeta. \tag{3.2}$$

Next the right equivalence operator E and the associated operator Ω are defined by

$$E = \frac{1}{2\pi i}\int_{\mathbb{T}}(1-\zeta^{-1})(\zeta G - A)^{-1}d\zeta \quad \text{and} \quad \Omega = \frac{1}{2\pi i}\int_{\mathbb{T}}(\zeta - \zeta^{-1})(\zeta G - A)^{-1}. \tag{3.3}$$

See [14, Equations (2.2)–(2.4) p. 389]. Then, (see [14, p. 398])

$$r_k = \begin{cases} -CE\Omega^k(I-P)B, & k = 1, 2, \dots, \\ I - CE(I-P)B, & k = 0, \\ CE\Omega^{-k-1}PB, & k = -1, -2, \dots . \end{cases}$$

The block entries of T_n^{-1} are now given as follows. Let $A^\times = A - BC$ and define $P^\times, E^\times$ and $\Omega^\times$ in a way analog to P, E and Ω, that is:

$$P^\times = \frac{1}{2\pi i}\int_{\mathbb{T}} G(\zeta G - A^\times)^{-1}d\zeta, \tag{3.4}$$

$$E^\times = \frac{1}{2\pi i}\int_{\mathbb{T}}(1-\zeta^{-1})(\zeta G - A^\times)^{-1}d\zeta, \tag{3.5}$$

and

$$\Omega^\times = \frac{1}{2\pi i}\int_{\mathbb{T}}(\zeta - \zeta^{-1})(\zeta G - A^\times)^{-1}. \tag{3.6}$$

Define moreover

$$Q = \frac{1}{2\pi i}\int_{\mathbb{T}}(\zeta G - A)^{-1}d\zeta, \tag{3.7}$$

$$\begin{aligned} V_n = {} & (I-Q)E^\times(I-P^\times) \\ & + (I-Q)E^\times(\Omega^\times)^{n+1}P^\times + QE^\times(\Omega^\times)^{n+1}(I-P^\times) + QE^\times P^\times, \end{aligned} \tag{3.8}$$

and

$$r_k^\times = \begin{cases} CE^\times(\Omega^\times)^k(I-P^\times)B, & k = 1, 2, \dots, n, \\ I + CE^\times(I-P^\times)B, & k = 0, \\ -CE^\times(\Omega^\times)^{-k}P^\times B, & k = -1, \dots, -n, \end{cases}$$

and

$$\begin{aligned} k_{kj}^{(n)} = {} & CE^\times(\Omega^\times)^{k+1}(I-P^\times)V_n^{-1}(I-Q)E^\times(\Omega^\times)^j P^\times B \\ & - CE^\times(\Omega^\times)^{n-k}P^\times V_n^{-1}QE^\times(\Omega^\times)^{n-j}(I-P^\times)B. \end{aligned}$$

Then, $T_n^{-1} = \left(\gamma_{kj}^{(n)}\right)_{k,j=1,\dots,n}$ with

$$\gamma_{kj}^{(n)} = r_{k-j}^{\times} + k_{kj}^{(n)}. \tag{3.9}$$

See [14, Theorem 8.2 p. 422].

4. Inverse spectral problem

We focus on the rational case and consider three cases:

1. The weight function is general: it is rational and strictly positive on $\mathbb{T}$.
2. We assume that the weight function is analytic at the origin and at infinity. Then we get concrete formulas.
3. We start from a spectral factor.

The uniqueness theorem (Theorem 2.4) is used in the proof of the following theorem.

Theorem 4.1. *Let $W(z)$ be a rational function without poles on the unit circle and which takes strictly positive values there, and which is normalized by*

$$\frac{1}{2\pi}\int_0^{2\pi} W(e^{it})dt = I_p. \tag{4.1}$$

Then, $W(z)$ is the spectral function of a uniquely determined first-order discrete system normalized by $\Delta_0 = I_{2p}$. The associated first-order discrete system is computed as follows: let

$$W(z) = I + C(zG - A)^{-1}B$$

be a realization of $W(z)$ which is regular on $\mathbb{T}$. Then,

$$\begin{aligned}
\alpha_n = CE^{\times}\big\{&(\Omega^{\times})^n(I-P^{\times}) + (\Omega^{\times})^{n+1}(I-P^{\times})V_n^{-1}(I-Q)E^{\times}P^{\times}\\
&-P^{\times}V_n^{-1}QE^{\times}(\Omega^{\times})^n(I-P^{\times})\big\}B\\
\times\big\{&I + CE^{\times}(I-P^{\times})B + CE^{\times}\Omega^{\times}(I-P^{\times})V_n^{-1}E^{\times}P^{\times}P^{\times}B\\
&-CE^{\times}(\Omega^{\times})^nP^{\times}V_n^{-1}QE^{\times}(\Omega^{\times})^n(I-P^{\times})B\big\}^{-1},\\
\beta_n = CE^{\times}\Big\{&(\Omega^{\times})^{(n-1)}P^{\times} + \Omega^{\times}(I-P^{\times})V_n^{-1}(I-Q)(\Omega^{\times})^nP^{\times}\\
&-P^{\times}V_n^{-1}QE^{\times}(I-P^{\times})\big\}B\\
\times\big\{&I + CE^{\times}(I-P^{\times})B\\
&+ CE^{\times}(\Omega^{\times})^{(n+1)}(I-P^{\times})V_n^{-1}(I-Q)(\Omega^{\times})^nP^{\times}B\\
&-CE^{\times}P^{\times}V_n^{-1}QE^{\times}(I-P^{\times})B\big\}^{-1},
\end{aligned} \tag{4.2}$$

with associated sequence of diagonal matrices given by

$$\Delta_n = \begin{pmatrix} d_{1,n} & 0\\ 0 & d_{2,n}\end{pmatrix} \tag{4.3}$$

where

$$d_{1,n} = I + CE^{\times}(I - P^{\times})B + CE^{\times}(\Omega^{\times})^{(n+1)}(I - P^{\times})V_n^{-1}(I - Q)(\Omega^{\times})^n P^{\times} B$$
$$- CE^{\times} P^{\times} V_n^{-1} Q E^{\times}(I - P^{\times})B,$$
$$d_{2,n} = I + CE^{\times}(I - P^{\times})B + CE^{\times}\Omega^{\times}(I - P^{\times})V_n^{-1}E^{\times}P^{\times}P^{\times}B$$
$$- CE^{\times}(\Omega^{\times})^n P^{\times} V_n^{-1} Q E^{\times}(\Omega^{\times})^n (I - P^{\times})B$$

for $n = 1, 2, \ldots$. *In these expressions, the quantities* P, E, Ω *and* Q *are given by* (3.2), (3.3) *and* (3.7) *respectively and* $P^{\times}, E^{\times}, Q^{\times}$ *and* V_n *are given by* (3.4), (3.5), (3.6) *and* (3.8) *respectively.*

Proof. We first prove the uniqueness of the associated first-order discrete system. Fix $n > 0$. For every $q > 0$ we have (recall that Ψ_n is defined by (1.11) and θ_n by (2.7))

$$\Psi_{n+q}(z) = \Psi_n(z)\theta_{n+1}(z)\cdots\theta_{n+q}(z),$$

and in particular

$$R(z) = \lim_{q\to\infty} T_{\Psi_n(z)}(T_{\theta_{n+1}(z)\cdots\theta_{n+q}(z)}(0)).$$

By Montel's theorem, the limit

$$R_n(z) = \lim_{q\to\infty} T_{\theta_{n+1}(z)\cdots\theta_{n+q}(z)}(0)$$

exists (via maybe a subsequence). The limit is analytic and contractive in the open unit disk. Thus

$$R(z) = T_{\Psi_n(z)}(R_n(z)).$$

By Theorem 2.5, the space $\mathcal{H}(\Psi_n)$ is built from the first n coefficients of the Taylor expansion of $R(z)$ at the origin.

Assume that there are two first-order discrete systems (normalized by $\Delta_0 = I_{2p}$) and with same spectral function $W(z)$. By formula (1.13) these two systems have the same reflection coefficient function $R(z)$. Denoting by a superscript $'$ the second one, we get $\mathcal{H}(\Psi_n) = \mathcal{H}(\Psi_n')$ for every $n \geq 0$. By Theorem 2.4 it follows that the two systems are equal.

We now turn to the existence of such a system. The function $W(z)$ is rational and has no poles on the unit circle. It belongs therefore to the Wiener algebra $\mathcal{W}^{p\times p}$. We set $W(e^{it}) = \sum_{\mathbb{Z}} r_j e^{ijt}$ (note that $r_0 = I_p$ in view of the normalization (4.1)). The block matrices T_n are strictly positive and it follows from [12] that the pair

$$\alpha_n = \gamma_{n0}^{(n)}(\gamma_{00}^{(n)})^{-1} \quad \text{and} \quad \beta_n = \gamma_{0n}^{(n)}(\gamma_{nn}^{(n)})^{-1}, \quad n = 1, 2, 3\ldots, \tag{4.4}$$

form an admissible sequence, with associated sequence of diagonal matrices given by

$$\Delta_n = \begin{pmatrix} \gamma_{nn}^{(n)} & 0 \\ 0 & \gamma_{00}^{(n)} \end{pmatrix}, \quad n = 0, 1, 2, \ldots. \tag{4.5}$$

The normalization (4.1) implies that $\Delta_0 = I_{2p}$. We now proceed in a number of steps:

STEP 1: *The limits* $\lim_{n\to\infty}\gamma_{00}^{(n)}$ *and* $\lim_{n\to\infty}\gamma_{nn}^{(n)}$ *exist and are strictly positive.*

Set $\Delta_n = \operatorname{diag}(d_{1,n}, d_{2,n})$. We have $\gamma_{00}^{(n)} = d_{1,n}$ and $\gamma_{nn}^{(n)} = d_{2,n}$. Formula (3.9) implies that the limits exist. Formulas (2.4)–(2.6) imply that $\gamma_{00}^{(n)}$ and $\gamma_{nn}^{(n)}$ are non-decreasing sequences of positive matrices, and so their limits are invertible since $\Delta_0 > 0$.

Alternatively, one can prove STEP 1 as follows: That the first limit exits follows from the projection method (see [17]). The invertibility of $\lim_{n\to\infty}\gamma_{00}^{(n)}$ is proved in [16, p. 123]. The second limit is reduced to the first one by considering $W(1/z)$. See the end of the proof of Theorem 1.8 in [3] for more information.

Thus (1.6) is in force. From (1.7) we have

$$\lim_{n\to\infty}\gamma_{00}^{(n)} = \delta_1 \quad \text{and} \lim_{n\to\infty}\gamma_{nn}^{(n)} = \delta_2.$$

STEP 2: *Condition* (1.3) *is in force.*

This follows from the explicit formulas (3.9) for $\gamma_{0n}^{(n)}$ and $\gamma_{nn}^{(n)}$.

As proved in [3] it follows from STEP 2 that the first-order discrete system (1.1) has a unique solution $X_n(z)$ such that (1.5) holds:

$$\lim_{n\to\infty}\begin{pmatrix} z^{-n}I_p & 0 \\ o & I_p \end{pmatrix} X_n(z) = \begin{pmatrix} I_p & 0 \\ 0 & I_p \end{pmatrix}, \quad |z| = 1.$$

We set (see [12, p. 80])

$$A_n(z) = \sum_{\ell=0}^{n} z^\ell \gamma_{\ell 0}^{(n)}, \qquad C_n(z) = \sum_{\ell=0}^{n} z^\ell \gamma_{\ell n}^{(n)},$$
$$A_n^\circ(z) = 2I_p - \sum_{\ell=0}^{n} p_\ell(z)\gamma_{\ell 0}^{(n)}, \quad C_n^\circ(z) = \sum_{\ell=0}^{n} p_\ell(z)\gamma_{\ell n}^{(n)},$$

where $p_\ell(z) = z^\ell r_0 + 2\sum_{s=1}^{\ell} z^{\ell-s} r_s^*$.

STEP 3: *It holds that*

$$\begin{aligned} \lim_{n\to\infty} A_n(\overline{z})^* &= \delta_2(Y_{21}(z) + Y_{22}(z)), \\ \lim_{n\to\infty} z^{-n} C_n(\overline{z})^* &= \delta_1(Y_{11}(z) + Y_{12}(z)), \quad |z| = 1. \end{aligned} \tag{4.6}$$

Indeed, set

$$\Theta_n(z) = \begin{pmatrix} zC_n(z) & A_n(z) \\ zC_n^\circ(z) & -A_n^\circ(z) \end{pmatrix}.$$

We have (see [12, Theorem 13.2 p. 127])

$$\Theta_n(z)\Delta_n^{-1} = \Theta_{n-1}(z)\Delta_{n-1}^{-1}\begin{pmatrix} I_p & \alpha_n \\ \beta_n & I_p \end{pmatrix}\begin{pmatrix} zI_p & 0 \\ 0 & I_p \end{pmatrix}, \quad n = 1, 2, \ldots.$$

It follows that the matrix-functions

$$X_n(z) = \Delta_n^{-1} \begin{pmatrix} z^{-1}I_p & 0 \\ 0 & I_p \end{pmatrix} \Theta_n(\overline{z})^* = \Delta_n^{-1} \begin{pmatrix} C_n(\overline{z})^* & C_n^\circ(\overline{z})^* \\ A_n(\overline{z})^* & -A_n^\circ(\overline{z})^* \end{pmatrix}$$

satisfy the recursion

$$X_n(z) = \begin{pmatrix} I_p & \alpha_n \\ \beta_n & I_p \end{pmatrix}^* \begin{pmatrix} zI_p & 0 \\ 0 & I_p \end{pmatrix} X_{n-1}(z), \quad n = 1, 2, \ldots.$$

Since, as already noticed, $\Delta_0 = I_{2p}$, we have:

$$\Delta_n^{-1} \begin{pmatrix} C_n(\overline{z})^* & C_n^\circ(\overline{z})^* \\ A_n(\overline{z})^* & -A_n^\circ(\overline{z})^* \end{pmatrix} \begin{pmatrix} I_p & I_p \\ I_p & -I_p \end{pmatrix} \frac{1}{2} = M_n(z),$$

where we recall that $M_n(z)$ is the solution of (1.1) subject to the initial condition $M_0(z) = I_{2p}$. Hence, with $Y(z)$ defined by (1.4),

$$M_n(z) = X_n(z)Y(z) = \Delta_n^{-1} \begin{pmatrix} C_n(\overline{z})^* & C_n^\circ(\overline{z})^* \\ A_n(\overline{z}) & -A_n^\circ(\overline{z})^* \end{pmatrix} \begin{pmatrix} I_p & I_p \\ I_p & -I_p \end{pmatrix} \frac{1}{2},$$

where $X_n(z)$ is the solution to (1.1) subject to the asymptotic (1.5). Recalling (1.7) we obtain:

$$\begin{aligned} &\lim_{n\to\infty} \begin{pmatrix} z^{-n}I_p & 0 \\ 0 & I_p \end{pmatrix} X_n(z)Y(z) \\ &\qquad = \begin{pmatrix} \delta_1^{-1} & 0 \\ 0 & \delta_2^{-1} \end{pmatrix} \begin{pmatrix} \lim_{n\to\infty} z^{-n}C_n(\overline{z})^* & \lim_{n\to\infty} z^{-n}C_n^\circ(\overline{z})^* \\ \lim_{n\to\infty} A_n(\overline{z}) & -\lim_{n\to\infty} A_n^\circ(\overline{z})^* \end{pmatrix} \begin{pmatrix} I_p & I_p \\ I_p & -I_p \end{pmatrix} \frac{1}{2}. \end{aligned}$$

Hence,

$$\begin{pmatrix} \delta_1 & 0 \\ 0 & \delta_2 \end{pmatrix} Y(z) \begin{pmatrix} I_p & I_p \\ I_p & -I_p \end{pmatrix} = \begin{pmatrix} \lim_{n\to\infty} z^{-n}C_n(\overline{z})^* & \lim_{n\to\infty} z^{-n}C_n^\circ(\overline{z})^* \\ \lim_{n\to\infty} A_n(\overline{z}) & -\lim_{n\to\infty} A_n^\circ(\overline{z})^* \end{pmatrix}.$$

In particular we have (4.6).

STEP 4: $W(z)$ *is the spectral function of the first-order discrete system associated to the pair* (4.7).

By [12, Theorem 10.4 p. 116], we have for $|z| = 1$

$$\begin{aligned} W(z) &= \lim_{n\to\infty} A_n(z)^{-*}\gamma_{00}^{(n)}A_n(z)^{-1} \\ &= \lim_{n\to\infty} C_n(z)^{-*}\gamma_{nn}^{(n)}C_n(z)^{-1}. \end{aligned}$$

and thus, still on the unit circle

$$\begin{aligned} W(1/z) &= \lim_{n\to\infty} A_n(\overline{z})^{-*}\gamma_{00}^{(n)}A_n(\overline{z})^{-1} \\ &= \lim_{n\to\infty} C_n(\overline{z})^{-*}\gamma_{nn}^{(n)}C_n(\overline{z})^{-1}. \end{aligned}$$

Hence, by the preceding two steps,

$$\begin{aligned} W(1/z) &= (Y_{21}(z)+Y_{22}(z))^{-1}\delta_1^{-1}\delta_1\delta_1^{-1}(Y_{21}(z)+Y_{22}(z))^{-*} \\ &= (Y_{21}(z)+Y_{22}(z))^{-1}\delta_1^{-1}(Y_{21}(z)+Y_{22}(z))^{-*} \\ &= (Y_{11}(z)+Y_{12}(z))^{-1}\delta_2^{-1}\delta_2\delta_2^{-1}(Y_{11}(z)+Y_{12}(z))^{-*} \\ &= (Y_{11}(z)+Y_{12}(z))^{-1}\delta_2^{-1}(Y_{11}(z)+Y_{12}(z))^{-*} \end{aligned}$$

and hence the result. □

In [3] we called admissible sequences of the form (4.4)–(4.5) Szegö admissible sequences.

In the next theorem we assume that the weight function is analytic at the origin and at infinity. This allows us to use formulas from [15].

Theorem 4.2. *Let $W(z)$ be a rational function analytic at infinity and at the origin, and without poles on the unit circle. Assume that $W(e^{it}) > 0$ for $t \in [0, 2\pi]$ and that the normalization* (4.1) *is in force. Then, $W(z)$ is the spectral function of a uniquely determined first-order system. The corresponding associated sequence is obtained as follows: let*

$$W(z) = D + zC(I - zA)^{-1}B$$

be a minimal realization of W. Then α_n and β_n are given by

$$\begin{aligned} \alpha_n &= (D - CA^{-1}B)^{-1}CA^{-1}\left((I-P_0)(A^\times)^{-n}\big|_{\ker P_0}\right)^{-1}(I-P_0)B, \\ \beta_n &= -D^{-1}C\left(P_0(A^\times)^n\big|_{\operatorname{Im} P_0}\right)^{-1}P_0A^{-1}B, \end{aligned} \tag{4.7}$$

and the associated sequence of diagonals is given by $\Delta_n = \operatorname{diag}(d_{1,n}, d_{2,n})$ with

$$\begin{aligned} d_{1,n} &= D^{-1} + D^{-1}C(A^\times)^n W_{n+1}^{-1}P_0A^{-1}BD^{-1}, \\ d_{2,n} &= D^{-1} + D^{-1}CW_{n+1}^{-1}P_0A^{-(n+1)}(A^\times)^nBD^{-1}, \end{aligned}$$

where P_0 denotes the Riesz projection corresponding to the spectrum of A outside the closed unit disk,

$$P_0 = I - \frac{1}{2\pi i}\int_{\mathbb{T}}(zI - A)^{-1}dz, \tag{4.8}$$

and where W_n is given by

$$W_n(I - P_0 + P_0A)^{-n}(I - P_0 + P_0A^{\times n}). \tag{4.9}$$

The proof is a special case of the previous theorem. Formulas (4.7) have been proved in our previous paper [3], and are the discrete analogue of [6, (3.1) p. 9], where the potential associated to a canonical differential expression was computed in terms of a minimal realization of the spectral function.

We now turn to the third case, where we start from a spectral factor.

Theorem 4.3. *Let $g_+(z)$ be a $\mathbb{C}^{p\times p}$-valued rational function analytic and invertible in the closed unit disk, and at infinity. Let*

$$W(z) = g_+(z)g_+(1/z^*)^*,$$

and assume that the normalization (4.1) *is in force. Then $W(z)$ is the spectral function of a first-order discrete system of type* (1.1)*. Let $g_+(z) = d+zc(I-za)^{-1}b$ be a minimal realization of $g_+(z)$ and let X and Y be the solutions of the Stein equations*

$$X - aXa^* = bb^* \tag{4.10}$$

and

$$Y - a^{\times *}Ya^{\times} = (d^{-1}c)^*(d^{-1}c). \tag{4.11}$$

Assume that a is invertible (that is, $W(z)$ is analytic at the origin and at infinity). Then the following formulas hold:

$$\begin{aligned}\alpha_n &= (d - ca^{-1}b)d^*ca^{-1}(a^{\times})^n(I + X(Y - (a^{\times})^{*n}Y(a^{\times})^n))^{-1}(bd^* + aXc^*),\\ \beta_n &= (d(d^*-b^*a^{-*}c^*))^{-1}(cX+db^*a^{-*})\left(I+(Y-(a^{\times *})^nY(a^{\times})^n)X\right)^{-1}(a^{\times *})^nc^*.\end{aligned} \tag{4.12}$$

The associated sequence of diagonals is given by $\Delta_n = \operatorname{diag}(d_{1,n}, d_{2,n})$ where

$$\begin{aligned}d_{1,n} &= (d(d^* - c^*a^{-*}b^*))^{-1}\\ &\quad\times\Big(I + (-c(a^{\times})^n(I + X(Y - (a^{\times})^{*n}Y(a^{\times})^n))^{-1}X(a^{\times})^{*(n+1)}\\ &\qquad -d^*a^{-*}c^*(a^{\times})^*(I + (Y - (a^{\times})^{*n}Y(a^{\times})^n)X)\\ &\quad\times(I + (Y- (a^{\times})^{*(n+1)}Y(a^{\times})^{(n+1)})X)^{-1}(a^{\times})^{(n+1)}a^{-*}c^*(d(d^* - b^*a^{-*}c^*)^{-1}\Big),\\ d_{2,n} &= (d(d^* - c^*a^{-*}b^*))^{-1}\\ &\quad\times\big(I - (cX + d^*b^*a^{-*})\\ &\quad\times(I + (Y - (a^{\times})^{*(n+1)}Y(a^{\times})^{(n+1)})X)^{-1}(a^{\times})^*\\ &\quad\times\big(b(d^* - b^*a^{-*}c^*) + ((a^{\times})^{*n}Y(a^{\times})^n - Y)a^{-*}c^*\big)(d(d^* - c^*a^{-*}b^*))^{-1}\big).\end{aligned}$$

Proof. The fact that $W(z)$ is the spectral function of a system (1.1) stems from Theorem 4.1. We now prove formulas (4.12). In the arguments to obtain a formula for the Schur coefficients α_n and β_n in terms of a minimal realization of $g_+(z)$ we make much use of computations from our previous paper [5].

Let $g_+(z) = d + zc(I - za)^{-1}b$ be a minimal realization of $g_+(z)$. By hypothesis the matrix a is invertible. Hence, a minimal realization of $g_+(1/z^*)^*$ is given by

$$\begin{aligned}g_+(1/z^*)^* &= d^* + b^*(zI - a^*)^{-1}c^*\\ &= d^* - b^*a^{-*}c^* + b^*\left((zI - a^*)^{-1} + a^{-*}\right)c^*\\ &= d^* - b^*a^{-*}c^* - zb^*(I - a^{-*})^{-1}c^*\\ &= d^* - b^*a^{-*}c^* - zb^*a^{-*}(I - za^{-*})^{-1}a^{-*}c^*,\end{aligned}$$

and hence the matrices

$$A = \begin{pmatrix} a & -bb^*a^{-*} \\ 0 & a^{-*} \end{pmatrix},\ B = \begin{pmatrix} b(d^* - b^*a^{-*}c^*) \\ a^{-*}c^* \end{pmatrix},\ C = \begin{pmatrix} c & -db^*a^{-*} \end{pmatrix}, \tag{4.13}$$

and

$$D = d(d^* - b^*a^{-*}c^*) \tag{4.14}$$

define a minimal realization $W(z) = D + zC(I - zA)^{-1}B$ of $W(z)$. See [5, Theorem 3.3 p. 155]). Furthermore, the Riesz projection (4.8) is given by

$$P_0 = \begin{pmatrix} 0 & -X \\ 0 & I \end{pmatrix},$$

where X is the solution of the Stein equation (4.10). We have (see [5, Equation (3.21) p. 156])

$$(A^\times)^n = \begin{pmatrix} (a^\times)^n & 0 \\ Y(a^\times)^n - (a^{\times *})^{-n}Y & (a^{\times *})^{-n} \end{pmatrix}, \tag{4.15}$$

where Y is the solution to the Stein equation (4.11). Therefore

$$P_0(A^\times)^n P_0 = \begin{pmatrix} 0 & X(Ya^{\times n} - (a^{\times *})^{-n}Y)X - X(a^{\times *})^{-n} \\ 0 & -(Ya^{\times n} - (a^{\times *})^{-n}Y)X + (a^{\times *})^{-n} \end{pmatrix},$$

and hence

$$(P_0(A^\times)^n\big|_{\operatorname{Im} P_0})^{-1} = (a^{\times *})^n \left(I + (Y - (a^{\times *})^n Y a^{\times n})X\right)^{-1}.$$

We remark that the matrix $I + (Y - (a^{\times *})^n Y a^{\times n})X$ is indeed invertible since $X > 0$ and since, for every $n \geq 0$,

$$Y - (a^{\times *})^n Y a^{\times n} \geq 0.$$

The formula for β_n follows.

To prove the formula for α_n we first note that (using (4.15))

$$(I - P_0)(A^\times)^{-n}(I - P_0) = \begin{pmatrix} I & X \\ 0 & 0 \end{pmatrix} \begin{pmatrix} (a^\times)^{-n} & 0 \\ Y(a^\times)^{-n} - a^{\times n}Y & (a^\times)^{*n} \end{pmatrix} \begin{pmatrix} I & X \\ 0 & 0 \end{pmatrix}$$
$$= \begin{pmatrix} (I + X(Y - (a^\times)^{*n}Y(a^\times)^n)(a^\times)^{-n} & (I + X(Y - (a^\times)^{*n}Y(a^\times)^n)(a^\times)^{-n}X \\ 0 & 0 \end{pmatrix}.$$

Moreover,

$$D - CA^{-1}B = W(\infty) = (d - ca^{-1}b)d^*,$$
$$CA^{-1}(I - P_0) = \begin{pmatrix} ca^{-1} & ca^{-1}X \end{pmatrix},$$

and (using the Stein equation (4.10))

$$(I - P_0)B = \begin{pmatrix} b(d^* - b^*a^{-*}c^*) + Xa^{-*}c^* \\ 0 \end{pmatrix}$$
$$= \begin{pmatrix} bd^* + aXc^* \\ 0 \end{pmatrix}.$$

The formula for α_n follows.

We now compute $d_{1,n} = \gamma_{nn}^{(n)}$. Using [15, p. p. 36] we have

$$\gamma_{nn}^{(n)} = D^{-1}(I + C(A^\times)^n W_{n+1}^{-1} P_0 A^{-(n+1)} B D^{-1}),$$

where A, B, C and D are given by (4.13)–(4.14) and where W_n is defined by (4.9). In [5, (4.8) p. 164] we proved that

$$W_{n+1} P_0 A^{-(n+1)} = \begin{pmatrix} 0 & a_{n+1} \\ 0 & b_{n+1} \end{pmatrix} \tag{4.16}$$

where

$$\begin{aligned} a_{n+1} &= -X(I + (Y - (a^\times)^{*(n+1)} Y (a^\times)^{(n+1)})X)^{-1}(a^\times)^{*(n+1)}, \\ b_{n+1} &= (I + (Y - (a^\times)^{*(n+1)} Y (a^\times)^{(n+1)})X)^{-1}(a^\times)^{*(n+1)}. \end{aligned}$$

Using (4.15) we have

$$(A^\times)^n W_{n+1} P_0 A^{-(n+1)} = \begin{pmatrix} 0 & (a^\times)^n a_{n+1} \\ 0 & h_n \end{pmatrix}$$

where

$$\begin{aligned} h_n &= (Y(a^\times)^n - (a^\times)^{-*n} Y) a_{n+1} + (a^\times)^{-*n} b_{n+1} \\ &= (a^\times)^{-*n} \Big\{ -((a^\times)^{*n} Y (a^\times)^n - Y)(I + X(Y - (a^\times)^{*(n+1)} Y (a^\times)^{(n+1)}))^{-1} \\ &\qquad \times X (a^\times)^{*(n+1)} \\ &\qquad\qquad + (I + (Y - (a^\times)^{*(n+1)} Y (a^\times)^{(n+1)} X)^{-1} (a^\times)^{*(n+1)} \Big\} \\ &= (a^\times)^{-*n} \Big\{ (Y - (a^\times)^{*n} Y (a^\times)^n)(I + X(Y - (a^\times)^{*(n+1)} Y (a^\times)^{(n+1)}))^{-1} X \\ &\qquad + (I + (Y - (a^\times)^{*(n+1)} Y (a^\times)^{(n+1)})X)^{-1} \Big\} (a^\times)^{*(n+1)} \\ &= (a^\times)^{-*} \\ &\times (I + (Y - (a^\times)^{*n} Y (a^\times)^n)X)(I + (Y - (a^\times)^{*(n+1)} Y (a^\times)^{(n+1)})X)^{-1} (a^\times)^{*(n+1)}. \end{aligned}$$

Since

$$C(A^\times)^n W_{n+1} P_0 A^{-(n+1)} B = (c(a^\times)^n a_{n+1} - d^* a^{-*} c^* h_n a^{-*} c^*,$$

we get the formula for $d_{1,n}$.

Finally, we compute the formula for $d_{2,n} = \gamma_{00}^{(n)}$. By the formula in [15, p. 36] we now have

$$\gamma_{00}^{(n)} = D^{-1} \left\{ I + C W_{n+1} P_0 A^{-(n+1)} (A^\times)^n B D^{-1} \right\}.$$

By (4.16) and [15, p. 36] we have

$$\begin{aligned} &C W_{n+1} P_0 A^{-(n+1)} \\ &= \begin{pmatrix} 0 & -(cX + db^* a^{-*})(I + (Y - (a^\times)^{*(n+1)} Y (a^\times)^{(n+1)})X)^{-1} (a^\times)^{*(n+1)} \end{pmatrix}. \end{aligned}$$

Hence, using (4.15) we obtain

$$\begin{aligned} CW_{n+1}P_0A^{-(n+1)}(A^\times)B = &-(cX + d^*b^*a^{-*}) \\ &\times (I + (Y - (a^\times)^{*(n+1)}Y(a^\times)^{(n+1)})X)^{-1}(a^\times)^* \\ &\times \big(b(d^* - b^*a^{-*}c^*) + ((a^\times)^{*n}Y(a^\times)^n - Y)a^{-*}c^*\big) \end{aligned}$$

and the formula for $\gamma_{00}^{(n)}$ follows. □

These formulas are the discrete analogs of the formula given in [6, Theorem 3.5 p.9], where we computed the potential associated to a canonical differential expression in terms of a minimal realization of a spectral factor of the spectral function. Connections with the formulas for Nehari admissible sequences given in [3, Section 1.3] will be explored in a separate publication.

5. Connection with the scattering function

The connection between the scattering function and the spectral function allows to reconstruct the discrete system from the scattering function by building first the associated spectral function. We are given two strictly positive matrices δ_1 and δ_2 in $\mathbb{C}^{p\times p}$, and consider a $\mathbb{C}^{p\times p}$-valued rational function $S(z)$ which admits a spectral factorization $S(z) = S_-(z)S_+(z)$ and satisfies the following two conditions:

$$S(z)^*\delta_1 S(z) = \delta_2, \quad |z| = 1, \tag{5.1}$$

and

$$\frac{1}{2\pi}\int_0^{2\pi} S_-(e^{it})\delta_1^{-1}S_-(e^{it})^* dt = I_p. \tag{5.2}$$

We also assume that the factors $S_+(z)$ and $S_-(z)$ are normalized by $S_+(0) = S_-(\infty) = I_p$. Note that for a given pair (δ_1, δ_2) there need not exist associated functions $S(z)$ with the required properties. For instance, in the scalar case we necessarily have $\delta_1 = \delta_2$ (see [3]) and then $S(z)$ is unitary on the unit circle.

Using (5.1) we define

$$S_-(1/z)\delta_1^{-1}S_-(1/z)^{-*} = S_+(1/z)\delta_2^{-1}S_+(1/z)^* \stackrel{\text{def.}}{=} W(z). \tag{5.3}$$

By Theorem (4.1) the function $W(z)$ is the spectral function of a uniquely defined first-order discrete system of the form (1.1) with Szegö admissible sequence defined by (4.4)–(4.5). We know from the proof of Step 1 of Theorem 4.1 that the limits

$$\lim_{n\to\infty}\gamma_{00}^{(n)} \quad \text{and} \quad \lim_{n\to\infty}\gamma_{nn}^{(n)}$$

exist and are strictly positive. For the moment being we denote these limits by k_1 and k_2. Let $Y(z)$ be defined by (1.4). Then (see Section 1)

$$\begin{aligned} W(z) &= (Y_{21} + Y_{22})(1/z)k_2^{-1}(Y_{21} + Y_{22}(1/z))^{-*} \\ &= (Y_{11} + Y_{12})(1/z)k_1^{-1}(Y_{11} + Y_{12}(1/z))^{-*}. \end{aligned} \tag{5.4}$$

By uniqueness of the spectral factorizations and comparing (5.3) and (5.4) we have

$$\begin{aligned}\delta_1 &= k_1,\\ \delta_2 &= k_2,\\ S_-(1/z) &= (Y_{11}+Y_{12})(1/z),\\ S_+(1/z) &= ((Y_{21}+Y_{22})(1/z))^{-1}.\end{aligned}$$

Hence, the associated scattering function is equal to

$$(Y_{11}(z)+Y_{12}(z))(Y_{21}(z)+Y_{22}(z))^{-1} = S_-(z)S_+(z) = S(z).$$

This way, we can reconstruct the system associated to the scattering function using the spectral function.

Theorem 5.1. *Let $S(z)$ be a rational matrix-function which admits a spectral factorization and satisfies conditions* (5.1) *and* (5.2) *for some pair of strictly positive matrices δ_1 and δ_2. Then $S(z)$ is the scattering function of the first-order discrete system with spectral function*

$$S_-(1/z)\delta_1^{-1}S_-(1/z)^{-*} = S_+(1/z)\delta_2^{-1}S_+(1/z)^*.$$

6. Connection with the reflection coefficient function

Let $R \in \mathcal{W}_+^{p\times p}$ be a rational function which is strictly contractive in the closed unit disk. The function

$$W(z) = (I_p - zR(z))^{-1}(I_p - R(z)R(z)^*)(I_p - zR(z))^{-*}, \quad |z| = 1, \tag{6.1}$$

is strictly positive on the unit circle and is the restriction there of the rational function

$$W(z) = \frac{1}{2i}(N(z) - N(1/z^*)^*) \quad \text{with} \quad N(z) = i(I_p - zR(z))(I_p + zR(z))^{-1}.$$

Hence $W(z)$ is the spectral function of a first-order discrete system. Since $R(z)$ defined uniquely $W(z)$ we have:

Theorem 6.1. *Let $R \in \mathcal{W}_+^{p\times p}$ be a rational function which is strictly contractive in the closed unit disk. Then it is the reflection coefficient function of the first-order canonical discrete system* (1.1) *with associated spectral function* (6.1).

Indeed, by Theorem 4.1 the function

$$W(z) = \frac{1}{2i}(N(z) - N(1/z^*)^*), \quad |z| = 1,$$

is the spectral function of a uniquely defined first-order discrete system and $R(z)$ is uniquely determined by $W(z)$.

References

[1] D. Alpay. *The Schur algorithm, reproducing kernel spaces and system theory*, volume 5 of *SMF/AMS Texts and Monographs*. American Mathematical Society, Providence, RI, 2001. Translated from the 1998 French original by Stephen S. Wilson.

[2] D. Alpay and H. Dym. On applications of reproducing kernel spaces to the Schur algorithm and rational J-unitary factorization. In I. Gohberg, editor, *I. Schur methods in operator theory and signal processing*, volume 18 of *Operator Theory: Advances and Applications*, pages 89–159. Birkhäuser Verlag, Basel, 1986.

[3] D. Alpay and I. Gohberg. Discrete systems and their characteristic spectral functions. Mediterranean Journal of Mathematics. To appear, 2007.

[4] D. Alpay and I. Gohberg. Unitary rational matrix functions. In I. Gohberg, editor, *Topics in interpolation theory of rational matrix-valued functions*, volume 33 of *Operator Theory: Advances and Applications*, pages 175–222. Birkhäuser Verlag, Basel, 1988.

[5] D. Alpay and I. Gohberg. Inverse spectral problems for difference operators with rational scattering matrix function. *Integral Equations Operator Theory* **20** no. 2 (1994), 125–170.

[6] D. Alpay and I. Gohberg. Inverse spectral problem for differential operators with rational scattering matrix functions. *Journal of differential equations* **118** (1995), 1–19.

[7] D. Alpay and I. Gohberg. Discrete analogs of canonical systems with pseudo–exponential potential. Definitions and formulas for the spectral matrix functions. In D. Alpay and I. Gohberg, editors, *The state space method. New results and new applications*, volume 161, pages 1–47. Birkhäuser Verlag, Basel, 2006.

[8] D. Alpay and I. Gohberg. Discrete analogs of canonical systems with pseudo–exponential potential. Inverse problems. In D. Alpay and I. Gohberg, editors, *Interpolation, Schur functions and moment problems*, volume 165, pages 31–65. Birkhäuser Verlag, Basel, 2006.

[9] N. Aronszajn. Theory of reproducing kernels. *Trans. Amer. Math. Soc.* **68** (1950), 337–404.

[10] H. Bart, I. Gohberg, and M.A. Kaashoek. *Minimal factorization of matrix and operator functions*, volume 1 of *Operator Theory: Advances and Applications*. Birkhäuser Verlag, Basel, 1979.

[11] L. de Branges and L.A. Shulman. Perturbation theory of unitary operators. *J. Math. Anal. Appl.* **23** (1968), 294–326.

[12] H. Dym. Hermitian block Toeplitz matrices, orthogonal polynomials, reproducing kernel Pontryagin spaces, interpolation and extension. In I. Gohberg, editor, *Orthogonal matrix-valued polynomials and applications (Tel Aviv, 1987–88)*, volume 34, pages 79–135. Birkhäuser, Basel, 1988.

[13] H. Dym. *J–contractive matrix functions, reproducing kernel Hilbert spaces and interpolation*. Published for the Conference Board of the Mathematical Sciences, Washington, DC, 1989.

[14] I. Gohberg and M.A. Kaashoek. Block Toeplitz operators with rational symbols. In I. Gohberg, J.W. Helton, and L. Rodman, editors, *Contributions to operator theory*

and its applications (Mesa, AZ, 1987), volume 35 of *Oper. Theory Adv. Appl.*, pages 385–440. Birkhäuser, Basel, 1988.

[15] I. Gohberg, M.A. Kaashoek, and F. van Schagen. Szegö–Kac–Achiezer formulas in terms of realizations of the symbol. *J. Funct. Anal.* **74** (1987), 24–51.

[16] I. Gohberg and S. Levin. On an open problem for block-Toeplitz matrices. *Integral Equations Operator Theory*, **2** no. 1 (1979), 121–129.

[17] I. C. Gohberg and I. A. Fel′dman. *Convolution equations and projection methods for their solution.* American Mathematical Society, Providence, R.I., 1974. Translated from the Russian by F. M. Goldware, Translations of Mathematical Monographs, Vol. 41.

[18] M. A. Kaashoek and A. L. Sakhnovich. Discrete skew self-adjoint canonical system and the isotropic Heisenberg magnet model. *J. Funct. Anal.* **228** no. 1 (2005), 207–233.

[19] S. Saitoh. *Theory of reproducing kernels and its applications*, volume 189. Longman scientific and technical, 1988.

Daniel Alpay
Department of Mathematics
Ben-Gurion University of the Negev
Beer-Sheva 84105
Israel
e-mail: dany@math.bgu.ac.il

Israel Gohberg
School of Mathematical Sciences
The Raymond and Beverly Sackler Faculty of Exact Sciences
Tel-Aviv University
Tel-Aviv, Ramat-Aviv 69989
Israel
e-mail: gohberg@post.tau.ac.il

Operator Theory:
Advances and Applications, Vol. 179, 25–34

Stability of Dynamical Systems via Semidefinite Programming

Mihály Bakonyi and Kazumi N. Stovall

Abstract. In this paper, we study stability of nonlinear dynamical systems by searching for Lyapunov functions of the form $\Lambda(x) = \sum_{i=1}^{m} \alpha_i x_i + \frac{1}{2} \sum_{i=1}^{m} \lambda_i x_1^2$, $\lambda_i > 0$, $i = 1, \ldots, m$, respectively $x^T A x$, where A is a positive definite real matrix. Our search for Lyapunov functions is based on interior point algorithms for solving certain positive definite programming problems and is applicable for non-polynomial systems not considered by similar methods earlier.

Mathematics Subject Classification (2000). 34D20, 34D45, 90C22.

Keywords. Lyapunov function, attractor, semidefinite programming.

1. Introduction

The aim of the paper is to use semidefinite programming methods in the study of stability of dynamical systems. It is well known that for linear systems $x' = Bx$, a matrix $A > 0$ such that $AB + B^T A < 0$ defines the Lyapunov function $\Lambda(x) = x^T A x$ which implies the stability of the system. A new efficient algorithm was introduced in [12] to search for Lyapunov functions that are sums of squares. This method was generalized in [10] and [11]. We claim in this paper a method which can be applied to even more general systems.

For definitions in the area of dynamical systems we mention [16] as a classical reference. A compact region Ω in $\mathbb{R}^n$ is called an *attractor region* for a dynamical system if any trajectory for the system starting outside Ω enters Ω after a finite time interval T, determined by the initial distance to Ω, and no trajectory starting in Ω leaves Ω. The existence of an attractor region gives precise information about the asymptotic behavior of the system. Even if an attractor region exists, it is not always possible to determine its shape. Research has shown that chaotic behavior and fractal attractors are common. The study of chaos and fractals are currently booming research areas, however, we do not want to enter into details here since

it is not the aim of the present work. Our goal is to study the existence of an attractor region within a hyperellipsoidal region, possibly of minimal diameter. That is equivalent to the existence of such a region Ω, such that for each solution $x(t)$ of the system there exists $T > 0$ which depends only on $\|x(0)\|$ such that $x(t) \in \Omega$ for $t > T$.

We say 0 is the *attractor trajectory* for a system if for each $\epsilon > 0$ there exists $T > 0$ such that for each solution $x(t)$, $\|x(t)\| < \epsilon$ for $t > T$. T must depend only on ϵ and $\|x(0)\|$.

In [6], Lyapunov functions of the form

$$\Lambda(x) = \sum_{i=1}^{m} \alpha_i x_i + \frac{1}{2} \sum_{i=1}^{m} \lambda_i x_1^2, \tag{1}$$

$\lambda_i > 0$, $i = 1, \ldots, m$, were considered and it was shown that if $\lambda = (\lambda_1, \ldots, \lambda_n)^T$ is a null-vector of a certain matrix determined by a dynamical system of a particular form, then there exists an attractor region for the system within a hyperellipsoidal region. Conditions on the sign-pattern of this matrix which guarantee the existence of an entry-wise positive null-vector were established in [7]. Based on an algorithm in Section 2, we can decide in polynomial time whether there exists a choice of Lyapunov function of type (1) which implies the existence of an attractor region.

We also consider Lyapunov functions of the form $\Lambda(x) = x^T Ax$, where A is a real $n \times n$ positive definite matrix. For a class of dynamical systems, 0 is the attractor trajectory for the system when $\mathrm{tr}(AB_r) = 0$, where $B_r, r = 1, \ldots, m$, are some symmetric matrices determined by the system. The existence of such A is decided by an algorithm in Section 2. Interior-point methods were previously used (see [14], [4], and [2]) for finding Lyapunov functions for linear time-variant dynamical systems, by solving a set of linear inequalities $B_k^T A + AB_k < 0$, $k = 1, \ldots, L$, for $A > 0$. In [12], a polynomial-time algorithm was developed for polynomial systems for finding Lyapunov functions that can be represented as sums of squares of polynomials. This method was extended in [10] for certain non-polynomial systems which can be transformed to equivalent polynomial ones. A review of the latter results can be found in [11]. The algorithm in this paper can be applied to systems not covered by [10]. Its implementation is also simpler.

2. Positive Definite Optimization Problems

In this section we present an interior point algorithm derived from an algorithm in [3]. It will be used in Section 3 for finding Lyapunov functions. For sake of completeness, we include here the details. We refer the reader to [15] for a survey on semidefinite programming.

Let $A_0, A_1, \ldots, A_m$ be $n \times n$ symmetric matrices. Assume $A_0 > 0$ and that there is no positive semidefinite matrix of the form $\sum_{i=1}^{m} x_i A_i$. Consider the set $\mathcal{S}= \{Q = A_0 + \sum_{i=1}^{m} x_i A_i : Q > 0\}$, which is nonempty and bounded.

Let $\phi : \mathcal{S} \to \mathbb{R}$ be defined by $\phi(Q) = \log\det Q$. It is known ([5], Theorem 7.6.7) that ϕ is a concave function, and since near the boundary of $\mathcal{S}$, ϕ approaches $-\infty$, ϕ takes on a maximum value at a unique point $P_0 \in \mathcal{S}$. It is well-known that P_0 is the only element of $\mathcal{S}$ which verifies $\mathrm{tr}(A_i P_0) = 0$ for $i = 1, \ldots, m$ (see, e.g., [1]). A method for approximating P_0 can be found in Section 1.2 of [9] (see also Section 3.4 of [3] and [1]).

For our applications, we are interested in solving the following feasibility problem.

Problem 1. "Given the matrices $A_i = A_i^T \in \mathbb{R}^{n\times n}$, $i = 1, \ldots, m$, determine whether there exists a positive definite matrix of the form $\sum_{i=1}^{m} x_i A_i$."

Problem 1 can be solved adapting an algorithm in [3] to solve:

$$\begin{cases} \text{Minimize } \mu \\ \mu I + \sum_{i=1}^{m} x_i A_i > 0 \\ |x_i| < 1. \end{cases} \tag{2}$$

If $\mu^{opt} < 0$, then there exist $|x_i| < 1$ such that $\sum_{i=1}^{m} x_i A_i > -\mu^{opt} I$, thus there exists a positive definite matrix of the form $\sum_{i=1}^{m} x_i A_i$. It is easy to see that $\mu^{opt} < 0$ is also a necessary condition for the existence of a positive definite matrix of the form $\sum_{i=1}^{m} x_i A_i$.

The algorithm to solve the problem works as follows. Initialize with $\mu^{(0)} = 1$ and $x^{(0)} = 0 \in \mathbb{R}^m$. Let

$$\mu^{(k+1)} = \frac{1}{2}\mu^{(k)} - \frac{1}{2}\lambda_{\min}(\sum_{i=1}^{m} x_i A_i)$$

$$x^{(k+1)} = x^*(\mu^{(k+1)})$$

Here $\lambda_{\min}(\sum_{i=1}^{m} x_i A_i)$ denotes the smallest eigenvalue of $\sum_{i=1}^{m} x_i A_i$, and $x^*(\mu^{(k+1)})$ is the vector in $\mathbb{R}^m$ which maximizes $\log\det(\mu^{(k+1)} I + \sum_{i=1}^{m} x_i A_i) + \sum_{i=1}^{m} \log(1 - x_i^2)$ subject to $\mu^{(k+1)} I + \sum_{i=1}^{m} x_i A_i > 0$ and $|x_i| < 1$. For approximating $x^*(\mu^{(k+1)})$ one can use the algorithm mentioned at the beginning of this section.

As shown in [3], we have that $\mu^{(k)} \longrightarrow \mu^{opt}$ at least geometrically, and $\mu^{opt} < 0$ or $\mu^{opt} = 0$ decides whether there is a positive definite matrix of the form $\sum_{i=1}^{m} x_i A_i$ or not.

The following is the dual of Problem 1.

Problem 2. "Given $B_i^T = B_i \in \mathbb{R}^{n\times n}$, $i = 1, \ldots, r$, find whether there exists $A \in \mathbb{R}^{n\times n}$, $A > 0$, such that $\mathrm{tr}(AB_i) = 0$, for $i = 1, \ldots, r$."

Indeed, let $A_i^T = A_i \in \mathbb{R}^{n\times n}$, $i = 1, \ldots, m$, be a linear basis for $\{Q = Q^T \in \mathbb{R}^{n\times n} : \mathrm{tr}(QB_i) = 0,\ i = 1, \ldots, r\}$. Then Problem 2 is equivalent to the existence of $A > 0$ of the form $A = \sum_{i=1}^{m} x_i A_i$.

A particular case of Problem 1 is to determine whether there exists an entry-wise positive vector that is a linear combination of some given vectors $v_1, \ldots, v_m \in \mathbb{R}^n$. The latter is equivalent to the existence of an entry-wise positive vector in the null-space of a given matrix. As suggested to us by Florian Potra, the latter problem can be solved in polynomial-time by applying Algorithm 2.1 in [13].

3. Stability of Nonlinear Dynamical Systems

Consider the dynamical system

$$x_i'(t) = -\epsilon_i x_i(t) + g_i(x) \tag{3}$$

where $\epsilon_i > 0$, $i = 1, \ldots, n$. Such systems are most commonly studied. We assume here the existence of a linearly independent set of functions $\{f_l(x)\}_{l=1}^m$ which span the set $\{g_i(x)\}_{i=1}^n$ as well as $\{x_k g_i(x) : i, k = 1, \ldots, n\}$. A typical situation for this is when each $g_i(x)$ is a polynomial in $x_1, x_2, \ldots, x_n$. We can thus assume the system (3) is of the form

$$x_i'(t) = -\epsilon_i x_i(t) + \sum_{l=1}^{m} k_{li} f_l(x) \tag{4}$$

for $i = 1, \ldots, n$. We are searching in this case for a Lyapunov function (see [6]) of the form

$$\Lambda(x) = \sum_{i=1}^{n} \alpha_i x_i + \frac{1}{2} \sum_{i=1}^{n} \lambda_i x_i^2 \tag{5}$$

where α_i and $\lambda_i > 0$ are unknown for $i = 1, \ldots, n$. The level sets of $\Lambda(x)$ are hyperellipsoids centered at $(-\frac{\alpha_1}{\lambda_1}, \ldots, -\frac{\alpha_n}{\lambda_n})$. Then (by denoting $\Lambda'(x) = \frac{d\Lambda(x(t))}{dt}$)

$$\begin{aligned}\Lambda'(x) = \sum_{i=1}^{n} \alpha_i x_i' + \sum_{i=1}^{n} \lambda_i x_i x_i' = -\sum_{i=1}^{n} \epsilon_i \alpha_i x_i - \sum_{i=1}^{n} \epsilon_i \lambda_i x_i^2 \\ + \sum_{l=1}^{m} (\sum_{i=1}^{n} \alpha_i k_{li}) f_l(x) + \sum_{l=1}^{m} (\sum_{i=1}^{n} \lambda_i k_{li} x_i f_l(x))\end{aligned} \tag{6}$$

By our assumption we have that

$$x_i f_l(x) = \sum_{j=1}^{m} \alpha_j^{(il)} f_j(x).$$

Then

$$\sum_{l=1}^{m}\sum_{i=1}^{n} \lambda_i k_{li} x_i f_l(x) = \sum_{j=1}^{m}\sum_{i=1}^{n}(\sum_{l=1}^{m} k_{li}\alpha_j^{(il)})\lambda_i f_j(x) = \sum_{j=1}^{m}(\sum_{i=1}^{n}\beta_{ji}\lambda_i) f_j(x),$$

where $\beta_{ji} = \sum\limits_{l=1}^{m} k_{li}\alpha_j^{(il)}$ for $i = 1, \ldots, n$ and $j = 1, \ldots, m$.

We try to find α_i and $\lambda_i > 0$ which reduce the last two terms of (6) to 0. This means that we want to make the coefficient of each $f_j(x)$, $j = 1, \ldots, m$, to vanish, namely that $\sum\limits_{i=1}^{n} k_{ji}\alpha_i + \sum\limits_{i=1}^{n} \beta_{ji}\lambda_i = 0$ for $j = 1, \ldots, m$.

Let $A = [\{k_{ji}\}_{i=1,j=1}^{n,m}, \{\beta_{ji}\}_{i=1,j=1}^{n,m}]$, which is an $m \times (2n)$ matrix and has $(\alpha_1, \ldots, \alpha_n, \lambda_1, \ldots, \lambda_n)^T$ as a null-vector. Let $\{v_1, \ldots, v_r\}$ be a linear basis for $\ker A$ and let for $t = 1, \ldots, r$, w_t be the vector in $\mathbb{R}^n$ which represents the last n entries of v_t. Since we want $\lambda_i > 0$, our problem reduces in finding a vector $\lambda = (\lambda_1, \ldots, \lambda_n)^T$ with $\lambda_i > 0$ of the form $\sum\limits_{t=1}^{r} x_t w_t$, which can be solved using a particular case of one of the algorithms described in Section 2. If the problem admits a solution, then we have

$$\Lambda'(x) = -\sum_{i=1}^{n} \epsilon_i \alpha_i x_i - \sum_{i=1}^{n} \epsilon_i \lambda_i x_i^2,$$

and $\Lambda'(x) < 0$ outside the hyperellipsoid of the equation $\Lambda'(x) = 0$. At points where $\Lambda'(x) < 0$, $\Lambda(x)$ decreases, thus the trajectory gets closer to the point $(-\frac{\alpha_1}{\lambda_1}, \ldots, -\frac{\alpha_n}{\lambda_n})$. Let $c \in \mathbb{R}$ be such that $\Omega = \{x \in \mathbb{R}^n : \Lambda(x) < c\}$ properly contains $\{x \in \mathbb{R}^n : \Lambda'(x) \geq 0\}$. Then Ω contains an attractor region for the dynamical system (4).

The best known example of a dynamical system of type (3) which admits an attractor region is the following one by Lorenz ([8]). This example triggered the research on attractor regions of the type considered in the present work, which are also called Lorenz attractors.

Example 1.

$$\begin{cases} x_1' = -10x_1 + 10x_2 \\ x_2' = -x_2 + 28x_1 - x_1x_3 \\ x_3' = -\frac{8}{3}x_3 + x_1x_2. \end{cases} \tag{7}$$

The simplest solution for which $\lambda_1, \lambda_2, \lambda_2 > 0$ is $\alpha_1 = \alpha_2 = 0$, $\alpha_3 = -38$, $\lambda_1 = \lambda_2 = \lambda_3 = 1$. So we can consider

$$\Lambda(x) = -38x_3 + \frac{1}{2}x_1^2 + \frac{1}{2}x_2^2 + \frac{1}{2}x_3^2,$$

for which

$$\Lambda'(x) = -10x_1^2 - x_2^2 - \frac{8}{3}(x_3 - 19)^2 + \frac{2888}{3}.$$

Then $\Lambda'(x) < 0$ outside an ellipsoid centered at $(0, 0, 19)$. Let $c \in \mathbb{R}$ be such that $\Omega = \{x \in \mathbb{R}^3 : \Lambda(x) < c\}$ properly contains $\{x \in \mathbb{R}^3 : \Lambda'(x) \geq 0\}$. Then Ω contains a Lorenz attractor for the dynamical system (7).

Let Φ_l, $l = 1, \ldots, M$, be monomials in the variables $x_1, \ldots, x_n$. Consider the dynamical system

$$x_i'(t) = -\epsilon_i x_i + \sum_{l=1}^{M} k_{li} \frac{\partial \Phi_l}{\partial x_i},$$

where $\epsilon_i > 0$ for $i = 1, \ldots, n$, and $\{k_{li}\}$ is an $M \times n$ real matrix. These systems are a slight generalization of the problem considered in [7]. We try to find a Lyapunov function of type $\Lambda(x) = \frac{1}{2} \sum\limits_{i=1}^{n} \lambda_i x_i^2$, $\lambda_i > 0$, $i = 1, \ldots, n$. Let $x_i \frac{\partial \Phi_l}{\partial x_i} = n_{li} \phi_l$ (n_{li} is the power of x_i in Φ_l), and then

$$\begin{aligned} \Lambda'(x) &= \sum_{i=1}^{n} \lambda_i x_i x_i' = -\sum_{i=1}^{n} \epsilon_i \lambda_i x_i^2 + \sum_{i=1}^{n} \sum_{l=1}^{M} \lambda_i k_{li} n_{li} \Phi_l(x) \\ &= -\sum_{i=1}^{n} \epsilon_i \lambda_i x_i^2 + \sum_{l=1}^{M} (\sum_{i=1}^{n} k_{li} n_{li} \lambda_i) \Phi_l(x). \end{aligned}$$

If $A = \{k_{li} n_{li}\}_{l=1,i=1}^{M,n}$, we can search for a null-vector $(\lambda_1, \ldots, \lambda_n)^T$, $\lambda_i > 0$, $i = 1, \ldots, n$, of A, using one of the algorithms mentioned in Section 2. For such a choice of λ_i, we have $\Lambda'(x) = -\sum\limits_{i=1}^{n} \epsilon_i \lambda_i x_i^2 < 0$, implying that 0 is an attractor trajectory for the system.

For a system of type (4), we consider next the existence of Lyapunov functions of type

$$\Lambda(x) = x^T A x = \sum_{i,j=1}^{n} a_{ij} x_i x_j,$$

where $A = \{a_{ij}\}_{i,j=1}^{n}$ is a positive definite real matrix. Then

$$\begin{aligned} \Lambda'(x) &= \sum_{i,j=1}^{n} a_{ij} x_i' x_j + \sum_{i,j=1}^{n} a_{ij} x_i x_j' \\ &= \sum_{i,j=1}^{n} a_{ij} x_j (-\epsilon_i x_i + \sum_{l=1}^{m} k_{li} f_l(x)) + \sum_{i,j=1}^{n} a_{ij} x_i (-\epsilon_j x_j + \sum_{l=1}^{m} k_{lj} f_l(x)). \end{aligned}$$

By our assumption we have that $x_i f_l(x) = \sum_{j=1}^{m} \alpha_j^{(il)} f_j(x)$, thus

$$\begin{aligned}\Lambda'(x) = &- \sum_{i,j=1}^{n} a_{ij}(\epsilon_i + \epsilon_j) x_i x_j + \sum_{i,j=1}^{n} a_{ij} \sum_{l=1}^{m} k_{li} \sum_{r=1}^{m} \alpha_r^{(jl)} f_r(x) \\ &+ \sum_{i,j=1}^{n} a_{ij} \sum_{l=1}^{m} k_{lj} \sum_{r=1}^{m} \alpha_r^{(il)} f_r(x) \\ = &- \sum_{i,j=1}^{n} a_{ij}(\epsilon_i + \epsilon_j) x_i x_j + \sum_{r=1}^{m} (\sum_{i,j=1}^{n} a_{ij} \sum_{l=1}^{m} k_{li} \alpha_r^{(jl)}) f_r(x) \\ &+ \sum_{r=1}^{m} (\sum_{i,j=1}^{n} a_{ij} \sum_{l=1}^{m} k_{lj} \alpha_r^{(il)}) f_r(x).\end{aligned}$$

Let us denote $\mu_{ij}^{(r)} = \sum_{l=1}^{m} k_{li} \alpha_r^{(jl)}$ and $E = \operatorname{diag}(\epsilon_1, \epsilon_2, \ldots, \epsilon_n)$. Then $AE = \{a_{ij}\epsilon_j\}_{i,j=1}^n$, $EA = \{\epsilon_i a_{ij}\}_{i,j=1}^n$. So

$$\Lambda'(x) = -x^T(AE + EA)x + \sum_{r=1}^{m} [\sum_{i,j=1}^{n} a_{ij}(\mu_{ij}^{(r)} + \mu_{ji}^{(r)})] f_r(x).$$

For $r = 1, \ldots, m$, let B_r denote the symmetric matrix $\{\mu_{ij}^{(r)} + \mu_{ji}^{(r)}\}_{i,j=1}^n$, so $\sum_{i,j=1}^{n} a_{ij}(\mu_{ij}^{(r)} + \mu_{ji}^{(r)}) = \operatorname{tr}(AB_r)$.

We try to find a matrix $A > 0$ such that $AE + EA > 0$ and $\operatorname{tr}(AB_r) = 0$ for $r = 1, \ldots, m$. This would then imply $\Lambda(x) > 0$ and $\Lambda'(x) = -x^T(AE + EA)x < 0$ for $x \neq 0$. As a consequence, 0 is an attractor trajectory for the system (4).

Consider the linear subspace $\mathcal{M} = \{X = X^T : \operatorname{tr}(AX) = 0\}$ and find a basis $C_1, C_2, \ldots, C_s$ for $\mathcal{M}$. Let

$$F_i = \begin{bmatrix} C_i & 0 \\ 0 & EC_i + C_i E \end{bmatrix},$$

$i = 1, \ldots, s$. Then our problem is equivalent to the existence of a positive definite matrix

$$\begin{bmatrix} A & 0 \\ 0 & B \end{bmatrix}$$

of the form $\sum_{i=1}^{s} x_i F_i$. If such a matrix exists, it is of the form

$$\begin{bmatrix} A & 0 \\ 0 & EA + AE \end{bmatrix},$$

where $A > 0$ and $EA + AE > 0$, and $A = \sum_{i=1}^{s} x_i C_i$, so $\operatorname{tr}(AB_r) = 0$ for $r = 1, \ldots, m$.

The existence of A can be determined by the algorithm in Section 2 to solve Problem 1. This represents a new way for finding in polynomial time a sufficient condition which guarantees 0 is an attractor trajectory for a system of type (4).

Example 2. Consider the dynamical system:

$$\begin{cases} x_1' = -x_1 + x_1x_2 + 6x_2x_3 + x_1x_3 \\ x_2' = -x_2 + x_1x_2 - 2x_2x_3 - 3x_1x_3 \\ x_3' = -x_3 - 3x_1x_2 - 2x_2x_3 + x_1x_3. \end{cases}$$

By considering a Lyapunov function of the type

$$\Lambda(x) = \lambda_1x_1^2 + \lambda_2x_2^2 + \lambda_3x_3^2 + \alpha_1x_1 + \alpha_2x_2 + \alpha_3x_3$$

we cannot cancel in $\Lambda'(x)$ all monomials of degree 3 for any $\lambda_1, \lambda_2, \lambda_3 > 0$.

Consider next $\Lambda(x) = x^TAx$, with $A > 0$. Since $E = I$, $A > 0$ automatically implies $AE + EA > 0$. The conditions that all monomials of degree three in $\Lambda'(x)$ cancel can be written as equations of the form $\mathrm{tr}(AB_r) = 0$. By the method in Section 2 we find

$$A = \begin{bmatrix} 2 & 1 & 1 \\ 1 & 2 & 1 \\ 1 & 1 & 2 \end{bmatrix} > 0,$$

for which $\Lambda'(x) = -2x^TAx$, implying 0 is an attractor trajectory for the dynamical system.

Example 3. Consider the dynamical system

$$\begin{cases} x' = -x - \frac{1}{2}y + (x - 2y)\sin x \\ y' = -\frac{1}{2}x - y + (2x - y)\sin x. \end{cases}$$

Let $B = \begin{bmatrix} -1 & -\frac{1}{2} \\ -\frac{1}{2} & -1 \end{bmatrix}$. We search for a Lyapunov function of the form $\Lambda(x, y) = a_{11}x^2 + 2a_{12}xy + a_{22}y^2$, with $A = \begin{bmatrix} a_{11} & a_{12} \\ a_{12} & a_{22} \end{bmatrix} > 0$. In the expression of $\Lambda'(x, y)$, we want all terms containing $\sin x$ to cancel, condition equivalent to $a_{11} + 2a_{12} = 0$ and $a_{22} + 2a_{12} = 0$. In this case, $\Lambda'(x, y) = x^T(AB + B^TA)x$, and the Lyapunov condition $AB + B^TA < 0$ is sufficient for the stability of the system. It is clear that $A = \begin{bmatrix} 2 & -1 \\ -1 & 2 \end{bmatrix}$, is a proper choice for A for implying that 0 is an attractor trajectory for the system.

The purpose of Example 3 is to illustrate the method; the matrix A could be easily determined. For similar systems in more variables, for finding the existence of a proper matrix A one needs the use of the algorithm in Section 2.

References

[1] M. Bakonyi and H.J. Woerdeman, Maximum entropy elements in the intersection of an affine space and the cone of positive definite matrices, *SIAM J. Matrix Anal. Appl.* **16** no. 2 (1995), 369–376.

[2] V. Balakrishnan and F. Wang, Semidefinite programming in systems and control theory, in Handbook of Semidefinite Programming, (H. Wolkowicz, R. Saigal and L. Vandenberghe, Editors), Kluwer Academic Publishers, Boston, 2000.

[3] S. Boyd and L. El-Ghaoui, Method of centers for minimizing generalized eigenvalues, *Linear Algebra Appl.* **188–190** (1993), 63–111.

[4] S. Boyd, L. El-Ghaoui, E. Feron, and V. Balakrisknan, *Linear Matrix Inequalities*, in System and Control Theory, SIAM, Philadelphia, 1994.

[5] R. Horn and C.R. Johnson, *Matrix Analysis*, Cambridge University Press, Cambridge, 1985.

[6] C. Jeffries, Qualitative stability of certain nonlinear systems, *Linear Algebra Appl.* **75** (1986), 133–144.

[7] C. Jeffries, C. Lee, and P. Van den Driessche, Hypergraphs, the qualitative solvability of $k \cdot \lambda = 0$, and Voltera mulitipliers for nonlinear dynamical systems, *J. Differential Equations* **105** (1993), 167–179.

[8] E. Lorenz, Deterministic nonperiodic flow, *J. Atmospheric Sci.* **29** (1963), 130–141.

[9] Yu. Nesterov and A.Nemirovsky, *Interior Point Polynomial Algorithms in Convex Programming*, SIAM, Philadelphia, 1994.

[10] A. Papachristodoulou and S. Prajna, On the construction of Lyapunov functions using the sum of squares decomposition, in: *Proceedings of the IEEE Conference on Decision and Control*, Las Vegas, NV, 2002, 1–6.

[11] A. Papachristodoulou and S. Prajna, A tutorial on sum of squares techniques for system analysis, in: *Proceedings of the American Control Conference*, Portland, Oregon, 2005.

[12] P.A. Parillo, *Structured Semidefinite Programming and Semialgebraic Geometry Methods in Robustness and Optimization*, Ph. D. Thesis, California Institute of Technology, Pasadena, CA, 2000.

[13] F. A. Potra, A quadratically convergent predictor-corrector method for solving linear programs from infeasible starting points, *Math Programming* **67** (1994), 383–406.

[14] L. Vandenberghe and S. Boyd, A polynomial-time algorithm for determining quadratic Lyapunov functions for nonlinear systems, in: *Proceedings of the European Conference on Circuit Theory and Design*, 1993, 1065–1068.

[15] L. Vandenberghe and S. Boyd, Semidefinite programming, *SIAM Review* **38** no. 1 (1996), 49–95.

[16] J.L.Willems, *Stability Theory of Dynamical Systems*, Nelson, London, 1970.

Mihály Bakonyi
Department of Mathematics and Statistics
Georgia State University
P.O. Box 4110
Atlanta, GA 30302-4110
USA
e-mail: mbakonyi@gsu.edu

Kazumi N. Stovall
Department of Management Science and Statistics
College of Business
One UTSA Circle
San Antonio, TX 78249-0632
USA
e-mail: kazumi.stovall@utsa.edu

Operator Theory:
Advances and Applications, Vol. 179, 35–46

Ranks of Hadamard Matrices and Equivalence of Sylvester–Hadamard and Pseudo-Noise Matrices

Tom Bella, Vadim Olshevsky and Lev Sakhnovich

Abstract. In this paper we obtain several results on the rank properties of Hadamard matrices (including Sylvester–Hadamard matrices) as well as generalized Hadamard matrices. These results are used to show that the classes of (generalized) Sylvester–Hadamard matrices and of (generalized) pseudo-noise matrices are equivalent, i.e., they can be obtained from each other by means of row/column permutations.

Mathematics Subject Classification (2000). Primary 15A57, 15A23; Secondary 05B15, 05B20 .

Keywords. Hadamard matrices, generalized Hadamard matrices, pseudo-random sequences, pseudo-noise sequences, pseudo-random matrices, pseudo-noise matrices, rank, equivalence.

1. Ranks of certain matrices related to classical Hadamard matrices

1.1. Hadamard and exponent Hadamard matrices

The classical $n \times n$ *Hadamard matrices* $H(2,n)$ are defined as those composed of ± 1's and satisfying

$$H(2,n)H(2,n)^T = nI_n, \tag{1.1}$$

that is, their distinct rows are orthogonal. Hadamard matrices are widely used in communication systems, data compression, error control coding, cryptography, linear filtering and spectral analysis, see, e.g., [5], [7], and the references therein. This popularity of Hadamard matrices is explained, among other reasons, by their simplicity and efficiency in a variety of concrete practical applications. For example, one simple way to construct a Hadamard matrix of the order $n = 2^m$ is due

to Sylvester. The method starts with defining a 1×1 matrix via $H(2,1) = 1$, and proceeds recursively:

$$H(2,2n) = \begin{bmatrix} H(2,n) & H(2,n) \\ H(2,n) & -H(2,n) \end{bmatrix}. \tag{1.2}$$

It is immediate to see that $H(2,2n)$ of (1.2) satisfies (1.1). Matrices generated in this fashion are referred to as *Sylvester–Hadamard* matrices. In addition to the Sylvester construction (1.2), there are alternate ways to construct Hadamard matrices, one of them is due to Paley, see, e.g., [5] and the references therein.

In many applications it is useful to consider matrices over $GF(2)$, so one typically changes -1's to 0's, e.g.,

$$H(2,2) = \begin{bmatrix} 1 & 1 \\ 1 & -1 \end{bmatrix} \longrightarrow \widetilde{H}(2,2) = \begin{bmatrix} 1 & 1 \\ 1 & 0 \end{bmatrix}, \tag{1.3}$$

or, alternatively, one replaces -1's by 1's and 1's by 0's, e.g.,

$$H(2,2) = \begin{bmatrix} 1 & 1 \\ 1 & -1 \end{bmatrix} \longrightarrow \widehat{H}(2,2) = \begin{bmatrix} 0 & 0 \\ 0 & 1 \end{bmatrix}. \tag{1.4}$$

We suggest to refer to matrices $\widehat{H}(2,2n)$ and $\widetilde{H}(2,2n)$ obtained in this fashion as *exponent* Hadamard matrices and *complimentary exponent* Hadamard matrices, respectively. (The justification for the above nomenclatures is in that using the entries of $\widehat{H}(2,2n)$ as exponents for -1 one obtains the entries of $H(2,n)$.)

In what follows we will adopt similar notations for any matrix A composed of ± 1's, and denote by $\widetilde{A}$ the matrix obtained from A by changing -1's to 0's, and denote by $\widehat{A}$ the matrix obtained from A by replacing -1's by 1's, and 1's by 0's.

1.2. General Hadamard matrices and ranks

In order to study the ranks of arbitrary Hadamard matrices we need to establish the following auxiliary result that applies to row/column scaled $H(2,n)$.

Lemma 1.1. *Let*

$$H(2,n) = \begin{bmatrix} 1 & e \\ e^T & H_{n-1} \end{bmatrix} \tag{1.5}$$

be a Hadamard matrix whose first column and top row contain only 1*'s, i.e.,*

$$e = \underbrace{\begin{bmatrix} 1 & \cdots & 1 \end{bmatrix}}_{n-1}.$$

Let $\widetilde{H}_{n-1}$ *denote the complimentary exponent matrix of* H_{n-1} *defined in Subsection* 1.1. *Then*

$$\widetilde{H}_{n-1}\widetilde{H}_{n-1}^T = tI_{n-1} + (t-1)J_{n-1}, \quad \text{where} \quad t = \frac{n}{4}, \quad \text{and} \quad J_{n-1} = ee^T. \tag{1.6}$$

Proof. It follows from (1.5) and the definition (1.1) that

$$e + eH_{n-1}^T = 0, \qquad J_{n-1} + H_{n-1}H_{n-1}^T = nI_{n-1}. \tag{1.7}$$

Consider an auxiliary matrix

$$B_{n-1} = H_{n-1} + J_{n-1}, \tag{1.8}$$

and observe, before proceeding further, that

$$\widetilde{H}_{n-1} = \frac{1}{2} B_{n-1}. \tag{1.9}$$

In view of (1.7) and (1.8) we have

$$\begin{aligned} B_{n-1}B_{n-1}^T &= H_{n-1}H_{n-1}^T + J_{n-1}H_{n-1}^T + H_{n-1}J_{n-1} + J_{n-1}J_{n-1} \\ &= (nI_{n-1} - J_{n-1}) - J_{n-1} - J_{n-1} + (n-1)J_{n-1} = nI_{n-1} + (n-4)J_{n-1}. \end{aligned} \tag{1.10}$$

Finally, (1.6) follows from (1.9) and (1.10). □

We are now ready to prove the following result.

Theorem 1.2. *Let us partition $H(2,n)$ by singling out its top row and first column:*

$$H(2,n) = \begin{bmatrix} h_{11} & r_1 \\ c_1 & H_{n-1} \end{bmatrix}.$$

Here h_{11} is a scalar, and H_{n-1} is an $(n-1)\times(n-1)$ submatrix of $H(2,n)$. Let $\widetilde{H}_{n-1}$ denote the complimentary exponent matrix of H_{n-1} defined in Section 1.1.

If 8 *divides n, then*

$$\operatorname{rank}\widetilde{H}_{n-1}(\operatorname{mod} 2) \le \frac{n}{2}. \tag{1.11}$$

If 8 *does not divide n, then*

$$\operatorname{rank}\widetilde{H}_{n-1}(\operatorname{mod} 2) = n-1. \tag{1.12}$$

Proof. Without loss of generality we may assume that $H(2,n)$ has the form shown in (1.5) and that the result in (1.6) holds. Let us consider two cases.

- If 8 divides n, then $t = \frac{n}{4}$ is even, and (1.6) implies

$$\widetilde{H}_{n-1}\widetilde{H}_{n-1}^T = J_{n-1}(\operatorname{mod} 2). \tag{1.13}$$

  If we denote by $k = \operatorname{rank}\widetilde{H}_{n-1}(\operatorname{mod} 2)$, then (1.13) implies

$$(n-1) - k \ge k - 1$$

  and (1.11) follows.
- If 8 does not divide n, then $t = \frac{n}{4}$ is odd, and (1.6) implies

$$\widetilde{H}_{n-1}\widetilde{H}_{n-1}^T = tI_{n-1}(\operatorname{mod} 2), \tag{1.14}$$

  so that (1.12) follows. □

1.3. Sylvester–Hadamard matrices, ranks and factorizations

In the previous subsection the result applied to arbitrary Hadamard matrices. Here we consider special Sylvester–Hadamard matrices. Here is the main result of this subsection.

Theorem 1.3. *Let $H(2,2^m)$ be a Sylvester–Hadamard matrix, i.e., one constructed via the recipe* (1.2)*. Then*

$$\operatorname{rank} \widehat{H}(2,2^m) = m(\operatorname{mod} 2), \tag{1.15}$$

where $\widehat{H}(2,2^m)$ denotes the exponent matrix of $H(2,2^m)$ defined in Section 1.1*.*

The result (1.15) follows from the following lemma.

Lemma 1.4. *The Sylvester–Hadamard matrix $H(2,2^m)$ admits the decomposition*

$$\widehat{H}(2,2^m) = L_m L_m^T (\operatorname{mod} 2), \tag{1.16}$$

where the rows of the $2^m \times m$ matrix

$$L_m = \begin{bmatrix} 0 & \cdots & 0 & 0 & 0 \\ 0 & \cdots & 0 & 0 & 1 \\ 0 & \cdots & 0 & 1 & 0 \\ 0 & \cdots & 0 & 1 & 1 \\ 0 & \cdots & 1 & 0 & 0 \\ \vdots & \cdots & \vdots & \vdots & \vdots \\ 1 & \cdots & 1 & 1 & 1 \end{bmatrix}$$

contain all possible binary m-tuples ordered naturally.

Proof. It is easy to see that for $m = 1$ we have

$$\widehat{H}(2,2) = \begin{bmatrix} 0 & 0 \\ 0 & 1 \end{bmatrix} = \begin{bmatrix} 0 \\ 1 \end{bmatrix} \begin{bmatrix} 0 & 1 \end{bmatrix}.$$

By applying an inductive argument we obtain

$$\widehat{H}(2,2^{m+1}) = \begin{bmatrix} \widehat{H}(2,2^m) & \widehat{H}(2,2^m) \\ \widehat{H}(2,2^m) & \widetilde{H}(2,2^m) \end{bmatrix} = \begin{bmatrix} \overrightarrow{0} & L_m \\ \overrightarrow{1} & L_m \end{bmatrix} \begin{bmatrix} \overrightarrow{0}^T & \overrightarrow{1}^T \\ L_m^T & L_m^T \end{bmatrix} \tag{1.17}$$

with

$$\overrightarrow{0} = \begin{bmatrix} 0 \\ 0 \\ \vdots \\ 0 \end{bmatrix}, \qquad \overrightarrow{1} = \begin{bmatrix} 1 \\ 1 \\ \vdots \\ 1 \end{bmatrix}.$$

The relations (1.17) and (1.16) coincide which completes the proof of the lemma. □

2. Pseudo-noise matrices

2.1. Linear recurrence relations and shift registers

Let m be a fixed positive integer, and $h_0, h_1, \ldots, h_{m-1} \in GF(2)$. Consider a linear m-term recurrence relation

$$a_i = a_{i-1}h_{m-1} + a_{i-2}h_{m-2} + \cdots + a_{i-m+1}h_1 + a_{i-m}h_0 \quad \text{for } i \geq m \tag{2.1}$$

over $GF(2)$. Observe that the above recurrence relation can be written in a matrix form:

$$\begin{bmatrix} a_{i-(m+1)} \\ a_{i-(m+2)} \\ a_{i-(m+3)} \\ \vdots \\ a_{i-1} \\ a_i \end{bmatrix} = \begin{bmatrix} 0 & 1 & 0 & 0 & \ldots & 0 \\ 0 & 0 & 1 & 0 & \ldots & 0 \\ 0 & 0 & 0 & 1 & \ddots & \vdots \\ \vdots & \vdots & \vdots & \ddots & \ddots & 0 \\ 0 & 0 & 0 & \ldots & 0 & 1 \\ h_0 & h_1 & h_2 & \ldots & h_{m-2} & h_{m-1} \end{bmatrix} \begin{bmatrix} a_{i-(m)} \\ a_{i-(m+1)} \\ a_{i-(m+2)} \\ \vdots \\ a_{i-2} \\ a_{i-1} \end{bmatrix}. \tag{2.2}$$

The m-tuple

$$\{a_{i-m}, \ldots, a_{i-2}, a_{i-1}, \}$$

is called the *state vector* corresponding to the time moment $i-m$. The semi-infinite sequence

$$a_0, a_1, a_2, a_3, a_4, a_5, \ldots \tag{2.3}$$

is called an *(m^{th} order) linear recurring sequence* corresponding to (2.1). Clearly, the latter is fully determined by the *initial state vector* $\{a_0, a_1, \ldots, a_{m-2}, a_{m-1}\}$ and the coefficients $h_0, h_1, \ldots, h_{m-1}$ of (2.1).

In order to define the concept of a pseudo-noise sequence it is useful to associate (2.1) with a *shift register*. As an example, consider a special case of (2.1), a 4-term linear recurrence relation with $h_0 = 1, h_1 = 0, h_2 = 0, h_3 = 1$, and visualize

$$a_i = a_{i-1} + a_{i-4} \tag{2.4}$$

with the help of the following figure:

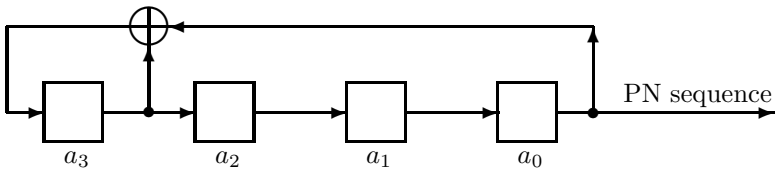

FIGURE 1. Shift register for (2.4). Time moment "zero".

The above figure corresponds to the time moment "zero", i.e., it is characterized by the initial state vector

$$\{a_0, a_1, a_2, a_3\}. \tag{2.5}$$

The next figure corresponds to the time moment "one",

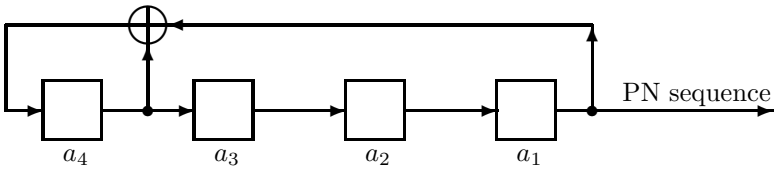

FIGURE 2. Shift register for (2.4). Time moment "one".

and its state vector

$$\{a_1, a_2, a_3, a_4\} \tag{2.6}$$

is obtained from the one in (2.5) by shifting entries to the left (a_0 disappears), and computing a_4 via (2.4). Figures 1 and 2 graphically express both the shift and computing a_4. That is, a_4 of Figure 2 is computed as $a_3 + a_0$ of Figure 1.

2.2. Pseudo-noise sequences and matrices

Recall that the semi-infinite sequence in (2.3) is fully determined by the *initial state vector* $\{a_0, a_1, \ldots, a_{m-2}, a_{m-1}\}$ and the coefficients $h_0, h_1, \ldots, h_{m-1}$ of (2.1). Indeed, the rule (2.1) maps the state vectors to the next ones, i.e.,

$$\begin{bmatrix} a_{m-1} \\ a_{m-2} \\ \vdots \\ a_2 \\ a_1 \\ a_0 \end{bmatrix} \longrightarrow \begin{bmatrix} a_m \\ a_{m-1} \\ \vdots \\ a_3 \\ a_2 \\ a_1 \end{bmatrix} \longrightarrow \begin{bmatrix} a_{m+1} \\ a_m \\ \vdots \\ a_4 \\ a_3 \\ a_2 \end{bmatrix} \longrightarrow$$

Since all the elements $\{a_i\}$ belong to $GF(2)$, none of the recurrence relations of the form (2.1) can generate more than $2^m - 1$ different state vectors (we exclude the trivial zero initial state vector). It follows that the sequence in (2.3) has to be periodic with the period not exceeding $2^m - 1$ (of course, there are coefficients $h_0, h_1, \ldots, h_{m-1}$ of (2.1) for which any sequence will have a period smaller that $2^m - 1$).

If the sequence (2.3) has the maximal possible period $2^m - 1$, then it is called a *pseudo-noise* sequence, see, e.g., [7]. Pseudo-noise sequences are useful in a number of applications, see, e.g., [4], [3].

A *pseudo-noise* matrix $T(2,n)$ with $n = 2^m$ is defined (see, e.g., [7]) as an $n \times n$ matrix of the form

$$T(2,n) = \begin{bmatrix} 0 & 0 & \dots & 0 \\ 0 & & & \\ \vdots & & \check{T} & \\ 0 & & & \end{bmatrix} \tag{2.7}$$

where $\check{T}$ is a circulant Hankel matrix whose top row is a $1 \times (2^m - 1)$ array that is a period of the pseudo-noise sequence, and whose first m entries coincide with the initial state.

Example. Let us again consider the 4-term recurrent relations (2.4) with with $h_0 = 1, h_1 = 0, h_2 = 0, h_3 = 1$ and the initial state

$$\begin{bmatrix} a_0 & a_1 & a_2 & a_3 \end{bmatrix} = \begin{bmatrix} 1 & 0 & 0 & 0 \end{bmatrix}.$$

This choice gives rise to the pseudo-noise sequence

$$\underbrace{\mathbf{1000}11110101100}_{\text{period } 15 = 2^3 - 1}\ \underbrace{\mathbf{1000}11110101100}_{\text{period } 15}\underbrace{\mathbf{1000}11110101100}_{\text{period } 15}\dots\dots$$

and the 15×15 matrix $\check{T}$ of (2.7) is given by

$$\check{T} = \begin{bmatrix} 100011110101100 \\ 000111101011001 \\ 001111010110010 \\ 011110101100100 \\ 111101011001000 \\ 111010110010001 \\ 110101100100011 \\ 101011001000111 \\ 010110010001111 \\ 101100100011110 \\ 011001000111101 \\ 110010001111010 \\ 100100011110101 \\ 001000111101011 \\ 010001111010110 \end{bmatrix}.$$

2.3. Equivalence of pseudo-noise and Sylvester–Hadamard exponent matrices

In order to establish the equivalence of the two classes of matrices we will need the following counterpart of Theorem 1.3.

Lemma 2.1. *For $n = 2^m$, the rank of any $n \times n$ pseudo-noise matrix T is m.*

Proof. This follows from the immediate observation that the rows of the matrix

$$\check{T} = \begin{bmatrix} \mathbf{a}_0 \\ \mathbf{a}_1 \\ \vdots \\ \mathbf{a}_{2^m-2} \end{bmatrix}$$

satisfy the m-term recurrence relations

$$\mathbf{a}_i = \mathbf{a}_{i-1}h_{m-1} + \mathbf{a}_{i-2}h_{m-2} + \cdots + \mathbf{a}_{i-m+1}h_1 + \mathbf{a}_{i-m}h_0 \tag{2.8}$$

(of the form (2.1)), and hence they are linearly dependent. □

The following theorem implies that the Sylvester–Hadamard matrices and Pseudo-noise matrices are equivalent, i.e., they can be obtained from each other via row/column permutations.

Theorem 2.2. *Let $H(2,2^m)$ be the Sylvester–Hadamard matrix, and let $T(2,2^m)$ be a $2^m \times 2^m$ pseudo-noise matrix. Then $H(2,2^m)$ is equivalent to $T(2,2^m)$; i.e., there exist permutation matrices P_1 and P_2 such that $H(2,2^m) = P_1T(2,2^m)P_2$.*

Proof. Recall that the matrix $H(2,2^m)$ admits a factorization (1.16) into the product of a $2^m \times m$ matrix L_m and a $m \times 2^m$ matrix L_m^T, each of which contains all possible binary m-tuples as rows/columns.

Secondly, by Lemma 2.1, $T(2,2^m)$ also has a similar factorization $T = MR$ where M is a $2^m \times m$ matrix, and R is an $m \times 2^m$ matrix. Further, the rows of M are all distinct and hence they must contain all possible binary m-tuples, and the same is true of the columns of R.

Hence these factorizations of $H(2,2^m)$ and of $T(2,2^m)$ differ only by the order in which the rows/columns appear, and this completes the proof. □

Theorem 2.2 was numerically checked to be valid for $n = 8$ and $n = 16$ in [6].

3. Generalized Hadamard matrices, ranks and factorizations

An $n \times n$ matrix $H(q,n)$ is called a generalized Hadamard matrix [1] if its elements coincide with one of the numbers

$$\epsilon^k = \exp\left(\frac{2\pi i}{q}k\right), \quad 0 \le k \le q-1, \tag{3.1}$$

and it satisfies

$$H(q,n)H(q,n)^* = nI_n, \tag{3.2}$$

where $*$ denotes the complex conjugate transposed. Clearly, in the case $q = 2$, the generalized Hadamard matrices $H(2,n)$ reduce to the classic Hadamard matrices $H(2,n)$.

Often we will be concerned with a matrix that contains not the entries ϵ^k of $H(q,n)$, but the values k from (3.1) corresponding to each entry. Specifically (as

in Section 1.1), for a generalized Hadamard matrix $H(q,n) = [h_{ij}]$ we define its *exponent generalized Hadamard* matrix $\widehat{H}(q,n) = [\widehat{h}_{ij}]$ such that $h_{ij} = \epsilon^{\widehat{h}_{ij}}$.

We will call a matrix $H(q,n)$ normalized if all the elements of the first row and first column are $\epsilon^0 = 1$. Without loss of generality, we will assume that all matrices henceforth are normalized.

The Sylvester method can be generalized to the generalized Hadamard matrices with a FFT-like construction:

Proposition 3.1. *Let $H(q,n)$ be a generalized Hadamard matrix. Then the matrix given by*

$$H(q,qn) = \begin{bmatrix} H(q,n) & H(q,n) & H(q,n) & \dots & H(q,n) \\ H(q,n) & \epsilon H(q,n) & \epsilon^2 H(q,n) & \dots & \epsilon^{q-1} H(q,n) \\ H(q,n) & \epsilon^2 H(q,n) & \epsilon^4 H(q,n) & \dots & \epsilon^{2(q-1)} H(q,n) \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ H(q,n) & \epsilon^{q-1} H(q,n) & \epsilon^{2(q-1)} H(q,n) & \dots & \epsilon^{(q-1)^2} H(q,n) \end{bmatrix} \tag{3.3}$$

is also a generalized Hadamard matrix.

As with the Sylvester method for classical Hadamard matrices, the previous proposition as well as the initial generalized Hadamard matrix (which is just the DFT matrix)

$$H(q,q) = \begin{bmatrix} 1 & 1 & 1 & 1 & \dots & 1 \\ 1 & \epsilon & \epsilon^2 & \epsilon^3 & \dots & \epsilon^{q-1} \\ 1 & \epsilon^2 & \epsilon^4 & \epsilon^6 & \dots & \epsilon^{2(q-1)} \\ 1 & \epsilon^3 & \epsilon^6 & \epsilon^9 & \dots & \epsilon^{3(q-1)} \\ \vdots & \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & \epsilon^{q-1} & \epsilon^{2(q-1)} & \epsilon^{2(q-1)} & \dots & \epsilon^{(q-1)^2} \end{bmatrix} \tag{3.4}$$

allows one to construct special generalized Hadamard matrices $H(q,q^m)$, which we will call generalized Sylvester–Hadamard matrices.

The following result is a generalization of Theorem 1.3.

Theorem 3.2. *Let $H(q,q^m)$ be a generalized Sylvester–Hadamard matrix. The rank of its exponent matrix $\widehat{H}(q,q^m)$ is m*(mod q).

This theorem follows from the following factorization result.

Theorem 3.3. *Let $\widehat{H}(q,q^m)$ be the exponent matrix corresponding to the generalized Sylvester–Hadamard matrix $H(q,q^m)$. Then $\widehat{H}(q,q^m)$ admits the decomposition*

$$\widehat{H}(q,q^m) = L_m L_m^T \pmod q \tag{3.5}$$

where L_m is a $q^m \times q$ matrix with elements from $\{0,1,\dots,q-1\}$. Further, the rows of L_m contain all possible q-ary m-tuples ordered naturally.

Proof. The proof is by induction on m. Letting $m = 1$, we have

$$\widehat{H}(q, q) = L_1 L_1^T \pmod{q} \tag{3.6}$$

with

$$L_1^T = \begin{bmatrix} 0 & 1 & 2 & \dots & q-1 \end{bmatrix} \tag{3.7}$$

which by construction contains all q-ary numbers as columns.

Proceeding inductively we see that

$$\widehat{H}(q, q^m) = L_m L_m^T (\mathrm{mod}\ q)$$

implies

$$\widehat{H}(q, q^{m+1}) = \begin{bmatrix} 0_m & L_m \\ 1_m & L_m \\ \vdots & \vdots \\ (q-1)_m & L_m \end{bmatrix} \begin{bmatrix} 0_m^T & 1_m^T & \dots & (q-1)_m^T \\ L_m^T & L_m^T & \dots & L_m^T \end{bmatrix} \tag{3.8}$$

modulo q, where we note that

$$r_m^T = \begin{bmatrix} r & r & \dots & r \end{bmatrix} \tag{3.9}$$

which are of size $1 \times m$.

The fact that L_{m+1} contains all possible q-ary $(m+1)$-tuples as columns is clear from the fact that all possible q-ary m-tuples are present in the columns of L_m by hypothesis, and L_m appears once beneath each of $0_m, 1_m, \dots, (q-1)_m$. This completes the proof. □

4. Generalized pseudo-noise matrices

In this section we generalize the results of Section 2 from $GF(2)$ to $GF(q)$.

Again, for a positive integer m, and $h_0, h_1, \dots, h_{m-1} \in GF(q)$ we define an (m^{th} order) linear recurring sequence

$$a_0, a_1, a_2, \dots$$

via

$$a_i = a_{i-1}h_{m-1} + a_{i-2}h_{m-2} + \cdots + a_{i-m+1}h_1 + a_{i-m}h_0 \quad \text{for } i \geq m. \tag{4.1}$$

As in Section 2, it is easy to see that every m^{th} order linear recurring sequence is periodic with period at least $r \leq q^m - 1$. A *pseudo-noise sequence* is an m^{th} order linear recurring sequence with the maximal possible period $q^m - 1$. Furthermore, a *pseudo-noise* matrix $T(q, q^m)$ is an $q^m \times q^m$ matrix of the form

$$T(q, q^m) = \begin{bmatrix} 0 & 0 & \dots & 0 \\ 0 & & & \\ \vdots & & \check{T} & \\ 0 & & & \end{bmatrix} \tag{4.2}$$

where $\check{T}$ is a circulant Hankel matrix with top row a pseudo-noise sequence

$$\left[\begin{array}{cccc} a_0 & a_1 & \cdots & a_{2^m-2} \end{array}\right].$$

The following result is a generalization of Lemma 2.1.

Lemma 4.1. *The rank of any $q^m \times q^m$ pseudo-noise matrix $T(q, q^m)$ is m.*

Proof. This follows from the immediate observation that the rows of the matrix

$$\check{T} = \left[\begin{array}{c} \mathbf{a}_0 \\ \mathbf{a}_1 \\ \vdots \\ \mathbf{a}_{2^m-2} \end{array}\right]$$

satisfy the m-term recurrence relations

$$\mathbf{a}_i = \mathbf{a}_{i-1}h_{m-1} + \mathbf{a}_{i-2}h_{m-2} + \cdots + \mathbf{a}_{i-m+1}h_1 + \mathbf{a}_{i-m}h_0$$

(of the form (4.1)), and hence they are linearly dependent. □

The above lemma implies the following result.

Theorem 4.2. *The $q^m \times q^m$ pseudo-noise matrix $T(q, q^m)$ admits the decomposition*

$$T(q, q^m) = MR, \tag{4.3}$$

where M is a $q^m \times m$ matrix, and R is an $m \times q^m$ matrix. Further, the rows of M are all distinct and contain all possible q-ary m-tuples, and the same is true of the columns of R.

Proof. The factorization (4.3) exists by Lemma 4.1. By the definition, the rows of $T(q, q^m)$ are distinct, and therefore so are the rows of M. Since M is over $GF(q)$ with size $q^m \times m$, we conclude that M must contain all possible q-ary m-tuples as rows.

Similarly, the columns of R are also distinct, and since R is also over $GF(q)$, we have that R contains all possible m-tuples as columns, which completes the proof. □

The following theorem implies that the generalized Sylvester–Hadamard matrices and generalized pseudo-noise matrices are equivalent, i.e., they can be obtained from each other via row/column permutations.

Theorem 4.3. *Let $H(q, q^m)$ be a $q^m \times q^m$ generalized Sylvester–Hadamard matrix, and let $T(q, q^m)$ be a $q^m \times q^m$ pseudo-noise matrix where q is prime. Then the exponent matrix $\widehat{H}(q, q^m)$ is equivalent to $T(q, q^m)$; i.e., there exist permutation matrices P_1 and P_2 such that $\widehat{H}(q, q^m) = P_1 T(q, q^m) P_2$.*

Proof. By Theorem 3.3, the exponent matrix $\widehat{H}(q, q^m)$ has a factorization into the product of a $q^m \times m$ matrix and an $m \times q^m$ matrix, each of which contains all possible q-ary m-tuples as rows/columns. By Theorem 4.2, $T(q, q^m)$ also has a factorization into the product of a $q^m \times m$ matrix and a $m \times q^m$ matrix which

contain all possible q-ary m-tuples as rows/columns. Thus the factorizations differ only by the order in which the rows/columns appear, and this completes the proof. □

Theorem 4.3 was announced in [2].

5. Conclusion

Several results for the ranks of generalized Hadamard matrices, Sylvester–Hadamard matrices, their exponent matrices and their generalizations were established. These rank properties were used to demonstrate that the two classes of matrices, those built from generalized pseudo-noise sequences, and the generalized Hadamard matrices are equivalent up to permutations of the rows and columns.

References

[1] A.T. Butson, *Generalized Hadamard matrices*, Proc. Am. Math. Soc. **13** (1962), 894–898.

[2] T. Bella, V. Olshevsky and L. Sakhnovich, *Equivalence of Hadamard matrices and pseudo-noise matrices*. In: Advanced Signal Processing Algorithms, Architectures, and Implementations XV. Franklin T. Luk (ed.), SPIE Publications, 2005, p. 265–271.

[3] J.H. Conway and N.J.A. Sloane, *Sphere Packings, Lattices and Groups*, Springer-Verlag, 1991.

[4] P. Hoeher and F. Tufvesson, *Channel Estimation with Superimposed Pilot Sequence Applied to Multi-Carrier Systems*, Proc. Advanced Signal Processing for Communications Symposium, 1999.

[5] K.J. Horadam, *Hadamard Matrices and Their Applications*, Princeton University Press, 2006.

[6] P. Mukhin, L. Sakhnovich and V. Timofeev, *About Equivalence of Hadamard's Matrices*, Praći UNDIRT **1** no. 17 (1999), 89–94. (In Russian.)

[7] N.J.A. Sloane and F.J. MacWilliams, *The Theory of Error-Correcting Codes*, North-Holland, 1977.

Tom Bella and Vadim Olshevsky
Department of Mathematics, University of Connecticut
Storrs CT 06269-3009
USA
e-mail: bella@math.uconn.edu
olshevsky@math.uconn.edu

Lev Sakhnovich
735 Crawford Avenue
Brooklyn, NY 11223
USA
e-mail: Lev.Sakhnovich@verizon.net

Operator Theory:
Advances and Applications, Vol. 179, 47–62

Image of a Jacobi Field

Yurij M. Berezansky and Artem D. Pulemyotov

Abstract. Consider the two Hilbert spaces H_- and T_-. Let $K^+ : H_- \to T_-$ be a bounded operator. Consider a measure ρ on H_-. Denote by ρ_K the image of the measure ρ under K^+. This paper aims to study the measure ρ_K assuming ρ to be the spectral measure of a Jacobi field. We present a family of operators whose spectral measure equals ρ_K. We state an analogue of the Wiener-Itô decomposition for ρ_K. Finally, we illustrate our constructions by offering a few examples and exploring a relatively transparent special case.

Mathematics Subject Classification (2000). 28C20, 60G20, 60H40, 47B36.

Keywords. Image measure; Jacobi field; spectral measure; Wiener-Itô decomposition; Lévy noise measure.

1. Introduction

Consider a real separable Hilbert space H and a rigging

$$H_- \supset H \supset H_+$$

with the pairing $\langle \cdot, \cdot \rangle_H$. We assume the embedding $H_+ \hookrightarrow H$ to be a Hilbert-Schmidt operator. Consider another real separable Hilbert space T and a rigging

$$T_- \supset T \supset T_+$$

with the pairing $\langle \cdot, \cdot \rangle_T$. Given a bounded operator $K : T_+ \to H_+$, define the operator $K^+ : H_- \to T_-$ via the formula

$$\langle K^+ \xi, f \rangle_T = \langle \xi, Kf \rangle_H, \qquad \xi \in H_-,\ f \in T_+.$$

Let ρ be a Borel probability measure on the space H_-. We denote by ρ_K the image of the measure ρ under the mapping K^+. This paper aims to study the measure ρ_K assuming ρ to be the spectral measure of a Jacobi field $J = (\tilde{J}(\phi))_{\phi \in H_+}$. We base ourselves upon the papers [9] and [13] dedicated to the same problem.

By definition, a Jacobi field $J = (\tilde{J}(\phi))_{\phi \in H_+}$ is a family of commuting self-adjoint three-diagonal operators $\tilde{J}(\phi)$ acting in the Fock space

$$\mathcal{F}(H) = \bigoplus_{n=0}^{\infty} \mathcal{F}_n(H), \qquad \mathcal{F}_n(H) = H_{\mathbb{C}}^{\hat{\otimes} n}$$

(we suppose $H_{\mathbb{C}}^{\hat{\otimes} 0} = \mathbb{C}$). The operators $\tilde{J}(\phi)$ are assumed to depend on the indexing parameter $\phi \in H_+$ linearly and continuously. In Section 2 of the present paper, we adduce the definition and the basic spectral theory of a Jacobi field. More details can be found in, e.g., [1], [2], [3], and [4]. Remark that the concept of a Jacobi field is relatively new, therefore the definitions given in different papers may differ in minor details.

Jacobi fields are actively used in non-Gaussian white noise analysis and theory of stochastic processes, see [7], [22], [2], [4], [5], [19], [11], [8], [12], [23], [24], [25], and [28]. In the case of a finite-dimensional H, the theory of Jacobi fields is closely related to some results in [15], [16], and [14].

The most principal examples of spectral measures of Jacobi fields are the Gaussian measure and the Poisson measure. The Jacobi field with the Gaussian spectral measure is the classical free field in quantum field theory, see, e.g., [6], [7], [22], [2], and [3]. The Jacobi field with the Poisson spectral measure is the so-called Poisson field, see, e.g., [22], [2], [3], and [5]. *De facto*, it has been independently discovered in [17] and [30]. Section 2 of the present paper contains the rigorous definitions of the classical free field and the Poisson field.

For other examples of spectral measures of Jacobi fields, see [2] and [4].

In Section 3 of the present paper, for a given operator K and a given Jacobi field J, we construct a Fock-type space

$$\mathcal{F}^{\mathrm{ext}}(T_+, K) = \bigoplus_{n=0}^{\infty} \mathcal{F}_n^{\mathrm{ext}}(T_+, K)$$

and a family $J_K = (\tilde{J}_K(f))_{f \in T_+}$ of operators in $\mathcal{F}^{\mathrm{ext}}(T_+, K)$ pursuing the three following goals:

- To show that ρ_K is the spectral measure of the family J_K.
- To show that the Fourier transform corresponding to the generalized joint eigenvector expansion of J_K coincides with the generalized Wiener-Itô-Segal transform associated with ρ_K.
- To obtain an analogue of the Wiener-Itô orthogonal decomposition for ρ_K employing the generalized Wiener-Itô-Segal transform associated with ρ_K.

A detailed description of the classical concept of the Wiener-Itô decomposition can be found in, e.g., [6] or [18]. Once again we emphasize that the space $\mathcal{F}^{\mathrm{ext}}(T_+, K)$ depends on the Jacobi field J as well as on T_+ and K. We use the shorter notation $\mathcal{F}^{\mathrm{ext}}(T_+, K)$ instead of $\mathcal{F}^{\mathrm{ext}}(T_+, K, J)$ for the sake of simplicity.

The Wiener-Itô-Segal transform I corresponding to the Gaussian measure γ on H_- is a unitary operator acting from $\mathcal{F}(H)$ to $L^2(H_-, d\gamma)$. It takes the

orthogonal sum $\bigoplus_{j=0}^{n} \mathcal{F}_j(H)$ to the set $\tilde{\mathcal{P}}_n$ of all ordinary polynomials on H_- with their degree less than or equal to n. The unitarity of I implies

$$L^2(H_-, d\gamma) = \bigoplus_{n=0}^{\infty} \left(\tilde{\mathcal{P}}_n \ominus \tilde{\mathcal{P}}_{n-1} \right) = \bigoplus_{n=0}^{\infty} I(\mathcal{F}_n(H))$$

(we suppose $\tilde{\mathcal{P}}_{-1} = \{0\}$). This formula constitutes the Wiener-Itô orthogonal decomposition for the Gaussian measure. It is a powerful technical tool for carrying out the calculations in the space $L^2(H_-, d\gamma)$. Remark that analogous results are possible to obtain considering the Poisson measure instead of the Gaussian measure γ.

The Fourier transform corresponding to the generalized joint eigenvector expansion of the classical free field coincides with the Wiener-Itô-Segal transform I. An analogous result is possible to obtain for the Poisson field. These facts give the basis for investigating the concept of the Wiener-Itô decomposition from the viewpoint of spectral theory of Jacobi fields.

The operator I can be represented as a sum of operators of multiple stochastic integration. An analogous result is possible to obtain considering the Poisson measure instead of the Gaussian measure γ.

Our generalization of the classical picture is as follows. One may introduce the generalized Wiener-Itô-Segal transform I_K associated with the measure ρ_K as an operator between $\mathcal{F}(T_+)$ and $L^2(T_-, d\rho_K)$. It takes $\bigoplus_{j=0}^{n} \mathcal{F}_j(T_+)$ to the set $\mathcal{Q}_n$ of all continuous polynomials on T_- with their degree less than or equal to n. We construct the space $\mathcal{F}^{\mathrm{ext}}(T_+, K)$ so that I_K could be extended to a unitary operator acting from $\mathcal{F}^{\mathrm{ext}}(T_+, K)$ to $L^2(T_-, d\rho_K)$. The orthogonal component $\mathcal{F}_n^{\mathrm{ext}}(T_+, K)$ has to be defined as the completion of $\mathcal{F}_n(T_+)$ with respect to a new scalar product $(\cdot, \cdot)_{\mathcal{F}_n^{\mathrm{ext}}(T_+, K)}$. Basically, the problem of constructing the space $\mathcal{F}^{\mathrm{ext}}(T_+, K)$ consists in identifying this scalar product explicitly.

The unitarity of I_K implies

$$L^2(T_-, d\rho_K) = \bigoplus_{n=0}^{\infty} \left(\tilde{\mathcal{Q}}_n \ominus \tilde{\mathcal{Q}}_{n-1} \right) = \bigoplus_{n=0}^{\infty} I_K(\mathcal{F}_n^{\mathrm{ext}}(T_+, K))$$

(the notation $\tilde{\mathcal{Q}}_n$ stands for the closure of $\mathcal{Q}_n$ and we suppose $\tilde{\mathcal{Q}}_{-1} = \{0\}$). This formula constitutes an analogue of the Wiener-Itô orthogonal decomposition for the measure ρ_K. It discovers the Fock-type structure of the space $L^2(T_-, d\rho_K)$ and enables one to carry out the calculations in $L^2(T_-, d\rho_K)$.

As mentioned above, the Wiener-Itô-Segal transform I can be represented as a sum of operators of multiple stochastic integration. Presumably, an analogous representation is possible to obtain for the generalized Wiener-Itô-Segal transform I_K. However, we do not concern ourselves with this problem in the present paper.

We illustrate our abstract constructions with a few concrete examples. Among others, we consider the case where K is the operator of multiplication by a function of a new independent variable and J is the Poisson field. Then ρ_K appears to

be a Lévy noise measure. Theorem 3.1 and Theorem 3.2 of the present paper explain the Fock-type structure of $L^2(T_-, d\rho_K)$ in this case. Theorem 3.2 shows that $\mathcal{F}^{\mathrm{ext}}(T_+, K)$ is similar to the extended Fock space investigated in [23]. A special form of this space has been introduced in [20] in the framework of Gamma white noise analysis. Its further study has been carried out in [10], [19], [11], and [12], see also [24] and [28].

A family of operators with a Lévy noise spectral measure has been constructed in [8], see also [23]. The case of the Gamma measure was studied in [19]. An analogue of the Wiener-Itô decomposition for a Lévy noise measure has been obtained in [23], see also [21], [26], [29], and [31]. The case of the Gamma measure was studied in [20], [11], and [12]. Remark that the works [26], [29], [23], and [31] (respectively, [11] and [12]) represent the generalized Wiener-Itô-Segal transform associated with a Lévy noise measure (respectively, the Gamma measure) as a sum of operators of stochastic integration. Once again we emphasize that the present paper does not attempt to obtain an analogous representation for the generalized Wiener-Itô-Segal transform associated with the measure ρ_K in the general case.

Our abstract considerations become much more transparent when the range of the operator K is dense in H_+. We explore this situation in Section 4 of the present paper providing the reduction of the general construction along with three examples. In particular, we study a Gaussian measure with a non-trivial correlation operator. Relevant results can be found in [9].

Remark that the riggings we consider in this paper are all quasinuclear. One may consider nuclear riggings instead.

2. Commutative Jacobi Fields

This section contains the definition, the basic spectral theory, and two examples of Jacobi fields.

Let H be a real separable Hilbert space. Denote by $H_{\mathbb{C}}$ the complexification of H. Let $\hat{\otimes}$ stand for the symmetric tensor product. Consider the symmetric Fock space

$$\mathcal{F}(H) = \bigoplus_{n=0}^{\infty} \mathcal{F}_n(H), \qquad \mathcal{F}_n(H) = H_{\mathbb{C}}^{\hat{\otimes} n}$$

(we suppose $H_{\mathbb{C}}^{\hat{\otimes} 0} = \mathbb{C}$). This space consists of the sequences $\Phi = (\Phi_n)_{n=0}^{\infty}$, $\Phi_n \in \mathcal{F}_n(H)$. In what follows, we identify $\Phi_n \in \mathcal{F}_n(H)$ with $(0, \dots, 0, \Phi_n, 0, 0, \dots) \in \mathcal{F}(H)$ (Φ_n standing at the nth position).

The finite vectors $\Phi = (\Phi_1, \dots, \Phi_n, 0, 0, \dots) \in \mathcal{F}(H)$ form a linear topological space $\mathcal{F}_{\mathrm{fin}}(H) \subset \mathcal{F}(H)$. The convergence in $\mathcal{F}_{\mathrm{fin}}(H)$ is equivalent to the uniform finiteness and coordinatewise convergence. The vector $\Omega = (1, 0, 0, \dots) \in \mathcal{F}_{\mathrm{fin}}(H)$ is called vacuum.

Let

$$H_- \supset H \supset H_+ \tag{2.1}$$

be a rigging of H with real separable Hilbert spaces H_+ and $H_- = (H_+)'$ (hereafter, X' denotes the dual of the space X). We suppose the embedding $H_+ \hookrightarrow H$ to be a Hilbert-Schmidt operator. The pairing in (2.1) can be extended naturally to a pairing between $\mathcal{F}_n(H_+)$ and $\mathcal{F}_n(H_-)$. The latter can be extended to a pairing between $\mathcal{F}_{\mathrm{fin}}(H_+)$ and $(\mathcal{F}_{\mathrm{fin}}(H_+))'$. In what follows, we use the notation $\langle\cdot,\cdot\rangle_H$ for all of these pairings. Note that $(\mathcal{F}_{\mathrm{fin}}(H_+))'$ coincides with the direct product of the spaces $\mathcal{F}_n(H_-)$, $n \in \mathbb{Z}_+$.

Throughout the paper, $\mathrm{Pr}_X F$ denotes the projection of a vector F onto a subspace X.

2.1. The definition and the spectral theory of a Jacobi field

In the Fock space $\mathcal{F}(H)$, consider a family $\mathcal{J} = (\mathcal{J}(\phi))_{\phi\in H_+}$ of operator-valued Jacobi matrices

$$\mathcal{J}(\phi) = \begin{pmatrix} b_0(\phi) & a_0^*(\phi) & 0 & 0 & 0 & \cdots \\ a_0(\phi) & b_1(\phi) & a_1^*(\phi) & 0 & 0 & \cdots \\ 0 & a_1(\phi) & b_2(\phi) & a_2^*(\phi) & 0 & \cdots \\ \vdots & \vdots & \vdots & \vdots & \vdots & \ddots \end{pmatrix}$$

with the entries

$$\begin{aligned} a_n(\phi) &: \mathcal{F}_n(H) \to \mathcal{F}_{n+1}(H), \\ b_n(\phi) = (b_n(\phi))^* &: \mathcal{F}_n(H) \to \mathcal{F}_n(H), \\ a_n^*(\phi) = (a_n(\phi))^* &: \mathcal{F}_{n+1}(H) \to \mathcal{F}_n(H), \\ \phi \in H_+,\ n \in \mathbb{Z}_+ &= 0, 1, \ldots. \end{aligned}$$

Each matrix $\mathcal{J}(\phi)$ gives rise to a Hermitian operator $J(\phi)$ in the space $\mathcal{F}(H)$ with its domain $\mathrm{Dom}(J(\phi)) = \mathcal{F}_{\mathrm{fin}}(H_+)$.

Consider the following assumptions.

1. The operators $a_n(\phi)$ and $b_n(\phi)$, $\phi \in H_+$, $n \in \mathbb{Z}_+$, are bounded and real (i.e., they take real vectors to real ones).
2. (Smoothness) The space $\mathcal{F}_{\mathrm{fin}}(H_+)$ is invariant with respect to $a_n(\phi)$, $b_n(\phi)$, and $a_n^*(\phi)$.
3. The operators $J(\phi)$, $\phi \in H_+$, are essentially selfadjoint and their closures $\tilde{J}(\phi)$ are strongly commuting.
4. The dependence of the entries $a_n(\phi)$, $b_n(\phi)$, and $a_n^*(\phi)$ on the parameter ϕ is linear and weakly continuous.
5. (Regularity) The linear operators $V_n : \mathcal{F}_n(H_+) \to \bigoplus_{j=0}^n \mathcal{F}_j(H_+)$ defined by the equalities

$$\begin{gathered} V_0 = \mathrm{Id}_{\mathbb{C}},\ V_n(\phi_1 \hat{\otimes} \cdots \hat{\otimes} \phi_n) = J(\phi_1)\ldots J(\phi_n)\Omega, \\ \phi_1, \ldots, \phi_n \in H_+,\ n \in \mathbb{N}, \end{gathered}$$

are continuous. Furthermore, the operators

$$\mathcal{F}_n(H_+) \ni F_n \mapsto V_{n,n}F_n = \mathrm{Pr}_{\mathcal{F}_n(H_+)}\, V_n F_n \in \mathcal{F}_n(H_+), \qquad n \in \mathbb{Z}_+,$$

are invertible. Remark that V_n play an essential role in the constructions of Section 3.

The family $J = (\tilde{J}(\phi))_{\phi \in H_+}$ is called a (commutative) *Jacobi field* if Assumptions 1–5 are satisfied (recall that $\tilde{J}(\phi)$ stands for the closure of the operator $J(\phi)$). For a more rigorous formulation of this definition, see, e.g., [4]. Once again we should emphasize that the operators $\tilde{J}(\phi)$ act in the Fock space $\mathcal{F}(H)$.

One can apply the projection spectral theorem, (see [6] and [27]) to the field $J = (\tilde{J}(\phi))_{\phi \in H_+}$. We only adduce the result of such an application here. Proofs can be found in [2].

Given $n \in \mathbb{Z}_+$, let $\mathcal{P}_n$ stand for the set of all continuous polynomials

$$H_- \ni \xi \mapsto \sum_{j=0}^{n} \langle \xi^{\otimes j}, a_j \rangle_H \in \mathbb{C}, \qquad a_j \in \mathcal{F}_j(H_+)$$

(we suppose $\xi^{\otimes 0} = 1$). The set $\mathcal{P} = \bigcup_{n=0}^{\infty} \mathcal{P}_n$ is a dense subset of $L^2(H_-, d\rho)$. The closure of $\mathcal{P}_n$ in $L^2(H_-, d\rho)$ is the set of ordinary polynomials. It will be denoted by $\tilde{\mathcal{P}}_n$.

Theorem 2.1. *There exist a vector-valued function $H_- \ni \xi \mapsto P(\xi) \in (\mathcal{F}_{\mathrm{fin}}(H_+))'$ and a Borel probability measure ρ on the space H_- (the spectral measure) such that the following statements hold:*

- *For every $\xi \in H_-$, the vector $P(\xi) \in (\mathcal{F}_{\mathrm{fin}}(H_+))'$ is a generalized joint eigenvector of J with eigenvalue ξ, i.e.,*

$$\langle P(\xi), \tilde{J}(\phi)\Phi \rangle_H = \langle \xi, \phi \rangle_H \langle P(\xi), \Phi \rangle_H, \qquad \phi \in H_+,\ \Phi \in \mathcal{F}_{\mathrm{fin}}(H_+).$$

- *The Fourier transform*

$$\mathcal{F}_{\mathrm{fin}}(H_+) \ni \Phi \mapsto I\Phi = \langle \Phi, P(\cdot) \rangle_H \in L^2(H_-, d\rho)$$

  can be extended to a unitary operator acting from $\mathcal{F}(H)$ to $L^2(H_-, d\rho)$. We preserve the notation I for this operator.
- *The Fourier transform I satisfies the formula*

$$I\Phi_n = \mathrm{Pr}_{\tilde{\mathcal{P}}_n \ominus \tilde{\mathcal{P}}_{n-1}} \langle V_{n,n}^{-1}\Phi_n, \cdot^{\otimes n} \rangle_H, \qquad \Phi_n \in \mathcal{F}_n(H_+),\ n \in \mathbb{Z}_+,$$

  (we suppose $\mathcal{P}_{-1} = \{0\}$).

Corollary 2.1. *The equality*

$$L^2(H_-, d\rho) = \bigoplus_{n=0}^{\infty} \left(\tilde{\mathcal{P}}_n \ominus \tilde{\mathcal{P}}_{n-1} \right) = \bigoplus_{n=0}^{\infty} I(\mathcal{F}_n(H_+))$$

holds true.

Along with the spectral measure ρ and the Fourier transform I introduced in Theorem 2.1, our further considerations will involve the characteristic functional

$$\hat{\rho}(\phi) = \int_{H_-} e^{i\langle \xi, \phi\rangle_H} d\rho(\xi), \qquad \phi \in H_+,$$

of the measure ρ.

2.2. Two principal examples of Jacobi fields

In this paper, we will mostly deal with the classical free field $J_{CF} = (\tilde{J}_{CF}(\phi))_{\phi\in H_+}$ and the Poisson field $J_P = (\tilde{J}_P(\phi))_{\phi\in H_+}$.

Consider the classical creation and annihilation operators

$$J_+(\phi)F_n = \sqrt{n+1}\,\phi \hat{\otimes} F_n,$$
$$J_-(\phi) = (J_+(\phi))^*, \qquad \phi \in H_+,\ F_n \in \mathcal{F}_n(H),\ n \in \mathbb{Z}_+,$$

in the space $\mathcal{F}(H)$. The classical free field is defined for an arbitrary rigging (2.1) by the formula

$$J_{CF}(\phi) = J_+(\phi) + J_-(\phi), \qquad \phi \in H_+.$$

The corresponding spectral measure is the standard Gaussian measure γ on H_-. Its characteristic functional is given by the formula

$$\hat{\gamma}(\phi) = \exp\left(-\frac{1}{2}\,\|\phi\|_H^2\right), \qquad \phi \in H_+.$$

The definition of the Poisson field is slightly more complicated. Namely, demanding that the space H in (2.1) equal $L^2(\mathbb{R}^d, d\mu)$ for a σ-finite Borel measure μ, we introduce the operators of the Poisson field as

$$J_P(\phi) = J_+(\phi) + J_0(\phi) + J_-(\phi), \qquad \phi \in H_+.$$

In order to define $J_0(\phi)$, consider the operator $b(\phi)$ of multiplication by the function $\phi \in H_+$ in the space $H_{\mathbb{C}}$. For an arbitrary $F_n \in \mathcal{F}_n(H)$, define

$$J_0(\phi)F_0 = 0,$$
$$J_0(\phi)F_n = (b(\phi)\otimes \mathrm{Id}_H \otimes \cdots \otimes \mathrm{Id}_H)F_n$$
$$+ (\mathrm{Id}_H \otimes b(\phi) \otimes \mathrm{Id}_H \otimes \cdots \otimes \mathrm{Id}_H)F_n + \cdots$$
$$+ (\mathrm{Id}_H \otimes \cdots \otimes \mathrm{Id}_H \otimes b(\phi))F_n, \qquad \phi \in H_+,\ n \in \mathbb{N}.$$

In other words, $J_0(\phi)$ equals the second (differential) quantization of $b(\phi)$.

Of course, we choose the rigging (2.1) so that J_P would satisfy the definition of a Jacobi field. One can see that the Poisson field is nothing but a perturbation of the classical free field by a family of neutral operators $J_0(\phi)$.

The spectral measure of J_P is the centered Poisson measure π on H_- with the intensity μ. Its characteristic functional is given by the formula

$$\hat{\pi}(\phi) = \exp\left(\int_{\mathbb{R}^d} \left(e^{i\phi(x)} - 1 - i\phi(x)\right)\, d\mu(x)\right), \qquad \phi \in H_+.$$

For both J_{CF} and J_P, the operator $V_{n,n}$ satisfies the equality

$$V_{n,n} = \sqrt{n!}\,\mathrm{Id}_{\mathcal{F}_n(H_+)}, \qquad n \in \mathbb{Z}_+.$$

The Fourier transform I coincides with the Wiener-Itô-Segal transform associated with the corresponding spectral measure. Corollary 2.1 constitutes the Wiener-Itô decomposition.

3. Image of the Spectral Measure

This section aims to study the image of the measure ρ under a bounded operator. Proofs of statements can be found in [13].

Consider a real separable Hilbert space T. Let

$$T_- \supset T \supset T_+ \tag{3.1}$$

be a rigging of T with real separable Hilbert spaces T_+ and $T_- = (T_+)'$. As in the case of the rigging (2.1), the pairing in (3.1) can be extended to a pairing between $\mathcal{F}_n(T_+)$ and $\mathcal{F}_n(T_-)$. The latter can be extended to a pairing between $\mathcal{F}_{\mathrm{fin}}(T_+)$ and $(\mathcal{F}_{\mathrm{fin}}(T_+))'$. We use the notation $\langle\cdot,\cdot\rangle_T$ for all of these pairings.

Consider a bounded operator $K : T_+ \to H_+$ such that $\mathrm{Ker}(K) = \{0\}$. We preserve the notation K for the extension of this operator to the complexified space $(T_+)_{\mathbb{C}}$.

The adjoint of K with respect to (2.1) and (3.1) is a bounded operator $K^+ : H_- \to T_-$ defined by the equality

$$\langle K^+\xi, f\rangle_T = \langle \xi, Kf\rangle_H, \qquad \xi \in H_-,\ f \in T_+.$$

One can prove that $\mathrm{Ran}(K^+)$ is dense in T_-.

We denote by ρ_K the image of the measure ρ under the mapping K^+. By definition, ρ_K is a probability measure on the σ-algebra

$$\mathcal{C} = \{\Delta \subset T_- | (K^+)^{-1}(\Delta) \text{ is a Borel subset of } H_-\}$$

($(K^+)^{-1}(\Delta)$ denoting the preimage of the set Δ).

Remark 3.1. The characteristic functional

$$\hat{\rho}_K(f) = \int_{T_-} e^{i\langle\omega, f\rangle_T} d\rho_K(\omega), \qquad f \in T_+,$$

of the measure ρ_K satisfies the equality

$$\hat{\rho}_K(f) = \hat{\rho}(Kf), \qquad f \in T_+.$$

Remark 3.2. The assumption $\mathrm{Ker}(K) = \{0\}$ is not essential. Indeed, the measure ρ_K proves to be lumped on the set of functionals which equal zero on $\mathrm{Ker}(K)$. This set can be naturally identified with $(\mathrm{Ker}(K)^\perp)'$. Thus we can always replace T_+ with $\mathrm{Ker}(K)^\perp \subset T_+$.

3.1. The space $\mathcal{F}^{\mathrm{ext}}(T_+, K)$ and the family J_K

Before constructing the space $\mathcal{F}^{\mathrm{ext}}(T_+, K)$, we have to introduce an auxiliary notation. Given $n \in \mathbb{Z}_+$, let $\mathcal{V}_n$ stand for the subspace of $\mathcal{F}(H)$ generated by the vectors

$$V_j K^{\otimes j} F_j, \qquad F_j \in \mathcal{F}_j(T_+),\ j = 0, \dots, n$$

(we suppose $K^{\otimes 0} = \mathrm{Id}_{\mathbb{C}}$). Roughly speaking, $\mathcal{V}_n$ is a "prototype" in $\mathcal{F}(H)$ for the set $\tilde{\mathcal{Q}}_n$ of ordinary polynomials on T_- with their degree not greater than n. The corresponding set of continuous polynomials will be denoted by $\mathcal{Q}_n$.

Introduce the mapping $A : \mathcal{F}_{\mathrm{fin}}(T_+) \to \mathcal{F}(H)$ via the formula

$$\mathcal{F}_n(T_+) \ni F_n \mapsto AF_n = \frac{1}{\sqrt{n!}} \Pr_{\mathcal{V}_n \ominus \mathcal{V}_{n-1}} V_n K^{\otimes n} F_n \in \mathcal{F}(H), \qquad n \in \mathbb{Z}_+$$

(we suppose $\mathcal{V}_{-1} = \{0\}$). It is easy to see that $\mathrm{Ker}(A) = \{0\}$. Let $\mathcal{F}_n^{\mathrm{ext}}(T_+, K)$ denote the completion of $\mathcal{F}_n(T_+)$ with respect to the scalar product

$$(F_n, G_n)_{\mathcal{F}_n^{\mathrm{ext}}(T_+,K)} = (AF_n, AG_n)_{\mathcal{F}(H)}, \qquad F_n, G_n \in \mathcal{F}_n(T_+),\ n \in \mathbb{Z}_+.$$

Define

$$\mathcal{F}^{\mathrm{ext}}(T_+, K) = \bigoplus_{n=0}^{\infty} \mathcal{F}_n^{\mathrm{ext}}(T_+, K).$$

Obviously, the mapping A can be extended to an isometric operator acting from $\mathcal{F}^{\mathrm{ext}}(T_+, K)$ to $\mathcal{F}(H)$. The notation A is preserved for this operator.

The structure of A will imply the unitarity of the map I_K mentioned in Section 1. The rigorous definition of this map will be given in Theorem 3.1. We emphasize that I_K is simultaneously the generalized Wiener-Itô-Segal transform for ρ_K and the Fourier transform for the family J_K that we are about to construct.

Now we have to describe a natural rigging for the space $\mathcal{F}^{\mathrm{ext}}(T_+, K)$. Consider a linear topological space

$$\mathcal{F}_+^{\mathrm{ext}}(T_+, K) = A^{-1}(\mathcal{F}_{\mathrm{fin}}(H_+) \cap \mathrm{Ran}(A)).$$

The sequence $(F_n)_{n=0}^{\infty}$ converges to F in $\mathcal{F}_+^{\mathrm{ext}}(T_+, K)$ if and only if the sequence $(AF_n)_{n=0}^{\infty}$ converges to AF in $\mathcal{F}_{\mathrm{fin}}(H_+)$. One can show that the space $\mathcal{F}_+^{\mathrm{ext}}(T_+, K)$ is a dense subset of $\mathcal{F}^{\mathrm{ext}}(T_+, K)$. This gives us the rigging

$$(\mathcal{F}_+^{\mathrm{ext}}(T_+, K))' \supset \mathcal{F}^{\mathrm{ext}}(T_+, K) \supset \mathcal{F}_+^{\mathrm{ext}}(T_+, K).$$

Denote the corresponding pairing by $\langle \cdot, \cdot \rangle_A$.

Let us construct the family J_K. The set $\mathcal{F}_{\mathrm{fin}}(H_+) \cap \mathrm{Ran}(A)$ is invariant with respect to every $J(Kf)$, $f \in T_+$. This allows us to introduce the operators

$$J_K(f) = A^{-1} J(Kf) A, \qquad \mathrm{Dom}(J_K(f)) = \mathcal{F}_+^{\mathrm{ext}}(T_+, K),\ f \in T_+.$$

in the space $\mathcal{F}^{\mathrm{ext}}(T_+, K)$. Evidently, they are essentially selfadjoint and their closures $\tilde{J}_K(f)$ are strong commuting. Define $J_K = (\tilde{J}_K(f))_{f \in T_+}$. Remark that $\mathcal{F}_+^{\mathrm{ext}}(T_+, K)$ is invariant with respect to $\tilde{J}_K(f)$, $f \in T_+$.

We now state an analogue of Theorem 2.1 for J_K. In particular, this would yield an analogue of the Wiener-Itô decomposition for the measure ρ_K.

Theorem 3.1. *Assume $\mathcal{Q} = \bigcup_{n=0}^{\infty} \mathcal{Q}_n$ to be a dense subset of $L^2(T_-, d\rho_K)$. There exists a vector-valued function $T_- \ni \omega \mapsto Q(\omega) \in (\mathcal{F}_+^{\mathrm{ext}}(T_+, K))'$ such that the following statements hold:*

- *For ρ_K-almost all $\omega \in T_-$, the vector $Q(\omega) \in (\mathcal{F}_+^{\mathrm{ext}}(T_+, K))'$ is a generalized joint eigenvector of the family J_K with the eigenvalue ω, i.e.,*

$$\langle Q(\omega), \tilde{J}_K(f) F \rangle_A = \langle \omega, f \rangle_T \langle Q(\omega), F \rangle_A, \qquad F \in \mathcal{F}_+^{\mathrm{ext}}(T_+, K).$$

- *The Fourier transform*

$$\mathcal{F}_+^{\mathrm{ext}}(T_+, K) \ni F \mapsto I_K F = \langle F, Q(\cdot) \rangle_A \in L^2(T_-, d\rho_K)$$

  can be extended to a unitary operator acting from $\mathcal{F}^{\mathrm{ext}}(T_+, K)$ to $L^2(T_-, d\rho_K)$. We preserve the notation I_K for this operator.
- *The Fourier transform I_K satisfies the equality*

$$I_K F_n = \frac{1}{\sqrt{n!}} \operatorname{Pr}_{\tilde{\mathcal{Q}}_n \ominus \tilde{\mathcal{Q}}_{n-1}} \langle F_n, \cdot^{\otimes n} \rangle_T, \qquad F_n \in \mathcal{F}_n(T_+),\ n \in \mathbb{Z}_+$$

  (we suppose $\mathcal{Q}_{-1} = \{0\}$ and $\omega^{\otimes 0} = 1$ for any $\omega \in T_-$).

Corollary 3.1. *If $\mathcal{Q} = \bigcup_{n=0}^{\infty} \mathcal{Q}_n$ is a dense subset of $L^2(T_-, d\rho_K)$, then the equality*

$$L^2(T_-, d\rho_K) = \bigoplus_{n=0}^{\infty} \left(\tilde{\mathcal{Q}}_n \ominus \tilde{\mathcal{Q}}_{n-1} \right) = \bigoplus_{n=0}^{\infty} I_K(\mathcal{F}_n^{\mathrm{ext}}(T_+, K)).$$

holds true.

Corollary 3.1 constitutes an analogue of the Wiener-Itô decomposition for the measure ρ_K.

3.2. A Lévy noise measure

We will now illustrate the preceding abstract constructions with some explicit calculations. Namely, we will obtain an explicit formula for the operator A and the scalar product $(\cdot, \cdot)_{\mathcal{F}^{\mathrm{ext}}(T_+, K)}$ in the case where K is the operator of multiplication by a function of a new independent variable and J equals J_P. Then ρ_K appears to be a Lévy noise measure. Remark that a slightly more complicated choice of K leads to a fractional Lévy noise measure.

Consider a real separable Hilbert space $S = L^2(\mathbb{R}^{d_1}, d\sigma)$. Let the space T equal $L^2(\mathbb{R}^{d_2}, d\tau)$. We assume the Borel measures σ and τ to be finite on compact sets. We also assume τ to be absolutely continuous with respect to the Lebesgue measure. Let the space H equal $S \otimes T$. Clearly, H can be identified with $L^2(\mathbb{R}^{d_1+d_2}, d(\sigma \otimes \tau))$. Suppose J to equal J_P.

The spaces T_+ and H_+ may be chosen arbitrarily provided that J_P satisfies the definition of a Jacobi field. Typically, the role of T_+ and H_+ is played by weighted Sobolev spaces.

Define K via the formula

$$T_+ \ni f(t) \mapsto (Kf)(s,t) = \kappa(s)f(t), \qquad s \in \mathbb{R}^{d_1},\ t \in \mathbb{R}^{d_2}.$$

The function $\kappa \in S$ has to be chosen so that K would be a bounded operator acting from T_+ to H_+.

The operator K^+ takes ρ to a probability measure ρ_K on T_-. According to Remark 3.1, the characteristic functional of ρ_K is now given by the formula

$$\hat{\rho}_K(f) = \exp\left(\int_{\mathbb{R}^{d_2}}\int_{\mathbb{R}^{d_1}} \left(e^{i\kappa(s)f(t)} - 1 - i\kappa(s)f(t)\right) d\sigma(s)d\tau(t)\right), \quad f \in T_+.$$

Denote by σ_κ the image of σ under κ. The above formula implies that ρ_K is the Lévy noise measure on T_- with the Lévy measure σ_κ and the intensity measure τ.

Before identifying the operator A and the scalar product $(\cdot,\cdot)_{\mathcal{F}^{\mathrm{ext}}(T_+,K)}$ explicitly, we have to carry out some preliminary constructions.

Given $n \in \mathbb{N}$, define $\kappa^n(s) = (\kappa(s))^n$, $s \in \mathbb{R}^{d_1}$. The function κ^n belongs to the space S for any $n \in \mathbb{N}$. Applying the Schmidt orthogonalization procedure to the sequence $(\kappa^n)_{n=1}^\infty$, we obtain an orthogonal sequence $(\kappa_n)_{n=0}^\infty$ in the space S. Each κ_n is a polynomial of degree n with respect to κ. We normalize κ_n so that the leading coefficient of this polynomial would equal 1.

A vector $F \in T_{\mathbb{C}}^{\otimes n}$, $n \in \mathbb{N}$, can be treated as a complex-valued function $F(t_1,\dots,t_n)$ depending on the variables $t_1,\dots,t_n \in \mathbb{R}^{d_2}$. Analogously, a vector $\Phi \in H_{\mathbb{C}}^{\otimes n}$, $n \in \mathbb{N}$, can be treated as a complex-valued function $\Phi(s_1,\dots,s_n,t_1,\dots,t_n)$ depending on the variables $s_1,\dots,s_n \in \mathbb{R}^{d_1}$ and $t_1,\dots,t_n \in \mathbb{R}^{d_2}$. Vectors from $\mathcal{F}_n(T)$ and $\mathcal{F}_n(H)$ appear as symmetric functions. We assume the set of all smooth compactly supported functions on $\mathbb{R}^{d_2 n}$ to be a dense subset of $(T_+)_{\mathbb{C}}^{\otimes n}$.

Consider an ordered partition $\omega = (\omega_1,\dots,\omega_k)$ of the set $\{1,\dots,n\}$ into k nonempty sets $\omega_1,\dots,\omega_k$. Let Ω_n^k stand for the set of all such partitions and let $|\omega_k|$ stand for the cardinality of ω_k. Introduce the mapping

$$\mathbb{R}^{d_2 k} \ni (t_1,\dots,t_k) \mapsto \pi_\omega(t_1,\dots,t_k) = (t_{i_1},\dots,t_{i_n}) \in \mathbb{R}^{d_2 n}$$

with $i_j = l$ for $j \in \omega_l$.

Given a smooth compactly supported symmetric function $F \in \mathcal{F}_n(T_+)$, $n \in \mathbb{N}$, denote $D_F = (0, D_F^1,\dots,D_F^n,0,0,\dots) \in \mathcal{F}(H)$ with

$$D_F^k(s_1,\dots,s_k,t_1,\dots,t_k) = \sum_{\omega\in\Omega_n^k} \frac{1}{\sqrt{k!}} \left(\kappa_{|\omega_1|}(s_1)\cdots\kappa_{|\omega_k|}(s_k)\right) F(\pi_\omega(t_1,\dots,t_k)), \qquad k = 1,\dots,n.$$

Theorem 3.2. *Under the assumptions of the present subsection, the operator A and the scalar product $(\cdot,\cdot)_{\mathcal{F}^{\rm ext}(T_+,K)}$ satisfy the equalities*

$$AF = \frac{1}{\sqrt{n!}}\, D_F,$$

$$(F,G)_{\mathcal{F}^{\rm ext}(T_+,K)} = \frac{1}{n!}\sum_{k=1}^{n} \int_{\mathbb{R}^{d_2 k}} \int_{\mathbb{R}^{d_1 k}} D_F^k(s,t)\overline{D_G^k(s,t)}\, d\sigma^{\otimes k}(s) d\tau^{\otimes k}(t)$$

for any smooth compactly supported symmetric functions $F, G \in \mathcal{F}_n(T_+)$, $n \in \mathbb{N}$ (the overbar denoting the complex conjugacy).

Theorem 3.1 and Theorem 3.2 explain the Fock-type structure of $L^2(T_-, d\rho_K)$. Theorem 3.2 shows that the space $\mathcal{F}^{\rm ext}(T_+,K)$ coincides with the extended Fock space investigated in [23] up to scalar weights at the orthogonal components. Note that one can construct an embedding of $\mathcal{F}^{\rm ext}(T_+,K)$ into a weighted orthogonal sum of function spaces. Using the arguments from [10], see also [19], [11], [12], [23], and [24], one can extend this embedding to a unitary operator.

Since ρ_K is now a Lévy noise measure, the family J_K is probably isomorphic in a certain sense to the operator family from [8]. (In the corresponding special case, this family is the Gamma field, see [19].) The measure ρ_K would then be the spectral measure of this family. We do not concentrate on these questions in the present paper.

4. The case of a dense range

If the range $\mathrm{Ran}(K)$ is dense in H_+, then the abstract constructions of Section 3 take a much simpler form. Before explaining the principal simplification, we should point out that several relevant results can be found in [9]. In particular, a statement similar to Corollary 3.1 is obtained there by means of an approximation procedure.

So we assume the range $\mathrm{Ran}(K)$ to be dense in H_+. Then $\mathcal{V}_n = \bigoplus_{j=0}^{n} \mathcal{F}_j(H)$ and

$$AF_n = \frac{1}{\sqrt{n!}}\, V_{n,n} K^{\otimes n} F_n, \qquad F_n \in \mathcal{F}_n(T_+),\ n \in \mathbb{Z}_+.$$

If, additionally, J is the classical free field or the Poisson field, then

$$AF_n = K^{\otimes n} F_n, \qquad F_n \in \mathcal{F}_n(T_+),\ n \in \mathbb{Z}_+,$$

and the scalar product $(\cdot,\cdot)_{\mathcal{F}^{\rm ext}(T_+,K)}$ satisfies the equality

$$(F_n, G_n)_{\mathcal{F}_n^{\rm ext}(T_+,K)} = (K^{\otimes n} F_n, K^{\otimes n} G_n)_{\mathcal{F}_n(H)},$$
$$F_n, G_n \in \mathcal{F}_n(T_+),\ n \in \mathbb{Z}_+.$$

This formula shows that $\mathcal{F}^{\rm ext}(T_+,K)$ may now be identified with $\mathcal{F}(T_0)$, the space T_0 being the completion of T_+ with respect to the scalar product

$$(f,g)_{T_0} = (Kf, Kg)_H, \qquad f, g \in T_+.$$

We proceed with three examples.

4.1. A Gaussian measure

Choose an arbitrary rigging (2.1) and suppose the original field J to equal J_{CF}. The space T_+ and the bounded operator K may be chosen arbitrarily. However, as before, we assume $\mathrm{Ran}(K)$ to be dense in H_+ and $\mathrm{Ker}(K)$ to equal $\{0\}$. According to Remark 3.1, the characteristic functional of ρ_K is given by the formula

$$\hat{\rho}_K(f) = \exp\left(-\frac{1}{2}\|Kf\|_H^2\right) = \exp\left(-\frac{1}{2}\langle K^+Kf, f\rangle_T\right), \quad f \in T_+$$

(since H_+ is a subset of H_-, the operator $K^+K : T_+ \to T_-$ is well defined). This means ρ_K is the Gaussian measure on T_- with the correlation operator K^+K.

Notice that we can choose the neutral space T in the rigging (3.1) to equal the space T_0 defined above. Then the restriction $K^+ \restriction \mathrm{Ran}(K) : \mathrm{Ran}(K) \to T_-$ coincides with the mapping $K^{-1} : \mathrm{Ran}(K) \to T_+ \subset T_-$ and the characteristic functional of ρ_K may be written in the form

$$\hat{\rho}_K(f) = \exp\left(-\frac{1}{2}\|f\|_T^2\right), \qquad f \in T_+.$$

Clearly, this means ρ_K is the standard Gaussian measure on T_-.

Theorem 3.1 produces a family J_K whose spectral measure is ρ_K. Obviously, this family coincides with the classical free field corresponding to the rigging (3.1) with the neutral space $T = T_0$. Originally, this fact has been pointed out by E.W. Lytvynov.

4.2. An operator of multiplication

Let H be $L^2(\mathbb{R}, dx)$ and let H_+ and T_+ be the Sobolev spaces $W_2^1(\mathbb{R}, (1+x^2)\,dx)$ and $W_2^1(\mathbb{R}, dx)$, respectively. Suppose J to be the Poisson field J_P and suppose $K : T_+ \to H_+$ to be the operator of multiplication by the function $\theta(x) = e^{-x^2}$. One can easily verify that K is bounded and $\mathrm{Ker}(K) = \{0\}$. The range $\mathrm{Ran}(K)$ is dense in H_+ because it contains all the smooth compactly supported functions. On the other hand, $\mathrm{Ran}(K) \neq H_+$ because, e.g., the function $\psi(x) = (1+x^2)^{-2} \in H_+$ does not belong to $\mathrm{Ran}(K)$.

Choose T_- to be the dual of $W_2^1(\mathbb{R}, dx)$ with respect to the neutral space $T_0 = L^2(\mathbb{R}, e^{-2x^2}\,dx)$. Evidently, one may realize T_- as the dual of $W_2^1(\mathbb{R}, dx)$ with respect to the neutral space $L^2(\mathbb{R}, dx)$, in which case T_- is the usual negative Sobolev space $W_2^{-1}(\mathbb{R}, dx)$. Applying Theorem 3.1 in this situation yields a family J_K whose spectral measure has the characteristic functional

$$\hat{\rho}_K(f(x)) = \exp\left(\int_{\mathbb{R}} \left(e^{ie^{-x^2}f(x)} - 1 - ie^{-x^2}f(x)\right)\,dx\right), \quad f \in T_+.$$

The operator A is now an orthogonal sum of multiplication operators. The space $\mathcal{F}^{\mathrm{ext}}(T_+, K)$ where J_K is defined appears as an orthogonal sum of L^2-spaces, each of them with respect to a weighted Lebesgue measure.

4.3. A differential operator

As before, let H be $L^2(\mathbb{R}, dx)$. Let H_+ and T_+ equal $W_2^1\big(\mathbb{R}, e^{\frac{x^2}{2}}\, dx\big)$ and $W_2^2\big(\mathbb{R}, e^{\frac{x^2}{2}}\, dx\big)$, respectively. Suppose J to be the Poisson field J_P.

Define the operator $K : T_+ \to H_+$ as the extension by continuity of the mapping

$$\mathrm{C}_0^\infty(\mathbb{R}) \ni p(x) \mapsto e^{-\frac{x^2}{2}} \frac{dp(x)}{dx} \in H_+$$

($\mathrm{C}_0^\infty(\mathbb{R})$ stands for the set of all smooth compactly supported functions on $\mathbb{R}$). Evidently, K is bounded and $\mathrm{Ker}(K) = \{0\}$. One can prove that the range $\mathrm{Ran}(K)$ is dense in H_+.

The space H_- is the negative Sobolev space $W_2^{-1}\big(\mathbb{R}, e^{\frac{x^2}{2}}\, dx\big)$, while T_- may be realized as the dual of $W_2^2(\mathbb{R}, e^{\frac{x^2}{2}}\, dx)$ with respect to the zero space $L^2(\mathbb{R}, dx)$. In this case, T_- is the usual negative Sobolev space $W_2^{-2}(\mathbb{R}, e^{\frac{x^2}{2}}\, dx)$. Applying Theorem 3.1 yields a family J_K whose spectral measure has the characteristic functional

$$\hat{\rho}_K(f(x)) = \exp\left(\int_{\mathbb{R}} \left(\exp\left(ie^{-\frac{x^2}{2}} \frac{df(x)}{dx}\right) - 1 - ie^{-\frac{x^2}{2}} \frac{df(x)}{dx}\right) dx\right), \quad f \in T_+.$$

The operator A is an orthogonal sum of differential operators. The structure of the space $\mathcal{F}^{\mathrm{ext}}(T_+, K)$ is now slightly more complicated than it was in the previous example.

Acknowledgement

Yurij M. Berezansky acknowledges the support of DFG, Project 436 UKR 113/78/0-1.

References

[1] Yu.M. Berezansky, *Direct and inverse spectral problems for Jacobi fields*, St. Petersburg Math. J. **9** (1998), 1053–1071.

[2] Yu.M. Berezansky, *Commutative Jacobi fields in Fock space*, Integr. Equ. Oper. Theory **30** (1998), 163–190.

[3] Yu.M. Berezansky, *On the theory of commutative Jacobi fields*, Methods Funct. Anal. Topology **4** no. 1 (1998), 1–31.

[4] Yu.M. Berezansky, *Spectral theory of commutative Jacobi fields: Direct and inverse problems*, Fields Inst. Commun. **25** (2000), 211–224.

[5] Yu.M. Berezansky, *Poisson measure as the spectral measure of Jacobi field*, Infin. Dimen. Anal. Quant. Prob. Rel. Top. **3** (2000), 121–139.

[6] Yu.M. Berezansky and Yu.G. Kondratiev, *Spectral Methods in Infinite-Dimensional Analysis* (Kluwer, 1995); (in Russian, Naukova Dumka, 1988).

[7] Yu.M. Berezansky, V.O. Livinsky, and E.W. Lytvynov, *A generalization of Gaussian white noise analysis*, Methods Funct. Anal. Topology **1** no. 1 (1995), 28–55.

[8] Yu.M. Berezansky, E.W. Lytvynov, and D.A. Mierzejewski, *The Jacobi field of a Lévy process*, Ukrainian Math. J. **55** (2003), 706–710.

[9] Yu.M. Berezansky, E.W. Lytvynov, and A.D. Pulemyotov, *Image of the spectral measure of a Jacobi field and the corresponding operators*, Integr. Equ. Oper. Theory **53** (2005), 191–208.

[10] Yu.M. Berezansky and D.A. Mierzejewski, *The structure of the extended symmetric Fock space*, Methods Funct. Anal. Topology **6** no. 4 (2000), 1–13.

[11] Yu.M. Berezansky and D.A. Mierzejewski, *The chaotic decomposition for the gamma field*, Funct. Anal. Appl. **35** (2001), 263–266.

[12] Yu.M. Berezansky and D.A. Mierzejewski, *The construction of chaotic representation for the gamma field*, Infin. Dimen. Anal. Quant. Prob. Rel. Top. **6** (2003), 33–56.

[13] Yu.M. Berezansky and A.D. Pulemyotov, *The spectral theory and the Wiener-Itô decomposition for the image of a Jacobi field*, submitted.

[14] C.F. Dunkl and Yuan Xu, *Orthogonal Polynomials of Several Variables*, Encyclopedia of Mathematics and its Applications, Vol. 81 (Cambridge Univ. Press, 2001).

[15] M.I. Gekhtman and A.A. Kalyuzhny, *Spectral theory of orthogonal polynomials in several variables*, Ukrainian Math. J. **43** (1991), 1334–1337.

[16] M.I. Gekhtman and A.A. Kalyuzhny, *On the orthogonal polynomials in several variables*, Integr. Equ. Oper. Theory **19** (1994), 404–418.

[17] R.L. Hudson and K.R. Parthasarathy, *Quantum Itô's formula and stochastic evolutions*, Comm. Math. Phys. **93** (1984), 301–323.

[18] S. Janson, *Gaussian Hilbert Spaces* (Cambridge Univ. Press, 1997).

[19] Yu.G. Kondratiev and E.W. Lytvynov, *Operators of gamma white noise calculus*, Infin. Dimen. Anal. Quant. Prob. Rel. Top. **3** (2000), 303–335.

[20] Yu.G. Kondratiev, J.L. Silva, L. Streit, and G.F. Us, *Analysis on Poisson and Gamma spaces*, Infin. Dimen. Anal. Quant. Prob. Rel. Top. **1** (1998), 91–117.

[21] Y.-J. Lee and H.-H. Shih, *The Segal-Bargmann transform for Lévy functionals*, J. Funct. Anal. **168** (1999), 46–83.

[22] E.W. Lytvynov, *Multiple Wiener integrals and non-Gaussian white noises: A Jacobi field approach*, Methods Funct. Anal. Topology **1** no. 1 (1995), 61–85.

[23] E.W. Lytvynov, *Orthogonal decompositions for Lévy processes with an application to the gamma, Pascal, and Meixner processes*, Infin. Dimen. Anal. Quant. Prob. Rel. Top. **6** (2003), 73–102.

[24] E.W. Lytvynov, *Polynomials of Meixner's type in infinite dimensions — Jacobi fields and orthogonality measures*, J. Funct. Anal. **200** (2003), 118–149.

[25] E.W. Lytvynov, *The square of white noise as a Jacobi field*, Infin. Dimen. Anal. Quant. Prob. Rel. Top. **7** (2004), 619–629.

[26] D. Nualart and W. Schoutens, *Chaotic and predictable representations for Lévy processes*, Stochastic Process. Appl. **90** (2000), 109–122.

[27] A.D. Pulemyotov, *Support of a joint resolution of identity and the projection spectral theorem*, Infin. Dimen. Anal. Quant. Prob. Rel. Top. **6** (2003), 549–561.

[28] I. Rodionova, *Analysis connected with generating functions of exponential type in one and infinite dimensions*, Methods Funct. Anal. Topology **11** (2005), 275–297.

[29] W. Schoutens, *Stochastic Processes and Orthogonal Polynomials*, Lecture Notes in Statist., Vol. 146 (Springer, 2000).

[30] D. Surgailis, *On multiple Poisson stochastic integrals and associated Markov semigroups*, Probab. Math. Statist. **3** (1984), 217–239.

[31] N. Tsilevich and A. Vershik, *Fock factorizations, and decompositions of the L^2 spaces over general Lévy processes*, Russian Mathematical Surveys **58** no. 3 (2003), 427–472.

Yurij M. Berezansky
Institute of Mathematics
National Academy of Sciences of Ukraine
3 Tereshchenkivs'ka Str.
01601 Kyiv
Ukraine
e-mail: berezan@mathber.carrier.kiev.ua

Artem D. Pulemyotov
Department of Mathematics
Cornell University
310 Malott Hall
Ithaca, NY 14853-4201
USA
e-mail: artem@math.cornell.edu

Operator Theory:
Advances and Applications, Vol. 179, 63–102

The Higher Order Carathéodory–Julia Theorem and Related Boundary Interpolation Problems

Vladimir Bolotnikov and Alexander Kheifets

Abstract. The higher order analogue of the classical Carathéodory-Julia theorem on boundary angular derivatives has been obtained in [7]. Here we study boundary interpolation problems for Schur class functions (analytic and bounded by one in the open unit disk) motivated by that result.

Mathematics Subject Classification (2000). 30E05, 46E22, 47A57, 47A20, 47A48.

Keywords. Boundary interpolation, angular derivatives, unitary extensions, characteristic function of a unitary colligation.

1. Introduction

We denote by $\mathcal{S}$ the Schur class of analytic functions mapping the open unit disk $\mathbb{D}$ into its closure. A well known property of Schur functions w is that the kernel

$$K_w(z,\zeta) = \frac{1 - w(z)\overline{w(\zeta)}}{1 - z\bar{\zeta}} \tag{1.1}$$

is positive on $\mathbb{D} \times \mathbb{D}$ and therefore, that the matrix

$$\mathbf{P}_n^w(z) := \left[\frac{1}{i!j!} \frac{\partial^{i+j}}{\partial z^i \partial \bar{z}^j} \frac{1 - |w(z)|^2}{1 - |z|^2} \right]_{i,j=0}^{n} \tag{1.2}$$

which will be referred to as to a *Schwarz-Pick matrix*, is positive semidefinite for every $n \geq 0$ and $z \in \mathbb{D}$. We extend this notion to boundary points as follows: *given a point $t_0 \in \mathbb{T}$, the boundary Schwarz-Pick matrix is*

$$\mathbf{P}_n^w(t_0) = \lim_{z \to t_0} \mathbf{P}_n^w(z), \tag{1.3}$$

provided the limit in (1.3) *exists.* It is clear that once the boundary Schwarz-Pick matrix $\mathbf{P}_n^w(t_0)$ exists for $w \in \mathcal{S}$, it is positive semidefinite. In (1.3) and in what

follows, all the limits are nontangential, i.e., $z \in \mathbb{D}$ tends to a boundary point nontangentially. Let us assume that $w \in \mathcal{S}$ possesses nontangential boundary limits

$$w_j(t_0) := \lim_{z \to t_0} \frac{w^{(j)}(z)}{j!} \quad \text{for } \ j = 0, \dots, 2n+1 \tag{1.4}$$

and let

$$\mathbb{P}_n^w(t_0) := \begin{bmatrix} w_1(t_0) & \cdots & w_{n+1}(t_0) \\ \vdots & & \vdots \\ w_{n+1}(t_0) & \cdots & w_{2n+1}(t_0) \end{bmatrix} \mathbf{\Psi}_n(t_0) \begin{bmatrix} w_0(t_0)^* & \dots & w_n(t_0)^* \\ & \ddots & \vdots \\ 0 & & w_0(t_0)^* \end{bmatrix}, \tag{1.5}$$

where the first factor is a Hankel matrix, the third factor is an upper triangular Toeplitz matrix and where $\mathbf{\Psi}_n(t_0) = [\Psi_{j\ell}]_{j,\ell=0}^n$ is the upper triangular matrix

$$\mathbf{\Psi}_n(t_0) = \begin{bmatrix} t_0 & -t_0^2 & t_0^3 & \cdots & (-1)^n \binom{n}{0} t_0^{n+1} \\ 0 & -t_0^3 & 2t_0^4 & \cdots & (-1)^n \binom{n}{1} t_0^{n+2} \\ \vdots & & t_0^5 & \cdots & (-1)^n \binom{n}{2} t_0^{n+3} \\ \vdots & & & \ddots & \vdots \\ 0 & \cdots & \cdots & 0 & (-1)^n \binom{n}{n} t_0^{2n+1} \end{bmatrix} \tag{1.6}$$

with entries

$$\Psi_{j\ell} = (-1)^\ell \binom{\ell}{j} t_0^{\ell+j+1}, \quad 0 \le j \le \ell \le n. \tag{1.7}$$

For notational convenience, in (1.5) and in what follows we use the symbol a^* for the complex conjugate of $a \in \mathbb{C}$.

We denote the lower diagonal entry in the Schwarz-Pick matrix $\mathbf{P}_n^w(z)$ by

$$d_{w,n}(z) := \frac{1}{(n!)^2} \frac{\partial^{2n}}{\partial z^n \partial \bar{z}^n} \frac{1 - |w(z)|^2}{1 - |z|^2}. \tag{1.8}$$

The following theorem was obtained in [7].

Theorem 1.1. *For $w \in \mathcal{S}$, $t_0 \in \mathbb{T}$ and $n \in \mathbb{Z}_+$, the following are equivalent:*

1. *The following limit inferior is finite:*

$$\liminf_{z \to t_0} d_{w,n}(z) < \infty \tag{1.9}$$

   where $z \in \mathbb{D}$ approaches t_0 unrestrictedly.
2. *The following nontangential boundary limit exists and is finite:*

$$d_{w,n}(t_0) := \lim_{z \to t_0} d_{w,n}(z) < \infty. \tag{1.10}$$

3. *The boundary Schwarz-Pick matrix $\mathbf{P}_n^w(t_0)$ defined via the nontangential boundary limit* (1.3) *exists.*

4. *The nontangential boundary limits* (1.4) *exist and satisfy*

$$|w_0(t_0)| = 1 \quad \textit{and} \quad \mathbb{P}_n^w(t_0) \geq 0, \tag{1.11}$$

where $\mathbb{P}_n^w(t_0)$ *is the matrix defined in* (1.5).

Moreover, when these conditions hold, then

$$\mathbf{P}_n^w(t_0) = \mathbb{P}_n^w(t_0). \tag{1.12}$$

In the case $n = 0$, Theorem 1.1 reduces to the classical Carathéodory-Julia theorem [9, 10]; this has been discussed in detail in [7]. The relation

$$d_{w,n}(t_0) = \begin{bmatrix} w_{n+1}(t_0) & \cdots & w_{2n+1}(t_0) \end{bmatrix} \mathbf{\Psi}_n(t_0) \begin{bmatrix} w_n(t_0)^* \\ \vdots \\ w_0(t_0)^* \end{bmatrix}$$

expresses equality of the lower diagonal entries in (1.12); upon separating the term containing w_{2n+1} it can be written as

$$d_{w,n}(t_0) = \sum_{i=0}^{n-1} \sum_{j=0}^{n} w_{n+i+1}(t_0) \Psi_{ij}(t_0) w_{n-j}(t_0)^* + (-1)^n t_0^{2n+1} w_{2n+1}(t_0) w_0(t_0)^*. \tag{1.13}$$

Theorem 1.1 motivates the following interpolation problem:

Problem 1.2. *Given points* $t_1, \ldots, t_k \in \mathbb{T}$*, given integers* $n_1, \ldots, n_k \geq 0$ *and given numbers* $c_{i,j}$ $(j = 0, \ldots, 2n_i + 1;\ i = 1, \ldots, k)$*, find all Schur functions* w *such that*

$$\liminf_{z \to t_i} d_{w,n_i}(z) < \infty \quad (i = 1, \ldots, k) \tag{1.14}$$

and

$$w_j(t_i) := \lim_{z \to t_i} \frac{w^{(j)}(z)}{j!} = c_{i,j} \quad (i = 1, \ldots, k;\ j = 0, \ldots, 2n_i + 1). \tag{1.15}$$

The problem makes sense since conditions (1.14) guarantee the existence of the nontangential limits (1.15); upon preassigning the values $w_j(t_i)$ for $i = 1, \ldots, k$ and $j = 0, \ldots, 2n_i + 1$, we come up with interpolation Problem 1.2. It is convenient to reformulate Problem 1.2 in the following form:

Problem 1.3. *Given points* $t_1, \ldots, t_k \in \mathbb{T}$*, given integers* $n_1, \ldots, n_k \geq 0$ *and given numbers*

$$c_{i,j} \quad \textit{and} \quad \gamma_i \quad (j = 0, \ldots, 2n_i;\ i = 1, \ldots, k),$$

find all Schur functions w *such that*

$$d_{w,n_i}(t_i) := \frac{1}{(n_i!)^2} \lim_{z \to t_i} \frac{\partial^{2n_i}}{\partial z^{n_i} \partial \bar{z}^{n_i}} \frac{1 - |w(z)|^2}{1 - |z|^2} = \gamma_i \tag{1.16}$$

and

$$w_j(t_i) = c_{i,j} \quad (i = 1, \ldots, k;\ j = 0, \ldots, 2n_i). \tag{1.17}$$

If w is a solution to Problem 1.2, then conditions (1.14) guarantee the existence of the nontangential limits (1.16) and by a virtue of (1.13),

$$\begin{aligned} d_{w,n_i}(t_i) &= \sum_{\ell=0}^{n_i-1}\sum_{j=0}^{n_i} w_{n_i+\ell+1}(t_i)\Psi_{\ell j}(t_0)w_{n_i-j}(t_i)^* \\ &\quad +(-1)^{n_i}t_i^{2n_i+1}w_{2n_i+1}(t_i)w_0(t_i)^*. \end{aligned} \tag{1.18}$$

Thus, for every Schur function w, satisfying (1.14) and (1.15), conditions (1.16) hold with

$$\gamma_i = \sum_{\ell=0}^{n-1}\sum_{j=0}^{n_i} c_{i,n_i+\ell+1}\Psi_{\ell j}(t_0)c_{i,n_i-j}^* + (-1)^{n_i}t_i^{2n_i+1}c_{i,2n_i+1}c_{i,0}^*. \tag{1.19}$$

Conversely, if w is a solution of Problem 1.3, then it clearly satisfies (1.14) and by Theorem 1.1, all the limits in (1.15) exist and satisfy relation (1.18). Since $w_0(t_i)$ is *unimodular*, the equation (1.18) can be solved for $w_{2n_i+1}(t_i)$; on account of interpolation conditions (1.17), we have

$$w_{2n_i+1}(t_i) = (-1)^n\overline{t}_i^{\,2n_i+1}\left(d_{w,n_i}(t_i) - \sum_{\ell=0}^{n_i-1}\sum_{j=0}^{n_i} c_{i,n_i+\ell+1}\Psi_{\ell j}(t_i)c_{i,n_i-j}^*\right)c_{i,0}. \tag{1.20}$$

It is readily seen now that w is a solution of Problem 1.2 with the data $c_{i,2n_i+1}$ chosen by

$$c_{i,2n_i+1} = (-1)^n\overline{t}_i^{\,2n_i+1}\left(\gamma_i - \sum_{\ell=0}^{n_i-1}\sum_{j=0}^{n_i} c_{i,n_i+\ell+1}\Psi_{\ell j}(t_i)c_{i,n_i-j}^*\right)c_{i,0}. \tag{1.21}$$

It is known that boundary interpolation problems become more tractable if they involve inequalities. Such a relaxed problem is formulated below; besides of certain independent interest it will serve as an important intermediate step in solving Problem 1.2.

Problem 1.4. *Given points $t_1,\dots,t_k\in\mathbb{T}$, given integers $n_1,\dots,n_k\geq 0$ and given numbers $c_{i,j}$ and γ_i $(j=0,\dots,2n_i;\ i=1,\dots,k)$, find all Schur functions w such that*

$$d_{w,n_i}(t_i)\leq\gamma_i, \tag{1.22}$$

$$w_j(t_i)=c_{i,j}\quad (i=1,\dots,k;\ j=0,\dots,2n_i). \tag{1.23}$$

By Theorem 1.1, for every solution w of Problem 1.3 there exists the limit $w_{2n_i+1}(t_i) := \lim\limits_{z\to t_i}\dfrac{w^{(2n_i+1)}(z)}{(2n_i+1)!}$ which satisfies (1.20). Let $c_{i,2n_i+1}$ be defined as in (1.21). Then it follows from (1.20), (1.21) and (1.22) that

$$0\leq\gamma_i - d_{w,n_i}(t_i) = (-1)^{n_i}t_i^{2n_i+1}\left(c_{i,2n_i+1}-w_{2n_i+1}(t_i)\right)c_{i,0}^*. \tag{1.24}$$

It is convenient to reformulate Problem 1.4 in the following equivalent form.

Problem 1.5. *Given the data*

$$t_i \in \mathbb{T} \quad \text{and} \quad c_{i,j} \in \mathbb{C} \quad (j = 0, \dots, 2n_i + 1;\ i = 1, \dots, k), \tag{1.25}$$

find all Schur functions w such that

$$d_{w,n_i}(t_i) \le \gamma_i \ , \tag{1.26}$$

$$w_j(t_i) = c_{i,j} \quad (i = 1, \dots, k;\ j = 0, \dots, 2n_i) \tag{1.27}$$

and

$$(-1)^{n_i} t_i^{2n_i+1} \left(c_{i,2n_i+1} - w_{2n_i+1}(t_i)\right) c_{i,0}^* \ge 0 \quad (i = 1, \dots, k), \tag{1.28}$$

where the γ_i's are defined by (1.19).

In Section 3 we will construct the Pick matrix P in terms of the interpolation data (1.25) (see formulas (3.1)–(3.2) below). Then we will show that Problem 1.5 has a solution if and only if $|c_{i,0}| = 1$ for $i = 1, \dots, k$ and $P \ge 0$. In case P is singular, Problem 1.5 has a unique solution w which is a finite Blaschke product of degree $r \le \operatorname{rank} P$. This unique w may or may not be a solution of Problem 1.2. The case when P is positive definite is more interesting.

Theorem 1.6. *Let $|c_{i,0}| = 1$ for $i = 1, \dots, k$ and $P > 0$. Then*

1. *Problem* 1.5 *has infinitely many solutions which are parametrized by the linear fractional transformation*

$$w(z) = s_0(z) + s_2(z) \left(1 - \mathcal{E}(z)s(z)\right)^{-1} \mathcal{E}(z)s_1(z) \tag{1.29}$$

where $\mathcal{E}$ is a free parameter running over the Schur class $\mathcal{S}$ and where the coefficient matrix

$$\mathbf{S}(z) = \begin{bmatrix} s_0(z) & s_2(z) \\ s_1(z) & s(z) \end{bmatrix} \tag{1.30}$$

is rational and inner in $\mathbb{D}$.

2. *A function w of the form* (1.29) *is a solution of Problem* 1.2 *if and only if either*

$$\liminf_{z \to t_i} \frac{1 - |\mathcal{E}(z)|^2}{1 - |z|^2} = \infty \quad \text{or} \quad \lim_{z \to t_i} \mathcal{E}(z) \ne s(t_i)^* \tag{1.31}$$

for $i = 1, \dots, k$, where the latter limit is understood as nontangential, and s is the right bottom entry of the coefficient matrix $\mathbf{S}(z)$.

Boundary interpolation problems for Schur class functions closely related to Problem 1.5 were studied previously in [3]–[6], [16]. Interpolation conditions (1.27) and (1.28) there were accompanied by various additional restrictions that in fact are equivalent to our conditions (1.26). Establishing these equivalences is a special issue which is discussed in [8]. A version of Problem 1.2 (with certain assumptions on the data that guarantee (1.14) to be in force) was studied in [4] for rational matrix-valued Schur functions. In this case, the parameters $\mathcal{E}$ in the parametrization formula (1.29) are also rational and therefore, the situation expressed by the first relation in (1.31) does not come into play. A similar matrix-valued problem was considered in [6] where the solvability criteria were established

rather than the description of all solutions. Problem 1.3 was considered in [21] in the case $n_1 = \ldots = n_k = 0$; the second part in Theorem 1.6 can be considered as a higher order generalization of some results in [21].

The paper is organized as follows. In Section 2 we recall some needed results from [7] and present some consequences of conditions (1.26) holding for a Schur class function. In Section 3 we introduce the Pick matrix P in terms of the interpolation data and establish the Stein equality this matrix satisfies. In Section 4 we imbed Problem 1.5 in the general scheme of the Abstract Interpolation Problem (AIP) developed in [11, 14, 15]. In Section 5 we recall some needed results on AIP and then prove the first part of Theorem 1.6 in Section 6. Explicit formulas for the coefficients in the parametrization formula (1.29) are derived in Theorem 6.3. An explicit formula for the unique solution of Problem 1.5 in case P is singular is given in Theorem 6.2. In Section 6 we also prove certain properties of the coefficient matrix (1.30) which enable us to prove the second part of Theorem 1.6 in Section 7.

2. Preliminaries

The proof of Theorem 1.1 presented in [7] relies on the de Branges-Rovnyak spaces L^w and H^w associated to a Schur function w. In this section we recall some needed definitions and results. We use the standard notation L_2 for the Lebesgue space of square integrable functions on the unit circle $\mathbb{T}$; the symbols H_2^+ and H_2^- stand for the Hardy spaces of functions with vanishing negative (respectively, nonnegative) Fourier coefficients. The elements in H_2^+ and H_2^- will be identified with their unique analytic (resp., conjugate-analytic) continuations inside the unit disk, and consequently H_2^+ and H_2^- will be identified with the Hardy spaces of the unit disk.

Let w be a Schur function. The nontangential boundary limits $w(t)$ exist and are bounded by one at a.e. $t \in \mathbb{T}$ and the matrix-valued function $\left[\begin{smallmatrix} 1 & w(t) \\ w(t)^* & 1 \end{smallmatrix}\right]$ is defined and positive semidefinite almost everywhere on $\mathbb{T}$. The space L^w is the range space $\left[\begin{smallmatrix} 1 & w \\ w^* & 1 \end{smallmatrix}\right]^{1/2}(L_2 \oplus L_2)$ endowed with the range norm. The set of functions $\left[\begin{smallmatrix} 1 & w \\ w^* & 1 \end{smallmatrix}\right] f$ where $f \in L_2 \oplus L_2$ is dense in L^w and

$$\left\| \begin{bmatrix} 1 & w \\ w^* & 1 \end{bmatrix} f \right\|_{L^w}^2 = \left\langle \begin{bmatrix} 1 & w \\ w^* & 1 \end{bmatrix} f,\ f \right\rangle_{L_2 \oplus L_2}. \tag{2.1}$$

Definition 2.1. A function $f = \left[\begin{smallmatrix} f_+ \\ f_- \end{smallmatrix}\right]$ is said to belong to the de Branges–Rovnyak space H^w if it belongs to L^w and if $f_+ \in H_2^+$ and $f_- \in H_2^-$.

As it was shown in [7], the vector-valued functions

$$K_z^{(j)}(t) = \frac{1}{j!}\frac{\partial^j}{\partial \bar{z}^j}\left(\begin{bmatrix} 1 & w(t) \\ w(t)^* & 1 \end{bmatrix} \begin{bmatrix} 1 \\ -w(z)^* \end{bmatrix} \cdot \frac{1}{1 - t\bar{z}} \right) \tag{2.2}$$

defined for $z \in \mathbb{D}$, $t \in \mathbb{T}$ and $j \in \mathbb{Z}_+$, belong to the space H^w and furthermore, for every $z \in \mathbb{D}$ and every $f = \begin{bmatrix} f_+ \\ f_- \end{bmatrix} \in H^w$,

$$\left\langle f,\ K_z^{(j)} \right\rangle_{H^w} = \frac{1}{j!} \frac{\partial^j}{\partial z^j} f_+(z). \tag{2.3}$$

Setting $f = K_\zeta^{(i)}$ in (2.3), we get

$$\left\langle K_\zeta^{(i)},\ K_z^{(j)} \right\rangle_{H^w} = \frac{1}{j!i!} \frac{\partial^{j+i}}{\partial z^j \partial \bar{\zeta}^i} \left(\frac{1 - w(z)\overline{w(\zeta)}}{1 - z\bar{\zeta}} \right). \tag{2.4}$$

Upon differentiating in (2.2) and taking into account that $|t| = 1$, we come to the following explicit formulas for $K_z^{(j)}$:

$$K_z^{(j)}(t) = \begin{bmatrix} 1 & w(t) \\ w(t)^* & 1 \end{bmatrix} \begin{bmatrix} t^j(1 - t\bar{z})^{-j-1} \\ -\sum_{\ell=0}^{j} w_\ell(z)^* t^{j-\ell}(1 - t\bar{z})^{\ell-j-1} \end{bmatrix}, \tag{2.5}$$

where $w_\ell(z)$ are the Taylor coefficients from the expansion

$$w(\zeta) = \sum_{\ell=0}^{\infty} w_\ell(z)(\zeta - z)^\ell, \quad w_\ell(z) = \frac{w^{(\ell)}(z)}{\ell!}.$$

The two next theorems (also proved in [7]) explain the role of condition (1.9).

Theorem 2.2. *Let $w \in \mathcal{S}$, $t_0 \in \mathbb{T}$, $n \in \mathbb{Z}_+$ and let*

$$\liminf_{z \to t_0} d_{w,n}(z) < \infty. \tag{2.6}$$

Then the nontangential boundary limits

$$w_j(t_0) := \lim_{z \to t_0} \frac{w^{(j)}(z)}{j!} \quad \text{exist for } j = 0, \dots, n \tag{2.7}$$

and the functions

$$K_{t_0}^{(j)}(t) = \begin{bmatrix} 1 & w(t) \\ w(t)^* & 1 \end{bmatrix} \begin{bmatrix} t^j(1 - t\bar{t}_0)^{-j-1} \\ -\sum_{\ell=0}^{j} w_\ell(t_0)^* t^{j-\ell}(1 - t\bar{t}_0)^{\ell-j-1} \end{bmatrix} \tag{2.8}$$

belong to the space H^w for $j = 0, \dots, n$. Moreover, the kernels $K_z^{(j)}$ defined in (2.5) converge to $K_{t_0}^{(j)}$ for $j = 1, \dots, n$ in norm of H^w as $z \in \mathbb{D}$ approaches t_0 nontangentially:

$$K_z^{(j)} \xrightarrow{H^w} K_{t_0}^{(j)} \quad \text{for} \quad j = 1, \dots, n \quad \text{as } z \to t_0.$$

Theorem 2.3. *Let $w \in \mathcal{S}$, $t_0 \in \mathbb{T}$, $n \in \mathbb{Z}_+$. If the numbers $c_0, \dots, c_n$ are such that the function*

$$F(t) = \begin{bmatrix} 1 & w(t) \\ w(t)^* & 1 \end{bmatrix} \begin{bmatrix} t^n(1-t\bar{t}_0)^{-n-1} \\ -\sum_{\ell=0}^{n} c_\ell^* t^{n-\ell}(1-t\bar{t}_0)^{\ell-n-1} \end{bmatrix}$$

belongs to H^w, then condition (2.6) holds, the limits (2.7) exist, and $w_j(t_0) = c_j$ for $j = 0, \dots, n$; consequently, F coincides with $K_{t_0}^{(n)}$.

Now the preceding analysis can be easily extended to a multi-point setting. Given a Schur function w and k-tuples $\mathbf{z} = (z_1, \dots, z_k)$ of points in $\mathbb{D}$ and $\mathbf{n} = (n_1, \dots, n_k)$ of nonnegative integers, define the *generalized Schwarz-Pick matrix*

$$\mathbf{P}_{\mathbf{n}}^w(\mathbf{z}) := \left[\left[\frac{1}{\ell! r!} \frac{\partial^{\ell+r}}{\partial z^\ell \partial \bar{\zeta}^r} \left(\frac{1 - w(z)\overline{w(\zeta)}}{1 - z\bar{\zeta}} \right) \Bigg|_{\substack{z = z_i, \\ \zeta = z_j}} \right]_{\substack{\ell = 0, \dots, n_i \\ r = 0, \dots, n_j}} \right]_{i,j=1}^{k}. \tag{2.9}$$

Given a tuple $\mathbf{t} = (t_1, \dots, t_k)$ of distinct points $t_i \in \mathbb{T}$, define the boundary generalized Schwarz-Pick matrix

$$\mathbf{P}_{\mathbf{n}}^w(\mathbf{t}) := \lim_{\mathbf{z} \to \mathbf{t}} \mathbf{P}_{\mathbf{n}}^w(\mathbf{z}) \tag{2.10}$$

provided the latter limit exists, where $\mathbf{z} \to \mathbf{t}$ means that $z_i \in \mathbb{D}$ approaches t_i for $i = 1, \dots, k$ nontangentially. It is readily seen that conditions

$$\liminf_{z \to t_i} d_{w,n_i}(z) < \infty \quad \text{for } \ i = 1, \dots, k, \tag{2.11}$$

(where d_{w,n_i} is defined via formula (1.8)) are necessary for the limit (2.10) to exist. They are also sufficient as the next theorem shows.

Theorem 2.4. *Let $\mathbf{t} = (t_1, \dots, t_k)$ be a tuple of distinct points $t_i \in \mathbb{T}$, let $\mathbf{n} = (n_1, \dots, n_k) \in \mathbb{Z}_+^k$ and let w be a Schur function satisfying conditions (2.11). Then:*

1. *The following nontangential boundary limits exist:*

$$w_j(t_i) := \lim_{z \to t_i} \frac{w^{(j)}(z)}{j!} \quad (j = 0, \dots, 2n_i + 1; \ \ i = 1, \dots, k). \tag{2.12}$$

2. *The functions*

$$K_{t_i}^{(j)}(t) = \begin{bmatrix} 1 & w(t) \\ w(t)^* & 1 \end{bmatrix} \begin{bmatrix} t^j(1-t\bar{t}_i)^{-j-1} \\ -\sum_{\ell=0}^{j} w_\ell(t_i)^* t^{j-\ell}(1-t\bar{t}_i)^{\ell-j-1} \end{bmatrix} \tag{2.13}$$

belong to the space H^w for $j = 0, \dots, n_i$ and $i = 1, \dots, k$.

3. *The boundary generalized Schwarz-Pick matrix* $\mathbf{P}_{\mathbf{n}}^{w}(\mathbf{t})$ *defined via the nontangential limit* (2.10) *exists and is equal to the Gram matrix of the set* $\{K_{t_i}^{(j)} : j = 0, \dots, n_i;\ i = 1, \dots, k\}$*:*

$$\mathbf{P}_{\mathbf{n}}^{w}(\mathbf{t}) := \left[\left[\left\langle K_{t_j}^{(r)},\ K_{t_i}^{(\ell)} \right\rangle_{H^w} \right]_{\substack{\ell = 0, \dots, n_i \\ r = 0, \dots, n_j}} \right]_{i,j=1}^{k}. \tag{2.14}$$

4. *The matrix* $\mathbf{P}_{\mathbf{n}}^{w}(\mathbf{t})$ *can be expressed in terms of the nontangential limits* (2.12) *as follows:*

$$\mathbf{P}_{\mathbf{n}}^{w}(\mathbf{t}) = \left[\mathbf{P}_{ij}^{w}\right]_{i,j=1}^{k} \tag{2.15}$$

where $\mathbf{P}_{ij}^{w}$ *is the* $(n_i + 1) \times (n_j + 1)$ *matrix defined by*

$$\mathbf{P}_{ij}^{w} = \mathbf{H}_{ij} \mathbf{\Psi}_{n_j}(t_j) \mathbf{W}_j^{*}, \tag{2.16}$$

where $\mathbf{\Psi}_{n_j}(t_j)$ *is defined as in* (1.6), $\mathbf{W}_j$ *is the lower triangular Toeplitz matrix given by*

$$\mathbf{W}_j = \begin{bmatrix} w_0(t_j) & 0 & \dots & 0 \\ w_1(t_j) & w_0(t_i) & \ddots & \vdots \\ \vdots & \ddots & \ddots & 0 \\ w_{n_j}(t_j) & \dots & w_1(t_j) & w_0(t_j) \end{bmatrix}, \tag{2.17}$$

and where $\mathbf{H}_{ij}$ *is the matrix with the entries*

$$\begin{aligned} [\mathbf{H}_{ij}]_{r,s} \quad &= \quad \sum_{\ell=0}^{r} (-1)^{r-\ell} \begin{pmatrix} s + r - \ell \\ s \end{pmatrix} \frac{w_\ell(t_i)}{(t_i - t_j)^{s+r-\ell+1}} \\ &\quad - \sum_{\ell=0}^{s} (-1)^{r} \begin{pmatrix} s + r - \ell \\ r \end{pmatrix} \frac{w_\ell(t_j)}{(t_i - t_j)^{s+r-\ell+1}}. \end{aligned} \tag{2.18}$$

if $i \neq j$*, and it is the Hankel matrix*

$$\mathbf{H}_{jj} = \begin{bmatrix} w_1(t_j) & w_2(t_j) & \cdots & w_{n_j+1}(t_j) \\ w_2(t_j) & w_3(t_j) & \cdots & w_{n+2}(t_j) \\ \vdots & \vdots & & \vdots \\ w_{n_j+1}(t_j) & w_{n_j+2}(t_i) & \cdots & w_{2n_j+1}(t_j) \end{bmatrix} \tag{2.19}$$

otherwise.

Proof. The two first statements follow by Theorems 1.1 and 2.2. Due to relation (2.4), the matrix in (2.9) can be written as

$$\mathbf{P}_{\mathbf{n}}^{w}(\mathbf{z}) := \left[\left[\left\langle K_{z_j}^{(r)},\ K_{z_i}^{(\ell)} \right\rangle_{H^w} \right]_{\substack{\ell = 0, \dots, n_i \\ r = 0, \dots, n_j}} \right]_{i,j=1}^{k}. \tag{2.20}$$

By Statement 3 in Theorem 2.2,

$$K_{z_i}^{(j)} \xrightarrow{H^w} K_{t_i}^{(j)} \quad \text{for} \quad j = 1, \dots, n_i;\ \ i = 1, \dots, k,$$

as z_i approaches t_i nontangentially. Passing to the limit in (2.20) we get the existence of the boundary generalized Schwarz-Pick matrix $\mathbf{P}_{\mathbf{n}}^{w}(\mathbf{t})$ and obtain its representation (2.14). Let us consider the block partitioning

$$\mathbf{P}_{\mathbf{n}}^{w}(\mathbf{z}) = \left[\mathbf{P}_{ij}^{w}(z_i, z_j)\right]_{i,j=1}^{k}$$

conformal with that in (2.15) so that

$$\mathbf{P}_{ij}^{w}(z_i, z_j) := \left[\begin{array}{c} \dfrac{1}{\ell! r!}\ \dfrac{\partial^{\ell+r}}{\partial z^{\ell}\partial\bar{\zeta}^{r}}\left(\dfrac{1-w(z)\overline{w(\zeta)}}{1-z\bar{\zeta}}\right)\Bigg|_{\substack{z=z_i,\\ \zeta=z_j}} \end{array}\right]_{\substack{\ell=0,\dots,n_i\\ r=0,\dots,n_j}}. \tag{2.21}$$

The direct differentiation in (2.21) gives

$$\left[\mathbf{P}_{ij}^{w}(z_i, z_j)\right]_{\ell,r} = \sum_{s=0}^{\min\{\ell,r\}} \frac{(\ell+r-s)!}{(\ell-s)!(r-s)!}\frac{z_i^{r-s}\bar{z}_j^{\ell-s}}{(1-z_i\bar{z}_j)^{\ell+r-s+1}} - \sum_{\alpha=0}^{\ell}\sum_{\beta=0}^{r}\sum_{s=0}^{\min\{\alpha,\beta\}} \frac{(\alpha+\beta-s)!}{(\alpha-s)!(\beta-s)!}\frac{z_i^{\beta-s}\bar{z}_j^{\alpha-s}w_{\ell-\alpha}(z_i)w_{r-\beta}(z_j)^*}{(1-z_i\bar{z}_j)^{\alpha+\beta-s+1}}.$$

For $i \neq j$, we pass to the limit in the latter equality as $z_i \to t_i$ and $z_j \to t_j$ and take into account (2.12):

$$\left[\mathbf{P}_{ij}^{w}\right]_{\ell,r} = \sum_{s=0}^{\min\{\ell,r\}} \frac{(\ell+r-s)!}{(\ell-s)!(r-s)!}\frac{t_i^{r-s}\bar{t}_j^{\ell-s}}{(1-t_i\bar{t}_j)^{\ell+r-s+1}} - \sum_{\alpha=0}^{\ell}\sum_{\beta=0}^{r}\sum_{s=0}^{\min\{\alpha,\beta\}} \frac{(\alpha+\beta-s)!}{(\alpha-s)!(\beta-s)!}\frac{t_i^{\beta-s}\bar{t}_j^{\alpha-s}w_{\ell-\alpha}(t_i)w_{r-\beta}(t_j)^*}{(1-t_i\bar{t}_j)^{\alpha+\beta-s+1}}.$$

Verification of the fact that the product on the right hand side of (2.15) gives the matrix with the same entries, is straightforward and will be omitted. Finally, it is readily seen from (2.21) and (1.2) that the j-th diagonal block $\mathbf{P}_{jj}^{w}(z_j, z_j)$ coincides with the Schwarz-Pick matrix $\mathbf{P}_{n_j}^{w}(z_j)$. Therefore, by Theorem 1.1 and formula (1.5), its nontangential boundary limit equals

$$\begin{aligned} \mathbf{P}_{jj}^{w} &= \mathbb{P}_{n_j}^{w}(t_j) \qquad (2.22)\\ &= \begin{bmatrix} w_1(t_j) & \cdots & w_{n_j+1}(t_j)\\ \vdots & & \vdots\\ w_{n_j+1}(t_j) & \cdots & w_{2n_j+1}(t_j)\end{bmatrix} \mathbf{\Psi}_{n_j}(t_j) \begin{bmatrix} w_0(t_j)^* & \dots & w_{n_j}(t_j)^*\\ & \ddots & \vdots\\ 0 & & w_0(t_j)^*\end{bmatrix}, \end{aligned}$$

which coincides with (2.16) for $j = i$. □

3. The Pick matrix and the Stein identity

The Pick matrix P defined and studied in this section is important for formulating a solvability criterion for Problem 1.5 and for parametrizing its solution set. The

definition of the Pick matrix is motivated by the formulas for the matrix $\mathbf{P}_{\mathbf{n}}^{w}(\mathbf{t})$ discussed in the previous section. Namely,

$$P = [P_{ij}]_{i,j=1}^{k} \in \mathbb{C}^{N\times N} \quad \text{where} \quad N = \sum_{i=1}^{k}(n_i+1), \tag{3.1}$$

and the block entries $P_{ij} \in \mathbb{C}^{(n_i+1)\times(n_j+1)}$ are defined by

$$P_{ij} = H_{ij} \cdot \mathbf{\Psi}_{n_j}(t_j) \cdot W_j^*, \tag{3.2}$$

where $\mathbf{\Psi}_{n_j}(t_j)$ is defined as in (1.6), where

$$W_i = \begin{bmatrix} c_{i,0} & 0 & \dots & 0 \\ c_{i,1} & c_{i,0} & \dots & 0 \\ \vdots & \ddots & \ddots & \vdots \\ c_{i,n_i} & \dots & c_{i,1} & c_{i,0} \end{bmatrix}, \tag{3.3}$$

$$H_{ii} = \begin{bmatrix} c_{i,1} & c_{i,2} & \cdots & c_{i,n_i+1} \\ c_{i,2} & c_{i,3} & \cdots & c_{i,n+2} \\ \vdots & \vdots & & \vdots \\ c_{i,n_i+1} & c_{i,n_i+2} & \cdots & c_{i,2n_i+1} \end{bmatrix} \tag{3.4}$$

for $i = 1, \dots, k$ and where the matrices H_{ij} (for $i \neq j$) are defined entrywise by

$$\begin{aligned} [H_{ij}]_{r,s} &= \sum_{\ell=0}^{r}(-1)^{r-\ell}\begin{pmatrix} s+r-\ell \\ s \end{pmatrix}\frac{c_{i,\ell}}{(t_i-t_j)^{s+r-\ell+1}} \\ &\quad -\sum_{\ell=0}^{s}(-1)^{r}\begin{pmatrix} s+r-\ell \\ r \end{pmatrix}\frac{c_{j,\ell}}{(t_i-t_j)^{s+r-\ell+1}} \end{aligned} \tag{3.5}$$

for $r = 0, \dots, n_i$ and $s = 0, \dots, n_j$. The latter formulas define P exclusively in terms of the interpolation data of (1.25). We also associate with the same data the following matrices:

$$T = \begin{bmatrix} T_1 & & 0 \\ & \ddots & \\ 0 & & T_k \end{bmatrix}, \qquad \text{where} \quad T_i = \begin{bmatrix} \bar{t}_i & 1 & \dots & 0 \\ 0 & \bar{t}_i & & \vdots \\ \vdots & \ddots & \ddots & 1 \\ 0 & \dots & 0 & \bar{t}_i \end{bmatrix}, \tag{3.6}$$

$$E = \begin{bmatrix} E_1 & \dots & E_k \end{bmatrix}, \qquad \text{where} \quad E_i = \begin{bmatrix} 1 & 0 & \dots & 0 \end{bmatrix}, \tag{3.7}$$

$$M = \begin{bmatrix} M_1 & \dots & M_k \end{bmatrix}, \qquad \text{where} \quad M_i = \begin{bmatrix} c_{i,0}^* & \dots & c_{i,n_i}^* \end{bmatrix}. \tag{3.8}$$

Note that $T_i \in \mathbb{C}^{(n_i+1)\times(n_i+1)}$ and E_i, $M_i \in \mathbb{C}^{1\times(n_i+1)}$. The main result of this section is:

Theorem 3.1. *Let $|c_{i,0}| = 1$ for $i = 1, \dots, k$ and let us assume that the diagonal blocks P_{ii} of the matrix P defined in* (3.1)–(3.5) *are Hermitian for $i = 1, \dots, k$.*

Then the matrix P is Hermitian and satisfies the Stein identity

$$P - T^*PT = E^*E - M^*M, \tag{3.9}$$

where the matrices T, E and M are defined in (3.6)–(3.8).

In view of (3.6)–(3.8), verifying (3.9) is equivalent to verifying

$$P_{ij} - T_i^*P_{ij}T_j = E_i^*E_j - M_i^*M_j \quad (i,j=1,\dots,k). \tag{3.10}$$

An identity like that is not totally surprising due to a special (Hankel and Toeplitz) structure of the factors H_{ij} and W_j in (3.2). Indeed, the identity verified in the next lemma (though, not exactly of the form (3.10)) follows from the structure of P_{ij} only (without any symmetry assumptions). Note that the right-hand side in (3.9) as well as the one in (3.10) is of rank 2 .

Lemma 3.2. *Let P_{ij} be defined as in* (3.2). *Then*

$$P_{ij} - T_i^*P_{ij}T_j = E_i^*\overline{M}_j\mathbf{\Psi}_{n_j}(t_j)W_j^*T_j - M_i^*M_j \tag{3.11}$$

where, according to (3.8),

$$\overline{M}_j = \begin{bmatrix} c_{j,0} & c_{j,1} & \dots & c_{j,n_j} \end{bmatrix} = E_jW_j^\top.$$

Proof. We shall make use of the equalities

$$W_j^*T_j = T_jW_j^*, \quad \overline{T}_j\mathbf{\Psi}_{n_j}(t_j)T_j = \mathbf{\Psi}_{n_j}(t_j), \quad E_j\mathbf{\Psi}_{n_j}(t_j)T_j = E_j. \tag{3.12}$$

The first equality follows by the Toeplitz triangular structure of W_j^* and T_j. The matrix $\overline{T}_j\mathbf{\Psi}_{n_j}(t_j)T_j$ is upper triangular as the product of upper triangular matrices, and due to (1.7) and (3.6), its $s\ell$-th entry (for $\ell \geq s$) equals

$$\begin{aligned}
\left[\overline{T}_j\mathbf{\Psi}_{n_j}(t_j)T_j\right]_{s,\ell} &= \Psi_{s,\ell} + t_j\Psi_{s,\ell-1} + \bar{t}_j\Psi_{s+1,\ell} + \Psi_{s+1,\ell-1} \\
&= (-1)^\ell t_j^{s+\ell+1}\left[\begin{pmatrix} \ell \\ s \end{pmatrix} - \begin{pmatrix} \ell-1 \\ s \end{pmatrix} + \begin{pmatrix} \ell \\ s+1 \end{pmatrix} - \begin{pmatrix} \ell-1 \\ s+1 \end{pmatrix}\right] \\
&= (-1)^\ell t_j^{s+\ell+1}\begin{pmatrix} \ell \\ s \end{pmatrix} = \Psi_{s,\ell}.
\end{aligned}$$

This completes the verification of the second equality in (3.12). The last relation in (3.12) follows by (1.7) and (3.6) and (3.7):

$$E_j\mathbf{\Psi}_{n_j}(t_j)T_j = \begin{bmatrix} t_j & -t_j^2 & \dots & (-1)^{n_j}t_j^{n_j+1} \end{bmatrix}T_j = \begin{bmatrix} 1 & 0 & \dots & 0 \end{bmatrix} = E_j.$$

We will also use the identity

$$H_{ij}\overline{T}_j - T_i^*H_{ij} = E_i^*\overline{M}_j - M_i^*E_j \tag{3.13}$$

which holds for every $i,j=1,\dots,k$ and is verified by straightforward calculations (separately for the cases $i=j$ and $i\neq j$). We have

$$\begin{aligned}
P_{ij} - T_i^*P_{ij}T_j &= H_{ij}\mathbf{\Psi}_{n_j}(t_j)W_i^* - T_i^*H_{ij}\mathbf{\Psi}_{n_j}(t_j)T_jW_j^* \\
&= \left(H_{ij}\overline{T}_j - T_j^*H_{ij}\right)\mathbf{\Psi}_{n_j}(t_j)T_jW_j^* \\
&= \left(E_i^*\overline{M}_j - M_i^*E_j\right)\mathbf{\Psi}_{n_j}(t_j)T_jW_j^*,
\end{aligned} \tag{3.14}$$

where the first equality follows by (3.2) and the first relation in (3.12), the second equality relies on the second relation in (3.12) and the last equality is a consequence of (3.13). Combining the third relation in (3.12) with formulas (3.3) and (3.8) we get

$$E_j \mathbf{\Psi}_{n_i}(t_i) T_j W_j^* = E_j W_j^* = M_j$$

which being substituted into (3.14) leads us to (3.11). □

Proof of Theorem 3.1. By Lemma 3.2 the structure of P implies (3.11). First we consider the case when $j = i$. Since, by assumption, matrices P_{ii} $(i = 1, \dots, k)$ are Hermitian, the left-hand sides in (3.11) are Hermitian, and hence the right-hand sides in (3.11) must be Hermitian. In other words,

$$E_i^* \overline{M}_i \mathbf{\Psi}_{n_i}(t_i) W_i^* T_i = (\overline{M}_i \mathbf{\Psi}_{n_i}(t_i) W_i^* T_i)^* E_i \quad \text{for } i = 1, \dots, k.$$

Multiplying the latter relation by E_i from the left and taking into account that

$$E_i E_i^* = 1 \quad \text{and} \quad E_i (\overline{M}_i \mathbf{\Psi}_{n_i}(t_i) W_i^* T_i)^* = (c_{i,0} t_i c_{i,0}^* \overline{t}_i)^* = 1,$$

we get

$$\overline{M}_i \mathbf{\Psi}_{n_i}(t_i) W_i^* T_i = E_i \quad \text{for } i = 1, \dots, k.$$

Therefore, relations (3.11) turn into (3.10), which is equivalent to (3.9). Furthermore, for $i \neq j$, the Stein equation

$$X - T_i^* X T_j = E_i^* E_j - M_i^* M_j$$

has a unique solution X. Taking adjoint of both sides in (3.10) we conclude that the matrix P_{ij}^* satisfies the same Stein equation as P_{ji} does and then, by the above uniqueness, $P_{ij}^* = P_{ji}$ for $i \neq j$. It follows now that P is Hermitian. □

Theorem 3.3. *Let $t_1, \dots, t_k \in \mathbb{T}$, $n_1, \dots, n_k \in \mathbb{Z}_+$, $N = \sum_{i=1}^k (n_1 + 1)$ and let us assume that a Schur function w satisfies conditions* (2.11)*. Then the matrix $\mathbf{P}_{\mathbf{n}}^w(\mathbf{t})$ defined via the limit* (2.10) *(that exists by Theorem* 3.1*) satisfies the Stein identity*

$$\mathbf{P}_{\mathbf{n}}^w(\mathbf{t}) - T^* \mathbf{P}_{\mathbf{n}}^w(\mathbf{t}) T = E^* E - (M^w)^* M^w, \tag{3.15}$$

where the matrices T and E are defined in (3.6), (3.7) *and*

$$M^w = \begin{bmatrix} M_1 & \dots & M_k \end{bmatrix}, \quad \text{where} \quad M_i^w = \begin{bmatrix} w_0(t_i)^* & \dots & w_{n_i}(t_i)^* \end{bmatrix}. \tag{3.16}$$

Proof. By Theorem 2.2, the matrix $\mathbf{P}_{\mathbf{n}}^w(\mathbf{t})$ admits the representation (2.15)–(2.18), that has the same structure as the Pick matrix P constructed in (3.1)–(3.5) but with parameters c_{ij} replaced by $w_j(t_i)$. Furthermore, it is positive semidefinite (and therefore, its diagonal blocks are Hermitian) due to representation (2.14), whereas $|w_0(t_i)| = 1$ for $i = 1, \dots, k$, by Theorem 1.1. Upon applying Theorem 3.1 we conclude that $\mathbf{P}_{\mathbf{n}}^w(\mathbf{t})$ satisfies the same Stein identity as P but with M^w instead of M, i.e., the Stein identity (3.15). □

4. Reformulation of Problem 1.5

The formula (2.14) for $\mathbf{P}_{\mathbf{n}}^{w}(\mathbf{t})$ motivates us to introduce the matrix function

$$\widetilde{\mathbf{F}}^{w}(t) = \begin{bmatrix} \widetilde{\mathbf{F}}_1^{w}(t) & \dots & \widetilde{\mathbf{F}}_k^{w}(t) \end{bmatrix}, \tag{4.1}$$

where

$$\widetilde{\mathbf{F}}_i^{w}(t) := \begin{bmatrix} K_{t_i}^{(0)}(t) & K_{t_i}^{(1)}(t) & \dots & K_{t_i}^{(n_i)}(t) \end{bmatrix} \quad (i = 1, \dots, k), \tag{4.2}$$

and $K_{t_i}^{(j)}(t)$ $(j = 0, \dots, n_i)$ are the functions defined in (2.13).

Theorem 4.1. *Let $t_1, \dots, t_k \in \mathbb{T}$, $n_1, \dots, n_k \in \mathbb{Z}_+$ and let us assume that a Schur function w satisfies conditions* (2.11)*. Then for $\widetilde{\mathbf{F}}^{w}$ defined in* (4.2), (4.3) *we have:*

1. *The function $\widetilde{\mathbf{F}}^{w}x$ belongs to the de Branges-Rovnyak space H^w for every vector $x \in \mathbb{C}^N$ and*
$$\|\widetilde{\mathbf{F}}^{w}x\|_{H^w}^2 = x^* \mathbf{P}_{\mathbf{n}}^{w}(\mathbf{t})x \tag{4.3}$$
where $\mathbf{P}_{\mathbf{n}}^{w}(\mathbf{t})$ is the boundary generalized Schwarz-Pick matrix (that exists due to conditions (2.11)*) and $N := \sum_{i=1}^{k}(n_i + 1)$.*
2. *$\widetilde{\mathbf{F}}^{w}$ admits the representation*
$$\widetilde{\mathbf{F}}^{w}(t) = \begin{bmatrix} 1 & w(t) \\ w(t)^* & 1 \end{bmatrix} \begin{bmatrix} E \\ -M^w \end{bmatrix} (\mathrm{I} - tT)^{-1}, \tag{4.4}$$

where the matrices T, E and M^w are defined in (3.6), (3.7) *and* (3.16)*, respectively.*

Proof. By Theorem 1.1, conditions (1.26) guarantee that the functions $K_{t_i}^{(j)}$ defined in (2.13) belong to H^w and the boundary Schwarz-Pick matrix $\mathbf{P}_{\mathbf{n}}^{w}(\mathbf{t})$ exists and admits a representation (2.14). Now it follows from (4.1) and (4.2) that for every $x \in \mathbb{C}^n$, the function $\mathbf{F}^w x$ belongs to H^w as a linear combination of the kernels $K_{t_i}^{(j)} \in H^w$, while relation (4.3) is an immediate consequence of (2.14). Furthermore, by definitions (3.6), (3.7) and (3.16) of T_i, E_i and M_i^w,

$$\begin{bmatrix} E_i \\ -M_i^w \end{bmatrix} (\mathrm{I} - tT_i)^{-1} = \begin{bmatrix} \dfrac{1}{1 - t\bar{t}_i} & \cdots & \dfrac{t^{n_i}}{(1 - t\bar{t}_i)^{n_i+1}} \\ -\dfrac{w_0(t_i)^*}{1 - t\bar{t}_i} & \cdots & -\displaystyle\sum_{\ell=0}^{n_i} \dfrac{w_\ell(t_i)^* t^{n_i - \ell}}{(1 - t\bar{t}_i)^{n_i+1-\ell}} \end{bmatrix}. \tag{4.5}$$

Multiplying both sides of (4.5) by the matrix $\begin{bmatrix} 1 & w(t) \\ w(t)^* & 1 \end{bmatrix}$ on the left and taking into account (2.13) and (4.2) we get

$$\begin{bmatrix} 1 & w(t) \\ w(t)^* & 1 \end{bmatrix} \begin{bmatrix} E \\ -M^w \end{bmatrix} (\mathrm{I} - tT)^{-1} = \begin{bmatrix} K_{t_i}^{(0)}(t) & K_{t_i}^{(1)}(t) & \dots & K_{t_i}^{(n_i)}(t) \end{bmatrix}$$
$$=: \widetilde{\mathbf{F}}_i^{w}(t) \qquad (i = 1, \dots, k). \tag{4.6}$$

Now representation formula (4.4) follows by definitions (block partitionings) (4.1), (3.6), (3.7) and (3.16) of $\widetilde{\mathbf{F}}^{w}$, T, E and M^w. □

Now we modify $\widetilde{\mathbf{F}}^w$ replacing M^w by M in (4.4): we introduce the function

$$\mathbf{F}^w(t) := \begin{bmatrix} 1 & w(t) \\ w(t)^* & 1 \end{bmatrix} \begin{bmatrix} E \\ -M \end{bmatrix} (\mathrm{I} - tT)^{-1} \tag{4.7}$$

with T, E and M defined in (3.6)–(3.8). The two next theorems show that Problem 1.5 can be reformulated in terms of this function and of the Pick matrix P.

Theorem 4.2. *Assume that w solves Problem* 1.5 *(i.e., $w \in \mathcal{S}$ and satisfies interpolation conditions* (1.26)–(1.28)*) and let $\mathbf{F}^w$ be defined as in* (4.7)*. Then*

1. *The function $\mathbf{F}^w x$ belongs to H^w for every vector $x \in \mathbb{C}^N$ and*
$$\|\mathbf{F}^w x\|^2_{H^w} \le x^* P x \tag{4.8}$$
where P is the Pick matrix defined in (3.1)–(3.5)*.*
2. *The numbers $c_{i,0}$ are unimodular for $i = 1, \dots, k$ and the matrix P is positive semidefinite,*
$$|c_{i,0}| = 1 \ \ (i = 1, \dots, k) \quad \textit{and} \quad P \ge 0. \tag{4.9}$$
3. *P satisfies the Stein identity* (3.9)*.*

Furthermore, if w is a solution of Problem 1.2*, then*

$$\|\mathbf{F}^w x\|^2_{H^w} = x^* P x \quad \textit{for every } \ x \in \mathbb{C}^N. \tag{4.10}$$

Proof. Conditions (1.26) guarantee (by Theorem 1.1) that the limits $w_0(t_i)$ are unimodular for $i = 1, \dots, k$; since $w_0(t_i) = c_{i,0}$ (according to (1.27)), the first condition in (4.9) follows.

Conditions (1.26) also guarantee (by Theorem 4.1), that for every $x \in \mathbb{C}^N$, the function $\widetilde{\mathbf{F}}^w x$ belongs to H^w for every vector $x \in \mathbb{C}^N$, and equality (4.3) holds, where $\widetilde{\mathbf{F}}^w$ is defined by the representation formula (4.4). On account of interpolation conditions (1.27) (only for $j = 0, \dots, n_i$ and for every $i = 1, \dots, k$) and by definitions (3.8) and (3.16), it follows that $M = M^w$. Then the formulas (4.4) and (4.7) show that $\mathbf{F}^w \equiv \widetilde{\mathbf{F}}^w$, so that equality (4.3) holds with $\mathbf{F}^w$ instead of $\widetilde{\mathbf{F}}^w$:

$$\|\mathbf{F}^w x\|^2_{H^w} = x^* \mathbf{P}^w_{\mathbf{n}}(\mathbf{t}) x. \tag{4.11}$$

Thus, to prove (4.8), it suffices to show that $\mathbf{P}^w_{\mathbf{n}}(\mathbf{t}) \le P$. We will use formulas (2.15)–(2.19) defining $\mathbf{P}^w_{\mathbf{n}}(\mathbf{t})$ in terms of the boundary limits $w_j(t_i)$. In view of these formulas and due to interpolation conditions (1.27), $\mathbf{P}^w_{\mathbf{n}}(\mathbf{t})$ can be expressed in terms of the interpolation data (1.25). Indeed, comparing (3.2)–(3.5) and (2.15)–(2.18) we conclude that

$$\mathbf{P}^w_{ij} = P_{ij} \quad (i \ne j) \tag{4.12}$$

and that formula (2.22) for the diagonal blocks of $\mathbf{P}^w$ turns into

$$\mathbf{P}^w_{ii} = \begin{bmatrix} c_{i,1} & \cdots & c_{i,n_i+1} \\ c_{i,2} & \cdots & c_{i,n_i+2} \\ \vdots & & \vdots \\ c_{i,n_i+1} & \cdots & w_{2n_i+1}(t_i) \end{bmatrix} \mathbf{\Psi}_{n_i}(t_i) \begin{bmatrix} c^*_{i,0} & c^*_{i,1} & \cdots & c^*_{i,n_i} \\ 0 & c^*_{i,0} & \cdots & c^*_{i,n_i-1} \\ \vdots & \ddots & \ddots & \vdots \\ 0 & \dots & 0 & c^*_{i,0} \end{bmatrix}. \tag{4.13}$$

Taking into account the upper triangular structure of $\mathbf{\Psi}_{n_i}(t_i)$, we conclude from (3.2), (3.3) and (4.13) that all the corresponding entries in P_{ii} and $\mathbf{P}^w_{ii}$ are equal except for the rightmost bottom entries that are equal to γ_i and to $d_{w,n_i}(t_i)$, respectively. Thus, by condition (1.26),

$$P_{ii} - \mathbf{P}^w_{ii} = \begin{bmatrix} 0 & \dots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \dots & \gamma_i - d_{w,n_i}(t_i) \end{bmatrix} \geq 0, \tag{4.14}$$

for $i = 1, \dots, k$ which together with (4.12) imply $P \geq \mathbf{P}^w$ and therefore, relation (4.8). If w is a solution of Problem 1.2 (or equivalently, of Problem 1.3), then $\gamma_i - d_{w,n_i}(t_i) = 0$ for $i = 1, \dots, k$ in (4.14) which proves the final statement in the theorem. Since $\mathbf{P}^w \geq 0$, we conclude from the inequality $P \geq \mathbf{P}^w$ that $P \geq 0$ which completes the proof of the second statement of the theorem. The third statement follows from (4.9) by Theorem 3.1. □

The next theorem is the converse to Theorem 4.2.

Theorem 4.3. *Let P, T, E and M be the matrices given by* (3.1)–(3.8). *Let $|c_{i,0}| = 1$ and $P \geq 0$. Let w be a Schur function such that*

$$\mathbf{F}^w x := \begin{bmatrix} 1 & w(t) \\ w(t)^* & 1 \end{bmatrix} \begin{bmatrix} E \\ -M \end{bmatrix} (\mathrm{I} - tT)^{-1} x \quad \textit{belongs to } H^w \tag{4.15}$$

for every $x \in \mathbb{C}^N$ and satisfies (4.8). *Then w is a solution of Problem* 1.5. *If moreover,* (4.10) *holds, then w is a solution of Problem* 1.2.

Proof. By the definitions (3.6)–(3.8) of T, E and M, the columns of the $2 \times N$ matrix $\mathbf{F}^w$ defined in (4.7), are of the form

$$\begin{bmatrix} 1 & w(t) \\ w(t)^* & 1 \end{bmatrix} \begin{bmatrix} t^j(1 - t\bar{t}_i)^{-j-1} \\ -\sum_{\ell=0}^{j} c^*_{i,\ell} t^{j-\ell}(1 - t\bar{t}_i)^{\ell-j-1} \end{bmatrix}$$

for $j = 1, \dots, n_i$ and $i = 1, \dots, k$, and all of them belong to H^w by the assumption (4.15) of the theorem. In particular, the functions

$$F_i(t) = \begin{bmatrix} 1 & w(t) \\ w(t)^* & 1 \end{bmatrix} \begin{bmatrix} t^{n_i}(1 - t\bar{t}_i)^{-n_i-1} \\ -\sum_{\ell=0}^{n_i} c^*_{i,\ell} t^{n_i-\ell}(1 - t\bar{t}_i)^{\ell-n_i-1} \end{bmatrix}$$

belong to H^w, which implies, by Theorems 3.2 and 2.4, that

$$\liminf_{z \to t_i} d_{w,n_i}(z) < \infty \quad \text{for} \quad i = 1, \dots, k, \tag{4.16}$$

and that the nontangential limits (2.12) exist and satisfy

$$w_j(t_i) = c_{ij} \quad \text{for} \quad j = 1, \dots, n_i \quad \text{and} \quad i = 1, \dots, k. \tag{4.17}$$

Therefore, w meets conditions (1.27) for $i = 1, \dots, k$ and $\ell_i = 0, \dots, n_i$. By Theorem 4.1, conditions (4.16) guarantee that the boundary generalized Schwarz-Pick matrix $\mathbf{P}_{\mathbf{n}}^{w}(\mathbf{t})$ exists and that

$$\|\widetilde{\mathbf{F}}^{w} x\|_{H^w}^2 = x^* \mathbf{P}_{\mathbf{n}}^{w}(\mathbf{t}) x \quad \text{for every } \ x \in \mathbb{C}^N, \tag{4.18}$$

where $\widetilde{\mathbf{F}}^{w}$ is the $2 \times N$ matrix function defined in (4.4). By Theorem 2.4, $\mathbf{P}_{\mathbf{n}}^{w}(\mathbf{t})$ is represented in terms of the boundary limits (2.12) by formulas (2.15)–(2.18). Equalities (4.17) along with definitions (3.8) and (3.16) of M and M^w show that the two latter matrices are equal and thus $\mathbf{F}^w \equiv \widetilde{\mathbf{F}}^w$, by (4.4) and (4.7). Now combining (4.18) and (4.15) gives $\mathbf{P}_{\mathbf{n}}^{w}(\mathbf{t}) \le P$ which implies inequalities for the diagonal blocks

$$\mathbf{P}_{ii}^{w} \le P_{ii} \quad (i = 1, \dots, k). \tag{4.19}$$

Since $d_{w,n_i}(t_i)$ and γ_i are (the lower) diagonal entries in $\mathbf{P}_{ii}^{w}$ and P_{ii}, respectively, the latter inequality implies (1.28).

By Theorems 3.1 and 3.3, the matrices P and $\mathbf{P}_{\mathbf{n}}^{w}(\mathbf{t})$ possess the Stein identities (3.9) and (3.15), respectively; since $M = M^w$, the matrix $\widetilde{P} := P - \mathbf{P}_{\mathbf{n}}^{w}(\mathbf{t})$ satisfies the homogeneous Stein identity

$$\widetilde{P} - T^* \widetilde{P} T = 0.$$

By the diagonal structure (3.6) of T and in view of (4.19) we have for the diagonal blocks $\widetilde{P}_{ii}$ of $\widetilde{P}$,

$$\widetilde{P}_{ii} - T_i^* \widetilde{P}_{ii} T_i = 0 \quad \text{and} \quad \widetilde{P}_{ii} \ge 0 \quad (i = 1, \dots, k). \tag{4.20}$$

By the Jordan structure (3.6) of T_i, it follows from (4.20) that $\widetilde{P}_{ii}$ is necessarily of the form

$$\widetilde{P}_{ii} = P_{ii} - \mathbf{P}_{ii}^{w} = \begin{bmatrix} 0 & \dots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \dots & \delta_i \end{bmatrix} \quad \text{with} \quad \delta_i \ge 0 \tag{4.21}$$

(for a simple proof see, e.g., [6, Corollary 10.7]). On the other hand, by the representations (2.16) and (3.2),

$$\mathbf{P}_{ii}^{w} = \mathbf{H}_{ii} \mathbf{\Psi}_{n_i}(t_i) \mathbf{W}_i^* \quad \text{and} \quad P_{ii} = H_{ii} \mathbf{\Psi}_{n_i}(t_i) W_i^*$$

and since by (4.17), $\mathbf{W}_i = W_i$ (which is readily seen from the definitions (2.17) and (3.3)), we conclude that

$$P_{ii} - \mathbf{P}_{ii}^{w} = (H_{ii} - \mathbf{H}_{ii})\, \mathbf{\Psi}_{n_i}(t_i) W_i^*.$$

Combining the last equality with (4.22) gives

$$H_{ii} - \mathbf{H}_{ii} = \begin{bmatrix} 0 & \dots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \dots & \delta_i \end{bmatrix} \left(\mathbf{\Psi}_{n_i}(t_i) W_i^* \right)^{-1}. \tag{4.22}$$

Since $|c_{i,0}| = 1$, it is seen from definitions (1.6) and (3.3) that the matrix $\mathbf{\Psi}_{n_i}(t_i)W_i^*$ is upper triangular and invertible and that its lower diagonal entry equals

$$g_i := (-1)^{n_i} t_i^{2n_i+1} c_{i,0}^*. \tag{4.23}$$

Therefore, the inverse matrix $(\mathbf{\Psi}_{n_i}(t_i)W_i^*)^{-1}$ is upper triangular with the lower diagonal entry equal g_i^{-1} so that the matrix on the right-hand side in (4.22) has all the entries equal to zero except the lower diagonal entry which is equal to $\delta_i g_i^{-1}$. Taking into account the definitions (2.19) and (3.4) we write (4.22) more explicitly as

$$[c_{i,j+k+1} - w_{j+k+1}(t_i)]_{j,k=0}^{n_i} = \begin{bmatrix} 0 & \dots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \dots & \delta_i g_i^{-1} \end{bmatrix}.$$

Upon equating the corresponding entries in the latter equality we arrive at

$$w_j(t_i) = c_{i,j} \quad (j = 1, \dots, 2n_i)$$

and

$$c_{i,2n_i+1} - w_{2n_i+1}(t_i) = \delta_i g_i^{-1}.$$

The first line (together with (4.17)) proves (1.27). The second one can be written as

$$(c_{i,2n_i+1} - w_{2n_i+1}(t_i))\, g_i = \delta_i \geq 0,$$

which implies (1.28), due to (4.23). In the case when equality (4.10) holds, we get from (4.21) that $\delta_i = 0$ for $i = 1, \dots, k$ and, therefore, that w is a solution of Problem 1.3 (or equivalently, of Problem 1.2). □

We recall now briefly the setting of the Abstract Interpolation Problem **AIP** (in a generality we need) for the Schur class $\mathcal{S}(\mathcal{E}, \mathcal{E}_*)$ of functions analytic on $\mathbb{D}$ whose values are contractive operators mapping a Hilbert space $\mathcal{E}$ into another Hilbert space $\mathcal{E}_*$. The data of the problem consists of Hilbert spaces $\mathcal{E}$, $\mathcal{E}_*$ and X, a positive semidefinite linear operator P on X, an operator T on X such that the operator $(I - zT)$ has a bounded inverse at every point $z \in \overline{\mathbb{D}}$ except for a finitely many points, and two linear operators $M : X \to \mathcal{E}$ and $E : X \to \mathcal{E}_*$ satisfying the identity

$$P - T^*PT = E^*E - M^*M. \tag{4.24}$$

Definition 4.4. A function $w \in \mathcal{S}(\mathcal{E}, \mathcal{E}_*)$ is said to be a solution of the **AIP** with the data

$$\{P, T, E, M\} \tag{4.25}$$

subject to above assumptions, if the function

$$(\mathbf{F}^w x)(t) := \begin{bmatrix} \mathrm{I}_{\mathcal{E}_*} & w(t) \\ w(t)^* & \mathrm{I}_{\mathcal{E}} \end{bmatrix} \begin{bmatrix} E \\ -M \end{bmatrix} (\mathrm{I} - tT)^{-1} x \tag{4.26}$$

belongs to the space H^w and

$$\|\mathbf{F}^w x\|_{H^w} \leq \|P^{\frac{1}{2}} x\|_X \quad \text{for every } x \in X.$$

The main conclusion of this section is that Problem 1.5 can be included into the **AIP** upon specifying the data in (4.24) in terms of the data (1.25) of Problem 1.5. Let $X = \mathbb{C}^N$ and $\mathcal{E} = \mathcal{E}_* = \mathbb{C}$ and let us identify the matrices P, T, E and M defined in (3.1)–(3.8) with operators acting between the corresponding finite dimensional spaces. For T of the form (3.6), the operator $(\mathrm{I} - tT)^{-1}$ is well defined on X for all $t \in \mathbb{T} \setminus \{t_1, \dots, t_k\}$. Also we note that when $X = \mathbb{C}^N$,

$$\|P^{\frac{1}{2}}x\|_X^2 = x^*Px.$$

Now Theorems 4.2 and 4.3 lead us to the following result.

Theorem 4.5. *Let the matrices P, T, E and M be given by* (3.1)–(3.8) *and let conditions* (4.9) *be satisfied. Then a Schur function w is a solution of Problem* 1.5 *if and only if it is a solution of the* **AIP** *with the data* (4.25).

Corollary 4.6. *Conditions $P \geq 0$ and $|c_{i,0}| = 1$ for $i = 1, \dots, k$ are necessary and sufficient for Problem* 1.5 *to have a solution.*

Proof. Necessity of the conditions was proved in Theorem 4.2. Sufficiency follows from Theorem 4.5 and from a general result [11] stating that **AIP** always has a solution. □

5. On the Abstract Interpolation Problem (AIP)

In this section we recall some results on the **AIP** formulated in Definition 4.4. Then in the next section we will specify these results for the setting of Problem 1.5, when $X = \mathbb{C}^N$, $\mathcal{E} = \mathcal{E}_* = \mathbb{C}$ and operators T, E, M and $P \geq 0$ are just matrices defined in terms of the data of Problem 1.5 via formulas (3.1)–(3.8). In this section they are assumed to be operators satisfying the Stein identity (4.24) for every $x \in X$. This identity means that the formula

$$\mathbf{V} : \begin{bmatrix} P^{\frac{1}{2}}x \\ Mx \end{bmatrix} \to \begin{bmatrix} P^{\frac{1}{2}}Tx \\ Ex \end{bmatrix}, \quad x \in X, \tag{5.1}$$

defines a linear map that can be extended by continuity to an isometry $\mathbf{V}$ acting from

$$\mathcal{D}_{\mathbf{V}} = \mathrm{Clos}\left\{ \begin{bmatrix} P^{\frac{1}{2}}x \\ Mx \end{bmatrix},\ x \in X \right\} \subseteq [X] \oplus \mathcal{E} \tag{5.2}$$

onto

$$\mathcal{R}_{\mathbf{V}} = \mathrm{Clos}\left\{ \begin{bmatrix} P^{\frac{1}{2}}Tx \\ Ex \end{bmatrix},\ x \in X \right\} \subseteq [X] \oplus \mathcal{E}_*, \tag{5.3}$$

where $[X] = \mathrm{Clos}\{P^{\frac{1}{2}}X\}$. One of the main results concerning the **AIP** is the characterization of the set of all solutions in terms of minimal unitary extensions of $\mathbf{V}$: let $\mathcal{H}$ be a Hilbert spaces containing $[X]$ and let

$$\mathbf{U} : \mathcal{H} \oplus \mathcal{E} \to \mathcal{H} \oplus \mathcal{E}_* \quad (\mathcal{H} \supset X) \tag{5.4}$$

be a unitary operator such that $\mathbf{U}|_{\mathcal{D}_\mathbf{V}} = \mathbf{V}$ and having no nonzero reducing subspaces in $\mathcal{H} \ominus [X]$. Then the *characteristic function* of $\mathbf{U}$ defined as

$$w(z) = \mathbf{P}_{\mathcal{E}_*} \mathbf{U} \left(I - z\mathbf{P}_\mathcal{H}\mathbf{U}\right)^{-1} |_\mathcal{E} \quad (z \in \mathbb{D}) \tag{5.5}$$

is a solution of the **AIP** and all the solutions to the **AIP** can be obtained in this way.

A parametrization of all the solutions can be obtained as follows: introduce the defect spaces

$$\Delta := \begin{bmatrix} [X] \\ \mathcal{E} \end{bmatrix} \ominus \mathcal{D}_\mathbf{V} \quad \text{and} \quad \Delta_* := \begin{bmatrix} [X] \\ \mathcal{E}_* \end{bmatrix} \ominus \mathcal{R}_\mathbf{V} \tag{5.6}$$

and let $\widetilde{\Delta}$ and $\widetilde{\Delta}_*$ be isomorphic copies of Δ and Δ_*, respectively, with unitary identification maps

$$i : \Delta \to \widetilde{\Delta} \quad \text{and} \quad i_* : \Delta_* \to \widetilde{\Delta}_*.$$

Define a unitary operator $\mathbf{U}_0$ from $\mathcal{D}_\mathbf{V} \oplus \Delta \oplus \widetilde{\Delta}_*$ onto $\mathcal{R}_\mathbf{V} \oplus \Delta_* \oplus \widetilde{\Delta}$ by the rule

$$\mathbf{U}_0|_{\mathcal{D}_\mathbf{V}} = \mathbf{V}, \quad \mathbf{U}_0|_\Delta = i, \quad \mathbf{U}_0|_{\widetilde{\Delta}_*} = i_*^{-1}. \tag{5.7}$$

This operator is called *the universal unitary colligation* associated to the Stein identity (4.24). Since $\mathcal{D}_\mathbf{V} \oplus \Delta = [X] \oplus \mathcal{E}$ and $\mathcal{R}_\mathbf{V} \oplus \Delta_* = [X] \oplus \mathcal{E}_*$, we can decompose $\mathbf{U}_0$ defined by (5.7) as

$$\mathbf{U}_0 = \begin{bmatrix} U_{11} & U_{12} & U_{13} \\ U_{21} & U_{22} & U_{23} \\ U_{31} & U_{32} & 0 \end{bmatrix} : \quad \begin{bmatrix} [X] \\ \mathcal{E} \\ \widetilde{\Delta}_* \end{bmatrix} \to \begin{bmatrix} [X] \\ \mathcal{E}_* \\ \widetilde{\Delta} \end{bmatrix}. \tag{5.8}$$

Note that $U_{33} = 0$, since (by definition (5.7)) for every $\widetilde{\delta}_* \in \widetilde{\Delta}_*$, the vector $\mathbf{U}_0\delta_*$ belongs to Δ_*, which is a subspace of $[X] \oplus \mathcal{E}_*$ and therefore is orthogonal to $\widetilde{\Delta}$. The *characteristic function* of $\mathbf{U}_0$ is defined as

$$\mathbf{S}(z) = \mathbf{P}_{\mathcal{E}_* \oplus \widetilde{\Delta}} \mathbf{U}_0 \left(I - z\mathbf{P}_{[X]}\mathbf{U}_0\right)^{-1} |_{\mathcal{E} \oplus \widetilde{\Delta}_*} \quad (z \in \mathbb{D}), \tag{5.9}$$

where $\mathbf{P}_{\mathcal{E}_* \oplus \widetilde{\Delta}}$ and $\mathbf{P}_{[X]}$ are the orthogonal projections of the space $[X] \oplus \mathcal{E}_* \oplus \widetilde{\Delta}$ onto $\mathcal{E}_* \oplus \widetilde{\Delta}$ and $[X]$, respectively. Upon substituting (5.8) into (5.9) we get a representation of the function $\mathbf{S}$ in terms of the block entries of $\mathbf{U}_0$:

$$\begin{aligned} \mathbf{S}(z) &= \begin{bmatrix} s_0(z) & s_2(z) \\ s_1(z) & s(z) \end{bmatrix} \\ &= \begin{bmatrix} U_{22} & U_{23} \\ U_{32} & 0 \end{bmatrix} + z \begin{bmatrix} U_{21} \\ U_{31} \end{bmatrix} (I_n - zU_{11})^{-1} \begin{bmatrix} U_{12} & U_{13} \end{bmatrix}. \end{aligned} \tag{5.10}$$

The next theorem was proved in [11].

Theorem 5.1. *Let* $\mathbf{S}$ *be the characteristic function of the universal unitary colligation partitioned as in* (5.10)*. Then all the solutions* w *of the* **AIP** *are parametrized by the formula*

$$w(z) = s_0(z) + s_2(z) \left(1 - \mathcal{E}(z)s(z)\right)^{-1} \mathcal{E}(z)s_1(z), \tag{5.11}$$

where $\mathcal{E}$ runs over the Schur class $\mathcal{S}(\widetilde{\Delta}, \widetilde{\Delta}_*)$.

Since $\mathbf{S}$ is the characteristic function of a unitary colligation, it belongs to the Schur class $\mathcal{S}(\mathcal{E} \oplus \widetilde{\Delta}_*,\ \mathcal{E}_* \oplus \widetilde{\Delta})$ (see [18], [1], [2]) and therefore one can introduce the corresponding de Branges–Rovnyak space $H^{\mathbf{S}}$ as it was explained in Section 2. The next result about realization of a unitary colligation in a function model space goes back to M. Livsits, B. Sz.-Nagy, C. Foias, L. de Branges and J. Rovnyak. In its present formulation it appears in [11]–[15].

Theorem 5.2. *Let $\mathbf{U_0}$ be a unitary colligation of the form* (5.8) *and let $\mathbf{S}$ be its characteristic function defined in* (5.9)*. Then the transformation $\mathcal{F}_{\mathbf{U_0}}$ defined as*

$$\left(\mathcal{F}_{\mathbf{U}_0}[x]\right)(z) = \begin{bmatrix} \left(\mathcal{F}^+_{\mathbf{U}_0}[x]\right)(z) \\ \left(\mathcal{F}^-_{\mathbf{U}_0}[x]\right)(z) \end{bmatrix} := \begin{bmatrix} \mathbf{P}_{\mathcal{E}_*\oplus\widetilde{\Delta}}\mathbf{U}_0\left(I - z\mathbf{P}_{[X]}\mathbf{U}_0\right)^{-1}[x] \\ \bar{z}\mathbf{P}_{\mathcal{E}\oplus\widetilde{\Delta}_*}\mathbf{U}_0^*\left(I - \bar{z}\mathbf{P}_{[X]}\mathbf{U}_0^*\right)^{-1}[x] \end{bmatrix} \tag{5.12}$$

maps $[X]$ onto the de Branges–Rovnyak space $H^{\mathbf{S}}$ and is a partial isometry.

The transformation $\mathcal{F}_{\mathbf{U_0}}$ is called *the Fourier representation* of the space $[X]$ associated with the unitary colligation $\mathbf{U_0}$. Note that the last theorem does not assume any special structure for $\mathbf{U}_0$. However, if $\mathbf{U_0}$ is the universal unitary colligation (5.7) associated to the partially defined isometry $\mathbf{V}$ given in (5.1), then $\mathcal{F}_{\mathbf{U_0}}$ can be expressed in terms of P, T, E and M. The formulation of the following theorem can be found (in a more general setting) in [12], [15]; the proof is contained in [13]. We reproduce it here since the source is hardly available.

Theorem 5.3. *Let $\mathbf{U_0}$ be the universal unitary colligation* (5.7) *associated to the isometry $\mathbf{V}$ given by* (5.1) *and let $\mathbf{S}$ be its characteristic function given by* (5.9)*. Then*

$$\left(\mathcal{F}_{\mathbf{U}_0}P^{\frac{1}{2}}x\right)(t) = \begin{bmatrix} I_{\mathcal{E}_*\oplus\widetilde{\Delta}} & \mathbf{S}(t) \\ \mathbf{S}(t)^* & I_{\mathcal{E}\oplus\widetilde{\Delta}_*} \end{bmatrix} \begin{bmatrix} E(I-tT)^{-1} \\ 0 \\ -M(I-tT)^{-1} \\ 0 \end{bmatrix} x \tag{5.13}$$

for almost every point $t \in \mathbb{T}$ and for every $x \in X$.

Proof. We will verify (5.13) for "plus" and "minus" components separately, i.e., we will verify the relations

$$\begin{aligned} \left(\mathcal{F}^+_{\mathbf{U}_0}P^{\frac{1}{2}}x\right)(t) &= \left(\begin{bmatrix} E \\ 0 \end{bmatrix} - \mathbf{S}(t)\begin{bmatrix} M \\ 0 \end{bmatrix}\right)(I-tT)^{-1}x, \\ \left(\mathcal{F}^-_{\mathbf{U}_0}P^{\frac{1}{2}}x\right)(t) &= \left(\mathbf{S}(t)^*\begin{bmatrix} E \\ 0 \end{bmatrix} - \begin{bmatrix} M \\ 0 \end{bmatrix}\right)(I-tT)^{-1}x, \end{aligned}$$

which are equivalent (upon analytic and conjugate-analytic continuations inside $\mathbb{D}$, respectively) to

$$\left(\mathcal{F}^+_{\mathbf{U}_0}P^{\frac{1}{2}}x\right)(z) = \left(\begin{bmatrix} E \\ 0 \end{bmatrix} - \mathbf{S}(z)\begin{bmatrix} M \\ 0 \end{bmatrix}\right)(I-zT)^{-1}x, \tag{5.14}$$

$$\left(\mathcal{F}^-_{\mathbf{U}_0}P^{\frac{1}{2}}x\right)(z) = \bar{z}\left(\mathbf{S}(z)^*\begin{bmatrix} E \\ 0 \end{bmatrix} - \begin{bmatrix} M \\ 0 \end{bmatrix}\right)(\bar{z}I-T)^{-1}x. \tag{5.15}$$

To prove (5.14), we pick an arbitrary vector

$$v = \begin{bmatrix} y \\ e \\ \delta_* \end{bmatrix} \in \begin{bmatrix} [X] \\ \mathcal{E} \\ \widetilde{\Delta}_* \end{bmatrix}$$

and note that by definitions (5.9) and (5.12),

$$\mathbf{P}_{\mathcal{E}_*\oplus\widetilde{\Delta}}\mathbf{U}_0\left(I - z\mathbf{P}_{[X]}\mathbf{U}_0\right)^{-1}\begin{bmatrix} y \\ e \\ \delta_* \end{bmatrix} = \left(\mathcal{F}^+_{\mathbf{U_0}}y\right)(z) + \mathbf{S}(z)\begin{bmatrix} e \\ \delta_* \end{bmatrix}. \tag{5.16}$$

Introduce the vector

$$v' = \begin{bmatrix} y' \\ e' \\ \delta'_* \end{bmatrix} := \left(I - z\mathbf{P}_{[X]}\mathbf{U}_0\right)^{-1}\begin{bmatrix} y \\ e \\ \delta_* \end{bmatrix}$$

so that $\left(I - z\mathbf{P}_{[X]}\mathbf{U}_0\right)v' = v$. Comparing the corresponding components in the latter equality we conclude that $e = e'$, $\delta_* = \delta'_*$ and

$$y = y' - z\mathbf{P}_{[X]}\mathbf{U}_0\begin{bmatrix} y' \\ e' \\ \delta'_* \end{bmatrix} = y' - z\mathbf{P}_{[X]}\mathbf{U}_0\begin{bmatrix} y' \\ e \\ \delta_* \end{bmatrix}, \tag{5.17}$$

so that

$$v' = \begin{bmatrix} y' \\ e \\ \delta_* \end{bmatrix} = \left(I - z\mathbf{P}_{[X]}\mathbf{U}_0\right)^{-1}\begin{bmatrix} y \\ e \\ \delta_* \end{bmatrix}. \tag{5.18}$$

Substituting (5.17) and (5.18), respectively into the right- and the left-hand side expressions in (5.16) we arrive at

$$\mathbf{P}_{\mathcal{E}_*\oplus\widetilde{\Delta}}\mathbf{U}_0\begin{bmatrix} y' \\ e \\ \delta_* \end{bmatrix} = \left(\mathcal{F}^+_{\mathbf{U_0}}y'\right)(z) - z\left(\mathcal{F}^+_{\mathbf{U_0}}\mathbf{P}_{[X]}\mathbf{U}_0\begin{bmatrix} y' \\ e \\ \delta_* \end{bmatrix}\right)(z) + \mathbf{S}(z)\begin{bmatrix} e \\ \delta_* \end{bmatrix}. \tag{5.19}$$

Since the vector v is arbitrary and $I - z\mathbf{P}_{[X]}\mathbf{U}_0$ is invertible, it follows by (5.17), that v' can be chosen arbitrarily in (5.19) . Fix a vector $x \in X$ and take

$$v' = \begin{bmatrix} y' \\ e \\ \delta_* \end{bmatrix} = \begin{bmatrix} P^{\frac{1}{2}}x \\ Mx \\ 0 \end{bmatrix}. \tag{5.20}$$

Then, by definition (5.7) of $\mathbf{U}_0$ and definition (5.1) of $\mathbf{V}$,

$$\mathbf{U_0}\begin{bmatrix} P^{\frac{1}{2}}x \\ Mx \\ 0 \end{bmatrix} = \begin{bmatrix} P^{\frac{1}{2}}Tx \\ Ex \\ 0 \end{bmatrix} \tag{5.21}$$

and thus,

$$\mathbf{P}_{[X]}\mathbf{U}_0\begin{bmatrix} P^{\frac{1}{2}}x \\ Mx \\ 0 \end{bmatrix} = P^{\frac{1}{2}}Tx \quad \text{and} \quad \mathbf{P}_{\mathcal{E}_*\oplus\widetilde{\Delta}}\mathbf{U}_0\begin{bmatrix} P^{\frac{1}{2}}x \\ Mx \\ 0 \end{bmatrix} = \begin{bmatrix} Ex \\ 0 \end{bmatrix}.$$

Plugging the two last relations and (5.20) into (5.19) we get

$$\begin{bmatrix} Ex \\ 0 \end{bmatrix} = \left(\mathcal{F}^+_{\mathbf{U_o}}P^{\frac{1}{2}}x\right)(z) - z\left(\mathcal{F}^+_{\mathbf{U_o}}P^{\frac{1}{2}}Tx\right)(z) + \mathbf{S}(z)\begin{bmatrix} Mx \\ 0 \end{bmatrix}.$$

By linearity of $\mathcal{F}^+_{\mathbf{U_o}}$, we have

$$\begin{bmatrix} Ex \\ 0 \end{bmatrix} = \left(\mathcal{F}^+_{\mathbf{U_o}}P^{\frac{1}{2}}(I-zT)x\right)(z) + \mathbf{S}(z)\begin{bmatrix} Mx \\ 0 \end{bmatrix}$$

and, upon replacing x by $(I-zT)x$, we rewrite the last relation as

$$\begin{bmatrix} E \\ 0 \end{bmatrix}(I-zT)^{-1}x = \left(\mathcal{F}^+_{\mathbf{U_o}}P^{\frac{1}{2}}x\right)(z) + \mathbf{S}(z)\begin{bmatrix} M \\ 0 \end{bmatrix}(I-zT)^{-1}x,$$

which is equivalent to (5.14). The proof of (5.15) is quite similar: we start with an arbitrary vector

$$v = \begin{bmatrix} y \\ e_* \\ \delta \end{bmatrix} \in \begin{bmatrix} [X] \\ \mathcal{E}_* \\ \widetilde{\Delta} \end{bmatrix}$$

and note that by definitions (5.9) and (5.12),

$$\bar{z}\mathbf{P}_{\mathcal{E}\oplus\widetilde{\Delta}_*}\mathbf{U}_0^*\left(I-\bar{z}\mathbf{P}_{[X]}\mathbf{U}_0^*\right)^{-1}\begin{bmatrix} y \\ e_* \\ \delta \end{bmatrix} = \left(\mathcal{F}^-_{\mathbf{U_o}}y\right)(z) + \bar{z}\mathbf{S}(z)^*\begin{bmatrix} e_* \\ \delta \end{bmatrix}. \tag{5.22}$$

Then we introduce the vector

$$v' := \begin{bmatrix} y' \\ e'_* \\ \delta' \end{bmatrix} = \left(I-\bar{z}\mathbf{P}_{[X]}\mathbf{U}_0^*\right)^{-1}\begin{bmatrix} y \\ e_* \\ \delta \end{bmatrix} \tag{5.23}$$

and check that

$$e'_* = e_*, \quad \delta' = \delta, \quad y = y' - \bar{z}\mathbf{P}_{[X]}\mathbf{U}_0^*\begin{bmatrix} y' \\ e_* \\ \delta \end{bmatrix}, \tag{5.24}$$

which allows us to rewrite (5.22) as

$$\bar{z}\mathbf{P}_{\mathcal{E}\oplus\widetilde{\Delta}_*}\mathbf{U}_0^*\begin{bmatrix} y' \\ e_* \\ \delta \end{bmatrix} = \left(\mathcal{F}^-_{\mathbf{U_o}}y'\right)(z) - \bar{z}\left(\mathcal{F}^-_{\mathbf{U_o}}\mathbf{P}_{[X]}\mathbf{U}_0^*\begin{bmatrix} y' \\ e_* \\ \delta \end{bmatrix}\right)(z) + \bar{z}\mathbf{S}(z)^*\begin{bmatrix} e_* \\ \delta \end{bmatrix}. \tag{5.25}$$

By the same arguments as above, v' can be chosen arbitrarily in $[X] \oplus \mathcal{E}_* \oplus \widetilde{\Delta}$ and we let

$$v' = \begin{bmatrix} y' \\ e \\ \delta_* \end{bmatrix} = \begin{bmatrix} P^{\frac{1}{2}}Tx \\ Ex \\ 0 \end{bmatrix}, \quad x \in X. \tag{5.26}$$

Since $\mathbf{U_0}$ is unitary, it follows from (5.21) that

$$\mathbf{U_0}^* \begin{bmatrix} P^{\frac{1}{2}}Tx \\ Ex \\ 0 \end{bmatrix} = \begin{bmatrix} P^{\frac{1}{2}}x \\ Mx \\ 0 \end{bmatrix}$$

and thus,

$$\mathbf{P}_{[X]}\mathbf{U}_0^* \begin{bmatrix} P^{\frac{1}{2}}Tx \\ Ex \\ 0 \end{bmatrix} = P^{\frac{1}{2}}x \quad \text{and} \quad \mathbf{P}_{\mathcal{E}\oplus\widetilde{\Delta}_*}\mathbf{U}_0^* \begin{bmatrix} P^{\frac{1}{2}}Tx \\ Ex \\ 0 \end{bmatrix} = \begin{bmatrix} Mx \\ 0 \end{bmatrix}.$$

Plugging the two last relations and (5.26) into (5.25) we get

$$\bar{z}\begin{bmatrix} Mx \\ 0 \end{bmatrix} = \left(\mathcal{F}^-_{\mathbf{U_0}} P^{\frac{1}{2}}Tx\right)(z) - \bar{z}\left(\mathcal{F}^-_{\mathbf{U_0}} P^{\frac{1}{2}}x\right)(z) + \bar{z}\mathbf{S}(z)^* \begin{bmatrix} Ex \\ 0 \end{bmatrix}.$$

By linearity of $\mathcal{F}^-_{\mathbf{U_0}}$, we have

$$\bar{z}\begin{bmatrix} Mx \\ 0 \end{bmatrix} = \left(\mathcal{F}^-_{\mathbf{U_0}} P^{\frac{1}{2}}(T - \bar{z}I)x\right)(z) + \bar{z}\mathbf{S}(z)^* \begin{bmatrix} Ex \\ 0 \end{bmatrix}$$

and, upon replacing x by $(\bar{z}I - T)^{-1}x$, we rewrite the latter relation as

$$\bar{z}\begin{bmatrix} M \\ 0 \end{bmatrix}(\bar{z}I - T)^{-1} = -\left(\mathcal{F}^-_{\mathbf{U_0}} P^{\frac{1}{2}}x\right)(z) + \bar{z}\mathbf{S}(z)^* \begin{bmatrix} E \\ 0 \end{bmatrix}(\bar{z}I - T)^{-1},$$

which is equivalent to (5.15). □

6. Description of all solutions of Problem 1.5

Since Problem 1.5 is equivalent to the **AIP** with a specific choice of the data (4.25), Theorem 5.1 gives, in fact, a parametrization of all solutions of Problem 1.5. However, the fact that in the context of Problem 1.5, $X = \mathbb{C}^N$ and $\mathcal{E} = \mathcal{E}_* = \mathbb{C}$, and that the matrices P, T, E and M are of special structure (3.1)-(3.8), allow us to rewrite the results from the previous section more transparently. We assume that the necessary conditions (4.9) for Problem 1.5 to have a solution are in force. Then P satisfies the Stein identity

$$P + M^*M = T^*PT + E^*E \tag{6.1}$$

(by Theorem 3.1) which in turn, gives raise to the isometry

$$\mathbf{V}: \begin{bmatrix} P^{\frac{1}{2}}x \\ Mx \end{bmatrix} \to \begin{bmatrix} P^{\frac{1}{2}}Tx \\ Ex \end{bmatrix}, \quad x \in \mathbb{C}^N$$

that maps

$$\mathcal{D}_{\mathbf{V}} = \operatorname{Ran} \begin{bmatrix} P^{\frac{1}{2}} \\ M \end{bmatrix} \subseteq \begin{bmatrix} [X] \\ \mathcal{E} \end{bmatrix} \quad \text{onto} \quad \mathcal{R}_{\mathbf{V}} = \operatorname{Ran} \begin{bmatrix} P^{\frac{1}{2}}T \\ E \end{bmatrix} \subseteq \begin{bmatrix} [X] \\ \mathcal{E}_* \end{bmatrix},$$

where $[X] = \operatorname{Ran} P^{\frac{1}{2}}$. In the present context, the defect spaces (5.6)

$$\Delta = \begin{bmatrix} [X] \\ \mathcal{E} \end{bmatrix} \ominus \mathcal{D}_{\mathbf{V}} \quad \text{and} \quad \Delta_* = \begin{bmatrix} [X] \\ \mathcal{E}_* \end{bmatrix} \ominus \mathcal{R}_{\mathbf{V}}$$

admit a simple characterization.

Lemma 6.1. *If P is nonsingular, then*

$$\Delta = \operatorname{Span} \begin{bmatrix} -P^{-\frac{1}{2}}M^* \\ 1 \end{bmatrix} \quad \text{and} \quad \Delta_* = \operatorname{Span} \begin{bmatrix} -P^{-\frac{1}{2}}(T^{-1})^*E^* \\ 1 \end{bmatrix}. \tag{6.2}$$

If P is singular, then $\Delta = \{0\}$ and $\Delta_* = \{0\}$.

Proof. A vector $\begin{bmatrix} [x] \\ e \end{bmatrix} \in \begin{bmatrix} [X] \\ \mathcal{E} \end{bmatrix}$ belongs to Δ if and only if

$$< [x], P^{\frac{1}{2}}y > + < e, My >= 0$$

for every $y \in X$, which is equivalent to

$$P^{\frac{1}{2}}[x] + M^*e = 0. \tag{6.3}$$

Equation (6.3) has a nonzero solution $\begin{bmatrix} [x] \\ e \end{bmatrix}$ if and only if the vector-column M^* belongs to $[X]$. If P is nonsingular, then $[X] = X$, therefore $M^* \in [X]$, and (6.3) implies the first relation in (6.2). The second relation is proved quite similarly.

Let now P be singular. Then $M^* \notin [X]$. Indeed assuming that $M^* \in \operatorname{Ran} P^{\frac{1}{2}}$ we get that $Mx = 0$ for every $x \in \operatorname{Ker} P$, which implies, in view of (6.1), that $Tx \in \operatorname{Ker} P$ and $Ex = 0$ for every $x \in \operatorname{Ker} P$. In particular, $\operatorname{Ker} P$ is T-invariant and therefore, at least one eigenvector x_0 of T belongs to $\operatorname{Ker} P$, and this vector must satisfy $Ex_0 = 0$. However, by definitions (3.6, (3.7) $Ex_0 \neq 0$ for every eigenvector x_0 of T. The contradiction means that $M^* \notin [X]$ and, therefore, equation (6.3) has only zero solution, i.e., $\Delta = \{0\}$ in case when P is singular. The result concerning Δ_* is established in much the same way. □

Theorem 6.2. *If P is singular, then Problem* 1.5 *has a unique solution*

$$w_0(z) = E\left(\widetilde{P} - zPT\right)^{-1} M^*, \tag{6.4}$$

(which is a finite Blaschke product of degree equal to rank P*), where*

$$\widetilde{P} := P + M^*M = T^*PT + E^*E. \tag{6.5}$$

The inverse in (6.4) *is well defined as an operator on $\widetilde{X} = \operatorname{Ran} \widetilde{P}$.*

Proof. By Lemma 6.1, if P is singular then $\mathcal{D}_{\mathbf{V}} = [X] \oplus \mathcal{E}$ and $\mathcal{R}_{\mathbf{V}} = [X] \oplus \mathcal{E}_*$ where $[X] = \operatorname{Ran} P^{\frac{1}{2}}$. Therefore, the isometry $\mathbf{V}$ defined by (5.1), is already a unitary operator from $[X] \oplus \mathcal{E}$ onto $[X] \oplus \mathcal{E}_*$. Therefore, the solution is unique and is given by the formula (5.5) with $\mathbf{V}$ and $[X]$ in place of $\mathbf{U}$ and $\mathcal{H}$, respectively:

$$w_0(z) = \mathbf{P}_{\mathcal{E}_*} \mathbf{V} \left(I - z\mathbf{P}_{[X]}\mathbf{V}\right)^{-1} |_{\mathcal{E}} \quad (z \in \mathbb{D}). \tag{6.6}$$

Since $\dim[X] < \infty$, it follows that w_0 is a finite Blaschke product of degree equal to $\dim[X] = \operatorname{rank} P$ (see, e.g., [19]). It remains to derive the realization formula (6.4) from (6.6).

Note that by definition of $\widetilde{X}$, it is $\widetilde{P}$-invariant. Since $\widetilde{P}$ is Hermitian, it is invertible on its range $\widetilde{X}$. In what follows, the symbol $\widetilde{P}^{-1}$ will be understood as an operator on $\widetilde{X}$. We define the mappings

$$A = \begin{bmatrix} P^{\frac{1}{2}} \\ M \end{bmatrix} : \widetilde{X} \to [X] \oplus \mathcal{E} \quad \text{and} \quad B = \begin{bmatrix} P^{\frac{1}{2}}T \\ E \end{bmatrix} : \widetilde{X} \to [X] \oplus \mathcal{E}_*.$$

Since

$$A^*A = B^*B = \widetilde{P} \tag{6.7}$$

and since $\widetilde{P}$ is invertible on $\widetilde{X}$, both A and B are nonsingular on $\widetilde{X}$. Since P is singular, it follows (by the proof of Lemma 6.1) that $M^* \notin [X]$ and thus $\dim \widetilde{X} = \dim[X] + 1$. Therefore, A is a bijection from $\widetilde{X}$ onto $[X] \oplus \mathcal{E}$ and B is a bijection from $\widetilde{X}$ onto $[X] \oplus \mathcal{E}_*$. Using (6.7), one can also write the formulas for the inverses

$$A^{-1} = \widetilde{P}^{-1}A^* : [X] \oplus \mathcal{E} \to \widetilde{X}, \quad B^{-1} = \widetilde{P}^{-1}B^* : [X] \oplus \mathcal{E}_* \to \widetilde{X}.$$

By definition (5.1), $\mathbf{V}A = B$, which can be rephrased as $\mathbf{V} = BA^{-1} = B\widetilde{P}^{-1}A^*$. Plugging this in (6.6) we get (6.4):

$$\begin{aligned} w_0(z) &= \mathbf{P}_{\mathcal{E}_*} B\widetilde{P}^{-1}A^* \left(I - z\mathbf{P}_{[X]}B\widetilde{P}^{-1}A^*\right)^{-1} |_{\mathcal{E}} \\ &= \mathbf{P}_{\mathcal{E}_*} B\widetilde{P}^{-1} \left(I - zA^*\mathbf{P}_{[X]}B\widetilde{P}^{-1}\right)^{-1} A^*|_{\mathcal{E}} \\ &= \mathbf{P}_{\mathcal{E}_*} B \left(\widetilde{P} - zA^*\mathbf{P}_{[X]}B\right)^{-1} A^*|_{\mathcal{E}} \\ &= \mathbf{P}_{\mathcal{E}_*} B \left(\widetilde{P} - zPT\right)^{-1} A^*|_{\mathcal{E}} \\ &= E \left(\widetilde{P} - zPT\right)^{-1} M^*. \end{aligned}$$

All the inverses in the latter chain of equalities (except the first one) are understood as operators on $\widetilde{X}$. They exist, since the first inverse in this chain does, which, in turn, is in effect since $\mathbf{V}$ is unitary. □

Theorem 6.3. *If P is nonsingular, then the set of all solutions of Problem* 1.5 *is parametrized by the formula*

$$w(z) = s_0(z) + s_2(z)\left(1 - \mathcal{E}(z)s(z)\right)^{-1} \mathcal{E}(z)s_1(z), \tag{6.8}$$

where the free parameter $\mathcal{E}$ runs over the Schur class $\mathcal{S}$,

$$s_0(z) = E(\widetilde{P} - zPT)^{-1}M^*, \tag{6.9}$$

$$s_1(z) = \alpha^{-1}\left(1 - zMT(\widetilde{P} - zPT)^{-1}M^*\right), \tag{6.10}$$

$$s_2(z) = \beta^{-1}\left(1 - zE(\widetilde{P} - zPT)^{-1}(T^{-1})^*E^*\right), \tag{6.11}$$

$$s(z) = z\alpha^{-1}\beta^{-1}MP^{-1}\widetilde{P}(\widetilde{P} - zPT)^{-1}(T^{-1})^*E^*, \tag{6.12}$$

the matrix $\widetilde{P}$ is given in (6.5) *and α and β are positive numbers given by*

$$\alpha = \sqrt{1 + MP^{-1}M^*} \quad \textit{and} \quad \beta = \sqrt{1 + ET^{-1}P^{-1}(T^{-1})^*E^*}. \tag{6.13}$$

The matrix $(\widetilde{P} - zPT)$ is invertible for every $z \in \mathbb{D}$ in this case.

Proof. By Theorem 5.1, all the solutions of Problem 1.5 are parametrized by the formula (6.8) where the coefficients s_0, s_1, s_2 and s are the entries of the characteristic function $\mathbf{S}$ of the universal unitary colligation $\mathbf{U_0}$. By Lemma 6.1, we have $\dim \Delta = \dim \Delta_* = 1$. Since, by the very construction of the universal colligation, $\widetilde{\Delta}$ and $\widetilde{\Delta}_*$ are isomorphic copies of Δ and Δ_*, respectively, we have also $\dim \widetilde{\Delta} = \dim \widetilde{\Delta}_* = 1$, and we will identify each of these two spaces with $\mathbb{C}$. However, we will keep the notations $\widetilde{\Delta}$ and $\widetilde{\Delta}_*$ for the spaces so that not to mix them up. Thus, in the present context, the characteristic function $\mathbf{S}$ of $\mathbf{U}_0$ is a 2×2 matrix-valued function and it remains to establish explicit formulas (6.9)–(6.12) for its entries which are scalar-valued functions. First we will write relations (5.7) defining the operator $\mathbf{U}_0 : X \oplus \mathcal{E} \oplus \widetilde{\Delta}_* \to X \oplus \mathcal{E}_* \oplus \widetilde{\Delta}$ more explicitly. The first relation in (5.7) can be written, by the definition (5.1) of $\mathbf{V}$, as

$$\mathbf{U}_0 \begin{bmatrix} P^{\frac{1}{2}} \\ M \\ 0 \end{bmatrix} = \begin{bmatrix} P^{\frac{1}{2}}T \\ E \\ 0 \end{bmatrix}. \tag{6.14}$$

By Lemma 6.1, the spaces Δ and Δ_* are spanned by the vectors

$$\delta = \begin{bmatrix} -P^{-\frac{1}{2}}M^* \\ 1 \\ 0 \end{bmatrix} \in \begin{bmatrix} X \\ \mathcal{E} \\ \widetilde{\Delta}_* \end{bmatrix} \quad \text{and} \quad \delta_* = \begin{bmatrix} -P^{-\frac{1}{2}}(T^{-1})^*E^* \\ 1 \\ 0 \end{bmatrix} \in \begin{bmatrix} X \\ \mathcal{E}_* \\ \widetilde{\Delta} \end{bmatrix},$$

respectively. Note that

$$\|\delta\|^2 = 1 + MP^{-1}M^* \quad \text{and} \quad \|\delta_*\|^2 = 1 + ET^{-1}P^{-1}(T^{-1})^*E^*. \tag{6.15}$$

By the second relation in (5.7), the vector $\mathbf{U}_0\delta$ belongs to $\widetilde{\Delta}$ and therefore, it is of the form

$$\mathbf{U}_0\delta = \begin{bmatrix} 0 \\ 0 \\ \alpha \end{bmatrix} \tag{6.16}$$

where $|\alpha| = \|\delta\|$, due to unitarity of $\mathbf{U}_0$. The latter equality and the first equality in (6.15) imply that $\alpha \neq 0$. In fact, we can choose the identification map $i : \Delta \to \widetilde{\Delta}$

so that α will be as in (6.13). Equality (6.16) is an explicit form of the second relation in (5.7). Similarly, the second identification map $i_* : \Delta_* \to \widetilde{\Delta}_*$ can be chosen so that

$$\mathbf{U}_0 \begin{bmatrix} 0 \\ 0 \\ \beta \end{bmatrix} = \delta_*, \tag{6.17}$$

where β is defined as in (6.13). Summarizing equalities (6.14), (6.16) and (6.17) we conclude that $\mathbf{U}_0$ satisfies (and is uniquely determined by) the equation

$$\mathbf{U}_0 A = B, \tag{6.18}$$

where

$$A = \begin{bmatrix} P^{\frac{1}{2}} & -P^{-\frac{1}{2}}M^* & 0 \\ M & 1 & 0 \\ 0 & 0 & \beta \end{bmatrix} \quad \text{and} \quad B = \begin{bmatrix} P^{\frac{1}{2}}T & 0 & -P^{-\frac{1}{2}}(T^{-1})^*E^* \\ E & 0 & 1 \\ 0 & \alpha & 0 \end{bmatrix} \tag{6.19}$$

are operators from $X \oplus E \oplus \mathcal{E}_*$ to $X \oplus E \oplus \widetilde{\Delta}_*$ and to $X \oplus E_* \oplus \widetilde{\Delta}$, respectively. Since $\mathbf{U}_0$ is unitary, it follows that $A^*A = B^*B$. We denote this matrix by $\widehat{P}$ and a straightforward calculation shows that

$$\widehat{P} := A^*A = B^*B = \begin{bmatrix} \widetilde{P} & 0 & 0 \\ 0 & |\alpha|^2 & 0 \\ 0 & 0 & |\beta|^2 \end{bmatrix} : \quad \begin{bmatrix} X \\ \mathcal{E} \\ \mathcal{E}_* \end{bmatrix} \to \begin{bmatrix} X \\ \mathcal{E} \\ \mathcal{E}_* \end{bmatrix}. \tag{6.20}$$

where $\widetilde{P}$ is given in (6.5). Since P is nonsingular so is $\widetilde{P}$ and since $\alpha \neq 0$ and $\beta \neq 0$, $\widehat{P}$ is nonsingular as well. Therefore, A and B are nonsingular. Now we proceed as in the proof of Theorem 6.2: it follows from (6.18) and (6.20) that $\mathbf{U}_0 = BA^{-1} = B\widehat{P}^{-1}A^*$ which being substituted into (5.9) leads us (recall that since P is nonsingular, $[X] = X = C^N$) to

$$\begin{aligned} \mathbf{S}(z) &= \mathbf{P}_{\mathcal{E}_* \oplus \widetilde{\Delta}} B\widehat{P}^{-1}A^* \left(I - z\mathbf{P}_X B\widehat{P}^{-1}A^*\right)^{-1} |_{\mathcal{E} \oplus \widetilde{\Delta}_*} \\ &= \begin{bmatrix} 0 & I_2 \end{bmatrix} B\widehat{P}^{-1} \left(I - zA^*\mathbf{P}_X B\widehat{P}^{-1}\right)^{-1} A^* \begin{bmatrix} 0 \\ I_2 \end{bmatrix} \\ &= \begin{bmatrix} E & 0 & 1 \\ 0 & \alpha & 0 \end{bmatrix} \left(\widehat{P} - zA^* \begin{bmatrix} I_N & 0 \\ 0 & 0 \end{bmatrix} B\right)^{-1} \begin{bmatrix} M^* & 0 \\ 1 & 0 \\ 0 & \beta \end{bmatrix}, \end{aligned} \tag{6.21}$$

where I_2 and I_N are 2×2 and $N \times N$ unit matrices, respectively. The first inverse in this chain of equalities exists for every $z \in \mathbb{D}$ since $\mathbf{U}_0$ is unitary, all the others exist since the first one does. By (6.19) and (6.20),

$$\widehat{P} - zA^* \begin{bmatrix} I_N & 0 \\ 0 & 0 \end{bmatrix} B = \begin{bmatrix} \widetilde{P} - zPT & 0 & z(T^{-1})^*E^* \\ zMT & |\alpha|^2 & -zMP^{-1}(T^{-1})^*E^* \\ 0 & 0 & |\beta|^2 \end{bmatrix}.$$

Upon inverting the latter triangular matrix and plugging it into (6.21), we eventually get

$$\begin{aligned}\mathbf{S}(z) &= \begin{bmatrix} s_0(z) & s_2(z) \\ s_1(z) & s(z) \end{bmatrix} \\ &= \begin{bmatrix} ER(z)M^* & \beta^{-1}\left(1 - zER(z)(T^{-1})^*E^*\right) \\ \alpha^{-1}\left(1 - zMTR(z)M^*\right) & z\alpha^{-1}\beta^{-1}MP^{-1}\widetilde{P}R(z)(T^{-1})^*E^* \end{bmatrix},\end{aligned}$$

where $R(z) = \left(\widetilde{P} - zPT\right)^{-1}$, which is equivalent to (6.9)–(6.12). □

In conclusion we will establish some important properties of the coefficient matrix $\mathbf{S}$ constructed in Theorem 6.3.

Theorem 6.4. *Let* $\mathbf{S} = \begin{bmatrix} s_0 & s_2 \\ s_1 & s \end{bmatrix}$ *be the characteristic function of the universal unitary colligation* $\mathbf{U_0}$ *defined in* (6.19), (6.20). *Then*

1. *The function* s_0 *is a solution of Problem* 1.2.
2. *The function* $\mathbf{S}(z)$ *is a rational inner matrix-function of degree at most* N.
3. *The functions* s_1 *and* s_2 *have zeroes of multiplicity* n_i+1 *at each interpolating point* t_i *and do not have other zeroes.*

Proof. By Theorem 6.3, s_0 is a solution of Problem 1.5 (corresponding to the parameter $\mathcal{E} \equiv 0$ in the parametrization formula (6.8)). Therefore, by Theorem 4.5, $\mathbf{F}^{s_0}x$ belongs to the space H^{s_0} for every $x \in \mathbb{C}^N = X$, where H^{s_0} is the de Branges–Rovnyak space associated to the Schur function s_0 and where

$$\mathbf{F}^{s_0}(t) := \begin{bmatrix} 1 & s_0(t) \\ s_0(t)^* & 1 \end{bmatrix} \begin{bmatrix} E \\ -M \end{bmatrix} (\mathrm{I} - tT)^{-1}. \tag{6.22}$$

Again, by Theorem 4.5, to show that s_0 is a solution of Problem 1.2, it remains to check that $\|\mathbf{F}^{s_0}x\|_{H^{s_0}} = x^*Px$. Letting for short

$$R_T(t) := (\mathrm{I} - tT)^{-1}, \tag{6.23}$$

we note that by (6.22) and by definition of the norm in the de Branges-Rovnyak space,

$$\|\mathbf{F}^{s_0}x\|^2_{H^{s_0}} = \left\langle \begin{bmatrix} 1 & s_0 \\ s_0^* & 1 \end{bmatrix} \begin{bmatrix} E \\ -M \end{bmatrix} R_T x,\ \begin{bmatrix} E \\ -M \end{bmatrix} R_T x \right\rangle_{L^2(\mathbb{C}^2)}. \tag{6.24}$$

By Theorems 5.2 and 5.3, the function

$$\left(\mathcal{F}_{\mathbf{U_o}} P^{\frac{1}{2}} x\right)(t) = \begin{bmatrix} \mathrm{I}_2 & \mathbf{S}(t) \\ \mathbf{S}(t)^* & \mathrm{I}_2 \end{bmatrix} \begin{bmatrix} E \\ 0 \\ -M \\ 0 \end{bmatrix} R_T(t) x \tag{6.25}$$

$$= \begin{bmatrix} E - s_0(t) M \\ -s_1(t) M \\ s_0(t)^* E - M \\ s_2(t)^* E \end{bmatrix} (\mathrm{I} - tT)^{-1} x$$

belongs to $H^{\mathbf{S}}$ for every vector $x \in \mathbb{C}^N$. Note that $E(\mathrm{I} - tT)^{-1} x \not\equiv 0$, unless $x = 0$. Indeed, letting

$$x = \begin{bmatrix} x_1 \\ \vdots \\ x_k \end{bmatrix}, \quad \text{where} \quad x_i = \begin{bmatrix} x_{i,1} \\ \vdots \\ x_{i,n_i} \end{bmatrix} \quad (i = 1, \dots, k), \tag{6.26}$$

we get, on account of definitions (3.6) and (3.7) of T and E, that

$$E(\mathrm{I} - tT)^{-1} x = \sum_{i=1}^{k} \sum_{j=0}^{n_i} \frac{t^j}{(1 - t\bar{t}_i)^{j+1}} x_{i,j} \not\equiv 0 \tag{6.27}$$

for every $x \neq 0$, since the functions

$$\frac{t^j}{(1 - t\bar{t}_i)^{j+1}} \quad (i = 1, \dots, k;\ j = 0, \dots, n_i)$$

are linearly independent (recall that all the points $t_1, \dots, t_k$ are distinct).

Note also that $s_2 \not\equiv 0$ (since $s_2(0) = \beta \neq 0$, by (6.11) and (6.13)) and therefore,

$$s_2(t) E(\mathrm{I} - tT)^{-1} x \not\equiv 0$$

for every $x \in X, x \neq 0$. It is seen from (6.25) that the latter function is the bottom component of $\mathcal{F}_{\mathbf{U_o}} P^{\frac{1}{2}} x$, which leads us to the conclusion that

$$\mathcal{F}_{\mathbf{U_o}} P^{\frac{1}{2}} x \not\equiv 0 \quad \text{for every} \quad x \neq 0.$$

The latter means that the linear map $\mathcal{F}_{\mathbf{U_o}} : [X] \to H^{\mathbf{S}}$ is a bijection. Since $\mathcal{F}_{\mathbf{U_o}}$ is a partial isometry (by Theorem 5.2), it now follows that this map is unitary, i.e., that

$$\left\| \mathcal{F}_{\mathbf{U_o}} P^{\frac{1}{2}} x \right\|_{H^{\mathbf{S}}}^2 = \left\| P^{\frac{1}{2}} x \right\|_X^2 = x^* P x. \tag{6.28}$$

Furthermore,

$$\left\| \mathcal{F}_{\mathbf{U_o}} P^{\frac{1}{2}} x \right\|_{H^{\mathbf{S}}}^2 = \left\langle \begin{bmatrix} \mathrm{I}_2 & \mathbf{S} \\ \mathbf{S}^* & \mathrm{I}_2 \end{bmatrix} \begin{bmatrix} E \\ 0 \\ -M \\ 0 \end{bmatrix} R_T x, \begin{bmatrix} E \\ 0 \\ -M \\ 0 \end{bmatrix} R_T x \right\rangle_{L^2(\mathbb{C}^4)},$$

by (6.25) and virtue of formula (2.1) for the norm in $H^{\mathbf{S}}$. Upon taking advantage of the zero entries in the last formula and the partition of the matrix $\mathbf{S}$, we get

$$\begin{aligned}\left\|\mathcal{F}_{\mathbf{U_0}}P^{\frac{1}{2}}x\right\|^2_{H^{\mathbf{S}}} &= \left\langle \begin{bmatrix} 1 & 0 & s_0 & s_2 \\ s_0^* & s_1^* & 1 & 0 \end{bmatrix} \begin{bmatrix} E \\ 0 \\ -M \\ 0 \end{bmatrix} R_T x,\ \begin{bmatrix} E \\ -M \end{bmatrix} R_T x \right\rangle_{L^2(\mathbb{C}^2)} \\ &= \left\langle \begin{bmatrix} 1 & s_0 \\ s_0^* & 1 \end{bmatrix} \begin{bmatrix} E \\ -M \end{bmatrix} R_T x,\ \begin{bmatrix} E \\ -M \end{bmatrix} R_T x \right\rangle_{L^2(\mathbb{C}^2)}. \qquad (6.29)\end{aligned}$$

Comparing (6.24) and (6.29) and taking into account (6.28) we arrive at

$$\|\mathbf{F}^{s_0}x\|_{H^{s_0}} = \left\|\mathcal{F}_{\mathbf{U_0}}P^{\frac{1}{2}}x\right\|_{H^{\mathbf{S}}} = x^*Px,$$

which proves the first assertion of the theorem. The second assertion follows since $\dim X = N = \sum_{i=0}^{k}(n_i+1) < \infty$ (see, e.g., [19]).

To prove the last assertion, we use (6.25) for x in the form (6.26) with the only nonzero entry $x_{i,n_i} = 1$. For this choice of x we have by definitions (3.6)–(3.8) of T, E and N,

$$E(I-tT)^{-1}x = \frac{t^{n_i}}{(1-t\bar{t}_i)^{n_i+1}} \quad \text{and} \quad M(I-tT)^{-1}x = \frac{\mathbf{c}_i(t)}{(1-t\bar{t}_i)^{n_i+1}}$$

where

$$\mathbf{c}_i(t) = \sum_{\ell=0}^{n_i} t^{n_i-\ell}(1-t\bar{t}_i)^{\ell}c_{i,\ell}^*. \qquad (6.30)$$

Now we conclude from (6.25) that

$$\frac{s_1(t)\mathbf{c}_i(t)}{(1-t\bar{t}_i)^{n_i+1}} \in H_2^+ \quad \text{and} \quad \frac{t^{n_i}s_2(t)^*}{(1-t\bar{t}_i)^{n_i+1}} = \bar{t}\frac{s_2(t)^*}{(\bar{t}-\bar{t}_i)^{n_i+1}} \in H_2^-. \qquad (6.31)$$

By (6.30), $\mathbf{c}_i(t_i) = t_i^{n_i}c_{i,0}^* \neq 0$ and thus, the first condition in (6.31) implies that s_1 has the zero of multiplicity at least n_i+1 at t_i. The second condition in (6.31) is equivalent to

$$\frac{s_2(t)}{(t-t_i)^{n_i+1}} \in H_2^+$$

which implies that s_2 has zero of multiplicity at least n_i+1 at t_i. On the other hand, since s_1 and s_2 are rational functions of degree at most $N = \sum_{i=0}^{k}(n_i+1)$ (the second assertion of this theorem) and since they do not vanish identically (by the proof of the first assertion of this theorem), they can not have more than N zeroes. Therefore, they have zeroes of multiplicities n_i+1 at t_i for $i = 1, \ldots, k$ and they do not have other zeroes. □

Some consequences of Theorem 6.4 needed in the next section are proved in the following lemma.

Lemma 6.5. *Let* $\mathbf{S} = \begin{bmatrix} s_0 & s_2 \\ s_1 & s \end{bmatrix}$ *be as in Theorem* 6.4. *Then* $|s(t_i)| = 1$,

$$\frac{s_1^{(n_i+1)}(t_i)}{(n_i+1)!} = \lim_{z\to t_i} \frac{s_1(z)}{(z-t_i)^{n_i+1}} \neq 0, \quad \frac{s_2^{(n_i+1)}(t_i)}{(n_i+1)!} = \lim_{z\to t_i} \frac{s_2(z)}{(z-t_i)^{n_i+1}} \neq 0, \tag{6.32}$$

and

$$s_2^{(n_i+1)}(t_i)^* = (-1)^{n_i} t_i^{2n_i+2} s(t_i)^* s_1^{(n_i+1)}(t_i) c_{i,0}{}^* \ . \tag{6.33}$$

Proof. By the third assertion of Theorem 6.4, the rational functions s_1 and s_2 have zeros of multiplicity n_i+1 at t_i. This implies (6.32). By the second assertion of Theorem 6.4, the matrix-function $\mathbf{S}$ is inner and rational. In particular, it is unitary at $t_i \in \mathbb{T}$ and therefore, $|s_2(t_i)|^2 + |s(t_i)|^2 = 1$ which implies $|s(t_i)| = 1$, since $s_2(t_i) = 0$. Furthermore, by the reflection principle, $\mathbf{S}(1/\bar{z})^*\mathbf{S}(z) \equiv \mathrm{I}_2$, or in more detail,

$$\begin{bmatrix} s_0(1/\bar{z})^* & s_1(1/\bar{z})^* \\ s_2(1/\bar{z})^* & s(1/\bar{z})^* \end{bmatrix} \begin{bmatrix} s_0(z) & s_2(z) \\ s_1(z) & s(z) \end{bmatrix} = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}.$$

In particular,

$$s_2(1/\bar{z})^* s_0(z) + s(1/\bar{z})^* s_1(z) = 0. \tag{6.34}$$

To verify (6.33), we note first that by the first relation in (6.32),

$$\lim_{z\to t_i} \frac{s(1/\bar{z})^* s_1(z)}{(z-t_i)^{n_i+1}} = \frac{s(t_i)^* s_1^{(n_i+1)}(t_i)}{(n_i+1)!}. \tag{6.35}$$

Since $|t_i| = 1$, the second relation in (6.32) gives

$$\frac{s_2^{(n_i+1)}(t_i)}{(n_i+1)!} = \lim_{z\to t_i} \frac{s_2(\bar{z}^{-1})}{(\bar{z}^{-1}-t_i)^{n_i+1}},$$

which is equivalent, on account of $\bar{z}^{-1} - t_i = -\frac{\bar{z}-\bar{t}_i}{\bar{z}\bar{t}_i}$, to

$$\frac{s_2^{(n_i+1)}(t_i)}{(n_i+1)!} = \lim_{z\to t_i} \frac{(-\bar{z}\bar{t}_i)^{n_i+1} s_2(\bar{z}^{-1})}{(\bar{z}-\bar{t}_i)^{n_i+1}} = (-1)^{n_i+1} \bar{t}_i^{2n_i+2} \lim_{z\to t_i} \frac{s_2(\bar{z}^{-1})}{(\bar{z}-\bar{t}_i)^{n_i+1}}.$$

Upon taking adjoints in the latter equality we get

$$\lim_{z\to t_i} \frac{s_2(\bar{z}^{-1})^*}{(z-t_i)^{n_i+1}} = (-1)^{n_i+1} \bar{t}_i^{2n_i+2} \frac{s_2^{(n_i+1)}(t_i)^*}{(n_i+1)!}$$

and, since $s_0(t_i) = c_{i,0}$ (recall that s_0 is a solution of Problem 1.2), we have also

$$\lim_{z\to t_i} \frac{s_2(\bar{z}^{-1})^* s_0(z)}{(z-t_i)^{n_i+1}} = (-1)^{n_i+1} \bar{t}_i^{2n_i+2} \frac{s_2^{(n_i+1)}(t_i)^*}{(n_i+1)!} c_{i,0}. \tag{6.36}$$

Now upon multiplying (6.34) by $\dfrac{(n_i+1)!}{(z-t_i)^{n_i+1}}$ and passing to limits as $z \to t_i$, we arrive, on account of (6.35) and (6.36), at the equality

$$s(t_i)^* s_1^{(n_i+1)}(t_i) + (-1)^{n_i+1} \bar{t}_i^{2n_i+2} s_2^{(n_i+1)}(t_i)^* c_{i,0} = 0,$$

which is equivalent to (6.33), since $|c_{i,0}| = |t_i| = 1$. □

7. Boundary interpolation problem with equality

In this section we establish a parametrization of all solutions of Problem 1.2. Recall that all solutions w of Problem 1.5 are parametrized by the linear fractional formula (6.8) with the free Schur class parameter $\mathcal{E}$. Thus, for every function w of the form (6.8), we have

$$\delta_{w,i} := \gamma_i - d_{w,n_i}(t_i) \geq 0 \quad (i = 1, \ldots, k).$$

Theorem 7.4 below will present the explicit formula for the gaps $\delta_{w,i}$ in terms of the parameter $\mathcal{E}$ leading to w via formula (6.8). As a consequence of this formula we will get a characterization of all the parameters $\mathcal{E}$, leading to functions w with zero gaps, i.e., to solutions of Problem 1.2. We start with some needed preliminaries. The proof of the first lemma can be found in [22] for the case when $n = 0$. For the case $n > 0$ the proof was given in [6] using pretty much the same ideas.

Lemma 7.1. *Let w be a function analytic in some nontangential neighborhood of a point $t_0 \in \mathbb{T}$ and let $w_0, \ldots, w_{2n+1}$ be complex numbers. Then equality*

$$\lim_{z \to t_0} \frac{w(z) - w_0 - (z-t_0)w_1 - \ldots - (z-t_0)^{2n} w_{2n}}{(z-t_0)^{2n+1}} = w_{2n+1}$$

holds if and only if the nontangential limits $\lim\limits_{z \to t_0} \dfrac{w^{(j)}(z)}{j!}$ exist and equal w_j for $j = 0, \ldots, 2n+1$.

With every triple (ω, t_0, b) consisting of a Schur function $\omega \in \mathcal{S}$, of a point $t_0 \in \mathbb{T}$ and a number $b \in \mathbb{C}$, we associate the quantity

$$\begin{aligned} D_{\omega,b}(t_0) &:= \int_{\mathbb{T}} \frac{1}{|1 - t\bar{t}_0|^2} \begin{bmatrix} 1 & -b \end{bmatrix} \begin{bmatrix} 1 & \omega(t) \\ \omega(t)^* & 1 \end{bmatrix} \begin{bmatrix} 1 \\ -b^* \end{bmatrix} m(dt) \\ &= \int_{\mathbb{T}} \left(\left| \frac{1 - \omega(t)\overline{b}}{1 - t\bar{t}_0} \right|^2 + |b|^2 \frac{1 - |\omega(t)|^2}{|1 - t\bar{t}_0|^2} \right) m(dt) \qquad (7.1) \\ &= \int_{\mathbb{T}} \left(\left| \frac{\omega(t) - b}{t - t_0} \right|^2 + \frac{1 - |\omega(t)|^2}{|t - t_0|^2} \right) m(dt), \end{aligned}$$

where $m(dt)$ is the normalized Lebesgue measure on $\mathbb{T}$. It follows from the very definition that

$$0 \leq D_{\omega,b}(t_0) \leq \infty.$$

The next theorem (which is a variation of the classical Julia-Carathéodory Theorem and can be mostly found in [22]) characterizes the cases when $D_{\omega,b}(t_0)$ is zero, positive or infinite.

Theorem 7.2. *Let $\omega \in \mathcal{S}$, $t_0 \in \mathbb{T}$, $b \in \mathbb{C}$ and let $D_{\omega,\,b}(t_0)$ be defined as in* (7.1). *Then:*

1. *$D_{\omega,b}(t_0) < \infty$ if and only if*
$$\liminf_{z\to t_0} \frac{1-|\omega(z)|^2}{1-|z|^2} < \infty \quad \text{and} \quad \lim_{z\to t_0} \omega(z) = b, \tag{7.2}$$
where the second limit is understood as nontangential. In this case $|b| = 1$.
2. *$D_{\omega,b}(t_0) = \infty$ if and only if either*
$$\liminf_{z\to t_0} \frac{1-|\omega(z)|^2}{1-|z|^2} = \infty,$$
or the function ω fails to have a nontangential limit b at t_0.
3. *$D_{\omega,b}(t_0) = 0$ if and only if $\omega(z) \equiv b$ and $|b| = 1$.*
4. *If $|b| \le 1$, then the equality*
$$\lim_{z\to t_0} \frac{1-\omega(z)b^*}{1-z\overline{t}_0} = D_{\omega,b}(t_0) \tag{7.3}$$
holds where the limit is understood as nontangential.

Proof. Let H^ω be the de Branges-Rovnyak space associated to the Schur class function ω and let us consider the function
$$K_{t_0,b}(t) = \begin{bmatrix} 1 & \omega(t) \\ \omega(t)^* & 1 \end{bmatrix} \begin{bmatrix} 1 \\ -b^* \end{bmatrix} \frac{1}{1-t\overline{t}_0} = \begin{bmatrix} K_{t_0,b,+}(t) \\ K_{t_0,b,-}(t) \end{bmatrix} \tag{7.4}$$
where
$$K_{t_0,b,+}(t) = \frac{1-\omega(t)b^*}{1-t\overline{t}_0} \quad \text{and} \quad K_{t_0,b,-}(t) = \overline{t}\,\frac{\overline{\omega(t)}-b^*}{\overline{t}-\overline{t}_0}. \tag{7.5}$$
By formula (2.1),
$$\|K_{t_0,b}\|^2_{H^\omega} = \left\langle \begin{bmatrix} 1 & \omega(t) \\ \omega(t)^* & 1 \end{bmatrix} \begin{bmatrix} 1 \\ -b^* \end{bmatrix} \frac{1}{1-t\overline{t}_0}, \begin{bmatrix} 1 \\ -b^* \end{bmatrix} \frac{1}{1-t\overline{t}_0} \right\rangle_{L^2\oplus L^2}$$
which is equal to the first integral in (7.1). Therefore, $D_{\omega,b}(t_0) < \infty$ if and only if $K_{t_0,b}$ belongs to L^ω, and in this case,
$$D_{\omega,b}(t_0) = \|K_{t_0,b}\|^2_{L^\omega}. \tag{7.6}$$
On the other hand, if $D_{\omega,\,b}(t_0) < \infty$, then it follows from the second form of $D_{\omega,b}(t_0)$ in (7.1) that
$$\int_{\mathbb{T}} \left| \frac{1-\omega(t)b^*}{1-t\overline{t_0}} \right|^2 m(dt) < \infty, \quad \text{i.e., that} \quad K_{t_0,b,+}(t) = \frac{1-\omega(t)b^*}{1-t\overline{t}_0} \in L_2.$$
Since $1-t\overline{t}_0$ is an outer function, it follows, by Smirnov's maximum principle [23], that $K_{t_0,b,+} \in H_2^+$. Similarly, it follows from the third representation of $D_{\omega,b}(t_0)$ in

(7.1) that $K_{t_0,b,-} \in H_2^-$. Therefore, $K_{t_0,b}$ belongs to H^ω by Definition 2.1. Thus, we have shown that

$$D_{\omega,b}(t_0) < \infty \iff K_{t_0,b} \in L^\omega \iff K_{t_0,b} \in H^\omega.$$

Now the first assertion of the lemma follows from Theorem 2.3 (the case when $n = 0$): the function $K_{t_0,b}$ of the form (7.4) belongs to H^ω if and only if conditions in (7.2) are satisfied. In this case $|b| = 1$, since

$$1 - |b|^2 = \lim_{z \to t_0} (1 - |\omega(z)|^2) = \lim_{z \to t_0} \frac{1 - |\omega(z)|^2}{1 - |z|^2}(1 - |z|^2) = 0.$$

The second assertion is simply the formal negation of the first one. To prove the third assertion, we observe that $D_{\omega,b}(t_0) = 0$ if and only if

$$\begin{bmatrix} 1 & -b \end{bmatrix} \begin{bmatrix} 1 & \omega(t) \\ \omega(t)^* & 1 \end{bmatrix} \begin{bmatrix} 1 \\ -b^* \end{bmatrix} = 0$$

almost everywhere on $\mathbb{T}$, which occurs if and only if

$$\begin{bmatrix} 1 & -b \end{bmatrix} \begin{bmatrix} 1 & \omega(t) \\ \omega(t)^* & 1 \end{bmatrix} = 0$$

almost everywhere on $\mathbb{T}$. The latter equality collapses to $\omega(t) - b = 1 - b\omega(t)^* = 0$ which implies the requisite.

The proof of the fourth assertion splits up into three cases.

Case 1: Let $D_{\omega,\,b}(t_0) < \infty$. Then by the first statement, conditions (7.2) are satisfied. Then by Theorem 2.2, the kernels

$$K_z(t) = \begin{bmatrix} 1 & \omega(t) \\ \omega(t)^* & 1 \end{bmatrix} \begin{bmatrix} 1 \\ -\omega(z)^* \end{bmatrix} \frac{1}{1 - t\overline{z}}$$

converge to $K_{t_0,\,b}$ in norm of H^ω:

$$K_z \xrightarrow{H^w} K_{t_0,\,b}, \tag{7.7}$$

as $z \to t_0$ nontangentially. By the reproducing property (2.3) (for $j = 0$),

$$\langle f,\ K_z \rangle_{H^\omega} = f_+(z) \quad \text{for every} \quad f = \begin{bmatrix} f_+ \\ f_- \end{bmatrix} \in H^\omega. \tag{7.8}$$

Then, upon making subsequent use of (7.6), (7.7), (7.8) and of the explicit formula (7.5) for $K_{t_0,b,+}$, we get (7.3):

$$\begin{aligned} D_{\omega,\,b}(t_0) = \|K_{t_0,\,b}\|^2_{H^\omega} &= \lim_{z \to t_0} \langle K_{t_0,\,b}, K_z \rangle_{H^\omega} \\ &= \lim_{z \to t_0} K_{t_0,b,+}(z) = \lim_{z \to t_0} \frac{1 - \omega(z)b^*}{1 - z\bar{t}_0}\ . \end{aligned} \tag{7.9}$$

Case 2: Let $D_{\omega,\,b}(t_0) = \infty$ and $\liminf\limits_{z \to t_0} \dfrac{1 - |\omega(z)|^2}{1 - |z|^2} < \infty$.

The second assumption guarantees (by Theorem 2.3), that there exists the nontangential limit $\omega(t_0) = \lim\limits_{z\to t_0}\omega(z)$ and that the function $K_{t_0,\omega(t_0)}$ defined via (7.4), belongs to H^ω. Then by virtue of (7.9), we have

$$\lim_{z\to t_0}\frac{1-\omega(z)\omega(t_0)^*}{1-z\bar{t}_0} = D_{\omega,\omega(t_0)}(t_0) = \|K_{t_0,\omega(t_0)}\|^2_{H^\omega} < \infty.$$

Since $D_{\omega,b}(t_0) = \infty$, it follows that $b \neq \omega(t_0)$. It remains to note that (7.3) again holds since

$$\lim_{z\to t_0}\frac{1-\omega(z)b^*}{1-z\bar{t}_0} = \lim_{z\to t_0}\frac{1-\omega(z)\omega(t_0)^*+\omega(z)(b^*-\omega(t_0)^*)}{1-z\bar{t}_0} = \infty.$$

Case 3: Let $D_{\omega,\,b}(t_0) = \infty$ and $\liminf\limits_{z\to t_0}\dfrac{1-|\omega(z)|^2}{1-|z|^2} = \infty$. Since

$$\begin{aligned} 2\Re(1-\omega(z)b^* &= (1-\omega(z)b^*)+(1-b\omega(z)^*) \\ &= |1-\omega(z)b^*|^2+1-|b|^2|\omega(z)|^2 \geq 1-|b|^2|\omega(z)|^2, \end{aligned}$$

it follows that if $|b|\leq 1$, then

$$|1-\omega(z)b^*| \geq \Re(1-\omega(z)b^*) \geq \frac{1}{2}(1-|\omega(z)|^2). \tag{7.10}$$

Furthermore, for every z in the following nontangential neighborhood

$$\Gamma_a(t_0) = \{z\in\mathbb{D}:\ |t_0-z| < a(1-|z|)\}, \quad a>1,$$

of t_0, we have

$$\frac{1-|z|^2}{|1-z\bar{t}_0|} \geq \frac{1-|z|}{|1-z\bar{t}_0|} > \frac{1}{a}$$

which together with (7.10) leads us to

$$\left|\frac{1-\omega(z)b^*}{1-z\bar{t}_0}\right| \geq \frac{1}{2}\,\frac{1-|\omega(z)|^2}{|1-z\bar{t}_0|} = \frac{1}{2}\,\frac{1-|\omega(z)|^2}{1-|z|^2}\cdot\frac{1-|z|^2}{|1-z\bar{t}_0|} > \frac{1}{2a}\,\frac{1-|\omega(z)|^2}{1-|z|^2}.$$

Therefore,

$$\lim_{z\to t_0}\frac{1-\omega(z)b^*}{1-z\bar{t}_0} = \infty = D_{\omega,b}(t_0),$$

which completes the proof of the theorem. □

Corollary 7.3. *If a Schur function ω is analytic in a neighborhood of $t_0\in\mathbb{T}$ and $|\omega(t_0)| = 1$, then $D_{\omega,\omega(t_0)}(t_0) < \infty$. In particular, $D_{\omega,\omega(t_0)}(t_0) < \infty$ for every rational $\omega\in\mathcal{S}$ with $|\omega(t_0)| = 1$.*

Proof. If w meets the assumed properties, then the limit

$$\lim_{z\to t_0}\frac{1-\omega(z)\overline{\omega(t_0)}}{1-z\bar{t}_0} = \lim_{z\to t_0}\left(\frac{\omega(t_0)-\omega(z)}{t_0-z}\right)\frac{\overline{\omega(t_0)}}{\bar{t}_0} = \omega'(t_0)\frac{\overline{\omega(t_0)}}{\bar{t}_0}$$

is finite, then, by the fourth assertion in Lemma 7.2, $D_{\omega,\,\omega(t_0)}(t_0) < \infty$. □

The next theorem presents an explicit formula for the gap $\gamma_i - d_{w,n_i}(t_i)$ for any solution w of Problem 1.5. Recall that by Theorem 6.3, all solutions of Problem 1.5 are parametrized by formula (6.8).

Theorem 7.4. *Let w be a solution of Problem* 1.5*, i.e., a function of the form* (6.8)*,*

$$w(z) = s_0(z) + s_2(z)\left(1 - \mathcal{E}(z)s(z)\right)^{-1}\mathcal{E}(z)s_1(z) \tag{7.11}$$

with a parameter $\mathcal{E} \in \mathcal{S}$. Then for $i = 1, \dots, k$,

$$\gamma_i - d_{w,n_i}(t_i) = \frac{1}{((n_i+1)!)^2} \cdot \frac{|s_2^{(n_i+1)}(t_i)|^2}{D_{\mathcal{E},s(t_i)^*}(t_i) + D_{s,s(t_i)}(t_i)}, \tag{7.12}$$

where $D_{\mathcal{E},s(t_i)^*}(t_i)$ and $D_{s,s(t_i)}(t_i)$ are defined according to (7.1)*.*

Proof. Since w is a solution of Problem 1.5 and therefore satisfies conditions (1.26)–(1.28), it follows by Lemma 7.1 that

$$w_{2n_i+1}(t_i) = \lim_{z\to t_i} \frac{w(z) - c_{i,0} - (z-t_i)c_{i,1} - \ldots - (z-t_i)^{2n}c_{i,2n_i}}{(z-t_i)^{2n_i+1}}$$

for $i = 1, \dots, k$. Since s_0 is a solution of Problem 1.2 (by the first statement in Theorem 6.4), we have (again by Lemma 7.1)

$$c_{i,2n_i+1} = \lim_{z\to t_i} \frac{s_0(z) - c_{i,0} - (z-t_i)c_{i,1} - \ldots - (z-t_i)^{2n}c_{i,2n_i}}{(z-t_i)^{2n_i+1}}.$$

Now it follows from the two latter equalities that

$$c_{i,2n_i+1} - w_{2n_i+1}(t_i) = \lim_{z\to t_i} \frac{s_0(z) - w(z)}{(z-t_i)^{2n_i+1}},$$

which being substituted into (1.24), leads us to to

$$\gamma_i - d_{w,n_i}(t_i) = (-1)^{n_i} t_i^{2n_i+1} \lim_{z\to t_i} \frac{s_0(z) - w(z)}{(z-t_i)^{2n_i+1}}\, {c_{i,0}}^*.$$

Substituting (7.11) into the latter equality gives

$$\gamma_i - d_{w,n_i}(t_i) = -(-1)^{n_i} t_i^{2n_i+1} \lim_{z\to t_i} \frac{s_2(z)(1-\mathcal{E}(z)s(z))^{-1}\mathcal{E}(z)s_1(z)}{(z-t_i)^{2n_i+1}}\, c_{i,0}^*. \tag{7.13}$$

Taking into account relations (6.32) (i.e., the fact that t_i is a zero of multiplicity n_i of s_1 and s_2), we rephrase (7.13) as

$$\gamma_i - d_{w,n_i}(t_i) = \frac{(-1)^{n_i} t_i^{2n_i+2}}{((n_i+1)!)^2} s_2^{(n_i+1)}(t_i) \lim_{z\to t_i} \frac{(1-z\bar{t}_i)\mathcal{E}(z)}{1-\mathcal{E}(z)s(z)} s_1^{(n_i+1)}(t_i) c_{i,0}^*. \tag{7.14}$$

Due to (6.33), the latter equality simplifies to

$$\gamma_i - d_{w,n_i}(t_i) = \frac{|s_2^{(n_i+1)}(t_i)|^2}{((n_i+1)!)^2} \lim_{z\to t_i} \frac{\mathcal{E}(z)s(t_i)(1-z\bar{t}_i)}{1-\mathcal{E}(z)s(z)}. \tag{7.15}$$

Since $|s(t_i)| = 1$ (by Lemma 6.5), we have

$$\frac{1-\mathcal{E}(z)s(z)}{1-z\bar{t}_i} = \frac{1-s(z)s(t_i)^*}{1-z\bar{t}_i} + s(z)\frac{1-\mathcal{E}(z)s(t_i)}{1-z\bar{t}_i}s(t_i)^*. \tag{7.16}$$

By the fourth assertion of Lemma 7.2,

$$\lim_{z\to t_i}\frac{1-s(z)s(t_i)^*}{1-z\bar{t}_i}=D_{s,s(t_i)}(t_i),\quad \lim_{z\to t_i}\frac{1-\mathcal{E}(z)s(t_i)}{1-z\bar{t}_i}=D_{\mathcal{E},s(t_i)^*}(t_i). \tag{7.17}$$

Taking advantage of (7.17) we pass to limits in (7.16) as $z\to t_i$ to get

$$\lim_{z\to t_i}\frac{1-\mathcal{E}(z)s(z)}{1-z\bar{t}_i}=D_{s,s(t_i)}(t_i)+D_{\mathcal{E},s(t_i)^*}(t_i). \tag{7.18}$$

Since s is rational and $|s(t_i)|=1$, it follows by Corollary 7.3 that $D_{s,s(t_i)}(t_i)$ is finite. Since, by Theorem 6.4, $\mathbf{S}(t)$ is unitary for $t\in\mathbb{T}$ and $s_1(z), s_2(z)$ are not identical zeros, then $s(z)$ is not a unimodular constant. Therefore, $D_{s,s(t_i)}(t_i)\neq 0$, by the third assertion in Lemma 7.2. Thus,

$$0<D_{s,s(t_i)}(t_i)<\infty.$$

If the second limit in (7.17) is also finite, then

$$\lim_{z\to t_i}\mathcal{E}(z)=s(t_i)^*,$$

by the first assertion in Lemma 7.2. Therefore, (7.15) turns into (7.12) in this case. If $D_{\mathcal{E},s(t_i)^*}(t_i)$ is infinite, then, in view of (7.18), the denominator in (7.15) tends to ∞. Since the numerator $\mathcal{E}(z)s(t_i)$ is bounded, the limit in (7.15) is 0. Thus, (7.12) holds in this case also. Theorem follows. □

Proof of Statement 2 *in Theorem* 1.6. As it was already pointed out, w is a solution of Problem 1.2 if and only if it is of the form (7.11) with some (uniquely determined) parameter $\mathcal{E}\in\mathcal{S}$ and satisfies

$$\delta_{w,i}:=\gamma_i-d_{w,n_i}(t_i)=0\quad(i=1,\dots,k).$$

The formula for $\delta_{w,i}$ is given in (7.12) and it is easily seen that $\delta_{w,i}=0$ if and only if

$$D_{\mathcal{E},\,s(t_i)^*}(t_i)+D_{s,\,s(t_i)}(t_i)=\infty.$$

Since $D_{s,\,s(t_i)}(t_i)<\infty$ (by Corollary 7.3), the latter is equivalent to $D_{\mathcal{E},\,s(t_i)^*}(t_i)=\infty$ which happens, by the second assertion in Lemma 7.2, if and only if either

$$\liminf_{z\to t_i}\frac{1-|\mathcal{E}(z)|^2}{1-|z|^2}=\infty,$$

or the function $\mathcal{E}$ fails to have the nontangential limit $s(t_i)^*$ at t_i. □

Note that vanishing of the gap at the point t_i depends on the local behavior of the parameter $\mathcal{E}$ at this point only. The number $s(t_i)^*$ absorbs all the interpolation data, though. Note also that the maximum value of the gap $\delta_{w,i}$ is assumed when $D_{\mathcal{E},s(t_i)^*}(t_i)=0$, which happens if and only if $\mathcal{E}(z)\equiv s(t_i)^*$.

References

[1] D. Z. Arov and L. Z. Grossman, *Scattering matrices in the theory of unitary extensions of isometric operators*, Soviet Math. Dokl. **270** (1983), 17–20.

[2] D. Z. Arov and L. Z. Grossman, *Scattering matrices in the theory of unitary extensions of isometric operators*, Math. Nachr. **157** (1992), 105–123.

[3] J. A. Ball, *Interpolation problems of Pick–Nevanlinna and Loewner type for meromorphic matrix functions*, Integral Equations Operator Theory **6** (1983), 804–840.

[4] J. A. Ball, I. Gohberg and L. Rodman, *Interpolation of rational matrix functions*, Birkhäuser Verlag, Basel, 1990.

[5] J. A. Ball and J. W. Helton, *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.

[6] V. Bolotnikov and H. Dym, *On boundary interpolation for matrix Schur functions*, Mem. Amer. Math. Soc. 181 (2006), no. 856.

[7] V. Bolotnikov and A. Kheifets, *A higher multiplicity analogue of the Carathéodory–Julia theorem*, J. Funct. Anal. **237** no. 1 (2006), 350–371.

[8] V. Bolotnikov and A. Kheifets, *Carathéodory–Julia type conditions and symmetries of boundary asymptotics for analytic functions on the unit disk*, Math. Nachr., to appear.

[9] C. Carathéodory, *Über die Winkelderivierten von beschränkten analytischen Funktionen*, Sitz. Preuss. Akad. Phys.-Math. **4** (1929), 1–18.

[10] G. Julia, *Extension d'un lemma de Schwartz*, Acta Math. **42** (1920), 349–355.

[11] V. Katsnelson, A. Kheifets and P. Yuditskii, *An abstract interpolation problem and extension theory of isometric operators*, in: *Operators in Spaces of Functions and Problems in Function Theory* (V.A. Marchenko, ed.), **146**, Naukova Dumka, Kiev, 1987, pp. 83–96. English transl. in: *Topics in Interpolation Theory* (H. Dym, B. Fritzsche, V. Katsnelson and B. Kirstein, eds.), Oper. Theory Adv. Appl., **OT 95**, Birkhäuser Verlag, Basel, 1997, pp. 283–298.

[12] A. Kheifets, *The Parseval equality in an abstract problem of interpolation, and the union of open systems*, Teor. Funktsii Funktsional. Anal. i Prilozhen. **49**, 1988, 112–120, and **50**, 1988, 98–103. English transl. in: *J. Soviet Math.* **49** no. 4 (1990=, 114–1120, and **49** no. 6 (1990), 1307–1310.

[13] A. Kheifets *Scattering matrices and Parseval Equality in Abstract Interpolation Problem*, Ph.D. Thesis, 1990, Khrakov State University (Russian)

[14] A. Kheifets and P. Yuditskii, *An analysis and extension approach of V. P. Potapov's approach to scheme interpolation problems with applications to the generalized bi–tangential Schur–Nevanlinna–Pick problem and J–inner–outer factorization*, in: *Matrix and Operator-Valued Functions* (I. Gohberg and L.A. Sakhnovich, eds.), Oper. Theory Adv. Appl., **OT 72**, Birkhäuser Verlag, Basel, 1994, pp. 133–161.

[15] A. Kheifets, *The abstract interpolation problem and applications*, in: *Holomorphic spaces* (Ed. D. Sarason, S. Axler, J. McCarthy), Cambridge Univ. Press, Cambridge, 1998, pp. 351–379.

[16] I. V. Kovalishina, *A multiple boundary interpolation problem for contractive matrix-valued functions in the unit circle*, Teoriya Funktsii, Funktsianal'nyi Analiz i Ikh Prilozheniya **51** (1989), 38–55. English transl. in: Journal of Soviet Mathematics **52** no. 6 (1990), 3467–3481.

[17] M. G. Kreĭn and A. A. Nudelman, *The Markov moment problem and extremal problems*, Translations of Mathematical Monographs **50**, Amer. Math. Soc., Providence, Rhode Island , 1977.

[18] B. Sz.-Nagy, C. Foias, *Harmonic analysis of operators on Hilbert space*, North-Holland Publishing Co., Amsterdam-London; American Elsevier Publishing Co., New York; Akadémiai Kiadó, Budapest, 1970

[19] N. Nikolskii, *Treatise on the Shift Operator*, Springer-Verlag, Berlin, 1986.

[20] N. Nikolskii and V. Vasyunin, *A unified approach to function models, and the transcription problem*, The Gohberg anniversary collection, Vol. II (Calgary, AB, 1988), Oper. Theory Adv. Appl. **41** (1989), 405–434, Birkhäuser Verlag, Basel.

[21] D. Sarason, *Nevanlinna–Pick interpolation with boundary data*, Integral Equations Operator Theory **30** (1998), 231–250.

[22] D. Sarason, *Sub–Hardy Hilbert Spaces in the Unit Disk*, Wiley, New York, 1994.

[23] V. I. Smirnov, *Sur les formules de Cauchy et de Green et quelques problèms qui s'y rattachent* (in French), Izv. Akad. Nauk SSSR, Ser. Mat. **3** (1932), 338 - 372.

Vladimir Bolotnikov
Department of Mathematics
The College of William and Mary
Williamsburg, VA 23187-8795
USA
e-mail: vladi@math.wm.edu

Alexander Kheifets
Department of Mathematics
University of Massachusetts Lowell,
Lowell, MA 01854
USA
e-mail: Alexander_Kheifets@uml.edu

Operator Theory:
Advances and Applications, Vol. 179, 103–109

A Generalization to Ordered Groups of a Kreĭn Theorem

Ramón Bruzual and Marisela Domínguez

Abstract. We give an extension result for positive definite operator-valued Toeplitz-Kreĭn-Cotlar triplets defined on an interval of an ordered group. When the triplet is positive definite and measurable we give a representation result.

Mathematics Subject Classification (2000). Primary 47A20; Secondary 43A35.

Keywords. Positive definite, Toeplitz kernel, ordered group.

1. Introduction

Let a be such that $0 < a \leq +\infty$ and let $I = (-a, a)$. A kernel on I is a function $K : I \times I \to \mathbf{C}$. The kernel K is said to be positive definite if for any positive integer n and any $x_1, \ldots, x_n$ in I, $\lambda_1, \ldots, \lambda_n$ in $\mathbf{C}$ we have

$$\sum_{i,j=1}^{n} K(x_i, x_j)\lambda_i\overline{\lambda_j} \geq 0,$$

and K is said to be a Toeplitz kernel if there exists a function $k : I - I \to \mathbf{C}$ such that $K(x, y) = k(x - y)$ for all x , y in I.

M.G. Kreĭn [10] proved that every continuous positive definite Toeplitz kernel on $I = (-a, a)$ can be extended to a continuous positive definite Toeplitz kernel on the whole line.

The concept of positive definite kernel on I can be extended to a more general context in the following natural way: Let $\mathcal{H}$ be a Hilbert space. A kernel $K : I \times I \to L(\mathcal{H})$ is said to be positive definite if

$$\sum_{x,y \in I} \langle K(x, y)h(x), h(y)\rangle_{\mathcal{H}} \geq 0$$

for each function $h : I \to \mathcal{H}$ of finite support.

Both authors were supported in part by the CDCH of the Univ. Central de Venezuela and by FONACIT grant G-97000668.

M.L. Gorbachuk [9] extended the result of Kreĭn for operator-valued continuous Toeplitz kernels defined on an interval $I = (-a, a)$. By the Naimark dilation theorem (see [13, Theorem 7.1]), we have that a continuous positive definite Toeplitz kernel on an interval I is of the form

$$K(x) = \tau^* U_x \tau \quad \text{for all} \quad x \in I$$

where $\{U_x\}_{x\in\mathbb{R}}$ is a strongly continuous unitary representation of $\mathbb{R}$ on a larger Hilbert space $\mathcal{G}$ and $\tau : \mathcal{H} \to \mathcal{G}$ is a bounded operator.

Several problems in analysis led Cotlar and Sadosky [6] to introduce the so-called generalized Toeplitz kernels, as kernels defined in $\mathbb{Z} \times \mathbb{Z}$ or in $\mathbb{R} \times \mathbb{R}$. These kernels satisfy a condition more general than being Toeplitz in their domain. Also generalized Toeplitz kernels with domain a product of intervals on the real line have been considered.

An approach to continuous operator-valued generalized Toeplitz kernels with domain a product of intervals on the real line was given in [3], where an extension result was obtained.

In this paper we will consider a more general concept than the generalized Toeplitz kernel on an interval of the real line; we will consider operator-valued Toeplitz-Kreĭn-Cotlar triplets defined on an interval of an ordered group. We obtain an extension result, which extends Kreĭn theorem, and a representation result which extends a Crum result for this forms ([7], see also [4]).

2. Preliminaries

If Ω is an abelian group, Λ is a subset of Ω and $L(\mathcal{H})$ stands for the space of the bounded linear operators of a Hilbert space $\mathcal{H}$, a function $F : \Lambda \to L(\mathcal{H})$ is said to be *positive definite* if

$$\sum_{x,y\in\Omega} \langle F(x-y)h(x), h(y)\rangle_{\mathcal{H}} \geq 0$$

for every function $h : \Omega \to \mathcal{H}$ with finite support, such that $\text{support}(h) - \text{support}(h)$ is contained in Λ.

Proposition 2.1. *Suppose that $F : \Omega \to L(\mathcal{H})$ is positive definite. Then*

(a) *If Ω is a topological group and F is weakly continuous on a neighborhood of 0, then F is weakly continuous on Ω.*
(b) *If Ω is a locally compact group and F is weakly measurable on a neighborhood of 0, then F is weakly measurable on Ω.*

Proof. For $h \in \mathcal{H}$ the scalar-valued function

$$\omega \mapsto \langle F(\omega)h, h\rangle$$

is positive definite.

From the corresponding results for scalar-valued functions (see [12, pages 24, 91]) and the polarization formula the result follows. □

3. Toeplitz-Kreĭn-Cotlar triplets on ordered groups

Let $(\Gamma, +)$ be an abelian group with neutral element 0_Γ. Γ is an *ordered group* if there exists a set $\Gamma_+ \subset \Gamma$ such that:

$$\Gamma_+ + \Gamma_+ = \Gamma_+, \qquad \Gamma_+ \bigcap (-\Gamma_+) = \{0_\Gamma\}, \qquad \Gamma_+ \bigcup (-\Gamma_+) = \Gamma.$$

In this case if $x, y \in \Gamma$, we write $x \leq y$ if $y - x \in \Gamma_+$, we also write $x < y$ if $x \leq y$ and $x \neq y$, so $\Gamma_+ = \{\gamma \in \Gamma : \gamma \geq 0_\Gamma\}$. If there is no possibility of confusion, we will use 0 instead of 0_Γ. When Γ is a topological group it is supposed that Γ_+ is closed.

If $a, b \in \Gamma$ and $a < b$,

$$(a, b) = \{x \in \Gamma : a < x < b\}, \qquad [a, b] = \{x \in \Gamma : a \leq x \leq b\}, \qquad \text{etc.}$$

In the following Γ is an ordered group, $\mathcal{H}_1$, $\mathcal{H}_2$ are Hilbert spaces and $L(\mathcal{H}_1, \mathcal{H}_2)$ stands for the space of the continuous linear operators from $\mathcal{H}_1$ to $\mathcal{H}_2$, and $L(\mathcal{H}_\alpha)$ indicates the space of the continuous linear operators from $\mathcal{H}_\alpha$ to itself (for $\alpha = 1, 2$).

For $a \in \Gamma$, $a > 0$ let $Q_1 = [0, a]$ and $Q_2 = [-a, 0]$.

Proposition 3.1. *Let $a \in \Gamma$, $a > 0$, and let $\mathcal{H}$ be a Hilbert space. If the function $B : [-a, a] \to L(\mathcal{H})$ satisfies*

$$\sum_{x,y \in [0,a]} \langle B(x - y)h(x), h(y)\rangle_{\mathcal{H}} \geq 0$$

for all finite support function $h : [0, a] \to \mathcal{H}$, then B is positive definite on $[-a, a]$.

Proof. Let $h : \Gamma \to \mathcal{H}$ with finite support such that $\text{support}(h) - \text{support}(h) \subset [-a, a]$. Suppose that $\text{support}(h) = \{\gamma_1, \ldots, \gamma_n\}$, where $\gamma_1 < \gamma_2 < \cdots < \gamma_n$, then $\gamma_n - \gamma_1 \leq a$.

Consider $h'(\gamma) = h(\gamma + \gamma_1)$; we have $\text{support}(h') \subset [0, a]$ and

$$\sum_{x,y \in \Gamma} \langle B(x - y)h(x), h(y)\rangle_{\mathcal{H}} = \sum_{x,y \in [0,b]} \langle B(x - y)h'(x), h'(y)\rangle_{\mathcal{H}} \geq 0. \qquad \square$$

Definition 3.2. Let $a \in \Gamma$, $a > 0$. A *Toeplitz-Kreĭn-Cotlar triplet*, $\mathbf{C}$, on $(\Gamma, [0, a], \mathcal{H}_1, \mathcal{H}_2)$ consists of three functions

$$C_{\alpha\beta} : Q_\alpha - Q_\beta \to L(\mathcal{H}_\alpha, \mathcal{H}_\beta) \quad \alpha, \beta = 1, 2, \ \alpha \leq \beta.$$

If $\mathbf{C}$ is a Toeplitz-Kreĭn-Cotlar triplet we define $C_{21}(\gamma) = C_{12}(-\gamma)^*$ for $\gamma \in Q_2 - Q_1$.

Definition 3.3. We shall say that the Toeplitz-Kreĭn-Cotlar triplet $\mathbf{C}$ on $(\Gamma, [0, a], \mathcal{H}_1, \mathcal{H}_2)$ is *positive definite* if

$$\sum_{\alpha,\beta=1}^{2} \sum_{(x,y) \in Q_\alpha \times Q_\beta} \langle C_{\alpha\beta}(x - y)h_\alpha(x), h_\beta(y)\rangle_{\mathcal{H}_\beta} \geq 0$$

for all pairs of functions $h_\alpha : \Gamma \to \mathcal{H}_\alpha$ with finite support, such that $\text{support}(h_\alpha) - \text{support}(h_\beta)$ is contained in $Q_\alpha - Q_\beta$, $\alpha, \beta = 1, 2$.

Remark 3.4. A Toeplitz-Kreĭn-Cotlar triplet is a particular case of a Toeplitz-Kreĭn-Cotlar form, according to the definition given in [1].

4. Extension results

Theorem 4.1. *Let Γ be an abelian ordered group and let $\mathcal{H}_1$,$\mathcal{H}_2$ be a pair of Hilbert spaces. If $\mathbf{C} = (C_{\alpha\beta})$ is a positive definite Toeplitz-Kreĭn-Cotlar triplet on $(\Gamma, [0,a], \mathcal{H}_1, \mathcal{H}_2)$, then there exist a Hilbert space $\mathcal{G}$, a unitary representation $(U_\gamma)_{\gamma\in\Gamma}$ of Γ on $L(\mathcal{G})$ and two bounded operators $\tau_\alpha : \mathcal{H}_\alpha \to \mathcal{G}$ such that*

(a) *$C_{\alpha\beta}(\gamma) = \tau_\beta^* U_\gamma \tau_\alpha$ for $\gamma \in Q_\alpha - Q_\beta$, $\alpha, \beta = 1, 2$;*
(b) *$\mathcal{G} = \bigvee\{U_\gamma \tau_1 h_1 : \gamma \in \Gamma,\ h_1 \in \mathcal{H}_1\} \vee \bigvee\{U_\gamma \tau_2 h_2 : \gamma \in \Gamma,\ h_2 \in \mathcal{H}_2\}$;*
(c) *if Γ is topological and $\mathbf{C}$ is weakly continuous, then U_γ is strongly continuous;*
(d) *if Γ is locally compact and $\mathbf{C}$ is weakly measurable, then U_γ is weakly measurable.*

In order to prove this theorem we need the following result.

Proposition 4.2. *Let $\mathbf{C}$ be a positive definite Toeplitz-Kreĭn-Cotlar triplet on $(\Gamma, [0,a], \mathcal{H}_1, \mathcal{H}_2)$. Then the function $B : [-a,a] \to L(\mathcal{H}_1 \oplus \mathcal{H}_2)$ defined by*

$$B(\gamma) = \begin{pmatrix} C_{11}(\gamma) & C_{21}(\gamma - a) \\ C_{12}(\gamma + a) & C_{22}(\gamma) \end{pmatrix}$$

is positive definite on $[-a,a]$.

Proof. Let $h : [0,a] \to \mathcal{H}_1 \oplus \mathcal{H}_2$ be a finite support function, thus $h = h_1 \oplus h_2$, where $h_1 : [0,a] \to \mathcal{H}_1$ and $h_2 : [0,a] \to \mathcal{H}_2$ are finite support functions. Then

$$\begin{aligned}
&\sum_{x,y\in[0,a]} \langle B(x-y)h(x), h(y)\rangle_{\mathcal{H}} \\
&= \sum_{x,y\in[0,a]} \langle C_{11}(x-y)h_1(x), h_1(y)\rangle_{\mathcal{H}_1} + \sum_{x,y\in[0,a]} \langle C_{21}(x-y-a)h_2(x), h_1(y)\rangle_{\mathcal{H}_1} \\
&+ \sum_{x,y\in[0,a]} \langle C_{12}(x-y+a)h_1(x), h_2(y)\rangle_{\mathcal{H}_2} + \sum_{x,y\in[0,a]} \langle C_{22}(x-y)h_2(x), h_2(y)\rangle_{\mathcal{H}_2} \\
&= \sum_{x,y\in\Gamma} \langle C_{11}(x-y)h_1(x), h_1(y)\rangle_{\mathcal{H}_1} + \sum_{x,y\in\Gamma} \langle C_{21}(x-y)h_2(x+a), h_1(y)\rangle_{\mathcal{H}_1} \\
&+ \sum_{x,y\in\Gamma} \langle C_{12}(x-y)h_1(x), h_2(y+a)\rangle_{\mathcal{H}_2} + \sum_{x,y\in\Gamma} \langle C_{22}(x-y)h_2(x+a), h_2(y+a)\rangle_{\mathcal{H}_2}.
\end{aligned}$$

This sum is nonnegative because $\mathbf{C}$ is positive definite, $\operatorname{support}(h_1) \subset [0,a]$ and $\operatorname{support}(g_2) \subset [-a,0]$, where $g_2(x) = h_2(x+a)$. □

Proof of Theorem 4.1. (a) Consider the function $B : [-a,a] \to L(\mathcal{H}_1 \oplus \mathcal{H}_2)$ defined by

$$B(\gamma) = \begin{pmatrix} C_{11}(\gamma) & C_{21}(\gamma - a) \\ C_{12}(\gamma + a) & C_{22}(\gamma) \end{pmatrix}.$$

From Proposition 4.2 it follows that B is positive definite on $[-a,a]$ and from Theorem 2.1 of [2] (with a natural modification in order to consider a closed interval) it follows that B can be extended to a positive definite function $F:\Gamma\to L(\mathcal{H}_1\oplus\mathcal{H}_2)$.

If $F=(F_{\alpha\beta})_{\alpha\beta=1,2}$, then the function

$$\widetilde{F}(\gamma)=\begin{pmatrix}F_{11}(\gamma) & F_{21}(\gamma+a)\\ F_{12}(\gamma-a) & F_{22}(\gamma)\end{pmatrix}$$

is also positive definite on Γ. From Naimark's theorem (see [13, Theorem 7.1]) it follows that there exists a Hilbert space $\mathcal{G}$, a unitary representation (U_γ) of Γ on $L(\mathcal{G})$ and a bounded operator $R:\mathcal{H}_1\oplus\mathcal{H}_2\to\mathcal{G}$ such that

$$\widetilde{F}(\gamma)=R^*U_\gamma R.$$

Let $i_\alpha:\mathcal{H}_\alpha\to\mathcal{H}_1\oplus\mathcal{H}_2$ be the canonical immersion ($\alpha=1,2$) and let $\gamma\in[-a,a]$. We have that

$$C_{\alpha\alpha}(\gamma)=i_\alpha^*\,R^*\,U_\gamma\,R\,i_\alpha,$$

so if $\tau_\alpha=R\,i_\alpha$, then τ_α is a bounded operator and

$$C_{\alpha\alpha}(\gamma)=\tau_\alpha^*\,U_\gamma\,\tau_\alpha.$$

We also have that

$$C_{12}(\gamma+a)=\tau_2^*\,U_{\gamma+a}\,\tau_1,$$

so

$$C_{12}(\sigma)=\tau_2^*\,U_\sigma\,\tau_1,$$

for $\sigma\in[0,2a]$, in the same way the result is obtained for C_{21}.

(b) In order to obtain the minimality condition it is enough to replace $\mathcal{G}$ by

$$\bigvee\{U_\gamma\tau_1h_1:\gamma\in\Gamma,\ h_1\in\mathcal{H}_1\}\vee\bigvee\{U_\gamma\tau_2h_2:\gamma\in\Gamma,\ h_2\in\mathcal{H}_2\}.$$

If $\mathbf{C}$ is weakly continuous, then B is weakly continuous and if $\mathbf{C}$ is weakly measurable, then B is weakly measurable.

Since $(-a,a)$ is a neighborhood of 0, from Proposition 2.1 and from the general fact that the minimal Naimark dilation of a weakly continuous positive definite function is is strongly continuous we obtain (c).

From the corresponding measurability result we obtain (d). □

5. Representation results

Theorem 5.1. *Let Γ be a locally compact abelian ordered group and let $\mathcal{H}_1$,$\mathcal{H}_2$ be a pair of separable Hilbert spaces. If $\mathbf{C}=(C_{\alpha\beta})$ is a weakly measurable positive definite Toeplitz-Kreĭn-Cotlar triplet on $(\Gamma,[0,a],\mathcal{H}_1,\mathcal{H}_2)$, then there exist two Toeplitz-Kreĭn-Cotlar triplets on $(\Gamma,[0,a],\mathcal{H}_1,\mathcal{H}_2)$, $\mathbf{C^c}=(C^c_{\alpha\beta})$ and $\mathbf{C^0}=(C^0_{\alpha\beta})$ such that*

(a) $C_{\alpha\beta}=C^c_{\alpha\beta}+C^0_{\alpha\beta}$ *for* $\alpha,\beta=1,2$;

(b) $\mathbf{C^c} = (C^c_{\alpha\beta})$ *is positive definite and weakly continuous;*
(c) $\mathbf{C^0} = (C^0_{\alpha\beta})$ *is positive definite and each* $C^0_{\alpha\beta}$ *is zero locally almost everywhere.*

Proof. From Theorem 4.1 there exist a Hilbert space $\mathcal{G}$, a weakly measurable unitary representation $(U_\gamma)_{\gamma\in\Gamma}$ of Γ on $L(\mathcal{G})$ and two bounded operators $\tau_\alpha : \mathcal{H}_\alpha \to \mathcal{G}$ such that $C_{\alpha\beta}(\gamma) = \tau^*_\beta U_\gamma \tau_\alpha$ for $\gamma \in Q_\alpha - Q_\beta$, $\alpha, \beta = 1, 2$.

Let $\tau : \mathcal{H}_1 \oplus \mathcal{H}_2 \to \mathcal{G}$ defined by $\tau(h_1 \oplus h_2) = \tau_1 h_1 + \tau_2 h_2$; then the function $F : \Gamma \to L(\mathcal{H}_1 \oplus \mathcal{H}_2)$ defined by

$$F(\gamma) = \tau^* U_\gamma \tau = \begin{pmatrix} \tau_1^* U_\gamma \tau_1 & \tau_1^* U_\gamma \tau_2 \\ \\ \tau_2^* U_\gamma \tau_1 & \tau_2^* U_\gamma \tau_2 \end{pmatrix},$$

is positive definite. Since $(U_\gamma)_{\gamma\in\Gamma}$ is weakly measurable, F is weakly measurable, so from the main result of [8] it follows that there exist two functions $F^c : \Gamma \to L(\mathcal{H}_1 \oplus \mathcal{H}_2)$ and $F^0 : \Gamma \to L(\mathcal{H}_1 \oplus \mathcal{H}_2)$ such that

(a) $F = F^c + F^0$;
(b) F^c is positive definite and weakly continuous;
(c) F^0 is positive definite and zero locally almost everywhere.

Considering $C^c_{\alpha\beta} = \tau^*_\beta F^c \tau_\alpha$ and $C^0_{\alpha\beta} = \tau^*_\beta F^0 \tau_\alpha$ we obtain the result. □

References

[1] R. Arocena, *On the Extension Problem for a class of translation invariant positive forms.* J. Oper. Theory **21** (1989), 323–347.

[2] M. Bakonyi, *The extension of positive definite operator-valued functions defined on a symmetric interval of an ordered group.* Proc. Am. Math. Soc. **130** no. 5 (2002), 1401–1406.

[3] R. Bruzual, *Local semigroups of contractions and some applications to Fourier representation theorems.* Integral Equations Oper. Theory **10** (1987), 780–801.

[4] R. Bruzual, *Representation of measurable positive definite generalized Toeplitz kernels in* $\mathbb{R}$. Integral Equations Oper. Theory **29** (1997), 251–260.

[5] R. Bruzual, M. Domínguez, *Extensions of operator-valued positive definite functions and commutant lifting on ordered groups.* J. Funct. Anal. **185** (2001), 456–473.

[6] M. Cotlar, C. Sadosky, *On the Helson-Szegö theorem and a related class of modified Toeplitz kernels.* Proc. Symp. Pure Math. AMS. **35-I** (1979), 383–407.

[7] M. Crum, *On positive definite functions.* Proc. London Math. Soc. **6** (1956), 548–560.

[8] A. Devinatz, *On measurable positive definite operator functions.* Journal London Math. Soc. **35** (1960), 417–424.

[9] M. L. Gorbachuck, *Representation of positive definite operator functions.* Ukrainian Math. J. **17** (1965), 29–46.

[10] M. G. Kreĭn, *Sur le problème du prolongement des fonctions hermitiennes positives et continues.* Dokl. Akad. Nauk. SSSR **26** (1940), 17–22.

[11] F. Riesz, *Uber satze vonStone und Bochner.* Acta Univ. Szeged **6** (1933), 184–198.

[12] Z. Sasvári, *Positive definite and definitizable functions.* Akademie Verlag, 1994.

[13] B. Sz.-Nagy, C. Foias, *Harmonic analysis of operators on Hilbert space.* North Holland Publishing Co. 1970.

Ramón Bruzual
Escuela de Matemática
Fac. Ciencias
Universidad Central de Venezuela
Mailing address:
Apartado Postal 47686
Caracas 1041-A
Venezuela
e-mail: rbruzual@euler.ciens.ucv.ve
ramonbruzual@cantv.net

Marisela Domínguez
Escuela de Matemática
Fac. Ciencias
Universidad Central de Venezuela
Mailing address:
Apartado Postal 47159
Caracas 1041-A
Venezuela
e-mail: mdomin@euler.ciens.ucv.ve

Operator Theory:
Advances and Applications, Vol. 179, 111–143

A Fast QR Algorithm for Companion Matrices

Shiv Chandrasekaran, Ming Gu, Jianlin Xia and Jiang Zhu

Abstract. It has been shown in [4, 5, 6, 31] that the Hessenberg iterates of a companion matrix under the QR iterations have low off-diagonal rank structures. Such invariant rank structures were exploited therein to design fast QR iteration algorithms for finding eigenvalues of companion matrices. These algorithms require only $O(n)$ storage and run in $O(n^2)$ time where n is the dimension of the matrix. In this paper, we propose a new $O(n^2)$ complexity QR algorithm for real companion matrices by representing the matrices in the iterations in their sequentially semi-separable (SSS) forms [9, 10]. The bulge chasing is done on the SSS form QR factors of the Hessenberg iterates. Both double shift and single shift versions are provided. Deflation and balancing are also discussed. Numerical results are presented to illustrate both high efficiency and numerical robustness of the new QR algorithm.

Mathematics Subject Classification (2000). 65F15, 65H17.

Keywords. Companion matrices, sequentially semi-separable matrices, structured QR iterations, structured bulge chasing, Givens rotation swaps.

1. Introduction

After nearly forty years since its introduction [18, 19], the QR algorithm is still the method of choice for small or moderately large nonsymmetric eigenvalue problems $Ax = \lambda x$ where A is an $n \times n$ matrix. At the moment of this writing, moderately large eigenvalue problems refer to matrices of order $1,000$ or perhaps a bit higher. The main reason for such a limitation in problem size is because the algorithm runs in $O(n^3)$ time and uses $O(n^2)$ storage.

The success of the algorithm lies on doing QR iterations repeatedly, which under mild conditions [29] leads to Schur form convergence. However, for a general nonsymmetric dense matrix A, one QR decomposition itself already takes $O(n^3)$ operations, so even if we are lucky enough to do only one iteration per eigenvalue, the cost would still be $O(n^4)$. To make the algorithm practical, it is necessary to first reduce A into an upper Hessenberg matrix H and then carry out QR iterations on H accordingly. It is also important to incorporate a suitable shift strategy (since

QR iteration is implicitly doing inverse iteration), which can dramatically reduce the number of QR iterations needed for convergence.

The rationale for reducing A to H is that the Hessenberg form is invariant under QR iterations. Such Hessenberg invariance structure enables us to implement QR iterations implicitly and efficiently by means of structured bulge chasing. In practice, with the use of shifts, convergence to the Schur form occurs in $O(n)$ bulge chasing passes, each pass consists of $O(n)$ local orthogonal similarity transformations, and each local similarity transformation takes $O(n)$ operations. Therefore the total cost of the algorithm is $O(n^3)$ operations. This new algorithm has been tested for many different types of examples and is stable in practice.

In this paper we consider the eigenvalue computation of a real companion matrix of the form

$$C = \begin{pmatrix} a_1 & a_2 & \dots & a_{n-1} & a_n \\ 1 & 0 & \dots & 0 & 0 \\ 0 & 1 & \dots & 0 & 0 \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & 0 & \dots & 1 & 0 \end{pmatrix} \in \mathbb{R}^{n\times n}. \tag{1}$$

Since the eigenvalues of C coincide with the zeros of a real univariate polynomial

$$p(x) = x^n - a_1 x^{n-1} - \cdots - a_{n-1}x - a_n, \tag{2}$$

algorithms for computing matrix eigenvalues can be used to approximate the zeros of $p(x)$. In fact, the Matlab function `roots` finds the zeros of $p(x)$ by applying the implicit shift QR algorithm to C_0, a suitably balanced version of C by means of a diagonal scaling (note that C_0 is not necessarily a companion matrix). The algorithm costs $O(n^3)$ operations as we mentioned.

The $O(n^3)$ cost and $O(n^2)$ storage are still expensive for a large n. In fact, it is possible to improve the performance of QR iterations by exploiting additional invariance structures of the Hessenberg iterates of C under QR iterations. It has been shown independently in [4] and in [5, 6] that the Hessenberg iterates of a companion matrix preserve an off-diagonal low-rank structure, called sequentially semi-separable structure and semi-separable structure, respectively. This fact was then exploited to design companion eigensolvers which require only $O(n^2)$ time and $O(n)$ storage.

In this paper, we present a new $O(n^2)$ QR variant algorithm for the real companion matrix, with experiments showing numerical stability. We implement both the single shift and double shift QR iterations with compact sequentially semi-separable structures. Instead of working on the similarity transformations of C, we work on the QR factors of these matrices. A swapping strategy for Givens rotation matrices is used to efficiently conduct structured bulge chasing. To maintain compact structured forms of those QR factors we introduce a structure recovery technique. We also provide a structured balancing strategy.

The paper is organized as follows. In Section 2, we describe the sequentially semi-separable representation and some related operations including matrix additions and matrix-matrix multiplications. In Section 3, we adopt the approach in [4] to prove why all Hessenberg iterates of C have off-diagonal blocks with ranks never exceeding 3. Similar off-diagonal rank results can be easily extended to the QR factors Q and R in the QR iterations. Thus Section 4 shows the representations of Q and R in compact SSS forms. In Section 5, we describe the deflation technique and the convergence criterion of the new QR algorithm, and then by using a concrete 5×5 matrix example, we demonstrate how to implicitly do both single and double shift QR iterations based on the compact representations of Q and R. Balancing strategy, which preserves the semi-separable structure, is discussed in Section 6. In Section 7, we present numerical results to demonstrate the performance. Finally, Section 8 draws some concluding remarks.

2. SSS representation

In this section we lay out some necessary background information about sequentially semi-separable (SSS) representations [9, 10]. Closely related matrix structures include quasiseparable matrices (e.g., [14, 15]), hierarchically semi-separable matrices [8], etc. Both the name "SSS" and "quasiseparable" refer to the same type of matrices. Related matrix properties and operations are discussed in the above references. Here we use SSS representations and some associated operations in [9, 10]. Similar results also appear in [14]. They will be used in our fast structured QR iterations.

2.1. SSS notations

We say that $A \in \mathbb{R}^{n\times n}$ is in *SSS* form if it is represented as

$$A = (A_{ij}), \quad \text{where } A_{ij} \in \mathbb{R}^{m_i \times m_j},\ A_{ij} = \begin{cases} \mathcal{D}_i & \text{if } i = j, \\ \mathcal{U}_i \mathcal{W}_{i+1} \cdots \mathcal{W}_{j-1} \mathcal{V}_j^T & \text{if } i < j, \\ \mathcal{P}_i \mathcal{R}_{i-1} \cdots \mathcal{R}_{j+1} \mathcal{Q}_j^T & \text{if } i > j. \end{cases} \tag{3}$$

Here the empty products are treated as the identity matrices, and the *partitioning sequence* $\{m_i\}_{i=1}^r$ satisfies $\sum_{i=1}^r m_i = n$, with r being the number of block rows (or columns) of the partitioning scheme. The SSS *generators* $\{\mathcal{D}_i\}_{i=1}^r$, $\{\mathcal{U}_i\}_{i=1}^{r-1}$, $\{\mathcal{V}_i\}_{i=2}^r$, $\{\mathcal{W}_i\}_{i=2}^{r-1}$, $\{\mathcal{P}_i\}_{i=2}^r$, $\{\mathcal{Q}_i\}_{i=1}^{r-1}$ and $\{\mathcal{R}_i\}_{i=2}^{r-1}$ are real matrices with dimensions specified in Table 2.1.

$\mathcal{D}_i$	$\mathcal{U}_i$	$\mathcal{V}_i$	$\mathcal{W}_i$	$\mathcal{P}_i$	$\mathcal{Q}_i$	$\mathcal{R}_i$
$m_i \times m_i$	$m_i \times k_i$	$m_i \times k_{i-1}$	$k_{i-1} \times k_i$	$m_i \times l_i$	$m_i \times l_{i+1}$	$l_{i+1} \times l_i$

TABLE 1. *Dimensions of matrices in* (3).

To illustrate the compactness of this SSS representation when the off-diagonal blocks of A have small ranks, assume $m_i = k_i = l_i = p \ll n$, then we only need to

store the SSS generators of A with about $7rp^2 (= 7pn)$ working precision numbers instead of storing every entry of A with n^2 numbers.

It should be noted that the SSS structure of a given matrix A depends on the partitioning sequence $\{m_i\}_{i=1}^r$. Different sequences will lead to different representations.

The power of SSS representation for matrices with low-rank off-diagonal blocks has been shown in [9, 10, 11, 30], where fast and stable linear system solvers based on SSS representation were designed with applications to many relevant engineering problems. In [9, 10], algorithms for SSS matrix operations have been systematically introduced, including constructions of the SSS representations, (LU-like) factorizations of SSS matrices, fast SSS matrix additions and fast matrix-matrix multiplications, etc. For our purpose of designing a new QR iteration method for companion matrices, we need to use two important SSS matrix operations, SSS addition and SSS multiplication. We present the results from [9, 10] without proofs.

2.2. SSS addition

Let A and B be two SSS matrices that are conformally partitioned, that is, $m_i(A) = m_i(B)$ for $i = 1, \ldots, r$. Then their sum $A + B$ is an SSS matrix with representation given by the following SSS generators [9, 10]:

$$
\begin{aligned}
&\mathcal{D}_i(A+B) = \mathcal{D}_i(A) + \mathcal{D}_i(B),\\
&\mathcal{U}_i(A+B) = \begin{pmatrix} \mathcal{U}_i(A) & \mathcal{U}_i(B) \end{pmatrix}, \qquad \mathcal{V}_i(A+B) = \begin{pmatrix} \mathcal{V}_i(A) & \mathcal{V}_i(B) \end{pmatrix},\\
&\mathcal{W}_i(A+B) = \begin{pmatrix} \mathcal{W}_i(A) & 0 \\ 0 & \mathcal{W}_i(B) \end{pmatrix},\\
&\mathcal{P}_i(A+B) = \begin{pmatrix} \mathcal{P}_i(A) & \mathcal{P}_i(B) \end{pmatrix}, \qquad \mathcal{Q}_i(A+B) = \begin{pmatrix} \mathcal{Q}_i(A) & \mathcal{Q}_i(B) \end{pmatrix},\\
&\mathcal{R}_i(A+B) = \begin{pmatrix} \mathcal{R}_i(A) & 0 \\ 0 & \mathcal{R}_i(B) \end{pmatrix}.
\end{aligned}
$$

Remark 2.1. Note that the computed SSS representation of the sum might be inefficient in the sense that the dimensions of the SSS generators are increasing additively, whereas in some cases the real ranks of the off-diagonal blocks might be far smaller. Ideally, these formulas should be followed by some sort of rank-reduction or compression step [9, 10].

2.3. SSS multiplication

Let A and B be two SSS matrices that are conformally partitioned. Define forward and backward recursions

$$
\begin{aligned}
&S_1 = 0, \quad S_{i+1} = \mathcal{Q}_i^T(A)\mathcal{U}_i(B) + \mathcal{R}_i(A)S_i\mathcal{W}_i(B), \quad \text{for } i = 1, 2, \ldots, r-1,\\
&T_n = 0, \quad T_{i-1} = \mathcal{V}_i^T(A)\mathcal{P}_i(B) + \mathcal{W}_i(A)T_i\mathcal{R}_i(B), \quad \text{for } i = r, r-1, \ldots, 2.
\end{aligned}
$$

Then the SSS generators of the matrix $A \cdot B$ can be computed through the following formulas [9, 10]:

$$\begin{aligned}
\mathcal{D}_i(A \cdot B) &= \mathcal{D}_i(A)\mathcal{D}_i(B) + \mathcal{P}_i(A)S_i\mathcal{V}_i^T(B) + \mathcal{U}_i(A)T_i\mathcal{Q}_i^T(B),\\
\mathcal{U}_i(A \cdot B) &= \begin{pmatrix} \mathcal{D}_i(A)\mathcal{U}_i(B) + \mathcal{P}_i(A)S_i\mathcal{W}_i(B) & \mathcal{U}_i(A) \end{pmatrix},\\
\mathcal{V}_i(A \cdot B) &= \begin{pmatrix} \mathcal{V}_i(B) & \mathcal{D}_i^T(B)\mathcal{V}_i(A) + \mathcal{Q}_i(B)T_i^T\mathcal{W}_i^T(A) \end{pmatrix},\\
\mathcal{W}_i(A \cdot B) &= \begin{pmatrix} \mathcal{W}_i(B) & 0 \\ \mathcal{V}_i^T(A)\mathcal{U}_i(B) & \mathcal{W}_i(A) \end{pmatrix},\\
\mathcal{P}_i(A \cdot B) &= \begin{pmatrix} \mathcal{D}_i(A)\mathcal{P}_i(B) + \mathcal{U}_i(A)T_i\mathcal{R}_i(B) & \mathcal{P}_i(A) \end{pmatrix},\\
\mathcal{Q}_i(A \cdot B) &= \begin{pmatrix} \mathcal{Q}_i(B) & \mathcal{D}_i^T(B)\mathcal{Q}_i(A) + \mathcal{V}_i(B)S_i^T\mathcal{R}_i^T(A) \end{pmatrix},\\
\mathcal{R}_i(A \cdot B) &= \begin{pmatrix} \mathcal{R}_i(B) & 0 \\ \mathcal{Q}_i^T(A)\mathcal{P}_i(B) & \mathcal{R}_i(A) \end{pmatrix}.
\end{aligned}$$

Remark 2.2. In the case where $m_i = k_i = l_i = p$, the total operation count of this fast multiplication algorithm is at most $40p^3n$, contrasting with $2n^3$ flops for doing ordinary matrix-matrix multiplication.

3. Invariant off-diagonal low-rank structure

The classical Hessenberg QR algorithm for finding eigenvalues computes a series of Hessenberg matrices H_k which are orthogonally similar to C in (1):

$$\begin{aligned}
&H^{(0)} = C,\\
&H^{(k)} = Q^{(k)}R^{(k)},\ H^{(k+1)} = R^{(k)}Q^{(k)},\ k = 0, 1, 2, \ldots
\end{aligned}$$

Generally, shifts are used in the iterations. It has been shown independently in [4] and [5] that each such Hessenberg matrix H_k (real or complex) maintains off-diagonal low-rank structures. More precisely, the following result holds.

Theorem 3.1. [4, 5] $\max_{1 \le j < n} \operatorname{rank}(H^{(k)}(1 : j, j + 1 : n)) \le 3$.

In what follows, we concentrate on real companion matrices. The proof of the theorem relies on the results in the following two lemmas [4].

Lemma 3.2. *For any Hessenberg matrix $H^{(k)}$ in the Hessenberg QR iterations, there exist an orthogonal matrix $Z^{(k)} \in \mathbb{R}^{n \times n}$ and two vectors $x^{(k)}$, $y^{(k)} \in \mathbb{R}^n$ so that*

$$H^{(k)} = Z^{(k)} + x^{(k)}y^{(k)T}. \tag{4}$$

($H^{(k)}$ is an orthogonal-plus-rank-one structure.)

It suffices to establish the equation for $H^{(0)}$ since the structure of a low-rank modification to an orthogonal matrix is preserved under orthogonal similarity

transformations. For $H^{(0)} = C$, we can write

$$C = \begin{pmatrix} 0 & 0 & \dots & 0 & \pm 1 \\ 1 & 0 & \dots & 0 & 0 \\ 0 & 1 & \dots & 0 & 0 \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & 0 & \dots & 1 & 0 \end{pmatrix} + \begin{pmatrix} 1 \\ 0 \\ 0 \\ \vdots \\ 0 \end{pmatrix} \begin{pmatrix} a_1 & a_2 & \dots & a_{n-1} & a_n \mp 1 \end{pmatrix}$$
$$\equiv Z^{(0)} + x^{(0)} y^{(0)T}.$$

For convenience, we choose the sign of the $(1, n)$-entry of $Z^{(0)}$ so that $\det(Z^{(0)}) = 1$.

Lemma 3.3. *An orthogonal matrix Z is rank-symmetric* [4]*, in the sense that for any* 2*-by-*2 *block partitioning*

$$Z = \begin{pmatrix} Z_{11} & Z_{12} \\ Z_{21} & Z_{22} \end{pmatrix},$$

where Z_{11} and Z_{22} are square, we have $\operatorname{rank}(Z_{12}) = \operatorname{rank}(Z_{21})$.

This is a direct outcome of the CS decomposition (see [17]). Actually not only $\operatorname{rank}(Z_{12}) = \operatorname{rank}(Z_{21})$, Z_{12} and Z_{21} have the same singular values as well. Therefore, we can expect that a slightly perturbed orthogonal matrix is still numerically rank-symmetric.

Now let us prove Theorem 3.1. For simplicity of the notation, we drop the superscript (k) from (4) in the rest of this section.

Proof of Theorem 3.1. Write $L = xy^T$. According to Lemma 3.2, we have $H = Z + L$. Partition H as

$$H = \begin{pmatrix} H_{11} & H_{12} \\ H_{21} & H_{22} \end{pmatrix},$$

where H_{11} and H_{22} are square, and partition Z and L conformally. Then H_{21} has rank at most 1, since there is only one possible nonzero in its upper right corner. In addition,

$$\begin{aligned} |\operatorname{rank}(H_{12}) - \operatorname{rank}(H_{21})| &= |(\operatorname{rank}(H_{12}) - \operatorname{rank}(Z_{12})) - (\operatorname{rank}(H_{21}) - \operatorname{rank}(Z_{12}))| \\ &\le |\operatorname{rank}(H_{12}) - \operatorname{rank}(Z_{12})| + |\operatorname{rank}(H_{21}) - \operatorname{rank}(Z_{21})| \\ &\qquad \text{(since } Z \text{ is rank-symmetric)} \\ &\le \operatorname{rank}(L_{12}) + \operatorname{rank}(L_{21}) \\ &\le 2 \cdot \operatorname{rank}(L) = 2. \end{aligned}$$

Thus

$$\operatorname{rank}(H_{12}) \le \operatorname{rank}(H_{21}) + 2 \le 3. \qquad \square$$

Theorem 3.1 indicates that all H in the QR iterations have low-rank off-diagonal blocks. Such a low-rank structure admits a compact representation for H.

Bini, Eidelman, et al. [6] take advantage of this property and represent each H in a quasiseparable form which can be represented by a linear number of parameters. Similarly, the new QR algorithm proposed by Bindel, Chandresekaran, et al. in [4] exploits this structure by writing the Hessenberg iterate H in terms of its SSS representation. Both type of schemes provide explicit formulas for QR iterations with single shifts.

Because during the structured bulge chasing passes only linear memory space and only local updating for the quasiseparable or SSS generators of H are required, those new QR algorithms are able to achieve $O(n^2)$ complexity and $O(n)$ storage. To maintain the compact quasiseparable or SSS representations for H, the algorithm in [6] involves some compression schemes, and the algorithm in [4] incurs merging and splitting SSS representations repeatedly during each bulge chasing pass.

In this paper we propose a different approach for QR iterations: instead of working explicitly on the compact representations of H, we choose to work on Q and R directly, and in the meantime, to maintain compact representations for them, where Q and R are QR factors of H. This allows more flexibility in handling the structured QR iterations. Partly because of this reason we are able to provide both single shift and double shift QR iterations, whereas [4] and [6] only provide single shift versions.

We use the following theorem to characterize the similar low-rank off-diagonal structures of Q and R.

Theorem 3.4. *Suppose that a nonsingular upper Hessenberg matrix H can be expressed as $H = Z + xy^T$, with Z being orthogonal and $x, y \in \mathbb{R}^n$, and suppose that it has QR factorization: $H = QR$. Then*

1. *Q has the form: $Q = Q_1 Q_2 \cdots Q_{n-1}$, where each Q_i is a Givens rotation;*
2. *R can be written as: $R = \widetilde{Z} + \widetilde{x}y^H$, with $\widetilde{Z}$ being orthogonal. Furthermore, if we partition R as*

$$R = \begin{pmatrix} R_{11} & R_{12} \\ 0 & R_{22} \end{pmatrix},$$

where R_{11} and R_{22} are square, then

$$\operatorname{rank}(R_{12}) \leq 2.$$

Proof. For any Hessenberg matrix H, its QR decomposition can be obtained by applying a sequence of Givens rotations $\{Q_i\}_{i=1}^{n-1}$ to zero out its subdiagonal entries from the top to bottom. Specifically, we will have $Q = Q_1 Q_2 \cdots Q_{n-1}$ and

$$R = Q_{n-1}^T \cdots Q_2^T Q_1^T \cdot H = Q^T(Z + xy^T) =: \widetilde{Z} + \widetilde{x}y^T$$

where $\widetilde{Z} := Q^T Z$ and $\widetilde{x} := Q^T x$. We can then finish the proof by using inequalities similar to those in the proof of Theorem 3.1. □

4. Compact Representations of Q and R

Theorem 3.4 implies that it is possible to represent Q and R in compact forms. We dedicate this section to the detailed description of such compact representations.

4.1. Compact representations of Q

Consider an orthogonal matrix Q which can be expressed in the form

$$Q = Q_1 Q_2 \cdots Q_{n-1} \tag{5}$$

where Q_k is a Givens rotation matrix

$$Q_k = \operatorname{diag}\left(I_{k-1}, \begin{pmatrix} c_k & s_k \\ -s_k & c_k \end{pmatrix}, I_{n-k-1}\right), \ c_k, s_k \in \mathbb{R}, \ c_k^2 + s_k^2 = 1. \tag{6}$$

For convenience we call Q_k the k-th Givens (rotation) matrix. Multiplying out the product (5), it is straightforward to verify that Q takes the following form (assuming $c_0 = c_n = 1$):

$$Q = Q_1 Q_2 \cdots Q_{n-1} = \begin{pmatrix} c_0c_1 & c_0s_1c_2 & c_0s_1s_2c_3 & \cdots & \cdots & c_0s_1\cdots s_{n-1}c_n \\ -s_1 & c_1c_2 & c_1s_2c_3 & \cdots & \cdots & c_1s_2\cdots s_{n-1}c_n \\ & -s_2 & c_2c_3 & \cdots & \cdots & c_2s_3\cdots s_{n-1}c_n \\ & & \ddots & \ddots & \vdots & \vdots \\ & & & -s_{n-2} & c_{n-2}c_{n-1} & c_{n-2}s_{n-1}c_n \\ & & & & -s_{n-1} & c_{n-1}c_n \end{pmatrix}.$$

It is evident that the maximum off-diagonal rank of Q is at most one. Hence an SSS representation for Q will come in handy when we need to conduct SSS matrix-matrix additions or multiplications. With the partitioning sequence $\{m_i = 1\}_{i=1}^n$, the SSS generators of Q are given by Table 2.

$\mathcal{D}_i(Q)$	$\mathcal{U}_i(Q)$	$\mathcal{V}_i(Q)$	$\mathcal{W}_i(Q)$	$\mathcal{P}_i(Q)$	$\mathcal{Q}_i(Q)$	$\mathcal{R}_i(Q)$
$c_{i-1}c_i$	$c_{i-1}s_i$	c_i	s_i	1	$-s_i$	0

TABLE 2. *SSS generators of Q.*

4.2. Compact representations of R

The off-diagonal low-rank structure of R in Theorem 3.4 admits a compact SSS representation. Using the partitioning sequence $\{m_i = 1\}_{i=1}^n$ and taking into account that R is upper triangular, we have

$$R = (R_{ij})_{N\times N}, \quad \text{where } R_{ij} = \begin{cases} d_i, & \text{if } i = j, \\ u_i w_{i+1} \cdots w_{j-1} v_j^T, & \text{if } i < j, \\ 0, & \text{if } i > j. \end{cases} \tag{7}$$

Again, the empty products above are treated as identity matrices. The dimensions of the (nonzero) SSS generators of R are specified in Table 3.

Generator matrix	$\mathcal{D}_i(R)$	$\mathcal{U}_i(R)$	$\mathcal{V}_i(R)$	$\mathcal{W}_i(R)$
	d_i	u_i	v_i	w_i
Size	1×1	$1 \times p$	$1 \times p$	$p \times p$

TABLE 3. *Dimensions of the SSS generators of* R.

According to Theorem 3.4, a compact SSS representation of R will have p not exceeding 2. During our new QR algorithm, however, we will allow not-so-compact (redundant) intermediate SSS generators of R but will compress them back to compact representations at the end of each QR iteration step.

Remark 4.1. As the SSS generators can be simply represented by a small number of vectors or parameters, later in most places of this paper for convenience we directly provide those vectors or parameters instead of writing the SSS forms.

5. A new QR algorithm for C

Consider the $n \times n$ companion matrix (1). Let

$$Z = \begin{pmatrix} 0 & \cdots & 0 & \pm 1 \\ 1 & \cdots & 0 & 0 \\ \vdots & \ddots & \vdots & \vdots \\ 0 & \cdots & 1 & 0 \end{pmatrix}, \quad e_1 = \begin{pmatrix} 1 \\ 0 \\ \vdots \\ 0 \end{pmatrix}, \quad \text{and} \quad y = \begin{pmatrix} a_1 \\ a_2 \\ \vdots \\ a_n \mp 1 \end{pmatrix},$$

and choose the sign of the $(1, n)$-entry of Z so that $\det(Z) = 1$. Clearly,

$$C = Z + e_1 y^T.$$

Instead of updating the Hessenberg iterates H in the standard QR algorithm, our new algorithm will carry out the implicit shift QR iterations based on the compact representations of Q and R mentioned in the previous section. The structured representations of Q and R will lead to a more delicate deflation scheme and a more convenient bulge chasing procedure, which are to be discussed in detail in the following subsections.

5.1. Swapping real Givens matrices

Before presenting the detailed QR iterations we first consider an important technique which swaps two or three Givens matrices and will be used in the structured bulge chasing. The notion of "swap" will become evident in a moment. Similar techniques can also be found in other places (e.g., [28]).

First consider the product $Q_i \cdot Q_j$, $1 \le i, j < n$, where Q_i and Q_j are two real Givens matrices as specified in (6).

- If $i = j$, then multiplying the product out we get $\widehat{Q}_i \equiv Q_i \cdot Q_j$, which is another Givens matrix, and

$$\begin{pmatrix} \widehat{c}_i & \widehat{s}_i \\ -\widehat{s}_i & \widehat{c}_i \end{pmatrix}, \ \widehat{c}_i = c_i c_j - s_i s_j, \ \widehat{s}_i = c_i s_j + s_i c_j. \tag{8}$$

- If $|i - j| \geq 2$, then

$$Q_i \cdot Q_j = Q_j \cdot Q_i, \tag{9}$$

which is literally swapping the two Givens matrices.

Next consider the product of the form: $Q_i Q_{i+1} G_i$, where Q_i and G_i are two i-th Givens matrices and Q_{i+1} is the $(i+1)$-st Givens matrix, with $1 \leq i \leq n-2$. Without loss of generality, we use $Q_1 Q_2 G_1$ as an example. Given the three Givens matrices in $\mathbb{R}^{3\times 3}$

$$Q_1 = \begin{pmatrix} c_1 & s_1 & \\ -s_1 & c_1 & \\ & & 1 \end{pmatrix}, \quad Q_2 = \begin{pmatrix} 1 & & \\ & c_2 & s_2 \\ & -s_2 & c_2 \end{pmatrix}, \quad G_1 = \begin{pmatrix} \alpha_1 & \beta_1 & \\ -\beta_1 & \alpha_1 & \\ & & 1 \end{pmatrix}, \tag{10}$$

we want to find another three Givens matrices in $\mathbb{R}^{3\times 3}$

$$\widehat{G}_2 = \begin{pmatrix} 1 & & \\ & \widehat{\alpha}_2 & \widehat{\beta}_2 \\ & -\widehat{\beta}_2 & \widehat{\alpha}_2 \end{pmatrix}, \quad \widehat{Q}_1 = \begin{pmatrix} \widehat{c}_1 & \widehat{s}_1 & \\ -\widehat{s}_1 & \widehat{c}_1 & \\ & & 1 \end{pmatrix}, \quad \widehat{Q}_2 = \begin{pmatrix} 1 & & \\ & \widehat{c}_2 & \widehat{s}_2 \\ & -\widehat{s}_2 & \widehat{c}_2 \end{pmatrix}, \tag{11}$$

so that

$$Q_1 Q_2 G_1 = \widehat{G}_2 \widehat{Q}_1 \widehat{Q}_2. \tag{12}$$

We present Algorithm 1 (next page) for the computation of $\widehat{G}_2$, $\widehat{Q}_1$ and $\widehat{Q}_2$.

Note that both approaches in the third step of Algorithm 1 for computing $\widehat{Q}_1$ and $\widehat{Q}_2$ (in exact arithmetic) yield $Q_1 Q_2 G_1 = \widehat{G}_2 \widehat{Q}_1 \widehat{Q}_2$. In a similar fashion, given three Givens matrices G_2, Q_1 and $Q_2 \in \mathbb{R}^{3\times 3}$, we can compute another three Givens matrices $\widehat{Q}_1$, $\widehat{Q}_2$ and $\widehat{G}_1 \in \mathbb{R}^{3\times 3}$ so that

$$G_2 Q_1 Q_2 = \widehat{Q}_1 \widehat{Q}_2 \widehat{G}_1, \tag{13}$$

where G_2 has a similar form as $\widehat{G}_2$ in (11) but without the hats in the notations, and the same situation holds for $\widehat{G}_1$ and G_1.

For the convenience of future reference, we call (12) a *backward Givens swap*, and (13) a *forward Givens swap*, according to the direction of G_1 (or G_2) being pushed. It is not hard to prove the backward stability of such swapping formulas.

Lastly, consider a special case of a backward Givens swap: $Q_{n-1} Q_n G_{n-1}$ with $Q_n = \mathrm{diag}\,[I_{n-1}, -1]$. We want to find another Givens matrix $\widehat{Q}_{n-1}$ so that

$$Q_{n-1} Q_n G_{n-1} = \widehat{Q}_{n-1} Q_n. \tag{14}$$

This boils down to inspect the products of their trailing 2×2 blocks:

$$\begin{pmatrix} c_{n-1} & s_{n-1} \\ -s_{n-1} & c_{n-1} \end{pmatrix} \begin{pmatrix} 1 & \\ & -1 \end{pmatrix} \begin{pmatrix} \alpha_{n-1} & \beta_{n-1} \\ -\beta_{n-1} & \alpha_{n-1} \end{pmatrix} = \begin{pmatrix} \widehat{c}_{n-1} & \widehat{s}_{n-1} \\ -\widehat{s}_{n-1} & \widehat{c}_{n-1} \end{pmatrix} \begin{pmatrix} 1 & \\ & -1 \end{pmatrix}$$

Algorithm 1 Givens swap of type I

(1) Compute

$$A := Q_1 Q_2 G_1 = \begin{pmatrix} c_1\alpha_1 - s_1c_2\beta_1 & c_1\beta_1 + s_1c_2\alpha_1 & s_1s_2 \\ -s_1\alpha_1 - c_1c_2\beta_1 & -s_1\beta_1 + c_1c_2\alpha_1 & c_1s_2 \\ -s_2\beta_1 & -s_2\alpha_1 & c_2 \end{pmatrix} = \begin{pmatrix} \times & \times & \times \\ \times & \times & \times \\ \times & \times & \times \end{pmatrix}.$$

(2) Compute a Givens matrix $\widehat{G}_2$ so that

$$A_1 := \widehat{G}_2^T A = \begin{pmatrix} \times & \times & \times \\ \times & \times & \times \\ & \times & \times \end{pmatrix}.$$

(3) *We have two different approaches to get* $\widehat{Q}_1$ *and* $\widehat{Q}_2$.

- *Either*, let

$$\begin{cases} \widehat{c}_1 = & A_1(1,1), \\ \widehat{s}_1 = & -A_1(2,1), \end{cases} \quad \text{and} \quad \begin{cases} \widehat{c}_2 = & A_1(3,3), \\ \widehat{s}_2 = & -A_1(3,2), \end{cases}$$

since if there holds $A_1 = \widehat{Q}_1\widehat{Q}_2$, A_1 must also have the form

$$A_1 = \begin{pmatrix} \widehat{c}_1 & \widehat{s}_1\widehat{c}_2 & \widehat{s}_1\widehat{s}_2 \\ -\widehat{s}_1 & \widehat{c}_1\widehat{c}_2 & \widehat{c}_1\widehat{s}_2 \\ & -\widehat{s}_2 & \widehat{c}_2 \end{pmatrix}.$$

- *Or*, continue to find $\widehat{Q}_1$ so that

$$A_2 := \widehat{Q}_1^T A_1 = \begin{pmatrix} \times & \times & \times \\ & \times & \times \\ & \times & \times \end{pmatrix};$$

and then find $\widehat{Q}_2$ so that

$$A_3 := \widehat{Q}_2^T A_2 = \begin{pmatrix} \times & \times & \times \\ & \times & \times \\ & & \times \end{pmatrix}.$$

Since A_3 is triangular and orthogonal, it must be an identity matrix.

which leads to

$$\begin{cases} \widehat{c}_{n-1} & = & c_{n-1}\alpha_{n-1} + s_{n-1}\beta_{n-1}, \\ \widehat{s}_{n-1} & = & -c_{n-1}\beta_{n-1} + s_{n-1}\alpha_{n-1}. \end{cases} \tag{15}$$

5.2. Initial QR factorization of C

We start the new QR algorithm by first finding the initial QR factorization of $C \equiv C^{(0)}$. This can be easily done by applying a sequence of (transposes of) Givens rotations $\{Q_i^T\}_{i=1}^{n-1}$ to C from the left side to zero out its subdiagonal entries (which are 1's) from top to bottom. The process can be expressed as

$$Q_{n-1}^T(Q_{n-2}^T(\cdots(Q_2^T(Q_1^T C))\cdots)) \Longrightarrow R^{(0)}, \tag{16}$$

where R is an upper triangular matrix and Q_k is the k-th Givens rotation matrix of the form (6).

Thus from equation (16), we can write

$$C = Q_1 Q_2 \cdots Q_{n-1} \cdot R^{(0)}.$$

Let $Q^{(0)} \equiv Q_1 Q_2 \ldots Q_{n-1}$. Then $Q^{(0)}$ is completely represented in terms of its cosine and sine parameters: $\{c_i,\ s_i\}_{i=1}^n$ (with the assumptions $c_n = 1$ and $s_n = 0$). As for $R^{(0)}$, it is straightforward to check that in terms of $\{c_i,\ s_i\}_{i=1}^n$ and $\{a_i\}_{i=1}^n$ we have:

$$R^{(0)} = (R_{ij}^{(0)}), \quad \text{where } R_{ij}^{(0)} = \begin{cases} c_i s_{i-1} \cdots s_1 a_i - s_i & \text{if } i = j, \\ c_i s_{i-1} \cdots s_1 a_j & \text{if } i < j, \\ 0 & \text{if } i > j. \end{cases}$$

Or equivalently, we can use the following SSS generators to completely describe $R^{(0)}$:

$$\begin{cases} \mathcal{D}(R^{(0)}) & \equiv d_i = c_i s_{i-1} \cdots s_1 a_i - s_i, & \text{if } 1 \le i \le n, \\ \mathcal{U}(R^{(0)}) & \equiv u_i = c_i s_{i-1} \cdots s_1, & \text{if } 1 \le i \le n-1, \\ \mathcal{V}(R^{(0)}) & \equiv v_i = a_i, & \text{if } 2 \le i \le n, \\ \mathcal{W}(R^{(0)}) & \equiv w_i = 1, & \text{if } 2 \le i \le n-1. \end{cases}$$

Note that for now, p, the common column dimension of SSS generators, is 1.

5.3. Structured QR iteration: single shift case

In this section, by using a concrete 5×5 example, we describe in detail how to implement the following implicit single shift QR iteration on an H as in Theorem 3.4, where $\sigma \in \mathbb{R}$ is a shift.

$$\begin{aligned} H - \sigma I &= QR, \\ \widehat{H} &= RQ + \sigma I = Q^T H Q. \end{aligned}$$

Contrasting with the standard QR algorithm, where we chase a bulge along the second subdiagonal of the Hessenberg iterate H, in our new QR algorithm, we create and chase a bulge along the subdiagonal of R.

Before we start, we make two notations clear:

$\widehat{G}_k$	:	the Givens rotation used to generate a bulge at $R(k+1,k)$,
$\widetilde{G}_k$	:	the Givens rotation used to eliminate the bulge at $R(k+1,k)$,

where $R(i,j)$ denotes the (i,j) entry of R.

Suppose that at the beginning of the QR iteration, we have

$$H = Q_1 Q_2 Q_3 Q_4 \cdot R = Z + xy^T,$$

where Z is orthogonal but not explicitly stored.

(1) **Initiate bulge chasing.** Let $H_0 = H$. Choose a Givens rotation $\bar{G}_1$ of the form

$$\bar{G}_1 = \operatorname{diag}\left(\begin{pmatrix} \bar{c}_1 & \bar{s}_1 \\ -\bar{s}_1 & \bar{c}_1 \end{pmatrix}, I_3\right), \quad \text{where } \bar{c}_1^2 + \bar{s}_1^2 = 1,$$

so that the first column of $\bar{G}_1$, that is, the vector $\begin{pmatrix} \bar{c}_1 & -\bar{s}_1 & 0 & 0 & 0 \end{pmatrix}^T$, is proportional to, $\begin{pmatrix} h_{11} - \sigma & h_{21} & 0 & 0 & 0 \end{pmatrix}^T$, the first column of $H_0 - \sigma I$. Let

$$H_1 \equiv \bar{G}_1^T H_0 \bar{G}_1 = (\bar{G}_1^T Q_1) Q_2 Q_3 Q_4 \cdot R\bar{G}_1.$$

Then a bulge is created at the $(2,1)$ entry of $R\bar{G}_1$. In fact, if we formed $R\bar{G}_1$ explicitly, we should expect

$$R\bar{G}_1 = \begin{pmatrix} \times & \times & \times & \times & \times \\ + & \times & \times & \times & \times \\ & & \times & \times & \times \\ & & & \times & \times \\ & & & & \times \end{pmatrix},$$

where the bulge is indicted by a plus sign. Next choose

$$\widetilde{G}_1 = \operatorname{diag}\left(\begin{pmatrix} \widetilde{c}_1 & \widetilde{s}_1 \\ -\widetilde{s}_1 & \widetilde{c}_1 \end{pmatrix}, I_3\right), \quad \text{where } \widetilde{c}_1^2 + \widetilde{s}_1^2 = 1$$

so that $R_1 \equiv \widetilde{G}_1^T (R\bar{G}_1)$ is upper triangular again. Let $\bar{Q}_1 = \bar{G}_1^T Q_1$, then

$$\begin{aligned} H_1 &= (\bar{G}_1^T Q_1) Q_2 Q_3 Q_4 \widetilde{G}_1 \cdot \widetilde{G}_1^T R_0 \bar{G}_1 \\ &= \bar{Q}_1 Q_2 Q_3 Q_4 \widetilde{G}_1 \cdot R_1 \\ &= (\bar{Q}_1 Q_2 \widetilde{G}_1) Q_3 Q_4 \cdot R_1 \qquad (\widetilde{G}_1 \text{ pushed forward}) \\ &= (\bar{G}_2 \widehat{Q}_1 \bar{Q}_2) Q_3 Q_4 \cdot R_1 \qquad (\text{backward Givens swap}) \\ &= \bar{G}_2 \cdot \widehat{Q}_1 \bar{Q}_2 Q_3 Q_4 \cdot R_1. \end{aligned}$$

(2) **Second chasing.** Let

$$H_2 \equiv \bar{G}_2^T H_1 \bar{G}_2 = \widehat{Q}_1 Q_2 Q_3 Q_4 \cdot R_1 \bar{G}_2,$$

where if explicitly formed,

$$R_1 \bar{G}_2 = \begin{pmatrix} \times & \times & \times & \times & \times \\ & \times & \times & \times & \times \\ & + & \times & \times & \times \\ & & & \times & \times \\ & & & & \times \end{pmatrix}.$$

Thus the bulge has been "chased" from the $(2,1)$ position to the $(3,2)$ position. To eliminate this bulge, we choose a Givens rotation $\widetilde{G}_2$ so that

$R_2 \equiv \widetilde{G}_2^T(R_1\bar{G}_2)$ becomes upper triangular again. Thus

$$\begin{aligned} H_2 &= \widehat{Q}_1\bar{Q}_2Q_3Q_4\widetilde{G}_2 \cdot \widetilde{G}_2^T R_1\bar{G}_2 \\ &= \widehat{Q}_1(\bar{Q}_2Q_3\widetilde{G}_2)Q_4 \cdot R_2 \qquad (\widetilde{G}_2 \text{ pushed forward}) \\ &= \widehat{Q}_1(\bar{G}_3\widehat{Q}_2\bar{Q}_3)Q_4 \cdot R_2 \qquad (\text{backward Givens swap}) \\ &= \bar{G}_3 \cdot \widehat{Q}_1\widehat{Q}_2\bar{Q}_3Q_4 \cdot R_2. \qquad (\bar{G}_3 \text{ pushed forward}). \end{aligned}$$

(3) **Third chasing.** Similarly, let

$$H_3 \equiv \bar{G}_3^T H_2\bar{G}_3 = \widehat{Q}_1\widehat{Q}_2\bar{Q}_3Q_4 \cdot R_2\bar{G}_3,$$

where if explicitly formed,

$$R_2\bar{G}_3 = \begin{pmatrix} \times & \times & \times & \times & \times \\ & \times & \times & \times & \times \\ & & \times & \times & \times \\ & & + & \times & \times \\ & & & & \times \end{pmatrix}.$$

Thus the bulge has been chased from the $(3,2)$ position to the $(4,3)$ position. To eliminate this bulge, we choose a Givens rotation $\widetilde{G}_3$ so that $R_3 \equiv \widetilde{G}_3^T(R_2\bar{G}_2)$ becomes upper triangular again. Thus

$$\begin{aligned} H_3 &= \widehat{Q}_1\widehat{Q}_2(\bar{Q}_3Q_4\widetilde{G}_3) \cdot \widetilde{G}_3^T R_2\bar{G}_3 \\ &= \widehat{Q}_1\widehat{Q}_2(\bar{G}_4\widehat{Q}_3\bar{Q}_4) \cdot R_3 \qquad (\text{backward Givens swap}) \\ &= \bar{G}_4 \cdot \widehat{Q}_1\widehat{Q}_2\widehat{Q}_3\bar{Q}_4 \cdot R_3. \qquad (\widetilde{G}_4 \text{ pushed forward}). \end{aligned}$$

(4) **Final chasing.** Let

$$H_4 \equiv \bar{G}_4^T H_3\bar{G}_4 = \widehat{Q}_1\widehat{Q}_2\widehat{Q}_3\bar{Q}_4 \cdot R_3\bar{G}_4,$$

where if explicitly formed,

$$R_3\bar{G}_4 = \begin{pmatrix} \times & \times & \times & \times & \times \\ & \times & \times & \times & \times \\ & & \times & \times & \times \\ & & & \times & \times \\ & & & + & \times \end{pmatrix}.$$

Thus the bulge has been chased from $(4,3)$ to $(5,4)$. This leads us to choose a Givens rotation $\widetilde{G}_4$ such that $R_4 \equiv \widetilde{G}_4^T(R_3\bar{G}_4)$ becomes upper triangular again. Let $\widehat{Q}_4 \equiv \bar{Q}_4\widetilde{G}_4$, then

$$\begin{aligned} H_4 &= \widehat{Q}_1\widehat{Q}_2\widehat{Q}_3(\bar{Q}_4\widetilde{G}_4) \cdot (\widetilde{G}_4^T R_3\bar{G}_4). \\ &= \widehat{Q}_1\widehat{Q}_2\widehat{Q}_3\widehat{Q}_4 \cdot R_4. \end{aligned}$$

Let $\widehat{H} = H_4$. A cycle of QR iteration with single shift is then completed.

Write $\bar{G} \equiv \bar{G}_1\bar{G}_2\bar{G}_3\bar{G}_4$, $\widetilde{G} \equiv \widetilde{G}_1\widetilde{G}_2\widetilde{G}_3\widetilde{G}_4$ and $\widehat{Q} = \widehat{Q}_1\widehat{Q}_2\widehat{Q}_3\widehat{Q}_4$. Then the structured single shift bulge chasing procedure presented above tells us

$$\begin{aligned} H_4 &= \bar{G}_4^T\bar{G}_3^T\bar{G}_2^T\bar{G}_1^T \cdot H_0 \cdot \bar{G}_1\bar{G}_2\bar{G}_3\bar{G}_4 = \bar{G}^T \cdot H_0 \cdot \bar{G}, \\ R_4 &= \widetilde{G}_4^T\widetilde{G}_3^T\widetilde{G}_2^T\widetilde{G}_1^T \cdot R_0 \cdot \bar{G}_1\bar{G}_2\bar{G}_3\bar{G}_4 = \widetilde{G}^T \cdot R_0 \cdot \bar{G}, \\ H_4 &= \widehat{Q} \cdot R_4. \end{aligned} \tag{17}$$

Remark 5.1. Since the first column of $\bar{G}$ is proportional to that of $H_0 - \sigma I$, according to the well known implicit Q theorem, $\bar{G}$ will be the same (up to sign differences in each column) as the Q-factor of the QR decomposition of $H_0 - \sigma I$.

Next we discuss the computation and elimination of the bulges in terms of the structured representations. Note that none of the R_k's are formed explicitly except certain entries. The explanation is as follows. R_k is represented via its SSS generators, $\{d_i, u_i, v_i, w_i\}$. Not all these generators are updated during the intermediate steps of a bulge chasing cycle. We need to form explicitly the main diagonal vector (d_i generators) and the first superdiagonal vector of R_k in order to compute the bulges. To simplify the notations we temporarily write R_k as R, in a general SSS form:

$$R = \begin{pmatrix} \ddots & & \cdots & & \vdots \\ & d_i & u_iv_{i+1}^T & u_iw_{i+1}v_{i+2}^T & \cdots & u_iw_{i+1}\cdots w_{n-1}v_n^T \\ & & d_{i+1} & u_{i+1}v_{i+2}^T & \cdots & u_{i+1}w_{i+2}\cdots w_{n-1}v_n^T \\ & & & d_{i+2} & \cdots & u_{i+2}w_{i+3}\cdots w_{n-1}v_n^T \\ & & & & \ddots & \vdots \end{pmatrix},$$

where the i-th through $(i+2)$-nd rows are shown. Let h be the first superdiagonal vector. That is, $h_i \equiv R_{i,i+1} = u_iv_{i+1}^T$. During the bulge chasing, a bulge b_i is created by right multiplying a Givens matrix $\bar{G}_j = \begin{pmatrix} c_i & -s_i \\ s_i & c_i \end{pmatrix}$ to a 2-by-2 upper triangular diagonal block:

$$\begin{pmatrix} \widehat{d}_i & \widehat{h}_i \\ b_i & \widehat{d}_{i+1} \end{pmatrix} = \begin{pmatrix} d_i & h_i \\ 0 & d_{i+1} \end{pmatrix} \begin{pmatrix} c_i & -s_i \\ s_i & c_i \end{pmatrix}. \tag{18}$$

A new Givens matrix $\widetilde{G}_j = \begin{pmatrix} \widetilde{c}_i & -\widetilde{s}_i \\ \widetilde{s}_i & \widetilde{c}_i \end{pmatrix}$ is now computed based on $\begin{pmatrix} \widehat{d}_i \\ b_i \end{pmatrix}$ so as to eliminate the bulge b_i:

$$\begin{pmatrix} \widetilde{d}_i & \widetilde{h}_i \\ 0 & \widetilde{d}_{i+1} \end{pmatrix} = \begin{pmatrix} \widetilde{c}_i & -\widetilde{s}_i \\ \widetilde{s}_i & \widetilde{c}_i \end{pmatrix} \begin{pmatrix} \widehat{d}_i & \widehat{h}_i \\ b_i & \widehat{d}_{i+1} \end{pmatrix}. \tag{19}$$

Then the i-th and $(i+1)$-st rows of R should be updated, which is done as follows:

$$\begin{aligned}
&\begin{pmatrix} \widetilde{c}_i & -\widetilde{s}_i \\ \widetilde{s}_i & \widetilde{c}_i \end{pmatrix} \left(\begin{array}{cc|ccc} \widehat{d}_i & \widehat{h}_i & u_i w_{i+1} v_{i+2}^T & \cdots & u_i w_{i+1} \cdots w_{n-1} v_n^T \\ b_i & \widehat{d}_{i+1} & u_{i+1} v_{i+2}^T & \cdots & u_{i+1} w_{i+2} \cdots w_{n-1} v_n^T \end{array}\right) \\
&= \left(\begin{pmatrix} \widetilde{d}_i & \widetilde{h}_i \\ 0 & \widetilde{d}_{i+1} \end{pmatrix} \begin{pmatrix} \widetilde{c}_i & -\widetilde{s}_i \\ \widetilde{s}_i & \widetilde{c}_i \end{pmatrix} \begin{pmatrix} u_i w_{i+1} \\ u_{i+1} \end{pmatrix} \begin{pmatrix} v_{i+2}^T & w_{i+2} v_{i+2}^T & \cdots & w_{i+2} \cdots w_{n-1} v_n^T \end{pmatrix} \right) \\
&= \left(\begin{pmatrix} \widetilde{d}_i & \widetilde{h}_i \\ 0 & \widetilde{d}_{i+1} \end{pmatrix} \begin{pmatrix} \widehat{u}_i \\ \widehat{u}_{i+1} \end{pmatrix} \begin{pmatrix} v_{i+2}^T & w_{i+2} v_{i+2}^T & \cdots & w_{i+2} \cdots w_{n-1} v_n^T \end{pmatrix} \right) \\
&= \begin{pmatrix} \widetilde{d}_i & \widetilde{h}_i & \widehat{u}_i v_{i+2}^T & \cdots & \widehat{u}_i w_{i+2} \cdots w_{n-1} v_n^T \\ 0 & \widetilde{d}_{i+1} & \widehat{u}_{i+1} v_{i+2}^T & \cdots & \widehat{u}_{i+1} w_{i+2} \cdots w_{n-1} v_n^T \end{pmatrix}. \qquad (20)
\end{aligned}$$

That is, we only need to find the updated $\widetilde{d}_i, \widetilde{d}_{i+1}, \widetilde{h}_i$, $\widehat{u}_i$, and $\widehat{u}_{i+1}$. After this step, the new superdiagonal entry $h_{i+1} = \widehat{u}_{i+1} v_{i+1}^T$ is formed. The next bulge will be generated with another Givens matrix applied on the right to the next 2-by-2 diagonal block

$$\begin{pmatrix} \widetilde{d}_{i+1} & h_{i+1} \\ 0 & d_{i+2} \end{pmatrix},$$

and the above process repeats. Therefore, during the bulge chasing cycle, $\{d_i, u_i\}$ are updated, and $\{h_i\}$ are formed. Clearly, we use each h_i once a time and do not need to store the entire h.

Equation (20) is sufficient for deriving h_{i+1} and thus further computing and eliminating the bulges. However, the $\widehat{u}_i$ it provides may not be an SSS generator of the final R. As an example, the updated value of $R_{i,i+1}$ is $\widetilde{h}_i$, which is generally not $\widehat{u}_i v_{i+1}^T$. Therefore, to get a final updated SSS form for R, we update all $\{u_i, v_i, w_i\}$ at the end of the bulge chasing cycle. For example, in the process (17) above, the SSS generators of R_4 are obtained by multiplying three SSS matrices $\widetilde{G}^T$, R_0, and $\bar{G}$ using the fast SSS matrix-matrix multiplication formulas in Subsection 2.3.

Remark 5.2. An outcome of using those multiplication formulas is that the column dimensions of R_4's SSS generators will grow additively by 2 (in case of single shift bulge chasing), since both $\bar{G}$ and $\widetilde{G}$ have the maximum off-diagonal rank 1. In Subsection 5.5 we will show how to recover a compact representation for R_4.

5.4. Structured QR iteration: double shift case

This section describes how to maintain real arithmetic by employing two shifts σ and $\bar{\sigma}$ at the same time, where $\bar{\sigma}$ is the complex conjugate of σ (although in this paper notations with bars do not necessarily mean complex conjugates). The

process of shifting σ and $\bar{\sigma}$ successively is like

$$\begin{aligned} H - \sigma I &= Q^{(1)} R^{(1)}, \\ H^{(1)} &= R^{(1)} Q^{(1)} + \sigma I = \left(Q^{(1)}\right)^T H \left(Q^{(1)}\right), \\ H^{(1)} - \bar{\sigma} I &= Q^{(2)} R^{(2)}, \\ \widehat{H} = H^{(2)} &= R^{(2)} Q^{(2)} + \bar{\sigma} I = \left(Q^{(1)} Q^{(2)}\right)^T H \left(Q^{(1)} Q^{(2)}\right), \end{aligned}$$

which leads to

$$M \equiv \left(Q^{(1)} Q^{(2)}\right) \left(R^{(2)} R^{(1)}\right) = (H - \sigma I)(H - \bar{\sigma} I) = H^2 - sH + tI, \tag{21}$$

with $s = 2\,\mathrm{Re}(\sigma)$, $t = |\sigma|^2$. Thus $\left(Q^{(1)} Q^{(2)}\right) \left(R^{(2)} R^{(1)}\right)$ is the QR decomposition of the real matrix M, and therefore $Q^{(1)} Q^{(2)}$, as well as $R^{(2)} R^{(1)}$, can be chosen real, which means that $\widehat{H} = \left(Q^{(1)} Q^{(2)}\right)^T H \left(Q^{(1)} Q^{(2)}\right)$ is also real.

While the rationale for maintaining real arithmetic is exactly the same, the difference of our new algorithm from the standard one lies in the use of the compact representations for Q and R. Contrasting with the standard implicit double shift QR algorithm where a 2-by-2 bulge is chased along the subdiagonal of the Hessenberg iterate H, in our new algorithm the 2-by-2 bulge is chased along the subdiagonal of R. Before we start, we make the following notations clear

F_{k+1}	:	the 1st Givens used to generate a nonzero at $R(k+2, k)$,
$\bar{G}_k$	:	the 2nd Givens used to generate nonzeros at $R(k+1 : k+2, k)$,
$\widetilde{F}_{k+1}$	:	the 1st Givens used to eliminate the nonzero at $R(k+2, k)$,
$\widetilde{G}_k$	:	the 2nd Givens used to eliminate the nonzero at $R(k+1, k)$.

Let us use the same 5-by-5 example from the last subsection. Suppose that at the beginning of the QR iteration, we have

$$H_0 \equiv H = Q_1 Q_2 Q_3 Q_4 \cdot R = Z + xy^T,$$

where Z is orthogonal but not explicitly stored.

(1) **Initiate bulge chasing.** Given a pair of complex conjugate shifts σ and $\bar{\sigma}$, we compute the first column of M in (21):

$$Me_1 = (H^2 - sH + tI)e_1 = \begin{pmatrix} x_1 & x_2 & x_3 & 0 & \cdots & 0 \end{pmatrix}^T,$$

where

$$\begin{cases} x_1 &= h_{11}^2 + h_{12} h_{21} - s h_{11} + t, \\ x_2 &= h_{21}(h_{11} + h_{22} - s), \\ x_3 &= h_{21} h_{32}. \end{cases} \tag{22}$$

Then find two Givens rotations $\bar{G}_1$ and $\bar{F}_2$ such that

$$\left(\bar{G}_1\right)^T \left(\bar{F}_2\right)^T \begin{pmatrix} x_1 \\ x_2 \\ x_3 \end{pmatrix} = \begin{pmatrix} \times \\ 0 \\ 0 \end{pmatrix}.$$

In other words, the first column of $\left(\bar{F}_2\bar{G}_1\right)$ should be made proportional to Me_1. Let

$$\begin{aligned}
H_1 &\equiv \left(\bar{F}_2\bar{G}_1\right)^T \cdot H_0 \cdot \left(\bar{F}_2\bar{G}_1\right) \\
&= \left(\bar{G}_1\right)^T \cdot \left((\bar{F}_2)^T Q_1 Q_2\right) Q_3 Q_4 \cdot R_0 \bar{F}_2 \bar{G}_1 \\
&= \left(\bar{G}_1\right)^T \cdot \left(\bar{Q}_1 \bar{Q}_2 \widetilde{F}_1\right) Q_3 Q_4 \cdot R_0 \bar{F}_2 \bar{G}_1 \qquad \text{(forward Givens swap)} \\
&= \left(\left(\bar{G}_1\right)^T \bar{Q}_1\right) \bar{Q}_2 Q_3 Q_4 \cdot \left(\widetilde{F}_1 R_0\right) \bar{F}_2 \bar{G}_1 \\
&= \widehat{Q}_1 \bar{Q}_2 Q_3 Q_4 \cdot \widetilde{R}_0 \bar{F}_2 \bar{G}_1,
\end{aligned}$$

where $\widehat{Q}_1 \equiv (\bar{G}_1)^T \bar{Q}_1$, $\widetilde{R}_0 \equiv \widetilde{F}_1 R_0$, and if formed explicitly,

$$\widetilde{R}_0 \bar{F}_2 \bar{G}_1 = \begin{pmatrix} \times & \times & \times & \times & \times \\ + & \times & \times & \times & \times \\ + & + & \times & \times & \times \\ & & & \times & \times \\ & & & & \times \end{pmatrix}.$$

We see that there is a 2-by-2 bulge, indicted by plus signs. Next choose two Givens rotations $\widetilde{F}_2$ and $\widetilde{G}_1$ to zero out entries $(3,1)$ and $(2,1)$ of $\widetilde{R}_0 \bar{F}_2 \bar{G}_1$ in order. Let $\widetilde{R}_1 \equiv \left(\widetilde{G}_1\right)^T \left(\widetilde{F}_2\right)^T \cdot \left(\widetilde{R}_0 \bar{F}_2 \bar{G}_1\right)$, then we may write

$$\begin{aligned}
H_1 &= \widehat{Q}_1 \bar{Q}_2 Q_3 Q_4 \left(\widetilde{F}_2 \widetilde{G}_1\right) \cdot \widetilde{R}_1 \\
&= \widehat{Q}_1 \left(\bar{Q}_2 Q_3 \widetilde{F}_2\right) Q_4 \widetilde{G}_1 \cdot \widetilde{R}_1 \qquad (\widetilde{F}_2 \text{ pushed forward}) \\
&= \widehat{Q}_1 \left(\bar{F}_3 \widetilde{Q}_2 \bar{Q}_3\right) Q_4 \widetilde{G}_1 \cdot \widetilde{R}_1 \qquad \text{(backward Givens swap)} \\
&= \bar{F}_3 \left(\widehat{Q}_1 \widetilde{Q}_2 \widetilde{G}_1\right) \bar{Q}_3 Q_4 \cdot \widetilde{R}_1 \qquad (\bar{F}_3 \text{ and } \widetilde{G}_1 \text{ pushed forward.}) \\
&= \bar{F}_3 \left(\bar{G}_2 \widehat{\widetilde{Q}}_1 \widehat{Q}_2\right) \bar{Q}_3 Q_4 \cdot \widetilde{R}_1 \qquad \text{(backward Givens swap)} \\
&= \bar{F}_3 \bar{G}_2 \cdot \widehat{\widetilde{Q}}_1 \widehat{Q}_2 \bar{Q}_3 Q_4 \cdot \widetilde{R}_1.
\end{aligned}$$

(2) **Second chasing.** Let

$$H_2 \equiv \left(\bar{F}_3\bar{G}_2\right)^T \cdot H_1 \cdot \left(\bar{F}_3\bar{G}_2\right) = \widehat{\widetilde{Q}}_1 \widehat{Q}_2 \bar{Q}_3 Q_4 \cdot \left(\widetilde{R}_1 \bar{F}_3 \bar{G}_2\right),$$

where if explicitly formed,

$$\widetilde{R}_1 \bar{F}_3 \bar{G}_2 = \begin{pmatrix} \times & \times & \times & \times & \times \\ & \times & \times & \times & \times \\ & + & \times & \times & \times \\ & + & + & \times & \times \\ & & & & \times \end{pmatrix}.$$

Thus compared with $\widetilde{R}_0\bar{F}_2\bar{G}_1$, the 2-by-2 bulge has been chased to the right for one column. Next choose two Givens rotations $\widetilde{F}_3$ and $\widetilde{G}_2$ to zero out $(4,2)$ and $(3,2)$ entries in order. Let $\widetilde{R}_2 \equiv \left(\widetilde{G}_2\right)^T \left(\widetilde{F}_3\right)^T \cdot \left(\widetilde{R}_1\bar{F}_3\bar{G}_2\right)$, then we may write

$$\begin{aligned}
H_2 &= \widehat{\widehat{Q}}_1\widehat{Q}_2\bar{Q}_3Q_4\left(\widetilde{F}_3\widetilde{G}_2\right)\cdot\widetilde{R}_2 \\
&= \widehat{\widehat{Q}}_1\widehat{Q}_2\left(\bar{Q}_3Q_4\widetilde{F}_3\right)\widetilde{G}_2\cdot\widetilde{R}_2 \\
&= \widehat{\widehat{Q}}_1\widehat{Q}_2\left(\bar{F}_4\widetilde{Q}_3\bar{Q}_4\right)\widetilde{G}_2\cdot\widetilde{R}_2 \qquad \text{(backward Givens swap)} \\
&= \bar{F}_4\widehat{\widehat{Q}}_1\left(\widehat{Q}_2\widetilde{Q}_3\widetilde{G}_2\right)\bar{Q}_4\cdot\widetilde{R}_2 \qquad (\bar{F}_4 \text{ and } \widetilde{G}_2 \text{ pushed forward}) \\
&= \bar{F}_4\widehat{\widehat{Q}}_1\left(\bar{G}_3\widehat{\widehat{Q}}_2\widehat{Q}_3\right)\bar{Q}_4\cdot\widetilde{R}_2 \qquad \text{(backward Givens swap)} \\
&= \bar{F}_4\bar{G}_3\cdot\widehat{\widehat{Q}}_1\widehat{\widehat{Q}}_2\widehat{Q}_3\bar{Q}_4\cdot\widetilde{R}_2. \qquad (\bar{G}_3 \text{ pushed forward}).
\end{aligned}$$

(3) **Final two steps of bulge chasing.** Let

$$H_3 \equiv \left(\bar{F}_4\bar{G}_3\right)^T\cdot H_2\cdot\left(\bar{F}_4\bar{G}_3\right) = \widehat{\widehat{Q}}_1\widehat{\widehat{Q}}_2\widehat{Q}_3\bar{Q}_4\cdot\left(\widetilde{R}_2\bar{F}_4\bar{G}_3\right),$$

where if explicitly formed,

$$\widetilde{R}_2\bar{F}_4\bar{G}_3 = \begin{pmatrix} \times & \times & \times & \times & \times \\ & \times & \times & \times & \times \\ & & \times & \times & \times \\ & & + & \times & \times \\ & & + & + & \times \end{pmatrix}.$$

Thus compared with $\widetilde{R}_1\bar{F}_3\bar{G}_2$, the 2-by-2 bulge has been chased by one column to the lower right. Next choose two Givens rotations $\widetilde{F}_4$ and $\widetilde{G}_3$ to zero out the $(5,3)$ and $(4,3)$ entries in order. Let $\widetilde{R}_3 \equiv \left(\widetilde{G}_3\right)^T\left(\widetilde{F}_4\right)^T\cdot\left(\widetilde{R}_2\bar{F}_4\bar{G}_3\right)$, then we may write

$$\begin{aligned}
H_3 &= \widehat{\widehat{Q}}_1\widehat{\widehat{Q}}_2\widehat{Q}_3\bar{Q}_4\left(\widetilde{F}_4\widetilde{G}_3\right)\cdot\widetilde{R}_3 \\
&= \widehat{\widehat{Q}}_1\widehat{\widehat{Q}}_2\left(\widehat{Q}_3\widetilde{Q}_4\widetilde{G}_3\right)\cdot\widetilde{R}_3 \qquad (\widetilde{Q}_4 \equiv \bar{Q}_4\widetilde{F}_4) \\
&= \widehat{\widehat{Q}}_1\widehat{\widehat{Q}}_2\left(\bar{G}_4\widehat{\widehat{Q}}_3\widehat{Q}_4\right)\cdot\widetilde{R}_3 \qquad \text{(backward Givens swap)} \\
&= \bar{G}_4\cdot\widehat{\widehat{Q}}_1\widehat{\widehat{Q}}_2\widehat{\widehat{Q}}_3\widehat{Q}_4\cdot\widetilde{R}_3. \qquad (\bar{G}_4 \text{ pushed forward}).
\end{aligned}$$

Lastly, let

$$H_4 \equiv \left(\bar{G}_4\right)^T\cdot H_3\cdot\left(\bar{G}_4\right) = \widehat{\widehat{Q}}_1\widehat{\widehat{Q}}_2\widehat{\widehat{Q}}_3\widehat{Q}_4\cdot\left(\widetilde{R}_3\bar{G}_4\right),$$

where if explicitly formed,

$$\widetilde{R}_3\bar{G}_4 = \begin{pmatrix} \times & \times & \times & \times & \times \\ & \times & \times & \times & \times \\ & & \times & \times & \times \\ & & & \times & \times \\ & & & + & \times \end{pmatrix}.$$

Next choose a Givens rotation $\widetilde{G}_4$ to zero out the $(5,4)$ entry above to get an upper triangular matrix $R_4 \equiv \left(\widetilde{G}_4\right)^T \cdot \widetilde{R}_3\bar{G}_4$. Now we may write

$$\begin{aligned} H_4 &= \widehat{\widehat{Q}}_1\widehat{\widehat{Q}}_2\widehat{\widehat{Q}}_3\widehat{Q}_4\widetilde{G}_4 \cdot R_4 \\ &= \widehat{\widehat{Q}}_1\widehat{\widehat{Q}}_2\widehat{\widehat{Q}}_3\widehat{\widehat{Q}}_4 \cdot R_4. \qquad (\widehat{\widehat{Q}}_4 \equiv \widehat{Q}_4\widetilde{G}_4). \end{aligned}$$

Let $\widehat{H} = H_4$. A cycle of QR iteration with a pair of complex conjugate shifts $\{\sigma, \bar{\sigma}\}$ is then completed. Define $\widehat{Q} \equiv \widehat{\widehat{Q}}_1\widehat{\widehat{Q}}_2\widehat{\widehat{Q}}_3\widehat{\widehat{Q}}_4$, and

$$\begin{aligned} \bar{W} &\equiv \bar{F}_2\bar{G}_1\bar{F}_3\bar{G}_2\bar{F}_4\bar{G}_3\bar{G}_4 \\ &= \left(\bar{F}_2\bar{F}_3\bar{F}_4\right) \cdot \left(\bar{G}_1\bar{G}_2\bar{G}_3\bar{G}_4\right) \\ &\equiv \bar{F} \cdot \bar{G}, \\ \widetilde{W} &\equiv \widetilde{F}_1\widetilde{F}_2\widetilde{G}_1\widetilde{F}_3\widetilde{G}_2\widetilde{F}_4\widetilde{G}_3\widetilde{G}_4 \\ &= \left(\widetilde{F}_1\widetilde{F}_2\widetilde{F}_3\widetilde{F}_4\right) \cdot \left(\widetilde{G}_1\widetilde{G}_2\widetilde{G}_3\widetilde{G}_4\right) \\ &\equiv \widetilde{F} \cdot \widetilde{G}. \end{aligned}$$

We can then summarize the structured double shift bulge chasing procedure as:

$$\begin{aligned} H_4 &= \bar{W}^T \cdot H_0 \cdot \bar{W} = (\bar{F}\bar{G}) \cdot H_0 \cdot (\bar{F}\bar{G}), \\ R_4 &= \widetilde{W}^T \cdot R_0 \cdot \bar{W} = (\widetilde{F}\widetilde{G}) \cdot H_0 \cdot (\bar{F}\bar{G}), \\ H_4 &= \widehat{Q} \cdot R_4. \end{aligned}$$

Remark 5.3. Since the first column of $\bar{W}$ is proportional to that of $H^2 - sH + tI$ (with $s = 2\,\mathrm{Re}(\sigma)$, $t = |\sigma|^2$), according to the well known implicit Q theorem, $\bar{W}$ will be the same (up to sign differences in each column) as the Q-factor of the QR decomposition of $H^2 - sH + tI$.

Remark 5.4. Similar to the single shift case, none of the R_k's are formed explicitly, except few diagonal vectors which are needed for computing the bulges. The SSS generators $\{d_i, u_i\}$ of R_k are updated during the process. At the end of a bulge chasing cycle, $\{v_i, w_i\}$, are updated (also $\{u_i\}$, in fact), and this can be done efficiently by applying the fast SSS matrix-matrix multiplication formulas. However, an outcome of using those multiplication formulas is that the column dimensions of R's SSS generators will grow by 4 in case of double shift bulge chasing, since

both $\bar{W}$ and $\widetilde{W}$ have the maximum off-diagonal rank to be 2. In the next subsection, we will show how to recover a compact representation for R_4, or in general R_{n-1}.

5.5. Recovery of the compact SSS representation of R

In both single and double shift cases, we computed the SSS representation of R_{n-1} ($n = 5$ for the 5-by-5 example we considered) through the formula

$$R_{n-1} = \widetilde{G}^T \cdot R_0 \cdot \bar{G},$$

where for simplicity of notation, we have written in case of double shift iteration: $\widetilde{W} = \widetilde{F}\widetilde{G}$ as $\widetilde{G}$, $\bar{W} = \bar{F}\bar{G}$ as $\bar{G}$. As pointed out in Remarks 5.2 and 5.4, the column dimensions of the SSS generators of R_{n-1} increase by 2 and 4 in single and double shift cases, respectively. However, the mathematical ranks of the off-diagonal blocks of R_{n-1} do not increase starting from $n = 2$. The reason is that given $H_0 = Z + xy^T$, where Z is orthogonal but never explicitly stored, we can represent R_{n-1} as a rank-one modification to an orthogonal matrix:

$$R_{n-1} = \widehat{Q}^T H_{n-1} = \widehat{Q}^T \bar{G}^T H_0 \bar{G} = \left(\widehat{Q}^T \bar{G}^T Z \bar{G}\right) + \left(\widehat{Q}^T \bar{G}^T x\right) \cdot \left(\bar{G}^T y\right)^T.$$

According to Theorem 3.4, $\operatorname{rank}(R_{12}) \leq 2$ for any 2-by-2 blocking partitioning.

To recover a compact representation of R_{n-1}, we do the following.

(1) Compute $\widehat{x} = \widehat{Q}^T \bar{G}^T x$ and $\widehat{y} = \bar{G}^T y$. As just shown, the computed R_{n-1} in a redundant SSS form can be viewed as a rank-one perturbation to an orthogonal matrix, that is,

$$R_{n-1} - \widehat{x}\widehat{y}^T \text{ is an orthogonal matrix.}$$

(2) Find a sequence of Givens rotations $\{X_1, X_2, \ldots, X_{n-1}\}$, and let

$$X \equiv X_1 X_2 \cdots X_{n-1},$$

so that

$$X\widehat{x} = e_1.$$

Apply X to $R_{n-1} - \widehat{x}\widehat{y}^T$ from the left-hand side. Now $XR_{n-1} - e_1\widehat{y}^T$ remains orthogonal. On the other hand, since R_{n-1} is upper triangular and X is upper Hessenberg, $XR_{n-1} - e_1\widehat{y}^T$ is also upper Hessenberg.

(3) Thus we can find another sequence of Givens rotations $\{Y_{n-1}, Y_{n-2}, \ldots, Y_1\}$, let $Y \equiv Y_1 Y_2 \cdots Y_{n-1}$, so that

$$\left(XR_{n-1} - e_1\widehat{y}^T\right) Y^T = I.$$

(4) The last equation provides an alternative way to express R_{n-1}, that is,

$$R_{n-1} = X^T Y + X^T e_1 \widehat{y}^T = X^T Y + \widehat{x}\widehat{y}^T.$$

Both X are Y have orthogonal upper Hessenberg matrices with similar structure as that of Q, so that they can be written as SSS matrices with the maximum off-diagonal rank to be 1. The rank-one matrix $\widehat{x}\widehat{y}^T$ can also be written in SSS form with off-diagonal rank to be 1. By applying the fast SSS matrix-matrix multiplication in Subsection 2.3 to $X^T Y$ we obtain an SSS form for

X^TY with generator sizes bounded by 2 (the sizes increase additively). Then another fast SSS addition (Subsection 2.2) makes $R_{n-1} = (X^TY) + (\widehat{x}\widehat{y}^T)$ a new SSS matrix with generator sizes bounded by 3. That means, we get a new compact representation for R_{n-1}. Here although theoretically, according to Theorem 3.4 it is possible to further make the generator sizes no larger than 2, it does not make a significant difference in practice. We allow the sizes to be 3 for the sake of convenience in the programming. The above recovery process also applies to all subsequent QR iterations and it guarantees the generators sizes to be bounded by 3. Another implication of the equation above is that in exact arithmetics, $X^TY + \widehat{x}\widehat{y}^T$ is an upper triangular matrix.

5.6. Deflation and Convergence Criterion

After showing the details of the fast structured bulge chasing schemes we provide the deflation technique and the convergence criterion in terms of SSS representations.

Deflation is an important concept in the practical implementation of the QR iteration method. It amounts to setting small subdiagonal elements of the Hessenberg matrix to zero. After deflation, it splits the Hessenberg matrix into two smaller subproblems which may be independently refined further. Theoretically, assume that deflation occurs to an intermediate Hessenberg matrix

$$H = Q_1 \cdots Q_{n-1} \cdot R,$$

and a subdiagonal entry $h_{i,i-1}$ of H becomes 0. This corresponds to the fact that the Givens matrix Q_{i-1} in the Q-factor sequence of H becomes an identity matrix:

$$\begin{aligned} H &= (Q_1 \cdots Q_{i-2}) \cdot Q_{i-1} \cdot (Q_i \cdots Q_{n-1}) \cdot R \\ &= (Q_1 \cdots Q_{i-2}) \cdot \quad I \quad \cdot (Q_i \cdots Q_{n-1}) \cdot R. \end{aligned} \tag{23}$$

In traditional deflation schemes H will be treated as two subproblems individually. That means here we have to look for a new orthogonal-plus-rank-one representation such as (4) for each subproblem. It is not obvious so far how we can quickly get those representations based on the original orthogonal-plus-rank-one representation. However, instead of seeking new representations, we will keep the original orthogonal-plus-rank-one representation, reuse the original Q- and R-factors, and in the meantime, keep track of the identity matrices such as Q_{i-1}. The identity matrix Q_{i-1} in (23) splits the Q_j factors into two subgroups (corresponding to the two subproblems in traditional deflation schemes). In later bulge chasing steps, operations will be done within each subgroup. That is, we maintain global representations for Q- and R-factors, but keep the actual structured operations locally within subgroups.

We also need to take care of deflation criteria based on the low-rank structures. In traditional computations there are various deflation criteria, such as the one proposed by Wilkinson which is used in LAPACK [2] and a new one proposed by Ahues and Tisseur [1]. For our new QR algorithm, we can adopt similar criteria. The difference is that since the Hessenberg iterate H is not explicitly formed, we

need to compute relevant elements of H on the fly through compact representations of Q and R. For example, Wilkinson's deflation criterion will set $h_{i,i-1}$ to zero if

$$|h_{i,i-1}| \leq \tau \cdot (|h_{i-1,i-1}| + |h_{i,i}|), \tag{24}$$

where τ is a given tolerance. In terms of the elements of Q and R we have

$$\begin{pmatrix} h_{i-1,i-1} & \times \\ h_{i,i-1} & h_{i,i} \end{pmatrix} = \begin{pmatrix} -s_{i-2} & c_{i-2}c_{i-1} & \times \\ & -s_{i-1} & c_{i-1}c_i \end{pmatrix} \begin{pmatrix} u_{i-2}v_{i-1}^T & \times \\ d_{i-1} & u_{i-1}v_i^T \\ & d_i \end{pmatrix},$$

where $\times$ denotes certain element in the corresponding matrix. This gives us

$$\begin{cases} h_{i,i-1} &= -s_{i-1}d_{i-1}, \\ h_{i-1,i-1} &= -s_{i-2}(u_{i-2}v_{i-1}^T) + c_{i-2}c_{i-1}d_{i-1}, \\ h_{i,i} &= -s_{i-1}(u_{i-1}v_i^T) + c_{i-1}c_i d_i. \end{cases}$$

When the criterion (24) is satisfied, we want to set $h_{i,i-1}$ to zero. However, since H is not explicitly stored, we choose to do this by making s_{i-1} zero. There are two possible scenarios:

1. If $|s_{i-1}| \leq O(\epsilon)$, with ϵ being the machine precision, it's straightforward: we will just set $s_{i-1} \equiv 0$ and $c_{i-1} \equiv \text{sign}(c_{i-1})$ without changing anything else.
2. If $|s_{i-1}| > O(\epsilon)$, things become tricky. We first multiply $(Q_{i-1}Q_i \cdots Q_{n-1})$ to R to get $H(i-1:n, i-1:n)$ in its SSS form. We then find another sequence of Givens rotation matrices $(\widehat{Q}_{i-1}\widehat{Q}_i \cdots \widehat{Q}_{n-1})$, whose transpose applied to the left side of $H(i-1:n, i-1:n)$ will yield a new upper triangular matrix $\widehat{R}$. Note that:
   (a) $\widehat{Q}_{i-1}$ is automatically an identity matrix, since $h_{i,i-1}$ is small enough to be ignored;
   (b) all matrix-matrix multiplications are done quickly by updating SSS generators.

In the standard QR algorithm, we say that the algorithm converges if the Hessenberg iterate H_k eventually becomes a real quasi triangular matrix (called the Schur form). In our new QR algorithm for real companion matrices, we say that the algorithm converges if the Q-factor in its trigonometric parametrization form $Q = Q_1Q_2 \cdots Q_{n-1}$ satisfies the following *convergence criterion*: for any two consecutive Givens rotations $\{Q_k, Q_{k+1}\}$ $(k = 1, 2, \ldots, n-2)$, one of them must be an identity matrix.

5.7. Summary of the new QR algorithm for C

The gist of our new QR algorithm for companion matrices is the usage of compact representations for Q (as a product of a sequence of Givens rotations) and for R (in terms of its SSS form) during the QR iteration process. The feasibility of such compact representations for Q and R is guaranteed by the fact that the Hessenberg iterates of the companion matrix during QR iteration process have low-rank off-diagonal blocks (the maximum off diagonal rank of H never exceeds 3). Similar low-rank properties extend to the Q- and R-factors of H.

In terms of compact representations of Q and R, rather than explicitly forming and updating structured matrices for the Hessenberg iterates H as done in [4] and [6], we may summarize our new QR iteration method in Algorithm 2.

6. Balancing Strategy

We also briefly mention the balancing strategy. Before QR iterations for the eigenvalues of a matrix A we usually apply a diagonal similarity transformation to A for the purpose of better accuracy and efficiency. That is, we compute the eigenvalues of DAD^{-1} where D is a diagonal matrix. The matrix D is often chosen such that the norms of each row and the corresponding column of DAD^{-1} are close.

A similar balancing strategy as in [4] can be used. In our new fast eigensolver for the companion matrix C, we have exploited the fact that the Hessenberg iterates under the QR iteration have low-rank off-diagonal blocks, so we are able to use compact representations for the Q- and and R-factors. However, after balancing these rank structures for the iterates of DCD^{-1} may be destroyed, where $D = \mathrm{diag}(d_1, \ldots, d_n)$. That is, The Hessenberg iterates for DCD^{-1} may no longer have low-rank off-diagonal blocks. However, notice

$$DCD^{-1} = \begin{pmatrix} a_1 & \frac{d_1}{d_2}a_2 & \cdots & \frac{d_1}{d_{n-1}}a_{n-1} & \frac{d_1}{d_n}a_n \\ \frac{d_2}{d_1} & 0 & \cdots & 0 & 0 \\ 0 & \frac{d_3}{d_2} & \cdots & 0 & 0 \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & 0 & \cdots & \frac{d_n}{d_{n-1}} & 0 \end{pmatrix}. \tag{25}$$

If we can select D such that

$$\frac{d_2}{d_1} = \frac{d_3}{d_2} = \cdots = \frac{d_n}{d_{n-1}} \equiv \alpha$$

for certain α, then DCD^{-1} becomes the multiple of a new companion matrix:

$$DCD^{-1} = \alpha \cdot \begin{pmatrix} \frac{a_1}{\alpha} & \frac{a_2}{\alpha^2} & \cdots & \frac{a_{n-1}}{\alpha^{n-1}} & \frac{a_n}{\alpha^n} \\ 1 & 0 & \cdots & 0 & 0 \\ 0 & 1 & \cdots & 0 & 0 \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & 0 & \cdots & 1 & 0 \end{pmatrix} \equiv \alpha \cdot \widehat{C},$$

where $\widehat{C}$ is the companion matrix corresponding to the polynomial $p(\alpha x)/\alpha^n$, with $p(x)$ being the polynomial (2) corresponding to the original companion matrix C. This means that we can choose a geometric scaling ($d_i = \alpha^i$), and apply the fast QR iterations to $\widehat{C}$ so as to preserve the low-rank structures. After the eigenvalues of the new companion matrix $\widehat{C}$ are obtained we can multiply them by α to get those of C.

Some efficient balancing algorithms for a given matrix A based on the approximation of Perron vectors of $|A|$ are developed in [12]. It was also shown

Algorithm 2 New structured QR algorithm for a real companion matrix C

Input: the first row of C: $\begin{pmatrix} a_1 & a_2 \dots & a_{n-1} & a_n \end{pmatrix}$
Output: Q: in terms of $\{c(Q), s(Q)\}$;
R: in terms of $\{d(R), u(R), v(R), w(R)\}$.

(1) **Initialization**
  (a) Compute QR factorization of C: $C = Q_1 Q_2 \dots Q_{n-1} \cdot R$.
  (b) Find x and y such that $C = Z + xy^T$. [*Note that only* $\{c_i(Q), s_i(Q)\}$, $\{d_i(R), u_i(R), v_i(R), w_i(R)\}$, x *and* y *are explicitly stored.*]
(2) **Repeat**
  (a) **Modified Bulge Chasing** with shift(s)
    (i) Determine what shift to use (Francis single or double shift or exceptional shift).
    (ii) For $i = 1$, find $\bar{G}_i$ to create a bulge on subdiagonal of R and then find $\widetilde{G}_i$ to eliminate it.
    (iii) For $i = 2, \dots, n-1$:
    (iv) Update Q by Givens swaps: $Q_{i-1} Q_j \widetilde{G}_{i-1} \Rightarrow \bar{G}_i \widehat{Q}_{i-1} \widehat{Q}_i$. Store $\bar{G}_i$.
    (v) Update R by bulge elimination: find $\widetilde{G}_i$ to eliminate the bulge in $R\bar{G}_i$. For example, for single shift:
      Update $d_i(R), d_{i+1}(R)$, form the bulge b_i in $R\bar{G}_i$, and update h_i, as in (18).
      Compute $\widetilde{G}_i$ and update $d_i(R), d_{i+1}(R)$ as in (19).
      Update $u_i(R), u_{i+1}(R)$ as in (20).
    (vi) Endfor
    (vii) Merge $\widetilde{G}_{n-1}$ into Q_{n-1}: $\widehat{Q}_{n-1} := Q_{n-1} \widetilde{G}_{n-1}$. Each $\widehat{Q}_i$ becomes the new Q_i.
    (viii) Get updated SSS representation for R by two SSS matrix multiplications (see, e.g. (17)).
  (b) **Deflation**:
    (i) If $H_{i+1,i}$ is small enough to be thrown away **and** if Q_i is not an identity matrix, update $Q_i, \dots, Q_{n-1}$ and the corresponding parts of SSS generators of $\widehat{R}$.
  (c) **Restore Compact Representation of** R
    (i) $\widehat{Q}$ and $\bar{G}$ are available through the parametric representations of $\widehat{Q}_i$ and $\bar{G}_i$, respectively. Let $\widehat{x} := \widehat{Q}^T \bar{G}^T x$, $\widehat{y} = \bar{G}^T y$, then $\widehat{R}$ satisfies: $\widehat{R} = \widehat{Z} + \widehat{x}\widehat{y}^T$ for some orthogonal $\widehat{Z}$.
    (ii) Find X so $X\widehat{x} = e_1 \implies X\widehat{R} - e_1 \widehat{y}^T$ is orthogonal and upper Hessenberg.
    (iii) Find Y so that $(X\widehat{R} - e_1 \widehat{y}^T) Y^T = I$.
    (iv) Compute SSS generators $\{d_i(R), u_i(R), v_i(R), w_i(R)\}$ of $R := X^T Y + \widehat{x}\widehat{y}^T$: first use SSS multiplications to obtain an SSS form for $X^T Y$ with generator sizes no larger than 2. Then use SSS additions to obtain an SSS form for R with generator sizes no larger than 3.
Until convergent

that if A is irreducible and x and y are the right and left Perron vectors of $|A|$, respectively, then $D = \mathrm{diag}(1/x_1, \ldots, 1/x_n)$ minimizes $\|DAD^{-1}\|_\infty$, and $D = \mathrm{diag}(\sqrt{y_1/x_1}, \ldots, \sqrt{y_n/x_n})$ minimizes $\|DAD^{-1}\|_2$. Here C is a companion matrix, and so is $|C|$. The matrix C has a right Perron vector with entries $x_i = \alpha^{n-i}$, where α is the maximum positive eigenvalue of $|C|$, or equivalently the largest positive root of $x^n - |a_1|x^{n-1} - \cdots - |a_{n-1}|x - |a_n|$. Therefore, a geometric scaling with such an α minimizes the infinity-norm of DCD^{-1}. In our algorithm, however, only orthogonal transformations are applied. Ideally, we should look for a geometric scaling strategy such that $\|DCD^{-1}\|_2$ is minimized. Empirically, we find the following criterion for choosing α to be useful: choosing α to make

$$\mathrm{Range}\{|\widehat{c}_1|, |\widehat{c}_2|, \ldots, |\widehat{c}_n|, 1\} \equiv \frac{\max\{|\widehat{c}_1|, \ldots, |\widehat{c}_n|, 1\}}{\min\{|\widehat{c}_1|, \ldots, |\widehat{c}_n|, 1\}}$$

as small as possible, where $\widehat{c}_i = \frac{a_i}{\alpha^i}$.

In practice, α is often selected to be a power of the machine radix so as to avoid errors in computing DCD^{-1}. In our numerical experiments we have tried different powers of 2 as α (see the next section), although more work needs to be done on a systematic way of choosing α.

7. Numerical Experiments

We have tested our new structured QR algorithm on many different examples and it is stable in practice, although it is still an open problem to show whether the new algorithm is stable or not. We implemented the new QR-iteration method in FORTRAN 90 for computing the eigenvalues of real companion matrices. The codes are available online.[1] Numerical experiments are run on a laptop with an Intel Pentium M 1.7GHz CPU and 512MB RAM. Results are summarized in the following two subsections to illustrate both the performance, i.e., $O(n^2)$ complexity and the stability in practice.

We first point out that among all our numerical tests, the program runs stably and we did not observe any significant failure or corruption of the orthogonal-plus-rank-one structures by using the compact SSS QR factors. The low-rank Hessenberg structures are well preserved in the experiments.

7.1. $O(n^2)$ complexity Tests

We use real polynomials with uniformly random coefficients as test polynomials. The degree of the polynomials doubles from 25 up to 102,400. We also show the relative backward error

$$\frac{\|\bar{G}^T \cdot C_0 \cdot \bar{G} - Q^{(m)}R^{(m)}\|_\infty}{\|C_0\|_\infty},$$

where C_0 denotes the initial companion matrix, m is the number of iterations needed for convergence, $Q^{(m)}$ and $R^{(m)}$ are explicitly formed Q- and R-factors

[1] http://www.math.ucla.edu/~jxia/work/companion/

n (size)	DGEEV(sec)	New SSS(sec)	iter. #	rel. BkErr
25	0.01	0.01	83	1×10^{-15}
50	0.03	0.03	161	2×10^{-15}
100	0.12	0.09	309	3×10^{-15}
200	0.33	0.22	584	7×10^{-15}
400	1.70	0.51	1200	2×10^{-14}
800	12.33	1.98	2165	3×10^{-14}
1,600	95.82	7.43	4170	1×10^{-13}
3,200	865.22	56.11	8125	
6,400	-	296.21	15569	
12,800	-	1,302.22	30551	
25,600	-	5,465.76	62080	
51,200	-	21,080.34	116708	
102,400	-	83,583.64	252822	

TABLE 4. *Numerical results on new $O(n^2)$ companion eigensolver.*

of the final convergent Schur form of C_0, and $\bar{G}$ is the accumulated orthogonal similarity transformation.

Remark 7.1. The break-even size of the current new companion eigensolver implementation versus LAPACK is about $n = 50$. For the test problem of size $102,400$, it took the new companion eigensolver about 23 hours to converge all the roots; on the other hand, the LAPACK routine DGEEV can't even run for problems of size about $8,000$ since it uses $O(n^2)$ storage; even if the memory was not an issue, it would take DGEEV more than 300 days to converge on the same machine since it's an $O(n^3)$ method.

Remark 7.2. From Table 4 and Figure 1, we clearly see the quadratic (i.e. $O(n^2)$) complexity of the new QR iteration Algorithm 1, see Figure 1 (a). The average iteration number needed per eigenvalue is less than 3, see Figure 1 (b). In the mean time, we observe nearly linear growth in both backward and forward errors.

Note that Figure 1 (a) reports the ratio between the running time for matrices of sizes $n = 25 \times 2^k$ and $n = 25 \times 2^{k-1}$. Since the new companion eigensolver is an $O(n^2)$ algorithm, we expect the ratio to be close to 4 for large n.

7.2. Backward Stability Tests

If the new QR algorithm for companion matrix is backward stable in eigenproblem sense, then according to error analysis by Van Dooren and Dewilde [13], and further by Edelman and Murakami [16], the new algorithm is also backward stable in polynomial sense, more precisely, the "calculus" definition holds: "the first order perturbations of the matrix lead to first order perturbations of the coefficients", see [16] for details.

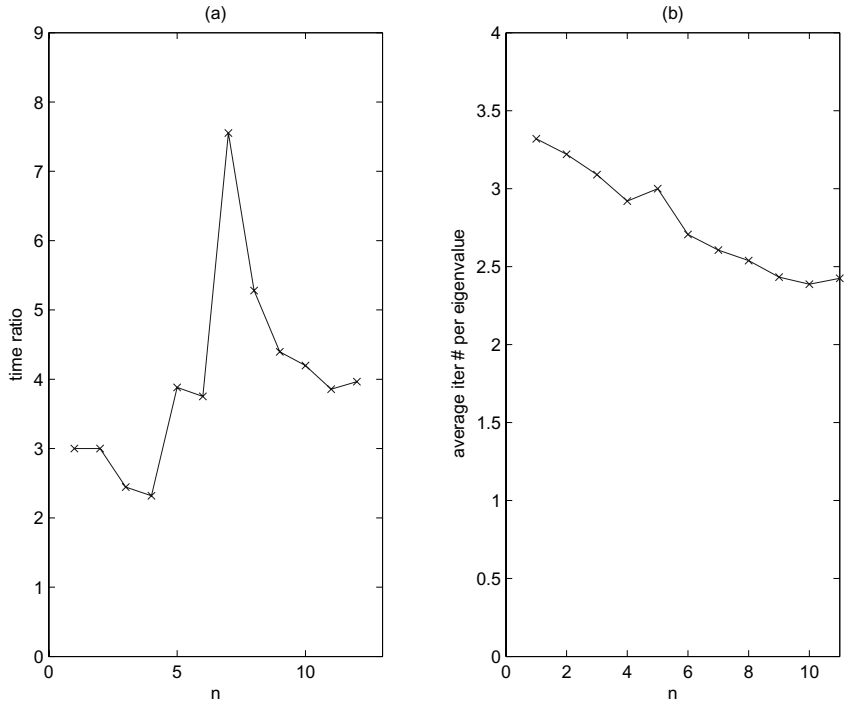

FIGURE 1. *New companion eigensolver, with test matrices of size* $25 \times 2^{n-1}$, $1 \le n \le 13$.

Following Toh and Trefethen [27] and Edelman and Murakami [16], we explore the following degree 20 monic real coefficient polynomials:

(1) "Wilkinson polynomial": zeros $1, 2, 3, \ldots, 20$.
(2) the monic polynomial with zeros $[-2.1 : 0.2 : 1.7]$.
(3) $p(z) = (20!) \sum_{k=0}^{20} z^k/k!$.
(4) the Bernoulli polynomial of degree 20.
(5) the polynomial $z^{20} + z^{19} + z^{18} + \cdots + z + 1$.
(6) the univariate polynomial with zeros $2^{-10}, 2^{-9}, 2^{-8}, \ldots, 2^9$.
(7) the Chebyshev polynomial of degree 20.

In addition, we tested some random polynomials of degree $100, 200, \ldots, 1600$:

(8) random coefficients with uniform distribution.

Like what Edelman and Murakami did in their paper [16], for each example above, we first computed the coefficients either exactly or with ultra-high precision using MPFUN90 (Multiple Precision package by David Bailey, [3]). Then we rounded these numbers to double precision (in F90). And we took the rounded polynomials stored in F90 to be our official test cases.

For all test cases, we computed two sets of relative backward errors. One is the norm-wise matrix relative backward error:

$$\frac{\|E\|_\infty}{\|\widehat{C}\|_\infty} \equiv \frac{\|\bar{G}^T \cdot \widehat{C} \cdot \bar{G} - Q^{(m)}R^{(m)}\|_\infty}{\|\widehat{C}\|_\infty},$$

where $\widehat{C}$ denotes the scaled companion matrix after balancing in (25), $\bar{G}$ is the accumulated orthogonal similarity transformation, and $Q^{(m)}R^{(m)}$ converges to the Schur form of $\widehat{C}$. The other is the component-wise coefficient relative backward error:

$$\frac{|\delta\,\widehat{c}_i|}{|\widehat{c}_i|} \equiv \frac{|\widetilde{c}_i - \widehat{c}_i|}{|\widehat{c}_i|},$$

where $\widehat{c}$ corresponds to the coefficient of the characteristic polynomial of $\widehat{C}$, and $\widetilde{c}_i$ is the ith coefficient of the polynomial recovered from the computed zeros by using ultra-high precision, e.g. MPFUN90.

Test	(1)	(2)	(3)	(4)	(5)	(6)	(7)
α	8	1	8	2	1	1/4	1/2
$\|\widehat{C}\|_\infty$	7	3	4	2	2	22	4
rel_bkerr	10^{-15}	10^{-15}	10^{-16}	10^{-15}	10^{-15}	10^{-16}	10^{-15}

TABLE 5. *Test* (1–7)*: matrix norm-wise backward errors.*

7.2.1. Test (1-7), degree 20.

Remark 7.3. 1. The last two rows of Table 6 show (1) x_{max}: the maximum positive root of $p_b(x) = x^n - |c_1|x^{n-1} - \cdots - |c_{n-1}|x_1 - |c_n|$, and (2) α: the particular scaling factor chosen so that the maximum coefficient backward error is minimized. As we can see, such α usually doesn't agree well with x_{max}. Although using x_{max} as scaling factor will minimize $\|DCD^{-1}\|_\infty$, the magnitudes of the coefficients of the new polynomial under such scaling could vary wildly.

2. The empty entries for Test 4 and 7 correspond to zero coefficients.

7.2.2. Test(8), random polynomials, degree $100, 200, \ldots 1600$**.**

Remark 7.4. 1. From Table 7, we can see that the new companion eigensolver has small backward error in matrix-norm sense, it also finds roots with small (coefficient) backward errors. In our random polynomial experiments, we choose $\alpha = 1$. When the size of polynomial gets bigger, to balance the corresponding companion matrix with geometric scaling limits our option.

2. Where the "average abs_bkerr" (average absolute backward error) is computed as average of $\{\log_{10}|c_i|\}$, and the "average rel_bkerr" (average relative backward error) is computed as average of $\left\{\log_{10}\frac{|\delta c_i|}{|c_i|}\right\}$.

index/Test	(1)	(2)	(3)	(4)	(5)	(6)	(7)
1	10^{-15}	10^{-14}	10^{-14}	10^{-14}	10^{-14}	10^{-14}	-
2	10^{-15}	10^{-14}	10^{-14}	10^{-14}	10^{-13}	10^{-14}	10^{-15}
3	10^{-14}	10^{-15}	10^{-14}	-	10^{-13}	10^{-13}	-
4	10^{-14}	10^{-12}	10^{-14}	10^{-14}	10^{-14}	10^{-13}	10^{-15}
5	10^{-14}	10^{-14}	10^{-14}	-	10^{-14}	10^{-13}	-
6	10^{-14}	10^{-13}	10^{-14}	10^{-14}	10^{-13}	10^{-13}	10^{-15}
7	10^{-14}	10^{-14}	10^{-14}	-	10^{-13}	10^{-13}	-
8	10^{-14}	10^{-14}	10^{-14}	10^{-14}	10^{-13}	10^{-13}	10^{-15}
9	10^{-14}	10^{-14}	10^{-14}	-	10^{-13}	10^{-13}	-
10	10^{-14}	10^{-15}	10^{-14}	10^{-14}	10^{-13}	10^{-13}	10^{-14}
11	10^{-14}	10^{-13}	10^{-14}	-	10^{-13}	10^{-13}	-
12	10^{-14}	10^{-14}	10^{-14}	10^{-14}	10^{-13}	10^{-13}	10^{-14}
13	10^{-13}	10^{-13}	10^{-14}	-	10^{-13}	10^{-13}	-
14	10^{-13}	10^{-14}	10^{-14}	10^{-14}	10^{-13}	10^{-13}	10^{-14}
15	10^{-13}	10^{-13}	10^{-14}	-	10^{-13}	10^{-13}	-
16	10^{-13}	10^{-13}	10^{-14}	10^{-14}	10^{-13}	10^{-13}	10^{-14}
17	10^{-13}	10^{-14}	10^{-14}	-	10^{-14}	10^{-13}	-
18	10^{-13}	10^{-13}	10^{-14}	10^{-13}	10^{-14}	10^{-13}	10^{-14}
19	10^{-13}	10^{-12}	10^{-14}	-	10^{-14}	10^{-12}	-
20	10^{-13}	10^{-13}	10^{-14}	10^{-14}	10^{-14}	10^{-12}	10^{-14}
max bkerr	10^{-13}	10^{-12}	10^{-14}	10^{-13}	10^{-13}	10^{-12}	10^{-14}
x_{max}	296.2	6.1	38.2	12.6	2.0	1319.8	2.6
α	8	1	8	2	1	1/4	1/2

TABLE 6. *Test* (1–7)*: coefficient-wise backward errors with appropriate* α*.*

size	matrix-wise		polynomial coeff.-wise			
	$\|\widehat{C}\|_\infty$	rel_bkerr	average abs_fwderr	average abs_fwderr		
100	5×10^1	3×10^{-15}	10^{-14}	10^{-13}		
200	9×10^1	7×10^{-15}	10^{-13}	10^{-13}		
400	2×10^2	2×10^{-14}	10^{-12}	10^{-12}		
800	4×10^2	3×10^{-14}	10^{-12}	10^{-11}		
1600	8×10^2	1×10^{-13}	10^{-11}	10^{-11}		

TABLE 7. *Test* (8)*: backward errors in matrix and polynomial coefficients.*

8. Conclusions

In this paper we presented a new fast QR algorithm for computing the eigenvalues of a real companion matrix. The algorithm is backward stable in practice. The success of the new method relies on (i) compact (SSS) representations for Q and

R, (ii) a new technique called Givens rotation swaps to update Q in an efficient fashion, and (iii) exploring the special rank structure of R for the purpose of efficient compression. The overall complexity is $O(n^2)$, though we have not yet derived the counts in detail. Our suspect is that the counts are similar to those in [6].

We also expect to propose a modified version with stability proof in the near future.

Acknowledgements

The authors are grateful to Professor Yuli Eidelman at Tel Aviv University and to the two anonymous referees for their valuable suggestions on this paper. We also thank Professor James Demmel and Doctor David Bindel at the University of California at Berkeley for their kind help in improving and testing the algorithm.

References

[1] Mario Ahues and Francoise Tisseur, *A new deflation criterion for the QR algorithm*, Technical Report CRPC-TR97713-S, Center for Research on Parallel Computation, January 1997.

[2] E. Anderson, Z. Bai, C. Bischof, S. Blackford, J. Demmel, J. Dongarra, J. Du Croz, A. Greenbaum, S. Hammarling, A. McKenney, S. Ostrouchov, D. Sorensen, *LAPACK Users' Guide, Release 2.0*, SIAM, Philadelphia, PA, USA, second edition, 1995.

[3] D. Bailey, *Software: MPFUN90 (Fortran-90 arbitrary precision package)*, available online at http://crd.lbl.gov/~dhbailey/mpdist/index.html

[4] D. Bindel, S. Chandresekaran, J. Demmel, D. Garmire, and M. Gu, *A fast and stable nonsymmetric eigensolver for certain structured matrices*, Technical report, University of California, Berkeley, CA, 2005.

[5] D. A. Bini, F. Daddi, and L. Gemignani, *On the shifted QR iteration applied to companion matrices, Electronic Transactions on Numerical Analysis* **18** (2004), 137–152.

[6] D. A. Bini, Y. Eidelman, L. Gemignani and I. Gohberg, *Fast QR eigenvalue algorithms for Hessenberg matrices which are rank-one perturbations of unitary matrices*, Technical Report no.1587, Department of Mathematics, University of Pisa, 2005.

[7] S. Chandrasekaran and M. Gu, *Fast and stable algorithms for banded plus semiseparable matrices*, SIAM J. Matrix Anal. Appl. **25** no. 2 (2003), 373–384.

[8] S. Chandrasekaran, M. Gu, and W. Lyons, *A fast and stable adaptive solver for hierarchically semi-separable representations*, Technical Report UCSB Math 2004-20, U.C. Santa Barbara, 2004.

[9] Chandrasekaran, P. Dewilde, M. Gu, T. Pals, X. Sun, A.-J. van der Veen, and D. White, *Fast stable solvers for sequentially semi-separable linear systems of equations and least squares problems*, Technical report, University of California, Berkeley, CA, 2003.

[10] S. Chandrasekaran, P. Dewilde, M. Gu, T. Pals, X. Sun, A.-J. van der Veen, and D. White, *Some fast algorithms for sequentially semiseparable representations*, SIAM J. Matrix Anal. Appl **27** (2005), 341–364.

[11] S. Chandrasekaran, M. Gu, X. Sun, J. Xia, J. Zhu, *A superfast algorithm for Toeplitz systems of linear equations*, SIAM J. Mat. Anal. Appl., to appear.

[12] T.-Y. Chen and J.W. Demmel, *Balancing sparse matrices for computing eigenvalues*, Lin. Alg. and Appl. **309** (2000), 261–287.

[13] P. Van Dooren and P. Dewilde, *The eigenstructure of an aribtrary polynomial matrix: Computational aspects*, Lin. Alg. and Appl. **50** (1983), 545–579.

[14] Y. Eidelman and I. Gohberg, *On a new class of structured matrices*, Integral Equations Operator Theory **34** (1999), 293–324.

[15] Y. Eidelman, I. Gohberg and V. Olshevsky, *The QR iteration method for Hermitian quasiseparable matrices of an arbitrary order*, Lin. Alg. and Appl. **404** (2005), 305–324.

[16] A. Edelman and H. Murakami, *Polynomial roots from companion matrix eigenvalues*, Mathematics of Computation **64** (1995), 763–776.

[17] G. Golub and C. V. Loan, *Matrix Computations*, The John Hopkins University Press, 1989.

[18] J. G. F. Francis, *The QR transformation. II*, Comput. J. **4** (1961/1962), 332–345.

[19] V. N. Kublanovskaya, *On some algorithms for the solution of the complete eigenvalue problem*, U.S.S.R. Comput. Math. and Math. Phys. **3** (1961), 637–657.

[20] C. Moler, *Roots – of polynomials, that is*, The Mathworks Newsletter **5** (1991), 8–9.

[21] V. Pan, *On computations with dense structured matrices*, Math. Comp. **55** (1990), 179–190.

[22] B. Parlett, *The symmetric eigenvalue problems*, SIAM, 1997.

[23] B. Parlett, *The QR algorithm*, Computing in Science and Engineering **2** (2000), 38–42. Special Issue: Top 10 Algorithms of the Century.

[24] G. Sitton, C. Burrus, J. Fox, and S. Treitel, *Factoring very-high-degree polynomials.* IEEE Signal Processing Mag. **20** no. 6 (2003), 27–42.

[25] M. Stewart, *An error analysis of a unitary Hessenberg QR algorithm*, Tech. Rep. TR-CS-98-11, Department of Computer Science, Australian National University, Canberra 0200 ACT, Australia, 1998.

[26] F. Tisseur, *Backward stability of the QR algorithm*, TR 239, UMR 5585 Lyon Saint-Etienne, October 1996.

[27] K.-C. Toh and L. N. Trefethen. *Pseudozeros of polynomials and pseudospectra of companion matrices.* Numer. Math. **68** (1994), 403–425.

[28] M. Van Barel and A. Bultheel, *Discrete Linearized Least Squares Rational Approximation on the Unit Circle*, J. Comput. Appl. Math. **50** (1994), 545–563.

[29] J. H. Wilkinson, *The algebraic eigenvalue problem*, Oxford University Press, London, 1965.

[30] J. Xia, *Fast Direct Solvers for Structured Linear Systems of Equations*, Ph.D. Thesis, University of California, Berkeley, 2006.

[31] J. Zhu, *Structured Eigenvalue Problems and Quadratic Eigenvalue Problems*, Ph.D. Thesis, University of California, Berkeley, 2005.

Shiv Chandrasekaran
Department of Electrical and Computer Engineering
University of California at Santa Barbara
USA
e-mail: shiv@ece.ucsb.edu

Ming Gu
Department of Mathematics
University of California at Berkeley
USA
e-mail: mgu@math.berkeley.edu

Jianlin Xia
Department of Mathematics
University of California at Los Angeles
USA
e-mail: jxia@math.ucla.edu

Jiang Zhu
Department of Mathematics
University of California at Berkeley
USA
e-mail: zhujiang.cal@gmail.com

Operator Theory:
Advances and Applications, Vol. 179, 145–155

The Numerical Range of a Class of Self-adjoint Operator Functions

Nurhan Çolakoğlu

Abstract. The structure of the numerical range and root zones of a class of operator functions, arising from one or two parameter polynomial operator pencils of waveguide type is studied. We construct a general model of such kind of operator pencils. In frame of this model theorems on distribution of roots and eigenvalues in some parts of root zones are proved. It is shown that, in general the numerical range and root zones are not connected but some connected parts of root zones are determined. It is proved that root zones, under some natural additional conditions which are satisfied for most of waveguide type multi-parameter spectral problems, are non-separated, i.e., they overlap.

Mathematics Subject Classification (2000). 47A56; 47A12.

Keywords. Waveguide, operator pencil, numerical range, root zone, eigenvalue.

1. Introduction

The purpose of this paper is to study the numerical range and the structure of root zones for a class of self-adjoint operator functions, arising from one or two parameter polynomial operator pencils of waveguide type (w.g.t.). These are questions mainly from the variational theory of the spectrum of operator pencils of w.g.t. The main difficulties in the variational theory of the spectrum of multi-parameter operator pencils are based on the fact that their root zones overlap, i.e., they are non-overdamped pencils. For this reason, we construct a general model of self-adjoint operator pencils which contains not only operator pencils of w.g.t., but also a wide class of non-overdamped pencils. Throughout, we study these problems in the frame of this model (see conditions **(I)**–**(IV)**).

The study of waveguiding systems of an arbitrary order often leads to the spectral theory of two parameter polynomial operator pencils, so-called pencils of w.g.t. (see [1], [2] and [9], see also for definition Examples 3.1, 3.2 and 3.3), in the

form

$$L(k,w) := A + \sum_{s=1}^{n} k^{2s} C_{s-1} + \sum_{s=0}^{n-1} k^{2s+1} B_s + iwD - w^2 I,$$

where A, D, B_s and C_s, $s = 0, 1, \ldots, n-1$, are symmetric operators in a Hilbert space H and all but C_{n-1} may be unbounded.

Such kind of operator pencils arise from a dynamical model of regular waveguiding system constructed by A. S. Silbergleit and Yu. I. Kopilevich in [7], [8] and [9]. These works are devoted to general spectral problems for quadratic operator pencils of w.g.t. It seems that [2] is the first paper devoted to the spectral theory (mainly variational theory of the spectrum) of two parameter polynomial operator pencils of w.g.t. Although variational principles for definite type eigenvalues for quadratic operator pencils of w.g.t. were studied in detail in [1], this paper also contains some new results (see Theorem 2.2) about the numerical range and root zones in this case.

In the spectral theory, especially in the variational theory of the spectrum of one or two parameter operator pencils of w.g.t., we often deal with operator functions whose root zones overlap in an interval $[a, b]$ (see [1], [2], [4], [5]). Namely we have an operator function L and the equation $(L(\lambda)x, x) = 0$ has only two roots $p_-(x)$ and $p_+(x)$ in $[a, b]$ for some x from a cone G' in a Hilbert space H. We recall that a real root λ of the equation $(L(\lambda)x, x) = 0$ is said to be of the first kind, of the second kind, and neutral if the number $(L'(\lambda)x, x)$ is greater than zero, less than zero, and equal to zero, respectively. We define also the cone

$$G = \{x \in G' \mid p_-(x) \neq p_+(x)\}$$

and the bounds of the ranges of functionals $p_\pm(x)$ on G and G':

$$\delta_- = \inf_G p_+(x), \qquad \delta_+ = \sup_G p_-(x),$$

$$k_- = \inf_G p_-(x), \qquad k_+ = \sup_G p_+(x),$$

$$k'_- = \inf_{G'} p_-(x), \qquad k'_+ = \sup_{G'} p_+(x).$$

In the spectral theory of one or two parameter operator pencils of w.g.t., in general, we have two different models for distribution of roots and curves $(L(\lambda)x, x)$ (see Figures 1 and 2).

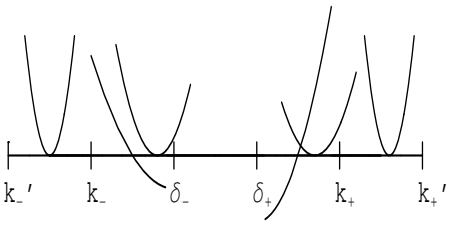

Figure 1. Model A

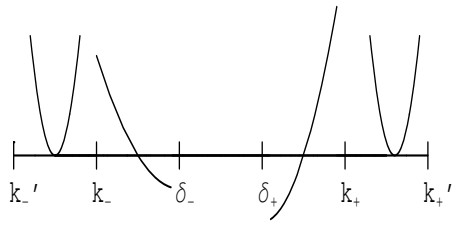

Figure 2. Model B

Especially in the solutions of variational problems, distribution of roots as in model B (Figure 2) and connectedness of the parts of root zones in the intervals $[k_-, \delta_-)$ and $(\delta_+, k_+]$ are very important.

In this paper we aim to give conditions under which such distribution of roots (Figure 2) occurs. This is considered in Section 2.

In addition, examples of one and two parameter operator functions satisfying these conditions are given in Section 3.

2. On the structure of root zones

Let A be a bounded linear operator on the Hilbert space H. The set of all numbers of the form (Ax, x), where $\|x\| = 1$, is called the *numerical domain* of A and denoted by $W(A)$. It is obvious that $W(A)$ is a nonempty subset of $\mathbb{C}$. This set is not closed in general. If $Ax = \lambda x$ $(\|x\| = 1)$, then $(Ax, x) = \lambda$, i.e., all the eigenvalues of A are in $W(A)$. The spectrum of A need not be contained in $W(A)$ but is necessarily in $\overline{W(A)}$. We know also that $W(A)$ is a convex set.

Let $A(\lambda)$ be an operator function whose values are bounded operators. The set of all roots of all possible functions $(A(\lambda)\, x, x)$ $(x \neq 0)$ is called the *numerical range* of the operator function $A(\lambda)$ and denoted by $R(A)$. In other words, $\lambda_0 \in R(A)$ if there exits a vector x_0 such that $\|x_0\| = 1$ and $(A(\lambda_0)x_0, x_0) = 0$. Obviously, each eigenvalue of $A(\lambda)$ is in the numerical range.

Note. In some resources the notion of 'numerical domain' for an operator, defined above, is called 'numerical range'. That is, the term 'numerical range' is used both for operators and operator functions (see [3]). But in order to differantiate between this two notions for operators and operator functions, we prefer to use both terms.

The numerical range of the operator function $A - \lambda I$ coincides with the numerical domain of the operator A, so the concept of the numerical range of an operator function is a natural generalization of the concept of the numerical domain of an operator.

The relation between numerical domain and spectrum of an operator can be generalized to operator functions. If $A(\lambda)$ is an operator function holomorphic in a domain U and there exists a number $z_0 \in U$ such that $0 \notin \overline{W(A(z_0))}$, then $\sigma(A) \subset \overline{R(A)}$ (see [6, p. 139]).

In contrast to the numerical domain of an operator, the numerical range of an operator function is nonconvex and even disconnected in general.

In the finite dimensional case, the following theorem exhibits a close relationship between the eigenvalues and numerical range of a monic self-adjoint matrix polynomial, namely, that every real boundary point of $R(L)$ is an eigenvalue of $L(\lambda)$.

Theorem 2.1 ([3, Theorem 10.15]**).** *Let $L(\lambda)$ be a monic self-adjoint matrix polynomial, and let $\lambda_0 \in R(L) \cap \overline{(\mathbb{R} \setminus R(L))}$. Then λ_0 is an eigenvalue of $L(\lambda)$.*

As mentioned above we are mainly interested in problems about the structure of root zones, which we encounter in the variational theory of operator pencils of w.g.t. in the nonoverdamped case. For this reason we prefer to deal, not with conditions on the coefficients, but with conditions that derive from them and are easier to apply in our case. So we construct a general model given by the conditions

(I) $L(k) : [a,b] \to S(H)$, $L \in C^1[a,b]$ and for all $x \neq 0$ from a cone G' in a Hilbert space H the equation $(L(k)x,x) = 0$ has only two roots $p_-(x)$, $p_+(x)$ in $[a,b]$ (multiplicities taken into account and $p_-(x) \leq p_+(x)$) and has no roots in $[a,b]$ for other $x \in H \setminus \{0\}$. Here $S(H)$ denotes the set of bounded self-adjoint operators in H.

(II) If $x \in G$, then $(L'(p_-(x))x,x) < 0$ and $(L'(p_+(x))x,x) > 0$, where

$$G = \{x \in G' \mid p_-(x) \neq p_+(x)\}.$$

(III) There exist a number $k \in [a,b]$ such that $(L(k)x,x) < 0$ if and only if $x \in G$.

(IV) If $\{x_n\} \subset G'$ weakly convergent to $x \in G'$, then

$$\liminf p_-(x_n) \geq p_-(x), \qquad \limsup p_+(x_n) \leq p_+(x).$$

Operator pencils of w.g.t. (see Examples 3.1, 3.3 and 3.2), as well as a wide class of nonoverdamped operator pencils, form a subclass of this model, i.e., they satisfy the conditions **(I)**–**(IV)**.

We set

$$W'_{p_\pm} := \{p_\pm(x) \mid x \in G'\}$$

and

$$W_{p_\pm} := \{p_\pm(x) \mid x \in G\}$$

which are called root zones of the pencil L.

Lemma 2.1. *The functionals $p_\pm$ are continuous on G'.*

Proof. Let $x_n, x \in G'$ and $x_n \to x$. We want to show that $p_+(x_n) \to p_+(x)$. Let $\beta_n = p_+(x_n)$. Since $\beta_n \in [a,b]$ it is bounded, it has a convergent subsequence. Let us denote it again by β_n and let $\beta_n \to \beta$. Now we must show that $\beta = p_+(x)$. Since

$$0 = (L(\beta_n)x_n, x_n) \to (L(\beta)x,x) = 0$$

it follows that $\beta = p_+(x)$ or $\beta = p_-(x)$. If we consider the condition **(II)**, then

$$0 \leq (L'(\beta_n)x_n, x_n) \to (L'(\beta)x,x) \geq 0.$$

Now there are two cases. If $(L'(\beta)x,x) > 0$, then $\beta = p_+(x)$. If $(L'(\beta)x,x) = 0$, then $\beta = p_+(x) = p_-(x)$. Consequently $p_+(x_n) \to p_+(x)$. □

The following theorem particularly shows that for operator pencils of w.g.t. we have distribution of roots as in Figure 2.

Theorem 2.2. *Let L be an operator function satisfying the conditions* **(I)**–**(IV)**. *Then we have the following properties:*

(i) $k_\pm \in \sigma_R(L) := \sigma(L) \cap \mathbb{R}$,

(ii) *If k_+ (k_-) is not a limit point of $\sigma(L)$, every $k \in W'_{p_+} \cap (\delta_+, k_+]$ $\big(k \in W'_{p_-} \cap [k_-, \delta_-)\big)$ is a root of the first kind (second kind). Particularly, all eigenvalues in $(\delta_+, k_+]$ $([k_-, \delta_-))$ are eigenvalues of first (second) kind.*

Proof. First we prove the property (i) for k_+. We select a sequence $\{x_n\}$ with the properties

$$x_n \in G, \quad \|x_n\| = 1, \quad p_+(x_n) \to k_+, \quad x_n \xrightarrow{w} x. \tag{1}$$

Since

$$|(L(k_+)x_n, x_n)| \le \|L(k_+) - L(p_+(x_n))\|,$$

we have

$$\lim_{n\to\infty} (L(k_+)x_n, x_n) = 0.$$

From the conditions **(II)**–**(III)** and the definition of k_+ follows that $L(k_+) \ge 0$. Consequently we have

$$\lim_{n\to\infty} L(k_+)x_n = 0, \quad L(k_+)x = 0. \tag{2}$$

From the existence of a sequence satisfying (1) and (2) it follows that $k_+ \in \sigma_R(L)$. In a similar way one shows that $k_- \in \sigma_R(L)$.

We now prove (ii). We show that roots in $W'_{p_+} \cap (\delta_+, k_+]$ are of the first kind. First we establish that k_+ is an eigenvalue and has an eigenvector of the first kind. We select a sequence having the properties (1) and (2). Since k_+ is not a limit point of $\sigma(L)$, the vector x cannot be zero, so from (2) it follows that k_+, x is an eigenpair. From the condition **(IV)** we have

$$p_+(x) - p_-(x) \ge \limsup p_+(x_n) - \liminf p_-(x_n).$$

Since $p_+(x_n) \to k_+$, choosing a subsequence we can write

$$p_+(x) - p_-(x) \ge \lim_{n\to\infty} p_+(x_n) - \lim_{n\to\infty} p_-(x_n).$$

We show that the right side of the inequality is strictly positive. Assume that

$$\lim_{n\to\infty} p_+(x_n) - \lim_{n\to\infty} p_-(x_n) = 0,$$

then

$$k_+ = \lim_{n\to\infty} p_+(x_n) = \lim_{n\to\infty} p_-(x_n) \le \delta_+.$$

So we obtain a contradiction to the fact that $\delta_+ < k_+$. So $p_-(x) < p_+(x)$ and $x \in G$. Since $k_+ \le p_+(x)$ by the condition **(IV)**, we have $k_+ = p_+(x)$ and the pair k_+, x is of the first kind by the condition **(II)**.

Now we show that k_+ is an eigenvalue of the first kind. Let z be an arbitrary (nonzero) eigenvector corresponding to k_+. If it is of the second kind, we have a contradiction with the fact that k_+ is the upper bound of p_+ on G. We show that z cannot be neutral. Assuming that $(L'(k_+)z, z) = 0$ we set $z_t = tx + (1-t)z$, $t \in [0,1]$ where x is the previously found eigenvector of the first kind for k_+. Since $L(k_+)z_t = 0$, we have $z_t \in G'$ and $p_+(z_t) = k_+$ for $t \in [0,1]$. Let $K = \{z_t \mid t \in [0,1]\}$, then $K \subset G'$ is a pathwise connected set. Note that the functional p_- is continuous on G', $p_-(z_0) = p_-(z) = k_+$ and $p_-(z_1) = p_-(x) \le \delta_+$. Since

$p_-(K)$ is connected, for every $k \in (p_-(x), p_-(z))$ there exist a $z_{t_*} \in K$ such that $p_-(z_{t_*}) = k$. If we choose k such that $\delta_+ < k < k_+$ we have $p_-(z_{t_*}) = k < k_+$, $p_+(z_{t_*}) = k_+$ and $z_{t_*} \in G$. Since $p_-(z_{t_*}) = k > \delta_+$ this leads to a contradiction with the fact that δ_+ is the upper bound of p_- on G. We conclude that z cannot be neutral and k_+ is an eigenvalue of the first kind.

Now let us prove that if $(L(k_+)z, z) = 0$, then $(L'(k_+)z, z) > 0$. Since $L(k_+) \geq 0$ we can write

$$\|L(k_+)z\|^2 \leq \|L(k_+)\|(L(k_+)z, z)$$

so z, k_+ is an eigenpair and it is of the first kind.

Let $\delta_+ < k < k_+$ and z be a corresponding vector such that $(L(k)z, z) = 0$. The vector z cannot be of second kind since from the condition $(L'(k)z, z) < 0$ follows that $z \in G$ and $k = p_-(z)$, contradicting the fact that $\delta_+ < k$. Assume that $(L'(k)z, z) = 0$, therefore $k = p_\pm(z)$. We consider an eigenvector x of the first kind corresponding to the eigenvalue k_+ and we set $z_\alpha = z + \alpha x$. Replacing x by $-x$, we can assume that $\operatorname{Re}(L(k)z, x) \leq 0$. We note that since $\delta_+ < k < k_+$, we have $(L(k)x, x) < 0$ and therefore

$$(L(k)z_\alpha, z_\alpha) = (L(k)z, z) + 2\alpha \operatorname{Re}(L(k)z, x) + \alpha^2(L(k)x, x) < 0,$$

if $\alpha > 0$. By the condition **(III)** we have $z_\alpha \in G$, $\alpha > 0$ and we obtain the contradiction

$$\delta_+ \geq \lim_{\alpha \to 0^+} p_-(z_\alpha) = p_-(z) = k$$

with the fact that $\delta_+ < k$. The case of $W'_{p_-} \cap [k_-, \delta_-)$ is analyzed in an analogous manner. □

Theorem 2.3. *Let L be an operator function satisfying the conditions* **(I)**–**(IV)**. *If $k \in W'_{p_+} \cap (k_+, k'_+]$ ($k \in W'_{p_-} \cap [k'_-, k_-)$), then k is a neutral eigenvalue.*

Proof. If $k \in W'_{p_+} \cap (k_+, k'_+]$, then there exist $x \in G'$ such that $(L(k)x, x) = 0$. Since $k > k_+$ we have $L(k) \geq 0$. Using the inequality

$$\|L(k)x\|^2 \leq \|L(k)\|\,(L(k)x, x)$$

we see that $L(k)x = 0$ and k, x is an eigenpair, since $x \in G' \setminus G$ it is a neutral eigenpair. □

As the example below shows the sets $W_{p_\pm}$ are not necessarily connected for every operator function L satisfying **(I)**–**(IV)**.

Example. Let $M(k) = A + kB + k^2C$ be a one parameter operator pencil, $H = \mathbb{C}^2$ and

$$A = \begin{pmatrix} 1 & 0 \\ 0 & 2 \end{pmatrix}, \quad B = \begin{pmatrix} 1 & 4 \\ 4 & 0 \end{pmatrix}, \quad C = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}.$$

The operator pencil M satisfies **(I)**–**(IV)** and the graphs of sets W_{p_-} and W_{p_+} (see Figure 3), obtained by computational methods, are disconnected.

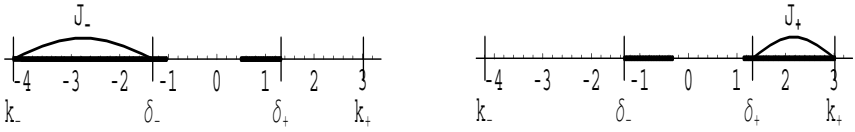

Figure 3. Graphics of W_{p_-} and W_{p_+}

Now we show that some parts of $W_{p_\pm}$ are connected. For this purpose we define the sets

$$\begin{aligned} G_+ &= \{x \in G \mid p_+(x) > \delta_+\} \\ G_- &= \{x \in G \mid p_-(x) < \delta_-\}. \end{aligned}$$

If G_+ (G_-) is nonempty then some part of $W_{p_\pm}$ turns out to be connected, as we can see in Figure 3.

We set $J_\pm = p_\pm(G_\pm)$. If the set G_+ (G_-) is empty, then we consider J_+ (J_-) is empty. In the case $G_\pm \neq \emptyset$ we have

$$\begin{aligned} J_+ &= W_{p_+} \cap (\delta_+, k_+] \\ J_- &= W_{p_-} \cap [k_-, \delta_-). \end{aligned}$$

Theorem 2.4. *Let L be an operator function satisfying* **(I)**–**(IV)**. *If H is a complex Hilbert space, then the sets $G_\pm$ are pathwise connected and $J_\pm$ are connected.*

Proof. Since the functionals $p_\pm$ are continuous on G, the connectedness of $J_\pm$ follows from pathwise connectedness of $G_\pm$. We show that G_+ is pathwise connected. Let $x, y \in G_+$, then $p_+(x), p_+(y) > \delta_+$ and $p_-(x), p_-(y) \leq \delta_+$. We select $\varepsilon > 0$ such that $p_+(x), p_+(y) > \delta_+ + \varepsilon$ and we set $k = \delta_+ + \varepsilon$. Then

$$k \in (p_-(x), p_+(x)) \cap (p_-(y), p_+(y)). \tag{3}$$

We select $\lambda = 1$ or -1 so that $\operatorname{Re}[\lambda (L(k)x, y)]$ is nonpositive, and we set $\tilde{x} = \lambda x$, $z_\alpha = \alpha\tilde{x} + (1-\alpha)y,\ \alpha \in [0,1]$. Then

$$(L(k)z_\alpha, z_\alpha) = \alpha^2(L(k)x, x) + 2\alpha(1-\alpha)\operatorname{Re}[\lambda(L(k)x, y)] + (1-\alpha)^2(L(k)y, y).$$

Note that if $x \in G$ from **(II)** and **(III)**,

$$(L(t)x, x) < 0 \Longleftrightarrow t \in (p_-(x), p_+(x)).$$

Taking $t = k$ from (3) we obtain that both $(L(k)x, x)$ and $(L(k)y, y)$ are negative. Therefore $(L(k)z_\alpha, z_\alpha)$ is negative for $\alpha \in [0,1]$. From this follows that $z_\alpha \in G$ and $k < p_+(z_\alpha)$. Since $k > \delta_+$ we have $z_\alpha \in G_+$ for $\alpha \in [0,1]$. Thus, $\tilde{x}$ can be joined with y in G_+ by a segment. Since H is a complex space, in the case $\lambda = -1$ the vectors x an $-x$ in G can be joined by $xe^{i\varphi},\ (0 \leq \varphi \leq \pi)$. For G_- the proof is similar. □

Now we show that the root zones, under some additional conditions, are not separated. An example was given above (see Figure 3). First we prove the following:

Lemma 2.2. *The set G is open and $\overline{G} = G'$.*

Proof. First we show that G^c is closed. Let $\{x_n\} \subset G^c$ and $x_n \to x$. Then for every $\alpha \in [a, b]$ we have

$$0 \leq (L(\alpha)x_n, x_n) \to (L(\alpha)x, x) \geq 0,$$

so $x \in G^c$ and G^c is closed.

Now let $\{x_n\} \subset G$ and $x_n \to x$. Then to every x_n corresponds an $\alpha_n \in [a, b]$ such that $(L(\alpha_n)x_n, x_n) < 0$. Since $\{\alpha_n\} \subset [a, b]$ is bounded, it has a convergent subsequence. Let us rename it again as $\{\alpha_n\}$ and let $\alpha_n \to \alpha$. Since

$$0 > (L(\alpha_n)x_n, x_n) \to (L(\alpha)x, x) \leq 0,$$

it follows that $x \in G'$ and $\overline{G} = G'$. □

Theorem 2.5. *Let L be an operator function satisfying the conditions* **(I)**–**(IV)**. *If $G \neq \emptyset$ and $G \neq H \setminus \{0\}$, then $\delta_- \leq \delta_+$.*

Proof. By the condition $G \neq \emptyset$ there exists $x_1 \in G$. On the other hand it follows from $G \neq H \setminus \{0\}$ that there exists $x_2 \notin G$. Define a path from x_1 to x_2 by $z_t = (1 - t)x_1 + tx_2$, $0 \leq t \leq 1$. Since $x_1 \in G$, $x_2 \notin G$ and G is open, there exists a number $t_* \in (0, 1]$ such that $z_t \in G$ for all $t \in [0, t_*)$. From $\overline{G} = G'$ follows that $z_{t_*} \in G'$. Now we can write

$$\begin{aligned} \delta_- &= \inf_{z \in G} p_+(z) \leq \lim_{t \to t_* - 0} p_+(z_t) = p_+(z_{t_*}) \\ &= p_-(z_{t_*}) = \lim_{t \to t_* - 0} p_-(z_t) \leq \sup_{z \in G} p_-(z) = \delta_+. \end{aligned}$$

□

3. Examples

Now we give some examples of classes of operator functions satisfying the conditions **(I)**–**(IV)**. Here, in the first two examples, we aim to show where our problem comes from. Note that, even if these two classes are studied extensively, some of the results (see Theorem 2.2) are new for these classes as well.

3.1. One parameter pencils of waveguide type

Definition 3.1 ([1, pp. 1278–1279]**).** An operator pencil of the form $L(k) = k^2C + kB + A$, where A, B and C are bounded and symmetric operators in a Hilbert space H is said to be an operator pencil of waveguide type if the following conditions are satisfied.

(A1) $C > 0$;
(A2) $A = A_1 - A_2$, $A_1 \gg 0$, $A_2 \in S_\infty$;
(A3) B and C are compact operators;
(A4) $G \neq \emptyset$, $G \neq H \setminus \{0\}$;

(A5) $-\infty < k'_-,\ k'_+ < \infty$.

We set $d(x) = (Bx, x)^2 - 4(Cx, x)(Ax, x)$. The sets G and G' are defined as

$$\begin{aligned} G' &= \{x \mid d(x) \geq 0\}, \\ G &= \{x \mid d(x) > 0\}, \end{aligned}$$

and the functionals $p_\pm(x)$ have the form

$$p_\pm(x) = \frac{-(Bx, x) + \sqrt{d(x)}}{2(Cx, x)}, \quad x \in G'.$$

Now we choose $[a, b] = [k'_-, k'_+]$ and the conditions **(I)**–**(IV)** follow from the conditions (A1)–(A3) [1, p. 1281].

3.2. Two parameter quadratic pencils of waveguide type

Definition 3.2 ([4, Definition 2.1]**).** An operator pencil of the form $L(k, w) = A + kB + k^2C - w^2I$ is called weak two parameter pencil of waveguide type if the following conditions are satisfied:

(B1) The operator A is nonnegative and $(A + I)^{-1} \in S_\infty$, where S_∞ is the set of compact operator.
(B2) C is a bounded and positive definite operator.
(B3) B is symmetric and $(A + I)^{-1/2}B(A + I)^{-1/2} \in S_\infty$.

Additionally, if the following condition:

(B4) $\exists\varepsilon$ satisfying $0 < \varepsilon < 1$ such that $(Au, u) + k(Bu, u) + \varepsilon^2 k^2 (Cu, u) \geq 0$, $\forall k \in \mathbb{R}$, $u \in D\left((A + I)^{1/2}\right)$ – the domain of the operator $(A + I)^{1/2}$ – called the energy stability condition is satisfied then we say that we have a weak operator pencil with the energy stability condition.

Here the coefficients A and B may be unbounded. We can transform the pencil $L(k, w)$ to the pencil $\tilde{L}(k, w) := (A + I)^{-1/2}L(k, w)(A + I)^{-1/2}$ which has bounded coefficients. Note that $\sigma(\tilde{L}) = \sigma(L)$. If we write

$$\tilde{L}(k, w) = k^2\tilde{C} + k\tilde{B} + \tilde{A}(w),$$

then

$$\tilde{A}(w) = I - (1 + w^2)(A + I)^{-1}, \quad \tilde{C} = (A + I)^{-1/2}C\,(A + I)^{-1/2},$$

$$\tilde{B} = (A + I)^{-1/2}C\,(A + I)^{-1/2}.$$

Now, for fixed $w \in \mathbb{R}$, let us check the conditions (A1)–(A3) and (A5).

(A1) By the condition (B2), $\tilde{C} > 0$.
(A2) $\tilde{A}(w) = I - (1 + w^2)(A + I)^{-1}$. $\tilde{A}_1 = I \gg 0$, $\tilde{A}_2 = (1 + w^2)(A + I)^{-1} \in S_\infty$ by (B1).
(A3) It follows from the conditions (B1)–(B3) that $\tilde{B}$ and $\tilde{C}$ are compact.

(A5) It follows from the conditions (B2) and (B4) that there exists a number $c_0 > 0$ such that for all $k \in \mathbb{R}$ and for all $u \in D\left((A+I)^{1/2}\right)$,

$$(Au,u) + k(Bu,u) + k^2(Cu,u) \geq c_0^2 k^2 (u,u).$$

For $v \in H$, $v \neq 0$ let $u = (A+I)^{-1/2}v$, then we have

$$([I-(A+I)^{-1}]v,v) + k(\tilde{B}v,v) + k^2(\tilde{C}v,v) \geq c_0^2 k^2((A+I)^{-1}v,v)$$

and

$$(\tilde{L}(k,w)v,v) \geq (c_0^2k^2 - w^2)((A+I)^{-1}v,v).$$

If k is a root of the equation $(\tilde{L}(k,w)v,v) = 0$, i.e, $k = p_-(v)$ or $k = p_+(v)$, then

$$0 \geq (c_0^2k^2 - w^2)((A+I)^{-1}v,v) = (c_0^2k^2 - w^2)\left\|(A+I)^{-1/2}v\right\|^2.$$

Consequently, we have $c_0^2k^2 \leq w^2$, hence $-\infty < k'_-,\ k'_+ < \infty$.

Now the conditions **(I)**-**(IV)** follow from Example 3.1.

3.3. Polynomial operator pencils of waveguide type

Definition 3.3. The two parameter operator pencil

$$L(k,w) := A + \sum_{s=1}^{n} k^{2s} C_{s-1} + \sum_{s=0}^{n-1} k^{2s+1} B_s + iwD - w^2 I$$

is said to be an operator pencil of waveguide type iff

(C1) A is a self-adjoint nonnegative operator satisfying $(A+I)^{-1} \in S_\infty$ and D is a symmetric operator which satisfies $D(A+I)^{-1/2} \in S_\infty$, where S_∞ is the set of compact operators,

(C2) C_{n-1} is a bounded and positive definite operator:

$$c_1(u,u) \leq (C_{n-1}u,u) \leq c_2(u,u), \qquad u \in H \text{ and } 0 < c_1 \leq c_2.$$

(C3) The operators B_s, $s = 0,1,\ldots,n-1$ and C_s, $s = 0,1,\ldots,n-2$ are symmetric and $(A+I)^{-1/2}B_s(A+I)^{-1/2} \in S_\infty$ and $(A+I)^{-1/2}C_s(A+I)^{-1/2} \in S_\infty$. Particularly, these conditions mean $D((A+I)^{1/2}) \subset D(B_s)$, $s = 0,1,\ldots,n-1$, and $D((A+I)^{1/2}) \subset D(C_s)$, $s = 0,1,\ldots,n-1$.

(C4) There exists a number $\mu \geq 0$ such that for all $k \in \mathbb{R}$ and $u \in D((A+I)^{1/2})$ the following inequality holds:

$$(Au,u) + \sum_{s=1}^{n} k^{2s}(C_{s-1}u,u) + \sum_{s=0}^{n-1} k^{2s+1}(B_su,u) \geq \mu^2(u,u).$$

In addition we say that a two parameter pencil of w.g.t. satisfies the energetic stability condition if the following condition is fulfilled:

(C5) There exist real numbers $\zeta \geq 0$ and $c_0 > 0$ such that for all $k \in \mathbb{R}$ and all $u \in D((A+I)^{1/2})$,

$$(Au,u) + \sum_{s=1}^{n} k^{2s}(C_{s-1}u,u) + \sum_{s=0}^{n-1} k^{2s+1}(B_s u,u) \geq ({c_0}^{2n}k^{2n} + \zeta)(u,u).$$

For this class, in the case $D = 0$, the conditions **(I)**–**(IV)** are fulfilled on some parts of the root domain.

References

[1] Yu. Abramov, Pencils of waveguide type and related extremal problems, *J. Soviet Math.* **64** no. 6 (1993), 1278–1289.

[2] N. Çolakoğlu, M. Hasanov, B. Ünalmış Uzun, Eigenvalues of two parameter polynomial operator pencils of waveguide type, *Integral Equations Operator Theory* **56** no. 3 (2006), 381–400.

[3] I. Gohberg, P. Lancaster, L. Rodman, *Matrix Polynomials,* Academic Press, New York (1982).

[4] M. Hasanov, On the spectrum of a weak class of operator pencils of waveguide type, *Math. Nachr.* **279** no. 8 (2006), 1–10.

[5] A.G. Kostyuchenko, M.B. Orazov, The problem of oscillations of an elastic half cylinder and related self-adjoint quadratic pencils, *J. Sov. Math.* **33** (1986), 1025–1065.

[6] A.S. Markus, *Introduction to the spectral theory of polynomial operator pencils,* Translations of Mathematical Monographs, Vol. 71, American Mathematical Society, Providence, RI (1988).

[7] A S. Silbergleit, Yu.I. Kopilevich, On properties of waves associated with quadratic operator pencils, *Dokl. Akad. Nauk SSSR* **256** no. 3 (1981), 565–570.

[8] A.S. Silbergleit, Yu.I. Kopilevich, On the dispersion curves of waveguide systems that are connected with quadratic operator pencils, *Dokl. Akad. Nauk SSSR* **259** no. 6 (1981), 1345–1349.

[9] A.S. Silbergleit, Yu.I. Kopilevich, *Spectral theory of guided waves,* Institute of Physics Publishing, Bristol (1996).

Nurhan Çolakoğlu
Istanbul Technical University
Department of Mathematics
34469, Maslak, Istanbul
Turkey
e-mail: `colakn@itu.edu.tr`

Operator Theory:
Advances and Applications, Vol. 179, 157–167

A Perturbative Analysis of the Reduction into Diagonal-plus-semiseparable Form of Symmetric Matrices

Dario Fasino

Abstract. It is known that any symmetric matrix can be transformed by an explicitly computable orthogonal transformation into diagonal-plus-semiseparable form, with prescribed diagonal term. In this paper, we present perturbation bounds for such transformations, under the condition that the diagonal term is close to (part of) the spectrum of the given matrix. As an application, we provide new iterative schemes for the simultaneous refinement of the eigenvalues of a symmetric matrix, having quadratic convergence.

Mathematics Subject Classification (2000). Primary 47A55; Secondary 65F15.

Keywords. Diagonal-plus-semiseparable matrices, perturbation analysis.

1. Introduction

In this paper, a *symmetric semiseparable matrix* is a real, symmetric $n \times n$ matrix S whose entries depend bilinearly on a set of $2n$ parameters, as follows:

$$S = \begin{pmatrix} u_1v_1 & u_2v_1 & \cdots & u_nv_1 \\ u_2v_1 & u_2v_2 & \ddots & u_nv_2 \\ \vdots & \ddots & \ddots & \vdots \\ u_nv_1 & u_nv_2 & \cdots & u_nv_n \end{pmatrix}. \tag{1.1}$$

The numbers u_i, v_j are arbitrary, and are called the *generators* of the matrix S. Actually, the above definition is referred to in the modern literature as *generator representable semiseparable matrix* [14], since the most general definition of a semiseparable matrix is given in terms of ranks of nondiagonal submatrices [3, 7].

Moreover, we call *diagonal-plus-semiseparable matrix* [4, 5, 8] (dpss, for short), any real, symmetric matrix A admitting a decomposition in the form $A = D + S$,

This work was partially supported by MIUR, grant number 2004015437 (PRIN 2004 fundings).

where S is as in (1.1) and $D = \mathrm{Diag}(d_1, \ldots, d_n)$ is any real, diagonal matrix. Although all the forthcoming discussions go almost unchanged in the complex Hermitian case, we prefer to stick with real symmetric matrices, just for notational simplicity.

Diagonal-plus-semiseparable matrices own interesting structural and computational properties, that make them a convenient tool for numerical linear algebra problems:

- numerically stable representations in $O(n)$ parameters [2, 14];
- fast algorithms for the solution of associated linear systems [5, 6, 7], computation of the characteristic polynomial [9], eigendecomposition [10], and basic factorizations (QR, LU) [3, 4, 12];
- structural invariance under (shifted) QR steps [8];
- relationships with orthogonal rational functions and rational Lanczos methods [8, 11];
- implicit-Q theorems and inverse eigenvalue problems [8, 9].

Many of the above-mentioned results admit certain generalizations to wider matrix classes, known as *rank structures*, recently found by authors including Bini, Chandrasekaran, Eidelman, Fiedler, Gemignani, Gohberg, Gu, Koltracht, Mastronardi, Olshevsky, Van Barel, Vandebril, Tyrtyshnikov (among others). The interested reader may consult the recent overview paper by Vandebril, Van Barel, Golub, Mastronardi [13], which contains a commented bibliography on 134 papers on the topic of semiseparable and rank-structured matrices.

Recently, Van Barel and co-authors found an $O(n^3)$ algorithm to reduce a generic symmetric matrix into dpss form via orthogonal transformations, $Q^T A Q = D + S$, where the diagonal term D can be prescribed in advance [15]. Extensive numerical experiments with that algorithm show that, if D is close to the exact spectrum of A, then S vanishes. Moreover, if D approximates only part of the spectrum of A, say,

$$|\lambda_i - d_i| \ll |\lambda_i - d_j|, \qquad i = 1, \ldots, k, \quad j \neq i,$$

where $\lambda_1, \ldots, \lambda_n$ are the eigenvalues of A, then S has an almost block diagonal structure:

$$S = \begin{pmatrix} S_{11} & S_{21}^T \\ S_{21} & S_{22} \end{pmatrix} \begin{matrix} \} k \\ \} n-k, \end{matrix}$$

where the submatrix S_{21} has a considerably small norm. This paper is basically motivated by these observations. Indeed, our aim is to give a rigorous explanation of these facts, on the basis of a perturbative analysis of the similarity reduction of a generic symmetric matrix into dpss form. After setting some basic notation and results, in Section 3 we develop a perturbative analysis of the matrices Q and S occurring in this reduction, under the condition that the diagonal term D approximates (part of) the spectrum of A. As an application of the forthcoming results, in Section 4 we devise a new numerical method for the simultaneous refinement of "good" initial approximations of the eigenvalues of A.

2. Notation and basic results

We will use the following notation: Let I_n denote the identity matrix of order n. Let $\lambda_1, \ldots, \lambda_n$ be the eigenvalues of the real symmetric $n \times n$ matrix A, and let $d_1, \ldots, d_n$ be pairwise distinct real numbers, such that all matrices $A - d_i I_n$ are nonsingular. Let $D = \mathrm{Diag}(d_1, \ldots, d_n)$, and let $v \in \mathbb{R}^n$ be arbitrary. Under these assumptions, the *rational Krylov matrix*

$$\mathcal{K}(A, v, D) = [\,(d_1 I_n - A)^{-1} v, \ldots, (d_n I_n - A)^{-1} v\,]$$

is well defined. We recall from [8] the following result, establishing necessary and sufficient conditions for nonsingularity of the rational Krylov matrix:

Lemma 2.1. *Let $A = U \Lambda U^T$ be the spectral decomposition of the symmetric matrix A, where $\Lambda = \mathrm{diag}(\lambda_1, \ldots, \lambda_n)$, and let $U^T v = w = (w_1, \ldots, w_n)^T$. Under the previously mentioned hypothesis on $d_1, \ldots, d_n$, the matrix $\mathcal{K}(A, v, D)$ is nonsingular if and only if $\lambda_i \neq \lambda_j$ for $i \neq j$ and all entries of w are nonzero.*

Proof. See [8, Lemma 1]. □

We will denote by $\mathcal{Q}(A, v, D)$ the orthogonal factor of the QR factorization of $\mathcal{K}(A, v, D)$, under a suitable condition ensuring continuity of $\mathcal{K}$ with respect to D (see later). This orthogonal factor is the key to define the orthogonal transformation of A into dpss form:

Theorem 2.2. *If the matrix $\mathcal{K}(A, v, D)$ is nonsingular, and $\mathcal{Q}(A, v, D) = Q$, then $Q^T A Q$ is a symmetric dpss matrix, $Q^T A Q = D + S$, with the diagonal matrix $D = \mathrm{diag}(d_1, \ldots, d_n)$.*

Proof. See [8, Thm. 1]. □

Remark 2.3. By construction, the first column of $\mathcal{Q}(A, v, D)$ is parallel to the one of $\mathcal{K}(A, v, D)$, whence

$$\mathcal{Q}(A, v, D) e_1 = \frac{q}{\|q\|}, \qquad q = (d_1 I - A)^{-1} v.$$

Moreover, we have also the following theorem of "Implicit-Q" type, stating that, under minor assumptions, there is a one-to-one correspondence between transformation into dpss form and the first column of $\mathcal{Q}(A, v, D)$, whence the vector v, by virtue of the previous remark:

Theorem 2.4. *Suppose that Q_1 and Q_2 are orthogonal matrices such that $Q_1^T A Q_1 = M_1$ and $Q_2^T A Q_2 = M_2$ are symmetric dpss matrices having the same diagonal term, that is, $M_1 = D + S_1$ and $M_2 = D + S_2$, with S_1 and S_2 semiseparable. Furthermore, suppose that $Q_1 e_1 = Q_2 e_1$. Then there exists a diagonal matrix $\Delta = \mathrm{Diag}(\pm 1, \ldots, \pm 1)$ such that $Q_2 = Q_1 \Delta$ and $M_1 = \Delta M_2 \Delta$.*

Proof. See [8, Thm. 2]. □

As a consequence, we can introduce another matrix-valued operator, whose value is the semiseparable part of the dpss form of A defined by the diagonal matrix D and the vector v, as in Theorem 2.2:

$$\mathcal{S}(A, v, D) = \mathcal{Q}(A, v, D)^T A \mathcal{Q}(A, v, D) - D. \tag{2.1}$$

Remark 2.5. The above-defined operators own the following invariance properties, whose proof is elementary:

1. $\mathcal{K}(UAU^T, Uv, D) = U\mathcal{K}(A, v, D)$
2. $\mathcal{Q}(UAU^T, Uv, D) = U\mathcal{Q}(A, v, D)$
3. $\mathcal{S}(UAU^T, Uv, D) = \mathcal{S}(A, v, D)$.

On the basis of the previous remark, in the analysis of these operators we can restrict ourselves to the case where $A = \Lambda$ is diagonal. In this way, we can identify the vector v appearing in the definition of $\mathcal{K}$, $\mathcal{Q}$, $\mathcal{S}$, with the vector w in the hypotheses of Lemma 2.1. From here on, we assume existence and uniqueness of all preceding matrices (i.e., the operators $\mathcal{K}$, $\mathcal{Q}$, $\mathcal{S}$, are well defined). In particular, the matrix $\mathcal{K}(\Lambda, w, D)$ is assumed nonsingular. Necessary and sufficient conditions ensuring this are those given in Lemma 2.1.

We close this section by recalling a perturbation bound for the orthogonal factor in the QR decomposition; here and in what follows, $\|\cdot\|_F$ denotes the Frobenius matrix norm.

Theorem 2.6. *Let $A = QR$ and $\tilde{A} = \tilde{Q}\tilde{R}$. Then, up to first order,*

$$\|Q - \tilde{Q}\|_F \lesssim \sqrt{2}\|A^{-1}\|\,\|A - \tilde{A}\|_F.$$

Proof. See [1, Thm. 4.2]. □

3. New Results

This section contains the main results of this paper. Assuming that the diagonal matrix D is "close" to the spectrum of A, in the following two subsections we consider individually the perturbative analysis of the matrices Q and S occurring in the transformation of A into dpss form. In the last subsection, we address the case where D approximates only part of the spectrum of A.

3.1. Perturbative analysis of Q

Theorem 3.1. *Under the previous notation and hypotheses, let $Q_\varepsilon = \mathcal{Q}(\Lambda, w, \Lambda + \varepsilon\Delta)$, with $\Delta = \mathrm{Diag}(\delta_1, \ldots, \delta_n)$, and let $E \equiv (e_{ij})$ where*

$$e_{ij} = \begin{cases} \frac{w_i}{w_j}\frac{\delta_j}{\lambda_j - \lambda_i} & i \neq j \\ 0 & i = j. \end{cases}$$

Up to first order terms in ε,

$$\|Q_\varepsilon - I_n\|_F \lesssim \sqrt{2}|\varepsilon|\|E\|_F.$$

In particular, $\lim_{\varepsilon \to 0} Q_\varepsilon = I_n$, independently on w.

Proof. By definition, Q_ε is the orthogonal factor of the rational Krylov matrix $K_\varepsilon = [(d_1 I_n - \Lambda)^{-1} w, \ldots, (d_n I_n - \Lambda)^{-1} w]$, where $d_i = \lambda_i + \varepsilon\delta_i$. Introduce the matrix $Z = \mathrm{Diag}(\varepsilon\delta_1/w_1, \ldots, \varepsilon\delta_n/w_n)$. For the (i,j)-th entry of the matrix $K_\varepsilon Z$ we obtain

$$(K_\varepsilon Z)_{i,j} = \frac{w_i}{w_j} \frac{\varepsilon\delta_j}{(\lambda_j + \varepsilon\delta_j - \lambda_i)} = \begin{cases} 1 & i = j, \\ \frac{w_i}{w_j} \frac{\varepsilon\delta_j}{\lambda_j - \lambda_i} + O(\varepsilon^2) & i \neq j. \end{cases}$$

Hence, we have the following expansion in powers of ε:

$$K_\varepsilon Z = I_n + \varepsilon E + O(\varepsilon^2).$$

The effect of the matrix Z is to scale the columns of K_ε, hence it does not affect its orthogonal factor. In other words, Q_ε is the orthogonal factor of the matrix $I_n + \varepsilon E + O(\varepsilon^2)$. In order to complete the proof, it is sufficient to apply Theorem 2.6 with $A = I_n$ and $\tilde{A} = I_n + \varepsilon E + O(\varepsilon^2)$. □

Owing to the formulas for the entries of the matrix E, we can supplement the foregoing theorem with the (rather crude) bound

$$\|Q_\varepsilon - I_n\|_F \lesssim \frac{\max_i |w_i|}{\min_i |w_i|} \frac{\sqrt{2(n-1)}}{\min_{i \neq j} |\lambda_i - \lambda_j|} \|\varepsilon\Delta\|_F.$$

One obvious extension of the preceding result is the following: If there exists a permutation π of the integers $1, \ldots, n$ such that $|d_i - \lambda_{\pi(i)}| = O(\varepsilon)$ then $\lim_{\varepsilon \to 0} Q_\varepsilon = P$, where P is the matrix representation of π. A neater statement is the following:

Corollary 3.2. *Let $P \in \mathbb{R}^{n \times n}$ be a permutation matrix, and let*

$$Q_\varepsilon = \mathcal{Q}(\Lambda, w, P^T \Lambda P + \varepsilon\Delta).$$

Then, $\lim_{\varepsilon \to 0} Q_\varepsilon = P$, independently on Δ.

Proof. Let $\tilde{\Lambda} = P^T \Lambda P$, $\tilde{w} = P^T w$, and $\tilde{Q}_\varepsilon = \mathcal{Q}(\tilde{\Lambda}, \tilde{w}, \tilde{\Lambda} + \varepsilon\Delta)$. Then, by the property of the operator $\mathcal{Q}$ mentioned in Remark 2.5, we have $Q_\varepsilon = P\tilde{Q}_\varepsilon$. Theorem 3.1 gives us $\|\tilde{Q}_\varepsilon - I_n\|_F = O(\varepsilon)$, and the claim follows. □

3.2. Perturbative analysis of S

Let $S_\varepsilon = \mathcal{S}(\Lambda, w, \Lambda + \varepsilon\Delta)$. Using Theorem 3.1, for $\varepsilon \to 0$ we have $Q_\varepsilon \to I_n$. Hence from (2.1) we obtain

$$\varepsilon \to 0 \quad \Longrightarrow \quad S_\varepsilon = Q_\varepsilon^T \Lambda Q_\varepsilon - (\Lambda + \varepsilon\Delta) \to O. \tag{3.1}$$

More precisely, we can prove the following result:

Theorem 3.3. *Under the previous notation and hypotheses, we have*

$$\mathcal{S}(\Lambda, w, \Lambda + \varepsilon\Delta) = \varepsilon\hat{S} + O(\varepsilon^2),$$

where $\hat{S}$ is the semiseparable matrix given by

$$\hat{S} = \begin{pmatrix} u_1 v_1 & \cdots & u_n v_1 \\ \vdots & \ddots & \vdots \\ u_n v_1 & \cdots & u_n v_n \end{pmatrix}, \quad u_i = -w_i, \quad v_j = \frac{\delta_j}{w_j}.$$

Proof. For $\varepsilon \neq 0$, the matrix $\varepsilon^{-1}\mathcal{S}(\Lambda, w, \Lambda + \varepsilon\Delta)$ is semiseparable. It is known that the closure (in any norm-induced metric) of the set of semiseparable matrices is made of all block diagonal matrices with semiseparable blocks, see [14]. Nevertheless, we set

$$\hat{S} = \lim_{\varepsilon \to 0} \frac{1}{\varepsilon} \mathcal{S}(\Lambda, w, \Lambda + \varepsilon\Delta)$$

and we look for an expression of $\hat{S}$ having the form (1.1). Having found it, the claim will follow by uniqueness of the limit.

Under our assumptions, QR factors are differentiable [1], whence

$$Q_\varepsilon = I_n + \varepsilon X + O(\varepsilon^2), \tag{3.2}$$

where $X = -X^T$. In fact, the Lie algebra of the Lie group of orthogonal matrices is the set of real skewsymmetric matrices [1]. Hence, neglecting $O(\varepsilon^2)$ terms, for the diagonal entries of S_ε we have:

$$\begin{aligned} e_i^T S_\varepsilon e_i &= e_i^T \left(Q_\varepsilon^T \Lambda Q_\varepsilon - (\Lambda + \varepsilon\Delta) \right) e_i \\ &\approx e_i^T \left(I_n - \varepsilon X \right) \Lambda \left(I_n + \varepsilon X \right) e_i - (\lambda_i + \varepsilon\delta_i) \\ &\approx e_i^T \Lambda e_i + \varepsilon e_i^T \left(\Lambda X - X\Lambda \right) e_i - (\lambda_i + \varepsilon\delta_i) \\ &= -\varepsilon\delta_i. \end{aligned}$$

As a consequence, $u_i v_i = \hat{S}_{i,i} = -\delta_i$, for $i = 1, \ldots, n$. In order to complete the description of $\hat{S}$, it is sufficient to compute its first column.

By Theorem 2.4, $Q_\varepsilon e_1$ characterizes $Q_\varepsilon = \mathcal{Q}(\Lambda, w, \Lambda + \varepsilon\Delta)$. Owing to Remark 2.3, $Q_\varepsilon e_1$ is proportional to $K_\varepsilon e_1$, and the latter is proportional to

$$\begin{pmatrix} 1 \\ 0 \\ \vdots \\ 0 \end{pmatrix} + \varepsilon \begin{pmatrix} 0 \\ \frac{w_2}{w_1} \frac{\delta_1}{\lambda_1 - \lambda_2} \\ \vdots \\ \frac{w_n}{w_1} \frac{\delta_1}{\lambda_1 - \lambda_n} \end{pmatrix} + O(\varepsilon^2).$$

By (3.2), this gives us the first column of X:

$$Xe_1 = \left(0, \frac{w_2}{w_1} \frac{\delta_1}{\lambda_1 - \lambda_2}, \ldots, \frac{w_n}{w_1} \frac{\delta_1}{\lambda_1 - \lambda_n} \right)^T.$$

Using the preceding equation, the task of characterizing the first column of $\hat{S}$ is accomplished as follows:

$$\begin{aligned} \hat{S}e_1 &= \lim_{\varepsilon \to 0} \frac{1}{\varepsilon} \left[Q_\varepsilon^T \Lambda Q_\varepsilon - (\Lambda + \varepsilon\Delta) \right] e_1 \\ &= \left(\Lambda X - X\Lambda \right) e_1 - \Delta e_1 \\ &= \left(\Lambda - \lambda_1 I_n \right) X e_1 - \delta_1 e_1 \\ &= -\left(\delta_1, \delta_1 \frac{w_2}{w_1}, \ldots, \delta_1 \frac{w_n}{w_1} \right)^T = -\frac{\delta_1}{w_1} w. \end{aligned}$$

Thus in (1.1) we have $u_i v_1 = u_i v_i w_i / w_1$, for $i = 1, \ldots, n$. Recall that all entries of w are different from zero, by hypothesis. Letting $u_i = -w_i$ and $v_i = \delta_i / w_i$ all the preceding equalities are fulfilled, and the proof is over. □

Corollary 3.4. *Let $w^+ = (1/w_1, \ldots, 1/w_n)$ denote the Moore-Penrose inverse of the vector $w = (w_1, \ldots, w_n)^T$. For the matrix $\hat{S}$ in Theorem* 3.3 *we have*

$$\|\hat{S}\|_F \leq \sqrt{2}\|w\|\|w^+\| \max_{1 \leq i \leq n} |\delta_i|.$$

Proof. Consider the rank-1 matrix $M = -ww^+\Delta$. The lower triangular part of $\hat{S}$ coincides with that of M, hence it is not difficult to realize that $\|\hat{S}\|_F \leq \sqrt{2}\|M\|_F$. The proof is completed by the inequality $\|M\|_F = \|w\|\|w^+\Delta\| \leq \|w\|\|w^+\|\|\Delta\|$. □

A straightforward consequence of Theorem 3.3 is the following:

Corollary 3.5. *For the diagonal part of the matrix $\mathcal{S}(\Lambda, w, \Lambda + \varepsilon\Delta)$ we have*

$$\mathrm{Diag}\left(\mathcal{S}(\Lambda, w, \Lambda + \varepsilon\Delta)\right) = -\varepsilon\Delta + O(\varepsilon^2),$$

independently of w.

3.3. Partial spectral approximation

Now we consider the case where D, Λ are partitioned consistently as

$$D = \begin{pmatrix} D_1 & O \\ O & D_2 \end{pmatrix}, \qquad \Lambda = \begin{pmatrix} \Lambda_1 & O \\ O & \Lambda_2 \end{pmatrix}, \tag{3.3}$$

where D_1, Λ_1 are $k \times k$, and $\Delta_1 = D_1 - \Lambda_1$ has "small" norm. Thus, we are considering here the case where D approximates only part of the entries of Λ. In this case, usually the matrix $\mathcal{S}(\Lambda, w, D)$ shows a small norm submatrix in its lower left corner, see [15]. In this subsection we will provide an explanation of this fact.

Let us define

$$K = \mathcal{K}(\Lambda, w, D)Z, \quad Q = \mathcal{Q}(\Lambda, w, D), \quad S = \mathcal{S}(\Lambda, w, D),$$

with the diagonal matrix $Z = \mathrm{Diag}(z_1, \ldots, z_n)$,

$$z_i = \begin{cases} \frac{d_i - \lambda_i}{w_i} & 1 \leq i \leq k \\ 1 & k+1 \leq i \leq n, \end{cases}$$

being chosen so that the first k diagonal entries of K are all ones:

$$K = \begin{pmatrix} I_k & K_{12} \\ O & K_{22} \end{pmatrix} + \begin{pmatrix} E_1 & O \\ E_2 & O \end{pmatrix} = K_0 + E.$$

Actually, the scaling operated by Z is analogous to the one exploited in the proof of Theorem 3.1. In fact, letting $E_1 \equiv (e_{ij}^{(1)})$ and $E_2 \equiv (e_{ij}^{(2)})$, we have the following formulas:

$$e_{ij}^{(1)} = \begin{cases} \frac{w_i}{w_j}\frac{d_j - \lambda_j}{d_j - \lambda_i} & i \neq j \\ 0 & i = j, \end{cases} \qquad e_{\ell j}^{(2)} = \frac{w_{k+\ell}}{w_j}\frac{d_j - \lambda_j}{d_j - \lambda_{k+\ell}}, \tag{3.4}$$

for $i,j=1,\dots,k$ and $\ell=1,\dots,n-k$. We partition the above matrices as $K\equiv(K_{ij})_{i,j=1,2}$ and so on, consistently with (3.3). Remark that K,Q,S are actually functions of Δ_1. For simplicity of notation, we refrain from indicating explicitly this dependence. By the way, the column scaling of K, which is irrelevant to Q and S, is introduced in order to obtain by continuity the equations

$$\lim_{\Delta_1\to 0} K = K_0, \quad \lim_{\Delta_1\to 0} Q = \begin{pmatrix} I_k & O \\ O & \hat{Q}_{22} \end{pmatrix}, \quad \lim_{\Delta_1\to 0} S = \begin{pmatrix} O & O \\ O & \hat{S}_{22} \end{pmatrix}, \tag{3.5}$$

with $\hat{Q}_{22}=\mathcal{Q}(\Lambda_2,(w_{k+1},\dots,w_n)^T,D_2)$ and $\hat{S}_{22}=\mathcal{S}(\Lambda_2,(w_{k+1},\dots,w_n)^T,D_2)$. Consider the rectangular QR factorization of the first k columns of K:

$$\begin{pmatrix} I_k+E_1 \\ E_2 \end{pmatrix} = \begin{pmatrix} Q_{11} \\ Q_{21} \end{pmatrix} \tilde{R} = \left[\begin{pmatrix} I_k \\ O \end{pmatrix} + \begin{pmatrix} X_1 \\ X_2 \end{pmatrix} \right] (I_k+Y),$$

where Y is upper triangular. Neglecting higher order terms, we can write

$$\begin{pmatrix} Q_{11} \\ Q_{21} \end{pmatrix} \tilde{R} \approx \left[\begin{pmatrix} I_k \\ O \end{pmatrix} + \begin{pmatrix} X_1+Y \\ X_2 \end{pmatrix} \right],$$

whence $Q_{21}=X_2\approx E_2$.

Now, consider the first k columns of the equation $Q(D+S)=\Lambda Q$, see (2.1). We have:

$$\begin{pmatrix} Q_{11} & Q_{12} \\ Q_{21} & Q_{22} \end{pmatrix} \left[\begin{pmatrix} D_1 \\ O \end{pmatrix} + \begin{pmatrix} S_{11} \\ S_{21} \end{pmatrix} \right] = \begin{pmatrix} \Lambda_1 Q_{11} \\ \Lambda_2 Q_{21} \end{pmatrix}. \tag{3.6}$$

Due to the limiting relations (3.5), we can rewrite the last $n-k$ rows of the left-hand side of (3.6) up to first-order terms in $\|E\|_F$ as

$$Q_{21}(D_1+S_{11})+Q_{22}S_{21} \approx Q_{21}D_1+\hat{Q}_{22}S_{21}.$$

By equating the last expression with the corresponding term of the right-hand side of (3.6), we obtain

$$\hat{Q}_{22}S_{21} \approx \Lambda_2 E_2 - E_2 D_1.$$

Since $\hat{Q}_{22}$ is orthogonal, we have $\|\hat{Q}_{22}S_{21}\|_F=\|S_{21}\|_F$. On the other hand, from the expression of the entries of E_2 found in (3.4) we see that

$$\Lambda_2 E_2 - E_2 D_1 \equiv \left(\frac{w_{k+i}}{w_j}(\lambda_j-d_j) \right)_{1\le i\le n-k,\, 1\le j\le k}$$

is a rank-1 matrix:

$$\Lambda_2 E_2 - E_2 D_1 = w_L w_H^+ (D_1-\Lambda_1),$$

where $w_L=(w_{k+1},\dots,w_n)^T$ and $w_H^+=(1/w_1,\dots,1/w_k)$ is the Moore-Penrose inverse of $w_H=(w_1,\dots,w_k)^T$. The Frobenius norm of a rank-1 matrix is exactly the product of the norms of its defining vectors. Hence, we arrive at the following result:

Theorem 3.6. *Let $S = \mathcal{S}(\Lambda, w, D)$. Under the hypotheses stated at the beginning of this section, we have:*

$$\|S_{21}\|_F \approx \|w_L\|\|w_H^+(D_1 - \Lambda_1)\| \leq \|w_L\|\|w_H^+\| \max_{1\leq i\leq k} |d_i - \lambda_i|,$$

where $w_L = (w_{k+1}, \ldots, w_n)^T$ and $w_H^+ = (1/w_1, \ldots, 1/w_k)$.

4. Simultaneous eigenvalue refinement

The results in the previous sections allow us to devise a possible numerical scheme to compute the eigenvalues of a symmetric matrix, starting from the knowledge of good initial approximations. Let A be a symmetric matrix, and let D be a diagonal matrix whose diagonal entries are "close" to the exact eigenvalues of A. Recall that the expression $\mathcal{S}(A, v, D) + D$ stands for the dpss matrix that is similar to A, has D as diagonal term, and whose transforming orthogonal matrix is the one whose first column is given in Remark 2.3.

We adopt the shortcut $\mathcal{S}^*(A, D)$ to denote *any* particular matrix $\mathcal{S}(A, v, D)$, under the sole hypotheses that $v = v(A, D)$ fulfills the hypotheses of Lemma 2.1 and the resulting map $(A, D) \mapsto \mathcal{S}^*(A, D) = \mathcal{S}(A, v(A, D), D)$ is sufficiently smooth. One such matrix can be computed in $O(n^3)$ operations by means of the previously mentioned algorithm in [15]. Indeed, for this algorithm, one obtains from [15, Thm. 6] that there exists a suitable polynomial $\pi(\lambda)$, whose coefficients depend polynomially on the entries of D, such that $v(A, D) = \pi(A)e_n$. Furthermore, the resulting semiseparable matrix $\mathcal{S}(A, v(A, D), D)$ is unreduced exactly when the hypotheses of Lemma 2.1 are met, see [15, Sect. 3.1] (when they are not met, $\mathcal{S}(A, v(A, D), D)$ splits into the direct sum of smaller semiseparable matrices, and one can deflate the eigenvalue problem for A into smaller subproblems). Finally, the smoothness of $(A, D) \mapsto \mathcal{S}(A, v(A, D), D)$ follows by the ones of the maps $(A, D) \mapsto \mathcal{K}(A, v(A, D), D)$ and $(A, D) \mapsto \mathcal{Q}(A, v(A, D), D)$.

Theorem 4.1. *Assume that the map $(A, D) \mapsto \mathcal{S}^*(A, D)$ fulfills the previously mentioned well-posedness and smoothness hypotheses. Let $A_0 = A$, and let D_0 be an arbitrary diagonal matrix. Consider the sequence $\{D_i\}$ of diagonal matrices generated by the iteration $D_i = \mathrm{Diag}(A_i)$, where A_i is defined according to one of the two following equations:*

$$A_i = \mathcal{S}^*(A_0, D_{i-1}) + D_{i-1}, \quad \text{or} \tag{4.1}$$

$$A_i = \mathcal{S}^*(A_{i-1}, D_{i-1}) + D_{i-1}. \tag{4.2}$$

Then, the sequence $\{D_i\}$ locally converges to the diagonal matrix Λ of eigenvalues of A. Moreover, this convergence is quadratic.

Proof. Firstly, observe that for $i > 0$ all matrices A_i are dpss and similar to A. Denote by $\mathbb{D}_n$ the set of all $n \times n$ diagonal matrices. The computation of the

eigenvalues of A can be accomplished by the solution of the fixed point problem $\Lambda = \Phi(\Lambda)$ in the unknown $\Lambda \in \mathbb{D}_n$, where

$$\Phi : \mathbb{D}_n \mapsto \mathbb{D}_n, \qquad \Phi(\Lambda) = \mathrm{Diag}(\mathcal{S}^*(A, \Lambda)) + \Lambda.$$

In fact, as shown in Equation (3.1), if Λ is a diagonal matrix containing the eigenvalues of A, then $\mathcal{S}^*(A, \Lambda) = O$, hence Λ is a fixed point of Φ.

By the assumed smoothness of the map $(A, D) \mapsto \mathcal{S}^*(A, D)$ we have that also Φ is sufficiently smooth. Moreover, the Fréchet derivative $\Phi'(\Lambda)$ is the zero operator. Indeed, consider an arbitrary $\Delta \in \mathbb{D}_n$. From Corollary 3.5 and the third part of Remark 2.5 we have:

$$\begin{aligned}
\Phi'(\Lambda)\Delta &= \lim_{\varepsilon\to 0} \frac{1}{\varepsilon}\Big(\Phi(\Lambda + \varepsilon\Delta) - \Phi(\Lambda)\Big) \\
&= \lim_{\varepsilon\to 0} \frac{1}{\varepsilon}\Big(\mathrm{Diag}(\mathcal{S}^*(A, \Lambda + \varepsilon\Delta)) - \mathrm{Diag}(\mathcal{S}^*(A, \Lambda)) + \varepsilon\Delta\Big) \\
&= \lim_{\varepsilon\to 0} \frac{1}{\varepsilon}\Big(- \varepsilon\Delta + \varepsilon\Delta + O(\varepsilon^2)\Big) = O.
\end{aligned}$$

Hence $\Phi'(\Lambda)$ vanishes. Now, equation (4.1) can be restated as $D_i = \Phi(D_{i-1})$. Thus, there exists a constant C such that

$$\|D_i - \Lambda\| = \|\Phi(D_{i-1}) - \Phi(\Lambda)\| \le C\|D_{i-1} - \Lambda\|^2 + o(\|D_{i-1} - \Lambda\|^2),$$

and we have the claim. Iteration (4.2) can be restated in a suitable product space as

$$\begin{pmatrix} A_i \\ D_i \end{pmatrix} = \hat{\Phi}\begin{pmatrix} A_{i-1} \\ D_{i-1} \end{pmatrix} = \begin{pmatrix} \mathcal{S}^*(A_{i-1}, D_{i-1}) + D_{i-1} \\ \mathrm{Diag}(\mathcal{S}^*(A_{i-1}, D_{i-1})) + D_{i-1} \end{pmatrix},$$

and the corresponding claim follows by analogous arguments. □

A numerical method based on the previous theorem could work as follows: After reducing the starting matrix to dpss form, using the initial eigenvalue approximations as diagonal term, the method constructs a sequence of dpss matrices, each step using the diagonal part of the previous step as diagonal term. The iteration proceeds possibly using a deflation technique based on Theorem 3.6, until all eigenvalues are resolved to a prescribed accuracy.

By the way, the resulting algorithm would not be competitive with the existing techniques for solving symmetric eigenvalue problems, because of the cubic cost of each iteration, if the algorithm in [15] is used as computational core. Nevertheless, iteration (4.2) requires only dpss matrices, hence in principle one can argue that fast algorithms could be devised for such task, acting only on generators. This remark prompts the following open question:

Is it possible to compute $\mathcal{S}^*(A, D)$ in $O(n^2)$ operations, when A is dpss but having a diagonal term different from D?

Acknowledgement

The author thanks Yuli Eidelman for a useful discussion on a preliminary version of this paper.

References

[1] R. Bhatia, *Matrix factorizations and their perturbations.* Linear Algebra and its Applications **197** (1994), 245–276.

[2] S. Delvaux, M. Van Barel, *A Givens-weight representation for rank structured matrices.* Report TW 453, Dept. of Computer Science, K. U. Leuven, Leuven, Belgium, March 2006.

[3] P. Dewilde, A.-J. van der Veen, *Time-varying systems and computations.* Kluwer, 1998.

[4] Y. Eidelman, I. Gohberg, *A look-ahead block Schur algorithm for diagonal plus semiseparable matrices.* Computers and Mathematics with Applications **35** (1997), 25–34.

[5] Y. Eidelman, I. Gohberg, *Fast inversion algorithms for diagonal plus semiseparable matrices.* Integr. Equ. Oper. Theory **27** (1997), 165–183.

[6] Y. Eidelman, I. Gohberg, *Linear complexity inversion algorithms for a class of structured matrices.* Integral Equations Operator Theory **35** (1999), 28–52.

[7] Y. Eidelman, I. Gohberg, *On a new class of structured matrices.* Integr. Equ. Oper. Theory **34** (1999), 293–324.

[8] D. Fasino, *Rational Krylov matrices and QR steps on Hermitian diagonal-plus-semiseparable matrices.* Numer. Linear Algebra Appl. **12** (2005), 743–754.

[9] D. Fasino, L. Gemignani, *Direct and inverse eigenvalue problems for diagonal-plus-semiseparable matrices.* Numerical Algorithms **34** (2003), 313–324.

[10] N. Mastronardi, M. Van Barel, E. Van Camp, *Divide and conquer algorithms for computing the eigendecomposition of symmetric diagonal-plus-semiseparable matrices.* Numerical Algorithms **39** (2005), 379–398.

[11] M. Van Barel, D. Fasino, L. Gemignani, N. Mastronardi, *Orthogonal rational functions and structured matrices.* SIAM J. Matrix Anal. Appl. **26** (2005), 810–829.

[12] E. Van Camp, N. Mastronardi, M. Van Barel, *Two fast algorithms for solving diagonal-plus-semiseparable systems.* Journal of Computational and Applied Mathematics **164-165** (2004), 731–747.

[13] R. Vandebril, M. Van Barel, G. Golub, N. Mastronardi, *A bibliography on semiseparable matrices.* Calcolo **42** (2005), 249–270.

[14] R. Vandebril, M. Van Barel, N. Mastronardi, *A note on the representation and definition of semiseparable matrices.* Numer. Linear Algebra Appl. **12** (2005), 839–858.

[15] R. Vandebril, E. Van Camp, M. Van Barel, N. Mastronardi, *Orthogonal similarity transformation of a symmetric matrix into a diagonal-plus-semiseparable one with free choice of the diagonal.* Numer. Math. **102** (2006), 709–726.

Dario Fasino
Dipartimento di Matematica e Informatica
Università di Udine
Via delle Scienze, 208
33100 Udine
Italy
e-mail: `fasino@dimi.uniud.it`

Operator Theory:
Advances and Applications, Vol. 179, 169–183

The Eigenstructure of Complex Symmetric Operators

Stephan Ramon Garcia

Abstract. We discuss several algebraic and analytic aspects of the eigenstructure (i.e., eigenvalues, eigenvectors, and generalized eigenvectors) of complex symmetric operators. In particular, we examine the relationship between the bilinear form $[x, y] = \langle x, Cy\rangle$ induced by a conjugation C on a complex Hilbert space $\mathcal{H}$ and the eigenstructure of a bounded linear operator $T : \mathcal{H} \to \mathcal{H}$ which is C-symmetric ($T = CT^*C$).

Mathematics Subject Classification (2000). 47A05, 47A07, 47A15.

Keywords. Complex symmetric operator, bilinear form, Toeplitz matrix, Hankel operator, Riesz idempotent, Riesz basis, generalized eigenvectors.

1. Introduction

In this note, we discuss several algebraic and analytic aspects of the eigenstructure (i.e., eigenvalues, eigenvectors, and generalized eigenvectors) of complex symmetric operators, a particular class of Hilbert space operators discussed in [3, 5, 6, 7]. Before proceeding, let us recall a few definitions.

A *conjugation* on a complex Hilbert space $\mathcal{H}$ is an antilinear operator $C : \mathcal{H} \to \mathcal{H}$ that is involutive ($C^2 = I$) and *isometric*, meaning that $\langle x, y\rangle = \langle Cy, Cx\rangle$ for all x, y in $\mathcal{H}$ (we assume that $\mathcal{H}$ is separable and that our operators are bounded, unless otherwise stated). We say that a linear operator $T : \mathcal{H} \to \mathcal{H}$ is *C-symmetric* if $T = CT^*C$ and *complex symmetric* if it is C-symmetric with respect to some conjugation C. For a fixed conjugation C, there exists an orthonormal basis $(e_n)_{n=1}^{\dim \mathcal{H}}$ of $\mathcal{H}$ such that $Ce_n = e_n$ for all n [6, Lem. 1]. We refer to such a basis as a *C-real* orthonormal basis and note that the matrix representation of a C-symmetric operator with respect to such a basis is symmetric (see [6, Prop. 2] or [5, Sect. 2.4]). In particular, an operator T is complex symmetric if and only if it is unitarily

Work partially supported by National Science Foundation Grant DMS-0638789.

equivalent to a symmetric matrix with complex entries, considered as an operator on an l^2 space of the appropriate dimension.

The class of complex symmetric operators includes all normal operators, operators defined by (finite or infinite) Hankel matrices, compressed Toeplitz operators (including the compressed shift), and the Volterra integration operator (see [3, 5, 6, 7]). Since we are more concerned here with eigenvectors and generalized eigenvectors of operators rather than with the operators which produce them, we could certainly consider *unbounded* complex symmetric operators as well (see [6, 7, 10] for details and references).

When dealing with C-symmetric operators, it turns out that the *bilinear* form

$$[x, y] = \langle x, Cy\rangle \tag{1}$$

induced by C is almost as important as the standard sesquilinear form $\langle \cdot\,, \cdot \rangle$. We will say that two vectors x and y are *C-orthogonal* if $[x, y] = 0$ (denoted by $x \perp_C y$). We shall also say that two subspaces $\mathcal{E}_1$ and $\mathcal{E}_2$ are C-orthogonal (denoted $\mathcal{E}_1 \perp_C \mathcal{E}_2$) if $[x_1, x_2] = 0$ for every x_1 in $\mathcal{E}_1$ and x_2 in $\mathcal{E}_2$. It is not hard to see that the bilinear form (1) is nondegenerate, in the sense that $[x, y] = 0$ for all y in $\mathcal{H}$ if and only if $x = 0$. Unlike the sesquilinear form $\langle \cdot\,, \cdot \rangle$, however, the bilinear form $[\,\cdot\,, \cdot\,]$ is not positive since $[e^{i\theta/2}x, e^{i\theta/2}x] = e^{i\theta}[x, x]$ for any θ.

With respect to $[\,\cdot\,, \cdot\,]$, C-symmetric operators somewhat resemble selfadjoint operators. For instance, an operator T is C-symmetric if and only if $[Tx, y] = [x, Ty]$ for all x, y in $\mathcal{H}$. As another example, the eigenvectors of a C-symmetric operator corresponding to distinct eigenvalues are orthogonal with respect to $[\,\cdot\,, \cdot\,]$, even though they are not necessarily orthogonal with respect to the original sesquilinear form $\langle \cdot\,, \cdot \rangle$.

Lemma 1. *The eigenvectors of a C-symmetric operator T corresponding to distinct eigenvalues are orthogonal with respect to the bilinear form $[\,\cdot\,, \cdot\,]$.*

Proof. The proof is essentially identical to the corresponding proof for selfadjoint operators. If $\lambda_1 \neq \lambda_2$, $Tx_1 = \lambda_1 x_1$, and $Tx_2 = \lambda_2 x_2$, then

$$\lambda_1[x_1, x_2] = [\lambda_1 x_1, x_2] = [Tx_1, x_2] = [x_1, Tx_2] = [x_1, \lambda_2 x_2] = \lambda_2[x_1, x_2].$$

Since $\lambda_1 \neq \lambda_2$, it follows that $[x_1, x_2] = 0$. □

There are some obvious differences between selfadjoint and complex symmetric operators. For instance, a complex symmetric matrix can have any possible Jordan canonical form (see [5, 6] and the references therein) while a selfadjoint matrix must be unitarily diagonalizable. Nevertheless, we will see in Section 2 that the generalized eigenspaces of an arbitrary C-symmetric operator are always C-orthogonal.

Another somewhat superficial resemblance between complex symmetric and selfadjoint operators concerns the relationship between the kernel and range. If T is a C-symmetric operator, then the subspaces $\ker T$ and $\mathrm{cl}(\mathrm{ran}\, T)$ are C-orthogonal subspaces. Indeed, this follows immediately from the definition of C-symmetry and the fact that $\ker T = (\mathrm{ran}\, T^*)^\perp$. In this respect, C-symmetric operators resemble

selfadjoint operators since the kernel and range of a selfadjoint operator are always orthogonal to each other. On the other hand, it turns out that $\ker T \cap \mathrm{cl}(\mathrm{ran}\, T)$ may be nontrivial for arbitrary complex symmetric operators. The vectors x in this intersection are *isotropic*, meaning that $[x, x] = 0$. The simplest example of this phenomenon occurs in two dimensions:

Example 1. If $T : \mathbb{C}^2 \to \mathbb{C}^2$ denotes the operator induced by a nilpotent 2×2 Jordan block, then T is C-symmetric with respect to $C(z_1, z_2) = (\overline{z_2}, \overline{z_1})$ (see [5, 6]). It is clear that $\ker T = \mathrm{ran}\, T = \mathrm{span}\{\binom{1}{0}\}$ and that $[\binom{1}{0}, \binom{1}{0}] = 0$.

As we will see in Section 4, isotropic eigenvectors of complex symmetric operators play an important role. To be specific, the existence of isotropic eigenvectors is directly related to the multiplicity of the corresponding eigenvalue.

Diagonalizable complex symmetric operators are naturally quite tractable objects of study. We discuss several aspects of the diagonalization of complex symmetric operators in Section 5. We conclude this note with a few basic remarks on Riesz bases of eigenvectors in Section 6.

Acknowledgments

The author would like to thank M. Putinar and D. Sarason for their helpful comments. Some of this work was conducted at the University of California, Santa Barbara.

2. Generalized eigenspaces

In this section, we show that the generalized eigenspaces of a complex symmetric operator (when they exist) are always mutually C-orthogonal. Thus although we do not necessarily have orthogonality with respect to the original hermitian form $\langle \cdot, \cdot \rangle$, we are able to separate the generalized eigenspaces via the bilinear form $[\cdot, \cdot]$. Our first proof is purely algebraic. A somewhat less general, but slicker and more sophisticated approach based on the Riesz functional calculus (which suffices for most cases of interest) is discussed later.

Theorem 1. *If T is a C-symmetric operator and $\lambda_1 \neq \lambda_2$, then*

$$\ker(T - \lambda_1 I)^{m_1} \perp_C \ker(T - \lambda_2 I)^{m_2}$$

for all $m_1, m_2 \geq 0$. In particular, generalized eigenspaces of a C-symmetric operator corresponding to distinct eigenvalues are mutually C-orthogonal.

Proof. Note that the case $m_1 = m_2 = 1$ is handled via Lemma 1. By subtracting a multiple of the identity from T, we may assume that $\lambda_1 = 0$ and that $\lambda_2 = \lambda$ is nonzero. Moreover, it suffices to show that any two subspaces of the form

$$\begin{aligned} \mathcal{E}_1 &= \mathrm{span}\{x, Tx, T^2x, \ldots, T^{m_1}x\} \\ \mathcal{E}_2 &= \mathrm{span}\{y, (T - \lambda I)y, (T - \lambda I)^2 y, \ldots, (T - \lambda I)^{m_2} y\}, \end{aligned}$$

where $T^{m_1}x \neq 0$, $(T-\lambda I)^{m_2}y \neq 0$, and $T^{m_1+1}x = (T-\lambda I)^{m_2+1}y = 0$, are mutually C-orthogonal.

Step 1: We first prove that the eigenvector $(T-\lambda I)^{m_2}y$ is C-orthogonal to $\mathcal{E}_1$ by showing that $[T^j x, (T-\lambda I)^{m_2}y] = 0$ for $0 \leq j \leq m_1$. Indeed, we have

$$\begin{aligned}\lambda^{m_1+1-j}[T^j x, (T-\lambda I)^{m_2}y] &= [T^j x, \lambda^{m_1+1-j}(T-\lambda I)^{m_2}y] \\ &= [T^j x, T^{m_1+1-j}(T-\lambda I)^{m_2}y] \\ &= [T^{m_1+1}x, (T-\lambda I)^{m_2}y] \\ &= 0.\end{aligned}$$

Since $\lambda \neq 0$, it follows that $[T^j x, (T-\lambda I)^{m_2}y] = 0$ for all $0 \leq j \leq m_1$ as claimed.

Step 2: Suppose now that we have proved $[T^j x, (T-\lambda I)^{m_2-k}g] = 0$ for all $0 \leq j \leq m_1$ and some $0 \leq k \leq m_2 - 1$. Under this assumption we have:

$$\begin{aligned}\lambda[T^j x, (T-\lambda I)^{m_2-k-1}y] &= [T^j x, \lambda(T-\lambda I)^{m_2-k-1}y + (T-\lambda I)^{m_2-k}y] \\ &= [T^j x, (\lambda I + (T-\lambda I))(T-\lambda I)^{m_2-k-1}y] \\ &= [T^j x, T(T-\lambda I)^{m_2-k-1}y] \\ &= [T^{j+1}x, (T-\lambda I)^{m_2-k-1}y].\end{aligned}$$

Iteration ultimately yields

$$\lambda^{m_1+1-j}[T^j x, (T-\lambda I)^{m_2-k-1}y] = [T^{m_1+1}x, (T-\lambda I)^{m_2-k-1}y] = 0$$

whence $[T^j x, (T-\lambda I)^{m_2-k-1}y] = 0$ for all $0 \leq j \leq m_1$. Thus the theorem is proved by induction. □

A slight modification of this argument shows that generalized eigenspaces can be separated from the "quasinilpotent vectors" as well (see Section 3).

Example 2. If $\mathbf{x} = (x_1, x_2, \ldots, x_n)$ and $\mathbf{y} = (y_1, y_2, \ldots, y_n)$ are generalized eigenvectors which correspond to distinct eigenvalues of a complex symmetric matrix acting on $\mathbb{C}^n$, then $\sum_{i=1}^n x_i y_i = [\mathbf{x}, \mathbf{y}] = 0$.

Example 3. If $\mathbf{x} = (x_1, x_2, \ldots, x_n)$ and $\mathbf{y} = (y_1, y_2, \ldots, y_n)$ are generalized eigenvectors which correspond to distinct eigenvalues of an $n \times n$ Toeplitz matrix, then $[\mathbf{x}, \mathbf{y}] = \sum_{i=1}^n x_i y_{n-i} = 0$ (see [5, 6] for details).

One of the classical techniques of spectral theory for dealing with generalized eigenspaces is the use of contour integrals involving the resolvent. Recall that the resolvent set $\rho(T)$ of a bounded linear operator T is the set of complex numbers z for which the *resolvent* $R(z,T) = (zI - T)^{-1}$ exists as a bounded operator. The spectrum $\sigma(T)$ of T is simply the complement of $\rho(T)$ in $\mathbb{C}$ and the resolvent is an analytic operator valued function on the open set $\rho(T)$.

If T is a bounded linear operator and f is a holomorphic function on a (not necessarily connected) neighborhood Ω of $\sigma(T)$, then the Riesz functional calculus

allows us to define an operator $f(T)$ via the Cauchy-type integral formula

$$f(T) = \frac{1}{2\pi i} \int_\Gamma f(z) R(z,T)\, dz \tag{2}$$

where Γ denotes a finite system of rectifiable Jordan curves, oriented in the positive sense, lying in Ω [4, p. 568]. The integral in (2) is to be interpreted in the sense of Riemann and hence $f(T)$ is approximable by Riemann sums involving the operator $R(z,T)$.

For each *clopen* (relatively open and closed) subset Δ of $\sigma(T)$, there exists a natural idempotent $P(\Delta)$ defined by the formula

$$P(\Delta) = \frac{1}{2\pi i} \int_\Gamma R(z,T)\, dz \tag{3}$$

where Γ is any rectifiable Jordan curve such that Δ is contained in the interior $\mathrm{int}(\Gamma)$ of Γ and $\sigma(T)\backslash\Delta$ does not intersect $\mathrm{int}(\Gamma)$. We refer to this idempotent as the *Riesz idempotent* corresponding to Δ.

If the spectrum of an operator T decomposes as the disjoint union of two clopen sets, then the corresponding Riesz idempotents are usually neither self-adjoint (i.e., they are not necessarily orthogonal projections), nor are their ranges necessarily orthogonal to each other. Indeed, any diagonalizable but non-normal operator on $\mathbb{C}^2$ shows that this is not the case. Nevertheless, the Riesz idempotents that arise from complex symmetric operators have some nice features.

Theorem 2. *Let T be a C-symmetric operator. If $\sigma(T)$ decomposes as the disjoint union $\sigma(T) = \Delta_1 \cup \Delta_2$ of two clopen sets, then the corresponding Riesz idempotents $P_1 = P(\Delta_1)$ and $P_2 = P(\Delta_2)$ (defined by (3)) are*

(i) *C-symmetric: $P_i = CP_i^*C$ for $i = 1, 2$,*

(ii) *C-orthogonal, in the sense that $\operatorname{ran} P_1 \perp_C \operatorname{ran} P_2$ (i.e., $P_1P_2 = P_2P_1 = 0$).*

Proof. For each z in $\rho(T)$, it is easy to see that the resolvent $R(z,T) = (zI - T)^{-1}$ of T is also C-symmetric. Indeed, it can be uniformly approximated by polynomials in T and such polynomials are clearly C-symmetric. Since the Riesz idempotents P_1 and P_2 corresponding to Δ_1 and Δ_2, respectively, are approximated by Riemann sums, it follows that P_1 and P_2 are C-symmetric. In particular, the Riesz idempotents P_1 and P_2 are C-symmetric and satisfy $P_1P_2 = P_2P_1 = 0$, whence their ranges are C-orthogonal. □

We will refer to a C-symmetric idempotent as a *C-projection*. In other words, a bounded linear operator P is a C-projection if and only if $P = CP^*C$ and $P^2 = P$. It is not hard to see that if P is a C-projection, then $\| P \| \geq 1$ and $\operatorname{ran} P$ is closed. Moreover, for any C-projection, we have $\ker P \cap \operatorname{ran} P = \{0\}$. This is not true for all C-symmetric operators, as Example 1 shows.

Example 4. If u is a nonisotropic vector, normalized so that $[u,u] = 1$, then $P_u x = [x,u]u$ is the C-projection onto $\operatorname{span}\{u\}$. On the other hand, if u is isotropic, then there can be no C-projection onto the subspace spanned by u. Indeed, such

an operator would have to be of the form $Px = \langle x, v\rangle u$ for some v. If $P = CP^*C$, then it would follow that u and Cv are multiples of each other and hence $Pu = 0$, a contradiction.

If P is a C-projection, then $\ker P$ and $\operatorname{ran} P$ are disjoint C-orthogonal subspaces and hence $I = P + (I - P)$ gives a C-orthogonal decomposition of $\mathcal{H}$ (the C-orthogonality of $\ker P$ and $\operatorname{ran} P$ follows from Lemma 1).

A classical theorem of spectral theory [4, p. 579] states that if T is a compact operator, then every nonzero point λ in $\sigma(T)$ is an eigenvalue of finite order $m = m(\lambda)$. For each such λ, the corresponding Riesz idempotent has a nonzero finite dimensional range given by $\operatorname{ran} P_\lambda = \ker(T - \lambda I)^m$. In particular, the nonzero elements of the spectrum of a compact operator correspond to generalized eigenspaces. Using Riesz idempotents, it is possible to give a much shorter proof of Theorem 1 if the complex symmetric operator T is assumed to be compact.

Theorem 3. *The generalized eigenspaces of a compact C-symmetric operator are C-orthogonal.*

Proof. It follows immediately from Theorem 2 and the preceding remarks that the generalized eigenspaces corresponding to nonzero eigenvalues of a compact C-symmetric operator T are mutually C-orthogonal. Since 0 is the only possible accumulation point of the eigenvalues of T, it follows that a generalized eigenvector corresponding to a nonzero eigenvalue is C-orthogonal to any vector in the range of

$$P_\epsilon = \frac{1}{2\pi i}\int_{|z|=\epsilon} R(z, T)\, dz$$

if $\epsilon > 0$ is taken sufficiently small. In particular, $\operatorname{ran} P_\epsilon$ contains the generalized eigenvectors for the eigenvalue 0 (if any exist). □

3. Quasinilpotent Vectors

Recall that a bounded linear operator $T : \mathcal{H} \to \mathcal{H}$ is called *quasinilpotent* if

$$\lim_{n\to\infty} \| T^n \|^{\frac{1}{n}} = 0.$$

In particular, any nilpotent operator is quasinilpotent and the spectral radius formula implies that a bounded operator is quasinilpotent if and only if $\sigma(T) = 0$.

There are many examples of quasinilpotent complex symmetric operators. For instance, quasinilpotent Hankel operators can be constructed using certain symbols with lacunary Fourier series [9, Section 10.3, p. 443–449]. A familiar example of a quasinilpotent complex symmetric operator is the Volterra integration operator [6, 7]. Moreover, the Fredholm alternative indicates that *any* Volterra operator is quasinilpotent [8, Pr. 187] and hence quasinilpotent complex symmetric operators are quite easy to produce.

We say that a vector q in $\mathcal{H}$ is a *quasinilpotent vector* (for T) if

$$\lim_{n\to\infty} \| T^n q \|^{\frac{1}{n}} = 0.$$

If T is compact, then it is not hard to show that the set $\mathcal{Q}$ of quasinilpotent vectors coincides with the orthogonal complement of the span of the generalized eigenspaces of T^*.

Theorem 4. *$\mathcal{Q} \perp_C \ker(T-\lambda I)^m$ for all $m \geq 0$ and $\lambda \neq 0$. In other words, every quasinilpotent vector is C-orthogonal to the generalized eigenspaces of T corresponding to nonzero eigenvalues.*

Proof. We proceed by induction on m. The case $m = 0$ is trivial. Now suppose that we have shown that $\mathcal{Q} \perp_C \ker(T-\lambda I)^m$ for some m. Let q denote an arbitrary quasinilpotent vector for T and let x be a unit vector in $\ker(T-\lambda I)^{m+1}$. It follows that the vector $y = (T-\lambda I)x$ belongs to $\ker(T-\lambda I)^m$ and hence

$$\lambda[q,x] = [q,\lambda x] = [q,Tx] - [q,y] = [Tq,x]$$

by the inductive hypothesis. Iteration of the preceding yields $\lambda^n[q,x] = [T^n q,x]$ from which it follows that

$$|\lambda|^n|[q,x]| = |[q,\lambda^n x]| = |[q,T^n x]| = |[T^n q,x]| \leq \| T^n q \|$$

holds for every $n \geq 0$. Taking nth roots shows that $|\lambda||[q,x]|^{\frac{1}{n}} \leq \| T^n q \|^{\frac{1}{n}}$, which tends to 0. Since $\lambda \neq 0$, it follows that $[q,x] = 0$ and hence q is C-orthogonal to $\ker(T-\lambda I)$. □

4. Isotropic eigenvectors and multiplicity

We say that a vector x is *isotropic* if $[x,x] = 0$. Although 0 is clearly an isotropic vector, it turns out that nonzero isotropic vectors are nearly unavoidable (see Lemma 2 below). However, isotropic eigenvectors are not mere algebraic inconveniences, for they often have meaningful interpretations. For example, isotropic eigenvectors of complex symmetric matrices are considered in [12] in the context of elastic wave propagation. In that theory, isotropic eigenvectors correspond to circularly polarized waves.

The following simple lemma implies that any subspace of dimension ≥ 2 contains isotropic vectors (see [2, Lem. 2]). In particular, this suggests the relationship between isotropy and multiplicity that we will explore in this section.

Lemma 2. *If C is a conjugation on a complex Hilbert space $\mathcal{H}$, then every subspace of dimension ≥ 2 contains isotropic vectors for the bilinear form $[x,y] = \langle x, Cy\rangle$.*

Proof. Let $\dim \mathcal{H} \geq 2$ and consider the span of two linearly independent vectors x_1 and x_2. If either x_1 or x_2 is isotropic, then we are done. If neither x_1 nor x_2 is isotropic, then we easily obtain C-orthogonal vectors y_1 and y_2 with the same span as x_1 and x_2:

$$y_1 = x_1, \quad y_2 = x_2 - \frac{[x_2,x_1]}{[x_1,x_1]}x_1.$$

In this case, either y_2 is isotropic (and hence we are done) or neither y_1 nor y_2 is isotropic. If this happens, then we may assume that y_1 and y_2 are normalized so

that $[y_1, y_1] = [y_2, y_2] = 1$. It is then easily verified that the vectors $y_1 \pm iy_2$ are both isotropic. □

The following lemma shows that the existence of an isotropic eigenvector for an isolated eigenvalue is completely determined by the multiplicity of the corresponding eigenvalue.

Theorem 5. *If T is a C-symmetric operator, then an isolated eigenvalue λ of T is simple if and only if T has no isotropic eigenvectors for λ.*

Proof. If λ is an isolated eigenvalue of T, then the Riesz idempotent P corresponding to λ is a C-projection. If λ is a simple eigenvalue, then the eigenspace corresponding to λ is spanned by a single unit vector x. If x is isotropic, then it is C-orthogonal to all of $\mathcal{H}$ since x is C-orthogonal to the range of the complementary C-projection $I - P$. This would imply that x is C-orthogonal to all of $\mathcal{H}$ and hence $x = 0$, a contradiction. Conversely, if λ is not a simple eigenvalue, then there are two cases to consider:

Case 1: If $\dim \ker(T-\lambda I) > 1$, then by Lemma 2, $\ker(T-\lambda I)$ contains an isotropic vector. Thus T has an isotropic eigenvector corresponding to the eigenvalue λ.

Case 2: If $\dim \ker(T - \lambda I) = 1$, then $\ker(T - \lambda I) = \operatorname{span}\{x\}$ for some $x \neq 0$ and $\dim \ker(T - \lambda I)^2 > 1$ since λ is not a simple eigenvalue. We can therefore find a nonzero generalized eigenvector y for λ such that $x = (T - \lambda I)y$. Thus

$$[x, x] = [x, (T - \lambda I)y] = [(T - \lambda I)x, y] = 0$$

and hence x is an isotropic eigenvector. □

We remark that the hypothesis that the eigenvalue λ is isolated is crucial. Indeed, $S \oplus S^*$ (where S is the unilateral shift on l^2) is complex symmetric (see [7]) and has each point in the open unit disk as a simple eigenvalue. The corresponding eigenvectors are all isotropic.

Example 5. If λ is an isolated eigenvalue of multiplicity ≥ 2 for a Hankel matrix (possibly infinite), then there exists an eigenvector $\mathbf{x}$ corresponding to λ so that $\sum_{i=1}^{\dim \mathcal{H}} x_i^2 = [\mathbf{x}, \mathbf{x}] = 0$. Here x_i denotes the ith entry of the vector $\mathbf{x}$.

Example 6. If λ is an eigenvalue of multiplicity ≥ 2 for an $n \times n$ Toeplitz matrix, then there exists an eigenvector $\mathbf{x} = (x_1, x_2, \ldots, x_n)$ corresponding to λ so that $\sum_{i=1}^{n} x_i x_{n-i} = [\mathbf{x}, \mathbf{x}] = 0$.

Example 7. Let H^2 denote the Hardy space of the open unit disk and let φ denote a nonconstant inner function. If λ is an eigenvalue of multiplicity ≥ 2 for the compression of a Toeplitz operator to $H^2 \ominus \varphi H^2$, then there exists an eigenfunction f in $H^2 \ominus \varphi H^2$ so that

$$\frac{1}{2\pi i} \int_{\partial \mathbb{D}} \frac{f^2(z)}{\varphi(z)}\, dz = 0,$$

where $\partial \mathbb{D}$ denotes the unit circle (oriented in the counter-clockwise sense). See [3, 5, 6, 7] for further details and various special cases.

Having seen that isotropic eigenvectors are related to the multiplicity of eigenvalues, we can go a bit further into the decomposition of generalized eigenspaces:

Theorem 6. *If T is a C-symmetric operator, then for each complex number λ there exists an increasing sequence of subspaces $\mathcal{K}_n^{(\lambda)}$ such that*

$$\ker(T-\lambda I)^n = [\ker(T-\lambda I)^n \cap \mathrm{cl}(\mathrm{ran}(T-\lambda I)^n)] \oplus \mathcal{K}_n^{(\lambda)}$$

is both an orthogonal and C-orthogonal direct sum. Furthermore, the subspaces $\ker(T-\lambda I)^n \cap \mathrm{cl}(\mathrm{ran}(T-\lambda I)^n)$ consist entirely of isotropic vectors.

Proof. It suffices to consider the case $\lambda = 0$. For each n, there exists a subspace $\mathcal{K}_n = \mathcal{K}_n^{(0)}$ of $\ker T^n$ (possibly the zero subspace) such that

$$\ker T^n = [\ker T^n \cap \mathrm{cl}(\mathrm{ran}\, T^n)] \oplus \mathcal{K}_n, \tag{4}$$

where $\oplus$ denotes the usual orthogonal direct sum. Since $\ker T^n$ is C-orthogonal to $\mathrm{cl}(\mathrm{ran}\, T^n)$ and $\mathcal{K}_n \subseteq \ker T^n$, it follows that (4) is also a C-orthogonal decomposition. In particular, every vector in $\ker T^n \cap \mathrm{cl}(\mathrm{ran}\, T^n)$ is isotropic. That the sequence $\mathcal{K}_n$ is increasing is clear from the fact that $\ker T^n \subseteq \ker T^{n+1}$ and $\mathrm{ran}\, T^{n+1} \subseteq \mathrm{ran}\, T^n$. □

Example 8. Consider the operator $T : \mathbb{C}^5 \to \mathbb{C}^5$ induced by a 5×5 nilpotent Jordan block and let $\{e_1, e_2, e_3, e_4, e_5\}$ denote the standard orthonormal basis for $\mathbb{C}^5$. The relevant subspaces $\mathcal{K}_n^{(0)}$ of the preceding theorem are readily exhibited:

n	$\ker T^n$	$\mathrm{ran}\, T^n$	$\ker T^n \cap \mathrm{ran}\, T^n$	$\mathcal{K}_n^{(0)}$
0	$\{0\}$	$\mathbb{C}^5$	$\{0\}$	$\mathbb{C}^5$
1	$\mathrm{span}\{e_1\}$	$\mathrm{span}\{e_1, e_2, e_3, e_4\}$	$\mathrm{span}\{e_1\}$	$\{0\}$
2	$\mathrm{span}\{e_1, e_2\}$	$\mathrm{span}\{e_1, e_2, e_3\}$	$\mathrm{span}\{e_1, e_2\}$	$\{0\}$
3	$\mathrm{span}\{e_1, e_2, e_3\}$	$\mathrm{span}\{e_1, e_2\}$	$\mathrm{span}\{e_1, e_2\}$	$\{e_3\}$
4	$\mathrm{span}\{e_1, e_2, e_3, e_4\}$	$\mathrm{span}\{e_1\}$	$\mathrm{span}\{e_1\}$	$\{e_2, e_3, e_4\}$
5	$\mathbb{C}^5$	$\{0\}$	$\{0\}$	$\mathbb{C}^5$

5. Diagonalization of complex symmetric operators

Suppose that T is a complex symmetric operator which has a complete system of (nonzero) eigenvectors $(u_n)_{n=1}^{\infty}$. By complete, we mean that the closed linear span of the u_n is all of $\mathcal{H}$. Implicitly, we will assume that $\dim \mathcal{H} = \infty$ since the finite dimensional case is somewhat trivial in comparison.

If the corresponding eigenvalues λ_n are distinct, then Lemma 1 tells us that the $(u_n)_{n=1}^{\infty}$ are mutually C-orthogonal. We may therefore assume that the system $(u_n)_{n=1}^{\infty}$ is *C-orthonormal*: $[u_j, u_k] = \delta_{jk}$, where δ_{jk} denotes the Kronecker δ-function. Indeed, if $[u_n, u_n] = 0$ for some n, then $[u_n, x] = 0$ would hold for all x since the system $(u_n)_{n=1}^{\infty}$ is complete, whence $u_n = 0$.

We consider here the linear extension of the map $u_n \mapsto Cu_n$. Since the u_n are not necessarily orthonormal with respect to the usual hermitian inner product $\langle \cdot, \cdot \rangle$, this map does not immediately extend (as a bounded linear operator)

further than the dense linear submanifold spanned by finite linear combinations of the u_n. To be specific, we say that a vector f in $\mathcal{H}$ is *finitely supported* if it is a finite linear combination of the u_n and we denote the linear manifold of finitely supported vectors by $\mathcal{F}$. Due to the C-orthonormality of the u_n, it follows immediately that each such f can be recovered via the *skew Fourier expansion*

$$f = \sum_{n=1}^{\infty} [f, u_n] u_n, \tag{5}$$

where all but finitely many of the *skew Fourier coefficients* $[f, u_n]$ are nonzero. We will let $A_0 : \mathcal{F} \to \mathcal{H}$ denote the linear extension of the map $A_0 u_n = C u_n$ to $\mathcal{F}$. Since $\mathcal{F}$ is a dense linear submanifold of $\mathcal{H}$, it follows that if $A_0 : \mathcal{F} \to \mathcal{H}$ is bounded on $\mathcal{F}$, then A_0 has a unique bounded extension (which we denote by A) to all of $\mathcal{H}$.

It turns out that the presence of the conjugation C ensures that the extension A will have several desirable algebraic properties. In particular, the following lemma shows that if A is bounded, then it is C-orthogonal. Specifically, we say that an operator $U : \mathcal{H} \to \mathcal{H}$ is *C-orthogonal* if $CU^*CU = I$. The terminology comes from the fact that, when represented with respect to a C-real orthonormal basis, the corresponding matrix will be complex orthogonal (i.e., $U^t U = I$ as matrices).

The importance of C-orthogonal operators lies in the fact that they preserve the bilinear form induced by C. To be specific, U is a C-orthogonal operator if and only if $[Ux, Uy] = [x, y]$ for all x, y in $\mathcal{H}$. Unlike unitary operators, C-orthogonal operators can have arbitrarily large norms. In fact, unbounded C-orthogonal operators are considered in [11], where they are called *J-unitary* operators.

Lemma 3. *If A_0 is bounded, then its extension $A : \mathcal{H} \to \mathcal{H}$ is positive and C-orthogonal. If this is the case, then A is invertible with $A^{-1} = CAC \geq 0$ and the operator $B = \sqrt{A}$ is also C-orthogonal.*

Proof. By (5), it follows that $\langle A_0 f, f \rangle = \sum_{n=1}^{\infty} |[f, u_n]|^2 \geq 0$ for all f in $\mathcal{F}$. If A_0 is bounded, then it follows by continuity that A will be positive. The fact that A is C-orthogonal (hence invertible) follows from the fact that $(CA^*C)Au_n = (CA)^2 u_n = u_n$ for all n. Since $(CBC)(CBC) = CAC = A^{-1}$ and $CBC \geq 0$, it follows that CBC is a positive square root of A^{-1}. By the uniqueness of the positive square root of a positive operator, we see that $CBC = B^{-1}$ and hence B is also C-orthogonal. □

We remark that Lemma 3 shows that if the map $u_n \mapsto C u_n$ is bounded, then its linear extension $A : \mathcal{H} \to \mathcal{H}$ is necessarily invertible. This property distinguishes C-orthonormal systems $(u_n)_{n=1}^{\infty}$ and their duals $(Cu_n)_{n=1}^{\infty}$ from general biorthogonal systems (which do not necessarily arise from conjugations on $\mathcal{H}$). Among other things, Lemma 3 also shows that if A_0 is bounded, then the *skew conjugation* $J\left(\sum_{n=1}^{\infty} c_n u_n\right) = \sum_{n=1}^{\infty} \overline{c_n} u_n$ (defined initially on $\mathcal{F}$) is given by

$$J = CA = CBB = B^{-1}CB.$$

In other words, the skew conjugation J is similar to our original conjugation C via the operator $B = \sqrt{A}$. Another consequence of the boundedness of A_0 is the existence of a natural orthonormal basis for $\mathcal{H}$:

Lemma 4. *If A_0 is bounded, then the vectors $(s_n)_{n=1}^{\infty}$ defined by $s_n = Bu_n$ (where $B = \sqrt{A}$) satisfy the following:*

(i) *$(s_n)_{n=1}^{\infty}$ is orthonormal: $\langle s_j, s_k \rangle = \delta_{jk}$ for all j, k,*

(ii) *$(s_n)_{n=1}^{\infty}$ is C-orthonormal: $[s_j, s_k] = \delta_{jk}$ for all j, k,*

(iii) *$Cs_n = s_n$ for all n.*

Furthermore, $(s_n)_{n=1}^{\infty}$ is an orthonormal basis for $\mathcal{H}$.

Proof. Conditions (i), (ii), and (iii) follow from direct computations:

$$\langle s_j, s_k \rangle = \langle Bu_j, Bu_k \rangle = \langle u_j, Au_k \rangle = \langle u_j, Cu_k \rangle = [u_j, u_k] = \delta_{jk},$$
$$[s_j, s_k] = \langle s_j, Cs_k \rangle = \langle Bu_j, CBu_k \rangle = \langle Bu_j, B^{-1}Cu_k \rangle = \langle u_j, Cu_k \rangle = \delta_{jk},$$
$$Cs_j = CBu_j = B^{-1}Cu_j = B^{-1}B^2u_j = Bu_j = s_j.$$

We now show that the system $(s_n)_{n=1}^{\infty}$ is complete. If f is orthogonal to each s_j, then $\langle Bf, u_j \rangle = \langle f, Bu_j \rangle = \langle f, s_j \rangle = 0$ for all j. Since B is invertible, it follows that $f = 0$ since $(u_n)_{n=1}^{\infty}$ is complete. □

If the operator A_0 is bounded, then its extension A is a positive, invertible operator whose spectrum is bounded away from zero. Thus $\Theta = -i \log A$ can be defined using the functional calculus for A and the principal branch of the logarithm. Since A is self-adjoint and the principal branch of the logarithm is real on $(0, \infty)$, it follows that Θ is skew-Hermitian: $\Theta^* = -\Theta$. Moreover, since A is a C-orthogonal operator, it follows that Θ is a *C-real* operator: $\overline{\Theta} = \Theta$, where $\overline{\Theta} = C\Theta C$.

Returning to our original C-symmetric operator T, we see that if A_0 is bounded, then T is similar to the diagonal operator $D : \mathcal{H} \to \mathcal{H}$ defined by $Ds_n = \lambda_n s_n$ since $T = B^{-1}DB$. Writing this in terms of the exponential representation $A = \exp(i\Theta)$ and inserting a parameter $t \in [0, 1]$, we obtain a family of operators

$$T_t = e^{-\frac{it}{2}\Theta} D e^{\frac{it}{2}\Theta}$$

which satisfies $T_0 = D$ and $T_1 = T$. This provides a continuous deformation of T to its diagonal model D. We also remark that the fact that Θ is C-real and skew-Hermitian implies that the operators $\exp(\pm\frac{it}{2}\Theta)$ are C-orthogonal for all t. From here, it is easy to show that each intermediate operator T_t is C-symmetric and that the path $t \mapsto T_t$ from $[0, 1]$ to $B(\mathcal{H})$ is norm continuous.

In particular, this framework applies to complex symmetric matrices (e.g., Hankel matrices) or to finite Toeplitz matrices. Moreover, the compressed shift corresponding to an interpolating Blaschke product produces exactly such a system $(u_n)_{n=1}^{\infty}$ (consisting of certain scalar multiples of reproducing kernels, see [6] for details and references). The boundedness of A_0 is guaranteed by Carleson's

interpolation theorem. It would be interesting to concretely identify the operators A, B, and Θ in such an example.

6. Riesz bases of eigenvectors

Recall that an arbitrary sequence of vectors $(u_n)_{n=1}^{\infty}$ is called a *Bessel sequence* if there exists a constant $M > 0$ (called a *Bessel bound*) such that

$$\sum_{n=1}^{\infty} |\langle x, u_n \rangle|^2 \leq M \| x \|^2$$

for all x in $\mathcal{H}$. Also recall that a sequence $(u_n)_{n=1}^{\infty}$ is called a *Riesz basis* if it is the image of an orthonormal basis of $\mathcal{H}$ under a bounded, invertible linear operator R. It is well known (see [1, Prop. 3.6.4]) that $(u_n)_{n=1}^{\infty}$ is a Riesz basis if and only if there exist there exist constants $M_1, M_2 > 0$ such that

$$M_1^2 \| x \|^2 \leq \sum_{n=1}^{\infty} |\langle x, u_n \rangle|^2 \leq M_2^2 \| x \|^2.$$

Furthermore, the optimal constants are $M_1 = \| R^{-1} \|^{-1}$ and $M_2 = \| R \|^2$.

Our final theorem consists of a number of equivalent statements concerning C-orthonormal systems. Our interest in such systems stems from the fact that they often arise as eigenvectors of C-symmetric operators.

Theorem 7. *If $(u_n)_{n=1}^{\infty}$ is a complete C-orthonormal system in $\mathcal{H}$, then the following are equivalent:*

(i) *$(u_n)_{n=1}^{\infty}$ is a Bessel sequence with Bessel bound M.*

(ii) *$(u_n)_{n=1}^{\infty}$ is a Riesz basis with lower and upper bounds M^{-1} and M.*

(iii) *The assignment $A_0 u_n = C u_n$ extends to a bounded linear operator $A : \mathcal{H} \to \mathcal{H}$ satisfying $\| A \| \leq M$.*

(iv) *There exists $M > 0$ satisfying*

$$\| \sum_{n=1}^{m} \overline{c_n} u_n \| \leq M \| \sum_{n=1}^{m} c_n u_n \| \tag{6}$$

for every finite sequence $c_1, c_2, \ldots, c_m$.

(v) *The Gram matrix $\left(\langle u_j, u_k \rangle\right)_{j,k=1}^{\infty}$, acting on finitely supported sequences, dominates its transpose:*

$$\left(M^2 \langle u_j, u_k \rangle - \langle u_k, u_j \rangle\right)_{j,k=1}^{\infty} \geq 0 \tag{7}$$

for some $M > 0$.

(vi) *The Gram matrix $G = \left(\langle u_j, u_k \rangle\right)_{j,k=1}^{\infty}$ is bounded on $l^2(\mathbb{N})$. Furthermore, $\| G \| \leq M$ and G is orthogonal ($G^t G = I$ as matrices).*

(vii) *For each f in $\mathcal{H}$, the skew Fourier expansion $\sum_{n=1}^{\infty}[f,u_n]u_n$ converges to f in norm and*

$$\frac{1}{M}\| f \|^2 \leq \sum_{n=1}^{\infty} |[f,u_n]|^2 \leq M\| f \|^2. \tag{8}$$

In all cases, the infimum over all such M equals the norm of A_0.

Proof. The proof consists of a number of parts. We first prove the implications (i) $\Rightarrow$ (iii) $\Rightarrow$ (ii) $\Rightarrow$ (i) and then establish the equivalences (iii) $\Leftrightarrow$ (iv) $\Leftrightarrow$ (v), (iii) $\Leftrightarrow$ (vi), and (ii) $\Leftrightarrow$ (vii).

(i) $\Rightarrow$ (iii) If $(u_n)_{n=1}^{\infty}$ is a Bessel sequence with Bessel bound M, then

$$\sum_{n=1}^{\infty} |[f,u_n]|^2 = \sum_{n=1}^{\infty} |\langle Cf, u_n\rangle|^2 \leq M\| Cf \|^2 = M\| f \|^2$$

holds for all f. It follows that the coordinate map $Lf = ([f,u_n])_{n=1}^{\infty}$ is a bounded linear operator from $\mathcal{H}$ into $l^2(\mathbb{N})$ whose norm satisfies $\| L \| \leq \sqrt{M}$. Since $[f,u_n] = \langle f, Cu_n\rangle$, it is not hard to see that $L^*Lu_n = Cu_n = A_0u_n$ and thus L^*L agrees with A_0 on the dense submanifold $\mathcal{F}$. This implies that A_0 extends to a bounded linear operator A satisfying $\| A \| = \| L^*L \| \leq M$.

(iii) $\Rightarrow$ (ii) $(u_n)_{n=1}^{\infty}$ is the image of the orthonormal basis $(s_n)_{n=1}^{\infty}$ under the bounded, bijective operator B^{-1} (see the preceding section for terminology). The bounds follow from [1, Prop. 3.6.4].

(ii) $\Rightarrow$ (i) This follows from the well-known fact that a Riesz basis is always a Bessel sequence (see [1, Prop. 3.6.4]).

(iii) $\Leftrightarrow$ (iv) This follows directly from the fact that the *antilinear* operator $J = CA_0$ fixes each u_n. Since $A_0 = CJ$ on $\mathcal{F}$ and C is isometric, the desired result follows.

(iv) $\Leftrightarrow$ (v) Upon squaring both sides of (6) and simplifying, one sees that (6) holds if and only if (7) holds (with the same M).

(iii) $\Leftrightarrow$ (vi) If A_0 is bounded, then A_0 extends to a bounded, invertible operator $A : \mathcal{H} \to \mathcal{H}$. Indeed, $A^{-1} = CAC$ since these two operators agree on the complete system $(Cu_n)_{n=1}^{\infty}$. The entries in the Gram matrix G are given by $\langle u_j, u_k\rangle = \langle B^{-1}s_j, B^{-1}s_k\rangle = \langle A^{-1}s_j, s_k\rangle$. Hence G is simply the matrix representation for the bounded operator A^{-1} with respect to the C-real basis $(s_n)_{n=1}^{\infty}$. In particular, this implies that G is bounded as an operator on $l^2(\mathbb{N})$. Since $\langle As_j, s_k\rangle = \langle Bs_j, Bs_k\rangle = \langle Cu_j, Cu_k\rangle = \langle u_k, u_j\rangle$ for all j,k, it follows that G^t is simply the matrix representation for A and hence $G^tG = I$. Conversely, if G is bounded, then a straightforward computation shows that A_0 is bounded.

(ii) $\Leftrightarrow$ (vii) Condition (ii) holds if and only if

$$\frac{1}{M}\| f \|^2 \leq \sum_{n=1}^{\infty} |\langle f, u_n\rangle|^2 \leq M\| f \|^2$$

for every f in $\mathcal{H}$. Upon substituting Cf for f and noting that $[f, u_n] = \langle f, Cu_n\rangle = \langle Cf, u_n\rangle$, these inequalities assume the form required by (vii). Once (8) is established, it is clear that each f in $\mathcal{H}$ can be represented as a norm-convergent skew Fourier series. □

We conclude this article with a simple, but illustrative, example:

Example 9. Let $w = \alpha + i\beta$ where α and β are real constants and consider $L^2[0,1]$, endowed with the conjugation $[Cf](x) = \overline{f(1-x)}$. A short computation shows that if w is not an integer multiple of 2π, then the vectors

$$u_n(x) = \exp[i(w + 2\pi n)(x - \tfrac{1}{2})], \quad n \in \mathbb{Z},$$

are eigenfunctions of the C-symmetric operator

$$[Tf](x) = e^{iw/2} \int_0^x f(y)\, dy + e^{-iw/2} \int_x^1 f(y)\, dy$$

(i.e., $T = e^{iw/2}V + e^{-iw/2}V^*$ where V denotes the Volterra operator) and that the system $(u_n)_{n=1}^{\infty}$ is complete and C-orthonormal. One the other hand, one might also say that the u_n are eigenfunctions of the derivative operator with boundary condition $f(1) = e^{iw} f(0)$.

We also see that the map $u_n \mapsto Cu_n$ extends to a bounded operator on all of $L^2[0,1]$. Indeed, this extension is simply the multiplication operator $[Af](x) = e^{2\beta(x-1/2)} f(x)$ whence $B = \sqrt{A}$ is given by

$$[Bf](x) = e^{\beta(x-1/2)} f(x).$$

As expected, the positive operators A and B are both C-orthogonal and the system $(u_n)_{n=1}^{\infty}$ forms a Riesz basis for $L^2[0,1]$. In fact, $(u_n)_{n=1}^{\infty}$ is the image of the C-real orthonormal basis $(s_n)_{n=1}^{\infty}$, defined by $s_n = Bu_n$, under the bounded and invertible operator B^{-1}. The s_n are given by

$$s_n(x) = \exp[i(\alpha + 2\pi n)(x - \tfrac{1}{2})]$$

and they are easily seen to be both orthonormal and C-real (see [5, Lem. 4.3]). Such bases and their relationship to the C-symmetric properties of the Volterra operator and the "compressed shift" corresponding to the atomic inner function $\varphi(z) = \exp[(z+1)/(z-1)]$ are discussed in [5].

References

[1] Christensen, O., *An Introduction to Frames and Riesz Bases*, Birkhäuser, Boston, 2003.

[2] Craven, B.D., *Complex symmetric matrices*, J. Austral. Math. Soc. **10** (1969), 341–354.

[3] Danciger, J., Garcia, S.R., Putinar, M., *Variational principles for symmetric bilinear forms*, Math. Nachr., to appear.

[4] Dunford, N., Schwartz, J.T., *Linear operators, Part I: General Theory*, Wiley, New York, 1988.

[5] Garcia, S.R., *Conjugation and Clark operators*, Contemp. Math. **393** (2006), 67–112.

[6] Garcia, S.R., Putinar, M., *Complex symmetric operators and applications*, Trans. Amer. Math. Soc. **358** (2006), 1285–1315.

[7] Garcia, S.R., Putinar, M., *Complex symmetric operators and applications II*, Trans. Amer. Math. Soc. **359** (2007), 3913–3931.

[8] Halmos, P.R., *A Hilbert Space Problem Book* (Second Edition), Springer-Verlag, 1982.

[9] Peller, V.V., *Hankel Operators and Their Applications*, Springer Monographs in Mathematics, Springer-Verlag, 2003.

[10] Prodan, E., Garcia, S.R., Putinar, M., *Norm estimates of complex symmetric operators applied to quantum systems*, J. Phys. A: Math. Gen. **39** (2006), 389–400.

[11] Riss, U.V., *Extension of the Hilbert space by J-unitary transformations*, Helv. Phys. Acta **71** (1998), 288–313.

[12] Scott, N.H., *A theorem on isotropic null vectors and its application to thermoelasticity*, Proc. Roy. Soc. London Ser. A **440** no. 1909 (1993), 431–442.

Stephan Ramon Garcia
Department of Mathematics
Pomona College
610 North College Avenue
Claremont, CA 91711
USA
e-mail: `Stephan.Garcia@pomona.edu`

Operator Theory:
Advances and Applications, Vol. 179, 185–196

Higher Order Asymptotic Formulas for Traces of Toeplitz Matrices with Symbols in Hölder-Zygmund Spaces

Alexei Yu. Karlovich

Abstract. We prove a higher order asymptotic formula for traces of finite block Toeplitz matrices with symbols belonging to Hölder-Zygmund spaces. The remainder in this formula goes to zero very rapidly for very smooth symbols. This formula refines previous asymptotic trace formulas by Szegő and Widom and complement higher order asymptotic formulas for determinants of finite block Toeplitz matrices due to Böttcher and Silbermann.

Mathematics Subject Classification (2000). Primary 47B35; Secondary 15A15, 47B10, 47L20, 47A68.

Keywords. Block Toeplitz matrix, determinant, trace, strong Szegő-Widom theorem, decomposing algebra, canonical Wiener-Hopf factorization, Hölder-Zygmund space.

1. Introduction and main result

1.1. Finite block Toeplitz matrices

Let $\mathbb{Z}, \mathbb{N}, \mathbb{Z}_+$, and $\mathbb{C}$ be the sets of integers, positive integers, nonnegative integers, and all complex numbers, respectively. Suppose $N \in \mathbb{N}$. For a Banach space X, let X_N and $X_{N\times N}$ be the spaces of vectors and matrices with entries in X. Let $\mathbb{T}$ be the unit circle. For $1 \le p \le \infty$, let $L^p := L^p(\mathbb{T})$ and $H^p := H^p(\mathbb{T})$ be the standard Lebesgue and Hardy spaces of the unit circle. For $a \in L^1_{N\times N}$ one can define

$$a_k = \frac{1}{2\pi}\int_0^{2\pi} a(e^{i\theta})e^{-ik\theta}d\theta \quad (k \in \mathbb{Z}),$$

the sequence of the Fourier coefficients of a. Let I be the identity operator, P be the Riesz projection of L^2 onto H^2, $Q := I - P$, and define I, P, and Q on L^2_N

This work is supported by Centro de Matemática da Universidade do Minho (Portugal) and by the Portuguese Foundation of Science and Technology through the research program POCTI.

elementwise. For $a \in L^\infty_{N\times N}$ and $t \in \mathbb{T}$, put $\widetilde{a}(t) := a(1/t)$ and $(Ja)(t) := t^{-1}\widetilde{a}(t)$. Define *Toeplitz operators*

$$T(a) := PaP|\operatorname{Im} P, \quad T(\widetilde{a}) := JQaQJ|\operatorname{Im} P$$

and *Hankel operators*

$$H(a) := PaQJ|\operatorname{Im} P, \quad H(\widetilde{a}) := JQaP|\operatorname{Im} P.$$

The function a is called the *symbol* of $T(a)$, $T(\widetilde{a})$, $H(a)$, $H(\widetilde{a})$. We are interested in the asymptotic behavior of *finite block Toeplitz matrices* $T_n(a) = [a_{j-k}]_{j,k=0}^n$ generated by (the Fourier coefficients of) the symbol a as $n \to \infty$. Many results in this direction are contained in the books by Grenander and Szegő [10], Böttcher and Silbermann [3, 4, 5], Simon [18], and Böttcher and Grudsky [1].

1.2. Szegő-Widom limit theorems

Let us formulate precisely the most relevant results. Let $K^2_{N\times N}$ be the Krein algebra [12] of matrix functions a in $L^\infty_{N\times N}$ satisfying

$$\sum_{k=-\infty}^{\infty} \|a_k\|^2(|k|+1) < \infty,$$

where $\|\cdot\|$ is any matrix norm on $\mathbb{C}_{N\times N}$. The following beautiful theorem about the asymptotics of finite block Toeplitz matrices was proved by Widom [21].

Theorem 1.1. (see [21, Theorem 6.1]). *If $a \in K^2_{N\times N}$ and the Toeplitz operators $T(a)$ and $T(\widetilde{a})$ are invertible on H^2_N, then $T(a)T(a^{-1}) - I$ is of trace class and, with appropriate branches of the logarithm,*

$$\log\det T_n(a) = (n+1)\log G(a) + \log\det T(a)T(a^{-1}) + o(1) \quad \text{as} \quad n \to \infty, \quad (1)$$

where

$$G(a) := \lim_{r\to 1-0} \exp\left(\frac{1}{2\pi}\int_0^{2\pi} \log\det \widehat{a}_r(e^{i\theta})d\theta\right), \widehat{a}_r(e^{i\theta}) := \sum_{n=-\infty}^{\infty} a_n r^{|n|} e^{in\theta}. \quad (2)$$

In formula (1), $\det T(a)T(a^{-1})$ refers to the determinant defined for operators on Hilbert space differing from the identity by an operator of trace class [9, Chap. 4].

The proof of the above result in a more general form is contained in [3, Theorem 6.11] and [5, Theorem 10.30] (in this connection see also [8]).

Let $\lambda_1^{(n)}, \ldots, \lambda_{(n+1)N}^{(n)}$ denote the eigenvalues of $T_n(a)$ repeated according to their algebraic multiplicity. Let $\operatorname{sp} A$ denote the spectrum of a bounded linear operator A and $\operatorname{tr} M$ denote the trace of a matrix M. Theorem 1.1 is equivalent to the assertion

$$\sum_i \log\lambda_i^{(n)} = \operatorname{tr}\log T_n(a) = (n+1)\log G(a) + \log\det T(a)T(a^{-1}) + o(1).$$

Widom [21] noticed that Theorem 1.1 yields even a description of the asymptotic behavior of $\operatorname{tr} f(T_n(a))$ if one replaces $f(\lambda) = \log\lambda$ by an arbitrary function f

analytic in an open neighborhood of the union $\operatorname{sp} T(a) \cup \operatorname{sp} T(\widetilde{a})$ (we henceforth call such f simply analytic on $\operatorname{sp} T(a) \cup \operatorname{sp} T(\widetilde{a})$).

Theorem 1.2. (see [21, Theorem 6.2]). *If $a \in K^2_{N\times N}$ and if f is analytic on $\operatorname{sp} T(a) \cup \operatorname{sp} T(\widetilde{a})$, then*

$$\operatorname{tr} f(T_n(a)) = (n+1)G_f(a) + E_f(a) + o(1) \quad as \quad n \to \infty, \tag{3}$$

where

$$\begin{aligned} G_f(a) &:= \frac{1}{2\pi}\int_0^{2\pi} (\operatorname{tr} f(a))(e^{i\theta})d\theta, \\ E_f(a) &:= \frac{1}{2\pi i}\int_{\partial\Omega} f(\lambda)\frac{d}{d\lambda}\log\det T[a-\lambda]T[(a-\lambda)^{-1}]d\lambda, \end{aligned}$$

and Ω is any bounded open set containing $\operatorname{sp} T(a)\cup\operatorname{sp} T(\widetilde{a})$ on the closure of which f is analytic.

The proof of Theorem 1.2 for continuous symbols a is also given in [5, Section 10.90]. In the scalar case ($N = 1$) Theorems 1.1 and 1.2 go back to Gabor Szegő (see [10] and historical remarks in [3, 4, 5, 18]).

1.3. Hölder-Zygmund spaces

Suppose g is a bounded function on $\mathbb{T}$. The *modulus of continuity* of g is defined for $s \geq 0$ by

$$\omega_1(g,s) := \sup\big\{|g(e^{i(x+h)}) - g(e^{ix})| : \ x, h \in \mathbb{R}, \ |h| \leq s\big\}.$$

By the *modulus of smoothness* (of order 2) of g is meant the function (see, e.g., [19, Section 3.3]) defined for $s \geq 0$ by

$$\omega_2(g,s) := \sup\big\{|g(e^{i(x+h)}) - 2g(e^{ix}) + g(e^{i(x-h)})| : \ x, h \in \mathbb{R}, \ |h| \leq s\big\}.$$

Let $C = C(\mathbb{T})$ be the set of all continuous functions on $\mathbb{T}$. Given $\gamma > 0$, write $\gamma = m + \delta$, where $m \in \mathbb{Z}_+$ and $\delta \in (0, 1]$. The Hölder-Zygmund space $C^\gamma = C^\gamma(\mathbb{T})$ is defined (see, e.g., [16, Section 3.5.4]) by

$$C^\gamma := \big\{f \in C : f^{(j)} \in C, \ 1 \leq j \leq m, \ [f^{(m)}]_\delta < \infty\big\}$$

with the norm

$$\|f\|_\gamma := \sum_{j=0}^{m} \|f^{(j)}\|_\infty + [f^{(m)}]_\delta,$$

where $f^{(j)}$ is the derivative of order j of f, $\|\cdot\|_\infty$ is the norm in L^∞, and

$$[g]_\delta := \sup_{s>0} \frac{\omega_2(g,s)}{s^\delta}, \quad 0 < \delta \leq 1.$$

Notice that if $\gamma > 0$ is not integer, then $[g]_\delta$ can be replaced by

$$[g]^*_\delta := \sup_{s>0} \frac{\omega_1(g,s)}{s^\delta}, \quad 0 < \delta < 1$$

in the above definition.

1.4. Böttcher-Silbermann higher order asymptotic formulas for determinants

Following [21] and [5, Sections 7.5–7.6], for $n \in \mathbb{Z}_+$ and $a \in L^\infty_{N\times N}$ define the operators P_n and Q_n on H^2_N by

$$P_n : \sum_{k=0}^{\infty} a_k t^k \mapsto \sum_{k=0}^{n} a_k t^k, \quad Q_n := I - P_n.$$

The operator $P_nT(a)P_n : P_nH^2_N \to P_nH^2_N$ may be identified with the finite block Toeplitz matrix $T_n(a) := [a_{j-k}]_{j,k=0}^n$. For a unital Banach algebra A we will denote by GA the group of all invertible elements of A. For $1 \le p \le \infty$, put

$$H^p_\pm \quad := \quad \big\{a \in L^p : a_{\mp n} = 0 \text{ for } n \in \mathbb{N}\big\}.$$

Böttcher and Silbermann [2] proved among other things the following result.

Theorem 1.3. *Let $p \in \mathbb{N}$ and $\alpha, \beta > 0$ satisfy $\alpha + \beta > 1/p$. Suppose $a = u_-u_+$, where $u_+ \in G(C^\alpha \cap H^\infty_+)_{N\times N}$ and $u_- \in G(C^\beta \cap H^\infty_-)_{N\times N}$, and the Toeplitz operator $T(\widetilde{a})$ is invertible on H^2_N. Then*

(a) *there exist $v_- \in G(H^\infty_-)_{N\times N}$ and $v_+ \in G(H^\infty_+)_{N\times N}$ such that $a = v_+v_-$;*

(b) *there exists a constant $\widetilde{E}(a) \ne 0$ such that*

$$\begin{aligned}\log\det T_n(a) \quad &= \quad (n+1)\log G(a) + \log\widetilde{E}(a)\\ &\quad +\mathrm{tr}\left[\sum_{\ell=1}^{n}\sum_{j=1}^{p-1}\frac{1}{j}\left(\sum_{k=0}^{p-j-1} G_{\ell,k}(b,c)\right)^j\right]\\ &\quad +O(1/n^{(\alpha+\beta)p-1})\end{aligned}$$

as $n \to \infty$, where the correcting terms $G_{\ell,k}(b,c)$ are given by

$$G_{\ell,k}(b,c) := P_0T(c)Q_\ell\big(Q_\ell H(b)H(\widetilde{c})Q_\ell\big)^kQ_\ell T(b)P_0 \quad (\ell,k \in \mathbb{Z}_+) \tag{4}$$

and the functions b, c are given by $b := v_-u_+^{-1}$ and $c := u_-^{-1}v_+$.

If, in addition, $p = 1$, then

(c) *the operator $T(a)T(a^{-1}) - I$ is of trace class and*

$$\log\det T_n(a) = (n+1)\log G(a) + \log\det T(a)T(a^{-1}) + O(1/n^{\alpha+\beta-1}) \tag{5}$$

as $n \to \infty$.

The sketch of the proof of parts (a) and (b) is contained in [3, Sections 6.18(ii)] and in [5, Theorem 10.35(ii)]. Part (c) is explicitly stated in [3, Section 6.18(ii)] or immediately follows from [5, Theorems 10.35(ii) and 10.37(ii)].

1.5. Our main result

Our main result is the following refinement of Theorem 1.2.

Theorem 1.4. *Let $\gamma > 1/2$. If $a \in C^\gamma_{N\times N}$ and if f is analytic on $\operatorname{sp}T(a) \cup \operatorname{sp}T(\widetilde{a})$, then (3) is true with $o(1)$ replaced by $O(1/n^{2\gamma-1})$.*

Clearly, this result is predicted by Theorem 1.3(c) with $\gamma = \alpha = \beta$. The key point in the Widom's proof of Theorem 1.2 is that (1) is valid for $a - \lambda$ in place of a, uniformly with respect to λ in a neighborhood of $\partial\Omega$. We will show that the same remains true for the higher order asymptotic formula (5) with $\gamma = \alpha = \beta$. In Section 2 we collect necessary information about right and left Wiener-Hopf factorizations in decomposing algebras and mention that a nonsingular matrix function belonging to a Hölder-Zygmund space $C^\gamma_{N\times N}$ ($\gamma > 0$) admits right and left Wiener-Hopf factorizations in $C^\gamma_{N\times N}$. In Section 3 we give the proof of Theorem 1.4 using an idea of Böttcher and Silbermann [2] of a decomposition of $\operatorname{tr}\log\{I - \sum_{k=0}^{\infty} G_{n,k}(b,c)\}$. We show that this decomposition can be made for $a - \lambda$ uniform with respect to λ in a neighborhood of $\partial\Omega$. This actually implies that (5) is valid with $\gamma = \alpha = \beta$ and a replaced by $a - \lambda$ uniformly with respect to λ in a neighborhood of $\partial\Omega$. Thus, Widom's arguments apply.

1.6. Higher order asymptotic trace formulas for Toeplitz matrices with symbols from other smoothness classes

Let us mention two other classes of symbols for which higher order asymptotic formulas for $\operatorname{tr} f(T_n(a))$ are available.

Theorem 1.5. *Suppose a is a continuous $N \times N$ matrix function on the unit circle and f is analytic on $\operatorname{sp} T(a) \cup \operatorname{sp} T(\widetilde{a})$. Let $\|\cdot\|$ be any matrix norm on $\mathbb{C}_{N\times N}$.*

(a) (see [20]). *If $\gamma > 1$ and*

$$\sum_{k=-\infty}^{\infty} \|a_k\| + \sum_{k=-\infty}^{\infty} \|a_k\|^2 |k|^\gamma < \infty,$$

then (3) is true with $o(1)$ replaced by $o(1/n^{\gamma-1})$.

(b) (see [11, Corollary 1.6]). *If $\alpha, \beta > 0$, $\alpha + \beta > 1$, and*

$$\sum_{k=1}^{\infty} \|a_{-k}\| k^\alpha + \sum_{k=1}^{\infty} \|a_k\| k^\beta < \infty,$$

then (3) is true with $o(1)$ replaced by $o(1/n^{\alpha+\beta-1})$.

2. Wiener-Hopf factorization in decomposing algebras of continuous functions

2.1. Definitions and general theorems

Let $\mathbb{D}$ be the open unit disk. Let $\mathcal{R}_-$ (resp. $\mathcal{R}_+$) denote the set of all rational functions with poles only in $\mathbb{D}$ (resp. in $(\mathbb{C}\cup\{\infty\})\setminus(\mathbb{D}\cup\mathbb{T})$). Let $C_\pm$ be the closure of $\mathcal{R}_\pm$ with respect to the norm of C. Suppose $\mathcal{A}$ is a Banach algebra of continuous functions on $\mathbb{T}$ that contains $\mathcal{R}_+ \cup \mathcal{R}_-$ and has the following property: if $a \in \mathcal{A}$ and $a(t) \neq 0$ for all $t \in \mathbb{T}$, then $a^{-1} \in \mathcal{A}$. The sets $\mathcal{A}_\pm := \mathcal{A} \cap C_\pm$ are subalgebras of $\mathcal{A}$. The algebra $\mathcal{A}$ is said to be *decomposing* if every function $a \in \mathcal{A}$ can be represented in the form $a = a_- + a_+$ where $a_\pm \in \mathcal{A}_\pm$.

Let $\mathcal{A}$ be a decomposing algebra. A matrix function $a \in \mathcal{A}_{N\times N}$ is said to admit a *right* (resp. *left*) *Wiener-Hopf* (WH) *factorization in* $\mathcal{A}_{N\times N}$ if it can be represented in the form $a = a_- d a_+$ (resp. $a = a_+ d a_-$), where

$$a_\pm \in G(\mathcal{A}_\pm)_{N\times N}, \quad d(t) = \operatorname{diag}\{t^{\kappa_1}, \dots, t^{\kappa_N}\}, \quad \kappa_i \in \mathbb{Z}, \quad \kappa_1 \le \dots \le \kappa_N.$$

The integers κ_i are usually called the *right* (resp. *left*) *partial indices* of a; they can be shown to be uniquely determined by a. If $\kappa_1 = \dots = \kappa_N = 0$, then the respective WH factorization is said to be *canonical*.

The following result was obtained by Budjanu and Gohberg [6, Theorem 4.3] and it is contained in [7, Chap. II, Corollary 5.1] and in [13, Theorem 5.7′].

Theorem 2.1. *Suppose the following two conditions hold for the algebra* $\mathcal{A}$*:*

(a) *the Cauchy singular integral operator*

$$(S\varphi)(t) := \frac{1}{\pi i} v.p. \int_{\mathbb{T}} \frac{\varphi(\tau)}{\tau - t} d\tau \quad (t \in \mathbb{T})$$

is bounded on $\mathcal{A}$*;*

(b) *for any function* $a \in \mathcal{A}$*, the operator* $aS - SaI$ *is compact on* $\mathcal{A}$*.*

Then every matrix function $a \in \mathcal{A}_{N\times N}$ *such that* $\det a(t) \ne 0$ *for all* $t \in \mathbb{T}$ *admits a right and left WH factorization in* $\mathcal{A}_{N\times N}$ *(in general, with different sets of partial indices).*

Notice that (a) holds if and only if $\mathcal{A}$ is a decomposing algebra.

The following theorem follows from a more general result due to Shubin [17]. Its proof can be found in [13, Theorem 6.15].

Theorem 2.2. *Let* $\mathcal{A}$ *be a decomposing algebra and let* $\|\cdot\|$ *be a norm in the algebra* $\mathcal{A}_{N\times N}$*. Suppose* $a, c \in \mathcal{A}_{N\times N}$ *admit canonical right and left WH factorizations in the algebra* $\mathcal{A}_{N\times N}$*. Then for every* $\varepsilon > 0$ *there exists a* $\delta > 0$ *such that if* $\|a - c\| < \delta$*, then for every canonical right WH factorization* $a = a_-^{(r)} a_+^{(r)}$ *and for every canonical left WH factorization* $a = a_+^{(l)} a_-^{(l)}$ *one can choose a canonical right WH factorization* $c = c_-^{(r)} c_+^{(r)}$ *and a canonical left WH factorization* $c = c_+^{(l)} c_-^{(l)}$ *such that*

$$\|a_\pm^{(r)} - c_\pm^{(r)}\| < \varepsilon, \qquad \|[a_\pm^{(r)}]^{-1} - [c_\pm^{(r)}]^{-1}\| < \varepsilon,$$

$$\|a_\pm^{(l)} - c_\pm^{(l)}\| < \varepsilon, \qquad \|[a_\pm^{(l)}]^{-1} - [c_\pm^{(l)}]^{-1}\| < \varepsilon.$$

2.2. Wiener-Hopf factorization in Hölder-Zygmund spaces

Theorem 2.3. (see [15, Section 6.25]). *Suppose* $\gamma > 0$*. Then*

(a) C^γ *is a Banach algebra;*

(b) $a \in C^\gamma$ *is invertible in* C^γ *if and only if* $a(t) \ne 0$ *for all* $t \in \mathbb{T}$*;*

(c) S *is bounded on* C^γ*;*

(d) *for* $a \in C^\gamma$*, the operator* aI *is bounded on* C^γ *and the operator* $aS - SaI$ *is compact on* C^γ*.*

For $\gamma \notin \mathbb{Z}_+$, parts (c) and (d) are proved in [6, Section 7] (see also [7, Chap. II, Section 6.2]). Note that a statement similar to (d) is proved in [14, Chap. 7, Theorem 4.3].

Theorem 2.4. *Let $\gamma > 0$ and Σ be a compact set in the complex plane. Suppose $a : \Sigma \to C^\gamma_{N\times N}$ is a continuous function and the Toeplitz operators $T(a(\lambda))$ and $T([a(\lambda)]^\sim)$ are invertible on H^2_N for all $\lambda \in \Sigma$. Then for every $\lambda \in \Sigma$ the function $a(\lambda) : \mathbb{T} \to \mathbb{C}$ admits canonical right and left WH factorizations*

$$a(\lambda) = u_-(\lambda)u_+(\lambda) = v_+(\lambda)v_-(\lambda)$$

in $C^\gamma_{N\times N}$. These factorizations can be chosen so that $u_\pm, v_\pm, u_\pm^{-1}, v_\pm^{-1} : \Sigma \to C^\gamma_{N\times N}$ are continuous.

Proof. Fix $\lambda \in \Sigma$ and put $a := a(\lambda)$. If $T(a)$ is invertible on H^2_N, then $\det a(t) \neq 0$ for all $t \in \mathbb{T}$ (see, e.g., [7, Chap. VII, Proposition 2.1]). Then, by [7, Chap. VII, Theorem 3.2], the matrix function a admits a canonical right generalized factorization in L^2_N, that is, $a = a_-a_+$, where $a_-^{\pm1} \in (H^2_-)_{N\times N}$, $a_+^{\pm1} \in (H^2_+)_{N\times N}$ (and, moreover, the operator $a_-Pa_-^{-1}I$ is bounded on L^2_N).

On the other hand, from Theorems 2.1 and 2.3 it follows that $a \in C^\gamma_{N\times N}$ admits a right WH factorization $a = u_-du_+$ in $C^\gamma_{N\times N}$. Then

$$u_\pm \in (C^\gamma_\pm)_{N\times N} \subset (H^2_\pm)_{N\times N}, \quad u_\pm^{-1} \in (C^\gamma_\pm)_{N\times N} \subset (H^2_\pm)_{N\times N}.$$

By the uniqueness of the partial indices in a right generalized factorization in L^2_N (see, e.g., [13, Corollary 2.1]), $d = 1$.

Let us prove that a admits also a canonical left WH factorization in the algebra $C^\gamma_{N\times N}$. In view of Theorem 2.3(b), $a^{-1} \in C^\gamma_{N\times N}$. By [5, Proposition 7.19(b)], the invertibility of $T(\widetilde{a})$ on H^2_N is equivalent to the invertibility of $T(a^{-1})$ on H^2_N. By what has just been proved, there exist $f_\pm \in G(C^\gamma_\pm)_{N\times N}$ such that $a^{-1} = f_-f_+$. Put $v_\pm := f_\pm^{-1}$. Then $v_\pm \in G(C^\gamma_\pm)_{N\times N}$ and $a = v_+v_-$ is a canonical left WH factorization in $C^\gamma_{N\times N}$.

We have proved that for each $\lambda \in \Sigma$ the matrix function $a(\lambda) : \mathbb{T} \to \mathbb{C}$ admits canonical right and left WH factorizations in $C^\gamma_{N\times N}$. By Theorem 2.2, these factorizations can be chosen so that the factors $u_\pm, v_\pm$ and their inverses $u_\pm^{-1}, v_\pm^{-1}$ are continuous functions from Σ to $C^\gamma_{N\times N}$. □

3. Proof of the main result

3.1. The Böttcher-Silbermann decomposition

The following result from [3, Section 6.16], [5, Section 10.34] is the basis for our asymptotic analysis.

Lemma 3.1. *Suppose $a \in L^\infty_{N\times N}$ satisfies the following hypotheses:*

(i) *there are two factorizations $a = u_-u_+ = v_+v_-$, where $u_+, v_+ \in G(H^\infty_+)_{N\times N}$ and $u_-, v_- \in G(H^\infty_-)_{N\times N}$;*

(ii) *$u_- \in C_{N\times N}$ or $u_+ \in C_{N\times N}$.*

Define the functions b, c by $b := v_- u_+^{-1}$, $c := u_-^{-1} v_+$ and the matrices $G_{n,k}(b,c)$ by (4). Suppose for all sufficiently large n (say, $n \ge N_0$) there exists a decomposition

$$\operatorname{tr} \log \left\{ I - \sum_{k=0}^{\infty} G_{n,k}(b,c) \right\} = -\operatorname{tr} H_n + s_n \tag{6}$$

where $\{H_n\}_{n=N_0}^{\infty}$ is a sequence of $N \times N$ matrices and $\{s_n\}_{n=N_0}^{\infty}$ is a sequence of complex numbers. If $\sum_{n=N_0}^{\infty} |s_n| < \infty$, then there exist a constant $\widetilde{E}(a) \ne 0$ depending on $\{H_n\}_{n=N_0}^{\infty}$ and arbitrarily chosen $N \times N$ matrices $H_1, \ldots, H_{N_0-1}$ such that for all $n \ge N_0$,

$$\log \det T_n(a) = (n+1) \log G(a) + \operatorname{tr} (H_1 + \cdots + H_n) + \log \widetilde{E}(a) + \sum_{k=n+1}^{\infty} s_k,$$

where the constant $G(a)$ is given by (2).

3.2. The best uniform approximation

Let $\mathcal{P}^n$ be the set of all Laurent polynomials of the form

$$p(t) = \sum_{j=-n}^{n} \alpha_j t^j, \quad \alpha_j \in \mathbb{C}, \quad t \in \mathbb{T}.$$

By the Chebyshev theorem (see, e.g., [19, Section 2.2.1]), for $f \in C$ and $n \in \mathbb{N}$, there is a Laurent polynomial $p_n(f) \in \mathcal{P}^n$ such that

$$\|f - p_n(f)\|_\infty = \inf_{p \in \mathcal{P}^n} \|f - p\|_\infty. \tag{7}$$

This polynomial $p_n(f)$ is called a polynomial of best uniform approximation.

By the Jackson-Ahiezer-Stechkin theorem (see, e.g., [19, Section 5.1.4]), if f has a bounded derivative $f^{(m)}$ of order m on $\mathbb{T}$, then for $n \in \mathbb{N}$,

$$\inf_{p \in \mathcal{P}^n} \|f - p\|_\infty \le \frac{C_m}{(n+1)^m} \omega_2 \left(f^{(m)}, \frac{1}{n+1} \right), \tag{8}$$

where the constant C_m depends only on m.

From (7) and (8) it follows that if $f \in C^\gamma$ and $n \in \mathbb{N}$, where $\gamma = m + \delta$ with $m \in \mathbb{Z}_+$ and $\delta \in (0,1]$, then there is a $p_n(f) \in \mathcal{P}^n$ such that

$$\|f - p_n(f)\|_\infty \le \frac{C_m}{(n+1)^m} \omega_2 \left(f^{(m)}, \frac{1}{n+1} \right) \le \frac{C_m [f^{(m)}]_\delta}{(n+1)^{m+\delta}} \le C_m \frac{\|f\|_\gamma}{n^\gamma}. \tag{9}$$

3.3. Norms of truncations of Toeplitz and Hankel operators

Let X be a Banach space. For the definiteness, let the norm of $a = [a_{ij}]_{i,j=1}^N$ in $X_{N\times N}$ is given by $\|a\|_{X_{N\times N}} = \max\limits_{1 \le i,j \le N} \|a_{ij}\|_X$. We will simply write $\|a\|_\infty$ and $\|a\|_\gamma$ instead of $\|a\|_{L^\infty_{N\times N}}$ and $\|a\|_{C^\gamma_{N\times N}}$, respectively. Denote by $\|A\|$ the norm of a bounded linear operator A on H^2_N.

A slightly less precise version of the following statement was used in the proof of [5, Theorem 10.35(ii)].

Proposition 3.2. *Let $\alpha, \beta > 0$. Suppose $b = v_- u_+^{-1}$ and $c = u_-^{-1} v_+$, where*

$$u_+ \in G(C^\alpha \cap H_+^\infty)_{N\times N}, \quad u_- \in G(C^\beta \cap H_-^\infty)_{N\times N}, \quad v_\pm \in G(H_\pm^\infty)_{N\times N}.$$

Then there exist positive constants M_α and M_β depending only on N and α and β, respectively, such that for all $n \in \mathbb{N}$,

$$\|Q_n T(b) P_0\| \le \frac{M_\alpha}{n^\alpha} \|v_-\|_\infty \|u_+^{-1}\|_\alpha, \quad \|Q_n H(b)\| \le \frac{M_\alpha}{n^\alpha} \|v_-\|_\infty \|u_+^{-1}\|_\alpha,$$

$$\|P_0 T(c) Q_n\| \le \frac{M_\beta}{n^\beta} \|v_+\|_\infty \|u_-^{-1}\|_\beta, \quad \|H(\widetilde{c}) Q_n\| \le \frac{M_\beta}{n^\beta} \|v_+\|_\infty \|u_-^{-1}\|_\beta.$$

Proof. Since $b = v_- u_+^{-1}$, $c = u_-^{-1} v_+$ and $v_\pm, u_\pm \in G(H_\pm^\infty)_{N\times N}$, one has

$$Q_n T(b) P_0 = Q_n T(v_-) Q_n T(u_+^{-1}) P_0, \tag{10}$$

$$Q_n H(b) = Q_n T(v_-) Q_n H(u_+^{-1}), \tag{11}$$

$$P_0 T(c) Q_n = P_0 T(u_-^{-1}) Q_n T(v_+) Q_n, \tag{12}$$

$$H(\widetilde{c}) Q_n = H(\widetilde{u_-^{-1}}) Q_n T(v_+) Q_n. \tag{13}$$

Let $p_n(u_+^{-1})$ and $p_n(u_-^{-1})$ be the polynomials in $\mathcal{P}_{N\times N}^n$ of best uniform approximation of u_+^{-1} and u_-^{-1}, respectively. Obviously,

$$Q_n T[p_n(u_+^{-1})] P_0 = 0, \qquad Q_n H[p_n(u_+^{-1})] = 0,$$

$$P_0 T[p_n(u_-^{-1})] Q_n = 0, \qquad H[(p_n(u_-^{-1}))\tilde{}] Q_n = 0.$$

Then from (9) it follows that

$$\begin{aligned} \|Q_n T(u_+^{-1}) P_0\| &= \|Q_n T[u_+^{-1} - p_n(u_+^{-1})] P_0\| \\ &\le \|P\| \, \|u_+^{-1} - p_n(u_+^{-1})\|_\infty \le \frac{M_\alpha}{n_\alpha} \|u_+^{-1}\|_\alpha \end{aligned} \tag{14}$$

and similarly

$$\|Q_n H(u_+^{-1})\| \le \frac{M_\alpha}{n^\alpha} \|u_+^{-1}\|_\alpha, \tag{15}$$

$$\|P_0 T(u_-^{-1}) Q_n\| \le \frac{M_\beta}{n^\beta} \|u_-^{-1}\|_\beta, \tag{16}$$

$$\|H(\widetilde{u_-^{-1}}) Q_n\| \le \frac{M_\beta}{n^\beta} \|u_-^{-1}\|_\beta, \tag{17}$$

where M_α and M_β depend only on α, β and N. Combining (10) and (14), we get

$$\|Q_n T(b) P_0\| \le \|T(v_-)\| \, \|Q_n T(u_-^{-1}) P_0\| \le \frac{M_\alpha}{n^\alpha} \|v_-\|_\infty \|u_+^{-1}\|_\alpha.$$

All other assertions follow from (11)–(13) and (15)–(17). □

3.4. The key estimate

The following proposition shows that a decomposition of Lemma 3.1 exists.

Proposition 3.3. *Suppose the conditions of Proposition* 3.2 *are fulfilled. If* $p \in \mathbb{N}$, *then there exists a constant* $C_p \in (0,\infty)$ *depending only on* p *such that*

$$\left| \operatorname{tr} \log\left\{ I - \sum_{k=0}^{\infty} G_{n,k}(b,c) \right\} + \operatorname{tr} \left[\sum_{j=1}^{p-1} \frac{1}{j} \left(\sum_{k=0}^{p-j-1} G_{n,k}(b,c) \right)^j \right] \right| \le C_p \left(\frac{M_\alpha M_\beta}{n^{\alpha+\beta}} \|u_+^{-1}\|_\alpha \|u_-^{-1}\|_\beta \|v_-\|_\infty \|v_+\|_\infty \right)^p$$

for all $n > \left(M_\alpha M_\beta \|u_+^{-1}\|_\alpha \|u_-^{-1}\|_\beta \|v_-\|_\infty \|v_+\|_\infty\right)^{1/(\alpha+\beta)}$.

Proof. From Proposition 3.2 it follows that

$$\|G_{n,k}(b,c)\| \le \left[\frac{M_\alpha M_\beta}{n^{\alpha+\beta}} \|u_+^{-1}\|_\alpha \|u_-^{-1}\|_\beta \|v_-\|_\infty \|v_+\|_\infty \right]^{k+1}$$

for all $k \in \mathbb{Z}_+$ and $n \in \mathbb{N}$. If $n > \left(M_\alpha M_\beta \|u_+^{-1}\|_\alpha \|u_-^{-1}\|_\beta \|v_-\|_\infty \|v_+\|_\infty\right)^{1/(\alpha+\beta)}$, then the expression in the brackets is less than 1. In view of these observations the proof can be developed as in [11, Proposition 3.3]. □

Theorem 1.3 (b) follows from the above statement and Lemma 3.1. In the next section we will use the partial case $p = 1$ of Proposition 3.3 as the key ingredient of the proof of our main result.

3.5. Proof of Theorem 1.4

Suppose $\gamma > 1/2$ and $\lambda \notin \operatorname{sp} T(a) \cup \operatorname{sp} T(\widetilde{a})$. Then

$$T(a) - \lambda I = T(a-\lambda), \quad T(\widetilde{a}) - \lambda I = T([a-\lambda]^{\sim})$$

are invertible on H^2_N. Since $a-\lambda$ is continuous with respect to λ as a function from a closed neighborhood Σ of $\partial\Omega$ to $C^\gamma_{N\times N}$, in view of Theorem 2.4, for each $\lambda \in \Sigma$, the function $a - \lambda : \mathbb{T} \to \mathbb{C}$ admits canonical right and left WH factorizations $a - \lambda = u_-(\lambda)u_+(\lambda) = v_+(\lambda)v_-(\lambda)$ in $C^\gamma_{N\times N}$ and these factorizations can be chosen so that the factors $u_\pm$, $v_\pm$ and their inverses $u_\pm^{-1}$, $v_\pm^{-1}$ are continuous from Σ to $C^\gamma_{N\times N}$. Then

$$A_\Sigma := \max_{\lambda\in\Sigma} \left(\|u_+^{-1}(\lambda)\|_\gamma \|u_-^{-1}(\lambda)\|_\gamma \|v_-(\lambda)\|_\gamma \|v_+(\lambda)\|_\gamma \right) < \infty.$$

Put $b = v_- u_+^{-1}$ and $c = u_-^{-1} v_+$. From Proposition 3.3 with $p = 1$ it follows that there exists $C_1 \in (0, \infty)$ such that

$$\begin{aligned}
&\left| \operatorname{tr} \log \left\{ I - \sum_{k=0}^{\infty} G_{n,k}(b(\lambda), c(\lambda)) \right\} \right| \\
&\le \frac{C_1 M_\gamma^2}{n^{2\gamma}} \|u_+^{-1}(\lambda)\|_\gamma \|u_-^{-1}(\lambda)\|_\gamma \|v_-(\lambda)\|_\infty \|v_+(\lambda)\|_\infty \\
&\le \frac{C_1 M_\gamma^2 A_\Sigma}{n^{2\gamma}} \qquad (18)
\end{aligned}$$

for all $n > (M_\gamma^2 A_\Sigma)^{1/(2\gamma)}$ and all $\lambda \in \Sigma$. Obviously

$$\sum_{k=n+1}^{\infty} \frac{1}{k^{2\gamma}} = O(1/n^{2\gamma-1}). \qquad (19)$$

From Lemma 3.1 and (18)–(19) it follows that there is a function $\widetilde{E}(a, \cdot) : \Sigma \to \mathbb{C} \setminus \{0\}$ such that

$$\log \det T_n(a - \lambda) = (n+1) \log G(a - \lambda) + \log \widetilde{E}(a, \lambda) + O(1/n^{2\gamma-1}) \qquad (20)$$

as $n \to \infty$ and this holds uniformly with respect to $\lambda \in \Sigma$. Theorem 1.3 (c) implies that $T(a - \lambda)T([a - \lambda]^{-1}) - I$ is of trace class and

$$\widetilde{E}(a, \lambda) = \det T(a - \lambda)T([a - \lambda]^{-1}) \qquad (21)$$

for all $\lambda \in \Sigma$. Combining (20) and (21), we deduce that

$$\log \det T_n(a-\lambda) = (n+1) \log G(a-\lambda) + \log \det T(a-\lambda)T([a-\lambda]^{-1}) + O(1/n^{2\gamma-1})$$

as $n \to \infty$ uniformly with respect to $\lambda \in \Sigma$. Hence, one can differentiate both sides of the last formula with respect to λ, multiply by $f(\lambda)$, and integrate over $\partial\Omega$. The proof is finished by a literal repetition of Widom's proof of Theorem 1.2 (see [21, p. 21] or [5, Section 10.90]) with $o(1)$ replaced by $O(1/n^{2\gamma-1})$. □

References

[1] A. Böttcher and S.M. Grudsky, *Spectral Properties of Banded Toeplitz Operators*. SIAM, Philadelphia, PA, 2005.

[2] A. Böttcher and B. Silbermann, *Notes on the asymptotic behavior of block Toeplitz matrices and determinants*. Math. Nachr. **98** (1980), 183–210.

[3] A. Böttcher and B. Silbermann, *Invertibility and Asymptotics of Toeplitz Matrices*. Akademie-Verlag, Berlin, 1983.

[4] A. Böttcher and B. Silbermann, *Introduction to Large Truncated Toeplitz Matrices*. Springer-Verlag, New York, 1999.

[5] A. Böttcher and B. Silbermann, *Analysis of Toeplitz Operators*. 2nd edition. Springer-Verlag, Berlin, 2006.

[6] M.S. Budjanu and I.C. Gohberg, *General theorems on the factorization of matrix-valued functions. II. Some tests and their consequences*. Amer. Math. Soc. Transl. (2), **102** (1973), 15–26.

[7] K.F. Clancey and I. Gohberg, *Factorization of Matrix Functions and Singular Integral Operators*. Birkhäuser Verlag, Basel, 1981.

[8] T. Ehrhardt, *A new algebraic approach to the Szegő-Widom limit theorem*. Acta Math. Hungar. **99** (2003), 233–261.

[9] I.C. Gohberg and M.G. Krein, *Introduction to the Theory of Linear Nonselfadjoint Operators*. AMS, Providence, RI, 1969.

[10] U. Grenander and G. Szegő, *Toeplitz Forms and Their Applications*. University of California Press, Berkeley, Los Angeles, 1958.

[11] A.Yu. Karlovich, *Asymptotics of determinants and traces of Toeplitz matrices with symbols in weighted Wiener algebras*. Z. Anal. Anwendungen **26** (2007), 43–56.

[12] M.G. Krein, *Certain new Banach algebras and theorems of the type of the Wiener-Lévy theorems for series and Fourier integrals*. Amer. Math. Soc. Transl. (2), **93** (1970), 177–199.

[13] G.S. Litvinchuk and I.M. Spitkovsky, *Factorization of Measurable Matrix Functions*. Birkhäuser Verlag, Basel, 1987.

[14] V.V. Peller, *Hankel Operators and Their Applications*. Springer, New York, 2003.

[15] S. Prössdorf and B. Silbermann, *Numerical Analysis for Integral and Related Operator Equations*. Birkhäuser Verlag, Basel, 1991.

[16] H.-J. Schmeisser and H. Triebel, *Topics in Fourier Analysis and Function Spaces*. John Wiley & Sons, Chichester, 1987.

[17] M.A. Shubin, *Factorization of matrix functions depending on a parameter in normed rings and related problems of the theory of Noetherian operators*. Matem. Sbornik **73(115)** (1967), 610–629 (in Russian).

[18] B. Simon, *Orthogonal Polynomials on the Unit Circle. Part 1*. AMS, Providence, RI, 2005.

[19] A.F. Timan, *Theory of Approximation of Functions of a Real Variable*. Pergamon Press, Oxford, 1963.

[20] V.A. Vasil'ev, E.A. Maksimenko, and I.B. Simonenko, *On a Szegő-Widom limit theorem*. Dokl. Akad. Nauk. **393** (2003), 307–308 (in Russian).

[21] H. Widom, *Asymptotic behavior of block Toeplitz matrices and determinants. II*. Advances in Math. **21** (1976), 1–29.

Alexei Yu. Karlovich
Universade do Minho, Centro de Matemática, Escola de Ciências, Campus de Gualtar
4710-057, Braga, Portugal
e-mail: oleksiy@math.uminho.pt
Current address:
Departamento de Matemática, Instituto Superior Técnico, Av. Rovisco Pais 1
1049–001, Lisbon, Portugal
e-mail: akarlov@math.ist.utl.pt

Operator Theory:
Advances and Applications, Vol. 179, 197–210

On an Eigenvalue Problem for Some Nonlinear Transformations of Multi-dimensional Arrays

Sawinder P. Kaur and Israel Koltracht

Abstract. It is shown that certain transformations of multi-dimensional arrays posses unique positive solutions. These transformations are composed of linear components defined in terms of Stieltjes matrices, and semi-linear components similar to $u \to ku^3$. In particular, the analysis of the linear components extends some results of the Perron-Frobenius theory to multi-dimensional arrays.

Mathematics Subject Classification (2000). Primary 47J10; Secondary 15A69, 15A90, 65N06.

Keywords. Nonlinear transformation, finite difference method, monotone operator, Perron-Frobenius theory, Kronecker product.

1. Introduction

In this paper we extend to multi-dimensional arrays, results for one-dimensional arrays presented in paper [2]. In the case of two-dimensional arrays we consider the following nonlinear eigenvalue problem for an $n \times n$ unknown matrix U,

$$AU + UB + F(U) = \lambda U, \tag{1.1}$$

where A and B are $n \times n$ symmetric positive definite, irreducible M-matrices. Such matrices are called Stieltjes matrices, and all entries of their inverses are strictly positive. It follows from the Perron-Frobenius theory that the smallest positive eigenvalue μ of a Stieltjes matrix has multiplicity 1 and that the corresponding eigenvector p has strictly positive entries. The function $F(U)$ is assumed to be "diagonal",

$$F(U) = \begin{bmatrix} f_{11}(u_{11}) & f_{12}(u_{12}) & \dots & f_{1n}(u_{1n}) \\ \vdots & \vdots & & \vdots \\ f_{n1}(u_{n1}) & f_{n2}(u_{n2}) & \dots & f_{nn}(u_{nn}) \end{bmatrix},$$

with the property that $f_{ij}(u_{ij}) > 0$ if $u_{ij} > 0$ and $F(0) = 0$.

Our objective is to characterize all (component-wise) positive solutions $U = [u_{ij}]$ of (1.1).

For the linear part of the operator in (1.1) we use the notation

$$T(U) = AU + UB.$$

It is well known (see e.g. [7]) that eigenvalues of T are all possible sums of eigenvalues of A and B, and that the corresponding eigen-matrices are the outer products of corresponding eigenvectors of A and B, that is, if $Ap^k = \mu^k p^k$ and $Bq^k = \nu^k q^k$, then $T(p^k(q^k)^t) = (\mu^k + \nu^k)p^k(q^k)^t$. In particular, if μ and ν are the smallest eigenvalues of A and B respectively then $\mu + \nu > 0$ is the smallest eigenvalue of T, it is simple and the corresponding eigen-matrix is pq^t. It is not hard to see that $\lambda > \mu + \nu$ is a necessary condition for (1) to have a non-negative eigenmatrix U. Indeed, given an $U \geq 0, U \neq 0$, multiply (1.1) by p^t on the left and by q on the right to get

$$p^t AUq + p^t UBq + \sum_{i=1}^{n}\sum_{j=1}^{n} f_{ij}(x_{ij})p_i q_j = \lambda p^t Uq.$$

Since

$$p^t AUq + p^t UBq = (\mu + \nu)p^t Uq > 0 \quad \text{and} \quad \sum_{i=1}^{n}\sum_{j=1}^{n} f_{ij}(u_{ij})p_i q_j > 0,$$

it follows that $(\mu + \nu) < \lambda$.

We give sufficient conditions on F under which the following result holds:

For any $\lambda > (\mu+\nu)$, there exists a unique positive solution $U(\lambda)$ of equation (1.1). *Moreover, if $(\mu + \nu) < \lambda_1 < \lambda_2$, then $U(\lambda_1) < U(\lambda_2)$*, component-wise, which means that each entry of $U(\lambda_1)$ is less than the corresponding entry of $U(\lambda_2)$.

Partial motivation for this extension comes from a discretization of the Gross-Pitaevskii partial differential equation which models a certain aspect of the Bose-Einstein condensation of matter at near absolute zero temperatures, (see [4] for more details and references). In the case of two spatial variables the Gross-Pitaevskii equation has the form

$$-\Delta u + V(x,y)u + ku^3 = \lambda u, \quad k > 0, \quad u > 0, \tag{1.2}$$

$$\lim_{|(x,y)|\to\infty} u = 0, \quad \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} u(x,y)^2 dxdy = 1,$$

where $u = u(x,y)$ is a function to be found together with λ s.t. the above normalization condition holds, and $\Delta u = u_{xx} + u_{yy}$ denotes the two-dimensional Laplacian. For some λ the solution may not exist. Here we are interested in λ for which there exists a unique positive solution (which corresponds to the stable state of the Bose-Einstein condensate, briefly BEC). The positive constant k of non-linearity is proportional to the number of atoms in the condensate and can be very large. Denoting the linear differential operator $-\Delta + V$ also by T,

$$Tu = -\Delta u(x,y) + V(x,y)u(x,y),$$

we can write (1.2) as

$$Tu + ku^3 = \lambda u.$$

We further assume that the potential $V(x, y)$ is separable,

$$V(x, y) = V_1(x) + V_2(y),$$

where V_1 and V_2 are non-negative functions of one variable. This assumption is there for BEC applications, where

$$V(x, y) = ax^2 + by^2, \quad a > 0,\ b > 0,$$

is the harmonic potential used to create a magnetic trap for BEC by experimentators. Given that $u(x, y)$ converges to zero at infinity, we restrict our equation to a finite domain $[-L, L] \times [-L, L]$. The discretization of second derivatives at points, $x_1, \ldots, x_n$ and $y_1, \ldots, y_n$, leads to the matrix equation

$$D_x U + U D_y + kU^3 = \lambda U,$$

where U is $n \times n$ matrix and U^3 is an $n \times n$ matrix whose entries are third powers of entries of U, and $x_i = y_i = ih, h = \frac{L}{n+1}, i = 1, \ldots, n$, and where

$$D_x = \frac{1}{h^2} \begin{pmatrix} 2 + h^2 V_1(x_1) & -1 & & & 0 \\ -1 & 2 + h^2 V_1(x_2) & -1 & & \\ & -1 & \ddots & \ddots & \\ & & \ddots & 2 + h^2 V_1(x_{n-1}) & -1 \\ 0 & & & -1 & 2 + h^2 V_1(x_n) \end{pmatrix}$$

corresponds to the discretized negative second derivative in x plus the potential $V_1(x)$, and similarly D_y. The matrix D_x is clearly a positive definite irreducible M-matrix.

In this paper we consider a somewhat more general equation

$$A_x U + U A_y + F(U) = \lambda U, \quad k > 0,$$

where A_x and A_y are Stieltjes matrices, or equivalently,

$$T(U) + F(U) = \lambda U.$$

The three-dimensional analog can be described as

$$A_x(U) + A_y(U) + A_z(U) + F(U) = \lambda U \tag{1.3}$$

where U is a triple array,

$$U = \{u_{ijk}\}_{i,j,k=1}^n,$$

and the linear transformation A_x is defined as follows: for a given (j, k), $j, k = 1, \ldots, n$,

$$\{[A_x(U)]_{i,j,k}\}_{i=1}^n = \begin{bmatrix} A_x(U)_{1,j,k} \\ \vdots \\ A_x(U)_{n,j,k} \end{bmatrix} = A \begin{bmatrix} u_{1,j,k} \\ \vdots \\ u_{n,j,k} \end{bmatrix} = \begin{bmatrix} \tilde{u}_{1,j,k} \\ \vdots \\ \tilde{u}_{n,j,k} \end{bmatrix}.$$

Thus $A_x(U)$ is the triple array composed of the entries $\tilde{u}_{i,j,k}, \quad i,j,k = 1,\ldots n$. $A_y(U)$ and $A_z(U)$ are defined similarly. Equation (1.3) includes a discretized version of the Gross-Pitaevskii equation in three spatial variables:

$$-\Delta u + V(x,y,z)u + ku^3 = \lambda u, \quad u > 0, \quad \lim_{|(x,y,z)|\to\infty} u = 0,$$
$$\int_{-\infty}^{\infty}\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} u(x,y,z)^2 dxdydz = 1.$$

Here $u = u(x,y,z), \quad \Delta u = u_{xx} + u_{yy} + u_{zz}$ is the three-dimensional Laplacian and $V(x,y,z) = ax^2 + by^2 + cz^2$.

We also comment on a similar case for more than three variables,

$$T(U) + F(U) = \lambda U, \tag{1.4}$$

where U is an N-dimensional array and $T(U) = A_{(1)}(U) + A_{(2)}(U) + \ldots + A_{(N)}(U)$, and $A_{(i)}(U)$ are defined similar to $A_x(U)$. We observe that T is monotone:

If $T(U) = P, T(V) = Q$ and $P < Q$ (in the componentwise sense), then $U < V$.

Moreover, eigenvalues of T, are all possible sums of eigenvalues of $A_{(1)}$, $A_{(2)}$, $A_{(3)},\ldots,A_{(N)}$ and for $\lambda > \sum_{i=1}^{N}\mu_i$, where for $i = 1,\ldots,N, \quad \mu_i$ are the smallest eigenvalues of $A_{(i)}$, (1.4) has a unique positive solution.

We first consider Equation (1.4) in R^2, where the matrix operations "Vec" and "$\otimes$" are used to extend proofs from [2]. In Section 3 we consider the case of $R^N, N > 2$, with most emphasis on $N = 3$. We extend the three-dimensional array into a vector, and the Kronecker product for three matrices is used (see [5] for the discussion of relevance of Kroneker products to the discretization of a Laplacian in three variables). The conversion of a three-dimensional array into an n^3-vector is not unique, and we give a definition of 'Vec' in three dimensions which allows extension of results for one and two-dimensional arrays to three-dimensional arrays. This definition can be extended to higher-dimensional arrays as well.

Finally we remark that not all discretization techniques lead to a Stieltjes matrix A. In particular, the collocation method based on interpolation at Legendre points gives an A whose inverse has entries with mixed signs. However, numerical experiments for $N = 3$ (see [3]) suggest that also in this case, for $\lambda > (\mu_x + \mu_y + \mu_z)$, Equation (1.4) has a unique positive solution.

2. Two Variable Case

Claim 2.1. *Let A and B be Stieltjes matrices and $T(U) = AU + UB$. Then T is monotone, that is, if $T(U) = P, T(V) = Q$ and $P < Q$, then $U < V$.*

Proof. Consider the Vec-function associated with the matrix $AU + UB$ (see [7, p. 409])

$$\mathrm{Vec}(AU + UB) = [(I_n \otimes A) + (B \otimes I_n)]\mathrm{Vec}(U)$$

where $I_n \otimes A = \mathrm{diag}[A, A, \ldots, A]$ and

$$B \otimes I_n = \begin{bmatrix} b_{11}I_n & b_{12}I_n & \ldots & b_{1n}I_n \\ \vdots & \vdots & & \vdots \\ b_{n1}I_n & b_{n2}I_n & \ldots & b_{nn}I_n \end{bmatrix}.$$

Eigenvalues of $\mathcal{D} = (I_n \otimes A) + (B \otimes I_n)$ are all possible sums of eigenvalues of A and B (see [7, p. 412]). Since A and B are symmetric positive definite it follows that $\mathcal{D}$ is also positive definite. Indeed it is symmetric by construction, and all its eigenvalues are positive.

Since sign pattern of A and B is preserved in $\mathcal{D}$, it follows that $\mathcal{D}$ is an M-matrix itself, and hence monotone. Therefore $T(U) > 0 \Rightarrow \mathrm{Vec}(U) > 0$ and hence $U > 0$. □

The proof of the existence is based on a certain reformulation of (1.1) as a fixed point problem, and on the fundamental result of L.V. Kantorovich, (see [8], [9] or [11] and references therein). Here we present for completeness the statement of the theorem by Kantorovich. By $[y, z]$ we denote the interval $y \le x \le z$ where the inequality is component-wise.

Theorem 2.2. *Let Z be a K-space (R^n in our case) and $V : Z \to Z$ be defined on $[y, z]$ which satisfies the following conditions:*

(i) $y < V(y) < z$.
(ii) $y < V(z) < z$.
(iii) $y \le x_1 < x_2 \le z$ *implies* $y < V(x_1) \le V(x_2) < z$.
(iv) *If* $y \le x_1 \le \ldots \le x_k \le \ldots \le z$ *and* $x_k \uparrow x$, *then* $V(x_k) \uparrow V(x)$.

Then

(a) *the fixed point iteration* $x_k = V(x_{k-1})$ *with* $x_0 = y$ *converges:* $x_k \to x_*$, $V(x_*) = x_*$, $y < x_* < z$;
(b) *the fixed point iteration* $x_k = V(x_{k-1})$ *with* $x_0 = z$ *converges:* $x_k \to x^*$, $V(x^*) = x^*$, $y < x^* < z$;
(c) *if x is a fixed point of V in $[y, z]$, then* $x_* < x < x^*$;
(d) *V has a unique fixed point in $[y, z]$ if and only if* $x_* = x^*$.

(The arrow $\uparrow$ means that x_k monotonically increases and converges to x.)

Theorem 2.3. *Let $(\mu + \nu)$ be the smallest positive eigenvalue of $T(U) = AU + UB$ and pq^t be the corresponding eigen-matrix, here A and B are Stieltjes matrices, μ and ν are the smallest positive eigenvalues and $p = [p_1, \ldots, p_n]^t$ and $q = [q_1, \ldots, q_n]^t$ are the corresponding positive eigenvectors of A and B, respectively. Let $\lambda > (\mu + \nu)$ and*

$$F(U) = \begin{bmatrix} f_{11}(u_{11}) & f_{12}(u_{12}) & \ldots & f_{1n}(u_{1n}) \\ \vdots & \vdots & & \vdots \\ f_{n1}(u_{n1}) & f_{n2}(u_{n2}) & \ldots & f_{nn}(u_{nn}) \end{bmatrix},$$

where, for $i,j = 1,\dots,n$, $f_{ij} : (0,\infty) \to (0,\infty)$ *are* C^1 *functions satisfying the conditions*

$$\lim_{t\to 0} \frac{f_{ij}(t)}{t} = 0, \qquad \lim_{t\to\infty} \frac{f_{ij}(t)}{t} = \infty. \tag{2.1}$$

Then $AU + UB + F(U) = \lambda U$ *has a positive solution. If in addition for* $i,j = 1,\dots,n$,

$$\frac{f_{ij}(s)}{s} < \frac{f_{ij}(t)}{t} \quad \text{whenever} \quad 0 < s < t, \tag{2.2}$$

then the positive solution is unique.

Proof. First take β_1 small enough so that $f_{ij}[(\beta_1 pq^t)_{ij}] < (\lambda - (\mu+\nu))(\beta_1 pq^t)_{ij}$, for $i,j = 1,\dots,n$ and $\beta_2 > \beta_1$ large enough so that $(\lambda - (\mu+\nu))(\beta_2 pq^t)_{ij} < f_{ij}[(\beta_2 pq^t)_{ij}]$ for $i,j = 1,\dots,n$. This is possible because of condition (2.1). Take a positive number $c > 0$ s.t.

$$c > \max_{1\le i,j\le n}\Big[\sup_{(\beta_1 pq^t)_{ij}\le t\le(\beta_2 pq^t)_{ij}} |f'_{ij}(t)|\Big] - \lambda. \tag{2.3}$$

Let $\tilde{T}(U) = (cI + A)U + UB$. Since $(cI + A)U + UB = (c+\lambda)U - F(U)$ and since eigenvalues of $\tilde{T}$ are bounded from below by eigenvalues of T, it follows that $\tilde{T}$ is invertible, and hence

$$U = \tilde{T}^{-1}[(c+\lambda)U - F(U)], \quad \text{or} \quad U = S(U),$$

where by definition $S(U) = \tilde{T}^{-1}[(c+\lambda)U - F(U)]$. To prove existence we show that S satisfies the conditions of Theorem 2.2 with $y = \beta_1 pq^t$ and $z = \beta_2 pq^t$.

For condition (i) of Theorem 2.2, note that $\tilde{T}(pq^t) = (c + (\mu+\nu))pq^t$ and therefore $pq^t = \tilde{T}^{-1}(c + (\mu+\nu))pq^t$.

Now $\tilde{T}^{-1}U > 0$ whenever $U > 0$, therefore it is suffices to show that

$$(c + (\mu+\nu))\beta_1 pq^t \le (c+\lambda)\beta_1 pq^t - F(\beta_1 pq^t)$$

or equivalently that

$$(c + (\mu+\nu))(\beta_1 pq^t)_{ij} \le (c+\lambda)(\beta_1 pq^t)_{ij} - f_{ij}((\beta_1 pq^t)_{ij}),$$

or

$$f_{ij}((\beta_1 pq^t)_{ij}) < (\lambda - (\mu+\nu))(\beta_1 pq^t)_{ij}, \quad i,j = 1,\dots,n.$$

The last inequality follows immediately from the choice of β_1.

In the same way one can show that for the above choice of β_2 one has $S(\beta_2 pq^t) \le \beta_2 pq^t$.

The conditions (iii) and (iv) can be verified in the similar way using condition (2.3) on c and by continuity of (1.1) in U. Now suppose that the condition (2.2) is satisfied. We show that in this case $U_* = U^*$, where U_* and U^* are the solutions of (1.1) obtained by the fixed point iteration starting at $\beta_1 pq^t$ and $\beta_2 pq^t$, respectively.

We have $AU_* + U_*B + F(U_*) = \lambda U_*$, and $AU^* + U^*B + F(U^*) = \lambda U^*$. Indeed,

$$\mathcal{D}\mathrm{Vec}(U_*) + \mathrm{Vec}(F(U_*)) = \lambda \mathrm{Vec}(U_*), \tag{2.4}$$

similarly,

$$\mathcal{D}\mathrm{Vec}(U^*) + \mathrm{Vec}(F(U^*)) = \lambda \mathrm{Vec}(U^*), \tag{2.5}$$

where $\mathcal{D}$ is defined in the proof of Claim 2.1. Since $\mathcal{D}$ is symmetric, pre-multipying (2.4) and (2.5) by $\mathrm{Vec}(U^*)^t$ and $\mathrm{Vec}(U_*)^t$, respectively and subtracting we get $\mathrm{Vec}(U^*)^t\mathrm{Vec}(F(U_*)) = \mathrm{Vec}(U_*)^t\mathrm{Vec}(F(U^*))$, or equivalently that

$$\sum_{i,j=1}^{n} [(u^*)_{ij}(u_*)_{ij}[\frac{f_{ij}((u^*)_{ij})}{(u^*)_{ij}} - \frac{f_{ij}((u_*)_{ij})}{(u_*)_{ij}}]] = 0.$$

Since all the terms are nonnegative the sum is zero only when $(u^*)_{ij} = (u_*)_{ij}$ for all $i, j = 1, \ldots, n$, implying that $U^* = U_*$ □

Claim 2.4. *Let* $(\mu + \nu) < \lambda_1 < \lambda_2 < \infty$, *and* $S_{\lambda_i,c}(U) = \tilde{T}^{-1}[(c + \lambda_i)U - F(U)]$, $i = 1, 2$, *where* $U > 0$. *Then* $S_{\lambda_1,c}(U) < S_{\lambda_2,c}(U)$.

Proof. Since $\tilde{T}^{-1}U > 0$ whenever $U > 0$, it is sufficient to show that

$$(c + \lambda_1)U - F(U) < (c + \lambda_2)U - F(U),$$

or $(c + \lambda_1)U < (c + \lambda_2)U$, which follows immediately from the fact that $U > 0$. Thus $S_{\lambda_1,c}(U) < S_{\lambda_2,c}(U)$. □

Theorem 2.5. *Let conditions* (2.1) *and* (2.2) *of Theorem* 2.3 *be satisfied, and let* $U(\lambda)$ *denote the unique positive eigenmatrix corresponding to* $\lambda\epsilon(\mu + \nu, \infty)$. *Then,*

(1) $U(\lambda_1) < U(\lambda_2)$ *if* $(\mu + \nu) < \lambda_1 < \lambda_2$;
(2) $U(\lambda)$ *is continuous on* $(\mu + \nu, \infty)$;
(3) $\lim_{\lambda\to\infty} U_{ij}(\lambda) = \infty$, $i, j = 1, \ldots, n$;
(4) $\lim_{\lambda\to(\mu+\nu)^+} U_{ij}(\lambda) = 0$, $i, j = 1, \ldots, n$.

Proof. (1) If c is sufficiently large and β' is sufficiently small so that $\beta' pq^t < \min(U(\lambda_1), U(\lambda_2))$, one can start both iterations at the same $\beta' pq^t$ such that

$$U_1^{(1)} = S_{\lambda_1,c}(\beta' pq^t) < S_{\lambda_2,c}(\beta' pq^t) = U_1^{(2)}.$$

Since $S_{\lambda,c}(U)$ satisfies the conditions of Kantorovich Theorem, it follows that $S_{\lambda,c}(U)$ is a monotone function of U, and so we have

$$U_2^{(1)} = S_{\lambda_1,c}(U_1^{(1)}) < S_{\lambda_2,c}(U_1^{(2)}) = U_2^{(2)}$$

$$\vdots$$

$$U_n^{(1)} < U_n^{(2)},$$

and, passing to the limits,

$$U(\lambda_1) < U(\lambda_2).$$

(2) To prove the left continuity let $\lambda_0 > (\mu + \nu)$ and $\{\lambda_k\}$ be a monotone increasing sequence converging to λ_0. Since $\{U(\lambda_k)\}$ is also monotone increasing and bounded by $U(\lambda_0)$, the sequence has a limit.

Suppose $\lim_{k\to\infty} U(\lambda_k) = W$. Since Equation (1.1) is continuous in U and in λ, it follows that the pair λ_0 and W satisfies (1.1). So by uniqueness, $W = U(\lambda_0)$.

Thus $U(\lambda)$ is left continuous. In the same way we can prove the right continuity. Therefore $U(\lambda)$ is continuous in $(\mu+\nu,\infty)$.

(3) Let (i,j), $1 \le i,j \le n$, be an ordered pair. Pre- and post-multiplying the equation

$$AU(\lambda) + U(\lambda)B + F(U(\lambda)) = \lambda U(\lambda)$$

by e_i^T (i-th unit vector) and e_j (j-th unit vector) and using the fact that A and B are M-matrices, we obtain

$$(a_{ii}+b_{jj})u_{ij}(\lambda) + \sum_{k=1,k\neq i}^{n}(a_{ik}u_{kj}(\lambda)) + \sum_{k=1,k\neq j}^{n}(u_{ik}(\lambda)b_{kj}) + f_{ij}(u_{ij}(\lambda)) = \lambda x_{ij}(\lambda).$$

For $\lambda \to \infty$, since $u_{ij}(\lambda)$ is bounded from below for increasing λ, it follows that the right-hand side increases to infinity and hence so does the left-hand side. Since $a_{ik}u_{kj}(\lambda) < 0$ and $u_{ik}(\lambda)b_{kj} < 0, k \neq i, k \neq j$, it follows that if $u_{ij}(\lambda)$ is bounded from above, then by continuity so is $f_{ij}(u_{ij}(\lambda))$, and therefore the left-hand side would be bounded from above, leading to a contradiction. Thus $u_{ij}(\lambda) \to \infty$.

(4) Pre- and post-multiplying (1.1) by p^t and q, we get $p^tAUq + p^tUBq + p^tF(U)q = \lambda p^tUq$, or, $p^tF(U)q = (\lambda - (\mu+\nu))p^tUq$. Since p,q are fixed positive vectors and $U(\lambda)$ are bounded from above for λ approaching $(\mu+\nu)$ from the right, $U(\lambda) > 0$, and $f_{ij} : (0,\infty) \to (0,\infty)$, it follows that

$$F(U(\lambda)) \to 0 \quad \text{as} \quad \lambda \to (\mu+\nu).$$

Since F is continuous and $U(\lambda)$ is monotone decreasing it implies that

$$U(\lambda) \to 0. \qquad \square$$

3. Three variable case

Claim 3.1. *Let $A = [a_{ij}]$, $B = [b_{ij}]$ and $C = [c_{ij}]$ be $n \times n$ irreducible Stieltjes matrices, let I_n denote the $n \times n$ identity matrix and let*

$$\mathcal{M} = A \otimes I_n \otimes I_n + I_n \otimes B \otimes I_n + I_n \otimes I_n \otimes C.$$

Then $\mathcal{M}$ is an $n^3 \times n^3$ irreducible Stieltjes matrix.

Proof. $A \otimes I_n \otimes I_n = A \otimes [I_n \otimes I_n] = A \otimes I_{n^2} = \begin{bmatrix} a_{11}I_{n^2} & a_{12}I_{n^2} & \dots & a_{1n}I_{n^2} \\ \vdots & \vdots & & \vdots \\ a_{n1}I_{n^2} & a_{n2}I_{n^2} & \dots & a_{nn}I_{n^2} \end{bmatrix}$

is an $n^3 \times n^3$ symmetric matrix, as $a_{ij} = a_{ji}$.

$$I_n \otimes B \otimes I_n = I_n \otimes [B \otimes I_n] = I_n \otimes \mathfrak{B}, \text{ where } \mathfrak{B} = \begin{bmatrix} b_{11}I_n & b_{12}I_n & \dots & b_{1n}I_n \\ \vdots & \vdots & & \vdots \\ b_{n1}I_n & b_{n2}I_n & \dots & b_{nn}I_n \end{bmatrix}.$$

$\mathfrak{B}$ is symmetric as $b_{ij} = b_{ji}$. Thus $I_n \otimes B \otimes I_n = diag(\mathfrak{B}, \dots \mathfrak{B})$ is an $n^3 \times n^3$ symmetric matrix. Finally, $I_n \otimes I_n \otimes C = \mathrm{diag}(C, \dots C)$ is an $n^3 \times n^3$ symmetric matrix as C is symmetric.

Since $A \otimes I_{n^2}$ is symmetric and its eigenvalues are just eigenvalues of A with increased multiplicity, it follows that $A \otimes I_n \otimes I_n$ is positive definite as A is positive definite. Similarly $I_n \otimes B \otimes I_n$ and $I_n \otimes I_n \otimes C$ are symmetric positive definite as B and C positive definite. Hence $\mathcal{M}$ is a symmetric positive definite matrix.

Also the sign pattern of A and B and C is preserved in $\mathcal{M} = [m_{ij}]$, $m_{ii} > 0$ and $m_{ij} \leq 0$ for $i \neq j$. Now it remains to show that $\mathcal{M}$ is irreducible. Writing

$$A \otimes I_n \otimes I_n = \begin{bmatrix} A_{11} & A_{12} & \dots & A_{1n} \\ \vdots & \vdots & & \vdots \\ A_{n1} & A_{n2} & \dots & A_{nn} \end{bmatrix}$$

where

$$A_{ij} = \begin{bmatrix} a_{ij} & & & \\ & \ddots & & \\ & & \ddots & \\ & & & a_{ij} \end{bmatrix},$$

we proceed by contradiction. Suppose A is reducible, then there exists a permutation matrix P such that

$$P^T(A \otimes I \otimes I)P = \begin{bmatrix} \tilde{B} & \tilde{C} \\ 0 & \tilde{D} \end{bmatrix}$$

$$= \left[\begin{array}{cccc|ccc} \tilde{A}_{11} & \tilde{A}_{12} & \dots & \tilde{A}_{1r} & \tilde{A}_{1r+1} & \dots & \tilde{A}_{1n} \\ \vdots & \vdots & & \vdots & \vdots & & \vdots \\ \tilde{A}_{r1} & \tilde{A}_{r2} & \dots & \tilde{A}_{rr} & \tilde{A}_{rr+1} & \dots & \tilde{A}_{rn} \\ \hline \tilde{A}_{r+1,1} & \tilde{A}_{r+1,2} & \dots & \tilde{A}_{r+1,r} & \tilde{A}_{r+1r+1} \dots & & \tilde{A}_{r+1,n} \\ \vdots & \vdots & & \vdots & \vdots & & \vdots \\ \tilde{A}_{n,1} & \tilde{A}_{n,2} & \dots & \tilde{A}_{n,r} & \tilde{A}_{nr+1} & & \tilde{A}_{nn} \end{array}\right]$$

where $\tilde{B}$ is an $r \times r$, $\tilde{D}$ is an $n-r \times n-r$, $\tilde{C}$ is an $r \times n-r$ matrix, 0 is an $n-r \times r$ zero matrix, and $\tilde{A_{ij}}$ are the elements of $(A \otimes I \otimes I)$ after permutation.

It is clear that if $(A \otimes I_n \otimes I_n)$ is reducible, then $\tilde{A_{ij}} = 0$ for $i = r+1, \dots n$ and $j = 1 \dots r$, which is a contradiction as $\tilde{A_{ij}}$ are diagonal matrices whose diagonal elements are the off diagonal elements a_{ij} of the matrix A which can not be equal to zero as A is an irreducible matrix.

Thus $A \otimes I_n \otimes I_n$ is irreducible. $I_n \otimes B \otimes I_n$, and $I \otimes I \otimes C$ are block diagonal matrices, where every block has positive main diagonal elements and non-positive off diagonal elements, it follows that $A \otimes I_n \otimes I_n + I_n \otimes B \otimes I_n + I_n \otimes I_n \otimes C$ is irreducible as well. □

Claim 3.2. *Let A, B and C be $n \times n$ irreducible Stieltjes matrices, $T(U) = A_x(U) + B_y(U) + C_z(U)$, where U is a triple array,*

$$U = \{u_{ijk}\}_{i,j,k=1}^{n},$$

and the linear transformation A_x is defined as follows: for a given (j,k), $j,k = 1,\ldots,n$,

$$\{[A_x(U)]_{i,j,k}\}_{i=1}^{n} = \begin{bmatrix} A_x(U)_{1,j,k} \\ \vdots \\ A_x(U)_{n,j,k} \end{bmatrix} = A \begin{bmatrix} u_{1,j,k} \\ \vdots \\ u_{n,j,k} \end{bmatrix} = \begin{bmatrix} \tilde{u}_{1,j,k} \\ \vdots \\ \tilde{u}_{n,j,k} \end{bmatrix},$$

and $B_y(U)$ and $C_z(U)$ are defined similarly. Then T is monotone.

Proof. To prove the claim we expand the three-dimensional array into an n^3-vector in the following way. In $U = \{u_{ijk}\}_{i,j,k=1}^{n}$ we first expand for index i, then corresponding to each fixed i write the expansion for the index j, and then for each fixed j expand for k. For example, for $i,j,k = 1,2$,

$$\mathrm{Vec}(U) = \mathrm{Vec}([u_{ijk}]_{ijk=1}^{2}) = \begin{bmatrix} u_{1jk} \\ u_{2jk} \end{bmatrix} = \begin{bmatrix} u_{11k} \\ u_{12k} \\ u_{21k} \\ u_{22k} \end{bmatrix} = \begin{bmatrix} u_{111} \\ u_{112} \\ u_{121} \\ u_{122} \\ u_{211} \\ u_{212} \\ u_{221} \\ u_{222} \end{bmatrix}.$$

In general,

$$\mathrm{Vec}(U) = \begin{bmatrix} u_{111} \\ \vdots \\ u_{11n} \\ \vdots \\ u_{1n1} \\ \vdots \\ u_{1nn} \\ u_{211} \\ \vdots \\ u_{2nn} \\ \vdots \\ u_{nn1} \\ \vdots \\ u_{nnn} \end{bmatrix}.$$

Now, to prove that $T(U)$ is monotone, we first show that

$$\mathrm{Vec}(A_x(U) + B_y(U) + C_z(U)) = (A \otimes I_n \otimes I_n + I_n \otimes B \otimes I_n + I_n \otimes I_n \otimes C)\mathrm{Vec}(U).$$

It is clear that $\mathrm{Vec}(U)$ defined above is linear, so it is sufficient to show that $\mathrm{Vec}(A_x(U)) = (A \otimes I_n \otimes I_n)\mathrm{Vec}(U)$, $\mathrm{Vec}(B_y(U)) = (I_n \otimes B \otimes I_n)\mathrm{Vec}(U)$, and $\mathrm{Vec}(C_z(U)) = (I_n \otimes I_n \otimes C)\mathrm{Vec}(U)$.

Here we show the above result for $i, j, k = 1, 2$, the general case can be proved in a similar way. Let

$$A = \begin{bmatrix} a_{11} & -a_{12} \\ -a_{12} & a_{22} \end{bmatrix}, \qquad a_{ij} > 0,\ i, j = 1, 2,$$

then

$$\begin{aligned} A_x(U) &= ([A_x(U)]_{ijk})_{ijk=1}^2 = \left(\begin{bmatrix} (A_x(U))_{1jk} \\ (A_x(U))_{2jk} \end{bmatrix} \right)_{j,k=1}^2 = \left(A \begin{bmatrix} u_{1jk} \\ u_{2jk} \end{bmatrix} \right)_{j,k=1}^2 \\ &= \left(\begin{bmatrix} a_{11}u_{1jk} - a_{12}u_{2jk} \\ -a_{12}u_{1jk} + a_{22}u_{2jk} \end{bmatrix} \right)_{j,k=1}^2 = \left(\begin{bmatrix} \hat{u}_{1jk} \\ \hat{u}_{2jk} \end{bmatrix} \right)_{j,k=1}^2, \end{aligned}$$

$$\mathrm{Vec}(A_x(U)) = \mathrm{Vec}\left(\begin{bmatrix} \hat{u}_{1jk} \\ \hat{u}_{2jk} \end{bmatrix} \right)_{j,k=1}^2 = \begin{bmatrix} a_{11}u_{111} - a_{12}u_{211} \\ a_{11}u_{112} - a_{12}u_{212} \\ a_{11}u_{121} - a_{12}u_{221} \\ a_{11}u_{122} - a_{12}u_{222} \\ -a_{12}u_{111} + a_{22}u_{211} \\ -a_{12}u_{112} + a_{22}u_{212} \\ -a_{12}u_{121} + a_{22}u_{221} \\ -a_{12}u_{122} + a_{22}u_{222} \end{bmatrix}$$

$$= \left[\begin{array}{cccc|cccc} a_{11} & & & & -a_{12} & & & \\ & \ddots & & & & \ddots & & \\ & & \ddots & & & & \ddots & \\ & & & a_{11} & & & & -a_{12} \\ \hline -a_{12} & & & & a_{22} & & & \\ & \ddots & & & & \ddots & & \\ & & \ddots & & & & \ddots & \\ & & & -a_{12} & & & & a_{22} \end{array}\right] \begin{bmatrix} u_{111} \\ u_{112} \\ u_{121} \\ u_{122} \\ u_{211} \\ u_{212} \\ u_{221} \\ u_{222} \end{bmatrix}$$

$$= (A \otimes I_n \otimes I_n)\mathrm{Vec}(U).$$

Let

$$B = \begin{bmatrix} b_{11} & -b_{12} \\ -b_{12} & b_{22} \end{bmatrix}, \qquad b_{ij} > 0,\ i, j = 1, 2,$$

then

$$\begin{aligned} B_y(U) &= ([B_y(U)]_{ijk})_{ijk=1}^2 = \left(\begin{bmatrix} (B_y(U))_{i1k} \\ (B_y(U))_{i2k} \end{bmatrix} \right)_{i,k=1}^2 = \left(B \begin{bmatrix} u_{i1k} \\ u_{i2k} \end{bmatrix} \right)_{i,k=1}^2 \\ &= \left(\begin{bmatrix} b_{11}u_{i1k} - b_{12}u_{i2k} \\ -b_{12}u_{i1k} + b_{22}u_{i2k} \end{bmatrix} \right)_{i,k=1}^2, \end{aligned}$$

$$\mathrm{Vec}(B_y(U)) = \mathrm{Vec}\left(\begin{bmatrix} b_{11}u_{i1k} - b_{12}u_{i2k} \\ -b_{12}u_{i1k} + b_{22}u_{i2k} \end{bmatrix}\right)_{i,k=1}^{2} = \begin{bmatrix} b_{11}u_{111} - b_{12}u_{121} \\ b_{11}u_{112} - b_{12}u_{122} \\ -b_{12}u_{111} + b_{22}u_{121} \\ -b_{12}u_{112} + b_{22}u_{122} \\ b_{11}u_{211} - b_{12}u_{221} \\ b_{11}u_{212} - b_{12}u_{222} \\ -b_{12}u_{211} + b_{22}u_{221} \\ -b_{12}u_{212} + b_{22}u_{222} \end{bmatrix}$$

$$= \left[\begin{array}{cccc|cccc} b_{11} & 0 & -b_{12} & 0 & & & & \\ 0 & b_{11} & 0 & -b_{12} & & & & \\ -b_{12} & 0 & b_{22} & 0 & & & & \\ 0 & -b_{12} & 0 & b_{22} & & & & \\ \hline & & & & b_{11} & 0 & -b_{12} & 0 \\ & & & & 0 & b_{11} & 0 & -b_{12} \\ & & & & -b_{12} & 0 & b_{22} & 0 \\ & & & & 0 & -b_{12} & 0 & b_{22} \end{array}\right] \begin{bmatrix} u_{111} \\ u_{112} \\ u_{121} \\ u_{122} \\ u_{211} \\ u_{212} \\ u_{221} \\ u_{222} \end{bmatrix}$$

$$= (I_n \otimes B \otimes I_n)\mathrm{Vec}(U).$$

Similarly $C_z(U)) = (I_n \otimes I_n \otimes C)\mathrm{Vec}(U)$. Let $\mathcal{M} = A \otimes I \otimes I + I \otimes B \otimes I + I \otimes I \otimes C$, then $\mathrm{Vec}(T(U) = \mathrm{Vec}(A_x(U) + B_y(U) + C_z(U)) = \mathcal{M}\mathrm{Vec}(U)$. Now

$$T(U) > 0 \Rightarrow \mathcal{M}\mathrm{Vec}(U) > 0.$$

Since from claim (3.1) $\mathcal{M}$ is a Stieltjes matrix and hence monotone, it follows that $\mathrm{Vec}(U) > 0$ and hence $U > 0$. □

The proofs of Theorems 3.3 and 3.4 below are straightforward extensions, without any significant changes, of the proofs of Theorems 2.3 and 2.5, and therefore are omitted here.

Theorem 3.3. *Let $(\mu + \nu + \sigma)$ be the smallest positive eigenvalue of $T = A_x(U) + B_y(U) + C_z(U)$ and $V = [v_{ijk}] = [p_i q_j r_k]$ be the corresponding eigenarray; here A, B and C are Stieltjes matrices, μ, ν and σ are the smallest positive eigenvalues and p, q, and r are the corresponding positive eigenvectors of A, B and C, respectively. Let $\lambda > (\mu + \nu + \sigma)$ and*

$$F(U) = [f_{ijk}(u_{ijk})],$$

where for $i, j, k = 1, \ldots, n$, $f_{ijk} : (0, \infty) \to (0, \infty)$ are C^1 functions satisfying the conditions

$$\lim_{t \to 0} \frac{f_{ijk}(t)}{t} = 0, \qquad \lim_{t \to \infty} \frac{f_{ijk}(t)}{t} = \infty. \tag{3.1}$$

Then $A_x(U) + B_y(U) + C_z(U) + F(U) = \lambda U$ has a positive solution. If in addition for $i, j, k = 1, \ldots, n$,

$$\frac{f_{ijk}(s)}{s} < \frac{f_{ijk}(t)}{t} \quad \text{whenever} \quad 0 < s < t, \tag{3.2}$$

then the positive solution is unique.

Theorem 3.4. *Let conditions* (3.1) *and* (3.2) *of Theorem* 3.3 *be satisfied, and let* $W(\lambda)$ *denote the unique positive eigenvector corresponding to* $\lambda\epsilon(\mu+\nu+\sigma,\infty)$. *Then,*

(1) $W(\lambda_1) < W(\lambda_2)$ *if* $(\mu+\nu+\sigma) < \lambda_1 < \lambda_2$;
(2) $W(\lambda)$ *is continuous on* $(\mu+\nu+\sigma,\infty)$;
(3) $\lim_{\lambda\to\infty} w_{ijk}(\lambda) = \infty$, $i,j,k = 1,\ldots,n$;
(4) $\lim_{\lambda\to(\mu+\nu+\sigma)^+} w_{ijk}(\lambda) = 0$, $i,j,k = 1,\ldots,n$.

N-variable case, $N > 3$

Analogous results can be obtained in a way similar to the case of $N = 3$. They are not presented here due to the cumbersome book-keeping.

Acknowledgment

We would like to thank Prof. Y.S. Choi for useful suggestions.

References

[1] R.L. Burden, J. Faires and Douglas, *Numerical Analysis*, 8th edition, Thomson Books, 2005.

[2] Y.S. Choi, I. Koltracht and P.J. McKenna, A Generalization of the Perron-Frobenius theorem for non-linear perturbations of Stiltjes Matrices, *Contemporary Mathematics* **281**, 2001.

[3] Y.S. Choi, J. Javanainen, I. Koltracht, M. Koštrum, P.J. McKenna and N. Savytska, N., A fast algorithm for the solution of the time-independent Gross-Pitaevskii equation, *Journal of Computational Physics* **190** (2003), 1–21.

[4] F. Dalfovo, S. Giorgini, L.P. Pitaevskii and S. Stringari, Theory of Bose-Einstein condensation in trapped gases, *Reviews of Modern Physics* **71** no. 3, April 1999.

[5] J.W. Demmel, *Applied Numerical Linear Algebra*, Siam, 1997.

[6] C.U. Huy, P.J. McKenna and W. Walter, Finite Difference Approximations to the Dirichlet Problem for Elliptic Systems, *Numer. Math.* **49** (1986), 227–237.

[7] P. Lancaster and M. Tismenetsky, *The Theory of Matrices, Second Edition with Applications*, Academic Press, 1985.

[8] L.V. Kantorovich, *Selected Works*, Amsterdam, the Netherlands, Gordon and Breach Pub., 1996.

[9] L.V. Kantorovich, B.Z. Vulikh B.Z. and A.G. Pinsker, *Funkzionalnyi Analiz v Poluuporiadochennych Prostranstvach*, Moscow, GosIzdat Techniko-Teoreticheskoi Literatury, 1950, (in Russian).

[10] P. Nozieres and D. Pines, *The Theory of Quantum Liquids*, Vol. II, Redwood City, CA, Addison-Wesley, 1990.

[11] J.M. Ortega and W.C. Rheinboldt, *Iterative Solution of Nonlinear Equations*, Academic Press, 1970.

[12] J. Stoer and R. Bulrisch, *Introduction to Numerical Analysis*, Springer-Verlag New York, 1980.

Sawinder P. Kaur and Israel Koltracht
Department of Mathematics
University of Connecticut
196 Auditorium Road
Storrs, CT 06269
USA
e-mail: kaur@math.uconn.edu
koltrachht@math.uconn.edu

Operator Theory:
Advances and Applications, Vol. 179, 211–227

On Embedding of the Bratteli Diagram into a Surface

Igor V. Nikolaev

Abstract. We study C^*-algebras $\mathcal{O}_\lambda$ which arise in dynamics of the interval exchange transformations and measured foliations on compact surfaces. Using Koebe-Morse coding of geodesic lines, we establish a bijection between Bratteli diagrams of such algebras and measured foliations. This approach allows us to apply K-theory of operator algebras to prove a strict ergodicity criterion and Keane's conjecture for the interval exchange transformations.

Mathematics Subject Classification (2000). Primary 46L40; Secondary 57R30.

Keywords. K-theory, dimension group, measured foliation.

Introduction

Let $\lambda = (\lambda_1, \ldots, \lambda_n)$ be a partition of the unit interval into a disjoint union of open subintervals. Let $\varphi : [0,1] \to [0,1]$ be an interval exchange transformation (with flips). Consider a unital C^*-algebra $\mathcal{O}_\lambda$ generated by the unitary operator $u(\zeta) = \zeta \circ \varphi^{-1}$ and characteristic operators $\chi_{\lambda_1}, \ldots, \chi_{\lambda_n}$ in the Hilbert space $L^2([0,1])$. This (noncommutative) C^*-algebra has an amazingly rich geometry.

$\mathcal{O}_\lambda$ is Morita equivalent to a groupoid C^*-algebra corresponding to measured foliations on a compact surface of genus greater than one. To this end, $\mathcal{O}_\lambda$ is an extension of the irrational rotation algebra A_θ whose theory experienced an extensive development in the last decades.

$\mathcal{O}_\lambda$ is closely related to simple C^*-algebras of minimal homeomorphisms on a Cantor set. These C^*-algebras were in focus of a brilliant series of works of I. F. Putnam starting with the papers [10], [11]. We refer the reader to our work [9] for discussion of connections between Putnam's algebras and $\mathcal{O}_\lambda$.

The K-groups of $\mathcal{O}_\lambda$ are finitely generated and can be obtained from the Pimsner-Voiculescu diagram for the crossed products. Namely, $K_0(\mathcal{O}_\lambda) = \mathbb{Z}^n$, $K_1(\mathcal{O}_\lambda) = \mathbb{Z}$, where n is the number of intervals in the partition of $[0,1]$. The

The work was partially supported by the NSERC grant RT733895.

dimension group $(K_0, K_0^+, [1])$ of $\mathcal{O}_\lambda$ was calculated in [9]. (The reader is referred to the appendix for the details of this construction.) When φ is minimal, the dimension group $(K_0, K_0^+, [1])$ is simple.

Recall that the state on dimension group is a positive homomorphism of $(K_0, K_0^+, [1])$ to $\mathbb{R}$ which respects the order units [1] and $1 \in \mathbb{R}$. The state space $S_\bullet$ of $(K_0, K_0^+, [1])$ is a Choquet simplex of dimension $\leq n-1$. The dimension of $S_\bullet$ is equal to the number of linearly independent invariant ergodic measures of φ. Each invariant measure corresponds to a 1-dimensional linear subspace of the state space, and $\dim S_\bullet = 1$ if and only if the interval exchange transformation φ is strictly (uniquely) ergodic.

It might be one of the most intriguing problems of topological dynamics since 25 years to indicate conditions of strict ergodicity of φ. Some results in this direction are due to Veech and Boshernitzan. In 1975 Keane conjectured that "typically" φ is strictly ergodic. Masur [6] and Veech [15] proved this conjecture in positive using methods of complex analysis and topological dynamics, respectively.

This note is an attempt to study dynamics of φ using the ideas and methods of operator algebras. A foundation to such an approach is given by the following main theorem (to be proved in Section 2):

Theorem 0.1. *Let $n \geq 2$ be an integer. Let $(P, P_+, [u])$ be a simple and totally ordered* [1] *dimension group of order $n \geq 2$. Then there exists an interval exchange transformation $\varphi = \varphi(\lambda, \pi, \varepsilon)$ of n intervals and a C^*-algebra $\mathcal{O}_\lambda$ with the group $(K_0, K_0^+, [1])$ which is order-isomorphic to $(P, P_+, [u])$. The transformation φ is minimal.*

The proof of the above theorem is based on the identification of the infinite paths of Bratteli diagram with the symbolic geodesics on a compact surface (so-called Koebe-Morse theory). This method has an independent interest since it provides direct links between geometry of geodesics and K-theory of operator algebras.

The paper is divided into five sections. In Section 1 we introduce notation and a lemma on positive cones in $K_0(\mathcal{O}_\lambda)$. In Section 2 we give the proof of main theorem. In Sections 3 and 4 we apply Theorem 0.1 to establish a strict ergodicity criterion and Keane's Conjecture, respectively. Section 5 is an Appendix containing quick review of dynamics of the interval exchanges, measured foliations, K-theory and rotation numbers associated to the C^*-algebra $\mathcal{O}_\lambda$. The reader is encouraged to read the "Conclusions and open problems" section at the end of this paper.

1. Notation

Let A be a unital C^*-algebra and $V(A)$ be the union (over n) of projections in the $n \times n$ matrix C^*-algebra with entries in A. Projections $p, q \in V(A)$ are equivalent

[1] The total ordering condition ensures that the Unimodular Conjecture is true, see Effros [2] and Elliott [3]. The author believes the condition is technical, but cannot drop it at this point.

if there exists a partial isometry u such that $p = u^*u$ and $q = uu^*$. The equivalence class of projection p is denoted by $[p]$.

Equivalence classes of orthogonal projections can be made to a semigroup by putting $[p] + [q] = [p + q]$. The Grothendieck completion of this semigroup to an abelian group is called a K_0-*group of algebra* A.

Functor $A \to K_0(A)$ maps a category of unital C^*-algebras into the category of abelian groups so that projections in algebra A correspond to a "positive cone" $K_0^+ \subset K_0(A)$ and the unit element $1 \in A$ corresponds to an "order unit" $[1] \in K_0(A)$. The ordered abelian group $(K_0, K_0^+, [1])$ with an order unit is called a *dimension (Elliott) group* of C^*-algebra A.

For the C^*-algebra $\mathcal{O}_\lambda$ one easily finds that $K_0(\mathcal{O}_\lambda) = \mathbb{Z}^n$, see the Appendix. It is harder to figure out the positive cone $K_0^+(\mathcal{O}_\lambda)$. The rest of the section is devoted to this specific question.

Let us fix the following notation:

$\mathbb{H}$	Lobachevsky complex half-plane $\{z = x + iy \mid y > 0\}$ endowed with the hyperbolic metric $ds = \lvert dz\rvert/y$;
$\partial\mathbb{H}$	absolute, i.e., line $y = 0$ of the Lobachevsky half-plane;
G	Fuchsian group of the first kind;
$M_{g,m}$	orientable surface of genus g with m boundary components;
$\mathcal{F}$	measured foliation of M_n obtained as suspension over interval exchange transformation $\varphi = \varphi(\lambda, \pi, \varepsilon)$ with n intervals;
Λ	geodesic lamination corresponding to $\mathcal{F}$;
γ	geodesic "generating" Λ, i.e. $\bar{\gamma} = \Lambda$.

Thurston has shown that each measured foliation $\mathcal{F}$ can be represented by a "geodesic lamination" Λ consisting of disjoint non-periodic geodesics, which lie in the closure of any of them; cf. Thurston [13]. Denote by $p : \mathbb{H} \to M_{g,m}$ a covering mapping corresponding to the action of a discrete group G.

The geodesic lamination Λ is a product $K \times \mathbb{R} \subset M_{g,m}$, where K is a (linear) Cantor set. The preimage $p^{-1}(\Lambda) \subset \mathbb{H}$ is a collection of geodesic half-circles without self-intersections except, possibly, at the absolute. The "footpoints" of these half-circles is a subset of $\partial\mathbb{H}$ homeomorphic to K.

Fix a Riemann surface $M_{g,m} = \mathbb{H}/G$ of genus g together with a point $p \in M_{g,m}$. Let γ be a "generating" geodesic of the lamination Λ, i.e. such that closure $\bar{\gamma} = \Lambda$. Consider the set

$$Sp\ (\gamma) = \{\gamma_0, \gamma_1, \gamma_2, \dots\} \tag{1}$$

of periodic geodesics γ_i based in p, which monotonically approximate γ in terms of "length" and "direction". The set $Sp\ (\gamma)$ is known as *spectrum* of γ and is defined uniquely upon γ.

Let $n=2g+m-1$. Then the (relative) integral homology $H_1(M_{g,m}, \partial M_{g,m}; \mathbb{Z}) \cong \mathbb{Z}^n$. Since each γ_i is a 1-cycle, there is an injective map $f : Sp\ (\gamma) \to H_1(M_{g,m}, \partial M_{g,m}; \mathbb{Z})$, which relates every closed geodesic its homology class. Note that $f(\gamma_i) = p_i \in \mathbb{Z}^n$ is "prime" in the sense that it is not an integer multiple of some other point of lattice $\mathbb{Z}^n$. Denote by $Sp_f(\gamma)$ the image of $Sp\ (\gamma)$ under the mapping f. Finally, let $SL(n, \mathbb{Z})$ be the group of $n \times n$ integral matrices of determinant 1 and $SL(n, \mathbb{Z}^+)$ its semigroup consisting of matrices with strictly positive entries. It is not hard to show, that in an appropriate basis in $H_1(M_{g,m}, \partial M_{g,m}; \mathbb{Z})$ the following is true:

(i) the coordinates of vectors p_i are non-negative;
(ii) there exists a matrix $A_i \in SL(n, \mathbb{Z}^+)$ such that $p_i = A_i(p_{i-1})$ for any pair of vectors p_{i-1}, p_i in $Sp_f(\gamma)$.

Definition 1.1. The ordered abelian group $(\mathbb{Z}^n, (\mathbb{Z}^n)^+, [1])$ defined as inductive limit of simplicially ordered groups:

$$\mathbb{Z}^n \xrightarrow{\mathrm{A}_1} \mathbb{Z}^n \xrightarrow{\mathrm{A}_2} \mathbb{Z}^n \xrightarrow{\mathrm{A}_3} \dots, \tag{2}$$

is called associated to the geodesic γ.

(We have shown in [9] that the order structure on $(\mathbb{Z}^n, (\mathbb{Z}^n)^+, [1])$ is independent of the choice of $M_{g,m}$ and γ.)

Lemma 1.2. *The dimension group $(K_0, K_0^+, [1])$ of the C^*-algebra $\mathcal{O}_\lambda$ is order-isomorphic to the associated group $(\mathbb{Z}^n, (\mathbb{Z}^n)^+, [1])$ of Definition* 1.1.

Proof. See [9]. □

2. Proof of Theorem 0.1

Let us outline main idea of the proof. To every dimension group $(P, P_+, [u])$ with $P \simeq \mathbb{Z}^n$ one can relate a Bratteli diagram (V, E). The path space X of (V, E) can be made a topological space by putting two paths "close" if and only if they coincide at the initial steps. (X is called Bratteli-Cantor compactum.) X can be embedded (as topological space) into the complex plane $\mathbb{H}$ by identification of each $x \in X$ with a geodesic in $\mathbb{H}$ via Morse coding of the geodesic lines. We show that $X = p^{-1}(\Lambda)$, where Λ is Thurston's geodesic lamination on the surface $M_n = \mathbb{H}/G$; cf. Thurston [13]. A concluding step is to recover $\mathcal{F}$ and φ from Λ.

Let $(P, P_+, [u])$ be a simple totally ordered dimension group with $P \simeq \mathbb{Z}^n$. Recall that a Bratteli diagram of $(P, P_+, [u])$ consists of a vertex set V and edge set E such that V is an infinite disjoint union $V_1 \sqcup V_2 \sqcup \dots$, where each V_i has cardinality n. The latter condition follows from the total ordering of $\mathbb{Z}^n$. Any pair V_{i-1}, V_i defines a non-empty set $E_i \subset E$ of edges with a pair of range and source functions r, s such that $r(E_i) \subseteq V_i$ and $s(E_i) \subseteq V_{i-1}$.

An *AF* C^*-algebra whose dimension group is order-isomorphic to $(P, P_+, [u])$ is an inductive limit of multi-matrix algebras

$$\lim M_{J_1}(\mathbb{C}) \oplus \dots \oplus M_{J_n}(\mathbb{C}).$$

We shall say that a Bratteli diagram (V, E) *corresponds to* the group $(P, P_+, [u])$ if the range and source functions of (V, E) represent the embedding scheme of the above multi-matrix algebras. (In other words, an AF-algebra defined by (V, E) has Elliott group $(P, P_+, [u])$.)

The equivalence class of Bratteli diagrams corresponding to a simple totally ordered dimension group of form $\mathbb{Z}^n$ has a representative (V, E) with no multiple edges, since every positive integral matrix decomposes into a finite product of non-negative matrices whose entries are zeros and ones. For the sake of simplicity, we always assume this case of Bratteli diagrams.

By an *infinite path* on (V, E) we shall mean an infinite sequence of edges $(e_0, e_1, \dots)$ such that $e_0 \in E_0, e_1 \in E_1$, etc. The set of all infinite paths on (V, E) is denoted by X. Let us identify "coordinates" x_i of $x \in X$ with vector $(e_0, e_1, \dots)$. Fix $x, y \in X$. This metric $d(x, y) = 1/2^k$, where

$$k = \max\{l \in \mathbb{N} \mid x_i = y_i \text{ for } i < l\},$$

turns X into an absolutely disconnected topological space which is called a *Bratteli-Cantor compactum*. To construct an embedding $X \to \mathbb{H}$ where each $x \in X$ represents a geodesic, a portion of symbolic dynamics is needed.

Koebe-Morse coding of geodesics. Let $M_{g,m}$ be a hyperbolic surface of genus g with m totally geodesic boundary components $v_1, \dots, v_m$. We dissect $M_{g,m}$ to a simply connected surface as follows [7]. Let P be an arbitrary point of v_m. One draws geodesic segments $h_1, \dots, h_{m-1}$ from P to some arbitrarily chosen points of $v_1, \dots, v_{m-1}$. (Thus, the h_i have only P as common point.) Next one dissects the handles of $M_{g,m}$ by closed geodesics $c_1, \dots, c_{2g}$ issued from point P. Clearly, the resulting surface is simply connected and has the boundary

$$c_1, \dots, c_{2g}; h_1, \dots, h_{m-1}. \tag{3}$$

Now given a geodesic half-circle $S \subset \mathbb{H}$ passing through the unique point $0 \in \tau$ one relates an infinite sequence of symbols

$$\sigma_1, \sigma_2, \sigma_3, \dots, \tag{4}$$

which "take values" in the set σ. One prescribes σ_p, $p = 1, \dots, \infty$ a "value" g_i, $1 \le i \le n$ if and only if S has a transversal intersection point with the side $a_i = b_i$ of p-th image of $G_\sigma \in \tau$. (In other words, code (4) "counts" points of intersection of S with "sides" of tessellation τ.) A sequence of symbols (4) is called a *Koebe-Morse code* of the geodesic S.

Morse showed that there is a bijective correspondence between sequences (4) satisfying some admissibility requirements[2] and the set of non-periodic geodesics on surfaces of negative curvature; see the bibliography to Morse and Hedlund [8].

Lemma 2.1. *Let S be a geodesic with the Koebe-Morse code $(\sigma_1, \sigma_2, \dots)$. Then any congruent to S geodesic S' will have the same Koebe-Morse code, except possibly in a finite number of terms.*

[2]Namely, there should be no words with the syllabi $a_i b_i$ or $b_i a_i$, where a_i and b_i are "dual" symbols from the alphabet G_σ.

Proof. This follows from the definition of coding and invariance of τ by the G-actions. □

Let X be a Bratteli-Cantor compactum. Let $V_i \to \sigma$ be a bijection between the vertices $V = V_1 \sqcup V_2 \sqcup \dots$ of (V, E) and the set of symbols σ. This bijection can be established by labeling each element of V_i from the left to the right by symbols $\{g_1, \dots, g_n\}$. Thus, every $x \in X$ is a "symbolic geodesic" $(x_1, x_2, \dots)$ whose "coordinates" take values in σ. Each sequence is admissible and by Morse's Theorem realized by a (class of congruent) geodesic whose Koebe-Morse code coincides with $(x_1, x_2, \dots)$. Where there is no confusion, we refer to $x \in X$ as a geodesic line in the complex plane $\mathbb{H}$.

Lemma 2.2. *Let l_x be an image of geodesic $x \in X$ on the surface $M_{g,m}$ under projection $\mathbb{H} \to \mathbb{H}/G$. If the Bratteli diagram (V, E) is simple, then $l_y \in Clos\ l_x$ for any $y \in X$.*

Proof. The simplicity of (V, E) means that every infinite path $x \in X$ is transitive, i.e. any finite "block" of symbols $\{x_n, x_{n+1}, \dots, x_{n+k}\}$ occurs "infinitely many times" in the sequence $x = (x_1, x_2, \dots)$. Indeed, simplicity of (V, E) means the absence of non-trivial ideals in the corresponding AF C^*-algebra. Using Bratteli's dictionary [1] between ideals and connectedness properties of (V, E), it can be easily shown that an arbitrary infinite path in (V, E) "visits" any given finite sequence of vertices infinitely often.

Suppose that B_k is a block of symbols of length $k \geq 1$. Let

$$x = (x_1, \dots, x_{n-1}, B_k, x_{n+k+1}, \dots), \qquad y = (y_1, \dots, y_{m-1}, B_k, y_{m+k+1}, \dots)$$

be the first time B_k appears in sequences $x, y \in X$. By a congruent transformation, the geodesics $x, y \in \mathbb{H}$ can be brought to the from

$$x' = (B_k, x_{k+1}, \dots), \qquad y' = (B_k, y_{k+1}, \dots).$$

This means that $dist\,(x', y') \leq 1/2^k$. Since B_k occurs in sequences x and y infinitely often. The lemma follows. □

Lemma 2.3. *The set $Clos\ l_x$ of Lemma 2.2 is a set $\Lambda \subset M_{g,m}$ consisting of continuum of irrational geodesic lines.*

Proof. This follows from the proof of Lemma 2.2. □

By Lemmas 2.2 and 2.3, Λ is homeomorphic to Thurston's geodesic lamination on a surface of genus $g \geq 2$; cf. Thurston [13]. To finish the proof of the theorem, one needs to "blow-down" Λ to a measured foliation $\mathbb{F}$. The required interval exchange transformation φ is the "mapping of first return" on a global transversal to $\mathbb{F}$. By the construction, φ is minimal and has $n = 2g+m-1$ intervals of continuity.

Theorem 0.1 is proven. □

3. Criterion of strict ergodicity for the interval exchange transformations

In general, the group $(K_0, K_0^+, [1])$ may be *not* totally ordered. The total order happens if and only if positive cone K_0^+ is bounded by a unique "hyperplane" in the "space" K_0. There exists up to $(n-2)$ hyperplanes in a group K_0 of rank n which constitute a boundary of K_0^+; cf Goodearl [4], p. 217.

A *state* is a homomorphism f from $(K_0, K_0^+, [1])$ to $\mathbb{R}$ such that $f(K_0^+) \subset \mathbb{R}_+$ and $f([1]) = 1$. The space of states $S_\bullet$ is dual to the linear space K_0^+. From this point of view, "hyperplanes" correspond to linearly independent "vectors" of the space $S_\bullet$. A total order is equivalent to the requirements dim $S_\bullet = 1$ and absence of "infinitesimals", cf. Effros [2] p. 26.

Let φ be an interval exchange transformation built upon $(K_0, K_0^+, [1])$. Invariant measures of φ form a vector space w.r.t. sums and multiplication of measures by positive reals. This vector space is isomorphic to $S_\bullet$. The requirement that φ be strictly (uniquely) ergodic is equivalent to the claim that $S_\bullet$ be one-dimensional. Strict ergodicity of the interval exchange transformations has been a challenging problem in the area for years. (Find a working criterion to determine whether given φ is strictly ergodic.)

This saga started in 1975 when examples of interval exchange transformations with two and three invariant ergodic measures became known due to Keynes and Newton. Keane made an assumption that the "majority" of transformations φ are strictly ergodic. This assumption was turned to a theorem independently by Masur and Veech who used for this purpose the Teichmüller theory and topological dynamics, respectively. The proof of Keane's conjecture based on Theorem 0.1 is given in Section 2.3.

In this section we establish strict ergodicity for a class of interval exchange transformations which we call "stationary". The name comes from theory of ordered abelian groups, because such transformations have stationary Bratteli diagrams; cf. Effros [2]. Foliations that correspond to such transformations are known as *pseudo-Anosov* or foliations whose leaves are 1-dimensional basic sets of the pseudo-Anosov homeomorphisms of a compact surface.

Definition 3.1. Let $\varphi = \varphi(\lambda, \pi, \varepsilon)$ be an interval exchange transformation whose Bratteli diagram is given by the infinite sequence of multiplicity matrices

$$\{P_{Y_1}, P_{Y_2}, P_{Y_3}, \dots\}. \tag{5}$$

If the set (5) can be divided into the blocks $B_k = \{P_{Y_n}, P_{Y_{n+1}}, \dots, P_{Y_{n+k}}\}$ such that $P_{Y_n} P_{Y_{n+1}} \dots P_{Y_{n+k}} = P$, then φ is called stationary. In particular, φ is stationary if $P_{Y_1} = P_{Y_2} = P_{Y_3} = \dots = P$.

Theorem 3.2. *Every stationary interval exchange transformation φ is strictly ergodic.*

Proof. The proof is based on the Perron-Frobenius Theorem. A dual (projective) limit

$$(\mathbb{R}^n)^* \xrightarrow{P_{Y_1}} (\mathbb{R}^n)^* \xrightarrow{P_{Y_2}} (\mathbb{R}^n)^* \xrightarrow{P_{Y_3}} \dots \tag{6}$$

consists of operators P_{Y_i} acting on the dual space $(\mathbb{R}^n)^*$ to $\mathbb{Z}^n \subset \mathbb{R}^n$. (In other words, we identify the space of positive homomorphisms $\mathbb{Z}^n \to \mathbb{R}$ and the space of linear functionals $\mathbb{R}^n \to \mathbb{R}$.) The diagram (6) converges to the state space $S_\bullet$ of dimension group $(K_0, K_0^+, [1])$. When $P_{Y_i} = P$, where P is a matrix with strictly positive entries, or can be reduced to this case, then there exists a maximal simple eigenvalue $\lambda > 0$ of matrix P (Perron-Frobenius Theorem). The eigenvector x_λ defines a 1-dimensional P-invariant subspace of $(\mathbb{R}^n)^*$ lying in the limit of diagram (6) and which is identified with $S_\bullet$. Let us formalize this idea.

For $1 \le i \le n$ denote by e_i and e_i^* the vectors of canonical bases in the vector space $\mathbb{R}^n$ and the dual space $(\mathbb{R}^n)^*$. By $(\alpha_1, \dots, \alpha_n)$ and $(\alpha_1^*, \dots, \alpha_n^*)$ we denote vectors in $\mathbb{R}^n$ and $(\mathbb{R}^n)^*$. $(\mathbb{R}^n)^+$ and $(\mathbb{R}^n)^{*+}$ are collections of vectors whose coordinates are $\alpha_i \ge 0$ and $\alpha_i^* \ge 0$, respectively. The same notation $(\mathbb{Z}^n)^+$ is reserved for the integer vectors of $\mathbb{R}^n$. By $\Delta_0 \subset (\mathbb{R}^n)^*$ we understand the n-dimensional simplex spanned by the vectors $0, e_1^*, \dots, e_n^*$. To each linear mapping $\phi : \mathbb{R}^n \to \mathbb{R}^n$ one associates a dual mapping $\phi^* : (\mathbb{R}^n)^* \to (\mathbb{R}^n)^*$.

Denote by P an dimension group corresponding to the limit

$$P \cong \lim_{k \to \infty} P_k, \tag{7}$$

where the P_k are ordered groups whose positive cone is defined to be an inverse of k-th iteration of the set $(\mathbb{Z}^n)^+$ under the automorphism ϕ:

$$P_k^+ = \phi^{-k}[(\mathbb{Z}^n)^+]. \tag{8}$$

The set Δ_0 has been introduced earlier. For $k = 1, \dots, \infty$ we let

$$\Delta_k = S_\bullet(P_k), \tag{9}$$

which is a state space of the group P_k. Define Δ_k to be a simplex spanned by the vectors $0, J_1(k), \dots, J_n(k)$, where

$$J_1(k) = \frac{X_1(k)}{||X_1(k)||}, \quad \dots \quad , J_n(k) = \frac{X_n(k)}{||X_n(k)||}, \tag{10}$$

and

$$X_1(k) = \phi^k(e_1^*), \quad \dots \quad , X_n(k) = \phi^k(e_n^*). \tag{11}$$

It is evident that

$$\Delta_0 \supseteq \Delta_1 \supseteq \Delta_2 \supseteq \dots \supseteq \Delta_\infty, \tag{12}$$

where

$$\Delta_\infty = \bigcap_{k=1}^{\infty} \Delta_k. \tag{13}$$

A recurrent formula linking $X_i(k-1)$ and $X_i(k)$ is given by the equation

$$X_i(k) = \sum_{j=1}^{n} p_{ij} X_j(k-1), \tag{14}$$

where the p_{ij} are the entries of matrix of "partial multiplicities" P_y.

Note that the simplex Δ_∞ has dimension $r \leq n$. The original problem of calculating the state space $S_\bullet$ is reduced to calculation of the asymptotic simplex Δ_∞ whose spanning vectors are linked by equation (14). We shall see that Δ_∞ can be completely calculated under the hypothesis of Theorem 3.2. The following lemma is basic.

Lemma 3.3. (Perron-Frobenius) *A strictly positive $n \times n$ matrix $P = (p_{ij})$ always has a real and positive eigenvalue λ which is a simple root of the characteristic equation and exceeds the moduli of all the other characteristic values. To this maximal eigenvalue λ there corresponds an eigenvector $x_\lambda = (x_\lambda^1, \ldots, x_\lambda^n)$ with positive coordinates $x_\lambda^i > 0,\ i = 1, \ldots, n.$* [3]

Proof. Let $x = (x_1, x_2, \ldots, x_n)$ be a fixed vector. A function

$$r_x = \min_{1 \leq i \leq n} \frac{(Px)_i}{x_i} \tag{15}$$

is introduced. We have $r_x \geq 0$ since

$$(Px)_i = \sum_{j=1}^{n} p_{ij} x_j, \tag{16}$$

[3] In fact, there exists a more general statement due to Frobenius which treats matrices with non-negative entries. Because of exceptional importance of this statement in understanding why the unique ergodicity may vanish, and also due to the clear connection of Frobenius theorem with the root systems of Coxeter-Dynkin, we give the formulation of this theorem below.

Theorem (Frobenius) *An irreducible non-negative $n \times n$ matrix $P = (p_{ij})$ always has a positive eigenvalue λ that is a simple root of the characteristic equation. The moduli of all the other eigenvalues do not exceed λ. To the maximal eigenvalue λ there corresponds an eigenvector with positive coordinates.*

Moreover, if P has r eigenvalues $\lambda_0 = \lambda,\ \lambda_1, \ldots, \lambda_{r-1}$ of modulus λ, then these numbers are all distinct and are roots of the equation

$$z^r - \lambda^r = 0.$$

More generally: The whole spectrum $\lambda_0, \lambda_1, \ldots, \lambda_{n-1}$ of P, regarded as a system of points in the complex plane, is mapped into itself under a rotation of the plane by the angle $2\pi/r$. If $r > 1$, then P can be put into the cyclic normal form

$$\begin{pmatrix} O & P_{12} & O & \ldots & O \\ O & O & P_{23} & \ldots & O \\ \vdots & & & & \vdots \\ O & O & O & \ldots & P_{r-1,r} \\ P_{r1} & O & O & \ldots & O \end{pmatrix}$$

where P_{ij} are non-zero square blocks along the main diagonal and O are zero square blocks elsewhere.

is a non-negative matrix. However, in the definition of minimum (15), the values of i for which $x_i = 0$ are excluded. The lemma follows from the variational principle for the continuous function r_x, which must assume a maximal value for some vector x with non-negative coordinates. □

Now we can finish the proof of the main theorem. By strict positivity of matrix P and Lemma 3.3, there is a positive maximal eigenvalue λ, whose eigenvector x_λ has positive coordinates. Notice that

$$\phi^k(x_\lambda) = (\lambda)^k x_\lambda, \tag{17}$$

so that the iterations of ϕ leave invariant a 1-dimensional linear subspace $\{\alpha\}$ spanned by x_λ. All other vectors in $(\mathbb{R}^n)^{*+}$ converge accordingly (14) to the subspace $\{\alpha\}$. We conclude that

$$\Delta_\infty = \{\alpha\}, \tag{18}$$

which is one-dimensional. Theorem 3.2 is proved. □

Remark 3.4. In the context of ordered abelian groups, Theorem 3.2 was known to Effros [2] and Elliott [3]. The main ingredients of proof can be traced in the work [14] of W. Veech.

4. Masur-Veech Theorem

The theorem of Masur and Veech is formulated in Section 5.2. There are two known proofs of this theorem, due to Masur [6] who used complex analysis and Teichmüller theory and Veech [15] who used methods of topological dynamics. In this section we suggest an independent proof using Theorem 0.1 and a lemma of Morse and Hedlund from symbolic dynamics; cf. Morse and Hedlund [8].

Parametrization of $(K_0, K_0^+, [1])$**.**[4] Let $\mathbb{H}, \tau$ and S be as in Definition 1.1 of Section 2. Without loss of generality we assume that S is a unit semi-circle in the complex plane $\mathbb{H}$. Consider a family S_t of the unit semi-circles parametrized by real numbers equal to a "horizontal shift" of S in $\mathbb{H}$. (In other words, t is equal to the x-coordinate of the centre of unit circle S_t.) A family of dimension groups which are defined by "positive cones" S_t, we shall denote by $(P, P_+^t, [u])$. By results of Sections 1–3 every dimension group of form $\mathbb{Z}^n$ has a representative in $(P, P_+^t, [u])$ and every measured foliation (with fixed singularity data) arises in this way.

Theorem 4.1. *Denote by $\mathbb{F}_t$ a family of measured foliations corresponding to $(P, P_+^t, [u])$ and by $t_1 \sim t_2$ an equivalence relation on $\mathbb{R}$ identifying topologically equivalent foliations $\mathbb{F}_{t_1}$ and $\mathbb{F}_{t_2}$. If $X = \mathbb{R}/\sim$ is a topological space, then for a residual set of the second category in X foliation $\mathbb{F}_t$ is strictly ergodic.*

[4] The idea of such a parametrization was communicated to the author by G. A. Elliott.

Proof. The idea is to apply Koebe-Morse coding to each geodesic $S \in S_t$. In this setting, X becomes a space of symbolic sequences with the topology described in Section 2. It is not hard to see that strict ergodicity of an individual geodesic S is equivalent to "uniform approximation" of S by periodic sequences of length N. (Using the terminology of Morse and Hedlund, such approximation property of a geodesic means that a transitivity index $\phi(N)$ tends to a covering index $\theta(N)$ of the geodesic as $N \to \infty$; cf. Morse and Hedlund [8].) The same authors proved that $\lim_{\mathbb{N}\to\infty} \inf \frac{\phi(N)}{\theta(N)} = 1$ for a residual set of the second category in X. Let us give the details of this construction.

Let S be a geodesic in the complex plane $\mathbb{H}$ and

$$\sigma_1, \sigma_2, \sigma_3, \ldots,$$

the Koebe-Morse code of S which we shall call a *ray*; cf. Section 3. The ray R is *transitive* if it contains a copy of each admissible block. (Block is shorthand for a finite sequence of symbols.) A function $\phi : \mathbb{N} \to \mathbb{N}$ of a transitive ray R is called a *transitivity index* if the initial block of R of the length $\phi(N)$ contains all admissible blocks of the length N and there are no shorter initial subblocks with this property. Dropping the claim that block $B \subset R$ is initial gives us function $\theta : \mathbb{N} \to \mathbb{N}$ which is called a *covering index* of the recurrent ray R. These functions satisfy an obvious inequality:

$$\phi(N) \geq \theta(N).$$

Lemma 4.2 (Morse-Hedlund). *A set of rays whose transitivity index and covering index satisfy the condition*

$$\lim_{N\to\infty} \inf \frac{\phi(N)}{\theta(N)} = 1,$$

is a residual set of the second category in the space X of all infinite rays endowed with the topology described in Section 2.

Proof. For a complete proof see [8]. Denote by Y the set of rays satisfying the condition of lemma. The following items will be proved consequently:

(i) Y is not empty;
(ii) Y is everywhere dense in X;
(iii) The complement of Y is nowhere dense in X.

(i) Let $H(n)$ be a block of minimum length containing all admissible blocks of length n. For a growing sequence of integers $r_0, r_1, \ldots, r_{k-1}$ consider a block

$$H(r_0)\sigma_1 H(r_1)\sigma_2 \ldots \sigma_{k-1} H(r_{k-1})\sigma_k \tag{19}$$

of length m_k. Let us choose r_k sufficiently large so that

$$\frac{\theta(r_k) + m_k}{\theta(r_k)} < 1 + \delta_k,$$

where δ_k is a vanishing positive real. The transitivity index of (19) satisfies the inequality

$$\phi(r_k) \leq \theta(r_k) + m_k.$$

By the construction, $\frac{m_k}{\theta(r_k)} \to 0$ as $k \to \infty$ so that (19) satisfies the condition of the lemma.

(ii) Let A be an arbitrary admissible block of length k and $R \in Y$. For a suitably chosen σ the ray

$$R' = A\sigma R$$

is admissible. We have the following inequalities for R and R':

$$\theta(N) \leq \phi'(N) \leq \phi(N) + k + 1,$$

where ϕ and ϕ' are the transitivity indices of R and R'. The condition of the lemma is satisfied and therefore $R' \in Y$.

(iii) This item follows from (ii) and an accurate construction of closed sets lying in the complement of Y; cf [8] for the details. This argument finishes the proof of the Morse-Hedlund lemma. □

Let R be a transitive ray and $B_1, \ldots, B_k$ admissible blocks of length N. We say that R is *uniformly distributed* relatively $B_1, \ldots, B_k$ if

$$\phi(N) = kN.$$

(In other words, each admissible block appears in the initial block of R with the "probability" $1/k$.) In the geometric terms this means that geodesic R is located at the same distance from periodic geodesics

$$B_1, B_1, \ldots; \quad B_2, B_2, \ldots; \quad \ldots; \quad B_k, B_k, \ldots.$$

Lemma 4.3. *Suppose that R is uniformly distributed relatively admissible blocks $B_1, \ldots, B_k$ for each integer $N > 0$. Then*

$$\lim_{N \to \infty} \inf \frac{\phi(N)}{\theta(N)} = 1.$$

Proof. This follows from the equality $\phi(N) = \theta(N)$. □

To finish the proof of Theorem 4.1 it remains to notice that strict ergodicity of R is equivalent to uniform distribution of periodic "blocks" in R and apply Lemmas 4.2 and 4.3. □

5. Appendix

5.1. Interval exchange transformations

Let $n \geq 2$ be a positive integer and let $\lambda = (\lambda_1, \ldots, \lambda_n)$ be a vector with positive components λ_i such that $\lambda_1 + \cdots + \lambda_n = 1$. One sets

$$\beta_0 = 0, \qquad \beta_i = \sum_{j=1}^{i} \lambda_j, \qquad v_i = [\beta_{i-1}, \beta_i) \subset [0, 1].$$

Let π be a permutation on the index set $N = \{1, \ldots, n\}$ and $\varepsilon = (\varepsilon_1, ..., \varepsilon_n)$ a vector with coordinates $\varepsilon_i = \pm 1, i \in N$. An *interval exchange transformation* is a mapping $\varphi(\lambda, \pi, \varepsilon) : [0, 1] \to [0, 1]$ which acts by piecewise isometries

$$\varphi(x) = \varepsilon_i x - \beta_{i-1} + \beta^{\pi}_{\pi(i)-1}, \qquad x \in v_i,$$

where β^{π} is a vector corresponding to $\lambda^{\pi} = (\lambda_{\pi^{-1}(1)}, \lambda_{\pi^{-1}(2)}, \ldots, \lambda_{\pi^{-1}(n)})$. The mapping φ preserves or reverses orientation of v_i depending on the sign of ε_i. If $\varepsilon_i = 1$ for all $i \in N$, then the interval exchange transformation is called *oriented*. Otherwise, the interval exchange transformation is said to have *flips*.

An interval exchange transformation is said to be *irreducible* if π is an irreducible permutation. An irreducible interval exchange transformation T is called *irrational* if the only rational relation between numbers $\lambda_1, \ldots, \lambda_n$ is given by the equality $\lambda_1 + \ldots + \lambda_n = 1$. Recall that measure μ on $[0, 1]$ is called invariant if $\mu(\varphi(A)) = \mu(A)$ for any measurable subset $A \subseteq [0, 1]$. The following theorem due to M. Keane [5] estimates the number of invariant measures.

Finiteness Theorem. *Let φ be an irrational interval exchange transformation of n intervals. Then there are at most finitely many ergodic invariant measures whose number cannot exceed $n + 2$.*

In case of the interval exchange transformations without flips, there exists an estimate of the number of invariant ergodic measures due to Veech [15].

Veech Theorem. *Let φ be an irrational interval exchange transformation without flips on $n \geq 2$ intevals. Then the number of invariant ergodic measures of φ is less or equal to $[\frac{n}{2}]$, where $[\bullet]$ is integer part of the number.*

5.2. Measured foliations

Measured foliations are suspensions over the interval exchange transformations which preserve the ergodic measure on intervals and such that their singularity set consists of p-prong saddles, $p \geq 3$. Measured foliations can be defined via closed 1-forms which is more elegant way due to Hubbard, Masur and Thurston.

Definition 5.1 (Hubbard-Masur-Thurston). Let M be a compact C^{∞} surface of genus $g > 1$, without boundary. A measured foliation $\mathbb{F}$ on M with singularities of order $k_1, \ldots, k_n$ at points $x_1, \ldots, x_n$ is given by an open cover U_i of $M \backslash \{x_1, \ldots, x_n\}$ and non-vanishing C^{∞} real-valued closed 1-form ϕ_i on each U_i, such that

(i) $\phi_i = \pm\phi_j$ on $U_i \cap U_j$;

(ii) at each x_i there is a local chart $(u, v) : V \to \mathbb{R}^2$ such that for $z = u + iv$, $\phi_i = \text{Im}\ (z^{k_i/2} dz)$ on $V \cap U_i$, for some branch of $z^{k_i/2}$ in $U_i \cap V$.

Pairs (U_i, ϕ_i) are called an atlas for $\mathbb{F}$.

As it follows from the definition, apart from the singular points, measured foliations look like a non singular volume preserving flows. In singularities, the substitution $z \mapsto re^{i\psi}$ brings ϕ_i, mentioned in (ii), to the form

$$\phi_i = r^{\frac{k_i}{2}} [\sin(\frac{k_i}{2} + 1)\psi dr + r \cos(\frac{k_i}{2} + 1)\psi d\psi].$$

It can be readily established, that ϕ_i are closed differential 1-forms, that is $d\phi_i = 0$ for all $k_i \geq 1$. To see which singularities are generated by the above formula, let us consider a vector field v_i, given by the system of differential equations

$$\frac{dr}{dt} = -r\cos(\frac{k_i}{2}+1)\psi, \qquad \frac{d\psi}{dt} = \sin(\frac{k_i}{2}+1)\psi.$$

Clearly, v_i is tangent to a foliation given by the equation $\phi_i = 0$. Our prior interest is to study the behavior of trajectories of v_i in a narrow stripe $\Pi = \{(r,\psi)| -\varepsilon \leq r \leq \varepsilon, 0 \leq \psi \leq 2\pi\}$. There are exactly $k_i + 2$ equilibria $p_n \in \Pi$, which have the coordinates $(0, \frac{2\pi n}{k_i+2})$, where $n \in \mathbb{N}$ varies from 0 to $k_i + 2$. The linearization of the vector field v_i in these points yields

$$A(p_n) = \begin{pmatrix} (-1)^{n+1} & 0 \\ \\ 0 & (-1)^n(\frac{k_i}{2}+1) \end{pmatrix}.$$

Therefore all p_n are saddle points. One maps the half-stripe $r \geq 0$ to the neighbourhood of the singular point x_i. Generally, a singular point x_i of the order k_i is a $(k_i + 2)$-prong saddle of a measured foliation $\mathbb{F}$.

Let M be a compact surface and $\mathbb{F}$ a measured foliation on M. By *measure* μ of $\mathbb{F}$ one understands a line element $||\phi||$ related with the point $x \in M$, induced in each $x \in U_i$ by $||\phi_i(x)||$. It measures a 'transversal length' of $\mathbb{F}$, since μ vanishes in direction tangent to the leaves of $\mathbb{F}$.

Take a cross-section of the measured foliation $\mathbb{F}$. $\mathbb{F}$ induces an interval exchange transformation φ on this cross-section. Depending on the orientability of $\mathbb{F}$, φ may have flips. Flips are excluded if $\mathbb{F}$ is an orientable measured foliation (in this case $\mathbb{F}$ is given by orbits of a measure-preserving flow). For orientable measured foliations, an estimate of number of invariant ergodic measures is due to Sataev [12].

Sataev Theorem. *Let n and k be a pair of natural numbers, such that $n \geq k$ and let M be a compact orientable surface of genus n. Then there exists a C^∞ orientable measured foliation $\mathbb{F}$ on M whose singularity set consists of 4-separatrix saddles and which has exactly k invariant ergodic measures.*

An important question arises when the measured foliation has a unique invariant measure. It was conjectured by M. Keane and proved by H. Masur and W. Veech that 'almost all' measured foliations have a unique invariant measure, which is a multiple of Lebesgue measure.

Masur-Veech Theorem. ([6], [15]) *Suppose that a family $\mathbb{F}_t$ of measured foliations is given by trajectories of a holomorphic quadratic differential $e^{it}\phi$ on the surface M. Then for 'almost all' values of t, the foliation $\mathbb{F}_t$ is strictly ergodic.*

5.3. $\mathcal{O}_\lambda$ as a crossed product C^*-algebra

Lemma 5.2. *Let $\varphi = \varphi(\lambda, \pi, \varepsilon)$ be an interval exchange transformation and $\lambda = (\lambda_1, \ldots, \lambda_n)$. Then $K_0(\mathcal{O}_\lambda) = \mathbb{Z}^n$ and $K_1(\mathcal{O}_\lambda) = \mathbb{Z}$.*

Proof. Let $p_1, \ldots, p_n$ be the set of discontinuous points of the mapping φ. Denote by $Orb\ \varphi = \{\varphi^m(p_i) :\ 1 \le i \le n,\ n \in \mathbb{Z}\}$ a set of full orbits of these points. When φ is irrational, the set $Orb\ \varphi$ is a dense subset in $[0,1]$. We replace every point $x \in Orb\ \varphi$ in the interior of $[0,1]$ by two points $x^- < x^+$ moving apart banks of the cut. The obtained set is a Cantor set denoted by X.

A mapping $\varphi : X \to X$ is defined to coincide with the initial interval exchange transformation on $[0,1]\backslash Orb\ \varphi \subset X$ prolonged to a homeomorphism of X. The mapping φ is a minimal homeomorphism of X, since there are no proper, closed, φ-invariant subsets of X except the empty set. Thus, $\mathcal{O}_\lambda = C(X) \rtimes_\varphi \mathbb{Z}$ is a crossed product C^*-algebra, where $C(X)$ denotes a C^*-algebra of continuous complex-valued functions on X. The following diagram of Pimsner and Voiculescu consists of exact sequences:

$$\begin{array}{ccccc} K_0(C(X)) & \xrightarrow{id-\varphi_*} & K_0(C(X)) & \xrightarrow{i_*} & K_0(C(X)\rtimes_\varphi \mathbb{Z}) \\ \uparrow & & & & \downarrow \\ K_1(C(X)\rtimes_\varphi \mathbb{Z}) & \xleftarrow{i_*} & K_1(C(X)) & \xleftarrow{id-\varphi_*} & K_1(C(X)) \end{array}$$

It was proved in [10] that $K_0(C(X)) \simeq \mathbb{Z}^n$ and $K_1(C(X)) \simeq 0$. To obtain the conclusion of Lemma 5.2 it remains to calculate all short exact sequences in the diagram of Pimsner and Voiculescu. □

5.4. Rotation numbers

One of the striking invariants of the algebra $\mathcal{O}_\lambda$ are rotation numbers associated to this algebra. In the dynamical context, rotation numbers are equal to "average inclination" of leaves of measured foliation relatively a coordinate system on M_n. (In fact, the original study of $\mathcal{O}_\lambda$ was motivated by the possibility to introduce such numbers; cf. [9].) Rotation numbers for $\mathcal{O}_\lambda$ play the same role as real numbers θ for the irrational rotation algebra A_θ.

Recall that the cone $K_0^+ \subset \mathbb{H}$ is a limit of "rational" cones $P_k^+ \subset \mathbb{H}$:

$$K_0^+ = \lim_{k\to\infty} P_k^+.$$

Each P_k^+ is represented by a periodic geodesic γ_k. Suppose that $g_k \in G$ is an isometry which moves the geodesic γ_{k-1} to the geodesic γ_k and let

$$g_k = \begin{pmatrix} a_i & b_i \\ c_i & d_i \end{pmatrix} \in PSL_2(\mathbb{Z})$$

be an integral matrix with non-negative entries and determinant ± 1, corresponding to g_k. The continued fraction

$$\theta_\lambda = \frac{a_1}{c_1} - \cfrac{c_1^{-2}}{\frac{d_1}{c_1} + \frac{a_2}{c_2} - \cfrac{c_2^{-2}}{\frac{d_2}{c_2} + \frac{a_3}{c_3} - \dots}}$$

converges to a real number θ_λ which is called a *rotation number* associated to the algebra $\mathcal{O}_\lambda$. The importance of rotation numbers is stipulated by the following theorem.

Theorem. *Let $\mathcal{O}_\lambda$ and $\mathcal{O}'_\lambda$ be two C^*-algebras whose rotation numbers are θ_λ and θ'_λ. Then $\mathcal{O}_\lambda$ is Morita equivalent to $\mathcal{O}'_\lambda$ if and only if θ_λ and θ'_λ are modular equivalent:*

$$\theta'_\lambda = \frac{a\theta_\lambda + b}{c\theta_\lambda + d}, \qquad a, b, c, d \in \mathbb{Z}, \qquad ad - bc = \pm 1.$$

Proof. This was proved in [9]. □

Corollary 5.3. *Suppose that φ is a stationary interval exchange transformation described in Section 3. Then the rotation number θ_λ is a quadratic surd (i.e., irrational root of a quadratic equation).*

Proof. By the results of Section 3, the dimension group of $\mathcal{O}_\lambda$ is stationary and must correspond to a periodic continued fraction (i.e., $g_1 = g_2 = \dots = \text{const}$). These fractions generate a field of quadratic algebraic numbers. □

Conclusions and open problems

The criterion of strict ergodicity of Section 3 is highly constructive and can be used in practice to check whether a given interval exchange transformation is strictly ergodic or not. (This can find applications in the theory of billiards in the rational polygons.) Of course, these conditions are only sufficient. The necessary conditions seem to be an open problem so far.

Another open problem is to relate the arithmetic of rotation numbers θ_λ with the number of invariant measures of the transformation φ. (In the case of strictly ergodic φ the answer is given by Corollary 5.3.)

References

[1] O. Bratteli, Inductive limits of finite dimensional C^*-algebras, *Trans. Amer. Math. Soc.* **171** (1972), 195–234.

[2] E.G. Effros, *Dimensions and C^*-Algebras*, in: Conf. Board of the Math. Sciences, Regional conference series in Math., No. 46, AMS, 1981.

[3] G.A. Elliott, On totally ordered groups and K_0, Ring theory (Proc. Conf., Univ. Waterloo, Waterloo, 1978), pp. 1–49, *Lecture Notes in Math.*, Vol. 734, Springer, Berlin, 1979.

[4] K.R. Goodearl, *Partially Ordered Abelian Groups With Interpolation*, Mathematical Surveys and Monographs, 20. American Mathematical Society, Providence, R.I., 1986. xxii+336 pp., ISBN: 0-8218-1520-2

[5] M. Keane, Interval exchange transformations, *Math. Z.* **141** (1975), 25–31.

[6] H. Masur, Interval exchange transformations and measured foliations, *Ann. of Math.* (2) **115** no. 1 (1982), 169–200.

[7] M. Morse, Recurrent geodesics on a surface of negative curvature, *Trans. Amer. Math. Soc.* **22** (1921), 84–100.

[8] M. Morse and G.A. Hedlund, Symbolic dynamics, *Amer. J. of Math.* **60** (1938), 815–866.

[9] I. Nikolaev, Invariant of minimal flows coming from the K_0-group of a crossed product C^*-algebra, *Ergodic Theory Dynam. Systems* **20** (2000), 1449–1468.

[10] Ian F. Putnam, The C^*-algebras associated with minimal homeomorphisms of the Cantor set, *Pacific J. Math.* **136** no. 2 (1989), 329–353.

[11] Ian F. Putnam, C^*-algebras arising from interval exchange transformations, *J. Operator Theory* **27** no. 2 (1992), 231–250.

[12] E.A. Sataev, On the number of invariant measures for flows on orientable surfaces, *Izv. Akad. Nauk SSSR Ser. Mat.* **39** no. 4 (1975), 860–878.

[13] W.P. Thurston, *The Geometry and Topology of Three-Manifolds*, MSRI 1997, electronic edition of 1980 Princeton Univ. notes, available at http://www.msri.org/gt3m/; alternative reference: Three-Dimensional Geometry and Topology, ed. by Silvio Levy, vol. 1, Princeton Univ. Press, 1997.

[14] W.A. Veech, Interval exchange transformations, *J. Analyse Math.* **33** (1978), 222–272.

[15] W.A. Veech, Gauss measures for transformations on the space of interval exchange maps, *Ann. of Math.* (2) **115** no.1 (1982), 201–242.

Acknowledgment

I wish to thank G. A. Elliott for many helpful discussions and ideas.

Igor V. Nikolaev
The Fields Institute
222 College street
Toronto M5T 3J1
Canada
e-mail: nikolaev@math.ucalgary.ca

Operator Theory:
Advances and Applications, Vol. 179, 229–240

Superfast Inversion of Two-Level Toeplitz Matrices Using Newton Iteration and Tensor-Displacement Structure

Vadim Olshevsky, Ivan Oseledets and Eugene Tyrtyshnikov

Abstract. A fast approximate inversion algorithm is proposed for two-level Toeplitz matrices (block Toeplitz matrices with Toeplitz blocks). It applies to matrices that can be sufficiently accurately approximated by matrices of low Kronecker rank and involves a new class of tensor-displacement-rank structured (TDS) matrices. The complexity depends on the prescribed accuracy and typically is $o(n)$ for matrices of order n.

Mathematics Subject Classification (2000). 15A12; 65F10; 65F15.

Keywords. Kronecker product, low-rank matrices, multilevel matrices, Toeplitz matrices, displacement rank.

1. Introduction

Dense matrices arise, for example, in numerical solution of multidimensional integral equations; their approximate inverses are often of interest either themselves or as preconditioners in iterative methods, and the size of matrices occurs to be about a few hundred of thousands or even millions. These cases are not very easy to handle. The standard Gaussian elimination has the $O(n^3)$ complexity and is unacceptable. Even a method with $O(n^2)$ complexity (an obvious lower bound) is still too slow for matrices on this scale. Luckily, in many cases the matrices possess some structure suggesting a way to make them tractable.

If A is a nonsingular Toeplitz matrix ($a_{ij} = a_{i-j}$), then all the entries of A^{-1} can be computed in $O(n^2)$ operations [13]. It is even more important that A^{-1} can be expressed by the Gohberg–Semencul formula [3] through some $O(n)$ parameters so that it can be multiplied by a vector in $O(n \log n)$ operations. A tremendous

E. Tyrtyshnikov supported by the Russian Fund of Basic Research (grant 05-01-00721) and a Priority Research Grant of the Department of Mathematical Sciences of the Russian Academy of Sciences.

impact of this formula on the field of structured matrices and numerical algorithms was systematically presented in the remarkable book by G. Heinig and K. Rost [5]. A direct but nontrivial generalization to block Toeplitz matrices is the Gohberg–Heinig formula [2].

In this paper we consider two-level Toeplitz matrices, which are block Toeplitz matrices with Toeplitz blocks. If p is simultaneously the block size and the size of blocks, then $n = p^2$ and such a matrix is defined by $O(n)$ parameters. In this case the Gohberg–Heinig formula contains as many as $O(p^3) = O(n^{3/2})$ parameters, which is viewed as too many, when compared with $O(n)$. A better approach can be one that we outlined and started to develop in [9]. However, it applies only to those two-level Toeplitz matrices that are of low tensor (Kronecker) rank. As a nice consequence of this combination of Toeplitz and tensor structure, such matrices are determined by $O(\sqrt{n})$ parameters, the same is expected from their approximate inverse matrices and may (and does, as we show) result in the $o(n)$ complexity. Luckily again, this special subclass of two-level Toeplitz matrices seems to cover all practically interesting matrices.

We will make use of the following iterative method attributed to Hotelling [6] and Schulz [12]:

$$X_i = 2X_{i-1} - X_{i-1}AX_{i-1}, \quad i = 0, 1, \ldots, \tag{1}$$

where X_0 is some initial approximation to A^{-1}. Since $I - AX_i = (I - AX_{i-1})^2$, the iterations (1) converge quadratically, provided that $||I - AX_0|| < 1$. This method is a special form of the Newton method for nonlinear equations and referred to as *Newton iteration*. It has some nice properties such as numerical stability and ease for parallel computations. All the same, each iteration requires two matrix multiplications, which is expensive for general matrices.

In order to perform the Newton iteration in a fast way, we need the following two ingredients:

- a fast matrix-by-matrix procedure;
- a method to preserve structure.

The first means that X_k and A must hold on some structure to facilitate the computation of matrix products. However, if the X_k do not belong to a commutative algebra (circulants, diagonal matrices etc), every next iterate X_{k+1} might be "less structured". As a consequence, the matrix-by-matrix complexity grows with every iteration. In order to slow down this growth, we should preserve the structure by "brute force" — using a method to substitute computed iterates with some approximations by "better structured matrices". We introduce a truncation operator $R(X)$ acting on $n \times n$ matrices as kind of a nonlinear projector. Then, the Newton iteration with approximations (truncations) reads

$$X_i = R(2X_{i-1} - X_{i-1}AX_{i-1}). \quad i = 0, 1, \ldots. \tag{2}$$

The Newton iteration was successfully applied to matrices with the displacement structure [1, 11] and matrices represented as a sum of tensor (Kronecker) products [10]. In the case of low-displacement-rank matrices, V. Pan [11] proved that the

quadratic convergence is maintained even after truncations. Then, it was discovered in [10] that the latter property holds true for many useful structures rather than one considered in [11]. A pretty general formulation stemming from [10] is given in [4].

Theorem 1.1. *Suppose that* $||(R(X) - A^{-1}|| \leq M||X - A^{-1}||$ *for all* X*. Then for any initial guess* X_0 *sufficiently close to* A^{-1}*, the truncated Newton iterates* (2) *converge quadratically:*

$$||A^{-1} - X_k|| \leq (1 + M)\,||A||||A^{-1} - X_{k-1}||^2, \qquad k = 1, 2, \ldots\ .$$

Now, with this encouraging result, we are going to propose an algorithm for computing an approximate inverse to a given two-level Toeplitz matrix. Our main idea is to combine two efficient matrix representations using the low-Kronecker-rank and low-displacement-rank properties. Thus, we introduce a new matrix format — the TDS format (tensor displacement structure), and therefore assume that A and A^{-1} should be in the TDS format, at least approximately. A rigorous theory behind this assumption is still lacking; however, all of our numerical experiments on various matrices show that the complexity of the proposed algorithm is $O(\sqrt{n}\log n)$.

The paper is organized as follows.

In Section 1 we define the TDS format and the transformation of a two-level Toeplitz matrix into this format. In Section 2 we describe all the basic matrix operations in the TDS format and propose a *fast recompression procedure* (in other words, define the operator R).

In Section 3 we discuss the Newton iteration with approximations and its modification which speeds up the computations dramatically. Also, we suggest a method for efficient selection of the initial guess X_0. In Section 4 we present some numerical experiments.

2. The TDS format

Below we recall a general notation of multilevel matrices introduced in [14] and the displacement rank constructions presented in [5] as a far-reaching development of the definition introduced first in [7].

Definition 2.1. A matrix T is considered as two-level with the size-vector (n_1, n_2) if it contains $n_1 \times n_1$ blocks and each block is of size $n_2 \times n_2$. Such a matrix is called *two-level Toeplitz matrix* if

$$T = [a(\mathbf{i} - \mathbf{j})], \tag{3}$$

where $\mathbf{i} = (i_1, i_2)$ and $\mathbf{j} = (j_1, j_2)$ define the place of the element in the two-level matrix: (i_1, j_1) specifies the block position and (i_2, j_2) does the element location inside the block.

Definition 2.2. The operator L is said to be *of Sylvester type* if

$$L(M) = \nabla_{A,B}(M) = AM - MB \tag{4}$$

and *of Stein type* if

$$L(M) = \triangle_{A,B}(M) = M - AMB. \tag{5}$$

The value $\alpha \equiv \mathrm{rank}(L(M))$ is called the *displacement rank* of M. Any $n \times \alpha$ matrices G and H from the skeleton decomposition

$$L(M) = GH^{\top}$$

are called the *generators* of M. A matrix defined by its generators is referred to as a *displacement-structured matrix*. By the very definition, displacement ranks and generators of a matrix depend on the choice of the displacement operator L.

We will use the Stein type operators. The Toeplitz matrices can be associated with the displacement operators $Z_a, Z_b^{\top}$, where

$$Z_a = Z + ae_0e_{n-1}^{\top}, \qquad Z_b = Z + be_0e_{n-1}^{\top},$$

Z is a unit lower shift matrix and a, b are some scalars. Let $\triangle_{Z_a,Z_b^{\top}}(M) = GH^{\top}$ and $G = [g_1, \ldots, g_\alpha]$, $H = [h_1, \ldots, h_\alpha]$. Then

$$(1 - ab)M = \sum_{j=1}^{\alpha} Z_a(g_j)Z_b^{\top}(h_j). \tag{6}$$

Here, $Z_a(g)$ and $Z_b(h)$ are defined as follows. Let c be a scalar and v a vector; then $Z_c(v)$ is a Toeplitz matrix with the entries

$$(Z_c(v))_{ij} = \begin{cases} v_{i-j}, & i-j \geq 0, \\ c\, v_{n+i-j}, & i-j < 0. \end{cases}$$

If M is nonsingular, then M^{-1} can be expressed by a formula of the same type as (6), considered in this case as one of possible generalizations of the Gohberg–Semencul formula to Toeplitz-like matrices. Both in the latter formula and in (6), a matrix is the sum of special Toeplitz matrices belonging to some algebras; however, the Gohberg–Semencul formula and (6) use different algebras. If M is a Toeplitz matrix, then $\alpha \leq 2$.

Definition 2.3. A matrix A is said to be in the *tensor format* of the *tensor rank* r, if

$$A = \sum_{k=1}^{r} A_k^1 \otimes A_k^2. \tag{7}$$

Given a two-level matrix A, we can try to approximate it by a low-tensor-rank matrix. Let

$$\mathcal{V}_{\mathbf{n}}(A) = [b_{(i_1,j_1)(i_2,j_2)}]$$

be a two-level matrix with the size-vectors (n_1, n_1) and (n_2, n_2), and define it by the rule

$$b_{(i_1,j_1)(i_2,j_2)} = a_{(i_1,i_2)(j_1,j_2)}.$$

Then, as is readily seen, the tensor rank of A is equal to the rank of $\mathcal{V}_{\mathbf{n}}(A)$. Moreover,

$$||A - A_r||_F = ||\mathcal{V}_{\mathbf{n}}(A) - \mathcal{V}_{\mathbf{n}}(A_r)||_F,$$

which reduces the problem of optimal tensor approximation to the problem of optimal lower-rank approximation. The latter can be solved using the SVD or the Lanzos bidiagonalization algorithm. However, in the case of two-level Toeplitz matrices we can solve this problem much easier [8] (for more general constructions see [9]).

Given $T = [a(\mathbf{i} - \mathbf{j})]$, we compose a smaller matrix

$$W(A) = [a_{\mu\nu}], \quad 1 - n_1 \le \mu \le n_1 - 1,\ 1 - n_2 \le \nu \le n_2 - 1, \tag{8}$$

construct an optimal rank-r approximation

$$W(A) \approx \sum_{k=1}^{r} u_k v_k^{\top},$$

$$U^k = [u^k_{i_1 - j_1}], \quad 0 \le i_1, j_1 \le n_1 - 1,$$

$$V^k = [v^k_{i_2 - j_2}], \quad 0 \le i_2, j_2 \le n_2 - 1,$$

and finish with the tensor approximation of the form

$$T \ \approx \ T_r = \sum_{k=1}^{r} U^k \otimes V^k. \tag{9}$$

It is proved that this is an optimal tensor-rank-r approximation to T in the Frobenius norm. The computational cost is that of finding a low-rank approximation to the matrix of size $(2n_1 - 1) \times (2n_2 - 1)$. Remarkably, the tensor factors are themselves Toeplitz matrices. A crucial parameter defining the complexity is the tensor rank r. It depends on the prescribed approximation accuracy and is directly related to the properties of the symbol (generating function) of T. Some upper estimates on r were proposed in [9] for asymptotically smooth symbols.

It is proved in [9] that a two-level Toeplitz matrix with an approximately separable symbol can be approximated by a sum of tensor products of Toeplitz matrices. Now we embed this format into a more general one which suits better to approximate the corresponding inverse matrices.

Definition 2.4. A two-level matrix A is said to be in the *TDS* (*tensor-displacement structure*) format if it is in the tensor format (7) with each factor being a displacement-structured matrix.

Let r be the tensor rank and s the maximal displacement rank of the factors. Obviously, the TDS format requires a storage of $O(\sqrt{n}rs)$ cells.

3. Matrix arithmetic in the TDS format

3.1. Basic operations in the displacement format

Consider matrices A and B of Toeplitz displacement rank α and β. Then it is is well known that

- a matrix-by-vector product Ax can be computed in $O(\alpha n \log n)$ operations;
- a matrix-by-matrix product AB can be computed in $O(\alpha\beta n \log n)$ operations, with the displacement rank of AB increasing at most to $\alpha + \beta$.

3.2. Basic operations in the tensor format

If two matrices M_1 and M_2 are in the tensor format

$$M^1 = \sum_{i=1}^{r_1} A_i^1 \otimes B_i^1, \quad M^2 = \sum_{i=1}^{r_2} A_i^2 \otimes B_i^2,$$

then the product

$$M^1 M^2 = \sum_{i=1}^{r_1}\sum_{j=1}^{r_2} (A_i^1 A_j^2) \otimes (B_i^1 B_j^2) \tag{10}$$

is already in the tensor format. However, it requires a larger storage as the tensor ranks are to be multiplied. The sum of two matrices in the tensor format is also in some tensor format (we even should not do anything — only merge two arrays). But again, the tensor rank grows. Thus, we should find a way to approximate the results of matrix operations by matrices of lower tensor rank. This task can be accomplished very efficiently through an SVD-based procedure called *recompression*. Since the problem of finding a low-tensor-rank approximation to matrix A is equivalent to the problem of finding a low-rank approximation to $\mathcal{V}_{\mathbf{n}}(A)$, we can expoit the following.

Given a low rank matrix $B = UV^\top$, $U, V \in \mathbb{R}^{n\times r}$, we can find $q \le r$ and matrices $\widetilde{U}, \widetilde{V}^\top \in \mathbb{R}^{n\times q}$ approximating A with the desired accuracy ε:

$$||B - \widetilde{U}\widetilde{V}^\top||_F \le \varepsilon ||B||_F. \tag{11}$$

All we need is to find the SVD of B. Since B is already in the low-rank format, we proceed as follows:

(1) Find the QR-decomposion of U and V: $U = Q_u R_u, V = Q_v R_v$.
(2) Find the SVD of a $r \times r$ matrix $R_u R_v^\top$: $R_u R_v^\top = U_1 \Sigma V_1^\top$.

Then, $B = (Q_u U_1)\Sigma(Q_v V_1)$ is the SVD of B. Now, take the smallest possible q so that

$$\sigma_{q+1}^2 + \cdots + \sigma_r^2 \le \varepsilon ||B||_F.$$

When r is small, the cost of this method is dominated by the QR-decomposition complexity, which is $O(nr^2)$, and is linear in matrix size. But, recall that the columns of U and V come from the reshaped tensor factors which are stored in the displacement format (to be extracted from the generators). Does it help to perform the recompression faster? The answer is yes, the algorithm being described in the next subsection.

3.3. The TDS recompression

Let us look more closely at the recompression steps. The QR-decomposition can be implemented through the Gram–Schmidt orthogonalization algorithm applied to the vectors $u_1, \dots, u_r$. The orthogonality is defined by the ordinary scalar product $(x, y) = \sum_{k=1}^{n} x_i \overline{y}_i$. Now, instead of working with vectors, we suggest to work directly with their matrix prototypes. The scalar product for matrices is defined as the *Frobenius scalar product*:

$$(A, B)_F = \mathrm{tr}(\mathrm{AB}^*).$$

Other operations required in the Gram–Schmidt algorithm, which are multiplication by numbers and addition, can be performed directly with matrices. Moreover, employing the displacement structure in these operations leads to the $O(\sqrt{n}\log n)$ complexity. Thus, we should focus on fast calculation of the Frobenius scalar product of two matrices given in the displacement formats.

Given $p \times p$ matrices A, B with displacement ranks α, β, we need to find $\mathrm{tr}(\mathrm{AB}^*)$. First, we calculate AB^*. As we know, that can be done in $O((\alpha+\beta)p\log p)$ operations and the displacement rank of the product does not exceed $\alpha + \beta$. It remains to calculate the trace of a Toeplitz-like matrix. Fortunately, this can be done by a simple formula involving the generators.

Lemma 3.1. *Let C be a $p \times p$ matrix and $\triangle_{Z_a, Z_b^\top}(C) = GH^T$, $G = [g^1, \dots, g^\alpha]$, $H = [h^1, \dots, h^\alpha]$, where $h^i, g^i \in \mathbb{R}^p$. Then*

$$\mathrm{tr}(C) = \frac{1}{1-ab} \sum_{r=1}^{\alpha} \sum_{k=0}^{p-1} h_k^r g_k^r (p - k + abk). \tag{12}$$

Proof. According to (6), the matrix C can be represented as

$$C = \frac{1}{1-ab} \sum_{j=1}^{\alpha} Z_a(g_j) Z_b^\top(h_j).$$

Therefore,

$$\mathrm{tr}(C) = \frac{1}{1-ab} \sum_{j=1}^{\alpha} \mathrm{tr}(Z_a(g_j) Z_b^\top(h_j)). \tag{13}$$

Each term in the sum (13) is of the form

$$\begin{aligned}\mathrm{tr}(Z_a(g) Z_b^\top(h)) &= \sum_{i=0}^{p-1} (Z_a(g) Z_b^\top(h))_{ii} = \sum_{i=0}^{p-1} \sum_{k=0}^{p-1} Z_a(g)_{ik} Z_b(h)_{ik} \\ &= \sum_{i=0}^{p-1} \sum_{k=0}^{i} g_{i-k} h_{i-k} + ab \sum_{i=0}^{p-1} \sum_{k=i+1}^{p-1} g_{p+i-k} h_{p+i-k}.\end{aligned}$$

The first summand is transformed as

$$\sum_{i=0}^{p-1} \sum_{k=0}^{i} g_{i-k} h_{i-k} = \sum_{i=0}^{p-1} \sum_{k=0}^{i} g_k h_k = \sum_{k=0}^{p-1} h_k g_k (p-k),$$

and, similarly, the second one is

$$\sum_{i=0}^{p-1}\sum_{k=i+1}^{p-1} g_{p+i-k}h_{p+i-k} = \sum_{k=0}^{p-1} h_k g_k k. \qquad \square$$

3.4. Truncation operator

The truncation operator $R(X)$ can be defined by setting either some bounds on the ranks or accuracy. Fixing the ranks, we find $R_{\rho,s}(X)$ through the following steps:

(1) Find the best tensor-rank-ρ approximation X_ρ to X using the fast recompression algorithm.
(2) Approximate tensor factors by some displacement-rank-s matrices.

It can be verified that such an operator satisfies the conjectures of Theorem 1.1. It follows that the Newton method with the truncation operator $R_{\rho,s}(X)$ retains an important property of quadratic convergence.

However, in practice it is expedient to prescribe the accuracy and let the rank vary. Denote the corresponding operator by R_ε. Formally the steps are the same, but the ranks are no longer constant. The first step ends with the best low-tensor approximation to X satisfying $||X - X_r|| \le \varepsilon||X||$, the second step produces an approximation with the preset accuracy and smallest possible displacement rank.

4. Newton iteration for approximate inversion of matrices

Let A be in the TDS format. If an initial approximation X_0 to A^{-1} is in the same format, then it can be fastly improved by the iteration (1). The residuals $R_k = I - AX_k$ satisfy $R_{k+1} = R_k^2$, which proves the quadratic convergence of the process provided that the spectral radius of R_0 is less than 1. The initial approximation can be always selected as

$$X_0 = \alpha A^*$$

with some $\alpha > 0$. In this case the estimated number of the operations to achieve accuracy $||A^{-1} - X_k||_2/||A^{-1}||_2 \le \varepsilon$ is

$$\log_2(c^2+1) + \log_2 \ln \frac{1}{\varepsilon},$$

where c is the spectral condition number of matrix A. For ill-conditioned matrices, the cost is dominated by $\log_2(c^2+1)$.

4.1. Modified Newton iteration

On each step of the Newton method (2), we replace X_k with $R_\varepsilon(X_k)$, where ε is the accuracy parameter. We can also use a modification [10] that works with approximations much better. Indeed, a typical tensor rank of matrices in our examples is about $10 \div 15$, so each Newton step involves about 200 multiplications

of Toeplitz-like matrices with the displacement ranks being typically about 10. Following [10], we consider the following modification of the Newton iteration:

$$X_k = X_{k-1}(2I - X_{k-1}), \quad Y_k = Y_{k-1}(2I - X_{k-1}), \qquad k = 1, 2, \ldots, \quad (14)$$

where Y_0 is an initial approximation to A^{-1} and $X_0 = AY_0$ is a nonsingular matrix of which we require that the spectral radius of $I - X_0$ is less than 1. The latter implies

$$\lim_{k\to\infty} X_k = I,$$

and since it is easy to derive from (14) that

$$Y_{k+1}X_{k+1}^{-1} = Y_kX_k^{-1} = \cdots = Y_0X_0^{-1} = A^{-1},$$

we conclude that

$$\lim_{k\to\infty} Y_k = A^{-1}.$$

In case of general matrices, it is easy to see that (14) is just another way of writing (1). However, in the approximate arithmetic the situation changes dramatically. The modified Newton method with approximations now reads

$$X_k = R_\varepsilon(X_{k-1}(2I - X_{k-1})), \; Y_k = R_\varepsilon(Y_{k-1}(2I - X_{k-1})), \quad k = 1, 2, \ldots. \quad (15)$$

A good argument in favour of this modification is the following. As long as X_k converges to the identity matrix, its tensor rank decreases and, hence, the displacement ranks of the factors become smaller and cause the complexity get down. (This should supposedly hold for any class of structured matrices in which the identity matrix is considered as one with "perfect" structure).

4.2. Selection of the initial approximation

Selection of the initial approximation X_0 to A^{-1} is crucial, in particular for ill-conditioned matrices. The common choice $X_0 = \alpha A^*$ with an appropriate $\alpha > 0$ is ever available, of course, but never good if we want a sufficiently accurate answer. However, we can play with the accuracy parameter ε. In the case of structured matrices it controls both the final accuracy and the truncation accuracy on iterations. Thus, it accounts for the ranks after truncation, and thence the speed of calculations. When the process is "far" from the fast convergence stage, we can carry out the truncation with a much lower accuracy ε. Consequently, the matrix operations become pretty fast in the beginning. On later stages ε must diminish and in the end stay on the level of the desired final accuracy.

This idea was used in [10] for a two-level Toeplitz matrix arising after discretization of a hypersingular integral equation. It can be summarized in the following scheme:

(1) Set $X_0 = \alpha A^*$ and perform the Newton iteration with the truncation accuracy $\delta \gg \varepsilon$. This results in a rough approximation M to the inverse, but the advantage is that the δ-truncated Newton iterations are expected to have a low complexity.

(2) Use the previous approximation M as a new guess to start the Newton iteration with finer accuracy ε.

Of course, this scheme can be extended to three or more steps with relative errors δ_1, δ_2, and so on.

5. Numerical results

Here two model numerical examples are presented. For simplicity we assume $n_1 = n_2 = \sqrt{n}$

First is the standard 5-point Laplacian. It is a two-level Toeplitz matrix $[a_{\mathbf{i}-\mathbf{j}}]$, with free parameters a_{ij} defined as $a_{ij} = 0$, for $-n_1 + 1 \le i \le n_1 - 1$, $j = -n_2 + 1 \le j \le n_2 - 1$, except for

$$a_{00} = 4, \quad a_{0,\pm 1} = -1, \quad a_{\pm 1,0} = -1.$$

Second is a dense two-level Toeplitz matrix with a_{ij} determined by formulas

$$a_{ij} = -f(i+0.5, j-0.5) + f(i-0.5, j-0.5) - f(i-0.5, j+0.5) + f(i+0.5, j+0.5),$$

where

$$f(x, y) = \frac{\sqrt{x^2 + y^2}}{xy}.$$

This matrix comes from the discretization of the hypersingular integral equation [10].

The results are given in Tables 1 and 2. We calculated tensor ranks for the approximate inverse and mean displacement ranks of the factors. All computations were conducted with $\varepsilon = 10^{-5}$ (this means that "tensor rank" and "mean displacement rank" in these tables stand for ε-ranks).

n	64^2	128^2	256^2	512^2
Running time	154 sec	333 sec	966 sec	2555 sec
Tensor rank of A^{-1}	9	10	11	12
Mean displacement rank of A^{-1}	13.5	13.5	16.8	18.6

Table 1. Numerical results for the case 1.

n	64^2	128^2	256^2	512^2
Running time	270 sec	433 sec	817 sec	1710 sec
Tensor rank of A^{-1}	13	13	12	11
Mean displacement rank of A^{-1}	8.5	9.3	9.5	9.7

Table 2. Numerical results for the case 2.

At least for these two examples we can see that the running time obeys the expected $\mathcal{O}(\sqrt{n} r_{\text{mean}}^2)$ asymptotics (where r_{mean} is a mean displacement rank; the dependence from tensor rank is hard to observe in these examples). However, we are not very satisfied with the absolute values: the constant seems to be quite large. After examining the program code it was found that the main computational efforts were spent while recompressing the results of the multiplication of two "large" (of tensor rank 5–10, approximately) TDS matrices. The multiplication using formula (10) was very fast. However, the recompression was much, much longer and it can

be explained why. We have to compress the matrix of tensor rank approximately 50–100. This involves computation of many scalar products. We do not take into account that the matrix is in fact of *much lower* tensor rank (say, 10). This surely can be used in some kind of *rank-revealing* approximation of such a matrix. In the current implementation we have to calculate, in fact, the Frobenius scalar products between all the factor matrices and that is approximately 100^2 scalar products and that leads to serious slowdown. The rank-revealing version of the structured recompression will be reported elsewhere.

References

[1] D.A. Bini and B. Meini, Solving block banded block Toeplitz systems with structured blocks: algorithms and applications, *Structured Matrices: Recent Developments in Theory and Computation. Advances in Computation* (Edited by D.A. Bini, E. Tyrtyshnikov and P. Yalamov), Nova Science Publishers, Inc., Huntington, New York, 2001.

[2] I. Gohberg and G. Heinig, Inversion of finite-section Toeplitz matrices consisting of elements of a non-commutative algebra, *Rev. Roum. Math. Pures et Appl.* **19** no. 5 (1974), 623–663.

[3] I. Gohberg and A.A. Semencul, On inversion of finite-section Toeplitz matrices and their continuous analogues, *Matem. Issled.* **7** no. 2 (1972), 201–224 (in Russian).

[4] W. Hackbusch, B.N. Khoromskij and E.E. Tyrtyshnikov, *Approximate iterations for structured matrices*, Max-Planck-Insitut, Leipzig, Preprint 112, 2005.

[5] G. Heinig and K. Rost, *Algebraic methods for Toeplitz-like matrices and operators*, Berliin, Akademie-Verlag, 1984.

[6] H. Hotelling, Analysis of a complex of statistical variables into principal components, *J. Educ. Psych.* **24** (1933), 417–441, 498–520.

[7] T. Kailath, S. Kung and M. Morf, Displacement ranks of matrcies and linear equations, *J. Math. Anal.and Appl.* **68** (1979), 395–407.

[8] J. Kamm and J.G. Nagy, Optimal Kronecker Product Approximations of Block Toeplitz Matrices, *SIAM J. Matrix Anal. Appl.* **22** no. 1 (2000), 155–172.

[9] V. Olshevsky, I. Oseledets and E. Tyrtyshnikov, Tensor properties of multilevel Toeplitz and related matrices, *Linear Algebra Appl.* **412** (2006), 1–21.

[10] I. Oseledets and E. Tyrtyshnikov, Approximate inversion of matrices in the process of solving a hypersingular integral equation, *Comp. Math. and Math. Phys.* **45** no. 2 (2005), 302–313. (Translated from *JVM i MF* **45** no. 2 (2005), 315–326.)

[11] V.Y. Pan, Y. Rami, Newton's iteration for the inversion of structured matrices, *Structured Matrices: Recent Developments in Theory and Computation* (Editors D.A. Bini, E.E. Tyrtyshnikov and P. Yalamov), Nova Science Publishers, Huntington, New York, 2001, 79–90.

[12] G. Schulz, Iterative Berechnung der reziproken Matrix, *Z. angew. Math. und Mech.* **13** no. 1 (1933), 57–59.

[13] W.F. Trench, An algorithm for the inversion of finite Toeplitz matrices, *SIAM J.Appl. Math.* **12** (1964), 515–521.

[14] E. Tyrtyshnikov, Optimal and superoptimal circulant preconditioners, *SIAM J. Matrix Anal. Appl.* **13** no. 2 (1992), 459-473.

[15] C.F. Van Loan and N.P. Pitsianis, Approximation with Kronecker products, *NATO Adv. Sci. Ser E Appl. Sci.* **232**, Kluwer, Dordrecht, 1993, 293–314.

Vadim Olshevsky
Department of Mathematics
University of Connecticut
Storrs CT 06269-3009
USA
e-mail: olshevsky@math.uconn.edu

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

Operator Theory:
Advances and Applications, Vol. 179, 241–256

On Generalized Numerical Ranges of Quadratic Operators

Leiba Rodman and Ilya M. Spitkovsky

Abstract. It is shown that the result of Tso-Wu on the elliptical shape of the numerical range of quadratic operators holds also for the essential numerical range. The latter is described quantitatively, and based on that sufficient conditions are established under which the c-numerical range also is an ellipse. Several examples are considered, including singular integral operators with the Cauchy kernel and composition operators.

Mathematics Subject Classification (2000). Primary 47A12; Secondary 45E05, 47B33, 47B35.

Keywords. Numerical range, essential numerical range, c-numerical range, quadratic operator, singular integral operator, composition operator.

1. Introduction

Let A be a bounded linear operator acting on a complex Hilbert space $\mathcal{H}$. Recall that the *numerical range* $W(A)$ of A is defined as

$$W(A) = \{\langle Ax, x\rangle \colon x \in \mathcal{H}, \|x\| = 1\}.$$

If c is a k-tuple of non-zero (in general, complex) numbers $c_1, \ldots, c_k$, then the *c-numerical range* of A is

$$W_c(A) = \left\{ \sum_{j=1}^{k} c_j \langle Ax_j, x_j\rangle \colon \{x_j\}_{j=1}^{k} \text{ is an orthonormal subset of } \mathcal{H} \right\}.$$

Of course, if c consists of just one number $c_1 = 1$, $W_c(A)$ is nothing but the regular numerical range of A. Also, for $c_1 = \ldots = c_k = 1$, the c-numerical range $W_c(A)$

The research of both authors was partially supported by NSF grant DMS-0456625.

turns into $W_k(A)$ – the so-called *k-numerical range*[1] introduced by Halmos; see [17]. Finally, the *essential numerical range* introduced in [31] can be defined [12] as

$$W_{\rm ess}(A) = \bigcap \mathrm{cl}\, W(A+K), \tag{1.1}$$

where the intersection is taken over all operators K that are compact on $\mathcal{H}$, and the symbol cl denotes the topological closure. Considering $W_c(A)$ or $W_{\rm ess}(A)$, we will implicitly suppose that $\dim \mathcal{H} \geq k$ or that $\mathcal{H}$ is infinite dimensional, respectively.

There are several monographs devoted to the numerical range and its various generalizations (including those mentioned above), see for example [5, 16]. We mention here only the results which are of direct relevance to the subject of this paper.

From the definitions it is clear that all three sets are unitarily invariant:

$$W(U^*AU) = W(A),\ W_c(U^*AU) = W_c(A),\ W_{\rm ess}(U^*AU) = W_{\rm ess}(A) \tag{1.2}$$

for any unitary operator U on $\mathcal{H}$. Also, they behave in a nice and predictable way under affine transformations of A:

$$W(\alpha A + \beta I) = \alpha W(A) + \beta,\ W_{\rm ess}(\alpha A + \beta I) = \alpha W_{\rm ess}(A) + \beta, \tag{1.3}$$

and

$$W_c(\alpha A + \beta I) = \alpha W_c(A) + \beta \sum_{j=1}^{k} c_j \tag{1.4}$$

for any $\alpha, \beta \in \mathbb{C}$.

It is a classical result (known as the Hausdorff-Toeplitz theorem) that the set $W(A)$ is convex. Clearly, $W_{\rm ess}(A)$ is therefore convex as well. The c-numerical range is convex if all c_j lie on the same line passing through the origin but not in general [34]. In what follows, we suppose that the c_j satisfy the above mentioned condition. Moreover, since

$$W_c(\alpha A) = W_{\alpha c}(A), \quad \alpha \in \mathbb{C},$$

we then may (and will) without loss of generality suppose that all c_j are real. We will also arrange them in non-increasing order:

$$c_1 \geq c_2 \ldots \geq c_k,$$

since permutations of the c_j leave $W_c(A)$ invariant.

When $\dim \mathcal{H} = 2$, the numerical range of A is the closed (as is always the case in the finite dimensional setting) elliptical disc with the foci at the eigenvalues λ_1, λ_2 of A and the minor axis $\sqrt{\mathrm{tr}(A^*A) - |\lambda_1|^2 - |\lambda_2|^2}$ (the elliptic range theorem, see, e.g., [16, Section 1.1]). According to the Cayley-Hamilton theorem, A in this setting satisfies the equation

$$A^2 - 2\mu A - \nu I = 0 \tag{1.5}$$

[1]We realize that there is a slight abuse of notation here, but both $W_c(A)$ and $W_k(A)$ are rather standard, and the meaning is usually clear from the content.

with

$$\mu = (\lambda_1 + \lambda_2)/2, \quad \nu = -\lambda_1\lambda_2.$$

For arbitrary $\mathcal{H}$, operators A satisfying (1.5) with some $\mu, \nu \in \mathbb{C}$ are called *quadratic operators*.

Rather recently, Tso and Wu showed that $W(A)$ is an elliptical disc (open or closed) for any quadratic operator A, independent of the dimension of $\mathcal{H}$ [32].

In this paper, we continue considering the (generalized) numerical ranges of quadratic operators. We start by stating Tso-Wu's result and outlining its proof (different from the one presented in [32]), in order to show how it can be modified to prove ellipticity of the *essential* numerical ranges of quadratic operators. We then use the combination of the two statements to derive some sufficient conditions for the c-numerical range to also have an elliptical shape. This is all done in Section 1. Section 2 is devoted to concrete implementations of the results obtained in Section 1.

2. Main results

2.1. Classical numerical range

We begin with the Tso-Wu result.

Theorem 2.1. *Let a non-scalar operator A satisfy equation* (1.5)*. Then $W(A)$ is the elliptical disc with the foci $\lambda_{1,2} = \mu \pm \sqrt{\mu^2 + \nu}$ and the major/minor axis of the length*

$$s \pm \left|\mu^2 + \nu\right| s^{-1}. \tag{2.1}$$

Here $s = \|A - \mu I\|$, and the set $W(A)$ is closed when the norm $\|A - \mu I\|$ is attained and open otherwise.

Proof. As in [32, Theorem 1.1], observe first that (1.5) guarantees unitary similarity of A to an operator of the form

$$\lambda_1 I \oplus \lambda_2 I \oplus \begin{bmatrix} \lambda_1 I & 2X \\ 0 & \lambda_2 I \end{bmatrix} \tag{2.2}$$

acting on $\mathcal{H}_1 \oplus \mathcal{H}_2 \oplus (\mathcal{H}_3 \oplus \mathcal{H}_3)$, where $\dim \mathcal{H}_j$ (≥ 0) is defined by A uniquely, and X is a positive definite operator on $\mathcal{H}_3$. According to the first of properties (1.2), we may suppose that A itself is of the form (2.2).

Using the first of formulas (1.3) we may further suppose that $\mu = 0$ and $\nu \geq 0$; in other words, that in (2.2)

$$\lambda_1 = -\lambda_2 := \lambda \geq 0, \quad \lambda^2 = \nu. \tag{2.3}$$

The case $\mathcal{H}_3 = \{0\}$ corresponds to the normal operator A when $W(A)$ is the closed line segment connecting λ_1 and λ_2. This is in agreement with formula (2.1) when $\nu \neq 0$, since in this case $s = \sqrt{\nu}$ is attained, and $s - \nu s^{-1} = 0$.

In the non-trivial case $\dim \mathcal{H}_3 > 0$ our argument is different from that in [32]. Namely, we will make use of the fact that the (directed) distance from the

origin to the support line ℓ_θ with the slope θ of $W(A)$ is the maximal point ω_θ of the spectrum of $\mathrm{Re}(ie^{-i\theta}A)$. Moreover, ℓ_θ actually contains points of $W(A)$ if and only if ω_θ belongs to the point spectrum of $\mathrm{Re}(ie^{-i\theta}A)$.

For A of the form (2.2) with λ_j as in (2.3),

$$\mathrm{Re}(ie^{-i\theta}A) = (\lambda\sin\theta)I \oplus (-\lambda\sin\theta)I \oplus \begin{bmatrix} (\lambda\sin\theta)I & ie^{-i\theta}X \\ -ie^{i\theta}X & (-\lambda\sin\theta)I \end{bmatrix}.$$

Thus,

$$\begin{aligned} &\mathrm{Re}(ie^{-i\theta}A) - \omega I \\ &= (\lambda\sin\theta - \omega)I \oplus (-\lambda\sin\theta - \omega)I \oplus \begin{bmatrix} (\lambda\sin\theta - \omega)I & ie^{-i\theta}X \\ -ie^{i\theta}X & -(\lambda\sin\theta + \omega)I \end{bmatrix}. \end{aligned} \tag{2.4}$$

For any $\omega \neq \lambda\sin\theta$, the last direct summand in (2.4) can be rewritten as

$$\begin{bmatrix} I & 0 \\ 0 & \frac{1}{\lambda\sin\theta - \omega}I \end{bmatrix} \begin{bmatrix} I & 0 \\ -ie^{i\theta}X & I \end{bmatrix} \begin{bmatrix} (\lambda\sin\theta - \omega)I & ie^{-i\theta}X \\ 0 & (\omega^2 - \lambda^2\sin^2\theta)I - X^2 \end{bmatrix}. \tag{2.5}$$

Therefore, $\omega_\theta = \sqrt{\lambda^2\sin^2\theta + \|X\|^2}$ is the rightmost point of the spectrum of $\mathrm{Re}(ie^{-i\theta}A)$. In other words, the support lines of $W(A)$ are the same as those of the numerical range of the 2×2 matrix

$$\begin{bmatrix} \lambda & 2\|X\| \\ 0 & -\lambda \end{bmatrix}.$$

The description of $W(A)$ as the elliptical disc with the foci and axes as given in the statement of the theorem follows from here and the elliptic range theorem.

Moreover, ω_θ is an eigenvalue of $\mathrm{Re}(ie^{-i\theta}A)$ if and only if the norm of X (or equivalently, of A itself) is attained, so that this either happens for all θ or for none of them. In the former case, every support line of $W(A)$ must contain at least one of its points, and the elliptical disc $W(A)$ is closed. In the latter case, the support lines are disjoint with $W(A)$, so that it is open. □

Remark. Formula (2.1) is formally different from the result of [32, Theorem 2.1], where the lengths of the axes of $W(A)$ are given in terms of $\|A - \lambda_1 I\|$, not $\|A - \mu I\|$. The two operators coincide when $\mu^2 + \nu = 0$. If this is not the case, the relation between their norms follows from the general property

$$\|P\| = \frac{1}{2}(\|S\| + \|S\|^{-1})$$

of any projection P and associated with it involution $S = 2P - I$ (see [29]) applied to $P = (A - \lambda_1 I)/(\lambda_2 - \lambda_1)$ and $S = (A - \mu I)/\sqrt{\mu^2 + \nu}$.

As a matter of fact, the relation between A and involution operators shows that A can be represented as a (rather simple) function of two orthogonal projections. This observation allows one to describe the spectra and norms of all

operators involved in the proof of Theorem 2.1 straightforwardly, using the machinery developed in [30]. We chose an independent exposition, in the interests of self-containment.

2.2. Essential numerical range

If A satisfies (1.5) and one of its eigenvalues (say λ_1) has finite multiplicity, then in representation (2.2) the spaces $\mathcal{H}_1$ and $\mathcal{H}_3$ are finite dimensional. Thus, A differs from $\lambda_2 I$ by a compact summand, and $W_{\mathrm{ess}}(A)$ is a single point. Let us exclude this trivial situation, that is, suppose that $\sigma_{\mathrm{ess}}(A) = \sigma(A) = \{\lambda_1, \lambda_2\}$.

From (1.1) it is clear that the support lines $\ell_\theta^{\mathrm{ess}}$ with the slope θ are at the distance $\omega_\theta^{\mathrm{ess}}$ from the origin. Here $\omega_\theta^{\mathrm{ess}}$ is the maximal point of the essential spectrum of $\mathrm{Re}(ie^{-i\theta}A)$. This observation allows us to repeat the statement and the proof of Theorem 2.1 almost literally, inserting the word "essential" where appropriate (of course, the last paragraph of the proof becomes irrelevant since the essential numerical range is always closed). We arrive at the following statement.

Theorem 2.2. *Let the operator A satisfy equation* (1.5)*, with both eigenvalues $\lambda_{1,2} = \mu \pm \sqrt{\mu^2 + \nu}$ having infinite multiplicity. Then $W_{\mathrm{ess}}(A)$ is the closed elliptical disc with the foci $\lambda_{1,2}$ and the major/minor axis of the length $s_0 \pm \left|\mu^2 + \nu\right| s_0^{-1}$, where s_0 is the essential norm of $A - \mu I$.*

In the trivial case $s_0 = 0$ (when A differs from μI by a compact summand, so that necessarily $\mu^2 + \nu = 0$) we by convention set $\left|\mu^2 + \nu\right| s_0^{-1} = 0$. This agrees with the fact that $W_{\mathrm{ess}}(A)$ then degenerates into a singleton μ.

Corollary 2.3. *Let the operator A satisfying* (1.5) *be such that*

$$\|A - \mu I\| > \|A - \mu I\|_{\mathrm{ess}}\,. \tag{2.6}$$

Then the elliptical disc $W(A)$ is closed.

Proof. Indeed, (2.6) holds if and only if $\|X\|_{\mathrm{ess}} < \|X\|$ for X from (2.2). Being positive definite, the operator X then has $\|X\|$ as an eigenvalue. In other words, the norm of X (and therefore of $A - \mu I$) is attained. It remains to invoke the last statement of Theorem 2.1. □

2.3. *c*-numerical range

The behavior of $W_c(A)$, even for quadratic operators, is more complicated; see [9] for some observations on the k-numerical range. With no additional assumptions on A, we give only a rather weak estimate. In what follows, it is convenient to use the notation $\|c\| = \sum_{j=1}^k |c_j|$.

Lemma 2.4. *Let A be as in Theorem* 2.2*. Denote by s and s_0 the norm and essential norm of $A - \mu I$ respectively, and by E and E_0 two elliptical discs with the foci at $\mu \sum_{j=1}^k c_j \pm \sqrt{\mu^2 + \nu}\,\|c\|$,*

- *the first closed, with the axes $(s \pm \left|\mu^2 + \nu\right| s^{-1})\,\|c\|$, and*
- *the second open, with the axes $(s_0 \pm \left|\mu^2 + \nu\right| s_0^{-1})\,\|c\|$.*

Then $W_c(A)$ contains E_0 and is contained in E.

Proof. Using (1.4) we may assume without loss of generality that $\mu = 0$, $\nu \geq 0$, as in the proof of Theorem 2.1. Since all the sets E, E_0 and $W_c(A)$ are convex, we need only to show that the support line to $W_c(A)$ in any direction lies between the respective support lines to E_0 and E. In other words, the quantity

$$\sup \left\{ \sum_{j=1}^{k} c_j \operatorname{Re}\langle ie^{-i\theta} A x_j, x_j \rangle \colon \{x_j\}_{j=1}^{k} \text{ is orthonormal} \right\} \tag{2.7}$$

must lie between

$$\|c\| \sqrt{\nu \sin^2 \theta + \|X\|_{\text{ess}}^2} \quad \text{and} \quad \|c\| \sqrt{\nu \sin^2 \theta + \|X\|^2}$$

with X given by (2.2). But this is indeed so, because (2.5) implies that the spectrum and the essential spectrum of $\operatorname{Re}(ie^{-i\theta} A)$ have the endpoints $\pm\sqrt{\nu \sin^2 \theta + \|X\|^2}$ and $\pm\sqrt{\nu \sin^2 \theta + \|X\|_{\text{ess}}^2}$, respectively. □

An interesting situation occurs when the norm of $A - \mu I$ coincides with its essential norm (equivalently, $\|X\| = \|X\|_{\text{ess}}$ for X from (2.2)), so that E is simply the closure of E_0. To state the explicit result, denote by $m_\pm$ the number of positive/negative coefficients c_j and let $m = \max\{m_+, m_-\}$.

Theorem 2.5. *Let A be as in Theorem* 2.2*, and on top of that*

$$\|A - \mu I\| = \|A - \mu I\|_{\text{ess}}. \tag{2.8}$$

Define E and E_0 as in Lemma 2.4*. Then $W_c(A)$ coincides with E if the norm of $A - \mu I$ is attained on a subspace of the dimension at least m, and with E_0 otherwise.*

Proof. Consider first a simpler case, when in (2.2) $\dim \mathcal{H}_3 < \infty$. Then due to (2.8), $\mathcal{H}_3 = \{0\}$, so that the operator A is normal. The norm $\left|\mu^2 + \nu\right|^{1/2}$ of $A - \mu I$ is attained on infinite dimensional subspaces $\mathcal{H}_1$ and $\mathcal{H}_2$, and $W_c(A)$ is the closed line segment connecting the points $\mu \sum_{j=1}^{k} c_j + \sqrt{\mu^2 + \nu}\, \|c\|$ and $\mu \sum_{j=1}^{k} c_j - \sqrt{\mu^2 + \nu}\, \|c\|$. This segment apparently coincides with E.

Let now $\mathcal{H}_3$ be infinite dimensional. From Lemma 2.4 it follows that $W_c(A)$ lies between E and its interior E_0, so that the only question is which points of the boundary of E belong to $W_c(A)$. It follows from (2.5) that the minimal and maximal points of the spectrum of $\operatorname{Re}(ie^{-i\theta} A)$ have the same multiplicity as its eigenvalues; this multiplicity does not depend on θ and coincides in fact with the dimension d (≥ 0) of the subspace on which the norm of X is attained. From (2.2) under conditions (2.3) it follows that the norm of $A - \mu I$ is attained on a d-dimensional subspace as well.

On the other hand, the supremum in (2.7) is attained if and only if this multiplicity is at least m. Thus, the boundary of E belongs to $W_c(A)$ if $d \geq m$ and is disjoint with $W_c(A)$ otherwise. □

3. Examples

We consider here several concrete examples illustrating the above-stated abstract results. All the operators A involved happen to be involutions which corresponds to the choice $\mu = 0$, $\nu = 1$ in (1.5). According to Theorems 2.1 and 2.2, the major/minor axes of the elliptical discs $W(A)$ and $W_{\mathrm{ess}}(A)$ then have the lengths

$$\|A\| \pm \|A\|^{-1} \quad \text{and} \quad \|A\|_{\mathrm{ess}} \pm \|A\|_{\mathrm{ess}}^{-1}, \tag{3.1}$$

respectively.

3.1. Singular integral operators on closed curves

Let Γ be the union of finitely many simple Jordan rectifiable curves in the extended complex plane $\dot{\mathbb{C}} = \mathbb{C} \cup \{\infty\}$. Suppose that Γ has only finitely many points of self-intersection and that it partitions $\dot{\mathbb{C}}$ into two open disjoint (not necessarily connected) sets D^+ and D^-. Moreover, we suppose that Γ is the common boundary of D^+ and D^-, and that it is oriented in such a way that the points of $D^{\pm}$ lie to the left/right of Γ.

The singular integral operator S with the Cauchy kernel is defined by

$$(S\phi)(t) = \frac{1}{\pi i} \oint_{\Gamma} \phi(\tau) \frac{d\tau}{\tau - t}. \tag{3.2}$$

It acts as an involution [14] on the linear manifold of all rational functions with the poles off Γ, dense in the Hilbert space $\mathcal{H} = L^2(\Gamma)$, with respect to the Lebesque measure on Γ. This operator is bounded in L^2 norm, and can therefore be continued to an involution acting on the whole $L^2(\Gamma)$, if and only if Γ is a so-called *Carleson curve*. This result, along with the definition of Carleson curves, as well as detailed proofs and the history of the subject, can be found in [6]. For our purposes it suffices to know that S is a bounded involution when the curve Γ is piecewise smooth, i.e., admits a piecewise continuously differentiable parametrization.

If Γ is a circle or a line, then S is in fact selfadjoint, and both its norm and essential norm are equal to 1. This situation is trivial from our point of view, since $W(S)$ and $W_{\mathrm{ess}}(S)$ then coincide with the closed interval $[-1, 1]$ and $W_c(S)$ is $[-\|c\|, \|c\|]$.

As it happens [19], circles and lines are the only simple closed curves in $\dot{\mathbb{C}}$ for which S is selfadjoint. On the other hand, for all smooth simple closed curves the essential norm of S is the same, that is, equal to 1 (see [14, Chapter 7] for Lyapunov curves; the validity of the result for general smooth curves rests on the compactness result from [15] and is well known within the singular integral community). Thus, lines and circles are the only smooth closed curves in $\dot{\mathbb{C}}$ for which the norm and the essential norm of S coincide. However, such a coincidence is possible for other *piecewise* smooth (even simple) curves.

One such case occurs when Γ is a bundle of m lines passing through a common point, or of m circles passing through two common points. According to [13], then

$$\|S\| = \|S\|_{\mathrm{ess}} \geq \cot \frac{\pi}{4m},$$

with the last inequality turning into equality for at least $m = 1, 2, 3$. Respectively, for such curves Γ the sets $W(S)$, $W_{\rm ess}(S)$ are the ellipses with the foci at ± 1, coinciding up to the boundary, and with the major axes of the length at least $2\csc\frac{\pi}{2m}$. This length equals $2\csc\frac{\pi}{2m}$ for $m = 2, 3$. The c-numerical range of S is the same ellipse, only scaled by $\|c\|$.

The equality $\|S\| = \|S\|_{\rm ess}$ also holds for Γ consisting of circular arcs (one of which can degenerate into a line segment) connecting the same two points in $\mathbb{C}$ [3, 4]; in order for an appropriate orientation on Γ to exist the number of these arcs must be even. If, in particular, there are two of them (that is, the curve Γ is simple), then

$$\|S\| = \|S\|_{\rm ess} = D_\phi + \sqrt{D_\phi^2 + 1},$$

where

$$D_\phi = \sup\left\{\frac{\sinh(\pi\phi\xi)}{\cosh(\pi\xi)} : \xi \geq 0\right\}$$

and $\pi(1-\phi)$ is the angle between the arcs forming Γ [3]. The ellipses $W(S)$, $W_{\rm ess}(S)$ therefore have the major axes of the length $2\sqrt{D_\phi^2 + 1}$.

For some particular values of ϕ the explicit value of D_ϕ can be easily computed, see [3]. If, for instance, Γ consists of a half circle and its diameter, that is $\phi = 1/2$, then $D_\phi = 1/2\sqrt{2}$. Respectively, the major axes of $W(S)$ and $W_{\rm ess}(S)$ have the length $3/\sqrt{2}$.

It would be interesting to describe all curves Γ for which the norm and the essential norm of the operator (3.2) are the same.

3.2. Singular integral operators on weighted spaces on the circle

Let now Γ be the unit circle $\mathbb{T}$. We again consider the involution (3.2), this time with $\mathcal{H}$ being the *weighted* Lebesgue space L^2_ρ. The norm on this space is defined by

$$\|f\|_{L^2_\rho} = \|\rho f\|_{L^2} := \frac{1}{\sqrt{2\pi}}\left(\int_0^{2\pi} |f(e^{i\theta})|^2 (\rho(e^{i\theta}))^2 d\theta\right)^{1/2},$$

where the weight ρ is an a.e. positive measurable and square integrable function on $\mathbb{T}$. In this setting, the operator S is closely related with the Toeplitz and Hankel operators on Hardy spaces, weighted or not. All needed definitions and "named" results used below and not supplied with explicit references can be conveniently found in the exhaustive recent monograph [25].

3.2.1. The involution S is bounded on L^2_ρ if and only if ρ^2 satisfies the Helson-Szegő condition, that is, can be represented as

$$\exp(\xi + \overline{\eta}) \text{ with } \xi, \eta \in L^\infty(\mathbb{T}) \text{ real valued and } \|\eta\|_\infty < \pi/2 \tag{3.3}$$

[25, p. 419]. This condition is equivalent to

$$\|H_\omega\| < 1, \tag{3.4}$$

where

$$\omega = \overline{\rho_+}/\rho_+, \tag{3.5}$$

ρ_+ is the outer function such that $|\rho_+| = \rho$ a.e. on $\mathbb{T}$, and H_ω denotes the Hankel operator H_ω with the symbol ω acting from the (unweighted) Hardy space H^2 to its orthogonal complement in L^2. It is also equivalent to the invertibility of the Toeplitz operator T_ω on H^2. Moreover [11],

$$\|S\|_{L^2_\rho} = \sqrt{\frac{1+\|H_\omega\|}{1-\|H_\omega\|}},$$

and a similar relation holds for the essential norms of S and H_ω. But

$$\|H_\omega\| = \operatorname{dist}(\omega, H^\infty)$$

(Nehari theorem [25, p. 3]) and

$$\|H_\omega\|_{\text{ess}} = \operatorname{dist}(\omega, H^\infty + C)$$

(Adamyan-Arov-Krein theorem [25, Theorem 1.5.3]), where H^∞ is the Hardy class of functions that are bounded analytic in $\mathbb{D}$, and its sum with the set C of continuous on $\mathbb{T}$ functions is the *Douglas algebra* $H^\infty + C$. Thus, the ellipses $W(S)$ and $W_{\text{ess}}(S)$ have the major axes

$$2/\sqrt{1-\operatorname{dist}(\omega, H^\infty)} \text{ and } 2/\sqrt{1-\operatorname{dist}(\omega, H^\infty + C)},$$

respectively.

The norm of S is attained only simultaneously with the norm of H_ω. This happens, in particular, if H_ω is compact, that is, $\omega \in H^\infty + C$. The latter condition can be restated directly in terms of ρ [11] and means that $\log \rho \in VMO$, where VMO (the class of functions with vanishing mean oscillation) is the sum of C with its harmonic conjugate $\widetilde{C}$.

Thus, for all the weights ρ such that $\log \rho \in VMO$ the ellipse $W(S)$ is closed, while $W_{\text{ess}}(S)$ degenerates into the line interval $[-1, 1]$.

A criterion for the norm of H_ω to be attained also can be given, though in less explicit form. Recall that the distance from ω to H^∞ is always attained on some $g \in H^\infty$ (this is part of Nehari's theorem). This g in general is not unique, and any f of the form $\omega - g$ is called a *minifunction*. By (another) theorem of Adamyan-Arov-Krein [25, Theorem 1.1.4], the norm of H_ω is attained if and only if the minifunction is unique and can be represented in the form

$$f(z) = \|H_\omega\| \, \overline{z\theta h}/h, \tag{3.6}$$

where θ and $h\,(\in H^2)$ are some inner and outer functions of z, respectively[2].

[2]Formally speaking, Theorem 1.1.4 in [25] contains only the "only if" part. The "if" direction is trivial, since the norm of H_ω is attained on h from (3.6); see Theorem 2.1 of the original paper [2].

3.2.2. We now turn to possible realizations of the outlined possibilities. If f admits a representation (3.6) with θ of an infinite degree (that is, being an infinite Blaschke product or containing a non-trivial singular factor), then $\|H_\omega\|$ is an s-number of H_ω having infinite multiplicity. In particular,

$$\|H_\omega\| = \|H_\omega\|_{\rm ess}\,. \tag{3.7}$$

According to Theorem 2.5, $W(S)$ in this case coincides with the closed ellipse $W_{\rm ess}(S)$, all c-numerical ranges also are closed and differ from $W(S)$ only by an appropriate scaling.

Now let θ in (3.6) be a finite Blaschke product of degree $b\,(\geq 0)$ while h is invertible in H^2. Suppose also that $|h|^2$ does not satisfy Helson-Szegő condition, that is, cannot be represented in the form (3.3) (such outer functions are easy to construct – take for example h with $|h|^{\pm 1} \in L^2$ but $|h| \notin L^{2+\epsilon}$ for any $\epsilon > 0$). Then the Toeplitz operator T_f has $(b+1)$-dimensional kernel, dense (but not closed) range [21, Corollary 3.1 and Theorem 3.16], and therefore is not left Fredholm. By Douglas-Sarason theorem [25, Theorem 1.1.15],

$$\operatorname{dist}(f, H^\infty + C) = |f| = \|H_\omega\| = \|H_f\|\,.$$

We conclude that (3.7) holds again. So, the ellipse $W(S)$ is closed and coincides with $W_{\rm ess}(S)$. According to Theorem 2.5, the c-numerical range of S is closed if the number of coefficients c_j of the same sign does not exceed $b+1$, and open otherwise.

Finally, if a unimodular function ω is such that the operator T_ω is invertible, (3.7) holds, but its minifunction is not constant a.e. in absolute value, then the norm of H_ω is not attained. Accordingly, all c-numerical ranges, $W(S)$ in particular, in this case are open.

A concrete realization of the latter possibility is given in the next subsection. All the other possibilities mentioned earlier also occur. To construct the respective weights ρ, the following procedure can be applied. Starting with any inner function θ and outer function $h \in H^2$, choose f as in (3.6) with $\|H_\omega\|$ changed to an arbitrary constant in $(0,1)$. Let ω be an 1-canonical function[3] of the Nehari problem corresponding to the Hankel operator H_f. As such, ω is unimodular, and can be represented as $\omega = g/\overline{g}$, where g is an outer function in H^2 [25, Theorem 5.1.8]. Since $\|H_\omega\| < 1$, the Toeplitz operator $T_{\omega^{-1}}$ is invertible [25, Theorem 5.1.10] (the last two cited theorems from [25] are again by Adamyan-Arov-Krein [2]). The desired weight is given by $\rho = |g|$.

By Treil's theorem [25, Theorem 12.8.1], any positive semi-definite noninvertible operator with zero or infinite dimensional kernel is unitarily similar to the modulus of a Hankel operator. Thus, the multiplicity of the norm of H_ω as its singular value can indeed assume any prescribed value, whether or not (3.7) holds.

[3]See [25, p. 156] for the definition.

3.2.3. Consider the concrete case of *power weights*

$$\rho(t) = \prod |t - t_j|^{\beta_j}, \quad t_j \in \mathbb{T},\ \beta_j \in \mathbb{R} \setminus \{0\}. \tag{3.8}$$

It is an old and well-known result that S is bounded on L^2_ρ with ρ given by (3.8) if and only if $|\beta_j| < 1/2$. This fact, along with other results about such weights cited and used below (and established by Krupnik-Verbitskii [33]) can be found in the monograph [20, Section 5].

The essential norm of S does not depend on the distribution of the nodes t_j along $\mathbb{T}$, and equals

$$\|S\|_{\text{ess}} = \cot \frac{\pi(1 - 2\tilde{\beta})}{4}, \text{ where } \tilde{\beta} = \max |\beta_j|. \tag{3.9}$$

In case of only one node (say t_0, with the corresponding exponent β_0), the norm of S is the same as (3.9). The function ω constructed by this weight ρ in accordance with (3.5) is simply $\omega(t) = t^{\beta_0}$, having a discontinuity at t_0. The distance from ω to H^∞ is the same as to $H^\infty + C$, it equals $\sin(\pi |\beta_0|)$ and is attained on a constant $\ell = \cos(\pi |\beta_0|)e^{i\pi\beta_0}$. A corresponding minifunction $f = \omega - \ell$ is not constant a.e. in absolute value; thus, it cannot admit representation (3.6). Consequently, the norm of H_ω is not attained. Accordingly, $W_c(S)$ is open for all c; the numerical range $W(S)$ has the major axis of the length $2\sec(\pi |\beta_0|)$. Other c-numerical ranges are scaled by $\|c\|$, as usual.

More generally, the norm of S coincides with (3.9) independently of the number of nodes, provided that one of the exponents (say β_0) differs by its sign from all others and at the same time exceeds or equals their sum by absolute value. The size and the shape of all the ellipses $W(S)$, $W_{\text{ess}}(S)$, $W_c(S)$ is then the same as for the weight with only one exponent β_0.

In case of two nodes (t_1 and t_2), the condition above holds if the respective exponents β_1, β_2 are of the opposite sign. If the signs are the same, the norm of S actually depends on $\arg t_1/t_2$. It takes its minimal value (for fixed β_j) when $t_1/t_2 < 0$. This value coincides with (3.9), thus making Theorem 2.5 applicable again.

3.3. Composition operators

For an analytic mapping of the unit disc $\mathbb{D}$ into itself, the *composition operator* C_ϕ is defined as

$$(C_\phi f)(z) = f(\phi(z)).$$

3.3.1. We consider this operator first on the Hardy space H^2. In this setting, the operator C_ϕ is bounded and, if ϕ is an inner function,

$$\|C_\phi\| = \sqrt{\frac{1 + |\phi(0)|}{1 - |\phi(0)|}}, \tag{3.10}$$

see [24], also [10]. It is easily seen from the proof of (3.10) given there that the norm of C_ϕ is not attained, unless $\phi(0) = 0$. As was shown in [27, 28], the essential norm

of C_ϕ for ϕ inner coincides with its norm; moreover, this property is characteristic for inner functions.

The numerical ranges of composition operators C_ϕ with ϕ being conformal automorphisms of $\mathbb{D}$ were treated in [8]. It was observed there, in particular, that $W(C_\phi)$ is an elliptical disc with the foci at ± 1 when C_ϕ is an involution, that is,

$$\phi(z) = \frac{p - z}{1 - \overline{p}z} \tag{3.11}$$

for some fixed $p \in \mathbb{D}$. The major axis of this disc E_p was computed in [1], where as a result of rather lengthy computations it was shown to equal $2/\sqrt{1 - |p|^2}$. For $p = 0$, C_ϕ is an involution of norm 1. Respectively, E_0 degenerates into the closed interval $[-1, 1]$. The question of openness or closedness of E_p for $p \neq 0$ was not discussed.

It follows from Theorem 2.1 that E_p is open (if $p \neq 0$); moreover, the length of its axes can be immediately seen from (3.1) and (3.10):

$$\sqrt{\frac{1 + |p|}{1 - |p|}} + \sqrt{\frac{1 - |p|}{1 + |p|}} = 2/\sqrt{1 - |p|^2}.$$

Furthermore, Theorem 2.2 implies that $W_{\rm ess}(C_\phi)$ is the closure of E_p. Finally, by Theorem 2.5 the c-numerical range of C_ϕ is E_p dilated by $\|c\|$.

3.3.2. These results, with some natural modifications, extend to the case of the operator C_ϕ with ϕ given by (3.11) acting on *weighted* spaces H^2_ρ. Namely, for a non-negative function $\rho \in L^2(\mathbb{T})$ with $\log \rho \in L^1$ we define the outer function ρ_+ as in (3.5). Then

$$H^2_\rho = \{f \colon \rho_+ f \in H^2\} \text{ and } \|f\|^2_{H^2_\rho} = \|\rho_+ f\|_{H^2}\,.$$

A change-of-variable argument, similar to that used in [24], shows the following equality:

$$\begin{aligned}\|C_\phi f\|^2_{H^2_\rho} &= \frac{1}{2\pi}\int_0^{2\pi} |f\left(\phi(e^{i\theta})\right)|^2 (\rho(e^{i\theta}))^2 d\theta \\ &= \frac{1}{2\pi}\int_0^{2\pi} |f(e^{i\mu})|^2 \left(\rho\left(\phi\left(e^{i\mu}\right)\right)\right)^2 \frac{1 - |p|^2}{|p - e^{i\mu}|^2} d\mu = \|f\chi\|^2_{H^2_\rho}\,, \end{aligned} \tag{3.12}$$

where

$$\chi(t) := \frac{\sqrt{1 - |p|^2}}{|p - t|}\frac{\rho(\phi(t))}{\rho(t)}, \quad t \in \mathbb{T}.$$

The norm of a multiplication operator on weighted and unweighted Hardy spaces is the same. According to (3.12) the operator C_ϕ is therefore bounded on H^2_ρ if and only if

$$\sup_{t \in \mathbb{T}} \frac{\rho(\phi(t))}{\rho(t)} < \infty. \tag{3.13}$$

Observe that (3.13) is equivalent to

$$\inf_{t\in\mathbb{T}} \frac{\rho(\phi(t))}{\rho(t)} > 0$$

because ϕ is an involution. Apparently, (3.13) holds if $\rho \in L^\infty$ is bounded below from 0, but there are plenty of unbounded weights ρ satisfying (3.13) as well.

Under this condition, $\|C_\phi\|_{H^2_\rho} = M$, where

$$M = \sqrt{1-|p|^2}\sup_{t\in\mathbb{T}} \frac{\rho(\phi(t))}{|p-t|\,\rho(t)}. \tag{3.14}$$

For any $\epsilon > 0$, consider a function $g \in H^2_\rho$ with the norm 1 and such that $\|C_\phi g\|_{H^2_\rho} > M - \epsilon$. Then $\|C_\phi g_n\|_{H^2_\rho} > M - \epsilon$ for $g_n(z) = z^n g(z)$, $n = 1, 2, \ldots$. Since the sequence g_n converges weakly to zero in H^2_ρ, from here it follows that the essential norm of C_ϕ also equals M. (We use here the well-known fact that compact operators on Hilbert spaces map weakly convergent sequences into strongly convergent sequences, see, for example, [26, Section 85].) Moreover, the norm of C_ϕ is attained if and only if there exist non-zero functions in H^2_ρ with absolute value equal zero a.e. on the subset of $\mathbb{T}$ where $|\chi(t)| \neq M$. Due to uniqueness theorem for analytic functions, a necessary and sufficient condition for this to happen is

$$\left|\frac{\rho(\phi(t))}{(p-t)\rho(t)}\right| = \text{const a.e. on } \mathbb{T}. \tag{3.15}$$

If (3.15) holds, then the norm is attained in particular on all inner functions, so that the respective subspace is infinitely dimensional. Consequently, $W_{\text{ess}}(C_\phi)$ is the closed ellipse with the foci at ± 1 and the axes $M \pm M^{-1}$, and $W(C_\phi)$ is the same ellipse when (3.15) holds or its interior when it does not. The c-numerical range is simply $\|c\|\, W(C_\phi)$.

Of course, for $\rho(t) \equiv t$ condition (3.13) holds, formula (3.14) turns into (3.10), and (3.15) is equivalent to $p = 0$. Thus, the results obtained match those already known in the unweighted setting.

3.3.3. One can also consider composition operators C_ϕ on weighted *Lebesgue* spaces L^2_ρ. Formulas for the norm and the essential norm of C_ϕ remain exactly the same, with no changes in their derivation[4]. The condition for the norm to be attained is different: in place of (3.15) it is required that the supremum in its left-hand side is attained on a set of positive measure. The respective changes in the statement about the numerical ranges are evident, and we skip them. We note only that for $\rho(t) \equiv t$ the supremum in the right-hand side of (3.15) either is attained everywhere (if $p = 0$) or just at one point (if $p \neq 0$). Thus, all the sets $W(C_\phi)$, $W_{\text{ess}}(C_\phi)$ and $W_c(C_\phi)$ are exactly the same whether the composition operator C_ϕ with the symbol (3.11) acts on H^2 or L^2.

[4]Moreover, condition $\log\rho \in L^1$ can be weakened simply to ρ being positive a.e. on $\mathbb{T}$, as was the case in Subsection 3.2.

3.3.4. Finally, we consider the operator C_ϕ on the Dirichlet space $\mathcal{D}$. Recall that the latter is defined as the set of all analytic functions f on $\mathbb{D}$ such that

$$\|f\|_{\mathcal{D}}^2 := |f(0)|^2 + \int_{\mathbb{D}} |f'(z)|^2 dA(z) < \infty,$$

where dA is the area measure.

It was shown in [22, Theorem 2] that for any univalent mapping ϕ of $\mathbb{D}$ onto a subset of full measure,

$$\|C_\phi\|_{\mathcal{D}} = \sqrt{\frac{L+2+\sqrt{L(4+L)}}{2}},$$

where $L = -\log(1-|\phi(0)|^2)$. This simplifies to

$$\|C_\phi\|_{\mathcal{D}} = \frac{\sqrt{L}+\sqrt{4+L}}{2},$$

and is of course applicable when ϕ is given by (3.11). Consequently, the elliptical disc $W(C_\phi)$ has the major axis

$$\sqrt{4+\log\frac{1}{1-|p|^2}}.$$

Moreover, the operators considered in [22, Theorem 2] attain their norms, so that $W(C_\phi)$ is closed.

It was further observed in [18, Proposition 2.4] that the essential norm of C_ϕ on $\mathcal{D}$ does not exceed 1, for any univalent ϕ. For ϕ given by (3.11), the essential norm of C_ϕ on $\mathcal{D}$ must be equal 1, since the essential norm of an involution on an infinite dimensional space is at least one. Thus, $W_{\rm ess}(C_\phi)$ in this setting is the closed interval $[-1,1]$.

Analogous remarks can be made in other contexts where the norms and essential norms of composition operators are known.

Added in proof. To illustrate this point: in [7], the composition operator C_ϕ was considered on Hardy spaces $H^2(B_N)$ and Bergman spaces $A^2(B_N)$, where B_N is the unit ball in the space C^N of N complex variables. Among other things, the length of the major axis of the ellipse $E_N = W(C_\phi)$ was computed there, for ϕ being the involutive linear-fractional transformation of B_N. It also follows from the results of [7] that, as in the setting of Subsection 3.3.1, the essential norm of C_ϕ coincides with its norm and the latter is not attained unless $\phi(0) = 0$. Thus, for $\phi(0) \neq 0$ the ellipse E_N is in fact open, $W_{\rm ess}(C_\phi)$ is its closure, and $W_c(C_\phi)$ is E_N dilated by $\|c\|$.

As we learned from the referee, the openness of the ellipse E_p in the setting of Subsection 3.3.1 was also shown in [23].

Acknowledgment

We thank V. Bolotnikov for helpful discussions concerning composition operators.

References

[1] A. Abdollahi, The numerical range of a composition operator with conformal automorphism symbol, *Linear Algebra Appl.* **408** (2005), 177–188.

[2] V.M. Adamjan, D.Z. Arov, and M.G. Krein, Infinite Hankel matrices and generalized problems of Carathéodory-Fejér and F. Riesz, *Funkcional. Anal. i Prilozhen.* **2** no. 1 (1968), 1–19 (in Russian), English translation: *Funct. Anal. and Appl.* **2** (1968), 1–18.

[3] R.E. Avendanõ, *Norm and essential norm estimates of singular integral operators*, Ph.D. thesis, Kishinev State University, 1988, 109 pp. (in Russian).

[4] R.E. Avendanõ and N.Ya. Krupnik, A local principle for calculating quotient norms of singular integral operators, *Funktsional. Anal. i Prilozhen.* **22** no. 2 (1988), 57–58 (in Russian), English translation: *Funct. Anal. Appl.* **22** (1988), 130–131.

[5] F.F. Bonsall and J. Duncan, *Numerical ranges. II*, Cambridge University Press, New York, 1973, London Mathematical Society Lecture Notes Series 10.

[6] A. Böttcher and Yu.I. Karlovich, *Carleson curves, Muckenhoupt weights, and Toeplitz operators*, Birkhäuser Verlag, Basel and Boston, 1997.

[7] P.S. Bourdon and B.D. MacCluer, Selfcommutators of automorphic composition operators, *Complex Var. Elliptic Equ.* **52** (2007), 85–104.

[8] P.S. Bourdon and J.H. Shapiro, The numerical ranges of automorphic composition operators, *J. Math. Anal. Appl.* **251** no. 2 (2000), 839–854.

[9] M.-T. Chien, S.-H. Tso, and P.Y. Wu, Higher-dimensional numerical ranges of quadratic operators, *J. Operator Theory* **49** no. 1 (2003), 153–171.

[10] C.C. Cowen and B.D. MacCluer, *Composition operators on spaces of analytic functions*, Studies in Advanced Mathematics, CRC Press, Boca Raton, FL, 1995.

[11] I. Feldman, N. Krupnik, and I.M. Spitkovsky, Norms of the singular integral operator with Cauchy kernel along certain contours, *Integral Equations and Operator Theory* **24** (1996), 68–80.

[12] P.A. Fillmore, J.G. Stampfli, and J.P. Williams, On the essential numerical range, the essential spectrum, and a problem of Halmos, *Acta Sci. Math. (Szeged)* **33** (1972), 179–192.

[13] J. Galperin and N. Krupnik, On the norms of singular integral operators along certain curves with intersections, *Integral Equations and Operator Theory* **29** no. 1 (1997), 10–16.

[14] I. Gohberg and N. Krupnik, *One-dimensional linear singular integral equations. Introduction*, vol. 1 and 2, OT 53, 54, Birkhäuser Verlag, Basel, 1992.

[15] S.M. Grudsky, *On the compactness of a certain integral operator*, No. 4856-80 dep., VINITI, Moscow, 1980 (in Russian).

[16] K.E. Gustafson and D.K.M. Rao, *Numerical range. The field of values of linear operators and matrices*, Springer, New York, 1997.

[17] P.R. Halmos, *A Hilbert space problem book*, Van Nostrand, Princeton, NJ, 1967.

[18] C. Hammond, The norm of a composition operator with linear symbol acting on the Dirichlet space, *J. Math. Anal. Appl.* **303** (2005), 499–508.

[19] N. Krupnik, The conditions of selfadjointness of the operator of singular integration, *Integral Equations and Operator Theory* **14** no. 5 (1991), 760–763.

[20] N.Ya. Krupnik, *Banach algebras with symbol and singular integral operators*, Birkhäuser, Basel and Boston, 1987.

[21] G.S. Litvinchuk and I.M. Spitkovsky, *Factorization of measurable matrix functions*, OT 25, Birkhäuser Verlag, Basel, 1987.

[22] M.J. Martın and D. Vukotić, Norms and spectral radii of composition operators acting on the Dirichlet space, *J. Math. Anal. Appl.* **304** (2005), 22–32.

[23] V. Matache, *Distances between composition operators*, *Extracta Math.* **22** (2007), 19–33.

[24] E.A. Nordgren, Composition operators, *Canad. J. Math.* **20** (1968), 442–449.

[25] V.V. Peller, *Hankel operators and their applications*, Springer, New York-Berlin-Heidelberg, 2003.

[26] F. Riesz and B. Sz.-Nagy, *Functional analysis*, Frederick Ungar Publishing Co., New York, 1955.

[27] J.H. Shapiro, The essential norm of a composition operator, *Annals of Math.* **125** (1987), 375–404.

[28] ———, What do composition operators know about inner functions?, *Monatsh. Math.* **130** no. 1 (2000), 57–70.

[29] I.M. Spitkovsky, Some estimates for partial indices of measurable matrix valued functions, *Mat. Sb. (N.S.)* **111(153)** no. 2 (1980), 227–248, 319 (in Russian), English translation: *Math. USSR Sbornik* **39** (1981), 207–226.

[30] ———, Once more on algebras generated by two projections, *Linear Algebra Appl.* **208/209** (1994), 377–395.

[31] J.G. Stampfli and J.P. Williams, Growth conditions and the numerical range in a Banach algebra, *Tôhoku Math. J.* (2) **20** (1968), 417–424.

[32] S.-H. Tso and P.Y. Wu, Matricial ranges of quadratic operators, *Rocky Mountain J. Math.* **29** no. 3 (1999), 1139–1152.

[33] I.E. Verbickii and N.Ya. Krupnik, Exact constants in theorems on the boundedness of singular operators in L_p spaces with a weight and their application, *Mat. Issled.* **54** (1980), 21–35, 165 (in Russian).

[34] R. Westwick, A theorem on numerical range, *Linear and Multilinear Algebra* **2** (1975), 311–315.

Leiba Rodman and Ilya M. Spitkovsky
Department of Mathematics
College of William and Mary
Williamsburg, VA 23185
USA
e-mail: lxrodm@math.wm.edu
ilya@math.wm.edu

Operator Theory:
Advances and Applications, Vol. 179, 257–288

Inverse Problems for Canonical Differential Equations with Singularities

James Rovnyak and Lev A. Sakhnovich

Abstract. The inverse problem for canonical differential equations is investigated for Hamiltonians with singularities. The usual notion of a spectral function is not adequate in this generality, and it is replaced by a more general notion of spectral data. The method of operator identities is used to describe a solution of the inverse problem in this setting. The solution is explicitly computable in many cases, and a number of examples are constructed.

Mathematics Subject Classification (2000). Primary 34A55; Secondary 47A57, 47B50, 47E05.

Keywords. Inverse problem, spectral function, singularity, operator identity, interpolation, generalized Nevanlinna function.

1. Introduction

By a canonical differential equation we understand a system of the form

$$\begin{aligned} \frac{dY}{dx} &= izJH(x)Y, \qquad 0 \le x < \ell, \\ D_2Y_1(0,z) + D_1Y_2(0,z) &= 0, \end{aligned} \tag{1.1}$$

where $H(x) = H(x)^*$ has $2m \times 2m$ matrix values and satisfies

$$H(x) \ge 0 \tag{1.2}$$

on $[0, \ell)$. Here ℓ is a finite positive number, z is a complex parameter,

$$J = \begin{bmatrix} 0 & I_m \\ I_m & 0 \end{bmatrix}, \qquad Y(x,z) = \begin{bmatrix} Y_1(x,z) \\ Y_2(x,z) \end{bmatrix}, \tag{1.3}$$

James Rovnyak was supported by the National Science Foundation Grant DMS-0100437.
Lev A. Sakhnovich thanks the University of Virginia for hospitality as a visitor.

where $Y_1(x,z)$, $Y_2(x,z)$ have $m\times 1$ matrix values, and D_1, D_2 are $m\times m$ matrices such that $D_1D_2^*+D_2D_1^*=0$ and $D_1D_1^*+D_2D_2^*=I_m$. Without loss of generality (see [21, p. 52]), we can take

$$D_1=0, \qquad D_2=I_m. \tag{1.4}$$

The **fundamental solution** is the $2m\times 2m$ matrix-valued function $W(x,z)$ such that

$$\frac{dW}{dx}=izJH(x)W, \qquad W(0,z)=I_{2m}. \tag{1.5}$$

With the aid of this function, we define a transform

$$Vf=F,$$

$$F(z)=\int_0^\ell \begin{bmatrix}0 & I_m\end{bmatrix} W(x,\bar z)^*\,H(x)f(x)\,dx,$$

where $f(x)$ is a $2m\times 1$ matrix-valued function on $[0,\ell)$ and $F(z)$ is an $m\times 1$ matrix-valued entire function. A nondecreasing $m\times m$ matrix-valued function $\tau(t)$ on the real line is called a **spectral function** for (1.1) if

$$\int_0^\ell f(x)^*H(x)f(x)\,dx=\int_{-\infty}^{\infty} F(t)^*\,[d\tau(t)]\,F(t) \tag{1.6}$$

for any transform pair $f(x), F(z)$. The direct problem of spectral theory is to find all spectral functions $\tau(t)$ for a given system (1.1). The inverse problem is find a system (1.1) having a given spectral function $\tau(t)$.

We recall how the inverse problem is solved in [18, 21] for systems (1.1) having locally integrable Hamiltonians $H(x)$. Let $v(z)$ be an $m\times m$ matrix-valued Nevanlinna function such that $v(iy)/y\to 0$ as $y\to\infty$. Then

$$v(z)=C_0+\int_{-\infty}^{\infty}\left[\frac{1}{t-z}-\frac{t}{1+t^2}\right]d\tau(t), \tag{1.7}$$

where $\tau(t)$ is a nondecreasing matrix-valued function such that $\int_{-\infty}^{\infty} d\tau(t)/(1+t^2)$ converges and C_0 is a constant selfadjoint $m\times m$ matrix. To construct a system (1.1) which has $\tau(t)$ as a spectral function, we choose a Hilbert space $\mathfrak{H}$, a Volterra operator $A\in\mathfrak{L}(\mathfrak{H})$, and an operator $\Phi_2\in\mathfrak{L}(\mathfrak{G},\mathfrak{H})$ where $\mathfrak{G}=\mathbf{C}^m$ in the Euclidean metric. Define operators $S=S_v$ in $\mathfrak{L}(\mathfrak{H})$ and $\Phi_1=\Phi_{1,v}$ in $\mathfrak{L}(\mathfrak{G},\mathfrak{H})$ by

$$S_v=\int_{-\infty}^{\infty}(I-At)^{-1}\Phi_2\,[d\tau(t)]\,\Phi_2^*(I-A^*t)^{-1}, \tag{1.8}$$

$$\Phi_{1,v}=-i\int_{-\infty}^{\infty}\left[A(I-At)^{-1}+\frac{tI}{t^2+1}\right]\Phi_2\,[d\tau(t)]+i\Phi_2C_0. \tag{1.9}$$

If the integral in (1.8) is weakly convergent, then so is the integral in (1.9). In this case,

$$AS-SA^*=i\left[\Phi_1\Phi_2^*+\Phi_2\Phi_1^*\right], \tag{1.10}$$

and $S\geq 0$. Let A^* have an eigenchain of projections P_x, $0\leq x\leq\ell$, with $P_0=0$ and $P_\ell=I$. Write $\mathfrak{H}_x=P_x\mathfrak{H}$, and assume that the operators $S_x=P_xSP_x|\mathfrak{H}_x$ are

invertible. Then under conditions detailed in in [21, pp. 54–55, Theorems 2.1, 2.2], the function

$$W(x,z) = I_{2m} + izJ\,\Pi^* P_x S_x^{-1} P_x (I - zA)^{-1}\Pi, \quad \Pi = \begin{bmatrix}\Phi_1 & \Phi_2\end{bmatrix}, \tag{1.11}$$

has a continuous product representation

$$\begin{aligned} W(x,z) &= \lim \exp\left\{\int_{t_{n-1}}^{t_n} izJH(t)\,dt\right\}\cdots\exp\left\{\int_{t_0}^{t_1} izJH(t)\,dt\right\} \\ &= \overset{\curvearrowright}{\int_0^x} \exp\left\{izJH(t)\,dt\right\}, \end{aligned} \tag{1.12}$$

where $0 = t_0 < t_1 < \cdots < t_n = x$ is a partition of the interval $[0,x]$ and where the limit is taken as the maximum length of the intervals in the partition tends to zero. The function $H(x)$ is extracted from this representation by the formula

$$H(x) = \frac{d}{dx}\,\Pi^* P_x S_x^{-1} P_x \Pi. \tag{1.13}$$

Moreover, the function $W(x,z)$ given by (1.11) is the fundamental solution of a canonical differential system (1.1) with Hamiltonian (1.13), and $\tau(t)$ is a spectral function for this system.

In this paper we generalize the preceding approach to the inverse problem. We retain the assumption of positivity but allow the Hamiltonian $H(x)$ to have singularities $0 < x_1 < x_2 < \cdots < \ell$ (that is, points where $H(x)$ is not locally integrable). Thus in place of (1.2) we have

$$H(x) \geq 0, \qquad x \neq x_1, x_2, \ldots\ . \tag{1.14}$$

Consider now a generalized Nevanlinna function $v(z)$ satisfying $v(iy)/y \to 0$ as $y \to \infty$. The representation (1.7) is replaced by the Kreĭn-Langer integral representation,

$$v(z) = \sum_{j=0}^{r} \int_{\Delta_j} \left[\frac{1}{t-z} - S_j(t,z)\right] d\tau(t) + R(z). \tag{1.15}$$

This representation depends on certain quantities

$$\boldsymbol{\tau} = \{\tau(t); \Delta_0, \ldots, \Delta_r; \alpha_1, \ldots, \alpha_r; \rho_1, \ldots, \rho_r; R(z)\} \tag{1.16}$$

that we call Kreĭn-Langer data (see Theorem 2.1). A transform V is defined for systems (1.1) as before. We say that (1.1) admits $\boldsymbol{\tau}$ as spectral data if

$$\int_0^\ell f(x)^* H(x) f(x)\,dx = \langle F(z), F(z)\rangle_{\boldsymbol{\tau}} \tag{1.17}$$

for all transform pairs $f(x)$ and $F(z)$, where $\langle\cdot,\cdot\rangle_{\boldsymbol{\tau}}$ is an inner product that generalizes the right side of (1.6).

To solve the inverse problem for systems with singularities, we use formulas from [12] that generalize (1.8) and (1.9) to construct an operator identity (1.10). Now we assume only that the operators S_x in the previous scheme are invertible except at certain points $0 < x_1 < x_2 < \cdots$. Then (1.11) and (1.13) define the

fundamental solution and Hamiltonian of a system (1.1) satisfying (1.2) but having singularities at the points $x_1, x_2, \ldots$. We emphasize that a system constructed in this way satisfies (1.14). Hence by the Parseval relation (1.17), the inner product $\langle \cdot, \cdot \rangle_{\boldsymbol{\tau}}$ is positive on the range of the transform V. In what follows, precise conditions will be given for the validity of the procedure just described.

It will be shown in examples that there are cases in which the calculations can be carried out explicitly. The singularities which occur are of pole type. The examples can be expanded to the complex domain, and in a number of cases it is possible to construct the global solutions to (1.5) in an explicit form.

The study of systems (1.1) has a long history, and we only mention a part of this development. Gohberg and Kreĭn [6] considered such systems on a finite interval, named them canonical differential equations, and introduced notions of eigenvalue and eigenfunction. L. de Branges [4] has obtained deep results on inverse problems by an analysis of families of Hilbert spaces of entire functions. The approach to inverse problems from the viewpoint of factorization problems and operator identities is given by L.A. Sakhnovich [18, 21]. The properties of spectral functions for canonical systems have also been investigated by A.L. Sakhnovich [16]. In a series of papers including [1] and [2], Arov and Dym have made thorough studies of inverse monodromy, inverse scattering, and inverse impedance problems for canonical systems with an emphasis on the strongly regular case. An indefinite theory is initiated in Kreĭn and Langer [9] and developed in an interesting paper by Langer and Winkler [11]. The indefinite case of canonical differential systems presents new technical difficulties. The theory of de Branges has a successful generalization to Pontryagin spaces, due to Kaltenbäck and Woracek [7]. Indefinite problems for canonical systems are studied in the simplest case of discrete systems by the authors [13], by the method of factorization and operator identities. Continuous systems are added to this theory in [15] under some simplifying assumptions. This list of references is not complete, and the sources cited here should be consulted for additional references.

The purpose of this paper is to describe classes of inverse problems in which the solution by means of operator identities produces examples of canonical differential systems (1.1) such that $H(x) \geq 0$ and $H(x)$ has singularities. We note how this paper differs from [15]. In [15] we considered systems (4.1) such that $B(x)$ has at most simple discontinuities at isolated points in $[0, \ell)$. Here we allow these points to be singularities, that is, points where $B(x)$ and $H(x) = B'(x)$ may fail to be locally integrable. Using the results of [12] and [14], we are also able to extend the theory to the full class of generalized Nevanlinna functions: in this paper we allow the points $\alpha_1, \ldots, \alpha_r$ in Theorem 2.1(1°), whereas such points are excluded in [15]. In [15] we also considered some problems with $\varkappa = \infty$, but such problems are not considered here.

In Sections 2 and 3 we formulate results from [12] and [14] that are needed for what follows. These concern the Kreĭn-Langer integral representation and operator identities associated with generalized Nevanlinna functions. In Section 4 we

construct a system with a given operator identity. The main scheme to solve the inverse problem is described in Section 5. Section 6 gives additional results for integral operators. Concrete examples are constructed in Section 7.

Notation. Throughout $\mathfrak{H}$ is a separable Hilbert space, m is a positive integer, and $\mathfrak{G} = \mathbf{C}^m$ in the Euclidean metric. By $\mathfrak{L}(\mathfrak{H})$ and $\mathfrak{L}(\mathfrak{G}, \mathfrak{H})$ we mean the usual spaces of bounded linear operators on $\mathfrak{H}$ into itself and on $\mathfrak{G}$ into $\mathfrak{H}$. Write $\mathbf{C}_\pm$ for the open upper and lower half-planes. The matrix J is as in (1.3). Let $\mathbf{N}_\varkappa$ be the **generalized Nevanlinna class** of $m \times m$ matrix-valued functions $v(z)$ which are meromorphic on $\mathbf{C}_+ \cup \mathbf{C}_-$ such that $v(z) = v(\bar{z})^*$ and the kernel

$$\frac{v(z) - v(\zeta)^*}{z - \bar{\zeta}}$$

has $\varkappa$ negative squares ($\varkappa$ a nonnegative integer). If $S \in \mathfrak{L}(\mathfrak{H})$ is a selfadjoint operator on a Hilbert space, $\varkappa_S$ is the dimension of the spectral subspace for the set $(-\infty, 0)$. Thus $\varkappa_S < \infty$ if and only if the negative spectrum consists of eigenvalues of finite total multiplicity.

2. The Kreĭn-Langer integral representation

The Kreĭn-Langer integral representation of a generalized Nevanlinna function $v(z)$ generalizes the integral formula (1.7) for classical Nevanlinna functions. The functions which occur in our applications satisfy the additional condition

$$\lim_{y \to \infty} \frac{v(iy)}{y} = 0, \tag{2.1}$$

and we state the result for this case. For the general case, see [3, 10, 14].

Theorem 2.1. *Every $m \times m$ matrix-valued function $v(z)$ which belongs to some class $\mathbf{N}_\varkappa$, $\varkappa \geq 0$, and satisfies* (2.1) *can be written as*

$$v(z) = \sum_{j=0}^{r} \int_{\Delta_j} \left[\frac{1}{t - z} - S_j(t, z) \right] d\tau(t) + R(z), \tag{2.2}$$

where $\Delta_1, \ldots, \Delta_r$ are bounded open intervals having disjoint closures, Δ_0 is the complement of their union in the real line, and

(1°) *there are points $\alpha_1, \ldots, \alpha_r$ and positive integers $\rho_1, \ldots, \rho_r$ such that $\alpha_j \in \Delta_j$, $j = 1, \ldots, r$, and*

$$\frac{1}{t - z} - S_j(t, z) = \frac{1}{t - z} \left(\frac{t - \alpha_j}{z - \alpha_j} \right)^{2\rho_j} \quad \text{on } \Delta_j, \quad j = 1, \ldots, r\,,$$

$$\frac{1}{t - z} - S_0(t, z) = \frac{1 + tz}{t - z} \, \frac{1}{1 + t^2} \quad \text{on } \Delta_0\,;$$

(2°) $\tau(t)$ *is an* $m \times m$ *matrix-valued function which is nondecreasing on each of the* $r+1$ *open intervals determined by* $\alpha_1, \dots, \alpha_r$ *such that the integral*

$$\int_{-\infty}^{\infty} \frac{(t-\alpha_1)^{2\rho_1} \cdots (t-\alpha_r)^{2\rho_r}}{(1+t^2)^{\rho_1+\cdots+\rho_r}} \, \frac{d\tau(t)}{1+t^2}$$

is convergent;

(3°) $R(z)$ *is an* $m \times m$ *matrix-valued rational function which is analytic at infinity and satisfies* $R(z) = R(\bar{z})^*$.

Conversely, every function of the form (2.2) *belongs to some class* $\mathbf{N}_\varkappa$ *and satisfies* (2.1).

Proof. This follows from Theorems 2.1 and 4.1 in [14]. □

The function $\tau(t)$ in (2.2) is essentially unique and can be recovered from $v(z)$ by a Stieltjes inversion formula [14, Corollary 3.3]. However, the other quantities in (2.2) are not unique.

We note that any function $R(z)$ satisfying (3°) can be written as

$$R(z) = C_0 - \sum_{k=1}^{s} \left[R_k\Big(\frac{1}{z-\lambda_k}\Big) + R_k\Big(\frac{1}{\bar{z}-\lambda_k}\Big)^* \right], \tag{2.3}$$

where C_0 is a constant selfadjoint $m \times m$ matrix, $\lambda_1, \dots, \lambda_s$ are distinct points in the closed upper half-plane, and $R_1(z), \dots, R_s(z)$ are polynomials such that $R_1(0) = \cdots = R_s(0) = 0$.

Definition 2.2. *The quantities*

$$\boldsymbol{\tau} = \{\tau(t); \Delta_0, \Delta_1, \dots, \Delta_r; \alpha_1, \dots, \alpha_r; \rho_1, \dots, \rho_r; R(z)\} \tag{2.4}$$

appearing in a representation (2.2) *are called* **Kreĭn-Langer data** *for* $v(z)$.

3. Operator identities

Generalized Nevanlinna functions $v(z)$ and their Kreĭn-Langer integral representations (2.2) are used to construct operator identities

$$\begin{aligned} AS - SA^* &= i\left[\Phi_1\Phi_2^* + \Phi_2\Phi_1^*\right], \\ A, S \in \mathfrak{L}(\mathfrak{H}), &\quad \Phi_1, \Phi_2 \in \mathfrak{L}(\mathfrak{G}, \mathfrak{H}), \end{aligned} \tag{3.1}$$

where $\mathfrak{H}$ is a Hilbert space, $\mathfrak{G} = \mathbf{C}^m$, and $S = S^*$. The method that we use here follows [12] and generalizes the formulas (1.8) and (1.9) from the definite case [20, 21]. In place of the inequality $S \geq 0$ which is used in [20, 21], it is assumed here and in [12] that $\varkappa_S < \infty$. We do not require the full generality of [12], since in the present applications A is a Volterra operator and $v(z)$ satisfies (2.1). In this section, we review background from [12] in the form needed in this paper.

By a Volterra operator A we mean a compact operator on a Hilbert space such that $\sigma(A) = \{0\}$.

Assumptions 3.1. *Let $A \in \mathfrak{L}(\mathfrak{H})$ and $\Phi_2 \in \mathfrak{L}(\mathfrak{G}, \mathfrak{H})$ be given operators, and let $v(z)$ be an $m \times m$ matrix-valued generalized Nevanlinna function satisfying* (2.1) *which is represented in the form* (2.2) *for associated Kreĭn-Langer data* (2.4). *Assume*

(i) *A is a Volterra operator, and*
(ii) *the integral $\int_{\Delta_0} (I - At)^{-1} \Phi_2 \, [d\tau(t)] \, \Phi_2^* (I - A^* t)^{-1}$ converges weakly.*

When these conditions are met, then following [12] we define operators $S_v \in \mathfrak{L}(\mathfrak{H})$ and $\Phi_{1,v} \in \mathfrak{L}(\mathfrak{G}, \mathfrak{H})$ by

$$S_v = \sum_{j=0}^{r} \int_{\Delta_j} \Big\{ (I - At)^{-1} \Phi_2 [d\tau(t)] \Phi_2^* \, (I - A^* t)^{-1} - d\tau_j(t; A, \Phi_2) \Big\} - \frac{1}{2\pi i} \int_{\Gamma} (I - \lambda A)^{-1} \Phi_2 R(\lambda) \Phi_2^* (I - \lambda A^*)^{-1} \, d\lambda \,, \tag{3.2}$$

$$i \, \Phi_{1,v} = \sum_{j=0}^{r} \int_{\Delta_j} \Big\{ A(I - At)^{-1} - \mathfrak{S}_j(t; A) \Big\} \Phi_2 \, [d\tau(t)] - \frac{1}{2\pi i} \int_{\Gamma} A (I - \lambda A)^{-1} \Phi_2 R(\lambda) \, d\lambda \, - \, \Phi_2 C_0. \tag{3.3}$$

In (3.2) and (3.3), Γ is any closed contour that winds once counterclockwise about each of the poles of $R(\lambda)$, that is, about each of the points $\lambda_1, \ldots, \lambda_s$ in a representation (2.3). Explicit formulas for these contour integrals are given in [12, Section 3]. The constant selfadjoint matrix C_0 plays no role and can be chosen arbitrarily. If C_0 is chosen is as in (2.3), that is, $C_0 = R(\infty)$, then (3.2) and (3.3) reduce to (1.8) and (1.9) when $\varkappa = 0$.

The convergence terms $d\tau_j(t; A, \Phi_2)$ and $\mathfrak{S}_j(t; A)$ in (3.2) and (3.3) are defined in this way. For $j = 0$, define

$$d\tau_0(t; A, \Phi_2) = 0, \qquad \mathfrak{S}_0(t; A) = -\frac{tI}{1 + t^2} \, .$$

For $j = 1, \ldots, r$, use the Taylor expansion of $(I - tA)^{-1}$ about α_j to write

$$(I - tA)^{-1} \Phi_2 [d\tau(t)] \Phi_2^* (I - tA^*)^{-1} = \sum_{\ell=0}^{\infty} (t - \alpha_j)^{\ell} \sum_{\substack{p+q=\ell \\ p,q \geq 0}} A_p(\alpha_j) \, \Phi_2 [d\tau(t)] \Phi_2^* \, A_q(\alpha_j)^*,$$

$$A(I - tA)^{-1} = \sum_{p=0}^{\infty} (t - \alpha_j)^p A_p(\alpha_j) A,$$

where $A_p(\alpha_j) = A^p (I - \alpha_j A)^{-p-1}$ for all $p \geq 0$. Then take

$$d\tau_j(t; A, \Phi_2) = \sum_{\ell=0}^{2\rho_j - 1} (t - \alpha_j)^{\ell} \sum_{\substack{p+q=\ell \\ p,q \geq 0}} A_p(\alpha_j) \, \Phi_2 [d\tau(t)] \Phi_2^* \, A_q(\alpha_j)^*,$$

$$\mathfrak{S}_j(t;A) = \sum_{p=0}^{2\rho_j-1} (t-\alpha_j)^p A_p(\alpha_j) A.$$

With these definitions, the integrals in (3.2) and (3.3) converge weakly.

The formulas (3.2) and (3.3) that define S_v and $\Phi_{1,v}$ agree with the corresponding formulas in [12]. From Theorems 3.4 and 3.5 in [12], we obtain:

Theorem 3.2. *Let A, Φ_2, and $v(z)$ satisfy Assumptions* 3.1. *Then*

(i) *the definitions of the operators S_v and $\Phi_{1,v}$ are independent of the choice of Kreĭn-Langer representation* (2.2) *for $v(z)$;*
(ii) *the operator S_v is selfadjoint, and $\varkappa_{S_v} < \infty$;*
(iii) *the operators A, Φ_2, $S = S_v$, and $\Phi_1 = \Phi_{1,v}$ satisfy*

$$AS - SA^* = i\left[\Phi_1\Phi_2^* + \Phi_2\Phi_1^*\right].$$

4. Systems associated with operator identities

The main result of this section, Theorem 4.1, shows how to construct a canonical differential equation from an operator identity. We first pass to an integral form of a system (1.1):

$$\begin{aligned} Y(x,z) &= Y(0,z) + izJ\int_0^x [dB(t)]\, Y(t,z), \\ D_2Y_1(0,z) &+ D_1Y_2(0,z) = 0, \end{aligned} \tag{4.1}$$

$0 \le x < \ell$. Here $Y(x,z)$ and J are as in (1.3). As before, we take

$$\begin{bmatrix} D_1 & D_2 \end{bmatrix} = \begin{bmatrix} 0 & I_m \end{bmatrix}.$$

In (4.1), we allow singularities at points $0 < x_1 < x_2 < \cdots$ which have no limit point in $[0,\ell)$. Thus we assume that $B(x)$ has selfadjoint $2m \times 2m$ matrix values and is continuous and nondecreasing on the intervals

$$[0,x_1), (x_1,x_2), (x_2,x_3), \dots . \tag{4.2}$$

For an interval (x_n, x_{n+1}) with $n \ge 1$, we interpret (4.1) to mean that

$$Y(b,z) - Y(a,z) = izJ\int_a^b [dB(t)]\, Y(t,z) \tag{4.3}$$

whenever $[a,b] \subseteq (x_n, x_{n+1})$. A similar meaning is attached to the equation

$$W(x,z) = I_{2m} + izJ\int_0^x [dB(t)]\, W(t,z), \tag{4.4}$$

where $W(x,z)$ is a $2m \times 2m$ matrix-valued function. On (x_n, x_{n+1}) with $n \ge 1$, we interpret (4.4) to mean that

$$W(b,z) - W(a,z) = izJ\int_a^b [dB(t)]\, W(t,z) \tag{4.5}$$

whenever $[a,b] \subseteq (x_n, x_{n+1})$. In particular, (4.3) and (4.5) hold in each of the intervals (4.2). In the usual way, (4.1) reduces to (1.1) when $B(x)$ is absolutely continuous and $H(x) = B'(x)$.

We call any solution $W(x,z)$ of (4.4) a **fundamental solution** for the system (4.1). The fundamental solution is not unique when singularities are present due to the way in which we interpret (4.4) in the intervals between the points $x_1, x_2, \dots$. A fundamental solution is continuous in x on the intervals (4.2) for fixed z, and entire in z for each fixed x. In Theorem 5.3 we show that for systems associated with operator identities, there is a distinguished choice of fundamental solution.

Theorem 4.1. *Let $A, S \in \mathfrak{L}(\mathfrak{H})$ and $\Phi_1, \Phi_2 \in \mathfrak{L}(\mathfrak{G}, \mathfrak{H})$ satisfy* (3.1) *with A Volterra, S selfadjoint, and $\varkappa_S < \infty$. Let A^* have a strongly continuous eigenchain of projections P_x, $0 \le x \le \ell$, satisfying an inequality*

$$\|(P_{x+\Delta x} - P_x)A(P_{x+\Delta x} - P_x)\| \le M\,\Delta x$$

whenever $0 \le x < x + \Delta x \le \ell$ for some $M > 0$. Assume:

(i) *there are points $0 < x_1 < x_2 < \cdots$ having no limit point in $[0,\ell)$ such that the operator $S_x = P_x S P_x|_{\mathfrak{H}_x}$ is invertible on $\mathfrak{H}_x = P_x\mathfrak{H}$ for each x in $[0,\ell) \setminus \{x_1, x_2, \dots\}$;*

(ii) *$S_x^{-1}P_x$ is a strongly continuous function of x on the intervals* (4.2).

Then the $2m \times 2m$ matrix-valued function

$$B(x) = \Pi^* P_x S_x^{-1} P_x \Pi, \qquad \Pi = \begin{bmatrix} \Phi_1 & \Phi_2 \end{bmatrix}, \tag{4.6}$$

is continuous and nondecreasing in each of the intervals (4.2), *and*

$$W(x,z) = I_{2m} + izJ\,\Pi^* P_x S_x^{-1} P_x (I - zA)^{-1}\Pi \tag{4.7}$$

is a fundamental solution for the system (4.1) *with $B(x)$ defined by* (4.6).

In Theorem 4.1, we assume that the eigenchain is indexed so that $P_0 = 0$ and $P_\ell = I$.

Lemma 4.2. *Under the assumptions in Theorem* 5.3,

$$P_\xi S_\xi^{-1} P_\xi S P_\eta S_\eta^{-1} P_\eta = P_\zeta S_\zeta^{-1} P_\zeta,$$

where $\zeta = \min\{\xi, \eta\}$ and ξ, η are any points in $[0,\ell]$ such that the inverses exist.

Proof of Lemma 4.2. If $\xi < \eta$, then $P_\xi S P_\eta|_{\mathfrak{H}_\eta} = P_\xi P_\eta S P_\eta|_{\mathfrak{H}_\eta} = P_\xi S_\eta$, and

$$P_\xi S_\xi^{-1} P_\xi S P_\eta S_\eta^{-1} P_\eta = P_\xi S_\xi^{-1} P_\xi S_\eta S_\eta^{-1} P_\eta = P_\xi S_\xi^{-1} P_\xi.$$

If $\xi \ge \eta$, then $P_\xi S P_\eta = P_\xi S P_\xi P_\eta = S_\xi P_\eta$, and

$$P_\xi S_\xi^{-1} P_\xi S P_\eta S_\eta^{-1} P_\eta = P_\xi S_\xi^{-1} S_\xi P_\eta S_\eta^{-1} P_\eta = P_\eta S_\eta^{-1} P_\eta,$$

as was to be shown. □

Proof of Theorem 4.1. By (ii), $B(x)$ is continuous in each of the intervals (4.2). To show that it is nondecreasing in these intervals, it is sufficient to show that for each $u \in \mathfrak{G} \times \mathfrak{G}$, the function

$$\beta(x) = u^* B(x) u$$

is nondecreasing in the intervals. We assume that $\beta(a) > \beta(b)$ for some compact subinterval $[a, b]$ of one of the intervals (4.2) and derive a contradiction. Since $\beta(x)$ is continuous on $[a, b]$, by the intermediate value theorem, for any positive integer $r > \varkappa_S$ we can find points $a_1 > b_1 > a_2 > b_2 > \cdots > a_r > b_r$ in $[a, b]$ such that

$$\beta(a_1) > \beta(b_1) > \beta(a_2) > \beta(b_2) > \cdots > \beta(a_r) > \beta(b_r).$$

For each $j = 1, \ldots, r$, set

$$\begin{aligned} \delta_j &= \beta(b_j) - \beta(a_j), \\ f_j &= P_{b_j} S_{b_j}^{-1} P_{b_j} \Pi u - P_{a_j} S_{a_j}^{-1} P_{a_j} \Pi u. \end{aligned}$$

By Lemma 4.2, if $\xi, \eta \in [a, b]$,

$$\left\langle S P_\eta S_\eta^{-1} P_\eta \Pi u, P_\xi S_\xi^{-1} P_\xi \Pi u \right\rangle = u^* \Pi^* P_\zeta S_\zeta^{-1} P_\zeta \Pi u = \beta(\zeta),$$

where $\zeta = \min\{\xi, \eta\}$. It follows that

$$\langle S f_j, f_k \rangle = \begin{cases} \delta_j, & j = k, \\ 0, & j \neq k. \end{cases}$$

For when $j = k$,

$$\langle S f_j, f_j \rangle = \beta(b_j) - \beta(a_j) - \beta(a_j) + \beta(a_j) = \delta_j.$$

If $j < k$,

$$\langle S f_j, f_k \rangle = \beta(b_j) - \beta(b_j) - \beta(a_j) + \beta(a_j) = 0,$$

and similarly if $j > k$. Since $\delta_j < 0$ for each j, $\mathfrak{H}$ contains an r-dimensional subspace $\mathfrak{N}$ which is the antispace of a Hilbert space in the inner product

$$\langle S f, g \rangle, \qquad f, g \in \mathfrak{N}.$$

This is impossible since $r > \varkappa_S$ (because the projection of $\mathfrak{N}$ into the spectral subspace of S for the negative axis is one-to-one). It follows that $B(x)$ is nondecreasing on each of the intervals (4.2). [In the case $S \geq 0$, a different argument to show that $B(x)$ is nondecreasing is given in [21, p. 42].]

The proof that $W(x, z)$ is a fundamental solution for the resulting system is essentially identical to the first part of the argument in [12, Theorem 3.3]. An extra condition is used in [15, Theorem 3.3], namely, that $\|S_x^{-1}\|$ is bounded on $[0, \ell]$. In our case, $\|S_x^{-1}\|$ is locally bounded by (ii) and the uniform boundedness principle, and this is all that is needed in the argument. □

5. Spectral data and the inverse problem

Consider a system (4.1) with fundamental solution $W(x,z)$. Define a transform

$$\begin{gathered} Vf = F, \\ F(z) = \int_0^\ell \begin{bmatrix} 0 & I_m \end{bmatrix} W(x,\bar{z})^* \, [dB(t)] \, f(x), \end{gathered} \tag{5.1}$$

where $f(x)$ is a $2m \times 1$ matrix-valued function on $[0,\ell)$. We assume that $f(x)$ is compactly supported, vanishes in an open interval about each point $x_1, x_2, \ldots$, and is continuous except for a finite number of simple discontinuities. For each such $f(x)$, the corresponding $F(z)$ is an $m \times 1$ matrix-valued entire function.

Let $v(z)$ be an $m \times m$ matrix-valued function in $\mathbf{N}_\varkappa$ satisfying (2.1) with Kreĭn-Langer data $\boldsymbol{\tau}$ given by (2.4). If $F(z)$ and $G(z)$ are $m \times 1$ matrix-valued entire functions such that the integrals

$$\int_{\Delta_0} F(t)^* \, [d\tau(t)] \, F(t), \qquad \int_{\Delta_0} G(t)^* \, [d\tau(t)] \, G(t) \tag{5.2}$$

converge, we define

$$\begin{aligned} \langle F(z), G(z) \rangle_{\boldsymbol{\tau}} = \sum_{j=0}^{r} \int_{\Delta_j} \Big\{ G(t)^* \, [d\tau(t)] \, F(t) - d\sigma_j(t; F, G) \Big\} \\ - \frac{1}{2\pi i} \int_\Gamma G(\bar{\lambda})^* R(\lambda) F(\lambda) \, d\lambda. \end{aligned} \tag{5.3}$$

In the last term, Γ is any closed contour that winds once counterclockwise about each of the poles of $R(\lambda)$. In the first integral term, we take

$$d\sigma_0(t; F, G) = 0.$$

Then the integral $\int_{\Delta_0}$ in (5.3) converges since the two integrals in (5.2) converge. For $j = 1, \ldots, r$, use the Taylor series $F(t) = \sum_{p=0}^{\infty} F_p(\alpha_j)(t-\alpha_j)^p$ and $G(t) = \sum_{q=0}^{\infty} G_q(\alpha_j)(t-\alpha_j)^q$ to formally write

$$G(t)^* \, [d\tau(t)] \, F(t) = \sum_{\ell=0}^{\infty} (t-\alpha_j)^\ell \sum_{\substack{p+q=\ell \\ p,q \geq 0}} G_q(\alpha_j)^* \, [d\tau(t)] \, F_p(\alpha_j).$$

Then choose

$$d\sigma_j(t; F, G) = \sum_{\ell=0}^{2\rho_j - 1} (t-\alpha_j)^\ell \sum_{\substack{p+q=\ell \\ p,q \geq 0}} G_q(\alpha_j)^* \, [d\tau(t)] \, F_p(\alpha_j).$$

With this choice, the integral $\int_{\Delta_j}$ in (5.3) converges by the condition (2°) in Theorem 2.1.

Lemma 5.1. *Let $\boldsymbol{\tau}$ be Kreĭn-Langer data for a function $v(z) \in \mathbf{N}_\varkappa$ which satisfies* (2.1)*. Suppose $F(z) = \Phi_2^*(I - zA^*)^{-1}f$ and $G(z) = \Phi_2^*(I - zA^*)^{-1}g$, $f, g \in \mathfrak{H}$, where $A \in \mathfrak{L}(\mathfrak{H})$ and $\Phi_2 \in \mathfrak{L}(\mathfrak{G}, \mathfrak{H})$ satisfy the conditions in Assumptions* 3.1 *and S_v is defined by* (3.2)*. Then*

$$\langle F(z), G(z)\rangle_{\boldsymbol{\tau}} = \langle S_v f, g\rangle\,.$$

Proof. If the integrals in (3.2) are interpreted in the weak sense, the inner product $\langle S_v f, g\rangle$ reduces to (5.3) by the definition of S_v in Section 3. □

Definition 5.2. *Consider a system* (4.1) *with fundamental solution $W(x, z)$ and transform V defined by* (5.1)*. Let*

$$\boldsymbol{\tau} = \{\tau(t); \Delta_0, \Delta_1, \dots, \Delta_r; \alpha_1, \dots, \alpha_r; \rho_1, \dots, \rho_r; R(z)\}$$

be Kreĭn-Langer data for an $m \times m$ matrix-valued function in $\mathbf{N}_\varkappa$ satisfying (2.1)*. We call $\boldsymbol{\tau}$* **spectral data** *for the system* (4.1) *if the Parseval identity*

$$\int_0^\ell g(t)^*\,[dB(t)]\,f(t) = \langle F(z), G(z)\rangle_{\boldsymbol{\tau}} \tag{5.4}$$

holds for all transform pairs $f(t), F(z)$ and $g(t), G(z)$.

We now describe a solution to the inverse problem for systems (4.1).

Theorem 5.3. *Let $v(z) \in \mathbf{N}_\varkappa$ be an $m \times m$ matrix-valued function satisfying* (2.1) *which has Kreĭn-Langer data $\boldsymbol{\tau}$. Choose operators $A \in \mathfrak{L}(\mathfrak{H})$ and $\Phi_2 \in \mathfrak{L}(\mathfrak{G}, \mathfrak{H})$ satisfying Assumptions* 3.1*, and define $S = S_v$ and $\Phi_1 = \Phi_{1,v}$ by* (3.2) *and* (3.3)*. Then S is selfadjoint, $\varkappa_S < \infty$, and A, S, Φ_1, Φ_2 satisfy* (3.1)*. Any system* (4.1) *constructed from these operators by means of the formulas* (4.6) *and* (4.7) *in Theorem* 4.1 *has spectral data $\boldsymbol{\tau}$.*

Proof. The stated properties of A, S, Φ_1, Φ_2 follow from Theorem 3.2. Let γ, δ of $[0, \ell)$ be subintervals of $[0, \ell)$ whose closures do not contain any of the points $x_1, x_2, \dots$. We first prove the identity

$$\int_0^\ell g_\delta(t)^*\,[dB(t)]\,f_\gamma(t) = \langle F_\gamma(z), G_\delta(z)\rangle_{\boldsymbol{\tau}} \tag{5.5}$$

for any transform pairs $f_\gamma(x), F_\gamma(z)$ and $g_\delta(x), G_\delta(z)$ such that

$$f_\gamma(x) = \chi_\gamma(x)u, \qquad g_\delta(x) = \chi_\delta(x)v, \tag{5.6}$$

where $u, v \in \mathfrak{G}$. We allow the possibility that γ and δ are contained different intervals in the list (4.2). Clearly,

$$\int_0^\ell g_\delta(t)^*\,[dB(t)]\,f_\gamma(t) = \int_{\gamma\cap\delta} v^*\,[dB(t)]\,u. \tag{5.7}$$

If $\gamma = [a, b]$, then by (4.5),

$$\begin{aligned} F_\gamma(z) &= \int_a^b \begin{bmatrix} 0 & I_m \end{bmatrix} W(t, \bar{z})^* \, [dB(t)] \, u \\ &= \begin{bmatrix} 0 & I_m \end{bmatrix} \frac{W(b, \bar{z})^* J - W(a, \bar{z})^* J}{-iz} \, u. \end{aligned}$$

Hence by (4.7),

$$\begin{aligned} F_\gamma(z) = \begin{bmatrix} 0 & I_m \end{bmatrix} \Big\{ & \Pi^* (I - zA^*)^{-1} P_b S_b^{-1} P_b \Pi u \\ & - \Pi^* (I - zA^*)^{-1} P_a S_a^{-1} P_a \Pi u \Big\} \\ = \Phi_2^* (I - zA^*)^{-1} (h_b - h_a), \end{aligned}$$

where $h_a = P_a S_a^{-1} P_a \Pi u$ and $h_b = P_b S_b^{-1} P_b \Pi u$. Similarly, if $\delta = [c, d]$, then

$$G_\delta(z) = \Phi_2^* (I - zA^*)^{-1} (k_d - k_c),$$

where $k_c = P_c S_c^{-1} P_c \Pi v$ and $k_d = P_d S_d^{-1} P_d \Pi v$. Now set $S = S_v$, and apply Lemma 5.1 to get

$$\begin{aligned} \langle F_\gamma(z), G_\delta(z) \rangle_{\boldsymbol{\tau}} &= \langle S(h_b - h_a), k_d - k_c \rangle \\ &= \left\langle S P_b S_b^{-1} P_b \Pi u, P_d S_d^{-1} P_d \Pi v \right\rangle - \left\langle S P_b S_b^{-1} P_b \Pi u, P_c S_c^{-1} P_c \Pi v \right\rangle \\ &\quad - \left\langle S P_a S_a^{-1} P_a \Pi u, P_d S_d^{-1} P_d \Pi v \right\rangle + \left\langle S P_a S_a^{-1} P_a \Pi u, P_c S_c^{-1} P_c \Pi v \right\rangle. \end{aligned} \tag{5.8}$$

Case 1: $\gamma \cap \delta = \emptyset$. If $a < b < c < d$, then by (5.7), (5.8), and Lemma 4.2,

$$\begin{aligned} \langle F_\gamma(z), G_\delta(z) \rangle_{\boldsymbol{\tau}} &= v^* \Pi^* P_b S_b^{-1} P_b \Pi u - v^* \Pi^* P_b S_b^{-1} P_b \Pi u \\ &\quad - v^* \Pi^* P_a S_a^{-1} P_a \Pi u + v^* \Pi^* P_a S_a^{-1} P_a \Pi u = 0 = \int_0^\ell g_\delta(t)^* \, [dB(t)] \, f_\gamma(t). \end{aligned}$$

Case 2: $\gamma \cap \delta \neq \emptyset$. Here we can assume that $a \le c \le b \le d$. As above,

$$\begin{aligned} \langle F_\gamma(z), G_\delta(z) \rangle_{\boldsymbol{\tau}} &= v^* \Pi^* P_b S_b^{-1} P_b \Pi u - v^* \Pi^* P_c S_c^{-1} P_c \Pi u \\ &\quad - v^* \Pi^* P_a S_a^{-1} P_a \Pi u + v^* \Pi^* P_a S_a^{-1} P_a \Pi u \\ &= \int_c^b g_\delta(t)^* \, [dB(t)] \, f_\gamma(t) = \int_0^\ell g_\delta(t)^* \, [dB(t)] \, f_\gamma(t). \end{aligned}$$

The general case follows by linearity and approximation. □

Corollary 5.4. *In the situation of Theorem* 5.3, $\langle F(z), G(z) \rangle_{\boldsymbol{\tau}}$ *is a strictly positive inner product on the range of the transform* (5.1).

Proof. The inner product is nonnegative by the Parseval formula (5.4) and the fact, established in Theorem 4.1, that $B(x)$ is nondecreasing in the intervals (4.2). If $\langle F(z), F(z) \rangle_{\boldsymbol{\tau}} = 0$ for some transform pair $f(x), F(z)$, the same identity implies

that $\int_0^\ell f(t)^* \, [dB(t)] \, f(t) = 0$. Then $F(z) \equiv 0$ by (5.1) and the Cauchy-Schwarz inequality. □

In many examples of Theorem 5.3, (4.1) is equivalent to a system

$$\begin{gathered} \frac{dY}{dx} = izJH(x)Y, \qquad 0 \le x < \ell, \\ Y_1(0,z) = 0, \end{gathered} \tag{5.9}$$

where $H(x)$ has the form (see Theorem 6.2)

$$H(x) = \begin{bmatrix} h_1(x)^* \\ h_2(x)^* \end{bmatrix} \begin{bmatrix} h_1(x) & h_2(x) \end{bmatrix}. \tag{5.10}$$

Then it is natural to consider an alternative form for the transform (5.1). In (5.1) write

$$f(x) = \begin{bmatrix} f_1(x) \\ f_2(x) \end{bmatrix}, \qquad W(x,z) = \begin{bmatrix} W_{11}(x,z) & W_{12}(x,z) \\ W_{21}(x,z) & W_{22}(x,z) \end{bmatrix}$$

and

$$\begin{aligned} g(x) &= h_1(x)f_1(x) + h_2(x)f_2(x), \\ \psi(x,z) &= W_{12}(x,\bar{z})^* \, h_1(x)^* + W_{22}(x,\bar{z})^* \, h_2(x)^*. \end{aligned}$$

Then (5.1) assumes the form

$$\begin{gathered} \widetilde{V} g = G, \\ G(z) = \int_0^\ell \psi(x,z) g(x) \, dx. \end{gathered} \tag{5.11}$$

The Parseval relation (5.4) becomes

$$\int_0^\ell g(t)^* g(t) \, dt = \langle G(z), G(z) \rangle_{\boldsymbol{\tau}} \tag{5.12}$$

in this case.

Paley-Wiener example. The simplest example of Theorem 5.3 yields the Paley-Wiener transform. Consider the spectral data $\tau(t) = t/(2\pi)$ on $\Delta_0 = (-\infty, \infty)$ and associated Nevanlinna function $v(z) = i/2$, $\operatorname{Im} z > 0$. We apply Theorem 5.3 with $\mathfrak{H} = L^2(0,\ell)$, $\mathfrak{G} = \mathbf{C}$, and

$$(Af)(x) = i \int_0^x f(t) \, dt \qquad \text{and} \qquad (\Phi_2 c)(x) = c$$

for all $f \in L^2(0,\ell)$ and $c \in \mathbf{C}$. We find that $S_v = I$ and $(\Phi_{1,v} c)(x) = \frac{1}{2} c$ for all $c \in \mathbf{C}$. If P_ξ is the projection onto $L^2(0,\xi)$, short calculations of the quantities (4.6) and (4.7) yield

$$B(\xi) = \begin{bmatrix} \frac{1}{4} & \frac{1}{2} \\ \frac{1}{2} & 1 \end{bmatrix} \xi, \qquad H(\xi) = \begin{bmatrix} \frac{1}{4} & \frac{1}{2} \\ \frac{1}{2} & 1 \end{bmatrix},$$

$$W(\xi,z)=\begin{bmatrix}\frac{1}{2} & -1\\ -\frac{1}{4} & \frac{1}{2}\end{bmatrix}+e^{iz\xi}\begin{bmatrix}\frac{1}{2} & 1\\ \frac{1}{4} & \frac{1}{2}\end{bmatrix}.$$

The transform (5.11) and Parseval relation (5.12) are given by

$$G(z)=\int_0^\ell e^{-izx}g(x)\,dx \tag{5.13}$$

and

$$\int_0^\ell |g(x)|^2\,dx=\frac{1}{2\pi}\int_{-\infty}^{\infty}|G(x)|^2\,dx. \tag{5.14}$$

The associated canonical differential equation, of course, has no singularities. Examples with singularities are given in Section 7.

6. Integral operators

In Theorem 6.1 we identify a large class of operator identities (3.1) for which the hypotheses of Theorem 4.1 are satisfied. Theorem 6.2 shows that in many cases a $2m\times 2m$ matrix-valued Hamiltonian $H(x)=B'(x)$ obtained from (4.6) satisfies $\operatorname{rank}H(x)\equiv m$ except at the points of singularity. For difference-kernel operators, the Hamiltonian has a special form, which is given in Theorem 6.3.

Let $\mathfrak{H}=L^2_m(0,\ell)$ and $\mathfrak{G}=\mathbf{C}^m$ for some positive integer m. Let P_ξ be the projection of $\mathfrak{H}$ onto $\mathfrak{H}_\xi=L^2_m(0,\xi)$, $0\le\xi\le\ell$. Assume that

$$(Af)(x)=i\int_0^x f(t)\,dt,\qquad f\in L^2_m(0,\ell), \tag{6.1}$$

and that Φ_1 and Φ_2 are operators on $\mathfrak{G}$ into $\mathfrak{H}$ given by

$$(\Phi_1 g)(x)=\varphi_1(x)g,\qquad (\Phi_2 g)(x)=\varphi_2(x)g,\qquad g\in\mathbf{C}^m, \tag{6.2}$$

where $\varphi_1(x)$ and $\varphi_2(x)$ are continuous $m\times m$ matrix-valued functions. Let

$$\begin{gathered}(Sf)(x)=f(x)+\int_0^\ell K(x,t)f(t)\,dt,\\ K(x,t)=K(t,x)^*,\quad x,t\in(0,\ell),\end{gathered} \tag{6.3}$$

where $K(x,t)$ is a bounded continuous $m\times m$ matrix-valued function. We assume that the identity

$$AS-SA^*=i\left[\Phi_1\Phi_2^*+\Phi_2\Phi_1^*\right] \tag{6.4}$$

is satisfied.

When the operators A,S,Φ_1,Φ_2 are as in (6.1)–(6.4), the formula (4.6) for the Hamiltonian takes the form

$$H(\xi)=B'(\xi)=\frac{d}{d\xi}\begin{bmatrix}\left\langle S_\xi^{-1}\varphi_1,\varphi_1\right\rangle_\xi & \left\langle S_\xi^{-1}\varphi_2,\varphi_1\right\rangle_\xi\\ \left\langle S_\xi^{-1}\varphi_1,\varphi_2\right\rangle_\xi & \left\langle S_\xi^{-1}\varphi_2,\varphi_2\right\rangle_\xi\end{bmatrix}, \tag{6.5}$$

where $\langle \cdot, \cdot \rangle_\xi$ denotes the inner product in $L^2_m(0,\xi)$. In (6.5) and below, we understand that operations on matrix-valued functions are performed as required on the columns of the functions.

The next result gives a sufficient condition for the technical hypotheses of Theorem 4.1 to be met.

Theorem 6.1. *The hypotheses of Theorem* 4.1 *are satisfied if* A, S, Φ_1, Φ_2 *are as in* (6.1)–(6.4)*, and if* $K(x,t)$ *has an extension to a function* $K(z,\bar{w})$ *which is bounded and analytic as functions of* z *and* w *in a region* G *such that* G *contains the interval* $(0,\ell)$ *and* $zt, \bar{z}t \in G$ *whenever* $z \in G$ *and* $0 < t \le 1$*.*

Proof. It is clear from (6.3) that S is selfadjoint, and $\varkappa_S < \infty$ since S is a compact perturbation of the identity operator. The assumptions on A are verified in a routine way. The main problem is to check the conditions (i) and (ii) in Theorem 4.1.

(i) For small ξ, the operator

$$(S_\xi f)(x) = f(x) + \int_0^\xi K(x,t) f(t)\, dt$$

on $L^2_m(0,\xi)$ differs from the identity operator by an operator of norm less than one. Therefore S_ξ is invertible for $0 \le \xi < \varepsilon$ for some $\varepsilon > 0$; for $\xi = 0$, S_ξ is the identity operator on the zero space and hence invertible.

For each ξ in $(0,\ell)$, define U_ξ from $L^2_m(0,\ell)$ to $L^2_m(0,\xi)$ by

$$(U_\xi f)(x) = \sqrt{\frac{\ell}{\xi}}\, f\left(\frac{\ell x}{\xi}\right), \qquad 0 < x < \xi.$$

Then U_ξ maps $L^2_m(0,\ell)$ isometrically onto $L^2_m(0,\xi)$, and

$$(U_\xi^{-1} g)(x) = \sqrt{\frac{\xi}{\ell}}\, g\left(\frac{\xi x}{\ell}\right), \qquad 0 < x < \ell.$$

Hence $U_\xi^{-1} S_\xi U_\xi$ is a bounded operator on $L^2_m(0,\ell)$ given by

$$(U_\xi^{-1} S_\xi U_\xi f)(x) = f(x) + \frac{\xi}{\ell} \int_0^\ell K\left(\frac{\xi x}{\ell}, \frac{\xi t}{\ell}\right) f(t)\, dt.$$

Clearly S_ξ is invertible if and only if $U_\xi^{-1} S_\xi U_\xi$ is invertible. Write

$$U_\xi^{-1} S_\xi U_\xi = I + T(\xi). \tag{6.6}$$

The assumptions on G allow us to define an operator $T(z)$ on $L^2_m(0,\ell)$ by

$$(T(z)f)(x) = \frac{z}{\ell} \int_0^\ell K\left(\frac{zx}{\ell}, \frac{\bar{z}t}{\ell}\right) f(t)\, dt, \qquad z \in G.$$

The operator $T(z)$ is compact and depends holomorphically on z, and $T(z)$ agrees with the operator $T(\xi)$ defined by (6.6) when $z = \xi$ is a point of $(0,\ell)$. Since $I + T(\xi)$ is invertible for small positive ξ, $I + T(z)$ is invertible except at isolated points of G (see Kato [8], Theorem 1.9 on p. 370). In particular, (i) follows.

(ii) In this condition, we interpret S_ξ^{-1} as acting from $L_m^2(0,\ell)$ into itself. Hence, for $0 < \xi < \ell$,

$$S_\xi^{-1} P_\xi = E_\xi U_\xi F(\xi) U_\xi^{-1} P_\xi, \tag{6.7}$$

where $F(\xi) = [I + T(\xi)]^{-1}$ and E_ξ is the natural embedding of $L_m^2(0,\xi)$ into $L_m^2(0,\ell)$. In any interval that does not include singularities, all of the operators on the right side of (6.7) are locally bounded. The function $F(\xi)$ is continuous in the operator norm, and one checks easily that $U_\xi^{-1} P_\xi$ and $E_\xi U_\xi$ are strongly continuous. It follows that $S_\xi^{-1} P_\xi$ is strongly continuous at any point ξ in $(0,\ell)$ which is not one of the singularities $x_1, x_2, \ldots$. The strong continuity of $S_\xi^{-1} P_\xi$ at the point $\xi = 0$ is clear because $\|S_\xi^{-1}\|$ is bounded for small ξ by (6.7), and $\|P_\xi f\| \to 0$ as $\xi \to 0$ for every f in $L_m^2(0,\ell)$. □

Theorem 6.2. *Let $B(x)$ be constructed by* (4.6) *for operators A, S, Φ_1, Φ_2 as in* (6.1)–(6.4). *Then $B(x)$ is continuously differentiable in the intervals between singularities, and in these intervals*

$$H(\xi) = B'(\xi) = \begin{bmatrix} h_1(\xi)^* \\ h_2(\xi)^* \end{bmatrix} \begin{bmatrix} h_1(\xi) & h_2(\xi) \end{bmatrix}, \tag{6.8}$$

where $h_1(\xi)$ and $h_2(\xi)$ are continuous $m \times m$ matrix-valued functions.

Proof. We use results from [6, Chapter IV, §7], which should be consulted for additional details. For each ξ,

$$(S_\xi f)(x) = f(x) + \int_0^\xi K(x,t) f(t)\, dt, \qquad f \in L_m^2(0,\xi).$$

Suppose that S_ξ is invertible for $x_1 < \xi < x_2$. Then

$$(S_\xi^{-1} f)(x) = f(x) + \int_0^\xi \Gamma_\xi(x,t) f(t)\, dt, \qquad f \in L_m^2(0,\xi),$$

where $\Gamma_\xi(x,t)$ is continuous in x and t and differentiable in ξ, and

$$\frac{\partial}{\partial \xi} \Gamma_\xi(x,t) = \Gamma_\xi(x,\xi) \Gamma_\xi(\xi,t). \tag{6.9}$$

The last formula is (7–10) in [6, Chapter IV, §7]. We shall compute $H(\xi)$ using (6.5). For any continuous functions f and g in $L_m^2(0,\ell)$,

$$\left\langle S_\xi^{-1} P_\xi f, P_\xi g \right\rangle_\xi = \int_0^\xi g(x)^* \left[f(x) + \int_0^\xi \Gamma_\xi(x,t) f(t)\, dt \right] dx.$$

Differentiation yields

$$\frac{d}{d\xi}\Big\langle S_\xi^{-1}P_\xi f, P_\xi g\Big\rangle_\xi = g(\xi)^* f(\xi) + \int_0^\xi g(\xi)^*\Gamma_\xi(\xi,t)f(t)\,dt + \int_0^\xi g(x)^*\Gamma_\xi(x,\xi)f(\xi)\,dx + \int_0^\xi\int_0^\xi g(x)^*\,\frac{\partial}{\partial\xi}\,\Gamma_\xi(x,t)\,f(t)\,dt\,dx\,.$$

By (6.9),

$$\begin{aligned}\frac{d}{d\xi}\Big\langle S_\xi^{-1}P_\xi f, P_\xi g\Big\rangle_\xi &= g(\xi)^* f(\xi) + \int_0^\xi g(\xi)^*\Gamma_\xi(\xi,t)f(t)\,dt \\ &\quad + \int_0^\xi g(x)^*\Gamma_\xi(x,\xi)f(\xi)\,dx \\ &\quad + \int_0^\xi\int_0^\xi g(x)^*\Gamma_\xi(x,\xi)\Gamma_\xi(\xi,t)f(t)\,dt\,dx \\ &= \left[g(\xi)^* + \int_0^\xi g(x)^*\Gamma_\xi(x,\xi)\,dx\right] \\ &\quad\cdot\left[f(\xi) + \int_0^\xi \Gamma_\xi(\xi,t)f(t)\,dt\right].\end{aligned}$$

By (6.5), on choosing $f = \varphi_j$ and $g = \varphi_k$, $j,k = 1,2$, we obtain (6.8) with

$$h_1(\xi) = \varphi_1(\xi) + \int_0^\xi \Gamma_\xi(\xi,t)\varphi_1(t)\,dt,$$
$$h_2(\xi) = \varphi_2(\xi) + \int_0^\xi \Gamma_\xi(\xi,t)\varphi_2(t)\,dt,$$

which yields the result. □

We suppose next that the operator (6.2) has a difference kernel: $K(x,t) = k(x-t)$. That is, we assume that S is defined on $L_m^2(0,\ell)$ by

$$\begin{aligned}(Sf)(x) &= f(x) + \int_0^\ell k(x-t)f(t)\,dt, \\ k(x) &= k(-x)^*, \quad x\in(-\ell,\ell),\end{aligned} \tag{6.10}$$

where $k(x)$ is a bounded continuous $m\times m$ matrix-valued function on $(-\ell,\ell)$. By writing

$$s(x) = \begin{cases} \frac{1}{2}I_m + \displaystyle\int_0^x k(u)\,du, & 0<x<\ell, \\ -\frac{1}{2}I_m + \displaystyle\int_0^x k(u)\,du, & -\ell<x<0, \end{cases} \tag{6.11}$$

we can bring (6.10) to the form (see [19]):

$$\begin{aligned}(Sf)(x) &= \frac{d}{dx}\int_0^\ell s(x-t)f(t)\,dt,\\ s(x) &= -s(-x)^*, \quad x \in (-\ell,\ell).\end{aligned} \tag{6.12}$$

Define A, Φ_1, and Φ_2 by (6.1) and (6.2), with

$$\varphi_1(x) = s(x), \qquad \varphi_2(x) = I_m, \qquad 0 < x < \ell. \tag{6.13}$$

The condition (6.4) is easily checked by direct calculation.

Theorem 6.3. *Let $B(x)$ be constructed by* (4.6) *for operators A, S, Φ_1, Φ_2 as in* (6.10)–(6.13). *Assume also that $k(x)$ has selfadjoint values. Then in the intervals between singularities,*

$$H(x) = \frac{1}{2}\begin{bmatrix} Q(x) & I_m \\ I_m & Q(x)^{-1}\end{bmatrix}, \tag{6.14}$$

where $Q(x)$ is a continuous $m \times m$ matrix-valued function whose values are nonnegative and invertible.

A similar result is obtained in [22, p. 507] under different assumptions, namely, $S \geq 0$ and S is factorable.

Lemma 6.4. *Define an involution U on $L^2_m(0,\ell)$ by*

$$(Uf)(x) = f(\ell - x), \qquad f \in L^2_m(0,\ell).$$

Let S have the form (6.10), *and assume also that $k(x) = k(x)^*$ on $(-\ell,\ell)$. Then $USU = S$.*

Proof of Lemma 6.4. Write S in the form (6.12) with $s(x)$ given by (6.11). The assumptions on $k(x)$ imply that $s(x)^* = s(x) = -s(-x)^*$ on $(-\ell,\ell)$. It is sufficient to show that $USUf = Sf$ whenever f is continuously differentiable on $[0,\ell]$ and $f(0) = f(\ell) = 0$. For such f, integration by parts yields

$$(Sf)(x) = \int_0^\ell s(x-t)f'(t)\,dt.$$

Therefore

$$(USUf)(x) = -\int_0^\ell s(-x+t)f'(t)\,dt = \int_0^\ell s(x-t)f'(t)\,dt = (Sf)(x),$$

as was to be shown. □

Proof of Theorem 6.3. By Theorem 6.2, $H(x)$ has the form (6.8). To deduce (6.14), it is sufficient to show that $h_2^*(\xi)h_1(\xi) = \frac{1}{2}\,I_m$, that is,

$$\frac{d}{d\xi}\left\langle S_\xi^{-1}\varphi_1, \varphi_2\right\rangle_\xi = \tfrac{1}{2}\,I_m \tag{6.15}$$

for ξ in any interval between singularities. For such ξ,

$$(S_\xi f)(x) = \frac{d}{dx}\int_0^\xi s(x-t)f(t)\,dt, \qquad f \in L^2_m(0,\xi). \tag{6.16}$$

Since $s(x)^* = s(x) = -s(-x)^*$ on $(-\ell,\ell)$,

$$\begin{aligned} S_\xi I_m = \frac{d}{dx}\int_0^\xi s(x-t)\,dt &= \frac{d}{dx}\int_{x-\xi}^x s(u)\,du \\ &= s(x) - s(x-\xi) = s(x) + s(\xi - x), \end{aligned}$$

$0 < x < \xi$. Therefore by (6.13),

$$S_\xi I_m = \varphi_1 + U_\xi \varphi_1,$$

where $(U_\xi f)(x) = f(\xi - x)$ for all $f \in L^2_m(0,\xi)$. By Lemma 6.4, $U_\xi S_\xi U_\xi = S_\xi$, and so

$$I_m = S_\xi^{-1}\varphi_1 + S_\xi^{-1}U_\xi\varphi_1 = I_m = S_\xi^{-1}\varphi_1 + U_\xi S_\xi^{-1}\varphi_1.$$

Writing $S_\xi^{-1}\varphi_1 = f$ and integrating, we get

$$2\left\langle S_\xi^{-1}\varphi_1, \varphi_2\right\rangle_\xi = 2\int_0^\xi f(x)\,dx = \int_0^\xi [f(x) + f(\xi - x)]\,dx = \xi I_m.$$

This yields (6.15) and hence the result. □

7. Examples

The examples in this section illustrate Theorems 4.1 and 5.3 in a number of ways. Each example features operators S, A, Φ_1, Φ_2 satisfying (3.1). We exhibit a corresponding canonical differential system (1.1) and spectral data. The systems which are constructed in the examples have Hamiltonians which are analytic except for poles. The calculations are straightforward but sometimes lengthy, and we only give the final results.

Let us first fix notation for the examples. In all cases, the underlying spaces are $\mathfrak{H} = L^2_m(0,\ell)$ and $\mathfrak{G} = \mathbf{C}^m$ for some positive integer m. The operators $A \in \mathfrak{L}(\mathfrak{H})$ and $\Phi_2 \in \mathfrak{L}(\mathfrak{G},\mathfrak{H})$ are the same in all of the examples:

$$(Af)(x) = i\int_0^x f(t)\,dt \quad \text{and} \quad (\Phi_2 c)(x) = \varphi_2(x)c, \quad \varphi_2(x) \equiv I_m. \tag{7.1}$$

The operators S and Φ_1 are special to each example. Since $\Phi_1 \in \mathfrak{L}(\mathfrak{G},\mathfrak{H})$, we always have

$$(\Phi_1 c)(x) = \varphi_1(x)c,$$

where $\varphi_1(x)$ is an $m \times m$ matrix-valued function. Let P_ξ be the projection of $\mathfrak{H}$ onto $\mathfrak{H}_\xi = L^2_m(0,\xi)$, $0 \le \xi \le \ell$, and let $S_\xi = P_\xi S P_\xi|_{\mathfrak{H}_\xi}$. Then according to Theorem 4.1,

the system (4.1) associated with the operator identity (3.1) is obtained with

$$B(\xi) = \Pi^* P_\xi S_\xi^{-1} P_\xi \Pi = \begin{bmatrix} \left\langle S_\xi^{-1}\varphi_1, \varphi_1 \right\rangle_\xi & \left\langle S_\xi^{-1}\varphi_2, \varphi_1 \right\rangle_\xi \\ \left\langle S_\xi^{-1}\varphi_1, \varphi_2 \right\rangle_\xi & \left\langle S_\xi^{-1}\varphi_2, \varphi_2 \right\rangle_\xi \end{bmatrix}. \tag{7.2}$$

Here $\langle \cdot, \cdot \rangle_\xi$ denotes an inner product in $L^2_m(0,\xi)$. In (7.2), we understand that φ_1 and φ_2 are first restricted to $(0,\xi)$, and we interpret $\left\langle S_\xi^{-1}\varphi_j, \varphi_k \right\rangle_\xi$ as an $m \times m$ matrix by viewing S_ξ^{-1} as acting on the columns of the matrix-valued functions φ_j and φ_k, $j,k = 1,2$.

In the examples we are mainly concerned with the scalar case, $m = 1$. In this case the underlying spaces are $\mathfrak{H} = L^2(0,\ell)$ and $\mathfrak{G} = \mathbf{C}$, and we use standard scalar notation.

Example 1. Assume the scalar case: $\mathfrak{H} = L^2(0,\ell)$ and $\mathfrak{G} = \mathbf{C}$. Define A and Φ_2 by (7.1), and let

$$(Sf)(x) = f(x) + \beta \int_0^\ell f(t)\, dt,$$
$$(\Phi_1 c)(x) = \varphi_1(x)c, \qquad \varphi_1(x) = \tfrac{1}{2} + \beta x,$$

where β is real and $\beta < 0$.

The operator identity (3.1) is satisfied, and $S = S_v$ and $\Phi_1 = \Phi_{1,v}$, where $v(z) = \frac{1}{2}\, i - \beta/z$ for $\operatorname{Im} z > 0$. The function $v(z)$ belongs to $\mathbf{N}_1$ and has the representation

$$v(z) = \int_{-\infty}^{\infty} \left[\frac{1}{t-z} - \frac{t}{1+t^2} \right] \frac{dt}{2\pi} - \frac{\beta}{z}.$$

Thus $v(z)$ has Kreĭn-Langer data $\boldsymbol{\tau} = \{\tau(t); \Delta_0; R(z)\}$, where $\tau(t) = t/(2\pi)$ on $\Delta_0 = (-\infty, \infty)$ and $R(z) = -\beta/z$. We obtain

$$(S_\xi f)(x) = f(x) + \beta \int_0^\xi f(t)\, dt,$$
$$(S_\xi^{-1} f)(x) = f(x) - \beta(1+\beta\xi)^{-1} \int_0^\xi f(t)\, dt,$$

on $L^2(0,\xi)$. The function (4.6) in Theorem 4.1 is given by

$$B(\xi) = \begin{bmatrix} (3\xi + 3\beta\xi^2 + \beta^2\xi^2)/12 & \frac{1}{2}\xi \\ \frac{1}{2}\xi & \xi/(1+\beta\xi) \end{bmatrix}.$$

The solution to the inverse problem in Theorem 5.3 is the system

$$\frac{dY}{dx} = izJH(x)Y, \qquad Y_1(0,z) = 0,$$

$$H(x) = B'(x) = \tfrac{1}{2}\begin{bmatrix} (1+\beta x)^2/2 & 1 \\ 1 & 2/(1+\beta x)^2 \end{bmatrix},$$

$0 \le x < \ell$. The Hamiltonian has a singularity at $x_1 = -1/\beta$ if $-1/\beta < \ell$. The transform (5.11) and Parseval relation (5.12) are given by

$$G(z) = \int_0^\ell \left(e^{-izx} - \frac{\beta}{1+\beta x}\,\frac{e^{-izx}-1}{-iz} \right) g(x)\,dx,$$

$$\int_0^\ell |g(t)|^2 dt = \frac{1}{2\pi}\int_{-\infty}^{\infty} |G(x)|^2\,dx + \beta\,|G(0)|^2.$$

Thus we obtain a perturbation of the Paley-Wiener example (5.13)–(5.14).

Example 2. Assume the scalar case as in Example 1. The operator identity (3.1) is satisfied with A and Φ_2 given by (7.1), and

$$(Sf)(x) = f(x) + \beta\int_0^\ell \left[e^{i\lambda(x-t)} + e^{-i\lambda(x-t)} \right] f(t)\,dt,$$

$$(\Phi_1 c)(x) = \varphi_1(x)c, \qquad \varphi_1(x) = \tfrac{1}{2} + \beta\,\frac{e^{i\lambda x} - e^{-i\lambda x}}{i\lambda},$$

where β and λ are real numbers such that $\beta < 0$ and $\lambda > 0$.

We have $S = S_v$ and $\Phi_1 = \Phi_{1,v}$, where

$$v(z) = \tfrac{1}{2}\,i - \frac{\beta}{z-\lambda} - \frac{\beta}{z+\lambda}, \qquad \operatorname{Im} z > 0.$$

This function belongs to $\mathbf{N}_2$ and has Kreĭn-Langer data $\boldsymbol{\tau} = \{\tau(t); \Delta_0; R(z)\}$, where $\tau(t) = t/(2\pi)$ on $\Delta_0 = (-\infty, \infty)$ and $R(z) = -\beta/(z-\lambda) - \beta/(z+\lambda)$. We find

$$(S_\xi f)(x) = f(x) + \beta\int_0^\xi \left[e^{i\lambda(x-t)} + e^{-i\lambda(x-t)} \right] f(t)\,dt,$$

$$(S_\xi^{-1} f)(x) = f(x) - K(x)T(\xi)^{-1}\int_0^\xi K(t)^* f(t)\,dt,$$

where $K(x) = \begin{bmatrix} e^{i\lambda x} & e^{-i\lambda x} \end{bmatrix}$, and

$$T(\xi) = \begin{bmatrix} \xi + \beta^{-1} & \gamma(\xi) \\ \overline{\gamma(\xi)} & \xi + \beta^{-1} \end{bmatrix}, \qquad \gamma(\xi) = \frac{e^{-2i\lambda\xi} - 1}{-2i\lambda}.$$

Using Theorem 5.3, we obtain a solution to the inverse problem given by

$$\frac{dY}{dx} = izJH(x)Y, \qquad Y_1(0,z) = 0,$$

$$H(x) = \begin{bmatrix} \overline{h_1(x)} \\ \overline{h_2(x)} \end{bmatrix}\begin{bmatrix} h_1(x) & h_2(x) \end{bmatrix},$$

where

$$h_1(x) = \tfrac{1}{2}\,\frac{x+\beta^{-1}+\lambda^{-1}\sin(\lambda x)}{x+\beta^{-1}-\lambda^{-1}\sin(\lambda x)},$$

$$h_2(x) = \frac{x+\beta^{-1}-\lambda^{-1}\sin(\lambda x)}{x+\beta^{-1}+\lambda^{-1}\sin(\lambda x)}.$$

Singularities occur when $\det T(x) = (x+\beta^{-1})^2 - \lambda^{-2}\sin^2(\lambda x) = 0$.

Example 3. Let $\mathfrak{H} = L^2_m(0,\ell)$ and $\mathfrak{G} = \mathbf{C}^m$. Define A and Φ_2 by (7.1), and let

$$(Sf)(x) = f(x) + \int_0^\ell \sum_{j=1}^r \beta_j e^{i\lambda_j(x-t)}\, f(t)\, dt,$$

$$(\Phi_1 c)(x) = \varphi_1(x)c, \qquad \varphi_1(x) = \tfrac{1}{2} + \sum_{j=1}^r \beta_j\,\frac{e^{i\lambda_j x}-1}{i\lambda_j},$$

where $\beta_1,\dots,\beta_r$ are invertible selfadjoint $m\times m$ matrices and $\lambda_1,\dots,\lambda_r$ are distinct real numbers; in the formula for $\varphi_1(x)$, if $\lambda_j = 0$ for some j, the expression $[e^{i\lambda_j x}-1]/(i\lambda_j)$ is interpreted as

$$\left.\frac{e^{i\lambda_j x}-1}{i\lambda_j}\right|_{\lambda_j=0} = x.$$

The operator identity (3.1) is satisfied, and $S = S_v$ and $\Phi_1 = \Phi_{1,v}$, where

$$v(z) = \tfrac{1}{2}\,i\,I_m - \sum_{j=1}^r \frac{\beta_j}{z-\lambda_j}, \qquad \operatorname{Im} z > 0.$$

This function has Kreĭn-Langer data $\boldsymbol{\tau} = \{\tau(t); \Delta_0; R(z)\}$, where $\tau(t) = t/(2\pi)$ on $\Delta_0 = (-\infty,\infty)$ and $R(z) = -\sum_{j=1}^r \beta_j/(z-\lambda_j)$. We find

$$(S_\xi f)(x) = f(x) + \int_0^\xi K(x)CK(t)^* f(t)\, dt, \tag{7.3}$$

$$(S_\xi^{-1} f)(x) = f(x) - \int_0^\xi K(x)\rho(\xi)^{-1}K(t)^* f(t)\, dt, \tag{7.4}$$

where $K(x) = \begin{bmatrix} e^{i\lambda_1 x} & \cdots & e^{i\lambda_r x}\end{bmatrix}$, $C = \operatorname{diag}\{\beta_1,\dots,\beta_r\}$, and

$$\rho(\xi) = C^{-1} + \int_0^\xi K(t)^* K(t)\, dt. \tag{7.5}$$

The inverse operator S_ξ^{-1} exists when $\det\rho(\xi) \neq 0$. Theorem 5.3 yields a solution to the inverse problem given by

$$\frac{dY}{dx} = izJH(x)Y, \qquad Y_1(0,z) = 0, \tag{7.6}$$

$$H(x) = \begin{bmatrix} h_1(x)^* \\ h_2(x)^* \end{bmatrix} \begin{bmatrix} h_1(x) & h_2(x) \end{bmatrix}, \tag{7.7}$$

where

$$h_1(x) = \varphi_1(x) - K(x)\rho(x)^{-1} \int_0^x K(t)^* \varphi_1(t)\, dt, \tag{7.8}$$

$$h_2(x) = 1 - K(x)\rho(x)^{-1} \int_0^x K(t)^*\, dt. \tag{7.9}$$

These functions are computable in closed form, but the expressions are not simple except in particular cases. We note that Examples 1 and 2 are special cases of this example with $m = 1$.

Example 4. In Examples 1–3, $v(z)$ is analytic for nonreal z. In this example, $v(z)$ has nonreal poles. Fix complex numbers $\lambda \neq \bar{\lambda}$ and $\beta \neq 0$. Again with $m = 1$, define A and Φ_2 by (7.1), and let

$$(Sf)(x) = f(x) + \int_0^\ell \left[\beta\, e^{i\lambda(x-t)} + \bar{\beta}\, e^{i\bar{\lambda}(x-t)}\right] f(t)\, dt,$$

$$(\Phi_1 c)(x) = \varphi_1(x)c, \qquad \varphi_1(x) = \tfrac{1}{2} + \beta\, \frac{e^{i\lambda x} - 1}{i\lambda} + \bar{\beta}\, \frac{e^{i\bar{\lambda} x} - 1}{i\bar{\lambda}}.$$

The identity (3.1) is satisfied, and $S = S_v$ and $\Phi_1 = \Phi_{1,v}$, where

$$v(z) = \tfrac{1}{2}\, i - \frac{\beta}{z - \lambda} - \frac{\bar{\beta}}{z - \bar{\lambda}}, \qquad \operatorname{Im} z > 0.$$

This function belongs to $\mathbf{N}_1$ and has Kreĭn-Langer data $\boldsymbol{\tau} = \{\tau(t); \Delta_0; R(z)\}$, where $\tau(t) = t/(2\pi)$ on $\Delta_0 = (-\infty, \infty)$ and $R(z) = -\beta/(z - \lambda) - \bar{\beta}/(z - \bar{\lambda})$. The inverse problem is solved by Theorem 5.3 using the identical formulas (7.3)–(7.9) from Example 3, but now taken with

$$K(x) = \begin{bmatrix} e^{i\lambda x} & e^{i\bar{\lambda} x} \end{bmatrix}, \quad C = \begin{bmatrix} 0 & \beta \\ \bar{\beta} & 0 \end{bmatrix},$$

$$\varphi_1(x) = \tfrac{1}{2} + \beta\, \frac{e^{i\lambda x} - 1}{i\lambda} + \bar{\beta}\, \frac{e^{i\bar{\lambda} x} - 1}{i\bar{\lambda}}.$$

Example 5. Let $\mathfrak{H} = L^2(0, \ell)$ and $\mathfrak{G} = \mathbf{C}$. Define A and Φ_2 by (7.1), and let

$$(Sf)(x) = f(x) + ia \int_0^x f(t)\, dt - ia \int_x^\ell f(t)\, dt,$$

$$(\Phi_1 c)(x) = \varphi_1(x)c, \qquad \varphi_1(x) = \tfrac{1}{2} + iax,$$

where $a \neq 0$ is a real number. The identity (3.1) is satisfied.

A priori we do not know a generalized Nevanlinna function $v(z)$ such that $S = S_v$ and $\Phi_1 = \Phi_{1,v}$, but we shall determine such a function later. Nevertheless,

we may apply Theorem 4.1 since S is a compact perturbation of the identity operator and hence $\varkappa_S < \infty$.

We find

$$(S_\xi f)(x) = f(x) + ia\int_0^x f(t)\,dt - ia\int_x^\xi f(t)\,dt\,,$$

$$(S_\xi^{-1} f)(x) = f(x) - 2ia\int_0^x e^{2ia(t-x)} f(t)\,dt + \frac{2ia}{e^{2ia\xi}+1}\int_0^\xi e^{2ia(t-x)} f(t)\,dt,$$

for all ξ such that $e^{2ia\xi}+1 \neq 0$, that is, for all points $\xi = (n-\frac{1}{2})\pi/a$, $n = 1, 2, \ldots$, that lie in $[0, \ell)$. The system constructed in Theorem 4.1 for the operator identity (3.1) is

$$\frac{dY}{dx} = izJH(x)Y, \qquad Y_1(0,z) = 0,$$

$$H(x) = \begin{bmatrix} \overline{h_1(x)} \\ \overline{h_2(x)} \end{bmatrix} \begin{bmatrix} h_1(x) & h_2(x) \end{bmatrix},$$

where

$$h_1(x) = \tfrac{1}{2}\,e^{iax} + \tfrac{1}{2}\,i\,\frac{ax}{\cos(ax)}, \qquad h_2(x) = \frac{1}{\cos(ax)}.$$

Next we determine $v(z)$ belonging to some class $\mathbf{N}_\varkappa$ such that $S = S_v$ and $\Phi_1 = \Phi_{1,v}$. This is an interpolation problem of a type first solved by A.L. Sakhnovich [17]. Alternatively, we may use [12, Theorem 5.3]. Thus we may choose

$$v(z) = ia(z)/c(z), \tag{7.10}$$

where

$$\begin{bmatrix} a(z) & b(z) \\ c(z) & d(z) \end{bmatrix} = I_2 - iz\Pi^*(I - zA^*)^{-1}S^{-1}\Pi J, \tag{7.11}$$

$\Pi = \begin{bmatrix} \Phi_1 & \Phi_2 \end{bmatrix}$. It can be shown that all of the conditions required in [12, Theorem 5.3] are met. We obtain $S = S_v$ and $\Phi_1 = \Phi_{1,v}$, where

$$v(z) = -\left(\frac{1}{2} + \frac{a}{z}\right)\cot\frac{(z+2a)\ell}{2}\,.$$

We remark that this provides a nontrivial example of the interpolation result in [12, Theorem 5.3].

Example 6. Let $\mathfrak{H} = L^2(0,\ell)$ and $\mathfrak{G} = \mathbf{C}$, and let α, β be real numbers, $\alpha\beta \neq 0$. The operator identity (3.1) is satisfied with A and Φ_2 defined by (7.1), and

$$(Sf)(x) = 2f(x) + \beta\int_0^\ell e^{-\alpha|x-t|} f(t)\,dt\,,$$

$$(\Phi_1 c)(x) = \varphi_1(x)c, \qquad \varphi_1(x) = 1 + \beta\,\frac{e^{-\alpha x}-1}{-\alpha}.$$

The hypotheses of Theorem 4.1 are satisfied. By Theorem 6.3, the Hamiltonian $H(\xi) = B'(\xi)$ constructed from (4.6) has the form

$$H(\xi) = \frac{1}{2}\begin{bmatrix} Q(\xi) & 1 \\ 1 & Q(\xi)^{-1} \end{bmatrix},$$

where $Q(\xi)$ is a positive continuous function in the intervals between singularities. The location of singularities and form of $Q(\xi)$ depend on cases.

Case 1: $\alpha^2 + \alpha\beta > 0$**.** Put $\omega^2 = \alpha^2 + \alpha\beta$, $\omega > 0$. We obtain

$$Q(\xi) = \left[\frac{\omega}{\alpha}\,\frac{(\omega+\alpha)e^{\frac{1}{2}\omega\xi} - (\omega-\alpha)e^{-\frac{1}{2}\omega\xi}}{(\omega+\alpha)e^{\frac{1}{2}\omega\xi} + (\omega-\alpha)e^{-\frac{1}{2}\omega\xi}}\right]^2. \tag{7.12}$$

There are no singularities if $\alpha > 0$. If $\alpha < 0$, there is one singularity at

$$\xi_1 = \frac{1}{\omega}\,\log\left|\frac{\omega-\alpha}{\omega+\alpha}\right|$$

if this point is less than ℓ. We have $S = S_v$ and $\Phi_1 = \Phi_{1,v}$ with

$$v(z) = i - \frac{\beta}{z+i\alpha}, \qquad \operatorname{Im} z > 0.$$

For $\alpha > 0$, $v(z)$ belongs to $\mathbf{N}_0$, that is, it is a classical Nevanlinna function; its Kreĭn-Langer (Nevanlinna) representation is

$$v(z) = \int_{-\infty}^{\infty}\left[\frac{1}{t-z} - \frac{t}{1+t^2}\right]d\tau(t),$$

where

$$d\tau(t) = \left[\frac{1}{\pi} + \frac{\beta}{\pi}\,\frac{\alpha}{t^2+\alpha^2}\right]dt. \tag{7.13}$$

For $\alpha < 0$, $v(z)$ belongs to $\mathbf{N}_1$ and has Kreĭn-Langer representation

$$v(z) = \int_{-\infty}^{\infty}\left[\frac{1}{t-z} - \frac{t}{1+t^2}\right]d\tau(t) - \frac{\beta}{z+i\alpha} - \frac{\beta}{z-i\alpha},$$

where $d\tau(t)$ has the same form (7.13).

Case 2: $\alpha^2 + \alpha\beta = 0$**.** In this case, $\beta = -\alpha$. We find that

$$Q(\xi) = (1 + \tfrac{1}{2}\alpha\xi)^{-2}.$$

There are no singularities if $\alpha > 0$. If $\alpha < 0$, there is one singularity at

$$\xi_1 = -\frac{2}{\alpha}$$

if this point is less than ℓ. We have $S = S_v$ and $\Phi_1 = \Phi_{1,v}$ for the same $v(z)$ as in Case 1 taken with $\beta = -\alpha$.

Case 3: $\alpha^2 + \alpha\beta < 0$**.** In this case,

$$Q(\xi) = \frac{\omega^2}{\alpha^2} \cot^2\left(\tfrac{1}{2}\,\omega\xi + \rho\right),$$

where $\omega^2 = |\alpha^2 + \alpha\beta|$, $\omega > 0$, and $\tan\rho = \omega/\alpha$. There are singularities at all of the points

$$\xi_k = \frac{1}{\omega}\,(k\pi - 2\rho), \qquad k = 0, \pm 1, \pm 2, \ldots,$$

which lie in $(0, \ell)$. An explicit choice of generalized Nevanlinna function $v(z)$ such that $S = S_v$ and $\Phi_1 = \Phi_{1,v}$ can be constructed as in Example 5 using the formulas (7.10) and (7.11). We obtain

$$v(z) = i - \frac{\beta}{z + i\alpha} + \frac{i}{c(z)}\left[1 + \frac{\beta\alpha^{-1}z}{z + i\alpha}\,c(-i\alpha)\right],$$

where

$$c(z) = -iz\int_0^\ell e^{-izt}\left[-\frac{\alpha^2}{2\omega^2} + \sigma^{-1}\cos\left(\omega(t - \tfrac{1}{2}\,\ell)\right)\right]\,dt$$

and $\sigma = 2\omega^2[\alpha\cos(\frac{1}{2}\omega\ell) - \omega\sin(\frac{1}{2}\omega\ell)]/[\alpha(\alpha^2 + \omega^2)]$. The integral in the formula for $c(z)$ is easily computed in terms of elementary functions.

Example 7. This example uses a generalized Nevanlinna function $v(z)$ whose Kreĭn-Langer representation (2.2) involves a nontrivial term with $\Delta_1 = (-1, 1)$. Let $\mathfrak{H} = L^2(0, \ell)$ and $\mathfrak{G} = \mathbf{C}$, and let α, β be real numbers with $\alpha \neq 0$. The operator identity (3.1) is satisfied with A and Φ_2 defined by (7.1), and

$$\begin{aligned} Sf &= f(x) + \int_0^\ell \left[\alpha|x - t| + \beta\right] f(t)\,dt \\ &= \frac{d}{dx}\int_0^\ell s(x - t)f(t)\,dt, \\ \Phi_1 g &= \varphi_1(x)g, \end{aligned}$$

where

$$s(x) = \begin{cases} \frac{1}{2} + \beta x + \frac{1}{2}\,\alpha x^2, & 0 < x < \ell, \\ -\frac{1}{2} + \beta x - \frac{1}{2}\,\alpha x^2, & -\ell < x < 0, \end{cases}$$

and $\varphi_1(x) = s(x)$ for $0 < x < \ell$. Theorems 4.1 and 6.3 produce a system with Hamiltonian $H(\xi) = B'(\xi)$ of the form

$$H(\xi) = \frac{1}{2}\begin{bmatrix} Q(\xi) & 1 \\ 1 & Q(\xi)^{-1} \end{bmatrix},$$

where $Q(\xi)$ is positive and continuous in the intervals between singularities.

Case 1: $\alpha < 0$**.** For sufficiently large ℓ there is one singularity, and

$$Q(\xi) = \frac{1}{2}\left[1 + \frac{\alpha\xi + 2\beta}{\sqrt{2|\alpha|}}\,\frac{e^{\xi\sqrt{|\alpha|/2}} - e^{-\xi\sqrt{|\alpha|/2}}}{e^{\xi\sqrt{|\alpha|/2}} + e^{-\xi\sqrt{|\alpha|/2}}}\right]^2$$
$$= \frac{1}{2}\left[1 + \frac{\alpha\xi + 2\beta}{\sqrt{2|\alpha|}}\,\tanh\left(\xi\sqrt{|\alpha|/2}\right)\right]^2.$$

In this case $S = S_v$ and $\Phi_1 = \Phi_{1,v}$, where $v(z) \in \mathbf{N}_1$ is given by

$$v(z) = \tfrac{1}{2}\,i - \frac{i\alpha}{z^2} - \frac{\beta}{z}, \qquad y > 0.$$

This function has the Kreĭn-Langer representation

$$v(z) = \int_{-1}^{1}\left[\frac{1}{t-z} - S_1(t,z)\right]d\tau(t)$$
$$+ \left(\int_{-\infty}^{-1} + \int_{1}^{\infty}\right)\left[\frac{1}{t-z} - \frac{t}{1+t^2}\right]d\tau(t) - \frac{C}{z},$$

where $C = (\pi\beta - 2|\alpha| - 1)/\pi$ and

$$\tau(t) = \frac{1}{2\pi}\,t - \frac{|\alpha|}{\pi}\,\frac{1}{t}, \qquad t \neq 0.$$

In the definition of $S_1(t,z)$, we take $\alpha_1 = 0$ and $\rho_1 = 1$:

$$\frac{1}{t-z} - S_1(t,z) = \frac{1}{t-z}\,\frac{t^2}{z^2}.$$

Case 2: $\alpha > 0$**.** In this case,

$$Q(\xi) = \frac{1}{2}\left[1 + \frac{\alpha\xi + 2\beta}{\sqrt{2\alpha}}\,\tan\left(\xi\sqrt{\alpha/2}\right)\right]^2.$$

For large ℓ there can be an arbitrarily large number of singularities. These occur at the points in $(0,\ell)$ where $Q(\xi)$ is zero or undefined. To find $v(z)$ in some class $\mathbf{N}_\varkappa$ such that $S = S_v$ and $\Phi_1 = \Phi_{1,v}$, we again use the formulas (7.10) and (7.11) as in Example 5. Setting $\alpha = \frac{1}{2}\omega^2$, we get

$$v(z) = \tfrac{1}{2}\,i - \frac{\beta}{z} - \frac{i\alpha}{z^2}$$
$$- \left(\frac{A}{z} + \frac{B}{z^2}\right)e^{\frac{1}{2}iz\ell}\frac{z^2 - \omega^2}{z\sin(\frac{1}{2}z\ell)\cos(\frac{1}{2}\omega\ell) - \omega\cos(\frac{1}{2}z\ell)\sin(\frac{1}{2}\omega\ell)},$$

where

$$A = \tfrac{1}{2}\cos(\tfrac{1}{2}\omega\ell) + \left(\tfrac{1}{2}\omega\ell + \frac{2\beta}{\omega}\right)\sin(\tfrac{1}{2}\omega\ell),$$
$$B = -\frac{i\alpha}{\omega}\sin(\tfrac{1}{2}\omega\ell).$$

Example 8. We return to Example 1 and show a connection with Bessel's equation. Write the system constructed in Example 1 as

$$\frac{dY}{dx} = izJH(x)Y, \qquad Y_1(0,z) = 0,$$
$$H(x) = \tfrac{1}{2}\begin{bmatrix} Q(x) & 1 \\ 1 & Q(x)^{-1} \end{bmatrix}, \tag{7.14}$$
$$Q(x) = \tfrac{1}{2}(1+\beta x)^2.$$

Here $\beta < 0$. In the regular case, the form of Hamiltonian in (7.14) occurs in the theory of dual systems [22]. We apply similar constructions and set

$$U(x,z) = \begin{bmatrix} U_1(x,z) \\ U_2(x,z) \end{bmatrix} = Y(2x,z)e^{-ixz}.$$

This leads to the system

$$\frac{dU}{dx} = izJ\begin{bmatrix} P(x) & 0 \\ 0 & P(x)^{-1} \end{bmatrix} U, \qquad U_1(0,z) = 0, \tag{7.15}$$
$$P(x) = Q(2x), \qquad 0 \le x < \tfrac{1}{2}\,\ell.$$

We refer to [5, 22] for the notion of dual equations. In our case, the dual equations derived from (7.15) have the form

$$U_1'' + \frac{P'(x)}{P(x)}\,U_1' + z^2 U_1 = 0,$$
$$U_2'' - \frac{P'(x)}{P(x)}\,U_2' + z^2 U_2 = 0,$$

with appropriate boundary conditions that play no role here. Writing

$$x_1 = -\frac{1}{\beta},$$

we obtain $Q(x) = \frac{1}{2}\,\beta^2(x-x_1)^2$ and $P(x) = 2\beta^2(x - \frac{1}{2}\,x_1)^2$. The dual equations become

$$U_1'' + \frac{2}{x - \frac{1}{2}\,x_1}\,U_1' + z^2 U_1 = 0,$$
$$U_2'' - \frac{2}{x - \frac{1}{2}\,x_1}\,U_2' + z^2 U_2 = 0.$$

On setting $U_1 = (x - \frac{1}{2}\,x_1)^{-1} y_1$ and $U_2 = (x - \frac{1}{2}\,x_1)y_2$, we obtain

$$y_1'' + z^2 y_1 = 0,$$
$$y_2'' + \left(z^2 - \frac{2}{(x - \frac{1}{2}\,x_1)^2}\right) y_2 = 0,$$

which are forms of Bessel's equation for the orders $\nu = \frac{1}{2}$ and $\nu = \frac{3}{2}$ (see Watson [23, §4.3, p. 95]).

8. Open problems

A number of open problems are suggested by our results, among them:

(1) Investigate the direct problem.

(2) Hamiltonians of the form

$$H(x) = \tfrac{1}{2}\begin{bmatrix} Q(x) & 1 \\ 1 & Q(x)^{-1} \end{bmatrix}$$

arise in Theorem 6.3. In the regular case such Hamiltonians give rise to a pair of dual equations [5, 22]. An example of a dual pair for a system with singularities is given in Example 8. The theory of dual equations should be generalized to systems with singularities. This requires introducing a new notion of spectral data for selfadjoint second order equations, and relating such a notion to spectral data for the associated canonical differential system.

(3) According to Theorem 6.1, the hypotheses of Theorem 4.1 hold when an analyticity condition is met. The examples in Section 7 show that the hypotheses of Theorem 4.1 also hold in situations which are not covered by Theorem 6.1. It is likely that there are general criteria that cover such examples.

Errata. *In* [15]*, Section* 3*, citations to theorems and definitions are shifted by one beginning with Theorem* 3.2*. For example, on p.* 130*, line* 1*, replace "conclusions of Theorem* 3.2*" by "conclusions of Theorem* 3.3*".*

References

[1] D.Z. Arov and H. Dym, *J-inner matrix functions, interpolation and inverse problems for canonical systems.* I. *Foundations*, Integral Equations Operator Theory, **29** no. 4 (1997), 373–454; II. *The inverse monodromy problem*, ibid. **36** no. 1 (2000), 11–70; III. *More on the inverse monodromy problem*, ibid. **36** no. 2 (2000), 127–181; IV. *Direct and inverse bitangential input scattering problems*, ibid. **43** no. 1 (2002), 1–67; V. *The inverse input scattering problem for Wiener class and rational $p \times q$ input scattering matrices*, ibid. **43** no. 1 (2002), 68–129.

[2] _____, *The bitangential inverse input impedance problem for canonical systems* I. *Weyl-Titchmarsh classification, existence and uniqueness*, Integral Equations Operator Theory **47** no. 1 (2003), 3–49; II. *Formulas and examples*, ibid. **51** no. 2 (2005), 155–213.

[3] K. Daho and H. Langer, *Matrix functions of the class N_κ*, Math. Nachr. **120** (1985), 275–294.

[4] L. de Branges, *Hilbert spaces of entire functions*, Prentice-Hall Inc., Englewood Cliffs, N.J., 1968.

[5] H. Dym and L.A. Sakhnovich, *On dual canonical systems and dual matrix string equations*, Operator theory, system theory and related topics (Beer-Sheva/Rehovot, 1997), Oper. Theory Adv. Appl., vol. 123, Birkhäuser, Basel, 2001, pp. 207–228.

[6] I.C. Gohberg and M.G. Kreĭn, *Theory and applications of Volterra operators in Hilbert space*, American Mathematical Society, Providence, R.I., 1970.

[7] M. Kaltenbäck and H. Woracek, *Pontryagin spaces of entire functions.* I, Integral Equations Operator Theory, **33** no. 1 (1999), 34–97; II, ibid. **33** no. 3 (1999), 305–380; III, Acta Sci. Math. (Szeged), **69** no. 1–2 (2003), 241–310; IV, ibid. **72** no. 3–4 (2006), 709–835.

[8] T. Kato, *Perturbation theory for linear operators*, second ed., Springer-Verlag, Berlin, 1976, Grundlehren der Mathematischen Wissenschaften, Band 132.

[9] M.G. Kreĭn and H. Langer, *On some extension problems which are closely connected with the theory of Hermitian operators in a space* Π_κ. III. *Indefinite analogues of the Hamburger and Stieltjes moment problems*, Part I, Beiträge Anal. **14** (1979), 25–40; Part II, ibid. **15**, (1981), 27–45.

[10] ———, *Über einige Fortsetzungsprobleme, die eng mit der Theorie hermitescher Operatoren im Raume* Π_κ *zusammenhängen.* I. *Einige Funktionenklassen und ihre Darstellungen*, Math. Nachr. **77** (1997), 187–236.

[11] H. Langer and H. Winkler, *Direct and inverse spectral problems for generalized strings*, Integral Equations Operator Theory **30** no. 4 (1998), 409–431, Dedicated to the memory of Mark Grigorievich Krein (1907–1989).

[12] J. Rovnyak and L.A. Sakhnovich, *On indefinite cases of operator identities which arise in interpolation theory*, The extended field of operator theory (Newcastle, 2004), Oper. Theory Adv. Appl., vol. 171, Birkhäuser, Basel, 2006, pp. 281–322.

[13] ———, *Some indefinite cases of spectral problems for canonical systems of difference equations*, Linear Algebra Appl. **343/344** (2002), 267–289.

[14] ———, *On the Kreĭn-Langer integral representation of generalized Nevanlinna functions*, Electron. J. Linear Algebra **11** (2004), 1–15 (electronic).

[15] ———, *Spectral problems for some indefinite cases of canonical differential equations*, J. Operator Theory **51** (2004), 115–139.

[16] A.L. Sakhnovich, *Spectral functions of a second-order canonical system*, Mat. Sb. **181** no. 11 (1990), 1510–1524, Engl. transl., USSR-Sb. **71** no. 2 (1992), 355–369.

[17] ———, *Modification of V. P. Potapov's scheme in the indefinite case*, Matrix and operator valued functions, Oper. Theory Adv. Appl., vol. 72, Birkhäuser, Basel, 1994, pp. 185–201.

[18] L.A. Sakhnovich, *Problems of factorization and operator identities*, Uspekhi Mat. Nauk **41** no. 1 (1986), (247), 4–55, Engl. transl., Russian Math. Surveys 41:1 (1986), 1–64.

[19] ———, *Integral equations with difference kernels on finite intervals*, Oper. Theory Adv. Appl., vol. 84, Birkhäuser Verlag, Basel, 1996.

[20] ———, *Interpolation theory and its applications*, Kluwer, Dordrecht, 1997.

[21] ———, *Spectral theory of canonical differential systems. Method of operator identities*, Oper. Theory Adv. Appl., vol. 107, Birkhäuser Verlag, Basel, 1999.

[22] ———, *On reducing the canonical system to two dual differential systems*, J. Math. Anal. Appl. **255** no. 2 (2001), 499–509.

[23] G.N. Watson, *A Treatise on the Theory of Bessel Functions*, second ed., Cambridge University Press, Cambridge, England, 1944.

James Rovnyak
University of Virginia
Department of Mathematics
P. O. Box 400137
Charlottesville, VA 22904–4137
USA
e-mail: rovnyak@Virginia.EDU

Lev A. Sakhnovich
735 Crawford Avenue
Brooklyn, NY 11223
USA
e-mail: Lev.Sakhnovich@verizon.net

Operator Theory:
Advances and Applications, Vol. 179, 289–308

On Triangular Factorization of Positive Operators

Lev A. Sakhnovich

Abstract. We investigate the problem of the triangular factorization of positive operators in a Hilbert space. We prove that broad classes of operators can be factorized.

Mathematics Subject Classification (2000). Primary 47A68; Secondary 47A05, 47A66.

Keywords. Triangular operators, operators with difference kernels, operator identity, homogeneous kernels.

1. Introduction

In the Hilbert space $L_m^2(a,b)$ we define the orthogonal projectors $P_\xi f = f(x)$, $a \le x < \xi$ and $P_\xi f = 0$, $\xi < x \le b$, where $f(x) \in L_m^2(a,b)$.

Definition 1.1. A bounded operator S_- on $L_m^2(a,b)$ is called lower triangular if for every ξ the relations

$$S_- Q_\xi = Q_\xi S_- Q_\xi, \tag{1.1}$$

are true, where $Q_\xi = I - P_\xi$.

Definition 1.2. A bounded operator S_+ on $L_m^2(a,b)$ is called upper triangular if for every ξ the relations

$$S_+ P_\xi = P_\xi S_+ P_\xi \tag{1.2}$$

are true.

Definition 1.3. A bounded, positive and invertible operator S on $L_m^2(a,b)$ is said to admit the right triangular factorization if it can be represented in the form

$$S = S_+ S_+^*, \tag{1.3}$$

where S_+ and S_+^{-1} are upper triangular, bounded operators.

Definition 1.4. A bounded, positive and invertible operator S on $L^2_m(a,b)$ is said to admit the left triangular factorization if it can be represented in the form

$$S = S_- S_-^*, \tag{1.4}$$

where S_- and S_-^{-1} are lower triangular, bounded operators.

I. Gohberg and M.G. Krein [5] studied the problem of factorization under the assumption

$$S - I \in \gamma_\infty, \tag{1.5}$$

where γ_∞ is the set of compact operators. The operators S_- and S_+ were assumed to have the form $S_+ = I + X_+$, $S_- = I + X_-$; $X_+, X_- \in \gamma_\infty$. The factorization method plays an important role in a number of analysis problems (for instance integral equations [17], spectral theory [18], nonlinear integrable equations). Giving up condition (1.5) and considering more general triangular operators would essentially widen the scope of the factorization method. D. Larson proved in his famous work [9] the existence of positive non-factorable operators. In Section 2 we formulate the necessary and sufficient conditions under which the positive operator S admits a triangular factorization. The factorizing operator $V = S_-^{-1}$ is constructed in an explicit form. In Section 3 we consider the class of positive operators S which satisfy the operator identity

$$AS - SA^* = \Pi J \Pi^*. \tag{1.6}$$

For operators of this class, the factorization conditions have a simpler form. The general results of Sections 2 and 3 are applied to operators with difference kernels (Section 4),

$$Sf = \frac{d}{dx} \int_0^a f(t) s(x-t) dt, \tag{1.7}$$

and to operators with sum-difference kernels (Section 5),

$$Sf = \frac{d^2}{dx^2} \int_0^b [s_1(x-t) + s_2(x+t)] f(t) dt, \tag{1.8}$$

where $f(t) \in L^2(0,b)$. In particular, we prove that the Dixon operator [4], [8], [19]

$$Sf = f(x) - \frac{\lambda}{\pi} \int_0^1 \frac{f(t)}{x+t} dt = g(x), \tag{1.9}$$

where $f(x) \in L^2(0,1)$ and $\lambda < 1$, admits a left triangular factorization. We note that the operators of the forms (1.7) and (1.8) play an important role in theoretical and applied problems (inverse problems, stationary processes, prediction theory). In Section 6 we investigate the case when

$$Af = i \int_0^x f(t) dt, \quad \operatorname{rank}(AS - SA^*) = 1. \tag{1.10}$$

In this case the factorizing operator S_- has the special form

$$S_- f = \frac{d}{dx} \int_0^x f(t) \phi(x-t) dt. \tag{1.11}$$

In Section 7 we consider a class of operators of the form

$$SF = F(x) - \int_0^1 F(y)k(\frac{y}{x})\frac{1}{x}dy = G(x), \tag{1.12}$$

where $F(x) \in L^2(0,1)$. The Dixon operator belongs to this class.

Remark 1.1. In our paper we consider triangular operators in the space $L^2_m(a,b)$ with the special set of projectors P_ξ. A general theory of triangular operators is constructed in the works [2], [3], [7], [9]–[13].

2. Triangular factorization

Let S be a linear, bounded and invertible operator S on $L^2_m(a,b)$. We introduce the notation

$$S_\xi = P_\xi S P_\xi, \quad (f,g)_\xi = \int_a^\xi g^*(x)f(x)dx, \tag{2.1}$$

where $f(x), g(x) \in L^2_m(a,b)$.

Theorem 2.1. *Let the bounded and invertible operator S on $L^2_m(a,b)$ be positive. For the operator S to admit the left triangular factorization it is necessary and sufficient that the following assertions are true.*

1. *There exists an $m\times m$ matrix function $F_0(x)$ such that*

$$Tr\int_a^b F_0^*(x)F_0(x)dx < \infty, \tag{2.2}$$

that the $m \times m$ matrix function

$$M(\xi) = (F_0(x), S_\xi^{-1}F_0(x))_\xi \tag{2.3}$$

is absolutely continuous, and almost everywhere

$$\det M'(\xi) \neq 0. \tag{2.4}$$

2. *The vector functions*

$$\int_a^x v^*(x,t)f(t)dt \tag{2.5}$$

are absolutely continuous. Here $f(x) \in L^2_m(a,b)$ and

$$v(\xi,t) = S_\xi^{-1}P_\xi F_0(x), \tag{2.6}$$

(In (2.3) the operator S_ξ^{-1} transforms the matrix column of the original into the corresponding column of the image.)

3. *The operator*

$$Vf = [R^*(x)]^{-1}\frac{d}{dx}\int_a^x v^*(x,t)f(t)dt \tag{2.7}$$

is bounded, invertible and lower triangular with its inverse V^{-1}. Here $R(x)$ is an $m\times m$ matrix function such that

$$R^*(x)R(x) = M'(x). \tag{2.8}$$

Proof. Necessity. We suppose that the operator S admits the left triangular factorization (1.4). Let $F_0(x) \in L^2_m(a,b)$ be a fixed $m\times m$ matrix function satisfying relation (2.2). We introduce the $m\times m$ matrix function

$$R(x) = VF_0(x), \tag{2.9}$$

where $V = S_-^{-1}$. We can choose $F_0(x)$ in such a way that almost everywhere the inequality

$$\det R(x) \neq 0 \tag{2.10}$$

is true. From relations (1.4), (2.3) and (2.9) we have

$$M(\xi) = \int_a^\xi R^*(x)R(x)dx. \tag{2.11}$$

Hence the function $M(\xi)$ is absolutely continuous and

$$M'(x) = R^*(x)R(x). \tag{2.12}$$

Now we use the equality

$$(f, S_\xi^{-1}F_0)_\xi = (Vf, VF_0)_\xi. \tag{2.13}$$

Relations (2.9) and (2.13) imply that

$$\frac{d}{dx}\int_a^x v^*(x,t)f(t)dt = R^*(x)(Vf). \tag{2.14}$$

The necessity is proved.

Sufficiency. Let the conditions 1–3 of Theorem 2.1 be fulfilled. It follows from (2.6)–(2.8) that

$$VF_0 = R(x). \tag{2.15}$$

From relations (2.6), (2.7) and (2.15) we deduce that $(Vf, VF_0)_\xi = (f, S_\xi^{-1}P_\xi F_0)_\xi$, i.e.,

$$V^*P_\xi VP_\xi F_0 = S_\xi^{-1}P_\xi F_0. \tag{2.16}$$

We define $v(\xi,t)$ in the domain $\xi \le t \le b$ by the equality $v(\xi,t) = 0$. It follows from the triangular structure of the operators V and V^{-1} that

$$P_\xi V^{-1}P_\xi VP_\xi = P_\xi. \tag{2.17}$$

Hence in view of (2.6) and (2.16) we have

$$P_\xi V^{-1}[V^*]^{-1}v(\xi,t) = P_\xi F_0. \tag{2.18}$$

It is easy to see that $P_\xi Sv(\xi,t) = P_\xi F_0$. Thus according to relations (2.17) and (2.18), the equality

$$(V^{-1}[V^*]^{-1}v(\xi,t), v(\mu,t)) = (Sv(\xi,t), v(\mu,t)) \tag{2.19}$$

is true. If there exists such a vector function $f_0(x) \in L_m^2(a,b)$ that $(f_0, v(\xi,t)) = 0$, then due to (2.7) the relation

$$Vf_0 = 0 \tag{2.20}$$

is valid. The operator V is invertible. Hence from (2.20) we deduce that $f_0 = 0$. This means that $v(\xi,t)$ is a complete system in $L_m^2(a,b)$. Using this fact and relation (2.19) we obtain the desired equality

$$S = V^{-1}[V^*]^{-1}. \tag{2.21}$$

The theorem is proved. □

Corollary 2.1. *If the conditions of Theorem* 2.1 *are fulfilled, then the corresponding operator* S^{-1} *can be represented in the form*

$$S^{-1} = V^*V. \tag{2.22}$$

We introduce the notation

$$C_\xi = Q_\xi S Q_\xi, \quad [f,g]_\xi = \int_\xi^b g^*(x) f(x) dx. \tag{2.23}$$

In the same way as Theorem 2.1 we deduce the following result.

Theorem 2.2. *Let the bounded and invertible operator* S *on* $L_m^2(a,b)$ *be positive. For the operator* S *to admit the right triangular factorization it is necessary and sufficient that the following assertions are true.*

1. *There exists an* $m\times m$ *matrix function* $F_0(x)$ *such that*

$$Tr \int_a^b F_0^*(x) F_0(x) dx < \infty, \tag{2.24}$$

that the $m\times m$ *matrix function*

$$N(\xi) = [F_0(x), C_\xi^{-1} F_0(x)]_\xi \tag{2.25}$$

is absolutely continuous, and almost everywhere

$$\det N'(\xi) \neq 0. \tag{2.26}$$

2. *The vector functions*

$$\int_x^b u^*(x,t) f(t) dt \tag{2.27}$$

are absolutely continuous. Here $f(x) \in L^2(a,b)$ *and*

$$u(\xi,t) = C_\xi^{-1} Q_\xi F_0. \tag{2.28}$$

3. *The operator*

$$Uf = -[Q^*(x)]^{-1} \frac{d}{dx} \int_x^b u^*(x,t) f(t) dt \tag{2.29}$$

is bounded, upper triangular and invertible together with its inverse U^{-1}. *Here*

$$Q^*(x) Q(x) = -N'(x). \tag{2.30}$$

Corollary 2.2. *If the conditions of Theorem* 2.2 *are fulfilled, then the corresponding operator* S^{-1} *can be represented in the form*

$$S^{-1} = U^*U. \tag{2.31}$$

Remark 2.1. Formulas (2.6), (2.7) and (2.28), (2.29) give the right and left factorization of the operator $T = S^{-1}$. It can be useful for solving operator equations of the form $Sf = g$. Using the notation

$$T = S^{-1}, \quad T_\xi = Q_\xi T Q_\xi, \quad w(\xi, t) = T_\xi^{-1} Q_\xi T F_0, \tag{2.32}$$

we introduce the operator

$$Wf = -[R^*(x)]^{-1} \frac{d}{dx} \int_x^b w^*(x,t) f(t) dt. \tag{2.33}$$

The connection between the operators V and W is given by the following assertion.

Proposition 2.1. *Let the operator* V *defined by formula* (2.7) *be bounded. Then the operator* W *defined by formula* (2.33) *is also bounded and*

$$WT = V. \tag{2.34}$$

Proof. It can be proved by linear algebra methods that (see [18], p. 41)

$$TQ_\xi T_\xi^{-1} Q_\xi T = T - S_\xi^{-1} P_\xi. \tag{2.35}$$

From relations (2.6), (2.32) and (2.35) we have

$$Tw(\xi, t) = TF_0 - v(\xi, t). \tag{2.36}$$

Hence the equality

$$[Tf, w(\xi, t)]_\xi = (Tf, F_0) - (f, v(\xi, t))_\xi \tag{2.37}$$

is true. From formulas (2.7) , (2.33) and (2.37) we obtain relation (2.34). The proposition is proved. □

Using Proposition 2.1 we deduce the following important assertion.

Proposition 2.2. *Let* S *be a bounded, positive, invertible operator and let the operator* V *defined by formula* (2.7) *be bounded. If the relations*

$$VF_0 = R(x), \tag{2.38}$$

and

$$Vf \neq 0, \quad ||f|| \neq 0 \tag{2.39}$$

are true, then the operator V *is invertible, the operator* V^{-1} *is lower triangular, and*

$$T = V^*V. \tag{2.40}$$

(Thus the operator T *admits the right triangular factorization.)*

Proof. It follows from the boundedness of the operator V and relation (2.34) that the operator W is also bounded. Let us consider

$$(Wf, R) = \int_a^b w^*(a,t)f(t)dt = (f, F_0), \tag{2.41}$$

i.e.,

$$W^*R = F_0. \tag{2.42}$$

Due to (2.38) and (2.42) we have

$$VW^*R = R. \tag{2.43}$$

From (2.34) we deduce that

$$WTW^* = VW^*. \tag{2.44}$$

Using (2.44) we see that the operator VW^* is selfadjoint and lower triangular. It means that the operator VW^* has the form

$$VW^*f = L(x)f, \tag{2.45}$$

where $L(x)$ is an $m\times m$ matrix function. Taking into account equality (2.43) we have $L(x) = I_m$, i.e.,

$$VW^* = I, \quad WV^* = I. \tag{2.46}$$

Let us introduce the notation $H = W^*L_m^2(a,b)$. If for all $h \in H$ the relation $(g,h) = 0$ is true, then $Wg = 0$. Hence in view of relation (2.34) we obtain that

$$Vf = 0 \quad (f = T^{-1}g). \tag{2.47}$$

From condition (2.39) we deduce that $g = 0$. Then the equality

$$H = L_m^2(a,b) \tag{2.48}$$

is valid. Due to (2.46) and (2.48) the operator W^* maps $L_m^2(a,b)$ onto $L_m^2(a,b)$ one-to-one. According to the classical Banach theorem [1] the operator W^* is invertible. It follows from (2.46) that the operator V is also invertible and

$$V^{-1} = W^*, \tag{2.49}$$

and

$$V^*W = I. \tag{2.50}$$

From (2.34) and (2.50) we directly obtain that $T = V^*V$. The proposition is proved. □

Example 2.1. Let us consider the operator

$$Sf = f(x) + \frac{i}{\pi}V.P.\int_a^b f(t)\frac{c(t)c(x)}{x-t}dt, \quad -\infty < a < b < \infty, \tag{2.51}$$

where $0 < m < c(t) < 1$. The operator (2.51) does not satisfy condition (1.5) but admits the left triangular factorization (see [15]).

3. Operator identity and factorization problems

We consider the operators A, S, Π and J satisfying the operator identity

$$AS - SA^\star = i\Pi J\Pi^*. \tag{3.1}$$

We suppose that the operators A and S act on the Hilbert space $L^2_m(0,b)$, the operator Π maps G ($\dim G = n < \infty$) into $L^2_m(0,b)$, the operator J acts on G, and $J = J^*$, and $J^2 = I_n$. We note that the operator Π has the form $\Pi g = [\phi_1(x), \phi_2(x), \dots, \phi_n(x)]g$, where $\phi_k(x)$ are $m\times 1$ vector functions, $g = col[g_1, g_2, \dots, g_n]$, $\phi_k(x) \in L^2_m(0,b)$. Relation (3.1) is fulfilled for the operators S which play an important role in the spectral theory of the canonical differential systems (see [18]). We shall use the following result ([18], Ch. 4).

Theorem 3.1. *Let the following conditions be fulfilled.*

1. *The operator S is bounded, positive and invertible.*
2. *The relations*
$$A^\star P_\xi = P_\xi A^\star P_\xi, \quad 0 \le \xi \le b \tag{3.2}$$
are true.
3. *The spectrum of the operator A is concentrated at the origin and there is a constant $M > 0$ such that*
$$||(P_{\xi+\Delta\xi} - P_\xi)A(P_{\xi+\Delta\xi} - P_\xi)|| \le M|\Delta\xi|, \quad 0 \le \xi \le b. \tag{3.3}$$

Then the $n\times n$ matrix function

$$W(\xi, z) = I_n + izJ\Pi^\star S_\xi^{-1}(I - zA)^{-1}P_\xi\Pi \tag{3.4}$$

satisfies the matrix integral equation

$$W(x,z) = I_n + izJ\int_0^x [dB(t)]W(t,z), \tag{3.5}$$

where

$$B(\xi) = \Pi^\star S_\xi^{-1}P_\xi\Pi. \tag{3.6}$$

From relations (1.4) and (3.6) we obtain the necessary conditions for the operator S to admit the left triangular factorization.

Proposition 3.1. *Let the operator S satisfy the relation* (3.1) *and let the conditions of Theorem* 3.1 *be fulfilled. If the operator S admits the left triangular factorization, then the matrix function $B(x)$ is absolutely continuous and*

$$\frac{d}{dx}B(x) = H(x) = \beta^*(x)\beta(x), \tag{3.7}$$

where

$$\beta(x) = [h_1(x), h_2(x), \dots, h_n(x)], \quad h_k(x) = V\phi_k(x), \quad V = S_-^{-1}. \tag{3.8}$$

Using relations (3.5) and (3.7) we obtain that

$$\frac{d}{dx}W(x,z) = izJH(x)W(x,z). \tag{3.9}$$

Lemma 3.1. *Let the conditions of Proposition* 3.1 *be fulfilled and let the* $m\times 1$ *vector functions*

$$F_j(x,z) = (I - Az)^{-1}\phi_j, \quad 1 \le j \le n \tag{3.10}$$

form a complete system in $L^2_m(a,b)$. *Then we have the equality*

$$\mathrm{mes}E = 0, \tag{3.11}$$

where the set E *is defined by the relation*

$$x \in E \quad if \quad H(x) = 0. \tag{3.12}$$

Proof. We use the following relation (see [18], Ch. 4):

$$\frac{J - W^*(\xi,\mu)JW(\xi,\lambda)}{i(\overline{\mu} - \lambda)} = \Pi^*(I - \overline{\mu}A^*)^{-1}S_\xi^{-1}(I - \lambda A^{-1}P_\xi\Pi. \tag{3.13}$$

Formula (3.13) implies that

$$(S_\xi^{-1}F_j(x,\lambda), F_\ell(x,\mu))_\xi = \frac{i[Y_\ell^*(\xi,\mu)JY_j(\xi,\lambda) - Y_\ell^*(0,\mu)JY_j(0,\lambda)]}{\overline{\mu} - \lambda}, \tag{3.14}$$

where $Y_j(x,\lambda) = col[W_{1,j}(x,\lambda), W_{2,j}(x,\lambda), \ldots, W_{n,j}(x,\lambda)]$. Here $W_{i,j}(x,\lambda)$ are entries of $W(x,\lambda)$. In view of (3.9) and (3.14) we have

$$\frac{d}{d\xi}(S_\xi^{-1}F_j(x,\lambda), F_\ell(x,\mu))_\xi = 0, \quad \xi \in E. \tag{3.15}$$

From (3.12) and (3.15) it follows that

$$\frac{d}{d\xi}(VF_j(x,\lambda), VF_\ell(x,\mu))_\xi = 0, \quad \xi \in E, \tag{3.16}$$

i.e., the relation

$$[VF_j](x,\lambda) = 0, \quad x \in E, \quad 1 \le j \le n, \tag{3.17}$$

is true. As the operator V is invertible and the system of functions $F_j(x,\lambda)$ is complete in $L^2_m(0,b)$, the system of the functions $VF_j(x,\lambda)$ is also complete in $L^2_m(0,b)$. The assertion of the lemma follows from this fact and equality (3.17). □

Further we suppose that the $n\times n$ matrix function $B(x)$ is absolutely continuous and that relations (3.7), (3.8) are true. Let us introduce the $m\times m$ matrix functions

$$R(x) = h_1(x)\alpha_1 + h_2(x)\alpha_2 + \cdots + h_n(x)\alpha_n, \tag{3.18}$$

$$F_0(x) = \phi_1(x)\alpha_1 + \phi_2(x)\alpha_2 + \cdots + \phi_n(x)\alpha_n, \tag{3.19}$$

$$v(\xi,x) = S_\xi^{-1}P_\xi F_0(x), \tag{3.20}$$

where α_k are constant $1\times m$ matrices. From Proposition 3.1 we deduce:

Corollary 3.1. *Let the conditions of Theorem* 3.1 *and Lemma* 3.1 *be fulfilled. If* $m = 1$, *then there exist numbers* $\alpha_1, \alpha_2, \ldots, \alpha_n$ *such that almost everywhere we have the inequality*

$$R(x) \ne 0. \tag{3.21}$$

Now we can formulate the main result of this section.

Theorem 3.2. *Let the following conditions be fulfilled.*

1. *The operator S satisfies relation* (3.1).
2. *The conditions of Theorem* 3.1 *are valid.*
3. *The matrix function $B(x)$ is absolutely continuous and formulas* (3.7) *and* (3.8) *are true.*
4. *The vector functions $F_j(x,\lambda)$ $(1 \le j \le n)$ form a complete system in $L^2_m(a,b)$.*
5. *Almost everywhere the inequality*

$$\det R(x) \neq 0 \tag{3.22}$$

holds.

Then the operator $T = S^{-1}$ admits the right triangular factorization

Proof. We introduce the operator

$$Vf = [R^*(x)]^{-1} \frac{d}{dx} \int_0^x v^*(x,t) f(t) dt. \tag{3.23}$$

From (3.4), (3.22) and (3.23) we deduce the equality

$$VF_j = [h_1(x), \dots, h_n(x)] Y_j(x,z). \tag{3.24}$$

Relation (3.24) implies that

$$(VF_j(x,\lambda), VF_\ell(x,\mu)) = \int_0^b Y_\ell^*(x,\mu) H(x) Y_j(x,\lambda) dx. \tag{3.25}$$

Using equality (3.24) and relation

$$\frac{d}{dx} Y_j(x,z) = izJH(x) Y_j(x,z) \tag{3.26}$$

we have

$$(VF_j(x,\lambda), VF_\ell(x,\mu)) = \frac{i[Y_\ell^*(b,\mu) J Y_j(b,\lambda) - Y_\ell^*(0,\mu) J Y_j(0,\lambda)]}{\overline{\mu} - \lambda}. \tag{3.27}$$

Comparing formulas (3.14) and (3.27) we obtain the equality

$$T = V^* V. \tag{3.28}$$

This means that the introduced operator V is bounded, $Vf \neq 0$, and $||f|| \neq 0$. Taking into account (3.18), (3.19) and (3.24) when $z = 0$ we obtain the relation

$$VF_0 = R. \tag{3.29}$$

Thus all conditions of Proposition 2.2 are fulfilled. The assertion of the theorem follows from Proposition 2.2. □

Proposition 3.2. *Let the following conditions be fulfilled.*

1. *Conditions* 1–3 *of Theorem* 3.2 *are valid.*
2. *The $m \times m$ blocks $b_{1,j}(x)$ $(1 \le j \le n)$ of the matrix $B(x)$ are absolutely continuous and*

$$b_{1,j}(x) = h_1^*(x) h_j(x). \tag{3.30}$$

3. *All the entries of the matrices $h_j(x)$ belong to $L^2(a,b)$.*

4. *Almost everywhere the inequality* (3.22) *holds. Here* $R(x) = h_1(x)$.

Then the operator V *defined by formula* (3.23) *and the equality*

$$v(\xi, x) = S_\xi^{-1} P_\xi \phi_1(x) \tag{3.31}$$

are bounded.

Proof. We introduce the matrix $H(x) = \beta^*(x)\beta(x)$, where $\beta(x) = [h_1(x), h_2(x), \ldots, h_n(x)]$. Relations (3.23)–(3.25) remain true. We use the formula

$$\int_0^b Y_\ell^*(x,\mu)[dB(x)]Y_j(x,\lambda)dx = \frac{i[Y_\ell^*(b,\mu)JY_j(b,\lambda) - Y_\ell^*(0,\mu)JY_j(0,\lambda)]}{\overline{\mu} - \lambda} \tag{3.32}$$

and the inequality $H(x)dx \le dB(x)$. From formulas (3.14), (3.25) and (3.32) we deduce that

$$V^*V \le T. \tag{3.33}$$

The proposition is proved. □

4. Operators with difference kernels

Let us consider the bounded, positive and invertible operator S with the difference kernel

$$Sf = \frac{d}{dx}\int_0^a f(t)s(x-t)dt. \tag{4.1}$$

Let us put

$$Af = i\int_0^x f(t)dt, \quad f \in L^2(0,a). \tag{4.2}$$

Equality (3.1) is valid (see [17], Ch. 1), if

$$J = \begin{bmatrix} 0 & 1 \\ 1 & 0 \end{bmatrix}, \tag{4.3}$$

$$\phi_1(x) = M(x), \quad \phi_2(x) = 1, \tag{4.4}$$

where $M(x) = s(x), \quad 0 \le x \le a$. In the case under consideration the matrix $B(\xi)$ has the form

$$B(\xi) = \begin{bmatrix} (S_\xi^{-1}M, M) & (S_\xi^{-1}1, M) \\ (S_\xi^{-1}M, 1) & (S_\xi^{-1}1, 1) \end{bmatrix}. \tag{4.5}$$

The corresponding function $F(x,\lambda)$ has the form

$$F(x,\lambda) = e^{ix\lambda}. \tag{4.6}$$

The operator A defined by formula (4.2) satisfies all the conditions of Theorem 3.1. The following fact is useful here.

Theorem 4.1. *Let the operator S be bounded, positive, invertible and have the form* (4.1). *If the matrix function $B(x)$ is absolutely continuous and*

$$B'(x) = \beta^*(x)\beta(x), \quad \beta(x) = [h_1(x), h_2(x)], \tag{4.7}$$

then the equality

$$h_1(x)\overline{h_2(x)} + h_2(x)\overline{h_1(x)} = 1 \tag{4.8}$$

is true almost everywhere.

Proof. Let us consider the expression

$$i_\xi = (S_\xi^{-1}P_\xi M, 1) + (1, S_\xi^{-1}P_\xi M). \tag{4.9}$$

Setting

$$N_1(x,\xi) = S_\xi^{-1}P_\xi M, \tag{4.10}$$

we rewrite formula (4.9) in the form $i_\xi = \int_0^\xi [N_1(x,\xi) + \overline{N_1(x,\xi)}dx$, i.e.,

$$i_\xi = \int_0^\xi [N_1(x,\xi) + \overline{N_1(\xi - x,\xi)}]dx. \tag{4.11}$$

We use the relation (see [17], Ch. 1)

$$N_1(x,\xi) + \overline{N_1(\xi - x,\xi)} = 1. \tag{4.12}$$

In view of (4.11) and (4.12) we obtain the equality

$$i_\xi = \xi. \tag{4.13}$$

Taking into consideration Equalities (2.1), (3.8), (4.1) and (4.9) we deduce that

$$i_\xi = \int_0^\xi [h_1(x)\overline{h_2(x)} + h_2(x)\overline{h_1(x)}]dx. \tag{4.14}$$

Relation (4.8) follows from (4.13) and (4.14). The theorem is proved. □

From equality (4.8) we have

$$h_2(x) \neq 0, \quad 0 \leq x \leq a. \tag{4.15}$$

Remark 4.1. The operators of the form

$$Sf = f(x) + \int_0^a f(t)k(x-t)dt, \tag{4.16}$$

where $k(x) \in L(-a,a)$, belong to class (4.1). For this case inequality (4.15) was deduced by M.G. Krein by another method (see [5], Ch. 4). The main result of this section follows directly from Proposition 3.1, Theorem 3.2 and Inequality (4.15).

Theorem 4.2. *Let the operator S be positive, invertible and have the form* (4.1). *The operator S admits the left triangular factorization if and only if the matrix $B(x)$ is absolutely continuous and relation* (4.7) *is valid.*

Remark 4.2. If a bounded operator S on $L^2(0,b)$ has the form (4.1) and admits one of the factorizations (left or right) then it admits another factorization too (see [15]).

Example 4.1. Let us consider the operator S_β of the form

$$S_\beta f = f + \frac{i\beta}{\pi} V.P. \int_0^b \frac{f(t)}{x-t} dt, \tag{4.17}$$

where $-1 < \beta < 1$. This operator with a difference kernel is bounded, invertible and positive (see [15]). The operator S_β does not satisfy condition (1.4). Nevertheless S_β admits the left triangular factorization $S_\beta = W_\alpha W_\alpha^*$, where

$$W_\alpha f = \frac{x^{i\alpha}}{\sqrt{ch(\pi\alpha)}\Gamma(i\alpha - 1)} \frac{d}{dx} \int_0^x f(t)(x-t)^{-i\alpha} dt. \tag{4.18}$$

Here $\alpha = \frac{1}{\pi}\text{arcth}\beta$, and $\Gamma(z)$ is the gamma function.

5. Operators with sum–difference kernels

Let us consider the following class of bounded and positive operators which can be represented in the form $((+,-)$-class):

$$Sf = \frac{d^2}{dx^2} \int_0^b [s_1(x-t) + s_2(x+t)] f(t) dt, \tag{5.1}$$

where $f(t) \in L^2(0,b)$. We introduce the operator

$$Af = \int_0^x (t-x) f(t) dt. \tag{5.2}$$

Then the operator identity (3.1) is valid. Here the 4×4 matrix J is defined by the relation

$$J = \begin{bmatrix} 0 & I_2 \\ I_2 & 0 \end{bmatrix}, \tag{5.3}$$

and the operator Π has the form

$$\Pi = [\Phi_1, \Phi_2], \tag{5.4}$$

the operators Φ_1 and Φ_2 are defined by the relations

$$\Phi_1 g = -iM(x) g_1 - iM_0(x) g_2, \tag{5.5}$$

$$\Phi_2 g = g_1 + x g_2, \tag{5.6}$$

where

$$M(x) = -[s_1(x) + s_2(x)], \quad M_0(x) = s_1'(x) - s_2'(x), \tag{5.7}$$

and a constant 2×1 vector g has the form $g = \text{col}[g_1, g_2]$. The main result of this section follows directly from Proposition 3.1, Lemma 3.1 and Theorem 3.2.

Theorem 5.1. *Let the operator S be positive, invertible and have the form* (5.1). *The operator S admits the left triangular factorization if and only if the matrix* $B(x)$ *is absolutely continuous and*

$$B'(x) = \beta^*(x)\beta(x), \quad \beta(x) = [h_1(x), h_2(x), h_3(x), h_4(x)]. \tag{5.8}$$

Example 5.1. Let us consider the equation

$$Sf = f(x) + \frac{i\mu}{\pi} V.P. \int_0^1 \frac{f(t)}{x-t} dt - \frac{\lambda}{\pi} \int_0^1 \frac{f(t)}{x+t} dt = g(x), \tag{5.9}$$

where $f(x) \in L^2(0,1)$, $\lambda = \overline{\lambda}$, $\mu = \overline{\mu}$, and $|\lambda| + |\mu| < 1$. It is well known ([5], Ch. 9) that the operator S is bounded, positive and invertible, i.e., the operator S belongs to the (+,-) class . We introduce the functions

$$v(x,\lambda,\mu) = S^{-1}1, \quad \alpha(\lambda,\mu) = \int_0^1 v(x,\lambda,\mu)dx = (S^{-1}1,1) > 0. \tag{5.10}$$

In view of (5.9) and (5.10) the relations

$$S_\xi^{-1} P_\xi 1 = v(\frac{x}{\xi},\lambda,\mu), \quad (S_\xi^{-1} P_\xi 1,1)_\xi = \xi\alpha(\lambda,\mu) \tag{5.11}$$

are true. We introduce the operator

$$Vf = \frac{1}{\sqrt{\alpha(\lambda,\mu)}} \frac{d}{dx} \int_0^x f(t) v(\frac{t}{x},\lambda,\mu)dt. \tag{5.12}$$

Using Proposition 3.2 we deduce that the operator V is bounded and $S^{-1} \geq V^*V$.

Open problem 5.1 Prove that

$$Vf \neq 0, \quad \text{when } ||f|| \neq 0. \tag{5.13}$$

Remark 5.1. If relation (5.13) is true, then $S^{-1} = V^*V$ and the operator S admits the left triangular factorization

$$S = V^{-1}[V^*]^{-1}. \tag{5.14}$$

Remark 5.2. Relation (5.13) is valid when $\lambda = 0$ (see Example 4.1). Now we consider separately the case when $\mu = 0$, i.e., the case of the Dixon equation [4], [8], [19]:

$$Sf = f(x) - \frac{\lambda}{\pi} \int_0^1 \frac{f(t)}{x+t} dt = g(x), \tag{5.15}$$

where $f(x) \in L^2(0,1)$, and $\lambda < 1$. M.G. Krein deduced the formula for the Dixon equation resolvent (see [8], Ch. 4). This formula can be written in the following way: $S^{-1} = V^*V$. Thus we obtain:

Proposition 5.1. *The Dixon operator S defined by* (5.15) *admits the left triangular factorization $S = V^{-1}[V^*]^{-1}$, where the operator V has the form* (5.12).

6. Triangular factorization, Class R_1

Let us consider the integral operators

$$Af = i \int_0^x f(t)dt, \quad A^*f = -i \int_x^b f(t)dt, \tag{6.1}$$

where $f(x) \in L^2(0,b)$.

Definition 6.1. We say that the linear bounded operator S acting in the Hilbert space $L^2(0,b)$ belongs to the class R_1 (rank 1) if the following conditions are fulfilled:

1)
$$m(f,f) \le (Sf,f) \le M(f,f), \quad 0 < m < M < \infty, \tag{6.2}$$

2) $\operatorname{rank}(AS - SA^*) = 1$, i.e.,
$$(AS - SA^*)f = i(f,\phi)\phi, \quad \phi(x) \in L^2(0,b). \tag{6.3}$$

We associate with the operator S the operator
$$S_- f = \frac{d}{dx}\int_0^x f(t)\phi(x-t)dt. \tag{6.4}$$

It is easy to see that
$$S_- 1 = \phi. \tag{6.5}$$

Lemma 6.1. *Let the bounded operator S satisfy relation* (6.3)*. If the corresponding operator S_- is bounded, then the representation*
$$S = S_- S_-^* \tag{6.6}$$
is true.

Proof. We consider the operator
$$X = S_- S_-^*. \tag{6.7}$$

Using formula (6.3) and relation $AS_- = S_- A$ we deduce the equality
$$AX - XA^* = S_-(A - A^*)S_-^* = AS - SA^*. \tag{6.8}$$

The equation $AX - XA^* = F$ has no more than one solution X (see [17], Ch. 1). Hence we deduce from (6.8) that $S = X$. The lemma is proved. □

Lemma 6.2. *If the bounded operator S satisfies the relation* (6.3)*, then this operator can be represented in the form* (6.6)*, where the operator S_- is defined by formula* (6.4)*.*

Proof. To prove that the operator S_- is bounded we introduce the operator
$$X_- f = AS_- f = i\int_0^x f(t)\phi(x-t)dt. \tag{6.9}$$

We note that
$$X_-^* f = S_-^* A^* f = -i\int_x^b f(t)\overline{\phi(t-x)}dt \tag{6.10}$$
where the operator S_-^* has the form
$$S_-^* f = -\frac{d}{dx}\int_x^b f(t)\overline{\phi(t-x)}dt. \tag{6.11}$$

According to Lemma 6.1 we have
$$ASA^* = X_- X_-^*. \tag{6.12}$$

It follows from relations (6.9) and (6.12) that $S = S_- S_-^*$. Hence the operator S_- is bounded. The lemma is proved. □

Now we shall deduce the main result of this section.

Theorem 6.1. *If the operator S belongs to the class R_1, then this operator admits the left triangular factorization.*

Proof. We suppose that for some $f_0(x) \in L^2(0,b)$ the relation

$$S_- f_0 = 0 \quad (||f_0|| \neq 0) \tag{6.13}$$

is true. In view of the well-known Titchmarsh theorem (see [19], Ch. 11) and (6.13) we have

$$\phi(x) = 0, \quad 0 \leq x \leq \delta. \tag{6.14}$$

Using (6.3) and (6.14) we deduce that

$$A_\delta S_\delta - S_\delta A_\delta^* = 0, \tag{6.15}$$

where $A_\delta f = i\int_0^x f(t)dt$, $0 \leq x \leq \delta$, and $S_\delta = P_\delta S P_\delta$. Operator equation (6.15) has only the trivial solution $S_\delta = 0$ (see [17], Ch. 1). The last equality contradicts relation (6.2). It means that equality (6.13) is impossible when $||f_0|| \neq 0$. Hence in view of (6.6) the operator S_- maps $L^2(0,b)$ one-to-one onto $L^2(0,b)$. This fact according to the classical Banach theorem [1] implies that the operator S_- is invertible. The operator S_-^{-1} is defined by formula (see [17], Ch. 1)

$$S_-^{-1} f = \frac{d}{dx}\int_0^x f(t)N(x-t)dt, \tag{6.16}$$

where $N(x) = S_-^{-1}1$. Thus the operators S_- and S_-^{-1} are bounded and lower triangular. The assertion of the theorem now follows directly from Definition 1.4. □

Example 6.1. We consider the case when

$$\phi(x) = \log(b-x). \tag{6.17}$$

In this case we have

$$S_- f = \frac{d}{dx}\int_0^x f(t)\log(b-x+t)dt = f(x)\log b - \int_0^x \frac{f(t)}{b-x+t}dt. \tag{6.18}$$

Let us introduce the operator

$$Kf = \int_0^x \frac{f(t)}{b-x+t}dt. \tag{6.19}$$

It is well known (see [19], Ch. 11) that $||K|| \leq \pi$. Hence the operator S_- defined by (6.18) and the operator S_-^{-1} are bounded, when $\log b > \pi$. From Lemma 6.1 we obtain the assertion.

Proposition 6.1. *If $\log(b) > \pi$, then the operator S defined by relations* (6.3) *and* (6.17) *admits the left triangular factorization* (6.6) *where the operator S_- has the form* (6.18).

7. Homogeneous kernels of degree (–1)

In this section we consider operators of the form

$$SF = F(x) - \int_0^1 F(y)k(\frac{y}{x})\frac{1}{x}dy = G(x), \tag{7.1}$$

where $F(x) \in L^2(0,1)$ and

$$k(\frac{y}{x})\frac{1}{x} = \overline{k(\frac{x}{y})}\frac{1}{y}. \tag{7.2}$$

We assume that

$$A = 2\int_0^1 |k(\frac{1}{x})|x^{-3/2}dx < \infty. \tag{7.3}$$

It follows from condition (7.2) that the operator S is selfadjoint. From condition (7.3) we deduce that the operator

$$KF = \int_0^1 F(y)k(\frac{y}{x})\frac{1}{x}dy \tag{7.4}$$

is bounded and (see [5], Ch. 9)

$$||K|| \leq A. \tag{7.5}$$

Theorem 7.1. *Let conditions* (7.2) *and* (7.3) *be fulfilled and let the corresponding operator S be positive and invertible, then the operator S admits the left triangular factorization.*

Proof. We introduce the change of variables $x = e^{-u}$ and $y = e^{-v}$. Hence equation (7.1) takes the form

$$Lf = f(u) - \int_0^\infty f(v)H(u-v)dv = g(u), \tag{7.6}$$

where

$$f(u) = F(e^{-u})e^{-u/2}, \quad g(u) = G(e^{-u})e^{-u/2}, \tag{7.7}$$

$$H(u) = \overline{H(-u)} = k(e^u)e^{u/2}, \quad u \geq 0. \tag{7.8}$$

It follows from relation (7.3) that

$$\int_{-\infty}^\infty |H(u)|du = A. \tag{7.9}$$

We denote by $\gamma(u)$ the solution of Equation (7.6) when $g(u) = H(u)$. In the theory of equations (7.6) the following function plays an important role (see [8], Ch. 2):

$$G_+(\lambda) = 1 + \int_0^\infty \gamma(u)e^{it\lambda}dt, \quad \mathrm{Im}\lambda \geq 0.$$

Let us consider the solution $\gamma_\xi(u)$ of equation (7.6) when $g(u) = e^{iu\xi}$ and $\mathrm{Im}\,\xi \geq 0$. We use the formula (see [8], Ch. 2)

$$\gamma_\xi(u) = \overline{G_+(-\overline{\xi})}[1 + \int_0^u \gamma(r)e^{-ir\xi}dr]e^{iu\xi}. \tag{7.10}$$

Further we need the particular case of $\gamma_\xi(u)$ when $\xi = i/2$. In this case we have

$$\gamma_{i/2}(u) = \beta[1 + \int_0^u \gamma(r)e^{r/2}dr]e^{-u/2}, \tag{7.11}$$

where

$$\beta = \overline{G_+(i/2)}. \tag{7.12}$$

Let us introduce the function $v(x)$, which satisfies Equation (7.1) when $G(x) = 1$. It is easy to see that

$$v(e^{-u}) = \gamma_{i/2}(u)e^{u/2}. \tag{7.13}$$

From (7.11) and (7.13) we deduce that

$$v'(x)x^2 = -\beta\gamma(t)e^{-t/2}, \tag{7.14}$$

and

$$v(1) = \beta. \tag{7.15}$$

Using relations (7.11) and (7.13) we can calculate the integral

$$\alpha = \int_0^1 v(x)dx = \beta[1 + \int_0^1 \int_0^{-\log x} \gamma(r)e^{r/2}drdx].$$

Hence the equalities

$$\alpha = \beta[1 + \int_0^\infty \gamma(r)e^{-r/2}drdx] = \beta\overline{\beta} \tag{7.16}$$

are true. The operator V in (7.1) has the form

$$Vf = \frac{1}{\beta}\frac{d}{dx}\int_0^x f(t)v(\frac{t}{x})dt. \tag{7.17}$$

In view of (7.14) and (7.15) we can represent the operator V in the form

$$Vf = f(x) + \int_0^x f(t)L(\frac{t}{x})\frac{1}{t}dt, \tag{7.18}$$

where

$$L(x) = \gamma(t)e^{-t/2}. \tag{7.19}$$

Now the assertion of the theorem follows from Proposition 2.2. □

Corollary 7.1. *Let the conditions of Theorem* 7.1 *be fulfilled. Then we have the equality*

$$S^{-1} = V^*V, \tag{7.20}$$

where the operator V is defined by relations (7.18) *and* (7.19).

Example 7.1. We obtain an interesting example when

$$k(u) = \frac{\lambda}{|1-u|^\alpha(1+u)^\beta}, \tag{7.21}$$

where $\lambda = \overline{\lambda}$, $\alpha \geq 0$, $\beta > 0$, and $\alpha + \beta = 1$. We note that $k(u)$ satisfies conditions (7.2) and (7.3). Equations (7.1) and (7.21) coincide with the Dixon equation when $\alpha = 0$.

References

[1] S. Banach, Sur les fonctionnelles lineares, I, II, *Studia Math* (1929), 211–216, 223–239.

[2] M.S. Brodskii, *Triangular and Jordan Representations of Linear Operators*, v. 32, Amer. Math. Soc., Providence, 1971.

[3] K.R. Davidson K.R., *Nest Algebras*, Pitnam, Res. Notes Math., 1988.

[4] A.C. Dixon, On the solving nuclei of certain integral equation etc., *Proc. London Math. Soc. (2)* **27** (1926), 233–272.

[5] I. Gohberg and M.G. Krein, *Theory and Applications of Volterra Operators in Hilbert Space*, Amer. Math. Soc., Providence, 1970.

[6] G.H. Hardy, J.E. Littlewood and G. Polya, *Inequalities*, London, 1951.

[7] R. Kadison and I. Singer I., Triangular Operator Algebras, *Amer. J. Math.* **82** (1960), 227–259.

[8] M.G. Krein, Integral Equations on a Halfline with a Kernel Depending on the Difference of the arguments, *Usp. Math. Nauk.* **13:5** (1958), 3–120 (in Russian).

[9] D.R. Larson, Nest Algebras and Similarity Transformation, *Ann. Math.* **125** (1985), 409–427.

[10] M.S. Livsic, *Operators, Oscillations, Waves, Open Systems*, v. 34, Amer. Math. Soc., Providence, 1973.

[11] L.A. Sakhnovich, Triangular Integro-Differential Operators with Difference Kernels, *Sib. Mat. Journ.* **19** no. 4 (1978), 871–877 (in Russian).

[12] L.A. Sakhnovich, On Reduction of Non-adjoint Operators to Triangular Form, *Izvest. Visch. Uch. Zav, ser. mat.* **1** (1959), 80–86 (in Russian).

[13] L.A. Sakhnovich, Investigation of Triangular Model of Non-adjoint Operators, *Izvest. Visch. Uch. Zav, ser. mat.* **14** (1959), 141–149 (in Russian).

[14] L.A. Sakhnovich, Factorization of Operators in $L^2(a,b)$, in: *Linear and Complex Analysis, Problem Book* (V.P. Havin, S.V. Hruscev and N.K. Nikol'ski, ed.), Springer Verlag, 172–174, 1984.

[15] L.A. Sakhnovich, Factorization of Operators in $L^2(a,b)$, *Functional Anal. and Appl.* **13** (1979), 187–192 (in Russian).

[16] L.A. Sakhnovich, Factorization Problems and Operator Identities, *Russian Math. Surv.* **41** no. 1 (1986), 1–64.

[17] L.A. Sakhnovich, *Integral Equations with difference Kernels on finite Intervals*, Operator Theory, Advances and Applications, v. 84, Birkhäuser, 1996.

[18] L.A. Sakhnovich, *Spectral Theory of Canonical Differential Systems. Method of Operator Identities*, Operator Theory, Advances and Applications, v. 107, Birkhäuser, 1999.

[19] E. Titchmarsh, *Introduction to the Theory of Fourier Integrals*, Oxford, 1970.

Lev A. Sakhnovich
735 Crawford Avenue
Brooklyn, NY 11223
USA
e-mail: Lev.Sakhnovich@verizon.net

Operator Theory:
Advances and Applications, Vol. 179, 309–328

Solutions for the $H^\infty(D^n)$ Corona Problem Belonging to $\exp(L^{\frac{1}{2n-1}})$

Tavan T. Trent

Abstract. For a countable number of input functions in $H^\infty(D^n)$, we find explicit analytic solutions belonging to the Orlicz-type space, $\exp(L^{\frac{1}{2n-1}})$. Note that $H^\infty(D^n) - BMO(D^n) \subsetneqq \exp(L^{\frac{1}{2n-1}}) \subsetneqq \bigcap_1^\infty H^p(D^n)$.

Mathematics Subject Classification (2000). 30A38, 46J20.

Keywords. Corona theorem, polydisk.

We give a solution, for general corona data on the polydisk, which, although not bounded or even in BMO, still belongs to a space better than $\bigcap_{p=1}^\infty H^p(D^n)$; namely the Orlicz-type space, $\exp(L^{\frac{1}{2n-1}})$. Also, we establish the $\mathcal{H}^p$-corona theorem on the polydisk. This paper is closely related to Trent [19]. The main purpose of the paper is to make explicit the algebraic requirements to represent solutions on D^n.

For the case of two functions in the input data, Chang [6] showed that solutions to the general corona problem for the bidisk can be found which belong to $\bigcap_{p=1}^\infty H^p(D^2)$. Again for two functions on the bidisk, Amar [1] and Cegrell [5] have found solutions to the general corona problem for the bidisk belonging to BMO. For a finite number of input functions, the $\overline{\partial}$-input data is more complicated and, for this case, first Varopoulos [20] and then Lin [14] found solutions to the general corona problem on the polydisk belonging to $\bigcap_{p=1}^\infty H^p(D^n)$. [See Chang and R. Fefferman [8] for a brief discussion of the difference (involving the Koszul complex) between two and a general finite number of input functions.] However, even in this case no relationship between the lower bound of the input data (denoted by ϵ) and the size of the solutions was obtained. An estimate will be given in this paper.

For a finite number of input functions, Li [13] and, independently, Lin [14] implicitly solved the $H^p(D^n)$ corona theorem ($1 \leq p < \infty$) based on the work of Lin. Again, for a finite number of input functions, Boo [4] gave an explicit solution

Partially supported by NSF Grant DMS-0400307.

to the $H^p(D^n)$ corona theorem ($1 \leq p < \infty$), which was based on integral formulas. For the case $p = 2$, we have already established the $\mathcal{H}^2(D^n)$ corona theorem in the vector-valued case; that is, with an infinite number of input functions. The current paper will establish the $\mathcal{H}^p(D^n)$ corona theorem in the vector-valued case.

We note that for the unit ball or even strictly pseudoconvex domains in $\mathbb{C}^n$, there are more complete results known. See Andersson and Carlsson [2] for these precise results and further references. Of course, when the dimension is greater than 1, bounded analytic solutions have not been found in any of the above cases, for general bounded analytic input data.

Our main technique is a linear algebra result, which enables us to exhibit explicit solutions (in the smooth case) which have the appropriate estimates. This is based on considering explicit mappings, arising from the Koszul complex. The basic idea for the corona estimates involves iterating the one-variable Littlewood-Paley results, motivated by T. Wolff's proof of Carleson's corona theorem on the unit disk. (See Garnett [10].) However, to simplify the estimates we will appeal to the remarkable $H^1(D^n)$ weak factorization theorem of Lacey-Terwilleger [12].

We will use the following notation:

D	open unit disk in the complex plane, $\mathbb{C}$
T	unit circle, $T = \partial D$
D^n	polydisk, $D^n = D \times \cdots \times D$, n times
T^n	distinguished boundary of D^n, $T^n = T \times \cdots \times T$, n times
$d\sigma$	normalized Lebesgue measure on $[-\pi, \pi]$
$d\boldsymbol{\sigma}_k$	denotes $d\sigma_1(t_1) \ldots d\sigma_k(t_k)$
dA	area Lebesgue measure on D
dL	measure on D defined by $dL(z) = \ln \dfrac{1}{\lvert z\rvert^2} \dfrac{d\,A(z)}{\pi}$
$d\mathbf{L}_k$	denotes $dL_1(z_1) \ldots dL_k(z_k)$
$H^p(D^n)$	Hardy space of analytic functions on $D^n, 1 \leq p \leq \infty$
	[We will also identify this space with
	$\{f \in L^p(T^n) \mid$ for $\{k_j\}_{j=1}^n \subset Z$ with at least one $k_j < 0$,
	$\displaystyle\int_{T^n} f(e^{it_1}, \ldots, e^{it_n}) e^{-ik_1 t_1} \ldots e^{-ik_n t_n}\, d\sigma(t_1) \ldots d\sigma(t_n) = 0\}]$
$\mathcal{L}^p(T^n)$	$\{f : T^n \to l^2 \mid f$ is strongly measurable and
	$\lvert f\rvert_p^p \stackrel{\text{def}}{=} \displaystyle\int_{T^n} \lVert f(e^{it_1}, \ldots, e^{it_n})\rVert_2^p\, d\sigma(t_1) \ldots d\sigma(t_n) < \infty\}$
	for $1 \leq p < \infty$

$\mathcal{L}^\infty(T^n)$ $\{f : T^n \to l^2 \mid f$ is strongly measurable and
$|f|_\infty \stackrel{\text{def}}{=} \operatorname{ess}\sup_{u_1,\dots,u_n \in T} \|f(u_1,\dots,u_n)\|_2 < \infty\}$

$\mathcal{H}^p(D^n)$ $\{f : D^n \to l^2 \mid f$ is analytic, l^2-valued on D^n and
$|f|_p^p \stackrel{\text{def}}{=} \sup_{r\uparrow 1}\left(\int_{T^n} \|f(r\,e^{it_1},\dots,r\,e^{it_n})\|_2^p \, d\sigma(t_1)\dots d\sigma(t_n)\right) < \infty\}$
for $1 \le p < \infty$

$\mathcal{H}^\infty(D^n)$ $\{f : D^n \to l^2 \mid f$ is analytic, l^2-valued on D^n and
$|f|_\infty \stackrel{\text{def}}{=} \sup_{z_1,\dots,z_n \in D} \|f(z_1,\dots,z_n)\|_2 < \infty\}$

$\exp(L^{\frac{1}{k}})$ $\{f \in \mathcal{H}^2(D^n) : \int_{T^n} e^{\left(\frac{\|f\|_2}{\lambda}\right)^{\frac{1}{k}}} d\sigma_1 \dots d\sigma_n \le 2$
for some $\lambda > 0$ (depending on f)$\}$

$|f|_{e,k}$ the smallest λ so that $\int_{T^n} e^{\left(\frac{\|f\|_2}{\lambda}\right)^{\frac{1}{k}}} d\sigma_1 \dots d\sigma_n \le 2$

$\widehat{\phi^j}(u)$ the j^{th} Cauchy transform of a (possibly l^2-valued) $C^{(1)}$ function on $\overline{D}^n$

$$\widehat{\phi^j}(u_1,\dots,\overset{jth}{\underset{z}{\simeq}},\dots,u_n) = -\frac{1}{\pi}\int_D \frac{\phi(u_1,\dots,\overset{jth}{z},\dots,u_n)}{w-z}\, dA(w)$$

T_F Toeplitz operator with symbol F acting on $\mathcal{H}^p(D^n)$ for any $1 \le p < \infty$

F the operator of pointwise multiplication by the matrix $[f_{jk}(e^{it_1},\dots,e^{it_n})]_{j,k=1}^\infty = F(e^{it_1},\dots,e^{it_n})$ on $\mathcal{L}^p(T^n)$

$F(z_1,\dots,z_n)$ the operator on l^2 gotten by applying the matrix $F(z_1,\dots,z_n) = [f_{jk}(z_1,\dots,z_n)]_{j,k=1}^\infty$ to the standard basis of l^2.

We will prove the following two theorems:

Theorem A. *Let $F \in \mathcal{H}^\infty(D^n)$ and assume that*

$$0 < \epsilon^2 \le F(\mathbf{z})\,F(\mathbf{z})^* \le 1$$

for all $\mathbf{z} \in D^n$. Then there exists $u \in \exp(L^{\frac{1}{2n-1}})$ satisfying

$$F\,u = 1 \;\; \textit{for} \;\; |u|_{e,2n-1} \le \frac{C_0}{\epsilon^7}.$$

Theorem B ($\mathcal{H}^p(D^n)$-corona theorem). *Let $F \in \mathcal{H}^\infty(D^n)$ satisfy $0 < \epsilon^2 \leq F(\mathbf{z})\, F(\mathbf{z})^* \leq 1$ for all $\mathbf{z} \in D^n$. Then T_F acting from $\mathcal{H}^p(D^n)$ to $H^p(D^n)$ is onto for each $1 < p < \infty$.*

Observe that the general corona problem for D^n has the hypothesis of Theorem A (or Theorem B), but the conclusion requires that a solution to $F\,u = 1$ belong to $\mathcal{H}^\infty(D^n)$. Both of these theorems follow trivially from a positive solution to the general corona problem. In fact, the raison d'etre of Theorems A and B is an attempt to understand some of the difficulties of the general corona problem for the polydisk.

We will give a proof of Theorem A and show how to modify it to get Theorem B. Several well-known lemmas will be required.

Lemma 1. *Let ϕ (possibly vector-valued) be $C^{(2)}$ in a neighborhood of $\overline{D}$. Then:*

$$\text{(a) } \phi(0) = \int_{-\pi}^{\pi} \phi(e^{it})\, d\sigma(t) - \int_D \triangle\phi(z)\, \frac{dL}{4}(z),$$

$$\text{(b) for } z \in D,\ \phi(z) = \frac{1}{2\pi i} \int_{\partial D} \frac{\phi(w)}{w - z}\, dw - \frac{1}{\pi} \int_D \frac{\overline{\partial}_z \phi(w)}{w - z}\, dA(w),$$

and

$$\text{(b}') \text{ for } z \in \overline{D},\ \phi(z) = \int_{-\pi}^{\pi} \frac{\phi(e^{it})}{1 - z\, e^{-it}}\, d\sigma(t) + \widehat{(\phi)_{\overline{z}}}^{\,1}(z),$$

$$\phi(z) = (P_z\phi)(z) + \widehat{\overline{\partial}_z(\phi)}^{\,1}(z).$$

See Koosis [11] for details. Notice that for smooth functions on $\overline{D^n}$, (b$'$) says that

$$\widehat{(\phi)_{\overline{z}_j}}^{\,j}(e^{it_1}, \ldots, e^{it_n}) = (P_j^\perp \phi)(e^{it_1}, \ldots, e^{it_n})$$

where P_j denotes the orthogonal projection of $L^2(d\sigma_1, \ldots, d\sigma_n)$ onto the subspace of functions whose Fourier coefficients, $a_{k_1,\ldots,k_n}$ are 0 if $k_j < 0$. For several variables, the order of application of the Cauchy transforms is irrelevant, so we may unambiguously write $\widehat{\phi}^{1,2,5}$, etc. to denote three applications of the Cauchy transforms on the 1st, 2nd, and 5th variables in any order.

The next lemma seems to be due to Uchiyama. See Nikolski [15] for the simple proof.

Lemma 2. *Assume that $a \in C^{(2)}(D)$, $\|a\|_{\infty,D} < \infty$, $a \geq 0$ and $\triangle a \geq 0$ on D. Then for p an analytic polynomial, we have*

$$\int_D \triangle a\, |p|\, dL \leq e\, \|a\|_\infty \int_{-\pi}^{\pi} |p|\, d\sigma.$$

To write down explicit solutions (in the smooth case), we need Cauchy transforms and the following representation theorem which appeared in Trent [18]. The proof will be provided in the Appendix for convenience. We also note that the

lemma is a purely linear algebra result, but we state it here in the context of algebras of bounded analytic functions on the polydisk, D^n.

Lemma 3. *Assume that $F \in \mathcal{H}^\infty(D^n)$. Then there exist operators $Q_l : D^n \to B(l^2)$ such that for all $\mathbf{z} \in D^n$ and $l = 0, 1, \ldots,$*

(a) $Q_l(\mathbf{z})\, Q_{l+1}(\mathbf{z}) = 0,$

(b) $(F(\mathbf{z})\, F(\mathbf{z})^*)\, I_{l^2} = Q_l^*(\mathbf{z}) Q_l(\mathbf{z}) + Q_{l+1}(\mathbf{z})\, Q_{l+1}^*(\mathbf{z}).$

Moreover, the entries of $Q_l(\mathbf{z})$ are 0, or else, for some n, either $f_n(\mathbf{z})$ or $-f_n(\mathbf{z})$.

The pertinent observation is that under the hypothesis that $0 < \epsilon^2 \leq F(\mathbf{z})\, F(\mathbf{z})^* \leq 1$, for $\mathbf{z} \in D^n$ fixed; we have for $l = 0, 1, \ldots$ and $Q_0(\mathbf{z}) = F(\mathbf{z})$:

(i) $\dfrac{Q_l(\mathbf{z})\, Q_l(\mathbf{z})^*}{F(\mathbf{z})F(\mathbf{z})^*}$ is the orthogonal projection of l^2 onto the kernel of $Q_{l-1}(\mathbf{z})$. Thus, range $Q_l(\mathbf{z}) =$ kernel $Q_{l-1}(\mathbf{z})$.

(ii) $Q_l(\mathbf{z}) Q_l(\mathbf{z})^* \leq (F(\mathbf{z})\, F(\mathbf{z})^*)\, I_{l^2} \leq I_{l^2}$.

(iii) Differentiating (b) with respect to $\mathbf{z}_j$ and $\overline{\mathbf{z}}_j$ gives us that

$$\partial_{\mathbf{z}_j} Q_l(\mathbf{z})\, (\partial_{\mathbf{z}_j} Q_l(\mathbf{z}))^* \leq \partial_{\mathbf{z}_j} F(\mathbf{z})\, (\partial_{\mathbf{z}_j} F(\mathbf{z}))^*\, I_{l^2}.$$

The proof can be found in the Appendix. By (iii),

$$\|\overline{\partial}_j Q_j(\mathbf{z})\|_{op} \leq \|\overline{\partial}_j F(\mathbf{z})\|_{l^2},$$

so all estimates involving $(Q_l)_j$ are replaced by F_j.

We give an example illustrating the finite case when we have four functions in $H^\infty(D^n)$. Then

$$F = (f_1, f_2, f_3, f_4),$$

$$Q_1 = \left[\begin{array}{ccc|ccc} f_2 & f_3 & f_4 & 0 & 0 & 0 \\ \hline -f_1 & 0 & 0 & f_3 & f_4 & 0 \\ 0 & -f_1 & 0 & -f_2 & 0 & f_4 \\ 0 & 0 & -f_1 & 0 & -f_2 & -f_3 \end{array}\right]$$

and

$$Q_2 = \left[\begin{array}{ccc|c} f_3 & f_4 & 0 & 0 \\ -f_2 & 0 & f_4 & 0 \\ 0 & -f_2 & -f_3 & 0 \\ \hline f_1 & 0 & 0 & f_4 \\ 0 & f_1 & 0 & -f_3 \\ 0 & 0 & f_1 & f_2 \end{array}\right].$$

The lemma below can be found in Stein [16, pp. 450–451]. We state the version we will require. For the proof when $n = 2$, see Trent [19].

Fix a p with $1 \le p \le \infty$ and assume that $\mathcal{T} \in B(L^p(T^n))$. We wish to define an operator $\mathcal{J}$, on $B(\mathcal{L}^p(T^n))$ as follows: For $H = (h_1, h_2, \dots) \in \mathcal{L}^p(T^n)$, we wish to define $\mathcal{J}\,H = (\mathcal{T}\,h_1, \mathcal{T}\,h_2, \dots)$. The following lemma tells us that $\mathcal{J} \in B(\mathcal{L}^p(T^n))$ and $\|\mathcal{J}\| = \|\mathcal{T}\|$.

Lemma 4. (a) *Let $\mathcal{T} \in B(L^p(T^n))$. Then $|\mathcal{J}\,H|_p \le \|\mathcal{T}\|\,|H|_p$ for all $H \in L^p(T^n)$. Thus, $\mathcal{J} \in B(\mathcal{L}^p(T^n))$ with $\|\mathcal{J}\| = \|\mathcal{T}\|$.*

(b) *The analogous result is true: $\mathcal{T} \in B(H^p(D^n))$, then $\mathcal{J} \in B(\mathcal{H}^p(D^n))$ and $\|\mathcal{J}\| = \|\mathcal{T}\|$.*

For notational purposes, we will use "$\mathcal{T}$" to denote both the operator in $B(L^p(D^n))$ and the operator in $B(\mathcal{L}^p(D^n))$. Thus, for example, "P_j" may denote the projection operator from $L^p(T^n)$ onto those $L^p(T^n)$ functions whose biharmonic extension into D^n is analytic in the jth variable or it may denote the corresponding operator from $\mathcal{L}^p(T^n)$. It should be clear from the context which operator is meant. Also, we may not always refer explicitly to this lemma, but it is clearly in the background for extending, for example, the usual Carleson measure results to the vector-valued case.

Recall that $d\boldsymbol{\sigma}_k = d\sigma_1 \dots d\sigma_k$ and $d\mathbf{L}_k = dL_1 \dots dL_k$. For an analytic function $A(\mathbf{z})$ on D^n and $i_1, \dots, i_k \subset \{i_1, \dots, n\}$, we will denote $\partial_{i_1} \dots \partial_{i_n} A$ by $A_{i_1,\dots,i_n}$.

Lemma 5. *Let $F \in H^\infty(D^n)$ with $\|F\|_\infty \le 1$. Fix $q \in H^2(D^n)$ for $1 \le j \le n$. Then there exists $C_0 < \infty$ such that*

$$\int_{D^j} \|F_{1,\dots,j}\|^2 \, \|q\|^2 \, d\mathbf{L}_j \le C_0^2 \int_{T^j} \|q\|^2 \, d\boldsymbol{\sigma}_j.$$

Proof. By induction, the case $j = 1$ is just the Paley-Littlewood estimate, Lemma 2. For $1 < j \le n$,

$$\begin{aligned}\int_{D^j} \|F_{1,\dots,j}\|_2^2 \, \|q\|_2^2 \, d\mathbf{L}_j &\le 2 \int_D \int_{D^{j-1}} \|F_{1,\dots,j-1}\|_2^2 \, \|q_j\|_2^2 \, d\mathbf{L}_{j-1} dL_j \\ &\quad + 2 \int_{D^{j-1}} \int_D \|(F_{1,\dots,j-1}\, q)_j\|^2 \, dL_j \, d\mathbf{L}_{j-1} \\ &\le 4\, C_{j-1}^2 \int_{T^j} \|q\|^2 \, d\boldsymbol{\sigma}_n,\end{aligned}$$

where C_{j-1} is the constant for j terms. Let $C_0 = 2\, C_{j-1}$. □

Lemma 6. *Let $\{1, \dots, n\} = I_1 \dot\cup I_2 \dot\cup \dots \dot\cup I_k \dot\cup J \dot\cup K$. Then there exists $C_1 < \infty$ so that for $h, k \in H^2(D^n)$ we have*

$$\int_{D^n} \|F_1\|_2 \dots \|F_n\|_2 \, \|F_{I_1}\|_2 \dots \|F_{I_k}\|_2 \, \|h_J\|_2 \, \|k_K\|_2 d\mathbf{L}_n \le C_1 \|h\|_2 \, \|k\|_2.$$

Proof. Let $\{1,\dots,n\} = J \dot\cup J' = K \dot\cup K'$.

$$\int_{D^n} \|F_1\|_2 \dots \|F_n\|_2 \, \|F_{I_1}\|_2 \dots \|F_{I_k}\|_2 \, \|h_J\|_2 \, \|k_K\|_2 \, d\mathbf{L}_n$$

$$\le \left(\int_{D^n} \prod_{j\notin J} \|F_j\|^2 \|h_J\|_2^2 d\mathbf{L}_n \right)^{\frac12} \left(\int_{D^n} \prod_{j\in J} \|F_j\|^2 \prod_{j=1}^{k} \|F_{I_j}\|_2^2 \|k_K\|^2 d\mathbf{L}_n \right)^{\frac12}$$

$$\le \left(C_0^{2|J'|} \int_{T^{J'}} \int_{D^J} \|h_J\|_2^2 d\mathbf{L}_J d\boldsymbol{\sigma}_{J'} \right)^{\frac12} \left(C_0^{2|J|} \int_{T^J} \int_{D^{J'}} \prod_{j=1}^{k} \|F_{I_j}\|_2^2 \|k_K\|^2 d\mathbf{L}_{J'} d\boldsymbol{\sigma}_J \right)^{\frac12}$$

$$\le \left(C_0^{2|J'|} \int_{T^{J'}} \int_{D^J} \|h_J\|_2^2 d\mathbf{L}_J d\boldsymbol{\sigma}_{J'} \right)^{\frac12} \left(C_0^{2|J|} \int_{T^J} \int_{D^{J'}} \prod_{j=1}^{k} \|F_{I_j}\|_2^2 \|k_K\|^2 d\mathbf{L}_{J'} d\boldsymbol{\sigma}_J \right)^{\frac12}$$

$$\le \left(C_0^{2|J'|} \int_{T^n} \|h\|_2^2 d\boldsymbol{\sigma}_n \right)^{\frac12} \left(C_0^{2|K'|} \int_{T^{K'}} \int_{D^K} \|k_K\|^2 d\mathbf{L}_K \, d\boldsymbol{\sigma}_{K'} \right)^{\frac12}$$

$$\le C_0^{|J'|+|K'|} \, |\!|h|\!|_2$$

by applying Lemmas 2 and 5. □

The next lemma is due to Chang [6].

Lemma 7. *Let $\{1,\dots,n\} = I_1 \dot\cup I_2 \dot\cup \dots \dot\cup I_k \dot\cup N$. Then there exists $C_2 < \infty$ so that for $H \in \mathcal{H}^1(D^n)$ we have*

$$\int_{D^n} \|F_1\|_2 \dots \|F_n\|_2 \, \|F_{I_1}\|_2 \dots \|F_{I_k}\|_2 \, \|H_N\|_2 \, d\mathbf{L}_n \le C_2 \, |\!|H|\!|_1. \tag{1}$$

Proof. By Lemma 4, we need only prove (1) for scalar $H \in H^1(D^n)$. In this case, by the Lacey-Terwilleger weak factorization result [12], there exists an $M < \infty$, so that for each $H \in H^1(D^n)$ there exist $\{k_j\}_{j=1}^\infty$ and $\{l_j\}_{j=1}^\infty$ contained in $H^2(D^n)$, satisfying $H = \sum k_j l_j$ in $H^1(D^n)$ (and pointwise in D^n); moreover,

$$\|H\|_1 \le \sum_{j=1}^{\infty} \|k_j\|_2 \|l_j\|_2 \le M \, \|H\|_1.$$

Now we apply Lemma 6 to $\|(k_j)_J\|_2 \|(l_j)_{N-J}\|_2$ for each j and each subset $J \subset N$; then add to get (1), where $C_2 = 2^N C_1 M$. □

The differential operators we need can be written in the following way. Fix n, the dimension of D^n. We will consider all operators as matrices of differential operators acting on vectors with entries in $C^\infty(\overline{D}^n)$. Let $\overline\partial_j = \frac12(\partial_{x_j} + i\,\partial_{y_j})$ for $1 \le j \le n$. Then

$$D_0(n) := I,$$

$$D_1(n) := \begin{pmatrix} \overline{\partial}_1 \\ \vdots \\ \overline{\partial}_n \end{pmatrix},$$
$$D_n(n) := (\overline{\partial}_n, -\overline{\partial}_{n-1}, \ldots, (-1)^{n-1}\overline{\partial}_1),$$

and, inductively, for $1 < k < n$,

$$D_k(n) := \left[\begin{array}{c|c} D^+_{k-1}(n-1) & (-1)^{k-1}\overline{\partial}_1 \otimes I \\ \hline 0 & D^+_k(n-1) \end{array}\right].$$

Here $D^+_k(n-1)$ is the operator $D_k(n-1)$, but with the $n-1$ terms numbered $2, \ldots, n$ instead of $1, 2, \ldots, n-1$. Note that $D_k(n)$ is an $\binom{n}{k} \times \binom{n}{k-1}$ matrix for $1 \le k \le n$.

For example,

$$D_2(3) = \left[\begin{array}{c|c} D^+_1(2) & (-1)\overline{\partial}_1 \otimes I \\ \hline 0 & D^+_2(2) \end{array}\right] = \left[\begin{array}{c|cc} \overline{\partial}_2 & -\overline{\partial}_1 & 0 \\ \overline{\partial}_3 & 0 & -\overline{\partial}_1 \\ \hline 0 & \overline{\partial}_3 & -\overline{\partial}_2 \end{array}\right].$$

We wish to find integral operators $K_l(n)$, so that

$$D_l(n)\, K_l(n) + K_{l+1}(n)\, D_{l+1}(n) = I \text{ for } 0 \le l \le n.$$

For the polydisk, we can achieve this by setting

$$K_0(n) = P_1 \ldots P_n$$
$$K_1(n) = (P_2 \ldots P_n \Lambda_1, P_3 \ldots P_n \Lambda_2, \ldots, \Lambda_n),$$
$$K_n(n) = \begin{pmatrix} \Lambda_n \\ 0 \\ \vdots \\ 0 \end{pmatrix},\ K_m(n) = 0,\ m > n,$$

and, inductively, for $1 < k < n$,

$$K_l(n) = \left[\begin{array}{c|c} K^+_{l-1}(n-1) & 0 \\ \hline 0 & K^+_l(n-1) \end{array}\right].$$

Note that $K_l(n)$ is an $\binom{n}{l-1} \times \binom{n}{l}$ matrix for $1 \le l \le n$.

For example,

$$K_2(3) = \left[\begin{array}{c|c} K^+_1(2) & 0 \\ \hline 0 & K^+_2(2) \end{array}\right] = \left[\begin{array}{cc|c} P_3\Lambda_2 & \Lambda_3 & 0 \\ \hline 0 & 0 & \Lambda_3 \\ 0 & 0 & 0 \end{array}\right].$$

Lemma 8.

(a) $D_{l+1}(n)\,D_l(n) = 0 \quad$ *for* $l = 0, 1, \ldots, n$

(b) $K_l(n)\,K_{l+1}(n) = 0 \quad$ *for* $l = 0, 1, \ldots, n$

(c) $D_l(n)\,K_l(n) + K_{l+1}(n)\,D_{l+1}(n) = I \quad$ *for* $l = 0, 1, \ldots, n.$

[Note that in (c) "I" denotes an $\binom{n}{l} \times \binom{n}{l}$ identity matrix acting on $\binom{n}{l}$ copies of $C^\infty(\overline{D}^n)$ and similarly for the "0" in (a) and (b).]

Proof. The proof is by induction on $(n + l)$. Now $n = 1, 2, \ldots$ and $0 \le l \le n$. For $n + l = 1$, we have $n = 1$ and $l = 0$. In this case we have

$$D_1(1)D_0(1) = \overline{\partial}_1(I) = 0;$$
$$K_0(1)K_1(1) = P_1\Lambda_1 = 0;$$
$$D_0(1)K_0(1) + K_1(1)D_1(1) = I - P_1 + \Lambda_1\overline{\partial}_1 = P_1 + P_1^\perp = I.$$

Assume that (a), (b), and (c) hold for $(n + l - 1) = j < 2n$, where $0 \le l \le n - 1$.

We show that (a), (b), and (c) hold for $n + l$. If $l = 0$, then

$$D_1(n)D_0(n) = \begin{pmatrix} \overline{\partial}_1 \\ \vdots \\ \overline{\partial}_n \end{pmatrix} I = \begin{pmatrix} 0 \\ \vdots \\ 0 \end{pmatrix} = 0$$

$$K_0(n)K_1(n) = P_1 \ldots P_n(P_2 \ldots P_n\Lambda_1, P_3 \ldots P_n\Lambda_2, \ldots, \Lambda_n) = 0$$

since $P_j\Lambda_j = 0$ for all $1 \le j \le n$.

$$\begin{aligned} D_0(n)K_0(n) + K_1(n)D_1(n) &= I \cdot P_1 \ldots P_n + \sum_{j=1}^{n-1} P_{j+1} \ldots P_n\Lambda_j\overline{\partial}_j + \Lambda_n\overline{\partial}_n \\ &= P_1 \ldots P_n + \sum_{j=1}^{n-1} P_j^\perp P_{j+1} \ldots P_n + P_n^\perp \\ &= P_1 \ldots P_n + P_1^\perp P_2 \ldots P_n \\ &\quad + \sum_{j=2}^{n-1} P_j^\perp P_{j+1} \ldots P_n + P_n^\perp \\ &= P_2 \ldots P_n + \sum_{j=2}^{n-1} P_j^\perp P_{j+1} \ldots P_n + P_n^\perp \\ &= \cdots = P_n + P_n^\perp = I. \end{aligned}$$

Now we may assume that $1 \le l \le n$. Then

$$D_{l+1}(n)D_l(n) = \begin{bmatrix} D_l^+(n-1) & (-1)^l\overline{\partial}_1 \otimes I \\ 0 & D_{l+1}^+(n-1) \end{bmatrix} \begin{bmatrix} D_{l-1}^+(n-1) & (-1)^{l-1}\overline{\partial}_1 \otimes I \\ 0 & D_l^+(n-1) \end{bmatrix}$$

$$= \begin{bmatrix} D_l^+(n-1)D_{l-1}^+(n-1) & (-1)^{l-1}D_l^+(n-1)\overline{\partial}_1 + (-1)^l\overline{\partial}_1 D_l^+(n-1) \\ 0 & D_{l+1}^+(n-1)D_l^+(n-1) \end{bmatrix}$$
$$= \begin{bmatrix} 0 & 0 \\ 0 & 0 \end{bmatrix},$$

since the diagonal terms vanish by induction and $\overline{\partial}_1 \otimes I$ commutes with $D_l^+(n-1)$, so the off-diagonal term is 0. Similarly, $K_l(n)K_{l+1}(n) = 0$.

Finally,

$$\begin{aligned} &D_l(n)K_l(n) + K_{l+1}(n)D_{l+1}(n) \\ &= \begin{bmatrix} D_{l-1}^+(n-1) & (-1)^{l-1}\overline{\partial}_1 \otimes I \\ 0 & D_l^+(n-1) \end{bmatrix} \begin{bmatrix} K_{l-1}^+(n-1) & 0 \\ 0 & K_l^+(n-1) \end{bmatrix} \\ &\quad + \begin{bmatrix} K_l^+(n-1) & 0 \\ 0 & K_{l+1}^+(n-1) \end{bmatrix} \begin{bmatrix} D_l^+(n-1) & (-1)^l\overline{\partial}_1 \otimes I \\ 0 & D_{l+1}^+(n-1) \end{bmatrix} \\ &= \begin{bmatrix} I & 0 \\ 0 & I \end{bmatrix}. \end{aligned}$$

This follows since, by induction, the diagonal terms equal I. Also, $K_l^+(n-1)$ only involves P_j and Λ_j for $j = 2, \ldots, n$, so $\overline{\partial}_1 \leftrightarrow K_l^+(n-1)$. Thus the off-diagonal term is 0 and this completes the induction step. □

Suppose that we have formally:

$$\begin{aligned} &Q_0E_0 = I && \text{and} && D_1K_1 + K_2D_2 = I \\ &E_0Q_0 + Q_1E_1 = I && && \vdots \\ &\quad\vdots && && D_{n-1}K_{n-1} + K_nD_n = I \\ &E_{n-1}Q_{n-1} + Q_nE_n = I && && D_nK_n = I (D_{n+1} = 0) \\ &\quad\vdots \end{aligned}$$

$$\text{and} \quad Q_jQ_{j+1} = E_{j+1}E_j = D_jD_{j+1} = Q_{j+1}Q_j = 0.$$

We form words by concatenation assuming, in addition, that

$$D_iQ_j = Q_jD_i \qquad \text{for all } i = 1, \ldots, n,\ j = 0, 1, \ldots$$

$$\text{and} \quad D_1I = 0.$$

Then we have the following:

Lemma 9. *Let*

$$u = E_0 + \sum_{j=1}^{\infty} (-1)^j Q_1 K_1 \dots Q_j K_j E_j D_j \dots E_1 D_1 E_0.$$

Then

$$D_1 u = 0.$$

Proof. Let

$$\begin{aligned} u_n &:= E_n D_n \dots E_1 D_1 E_0 \\ u_{n-1} &:= E_{n-1} D_{n-1} \dots E_1 D_1 E_0 - Q_n K_n u_n \\ &\vdots \\ u_1 &:= E_1 D_1 E_0 - Q_2 K_2 u_2 \\ u_0 &:= E_0 - Q_1 K_1 u_1. \end{aligned}$$

Then $u = u_0$; so we must show that $D_1 u_0 = 0$. Since $D_{n+1} \equiv 0$, $D_{n+1} u_n = 0$. Assume $D_{j+2} u_{j+1} = 0$. For $0 \le j \le n-1$, we show that $D_{j+1} u_j = 0$ and this will complete the proof.

$$\begin{aligned} D_{j+1} u_j &= D_{j+1} \left[E_j D_j \dots E_1 D_1 E_0 - Q_{j+1} K_{j+1} u_{j+1} \right] \\ &= D_{j+1} E_j D_j \dots E_1 D_1 E_0 - Q_{j+1} (D_{j+1} K_{j+1}) u_{j+1} \\ &= D_{j+1} E_j D_j \dots E_1 D_1 E_0 - Q_{j+1} [I - K_{j+2} D_{j+2}] u_{j+1} \\ &= D_{j+1} E_j D_j \dots E_1 D_1 E_0 - Q_{j+1} u_{j+1} \\ &= D_{j+1} E_j D_j \dots E_1 D_1 E_0 - Q_{j+1} [E_{j+1} D_{j+1} \dots E_1 D_1 E_0 - Q_{j+2} K_{j+2} u_{j+2}] \\ &= (I - Q_{j+1} E_{j+1}) D_{j+1} E_j D_j \dots E_1 D_1 E_0 \\ &= (E_j Q_j) D_{j+1} E_j D_j \dots E_1 D_1 E_0 \\ &= E_j D_{j+1} (Q_j E_j) D_j E_{j-1} D_{j-1} \dots E_1 D_1 E_0 \\ &= E_j D_{j+1} (I - E_{j-1} Q_{j-1}) D_j E_{j-1} D_{j-1} \dots E_1 D_1 E_0 \\ &= (-1) E_j D_{j+1} E_{j-1} D_j (Q_{j-1} E_{j-1}) D_{j-1} \dots E_1 D_1 E_0 \\ &\vdots \\ &= (-1)^j E_j D_{j+1} E_{j-1} D_j \dots D_1 (Q_0 E_0) \\ &= 0 \quad \text{since } Q_0 E_0 = I \text{ and } D_1(I) = 0. \end{aligned}$$

□

Up to an analytic perturbation, which will be specified later, we can now write our solution. Let

$$E_0(\mathbf{z}) = \frac{F(\mathbf{z})^*}{F(\mathbf{z}) F(\mathbf{z})^*} \text{ and } E_l(\mathbf{z}) = \frac{Q_l^*(\mathbf{z})}{F(\mathbf{z}) F(\mathbf{z})^*} \text{ for } l = 1, \dots, n.$$

Suppressing the $\mathbf{z} \in D^n$, we set

$$u_0 = E_0 + \sum_{j=1}^{n} (-1)^j Q_1 K_1 \dots Q_j K_j E_j D_j \dots E_1 D_1 E_0.$$

Clearly $F\,u_0 \equiv 1$ in $\overline{D}^n$, since $F\,Q_1 \equiv 0$ there. By Lemma 9, $D_1 u_0 \equiv 0$ in $\overline{D}^n$. It only remains to estimate the size of the solution u_0.

Lemma 10. (a) *There exist operators* $\{B_\pi\}_{\pi \in \Pi(j)}$, *so that*

$$Q_1 K_1 \dots Q_j K_j = [B_\pi \Lambda_\pi]_{\pi \in \Pi(j)}.$$

This is a $1 \times \binom{n}{j}$ *row vector with operators as entries. For* $\pi = (i_1, \dots, i_j)$, Λ_π *denotes* $\Lambda_{i_1,\dots,i_j}$. *Here* B_π *is a finite product of operators belonging to* $\{Q_l, P_k, P_k^\perp : 1 \le l \le j,\ 1 \le k \le n\}$.

(b) $E_j D_j \dots E_1 D_1 E_0 = [j! \dfrac{Q_j^*}{(FF^*)^{j+1}} \overline{\partial}_{i_j} Q_{j-1}^* \dots \overline{\partial}_{i_1} F^*]_{\substack{\pi \in \Pi(j) \\ \pi = (i_1,\dots,i_j)}}$. *This is an* $\binom{n}{j} \times 1$ *column vector, whose entries are vectors of functions.*

(c) *Each* B_π *appearing in* (a) *can be written as a finite sum of terms involving no repetitions of the projections* $\{P_j\}_{j=1}^n$. *Thus each term involves at most* n *different projections.*

Proof. Consider (a). For $n = 1$ analytic variable,

$$Q_1 K_1(1) = Q_1 \Lambda_1,$$

so we are done. For $n > 1$ and $1 \le j \le n$,

$$\begin{aligned} Q_1 K_1 \dots Q_j K_j &= [Q_1 P_2 \dots P_n \Lambda_1, K_1^+(n-1)]\, Q_2 \begin{bmatrix} K_1^+(n-1) & 0 \\ 0 & K_2^+(n-1) \end{bmatrix} \\ &\quad \dots Q_j \begin{bmatrix} K_{j-1}^+(n-1) & 0 \\ 0 & K_{j-1}^+(n-1) \end{bmatrix} \\ &= [Q_1 P_2 \dots P_n \Lambda_1 (Q_2 K_1^+(n-1) \dots Q_j K_{j-1}^+(n-1), \\ &\quad Q_1 K_1^+(n-1) \dots Q_j K_j^+(n-1)]. \end{aligned}$$

By induction on the number of analytic variables,

$$Q_1 K_1^+(n-1) \dots Q_j K_j^+(n-1) = [B_\pi \Lambda_\pi]_{\substack{\pi \in \Pi(j) \\ 1 \notin \pi}}$$

$$\text{and } \ Q_2 K_1^+(n-1) \dots Q_j K_{j-1}^+(n-1) = [C_\sigma \Lambda_\sigma]_{\substack{\sigma \in \Pi(j-1) \\ 1 \notin \sigma}}.$$

Here B_π and C_σ denote operators formed from finite products of Q_l's, P_k's, and $P_k^\perp$'s for $1 \le l \le j$ and $2 \le k \le n$.

But

$$\Lambda_1 C_\sigma \Lambda_\sigma = B_{1,\sigma} \Lambda_1 \Lambda_\sigma = B_\pi \Lambda_\pi,$$

where $\pi \in \Pi(j)$ and $1 \in \pi$. To see this, first notice that neither expression $Q_1K_1^+(n-1)\dots Q_jK_j^+(n-1)$ or $Q_2K_1^+(n-1)\dots Q_jK_{j-1}^+(n-1)$ involves the projection P_1 and $\Lambda_1 P_k = P_k\Lambda_1$ for $2 \le k \le n$. Also, we have

$$X\,\Lambda_1 Q_l Y = X\,P_1^\perp Q_l \Lambda_1 Y. \tag{2}$$

Continuing this procedure with Λ_1 commuting across the P_k's and using (2) to move Λ_1 to the right of the Q_l's, (a) follows.

As for (b), first note that

$$\frac{Q_j^*}{FF^*}D_j\frac{Q_{j-1}}{FF^*}\dots D_1\frac{F^*}{FF^*} = \frac{Q_j^*}{(FF^*)^{j+1}}D_jQ_{j-1}^*\dots D_2Q_1^*D_1F^*.$$

This follows since

$$\begin{aligned} X\frac{Q_l^*}{FF^*}D_l\frac{Q_{l-1}^*}{FF^*}Y &= X\frac{Q_l^*}{FF^*}[D_l(\frac{1}{FF^*})Q_{l-1}^*Y + \frac{1}{FF^*}D_lQ_{l-1}^*Y] \\ &= X\frac{Q_l^*}{(FF^*)^2}D_lQ_{l-1}^*Y, \text{ using } Q_l^*Q_{l-1}^* = 0. \end{aligned}$$

Now $D_jQ_{j-1}^*\dots D_1F^* = (D_jQ_{j-1}^*)\dots(D_1F^*)$ is formally (as in the Appendix) the same as

$$\begin{aligned} &\begin{pmatrix} \overline{\partial}_1 Q_{j-1}^* \\ \vdots \\ \overline{\partial}_n Q_{j-1}^* \end{pmatrix} \wedge\dots\wedge \begin{pmatrix} \overline{\partial}_1 F^* \\ \vdots \\ \overline{\partial}_n F^* \end{pmatrix} \qquad (3) \\ &= \sum_{\substack{\pi\in\Pi(j) \\ \pi=(i_1,\dots,i_j)}} \left[\sum_{\alpha\in P(\pi)} (-1)^{sgn\,\alpha}\overline{\partial}_{\alpha(i_j)}Q_{j-1}^*\dots\overline{\partial}_{\alpha(i_1)}F^*\right] e_\pi \\ &= j!\sum_{\pi\in\Pi(j)} \overline{\partial}_{i_j}Q_{j-1}^*\dots\overline{\partial}_{i_1}F^*\,e_\pi. \end{aligned}$$

This last equality comes from the fact that

$$\begin{aligned} \overline{\partial}_{\alpha(i_j)}Q_{j-1}^*\dots\overline{\partial}_{\alpha(i_1)}F^* &= (\overline{\partial}_{\alpha(i_j)}F^*)\wedge\dots\wedge(\overline{\partial}_{\alpha(i_1)}F^*) \\ &= (-1)^{sgn\,\alpha}\,\overline{\partial}_{i_j}F^*\wedge\dots\wedge\overline{\partial}_{i_1}F^* \\ &= (-1)^{sgn\,\alpha}\,\overline{\partial}_{i_j}Q_{j-1}^*\dots\overline{\partial}_{i_1}F^*. \end{aligned}$$

Thus

$$\frac{Q_j^*}{FF^*}D_j\frac{Q_{j-1}^*}{FF^*}\dots D_1\frac{F^*}{FF^*} = [j!\frac{Q_j^*}{(FF^*)^{j+1}}\overline{\partial}_{i_j}Q_{j-1}^*\dots\overline{\partial}_{i_1}F^*]_{\substack{\pi\in\Pi(n) \\ \pi=(i_1,\dots,i_n)}}$$

as an $\binom{n}{j}\times 1$ vector.

Consider elements from B_π, which from (a) involves finite products using only elements from $\{Q_l\}_{l=1}^j$, $\{P_k\}_{k=1}^n$, and $\{P_k^\perp\}_{k=1}^n$. Replace any $P_k^\perp$'s by $I-P_k$ and write B_π as a finite sum of terms containing no $P_k^\perp$'s. Consider a term in B_π of the form $X\,P_kS\,P_kY$. Then P_kY is analytic in the kth variable. Since S

involves Q_l's which are analytic and P_l's, we have that $S\,P_kY$ is analytic in the kth variable. Thus $X\,P_kS\,P_kY = X\,S\,P_kY$. Using this procedure, we may assume that in each term of B_π at most one occurrence of P_k for $1 \le k \le n$ appears. Thus (c) follows. □

Lemma 11. *For each $Q_1K_1 \dots Q_jK_j$ from Lemma 10, there is an operator A_j such that*

(1) $F\,A_j \equiv 0$ *in* $\overline{D}^n$,
(2) A_j *is analytic in* $\overline{D}^n$,
(3) *terms of* $Q_1K_1 \dots Q_jK_j + A_j$ *involve at most* $n-1$ *of the projections* $\{P_j\}_{j=1}^n$.

Proof. By construction,

$$Q_1K_1 \dots Q_jK_j = [Q_1P_2 \dots P_nP_1^\perp(Q_2K_1^+(n-1) \dots Q_jK_{j-1}^+(n-1))\Lambda_1, \\ Q_1K_1^+(n-1) \dots Q_jK_j^+(n-1)].$$

Let

$$A_j = [Q_1P_2 \dots P_nP_1(Q_2K_1^+(n-1) \dots Q_jK_{j-1}^+(n-1))\Lambda_1, 0].$$

Then (1) and (2) are clear.

$$Q_1K_1 \dots Q_jK_j + A_j = [Q_1P_2 \dots P_n(Q_2K_1^+(n-1) \dots Q_jK_{j-1}^+(n-1))\Lambda_1, \\ Q_1K_1^+(n-1) \dots Q_jK_j^+(n-1)].$$

By Lemma 10,

$$Q_1K_1^+(n-1) \dots Q_jK_{j-1}^+(n-1) = \{B_\pi\Lambda_\pi\}_{\substack{\pi\in\Pi(j)\\ 1\notin\pi}}$$

and

$$Q_2K_1^+(n-1) \dots Q_jK_{j-1}^+(n-1)\Lambda_1 = \{C_\sigma\Lambda_\sigma\Lambda_1\}_{\substack{\sigma\in\Pi(j-1)\\ 1\notin\sigma}}.$$

Since B_π and C_σ do not contain the projection P_1 and thus $Q_1P_2 \dots P_nC_\sigma$ does not contain P_1 either, we see that the terms of $Q_1K_1 \dots Q_jK_j + A_j$ do not involve P_1 and by Lemma 10 can thus be written as sums of terms involving at most $(n-1)$ projections. □

We are now ready to complete our estimates.

Proof of Theorem A. Suppose that $F \in \mathcal{H}^\infty(D^n)$ and $0 < \epsilon^2 \le F(\mathbf{z})F(\mathbf{z})^* \le 1$ for all $\mathbf{z} \in D^n$. We lose no generality (by considering $F_r(\mathbf{z}) = F(r\mathbf{z})$) in assuming that $F \in \mathcal{H}^\infty(D^n_{\frac{1}{r}})$, for some $\frac{1}{r} > 1$. Then we must show that our estimates are independent of r and apply a compactness argument to complete the proof.

Let

$$u_r = u_0 + \sum_{j=1}^{n}(-1)^j A_jE_jD_j \dots E_1D_1E_0.$$

We use the subscript "r" to remind us that all the terms are defined using F_r in place of F. Suppose we show that $|u_r|_p \leq \frac{C_0}{\epsilon^{3n+1}} p^{2n-1}$ with C_0 independent of r and $p \geq 2$. This suffices for u_r to belong to $\exp(L^{\frac{1}{2n-1}})$, since u_r is analytic in $\overline{D}^n$ by Lemma 9 and

$$\begin{aligned}
\int_{T^n} e^{(\frac{\|u_r\|_2}{\lambda})^{\frac{1}{2n-1}}} d\sigma_1 \dots d\sigma_n &= \sum_{k=0}^{\infty} \frac{1}{k!\,\lambda^{\frac{k}{2n-1}}} |u_r|_{\frac{k}{2n-1}}^{\frac{k}{2n-1}} \\
&\leq \sum_{k=0}^{4n} \frac{1}{k!\,\lambda^{\frac{k}{2n-1}}} |u_r|_2^{\frac{k}{2n-1}} \\
&\quad + \sum_{k=4n+1}^{\infty} \frac{1}{k!\,\lambda^{\frac{k}{2n-1}}} \left(\frac{C_0}{\epsilon^{3n+1}} (\frac{k}{2n-1})^{2n-1} \right)^{\frac{k}{2n-1}} \\
&\leq \sum_{k=0}^{4n} \frac{1}{k!\,\lambda^{\frac{k}{2n-1}}} \left(\frac{C_0}{\epsilon^7} 2^3 \right)^{\frac{k}{2n-1}} \\
&\quad + \sum_{k=4n+1}^{\infty} \frac{k^k}{k!} \left[\frac{C_0}{\epsilon^{3n+1}(2n-1)^{2n-1}\lambda} \right]^{\frac{k}{2n-1}}. \qquad (4)
\end{aligned}$$

Thus (3) is finite for $\lambda \approx \frac{1}{\epsilon^{3n+1}}$. So

$$|u_r|_{e,2n-1} \leq \frac{C_0}{\epsilon^{3n+1}}.$$

This estimate is independent of r, so a compactness argument gives us our analytic u with $Fu \equiv 1$ in D^n, $u \in \exp(L^{\frac{1}{2n-1}})$ and $|u|_{e,2n-1} \leq \frac{C_0}{\epsilon^{3n+1}}$.

It remains to show that if

$$u_r = u_0 + \sum_{j=1}^{n} (-1)^j A_j E_j D_j \dots E_1 D_1 E_0,$$

then there exists $C_0 < \infty$, independent of r so that

$$|u_r|_p \leq \frac{C_0}{\epsilon^{3n+1}} p^{2n-1},\ p \geq 2.$$

Fix $p \geq 2$. Now by Lemma 10

$$u_r = \sum_{j=1}^{n} (-1)^j \sum_{\substack{\pi \in \Pi(j) \\ \pi=(i_1,\dots,i_j)}} j!\, B_\pi \left[\frac{Q_j^*}{(FF^*)^{j+1}} \overline{\partial}_{i_j} Q_{j-1}^* \dots \overline{\partial}_{i_1} F^* \right]^{\Lambda_\pi},$$

where $\Lambda_\pi = \Lambda_{i_1,\dots,i_j}$. By Lemma 11, B_π is a finite sum (of at most $n!$ terms) of finite products of contractions Q_l and of at most $(n-1)$ projections among $\{P_k\}_{k=1}^n$. Now as an operator on $\mathcal{L}^p(T^n)$, $p \geq 2$, P_k has norm $\|P_k\| \leq C_0 p$. [See Garnett [10] for this fact on $L^p(T)$.] Thus

$$\|B_\pi\|_{B(\mathcal{L}^p(T^n))} \leq C_0 p^{n-1},$$

where C_0 is independent of r and p (but depends on n).

Estimating, we get

$$|u_r|_p \leq \left(\sum_{j=1}^{n} \sum_{\substack{\pi \in \Pi(j) \\ \pi=(i_1,\ldots,i_j)}} j! \, \| [\frac{Q_j^*}{(FF^*)^{j+1}} \overline{\partial}_{i_j} Q_{j-1}^* \ldots \overline{\partial}_{i_1} F^*]^{\Lambda_\pi} \|_{\mathcal{L}^p(T^n)} \right) C_0 p^{n-1}.$$

We estimate the worst term in growth rate; the others follow more easily.

Let

$$l = \left[\frac{Q_n^*}{(FF^*)^{n+1}} \overline{\partial}_n Q_{n-1}^* \ldots \overline{\partial}_1 F^* \right].$$

We must show that

$$|l^{\Lambda_1,\ldots,n}|_p \leq \frac{C_0}{\epsilon^{3n+1}} p^n$$

to complete the proof.

By duality, since $\mathcal{L}^p(T^n)^* \approx \mathcal{L}^q(T^n)$ for $1 \leq p < \infty$ (see Edwards [9], p. 607) and $l^{\Lambda_1,\ldots,n} = P_1^\perp \ldots P_n^\perp l^{\Lambda_1,\ldots,n}$, we have

$$\begin{aligned} |l^{\Lambda_1,\ldots,n}|_p &= \sup_{\substack{k \in \mathcal{L}^q(T^n) \\ |k|_q \leq 1}} |\langle l^{\Lambda_1,\ldots,n}, k \rangle| \\ &= \sup_{\substack{k \in \mathcal{L}^q(T^n) \\ |k|_q \leq 1}} |\langle l^{\Lambda_1,\ldots,n}, P_1^\perp \ldots P_n^\perp k \rangle| \\ &\leq \sup_{\substack{H_0 \in z_1 \ldots z_n \mathcal{H}^q(T^n) \\ |H_0|_q \leq C_0 p^n}} |\langle l^{\Lambda_1,\ldots,n}, \overline{H}_0 \rangle| \\ &\leq \sup_{\substack{H_0 \in z_1 \ldots z_n \mathcal{H}^1(T^n) \\ |H_0|_1 \leq C_0 p^n}} |\langle l^{\Lambda_1,\ldots,n}, \overline{H}_0 \rangle|. \end{aligned}$$

Applying Lemma 1 n times, we must estimate

$$\begin{aligned} \langle l^{\Lambda_1,\ldots,n}, \overline{H}_0 \rangle &= \int_{D_n} \partial_1 \ldots \partial_n \langle l, \overline{H}_0 \rangle \, dL_1 \ldots dL_n \\ &= \int_{D_n} \partial_1 \ldots \partial_n \langle \frac{Q_n^*}{(FF^*)^{n+1}} \overline{\partial}_n Q_{n-1}^* \ldots \overline{\partial}_1 F^*, \overline{H}_0 \rangle \, d\mathbf{L}_n \end{aligned} \tag{5}$$

for $H_0 \in z_1 \ldots z_n \mathcal{H}^1(T^n)$ and $|H_0|_1 \leq C_0 p^n$.

Since $Q_n^* \overline{\partial}_n Q_{n-1}^* \ldots \overline{\partial}_1 F^*$ is coanalytic in D^n, all $\partial_1, \ldots, \partial_n$ derivatives in (4) apply to either $\frac{1}{(FF^*)^{n+1}}$ or to $\overline{H}_0$. Let $\{1, \ldots, n\} = I \dot{\cup} J$. If $J = \{j_1, \ldots, j_p\}$, let ∂_J denote $\partial_{j_1} \ldots \partial_{j_p}$ and similarly for I.

Then we must estimate sums of terms of the form

$$\left| \int_{D^n} \langle \partial_I (\frac{1}{(FF^*)^{n+1}}) Q_n^* \overline{\partial}_n Q_{n-1}^* \ldots \overline{\partial}_1 F^*, \overline{\partial_J H_0} \rangle \, d\mathbf{L}_n \right|$$

$$\leq \int_{D^n} \|\partial_I(\frac{1}{(FF^*)^{n+1}})\|_{op}\|Q_n^*\|_{op}\|\overline{\partial}_n Q_{n-1}^*\|_{op}\dots\|\overline{\partial}_1 F^*\|_2\|\partial_J H_0\|_2\, d\mathbf{L}_n.$$

Recall that $\|\overline{\partial}_k Q_j^*\|_{op} \leq \|\overline{\partial}_k F^*\|_2$, $k = 0,\dots,n$ and $j = 1,\dots,n$. The result now follows from Lemma 7. This completes the proof of Theorem A. □

The above argument shows that $l_r^{\Lambda_1,\dots,n} \in BMO(T^n) \cap \overline{\mathcal{H}^2(D^n)}$ and with BMO norm independent of $0 \leq r < 1$. Thus, in the case that $n = 1$, $\widehat{l_r^{-1}} \in BMO\, A$ and $\sup_{0\leq r<1} \|\widehat{l_r^{-1}}\|_{BMO} \triangleq C_0 < \infty$. So there exist $\varphi_{(r)} \in L^\infty(T)$ and $h_{(r)} \in H_0^2(T)$ so that $P_{H^2}^\perp(\varphi_{(r)}) = \widehat{l_r^{-1}}$, $\varphi_{(r)} = \widehat{l_r^{-1}} + h_{(r)}$ and $\sup_{0\leq r<1} \|\varphi_{(r)}\|_\infty = C_0 < \infty$. Then $u_r = \frac{F_r^*}{F_r F_r^*} - Q_r[\varphi_{(r)}]$ gives a corona solution with $\|u_r\|_\infty \leq \frac{1}{\epsilon} + C_0$ for all $0 \leq r < 1$.

We will outline the modifications of Theorem A necessary to establish Theorem B. To show that $T_F \in B(\mathcal{H}^p(T^n))$ is onto, fix any $G \in \mathcal{H}^p(T^n)$. Then for $0 < r < 1$,

$$u_{G_r} = \sum_{j=1}^{n} (-1)^j \sum_{\substack{\pi\in\Pi(j)\\ \pi=(i_1,\dots,i_j)}} j!\, B_\pi[\frac{Q_j^*}{(FF^*)^{j+1}}\overline{\partial}_{j_1} Q_{j-1}^* \dots \overline{\partial}_{i_1} F^* G_r]^{\Lambda_\pi}$$

satisfies

$$T_{F_r}\, u_{G_r} = G_r \in \mathcal{H}^p(T^n)$$
$$\text{and} \qquad |u_{G_r}|_p \leq C_0\, |G_r|_p, \tag{6}$$

where C_0 is independent of r.

To verify (5), we estimate $|u_{G_r}|_p$ as we did $|u_r|_p$ in Theorem A. The proof is completed with a compactness argument. More details for the case $n = 2$ and the compactness argument can be found in Trent [19].

Appendix

We finish this paper by providing a proof of the linear algebra result, Lemma 3. Certainly, the basic exterior algebra idea in Lemma 3 is classic. See, for example, Birkhoff-MacLane [3, Problem 4, p. 566].

We will sketch the basic idea. Note that although our operators defined below are (of course) "bases free", it is only with respect to a particular fixed basis that the entries of the corresponding matrices belong to the algebras in question; which for us is $H^\infty(D^N)$.

For our notation, $l_{(n)}^2$ will denote the exterior product of l^2 with itself n-times, i.e., $l_{(n)}^2 = l^2 \wedge \dots \wedge l^2$ (n times). For $n = 0$, $l_{(0)}^2 = \mathbb{C}$. Let $\{e_j\}_{j=1}^\infty$ denote the standard basis in l^2. If $\Pi(n)$ denotes increasing n-tuples of positive integers

and if $(i_1, \ldots, i_n) \in \Pi(n)$, we let $\pi_n = \{i_1, \ldots, i_n\}$ and, abusing notation, we write $\pi_n \in \Pi(n)$.

Define $e_{\pi_n} = e_{i_1} \wedge \cdots \wedge e_{i_n}$. Then $\{e_{\pi_n}\}_{\pi_n \in \Pi(n)}$ denotes the standard basis for $l^2_{(n)}$.

For $f \sim \{f_n\}_{n=1}^{\infty}$ and $f_n \in H^\infty(D^N)$, we assume that $\epsilon^2 \leq F(\mathbf{z})F(\mathbf{z})^* \leq 1$ for all $\mathbf{z} \in D^N$. Fix $\mathbf{z} \in D^N$. For $n = 0, 1, \ldots$ define

$$Q_n^*(\mathbf{z}) : \ l^2_{(n)} \to l^2_{(n+1)}$$

by

$$Q_n^*(\mathbf{z})(w_n) = \overline{F(\mathbf{z})} \wedge w_n, \text{ where } w_n \in l^2_{(n)}.$$

Now

$$Q_n^*(\mathbf{z})(e_{\pi_n}) = \sum_{j=1}^{\infty} \overline{f_j(\mathbf{z})}\, e_j \wedge e_{\pi_n}.$$

So with respect to the standard basis, then entries of $Q_n^*(\mathbf{z})$ are 0 or else $\pm\overline{f_n(\mathbf{z})}$ for some n. Thus $Q_n(\cdot)$ has analytic entries with respect to the standard basis. This is the only place where we are using the particular algebra $H^\infty(D^N)$.

Proof. Fix $\mathbf{z} \in D^N$ and let $\underline{a} = \overline{F(\mathbf{z})}$ and $Q_n^* = Q_n^*(\mathbf{z})$. Then $Q_n^*(w_n) = \underline{a} \wedge w_n$. Choose an orthonormal basis $\{u_n\}_{n=1}^{\infty}$ of l^2 with $u_1 = \frac{\underline{a}}{\|\underline{a}\|}$. (Note $\|\underline{a}\|^2 \geq \epsilon^2$.) Then it follows that for $\pi_n \in \Pi(n)$ and $u_{\pi_n} = u_{i_1} \wedge \cdots \wedge u_{i_n}$, we have that $\{u_{\pi_n}\}_{\pi_n \in \Pi(n)}$ is an orthonormal basis for $l^2_{(n)}$. Thus

$$\begin{aligned} Q_n(w_{n+1}) &= \sum_{\pi_n \in \Pi(n)} \langle Q_n(w_{n+1}), u_{\pi_n} \rangle u_{\pi_n} \\ &= \sum_{\pi_n \in \Pi(n)} \langle w_{n+1}, \underline{a} \wedge u_{\pi_n} \rangle u_{\pi_n} \\ &= \|\underline{a}\| \sum_{\pi_n \in \Pi(n)} \langle w_{n+1}, u_1 \wedge u_{\pi_n} \rangle u_{\pi_n}. \end{aligned} \tag{7}$$

We wish to show that for $n = 0, 1, \ldots,$

$$Q_n^* Q_n + Q_{n+1} Q_{n+1}^* = \|\underline{a}\|^2 I_{l^2_{(n+1)}}. \tag{8}$$

For $n = 0$, $\frac{Q_0^* Q_0}{\|\underline{a}\|^2}$ is the rank one projection of l^2 onto $\underline{a}$. So given (7), $\frac{Q_1 Q_1^*}{\|\underline{a}\|^2}$ is a projection. But then $\frac{Q_1^* Q_1}{\|\underline{a}\|^2}$ must be a projection. Applying (7) again, we see that $\frac{Q_2 Q_2^*}{\|\underline{a}\|^2}$ is a projection. Repeating this procedure, we conclude that $\frac{Q_n Q_n^*}{\|\underline{a}\|^2}$ is the projection onto the range of Q_n. Also, given (7), it follows that $\operatorname{Ker} Q_n = \operatorname{ran} Q_{n+1}$.

To prove (7) it suffices to check that for $w_{n+1} \in l^2_{(n+1)}$,

$$\|Q_n(w_{n+1})\|^2 + \|Q_{n+1}^*(w_{n+1})\|^2 = \|\underline{a}\|^2 \|w_{n+1}\|^2. \tag{9}$$

Denote w_{n+1} by w. Then from (6), we see that

$$\|Q_n(w)\|^2 = \|\underline{a}\|^2 \sum_{\substack{\pi_n\in\Pi(n)\\ 1\notin\pi_n}} |\langle w, u_{1,\pi_n}\rangle|^2$$
$$= \|\underline{a}\|^2 \sum_{\substack{\pi_{n+1}\in\Pi(n+1)\\ 1\in\pi_{n+1}}} |\langle w, u_{\pi_{n+1}}\rangle|^2.$$

Also, since

$$Q^*_{n+1}(w) = \underline{a} \wedge \underline{w} = \|\underline{a}\|\, u_1 \wedge \sum_{\pi_{n+1}\in\Pi(n+1)} \langle w, u_{\pi_{n+1}}\rangle\, u_{\pi_{n+1}}$$
$$= \|\underline{a}\| \sum_{\substack{\pi_{n+1}\in\Pi(n+1)\\ 1\notin\pi_{n+1}}} \langle w, u_{\pi_{n+1}}\rangle u_1 \wedge u_{\pi_{n+1}},$$

we compute that

$$\|Q^*_{n+1}(w)\|^2 = \|\underline{a}\|^2 \sum_{\substack{\pi_{n+1}\in\Pi(n+1)\\ 1\notin\pi_{n+1}}} |\langle w, u_{\pi_{n+1}}\rangle|^2.$$

So (8) holds. This completes the proof of Lemma 3. □

References

[1] E. Amar, Big Hankel operator and $\overline{\partial}_b$-equation, *J. Oper. Theory* **33** (1995), 223–233.

[2] M. Andersson and H. Carlsson, Estimates of solutions of the H^p and BMOA corona problem, *Math. Ann.* **316** (2000), 83–102.

[3] G. Birkhoff and S. MacLane, *Algebra*, MacMillan, Toronto, 1971.

[4] J. Boo, The H^p corona theorem in analytic polyhedra, *Ark. Mat.* **35** (1997), 225–251.

[5] U. Cegrell, On ideals generated by bounded analytic functions in the bi-disc, *Bull. Soc. Math. Frances* **121** (1993), 109–116.

[6] S.A. Chang, Carleson measures on the bi-disc, *Annals of Math.* **109** (1979), 613–620.

[7] ———, *Two remarks about H^1 and BMO on the bidisk*, Conference on Harmonic Analysis in Honor of Antoni Zygmund, Vol. II, Wadsworth, Inc., 1983.

[8] S.A. Chang and R. Fefferman, Some recent developments in Fourier analysis and H^p-theory on product domains, *Bull. Amer. Math. Soc.* **12** (1985), 1–44.

[9] R.E. Edwards, *Functional Analysis, Theory and Applications*, Dover Pub., New York, 1995.

[10] J.B. Garnett, *Bounded Analytic Functions*, Academic Press, New York, 1981.

[11] P. Koosis, *Introduction to H^p Spaces*, Cambridge University Press, New York, 1980.

[12] M. Lacey and E. Terwilleger, *Hankel operators in several complex variables and product BMO*($\bigotimes_1^n C_+$), preprint.

[13] S.-Y. Li, *Corona problems of several complex variables*, Madison Symposium of Complex Analysis: Contemporary Mathematics, vol. 137, Amer. Math. Soc., 1991.

[14] K.C. Lin, H^p solutions for the corona problem on the polydisc in $\mathbb{C}^n$, *Bull. Sci. Math.* **110** (1986), 69–84.

[15] N.K. Nikolski, *Treatise on the Shift Operator*, Springer-Verlag, New York, 1985.

[16] E.M. Stein, *Harmonic Analysis*, Princeton University Press, Princeton, New Jersey, 1993.

[17] T.T. Trent, A new estimate for the vector-valued corona theorem, *J. Func. Anal.* **189** (2002), 267–282.

[18] ———, An H^2-corona theorem on the bidisk for infinitely many functions, *Linear Alg. and its Appl.* **379** (2004), 213–227.

[19] ———, A vector-valued H^p-corona theorem on the polydisk, *Int. Equa. and Op. Theory*, to appear.

[20] N.Th. Varopoulos, Probabilistic approach to some problems in complex analysis, *Bull. Sci. Math.* **105** (1981), 181–224.

Tavan T. Trent
Department of Mathematics
The University of Alabama
Box 870350
Tuscaloosa, AL 35487-0350
USA
e-mail: ttrent@gp.as.ua.edu

Operator Theory:
Advances and Applications, Vol. 179, 329–338

A Matrix and its Inverse: Revisiting Minimal Rank Completions

Hugo J. Woerdeman

Abstract. We revisit a formula that connects the minimal ranks of triangular parts of a matrix and its inverse and relate the result to structured rank matrices. We also address a generic minimal rank problem that was proposed by David Ingerman and Gilbert Strang.

Mathematics Subject Classification (2000). 15A09, 15A15, 65F05.

Keywords. Minimal rank, matrix completion, nullity theorem, band matrix, semi-separable, quasi-separable.

1. Introduction

In this paper we revisit the following result from [22]:

Let $[(T_{ij})_{i,j=1}^n]^{-1} = (S_{ij})_{i,j=1}^n$ *be block matrices with sizes that are compatible for multiplication. Other than the full matrix (which is of size* N*, say), none of the blocks need to be square. Then*

$$\min \ \operatorname{rank} \begin{pmatrix} T_{11} & ? & \cdots & ? \\ T_{21} & T_{22} & \cdots & ? \\ \vdots & & \ddots & \vdots \\ T_{n1} & T_{n2} & \cdots & T_{nn} \end{pmatrix} + \min \ \operatorname{rank} \begin{pmatrix} ? & ? & \cdots & ? \\ S_{21} & ? & \cdots & ? \\ \vdots & \ddots & \ddots & \vdots \\ S_{n1} & \cdots & S_{n,n-1} & ? \end{pmatrix} = N. \tag{1.1}$$

With the recent interest in numerical algorithms that make effective use of matrices with certain rank structures (see, e.g., [4], [21], [18], [7], [9], and references therein), it seems appropriate to revisit this formula that captures many of the rank considerations that go into these algorithms. The nullity theorem due to [10] is a particular case. The papers [17] and [20] show the recent interest in the nullity theorem. It is our hope that this general formula (1.1) enhances the insight in rank structured matrices.

Supported in part by National Science Foundation grant DMS-0500678.

In addition, in Section 3 we will address the so-called "generic minimal rank problem". This problem was introduced by Gilbert Strang and David Ingerman.

2. Minimal ranks of matrices and their inverses

Let us recall the notion of partial matrices and their minimal rank. Let $\mathbb{F}$ be a field and let $n, m, \nu_1, \ldots, \nu_n, \mu_1, \ldots, \mu_m$ be nonnegative integers. The *pattern* of specified entries in a partial matrix will be described by a set $J \subset \{1, \ldots, n\} \times \{1, \ldots, m\}$. A pattern K that is a subset of J will be called a *subpattern* of J. Let now $A_{ij}, (i,j) \in J$, be given matrices with entries in $\mathbb{F}$ of size $\nu_i \times \mu_j$. We will allow ν_i and μ_j to equal 0. The collection of matrices $\mathcal{A} = \{ A_{ij};\ (i,j) \in J \}$ is called a *partial block matrix with the pattern J*. When all the blocks are of size 1×1 (i.e., $\nu_i = \mu_j = 1$ for all i and j), we will simply talk about a *partial matrix*. Clearly, any block matrix as above may be viewed as a partial matrix of size $N \times M$ as well, where $N = \nu_1 + \ldots + \nu_n, M = \mu_1 + \ldots + \mu_m$. It will be convenient to represent partial block matrices in matrix format. As usual a question mark will represent an unknown block. For instance, $\mathcal{A} = \{A_{ij} : 1 \leq j \leq i \leq n\}$ will be represented as

$$\mathcal{A} = \begin{pmatrix} A_{11} & ? & \ldots & ? \\ \vdots & \ddots & \ddots & \vdots \\ \vdots & & \ddots & ? \\ A_{n1} & \ldots & \ldots & A_{nn} \end{pmatrix}.$$

Let a partial matrix $\mathcal{A} = \{ A_{ij};\ (i,j) \in J \}$ be given. A block matrix $B = (B_{ij})_{i=1,j=1}^{n\ \ m}$ with $B_{ij} \in \mathbb{F}^{\nu_i \times \mu_j}$ is called a *completion* of $\mathcal{A}$ if $B_{ij} = A_{ij}, (i,j) \in J$. The *minimal rank* of $\mathcal{A}$ (notation: min rank$(\mathcal{A})$) is defined by

$$\text{min rank}(\mathcal{A}) = \min\{\text{rank}\, B : B \text{ is a completion of } \mathcal{A}\}.$$

The formula that connects the minimal ranks of triangular parts of a matrix and its inverse is the following. The result appeared originally in [22] (see also [24] and Chapter 5 of [23]).

Theorem 2.1. [22] *Let $T = (T_{ij})_{i,j=1}^n$ be an invertible block matrix with T_{ij} of size $\nu_i \times \mu_j$, where $\nu_i \geq 0$, $\mu_j \geq 0$ and $N = \nu_i + \ldots + \nu_n = \mu_i + \ldots + \mu_n$. Put $T^{-1} = (S_{ij})_{i,j=1}^n$ where S_{ij} is of size $\mu_i \times \nu_j$. Then*

$$\text{min rank} \begin{pmatrix} T_{11} & ? & \cdots & ? \\ T_{21} & T_{22} & \cdots & ? \\ \vdots & & \ddots & \vdots \\ T_{n1} & T_{n2} & \cdots & T_{nn} \end{pmatrix} + \text{min rank} \begin{pmatrix} ? & ? & \cdots & ? \\ S_{21} & ? & \cdots & ? \\ \vdots & \ddots & \ddots & \vdots \\ S_{n1} & \cdots & S_{n,n-1} & ? \end{pmatrix} = N.$$

As we will see, one easily deduces from Theorem 2.1 that the inverse of a lower Hessenberg matrix has the upper triangular part of a rank 1 matrix. The strength of Theorem 2.1 lies in that one easily deduces a multitude of such results from it.

From the same paper [22] we would also like to recall the following result.

Theorem 2.2. [22] *The partial matrix* $\mathcal{T} = \{T_{ij} : 1 \leq j \leq i \leq n\}$ *has minimal rank*

$$\min \operatorname{rank} \mathcal{T} = \sum_{i=1}^{n} \operatorname{rank} \begin{pmatrix} T_{i1} & \dots & T_{ii} \\ \vdots & & \vdots \\ T_{n1} & \dots & T_{ni} \end{pmatrix} - \sum_{i=1}^{n-1} \operatorname{rank} \begin{pmatrix} T_{i+1,1} & \dots & T_{i+1,i} \\ \vdots & & \vdots \\ T_{n1} & \dots & T_{ni} \end{pmatrix}.$$

For the 2×2 case of Theorem 2.2 one needs to observe that the minimal rank of

$$\begin{pmatrix} T_{11} & ? \\ T_{21} & T_{22} \end{pmatrix}$$

will at least be the rank of $\begin{pmatrix} T_{11} \\ T_{21} \end{pmatrix}$ plus the minimal number of columns in T_{22} that together with the columns of T_{21} span the column space of $\begin{pmatrix} T_{21} & T_{22} \end{pmatrix}$. Once such a minimal set of columns in T_{22} has been identified, put any numbers on top of these columns. Now any other columns in T_{22} can be completed to be a linear combination of fully completed columns. Doing this leads to a completion of rank

$$\operatorname{rank} \begin{pmatrix} T_{11} \\ T_{21} \end{pmatrix} + \operatorname{rank} \begin{pmatrix} T_{21} & T_{22} \end{pmatrix} - \operatorname{rank} T_{21},$$

yielding the case $n = 2$ of Theorem 2.2. The general case now follows easily by induction.

The proof of Theorem 2.1, which can be found in [22] is easily derived from Theorem 2.2 and the nullity theorem, which we recall now.

Theorem 2.3. [10] *Consider*

$$\begin{pmatrix} A & B \\ C & D \end{pmatrix}^{-1} = \begin{pmatrix} P & Q \\ R & S \end{pmatrix}.$$

Then $\dim \ker C = \dim \ker R$.

Proof. Since $CP = -DR$, $P[\ker R] \subseteq \ker C$. Likewise, since $RA = -SC$, we get $A[\ker C] \subseteq \ker R$. Consequently,

$$AP[\ker R] \subseteq A[\ker C] \subseteq \ker R.$$

Since $AP + BR = I$, $AP[\ker R] = \ker R$, thus

$$A[\ker C] = \ker R.$$

This yields $\dim \ker C \geq \dim \ker R$. By reversing the roles of C and R one obtains also that $\dim \ker R \geq \dim \ker C$. This gives $\dim \ker R = \dim \ker C$, yielding the lemma. □

The nullity theorem is in fact the case $n = 2$ of Theorem 2.1. Indeed, if

$$T^{-1} = \begin{pmatrix} T_{11} & T_{12} \\ T_{21} & T_{22} \end{pmatrix}^{-1} = \begin{pmatrix} S_{11} & S_{12} \\ S_{21} & S_{22} \end{pmatrix},$$

we get from Theorem 2.1 that

$$\operatorname{rank}\begin{pmatrix} T_{11} \\ T_{21} \end{pmatrix} + \operatorname{rank}\begin{pmatrix} T_{21} & T_{22} \end{pmatrix} - \operatorname{rank} T_{21} + \operatorname{rank} S_{21} = N. \tag{2.1}$$

As T is invertible we have that $\begin{pmatrix} T_{11} \\ T_{21} \end{pmatrix}$ and $\begin{pmatrix} T_{21} & T_{22} \end{pmatrix}$ are full rank, so (2.1) gives

$$\mu_1 + \nu_2 - \operatorname{rank} T_{21} + \operatorname{rank} S_{21} = \mu_1 + \mu_2 = \nu_1 + \nu_2,$$

and thus

$$\nu_2 - \operatorname{rank} T_{21} = \mu_2 - \operatorname{rank} S_{21},$$

which is exactly Theorem 2.3.

To make the connection with some of the results in the literature we need the following proposition.

Proposition 2.4. *Let $\mathcal{T} = \{t_{ij} : 1 \le j \le i \le n\}$ be a scalar-valued partial matrix. Then $\min\operatorname{rank}(\mathcal{T}) = n$ if and only if $t_{ii} \neq 0, i = 1, \ldots, n$, and $t_{ij} = 0$ for $i > j$.*

Proof. The "if" part is immediate. For the only if part write

$$\min\operatorname{rank}\mathcal{T} = \operatorname{rank}\begin{pmatrix} t_{11} \\ \vdots \\ t_{n1} \end{pmatrix} + \sum_{i=2}^{n} s_i, \tag{2.2}$$

where

$$s_i = \operatorname{rank}\begin{pmatrix} t_{i1} & \ldots & t_{ii} \\ \vdots & & \vdots \\ t_{n1} & \ldots & t_{ni} \end{pmatrix} - \operatorname{rank}\begin{pmatrix} t_{i1} & \ldots & t_{i,i-1} \\ \vdots & & \\ t_{n1} & \ldots & t_{n,i-1} \end{pmatrix}.$$

All the terms in (2.2) are at most 1, and as there are exactly n terms they need to all be equal to 1 for $\min\operatorname{rank}(\mathcal{T}) = n$ to be satisfied. But then $s_n = 1$ implies $t_{n1} = \ldots = t_{n,n-1} = 0$ and $t_{nn} \neq 0$. Inductively, one can then show that $s_k = 1$ implies $t_{k1} = \ldots = t_{k,k-1} = 0$ and $t_{kk} \neq 0, k = n-1, \ldots, 2$. Finally the first column of $\mathcal{T}$ needs to have rank 1. As $t_{ij} = 0, j = 2, \ldots, n$, was already established we get that $t_{11} \neq 0$. This proves the result. □

We now easily obtain the following corollary, due to Asplund [1].

Corollary 2.5. [1] *Let $p \ge 0$ and $A = (a_{ij})_{i,j=1}^{N}$ be an $N \times N$ scalar matrix with inverse $B = (b_{ij})_{i,j=1}^{N}$. Then $a_{ij} = 0$ for all i and j with $j > i + p$, and $a_{ij} \neq 0$, $j = i + p$, if and only if there exist an $N \times p$ matrix F and a $p \times N$ matrix G so that $b_{ij} = (FG)_{ij}$, $i < j + p$. In particular, if $p = 1$ (so A is lower Hessenberg), then $b_{ij} = F_i G_j$, $1 \le i \le j \le N$, where $F_1, \ldots, F_N, G_1, \ldots, G_N$ are scalars.*

Proof. Let $(S_{ij})_{i,j=1}^{N-p+1} = A$, where S_{i1} is of size $1 \times p$, $i = 1, \ldots, n-p$, $S_{N-p+1,1}$ has size $p \times p$, $S_{N-p+1,j}$ has size $p \times 1$, $j = 2, \ldots, N-p+1$, and all the other S_{ij}

are 1×1. Let $B = (T_{ij})_{i,j=1}^{N-p+1}$ be partitioned accordingly. Then, it follows from (1.1) that

$$\min \ \operatorname{rank} \begin{pmatrix} ? & ? & \cdots & ? \\ S_{21} & ? & \cdots & ? \\ \vdots & \ddots & \ddots & \vdots \\ S_{n1} & \cdots & S_{n,n-1} & ? \end{pmatrix} = N - p$$

if and only if

$$\min \ \operatorname{rank} \begin{pmatrix} T_{11} & ? & \cdots & ? \\ T_{21} & T_{22} & \cdots & ? \\ \vdots & & \ddots & \vdots \\ T_{n1} & T_{n2} & \cdots & T_{nn} \end{pmatrix} = p.$$

Using Proposition 2.4 the result now follows. □

Corollary 2.5 shows that Theorem 2.1 is useful in the contexts of semi-separability and quasi-separability (see, e.g., [19] and [6] for an overview of these notions). In a similar way it is easy to deduce results by [3], [13], [14], [15], [16] and [8] from Theorem 2.1.

3. The generic minimal rank completion problem

Let $r \in \mathbb{N}$. We will call a pattern $J \subseteq \{1, \ldots, n\} \times \{1, \ldots, m\}$ *r-sparse* if for every $K \subset \{1, \ldots, n\}$ and every $L \subseteq \{1, \ldots, m\}$ with $|K| = |L| \geq r$ we have that

$$|J \cap (K \times L)| \leq (2|K| - r)r.$$

Thus square submatrices of size N intersect the pattern in at most $2Nr - r^2$ positions. The number $2Nr - r^2$ comes from the situation where exactly r rows and r columns in an $N \times N$ submatrix are prescribed.

Recently D. Ingerman and G. Strang observed that a 1-sparse matrix with nonzero entries has a completion of rank 1. We will give a proof of this fact below. This observation led Ingerman and Strang to ask: Is it true that one can "generically" complete a r-sparse partial matrix to a matrix of rank $\leq r$? Of course, to be able to answer the question one needs to define "generically". The 1-sparse case suggests that it may suffice to require that all fully specified matrices are nonsingular. The following example, however, shows that this is not the right formulation of "generic".

Example 3.1. Consider the matrix

$$A := \begin{pmatrix} 6 & 3 & x & 1 \\ 3 & 1 & 1 & y \\ z & 1 & 2 & 3 \\ 1 & w & 1 & 1 \end{pmatrix},$$

where x, y, z and w are the unknowns. Note that this partial matrix has a pattern that is 2-sparse. However there is no completion of rank 2. Indeed, suppose that $\text{rank} A = 2$. Then we have that

$$\begin{pmatrix} 6 & 3 \\ 3 & 1 \end{pmatrix} - \begin{pmatrix} x & 1 \\ 1 & y \end{pmatrix} \begin{pmatrix} 2 & 3 \\ 1 & 1 \end{pmatrix}^{-1} \begin{pmatrix} z & 1 \\ 1 & w \end{pmatrix} = 0,$$

and since the rank of the first term is 2, the second term must also have rank 2. Thus, we have that $xy \neq 1$ and $zw \neq 1$. Next, we also have that

$$\begin{pmatrix} z & 1 \\ 1 & w \end{pmatrix} - \begin{pmatrix} 2 & 3 \\ 1 & 1 \end{pmatrix} \begin{pmatrix} x & 1 \\ 1 & y \end{pmatrix}^{-1} \begin{pmatrix} 6 & 3 \\ 3 & 1 \end{pmatrix} = 0.$$

Multiplying on both sides with $xy - 1$, the off-diagonal entries yield the equations

$$xy - 6y - 3x + 10 = 0, \quad xy - 6y - 3x + 8 = 0.$$

These are not simultaneously solvable (as long as we are in a field where $8 \neq 10$). It should be noted that this is a counterexample for any field in which $6 \neq 9$, $6 \neq 1$, $3 \neq 1$, $9 \neq 1$ (so that we have full rank-specified submatrices) and $8 \neq 10$. As an aside, we note that for some of the small fields it may be impossible to fulfill the nondegeneracy requirement on the data. E.g., when $\mathbb{F} = \{0, 1\}$, a 2×2 matrix can only be nonsingular if zeroes are allowed in the matrix.

In order to formulate our results and conjectures it is useful to introduce the *bipartite graph* associated with a pattern. Given a pattern J the corresponding (undirected) bipartite graph $G(J)$ has vertices $\{v_1, \ldots, v_n, u_1, \ldots, u_m\}$, and (v_i, u_j) is an edge in $G(J)$ if and only if $(i, j) \in J$. A subset $K \subseteq J$ is called a *cycle* in $G(J)$ if K is of the form

$$K = \{(i_1, j_1), (i_2, j_1), (i_2, j_2), \ldots, (i_k, j_k), (i_1, j_k)\} \tag{3.1}$$

for distinct $i_1, \ldots, i_k$ and distinct $j_1, \ldots, j_k$, where $k \geq 2$. In other words, $G(K)$ is a cycle in $G(J)$. The *length* of the cycle K is its number of elements; so for K in (3.1) we have that its length is $2k$. Note that in a bipartite graph all cycles have even length. When a pattern is 1-sparse, it cannot have any cycles as a cycle of length k requires the corresponding partial matrix to have $2k$ specified entries in a $k \times k$ submatrix. This observation makes it easy to apply a result in [5] to prove the statement by Ingerman and Strang.

Proposition 3.2. [11] *Any partial matrix whose pattern is* 1*-sparse and whose given entries are nonzero, has a rank* 1 *completion.*

Proof. Let J denote the pattern of specified entries of the matrix. By the observation before the proposition we have that $G(J)$ does not contain any cycles. Therefore by Lemma 6.2 in [5] it suffices to observe that the absence of zero specified entries implies trivially that the partial matrix "is singular with respect to all three lines in $G(J)$" (the latter condition can be rephrased as saying that for all 2×2 submatrices there is a rank one completion). □

We say that the cycle K in (3.1) has a *chord* in J if for some $1 \leq p, q \leq k$ we have that $(i_p, j_q) \in J \setminus K$. One can think of a chord as a "shortcut" in the cycle. The cycle K is called *minimal* in J if it does not have a chord in J. Notice that a 4-cycle is a complete bipartite subgraph, and by definition it is automatically minimal. Cycles of larger (necessarily even) length can be minimal or not. We say that a bipartite graph is called *chordal* if it does not have minimal cycles of length 6 or larger.

Notice that the bipartite graph associated with the partial matrix in Example 3.1 is a minimal cycle of length 8, and therefore the graph is not chordal. Therefore it could be that the obstruction of finding a solution in Example 3.1 lies in the non-chordality of the underlying bipartite graph. As we have seen in [5, Theorem 3.1] it is also the non-chordality that prevents graphs from being so-called "rank determined". In addition, notice that in the case $r = 1$ the 1-sparse property prevents the existence of minimal cycles of length 6 or more, and thus in that case all patterns are automatically bipartite chordal. Thus we arrive at the following conjecture.

Conjecture 3.3. *Consider an r-sparse partial matrix $\mathcal{A}$ with a bipartite graph that is chordal. If all fully specified submatrices have full rank, then $\mathcal{A}$ has a completion of rank at most r.*

We can prove the conjecture for the subclass of banded patterns. Recall (cf. [25]) that a pattern $J \subset \{1, \ldots, n\} \times \{1, \ldots, m\}$ is called *banded* if there exist permutations σ on $\{1, \ldots, n\}$ and τ on $\{1, \ldots, m\}$ so that

$$J_{\sigma,\tau} := \{(\sigma(i), \tau(j)) \ ; \ (i, j) \in J\}$$

satisfies

$$(i, j), (k, l) \in J_{\sigma,\tau}, i \leq k, j \geq l \Rightarrow \{i, \ldots, k\} \times \{l, \ldots, j\} \subset J_{\sigma,\tau}.$$

Theorem 3.4. *Let $\mathbb{F}$ be an infinite subfield of $\mathbb{C}$. Consider a r-sparse partial matrix with a banded pattern and suppose that all fully specified submatrices have full rank. Then there exists a completion of rank at most r.*

To prove this result we need the notion of triangular pattern: a pattern J is called *triangular* if there exist permutations σ and τ so that $J_{\sigma,\tau}$ satisfies

$$(i, j) \in J_{\sigma,\tau} \Rightarrow \{i, \ldots, n\} \times \{1, \ldots, j\} \subset J_{\sigma,\tau}.$$

Proof of Theorem 3.4. By Theorem 1.1 in [25] it suffices to show that for every triangular subpattern (for the definition, see [25]) we have that the minimal rank is $\leq r$. But a triangular subpattern can always be embedded in a pattern that corresponds to r rows and columns specified (due to the condition that in any $k \times k$ submatrix at most $(2k - r)r$ entries are specified). But then the result follows. □

Observe that the proof shows that if the bipartite chordal minimal rank conjecture (Conjecture 3.3 in [5]; see also Chapter 5 in [23]) is true, then Conjecture

3.3 above is true as well. The techniques developed in [2] and/or [12] may be helpful in proving this conjecture above.

While Example 3.1 shows that not every r-sparse partial matrix with full rank-specified submatrices has a rank r completion, it should be noticed that if we consider the example as one over an infinite subfield of $\mathbb{C}$ and we perturb the data slightly, then there is a rank r completion. This observation leads to the following.

Let $\mathbb{F}$ be an infinite subfield of $\mathbb{C}$, and let J be a pattern. Consider the set $\mathcal{P}_J$ of partial matrices over $\mathbb{F}$ with pattern J. We can identify $\mathcal{P}_J$ with the set $\mathbb{F}^{|J|}$, and we use this correspondence and the usual topology on $\mathbb{F}^{|J|}$ to define a topology on $\mathcal{P}_J$. We can now formulate the following conjecture, which probably best describes the conjecture that Ingerman and Strang had in mind.

Conjecture 3.5. *Let J be an r-sparse pattern. Then there is a dense subset $\mathcal{P}'$ of $\mathcal{P}_J$, so that all partial matrices in $\mathcal{P}'$ have a completion of rank at most r.*

Analyzing Example 3.1 it is not hard to convince oneself that Conjecture 3.5 is true for the 2-sparse pattern $J = \{1,\dots,4\}\times\{1,\dots,4\}\backslash\{(1,4),(2,3),(3,2),(4,1)\}$. Indeed, a desired set $\mathcal{P}'$ is described by all partial matrices for which the given entries satisfy 4 nonequalities. These nonequalities are obtained by eliminating three out of four unknowns, and requiring that the coefficient in front of the highest power is not equal to 0; doing this for all possible choices of the remaining unknown, we get 4 nonequalities. Also, by the results stated earlier, the conjecture is true for $r = 1$, and for banded J.

One may perhaps prove Conjecture 3.5 in the following way. View the partial matrices in $\mathcal{P}_J$ as a regular matrix where the unknowns are represented by variables $x_1,\dots,x_k$, $k = nm - |J|$. Let I be the ideal generated by the determinants $p_1,\dots,p_N$ of $(r+1)\times(r+1)$ submatrices as polynomials in the unknowns $x_1,\dots,x_k$. We now need to show that generically, the constant polynomial 1 is not in I. Use of elimination theory may perhaps be used to now show that one can only eliminate all variables when the coefficients satisfy certain equalities. Finally, observing that these equalities are generically not satisfied one may finish the proof.

References

[1] E. Asplund, *Inverses of matrices $\{a_{ij}\}$ which satisfy $a_{ij} = 0$ for $j > i + p$.* Math. Scand. **7** (1959), 57–60.

[2] M. Bakonyi and A. Bono, *Several results on chordal bipartite graphs.* Czechoslovak Math. J. **47** (1997), 577–583.

[3] W. W. Barrett and Ph. J. Feinsilver, *Inverses of banded matrices.* Linear Algebra Appl. **41** (1981), 111–130.

[4] D. A. Bini, L. Gemignani, and V. Y. Pan, *Fast and stable QR eigenvalue algorithms for generalized companion matrices and secular equations.* Numer. Math. **100** (2005), 373–408.

[5] N. Cohen, C. R. Johnson, L. Rodman, and H. J. Woerdeman, *Ranks of completions of partial matrices.* In "The Gohberg anniversary collection", Vol. I (Calgary, AB, 1988), Oper. Theory Adv. Appl. (Birkhäuser, Basel) **40** (1989), 165–185.

[6] Y. Eidelman and I. Gohberg, *On generators of quasiseparable finite block matrices.* Calcolo **42** (2005), 187–214.

[7] Y. Eidelman, I. Gohberg, and V. Olshevsky, *Eigenstructure of order-one-quasiseparable matrices. Three-term and two-term recurrence relations.* Linear Algebra Appl. **405** (2005), 1–40.

[8] L. Elsner, *A note on generalized Hessenberg matrices.* Linear Algebra Appl. **409** (2005), 147–152.

[9] D. Fasino and L. Gemignani, *A Lanczos-type algorithm for the QR factorization of Cauchy-like matrices.* In "Fast algorithms for structured matrices: theory and applications" (South Hadley, MA, 2001), Contemp. Math. (Amer. Math. Soc., Providence, RI) **323** (2003), 91–104.

[10] W. H. Gustafson, *A note on matrix inversion.* Linear Algebra Appl. **57** (1984), 71–73.

[11] D. Ingerman and G. Strang, private communication.

[12] C. R. Johnson and J. Miller, *Rank decomposition under combinatorial constraints.* Linear Algebra Appl. **251** (1997), 97–104.

[13] P. Rózsa, *Band matrices and semiseparable matrices.* In "Numerical methods" (Miskolc, 1986), Colloq. Math. Soc. János Bolyai (North-Holland, Amsterdam) **50** (1988), 229–237.

[14] P. Rózsa, R. Bevilacqua, P. Favati, and F. Romani, *On the inverse of block tridiagonal matrices with applications to the inverses of band matrices and block band matrices.* In "The Gohberg anniversary collection", Vol. I (Calgary, AB, 1988), Oper. Theory Adv. Appl. (Birkhäuser, Basel) **40** (1989), 447–469.

[15] P. Rózsa, R. Bevilacqua, F. Romani, and P. Favati, *On band matrices and their inverses.* Linear Algebra Appl. **150** (1991), 287–295.

[16] P. Rózsa, F. Romani, and R. Bevilacqua, *On generalized band matrices and their inverses.* In "Proceedings of the Cornelius Lanczos International Centenary Conference" (Raleigh, NC, 1993), SIAM, Philadelphia, PA, (1994), 109–121.

[17] G. Strang and T. Nguyen, *The interplay of ranks of submatrices.* SIAM Rev. **46** (2004), 637–646.

[18] E. Tyrtyshnikov, *Piecewise separable matrices.* Calcolo **42** (2005), 243–248.

[19] R. Vandebril, M. Van Barel, G. Golub, and N. Mastronardi, *A bibliography on semiseparable matrices.* Calcolo **42** (2005), 249–270.

[20] R. Vandebril and M. Van Barel, *A note on the nullity theorem.* J. Comput. Appl. Math. **189** (2006), 179–190.

[21] R. Vandebril, M. Van Barel, and N. Mastronardi, *An implicit QR algorithm for symmetric semiseparable matrices.* Numer. Linear Algebra Appl. **12** (2005), 625–658.

[22] H. J. Woerdeman, *The lower order of lower triangular operators and minimal rank extensions.* Integral Equations Operator Theory **10** (1987), 859–879.

[23] H. J. Woerdeman, *Matrix and operator extensions.* Stichting Mathematisch Centrum voor Wiskunde en Informatica, Amsterdam, 1989.

[24] H. J. Woerdeman, *Minimal rank completions for block matrices.* Linear Algebra Appl. **121** (1989), 105–122.

[25] H. J. Woerdeman, *Minimal rank completions of partial banded matrices.* Linear and Multilinear Algebra **36** (1993), 59–68.

Hugo J. Woerdeman
Department of Mathematics
Drexel University
3141 Chestnut Street
Philadelphia, PA 19104
USA
e-mail: hugo@math.drexel.edu