Operator Theory: Advances and Applications

Vol. 197

Characteristic Functions, Scattering Functions and Transfer Functions

The Moshe Livsic Memorial Volume

Daniel Alpay
Victor Vinnikov
Editors

Birkhäuser
Basel · Boston · Berlin

Editors:

Daniel Alpay
Victor Vinnikov
Deaprtment of Mathematics
Ben-Gurion University of the Negev
P.O. Box 653
Beer Sheva 84105
Israel
e-mail: dany@math.bgu.ac.il
vinnikov@math.bgu.ac.il

2000 Mathematics Subject Classification: 47A, 60G, 93C

Library of Congress Control Number: 2009929500

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

ISBN 978-3-0346-0182-5 Birkhäuser Verlag AG, Basel – Boston – Berlin

ISBN 978-3-0346-0182-5 e-ISBN 978-3-0346-0183-2

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

Contents

Editorial Introduction

Daniel Alpay and Victor Vinnikov

During the period July 9 to July 13, 2007, a conference called **Characteristic functions and transfer functions in operator theory and system theory: a conference dedicated to Paul Fuhrmann on his 70th anniversary and to the memory of Moshe Livsic on his 90th anniversary** was held at the Department of Mathematics of Ben–Gurion University of the Negev. The notions of transfer function and characteristic functions proved to be fundamental in the last fifty years in operator theory and in system theory. This conference was envisaged to pay tribute to our colleagues Paul Fuhrmann and Moshe Livsic who played a central role in developing these notions. Sadly, Moshe Livsic passed away on the 30th of March, 2007 (11th of Nissan 5767), so the conference was dedicated to his memory. It is a pleasure to thank all the participants, who contributed to a very exciting and fruitful conference, and especially those who submitted papers to the present volume.

The volume contains a selection of thirteen research papers dedicated to the memory of Moshe Livsic. The topics addressed can be divided into the following categories:

Classical operator theory and its applications: This pertains to the paper *Differential-difference equations in entire functions* by G. Belitskii and V. Tkachenko, the paper *Bi-Isometries and Commutant Lifting* by H. Bercovici, R.G. Douglas. and C. Foias and the paper *Convexity of ranges and connectedness of level sets of quadratic forms* by I. Feldman, N. Krupnik and A. Markus.

Ergodic theory and stochastic processes: We have the papers *The one-sided ergodic Hilbert transform of normal contractions* by G. Cohen and M. Lin, and *Integral Equations in the Theory of Levy Processes* by L. Sakhnovich.

Geometry of smooth mappings: This is covered by the paper of Y. Yomdin entitled *β-Spread of sets in metric spaces and critical values of smooth functions*.

Mathematical physics: This topic is covered by the paper *Solvable models for quantum networks and a modified analytic perturbation procedure* by B. Pavlov.

Schur analysis: This is covered by the paper of H. Dym and D. Volok *Pick Matrices for Schur Multipliers* and the papers *The Schur Algorithm in Terms of System*

Realizations by B. Fritzsche, V. Katsnelson and B. Kirstein and *On some interrelations between J-Potapov functions and J-Potapov sequences* by B. Fritzsche, B. Kirstein and U. Raabe.

System theory: This topic is covered by the papers *Inverse Stieltjes like functions and inverse problems for systems with Schrödinger operator* by S. Belyi and E. Tsekanovskii, the paper *Model reduction in symbolically semiseparable systems with application to building preconditioners for 3D sparse systems of equations* by P. Dewilde, H. Jiao and S. Chandrasekaran, and *The Stable rank of a Nest Algebra and Strong Stabilization of Linear Time-Varying Systems* by A. Feintuch.

The variety of topics attests well to the breadth of Moshe Livsic's mathematical vision and the deep impact of his work. Another volume, entitled **Selected translations of papers of Moshe Livsic**, is planned in the same book series, with editors Victor Katsnelson, Israel Gohberg and the present editors. The volume will also contain memorial material about Moshe.

The conference in July 2007 was supported by the following sources: The Center for Advanced Studies in Mathematics of Ben–Gurion University of the Negev (which has also supported three previous workshops on operator and system theory in 2001, 2003, and 2005), the President of Ben–Gurion University of the Negev, the Rector of Ben–Gurion University of the Negev, the Dean of the Faculty of Natural Sciences of Ben–Gurion University of the Negev, and the Earl Katz Family Chair in Algebraic System Theory. The visit and participation Prof. J. Ball was supported by the US-Israel Binational Science Foundation. The participation of Prof. H. Woerdeman was supported by the Visiting Fellowship Program at the Faculty of Natural Sciences of Ben–Gurion University of the Negev.

Daniel Alpay
Earl Katz Family Chair in algebraic system theory
Department of Mathematics
Ben–Gurion University of the Negev
Beer-Sheva 84105, Israel
e-mail: dany@math.bgu.ac.il

Victor Vinnikov
Department of Mathematics
Ben–Gurion University of the Negev
Beer-Sheva 84105, Israel
e-mail: vinnikov@math.bgu.ac.il

Operator Theory:
Advances and Applications, Vol. 197, 1–19

Differential-difference Equations in Entire Functions

Genrich Belitskii and Vadim Tkachenko

To the memory of Moshe Livshits

Abstract. For a linear differential-difference equation with real shifts in the complex plane we prove a theorem of existence of entire solutions for an arbitrary entire function in the r.h.s. and, using it, show that the space of entire solutions of the corresponding homogeneous equation is infinite dimensional.

Mathematics Subject Classification (2000). Primary:30D05. Secondary:34K06.

Keywords. Linear functional equations, entire functions.

1. Introduction

We consider a differential-difference equation with

$$\sum_{k=0}^{m}\sum_{j=0}^{p_k} a_{jk}(z)\varphi^{(j)}(z+\alpha_k) = \gamma(z), \quad z \in \mathbb{C}, \tag{1.1}$$

with real shifts $0 = \alpha_0 < \alpha_1 < \cdots < \alpha_m$ and entire functions γ, a_{jk}, $k = 0, \dots, m$; $j = 0, \dots, p_k$. Our aim is to prove that under some restrictions imposed on the coefficients a_{jk} there exists an entire solution φ of the above equation for each entire function γ, and that the space of entire solutions of the corresponding homogeneous equation is infinite dimensional.

Difference equation (1.1) with constant coefficients, i.e., equations of the form (1.1) with $p_0 = \cdots = p_m = 0$, belongs to a branch of the complex analysis originated from the works by Euler, Bernoulli and other classics of mathematics, see [1]. Differential-difference equations with constant coefficients (even with complex shifts α_k) may be treated in the framework of differential equations of infinite order with constant coefficients [2]. The main result here states that the space of solutions of homogeneous equation is spanned by its elementary solutions $z^k e^{\lambda z}$, while the non-homogeneous equation is solvable in entire functions for arbitrary entire functions γ.

Much less is known about solutions of equations (1.1) in the complex plane with non-constant coefficients a_{jk}. Naftalevich [3] studied equations

$$f(z+1) = \exp[P(z)]f'(z) \tag{1.2}$$

where $P(z)$ is a polynomial and showed that its entire solutions form an infinite-dimensional space. He also proved some results on the solvability of the respective non-homogeneous equation in meromorphic functions.

Our recent interest in the theory of equations (1.1) is connected with the following old problem formulated by Hurwitz [5].

Let f_0 be an analytic germ at the point $z = 1$ and let f_0 may be analytically continued along the unit circle $\{\,\xi : |\xi| = 1\,\}$. Assume f_1 be the analytic germ of this continuation at $z = 1$ and $f_1'(\xi)/f_0(\xi) \equiv \text{const}$. Does the identity $f_0(\xi) = Ce^{\xi}$ now follow? The question was answered in negative by H.Lewy (see the paper [5] cited above) who explicitly constructed the requested function $f_0(\xi) \neq Ce^{\xi}$.

The substitution $\xi = e^{iz}$ leads us to the equation

$$\psi'(z + 2\pi) = ie^{iz}\psi(z) \tag{1.3}$$

and transforms the Hurwitz problem into the question whether there exists real-analytic solutions $\psi(z) \not\equiv Cexp\ e^{iz}$? The example by H.Lewy and results by Naftalevich give the negative answer to the question.

Our efforts to extend this information to homogeneous equations of more general form than Eq. (1.3) led us to the problem of solvability of Eq. (1.1). We prove that if the principal coefficients $a_{p_0 0}$ and $a_{p_m m}$ are nowhere degenerate entire functions and the growth of $a_{p_0 0}^{-1}$, $a_{p_m m}^{-1}$ $a_{jk}/a_{p_0 0}$ and $a_{jk}/a_{p_m m}$ along the real axis is not too fast then Equation (1.1) has an entire solution φ for each entire r.h.s. γ, and the space of entire solutions of the corresponding homogeneous equation is infinite dimensional.

It gives us the great pleasure to thank Professor Alexandre Eremenko, who initiated the present research, for very useful discussions.

2. Statement of results

We will formulate restrictions on the coefficients a_{jk} of Eq. (1.1) in terms of a real-valued increasing function $w(t) \geq 2$, $t \geq 0$, satisfying conditions

$$\lim_{t\to\infty} w(t) = +\infty,\ w(t+s) \leq w(t) + w(s), \quad \int_0^{\infty} \frac{w(t)}{1+t^2}dt < \infty. \tag{2.1}$$

The main result of the paper is following.

Theorem 1. *Let $a_{p_0 0}$ and $a_{p_m m}$ be nowhere vanishing entire functions such that*

$$\sup_{|\operatorname{Im} z|\leq h} \left(|a_{p_0 0}(z)|^{-1} + |a_{p_m m}(z)|^{-1}\right) \exp\left(-e^{w(|\Re z|)}\right) < \infty,$$

and let conditions

$$\sup_{|\mathrm{Im}\ z|\le h}\left(\sum_{j=0}^{p_0-1}\left|\frac{a_{j0}(z)}{a_{p_00}(z)}\right| + \sum_{j=0}^{p_m-1}\left|\frac{a_{jp_m}(z)}{a_{p_mm}(z)}\right|\right)e^{-w(|\Re\, z|)} < \infty$$

$$\sup_{|\mathrm{Im}\ z|\le h}\left(\sum_{k=1}^{m}\sum_{j=0}^{p_k}\left|\frac{a_{jk}(z)}{a_{p_00}(z)}\right| + \sum_{k=0}^{m-1}\sum_{j=0}^{p_k}\left|\frac{a_{jk}(z)}{a_{p_mm}(z)}\right|\right)\exp\left(-e^{w(|\Re\, z|)}\right) < \infty$$

be fulfilled for every $h > 0$. *Then*

i) *Equation* (1.1) *has an entire solution* φ *for every entire function* γ;

ii) *The space of entire solutions of the corresponding homogeneous equation is infinite dimensional.*

It is easy to see that conditions of Theorem 1 are fulfilled for Eq. (1.2) with $w(t) = n\ln(t+1)$, $n = \deg\ P$.

To prove part (i) of Theorem 1 we split an arbitrary entire function γ in a sum $\gamma = \gamma_+ + \gamma_-$ of entire functions γ_+ and γ_- vanishing faster than a permitted growth of coefficients $a_{jk}(z)$ as $\Re\, z \to +\infty$ and $\Re\, z \to -\infty$, respectively. At the next step we adjust equations (1.1) with functions $\gamma = \gamma_+$ and $\gamma = \gamma_-$ to iterations according to the directions of their decay. At last, we prove that the corresponding Neumann series converge to entire solutions φ_+ and φ_- of respective equations which produces the entire solution $\varphi = \varphi_+ + \varphi_-$ of Eq. (1.1).

The idea to split the function γ from Eq. (1.1) in a sum of functions vanishing in opposite directions was first used by Naftalevich [4] in the study of solvability of difference equations in entire functions. We applied the same method to difference equations in classes of smooth and real-analytic functions (see [6] and references therein). Here we notice that restrictions imposed by this method on the growth of coefficients of Eq. (1.1) are dictated by the dynamical properties of the shifts and do not involve their behavior along the rays transversal to the real axis. It may be modified to relax conditions of Theorem 1 and to treat the above equations with complex shifts α_k; the related results will be published elsewhere.

To prove part (ii) of Theorem 1 we start from a system $\{\psi_s\}_{s=0}^{l}$ of entire functions (e.g., polynomials) with the Jacobian

$$\det\ \|\psi_s^{(p)}(0)\|_{p,s=0}^{l} = 1$$

and satisfying Eq. (1.1) up to the order $l+1$ at $z = 0$:

$$\gamma_s(z) \equiv \sum_{k=0}^{m}\sum_{j=0}^{p_k} a_{kj}(z)\psi_s^{(j)}(z+\alpha_k) = o(z^{l+1}).$$

Estimates of a solution to a non-homogeneous Eq. (1.1) derived in the proof of part (i) permit us to find solutions φ_s of Eq. (1.1) with $\gamma = \gamma_s$ so small at $z = 0$ that

$$\det\ \|\psi_s^{(p)}(0) - \varphi_s^{(p)}(0)\|_{p,s=0}^{l} \neq 0.$$

Thus $\{\psi_s - \varphi_s\}_{s=0}^{l}$ is a linear independent system of solution of the homogeneous equation (1.1) completing the proof of Theorem 1.

3. Decomposition of an entire function

In this section we decompose an entire function γ in a sum of entire functions γ_+ and γ_- which are vanishing, as $z \to \infty$ in each semi-strip of half-planes $P_+ = \{ z : \Re z \geq 0 \}$ and $P_- = \{ z : \Re z \leq 0 \}$, respectively.

It is well known [7] that a function holomorphic and bounded in a semi-strip cannot decay too fast in it. The best possible rate of their decay may be described in terms of real-valued strictly monotonic continuous functions $w(t)$, $t \geq 0$, satisfying conditions (2.1). We will call every such function w a *weight function* and, without loss of generality, assume

$$w(0) > 2, \quad w(t) \geq 4\ln(t+1). \tag{3.1}$$

The required decomposition of γ is based on the following auxiliary statement.

Lemma 1. *For every weight function $w(t)$ there exists an entire function $\Omega(z)$ such that for every positive number H an estimate*

$$\begin{gathered} |\Omega(z)| < C(H)\exp(-\exp w(|\Re z|)), \\ |\operatorname{Im} z| \leq H \end{gathered} \tag{3.2}$$

holds with a coefficient $C(H)$ not depending on z.

Proof. Let us set $w^*(t) = 4w(t) + 1$ and define a sequence $\Lambda = \{\lambda_n\}_{n_0+1}^{\infty}$ of positive numbers by the equation

$$w^*(\lambda_n) = n, \quad n > [w^*(0)] = n_0.$$

Since w^* is monotonic, we have

$$n \leq w^*(t) \leq n+1, \quad \lambda_n \leq t \leq \lambda_{n+1},$$

and the counting function $n(t) = \{\#\lambda_k : \lambda_k \leq t\}$ of Λ satisfies the inequalities

$$w^*(t) - 1 \leq n(t) \leq w^*(t), \quad t \geq 0.$$

Condition (2.1) implies

$$\int_{n_0}^{\infty} \frac{n(t)}{t^2} dt < \infty \qquad \text{and} \qquad \int_{x}^{\infty} \frac{n(t)}{t^2} dt \geq \frac{n(x)}{x},$$

and hence $\lim\limits_{x\to+\infty} n(x)/x = 0$. Therefore

$$\theta(z) = \prod_{n=n_0+1}^{\infty} \left(1 + \frac{z^2}{\lambda_n^2}\right)$$

is an entire function of zero order [7]. Let us set

$$\Omega(z) = e^{-\theta(z)}$$

and check the property (3.2).

Assume first $z = x \geq \lambda_{n_0}$. Using integration by parts we find

$$\begin{aligned}
\ln\,\theta(x) &= \sum_{n=n_0+1}^{\infty} \ln\Big(1 + \frac{x^2}{\lambda_n^2}\Big) = \int_{n_0}^{\infty} \ln\left(1 + \frac{x^2}{t^2}\right) dn(t) \\
&= \ln\big(1 + \frac{x^2}{t^2}\big) n(t)\big|_{n_0}^{\infty} + 2x^2 \int_{n_0}^{\infty} \frac{n(t)}{(t^2 + x^2)t} dt \\
&\geq 2x^2 \int_{x}^{\infty} \frac{n(t)}{t(t^2 + x^2)} dt \geq 2n(x) \int_{x}^{\infty} \frac{dt}{t(t^2 + 1)} \\
&\geq \frac{n(x)}{2} \geq \frac{w^*(x) - 1}{2} = 2w(x).
\end{aligned}$$

For arbitrary $z = x + iy,\ x > 0,\ |y| \leq H$, we set $\theta(z) = \theta(x)\sigma(z)$ and obtain

$$\sigma(z) = \prod_{n=n_0+1}^{\infty} \frac{1 + \frac{x^2 - y^2 + 2ixy}{\lambda_n^2}}{1 + \frac{x^2}{\lambda_n^2}} = \prod_{n=n_0+1}^{\infty} \left(1 + \frac{2ixy - y^2}{\lambda_n^2 + x^2}\right).$$

Since

$$\Big|\frac{2ixy - y^2}{t^2 + x^2}\Big| \leq \frac{2xH + H^2}{x^2} \leq \frac{3H}{x} \leq \frac{3}{4}, \quad x \geq 4H,$$

we have

$$\begin{aligned}
\ln\,\sigma(z) &= \int_{n_0}^{\infty} \ln\left(1 + \frac{2ixy - y^2}{t^2 + x^2}\right) dn(t) \\
&= 2(2ixy - y^2) \int_{n_0}^{\infty} \frac{tn(t)}{(1 + \frac{2ixy - y^2}{t^2 + x^2})(t^2 + x^2)^2} dt
\end{aligned}$$

and

$$|\ln\,\sigma(z)| \leq 18Hx \int_{n_0}^{\infty} \frac{tn(t)}{(t^2 + x^2)^2} dt, \quad x \geq 4H.$$

Furthermore,

$$x \int_{n_0}^{x} \frac{tn(t)}{(t^2 + x^2)^2} dt \leq xn(x) \int_{0}^{x} \frac{t}{(t^2 + x^2)^2} dt = \frac{n(x)}{4x}$$

and

$$x \int_{x}^{\infty} \frac{tn(t)}{(t^2 + x^2)^2} dt \leq x \sup_{t \geq x} \frac{t^3}{(t^2 + x^2)^2} \int_{x}^{\infty} \frac{n(t)}{t^2} dt \leq \int_{x}^{\infty} \frac{n(t)}{t^2} dt$$

proving

$$\lim_{\Re z \to \infty,\, |\Im\, z| \leq H} \ln\,\sigma(z) = 0, \qquad \lim_{\Re z \to \infty,\, |\Im\, z| \leq H} \sigma(z) = 1.$$

At last

$$\begin{aligned}\ln|\Omega(z)| &= -\Re\theta(z) = -\theta(x)\Re\sigma(z) \le -e^{2w(x)}(1+o(1))\\ &= -e^{w(x)} + \big(e^{w(x)} - e^{2w(x)}(1+o(1))\big).\end{aligned}$$

Thus, for all sufficiently large values of $|x|$ we have $\ln|\Omega(z)| \le -e^{w(|x|)}$ and the estimate (3.2) follows.

For a given real-valued function $p(t) > 0$, $t \in \mathbb{R}$, we denote by $\mathcal{E}(p)$ the Frechêt space of entire functions endowed with the norms

$$\|\varphi\|_{p,S(h)} = \sup_{z\in S(h)} |\varphi(z)|\, e^{-p(|\Re z|)}$$

where

$$S(h) = \{\, z : |\Im z| \le h \,\}, \quad h \in \mathbb{R}_+. \tag{3.3}$$

In addition, we denote by $\mathcal{E}_+(p)$ the space of entire functions with the norms

$$\|\varphi\|_{p,S_+(a,h)} = \sup_{z\in S_+(a,h)} |\varphi(z)|e^{-p(\Re z)}$$

where

$$S_+(a,h) = \{\, z : \Re z \ge a, \quad |\Im z| \le h \,\}, \quad a \in \mathbb{R}, \ h \in \mathbb{R}_+.$$

At last, we set $\mathcal{E}_-(p) = \{\, \varphi : \varphi(-z) \in \mathcal{E}_+(p) \,\}$ and

$$\|\varphi\|_{p,S_-(a,h)} = \sup_{z\in S_-(a,h)} |\varphi(z)|e^{-p(-\Re z)}$$

where

$$S_-(a,h) = \{\, z : \Re z \le a, \quad |\Im z| \le h \,\}, \quad a \in \mathbb{R}, \ h \in \mathbb{R}_+.$$

In what follows we will need the above spaces corresponding to $p = w$, $p = \exp w$ and $p = -\exp w$ with weight functions w extended to the whole real axis as a positive monotonic functions.

Theorem 2. *Let w be a weight function and let l be an integer. Then for every entire function γ with $\gamma(0) = \gamma'(0) = \cdots = \gamma^{(l)}(0) = 0$ there exist entire functions γ_+ and γ_- with*

$$\gamma_+(0) = \gamma_-(0) = \gamma'_+(0) = \gamma'_-(0) = \cdots = \gamma^{(l)}_+(0) = \gamma^{(l)}_-(0) = 0$$

such that $\gamma = \gamma_+ + \gamma_-$ and the inclusions are valid

$$\gamma_+ \in \mathcal{E}_+(-e^w), \quad \gamma_- \in \mathcal{E}_-(-e^w). \tag{3.4}$$

In the proof we will use the following proposition well known in the mathematical folklore. We give its proof here for a completeness of exposition, moreover that, except Problem 1 from Chapter 1 in the monograph [7], we have no satisfactory reference to it at our disposition.

Lemma 2. *Let $M(r), r \ge 0$, be a strictly logarithmically convex function. Then there exists an entire function $\mu(z)$ such that*

$$\mu(r) \ge M(r), \qquad r \ge 0. \tag{3.5}$$

Proof. Since $M(r)$ is a strictly logarithmically convex function, for every integer $n \geq 0$ there exists the unique number r_n such that

$$\ln M(r) \geq \ln M(r_n) + n(\ln r - \ln r_n), \quad r > 0,$$

and

$$\inf_{r \geq 0} \frac{M(r)}{r^n} = \frac{M(r_n)}{r_n^n}.$$

Assume that for some integer $n > 0$ there exists a logarithmically convex function $\tilde{M}(r)$, $0 \leq r \leq r_n$, such that

$$\min_{r_n \geq r > 0} \frac{\tilde{M}(r)}{r^k} = \frac{\tilde{M}(r_k)}{r_k^k}, \quad 0 \leq k \leq n,$$

and conditions

$$\tilde{M}(r_k) \geq M(r), \quad r_k \leq r \leq r_{k+1}, \quad 0 \leq k \leq n-1,$$

are satisfied. We set

$$\tilde{M}(r_{n+1}) = \max\Big\{M(r_{n+1}), \tilde{M}(r_n)\Big(\frac{r_{n+1}}{r_n}\Big)^n\Big\} + 1$$

and define $\tilde{M}(r)$ for $r \in [r_n, r_{n+1}]$ in such a way that it is a logarithmically convex function in $[0, r_{n+1}]$ and

$$\min_{r_{n+1} \geq r > 0} \frac{\tilde{M}(r)}{r^k} = \frac{\tilde{M}(r_k)}{r_k^k}, \quad 0 \leq k \leq n+1.$$

By induction we obtain a logarithmically convex function $\tilde{M}$ on the semi-axis $\mathbb{R}_+$ for which relations

$$\inf_{r > 0} \frac{\tilde{M}(r)}{r^k} = \frac{\tilde{M}(r_k)}{r_k^k}; \quad \tilde{M}(r_k) \geq M(r), \quad r_k \leq r \leq r_{k+1},$$

are fulfilled for $k = 0, 1, 2, \ldots$.

Let us now define the entire function

$$\mu(z) = \sum_{k=0}^{\infty} c_k z^k, \quad c_k = \inf_{r > 0} \frac{\tilde{M}(r)}{r^k}.$$

For arbitrary positive r we find an integer n such that $r_n \leq r < r_{n+1}$ and obtain

$$\mu(r) \geq c_n r_n = \frac{\tilde{M}(r_n)}{r_n^n} r^n \geq M(r), \quad r > 0,$$

proving Eq. (3.5).

Proof of Theorem 2. Under assumptions of Theorem 2, let Ω be an entire function satisfying conditions (3.2) and let

$$M(r) \;=\; M(r,\gamma)M(r,\Omega)$$

where

$$M(r,\varphi) \;=\; \max_{|z|=r}|\varphi(z)|.$$

According to the Hadamard Theorem [8] the function $M(r)$ is strictly logarithmically convex and by Lemma 2 there exists an entire function $\hat{\mu}(z)$ satisfying the lower estimate such that $\hat{\mu}(r) \;\geq\; M(r),\; r>0$. We set

$$\mu(z) \;=\exp\,\hat{\mu}(\cos^2 z)$$

and obtain

$$\mu(it)=\hat{\mu}(\cosh^2 t)\geq M(\cosh^2 t)\geq M(|t|),\quad t\in\mathbb{R}.$$

The entire function $\mu(z)$ is 2π-periodic and hence bounded in every strip $S(h)$. We fix an arbitrary number $N>0$ and define the functions γ_+ and γ_- by the formulae

$$\gamma_{\pm}(z) \;=\; \pm z^{l+1}e^{-Nz^2}\Omega(z)\mu(z)\frac{1}{2\pi i}\int\limits_{\Re\, z=0}\frac{t^{-(l+1)}e^{Nt^2}\Omega^{-1}(t)\mu^{-1}(t)\gamma(t)}{t-z}dt \qquad (3.6)$$

in the half-planes P_+ and P_-, respectively.

The numerator of the integrand here is an entire function, which permits us to replace an interval $\{\Re t=0,\;|\Im t|\leq b\}$ from the integration path by the semicircle $\{\;t\;:|t|=b,\;\Re\, t\;\leq\;0\,\}$ and to extend analytically the function γ_+ to the domain on the right side of the deformed integration path. Since b is an arbitrary positive number, it means that γ_+ is extended from P_+ to the complex plane as an entire function.The similar arguments prove that γ_- is an entire function as well. The Cauchy formula yields $\gamma \;=\; \gamma_+ \;+\; \gamma_-$. At last, inclusions (3.4) immediately follow from the properties of the function Ω described in Lemma 2.

4. Auxiliary equation

The general equation (1.1) may be represented in the operator form

$$\mathcal{L}\varphi \;=\; \gamma$$

where

$$\mathcal{L}\varphi(z) \;=\; \sum_{k=1}^{m}\Big(\;\sum_{j=0}^{p_k} f_{jk}(z)\varphi^{(j)}(z\;+\;\alpha_k)\Big) \qquad (4.1)$$

In this section we consider an equation

$$\varphi \;=\; \mathcal{L}_+\varphi \;+\; \gamma \qquad (4.2)$$

with

$$\mathcal{L}_+\varphi(z) \;=\; \sum_{k=1}^{m}\Big(\;\sum_{j=0}^{P} f_{jk}(z)\varphi^{(j)}(z\;+\;\alpha_k)+\;g_k(z)\int\limits_z^{+\infty} h_k(t)\varphi(t\;+\;\alpha_k)dt\;\Big). \quad (4.3)$$

We assume that the integrand here decays along every ray $R_+(a,h)=\{z:\Re z\geq a,\Im z=h\}$ and the path of integration is $R_+(\Re z,\Im z)$.

In what follows we set

$$\Delta=\alpha_m,\quad \delta=\min\{\alpha_1,\alpha_m-\alpha_{m-1}\},\quad \kappa=\Delta\delta^{-1}.$$

Theorem 3. *Assume that w and w^* are weight functions such that*

$$w^*(t)\;=\;4w(2\kappa t),\quad t\in\mathbb{R}, \quad (4.4)$$

that f_{jk}, g_k, $h_k\;\in\;\mathcal{E}_+(\exp\,w)$, $k\;=\;1,\,\ldots,\,m$, $j\;=\;0,\,\ldots,\,P$, and that $\gamma\;\in\;\mathcal{E}_+(-\exp\,w^*)$. Then there exists an entire solution φ of Eq. (4.2) such that for every $a\;\in\;\mathbb{R}$, $h\;\in\;\mathbb{R}_+$, $\tau\;\in\;\mathbb{R}_+$ an estimate

$$|\varphi(z)|\;\leq\;C(a,h,\tau)\|\gamma\|_{-\exp\,w^*,S_+(a-\tau,h+\tau)}e^{-\exp\,w(\Re\,z)},\quad z\;\in\;S_+(a,h) \quad (4.5)$$

holds.

Proof. To simplify the following calculations we introduce some notation.

First, the symbol A^n will stay for a vector $(a_1,\,\ldots,\,a_n)\;\in\;\mathbb{Z}^n$. In particular, we set

$$O^n\;=\;\underbrace{\{0,\,\ldots,\,0\,\}}_{n\text{ times}},\quad U^n\;=\;\underbrace{\{1,\,\ldots,\,1\,\}}_{n\text{ times}},\quad M^n\;=\;\underbrace{\{m,\,\ldots,\,m\,\}}_{n\text{ times}}$$

where m is the integer from Eq. (4.1).

Furthermore, for every vector $A^n\;\in\;\mathbb{Z}^n$ we set

$$|A^n|\;=\;\sum_{k=1}^{n}|a_k|$$

and if $B^n\;=\;\underbrace{\{b_1,\,\ldots,b_n\,\}}_{n\text{ times}}\in\mathbb{Z}^n$ is another vector we write $A^n\;\prec\;B^n$ if $a_k\;\leq\;b_k$ for all $k\;=\;1,\,\ldots,\,n$.

Given vectors $K^n,J^n\;\in\;\mathbb{Z}^n,U^n\;\prec\;K^n\;\prec\;M^n$, let $\sigma_p(z;K^n,J^n)$ and $\theta_p(z;K^n,J^n)$, $p=1,\ldots,n$, be one of the functions

$$f_{jk}^{(\nu)}\left(z\;+\;\sum_{s=1}^{p-1}\alpha_{k_s}\right),\quad g_k^{(\nu)}\left(z\;+\;\sum_{s=1}^{p-1}\alpha_{k_s}\right),\quad h_k^{(\nu)}\left(z\;+\;\sum_{s=1}^{p-1}\alpha_{k_s}\right),$$

where $1\;\leq\;k\;\leq\;m$, $0\;\leq\;j\;\leq\;P$; $0\;\leq\;\nu\;\leq\;|J^n|$. Denote by $\sigma_p(.;K^n,J^n)$ and $\theta_p(.;K^n,J^n)$ operators of multiplication by $\sigma_p(z;K^n,J^n)$ and $\theta_p(z;K^n,J^n)$, respectively, and set

$$I_+\varphi(z)=\int\limits_z^{+\infty}\varphi(t)\,dt.$$

With $R^n = \{r_1, \dots, r_n\}$ such that $O^n \prec R^n \prec U^n$ we define the operators

$$\Phi(K^n, J^n, R^n) = \prod_{p=1}^{n} \sigma_p(.; K^n, J^n) I_+^{r_p} \theta_p(.; K^n, J^n) \tag{4.6}$$

where the factors are ordered from the left to the right according to the growth of p from $p = 1$ through $p = n$. Let us show now that

$$\mathcal{L}_+^n = \sum_{U^n \prec K^n \prec M^n} \sum_{j=0}^{nP} F_j(K^n) D^j T(K^n), \ D = \frac{d}{dz}, \tag{4.7}$$

where

$$T(K^n)\varphi(z) = \varphi\Big(z + \sum_{s=1}^{n} \alpha_{k_s}\Big)$$

and

$$F_j(K^n) = \sum_{|J^n|=j} \sum_{O^n \prec \mathbb{R}^n \prec U^n} c(K^n, J^n, R^n) \Phi(K^n, J^n, R^n) \tag{4.8}$$

with real-valued coefficients $c(K^n, J^n, R^n)$ satisfying estimates

$$C_n \equiv \sum_{U^n \prec K^n \prec M^n} \sum_{|J^n|=0}^{nP} \sum_{O^n \prec R^n \prec U^n} |c(K^n, J^n, R^n)| \le Ce^{n^3}. \tag{4.9}$$

Indeed, for $n = 1$ the representation (4.8) is an immediate consequence of (4.3) and we can accept $C_1 = (m+1)(P+1)$.

For arbitrary $n > 1$ and $j \ge 0$ we have

$$D^j T(K^n) \mathcal{L}_+ \psi(z)$$

$$= D^j \sum_{k_{n+1}=1}^{m} \Big(\sum_{j_{n+1}=0}^{P} f_{j_{n+1},k_{n+1}} \left(z + \sum_{s=1}^{n} \alpha_{k_s} \right) \psi^{(j_{n+1})} \left(z + \sum_{s=1}^{n+1} \alpha_{k_s} \right)$$

$$+ g_{k_{n+1}} \left(z + \sum_{s=1}^{n} \alpha_{k_s} \right) \int_z^{+\infty} h_{k_{n+1}} \left(t + \sum_{s=1}^{n} \alpha_{k_s} \right) \psi \left(t + \sum_{s=1}^{n+1} \alpha_{k_s} \right) dt \Big).$$

The r.h.s of the last equation is a sum of $(m+1)(P+1)(|J^n|+1)^2$ expressions of one of the following forms

(i) $$\binom{l}{j} f_{j_{n+1},k_{n+1}}^{(l)} \left(z + \sum_{s=1}^{n} \alpha_{k_s} \right) \psi^{(j-l+j_{n+1})} \left(z + \sum_{s=1}^{n+1} \alpha_{k_s} \right),$$

(ii) $$g_{k_{n+1}}^{(j)} (z + \sum_{s=1}^{n} \alpha_{k_s}) \int_z^{\infty} h_{k_{n+1}} \left(t + \sum_{s=1}^{n} \alpha_{k_s} \right) \psi \left(t + \sum_{s=1}^{n+1} \alpha_{k_s} \right) dt,$$

(iii) $$-\binom{l}{j} \binom{r}{j-l-1} g_{k_{n+1}}^{(l)} \left(z + \sum_{s=1}^{n} \alpha_{k_s} \right)$$
$$\times h_{k_{n+1}}^{(r)} \left(z + \sum_{s=1}^{n} \alpha_{k_s} \right) \psi^{(j-l-r-1)} \left(z + \sum_{s=1}^{n+1} \alpha_{k_s} \right)$$

with $0 \le l \le j-1$ in (i) and $0 \le r \le j-l-1$, $0 \le l \le j-1$, in (iii). Assuming that Eq. (4.8) is valid for some $n \ge 1$ we find that it is valid for $n+1$ with $K^{n+1} = \{\underbrace{k_1, \ldots, k_n}_{n \text{ times}}, k_{n+1}\} = \{K^n, k_{n+1}\}$, $1 \le k_{n+1} \le m$; $|J^{n+1}| \le |J^n| + P \le P(n+1)$; $R^{n+1} = \{R^n, r_{n+1}\}$, $0 \le r_{n+1} \le 1$, and with operator $\Phi(K^{n+1}, J^{n+1}, R^{n+1})$ obtained by multiplying $\Phi(K^n, J^n, R^n)$ from the right by one of the following operators

$$\text{(i)} \quad f^{(l)}_{j_{n+1},k_{n+1}}\left(\cdot + \sum_{s=1}^{n} \alpha_{k_s}\right), \qquad 0 \le l \le Pn,$$

$$\text{(ii)} \quad g^{(j)}_{k_{n+1}}\left(\cdot + \sum_{s=1}^{n} \alpha_{k_s}\right) I_+ h_{k_{n+1}}\left(\cdot + \sum_{s=1}^{n} \alpha_{k_s}\right), \quad 0 \le j \le Pn,$$

$$\text{(iii)} \quad g_{k_{n+1}}\left(\cdot + \sum_{s=1}^{n} \alpha_{k_s}\right) h^{(r)}_{k_{n+1}}\left(\cdot + \sum_{s=1}^{n} \alpha_{k_s}\right), \qquad 0 \le r \le j-l-1.$$

It is easy to see that the number of these operators does not exceed $3(Pn)^2$. The elementary inequality

$$\begin{pmatrix} l \\ j \end{pmatrix} + \begin{pmatrix} l \\ j \end{pmatrix} \begin{pmatrix} r \\ j-l-1 \end{pmatrix} \le 2^{2Pn+1},$$

yields

$$C_{n+1} \le 2C_n(3Pn)^2 2^{2Pn} \le Ce^{(n+1)^3}$$

with

$$C = C_1 \sup_{n \ge 1} (6Pn)^{2n} 2^{2Pn^2} e^{-(n+1)^3},$$

proving (4.9) for all $n \ge 1$.

Let now real numbers $a, h > 0$ and τ be fixed such that $0 < \tau < \delta/2$. With vectors K^n, J^n, $R^n \in \mathbb{Z}^n_+$ and functions f_{jk}, g_k, h_k from Eq. (4.3) also being fixed we have, using the Cauchy formulae for derivatives of analytic functions,

$$\begin{aligned} &|\sigma_p(z; K^n, J^n)| + |\theta_p(z; K^n, J^n)| \\ &\le C(a, h, \tau) \frac{|J^n|!}{\tau^{|J^n|}} \exp e^{w(\Re\, z + (n-1)\Delta + \tau)}, \quad z \in S_+(a, h). \end{aligned} \tag{4.10}$$

Furthermore, let $n_0 = 2|a-\tau|\delta^{-1}$. Then for $t \ge a - \tau$ and $n \ge n_0$ we have

$$\begin{aligned} &2\Delta\delta^{-1}(t-\tau) + 2n\Delta - ((t-\tau) + n\Delta) \\ &\ge n\Delta - (2\Delta\delta^{-1})|a-\tau| \ge (n-n_0)\Delta \ge 0 \end{aligned} \tag{4.11}$$

and

$$\begin{aligned} &2\Delta\delta^{-1}(t-\tau) + 2n\Delta \\ &\ge n\Delta + (n\Delta - 2\Delta\delta^{-1}|a-\tau|) \ge n\Delta + (n-n_0)\Delta \ge n\Delta. \end{aligned} \tag{4.12}$$

Assume now $\gamma \in \mathcal{E}(-e^{w^*})$ and estimate the function

$$\Phi(K^n, J^n, R^n)\gamma(z), \qquad z \in S_+(a-\tau, h+\tau).$$

We use the monotonicity of w and w^* and estimates Eq. (4.10) to obtain an inequality

$$\begin{aligned}&\left|\Phi(K^n,J^n,R^n)\gamma^{(|J^n|)}\left(z+\sum_{s=1}^{n}\alpha_{k_s}\right)\right|\\&\le\left(C(a,h,r)\frac{|J^n|!}{\tau^{|J^n|}}\right)^{3n}\cdot\|\gamma\|_{-\exp w^*,\,S_+(a-\tau,h+\tau)}\int\limits_{\Re z}^{+\infty}dt_1\dots dt_{r-1}\int\limits_{t_{r-1}}^{+\infty}dt_r e^{E(t_r,n)}\end{aligned}$$

where $r=|R^n|\le n$ and

$$E(t,n)=2ne^{w(t+n\Delta-\tau)}-e^{w^*(t+n\delta-\tau)}.$$

According to (4.11), (4.12) and the second inequality from (3.1) we have

$$\begin{aligned}w^*(t-\tau+n\delta)&\ge w(2\Delta\delta^{-1}(t-\tau)+2n\Delta)+w(n\Delta)\\&\ge w(t-\tau+n\Delta)+4\ln(n\Delta)\end{aligned}$$

and using the first inequality from (3.1) and (4.4) we obtain

$$\begin{aligned}E(t,n)&\le-e^{w^*(t+n\delta-\tau)}(1-2/\Delta^2)\\&\le-e^{3w(2\Delta\delta^{-1}(t-\tau)+2n\Delta)}\le-3e^{w(2\Delta\delta^{-1}(t-\tau)+2n\Delta)}.\end{aligned}$$

It is easy to see that

$$\sup_{t\ge a-\tau}(t^2-e^{w(t)})=C(a,\tau)<\infty,$$

$$\int\limits_{\Re z}^{+\infty}dt_1\dots dt_{r-1}\int\limits_{t_{r-1}}^{+\infty}e^{-t_r^2}dt_r\le\int\limits_{a}^{+\infty}\frac{(t-a)^{r-1}}{(r-1)!}e^{-t^2}dt\le C(a)<\infty.$$

At last,

$$w(2\Delta\delta^{-1}(t-\tau)+2n\Delta)\ge w(n\Delta)\ge 4\ln n$$

and, since $0<\tau<n\delta$,

$$w(2\Delta\delta^{-1}(t-\tau)+2n\Delta)\ge w(t-\tau+n\delta)\ge w(t),$$

giving us an estimate

$$\begin{aligned}&|\Phi(K^n,J^n,R^n)D^{(|J^n|)}T(K^n)\gamma(z)|\\&\le C(a,h,\tau)^n n^{3Pn^2}e^{-n^4}\|\gamma\|_{-\exp w^*,S_+(a-\tau,h-\tau)}\exp(-e^{w(\Re z)}),\end{aligned}\tag{4.13}$$

for all $n\ge n_0$, $z\in S_+(a,h)$. It is evident that the same estimates are valid for $n<n_0$ and sufficiently large values of $\Re z$. Now we note that there are not more than $n_0\le 2\delta^{-1}(|a|+1)$ indices $n<n_0$ and hence Eq. (4.13) is fulfilled for all $n\ge 0$ with some number $C(a,h,\tau)$.

Taking into account estimates (4.9) we conclude that the series

$$\sum\nolimits_{n=0}^{+\infty}\mathcal{L}_+^n\gamma$$

converges to an entire solution φ of Eq. (4.2), and Eq. (4.5) holds.

5. Proof of Theorem 1, Part (i)

First we set

$$\tilde{a}_{j0} = a_{p_0 0}^{-1} a_{j0}, \qquad 0 \le j \le p_0;$$
$$\tilde{a}_{jk} = -a_{p_0 0}^{-1} a_{jk}, \qquad 1 \le k \le m, \quad 0 \le j \le p_k,$$

and represent Eq. (1.1) in the form

$$\mathcal{L}_0 \varphi(z) \;=\; \sum_{k=1}^{m} \mathcal{L}_k \varphi(z + \alpha_k) \;+\; a_{p_0 0}^{-1}(z)\gamma(z) \tag{5.1}$$

where $\mathcal{L}_k \;=\; \sum\limits_{j=0}^{p_k} \tilde{a}_{jk} D^j, \quad k = 1, \dots, m.$

Let $K_0(z,t)$ be the Cauchy kernel of operator $\mathcal{L}_0$. It means that if ψ is an entire function vanishing sufficiently fast as $\Re\, z \;\to\; +\infty$ in every half-strip $S_+(a,h)$ then the function

$$\varphi(z) \;=\; -\int_{z}^{+\infty} K_0(z,t)\psi(t)dt \tag{5.2}$$

is a solution of the equation $\mathcal{L}_0\varphi \;=\; \psi$. On the other hand, if ψ is an entire solution of the equation

$$\psi(z) \;= -\sum_{k=1}^{m} \mathcal{L}_k \int_{z+\alpha_k}^{+\infty} K_0(z \;+\; \alpha_k, t)\psi(t)dt \;+\; a_{p_0 0}^{-1}(z)\gamma(z) \tag{5.3}$$

then the function φ defined (formally, at least) by (5.2) is a solution of Eq. (5.1).

To find an entire solution of Eq. (5.3) we use (cf., [9]) the explicit representation

$$K_0(z,t) \;=\; \begin{vmatrix} u_1(t) & \dots & u_{p_0}(t) \\ \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot \\ u_1^{(p_0-1)}(t) & \dots & u_{p_0}^{(p_0-2)}(t) \\ u_1(z) & \dots & u_{p_0}(z) \end{vmatrix} \exp\left(-\int_0^z a_{p_0-1,0}(s)ds \right).$$

where $\{\, u_j(z) \,\}_{j=1}^{p_0}$ is the fundamental system of solutions of the homogeneous equation $\mathcal{L}_0 u = 0$, normalized by the unit matrix of initial conditions at $z = 0$. It follows from the Gronwall inequality [10], that the functions of this system satisfy the inequality

$$|u_j(z)| \le \exp\left(p_0 |z| e^{w(\Re z)}\right)$$

where w is the weight function from Theorem 1. Hence

$$K_0(z,t) \;=\; \sum_{j=1}^{p_0} u_j(z) v_j(t) \tag{5.4}$$

where $u_j,\; v_j \in \mathcal{E}_+(\exp(2w)), \quad j = 1, \dots, p_0.$

It is easy to see that every expression

$$\frac{d^j}{dz^j}\Big(u(z)\int\limits_z^\infty v(t)\psi(t)dt\Big)$$

is a linear combination of functions

$$u^{(p)}(z)v^{(q)}(z)\psi^{(r)}(z), \quad p,q,r\,<\,j; \quad u^{(j)}(z)\int\limits_z^\infty v(t)\psi(t)dt.$$

Therefore Eq. (5.3) may be written in the form

$$\psi(z)\,=\,\mathcal{L}_+\psi(z)\,+\,a_{p_0 0}^{-1}(z)\gamma(z) \tag{5.5}$$

where $\mathcal{L}_+$ is an operator defined by Eq. (4.3) whose coefficients are entire functions $f_{jk}(z)$, $g_k(z)$, $h_k(z)$ belonging to the linear span of functions $\tilde a_{jk}(z)u_s^{(p)}(z\,+\,\alpha_k)v_s^{(p)}(z\,+\,\alpha_k)$, $j,p\,\le\,\max\limits_{1\le k\le m} p_k\,=\,P$. It is easy to see that $f_{jk},\ g_k,\ h_k\,\in\,\mathcal{E}_+(\exp\tilde w)$ with $\tilde w = 3w$. With a function w (and hence $\tilde w$) being fixed we define $\tilde w^*(t) = 4\tilde w(2\kappa t)$ and represent γ in the form $\gamma = \gamma_+ + \gamma_-$ where the summands γ_+ and γ_- are given by Eq. (3.6) in which w is replaced by $\tilde w^*$ and N is a fixed number.

Since $\tilde w^* > w^*$, Theorem 3 implies that there exists an entire solution $\psi_+(z)$ of Eq. (4.2) satisfying

$$|\psi_+(z)|\,<\,C(a,h,\tau)\|a_{p_0 0}^{-1}\gamma_+\|_{\,-\exp w^*,\,S_+(a-\tau,\,h+\tau)}\exp(-e^{\tilde w(\Re z)}),$$
$$z\,\in\,S_+(a,h).$$

The function

$$\varphi_+(z)\,=\,-\int\limits_z^\infty K_0(z,t)\psi_+(t)dt$$

is an entire solution of Eq. (1.1) with $\gamma\,=\,\gamma_+$ and the previous inequality yields

$$|\varphi_+(z)|\,<\,C(a,h,\tau)\|a_{p_0 0}^{-1}\gamma_+\|_{\,-\exp w^*,\,S_+(a-\tau,\,h+\tau)}\exp(-e^{w(\Re z)}),$$
$$z\,\in\,S_+(a,h). \tag{5.6}$$

Our next step is to consider Equation (1.1) with $\gamma\,=\,\gamma_-$. First, we substitute $-z$ in it instead of z and set $\tilde\gamma_+(z)\,=\gamma_-(-z)$, $\tilde\varphi(z) = \varphi(-z+\alpha_m)$. Then $\varphi(-z+\alpha_k) = \varphi(-(z+\alpha_m-\alpha_k)+\alpha_m) = \tilde\varphi(z+\alpha'_{m-k})$ where$\alpha'_k = \alpha_m-\alpha_{m-k}$, $k = 0,1,\,\dots,\,m$, and we arrive at the equation

$$\sum_{k=0}^{m}\sum_{j=0}^{p_{m-k}}\tilde a_{jk}(z)\tilde\varphi^{(j)}(z+\alpha'_k)\,=\,\tilde\gamma_+(z) \tag{5.7}$$

with $\tilde a_{j\,k}(z)\,=\,(-1)^j a_{j,\,m-k}(-z)$. It is evident that

$$\alpha'_0\,=\,0\,<\,\alpha'_1\,<\,\cdots\,<\,\alpha'_m,\ \ \alpha'_1\,=\,\alpha_m-\alpha_{m-1}\,\ge\,\delta,\ \ \alpha'_m\,=\,\alpha_m\,=\,\Delta,$$

and $\tilde{\gamma}_+ \in \mathcal{E}_+(-e^w)$. Hence this is an equation of the type Eq. (4.2). According to (5.6) there exists its entire solution $\tilde{\varphi}_+$ satisfying

$$\begin{aligned} |\tilde{\varphi}_+(z)| \;&\le\; C(a,h,\tau)\|a_{p_m m}^{-1}\gamma_-\|_{-\exp w^*,\, S_-(-a+\tau,\, h+\tau)} \exp(-e^{w(\Re z)}) \\ z \;&\in\; S_-(-a,h) \;=\; \{\, z \,:\, \Re\, z \;\le\; -a, \;\; |\Im\, z| \le h.\} \end{aligned} \tag{5.8}$$

It remains to note that $\varphi_-(z) = \tilde{\varphi}_+(-z + \alpha_m)$ is an entire solution of Eq. (4.3) with $\gamma = \gamma_-$ and $\varphi = \varphi_+ + \varphi_-$ is an entire solution of Eq. (1.1), completing the proof of part (ii) of Theorem 3.

Combining estimates (5.8) and (5.6) we find that for every $\tau > 0$ and $R > 0$ there exists $C(R,\tau)$ such that, whatever is an entire function γ, there exists an entire solution φ of Eq. (1.1) satisfying inequalities

$$\begin{aligned} \max_{|z|\le R} |\varphi(z)| \le C(R,\tau)(&\|a_{p_0 0}^{-1}\gamma_+\|_{-\exp w^*,\, S_+(-R-\tau, R+\tau)} \\ &+ \|a_{p_m m}^{-1}\gamma_-\|_{-\exp w^*,\, S_-(R+\tau, R+\tau)}) \end{aligned}$$

The method used to prove part (i) of Theorem 1 permits us to relate effectively to every entire function γ an entire solution of Eq. (1.1). As a matter of fact, the correspondence between γ and φ depends on a choice of a function μ in Eq. (3.6) and is not linear with respect to it. We can slightly "improve" the situation by restricting γ in Eq. (1.1) to the subspace $\mathcal{F}(p)$ of entire functions endowed, in addition to the traditional system of norms

$$\|\gamma\|_R \;=\; \max_{|z|\le R} |\gamma(z)|, \quad R \;>\; 0,$$

with the norm

$$\|\gamma\|_p \;=\; \sup_{t\in\mathbb{R}} |\gamma(it)| p^{-1}(|t|)$$

where $p = p(t) > 0$, $t > 0$, is a fixed logarithmically convex function.

For a given equation $\mathcal{L}$ of the form (1.1) satisfying conditions of Theorem 1 we choose an arbitrary (but fixed!) number N and a nowhere degenerate entire function $\tilde{\mu}(z)$ such that

$$|\tilde{\mu}(it)| \;\ge\; p(|t|) M(|t|, e^\Omega).$$

The representation (3.6) defines a linear mapping $\gamma \to (\gamma_+, \gamma_-)$ from $\mathcal{F}(p)$ to the space of entire 2-vectors and it is easy to see that the solution $\varphi = \varphi_+ + \varphi_-$ depends linearly on a function γ. Moreover, since the function μ^{-1} is bounded on every compact set in $\mathbb{C}$, we obtain, in the notation used in the proof of part (i),

$$\begin{aligned} &\|a_{p_0 0}^{-1}\gamma_+\|_{-\exp w^*,\, S_+(-R-\tau,\, R+\tau)} \;+ \|a_{p_m m}^{-1}\gamma_-\|_{-\exp w^*,\, S_-(R+\tau,\, R+\tau)} \\ &\le C(R)(\|\gamma\|_{R+\tau} \;+\; \|\gamma\|_p), \end{aligned}$$

and hence

$$\|\varphi\|_R \;\le\; C(R)(\|\gamma\|_{R+\tau} \;+\; \|\gamma\|_p).$$

In other words, the correspondence $\gamma \to \varphi$ defines a bounded linear mapping from $\mathcal{F}(p)$ to the space of all entire functions and therefore is a right inverse to $\mathcal{L}$.

6. Proof of Theorem 1, Part (ii)

At the first step we construct a system of linear independent functions which solve the homogeneous equation $\mathcal{L}\varphi = 0$ up to a finite order at the point $z = 0$ only.

Lemma 3. *For every integer $l > 0$ there exists a system $\Psi = \{\psi_j\}_{j=0}^{l}$ of entire functions satisfying conditions*

$$\begin{aligned} \psi_j^{(s)}(0) &= \delta_{js} && j, s = 0, \dots, l; \\ \mathcal{L}\psi_j(z) &= o(z^{l+1}), && j = 0, \dots, l. \end{aligned}$$

Proof. Let us first assume that for a fixed integer s, $0 \le s \le l$, a function $\psi(z)$ satisfies conditions

$$\psi^{(j)}(\alpha_k) = \begin{cases} \delta_{js}, & k = 0, & j = 0, 1, \dots, p_0 + l + 1 \\ 0, & k = 1, \dots, m-1, & j = 0, \dots, p_k + l + 1 \\ 0, & k = m, & j = 0, \dots, p_m - 1. \end{cases} \tag{6.1}$$

Then

$$\begin{aligned} \mathcal{L}\psi(z) = a_{p_m m}(z)\psi^{(p_m)}(z + \alpha_m) \\ + \sum_{j=0}^{p_m-1} a_{jm}(z)\psi^{(j)}(z + \alpha_m) + \sum_{j=0}^{p_0} a_{j0}(z)\psi^{(j)}(z) + o(z^{l+1}), \end{aligned}$$

and hence the system of equations

$$(\mathcal{L}\psi(z))^{(j)}|_{z=0} = 0, \quad j = 0, \dots, l+1,$$

may be written in the form

$$\begin{cases} a_{p_m m}(0)\xi_0 & = \mu_0(\eta_0, \dots, \eta_{p_0}) \\ a_{p_m m}(0)\xi_j + \lambda_j(\xi_0, \dots, \xi_{j-1}) & = \mu_j(\eta_0, \dots, \eta_{p_0+j}), \quad 1 \le j \le l+1, \end{cases}$$

where $\xi_j = \psi^{(p_m+j)}(\alpha_m)$, $\eta_j = \psi^{(j)}(0)$ and λ_j and μ_j are linear functions. Since $a_{p_m m}(0) \neq 0$, for every fixed s, $0 \le s \le l$, and η_j defined by the first line of Eq. (6.1) the last system has the unique solution $\{\xi_j\}_{j=0}^{l+1}$. Denote now by $\psi_s(z)$ an entire function (e.g., polynomial) satisfying the following interpolation conditions

$$\psi_s^{(j)}(\alpha_k) = \begin{cases} \delta_{js}, & k = 0, & j = 0, \dots, p_0 + l + 1 \\ 0, & k = 2, \dots, p-1, & j = 0, \dots, p_k + l + 1 \\ 0, & k = m, & j = 0, \dots, p_m - 1 \\ \xi_{j-p_m}, & k = m, & j = p_m, \dots, p_m + l + 1. \end{cases}$$

It is evident that $(\mathcal{L}\psi_s)(z) = o(z^{l+1})$, $\det \|\psi_s^{(j)}(0)\|_{j,s=0}^{l} = 1$ and $\Psi = \{\psi_s\}_{s=0}^{l}$ is a linear independent system. Hence there exists a number $\epsilon(\Psi) > 0$ such that for every system $\Theta = \{\theta_s\}_{s=0}^{l}$ satisfying conditions $|\theta_s^{(j)}(0)| \le \epsilon(\Psi)$, $s, j = 0, 1, \dots, l$, the system $\Psi + \Theta = \{\psi_s + \theta_s\}_{s=0}^{l}$ is linear independent as well.

At the second step we fix an integer $l > 0$, a system $\Psi = \{\psi_s\}_{s=0}^{l}$ described in Lemma 3, the corresponding number $\epsilon(\Psi)$ and the set $\{\gamma_j\}_{j=0}^{l}$ of entire functions

$$\gamma_j = \mathcal{L}\psi_j, \quad j = 0, \dots, l. \tag{6.2}$$

Denote by φ_j the solution of equation $\mathcal{L}\varphi = \gamma_j$ constructed in part (i) with some $N > 0$. Our aim is to show that if $N > 0$ is sufficiently large then

$$|\varphi_j^{(s)}(0)| \le \epsilon(\Psi), \quad j, s = 0, \dots, l, \tag{6.3}$$

and hence $\{\psi_j - \varphi_j\}_{j=0}^{l}$ is a linear independent system of solutions to the homogeneous equation (1.1).

To this end let γ be one of the functions γ_j, and let $\gamma = \gamma_+ + \gamma_-$ with the summands γ_+ and γ_- being defined by (3.6). Then the solution φ of Eq. (1.1) has the form

$$\varphi(z) = \varphi_+(z) + \varphi_-(z)$$

where

$$\begin{aligned} \varphi_+(z) &= -\int_z^{\infty} K_0(z,t) a_{p_0 0}^{-1}(t)\gamma_+(t)dt - \int_z^{\infty} K_0(z,t)\mathcal{L}_+\psi_+(t)dt \\ \varphi_-(z) &= \tilde{\varphi}_+(-z + \alpha_m) \end{aligned} \tag{6.4}$$

and the estimates (5.6) and (5.8) are valid.

Since $\gamma_+(0) = \gamma_-(0) = \cdots = \gamma_+^{(l+1)}(0) = \gamma_-^{(l+1)}(0) = 0$, the jth derivative of the first term on the r.h.s. of (6.4) is

$$(-1)^{j+1}\int_0^{+\infty} \frac{\partial^j K_0(0,t)}{\partial z^j} a_{p_0 0}^{-1}(t)\gamma_+(t)dt, \quad j = 0, 1, \dots, l. \tag{6.5}$$

The representation (3.6) yields

$$|\gamma_+(t)| \le CN^{-\frac{l+1}{2}}|\Phi_+(t)|\exp(-w^*(t)), \quad t \ge 0,$$

with the function

$$\Phi_+(t) = \int_{\Re s=0} e^{Ns^2} e^{\Omega(s)}\gamma(s)s^{-(l+1)}\mu^{-1}(s)\frac{ds}{s-t}, \quad \Re t > 0.$$

Here and in the estimates which follow we denote by C (maybe different) numbers not depending on parameters and variables explicitly written in the relevant formulas.

Since $\gamma(z) = o(z^{l+1})$ as $z \to 0$, we have

$$|\Phi_+(t)| \le C\int_{-\infty}^{\infty} e^{-Nx^2}\Big|\frac{x}{ix+t}\Big|\,dx \le C, \quad \Re t \ge 0,\ \Im t = 0.$$

Therefore

$$|\gamma_+(t)| \le CN^{-(l+1)/2}\exp(-w^*(t)), \quad t \ge 0.$$

Furthermore, it follows from (5.4)

$$\frac{\partial^j K_0(0,t)}{\partial z^j} a_{p_0 0}^{-1}(t) \ \in \ \mathcal{E}_+(\exp(3w(t))$$

and

$$\Big|(-1)^j \int_0^\infty \frac{\partial^j K_0(0,t)}{\partial z^j} a_{p_0 0}^{-1}(t)\gamma_+(t)dt\Big| \ \le \ CN^{-(l+1)/2}. \tag{6.6}$$

To estimate derivatives at $z = 0$ of the second term in (6.4), let us apply Eq. (5.6) to functions $\psi_+(t+\alpha_k)$, $k = 1, \dots, m$ with $t \in S_+(-\delta/4, \delta/4)$. Since $(t + \alpha_k) \ge -\delta/4 + \alpha_k \ge 3\delta/4$, we obtain

$$|\psi_+(t+\alpha_k)| \ \le \ C\|a_{p_0 0}^{-1}\gamma_+\|_{-\exp w^*,\, S_+(3\delta/4,\, \delta/2)} \exp(-e^{w(\Re t)})$$

and hence

$$|\mathcal{L}_+\psi_+(t)| \ \le \ C\|a_{p_0 0}^{-1}\gamma_+\|_{-\exp w^*,\, S_+(3\delta/4,\, \delta/2)} \exp(-e^{w(\Re t - \delta/4)}),$$
$$t \ \in \ S_+(-\delta/4, \delta/4).$$

With account taken of Eq. (5.6) we find

$$\Big|\int_z^\infty K_0(z,t)\mathcal{L}_+\psi_+(t)dt\Big| \ \le \ C\|a_{p_0 0}^{-1}\gamma_+\|_{-\exp w^*,\, S_+(3\delta/4,\, \delta/2)}.$$

Now we have

$$|\Phi_+(z)| \ \le \ C, \quad z \ \in \ S_+(3\delta/4, \delta/2)$$

which together with (6.6) results in the estimates

$$|\varphi_+^{(j)}(0)| \ \le \ CN^{-(l+1)/2}, \quad j \ = \ 0, \dots, l.$$

To estimate $|\varphi_-^{(j)}(0)|$, we use relations $\varphi_-^{(j)}(0) \ = \ (-1)^j\tilde{\varphi}_+(\alpha_m)$ and estimates (5.6). If we accept $\tau \ = \ \alpha_m/4$, $a \ = \ -\alpha_m/2$, then we find

$$|\tilde{\varphi}_+(z)| \ \le \ C\|a_{p_m m}^{-1}\gamma_-\|_{-\exp w^*,\, S_-(-\alpha_m/2, \alpha_m/2)} \ \le \ CN^{-(l+1)/2}$$

and

$$|\varphi_-^{(j)}(0)| \ \le \ CN^{-(l+1)/2}, \quad j = 0, \dots, l.$$

With N sufficiently large it leads us to (6.3) completing the proof of part (ii).

References

[1] Nörlund, N.E. Differenzenrechnung, Springer Verlag, Berlin, 1924.

[2] Gel'fond, A.O. Calculus of Finite Differences, Third corrected edition, Nauka, Moscow, 1967; English translation in: International Monographs on Advanced Mathematics and Physics, Hindustan Publishing Corp., Delhi, 1971.

[3] Naftalevich, A. On a differential-difference equation, Michigan Math. J. **22**, no. 3, 205–223, 1975.

[4] Naftalevich, A. Application of an iteration method for the solution of a difference equation, Matematicheskii Sbornik, **57 (99)**, 151–178, 1962.

[5] Hurwitz, A. Mathematische Werke, Birkhäuser Verlag, Basel, Bd. II, S. 752, 1933.

[6] Belitskii G., and Tkachenko V. One-dimensional Functional Equations, Birkhäuser Verlag, Basel-Boston-Berlin, 2003.

[7] Levin, B.Ya. Lectures on Entire Functions, AMS, Translations of Mathematical Monographs, **150**, 1996.

[8] Markuschevich, A.I. Theory of Analytic Functions, v. 2, Nauka, Moscow, 1968.

[9] Naimark M.A. Linear Differential Operators, Ungar, New York, 1968.

[10] Hartman, F. Ordinary Differential Equations, John Wiley & Sons, New York-London-Sydney, 1964.

Genrich Belitskii and Vadim Tkachenko
Dept. of Mathematics
Ben-Gurion University
Beer–Sheva 84105, Israel
e-mail: genrich@math.bgu.ac.il
tkachenk@math.bgu.ac.il

Operator Theory:
Advances and Applications, Vol. 197, 21–49

Inverse Stieltjes-like Functions and Inverse Problems for Systems with Schrödinger Operator

Sergey V. Belyi and Eduard R. Tsekanovskii

To the memory of Moshe Livšic, a remarkable human being and a great mathematician

Abstract. A class of scalar inverse Stieltjes-like functions is realized as linear-fractional transformations of transfer functions of conservative systems based on a Schrödinger operator T_h in $L_2[a, +\infty)$ with a non-selfadjoint boundary condition. In particular it is shown that any inverse Stieltjes function of this class can be realized in the unique way so that the main operator $\mathbb{A}$ possesses a special semi-boundedness property. We derive formulas that restore the system uniquely and allow to find the exact value of a non-real boundary parameter h of the operator T_h as well as a real parameter μ that appears in the construction of the elements of the realizing system. An elaborate investigation of these formulas shows the dynamics of the restored parameters h and μ in terms of the changing free term α from the integral representation of the realizable function.

Mathematics Subject Classification (2000). Primary 47A10, 47B44; Secondary 46E20, 46F05.

Keywords. Operator colligation, conservative system, transfer (characteristic) function.

1. Introduction

The role of realizations of different classes of holomorphic operator-valued functions is universally recognized in the spectral analysis of non-self-adjoint operators, interpolation problems, and system theory, with the attention to them growing over the years. The literature on realization theory is too extensive to be discussed thoroughly in this paper. We refer a reader, however, to [2], [3], [7], [8], [9], [10], [11], [12], [20], [27], [26], and the literature therein. This paper is the second in a series

where we study realizations of a subclass of Herglotz-Nevanlinna functions with the systems based upon a Schrödinger operator. In [14] we have considered a class of scalar Stieltjes-like functions. Here we focus our attention on another important subclass of Herglotz-Nevanlinna functions, the so-called inverse Stieltjes-like functions.

We recall that an operator-valued function $V(z)$ acting on a finite-dimensional Hilbert space E belongs to the class of operator-valued Herglotz-Nevanlinna functions if it is holomorphic on $\mathbb{C} \setminus \mathbb{R}$, if it is symmetric with respect to the real axis, i.e., $V(z)^* = V(\bar{z})$, $z \in \mathbb{C} \setminus \mathbb{R}$, and if it satisfies the positivity condition

$$\operatorname{Im} V(z) \geq 0, \quad z \in \mathbb{C}_+.$$

It is well known (see, e.g., [18], [19]) that operator-valued Herglotz-Nevanlinna functions admit the following integral representation:

$$V(z) = Q + Lz + \int_{\mathbb{R}} \left(\frac{1}{t-z} - \frac{t}{1+t^2} \right) dG(t), \quad z \in \mathbb{C} \setminus \mathbb{R}, \tag{1.1}$$

where $Q = Q^*$, $L \geq 0$, and $G(t)$ is a nondecreasing operator-valued function on $\mathbb{R}$ with values in the class of nonnegative operators in E such that

$$\int_{\mathbb{R}} \frac{(dG(t)x, x)_E}{1+t^2} < \infty, \quad x \in E. \tag{1.2}$$

The realization of a selected class of Herglotz-Nevanlinna functions is provided by a linear conservative system Θ of the form

$$\begin{cases} (\mathbb{A} - zI)x = KJ\varphi_- \\ \varphi_+ = \varphi_- - 2iK^*x \end{cases} \tag{1.3}$$

or

$$\Theta = \begin{pmatrix} \mathbb{A} & K & J \\ \mathcal{H}_+ \subset \mathcal{H} \subset \mathcal{H}_- & & E \end{pmatrix}. \tag{1.4}$$

In this system $\mathbb{A}$, the *main operator* of the system, is a so-called $(*)$-extension, which is a bounded linear operator from $\mathcal{H}_+$ into $\mathcal{H}_-$ extending a symmetric operator A in $\mathcal{H}$, where $\mathcal{H}_+ \subset \mathcal{H} \subset \mathcal{H}_-$ is a rigged Hilbert space. Moreover, K is a bounded linear operator from the finite-dimensional Hilbert space E into $\mathcal{H}_-$, while $J = J^* = J^{-1}$ is acting on E, are such that $\operatorname{Im} \mathbb{A} = KJK^*$. Also, $\varphi_- \in E$ is an input vector, $\varphi_+ \in E$ is an output vector, and $x \in \mathcal{H}_+$ is a vector of the state space of the system Θ. The system described by (1.3)–(1.4) is called a rigged canonical system of the *Livšic type* [24] or (in operator theory) the *Brodskiĭ-Livšic rigged operator colligation*, cf., e.g., [11], [12], [15]. The operator-valued function

$$W_\Theta(z) = I - 2iK^*(\mathbb{A} - zI)^{-1}KJ \tag{1.5}$$

is a transfer function (or characteristic function) of the system Θ. It was shown in [11] that an operator-valued function $V(z)$ acting on a Hilbert space E of the form (1.1) can be represented and realized in the form

$$V(z) = i[W_\Theta(z) + I]^{-1}[W_\Theta(z) - I] = K^*(\mathbb{A}_R - zI)^{-1}K, \tag{1.6}$$

where $W_\Theta(z)$ is a transfer function of some canonical scattering ($J = I$) system Θ, and where the "*real part*" $\mathbb{A}_R = \frac{1}{2}(\mathbb{A} + \mathbb{A}^*)$ of $\mathbb{A}$ satisfies $\mathbb{A}_R \supset \hat{A} = \hat{A}^* \supset A$ if and only if the function $V(z)$ in (1.1) satisfies the following two conditions:

$$\begin{cases} L = 0, \\ Qx = \int_{\mathbb{R}} \frac{t}{1+t^2}\, dG(t)x \quad \text{when} \quad \int_{\mathbb{R}} (dG(t)x, x)_E < \infty. \end{cases} \tag{1.7}$$

In the current paper we specialize in an important subclass of Herglotz-Nevanlinna functions, the class of inverse Stieltjes-like functions that also includes inverse Stieltjes functions (see [13]). In Section 4 we specify a subclass of realizable inverse Stieltjes operator-functions and show that any member of this subclass can be realized by a system of the form (1.4) whose main operator $\mathbb{A}$ satisfies inequality

$$(\mathbb{A}_R f, f) \le (A^* f, f) + (f, A^* f), \quad f \in \mathcal{H}_+.$$

In Section 5 we introduce a class of scalar inverse Stieltjes-like functions. Then we rely on the general realization results developed in Section 4 (see also [13] and [14]) to restore a system Θ of the form (1.4) containing the Schrödinger operator in $L_2[a, +\infty)$ with non-self-adjoint boundary conditions

$$\begin{cases} T_h y = -y'' + q(x)y \\ y'(a) = hy(a) \end{cases}, \quad \Big(q(x) = \overline{q(x)},\, \operatorname{Im} h \ne 0\Big).$$

We show that if a non-decreasing function $\sigma(t)$ is the spectral distribution function of a positive self-adjoint boundary value problem

$$\begin{cases} A_\theta y = -y'' + q(x)y \\ y'(a) = \theta y(a) \end{cases}$$

and satisfies conditions

$$\int_0^\infty d\sigma(t) = \infty, \quad \int_0^\infty \frac{d\sigma(t)}{t + t^2} < \infty,$$

then for every real α an inverse Stieltjes-like function

$$V(z) = \alpha + \int_0^\infty \left(\frac{1}{t - z} - \frac{1}{t} \right) d\sigma(t)$$

can be realized in the unique way as $V(z) = V_\Theta(z) = i[W_\Theta(z) + I]^{-1}[W_\Theta(z) - I]$, where $W_\Theta(z)$ is the transfer function of a rigged canonical system Θ containing some Schrödinger operator T_h. In particular, it is shown that for every $\alpha \le 0$ an inverse Stieltjes function $V(z)$ with integral representation above can be realized by a system Θ whose main operator $\mathbb{A}$ is a $(*)$-extension of a Schrödinger operator T_h and satisfies (2.7).

In addition to the general realization results, Section 5 provides the reader with formulas that allow to find the exact value of a non-real parameter h in the definition of T_h of the realizing system Θ. A somewhat similar study is presented in Section 6 to describe the real parameter μ that appears in the construction of the elements of the realizing system. An elaborate investigation of these formulas

shows the dynamics of the restored parameters h and μ in terms of a changing free term α in the integral representation of $V(z)$ above. It will be shown and graphically presented that the parametric equations for the restored parameter h represent different circles whose centers and radii are completely determined by the function $V(z)$. Similarly, the behavior of the restored parameter μ are described by straight lines.

2. Some preliminaries

For a pair of Hilbert spaces $\mathcal{H}_1$, $\mathcal{H}_2$ we denote by $[\mathcal{H}_1, \mathcal{H}_2]$ the set of all bounded linear operators from $\mathcal{H}_1$ to $\mathcal{H}_2$. Let A be a closed, densely defined, symmetric operator in a Hilbert space $\mathcal{H}$ with inner product $(f, g), f, g \in \mathcal{H}$. Consider the rigged Hilbert space

$$\mathcal{H}_+ \subset \mathcal{H} \subset \mathcal{H}_-,$$

where $\mathcal{H}_+ = D(A^*)$ and

$$(f, g)_+ = (f, g) + (A^* f, A^* g), \quad f, g \in D(A^*).$$

Note that identifying the space conjugate to $\mathcal{H}_\pm$ with $\mathcal{H}_\mp$, we get that if $\mathbb{A} \in [\mathcal{H}_+, \mathcal{H}_-]$ then $\mathbb{A}^* \in [\mathcal{H}_+, \mathcal{H}_-]$.

Definition 2.1. An operator $\mathbb{A} \in [\mathcal{H}_+, \mathcal{H}_-]$ is called a self-adjoint bi-extension of a symmetric operator A if $\mathbb{A} = \mathbb{A}^*$, $\mathbb{A} \supset A$, and the operator

$$\widehat{A} f = \mathbb{A} f, \ f \in D(\widehat{A}) = \{f \in \mathcal{H}_+ : \mathbb{A} f \in \mathcal{H}\}$$

is self-adjoint in $\mathcal{H}$.

The operator $\widehat{A}$ in the above definition is called a *quasi-kernel* of a self-adjoint bi-extension $\mathbb{A}$ (see [30]) .

Definition 2.2. An operator $\mathbb{A} \in [\mathcal{H}_+, \mathcal{H}_-]$ is called a $(*)$-extension (or correct bi-extension) of an operator T (with non-empty set $\rho(T)$ of regular points) if

$$\mathbb{A} \supset T \supset A, \mathbb{A}^* \supset T^* \supset A$$

and the operator $\mathbb{A}_R = \frac{1}{2}(\mathbb{A} + \mathbb{A}^*)$ is a self-adjoint bi-extension of an operator A.

The existence, description, and analog of von Neumann's formulas for self-adjoint bi-extensions and $(*)$-extensions were discussed in [30] (see also [4], [5], [11]). For instance, if Φ is an isometric operator from the defect subspace $\mathfrak{N}_i$ of the symmetric operator A onto the defect subspace $\mathfrak{N}_{-i}$, then the formulas below establish a one-to one correspondence between $(*)$-extensions of an operator T and Φ

$$\mathbb{A} f = A^* f + iR(\Phi - I)x, \ \mathbb{A}^* f = A^* f + iR(\Phi - I)y, \tag{2.1}$$

where $x, y \in \mathfrak{N}_i$ are uniquely determined from the conditions

$$f - (\Phi + I)x \in D(T), \ f - (\Phi + I)y \in D(T^*)$$

and R is the Riesz-Berezanskii operator of the triplet $\mathcal{H}_+ \subset \mathcal{H} \subset \mathcal{H}_-$ that maps $\mathcal{H}_+$ isometrically onto $\mathcal{H}_-$ (see [30]). If the symmetric operator A has deficiency indices (n, n), then formulas (2.1) can be rewritten in the following form

$$\mathbb{A}f = A^*f + \sum_{k=1}^{n} \Delta_k(f)V_k, \quad \mathbb{A}^*f = A^*f + \sum_{k=1}^{n} \delta_k(f)V_k, \tag{2.2}$$

where $\{V_j\}_1^n \in \mathcal{H}_-$ is a basis in the subspace $R(\Phi - I)\mathfrak{N}_i$, and $\{\Delta_k\}_1^n$, $\{\delta_k\}_1^n$, are bounded linear functionals on $\mathcal{H}_+$ with the properties

$$\Delta_k(f) = 0, \ \ \forall f \in D(T), \quad \delta_k(f) = 0, \ \ \forall f \in D(T^*). \tag{2.3}$$

Let $\mathcal{H} = L_2[a, +\infty)$ and $l(y) = -y'' + q(x)y$ where q is a real locally summable function. Suppose that the symmetric operator

$$\begin{cases} Ay = -y'' + q(x)y \\ y(a) = y'(a) = 0 \end{cases} \tag{2.4}$$

has deficiency indices (1,1). Let D^* be the set of functions locally absolutely continuous together with their first derivatives such that $l(y) \in L_2[a, +\infty)$. Consider $\mathcal{H}_+ = D(A^*) = D^*$ with the scalar product

$$(y, z)_+ = \int_a^\infty \left(y(x)\overline{z(x)} + l(y)\overline{l(z)} \right) dx, \ \ y, \ z \in D^*.$$

Let

$$\mathcal{H}_+ \subset L_2[a, +\infty) \subset \mathcal{H}_-$$

be the corresponding triplet of Hilbert spaces. Consider operators

$$\begin{cases} T_h y = l(y) = -y'' + q(x)y \\ hy(a) - y'(a) = 0 \end{cases}, \quad \begin{cases} T_h^* y = l(y) = -y'' + q(x)y \\ \overline{h}y(a) - y'(a) = 0 \end{cases}, \tag{2.5}$$

$$\begin{cases} \widehat{A}y = l(y) = -y'' + q(x)y \\ \mu y(a) - y'(a) = 0 \end{cases}, \ \ \operatorname{Im}\mu = 0.$$

It is well known [1] that $\widehat{A} = \widehat{A^*}$. The following theorem was proved in [6].

Theorem 2.3. *The set of all $(*)$-extensions of a non-self-adjoint Schrödinger operator T_h of the form (2.5) in $L_2[a, +\infty)$ can be represented in the form*

$$\begin{aligned} \mathbb{A}y &= -y'' + q(x)y - \frac{1}{\mu - h}\left[y'(a) - hy(a)\right]\left[\mu\delta(x-a) + \delta'(x-a)\right], \\ \mathbb{A}^*y &= -y'' + q(x)y - \frac{1}{\mu - \overline{h}}\left[y'(a) - \overline{h}y(a)\right]\left[\mu\delta(x-a) + \delta'(x-a)\right]. \end{aligned} \tag{2.6}$$

In addition, the formulas (2.6) establish a one-to-one correspondence between the set of all $(*)$-extensions of a Schrödinger operator T_h of the form (2.5) and all real numbers $\mu \in [-\infty, +\infty]$.

Definition 2.4. An operator T with the domain $D(T)$ and $\rho(T) \neq \emptyset$ acting on a Hilbert space $\mathcal{H}$ is called *accretive* if

$$\operatorname{Re}(Tf, f) \geq 0, \ \ \forall f \in D(T).$$

Definition 2.5. An accretive operator T is called [22] *α-sectorial* if there exists a value of $\alpha \in (0, \pi/2)$ such that

$$\cot\alpha\,|\mathrm{Im}\,(Tf,f)| \le \mathrm{Re}\,(Tf,f), \qquad f \in \mathcal{D}(T).$$

An accretive operator is called *extremal accretive* if it is not α-sectorial for any $\alpha \in (0, \pi/2)$.

Definition 2.6. A $(*)$-extensions $\mathbb{A}$ in Definition 2.2 is called *accumulative* if

$$(\mathbb{A}_R f, f) \le (A^* f, f) + (f, A^* f), \quad f \in \mathcal{H}_+. \tag{2.7}$$

Consider the symmetric operator A of the form (2.4) with defect indices (1,1), generated by the differential operation $l(y) = -y'' + q(x)y$. Let $\varphi_k(x,\lambda)\,(k = 1,2)$ be the solutions of the following Cauchy problems:

$$\begin{cases} l(\varphi_1) = \lambda\varphi_1 \\ \varphi_1(a,\lambda) = 0 \\ \varphi_1'(a,\lambda) = 1 \end{cases}, \quad \begin{cases} l(\varphi_2) = \lambda\varphi_2 \\ \varphi_2(a,\lambda) = -1 \\ \varphi_2'(a,\lambda) = 0 \end{cases}.$$

It is well known [1] that there exists a function $m_\infty(\lambda)$ (called the Weyl-Titchmarsh function) for which

$$\varphi(x,\lambda) = \varphi_2(x,\lambda) + m_\infty(\lambda)\varphi_1(x,\lambda)$$

belongs to $L_2[a,+\infty)$.

Suppose that the symmetric operator A of the form (2.4) with deficiency indices (1,1) is nonnegative, i.e., $(Af, f) \ge 0$ for all $f \in D(A)$). It was shown in [28] that the Schrödinger operator T_h of the form (2.5) is accretive if and only if

$$\mathrm{Re}\,h \ge -m_\infty(-0). \tag{2.8}$$

For real h such that $h \ge -m_\infty(-0)$ we get a description of all nonnegative self-adjoint extensions of an operator A. For $h = -m_\infty(-0)$ the corresponding operator

$$\begin{cases} A_K\, y = -y'' + q(x)y \\ y'(a) + m_\infty(-0)y(a) = 0 \end{cases} \tag{2.9}$$

is the Kreĭn-von Neumann extension of A and for $h = +\infty$ the corresponding operator

$$\begin{cases} A_F\, y = -y'' + q(x)y \\ y(a) = 0 \end{cases} \tag{2.10}$$

is the Friedrichs extension of A (see [28], [6]).

3. Rigged canonical systems with Schrödinger operator

Let $\mathbb{A}$ be $(*)$-extension of an operator T, i.e.,

$$\mathbb{A} \supset T \supset A, \quad \mathbb{A}^* \supset T^* \supset A$$

where A is a symmetric operator with deficiency indices (n,n) and $D(A) = D(T) \cap D(T^*)$. In what follows we will only consider the case when the symmetric operator A has dense domain, i.e., $\overline{\mathcal{D}(A)} = \mathcal{H}$.

Definition 3.1. A system of equations

$$\begin{cases} (\mathbb{A} - zI)x = KJ\varphi_- \\ \varphi_+ = \varphi_- - 2iK^*x \end{cases},$$

or an array

$$\Theta = \begin{pmatrix} \mathbb{A} & K & J \\ \mathcal{H}_+ \subset \mathcal{H} \subset \mathcal{H}_- & & E \end{pmatrix} \tag{3.1}$$

is called a *rigged canonical system of the Livsic type* if:

1) E is a finite-dimensional Hilbert space with scalar product $(\cdot,\cdot)_E$ and the operator J in this space satisfies the conditions $J = J^* = J^{-1}$,
2) $K \in [E, \mathcal{H}_-]$, $\ker K = \{0\}$,
3) $\operatorname{Im}\mathbb{A} = KJK^*$, where $K^* \in [\mathcal{H}_+, E]$ is the adjoint of K.

In the definition above $\varphi_- \in E$ stands for an input vector, $\varphi_+ \in E$ is an output vector, and x is a state space vector in $\mathcal{H}$. An operator $\mathbb{A}$ is called a *main operator* of the system Θ, J is a *direction operator*, and K is a *channel operator*. A system Θ of the form (3.1) is called an *accretive system* [14] if its main operator $\mathbb{A}$ is accretive and *accumulative* if its main operator $\mathbb{A}$ is accumulative, i.e., satisfies (2.7).

An operator-valued function

$$W_\Theta(\lambda) = I - 2iK^*(\mathbb{A} - \lambda I)^{-1}KJ \tag{3.2}$$

defined on the set $\rho(T)$ of regular points of an operator T is called the *transfer function* (*characteristic function*) of the system Θ, i.e., $\varphi_+ = W_\Theta(\lambda)\varphi_-$. It is known [28], [30] that any $(*)$-extension $\mathbb{A}$ of an operator T $(A^* \supset T \supset A)$, where A is a symmetric operator with deficiency indices (n,n) $(n < \infty)$, $D(A) = D(T) \cap D(T^*)$, can be included as a main operator of some rigged canonical system with $\dim E < \infty$ and invertible channel operator K.

It was also established [28], [30] that

$$V_\Theta(\lambda) = K^*(\operatorname{Re}\mathbb{A} - \lambda I)^{-1}K \tag{3.3}$$

is a Herglotz-Nevanlinna operator-valued function acting on a Hilbert space E, satisfying the following relation for $\lambda \in \rho(T)$, $\operatorname{Im}\lambda \neq 0$

$$V_\Theta(\lambda) = i[W_\Theta(\lambda) - I][W_\Theta(\lambda) + I]^{-1}J. \tag{3.4}$$

Alternatively,

$$\begin{aligned} W_\Theta(\lambda) &= (I + iV_\Theta(\lambda)J)^{-1}(I - iV_\Theta(\lambda)J) \\ &= (I - iV_\Theta(\lambda)J)(I + iV_\Theta(\lambda)J)^{-1}. \end{aligned} \tag{3.5}$$

Let us recall (see [30], [6]) that a symmetric operator with dense domain $\mathcal{D}(A)$ is called *prime* if there is no reducing, nontrivial invariant subspace on which A induces a self-adjoint operator. It was established in [29] that a symmetric operator A is prime if and only if

$$\operatorname*{c.l.s.}_{\lambda \neq \overline{\lambda}} \mathfrak{N}_\lambda = \mathcal{H}. \tag{3.6}$$

We call a rigged canonical system of the form (3.1) *prime* if

$$\underset{\lambda\neq\overline{\lambda},\,\lambda\in\rho(T)}{c.l.s.}\ \mathfrak{N}_\lambda=\mathcal{H}.$$

One easily verifies that if system Θ is prime, then a symmetric operator A of the system is prime as well.

The following theorem [6], [14] and corollary [14] establish the connection between two rigged canonical systems with equal transfer functions.

Theorem 3.2. *Let* $\Theta_1=\begin{pmatrix}\mathbb{A}_1 & K_1 & J\\ \mathcal{H}_{+1}\subset\mathcal{H}_1\subset\mathcal{H}_{-1} & & E\end{pmatrix}$ *and* $\Theta_2=\begin{pmatrix}\mathbb{A}_2 & K_2 & J\\ \mathcal{H}_{+2}\subset\mathcal{H}_2\subset\mathcal{H}_{-2} & & E\end{pmatrix}$ *be two prime rigged canonical systems of the Livsic type with*

$$\begin{aligned}&\mathbb{A}_1\supset T_1\supset A_1,\quad \mathbb{A}_1^*\supset T_1^*\supset A_1,\\ &\mathbb{A}_2\supset T_2\supset A_2,\quad \mathbb{A}_2^*\supset T_2^*\supset A_2,\end{aligned}\tag{3.7}$$

and such that A_1 *and* A_2 *have finite and equal defect indices.*

If

$$W_{\Theta_1}(\lambda)=W_{\Theta_2}(\lambda),\quad \lambda\in\rho(T_1)\cap\rho(T_2),\quad (\rho(T_1)\cap\rho(T_2))\cap\mathbb{C}_\pm\neq\emptyset,\tag{3.8}$$

then there exists an isometric operator U *from* $\mathcal{H}_1$ *onto* $\mathcal{H}_2$ *such that* $U_+=U|_{\mathcal{H}_{+1}}$ *is an isometry*[1] *from* $\mathcal{H}_{+1}$ *onto* $\mathcal{H}_{+2}$, $U_-^*=U_+^*$ *is an isometry from* $\mathcal{H}_{-1}$ *onto* $\mathcal{H}_{-2}$, *and*

$$UT_1=T_2U,\quad \mathbb{A}_2=U_-\mathbb{A}_1U_+^{-1},\quad U_-K_1=K_2.\tag{3.9}$$

Corollary 3.3. *Let* Θ_1 *and* Θ_2 *be the two prime systems from the statement of Theorem* 3.2. *Then the mapping* U *described in the conclusion of the theorem is unique.*

Now we shall construct a rigged canonical system based on a non-self-adjoint Schrödinger operator. One can easily check that the $(*)$-extension

$$\mathbb{A}y=-y''+q(x)y-\frac{1}{\mu-h}\,[y'(a)-hy(a)]\,[\mu\delta(x-a)+\delta'(x-a)],\ \ \operatorname{Im}h>0$$

of the non-self-adjoint Schrödinger operator T_h of the form (2.5) satisfies the condition

$$\operatorname{Im}\mathbb{A}=\frac{\mathbb{A}-\mathbb{A}^*}{2i}=(.,g)g,\tag{3.10}$$

where

$$g=\frac{(\operatorname{Im}h)^{\frac{1}{2}}}{|\mu-h|}\,[\mu\delta(x-a)+\delta'(x-a)]\tag{3.11}$$

[1] It was shown in [6] that the operator U_+ defined this way is an isometry from $\mathcal{H}_{+1}$ onto $\mathcal{H}_{+2}$. It is also shown there that the isometric operator $U^*:\mathcal{H}_{+2}\to\mathcal{H}_{+1}$ uniquely defines operator $U_-=(U^*)^*:\mathcal{H}_{-1}\to\mathcal{H}_{-2}$.

and $\delta(x-a), \delta'(x-a)$ are the delta-function and its derivative at the point a. Moreover,

$$(y,g) = \frac{(\operatorname{Im} h)^{\frac{1}{2}}}{|\mu - h|}\ [\mu y(a) - y'(a)], \tag{3.12}$$

where

$$y \in \mathcal{H}_+, g \in \mathcal{H}_-, \mathcal{H}_+ \subset L_2(a,+\infty) \subset \mathcal{H}_-$$

and the triplet of Hilbert spaces is as discussed in Theorem 2.3. Let $E = \mathbb{C}$, $Kc = cg$ $(c \in \mathbb{C})$. It is clear that

$$K^* y = (y,g), \quad y \in \mathcal{H}_+ \tag{3.13}$$

and $\operatorname{Im} \mathbb{A} = KK^*$. Therefore, the array

$$\Theta = \begin{pmatrix} \mathbb{A} & K & 1 \\ \mathcal{H}_+ \subset L_2[a,+\infty) \subset \mathcal{H}_- & & \mathbb{C} \end{pmatrix} \tag{3.14}$$

is a rigged canonical system with the main operator $\mathbb{A}$ of the form (2.6), the direction operator $J = 1$ and the channel operator K of the form (3.13). Our next logical step is finding the transfer function of (3.14). It was shown in [6] that

$$W_\Theta(\lambda) = \frac{\mu - h}{\mu - \overline{h}}\ \frac{m_\infty(\lambda) + \overline{h}}{m_\infty(\lambda) + h}, \tag{3.15}$$

and

$$V_\Theta(\lambda) = \frac{(m_\infty(\lambda) + \mu) \operatorname{Im} h}{(\mu - \operatorname{Re} h)\, m_\infty(\lambda) + \mu \operatorname{Re} h - |h|^2}. \tag{3.16}$$

4. Realization of inverse Stieltjes functions

Let E be a finite-dimensional Hilbert space. The scalar versions of the definitions below can be found in [21]. We recall (see [14], [21]) that an operator-valued Herglotz-Nevanlinna function $V(z)$ is Stieltjes if it is holomorphic in $\operatorname{Ext}[0,+\infty)$ and

$$\frac{\operatorname{Im}[zV(z)]}{\operatorname{Im} z} \geq 0.$$

Definition 4.1. We will call an operator-valued Herglotz-Nevanlinna function $V(z) \in [E,E]$ by an *inverse Stieltjes* if $V(z)$ admits the following integral representation

$$V(z) = \alpha + \beta \cdot z + \int_0^\infty \left(\frac{1}{t-z} - \frac{1}{t} \right) dG(t), \tag{4.1}$$

where $\alpha \leq 0$, $\beta \geq 0$, and $G(t)$ is a non-decreasing on $[0,+\infty)$ operator-valued function such that

$$\int_0^\infty \frac{(dG(t)e,e)}{t+t^2} < \infty, \quad \forall e \in E.$$

Alternatively (see [21]) an operator-valued function $V(z)$ is inverse Stieltjes if it is holomorphic in $\mathrm{Ext}[0,+\infty)$ and $V(z)\le 0$ in $(-\infty,0)$. It is known [21] that a function $V(z)\not\equiv 0$ is an inverse Stieltjes function iff the function $-(V(z))^{-1}$ is Stieltjes.

The following definition was given in [13] and provides the description of all realizable inverse Stieltjes operator-valued functions.

Definition 4.2. An operator-valued inverse Stieltjes function $V(z)\in[E,E]$ is said to be a member of the class $S^{-1}(R)$ if in the representation (4.1) we have

$$\text{i)}\qquad \beta=0,$$

$$\text{ii)}\qquad \alpha e=\int_0^\infty \frac{1}{t}dG(t)e=0,$$

for all $e\in E$ with

$$\int_0^\infty (dG(t)e,e)_E<\infty. \tag{4.2}$$

In what follows we will, however, be mostly interested in the following subclass of $S^{-1}(R)$ that was also introduced in [13].

Definition 4.3. An operator-valued inverse Stieltjes function $V(z)\in S^{-1}(R)$ is a member of the class $S_0^{-1}(R)$ if

$$\int_0^\infty (dG(t)e,e)_E=\infty, \tag{4.3}$$

for all $e\in E$, $e\neq 0$.

It is not hard to see that $S_0^{-1}(R)$ is the analogue of the class $N_0(R)$ introduced in [12] and of the class $S_0(R)$ discussed in [14].

The following statement [13] is the direct realization theorem for the functions of the class $S_0^{-1}(R)$.

Theorem 4.4. *Let Θ be an accumulative system of the form* (3.1)*. Then the operator-function $V_\Theta(z)$ of the form* (3.3), (3.4) *belongs to the class $S_0^{-1}(R)$.*

The inverse realization theorem can be stated and proved (see [13]) for the class $S_0^{-1}(R)$ as follows.

Theorem 4.5. *Let a operator-valued function $V(z)$ belong to the class $S_0^{-1}(R)$. Then $V(z)$ admits a realization by an accumulative prime system Θ of the form* (3.1) *with $J=I$.*

Proof. It was shown in [13] that any member of the class $S_0^{-1}(R)$ is realizable by an accumulative system Θ of the form (3.1) with $J=I$. Thus all we actually have to show is that the model system Θ that was constructed in [13] is prime.

As it was also shown in [11], [12], and [13], the symmetric operator A of the model system Θ is prime and positive, and hence (3.6) takes place. We are going to show that in this case the system Θ is also prime, i.e.,

$$\underset{\lambda\neq\bar\lambda,\ \lambda\in\rho(T)}{c.l.s.}\ \mathfrak{N}_\lambda=\mathcal{H}. \tag{4.4}$$

Consider the operator $U_{\lambda_0\lambda} = (\tilde{A} - \lambda_0 I)(\tilde{A} - \lambda I)^{-1}$, where $\tilde{A}$ is an arbitrary self-adjoint extension of A. By a simple check one confirms that $U_{\lambda_0\lambda}\mathfrak{N}_{\lambda_0} = \mathfrak{N}_\lambda$. To prove (4.4) we assume that there is a function $f \in \mathcal{H}$ such that

$$f \perp \underset{\lambda \neq \bar{\lambda},\, \lambda \in \rho(T)}{c.l.s.} \mathfrak{N}_\lambda.$$

Then $(f, U_{\lambda_0\lambda} g) = 0$ for all $g \in \mathfrak{N}_{\lambda_0}$ and all $\lambda \in \rho(T)$. But since the system Θ is accumulative, it follows that there are regular points of T in the upper and lower half-planes. This leads to a conclusion that the function $\phi(\lambda) = (f, U_{\lambda_0\lambda} g) \equiv 0$ for all $\lambda \neq \bar{\lambda}$. Combining this with (3.6) we conclude that $f = 0$ and thus (4.4) holds. □

5. Restoring a non-self-adjoint Schrödinger operator T_h

In this section we are going to use the realization technique and results developed for inverse Stieltjes functions in section 4 to obtain the solution of inverse spectral problem for Schrödinger operator of the form (2.5) in $L_2[a, +\infty)$ with non-self-adjoint boundary conditions

$$\begin{cases} T_h y = -y'' + q(x)y \\ y'(a) = hy(a) \end{cases}, \quad \left(q(x) = \overline{q(x)},\, \operatorname{Im} h \neq 0\right). \tag{5.1}$$

Following the framework of [14] we let $\mathcal{H} = L_2[a, +\infty)$ and $l(y) = -y'' + q(x)y$ where q is a real locally summable function. We consider a symmetric operator with defect indices $(1, 1)$

$$\begin{cases} \tilde{B} y = -y'' + q(x)y \\ y'(a) = y(a) = 0 \end{cases} \tag{5.2}$$

together with its positive self-adjoint extension of the form

$$\begin{cases} \tilde{B}_\theta y = -y'' + q(x)y \\ y'(a) = \theta y(a) \end{cases} \tag{5.3}$$

defined in $\mathcal{H} = L_2[a, +\infty)$. A non-decreasing function $\sigma(\lambda)$ defined on $[0, +\infty)$ is called the *distribution function* (see [25]) *of an operator pair* $\tilde{B}_\theta$, $\tilde{B}$, where $\tilde{B}_\theta$ of the form (5.3) is a self-adjoint extension of symmetric operator $\tilde{B}$ of the form (5.2), and if the formulas

$$\begin{aligned} \varphi(\lambda) &= U f(x), \\ f(x) &= U^{-1} \varphi(\lambda), \end{aligned} \tag{5.4}$$

establish one-to-one isometric correspondence U between

$$L_2^\sigma[0, +\infty) \quad \text{and} \quad L_2[a, +\infty).$$

Moreover, this correspondence is such that the operator $\tilde{B}_\theta$ is unitarily equivalent to the operator

$$\Lambda_\sigma \varphi(\lambda) = \lambda \varphi(\lambda), \quad (\varphi(\lambda) \in L_2^\sigma[0, +\infty)) \tag{5.5}$$

in $L_2^\sigma[0,+\infty)$ while symmetric operator $\tilde{B}$ in (5.2) is unitarily equivalent to the symmetric operator

$$\Lambda_\sigma^0\varphi(\lambda)=\lambda\varphi(\lambda),\quad D(\Lambda_\sigma^0)=\left\{\varphi(\lambda)\in L_2^\sigma[0,+\infty):\ \int_0^{+\infty}\varphi(\lambda)d\sigma(\lambda)=0\right\}. \tag{5.6}$$

Definition 5.1. A scalar Herglotz-Nevanlinna function $V(z)$ is called an *inverse Stieltjes-like function* if it has an integral representation

$$V(z)=\alpha+\int_0^\infty\left(\frac{1}{t-z}-\frac{1}{t}\right)d\tau(t),\quad \int_0^\infty\frac{d\tau(t)}{t+t^2}<\infty \tag{5.7}$$

similar to (4.1) but with an arbitrary (not necessarily non-positive) constant α and $\beta=0$.

We are going to introduce a new class of realizable scalar inverse Stieltjes-like functions whose structure is similar to that of $S_0^{-1}(R)$ of Section 4.

Definition 5.2. An inverse Stieltjes-like function $V(z)$ is said to be a member of the *class* $SL_0^{-1}(R)$ if it admits an integral representation

$$V(z)=\alpha+\int_0^\infty\left(\frac{1}{t-z}-\frac{1}{t}\right)d\tau(t), \tag{5.8}$$

where non-decreasing function $\tau(t)$ satisfies the following conditions

$$\int_0^\infty d\tau(t)=\infty,\quad \int_0^\infty\frac{d\tau(t)}{t+t^2}<\infty. \tag{5.9}$$

Consider the following subclasses of $SL_0^{-1}(R)$.

Definition 5.3. A function $V(z)\in SL_0^{-1}(R)$ belongs to the *class* $SL_0^{-1}(R,K)$ if

$$\int_0^\infty\frac{d\tau(t)}{t}=\infty. \tag{5.10}$$

Definition 5.4. A function $V(z)\in SL_0^{-1}(R)$ belongs to the *class* $SL_{01}^{-1}(R,K)$ if

$$\int_0^\infty\frac{d\tau(t)}{t}<\infty. \tag{5.11}$$

The following theorem describes the realization of the class $SL_0^{-1}(R)$.

Theorem 5.5. *Let $V(z)\in SL_0^{-1}(R)$. Then it can be realized by a prime system Θ of the form* (3.1).

Proof. We start by applying the general realization theorems from [11] and [13] to a Herglotz-Nevanlinna function $V(z)$ and obtain a rigged canonical system of the Livsic type

$$\Theta_\Lambda=\begin{pmatrix}\mathbf{\Lambda} & K^\tau & 1\\ \mathcal{H}_+^\tau\subset L_2^\tau[0,+\infty)\subset\mathcal{H}_-^\tau & & \mathbb{C}\end{pmatrix}, \tag{5.12}$$

such that $V(z) = V_{\Theta_\Lambda}(z)$. Following the steps for construction of the model system described in [11] and [13], we note that

$$\mathbf{\Lambda} = \operatorname{Re}\mathbf{\Lambda} + iK^\tau(K^\tau)^*$$

is a correct $(*)$-extension of an operator T^τ such that $\mathbf{\Lambda} \supset T^\tau \supset \Lambda_\tau^0$ where Λ_τ^0 is defined in (5.6). The real part $\operatorname{Re}\mathbf{\Lambda}$ is a self-adjoint bi-extension of Λ_τ^0 that has a quasi-kernel Λ_τ of the form (5.5). It was also shown in [13] that the operator $\mathbf{\Lambda}$ possess the accumulative property (2.7). The operator K^τ in the above system (see [11], [13]) is defined by

$$K^\tau c = c \cdot \alpha, \quad (K^\tau)^* x = (x, \alpha) \quad c \in \mathbb{C},\ \alpha \in \mathcal{H}_-^\tau,\ x(t) \in \mathcal{H}_+^\tau.$$

In addition we can observe that the function $\eta(\lambda) \equiv 1$ belongs to $\mathcal{H}_-^\tau$. To confirm this we need to show that $(x, 1)$ defines a continuous linear functional for every $x \in \mathcal{H}_+^\tau$. It was shown in [11], [12] that

$$\mathcal{H}_+^\tau = \mathcal{D}(\Lambda_\tau^0) \dot{+} \left\{\frac{c_1}{1+t^2}\right\} \dot{+} \left\{\frac{c_2 t}{1+t^2}\right\}, \quad c_1, c_2 \in \mathbb{C}. \tag{5.13}$$

Consequently, every vector $x \in \mathcal{H}_+^\tau$ has three components $x = x_1 + x_2 + x_3$ according to the decomposition (5.13) above. Obviously, $(x_1, 1)$ and $(x_2, 1)$ yield convergent integrals while $(x_3, 1)$ boils down to

$$\int_0^\infty \frac{t}{1+t^2}\, d\tau(t).$$

The convergence of the latter is guaranteed by the definition of inverse Stieltjes-like function. The state space of the system Θ_Λ is $\mathcal{H}_+^\tau \subset L_2^\tau[0, +\infty) \subset \mathcal{H}_-^\tau$, where $\mathcal{H}_+^\tau = \mathcal{D}\big((\Lambda_\tau^0)^*\big)$.

We can also show that the system Θ_Λ is a prime system. In order to do so we need to show that

$$\underset{\lambda \neq \bar{\lambda},\ \lambda \in \rho(T^\tau)}{c.l.s.} \mathfrak{N}_\lambda = L_2^\tau[0, +\infty), \tag{5.14}$$

where $\mathfrak{N}_\lambda$ are defect subspaces of the symmetric operator Λ_τ^0. It is known (see [11], [13]) that Λ_τ^0 is a non-negative prime operator. Hence we can follow the reasoning of the proof of theorem 4.5 and only confirm that operator T^τ has regular points in the upper and lower half-planes. To see this we first note that non-negative operator Λ_τ^0 has no kernel spectrum [1] on the left real half-axis. Then we apply Theorem 1 of [1] (see page 149 of vol. 2 of [1]) that gives the complete description of the spectrum of T^τ. This theorem implies that there are regular points of T^τ on the left real half-axis. Since $\rho(T^\tau)$ is an open set we confirm the presence of non-real regular points of T^τ in both half-planes. Thus (5.14) holds and Θ_Λ is a prime system.

In order to complete the proof of the theorem we merely set

$$\mathbb{A} = \mathbf{\Lambda} = \operatorname{Re}\mathbf{\Lambda} + iK^\tau(K^\tau)^* \quad \text{and} \quad K = K^\tau. \qquad \square$$

At this point we are ready to state and prove the main realization result of this paper.

Theorem 5.6. *Let $V(z) \in SL_0^{-1}(R)$ and the function $\tau(t)$ be the distribution function of an operator pair $\tilde{B}_\theta$ of the form* (5.2) *and $\tilde{B}$ of the form* (5.3)*. Then there exist unique Schrödinger operator T_h ($\operatorname{Im} h > 0$) of the form* (5.1)*, operator $\mathbb{A}$ given by* (2.6)*, operator K as in* (3.13)*, and the rigged canonical system of the Livsic type*

$$\Theta = \begin{pmatrix} \mathbb{A} & K & 1 \\ \mathcal{H}_+ \subset L_2[a,+\infty) \subset \mathcal{H}_- & & \mathbb{C} \end{pmatrix}, \tag{5.15}$$

of the form (3.14) *so that $V(z)$ is realized by Θ, i.e., $V(z) = V_\Theta(z)$.*

Proof. Since $\tau(t)$ is the distribution function of the positive self-adjoint operator, then (see [25]) we can completely restore the operator $\tilde{B}_\theta$ of the form (5.3) as well as a symmetric operator $\tilde{B}$ of the form (5.2). It follows from the definition of the distribution function above that there is operator U defined in (5.4) establishing one-to-one isometric correspondence between $L_2^\tau[0,+\infty)$ and $L_2[a,+\infty)$ while providing for the unitary equivalence between the operator $\tilde{B}_\theta$ and operator of multiplication by independent variable Λ_τ of the form (5.5).

Let us consider the system Θ_Λ of the form (5.12) constructed in the proof of Theorem 5.5. Applying Theorem 3.2 on unitary equivalence to the isometry U defined in (5.4) we obtain a triplet of isometric operators U_+, U, and U_-, where

$$U_+ = U\big|_{\mathcal{H}_+^\tau}, \quad U_-^* = U_+^*.$$

This triplet of isometric operators will map the rigged Hilbert space of Θ_Λ, that is $\mathcal{H}_+^\tau \subset L_2^\tau[0,+\infty) \subset \mathcal{H}_-^\tau$, into another rigged Hilbert space $\mathcal{H}_+ \subset L_2^\tau[a,+\infty) \subset \mathcal{H}_-$. Moreover, U_+ is an isometry from $\mathcal{H}_+^\tau = \mathcal{D}(\Lambda_\tau^{0*})$ onto $\mathcal{H}_+ = \mathcal{D}(\tilde{B}^*)$, and $U_-^* = U_+^*$ is an isometry from $\mathcal{H}_+^\tau$ onto $\mathcal{H}_-$. This is true since the operator U provides the unitary equivalence between the symmetric operators $\tilde{B}$ and Λ_τ^0.

Now we construct a system

$$\Theta = \begin{pmatrix} \mathbb{A} & K & 1 \\ \mathcal{H}_+ \subset L_2[a,+\infty) \subset \mathcal{H}_- & & \mathbb{C} \end{pmatrix}$$

where $K = U_- K^\tau$ and $\mathbb{A} = U_- \mathbf{\Lambda} U_+^{-1}$ is a correct $(*)$-extension of operator $T = UT^\tau U^{-1}$ such that $\mathbb{A} \supset T \supset \tilde{B}$. The real part $\operatorname{Re}\mathbb{A}$ contains the quasi-kernel $\tilde{B}_\theta$. This construction of $\mathbb{A}$ is unique due to the theorem on the uniqueness of a $(*)$-extension for a given quasi-kernel (see [30]). On the other hand, all $(*)$-extensions based on a pair $\tilde{B}$, $\tilde{B}_\theta$ must take form (2.6) for some values of parameters h and μ. Consequently, our function $V(z)$ is realized by the system Θ of the form (5.15) and

$$V(z) = V_{\Theta_\Lambda}(z) = V_\Theta(z). \qquad \square$$

The theorem below gives the criteria for the operator T_h of the realizing system to be accretive.

Theorem 5.7. *Let $V(z) \in SL_0^{-1}(R)$ satisfy the conditions of Theorem* 5.5. *Then the operator T_h in the conclusion of the Theorem* 5.5 *is accretive if and only if*

$$\alpha^2 - \alpha \int_0^\infty \frac{d\tau(t)}{t} + 1 \geq 0. \tag{5.16}$$

The operator T_h is ϕ-sectorial for some $\phi \in (0, \pi/2)$ if and only if the inequality (5.16) *is strict. In this case the exact value of angle ϕ can be calculated by the formula*

$$\tan\phi = \frac{\int_0^\infty \frac{d\tau(t)}{t}}{\alpha^2 - \alpha \int_0^\infty \frac{d\tau(t)}{t} + 1}. \tag{5.17}$$

Proof. It was shown in [29] that for the system Θ in (5.15) described in the previous theorem the operator T_h is accretive if and only if the function

$$\begin{aligned} V_h(z) &= -i[W_\Theta^{-1}(-1)W_\Theta(z) + I]^{-1}[W_\Theta^{-1}(-1)W_\Theta(z) - I] \\ &= -i\frac{1 - [(m_\infty(z) + \bar{h})/(m_\infty(z) + h)][(m_\infty(-1) + h)/(m_\infty(-1) + \bar{h})]}{1 + [(m_\infty(z) + \bar{h})/(m_\infty(z) + h)][(m_\infty(-1) + h)/(m_\infty(-1) + \bar{h})]}, \end{aligned} \tag{5.18}$$

is holomorphic in $\mathrm{Ext}[0, +\infty)$ and satisfies the following inequality

$$1 + V_h(0)\, V_h(-\infty) \geq 0. \tag{5.19}$$

Here $W_\Theta(z)$ is the transfer function of (5.15). It is also shown in [29] that the operator T_h is α-sectorial for some $\alpha \in (0, \pi/2)$ if and only if the inequality (5.19) is strict while the exact value of angle α can be calculated by the formula

$$\cot\alpha = \frac{1 + V_h(0)\, V_h(-\infty)}{|V_h(-\infty) - V_h(0)|}. \tag{5.20}$$

According to Theorem 5.5 and equation (3.5)

$$W_\Theta(z) = (I - iV(z)J)(I + iV(z)J)^{-1}.$$

By direct calculations one obtains

$$W_\Theta(-1) = \frac{1 - i\left[\alpha - \int_0^\infty \frac{d\tau(t)}{t+t^2}\right]}{1 + i\left[\alpha - \int_0^\infty \frac{d\tau(t)}{t+t^2}\right]},\ W_\Theta^{-1}(-1) = \frac{1 + i\left[\alpha - \int_0^\infty \frac{d\tau(t)}{t+t^2}\right]}{1 - i\left[\alpha - \int_0^\infty \frac{d\tau(t)}{t+t^2}\right]}. \tag{5.21}$$

Using the following notations

$$c = \alpha - \int_0^\infty \frac{d\tau(t)}{t+t^2} \quad \text{and} \quad d = \alpha - \int_0^\infty \frac{d\tau(t)}{t},$$

and performing straightforward calculations we obtain

$$W_\Theta(-1) = \frac{1 - i\,c}{1 + i\,c}, \quad W_\Theta(-\infty) = \frac{1 - i\,d}{1 + i\,d},$$

and

$$V_h(0) = \frac{c - \alpha}{1 + c\,\alpha} \quad \text{and} \quad V_h(-\infty) = \frac{c - d}{1 + c\,d}. \tag{5.22}$$

Substituting (5.22) into (5.20) and performing the necessary steps we get

$$\cot\phi = \frac{1+\alpha\, d}{\alpha - d} = \frac{\alpha^2 - \alpha \int_0^\infty \frac{d\tau(t)}{t} + 1}{\int_0^\infty \frac{d\tau(t)}{t}}. \tag{5.23}$$

Taking into account that $\alpha - d > 0$ we combine (5.19), (5.20) with (5.23) and this completes the proof of the theorem. □

Below we will derive the formulas for calculation of the boundary parameter h in the restored Schrödinger operator T_h of the form (5.1). We consider two major cases.

Case 1. In the first case we assume that $\int_0^\infty \frac{d\tau(t)}{t} < \infty$. This means that our function $V(z)$ belongs to the class $SL_{01}^{-1}(R, K)$. In what follows we denote

$$b = \int_0^\infty \frac{d\tau(t)}{t} \quad \text{and} \quad m = m_\infty(-0).$$

Suppose that $b \geq 2$. Then the quadratic inequality (5.16) implies that for all α such that

$$\alpha \in \left(-\infty, \frac{b - \sqrt{b^2-4}}{2}\right] \cup \left[\frac{b + \sqrt{b^2-4}}{2}, +\infty\right) \tag{5.24}$$

the restored operator T_h is accretive. Clearly, this operator is extremal accretive if

$$\alpha = \frac{b \pm \sqrt{b^2-4}}{2}.$$

In particular if $b = 2$ then $\alpha = 1$ and the function

$$V(z) = 1 + \int_0^\infty \left(\frac{1}{t-z} - \frac{1}{t}\right) d\tau(t)$$

is realized using an extremal accretive T_h.

Now suppose that $0 < b < 2$. Then for every $\alpha \in (-\infty, +\infty)$ the restored operator T_h will be accretive and ϕ-sectorial for some $\phi \in (0, \pi/2)$. Consider a function $V(z)$ defined by (5.8). Conducting realizations of $V(z)$ by operators T_h for different values of $\alpha \in (-\infty, +\infty)$ we notice that the operator T_h with the largest angle of sectoriality occurs when

$$\alpha = \frac{b}{2}, \tag{5.25}$$

and is found according to the formula

$$\phi = \arctan \frac{b}{1 - b^2/4}. \tag{5.26}$$

This follows from the formula (5.17), the fact that $\alpha^2 - \alpha\, b + 1 > 0$ for all α, and the formula

$$\alpha^2 - \alpha\, b + 1 = \left(\alpha - \frac{b}{2}\right)^2 + \left(1 - \frac{b^2}{4}\right).$$

Now we will focus on the description of the parameter h in the restored operator T_h. It was shown in [6] that the quasi-kernel $\hat{A}$ of the realizing system Θ from theorem 5.5 takes a form

$$\begin{cases} \widehat{A}y = -y'' + qy \\ y'(a) = \eta y(a) \end{cases}, \ \eta = \frac{\mu \mathrm{Re}\, h - |h|^2}{\mu - \mathrm{Re}\, h} \tag{5.27}$$

On the other hand, since $\sigma(t)$ is also the distribution function of the positive self-adjoint operator, we can conclude that $\hat{A}$ equals to the operator $\tilde{B}_\theta$ of the form (5.3). This connection allows us to obtain

$$\theta = \eta = \frac{\mu \mathrm{Re}\, h - |h|^2}{\mu - \mathrm{Re}\, h}. \tag{5.28}$$

Assuming that

$$h = x + iy$$

we will use (5.28) to derive the formulas for x and y in terms of γ. First, to eliminate parameter μ, we notice that (3.15) and (3.5) imply

$$W_\Theta(\lambda) = \frac{\mu - h}{\mu - \overline{h}} \, \frac{m_\infty(\lambda) + \overline{h}}{m_\infty(\lambda) + h} = \frac{1 - iV(z)}{1 + iV(z)}. \tag{5.29}$$

Passing to the limit in (5.29) when $\lambda \to -\infty$ and taking into account that $V(-\infty) = \alpha - b$ and $m_\infty(-\infty) = \infty$ (see [14]) we obtain

$$\frac{\mu - h}{\mu - \overline{h}} = \frac{1 - i(\alpha - b)}{1 + i(\alpha - b)}.$$

Let us denote

$$a = \frac{1 - i(\alpha - b)}{1 + i(\alpha - b)}. \tag{5.30}$$

Solving (5.30) for μ yields

$$\mu = \frac{h - a\bar{h}}{1 - a}.$$

Substituting this value into (5.28) after simplification produces

$$\frac{x + iy - a(x - iy)x - (x^2 + y^2)(1 - a)}{x + iy - a(x - iy) - x(1 - a)} = \theta.$$

After straightforward calculations targeting to represent numerator and denominator of the last equation in standard form one obtains the following relation

$$x - (\alpha - b)\, y = \theta. \tag{5.31}$$

It was shown in [29] that the ϕ-sectoriality of the operator T_h and (5.20) lead to

$$\tan\phi = \frac{\mathrm{Im}\, h}{\mathrm{Re}\, h + m_\infty(-0)} = \frac{y}{x + m_\infty(-0)}. \tag{5.32}$$

Combining (5.31) and (5.32) one obtains

$$x - (\alpha - b)(x \tan\phi + m_\infty(-0)\, \tan\phi) = \theta,$$

or

$$x = \frac{\theta + (\alpha - b)m_\infty(-0)\,\tan\phi}{1 - (\alpha - b)\,\tan\phi}.$$

But $\tan\phi$ is also determined by (5.17). Direct substitution of

$$\tan\phi = \frac{b}{1 + \alpha(\alpha - b)}$$

into the above equation yields

$$x = \theta + \frac{\big[\theta + m_\infty(-0)\big]b(\alpha - b)}{1 + (\alpha - b)^2}.$$

Using the short notation and finalizing calculations we get

$$h = x + iy, \quad x = \theta + \frac{(\alpha - b)[\theta + m]b}{1 + (\alpha - b)^2}, \quad y = \frac{[\theta + m]b}{1 + (\alpha - b)^2}. \tag{5.33}$$

At this point we can use (5.33) to provide analytical and graphical interpretation of the parameter h in the restored operator T_h. Let

$$c = (\theta + m)b.$$

Again we consider three subcases.

Subcase 1. $b > 2$ Using basic algebra we transform (5.33) into

$$(x - \theta)^2 + \Big(y - \frac{c}{2}\Big)^2 = \frac{c^2}{4}. \tag{5.34}$$

Since in this case the parameter α belongs to the interval in (5.24), we can see that h traces the highlighted part of the circle on Figure 1 as α moves from $-\infty$ towards $+\infty$. We also notice that the removed point $(\theta, 0)$ corresponds to the value of $\alpha = \pm\infty$ while the points h_1 and h_2 correspond to the values $\alpha_1 = \frac{b - \sqrt{b^2-4}}{2}$ and $\alpha_2 = \frac{b + \sqrt{b^2-4}}{2}$, respectively (see Figure 1).

Subcase 2. $b < 2$ For every $\alpha \in (-\infty, +\infty)$ the restored operator T_h will be accretive and ϕ-sectorial for some $\phi \in (0, \pi/2)$. As we have mentioned above, the operator T_h achieves the largest angle of sectoriality when $\alpha = \frac{b}{2}$. In this particular case (5.33) becomes

$$h = x + iy, \quad x = \theta - \frac{2(\theta + m)b^2}{4 + b^2}, \quad y = \frac{4(\theta + m)b}{4 + b^2}. \tag{5.35}$$

The value of h from (5.35) is marked on Figure 2.

Subcase 3. $b = 2$ The behavior of parameter h in this case is depicted on Figure 3. It shows that in this case the function $V(z)$ can be realized using an extremal accretive T_h when $\alpha = 1$. The value of the parameter h according to (5.33) then becomes

$$h = x + iy, \quad x = -m, \quad y = \theta + m. \tag{5.36}$$

Clockwise direction of the circle again corresponds to the change of α from $-\infty$ to $+\infty$ and the marked value of h occurs when $\alpha = 1$.

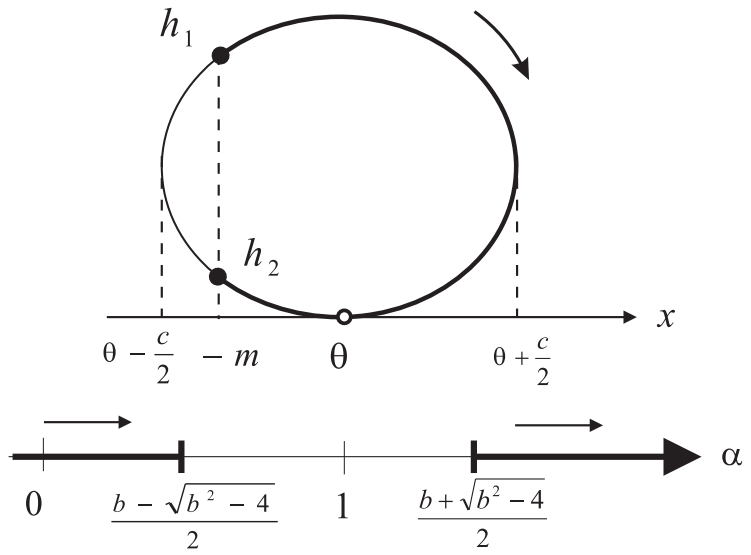

FIGURE 1. $b > 2$

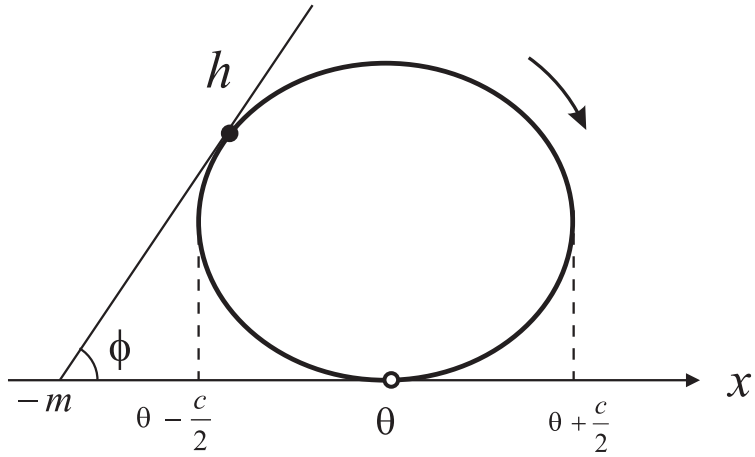

FIGURE 2. $b < 2$

Now we consider the second case.

Case 2. Here we assume that $\int_0^\infty \frac{d\tau(t)}{t} = \infty$. This means that our function $V(z)$ belongs to the class $SL_0^{-1}(R, K)$ and $b = \infty$. According to Theorem 5.7 and formulas (5.16) and (5.17), the restored operator T_h is accretive if and only if

$$\alpha \leq 0,$$

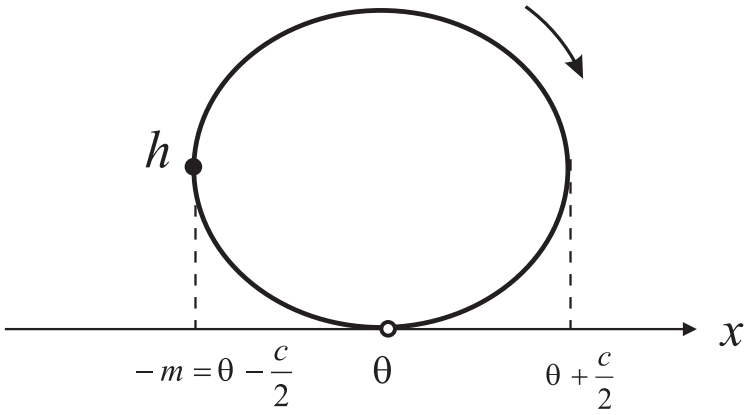

FIGURE 3. $b = 2$

and ϕ-sectorial if and only if $\alpha < 0$. It directly follows from (5.17) that the exact value of the angle ϕ is then found from

$$\tan\phi = -\frac{1}{\alpha}. \tag{5.37}$$

The latter implies that the restored operator T_h is extremal if $\alpha = 0$. This means that a function $V(z) \in SL_0^{-1}(R, K)$ is realized by a system with an extremal operator T_h if and only if

$$V(z) = \int_0^\infty \left(\frac{1}{t-z} - \frac{1}{t}\right) d\tau(t). \tag{5.38}$$

On the other hand since $\alpha \leq 0$ the function $V(z)$ is an inverse Stieltjes function of the class $S_0^{-1}(R)$. Applying realization theorems from [13] we conclude that $V(z)$ admits realization by an accumulative system Θ of the form (3.1) with $\mathbb{A}_R$ containing the Friedrichs extension A_F as a quasi-kernel. Here A_F is defined by (2.10). This yields

$$\theta = \frac{\mu x - (x^2 + y^2)}{\mu - x} = \infty, \tag{5.39}$$

and hence $\mu = x$. As in the beginning of the previous case we derive the formulas for x and y, where $h = x + iy$. Assuming that $\alpha \neq 0$ and using (5.32) and (5.37) leads to

$$x = \mu, \quad y = -\frac{x+m}{\alpha}. \tag{5.40}$$

To proceed, we first notice that our function $V(z)$ satisfies the conditions of Theorem 4.9 of [6]. Indeed, the inequality

$$\mu \geq \frac{(\operatorname{Im} h)^2}{m_\infty(-0) + \operatorname{Re} h} + \operatorname{Re} h,$$

that is required to apply this theorem, in our case turns into

$$\mu \geq -\frac{1}{\alpha} + \mu,$$

that is obvious if $\alpha < 0$. Applying Theorem 4.9 of [6] yields

$$\int_0^\infty \frac{d\tau(t)}{1+t^2} = \frac{\operatorname{Im} h}{|\mu - h|^2} \left(\sup_{y \in D(A_F)} \frac{|\mu y(a) - y'(a)|}{\left(\int_a^\infty (|y(x)|^2 + |l(y)|^2)\, dx \right)^{\frac{1}{2}}} \right)^2. \tag{5.41}$$

Taking into account that for the case of A_F

$$|\mu y(a) - y'(a)| = |y'(a)|$$

and setting

$$d^{1/2} = \sup_{y \in D(A_F)} \frac{|y'(a)|}{\left(\int_a^\infty (|y(x)|^2 + |l(y)|^2)\, dx \right)^{\frac{1}{2}}}, \tag{5.42}$$

we obtain

$$\frac{\operatorname{Im} h}{|\mu - h|^2}\, d = \int_0^\infty \frac{d\tau(t)}{1+t^2}. \tag{5.43}$$

Considering that $\operatorname{Im} h = y$ and $\mu = x$, solving (5.43) for y yields

$$y = \frac{d}{\int_0^\infty \frac{d\tau(t)}{1+t^2}}. \tag{5.44}$$

Consequently, equations (5.40) describing $h = x + iy$ take form

$$x = -m + \frac{\alpha\, d}{\int_0^\infty \frac{d\tau(t)}{1+t^2}}, \quad y = \frac{d}{\int_0^\infty \frac{d\tau(t)}{1+t^2}}. \tag{5.45}$$

The equations (5.45) above provide parametrical equations of the straight horizontal line shown on Figure 4. The connection between the parameters α and h in the accretive restored operator T_h is depicted in bold.

As we mentioned earlier the restored operator T_h is extremal if $\alpha = 0$. In this case formulas (5.45) become

$$x = -m, \quad y = \frac{d}{\int_0^\infty \frac{d\tau(t)}{1+t^2}}. \tag{5.46}$$

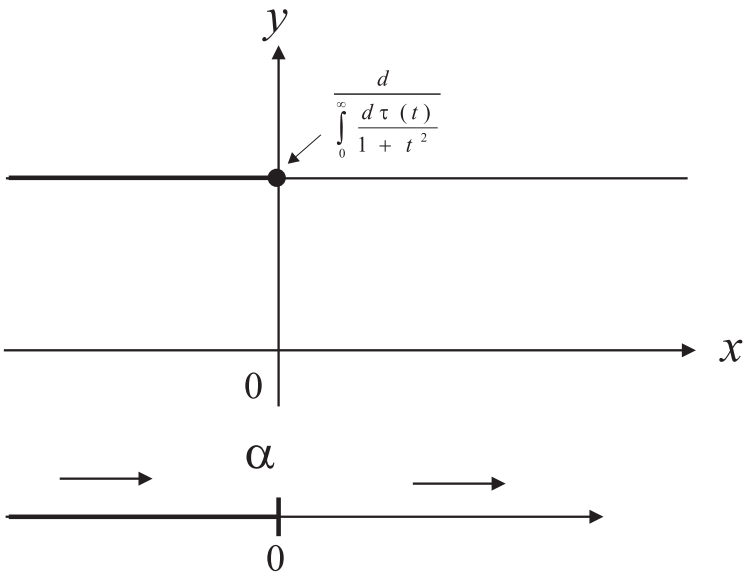

FIGURE 4. $b = \infty$

6. Realizing systems with Schrödinger operator

Now once we described all the possible outcomes for the restored accretive operator T_h, we can concentrate on the main operator $\mathbb{A}$ of the system (5.15). We recall that $\mathbb{A}$ is defined by formulas (2.6) and beside the parameter h above contains also parameter μ. We will obtain the behavior of μ in terms of the components of our function $V(z)$ the same way we treated the parameter h. As before we consider two major cases dividing them into subcases when necessary.

Case 1. Assume that $b = \int_0^\infty \frac{d\tau(t)}{t} < \infty$. In this case our function $V(z)$ belongs to the class $SL_{01}^{-1}(R, K)$. First we will obtain the representation of μ in terms of x and y, where $h = x + iy$. We recall that

$$\mu = \frac{h - a\bar{h}}{1 - a},$$

where a is defined by (5.30). By direct computations we derive that

$$a = \frac{1 - (\alpha - b)^2}{1 + (\alpha - b)^2} - \frac{2(\alpha - b)}{1 + (\alpha - b)^2} i, \quad 1 - a = \frac{2(\alpha - b)^2}{1 + (\alpha - b)^2} + \frac{2(\alpha - b)}{1 + (\alpha - b)^2} i,$$

and

$$h - a\bar{h} = \left(\frac{2(\alpha - b)^2}{1 + (\alpha - b)^2} x + \frac{2(\alpha - b)}{1 + (\alpha - b)^2} y \right) + \left(\frac{2}{1 + (\alpha - b)^2} y + \frac{2(\alpha - b)}{1 + (\alpha - b)^2} x \right) i.$$

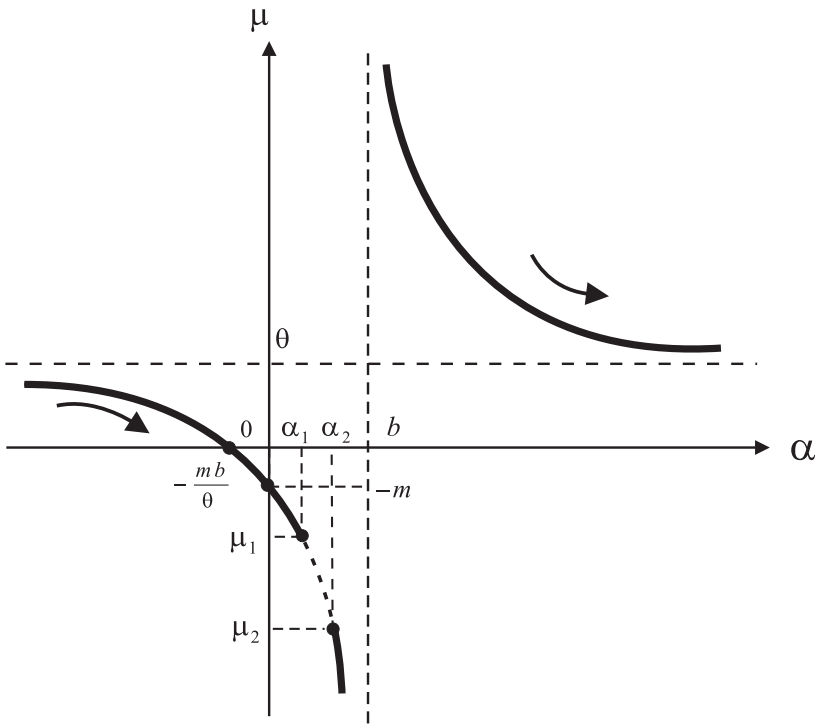

FIGURE 5. $b > 2$

Plugging the last two equations into the formula for μ above and simplifying we obtain

$$\mu = x + \frac{y}{\alpha - b}. \tag{6.1}$$

We recall that during the present case x and y parts of h are described by the formulas (5.33).

Once again we elaborate in three subcases.

Subcase 1. $b > 2$ As we have shown this above, the formulas (5.33) can be transformed into equation of the circle (5.34). In this case the parameter α belongs to the interval in (5.24), the accretive operator T_h corresponds to the values of h shown in the bold part of the circle on Figure 1 as α moves from $-\infty$ towards $+\infty$.

Substituting the expressions for x and y from (5.33) into (6.1) and simplifying we get

$$\mu = \theta + \frac{(\theta + m)b}{\alpha - b}. \tag{6.2}$$

The connection between values of α and μ is depicted on Figure 5.

We note that $\mu = 0$ when $\alpha = -\frac{mb}{\theta}$. Also, the endpoints

$$\alpha_1 = \frac{b - \sqrt{b^2 - 4}}{2} \quad \text{and} \quad \alpha_2 = \frac{b + \sqrt{b^2 - 4}}{2}$$

of α-interval (5.24) are responsible for the μ-values

$$\mu_1 = \theta + \frac{(\theta + m)b}{\alpha_1} \quad \text{and} \quad \mu_2 = \theta + \frac{(\theta + m)b}{\alpha_2}.$$

The values of μ that are acceptable parameters of operator $\mathbb{A}$ of the restored system with an accretive operator T_h make the bold part of the hyperbola on Figure 5. It follows from Theorems 4.4 and 4.4 that the operator $\mathbb{A}$ of the form (2.6) is accumulative if and only if $\alpha \leq 0$ and thus μ belongs to the part of the left branch on the hyperbola where $\alpha \in (-\infty, 0]$. We note that Figure 5 shows the case when $-m < 0$, $\theta > 0$, and $\theta > -m$. Other possible cases, such as $(-m < 0, \theta < 0, \theta > -m)$, $(-m < 0, \theta = 0)$, and $(m = 0, \theta > 0)$ require corresponding adjustments to the graph shown in the picture 5.

Subcase 2. $b < 2$ For every $\alpha \in (-\infty, +\infty)$ the restored operator T_h will be accretive and ϕ-sectorial for some $\phi \in (0, \pi/2)$. As we have mentioned above, the operator T_h achieves the largest angle of sectoriality when $\alpha = \frac{b}{2}$. In this particular case (5.33) becomes (5.35). Substituting $\alpha = b/2$ and (5.35) into (6.1) we obtain

$$\mu = -(\theta + 2m). \tag{6.3}$$

This value of μ from (6.3) is marked on Figure 6. The corresponding operator $\mathbb{A}$ of the realizing system is based on these values of parameters h and μ.

Subcase 3. $b = 2$ The behavior of parameter μ in this case is also shown on Figure 6. It was shown above that in this case the function $V(z)$ can be realized using an extremal accretive T_h when $\alpha = 1$. The values of the parameters h and μ then become

$$h = x + iy, \quad x = -m, \quad y = \theta + m, \quad \mu = -(\theta + 2m).$$

The value of μ above is marked on the left branch of the hyperbola and occurs when $\alpha = 1 = b/2$.

Case 2. Again we assume that $\int_0^\infty \frac{d\tau(t)}{t} = \infty$. Hence $V(z) \in SL_0^{-1}(R, K)$ and $b = \infty$. As we mentioned above the restored operator T_h is accretive if and only if $\alpha \leq 0$ and ϕ-sectorial if and only if $\alpha < 0$. It is extremal if $\alpha = 0$. The values of x and y, were already calculated and are given in (5.45). In particular, the value for μ is given by

$$\mu = x = -m + \frac{\alpha\, d}{\int_0^\infty \frac{d\tau(t)}{1+t^2}}. \tag{6.4}$$

where d is defined in (5.42). Figure 7 gives graphical representation of this case. The left bold part of the line corresponds to the values of μ that yield an accumulative realizing system. If $m = 0$ then the line passes through the origin and the graph

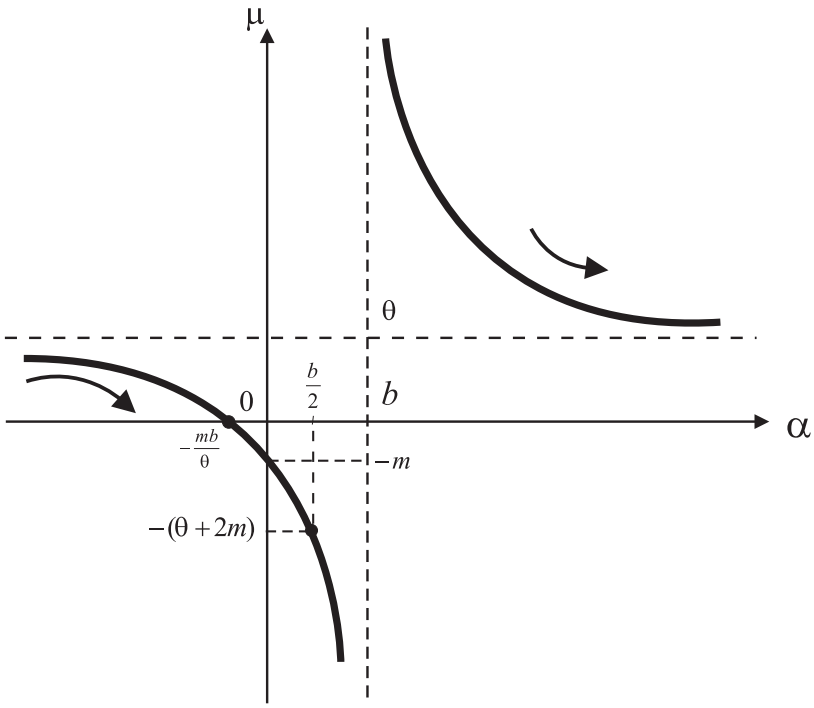

FIGURE 6. $b < 2$ and $b = 2$

should be adjusted accordingly. In the case when $\alpha = 0$ and T_h is extremal we have $\mu = m$.

Example

We conclude this paper with simple illustration. Consider a function

$$V(z) = i\sqrt{z}. \tag{6.5}$$

A direct check confirms that $V(z)$ is an inverse Stieltjes function. It can be shown (see [25] pp. 140–142) that the inversion formula

$$\tau(\lambda) = C + \lim_{y \to 0} \frac{1}{\pi} \int_0^{\lambda} \operatorname{Im} \left(i\sqrt{x + iy} \right) dx \tag{6.6}$$

describes the distribution function for a self-adjoint operator

$$\begin{cases} \tilde{B}_\infty y = -y'' \\ y(0) = 0. \end{cases}$$

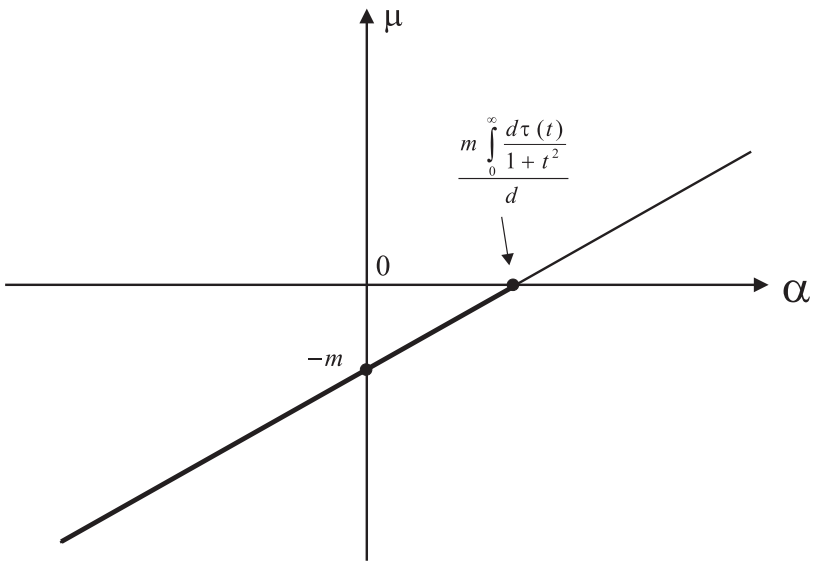

FIGURE 7. $b = \infty$

The corresponding to $\tilde{B}_\infty$ symmetric operator is

$$\begin{cases} B_\infty y = -y'' \\ y(0) = y'(0) = 0. \end{cases} \tag{6.7}$$

It was also shown in [25] that $\tau(\lambda) = 0$ for $\lambda \le 0$ and

$$\tau'(\lambda) = \frac{1}{\pi}\sqrt{\lambda} \quad \text{for} \quad \lambda > 0. \tag{6.8}$$

By direct calculations one can confirm that

$$V(z) = \int_0^\infty \left(\frac{1}{t-z} - \frac{1}{t}\right) d\tau(t) = i\sqrt{z},$$

and that

$$\int_0^\infty \frac{d\tau(t)}{t} = \int_0^\infty \frac{dt}{\pi\sqrt{t}} = \infty.$$

It is also clear that the constant term in the integral representation (5.7) is zero, i.e., $\alpha = 0$.

Let us assume that $\tau(t)$ satisfies our definition of spectral distribution function of the pair B_∞, $\tilde{B}_\infty$ given in Section 5. Operating under this assumption, we proceed to restore parameters h and μ and apply formulas (5.45) for the values $\alpha = 0$ and $m = m_\infty(-0) = 0$ (see [6]). This yields $x = 0$. To obtain y we first find the value of

$$\int_0^\infty \frac{d\tau(t)}{1+t^2} = \frac{1}{\sqrt{2}},$$

and then use formula (5.42) to get the value of d. This yields $d = 1/\sqrt{2}$. Consequently,

$$y = \frac{d}{\int_0^\infty \frac{d\tau(t)}{1+t^2}} = 1,$$

and hence $h = yi = i$. From (6.4) we have that $\mu = 0$ and (2.6) becomes

$$\mathbb{A}\, y = -y'' - [iy'(0) + y'(0)]\delta'(x). \tag{6.9}$$

The operator T_h in this case is

$$\begin{cases} T_h y = -y'' \\ y'(0) = iy(0). \end{cases}$$

The channel vector g of the form (3.11) then equals

$$g = \delta'(x), \tag{6.10}$$

satisfying

$$\operatorname{Im} \mathbb{A} = \frac{\mathbb{A} - \mathbb{A}^*}{2i} = KK^* = (., g)g,$$

and channel operator $Kc = cg$, $(c \in \mathbb{C})$ with

$$K^* y = (y, g) = y'(0). \tag{6.11}$$

The real part of $\mathbb{A}$

$$\operatorname{Re} \mathbb{A}\, y = -y'' - y(0)\delta'(x)$$

contains the self-adjoint quasi-kernel

$$\begin{cases} \widehat{A} y = -y'' \\ y(0) = 0. \end{cases}$$

A system of the Livšic type with Schrödinger operator of the form (5.15) that realizes $V(z)$ can now be written as

$$\Theta = \begin{pmatrix} \mathbb{A} & K & 1 \\ \mathcal{H}_+ \subset L_2[a, +\infty) \subset \mathcal{H}_- & & \mathbb{C} \end{pmatrix}.$$

where $\mathbb{A}$ and K are defined above. Now we can back up our assumption on $\tau(t)$ to be the spectral distribution function of the pair B_∞, $\tilde{B}_\infty$. Indeed, calculating the function $V_\Theta(z)$ for the system Θ above directly via formula (3.16) with $\mu = 0$ and comparing the result to $V(z)$ gives the exact value of $h = i$. Using the uniqueness of the unitary mapping U in the definition of spectral distribution function (see Remark 5.6 of [14]) we confirm that $\tau(t)$ is the spectral distribution function of the pair B_∞, $\tilde{B}_\infty$.

Remark 6.1. All the derivations above can be repeated for an inverse Stieltjes-like function

$$V(z) = \alpha + i\sqrt{z}, \qquad -\infty < \alpha < +\infty,$$

with very minor changes. In this case the restored values for h and μ are described as follows:

$$h = x + iy, \quad x = \alpha, \quad y = 1, \quad \mu = \alpha.$$

The dynamics of changing h according to changing α is depicted on Figure 4 where the horizontal line has a y-intercept of 1. The behavior of μ is described by a sloped line $\mu = \alpha$ (see Figure 7 with $m = 0$). In the case when $\alpha \leq 0$ our function becomes inverse Stieltjes and the restored system Θ is accretive. The operators $\mathbb{A}$ and K of the restored system are given according to the formulas (2.6) and (3.13), respectively.

References

[1] N.I. Akhiezer and I.M. Glazman. *Theory of linear operators.* Pitman Advanced Publishing Program, 1981.

[2] D. Alpay, I. Gohberg, M.A. Kaashoek, A.L. Sakhnovich, "Direct and inverse scattering problem for canonical systems with a strictly pseudoexponential potential", Math. Nachr. 215 (2000), 5–31.

[3] D. Alpay and E.R. Tsekanovskiĭ, "Interpolation theory in sectorial Stieltjes classes and explicit system solutions", Lin. Alg. Appl., 314 (2000), 91–136.

[4] Yu.M. Arlinskiĭ. *On regular* $(*)$*-extensions and characteristic matrix-valued functions of ordinary differential operators.* Boundary value problems for differential operators, Kiev, 3–13, 1980.

[5] Yu. Arlinskiĭ and E. Tsekanovskiĭ. *Regular* $(*)$*-extension of unbounded operators, characteristic operator-functions and realization problems of transfer functions of linear systems.* Preprint, VINITI, Dep.-2867, 72p., 1979.

[6] Yu.M. Arlinskiĭ and E.R. Tsekanovskiĭ, "Linear systems with Schrödinger operators and their transfer functions", Oper. Theory Adv. Appl., 149, 2004, 47–77.

[7] D. Arov, H. Dym, "Strongly regular J-inner matrix-valued functions and inverse problems for canonical systems", Oper. Theory Adv. Appl., 160, Birkhäuser, Basel, (2005), 101–160.

[8] D. Arov, H. Dym, "Direct and inverse problems for differential systems connected with Dirac systems and related factorization problems", Indiana Univ. Math. J. 54 (2005), no. 6, 1769–1815.

[9] J.A. Ball and O.J. Staffans, "Conservative state-space realizations of dissipative system behaviors", Integr. Equ. Oper. Theory, 54 (2006), no. 2, 151–213.

[10] H. Bart, I. Gohberg, and M.A. Kaashoek, *Minimal Factorizations of Matrix and Operator Functions*, Operator Theory: Advances and Applications, Vol. 1, Birkhäuser, Basel, 1979.

[11] S.V. Belyi and E.R. Tsekanovskiĭ, "Realization theorems for operator-valued R-functions", Oper. Theory Adv. Appl., 98 (1997), 55–91.

[12] S.V. Belyi and E.R. Tsekanovskiĭ, "On classes of realizable operator-valued R-functions", Oper. Theory Adv. Appl., 115 (2000), 85–112.

[13] S.V. Belyi, S. Hassi, H.S.V. de Snoo, and E.R. Tsekanovskiĭ, "On the realization of inverse of Stieltjes functions", Proceedings of MTNS-2002, University of Notre Dame, CD-ROM, 11p., 2002.

[14] S.V. Belyi and E.R. Tsekanovskiĭ, "Stieltjes-like functions and inverse problems for systems with Schrödinger operator", Operators and Matrices., vol. **2**, no. 2, (2008), pp. 265–296.

[15] Brodskiĭ, M.S., *Triangular and Jordan Representations of Linear Operators*, Amer. Math. Soc., Providence, RI, 1971.

[16] M.S. Brodskiĭ, M.S. Livšic. *Spectral analysis of non-selfadjoint operators and intermediate systems*, Uspekhi Matem. Nauk, **XIII**, no. 1 (79), [1958], 3–84.

[17] V.A. Derkach, M.M. Malamud, and E.R. Tsekanovskiĭ, *Sectorial extensions of a positive operator, and the characteristic function*, Sov. Math. Dokl. **37**, 106–110 (1988).

[18] F. Gesztesy and E.R. Tsekanovskiĭ, "On matrix-valued Herglotz functions", Math. Nachr., 218 (2000), 61–138.

[19] F. Gesztesy, N.J. Kalton, K.A. Makarov, E. Tsekanovskiĭ, "Some Applications of Operator-Valued Herglotz Functions", Operator Theory: Advances and Applications, **123**, Birkhäuser, Basel, (2001), 271–321.

[20] S. Khrushchev, "Spectral Singularities of dissipative Schrödinger operator with rapidly decreasing potential", Indiana Univ. Math. J., 33 no. 4, (1984), 613–638.

[21] I.S. Kac and M.G. Krein, *R-functions – analytic functions mapping the upper halfplane into itself*, Amer. Math. Soc. Transl. (2) **103**, 1–18 (1974).

[22] Kato T.: Perturbation Theory for Linear Operators, Springer-Verlag, 1966

[23] B.M. Levitan, *Inverse Sturm-Liouville Problems*, VNU Science Press, Utrecht, 1987.

[24] M.S. Livšic, *Operators, oscillations, waves*, Moscow, Nauka, 1966 (Russian).

[25] M.A. Naimark, *Linear Differential Operators II*, F. Ungar Publ., New York, 1968.

[26] O.J. Staffans, "Passive and conservative continuous time impedance and scattering systems, Part I: Well-posed systems", Math. Control Signals Systems, 15, (2002), 291–315.

[27] O.J. Staffans, *Well-posed linear systems*, Cambridge University Press, Cambridge, 2005.

[28] E.R. Tsekanovskiĭ, "Accretive extensions and problems on Stieltjes operator-valued functions realizations", Oper. Theory Adv. Appl., 59 (1992), 328–347.

[29] E.R. Tsekanovskiĭ. "Characteristic function and sectorial boundary value problems", Investigation on geometry and math. analysis, Novosibirsk, **7**, (1987), 180–194.

[30] E.R. Tsekanovskiĭ and Yu.L. Shmul'yan, "The theory of bi-extensions of operators on rigged Hilbert spaces. Unbounded operator colligations and characteristic functions", Russ. Math. Surv., 32 (1977), 73–131.

Sergey V. Belyi
Department of Mathematics
Troy State University
Troy, AL 36082, USA
e-mail: sbelyi@troy.edu

Eduard R. Tsekanovskii
Department of Mathematics
Niagara University, NY 14109, USA
e-mail: tsekanov@niagara.edu

Operator Theory:
Advances and Applications, Vol. 197, 51–76

Bi-Isometries and Commutant Lifting

Hari Bercovici, Ronald G. Douglas and Ciprian Foias

In memory of M.S. Livsic, one of the founders of modern operator theory

Abstract. In a previous paper, the authors obtained a model for a bi-isometry, that is, a pair of commuting isometries on complex Hilbert space. This representation is based on the canonical model of Sz.-Nagy and the third author. One approach to describing the invariant subspaces for such a bi-isometry using this model is to consider isometric intertwining maps from another such model to the given one. Representing such maps requires a careful study of the commutant lifting theorem and its refinements. Various conditions relating to the existence of isometric liftings are obtained in this note, along with some examples demonstrating the limitations of our results.

Mathematics Subject Classification (2000). 46G15, 47A15, 47A20, 47A45, 47B345.

Keywords. Bi-isometries, commuting isometries, canonical model, commutant lifting, intertwining maps, invariant subspaces.

1. Introduction

The geometry of complex Hilbert space is especially transparent. In particular, all Hilbert spaces of the same dimension are isometrically isomorphic. One consequence is the simple structure of isometric operators on Hilbert space as was discovered by von Neumann in his study of symmetric operators in connection with quantum mechanics. A decade later, this decomposition was rediscovered by Wold who made it the basis for his study of stationary stochastic processes. Another decade later, Beurling obtained his iconic result on invariant subspaces for the unilateral shift operator. While his proof did not rely on the structure of isometries, later works showed that the result could be established using it. In the fifties, Sz.-Nagy demonstrated that all contraction operators on Hilbert space have a unique minimal unitary dilation. The application of structure theory for isome-

Research partially supported by grants from the National Science Foundation.

tries to this unitary operator is one starting point for the canonical model theory of Sz.-Nagy and the third author [11]. Much of the development of this theory, including the lifting theorem for intertwining operators and the parametrization of the possible lifts, can be viewed as exploiting and refining the structure theory of isometric operators on complex Hilbert space.

The study of commuting n-tuples of isometries is not so simple, even for $n = 2$. This paper makes a contribution to this theory. The starting point is the model introduced implicitly in [1] for a bi-isometry or a pair of commuting isometries. We now describe the model explicitly. Let $\{\Theta(z), \mathcal{E}, \mathcal{E}\}$ be a contractive operator-valued analytic function ($z \in \mathbb{D}$) and set $\Delta(\zeta) = (I - \Theta(\zeta)^*\Theta(\zeta))^{1/2}$, $\zeta \in \partial\mathbb{D}$. Define the Hilbert space

$$\mathcal{H}_\Theta = H^2(\mathcal{E}) \oplus H^2(\overline{\Delta L^2(\mathcal{E})}) \tag{1.1}$$

and the operators

$$V_\Theta(f \oplus g) = f_1 \oplus g_1, W_\Theta(f \oplus g) = f_2 \oplus g_2, \tag{1.1a}$$

where

$$f_1(z) = zf(z), \quad f_2(z) = \Theta(z)f(z) \qquad (z \in \mathbb{D}) \tag{1.1b}$$

$$g_1(w, \zeta) = \zeta g(w, \zeta), g_2(w, \zeta) = \Delta(\zeta)f(\zeta) + wg(w, \zeta) \qquad (w \in \mathbb{D}, \zeta \in \partial\mathbb{D}). \tag{1.1c}$$

Then (V_Θ, W_Θ) is a bi-isometry such that there is no nonzero reducing subspace for (V_Θ, W_Θ) on which V_Θ is unitary.

In [2] we have shown that *any bi-isometry (V, W), for which there is no nonzero reducing subspace $\mathcal{N}$ such that $V|\mathcal{N}$ is unitarily equivalent to a bi-isometry (V_Θ, W_Θ), where $\Theta(\cdot)$ is uniquely determined up to unitary equivalence.* (Note that the terminology and the notations are as in [11].)

An important part of the study of this model is a description of all invariant subspaces of the bi-isometry (V_Θ, W_Θ). To this end we first describe all the contractive operators Y intertwining two bi-isometries $(V_{\Theta_1}, W_{\Theta_1})$ and (V_Θ, W_Θ); that is, $Y \in \mathcal{L}(\mathcal{H}_{\Theta_1}, \mathcal{H}_\Theta)$ and

$$YV_{\Theta_1} = V_\Theta Y, \quad YW_{\Theta_1} = W_\Theta Y. \tag{1.2}$$

Let P denote the orthogonal projection of $\mathcal{H}_\Theta$ onto $H^2(\mathcal{E})(\simeq H^2(\mathcal{E}) \oplus \{0\} \subset \mathcal{H}_\Theta)$. Then there exists a unique contractive analytic operator-valued function $\{A(\cdot), \mathcal{E}_1, \mathcal{E}\}$ such that

$$(PYh_1)(z) = A(z)h_1(z) \qquad (z \in \mathbb{D})$$

for all $h_1 \in H^2(\mathcal{E}_1)(\simeq H^2(\mathcal{E}_1) \oplus \{0\} \subset \mathcal{H}_{\Theta_1})$. Conversely, given such a contractive analytic function $A(\cdot)$, there exists a contractive intertwining operator Y, but it is not unique. Using the Commutant Lifting Theorem, one can describe completely the set of such intertwining contractions. The description involves an analytic operator-valued function $\{R(\cdot), \mathcal{R}, \mathcal{R}'\}$, called the free Schur contraction in Section 2. Here, the spaces $\mathcal{R}$ and $\mathcal{R}'$ are called residual spaces, and they are entirely determined by the functions Θ_1, Θ and A. If $\mathcal{M}$ is a common invariant subspace for

the bi-isometry (V_Θ, W_Θ), then defining $U_1 = V_\Theta|\mathcal{M}$ and $U_2 = W_\Theta|\mathcal{M}$ yields a bi-isometry (U_1, U_2) on $\mathcal{M}$. Moreover, the inclusion map $X\colon\ \mathcal{M} \to \mathcal{H}_\Theta$ is an isometric intertwining map. Conversely, if Y is an isometric intertwining map from a model $(V_{\Theta_1}, W_{\Theta_1})$ on $\mathcal{H}_{\Theta_1}$ to $\mathcal{H}_\Theta$, then the range of Y is a common invariant subspace for the bi-isometry (V_Θ, W_Θ). Hence, the problem of describing the common invariant subspaces for (V_Θ, W_Θ) is closely related to describing the isometric intertwining maps from some model $(V_{\Theta_1}, W_{\Theta_1})$ to (V_Θ, W_Θ).

Thus the description of all the invariant subspaces of (V_Θ, W_Θ) is intimately connected to the determination of the class of those free Schur contractions for which the corresponding operator Y is an isometry. As yet we have not found a completely satisfactory characterization of that set. In this note we present our contributions to this problem with the hope that they may be instrumental in the discovery of an *easily applicable* characterization.

In the next section we provide a description of the commutant lifting theorem focusing on the aspects relevant to our problem. In Section 3 the analytical details are taken up, while in the fourth section we state our results on isometric intertwining maps. In the final section, we apply these results to the question of invariant subspaces in those cases in which our results are effective. We conclude with a number of open questions and future directions for study.

2. A short review of the Commutant Lifting Theorem

Let $T' \in \mathcal{L}(\mathcal{H}')$ be a completely nonunitary (c.n.u.) contraction and $T \in \mathcal{L}(\mathcal{H})$ an isometry; here $\mathcal{H}, \mathcal{H}'$ are (separable) Hilbert spaces. Furthermore, let $X \in \mathcal{L}(\mathcal{H}, \mathcal{H}')$ be a contraction intertwining T and T'; that is,

$$T'X = XT, \quad \|X\| \leq 1. \tag{2.1}$$

Let $U' \in \mathcal{L}(\mathcal{K}')$ be a minimal isometric lifting of T'. In other words, if P' denotes the orthogonal projection $\mathcal{K}'$ onto $\mathcal{H}'$ and I' denotes the identity operator on $\mathcal{K}'$, we have

$$P'U' = T'P', \tag{2.2}$$

$$U'^*U' = I', \tag{2.2a}$$

and

$$\mathcal{K}' = \bigvee_{n=0}^{\infty} U'^n \mathcal{H}'. \tag{2.2b}$$

Since $\mathcal{K}'$ is essentially unique, one can take

$$\mathcal{K}' = \mathcal{H}' \oplus H^2(\mathcal{D}_{T'}) \tag{2.2c}$$

$$U'(h' \oplus f(\cdot)) = T'h' \oplus (D_{T'}h' + \cdot f(\cdot)) \tag{2.2d}$$

for all $h' \in \mathcal{H}', f(\cdot) \in H^2(\mathcal{D}_{T'})$. Recall that $D_{T'} = (I - T'^*T')^{1/2}$ and $\mathcal{D}_{T'} = (D_{T'}\mathcal{H}')^-$.

In its original form [9, 10], *the Commutant Lifting Theorem asserts that there exists an operator* $X \in \mathcal{L}(\mathcal{H}, \mathcal{K}')$ *satisfying the following properties*

$$U'Y = YT, \tag{2.3}$$

$$\|Y\| \leq 1, \tag{2.3a}$$

$$P'Y = X. \tag{2.3b}$$

Such an operator Y is called a *contractive intertwining lifting of* X. In this study we need a tractable classification of all these liftings. To this aim we introduce the *isometry*

$$\omega : (D_X T\mathcal{H})^- \to \{D_{T'} Xh \oplus D_X h\colon\ h \in \mathcal{H}\}^-, \tag{2.4}$$

obtained by closing the linear operator

$$\omega_0\colon\ D_X Th \mapsto D_{T'} Xh \oplus D_X h \qquad (h \in \mathcal{H}); \tag{2.4a}$$

and the partial isometry operator $\bar{\omega} \in \mathcal{L}(\mathcal{D}_X, \mathcal{D}_{T'})$ defined by

$$\bar{\omega}|(D_X T\mathcal{H})^- = \omega, \quad \bar{\omega}|(\mathcal{D}_X \ominus (D_X T\mathcal{H})^-) = 0. \tag{2.4b}$$

This operator obviously satisfies

$$\ker \bar{\omega} = \mathcal{D}_X \ominus (D_X T\mathcal{H})^-, \quad \ker \bar{\omega}^* = (\mathcal{D}_{T'} \oplus \mathcal{D}_X) \ominus \operatorname{ran} \omega. \tag{2.4c}$$

Also we will denote by Π and Π' the operators on $\mathcal{D}_{T'} \oplus \mathcal{D}_X$ defined by

$$\Pi'(d' \oplus d) = d', \quad \Pi(d' \oplus d) = d \qquad (d' \in \mathcal{D}_{T'}, d \in \mathcal{D}_X). \tag{2.4d}$$

With this preparation we can state the needed description (see [4], Ch. VI).

Proposition 2.1.

(i) *Any contractive intertwining lifting* Y *of* X *is of the form*

$$Y = \begin{bmatrix} X \\ \Gamma(\cdot) D_X \end{bmatrix}\colon\ \mathcal{H} \mapsto \mathcal{H} \oplus H^2(\mathcal{D}_{T'}) = \mathcal{K}', \tag{2.5}$$

where $\Gamma(\cdot)$ *is given by the formula*

$$\Gamma(z) = \Pi' W(z)(1 - z\Pi W(z))^{-1} \qquad (z \in \mathbb{D}), \tag{2.5a}$$

where

$$W(z)\colon\ \mathcal{D}_X \mapsto \mathcal{D}_{T'} \oplus \mathcal{D}_X \qquad (z \in \mathbb{D}) \tag{2.5b}$$

is a contractive analytic function satisfying

$$W(z)|(D_X T\mathcal{H})^- = \omega \qquad (z \in \mathbb{D}). \tag{2.5c}$$

(ii) *Conversely, for any* $W(\cdot)$ *satisfying the above conditions, the formulas* (2.5) *and* (2.5a) *yield a contractive intertwining lifting* Y *of* X.

(iii) *The correspondence between* Y *and* $W(\cdot)$ *is one-to-one.*

The function $W(\cdot)$ is called the *Schur contraction* of Y, and Y is called the contractive intertwining lifting associated to $W(\cdot)$. It is immediate that the Schur contraction is uniquely determined by its restriction

$$R(z) = W(z)|\ker\bar{\omega} : \ker\bar{\omega} \to \ker\bar{\omega}^* \qquad (z \in \mathbb{D}). \tag{2.6}$$

Any contractive analytic function $\{R(\cdot), \ker\bar{\omega}, \ker\bar{\omega}^*\}$ determines a Schur contraction. The function $R(\cdot)$ will be called the *free Schur contraction* of Y. Thus we have the following

Corollary 2.2. *The formulas* (2.5), (2.5a), (2.6) *establish a bijection between the set of all contractive intertwining liftings of X and the set of all free Schur contractions.*

As already stated, the main purpose of this paper is to study the free Schur contractions for which the associated contractive intertwining lifting is isometric; in particular, to find necessary conditions on T, T' and X for such an isometric lifting to exist.

3. Analytic considerations

Let $\mathcal{D}, \mathcal{D}'$ be two (separable) Hilbert spaces and let

$$W(z) = \begin{bmatrix} A(z) \\ B(z) \end{bmatrix} \in \mathcal{L}(\mathcal{D}, \mathcal{D} \oplus \mathcal{D}'), \left\| \begin{bmatrix} A(z) \\ B(z) \end{bmatrix} \right\| \leq 1 \qquad (z \in \mathbb{D}) \tag{3.1}$$

be analytic. Define an analytic function Γ by setting

$$\Gamma(z) = B(z)(I - zA(z))^{-1} \in \mathcal{L}(\mathcal{D}, \mathcal{D}') \qquad (z \in \mathbb{D}). \tag{3.2}$$

Lemma 3.1. *For all $d \in \mathcal{D}$ the function Γd defined by $\Gamma d(z) = \Gamma(z)d$ belongs to $\in H^2(\mathcal{D}')$, and*

$$\begin{aligned} \|\Gamma d\|^2_{H^2(\mathcal{D}')} &= \lim_{\rho \nearrow 1} \frac{1}{2\pi} \int_0^{2\pi} \|\Gamma(\rho e^{i\theta})d\|^2 d\theta \qquad (3.3) \\ &= \|d\|^2 - \lim_{\rho \nearrow 1} \left[\left(\frac{1}{\rho^2} - 1 \right) \frac{1}{2\pi} \int_0^{2\pi} \|(I - \rho e^{i\theta} A(\rho e^{i\theta}))^{-1} d\|^2 d\theta \right. \\ &\quad \left. + \frac{1}{2\pi} \int_0^{2\pi} \|D_{W(\rho e^{i\theta})}(I - \rho e^{i\theta} A(\rho e^{i\theta}))^{-1} d\|^2 d\theta \right]. \end{aligned}$$

Proof. For $\rho \in (0,1)$, the function $(1 - \rho z A(\rho z))^{-1} d$ is bounded, and in particular it belongs to $H^2(\mathcal{D})$. Moreover, it can be decomposed as a sum of two orthogonal vectors in $H^2(\mathcal{D})$ as follows:

$$(1 - \rho z A(\rho z))^{-1} d = d + \rho z A(\rho z)(1 - \rho z A(\rho z))^{-1} d.$$

Thus we have

$$\begin{aligned}
&\frac{1}{2\pi}\int_0^{2\pi}\|\Gamma(\rho e^{i\theta})d\|^2 d\theta \\
&\quad=\frac{1}{2\pi}\int_0^{2\pi}\|W(\rho e^{i\theta})(I-\rho e^{i\theta}A(\rho e^{i\theta}))^{-1}d\|^2 d\theta \\
&\qquad-\frac{1}{2\pi}\int_0^{2\pi}\|A(\rho e^{i\theta})(I-\rho e^{i\theta}A(\rho e^{i\theta}))^{-1}d\|^2 d\theta \\
&\quad=\frac{1}{2\pi}\int_0^{2\pi}\|W(\rho e^{i\theta})(I-\rho e^{i\theta}A(\rho e^{i\theta}))^{-1}d\|^2 d\theta \\
&\qquad-\frac{1}{2\pi}\int_0^{2\pi}\frac{1}{\rho^2}\|-d+(I-\rho e^{i\theta}A(\rho e^{i\theta}))^{-1}d\|^2 d\theta \\
&\quad=\frac{1}{2\pi}\int_0^{2\pi}\|W(\rho e^{i\theta})(I-\rho e^{i\theta}A(\rho e^{i\theta}))^{-1}d\|^2+\frac{1}{\rho^2}\|d\|^2 \\
&\qquad-\frac{1}{\rho^2}\frac{1}{2\pi}\int_0^{2\pi}\|(I-\rho e^{i\theta}A(\rho e^{i\theta}))^{-1}d\|^2 d\theta \\
&\quad=\frac{1}{\rho^2}\|d\|^2-\frac{1}{2\pi}\int_0^{2\pi}\|D_{W(\rho e^{i\theta})}(I-\rho e^{i\theta}A(\rho e^{i\theta})^{-1}d\|^2 d\theta \\
&\qquad-\left(\frac{1}{\rho^2}-1\right)\frac{1}{2\pi}\int_0^{2\pi}\|(I-\rho e^{i\theta}A(e^{i\theta}))^{-1}d\|^2 d\theta+\left(\frac{1}{\rho^2}-1\right)\|d\|^2,
\end{aligned}$$

from which (3.3) follows by letting $\rho\nearrow 1$. □

The following equality is an obvious consequence of Lemma 3.1.

Corollary 3.2. *The map* $\Gamma(\cdot)\colon\ d(\in\mathcal{D})\mapsto\Gamma d$ *is a contraction from* $\mathcal{D}$ *into* $H^2(\mathcal{D}')$ *and*

$$\begin{aligned}
\|D_{\Gamma(\cdot)}d\|^2=\lim_{\rho\nearrow 1}\Bigg[&\frac{1}{2\pi}\int_0^{2\pi}\|D_{W(\rho e^{i\theta})}(I-\rho e^{i\theta}A(\rho e^{i\theta}))^{-1}d\|^2 d\theta \qquad (3.3a)\\
&+\left(\frac{1}{\rho^2}-1\right)\frac{1}{2\pi}\int_0^{2\pi}\|(I-\rho e^{i\theta}A(\rho e^{i\theta}))^{-1}d\|^2 d\theta\Bigg] \qquad (d\in\mathcal{D}).
\end{aligned}$$

Lemma 3.3. *For all* $d\in\mathcal{D}$ *we have*

$$\begin{aligned}
\|d\|^2=\lim_{\rho\nearrow 1}\Bigg[&\frac{1}{2\pi}\int_0^{2\pi}\|D_{A(\rho e^{i\theta})}(I-\rho e^{i\theta}A(\rho e^{i\theta}))^{-1}d\|^2 d\theta \qquad (3.4)\\
&+\left(\frac{1}{\rho^2}-1\right)\frac{1}{2\pi}\int_0^{2\pi}\|(I-\rho e^{i\theta}A(\rho e^{i\theta}))^{-1}d\|^2 d\theta\Bigg].
\end{aligned}$$

Proof. For $d \in \mathcal{D}$ we have

$$\begin{aligned}
&\frac{1}{2\pi}\int_0^{2\pi} \|D_{A(\rho e^{i\theta})}(I-\rho e^{i\theta}A(\rho e^{i\theta}))^{-1}d\|^2 d\theta \\
&\quad = \frac{1}{2\pi}\int_0^{2\pi}\left[\|(I-\rho e^{i\theta}A(\rho e^{i\theta}))^{-1}d\|^2 - \frac{1}{\rho^2}\| - d + (I-\rho e^{i\theta}A(\rho e^{i\theta}))^{-1}d\|^2\right] d\theta \\
&\quad = \frac{1}{\rho^2}\|d\|^2 + \left(1-\frac{1}{\rho^2}\right)\frac{1}{2\pi}\int_0^{2\pi}\|(I-\rho e^{i\theta}A(\rho e^{i\theta}))^{-1}d\|^2 d\theta,
\end{aligned}$$

from which (3.4) readily follows. □

Lemma 3.4. *Let $d \in \mathcal{D}$ and set*

$$d(z) = (I - zA(z))^{-1}d = d_0 + zd_1 + \cdots + z^n d_n + \cdots \quad (z \in \mathbb{D}), \tag{3.5}$$

where $d_n \in \mathcal{D}$ and $d_0 = d$. If

$$\|d_n\| \to 0 \quad \textit{for} \quad n \to \infty, \tag{3.5a}$$

then we also have

$$I_\rho = \left(\frac{1}{\rho^2} - 1\right)\frac{1}{2\pi}\int_0^{2\pi}\|(I-\rho e^{i\theta}A(\rho e^{i\theta}))^{-1}d\|^2 d\theta \to 0 \quad \textit{for} \quad \rho \nearrow 1. \tag{3.5b}$$

Proof. Observe that

$$I_\rho = \frac{1+\rho}{\rho^2}(1-\rho)\sum_{n=0}^{\infty}\rho^{2n}\|d_n\|^2,$$

so for any $N = 1, 2, \ldots$ we have

$$\limsup_{\rho\nearrow 1} I_\rho \le 2\limsup_{\rho\nearrow 1}(1-\rho)\sum_{n=N}^{\infty}\rho^{2n}\|d_n\|^2 \le \max\{\|d_n\|^2 \colon\ n \ge N\},$$

which tends to 0 as $N \to \infty$. □

Lemma 3.5. *Let d and $d(z)$ $(z \in \mathbb{D})$ be as in Lemma 3.4. Then the equality $\|\Gamma d\|^2_{H^2(\mathcal{D}')} = \|d\|^2$ implies the convergence in* (3.5a).

Proof. Let

$$\Gamma(z)d = g_0 + zg_1 + \cdots + z^n g_n + \cdots \qquad (z \in \mathbb{D}),$$

where $g_n \in \mathcal{D}'$ and note that

$$W(z)d(z) = \begin{bmatrix} d_1 + zd_2 + z^2 d_2 + \cdots \\ g_0 + zg_1 + z^2 g_2 + \cdots \end{bmatrix} \qquad (z \in \mathbb{D}).$$

Thus for $n = 2, 3, \ldots$, we have

$$\begin{aligned}
&\left\| \begin{bmatrix} d_1 + zd_2 + \cdots + z^{n-1}d_n \\ g_0 + zg_1 + \cdots + z^{n-1}g_{n-1} \end{bmatrix} \right\|_{H^2}^2 \\
&\quad = \frac{1}{2\pi} \int_0^{2\pi} \left\| \begin{bmatrix} d_1 + e^{i\theta}d_2 + \cdots + e^{i(n-1)\theta}d_n \\ g_0 + e^{i\theta}g_2 + \cdots + e^{i(n-1)\theta}g_{n-1} \end{bmatrix} \right\|^2 d\theta \\
&\quad = \|d_1\|^2 + \|d_2\|^2 + \cdots + \|d_n\|^2 + \|g_0\|^2 + \cdots + \|g_{n-1}\|^2 \\
&\quad \leq \frac{1}{2\pi} \int_0^{2\pi} \|W(e^{i\theta})(d_0 + e^{i\theta}d_1 + \cdots + e^{i(n-1)\theta}d_{n-1})\|^2 d\theta \\
&\quad \leq \frac{1}{2\pi} \int_0^{2\pi} \|d_0 + e^{i\theta}d_1 + \cdots + e^{i(n-1)}d_{n-1}\|^2 d\theta \\
&\quad = \|d_0\|^2 + \|d_1\|^2 + \cdots + \|d_{n-1}\|^2,
\end{aligned}$$

where $d_0 = d$. Thus we obtain

$$\|d_n\|^2 + \|g_0\|^2 + \cdots + \|g_{n-1}\|^2 \leq \|d\|^2.$$

But since $\|\Gamma d\|_{H^2}^2 = \|g_0\|^2 + \cdots + \|g_{n-1}\|^2 + \cdots$, the above inequality and the assumption that $\|\Gamma d\|_{H^2} = \|d\|$ implies (3.5a). □

We can now state and prove the main result of this section.

Proposition 3.6. *The following sets of properties* (a), (b), (c) *and* (d) *are equivalent:*

(a) $\Gamma(\cdot)$ *is an isometry;*

(b) $W(\cdot)$ *and* $A(\cdot)$ *satisfy the conditions*

$$\lim_{\rho \nearrow 1} \int_0^{2\pi} \|D_{W(\rho e^{i\theta})}(I - \rho e^{i\theta}A(\rho e^{i\theta}))^{-1}d\|^2 = 0 \qquad (d \in \mathcal{D}) \tag{3.6}$$

and

$$\lim_{\rho \nearrow 1} (1 - \rho) \int_0^{2\pi} \|(I - \rho e^{i\theta}A(\rho e^{i\theta})^{-1}d\|^2 = 0 \qquad (d \in \mathcal{D}); \tag{3.6a}$$

(c) $W(\cdot)$ *and* $A(\cdot)$ *satisfy the condition* (3.6) *and*

$$\lim_{\rho \nearrow 1} \frac{1}{2\pi} \int_0^{2\pi} \|D_{A(\rho e^{i\theta})}(I - \rho e^{i\theta}A(\rho e^{i\theta}))^{-1}d\|^2 d\theta = \|d\|^2 \qquad (d \in \mathcal{D}); \tag{3.6b}$$

and

(d) $W(\cdot)$ *and* $A(\cdot)$ *satisfy* (3.6)*, and in the Taylor expansion*

$$(1 - zA(z))^{-1} = I + zD_1 + \cdots + z^n D_n + \cdots \qquad (z \in \mathbb{D})$$

we have $D_n \to 0$ *strongly* (*that is,* $\lim_{n\to\infty} \|D_n d\| = 0$ *for all* $d \in \mathcal{D}$).

Proof. The equivalence of properties (a) and (b) follows directly from the Corollary 3.2. This corollary and Lemma 3.5 show that property (d) implies property (a). The converse implication follows readily from the same corollary and Lemma 3.4. Finally, Lemma 3.3 shows that conditions (3.6a) and (3.6b) are equivalent and thus so are the sets of properties (b) and (c). □

Corollary 3.7. *Assume that $W(z_0) = 0$ for some $z_0 \in \mathbb{D}$. Then $\Gamma(\cdot)$ is an isometry if and only if* (3.6) *is valid.*

Proof. We have

$$\|W(z)\| \leq |z - z_0|/|1 - \bar{z}_0 z| \qquad (z \in \mathbb{D})$$

and consequently, for all $d \in \mathcal{D}$, we also have

$$\begin{aligned} \|D_{W(z)}d\|^2 &\geq (1 - |z - z_0|^2/|1 - \bar{z}_0 z|^2)\|d\|^2 \\ &\geq ((1 - |z_0|^2)/(1 + |z_0 z|)^2)(1 - |z|^2)\|d\|^2 \quad (z \in \mathbb{D}). \end{aligned}$$

It follows that

$$\begin{aligned} &\int_0^{2\pi} \|D_{W(\rho e^{i\theta})}(I - \rho e^{i\theta} A(\rho e^{i\theta}))^{-1} d\|^2 \\ &\quad \geq \frac{(1 - |z_0|^2)(1 + \rho)}{(1 + |z_0|\rho)^2}(1 - \rho) \int_0^{2\pi} \|(I - \rho e^{i\theta} A(\rho e^{i\theta}))^{-1} d\|^2 d\theta \end{aligned}$$

for all $d \in \mathcal{D}$. Thus (3.6) implies (3.6a). □

Another special case of Proposition 3.6 is given by the following.

Corollary 3.8. *Assume that $W(z) \equiv W_0 \in \mathcal{L}(\mathcal{D}, \mathcal{D} \oplus \mathcal{D}')$ $(z \in \mathbb{D})$. Then $\Gamma(\cdot)$ is an isometry if and only if*

(d′) *W_0 is an isometry and $A_0 = A(0)(= A(z)\ (z \in \mathbb{D}))$ is a $C_{0\bullet}$-contraction; that is,*

$$\|A_0^n d\| \to 0 \quad \text{for} \quad n \to \infty \qquad (d \in \mathcal{D}). \tag{3.7}$$

Proof. In this case $D_{W(z)}(1 - zA(z))^{-1}d = D_{W_0}(1 - zA(z))^{-1}d$ (as a $\mathcal{D}$-valued function of z) belongs to $H^2(\mathcal{D})$ for any $d \in \mathcal{D}$. Therefore, (3.6) implies that $D_{W_0}(1 - zA(z))^{-1}d = 0$ $(z \in \mathbb{D})$ and, in particular, $D_{W_0} = 0$; that is, W_0 is an isometry. Clearly, if this last property holds for W_0, then (3.6) is trivially true. According to the equivalence of the properties (a) and (d) in Proposition 3.6 and the fact that in the present case $D_n = A_0^n$ $(n = 0, 1, \ldots)$, the property (d′) above is equivalent to (a). □

Remark 3.9. In Corollary 3.8, the condition (3.7) is not superfluous. Indeed, if A_0 is any contraction for which (3.7) fails, then define $W = W_0 = \left[\begin{smallmatrix} C \\ D_C \end{smallmatrix}\right]$. This W is an isometry but $\Gamma(\cdot)$ is not isometric. Thus, in general, the (actually equivalent) conditions (3.6a) and (3.6b) are not superfluous.

Proposition 3.6 has an interesting connection to the Herglotz representation (cf. [3, p. 3]) of an analytic operator-valued function

$$F(z) \in \mathcal{L}(\mathcal{H}), \qquad z \in \mathbb{D}$$

(where $\mathcal{H}$ is a Hilbert space) such that

$$\text{Re } F(z)(= (F(z) + F(z)^*)/2) \geq 0, \quad \text{Im } F(0) = \left(\frac{F(0) - F(0)^*}{2}\right) = 0. \tag{3.8}$$

This representation is

$$F(z) = \int_{\partial\mathbb{D}} \frac{\zeta + z}{\zeta - z} E(d\zeta) \qquad (z \in \mathbb{D}), \tag{3.8a}$$

where $E(\cdot)$ is a positive operator-valued measure on $\partial\mathbb{D} = \{\zeta : |\zeta| = 1\}$, uniquely determined by F (an early occurrence of this representation is in [8, Theorem 3]).

To explicate that connection, we first observe that the function

$$F(z) = (I + zA(z))(I - zA(z))^{-1} \qquad (z \in \mathbb{D}), \tag{3.8b}$$

satisfies the inequality (3.8). Indeed, we have (with $d(z) = (1 - zA(z))^{-1}d, z \in \mathbb{D}$)

$$\begin{aligned} h_d(z) &:= ((\text{Re } F(z))d, d) = \text{Re}(F(z)d, d) \\ &= \text{Re}((I + zA(z))d(z), (I - zA(z))d(z)) = \\ &= \|d(z)\|^2 - \|zAd(z)\|^2 = \|D_{zA(z)}d(z)\|^2 \geq 0. \end{aligned} \tag{3.8c}$$

Thus our particular $F(\cdot)$ has a representation of the form (3.8a). Now from (3.8a) we easily infer

$$h_d(z) = \int_{\partial\mathbb{D}} \text{Re } \frac{\zeta + z}{\zeta - z}(E(d\zeta)d, d) \qquad (z \in \mathbb{D}). \tag{3.8d}$$

Therefore

$$\begin{aligned} \frac{1}{2\pi}\int_0^{2\pi} h_d(\rho e^{i\theta})d\theta &= \int_{\partial D} \left(\frac{1}{2\pi}\int_0^{2\pi} \text{Re } \frac{\zeta + \rho e^{i\theta}}{\zeta - \rho e^{i\theta}} d\theta\right)(E(d\zeta)d, d) = \\ &= \int_{\partial\mathbb{D}} (E(d\zeta)d, d) = \|d\|^2 \quad (\text{for } \rho \in (0, 1)). \end{aligned} \tag{3.8e}$$

Since $h_d(z)$ is a nonnegative harmonic function in $\mathbb{D}$, the limit

$$\lim_{\rho \nearrow 1} h_d(\rho e^{i\theta}) = h_d(e^{i\theta}) \tag{3.8f}$$

exists a.e. in $[0, 2\pi)$, and the absolutely continuous part of the measure $(E(\cdot)d, d)$ has density equal to $h_d(e^{i\theta})/2\pi$ a.e. (cf. [3, p. 5–6] or [6, Chapter 2])

It follows that the singular part $\mu_d(\cdot)$ of $(E(\cdot)d, d)$ satisfies

$$\mu_d(\partial\mathbb{D}) = \|d\|^2 - \int_0^{2\pi} h_d(e^{i\theta})\frac{d\theta}{2\pi}. \tag{3.8g}$$

Consequently, the following facts (a) and (b) are equivalent:

(a) the measure $E(\cdot)$ is absolutely continuous; and

(b) the relation

$$\|d\|^2 = \int_0^{2\pi} h_d(e^{i\theta})\frac{d\theta}{2\pi} \tag{3.8h}$$

holds for all $d \in \mathcal{D}$.

We recall that the function $\Gamma d(z) = \Gamma(z)d$ is in $H^2(\mathcal{D}')$ for every $d \in \mathcal{D}$. In particular, this function has radial limits a.e. on $\partial\mathbb{D}$.

Lemma 3.10. *Let $d \in \mathcal{D}$ be fixed, and define*

$$k_d(z) = \frac{1-|z|^2}{|z|^2}\|d(z)\|^2 + \frac{1}{|z|^2}\|D_{W(z)}d(z)\|^2 \qquad (z \in \mathbb{D}\backslash\{0\}), \tag{3.8i}$$

where

$$d(z) = (1 - zA(z))^{-1}d \qquad (z \in \mathbb{D}). \tag{3.8ia}$$

Then

$$\lim_{\varphi \nearrow 1} k_d(\rho e^{i\theta}) := k_d(e^{i\theta}) \tag{3.8j}$$

exists a.e. and

$$k_d(e^{i\theta}) + \|\Gamma d(e^{i\theta})\|^2 = h_d(e^{i\theta}) \quad \text{a.e.} \tag{3.8k}$$

Proof. We first observe that

$$k_d(z) + \|\Gamma(z)d\|^2 = \frac{1}{|z|^2}h_d(z) \qquad (z \in \mathbb{D}\backslash\{0\}). \tag{3.8l}$$

Indeed, for $z \in \mathbb{D}, z \neq 0$ we have

$$\begin{aligned}\|D_{W(z)}d(z)\|^2 + \|\Gamma(z)d\|^2 = \|D_{A(z)}d\|^2 = \|d(z)\|^2 \\ - \frac{1}{|z|^2}\|zA(z)d(z)\|^2 = \left(1 - \frac{1}{|z|^2}\right)\|d(z)\|^2 \\ + \frac{1}{|z|^2}\|D_{zA(z)}d(z)\|^2 = \left(1 - \frac{1}{|z|^2}\right)\|d(z)\|^2 + \frac{1}{|z|^2}h_d(z).\end{aligned}$$

Thus (3.8f) and (3.8l) imply (3.8j) and (3.8k). □

We can now give the following complement to Proposition 3.6 which establishes the connection between the absolute continuity of the measure $E(\cdot)$ in (3.8a) and the fact that Γ is an isometry when viewed as an operator from $\mathcal{D}$ to $H^2(\mathcal{D}')$.

Proposition 3.11. *The operator $\Gamma \in \mathcal{L}(\mathcal{D}, H^2(\mathcal{D}'))$ is an isometry if and only if the following two conditions are satisfied:*

The measure $E(\cdot)$ in the representation (3.8a) *of the function $F(\cdot)$* (3.8m)
defined in (3.8b) *is absolutely continuous, and*

$$k_d(e^{i\theta}) = 0 \quad \text{a.e.} \qquad (d \in \mathcal{D}), \tag{3.8n}$$

where $k_d(\cdot)$ is defined in (3.8i), (3.8j).

Proof. Assume first that Γ is an isometry. Then from (3.8k) we infer (for any $d \in \mathcal{D}$)

$$\frac{1}{2\pi}\int_0^{2\pi} k_d(e^{i\theta})d\theta + \|d\|^2 = \frac{1}{2\pi}\int_0^{2pi} h_d(e^{i\theta})d\theta \leq \|d\|^2. \tag{3.8o}$$

Since $k_d(e^{i\theta}) \geq 0$ a.e., (3.8o) implies (3.8n) and (3.8h). Consequently, (due to the equivalence of the facts (a) and (b) above; see the discussion preceding Lemma

3.10), we have that (3.8m) is also valid. Conversely, assume that both statements (3.8m) and (3.8n) are valid. Then using again (3.8k) we have

$$\|\Gamma d\|^2_{H^2(\mathcal{D}')} = \frac{1}{2\pi}\int_0^{2\pi} \|(\Gamma d)(e^{i\theta})\|^2 d\theta = \frac{1}{2\pi}\int_0^{2\pi} h_d(e^{i\theta})d\theta = \|d\|^2$$

for all $d \in \mathcal{D}$. Consequently, Γ is an isometry. □

For the investigation of the basic condition (3.6) we need to study first the operator-valued functions

$$K(z) = 1 - zA(z) \qquad (z \in \mathbb{D}) \tag{3.9}$$

and

$$J(z) = K(z)^{-1} \qquad (z \in \mathbb{D}). \tag{3.9a}$$

Lemma 3.12.

a) $K(z)$ $(z \in \mathbb{D})$ *is an outer function.*

b) $\|K(z)\| \leq 1$ $(z \in \mathbb{D})$ *if and only if* $A(z) = 0$ $(z \in \mathbb{D})$.

Proof. For $n = 1, 2, 3, \ldots$ we have

$$I - (zA(z))^n = (1 - zA(z))(1 + zA(z) + \cdots + (zA(z))^{n-1}) \qquad (z \in \mathbb{D}).$$

Consequently, for any function $h \in H^2(\mathcal{D})$ the function

$$h(z) - z^n A(z)^n h(z)$$

belongs to the range $\mathcal{R}$ of the operator of multiplication by $(1 - zA(z))$ on $H^2(\mathcal{D})$. But, since $\|A(z)^n h(z)\| \leq \|h(z)\|$ the functions

$$z^n A(z)^n h(z)$$

converge to zero weakly in $H^2(\mathcal{D})$. It follows that $\mathcal{R}$ is weakly dense in $H^2(\mathcal{D})$ and hence (being a subspace of $H^2(\mathcal{D})$) also strongly dense in $H^2(\mathcal{D})$. This proves the first part of the lemma. For the second part, assume that $\|K(z)\| \leq 1$ for all $z \in \mathbb{D}$. Then

$$\|K(z)d\| \leq \|d\| = \|K(0)d\|, \quad d \in \mathcal{D},$$

and the maximum principle forces $K(z)d = K(0)d = d$ $(d \in \mathcal{D})$. Thus $A(z) = 0$ for $(z \in \mathbb{D})$. □

Remark 3.13. Proposition 3.6 in the case $A(z) = 0$ $(z \in \mathbb{D})$ takes the following trivial form: $\Gamma(\cdot)$ is an isometry if and only if $W(\cdot)$ is inner. Therefore, from now on we will assume that $A(z) \not\equiv 0$, or equivalently that

$$\operatorname*{ess\,sup}_{|\zeta|=1} \|K(\zeta)\| > 1. \tag{3.9b}$$

Remark 3.14. It is worth noticing that the basic condition (3.6) implies that if $F(z) \in \mathcal{L}(\mathcal{D}, \mathcal{D}'')$ $(z \in \mathbb{D})$ is a bounded operator-valued analytic function, where $\mathcal{D}''$ is any Hilbert space, such that

$$F(e^{it})^* F(e^{it}) \leq D^2_{W(e^{it})} = I_{\mathcal{D}} - W(e^{it})^* W(e^{it}) \text{ a.e.},$$

then $F(z) \equiv 0$ $(z \in \mathbb{D})$. Indeed, for the bounded analytic function

$$G(z) = \begin{bmatrix} W(z) \\ F(z) \end{bmatrix} \qquad (z \in \mathbb{D}),$$

we have

$$G(e^{i\theta})^* G(e^{i\theta}) \leq I_{\mathcal{D}} \quad \text{a.e.}$$

Therefore,

$$\begin{aligned} G(z)^* G(z) &\leq I_{\mathcal{D}} && (z \in \mathbb{D}), \\ F(z)^* F(z) &\leq D^2_{W(z)} && (z \in \mathbb{D}), \\ \|F(z)J(z)d\|^2 &\leq \|D_{W(z)} J(z)d\|^2 && (z \in \mathbb{D}, d \in \mathcal{D}), \end{aligned} \tag{3.9c}$$

and hence by virtue of (3.6)

$$\lim_{\rho \nearrow 1} \int_0^{2\pi} \|F(\rho e^{i\theta})J(\rho e^{i\theta})d\|^2 d\theta \leq \lim_{\rho \nearrow 1} \int_0^{2\pi} \|D_{W(\rho e^{i\theta})} J(e^{i\theta})d\|^2 d\theta = 0.$$

It follows that the $\mathcal{D}''$-valued function $F(z)J(z)$ (in $H^2(\mathcal{D}'')$) is identically 0. Thus

$$F(z) = F(z) \cdot J(z)K(z) \equiv 0 \qquad (z \in \mathbb{D}).$$

Note that by virtue of ([11, p. 201–203]), the result we just established is equivalent to

$$\overline{D_{W(\cdot)} H^2(\mathcal{D})}^{L^2(\mathcal{D})} = \overline{D_{W(\cdot)} L^2(\mathcal{D})}^{L^2(\mathcal{D})}, \tag{3.10}$$

where both closures are in $L^2(\mathcal{D})$. Thus (3.10) is a necessary condition for $\Gamma(\cdot)$ to be an isometry.

It is obvious that if $W(\cdot)$ is inner (that is, $D_{W(e^{i\theta})} = 0$ a.e.), then (3.10) is satisfied. We will give now a case in which the basic condition (3.6) in Proposition 3.6 can be replaced with the condition

$$D_{W(e^{it})} = 0 \quad \text{(a.e.)}; \tag{3.11}$$

that is, $W(\cdot)$ is an inner (analytic) function. To this end we recall that the analytic operator-valued function $K(z)$, $(z \in \mathbb{D})$, is said to have a scalar multiple if there exist a nonzero function $\delta(\cdot) \in H^\infty$ and a bounded operator-valued analytic function $G(z)$ $(z \in \mathbb{D})$ such that

$$K(z)G(z) = G(z)K(z) = \delta(z) I_{\mathcal{D}} \qquad (z \in \mathbb{D}) \tag{3.12}$$

(cf. [11, Ch. V, Sec. 6]). By adapting the proof of Theorem 6.2 (loc. cit.) to our situation, we can assume due to Lemma 3.12 a) that δ is an outer function; that is,

$$\overline{\delta H^2} = H^2. \tag{3.12a}$$

Consequently, so is $G(\cdot)$; that is,

$$\overline{G(\cdot)H^2(\mathcal{D})} = H^2(\mathcal{D}), \tag{3.12b}$$

and hence (cf. [11, Ch. V, Proposition 2.4 (ii)])

$$G(e^{it})\mathcal{D} = \mathcal{D} \quad \text{(a.e.)} \tag{3.12c}$$

Note that (3.12), (3.12a), (3.12b), and (3.12c) imply that

$$J(e^{it}) = K(e^{it})^{-1} \text{ exists in } \mathcal{L}(\mathcal{D}) \text{ a.e.,} \tag{3.12d}$$

and is in fact equal to $G(e^{it})/\delta(e^{it})$ a.e. Moreover,

$$\|J(\rho e^{it})d - J(e^{it})d\| \to 0 \text{ for } \rho \nearrow 1 \text{ for all } d \in \mathcal{D}, \text{ a.e.} \tag{3.12e}$$

We will now consider the slightly more general case in which (3.12d), (3.12a) hold regardless of whether a scalar multiple exists for $K(\cdot)$.

Lemma 3.15. *Assume that* (3.12d) *and* (3.12e) *hold. Then* (*see the notation in Proposition* 3.11)

$$k_d(e^{i\theta}) = \|D_{W(e^{i\theta})}J(e^{i\theta})d\|^2 \quad \textit{a.e.} \tag{3.13}$$

and

$$h_d(e^{i\theta}) = \|D_{A(e^{i\theta})}J(e^{i\theta})d\|^2 \quad \textit{a.e.} \tag{3.13a}$$

Proof. Since $A(z)$ and $\tilde{A}(z) := A(\bar{z})^*$ are analytic, we have

$$A(\rho e^{i\theta}) \to A(e^{i\theta}), \qquad A(\rho e^{i\theta})^* \to A(e^{i\theta})^* \quad \text{strongly a.e.}$$

and consequently

$$\begin{aligned} D_{A(\rho e^{i\theta})} &= (I - A(\rho e^{i\theta})^* A(\rho e^{i\theta}))^{1/2} \\ &\to (I - A(e^{i\theta})^* A(e^{i\theta}))^{1/2} = D_{A(e^{i\theta})} \quad \text{strongly a.e.} \end{aligned}$$

A similar argument holds for the strong convergence

$$D_{W(\rho e^{i\theta})} \to D_{W(e^{i\theta})} \quad \text{a.e.}$$

Relations (3.13) and (3.13a) are direct consequences of the above strong convergences and of (3.12e). □

We can now give the following corollary to Proposition 3.11.

Proposition 3.16. *Assume that* (3.12d) *and* (3.12e) *hold. Then* $\Gamma(\cdot) \in \mathcal{L}(\mathcal{D}, H^2(\mathcal{D}'))$ *is an isometry if and only if the following property holds:*

(e) $W(\cdot)$ *satisfies the condition* (3.11) *and* $A(\cdot)$ *satisfies the condition*

$$\frac{1}{2\pi}\int_0^{2\pi} \|D_{A(e^{i\theta})}(I - e^{i\theta}A(e^{i\theta}))^{-1}d\|^2 d\theta = \|d\|^2 \qquad (d \in \mathcal{D}). \tag{3.14}$$

Proof. In the present situation (due to Lemma 3.15) we have that $\Gamma(\cdot)$ is isometric if and only if relation (3.14) holds, and

$$D_{W(e^{i\theta})}J(e^{i\theta})d = 0 \quad \text{a.e.} \quad (d \in \mathcal{D}) \tag{3.15}$$

hold. Thus property (e) clearly implies that Γ is an isometry. Now, let $\mathcal{D}_0 \subset \mathcal{D}$ be a countable dense subset of $\mathcal{D}$ and denote by $Ex(d)$ the null set on which (3.15) fails. If Ex_0 denotes the set of the $e^{i\theta}$ is for which at least one of the relations (3.12d), (3.12e) is not valid, then

$$Ex = Ex_0 \cup \left(\bigcup\nolimits_{d\in\mathcal{D}_0} Ex(d)\right)$$

is also a null set and

$$D_{W(e^{i\theta})}J(e^{i\theta})\mathcal{D}_0 = \{0\} \qquad (e^{i\theta} \notin Ex). \tag{3.15a}$$

But for $e^{i\theta} \notin Ex$, the operator $D_{W(e^{i\theta})}J(e^{i\theta})$ is bounded. Therefore (3.15a) implies

$$D_{W(e^{i\theta})} = D_{W(e^{i\theta})}J(e^{i\theta}) \cdot K(e^{i\theta}) = 0 \cdot K(e^{i\theta}) = 0 \quad (e^{i\theta} \notin Ex),$$

i.e., (3.11). This concludes the proof since (3.11) obviously implies (3.15). □

Corollary 3.17. *Let $a, b \in H^\infty$ satisfy the condition*

$$\left\| \begin{bmatrix} a(z) \\ b(z) \end{bmatrix} \right\| \leq 1 \qquad (z \in \mathbb{D})$$

and define

$$w(z) = \begin{bmatrix} a(z) \\ b(z) \end{bmatrix} \quad \text{and} \quad \gamma(z) = \frac{b(z)}{1 - za(z)} \qquad (z \in \mathbb{D}).$$

Then

$$\|\gamma(\cdot)\|^2 = \frac{1}{2\pi}\int_0^{2\pi} |\gamma(e^{i\theta})|^2 d\theta \leq 1 \tag{3.16}$$

and equality holds in (3.16) *if and only if w and a satisfy the following conditions:*

$$w(e^{i\theta})^* w(e^{it}) = 1 \quad \text{a.e.} \quad \text{and} \tag{3.16a}$$

$$\frac{1}{2\pi}\int_0^{2\pi} \frac{1 - |a(e^{i\theta})|^2}{|1 - e^{i\theta}a(e^{i\theta})|^2} d\theta = 1. \tag{3.16b}$$

Proof. The result follows readily from Proposition 3.11, by taking $\mathcal{D} = \mathcal{D}' = \mathbb{C}$. □

Remark 3.18. In Proposition 3.11, neither one of the equalities (3.16) or (3.16b) implies the other, as is shown by the following two examples.

Example. Define

$$u(z) = \frac{3/4}{2\pi}\int_0^{\pi} \frac{e^{i\theta} + z}{e^{i\theta} - z} d\theta + \frac{1/4}{2\pi}\int_{\pi}^{2\pi} \frac{e^{i\theta} + z}{e^{i\theta} - z} d\theta + \frac{1}{2}\frac{1+z}{1-z} \quad (z \in \mathbb{D}).$$

Then

$$v(z) = \operatorname{Re} u(z) \geq 0 \qquad (z \in \mathbb{D})$$

and

$$v(e^{i\theta}) = \lim_{\rho \nearrow 1} v(\rho e^{i\theta}) = \begin{cases} 3/4 & \text{for } 0 < e^{i\theta} < \pi \\ 1/4 & \text{for } \pi < e^{i\theta} < 2\pi \end{cases}.$$

Set

$$a(z) = \frac{1}{z}\frac{u(z) - 1}{u(z) + 1} \quad (z \in \mathbb{D}\backslash\{0\}) \quad \text{and} \quad a(0) = \frac{1}{2}u'(0).$$

Then $a(z) \in H^\infty$ and

$$1 - |a(e^{i\theta})|^2 = \frac{4v(e^{i\theta})}{|u(e^{i\theta}) + 1|^2} = v(e^{i\theta})|1 - e^{i\theta}a(e^{i\theta})|^2 \text{ a.e.} \tag{3.17}$$

In particular, this relation shows that $\|a(\cdot)\|_{H^\infty} \leq 1$; moreover,

$$1 - |a(e^{i\theta})|^2 \geq \frac{1}{4}|1 - e^{i\theta}a(e^{i\theta})|^2 \quad \text{a.e.}$$

Since $1 - za(z) (z \in \mathbb{D})$ is an outer function, there exists an outer function $b \in H^\infty$ such that

$$|b(e^{i\theta})|^2 = 1 - |a(e^{i\theta})|^2 \quad \text{a.e.}$$

Therefore (3.16) is valid,

$$\|w(z)\| \leq 1 \quad (z \in \mathbb{D}), \text{ where } w(z) = \begin{bmatrix} a(z) \\ b(z) \end{bmatrix} \quad (z \in \mathbb{D}),$$

and (3.16a) is satisfied. However, due to (3.17)

$$\frac{1}{2\pi}\int_0^{2\pi} \frac{1 - |a(e^{i\theta})|^2}{|1 - e^{i\theta}a(e^{i\theta})|^2} d\theta = \frac{1}{2\pi}\int_0^{2\pi} v(e^{i\theta})d\theta = \frac{1}{2}, \quad \text{and}$$

hence (3.16b) is not valid.

Example. Define

$$w(z) = \begin{bmatrix} 1/2 \\ 1/2 \end{bmatrix} \quad (z \in \mathbb{D}).$$

Then $\|w(z)\| \leq 1/\sqrt{2} < 1$ $(z \in \overline{\mathbb{D}})$ and (3.16) is not satisfied, but

$$\frac{1}{2\pi}\int_0^{2\pi} \frac{1 - (1/2)^2}{|1 - e^{i\theta}/2|^2} d\theta = 1,$$

i.e., (3.16b) is valid.

4. Isometric intertwining lifting

We return now to the commutant lifting theorem setting presented in Section 3. We recall that if we denote

$$A(z) = \Pi W(z), \quad d(z) = (I - zA(z))^{-1}d \qquad (z \in \mathbb{D}), \tag{4.1}$$

where $d \in \mathcal{D} = \mathcal{D}_X$, then the contractive intertwining liftings of X are given by the formula

$$B = \begin{bmatrix} X \\ \Gamma(\cdot)D_X \end{bmatrix} \in \mathcal{L}(\mathcal{H}, \mathcal{H}' \oplus H^2(\mathcal{D})), \tag{4.1a}$$

where

$$\Gamma(z)d = \Pi' W(z)d(z) \qquad (z \in \mathbb{D}) \tag{4.1b}$$

and

$$W(z) = \bar{\omega}d(z) + R(z)(I - \bar{\omega}^*\bar{\omega})d(z) \quad (z \in \mathbb{D}). \tag{4.1c}$$

In (4.1c), $\bar{\omega}$ is the partial isometry $\in \mathcal{L}(\mathcal{D}, \mathcal{D} \oplus \mathcal{D}_T)$ defined in Section 2 and

$$R(z) \in \mathcal{L}(\ker \bar{\omega}, \ker \bar{\omega}^*) \qquad (z \in \mathbb{D})$$

is an arbitrary analytic operator-valued function such that

$$\|R(z)\| \leq 1 \qquad (z \in \mathbb{D}).$$

We also recall that B is isometric if and only if $\Gamma(\cdot)$ is isometric. According to Proposition 3.6 this can happen if and only if the (d) set of properties holds for our current $W(\cdot)$. By noticing that

$$\begin{aligned}\|D_{W(z)}d(z)\|^2 &= \|d(z)\|^2 - \|\bar\omega d(z)\|^2 - \|R(z)(1-\bar\omega^*\bar\omega)d(z)\|^2 \\ &= \|(1-\bar\omega^*\bar\omega)d(z)\|^2 - \|R(z)(1-\bar\omega^*\bar\omega)d(z)\|^2 \qquad (4.1\text{d}) \\ &= \|D_{R(z)}(1-\bar\omega^*\bar\omega)d(z)\|^2 \qquad (z\in\mathbb{D}),\end{aligned}$$

we have the following result as a direct consequence of Proposition 3.6.

Proposition 4.1. *The contractive intertwining lifting B associated to $R(\cdot)$ is isometric if and only if the following two properties hold for all $d\in\mathcal{D}$:*

$$\lim_{\rho\nearrow 1}\frac{1}{2\pi}\int_0^{2\pi}\|D_{R(\rho e^{i\theta})}(1-\bar\omega^*\bar\omega)d(e^{i\theta})\|^2 d\theta = 0, \tag{4.2}$$

and the Taylor coefficients $d_n (n=0,1,\ldots)$ of $d(\cdot)$ satisfy the condition $\|d_n\|\to 0$ for $n\to\infty$. (4.2a)

Due to (4.1d) (as well as to the equivalence of the facts (a), (b) observed before Lemma 3.10) we can also reformulate Proposition 3.11 as follows

Proposition 4.2. *$\Gamma(\cdot)\in\mathcal{L}(\mathcal{D},H^2(\mathcal{D}'))$ is an isometry if and only if the following two conditions are satisfied:*

$$\int_0^{2\pi} h_d(e^{i\theta})\frac{d\theta}{2\pi} = \|d\|^2 \qquad (d\in\mathcal{D}), \tag{4.3}$$

where (see also (4.1))

$$h_d(e^{i\theta}) = \lim_{\rho\nearrow 1}\|D_{W(z)}d(z)\|\big|_{z=\rho e^{i\theta}} \text{ a.e.}, \tag{4.3a}$$

and

$$\lim_{\rho\nearrow 1}[(1-|z|^2)\|d(z)\|^2 + \|D_{R(z)}(1-\bar\omega^*\bar\omega)d(z)\|\big|_{z=\rho e^{i\theta}} = 0 \text{ a.e.} \tag{4.4}$$

for each $d\in\mathcal{D}$.

The problem with these two propositions is that one cannot always apply them. None of the conditions (4.2), (4.2a), (4.3) or (4.4) is easy to analyse or check. To illustrate this difficulty we will now give two results.

Proposition 4.3. *With the notation of Section 2, assume $\|X\|<1$ and that there is an isometric intertwining lifting Y_1 of X. Then T is a unilateral shift.*

Proof. Let $\mathcal{H}=\mathcal{H}_0\oplus\mathcal{H}_1$ be the Wold decomposition for T; that is, $T\mathcal{H}_0\subset\mathcal{H}_0$, $T\mathcal{H}_1\subset\mathcal{H}_1$, $T|\mathcal{H}_0$ is a unilateral shift and $T|\mathcal{H}_1$ is unitary, also

$$\mathcal{H}_1 = \bigcap_{n=0}^{\infty} T^n\mathcal{H}.$$

Therefore, if Y is any intertwining lifting for X (that is,

$$U'Y = YT, \quad P'Y = X),$$

then we have

$$Y\mathcal{H}_1 \subseteq \bigcap_{n=0}^{\infty} U'^n \mathcal{K}' = \mathcal{R},$$

where

$$\mathcal{K}' = \mathcal{R}^\perp \oplus \mathcal{R}$$

is the Wold decomposition for U'. If $U'_1 = U'|\mathcal{R}$ and Y_1 is an isometric lifting of X, then $Z = Y - Y_1$ satisfies

$$U'_1(Z|\mathcal{H}_1) = Z(T|\mathcal{H}_1),$$
$$U'^*_1(Z|\mathcal{H}_1) = (Z|\mathcal{H}_1)(T|\mathcal{H}_1)^*$$

since U'_1 and $T|\mathcal{H}_1$ are unitary. Therefore, $\overline{Z\mathcal{H}_1}$ is a reducing subspace for U' and is orthogonal to $\mathcal{H}'$ (because $P'Z = P'Y - P'Y_1 = 0$). Due to the minimality of U' (that is,

$$\mathcal{K}' = \bigvee_{n>0} U'^n \mathcal{H}')$$

we have $\overline{Z\mathcal{H}_1} = \{0\}$ and hence

$$Y|\mathcal{H}_1 = Y_1|\mathcal{H}_1. \tag{4.5}$$

But the Commutant Lifting Theorem applied to $X_0 = X/\|X\|$ yields a contractive intertwining lifting Y_0 of X_0. It follows that $Y := \|X\|Y_0$ is an intertwining lifting of X such that $\|Y\| \leq \|X\| < 1$. From the relation (4.5) and the hypothesis that Y_1 is isometric, we conclude that $\mathcal{H}_1 = \{0\}$ and so $T = T_0$ is a unilateral shift. □

This result shows that if X is a strict contraction, there cannot exist an isometric Γ, unless T is a unilateral shift.

Example. Let U denote the canonical bilateral shift on $L^2(T)$; that is,

$$(Uf)(e^{it}) = e^{it} f(e^{it}) \quad \text{a.e.} \quad (f \in L^2(T))$$

and let $S = U|H^2$ be the canonical unilateral shift on H^2. Let $V = U|L^2[(0,\pi))$ and Q be the orthogonal projection of $L^2(T) = L^2([0,2\pi))$ onto $L^2([0,\pi))$. Then the following properties are immediate:

$$VQ = QS, \quad \ker Q = \{0\}, \quad \text{and} \quad \ker Q^* = \{0\}. \tag{4.6}$$

Now set

$$T = V^*, \quad T' = S^*, \quad \text{and} \quad X = Q^*/2. \tag{4.6a}$$

Then T and $U' = U^*$ are unitary. Hence, there exists a unique intertwining lifting Y of X and its norm is equal to $\|X\| = 1/2 < 1$.

Thus even when the operators T and T' are very elementary, (in this case, T is a unitary operator of multiplicity one and T' is the backward shift of multiplicity one), there may not exist any free Schur contraction that makes Γ isometric. Again we don't see how to deduce this fact easily from Propositions 4.1 and 4.2.

In the study of the parametrization of all contractive intertwining liftings of a given intertwining contraction X, the case when $\|X\| < 1$ is the most amenable to study. Proposition 4.3 shows that in this case our present study reduces to the case when T is a unilateral shift. Related to this case we have the following .

Lemma 4.4. *Assume T is a unilateral shift, T' is a $C_{\bullet 0}$-contraction (that is, $T^{*n} \to 0$ strongly) with dense range, and $\|X\| < 1$. If an isometric intertwining lifting Y of X exists, then we have*

$$\dim \ker \bar{\omega} \leq \dim \ker \bar{\omega}^*. \tag{4.7}$$

Proof. In this case the space $\mathcal{R}$ introduced in the proof of Proposition 4.3 is $\{0\}$, or equivalently, U' is also a unilateral shift. Consider the minimal unitary extensions $\widehat{U}' \in \mathcal{L}(\widehat{\mathcal{K}}')$ and $\widehat{T} \in \mathcal{L}(\widehat{\mathcal{H}})$ of U' and T, respectively. Let $\widehat{Y} \in \mathcal{L}(\widehat{\mathcal{H}}, \widehat{\mathcal{K}}')$ be the unique extension of Y satisfying

$$\widehat{U}'\widehat{Y} = \widehat{Y}\widehat{T}.$$

It is easy to see that $\widehat{Y}$ is isometric and thus the multiplicities ν and μ of the bilateral shifts $\widehat{U}'$ and $\widehat{T}$, respectively, satisfy

$$\mu \leq \nu. \tag{4.7a}$$

The inequality (4.7) follows directly from the equalities

$$\dim \ker \bar{\omega} = \mu \tag{4.7b}$$

$$\dim \ker \bar{\omega}^* = \nu. \tag{4.7c}$$

To prove (4.7b) and (4.7c) we notice, using the fact that D_X is invertible, that

$$\ker \bar{\omega} = D_X^{-1} \ker T^* \quad \text{and}$$

$$\ker \bar{\omega}^* = \{D_X^{-1} X^* D_{T'} d' \oplus (d') \colon \ d' \in \mathcal{D}_{T'}\}.$$

Thus

$$\dim \ker \bar{\omega} = \dim ker T^* = \mu$$

and

$$\dim \ker \bar{\omega}^* = \dim \mathcal{D}_{T'} = \dim \mathcal{D}_{T'^*} = \nu,$$

where the second equality follows from the fact that $\ker T'^* = \{0\}$. □

The preceding lemma shows that the case when the inequality (4.7) holds is of some interest. Therefore, *throughout the remaining part of this section we will assume that* (4.7) *is valid*. Under this assumption, we will study only the case when the free Schur contraction $R(z)$ is independent of z; that is, when $W(z) = W(0)$ $(z \in \mathbb{D})$. According to Corollary 3.8, in this case $\Gamma(\cdot)$ is an isometry if and only if $A_0 = \Pi W(0)$ is a $C_{0\bullet}$-contraction and $W(0)$ is an isometry. This last restriction is obviously equivalent to the free Schur contraction $R(z) \equiv R(0)$ $(z \in \mathbb{D})$ being an

isometry. Therefore (see Corollary 3.8), if there exists such a free Schur contraction for which the corresponding $\Gamma(\cdot)$ is not an isometry, the operator

$$V = A_0^* = W(0)^* \pi^* \in \mathcal{L}(\mathcal{D}) \tag{4.8}$$

would not be a $C_{\bullet 0}$-contraction. Let $\widehat{V} \in \mathcal{L}(\widehat{\mathcal{D}})$ denote the minimal isometric lifting of V and let

$$\widehat{\mathcal{D}} = \mathcal{R}^\perp \oplus \mathcal{R}$$

be the Wold decomposition for $\widehat{V}$, where $\widehat{V}|\mathcal{R}$ is the unitary part of $\widehat{V}$. Since V is not a $C_{\bullet 0}$-contraction, there exists an $r_0 \in \mathcal{R}$ satisfying $d_0 = Pr_0 \neq 0$, where P denotes the orthogonal projection of $\widehat{\mathcal{D}}$ onto $\mathcal{D}$. Let

$$d_n = P\widehat{V}^{*n} r_0 \qquad (n = 0, 1, \ldots). \tag{4.8a}$$

Then

$$V d_{n+1} = P\widehat{V}\widehat{V}^{*n+1} r_0 = P\widehat{V}^{*n} r_0 = d_n, \tag{4.8b}$$

$$0 \leq \|d_0\| \leq \|d_1\| \leq \cdots \leq \|d_n\| \leq \cdots, \tag{4.8c}$$

and

$$\|d_n\| \leq \|r_0\| \qquad (n = 0, 1, 2, \ldots). \tag{4.8d}$$

At this moment it is worth noticing that we have actually proven part of the following characterization of a contraction which is not a $C_{\bullet 0}$-contraction, a fact which may be useful elsewhere.

Lemma 4.5. *Let $T \in \mathcal{L}(\mathcal{H})$ be a contraction. Then T is not a $C_{\bullet 0}$-contraction if and only if there exists a bounded sequence $\{h_n\}_{n=0}^{\infty} \subset \mathcal{H}$ such that*

$$h_0 \neq 0, \quad h_n = T h_{n+1} \qquad (n = 0, 1, \ldots). \tag{4.9}$$

Proof. It remains to prove that if such a sequence exists then T is not a $C_{\bullet 0}$-contraction. To this end note that

$$\|h_0\| \leq \|h_1\| \leq \cdots \leq \|h_n\| \leq \|h_{n+1}\| \leq \cdots \leq M < \infty,$$

where M is the supremum in (4.9). Choose n_0 large enough for $\|h_{n_0}\| \geq \sqrt{63}\, M/8$ to hold. Then for any $N = 0, 1, \ldots$, we have

$$\begin{aligned} \|(I - T^{*N} T^N) h_{n_0+N}\|^4 &\leq \|(I - T^{*N} T^N)^{1/2} h_{n_0+N}\|^2 \|h_{n_0+N}\|^2 \\ &\leq ((I - T^{*N} T^N) h_{n_0+N}, h_{n_0+N}) M^2 \\ &= (\|h_{n_0+N}\|^2 - \|h_{n_0}\|^2) M^2 \\ &\leq M^4/64. \end{aligned}$$

Hence

$$\begin{aligned} \|T^N h_{n_0}\| &\geq \|h_{n_0+N}\| - \|(I - T^{*N} T^N) h_{n_0+N}\| \\ &\geq \sqrt{63}\, M/8 - M/2\sqrt{2} > 0 \end{aligned}$$

for all $N = 0, 1, \ldots$. This proves that T is not a $C_{\bullet 0}$-contraction. □

We return now to our particular considerations. The relation (4.8b) can be written as

$$W(0)^*\Pi^* d_{n+1} = d_n \qquad (n = 0, 1, \ldots). \tag{4.10}$$

Applying $\bar{\omega}$ to these last equalities we obtain

$$\bar{\omega}\bar{\omega}^*\Pi^* d_{n+1} = \bar{\omega} d_n \qquad (n = 0, 1, \ldots). \tag{4.10a}$$

Note that (4.10) also implies

$$\|d_0\| \leq \|d_1\| \leq \|d_2\| \leq \ldots \ . \tag{4.10b}$$

Thus we obtain the following.

Lemma 4.6. *Let ω have (besides (4.7)) the following property:*

(a) *Any sequence $\{d_n\}_{n=0}^{\infty} \subset \mathcal{D}$ for which (4.10a) and (4.10b) are valid is either identically zero or unbounded.*

Then for any isometric free Schur contraction $R(z) \equiv R(0)$ $(z \in \mathbb{D})$, the corresponding operator $\Gamma(\cdot)$ is also an isometry.

This lemma does not preclude the possibility that its conclusion may hold under a weaker hypothesis than condition (a).

Indeed, let us assume that we have a sequence $\{d_n\}_{n=0}^{\infty} \subset \mathcal{D}$ satisfying the condition (4.10). We extend recursively the definition of the d_n's as follows:

$$d_{n-1} = W(0)^*\Pi^* d_n \tag{4.11}$$

for $n = 0, n = -1, \ldots$. Let $\mathcal{D}_0$ be the linear space spanned by $\{d_n\}_{n=-\infty}^{\infty}$. Then the linear map C defined from $(I - \bar{\omega}\bar{\omega}^*)\Pi^*\mathcal{D}_0$ into $(I - \bar{\omega}^*\bar{\omega})\mathcal{D}_0$ by

$$C(I - \bar{\omega}\bar{\omega}^*)\Pi^* d_{n+1} = (I - \bar{\omega}^*\bar{\omega}) d_n \qquad (n \in \mathbb{Z}) \tag{4.11a}$$

extends by continuity to $\overline{C} = R(0)^*|((I - \bar{\omega}\bar{\omega}^*)\Pi^*\mathcal{D}_0)^-$. Clearly, $\overline{C}$ *is a contraction and its definition depends only on ω and the sequence $\{d_n\}_{n=0}^{\infty}$ satisfying* (4.10). Moreover, by its construction $\overline{C}$ extends to a co-isometry (namely $R(0)^*$) from $\ker \bar{\omega}^*$ onto $\ker \bar{\omega}$.

To continue our analysis we now need the following.

Lemma 4.7. *Let $\mathcal{H}$ and $\mathcal{H}'$ be two Hilbert spaces with subspaces $\mathcal{M} \subset \mathcal{H}$ and $\mathcal{M}' \subset \mathcal{H}'$. Let $C \in \mathcal{L}(\mathcal{M}', \mathcal{M})$ be a contraction with dense range in $\mathcal{M}$. Then C has a coisometric extension $\widehat{C} \in \mathcal{L}(\mathcal{H}', \mathcal{H})$ if and only if*

$$\dim(\mathcal{H}' \ominus \mathcal{M}') \geq \dim((\mathcal{H} \ominus \mathcal{M}) \oplus \mathcal{D}_{C^*}). \tag{4.12}$$

Proof. If a coisometric extension $\widehat{C}$ of C exists, then for $h \in \mathcal{H} \ominus \mathcal{M}$ we have

$$(\widehat{C}^* h, m') = (h, Cm') = 0 \qquad (m' \in \mathcal{M}')$$

and so

$$\widehat{C}^*(\mathcal{H} \ominus \mathcal{M}) \subset \mathcal{H} \ominus \mathcal{M}'. \tag{4.12a}$$

Clearly, we also have

$$\widehat{C}^*\mathcal{M} \perp \widehat{C}^*(\mathcal{H} \ominus \mathcal{M}), \tag{4.12b}$$

and

$$P'_{\mathcal{M}'}\widehat{C}^*|\mathcal{M} = C^*, \tag{4.12c}$$

where $P'_{\mathcal{M}}$ is the orthogonal projection of $\mathcal{H}'$ onto $\mathcal{M}'$. Thus

$$\widehat{C}^* m = C^* m + X D_{C^*} m \qquad (m \in \mathcal{M}), \tag{4.12d}$$

where $X \in \mathcal{L}(\mathcal{D}_{C^*}, \mathcal{H}' \ominus \mathcal{M}')$ is an isometry. Due to (4.12b) we have

$$X\mathcal{D}_{C^*} \perp \widehat{C}^*(\mathcal{H} \ominus \mathcal{M})$$

and therefore (4.12) holds. Conversely, if (4.12) holds we can define an isometric operator C_1 from $\mathcal{H}$ into $\mathcal{H}'$ in the following way. First, due to (4.12) we can find two mutually orthogonal subspaces $\mathcal{X}$ and $\mathcal{Y}$ of $\mathcal{H}' \ominus \mathcal{M}'$ such that

$$\dim \mathcal{X} = \dim \mathcal{D}_{C^*}, \quad \dim \mathcal{Y} = \dim \mathcal{H} \ominus \mathcal{M}.$$

Choose for $C_1|\mathcal{H} \ominus \mathcal{M}$ any unitary operator $\in \mathcal{L}(\mathcal{H} \ominus \mathcal{M}, \mathcal{Y})$ and define $C_1|\mathcal{M}$ by

$$C_1 m = C^* m + X D_{C^*} m \qquad (m \in \mathcal{M}), \tag{4.12e}$$

where X is any unitary operator in $\mathcal{L}(\mathcal{D}_{C^*}, \mathcal{X})$. The operator thus defined on $\mathcal{H}$ is isometric and

$$P'_{\mathcal{M}'} C_1|\mathcal{M} = C^*.$$

Consequently, C_1^* is coisometric and

$$\begin{aligned}(C_1^* m', h) &= (m', C_1 h) = (m', C_1 P_{\mathcal{M}} h) = \\ &= (m', C^* P_{\mathcal{M}} h) = (Cm', P_{\mathcal{M}} h) = (Cm', h)\end{aligned}$$

for all $m' \in \mathcal{M}'$, $h \in \mathcal{H}$, and hence $C_1^*|\mathcal{M}' = C$, where $P_{\mathcal{M}}$ is the orthogonal projection of $\mathcal{H}$ onto $\mathcal{M}$. □

Returning to the discussion preceding the above lemma, we deduce that the contraction $\overline{C}$ must satisfy the condition

$$\dim(\ker \bar{\omega}^*) \geq \dim(\ker \bar{\omega} \oplus \mathcal{D}_{C^*}). \tag{4.13}$$

Thus we have proved the following. □

Proposition 4.8. *Assume that there exists a not identically zero sequence* $\{d_n\}_{n=0}^{\infty} \subset \mathcal{D}$ *satisfying* (4.10a) *and* (4.10b). *In order that this sequence also satisfies* (4.10) *for an appropriate free Schur contraction* $R(z) = R(0)$ $(z \in \mathbb{D})$ *where* $R(0)$ *is isometric, the following set of properties is necessary and sufficient:*

(a) *The sequence* $\{d_n\}_{n=0}^{\infty}$ *can be extended to a bilateral sequence* $\{d_n\}_{n=-\infty}^{\infty}$ *satisfying*

$$\bar{\omega}\bar{\omega}^* \Pi^* d_{n+1} = \bar{\omega} d_n \qquad (n \in \mathbb{Z}); \tag{4.14}$$

(b) *the definition* (4.11a) *yields, by linearity and continuity, a contraction in* $\mathcal{L}(((I - \bar{\omega}\bar{\omega}^*)\Pi\mathcal{D}_0)^-, ((I - \bar{\omega}^*\bar{\omega})\mathcal{D}_0)^-)$, *where* $\mathcal{D}_0$ *is the linear span of* $\{d_n\}_{n=-\infty}^{\infty}$;

(c) *the inequality*

$$\dim(\ker \bar{\omega}^*) \geq \dim(\ker \bar{\omega} \oplus \mathcal{D}_{C^*}) \tag{4.14a}$$

holds.

Note that (4.14a) is a more stringent condition than (4.7).

Finally, the proof of Lemma 4.6 allows us to infer the following complement to Proposition 4.8 and Lemma 4.6.

Proposition 4.9. *Let* $\{d_n\}$ *be a sequence satisfying* (4.10a), (4.10b), *all the properties* (a), (b), *and* (c) *in Proposition* 4.8 *and*

$$\sup_{n\geq 0}\|d_n\| < \infty. \tag{4.14b}$$

Then no operator $\Gamma(\cdot)$ *corresponding to a free Schur contraction provided by Proposition* 4.8 *is isometric.*

We conclude this note with a closer look at the case

$$\|X\| < 1, \tag{4.15}$$

in which the partial isometries $\bar{\omega}$ and $\bar{\omega}^*$ can be given in an explicit form. Indeed, in this case D_X and $T^*D_X^2T$ are invertible operators in $\mathcal{H}$ and $\mathcal{D} = \mathcal{D}_X = \mathcal{H}$,

$$\bar{\omega}^*\bar{\omega} = D_XT(T^*D_X^2T)^{-1}T^*D_X, \tag{4.15a}$$

$$\bar{\omega} = \begin{bmatrix} D_X \\ D_{T'}X \end{bmatrix}(T^*D_X^2T)^{-1}T^*D_X, \quad \text{and} \tag{4.15b}$$

$$\bar{\omega}^* = D_XT(T^*D_X^2T)^{-1}[D_X\ X^*D_{T'}]. \tag{4.15c}$$

With this preparation we can now prove the following result.

Proposition 4.10. *Assume* T *is a unilateral shift* (*of any multiplicity*) *and that the relations* (4.7) *and* (4.15) *are satisfied. Let* $R_0 \in \mathcal{L}(\ker\bar{\omega}, \ker\bar{\omega}^*)$ *be any isometry. Define the free Schur contraction by* $R(z) = R_0$ $(z \in \mathbb{D})$ *and let* Y *be the corresponding intertwining lifting of* X. *Then* Y *is an isometry.*

Proof. It will be sufficient to prove that in the present case Property (a) in Lemma 4.6 is satisfied. So, let $\{d_n\}_{n=0}^{\infty}$ be a sequence in $\mathcal{H}$ satisfying the relations (4.10a) and (4.10b). Applying $\bar{\omega}^*$ on both sides of identity (4.10a) we obtain

$$\bar{\omega}^*\Pi^*d_{n+1} = \bar{\omega}^*\bar{\omega}d_n \qquad (n = 0, 1, 2, \ldots,). \tag{4.16}$$

Introducing in (4.16) the explicit forms (4.15a) of $\bar{\omega}^*\bar{\omega}$; and (4.15c) of $\bar{\omega}^*$, respectively, we obtain

$$D_XT(T^*D_X^2T)^{-1}D_Xd_{n+1} = D_XT(T^*D_X^2T)^{-1}T^*D_Xd_n$$

$(n = 0, 1, \ldots)$, whence

$$d_{n+1} = D_X^{-1}T^*D_Xd_n \qquad (n = 0, 1, 2, \ldots).$$

We infer

$$d_n = D_X^{-1}T^{*n}D_Xd_0 \qquad (n = 0, 1, 2, \ldots),$$

where $T^{*n} \to 0$ strongly. This together with (4.10b) forces $d_n = 0$ for all $n \geq 0$. □

Remark 4.11. Under the assumptions of Proposition 4.10, the inequality (4.7) obtains an explicit form. Indeed, since

$$\ker \bar{\omega}^* \bar{\omega} = D_X^{-1} \ker T^*$$

and

$$\ker \bar{\omega} \bar{\omega}^* = \left\{ \begin{bmatrix} D_X^{-1} X^* D_T d' \\ d' \end{bmatrix} : \ d' \in \mathcal{D}_{T'} (= \mathcal{D}') \right\},$$

we have

$$\dim \ker \bar{\omega}^* \bar{\omega} = \dim \ker T^* = \dim \mathcal{D}_{T^*}, \tag{4.17}$$

and

$$\dim \ker \bar{\omega} \bar{\omega}^* = \dim \mathcal{D}_{T'}. \tag{4.17a}$$

Consequently, introducing (4.17) and (4.17a) in (4.7), the last relation takes the form

$$\dim \mathcal{D}_{T'} \geq \dim \mathcal{D}_{T^*}. \tag{4.17b}$$

In fact, the previous proof can be modified to yield the following slight improvement of Proposition 4.10.

Proposition 4.12. *Assume that T is a unilateral shift and* (4.7) *is satisfied. Assume also that*

$$D_X \text{ and } D_X T \text{ both have closed range.} \tag{4.18}$$

Then the conclusion of Proposition 4.10 *is valid.*

Proof. We will use the proof of Proposition 4.10 replacing the inverse for $T^* D_X^2 T$ by a left inverse which the fact that $D_X T$ and D_X have closed range will allow us to define. The key here is to show that the range of $T^* D_X$ is contained in the support of $T^* D_X^2 T$. □

Remark 4.13. One can verify using the definition that $\ker \bar{\omega} = D_X \mathcal{H} \cap D_{T^*} \mathcal{H}$ and $\ker \bar{\omega}^* = D_{T'} \mathcal{H}' \cap D_{X^*} \mathcal{H}'$. Therefore, in the context of Proposition 4.12, (4.7) becomes

$$\dim (D_X \mathcal{H} \cap D_{T^*} \mathcal{H}) \leq \dim (D_{T'} \mathcal{H}' \cap D_{X^*} \mathcal{H}'). \tag{4.18a}$$

Obviously, (4.18a) reduces to (4.17b) if $\|X\| < 1$ and hence D_X is invertible.

Remark 4.14. We begin by noting that (4.18) in Proposition 4.12 is equivalent to the assumption that D_X and D_{XT} have closed range. Proposition 4.12 has a direct consequence concerning an apparently more general setting of the Commutant Lifting Theorem. Indeed, let $T_0 \in \mathcal{L}(\mathcal{H}_0)$ be a contraction, $X_0 \in \mathcal{L}(\mathcal{H}_0, \mathcal{H}')$ satisfying

$$X_0 T_0 = T' X_0, \text{ and both } D_{X_0} \text{ and } D_{X_0 T_0} \text{ have closed range.} \tag{4.19}$$

If $T \in \mathcal{L}(\mathcal{H})$ is the minimal isometric lifting of T_0, then

$$X = X_0 P_0, \tag{4.19a}$$

where P_0 is the orthogonal projection of $\mathcal{H}$ onto $\mathcal{H}_0$, will satisfy

$$XT = T'X \text{ and both } D_X \text{ and } D_{XT} \text{ will have closed range.} \tag{4.19b}$$

Let Y be any contractive intertwining lifting of X. We have

$$P'Y = X_0P_0 \quad \text{and} \quad YT = T'Y. \tag{4.19c}$$

Moreover, any contraction $Y \in \mathcal{L}(\mathcal{H}, \mathcal{K}')$ satisfying (4.19b) (actually referred to as a contractive intertwining lifting of X_0) will be a contractive intertwining lifting of X. Now assume that T_0 is a $C_{\bullet 0}$-contraction. This implies that T is a shift such that

$$\dim \mathcal{D}_{T^*} = \dim \mathcal{D}_{T_0^*}.$$

(see [11, Ch. II]). Moreover

$$D_X\mathcal{H} \cap D_{T^*}\mathcal{H} = D_{X_0}\mathcal{H}_0 \cap D_{T_0^*}\mathcal{H}, \quad D_{X^*}\mathcal{H}' = D_{X_0^*}\mathcal{H}';$$

thus, if

$$\dim(D_{X_0}\mathcal{H}_0 \cap D_{T_0}\mathcal{H}_0) \le \dim(D_{T'}\mathcal{H}' \cap D_{X_0^*}\mathcal{H}'), \tag{4.20}$$

we can apply Proposition 4.12 to the present setting and conclude *that the set of the isometric intertwining liftings of X_0 is not empty and, moreover, that for every free Schur contraction of the form $R(z) = R_0$ $(z \in \mathbb{D})$ with R_0 an isometry, the corresponding contractive intertwining lifting Y of X_0 is also an isometry*; in connection with this result see [5, 7].

References

[1] H. Bercovici, R.G. Douglas and C. Foias, *On the Classification of Multi-isometries.* Acta Sci. Math. (Szeged) **72** (2006), no. 3-4, 639–661.

[2] H. Bercovici, R.G. Douglas and C. Foias, *On the Classification of Multi-isometries II: Functional Models* (in progress).

[3] P.L. Duren, *Theory of H^p Spaces.* Academic Press, New York-London, 1970.

[4] C. Foias, A.E. Frazho, I. Gohberg, and M.A. Kaashoek, *Metric Constrained Interpolation, Commutant Lifting and Systems.* Operator Theory: Advances and Applications, 100, Birkhäuser Verlag, Basel, 1998, 587 pp.

[5] C. Foias, A.E. Frazho and A. Tannenbaum, *On Certain Minimal Entropy Extensions Appearing in Dilation Theory I.* Lin. Alg. and its Applic. **137/138** (1990), 213–238.

[6] K. Hoffman, *Banach Spaces of Analytic Functions.* Prentice-Hall Inc., 1962; Dover Publ., Inc., 1988, 215 pp.

[7] W.-S. Li and D. Timotin, *On Isometric Intertwining Liftings.* Non-selfadjoint Operator Algebras, Operator Theory and Related Topics, 155–167. Oper. Theory Adv. Appl., Vol. 104, Birkhäuser, 1998.

[8] M.S. Livsic, *Isometric Operators with Equal Deficiency Indices, Quasi-unitary Operators.* Transl. A.M.S. Series II **13** (1960), 85–103.

[9] D.E. Sarason, *Generalized Interpolation in H^∞.* Trans. Amer. Math. Soc. **127** (1967), 179–203.

[10] B. Sz.-Nagy and C. Foias, *Dilatation des Commutants d'Opérateurs.* C.R. Acad. Sci. Paris, série A **266** (1986), 493–495.

[11] B. Sz.-Nagy and C. Foias, *Harmonic Analysis of Operators on Hilbert Spaces.* Akadémiai Kiadó, & North Holland Publ. Co., 1970, 389 pp.

Hari Bercovici
Department of Mathematics
Indiana University
Bloomington, Indiana 47405, USA
e-mail: bercovic@indiana.edu

Ronald G. Douglas and Ciprian Foias
Department of Mathematics
Texas A&M University
College Station, Texas 77843, USA
e-mail: rdouglas@math.tamu.edu
foias@math.tamu.edu

Operator Theory:
Advances and Applications, Vol. 197, 77–98

The One-sided Ergodic Hilbert Transform of Normal Contractions

Guy Cohen and Michael Lin

Dedicated to the memory of Moshe Livšic

Abstract. Let T be a normal contraction on a Hilbert space H. For $f \in H$ we study the one-sided ergodic Hilbert transform $\lim_{n\to\infty} \sum_{k=1}^{n} \frac{T^k f}{k}$. We prove that weak and strong convergence are equivalent, and show that the convergence is equivalent to convergence of the series $\sum_{n=1}^{\infty} \frac{\log n \| \sum_{k=1}^{n} T^k f\|^2}{n^3}$. When $H = \overline{(I-T)H}$, the transform is shown to be precisely minus the infinitesimal generator of the strongly continuous semi-group $\{(I-T)^r\}_{r\geq 0}$.

The equivalence of weak and strong convergence of the transform is proved also for T an isometry or the dual of an isometry.

For a general contraction T, we obtain that convergence of the series $\sum_{n=1}^{\infty} \frac{\langle T^n f, f\rangle \log n}{n}$ implies strong convergence of $\sum_{n=1}^{\infty} \frac{T^n f}{n}$.

Mathematics Subject Classification (2000). Primary: 47A35, 47B15; Secondary: 37A30, 42A16.

Keywords. Normal contractions, one-sided ergodic Hilbert transform.

1. Introduction

Let θ be a measure preserving invertible transformation of a probability space $(\mathcal{S}, \Sigma, m)$, and let U be the unitary operator induced on $L_2(m)$. For θ ergodic, Izumi [I] raised the question of almost everywhere (a.e.) convergence of $\sum_{k=1}^{\infty} \frac{U^k f}{k}$ for all functions $f \in L_2(m)$ with zero integral. Halmos [H] proved that when the probability space $(\mathcal{S}, \Sigma, m)$ is non-atomic, there is always a function $f \in L_2(m)$ with zero integral for which the above series fails to converge in L_2-norm. For additional background and references see [AL].

For T power-bounded on a Banach space X we have $\|\frac{1}{n}\sum_{k=1}^{n} T^k f\| \to 0$ if and only if $f \in \overline{(I-T)X}$, and it is known that weak and strong convergence of the averages are equivalent (e.g., [Kr, §2.1]). Hence, by Kronecker's lemma, *weak* convergence of the *one-sided ergodic Hilbert transform* $\sum_{k=1}^{\infty} \frac{T^k f}{k}$ strengthens the strong convergence to zero of the averages.

Theorem 1.1. *Let T be a power-bounded operator on a Banach space X, put $Y := \overline{(I-T)X}$, and denote by S the restriction of T to Y. Then the following are equivalent:*

(i) *$(I-T)X$ is closed in X.*
(ii) *The series $\sum_{k=1}^{\infty} \frac{S^k}{k}$ converges in operator norm.*
(iii) *The series $\sum_{k=1}^{\infty} \frac{T^k f}{k}$ converges in norm for every $f \in Y$.*
(iv) *The series $\sum_{k=1}^{\infty} \frac{T^k f}{k}$ converges weakly for every $f \in Y$.*

Proof. (i)$\Longrightarrow$(ii): It is easy to compute that we always have operator-norm convergence of $\sum_{k=1}^{\infty} \frac{T^k}{k}(I-T)$. By [L], condition (i) implies that $(I-S)$ is invertible on Y, so (ii) holds.

Clearly (ii)$\Longrightarrow$(iii)$\Longrightarrow$(iv).

By [AL, Proposition 4.1], (iv) implies that $Gf := -\sum_{k=1}^{\infty} \frac{T^k f}{k}$ (weak convergence) is a bounded operator on Y which is the infinitesimal generator of a semi-group. Now the proof of [DL, Theorem 2.23] yields (i). □

Remarks

1. Condition (i) implies that $\frac{1}{n}\sum_{k=1}^{n} T^k$ converges in operator norm, even for non-reflexive spaces [L].
2. The equivalence of the first three conditions is implicit in [DL], since (iii)$\Longrightarrow$(i) by [DL, Theorem 2.23].
3. The result of [H] follows from the theorem, since condition (i) is not satisfied by unitary operators induced by aperiodic probability preserving transformations (for which the spectrum is the whole unit circle).

Since for a contraction T in H the fixed points of T and T^* are the same [RN, §144], we have $\overline{(I-T)H} = \overline{(I-T^*)H}$, so $\frac{1}{n}\sum_{k=1}^{n} T^k f \to 0$ if and only if $\frac{1}{n}\sum_{k=1}^{n} T^{*k} f \to 0$.

Proposition 1.2. *Let T be a contraction in a Hilbert space H. Then $\sum_{k=1}^{\infty} \frac{T^k f}{k}$ converges (weakly) if and only if $\sum_{k=1}^{\infty} \frac{T^{*k} f}{k}$ converges (weakly).*

Proof. Using the unitary dilation of T, Campbell [Ca] proved that for every $f \in H$ the series $\sum_{k=1}^{\infty} \frac{T^k f - T^{*k} f}{k}$ converges in norm. □

For a power-bounded operator T on a Banach space X, Derriennic and Lin [DL] defined the operator $(I-T)^\alpha$ for $0 < \alpha < 1$ by the series $I - \sum_{k=1}^{\infty} a_k^{(\alpha)} T^k$, where $a_k^{(\alpha)} > 0$ with $\sum_{k=1}^{\infty} a_k^{(\alpha)} = 1$ are the coefficients of the power-series $(1-t)^\alpha =$

$1-\sum_{k=1}^{\infty} a_k^{(\alpha)} t^k$ for $|t| \le 1$. They proved that $(I-T)X \subset (I-T)^\alpha X \subset \overline{(I-T)X}$, and when $(I-T)X$ is not closed both inclusions are strict. For T mean ergodic (e.g., X is reflexive) we have $f \in (I-T)^\alpha X$ if and only if $\displaystyle\sum_{k=1}^{\infty} \frac{T^k f}{k^{1-\alpha}}$ converges strongly ([DL, Theorem 2.11]), and then $\sum_{k=1}^{\infty} \frac{T^k f}{k}$ converges strongly.

When $\overline{(I-T)X} = X$ we have that $\{(I-T)^r : r \ge 0\}$ is a strongly continuous one-parameter semi-group [DL, Theorem 2.22], and the domain of its infinitesimal generator G contains all f for which $\sum_{k=1}^{\infty} \frac{T^k f}{k}$ converges weakly, and then the sum of the series is $-Gf$ [AL, Proposition 4.1].

For fixed $t \in [-1,1)$ the infinitesimal generator of $(1-t)^r = e^{r\log(1-t)}$ is obviously $\log(1-t) = -\sum_{k=1}^{\infty} \frac{t^k}{k}$, so two natural questions arise (when $(I-T)X$ is not closed, but dense):

(i) If f is in the domain of G, does the series $\sum_{k=1}^{\infty} \frac{T^k f}{k}$ converge weakly?

(ii) If $\sum_{k=1}^{\infty} \frac{T^k f}{k}$ converges weakly, does it converge strongly?

We answer both questions positively for normal contractions in a (complex) Hilbert space; for T unitary or self-adjoint this was proved in [AL].

2. Preliminaries

Lemma 2.1. *Let $\{f_k\}$ be a sequence in a Banach space. If the series $\displaystyle\sum_{k=1}^{\infty} \frac{f_k}{k}$ converges, then for every $\alpha > 0$ the series $\displaystyle\sum_{k=1}^{\infty} \frac{f_k}{k^{1+\alpha}}$ converges and we have*

$$\lim_{\alpha \to 0^+} \sum_{k=1}^{\infty} \frac{f_k}{k^{1+\alpha}} = \sum_{k=1}^{\infty} \frac{f_k}{k}.$$

Proof. We put $S_n = \sum_{k=1}^{n} \frac{f_k}{k}$. By Abel's summation by parts we have

$$\sum_{k=1}^{n} \frac{f_k}{k^{\alpha+1}} = \frac{S_n}{n^\alpha} + \sum_{k=1}^{n-1} S_k \Big[\frac{1}{k^\alpha} - \frac{1}{(k+1)^\alpha}\Big].$$

Since S_n converges we have $\sup_n \|S_n\| < \infty$, so the first term on the right-hand side above tends to zero as n tends to infinity. The second term is absolutely summable as the factor of S_k there behaves like $1/k^{1+\alpha}$. Hence we obtain the first assertion. In particular we obtain

$$\sum_{k=1}^{\infty} \frac{f_k}{k^{\alpha+1}} = \sum_{k=1}^{\infty} S_k \Big[\frac{1}{k^\alpha} - \frac{1}{(k+1)^\alpha}\Big]. \tag{1}$$

It remains to prove the second assertion. We are going to define a Toeplitz summability matrix.

Let $\alpha_j \to 0^+$ be an arbitrary sequence and define a summability matrix (with positive entries!) $a_{j,k} = \frac{1}{k^{\alpha_j}} - \frac{1}{(k+1)^{\alpha_j}}$. Clearly, (i): $\lim_{j\to\infty} a_{j,k} = 0$ for every $k \geq 1$ and (ii): $\sum_{k=1}^{\infty} a_{j,k} = 1$ for every j. Put $S = \lim_{n\to\infty} S_n$.

Using (1) above we have

$$\Big\| \sum_{k=1}^{\infty} \frac{f_k}{k^{1+\alpha_j}} - S \Big\| = \Big\| \sum_{k=1}^{\infty} a_{j,k}(S_k - S) \Big\| \leq \sum_{k=1}^{\infty} a_{j,k} \|S_k - S\|.$$

Since $\|S_k - S\| \underset{k\to\infty}{\longrightarrow} 0$, properties (i) and (ii) yield

$$\lim_{j\to\infty} \Big\| \sum_{k=1}^{\infty} \frac{f_k}{k^{1+\alpha_j}} - S \Big\| = 0.$$

Since $\{\alpha_j\}$ is arbitrary, the assertion follows. □

Corollary 2.2. *Let (X, μ) be a measure space and let T be an operator on $L_p(X)$ $(p \geq 1)$. If for some $f \in X$, the series $\sum_{k=1}^{\infty} \frac{T^k f}{k}$ converges a.e., then*

$$\lim_{\alpha\to 0^+} \sum_{k=1}^{\infty} \frac{T^k f}{k^{1+\alpha}} = \sum_{k=1}^{\infty} \frac{T^k f}{k} \quad a.e.$$

Proof. For a.e. $x \in X$ we put $f_k = [T^k f](x)$. Now, apply Lemma 2.1 in the normed space $\mathbb{C}$. □

Corollary 2.3. *For every $|z| \leq 1$, with $z \neq 1$, we have*

$$\lim_{\alpha\to 0^+} \sum_{n=1}^{\infty} \frac{z^n}{n^{1+\alpha}} = \sum_{n=1}^{\infty} \frac{z^n}{n}\,.$$

Corollary 2.4. *Let T be an operator in a Banach space X and let $f \in X$. If the series $\sum_{k=1}^{\infty} \frac{T^k f}{k}$ converges, then $\lim_{\alpha\to 0^+} \sum_{k=1}^{\infty} \frac{T^k f}{k^{1+\alpha}} = \sum_{k=1}^{\infty} \frac{T^k f}{k}$.*

Remarks

1. When the series $\sum_{k=1}^{\infty} \frac{T^k f}{k}$ converges weakly, the proof of the lemma still yields norm convergence of $\sum_{k=1}^{\infty} \frac{T^k f}{k^{1+\alpha}}$ for each $\alpha > 0$, but the lemma yields only weak convergence of these series, as $\alpha \to 0^+$.
2. Combining Proposition 4.1 and Corollary 4.5 of [AL], we obtain the more difficult result (not used in the sequel) that for T power-bounded, *weak* convergence of the series $\sum_{k=1}^{\infty} \frac{T^k f}{k}$ implies the full conclusion of Corollary 2.4.

By considering the power series $\sum_{k=1}^{\infty} \frac{z^k}{k} = -\log(1-z)$ for $|z| < 1$, we conclude that for $z = re^{ix}$, with $r < 1$, $0 \le x < 2\pi$, we have (see [Z, Ch. I, p. 2]):

$$\sum_{n=1}^{\infty} \frac{r^n \cos nx}{n} = \frac{1}{2} \log \frac{1}{1 - 2r\cos x + r^2} = -\log|1-z|$$
$$\sum_{n=1}^{\infty} \frac{r^n \sin nx}{n} = \arctan \frac{r \sin x}{1 - r\cos x} = -\arg(1-z).$$

On the other hand, since the series $\sum_{n=1}^{\infty} n^{-1} \cos nx$ and $\sum_{n=1}^{\infty} n^{-1} \sin nx$ converge for $x \neq 0$ (the latter even everywhere and both converge uniformly for $\epsilon \le x \le 2\pi - \epsilon$), we have continuity at $r = 1^-$ by Abel's summability, so

$$\sum_{n=1}^{\infty} \frac{\cos nx}{n} = \log \frac{1}{|2\sin \frac{1}{2}x|} \quad \text{and} \quad \sum_{n=1}^{\infty} \frac{\sin nx}{n} = \frac{1}{2}(\pi - x), \tag{2}$$

for $0 < x < 2\pi$ (see [Z, Ch. I, p. 5]).

We also have (see [Z, Ch. II, p. 61] and [Z, Ch. V, p. 191], respectively)

$$\sup_{n\ge 1} \max_{0\le x\le 2\pi} \left| \sum_{k=1}^{n} \frac{\sin kx}{k} \right| < \infty. \tag{3}$$
$$\text{and} \quad \sup_{n\ge 1} \left| \sum_{k=1}^{n} \frac{\cos kx}{k} \right| \le \log \frac{1}{x} + C \text{ for } 0 < x \le \pi.$$

Put $S_n(x) = \sum_{k=1}^{n} \frac{\cos kx}{k}$. Abel's summation by parts (with $S_0 \equiv 0$) yields

$$\sum_{k=1}^{n} \frac{r^k \cos kx}{k} = r^n S_n(x) + \sum_{k=1}^{n-1} (r^k - r^{k+1}) S_k(x)$$
$$= r^n S_n(x) + r(1-r) \sum_{k=1}^{n-1} r^{k-1} S_k(x).$$

Hence for $0 \le r \le 1$ and $0 < x \le \pi$ we have

$$\sup_{n\ge 1} \left| \sum_{k=1}^{n} \frac{r^k \cos kx}{k} \right| \le 2r \log \frac{1}{x} + C. \tag{4}$$

Similar summation by parts for $\sum_{k=1}^{n} \frac{r^k \sin kx}{k}$ yields

$$\sup_{0\le r\le 1} \sup_{n\ge 1} \max_{0\le x\le 2\pi} \left| \sum_{k=1}^{n} \frac{r^k \sin kx}{k} \right| < \infty. \tag{5}$$

We also notice that for $0 \le r < 1$ and any x we have

$$\sup_{n\ge 1} \left| \sum_{k=1}^{n} \frac{r^k \cos kx}{k} \right| \le \sum_{k=1}^{\infty} \frac{r^k}{k} = -\log(1-r),$$

so we obtain by (4) that for every $0 \le r \le 1$ and $0 \le |x| \le \pi$ we have

$$\sup_{n\ge 1}\left|\sum_{k=1}^{n}\frac{r^k\cos kx}{k}\right| \le C + 2\min\left\{\log\frac{1}{|x|}, -\log(1-r)\right\}. \tag{6}$$

Note that only when $x = 0$ *and* $r = 1$ both sides of (6) are infinite; in all other cases they are finite.

3. The ergodic Hilbert transform for normal contractions

Let T be a normal operator on a complex Hilbert space H with resolution of the identity $E(dz)$. For $f \in H$ denote by $\sigma_f(dz) = \langle E(dz)f, f\rangle$ the spectral measure of T with respect to f. By the mean ergodic theorem, $f \in \overline{(I-T)H}$ if and only if $\sigma_f(\{1\}) = 0$.

Theorem 3.1. *Let T be a normal contraction on H and let $0 \ne f \in H$ with spectral measure σ_f. Put $D = \{z : |z| \le 1\}$. The following conditions are equivalent:*

(i) $\displaystyle\sum_{n=1}^{\infty}\frac{T^n f}{n}$ *converges strongly*;

(ii) $\displaystyle\sum_{n=1}^{\infty}\frac{T^n f}{n}$ *converges weakly*;

(iii) $\displaystyle\sup_N\left\|\sum_{j=1}^{N}\frac{T^j f}{j}\right\| < \infty$;

(iv) $\displaystyle\int_D \log^2|1-z|\sigma_f(dz) < \infty$.

If either condition holds, then $f \in \overline{(I-T)H}$,

$$\left\langle\sum_{n=1}^{\infty}\frac{T^n f}{n}, g\right\rangle = -\int_D \log(1-z)\langle E(dz)f, g\rangle \quad \textit{for every} \quad g \in H,$$

and

$$\left\|\sum_{n=1}^{\infty}\frac{T^n f}{n}\right\|^2 = \int_D |\log(1-z)|^2\sigma_f(dz).$$

Proof. Clearly, (i)⇒(ii) and by the uniform boundedness principle (ii)⇒(iii).

(iii)⇒(iv). Clearly (iii) implies that f is orthogonal to the fixed points of T^*, so $f \in \overline{(I-T)H}$, and we have $\sigma_f(\{1\}) = 0$. Hence all integrals below with respect to σ_f are in fact over $\tilde{D} = \{z : |z| \le 1,\ z \ne 1\}$.

The spectral theorem gives us the equality

$$\left\|\sum_{j=1}^{N}\frac{T^j f}{j}\right\|^2 = \int_{\tilde{D}}\left[\left(\Re\Big\{\sum_{j=1}^{N}\frac{z^j}{j}\Big\}\right)^2 + \left(\Im\Big\{\sum_{j=1}^{N}\frac{z^j}{j}\Big\}\right)^2\right]\sigma_f(dz).$$

The imaginary part is uniformly bounded on the whole closed unit disk $D = \{|z| \le 1\}$, so we just need to take care of the real part. By Fatou's lemma and the previous equality we have

$$\int_{\tilde D} \log^2|1-z|\sigma_f(dz) = \int_{\tilde D} \liminf_{N\to\infty} \Big(\Re\Big\{\sum_{n=1}^{N} \frac{z^n}{n}\Big\}\Big)^2 \sigma_f(dz) \le \sup_N \Big\| \sum_{j=1}^{N} \frac{T^j f}{j}\Big\|^2 < \infty.$$

(iv)⇒(i). The convergence of the integral yields $\sigma_f(\{1\}) = 0$. Hence all integrals with respect to σ_f are actually over $\tilde D$. By the spectral theorem, we have

$$\Big\| \sum_{j=N}^{M} \frac{T^j f}{j}\Big\|^2 = \int_{\tilde D} \Big[\Big(\Re\Big\{\sum_{j=N}^{M} \frac{z^j}{j}\Big\}\Big)^2 + \Big(\Im\Big\{\sum_{j=N}^{M} \frac{z^j}{j}\Big\}\Big)^2\Big]\sigma_f(dz).$$

We will show that $\lim_{N,M\to\infty} \| \sum_{j=N}^{M} \frac{T^j f}{j}\| = 0$.

The series $\sum_{n=1}^{\infty} \frac{z^n}{n}$ converges at each point of $\tilde D$, so we conclude that $\lim\limits_{k\to\infty} \sup\limits_{N,M\ge k} |\sum_{n=N}^{M} \frac{z^n}{n}| = 0$. Furthermore, $\Im\{\sum_{n=1}^{N} \frac{z^n}{n}\}$ is uniformly bounded on $\tilde D$, hence $\sup\limits_{k\ge 1} \sup\limits_{N,M\ge k} |\Im\{\sum_{n=N}^{M} \frac{z^n}{n}\}| < \infty$ uniformly on $\tilde D$.

Using Lebesgue's monotone convergence theorem (by considering $\sup\limits_{k\le N,M\le K}(\cdot)$ and letting $K\to\infty$), we obtain

$$\sup_{N,M\ge k} \int_D \Big(\Im\Big\{\sum_{n=N}^{M} \frac{z^n}{n}\Big\}\Big)^2 \sigma_f(dz) \le \int_D \sup_{N,M\ge k} \Big(\Im\Big\{\sum_{n=N}^{M} \frac{z^n}{n}\Big\}\Big)^2 \sigma_f(dz).$$

Using Lebesgue's dominated convergence theorem we conclude that

$$\lim_{k\to\infty} \sup_{N,M\ge k} \int_{\tilde D} \Big(\Im\Big\{\sum_{j=N}^{M} \frac{z^j}{j}\Big\}\Big)^2 \sigma_f(dz) = 0\,.$$

So, it only remains to check the assertion for $\Re\{\sum_{n=1}^{N} \frac{z^n}{n}\}$. We split $\tilde D$ into two disjoint parts by putting

$$D' = \{z\in\tilde D : |\arg z| > 1\} \quad \text{and} \quad D'' = \{z\in\tilde D : |\arg z| \le 1\}.$$

Using (6) we conclude that

$$\sup_N \max_{z\in D'} \Big|\Re\{\sum_{n=1}^{N} \frac{z^n}{n}\}\Big| \le C.$$

Again, the same arguments and using Lebesgue's dominated convergence theorem we conclude that

$$\lim_{N,M\to\infty} \int_{D'} \Big(\Re\Big\{\sum_{j=N}^{M} \frac{z^j}{j}\Big\}\Big)^2 \sigma_f(dz) = 0\,.$$

On D'' we have the following consideration. By (6) we have

$$\int_{D''} \sup_{N\geq 1} \Big(\Re\Big\{\sum_{n=1}^{N} \frac{z^n}{n}\Big\}\Big)^2 \sigma_f(dz)$$
$$\leq C_1\|f\|^2 + C_2 \int_{D''} \min\{\log^2[|\log(z/|z|)|], \log^2(1-|z|)\}\sigma_f(dz)\,.$$

Now, let $z = |z|\mathrm{e}^{ix}$ with $z \in D''$. Since $|x| \leq 1$ and $1-|z| \leq 1$ we have,

$$\min\{\log^2|x|, \log^2(1-|z|)\} = \log^2[\max\{|x|, 1-|z|\}]\,.$$

On the other hand, since $4|z|\sin^2\frac{x}{2} \leq x^2$ we obtain

$$2[\max\{|x|, 1-|z|\}]^2 \geq (1-|z|)^2 + x^2 \geq (1-|z|)^2 + 4|z|\sin^2\frac{x}{2} = |1-z|^2.$$

Since all the arguments of the logarithms below are less than or equal 1, this yields using our assumption,

$$\int_{D''} \min\{\log^2[|\log(z/|z|)|], \log^2(1-|z|)\}\sigma_f(dz)$$
$$\leq \int_{D''} \log^2\left[\frac{1}{\sqrt{2}}|1-z|\right]\sigma_f(dz) \leq C\|f\|^2 + C'\int_{D''}\log^2|1-z|\sigma_f(dz) < \infty\,.$$

Hence, using the same arguments as we have considered in the case of the imaginary part and applying Lebesgue's dominated convergence theorem, we conclude the implication $(iv) \Rightarrow (i)$.

If any of the conditions in the theorem holds, then the last assertion follows from what we have done and the convergence $\sum_{n=1}^{\infty}\frac{z^n}{n} = -\log(1-z)$ on $\tilde{D}$. □

Remarks

1. The equivalence of conditions (i) and (iv) in Theorem 3.1 is implicit (without proof) in [G4]; an explicit statement is given and proved there for T unitary.
2. For the particular cases of T unitary or self-adjoint, the equivalence of the four conditions in the theorem was proved in [AL].

Proposition 3.2. *Let T be a normal contraction on H and let $0 \neq f \in H$ with spectral measure σ_f. If*

$$\sum_{n=1}^{\infty} \frac{\|\sum_{k=1}^{n} T^k f\|^2 \log n}{n^3} < \infty\,,$$

then $\int_D \log^2|1-z|\sigma_f(dz) < \infty$.

Proof. For every $n \geq 1$ put

$$D_n := \left\{z = r\mathrm{e}^{2i\pi\theta} :\ 1-\frac{1}{n} \leq r \leq 1,\ -\frac{1}{2n} \leq \theta \leq \frac{1}{2n}\right\}.$$

Then $\{D_n\}$ is decreasing, $D_1 = D$, and $\bigcup_{n=1}^{\infty}(D_n - D_{n+1}) = D - \{1\}$.

Let $n \geq 2$. Since $(1-\frac{1}{n})^{n-1}$ decreases to 1/e, for $1-1/n \leq r \leq 1$ we have

$$r^n \geq r\left(1-\frac{1}{n}\right)^{n-1} > r/3$$

$$1-r^n = (1-r)\sum_{k=0}^{n-1} r^k \geq (1-r)nr^{n-1} \geq n(1-r)/3.$$

For $|\theta| \leq \frac{1}{2n}$ we have $|\sin(\pi n\theta)| \geq 2n|\theta| \geq \frac{2n}{\pi}|\sin(\pi\theta)|$.

For $z = re^{2i\pi\theta} \in D_n$, $n \geq 2$, since $r \geq \frac{1}{2}$, we thus obtain

$$\left|\sum_{k=1}^{n} z^k\right|^2 = |z|^2\left|\frac{1-z^n}{1-z}\right|^2 = r^2\frac{1-2r^n\cos(2\pi n\theta)+r^{2n}}{1-2r\cos(2\pi\theta)+r^2}$$

$$\geq \frac{1}{4}\frac{(1-r^n)^2+4r^n\sin^2(\pi n\theta)}{(1-r)^2+4r\sin^2(\pi\theta)} \geq \frac{n^2}{36}.$$

So, by the spectral theorem we obtain

$$\sigma_f(D_n) \leq \frac{36}{n^2}\int_{D_n}\left|\sum_{k=1}^{n} z^k\right|^2 \sigma_f(dz) \leq \frac{36}{n^2}\Big\|\sum_{k=1}^{n} T^k f\Big\|^2. \tag{7}$$

For $j \geq 2$ and $z \in D_j - D_{j+1}$ we have $\frac{j}{4} \leq \frac{1}{|1-z|} \leq j+1$, so $\int_D \log^2|1-z|\sigma_f(dz) < \infty$ if and only if $\sum_{n=1}^{\infty}(\sigma_f(D_n)-\sigma_f(D_{n+1}))\log^2 n < \infty$.

Assume that $\sum_{n=1}^{\infty}\frac{\|\sum_{k=1}^{n} T^k f\|^2 \log n}{n^3} < \infty$. Then (7) yields

$$\sum_{n=1}^{\infty}\frac{\log n\,\sigma_f(D_n)}{n} \leq 36\sum_{n=1}^{\infty}\frac{\log n\|\sum_{k=1}^{n} T^k f\|^2}{n^3} < \infty.$$

Abel's summation by parts yields

$$\sum_{n=1}^{N-1}(\sigma_f(D_n)-\sigma_f(D_{n+1}))\log^2 n \leq C'\sum_{n=1}^{N}\frac{\log n\,\sigma_f(D_n)}{n}.$$

So $\int_D \log^2|1-z|\sigma_f(dz) < \infty$. □

Remarks

1. The proposition, suggested by Christophe Cuny, leads (see below) to a characterization of the convergence of the transform by a condition on the norms of the sums (or of the averages).
2. The computations leading to (7) (and (10) below) were made in [CL], and are included for the sake of completeness. Computations of this type on the unit circle (for unitary operators) appear in [G2] and [G3].

Theorem 3.3. *Let T be a normal contraction on H and let $0 \neq f \in H$. Then the following are equivalent:*

(i) $\displaystyle\int_D \log^2|1-z|\sigma_f(dz) < \infty$

(ii) $\displaystyle\sum_{n=1}^{\infty} \frac{\langle T^n f, f\rangle \log n}{n}$ *converges.*

(iii) $\displaystyle\sum_{n=1}^{\infty} \frac{\|\sum_{k=1}^{n} T^k f\|^2 \log n}{n^3} < \infty\,.$

Proof. Proposition 3.2 shows (iii)$\Longrightarrow$(i).

(i)$\Longrightarrow$(ii): Assume $\displaystyle\int_D \log^2|1-z|\sigma_f(dz) < \infty$. By the spectral theorem,

$$\sum_{k=1}^{n} \frac{\langle T^k f, f\rangle \log k}{k} = \int_D \sum_{k=1}^{n} \frac{z^k \log k}{k} \sigma_f(dz). \tag{8}$$

We continue to denote $\tilde{D} = \{z : |z| \le 1,\ z \ne 1\}$. For every $n \ge 1$ and $z \in \tilde{D}$ we have $|\sum_{k=1}^{n} z^k| \le 2/|1-z|$. Since the sequence $\{\log n/n\}_{n>2}$ is decreasing to zero, Abel's summation by parts yields that the series $\sum_{n=1}^{\infty} \frac{z^n \log n}{n}$ converges for every $z \in \tilde{D}$. Actually, the partial sums are uniformly bounded on $\{z \in D : |1-z| \ge \epsilon > 0\}$. By our assumption $\sigma_f(\{1\}) = 0$, so $\sum_{n=1}^{\infty} \frac{z^n \log n}{n}$ converges σ_f-a.e. To prove convergence in (8), we will prove σ_f-integrability of $\sup_{n\ge1} |\sum_{k=1}^{n} \frac{z^k \log k}{k}|$.

Recall that using (3) and (6) we have already shown in the proof of Theorem 3.1 that for every $z \in \tilde{D}$,

$$\sup_{n\ge1} \Big|\sum_{k=1}^{n} \frac{z^k}{k}\Big| \le C + \Big|\log \frac{1}{|1-z|}\Big|\,.$$

Now, we majorize $\sup_{n\ge1} |\sum_{k=1}^{n} \frac{z^k \log k}{k}|$ for $z \in \tilde{D}$ with $0 < |1-z| \le \frac{1}{3}$. We fix z and put $n' = [1/|1-z|]$. For $n > n'$ write

$$\sum_{k=1}^{n} \frac{z^k \log k}{k} = \sum_{k=1}^{n'} \frac{z^k \log k}{k} + \sum_{k=n'+1}^{n} \frac{z^k \log k}{k} = P_1 + P_2.$$

We deal with two cases: (i) $n \le n'$ and (ii) $n > n'$.

Case (i): put $S_j = \sum_{k=1}^{j} \frac{z^k}{k}$. Since $\log n \le \log n' \le \log(1/|1-z|)$, we have

$$\Big|\sum_{k=1}^{n} \frac{z^k \log k}{k}\Big| \le \sum_{k=1}^{n} \frac{\log k}{k} \le C \log^2 n \le C \log^2(1/|1-z|).$$

Case (ii): put $S'_j = \sum_{k=1}^{j} z^k$. We use the decomposition $P_1 + P_2$, with P_1 estimated in case (i). Using Abel's summation, we obtain

$$\Big| \sum_{k=n'+1}^{n} \frac{z^k \log k}{k} \Big| \\ \leq \frac{\log n}{n} |S'_n| + \sum_{k=n'+1}^{n-1} \Big(\frac{\log(k)}{k} - \frac{\log(k+1)}{k+1} \Big) |S'_k| + \frac{\log(n'+1)}{n'+1} |S'_{n'}|.$$

Since $n \geq n' + 1 > 1/|1-z|$ and $\log x/x$ is decreasing, we obtain

$$|P_2| \leq \frac{\log n}{n} \frac{2}{|1-z|} + 2 \frac{\log(n'+1)}{n'+1} \frac{2}{|1-z|} \\ \leq 3 \frac{\log(1/|1-z|)}{1/|1-z|} \frac{2}{|1-z|} = 6 \log \Big(\frac{1}{|1-z|} \Big).$$

Putting the two cases together, for $z \in \tilde{D}$ we obtain

$$\sup_{n \geq 1} \Big| \sum_{k=1}^{n} \frac{z^k \log k}{k} \Big| \leq C' |\log(1/|1-z|)| + C \log^2(1/|1-z|). \qquad (*)$$

Now we prove our claim. For z close to 1, the dominant part in (*) is $\log^2(1/|1-z|)$. Hence by our assumption, the convergence in (8) follows from the Lebesgue bounded convergence theorem.

(ii) implies (iii): We first prove the implication for T unitary. In this case, it follows from the general Lemma 3 of [G2] (see also [V], [G4]), but the proof of its applicability to our case is omitted; for the sake of completeness we give the full proof.

$$\sum_{n=1}^{N} \frac{\log n \Big\| \sum_{k=1}^{n} T^k f \Big\|^2}{n^3} = \sum_{n=1}^{N} \frac{\log n \big(n \|f\|^2 + 2\Re \sum_{k=1}^{n-1} (n-k) \langle T^k f, f \rangle \big)}{n^3} \\ = \sum_{n=1}^{N} \frac{\log n \|f\|^2}{n^2} + 2\Re \sum_{n=1}^{N} \frac{\log n}{n^3} \sum_{k=1}^{n} (n-k) \langle T^k f, f \rangle.$$

The first series converges, and we show convergence of the second. Write

$$\sum_{n=1}^{N} \frac{\log n}{n^3} \sum_{k=1}^{n} (n-k) \langle T^k f, f \rangle = \sum_{k=1}^{N} \langle T^k f, f \rangle \Big[\sum_{n=k}^{N} \frac{\log n}{n^2} - k \sum_{n=k}^{N} \frac{\log n}{n^3} \Big].$$

For $h(x) > 0$ non-increasing we have

$$\int_N^{N+1} h(x) dx \leq \sum_{n=k}^{N} h(k) - \int_k^N h(x) dx \leq \int_{k-1}^{k} h(x) dx \leq h(k-1).$$

We fix K large, and approximating sums by integrals we obtain

$$\sum_{k=K}^{N}\langle T^k f,f\rangle\sum_{n=k}^{N}\frac{\log n}{n^2}=\sum_{k=K}^{N}\langle T^k f,f\rangle\left(\frac{\log k}{k}+\frac{1}{k}-\frac{\log N}{N}-\frac{1}{N}+\mathcal{O}\left(\frac{\log k}{k^2}\right)\right),$$

$$\sum_{k=K}^{N}\langle T^k f,f\rangle k\sum_{n=k}^{N}\frac{\log n}{n^3}$$
$$=\sum_{k=K}^{N}\langle T^k f,f\rangle k\left(\frac{\log k}{2k^2}+\frac{1}{4k^2}-\frac{\log N}{2N^2}-\frac{1}{4N^2}+\mathcal{O}\left(\frac{\log k}{k^3}\right)\right).$$

By Abel's summation (ii) implies convergence of $\sum_{k=2}^{\infty}\frac{\langle T^k f,f\rangle}{k}$, and by Kronecker's lemma, (ii) implies

$$\frac{\log N}{N}\sum_{k=K}^{N}\langle T^k f,f\rangle\to 0\qquad\text{and}\qquad\frac{\log N}{N^2}\sum_{k=K}^{N}k\langle T^k f,f\rangle\to 0.$$

Letting $N\to\infty$ we obtain (iii) for T unitary.

Now let T be a contraction, and let U be its unitary dilation, defined on a larger space H_1. Since $\langle U^n f,f\rangle=\langle T^n f,f\rangle$ for $f\in H$, condition (ii) for T implies the same for U, so by the above we have that $\sum_{n=1}^{\infty}\frac{\|\sum_{k=1}^{n}U^k f\|^2\log n}{n^3}$ converges. Continuity of the projection from H_1 onto H yields convergence of $\sum_{n=1}^{\infty}\frac{\|\sum_{k=1}^{n}T^k f\|^2\log n}{n^3}$. □

Remarks

1. The proof shows that the implication (ii)$\Longrightarrow$(iii) is in fact true for any contraction.
2. It follows from Theorems 3.1 and 3.3 that, when T is normal, a sufficient condition for the convergence of the one-sided ergodic Hilbert transform is $\|\frac{1}{n}\sum_{k=1}^{n}T^k f\|=\mathcal{O}\big(1/\log n(\log\log n)^{\delta}\big)$ for some $\delta>\frac{1}{2}$; this is weaker than the general assumption in [AL, remark to Corollary 2.2] (for arbitrary contractions in Banach spaces), which requires $\delta>1$.
3. For T a normal contraction, the equality $\|T^n f\|=\|T^{*n}f\|$ yields that the series in (ii) converges absolutely when
$$\sum_{n=1}^{\infty}\frac{\|T^n f\|^2\log n}{n}<\infty$$
(by separating the series in (ii) to summations on odd and even integers). When T is self-adjoint non-negative definite, the converse implication also holds.

Proposition 3.4. *Let T be a normal contraction in H. If $\sum_{k=1}^{\infty}\frac{T^k f}{k}$ converges, then*

$$\Big\|\frac{1}{n}\sum_{k=1}^{n}T^k f\Big\|=\mathcal{O}\Big(\frac{1}{\log n}\Big)\ .\tag{9}$$

Proof. If $z \in D_j - D_{j+1}$, then $1-|z| \geq \frac{1}{j+1}$ or $|1-z| \geq |z| \sin \frac{\pi}{j+1} \geq \frac{2|z|}{j+1}$, so

$$\Big|\sum_{k=1}^{n} z^k\Big| \leq \sum_{k=0}^{\infty} |z|^k = \frac{1}{1-|z|} \leq j+1 \quad \text{or} \quad \Big|\sum_{k=1}^{n} z^k\Big| \leq \frac{2|z|}{|1-z|} \leq j+1\,.$$

For $n \geq 2$ we obtain

$$\begin{aligned}
\Big\|\sum_{k=1}^{n} T^k f\Big\|^2 &= \int_D \Big|\sum_{k=1}^{n} z^k\Big|^2 \sigma_f(dz) \\
&= \int_{D_n} \Big|\sum_{k=1}^{n} z^k\Big|^2 \sigma_f(dz) + \sum_{j=1}^{n-1} \int_{D_j - D_{j+1}} \Big|\sum_{k=1}^{n} z^k\Big|^2 \sigma_f(dz) \\
&\leq n^2 \sigma_f(D_n) + \sum_{j=1}^{n-1} (j+1)^2 (\sigma_f(D_j) - \sigma_f(D_{j+1})) \\
&\leq n^2 \sigma_f(D_n) + \sum_{j=2}^{n-1} \sigma_f(D_j)((j+1)^2 - j^2) - n^2 \sigma_f(D_n) + 4\sigma_f(D_1).
\end{aligned}$$

Hence, for $n \geq 2$, we have

$$\Big\|\sum_{k=1}^{n} T^k f\Big\|^2 \leq 4 \sum_{j=1}^{n-1} j \sigma_f(D_j)\,. \tag{10}$$

Since for $z \in D_n$ we have $|1-z| \leq \frac{4}{n}$, condition (iv) of Theorem 3.1 yields $\sup_n \sigma_f(D_n) \log^2 n < \infty$. Using (10) we obtain

$$\Big\|\sum_{k=1}^{n} T^k f\Big\|^2 \leq 4 \sum_{j=1}^{n-1} j \sigma_f(D_j) \leq K \sum_{j=1}^{n-1} \frac{j}{\log^2 j} \sim K' \frac{n^2}{\log^2 n}. \qquad \square$$

Remarks

1. For unitary operators the proposition is proved in [AL].
2. Assani [A] has constructed a unitary operator T, induced on L_2 by an ergodic probability preserving transformation, and a function f satisfying (9) for which $\sum_{k=1}^{n} \frac{T^k f}{k}$ does not converge.

Theorem 3.5. *Let T be a normal contraction on H and let $0 \neq f \in H$. The following assertions are equivalent*

(i) $\displaystyle\sum_{n=1}^{\infty} \frac{T^n f}{n}$ *converges strongly*

(ii) $\displaystyle\lim_{\alpha \to 0^+} \sum_{n=1}^{\infty} \frac{T^n f}{n^{1+\alpha}}$ *converges strongly*

(iii) $\displaystyle\lim_{\alpha \to 0^+} \sum_{n=1}^{\infty} \frac{T^n f}{n^{1+\alpha}}$ *converges weakly*

(iv) $\sup_{0<\alpha<1/2} \left\| \sum_{j=1}^{\infty} \frac{T^j f}{j^{\alpha+1}} \right\| < \infty.$

If either condition holds, then $f \in \overline{(I-T)H}$ *and*

$$\lim_{\alpha\to 0^+} \sum_{n=1}^{\infty} \frac{T^n f}{n^{1+\alpha}} = \sum_{n=1}^{\infty} \frac{T^n f}{n} \quad \textit{strongly.}$$

Proof. (i)⇒(ii) follows from Corollary 2.4. Clearly, (ii)⇒(iii). The uniform boundedness principle yields (iii)⇒(iv).

Now we prove (iv)⇒(i). By (iv) we have

$$\sup_{0<\alpha<1/2} \left\| \sum_{j=1}^{\infty} \frac{T^j f}{j^{\alpha+1}} \right\| \le M < \infty,$$

so by the spectral theorem we have

$$\sup_{0<\alpha<1/2} \int_{\tilde{D}} \left(\Re\Big\{ \sum_{k=1}^{\infty} \frac{z^k}{k^{1+\alpha}} \Big\} \right)^2 \sigma_f(dz) \le M^2 < \infty.$$

Corollary 2.3 and Fatou's lemma yield

$$\begin{aligned} \int_{\tilde{D}} \log^2 |1-z| \sigma_f(dz) &= \int_{\tilde{D}} \left(\Re\Big\{ \sum_{k=1}^{\infty} \frac{z^k}{k} \Big\} \right)^2 \sigma_f(dz) \\ &= \int_{\tilde{D}} \liminf_{\alpha\to 0^+} \left(\Re\Big\{ \sum_{k=1}^{\infty} \frac{z^k}{k^{1+\alpha}} \Big\} \right)^2 \sigma_f(dz) \le M^2 < \infty\,. \end{aligned}$$

This proves (i) via Theorem 3.1.

Clearly if either condition holds, then $f \in \overline{(I-T)H}$. The last assertion follows from Corollary 2.4. □

Theorem 3.6. *Let* T *be a normal contraction or an isometry on* H *with* $\overline{(I-T)H} = H$. *For* $0 \ne f \in H$ *the following assertions are equivalent:*

(i) $\displaystyle\sum_{n=1}^{\infty} \frac{T^n f}{n}$ *converges strongly*

(ii) $\displaystyle\sum_{n=1}^{\infty} \frac{T^n f}{n}$ *converges weakly*

(iii) $\displaystyle\sum_{n=1}^{\infty} \frac{\| \sum_{k=1}^{n} T^k f\|^2 \log n}{n^3} < \infty\,.$

(iv) $\displaystyle\sum_{n=1}^{\infty} \frac{\langle T^n f, f\rangle \log n}{n}$ *converges.*

(v) f *is in the domain of* G, *the infinitesimal generator of the semi-group* $\{(I-T)^r : r \ge 0\}$.

If either condition holds, then

$$Gf = -\sum_{n=1}^{\infty} \frac{T^n f}{n}.$$

Proof. Assume first that T is a normal contraction. We already know by Theorem 3.1 and Theorem 3.3 that the first four conditions are equivalent. By Theorem 3.5 (i) is equivalent to the convergence of $\lim_{\alpha \to 0^+} \sum_{n=1}^{\infty} \frac{T^n f}{n^{1+\alpha}}$ and by Corollary 4.5 in [AL] this last convergence is equivalent to (v).

When either condition holds, we apply [AL, Proposition 4.1] (or [DL, Theorem 2.22(ii)]) to obtain $Gf = -\sum_{n=1}^{\infty} \frac{T^n f}{n}$.

Assume now that T is an isometry, and let U be its unitary dilation (on a larger space H_1). By the construction, $T^n f = EU^n f$ for $f \in H$ and $n > 0$, where E is the orthogonal projection from H_1 onto H, and since T is an isometry we have $T^n f = U^n f$. An application of Theorem 3.5 to U yields that it is in fact valid also for the isometry T. Similarly, by Theorem 3.1 and Theorem 3.3, the first four conditions of the theorem are equivalent for the isometry T. Now the first part of the proof yields the result. □

Corollary 3.7. *Let T be a contraction on H such that T^* is an isometry. Then $\sum_{n=1}^{\infty} \frac{T^n f}{n}$ converges weakly if and only if it converges strongly.*

Proof. We may restrict ourselves to $\overline{(I-T)H} = \overline{(I-T^*)H}$. If $\sum_{n=1}^{\infty} \frac{T^n f}{n}$ converges weakly, then by Proposition 1.2 and the previous theorem applied to T^* we have strong convergence. □

Remark. Similarly, if T^* is an isometry, $\sum_{n=1}^{\infty} \frac{T^n f}{n}$ converges if and only if $\sum_{n=1}^{\infty} \frac{\langle T^n f, f\rangle \log n}{n}$ converges.

Corollary 3.8. *Let T be a normal contraction on H such that $\overline{(I-T)H} = H$ (so also $\overline{(I-T^*)H} = H$). Then the infinitesimal generators of $\{(I-T)^r : r \geq 0\}$ and $\{(I-T^*)^r : r \geq 0\}$ have the same domain of definition.*

Proof. Use Proposition 1.2 and the characterization of the domain of the generator given by Theorem 3.6. □

4. The ergodic Hilbert transform for general contractions

For any contraction T on H, weak convergence of $\sum_{k=1}^{\infty} \frac{T^k f}{k}$ implies convergence of the series

$$\sum_{k=1}^{\infty} \frac{\langle T^k f, f\rangle}{k}. \tag{11}$$

Convergence of (11) yields $\|\frac{1}{n}\sum_{k=1}^{n} T^k f\| \to 0$, by Kronecker's lemma and the next proposition.

Proposition 4.1. *Let T be a contraction on a Hilbert space H and $f \in H$. If $\frac{1}{n}\sum_{k=1}^{n}\langle T^k f, f\rangle \to 0$, then $\|\frac{1}{n}\sum_{k=1}^{n} T^k f\| \to 0$.*

Proof. By the mean ergodic theorem, $\frac{1}{n}\sum_{k=1}^{n} T^k f$ converges to some $g \in H$, and $Tg = g$, so also $T^*g = g$ [RN, §144]. Hence the assumption yields

$$\|g\|^2 = \lim_{n\to\infty}\left\langle g, \frac{1}{n}\sum_{k=1}^{n} T^k f\right\rangle = \lim_{n\to\infty}\left\langle \frac{1}{n}\sum_{k=1}^{n} T^{*k} g, f\right\rangle = \langle g, f\rangle = 0. \qquad \square$$

Remark. Foguel [F] proved that if $\langle T^n f, f\rangle \to 0$, then $T^n f \to 0$ weakly. As mentioned above, for the averages weak and strong convergence are the same.

Since the condition of Theorem 3.3(ii) is stronger than convergence of the series (11), the latter convergence is not expected to imply convergence of $\sum_{k=1}^{\infty}\frac{T^k f}{k}$. This will be exhibited in the examples below.

Theorem 4.2. *Let T be a contraction on a complex Hilbert space H and $f \in H$. If $\sum_{k=1}^{\infty}\frac{\langle T^k f, f\rangle \log k}{k}$ converges, then $\sum_{k=1}^{\infty}\frac{T^k f}{k}$ converges strongly.*

Proof. For T unitary the assertion follows from Theorem 3.6.

Now let T be a contraction, and let U be its unitary dilation, defined on a larger space H_1. Since $\langle U^n f, f\rangle = \langle T^n f, f\rangle$ for $f \in H$, by Theorem 3.6 applied to U the assumption yields strong convergence of $\sum_{k=1}^{\infty}\frac{U^k f}{k}$, and continuity of the projection from H_1 onto H yields convergence of $\sum_{k=1}^{\infty}\frac{T^k f}{k}$. $\square$

When $H = L_2(\mathcal{S}, \Sigma, m)$ of a σ-finite measure space, it is of interest to investigate also the almost everywhere (a.e.) convergence of the one-sided ergodic Hilbert transform of a contraction T. For T unitary there are extensive studies by Gaposhkin ([G2], [G3], [G4]). Gaposhkin assumes m to be a probability, but this is not a restriction, since (e.g., [Kr, p. 189]) when m is not finite we take an equivalent probability m' and the map $Vf := f/\sqrt{\psi}$, with $\psi = dm'/dm$, is an order-preserving linear isometry of $L_2(m)$ onto $L_2(m')$ which preserves also pointwise convergence.

Theorem 4.3. *Let T be a contraction of $L_2(\mathcal{S}, m)$ of a σ-finite measure space, and $f \in L_2(m)$. If*

$$\sum_{n=1}^{\infty}\frac{\langle T^n f, f\rangle \log n(\log\log\log n)^2}{n} \quad \text{converges} \tag{12}$$

then $\sum_{k=1}^{\infty}\frac{T^k f}{k}$ converges a.e. (and in norm).

Proof. The norm convergence follows from Theorem 4.2.

If T is unitary, this is Theorem 3a of [G4] (see also [G2, Theorem 7]).

Now let T be a contraction of $L_2(\mathcal{S}, m)$. We may assume that m is a probability. We will use Schäffer's construction of the unitary dilation [Sc]: Let $(\mathcal{S}_n, m_n)$

be disjoint copies $(\mathcal{S}, m)$, put $\Omega = \bigcup_{n\in\mathbb{Z}} \mathcal{S}_n$ with the obvious σ-algebra, and define $\mu(A) = \sum_{n\in\mathbb{Z}} m_n(A \cap \mathcal{S}_n)$. Then $L_2(\Omega, \mu) = \sum_n \oplus L_2(\mathcal{S}_n, m_n)$, and the unitary dilation U is defined on $L_2(\mu)$. The orthogonal projection on $L_2(\mathcal{S}_0, m_0)$ is in fact multiplication by the indicator function $1_{\mathcal{S}_0}$. If (12) is satisfied, then also $\tilde{f}$, the extension by zero to Ω of f on $\mathcal{S}_0$, satisfies (12) with T replaced by U. Now we apply Gaposhkin's result to obtain μ-a.e. convergence of $\sum_{n=1}^{\infty} \frac{U^n \tilde{f}}{n}$ on Ω, which yields m_0-a.e. convergence on $\mathcal{S}_0$ of $\sum_{n=1}^{\infty} \frac{T^n f}{n}$. □

Note that there are contractions on L_2 for which even the averages may fail to converge a.e. ([B, p. 128]; for examples of unitary operators see [G1] or [Kr, p. 191]). The proof of [Kr, Lemma 5.2.1] can be adapted to show that if T is power-bounded on L_2 and $f \in (I - T)^{\alpha} L_2$ with $\alpha > \frac{1}{2}$, then $\frac{1}{n}\sum_{k=1}^{n} T^k f \to 0$ a.e. The next proposition shows that for contractions we can do better.

Proposition 4.4. *Let T be a contraction of $L_2(\mathcal{S}, m)$ of a σ-finite measure space, and $f \in L_2(m)$. If the series* (11) *converges, in particular if $\sum_{k=1}^{\infty} \frac{T^k f}{k}$ converges weakly, then $\frac{1}{n}\sum_{k=1}^{n} T^k f \to 0$ a.e.*

Proof. For T unitary, this is due to Gaposhkin [G1, Theorem 2]. In the general case, we use the unitary dilation of [Sc] as in the previous proof. □

Theorem 4.5. *Let T be a contraction of $L_2(\mathcal{S}, m)$ of a σ-finite measure space, and $f \in L_2(m)$. If*

$$\sum_{n=1}^{\infty} \frac{\|\frac{1}{n}\sum_{k=1}^{n} T^k f\|}{n} \quad \textit{converges} \tag{13}$$

then $\sum_{k=1}^{\infty} \frac{T^k f}{k}$ converges a.e. (and in norm).

Proof. We may assume that m is a probability. Denote $S_n f := \sum_{k=1}^{n} T^k f$. The mean ergodic theorem, (13), and the identity

$$\sum_{k=1}^{n} \frac{T^k f}{k} = \frac{S_n f}{n} + \sum_{k=1}^{n-1} \frac{1}{k(k+1)} S_k f$$

yield that $\sum_{k=1}^{n} \frac{T^k f}{k}$ converges strongly in L_2. By Proposition 4.4, $\frac{1}{n} S_n f \to 0$ a.e.; since $\|S_n f\|_1 \le \|S_n f\|_2$ in a probability space, by (13) and Beppo Levi's theorem $\sum_{k=1}^{\infty} \frac{1}{k(k+1)} S_k f$ converges a.e. Thus $\sum_{k=1}^{n} \frac{T^k f}{k}$ converges a.e. □

Remark. The previous theorem is true also for isometries or *order-preserving* contractions of L_p, $1 < p < \infty$. The proof is the same, except that instead of Proposition 4.4, we use Kan's pointwise ergodic theorem [Kn, Corollary 5.1] for isometries of L_p, $p \neq 2$, and Akcoglu's pointwise ergodic theorem (e.g., [Kr, p. 190]) for order-preserving contractions.

For T unitary on L_2, Gaposhkin [G4] proved that

$$\sum_{n=1}^{\infty} \frac{\|\frac{1}{n}\sum_{k=1}^{n} T^k f\|^2 \log n (\log\log\log n)^2}{n} < \infty \tag{14}$$

is sufficient for a.e. (and norm) convergence of the ergodic Hilbert transform, and showed that (12) implies (14). We do not know if this latter condition implies a.e. convergence of the transform for general contractions on L_2. We even do not know if $\|\frac{1}{n}\sum_{k=1}^{n} T^k f\| = \mathcal{O}\big(\frac{1}{\log n(\log\log n)^{\delta}}\big)$ for some $\delta > \frac{1}{2}$, which implies (14), is sufficient for a.e. convergence of the transform for general contractions in L_2.

Example 1. *U unitary on L_2, with $f \in L_2$ satisfying* (14)*, but not* (13).

Put $H = L_2([0,1), dt)$ and for $h \in H$ define $Uh(t) = \mathrm{e}^{2\pi i t} h(t)$ (the operator induced by the shift). Denote $\log_2 x := \log(\log x)$, and for $k \geq 4 > \mathrm{e}+1$ put $c_k := (\sqrt{k \log^3 k}\log_2 k)^{-1}$, $c_k = 0$ for $k < 4$. Let $f := \sum_{k=1}^{\infty} c_k \mathrm{e}^{2\pi i k t}$. Clearly,

$$\begin{aligned}\Big\|\sum_{j=1}^{n} U^j f\Big\|^2 &= \Big\|\sum_{k=1}^{\infty} \mathrm{e}^{2\pi i k t} \sum_{j=1}^{n} c_{k-j}\Big\|^2 = \sum_{l=0}^{\infty} \sum_{k=ln+1}^{(l+1)n} \Big(\sum_{j=1}^{n} c_{k-j}\Big)^2 \\ &= \sum_{k=1}^{n} \Big(\sum_{j=1}^{n} c_{k-j}\Big)^2 + \sum_{k=n+1}^{2n} \Big(\sum_{j=1}^{n} c_{k-j}\Big)^2 + \sum_{l=2}^{\infty} \sum_{k=ln+1}^{(l+1)n} \Big(\sum_{j=1}^{n} c_{k-j}\Big)^2 \\ &= \Sigma_I + \Sigma_{II} + \Sigma_{III}.\end{aligned}$$

We start with Σ_{III}. For $n \geq 4$, the monotonicity of $\{c_k\}_{k\geq 4}$ yields

$$\begin{aligned}\Sigma_{III} &\leq \sum_{l=2}^{\infty} \sum_{k=ln+1}^{(l+1)n} (n \cdot c_{(l-1)n})^2 = \sum_{l=2}^{\infty} n^3 (c_{(l-1)n})^2 \\ &= n^2 \sum_{l=1}^{\infty} \frac{1}{l \log^3(ln) \log_2^2(ln)} \leq n^2 C \int_{n-1}^{\infty} \frac{dx}{x \log^3 x \log_2^2 x} \leq \frac{C_1 n^2}{(\log n \log_2 n)^2}.\end{aligned}$$

For Σ_{II} we have, using monotonicity,

$$\Sigma_{II} \leq n\Big(\sum_{k=4}^{n+3} c_k\Big)^2 \leq nC\Big(\int_3^{n+3} \frac{dx}{\sqrt{x \log^3 x}\log_2 x}\Big)^2 \leq \frac{C_2 n^2}{\log^3 n (\log_2 n)^2}.$$

The same estimate holds for Σ_I, since $\Sigma_I \leq n\big(\sum_{k=4}^{n-1} c_k\big)^2 \leq n\big(\sum_{k=4}^{n+3} c_k\big)^2$.

On the other hand, $\big\|\sum_{j=1}^{n} U^j f\big\|^2 \geq \Sigma_{III} \geq \frac{C'' n^2}{(\log n \log_2 n)^2}$, by a similar computation. Hence

$$\frac{C'' n^2}{(\log n \log_2 n)^2} \leq \Big\|\sum_{j=1}^{n} U^j f\Big\|^2 \leq \frac{C' n^2}{(\log n \log_2 n)^2}$$

and the assertion clearly follows. In fact, also (12) holds, since for $n \geq 4$

$$\langle U^n f, f\rangle = \sum_{l=0}^{\infty} \sum_{k=ln+1}^{(l+1)n} c_k c_{k+n} \leq c_n \sum_{k=1}^{n} c_k + \sum_{l=1}^{\infty} n(c_{ln})^2.$$

Definition. A *Dunford-Schwartz operator* is a contraction T of $L_1(\mathcal{S}, m)$ which is also a contraction of L_∞ (if m is σ-finite infinite, T is extended to L_∞ from $L_1 \cap L_\infty$). By the Riesz-Thorin theorem (see also [Kr, p. 65]), T is also (extendable to) a contraction of $L_2(\mathcal{S}, m)$.

Measure preserving transformations, and more generally Markov operators with a subinvariant measure, induce order-preserving Dunford-Schwartz operators. If T is a Dunford-Schwartz operator, then $\frac{1}{n}\sum_{k=1}^{n} T^k f$ converges a.e. for any $f \in L_p$, $1 \leq p < \infty$ (e.g., [DuS, p. 675]).

Example 2. *A self-adjoint Dunford-Schwartz operator T, $f \in L_2$ with* (11) *convergent, $\sum_{k=1}^{\infty} \frac{T^k f}{k}$ converges a.e. but not weakly in L_2.*

Let m be the finite measure on the Borel sets of $[0,1)$ with density $\frac{dm}{dt} = 1/(1-t)|\log(1-t)|^3$ for $t > \frac{1}{2}$ and $\frac{dm}{dt} = c$ for $0 \leq t \leq \frac{1}{2}$. On $L_1([0,1), m)$ define the operator $Th(t) = th(t)$, which is obviously Dunford-Schwartz. Since $\langle T^n h, h\rangle = \int t^n |h|^2 dm$ for $h \in H = L_2([0,1), m)$, the function $f \equiv 1$ has the spectral measure $\sigma_f = m$, and by Beppo Levi

$$\sum_{k=1}^{\infty} \frac{\langle T^k f, f\rangle}{k} = \sum_{k=1}^{\infty} \frac{\int_{[0,1)} t^k dm}{k} = \int_{[0,1)} |\log(1-t)| dm < \infty.$$

However, the one-sided ergodic Hilbert transform does not converge weakly by Theorem 3.1, since

$$\int_{[0,1)} |\log(1-t)|^2 dm \geq \int_{1/2}^{1} \frac{1}{(1-t)\,|\log(1-t)|} dt = \infty\ .$$

Remark. In this example $H = \overline{(I-T)H}$, and $\sum_{k=1}^{\infty} \frac{T^k h}{k}$ converges a.e. for *every* $h \in H$, although $(I-T)H$ is not closed. Note that T is order-preserving.

Example 3. *U unitary on L_2, $f \in L_2$ with $\sum_{k=1}^{\infty} \frac{U^k f}{k}$ convergent a.e., but not weakly in L_2.*

Let T be the operator on $L_2([0,1), m)$ described in the previous example, and let U be the unitary dilation constructed by Schäffer [Sc], which is defined on $L_2(\mathbb{R}, \mu)$ (see the proof of Theorem 4.3). For f of the previous example we define $\tilde{f}$ on $\mathbb{R}$ by $\tilde{f}(x) = f(x)$ for $0 \leq x < 1$ and $\tilde{f}(x) = 0$ otherwise. Schäffer's definition of U yields that $U^n \tilde{f}$ is zero on $[1, \infty)$ and on $(-\infty, -n)$, on the interval $[-j, -j+1)$ $(1 \leq j \leq n)$ we have $U^n \tilde{f}(x) = \sqrt{1-(x+j)^2}(x+j)^{n-j} f(x+j)$, and $U^n \tilde{f}(x) = x^n f(x)$ on $[0,1)$. This yields that $\sum_{n=1}^{\infty} \frac{U^n \tilde{f}}{n}$ converges a.e. However, L_2-weak convergence of $\sum_{k=1}^{\infty} \frac{U^k \tilde{f}}{k}$ would imply that of $\sum_{k=1}^{\infty} \frac{T^k f}{k}$, a contradiction.

Note that since $\langle U^n \tilde{f}, \tilde{f}\rangle = \langle T^n f, f\rangle$, the series $\sum_{n=1}^{\infty} \frac{\langle U^n \tilde{f}, \tilde{f}\rangle}{k}$ converges.

Remarks

1. In the example convergence a.e. of the transform does not imply weak (norm) convergence, which shows that for unitary operators, Gaposhkin's sufficient condition (14) for a.e. convergence of the transform is not necessary; neither is condition (13), as either condition implies also norm convergence.
2. Gaposhkin [G4, pp. 253–254] constructed an example of U unitary on $L_2[0,1]$ and a function f such that $\sum_{k=1}^{\infty} \frac{U^k f}{k}$ converges in norm, but not a.e. Thus, for unitary operators on L_2, a.e. and norm convergence of the series are not comparable in general. It is worth to mention that in his example the ergodic averages do converge a.e. to 0, since the sufficient conditions of Theorem 3A in [G3] are satisfied, and also the *two-sided Hilbert transform* converges a.e.
3. Almost everywhere convergence of the transform for every function (for which the averages converge to 0), as in Example 2, cannot occur for U induced by an invertible ergodic measure preserving transformation of a separable non-atomic probability space: Kakutani and Petersen [KP] proved that there always exists a bounded function of zero integral for which the one-sided ergodic Hilbert transform is a.e. non-convergent; for references to earlier related results see [AL].

Let T be a contraction in H such that $(I-T)H$ is not closed. By Theorem 1.1 there exists $f \in \overline{(I-T)H}$ such that $\sum_{n=1}^{\infty} \frac{T^n f}{n}$ does not converge. A natural question (raised in the context of Fourier series – see [Z, Theorem V(8.12)] with remarks and references on [Z, p. 380]), is the convergence of the one-sided ergodic Hilbert transform for almost every random choice of signs, i.e., the convergence of $\sum_{n=1}^{\infty} \pm\frac{T^n f}{n}$ for every $f \in \overline{(I-T)H}$. This is made precise in the following theorem, when we take for $\{\xi_n\}$ the Rademacher functions.

Theorem 4.6. *Let $\{\xi_n\}$ be independent identically distributed random variables on a probability space $(\Omega, \mathcal{F}, \mathbf{P})$ such that $\mathbb{E}(|\xi_1| \log^+ |\xi_1|) < \infty$ and $\mathbb{E}\xi_1 = 0$. Then there exists a set $\Omega_1 \in \mathcal{F}$ with $\mathbf{P}(\Omega_1) = 1$ such that when $\omega \in \Omega_1$, for every contraction T on a Hilbert space H and any $f \in H$ the "modulated" transform $\sum_{n=1}^{\infty} \xi_n(\omega)\frac{T^n f}{n}$ converges in norm.*

Proof. Cuzick and Lai [CuLa, Theorem 1(iv)] proved that for $\{\xi_n\}$ as in the theorem there exists $\Omega_1 \in \mathcal{F}$ with $\mathbf{P}(\Omega_1) = 1$ such that for $\omega \in \Omega_1$ the "random Fourier series"

$$\sum_{n=1}^{\infty} \frac{\xi_n(\omega)}{n} \lambda^n$$

converges uniformly for complex $|\lambda| = 1$. We fix $\omega \in \Omega_1$.

For a unitary operator U on H and $f \in H$ the spectral theorem yields

$$\Big\| \sum_{n=j}^{k} \frac{\xi_n(\omega) U^n f}{n} \Big\|^2 = \int_{\{|\lambda|=1\}} \Big| \sum_{n=j}^{k} \frac{\xi_n(\omega)\lambda^n}{n} \Big|^2 d\sigma_f \le \|f\|^2 \sup_{|\lambda|=1} \Big| \sum_{n=j}^{k} \frac{\xi_n(\omega)\lambda^n}{n} \Big|^2$$

which converges to 0 as $k > j \to \infty$ by the choice of ω. Hence $\sum_{n=1}^{\infty} \xi_n(\omega)\frac{U^n f}{n}$ converges in norm.

For T a contraction on H let U be its unitary dilation on H_1 containing H. Then

$$\Big\| \sum_{n=j}^{k} \frac{\xi_n(\omega) T^n f}{n} \Big\|^2 \le \Big\| \sum_{n=j}^{k} \frac{\xi_n(\omega) U^n f}{n} \Big\|^2 \underset{k>j\to\infty}{\longrightarrow} 0$$

so $\sum_{n=1}^{\infty} \xi_n(\omega)\frac{T^n f}{n}$ converges in norm. □

Acknowledgement

The authors are grateful to Christophe Cuny for many valuable remarks and to Idris Assani for sending them a copy of [A].

References

[A] I. Assani, Pointwise convergence of the one-sided ergodic Hilbert transform, preprint.

[AL] I. Assani and M. Lin, On the one-sided ergodic Hilbert transform, *Contemp. Math.* 430 (2007), 221–39.

[B] D. Burkholder, Semi-Gaussian spaces, *Trans. Amer. Math. Soc.* 104 (1962), 123–131.

[Ca] J. Campbell, Spectral analysis of the ergodic Hilbert transform, *Indiana Univ. Math. J.* 35 (1986), 379–390.

[CL] C. Cuny and M. Lin, Pointwise ergodic theorems with rates and application to the CLT for Markov chains, *Ann. Inst. Poincaré Proba. Stat.*, to appear.

[CuLa] J. Cuzick and T.L. Lai, On random Fourier series, *Trans. Amer. Math. Soc.* 261 (1980), 53–80.

[DL] Y. Derriennic and M. Lin, Fractional Poisson equations and ergodic theorems for fractional coboundaries, *Israel J. Math.* 123 (2001), 93–130.

[DuS] N. Dunford and J. Schwartz, *Linear operators*, part I, Wiley Interscience, New York, 1958.

[F] S. Foguel, Powers of a contraction in Hilbert space, *Pacific J. Math.* 13 (1963), 331–562.

[G1] V. Gaposhkin, On the strong law of large numbers for second order stationary processes and sequences, *Theory of probability and its appl.* 18 (1973), 372–375.

[G2] V. Gaposhkin, Convergence of series connected with stationary sequences, *Math. USSR Izv.* 9 (1975), 1297–1321.

[G3] V. Gaposhkin, Criteria for the strong law of large numbers for some classes of weakly stationary processes and homogeneous random fields, *Theory of probability and its appl.* 22 (1977), 286–310.

[G4] V. Gaposhkin, Spectral criteria for existence of generalized ergodic transforms, *Theory of probability and its appl.* 41 (1996), 247–264.

[H] P.R. Halmos, A non-homogeneous ergodic theorem, *Trans. Amer. Math. Soc.* 66 (1949), 284–288.

[I] S. Izumi, A nonhomogeneous ergodic theorem, *Proc. Imp. Acad. Tokyo* 15 (1939), 189–192.

[KP] S. Kakutani and K. Petersen, The speed of convergence in the ergodic theorem, *Monatshefte Math.* 91 (1981), 11–18.

[Kn] C.H. Kan, Ergodic properties of Lamperti operators, *Canadian J. Math.* 30 (1978), 1206–1214.

[Kr] U. Krengel, *Ergodic theorems*, De Gruyter, Berlin, 1985.

[L] M. Lin, On the Uniform ergodic theorem, *Proc. Amer. Math. Soc.* 43 (1974), 337–340.

[RN] F. Riesz and B. Sz-Nagy. (1990). *Functional analysis*, Translated from the 2nd French edition by L.F. Boron, Dover Publications inc., New York.

[Sc] J.J. Schäffer, On unitary dilations of contractions, *Proc. Amer. Math. Soc.* 6 (1955), 322.

[V] I.N. Verbitskaya (Verbickaja), On conditions for the applicability of the strong law of large numbers to wide sense stationary processes, *Theory of probability and its appl.* 11 (1966), 632–636.

[Z] A. Zygmund, *Trigonometric series*, corrected 2nd ed., Cambridge University Press, Cambridge, 1968.

Guy Cohen
Dept. of Electrical Engineering
Ben-Gurion University
e-mail: guycohen@ee.bgu.ac.il

Michael Lin
Dept. of Mathematics
Ben-Gurion University
e-mail: lin@math.bgu.ac.il

Operator Theory:
Advances and Applications, Vol. 197, 99–132

Model Reduction in Symbolically Semi-separable Systems with Application to Pre-conditioners for 3D Sparse Systems of Equations

Patrick Dewilde, Haiyan Jiao and Shiv Chandrasekaran

In dear memory of Prof. M. Livsic

Abstract. Preconditioned iterative solvers are considered to be one of the most promising methods for solving large and sparse linear systems. It has been shown in the literature that their impact can be fairly easily extended to semi-separable systems or even larger classes build on semi-separable ideas. In this paper, we propose and evaluate a new type of preconditioners for the class of matrices that have a two level deep 'symbolically hierarchical semi-separable form' meaning that the matrices have a semi-separable like block structure with blocks that are (sequentially) semi-separable themselves. The new preconditioners are based on approximations of Schur complements in a sequential or hierarchical decomposition of the original block matrix. The type of matrices considered commonly occur in 3D modeling problems.

Mathematics Subject Classification (2000). 65F10, 65F15; 65F50; 93A13.

Keywords. Preconditioners, semi-separable systems, model reduction, Poisson equation.

1. Introduction

The importance of preconditioners to solve large systems of sparse equations has been amply demonstrated in the literature, an excellent survey is to be found in [3]. However, in a number of crucial cases, finding good preconditioners has proved to be very difficult, if not impossible, lacking a systematic method to construct them either from basic principles or from the physical circumstances leading to the system to be solved. In this paper we propose a method based on algebraic principles, but which can also accommodate physical considerations to some extent. The proposed method is adequate to handle systems that extend in 2 dimensions

(2D-systems) but we want to show that the ideas will extent to 3D systems as well. Although the method applies to a fairly general class of systems, we are only able to validate it on systems that can be solved explicitly. In this paper we consider one type of such systems: positive definite, Hermitian of the block-Toeplitz-block-Toeplitz (BTBT) kind and we consider two cases: purely Toeplitz and circulant. Validation to larger classes has necessarily to be experimental, but the physical connotations of the method makes it a very good candidate for future use in 3D systems in general.

The generic system solver solves a set of equations

$$\Phi u = b \tag{1}$$

in which Φ is a square matrix, b a conformal vector of data and u is the solution vector to be found. We assume that Φ is non-singular and even that it can be LU-factored (the method can be extended to the Moore-Penrose case, but that is beyond our present scope). We put further assumptions on the structure of Φ that make the class considered adequate for fairly general modeling problems that lead to 3D sparse matrices.

A good preconditioner P is a matrix of the same dimensions as Φ such that (1) $I - \Phi P$ is small and (2) multiplication with P is computationally cheap. Iterative solvers are adequate when (1) also the multiplication with Φ is cheap and (2) a good P is known or can easily be determined. The iterative solver will then iterate on the error residue and converge quickly when the eigenvalues of $I - \Phi P$ are close to zero – we refer to the literature for more details [3].

Solvers based on preconditioners are obviously attractive when Φ is a sparse matrix, for then the condition of 'cheap multiplication with Φ' is automatically fulfilled. However, this is certainly not the only class that leads to cheap multiplication. Another is the class of 'sequentially semi-separable matrices' [5, 7], or the class of 'hierarchically semi-separable matrices' [1]. These classes are distinct, sparse matrices are not semi-separable in general (only banded matrices are). Hierarchically semi-separable matrices can be transformed into specific classes of sparse matrices, making it an attractive class because the extra structure allows for efficient solving, either in a direct or a preconditioned way. The problem with general classes of sparse matrices is the difficulty of finding a good preconditioner. Our approach is to extend the class of structured matrices of the semi-separable type so that it covers a wider collection of sparse matrices and transformations thereof. The extension that we consider in this paper (and that is described in the next paragraph) is able to cover most, if not all, 2D type modeling problems, whether of the sparse type or the so-called 'multipole' type.

The matrix structure that we consider in this paper can be termed 'symbolically semi-separable'. We shall treat the semi-separable structure extensively in a further section. A semi-separable matrix is characterized by a so-called 'realization', i.e., an ordered sets of seven (small) diagonal block-matrices denoted, e.g., as $\{A, B, C, D, A', B', C'\}$. We say that the structure is 'symbolical' if the characterization has the same form, but the characterizing set of matrices has further struc-

ture, namely all the submatrices are themselves either sequentially semi-separable or symbolically semi-separable again. E.g., if $A = \mathrm{diag}[\cdots A_k \cdots]$ where each A_k is sequentially semi-separable (and hence characterized again by a realization at a lower hierarchical level) then the symbolical hierarchy will have two layers.

Hence, our goal will be the construction of preconditioners, assuming the underlying matrix structure to be given in terms of blocks that themselves have a sequential or symbolical semi-separable structure. We shall formulate the theory and the results at a 'medium complexity level' – to keep things as simple as possible without endangering the generality needed to handle significant 3D modeling cases. In particular, we shall assume a block tri-diagonal form for the top level hierarchy. This structure is less general than full blown symbolic semiseparability, but it does cover the main application, namely systems originating from 3D finite element modeling. A special case is obtained when second-order 3D partial differential equation is considered on a regular (finite 3D) grid. We shall develop this case for Laplace's (or Poisson's) equation in the next section and carry it as a test case throughout the paper, comparing the performance of the various preconditioners proposed. In particular, we use a 27-point stencil to discretized the PDF, basic cells of dimension 8×8 resulting in an overall matrix of dimension $8^3 \times 8^3$. Measures for performance of the preconditioner P are norm differences between I and ΦP and the largest eigenvalue of the matrix $I - \Phi P$ because it determines the rate of convergence (we wish it typically smaller than 0.1).

2. Prototype example

As prototype example and to fix ideas, we consider Poisson's equation in a homogeneous medium, discretized on a uniform 3D grid. A formulation of Poisson's equation requires the solution of

$$-\left(\frac{\partial^2}{\partial x}u(x,y,z) + \frac{\partial^2}{\partial y}u(x,y,z) + \frac{\partial^2}{\partial z}u(x,y,z)\right) = f(x,y,z)$$

for $(x, y, z) \in \Omega$ where $\Omega = [0,1] \times [0,1] \times [0,1]$ with boundary conditions that after discretization with a 27 point stencil results in either a hierarchical $n^3 \times n^3$ block-tridiagonal block-Toeplitz or block circulant system of equations. Let us define a parameter $\epsilon = 0$ for the block-tridiagonal case and $\epsilon = 1$ for the circulant case, then the discretized equations to be solved take the form

$$\Phi u = b \tag{2}$$

$$\begin{pmatrix} M & -L^H & & & -\epsilon L \\ -L & M & -L^H & & \\ & -L & M & \ddots & \\ & & \ddots & \ddots & -L^H \\ -\epsilon L^H & & & -L & M \end{pmatrix} \begin{pmatrix} u_0 \\ u_0 \\ \vdots \\ u_{m-1} \end{pmatrix} = \begin{pmatrix} b_0 \\ b_1 \\ \vdots \\ b_{m-1} \end{pmatrix}$$

where we have assumed that u_i are the discretized unknowns along the ith column of the $n \times n \times n$ grid. Φ is a symmetric positive definite matrix with n block columns, and has the same sub-blocks on each of the tri-diagonal given by

$$M = \begin{pmatrix} O & -P^H & & & -\epsilon P \\ -P & O & -P^H & & \\ & -P & O & -P^H & \\ & & \ddots & \ddots & \ddots \\ -\epsilon P^H & & & -P & O \end{pmatrix} \tag{3}$$

$$L = \begin{pmatrix} R & Q^H & & & \epsilon Q \\ Q & R & Q^H & & \\ & Q & R & Q^H & \\ & & \ddots & \ddots & \ddots \\ \epsilon Q^H & & & Q & R \end{pmatrix} \tag{4}$$

$$O = \frac{1}{30} \begin{pmatrix} 128 & -14 & & & -14\epsilon \\ -14 & 128 & -14 & & \\ & -14 & 128 & -14 & \\ & & \ddots & \ddots & \ddots \\ -14\epsilon & & & -14 & 128 \end{pmatrix} \tag{5}$$

$$P = \frac{1}{30} \begin{pmatrix} 14 & 3 & & & \epsilon 3 \\ 3 & 14 & 3 & & \\ & 3 & 14 & 3 & \\ & & \ddots & \ddots & \ddots \\ \epsilon 3 & & & 3 & 14 \end{pmatrix} \tag{6}$$

$$Q = \frac{1}{30} \begin{pmatrix} 3 & 1 & & & \epsilon 1 \\ 1 & 3 & 1 & & \\ & 1 & 3 & 1 & \\ & & \ddots & \ddots & \ddots \\ \epsilon 1 & & & 1 & 3 \end{pmatrix} \tag{7}$$

$$R = \frac{1}{30} \begin{pmatrix} 14 & 3 & & & \epsilon 3 \\ 3 & 14 & 3 & & \\ & 3 & 14 & 3 & \\ & & \ddots & \ddots & \ddots \\ \epsilon 3 & & & 3 & 14 \end{pmatrix} \tag{8}$$

The example exhibits a strong hierarchical structure. At the top level we have a tri-diagonal or circulant block structure, whereby each of the component blocks again has a tri-diagonal or circulant block structure of scalar entries. The overall resulting matrix is therefore very sparse with a sparsity pattern characterized by small bunches of non-diagonals clustered in bands. Such a situation is typical for

3D systems in which there is only local interaction between the quantities (as is the case with a differential equation). The regularity produces a Toeplitz or at least a block-Toeplitz structure, but in the more general case the sparsity pattern keeps the same general structure in which many diagonals are zero, with big gaps between significant diagonals. It is those big gaps that make the elimination procedures tricky because of the systematic occurrence of fill ins in the gaps. In the next section we propose a strategy that consists in forcing only partial or approximated elimination steps so that an explosion of fill ins is avoided and replaced by approximations based on a small amount of data.

3. The basic procedure: decoupling

The preconditioners we propose in this paper are based on partitioning the set of equations and decoupling them by estimating (approximating rather than calculating) the perturbation one set exerts on the other. This approach is somewhat similar to what has been termed incomplete LU factorization in the literature [9]. The difference with this traditional ad hoc approach is in how the perturbation is gauged. An efficient realization of the perturbed matrix (actually a Schur complement) is the key in the reduced modeling. In this section we review the basis for the decoupling strategy and introduce some notation that will allow for hierarchical recursion of the procedure.

Assume that we split the set of unknowns $u \in \mathbf{V}$ into two nonintersecting subsets $u_1 \in \mathbf{V}^{(1)}$ of size n_1 and $u_2 \in \mathbf{V}^{(2)}$ of size n_2, $\mathbf{V}^{(1)} \cap \mathbf{V}^{(2)} = \emptyset$ and $n = n_1 + n_2$, as

$$u = \begin{pmatrix} u_1 \\ u_2 \end{pmatrix}.$$

This splitting induces in a natural way a 2-by-2 block splitting of the matrix Φ,

$$\Phi = \begin{pmatrix} \Phi_{11} & \Phi_{12} \\ \Phi_{21} & \Phi_{22} \end{pmatrix}.$$

Then the matrix can be decomposed into a two level structure by a block LU factorization,

$$\begin{pmatrix} \Phi_{11} & \Phi_{12} \\ \Phi_{21} & \Phi_{22} \end{pmatrix} = \begin{pmatrix} I & 0 \\ \Phi_{21}\Phi_{11}^{-1} & I \end{pmatrix} \begin{pmatrix} \Phi_{11} & \Phi_{12} \\ 0 & S \end{pmatrix}$$

where I and 0 are generic identity and zero matrices of appropriate dimensions and

$$S = \Phi_{22} - \Phi_{21}\Phi_{11}^{-1}\Phi_{12}$$

is the Schur complement of Φ_{11} in Φ (as stated already, we assume existence of all relevant Schur complements).

Suppose the right-hand side vector b partitioned as above then the linear system decouples in two systems of reduced dimensions

$$\begin{aligned} \Phi_{11}\ u_1' &= b_1' \\ S\ u_2' &= b_2' \end{aligned} \tag{9}$$

with $b_1' = b_1$, $b_2' = b_2 - \Phi_{22}\Phi_{11}^{-1}b_1$, $u_1 = u_1' - \Phi_{11}^{-1}\Phi_{12}u_2'$ and $u_2 = u_2'$.

Our strategy for preconditioning consists in setting up a recursive schema of partitioning the variables and then decoupling the respective linear systems, and we do this not only at the top level of the hierarchy (as is discussed here), but recursively at lower levels as well. At each step in the procedure we approximate (or if one wishes, model reduce) the Schur complement systematically. The motivation for this is that the determination of the Schur complement is the step in the procedure where the fill ins are produced and the model complexity of the system hence increases. In many cases (and in particular the model case we are considering) approximating at this point is both physically and numerically justifiable, provided the partitioning is done in a justifiable way.

The recursive procedure can be set up in either a linear or a hierarchical manner. The linear recursion is of course the same as in the common LU factorization. In the block tri-diagonal case it reduces to a recursive determination of Schur complements, e.g., in the kth step written as

$$\begin{cases} S_0 = M_0, \\ S_{k+1} = M_{k+1} - L_k S_k^{-1} L_k^H. \end{cases} \tag{10}$$

In our model case, the recursion starts out with a block tridiagonal matrix. In the 2D case each of these blocks is again a tri-diagonal matrix. After the first step, the Schur complement then already has nine diagonals and at every step the number more than doubles, filling up the matrix quickly. It is not difficult to show that also the more general 'degree of semi-separability' [4] increases at the same rate, but at the same time it can be shown that there is a system with a low degree of semi-separability close by in operator norm. It is this model reduction that allows the determination of a low complexity approximant (in the semi-separable sense) in the 2D case. In the 3D case, however, one more level of hierarchy has to be dealt with – we discuss how to do this further on.

As discussed in the previous paragraph, a partitioning of the network (data and unknowns) leads to decoupling. This procedure can of course be repeated on each of the two sets, and then again, leading to a hierarchical decomposition tree representing the partitioning (still at this top level of the original hierarchy). Attached to each node of the tree there is the decoupled system of equations (and, of course, the corresponding primed and unprimed data sets which can be converted to each other according to the elimination formulas of the previous section). We use a level ordering notation as in the papers on 'HSS = Hierarchical Semi Separable' decompositions: the ordered index pair (k, ℓ) indicates node ℓ at level k ($\ell \in (1 \cdots 2^k)$). Node (k, ℓ), if it is not a leaf node, gets decomposed in two

nodes $(k+1, 2\ell-1)$ and $(k+1, 2\ell)$. To such a level decomposition there is a four block decomposition of the system attached to the node being decomposed. The decoupled system attached to the uneven child node is the 11 block of the parent system, while the system attached to the even child node is the Schur complement of that 11 block within the system defined by the parent node. In the sequel we shall mark the (eventually approximate) Schur complements with the index pairs indicating the level at which they define the decoupled system.

Because of the block triangular structure of the original system (and eventual semi separable generalizations thereof not considered here) there is a further hierarchical relation between Schur complements at various levels of the hierarchy.

Let Φ_α be a block triangular matrix at any given level $\alpha < \log m$ higher than the bottom level, dropping the index α for a simple notation and applying the two-by-two LU factorization on Φ we obtain Φ_{11} and Φ_{22} as block triangular matrices and Φ_{12} and Φ_{21} as low rank matrices with only one block at the left lower corner and the right upper corner respectively.

$$\Phi = \begin{pmatrix} \Phi_{11} & \Phi_{12} \\ \Phi_{21} & \Phi_{22} \end{pmatrix}.$$

Factorizing the matrix one more level we get:

$$\Phi = \begin{pmatrix} \Phi_{11_{11}} & \Phi_{11_{12}} & 0 & 0 \\ \Phi_{11_{21}} & \Phi_{11_{22}} & \Phi_{12_{21}} & 0 \\ 0 & \Phi_{21_{12}} & \Phi_{22_{11}} & \Phi_{22_{12}} \\ 0 & 0 & \Phi_{22_{21}} & \Phi_{22_{22}} \end{pmatrix}$$

with Φ_{11} and Φ_{22} the matrix decomposition for the next level. Then the Schur complement S of matrix Φ_{11} in Φ is

$$\begin{aligned} S &= \Phi_{22} - \Phi_{21}\Phi_{11}^{-1}\Phi_{12} \\ &= \begin{pmatrix} \Phi_{22_{11}} - \Phi_{21_{12}}(\Phi_{11}^{-1})_{22}\Phi_{12_{21}} & \Phi_{22_{12}} \\ \Phi_{22_{21}} & \Phi_{22_{22}} \end{pmatrix}. \end{aligned}$$

Let us put temporarily

$$E = \Phi_{11}$$

then with the two-by-two factorization block structure we find

$$E^{-1} = \begin{pmatrix} E_{11}^{-1} + E_{11}^{-1}E_{12}S_E^{-1}E_{21}E_{11}^{-1} & -E_{11}^{-1}E_{21}S_E^{-1} \\ -S_E^{-1}E_{21}E_{11}^{-1} & S_E^{-1} \end{pmatrix},$$

which shows that

$$(E^{-1})_{22} = S_E^{-1}.$$

We substitute

$$(\Phi_{11}^{-1})_{22} = S_{\Phi_{11}}^{-1}$$

back, where $S_{\Phi_{11}}$ is the Schur complement of block $\Phi_{11_{11}}$ in Φ_{11}. Therefore the Schur complement S of Φ_{11} in Φ becomes:

$$S = \begin{pmatrix} \Phi_{22_{11}} - \Phi_{21_{12}}S_{\Phi_{11}}^{-1}\Phi_{12_{21}} & \Phi_{22_{12}} \\ \Phi_{22_{21}} & \Phi_{22_{22}} \end{pmatrix}$$

in which the (11)-entry is actually the Schur complement of S_E in the submatrix

$$\begin{pmatrix} S_E & \Phi_{12_{21}} \\ \Phi_{21_{12}} & \Phi_{22_{11}} \end{pmatrix}.$$

Hence, the higher level Schur complement is constituted of lower level Schur complements and other lower level matrices.

In this fashion, the causality relations get to be very simple when the recursion is spun out to the bottom decomposition level, no calculations are needed to move to higher levels in the tree, only assembly of submatrices. In an exact calculation, a chain of Schur complements only involving local matrices is obtained at the bottom level of the hierarchy – see Fig. 3 – typically the level of the size of the block entries in the original tri-diagonal matrix, but any higher level may serve as bottom level just as well. The recursion starts out with a tridiagonal matrix and then doubles in theoretical semi-separable complexity at each step. Keeping this complexity increase under control is the key to the systematic construction of preconditioners based on Schur complementation. That is the topic of Section 7.

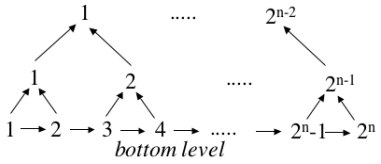

FIGURE 1. Causality relations for the Schur elimination schema. Only the bottom line requires computations, all the upward arrows only involve assembly of matrices.

4. Algorithms for sequentially semi-separable matrices

In this section we treat the case in which the Schur complement is approximated by a low-order semi-separable representation. This would be the method to be followed in the 2D case, or at the bottom level of the hierarchy in the 3D case. In [4, 7] it is shown that subsequent Schur complements occurring in the solution of the regular 2D Poisson problem are close to a low degree semi-separable matrix. In particular, in [7] the convergence in terms of the ϵ-rank of Hankel blocks is shown to be bounded with some low bound. These bounds are confirmed by experiments, which actually exhibit very close approximation even at low semi separable degree. The precise results for the regular Poisson equation are of course due to the fact that in this case the system can be solved in closed form. We start out with a brief summary of the basic properties of semi-separable representations we use. For a comprehensive treatment of the basics we refer to [5].

Matrices that have the following structure, defined through a set of small matrices $\{A_k, B_k, C_k, D_k, A'_k, B'_k, C'_k\}$

$$\begin{pmatrix} D_1 & B_1C_2 & B_1A_2C_3 & \dots & B_1A_2\dots A_{n-1}C_n \\ B'_2C'_1 & D_2 & B_2C_3 & \dots & B_2A_3\dots A_{n-1}C_n \\ B'_3A'_2C'_1 & B'_3C'_2 & D_3 & \dots & B_3A_4\dots A_{n-1}C_n \\ & & \ddots & \ddots & \ddots \\ & & & B'_nC'_{n-1} & D_n \end{pmatrix}$$

are called matrices with sequentially semi-separable structure, and the sequence of matrices is called the state realization of the sequentially semi-separable structure of the matrix [5]. Let T be such a matrix, then the realization matrices correspond to a computational schema for the input-output product $y = uT$ involving a set of intermediate so-called state vectors $\{x_k, x'_k\}$ that are computed recursively

$$\begin{cases} x_{k+1} &= x_kA_k + u_kB_k \\ x'_{k-1} &= x'_kA'_k + u_kB'_k \\ y_k &= x_kC_k + x'_kC'_k + u_kD_k. \end{cases}$$

Rewritten in global operator form by assembling the matrices A_k, B_k etc... as diagonal operators on spaces of sequences of appropriate dimensions

$$A = \begin{bmatrix} \ddots & & 0 \\ & A_k & \\ 0 & & \ddots \end{bmatrix} B = \begin{bmatrix} \ddots & & 0 \\ & B_k & \\ 0 & & \ddots \end{bmatrix}$$

etc... and defining the shift-operator Z as $(uZ)_i = u_{i-1}$ we obtain a compact representation of T in terms of its structural matrices as

$$T = D + BZ(I - AZ)^{-1}C + B'Z^{-1}(I - A'Z^{-1})^{-1}C'.$$

Of course, all dimensions of matrices and vectors have to match wherever needed. The structural matrices are often brought together in view of this as

$$\mathbf{T}_c = \begin{bmatrix} A & C \\ B & D \end{bmatrix}, \ \mathbf{T}_a = \begin{bmatrix} A' & C' \\ B' & 0 \end{bmatrix}, \tag{11}$$

which is a 4×4 block matrix with diagonal entries. We now briefly discuss matrix operations using the semi separable structure. We first concentrate on upper triangular matrices for which the accented quantities are zero. Let us, for convenience, define the diagonal shift operator $T^{(1)}$ by

$$Z_M T^{(1)} = TZ_N$$

that is, $T^{(1)} = Z^{-1}TZ$, then $T^{(1)}$ is the operator T whose representation is shifted one position into the South-East direction: $(T^{(1)})_{i,j} = T_{i-1,j-1}$. More generally,

the kth diagonal shift of T into the southeast direction along the diagonals of T is defined by

$$T^{(k)} = (Z^k)^{-1} T Z^k.$$

Equivalently, $(T^{(k)})_{i,j} = T_{i-k,j-k}$.

4.1. State transformations

Two realizations $\{A_1, B_1, C_1, D\}$ and $\{A_2, B_2, C_2, D\}$ are called equivalent if their respective state vectors are related through an invertible transformation R. We have then

$$\begin{bmatrix} A_2 & C_2 \\ B_2 & D \end{bmatrix} = \begin{bmatrix} R & \\ & I \end{bmatrix} \begin{bmatrix} A_1 & C_1 \\ B_1 & D \end{bmatrix} \begin{bmatrix} [R^{(-1)}]^{-1} & \\ & I \end{bmatrix}$$
$$R^{(-1)} = ZRZ^{-1}.$$

We say that a realization is minimal if none of the dimensions of the state vectors can be reduced further. It is known [5] that these minimal dimensions form a unique sequence and that two minimal realizations are related through an invertible transformation matrix.

4.2. Sum of two realizations

let T_1, T_2 be two upper triangular matrices, with realizations A_1, B_1, C_1, D_1 and A_2, B_2, C_2, D_2, respectively. Then the sum of these two operators, $T = T_1 + T_2$, has a realization given directly in terms of these two realizations as

$$\begin{bmatrix} A & C \\ B & D \end{bmatrix} = \left[\begin{array}{cc|c} A_1 & 0 & C_1 \\ 0 & A_2 & C_2 \\ \hline B_1 & B_2 & D_1 + D_2 \end{array}\right].$$

The state dimension sequence of this realization is equal to the sum of the state dimension sequences of T_1 and T_2. It is, however, not necessarily minimal even if the component dimensions are.

4.3. Product of two realizations

The product of $T = T_1 T_2$ can also be obtained using realizations by

$$\begin{bmatrix} A & C \\ B & D \end{bmatrix} = \left[\begin{array}{cc|c} A_1 & C_1 B_2 & C_1 D_2 \\ 0 & A_2 & C_2 \\ \hline B_1 & D_1 B_2 & D_1 D_2 \end{array}\right].$$

In this case also the dimension of the given product realization is the sum of the dimensions of the components. It is not necessarily minimal even though the realizations of the factors are – there may be cancellations between the factors.

4.4. Realization of an upper inverse

Let T be an invertible upper triangular matrix, and suppose that it is known that T^{-1} is also upper, then the D matrix in the realization has to be square invertible

and a realization for T^{-1} is given by [5]

$$\begin{bmatrix} A' & C' \\ B' & D' \end{bmatrix} = \begin{bmatrix} A - CD^{-1}B & -CD^{-1} \\ D^{-1}B & D^{-1} \end{bmatrix}.$$

This realization for T^{-1} will be minimal if the realization for T is.

4.5. Cholesky factorization

We now return to the mixed upper-lower case. Given $T > 0$ and let the upper triangular part of T have a minimal state space realization A_k, B_k, C_k, D_k. Let $T = F^H F$ where F is upper triangular. The main property of relevance here is that the upper factor F has a minimal state space realization of the same dimensions as the upper part of T. It is given by
Then a realization $A_{F,k}, B_{F,k}, C_{F,k}, D_{F,k}$ of F is given by (superscript $\cdot^H$ indicates Hermitian conjugation)

$$\begin{cases} A_{F,k} = A_k \\ C_{F,k} = C_k \\ D_{F,k} = (D_k - C_k^H \Lambda_k C_k)^{-1/2} \\ B_{F,k} = D_{F,k}^{-1}(B_k - C_k^H \Lambda_k A_k) \end{cases}$$

where Λ_k is given by the recursion

$$\Lambda_{k+1} = A_k^H \Lambda_k A_k + B_{F,k}^H B_{F,k}.$$

T has to be positive definite for this recursion to work out. In case that turns out not to be so, then at a certain point k in the recursion $D_k - C_k^H \Lambda_k C_k$ will turn out to be non-positive definite, leading to a non positive square root.

5. Efficient Schur reduction: the semi-separable case

In this section we consider the rather more general case where the matrix to be reduced has the form

$$\Phi = \begin{pmatrix} M_1 & -L_1^H & & & \\ -L_1 & M_2 & -L_1^H & & \\ & \ddots & \ddots & \ddots & \\ & & \ddots & M_{m-1} & -L_{m-1}^H \\ & & & -L_{m-1} & M_m \end{pmatrix} \tag{12}$$

5.1. State space realizations of matrices M and L

Φ consists of block matrices M_k and L_k. Dropping the index k whenever clear from the context, we assume further that the matrices M and L have general

sequentially semi-separable realizations of the type

$$\begin{aligned} M_k &= \{A_{M_c}, A_{M_a}, B_{M_c}, B_{M_a}, C_{M_c}, C_{M_a}, D_M\} \\ &= D_M + B_{M_c} Z (I - A_{M_c} Z)^{-1} C_{M_c} \\ &\quad + B_{M_a} Z^{-1} (I - A_{M_a} Z^{-1})^{-1} C_{M_a} \\ L_k &= \{A_{L_c}, A_{L_a}, B_{L_c}, B_{L_a}, C_{L_c}, C_{L_a}, D_L\} \\ &= D_L + B_{L_c} Z (I - A_{L_c} Z)^{-1} C_{L_c} \\ &\quad + B_{L_a} Z^{-1} (I - A_{L_a} Z^{-1})^{-1} C_{L_a}. \end{aligned}$$

The realization of a sequentially semi-separable structure matrix can be computed by a low rank factorizations of some off-diagonal blocks called Hankel blocks, as described in [8]. In the tri-diagonal case the derivation is trivial.

5.2. Schur complements in the state space formalism

Because of the Hermitian structure of Φ, M and L, we may assume the realizations of M and L to have

$$\begin{aligned} A_{M_a} &= A_{M_c}^H, & & & A_{L_a} &= A_{L_c}^H, \\ B_{M_a} &= C_{M_c}^H, & \text{and} & & B_{L_a} &= C_{L_c}^H, \\ C_{M_a} &= B_{M_c}^H, & & & C_{L_a} &= B_{L_c}^H. \end{aligned}$$

From equation (10), each S_k will be Hermitian matrix as well, and hence can be Cholesky factorized as $S_k = F_k^H F_k$.

Assuming that the upper triangular matrix F_k has a realization

$$\{A_{Fc}, B_{Fc}, C_{Fc}, D_F\},$$

we now try to find the realization of upper triangular matrix F_{k+1} using only state space data.

$$\begin{aligned} S_{k+1} &= F_{k+1}^H F_{k+1} \\ &= M_{k+1} - L_k (F_k^H F_k)^{-1} L_k^H \\ &= M_{k+1} - L_k F_k^{-1} F_k^{-H} L_k^H. \end{aligned}$$

Let $E_k = F_k^{-1}$ and let us factorize L_k to be $L_k = L_{1_k} L_{2_k}$ and let the Cholesky factorization of M_k be $M_k = X_k X_k^H$, with E_k and L_{2_k} upper triangular matrices and L_{1_k} and X_k lower triangular. We get

$$F_{k+1}^H F_{k+1} = X_{k+1} X_{k+1}^H - L_{1_k} L_{2_k} E_k E_k^H L_{2_k}^H L_{1_k}^H$$

Let $G_k = L_{2_k} E_k$, which is an upper triangular matrix. To make the computation simple, we convert each G_k matrix to a lower triangular matrix H_k, where $G_k G_k^H = H_k H_k^H$. We get

$$\begin{aligned} F_{k+1}^H F_{k+1} &= X_{k+1} X_{k+1}^H - L_{1_k} G_k G_k^H L_{1_k}^H \\ &= X_{k+1} X_{k+1}^H - L_{1_k} H_k H_k^H L_{1_k}^H. \end{aligned}$$

Let $Y_k = L_{1_k} H_k$ (lower triangular), then

$$F_{k+1}^H F_{k+1} = X_{k+1} X_{k+1}^H - Y_k Y_k^H.$$

Dropping the index k in the following sections for simplification, and assuming the realization of any upper matrix O to be

$$\begin{aligned} O &= \{A_O, B_O, C_O, D_O\} \\ &= D_O + B_O Z (I - A_O Z)^{-1} C_O \end{aligned}$$

and the realization of any lower matrix P

$$\begin{aligned} P &= \{A_P, B_P, C_P, D_P\} \\ &= D_P + B_P Z^{-1} (I - A_P Z^{-1})^{-1} C_P. \end{aligned}$$

The following subsections explain how to get the state space of

$$F_{k+1} = \{A'_F, B'_F, C'_F, D'_F\} \quad \text{from} \quad F_k = \{A_F, B_F, C_F, D_F\}$$

using the basic steps explained above.

5.2.1. From F_{k-1} to E_k. As F_{k-1} is an upper triangular matrix, we can get the realization of $E_k = F_{k-1}^{-1}$ directly by time varying system theory as:

$$\begin{aligned} A_E &= A_F - C_F D_F^{-1} B_F \\ B_E &= D_F^{-1} B_F \\ C_E &= -C_F D_F^{-1} \\ D_E &= D_F^{-1}. \end{aligned}$$

5.2.2. Factorize L_k to L_{1_k} and L_{2_k}. where each L_{1_k} is an lower triangular matrix and L_{2_k} is a upper triangular matrix.

$$\begin{aligned} A_{l_1} &= A_{L_a}; \\ A_{l_2} &= A_{L_c}; \\ B_{l_1} &= B_{L_a}; \\ C_{l_2} &= C_{L_c}; \\ D_{l_2} &= I; \\ D_{l_1} &= (D_L - B_{L_a} \Lambda C_{L_c}) D_{l_2}^{-1} \\ B_{l_2} &= D_{l_1}^{-1} (B_{L_c} - B_{L_a} \Lambda A_{L_c}) \\ C_{l_1} &= (C_{L_a} - A_{L_a} \Lambda C_{L_c}) D_{l_2}^{-1} \\ \Lambda^{(1)} &= C_{l_1} B_{l_2} + A_{l_1} \Lambda A_{l_2} \end{aligned}$$

where $\Lambda^{(1)}$ stands for Λ_{k+1} and $\Lambda^{(-1)}$ stands for Λ_{k-1}.

5.2.3. $G_k = L_{2_k} E_k$**.** Both L_{2_k} and E_k are upper triangular matrices, so does G_k.

$$\begin{aligned} A_G &= \begin{bmatrix} A_{l_2} & C_{l_2} B_E \\ 0 & A_E \end{bmatrix} \\ B_G &= \begin{bmatrix} B_{l_2} & D_{l_2} B_E \end{bmatrix} \\ C_G &= \begin{bmatrix} C_{l_2} D_E \\ C_E \end{bmatrix} \\ D_G &= D_{l_2} D_E. \end{aligned}$$

5.2.4. From G_k **to** H_k**.** G_k are upper triangular matrices while H_k are lower triangular matrices with $G_k G_k^H = H_k H_k^H$.

$$\begin{aligned} A_H &= A_G^H; \\ B_H &= D_G C_G^H + B_G \Lambda A_G^H; \\ D_H &= (D_G D_G^H + B_G \Lambda B_G^H - B_H \Gamma B_H^H)^{1/2} \\ C_H &= (B_G^H - A_G^H \Gamma B_H^H) D_H^{-1} \end{aligned}$$

where

$$\begin{aligned} \Lambda^{(-1)} &= C_G C_G^H + A_G \Lambda A_G^H \\ \Gamma^{(1)} &= C_H C_H^H + A_H \Gamma A_H^H. \end{aligned}$$

5.2.5. $Y_k = L_{1_k} H_k$**.** Y_k, L_{1_k}, and H_k are all lower triangular matrices,

$$\begin{aligned} A_Y &= \begin{bmatrix} A_H & 0 \\ C_{l_1} B_H & A_{l_1} \end{bmatrix} \\ B_Y &= \begin{bmatrix} D_{l_1} B_H & B_{l_1} \end{bmatrix} \\ C_Y &= \begin{bmatrix} C_H \\ C_{l_1} D_H \end{bmatrix} \\ D_Y &= D_{l_1} D_H. \end{aligned}$$

5.2.6. Factorize $M_k = X_k X_k^H$**.** We can get lower triangular matrix X_k from M_k by Cholesky factorization again.

$$\begin{aligned} A_X &= A_M^H \\ B_X &= C_M^H \\ D_X &= (D_M - C_M^H \Gamma C_M)^{1/2} \\ C_X &= (B_M^H - A_M^H \Gamma C_M) D_X^{-1} \end{aligned}$$

where

$$\Gamma^{(1)} = C_X C_X^H + A_M^H \Gamma A_M.$$

5.2.7. Get F_{k+1} from $F_{k+1}^H F_{k+1} = X_{k+1} X_{k+1}^H - Y_k Y_k^H$. Compute lower triangular matrix $W_k = X_k^{-1} Y_k$,

$$\begin{aligned}
A_{X^{-1}} &= A_X - C_X D_X^{-1} B_X \\
B_{X^{-1}} &= -D_X^{-1} B_X \\
C_{X^{-1}} &= C_X D_X^{-1} \\
D_{X^{-1}} &= D_X^{-1}
\end{aligned}$$

$$\begin{aligned}
A_W &= \begin{bmatrix} A_Y & 0 \\ C_{X^{-1}} B_Y & A_{X^{-1}} \end{bmatrix} \\
B_W &= \begin{bmatrix} D_{X^{-1}} B_Y & B_{X^{-1}} \end{bmatrix} \\
C_W &= \begin{bmatrix} C_Y \\ C_{X^{-1}} D_Y \end{bmatrix} \\
D_W &= D_{X^{-1}} D_Y .
\end{aligned}$$

Let lower triangular matrix Δ satisfy $\Delta\Delta^H = I - WW^H$, and after Cholesky factorization, we obtain

$$\begin{aligned}
A_\Delta &= A_W \\
B_\Delta &= -B_W \\
D_\Delta &= I - D_W D_W^H - B_W \Lambda B_W^H - B_W \Gamma B_W^H)^{1/2} \\
C_\Delta &= (C_W D_W^H + A_W \Lambda B_W^H + A_W \Gamma B_W^H) D_\Delta^{-1}
\end{aligned}$$

where

$$\begin{aligned}
\Lambda^{(1)} &= C_W C_W^H + A_W \Lambda A_W^H \\
\Gamma^{(1)} &= C_\Delta C_\Delta^H + A_W \Gamma A_W^H .
\end{aligned}$$

Multiply $F = X\Delta$. We finally get the lower triangular matrix F'_{k+1} and upper triangular matrix $F_{k+1} = F'_{k+1}$ as

$$\begin{aligned}
A_F &= \begin{bmatrix} A_\Delta^H & B_\Delta^H C_X^H \\ 0 & A_X^H \end{bmatrix} \\
B_F &= \begin{bmatrix} C_\Delta^H & D_\Delta^H C_X^H \end{bmatrix} \\
C_F &= \begin{bmatrix} B_\Delta^H D_X^H \\ B_X^H \end{bmatrix} \\
D_F &= D_\Delta^H D_X^H .
\end{aligned}$$

5.3. Model reduction

The resulting Schur complement factor F_k at each step in the recursion can be model reduced by reduction on both reachability Gramian and observability Gramian in the state space. Let us assume that the realization of F_k obtained in the recursion step is $\{A, B, C, D\}$, and do the model reduction on reachability

Gramian Λ_c first, which results the reduced realization $\{A'_{11}, B'_1, C'_1, D'\}$. Then we reduce it again on observability Gramian Λ_o on $\{A'_{11}, B'_1, C'_1, D'\}$, resulting the final reduced realization $\{A''_{11}, B''_1, C''_1, D''\}$. More details can be found in [2].

5.3.1. On reachability Gramians. Let Λ_c be the reachability Gramian of the given realization $\{A, B, C, D\}$ of T, then it has the eigenvalue decomposition:

$$\Lambda_c = R_c^H \begin{bmatrix} \widetilde{\Lambda}_c & 0 \\ 0 & 0 \end{bmatrix} R_c, \quad R_c = \begin{bmatrix} \widetilde{R}_c \\ * \end{bmatrix},$$

where $\widetilde{\Lambda}_c$ is a diagonal matrix containing the nonzero eigenvalues of Λ_c, while R_c can be chosen unitary, and $\widetilde{R}_c$ contains the columns of R_c corresponding to the entries in $\widetilde{\Lambda}_c$. Apply R_c as a state transformation to T, we get an equivalent realization $T' = \{A', B', C', D'\}$,

$$\begin{aligned} \begin{bmatrix} A' & C' \\ B' & D' \end{bmatrix} &= \begin{bmatrix} R_c & 0 \\ 0 & I \end{bmatrix} \begin{bmatrix} A & C \\ B & D \end{bmatrix} \begin{bmatrix} R_c^{(-1)H} & 0 \\ 0 & I \end{bmatrix} \\ &= \begin{bmatrix} R_c A R_c^{(-1)H} & R_c C \\ B R_c^{(-1)H} & D \end{bmatrix} \\ \Lambda'_c &= \begin{bmatrix} \widetilde{\Lambda}_c & 0 \\ 0 & 0 \end{bmatrix} \end{aligned}$$

where Λ'_c is the reachability Gramian of T', and satisfies the Lyapunov equation $\Lambda_c'^{(-1)} = A'^H \Lambda'_c A' + B'^H B'$. Partition A', B', and C' according to that of R_c,

$$\begin{aligned} A' &= \begin{bmatrix} A'_{11} & 0 \\ A'_{21} & A'_{22} \end{bmatrix}, \quad C' = \begin{bmatrix} C'_1 \\ C'_2 \end{bmatrix}, \\ B' &= \begin{bmatrix} B'_1 & 0 \end{bmatrix}, \qquad D' = D. \end{aligned}$$

Because $B_2'^H B'_2 + A_{12}'^H \Lambda_{11} A'_{12} = 0$ and $\Lambda_{11} > 0$, we have $B_2 = A_{12} = 0$.

Then the model reduced realization is

$$\begin{bmatrix} A'_{11} & C'_1 \\ B'_1 & D' \end{bmatrix} = \begin{bmatrix} \widetilde{R}_c A \widetilde{R}_c^{(-1)H} & \widetilde{R}_c C \\ B \widetilde{R}_c^{(-1)H} & D \end{bmatrix}.$$

5.3.2. On observability Gramians. Similarly, the observability Gramian Λ_o of realization $\{A'_{11}, B'_1, C'_1, D'\}$ can be decomposed as:

$$\Lambda_o = R_o^{-1} \begin{bmatrix} \widetilde{\Lambda}_o & 0 \\ 0 & 0 \end{bmatrix} R_o^{-H}, \quad R_o = \begin{bmatrix} \widetilde{R}_o & * \end{bmatrix},$$

Apply the unitary matrix R_o to the realization $\{A'_{11}, B'_1, C'_1, D'\}$, we get realization,

$$\begin{aligned}
\begin{bmatrix} A'' & C'' \\ B'' & D'' \end{bmatrix} &= \begin{bmatrix} R_o & 0 \\ 0 & I \end{bmatrix} \begin{bmatrix} A'_{11} & C'_1 \\ B'_1 & D' \end{bmatrix} \begin{bmatrix} R_o^{(-1)H} & 0 \\ 0 & I \end{bmatrix} \\
&= \begin{bmatrix} R_o A'_{11} R_o^{(-1)H} & R_o C'_1 \\ B'_1 R_o^{(-1)H} & D' \end{bmatrix} \\
\Lambda''_o &= \begin{bmatrix} \widetilde{\Lambda}_o & 0 \\ 0 & 0 \end{bmatrix}
\end{aligned}$$

and

$$\begin{aligned}
A'' &= \begin{bmatrix} A''_{11} & A''_{12} \\ 0 & A''_{22} \end{bmatrix}, \quad C'' = \begin{bmatrix} C''_1 \\ 0 \end{bmatrix}, \\
B'' &= \begin{bmatrix} B''_1 & B''_2 \end{bmatrix}, \quad D'' = D.
\end{aligned}$$

So the final reduced realization is

$$\begin{bmatrix} A''_{11} & C''_1 \\ B''_1 & D'' \end{bmatrix} = \begin{bmatrix} \widetilde{R}_o A''_{11} \widetilde{R}_o^{(-1)H} & \widetilde{R}_o C''_1 \\ B''_1 \widetilde{R}_o^{(-1)H} & D'' \end{bmatrix}.$$

6. Exact solutions for the 3D model cases

In this section we present the exact solutions (in closed form) for the two regular 3D-cases announced earlier. The method we use extends the method presented in [4, 7] to the 3D case. In particular, we compute the exact value of the Schur complements and their limiting fixed point. In the next section we shall use these results to evaluate pre-conditioners in a number of situations.

Reverting back to the notation of Section 2 we have that the sequence of Schur complements from the top-left subblock to the down-right subblock is given by equation (10).

Because of the positive definiteness of the Poisson matrix, all subsequent Schur complements are positive definite as well, and the recursion will converge, as we shall see, to a fixed point matrix S_∞, which we shall evaluate. It turns out that the latter is actually a very good approximant of the actual Schur complements for larger values of k, the convergence being pretty fast.

6.1. Block symmetric tri-diagonal Toeplitz matrix

In this section we consider only real symmetric (block-) tri-diagonal Toeplitz matrices, in which the blocks themselves are real symmetric (block) tri-diagonal and

Toeplitz. For brevity, we introduce the shorthand tridiag(A, B) for the matrix

$$\text{tridiag}(A,B) = \begin{pmatrix} A & B^H & & & \\ B & A & B^H & & \\ & B & A & \ddots & \\ & & \ddots & \ddots & B \\ & & & B & A \end{pmatrix}. \tag{13}$$

In this hierarchy we indicate the depth d by a super-index, and the corresponding size of the matrices by n_d. At level 0 we then have

$$A^{(0)} = \text{tridiag}(A^{(1)}, B^{(1)}). \tag{14}$$

The hierarchy terminates when a certain depth $d = D - 1$ is reached, where each of the sub-blocks is a symmetric tri-diagonal Toeplitz matrix having the form

$$A^{(D-1)} = \text{tridiag}(a, b) \tag{15}$$

with scalar entries.

Theorem 1. *Given a block real symmetric tri-diagonal Toeplitz matrices $T = tridiag(A, B)$ at a hierarchical depth d with n_d number of sub-blocks on its diagonal, and let the $d-1$ depth sub-block matrices A and B have eigenvalues $\Lambda_A = \text{diag}(\lambda_{A;1}, \ldots, \lambda_{A;p})$ and $\Lambda_B = \text{diag}(\lambda_{B;1}, \ldots, \lambda_{B;p})$ respectively, with dimension $p \times p$ and the same normalized eigenvector matrix U, then the matrix T has eigenvalues $\Lambda_T = \Lambda_A + 2\Lambda_B \cos\frac{k\pi}{n_d+1}$, where $1 \le k \le n_d$ with corresponding normalized block eigenvectors given by*

$$V_k = \begin{pmatrix} U\sin(\frac{1k\pi}{n_d+1})/\sigma \\ U\sin(\frac{2k\pi}{n_d+1})/\sigma \\ \vdots \\ U\sin(\frac{n_dk\pi}{n_d+1})/\sigma \end{pmatrix} \qquad \textit{where} \qquad \sigma = \sqrt{\sum_{i=1}^{n_d} \sin^2\left(\frac{ik\pi}{n_d+1}\right)}.$$

The theorem generalizes a result of [6] about tri-diagonal Toeplitz matrices reproduced hereunder.

Theorem 2. *If C is an $n \times n$ tri-diagonal Toeplitz matrix with*

$$C = \begin{pmatrix} a & b & & & \\ c & a & b & & \\ & c & a & \ddots & \\ & & \ddots & \ddots & b \\ & & & c & a \end{pmatrix} \tag{16}$$

then the eigenvalues of C are given by

$$\lambda_j = a + 2b\sqrt{\frac{c}{b}}\cos(\frac{j\pi}{n+1}), \tag{17}$$

where $1 \leq j \leq n$, *and corresponding eigenvectors are given by*

$$x_j = \begin{pmatrix} (\frac{c}{b})^{1/2} \sin(\frac{1j\pi}{n+1}) \\ (\frac{c}{b})^{2/2} \sin(\frac{2j\pi}{n+1}) \\ (\frac{c}{b})^{3/2} \sin(\frac{3j\pi}{n+1}) \\ \vdots \\ (\frac{c}{b})^{n/2} \sin(\frac{nj\pi}{n+1}) \end{pmatrix}. \tag{18}$$

Proof of Theorem 1. The theorem follows from the fact that all matrices involved commute. This is true at the bottom level $D-1$ of the hierarchy since at that level the off-diagonal entries of the matrices are equal resulting in eigenvectors that can be chosen equal. Moving up the hierarchy, the constitutive block-matrices keep commuting. We may assume recursively that A and B are diagonalized by the same eigenvector matrix U,

$$A = U\Lambda_A U^H$$
$$B = U\Lambda_B U^H$$

then

$$T = \begin{pmatrix} U & & & \\ & U & & \\ & & \ddots & \\ & & & U \end{pmatrix} \begin{pmatrix} \Lambda_A & \Lambda_B & & \\ \Lambda_B & \Lambda_A & & \\ & & \ddots & \Lambda_B \\ & & \Lambda_B & \Lambda_A \end{pmatrix} \begin{pmatrix} U^H & & & \\ & U^H & & \\ & & \ddots & \\ & & & U^H \end{pmatrix}.$$

Let us denote

$$T' = \begin{pmatrix} \Lambda_A & \Lambda_B & & \\ \Lambda_B & \Lambda_A & & \\ & & \ddots & \Lambda_B \\ & & \Lambda_B & \Lambda_A \end{pmatrix} \tag{19}$$

and apply Theorem 2 on T', (notice that after proper permutation P, $T" = PT'P$ is block diagonal matrix with each sub-block a symmetric tri-diagonal Toeplitz matrix,) we get the eigenvalues of T':

$$\Lambda_{T'_k} = \Lambda_A + 2\Lambda_B \cos\left(\frac{k\pi}{n_d + 1}\right), \tag{20}$$

where $1 \leq k \leq n_d$, and the corresponding eigenvector block is

$$V_j' = \begin{pmatrix} I_{p\times p} \sin(\frac{1k\pi}{n_d+1})/\sigma \\ I_{p\times p} \sin(\frac{2k\pi}{n_d+1})/\sigma \\ I_{p\times p} \sin(\frac{3k\pi}{n_d+1})/\sigma \\ \vdots \\ I_{p\times p} \sin(\frac{nk\pi}{n_d+1})/\sigma \end{pmatrix}. \tag{21}$$

where

$$\sigma = \sqrt{\sum_{i=1}^{n_d} \sin^2\left(\frac{ik\pi}{n_d+1}\right)}$$

$$T = \begin{pmatrix} U & & & \\ & U & & \\ & & \ddots & \\ & & & U \end{pmatrix} V' \Lambda_{T'} V'^H \begin{pmatrix} U^H & & & \\ & U^H & & \\ & & \ddots & \\ & & & U^H \end{pmatrix}.$$

Therefore, the eigenvalues and eigenvectors of T are: $\Lambda_T = \Lambda_A + 2\Lambda_B \cos\frac{k\pi}{n_d+1}$, where $1 \le k \le n_d$. The corresponding block eigenvector is

$$V_k = \begin{pmatrix} U\sin(\frac{1k\pi}{n_d+1})/\sigma \\ U\sin(\frac{2k\pi}{n_d+1})/\sigma \\ \vdots \\ U\sin(\frac{n_dk\pi}{n_d+1})/\sigma \end{pmatrix}.$$

The block eigen-structure can hence be applied hierarchically in this case leading to a diagonalized matrix at the top level of the hierarchy. From this it follows that subsequent Schur complements and the fixed point solution of the Schur-complement equation can be found explicitly. On the diagonalized matrix all the Schur complements are diagonal as well, as well as the fixed point solution. This is stated in the following (almost obvious) theorem.

Theorem 3. *Let Φ be given by* (2)*, and assume that for any block-dimension n, Φ is strictly positive definite. Then $\Lambda_M > 2\Lambda_L$ and the fixed point solution for the Schur complement S_∞ is given by*

$$S_\infty = \frac{1}{2} V \left(\Lambda_M + \sqrt{\Lambda_M^2 - 4\Lambda_L^2}\right) V^H. \tag{22}$$

where

$$\Lambda_{M_j} = \Lambda_O + 2\Lambda_P \cos\left(\frac{j\pi}{n+1}\right),$$

$$\Lambda_{L_j} = \Lambda_R + 2\Lambda_Q \cos\left(\frac{j\pi}{n+1}\right),$$

$$\lambda_{O_i} = a_O + 2b_O \cos\left(\frac{i\pi}{n+1}\right),$$

$$\lambda_{P_i} = a_P + 2b_P \cos\left(\frac{i\pi}{n+1}\right),$$

$$\lambda_{Q_i} = a_Q + 2b_Q \cos\left(\frac{i\pi}{n+1}\right),$$

$$\lambda_{R_i} = a_R + 2b_R \cos\left(\frac{i\pi}{n+1}\right),$$

$$v_j = \frac{1}{\sqrt{\sum_{k=1}^{n} \sin^2(\frac{kj\pi}{n+1})}} \begin{pmatrix} U\sin(\frac{1j\pi}{n+1}) \\ U\sin(\frac{2j\pi}{n+1}) \\ \vdots \\ U\sin(\frac{nj\pi}{n+1}) \end{pmatrix}.$$

$$u_i = \frac{1}{\sqrt{\sum_{k=1}^{n} \sin^2(\frac{ki\pi}{n+1})}} \begin{pmatrix} \sin(\frac{1i\pi}{n+1}) \\ \sin(\frac{2i\pi}{n+1}) \\ \vdots \\ \sin(\frac{ni\pi}{n+1}) \end{pmatrix}$$

and $1 \le j \le n, 1 \le i \le n$.

Proof. In this case the hierarchical decomposition is just two levels deep (corresponding to the 3D case) and the matrices involved can easily be written down explicitly. It is easy to tell from the theorem that M and L share the same eigenvector matrix V and $M = V\Lambda_M V^H$, $L = V\Lambda_L V^H$. Hence we have $S_0 = U\Lambda_0 U^H$ and $S_k = U\Lambda_k U^H$ for the same collection of normalized eigenvectors assembled in U. The Schur recursion then becomes a collection of scalar recursions given by

$$\begin{cases} \Lambda_0 = \Lambda_M, \\ \Lambda_{k+1} = \Lambda_M - \Lambda_L \Lambda_k^{-1} \Lambda_L^H \ (k = 0, 1, \dots). \end{cases}$$

The intermediate matrices at the hierarchical level 2 have a diagonalization represented by $\Phi' = \text{tridiag}(\Lambda_M, \Lambda_L)$ which transforms, by reordering of rows and columns into a direct sum of Toeplitz matrices of the form $\text{tridiag}(m_k, \ell_k)$. Since all these are strictly positive definite for any dimension by assumption, it must be that $m_k > 2\ell_k$, by a well-known property of Toeplitz matrices (corresponding to the fact that the limiting spectrum must be (strictly) positive definite – see [6]). The scalar iteration

$$x_{k+1} = m - \ell x_k^{-1} \ell \quad \text{with} \quad m > 2\ell > 0$$

converges to

$$x = \frac{1}{2}(m + \sqrt{m^2 - 4l^2}).$$

Since $\Lambda_M > 2\Lambda_L$, the iteration converges for each entry in the recursion, leading to the result claimed in the theorem. □

6.2. The block-circulant case

A theory similar to the block triangular Toeplitz case can be set up for (block-) circulant matrices. It covers the case of a regular grid on which periodic boundary conditions are in effect. The eigenvalue analysis in this case leads to Fourier transformation. We briefly summarize the results.

A block circulant matrix is completely determined by its first block row. This we denote as follows

$$\mathrm{circ}(A_k : k = 0 \cdots n-1) = \begin{pmatrix} A_0 & A_1 & A_2 & \dots & A_{n-1} \\ A_{n-1} & A_0 & A_1 & \dots & \vdots \\ A_{n-2} & A_{n-1} & A_0 & \dots & \vdots \\ \vdots & \dots & \dots & \ddots & A_1 \\ A_1 & A_2 & \dots & A_{n-1} & A_0 \end{pmatrix}.$$

We consider a hierarchy of such matrices, meaning that at each level we dispose of a circulant block matrix with blocks that are themselves circulant block matrices up to a level where the components are just scalar circulant matrices – this corresponds to a physical situation in which periodic boundary conditions are in effect in every dimension. We shall indicate the hierarchical level with a super-index. Assuming the overall hierarchical depth is D, $A^{(0)}$ indicates the top matrix in the hierarchy. Then

$$A^{(0)} = \mathrm{circ}(A_k^{(1)} : k = 1 \cdots n_1)$$

in which

$$A_k^{(1)} = \mathrm{circ}(A_{k;\ell}^{(2)} : \ell = 1 \cdots n_2)$$

etc... This keeps going until depth $d = D - 1$, where each of the sub-blocks is a regular circulant matrix with scalar entries.

Theorem 4. *Given a one depth block circulant matrix with $n \times n$ sub-blocks*

$$A = \begin{pmatrix} A_0 & A_1 & A_2 & \dots & A_{n-1} \\ A_{n-1} & A_0 & A_1 & \dots & \vdots \\ A_{n-2} & A_{n-1} & A_0 & \dots & \vdots \\ \vdots & \dots & \dots & \ddots & A_1 \\ A_1 & A_2 & \dots & A_{n-1} & A_0 \end{pmatrix}.$$

Assume that every sub-block A_k, $k = 1, \dots, n$ has the same dimension $p \times p$ and the same complete set of orthonormal eigenvectors assembled in the matrix U. Let the eigenvalue matrix of sub-block matrix A_k be $\Gamma_k = \mathrm{diag}(\gamma_{k;1}, \dots, \gamma_{k;p})$. Then the matrix A has eigenvalue $\Lambda_m = \mathrm{diag}(\lambda_{m;1}, \dots, \lambda_{m;p})$, with corresponding eigenvector block $x^{(m)}$, where

$$\lambda_{m;i} = \sum_{k=0}^{n-1} \gamma_{k;i} e^{\frac{-2\pi j m k}{n}},$$

$$x^{(m)} = \frac{1}{\sqrt{n}} [U, Ue^{\frac{-2\pi j m k}{n}}, \dots, Ue^{\frac{-2\pi j (n-1) k}{n}}]^T.$$

Proof. The equation $Ax = \lambda x$ for the eigenvalues and eigenvectors of A specializes to

$$\begin{pmatrix} A_0 & A_1 & \dots & A_{n-1} \\ A_{n-1} & A_0 & \dots & \vdots \\ A_{n-2} & A_{n-1} & \dots & \vdots \\ \vdots & \dots & \ddots & A_1 \\ A_1 & \dots & A_{n-1} & A_0 \end{pmatrix} \begin{pmatrix} \vec{x}_1 \\ \vec{x}_2 \\ \vdots \\ \vec{x}_n \end{pmatrix} = \begin{pmatrix} \Lambda_1 & & \\ & \ddots & \\ & & \Lambda_n \end{pmatrix} \begin{pmatrix} \vec{x}_1 \\ \vec{x}_2 \\ \vdots \\ \vec{x}_n \end{pmatrix} \tag{23}$$

where

$$\vec{x}_m = \begin{pmatrix} x_{m;1} \\ x_{m;2} \\ \vdots \\ x_{m;p} \end{pmatrix} \tag{24}$$

$$\Lambda_m = \begin{pmatrix} \lambda_{m;1} & & \\ & \ddots & \\ & & \lambda_{m;p} \end{pmatrix} \tag{25}$$

$$\sum_{k=0}^{n-1-m} A_k \vec{x}_{k+m} + \sum_{k=n-m}^{n-1} A_k \vec{x}_{k-n+m} = \vec{x}_m \Lambda_m \tag{26}$$

for $m = 1, \dots, n$.

Because

$$A_k = U \Gamma_k U^*. \tag{27}$$

Let

$$\vec{y}_m = U^* \vec{x}_m = \begin{pmatrix} y_{m;1} \\ y_{m;2} \\ \vdots \\ y_{m;p} \end{pmatrix} \tag{28}$$

then

$$\vec{x}_m = U \vec{y}_m, \tag{29}$$

$$\sum_{k=0}^{n-1-m} \Gamma_k \vec{y}_{k+m} + \sum_{k=n-m}^{n-1} \Gamma_k \vec{y}_{k-n+m} = \vec{y}_m \Lambda_m, \tag{30}$$

$$\sum_{k=0}^{n-1-m} \gamma_{k;i} y_{k+m;i} + \sum_{k=n-m}^{n-1} \gamma_{k;i} y_{k-n+m;i} = y_{m;i} \lambda_{m;i}, \tag{31}$$

where $m = 1, \dots, n$ and $i = 1, \dots, p$.

As this system of equations involves a scalar circulant matrix, namely

$$\operatorname{circ}(\gamma_{k;i} : k = 0 \cdots n-1),$$

the solution can be written down directly (the discrete Fourier transform of order n diagonalizes the matrix)

$$y_{k;i} = \rho^k, \tag{32}$$

with $\rho = e^{-\frac{2m\pi j}{n}}$. Furthermore

$$\lambda_{m;i} = \sum_{k=0}^{n-1} \gamma_{k;i} e^{\frac{-2\pi jmk}{n}}; \tag{33}$$

$$y_i = \frac{1}{\sqrt{n}}[1, e^{\frac{-2\pi jm}{n}}, \ldots, e^{\frac{-2\pi jm(n-1)}{n}}]^T. \tag{34}$$

Thus, for the sub-blocks,

$$\Lambda_m = \begin{pmatrix} \lambda_{m;1} & & \\ & \ddots & \\ & & \lambda_{m;p} \end{pmatrix}; \tag{35}$$

$$x^{(m)} = \begin{pmatrix} Uy_{i,1} \\ Uy_{i,2} \\ \vdots \\ Uy_{i,n} \end{pmatrix} \tag{36}$$

where I is the $p \times p$ identity matrix.

Therefore,

$$\Lambda_m = \begin{pmatrix} \lambda_{m;1} & & \\ & \ddots & \\ & & \lambda_{m;p} \end{pmatrix}, \tag{37}$$

$$\lambda_{m;i} = \sum_{k=0}^{n-1} \gamma_{k;i} e^{\frac{-2\pi jmk}{n}}, \tag{38}$$

$$x^{(m)} = \frac{1}{\sqrt{n}}[U^T, U^T e^{\frac{-2\pi jmk}{n}}, \ldots, U^T e^{\frac{-2\pi jm(n-1)k}{n}}]^T. \tag{39}$$

□

6.2.1. Application to the 3D Poisson equation on a regular grid. Using the notation of Section 2 we take the circulant version of the matrices $\Phi, M, L, \ldots$, i.e., the case $\epsilon = 1$. Application of Theorem 4 immediately provides the eigenvalues as samples of the Fourier transforms (it is just as convenient to handle the transforms directly):

$$\mathcal{F}(O) = -\frac{64}{15} + \frac{14}{15}\cos\theta$$
$$\mathcal{F}(P) = \frac{7}{15} + \frac{1}{5}\cos\theta$$
$$\mathcal{F}(R) = \frac{7}{15} + \frac{1}{5}\cos\theta$$

$$\begin{aligned}
\mathcal{F}(Q) &= \frac{1}{10} + \frac{1}{15}\cos\theta \\
\mathcal{F}(M) &= \mathcal{F}(O) + 2\mathcal{F}(P)\cos n\theta \\
&= -\frac{64}{15} + \frac{14}{15}\cos\theta + 2\left(\frac{7}{15} + \frac{1}{5}\cos\theta\right)\cos n\theta \\
\mathcal{F}(L) &= \mathcal{F}(R) + 2\mathcal{F}(Q)\cos n\theta \\
&= \frac{7}{15} + \frac{1}{5}\cos\theta + 2\left(\frac{1}{10} + \frac{1}{15}\cos\theta\right)\cos n\theta.
\end{aligned}$$

As in the previous schema, we can approximate the Schur complement sequence by calculating S_∞,

$$S_\infty = \frac{M + \sqrt{M^2 - 4LL^H}}{2} \tag{40}$$

with Fourier transform of S_∞

$$\begin{aligned}
\mathcal{F}(S_\infty) &= \mathcal{F}\left(\frac{M + \sqrt{M^2 - 4LL^H}}{2}\right) \qquad (41)\\
&= \frac{1}{2}(\mathcal{F}(M) + \sqrt{\mathcal{F}^2(M) - 4\mathcal{F}^2(L)}) \\
&= \frac{1}{15}[-32 + 7\cos\theta + 7\cos n\theta + 3\cos\theta\cos n\theta \\
&\quad + [5(-5 + 2\cos\theta + 2\cos n\theta + \cos\theta\cos n\theta) \\
&\qquad \cdots (-39 + 4\cos\theta + 4\cos n\theta + \cos\theta\cos n\theta)]^{\frac{1}{2}}].
\end{aligned}$$

This explicit representation of the fixed point for the Schur complement recursion, although interesting in its own right, allows us to evaluate various approximation schemas that lead to candidate pre-conditioners. We report on a number of results in the next section. Looking at the spectrum (formula 41) it is obvious that it oscillates strongly with a period π/m – in contrast to the 2D-case where the spectrum is very smooth with just a discontinuity in the first derivative at zero frequency [4]. As a consequence, any reasonable rational (finite state space) approximation of such a spectrum will need an order that is at least m. This shows that the 3D case is essentially different from 2D, where low order semi-separable approximation yield very good results. Hence, a different pre-conditioning strategy will have to be used in 3D and higher dimensions. The following section presents a preliminary investigation of this issue.

7. Decoupling strategies for matrices with 3D sparsity patterns

The goal of this section is to obtain insight in the possibilities of constructing pre-conditioners through decoupling, whereby the Schur complements that have to be introduced are being approximated by simpler matrices. This can happen at various positions in the schema, as we shall indicate further on in this section.

Since we dispose of exact inverses for special types of matrices (Poisson's equation on a regular grid) as presented in the previous sections, we can evaluate the various strategies on this very well-conditioned example. The idea behind this approach is that a pre-conditioning strategy should at least work well on simple straight examples, for there would be no hope for it to work in more complex cases if it already breaks down on the most well-conditioned realistic examples. We first describe the general strategies and then apply them on the Poisson case. The chapter has a very preliminary nature, as many possibilities have yet to be investigated.

As explained in the section "Hierarchical Decoupling Structure", the matrix on one layer of the symbolic hierarchy can be decoupled into several levels (all within that layer):

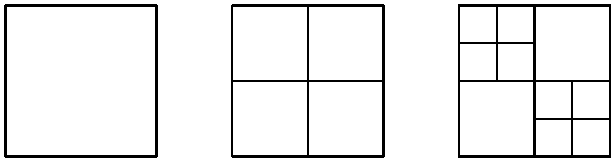

FIGURE 2. Level decomposition: level 0, level 1 and level 2 are shown

To keep terminology consistent we use the term 'layer' systematically to indicate the position in the symbolic hierarchy, while the term 'level' will be used for subdivisions within one layer as indicated by the diagrams above.

Positioning ourselves at a given layer of the hierarchy (say the top layer), if it consists of $n \times n$ sub-blocks from its next layer, then we dispose of $\log_2 n$ levels for decoupling totally at the given layer. Consider a point $k+1$ where an approximation of a Schur inverse has to be introduced. The exact solution is of course given by equation (10). The simplest approximant would of course be $S_{k+1} = M$, which would amount to putting $L = 0$. This we would refer to as a 'Jacobi step'. Notice that doing so, the resulting decoupled matrices remain positive definite (they are in fact more positive than the exact ones, as the difference is a semi-positive operator). This amount to a 'zero'th order' approximation. One step more (first order) would take $S_{k+1} = M - LM^{-1}L^H$, which again would produce a positive definite decoupling as can be seen from considering the positive definite block submatrix with rows and columns indexed by $k, k+1, \ldots$. This we would call a 'first-order approximant'. This process can of course be continued leading to second, third etc... order approximants. We show later that for the example the approximation error decreases quickly when higher orders are involved - the approximation error at the top layer of the hierarchy will appear to be the most significant factor.

The first-order approximation involves the inverse M^{-1}. This is a matrix at the next layer of the hierarchy. We may denote it as $S^{-1}_{k+1;-1}$ as we regress one stage in the Schur recursion. Similarly, the second order will involve a matrix $S_{k+1;-2}$, which would be obtained by regressing two stages – we have not studied

this case yet. As these inverses involve a lower layer of the hierarchy, again a Schur approximation can be introduced, but now involving matrices with a lower hierarchical structure. In the case of hierarchical depth $D = 3$, this may already involve a semi-separable approximation – we do present numerical results for this case where we restrict ourselves to a one-stage recursion (i.e., we approximate the relevant S_{k+1} as $M - LM^{-1}L^H$).

Matrix Φ is then first decoupled in the middle, where $k = \lfloor \frac{n}{2} \rfloor$. The process can then be repeated at the next levels on the same layer. We specify a level number $lvl1$ and decouple matrix Φ into 2^{lvl1} matrices as in the picture above, and approximate the Schur complement by $S_{k+1} = M - LM^{-1}L^H$ at the decoupling points while performing the exact recursion in between:

$$\begin{cases} S_0 = M, \\ S_{k+1} = M - LM^{-1}L^H, & k = \lfloor \frac{n}{2^{lvl1}} \rfloor, \lfloor \frac{2n}{2^{lvl1}} \rfloor, \dots \\ S_{k+1} = M - LS_k^{-1}L^H, & k = \text{else}. \end{cases} \tag{42}$$

Let us now specialize these ideas to our prototype example, the matrix representing the 3D regular Poisson problem with discretization based on the 27 points stencil. It has the form given in Fig. 3 (here specialized to $n = 8$). Each block in the matrix

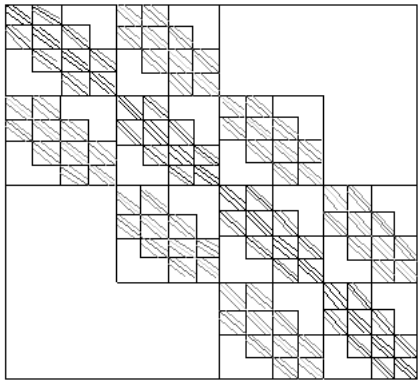

FIGURE 3. Structure of matrix Φ

represents a 2D plane, while each row corresponds to an array of points in the x-direction. The interactions in the y-direction take place within the blocks while the interactions in the z-direction create entries between blocks in the top layer. These interactions are shown in Fig. 4. where each of the xy-plane corresponds to a diagonal 2D block, each x line corresponds to a diagonal 1D block, and the lines between these points correspond to the off-diagonal blocks or entries, representing their relations. Introducing a cut in the top layer matrix (e.g., in the middle as shown in Fig. 5) forces the elimination of the cross dependencies between the layers, which are then incorporated as a (Schur-) correction $(-LS_k^{-1}L^H)$ on

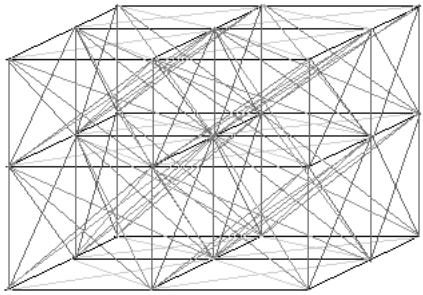

FIGURE 4. Physical model of matrix Φ

the next diagonal block. Our strategy hence consists in estimating rather than calculating that correction.

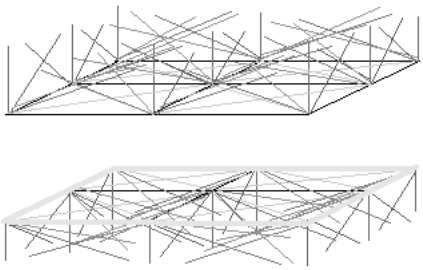

FIGURE 5. Cut on 3D layer (0 order)

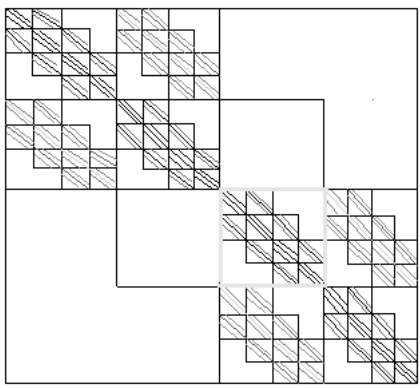

FIGURE 6. Matrix cut on 3D layer (0 order)

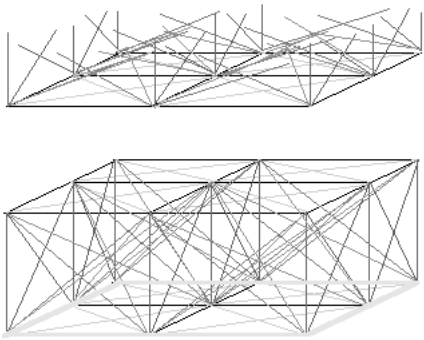

FIGURE 7. Cut on 3D layer (1 order)

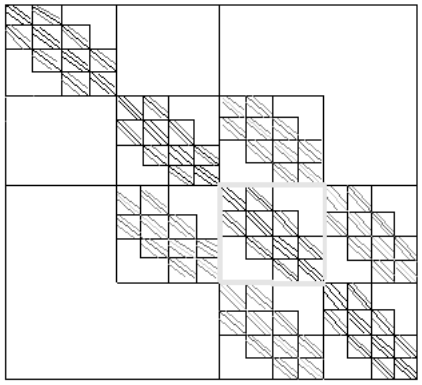

FIGURE 8. Matrix cut on 3D layer (1 order)

The first layer steps (up to a certain level) hence consist in the estimation of the Schur correction term, which in the first-order approximation is $-LM^{-1}L^H$ in which M and L again have structure, in our example they are block tridiagonal, with blocks that are tridiagonal with scalar entries. A central step is then the computation of M^{-1} with such a structure. One way of proceeding is to assume semi-separability for this layer of computation, motivated by the fact that the 2D case is (extremely) well approximated by such systems, in contrast to 3D systems. Assuming low-order semi-separability, S_k for $k = 1, \ldots, n$ will also be semi-separable, and we can calculate the Schur complements S_k using the state space theory expounded in Sections 4 and 5. In the experiments we report on here, we use a rather global semi-separable strategy, restricting the semi-separable order to $n_{kp} = rm$ (i.e., the size of a block) with a low value of r (called MRsize for 'Model Reduction Size'). It turns out that with very low values the approximation error is already negligible.

Simulation results

To characterize the various experiments we introduce the notation

'3D(level,order)2D(level,order,MRsize)'

to indicate an experiment in which the indicated levels and orders have been used in the 3D block matrix, respect. 2D block submatrices, with an eventual semi separable model order reduction size in the 2D case. Figure 9 shows $\|I - \Phi P\|_F$ for a number of situations (if a 2D specification is not shown, it means that the 2D calculation has been done on the full matrix without approximations). Each schema is carried out on orders from 0 to 3 (the latter meaning a regression of 3 layers). The X-axis shows the 3D level, while y-axis shows $\|I - \Phi P\|_F$, where the preconditioner $P \approx \Phi^{-1}$ is computed as

$$P = \begin{pmatrix} I & 0 \\ \Phi_{21}\Phi_{11}^{-1} & I \end{pmatrix} \begin{pmatrix} \Phi_{11} & \Phi_{12} \\ 0 & S^{(3D)} \end{pmatrix}$$

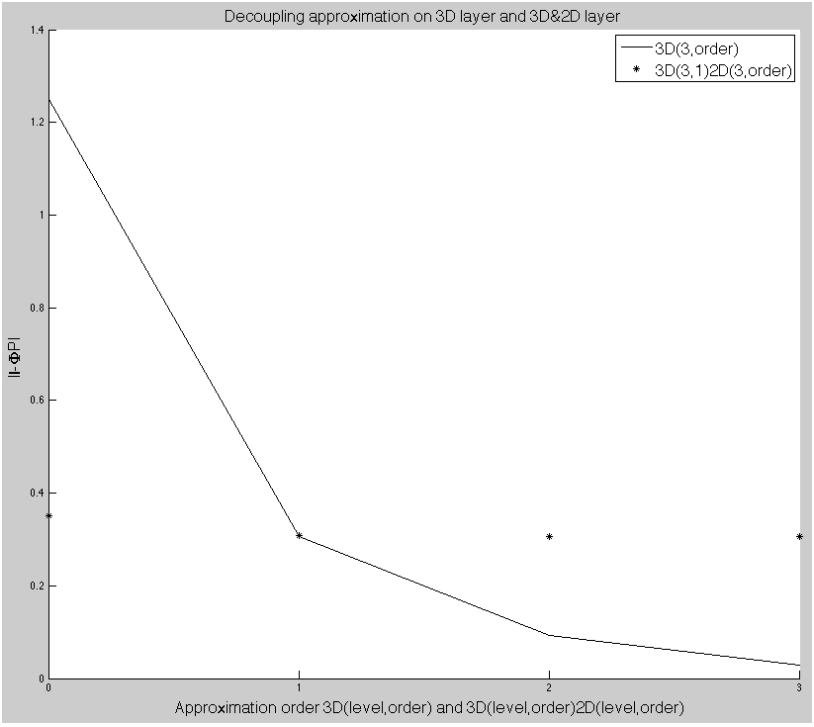

FIGURE 9. $\|I - \Phi P\|_F$ for decoupling approximation on 3D layer and 3D & 2D layers, where $3D(level, order)$ is by only decoupling on 3D layer with decoupling level *level* and approximation order *order*, while $3D(level, order)2D(level, order)$ is by decoupling on 2D layer when the *level* and *order* on 3D layer is 3 and 1

in which $S^{(3D)}$ is the approximated Schur complement by the difference approximation schemas described above. The order on the 3D layer in this figure is just one, and three levels approximation are shown as explained before. So the x-axis shows the level used on the 2D submatrices. We compare the different approximation orders on 2D taking the one without 2D approximation as a reference. As expected, cutting on both the 2D and 3D layers gives more error, but the influence of the approximation on the lower layer is much smaller compared with that on the 3D layer. This means that for a good preconditioner, the top layer must use a higher order of approximation than the lower layers. This situation is expected and it is quite pronounced.

Figure 10 shows the approximations in Frobenius norm, if in addition model order reduction with sizes 1 and 2 are applied on the 2D submatrices, and it shows the comparison with the no model reduction case. Here $r = 1$ means the realization matrices considered size 8×8, while they are 16×16 when $r = 2$.

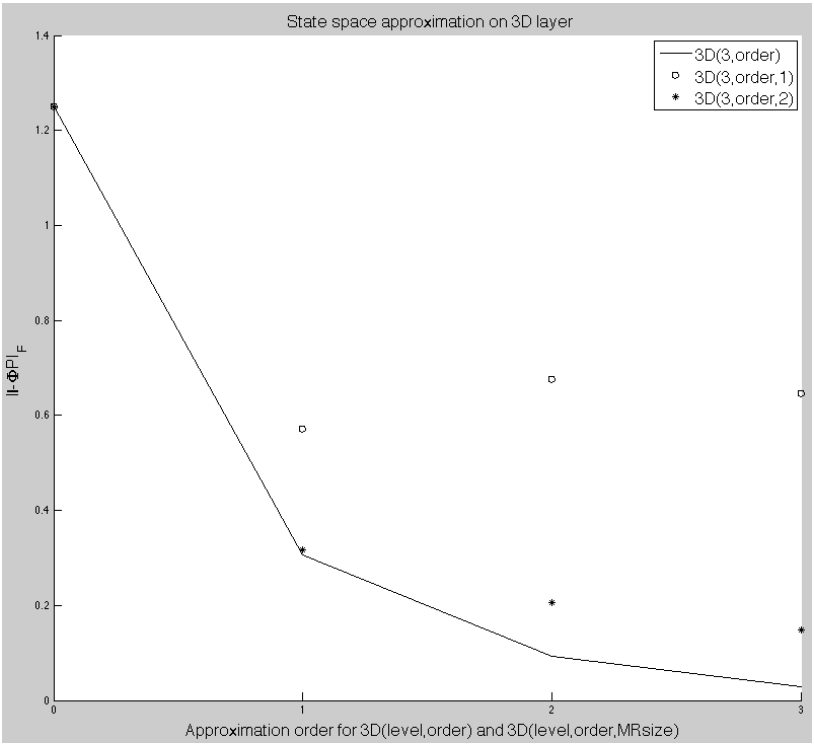

FIGURE 10. $\|I - \Phi P\|_F$ for state space approximation on 3D layer, where $3D(\text{level}, \text{order})$ is by matrix computation and

$$3D(\text{level}, \text{order}, \text{reduction rate})_{\text{state space}}$$

is by state space realization with model reduction rate *state space*

Eigenvalues of the preconditioning error. the preconditioner P is supposed to approximate Φ^{-1}. Its performance can be checked by evaluating the maximum eigenvalue $\lambda_{\max}$ of $(I-\Phi P)$ for all the proposed cases. These are listed below from table 1 to 5 for approximation schemas "Fixed point", "Cutting on 3D layer", "Cutting on 3D layer in state space with model reduction rate $r = 2$", "Cutting on 3D layer but keeping the diagonal values on off-diagonal 3D blocks", and "Cutting on both 2D and 3D layer". In these tables *level* stands for the number of level cut in 3D/2D layer; and *order* stands for the approximation order.

$\lambda_{\max}(I - \Phi P)$
0.16

TABLE 1. Fixed point

$\lambda_{\max}(I - \Phi P)$			
Order	level=1	level=2	level=3
0	0.3636	0.4420	0.5413
1	0.1304	0.1629	0.2154
2	0.0440	0.0554	0.0754
3	0.0119	0.0178	0.0244

TABLE 2. Cutting on 3D layer

$\lambda_{\max}(I - \Phi P)$			
Order	level=1	level=2	level=3
0	0.3636	0.4420	0.5413
1	0.1304	0.1629	0.2154
2	0.0440	0.0554	0.0754
3	0.0119	0.0178	0.0244

TABLE 3. Cutting on 3D layer with rate $r = 2$

$\lambda_{\max}(I - \Phi P)$			
Order	level=1	level=2	level=3
0	0.3205	0.3935	0.4884
1	0.1145	0.1428	0.1902
2	0.0381	0.0485	0.0658
3	0.0098	0.0153	0.0212

TABLE 4. Partially cutting on 3D layer

If we require the preconditioner to produce error eigenvalues smaller than 0.1, then schemas that qualify are "Cutting on 3D layer", "Cutting on 3D layer

$\lambda_{\max}(I - \Phi P)$				
Order	level=0	level=1	level=2	level=3
0	0.2154	0.2218	0.2292	0.2419
1	0.2154	0.2157	0.2160	0.2165
2	0.2154	0.2155	0.2155	0.2155
3	0.2154	0.2154	0.2154	0.2155

TABLE 5. Cutting on both 2D and 3D layer

$\lambda_{\max}(I - \Phi P)$			
Order	level=1	level=2	level=3
0	0.3636	0.4420	0.5413
1	0.1304	0.1629	0.2154
2	0.0440	0.0554	0.0754
3	0.0119	0.0178	0.0244

TABLE 6. Cutting 3D layer & 2D$\approx$ SSS

in state space with model reduction" and "Cutting on 3D layer but keeping the diagonal values on off-diagonal 3D blocks" when the approximation order is larger than one.

8. Discussion

Many more numerical results are available than those presented in this paper. We have tried in particular to use the fixed point solution as preconditioner with excellent results (almost no approximation error). If such a strategy is desired, then an efficient method must be devised to compute the fixed point solution – a question that is solvable using the exact solutions presented here, but it goes beyond the purposes of the present paper. The numerical results that we do present show that there exist very good schemas in which the preconditioning error (defined as $\lambda_{\max}(I - \Phi P)$) is below .1 or .2, resulting in very good low complexity preconditioners. This is to some extent due to the good conditioning of the Poisson matrix Φ. In future work we wish to investigate whether these properties can be extended to system matrices that are much less well conditioned, such as matrices resulting from full Maxwell equations. This will be the next effort on the program.

References

[1] S. Chandrasekaran, M. Gu, and T. Pals. Fast and stable algorithms for hierarchically semi-separable representations. In *Technical Report*. University of California at Santa Barbara, April 2004.

[2] A.-J. van der Veen. Approximant inversion of a large semi-separable matrix. In *Proc. MTNS'*04, 2004.

[3] H.A. van der Vorst. *Iterative Krylov Methods for Large Linear Systems*. Cambridge University Press, 2003.

[4] P. Dewilde, K. Diepold, and W. Bamberger. A semi-separable approach to a tridiagonal hierarchy of matrices with application to image flow analysis. In *Proceedings MTNS*, 2004.

[5] P. Dewilde and A.-J. van der Veen. *Time-varying Systems and Computations*. Kluwer, 1998.

[6] Carl D. Meyer. *Matrix Analysis and Applied Linear Algebra*. SIAM, 2000.

[7] P. Dewilde, S. Chandrasekaran and M. Gu. On the numerical rank of the off-diagonal blocks of Schur complements of discretized elliptic pde's. *submitted*, March, 2007.

[8] S. Chandrasekaran, P. Dewilde, M. Gu, T. Pals, A.-J. van der Veen and J. Xia. *A fast backward stable solver for sequentially semi-separable matrices*, volume HiPC202 of *Lecture Notes in Computer Science*, pages 545–554. Springer Verlag, Berlin, 2002.

[9] J. Zhang. A class of multilevel recursive incomplete lu preconditioning techniques. *Korean J. Comp. Appl. Math.*, 8(2):213–234, 2001.

Patrick Dewilde and Haiyan Jiao
Circuits and Systems
Faculty of EEMCS
Mekelweg 4
Delft University of Technology
NL-2628 CD Delft, The Netherlands
e-mail: p.dewilde@ewi.tudelft.nl
h.jiao@tudelft.nl

Shiv Chandrasekaran
Department of Electrical and Computer Engineering
University of California,
Santa Barbara
Santa Barbara, CA 93106, USA
e-mail: shiv@ece.ucsb.edu

Operator Theory:
Advances and Applications, Vol. 197, 133–138

Pick Matrices for Schur Multipliers

Harry Dym and Dan Volok

Dedicated to the memory of Moshe Livsic, a gentleman and a scholar

Abstract. It is well known that the classical Nevanlinna-Pick problem for holomorphic contractive functions in the open unit disk is solvable if and only if a matrix P with entries of the form

$$p_{jk} = \frac{1 - \overline{\eta_j}\eta_k}{1 - \omega_j\overline{\omega_k}}, \quad j.k = 1, \dots, n,$$

that is based on the data of the problem is positive semidefinite. The purpose of this purely expository note is to draw attention to another matrix that arises in the theory of interpolation problems for multipliers which deserves to be better known. This matrix and other more general forms are discussed in [AlB97] and [AlBL96]. Our interest in this problem was aroused by the formula

$$p_{jk} = \frac{1 - \overline{\eta_j}(1 - \omega_j^{\ell}\overline{\omega_k}^{\ell})\eta_k}{1 - \omega_j\overline{\omega_k}}, \quad j.k = 1, \dots, n,$$

that was mentioned by M.A. Kaashoek in a lecture at the IWOTA conference in Blacksburg, Virginia, as a byproduct of his joint investigations with C. Foias and A.E. Frazho [FFK02] into constrained lifting problems.

Mathematics Subject Classification (2000). Primary 30E05; Secondary 47B32, 47B38, 46E22.

Keywords. Tangential interpolation, Schur class, multipliers, Pick matrices, reproducing kernel Hilbert spaces.

1. Introduction

Let Ω_+ denote either the open unit disc $\mathbb{D} = \{\lambda \in \mathbb{C} : |\lambda| < 1\}$ or the open upper half-plane $\mathbb{C}_+ = \{\lambda \in \mathbb{C} : -i(\lambda - \overline{\lambda}) > 0\}$, let

$$\rho_\omega(\lambda) = \begin{cases} (1 - \lambda\overline{\omega}) & \text{if } \Omega_+ = \mathbb{D} \\ -2\pi i(\lambda - \overline{\omega}) & \text{if } \Omega_+ = \mathbb{C}_+ \end{cases}$$

and let $\mathcal{S}^{p\times q}(\Omega_+)$ denote the Schur class of $p \times q$ mvf's (matrix-valued functions) $s(\lambda)$ that are holomorphic and contractive (i.e., $\|s(\lambda)\| \le 1$) in Ω_+.

It is well known that:

Theorem 1.1. *If*

$$\omega_1, \dots, \omega_n \in \Omega_+, \quad \xi_1, \dots, \xi_n \in \mathbb{C}^p \quad \text{and} \quad \eta_1, \dots, \eta_n \in \mathbb{C}^q, \tag{1.1}$$

then there exists a mvf $s(\lambda)$ in the Schur class $\mathcal{S}^{p\times q}(\Omega_+)$ such that

$$\xi_j^* s(\omega_j) = \eta_j^* \quad \text{for } j = 1, \dots, n \tag{1.2}$$

if and only if the $n \times n$ matrix P with jk entry

$$p_{jk} = \frac{\xi_j^* \xi_k - \eta_j^* \eta_k}{\rho_{\omega_k}(\omega_j)} \tag{1.3}$$

is positive semidefinite.

A proof of Theorem 1.1 based on reproducing kernel Hilbert spaces may be found, e.g., in [Dy89] and [Dy03]; see also [Dy94] for a sketch of other approaches and additional references.

The purpose of this expository note is to draw attention to another less well-known Pick matrix that is encountered in the theory of interpolation problems for multipliers.

2. Preliminaries

Let $H_2^k(\Omega_+)$ denote the set of $k \times 1$ vector-valued functions with entries in the Hardy space $H_2(\Omega_+)$. The vector-valued functions in $H_2^k(\Omega_+)$ have nontangential limits a.e. on the boundary of Ω_+ and are Hilbert spaces with respect to the inner product

$$\langle f, g\rangle_{st} = \begin{cases} \frac{1}{2\pi}\int_0^{2\pi} g(e^{i\theta})^* f(e^{i\theta}) d\theta & \text{if } \Omega_+ = \mathbb{D} \\ \int_{-\infty}^{\infty} g(\mu)^* f(\mu) d\mu & \text{if } \Omega_+ = \mathbb{C}_+. \end{cases}$$

If $b(\lambda)$ is a $q \times q$ inner function, then $\mathcal{H}(b) = H_2^q \ominus bH_2^q$ is a closed subspace of H_2^q.

A Hilbert space $\mathcal{H}$ of $m \times 1$ vector-valued functions defined on Ω_+ is said to be a reproducing kernel Hilbert space if there exists an $m \times m$ matrix-valued function $K_\omega(\lambda)$ on $\Omega_+ \times \Omega_+$ such that for every choice of $f \in \mathcal{H}$, $\omega \in \Omega_+$ and $u \in \mathbb{C}^m$

$$K_\omega u \in \mathcal{H} \quad \text{and} \quad \langle f, K_\omega u\rangle_{\mathcal{H}} = u^* f(\omega).$$

A matrix-valued function $K_\omega(\lambda)$ with these two properties is called a reproducing kernel for the reproducing kernel Hilbert space $\mathcal{H}$. It is readily checked that:

(1) A reproducing kernel Hilbert space has exactly one reproducing kernel.

(2) If $K_\omega(\lambda)$ is a reproducing kernel, then $K_\omega(\lambda)^* = K_\lambda(\omega)$.

(3) A reproducing kernel $K_\omega(\lambda)$ on $\Omega_+ \times \Omega_+$ is positive in the sense that

$$\sum_{j,k=1}^{n} u_k^* K_{\omega_j}(\omega_k) u_j \geq 0$$

for every choice of vectors $u_, \ldots, u_n \in \mathbb{C}^m$ and points $\omega_1, \ldots, \omega_n \in \Omega_+$.

(4) Finite linear combinations of the form $\sum_{j=1}^n K_{\omega_j} u_j$ with $u_j \in \mathbb{C}^m$ and $\omega_j \in \Omega_+$ are dense in $\mathcal{H}$.

The Hilbert spaces H_2^p and $\mathcal{H}(b) = H_2^q \ominus bH_2^q$ are both reproducing kernel Hilbert spaces with respect to the standard inner product with reproducing kernels

$$\frac{I_p}{\rho_\omega(\lambda)} \quad \text{and} \quad \Lambda_\omega^b(\lambda) = \frac{I_q - b(\lambda)b(\omega)^*}{\rho_\omega(\lambda)},$$

respectively as is easily checked with the help of Cauchy's formula for the Hardy space H_2.

3. The main formula

Let M_s denote the operator of multiplication by a $p \times q$ mvf $s(\lambda)$ that is holomorphic on Ω_+, from a subspace of H_2^q into H_2^p.

Theorem 3.1. *Let $b(\lambda)$ be a $q \times q$ mvf that is inner with respect to Ω_+ and $\|b(\lambda)\| < 1$ for each point $\lambda \in \mathbb{C}_+$. Then there exists a $p \times q$ mvf $s(\lambda)$ that is holomorphic in Ω_+ such that*

(1) *The interpolation conditions* (1.2) *are met and*
(2) *M_s maps the Hilbert space $\mathcal{H}(b) = H_2^q \ominus bH_2^q$ contractively into the Hilbert space H_2^p*

if and only if the $n \times n$ matrix $\widetilde{P}$ with jk entry

$$\widetilde{p}_{jk} = \frac{\xi_j^* \xi_k - \eta_j^* (I_q - b(\omega_j) b(\omega_k)^*) \eta_k}{\rho_{\omega_k}(\omega_j)} \tag{3.1}$$

is positive semidefinite.

Proof. Suppose first that conditions (1) and (2) are met and let

$$f = \sum_{j=1}^{n} c_j \Lambda_{\omega_j}^b \eta_j \quad \text{and} \quad g = \sum_{i=1}^{n} c_i \frac{\xi_i}{\rho_{\omega_i}} \quad \text{for any choice of} \quad c_1, \ldots, c_n \in \mathbb{C}. \tag{3.2}$$

Then

$$\begin{aligned} \langle M_s f, g \rangle_{st} &= \left\langle M_s f, \sum_{i=1}^{n} c_i \frac{\xi_i}{\rho_{\omega_i}} \right\rangle_{st} = \sum_{i=1}^{n} \overline{c_i} \xi_i^* s(\omega_i) f(\omega_i) \\ &= \sum_{i,j=1}^{n} \overline{c_i} \xi_i^* s(\omega_i) \Lambda_{\omega_j}^b(\omega_i) \eta_j c_j = \sum_{i,j=1}^{n} \overline{c_i} \eta_i^* \Lambda_{\omega_j}^b(\omega_i) \eta_j c_j = \|f\|_{st}^2. \end{aligned}$$

Moreover, since M_s is contractive,

$$\|f\|_{st}^2 = |\langle M_s f, g\rangle_{st}| \le \|M_s f\|_{st}\|g\|_{st} \le \|f\|_{st}\|g\|_{st}, \quad \text{i.e.,} \quad \|f\|_{st} \le \|g\|_{st}.$$

Thus, as

$$\|f\|_{st}^2 = \sum_{i,j=1}^{n} \overline{c_i}\eta_i^* \Lambda^b_{\omega_j}(\omega_i)\eta_j c_j \qquad \text{and} \qquad \|g\|_{st}^2 = \sum_{i,j=1}^{n} \overline{c_i}\frac{\xi_i^*\xi_j}{\rho_{\omega_j}(\omega_i)}c_j,$$

it follows that

$$\sum_{i,j=1}^{n} \overline{c_i}\left\{\frac{\xi_i^*\xi_j - \eta_i^*(I_q - b(\omega_i)b(\omega_j)^*)\eta_j}{\rho_{\omega_j}(\omega_i)}\right\}c_j \ge 0$$

for every choice of $c_1, \dots, c_n \in \mathbb{C}$. and hence that the modified Pick matrix $\widetilde{P}$ defined by (3.1) is positive semidefinite.

Now suppose conversely that $\widetilde{P}$ is positive semidefinite and let

$$x_j = \begin{bmatrix} \xi_j \\ b(\omega_j)^*\eta_j \end{bmatrix} \quad \text{and} \quad y_j = \eta_j \quad \text{for } j = 1, \dots, n.$$

Then, since $\widetilde{p}_{jk} = (x_j^* x_k - y_j^* y_k)/\rho_{\omega_k}(\omega_j)$, Theorem 1.1 guarantees that there exists a mvf $\Sigma \in \mathcal{S}^{(p+q)\times q}$ such that

$$x_j^*\,\Sigma(\omega_j) = y_j^* \quad \text{for } j = 1, \dots, n. \tag{3.3}$$

Thus, upon writing

$$\Sigma(\lambda) = \begin{bmatrix} \Sigma_{11}(\lambda) \\ \Sigma_{12}(\lambda) \end{bmatrix}$$

with blocks $\Sigma_{11}(\lambda)$ of size $p\times q$ and $\Sigma_{12}(\lambda)$ of size $q\times q$, the interpolation conditions (3.3) can be reexpressed as

$$\xi_j^*\Sigma_{11}(\omega_j) + \eta_j^* b(\omega_j)\Sigma_{12}(\omega_j) = \eta_j^* \quad \text{for } j = 1, \dots, n,$$

and hence, upon setting

$$s(\lambda) = \Sigma_{11}(\lambda)(I_q - b(\lambda)\Sigma_{12}(\lambda))^{-1},$$

that

$$\xi_j^* s(\omega_j) = \eta_j^* \quad \text{for } j = 1, \dots, n.$$

Moreover, the kernel

$$\begin{aligned} L_\omega(\lambda) &= \frac{I_p - s(\lambda)\{I_q - b(\lambda)b(\omega)^*\}s(\omega)^*}{\rho_\omega(\lambda)} \\ &= \frac{[I_p \quad s(\lambda)b(\lambda)]\begin{bmatrix} I_p \\ b(\omega)^*s(\omega)^* \end{bmatrix} - s(\lambda)s(\omega)^*}{\rho_\omega(\lambda)} \\ &= \frac{[I_p \quad s(\lambda)b(\lambda)]\,\{I_{p+q} - \Sigma(\lambda)\Sigma(\omega)^*\}\begin{bmatrix} I_p \\ b(\omega)^*s(\omega)^* \end{bmatrix}}{\rho_\omega(\lambda)} \end{aligned}$$

is clearly positive on $\Omega_+ \times \Omega_+$.

It remains to show that the multiplication operator M_s maps $\mathcal{H}(b)$ contractively into H_2^p.
Let T denote the linear operator from H_2^p into $\mathcal{H}(b)$ that is defined by the formula

$$T\frac{x}{\rho_\alpha} = \Lambda_\alpha^b s(\alpha)^* x \quad \text{for } \alpha \in \Omega_+ \quad \text{and} \quad x \in \mathbb{C}^p.$$

Then, since the kernel $L_\omega(\lambda)$ is positive on $\Omega_+ \times \Omega_+$,

$$\left\| T\sum_{j=1}^n c_j \frac{x_j}{\rho_{\alpha_j}} \right\|_{st}^2 = \left\| \sum_{j=1}^n c_j \Lambda_{\alpha_j}^b s(\alpha_j)^* x_j \right\|_{st}^2 = \sum_{j,k=1}^n \overline{c_k} x_k^* s(\alpha_k) \Lambda_{\alpha_j}^b(\alpha_k) s(\alpha_j)^* x_j c_j$$
$$\le \sum_{j,k=1}^n \overline{c_k} \frac{x_k^* x_j}{\rho_{\alpha_k}(\alpha_j)} c_j = \left\| \sum_{j=1}^n c_j \frac{x_j}{\rho_{\alpha_j}} \right\|_{st}^2,$$

which clearly displays the fact that $\|T\| \le 1$. Moreover, the formula

$$\left\langle T\frac{x}{\rho_\alpha}, \Lambda_\beta^b y \right\rangle_{st} = y^* \left(T\frac{x}{\rho_\alpha} \right)(\beta) = y^* \Lambda_\alpha^b(\beta) s(\alpha)^* x.$$

implies that

$$x^* s(\alpha) \Lambda_\beta^b(\alpha) y = \left\langle \Lambda_\beta^b y, T\frac{x}{\rho_\alpha} \right\rangle_{st} = \left\langle T^* \Lambda_\beta^b y, \frac{x}{\rho_\alpha} \right\rangle_{st}$$
$$= x^* (T^* \Lambda_\beta y)(\alpha),$$

i.e.,

$$(T^* \Lambda_\beta y)(\alpha) = s(\alpha) \Lambda_\beta^b(\alpha) y = (M_s \Lambda_\beta^b y)(\alpha).$$

Thus,

$$\|M_s\| = \|T^*\| = \|T\| \le 1$$

and the proof is complete. □

Remark 3.2. The proof of the preceding theorem is easily adapted to show that

$$\|M_s\| \le 1 \Longleftrightarrow L_\omega(\lambda) \quad \text{is a positive kernel on } \Omega_+ \times \Omega_+. \tag{3.4}$$

Thus, if

$$f = \sum_{j=1}^n c_j \Lambda_{\alpha_j}^b s(\alpha_j)^* x_j \quad \text{and} \quad g = \sum_{i=1}^n c_i \frac{x_i}{\rho_{\alpha_i}}$$

for any choice of $\alpha_1, \dots, \alpha_n \in \Omega_+$, $x_1, \dots, x_n \in \mathbb{C}^p$ and $c_1, \dots, c_n \in \mathbb{C}$, then the implication $\Longrightarrow$ follows from the formulas

$$\|f\|_{st}^2 = \langle M_s f, g \rangle_{st} \le \|M_s\| \|f\|_{st} \|g\|_{st}$$

just as in the first part of the proof of the theorem. The opposite implication is verified in the last few lines of the proof of the theorem.

References

[AlB97] D. Alpay and V. Bolotnikov, On tangential interpolation in reproducing kernel Hilbert modules and applications; in *Topics in Interpolation Theory* (Leipzig, 1994) (H. Dym, B. Fritzsche, V. Katsnelson and B. Kirstein, eds.), Oper. Theory Adv. Appl. **95**, Birkhäuser, Basel, 1997, pp. 37–68.

[AlBL96] D. Alpay, V. Bolotnikov and P. Loubaton, On tangential H_2 interpolation with second order norm constraints. Integ. Equat. Oper. Th. **24** (1996), no. 2, 156–178.

[Dy89] H. Dym, *J Contractive Matrix Functions, Reproducing Kernel Hilbert Spaces and Interpolation*. CBMS Regional Conference Series in Mathematics, **71**, Amer. Math. Soc., Providence, RI, 1989.

[Dy94] H. Dym, Expository Review of *The Commutant Lifting Approach to Interpolation* by C. Foias and A.E. Frazho, Bull. Amer. Math. Soc. **31** (1994), no. 1, 125–140.

[Dy03] H. Dym, Riccati equations and bitangential interpolation problems with singular Pick matrices; in *Fast Algorithms for Structured Matrices: Theory and Applications* (South Hadley, MA, 2001) (V. Olshevsky, ed.), Contemp. Math. **323**, Amer. Math. Soc., Providence, RI, 2003, pp. 361–391.

[FFK02] C. Foias, A. Frazho and M.A. Kaashoek Relaxation of metric constrained interpolation and a new lifting theorem. Integ. Equat. Oper. Th. **42** (2002), no. 3, 253–310.

Harry Dym
Department of Mathematics
The Weizmann Institute, Rehovot 76100, Israel
e-mail: harry.dym@weizmann.ac.il

Dan Volok
Department of Mathematics
Kansas State University, Manhattan
Kansas 66506, USA
e-mail: danvolok@math.ksu.edu

Operator Theory:
Advances and Applications, Vol. 197, 139–148

The Stable Rank of a Nest Algebra and Strong Stabilization of Linear Time-varying Systems

Avraham Feintuch

This paper is dedicated to the memory of my teacher, colleague and friend, Moshe Livsic, one of the great operator theorists of our time. Conversations with Moshe, about mathematics or any other topic always reminded me of a statement of the Rabbis of the Talmud (Tractate Berachot, 38, a) that one who sees a wise man is required to thank God "who gave of his wisdom to humanity".

Abstract. It is shown that the stable rank of a continuous time nest algebra is infinite. From this it follows that there exist stabilizable continuous time systems which are not strongly stabilizable.

Keywords. Time-varying linear systems, coprime factorizations, stabilization, causality, continuous nests, stable rank.

1. Introduction

The notion of stable rank of a ring, was introduced by Bass ([1]) to study stabilization results in algebraic K-theory. Bass' results turned out to have several applications extending to topics in topology and analysis and in the early 1980's Corach and Suarez ([2]) computed the stable rank of various Banach algebras. In particular they showed that the stable rank of the algebra of bounded linear operators on an infinite-dimensional Hilbert space $\mathcal{H}$ is infinite. One can look at this theorem as giving the stable rank of the nest algebra determined by the trivial nest $\{\{0\}, \mathcal{H}\}$ and therefore raising the problem of computing the stable rank of nest algebras with nontrivial nests. In this paper we consider this problem. In particular we show that for any continuous nest its nest algebra also has infinite stable rank. Our motivation for studying this problem comes from a well-known problem in the stability theory of linear feedback systems, the strong stabilization problem: given a linear system which is stabilizable by linear dynamic feedback, can one choose the stabilizing compensator to be itself a stable system? This problem seems to have been initially posed by Vidyisagar ([15])in the framework of linear

time-invariant finite-dimensional systems and has since been studied in various frameworks by numerous authors (e.g., [10]). The deep result of S. Treil ([13]) that H^∞ has stable rank one and its immediate extension to the algebra $M_n(H^\infty)$ of n by n matrices with entries from H^∞ by A. Quadrat gives that stabilizable linear time invariant multi-input multi-output systems are strongly stabilizable. In this paper we consider this problem for stabilizable linear time-varying systems. We obtain that for continuous time, there exist such systems which are not strongly stabilizable. We discuss the implications of this result and define a more tractable problem. The results here do not shed any light on the discrete time case.

2. Preliminaries

Let $\mathcal{H}$ denote the complex infinite-dimensional Hilbert function space $L_2[0,\infty)$, the space of square integrable complex-valued functions on the interval $[0,\infty)$, with the standard inner product and norm. For each $0 \leq t < \infty$, P_t denotes the standard truncation projection on the subspace $L_2[0,t]$ defined on $\mathcal{H}$. For each t we consider the seminorm defined for $x \in \mathcal{H}$ by $\| x \|_t = \| P_t x \|$. This family of semi-norms defines the "resolution topology" on $\mathcal{H}$ (see [6], Chapter 5). The completion of $\mathcal{H}$ with respect to this topology is called the extended space of $\mathcal{H}$ and is denoted by $\mathcal{H}_e$. In fact this is just

$$\{f : [0,\infty) \to C : f \in L_2[0,t] \forall 0 \leq t < \infty\}.$$

The totally ordered continuous set $\{P_t : 0 \leq t \leq \infty\}$ (where $P_\infty = I$) is used to define the physical notion of causality for linear systems. A continuous linear transformation T on $\mathcal{H}_e$ (with the resolution topology) is a causal linear system if for each $0 \leq t < \infty$, $P_t T = P_t T P_t$. A causal linear system T is stable if its restriction to $\mathcal{H}$ is a bounded operator ([6], Chapter 5).This means that there exists a positive constant c such that for every f in $\mathcal{H}$, Tf is in $\mathcal{H}$ and $\|Tf\| \leq c\|f\|$. The set of stable systems is a weakly closed Banach algebra $\mathcal{C}$ containing the identity. This is the algebra of operators which leave invariant the (complete) nest $\{L_2[t,\infty)\}$ of subspaces. Such algebras are called nest algebras ([6]), and in our case the nest is continuous, a fact of fundamental importance for our analysis.

Let L and C be given causal linear systems and consider the standard feedback configuration with plant L and compensator C, where the closed loop system equation for this configuration is

$$\begin{bmatrix} u_1 \\ u_2 \end{bmatrix} = \begin{bmatrix} I & C \\ L & -I \end{bmatrix} \begin{bmatrix} e_1 \\ e_2 \end{bmatrix}.$$

The system is well posed if the internal input $e = \begin{bmatrix} e_1 & e_2 \end{bmatrix}$ can be expressed as a causal function of the external input $\begin{bmatrix} u_1 & u_2 \end{bmatrix}$. This is equivalent to ([6], chapter 6) requiring that $\begin{bmatrix} I & C \\ L & -I \end{bmatrix}$ be invertible. This inverse is easily computed

formally and is given by the transfer matrix

$$H(L,C) = \begin{bmatrix} (I+CL)^{-1} & C(I+LC)^{-1} \\ L(I+CL)^{-1} & -(I+LC)^{-1} \end{bmatrix}.$$

L and C may not be stable. This means that there may be an input u in $\mathcal{H}$ such that Lu or Cu may not be in $\mathcal{H}$. Let $\mathcal{D}(L) = \{u \in \mathcal{H} : Lu \in \mathcal{H}\}$ and $\mathcal{D}(C) = \{u \in \mathcal{H} : Cu \in \mathcal{H}\}$. Then $\begin{bmatrix} I & C \\ L & -I \end{bmatrix}$ can be seen as a linear transformation from $\mathcal{D}(L) \bigoplus \mathcal{D}(C)$ into $\mathcal{H} \bigoplus \mathcal{H}$.

Definition 2.1. *The closed loop system determined by the plant L and compensator C is stable if all the entries of $H(L,C)$ are stable systems on $\mathcal{H}$. The plant L is stabilizable if there exists a causal linear system C such that the closed loop system determined by L and C is stable.*

In order to characterize the stabilizable and strongly stabilizable systems we need the notions of right and left strong representations for a causal linear system. Recall that the graph of a linear transformation L with domain $\mathcal{D}(L)$ in $\mathcal{H}$ is $\mathcal{G}(L) = \left\{ \begin{bmatrix} x \\ Lx \end{bmatrix} : x \in \mathcal{D}(L) \right\}$. The following definitions are from [6], Chapter 6.

Definition 2.2. *A plant L has a strong right representation $\begin{bmatrix} A \\ B \end{bmatrix}$ with A and B stable if*

(1) $\mathcal{G}(L) = Im \begin{bmatrix} A \\ B \end{bmatrix}$,

(2) $\begin{bmatrix} A \\ B \end{bmatrix}$ *has a stable left inverse: there exist X, Y stable such that*

$$\begin{bmatrix} X & Y \end{bmatrix} \begin{bmatrix} A \\ B \end{bmatrix} = I.$$

L has a strong left representation $\begin{bmatrix} -\hat{B} & \hat{A} \end{bmatrix}$ with $\hat{A}, \hat{B}$ stable if

(1) $\mathcal{G}(L) = Ker \begin{bmatrix} -\hat{B} & \hat{A} \end{bmatrix}$.

(2) $\begin{bmatrix} -\hat{B} & \hat{A} \end{bmatrix}$ *has a stable right inverse; there exist $\hat{X}, \hat{Y}$ stable such that*

$$\begin{bmatrix} -\hat{B} & \hat{A} \end{bmatrix} \begin{bmatrix} \hat{Y} \\ \hat{X} \end{bmatrix} = I.$$

The following theorem is well known (see [5], Chapter 9, [6], Chapter 6).

Theorem 2.3. *A causal linear system L is stabilizable if L has a strong right and a strong left representation. If this is the case the representations can be chosen so that the double Bezout identity*

$$\begin{bmatrix} X & Y \\ -\hat{B} & \hat{A} \end{bmatrix} \begin{bmatrix} A & -\hat{Y} \\ B & \hat{X} \end{bmatrix} = \begin{bmatrix} A & -\hat{Y} \\ B & \hat{X} \end{bmatrix} \begin{bmatrix} X & Y \\ -\hat{B} & \hat{A} \end{bmatrix} = \begin{bmatrix} I & 0 \\ 0 & I \end{bmatrix}.$$

holds. The causal linear systems which stabilize L are given by the strong right representations $\begin{bmatrix} \hat{Y} - BQ \\ \hat{X} + AQ \end{bmatrix}$ *and the strong left representation* $\begin{bmatrix} -(X + Q\hat{A} & Y - Q\hat{B} \end{bmatrix}$ *for some stable Q.*

3. Strong stabilization

Practicing control engineers are reluctant to use unstable compensators for the purpose of stabilization. This motivated considering whether among the stabilizing compensators parametrized by the Youla parametrization for a given stabilizable plant L that were described above, there exist stable ones. If there exists such a compensator, L is said to be strongly stabilizable. We continue with the setup and notation used in the previous theorem. The following theorem is from [6], Chapter 6.

Theorem 3.1. *Suppose L satisfies the hypothesis of Theorem 2.3. The L is stabilized by a stable C if and only if $\hat{A} + \hat{B}C$ is invertible in $\mathcal{C}$. Equivalently, C stabilizes L if and only if $A + CB$ is invertible in $\mathcal{C}$.*

It is interesting that while the two formulations are independent the same C will in fact work for both. (See [7], Theorem 3.2.)

Theorem 3.2. *Given a causal linear system L with doubly coprime factorization as above. For $C \in \mathcal{C}$, $\hat{A} + \hat{B}C$ is invertible in $\mathcal{C}$ if and only if $A + CB$ is invertible in $\mathcal{C}$.*

As a result of the previous discussion, we can now give a simple formulation of the strong stabilization problem: Given $A, B, X, Y \in \mathcal{C}$, which satisfy the Bezout equation $AX + BY = I$, does there exist $C \epsilon \mathcal{C}$ such that $A + CB$ is invertible in $\mathcal{C}$? It is of interest to point out that if $\mathcal{C}$ is replaced by $\mathcal{L}(\mathcal{H})$, the algebra of all bounded linear operators on $\mathcal{H}$, then the answer is negative. This follows immediately from a corollary to Theorem 2 of [2]. Our proof for $\mathcal{C}$ will use this approach.

4. Nest algebras

A nest is a chain $\mathcal{N}$ of closed subspaces of a Hilbert space which is closed under intersection and closed linear span.The nest algebra $\mathcal{T}(\mathcal{N})$ is the set of all operators T on $\mathcal{H}$ such that $TN \subset N$ for each $N \in \mathcal{N}$. $\mathcal{T}(\mathcal{N})$ is a weakly closed subalgebra of the algebra $\mathcal{B}(\mathcal{H})$ of operators on $\mathcal{H}$. The theory of nest algebras is described in [4].

Let $\mathcal{N}$ be a nest and $N \in \mathcal{N}$. Define

$$N_- = \vee\{N' \in \mathcal{N} : N' \subset N\}.$$

N_- is the immediate predecessor to N if there is one. Otherwise $N_- = N$. The subspaces $N \ominus N_-$ are called the atoms of $\mathcal{N}$. If there are no atoms, $\mathcal{N}$ is called continuous. We will be concerned here with nest algebras for continuous nests.

Consider the following continuous nests:

$$\mathcal{N} = \{L^2[0,t] : 0 \le t \le 1\},$$
$$\mathcal{M} = \{L^2[0,t] \oplus L^2[0,t] : 0 \le t \le 1\}.$$

The first is a continuous nest of subspaces in $L^2[0,1]$ (this nest, the Volterra nest, was studied by Moshe Livsic in [8]). and the second in $L^2[0,1] \oplus L^2[0,1]$. These nests are not unitarily equivalent. There doesn't exist a unitary operator U from $L^2[0,1]$ onto $L^2[0,1] \oplus L^2[0,1]$ such that $U\mathcal{N} = \{UN; N \in \mathcal{N}\}$ equals $\mathcal{M}$. They are however similar. There exists a bounded invertible operator S from $L^2[0,1]$ onto $L^2[0,1] \oplus L^2[0,1]$ such that $S\mathcal{N} = \{SN; N \in \mathcal{N}\}$ equals $\mathcal{M}$. In fact D. Larson showed that this is true for any two continuous nests ([4], Theorem 13.10):

Theorem 4.1. *Any two continuous nests on separable spaces are similar and therefore the corresponding nest algebras are isomorphic Banach algebras.*

This fact will play a fundamental role in our analysis.

5. The stable rank of $\mathcal{C}$

Let $\mathcal{A}$ be a ring with identity e, and n a positive integer. $\mathcal{A}^n$ will denote the set

$$\{(a_1, a_2, \ldots, a_n) : a_i \in \mathcal{A}, 1 \le i \le n\}.$$

Definition 5.1. *An element $a = (a_1, a_2, \ldots, a_n) \in \mathcal{A}^n$ is called (left) unimodular if there exists $b = (b_1, b_2, \ldots, b_n) \in \mathcal{A}^n$ such that $\Sigma_{k=1}^n b_k a_k = e$.*

The set of unimodular elements of $\mathcal{A}^n$ will be denoted by $\mathcal{U}_n(\mathcal{A})$.

Definition 5.2. *An element $a \in \mathcal{U}_{n+1}(\mathcal{A})$ is called (left) reducible if there exists $x = (x_1, x_2, \ldots, x_n) \in \mathcal{A}^n$ such that*

$$(a_1 + x_1 a_{n+1}, \ldots, a_n + x_n a_{n+1}) \in \mathcal{U}_n(\mathcal{A}).$$

Definition 5.3. *The (left) (Bass) stable rank of $\mathcal{A}$, denoted $sr(\mathcal{A})$, is the least integer $n \ge 1$ such that every $a \in \mathcal{U}_{n+1}(\mathcal{A})$ is reducible, and it is infinite if no such integer exists.*

It is well known ([1]) that for any ring $\mathcal{A}$, the left stable rank and right stable rank are equal, and this number is called the stable rank of $\mathcal{A}$. It was shown in [2] that for an infinite-dimensional separable Hilbert space $\mathcal{H}$, the stable rank of the algebra of bounded linear operators on $\mathcal{H}$ is infinite. The connection between the stable rank of $\mathcal{C}$, the nest algebra of stable systems and the question whether every continuous time stabilizable system is strongly stabilizable is clear. If the stable rank of $\mathcal{C}$ is greater than one, then there are stabilizable systems which are not strongly stabilizable. We will show that the stable rank of $\mathcal{C}$ is infinite.

We will need two known results. The first is the deep result of Larson ([4], Theorem 13.10) mentioned above.

The second result we need is due to Vasershtein ([14], Theorem 3)

Theorem 5.4. *Given a ring $\mathcal{A}$ with stable rank k, let $M_n(\mathcal{A})$ denote the ring of $n \times n$ matrices with entries from $\mathcal{A}$. The stable rank of $M_n(\mathcal{A})$ is $1-[-\frac{k-1}{n}]$, where $[r]$ denotes the largest integer smaller than or equal to r.*

Let $\mathcal{A}$ be a nest algebra with continuous nest $\{P_t : 0 \leq t \leq 1\}$ on the Hilbert space $\mathcal{H}$ (Any continuous nest can be parametrized in this way. See [4], Chapter 2). It is easy to see that for any positive integer n, the nest algebra associated with the nest $\{P_t \oplus P_t \oplus \cdots \oplus P_t\}$ (n copies) on the space $\mathcal{H} \oplus \mathcal{H} \oplus \cdots \oplus \mathcal{H}$ is $M_n(\mathcal{A})$. Since, by Larson's theorem, these algebras are similar and therefore isomorphic Banach algebras, they have the same stable rank for any n. It follows from Vasershtein's formula that if k is not one it must be infinite. We will show that for a nest algebra with continuous nest its stable rank can not be one and it will then follow that it is infinite.

We begin with the nest algebra $\mathcal{C}$ with continuous nest $\{P_t\}$ and consider the nest $\mathcal{N} = \{P_t \oplus P_t \oplus \cdots\}$ of countably infinitely many copies of P_t acting on the Hilbert space $l^2_+(\mathcal{H}) = \{(x_1, x_2, \dots) : \Sigma_{i=1}^{\infty} \parallel x_i \parallel^2 < \infty\}$, the space of (one-sided) square summable infinite sequences of vectors from $\mathcal{H}$. We denote a projection from this nest by Q_t. Every bounded linear operator on this space can be represented as an (one-sided) infinite matrix whose entries are bounded linear operators on $\mathcal{H}$ and which define bounded linear operators on $l^2_+(\mathcal{H})$. The nest algebra for this continuous nest is the subalgebra $\mathcal{B}$ of those one-sided infinite matrices which satisfy $Q_t A Q_t = Q_t A$. This is equivalent to requiring that each entry A_{ij} satisfy $P_t A_{ij} P_t = P_t A_{ij}$ for each t.

By Larson's theorem this algebra is isomorphic to $\mathcal{C}$.

Proposition 5.5. *The stable rank of $\mathcal{B}$ is greater than one.*

Proof. We use ideas from [12] and [2]. Let

$$U_1 = \begin{bmatrix} I & 0 & 0 & 0 & \dots \\ 0 & 0 & 0 & 0 & \dots \\ 0 & I & 0 & 0 & \dots \\ 0 & 0 & 0 & 0 & \dots \\ 0 & 0 & I & 0 & \dots \\ \vdots & \vdots & \vdots & \vdots & \ddots \end{bmatrix}, \qquad U_2 = \begin{bmatrix} 0 & 0 & 0 & 0 & \dots \\ I & 0 & 0 & 0 & \dots \\ 0 & 0 & 0 & 0 & \dots \\ 0 & I & 0 & 0 & \dots \\ 0 & 0 & 0 & 0 & \dots \\ 0 & 0 & I & 0 & \dots \\ \vdots & \vdots & \vdots & \vdots & \ddots \end{bmatrix},$$

$$V_1 = \begin{bmatrix} I & 0 & 0 & 0 & 0 & \dots \\ 0 & 0 & I & 0 & 0 & \dots \\ 0 & 0 & 0 & 0 & I & \dots \\ 0 & 0 & 0 & 0 & 0 & \dots \\ 0 & 0 & 0 & 0 & 0 & \dots \\ \vdots & \vdots & \vdots & \vdots & \vdots & \ddots \end{bmatrix}, \quad \text{and,} \quad V_2 = \begin{bmatrix} 0 & I & 0 & 0 & 0 & 0 \dots \\ 0 & 0 & 0 & I & 0 & 0 \dots \\ 0 & 0 & 0 & 0 & 0 & I \dots \\ 0 & 0 & 0 & 0 & 0 & 0 \dots \\ \vdots & \vdots & \vdots & \vdots & \vdots & \ddots \end{bmatrix}.$$

Then U_1 and U_2 are isometries and $U_i, V_i \in \mathcal{B}$ for $i = 1, 2$. Also

$$U_1V_1 + U_2V_2 = I \quad \text{and} \quad V_iU_j = \delta_{ij}I,$$

for $i, j, = 1, 2$. We show that the mapping

$$\phi(T) = [TU_1, TU_2]$$

from $\mathcal{B}$ to $\mathcal{B}^2$ is an isomorphism with inverse

$$\varphi([S, T]) = SV_1 + TV_2.$$

That ϕ is injective follows immediately from the existence of a left inverse. For surjectivity, let

$$A = \begin{bmatrix} A_{11} & A_{12} & A_{13} & \dots \\ A_{21} & A_{22} & A_{23} & \dots \\ \vdots & \vdots & \vdots & \dots \end{bmatrix},$$

and

$$B = \begin{bmatrix} B_{11} & B_{12} & B_{13} & \dots \\ B_{21} & B_{22} & B_{23} & \dots \\ \vdots & \vdots & \vdots & \ddots \end{bmatrix},$$

belong to $\mathcal{B}$ with $(A, B) \in \mathcal{B}^2$. It is easy to see that

$$T = \begin{bmatrix} A_{11} & B_{11} & A_{12} & B_{12} & \dots \\ A_{21} & B_{21} & A_{22} & B_{22} & \dots \\ \vdots & \vdots & \vdots & \vdots & \ddots \end{bmatrix},$$

defines a bounded operator and belongs to $\mathcal{B}$. Since $TU_1 = A, TU_2 = B$ we have $\phi(T) = (A, B)$.

If $sr(\mathcal{B}) = 1$, then, since $U_1V_1 + U_2V_2 = I$, there exists $C \in \mathcal{B}$ such that $V_1 + CV_2$ is left invertible in $\mathcal{B}$. If its left inverse is F, define μ on $\mathcal{B}$ by

$$\mu(T) = T(V_1 + CV_2).$$

Since $\mu(AF) = A$ for all $A \in \mathcal{B}$, μ is surjective. Now consider the mapping ρ from $\mathcal{B}$ to $\mathcal{B}^2$ defined by

$$\rho(T) = [T, TC] = T[I, C].$$

Then, for all $T \in \mathcal{B}$, we obtain

$$\begin{aligned} \varphi\rho(T) = \varphi[T, TC] &= TV_1 + TCV_2 \\ &= T(V_1 + CV_2) = \mu(T). \end{aligned}$$

Since φ is an isomorphism and μ is surjective, so is ρ. Then there exists $X \in \mathcal{B}$ such that $\rho(X) = [0, I]$. By the definition of ρ this is impossible. Thus $sr(\mathcal{B})$ is not one. □

The immediate consequence of this theorem is our main result:

Theorem 5.6. *The stable rank of $\mathcal{C}$ is infinite.*

Remark 5.7. *What happens in discrete time? In this case we consider the sequence Hilbert space $\mathcal{H} = l_2[0, \infty)$ as the input output space and the algebra of stable systems $\mathcal{C}$ is the algebra of the lower triangular (one-sided) infinite matrices which define bounded linear operators on $\mathcal{H}$. In this case Larson's theorem doesn't hold. Thus finding the stable rank in the discrete time case would require a different approach.*

6. A weak type of strong stabilization

Theorem 5.6 of the previous section shows that for continuous time the existence of a doubly coprime factorization doesn't guarantee strong stabilization. For if this was the case the stable rank of $\mathcal{C}$ would be 1. Thus the problem that must be considered is the identification of those stabilizable systems which are strongly stabilizable. In order to give a clear mathematical formulation of this problem we consider the diagonal seminorm function which has played a fundamental role in the theory of nest algebras beginning with the seminal paper of Ringrose [11]. We restrict ourselves to continuous time and take our parameter set to be the interval [0, 1].

Definition 6.1. *Given $A \in (C)$, for each $t \in (0, 1]$,*

$$i_t(A) = \lim_{h \to 0} \|(P_t - P_{t-h})A(P_t - P_{t-h})\|,$$

for $h > 0$.

If the elements of a nest algebra are an abstraction of triangular matrices with respect to a given basis on a finite-dimensional space,then i_t is an abstract way of relating to the "diagonal part" of the operators in the nest algebra. It is easy to see that, for fixed t, the function $i_t(A)$ is a seminorm on $\mathcal{C}$. Also, since i_t is a limit of norms of compressions to semi-invariant projections for the elements of $\mathcal{C}$,

$$i_t(CAB) \leq \|C\|\|B\|i_t(A)$$

for all $C, B, \in \mathcal{C}$. It is shown in [3] that for fixed $A \in \mathcal{C}$, i_t is left upper semicontinuous:

$$\limsup_{s \to t} i_s(A) \leq i_t$$

for $s < t$. The parametrization of the nest $\{L_2[0, t)\}$ gives rise to a projection-valued Borel measure $E(\cdot)$ on the interval $[0, 1]$. This has the defining property $E((0, t)) = P_t$.

The following theorem is based on [3], Lemma 2.3, Prop. 2.4. An analogous argument can be given in discrete time.

Theorem 6.2. *Suppose $A, B \in \mathcal{C}$ such that $[A, B]$ is left coprime (there exist $X, Y \in \mathcal{C}$ such that $XA + YB = I$). Then there exists $C \in \mathcal{C}$ and $\delta > 0$ such that $i_t(A + CB) \geq \delta$.*

Proof. We first show that there exists $\delta > 0$ such that $\max\{i_t(A), i_t(B)\} \geq \delta$ for all t in our interval. Since $XA+YB = I$, it follows from the semi-norm properties of i_t that

$$\|X\| i_t(A) + \|Y\| i_t(B) \geq 1$$

for all t. Thus $\max\{i_t(A), i_t(B)\} \geq \delta = (\|X\| + \|Y\|)^{-1} > 0$ for all t. Now let $U = \{t : i_t(A+B) < \delta\}$, a Borel subset of $[0,1]$. Since i_t is subadditive on $\mathcal{C}$,

$$i_t(A+B) + i_t(A-B) \geq 2\max\{i_t(A), i_t(B)\}.$$

Let C be the operator $I - 2E(U)$ (recall that $E(\cdot)$ is the projection-valued Borel measure determined by the nest). For $t \in U$, there is an $h > 0$ such that $(t-h, t) \subseteq U$, so $(P_t - P_{t-h})C = -(P_t - P_{t-h})$. Thus $i_t(A + CB) = i_t(A - B)$ which is at least δ since $i_t(A+B) < \delta$. If t is in the interior of U^c, the compliment of U, then $i_t(A + CB) = i_t(A + B) \geq \delta$. Thus $i_t(A + CB) \geq \delta$ on the dense set $U \cup U^c$ and by left upper semicontinuity, $i_t(A + CB) \geq \delta$ for all t. □

It is shown by Orr in [9] that there exist operators in $\mathcal{C}$ with the property that $i_t(A) \geq \delta$ for all t but which aren't left invertible in the algebra. In fact these operators are characterized by the following deep result of Orr ([9]):

Theorem 6.3. *$A \in \mathcal{C}$ satisfies $i_t(A) \geq \delta$ for all $t \in [0,1]$ if and only if there exist $S, T \in \mathcal{C}$ such that $SAT = I$.*

Thus it is natural to consider the question: Given $A \in \mathcal{C}$ such that $i_t(A)$ is bounded below for all t, when is A invertible in $\mathcal{C}$? As a consequence of Theorem 5.2, the answer to this question will allow us to determine when a given continuous time time-varying stabilizable linear system is strongly stabilizable.

References

[1] H. Bass, *Algebraic K-Theory*, Benjamin, New York, 1967.

[2] G. Corach, A.R. Larotonda, *Stable range in Banach algebras*, Journal of Pure and Applied Algebra, **32** (1984), 289–300.

[3] K.R. Davidson, J.L. Orr, *The invertibles are connected in infinite multiplicity nest algebras*, Bull. London Math. Soc., 27, (1995), 155–161.

[4] K.R. Davidson, *Nest Algebras*, Pitman Res. Notes Math. Ser. 191, Longman Scientific and Technical, London, 1988.

[5] A. Feintuch, R. Saeks *System Theory, A Hilbert Space Approach*, Academic Press, Series in Pure and Applied Mathematics 102, New York, London, 1982.

[6] A. Feintuch, *Robust Control theory in Hilbert Space*, Applied Mathematical Sciences 130, Springer, 1998.

[7] A. Feintuch, *On strong stabilization for linear time-varying systems*, System and Control Letters, **54**, (2005), 1091–1095.

[8] M. Livsic, *On a class of linear operators in a Hilbert space*, Mat. Sb, **19 (61)**, (1946), 239–262. A.M.S. Transl. **(2) 5**, (1957), 67–114.

[9] J.L. Orr, *Triangular Algebras and Ideals of Nest Algebras*, Memoirs of the A.M.S., **117, 562**, 1995.

[10] A. Quadrat, *On a General Structure of the Stabilizing Controllers Based on a Stable Range*, SIAM J. Cont. and Optim., **24, 6**, (2004) 2264–2285.

[11] J.R. Ringrose, *On some algebras of operators*, Proc. London Math. Soc., **15, 3** (1965), 61–83.

[12] A. Sasane, *Stable ranks of Banach algebras of operator valued analytic functions*, Complex Analysis and Operator Theory, to appear (issue 3-1, 2009).

[13] S. Treil, *The Stable Rank of the Algebra H^∞ Equals* 1, J. Funct. Anal., **109** (1992), 130–154.

[14] L.N. Vasershtein, *Stable ranks of rings and dimensionality of topological spaces*, Func. Anal. Applic. **5** 1972, 102–110.

[15] M. Vidyasagar, *Control System Synthesis: A Factorization Approach*, MIT Press, Cambridge Mass., 1985.

Avraham Feintuch
Department of Mathematics
Ben Gurion University of the Negev
POB 653
84105 Beer-Sheva, Israel
e-mail: abie@math.bgu.ac.il

Operator Theory:
Advances and Applications, Vol. 197, 149–179

Convexity of Ranges and Connectedness of Level Sets of Quadratic Forms

I. Feldman, N. Krupnik and A. Markus

Dedicated to the memory of Moshe Livsic,
a great mathematician and a great human being

Abstract. O. Toeplitz and F. Hausdorff proved that the range of any quadratic form on the unit sphere S of an inner product space X is convex and the level sets of any Hermitian form on S are connected. We consider the question: Which subsets of X, besides S, have these properties?

Mathematics Subject Classification (2000). Primary 15A63, 47A12.

Keywords. Quadratic forms, connectedness, convexity, numerical range, Toeplitz-Hausdorff Theorem.

1. Introduction

1. Let A be a complex $n \times n$ matrix and (Af, f) be the corresponding quadratic form on the space $\mathbb{C}^n$. In 1918, O. Toeplitz [T] considered the range $W(A)$ of this quadratic form on the unit sphere S and proved that the outer boundary of $W(A)$ is a convex curve. He also suggested that the set $W(A)$ by itself is convex. This conjecture was proved in 1919 by F. Hausdorff [H]. As an auxiliary result he proved that, in the case of Hermitian form (Af, f), all its level sets on S are connected.

The results of Toeplitz and Hausdorff were developed and generalized by many authors. This area continues to attract the attention of researchers, and one of the main reasons is its numerous applications in various domains. Some information on the contemporary state of this direction can be found in the book [GR]. See also the papers [FM] and [G] intended for the general reading public.

This paper is devoted to a systematic study of the following question: Which sets, besides the unit sphere, have the properties of convexity and connectedness discovered by Toeplitz and Hausdorff? We consider an arbitrary inner product (pre-Hilbert) space X. In the question that we consider, there is no substantial

difference neither between the finite and infinite dimensions, nor between complete and incomplete spaces (if $\dim X = \infty$). This is the reason that we sometimes restrict ourselves to the case $\dim X < \infty$, even if the generalization to the infinite dimension does not require much effort. On the contrary, some essential differences arise sometimes between the cases $\dim X = 1$ and $\dim X > 1$ (Section 7), and more often – between the cases $\dim X = 2$ and $\dim X > 2$ (Sections 3, 4, 5).

Consider a set $M \subset X$. If the level sets of arbitrary Hermitian form of the set M are connected, we say that M has the Hausdorff property (or is an H-set). If the range of arbitrary quadratic form on the set M is convex, we say that M has the Toeplitz-Hausdorff property (or is a TH-set).

Some examples of TH- and H-sets appeared (explicitly or implicitly) in various papers. Without any claim on completeness, we mention here [K], [AP], [LTU], [LM], [GM], [BM].

2. The paper contains 7 sections (including this Introduction). In the next section, we consider some examples and prove some properties of TH- and H-sets. In particular, we prove that each H-set is also a TH-set and give various examples of TH-sets which are not H-sets.

Section 3 is devoted to the case $\dim X = 2$. Let X be realized as $\mathbb{C}^2$. A set $M \subset \mathbb{C}^2$ is called *bicircular* (or *toroidal*) if it contains, together with each its vector (u, v), all vectors of the form $(z_1 u, z_2 v)$ where $|z_1| = |z_2| = 1$. We prove that for any bicircular set M the Hausdorff property and the Toeplitz-Hausdorff property are equivalent. We give also a suitable test for these properties (Theorem 3.5).

In Section 4, we discuss the similarity and difference between the cases $\dim X = 2$ and $\dim X > 2$. We prove that in any dimension the annulus

$$\{f \in X : r \le \|f\| \le R\} \tag{1.1}$$

is an H-set. In addition, we consider sets defined through ℓ_p norms in $\mathbb{C}^n$ (for simplicity, we restrict ourselves to finite dimension). We give our main attention to the ℓ_p unit balls in $\mathbb{C}^n$. We show that there is an essential difference between the cases $n = 2$ and $n > 2$. It follows also from these results that the test for TH- and H-properties of bicircular sets does not admit generalization to the case $n > 2$.

In Section 5, we prove the Hausdorff property for various subsets of the unit sphere S. In particular, we show that if $\dim X > 2$ and G is a selfadjoint operator, then the set

$$\{f \in S : p_1 \ge (Gf, f) \ge p_2\} \quad (p_1, p_2 \in \mathbb{R})$$

is an H-set. Some unbounded H-sets also are considered in this section.

In Section 6, we bring some applications. At first we prove some operator inequalities. We formulate one of them. Let A and B be two operators in X ($\dim X > 2$). Suppose that B is an indefinite selfadjoint operator and that

$$(Bf, f) = 0 \Rightarrow (Af, f) \ne 0 \quad (f \ne 0).$$

Then there exist real numbers θ and μ such that $\mathrm{Re}(\exp(i\theta)A) \ge \mu B$.

The last section is connected with the recent paper by Guttierrez and de Merdano [GM]. They considered the ranges and level sets on the sphere for arbitrary (non-homogeneous in general) quadratic functionals. We generalize their results to the case of the annulus (1.1). Exactly as in [GM], this allows us to show that an arbitrary shift of the annulus (1.1) is an H-set.

2. Definitions, examples and first results

1. Let X be a complex vector space provided with a positive definite inner product (f, g) (pre-Hilbert space). We denote by $\mathcal{L}(X)$ the set of all linear operators in X bounded with respect to the norm $\|f\| = (f, f)^{1/2}$. Denote by S the unit sphere in X :

$$S = \{f \in X : \|f\| = 1\}.$$

Let M be a set in X and $A \in \mathcal{L}(X)$. Denote

$$W(A, M) = \{(Af, f) : f \in M\}.$$

In other words, $W(A, M)$ is the range of the quadratic form (Af, f) on the set M.

Instead of $W(A, S)$ we will use the usual notation $W(A)$. We say that the set M has the *Toeplitz-Hausdorff property* (in short, M is a TH*-set*) if the set $W(A, M)$ is a convex subset of $\mathbb{C}$ for arbitrary $A \in (X)$.

For a selfadjoint operator $B \in \mathcal{L}(X)$ and for a real number p, denote

$$Z(B, M, p) = \{f \in M : (Bf, f) = p\}.$$

In other words, the sets $Z(B, M, p)$ are the level sets on the Hermitian form (Bf, f) on the set M.

We say that the set M has the *Hausdorff property* (in short, M is an H*-set*) if the sets $Z(B, M, p)$ are connected for all $p \in \mathbb{R}$ and for all selfadjoint operators $B \in \mathcal{L}(X)$.

Obviously, the empty set is both a TH- and an H-set. The simplest example of nonempty TH- and H-sets gives the singleton $\{h\}$. Before considering some substantive examples, we prove a statement on the connection between the two main subjects of this paper. We use here the Halmos approach [Ha, pp. 314–315] to the proof of the Teoplitz-Hausdorff Theorem.

Theorem 2.1. *Every H-set is also a TH-set.*

Proof. We identify $\mathbb{C}$ with $\mathbb{R}^2$. To prove the convexity of a set $W(A, M)$ ($A \in \mathcal{L}(X)$), it is enough to show that the intersection Δ of $W(A, M)$ with any straight line $ax + by = c$ is connected. Let $G_1 = \frac{1}{2}(A + A^*)$, $G_2 = \frac{1}{2i}(A - A^*)$ and $B = aG_1 + bG_2$. It is easy to see that a point (x, y) belongs to Δ if and only if $(Bf, f) = c$. Since M is an H-set, the set $Z(B, M, c)$ is connected. But the set Δ is the continuous image of $Z(B, M, c)$ under the mapping $f \to ((G_1 f, f), (G_2 f, f))$. Hence Δ also is connected. □

The next statement gives a simple necessary condition for the Hausdorff property.

Proposition 2.2. *If M is an H-set, then it is connected.*

Indeed, if we put $B = 0$ and $p = 0$, we obtain $Z(0, M, 0) = M$.

For TH-sets we can prove only a weaker statement.

Proposition 2.3. *If M is a TH-set, then*

(1) *for any $h \in X$ the set $\{|(f, h)| : f \in M\}$ is connected;*
(2) *the set $\{\|f\| : f \in M\}$ is connected.*

Proof. If $A = (\cdot, h)h$ then the set $W(A, M) = \{|(f, h)|^2 : f \in M\}$ is convex, and this implies (1). For (2) it is enough to consider $A = I$ (the identity operator). □

Proposition 2.2 allows to give an example of a TH-set which is not an H-set.

Example 2.4. Let $f_0 \neq 0$ and $M = \{f_0, -f_0\}$. For each $A \in \mathcal{L}(X)$ the set $W(A, M)$ is a singleton $\{(Af_0, f_0)\}$, and hence M is a TH-set. But M is disconnected, and by Proposition 2.2 it not an H-set.

We show below that it can happen that even a connected TH-set is not an H-set (see Examples 2.16–2.18).

Now we consider the classical example of an H- and a TH-set, namely, the unit sphere S. These results were obtained by Toeplitz and Hausdorff, and we prove them for the convenience of the reader.

Theorem 2.5. (Hausdorff) *S is an H-set.*

Proof. Since $Z(B, S, p) = Z(B - pI, S, 0)$, it is enough to prove the connectedness of $Z(B, S, 0)$ for arbitrary selfadjoint B.

Let $f, g \in S$ and $(Bf, f) = (Bg, g) = 0$. If f and g are linearly dependent, then $g = e^{i\theta} f$ for some $\theta \in \mathbb{R}$, and $s_0(t) = e^{it\theta} f$ $(0 \leq t \leq 1)$ is a path in $Z(B, S, 0)$ connecting f and g. So, we can consider the case when f and g are linearly independent.

Choose $\alpha \in \mathbb{R}$ such that $\mathrm{Re}\{e^{i\alpha}(Bf, g)\} = 0$ and denote $h = e^{i\alpha} f$. Let

$$s(t) = \frac{th + (1 - t)g}{\|th + (1 - t)f\|} \quad (0 \leq t \leq 1).$$

It is easy to see that $(Bs(t), s(t)) \equiv 0$, and hence $s(t)$ is a path in $Z(B, S, 0)$ connecting h and g. Since the path $s_1(\tau) = e^{i\tau\alpha} f$ $(0 \leq \tau \leq 1)$ connects f and h, the vectors f and g are path connected in $Z(B, S, 0)$. □

Theorems 2.1 and 2.5 imply

Theorem 2.6. (Toeplitz-Hausdorff) *The set $W(A)$ is convex for any $A \in \mathcal{L}(X)$ (i.e., S is a TH-set).*

Now consider the simplest case of a one-dimensional space (i.e., $X = \mathbb{C}$).

Example 2.7. Let $\dim X = 1$.

(a) The set $M \subset X$ is a TH-set if and only if the set $\{|z| : z \in M\}$ is connected.
(b) The set M is an H-set if and only if it is connected and the set

$$M \cap \{re^{i\theta} : -\pi \le \theta \le \pi\}$$

is connected for each $r > 0$.

Proof. (a) M is a TH-set if and only if for each $a \in \mathbb{C}$ the set $\{a|z|^2 : z \in M\}$ is convex. This is equivalent to the connectedness of $\{|z| : z \in M\}$.

(b) We have to check the connectedness of the sets

$$Z(b, M, p) = \{z \in M : b|z|^2 = p\}$$

for arbitrary $b, p \in \mathbb{R}$. If at least one of the numbers b, p equals 0, then the set $Z(b, M, p)$ is either M, or $M \cap \{0\}$, or $\emptyset$. If $b, p \neq 0$ and $\frac{p}{b} < 0$, then $Z(b, M, p) = \emptyset$. Finally, if $\frac{p}{b} > 0$, then $Z(b, M, p) = M \cap \left\{\sqrt{\frac{p}{b}}e^{i\theta} : 0 \le \theta \le 2\pi\right\}$. □

2. Here we bring some simple properties of TH- and H-sets.

Proposition 2.8. *If M is a TH-set (resp. H-set) and $D \in \mathcal{L}(X)$, then DM also is a TH-set (resp. H-set).*

Proof. It follows from $(ADf, Df) = (D^*ADf, f)$ that $W(A, DM) = W(D^*AD, M)$ and $Z(B, DM, p) = Z(D^*BD, M, p)$. These equalities imply both statements. □

Taking $D = zI$ and $D = \operatorname{diag}[a_1, \ldots, a_n]$ in Proposition 2.8, we obtain the following two statements.

Corollary 2.9. *If M is a TH-set (resp. H-set) and $z \in \mathbb{C}$, then zM also is a TH-set (resp. H-set).*

Corollary 2.10. *Let M $(\subset \mathbb{C}^n)$ be a TH-set (resp. H-set) and $a_k \in \mathbb{C}$. Then the set*

$$M_1 = \{(a_k u_k)_1^n : (u_k)_1^n \in M\}$$

also is a TH-set (resp. H-set).

Proposition 2.11. *Let X_0 be a subspace of X and $M \subset X_0$. Then M is a TH-set (resp. H-set) in X if and only if M is a TH-set (resp. H-set) in X_0.*

Proof. Let P $(\in \mathcal{L}(X))$ be the orthogonal projection on X_0. If $A \in \mathcal{L}(X)$, $B = B^* \in \mathcal{L}(X)$ and $f \in M$, then

$$W(A, M) = W(PAP|X_0, M), \quad Z(B, M, p) = Z(PBP|X_0, M, p).$$

Since $\{PAP|X_0 : A \in \mathcal{L}(X)\} = \mathcal{L}(X_0)$, the statements follow. □

Proposition 2.12. *If M is a TH-set (resp. H-set) and $P \in \mathcal{L}(X)$ is a projection on a subspace X_0, then the set PM is a TH-set (resp. H-set) in X_0.*

Proof. Follows from Propositions 2.8 and 2.11. □

Proposition 2.13. *Let $(f,g)_1$ be a new inner product in X equivalent to the initial inner product (f,g). Then a set M is a TH-set (resp. H-set) with respect to $(f,g)_1$ if and only if it is a TH-set (resp. H-set) with respect to (f,g).*

Proof. There exists an invertible operator $G \in \mathcal{L}(X)$ such that

$$(f,g)_1 = (Gf, Gg) \quad (f,g \in X).$$

Hence $(Af,f)_1 = p$ if and only if $(G^*AGf, f) = p$. If we denote the sets $W(\cdot,\cdot)$ and $Z(\cdot,\cdot,\cdot)$ with respect to $(f,g)_1$ by $W_1(\cdot,\cdot)$ and $Z_1(\cdot,\cdot,\cdot)$, we obtain

$$W_1(A,M) = W(G^*AG, M), \quad Z_1(B,M,p) = Z(G^*BG, M, p),$$

and the statements follow. □

3. Here we consider one simple but useful example.

Let $f, g, \in X$. We will use the notation $[f,g]$ (or $[g,f]$) for the closed interval joining f and g, i.e.,

$$[f,g] = \{tf + (1-t)g : 0 \le t \le 1\}. \tag{2.1}$$

Proposition 2.14. *The set $M = [f,g]$ is a TH-set if and only if the vectors f and g are linearly dependent.*

Proof. If f and g are linearly independent, then there exists a pair $\{\tilde{f}, \tilde{g}\}$ which is biorthogonal to the pair $\{f,g\}$, i.e.,

$$(f,\tilde{f}) = (g,\tilde{g}) = 1, \qquad (f,\tilde{g}) = (g,\tilde{f}) = 0.$$

Consider the operator $A = (\cdot,\tilde{f})\tilde{f} + i(\cdot,\tilde{g})\tilde{g}$. Then

$$(A(tf + (1-t)g), tf + (1-t)g) = t^2 + i(1-t)^2,$$

and $W(A,M)$ is the curve $z = t^2 + i(1-t^2)$ $(0 \le t \le 1)$. Obviously, $W(A,M)$ is not convex, and hence M is not a TH-set.

Let the vectors f, g be linearly dependent. The case $f = g = 0$ is trivial, and we suppose for definiteness that $g \ne 0$ and $f = \lambda g$ $(\lambda \in \mathbb{C})$. Then for any $A \in \mathcal{L}(X)$

$$(A(tf + (1-t)g), tf + (1-t)g) = |(\lambda - 1)t + 1|^2 (Ag, g).$$

The set $\{|(\lambda-1)t+1|^2 : 0 \le t \le 1\}$ is a segment in $[0,\infty)$, and hence $W(A,M)$ is convex for each $A \in \mathcal{L}(X)$. Hence M is a TH-set. □

Now we determine when the set $[f,g]$ is an H-set. By Proposition 2.14 and Theorem 2.1 this can happen only if f and g are linearly dependent. Since the case $f = g = 0$ is trivial, we can suppose that $g \ne 0$ and then $f = \lambda g$ $(\lambda \in \mathbb{C})$.

Proposition 2.15. *Let $f = \lambda g$ and $M = [f,g]$. The set M is an H-set if and only if the number λ belongs to the union of the half-plane $\operatorname{Re}\lambda \ge 1$ and the disk $\left|\lambda - \frac{1}{2}\right| \le \frac{1}{2}$.*

Proof. Let $B = B^* \in \mathcal{L}(X)$ and $p \in \mathbb{R}$. Since

$$(B(tf + (1-t)g), tf + (1-t)g) = |(\lambda - 1)t + 1|^2 (Bg, g),$$

the set $Z(B, M, p)$ is the set of all vectors of the form $tf + (1-t)g$ where t ($\in [0,1]$) is a solution of the equation

$$|(\lambda - 1)t + 1|^2 (Bg, g) = p. \tag{2.2}$$

If $(Bg, g) = 0$ then this equation either has no solutions or each $t \in [0,1]$ is a solution of (2.2), and the choice depends on p ($p \neq 0$ or $p = 0$). Obviously, in both cases $Z(B, M, p)$ is connected. Let $(Bg, g) \neq 0$. Denote $\sigma = \operatorname{Re}\lambda$, $\tau = \operatorname{Im}\lambda$ and rewrite (2.2) in the form

$$((\sigma - 1)t + 1)^2 + \tau^2 t^2 = \frac{p}{(Bg, g)}$$

or

$$((\sigma - 1)^2 + \tau^2)t^2 + 2(\sigma - 1)t = \frac{p}{(Bg, g)} - 1. \tag{2.3}$$

In the case $\lambda = 1$, the set $M = \{g\}$ is an H-set. Suppose that $\lambda \neq 1$. Then equation (2.3) (with respect to t) cannot have more than 2 solutions. Obviously, the set $Z(B, M, p)$ is connected if and only if the equation (2.3) has at most one solution on the segment $[0,1]$. The left-hand side of (2.3) is a monotone function on $[0,1]$ if and only if

$$\frac{1 - \sigma}{(\sigma - 1)^2 + \tau^2} \notin (0,1). \tag{2.4}$$

This implies that the set $Z(B, M, p)$ is connected for all $p \in \mathbb{R}$ if and only if the condition (2.4) holds. It is easy to check that this condition holds if and only if either $\sigma \geq 1$ or $\sigma^2 + \tau^2 - \sigma \leq 0$. □

From Propositions 2.14 and 2.15, we see that for $g \neq 0$ and for λ that does not satisfy the conditions of Proposition 2.15, the set $[\lambda g, g]$ is a TH-set but not an H-set. In particular, for $\lambda = -1$ we obtain the following simple example.

Example 2.16. Let $g \neq 0$ and

$$M = \{sg : -1 \leq s \leq 1\}.$$

Then the set M is a TH-set but not an H-set.

It is worth noting that by Proposition 2.11, it is enough to prove Proposition 2.15 for the case $\dim X = 1$, and then we can use Example 2.7. However, this approach to the proof of Proposition 2.15 also needs some considerations.

4. Example 2.16 is suitable for arbitrary $\dim X \geq 1$. Here we give two more examples of TH-sets which are not H-sets. The first relates to the case $\dim X = 2$, and the second to the case $\dim X > 2$.

Example 2.17. Let M ($\subset \mathbb{C}^2$) be defined by the equality

$$M = \{(x, y) : x, y \in \mathbb{R},\ x^2 + y^2 \le 1\}.$$

Consider also the set

$$M_0 = \{(x, y) : x, y \in \mathbb{R},\ x^2 + y^2 = 1\}.$$

Let us prove first that for any matrix $A \in \mathbb{C}^{2\times 2}$ the set $W(A, M_0)$ is an ellipse in $\mathbb{C} = \mathbb{R}^2$. Since $W(A - \lambda I, M_0) = W(A, M_0) - \lambda$ for any $\lambda \in \mathbb{C}$, we can consider here the case when $\operatorname{tr} A = 0$, i.e.,

$$A = \begin{bmatrix} a & b \\ c & -a \end{bmatrix}.$$

Each vector $g \in M_0$ can be written in the form $g = (\cos\theta, \sin\theta)$ for some $\theta \in \mathbb{R}$. Then

$$(Ag, g) = a(\cos^2\theta - \sin^2\theta) + (b + c)\cos\theta\sin\theta.$$

If $a = a_1 + ia_2$, $\frac{1}{2}(b + c) = d_1 + id_2$, then

$$(Ag, g) = a_1\cos 2\theta + d_1 \sin 2\theta + i(a_2\cos 2\theta + d_2 \sin 2\theta).$$

Denote $\sigma = a_1\cos 2\theta + d_1 \sin 2\theta$, $\tau = a_2\cos 2\theta + d_2 \sin 2\theta$. Eliminating the parameter θ, we obtain that

$$W(A, M_0) = \{(\sigma, \tau) \in \mathbb{R}^2 : (a_2\sigma - a_1\tau)^2 + (d_2\sigma - d_1\tau)^2 = (a_1d_2 - a_2d_1)^2\},$$

and this is an ellipse (maybe, degenerate).

Since the general form of a vector $f \in M$ is $f = rg$ where $g \in M_0$ and $r \in [0, 1]$, the set $W(A, M)$ is the convex hull of $W(A, M_0)$ and the point $(0, 0)$. This proves that M is a TH-set.

Consider now the matrix $B = \left[\begin{smallmatrix} 0 & 0 \\ 0 & 1 \end{smallmatrix}\right]$. It is easy to check that the set $Z(M, B, 1)$ is disconnected:

$$Z(M, B, 1) = \{(0, 1), (0, -1)\}.$$

Hence M is not an H-set.

The next example is based on a result of Brickman [B].

Example 2.18. Consider the set

$$M = \left\{(x_1, \dots, x_n) \in \mathbb{R}^n : \sum_{k=1}^{n} x_k^2 = 1\right\}.$$

in $\mathbb{C}^n$ ($n \ge 3$). If $A = \operatorname{diag}[1, 0, \dots, 0]$, then

$$Z(A, M, 1) = \{(1, 0, \dots, 0), (-1, 0, \dots, 0)\}$$

and hence M is not an H-set.

On the other hand, if C is an arbitrary complex $n \times n$-matrix, then

$$W(C, M) = W\left(\frac{1}{2}(C + C^T)\right)$$

where C^T is the transpose (see [B, Corollary]). Hence from Theorem 2.6 it follows that M is a TH-set.

3. Two-dimensional case

1. In this section we assume that $X = \mathbb{C}^2$. To each subset $M \subset \mathbb{C}^2$ we assign the following subset of the first quadrant $\mathbb{R}^2_+$ of the real plane $\mathbb{R}^2$:

$$K(M) := \{(|u|^2, |v|^2) : (u, v) \in M\}.$$

Lemma 3.1. *If M is a TH-set, then the set $K(M)$ is convex.*

Proof. If $A = \operatorname{diag}[1, i]$, then

$$W(A, M) = \{|u|^2 + i|v|^2 : (u, v) \in M\},$$

and this set must be convex. Identifying $\mathbb{C}$ and $\mathbb{R}^2$, we have $W(A,M)=K(M)$. □

The inverse statement is not true. The next lemma gives (for a wide class of sets M) an additional necessary condition to be a TH-set. First of all we define this class.

We say that a set $M \subset \mathbb{C}^2$ is *bicircular* if for each vector $(u, v) \in M$ all vectors of the form $(e^{i\alpha}u, e^{i\beta}v)$ $(\alpha, \beta \in \mathbb{R})$ also belong to M.

Lemma 3.2. *If M is a nonempty bicircular TH-set, then the set $K(M)$ contains at least one point (x', y') such that $x'y' = 0$.*

Proof. Since the set M is bicircular,

$$M = \{(\sqrt{x}e^{i\alpha}, \sqrt{y}e^{i\beta}) : (x, y) \in K(M); \alpha, \beta \in \mathbb{R}\}.$$

If $A = \left[\begin{smallmatrix} 0 & 0 \\ 1 & 0 \end{smallmatrix}\right]$, then

$$W(A, M) = \{\sqrt{xy}\exp(i\gamma) : (x, y) \in K(M), \gamma \in \mathbb{R}\}.$$

Since the set $W(A, M)$ is convex, it must contain the point 0. □

The main result of this section is the proof that for a bicircular set M two necessary conditions of Lemmas 3.1 and 3.2 are together sufficient. We prove also that for the bicircular sets the H- and TH-properties are equivalent.

2. Let M $(\subset \mathbb{C}^2)$ be a bicircular set, B be a selfadjoint operator and $p \in \mathbb{R}$. Since the set M is invariant with respect to all diagonal unitary operators, we may assume that

$$B = \begin{bmatrix} a & c \\ c & b \end{bmatrix} \tag{3.1}$$

where $a, b, c \in \mathbb{R}$. Then

$$Z(B, M, p) = \{(\sqrt{x}\exp(i\varphi), \sqrt{y}\exp(i\psi)) : (x, y) \in K(M);\ \varphi, \psi \in \mathbb{R};\\ ax + by + 2c\sqrt{xy}\cos(\varphi - \psi) = p\}.$$

We start with a technical lemma.

Lemma 3.3. *Suppose that M is a bicircular set, the set $K(M)$ is convex and it contains a point (x', y') such that $x'y' = 0$. Then for each vector $f_0 \in Z(B, M, p)$, there exists a path connecting f_0 within $Z(B, M, p)$ with a vector $f_1 = (\sqrt{x}\exp(i\beta), \sqrt{y})$ where $(x, y) \in K(M)$ and $\beta \in [0, \pi]$.*

Note that the main point here is that $\beta \in [0, \pi]$ (and not only real). It is easy to see from the proof that we can even claim that β equals either 0 or π, but this is not relevant for us.

Proof. Let $f_0 = (\sqrt{x}_0 \exp(i\varphi), \sqrt{y_0} \exp(i\psi))$. If $c = 0$ then

$$(\sqrt{x_0} \exp(i\varphi t), \sqrt{y_0} \exp(i\psi t)) \ (0 \le t \le 1)$$

gives the required path. Now we mention that the path

$$s(t) := (\sqrt{x_0} \exp(i(\varphi - t\psi)), \sqrt{y_0} \exp(i(\psi - t\psi)) \ (0 \le t \le 1)$$

connects f_0 within $Z(B, M, p)$ with the vector $s(1) = (\sqrt{x_0} \exp(i\alpha), \sqrt{y_0})$ ($\alpha = \varphi - \psi$). It is easy to see that in the case $x_0 y_0 = 0$ this implies the statement of the lemma.

Below we assume that $c \neq 0$ and $x_0 y_0 \neq 0$. Of course, it is enough to consider only the case $\alpha \in (-\pi, 0)$.

For definiteness, we suppose that $(x', y') = (0, d)$ ($d \ge 0$). Let $y = kx + d$ be the equation of the straight line which passes through the points $(0, d)$ and (x_0, y_0). Then the condition $s(1) \in Z(B, M, p)$ can be written as

$$ax_0 + b(kx_0 + d) + 2c\sqrt{x_0(kx_0 + d)} \cos\alpha = p. \tag{3.2}$$

Define

$$h(x) = \frac{p - bd - (a + bk)x}{2c\sqrt{x(kx + d)}} \quad (0 < x \le x_0). \tag{3.3}$$

Since $x_0 y_0 > 0$, the function $h(x)$ is well defined. It follows from (3.2) that $h(x_0) = \cos\alpha$.

If $|h(x)| \le 1$ for some $x \in (0, x_0]$, we define

$$\alpha(x) = -\arccos h(x), \tag{3.4}$$

$$g(x) = (\sqrt{x} \exp(i\alpha(x)), \sqrt{kx + d}). \tag{3.5}$$

Since $(0, d), (x_0, y_0) \in K(M)$ and $K(M)$ is convex, $(x, kx + d) \in K(M)$ for all $x \in [0, x_0]$. It follows from (3.2)–(3.5) that $g(x) \in Z(B, M, p)$ for all x such that $|h(x)| \le 1$. Observe also that the definition (3.4) and the condition $\alpha \in (-\pi, 0)$ imply $g(x_0) = s(1)$.

Consider two cases.

1°. $|h(x)| < 1$ for all $x \in (0, x_0)$. Then the vector function $g(x)$ is defined on $(0, x_0]$. It is easy to check that $\lim\limits_{x \to 0} g(x) = (0, \sqrt{d})$, and hence we have a path connecting $g(x_0) = s(1)$ and $(0, \sqrt{d})$ within $Z(B, M, p)$. Of course, the vector $f_1 = (0, \sqrt{d})$ has the required form.

2°. There exists a number $x \in (0, x_0)$ such that $|h(x)| \ge 1$. In this case we denote

$$x_1 = \max\{x < x_0 : |h(x)| = 1\}.$$

Then $|h(x)| < 1$ on $(x_1, x_0]$, $|h(x_1)| = 1$ and the path $g(x)$ ($x_1 \le x \le x_0$) connects $g(x_0) = s(1)$ and

$$g(x_1) = (\sqrt{x_1} \exp(i\alpha(x_1)), \sqrt{kx_1 + d}).$$

It is easy to see that $\exp(i\alpha(x_1)) = \pm 1$, and in both cases the vector $f_1 = g(x_1)$ has the required form. □

3. Lemma 3.3 enables us to obtain the following result.

Lemma 3.4. *If a set M satisfies the conditions of Lemma* 3.3*, it is an H-set.*

Proof. We have to prove the connectedness of the set $Z(B, M, p)$ for any operator (3.1) and any $p \in \mathbb{R}$. By Lemma 3.3 it is enough to find a path in $Z(B, M, p)$ which joins two vectors of the form

$$f_1 = (\sqrt{x_1}\exp(i\alpha), \sqrt{y_1}), \qquad f_2 = (\sqrt{x_2}\exp(i\beta), \sqrt{y_2})$$

where $(x_j, y_j) \in K(M)$ $(j = 1, 2)$ and $\alpha, \beta \in [0, \pi]$. Define for $t \in [0, 1]$

$$x(t) = tx_2 + (1-t)x_1, \qquad y(t) = ty_2 + (1-t)y_1,$$

$$q(t) = \frac{(1-t)\sqrt{x_1y_1}\cos\alpha + t\sqrt{x_2y_2}\cos\beta}{\sqrt{x(t)y(t)}}.$$

Consider first the case when $x_1y_1x_2y_2 \neq 0$. Then $x(t)$ and $y(t)$ do not vanish and therefore $q(t)$ is well defined.

We need the inequality

$$(1-t)\sqrt{x_1, y_1} + t\sqrt{x_2y_2} \leq \sqrt{x(t)y(t)} \tag{3.6}$$

which can be reduced by squaring to the obvious inequality

$$2\sqrt{x_1y_1x_2y_2} \leq x_1y_2 + x_2y_1.$$

It is easy to see that (3.6) implies $|q(t)| \leq 1$ $(t \in [0, 1])$, and we can define

$$\gamma(t) = \arccos q(t), \quad f(t) = (\sqrt{x(t)}\exp(i\gamma(t)), \sqrt{y(t)}). \tag{3.7}$$

It can be directly checked that

$$ax(t) + by(t) + 2c\sqrt{x(t)y(t)}\cos\gamma(t) = (1-t)(Bf_1, f_1) + t(Bf_2, f_2) = p.$$

Using the condition $\alpha, \beta \in [0, \pi]$, we obtain that $f(0) = f_1$, $f(1) = f_2$. So $f(t)$ $(0 \leq t \leq 1)$ is a path joining f_1 and f_2 within $Z(B, M, p)$.

We turn now to the cases when some of the numbers x_1, x_2, y_1, y_2 equal 0.

If one of the vectors f_1, f_2 (say, f_1) equals 0, then $p = 0$ and the path $f(t) = tf_2$ solves the problem.

If $x_1 = x_2 = 0$, we set $f(t) = (0, \sqrt{y(t)})$.

If $y_1 = y_2 = 0$, then the path

$$(\sqrt{x_1}\exp(i(\alpha - t\alpha)), 0) \tag{3.8}$$

joins f_1 and $(\sqrt{x_1}, 0)$, and then the path

$$(\sqrt{x_2}\exp(i(\alpha - t\alpha)), 0)$$

joins f_2 and $(\sqrt{x_2}, 0)$. Now use the path $(\sqrt{x(t)}, 0)$.

If $y_1 = x_2 = 0$, the path (3.8) joins f_1 and $(\sqrt{x_1}, 0)$. Now define

$$g(t) = (i\sqrt{(1-t)x_1}, \sqrt{ty_2}).$$

It is easy to check that

$$(Bg(t), g(t)) = a(1-t)x_1 + bty_2 + 2c\sqrt{t(1-t)x_1y_2}\cos\frac{\pi}{2} = p,$$

and hence the path $g(t)$ joins $i(\sqrt{x_1}, 0)$ and f_2 within $Z(B, M, p)$. The case $x_1 = y_2 = 0$ is completely analogous.

It remains to consider the case when exactly one of the numbers x_1, x_2, y_1, y_2 equals 0. If $x_1 = 0$, we use the path $f(t)$ (see (3.7)) where

$$q(t) := \frac{\sqrt{ty_2}\cos\beta}{\sqrt{y(t)}}.$$

If $y_1 = 0$, we first take the path (3.8) to pass to the vector $(\sqrt{x_1}, 0)$, and then use the path $f(t)$ where

$$q(t) := \frac{\sqrt{tx_2}\cos\beta}{\sqrt{x(t)}}.$$

The cases $x_2 = 0$ and $y_2 = 0$ are completely analogous. □

4. Now we are able to obtain the main result of this section.

Theorem 3.5. *For any bicircular set M, the following properties are equivalent:*

(a) *M is an H-set;*
(b) *M is a TH-set;*
(c) *The set $K(M)$ is convex and contains at least one point (x', y') such that $x'y' = 0$.*

Proof. (a) implies (b) by Theorem 2.1; (b) implies (c) by Lemmas 3.1 and 3.2. Finally, (c) implies (a) by Lemma 3.4. □

Theorem 3.5 has a number of consequences. We start with several statements connected with ℓ_p norms (quasinorms, if $p < 1$) in $\mathbb{C}^2$. Recall the definition of $\|\cdot\|_p$ in $\mathbb{C}^n$:

$$\|f\|_p = \left(\sum_{k=1}^{n} |f_k|^p\right)^{1/p} \quad (0 < p < \infty), \qquad \|f\|_\infty = \max_{1\le k\le n} |f_k|.$$

Corollary 3.6. *Let $0 < r \le R < \infty$. The set*

$$\{(u, v) \in \mathbb{C}^2 : r \le \|(u, v)\|_p \le R\} \tag{3.9}$$

is a TH-set (H-set) if and only if $p = 2$.

The next two statements show that the restrictions $r > 0$ and $R < \infty$ in Corollary 3.6 are essential.

Corollary 3.7. *The set*

$$\{(u, v) \in \mathbb{C}^2 : \|(u, v)\|_p \le 1\} \tag{3.10}$$

is a TH-set (H-set) if and only if $p \ge 2$.

Corollary 3.8. *The set*

$$\{(u,v) \in \mathbb{C}^2 : \|(u,v)\|_p \geq 1\} \tag{3.11}$$

is a TH-set (H-set) if and only if $p \leq 2$.

The next statement is a generalization of Corollary 3.6.

Corollary 3.9. *Let* $0 < r_1,\ r_2 < \infty$, $0 \leq p_1 \leq p_2 \leq \infty$. *The set*

$$\{(u,v) \in \mathbb{C}^2 : \|(u,v)\|_{p_1} \geq r_1, \quad \|(u,v)\|_{p_2} \leq r_2\} \tag{3.12}$$

is a TH-set (H-set) if and only if $r_1 \leq r_2$ *and* $p_1 \leq 2 \leq p_2$.

Remark 3.10. Corollaries 3.6–3.9 remain true if one or both the sings $\leq$ in definitions (3.9)–(3.12) are replaced by the sign $<$. Analogous remarks about the replacement of some non-strict inequalities with strict one hold for a number of statements below. We mention here Theorems 4.3, 4.4, 5.2, 5.6, 5.10, 7.2. These modifications do not require substantial changes in the proofs.

The following statement gives a generalization of the "if" part in Corollary 3.7.

Corollary 3.11. *Let* $n(x,y)$ *be another norm on* $\mathbb{R}^2$ *which satisfies*

$$n(x,y) = n(|x|,|y|).$$

Then the set

$$M = \{(u,v) \in \mathbb{C}^2 : n(|u|^2,|v|^2) \leq 1\}$$

is an H-set.

Proof. The set $K(M)$ is the intersection of $\mathbb{R}^2_+$ and of the unit ball in the norm $n(x,y)$. Hence $K(M)$ is convex, and Theorem 3.5 implies that M is an H-set. □

We conclude this section with a characterization of unit spheres which correspond to norms generated by inner products.

Theorem 3.12. *Let* M $(\subset \mathbb{C}^2)$ *be a bicircular set, and for each nonzero vector* $f \in \mathbb{C}^2$ *there exists exactly one positive number* t *such that* $tf \in M$. *If the set* M *is a TH-set, then it has the form*

$$M = \left\{(u,v) \in \mathbb{C}^2 : \frac{|u|^2}{a} + \frac{|v|^2}{b} = 1\right\} \tag{3.13}$$

for some positive numbers a, b.

Proof. Let us show that for each nonzero vector $g = (x,y) \in \mathbb{R}^2_+$ there exists exactly one $t > 0$ such that $tg \in K(M)$.

Indeed, if $t_j g \in K(M)$ $(j = 1,2)$, then $t_j^{1/2}(x^{1/2}, y^{1/2}) \in M$ $(j = 1,2)$ (we use here that M is bicircular). Hence $t_1^{1/2} = t_2^{1/2}$ and $t_1 = t_2$.

This property of the set $K(M)$ implies that $K(M)$ has no inner points. From the other hand, by Lemma 3.1 the set $K(M)$ is convex. The only convex sets in $\mathbb{R}^2$ which have no inner points are intervals. Since the set $K(M)$ contains points of

the form $(a,0)$ and $(0,b)$ and $K(M) \subset \mathbb{R}^2_+$, we obtain that $K(M)$ is the segment joining the points $(a,0)$ and$(0,b)$, i.e.,

$$K(M) = \left\{(x,y) \in \mathbb{R}^2_+ : \frac{x}{a} + \frac{y}{b} = 1\right\}.$$

This equality implies (3.13). □

Corollary 3.13. *Let $n(u,v)$ be another norm (or quasi-norm) on $\mathbb{C}^2$ which satisfies $n(u,v) = n(|u|,|v|)$. The corresponding unit sphere*

$$\{(u,v) \in \mathbb{C}^2 : n(u,v) = 1\}$$

is a TH-set (H-set) if and only if

$$n(u,v) = \left(\frac{|u|^2}{a} + \frac{|v|^2}{b}\right)^{1/2}$$

for some $a, b > 0$.

4. Passage to $\dim X > 2$: similarities and differences

1. We start with the following simple statement which allows sometimes to reduce the problems from the case of arbitrary dimension to the case $\dim X = 2$.

Lemma 4.1. *Let $M \subset X$, $\dim X > 2$. If for each two-dimensional subspace L ($\subset X$) the set $M \cap L$ is a TH-set (resp. H-set), then M is a TH-set (resp. H-set).*

Proof. Consider the case of TH-sets (the case of H-sets is completely analogous).

To prove that $W(A,M)$ is convex for any $A \in \mathcal{L}(X)$, it is enough to show that each two points $(Af,f), (Ag,g) \in W(A,M)$ are contained in some convex subset of $W(A,M)$. Let $L = \operatorname{span}(f,g)$ and P be the orthogonal projection in X on the subspace L. By the condition of the lemma, the set $W(PA|L, M \cap L)$ has the required properties. □

The following example shows that the inverse of Lemma 4.1 does not hold. Moreover, there exists an H-set M and a two-dimensional subspace L ($\subset X$) such that $M \cap L$ is not a TH-set.

Example 4.2. Let $X = \mathbb{C}^3$, $\{e_1, e_2, e_3\}$ be the standard basis in $\mathbb{C}^3$ and S, as always, be the unit sphere. Denote $G = \operatorname{diag}[3/2, 1/2, 1]$, $M = \{f \in S : (Gf,f) = 1\}$ and $L = \operatorname{span}(e_1, e_2)$. It follows from Theorem 5.4 below (or directly from [LM]) that M is an H-set. It is easy to check that

$$M \cap L = \{(u,v) \in \mathbb{C}^2 : |u| = |v| = 1/\sqrt{2}\}.$$

By Lemma 3.2 $M \cap L$ is not a TH-set.

2. Since every two-dimensional subspace of an inner product space X is isometric to $\mathbb{C}^2$, Lemma 4.1 and Corollaries 3.6–3.8 (for $p = 2$) imply the following result.

Theorem 4.3. *If $0 \le r \le R \le \infty$, then the set*

$$\{f \in X : r \le \|f\| \le R\}$$

is an H-set.

Of course, if $R = \infty$ the inequality $\|f\| \le R$ disappears.

Now we can obtain also a generalization of Corollary 3.6.

Theorem 4.4. *Let $0 < p \le \infty$, $0 < r \le R < \infty$ and*

$$M_p = \{f \in \mathbb{C}^n : r \le \|f\|_p \le R\}.$$

Then the following statements are equivalent:

(a) *M_p is an H-set;*
(b) *M_p is a TH-set;*
(c) *$p = 2$.*

Proof. Implication (a) $\Rightarrow$ (b) follows from Theorem 2.1 and implication (c) $\Rightarrow$ (a) follows from Theorem 4.3

(b) $\Rightarrow$ (a) Define $A_1 = \operatorname{diag}[1, i, 0, \dots, 0]$. It is clear that

$$W(A_1, M_p) = \{|f_1|^2 + i|f_2|^2 : f \in M_p\}.$$

Denote $|f_1|^2 = x$, $|f_2|^2 = y$ and identify $\mathbb{C}$ with $\mathbb{R}^2$. Then, obviously,

$$W(A_1, M_p) \subset \{(x, y) \in \mathbb{R}^2_+ : \|(x, y)\|_{p/2} \le R^2\}.$$

Since $(R^2, 0), (0, R^2) \in W(A_1, M_p)$ and the set $W(A_1, M_p)$ is convex, it follows that $p \ge 2$.

Now consider $A_2 = \operatorname{diag}[1, \dots, 1, i]$. Then

$$W(A_2, M_p) = \left\{ \sum_{k=1}^{n-1} |f_k|^2 + i|f_n|^2 : f \in M_p \right\}.$$

Denote $\sum_1^{n-1} |f_k|^2 = x$, $|f_n|^2 = y$. Since $p \ge 2$, we have

$$x = \|(f_1, \dots, f_{n-1})\|_2^2 \ge \|(f_1, \dots, f_{n-1}\|_p^2,$$

and hence

$$\|(x, y)\|_{p/2} \ge \|f\|_p^2 \ge r^2 \quad (f \in M_p).$$

Thus,

$$W(A_2, M_p) \subset \{(x, y) \in \mathbb{R}^2_+ : \|(x, y)\|_{p/2} \ge r^2\}.$$

Since $(r^2, 0), (0, r^2) \in W(A_2, M_p)$ and the set $W(A_2, M_p)$ is convex, it follows that p cannot be > 2. Hence $p = 2$. □

Let us formulate separately the assertion for the case of the ℓ_p spheres.

Corollary 4.5. *Let $0 < p \le \infty$ and*

$$S_p = \{f \in \mathbb{C}^n : \|f\|_p = 1\}.$$

Then the following statements are equivalent:

(a) S_p *is an* H*-set;*
(b) S_p *is a* TH*-set;*
(c) $p = 2$.

3. To each subset $M \subset \mathbb{C}^n$ we assign the following subset of $\mathbb{R}^n_+$:

$$K(M) := \left\{(|f_k|^2)_1^n : f = (f_k)_1^n \in M\right\}.$$

We already used this notion for $n = 2$. Unfortunately, for $n > 2$ it is not very useful.

We do not know if Lemma 3.1 remains true for $n > 2$, but we prove the following weaker statement.

Lemma 4.6. *Let* M $(\subset \mathbb{C}^n)$ *be a* TH*-set. Then for any linear operator* $T : \mathbb{R}^n \to \mathbb{R}^2$*, the set* $T(K(M))$ *is convex.*

Proof. Consider the matrix representation of T :

$$T = \begin{bmatrix} t_{11} & \cdots & t_{1n} \\ t_{21} & \cdots & t_{2n} \end{bmatrix}$$

and define $A := \operatorname{diag}[t_{1k} + it_{2k}]_1^n$. It is easy to check (we identify $\mathbb{C}$ with $\mathbb{R}^2$) that

$$\begin{aligned} W(A, M) &= \left\{ \sum_{k=1}^{n} (t_{1k} + it_{2k})|f_k|^2 : f \in M \right\} \\ &= \left\{ \left(\sum_1^n t_{1k}|f_k|^2, \sum_1^n t_{2k}|f_k|^2 \right) : f \in M \right\} = T(K(M)). \end{aligned}$$

Since M is a TH-set, the set $W(A, M)$ is convex. □

Let $\{e_j\}_1^n$ be the standard basis in $\mathbb{C}^n$. Denote by B_p^n the ℓ_p unit ball in $\mathbb{C}^n$, i.e.,

$$B_p^n = \{f \in \mathbb{C}^n : \|f\|_p \le 1\} \qquad (0 < p \le \infty).$$

Prove that for $p < 2$ many subsets of B_p^n cannot be TH-sets.

Lemma 4.7. *Let* M *be a subset of* B_p^n *which contains two vectors* e_j, e_k $(j \ne k)$*. If* $p < 2$*, then* M *is not a* TH*-set.*

Proof. Let T be the orthogonal projection in $\mathbb{R}^n$ on the subspace $\operatorname{span}\{e_j, e_k\}$. Then

$$K(M) \subset K(B_p^n) = \left\{ x \in \mathbb{R}^n_+ : \sum_{k=1}^{n} x_k^{p/2} \le 1 \right\}$$

and

$$T(K(M)) \subset \left\{ (x_1, x_2) \in \mathbb{R}^2_+ : x_1^{p/2} + x_2^{p/2} \le 1 \right\}.$$

Since the vectors $(1, 0)$ and $(0, 1)$ belong to $T(K(M))$ and $p < 2$, the set $T(K(M))$ is not convex. By Lemma 4.6, M is not a TH-set. □

We say that a set $M \subset \mathbb{C}^n$ is *n-circular* if for any vector $f = (f_k)_1^n$ from M, all vectors of the form $(z_k f_k)_1^n$ $(|z_k| = 1;\ k = 1, \dots, n)$ also belong to M.

Recall that for $n = 2$ we already used this notion in Section 3 under the name bicircular. The next statement is a generalization of Lemma 3.2.

Lemma 4.8. *Let M be an n-circular TH-set. Then for any two different indices $j, k \in \{1, \dots, \}$, there exists a vector $x = (x_1^0, \dots, x_n^0) \in K(M)$ such that $x_j^0 x_k^0 = 0$.*

Proof. Assume that there exists a pair j, k $(j \neq k)$ such that $x_j x_k \neq 0$ for all $x = (x_i)_1^n \in K(M)$. Let A be the $n \times n$ matrix which has exactly one non-zero element, and this element equals 1 and lies on the intersection of the jth row and the kth column. Since M is n-circular,

$$W(A, M) = \left\{ \sqrt{x_j x_k} \exp(i\varphi) : x \in K(M), \varphi \in [-\pi, \pi] \right\}.$$

This set contains a circle with center in the point 0 but does not contain this point. Hence $W(A, M)$ is not convex. Contradiction. □

4. Here we give some examples of n-circular sets which are not TH-sets. This allows us to show that some results from Section 3 do not admit generalizations for $n > 2$. In particular, this concerns the properties of the balls B_p^n $(2 < p \leq \infty)$.

By Theorem 4.3 B_2^n is an H-set for all n, by Corollary 3.7 B_p^2 is a TH-set (H-set) if and only if $p \geq 2$, and by Lemma 4.7 B_p^n is not a TH-set if $p < 2$. Now we pass to the case $n > 2$, $p > 2$, and we begin with B_∞^3.

Theorem 4.9. *Let M be a 3-circular subset of B_∞^3. If $(1, 1, 1) \in M$ then M is not a TH-set.*

Proof. Consider the matrix

$$A = \begin{bmatrix} 0 & 0 & 0 \\ 1 & 0 & 0 \\ -1 & 1 & 0 \end{bmatrix}.$$

If

$$f = (x \exp(i\theta), y \exp(i\varphi), z \exp(i\psi)) \quad (x, y, z \geq 0;\ \theta, \varphi, \psi \in [-\pi, \pi]),$$

then

$$(Af, f) = xy \exp(i(\theta - \varphi)) + yz \exp(i(\varphi - \psi)) - xz \exp(i(\theta - \psi)).$$

Hence

$$W(A, B_\infty^3) = \{xy \exp(i\alpha) + yz \exp(i\beta) - xz \exp(i(\alpha + \beta)) : \quad (4.1)$$
$$0 \leq x, y, z \leq 1;\ -\pi \leq \alpha, \beta \leq \pi\}.$$

Denote

$$f(x, y, z, \alpha, \beta) := xy \exp(i\alpha) + yz \exp(i\beta) - xz \exp(i(\alpha + \beta)),$$
$$g(x, y, z, \alpha, \beta) := \operatorname{Re} f = xy \cos\alpha + yz \cos\beta - xz \cos(\alpha + \beta),$$
$$h(x, y, z, \alpha, \beta) := \operatorname{Im} f = xy \sin\alpha + yz \sin\beta - xz \sin(\alpha + \beta).$$

By (4.1) $W(A, B^3_\infty)$ coincide with the range of the function f on the domain

$$D = \{(x, y, z, \alpha, \beta) : 0 \le x, y, z \le 1; -\pi \le \alpha, \beta \le \pi\}.$$

We want to prove that this range is not convex. To this end we calculate the maximal value of the function g on the domain D. We begin with the remark that if at least one of the variables x, y, z equals 0 then $g(x, y, z, \alpha, \beta) \le 1$, and these values of g are not maximal since, e.g., $g(1, 1, 1, \pi/3, \pi/3) = 3/2$.

Suppose that the function g attains its maximal value on D in a point $(x_0, y_0, z_0, \alpha_0, \beta_0)$. Then the function $\hat{g}(x) = g(x, y_0, z_0, \alpha_0, \beta_0)$ is a linear function of x defined on the segment $[0, 1]$ which reaches its maximal value in a point x_0. Hence x_0 equals either 0 or 1. By the remark above $x_0 \ne 0$, and then $x_0 = 1$. Precisely the same argument shows that $y_0 = z_0 = 1$. Therefore the function g attains its maximal value on D exactly when $x_0 = y_0 = z_0 = 1$ and (α_0, β_0) is the point where the function

$$g_0(\alpha, \beta) = \cos\alpha + \cos\beta - \cos(\alpha + \beta) \quad (-\pi \le \alpha, \beta \le \pi)$$

obtains its maximal value. It is easy to check that there are exactly two such points (α_0, β_0), namely, $(\pi/3, \pi/3)$ and $(-\pi/3, -\pi/3)$, and the corresponding value of the function g_0 equals $3/2$. Thus we obtain that

$$\max\{g(x, y, z, \alpha, \beta) : (x, y, z, \alpha, \beta) \in D\} = 3/2,$$

and this value is reached only in two points $N^\pm = (1, 1, 1, \pm\pi/3, \pm\pi/3)$.

Since $h(N^\pm) = \pm\sqrt{3}/2$, we see that a complex number $\lambda = 3/2 + i\tau$ $(\tau \in \mathbb{R})$ belongs to $W(A, B^3_\infty)$ if and only if $\tau = \pm\sqrt{3}/2$. In particular, $3/2 \notin W(A, B^3_\infty)$.

If M is a subset of B^3_∞, then $3/2 \notin W(A, M)$. On the other hand, if M contains the point $(1, 1, 1)$ and M is 3-circular, then $N^\pm \in M$ and hence $f(N^\pm) = 3/2 \pm i\sqrt{3}/2 \in W(A, M)$. This implies that the set $W(A, M)$ is not convex, and hence M is not a TH-set. □

Propositions 2.8 and 2.12 allow us to obtain from Theorem 4.9 various examples of sets which are not TH-sets. We formulate the following corollary of Theorem 4.9 and Proposition 2.12.

Corollary 4.10. *Let M be a subset of $\mathbb{C}^n$ $(n > 3)$ and P be the orthogonal projection in $\mathbb{C}^n$ on $\mathbb{C}^3$. If the set PM satisfies the conditions of Theorem* 4.9*, then M is not a TH-set.*

In particular, we have

Corollary 4.11. *The set B^n_∞ is not a TH-set for all $n \ge 3$.*

Corollary 4.11 gives an example of an n-circular set M $(n \ge 3)$ with the following properties:

(a) $K(M)$ is convex;
(b) $K(M)$ intersects all coordinate axes;
(c) M is not a TH-set.

Such an example is impossible if $n = 2$ (see Theorem 3.5).

5. Now we are going to consider the case $2 < p < \infty$. We need some notations besides the matrix A and the functions f, g, h defined in the previous subsection. Denote

$$x_p = 3^{-1/p},\ N_p^{\pm} = (x_p, x_p, x_p, \pm\pi/3, \pm\pi/3),\ a_p = g(N_p^+),\ b_p = h(N_p^+),$$
$$c_p = \max\{g(x,y,z,\alpha,\beta) : h(x,y,z,\alpha,\beta) = 0, \|(x,y,z)\|_p \le 1;$$
$$x, y, z \ge 0; -\pi \le \alpha, \beta \le \pi\}$$

(the last notation will be used also for the case $p = \infty$).

It is easy to check that

$$W(A, B_p^3) = \{f(x,y,z,\alpha,\beta) : \|(x,y,z)\|_p \le 1; x, y, z \ge 0; -\pi \le \alpha, \beta \le \pi\},$$
$$(x_p, x_p, x_p) \in B_p^3, \quad 2a_p = 3^{\frac{p-2}{p}}, \quad 2b_p = 3^{\frac{p-4}{2p}}, \quad a_p \pm ib_p \in W(A, B_p^3). \qquad (4.2)$$

Lemma 4.12.

(i) *If $c_p < a_p$ for some $p > 2$, then B_p^n is not a TH-set for all $n \ge 3$.*
(ii) *Let $2 < p_1 < p_2$ and $c_{p_2} < a_{p_1}$. Then B_p^n is not a TH-set for all $p \in [p_1, p_2]$ and all $n \ge 3$.*

Proof. Using Proposition 2.12 we may assume that $n = 3$.

(i) It follows from the definition of c_p and the inequality $c_p < a_p$ that $a_p \notin W(A, B_p^3)$. Since $a_p \pm ib_p \in W(A, B_p^3)$, we see that $W(A, B_p^3)$ is not convex, and hence B_p^3 is not a TH-set.

(ii) Let $p \in [p_1, p_2]$. Since $B_p^3 \subset B_{p_2}^3$, it follows that $c_p \le c_{p_2}$. Also it follows from (4.2) that $a_{p_1} \le a_p$. Thus $c_p \le c_{p_2} < a_{p_1} \le a_p$, and the statement follows from part (i). □

Proposition 4.13. *If $p \ge 6$ and $n \ge 3$, then B_p^n is not a TH-set.*

Proof. Using a computer program, one can obtain with great precision that $c_\infty = 1$ (notice that this value is attained at the point $(1, 1, 1, 0, 0)$). Thus $c_\infty < 3^{2/3}/2 = a_6$, and the statement follows from Lemma 4.12. □

We conjecture that $c_p < a_p$ for all $p > 2$, and then B_p^n is not a TH-set for all $p > 2$ and $n \ge 3$. The confirmation of this conjecture using standard calculus seems to be very cumbersome, and we leave it as an open question. The next statement shows how the confirmed intervals can be enlarged step by step by using numerical arguments.

Proposition 4.14. *If $p \ge 2.5$ and $n \ge 3$, then B_p^n is not a TH-set.*

Proof. By definition of a_p, we have

$$2a_4 = 3^{1/2}, \qquad 2a_3 = 3^{1/3}, \qquad 2a_{8/3} = 3^{1/4}, \qquad 2a_{5/2} = 3^{1/5}.$$

It can be checked by computer programs (we used MAPLE 11) that

$$c_6 \cong 0.797 < a_4, \quad c_4 \cong 0.707 < a_3, \quad c_3 \cong 0.629 < a_{8/3}, \quad c_{8/3} \cong 0.5946 < a_{5/2}.$$

Using Lemma 4.12 step by step, we cover the segment $[2.5, 6]$ and then use Proposition 4.13. □

5. Subsets of the sphere and unbounded set

1. The first and most important example of H- and TH-sets is the unit sphere S. Here we consider some examples of subsets of S which are H-sets, and hence TH-sets. We start with a very simple but useful statement.

Lemma 5.1. *If M is a subset of S and $Z(B, M, 0)$ is connected for any $B = B^*$, then M is an H-set.*

Proof. Follows from the equality

$$Z(B, M, s) = Z(B - sI, M, 0) \qquad (s \in \mathbb{R}). \qquad \square$$

Theorem 5.2. *For any selfadjoint operator G and any real number p, the set*

$$M = \{f \in S : (Gf, f) \geq p\} \tag{5.1}$$

is an H-set.

Proof. Let B be a selfadjoint operator. By Lemma 5.1 it is sufficient to prove the connectedness of the set $Z(B, M, 0)$. Suppose that $f, g \in Z(B, M, 0)$ and prove that there exists a path in $Z(B, M, 0)$ joining f and g. If f and g are linearly dependent, then $f = e^{i\alpha}g$, and we put $h(t) = e^{i\alpha t}g$ $(0 \leq t \leq 1)$.

Let f and g be linearly independent. Choose a real number θ such that

$$\operatorname{Re}\{e^{i\theta}(Bf, g)\} = 0. \tag{5.2}$$

We can also suppose that

$$\operatorname{Re}\{e^{i\theta}((G - pI)f, g)\} \geq 0, \tag{5.3}$$

since if it is not the case we change θ by $\theta + \pi$. Denote $e^{i\theta}f$ by h. Then h also belongs to $Z(B, M, 0)$, and it is sufficient to join h and g in $Z(B, M, 0)$. Let

$$h(t) = th + (1 - t)g \quad (0 \leq t \leq 1), \qquad h_0(t) = h(t)/\|h(t)\|.$$

Using (5.3) we obtain

$$\begin{aligned} &((G - pI)h(t), h(t)) \\ &= t^2((G - pI)f, f) + (1 - t)^2((G - pI)g, g) + 2t(1 - t)\operatorname{Re}((G - pI)h, g) \geq 0, \end{aligned} \tag{5.4}$$

since

$$((G - pI)f, f) \geq 0, \quad ((G - pI)g, g) \geq 0, \qquad \operatorname{Re}((G - pI)h, g) \geq 0.$$

On the other hand,

$$(Bh(t), h(t)) = 2t(1 - t)\operatorname{Re}(Bh, g) = 0 \tag{5.5}$$

by (5.2). It follows from (5.4) and (5.5) that $h_0(t) \in Z(B, M, 0)$ $(0 \leq t \leq 1)$. $\square$

Remark 5.3. If we take $G = AA^*$ in Theorem 5.2, we obtain that the set

$$\{f \in S : \|Af\| \geq \delta\} \tag{5.6}$$

is an H-set for any $A \in \mathcal{L}(X)$ and $\delta \geq 0$. It should be noted that in [AP], [K, Corollary 5.5-11], it was proved that the sets (5.1) and (5.6) are TH-sets.

2. Here we consider a more delicate case when the sign ≥ 0 in the definition (5.1) is replaced by $=$.

Theorem 5.4. *Let*

$$M = \{f \in S : (G, f, f) = s\}$$

where G is a selfadjoint operator and $s \in \mathbb{R}$. If $\dim X > 2$ then M is an H-set. If $s \notin W(G)$ then the set M is empty and hence is an H-set. If $\dim X = 2$ and $s \in W(G)$ then M is a TH-set (H-set) exactly when s is an eigenvalue of G (i.e., s is one of the endpoints of the segments $W(G)$).

Proof. In our notations $M = Z(G, S, s)$, and hence

$$Z(B, M, p) = Z(B + iG, S, p + is).$$

It is proved in [LM] (see also [BM]) that the last set is connected if $\dim X > 2$. Hence in this case M is an H-set. If $s = \max W(G)$ (resp. $\min W(G)$) then (in any dimension) the equality $(Gf, f) = s$ is equivalent to the inequality $(Gf, f) \geq s$ (resp. $(Gf, f) \leq s$), and M is an H-set by Theorem 5.2.

Let now $\dim X = 2$. Choose in X an orthonormal basis e_1, e_2 consisting of the eigenvectors of the operator G, and let λ, μ be the corresponding eigenvalues. Then the set M has the form

$$M = \{f \in X : |(f, e_1)|^2 + |(f, e_2)|^2 = 1, \lambda|(f, e_1)|^2 + \mu|(f, e_2)|^2 = s\}.$$

This set is bicircular, and

$$K(M) = \{(x, y) \in \mathbb{R}^2_+ : x + y = 1, \lambda x + \mu y = s\}.$$

By Theorem 3.5 the set M is a TH-set (H-set) if and only if $K(M)$ contains a point (x_0, y_0) such that $x_0 y_0 = 0$. If $x_0 = 0$ then $y_0 = 1$ and $s = \mu$. Analogously, if $y_0 = 0$ then $s = \lambda$. □

3. We start with a simple statement.

Lemma 5.5. *Let $f(t) : [a, b] \to S$ be a continuous vector function, $G = G^*$ and $g(t) = (Gf(t), f(t))$. If $g(a) > q$ and $g(b) < q$, then there exists $c \in (a, b)$ such that $g(t) \geq q$ on $[a, c]$ and $g(c) = q$.*

Proof. Put $c = \inf\{t : g(t) = q\}$. □

Theorem 5.6. *Let*

$$M = \{f \in S : p \geq (Gf, f) \geq q\}$$

where $G = G^*$ and $p > q$. If $\dim X > 2$ then M is an H-set. In the case $\dim X = 2$ the set M is a TH-set (H-set) if and only if either

$$(Gf, f) \geq q \quad (\forall f \in S) \tag{5.7}$$

or

$$(Gf, f) \leq p \quad (\forall f \in S). \tag{5.8}$$

Proof. We start with the case $\dim X > 2$. Denote

$$M_1 = \{f \in S : (Gf, f) \le p\},$$

$$M_0 = \{f \in S : (Gf, f) = q\}.$$

By Theorems 5.2 and 5.4, M_1,M_0 are H-sets. Let B be a selfadjoint operator and

$$(Bf_1, f_1) = (Bf_2, f_2) = s$$

for some $s \in \mathbb{R}$ and for some vectors $f_1, f_2 \in M$. It is clear that $f_1, f_2 \in M_1$. Since M_1 is an H-set, there exists a continuous vector function $f(t) : [0,1] \to M_1$ such that

$$f(0) = f_1, \quad f(1) = f_2, \quad (Bf(t), f(t)) = s \quad (0 \le t \le 1).$$

If $f(t) \in M$ for all $t \in [0,1]$, then the theorem is proved. Assume that there exists $t_0 \in [0,1]$ such that $(Gf(t_0), f(t_0)) < q$. By Lemma 5.5 there exist $c_1, c_2 \in [0,1]$ such that

$$c_1 \le c_2; \quad f(c_1), f(c_2) \in M_0, \quad f(t) \in M \quad (t \in [0, c_1] \cup [c_2, 1]).$$

Since M_0 is an H-set, there exists a continuous vector function $h(t) : [c_1, c_2] \to M_0$ such that $(Bh(t), h(t)) = s$ for all t. Define

$$g(t) = \begin{cases} f(t) : t \in [0, c_1] \cup [c_2, 1] \\ h(t) : t \in [c_1, c_2] \end{cases}.$$

Then $g(t)$ gives a path in $Z(B, M, s)$ which joins f_1 and f_2. Hence M is an H-set.

Let now $\dim X = 2$. If one of the conditions (5.7), (5.8) holds, then by Theorem 5.2 the set M is an H-set.

Consider the opposite case. Then the eigenvalues λ, μ $(\lambda \ge \mu)$ of G satisfy

$$\lambda < p, \quad \mu > q. \tag{5.9}$$

As in the proof of Theorem 5.4, choose in X an orthonormal basis e_1, e_2 such that $Ge_1 = \lambda e_1$, $Ge_2 = \mu e_2$. Then the set M is a bicircular set, and

$$K(M) = \{(x, y) \in \mathbb{R}^2_+ : x + y = 1, q \le \lambda x + \mu y \le p\}.$$

It follows from (5.9) that $K(M)$ does not contain a point (x_0, y_0) such that $x_0 y_0 = 0$. By Theorem 3.5 M is not a TH-set. □

4. Recall that Corollary 3.8 contains some two-dimensional examples of unbounded H-sets. Here we obtain some results on unbounded H-sets using the approach and reasoning from Theorems 5.2, 5.4 and 5.6.

Theorem 5.7. *For any selfadjoint operator G and any real p the set*

$$M = \{f \in X : (Gf, f) \ge p\} \tag{5.10}$$

is an H-set.

Proof. Consider the set

$$F = Z(B, M, s) = \{f \in X : (Gf, f) \geq p, (Bf, f) = s\}$$

where $B = B^*$ and $s \in \mathbb{R}$. Let $f, g \in F$. Choose a real number θ such that $\mathrm{Re}\{e^{i\theta}(Bf, g)\} = 0$. We can also suppose that

$$\mathrm{Re}\{e^{i\theta}(Gf, g)\} \geq 0 \tag{5.11}$$

(if not, we change θ by $\theta + \pi$).

Denote $e^{i\theta}f$ by h. Then h also belongs to F, and it is sufficient to join h and g by a path within F. Let

$$h(t) = \sqrt{t}h + \sqrt{1-t}g \qquad (0 \leq t \leq 1).$$

Then

$$(Bh(t), h(t)) = t(Bh, h) + (1-t)(Bg, g) + 2\sqrt{t(1-t)}\,\mathrm{Re}(Bh, g) = s$$

since $(Bh, h) = (Bg, g) = s$, $\mathrm{Re}(Bh, g) = 0$.

On the other hand, by (5.11)

$$(Gh(t), h(t)) = t(Gh, h) + (1-t)(Gg, g) + 2\sqrt{t(1-t)}\,\mathrm{Re}(Gh, g) \geq tp + (1-t)p = p.$$

Therefore $h(t) \in F$ for any $t \in [0, 1]$. This means that $F = Z(B, M, s)$ is connected, and hence M is an H-set. □

Now we consider the case when the sign ≥ 0 in definition (5.10) is replaced by $=$.

Theorem 5.8. *For any selfadjoint operator G and any real number p, the set*

$$M = \{f \in X : (Gf, f) = p\} \tag{5.12}$$

is an H-set.

Proof. Let $B = B^*$ and $s \in \mathbb{R}$. Denote $B + iG$ by A and $s + ip$ by z. Since

$$Z(B, M, s) = \{f \in X : (Af, f) = z\},$$

we have to prove that for arbitrary $A \in \mathcal{L}(X)$ and $z \in \mathbb{C}$, the set

$$\{f \in X : (Af, f) = z\} \tag{5.13}$$

is connected. For $z = 0$ this is obvious. Indeed, if $(Af, f) = (Ag, g) = 0$, then the path

$$\{tf : 1 \geq t \geq 0\} \cup \{tg : 0 \leq t \leq 1\}$$

joins f and g within the set (5.13).

The case $z \neq 0$ is easy to reduce to the case $z = 1$. So we have to prove that for arbitrary selfadjoint operators B and G the set

$$F_1 := \{f \in X : (Bf, f) = 1, (Gf, f) = 0\}$$

is connected. Let $f, g \in F_1$. Then, obviously,

$$f, g \in F_2 := \{f \in X : (Bf, f) \geq 1, (Gf, f) = 0\}.$$

By Theorem 5.7 the set

$$M_1 := \{f \in X : (Bf, f) \geq 1\}$$

is an H-set, and hence the set $F_2 = Z(G, M_1, 0)$ is connected. Let $u(t)$ be a path in F_2 which joins f and g. Then

$$h(t) = \frac{u(t)}{(Bu(t), u(t))^{1/2}}$$

is a path in F_1 joining f and g. □

Remark 5.9. In the paper [LTU] it was proved that the set (5.12) is a TH-set.

We conclude this section with the following statement.

Theorem 5.10. *For an arbitrary selfadjoint operator G and real numbers q, p $(q \leq p)$, the set*

$$M = \{f \in X : p \geq (Gf, f) \geq q\}$$

is as H-set.

Proof. We use the reasoning from the proof of Theorem 5.6. If $\dim X > 2$ the proof is completely analogous, except that instead of Theorems 5.2 and 5.4, we have to use here Theorems 5.7 and 5.8.

In the case $\dim X = 2$, we also use the approach from the proof of Theorem 5.6, and we obtain that

$$K(M) = \{(x, y) \in \mathbb{R}^2_+ : p \geq \lambda x + \mu y \geq q\}.$$

Obviously, this set is convex. We can suppose that at least one of the eigenvalues λ, μ (say, λ) is non-zero. Then the point $(p/\lambda, 0)$ belongs to $K(M)$, and by Theorem 3.5 M is an H-set. □

6. Some applications

1. We start with two inequalities for Hermitian forms. The first is well known and has a number of applications (see, e.g., [IKL, Lemma 6.1] and [LMM, Lemma 4.2]). We give here a different proof.

Lemma 6.1. *Let B, G $(\in \mathcal{L}(X))$ be selfadjoint operators, and suppose that B is indefinite, that is, for some vectors $f_0, g_0 \in X$, it holds*

$$(Bf_0, f_0) > 0, \qquad (Bg_0, g_0) < 0. \tag{6.1}$$

If for any vector $f \in X$

$$(Bf, f) = 0 \;\Rightarrow\; (Gf, f) \geq 0, \tag{6.2}$$

then there exists a real number μ such that

$$G \geq \mu B. \tag{6.3}$$

Proof. Define $A = B + iG$. Condition (6.2) implies that the imaginary negative semiaxis $\ell := \{-iy : 0 < y < \infty\}$ does not intersect the numerical range $W(A)$ of the operator A. Hence there exists a straight line Γ which passes through the origin and such that $W(A)$ and ℓ are located in different closed half-planes determined by this line.

By condition (6.1), $W(A)$ has points both in the right and in the left open half-planes; hence Γ does not coincide with the imaginary axis. Therefore the equation of Γ has the form $y = \mu x$ ($\mu \in \mathbb{R}$), and the set $W(A)$ is located in the half-plane $y \geq \mu x$. This proves (6.3). □

Lemma 6.2. *Let B, G ($\in \mathcal{L}(X)$) be selfadjoint operators and B be not negative semidefinite, that is,*

$$(Bf_0, f_0) > 0 \tag{6.4}$$

for some vector $f_0 \in X$. If for any vector $f \in X$

$$(Bf, f) > 0 \;\Rightarrow\; (Gf, f) \geq 0,$$

then there exists a non-negative number μ such that the inequality (6.3) *holds.*

Proof. Like in the proof of Lemma 6.1, we see that

$$x > 0 \;\Rightarrow\; y \geq 0 \quad (z = x + iy \in W(A)).$$

This means that the sets $W(A)$ and

$$Q := \{x + iy : x > 0, y < 0\}$$

do not intersect. Consider the straight line Γ which passes through the origin and such that $W(A)$ and Q are located in different closed half-planes determined by this line. By (6.4) the open right half-plane contains a point from $W(A)$, and therefore Γ does not coincide with the imaginary axis. Hence the equations of Γ has the form $y = \mu x$ and since $\Gamma \cap Q = \emptyset$, we obtain $\mu \geq 0$. It is clear that the set $W(A)$ is located in the half-plane $y \geq \mu x$. □

2. Here we obtain some applications of the results of Section 5.

Theorem 6.3. *Suppose that a selfadjoint operator B ($\in \mathcal{L}(X)$) is not negative semidefinite. Let $A \in \mathcal{L}(X)$) and for any $f \in X$*

$$(Bf, f) > 0 \;\Rightarrow\; (Af, f) \neq 0. \tag{6.5}$$

Then there exists a real number θ and a non-negative number μ such that

$$\mathrm{Re}(e^{i\theta} A) \geq \mu B. \tag{6.6}$$

Proof. Define

$$M = \{f \in S : (Bf, f) > 0\}.$$

By Theorem 5.2 and Remark 3.10, the set M is an H-set and hence a TH-set. Therefore $W(A, M)$ is convex.

Condition (6.5) means that $0 \notin W(A, M)$. Hence there exists a straight line which passes through the origin and such that $W(A, M)$ is located in one of the

two closed half-planes determined by this line. Therefore there exists a number $\theta \in \mathbb{R}$ such that

$$\mathrm{Re}(e^{i\theta}Af, f) \geq 0$$

for all $f \in M$. This means that the conditions of Lemma 6.2 hold for the operators B and $\mathrm{Re}(e^{i\theta}A)$, and we obtain (6.6) □

Theorem 6.4. *Suppose that* $\dim X > 2$ *and that a selfadjoint operator* B $(\in \mathcal{L}(X))$ *is indefinite. Let* $A \in \mathcal{L}(X)$ *and let for any non-zero* $f \in X$

$$(Bf, f) = 0 \;\Rightarrow\; (Af, f) \neq 0. \tag{6.7}$$

Then there exist real numbers θ *and* μ *such that*

$$\mathrm{Re}(e^{i\theta}A) \geq \mu B. \tag{6.8}$$

Proof. Define

$$M_0 = \{f \in S : (Bf, f) = 0\}.$$

Condition (6.7) means that $0 \notin W(A, M_0)$. By Theorem 5.4 the set $W(A, M_0)$ is convex. Repeating the argument from the proof of Theorem 6.3, we obtain that there exists a number $\theta \in \mathbb{R}$ such that

$$(Bf, f) = 0 \;\Rightarrow\; \mathrm{Re}(e^{i\theta}Af, f) \geq 0.$$

Hence for the operators B and $\mathrm{Re}(e^{i\theta}A)$, the conditions of Lemma 6.1 hold, and we obtain the inequality (6.8) for some $\mu \in \mathbb{R}$. □

Example 6.5. We show that for $\dim X = 2$ Theorem 6.4 is false. Let

$$A = \begin{bmatrix} 0 & 0 \\ 1 & 0 \end{bmatrix}, \qquad B = \begin{bmatrix} 1 & 0 \\ 0 & -1 \end{bmatrix}.$$

It is easy to check that here condition (6.7) holds. Suppose that the inequality (6.8) holds for some $\theta, \mu \in \mathbb{R}$. Then

$$\mathrm{Re}\left(e^{i\theta} f_1 \overline{f}_2\right) \geq \mu\left(|f_1|^2 - |f_2|^2\right) \tag{6.9}$$

for all $f_1, f_2 \in \mathbb{C}$. If we set in (6.9) $f_1 = 1$, $f_2 = 0$ (resp. $f_1 = 0$, $f_2 = 1$), we obtain $\mu \leq 0$ (resp. $\mu \geq 0$). Hence $\mu = 0$. Now set $f_1 = f_2 = 1$ (resp. $f_1 = 1$, $f_2 = -1$), and we obtain $\cos\theta \geq 0$ (resp. $\cos\theta \leq 0$.) Hence $\cos\theta = 0$. Finally, set $f_1 = 1$, $f_2 = i$ (resp. $f_1 = 1$, $f_2 = -i$), and we obtain $\sin\theta = 0$. Contradiction.

3. Consider a monic selfadjoint operator polynomial of degree n≥ 2

$$A(\lambda) = \lambda^n I + \sum_{k=0}^{n-1} \lambda^k A_k \qquad (A_k = A_k^*).$$

Let M be a subset of S which is an H-set, and suppose that for each $f \in M$ the scalar polynomial $(A(\lambda)f, f)$ has real and distinct roots

$$p_1(f) > p_2(f) > \cdots > p_n(f).$$

We denote the range of the functional $p_k(f)$ $(f \in M)$ by $\Delta_k(M)$. Since the functional p_k is continuous and the set M is connected (see Proposition 2.2), the sets $\Delta_k(M)$ $(k = 1, \dots, n)$ are intervals.

Theorem 6.6. *The intervals $\Delta_k(M)$ $(k = 1, \dots, n)$ are mutually disjoint.*

For the classical case $M = S$, this theorem becomes the well-known Duffin Theorem [D]. We apply below the argument from the proof of Duffin's Theorem used in [M, Theorem 31.3].

Proof. Since all the roots of the polynomial $(A(\lambda)f, f)$ $(f \in M)$ are real and distinct, its derivative has opposite signs at adjacent roots, and hence

$$(-1)^{k-1}(A'(p_k(f))f, f) > 0 \quad (k = 1, \dots, n). \tag{6.10}$$

Assume that the theorem is false. Then $\Delta_k(M) \cap \Delta_{k+1}(M)$ is nonempty for some k, i.e., there exist a real number α and vectors $g, h \in M$ such that

$$(A(\alpha)g, g) = (A(\alpha)h, h) = 0 \tag{6.11}$$

and $\alpha = p_k(g) = p_{k+1}(h)$. By (6.10)

$$(A'(\alpha)g, g)(A'(\alpha)h, h) < 0, \tag{6.12}$$

and by (6.11) $g, h \in Z(A(\alpha), M, 0)$. Since M is an H-set, the set $Z(A(\alpha), M, 0)$ is connected. By (6.12) the continuous functional $(A'(\alpha)f, f) = 0$ changes its sign on the set $Z(A(\alpha), M, 0)$, and hence it must vanish there. So, there exists a vector $u \in Z(A(\alpha), M, 0)$ such that $(A'(\alpha)u, u) = 0$. But then $(A(\alpha)u, u) = (A'(\alpha)u, u) = 0$, and α is a multiple root of the polynomial $(A(\lambda)u, u)$. Since $u \in M$, this contradicts the main property of the set M formulated above. □

Let us remark in the conclusion of this section that it is not difficult to obtain here generalizations of some other results stated in [M, Section 31] (e.g., of Theorem 31.5).

7. Non-homogeneous quadratic functionals

1. A complex-valued functional q defined on X will be called *quadratic* if it has the form

$$q(f) = (Af, f) + (f, g) + (h, f) \quad (f \in X)$$

where $A \in \mathcal{L}(X)$ and $g, h \in X$. This functional will be called *Hermitian* if $A = A^*$ and $h = g$. Of course, Hermitian functional is real-valued.

Denote by $W(q, M)$ the range of a quadratic functional q on a set M $(\subset X)$:

$$W(q, M) = \{q(f) : f \in M\}.$$

Denote also by $Z(q, M, p)$ $(p \in \mathbb{R})$ the level set of a Hermitian functional q on the set M:

$$Z(q, M, p) = \{f \in M : q(f) = p\}.$$

Under the condition $\dim X > 1$, it was proved in [GM] that the sets $Z(q, S, p)$ are connected for any Hermitian functional q and any $p \in \mathbb{R}$, and the sets $W(q, S)$ are convex for any quadratic functional q. The second statement was obtained in [GM] as a corollary of the first one, and actually they proved the following result (cf. Theorem 2.1).

Lemma 7.1. *Let $M \subset X$. If the set $Z(q, M, p)$ is connected for any Hermitian functional q and any $p \in \mathbb{R}$, then the set $W(q, M)$ is convex for any quadratic functional q.*

2. The main result of this section is the following theorem.

Theorem 7.2. *Let* $\dim X > 1$, $0 \le r \le R \le \infty$ *and*

$$M = \{f \in X : r \le \|f\| \le R\}. \tag{7.1}$$

Then the set $Z(q, M, p)$ is connected for any Hermitian functional q and any $p \in \mathbb{R}$.

This theorem generalizes both Theorem 4.3 and the main result of [GM]. It should be observed that in our proof we use the result of [GM].

Proof. First of all we remark that, exactly as in Theorem 4.3, it is sufficient to consider the case $\dim X = 2$.

Write a Hermitian functional q in the form

$$q(f) = (Bf, f) + \mathrm{Re}(f, g) \quad (f \in X)$$

where $B = B^*$ and $g \in X$. We choose in X an orthonormal basis $\{e_1, e_2\}$ such that e_1, e_2 are eigenvectors of B and $(g, e_1) \ge 0$, $(g, e_2) \ge 0$. With respect to this basis, we have

$$B = \mathrm{diag}[a, b] \quad (a, b \in \mathbb{R}), \quad g = (g_1, g_2) \quad (g_1, g_2 \ge 0).$$

Consider a vector $f \in M$:

$$f = (c\exp(i\alpha), d\exp(i\beta)) \quad (c, d \ge 0; \alpha, \beta \in [-\pi, \pi]). \tag{7.2}$$

This vector belongs to the set $Z(q, M, p)$ if and only if

$$ac^2 + bd^2 + g_1 c\cos\alpha + g_2 d\cos\beta = p.$$

Hence, if the vector (7.2) belongs to the set $Z(q, M, p)$, then all four vectors

$$(c\exp(\pm i\alpha), d\exp(\pm i\beta)) \tag{7.3}$$

belong to this set. Obviously, they also belong to the sphere

$$\{f \in X : \|f\|^2 = c^2 + d^2\}.$$

It follows from [GM] that each pair of vectors from the quadruple (7.3) can be connected within the set $Z(q, M, p)$. Hence we can assume that $\alpha, \beta \in [0, \pi]$ in (7.2).

Consider now two vectors

$$f_k = (c_k\exp(i\alpha_k), d_k\exp(i\beta_k)) \quad (k = 1, 2)$$

from the set $Z(q, M, p)$, where $c_k, d_k \ge 0$ and $\alpha_k, \beta_k \in [0, \pi]$.

Define for $t \in [0,1]$

$$x(t) = tc_1^2 + (1-t)c_2^2, \qquad y(t) = td_1^2 + (1-t)d_2^2,$$

$$f(t) = (\sqrt{x(t)}\exp(i\alpha(t)), \sqrt{y(t)}\exp(i\beta(t))).$$

It is easy to see that $f(t) \in M$ for arbitrary real functions $\alpha(t)$ and $\beta(t)$. It remains to choose continuous functions $\alpha(t)$ and $\beta(t)$ from the conditions that $f(t) \in Z(q, M, p)$ and

$$\alpha(0) = \alpha_2, \quad \alpha(1) = \alpha_1, \quad \beta(0) = \beta_2, \quad \beta(1) = \beta_1.$$

Let for now $c_1c_2d_2d_2 \neq 0$. We define

$$\alpha(t) = \arccos \frac{tc_1\cos\alpha_1 + (1-t)c_2\cos\alpha_2}{\sqrt{x(t)}}.$$

This definition is correct since

$$(tc_1 + (1-t)c_2)^2 \leq tc_1^2 + (1-t)c_2^2$$

(the last inequality can be reduced to $2c_1c_2 \leq c_1^2 + c_2^2$). Similarly, we define

$$\beta(t) = \arccos \frac{td_1\cos\beta_1 + (1-t)d_2\cos\beta_2}{\sqrt{y(t)}}.$$

It is easy to see that $\alpha(t)$ and $\beta(t)$ have all the required properties.

If $c_1 = c_2 = 0$, we can take, e.g., $\alpha(t) = t\alpha_1 + (1-t)\alpha_2$, and if $d_1 = d_2 = 0$, we can take $\beta(t) = t\beta_1 + (1-t)\beta_2$. If only one of the numbers c_1, c_2 (say, c_1) equals zero, we take

$$\alpha(t) = \arccos(\sqrt{1-t}\cos\alpha_2).$$

If only one of the numbers d_1, d_2 (say, d_1) equals zero, we take

$$\beta(t) = \arccos(\sqrt{1-t}\cos\beta_2). \qquad \square$$

Theorem 7.2 and Lemma 7.1 imply the following statement.

Corollary 7.3. *Under the conditions of Theorem* 7.2*, the set* $W(q, M)$ *is convex for any quadratic functional* q*.*

3. Let $M \subset X$, $B = B^* \in \mathcal{L}(X)$ and $h \in X$. Define the Hermitian functional:

$$q(f) = (Bf, f) + (f, Bh) + (Bh, f).$$

It is easy to check that for any $p \in \mathbb{R}$

$$Z(B, M+h, p) = Z(q, M, p - (Bh, h)),$$

and we obtain the following result from Theorem 7.2.

Corollary 7.4. *Under the conditions of Theorem* 7.2 *the set* $M + h$ *is an* H*-set for each* $h \in M$*.*

We use here the same approach which allowed to show that an arbitrary shift of a sphere is an H-set [GM, Corollary 2.3].

It is easy to see that the condition $\dim X > 1$ in Theorem 7.2 and Corollaries 7.3, 7.4 cannot be rejected. Indeed, if we consider in $\mathbb{C}$ the set of numbers

$$\{z = e^{i\varphi} + h : \varphi \in [-\pi, \pi]\}$$

where $h \in \mathbb{C}$ is fixed, then Example 2.7 implies that this set is an H-set if and only if $h = 0$.

In conclusion we show that it can happen that the shift of an H-set is not even a TH-set.

Example 7.5. Let

$$M = \{(e^{i\varphi}, 0) : \varphi \in [-\pi, \pi]\} \ (\subset \mathbb{C}^2).$$

This set is bicircular and $K(M) = \{(1, 0)\}$. By Theorem 3.5 M is an H-set.

On the other hand, if

$$h = (0, 1) \quad \text{and} \quad A = \begin{bmatrix} 0 & 0 \\ 1 & 0 \end{bmatrix},$$

then $W(A, M + h) = \{e^{i\varphi} : \varphi \in [-\pi, \pi]\}$. This set is not convex, and hence $M + h$ is not a TH-set

Acknowledgment

The work of the second-named author was partially supported by Retalon Inc., Toronto, ON, Canada.

References

[AP] J.-H. Au-Jeung and J.-T. Poon, *A remark on the convexity and positive definiteness concerning Hermitian matrices.* Southeast Bull. Math. **3** (1979), 85–92.

[BM] P. Binding and A. Markus, *Joint zero sets and ranges of several Hermitian forms over complex and quaternionic scalars.* Linear Algebra Appl. **385** (2004), 63–72.

[B] L. Brickman, *On the field of values of a matrix.* Proc. Amer. Math. Soc. **12** (1961), 61–66.

[D] R.J. Duffin, *A minimax theory for overdamped networks.* J. Rational Mech. Appl. **4** (1955), 221–233.

[FM] A. Feintuch and A. Markus, *The Toeplitz-Hausdorff theorem and robust stability theory.* Math. Intell. **21** (1999), no. 3, 33–26.

[GR] K.E. Gustafson and D.K.M. Rao, *Numerical Range.* Springer-Verlag, Berlin, 1997.

[G] E. Gutkin, *The Toeplitz-Hausdorff theorem revisited: relating linear algebra and geometry.* Math. Intell. **26** (2004), no. 1, 8–14.

[GM] V.G. Guttierrez and S.L. de Merdano, *An extension of the Toeplitz-Hausdorff theorem.* Bol. Soc. Mat. Mexicana (3) **9** (2003), 273–278.

[Ha] P.R. Halmos, *A Hilbert Space Problem Book.* Springer-Verlag, New York, 1982.

[H] F. Hausdorff, *Der Wertvorrat einer Bilinearform.* Math. Z. **3** (1919), 314–316.

[IKL] I.S. Iokhvidov, M.G. Krein and H. Langer, *Introduction to the Spectral Theory in Spaces with Indefinite Metric.* Akademie-Verlag, Berlin, 1992.

[K] J. Kyle, $W_\delta(T)$ *is convex.* Pacific J. Math. **72** (1977), 483–485.

[LMM] H. Langer, A. Markus and V. Matsaev, *Self-adjoint analytic operator functions and their local spectral function.* J. Funct. Anal. **235** (2006), 193–225.

[LTU] G.-K. Li, N.-K. Tsing and F. Uhlig, *Numerical ranges of an operator on an indefinite inner product space.* Electronic J. Linear Algebra **1** (1996), 1–17.

[LM] Ju. Lyubich and A. Markus, *Connectivity of level sets of quadratic forms and Hausdorff-Toeplitz type theorems.* Positivity **1** (1997), 239–254.

[M] A. Markus, *Introduction to the Spectral Theory of Polynomial Operator Pencils.* Amer. Math. Soc., Providence, RI, 1988.

[T] O. Toeplitz, *Das algebraische Analogen zu einem Satze von Fejér.* Math. Z. **2** (1918), 187–197.

I. Feldman and N. Krupnik
Dept. of Mathematics
Bar-Ilan University
Ramat-Gan 52900, Israel
e-mail: felmni@math.biu.ac.il
krupnik13@rogers.com

A. Markus
Dept. of Mathematics
Ben-Gurion University of the Negev
Beer-Sheva 84105, Israel
e-mail: markus@math.bgu.ac.il

Operator Theory:
Advances and Applications, Vol. 197, 181–250

The Schur Algorithm in Terms of System Realizations

Bernd Fritzsche, Victor Katsnelson and Bernd Kirstein

Dedicated to Moshe Livšic: A great man, thinker, philosopher and mathematician

Abstract. The main goal of this paper is to demonstrate the usefulness of certain ideas from System Theory in the study of problems from complex analysis. With this paper, we also aim to encourage analysts, who might not be familiar with System Theory, colligations or operator models to take a closer look at these topics. For this reason, we present a short introduction to the necessary background. The method of system realizations of analytic functions often provides new insights into and interpretations of results relating to the objects under consideration. In this paper we will use a well-studied topic from classical analysis as an example. More precisely, we will look at the classical Schur algorithm from the perspective of System Theory. We will confine our considerations to rational inner functions. This will allow us to avoid questions involving limits and will enable us to concentrate on the algebraic aspects of the problem at hand. Given a non-negative integer n, we describe all system realizations of a given rational inner function of degree n in terms of an appropriately constructed equivalence relation in the set of all unitary $(n+1)\times(n+1)$-matrices. The concept of Redheffer coupling of colligations gives us the possibility to choose a particular representative from each equivalence class. The Schur algorithm for a rational inner function is, consequently, described in terms of the state space representation.

Mathematics Subject Classification (2000). Primary 30D50, 47A48, 47A57; Secondary 93B28.

Keywords. Schur algorithm, rational inner functions, state space method, characteristic functions of unitary colligations, Redheffer coupling of colligations, Hessenberg matrices.

Notation

$\mathbb{T}$	is the unit circle in the complex plane: $\mathbb{T} = \{t \in \mathbb{C} : \ |t| = 1\}$
$\mathbb{D}$	is the unit disc in the complex plane: $\mathbb{D} = \{z \in \mathbb{C} : \ |z| < 1\}$
$\mathbb{D}^-$	is the exterior of the unit circle: $\mathbb{D}^- = \{z : 1 < |z| \leq \infty\}$.
$\mathfrak{M}_{p\times q}$	is the set of all $p \times q$ (p rows, q columns) matrices with complex entries.
I_n	is the identity $n \times n$ matrix.

Table of Contents

0. Introduction

Up until the 1960s System Theory suggested that a system be considered only in terms of its input and output. A system was treated as a 'black box' with input and output terminals. Associated with each system was an 'input-output' mapping, considered to be of primary importance to the theory at the time. This approach, however, did not take the internal state of the system into account. It is to be assumed that an input signal will, in some way, influence the internal state of a system. Nevertheless, there was little discussion of the relationship between input and the inner state of a system until the introduction of State Space System Theory. This theory not only incorporated input and output spaces, serving, respectively, as 'domains' for input and output 'signals', but also a 'state space'. This state space was introduced to describe the interior state of the system.

State Space System Theory (both the linear and general non-linear variations of the theory) was developed in the early 1960s. Two names closely associated with

the early development of this theory are those of R. Kalman and M.S. Livshitz. Kalman's first publications pertaining to State Space System Theory include [Kal1, Kal2, Kal3]. The monograph [KFA] summarizes these papers, among others. R. Kalman's approach to State Space System Theory was from the perspective of Control Theory. This approach suggested that the questions of a system's controllability and observability be given the most attention. Control Theory does not, however, put much emphasis on energy relations and, as a result, Kalman's work does not address the subject of energy balance relations (Kalman's approach to System Theory was abstract. He develops the theory over arbitrary fields, not specifically over the field of complex numbers). In Kalman's theory, one first starts from the input-output behavior (i.e., transfer function) and then constructs the state operator. In Livshitz's theory, the reverse approach is used: The characteristic function (which is the analogue of the transfer function) is produced from the main operator (which is the analogue of the state operator). Kalman's theory is mainly finite-dimensional and affine, whereas Livshitz's theory is mainly infinite-dimensional and metric. It took some decades before the connections between these two theories were discovered in the 1970s. Among others, Dewilde [Dew1, Dew2] and Helton [He1, He2, He3] produced much of the work leading to this discovery. The connections between the two approaches were made explicit in the monograph [BGK].

M.S. Livshitz, a pioneer in the theory of non-self-adjoint operators, chose to approach State Space Theory from the perspective of Operator Theory. For a particular class of non-self-adjoint operators, Livshitz was able to associate each operator of this class with an analytic function in the upper half-plane or unit disc. These analytic functions were called 'characteristic functions'. Livshitz was, furthermore, able to determine a correspondence between the invariant subspaces of a linear operator and the factors of its characteristic function (See [Liv3] and references within [Liv3]). Using the framework provided for by these results for characteristic functions, Livshitz constructed triangular models of non-self-adjoint operators (Triangular models were later partially supplanted by functional models. See [SzNFo]). Following this, Livshitz focused on questions in both mathematics and physics. Oscillation and wave propagation problems in linear isolated systems are related to self-adjoint operators. In the mid-1950s M.S. Livshitz began to look for a physical example to which his theory of non-self-adjoint operators could be applied. This lead him to consider a number of concrete linear systems. These systems were not isolated systems, but were such that they allowed for the exchange of energy with the 'external world'. The model of the dynamical behavior of a system of this type makes use of an operator and this 'principal' operator is, in general, non-self-adjoint. The energy exchange of the system is reflected in the non-self-adjointness of the operator. Livshitz worked on problems involving the scattering of elementary particles (see [Liv4], [Liv5], [BrLi]), problems in electrical networks (See [LiFl]) and questions dealing with wave propagation in wave-guides (see [Liv6]). It was at this juncture that the notion of an 'operator colligation' (also common are the terms 'operator node' and 'operator cluster') was introduced to provide further clarity. An operator colligation consists of the aforementioned

'principal' operator, but also 'channel' spaces and 'channel' operators, of which the latter two objects describe the non-self-adjointness of the 'principal' operator. The introduction of this concept allowed a characteristic function to be associated with an operator colligation, as opposed to its respective 'principal' operator (see [BrLi], [Br], [LiYa] and references therein). At much the same time, the concept of an 'open system' was then being established. (What Livshitz then referred to as an 'open system' was, in essence, what is now known as a stationary linear dynamical system.) Livshitz first introduced the notion of an 'open system' in his influential paper, [Liv9] (see Definition 1 on p. 1002 of the original Russian paper [Liv7]). To each system there is an associated colligation and in [Liv7] it is shown that a system's transfer operator coincides with the characteristic function of the system's colligation. [Liv8] introduces the operation of coupling open systems as well as the concept of closing coupling channels. [Liv8] furthermore introduces the 'kymological resolution' of an open system, i.e., the resolution of this system into a chain of simpler coupled open systems. These simpler systems correspond to the invariant subspaces of the 'inner-state' operator of the original open system. To emphasize that the notion of an open system is closely related to oscillations and to wave-propagation processes, Livshitz uses the terminology 'kymological', 'kymmer' and 'kymmery', derived from the Greek word '$\kappa\upsilon\mu\alpha$', meaning 'wave'. We quote from page 15 of the English translation of [Liv9] and mention that: "the appropriate representation of an open system, transforming a known input into a known output, depends on which are known and unknown variables, so that the concept of an open system is 'physico-logical' rather than purely physical in nature."

The relevant theory of open systems and operator colligations, as developed by Livshitz and other mathematicians, is presented in the monographs [Liv9], [LiYa] and [Br]. Chapter 2 of the monograph [Liv9] deals with the details of the kymological resolution of open systems (a concept of which much use is made in the following). A detailed presentation of Scattering Theory for linear stationary dynamical systems (with an emphasis on applications to the Wave Equation in $\mathbb{R}^n$) can be found in [LaPhi].

General State Space System Theory, as developed by R. Kalman and M.S. Livshitz provides us with the proper setting and the necessary language for the further study of physical systems and various aspects of Control Theory. Despite the fact that State Space System Theory does not immediately lead to a solution of the initial physical or control problem, it does lead to some interesting related questions (mostly analytic). It should, furthermore, be noted that general State Space Theory's importance extends beyond its significance within Control Theory and when applied to physical systems. M.S. Livshitz was very likely the first to understand that this theory had wide-reaching applications within mathematics, e.g., in Complex Analysis.

Analytic functions can be represented or specified in many ways, e.g., as Taylor-series, by decomposition into continuous fractions, or via representations as Cauchy or Fourier integrals. In the early half of the 1970s an additional method for representing an analytic function was introduced, namely the method of 'system

realization'. This theory has its origins in Synthesis Theory for linear electrical networks, the theory of linear control systems and the theory of operator colligations (and associated characteristic functions). M.S. Livshitz established the Theory of System Realizations and L.A. Sakhnovich, a former Ph.D. student of Livshitz's, later made further important progress in the theory (see [Sakh1] and also [Sakh2] for a more detailed presentation of these results). L.A. Sakhnovich studied the spectral factorization of a given rational matrix-function R, where both R and the inverse function R^{-1} are transfer functions corresponding to linear systems (operator colligations). Unfortunately, the paper [Sakh1] did not garner the attention it deserved at the time. L.A. Sakhnovich's factorization theorem is a predecessor to a fundamental result due to Bart/Gohberg/Kaashoek/van Dooren [BGKV], which was remembered as Theorem 2 in the Editorial Introduction to [CWHF], where one can also find a detailed account of the history of the state space factorization theorem.

Our goal is not to provide a comprehensive survey of the history of System Theory, so that we have focused on the period leading up to the mid-1970s (with particular emphasis on the contributions of M.S. Livshitz and his co-workers). His work on open systems was unkown in the western world until his monograph [Liv9] was translated in 1973. His fundamental papers [Liv7] and [Liv8] remained untranslated up until this memorial volume.

The subsequent development of the Theory of System Realizations is generally associated with the name I. Gohberg, who produced and inspired much in the way of new work and results for this theory and its applications. As a result, the theory experienced a period of accelerated growth, beginning in the late 1970s. Published in 1979, the monograph [BGK][1] dealt with general factorizations of a rational matrix-functions as well as with the Wiener-Hopf factorization of rational matrix function, where, in both cases, this function is a transfer function for a linear system (operator colligation).

I. Gohberg and his co-workers have shown that State Space Theory has a much wider range and goes far beyond System Theory and the theory of operator colligations. We list a few topics to which State Space Theory can be applied:

1. Methods of factorization of matrix- and operator-valued functions; solutions of Wiener-Hopf and singular integral equations.
2. Interpolation in the complex plane and generalizations.
3. Limit formulas of Akhiezer/Kac/Widom type.
4. Projection methods, Bezoutiants, resultants.
5. Inverse problems.

The monograph [BGR] offers a detailed discussion of interpolation problems and many other questions. Matrix function factorization is a tool applied in discussions of many other problems as well, e.g., in the theory of inverse problems for differential equations and also in prediction theory for stationary stochastic

[1] N.b. There is now an extended version of this monograph. See [BGKR].

processes. If a matrix function is rational, then this factorization can be attained using system realizations. These system realizations, in turn, play a certain role in the solution of the original problem (see, for example, [AG]). The Theory of Isoprincipal Deformations of Rational Matrix-Functions (which is, in particular, a useful tool for investigating rational solutions of Schlesinger systems) is formulated in terms of the Theory of System Realizations (See [KaVo1] and [KaVo2]. For our purposes, the theory developed in [Ka] is most relevant). The current state of System Theory, as a branch of pure mathematics, is presented in [Nik].

In the present paper we show how the Schur algorithm for contractive holomorphic functions in the unit disc can be described in terms of system realizations. In the following, we consider only rational inner functions, which allows us to avoid questions involving limits and enables us to concentrate on the algebraic aspects of the problem at hand. At first glance the formulas here presented might seem rather complicated and, to some degree, less than intuitive. This is, however, from the perspective of System Theory, not the case. The aforementioned formulas serve as the function-theoretical counterpart to Livshitz's kymological resolution as applied to the system (represented by the original inner function) corresponding to the cascade coupling, i.e., the Redheffer coupling, of open systems. The elementary open systems, which make up this cascade (or chain) correspond to the steps of the Schur algorithm.

This paper is organized as follows. In Section 1, we state some facts relating to rational inner functions. In Section 2, we discuss some aspects of the classical Schur algorithm. Section 3 is devoted to a short introduction to operator colligations and their characteristic functions, where particular attention is paid to finite-dimensional unitary colligations. The characteristic functions of finite-dimensional unitary colligations are shown to be rational inner matrix-functions (see Theorem 3.5). Theorem 3.6 shows that an arbitrary rational inner matrix function can, on the other hand, be realized as a characteristic function of a finite-dimensional minimal unitary colligation. The scalar rational inner functions of degree n are just the finite Blaschke products of n elementary Blaschke factors. The essential facts on the realization of scalar inner rational functions of degree n as characteristic functions of minimal unitary colligations are summarized in Theorem 3.10. These minimal unitary colligations can be equivalently described by equivalence classes of minimal unitary $(n+1)\times(n+1)$-matrices. A proof for Theorem 3.10 can be found in the Appendix at the end of the paper.

The main objective of this paper can be described as follows. The application of the Schur algorithm to a given rational inner function $s(z)$ of degree n produces a sequence $s_k(z)\,,\ k=0,\,1,\,\ldots\,,n$ of rational inner functions with $s_0(z)=s(z)$ and $\deg s_i(z)=n-k$. In particular, the function $s_n(z)$ is constant with unimodular value. In Section 3, it will be shown that each of the functions $s_k(z)$ admits a system representation

$$s_k(z)=A_k+zB_k\,(I-zD_k)^{-1}\,C_k$$

in terms of the blocks of some minimal unitary matrix $U_k \in \mathfrak{M}_{(n+1)\times(n+1)}$,

$$U_k = \begin{pmatrix} A_k & B_k \\ C_k & D_k \end{pmatrix}.$$

We assume that U_0 is given. The goal is to recursively produce the sequence matrices U_k. In other words, the steps of the Schur algorithm have to be described in terms of the state space representation. Since the unitary matrices U_k are defined only up to an equivalence relation, we have to find corresponding operations for the arithmetic of these equivalence classes.

In Section 4 we discuss the means by which the linear-fractional transformation associated with the Schur algorithm can be described in terms of the input-output mapping of linear systems. The Redheffer coupling of linear systems will be introduced as a useful tool in these considerations.

In Section 5, the Redheffer coupling of linear systems will be translated into the language of unitary systems.

In Section 6, we apply the concept of Redheffer couplings of colligations to the linear-fractional transformation associated with the inverse of the Schur algorithm. In so doing, we will describe the 'degrees of freedom' of unitary equivalence. A closer look shows us that amongst all the unitary matrices which realize the desired system realization, there are some distinguished by the fact that they are, in a sense, associated with the concept of Redheffer coupling.

In Section 7, the basic step of the Schur algorithm will be described in the language of colligations. This requires that we solve a particular equation for unitary matrices, suggested by the results of Section 6. The solution to this matrix equation is given in Theorem 7.1. Together with Lemma 7.2, Theorem 7.2 describes the basic step of the Schur algorithm in terms of system representations.

The investigations of Section 8 show that a certain normalization procedure has to be performed at every step of the Schur algorithm if the Schur algorithm is to be dealt with in the language of system realizations. We consider the degrees of freedom for this normalization procedure. It turns out that we can use these degrees of freedom to make the normalization procedure a one-time-procedure, so that it might be dealt with during preprocessing for further step-by-step recurrence. A one-time-normalization of this kind is related to the reduction of the 'initial' colligation matrix to the lower Hessenberg matrix.

In Section 9, we will be well positioned to present the Schur algorithm in terms of unitary colligations representing the appropriate functions.

In Section 10 we express the colligation matrix in terms of the Schur parameters.

In the final section (Section 11) we discuss some connections between the present work and other work relating to the Schur algorithm as expressed in terms of system realizations. In particular, we discuss the results presented in Alpay/Azizov/Dijksma/Langer [AADL] and Killip/Nenciu [KiNe].

1. Rational Inner Functions

We say that a function $s : \mathbb{D} \to \mathbb{C}$, where s is holomorphic in $\mathbb{D}$, is contractive if

$$|s(z)| \leq 1 \text{ for every } z \in \mathbb{D}.$$

A contractive function s is called an inner function if

$$|s(t)| = 1 \text{ for every } t \in \mathbb{T}.$$

In the following we consider rational inner functions, so that $s(t)$ is defined for every $t \in \mathbb{T}$.

A rational function is representable as a quotient of irreducible polynomials and we call the order of the highest-degree polynomial the degree of the rational function.

If a rational function s is an inner function, then the degree of its numerator and the degree of its denominator are equal.

An inner rational function s is representable as a finite Blaschke product, i.e., in the form

$$s(z) = c \prod_{1 \leq k \leq n} \frac{z_k - z}{1 - z\overline{z}_k}. \tag{1.1}$$

$z_1, \ldots, z_n$ are points in $\mathbb{D}$, or, in other words, complex numbers satisfying the condition

$$|z_1| < 1, \ldots, |z_n| < 1, \tag{1.2}$$

c is a unimodular complex number, i.e.,

$$|c| = 1. \tag{1.3}$$

Conversely, given complex numbers $z_1, \ldots, z_n$ and c satisfying the conditions (1.2) and (1.3), respectively, the function s in (1.1) is an inner rational function of degree n.

The number c and the *set* $\{z_1, \ldots, z_n\}$ are uniquely defined by the inner function s (the *sequence* of numbers $(z_1, \ldots, z_n)$) up to permutation.

The notions of contractive and inner functions can also be defined for matrix-functions:

We say that a matrix-function $S : \mathbb{D} \to \mathfrak{M}_{p\times p}$, where S is holomorphic in $\mathbb{D}$, is contractive if

$$I_p - S^*(z)S(z) \geq 0 \text{ for every } z \in \mathbb{D}.$$

A contractive matrix-function $S : \mathbb{D} \to \mathfrak{M}_{p\times p}$, is called an inner function if[2]

$$I_p - S^*(t)S(t) = 0 \text{ for almost every } t \in \mathbb{T}.$$

[2]For a contractive holomorphic function S in $\mathbb{D}$, the boundary values $S(t) \stackrel{\text{def}}{=} \lim_{r\to 1-0} S(rt)$ exist for almost every $t \in \mathbb{T}$ (with respect to the Lebesgue measure).

2. The Schur algorithm

In this section, we present a short introduction to the classical Schur algorithm, which originated in Issai Schur's renowned paper, [Sch]. In so doing, we will mainly emphasize those aspects of the Schur algorithm, which are essential for this paper. For comprehensive treatments of the Schur algorithm and its matricial generalizations, we refer the reader to [BFK1], [BFK2], [Con2], [DFK], [S:Meth] and the references therein.

Let $s(z)$ be a contractive holomorphic function in $\mathbb{D}$ and

$$s_0 = s(0). \tag{2.1}$$

Then $|s_0| \leq 1$, where $|s_0| = 1$ only if $s(z) \equiv s_0$. If $|s_0| < 1$, then the function

$$\omega(z) = \frac{1}{z} \frac{s(z) - s_0}{1 - s(z)\overline{s_0}} \tag{2.2}$$

is well defined. Moreover, it is contractive holomorphic in $\mathbb{D}$. The function $s(z)$ can be expressed in terms of these $\omega(z)$ and s_0:

$$s(z) = \frac{s_0 + z\,\omega(z)}{1 + z\,\overline{s_0}\,\omega(z)}\,. \tag{2.3}$$

If the function $s(z)$ is an inner function, then $\omega(z)$ is also an inner function. If $s(z)$ is an inner rational function of degree n, then $\omega(z)$ is an inner rational function of degree $n - 1$.

Conversely, if $\omega(z)$ is an *arbitrary* contractive holomorphic function in $\mathbb{D}$ and s_0 is an *arbitrary* complex number satisfying the condition $|s_0| < 1$, then the expression on the right-hand side of (2.3) defines the function $s(z)$, which is holomorphic and contractive in $\mathbb{D}$. Furthermore, if $\omega(z)$ is an inner function, then $s(z)$ is an inner function as well.

Definition 2.1.

I. *We call the transformation* $s(z) \longmapsto \omega(z)$, *defined by* (2.2), *where* $s_0 = s(0)$, *the* (*direct*) Schur transformation.

II. *We call the transformation* $\omega(z) \longmapsto s(z)$, *defined by* (2.3), *where* s_0 *is a given complex number, the* inverse Schur transformation.

The correspondence $s(z) \Longleftrightarrow (s(0), \omega(z))$ describes the elementary step of the Schur algorithm.

The Schur algorithm is applied to a holomorphic function $s(z)$, which is contractive in $\mathbb{D}$. This algorithm inductively produces the sequence (finite or infinite) of contractive holomorphic functions $s_k(z)$ in $\mathbb{D}$ and contractive numbers $s_k = s_k(0)$, $k = 0, 1, 2, \ldots$. The algorithm terminates only if $s(z)$ is a rational inner function. Starting from $s(z)$, we define

$$s_0(z) = s(z), \qquad s_0 = s_0(0)\,.$$

If the functions $s_i(z)$, $i = 0, 1, \ldots, k$ are already constructed and $|s_k(0)| < 1$, then we construct the function $s_{k+1}(z)$ as follows:

$$s_{k+1}(z) = \frac{1}{z}\,\frac{s_k - s_k(z)}{1 - s_k(z)\overline{s_k}}\,, \quad s_{k+1} = s_{k+1}(0)\,. \tag{2.4}$$

If $s(z)$ is not a rational inner function, then the algorithm does not terminate: On the kth step we obtain the function $s_k(z)$, for which $|s_k(0)| < 1$, so that we can construct the function $s_{k+1}(z)$ and still have $|s_{k+1}(0)| < 1$.

If $s(z)$ is a rational inner function of degree n, then we can define the functions $s_i(z)$ for $i = 0, 1, \ldots, n$ such that

$$\deg s_i(z) = n - i, \qquad i = 0, 1, \ldots, n\,.$$

The numbers $s_i = s_i(0)$ satisfy the conditions

$$|s_i| < 1, \qquad i = 0, 1, \ldots, n-1\,.$$

However, in this case

$$|s_n| = 1, \qquad s_n(z) \equiv s_n.$$

So, for $k = n$ the numerator and the denominator of the expression on the right-hand side of (2.4) vanish identically. The function $s_{n+1}(z)$ is thus not defined and the Schur algorithm terminates.

The numbers $s_k = s_k(0)$ are called the Schur parameters of the function $s(z)$.

If $s(z)$ is not an inner rational function, then the sequence of its Schur parameters is infinite and these parameters s_k satisfy the inequality $|s_k| < 1$ for all $k : 0 \le k < \infty$. If $s(z)$ is an inner rational function with $\deg s(z) = n$, then its Schur parameters s_k are defined only for $k = 0, 1, \ldots, n$ and

$$|s_k| < 1, \quad k = 0, 1, \ldots, n-1, \qquad |s_n| = 1. \tag{2.5}$$

Conversely, given complex numbers $s_0, s_1, \ldots, s_n$ satisfying the conditions (2.5), one can construct the inner rational function of degree n, having Schur parameters $s_0, s_1, \ldots, s_n$. This function $s(z)$ can be constructed inductively: First, we set

$$s_n(z) \equiv s_n\,.$$

If the functions $s_i(z)$ for $i = n, n-1, \ldots, k$ are already constructed, then we set

$$s_{k-1}(z) = \frac{s_{k-1} + z\,s_k(z)}{1 + z\,\overline{s_{k-1}}\,s_k(z)}\,.$$

In the final step we construct the function $s_0(z)$ and set

$$s(z) = s_0(z).$$

Thus, *there exists a one-to-one correspondence between rational inner functions of degree n and sequences of complex numbers $\{s_k\}_{0\le k\le n}$ satisfying the conditions* (2.5).

3. The system representation of a rational inner function

Contractive holomorphic functions appear in several roles. In particular, such functions appear in Operator Theory as the *characteristic functions of operator colligations*. The notion of an operator colligation is closely related to that of a linear stationary dynamical system. There is a correspondence between the theory of operator colligations and the theory of linear stationary dynamical systems. The concepts and results of one theory can be translated into the language of the other. There are interesting connections to be made between these theories. Definitions and constructions, which are well motivated and natural in the framework of one theory may look artificial in the framework of the other. In particular, the notion of the characteristic function of a colligation and of the coupling of colligations are more transparent in the language of System Theory.

In this section, the term 'operator' means 'continuous linear operator'.

Definition 3.1. *Let $\mathcal{H}$, $\mathcal{E}^{\text{in}}$, $\mathcal{E}^{\text{out}}$ be Hilbert spaces and U be an operator:*

$$U : \mathcal{E}^{\text{in}} \oplus \mathcal{H} \to \mathcal{E}^{\text{out}} \oplus \mathcal{H}\,, \tag{3.1}$$

Let

$$U = \begin{bmatrix} A & B \\ C & D \end{bmatrix} \tag{3.2}$$

be the block decomposition of the operator U, corresponding to (3.1)*:*

$$A : \mathcal{E}^{\text{in}} \to \mathcal{E}^{\text{out}},\ B : \mathcal{H} \to \mathcal{E}^{\text{out}},\ C : \mathcal{E}^{\text{in}} \to \mathcal{H}, D : \mathcal{H} \to \mathcal{H}\,. \tag{3.3}$$

The quadruple $(\mathcal{E}^{\text{in}},\ \mathcal{E}^{\text{out}},\ \mathcal{H}, U)$ is called an operator colligation.

$\mathcal{E}^{\text{in}}$ and $\mathcal{E}^{\text{out}}$ are, respectively, the input *and* output spaces *of the colligation. We call $\mathcal{H}$ the* state space *of the colligation and A the* exterior operator. *We call B and C* channel operators, *while D is referred to as the* principal operator *of the colligation. Finally, we call U the* colligation operator.

If the input and the output spaces $\mathcal{E}^{\text{in}}$ and $\mathcal{E}^{\text{out}}$ coincide: $\mathcal{E}^{\text{in}} = \mathcal{E}^{\text{out}} = \mathcal{E}$, we call the space $\mathcal{E}$ the exterior space *of the colligation and denote the colligation by the triple $(\mathcal{E}, \mathcal{H}, U)$*

Definition 3.2. *Let $(\mathcal{E}^{\text{in}}, \mathcal{E}^{\text{out}}, \mathcal{H}, U)$ be an operator colligation.*

The operator-function

$$S(z) = A + zB(I_{\mathcal{H}} - zD)^{-1}C \tag{3.4}$$

is called the characteristic function *of the colligation $(\mathcal{E}^{\text{in}}, \mathcal{E}^{\text{out}}, \mathcal{H}, U)$.*

The function $S(z)$ is defined for the $z \in \mathbb{C}$ where the operator $(I_{\mathcal{H}} - zD)^{-1}$ exists. The values of S are operators acting from $\mathcal{E}^{\text{in}}$ into $\mathcal{E}^{\text{out}}$.

Remark 3.1. *The function $S(z)$ is defined and holomorphic in some neighborhood of the point $z = 0$. Furthermore, $S(0) = A$.*

The notion of a colligation's characteristic function draws on the framework of the theory of linear stationary dynamical systems (LSDS). (The theory of open systems, in the terminology of M.S. Livŝic). The theory of LSDS, which we are

dealing with is not a 'black box theory', where only the input signals, output signals and the mapping 'input → output' are considered. The theory of LSDS also takes 'interior states' of the system into account. The input and output signals are described (in the discrete time case, where the index k serves as time) by sequences $\{\varphi_k\}_{0\le k<\infty}$ and $\{\psi_k\}_{0\le k<\infty}$ of vectors belonging to some Hilbert spaces $\mathcal{E}^{\text{in}}$ and $\mathcal{E}^{\text{out}}$ (the *input* and the *output* spaces of the system). The 'interior states' are described by vectors h of a Hilbert space $\mathcal{H}$, called *the state space* of the system.

The dynamics of a *linear stationary* system is described by the linear equations

$$\begin{bmatrix} \psi_k \\ h_{k+1} \end{bmatrix} = \begin{bmatrix} A & B \\ C & D \end{bmatrix} \begin{bmatrix} \varphi_k \\ h_k \end{bmatrix}, \quad k = 0, 1, 2, \dots, \tag{3.5}$$

where the operators A, B, C, D do not depend on k ('time') and are defined in (3.3).

It is natural to consider the four operators A, B, C, D as blocks of the 'unified' operator, say U as in (3.2), from the space $\mathcal{E}^{\text{in}} \oplus \mathcal{H}$ into the space $\mathcal{E}^{\text{out}} \oplus \mathcal{H}$. The operator colligation $(\mathcal{E}^{\text{in}}, \mathcal{E}^{\text{out}}, \mathcal{H}, U)$ then corresponds to the LSDS (3.5), (3.3). Given the sequence $\{\varphi_k\}_{0\le k\le m}$ and the initial value h_0, the system (3.5) uniquely determines the sequences $\{\psi_k\}_{0\le k\le m}$ and $\{h_k\}_{0\le k\le m+1}$. In the case $h_0 = 0$,

$$\psi_0 = A\varphi_0, \quad \psi_m = A\varphi_m + \sum_{1\le k\le m-1} BD^k C\varphi_{m-k-1}\,, \ m \ge 1\,. \tag{3.6}$$

The relation (3.6) can be considered as the description of the evolution of the LSDS (3.5) in the *time domain*. The description of the evolution is, however, especially transparent in the *frequency domain*. Since the considered sequences are unilateral, the Fourier transforms of these sequences are the (formal) power series

$$\varphi(z) = \sum_{0\le k<\infty} \varphi_k z^k\,, \quad \psi(z) = \sum_{0\le k<\infty} \psi_k z^k\,, \quad h(z) = \sum_{0\le k<\infty} h_k z^k\,. \tag{3.7}$$

The complex variable z can be interpreted as the frequency. Under the extra assumption that $h_0 = 0$ we can rewrite (3.5) in terms of the Fourier representations:

$$\begin{bmatrix} \psi(z) \\ z^{-1}h(z) \end{bmatrix} = \begin{bmatrix} A & B \\ C & D \end{bmatrix} \begin{bmatrix} \varphi(z) \\ h(z) \end{bmatrix}. \tag{3.8}$$

From (3.8) we obtain

$$\psi(z) = A\varphi(z) + Bh(z), \tag{3.9a}$$

$$h(z) = z\,(I - zD)^{-1}C\varphi(z)\,. \tag{3.9b}$$

Eliminating $h(z)$, we get

$$\psi(z) = S(z)\varphi(z)\,, \tag{3.10}$$

where $S(z)$ is expressed in terms of the matrix (3.2) as in (3.4):

$$S(z) = A + zB(I - zD)^{-1}C\,.$$

The operator function $S(z)$ describes the input-output mapping corresponding to LSDS (3.5).

Definition 3.3. *In the framework of System Theory, the function $S(z)$ in* (3.10) *is called the* transfer matrix *of the LSDS* (3.5).

In the theory of operator colligations the operator function $S(z)$ is called the *characteristic function*, while in the theory of LSDS it is called the *transfer function*. This notion, however, makes more sense in the theory of LSDS. Along with the input-output mapping described by the transfer function $S(z)$, the input-state mapping:

$$\varphi(z) \to h(z)\,, \quad \text{where} \quad h(z) = z\,(I - zD)^{-1}C\varphi(z)\,,$$

is also naturally related to the system (3.5).

If the dimensions $\dim \mathcal{E}^{\text{in}}$ and $\dim \mathcal{E}^{\text{out}}$ of the input and output spaces are finite, then, choosing bases in $\mathcal{E}^{\text{in}}$ and $\mathcal{E}^{\text{out}}$, we can consider $S(z)$ as a matrix-valued function. If, moreover, the dimension $\dim \mathcal{H}$ of the state space is finite, then $S(z)$ is a rational matrix-function.

Definition 3.4. *The colligation $(\mathcal{E}^{\text{in}}, \mathcal{E}^{\text{out}}, \mathcal{H}, U)$ is said to be* finite-dimensional *if* $\dim \mathcal{E}^{\text{in}} < \infty$, $\dim \mathcal{E}^{\text{out}} < \infty$ *and* $\dim \mathcal{H} < \infty$.

The dimension $\dim \mathcal{H}$ of the state space of the finite-dimensional colligation $(\mathcal{E}^{\text{in}}, \mathcal{E}^{\text{out}}, \mathcal{H}, U)$ is related to the degree of its characteristic function. Here we use the notion of the *McMillan degree* of a rational matrix-valued function as it is defined in [McM]. The notion of the degree of a rational matrix-function is discussed in [DuHa] and [Kal4]. See also [BGK]. In the case when $\dim \mathcal{E} = 1$, i.e., in the case when the considered rational function is scalar (or $\mathbb{C}$-valued), the McMillan degree of this function coincides with its 'standard' degree.

To precisely formulate how the dimension of the state space $\mathcal{H}$ and the degree of the characteristic function $S(z)$ are related, we need to introduce the notion of a *minimal* colligation $(\mathcal{E}^{\text{in}}, \mathcal{E}^{\text{out}}, \mathcal{H}, U)$.

Definition 3.5. *Let $(\mathcal{E}^{\text{in}}, \mathcal{E}^{\text{out}}, \mathcal{H}, U)$ be a colligation. We define the following subspaces of the state space $\mathcal{H}$:*

$$\mathcal{H}^c = \operatorname{clos}\left(\bigvee_{0\le k<\infty} (D^k C)\mathcal{E}^{\text{in}}\right), \quad \mathcal{H}^o = \operatorname{clos}\left(\bigvee_{0\le k<\infty} (D^{*k} B^*)\mathcal{E}^{\text{out}}\right), \tag{3.11}$$

where $\bigvee_k f_k$ denotes the linear hull of the vectors f_k and $\operatorname{clos}(M)$ denotes the closure of the set M.

The subspaces $\mathcal{H}^c$ and $\mathcal{H}^o$ are, respectively, called the controllability *and* observability subspaces *of the colligation $(\mathcal{E}^{\text{in}}, \mathcal{E}^{\text{out}}, \mathcal{H}, U)$.*

Remark 3.2. *If the state space $\mathcal{H}$ is finite-dimensional, say $\dim \mathcal{H} = n < \infty$, then it is enough to restrict our considerations in* (3.11) *to the linear hull of the vectors $(D^k C)\mathcal{E}^{\text{in}}$ and $(D^{*k} B^*)\mathcal{E}^{\text{out}}$ with $k < n$. In this case there is no need to make use of the closure in* (3.11).

Definition 3.6. *We say that a colligation* $(\mathcal{E}^{\rm in}, \mathcal{E}^{\rm out}, \mathcal{H}, U)$ *is* controllable *if* $\mathcal{H}^c = \mathcal{H}$ *and* observable *if* $\mathcal{H}^o = \mathcal{H}$.

We say that a colligation is simple *if the sum of the controllability and the observability subspaces is dense in the state space, i.e., if*

$$\operatorname{clos}\bigl(\mathcal{H}^c + \mathcal{H}^o\bigr) = \mathcal{H}\,.$$

We say that a colligation $(\mathcal{E}^{\rm in}, \mathcal{E}^{\rm out}, \mathcal{H}, U)$ *is* minimal *if it is both controllable and observable, i.e., if*

$$\mathcal{H}^c = \mathcal{H} \quad \textit{and} \quad \mathcal{H}^o = \mathcal{H}\,.$$

Theorem 3.1. *Let* $(\mathcal{E}^{\rm in}, \mathcal{E}^{\rm out}, \mathcal{H}, U)$ *be a finite-dimensional colligation and let* $S(z)$ *be the characteristic function of this colligation.*

$S(z)$ is then a rational matrix-function, which is holomorphic at $z = 0$ and such that

$$\deg S \le \dim \mathcal{H} \tag{3.12}$$

Equality holds in (3.12) *if and only if the colligation* $(\mathcal{E}^{\rm in}, \mathcal{E}^{\rm out}, \mathcal{H}, U)$ *is minimal.*

Theorem 3.2. *Let $\mathcal{E}_1$ and $\mathcal{E}_2$ be finite-dimensional spaces and let $S(z)$ be a rational function, whose values are operators acting from $\mathcal{E}_1$ to $\mathcal{E}_2$ and which is holomorphic at the point $z = 0$.*

There then exists a finite-dimensional minimal operator colligation $(\mathcal{E}_1^{\rm in}, \mathcal{E}_1^{\rm out}, \mathcal{H}, U)$, (3.1)–(3.2)–(3.3), *with* $\mathcal{E}^{\rm in} = \mathcal{E}_1$ *and* $\mathcal{E}^{\rm out} = \mathcal{E}_2$, *whose characteristic function* $S_U(z) = A + zB(I - zD)^{-1}C$ *coincides with the original function $S(z)$. In other words, S can be expressed in the form* (3.4).

Definition 3.7. *The representation of a given function $S(z)$ as a characteristic function of an operator colligation is called the* state space representation of $S(z)$ *or the* state space realization of $S(z)$. *If the representative operator colligation is minimal, then we say that the state space realization of $S(z)$ is* minimal.

Let us discuss the uniqueness of the state space representation.

Definition 3.8. *Let* $(\mathcal{E}_1^{\rm in}, \mathcal{E}_1^{\rm out}, \mathcal{H}_1, U_1)$ *and* $(\mathcal{E}_2^{\rm in}, \mathcal{E}_2^{\rm out}, \mathcal{H}_2, U_2)$ *be two operator colligations:*

$$U_1 = \begin{bmatrix} A_1 & B_1 \\ C_1 & D_1 \end{bmatrix}, \qquad U_2 = \begin{bmatrix} A_2 & B_2 \\ C_2 & D_2 \end{bmatrix}, \tag{3.13}$$

where

$$A_i : \mathcal{E}_i^{\rm in} \to \mathcal{E}_i^{\rm out}\,, \quad B_i : \mathcal{H}_i \to \mathcal{E}_i^{\rm out}, \quad C_i : \mathcal{E}_i^{\rm in} \to \mathcal{H}_i, \quad D_i : \mathcal{H}_i \to \mathcal{H}_i, \quad i = 1, 2\,. \tag{3.14}$$

We consider the colligations $(\mathcal{E}_1^{\rm in}, \mathcal{E}_1^{\rm out}, \mathcal{H}_1, U_1)$ *and* $(\mathcal{E}_2^{\rm in}, \mathcal{E}_2^{\rm out}, \mathcal{H}_2, U_2)$ *to be* equivalent *if* invertible *operators* $E^{\rm in}$, $E^{\rm out}$ *and* V*:*

$$E^{\rm in} : \mathcal{E}_2^{\rm in} \to \mathcal{E}_1^{\rm in}, \qquad E^{\rm out} : \mathcal{E}_2^{\rm out} \to \mathcal{E}_1^{\rm in}, \qquad V : \mathcal{H}_2 = \mathcal{H}_1, \tag{3.15}$$

exist, such that the intertwining relation

$$\begin{bmatrix} E^{\text{out}} & 0 \\ 0 & V \end{bmatrix} \begin{bmatrix} A_2 & B_2 \\ C_2 & D_2 \end{bmatrix} = \begin{bmatrix} A_1 & B_1 \\ C_1 & D_1 \end{bmatrix} \begin{bmatrix} E^{\text{in}} & 0 \\ 0 & V \end{bmatrix} \tag{3.16}$$

holds.

Clearly, given two equivalent operator colligations, one of these colligations is controllable, observable, simple or minimal if and only if the other colligation possesses the same respective property.

The following result is evident:

Theorem 3.3. *Let* $(\mathcal{E}_1^{\text{in}}, \mathcal{E}_1^{\text{out}}, \mathcal{H}_1, U_1)$ *and* $(\mathcal{E}_2^{\text{in}}, \mathcal{E}_2^{\text{out}}, \mathcal{H}_2, U_2)$ *be operator colligations. Assume that these colligations are equivalent, i.e., that the intertwining relation* (3.16) *holds with some invertible operators* E^{in}, E^{out} *and* V.

Then the characteristic functions $S_1(z)$ *and* $S_2(z)$ *of these colligations,*

$$S_i(z) = A_i + zB_i(I - zD_i)^{-1}C_i, \qquad i = 1, 2, \tag{3.17}$$

satisfy the intertwining relation:

$$E^{\text{out}} S_2(z) = S_1(z) E^{\text{in}}. \tag{3.18}$$

for all z *where* S_1 *and* S_2 *are defined.*

Under the extra assumptions that the colligations are minimal and finite-dimensional we can show that for Theorem 3.3 the converse assertion also holds.

Theorem 3.4. *Let* $(\mathcal{E}_1^{\text{in}}, \mathcal{E}_1^{\text{out}}, \mathcal{H}_1, U_1)$ *and* $(\mathcal{E}_2^{\text{in}}, \mathcal{E}_2^{\text{out}}, \mathcal{H}_2, U_2)$ *be* finite-dimensional *operator colligations. Let* $S_1(z)$ *and* $S_2(z)$, (3.17), *be the characteristic functions of these colligations. We make the following assumptions:*

1. *The functions* $S_1(z)$ *and* $S_2(z)$ *satisfy the intertwining relation* (3.18) *for all* z *small enough, where* $E^{\text{in}} : \mathcal{E}_2^{\text{in}} \to \mathcal{E}_1^{\text{in}}$ *and* $E^{\text{out}} : \mathcal{E}_2^{\text{out}} \to \mathcal{E}_1^{\text{out}}$ *are some invertible operators.*
2. *The colligations* $(\mathcal{E}_1^{\text{in}}, \mathcal{E}_1^{\text{out}}, \mathcal{H}_1, U_1)$ *and* $(\mathcal{E}_2^{\text{in}}, \mathcal{E}_2^{\text{out}}, \mathcal{H}_2, U_2)$ *are minimal.*

These colligations are then equivalent, i.e., there exists an invertible operator $V : \mathcal{H}_2 \to \mathcal{H}_1$ *such that the intertwining relation* (3.16) *holds.*

Up to this point, we have not taken advantage of any scalar products that may be defined in the input, output and state spaces. From this point forward, we will focus more on these scalar products and the benefits they bring when we have them at our disposal. In what follows, we consider rational *inner* functions. Operator colligations representing such functions are *unitary, finite-dimensional* operator colligations.

For convenience, we recall the definition of a unitary operator:

Let $\mathcal{L}_1$ *and* $\mathcal{L}_2$ *be Hilbert spaces and* $T : \mathcal{L}_1 \to \mathcal{L}_2$ *be an operator. We say that* T *is* unitary *if it satisfies the following two conditions:*

a) *T preserves the scalar product, i.e.,*

$$\langle Tx, Ty\rangle_{\mathcal{L}_2} = \langle x, y\rangle_{\mathcal{L}_1} \quad \forall x \in \mathcal{L}_1,\, y \in \mathcal{L}_1\,.$$

b) *T maps $\mathcal{L}_1$ onto $\mathcal{L}_2$, i.e., T is invertible.*

The unitarity property of a linear operator T can also be characterized as follows:

$$T^*T = I_{\mathcal{L}_1}, \qquad TT^* = I_{\mathcal{L}_2}\,.$$

Definition 3.9. *Let $(\mathcal{E}^{\rm in}, \mathcal{E}^{\rm out}, \mathcal{H}, U)$, (3.2) - (3.3), be an operator colligation. We call $(\mathcal{E}^{\rm in}, \mathcal{E}^{\rm out}, \mathcal{H}, U)$ a unitary colligation if the colligation operator U is a unitary operator, i.e., if*

$$U^*U = I_{\mathcal{E}^{\rm in}\oplus\mathcal{H}}, \qquad UU^* = I_{\mathcal{E}^{\rm out}\oplus\mathcal{H}}\,. \tag{3.19}$$

Definition 3.10. *Let $\mathcal{E}_1$ and $\mathcal{E}_2$ be finite-dimensional Hilbert spaces and let $S(z)$ be a rational function whose values are operators acting from $\mathcal{E}_1$ to $\mathcal{E}_2$.*

The matrix-function S is called an inner function if its values $S(z)$ are contractive operators for $z \in \mathbb{D}$ and unitary operators for $t \in \mathbb{T}$, i.e., if the conditions

$$I_{\mathcal{E}_1} - S^*(z)S(z) \geq 0\,,\;\; I_{\mathcal{E}_2} - S(z)S^*(z) \geq 0, \quad \text{for} \;\; z \in \mathbb{D}, \tag{3.20a}$$

$$I_{\mathcal{E}_1} - S^*(t)S(t) \,= 0\,,\;\; I_{\mathcal{E}_2} - S(t)S^*(t) \,= 0\;, \quad \text{for} \;\; t \in \mathbb{T}\,. \tag{3.20b}$$

hold. (In particular, S has no singularities in $\mathbb{D}\cup\mathbb{T}$.)

Remark 3.3. *Since unitary operators are invertible, $\mathcal{E}_1$ - $\mathcal{E}_2$ inner functions exist only if $\dim\mathcal{E}_1 = \dim\mathcal{E}_1$.*

Theorem 3.5. *Let $(\mathcal{E}^{\rm in}, \mathcal{E}^{\rm out}, \mathcal{H}, U)$, (3.2) - (3.3), be a finite-dimensional unitary colligation and $S(z)$, (3.4), be its characteristic function.*

Then the function $S(z)$ is a rational inner function.

Proof. The proof of this lemma is based on identity (3.8), where $h(z)$ is expressed in terms of $\varphi(z)$ as in (3.9b). Let z and ζ be such that the operators $I - zD$ and $I - \zeta D$ are invertible. (These operators are invertible if $z \in \mathbb{D}, \zeta \in \mathbb{D}$. Also, since the spectrum of the operator D is a finite set, the operators $I - zD$ and $I - \zeta D$ are invertible for all but finitely many $z \in \mathbb{T}, \zeta \in \mathbb{T}$.) Because the operator U is unitary, (3.8) yields

$$\langle\psi(z), \psi(\zeta)\rangle_{\mathcal{E}^{\rm out}} + (z\overline{\zeta})^{-1}\langle\, h(z), h(\zeta)\,\rangle_{\mathcal{H}} = \langle\varphi(z), \varphi(\zeta)\rangle_{\mathcal{E}} + \langle\, h(z), h(\zeta)\,\rangle_{\mathcal{H}}\,,$$

or

$$\begin{aligned}\big\langle\varphi(z), \varphi(\zeta)\big\rangle_{\mathcal{E}^{\rm in}} - \big\langle S(z)\varphi(z), S(\zeta)\varphi(\zeta)\big\rangle_{\mathcal{E}^{\rm out}}\\ = (1 - z\overline{\zeta})\big\langle(I - zA)^{-1}C\varphi(z), (I - \zeta A)^{-1}C\varphi(\zeta)\big\rangle_{\mathcal{H}}\,.\end{aligned} \tag{3.21}$$

In particular, taking $\varphi(z) \equiv \varphi'$ and $\varphi(\zeta) \equiv \varphi''$, where φ' and φ'' are arbitrary vectors in $\mathcal{E}^{\rm in}$, we obtain the equality

$$\frac{I_{\mathcal{E}^{\rm in}} - S^*(\zeta)S(z)}{1 - \overline{\zeta}z} = C^*(I - \overline{\zeta}D^*)^{-1}(I - zD)^{-1}C\,. \tag{3.22}$$

In the same way we obtain the equality

$$\frac{I_{\mathcal{E}^{\text{out}}} - S(z)S^*(\zeta)}{1 - z\overline{\zeta}} = B(I - zD)^{-1}(I - \overline{\zeta}D^*)^{-1}B^*\,. \tag{3.23}$$

Using the identity $\frac{\zeta(I-\zeta D)^{-1} - z(I-zD)^{-1}}{\zeta - z} = (I - \zeta D)^{-1}(I - zD)^{-1}$, we obtain

$$\frac{S(\zeta) - S(z)}{\zeta - z} = B(I - \zeta D)^{-1}(I - zD)^{-1}C\,, \tag{3.24}$$

and

$$\frac{S^*(\zeta) - S^*(z)}{\overline{\zeta} - \overline{z}} = C^*(I - \overline{\zeta}D^*)^{-1}(I - \overline{z}D^*)^{-1}B^*\,, \tag{3.25}$$

To get (3.20) we let $\zeta = z$ in (3.22)–(3.23):

$$I_{\mathcal{E}^{\text{in}}} - S^*(z)S(z) = (1 - |z|^2)\, C^*(I - \overline{z}A^*)^{-1}(I - zA)^{-1}C, \tag{3.26a}$$

$$I_{\mathcal{E}^{\text{out}}} - S(z)S^*(z) = (1 - |z|^2)\, B(I - zA)^{-1}(I - \overline{z}A^*)^{-1}B^*\,. \tag{3.26b}$$

The inequalities (3.20a) follow from equalities (3.26), which hold for all $z \in \mathbb{D}$. The equalities (3.26) furthermore hold for all but finitely many $z \in \mathbb{T}$. Thus, the rational function $S(z)$ is bounded in $\mathbb{T}$, except on a finite set. S therefore has no singularities in $\mathbb{T}$ and takes unitary values there.

The following theorem serves as a 'unitary' counterpart to Theorem 3.2.

Theorem 3.6. *Let $S(z)$ be a* rational inner *function whose values are operators acting from $\mathcal{E}_1$ into $\mathcal{E}_2$, where $\mathcal{E}_1$ and $\mathcal{E}_2$ are finite-dimensional Hilbert spaces.*

Then there exists a finite-dimensional, minimal*, unitary operator colligation $(\mathcal{E}_1^{\text{in}}, \mathcal{E}_1^{\text{out}}, \mathcal{H}, U)$,* (3.1)–(3.3)*, with $\mathcal{E}^{\text{in}} = \mathcal{E}_1$ and $\mathcal{E}^{\text{out}} = \mathcal{E}_2$, whose characteristic function $S_U(z) = A + zB(I - zD)^{-1}C$ coincides with the original function $S(z)$. In other words, the function S is representable in the form* (3.4)*.*

Definition 3.11. *Let $(\mathcal{E}_1^{\text{in}}, \mathcal{E}_1^{\text{out}}, \mathcal{H}_1, U_1)$ and $(\mathcal{E}_2^{\text{in}}, \mathcal{E}_2^{\text{out}}, \mathcal{H}_2, U_2)$ be operator colligations,* (3.13)–(3.14)*. If these colligations are equivalent (i.e., if they satisfy the intertwining relation* (3.16)–(3.15)*) and each of the operators E^{in}, E^{out}, V is unitary, we say that $(\mathcal{E}_1^{\text{in}}, \mathcal{E}_1^{\text{out}}, \mathcal{H}_1, U_1)$ and $(\mathcal{E}_2^{\text{in}}, \mathcal{E}_2^{\text{out}}, \mathcal{H}_2, U_2)$ are* unitarily equivalent*.*

Clearly, if two operator colligations are unitarily equivalent and one of these colligations is unitary, then the second colligation is also unitary.

The following theorem provides us with a 'unitary' counterpart to Theorem 3.3.

Theorem 3.7. *Let $(\mathcal{E}_1^{\text{in}}, \mathcal{E}_1^{\text{out}}, \mathcal{H}_1, U_1)$ and $(\mathcal{E}_2^{\text{in}}, \mathcal{E}_2^{\text{out}}, \mathcal{H}_2, U_2)$ be unitary colligations,* (3.13)*. Furthermore, let these colligations be unitarily equivalent, i.e., suppose that the intertwining relation* (3.16) *holds for some* unitary *operators E^{in}, E^{out} and V.*

The respective characteristic functions $S_1(z)$ and $S_2(z)$ of these colligations, (3.17)*, then satisfy the intertwining relation* (3.18) *with these very same* unitary *operators E^{in} and E^{out}.*

If we, furthermore, assume that both unitary colligations are simple, we can show that the converse to Theorem 3.7 also holds.

The next theorem serves as a 'unitary' counterpart to Theorem 3.4.

Theorem 3.8. *Let* $(\mathcal{E}_1^{\text{in}}, \mathcal{E}_1^{\text{out}}, \mathcal{H}_1, U_1)$ *and* $(\mathcal{E}_2^{\text{in}}, \mathcal{E}_2^{\text{out}}, \mathcal{H}_2, U_2)$ *be finite-dimensional* unitary *operator colligations,* (3.13)*. Let* $S_1(z)$ *and* $S_2(z)$*,* (3.17)*, be the characteristic functions of* $(\mathcal{E}_1^{\text{in}}, \mathcal{E}_1^{\text{out}}, \mathcal{H}_1, U_1)$ *and* $(\mathcal{E}_2^{\text{in}}, \mathcal{E}_2^{\text{out}}, \mathcal{H}_2, U_2)$*, respectively. We now make the following assumptions:*

1. *The functions* $S_1(z)$ *and* $S_2(z)$ *satisfy the intertwining relation* (3.18) *for* $z \in \mathbb{D}$*, where* $E^{\text{in}} : \mathcal{E}_2^{\text{in}} \to \mathcal{E}_1^{\text{in}}$, $E^{\text{out}} : \mathcal{E}_2^{\text{out}} \to \mathcal{E}_1^{\text{out}}$ *are some unitary operators.*
2. *The colligations* $(\mathcal{E}_1^{\text{in}}, \mathcal{E}_1^{\text{out}}, \mathcal{H}_1, U_1)$ *and* $(\mathcal{E}_2^{\text{in}}, \mathcal{E}_2^{\text{out}}, \mathcal{H}_2, U_2)$ *are simple.*

The colligations $(\mathcal{E}_1^{\text{in}}, \mathcal{E}_1^{\text{out}}, \mathcal{H}_1, U_1)$ *and* $(\mathcal{E}_2^{\text{in}}, \mathcal{E}_2^{\text{out}}, \mathcal{H}_2, U_2)$ *are then unitarily equivalent, i.e., there exists a unitarily operator* $V : \mathcal{H}_2 \to \mathcal{H}_1$ *such that the intertwining relation* (3.16) *holds.*

Let us compare the assumptions of Theorems 3.4 and 3.8. In Theorem 3.4 we assume that the colligations $(\mathcal{E}_i^{\text{in}}, \mathcal{E}_i^{\text{out}}, \mathcal{H}_i, U_i)$, $i = 1, 2$, are minimal, however it is not assumed that these colligations are unitary. In Theorem 3.8 we assume that the colligations are unitary and simple, but we do not explicitly assume that these colligations are minimal, because they *are*, in fact, already *minimal*.

Theorem 3.9. *Let* $(\mathcal{E}^{\text{in}}, \mathcal{E}^{\text{out}}, \mathcal{H}, U)$ *be a finite-dimensional,* unitary *operator colligation. The following statements are then equivalent:*

1. *The colligation is simple.*
2. *The colligation is minimal.*
3. *The colligation is controllable.*
4. *The colligation is observable.*

In what follows we deal only with scalar-valued inner functions $S(z)$, i.e., with functions whose values are complex numbers. The input space $\mathcal{E}^{\text{in}}$ and the output space $\mathcal{E}^{\text{out}}$ of the unitary colligation $(\mathcal{E}^{\text{in}}, \mathcal{E}^{\text{out}}, \mathcal{H}, U)$ representing this $S(z)$ can be identified with the space $\mathbb{C}$: $\mathcal{E}^{\text{in}} = \mathcal{E}^{\text{out}} = \mathbb{C}$. The finite-dimensional state space $\mathcal{H}$, with, say $\dim \mathcal{H} = n$, can be identified with the space $\mathbb{C}^n$ (with the standard scalar product): $\mathcal{H} = \mathbb{C}^n$. With these conventions in place, the orthogonal sums $\mathcal{E}^{\text{in}} \oplus \mathcal{H}$ and $\mathcal{E}^{\text{out}} \oplus \mathcal{H}$ can be identified naturally with the space $\mathbb{C} \oplus \mathbb{C}^n$.

We note that $\mathbb{C} \oplus \mathbb{C}^n$ represents a canonical decomposition of the space $\mathbb{C}^{n+1}$ into an orthogonal sum. We consider the space $\mathbb{C}^{n+1}$ as the set $\mathfrak{M}_{(n+1)\times 1}$ of all $(n+1)$-column-vectors, along with the standard linear operations and scalar product:

$$\langle f\,,\, g\rangle = g^* f, \quad f,\, g \in \mathfrak{M}_{(n+1)\times 1}, \tag{3.27}$$

where the asterisk * denotes Hermitian conjugation.

A unitary operator, U, acting in $\mathbb{C}^{n+1}$ is described by a unitary $(1+n) \times (1+n)$-matrix, which will also be denoted by U. U maps the column-vector f to the column-vector Uf, where Uf is the usual matrix product. The decomposition

$\mathbb{C}^{n+1} = \mathbb{C} \oplus \mathbb{C}^n$ of the space $\mathbb{C}^{n+1}$ suggest that we consider the following block-matrix decomposition of U:

$$U = \begin{bmatrix} A & B \\ C & D \end{bmatrix}, \tag{3.28a}$$

$$A \in \mathfrak{M}_{1\times 1},\ B \in \mathfrak{M}_{1\times n},\ C \in \mathfrak{M}_{n\times 1},\ D \in \mathfrak{M}_{n\times n}\,. \tag{3.28b}$$

The matrix entries are considered as operators:

$$A : \mathcal{E} \to \mathcal{E}\,,\ B : \mathcal{H} \to \mathcal{E}\,,\ C : \mathcal{E} \to \mathcal{H}\,,\ D : \mathcal{H} \to \mathcal{H}\,. \tag{3.29}$$

where

$$\mathcal{E} = \mathfrak{M}_{1\times 1}\ \ (= \mathbb{C}),\quad \mathcal{H} = \mathfrak{M}_{n\times 1}\ \ (= \mathbb{C}^n)\,. \tag{3.30}$$

Definition 3.12. *Given a unitary matrix $U \in \mathfrak{M}_{(n+1)\times(n+1)}$ with block decomposition* (3.28)*, we associate the unitary colligation $(\mathcal{E}, \mathcal{H}, U)$ with U. The exterior space $\mathcal{E}$ and the state space $\mathcal{H}$ of this colligation are as in* (3.30)*, where the spaces $\mathbb{C}$ and $\mathbb{C}^n$ have the standard scalar products. The exterior, principal and channel operators A, D, B, C correspond to the block-matrix entries in* (3.28) *and satisfy* (3.29)*.*

We call this colligation the unitary colligation associated with the unitary matrix U*.*

Given two unitary colligations associated with unitary matrices U' and U'', how do we express that these colligations are unitarily equivalent? The exterior spaces of both colligations are 'copies' of the same space $\mathbb{C}$. To identify the exterior spaces $\mathbb{C}$ of two different colligations, we should specify the unitary operators E^{in} and E^{out} for the two copies of $\mathbb{C}$ (These operators, E^{in} and E^{out}, appear in (3.15) and in the intertwining relations (3.16) and (3.18).) We can naturally choose these identification operators as the identity operators, i.e., such that each of operators E^{in} and E^{out} is represented by the 1×1-matrix whose (unique) entry is the number 1 (Such operators can be represented by 1×1-matrices, where the matrices corresponding to E^{in} and E^{out} consist, respectively, of an arbitrary number ν^{in} and ν^{out} with $|\nu^{\mathrm{in}}| = 1$ and $|\nu^{\mathrm{out}}| = 1$.)

With this convention in place, the unitary equivalence of the colligations associated with the block-matrices

$$U' = \begin{bmatrix} A' & B' \\ C' & D' \end{bmatrix} \in \mathfrak{M}_{(n+1)\times(n+1)} \quad \text{and} \quad U'' = \begin{bmatrix} A'' & B'' \\ C'' & D'' \end{bmatrix} \in \mathfrak{M}_{(n+1)\times(n+1)} \tag{3.31}$$

means that these matrices satisfy the intertwining relation:

$$\begin{bmatrix} 1 & 0 \\ 0 & V \end{bmatrix} \begin{bmatrix} A'' & B'' \\ C'' & D'' \end{bmatrix} = \begin{bmatrix} A' & B' \\ C' & D' \end{bmatrix} \begin{bmatrix} 1 & 0 \\ 0 & V \end{bmatrix} \tag{3.32}$$

where $V \in \mathfrak{M}_{n\times n}$ is a unitary matrix. The equality (3.18) then becomes:

$$S_1(z) = S_2(z).$$

Definition 3.13. *We say that the unitary matrices* $U' \in \mathfrak{M}_{(n+1)\times(n+1)}$ *and* $U'' \in \mathfrak{M}_{(n+1)\times(n+1)}$, (3.31), *are* equivalent *if there exists a unitary matrix* $V \in \mathfrak{M}_{n\times n}$ *such that the intertwining relation* (3.32) *holds.*

Let $U = \begin{bmatrix} A & B \\ C & D \end{bmatrix} \in \mathfrak{M}_{(n+1)\times(n+1)}$. We now consider the following matrices associated with the unitary matrix U:

$$\mathcal{C}(U) = [C \quad DC \quad \dots \quad D^{n-1}C], \qquad \mathcal{C}(U) \in \mathfrak{M}_{n\times n}\,, \tag{3.33a}$$

$$\mathcal{B}(U) = [B^* \;\; D^*B^* \;\; \dots \;\; (D^*)^{n-1}B^*], \quad \mathcal{B}(U) \in \mathfrak{M}_{n\times n}\,, \tag{3.33b}$$

and

$$\mathcal{S}(U) = [C \;\; DC \;\; \dots \;\; D^{n-1}C,\, B^* \;\; D^*B^* \;\; \dots \;\; (D^*)^{n-1}B^*]\,,$$
$$\mathcal{S}(U) \in \mathfrak{M}_{n\times 2n}\,. \tag{3.33c}$$

If the unitary colligation associated with the matrix U is controllable, observable or simple, this means that the matrix (3.33a), (3.33b) or (3.33c) is, respectively, of rank n.

Remark 3.4. *If one of the matrices* (3.33) *has rank* n*, then its columns* (*considered as vectors in* $\mathbb{C}^n = \mathfrak{M}_{n\times 1}$) *generate the whole space. The columns of these matrices are of the form* D^kC *or* $(D^*)^kB^*$*, where* k *takes values in the interval* $[0, \dots, (n-1)]$*. It is possible to consider matrices of this kind for* k *over a larger interval. Extending the interval* $[0, \dots, (n-1)]$ *does not, however, lead to an increase in rank for these matrices: The Cayley-Hamilton Theorem tells us that the column-vectors,* D^pC *and* $(D^*)^kB^*$ *with* $p \geq n$*, are, respectively, linear combinations of the column-vectors* D^kC *and* $(D^*)^kB^*$ *with* $k \in [0, \dots, (n-1)]$.

Definition 3.14. *We say that a unitary matrix* $U \in \mathfrak{M}_{(n+1)\times(n+1)}$*, expressed using the block-decomposition in* (3.28)*, is* controllable *if* rank $\mathcal{C}(U) = n$*,* observable *if* rank $\mathcal{B}(U) = n$ *and* simple *if* rank $\mathcal{S}(U) = n$*. If the matrix* U *is both controllable and observable, we say that it is* minimal.

(We note that any one of the matrices (3.33) is of rank n if and only if the other two have rank n. See Theorem 3.9.)

The results of this section on the state space representation of scalar (i.e., complex-valued) rational inner functions can be summarized in the following way:

Theorem 3.10. (Rational Inner Functions $\Longleftrightarrow$ Equivalence Classes of Unitary Matrices)

1. *Let* $S(z)$ *be an inner rational function of degree* n*. Then* $S(z)$ *can be represented in the form:*

$$S(z) = A + zB(I_n - zD)^{-1}C, \tag{3.34}$$

where A, B, C, D *are blocks of some unitary minimal matrix* U*,*
$U \in \mathfrak{M}_{(n+1)\times(n+1)}$, (3.28).

2. *Let* $U \in \mathfrak{M}_{(n+1)\times(n+1)}$ *be a unitary matrix with block-decomposition* (3.28) *and let the function* $S(z)$ *be defined in terms of* U *by* (3.34). *Then the function* $S(z)$ *is a rational inner function with* $\deg S \le n$. *If the matrix* U *is minimal, then* $\deg S = n$.
3. *Let* $U' \in \mathfrak{M}_{(n+1)\times(n+1)}$ *and* $U'' \in \mathfrak{M}_{(n+1)\times(n+1)}$ *be unitary matrices with block-decomposition* (3.31) *and let* $S'(z)$ *and* $S''(z)$ *be the functions defined in terms of* U' *and* U'' *by:*

$$\begin{aligned} S'(z) &= A' + zB'(I_n - zD')^{-1}C', \\ S''(z) &= A'' + zB'(I_n - zD'')^{-1}C'', \end{aligned} \tag{3.35}$$

If the matrices U' *and* U'' *are equivalent, then* $S'(z) \equiv S''(z)$. *If* $S'(z) \equiv S''(z)$ *and the matrices* U' *and* U'' *are minimal, then* U' *and* U'' *are equivalent.*

The substance of this theorem can be summarized as follows:

- There exists a one-to-one correspondence between the set of all rational inner functions of degree $\le n$ and the set of all equivalence classes of unitary matrices in $\mathfrak{M}_{(n+1)\times(n+1)}$.
- This correspondence can be expressed as a mapping from the set of all rational inner functions of degree n onto the set of all equivalence classes of minimal unitary matrices in $\mathfrak{M}_{(n+1)\times(n+1)}$.

For a proof of Theorem 3.10, see the Appendix at the end of this paper.

The main objective of this paper

Applying the Schur algorithm to a given rational inner function $s(z)$ of degree n produces the sequence $s_k(z)$, $k = 0, 1, \dots, n$, of rational inner functions with $s_0(z) = s(z)$ and $\deg s_k(z) = n - i$. In particular, $s_n(z) \equiv s_n$ is a unitary constant. According to what was stated in Section 3, each of the functions $s_k(z)$ admits a system representation,

$$s_k(z) = A_k + zB_k(I - zD_k)^{-1}C_k\,, \tag{3.36}$$

in terms of the blocks of some minimal unitary matrix $U_k \in \mathfrak{M}_{(1+n-k)\times(1+n-k)}$:

$$U_k = \begin{bmatrix} A_k & B_k \\ C_k & D_k \end{bmatrix}. \tag{3.37}$$

We assume that from the very beginning, the given inner rational function $s(z) = s_0(z)$ is determined in terms of its state space representation, so that the matrix U_0 is given. The goal is to recursively produce the sequence of matrices U_k representing the functions $s_k(z)$, $k = 1, 2, \dots, n$. The matrix U_{k+1}, representing the function $s_{k+1}(z)$, is thus constructed from the matrix U_k, representing the function $s_k(z)$. In other words, the steps (2.4) of the Schur algorithm must be described in terms of the state space representation (3.36).

It should be noted that the unitary matrices in the system representation of a rational inner function are determined only up to the equivalence

$$\begin{bmatrix} A_k & B_k \\ C_k & D_k \end{bmatrix} \sim \begin{bmatrix} 1 & 0 \\ 0 & V_k^{-1} \end{bmatrix} \cdot \begin{bmatrix} A_k & B_k \\ C_k & D_k \end{bmatrix} \cdot \begin{bmatrix} 1 & 0 \\ 0 & V_k \end{bmatrix}, \tag{3.38}$$

where V_k is an arbitrary unitary $k \times k$ matrix. So we have to find a rule for constructing a matrix U_{k+1}, which belongs to the equivalence class of matrices representing the function $s_{k+1}(z)$, from an arbitrary element U_k of the equivalence class of matrices representing the function $s_k(z)$.

The Schur algorithm in the framework of system representations is described in Section 9.

Historical Remarks

The definition of a characteristic function was developed gradually, starting from the pioneering works of M.S.Livshitz. The first definition appeared in [Liv1] (for operators for which $I - T^*T$ and $I - TT^*$ have rank one) and in [Liv2] (for the case that these operators have finite rank). M.S. Livshitz and those working in the same field, subsequently turned their attention to bounded operators T for which $T - T^*$ is of finite rank or at least of finite trace. For these operators T, a characteristic function was defined in an analogous way and by means of this function, a wide-reaching theory for these operators developed. In particular, triangular models of non-self-adjoint operators were introduced. See [Liv3], [BrLi], [Br]. In the course of the evolution of the concept of characteristic functions, it became clear that it was advantageous to consider, not just non-self-adjoint operators, but also more general objects: operator nodes (or operator colligations). The notion of an operator colligation was prompted by physical applications of the Livshitz theory of non-selfadjoint operators. (See [BrLi], [Liv9] and references there.)

B. Sz. Nagy and C. Foias used a different approach to characteristic functions in 1962. Their work involved harmonic analysis of the unitary dilation of the contractive operator T. Moreover, they simultaneously obtained a functional model of T depending explicitly and exclusively on the characteristic function of T. See [SzNFo, especially Chapter VI] and references therein.

The version of operator colligations, which appears in Definition 3.1 goes back to a remark of M.G. Krein to the work [BrSv1]. In [BrSv1], the notion of a contractive operator colligation (node) was defined as the collection of Hilbert spaces $\mathcal{H}, \mathcal{F}, \mathcal{G}$ and operators

$$T_0 : \mathcal{G} \to \mathcal{F},\ F : \mathcal{F} \to \mathcal{H},\ G : \mathcal{G} \to \mathcal{H},\ T : \mathcal{H} \to \mathcal{H}, \tag{3.39}$$

satisfying the conditions

$$I - TT^* = FF^*,\ I - T^*T = GG^*,$$
$$I - T_0T_0^* = F^*F,\ I - T_0^*T_0 = G^*G,\ TG = FT_0, \tag{3.40}$$

The results presented in the paper [BrSv1] were reported on in a seminar of Krein's in Odessa. In the remark to this talk, M.G. Krein noticed that the conditions (3.39)–(3.40) mean that the block-operator

$$\begin{bmatrix} T_0^* & G^* \\ -F & T \end{bmatrix} : \quad \begin{bmatrix} \mathcal{F} \\ \mathcal{H} \end{bmatrix} \to \begin{bmatrix} \mathcal{G} \\ \mathcal{H} \end{bmatrix} , \tag{3.41}$$

acting in the appropriate orthogonal sums of Hilbert spaces, is a unitary operator. Starting from this remark of M.G. Krein's, mathematicians belonging to the Odessa school as well as other mathematicians, defined the operator colligation as the block operator acting from the direct sum $\begin{bmatrix} \textit{input space} \\ \textit{state space} \end{bmatrix}$ into the direct sum $\begin{bmatrix} \textit{output space} \\ \textit{state space} \end{bmatrix}$. If the spaces have scalar products and the block operator is a unitary operator with respect to this product, then the operator colligation is called an *unitary colligation*.

It should be mentioned that the paper [BrSv1] has connections to the theory of functional models of contractive operators developed in [SzNFo]. The definition (3.4) of the characteristic function of the colligation (3.2)–(3.3) agrees with the definition of the characteristic function in [BrSv1].

The notions of controllability and observability (and minimality) in the setting of State Space Theory were introduced by R.Kalman in [Kal1]. The study of controllability and observability of composite systems was first dealt with in [Gil]. Under other names, the notion of controllability also appears in the Livshitz theory of open systems. See the notions of the simple system and of the complementary component in section 1.3 of [Liv9]. (See pages 36–37 of the Russian original, or pages 27–29 of the English translation.)

The fact that every rational matrix-function S can be realized as the transfer function of some *minimal* stationary linear system (which here appears as Theorem 3.2), the uniqueness of the state space representation (Theorem (3.4)) and the equality $\dim \mathcal{H} = \deg S$ were all established by R. Kalman in a very general setting. These results, as well as many other results, can be found in Chapter 10 of [KFA]. See also Chapter 1 of [Fuh].

Some algorithms for the system realization of a given rational function were proposed by R. Kalman and his collaborators. (See Chapter 10 of the monograph [KFA] and references there.) R. Kalman did not consider questions related to the realization of contractive or inner matrix-functions: He developed system theory over arbitrary fields rather than over the field of complex numbers.

An excellent (and short!) presentation of the state space approach to the problems of minimal realization and factorization of rational functions can be found in [Kaa].

Realizations of contractive or inner rational matrix-functions (rational and more general) were later considered in the framework of the Sz. Nagy-Foias model for contractive operators. These and also more general results can be found in many publications now. For convenience, we present some basic facts on system

realizations for inner rational functions (scalar) in the Appendix to the present paper.

The state space description of the composite system, which is formed by the cascade (or Redheffer) coupling of several state space systems, was dealt with in [HeBa] in more generality. We make use of these results, but prefer to derive them independently of [HeBa] in the form and in the generality which is most suitable for our goal.

4. Coupled systems and the Schur transformation: Input-output mappings

To describe the Schur algorithm using system representations, we must first consider how the linear-fractional Schur transformation

$$\omega(z) \to s(z), \quad s(z) = \frac{s_0 + z\omega(z)}{1 + z\omega(z)\overline{s_0}} \qquad (s_0 \text{ is a complex number, } |s_0| < 1) \tag{4.1}$$

can be described in terms of the *input-output mappings of linear systems.* The linear-fractional transform (4.1) is of the form

$$s(z) = \frac{w_{11}(z)\omega(z) + w_{12}(z)}{w_{21}(z)\omega(z) + w_{22}(z)}\,. \tag{4.2}$$

This form of a linear-fractional transform is the most familiar to the classical analyst. In the theory of unitary operator colligations, the Redheffer[3] form for linear-fractional transforms, i.e.,

$$s(z) = s_{11}(z) + s_{12}(z)\omega(z)(I - s_{22}(z)\omega(z))^{-1}s_{21}(z)\,. \tag{4.3}$$

is often more convenient. Every linear-fractional transformation of the form (4.2) can be rewritten in the Redheffer form (4.3), but not every transformation in Redheffer form can be expressed in linear-fractional form.

The matrix $W(z) = \begin{bmatrix} w_{11}(z) & w_{12}(z) \\ w_{21}(z) & w_{21}(z) \end{bmatrix}$ for the transformation (4.1) and (4.2) (under the appropriate normalization[4]) is

$$W(z) = (1 - |s_0|^2)^{-1/2} \begin{bmatrix} z & s_0 \\ z\overline{s_0} & 1 \end{bmatrix}. \tag{4.4}$$

$W(z)$ in (4.4) is not an inner matrix but it is a j-inner matrix:

$$j - W^*(z)jW(z) \geq 0, \quad z \in \mathbb{D}, \qquad j - W^*(t)jW(t) = 0, \quad t \in \mathbb{T}\,,$$

where $j = \begin{bmatrix} 1 & 0 \\ 0 & -1 \end{bmatrix}$.

[3]Raymond Redheffer (1921–2005) was a US mathematician working at UCLA.

[4]The matrix of the linear-fractional transform (4.2) is determined only up to the proportionality $W(z) \to \lambda(z)W(z)$, where $\lambda \in \mathbb{C} \setminus \{0\}$.

Let us express the fractional-linear transformation (4.1) in the Redheffer form (4.3), where the 2×2-matrix-function $S(z) = \left[\begin{smallmatrix} s_{11}(z) & s_{12}(z) \\ s_{21}(z) & s_{22}(z) \end{smallmatrix}\right]$ is:

$$S(z) = \begin{bmatrix} s_0 & z\,(1 - |s_0|^2)^{1/2} \\ (1 - |s_0|^2)^{1/2} & -z\,\overline{s_0} \end{bmatrix} \tag{4.5}$$

Unlike $W(z)$, (4.4), the matrix-function $S(z)$, (4.5), is an *inner* function.

The transformation in the Redheffer form (4.3) admits an interpretation in System Theory. We discuss this in more generality than is needed for our considerations, which are centered on the linear-fractional Schur transformation.

Suppose that $\mathrm{LSDS}^{\mathrm{I}}$ and $\mathrm{LSDS}^{\mathrm{II}}$ are two linear stationary dynamical systems. In this section, we focus on the input-output mapping and do not touch on considerations related to state spaces.

Let $S(z) : \mathcal{E}^{\mathrm{I}} \to \mathcal{E}^{\mathrm{I}}$ be the transfer matrix-function of the system $\mathrm{LSDS}^{\mathrm{I}}$. Furthermore, let

$$\psi(z) = S(z)\varphi(z)$$

be the input-output mapping corresponding to the system $\mathrm{LSDS}^{\mathrm{I}}$, where $\varphi(z) : \mathbb{D} \to \mathcal{E}^{\mathrm{I}}$ is the input signal and $\psi(z) : \mathbb{D} \to \mathcal{E}^{\mathrm{I}}$ is the output signal. Suppose now that the exterior space $\mathcal{E}^{\mathrm{I}}$ of the system $\mathrm{LSDS}^{\mathrm{I}}$ is the orthogonal sum of the subspaces $\mathcal{E}^{\mathrm{I}}_1$ and $\mathcal{E}^{\mathrm{I}}_2$:

$$\mathcal{E}^{\mathrm{I}} = \mathcal{E}^{\mathrm{I}}_1 \oplus \mathcal{E}^{\mathrm{I}}_2\,. \tag{4.6}$$

Equation (4.6) suggests that the input and output signals be decomposed as follows:

$$\varphi(z) = \begin{bmatrix} \varphi_1(z) \\ \varphi_2(z) \end{bmatrix}, \qquad \psi(z) = \begin{bmatrix} \psi_1(z) \\ \psi_2(z) \end{bmatrix}, \tag{4.7}$$

And furthermore that the matrix $S(z)$ be decomposed accordingly:

$$S(z) = \begin{bmatrix} s_{11}(z) & s_{12}(z) \\ s_{21}(z) & s_{22}(z) \end{bmatrix}, \tag{4.8}$$

So that

$$\begin{bmatrix} \psi_1(z) \\ \psi_2(z) \end{bmatrix} = \begin{bmatrix} s_{11}(z) & s_{12}(z) \\ s_{21}(z) & s_{22}(z) \end{bmatrix} \begin{bmatrix} \varphi_1(z) \\ \varphi_2(z) \end{bmatrix}. \tag{4.9}$$

The system $\mathrm{LSDS}^{\mathrm{I}}$ can be considered as a linear stationary dynamical system with two input channels, corresponding to the input signals $\varphi_1(z)$ and $\varphi_2(z)$, and two output channels, corresponding to the output signals $\psi_1(z)$ and $\psi_2(z)$:

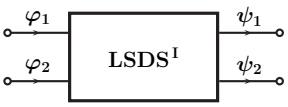

Figure 1

Let

$$\tau(z) = \omega(z)\sigma(z) \tag{4.10}$$

be the input-output mapping corresponding to the system $\mathrm{LSDS}^{\mathrm{II}}$, where $\sigma(z) : \mathbb{D} \to \mathcal{E}^{\mathrm{II}}$ is the input signal and $\tau(z) : \mathbb{D} \to \mathcal{E}^{\mathrm{II}}$ is the output signal. The system $\mathrm{LSDS}^{\mathrm{II}}$ can be considered as a linear stationary dynamical system with one input channel, corresponding to the input signal $\sigma(z)$, and one output channel, corresponding to the output signal $\tau(z)$:

Figure 2

Suppose now that

$$\mathcal{E}_2^{\mathrm{I}} = \mathcal{E}^{\mathrm{II}} \tag{4.11}$$

This allows us to 'link' the systems $\mathrm{LSDS}^{\mathrm{I}}$ and $\mathrm{LSDS}^{\mathrm{II}}$. We connect the output channel of the system $\mathrm{LSDS}^{\mathrm{II}}$ with the second $\mathrm{LSDS}^{\mathrm{I}}$ input channel and the $\mathrm{LSDS}^{\mathrm{II}}$ input channel with the second $\mathrm{LSDS}^{\mathrm{I}}$ output channel, as shown in Figure 3.

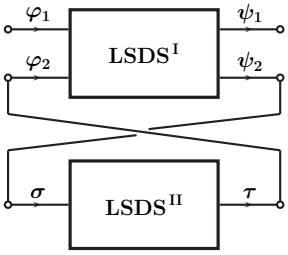

Figure 3

The resulting linear stationary dynamical system LSDS has exterior space $\mathcal{E}_1^{\mathrm{I}}$, input signal $\varphi_1(z)$ and output signal $\psi_1(z)$. The output signal $\psi_1(z)$ is linearly dependent on the input signal $\varphi_1(z)$:

$$\psi_1(z) = s(z)\,\varphi_1(z)\,, \tag{4.12}$$

where $s(z)$ is the transfer function for LSDS.

We call LSDS the Redheffer coupling of the systems $\mathrm{LSDS}^{\mathrm{I}}$ and $\mathrm{LSDS}^{\mathrm{II}}$.

We now look to express $s(z)$ in terms of $S(z)$ and $\omega(z)$. The above-described connection between the systems $\mathrm{LSDS}^{\mathrm{I}}$ and $\mathrm{LSDS}^{\mathrm{II}}$ can be formally expressed by means of the constraints

$$\varphi_2(z) = \tau(z), \qquad \psi_2(z) = \sigma(z)\,. \tag{4.13}$$

Eliminating $\varphi_2(z)$, $\psi_2(z)$, $\sigma(z)$, $\tau(z)$ from the system of linear equations (4.9), (4.10) and (4.13), we obtain the equation (4.12), where $s(z)$ has the form

$$s(z) = s_{11}(z) + s_{12}(z)\omega(z)\big(I - s_{22}(z)\omega(z)\big)^{-1}s_{21}(z)\,. \tag{4.14}$$

We now turn our attention to the '**energy relation**' associated with the linear fractional transformation (4.14): $\omega(z) \to s(z)$.

Equation (4.9) yields,

$$\varphi_1^*\varphi_1 + \varphi_2^*\varphi_2 - \psi_1^*\psi_1 - \psi_2^*\psi_2 = \begin{bmatrix}\varphi_1^* & \varphi_2^*\end{bmatrix}(I - S^*S)\begin{bmatrix}\varphi_1\\ \varphi_2\end{bmatrix}.$$

Making the substitutions $\psi_1 = s\varphi_1$, $\psi_2 = \sigma$ and $\varphi_2 = \omega\sigma$, we obtain

$$\varphi_1^*(1 - s^*s)\varphi_1 = \begin{bmatrix}\varphi_1^* & \varphi_2^*\end{bmatrix}(I - S^*S)\begin{bmatrix}\varphi_1\\ \varphi_2\end{bmatrix} + \sigma^*(1 - \omega^*\omega)\sigma\,, \tag{4.15}$$

where

$$\sigma = (1 - s_{22}\omega)^{-1}s_{21}\varphi_1, \qquad \varphi_2 = \omega(1 - s_{22}\omega)^{-1}s_{21}\varphi_1\,. \tag{4.16}$$

It follows from equation (4.15) that if $I - S^*S \geq 0$ and $1 - \omega^*\omega \geq 0$, then $1 - s^*s \geq 0$. If $I - S^*S = 0$ and $1 - \omega^*\omega = 0$, then $1 - s^*s = 0$. In particular, this brings us to:

Theorem 4.1. *Let $S(z)$ and $\omega(z)$ be rational inner matrix-functions. Furthermore, let $s(z)$ be given by the Redheffer linear-fractional transform* (4.14)*. Then $s(z)$ is a rational inner matrix-function.*

We note that the linear-fractional transform, in its classical form (4.2), is related to another kind of coupling. The relevant connection is shown in Figure 4.

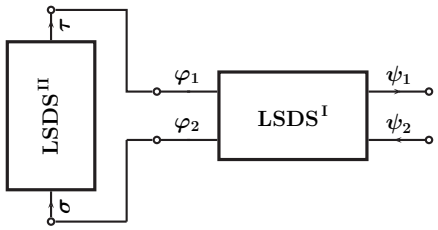

Figure 4

LSDS^{I} has two input channels with input signals $\varphi_1(z)$ and $\varphi_2(z)$. LSDS^{I} also has two output channels with output signals $\psi_1(z)$ and $\psi_2(z)$ (with frequency representation). The system LSDS^{II} has one input channel with input signal $\sigma(z)$ and output signal $\tau(z)$. We connect the LSDS^{II} output channel with the first input channel of the system LSDS^{I} as well as the LSDS^{II} input channel with the second input channel of the system LSDS^{I}. (We assume that the systems are compatible with respect to these connections, i.e., that the appropriate subspaces coincide.) We consider the second output channel of the system LSDS^{I} as the input channel of the new coupled system LSDS and the first output channel of LSDS^{I} as the

output channel of LSDS (shown in Figure 4). Let $W(z) = \left[\begin{smallmatrix} w_{11}(z) & w_{12}(z) \\ w_{21}(z) & w_{21}(z) \end{smallmatrix}\right]$ be the transfer matrix for $\mathrm{LSDS}^{\mathrm{I}}$ and $\omega(z)$ be the transfer matrix for $\mathrm{LSDS}^{\mathrm{II}}$:

$$\begin{bmatrix} \psi_1(z) \\ \psi_2(z) \end{bmatrix} = \begin{bmatrix} w_{11}(z) & w_{12}(z) \\ w_{21}(z) & w_{22}(z) \end{bmatrix} \begin{bmatrix} \varphi_1(z) \\ \varphi_2(z) \end{bmatrix}, \qquad \tau(z) = \omega(z)\sigma(z)\,.$$

The link between the systems $\mathrm{LSDS}^{\mathrm{I}}$ and $\mathrm{LSDS}^{\mathrm{II}}$, shown in Figure 3, is described by the constraints

$$\sigma(z) = \varphi_2(z), \qquad \tau(z) = \varphi_1(z)\,.$$

In which the input and the output signals of the system LSDS are denoted by $\varphi(z)$ and $\psi(z)$, respectively:

$$\varphi(z) = \psi_2(z)\,, \qquad \psi(z) = \psi_1(z)\,,$$

so that:

$$\psi(z) = s(z)\varphi(z)\,,$$

where

$$s(z) = \big(w_{11}(z)\omega(z) + w_{12}(z)\big) \cdot \big(w_{21}(z)\omega(z) + w_{22}(z)\big)^{-1}\,.$$

Historical Remark

The coupling of input-output systems having four terminals, considered in this section (See Figures 1–3), is sometimes called *cascade coupling*. This kind of coupling (as well as related mathematical questions) was investigated by R. Redheffer in [Red1]–[Red5]. Because of this, we use the name *Redheffer coupling*. Redheffer did not consider questions related to cascade coupling of state space linear systems. These questions were later addressed in [HeBa] (Without any reference to Redheffer.)

The results presented in [HeBa] are more general than here needed. We have tailored our approach to the theory of Redheffer coupling in the next two sections to fit our needs.

5. The Redheffer coupling of unitary colligations

As rational inner functions, $S(z)$, $\omega(z)$ and $s(z)$ admit system representations as characteristic functions of the unitary operator colligations with colligation operators U^{I}, U^{II} and U, respectively. We now turn to the question of how we might express U in terms of the operators U^{I} and U^{II}.

Our approach to this problem will be more general than is here called for, our goal being to describe the colligations related to Schur transformations. We assume that the unitary colligations corresponding to the systems $\mathrm{LSDS}^{\mathrm{I}}$ and $\mathrm{LSDS}^{\mathrm{II}}$ are given. We look to obtain the unitary colligation corresponding to the system LSDS, the Redheffer coupling of the systems $\mathrm{LSDS}^{\mathrm{I}}$ and $\mathrm{LSDS}^{\mathrm{II}}$. The system $\mathrm{LSDS}^{\mathrm{I}}$ is not assumed to be related to the Schur transformation. $\mathrm{LSDS}^{\mathrm{I}}$ and $\mathrm{LSDS}^{\mathrm{II}}$ can be generic systems. The only condition imposed on these systems is that the exterior space $\mathcal{E}^{\mathrm{II}}$ of the system $\mathrm{LSDS}^{\mathrm{II}}$ is identified with the subspace

$\mathcal{E}_1^{\mathrm{I}}$ of the exterior space $\mathcal{E}^{\mathrm{I}}$ belonging to $\mathrm{LSDS}^{\mathrm{I}}$. To avoid technical complications we assume that the exterior and state spaces of the systems $\mathrm{LSDS}^{\mathrm{I}}$ and $\mathrm{LSDS}^{\mathrm{II}}$ are finite-dimensional.

To simplify the notation, we denote the matrix entries of the colligation operator U^{I}, corresponding to the system $\mathrm{LSDS}^{\mathrm{I}}$, as follows

$$U^{\mathrm{I}} = \begin{bmatrix} a_{11} & a_{12} & b_1 \\ a_{21} & a_{22} & b_2 \\ c_1 & c_2 & d \end{bmatrix}, \tag{5.1}$$

where

$$a_{p,q} : \mathcal{E}_p^{\mathrm{I}} \to \mathcal{E}_q^{\mathrm{I}}, \quad b_q : \mathcal{H}^{\mathrm{I}} \to \mathcal{E}_q^{\mathrm{I}}, \quad c_p : \mathcal{E}_p^{\mathrm{I}} \to \mathcal{H}^{\mathrm{I}}, \quad d : \mathcal{H}^{\mathrm{I}} \to \mathcal{H}^{\mathrm{I}}.$$

The matrix entries for the colligation operator U^{II}, corresponding to the system $\mathrm{LSDS}^{\mathrm{II}}$, are denoted as follows:

$$U^{\mathrm{II}} = \begin{bmatrix} \alpha & \beta \\ \gamma & \delta \end{bmatrix}, \tag{5.2}$$

where

$$\alpha : \mathcal{E}^{\mathrm{II}} \to \mathcal{E}^{\mathrm{II}}, \quad \beta : \mathcal{H}^{\mathrm{II}} \to \mathcal{E}^{\mathrm{II}}, \quad \gamma : \mathcal{E}^{\mathrm{II}} \to \mathcal{H}^{\mathrm{II}}, \quad \delta : \mathcal{H}^{\mathrm{II}} \to \mathcal{H}^{\mathrm{II}}.$$

The linear equations describing the dynamics of the system $\mathrm{LSDS}^{\mathrm{I}}$ are

$$\begin{bmatrix} \psi_1(z) \\ \psi_2(z) \\ z^{-1}h(z) \end{bmatrix} = \begin{bmatrix} a_{11} & a_{12} & b_1 \\ a_{21} & a_{22} & b_2 \\ c_1 & c_2 & d \end{bmatrix} \begin{bmatrix} \varphi_1(z) \\ \varphi_2(z) \\ h(z) \end{bmatrix}, \tag{5.3}$$

where

$$\varphi(z) = \begin{bmatrix} \varphi_1(z) \\ \varphi_2(z) \end{bmatrix}, \quad \psi(z) = \begin{bmatrix} \psi_1(z) \\ \psi_2(z) \end{bmatrix} \quad \text{and} \quad h(z)$$

are, respectively, the input signal, the output signal and the inner state signal corresponding to the system $\mathrm{LSDS}^{\mathrm{I}}$.

The linear equations describing the dynamics of the system $\mathrm{LSDS}^{\mathrm{II}}$ are

$$\begin{bmatrix} \tau(z) \\ z^{-1}l(z) \end{bmatrix} = \begin{bmatrix} \alpha & \beta \\ \gamma & \delta \end{bmatrix} \begin{bmatrix} \sigma(z) \\ l(z) \end{bmatrix}, \tag{5.4}$$

where $\sigma(z)$, $\tau(z)$ and $l(z)$ are, respectively, the input signal, output signal and the interior state signal corresponding to the system $\mathrm{LSDS}^{\mathrm{II}}$.

The constraints

$$\tau(z) = \varphi_2(z), \quad \sigma(z) = \psi_2(z) \tag{5.5}$$

correspond to the Redheffer coupling of the systems $\mathrm{LSDS}^{\mathrm{I}}$ and $\mathrm{LSDS}^{\mathrm{II}}$.

We now aim to eliminate the variables $\varphi_2(z)$, $\psi_2(z)$, $\sigma(z)$, $\tau(z)$ from the systems (5.3), (5.4), (5.5). To this end, we substitute the expressions $\alpha\sigma(z) + \beta l(z)$ and $\sigma(z)$ for the variables $\varphi_2(z)$ and $\psi_2(z)$ into the equation

$$\psi_2(z) = a_{21}\varphi_1(z) + a_{22}\varphi_2(z) + \beta h(z)\,.$$

With this we can express $\sigma(z)$ in terms of $\varphi_1(z)$, $h(z)$ and $l(z)$:

$$\begin{aligned}&\sigma(z)\\ &= (1 - a_{22}\,\alpha)^{-1}\, a_{21}\,\varphi_1(z) + (1 - a_{22}\,\alpha)^{-1}\, b_2\, h(z) + (1 - a_{22}\,\alpha)^{-1}\,\beta\, l(z)\,. \quad (5.6)\end{aligned}$$

Substituting this expressions for σ into (5.3), (5.4), (5.5), we obtain

$$\begin{bmatrix} \psi_1(z) \\ z^{-1}h(z) \\ z^{-1}l(z) \end{bmatrix} = \begin{bmatrix} A & B_1 & B_2 \\ C_1 & D_{11} & D_{12} \\ C_2 & D_{21} & D_{21} \end{bmatrix} \begin{bmatrix} \varphi_1(z) \\ h(z) \\ l(z) \end{bmatrix}\,, \quad (5.7)$$

where

$$\begin{aligned} A : \mathcal{E}_1^{\mathrm{I}} \to \mathcal{E}_1^{\mathrm{I}},\ B_1 : \mathcal{H}^{\mathrm{I}} \to \mathcal{E}_1^{\mathrm{I}},\ B_2 : \mathcal{H}^{\mathrm{II}} \to \mathcal{E}_1^{\mathrm{I}},\\ C_1 : \mathcal{E}_1^{\mathrm{I}} \to \mathcal{H}^{\mathrm{I}},\ C_2 : \mathcal{H}^{\mathrm{II}} \to \mathcal{E}_1^{\mathrm{I}},\\ D_{11} : \mathcal{H}^{\mathrm{I}} \to \mathcal{H}^{\mathrm{I}},\ \ D_{12} : \mathcal{H}^{\mathrm{II}} \to \mathcal{H}^{\mathrm{I}},\ \ D_{21} : \mathcal{H}^{\mathrm{I}} \to \mathcal{H}^{\mathrm{II}},\ \ D_{22} : \mathcal{H}^{\mathrm{II}} \to \mathcal{H}^{\mathrm{II}}\,. \end{aligned}$$

The matrix

$$U = \begin{bmatrix} A & B_1 & B_2 \\ C_1 & D_{11} & D_{12} \\ C_2 & D_{21} & D_{21} \end{bmatrix} \quad (5.8)$$

can be expressed using the entries of the matrices U^{I}, (5.1), and U^{II}, (5.2), as follows:

$$U = \begin{bmatrix} a_{11} & b_1 & a_{12}\,\beta \\ c_1 & d & c_2\,\beta \\ 0 & 0 & \delta \end{bmatrix} + \begin{bmatrix} a_{12}\,\alpha \\ c_2\,\alpha \\ \gamma \end{bmatrix} \cdot \big(1 - a_{22}\,\alpha\big)^{-1} \cdot \begin{bmatrix} a_{21} & b_2 & a_{22}\,\beta \end{bmatrix}\,. \quad (5.9)$$

The operator U is called the Redheffer product of the operators U_1 and U_2.

We again turn our attention to the 'energy relation' associated with the operators U^{I}, U^{II} and U. Suppose that U^{I} and U^{II} are unitary. Let $\varphi_1 \in \mathcal{E}_1^{\mathrm{I}}$, $\varphi_2 \in \mathcal{E}_2^{\mathrm{I}}$, $h \in \mathcal{H}^{\mathrm{I}}$, $\sigma \in \mathcal{E}^{\mathrm{II}}$ and $l \in \mathcal{H}^{\mathrm{II}}$ be arbitrary vectors. If $\psi_1 \in \mathcal{E}_1^{\mathrm{I}}$, $\psi_2 \in \mathcal{E}_2^{\mathrm{I}}$, $k \in \mathcal{H}^{\mathrm{I}}$, $\tau \in \mathcal{E}^{\mathrm{II}}$ and $m \in \mathcal{H}^{\mathrm{II}}$ are defined by the equalities

$$\begin{bmatrix} \psi_1 \\ \psi_2 \\ k \end{bmatrix} = U^{\mathrm{I}} \begin{bmatrix} \varphi_1 \\ \varphi_2 \\ h \end{bmatrix}\,, \quad \begin{bmatrix} \tau \\ m \end{bmatrix} = U^{\mathrm{II}} \begin{bmatrix} \sigma \\ l \end{bmatrix}\,,$$

then

$$||\psi_1||^2 + ||\psi_2||^2 + ||k||^2 = ||\varphi_1||^2 + ||\varphi_2||^2 + ||h||^2\,, \quad (5.10)$$

and

$$||\tau||^2 + ||m||^2 = ||\sigma||^2 + ||l||^2\,. \quad (5.11)$$

For arbitrary φ_1, h, l and

$$\sigma = (1 - a_{22}\,\alpha)^{-1}\, a_{21}\,\varphi_1 + (1 - a_{22}\,\alpha)^{-1}\, b_2 \, + (1 - a_{22}\,\alpha)^{-1}\,\beta\,, \tag{5.12}$$

$$\varphi_2 = \alpha\big((1 - a_{22}\,\alpha)^{-1}\, a_{21}\,\varphi_1 + (1 - a_{22}\,\alpha)^{-1}\, b_2 \, + (1 - a_{22}\,\alpha)^{-1}\,\beta\big) + \beta l\,, \tag{5.13}$$

it follows that

$$\psi_2 = \sigma, \qquad \tau = \varphi_2\,,$$

and

$$||\psi_1||^2 + ||l||^2 + ||m||^2 = ||\varphi_1||^2 + ||h||^2 + ||l||^2\,. \tag{5.14}$$

According to the definition of the operator U,

$$\begin{bmatrix} \psi_1 \\ k \\ m \end{bmatrix} = U \begin{bmatrix} \varphi_1 \\ h \\ l \end{bmatrix}\,. \tag{5.15}$$

Since φ_1, h, l are arbitrary, equality (5.14) means that U is unitary. This operator, partitioned into blocks according to (5.8), is related to the unitary colligation $(\mathcal{E}, \mathcal{H}, U)$, where $\mathcal{E} = \mathcal{E}^{\mathrm{I}}$, $\mathcal{H} = \mathcal{H}^{\mathrm{I}} \oplus \mathcal{H}^{\mathrm{II}}$.

Definition 5.1. *The colligation $(\mathcal{E}, \mathcal{H}, U)$ is called the Redheffer coupling of the colligations $(\mathcal{E}^{\mathrm{I}}, \mathcal{H}^{\mathrm{I}}, U^{\mathrm{I}})$ and $(\mathcal{E}^{\mathrm{II}},\ \mathcal{H}^{\mathrm{II}}, U^{\mathrm{II}})$.*

Theorem 5.1. *Let $S(z) = \left[\begin{smallmatrix} s_{11}(z) & s_{12}(z) \\ s_{21}(z) & s_{22}(z) \end{smallmatrix}\right]$, $\omega(z)$ and $s(z)$ be the characteristic functions of the colligations $(\mathcal{E}^I, \mathcal{H}^I, U^I)$, $(\mathcal{E}^{II}, \mathcal{H}^{II}, U^{II})$ and their Redheffer coupling $(\mathcal{E}, \mathcal{H}, U)$, respectively:*

$$\begin{bmatrix} s_{11}(z) & s_{12}(z) \\ s_{21}(z) & s_{22}(z) \end{bmatrix} = \begin{bmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{bmatrix} + z \begin{bmatrix} b_1 \\ b_2 \end{bmatrix} (I - zd)^{-1} \begin{bmatrix} c_1 & c_2 \end{bmatrix}\,, \tag{5.16}$$

$$\omega(z) = \alpha + z\beta(1 - z\delta)^{-1}\gamma\,, \tag{5.17}$$

$$s(z) = A + z \begin{bmatrix} B_1 \\ B_2 \end{bmatrix} \left(\begin{bmatrix} I_{\mathcal{H}^{\mathrm{I}}} & 0 \\ 0 & I_{\mathcal{H}^{\mathrm{II}}} \end{bmatrix} - z \begin{bmatrix} D_{11} & D_{12} \\ D_{21} & D_{22} \end{bmatrix} \right)^{-1} \begin{bmatrix} C_1 & C_2 \end{bmatrix}\,. \tag{5.18}$$

(*The notation for the entries of the matrices U^I, U^{II} and U is taken from* (5.1), (5.2) *and* (5.8), *respectively.*)

Then

$$s(z) = s_{11}(z) + s_{12}(z)\omega(z)(I - s_{22}(z)\omega(z))^{-1}s_{21}(z)\,. \tag{5.19}$$

6. The inverse Schur transformation and Redheffer couplings of colligations

We now focus again on the linear-fractional transformation (4.1) in the Redheffer form (4.3), where $\omega(z)$ is a rational inner matrix-function of degree $n-1$, so that $s(z)$ is a rational inner matrix-function of degree n.

The function $S(z)$, which appears in (4.5) is a rational inner function. It is a characteristic matrix-function for the operator colligation $(\mathcal{E}^{\mathrm{I}},\ \mathcal{H}^{\mathrm{I}},\ U^{\mathrm{I}})$, which we now describe.

The outer space $\mathcal{E}^{\mathrm{I}}$ is two-dimensional. We identify $\mathcal{E}^{\mathrm{I}}$ with $\mathbb{C}^2$. The space $\mathcal{E}^{\mathrm{I}}$ is considered as the orthogonal sum $\mathcal{E}^{\mathrm{I}} = \mathcal{E}^{\mathrm{I}}_1 \oplus \mathcal{E}^{\mathrm{I}}_2$, where $\mathcal{E}^{\mathrm{I}}_1$ is identified with $\mathbb{C}$ and $\mathcal{E}^{\mathrm{I}}_2$ is identified with $\mathbb{C}$. The orthogonal decomposition $\mathcal{E}^{\mathrm{I}} = \mathcal{E}^{\mathrm{I}}_1 \oplus \mathcal{E}^{\mathrm{I}}_2$ is thus the canonical decomposition $\mathbb{C}^2 = \mathbb{C} \oplus \mathbb{C}$. The inner space $\mathcal{H}^{\mathrm{I}}$ is one-dimensional. We identify $\mathcal{H}^{\mathrm{I}}$ with $\mathbb{C}^1$. The colligation operator U^{I} is defined by the unitary $3 \times 3 = (2+1) \times (2+1)$-matrix considered as an operator acting in $\mathbb{C}^3 = \mathbb{C}^2 \oplus \mathbb{C}^1$:

$$U^{\mathrm{I}} = \begin{bmatrix} A^{\mathrm{I}} & B^{\mathrm{I}} \\ C^{\mathrm{I}} & D^{\mathrm{I}} \end{bmatrix} \tag{6.1}$$

with

$$A^{\mathrm{I}} = \begin{bmatrix} s_0 & 0 \\ (1-|s_0|^2)^{1/2} & 0 \end{bmatrix},\ B^{\mathrm{I}} = \begin{bmatrix} (1-|s_0|^2)^{1/2} \\ -\overline{s_0}, \end{bmatrix},$$

$$C^{\mathrm{I}} = \begin{bmatrix} 0 & 1 \end{bmatrix},\quad D^{\mathrm{I}} = \begin{bmatrix} 0 \end{bmatrix}.$$

The characteristic function of the colligation $(\mathcal{E}^{\mathrm{I}}, \mathcal{H}^{\mathrm{I}}, U^{\mathrm{I}})$ is the matrix-function $S(z)$ of the form (4.5):

$$\begin{bmatrix} s_0 & z\,(1-|s_0|^2)^{1/2} \\ (1-|s_0|^2)^{1/2} & -z\,\overline{s_0} \end{bmatrix} = A^{\mathrm{I}} + zB^{\mathrm{I}}(I - zD^{\mathrm{I}})^{-1}C^{\mathrm{I}}. \tag{6.2}$$

The rational inner function $\omega(z)$ of degree $n-1$ is the characteristic function of the colligation $(\mathcal{E}^{\mathrm{II}}, \mathcal{H}^{\mathrm{II}}, U^{\mathrm{II}})$. The outer space $\mathcal{E}^{\mathrm{II}}$ is one-dimensional and is identified with $\mathbb{C}$ and the inner space $\mathcal{H}^{\mathrm{II}}$ is $(n-1)$-dimensional and is identified with $\mathbb{C}^{n-1}$. The colligation operator U^{II} thus acts in $\mathbb{C}^n = \mathbb{C} \oplus \mathbb{C}^{n-1}$. We identify the operator U^{II} with its matrix in the canonical basis of $\mathbb{C}^n$:

$$U^{\mathrm{II}} = \begin{bmatrix} \alpha & \beta \\ \gamma & \delta \end{bmatrix}, \tag{6.3}$$

where

$$\alpha \in \mathfrak{M}_{1\times 1},\ \beta \in \mathfrak{M}_{1\times(n-1)},\ \gamma \in \mathfrak{M}_{(n-1)\times 1},\ \delta \in \mathfrak{M}_{(n-1)\times(n-1)}.$$

α is simply a complex number. The matrix U^{II} is unitary. The system representation of the function $\omega(z)$ is given by:

$$\omega(z) = \alpha + \beta(1 - z\delta)^{-1}\gamma. \tag{6.4}$$

In particular,

$$\omega(0) = \alpha. \tag{6.5}$$

The function

$$s(z) = \frac{s_0 + z\omega(z)}{1 + z\omega(z)\overline{s_0}}, \tag{6.6}$$

written as a Redheffer fractional-linear transform, takes the form:

$$s(z) = s_0 + z(\,1 - |s_0|^2\,)^{1/2}\,\omega(z)\,(\,1 + z\omega(z)\overline{s_0}\,)^{-1}\,(1 - |s_0|^2)^{1/2}\,, \tag{6.7}$$

and admits a system realization by means of the operator colligation $(\mathcal{E}, \mathcal{H}, U)$, where $(\mathcal{E}, \mathcal{H}, U)$ is the Redheffer coupling of the colligations $(\mathcal{E}^{\mathrm{I}}, \mathcal{H}^{\mathrm{I}}, U^{\mathrm{I}})$, representing the function $S(z)$, and $(\mathcal{E}^{\mathrm{II}}, \mathcal{H}^{\mathrm{II}}, U^{\mathrm{II}})$, representing the function $\omega(z)$.

Clearly, $\mathcal{E} = \mathbb{C}$ and $\mathcal{H} = \mathbb{C}^n$. U is the Redheffer coupling of the matrices U^{I} and U^{II}. Applying formula (5.9) to $U^{\,\mathrm{I}}$ and $U^{\,\mathrm{II}}$, we obtain:

$$U = \begin{bmatrix} s_0 & (1 - |s_0|^2)^{1/2} & 0_{1\times(n-1)} \\ \alpha\,(1 - |s_0|^2)^{1/2} & -\alpha\,\overline{s_0} & \beta \\ \gamma\,(1 - |s_0|^2)^{1/2} & -\gamma\,\overline{s_0} & \delta \end{bmatrix}, \tag{6.8}$$

so that U takes the form:

$$U = \begin{bmatrix} A & B \\ C & D \end{bmatrix}, \tag{6.9}$$

where

$$A = s_0, \qquad B = \begin{bmatrix} (1 - |s_0|^2)^{1/2} & 0_{1\times(n-1)} \end{bmatrix},$$

$$C = \begin{bmatrix} \alpha\,(1 - |s_0|^2)^{1/2} \\ \gamma\,(1 - |s_0|^2)^{1/2} \end{bmatrix}, \qquad D = \begin{bmatrix} -\alpha\,\overline{s_0} & \beta \\ -\gamma\,\overline{s_0} & \delta \end{bmatrix},$$

$$A \in \mathfrak{M}_{1\times 1}\,,\ B \in \mathfrak{M}_{1\times n}\,,\ B \in \mathfrak{M}_{n\times 1},\ D \in \mathfrak{M}_{n\times n}\,. \tag{6.10}$$

Clearly, U in (6.8)–(6.9) can be expressed as follows:

$$U = \begin{bmatrix} 1 & 0 & 0_{1\times(n-1)} \\ 0 & \alpha & \beta \\ 0_{(n-1)\times 1} & \gamma & \delta \end{bmatrix} \begin{bmatrix} s_0 & (1 - |s_0|^2)^{1/2} & 0_{1\times(n-1)} \\ (1 - |s_0|^2)^{1/2} & -\overline{s_0} & 0_{1\times(n-1)} \\ 0_{(n-1)\times 1} & 0_{(n-1)\times 1} & 1_{(n-1)\times(n-1)} \end{bmatrix}. \tag{6.11}$$

Applying Theorem 5.1 to the Redheffer coupling of the colligations U^{I}, (6.1), and U^{II}, (6.1), yields:

Theorem 6.1. *Let $\omega(z)$ be an rational inner matrix-function of degree $n-1$ and let*

$$\begin{bmatrix} \alpha & \beta \\ \gamma & \delta \end{bmatrix},\ \alpha \in \mathfrak{M}_{1\times 1},\, \beta \in \mathfrak{M}_{1\times(n-1)},\, \gamma \in \mathfrak{M}_{(n-1)\times 1},\, \delta \in \mathfrak{M}_{(n-1)\times(n-1)}\,, \tag{6.12}$$

be a unitary matrix so that the system representation (6.4) *for $\omega(z)$ holds. Let s_0 be a complex number with $|s_0| < 1$. Let the function $s(z)$ be defined as the inverse Schur transform* (6.6) (*using s_0 and $\omega(z)$*) *and let the matrix U,*

$$U = \begin{bmatrix} A & B \\ C & D \end{bmatrix},\ A \in \mathfrak{M}_{1\times 1},\, B \in \mathfrak{M}_{1\times(n)},\, C \in \mathfrak{M}_{(n)\times 1},\, D \in \mathfrak{M}_{(n)\times(n)}\,, \tag{6.13}$$

be defined by equation (6.11).

U is then unitary and yields the system representation of $s(z)$:

$$s(z) = A + zB(I - zD)^{-1}C\,. \tag{6.14}$$

Unitary equivalence freedom

The same function $s(z)$, for which we earlier found a representation using the matrix U in (6.11), can also be represented with the help of a matrix having the form:

$$U^{\,V} = \begin{bmatrix} 1 & 0_{n\times n} \\ 0_{n\times n} & V^* \end{bmatrix} U \begin{bmatrix} 1 & 0_{n\times n} \\ 0_{n\times n} & V \end{bmatrix}, \tag{6.15}$$

where $V \in \mathfrak{M}_{n\times n}$ is a unitary matrix.

The matrix representing $s(z)$ and which, furthermore, *appears as the Redheffer coupling matrix* for the matrices representing $S(z)$ and $\omega(z)$, can be considered to have fewer 'degrees of freedom' than matrices of the form (6.15). The degree of freedom for the Redheffer coupling matrix is derived from this same property in the Redheffer coupled matrices. The more general form of the matrix, which represents the 2×2-matrix-function $S(z)$, is the 'transformed' matrix:

$$U^{\,\mathrm{I},\,\varepsilon} = \begin{bmatrix} 1_{\,2\times 2} & 0_{\,2\times 1} \\ 0_{\,1\times 2} & \overline{\varepsilon} \end{bmatrix} U^{\,\mathrm{I}} \begin{bmatrix} 1_{\,2\times 2} & 0_{\,2\times 1} \\ 0_{\,1\times 2} & \varepsilon \end{bmatrix}, \tag{6.16}$$

i.e.,

$$U^{\,\mathrm{I},\,\varepsilon} = \begin{bmatrix} A^{\,\mathrm{I}} & B^{\,\mathrm{I}}\varepsilon \\ \overline{\varepsilon}\,C^{\,\mathrm{I}} & D^{\,\mathrm{I}} \end{bmatrix}, \tag{6.17}$$

where $U^{\,\mathrm{I}}$ is the matrix from (6.1) and ε is an arbitrary unimodular complex number. A more general form of the colligation matrix representing the function $\omega(z)$ is given by:

$$U^{\,\mathrm{II},\,v} = \begin{bmatrix} 1 & 0_{1\times(n-1)} \\ 0_{(n-1)\times 1} & v^* \end{bmatrix} U^{\,\mathrm{II}} \begin{bmatrix} 1 & 0_{1\times(n-1)} \\ 0_{(n-1)\times 1} & v \end{bmatrix}, \tag{6.18}$$

i.e.

$$,U^{\,\mathrm{II},\,v} = \begin{bmatrix} \alpha & \beta^{\,v} \\ \gamma^{\,v} & \delta^{\,v} \end{bmatrix}, \tag{6.19}$$

where $U^{\,\mathrm{II}}$, (6.3) , is some $n\times n$ unitary colligation matrix representing the function $\omega(z)$,

$$\beta^{\,v} = \beta\, v, \qquad \gamma^{\,v} = v^*\gamma, \qquad \delta^{\,v} = v^*\delta\, v\,, \tag{6.20}$$

and v is an arbitrary unitary $(n-1)\times(n-1)$-matrix. Applying formula (5.9) to the matrices $U^{\,\mathrm{I},\,\varepsilon}$ and $U^{\,\mathrm{II},\,v}$, we obtain the Redheffer coupling matrix:

$$U^{\,\varepsilon,\,v} = \begin{bmatrix} s_0 & \varepsilon\,(1-|s_0|^2)^{1/2} & 0_{1\times(n-1)} \\ \alpha\,\overline{\varepsilon}\,(1-|s_0|^2)^{1/2} & -\alpha\,\overline{s_0} & \overline{\varepsilon}\,\beta^{\,v} \\ \gamma^{\,v}\,(1-|s_0|^2)^{1/2} & -\gamma\,\varepsilon\,\overline{s_0} & \delta^{\,v} \end{bmatrix}. \tag{6.21}$$

Clearly,

$$U^{\varepsilon,v} = \begin{bmatrix} 1 & 0 & 0_{1\times(n-1)} \\ 0 & \alpha & \overline{\varepsilon}\,\beta^{v} \\ 0_{(n-1)\times 1} & \gamma^{v}\varepsilon & \delta^{v} \end{bmatrix} \times \begin{bmatrix} s_0 & \varepsilon\,(1-|s_0|^2)^{1/2} & 0_{1\times(n-1)} \\ \overline{\varepsilon}\,(1-|s_0|^2)^{1/2} & -\overline{s_0} & 0_{1\times(n-1)} \\ 0_{(n-1)\times 1} & 0_{(n-1)\times 1} & 1_{(n-1)\times(n-1)} \end{bmatrix}, \quad (6.22)$$

and finally

$$U^{\varepsilon,\,v} = \begin{bmatrix} 1 & 0_{n\times n} \\ 0_{n\times n} & V^{*}_{\varepsilon,\,v} \end{bmatrix} U \begin{bmatrix} 1 & 0_{n\times n} \\ 0_{n\times n} & V_{\varepsilon,\,v} \end{bmatrix}, \quad (6.23)$$

where

$$V_{\varepsilon,v} = \begin{bmatrix} \varepsilon & 0_{1\times(n-1)} \\ 0_{(n-1)\times 1} & v \end{bmatrix}, \quad (6.24)$$

ε is an arbitrary unimodular complex number and v is an arbitrary unitary $(n-1)\times(n-1)$-matrix (ε and v are the same as in (6.18)).

Comparing formulas (6.15) and (6.23)–(6.24), we see that the matrices $U^{\varepsilon,\,v}$ which come from Redheffer coupling of the matrices representing $S(z)$ and $\omega(z)$ are special. The distinguishing feature of the matrices $U^{\varepsilon,\,v}$ can be summarized as follows:

Among all of the $(n+1)\times(n+1)$-matrices $U^{V} = \begin{bmatrix} s_0 & B^{V} \\ C^{V} & D^{V} \end{bmatrix}$ of the form (6.15)*, it is precisely those for which the block-matrix entry B^{V} takes the form*

$$B^{V} = \begin{bmatrix} \varepsilon\,(1-|s_0|^2)^{1/2} & 0_{1\times(n-1)} \end{bmatrix},$$

where ε is an arbitrary unimodular complex number, that can be expressed as in (6.23)–(6.24)*.*

7. One step of the Schur algorithm, expressed in the language of colligations

The results from Section 6 can be summarized as follows: Starting from the unitary $n\times n$-matrix $\left[\begin{smallmatrix} \alpha & \beta \\ \gamma & \delta \end{smallmatrix}\right]$ representing a given inner rational matrix-function $\omega(z)$ of degree n,

$$\omega(z) = \alpha + z\beta(I - z\delta)^{-1}\gamma\,,$$

we constructed the unitary $(n+1)\times(n+1)$-matrix $\left[\begin{smallmatrix} s_0 & B \\ C & D \end{smallmatrix}\right]$ representing the function $s(z)$:

$$s(z) = s_0 + zB(I - zD)^{-1}C\,,$$

where $s(z)$ is the inverse Schur transform (6.6).

Our goal is not, however, to determine $s(z)$ from $\omega(z)$, but instead to start with $s(z)$ and determine $\omega(z)$. We look to describe a step of the Schur algorithm when applied to a rational inner function $s(z)$,

$$s(z) \longrightarrow \omega(z) = \frac{1}{z}\frac{s(z)-s_0}{1-s(z)\overline{s_0}}\,, \qquad s_0 = s(0)\,,$$

in terms of system representations. In other words, we would like to find the unitary matrix $\begin{bmatrix}\alpha & \beta\\ \gamma & \delta\end{bmatrix}$ representing $\omega(z)$, starting from the matrix U representing the function $s(z)$.

Equation (6.11) serves as a heuristic argument. Until now, $\begin{bmatrix}\alpha & \beta\\ \gamma & \delta\end{bmatrix}$ was given and U was unknown. Now we assume that the unitary matrix U is given and that the matrix $\begin{bmatrix}\alpha & \beta\\ \gamma & \delta\end{bmatrix}$ is unknown. We consider (6.11) as an equation with respect to the matrix $\begin{bmatrix}\alpha & \beta\\ \gamma & \delta\end{bmatrix}$ and U as given. Because the second factor on the right-hand side of (6.11) is a unitary matrix, the solution matrix $\begin{bmatrix}\alpha & \beta\\ \gamma & \delta\end{bmatrix}$ (if it exists) for equation (6.11) is also a unitary matrix.

For a *general* unitary matrix U, equation (6.11) has *no solution* with respect to the matrix $\left[\begin{array}{cc}\alpha & \beta\\ \gamma & \delta\end{array}\right]$: We know that the block-matrix entry B in $U = \begin{bmatrix}s_0 & B\\ C & D\end{bmatrix}$ (U as in (6.11)) is necessarily of the form $B = \begin{bmatrix}(1-|s_0|^2)^{1/2} & 0_{1\times(n-1)}\end{bmatrix}$.

Since the characteristic functions of unitarily equivalent colligations coincide, it is enough to find a solution for (6.11) with U replaced by some matrix U^V of the form (6.15):

$$U^V = \begin{bmatrix} 1 & 0 & 0_{1\times(n-1)} \\ 0 & \alpha & \beta \\ 0_{(n-1)\times 1} & \gamma & \delta \end{bmatrix} \begin{bmatrix} s_0 & (1-|s_0|^2)^{1/2} & 0_{1\times(n-1)} \\ (1-|s_0|^2)^{1/2} & -\overline{s_0} & 0_{1\times(n-1)} \\ 0_{(n-1)\times 1} & 0_{(n-1)\times 1} & 1_{(n-1)\times(n-1)} \end{bmatrix}, \tag{7.1}$$

Lemma 7.1. *Given a unitary $(n+1)\times(n+1)$-matrix U, the unitary $n\times n$-matrix V can be chosen such that equation* (7.1) *has a solution.*

Lemma 7.2. *Given a unitary $(n+1)\times(n+1)$-matrix U:*

$$U = \left[\begin{array}{cc} s_0 & B \\ C & D \end{array}\right], \tag{7.2}$$

we can find a unitary $n\times n$-matrix V_0 such that U^{V_0}, given by

$$U^{V_0} = \begin{bmatrix} 1 & 0_{n\times n} \\ 0_{n\times n} & V_0^* \end{bmatrix} U \begin{bmatrix} 1 & 0_{n\times n} \\ 0_{n\times n} & V_0 \end{bmatrix},$$

takes the form $U^{V_0} = U^0$*, where*

$$U^0 = \begin{bmatrix} s_0 & B_0 \\ C_0 & D_0 \end{bmatrix}, \tag{7.3}$$

and the block-matrix entry $B_0 \in \mathfrak{M}_{1\times n}$ *is*

$$B_0 = \begin{bmatrix} (1-|s_0|^2)^{1/2} & \cdots & 0_{1\times(n-1)} \end{bmatrix}. \tag{7.4}$$

Proof. The row-vectors B and B_0 satisfy the condition

$$BB^* = B_0 B_0^* \quad \big(= (1-|s_0|^2)\big)$$

The equality $B_0 B_0^* = 1-|s_0|^2$ follows from the definition of B_0, (7.4). The equality $s_0\overline{s_0}+BB^* = 1$ holds, since the matrix U, (7.2), is unitary. Applying Lemma 8.1 to the row-vectors B and B_0, we find the unitary $n\times n$-matrix V_0 such that $BV_0 = B_0$. For every such choice of V_0, the matrix U^{V_0} has the form (7.3)–(7.4).

Remark 7.1. *If* $n > 1$*, the matrices* U^0 *and* V_0 *are not uniquely defined. The row-vector* B *of any matrix* U^V *with* V *of the form* $V = V_0 \left[\begin{smallmatrix} 1 & 0 \\ 0 & v \end{smallmatrix}\right]$*, where* v *is an arbitrary unitary* $(n-1)\times(n-1)$*-matrix is also of the form* (7.4)*.*

Theorem 7.1. *Given a unitary* $(n+1)\times(n+1)$*-matrix of the form*

$$U^0 = \begin{bmatrix} s_0 & (1-|s_0|^2)^{1/2} & 0_{1\times(n-1)} \\ c_1 & d_{11} & d_{12} \\ c_2 & d_{21} & d_{22} \end{bmatrix}, \tag{7.5}$$

where $s_0 \in \mathbb{C}$, $|s_0| \le 1$, $n \ge 2$,

$$\begin{array}{ccc} c_1 \in \mathfrak{M}_{1\times 1}, & d_{11} \in \mathfrak{M}_{1\times 1}, & d_{12} \in \mathfrak{M}_{1\times(n-1)}, \\ c_2 \in \mathfrak{M}_{(n-1)\times 1}, & d_{21} \in \mathfrak{M}_{(n-1)\times 1}, & d_{22} \in \mathfrak{M}_{(n-1)\times(n-1)}, \end{array}$$

the equation

$$U^0 = \begin{bmatrix} 1 & 0 & 0_{1\times(n-1)} \\ 0 & \alpha & \beta \\ 0_{(n-1)\times 1} & \gamma & \delta \end{bmatrix} \begin{bmatrix} s_0 & (1-|s_0|^2)^{1/2} & 0_{1\times(n-1)} \\ (1-|s_0|^2)^{1/2} & -\overline{s_0} & 0_{1\times(n-1)} \\ 0_{(n-1)\times 1} & 0_{(n-1)\times 1} & 1_{(n-1)\times(n-1)} \end{bmatrix}, \tag{7.6}$$

where

$$\begin{array}{cc} \alpha \in \mathfrak{M}_{1\times 1}, & \beta \in \mathfrak{M}_{1\times(n-1)}, \\ \gamma \in \mathfrak{M}_{(n-1)\times 1}, & \delta \in \mathfrak{M}_{(n-1)\times(n-1)}, \end{array}$$

has a solution with respect to the matrix

$$U^1 = \begin{bmatrix} \alpha & \beta \\ \gamma & \delta \end{bmatrix}. \tag{7.7}$$

The solution of this equation can be expressed as

$$\begin{bmatrix} \alpha & \beta \\ \gamma & \delta \end{bmatrix} = \begin{bmatrix} -d_{11}\, s_0 + c_1\,(1-|s_0|^2)^{1/2} & d_{12} \\ -d_{21}\, s_0 + c_2\,(1-|s_0|^2)^{1/2} & d_{22} \end{bmatrix} \tag{7.8}$$

Proof of Theorem 7.1. We consider equation (7.6) in further detail. If this equation is solvable, then

$$\begin{bmatrix} s_0 & (1-|s_0|^2)^{1/2} & 0_{1\times(n-1)} \\ c_1 & d_{11} & d_{12} \\ c_2 & d_{21} & d_{22} \end{bmatrix} \times \begin{bmatrix} \overline{s_0} & (1-|s_0|^2)^{1/2} & 0_{1\times(n-1)} \\ (1-|s_0|^2)^{1/2} & -s_0 & 0_{1\times(n-1)} \\ 0_{(n-1)\times 1} & 0_{(n-1)\times 1} & 1_{(n-1)\times(n-1)} \end{bmatrix} = \begin{bmatrix} 1 & 0 & 0_{1\times(n-1)} \\ 0 & \alpha & \beta \\ 0_{(n-1)\times 1} & \gamma & \delta \end{bmatrix} \tag{7.9}$$

Multiplying the matrices on the left-hand side of (7.9), we see that their product is of the form $\left[\begin{smallmatrix} 1 & 0 & 0 \\ * & * & * \\ * & * & * \end{smallmatrix}\right]$. Since the matrix U^0, (7.5), is unitary, the scalar product of its different rows vanishes. The fact that the first row of this matrix is orthogonal to each other row can be expressed as

$$\begin{bmatrix} c_1 \\ c_2 \end{bmatrix} \overline{s_0} + \begin{bmatrix} d_{11} \\ d_{21} \end{bmatrix} (1-|s_0|^2)^{1/2} = 0\,. \tag{7.10}$$

The latter equalities mean that the product of the matrices on the left-hand side of (7.9) takes the form $\left[\begin{smallmatrix} * & * & * \\ 0 & * & * \\ 0 & * & * \end{smallmatrix}\right]$. Thus, the product of the matrices on the left-hand side of (7.9) has the desired form $\left[\begin{smallmatrix} 1 & 0 & 0 \\ 0 & * & * \\ 0 & * & * \end{smallmatrix}\right]$. Multiplying out the matrices in (7.9), we obtain (7.8). □

Remark 7.2. *In view of* (7.10)*, the solution* $\left[\begin{smallmatrix} \alpha & \beta \\ \gamma & \delta \end{smallmatrix}\right]$ *of equation* (7.6) *can also be written as:*

$$\begin{bmatrix} \alpha & \beta \\ \gamma & \delta \end{bmatrix} = \begin{bmatrix} (1-|s_0|^2)^{-1/2} c_1 & d_{12} \\ (1-|s_0|^2)^{-1/2} c_2 & d_{22} \end{bmatrix} \quad \text{if} \quad |s_0| < 1, \tag{7.11}$$

and

$$\begin{bmatrix} \alpha & \beta \\ \gamma & \delta \end{bmatrix} = \begin{bmatrix} -(\overline{s_0})^{-1} d_{11} & d_{12} \\ -(\overline{s_0})^{-1} d_{21} & d_{22} \end{bmatrix} \quad \text{if} \quad s_0 \neq 0\,. \tag{7.12}$$

Remark 7.3. *If* $n = 1$*, then there is no room for the matrices* d_{12}, d_{21}, d_{22} *and* β, γ, δ*. In this case* U^0*,* (7.5)*, should be replaced by the matrix:*

$$U^0 = \begin{bmatrix} s_0 & (1-|s_0|^2)^{1/2} \\ c_1 & d_{11} \end{bmatrix}, \tag{7.13}$$

where $s_0 \in \mathbb{C}$ *with* $|s_0| \leq 1$*,*

$$c_1 \in \mathfrak{M}_{1\times 1}, \quad d_{11} \in \mathfrak{M}_{1\times 1}\,,$$

and the matrix U^1, (7.7), *should be replaced with: matrix* U^1

$$U^1 = \begin{bmatrix} \alpha \end{bmatrix} , \tag{7.14}$$

where

$$\alpha \in \mathfrak{M}_{1\times 1} .$$

Equation (7.6) *takes the form*

$$U^0 = \begin{bmatrix} 1 & 0 \\ 0 & \alpha \end{bmatrix} \begin{bmatrix} s_0 & (1-|s_0|^2)^{1/2} \\ (1-|s_0|^2)^{1/2} & -\overline{s_0} \end{bmatrix}. \tag{7.15}$$

The solution of this equation can be expressed as

$$\begin{bmatrix} \alpha \end{bmatrix} = \begin{bmatrix} -d_{11}\, s_0 + c_1\, (1-|s_0|^2)^{1/2} \end{bmatrix} , \tag{7.16}$$

as well as in the forms:

$$\begin{bmatrix} \alpha \end{bmatrix} = \begin{bmatrix} (1-|s_0|^2)^{-1/2} c_1 \end{bmatrix} \quad \textit{if} \ \ |s_0| < 1, \tag{7.17}$$

and

$$\begin{bmatrix} \alpha \end{bmatrix} = \begin{bmatrix} -(\overline{s_0})^{-1} d_{11} \end{bmatrix} \quad \textit{if} \ \ s_0 \neq 0 . \tag{7.18}$$

Since both factors on the right-hand side of (7.6) are unitary matrices, we have that U^1 is also a unitary matrix. The matrix U^0 in (7.5) can be considered as a matrix of the unitary colligation $(\mathcal{E}^0, \mathcal{H}^0, U^0)$ with outer space $\mathcal{E}^0 = \mathbb{C}$ and with inner space $\mathcal{H}^0 = \mathbb{C}^n$. The matrix U^1 in (7.7) can, in turn, be considered as a matrix of the the unitary colligation $(\mathcal{E}^1, \mathcal{H}^1, U^1)$ with outer space $\mathcal{E}^1 = \mathbb{C}$ and with inner space $\mathcal{H}^1 = \mathbb{C}^{n-1}$.

Lemma 7.3.

I. *If the colligation* $(\mathcal{E}^0, \mathcal{H}^0, U^0)$ *is controllable, then the colligation* $(\mathcal{E}^1, \mathcal{H}^1, U^1)$ *is also controllable.*

II. *If the colligation* $(\mathcal{E}^0, \mathcal{H}^0, U^0)$ *is observable, then the colligation* $(\mathcal{E}^1, \mathcal{H}^1, U^1)$ *is also observable.*

Proof. Without loss of generality, we assume that $|s_0| < 1$. Otherwise the colligation $(\mathcal{E}^0, \mathcal{H}^0, U^0)$ can not be neither controllable nor observable. Our reasoning is based on the equalities

$$\begin{bmatrix} d_{11} & d_{12} \\ d_{21} & d_{22} \end{bmatrix} = \begin{bmatrix} (-\overline{s_0})\alpha & \beta \\ (-\overline{s_0})\gamma & \delta \end{bmatrix} . \tag{7.19}$$

and

$$\begin{bmatrix} c_1 \\ c_2 \end{bmatrix} = (1-|s_0|^2)^{1/2} \begin{bmatrix} \alpha \\ \gamma \end{bmatrix} . \tag{7.20}$$

Proof of Statement I. The condition that the colligation $(\mathcal{E}^0, \mathcal{H}^0, U^0)$ be controllable means that

$$\bigvee_{0\leq k} \begin{bmatrix} d_{11} & d_{12} \\ d_{21} & d_{22} \end{bmatrix}^k \begin{bmatrix} c_1 \\ c_2 \end{bmatrix} = \mathfrak{M}_{n\times 1} \quad \left(\ = \begin{bmatrix} \mathfrak{M}_{1\times 1} \\ \mathfrak{M}_{(n-1)\times 1} \end{bmatrix} \ \right) . \tag{7.21}$$

And the controllability of the colligation $(\mathcal{E}^1, \mathcal{H}^1, U^1)$ can be expressed as:

$$\bigvee_{0\le k} \delta^k \gamma = \mathfrak{M}_{(n-1)\times 1}\,. \tag{7.22}$$

We look to show that (7.22) follows from (7.21). In view of (7.19) and (7.20), we can express (7.21) as:

$$\bigvee_{0\le k} \begin{bmatrix} (-\overline{s_0})\alpha & \beta \\ (-\overline{s_0})\beta & \delta \end{bmatrix}^k \begin{bmatrix} \alpha \\ \gamma \end{bmatrix} = \mathfrak{M}_{n\times 1} \quad \left(= \begin{bmatrix} \mathfrak{M}_{1\times 1} \\ \mathfrak{M}_{(n-1)\times 1} \end{bmatrix} \right). \tag{7.23}$$

Let

$$\begin{bmatrix} (-\overline{s_0})\alpha & \beta \\ (-\overline{s_0})\gamma & \delta \end{bmatrix}^k \begin{bmatrix} \alpha \\ \gamma \end{bmatrix} = \begin{bmatrix} * \\ f_k \end{bmatrix}, \quad k = 0,\, 1,\, 2\ \ldots\,, \tag{7.24}$$

where $f_k \in \mathfrak{M}_{(n-1)\times 1}$. In view of (7.23),

$$\bigvee_{0\le k} [f_k] = \mathfrak{M}_{(n-1)\times 1}\,. \tag{7.25}$$

Clearly, we have that for every $k = 0,\, 1,\, 2,\, \ldots$

$$f_k = \xi_{0,k}\delta^0\gamma + \cdots + \xi_{k-1,k}\delta^{k-1}\gamma + \delta^k\gamma\,, \tag{7.26}$$

where $\xi_{j,k},\ 0 \le j \le k-1$, are some complex numbers. Therefore,

$$\bigvee_{0\le k} [f_k] = \bigvee_{0\le k} \delta^k\gamma\,.$$

We have thus proved Statement I.

Proof of Statement II. The condition that the colligation be observable $(\mathcal{E}^0, \mathcal{H}^0, U^0)$ can be written as:

$$\bigvee_{0\le k} \begin{bmatrix} (1-|s_0|^2)^{1/2} & 0_{1\times(n-1)} \end{bmatrix} \begin{bmatrix} d_{11} & d_{12} \\ d_{21} & d_{22} \end{bmatrix}^k = \mathfrak{M}_{1\times n} \quad \left(= \begin{bmatrix} \mathfrak{M}_{1\times 1} & \mathfrak{M}_{1\times(n-1)} \end{bmatrix} \right). \tag{7.27}$$

And the observability of the colligation $(\mathcal{E}^1, \mathcal{H}^1, U^1)$ can be expressed as:

$$\bigvee_{0\le k} \beta\delta^k = \mathfrak{M}_{1\times(n-1)}\,. \tag{7.28}$$

We aim to show that (7.28) follows from the formulas (7.27), (7.8) and (7.10). In view of (7.19), we can express (7.27) as follows:

$$\bigvee_{0\le k} \begin{bmatrix} 1 & 0_{1\times(n-1)} \end{bmatrix} \begin{bmatrix} (-\overline{s_0})\alpha & \beta \\ (-\overline{s_0})\beta & \delta \end{bmatrix}^k = \mathfrak{M}_{1\times n} \quad \left(= \begin{bmatrix} \mathfrak{M}_{1\times 1} & \mathfrak{M}_{1\times(n-1)} \end{bmatrix} \right). \tag{7.29}$$

Let

$$\begin{bmatrix} 1 & 0_{1\times(n-1)} \end{bmatrix} \begin{bmatrix} (-\overline{s_0})\alpha & \beta \\ (-\overline{s_0})\gamma & \delta \end{bmatrix}^k = \begin{bmatrix} * & g_k \end{bmatrix}, \quad k = 0,\, 1,\, 2\ \ldots\,, \tag{7.30}$$

where $g_k \in \mathfrak{M}_{1\times(n-1)}$. In view of (7.29), we have that

$$\bigvee_{0\le k} [g_k] = \mathfrak{M}_{1\times(n-1)}\,. \tag{7.31}$$

Clearly, $g_0 = 0_{1\times(n-1)}$ and for every $k = 0, 1, 2, \ldots$

$$g_{k+1} = \eta_{0,k}\,\beta\delta^0 + \cdots + \eta_{k-1,k}\,\beta\delta^{k-1} + \beta\delta^k\,, \tag{7.32}$$

where $\eta_{j,k},\ 0 \le j \le k-2$, are some complex numbers. Therefore,

$$\bigvee_{0\le k} [g_k] = \bigvee_{0\le k} \beta\delta^k\,.$$

We have thus proved Statement II. □

The following lemma is an immediate consequence of Lemma 7.3.

Lemma 7.4. *Let $s_0 \in \mathbb{C}$ with $|s_0| < 1$ and U^0 be a unitary $(n+1)\times(n+1)$-matrix of the form (7.5). Suppose that the $n\times n$-matrix U^1, (7.7), is related to the matrix U^0 by equation (7.6). Let $(\mathcal{E}^0, \mathcal{H}^0, U^0)$ and $(\mathcal{E}^1, \mathcal{H}^1, U^1)$ be the above-described operator colligations related to the matrices U^0 and U^1. If the colligation $(\mathcal{E}^0, \mathcal{H}^0, U^0)$ is minimal, then the colligation $(\mathcal{E}^1, \mathcal{H}^1, U^1)$ is also minimal.*

Theorem 7.2. *Let $s(z)$ be a rational inner matrix-function of degree $n > 1$ ($s(z)$ is thus non-constant and $|s_0| < 1$, where $s_0 = s(0)$) and let*

$$\omega(z) = \frac{1}{z}\cdot\frac{s(z) - s_0}{1 - s(z)\overline{s_0}}\,, \qquad s_0 = s(0), \tag{7.33}$$

be the Schur transformation of the function $s(z)$.
Let the unitary matrix U,

$$U = \begin{bmatrix} s_0 & B_0 \\ C_0 & D_0 \end{bmatrix}\,, \quad B_0 \in \mathfrak{M}_{1\times n},\ C_0 \in \mathfrak{M}_{n\times 1},\ D_0 \in \mathfrak{M}_{n\times n}, \tag{7.34}$$

which yields the minimal system representation

$$s(z) = s_0 + zB_0(I - zD_0)^{-1}C_0\,, \tag{7.35}$$

have row B_0 of the special form

$$B_0 = [b \quad 0_{1\times(n-1)}]\,, \qquad b > 0\,. \tag{7.36}$$

Then the function $\omega(z)$ admits the system representation

$$\omega(z) = \alpha + z\beta(I - z\delta)^{-1}\gamma\,, \tag{7.37}$$

where the unitary $n\times n$-matrix $\begin{bmatrix} \alpha & \beta \\ \gamma & \delta \end{bmatrix}$,

$$\alpha \in \mathfrak{M}_{1\times 1},\quad \beta \in \mathfrak{M}_{1\times(n-1)},\quad \gamma \in \mathfrak{M}_{(n-1)\times 1},\quad \delta \in \mathfrak{M}_{(n-1)\times(n-1)}\,,$$

can be determined from the matrix U^0 using

$$\begin{bmatrix} \alpha & \beta \\ \gamma & \delta \end{bmatrix} = \begin{bmatrix} (1-|s_0|^2)^{-1/2}c_1 & d_{12} \\ (1-|s_0|^2)^{-1/2}c_2 & d_{22} \end{bmatrix}\,, \tag{7.38}$$

where c_j and d_{jk} are the block-matrix entries of the block-matrix decompositions

$$C_0 = \begin{bmatrix} c_1 \\ c_2 \end{bmatrix} \qquad D_0 = \begin{bmatrix} d_{11} & d_{12} \\ d_{21} & d_{22} \end{bmatrix}, \tag{7.39}$$

$$c_1 \in \mathfrak{M}_{1\times 1},\ c_2 \in \mathfrak{M}_{(n-1)\times 1},$$

$$D_{11} \in \mathfrak{M}_{1\times 1},\ D_{12} \in \mathfrak{M}_{1\times(n-1)},\ D_{21} \in \mathfrak{M}_{(n-1)\times 1},\ D_{22} \in \mathfrak{M}_{(n-1)\times(n-1)}\,.$$

The unitary colligation associated with the matrix $\tilde{U}^1$ is minimal.

Proof. The matrix U^0, (7.34), the matrix $\left[\begin{smallmatrix} \alpha & \beta \\ \gamma & \delta \end{smallmatrix}\right]$, (7.38), and the number s_0 are related by equation (6.11). According to Theorem 6.1, the function $\omega(z)$, defined by (7.37), and the function $s(z)$ are related by the equality (6.6). □

Theorem 7.2 together with Lemma 7.2 describe a step of the Schur algorithm in terms of system representations. Before applying the direct Schur transform (7.33), which is a step of the Schur algorithm, we should first 'normalize' the colligation matrix U representing the 'initial' function $s(z)$. This normalization starts with the matrix U, from which we determine the unitarily equivalent matrix U^0, (7.5), whose row B^0 is of the special form (7.3). We then aim to solve the equation (7.6) with respect to the matrix $\left[\begin{smallmatrix} \alpha & \beta \\ \gamma & \delta \end{smallmatrix}\right]$. The solution of (7.6) is given by (7.38). The unitary matrix $\left[\begin{smallmatrix} \alpha & \beta \\ \gamma & \delta \end{smallmatrix}\right]$ yields the system representation of the function $\omega(z)$.

It should be emphasized that, in general, *the matrix $\left[\begin{smallmatrix} \alpha & \beta \\ \gamma & \delta \end{smallmatrix}\right]$ is not normalized, i.e., its row β is not of the form $\beta = [*\ \ 0_{1\times(n-2)}]$.* To perform the next step of the Schur algorithm, we must therefore 'normalize' the matrix $\left[\begin{smallmatrix} \alpha & \beta \\ \gamma & \delta \end{smallmatrix}\right]$, obtaining the 'normalized' form U^1. We then have to solve the equation of the form (7.6), where U^0 is replaced by U^1, etc. The normalization procedure must therefore be performed *at every step of the Schur algorithm.* This normalization procedure is, however, not quite unique. It has some degrees of freedom (see Remark 7.1). It turns out that we can use these degrees of freedom to make the normalization procedure a *one-time* procedure, so that it might be dealt with during *preprocessing* for the further step-by-step recurrence. In further processing there is then no need for normalization and one only has to solve the recurrent chain of equations of the form (7.6). A *one-time normalization* of this kind is related to the reduction of the 'initial' colligation matrix to the *lower Hessenberg form.*

8. Hessenberg matrices. The Householder algorithm

Roughly speaking, the lower (upper) Hessenberg matrix, is a matrix which is almost lower (upper) triangular. The precise definition is:

Definition 8.1.

I. *We say that a square matrix H is a* lower Hessenberg matrix *if it has zero-entries above the first superdiagonal. If $H = ||h_{jk}||_{0\le j,k\le n}$, then H is lower Hessenberg matrix if $h_{jk} = 0$ for $k > j+1$, $0 \le j \le n-1$.*

II. *We say that a lower Hessenberg matrix* $H = ||h_{jk}||_{0\le j,k\le n}$ *is* special *if all entries of its first superdiagonal are non-negative:* $h_{j,j+1} \ge 0\,, 0 \le j \le n-1$.

III. *We say that a Hessenberg matrix* $H = ||h_{jk}||_{0\le j,k\le n}$ *is* HL-non-singular *if all entries of its first superdiagonal are non-zero:* $h_{j,j+1} \ne 0\,, 0 \le j \le n-1$.

The definition of an upper Hessenberg matrix, special upper Hessenberg matrix and non-singular upper Hessenberg matrix is similar to Definition 8.1:

Definition 8.2.

I. *We say that a square matrix* H *is an* upper Hessenberg matrix *if it has zero-entries below the first subdiagonal. If* $H = ||h_{jk}||_{0\le j,k\le n}$, *then* H *is upper Hessenberg matrix if* $h_{jk} = 0$ *for* $k < j-1$, $0 \le j \le n-1$.

II. *We say that an upper Hessenberg matrix* $H = ||h_{jk}||_{0\le j,k\le n}$ *is* special *if all entries of its first subdiagonal are non-negative:* $h_{j,j-1} \ge 0\,, 1 \le j \le n$.

III. *We say that an upper Hessenberg matrix* $H = ||h_{jk}||_{0\le j,k\le n-1}$ *is* HU-non-singular *if all entries of its first subdiagonal are non-zero:* $h_{j,j-1} \ne 0\,,\ 1 \le j \le n$.

Hessenberg matrices were investigated by Karl Hessenberg (1904–1959), a German engineer whose dissertation dealt with the computation of eigenvalues and eigenvectors of linear operators.

Theorem 8.1.

I. *Given an* $(n+1)\times(n+1)$*-matrix* $M = ||M_{j,k}||_{0\le j,k\le n}$, *there exists a unitary* $n\times n$*-matrix* V *such that the matrix* H^L,

$$H^L = \begin{bmatrix} 1 & 0_{1\times n} \\ 0_{n\times 1} & V^* \end{bmatrix} M \begin{bmatrix} 1 & 0_{1\times n} \\ 0_{n\times 1} & V \end{bmatrix} \tag{8.1}$$

is a special lower Hessenberg matrix.

II. *If the matrix* M *is HL-non-singular, then both matrices* H^L *and* V *are uniquely determined. From the equalities*

$$H_j^L = \begin{bmatrix} 1 & 0_{1\times n} \\ 0_{n\times 1} & V_j^* \end{bmatrix} M \begin{bmatrix} 1 & 0_{1\times n} \\ 0_{n\times 1} & V_j \end{bmatrix}, \quad j = 1,\, 2\,, \tag{8.2}$$

where H_1^L *and* H_2^L *are special upper Hessenberg matrices,* V_1 *and* V_2 *are unitary matrices and the Hessenberg matrix* H_1^L *is HL-non-singular, it follows that* $H_2^L = H_1^L$ *and* $V_2 = V_1$.

Definition 8.3. *Given a square matrix* M, *a lower Hessenberg matrix* H^L *to which* M *can be reduced,* (8.1), *is called a lower Hessenberg form of the matrix* M.

Theorem 8.2.

I. *Given an* $(n+1)\times(n+1)$*-matrix* $M = ||M_{j,k}||_{0\le j,k\le n}$, *there exists a unitary* $n\times n$*-matrix* V *such that the matrix* H^U,

$$H^U = \begin{bmatrix} 1 & 0_{1\times n} \\ 0_{n\times 1} & V^* \end{bmatrix} M \begin{bmatrix} 1 & 0_{1\times n} \\ 0_{n\times 1} & V \end{bmatrix} \tag{8.3}$$

is a special upper Hessenberg matrix.

II. *If the matrix M is HU-non-singular, then both matrices H^U and V are uniquely determined. From the equalities*

$$H_j^U = \begin{bmatrix} 1 & 0_{1\times n} \\ 0_{n\times 1} & V^* \end{bmatrix} M \begin{bmatrix} 1 & 0_{1\times n} \\ 0_{n\times 1} & V_j \end{bmatrix}, \quad j = 1, 2, \tag{8.4}$$

where H_1^U and H_2^U are upper Hessenberg matrices, V_1 and V_2 are unitary matrices and the Hessenberg matrix H_1^U is HU-non-singular, it follows that $H_2^U = H_1^U$ and $V_2 = V_1$.

Definition 8.4. *Given a square matrix M, an upper Hessenberg matrix H^U to which M can be reduced,* (8.3)*, is called an upper Hessenberg form of the matrix M.*

Theorem 8.3. *Let U be an $(n+1)\times(n+1)$-unitary matrix.*

I. *The unitary colligation associated with the matrix U is observable if and only if lower Hessenberg form of U is HL-non-singular.*

II. *The unitary colligation associated with the matrix U is controllable if and only if the upper Hessenberg form of U is HU-non-singular.*

Corollary 8.1. According to Theorem 3.9, the finite-dimensional unitary colligation is observable if and only if it is controllable. Thus, *for a unitary matrix U, the lower Hessenberg form of U is HL-nonsingular if and only if the upper Hessenberg form of U is HU-nonsingular.*

Lemma 8.1. *Given two row-vectors*

$$B' = [b_1'\ b_2' \dots b_n'] \in \mathfrak{M}_{1\times n} \quad \text{and} \quad B'' = [b_1''\ b_2'' \dots b_n''] \in \mathfrak{M}_{1\times n}$$

having same norm:

$$B'B'^* = B''B''^*, \tag{8.5}$$

there exists a unitary $n\times n$-matrix V such that

$$B'V = B''. \tag{8.6}$$

Proof of Lemma 8.1. We first consider the question in a more general setting. Assume that $\mathfrak{H}$ is a *complex* Hilbert space with scalar product $\langle u, v\rangle$, where $\langle u, v\rangle$ is linear with respect to the argument u and antilinear with respect to v. Let $x \in \mathfrak{H}$ and $y \in \mathfrak{H}$ be two vectors such that $\langle x, x\rangle = \langle y, y\rangle \neq 0$. Let $||u||$ denote the norm of the vector u: $||u|| = \langle u, u\rangle^{1/2}$. Given two vectors $x \in \mathfrak{H}$, $y \in \mathfrak{H}$ such that $||x|| = ||y|| \neq 0$, our goal is to construct a unitary operator $V : \mathfrak{H} \to \mathfrak{H}$ such that $Vx = y$. If the vector y is proportional to the vector x: $x = \lambda y$ for some $\lambda \in \mathbb{C}$, we put $Vz = \lambda z\ \ \forall z \in \mathbb{C}$. This operator is unitary: $|\lambda| = 1$, because $||x|| = ||y|| \neq 0$. If the vectors x and y are not proportional, we choose $\lambda \in \mathbb{C}$, $|\lambda| = 1$ such that $\lambda\langle x, y\rangle \geq 0$. (If $\langle x, y\rangle \neq 0$, then this λ is unique. If $\langle x, y\rangle = 0$, we can choose arbitrary λ with $|\lambda| = 1$.) Let

$$Vz = \lambda z - 2\frac{\langle z, x - \overline{\lambda}y\rangle}{||x - \overline{\lambda}y||^2}(\lambda x - y) \qquad \forall\ z \in \mathfrak{H}. \tag{8.7}$$

The vectors

$$e_1 = x + \overline{\lambda}y \quad \text{and} \quad e_2 = x - \overline{\lambda}y$$

are non-zero (x and y are not proportional to one another) and orthogonal:

$$\langle e_1,\, e_2\rangle = 0\,, \tag{8.8}$$

because

$$\langle x+\overline{\lambda}y\,,\, x-\overline{\lambda}y\rangle = \langle x,\, x\rangle - \overline{\lambda}\lambda\langle y,\, y\rangle + \overline{\lambda}\langle y,\, x\rangle - \lambda\langle x,\, y\rangle$$

and $\langle x,\, x\rangle = \langle y,\, y\rangle$, $\overline{\lambda}\lambda = 1$, $\lambda\langle x,\, y\rangle = \overline{\lambda\langle x,\, y\rangle} = \overline{\lambda}\langle y,\, x\rangle$. From (8.7) and (8.8) it follows that

$$Ve_1 = \lambda e_1\,. \tag{8.9}$$

From (8.7) it follows that

$$Ve_2 = -\lambda e_2\,, \tag{8.10}$$

and

$$Vz = \lambda z \tag{8.11}$$

$\forall\, z \in \mathfrak{H} : \langle z,\, e_1\rangle = 0$, $\langle z,\, e_2\rangle = 0$. Therefore the operator V is unitary. Since

$$x = \frac{1}{2}(e_1 + e_2), \quad y = \frac{\lambda}{2}(e_1 - e_2)$$

from (8.9) and (8.10) it follows that $Vx = y$.

Let us turn to the proof of the statement of Lemma 8.1. Let $\mathfrak{H}$ be the set of all n-row-vectors with complex entries (in other words, $\mathfrak{H} = \mathfrak{M}_{1\times n}$) and with the following scalar product: if $u = [u_1,\, \dots\,,\, u_n]$ and $v = [v_1,\, \dots\,,\, v_n]$ are vectors in $\mathfrak{H}$, then their scalar product $\langle u\,,\, v\rangle$ is defined as

$$\langle u\,,\, v\rangle = u\, v^*$$

where v^* is the Hermitian conjugate of the row-vector v. If H is some $n\times n$-matrix, then it generates an operator in $\mathfrak{H}$. This operator maps the row-vector u to the row-vector uH, where uH is the product of the *matrices* u and H. This operator is unitary if and only if H is unitary.

In the notation of Lemma 8.1: $x = B' = [b'_1\ b'_2 \dots b'_n]$, $y = B'' = [b''_1\ b''_2 \dots b''_n]$. Thus the matrix V corresponding to the operator (8.7) takes the form

$$V = ||v_{jk}||_{1\le j,k\le n}, \tag{8.12}$$

where

$$v_{jk} = \lambda\delta_{jk} - 2(\overline{b'_j} - \lambda\overline{b''_j})\langle\, B' - \overline{\lambda}B''\,,\, B' - \overline{\lambda}B''\rangle^{-1}(\lambda b'_k - b''_k)\,. \tag{8.13}$$

and λ is such that

$$\lambda B'(B'')^* \ge 0, \quad |\lambda| = 1\,.$$

δ_{jk} is the Kronecker symbol. □

Remark 8.1. *In the case when the rows B' and B'' are real, the matrix V,* (8.12)–(8.13)*, is also real. In this case matrices of the form* (8.12)–(8.13) *are known as* Householder reflection matrices*. Householder reflection matrices and the Householder Algorithm* (*which is based on matrices of this type*) *are widely used in numerical linear algebra. See* [Wil], [Str], [GolV] and [Hou].

Remark 8.2. *A unitary matrix V satisfying the condition* (8.6) *is not unique. The process of constructing such matrices* (8.12)–(8.13) *is constructive.*

We will apply Lemma 8.1 *to the following special situation: Let $B' \neq 0$ be an arbitrary $1 \times n$-column and B'' be of the special form $B'' = [b'' \; 0_{1\times(n-1)}]$, where $b'' > 0$ and thus $b'' = (B'(B')^*)^{1/2}$. For these B', B'', the first column of the unitary matrix V satisfying* (8.6) *is uniquely determined:*

$$v_{j1} = \overline{b'_j}(B'(B')^*)^{-1/2},\ 1 \leq j \leq n.$$

The construction of the desired matrix V is thus reduced to the following problem: Given the first column of an $n \times n$-matrix, one needs to extend this column to a full unitary *matrix. The Householder reflection procedure is one way of doing this.*

We use the Householder reflection matrices to reduce an arbitrary matrix to a Hessenberg matrix.

Proof of Theorem 8.1. Let $M = M^0$ and let $m^0_{j,k}$ be entries of the matrix M^0:

$$M^0 = ||m^0_{j,k}||_{0\leq j,k\leq n} \tag{8.14}$$

Applying Lemma 8.1, we choose the unitary matrix $V_1 \in \mathfrak{M}_{n,n}$ such that

$$[m^0_{0,1},\, m^0_{0,2},\, \ldots,\, m^0_{0,n}]V_1 = [m^1_{0,1},\, m^1_{0,2},\, \ldots,\, m^1_{0,n}]\,, \tag{8.15}$$

where

$$m^1_{0,1} \geq 0,\ m^1_{0,k} = 0,\ 2 \leq k \leq n. \tag{8.16}$$

(So that $m_{0,1} = \left([m^0_{0,1},\, m^0_{0,2},\, \ldots,\, m^0_{0,n}] \cdot [m^0_{0,1},\, m^0_{0,2},\, \ldots,\, m^0_{0,n}]^*\right)^{1/2}$.)
V_1 can be considered as an appropriate Householder rotation, for instance. Let us consider the matrix

$$M^1 = \begin{bmatrix} 1 & 0_{1\times n} \\ 0_{n\times 1} & V_1^* \end{bmatrix} M^0 \begin{bmatrix} 1 & 0_{1\times n} \\ 0_{n\times 1} & V_1 \end{bmatrix}, \tag{8.17}$$

and let $m^1_{j,k}$ denote the entries of the matrix M^1:

$$M^1 = ||m^1_{j,k}||_{0\leq j,k\leq n} \tag{8.18}$$

Clearly,

$$m^1_{0,0} = m^0_{0,0}\,. \tag{8.19}$$

We continue this procedure inductively. We next turn to the inductive step from l to $l+1$.

Suppose that the matrices $M^p \in \mathfrak{M}_{n+1,n+1}$ and $V_p \in \mathfrak{M}_{n-p+1,n-p+1}$ with $0 \leq p \leq l$ are already known and that the following condition for the entries of the matrix M^p,

$$M^p = ||m^p_{j,k}||_{0\leq j,k\leq n}\,, \tag{8.20}$$

are satisfied:

$$m^p_{j,j+1} \geq 0, \quad m^p_{j,k} = 0, \quad j+2 \leq k \leq n\,, \quad j = 0,\, 1,\, \ldots,\, p-1\,, \tag{8.21}$$

The matrices V_p, $1 \le p \le l$ are unitary and we have

$$M^p = \begin{bmatrix} I_p & 0_{p\times(n-p+1)} \\ 0_{(n-p+1)\times p} & V_p^* \end{bmatrix} M^{p-1} \begin{bmatrix} I_p & 0_{p\times(n-1p+1)} \\ 0_{(n-p+1)\times p} & V_p \end{bmatrix} \tag{8.22}$$

for every $p : p \le l$.

We choose the unitary $(n-l)\times(n-l)$-matrix V_{l+1} such that

$$[m_{l,l+1}^l,\, m_{l,l+2}^l,\, \dots,\, m_{l,n}^l]V_{l+1} = [m_{l,l+1}^{l+1},\, m_{l,l+2}^{l+1},\, \dots,\, m_{l,n}^{l+1}]\,, \tag{8.23}$$

where

$$m_{l,l+1}^{l+1} \ge 0, \quad m_{l,k}^{l+1} = 0, \quad l+2 \le k \le n. \tag{8.24}$$

Lemma 8.1 ensures that this choice is possible. We then define the matrix M^{l+1},

$$M^{l+1} = ||m_{j,k}^{l+1}||_{0\le j,k\le n} \tag{8.25}$$

as

$$M^{l+1} = \begin{bmatrix} I_{l+1} & 0_{l+1\times(n-l)} \\ 0_{(n-l)\times(l+1)} & V_{l+1}^* \end{bmatrix} M^l \begin{bmatrix} I_{l+1} & 0_{(l+1)\times(n-l)} \\ 0_{(n-l)\times(l+1)} & V_{l+1} \end{bmatrix}. \tag{8.26}$$

The entries of the matrix M^{l+1} satisfy the condition

$$m_{j,j+1}^{l+1} \ge 0, \quad m_{j,k}^{l+1} = 0, \quad j = 0,\, 1,\, \dots,\, l\,, \quad j+2 \le k \le n\,. \tag{8.27}$$

For $j = l$, condition (8.24) holds in view of (8.23) (ensuring this was our goal in choosing the matrix V_{l+1} as we did).
For $0 \le j \le l-1$, condition (8.24) holds, because going from the matrix M^l to the matrix M^{l+1} we do not change the rows with indices $j:\ 0 \le j \le l-1$:

$$m_{j,k}^{l+1} = m_{j,k}^l, \quad 0 \le j \le l-1,\ 0 \le k \le n\,. \tag{8.28}$$

The equality (8.28) holds, firstly because the identity matrix of size $l+1$ is the left upper corner of the block-matrix $\left[\begin{smallmatrix} I_{l+1} & 0_{l+1\times(n-l)} \\ 0_{(n-l)\times(l+1)} & V_{l+1} \end{smallmatrix}\right]$ and secondly, because

$$m_{j,k}^l = 0 \quad \forall\, j,k :\ 0 \le j \le l-1,\, l+1 \le k \le n$$

(The latter is a consequence of the induction hypothesis (8.21) for $p = l-1$.)
The inductive process finishes when we construct the matrix $M_n = M^{l+1}$ for $l = n-1$.

The matrix V satisfying (8.1) appears as the product

$$V = V_1 \cdot \begin{bmatrix} I_1 & 0_{1\times(n-1)} \\ 0_{(n-2)\times 2} & V_2 \end{bmatrix} \cdot \begin{bmatrix} I_2 & 0_{2\times(n-2)} \\ 0_{(n-2)\times 2} & V_3 \end{bmatrix} \cdot \ \dots\ \cdot \begin{bmatrix} I_{n-2} & 0_{(n-2)\times 2} \\ 0_{2\times(n-2)} & V_{n-1} \end{bmatrix}. \tag{8.29}$$

According to the above construction, the entries of the matrix $H = ||h_{j,k}||_{0\le j,k\le n}$, (8.1), satisfy:

$$h_{j,k} = m_{j,k}^{j+1}, \quad j \le k \le n\,, \tag{8.30}$$

and thus we have:

$$h_{j,j+1} = m_{j,j+1}^{j+1} \geq 0, \quad h_{j,k} = 0, \ j+2 \leq k \leq n\,, \tag{8.31}$$

□

The reduction of matrices to the Hessenberg form is a tool often applied in numerical linear algebra as a preliminary step for further numerical algorithms. See [Wil], [Str], [GolV] and other sources in numerical linear algebra.

The Householder algorithm is implemented in the programming system `MATLAB`. The `MATLAB` command `H=hess(A)` reduces the matrix `A` to the upper Hessenberg form `H`.

In the next section we discuss the Schur algorithm for rational inner functions in terms of the unitary colligation for the system representation of this function. Reducing the colligation matrix to the upper Hessenberg form is a preliminary step for further developing the Schur algorithm in terms of system representations.

Remark 8.3. *In* [KiNe]*, the Householder algorithm and the Hessenberg form for unitary matrices are used to study the probability measures associated with finite Blaschke products via Cayley transform.*

9. The Schur algorithm in terms of system representations

We have now finished all necessary preparations and we are well positioned to present the Schur algorithm in terms of unitary colligations representing the appropriate functions.

Let $s(z)$ be a rational inner matrix-function of degree $n > 0$ ($s(z)$ is thus non-constant and $|s_0| < 1$, where $s_0 = s(0)$) and let

$$s_0(z) = s(z), \quad s_k(z), \ k = 1, 2, \ldots, n\,,$$

be the sequence of rational inner functions constructed according to (2.4) ($\deg s_k(z) = n - k$, so that $s_n(z) = s_n$ is a unitary constant).

Let

$$s(z) = A + zB(I_n - zD)^{-1}C \tag{9.1}$$

be the system representation of $s(z)$, where

$$U = \begin{bmatrix} A & B \\ C & D \end{bmatrix} \tag{9.2}$$

is the matrix of the minimal unitary colligation representing $s(z)$:

$$A \in \mathfrak{M}_{1\times 1}, \ B \in \mathfrak{M}_{1\times n}, \ C \in \mathfrak{M}_{n\times 1}, \ D \in \mathfrak{M}_{n\times n}\,. \quad (\text{So}, A = s_0.)$$

We first reduce U to the lower Hessenberg form. Let V be a unitary $n \times n$-matrix such that the matrix U^0 (also unitary):

$$U^0 = \begin{bmatrix} 1 & 0_{1\times n} \\ 0_{n\times 1} & V^* \end{bmatrix} U \begin{bmatrix} 1 & 0_{1\times n} \\ 0_{n\times 1} & V \end{bmatrix}, \quad j = 1, 2\,, \tag{9.3}$$

is an upper Hessenberg matrix. The block entries of the matrix

$$U^0 = \begin{bmatrix} A^0 & B^0 \\ C^0 & D^0 \end{bmatrix} \tag{9.4}$$

are:

$$A^0 \in \mathfrak{M}_{1\times 1},\ B^0 \in \mathfrak{M}_{1\times n},\ C^0 \in \mathfrak{M}_{n\times 1},\ D^0 \in \mathfrak{M}_{n\times n}\,. \quad (\text{So}, A^0 = A = s_0.)$$

The unitary colligations associated with the matrices U and U^0 are unitarily equivalent. The unitary colligation associated with the unitary matrix U^0 is therefore minimal and represents the function $s_0(z) = s(z)$:

$$s_0(z) = A^0 + zB^0(I_n - zD^0)^{-1}C^0\,. \tag{9.5}$$

Inductively, we construct the sequence $U^p,\ p = 0, 1, \ldots, n-1$ of unitary upper Hessenberg matrices such that the unitary colligation associated with the matrix U^p is minimal and represents the function $s_p(z)$, which appears in the pth step of the Schur algorithm.

For $p = 0$, the representation in (9.5) holds. We consider the step from p to $p+1$.

Suppose that $U^p,\ 0 \le p < (n-1)$ is a unitary lower HL-non-singular $(n-p+1)\times(n-p+1)$ Hessenberg matrix with the block-matrix decomposition:

$$U^p = \begin{bmatrix} A^p & B^p \\ C^p & D^p \end{bmatrix}\,, \tag{9.6}$$

where

$$A^p \in \mathfrak{M}_{1\times 1},\ B^p \in \mathfrak{M}_{1\times(n-p)},\ C^p \in \mathfrak{M}_{(n-p)\times 1},\ D^p \in \mathfrak{M}_{(n-p)\times(n-p)}\,.$$

The unitary colligation associated with the matrix U^p is minimal and represents the function $s_p(z)$, which appears in the pth step of the Schur algorithm:

$$s_p(z) = A^p + zB^p(I_{n-p} - zD^p)^{-1}C^p\,. \tag{9.7}$$

Let

$$C^p = \begin{bmatrix} C_1^p \\ C_2^p \end{bmatrix} \qquad D^p = \begin{bmatrix} D_{11}^p & D_{12}^p \\ D_{21}^p & D_{22}^p \end{bmatrix}\,, \tag{9.8}$$

be the more refined block matrix decomposition of the block-matrix entries C^p and D^p:

$$C_1^p \in \mathfrak{M}_{1\times 1},\ C_2^p \in \mathfrak{M}_{(n-1-p)\times 1},$$

$$D_{11} \in \mathfrak{M}_{1\times 1},\ D_{12} \in \mathfrak{M}_{1\times(n-1-p)},\ D_{21} \in \mathfrak{M}_{(n-1-p)\times 1},\ D_{22} \in \mathfrak{M}_{(n-1-p)\times(n-1-p)}\,.$$

Since U^p is an upper Hessenberg matrix and also an HU-non-singular matrix, we have that $B^p \neq 0$. Because U^p is also unitary, it follows that $|A^p| < 1$, i.e., that

$$|s_p| < 1 \ \text{ where } \ s_p = s_p(0)\,. \tag{9.9}$$

The row B^p is of the form

$$B^p = [(1 - |s_p|^2)^{1/2},\, 0_{1\times(n-p-1)}] \tag{9.10}$$

We construct the $(n-p)\times(n-p)$-matrix U^{p+1}:

$$U^{p+1}=\begin{bmatrix} A^{p+1} & B^{p+1} \\ C^{p+1} & D^{p+1}\end{bmatrix}, \tag{9.11}$$

$$\begin{aligned} &A^{p+1}\in\mathfrak{M}_{1\times 1}, && B^{p+1}\in\mathfrak{M}_{1\times(n-p-1)},\\ &C^{p+1}\in\mathfrak{M}_{(n-p-1)\times 1}, && D^{p+1}\in\mathfrak{M}_{(n-p-1)\times(n-p-1)}, \end{aligned}$$

where

$$\begin{bmatrix} A^{p+1} & B^{p+1} \\ C^{p+1} & D^{p+1}\end{bmatrix}\stackrel{\text{def}}{=}\begin{bmatrix}(1-|s_p|^2)^{-1/2}C_1^p & D_{12}^p\\ (1-|s_p|^2)^{-1/2}C_2^p & D_{22}^p\end{bmatrix}. \tag{9.12}$$

To obtain the matrix U^p from U^{p+1}, one should delete the left column and the upper row of the matrix U^p and then recalculate the first column of the resulting matrix. The matrix U^{p+1} is then an upper Hessenberg matrix. The matrix U^{p+1} is HL-non-degenerate, because U^p is HL-non-degenerate and because the first superdiagonal of the matrix U^{p+1} is a subset of the first superdiagonal of the matrix U^p. According to Theorem 7.2 (which can be applied to the matrix U^p in view of (9.10)), the matrix U^{p+1} is unitary and the unitary colligation associated with U^{p+1} represents the function $s_{p+1}(z)$ appearing in the $p+1$th step of the Schur algorithm:

$$s_{p+1}(z)=A^{p+1}+zB^{p+1}(I_{n-p-1}-zD^{p+1})^{-1}C^{p+1}. \tag{9.13}$$

These considerations do not directly apply when $p=n-1$. In this case, there is no room for B^n, C^n, D^n. However, we can construct 'part' of the matrix (9.12):

$$A^n=(1-|s_{n-1}|^2)^{-1/2}C_1^{n-1}. \tag{9.14}$$

(See Remark 7.3.) The 1×1-matrix A^n is unitary, hence it is a unitary constant. Clearly, $A^n=s_n$, where s_n is the nth Schur parameter. This completes the description of the Schur algorithm for inner rational matrix-functions in terms of system representations. □

Remark 9.1. *It is particularly easy to determine the sequence $\{D^p\}_{p=0,1,\dots,n}$ of matrices representing the inner operators of the unitary colligations associated with the colligation matrices U^p. The matrix D^p makes up the $(n-p)\times(n-p)$ lower-right corner of the matrix D^0. The inner rational matrix-function $s(z)$ is the ratio of two polynomials:*

$$s_p(z)=c_p\frac{z^{n-p}\overline{\chi_p(1/\overline{z})}}{\chi_p(z)},\quad \deg\chi_p(z)=n-p\,,\ \chi_p(0)=1\quad |c_p|=1. \tag{9.15}$$

Clearly,

$$\chi_p(z)=\det(I_{n-p}-zD^p)\,,\quad z^{n-p}\overline{\chi_p(1/\overline{z})}=\det\left(zI_{n-p}-(D^p)^*\right), \tag{9.16a}$$

thus

$$s_p(z)=c_p\det\left(\left(zI_{n-p}-(D^p)^*\right)\left(I_{n-p}-zD^p\right)^{-1}\right). \tag{9.16b}$$

10. An expression for the colligation matrix in terms of the Schur parameters

Let $s(z)$ be a rational inner matrix-function of degree n. Let $s_p(z)$, $p = 0, 1, \dots, n$ be the sequence of rational inner functions produced by the Schur algorithm from the function $s(z)$, as described in (2.4), $\deg s_p(z) = n - p$. Let U^p , (9.6), be the colligation matrix of the minimal unitary colligation, which yields the system representation (9.7) of the function s_p. Among all unitary $(n-p+1)\times(n-p+1)$-matrices representing the function s_p we choose a lower Hessenberg matrix U^p. Such a matrix U^p exists and is unique.

The equality (7.6), where U^p is taken as the matrix U^0 and U^{p+1} is taken as the matrix $\left[\begin{smallmatrix}\alpha & \beta\\ \gamma & \delta\end{smallmatrix}\right]$ takes the form

$$U^p = \begin{bmatrix} 1 & 0_{1\times 1} & 0_{1\times(n-1-p)} \\ 0_{1\times 1} & & \\ 0_{(n-1-p)\times 1} & U^{p+1} & \end{bmatrix} \cdot \begin{bmatrix} s_p & (1-|s_p|^2)^{1/2} & 0_{1\times(n-1-p)} \\ (1-|s_p|^2)^{1/2} & -\overline{s_p} & 0_{1\times(n-1-p)} \\ 0_{(n-1-p)\times 1} & 0_{(n-1-p)\times 1} & I_{(n-1-p)\times(n-1-p)} \end{bmatrix},$$
$$p = 0, 1, \dots, n-2.$$

The latter formula can be rewritten in the equivalent but more convenient form:

$$\begin{bmatrix} I_p & 0_{p\times(n-p+1)} \\ 0_{(n-p+1)\times p} & U^p \end{bmatrix} = \begin{bmatrix} I_{p+1} & 0_{(p+1)\times(n-p)} \\ 0_{(n-p)\times(p+1)} & U^{p+1} \end{bmatrix}$$
$$\cdot \begin{bmatrix} I_p & 0_{p\times 2} & 0_{p\times(n-1-p)} \\ 0_{2\times p} & \begin{matrix} s_p & (1-|s_p|^2)^{1/2} \\ (1-|s_p|^2)^{1/2} & -\overline{s_p} \end{matrix} & 0_{2\times(n-p-1)} \\ 0_{(n-p-1)\times p} & 0_{(n-p-1)\times 2} & I_{n-1-p} \end{bmatrix},$$
$$0 \le p \le n-1\,. \quad (10.1)$$

For $p = 0$, the matrix on the left-hand side of (10.1) takes the form $\begin{bmatrix} U^0 \end{bmatrix}$. For $p = n-1$, the matrix U^{p+1} takes the form $U^n = s_n$ and the second factor on the right-hand side of (10.1) takes the form

$$\begin{bmatrix} I_{n-2} & 0_{(n-2)\times 2} \\ 0_{2\times(n-2)} & \begin{matrix} s_{n-1} & (1-|s_{pn-1}|^2)^{1/2} \\ (1-|s_{n-1}|^2)^{1/2} & -\overline{s_{n-1}} \end{matrix} \end{bmatrix} \cdot$$

From (10.1) it follows that

$$U^0 = \prod_{0\le p\le n-1}^{\curvearrowleft} \begin{bmatrix} I_p & 0_{p\times 2} & 0_{p\times(n-1-p)} \\ 0_{2\times p} & \begin{matrix} s_p & (1-|s_p|^2)^{1/2} \\ (1-|s_p|^2)^{1/2} & -\overline{s_p} \end{matrix} & 0_{2\times(n-p-1)} \\ 0_{(n-p-1)\times p} & 0_{(n-p-1)\times 2} & I_{n-1-p} \end{bmatrix}. \tag{10.2}$$

Multiplying the matrices in (10.2), we obtain an expression for the entries of the matrix U^0, which gives us the system representation of the function $s(z)$ in terms of the Schur parameters of $s(z)$:

$$U^0 = ||u^0_{j,k}||_{0\le j,k\le n}\,, \tag{10.3}$$

where

$$u^0_{j,k} = \begin{cases} s_0, & j=0, \quad k=0, \\ s_j\,\Delta_{j-1}\Delta_{j-2}\cdot\,\cdots\,\cdot\Delta_1\Delta_0\,, & 1\le j\le n\,, \quad k=0\,, \\ -s_j\,\Delta_{j-1}\Delta_{j-2}\cdot\,\cdots\,\cdot\Delta_k\,\overline{s_{k-1}}, & 1\le j\le n\,, \quad 1\le k\le j, \\ \Delta_j\,, & 0\le j\le n-1\,, \ k=j+1\,, \\ 0\,, & 0\le j<n-1, \ j+1<k\le n\,, \end{cases} \tag{10.4}$$

with

$$\Delta_j = (1-|s_j|^2)^{1/2}\,. \tag{10.5}$$

One can, in the same way, obtain expressions for the matrices U^j of the unitary colligations representing the functions s_j, $1\le j\le n$.

It should be mentioned that a matrix of the form (10.3), (10.4) appeared in the paper [Ger, formula (66′)] and was then rediscovered a number times. See [Grg], [Con1], [Con2, Section 2.5], [Tep], [Sim, Chapter 4], [Dub, Theorem 2.17].

11. On work related to system theoretic interpretations of the Schur algorithm

In this section we discuss the connections between the present work and other work relating to the Schur algorithm as expressed in terms of system realizations. In particular, we discuss the results presented in [AADL] and in [KiNe].

The paper [AADL] deals with functions of the class $\boldsymbol{S}_{\boldsymbol{\kappa}}$, i.e., with the functions $s(z)$ meromorphic in the unit disc and possessing the properties:

1). For every N and for all points $z_1, \dots, z_N \in \mathbb{D}$ which are holomorphicity points for s, the matrix $\|K(z_p, z_q)\|_{1\le p,q\le N}$, $K(z,\zeta) = \frac{1-s(z)\overline{s(\zeta)}}{1-z\overline{\zeta}}$, does not have more than κ negative squares.

2). There exists an N and points $z_1, \dots, z_N \in \mathbb{D}$ such that this matrix has precisely κ negative squares.

One of the goals of the paper [AADL] is to discuss the Schur algorithm for functions from the class $\boldsymbol{S_\kappa}$ in terms of system realizations. In particular, the results of [AADL] are applicable to the special case[5] $\kappa = 0$, in which they can be simplified. In our considerations on the algebraic structure of a step of the Schur algorithm we will, for the sake of simplicity, restrict ourselves to finite-dimensional systems, which correspond to rational inner functions (of, say, degree n). We now describe the relevant result from [AADL], adopting the notation used there (to make the comparison with the results presented in our paper easier). In [AADL] the function $s(z)$ is given by

$$s(z) = s_0 + zB(I - zD)^{-1}C, \tag{11.1}$$

where B, C, D are entries of a unitary matrix U,

$$U = \begin{bmatrix} s_0 & B \\ C & D \end{bmatrix}, \quad B \in \mathfrak{M}_{1\times n},\, C \in \mathfrak{M}_{n\times 1},\, D \in \mathfrak{M}_{n\times n} \tag{11.2}$$

It is not *explicitly* assumed from the very beginning that the entry B of the matrix U has the special form (7.36). The matrix U appears as the matrix V, (1.2), in [AADL]. Our notation corresponds to that of [AADL] as follows: The objects, which appear as γ, v, u, T in formula (1.2) of [AADL] are s_0, B^*, C, D in our formulas (7.34)–(7.35). The state space which is denoted by $\mathcal{K}$ in (1.2) of [AADL] is the space $\mathcal{H} = \mathfrak{M}_{n\times 1}$ $(= \mathbb{C}^n)$ in our paper.

Let $s_1(z)$ be the Schur transform of the function $s(z)$,

$$s_1(z) = \frac{1}{z} \cdot \frac{s(z) - s_0}{1 - s(z)\overline{s_0}}, \qquad s_0 = s(0) \tag{11.3}$$

(or (7.33) in our paper). According to [AADL], $s_1(z)$ is representable in the form

$$s_1(z) = \alpha + z\beta(I - z\delta)^{-1}\gamma, \tag{11.4}$$

with

$$\alpha = \frac{1}{1 - |s_0|^2}\, BC, \qquad \beta = \frac{1}{\sqrt{1 - |s_0|^2}}\, BDP,$$
$$\gamma = \frac{1}{\sqrt{1 - |s_0|^2}}\, PC, \quad \delta = PDP, \tag{11.5}$$

where P is the matrix of the orthogonal projector onto the orthogonal complement of the vector B^* in $\mathcal{H}$, i.e.,

$$P \in \mathfrak{M}_{n\times n},\, \operatorname{rank} P = n - 1,\, P^2 = P,\, P = P^*,\, BP = 0\,.$$

(Formulas (11.5) are the formulas for the entries of the matrix V_1 which appear on page 11 of [AADL].) If we would like to represent the image space $P\mathcal{H}$ as the space $\mathcal{H}_1 = \mathfrak{M}_{(n-1)\times 1}$ $(= \mathbb{C}^{n-1})$, $\mathcal{H} = \mathbb{C} \oplus \mathcal{H}_1$, that is, if we would like the matrix

[5] $\boldsymbol{S_0}$ is the class of contractive functions holomorphic in the unit disc.

of the projector P to be of the form $P = \begin{bmatrix} 0 & 0 \\ 0 & I_{n-1} \end{bmatrix}$, then we have to replace the original matrix U with the matrix

$$U^0 = \begin{bmatrix} 1 & 0_{n\times n} \\ 0_{n\times n} & V^* \end{bmatrix} U \begin{bmatrix} 1 & 0_{n\times n} \\ 0_{n\times n} & V \end{bmatrix}, \quad U^0 = \begin{bmatrix} s_0 & B_0 \\ C_0 & D_0 \end{bmatrix} \tag{11.6}$$

where $V \in \mathfrak{M}_{n\times n}$ is a unitary matrix such that

$$V^* P V = \begin{bmatrix} 0 & 0 \\ 0 & I_{n-1} \end{bmatrix}. \tag{11.7}$$

The condition $BP = 0$ implies the condition $B_0 \begin{bmatrix} 1 & 0 \\ 0 & I_{n-1} \end{bmatrix} = 0$. The last equality means that B_0 is of the form $B_0 = \begin{bmatrix} b & 0_{1\times(n-1)} \end{bmatrix}$. Since the matrix U^0 is unitary, we have $|s_0|^2 + B_0 B_0^* = 1$. Therefore, B_0 must be of the form

$$B_0 = \begin{bmatrix} \delta(1-|s_0|^2)^{1/2} & 0_{1\times(n-1)} \end{bmatrix},$$

where δ is a unimodular complex number. The unitary matrix V from (11.6) is not unique: In this case, the degrees of freedom are clear, when we consider the replacement $V \to V \cdot \begin{bmatrix} \varepsilon & 0 \\ 0 & v \end{bmatrix}$, where ε is an arbitrary unimodular complex number and $v,\ v \in \mathfrak{M}_{(n-1)\times(n-1)}$ are unitary matrices. Choosing the number ε appropriately, we can ensure that B_0 is of the form

$$B_0 = \begin{bmatrix} (1-|s_0|^2)^{1/2} & 0_{1\times(n-1)} \end{bmatrix}. \tag{11.8}$$

Let us decompose the matrices C_0, D_0, which appear as the entries of the matrix U_0 from (11.5):

$$C_0 = \begin{bmatrix} c_1 \\ c_2 \end{bmatrix}, \qquad D_0 = \begin{bmatrix} d_{11} & d_{12} \\ d_{21} & d_{22} \end{bmatrix}, \tag{11.9}$$

$$c_1 \in \mathfrak{M}_{1\times 1},\ c_2 \in \mathfrak{M}_{(n-1)\times 1},$$

$$D_{11} \in \mathfrak{M}_{1\times 1},\ D_{12} \in \mathfrak{M}_{1\times(n-1)},\ D_{21} \in \mathfrak{M}_{(n-1)\times 1},\ D_{22} \in \mathfrak{M}_{(n-1)\times(n-1)}.$$

The equalities (11.5) (where B, C, D are replaced by B_0, C_0, D_0) now take the form

$$\begin{aligned} \alpha &= \frac{1}{\sqrt{1-|s_0|^2}} c_1, \quad \beta = d_{12}, \\ \gamma &= \frac{1}{\sqrt{1-|s_0|^2}}\, c_2, \quad \delta = d_{22}\,, \end{aligned} \tag{11.10}$$

Thus, the matrix

$$U^1 = \begin{bmatrix} \alpha & \beta \\ \gamma & \delta \end{bmatrix}, \tag{11.11}$$

from [AADL], whose entries appear in the representation (11.4) of the function $s_1(z)$ is the same matrix which appears in our Theorem 7.2 as the matrix (7.38).

(The matrix V_1 from page 11 of [AADL] can be considered as a *coordinate-free* expression for the colligation matrix representing the function $s_1(z)$.) The difference between our work and the work [AADL] is not in the results but in the methods. The reason for choosing the expression for the colligation matrix U_1 given in [AADL] is not fully explained. The facts that the matrix U_1 is unitary and that the matrix U_1 represents the Schur transform s_1 of the function s are obtained as the result of a long chain of *formal* calculations. These calculations come across as somewhat contrived and do not serve to further our understanding of the subject at hand.

The state system approach is much more transparent. The fact that the matrix U_1 is unitary is an immediate consequence of our formula (7.6). The fact that the matrix U_1 represents the function s_1 is a consequence of the interpretation of the linear fractional transform (6.6)–(6.7) in terms of the Redheffer coupling of the appropriate colligation.

The paper [KiNe] can also be considered as relevant to our paper. In [KiNe] the system representation of Schur functions is not considered at all. Nevertheless, in this work the Householder algorithm is used to calculate the sequence of numbers, which can be identified with the Schur parameters of the rational inner function naturally related to the appropriate unitary matrix. Namely, given a unitary matrix $U \in \mathfrak{M}_{(n+1)\times(n+1)}$, the measure μ on the unit circle is related to U in the following way: $\mu(dt) = (E(dt)e_1,\, e_1)$, where $E(dt)$ is the spectral measure of the matrix U and $e_1 = (1, 0, \ldots, 0)^T$, $e_1 \in \mathfrak{M}_{(n+1)\times 1}$. It is assumed that e_1 is a cyclic vector of U. The following equality holds:

$$e_1^* \frac{I + zU}{I - zU} e_1 = \int_{\mathbb{T}} \frac{1 + zt}{1 - zt} \mu(dt) \tag{11.12}$$

The measure μ generates the (finite) sequence of polynomials orthogonal on the unit circle. These orthogonal polynomials (Φ_k is monic of degree k) satisfy the recurrence relations

$$\Phi_{k+1}(z) = z\Phi_k(z) - \overline{s}_k \Phi_k^*(z) \tag{11.13}$$

$$\Phi_{k+1}^*(z) = z\Phi_k^*(z) - s_k z\Phi_k(z) \tag{11.14}$$

where s_k, $k = 0, 1, \ldots, n$ are some recurrence coefficients. There are many different names for these coefficients. Recently dubbed 'Verblunsky parameters' by Barry Simon in [Sim]. On the other hand, the function in (11.12), which we denote by $p(z)$ is holomorphic in the unit disc $\mathbb{D}$ and has the following properties.

$$p(0) = 1, \qquad p(z) + \overline{p(z)} \geq 0 \quad (z \in \mathbb{D}).$$

Therefore $p(z)$ is representable in the form

$$p(z) = \frac{1 + zs(z)}{1 - zs(z)}, \tag{11.15}$$

where $s(z)$ is a function holomorphic and contractive in $\mathbb{D}$. Ya.L. Geronimus established that the Verblunsky coefficients $s(z)$ in the recurrence relations (11.13)–(11.14) are also the Schur parameters of the functions $s(z)$, which appear in (11.15).

From (11.12) and (11.15) it follows that

$$e_1^* \frac{I + zU}{I - zU} e_1 = \frac{1 + zs(z)}{1 - zs(z)}. \tag{11.16}$$

In Lemma 3.2 of [KiNe], the following method for finding Schur (=Verblunsky) parameters was proposed: First, the given unitary matrix U should be converted to Hessenberg form:

$$U^0 = \begin{bmatrix} 1 & 0_{1\times n} \\ 0_{n\times 1} & V^* \end{bmatrix} U \begin{bmatrix} 1 & 0_{1\times n} \\ 0_{n\times 1} & V \end{bmatrix}, \tag{11.17}$$

In [KiNe] it is claimed that the entries of the (lower Hessenberg) matrix U^0 are of the form (10.3)–(10.5), from which the Schur-Verblunsky parameters s_k can be found. However, it follows from (11.16) that

$$s(z) = A + sB(I - zD)^{-1}C, \tag{11.18}$$

where

$$U = \begin{bmatrix} A & B \\ C & D \end{bmatrix} \tag{11.19}$$

$$A \in \mathfrak{M}_{1\times 1}, \quad B \in \mathfrak{M}_{1\times n}, \quad C \in \mathfrak{M}_{n\times 1}, \quad D \in \mathfrak{M}_{n\times n}.$$

Thus the formula (11.18) can be interpreted as the system representation of the function $s(z)$. The formula (11.18), where $s(z)$ is defined by (11.16) from U was unfamiliar to us, but we do not think that this formula is new.

In his forthcoming paper [Arl] Yu.M. Arlinskii studied a related question for operator-valued Schur functions Θ acting between separable Hilbert spaces. These investigations correspond to the operator generalization of the classical Schur algorithm which is due to Constantinescu (see Section 1.3 in [BC].) Yu.M. Arlinskii presents a construction of conservative and simple realizations of the Schur algorithm iterates Θ_n of Θ by means of the conservative and simple realization of Θ.

Appendix: System realizations of inner rational functions

We prove that every complex-valued (i.e., scalar) inner rational function of degree n can be represented as the characteristic function of the minimal unitary colligation associated with some unitary matrix $U \in \mathfrak{M}_{(n+1)\times(n+1)}$. Let us denote a given rational inner function by S. The operator colligation whose characteristic function is S will be constructed as the 'left shift' operator in the appropriate space of analytic functions constructed from S. A similar construction appears in a paper by B.Sz. Nagy-C.Foias. See [SzNFo, Chapter VI]. The construction of B.Sz. Nagy-C.Foias was adapted to unitary colligations in [BrSv2].

1. The space K_S. The most important part of our construction is the Hilbert space K_S of rational functions. We consider S as a function defined on the unit circle $\mathbb{T}$, i.e., $S : \mathbb{T} \to \mathbb{T}$. As usual,

$$L^2 = \{x : \mathbb{T} \to \mathbb{C}, \|x\| < \infty\},$$

where $\|x\|^2 = \langle x\,,\, x\rangle$ and

$$\langle x\,,\, y\rangle = \int\limits_{\mathbb{T}} x(t)\,\overline{y(t)}\,m(dt)\,,$$

$m(dt)$ is the normalized Lebesgue measure on $\mathbb{T}$. Let H^2_+ and H^2_- be the Hardy subspaces of the space L^2:

$$H^2_+ = \{x \in L^2 : \langle x(t)\,,\, t^k\rangle = 0,\quad k = -1, -2, \dots\}\,.$$

$$H^2_- = \{x \in L^2 : \langle x(t)\,,\, t^k\rangle = 0,\quad k = 0,\, 1,\, 2,\, \dots\ \}\,.$$

Clearly,

$$L^2 = H^2_+ \oplus H^2_-\,.$$

It is also convenient to consider the functions from H^2_+ and from H^2_- as functions holomorphic in $\mathbb{D}$ and in $\mathbb{D}^-$, respectively. In particular, the evaluation $f \to f(0)$ is defined for every f in H^2_+ and $f(\infty) = 0$ for every f in H^2_-.

The space K_S is defined as

$$K_S = H^2_+ \ominus S\,H^2_+\,, \tag{A.1}$$

where $S\,H^2_+ = \{S(t)\,h(t) :\, h \in H^2_+\}$. Another description of the space K_S is:

$$K_S = \{x \in L^2 :\, x \in H^2_+,\, xS^{-1} \in H^2_-\,\}. \tag{A.2}$$

It can be shown that the space K_S consists of rational functions whose poles are contained in the set of poles of the function S and that $\dim K_S = \deg S$. If all zeros z_k of S are simple (see (1.1)), then the space K_S is generated by the functions $\{(1 - t\overline{z_k})^{-1}\}_{1\le k\le n}$. If S has non-simple zeros, the modification of this statement is clear. The space K_S is a reproducing kernel Hilbert space. If $f \in K_S$, then

$$f(z) = \langle f(t), K(t,z)\rangle\,, \tag{A.3}$$

where the reproducing kernel $K(t,z)$ is:

$$K(t,z) = \frac{1 - S(t)\overline{S(z)}}{1 - t\overline{z}}\,. \tag{A.4}$$

2. The left shift operator. The left shift operator T is defined as

$$T(f)(t) = (f(t) - f(0)e(t))\cdot t^{-1} \quad \text{for } f \in H^2_+\,. \tag{A.5}$$

where

$$e(t) = 1 \qquad \forall t \in \mathbb{T}\,. \tag{A.6}$$

This operator is contractive:

$$\|Tf\|^2 = \|f\|^2 - |f(0)|^2 \quad \forall f \in H^2_+\,. \tag{A.7}$$

The space K_S, considered as a subspace of H^2_+, is an invariant subspace of the left shift operator T. This is evident from the description (A.2) of the space K_S.

3. The construction of the unitary colligation U. The unitary colligation $(\mathcal{E},\ \mathcal{H},\ U)$ (see Definition 3.1) is defined as follows: Let the state space $\mathcal{H}$ be the space K_S and

let the principal operator D be the left shift operator T, (A.5), restricted to K_S:

$$\mathcal{H} = K_S, \quad Df(t) = (f(t) - f(0)e(t)) \cdot t^{-1} \quad \forall f \in \mathcal{H}\,. \tag{A.8}$$

The equality $\|Df\|^2 + |f(0)|^2 = \|f\|^2$, together with the requirement that the colligation operator U, (3.1)–(3.2), be unitary, prompts us to define the exterior space $\mathcal{E}$ and the channel operator $B : \mathcal{H} \to \mathcal{E}$ as follows:

Let $\mathcal{E}$ be a one-dimensional Hilbert space which is identified with the *vector space* $\mathbb{C}$ over the *field* $\mathbb{C}$ *of scalars*. We choose the number $\beta = 1$ as a basis *vector* in $\mathbb{C}$ and will denote this basis vector by $\mathbb{1}$. Every number $\varepsilon \in \mathbb{C}$, considered as an element $\boldsymbol{\varepsilon}$ of the *vector space* $\mathbb{C}$, can be presented as $\boldsymbol{\varepsilon} = \varepsilon\mathbb{1}$, where the factor in front of $\mathbb{1}$ is the same *number* ε, but considered as an element of the *field of scalars* $\mathbb{C}$.

The channel operator B is:

$$(Bf)(t) = f(0)\mathbb{1}, \quad \forall f \in \mathcal{H}\,. \tag{A.9}$$

Equation (A.7) ensures that

$$\|Bf\|^2_{\mathcal{E}} + \|Df\|^2_{\mathcal{H}} = \|f\|^2_{\mathcal{H}}\,, \quad \forall f \in \mathcal{H}\,.$$

$f(0)$, which appears in (A.8) and (A.9), can be represented using the reproducing kernel (A.3)–(A.4). Let

$$k(t) = 1 - s(t)\,\overline{s(0)}, \quad (= K(t,0)\,)\,. \tag{A.10}$$

Then

$$Bf = \langle\, f, k\,\rangle\, \mathbb{1}\,. \tag{A.11}$$

The operator $A : \mathcal{E} \to \mathcal{E}$ (as is the case for every operator in $\mathcal{E} : \dim \mathcal{E} = 1$) is of the form

$$A\boldsymbol{\varepsilon} = \alpha\langle\, \boldsymbol{\varepsilon}, \mathbb{1}\,\rangle\, \mathbb{1}, \quad \boldsymbol{\varepsilon} \in \mathcal{E},$$

where $\alpha \in \mathbb{C}$. Since the vector $\mathbb{1}$, which generates $\mathcal{E}$, is orthogonal to $\mathcal{H}$ in the orthogonal sum $\mathcal{E} \oplus \mathcal{H}$, the unitary property of U implies that

$$\langle Bf\,,\, A\mathbb{1}\rangle + \langle Df\,, C\mathbb{1}\,\rangle = 0 \quad \forall\, f \in \mathcal{H}\,. \tag{A.12}$$

Therefore

$$\alpha\langle Bf\,,\, \mathbb{1}\rangle + \langle Df\,, C\mathbb{1}\,\rangle = 0 \quad \forall\, f \in \mathcal{H}\,.$$

Let us denote

$$C\mathbb{1} = l, \quad l \in \mathcal{H}\,.$$

Equation (A.12) means that

$$\overline{\alpha}\langle\, f, k\,\rangle + \langle\, Df, l\,\rangle = 0\,, \quad \forall f \in \mathcal{H}\,.$$

Thus, one should take

$$l = -\alpha(D^*)^{-1}k, \tag{A.13}$$

where D^* is the adjoint to the operator D, with respect to the scalar product $\langle\,,\,\rangle$.

We now look to determine the operator D^*. The equality

$$\langle\, Df, g\,\rangle = \langle\, f, D^*g\,\rangle \quad \forall f, g \in \mathcal{H}$$

means that

$$\langle\,(f(t)-f(0)e(t))\,t^{-1},g(t)\,\rangle=\langle\,f(t),(D^*g)(t)\,\rangle\,.$$

The last equality implies that

$$(D^*g)(t)=P(tg(t)),\quad\forall g\in K_S \tag{A.14}$$

where P is the orthogonal projector from L^2 onto K_S. Clearly,

$$\langle\,h(t),S(t)\,\rangle=0\ \ \forall h\in K_S\,,$$

and

$$tg(t)-\langle\,tg(t),S(t)\,\rangle S(t)\in K_S\ \forall g\in K_S\,.$$

Therefore,

$$P(tg(t))=tg(t)-\langle\,tg(t),S(t)\,\rangle S(t)\ \ \forall g\in K_S\,,$$

that is

$$(D^*g)(t)=tg(t)-\langle\,tg(t),S(t)\,\rangle S(t)\ \ \forall g\in\mathcal{H}\,. \tag{A.15}$$

From (A.13), we obtain

$$l(t)=\frac{\alpha}{S(0)}\frac{S(t)-S(0)e(t)}{t}\,.$$

$\|Ae\|^2+\|Ce\|^2=1$ gives us $|\alpha|=|S(0)|$. We choose

$$\alpha=S(0).$$

(Later we see that this is the only possible choice for α.) We set

$$l(t)=\frac{S(t)-S(0)e(t)}{t}\,. \tag{A.16}$$

(In intermediate steps we assumed that $S(0)\neq 0$, but this does not appear in the final expression (A.16) for $l(t)$.) Thus,

$$A\boldsymbol{\varepsilon}=S(0)\langle\,\boldsymbol{\varepsilon},\mathbb{1}\,\rangle\,\mathbb{1},\quad Bf=\langle\,f,k\,\rangle\,\mathbb{1},\quad C\boldsymbol{\varepsilon}=\langle\,\boldsymbol{\varepsilon},\mathbb{1}\,\rangle\,l,$$

$$(Df)(t)=(\,f-\langle\,f,k\,\rangle e(t)\,)\,t^{-1}\quad\forall\boldsymbol{\varepsilon}\in\mathcal{E},\ f\in\mathcal{H}\,. \tag{A.17}$$

or

$$A\boldsymbol{\varepsilon}=\varepsilon S(0)\mathbb{1}\,,\quad Bf=f(0)\mathbb{1},\quad (C\boldsymbol{\varepsilon}))=\varepsilon l(t),$$

$$(Df)(t)=\big(\,f(t)-f(0)e(t)\,\big)\,t^{-1}\,,\quad\forall\boldsymbol{\varepsilon}=\varepsilon\mathbb{1}\in\mathcal{E},\ f\in\mathcal{H}\,. \tag{A.18}$$

From (A.18) it follows that the block-operator

$$U=\begin{bmatrix}A & B\\ C & D\end{bmatrix} \tag{A.19}$$

is unitary. (After the block D was chosen, the other blocks A, B, C were chosen to ensure that U be a unitary operator.) The characteristic function $S_U(z)$,

$$S_U(z)=A+zB(I-zD)^{-1}C\,, \tag{A.20}$$

of the colligation U coincides with the original rational inner function $S(z)$. This can be checked by direct calculation of $S_U(z)$ using the expression (A.18) for blocks

of the colligation operator U. The expression for the operator $(I - zD)^{-1}$, which is needed for this calculation, is

$$\big((I - zD)^{-1} f\big)(t) = \frac{tf(t) - zf(z)}{t - z}, \quad \forall f \in \mathcal{H}\,. \tag{A.21}$$

In what follows we also need the expression for the operator $(I - zD^*)^{-1}$:

$$\big((I - zD^*)^{-1} f\big)(t) = \frac{f(t) - (fS^{-1})(z^{-1})S(t)}{1 - tz}, \quad \forall f \in \mathcal{H}\,. \tag{A.22}$$

(Since the function fS^{-1} belongs to H^2_-, the evaluation $fS^{-1} \to (fS^{-1})(z^{-1})$ is defined for $z \in \mathbb{D}$.)

Choosing an orthogonal basis in the (n-dimensional) Hilbert space K_S, we realize that the unitary operator U, (A.18)–(A.19), originally constructed as an operator acting in *a functional* space, is a matrix operator acting in $\mathbb{C}^{n+1} = \mathbb{C} \oplus \mathbb{C}^n$.

Definition 11.1. *The colligation* (A.18)–(A.19) *is called the model unitary colligation constructed from the rational inner function* S.

4. Minimality of the model unitary colligation ($\mathcal{E}$, $\mathcal{H}$, U). We look to prove that the model colligation U, (A.18)–(A.19), is controllable and observable. In view of the expression for the channel operator C (one-dimensional), controllability of U can be formulated as follows:

$$\textit{The set of vectors } \{(I - zD)^{-1} l\}_{z \in \mathbb{D}} \ \textit{ generates the space } K_S. \tag{A.23}$$

From (A.16) and (A.21) it follows that

$$\big((I - zD)^{-1} l\big)(t) = \frac{S(t) - S(z)}{t - z}\,.$$

Let $f \in L^2$ be such that

$$\int_{\mathbb{T}} \frac{S(t) - S(z)}{t - z}\, \overline{f(t)}\, m(dt) = 0 \ \ \forall\, z \in \mathbb{D} \tag{A.24}$$

If $f \in H^2_+$, then $\int_{\mathbb{T}} \frac{\overline{f(t)}}{t - z}\, m(dt) = 0\ \forall z \in \mathbb{D}$, hence, $\int_{\mathbb{T}} \frac{S(t)\overline{f(t)}}{t - z}\, m(dt) = 0\ \forall\, z \in \mathbb{D}$. The last equality implies that $\overline{f(t)S^{-1}(t)} \in H^2_-$. (Here we use that $S(t) = \overline{S^{-1}(t)}$ for $t \in \mathbb{T}$.) If also $f(t)S^{-1}(t) \in H^2_-$, then $f(t)S^{-1}(t) \equiv 0$ and $f \equiv 0$. Therefore, if the condition (A.24) holds for some $f \in K_S$, then $f \equiv 0$. Controllability of the colligation ($\mathcal{E}$, $\mathcal{H}$, U) is thus proved.

Observability of this colligation can be proved analogously. According to (A.11), $B^* f = \langle f, e \rangle\, k$. Therefore the observability criterion is reduced to the statement:

$$\textit{The set of vectors } \{(I - zD^*)^{-1} k\}_{z \in \mathbb{D}} \ \textit{ generates the space } K_S. \tag{A.25}$$

Using expressions (A.22) and (A.10), we obtain:

$$\big((I - zD^*)^{-1} k\big)(t) = \frac{1 - S(t)\, S^{-1}(z^{-1})}{1 - tz}\,.$$

Let $f \in L^2$ is such that

$$\int_{\mathbb{T}} \frac{1 - S(t)S^{-1}(z^{-1})}{1 - tz}\, \overline{f(t)}\, m(dt) = 0 \quad \forall\, z \in \mathbb{D}\,. \tag{A.26}$$

If $f(t)S^{-1}(t) \in H^2_-$, then $\int\limits_{\mathbb{T}} \frac{S(t)\overline{f(t)}}{1-tz} = 0$, hence $\int\limits_{\mathbb{T}} \frac{\overline{f(t)}}{1-tz} m(dt) = 0\ \forall z \in \mathbb{D}$ and $f(t) \in H^2_-$. If also $f \in H^2_+$, then $f \equiv 0$. Therefore if the condition (A.26) holds for some $f \in K_S$, then $f \equiv 0$.

5. Uniqueness of simple realization. The uniqueness of the minimal realization is, in fact, a version of a result by M.S. Livshitz, which, in the language of M.S. Livshitz, claims that the characteristic function uniquely determines (up to unitary equivalence) the operator colligation without complementary component.

Let $U_1 = \begin{bmatrix} A_1 & B_1 \\ C_1 & D_1 \end{bmatrix}$ and $U_2 = \begin{bmatrix} A_2 & B_2 \\ C_2 & D_2 \end{bmatrix}$ be two unitary matrices divided into blocks,

$$A_j \in \mathfrak{M}_{1\times 1},\quad B_j \in \mathfrak{M}_{1\times n_j},\quad C_j \in \mathfrak{M}_{n_j\times 1},\quad D_j \in \mathfrak{M}_{n_j\times n_j},$$
$$\text{where } n_j,\ j = 1,\ 2,\ \text{are natural numbers.} \tag{A.27}$$

We do not assume that $n_1 = n_2$.

Let

$$S_i(z) = A_i + B_i(I - zD_i)^{-1}C_i, \quad i = 1, 2,$$

be the characteristic functions of the unitary colligations associated with the matrices U_1 and U_2 respectively. Suppose that

1. The characteristic functions are equal.

$$S_1(z) \equiv S_2(z). \tag{A.28}$$

2. Each of the matrices U_1 and U_2 is simple in the sense of Definition 3.14.

We prove that under these assumptions the matrices U_1 and U_2 are equivalent in the sense of Definition 3.13, in particular, $n_1 = n_2$.

To prove this, we have to first of all construct a unitary mapping V of the space $\mathbb{C}^{n_2}$ onto the space $\mathbb{C}^{n_1}$. Assume that the matrices U_1 and U_2 are simple. Let us consider the vectors $f_j^k,\ g_j^l \in \mathbb{C}^{n_j}\ (= \mathfrak{M}_{n_j\times 1}),\ j = 1, 2,\ 0 \le k, l$:

$$f_j^k = D_j^k C_j,\quad g_j^l = (D_j^*)^l B_j^*,\ j = 1,\ 2,\ 0 \le k, l\,. \tag{A.29}$$

By the assumption, for each $j = 1, 2$, the vectors $f_j^k,\ g_j^k,\ 0 \le k \le \max(n_1, n_2)$ generate the space $\mathbb{C}^{n_j}$. The equality (A.28) implies

$$A_1 = A_2\,, \tag{A.30}$$

and the equalities

$$B_1 D_1^p C_1 = B_2 D_2^p C_2,\ 0 \le p, \tag{A.31}$$

or

$$B_1 D_1^l D_1^k C_1 = B_2 D_2^l D_2^k C_2,\quad 0 \le k, l\,.$$

The latter equalities can be interpreted as

$$\langle f_1^k, g_1^l\rangle_{\mathbb{C}^{n_1}} = \langle f_2^k, g_2^l\rangle_{\mathbb{C}^{n_2}}, \quad \forall k,l : 0 \le k,\, 0 \le l\,. \tag{A.32a}$$

Moreover, the equalities (A.28) imply that

$$1 - S_1^*(\zeta)S_1(z) \equiv 1 - S_2^*(\zeta)S_2(z), \quad 1 - S_1(z)S_1^*(\zeta) \equiv 1 - S_2(z)S_2^*(\zeta)\,.$$

In view of (3.22), (3.23), the latter equalities imply that

$$C_1^*(D_1^*)^q D_1^p C_1 = C_2^*(D_2^*)^q D_2^p C_2, \quad \text{and} \quad B_1 D_1^q (D_1^*)^p C_1^* = B_2 D_2^q (D_2^*)^p C_2^*, \qquad 0 \le p,\, q\,.$$

This can, in turn, be interpreted as

$$\langle f_1^p, f_1^q\rangle_{\mathbb{C}^{n_1}} = \langle f_2^p, f_2^q\rangle_{\mathbb{C}^{n_2}}, \quad \text{and} \quad \langle g_1^p, g_1^q\rangle_{\mathbb{C}^{n_1}} = \langle g_2^p, g_2^q\rangle_{\mathbb{C}^{n_2}}, \qquad \forall p,q : 0 \le p,\, 0 \le q\,. \tag{A.32b}$$

From (A.32) it follows that for arbitrary α_k, β_l (such that only finitely many of them differ from zero),

$$\big\| \sum \alpha_k f_1^k + \sum \beta_l g_1^l \big\|_{\mathbb{C}^{n_1}} = \big\| \sum \alpha_k f_2^k + \sum \beta_l g_2^l \big\|_{\mathbb{C}^{n_2}}\,. \tag{A.33}$$

Let us define the operator $V : \mathbb{C}^{n_2} \to \mathbb{C}^{n_1}$ first as

$$V f_2^k = f_1^k, \quad V g_2^l = g_1^l, \quad \forall \;\; k \ge 0,\, l \ge 0, \tag{A.34a}$$

and then extend this operator by linearity to all vector columns $h \in \mathbb{C}^{n_2}$ representable as a finite linear combination of the form $h = \sum \alpha_k f_2^k + \sum \beta_l g_2^l$. Thus,

$$V\Big(\sum \alpha_k f_2^k + \sum \beta_l g_2^l\Big) = \sum \alpha_k f_1^k + \sum \beta_l g_1^l\,. \tag{A.34b}$$

If some $h \in \mathbb{C}^{n_2}$ admits two different representations, say

$$h = \sum \alpha'_k f_2^k + \sum \beta'_l g_2^l, \quad \text{and} \quad h = \sum \alpha''_k f_2^k + \sum \beta''_l g_2^l\,,$$

then Vh also admits two different representations:

$$Vh = \sum \alpha'_k f_1^k + \sum \beta'_l g_1^l, \quad \text{and} \quad Vh = \sum \alpha''_k f_1^k + \sum \beta''_l g_1^l\,.$$

However, since $\sum \alpha_k f_2^k + \sum \beta_l g_2^l = 0$, where $\alpha_k = \alpha''_k - \alpha'_k$, $\beta_l = \beta''_l - \beta'_l$, the equality (A.33) implies that $\sum \alpha_k f_1^k + \sum \beta_l g_1^l = 0$, i.e.,

$$\sum \alpha'_k f_1^k + \sum \beta'_l g_1^l = \sum \alpha''_k f_1^k + \sum \beta''_l g_1^l\,.$$

The definition (A.34) of V is thus non-contradictory.

The operator V is defined on the linear hull of all vectors $\{f_2^k, g_2^l\}_{k,l}$ and isometrically maps its definition domain onto the linear hull of all vectors $\{f_1^k, g_1^l\}_{k,l}$. If both the matrices U^2, U^1 are simple, then these linear hulls are the whole spaces $\mathbb{C}^{n_2}$ and $\mathbb{C}^{n_1}$, respectively. In this case $n_1 = n_2 \,(\stackrel{\text{def}}{=} n)$ and

$$V^*V = I_n, \quad VV^* = I_n \tag{A.35}$$

We now prove the intertwining relation

$$\begin{bmatrix} A_1 & B_1 \\ C_1 & D_1 \end{bmatrix} \begin{bmatrix} 1 & 0 \\ 0 & V \end{bmatrix} = \begin{bmatrix} 1 & 0 \\ 0 & V \end{bmatrix} \begin{bmatrix} A_2 & B_2 \\ C_2 & D_2 \end{bmatrix},$$

which can be rewritten as follows:

$$C_1 = VC_2, \quad B_2 = B_1V, \tag{A.36}$$

and

$$VD_2 = D_1V. \tag{A.37}$$

The first of the equalities (A.36) corresponds to the first of the equalities (A.29) for $k = 0$ (see (A.34a) for $k = 0$). The second of the equalities (A.36) relates to the second of equality in (A.29) for $l = 0$ (see (A.35)).

To check the splitting relation (A.37), it is enough to check that

$$VD_2f_2^k = D_1Vf_2^k \quad \text{for} \ \ \forall k \geq 0, \tag{A.38a}$$

and

$$VD_2g_2^l = D_1Vg_2^l \quad \text{for} \ \ \forall l \geq 0. \tag{A.38b}$$

The equality (A.38a) is an obvious consequence of the definitions of the operator V and vectors f_j^k. Indeed, $D_2f_2^k = f_2^{k+1}$, $Vf_2^{k+1} = f_1^{k+1}$. On the other hand, $Vf_2^k = f_1^k$, $D_1f_1^k = f_1^{k+1}$. Therefore, (A.38a) holds.

Our approach to checking the condition (A.38b) will be different in the two cases $l = 0$ and $l > 0$. For $l = 0$ the equality (A.38b) takes the form $VD_2B_2^* = D_1VB_2^*$. (A.34) for $l = 0$ means that $VB_2^* = B_1^*$, so we should check that $VD_2B_2^* = D_1B_1^*$. Since $D_jB_j^* = -C_jA_j^*$, $j = 1, 2$, the last equality is equivalent to $VC_2A_2^* = C_1A_1^*$. The latter equation is a consequence of the first of the equalities (A.36). Thus, (A.38b) holds for $l = 0$.

We check condition (A.38b) for $l > 0$. Since the matrices U_j are unitary, we have that $D_jD_j^* = I - C_jC_j^*$. Thus, (A.38b) is equivalent to

$$V(I - C_2C_2^*)(D_2^*)^{l-1}B_2^* = (I - C_1C_1^*)(D_1^*)^{l-1}B_1^*.$$

This equation is a consequence of the following three equalities:

$$V(D_2^*)^{l-1}B_2^* = (D_1^*)^{l-1}B_1^*, \tag{A.39a}$$

$$VC_2 = C_1, \tag{A.39b}$$

and

$$C_2^*(D_2^*)^{l-1}B_2^* = C_1^*(D_1^*)^{l-1}B_1^*. \tag{A.39c}$$

(A.39a) holds, because it can be written as $Vg_2^{l-1} = g_1^{l-1}$, which is part of the definition (A.34) of the operator V. (A.39b) has already been checked: This is the first of the relations (A.36). (A.39c) is the same as (A.31) for $p = l - 1$. The condition (A.38b) has also been checked for $l > 0$.

6. Simple realization is minimal. Let $U \in \mathfrak{M}_{(1+n)\times(1+n)}$ be a simple unitary matrix. The matrix U is then minimal. Indeed, let $S(z)$ be the characteristic function of the unitary colligation associated with U. $S(z)$ is a rational inner function, $\deg S \leq n$. Let $(\mathbb{C}, K_S, T)$ be the model colligation constructed from this S. We established

that the model unitary colligation is minimal (in particular, simple) and that its characteristic function is the function S, from which it was constructed. Both colligations (the original colligation and the model colligation) have the same characteristic function and both are simple. Hence, these colligations are equivalent. Since the model colligation is minimal, the original colligation is also minimal.

Acknowledgement

We thank Professors I. Gohberg and M.A. Kaashoek for useful remarks and suggestions on the history and scope of the state space method and its applications.

Moreover, we thank Professor D. Alpay for drawing our attention to the paper [AADL].

We thank Armin Rahn for his careful reading of the manuscript and his help in improving the English in this paper.

References

[AADL] Alpay, D., Azizov, T., Dijksma, A., Langer, H. *The Schur algorithm for generalized Schur functions I: coisometric realizations.* Pp. 1–36 in: Operator Theory: Adv. Appl. **129**, Borichev, A., Nikolski, N. – editors. Birkhäuser Verlag, Basel 2001.

[AG] Alpay, D., Gohberg, I. *Inverse problem for Sturm-Liouville operators with rational reflection-coefficient.* Integral Equations Operator Theory, **30**:3, (1998), pp. 317–325.

[Arl] Arlinskii, Yu.M. *Iterates of the Schur class operator-valued functions and their conservative realizations.* arXiv:0801.4267 [math.FA]
Submitted to: Operators and Matrices.

[Ar] Аров, д.З. *Пассивные линейные стационарные динамические системы.* Сибирск. Мат. Журн., **20**:2 (1979), 211–228.
English transl.:
Arov, D.Z. *Passive linear stationary dynamic systems.* Siberian Math. J., **20**:2 (1979), 149–162.

[Ausg] *Ausgewählte Arbeiten zu den Ursprüngen der Schur-Analysis* (German). [*Selected papers on the origins of Schur analysis*]. (Fritzsche, B. and B. Kirstein – editors) (German). (Series: Teubner-Archiv zur Mathematik, Volume 16). B.G. Teubner Verlagsgesellschaft, Stuttgart·Leipzig 1991, 290pp.

[ARC] *Automatic and Remote Control.* (Proc. of the First Int. Congress of International Federation of Autom. Control (IFAC). Moscow 1960.) Coales, J.F., Ragazzini, J.R., Fuller, A.T. – editors. Botterworth, London 1961.

[BC] Baconyi, M., Constantinescu, T. *Schur's algorithm and several applications.* Pitman Research Notes in Math., Volume **261**, Longman, Harlow 1992.

[BGR] Ball, J.A., Gohberg, I., Rodman, L. *Interpolation of rational matrix-functions.* Operator Theory: Advances and Applications, **OT 45**. Birkhäuser Verlag, Basel, 1990. xii+605 pp.

[BGK] Bart, H., Gohberg, I., Kaashoek, M.A. *Minimal Factorization of Matrix and Operator Functions.* Operator Theory: Advances and Applications, **OT 1**. Birkhäuser, Basel·Boston·Stuttgart, 1979.

[BGKR] Bart, H., Gohberg, I., Kaashoek, M.A., Ran, A. *Factorization of matrix and operator functions: the state space method.* Operator Theory: Advances and Applications, **OT 178**. Birkhäuser, Basel·Boston·Stuttgart, 2008. xii+409 pp.

[BGKV] Bart, H., Gohberg, I., Kaashoek, M.A., Van Dooren, P. *Factorizations of Transfer Functions.* SIAM J. Control and Optimization. **18**:6 (1980), 675–696.

[BFK1] Bogner, S., Fritzsche, B., Kirstein, B. *The Schur-Potapov algorithm for sequences of complex $p \times q$-matrices.* I. Compl. Anal. Oper. Theory 1 (2007), 55–95.

[BFK2] Bogner, S., Fritzsche, B., Kirstein, B. *The Schur-Potapov algorithm for sequences of complex $p \times q$-matrices.* II. Compl. Anal. Oper. Theory 1 (2007), 235–278.

[Br] Бродский, М.С. *Треугольные и жордановы представления линейных операторов* . Наука, Москва 1969, 287 сс.
English transl.:
Brodskiĭ, M.S. *Triangular and Jordan Representation of linear operators.* Transl. of Math. Monogr. **32**. Amer. Mat. Soc, Providence, RI, 1971. viii + 246 pp.

[BrSv1] Бродский, В.М., Шварцман, Я.С. *Об инвариантных подпространствах сжатий.* доклады Академии наук СССР, **201**:3 (1957), 519–522. English transl.:
Brodskiĭ, V.M., Švarcman, Ja.S. *On invariant subspaces of contractions.* Soviet. Math. Dokl., **12**:6 (1971), 1659–1663.

[BrSv2] Бродский, В.М., Шварцман, Я.С. *Инвариантные подпространства сжатия и факторизация характеристической функции.* (Russian). Теория функций, функциональный анализ и их приложения. [Teor. Funkciĭ, Funkcional. Anal. i Priložen.] **12**:6 (1971), 15–35, 160.

[BrLi] Бродский, М.С., Лившиц, М.С. *Спектральный анализ несамосопряженных операторов и промежуточные системы.* Успехи Мат. Наук, том /**13**:1 (1957), 3–85. English transl.:
Brodskiĭ, M.S., Livšic, M.S. *Spectral analysis of non-selfadjoint operators and intermediate systems.* Amer. Math. Soc. Transl. (**2**), **13** (1958), 265–346.

[CWHF] *Constructive Methods of Wiener-Hopf Factorization.*
Gohberg, I., Kaashoek, M.A. – editors. Birkhäuser, Basel·Boston·Stuttgart 1986. 324 pp.

[Con1] Constantinescu, T. *On the structure of the Naimark dilation.* Journ. Operator Theory, **12** (1984), pp. 159–175.

[Con2] Constantinescu, T. *Schur Parameters, Factorization and Dilations Problems.* Operator Theory: Advances and Applications, **OT 82**. Birkhäuser Verlag, Basel 1996. ix+253pp.

[Dew1] Dewilde, P. *Cascade scattering matrix synthesis.* Tech. Rep. 6560-21, Information Systems Lab., Stanford University, Stanford 1970.

[Dew2] Dewilde, P. *Input-output description of roomy systems.* SIAM Journ. Control and Optim., **14**:4 (1976), 712–736.

[Dub] Dubovoy, V.K. *Shift operators contained in contractions, Schur parameters and pseudocontinuable Schur functions.* Pp. 175–250 in: *Interpolation, Schur Functions and Moment Problems.*
Alpay, D. and Gohberg, I. – eds., Operator Theory: Advances and Applications, **165**. Birkhäuser Verlag, Basel 2006. xi+302pp.

[DFK] Dubovoy, V.K., Fritzsche, B., Kirstein, B.
Matricial Version of the Classical Schur Problem, Teubner Texte zur Mathematik, Band 129, B. 6. Teubner Verlagsgesellschaft, Stuttgart-Leipzig, 1992.

[DuHa] Duffin, R.J., Hazony D. *The degree of a rational matrix-function.* Journ. of Soc. for Industr. Appl. Math.**11**:3 (1963), pp. 645–658.

[Ger] Геронимус, Я.Л. *О полиномах, ортогональных на круге, о тригонометрической проблеме моментов и об ассоцированных с нею функциях типа Carathéodory и Schur'a.* (In Russian.) Матем. Сборник, **15**(57):11 (1944), 99–130.

[Gil] Gilbert, E.G. *Controllability and observability in multivariate systems.* Journ. of SIAM, Ser. **A**: Control. Vol. **1** (1962–1963), 128–151.

[GolV] Golub, G.H., van Loan, C.F. *Matrix Computations.* 2nd edition. John Hopkins University Press. Baltimore 1989.

[Grg] Gragg, W.B. *Positive definite Toeplitz matrices, the Arnoldi process for isometric operators, and Gaussian quadrature on the unit circle.* Journ. of Comput. and Appl. Math., **46** (1993), pp. 183–198.

[He1] Helton, J.W. *The characteristic function of operator theory and electrical network realization.* Indiana Univ. Math. Journ., **22**:5, (1972), 403–414.

[He2] Helton, J.W. *Discrete time systems, operator models, and scattering theory.* Indiana Journal of Funct. Anal., **16**, (1974), 15–38.

[He3] Helton, J.W. *Systems with infinite-dimensional state space: the Hilbert space approach.* Proc. IEEE, **64**:1, (1976), 145–160.

[HeBa] Helton, J.W., Ball, J.A. *The cascade decomposition of a given system vs. the linear fractional decompositions of its transfer function.* Integral Equations and Operator Theory, **5** (1982), pp. 341–385.

[Hou] Householder, A.S. *The Theory of Matrices in Numerical Analysis.* Blasdell Publishing, New York·Toronto·London 1964. xi+257 pp. Reprint: Dover Publications, Inc., New York, 1974. x+274 pp.

[Fuh] Fuhrmann, P.A. *Linear Systems and Operators in Hilbert Space.* McGraw Hill, 1981, x+325 pp.

[Kaa] Kaashoek, M.A. *Minimal factorization, linear systems and integral operators.* Pp. 41–86 in: Operators and function theory (Lancaster, 1984), (edited by S.C. Power). NATO Adv. Sci. Inst. Ser. C Math. Phys. Sci., **153**, Reidel, Dordrecht, 1985.

[Kai] Kailath, T. *A theorem of I. Schur and its impact on modern signal processing.* In [S:Meth], pp. 9–30.

[KFA] Kalman, R.E., Falb, P.L., Arbib, M.A. *Topics in mathematical system theory*. McGraw-Hill, New York-Toronto-London, 1969. xiv+358.

[Kal1] Kalman, R.E. *On the general theory of control systems.* In: [ARC], Vol. 1, pp. 481–492.

[Kal2] Kalman, R.E. *Canonical structure of linear dynamical system.* Proc. Nat. Acad. Sci. USA, Vol. **48**, No. 4 (1962), pp. 596–600.

[Kal3] Kalman, R.E. *Mathematical description of linear dynamical systems.* Journ. of Soc. for Industr. Appl. Math., ser. **A**: Control, Vol. **1**, No. 2 (1962–1963), pp. 152–192.

[Kal4] Kalman, R.E. *Irreducible realizations and the degree of a rational matrix.* Journ. of Soc. for Industr. Appl. Math., **13**:2 (1965), pp. 520–544.

[Ka] Katsnelson, V.E. *Right and left joint system representation of a rational matrix-function in general position* (*system representation theory for dummies*). Pp. 337–400 in: *Operator theory, system theory and related topics* (*Beer-Sheva/Rehovot,* 1997). Oper. Theory Adv. Appl., **OT 123**, Birkhäuser, Basel, 2001.

[KaVo1] Katsnelson, V., Volok, D. *Rational solutions of the Schlesinger system and isoprincipal deformations of rational matrix-functions. II.* pp. 165–203 in: *Operator theory, systems theory and scattering theory: multidimensional generalizations.* Oper. Theory Adv. Appl., **OT 157**, Birkhäuser, Basel, 2005.

[KaVo2] Katsnelson, V., Volok, D. *Deformations of Fuchsian systems of linear differential equations and the Schlesinger system.* Math. Phys. Anal. Geom. **9** (2006), no. 2, 135–186.

[KiNe] Killip, R. and Nenciu, I. *Matrix models for circular ensembles.* Intern. Math. Research Notes, **50** (2004), 2665–2701.

[LaPhi] Lax, P. and R. Phillips. *Scattering Theory*. Academic Press, New York · London 1967.

[Liv1] Лившиц, М.С. *Об одном классе линейных операторов в гильбертовом пространстве.*
Математический Сборник, том **(19) (61)**:2 (1946), 239–260.
English transl.:
Livšic, M.S. *On a class of linear operators in Hilbert space.* Amer. Math. Soc. Transl. (**2**), **13** (1960), 61–82.

[Liv2] Лившиц, М.С. *Изометрические операторы с равными дефектными числами, квазиунитарные операторы.*
Математический Сборник, том **26 (68)**:1 (1950), 247–264.
English transl.:
Livšic, M.S. *Isometric operators with equal deficiency indices, quasiunitary operators.*
Amer. Math. Soc. Transl. (**2**), **13** (1960), 85–102.

[Liv3] Лившиц, М.С. *О спектральном разложении линейных несамосопряженных операторов.* Математический Сборник, том **34**:1 (1954), 145–199. English transl.:
Livšic, M.S. *On the spectral resolution of linear non-selfadjoint operator.* Amer. Math. Soc. Transl. (**2**), **5** (1957), 67–114.

[Liv4] Лившиц, М.С. *О применении теории несамосопряженных операторов в теории рассеяния.* ЖЭТФ, том **31**:1 (1956,), 121–131. English transl.:
Livshitz, M.S. *The application of non-self-adjoint operators to scattering theory.* Soviet Physics JETP **4**:1 (1957), 91–98.

[Liv5] Лившиц, М.С. *Метод несамосопряженных операторов в теории рассеяния.* Успехи Мат. Наук, том **12**:1 (1957), 212–218. English transl.:
Livšic, M.S. *The method of non-selfadjoint operators in dispersion theory.* Amer. Math. Soc. Transl. (**2**), **16** (1960), 427–434.

[Liv6] Лившиц, М.С. *Метод несамосопряженных операторов в теории волноводов.* Радиотехника и Электроника, том **7** (1962), 281–297. English transl.:
Livšic, M.S. *The method of non-selfadjoint operators in the theory of waveguides.* Radio Engineering and Electronic Physics, **7** (1962), 260–276.

[Liv7] Лившиц, М.С. *О линейных физических системах, соединённых с внешним миром каналами связи.* Известия АН СССР, сер. математическая, **27** (1963), 993–1030.

[Liv8] Лившиц, М.С. *Открытые системы как линейные автоматы.* Известия АН СССР, сер. математическая, **27** (1963), 1215–1228.

[Liv9] Лившиц, М.С. *Операторы, колебания, волны. Открытые Системы.* Наука. Москва 1966. English transl.:
Livshitz, M.S. *Operators, Oscillations, Waves. Open Systems.* (Transl. of Math. Monogr., **34**.) Amer. Math. Soc., Providence, RI, 1973. vi+274 pp.

[LiFl] Лившиц, М.С., Флексер, М.Ш. *Разложение реактивного четырехполюсника в цепочку простейших четырехполюсников.* доклады Акад. Наук СССР (кибернетика и теория регулирования), **135**:3 (1960, 542–544. English transl.:
Livshitz, M.S., Flekser, M.S. *Expansion of a reactive four-terminal network into a chain of simplest four-terminal networks.* Soviet Physics – Doklady (cybernetics and control theory), **135**:3 (1960), 1150–1152.

[LiYa] Лившиц, М.С., Янцевич, А.А. *Теория Операторных Узлов в Гильбертовом пространстве.* Изд-во Харьковского Унив-та, Харьков 1971.
English transl.:
Livshitz, M.S., Yantsevich, A.A. *Operator colligations in Hilbert Spaces.* Winston & Sons, Washington, D.C., 1979, x+212.

[McM] McMillan, B. *Introduction to formal realizability theory.* Bell Syst. Techn. Journ., textbf31 (1952), Part I – pp. 217–279. Part II – pp. 541–600.

[Nik] Nikol'skiĭ, N.K. (= Nikolski, N.K.) *Operators, functions, and systems: an easy reading.*
Vol. 1. *Hardy, Hankel, and Toeplitz.* Mathematical Surveys and Monographs, **92**. American Mathematical Society, Providence, RI, 2002. xiv+461 pp.
Vol. 2. *Model operators and systems.* Mathematical Surveys and Monographs, **93**. American Mathematical Society, Providence, RI, 2002. xiv+439 pp.

[Red1] Redheffer, R. *Remarks on the basis of network theory.* J. Math. and Phys. **28** (1949), 237–258.

[Red2] Redheffer, R. *Inequalities for a matrix Riccati equation.* J. Math. Mech. **8** (1959), pp. 349–367.

[Red3] Redheffer, R. *On a certain linear fractional transformation.* J. Math. and Phys. **39** (1960), 269–286.

[Red4] Redheffer, R. *Difference equations and functional equations in transmission-line theory.* Pp. 282–337 in: *Modern mathematics for the engineer: Second series*, Beckenbach, E.F. – ed., McGraw-Hill, New York, 1961.

[Red5] Redheffer, R. *On the relation of transmission-line theory to scattering and transfer.* J. Math. and Phys. **41** (1962), 1–41.

[Sakh1] Сахнович, Л.А. *О факторизации передаточной оператор-функции.* доклады Акад. Наук СССР, **226**:4 (1976), 781–784.
English transl.:
Sahnovič, L.A.(=Sakhnovich, L.A.) *On the factorization of an operator-valued transfer function.* Soviet Math. – Doklady, **17**:1 (1976), 203–207.

[Sakh2] Сахнович, Л.А. *Задачи факторизации и операторные тождества.*, Успехи Матем. Наук, **41**:1 (1986), 3–55.
English transl.:
Sahnovič, L.A.(=Sakhnovich, L.A.) *Factorization problems and operator identities.* Russian Math. Surveys, **41**:1 (1986), pp. 1–64.

[Sakh3] Sahnovič, L.A.(=Sakhnovich, L.A.) *Spectral Theory of Canonical Differential Systems: Method of Operator Identities.*
Birkhäuser, Basel 1999, vi+202 pp.

[Sim] Simon, B. *Orthogonal polynomials on the unit circle.*
Part 1. *Classical theory.* American Mathematical Society Colloquium Publications, **54**: 1. American Mathematical Society, Providence, RI, 2005. xxvi+466 pp.
Part 2. *Spectral theory.* American Mathematical Society Colloquium Publications, **54**:2. American Mathematical Society, Providence, RI, 2005. pp. i–xxii and 467–1044.

[Sch] Schur, I. *Über Potenzreihen, die im Innern des Einheitskreises beschränkt sind, I* . (in German) J. reine und angewandte Math. **147**(1917), 205–232. Reprinted in: [Sch: Ges], Vol. II, pp. 137–164. Reprinted also in: [Ausg], pp. 22–49.
English translation: *On power series which are bounded in the interior of the unit circle. I.*, In: [S:Meth], pp. 31–59.

[S:Meth] *I. Schur Methods in Operator Theory and Signal Processing.*
Operator Theory: Advances and Applications. Vol. **18**.
I. Gohberg – editor. Birkhäuser, Basel·Boston·Stuttgart 1986.

[Sch: Ges] Schur, I.: *Gesammelte Abhandlungen* [*Collected Works*]. *Vol. II.* Springer-Verlag, Berlin· Heidelberg·New York, 1973.

[Str] Strang, G. *Linear Algebra and its Applications.* Academic Press, 1976.

[SzNFo] Sz.-Nagy, B. and C. Foias. *Analyse Harmonique des Opérateurs de l'espace de Hilbert* (French). Masson and Académiae Kiado, 1967. English transl.: *Harmonic Analysis of Operators in Hilbert Space.* North Holland, Amsterdam 1970.

[Tep] Тепляев, А.В. *Чисто точечный спектр случайных ортогональных на окружности многочленов.* доклады Акад. Наук СССР, **320**:1 (1991), 49–53. English Transl.:
Teplyaev, A.V. *The pure point spectrum of random polynomials orthogonal on the circle.* Sov. Math., Dokl. **44**:2 (1992), 407–411.

[Wil] Wilkinson, J.H. *The Algebraic Eigenvalue Problem.* Clarendon Press, Oxford, 1965.

Bernd Fritzsche and Bernd Kirstein
Mathematisches Institut
Universität Leipzig
D-04009, Leipzig, Germany
e-mail: bernd.fritzsche@mathematik.uni-leipzig.de
bernd.kirstein@mathematik.uni-leipzig.de

Victor Katsnelson
Department of Mathematics
the Weizmann Institute
Rehovot 76100, Israel
e-mail: victor.katsnelson@weizmann.ac.il
victorkatsnelson@gmail.com

Operator Theory:
Advances and Applications, Vol. 197, 251–279

On Some Interrelations Between J-Potapov Functions and J-Potapov Sequences

Bernd Fritzsche, Bernd Kirstein and Uwe Raabe

Dedicated to the memory of M.S. Livšic

Abstract. This paper deals with the J-Potapov class $\mathcal{P}_J(\mathbb{D})$ in the unit disk $\mathbb{D}$. Particular emphasis is laid on the subclass $\mathcal{P}_{J,0}(\mathbb{D})$ of all functions $f \in \mathcal{P}_J(\mathbb{D})$ which are holomorphic at the origin. A complete characterization of the sequences of Taylor coefficients of functions from $\mathcal{P}_{J,0}(\mathbb{D})$ is given. Moreover, the generalization of the matricial Schur problem for the class $\mathcal{P}_{J,0}(\mathbb{D})$ is treated. A complete description of the set of solutions is given in the nondegenerate and degenerate cases.

Mathematics Subject Classification (2000). Primary 30E05, 47A57.

Keywords. J-Potapov functions, J-Potapov sequences, J-inner sequences, J-inner functions, Taylor coefficient problem for J-Potapov functions.

1. Introduction

The central object of this paper is a distinguished class of meromorphic $m \times m$ matrix-valued functions in the open unit disk which originates in the fundamental paper [25] by V.P. Potapov. The investigations of V.P. Potapov were initiated by the studies of M.S. Livšic on characteristic functions of nonunitary (respectively, nonselfadjoint) operators (see [21], [22], [9], [24]). In the paper [23] M.S. Livšic sketches some impressions on his interactions with V.P. Potapov. The theory of characteristic functions is one of the cornerstones of the spectral theory of nonunitary and nonselfadjoint operators. It allows to apply the whole theory of bounded analytic functions to operator theoretic problems. Of equal and perhaps even greater importance, operator theory leads to new problems in function theory and provides new methods for a solution of old ones. The interplay between operator theory and complex function theory is one of the most significant features

The work of the third author of the present paper was supported by the EU project "Geometric Analysis on Lie Groups and Applications" (GALA).

of mathematics in the period since 1950. In the framework of these developments, the Potapov class of meromorphic matrix-valued functions plays an important role. This is caused by several reasons. After proving his famous factorization theorem for functions of this class in [25] V.P. Potapov applied these functions in collaboration with A.V. Efimov (see [15]) to electrical networks. Further important fields of application of the Potapov class are inverse scattering (see, e.g., Dewilde/Dym [10] and [11], Alpay/Dym [2] and [3], Alpay [1]) and Darlington synthesis (see Arov [4]). The role of the Potapov class in the context of matrix versions of classical interpolation and moment problems has been discussed in detail in several monographs (see Ball/Gohberg/Rodman [6], Dubovoj/Fritzsche/Kirstein [13], Dym [14], Katsnelson [19], Sakhnovich [28]) and the seminal paper Kovalishina [20].

The main theme of this paper is to characterize the subclass of the Potapov class which consists of all its members, which are holomorphic at $z = 0$, in terms of its Taylor coefficients. This topic is a part of a branch of geometric function theory which was started by the nowadays classical investigations by C. Carathéodory [5] and I. Schur [29] on the scalar holomorphic functions in the unit disk which are named after them now. The results due to Carathéodory and Schur were generalized to the matrix case via Schur analysis methods (see [12], [13], [16], [17], [7], [8]). Our approach is based on the interplay between the Potapov class and the matricial Schur class. This enables us to derive the desired results on the Potapov class from former results on the matricial Schur class (see [13], [17]). It would be also possible to realize a selfcontained treatment of the Potapov class by imitating the strategy used in [13], [16], [17], [7], [8]. This will be done somewhere else.

This paper is organized as follows. In Section 2, we state some facts on J-contractive matrices and some interrelations to contractive matrices. The main tool in establishing these interrelations is the Potapov-Ginzburg transform.

In Section 3, we discuss the J-Potapov class in the open unit disk and draw particular attention to some remarkable subclasses of it. Here we emphasize interrelations to the Schur class $\mathcal{S}_{m\times m}(\mathbb{D})$ and its corresponding subclasses.

In Section 4, we recall some interactions between the Schur class $\mathcal{S}_{m\times m}(\mathbb{D})$ and $m \times m$ Schur sequences.

In Section 5, we introduce the indefinite analogue of $m \times m$ Schur sequences. These are the J-Potapov sequences which will turn out to be one of the central objects of this paper.

Section 6 is devoted to a detailed analysis of the structure of the sequences of Taylor coefficients belonging to the class $\mathcal{P}_{J,0}(\mathbb{D})$ of all J-Potapov functions which are holomorphic at the origin. Theorem 6.2 provides a complete answer to this question. The corresponding result for the subclass $\underline{\mathcal{P}}_{J,0}(\mathbb{D})$ of all J-inner functions which are holomorphic at the origin is handled in Theorem 6.5.

The final Section 7 is devoted to a first study of the generalization of the matricial Schur problem to the class $\mathcal{P}_{J,0}(\mathbb{D})$. Using former results on the matricial Schur problem obtained in [13] and [18] the solvability of the interpolation problem for the J-Potapov class will be characterized in Theorem 7.2 whereas a complete description of the solution set will be stated in Theorem 7.4.

2. Some preliminaries on J-contractive matrices and the J-Potapov-Ginzburg transform

Throughout this paper, let m be a positive integer and let J be an $m\times m$ signature matrix, i.e., J is a complex $m\times m$ matrix such that $J^*=J$ and $J^2=I_m$ hold. Obviously, the matrices I_m and $-I_m$ are $m\times m$ signature matrices. Here I_m stands for the $m\times m$ identity matrix. A complex $m\times m$ matrix X is said to be *J-contractive* (respectively, *strictly J-contractive*) if the matrix $J-X^*JX$ is nonnegative Hermitian (respectively, positive Hermitian). A complex $m\times m$ matrix X is said to be *J-unitary* if $X^*JX=J$.

For our further considerations, some particular interrelations between J-contractive and contractive matrices turn out to be important. To explain this in more detail we need some basic facts on linear fractional transformations of matrices which are taken from Potapov [26] (see also [13, Section 1.6]). Let A and B be complex $2m\times 2m$ matrices and let

$$A=\begin{pmatrix} a & b\\ c & d\end{pmatrix} \qquad \text{and} \qquad B=\begin{pmatrix} \alpha & \gamma\\ \beta & \delta\end{pmatrix} \tag{2.1}$$

be the $m\times m$ block representations of A and B. If the set

$$\mathcal{Q}_{(c,d)}:=\big\{x\in\mathbb{C}^{m\times m}:\det(cx+d)\neq 0\big\}$$

is nonempty, then let $\mathcal{S}_A^{(m,m)}:\mathcal{Q}_{(c,d)}\to\mathbb{C}^{m\times m}$ be defined by

$$\mathcal{S}_A^{(m,m)}(x)=(ax+b)(cx+d)^{-1}. \tag{2.2}$$

If the set

$$\mathcal{R}_{\left(\begin{smallmatrix}\gamma\\ \delta\end{smallmatrix}\right)}:=\big\{x\in\mathbb{C}^{m\times m}:\det(x\gamma+\delta)\neq 0\big\}$$

is nonempty, then let $\mathcal{T}_B^{(m,m)}:\mathcal{R}_{\left(\begin{smallmatrix}\gamma\\ \delta\end{smallmatrix}\right)}\to\mathbb{C}^{m\times m}$ be defined by

$$\mathcal{T}_B^{(m,m)}(x)=(x\gamma+\delta)^{-1}(x\alpha+\beta). \tag{2.3}$$

Observe that $\mathcal{Q}_{(c,d)}\neq\emptyset$ if and only if $\operatorname{rank}(c,d)=m$. Moreover, $\mathcal{R}_{\left(\begin{smallmatrix}\gamma\\ \delta\end{smallmatrix}\right)}\neq\emptyset$ if and only if $\operatorname{rank}\begin{pmatrix}\gamma\\ \delta\end{pmatrix}=m$.

Let J be an $m\times m$ signature matrix and let

$$P_J:=\frac{1}{2}\left(I_m+J\right),\qquad Q_J:=\frac{1}{2}\left(I_m-J\right). \tag{2.4}$$

Then the matrices

$$\mathcal{A}_J=\begin{pmatrix} P_J & Q_J\\ Q_J & P_J\end{pmatrix} \qquad \text{and} \qquad \mathcal{B}_J=\begin{pmatrix} -P_J & Q_J\\ Q_J & -P_J\end{pmatrix} \tag{2.5}$$

are both nonsingular. For each $x\in\mathcal{Q}_{(Q_J,P_J)}$ the matrix $\mathcal{S}_{\mathcal{A}_J}^{(m,m)}(x)$ is called the *right J-Potapov-Ginzburg transform of x*. Furthermore, for each $x\in\mathcal{R}_{\left(\begin{smallmatrix}Q_J\\ -P_J\end{smallmatrix}\right)}$, the matrix $\mathcal{T}_{\mathcal{B}_J}^{(m,m)}(x)$ is said to be the *left J-Potapov-Ginzburg transform of x*.

Because of $\mathcal{A}_J^2 = I_{2m}$ and $\mathcal{B}_J^2 = I_{2m}$, [13, Proposition 1.6.2] implies that the mappings $\mathcal{S}_{\mathcal{A}_J}^{(m,m)}$ and $\mathcal{T}_{\mathcal{B}_J}^{(m,m)}$ are both injective and that

$$\left[\mathcal{S}_{\mathcal{A}_J}^{(m,m)}\right]^{-1} = \mathcal{S}_{\mathcal{A}_J}^{(m,m)}, \qquad \left[\mathcal{T}_{\mathcal{B}_J}^{(m,m)}\right]^{-1} = \mathcal{T}_{\mathcal{B}_J}^{(m,m)}. \tag{2.6}$$

Remark 2.1. Let J be an $m \times m$ signature matrix. Then a straightforward calculation yields that the matrices $\mathcal{A}_J$ and $\mathcal{B}_J$ given by (2.5) fulfill the identity

$$\mathcal{B}_J U_{m,m} \mathcal{A}_J = -U_{m,m}$$

where

$$U_{m,m} = \begin{pmatrix} 0 & I_m \\ -I_m & 0 \end{pmatrix}.$$

Hence a well-known result on linear fractional transformations (see [26] or Proposition 1.6.1 from [13]) provides us that

$$\mathcal{Q}_{(Q_J,P_J)} = \mathcal{R}_{\left(\begin{smallmatrix} Q_J \\ -P_J \end{smallmatrix}\right)} \qquad \text{and} \qquad \mathcal{S}_{\mathcal{A}_J}^{(m,m)} = \mathcal{T}_{\mathcal{B}_J}^{(m,m)}.$$

The following result can be verified by straightforward computation.

Lemma 2.2. *Let J be an $m \times m$ signature matrix, let $X \in \mathcal{Q}_{(Q_J,P_J)}$ and let $Y := \mathcal{S}_{\mathcal{A}_J}^{(m,m)}(X)$. Then* $\det(Q_J X + P_J) \neq 0$,

$$I_m - Y^* Y = (Q_J X + P_J)^{-*}(J - X^* J X)(Q_J X + P_J)^{-1}$$

and

$$J - Y^* J Y = (Q_J X + P_J)^{-*}(I_m - X^* X)(Q_J X + P_J)^{-1}.$$

Moreover, $\det(X Q_J - P_J) \neq 0$,

$$I_m - Y Y^* = (X Q_J - P_J)^{-1}(J - X J X^*)(X Q_J - P_J)^{-*}$$

and

$$J - Y J Y^* = (X Q_J - P_J)^{-1}(I_m - X X^*)(X Q_J - P_J)^{-*}.$$

Lemma 2.2 implies immediately the most of the following result which is taken from Potapov [27].

Proposition 2.3. *Let J be an $m \times m$ signature matrix. Then:*

(a) *Let X be a J-contractive complex $m \times m$ matrix. Then $X \in \mathcal{Q}_{(Q_J,P_J)}$ and the matrix $Y := \mathcal{S}_{\mathcal{A}_J}^{(m,m)}(X)$ is contractive. If X is strictly J-contractive (respectively, J-unitary), then the matrix Y is strictly contractive (respectively, unitary).*

(b) *Let $X \in \mathcal{Q}_{(Q_J,P_J)}$ and let $Y := \mathcal{S}_{\mathcal{A}_J}^{(m,m)}(X)$. If X is contractive (respectively, strictly contractive), then Y is J-contractive (respectively, strictly J-contractive). If X is unitary, then Y is J-unitary.*

Outgoing from Lemma 2.2 V.P. Potapov reproved in [27, Theorem 2.1] the following result, which he originally proved in [25, Ch. 2, Sect. 2, Theorem 7] using a more complicated alternate approach.

Proposition 2.4. *Let J be an $m \times m$ signature matrix. Let X be a J-contractive (respectively, strictly J-contractive) complex $m \times m$ matrix. Then X^* is a J-contractive (respectively, strictly J-contractive) complex $m \times m$ matrix. If X is J-unitary, then X^* is J-unitary.*

3. On the J-Potapov class in the open unit disk

Throughout this section, we again assume that m is a positive integer. If f is an $m \times m$ matrix-valued function, which is meromorphic in the open unit disk $\mathbb{D} := \{z \in \mathbb{C} : |z| < 1\}$, then let $\mathbb{H}_f$ be the set of all points at which f is holomorphic.

Definition 3.1. Let J be an $m \times m$ signature matrix and let f be a $\mathbb{C}^{m\times m}$-valued function which is meromorphic in the open unit disk $\mathbb{D}$. Then f is called a *J-Potapov function in $\mathbb{D}$* (respectively, a *strong J-Potapov function in $\mathbb{D}$*), if for each $w \in \mathbb{H}_f$ the matrix $f(w)$ is J-contractive (respectively, strictly J-contractive).

For each $m \times m$ signature matrix J, we will use the notation $\mathcal{P}_J(\mathbb{D})$ (respectively, $\mathcal{P}'_J(\mathbb{D})$) to denote the set of all J-Potapov functions in $\mathbb{D}$ (respectively, strong J-Potapov functions in $\mathbb{D}$). We will turn particular attention to a distinguished subclass of $\mathcal{P}_J(\mathbb{D})$, namely the class

$$\mathcal{P}_{J,0}(\mathbb{D}) := \{f \in \mathcal{P}_J(\mathbb{D}) : \ 0 \in \mathbb{H}_f\}.$$

From [13, Lemma 1.3.13] it can be seen that all classes $\mathcal{P}_J(\mathbb{D})$, $\mathcal{P}'_J(\mathbb{D})$, and $\mathcal{P}_{J,0}(\mathbb{D})$ are multiplicative. V.P. Potapov obtained in his landmark paper [25] a multiplicative decomposition of a function $f \in \mathcal{P}_J(\mathbb{D})$ into special factors belonging to $\mathcal{P}_J(\mathbb{D})$. Perhaps the simplest functions belonging to the class $\mathcal{P}_J(\mathbb{D})$ are particular rational $m \times m$ matrix-valued functions which have exactly one pole of order 1 in the extended complex plane. These functions will be discussed in the following two examples.

For $\alpha \in \mathbb{D}$ by b_α we denote the normalized elementary Blaschke factor associated with α, i.e., for $w \in \mathbb{C}\backslash\{\frac{1}{\overline{\alpha}}\}$ we have

$$b_\alpha(w) := \begin{cases} w, & \text{if } \alpha = 0 \\ \frac{|\alpha|}{\alpha}\frac{\alpha - w}{1 - \overline{\alpha}w}, & \text{if } \alpha \neq 0. \end{cases}$$

Example 3.2. Let J be an $m \times m$ signature matrix and let $\alpha \in \mathbb{D}$.

(a) Let P be a complex $m \times m$ matrix satisfying $P \neq 0_{m\times m}$, $JP \geq 0$, and $P^2 = P$. Let the function $B_{\alpha,P} : \mathbb{C}\backslash\{\frac{1}{\overline{\alpha}}\} \to \mathbb{C}^{m\times m}$ be defined by

$$B_{\alpha,P}(w) := I_m + [b_\alpha(w) - 1]\, P.$$

For $w \in \mathbb{C}\backslash\{\frac{1}{\overline{\alpha}}\}$, then the identity

$$J - [B_{\alpha,P}(w)]^*\, J\, [B_{\alpha,P}(w)] = \left(1 - |b_\alpha(w)|^2\right) JP$$

holds. Thus, the restriction of the function $B_{\alpha,P}$ onto $\mathbb{D}$ belongs to $\mathcal{P}_{J,0}(\mathbb{D})$. The function $B_{\alpha,P}$ is called *the Blaschke-Potapov J-elementary factor of first kind associated with α and P.*

(b) Let Q be a complex $m \times m$ matrix satisfying $Q \neq 0_{m\times m}$, $-JQ \geq 0$ and $Q^2 = Q$. Let the function $C_{\alpha,Q} : \mathbb{C}\backslash\{\alpha\} \to \mathbb{C}^{m\times m}$ be defined by

$$C_{\alpha,Q}(w) := I_m + [(1/b_\alpha(w)) - 1]\, Q.$$

For $w \in \mathbb{C}\backslash\{\alpha\}$, then the identity

$$J - [C_{\alpha,Q}(w)]^*\, J\, [C_{\alpha,Q}(w)] = \frac{1 - |b_\alpha(w)|^2}{|b_\alpha(w)|^2}(-JQ)$$

holds. Thus, the restriction of the function $C_{\alpha,Q}$ onto $\mathbb{D}$ belongs to $\mathcal{P}_J(\mathbb{D})$. In the case $\alpha \neq 0$ this restriction even belongs to $\mathcal{P}_{J,0}(\mathbb{D})$. The function $C_{\alpha,Q}$ is called *the Blaschke-Potapov J-elementary factor of second kind associated with α and Q.*

Observe that if $J = -I_m$ (respectively, $J = I_m$), then there is no matrix $P \in \mathbb{C}^{m\times m}\backslash\{0_{m\times m}\}$ (respectively, $Q \in \mathbb{C}^{m\times m}\backslash\{0_{m\times m}\}$) such that $JP \geq 0$ and $P^2 = P$ (respectively, $-JQ \geq 0$ and $Q^2 = Q$). Consequently, if $J = -I_m$ (respectively, $J = I_m$), then there is no Blaschke-Potapov J-elementary factor of first (respectively, second) kind.

Example 3.3. Let $\mathbb{T} := \{z \in \mathbb{C} : |z| = 1\}$. Let J be an $m \times m$ signature matrix and let $u \in \mathbb{T}$. Moreover, let R be a complex $m \times m$ matrix satisfying $R \neq 0_{m\times m}$, $JR \geq 0$, and $R^2 = 0_{m\times m}$. Let the function $D_{u,R} : \mathbb{C}\backslash\{u\} \to \mathbb{C}^{m\times m}$ be defined by

$$D_{u,R}(w) := I_m - \frac{u+w}{u-w}R.$$

For $w \in \mathbb{C}\backslash\{u\}$, then the identity

$$J - [D_{u,R}(w)]^*\, J\, [D_{u,R}(w)] = \frac{2(1-|w|^2)}{|u-w|^2}JR$$

holds. Thus, the restriction of the function $D_{u,R}$ onto $\mathbb{D}$ belongs to $\mathcal{P}_{J,0}(\mathbb{D})$. The function $D_{u,R}$ is called *the Blaschke-Potapov J-elementary factor of third kind associated with u and R.* Observe that in the cases $J = I_m$ and $J = -I_m$ there does not exist a complex $m \times m$ matrix $R \neq 0_{m\times m}$ satisfying $JR \geq 0$ and $R^2 = 0_{m\times m}$.

Example 3.3 leads us to some special principle of constructing functions belonging to $\mathcal{P}_{J,0}(\mathbb{D})$ by using functions of the Carathéodory class $\mathcal{C}(\mathbb{D})$ of all complex-valued functions holomorphic in $\mathbb{D}$ with nonnegative real part. Let R be a complex $m \times m$ matrix satisfying $JR \geq 0$ and $R^2 = 0_{m\times m}$. Moreover, let $h \in \mathcal{C}(\mathbb{D})$. Then the function $f : \mathbb{D} \to \mathbb{C}^{m\times m}$ defined by

$$f(w) := I_m - [h(w)]R$$

is holomorphic in $\mathbb{D}$ and for all $w \in \mathbb{D}$ the relation

$$J - [f(w)]^*\, J\, [f(w)] = 2\Re[h(w)]JR$$

is fulfilled. Here $\Re[h(w)]$ stands for the real part of $h(w)$. Thus, $f \in \mathcal{P}_{J,0}(\mathbb{D})$.

There are close interrelations between the Potapov class $\mathcal{P}_J(\mathbb{D})$ and the Schur class $\mathcal{S}_{m\times m}(\mathbb{D})$ of all $m \times m$ matrix-valued functions $f : \mathbb{D} \to \mathbb{C}^{m\times m}$ which are holomorphic in $\mathbb{D}$ and for which the matrix $f(w)$ is contractive for each $w \in \mathbb{D}$. The class $\mathcal{S}_{m\times m}(\mathbb{D})$ coincides with the I_m-Potapov class in $\mathbb{D}$. Indeed, the relation $\mathcal{S}_{m\times m}(\mathbb{D}) = \mathcal{P}_{I_m}(\mathbb{D})$ can be easily seen from a matrix version of Riemann's theorem on removable singularities.

Proposition 2.3 implies the following important interrelations between the Potapov class $\mathcal{P}_J(\mathbb{D})$ and the Schur class $\mathcal{S}_{m\times m}(\mathbb{D})$ on the one-hand side and between the strong J-Potapov class $\mathcal{P}'_J(\mathbb{D})$ and the strong Schur class $\mathcal{S}'_{m\times m}(\mathbb{D})$ of all $m \times m$ matrix-valued functions $f : \mathbb{D} \to \mathbb{C}^{m\times m}$ which are holomorphic in $\mathbb{D}$ and for which the matrix $f(w)$ is strictly contractive for each $w \in \mathbb{D}$ on the other side. Here for an $m \times m$ matrix X we will write $\det X$ for the determinant of X.

Proposition 3.4. *Let J be an $m \times m$ signature matrix and let the matrices P_J and Q_J be given by* (2.4)*. Then:*

(a) *If f belongs to $\mathcal{P}_J(\mathbb{D})$, then $\det(Q_J f(w) + P_J) \neq 0$ for each $w \in \mathbb{H}_f$ and the matrix-valued function*

$$g := (P_J f + Q_J)(Q_J f + P_J)^{-1} \tag{3.1}$$

belongs to the class $\mathcal{S}_{m\times m}(\mathbb{D})$. Moreover, $\mathbb{H}_f = \{w \in \mathbb{D} : \det(Q_J g(w)+P_J) \neq 0\}$ and

$$f = (P_J g + Q_J)(Q_J g + P_J)^{-1}. \tag{3.2}$$

(b) *If g is a function from $\mathcal{S}_{m\times m}(\mathbb{D})$ such that the function $\det(Q_J g + P_J)$ does not vanish identically, then*

$$f := (P_J g + Q_J)(Q_J g + P_J)^{-1} \tag{3.3}$$

belongs to $\mathcal{P}_J(\mathbb{D})$.

(c) *If g belongs to $\mathcal{S}_{m\times m}(\mathbb{D})$ and satisfies $\det(Q_J g(0)+P_J) \neq 0$, then the matrix-valued function f defined by* (3.3) *belongs to $\mathcal{P}_{J,0}(\mathbb{D})$.*

(d) *If f belongs to $\mathcal{P}'_J(\mathbb{D})$, then $\det(Q_J f(w) + P_J) \neq 0$ for each $w \in \mathbb{H}_f$ and the matrix-valued function g defined by* (3.1) *belongs to the class $\mathcal{S}'_{m\times m}(\mathbb{D})$.*

(e) *If g is a function belonging to $\mathcal{S}'_{m\times m}(\mathbb{D})$ such that the function $\det(Q_J g+P_J)$ does not vanish identically, then the matrix-valued function f defined by* (3.3) *belongs to the class $\mathcal{P}'_J(\mathbb{D})$.*

Corollary 3.5. *Let J be an $m\times m$ signature matrix. Let f be a function from $\mathcal{P}_J(\mathbb{D})$ for which there exists some point $w_0 \in \mathbb{H}_f$ such that the matrix $f(w_0)$ is strictly J-contractive. Then $f \in \mathcal{P}'_J(\mathbb{D})$.*

Proof. In view of part (a) of Proposition 3.4 the function g defined via (3.1) belongs to $\mathcal{S}_{m\times m}(\mathbb{D})$. Because of the choice of $f(w_0)$ we infer from part (a) of Proposition 2.3 that the matrix $g(w_0)$ is strictly contractive. Thus [13, Lemma 2.1.5] implies that g belongs to $\mathcal{S}'_{m\times m}(\mathbb{D})$. Taking into account (3.2) we see from part (e) of Proposition 3.4 that f belongs to $\mathcal{P}'_J(\mathbb{D})$. □

Corollary 3.6. *Let J be an $m \times m$ signature matrix and let $f \in \mathcal{P}_J(\mathbb{D})$. Then each entry function of f belongs to the Nevanlinna class $\mathcal{NM}(\mathbb{D})$ of meromorphic functions of bounded type in $\mathbb{D}$.*

Proof. The assertion is an immediate consequence of part (a) of Proposition 3.4. □

In view of Corollary 3.6, the functions belonging to the class $\mathcal{P}_J(\mathbb{D})$ have radial boundary values almost everywhere with respect to the normalized Lebesgue measure $\frac{1}{2\pi}\underline{\lambda}$ of the unit circle. This observation leads to the following important notion. If J is an $m \times m$ signature matrix and if $f \in \mathcal{P}_J(\mathbb{D})$, then the function f is called *J-inner*, if f has J-unitary radial boundary values $\frac{1}{2\pi}\underline{\lambda}$-almost everywhere. For each $m \times m$ signature matrix J, we will use the notation $\underline{\mathcal{P}}_J(\mathbb{D})$ to denote the class of all J-inner functions. From [13, Lemma 1.3.13] it can be seen that the class $\underline{\mathcal{P}}_J(\mathbb{D})$ is multiplicative

Remark 3.7. Let J be an $m \times m$ signature matrix and let $f \in \underline{\mathcal{P}}_J(\mathbb{D})$. Since every J-unitary matrix is nonsingular the function $\det f$ does not vanish identically.

A closer look at the rational $m \times m$ matrix-valued functions studied in Example 3.2 and Example 3.3 shows that their restrictions onto the unit disk $\mathbb{D}$ are J-inner functions.

We recall that a function $g \in \mathcal{S}_{m\times m}(\mathbb{D})$ is called *inner*, if g has unitary radial boundary values $\frac{1}{2\pi}\underline{\lambda}$-almost everywhere. We use the symbol $\underline{\mathcal{S}}_{m\times m}(\mathbb{D})$ to denote the subclass of all inner functions g belonging to $\mathcal{S}_{m\times m}(\mathbb{D})$. The following result complements Proposition 3.4. It is an immediate consequence of Lemma 2.2.

Proposition 3.8. *Let J be an $m \times m$ signature matrix and let the matrices P_J and Q_J be given by* (2.4)*. Then:*

(a) *If f belongs to $\underline{\mathcal{P}}_J(\mathbb{D})$, then $\det(Q_J f(w) + P_J) \neq 0$ for each $w \in \mathbb{H}_f$ and the function g defined via* (3.1) *belongs to the class $\underline{\mathcal{S}}_{m\times m}(\mathbb{D})$.*

(b) *If g is a function from $\underline{\mathcal{S}}_{m\times m}(\mathbb{D})$ such that the function $\det(Q_J g + P_J)$ does not vanish identically, then the function f defined via* (3.3) *belongs to $\underline{\mathcal{P}}_J(\mathbb{D})$.*

Let $\mathcal{M}$ be a nonempty subset of $\mathbb{C}$ and let $f : \mathcal{M} \to \mathbb{C}^{m\times m}$. Then we set $\mathcal{M}^\vee := \{\overline{z} : z \in \mathcal{M}\}$ and define the mapping $f^\vee : \mathcal{M}^\vee \to \mathbb{C}^{m\times m}$ by $f^\vee(z) = [f(\overline{z})]^*$. Obviously, $(f^\vee)^\vee = f$.

Remark 3.9. Let J be an $m \times m$ signature matrix. If f belongs to $\mathcal{P}_J(\mathbb{D})$(respectively, $\underline{\mathcal{P}}_J(\mathbb{D})$), then Proposition 2.4 shows that the function $f^\vee$ belongs to $\mathcal{P}_J(\mathbb{D})$(respectively, $\underline{\mathcal{P}}_J(\mathbb{D})$). If $f \in \mathcal{P}_{J,0}(\mathbb{D})$, then $f^\vee \in \mathcal{P}_{J,0}(\mathbb{D})$.

It is interesting to study the image of further important subclasses of the Schur class $\mathcal{S}_{m\times m}(\mathbb{D})$ under the Potapov-Ginzburg transformation. For this reason, we consider the class $\mathcal{S}''_{m\times m}(\mathbb{D})$ of all $g \in \mathcal{S}_{m\times m}(\mathbb{D})$ satisfying $\|g\|_\infty < 1$. Let $\mathcal{P}''_J(\mathbb{D})$ be the set of all $f \in \mathcal{P}_J(\mathbb{D})$ for which the function g defined via (3.1) belongs to $\mathcal{S}''_{m\times m}(\mathbb{D})$.

Proposition 3.10. *Let J be an $m \times m$ signature matrix and let $f \in \mathcal{P}''_J(\mathbb{D})$. Then there exists some positive real constant α such that for all $w \in \mathbb{H}_f$ the inequality*

$$J - f^*(w)Jf(w) \geq \alpha I_m$$

is satisfied.

Proof. Let g be defined via (3.1). In view of $f \in \mathcal{P}''_J(\mathbb{D})$ we have $g \in \mathcal{S}''_{m\times m}(\mathbb{D})$. Hence, there exists a positive constant β such that for each $w \in \mathbb{D}$ the inequality

$$I_m - g^*(w)g(w) \geq \beta I_m$$

is satisfied. Hence, for each $w \in \mathbb{D}$ we have $\det[I_m - g^*(w)g(w)] \neq 0$ and

$$[I_m - g^*(w)g(w)]^{-1} \leq \frac{1}{\beta} I_m. \tag{3.4}$$

In view of (3.2) we obtain from Lemma 2.2 for each $w \in \mathbb{H}_f$ the relations

$$\det(Q_J g(w) + P_J) \neq 0$$

and

$$J - f^*(w)Jf(w) = (Q_J g(w) + P_J)^{-*}[I_m - g^*(w)g(w)](Q_J g(w) + P_J)^{-1}.$$

Thus, for $w \in \mathbb{H}_f$, we get

$$\det(J - f^*(w)Jf(w)) \neq 0$$

and

$$[J - f^*(w)Jf(w)]^{-1} = (Q_J g(w) + P_J)[I_m - g^*(w)g(w)]^{-1}(Q_J g(w) + P_J)^*. \tag{3.5}$$

In view of $g \in \mathcal{S}''_{m\times m}(\mathbb{D})$ the function $Q_J g + P_J$ is bounded. Combining this with (3.4) and (3.5) we see that there exists some $\gamma \in (0, \infty)$ such that for each $w \in \mathbb{H}_f$ the inequality

$$[J - f^*(w)Jf(w)]^{-1} \leq \gamma I_m$$

is satisfied. Choosing $\alpha := \frac{1}{\gamma}$ we obtain the assertion. □

4. On some interrelations between $m \times m$ Schur functions and $m \times m$ Schur sequences

One of the main goals of this paper is to characterize the Taylor coefficient sequences of the functions belonging to the class $\mathcal{P}_{J,0}(\mathbb{D})$. For this purpose, we will mainly use former results on the structure of the Taylor coefficient sequences of functions belonging to $\mathcal{S}_{m\times m}(\mathbb{D})$ (see [13], [17], [7], and [8]). To recall these results we need some preparation.

In the following, we will use $\mathbb{N}_0$ (respectively, $\mathbb{N}$) to denote the set of all nonnegative (respectively, positive) integers. If $\kappa \in \mathbb{N}_0 \cup \{\infty\}$, then let $\mathbb{N}_{0,\kappa}$ be the set of all integers k which satisfy $0 \leq k \leq \kappa$.

For each $n \in \mathbb{N}_0$ and each sequence $(B_j)_{j=0}^n$ of complex $m \times m$ matrices, we will use the notation $S_n^{(B)}$ for the block Toeplitz matrix

$$S_n^{(B)} := \begin{pmatrix} B_0 & 0_{m\times m} & \dots & 0_{m\times m} \\ B_1 & B_0 & \dots & 0_{m\times m} \\ \vdots & \vdots & & \vdots \\ B_n & B_{n-1} & \dots & B_0 \end{pmatrix}. \tag{4.1}$$

For short, if there will not arise misunderstandings, then we will also write S_n instead of $S_n^{(B)}$.

If $n \in \mathbb{N}_0$, then a sequence $(B_j)_{j=0}^n$ of complex $m \times m$ matrices is called an $m \times m$ *Schur sequence* (respectively, a *strict* $m \times m$ *Schur sequence*) if the matrix S_n is contractive (respectively, strictly contractive). A sequence $(B_j)_{j=0}^\infty$ of complex $m \times m$ matrices is said to be an $m \times m$ *Schur sequence* (respectively, a *strict Schur sequence*) if for each $n \in \mathbb{N}_0$, the sequence $(B_j)_{j=0}^n$ is an $m \times m$ Schur sequence (respectively, a strict $m \times m$ Schur sequence).

If g is a matrix-valued function which is holomorphic at $w = 0$, then we consider the Taylor series representation

$$g(w) = \sum_{j=0}^{\infty} B_j w^j \tag{4.2}$$

for each w belonging to some neighborhood of 0 and the sequence $(B_j)_{j=0}^\infty$ of complex $m \times m$-matrices will be shortly called the sequence of Taylor coefficients of g.

The following result is taken from [13, Theorem 5.1.1]. The scalar case is due to I. Schur [29].

Theorem 4.1. *Let $m \in \mathbb{N}$. Then:*

(a) *If $g \in \mathcal{S}_{m\times m}(\mathbb{D})$, then the sequence $(B_j)_{j=0}^\infty$ of Taylor coefficients of g is an $m \times m$ Schur sequence.*
(b) *If $(B_j)_{j=0}^\infty$ is an $m \times m$ Schur sequence, then there is a unique $g \in \mathcal{S}_{m\times m}(\mathbb{D})$ such that $(B_j)_{j=0}^\infty$ is the sequence of Taylor coefficients of g.*

On the basis of Theorem 4.1 we characterize now the Taylor coefficient sequences of functions belonging to the class $\underline{\mathcal{S}}_{m\times m}(\mathbb{D})$. For this reason, we introduce the following notion.

Definition 4.2. Let $(B_j)_{j=0}^\infty$ be a sequence of complex $m\times m$ matrices. Then $(B_j)_{j=0}^\infty$ is called an *inner sequence*, if for all $k \in \mathbb{N}_0$ the identity

$$\sum_{j=0}^{\infty} B_{j+k}^* B_j = \begin{cases} I_m, & \text{if } k = 0 \\ 0_{m\times m}, & \text{if } k \in \mathbb{N}. \end{cases}$$

holds true.

Denote $H^2(\mathbb{D})$ the usual Hardy space of all holomorphic functions h in the unit disk which satisfy

$$\sup_{r\in[0,1)} \frac{1}{2\pi}\int_{\mathbb{T}} |h(rz)|^2 \underline{\lambda}(dz) < \infty.$$

Furthermore, let $[H^2(\mathbb{D})]^{m\times m}$ be the set of all $m \times m$ matrix-valued functions defined in $\mathbb{D}$ each entry function of which belongs to $H^2(\mathbb{D})$. Let $g \in [H^2(\mathbb{D})]^{m\times m}$ and denote by $\underline{g}$ the radial boundary function of g. Moreover, let $(B_j)_{j=0}^{\infty}$ be the Taylor coefficient sequence of g. Then the matrix version of the Parseval equality provides for each $k \in \mathbb{N}_0$ the identity

$$\frac{1}{2\pi}\int_{\mathbb{T}} z^{-k}[\underline{g}(z)]^*[\underline{g}(z)]\underline{\lambda}(dz) = \sum_{j=0}^{\infty} B_{j+k}^* B_j \tag{4.3}$$

(see, e.g., [30, Theorem 3.9, part (c)]).

Proposition 4.3. *Let $m \in \mathbb{N}$. Then:*

(a) *Let $g \in \underline{\mathcal{S}}_{m\times m}(\mathbb{D})$. Then the sequence $(B_j)_{j=0}^{\infty}$ of Taylor coefficients of g is an $m \times m$ inner sequence.*

(b) *Let $(B_j)_{j=0}^{\infty}$ be an $m \times m$ inner sequence. Then:*

(b1) *There is a unique $g \in \underline{\mathcal{S}}_{m\times m}(\mathbb{D})$ such that $(B_j)_{j=0}^{\infty}$ is the sequence of Taylor coefficients of g.*

(b2) *The sequence $(B_j)_{j=0}^{\infty}$ is an $m \times m$ Schur sequence.*

Proof. (a) This follows immediately from (4.3).

(b1) In view of the choice of the sequence $(B_j)_{j=0}^{\infty}$ the matrix version of the Riesz-Fischer theorem implies that there is a unique function $g \in [H^2(\mathbb{D})]^{m\times m}$ with Taylor series representation (4.2) for each $w \in \mathbb{D}$. Then taking into account (4.3) and the choice of the sequence $(B_j)_{j=0}^{\infty}$, we obtain that the radial boundary function $\underline{g}$ of g fulfills for each $k \in \mathbb{N}_0$ the equation

$$\frac{1}{2\pi}\int_{\mathbb{T}} z^{-k}[\underline{g}(z)]^*[\underline{g}(z)]\underline{\lambda}(dz) = \begin{cases} I_m, & \text{if } k = 0 \\ 0_{m\times m}, & \text{if } k \in \mathbb{N}. \end{cases}$$

Thus, the unicity theorem of Fourier analysis on the unit circle yields that the function $\underline{g}$ has unitary values $\frac{1}{2\pi}\underline{\lambda}$-almost everywhere. Hence, in view of $g \in [H^2(\mathbb{D})]^{m\times m}$, we obtain that $g \in \underline{\mathcal{S}}_{m\times m}(\mathbb{D})$.

(b2) Combine (b1) with part (a) of Theorem 4.1. □

5. On J-Potapov sequences

Let J be an $m\times m$ signature matrix. To generalize Theorem 4.1 to the class $\mathcal{P}_{J,0}(\mathbb{D})$ we have to look for an appropriate generalization of the notion of an $m \times m$ Schur sequence. Obviously, for each $n \in \mathbb{N}_0$, the complex $(n+1)m \times (n+1)m$ matrix

$$J_{[n]} := \operatorname{diag}(J, \ldots, J) \tag{5.1}$$

is an $(n+1)m \times (n+1)m$ signature matrix. Observe that $J_{[n]}$ can be written in terms of Kronecker products as

$$J_{[n]} = I_{n+1} \otimes J. \tag{5.2}$$

Definition 5.1. If $n \in \mathbb{N}_0$, then a sequence $(A_j)_{j=0}^n$ of complex $m \times m$ matrices is called a *J-Potapov sequence* (respectively, a *strict J-Potapov sequence*) if the matrix $S_n^{(A)}$ is $J_{[n]}$-contractive (respectively, strictly $J_{[n]}$-contractive). For each $n \in \mathbb{N}_0$, let $\mathcal{P}_{J,n}^{\leq}$ (respectively, $\mathcal{P}_{J,n}^{<}$) be the set of all J-Potapov sequences (respectively, strict J-Potapov sequences) $(A_j)_{j=0}^n$.

For each $n \in \mathbb{N}_0$ and each sequence $(A_j)_{j=0}^n$ of complex $m \times m$ matrices, let

$$Q_{n,J} := J_{[n]} - S_n^* J_{[n]} S_n, \tag{5.3}$$

where $S_n := S_n^{(A)}$. Then one can easily see that in the case $n \geq 1$ the matrix $Q_{n,J}$ admits the block representation

$$Q_{n,J} = \begin{pmatrix} J - A_0^* J A_0 - y_n^* J_{[n-1]} y_n & -y_n^* J_{[n-1]} S_{n-1} \\ -S_{n-1}^* J_{[n-1]} y_n & Q_{n-1,J} \end{pmatrix}, \tag{5.4}$$

where $y_n := (A_1^*, A_2^*, \ldots, A_n^*)^*$. Thus one can easily see that if $(A_j)_{j=0}^n \in \mathcal{P}_{J,n}^{\leq}$ (respectively, $(A_j)_{j=0}^n \in \mathcal{P}_{J,n}^{<}$), then $(A_j)_{j=0}^k \in \mathcal{P}_{J,k}^{\leq}$ (respectively, $(A_j)_{j=0}^k \in \mathcal{P}_{J,k}^{<}$) for each $k \in \mathbb{N}_{0,n}$.

Definition 5.2. A sequence $(A_j)_{j=0}^\infty$ of complex $m \times m$ matrices is said to be a *J-Potapov sequence* (respectively, a *strict J-Potapov sequence*) if for each $n \in \mathbb{N}_0$ the sequence $(A_j)_{j=0}^n$ is a J-Potapov sequence (respectively, a strict J-Potapov sequence). Let $\mathcal{P}_{J,\infty}^{\leq}$ be the set of all J-Potapov sequences $(A_j)_{j=0}^\infty$ and let $\mathcal{P}_{J,\infty}^{<}$ be the set of all strict J-Potapov sequences $(A_j)_{j=0}^\infty$.

Obviously, for each $\kappa \in \mathbb{N}_0 \cup \{\infty\}$, the set $\mathcal{P}_{I_m,\kappa}^{\leq}$ coincides with the set of all $m \times m$ Schur sequences $(A_j)_{j=0}^\kappa$ and $\mathcal{P}_{I_m,\kappa}^{<}$ is exactly the set of strict $m \times m$ Schur sequences $(A_j)_{j=0}^\kappa$.

Lemma 5.3. *Let J be an $m \times m$ signature matrix and let $\kappa \in \mathbb{N}_0 \cup \{\infty\}$. If $(A_j)_{j=0}^\kappa$ is a J-Potapov sequence (respectively, a strict J-Potapov sequence), then $(A_j^*)_{j=0}^\kappa$ is a J-Potapov sequence (respectively, a strict J-Potapov sequence).*

Proof. Let $n \in \{0, \ldots, \kappa\}$ and let

$$P_{n,J}^{(A^*)} := J_{[n]} - S_n^{(A^*)} J_{[n]} [S_n^{(A^*)}]^*$$

where the matrix $S_n^{(A^*)}$ is defined via (4.1) using the sequence $(A_j^*)_{j=0}^n$ instead of $(B_j)_{j=0}^n$. Then a straightforward computation shows the identity

$$P_{n,J}^{(A^*)} = J_{[n,m]} Q_{n,J} J_{[n,m]} \tag{5.5}$$

where

$$J_{[n,m]} := \begin{cases} I_m, & \text{if } n = 0 \\ \begin{pmatrix} 0_{m\times m} & \dots & I_m \\ \vdots & \iddots & \vdots \\ I_m & \dots & 0_{m\times m} \end{pmatrix}, & \text{if } n \neq 0 \end{cases} \tag{5.6}$$

and where $Q_{n,J}$ is defined via (5.3). Now from (5.5), (5.6) and Proposition 2.4 both assertions can be seen immediately. □

In order to describe some important property of J-Potapov sequences we need some notation. If $\kappa \in \mathbb{N}_0 \cup \{\infty\}$ and if $(A_j)_{j=0}^{\kappa}$ and $(B_j)_{j=0}^{\kappa}$ are sequences of complex $m \times m$ matrices, then the sequence $(C_j)_{j=0}^{\kappa}$ given by $C_j := \sum_{k=0}^{j} A_k B_{j-k}$ for each $j \in \mathbb{N}_{0,\kappa}$ is called the Cauchy product of $(A_j)_{j=0}^{\kappa}$ and $(B_j)_{j=0}^{\kappa}$.

It is obvious that the sequence $(C_j)_{j=0}^{\infty}$ is the Cauchy product of $(A_j)_{j=0}^{\infty}$ and $(B_j)_{j=0}^{\infty}$ if and only if for each $\kappa \in \mathbb{N}_0$ the sequence $(C_j)_{j=0}^{\kappa}$ is the Cauchy product of $(A_j)_{j=0}^{\kappa}$ and $(B_j)_{j=0}^{\kappa}$.

Remark 5.4. Let $n \in \mathbb{N}_0$. Let $(A_j)_{j=0}^{n}$, $(B_j)_{j=0}^{n}$, and $(C_j)_{j=0}^{n}$ be sequences of complex $m \times m$ matrices. Then it is readily checked that the following statements are equivalent:

(i) $(C_j)_{j=0}^{n}$ is the Cauchy product of $(A_j)_{j=0}^{n}$ and $(B_j)_{j=0}^{n}$.
(ii) $S_n^{(C)} = S_n^{(A)} S_n^{(B)}$.
(iii) For each $k \in \mathbb{N}_{0,n}$, $(C_j)_{j=0}^{k}$ is the Cauchy product of $(A_j)_{j=0}^{k}$ and $(B_j)_{j=0}^{k}$.

Lemma 5.5. *Let J be an $m\times m$ signature matrix and let $\kappa \in \mathbb{N}_0\cup\{\infty\}$. Moreover, let $(A_j)_{j=0}^{\kappa}$ and $(B_j)_{j=0}^{\kappa}$ be sequences of complex $m \times m$ matrices. Denote by $(C_j)_{j=0}^{\kappa}$ the Cauchy product of $(A_j)_{j=0}^{\kappa}$ and $(B_j)_{j=0}^{\kappa}$. If $(A_j)_{j=0}^{\kappa}$ and $(B_j)_{j=0}^{\kappa}$ are J-Potapov sequences, then $(C_j)_{j=0}^{\kappa}$ is a J-Potapov sequence as well. If $(A_j)_{j=0}^{\kappa}$ and $(B_j)_{j=0}^{\kappa}$ are J-Potapov sequences and, additionally, at least one of them is a strict J-Potapov sequence, then $(C_j)_{j=0}^{\kappa}$ is a strict J-Potapov sequence.*

Proof. Combine Remark 5.4 with Lemma 1.3.13 from [13]. □

Our next considerations are aimed at studying interrelations between J-Potapov sequences and $m \times m$ Schur sequences. For this reason, we introduce first some general construction for sequences of matrices.

Lemma 5.6. *Let $n \in \mathbb{N}_0$ and let $(A_j)_{j=0}^{n}$ be a sequence of complex $m \times m$ matrices with $\det A_0 \neq 0$. Then there is a unique sequence $(B_j)_{j=0}^{n}$ of complex $m \times m$ matrices such that $S_n^{(B)} = (S_n^{(A)})^{-1}$. This sequence is given by $B_0 := A_0^{-1}$ and recursively for each $k \in \mathbb{N}_{1,n}$ by*

$$B_k = -\sum_{j=0}^{k-1} A_0^{-1} A_{k-j} B_j.$$

Moreover, for each $k \in \mathbb{N}_{1,n}$, the matrix B_k can be represented by

$$B_k = -\sum_{j=1}^{k} B_{k-j} A_j A_0^{-1}.$$

If $(A_j)_{j=0}^n$ is a sequence of Hermitian matrices, then $(B_j)_{j=0}^n$ is a sequence of Hermitian matrices as well.

Proof. The case $n = 0$ is trivial. Let $n \geq 1$. For each $k \in \mathbb{N}_{1,n}$, we have the block representations

$$S_k^{(A)} = \begin{pmatrix} S_{k-1}^{(A)} & 0 \\ z_k^{(A)} & A_0 \end{pmatrix} \qquad \text{and} \qquad S_k^{(A)} = \begin{pmatrix} A_0 & 0 \\ y_k^{(A)} & S_{k-1}^{(A)} \end{pmatrix}, \tag{5.7}$$

where

$$z_k^{(A)} := (A_k, A_{k-1}, \ldots, A_1) \qquad \text{and} \qquad y_k^{(A)} := (A_1^*, A_2^*, \ldots, A_k^*)^*. \tag{5.8}$$

From (4.1) and $\det A_0 \neq 0$ we see that $\det S_l^{(A)} \neq 0$ for each $l \in \mathbb{N}_{0,n}$. Therefore we obtain for each $k \in \mathbb{N}_{1,n}$ the block representations

$$(S_k^{(A)})^{-1} = \begin{pmatrix} (S_{k-1}^{(A)})^{-1} & 0 \\ -A_0^{-1} z_k^{(A)} (S_{k-1}^{(A)})^{-1} & A_0^{-1} \end{pmatrix}$$

and

$$(S_k^{(A)})^{-1} = \begin{pmatrix} A_0^{-1} & 0 \\ -(S_{k-1}^{(A)})^{-1} y_k^{(A)} A_0^{-1} & (S_{k-1}^{(A)})^{-1} \end{pmatrix}$$

of $(S_k^{(A)})^{-1}$. Thus one gets inductively the assertion. We omit the details. □

Remark 5.7. Let $n \in \mathbb{N}_0$ and let $(A_j)_{j=0}^n$ be a sequence of complex $m \times m$ matrices with $\det A_0 \neq 0$. We will call the sequence $(B_j)_{j=0}^n$ described in Lemma 5.6 *the reciprocal sequence corresponding to* $(A_j)_{j=0}^n$ and we will write $(A_j^\sharp)_{j=0}^n$ for $(B_j)_{j=0}^n$. Observe that $det A_0^\sharp \neq 0$ and that $(A_j)_{j=0}^n$ is the reciprocal sequence corresponding to $(A_j^\sharp)_{j=0}^n$. Moreover, for each $k \in \mathbb{N}_{0,n}$, the sequence $(A_j^\sharp)_{j=0}^k$ is the reciprocal sequence of $(A_j)_{j=0}^k$.

Remark 5.8. Let $(A_j)_{j=0}^\infty$ be a sequence of complex $m \times m$ matrices with $\det A_0 \neq 0$. Then one can easily see that there is a unique sequence $(B_j)_{j=0}^\infty$ of complex $m \times m$ matrices such that for each $n \in \mathbb{N}_0$ the sequence $(B_j)_{j=0}^n$ is the reciprocal sequence corresponding to $(A_j)_{j=0}^n$. This sequence $(B_j)_{j=0}^\infty$ is said to be *the reciprocal sequence corresponding to* $(A_j)_{j=0}^\infty$ and we will write $(A_j^\sharp)_{j=0}^\infty$ instead of $(B_j)_{j=0}^\infty$. Moreover, it is easily checked that $det A_0^\sharp \neq 0$ and that $(A_j)_{j=0}^\infty$ is the reciprocal sequence corresponding to $(A_j^\sharp)_{j=0}^\infty$.

For $j \in \mathbb{N}_0$ we set

$$\delta_{j0} := \begin{cases} 1, & \text{if } j = 0 \\ 0, & \text{if } j \in \mathbb{N}. \end{cases}$$

The following notion plays a key role in our further considerations.

Definition 5.9. Let J be an $m \times m$ signature matrix, let P_J and Q_J be given by (2.4), and let $\kappa \in \mathbb{N}_0 \cup \{\infty\}$.

(a) Let $(A_j)_{j=0}^{\kappa}$ be a sequence of complex $m \times m$ matrices with $\det(Q_J A_0 + P_J) \neq 0$. For each $j \in \mathbb{N}_{0,\kappa}$ let

$$X_j := P_J A_j + \delta_{j0} Q_J \qquad \text{and} \qquad W_j := Q_J A_j + \delta_{j0} P_J. \tag{5.9}$$

Further, let $(W_j^{\sharp})_{j=0}^{\kappa}$ be the reciprocal sequence corresponding to $(W_j)_{j=0}^{\kappa}$. Then the Cauchy product $(B_j)_{j=0}^{\kappa}$ of $(X_j)_{j=0}^{\kappa}$ and $(W_j^{\sharp})_{j=0}^{\kappa}$ is called the *right J-Potapov–Ginzburg transform* (short: *right J-PG transform*) *of* $(A_j)_{j=0}^{\kappa}$.

(b) Let $(\tilde{A}_j)_{j=0}^{\kappa}$ be a sequence of complex $m \times m$ matrices with $\det(\tilde{A}_0 Q_J - P_J) \neq 0$. For each $j \in \mathbb{N}_{0,\kappa}$ let

$$\tilde{X}_j := -\tilde{A}_j P_J + \delta_{j0} Q_J \qquad \text{and} \qquad \tilde{W}_j := \tilde{A}_j Q_J - \delta_{j0} P_J.$$

Further, let $(\tilde{W}_j^{\sharp})_{j=0}^{\kappa}$ be the reciprocal sequence corresponding to $(\tilde{W}_j)_{j=0}^{\kappa}$. Then the Cauchy product $(\tilde{B}_j)_{j=0}^{\kappa}$ of $(\tilde{W}_j^{\sharp})_{j=0}^{\kappa}$ and $(\tilde{X}_j)_{j=0}^{\kappa}$ is said to be the *left J-Potapov–Ginzburg transform* (short: *left J-PG transform*) *of* $(\tilde{A}_j)_{j=0}^{\kappa}$.

Remark 5.10. Let J be an $m \times m$ signature matrix and let the matrices P_J and Q_J be given by (2.4). Furthermore, let $\kappa \in \mathbb{N}_0 \cup \{\infty\}$ and let $(A_j)_{j=0}^{\kappa}$ be a sequence of complex $m \times m$ matrices with $\det(Q_J A_0 + P_J) \neq 0$ (respectively, $\det(A_0 Q_J - P_J) \neq 0$). Then one can easily see that there is a unique sequence $(B_j)_{j=0}^{\kappa}$ of complex $m \times m$ matrices such that for each $n \in \mathbb{N}_{0,\kappa}$ the sequence $(B_j)_{j=0}^{n}$ is the right (respectively, left) J-PG transform of $(A_j)_{j=0}^{n}$. This sequence $(B_j)_{j=0}^{\kappa}$ is exactly the right (respectively, left) J-PG transform of $(A_j)_{j=0}^{\kappa}$.

Our following considerations are aimed at translating the notion of J-PG transform for sequences of $m \times m$ matrices into the language of linear fractional transformations of matrices (see Section 2).

Proposition 5.11. *Let J be an $m \times m$ signature matrix, let $\kappa \in \mathbb{N}_0 \cup \{\infty\}$, and let $(A_j)_{j=0}^{\kappa}$ be a sequence of complex $m \times m$ matrices such that* $\det(Q_J A_0 + P_J) \neq 0$. *Let $(B_j)_{j=0}^{\kappa}$ be the right J-PG transform of $(A_j)_{j=0}^{\kappa}$. Then:*

(a) *For each $n \in \mathbb{N}_{0,\kappa}$, the matrix $S_n^{(A)}$ belongs to $\mathcal{Q}_{\left(Q_{J_{[n]}}, P_{J_{[n]}}\right)}$ and*

$$\mathcal{S}_{\mathcal{A}_{J_{[n]}}}^{((n+1)m,(n+1)m)}(S_n^{(A)}) = S_n^{(B)}. \tag{5.10}$$

(b) *For each $n \in \mathbb{N}_{0,\kappa}$, the matrix $S_n^{(B)}$ belongs to $\mathcal{Q}_{\left(Q_{J_{[n]}}, P_{J_{[n]}}\right)}$ and*

$$\mathcal{S}_{\mathcal{A}_{J_{[n]}}}^{((n+1)m,(n+1)m)}(S_n^{(B)}) = S_n^{(A)}.$$

(c) *The matrix $Q_J B_0 + P_J$ is nonsingular and the sequence $(A_j)_{j=0}^{\kappa}$ is the right J-PG transform of $(B_j)_{j=0}^{\kappa}$.*

Proof. Let $n \in \mathbb{N}_{0,\kappa}$. In view of (5.2) and (2.4) we obtain

$$P_{J_{[n]}} = \frac{1}{2}\left(I + J_{[n]}\right) = I_{n+1} \otimes P_J$$

and analogously $Q_{J_{[n]}} = I_{n+1} \otimes Q_J$. For each $j \in \mathbb{N}_{0,\kappa}$, let the matrices W_j and X_j be given by (5.9). Then

$$S_n^{(W)} = (I_{n+1} \otimes Q_J)\, S_n^{(A)} + (I_{n+1} \otimes P_J) = Q_{J_{[n]}} S_n^{(A)} + P_{J_{[n]}} \tag{5.11}$$

and similarly

$$S_n^{(X)} = P_{J_{[n]}} S_n^{(A)} + Q_{J_{[n]}}. \tag{5.12}$$

In view of $\det W_0 = \det(Q_J A_0 + P_J) \neq 0$, we get

$$\det S_n^{(W)} \neq 0 \tag{5.13}$$

and the reciprocal sequence $(Y_j)_{j=0}^{\kappa}$ corresponding to $(W_j)_{j=0}^{\kappa}$ fulfills $\left(S_n^{(Y)}\right)^{-1} = S_n^{(W)}$. Since $(B_j)_{j=0}^{\kappa}$ is the right J-PG transform of $(A_j)_{j=0}^{\kappa}$ from Remark 5.4, Remark 5.8, and Remark 5.7 we have then

$$S_n^{(B)} = S_n^{(X)} S_n^{(Y)} = S_n^{(X)} \left(S_n^{(W)}\right)^{-1}.$$

Consequently, by virtue of (5.11) and (5.12) we can conclude

$$S_n^{(B)} = \left(P_{J_{[n]}} S_n^{(A)} + Q_{J_{[n]}}\right) \left(Q_{J_{[n]}} S_n^{(A)} + P_{J_{[n]}}\right)^{-1}. \tag{5.14}$$

From (5.11) and (5.13) we see that $S_n^{(A)}$ belongs to $\mathcal{Q}_{\left(Q_{J_{[n]}}, P_{J_{[n]}}\right)}$, whereas (5.14) yields that (5.10) is true. Thus part (a) is checked. Part (b) follows from using $\mathcal{A}^2_{J_{[n]}} = I$, (2.6), and part (a). Part (c) is a consequence of (a) and (b). □

Part (a) of Proposition 5.11 says that for each $n \in \mathbb{N}_{0,\kappa}$ the matrix $S_n^{(B)}$ is the right $J_{[n]}$-PG transform of the matrix $S_n^{(A)}$. This fact motivated us to choose the terminology introduced in Definition 5.9.

Remark 5.12. Let $\kappa \in \mathbb{N}_0 \cup \{\infty\}$ and let $(A_j)_{j=0}^{\kappa}$ be a sequence of complex $m \times m$ matrices. Considering the case $J = I_m$ one can easily see from Proposition 5.11 that $(A_j)_{j=0}^{\kappa}$ is the right I_m-PG transform of $(A_j)_{j=0}^{\kappa}$.

Proposition 5.13. *Let J be an $m \times m$ signature matrix, let $\kappa \in \mathbb{N}_0 \cup \{\infty\}$, and let $(A_j)_{j=0}^{\kappa}$ be a sequence of complex $m \times m$ matrices such that $\det(A_0 Q_J - P_J) \neq 0$. Let $(B_j)_{j=0}^{\kappa}$ be the left J-PG transform of $(A_j)_{j=0}^{\kappa}$. Then for each $n \in \mathbb{N}_{0,\kappa}$, the matrices $S_n^{(A)}$ and $S_n^{(B)}$ belong to $\mathcal{R}_{\begin{pmatrix} Q_{J_{[n]}} \\ -P_{J_{[n]}} \end{pmatrix}}$ and*

$$\mathcal{T}^{((n+1)m,(n+1)m)}_{\mathcal{B}_{J_{[n]}}}(S_n^{(A)}) = S_n^{(B)} \qquad \text{and} \qquad \mathcal{T}^{((n+1)m,(n+1)m)}_{\mathcal{B}_{J_{[n]}}}(S_n^{(B)}) = S_n^{(A)}.$$

Moreover, the matrix $B_0Q_J - P_J$ is nonsingular and the sequence $(A_j)_{j=0}^{\kappa}$ is the left J-PG transform of $(B_j)_{j=0}^{\kappa}$.

Proof. Proposition 5.13 can be proved analogously to Proposition 5.11. □

In order to prove that the notion of the right and the left J-PG transform of a sequence of complex $m \times m$ matrices coincide we note that in view of Remark 2.1, for every nonnegative integer n, we have the relations

$$\mathcal{Q}_{\left(Q_{J_{[n]}},P_{J_{[n]}}\right)} = \mathcal{R}_{\left(\begin{smallmatrix} Q_{J_{[n]}} \\ -P_{J_{[n]}} \end{smallmatrix}\right)} \qquad \text{and} \qquad \mathcal{S}_{\mathcal{A}_{J_{[n]}}}^{((n+1)m,(n+1)m)} = \mathcal{T}_{\mathcal{B}_{J_{[n]}}}^{((n+1)m,(n+1)m)}. \tag{5.15}$$

Proposition 5.14. *Let J be an $m \times m$ signature matrix, let $\kappa \in \mathbb{N}_0 \cup \{\infty\}$, and let $(A_j)_{j=0}^{\kappa}$ be a sequence of complex $m \times m$ matrices. Then:*

(a) *The matrix $Q_JA_0 + P_J$ is nonsingular if and only if the matrix $A_0Q_J - P_J$ is nonsingular.*

(b) *Let the matrix $Q_JA_0 + P_J$ be nonsingular. Then the right J-PG transform of $(A_j)_{j=0}^{\kappa}$ and the left J-PG transform of $(A_j)_{j=0}^{\kappa}$ coincide.*

Proof. Use Proposition 5.11 and 5.13 as well as (5.15). □

Taking into account Proposition 5.14, in the following we will shortly speak of the *J-PG transform* instead of the right J-PG transform.

Lemma 5.15. *Let J be an $m \times m$ signature matrix, let $\kappa \in \mathbb{N}_0 \cup \{\infty\}$, and let $(B_j)_{j=0}^{\kappa}$ be a sequence of complex $m \times m$ matrices such that $\det(Q_JB_0 + P_J) \neq 0$. Let $(A_j)_{j=0}^{\kappa}$ be the J-PG transform of $(B_j)_{j=0}^{\kappa}$. For each $n \in \mathbb{N}_{0,\kappa}$, then the matrices $Q_{J_{[n]}}S_n^{(B)} + P_{J_{[n]}}$ and $S_n^{(B)}Q_{J_{[n]}} - P_{J_{[n]}}$ are both nonsingular and*

$$\begin{aligned} &J_{[n]} - (S_n^{(A)})^* J_{[n]} S_n^{(A)} \\ &\quad = \left(Q_{J_{[n]}}S_n^{(B)} + P_{J_{[n]}}\right)^{-*} \left(I - (S_n^{(B)})^* S_n^{(B)}\right) \left(Q_{J_{[n]}}S_n^{(B)} + P_{J_{[n]}}\right)^{-1} \end{aligned} \tag{5.16}$$

and

$$\begin{aligned} &J_{[n]} - S_n^{(A)} J_{[n]} (S_n^{(A)})^* \\ &\quad = \left(S_n^{(B)}Q_{J_{[n]}} - P_{J_{[n]}}\right)^{-1} \left(I - S_n^{(B)}(S_n^{(B)})^*\right) \left(S_n^{(B)}Q_{J_{[n]}} - P_{J_{[n]}}\right)^{-*} \end{aligned} \tag{5.17}$$

hold.

Proof. Use Proposition 5.11 and Lemma 2.2. □

Now we are going to turn our attention to some interrelations between J-Potapov sequences and Schur sequences.

Proposition 5.16. *Let J be an $m \times m$ signature matrix, let $\kappa \in \mathbb{N}_0 \cup \{\infty\}$, and let $(A_j)_{j=0}^{\kappa}$ be a J-Potapov sequence (respectively, a strict J-Potapov sequence). Then $\det(Q_JA_0 + P_J) \neq 0$ and the J-PG transform $(B_j)_{j=0}^{\kappa}$ of $(A_j)_{j=0}^{\kappa}$ is an*

$m \times m$ Schur sequence (*respectively, a strict $m \times m$ Schur sequence*) *which fulfills* $\det(Q_J B_0 + P_J) \neq 0$. *Furthermore,* $(A_j)_{j=0}^{\kappa}$ *is the J-PG transform of* $(B_j)_{j=0}^{\kappa}$.

Proof. Apply Lemma 5.15. □

Proposition 5.17. *Let J be an $m \times m$ signature matrix, let $\kappa \in \mathbb{N}_0 \cup \{\infty\}$, and let $(B_j)_{j=0}^{\kappa}$ be an $m \times m$ Schur sequence* (*respectively, a strict $m \times m$ Schur sequence*) *such that* $\det(Q_J B_0 + P_J) \neq 0$. *Then the J-PG transform $(A_j)_{j=0}^{\kappa}$ of $(B_j)_{j=0}^{\kappa}$ is a J-Potapov sequence* (*respectively, a strict J-Potapov sequence*) *and $(B_j)_{j=0}^{\kappa}$ is the J-PG transform of $(A_j)_{j=0}^{\kappa}$.*

Proof. Apply Lemma 5.15. □

6. On the sequence of Taylor coefficients of functions from the class $\mathcal{P}_{J,0}(\mathbb{D})$

In this section, we will present a detailed analysis of the structure of the sequences of Taylor coefficients of functions belonging to $\mathcal{P}_{J,0}(\mathbb{D})$. In particular, we shall obtain a generalization of Theorem 4.1.

If g is an $m \times m$ matrix-valued function which is holomorphic at 0, then for each $n \in \mathbb{N}_0$, we will use the notation $S_n^{[g]}$ to denote the block Toeplitz matrix

$$S_n^{[g]} := S_n^{(B)}, \tag{6.1}$$

where $(B_j)_{j=0}^{n}$ is the sequence of the first $(n+1)$ Taylor coefficients of g and where $S_n^{(B)}$ is given by (4.1).

Remark 6.1. Let J be an $m \times m$ signature matrix and let the matrices P_J and Q_J be given by (2.4). Let f be a $\mathbb{C}^{m\times m}$-valued function which is holomorphic at $w = 0$ and which fulfills $\det(Q_J f(0) + P_J) \neq 0$. Then it is readily checked that the matrix-valued function g defined by (3.1) is holomorphic at $w = 0$ and that, for each nonnegative integer n, the matrix $Q_{J_{[n]}} S_n^{[f]} + P_{J_{[n]}}$ is nonsingular and that the matrix $S_n^{[g]}$ coincides with the $J_{[n]}$-Potapov–Ginzburg transform of $S_n^{[f]}$.

The following theorem is the first main result of this paper.

Theorem 6.2. *Let J be an $m \times m$ signature matrix. Then:*

(a) *If $f \in \mathcal{P}_{J,0}(\mathbb{D})$, then the sequence $(A_j)_{j=0}^{\infty}$ of Taylor coefficients of f is a J-Potapov sequence.*
(b) *If $(A_j)_{j=0}^{\infty}$ is a J-Potapov sequence, then there is a unique $f \in \mathcal{P}_{J,0}(\mathbb{D})$ such that $(A_j)_{j=0}^{\infty}$ is the sequence of Taylor coefficients of f.*

Proof. (a) Let $f \in \mathcal{P}_{J,0}(\mathbb{D})$. According to part (a) of Proposition 3.4, then the matrix $Q_J f(0) + P_J$ is nonsingular and g defined by (3.1) belongs to $\mathcal{S}_{m\times m}(\mathbb{D})$. Because of $f(0) = A_0$ we also have $\det(Q_J A_0 + P_J) \neq 0$ and hence $\det(Q_{J_{[n]}} S_n^{(A)} +$

$P_{J_{[n]}}) \neq 0$ for each $n \in \mathbb{N}_0$. Taking into account $S_n^{(A)} = S_n^{[f]}$ for each $n \in \mathbb{N}_0$, from Remark 6.1 we get

$$\left(P_{J_{[n]}} S_n^{(A)} + Q_{J_{[n]}}\right) \left(Q_{J_{[n]}} S_n^{(A)} + P_{J_{[n]}}\right)^{-1} = S_n^{[g]} \tag{6.2}$$

for each $n \in \mathbb{N}_0$. Let $(B_j)_{j=0}^{\infty}$ be the J-PG transform of $(A_j)_{j=0}^{\infty}$. Application of Proposition 5.11 yields then

$$S_n^{(B)} = \left(P_{J_{[n]}} S_n^{(A)} + Q_{J_{[n]}}\right) \left(Q_{J_{[n]}} S_n^{(A)} + P_{J_{[n]}}\right)^{-1} \tag{6.3}$$

for each $n \in \mathbb{N}_0$. Comparing (6.2) and (6.3) we obtain $S_n^{[g]} = S_n^{(B)}$ for each $n \in \mathbb{N}_0$. Thus $(B_j)_{j=0}^{\infty}$ is the sequence of Taylor coefficients of g. Since g belongs to the Schur class $\mathcal{S}_{m\times m}(\mathbb{D})$ from part (a) of Theorem 4.1 we know then that $(B_j)_{j=0}^{\infty}$ is an $m \times m$-Schur sequence. Part (c) of Proposition 5.11 shows that $(A_j)_{j=0}^{\infty}$ is the J-PG transform of $(B_j)_{j=0}^{\infty}$. Since $(B_j)_{j=0}^{\infty}$ is an $m \times m$-Schur sequence from Proposition 5.17 we get thus that $(A_j)_{j=0}^{\infty}$ is a J-Potapov sequence.

(b) Let $(A_j)_{j=0}^{\infty}$ be a J-Potapov sequence. Thus, from Proposition 5.16 we can conclude that $\det(Q_J A_0 + P_J) \neq 0$ and that the J-PG transform $(B_j)_{j=0}^{\infty}$ of $(A_j)_{j=0}^{\infty}$ is an $m \times m$-Schur sequence. From part (b) of Theorem 4.1 we obtain then that there is a $g \in \mathcal{S}_{m\times m}(\mathbb{D})$ such that $(B_j)_{j=0}^{\infty}$ is the sequence of Taylor coefficients of g. We have $S_n^{[g]} = S_n^{(B)}$ for each $n \in \mathbb{N}_0$. In particular $g(0) = B_0$. Hence part (c) of Proposition 5.11 provides us $\det(Q_J B_0 + P_J) \neq 0$. Consequently, $\det(Q_{J_{[n]}} S_n^{[g]} + P_{J_{[n]}}) \neq 0$ for each $n \in \mathbb{N}_0$ and $\det(Q_J g(0) + P_J) \neq 0$. Since g belongs to $\mathcal{S}_{m\times m}(\mathbb{D})$ thus from part (c) of Proposition 3.4 we get that f defined by (3.3) belongs to the class $\mathcal{P}_{J,0}(\mathbb{D})$. From Remark 6.1 we can conclude

$$\begin{aligned} S_n^{[f]} &= \left(P_{J_{[n]}} S_n^{[g]} + Q_{J_{[n]}}\right) \left(Q_{J_{[n]}} S_n^{[g]} + P_{J_{[n]}}\right)^{-1} \\ &= \left(P_{J_{[n]}} S_n^{(B)} + Q_{J_{[n]}}\right) \left(Q_{J_{[n]}} S_n^{(B)} + P_{J_{[n]}}\right)^{-1} \end{aligned} \tag{6.4}$$

for each $n \in \mathbb{N}_0$. On the other hand, using Proposition 5.11 we get that, for each $n \in \mathbb{N}_0$, the right-hand side of (6.4) coincides with $S_n^{(A)}$. Therefore $S_n^{(A)} = S_n^{[f]}$ for every nonnegative integer n. Consequently, $(A_j)_{j=0}^{\infty}$ is exactly the sequence of Taylor coefficients of f. □

Theorem 6.2 shows that remarkable subclasses of the class $\mathcal{P}_{J,0}(\mathbb{D})$ can be introduced on the basis of distinguished subclasses of J-Potapov sequences. Let $\kappa \in \mathbb{N}_0 \cup \{\infty\}$. Then we will use the symbol $\mathcal{P}_{J,0,\kappa}(\mathbb{D})$ to denote the set of all functions $f \in \mathcal{P}_{J,0}(\mathbb{D})$ with Taylor coefficient sequence $(A_j)_{j=0}^{\infty}$ for which the sequence $(A_j)_{j=0}^{\kappa}$ is a strict J-Potapov sequence. Moreover, we will use the notation $\mathcal{S}_{m\times m,\kappa}(\mathbb{D})$ to denote the set of all functions $g \in \mathcal{S}_{m\times m}(\mathbb{D})$ with Taylor coefficient sequence $(B_j)_{j=0}^{\infty}$ for which the sequence $(B_j)_{j=0}^{\kappa}$ is a strict $m \times m$ Schur sequence.

Proposition 6.3. *Let J be an $m \times m$ signature matrix and let $f \in \mathcal{P}_{J,0}(\mathbb{D})$. Let the function g be defined via* (3.1)*. If $\kappa \in \mathbb{N}_0 \cup \{\infty\}$, then $f \in \mathcal{P}_{J,0,\kappa}(\mathbb{D})$ if and only if $g \in \mathcal{S}_{m\times m,\kappa}(\mathbb{D})$.*

Proof. Let $(A_j)_{j=0}^{\infty}$ be the Taylor coefficient sequence of f and let $(B_j)_{j=0}^{\infty}$ be the J-PG transform of $(A_j)_{j=0}^{\infty}$. Then it was shown in the proof of part (a) of Theorem 6.2 that $(B_j)_{j=0}^{\infty}$ is the sequence of Taylor coefficients of g. Combining this with (5.16) we get the asserted equivalence. □

In connection with Proposition 6.3 it should be mentioned that Theorem 8.4 in [7] shows that for each $\kappa \in \mathbb{N}_0 \cup \{\infty\}$ the class $\mathcal{S}_{m\times m,\kappa}(\mathbb{D})$ coincides with the set of all functions $g \in \mathcal{S}_{m\times m}(\mathbb{D})$ for which the Schur-Potapov algorithm does not break down immediately after the κth step.

Let $\underline{\mathcal{P}}_{J,0}(\mathbb{D}) := \underline{\mathcal{P}}_J(\mathbb{D}) \cap \mathcal{P}_{J,0}(\mathbb{D})$. To characterize the sequence of Taylor coefficients of functions belonging to $\underline{\mathcal{P}}_{J,0}(\mathbb{D})$, we introduce the following notion.

Definition 6.4. Let J be an $m \times m$ signature matrix and let the matrices P_J and Q_J be given by (2.4). Then a sequence $(A_j)_{j=0}^{\infty}$ of complex $m\times m$ matrices is called a *J-inner sequence*, if $\det(Q_J A_0 + P_J) \neq 0$ and if the J-PG transform $(B_j)_{j=0}^{\infty}$ of $(A_j)_{j=0}^{\infty}$ is an $m \times m$ inner sequence.

The following theorem, which contains a complete description of the Taylor coefficient sequences of J-inner functions, is the second main result of this paper.

Theorem 6.5. *Let J be an $m \times m$ signature matrix. Then:*

(a) *Let $f \in \underline{\mathcal{P}}_{J,0}(\mathbb{D})$. Then the sequence $(A_j)_{j=0}^{\infty}$ of Taylor coefficients of f is a J-inner sequence.*

(b) *Let $(A_j)_{j=0}^{\infty}$ be J-inner sequence. Then:*

(b1) *There is a unique $f \in \underline{\mathcal{P}}_{J,0}(\mathbb{D})$ such that $(A_j)_{j=0}^{\infty}$ is the sequence of Taylor coefficients of f.*

(b2) *The sequence $(A_j)_{j=0}^{\infty}$ is a J-Potapov sequence.*

Proof. Let the matrices P_J and Q_J be given by (2.4).

(a) From part (a) of Proposition 3.8, we obtain that $\det(Q_J f(w) + P_J) \neq 0$ for $w \in \mathbb{H}_f$ and that the function g defined by (3.1) belongs to the class $\underline{\mathcal{S}}_{m\times m}(\mathbb{D})$. Thus, if $(B_j)_{j=0}^{\infty}$ denotes the sequence of Taylor coefficients of g, then from part (a) of Proposition 4.3 we get that $(B_j)_{j=0}^{\infty}$ is an $m \times m$ inner sequence. Since we have shown in the proof of part (a) of Theorem 6.2 that $(B_j)_{j=0}^{\infty}$ is the J-PG transform of $(A_j)_{j=0}^{\infty}$, we see that $(A_j)_{j=0}^{\infty}$ is a J-inner sequence.

(b1) Because of the choice of $(A_j)_{j=0}^{\infty}$ we know that $\det(Q_J A_0 + P_J) \neq 0$ and that the J-PG transform $(B_j)_{j=0}^{\infty}$ of $(A_j)_{j=0}^{\infty}$ is an $m \times m$ inner sequence. Hence part (b1) of Proposition 4.3 implies that there is a unique $g \in \underline{\mathcal{S}}_{m\times m}(\mathbb{D})$ such that $(B_j)_{j=0}^{\infty}$ is the sequence of Taylor coefficients of g. In particular, $g(0) = B_0$. In view of the construction of the sequence $(B_j)_{j=0}^{\infty}$, part (c) of Proposition 5.11 yields $\det(Q_J g(0) + P_J) = \det(Q_J B_0 + P_J) \neq 0$. Combining this with $g \in \underline{\mathcal{S}}_{m\times m}(\mathbb{D})$, we obtain by using part (c) of Proposition 3.4 and part (b) of Proposition 3.8, that the function f defined by (3.3) belongs to $\underline{\mathcal{P}}_{J,0}(\mathbb{D})$. Because of (3.3) we have

(6.4) for each $n \in \mathbb{N}_0$. Consequently, Proposition 5.11 yields $S_n^{[f]} = S_n^{(A)}$ for each $n \in \mathbb{N}_0$. Hence $(A_j)_{j=0}^{\infty}$ is the sequence of Taylor coefficients of f. The proof of (b1) is complete.

(b2) Combine (b1) with part (a) of Theorem 6.2. □

At the end of this section we investigate the sequence of Taylor coefficients of a function belonging to $\mathcal{P}_{J,0}(\mathbb{D}) \cap \mathcal{P}_J''(\mathbb{D})$.

Proposition 6.6. *Let J be an $m \times m$ signature matrix and let $f \in \mathcal{P}_{J,0}(\mathbb{D}) \cap \mathcal{P}_J''(\mathbb{D})$. Denote by $(A_j)_{j=0}^{\infty}$ the sequence of Taylor coefficients of f. Then there exists some positive real constant α such that for all $n \in \mathbb{N}_0$ the inequality*

$$J_{[n]} - [S_n^{(A)}]^* J_{[n]} S_n^{(A)} \geq \alpha I_{(n+1)m}$$

is satisfied.

Proof. Let g be defined via (3.1). Then in view of $f \in \mathcal{P}_J''(\mathbb{D})$ we have $g \in \mathcal{S}_{m\times m}''(\mathbb{D})$. Denote by $(B_j)_{j=0}^{\infty}$ the sequence of Taylor coefficients of g. Because of $g \in \mathcal{S}_{m\times m}''(\mathbb{D})$ from [18, Theorem 5.5] we infer that there exists some positive constant β such that for all $n \in \mathbb{N}_0$ the inequality

$$I_{(n+1)m} - [S_n^{(B)}]^* S_n^{(B)} \geq \beta I_{(n+1)m}$$

is satisfied. Hence, for $n \in \mathbb{N}_0$ we have $\det[I_{(n+1)m} - [S_n^{(B)}]^* S_n^{(B)}] \neq 0$ and

$$[I_{(n+1)m} - [S_n^{(B)}]^* S_n^{(B)}]^{-1} \leq \frac{1}{\beta} I_{(n+1)m}. \tag{6.5}$$

As it was shown in the proof of Theorem 6.2 the sequence $(A_j)_{j=0}^{\infty}$ is the J-PG transform of $(B_j)_{j=0}^{\infty}$. Hence, from Lemma 5.15 we obtain for $n \in \mathbb{N}_0$ that $\det(Q_{J_{[n]}} S_n^{(B)} + P_{J_{[n]}}) \neq 0$ and

$$\begin{aligned} &J_{[n]} - [S_n^{(A)}]^* J_{[n]} S_n^{(A)} \\ &\quad = (Q_{J_{[n]}} S_n^{(B)} + P_{J_{[n]}})^{-*} [I_{(n+1)m} - [S_n^{(B)}]^* S_n^{(B)}] (Q_{J_{[n]}} S_n^{(B)} + P_{J_{[n]}})^{-1}. \end{aligned}$$

Thus, for $n \in \mathbb{N}_0$, we get $\det(J_{[n]} - [S_n^{(A)}]^* J_{[n]} S_n^{(A)}) \neq 0$ and

$$\begin{aligned} &[J_{[n]} - [S_n^{(A)}]^* J_{[n]} S_n^{(A)}]^{-1} \\ &\quad = (Q_{J_{[n]}} S_n^{(B)} + P_{J_{[n]}}) [I_{(n+1)m} - [S_n^{(B)}]^* S_n^{(B)}]^{-1} (Q_{J_{[n]}} S_n^{(B)} + P_{J_{[n]}})^*. \end{aligned} \tag{6.6}$$

Let $\|X\|$ be the operator norm of a $p \times q$ matrix X, where $p, q \in \mathbb{N}$. In view of $g \in \mathcal{S}_{m\times m}''(\mathbb{D})$ we infer from part (a) of Theorem 4.1 that $\|S_n^{(B)}\| \leq 1$ and consequently taking into account that $\|Q_{J_{[n]}}\| \leq 1$ and $\|P_{J_{[n]}}\| \leq 1$ then

$$\|Q_{J_{[n]}} S_n^{(B)} + P_{J_{[n]}}\| \leq 2. \tag{6.7}$$

Combining (6.5), (6.6), and (6.7) we see that there exists some $\gamma \in (0, \infty)$ such that for $n \in \mathbb{N}_0$ the inequality

$$[J_{[n]} - [S_n^{(A)}]^* J_{[n]} S_n^{(A)}]^{-1} \leq \gamma I_{(n+1)m}$$

is satisfied. Choosing $\alpha := \frac{1}{\gamma}$ we obtain the assertion. □

7. First considerations on an interpolation problem for functions belonging to the class $\mathcal{P}_{J,0}(\mathbb{D})$

In this section we consider the following interpolation problem for functions belonging to the class $\mathcal{P}_{J,0}(\mathbb{D})$.

Potapov problem (P): Let J be an $m \times m$ signature matrix, let $n \in \mathbb{N}_0$, and let $(A_j)_{j=0}^n$ be a sequence of complex $m \times m$ matrices. Describe the set $\mathcal{P}_{J,0}\left[\mathbb{D}, (A_j)_{j=0}^n\right]$ of all matrix-valued functions $f \in \mathcal{P}_{J,0}(\mathbb{D})$ such that

$$\frac{f^{(j)}(0)}{j!} = A_j$$

for each $j \in \mathbb{N}_{0,n}$ where the notation $f^{(j)}$ stands for the jth derivative of f. In particular, characterize the case that the set $\mathcal{P}_{J,0}\left[\mathbb{D}, (A_j)_{j=0}^n\right]$ is nonempty.

Observe that in the particular case $J = I_m$ this problem (P) is exactly the matricial Schur problem for $m \times m$ Schur functions.

First we treat the solvability of the problem (P). Our method is based on a transformation into an associated Schur problem. The following result is proved, e.g., in Theorem 3.5.2 in [13].

Theorem 7.1. *Let $n \in \mathbb{N}_0$ and let $(B_j)_{j=0}^n$ be a sequence of complex $m \times m$ matrices. Then the set $\mathcal{S}_{m\times m}\left[\mathbb{D}, (B_j)_{j=0}^n\right]$ of all matrix-valued functions $g \in \mathcal{S}_{m\times m}(\mathbb{D})$ such that*

$$\frac{g^{(j)}(0)}{j!} = B_j \tag{7.1}$$

for each $j \in \mathbb{N}_{0,n}$ is nonempty if and only if $(B_j)_{j=0}^n$ is an $m \times m$ Schur sequence.

Now we can prove a criterion for the solvability of Problem (P).

Theorem 7.2. *Let J be an $m \times m$ signature matrix, let $n \in \mathbb{N}_0$, and let $(A_j)_{j=0}^n$ be a sequence of complex $m \times m$ matrices. Then the set $\mathcal{P}_{J,0}\left[\mathbb{D}, (A_j)_{j=0}^n\right]$ is nonempty if and only if $(A_j)_{j=0}^n$ is a J-Potapov sequence.*

Proof. If the set $\mathcal{P}_{J,0}\left[\mathbb{D}, (A_j)_{j=0}^n\right]$ is nonempty, then part (a) of Theorem 6.2 shows that $(A_j)_{j=0}^n$ is a J-Potapov sequence.

Conversely now suppose that $(A_j)_{j=0}^n$ is a J-Potapov sequence. From Proposition 5.16 we obtain then that $\det(Q_J A_0 + P_J) \neq 0$ and that the J-PG transform $(B_j)_{j=0}^n$ of $(A_j)_{j=0}^n$ is an $m \times m$ Schur sequence. In view of Theorem 7.1, there exists some function g which belongs to $\mathcal{S}_{m\times m}\left[\mathbb{D}, (B_j)_{j=0}^n\right]$. In particular, $g(0) = B_0$. Thus, from part (c) of Proposition 5.11 we see that $\det(Q_J g(0) + P_J) \neq 0$ and that $(A_j)_{j=0}^n$ is the J-PG transform of $(B_j)_{j=0}^n$. Consequently, part (b) of Proposition 3.4 yields that the function

$$f := (P_J g + Q_J)(Q_J g + P_J)^{-1} \tag{7.2}$$

belongs to $\mathcal{P}_{J,0}(\mathbb{D})$. Because of (7.2) we get

$$S_n^{[f]} = \left(P_{J_{[n]}} S_n^{[g]} + Q_{J_{[n]}}\right)\left(Q_{J_{[n]}} S_n^{[g]} + P_{J_{[n]}}\right)^{-1}. \tag{7.3}$$

In view of $g \in \mathcal{S}_{m\times m}\left[\mathbb{D}, (B_j)_{j=0}^n\right]$ we have

$$S_n^{[g]} = S_n^{(B)}. \tag{7.4}$$

Since $(A_j)_{j=0}^n$ is the J-PG transform of $(B_j)_{j=0}^n$ Proposition 5.11 provides

$$S_n^{(A)} = \left(P_{J_{[n]}} S_n^{(B)} + Q_{J_{[n]}}\right)\left(Q_{J_{[n]}} S_n^{(B)} + P_{J_{[n]}}\right)^{-1}. \tag{7.5}$$

Combining formulas (7.3), (7.4), and (7.5) we obtain $S_n^{[f]} = S_n^{(A)}$. Thus f belongs to $\mathcal{P}_{J,0}\left[\mathbb{D}, (A_j)_{j=0}^n\right]$. □

In a forthcoming paper we will prove that for each $n \in \mathbb{N}_0$ and each J-Potapov sequence $(A_j)_{j=0}^n$ there can be found a matrix $A_{n+1} \in \mathbb{C}^{m\times m}$ such that $(A_j)_{j=0}^{n+1}$ is a J-Potapov sequence. This fact provides in combination with Theorem 6.2 an alternative proof of Theorem 7.2.

Supposing that J is an $m\times m$ signature matrix, that $n \in \mathbb{N}_0$, and that $(A_j)_{j=0}^n$ is a J-Potapov sequence we are going now to describe the set $\mathcal{P}_{J,0}[\mathbb{D}, (A_j)_{j=0}^n]$ via a linear fractional transformation of matrices. Here we will use the corresponding statement in the particular case $J = I_m$ which was obtained in [18]. To state this description of the solution set of the matricial Schur problem we need some notations.

Let $n \in \mathbb{N}_0$ and let $(B_j)_{j=0}^n$ be an $m \times m$ Schur sequence. Then let the constant matrix-valued functions π_0, ρ_0, σ_0, and τ_0 be given for each $w \in \mathbb{C}$ by

$$\pi_0(w) := B_0, \qquad \rho_0(w) := I_m, \qquad \sigma_0(w) := B_0, \qquad \text{and} \qquad \tau_0(w) := I_m. \tag{7.6}$$

If $n \geq 1$, then let π_n, ρ_n, σ_n, and τ_n be the $m \times m$ matrix polynomials which are defined on $\mathbb{C}$ and which are given for each $w \in \mathbb{C}$ by

$$\pi_n(w) := B_0 + we_{n-1}(w)\left[I_{nm} - S_{n-1}^{(B)}(S_{n-1}^{(B)})^*\right]^+ y_n^{(B)}, \tag{7.7}$$

$$\rho_n(w) := I_m + we_{n-1}(w)(S_{n-1}^{(B)})^*\left[I_{nm} - S_{n-1}^{(B)}(S_{n-1}^{(B)})^*\right]^+ y_n^{(B)}, \tag{7.8}$$

$$\sigma_n(w) := z_n^{(B)}\left[I_{nm} - (S_{n-1}^{(B)})^* S_{n-1}^{(B)}\right]^+ w\mathcal{E}_{n-1}(w) + B_0, \tag{7.9}$$

and

$$\tau_n(w) := z_n^{(B)}\left[I_{nm} - (S_{n-1}^{(B)})^* S_{n-1}^{(B)}\right]^+ (S_{n-1}^{(B)})^* w\mathcal{E}_{n-1}(w) + I_m, \tag{7.10}$$

where $e_{n-1} : \mathbb{C} \to \mathbb{C}^{m\times nm}$ and $\mathcal{E}_{n-1} : \mathbb{C} \to \mathbb{C}^{nm\times m}$ are defined by

$$e_{n-1}(w) := (I_m, wI_m, \ldots, w^{n-1}I_m) \tag{7.11}$$

and

$$\mathcal{E}_{n-1}(w) := (\overline{w}^{n-1}I_m, \overline{w}^{n-2}I_m, \ldots, I_m)^*, \tag{7.12}$$

respectively, and where

$$y_n^{(B)} := (B_1^*, B_2^*, \ldots, B_n^*)^*, \qquad z_n^{(B)} := (B_n, B_{n-1}, \ldots, B_1).$$

Furthermore, we used the notation X^+ for the Moore-Penrose inverse of a $p \times q$ matrix X, where $p, q \in \mathbb{N}$. Observe that since $(B_j)_{j=0}^n$ is an $m \times m$ Schur sequence, the matrices

$$l_{n+1} := \begin{cases} I_m - B_0 B_0^*, & \text{if } n = 0 \\ I_m - B_0 B_0^* - z_n^{(B)} \left(I_{nm} - (S_{n-1}^{(B)})^* S_{n-1}^{(B)} \right)^+ (z_n^{(B)})^*, & \text{if } n \geq 1 \end{cases} \tag{7.13}$$

and

$$r_{n+1} := \begin{cases} I_m - B_0^* B_0, & \text{if } n = 0 \\ I_m - B_0^* B_0 - (y_n^{(B)})^* \left(I_{nm} - S_{n-1}^{(B)} (S_{n-1}^{(B)})^* \right)^+ y_n^{(B)}, & \text{if } n \geq 1 \end{cases} \tag{7.14}$$

are both nonnegative Hermitian (see, e.g., Lemma 3.3.1 and the proof of Theorem 3.5.1 in [13]).

Let P be an $m \times m$ matricial polynomial, i.e., there exist a nonnegative integer k and a complex $(k+1)m \times m$ matrix B such that $P(w) = e_k(w)B$ where the particular matrix polynomial e_k is defined as in (7.11). Then the reciprocal matrix polynomial $\tilde{P}^{[k]}$ of P with respect to the unit circle $\mathbb{T}$ and the formal degree k is given by

$$\tilde{P}^{[k]}(w) := B^* \mathcal{E}_k(w)$$

where the matrix polynomial $\mathcal{E}_k$ is defined as in (7.12).

The following theorem gives a parametrization of the solution set of the matricial Schur problem. To be more precise, it describes the set $\mathcal{S}_{m\times m}[\mathbb{D}, (B_j)_{j=0}^n]$ of all $g \in \mathcal{S}_{m\times m}(\mathbb{D})$ which satisfy (7.1) for each $j \in \mathbb{N}_{0,n}$.

Theorem 7.3. *Let $n \in \mathbb{N}_0$ and let $(B_j)_{j=0}^n$ be an $m \times m$ Schur sequence. Then:*

(a) *For each $G \in \mathcal{S}_{m\times m}(\mathbb{D})$ and each $w \in \mathbb{D}$, the matrix*

$$w\tilde{\sigma}_n^{[n]}(w)\sqrt{l_{n+1}}^+ G(w)\sqrt{r_{n+1}} + \rho_n(w)$$

is nonsingular and the function $g : \mathbb{D} \to \mathbb{C}^{m\times m}$ defined by

$$\begin{aligned} g(w) := \; & \left(w\tilde{\tau}_n^{[n]}(w)\sqrt{l_{n+1}}^+ G(w)\sqrt{r_{n+1}} + \pi_n(w) \right) \\ & \cdot \left(w\tilde{\sigma}_n^{[n]}(w)\sqrt{l_{n+1}}^+ G(w)\sqrt{r_{n+1}} + \rho_n(w) \right)^{-1} \end{aligned} \tag{7.15}$$

belongs to $\mathcal{S}_{m\times m}[\mathbb{D}, (B_j)_{j=0}^n]$.

(b) *If $g \in \mathcal{S}_{m\times m}[\mathbb{D}, (B_j)_{j=0}^n]$, then there is a $G \in \mathcal{S}_{m\times m}(\mathbb{D})$ such that*

$$\begin{aligned} g(w) = \; & \left(w\tilde{\tau}_n^{[n]}(w)\sqrt{l_{n+1}}^+ G(w)\sqrt{r_{n+1}} + \pi_n(w) \right) \\ & \cdot \left(w\tilde{\sigma}_n^{[n]}(w)\sqrt{l_{n+1}}^+ G(w)\sqrt{r_{n+1}} + \rho_n(w) \right)^{-1} \end{aligned}$$

holds for each $w \in \mathbb{D}$.

A proof of Theorem 7.3 is given in [18, Theorem 1.1]. (Observe that Theorem 1.1 in [18] includes moreover the case of not necessarily quadratic matricial Schur functions.)

We again assume that J is an $m\times m$ signature matrix, that n is a nonnegative integer, and that $(B_j)_{j=0}^n$ is an $m\times m$ Schur sequence. Then we will use $\hat{\mathcal{C}}_n$ to denote the $2m\times 2m$ matricial polynomial which is given by

$$\hat{\mathcal{C}}_n := \begin{pmatrix} \hat{\mathcal{C}}_n^{[1,1]} & \hat{\mathcal{C}}_n^{[1,2]} \\ \hat{\mathcal{C}}_n^{[2,1]} & \hat{\mathcal{C}}_n^{[2,2]} \end{pmatrix} \tag{7.16}$$

where the matrix polynomials $\hat{\mathcal{C}}_n^{[1,1]}$, $\hat{\mathcal{C}}_n^{[1,2]}$, $\hat{\mathcal{C}}_n^{[2,1]}$, and $\hat{\mathcal{C}}_n^{[2,2]}$ are defined for each $w\in\mathbb{C}$ by

$$\hat{\mathcal{C}}_n^{[1,1]}(w) := w\left(P_J\tilde{\tau}_n^{[n]}(w) + Q_J\tilde{\sigma}_n^{[n]}(w)\right), \tag{7.17}$$

$$\hat{\mathcal{C}}_n^{[1,2]}(w) := P_J\pi_n(w) + Q_J\rho_n(w), \tag{7.18}$$

$$\hat{\mathcal{C}}_n^{[2,1]}(w) := w\left(Q_J\tilde{\tau}_n^{[n]}(w) + P_J\tilde{\sigma}_n^{[n]}(w)\right), \tag{7.19}$$

and

$$\hat{\mathcal{C}}_n^{[2,2]}(w) := Q_J\pi_n(w) + P_J\rho_n(w). \tag{7.20}$$

Moreover, let $\mathcal{C}_n$, $\mathcal{C}_n^{[1,1]}$, $\mathcal{C}_n^{[1,2]}$, $\mathcal{C}_n^{[2,1]}$, and $\mathcal{C}_n^{[2,2]}$ be the restrictions of $\hat{\mathcal{C}}_n$, $\hat{\mathcal{C}}_n^{[1,1]}$, $\hat{\mathcal{C}}_n^{[1,2]}$, $\hat{\mathcal{C}}_n^{[2,1]}$, and $\hat{\mathcal{C}}_n^{[2,2]}$, respectively, onto $\mathbb{D}$.

The following theorem is the third main result of this paper. It contains a complete answer to the Potapov problem (P) associated with a J-Potapov sequence.

Theorem 7.4. *Let J be an $m\times m$ signature matrix, let $n\in\mathbb{N}_0$, and let $(A_j)_{j=0}^n$ be a J-Potapov sequence. Let $(B_j)_{j=0}^n$ be the J-PG transform of $(A_j)_{j=0}^n$. Then the following statements hold:*

(a) *For each $G\in\mathcal{S}_{m\times m}(\mathbb{D})$, the matrix*

$$\mathcal{C}_n^{[2,1]}(0)\sqrt{l_{n+1}}^{+}G(0)\sqrt{r_{n+1}} + \mathcal{C}_n^{[2,2]}(0)$$

is nonsingular and the matrix-valued function

$$\begin{aligned} f :\ = \ & \left(\mathcal{C}_n^{[1,1]}\sqrt{l_{n+1}}^{+}G\sqrt{r_{n+1}} + \mathcal{C}_n^{[1,2]}\right) \\ & \cdot\left(\mathcal{C}_n^{[2,1]}\sqrt{l_{n+1}}^{+}G\sqrt{r_{n+1}} + \mathcal{C}_n^{[2,2]}\right)^{-1} \end{aligned} \tag{7.21}$$

belongs to $\mathcal{P}_{J,0}[\mathbb{D},(A_j)_{j=0}^n]$.

(b) *If f belongs to $\mathcal{P}_{J,0}[\mathbb{D},(A_j)_{j=0}^n]$, then there is a $G\in\mathcal{S}_{m\times m}(\mathbb{D})$ such that*

$$\begin{aligned} f \ = \ & \left(\mathcal{C}_n^{[1,1]}\sqrt{l_{n+1}}^{+}G\sqrt{r_{n+1}} + \mathcal{C}_n^{[1,2]}\right) \\ & \cdot\left(\mathcal{C}_n^{[2,1]}\sqrt{l_{n+1}}^{+}G\sqrt{r_{n+1}} + \mathcal{C}_n^{[2,2]}\right)^{-1}. \end{aligned} \tag{7.22}$$

Proof. Since $(A_j)_{j=0}^n$ is a J-Potapov sequence, Theorem 7.2 shows that the set $\mathcal{P}_{J,0}[\mathbb{D},(A_j)_{j=0}^n]$ is nonempty and moreover Proposition 5.16 yields that $(B_j)_{j=0}^n$

is an $m \times m$ Schur sequence which fulfills

$$\det(Q_J B_0 + P_J) \neq 0 \tag{7.23}$$

and that $(A_j)_{j=0}^n$ is the J-PG transform of $(B_j)_{j=0}^n$. Let $V_n : \mathbb{C} \to \mathbb{C}^{2m\times 2m}$ be defined for each $w \in \mathbb{D}$ by

$$V_n(w) := \begin{pmatrix} w\tilde{\tau}_n^{[n]}(w) & \pi_n(w) \\ w\tilde{\sigma}_n^{[n]}(w) & \rho_n(w) \end{pmatrix}. \tag{7.24}$$

Then using (2.5) and (7.16)–(7.20) it is readily checked that

$$\mathcal{A}_J V_n(w) = \mathcal{C}_n(w) \tag{7.25}$$

holds for each $w \in \mathbb{D}$.

(a) Let $G \in \mathcal{S}_{m\times m}(\mathbb{D})$. According to Theorem 7.3, then

$$\det\left(w\tilde{\sigma}_n^{[n]}(w)\sqrt{l_{n+1}}^{+} G(w)\sqrt{r_{n+1}} + \rho_n(w)\right) \neq 0 \tag{7.26}$$

for each $w \in \mathbb{D}$ and the function $g : \mathbb{D} \to \mathbb{C}^{m\times m}$ defined by (7.15) belongs to $\mathcal{S}_{m\times m}[\mathbb{D}, (B_j)_{j=0}^n]$. Then one can easily see that

$$g(w) = \mathcal{S}_{V_n(w)}^{(m,m)}\left(\sqrt{l_{n+1}}^{+} G(w)\sqrt{r_{n+1}}\right) \tag{7.27}$$

for each $w \in \mathbb{D}$. Moreover, in view of (7.23), we have

$$\det(Q_J g(0) + P_J) = \det(Q_J B_0 + P_J) \neq 0. \tag{7.28}$$

Hence,

$$\mathcal{M}_g := \{z \in \mathbb{D} : \det(Q_J g(z) + P_J) = 0\} \tag{7.29}$$

is a discrete subset of $\mathbb{D}$ and $0 \in \mathbb{D}\backslash\mathcal{M}_g$. According to part (c) of Proposition 3.4, the function $\hat{f} := (P_J g + Q_J)(Q_J g + P_J)^{-1}$ belongs to $\mathcal{P}_{J,0}(\mathbb{D})$. Taking into account (7.24), (7.26), (7.27), (7.29), and Lemma 1.6.1 and Proposition 1.6.3 in [13], (7.25), and (7.16), for each $w \in \mathbb{D}\backslash\mathcal{M}_g$ we can conclude then

$$\det\left(\mathcal{C}_n^{[2,1]}(w)\sqrt{l_{n+1}}^{+} G(w)\sqrt{r_{n+1}} + \mathcal{C}_n^{[2,2]}(w)\right) \neq 0$$

and

$$\begin{aligned} &\mathcal{S}_{\mathcal{C}_n(w)}^{(m,m)}\left(\sqrt{l_{n+1}}^{+} G(w)\sqrt{r_{n+1}}\right) = \mathcal{S}_{\mathcal{A}_J V_n(w)}^{(m,m)}\left(\sqrt{l_{n+1}}^{+} G(w)\sqrt{r_{n+1}}\right) \\ = \;& \mathcal{S}_{\mathcal{A}_J}^{(m,m)}\left(\mathcal{S}_{V_n(w)}^{(m,m)}\left(\sqrt{l_{n+1}}^{+} G(w)\sqrt{r_{n+1}}\right)\right) = \mathcal{S}_{\mathcal{A}_J}^{(m,m)}(g(w)) \\ = \;& \hat{f}(w) \end{aligned} \tag{7.30}$$

From (7.16), (7.21), and (7.30) we get then $f(w) = \hat{f}(w)$ for each $w \in \mathbb{D}\backslash\mathcal{M}_g$. Having in mind that $\mathcal{M}_g$ is a discrete subset of $\mathbb{D}$, this implies

$$f = \hat{f} = (P_J g + Q_J)(Q_J g + P_J)^{-1}. \tag{7.31}$$

In particular, f belongs to $\mathcal{P}_{J,0}(\mathbb{D})$. Let $(C_j)_{j=0}^{\infty}$ be the sequence of Taylor coefficients of f. Since g belongs to $\mathcal{S}_{m\times m}[\mathbb{D},(B_j)_{j=0}^n]$ and because of (7.31) we have then

$$\begin{aligned} S_n^{(C)} = S_n^{[f]} &= \left(P_{J_{[n]}}S_n^{[g]} + Q_{J_{[n]}}\right)\left(Q_{J_{[n]}}S_n^{[g]} + P_{J_{[n]}}\right)^{-1} \\ &= \left(P_{J_{[n]}}S_n^{(B)} + Q_{J_{[n]}}\right)\left(Q_{J_{[n]}}S_n^{(B)} + P_{J_{[n]}}\right)^{-1}. \end{aligned}$$

Hence, from Proposition 5.11 we get that $(C_j)_{j=0}^n$ is the J-PG transform of $(B_j)_{j=0}^n$. Consequently, $(C_j)_{j=0}^n = (A_j)_{j=0}^n$. Thus f belongs to $\mathcal{P}_{J,0}[\mathbb{D},(A_j)_{j=0}^n]$.

(b) Conversely, now suppose that f is an arbitrary matrix-valued function which belongs to $\mathcal{P}_{J,0}[\mathbb{D},(A_j)_{j=0}^n]$. Proposition 3.4 yields then that g given by (3.1) belongs to $\mathcal{S}_{m\times m}(\mathbb{D})$, that

$$\mathbb{H}_f = \{w \in \mathbb{D} : \det(Q_J g(w) + P_J) \neq 0\} \tag{7.32}$$

and that

$$f(w) = (P_J g(w) + Q_J)(Q_J g(w) + P_J)^{-1} = \mathcal{S}_{\mathcal{A}_J}^{(m,m)}(g(w)) \tag{7.33}$$

for each $w \in \mathbb{H}_f$. Since f is holomorphic at 0, from (7.32) it follows that

$$\det(Q_J g(0) + P_J) \neq 0. \tag{7.34}$$

Let $(D_j)_{j=0}^{\infty}$ be the sequence of Taylor coefficients of g. From (7.34), (3.1), and $f \in \mathcal{P}_{J,0}[\mathbb{D},(A_j)_{j=0}^n]$ we get then $\det(Q_J D_0 + P_J) \neq 0$ and

$$\begin{aligned} S_n^{(D)} = S_n^{[g]} &= \left(P_{J_{[n]}}S_n^{[f]} + Q_{J_{[n]}}\right)\left(Q_{J_{[n]}}S_n^{[f]} + P_{J_{[n]}}\right)^{-1} \\ &= \left(P_{J_{[n]}}S_n^{(A)} + Q_{J_{[n]}}\right)\left(Q_{J_{[n]}}S_n^{(A)} + P_{J_{[n]}}\right)^{-1}. \end{aligned}$$

Hence, from Proposition 5.11 we get that $(D_j)_{j=0}^n$ is the J-PG transform of $(A_j)_{j=0}^n$. Consequently, $(D_j)_{j=0}^n = (B_j)_{j=0}^n$. In particular, g belongs to the class $\mathcal{S}_{m\times m}[\mathbb{D},(B_j)_{j=0}^n]$. Using Theorem 7.3 we can conclude then that there is a $G \in \mathcal{S}_{m\times m}(\mathbb{D})$ such that (7.26) and (7.27) hold for each $w \in \mathbb{D}$. Taking into account (7.33), Lemma 1.6.1 and Proposition 1.6.3 in [13], and (7.25), it follows

$$\begin{aligned} f(w) &= \mathcal{S}_{\mathcal{A}_J}^{(m,m)}\left(\mathcal{S}_{V_n(w)}^{(m,m)}(\sqrt{l_{n+1}}^{+}G(w)\sqrt{r_{n+1}})\right) \\ &= \mathcal{S}_{\mathcal{A}_J V_n(w)}^{(m,m)}\left(\sqrt{l_{n+1}}^{+}G(w)\sqrt{r_{n+1}}\right) \\ &= \mathcal{S}_{\mathcal{C}_n(w)}^{(m,m)}\left(\sqrt{l_{n+1}}^{+}G(w)\sqrt{r_{n+1}}\right), \end{aligned}$$

for each $w \in \mathbb{H}_f$. Taking into account (7.16), then (7.22) is proved. □

It should be mentioned that in the case $J = I_m$ the description of the solution set of the Potapov problem given in Theorem 7.4 coincides with the description of the solution set of the matricial Schur problem given in [18, Theorem 1.1].

References

[1] D. Alpay: *The Schur Algorithm, Reproducing Kernel Spaces and System Theory*, SMF/AMS Texts and Monographs, Volume 5, Amer. Math. Soc., Providence, R.I. 2001.

[2] D. Alpay, H. Dym: *Hilbert spaces of analytic functions, inverse scattering and operator models I*, Integral Equations and Operator Theory 7 (1984), 589–641.

[3] D. Alpay, H. Dym: *Hilbert spaces of analytic functions, inverse scattering and operator models II*, Integral Equations and Operator Theory 8 (1985), 145–180.

[4] D.Z. Arov: *Darlington realization of matrix-valued functions* (Russian), Izv. Akad. Nauk SSSR, Ser. Mat. 37 (1973), 1299–1326,
English transl.: Math. USSR Izvestija 7 (1973), 1295–1326.

[5] C. Carathéodory: *Über den Variabilitätsbereich der Koeffizienten von Potenzreihen, die gegebene Werte nicht annehmen*, Math. Am. 64 (1907), 95–115.

[6] J.A. Ball, I. Gohberg, L. Rodman: *Interpolation of Rational Matrix Functions*, OT Series, Volume 45, Birkhäuser, Basel – Boston – Berlin 1990.

[7] S. Bogner, B. Fritzsche, B. Kirstein: *The Schur-Potapov algorithm for sequences of complex $p \times q$ matrices I*, Compl. Anal. Oper. Theory 1 (2007), 55–95.

[8] S. Bogner, B. Fritzsche, B. Kirstein: *The Schur-Potapov algorithm for sequences of complex $p \times q$ matrices II*, Compl. Anal. Oper. Theory 1 (2007), 235–278.

[9] M.S. Brodskii, M.S. Livšic: *Spectral analysis of non-selfadjoint operators and intermediate systems* (Russian), Uspekhi Mat. Nauk 13 (1958), Issue 1, 3–85,
English transl.: Amer. Math. Soc. Transl., Series 2, Volume 13 (1960), 256–346.

[10] P. Dewilde, H. Dym: *Schur recursion error formulas and convergence of rational estimations for stationary stochastic processes*, IEEE Trans. Inf. Theory 17 (1981), 416–461.

[11] P. Dewilde, H. Dym: *Lossless chain scattering matrices and optimum linear prediction: The vector case*, Intern. J. Circuit Theory Appl. 9 (1981), 135–175.

[12] V.K. Dubovoj: *Indefinite metric in the interpolation problem of Schur for analytic matrix functions I* (Russian), Teor. Funktsii, Funktsional. Anal. i Prilozhen. 37 (1982), 14–26,
English transl.: Amer. Math. Soc. Transl., Series 2, Volume 144 (1989), 47–60.

[13] V.K. Dubovoj, B. Fritzsche, B. Kirstein: *Matricial Version of the Classical Schur Problem*, Teubner Texte zur Mathematik, Band 129, B.G. Teubner, Stuttgart – Leipzig 1992.

[14] H. Dym: *J-contractive Matrix Functions, Reproducing Hilbert Spaces and Interpolation*, CBMS Regional Conf. Ser. Math., Volume 71, Amer. Math. Soc., Providence, R.I. 1989.

[15] A.V. Efimov, V.P. Potapov: *J-expansive matrix-valued functions and their role in the analytic theory of electrical circuits* (Russian), Uspekhi Mat. Nauk 28 (1973), Issue 1, 65–130,
English transl.: Russian Math. Surveys 28 (1973), Issue 1, 69–140.

[16] B. Fritzsche, B. Kirstein: *An extension problem for nonnegative Hermitian block Toeplitz matrices*, Math. Nachr. 130 (1987), 121–135.

[17] B. Fritzsche, B. Kirstein: *A Schur type matrix extension problem*, Math. Nachr. 134 (1987), 257–271.

[18] B. Fritzsche, B. Kirstein, A. Lasarow: *The matricial Schur problem in both nondegenerate and degenerate cases*, Math. Nachr. 282, No. 2 (2009), 1–31.

[19] V.E. Katsnelson: *Methods of J-theory in continuous interpolation theorems of analysis* (Russian), deposited in VINITI 1983,
English transl. with a foreword by T. Ando, Hokkaido University, Sapporo 1985.

[20] I.V. Kovalishina: *Analytic theory of a class of interpolation problems* (Russian), Izv. Akad. Nauk SSSR, Ser. Mat. 47 (1983), 455–497,
English transl.: Math. USSR Izvestija 22 (1984), 419–463.

[21] M.S. Livšic: *On a class of linear operators in Hilbert space* (Russian), Mat. Sbornik 19 (1946), 239–262,
English transl.: Amer. Math. Soc. Transl., Series 2, Volume 13 (1960), 61–83.

[22] M.S. Livšic: *Isometric operators with equal deficiency indices, quasi-unitary operators* (Russian), Mat. Sbornik 26 (1950), 247–264,
English transl.: Amer. Math. Soc. Transl., Series 2, Volume 13 (1960), 85–103.

[23] M.S. Livšic: *The Blaschke–Potapov factorization theory and the theory of nonselfadjoint operators*, in: Topics in Interpolation Theory (Eds.: H. Dym, B. Fritzsche, V.E. Katsnelson, B. Kirstein), OT Series, Volume 95, Birkhäuser, Basel – Boston – Berlin 1997, pp. 391–396.

[24] M.S. Livšic, V.P. Potapov: *A theorem on the multiplication of characteristic matrix functions* (Russian), Doklady Akad. Nauk SSSR 72 (1950), Issue 4, 625–628.

[25] V.P. Potapov: *The multiplicative structure of J-contractive matrix functions* (Russian), Trudy Moskov. Mat. Obsh. 4 (1955), 129–236,
English transl.: Amer. Math. Soc. Transl., Series 2, Volume 15 (1960), 131–243.

[26] V.P. Potapov: *Linear fractional transformations of matrices* (Russian), in: Investigations on Operator Theory and Their Applications (Ed.: V.A. Marchenko), Naukova Dumka, Kiev 1979, 75–97,
English transl.: Amer. Math. Soc. Transl., Series 2, Volume 138 (1988), 21–35.

[27] V.P. Potapov: *A theorem on the modulus. I: Main concepts. The modulus* (Russian), Teor. Funktsii, Funktsional. Anal. i Prilozhen. 38 (1982), 91–101,
English transl.: Amer. Math. Soc. Transl., Series 2, Volume 138 (1988), 55–65.

[28] L.A. Sakhnovich: *Interpolation Theory and Its Applications*, Kluwer, Dordrecht 1997.

[29] I. Schur: *Über Potenzreihen, die im Inneren des Einheitskreises beschränkt sind*, J. reine u. angew. Math., Part I: 147 (1917), 205–232; Part II: 148 (1918), 122–145.

[30] N. Wiener, P.R. Masani: *The prediction theory of multivariate stochastic processes. I. The regularity condition*, Acta Math. (1957), 111–150.

Bernd Fritzsche and Bernd Kirstein
Mathematisches Institut, Universität Leipzig
D-04109 Leipzig, Germany
e-mail: fritzsche@math.uni-leipzig.de
kirstein@math.uni-leipzig.de

Uwe Raabe
Fachbereich Mathematik, Universität Siegen
D-57068 Siegen, Germany
e-mail: raabe@math.uni-leipzig.de

Operator Theory:
Advances and Applications, Vol. 197, 281–336

A Solvable Model for Scattering on a Junction and a Modified Analytic Perturbation Procedure

B. Pavlov

The paper is dedicated to Mihail Samoilovich Livshits, who was first to consider a nonselfadjoint operator as a part of an extended selfadjoint scattering system.

Abstract. We consider a one-body spin-less electron spectral problem for a resonance scattering system constructed of a quantum well weakly connected to a noncompact exterior reservoir, where the electron is free. The simplest kind of the resonance scattering system is a quantum network, with the reservoir composed of few disjoint cylindrical quantum wires, and the Schrödinger equation on the network, with the real bounded potential on the wells and constant potential on the wires. We propose a Dirichlet-to-Neumann – based analysis to reveal the resonance nature of conductance across the star-shaped element of the network (a junction), derive an approximate formula for the scattering matrix of the junction, construct a fitted zero-range solvable model of the junction and interpret a phenomenological parameter arising in Datta-Das Sarma boundary condition, see [14], for T-junctions. We also propose using of the fitted zero-range solvable model as the first step in a modified analytic perturbation procedure of calculation of the corresponding scattering matrix.

Mathematics Subject Classification (2000). Primary 47A40, 47A48, 47A55; Secondary 47N50, 47N70, 35Q40.

Keywords. Junction, Fitted zero-range model, Dirichlet-to-Neumann map.

Outline

1. Introduction

A typical quantum resonance scattering system is composed of a compact inner region surrounded by barriers and an exterior reservoir, where the quantum dynamics is free. These components are weakly connected due to tunneling across the barriers or via a narrow connecting channels. Non-compact quantum networks (QN) are typical resonance scattering systems. Manufacturing of QN with prescribed transport properties is now a most challenging problem of computational nano-electronics. While physical laws defining transport properties of the QN are mostly represented in form of partial differential equations, the direct computing can't help optimization of design of the QN, because it requires expensive and resource consuming scanning over the space of physical and geometrical parameters of the network. The domain of scanning could be essentially reduced in the case when there exist an approximate explicit formula connecting directly the transport characteristics with the parameters defining the geometry and the physical properties of the network.

We derive an explicit approximate formula for the scattering matrix of a simplest QN – a junction – consisting of a vertex domain – a quantum well connected to the outer reservoir *decomposed geometrically into a sum of cylindrical leads*. The corresponding model Hamiltonian is obtained based on Glazman splitting, [4], $\mathcal{L} \to L_\Lambda \oplus l_\Lambda$ of the original Hamiltonian, depending on the Fermi level Λ, into the sum of two operators with complementary branches of the continuous spectra. The model proves to be fitted because the corresponding model Dirichlet-to-Neumann map (DN-map, see [74]) serves a rational approximation of the DN-map of the non-trivial component L_Λ of the split system.

In an important alternative class of the resonance scattering systems, represented by the Helmholtz resonator, the reservoir can't be decomposed into simple components similar to the cylindrical leads, but the finite leads connecting the compact subsystem – the resonator – with the reservoir, admit a similar decomposition. Then again, a fitted solvable model can be constructed, see [33], based on the splitting of the spectral channels in the leads. The model obtained can serve again as a first step – a jump start – of the corresponding analytic perturbation procedure. We postpone the discussion of the Helmholtz resonator and other systems with nontrivial reservoirs to oncoming publications.

Main difficulty of analysis of resonance scattering systems of both above kinds on the networks is defined by presence of the eigenvalues of the isolated

compact subsystem, embedded into the continuous spectrum of the reservoir, separated from the compact subsystem. Indeed, for a selfadjoint operator A_0 in the Hilbert space E, *with discrete spectrum*, and small εV, the selfadjoint perturbation, $\| \varepsilon V \| \leq \varepsilon$, defines, for each simple isolated eigenvalue λ_s^0 of A_0

$$2\varepsilon < \min_{t \neq s} |\lambda_s - \lambda_t| \equiv \rho_s$$

a branch of eigenvalues λ_s^ε of the perturbed operator $A_\varepsilon := A_0 + \varepsilon V$ represented in form of a geometrically convergent series

$$\lambda_s^\varepsilon = \lambda_s^0 + \varepsilon\, \lambda_s^0(1) + \varepsilon^2\, \lambda_s^0(2) + \varepsilon^3\, \lambda_s^0(3) + \cdots ,$$

and the corresponding branch of eigenfunctions, see [29].

This standard analytic perturbation approach is not applicable, generally, to operators with eigenvalues embedded into the continuous spectrum, in particular to non-compact QN where the "spacing" ρ is zero. Development of radio-location during WWII required analysis of scattering problems on the networks of electromagnetic wave guides, in particular on junctions. The scattering on the junction is a typical perturbation problem for embedded eigenvalues. The perturbation of the problem causes the transformation of real eigenvalues on the vertex domain of the junction into complex resonances. This problem can't be solved by methods of the standard spectral theory of selfadjoint operators. In the paper [41] M.S. Livshits proposed an elegant approach to the problem of transmission of electro-magnetic signals across the junction, taking into account only oscillatory electro-magnetic modes in the wave-guides and neglecting the "evanescent"-exponentially decreasing modes. He reduced the calculation of the scattering matrix to calculation of the corresponding characteristic function and found a real wave conductance for the oscillatory modes and pure imaginary wave conductance for evanescent modes. The discovery, based on M.S. Livshits ideas, of the connection between the scattering matrix and the characteristic function of the corresponding non-selfadjoint operator, see [1], was an extraordinary achievement and became eventually a source/basement of a series of important results in the theory of the functional models of the dissipative operators, see [40, 51, 50]. The approach to the perturbation theory developed in these papers permitted to understand the spectral nature of the resonances, but it does not help practical problem of optimization of design of quantum or electro-magnetic networks with prescribed transport properties. One more detail in the above paper [41] was important in this respect. M.S. Livshits completely disregarded the evanescent waves in the wave-guides, which was usual in the papers of engineers and physicists on the electro-magnetic wave-guides. But he emphasized in [41] importance of accurate elimination of the evanescent waves. This was not done till now. We see now, that the absence of analysis of the evanescent waves prevented M.S. Livshits from establishing, at that time, an effective connection between the geometry of the junction and the transmisson/reflection coefficients, see also our comments below, Section 3.

In the case of operators with continuous or dense discrete spectrum one can substitute the Hamiltonian A_0 of an unperturbed system by a fitted solvable model

A^ε, and then develop an analytic perturbation procedure between the perturbed Hamiltonian A_ε and the model A^ε. This two-steps idea $A_0 \to A^\varepsilon \to A_\varepsilon$ of the modified analytic perturbation procedure was suggested, in implicit form, by H. Poincaré for relevant problems of celestial mechanics, see [63], and formulated in an explicit form in 1972 by I. Prigogine. In 1972 I. Prigogine, [65], declared importance of the search of a general practical algorithm for the two-step analytic perturbation procedure

$$A_0 \longrightarrow A^\varepsilon \longrightarrow A^\varepsilon$$

implementing the above Poincaré idea. Prigogine attempted to find an intermediate operator A^ε as a function of the unperturbed operator $A^\varepsilon = \Phi(A_0)$, and he wanted to have the above two step analytic perturbation procedure on the whole Hilbert space. The search of the corresponding "intermediate operator" A^ε continued for almost 20 years, but did not give any results. Finally Prigogine declared that the intermediate operator with the expected properties does not exist and can't be constructed.

We guess now, that I. Prigogine's suggestion based on the intermediate operator A^ε was very close to success. The idea of Prigogine was commonly used by physicists in form of effective Hamiltonian of a complex quantum systems, and, after essential modification, in [39] for "geometrical integration" in dynamical problems of classical mechanics.

In our recent papers [7, 44, 28], see also an extended list of references below, we suggested a method of accurate elimination of the evanescent waves based on the idea of the intermediate Hamiltonian. Our method also permits to accurately eliminate the evanescent waves in the case studied by M.S. Livshits. In this paper we provide, following the quoted papers, a review of the corresponding modified approach to the analytic perturbation procedure and describe, based on [60], an algorithm of construction of the solvable model and the procedure of fitting. We developed the corresponding general approach to the spectral problems with embedded eigenvalues in the series of papers [7, 57, 58, 28, 59, 44, 60, 62]. Contrary to the original Prigogine's idea, we do a couple of changes:

1. We search for the intermediate operator – the "jump start", see [56] – A^ε, on the first step of the above-mentioned two-step procedure, not among functions $\Phi(A_0)$ of A_0, as I. Prigogine suggested, but among weak (finite-dimensional) perturbations of the non-perturbed Hamiltonian A_0, which is close to the method suggested by Livshits.

2. We restricted our analysis to the part of the unperturbed operator on a spectral subspace which corresponds to some "essential spectral interval", contrary to I. Prigogine who attempted to find a global intermediate operator on the whole space. Thus we develop our modified analytic perturbation procedure *locally*. A similar requirement of locality is applied in [39] on the space of initial data of the structure-preserving model.

We begin with two classical examples of the resonance scattering systems, to reveal typical difficulties arising from the very beginning when considering pertur-

bations of systems with eigenvalues embedded into the continuous spectrum, and discuss nearest prospects of the perturbative analysis of these systems.

Example 1: Helmholtz Resonator. Helmholtz resonator was probably the first resonance scattering system discussed mathematically, see [66]. It is composed of the typical details: the inner domain $\Omega_{\rm int}$, the shell $\Omega_{\rm shell}$, and the reservoir $\Omega_{\rm out}$ separated by the shall from $\Omega_{\rm int}$. Consider the Helmholtz equation $-\Delta u = \lambda u$ in a domain $\Omega \in R_3$ represented as a sum of two disjoint parts $\Omega = \Omega_{\rm int} \cup \Omega_{\rm out}$ and a shell $\Omega_{\rm shell}$ with a small opening. Kirchhoff suggested to substitute the problem by the model where the opening is pointwise, so that there exist only one common point $a \in \bar\Omega_{\rm int} \cup \bar\Omega_{\rm out} \cup \bar\Omega_{\rm shell}$. Kirchhoff suggested an Ansatz for the Green-function $G_\lambda(x,y)$ of the Helmholtz equation in Ω with Neumann boundary condition

$$-\Delta G_\lambda(x,y) - \lambda G_\lambda(x,y) = \delta(x-y),\ \left.\frac{\partial G}{\partial n_x}\right|_{\partial\Omega} = 0,\ x,y \in \bar\Omega.$$

in the form of a linear combination of the Green functions $G^{\rm int}_\lambda(x,y)$, $G^{\rm out}_\lambda(x,y)$ of the inner and the outer problems:

$$-\Delta G^{\rm in,out}_\lambda(x,y) - \lambda G^{\rm in,out}_\lambda(x,y) = \delta(x-y),\ \left.\frac{\partial G}{\partial n_x}\right|_{\Omega_{\rm in,out}} = 0.$$

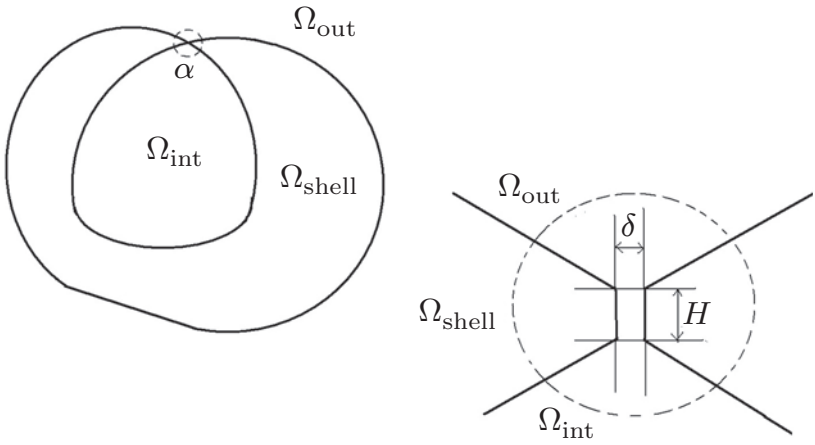

FIGURE 1. Helmholtz Resonator with a point-wise opening at the point a and the enlarged detail of the resonator with a narrow short channel, $\delta \ll H \ll \lambda^{-1/2}$

$$G_\lambda(x,y) - \left\{ \begin{array}{ccc} G^{\rm out}_\lambda(x,y) & + & A^{\rm out}\, G^{\rm out}_\lambda(x,a),\ x,y \in \Omega_{\rm out} \\ A^{\rm int}\, G^{\rm in}_\lambda(x,a), & \text{if} & y \in \Omega_{\rm out},\ x \in \Omega_{\rm int}, \end{array} \right. ,$$

with undefined constants – the Kirchhoff coefficients $A^{\rm out}$, $A^{\rm int}$. This *Kirchhoff Ansatz* satisfies the equation and the Neumann boundary conditions everywhere on $\partial\Omega$, except the point a, where the Ansatz is singular. The problem of choice of

the Kirchhoff constant and other interesting problems concerning the resonator, see [66] remained open until recent time, see in this connection the preprint [11].

Example 2: Zero-range potential. In 1933 E. Fermi, [19], considered the problem of scattering of neutrons n by the nucleon S of Sulphur. He suggested to choose for this problem the model Hamiltonian in the form of Laplacian in $L_2(R_3)$ defined on smooth functions $u \in L_2(R_3)$ with a singularity at the origin

$$u(x) = \frac{A^u}{4\pi|x|} + B^u + \cdots ,$$

and a special boundary condition imposed on the asymptotic boundary values A^u, B^u:

$$A^u = \gamma B^u,\ \gamma = \bar{\gamma}.$$

The Laplacian is symmetric and even selfadjoint with this boundary condition, and admits explicit calculation of the eigenfunctions: this model is "solvable". Fermi suggested to "fit" this model choosing $\gamma = -4\pi p_0^{-1}$, if $-p_0^2$ is a small negative eigenvalue in the system n, S.

The model can be extended to the case when $\gamma > 0$, and fit to the purely imaginary resonance $p_0 = i\gamma$. The resulting mysterious "zero-range potential" suggested by Fermi was interpreted by F. Berezin and L. Faddeev [8] in terms of von Neumann Operator Extensions Theory, [48]. Later this "zero-range potential" was used in numerous physical and mathematical papers and books, see, e.g., [5].

In both above examples the reservoirs are either a large exterior domain, or the whole space with single point $x = 0$ removed. The first example was also treated by the operator extension methods in [18], producing a *zero-range solvable model* of the resonator immersed into 3D space. The role of the unperturbed operator in [18] played an orthogonal sum of the Neumann Laplacian $L_{\rm int}$ in $L_2(\Omega_{\rm int})$ and $L_{\rm out}$ in $L_2(\Omega_{\rm out})$. The basic difficulty of the original perturbation problem, with a thin short channel, is caused by presence of the eigenvalues of $L_{\rm int}$ embedded into the continuous spectrum of $L_{\rm out}$. The standard selfadjoint spectral theory is *generally* unable to treat the problem of embedded eigenvalues, by observing transformation of them into the corresponding complex resonances. The elegant Lax-Phillips approach to resonance scattering problems reveals the spectral nature or resonances, see [40], but does not help to calculate them. In [18] the resonances can be easily calculated via solving an algebraic equation, but yet the *fitting* of the suggested zero-range model remained a problem. In [11] an approach to the problem of fitting of the model is suggested based on an explicit formula connecting the "full" scattering matrix of the Helmholtz Resonator with the Neumann-to-Dirichlet map, see [55], and a subsequent rational approximation of the Neumann-to-Dirichlet map for the inner domain of the resonator (the cavity $\Omega_{\rm int}$). Fortunately the problem of search of the resonances, in the case of small opening, becomes finite-dimensional after replacement of the Neumann-to-Dirichlet map of the cavity by the corresponding rational approximation, see more details in [11].

In the second example just a selfadjoint operator $-\Delta_\gamma$ is suggested, with only parameter γ, which can be interpreted in spectral terms. This operator plays a role of an effective Hamiltonian of the original scattering problem for the neutrons and the nuclei, see [19]. Yet again, the substitution of the original perturbed (*full*) Hamiltonian by the effective solvable Hamiltonian $-\Delta_\gamma$ requires fitting of the model, at least on some essential interval of energy.

Note that the role of the effective Hamiltonian is played, in the second example, by a selfadjoint extension of the unperturbed Hamiltonian $-\Delta$. Numerous effective Hamiltonians in quantum mechanics are constructed as zero-range solvable models of quantum systems see for instance [54] and our recent papers quoted above.

In this review we represent some results of our recent papers quoted above (see the text preceding the Example 1) where the effective Hamiltonians are constructed as zero-range solvable models. To make the text easily readable, we omit some proofs and most of technical details, which can be found in the original publications. But we pay additional attention to the interconnections of our constructions previously spread in different publications.

2. Scattering on quantum networks and junctions via DN-map

The basic idea of analysis of partial differential equations on quantum networks is that the corresponding Schrödinger problem can be divided in two parts: a Schrödinger equation on the region surrounded by barriers (e.g., a quantum well) and one on the reservoir the two being weakly coupled by tunneling, see for instance [64] or by a thin channel. It is noticed in [64] that this decomposition "corresponds to the schematization of the transport process as a coherent process on the quantum well, fed by the exterior reservoirs – quantum wires". On the reservoirs, assumed to be homogeneous and neutral, the electron-electron interaction is neglected, and the single electron is free. But the resonance properties of the quantum well and the tunneling on the contacts define the transport properties of the whole network. It is a common belief that thin quantum network can be modeled by a 1d graph, see [52], with either Kirchhoff boundary conditions, or just non-specified selfadjoint boundary conditions at the vertices. There exist an extended bibliography concerning one-dimensional models of quantum networks, see for instance [13, 37, 38, 26, 73]. Notice that even sharp resonance effects on 2d wave-guides and networks were studied theoretically mainly by numerical methods, see for instance [24, 77].

Quantum network Ω which is being studied in this paper, is composed of straight leads (quantum wires) width δ, some of them semi-infinite, and vertex domains Ω_s (quantum wells), see Fig. 2. An important basic detail of the quantum network is a junction, see Fig. 3. The junction is a non-compact quantum network composed of a quantum well and few semi-infinite quantum wires, of constant width, attached to it. The junction is usually called thin, if the diameter

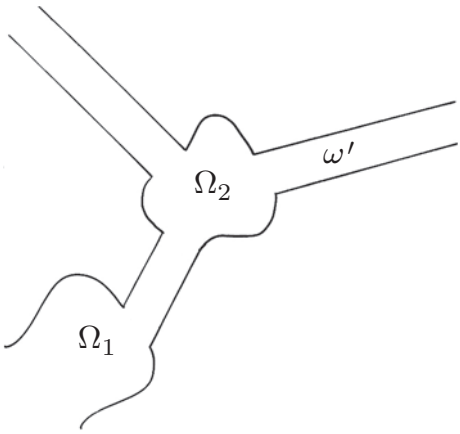

FIGURE 2. Quantum Network: a detail

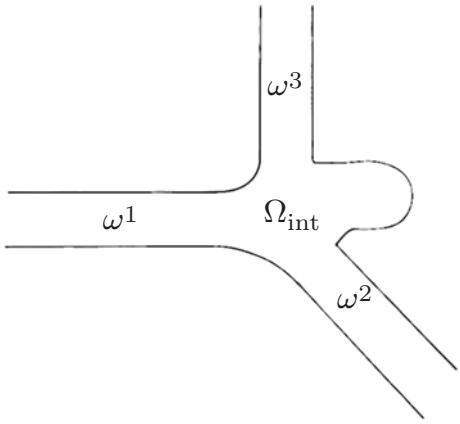

FIGURE 3. General junction

of the quantum well Ω_{int} strongly dominates the width δ of the wires ω attached to it: $\delta \ll \text{diam } \Omega_{\text{int}}$. Calculation of the scattering matrix of a junction is a challenging computational problem, yet accessible for standard commercial programs, see the discussion below. Physicists have certain preferences about the boundary conditions at the vertices, see the discussion below, Example 3.

Example 3: Thin symmetric T-junction. For thin symmetric T-junction, with the "bar" orthogonal to the "leg", a reasonably simple explicit formula for the scattering matrix was suggested in [14] based on reduction of the 2D scattering problem on the junction to the corresponding 1D scattering problem on the corresponding

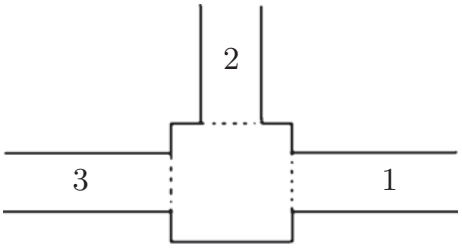

FIGURE 4. Simplest symmetric T-junction with a square vertex domain

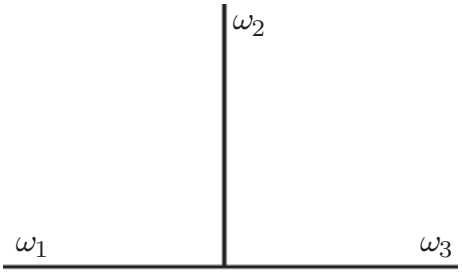

FIGURE 5. Model symmetric T-junction

quantum graph, see Fig. 5. The boundary conditions for the model T-junction suggested in [14], is presented in terms of limit values of the wave-function on the 1D wires $\{\psi_i\}_{i=1}^3 := \vec{\psi}$ and the values of the corresponding outward derivative (boundary currents) $\{\psi_i'\}_{i=1}^3 := \vec{\psi}'$ at the node:

$$\psi_1 = \beta^{-1}\psi_2 = \psi_3\,, \quad \psi_1' + \beta\psi_2' + \psi_3' = 0, \tag{1}$$

or in the form

$$P_\beta^\perp \vec{\psi} = 0\,, \quad P_\beta \vec{\psi}' = 0 \tag{2}$$

with the projection

$$P_\beta = \frac{1}{\beta^2+2}\begin{pmatrix} 1 & \beta & 1 \\ \beta & \beta^2 & \beta \\ 1 & \beta & 1 \end{pmatrix}. \tag{3}$$

The scattering matrix of such a junction is constant $S_\beta = I - 2P_\beta$, with the phenomenological parameter β responsible for connection between the bar and the leg. This formula was intensely used, see for instance [69, 72], despite unclear meaning of the parameter β. See further discussion of transmission across the junction in [16, 17, 25] and find more references therein.

In this paper, based on the resonance conception of conductance on the junction, we suggest a semi-analytic procedure of calculation of the scattering matrix

and a method of reduction of a general thin junction to a quantum graph. Moreover, we suggest a solvable model of a thin junction and reveal the meaning of the projection P_β. In this paper we do not take into account the spin-orbital interaction, just by disregarding the spin of the electron.

We consider a junction Ω constructed of a few straight leads ω^m, $\cup_{m=1}^M \omega^m = \omega$, width δ, attached orthogonally to the flat pieces Γ_m of the piecewise-smooth boundary of the vertex domain $\Omega_{\rm int}$, $\Omega = \Omega_{\rm int} \cup \omega$. On smooth functions $u \in W_2^2(\Omega)$ satisfying the homogeneous Neumann boundary condition, we define the Schrödinger operator

$$-\triangle u + Vu \ =: \mathcal{L}$$

with the potential V equal to the constant V_δ on the leads and equal to a real bounded piecewise-continuous function on $\Omega_{\rm int}$ supplied with Dirichlet boundary condition. The operator $\mathcal{L}$ is essentially selfadjoint, and it can be considered as a perturbation of the corresponding operator $\mathcal{L}_0$ defined by the same differential expression and an additional Dirichlet boundary condition on $\cup_{m=1}^M \Gamma_m =: \Gamma$:

$$\mathcal{L} \longrightarrow l^\omega \oplus L_{\rm int} \ = \ \mathcal{L}_0 \ .$$

The spectrum $\sigma(L_{\rm int})$ of $L_{\rm int}$ is discrete, and the spectrum σ^ω of L^ω is absolutely continuous, consists of a countable set of branches $\sigma^\omega = \cup_{m=1}^M \cup_{l=1}^\infty \sigma_l^m$ corresponding to the parts l_l^m of l^ω

$$l_l^m \ = \ -\frac{d^2}{dx^2} + \frac{\pi^2 l^2}{\delta^2} + V_\delta, \quad l \geq 1 \ ,$$

with the homogeneous Dirichlet boundary condition $u\big|_\Gamma = 0$ at the bottom sections $x^m\big|_{\Gamma_m} = 0$. The operators L_l^m on the wires are obtained from L^ω via separation of variables, with the basis of cross-section eigenfunctions $\{e_l^m\} = \left\{\sqrt{\frac{2}{\delta}} \sin \frac{\pi l y}{\delta}\right\}$, $l = 1, 2, \ldots$, $m = 1, 2, \ldots, M$. Here the local transversal coordinate on ω^m is denoted by y. The eigenfunctions of L^ω are scattered waves on each lead ω^m:

$$\psi_l^m(x) \ = \ \chi_+^l(x) \quad - \quad \chi_-^l(x), \quad x = x_m \geq 0 \ ,$$

represented as linear combinations of oscillating exponential modes

$$\chi_\pm^{m,l} \ = e^{\pm i \sqrt{\lambda - \lambda_l}\ x}\ e_l^m(y) := \ e^{\pm i K_+^{m,l} x}\, e_l^m(y), \quad \lambda > \lambda_l = \pi^2 l^2 \delta^{-2}, \qquad (4)$$

with the reflection coefficients $S_l = 1$. The perturbed operator $\mathcal{L}$ is obtained from $\mathcal{L}_0$ by replacement of the homogeneous Dirichlet boundary condition on the bottom sections Γ by the smooth matching condition. The corresponding scattered waves are obtained via matching on Γ a solution of the Schrödinger equation on the vertex domain with the scattering Ansatz (see for instance [46, 47]):

$$\psi_l^m(x) = \begin{cases} \chi_+^l(x) + \sum_{\pi^2 r^2/\delta^2 < \lambda} S_{l,r}^{m,m} \chi_-^r(x) + \sum_{\pi^2 r^2/\delta^2 > \lambda} s_{l,r}^{m,m} \xi^r(x), x \in \omega^m \\ \sum_{\pi^2 r^2/\delta^2 < \lambda} S_{l,r}^{m,n} \chi_-^r(x) + \sum_{\pi^2 r^2/\delta^2 > \lambda} s_{l,r}^{m,n} \xi^r(x), x \in \omega^n,\ n \neq m \ , \end{cases} \qquad (5)$$

composed, for given λ, of the above oscillating modes $\chi^r_\pm$ in the open channels, with the thresholds below λ, $\lambda^m_r < \lambda$, and the exponentially decreasing ("evanescent") modes in the closed channels

$$\xi^{m,s}_- \;=\; e^m_s(y)\; e^{-\;\sqrt{\lambda_s-\lambda}\;\;x} := \; e^{-K^{m,s}_-\,x}\, e^m_s(y), \quad \lambda < \lambda_s, \tag{6}$$

associated with the thresholds $\lambda^m_s = \pi^2\, s^2\, \delta^{-2}$ of the closed channels in the leads – see [44] for details.

The quantum wells and the quantum wires are usually manufactured as a certain relief of the surface of the semiconductor. We assume in this paper that the scaled Fermi level $\Lambda = 2m^* E_F \hbar^{-2}$ of the semiconductor is situated in the middle of the first spectral band $\Delta_1 = [V_\delta + \frac{\pi^2}{\delta^2},\, V_\delta + 4\frac{\pi^2}{\delta^2}]$ of the wire, $\Lambda = V_\delta + \frac{5}{2}\frac{\pi^2}{\delta^2}$. Then the first spectral band plays the role of the conductivity band and the junction has metallic properties. At low temperature T, the scattering processes are observed only on the essential spectral interval

$$\Delta^T = [\Lambda - 2m^*\kappa T\hbar^{-2},\, \Lambda - 2m^*\kappa T\hbar^{-2}] \subset \Delta_1. \tag{7}$$

If the electron's density is low, the main contribution to the scattering picture is defined by one-body processes. In this paper we focus on one-body scattering on the essential spectral interval. We disregard the spin-orbital interaction and neglect all effects connected with electrons spin. Hence we study the scattering on the first spectral band $\Delta_1 = [\pi^2\,\delta^{-2},\, 4\pi^2\,\delta^{-2}]$ of the open channel, and represent the cross-section space $L_2(\Gamma) =: E$ as an orthogonal sum of the entrance subspaces $E_\pm$ of the open and closed spectral channels respectively:

$$E_+ \;=\; \bigvee_{m=1}^{M} e^m_1, \quad E_- \;=\; \bigvee_{m=1}^{M}\bigvee_{l=2}^{\infty} e^m_l, \quad P_{E_\pm} =: P_\pm\ . \tag{8}$$

The infinite linear system for the coefficients of the scattering Ansatz, obtained from the matching conditions, can be solved, if the Green functions $G^D_\Gamma = G_{\rm int}$ of the Schrödinger operators $L^D_\Gamma = L_{\rm int}$ in $L_2(\Omega_{\rm int})$, with Dirichlet boundary conditions is constructed. The operator L^D_Γ is defined on W^2_2-functions in $\Omega_{\rm int}$, with the Meixner conditions at the inner corner points:

$$L_{\rm int}u = -\Delta u + Vu = \lambda u,\ u\big|_{\partial\Omega_{\rm int}} = 0. \tag{9}$$

The Green function is found from the equation:

$$L^D_\Gamma G^D_\Gamma \;=\; -\triangle G^D_\Gamma + VG^D_\Gamma \;=\; \lambda G^D_\Gamma + \delta(x-y),\ G^D_\Gamma\big|_{\partial\Omega_{\rm int}} \;=\; 0. \tag{10}$$

Hereafter we denote by σ^D the spectrum of L^D_Γ. According to the general theory of second-order elliptic equations, the solution u of the boundary problem

$$-\Delta u + Vu = \lambda u,\ u\big|_\Gamma = u_\Gamma,\ u\big|_{\partial\Omega_{\rm int}\setminus\Gamma} = 0. \tag{11}$$

is represented by the Poisson map

$$u(x) \;=\; \int_\Gamma \mathcal{P}_\Gamma(x,\gamma,\lambda)u_\Gamma(\gamma)\; d\gamma\ ,$$

involving the kernel $\mathcal{P}_{\rm int}(x,\gamma) = -\partial G^D_\Gamma(x,\gamma)/\partial n_\gamma$. The corresponding boundary current on Γ is calculated as

$$\left.\frac{\partial u}{\partial n}\right|_{x\in\Gamma} = -\int_\Gamma \frac{\partial^2 G^D_\Gamma(x,\gamma,\lambda)}{\partial n_x \partial n_\gamma} u_\Gamma(\gamma)\, d\Gamma \ =: \mathcal{DN}_\Gamma(\lambda) u_\Gamma \ .$$

This formal integral operator is restriction onto Γ of the Dirichlet-to-Neumann map, see [74, 21, 22]. For the sake of brevity we call it here "relative DN-map". The relative DN-map is also a Nevanlinna class function $\mathcal{DN}(\lambda)$ for Im $\lambda \le 0$, with poles at the eigenvalues of the corresponding Schrödinger operator $L^D_\Gamma = L_{\rm int}$. The relative DN-map is a pseudo-differential operator of order 1: for $W^2_2(\Omega)$ solutions u the DN-map acts from $W^{3/2}_2(\Gamma)$ to $W^{1/2}_2(\Gamma)$ and for $W^{3/2}_2(\Omega)$ generalized solutions the D-map acts from $W^1_2(\Gamma)$ to $L_2(\Gamma)$.

We consider also the boundary problem

$$-\Delta u + Vu = \lambda u, \ \left.\frac{\partial u}{\partial n}\right|_\Gamma = \rho_\Gamma, \ \left. u\right|_{\partial\Omega_{\rm int}\setminus\Gamma} = 0. \tag{12}$$

and the operator

$$L^N_\Gamma = -\Delta u + Vu, \ \left.\frac{\partial u}{\partial n}\right|_\Gamma = 0, \ \left. u\right|_{\partial\Omega_{\rm int}\setminus\Gamma} = 0. \tag{13}$$

with the relative Neumann Green function G^N_Γ:

$$L^N_\Gamma G^N_\Gamma \ = \ -\triangle G^N_\Gamma + VG^N_\Gamma \ = \ \lambda G^N_\Gamma + \delta(x-y), \ \left. G^N_\Gamma\right|_{\partial\Omega_{\rm int}\setminus\Gamma} = 0, \ \left.\frac{\partial G^N}{\partial n_x}\right|_{\partial\Omega_{\rm int}\setminus\Gamma} = 0. \tag{14}$$

The map

$$u(x) \ = \ \int_\Gamma G^N_\Gamma(x,\gamma,\lambda)\rho_\Gamma(\gamma) d\,\Gamma \ =: Q^\Gamma_1 \rho_\Gamma, \quad x\in\Omega_{\rm int} \ ,$$

gives a solution of the relative Neumann boundary problem (12). The trace of the solution on Γ

$$\left. u(x)\right|_\Gamma \ = \ \int_\Gamma G^N_\Gamma(x,\gamma)\rho_\Gamma d\,\Gamma \ =: \mathcal{ND}^{\rm int}_\Gamma \left.\frac{\partial\psi}{\partial n}\right|_\Gamma , \ x\in\Gamma \ ,$$

defines the relative Neumann-to-Dirichlet map which is inverse to the relative Dirichlet-to-Neumann map defined above,

$$\mathcal{ND}_\Gamma \ \mathcal{DN}_\Gamma \ = \ I_\Gamma.$$

For W^2_2 solutions u the corresponding DN-map acts, on the set of all regular spectral points λ of the Neumann Schrödinger, from $W^{1/2}_2(\Gamma)$ onto $W^{3/2}_2(\Gamma)$. For $W^{3/2}_2$ solutions the ND-map acts acts from $L_2(\Gamma)$ onto $W^1_2(\Gamma)$.

The coefficients of the scattering Ansatz (5) can be found, in principle, from the infinite linear system which is obtained by substitution of the scattering Ansatz into the matching condition (see [44]). An important part of the calculation is the proof of the formula for the DN-map in terms of the G^D_Γ (see [44]), or, respectively, a similar formula for the ND–map in terms of G^N_Γ. Selecting $E_\pm$ as indicated

above, (8), represent the ND-map of L_Γ^N by 2×2 operator matrix with elements $\mathcal{ND}_{\pm,\pm} = P_\pm \mathcal{ND}_\Gamma P_\pm$

$$\mathcal{ND}_\Gamma = \begin{pmatrix} \mathcal{ND}_{++} & \mathcal{ND}_{+-} \\ \mathcal{ND}_{-+} & \mathcal{ND}_{--} \end{pmatrix}. \tag{15}$$

The similar decomposition of the DN-map of the Schrödinger operator L_Γ^D on $\Omega_{\rm int}$

$$\mathcal{DN}_\Gamma = \begin{pmatrix} \mathcal{DN}_{++} & \mathcal{DN}_{+-} \\ \mathcal{DN}_{-+} & \mathcal{DN}_{--} \end{pmatrix} \tag{16}$$

was used in [44] in the course of construction of a convenient representation for the scattering matrix on the open spectral bands. We set, in agreement with the above notations in (4,5,6):

$$\begin{aligned} K_+ &= \textstyle\sum_m \sum_{\rm open} \sqrt{\lambda-\lambda_l}\, e_l^m\rangle\langle e_l^m = \sum_m \sqrt{\lambda-\frac{\pi^2}{\delta^2}}\, e_1^m\rangle\langle e_1^m\ , \\ K_- &= \textstyle\sum_m \sum_{\rm closed} \sqrt{\lambda_l-\lambda}\, e_l^m\rangle\langle e_l^m = \sum_m \sum_{l\geq 2} \sqrt{\lambda_l-\lambda}\, e_l^m\rangle\langle e_l^m\ . \end{aligned}$$

Hereafter we use the standard bracket notations, $e\rangle\langle e' : u \to e\,\langle e',\, u\rangle$, with the bar on the first factor of the dot-product in $E = L_2(\Gamma)$. The exponents of oscillating and decreasing modes on the first spectral band spanned by the vectors $e_\pm \in E_\pm$ are represented as:

$$\chi_\pm e_+ = e^{\pm iK_+ x} e_+\ , \qquad \xi_- e_- = e^{-K_- x} e_-\ .$$

The matrices $S_{l,r}^{m,n}$ and $s_{l,r}^{m,n}$, which are defined by the matching of the scattering Ansatz to the solution of the homogeneous equation on $\Omega_{\rm int}$, constitute respectively the *scattering matrix* – the square table of amplitudes in front of the oscillating modes in open channels ($l = 1$):

$$S = \sum_{m,n=1}^{M} \sum S_{1,1}^{m,n} e_1^m\rangle\ \langle e_1^n\ ,$$

and the table of amplitudes in front of the evanescent modes

$$s = \sum_{m,n=1}^{M} \sum_{1,\, r\geq 2} s_{1,r}^{m,n} e_1^m\rangle\ \langle e_r^n\ .$$

The scattering matrix of the junction is represented (see [44] and Theorem 2.1 below) in terms of the matrix elements $\mathcal{DN}$, $\mathcal{ND}$ combined in aggregates

$$\mathcal{M} = \mathcal{DN}_{++} - \mathcal{DN}_{+-} \frac{I}{\mathcal{DN}_{--} + K_-} \mathcal{DN}_{-+} \tag{17}$$

$$\mathcal{N} = \mathcal{ND}_{++} - \mathcal{ND}_{+-} K_- \frac{I}{I_- + \mathcal{ND}_{--} K_-} \mathcal{ND}_{-+}\ , \tag{18}$$

The width δ of the leads can serve as a small parameter in the course of calculation of the scattering matrix. Thin networks, with small δ, are characterized by large

distance between the neighboring spectral thresholds:

$$\frac{\pi^2(l+1)^2}{\delta^2} - \frac{\pi^2 l^2}{\delta^2} \;=\; \frac{(2l+1)\pi^2}{\delta^2}\ .$$

One can prove following [44] that, for a "thin junction", the denominator $\mathcal{DN}_{--}+K_-$ is invertible on a major part of a properly selected auxiliary spectral interval Δ, where the DN-map is represented as a sum of a rational function and a regular correcting term:

$$\mathcal{DN}_\Gamma = \sum_{\lambda_s\in\Delta} \frac{\frac{\partial\varphi_s}{\partial n}\rangle\langle\frac{\partial\varphi_s}{\partial n}}{\lambda_s-\lambda} + \mathcal{K}^{\Delta_1} =: \mathcal{DN}^\Delta + \mathcal{K}^\Delta, \tag{19}$$

The zeros of the denominator $\mathcal{DN}_{--}+K_-$ on Δ have an important operator-theoretic meaning: they are eigenvalues of the *intermediate Hamiltonian*. Hereafter we consider the rational approximation (19) and the corresponding rational approximation of $\mathcal{DN}_{--}=P_-\mathcal{DN}P_-$:

$$\mathcal{DN}_{--} = \mathcal{DN}^\Delta_{--} + \mathcal{K}^\Delta_{--}, \tag{20}$$

with a regular "error" $\mathcal{K}^\Delta_{--}$ on a complex neighborhood $G(\Delta)$ of Δ.

We call the junction Ω thin in closed channels, either in $W_2^1(\Gamma)$ or in $W_2^{3/2}(\Gamma)$, if, respectively,

$$\sup_\Delta \parallel K_-^{-1}\mathcal{K}^\Delta_{--} \parallel_{W_2^1(\Gamma)} < 1\ , \qquad or \qquad \sup_\Delta \parallel K_-^{-1}\mathcal{K}^\Delta_{--} \parallel_{W_2^{3/2}(\Gamma)} < 1 \tag{21}$$

This implies the following statement (see [44]):

Lemma 2.1. *If the junction is thin on closed channels, then the denominator of* (17) *is invertible*

$$\left[\mathcal{K}^\Delta_{--}+K_-\right]^{-1} : L_2(\Gamma)\to W_2^1(\Gamma)\ ,$$

on a corresponding "major part of the essential spectral interval" – the complement of the set of zeros $Z_\Delta\subset\Delta$ of the determinant of the finite-dimensional matrix-function:

$$Z_\Delta = \left\{\lambda : det\left[I+\left(\mathcal{K}^\Delta_{--}+K_-\right)^{-1}\mathcal{DN}^\Delta_{--}(\lambda)\right]=0\right\}$$

A similar statement holds for the above denominator as an operator from $W_2^{1/2}(\Gamma)$ to $W_2^{3/2}(\Gamma)$.

Theorem 2.1. *The substitution of the scattering Ansatz* (5) *into the matching conditions on Γ gives the following formulae for the scattering matrix on $\hat\Delta$*

$$S = \left[iK_+ + \mathcal{M}\right]^{-1}\left[iK_+ - \mathcal{M}\right]\ , \tag{22}$$

$$S = \left[\mathcal{N}iK_+ + 1\right]^{-1}\left[\mathcal{N}iK_+ - 1\right]\ . \tag{23}$$

Proof. The scattering Ansatz generated by the entrance vector $e \in E_+$ is constituted by the incoming wave $e^{iK_+ x}e$, the transmitted/reflected wave $e^{-iK_+ x}Se$ and the evanescent wave $e^{-K_- x}se$:

$$\Psi_e \;=\; e^{iK_+ x}e + e^{-iK_+ x}Se + e^{-K_- x}se \;.$$

The boundary data of the scattering Ansatz at the bottom sections Γ should match on Γ the boundary data of the solution of the homogeneous Schrödinger equation inside $\Omega_{\rm int}$:

$$L_{\rm int}\psi \;=\; \lambda\psi, \qquad \psi\Big|_{\partial\Omega_{\rm int}\setminus\Gamma} \;=\; 0 \;,$$

$$\psi\big|_\Gamma = \psi_e(0) = e+Se+se \;, \qquad \frac{\partial\psi}{\partial n}\Big|_\Gamma \;=\; \psi'_e(0) \;=\; iK_+e - iK_+Se - K_-se \;. \quad (24)$$

Using the matrix representations (16, 15) for $\mathcal{DN}$, $\mathcal{ND}$, we obtain from (24) two equivalent linear systems which describe matching conditions on Γ

$$\begin{aligned} iK_+(1-S)e &= \mathcal{DN}_{++}(1+S)e + \mathcal{DN}_{+-}se \;, \\ -K_-se &= \mathcal{DN}_{-+}(1+S)e + \mathcal{DN}_{--}se \;, \end{aligned} \quad (25)$$

and

$$\begin{aligned} (I+S)e &= \mathcal{ND}_{++}\, i\,K_+(I-S)e - \mathcal{ND}_{+-}K_-se, \\ [I+\mathcal{ND}_{--}K_-]se &= \mathcal{ND}_{-+}\, i\,K_+(I-S)e\mathcal{ND}_{--}se. \end{aligned} \quad (26)$$

Eliminating the component se from them and using the former notations $\mathcal{M}$, $\mathcal{N}$ we obtain the announced representation for the scattering matrix (22, 23). □

Consider the operator $\mathcal{L}$ defined by the above Schrödinger differential expression on the junction $\Omega = \Omega_{\rm int} \cup \omega$, with zero Dirichlet condition on the boundary $\partial\Omega$. It is essentially selfadjoint on the domain of smooth functions u, subject to the Meixner restriction $u \in W_2^1(\Omega)$. Assume that the entrance space $E = L_2(\Gamma)$ on the cross-sections Γ is decomposed as $E_+ \oplus E_-$. We use the former notations $P_\pm$ for the orthogonal projections in E onto $E_\pm$. Consider the Glazman splitting $\mathcal{L}_\Gamma$ obtained from $\mathcal{L}$ by imposing an additional *partial zero boundary condition* on the bottom sections Γ of the leads:

$$P_+u\big|_\Gamma \;=\; 0 \;, \quad (27)$$

complemented by the standard smooth matching condition on Γ in closed channels. The operator $\mathcal{L}$ is split by this boundary conditions into an orthogonal sum of two operators:

$$\mathcal{L} \longrightarrow L_\Lambda \oplus l_\Lambda \;=\; \mathcal{L}_\Lambda \;.$$

Here $l_\Lambda = -\dfrac{d^2}{dx^2} + \frac{\pi^2}{\delta^2} + V_\delta$ in $L^2(0,\infty)\times E_+) =: \mathcal{H}_+$, with zero boundary condition at the origin $u(0) = 0$, and L_Λ is defined in the orthogonal sum of the channel space $L^2(0,\infty)\times E_- =: \mathcal{H}_-$ of the closed channels and $L_2(\Omega_{\rm int})$ on W_2^2-smooth

functions, subject to the Meixner condition and the matching condition on Γ in closed channels:

$$L_\Lambda : D(\Lambda) \longrightarrow L_2(\Omega_{\rm int}) \oplus \mathcal{H}_- \, .$$

Theorem 2.2. *The operators L_Λ, l_Λ are essentially selfadjoint. The absolutely continuous components of spectra of the corresponding selfadjoint extensions are*

$$\begin{aligned} \sigma_a(l_\Lambda) &= [\lambda_1, \infty), \qquad \text{with multiplicity } M, \\ \sigma_a(L_\Lambda) &= \bigcup_{l=2}^{\infty} [\lambda_l, \infty) =: \bigcup_{l\geq 2} \sigma_a^l \, . \end{aligned} \tag{28}$$

where each branch σ_a^l has multiplicity M, and the total multiplicity is growing stepwise on the thresholds λ_l separating the spectral bands $\Delta_l = [\lambda_l, \lambda_{l+1}]$. The spectral multiplicity of the absolutely continuous spectrum of L_Λ on the spectral bands Δ_l is equal to $Ml(l+1)/2$. The discrete spectrum of L_Λ consists of a countable set of eigenvalues λ_s^Λ accumulating at infinity. The singular spectrum of L_Λ is empty.

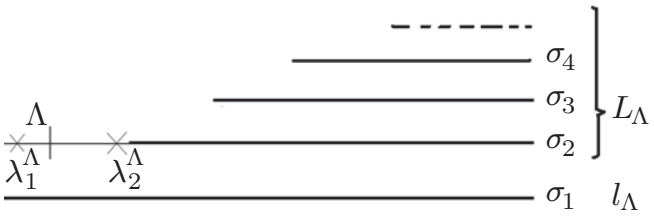

FIGURE 6. The intermediate Hamiltonian L_Λ inherits the closed branches of the continuous spectrum of the unperturbed operator. The part l_Λ of the split operator inherits the first – open – branch of the spectrum of the split operator. The resonance eigenvalues of the intermediate Hamiltonian define the resonance conductance of the junction.

The relation $\mathcal{M}\mathcal{N} = I$ observed from comparison of the formulæ (23, 22) has an important operator-theoretic meaning. It is derived from the fact that $\mathcal{M}, \mathcal{N}$ are respectively DN and ND-maps of the intermediate Hamiltonian – the part L^0 of the Glazman splitting

$$\mathcal{L} \longrightarrow \mathcal{L}_\Lambda \;=\; L_\Lambda \oplus l_\Lambda. \tag{29}$$

defined by the partial boundary condition $P_+ u\big|_\Gamma = 0$ – see [46]. Contrary to the standard splitting $\mathcal{L} \longrightarrow L_{\rm int} \oplus L^\omega$, this splitting (29) is finite-dimensional – see [4]. The poles of $\mathcal{M}$ on the first spectral band, below $\lambda_{\min}$, are the eigenvalues of $_\Lambda$.

3. Krein formulae for the intermediate DN-map and ND-map, with the compensated singularities

Expressions $\mathcal{M}$, $\mathcal{N}$ in the formulae (17,18) contain, at least formally, the singularities at the eigenvalues of the operators L_Γ^D, L_Γ^N. Presence of these singularities in the conditions of the last theorem of the previous section looks strange. In fact we were able to prove, see [3], that the singularities in the first and second terms of the above Krein formula for $\mathcal{M}$, inherited from L_Γ^D, compensate each other, so that only the singularities of the denominators of (17) play a role. Similar statement can be proved, see Theorem 3.3 below, for $\mathcal{N}$. But in fact the compensation of singularities permits to obtain more convenient representations for $\mathcal{M}$, $\mathcal{N}$ – that is for DN and ND maps of the intermediate Hamiltonian L_Λ. These new representations imply also the corresponding exact formulae for the scattering matrix of $\mathcal{L}$, and convenient approximate expressions for the scattering matrix as well. This approximate expressions can serve a base for construction of a fitted solvable model of the junction in form of star-shaped 1D quantum graph, and for derivation of the boundary condition at the vertex.

We begin with the discussion of compensation of singularities in the formula (17) for the DN-map $\mathcal{M}$ of the intermediate Hamiltonian. It appeared, that the singularities of the first and second term at the eigenvalues of $L_{\rm int}$ compensate each other, so that only the zeros of the denominator $\mathcal{DN}_{--} + K_-$ arise as singularities of $\mathcal{DN}^\Lambda$ on Δ. A one-dimensional version of the statement can be found in [10] and a rescription of the classical Krein formula with compensated singularities is given in [45]. In this paper we review the compensation singularities in Theorem 3.1, following [3] for a general thin junction and prove a similar statement, see Theorem 3.3 for the intermediate ND-map. We also obtain, in course of calculations, an important "byproduct": a version of analytic perturbation procedure for groups of eigenpairs. Note that the standard analytic perturbation procedure is aimed on calculation of an individual perturbed eigenvalue.

Usually the convergence of the corresponding perturbation series is limited by the condition of non-intersection of corresponding terms $\lambda_s(\varepsilon) \neq \lambda_t(\varepsilon)$. Our technique, based on compensation of singularities, can be used even to study overlapping terms and to study the transformation, under small perturbations, of intersections of terms into quasi-intersections. This technique is aimed not on the calculation of an individual perturbed eigenvalue, but rather on derivation of an approximate algebraic equation for the perturbed eigenvalues and calculation of the corresponding residues at the poles of the perturbed DN-map. We also calculate, based on (22), the scattering matrix of a "relatively thin" junction in the quantum network. We also develop similar technique for ND-map. In fact our technique can be modified to calculate the scattering matrix for arbitrary junction, see [3].

For given temperature T we consider an *essential spectral interval* centered at the scaled Fermi level Λ : $\Delta_T = [\Lambda - 2m^* T\hbar^{-2}, \Lambda + 2m^* T\hbar^{-2}]$. We assume that the temperature is *low*, so that Δ_T is situated inside the auxiliary spectral

interval or an open spectral set Δ

$$\Delta_T \subset \Delta \subset \left(\frac{\pi^2}{\delta^2} + V_\delta,\ \frac{4\pi^2}{\delta^2} + V_\delta\right) =: \Delta_1.$$

Our prime aim is: to construct on Δ_T a convenient local "quasi-one-dimensional" representation for the intermediate DN-map and one for the scattering matrix[1] of the junction *with compensated singularities* inherited from the $L_{\rm int}$, to substitute previous formula (22).

Selecting an appropriate spectral interval (or just a spectral set) $\Delta : \Delta_T \subset \Delta$, we represent the DN-map $\mathcal{DN}$ of $L_{\rm int}$ on the essential spectral interval Δ_T as a sum

$$\mathcal{DN}_{\rm int} = \sum_{\lambda_s \in \Delta} \frac{\frac{\partial \varphi_s}{\partial n}\Big|_\Gamma \rangle\langle \frac{\partial \varphi_s}{\partial n}\Big|_\Gamma}{\lambda - \lambda_s} + \mathcal{K}^\Delta =: \mathcal{DN}^\Delta\ + \mathcal{K}^\Delta. \tag{30}$$

of the rational expression constituted by the polar terms with singularities at the eigenvalues $\lambda_s \in \Delta$, $s = 1, 2, \dots, N$ of the operator $L_{\rm int}$ and the analytic operator function $\mathcal{K}^\Delta$ on a complex neighborhood G_{Δ_T} of Δ_T.

We will also use the operators obtained from $\mathcal{DN}_{\rm int}$ via framing it by the projections $P_\pm$, for instance:

$$P_+\mathcal{DN}_{\rm int}P_- = P_+\mathcal{DN}^\Delta P_- + P_+\mathcal{K}^\Delta P_- = \mathcal{DN}^\Delta_{+-}\ +\ \mathcal{K}^\Delta_{+-}.$$

We introduce also the linear hull $E^\Delta = \bigvee_{s=1}^N \{\varphi_s\}$ – an invariant subspace of $L_{\rm int}$, $\dim E^\Delta = N$, corresponding to the spectrum of $L_{\rm int}$ contained in Δ and the part $L^\Delta := \sum_{\lambda_s \in \Delta} \lambda_s \varphi_s\rangle\langle\varphi_s$ of $L_{\rm int}$ in it. To calculate the intermediate DN-map $\mathcal{M}$ in terms of the standard DN-map of $L_{\rm int}$ we have to solve, see (3.1) the equation:

$$[\mathcal{DN}_{--} + K_-]u = \mathcal{DN}_{-+}g \tag{31}$$

on the essential spectral interval Δ_T. It can be solved based on Banach principle if K_- can play a role of a large parameter.

Definition. *The junction, for which the operator*

$$\left[\mathcal{K}^\Delta_{--} + K_-\right]^{-1} \tag{32}$$

exists on Δ_T is called hereafter "relatively thin junction", for the selected spectral set Δ and given temperature T. A general network is called thin, if all junctions of the network are thin.

For a relatively thin junction, due to continuity of $\mathcal{K}^\Delta_{--}, K_-$ there exist also a complex neighborhood G_{Δ_T} of Δ_T, where $\mathcal{K}^\Delta_{--} + K_-$ is invertible.

Hereafter we assume that the junction is thin. The case of an arbitrary junction is considered in [3].

[1] Compare with the popular one-dimensional formula for the scattering matrix in terms of the Weyl function, derived in [53] and intensely used by B. Simon and F. Gezstezy in their approach to the spectral inverse problem, see [20].

The above definition of thin junctions (and networks) is based on the following motivation. The DN-map of $L_{\rm int}$ is homogeneous degree -1. It acts from $W_2^{3/2}(\Gamma)$ to $W_2^{1/2}(\Gamma)$, see [74]. If $\Omega_{\rm int}$ has a small diameter d then, the norm of the correcting term $\mathcal{K}$ is estimated, generically, on the complement of the spectrum, as Const $1/d$. The same estimate remains true for $P_-\mathcal{K}^\Delta P_- := \mathcal{K}^\Delta_{--}$. The exponent K_- on the essential spectral band Δ acts from $W_2^{3/2}(\Gamma)$ to $W_2^{1/2}(\Gamma)$ and the norm of its inverse is estimated as Const δ. Then the $W_2^{3/2}$-norm of $K_-^{-1}\,\mathcal{K}^\Delta_{--}$ is estimated generically as Const δ/d. Hence, in particular, $K_- + \mathcal{K}^\Delta_{--} = K_-\left[I + K_-^{-1}\,\mathcal{K}^\Delta_{--}\right]$ is invertible if $\delta/d \ll 1$, see more comments in [44]. Notice, that for *arbitrary junction* an auxiliary Fermi level $\Lambda_1^F := \Lambda_1$ can be selected, see [3], such that the condition (32) is fulfilled. Now we proceed assuming that (32) is fulfilled.

Consider the part $L^\Delta_{\rm int}$ of $L_{\rm int}$ in the subspace $E^\Delta = \oplus\sum_{\lambda_s\in\Delta}\varphi_s$:

$$L^\Delta_{\rm int} = \sum_{\lambda_s\in\Delta}\lambda_s\ \varphi_s\rangle\,\langle\varphi_s : E^\Delta \to E^\Delta,\ \dim E^\Delta = N.$$

Assume that $\left\{\frac{\partial\varphi_s}{\partial n}\right\}\Big|_\Gamma$ are linearly independent. Then

$$\dim \bigvee_{\lambda_s\in\Delta}\varphi_s = \dim \bigvee_{\lambda_s\in\Delta}\frac{\partial\varphi_s}{\partial n}\Big|_\Gamma = N.$$

Denote by $\mathcal{T}$ the map

$$\mathcal{T} = \sum_{\lambda_s\in\Delta}\varphi_s\rangle\langle\frac{\partial\varphi_s}{\partial n}\Big|_\Gamma,$$

and introduce

$$\left(P_+ - \mathcal{K}^\Delta_{+-}\frac{I}{\mathcal{K}^\Delta_{--}+K_-}P_-\right) := \mathcal{J}(\lambda) : E \to E_+.$$

It is obvious that $\dim\left\{\mathcal{J}\frac{\partial\varphi_s}{\partial n}\Big|_\Gamma\right\}_s \le \dim E^\Delta$. Later we will utilize a stronger

Assumption 1. *The vectors* $\mathcal{J}\frac{\partial\varphi_t}{\partial n}\Big|_\Gamma$ *are linearly independent in* E_+ *for any* $\lambda\in\Delta$, *hence both:* $\dim\left\{\frac{\partial\varphi_t}{\partial n}\Big|_\Gamma\right\} = \dim E^\Delta = N$ *and*

$$W_J(\lambda) := \det\left\{\left\langle\mathcal{J}\frac{\partial\varphi_s}{\partial n}\Big|_\Gamma,\ \mathcal{J}\frac{\partial\varphi_t}{\partial n}\Big|_\Gamma\right\rangle\right\}^N_{s,t=1}(\lambda) > 0. \tag{33}$$

The above condition (33) is equivalent to the pair of conditions:

1. The functions $\frac{\partial\varphi_s}{\partial n}\Big|_\Gamma$, $s = 1,2,3,\dots,N$ are linearly independent.
2. The operator $\mathcal{J}^+\mathcal{J}$ is invertible in the linear hull $\bigvee_{s=1}^{s=N}\frac{\partial\varphi_s}{\partial n}\Big|_\Gamma$ for any $\lambda\in\Delta$.

Hereafter we reduce the problem of compensation of singularities to the spectral analysis of the Schrödinger-type equation

$$\left[L^\Delta - Q(\lambda)\right]\psi = \lambda\psi$$

in E^Δ with the λ-dependent "potential"

$$Q(\lambda) := \mathcal{T}\frac{I}{\mathcal{K}^\Delta_{--} + K_-}\mathcal{T}^+ : E^\Delta \to E^\Delta.$$

Lemma 3.1. *For thin junction, on the essential spectral interval Δ the derivative $\frac{\partial Q}{\partial\lambda}$ is a positive matrix $N \times N$.*

Proof. Recall that the branch of the square root $\sqrt{\pi^2\,\delta^{-2} + V_\delta - \lambda}$ is defined such that $\frac{dK_-}{d\lambda} < 0$ on the conductivity band. The correcting term $\mathcal{K}^\Delta$ is a meromorphic operator function with a negative imaginary part in the upper half-plane $\Im\lambda > 0$ and a positive imaginary part in the lower half-plane, hence $\frac{d\mathcal{K}^\Delta_{--}}{d\lambda} < 0$ on the conductivity band Δ_1. This implies:

$$\frac{\partial Q}{\partial\lambda} = -\mathcal{T}\frac{I}{\mathcal{K}^\Delta_{--} + K_-}\left[\frac{dK_-}{d\lambda} + \frac{d\mathcal{K}^\Delta_{--}}{d\lambda}\right]\frac{I}{\mathcal{K}^\Delta_{--} + K_-}\mathcal{T}^+ > 0, \text{ for } \lambda \in \Delta_T \in \Delta_1.$$

□

Based on some cumbersome calculation, we are able to derive, see [3], that all singularities in the Krein formula, inherited from the eigenvalues λ_s of the unperturbed operator $L_{\rm int}$, are compensated.

Theorem 3.1 (Compensation of Singularities M). *The Krein formula* (17) *for the intermediate DN-map, can be re-written, for a thin junction, on the spectral interval Δ, as:*

$$\begin{aligned}\mathcal{M} = \mathcal{D}\mathcal{N}^\Lambda &= \mathcal{M}_{\rm reg} + \mathcal{J}\mathcal{T}^+\rangle\frac{I}{\lambda I^\Delta - L^\Delta + Q(\lambda)}\langle\mathcal{T}\mathcal{J}^+ \\ &=: k(\lambda) + \mathcal{J}\mathcal{T}^+\rangle\frac{I}{\lambda I^\Delta - L^\Delta + Q(\lambda)}\langle\mathcal{T}\mathcal{J}^+, \qquad (34)\end{aligned}$$

where $\mathcal{K}^\Delta_{++} - \mathcal{K}^\Delta_{+-}\frac{I}{\mathcal{K}^\Delta_{--}+K_-}\mathcal{K}^\Delta_{-+} =: k(\lambda)$ is a regular part of $\mathcal{M}$ on Δ_T. The representation (34) *remains valid on a complex neighborhood G_Δ of the spectral interval Δ.*

Remark 1. The announced rescription (34) of the Krein formula (17) for the DN-map of the intermediate Hamiltonian, has on the essential spectral interval only non-compensated singularities, at the eigenvalues of the intermediate Hamiltonian, calculated as zeros λ^Q_s of the denominator $\lambda I^\Delta - L^\Delta + Q(\lambda) := \mathbf{d}(\lambda)$:

$$\mathbf{d}(\lambda^Q_s)\ \nu^Q_s = 0.$$

These singularities coincide with the eigenvalues of the intermediate Hamiltonian. We call the above formula (34) for $\mathcal{D}\mathcal{N}^\Lambda$ *the modified Krein formula*. Inserting (34) into the above formula (22) gives a convenient representation for the scattering matrix of the relatively thin junction, which permits, in particular, to calculate

sharp resonances, situated near the continuous spectrum, based on eigenvalues of the intermediate operator.

In the case of one-dimensional zeros of the denominator $\mathbf{d}(\lambda_s^Q)(\nu_s^Q) = 0, \mathcal{J}^+ \mathcal{T}^+ \nu_s^\Lambda \neq 0$, the corresponding residues are calculated as projections onto the subspaces

$$\mathcal{E}_s^Q = \mathcal{J}(\lambda_s^Q)\mathcal{T}^+ \ \nu_s^Q.$$

For multidimensional zeros of the denominator, $\mathbf{d}(\lambda_s^Q)N_s^Q = 0$, $\dim N_s^Q \ > 1$ the residues are projections onto the images of the corresponding null-spaces $N_s^Q = \bigvee_s \nu_s^Q$

$$\mathcal{E}_s^Q = \mathcal{J}(\lambda_s^Q)\mathcal{T}^+ \ N_s^Q.$$

The above expression (34) is analytic in Ω_Δ on the complement of the set of zeros of the denominator $\mathbf{d}(\lambda)$. This means that the eigenvalues of the intermediate Hamiltonian are selected from the set. Assume that λ_s^Λ, $s = 1, 2, 3, \ldots, N$ are simple zeros of the denominator.

Theorem 3.2. *If the Wronskian* (33) *does not vanish at the algebraically simple* (*first-order*) *zero* λ_1^Q *of the denominator,* $W_J(\lambda_1^Q) \neq 0$*, then the zero is an eigenvalue of the intermediate Hamiltonian, with the same spectral multiplicity.*

Proof. It is sufficient to prove, that the zero is a first-order pole of the intermediate DN-map, with a finite-dimensional residue having the same dimension as the zero of the denominator. Consider the equation

$$\mathbf{d}(\lambda)u = \left[\lambda I^\Delta - L^\Delta + Q(\lambda)\right] u = f, \ \lambda \in E^\Delta. \tag{35}$$

Assume that $\mathbf{d}(\lambda_1^Q)e_s^Q = 0$, and denote by P_1^Q an orthogonal projection onto the multiple eigenspace $\bigvee_s e_s^Q$ of the operator $L^\Delta - Q(\lambda_1^Q) =: L_Q^\Delta$, and by R_λ^Q the corresponding resolvent:

$$\left[L^\Delta - Q(\lambda_1^Q) - \lambda I^\Delta\right]^{-1} = R_\lambda^Q(\lambda).$$

Then, for λ close to λ_1^Q we can substitute the "potential" Q in the above equation by the Taylor expansion:

$$Q(\lambda) = Q(\lambda_1^Q) + (\lambda - \lambda_1^Q)\frac{dQ}{d\lambda}(\lambda_1^Q) + \frac{(\lambda - \lambda_1^Q)^2}{2}\frac{d^2Q}{d\lambda^2}(\lambda_1^Q) + \cdots$$

and represent the resolvent near the pole λ_1^Q as

$$R_\lambda^Q = \frac{P_1}{(\lambda_1^Q - \lambda)} + R^Q_{\perp,\lambda_1^Q}, \tag{36}$$

where $R^Q_{\perp,\lambda_1^Q}$ is the part of the resolvent in the complementary invariant subspace $E_1^\perp = (I - P_1)E$. Replacing $Q(\lambda)$ by the corresponding first-order Taylor formula, we rewrite the equation $\mathbf{d}(\lambda)u = f$ by the equation

$$u + R_\lambda^Q(\lambda_1^Q - \lambda)\frac{dQ}{d\lambda}u = -R_\lambda^Q f. \tag{37}$$

To calculate the residue of the solution u at the pole λ_1^Q we multiply (37) by $\frac{dQ}{d\lambda}$ and by the spectral projection P_1^Q of L_Q^Δ at λ_1^Q and take into account that the resolvent R_λ^Q of L_Q^Δ has a simple pole and neglect the terms vanishing at λ_1^Q, in particular all terms arising from the above Taylor expansion beginning from the second $(\lambda-\lambda_1^Q)^2\frac{d^2Q}{d\lambda^2}(\lambda_1^Q)$. Due to positivity of $\frac{dQ}{d\lambda}$ the operator $I+P_1^Q\frac{dQ}{d\lambda}P_1^Q$ is invertible. Then, due to (36)

$$P_1^Q\frac{dQ}{d\lambda}u=\frac{I}{I+P_1^Q\frac{dQ}{d\lambda}P_1^Q}P_1^Q\frac{dQ}{d\lambda}R_\lambda^Q f$$

has the polar part at λ_1^Q

$$u=-\frac{P_1^Q\left[I+P_1^Q\frac{dQ}{d\lambda}P_1^Q\right]^{-1}P_1^Q}{\lambda_1^Q-\lambda}f+\cdots. \tag{38}$$

The operator in the square bracket is positive and hence has the spectral form:

$$P_1^Q\left[I+P_1^Q\frac{dQ}{d\lambda}P_1^Q\right]^{-1}P_1^Q=\sum_r \alpha_r\ \nu_r\rangle\,\langle\nu_r.$$

Then the polar term of $\mathbf{d}^{-1}=-\left[L_Q^\Delta(\lambda)\right]^{-1}$ at λ_1^Q is $(\lambda-\lambda_1^Q)^{-1}\sum_r \alpha_r\ \nu_r\rangle\,\langle\nu_r$ and the pole part of the intermediate DN-map at λ_1^Q is

$$\mathcal{DN}^\Lambda_{pole}=\frac{\sum_r \mathcal{JT}^+\ \nu_r\rangle\,\alpha_r\,\langle\mathcal{JT}^+\ \nu_r}{\lambda-\lambda_1^Q}.$$

Due to the orthogonality of ν_s and non-degeneracy of the Wronskian W_Δ^Q, the vectors $\mathcal{J}T^+\ \nu_r$ are linearly independent, hence λ_1^Q is a simple pole of the intermediate DN-map, with the spectral multiplicity $\dim\bigvee_s\ e_s^Q$. □

The scattering matrix of the original problem on the essential spectral interval can be obtained via replacement in (22) the intermediate DN-map by the expression (34) with compensated singularities. This substitution is possible for thin junctions, when the exponent K_- in closed channels can play a role of a large parameter, compared with the error $\mathcal{K}_{--}^\Delta$ of the rational approximation $\mathcal{DN}^\Delta$ of $\mathcal{DN}$.

This condition may be not satisfied, for given quantum network, at the scaled Fermi level Λ. In that case another representation of the scattering matrix (23) can help. We consider now the problem of compensation singularities for the intermediate ND-map $\mathcal{N}$ on the essential spectral interval.

Denote by ψ_s,λ_s^N the eigenpairs of the operator L_Γ^N, see (13). Select the eigenvalues from the spectral interval Δ_2, to be defined later, and introduce $E_{\Delta_2}^N:=\bigvee_{\lambda_s^N\in\Delta_2}\psi_s$ and $E_\Gamma^N:=\bigvee_{\lambda_s^N\in\Delta_2}\psi_s\big|_\Gamma$, and consider the map

$$\tilde{\mathcal{T}}:\sum_{\lambda_s^N\in\Delta_2}\psi_s\big|_\Gamma\rangle\,\langle\psi_s:\ E_{\Delta_2}\to E_\Gamma^N, s=1,2,\dots,\tilde{N} \tag{39}$$

Assumption 2.

1. *We assume that the families* $\{\psi_s\}_{s=1}^{\tilde N}$, $\{\psi_s\}_{s=1}^{\tilde N}\big|_\Gamma$ *are linearly independent, thus are bases in their linear hulls* $\tilde E^N_{\Delta_2}$, $\tilde E^N_\Gamma$, $\dim \tilde E^N_{\Delta_2} = \dim \tilde E^N_\Gamma = \tilde N$.

Represent the compact in $L_2(\Gamma)$ relative ND-map $\mathcal{ND}^\Gamma_{\rm int} =: \mathcal{ND}^\Gamma$ as

$$\mathcal{ND}^\Gamma_{\rm int} = \begin{pmatrix} \mathcal{ND}_{++} & \mathcal{ND}_{+-} \\ \mathcal{ND}_{-+} & \mathcal{ND}_{--} \end{pmatrix} = \sum_{\lambda^N_s \in \Delta_2} \frac{\psi_s\big|_\Gamma\rangle\langle\psi_s\big|_\Gamma}{\lambda^N_s - \lambda} + \tilde{\mathcal{K}}^{\Delta_2} =: \mathcal{ND}^{\Delta_2} + \tilde{\mathcal{K}}^{\Delta_2},$$

and consider the corresponding matrices

$$\mathcal{ND}^{\Delta_2} = \begin{pmatrix} \mathcal{ND}^{\Delta_2}_{++} & \mathcal{ND}^{\Delta_2}_{+-} \\ \mathcal{ND}^{\Delta_2}_{-+} & \mathcal{ND}^{\Delta_2}_{--} \end{pmatrix},$$

and

$$\tilde{\mathcal{K}}^{\Delta_2} = \begin{pmatrix} \tilde{\mathcal{K}}^{\Delta_2}_{++} & \tilde{\mathcal{K}}^{\Delta_2}_{+-} \\ \tilde{\mathcal{K}}^{\Delta_2}_{-+} & \tilde{\mathcal{K}}^{\Delta_2}_{--}, \end{pmatrix} =: \begin{pmatrix} \tilde{\mathcal{K}}_{++} & \tilde{\mathcal{K}}_{+-} \\ \tilde{\mathcal{K}}_{-+} & \tilde{\mathcal{K}}_{--}, \end{pmatrix} =: \tilde{\mathcal{K}},$$

in the basis E_+, E_- of $E = L_2(\Gamma)$. Hereafter we omit the upper index Δ_2 on matrix elements $\tilde{\mathcal{K}}_{\pm\,\pm}$. To represent $\mathcal{ND}^\Lambda =: \mathcal{N}$ in the form with already compensated singularities inherited from the resolvent of L^N, we have to solve the equation

$$(I + \mathcal{ND}_{--}\, K_-)\, u = \mathcal{ND}_{-+} f \tag{40}$$

Our second basic assumption is the following:

2. *We assume, that the width* δ *of the leads, the essential spectral interval* $\Delta_T =: \Delta \subset \Delta_2$ *and the rational approximation* $\tilde{\mathcal{K}}$ *are selected such that*

$$I + \tilde{\mathcal{K}}^{\Delta_2}_{--}\, K_- \tag{41}$$

is invertible for $\lambda \in \Delta$.

For low temperature (that is for a relatively small essential spectral interval Δ_T) this condition is equivalent to the corresponding condition imposed just at the scaled Fermi level Λ. It is satisfied, if

$$\sup_{\lambda\in\Delta} \parallel \tilde{\mathcal{K}}_{--} K_- \parallel < 1. \tag{42}$$

We will not give here a formal condition which guarantees 2, but just notice that due to compactness of the resolvent on L^N_Γ for any Δ_T there exist $\Delta_2 \supset \Delta_T$ and the corresponding number $\tilde N_{\Delta_2} =: \tilde N$ such that the error of the finite rational approximation of the resolvent is small:

$$\begin{aligned} G^N(x,s,\lambda) &= G^N(x,s,\mu) + (\lambda-\mu) Q_2^{\Delta_2} \qquad (43)\\ &= G^N(x,s,\mu) + (\lambda-\mu)\sum_{l=1}^{\tilde N} \frac{\varphi_l(x)\rangle\,\langle\varphi_l(s)}{(\lambda_l-\mu)^2} + (\lambda-\mu)^2 \sum_{l=\tilde N+1}^{\infty} \frac{\varphi_l(x)\rangle\,\langle\varphi_l(s)}{(\lambda_l-\lambda)(\lambda_l-\mu)^2} \\ &= G^N(x,s,\mu) + \tilde Q_2(x,s,\lambda,\mu) + \tilde{\mathcal{K}}^{\Delta_2}(x,s,\lambda,\mu) =: \tilde Q(x,s,\lambda,\mu) + \tilde{\mathcal{K}}(x,s,\lambda,\mu). \end{aligned}$$

Here we choose μ large negative, so that $G^N(x,s,\mu)$ is a kernel of a small integral operator and denote hereafter

$$G^N(x,s,\mu)+\tilde{\mathcal{K}}^{\Delta_2}(x,s,\lambda,\mu)=:\tilde{\mathcal{K}}(x,s,\lambda,\mu),\ \tilde{Q}_2(x,s,\lambda,\mu)=\tilde{\mathcal{N}\mathcal{D}}(x,s,\lambda,\mu).$$

Then for the error $P_-\tilde{\mathcal{K}}P_-=:\tilde{\mathcal{K}}_{--}$ of the rational approximation $\tilde{Q}$ framed by the projections onto E_- the corresponding estimate (42) is valid. Denote

$$\begin{aligned}
\tilde{\mathcal{K}}_{++}-\tilde{\mathcal{K}}_{+-}K_-(I+\tilde{\mathcal{K}}_{--}K_-)^{-1}\tilde{\mathcal{K}}_{-+}&=:\tilde{\mathcal{N}}_{\mathrm{reg}},\\
\left\{P_+-\tilde{\mathcal{K}}_{+-}K_-(I+\tilde{\mathcal{K}}_{--}K_-)^{-1}\right\}\tilde{\mathcal{T}}^+&=:\tilde{\mathcal{J}}\tilde{\mathcal{T}}^+,\\
\tilde{\mathcal{T}}K_-(I+\tilde{\mathcal{K}}_{--}K_-)^{-1}\tilde{\mathcal{T}}^+&=:V(\lambda)\\
L^{\Delta_2}-\lambda I^{\Delta_2}+V(\lambda)&=:L^{\Delta_2}(\lambda),\\
\tilde{\mathcal{T}}\left\{P_+-K_-(I+\tilde{\mathcal{K}}_{--}K_-)^{-1}\tilde{\mathcal{K}}_{-+}\right\}&=:\tilde{\mathcal{T}}\tilde{\mathcal{J}}^+.
\end{aligned}\tag{44}$$

Theorem 3.3 (Compensation of Singularities N).

$$\mathcal{N}=\mathcal{N}_{\mathrm{reg}}-\tilde{\mathcal{J}}\tilde{\mathcal{T}}^+\frac{I}{L^{\Delta_2}(\lambda)}\tilde{\mathcal{T}}\tilde{\mathcal{J}}^+.\tag{45}$$

Proof. We treat the equation (40) with use of the Banach principle under assumption (42):

$$K_-u+K_-(I+\tilde{\mathcal{K}}_{--}K_-)^{-1}\mathcal{N}\mathcal{D}^{\Delta_2}_{--}K_-u=K_-(I+\tilde{\mathcal{K}}_{--}K_-)^{-1}\mathcal{N}\mathcal{D}^{\Delta_2}_{-+}f.\tag{46}$$

Recall that $\mathcal{N}\mathcal{D}=\mathcal{N}\mathcal{D}^{\Delta_2}+\tilde{\mathcal{K}}$ and notice that $\mathcal{N}\mathcal{D}^{\Delta_2}$ is connected with the part L^{Δ_2} of L^N in E^{Δ_2} as

$$\mathcal{N}\mathcal{D}^{\Delta_2}=\tilde{\mathcal{T}}^+\frac{I}{L^{\Delta_2}-\lambda I^{\Delta}}\tilde{\mathcal{T}}.$$

Then, denoting

$$\frac{I}{L^{\Delta_2}-\lambda I^{\Delta_2}}\tilde{\mathcal{T}}K_-u=:v$$

we rewrite the above equation (46) as an equation for v and obtain the solution of it in terms of the inverse of the matrix $L^{\Delta_2}-\lambda I^{\Delta_2}+\mathcal{T}K_-(I+\mathcal{K}^{\Delta_2}_{--}\tilde{K^{\Delta_2}_{-+}})^{-1}\tilde{\mathcal{T}}=:L^{\Delta_2}(\lambda)$:

$$v=[L^{\Delta_2}(\lambda)]^{-1}V\frac{I}{L^{\Delta_2}-\lambda I^{\Delta_2}}\tilde{\mathcal{T}}P_+f.$$

Substituting the result into (46)

$$\begin{aligned}
K_-u=K_-(I+\tilde{\mathcal{K}}_{--}K_-)^{-1}&\left\{\left[\tilde{\mathcal{T}}^+\frac{I}{L^{\Delta_2}-\lambda I^{\Delta_2}}\tilde{\mathcal{T}}+\tilde{\mathcal{K}}_{-+}\right]P_+f\right.\\
&\left.-\tilde{\mathcal{T}}^+[L^{\Delta_2}(\lambda)]^{-1}\tilde{\mathcal{T}}K_-(I+\tilde{\mathcal{K}}_{--}K_-)^{-1}\left[\tilde{\mathcal{T}}^+\frac{I}{L^{\Delta_2}-\lambda I^{\Delta_2}}\tilde{\mathcal{T}}+\tilde{\mathcal{K}}_{-+}\right]P_+f.\right.
\end{aligned}$$

Then

$$\mathcal{N}f = P_+\tilde{\mathcal{T}}^+\frac{I}{L^{\Delta_2}-\lambda I^{\Delta_2}}\tilde{\mathcal{T}}P_+f + \tilde{\mathcal{K}}_{++}f - P_+\left[\tilde{\mathcal{T}}^+\frac{I}{L^{\Delta_2}-\lambda I^{\Delta_2}}\tilde{\mathcal{T}}P_- + \tilde{\mathcal{K}}_{+-}\right]$$
$$\times K_-(I+\tilde{\mathcal{K}}_{--}K_-)^{-1}\left\{\left[\tilde{\mathcal{T}}^+\frac{I}{L^{\Delta_2}-\lambda I^{\Delta_2}}\tilde{\mathcal{T}}+\tilde{\mathcal{K}}\right]P_+f\right.$$
$$\left.-\tilde{\mathcal{T}}^+[L^{\Delta_2}(\lambda)]^{-1}\tilde{\mathcal{T}}K_-(I+\tilde{\mathcal{K}}_{--}K_-)^{-1}\left[\tilde{\mathcal{T}}^+\frac{I}{L^{\Delta_2}-\lambda I^{\Delta_2}}\tilde{\mathcal{T}}+\tilde{\mathcal{K}}_{-+}\right]P_+f\right\}.$$

Leading terms inside parentheses give:

$$K_-(I+\tilde{\mathcal{K}}_{--}K_-)^{-1}\tilde{\mathcal{T}}^+\left[\frac{I}{L^{\Delta_2}-\lambda I^{\Delta_2}}-\frac{I}{L^{\Delta_2}(\lambda)}V\frac{I}{L^{\Delta_2}-\lambda I^{\Delta_2}}\right]$$
$$= K_-(I+\tilde{\mathcal{K}}_{--}K_-)^{-1}\tilde{\mathcal{T}}^+\frac{I}{L^{\Delta_2}(\lambda)}. \tag{47}$$

Taking into account the leading term of the first addendum, we obtain:

$$P_+\tilde{\mathcal{T}}^+\frac{I}{L^{\Delta_2}-\lambda I^{\Delta_2}}\tilde{\mathcal{T}}P_+f - \tilde{\mathcal{T}}^+\frac{I}{L^{\Delta_2}-\lambda I^{\Delta_2}}V\,\frac{I}{L^{\Delta_2}(\lambda)}(\lambda)\tilde{\mathcal{T}}P_+f$$
$$= P_+\tilde{\mathcal{T}}^+\frac{I}{L^{\Delta_2}(\lambda)}\tilde{\mathcal{T}}P_+f. \tag{48}$$

Lower-order terms containing $\frac{I}{L^{\Delta_2}-\lambda I^{\Delta_2}}$ in parentheses give:

$$-\tilde{\mathcal{K}}_{+-}K_-(I+\tilde{\mathcal{K}}_{--}K_-)^{-1}\tilde{\mathcal{T}}^+\left[\frac{I}{L^{\Delta_2}-\lambda I^{\Delta_2}}-\frac{I}{L^{\Delta_2}(\lambda)}V\right]\tilde{\mathcal{T}}P_+f$$
$$= -\tilde{\mathcal{K}}_{+-}K_-(I+\tilde{\mathcal{K}}_{--}K_-)^{-1}\tilde{\mathcal{T}}^+\frac{I}{L^{\Delta_2}(\lambda)}\tilde{\mathcal{T}}P_+f, \tag{49}$$

and the adjoint expression. The term which contains only the main singularity $[L^{\Delta_2}(\lambda)]^{-1}$ is

$$\tilde{\mathcal{K}}_{+-}K_-(I+\tilde{\mathcal{K}}_{--}K_-)^{-1}\,\tilde{\mathcal{T}}^+\frac{I}{L^{\Delta_2}(\lambda)}\tilde{\mathcal{T}}\;K_-(I+\tilde{\mathcal{K}}_{--}K_-)^{-1}\tilde{\mathcal{K}}_{-+}$$

The terms which do not contain singularities result in:

$$\tilde{\mathcal{K}}_{++}-\tilde{\mathcal{K}}_{+-}K_-(I+\tilde{\mathcal{K}}_{--}K_-)^{-1}\tilde{\mathcal{K}}^{\Delta}_{-+} =: \mathcal{N}_{\mathrm{reg}} \tag{50}$$

Note that the operator $K_-(I+\tilde{\mathcal{K}}_{--}K_-)^{-1}$ is selfadjoint on the first spectral band. Then, collecting all terms we obtain the announced result. □

Based on Theorems 3.1, 3.3 we can calculate the scattering matrix either in the form (22) or in the form (23). One of these formulae can be more convenient on the essential spectral interval, than another, depending on localization of singularities of $\mathcal{D}\mathcal{N}^{\Lambda}$ and $\mathcal{N}\mathcal{D}^{\Lambda}$. Luckily, due to to $\mathcal{D}\mathcal{N}^{\Lambda}\;\mathcal{N}\mathcal{D}^{\Lambda} = I_+$, the singularities of the factors do not overlap, hence for any point $\lambda_0 \in \Delta$ one can select an interval centered at λ_0 where at least one of the factors $\mathcal{D}\mathcal{N}^{\Lambda}$ or $\mathcal{N}\mathcal{D}^{\Lambda}$ can be substituted by the corresponding approximate expression based on the bi-linear formulae suggested in [61].

4. Approximate scattering matrix and the boundary condition at the vertex of the quantum graph

The standard method of calculation of the scattering matrix requires solving of an infinite algebraic system anyway, though practically admits a certain simplification in closed channels, see [46][2], where it was done for wave guides with simple geometry. The approach based on the intermediate DN-map gives a finite linear system for the Scattering matrix derived from matching of the component Ψ_+ of the scattering Ansatz on the first (open) channel in wires with $p = \sqrt{\lambda - V_\delta - \frac{\pi^2}{\delta^2}}$, on the first spectral band Δ_1:

$$\Psi_+ e_+ := e^{ip\xi} e_+ + e^{-ip\xi} S(p) e_+ \tag{51}$$

to the limit values on the spectrum, $\Im\lambda \to 0$, of the solution of an intermediate boundary problem with the boundary data on Γ defined by the scattering Ansatz Ψ. The boundary data $\Psi\big|_\Gamma$, $\frac{\partial\Psi}{\partial n}\big|_\Gamma$ are connected by the intermediate DN-map, hence the required matching gives a finite linear system for S:

$$ip\,[e_+ - S(p)e_+] = \mathcal{M}(\lambda)\ [e_+ + S(p)e_+]\,. \tag{52}$$

Solving this equation we obtain the formula for the scattering matrix of the operator $\mathcal{L}$ on the first spectral band Δ_1 in terms of $\mathcal{L}_\Lambda$ by the formula, see (22). Due to Theorem 3.2 one can substitute, for a thin junction, the intermediate DN-map $\mathcal{M}(\lambda)$ in (22) by an approximate expression for $\mathcal{M} = \mathcal{M}^\Delta + \mathcal{K}^\Delta$, where $\mathcal{M}^\Delta =: \mathcal{M}_{\rm approx}$ is a rational approximation $\mathcal{DN}^\Delta$ of $\mathcal{M}$ on the essential spectral interval Δ_T, containing only polar terms with poles on an auxiliary spectral interval Δ, and $\mathcal{K}^\Delta$ is a regular part of $\mathcal{M}$ on Δ.

Theorem 4.1. *The resulting approximate expression for the scattering matrix*

$$S \approx [ipP_+ + \mathcal{M}^\Delta]^{-1}[ipP_+ - \mathcal{M}^\Delta] =: S_{\rm approx}, \tag{53}$$

with $p = \sqrt{\lambda - V_\delta - \pi^2\ \delta^{-2}}$, can be used as a first step for the calculation of the exact scattering matrix via an analytic perturbation procedure.

Proof. Indeed, due to the above Theorem 3.2 the error $\mathcal{M} - \mathcal{M}_{\rm approx} = \mathcal{K}^\Delta$, with $\mathcal{K}^\Delta$ containing the regular term $\mathcal{M}_{\rm reg}$ too, is real and estimated by $O(\delta\, d_{\rm int}^{-2})$. Then, due to 3.1 we can represent the exact scattering matrix in form of a product:

$$S = \left(I + [ipP_+ + \mathcal{M}^\Delta]^{-1}\mathcal{K}^\Delta\right)^{-1}\ S_{\rm approx}\ \left(I - [ipP_+ - \mathcal{M}^\Delta]^{-1}\mathcal{K}^\Delta\right). \tag{54}$$

Here $\mathcal{M}^\Delta$, $\mathcal{K}^\Delta$ are hermitian on Δ_1, hence $\|\ [ipP_+ \pm \mathcal{M}^\Delta]^{-1}\ \| \ll \delta$, for thin junctions. Hence the analytic perturbation procedure of the calculation the left and right factors of the expression in (54) is geometrically convergent due to $\delta\ O(\delta\, d_{\rm int}^{-2}) \ll 1$. Thus the exact scattering matrix can be obtained from $S_{\rm approx}$ by an analytic perturbation procedure. □

[2]The author is grateful to V. Katsnelson for important comments in that connection.

One can construct various approximations for the scattering matrix based on S_{approx}, replacing M by various approximate expressions, with controllable errors, see for instance (58, 59).

4.1. Simple resonance eigenvalue of the intermediate Hamiltonian

The simplest approximate formula for the scattering matrix can be obtained in the case when there exist a single simple eigenvalue λ_1^Λ of the intermediate Hamiltonian on the auxiliary spectral interval Δ. Indeed, substituting the intermediate DN-map $\mathcal{M} = \mathcal{K}^\Lambda + \alpha_1^2 \frac{P_1^\Lambda}{\lambda - \lambda_1^\Lambda}$ by the corresponding polar approximation generated by the resonance eigenvalue λ_1^Λ of the intermediate Hamiltonian and the boundary current of the corresponding eigenfunction

$$\alpha_1^2 P_1^\Lambda = P_+ \frac{\partial \varphi_1^\Lambda}{\partial n}\Big|_\Gamma \rangle \, \langle P_+ \frac{\partial \varphi_1^\Lambda}{\partial n}\Big|_\Gamma =: \vec{\psi_1^\Lambda}\rangle \, \langle \vec{\psi_1^\Lambda},$$

we are able to obtain, due to preceding Theorem (4.1), the scattering matrix of a thin junction via an analytic perturbation procedure based on the jump-start

$$S_{\text{jump-start}} = \left[iK_+ + k(\lambda) + \alpha_1^2 \frac{P_1^\Lambda}{\lambda - \lambda_1^\Lambda} \right]^{-1} \left[iK_+ - k(\lambda) - \alpha_1^2 \frac{P_1^\Lambda}{\lambda - \lambda_1^\Lambda} \right]. \qquad (55)$$

In the case when the first spectral band Δ_1 is the conductivity band, the above approximate expression for the scattering matrix can be represented, with $P_1^\perp = P_+ \ominus P_1^\Lambda$ and $p_1 = \sqrt{\lambda - \pi^2 \delta^{-2} - V_\delta}$ and $\lambda_1^\Lambda \approx \Lambda$ as:

$$S_{\text{jump-start}} = P_1^\perp + \frac{ip_1 - k(\lambda) - \alpha_1^2 \frac{1}{\lambda - \lambda_1^\Lambda}}{ip_1 + k(\lambda) + \alpha_1^2 \frac{1}{\lambda - \lambda_1^\Lambda}} P_1^\Lambda. \qquad (56)$$

It corresponds to the one-dimensional solvable model of the junction, obtained via attachment an appropriate inner structure to the vertex, see Fig. 7. This approximate expression for the scattering matrix can be obtained via imposing on the Scattering Ansatx a λ-dependent boundary condition at the vertex, see a discussion in [44]. Unfortunately this boundary condition does not correspond to a selfadjoint operator, so that it can't be interpreted in terms of Quantum Mechanics, the same as prominent Wigner boundary condition, see [76]. We are also able to represent the scattering matrix with use of a single Blaschke factor: $S_{\text{approx}_1}(\lambda) =$

$$P_1^\perp + \left[\frac{ip(\lambda - \lambda_1^\Lambda) - \alpha_1^2}{ip(\lambda - \lambda_1^\Lambda) + \alpha_1^2} \right] P_1^\Lambda \equiv P_1^\perp + \Theta_1^\Lambda(\lambda) \, P_1^\Lambda. \qquad (57)$$

Notice that the scalar Blaschke factor Θ_1^Λ is close to -1 on the essential spectral interval

$$\Delta_T : \left\{ \lambda : |\lambda - \lambda_1^\Lambda| \le 2m^* \, \kappa T \, \hbar^{-2} < \alpha_1^2 \, p^{-1}(\lambda_1^\Lambda) \right\},$$

for low temperature T, and it is close to 1 on the complement. For thin junction and low temperature the boundary condition can be reduced to Datta-type boundary condition,[14], see below, formula (64) represented in terms of boundary currents of the resonance eigenfunction of the intermediate Hamiltonian.

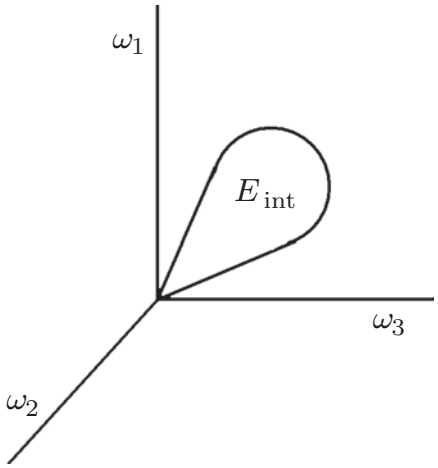

FIGURE 7. 1-d model of T-junction

Once we already developed the compensation procedure based on the representation of the intermediate DN-map in terms of classical DN-map, we can do one more step, expressing the approximate scattering matrix (55) in spectral terms of the unperturbed operator $L_{\rm int}$ on the vertex domain $\Omega_{\rm int}$, under assumption that it has a single resonance eigenvalue $\lambda_1 \in \Delta_T \subset \Delta$,

$$L_{\rm int}\varphi_1 = \lambda_1\varphi_1.$$

Then, for thin junction, the intermediate Hamiltonian also has a simple eigenvalue near to λ_1. We assume that the major part of the correcting term $\mathcal{K}^\Delta$ in the corresponding rational approximation of the DN-map of $L_{\rm int}$ is defined by a finite sum of polar terms and a regular term $\mathcal{M}_{\rm reg} = \mathcal{K}^\Delta_{++} - \mathcal{K}^\Delta_{+-}\frac{I}{\mathcal{K}^\Delta_{--}+K_-}\mathcal{K}^\Delta_{-+}$, see (3.1)

$$\mathcal{K}^\Delta = \sum_{s=2}^{s=M} \frac{I}{\lambda-\lambda_s}\frac{\partial\varphi_s}{\partial n}\rangle\langle\frac{\partial\varphi_s}{\partial n} + \mathcal{M}_{\rm reg},$$

$$\mathcal{DN}(\lambda) = \frac{I}{\lambda-\lambda_1}\frac{\partial\varphi_1}{\partial n}\rangle\langle\frac{\partial\varphi_1}{\partial n} + \sum_{s=2}^{s=M} \frac{I}{\lambda-\lambda_s}\frac{\partial\varphi_s}{\partial n}\rangle\langle\frac{\partial\varphi_s}{\partial n} =: \mathcal{DN}^\Delta + \mathcal{K}^\Delta,$$

$$\mathcal{T} = \varphi_1\rangle\langle\frac{\partial\varphi_1}{\partial n},\ \ \mathcal{J} = P_+ - \mathcal{K}_{+-}K_-^{-1}\left[I + \mathcal{K}_{--}K_-^{-1}\right]^{-1}P_-.$$

Hereafter we neglect the contribution from higher terms of the geometrically convergent series

$$\left[I + \mathcal{K}_{--}K_-^{-1}\right]^{-1}P_- \approx I_- = P_-;$$

$$\mathcal{J} = P_+ - \mathcal{K}_{+-}K_-^{-1} \approx P_+ - \sum_{s=2}^{s=M} \frac{I}{\lambda-\lambda_s}P_+\frac{\partial\varphi_s}{\partial n}\rangle\langle\ K_-^{-1}\ \frac{\partial\varphi_s}{\partial n}.$$

$$Q(\lambda) = \varphi_1\rangle\langle\frac{\partial\varphi_1}{\partial n}K_-^{-1}\left[I + \mathcal{K}_{--}K_-^{-1}P_-\right]^{-1}\frac{\partial\varphi_1}{\partial n}\rangle\langle\varphi_1 \approx \varphi_1\rangle\langle\varphi_1\ \ \langle\frac{\partial\varphi_1}{\partial n}K_-^{-1}\frac{\partial\varphi_1}{\partial n}\rangle.$$

Then, with only terms containing K_-^{-1} taken into account, we obtain for $\mathcal{M}$, based on Theorem 3.1, an approximate expression for

$$\mathcal{M}_{\text{approx}} = \mathcal{K}_{++}^{\Delta} - \mathcal{K}_{+-}^{\Delta}K_-^{-1}\mathcal{K}_{-+}^{\Delta} +$$

$$\left(P_+ - \mathcal{K}_{+-}K_-^{-1}\right)\frac{\partial\varphi_1}{\partial n}\Big|_\Gamma\rangle\frac{I}{\lambda - \lambda_1 + \langle\frac{\partial\varphi_1}{\partial n}\Big|_\Gamma K_-^{-1}\frac{\partial\varphi_1}{\partial n}\Big|_\Gamma\rangle}\langle\left(P_+ - \mathcal{K}_{+-}K_-^{-1}\right)\frac{\partial\varphi_1}{\partial n}\Big|_\Gamma. \tag{58}$$

For "very thin" junction one can neglect even first-order terms containing K_-^{-1}, everywhere, except the expression staying in the denominator, and obtain from (58) a simpler approximate formula:

$$\mathcal{M}_{\text{thin}} = \mathcal{K}_{++}^{\Delta} + P_+\frac{\partial\varphi_1}{\partial n}\Big|_\Gamma\rangle\frac{I}{\lambda - \lambda_1 + \langle\frac{\partial\varphi_1}{\partial n}\Big|_\Gamma K_-^{-1}\frac{\partial\varphi_1}{\partial n}\Big|_\Gamma\rangle}\langle P_+\frac{\partial\varphi_1}{\partial n}\Big|_\Gamma$$

$$=: \mathcal{K}_{++}^{\Delta} + \alpha_1^2\frac{P_1^Q}{\lambda - \lambda_1^Q} \tag{59}$$

where

$$P_1^Q = e_1^Q\rangle\langle e_1^Q,\ e_1^Q = \alpha_1^{-1}P_+\frac{\partial\varphi_1}{\partial n}\Big|_\Gamma,$$

$$\alpha_1 = \|\ P_+\frac{\partial\varphi_1}{\partial n}\Big|_\Gamma\ \|,\ \lambda_1^Q = \lambda_1 - \langle\frac{\partial\varphi_1}{\partial n}\Big|_\Gamma K_-^{-1}\frac{\partial\varphi_1}{\partial n}\Big|_\Gamma\rangle.$$

This implies an approximate formula for the scattering matrix on the major part of the essential spectral interval Δ_T[3], for low temperature, once the junction is *thin on the open channel*

$$p^2(\Lambda) = |\Lambda - (\pi^2\,\delta^{-2} + V_\delta)| \gg \|\ K^{\Delta}| \ \| \ . \tag{60}$$

Notice that arising of the shape of the resonance eigenfunction of the unperturbed operator L_{int} in the corresponding approximate jump-start formula (61) for the scattering matrix corresponds to folklore observation of physicists, that the eigenfunctions react to perturbation slower that the eigenvalues. To derive the jump-start approximation for the scattering matrix, denote $\sqrt{\lambda - \pi^2\,\delta^{-2} - V_\delta} =: p$. Then we obtain on Δ_T, similarly to Theorem 4.1:

$$S(\lambda) \approx P_1^{\perp} + \frac{ip - k - \frac{\alpha_1^2}{\lambda - \lambda_1^Q}}{ip + k + \frac{\alpha_1^2}{\lambda - \lambda_1^Q}}P_1 =: S_{\text{jump-start}}^Q. \tag{61}$$

[3]Roughly speaking, on a complement of a certain small neighborhood of the zero λ_1^Q of the denominator.

In the case when $\lambda_1^Q \approx \Lambda$ we can replace on Δ_T the Blaschke factor in front of P_1 by -1, which implies on Δ_T, for sufficiently low temperature:

$$S^Q_{\text{jump-start}}(\lambda) \approx P_1^\perp - P_1 \equiv S_{\text{Datta}} \tag{62}$$

When modeling the quantum network by a one-dimensional graph, one can attempt to define a boundary condition at the vertex which implies the scattering matrix (62). Indeed, forming the component Ψ_+ of the scattering Ansatz based on (62), we see that the boundary values of the Ansatz

$$\Psi_+(x)\nu = e^{iK_+x}\nu + e^{-iK_+x}\ S\nu$$

satisfy the following boundary condition, similar[4] to one suggested in [14]:

$$\Psi_+(0)\ \nu = 2P_1^\perp\nu,\ \frac{d}{dx}\,\Psi_+(0)\nu = 2P_1\nu.$$

In terms of the components $\Psi_+^n(0)\nu$, $\frac{d}{dx}\,\Psi_+^n(0)\nu$ of the boundary values of the Ansatz on the bottom sections Γ_n of the wires and with use of the components of the boundary currents $\vec{\psi}_1 = \{\psi_1^n\}_{n=1}^N$

$$P_+\frac{\partial\varphi_1}{\partial n}\bigg|_{\Gamma_n} \equiv \psi_1^n \tag{63}$$

we obtain: $\langle\vec{\psi}, \Psi_+(0)\nu\rangle = 0$, $\frac{d}{dx}\Psi_+(0)\nu \parallel \vec{\psi}$, or

$$\langle\vec{\psi}, \Psi_+(0)\nu\rangle = \sum_n^N \Psi_+^n\ \nu\ \bar{\psi^n} = 0,$$

$$\frac{\frac{d}{dx}\,\Psi_+^1(0)\nu}{\psi_1^1} = \frac{\frac{d}{dx}\,\Psi_+^2(0)\nu}{\psi_1^2} = \cdots = \frac{\frac{d}{dx}\,\Psi_+^n(0)\nu}{\psi_1^n} = \cdots = \frac{\frac{d}{dx}\,\Psi_+^N(0)\nu}{\psi_1^N} \tag{64}$$

Example 4: asymmetric T-junction. Consider a two-dimensional quantum network Ω constructed as a simplest asymmetric T-junction of three straight semi-infinite quantum wires width $\pi/2$ attached as shown in Fig. 8 to the quantum well – the square $\Omega_{\text{int}} : 0 < x < \pi, 0 < y < \pi$ on x, y-plane. The role of the one-electron Hamiltonian on Ω is played the Dirichlet Laplacian. The spectrum of the corresponding unperturbed Hamiltonian on the quantum well is discrete and the eigenpairs, e.g.,

$$\lambda_1 = 2 : \varphi_{1,1} = \frac{2}{\pi}\sin x \times \sin y;$$

$$\lambda_2 = \lambda_3 = 5; \varphi_{1,2} = \frac{2}{\pi}\sin x\ \times\ \sin 2y \text{ and } \varphi_{2,1} = \frac{2}{\pi}\sin 2x\ \times\ \sin y;$$

$$P_5 = P_{1,2} + P_{2,1},\ P_{1,2} = \varphi_{1,2}\rangle\ \langle\varphi_{1.2} : P_{2,1} = \varphi_{2,1}\rangle\ \langle\varphi_{2,1};$$

$$\lambda_4 = 8; \varphi_{2,2} = \frac{2}{\pi}\sin 2x \times \sin 2y; P_8 = P_{2,2} = \varphi_{2,2}\rangle\,\langle\varphi_{2,2};$$

[4] but not yet equivalent, see an extended discussion below, in next subsection.

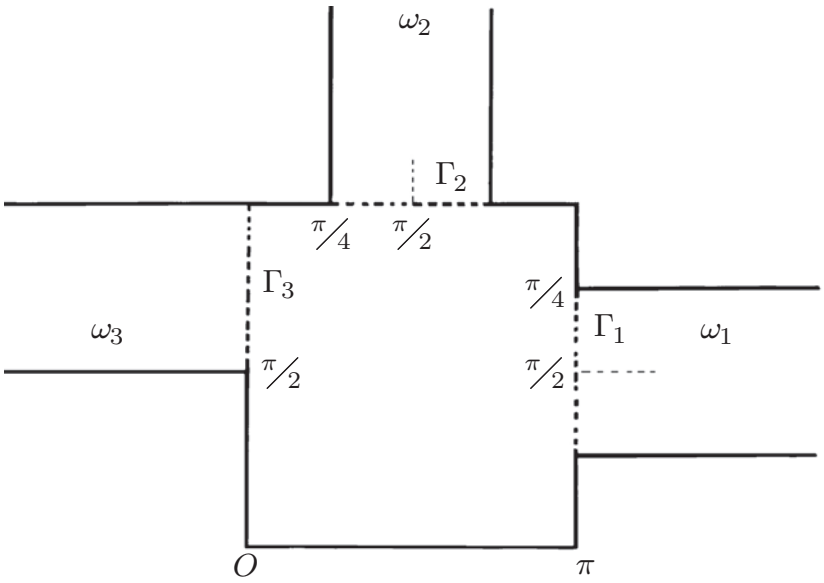

FIGURE 8. Simplest asymmetric T-junction

$$\lambda_5 = \lambda_6 = 10; \varphi_{1,3} = \frac{2}{\pi}\sin x \times \sin 3y \text{ and } \varphi_{3,1} = \frac{2}{\pi}\sin 3x \times \sin y; \dots$$
$$\lambda_6 = \lambda_7 = 13; \varphi_{2,3} = \frac{2}{\pi}\sin 2x \times \sin 3y \text{ and } \varphi_{3,2} = \frac{2}{\pi}\sin 3x \times \sin 2y; \dots$$
$$\dots\dots\dots \tag{65}$$

are obtained via separation of variables. We choose the basic spectral interval $\Delta = [4, 6]$, so that there is only one multiple eigenvalue $\lambda_2 = \lambda_3 = 5$ of L_{int} on that interval, and use the approximate formula (59) for $\mathcal{M}$. The normal boundary current $J_{1,2}$ of the resonance eigenfunctions $\varphi_{1,2}$ and $\varphi_{2,1}$ is calculated as

$$J_{1,2} = \begin{pmatrix} \frac{\partial\varphi_{1,2}}{\partial n}\Big|_{\Gamma_1} \\ \frac{\partial\varphi_{1,2}}{\partial n}\Big|_{\Gamma_2} \\ \frac{\partial\varphi_{1,2}}{\partial n}\Big|_{\Gamma_3} \end{pmatrix} = \begin{pmatrix} -\frac{2}{\pi}\sin 2y \\ \frac{4}{\pi}\sin x \\ -\frac{2}{\pi}\sin 2y \end{pmatrix},$$

$$J_{2,1} = \begin{pmatrix} \frac{\partial\varphi_{2,1}}{\partial n}\Big|_{\Gamma_1} \\ \frac{\partial\varphi_{2,1}}{\partial n}\Big|_{\Gamma_2} \\ \frac{\partial\varphi_{2,1}}{\partial n}\Big|_{\Gamma_3} \end{pmatrix} = \begin{pmatrix} \frac{4}{\pi}\sin y \\ -\frac{2}{\pi}\sin 2x \\ -\frac{4}{\pi}\sin y \end{pmatrix}. \tag{66}$$

The spectrum of the unperturbed Hamiltonian on the wires is absolutely continuous and has a band structure, with thresholds $\{l^2\}$, $l = 1, 2, 3, \ldots$ separating the spectral bands. The multiplicity of the continuous spectrum jumps up by three units on each threshold. We assume that the first spectral band $\Delta_1 = [4, 16]$ is the conductivity band. The entrance subspace of the open channel is spanned by the inferior cross-section eigenfunctions $e^{\perp}_{s,1} = (4/\pi)^{1/2} \sin 2x^{\perp}_s$, $s = 1, 2, 3$ on the bottom cross-section $\Gamma_s = [0 < x^{\perp}_s < \pi/2], s = 1, 2, 3$. The entrance subspace of closed channels is the linear hull of the superior cross-section eigenfunctions $e^{\perp}_{s,l} = (4/\pi)^{1/2} \sin 2l\, x^{\perp}_s$, $s = 1, 2, 3,$ with $l = 2, 3, \ldots$. We will calculate the approximate scattering matrix (jump-start) of the thin junction based on the approximate formulae for (58, 59). The approximate eigenvalues of the intermediate Hamiltonian are found as zeros of the denominator $\mathbf{d}(\lambda)$ represented as a 2×2-matrix with respect to the basis $\varphi_{1,2}, \varphi_{2,1}$;

$$\mathbf{d}(\lambda) = \begin{pmatrix} \lambda - 5 & 0 \\ 0 & \lambda - 5 \end{pmatrix} + \begin{pmatrix} \langle P_- \frac{\partial \varphi_{1,2}}{\partial n}\big|_\Gamma K_-^{-1} P_- \frac{\partial \varphi_{1,2}}{\partial n}\big|_\Gamma \rangle & \langle P_- \frac{\partial \varphi_{1,2}}{\partial n}\big|_\Gamma K_-^{-1} P_- \frac{\partial \varphi_{2,1}}{\partial n}\big|_\Gamma \rangle \\ \langle P_- \frac{\partial \varphi_{2,1}}{\partial n}\big|_\Gamma K_-^{-1} P_- \frac{\partial \varphi_{1,2}}{\partial n}\big|_\Gamma \rangle & \langle P_- \frac{\partial \varphi_{2,1}}{\partial n}\big|_\Gamma K_-^{-1} P_- \frac{\partial \varphi_{2,1}}{\partial n}\big|_\Gamma \rangle \end{pmatrix}. \tag{67}$$

Taking into account only components of the currents in the second spectral channel

$$K_- \frac{\partial \varphi_{1,2}}{\partial n}\Big|_\Gamma \approx e_{s,2} \langle \frac{\partial \varphi_{1,2}}{\partial n}\Big|_\Gamma, e_{s,2} \rangle$$

and introducing the following notations for the integrals

$$2\int_0^{\pi/4} \sin 2x \sin 4x dx = 2/3 =: \alpha,$$
$$\int_0^{\pi/2} \sin x \sin 4x dx = -4/15 =: \gamma,\ \alpha + \gamma = 2/5 =: \beta,$$

we represent the denominator (67) and the inverse $[\mathbf{d}(\lambda)]^{-1}$ as

$$\mathbf{d}(\lambda) = \begin{pmatrix} \lambda - 5 & 0 \\ 0 & \lambda - 5 \end{pmatrix} + \frac{\pi}{4\sqrt{16 - \lambda}} \begin{pmatrix} \alpha^2 & -\alpha\beta \\ -\alpha\beta & \beta^2 \end{pmatrix}$$
$$= (\lambda - 5) \frac{1}{\alpha^2 + \beta^2} \begin{pmatrix} \beta \\ \alpha \end{pmatrix} \rangle \langle \begin{pmatrix} \beta \\ \alpha \end{pmatrix}$$
$$+ \left(\lambda - 5 + [\alpha^2 + \beta^2] \frac{\pi}{4\sqrt{16 - \lambda}}\right) \frac{1}{\alpha^2 + \beta^2} \begin{pmatrix} -\alpha \\ \beta \end{pmatrix} \rangle \langle \begin{pmatrix} -\alpha \\ \beta \end{pmatrix},$$
$$[\mathbf{d}(\lambda)]^{-1} =: \frac{P_5}{\lambda - 5} + \frac{P_{5-\delta^Q}}{\lambda - 5 + \delta^Q}, \tag{68}$$

with $\delta^Q = \pi\ [\alpha^2 + \beta^2]/4\ \sqrt{16 - \lambda}$. Here K_- is substituted by the contribution $4/\pi\ \ \sin 4x^{\perp}\rangle\ \langle \sin 4x^{\perp}$ from the second spectral branch in the wires.

If the scaled Fermi-level is 5, then the corresponding multiple resonance eigenvalue of L_{int} is split into pair of eigenvalues $\lambda^{\Lambda}_1 = 5$, $\lambda^{\Lambda}_2 \approx 5 - \frac{\pi\ [\alpha^2 + \beta^2]}{4\sqrt{16 - 5}}$,

and the jump-start approximation of the scattering matrix can be calculated in terms of P_+-projections of the boundary currents of the resonance eigenfunctions $\varphi_{1,2}, \varphi_{2,1} :=$

$$P_+ \frac{\partial \varphi_{1,2}}{\partial n}\Big|_\Gamma = P_+ J_{1,2} = \begin{pmatrix} P_+ \frac{\partial \varphi_{1,2}}{\partial n}\Big|_{\Gamma_1} \\ P_+ \frac{\partial \varphi_{1,2}}{\partial n}\Big|_{\Gamma_2} \\ P_+ \frac{\partial \varphi_{1,2}}{\partial n}\Big|_{\Gamma_3} \end{pmatrix} = \begin{pmatrix} -\frac{4}{\pi^2} \ \sin 2x_1^\perp \int_{\Gamma_1} \sin 2x_1^\perp \sin 2y d\Gamma_1 \\ \frac{8}{\pi^2} \sin 2x_2^\perp \int_{\Gamma_2} \sin 2x_2^\perp \sin x d\Gamma_3 \\ -\frac{4}{\pi^2} \ \sin 2x_3^\perp \int_{\Gamma_3} \sin 2x_3^\perp \sin 2y d\Gamma_3 \end{pmatrix}$$

$$= \sin 2x_2^\perp \begin{pmatrix} 0 \\ 16\sqrt{2}/3\pi^2 \\ -4/\pi \end{pmatrix} =: \frac{2}{\sqrt{\pi}} \sin 2x_2^\perp \ \vec{\psi}_{1,2}.$$

$$P_+ \frac{\partial \varphi_{2,1}}{\partial n}\Big|_\Gamma = P_+ J_{2,1} = \begin{pmatrix} P_+ \frac{\partial \varphi_{2,1}}{\partial n}\Big|_{\Gamma_1} \\ P_+ \frac{\partial \varphi_{2,1}}{\partial n}\Big|_{\Gamma_2} \\ P_+ \frac{\partial \varphi_{2,1}}{\partial n}\Big|_{\Gamma_3} \end{pmatrix} = \begin{pmatrix} \frac{8}{\pi^2} \ \sin 2x_1^\perp \int_{\Gamma_1} \sin 2x_1^\perp \sin y \, d\Gamma_1 \\ -\frac{4}{\pi^2} \ \sin 2x_2^\perp \int_{\Gamma_2} \sin 2x_2^\perp \sin 2x \, d\Gamma_2 \\ -\frac{8}{\pi^2} \ \sin 2x_3^\perp \int_{\Gamma_3} \sin 2x_3^\perp \sin y \, d\Gamma_2 \end{pmatrix}$$

$$= \sin 2x_2^\perp \begin{pmatrix} 16/3\pi^2 \\ 0 \\ -16/3\pi^2 \end{pmatrix} = \frac{2}{\sqrt{\pi}} \sin 2x_2^\perp \ \vec{\psi}_{2,1}. \tag{69}$$

The exponent K_+ of the first (open) channel is represented as $K_+(\lambda) = \sqrt{\lambda - 4} P_+$, where the projection P_+ onto the entrance subspace of open channels plays the role of unity $I_+ = I_1 + I_2 + I_3$ in E_+ and is represented as

$$P_+ = 4/\pi \left[\sin 2x_1^\perp \rangle \langle \sin 2x_1^\perp + \sin 2x_2^\perp \rangle \langle \sin 2x_2^\perp + \sin 2x_3^\perp \rangle \langle \sin 2x_3^\perp \right].$$

Then taking into account that the contribution $\mathcal{K}_{++}$ from the major polar part of $\mathcal{K} \approx \frac{P_8}{\lambda - 8}$ is

$$\mathcal{K}_{++} \approx \frac{16}{\pi^3} \ \frac{\sin 2x_3^\perp \rangle \, \langle \sin 2x_3^\perp}{\lambda - 8} = \frac{2}{\sqrt{\pi}} \sin 2x_3^\perp \rangle \frac{4}{\pi^2 (\lambda - 8)} \langle \frac{2}{\sqrt{\pi}} \sin 2x_3^\perp,$$

and omitting the vector factors $2/\sqrt{\pi} \sin 2x_1^\perp \rangle$ and $\langle 2/\sqrt{\pi} \sin 2x_1^\perp$ on the right and left side of the jump-start scattering matrix,

$$S_{\text{jump-start}} = \frac{ipI_+ - \frac{4I_3}{\lambda - 8} - \langle \begin{pmatrix} \vec{\psi}_{1,2} \\ \vec{\psi}_{2,1} \end{pmatrix} \left[\frac{P_5}{\lambda - 5} + \frac{P_{5-\delta^Q}}{\lambda - 5 + \delta^Q} \right] \begin{pmatrix} \vec{\psi}_{1,2} \\ \vec{\psi}_{2,1} \end{pmatrix} \rangle}{ipI_+ + \frac{4I_3}{\lambda - 8} + \langle \begin{pmatrix} \vec{\psi}_{1,2} \\ \vec{\psi}_{2,1} \end{pmatrix} \left[\frac{P_5}{\lambda - 5} + \frac{P_{5-\delta^Q}}{\lambda - 5 + \delta^Q} \right] \begin{pmatrix} \vec{\psi}_{1,2} \\ \vec{\psi}_{2,1} \end{pmatrix} \rangle} \tag{70}$$

For better approximation of the scattering matrix we should use better approximation for $K_-, \mathcal{K}^\Delta$.

4.2. Symmetric junction

The Datta-type boundary condition (64) does not coincide with the original Datta-Das Sarma boundary condition (1,2), suggested in [14], because the phenomenological Datta-Das Sarma boundary condition was suggested for a T-junction which is symmetric with respect to the left-right reflection. The resonance concept of the conductance permits to derive, for a symmetric junction, the original Datta-Das Sarma condition and interpret the phenomenological parameter β.

Consider a symmetric junction Ω consisting of a square $(0,\pi)\times(0,\pi)$, and the quantum wires width $\pi/2$, attached in the middle of the sides $\Gamma_1,\Gamma_2,\Gamma_3$, see (4). The role of the one-electron Hamiltonian is played by the Laplacian on Ω with zero boundary conditions. Similarly to above example we assume that the electrons are supplied from the second wire, in the first spectral channel. We assume that the scaled Fermi level is Λ =10 and the conductivity band is $4\leq\lambda\leq 16$, the eigenvalues of $L_{\rm int}$ embedded into the conductivity band are $\lambda_0=5$, $\lambda_1=8$, $\lambda_2=10$, $\lambda_3=13$. The corresponding eigenfunctions of $L_{\rm int}$ are found in previous subsection via separation of variables, see (65). The role of the resonance eigenfunctions is played by $\varphi_{1,3},\varphi_{3,1}$, with the eigenvalue $\lambda_5=\lambda_6=10$. We also use the symmetric and antisymmetric linear combinations of them

$$2^{-1/2}[\varphi_{1,3}+\varphi_{3,1}]=:\varphi_s,\ 2^{-1/2}[\varphi_{1,3}-\varphi_{3,1}]=:\varphi_a.$$

We denote the corresponding boundary currents as

$$\frac{\partial\varphi_s}{\partial n}\Big|_\Gamma=:J_s,\ \frac{\partial\varphi_a}{\partial n}\Big|_\Gamma=:J_a$$

and consider the projections of them $P_+J_{\rm sym}$, $P_+J_{\rm asym}$ onto the entrance space of the first (open) channel. We assume that the temperature is low, so that the role of an essential spectral interval is played by $\Delta_T=[9,11]$. Then the eigenfunctions and eigenvalues of the intermediate Hamiltonian can be found based on Theorem 3.1, taking into account the approximate calculation of the potential $Q(\lambda)$ of $L^\Delta(\lambda)$:

$$Q(\lambda)\approx\begin{pmatrix}\varphi_{1,3}\\ \varphi_{3,1}\end{pmatrix}\rangle\mathbf{D}\langle\begin{pmatrix}\varphi_{1,3}\\ \varphi_{3,1}\end{pmatrix},$$

with

$$\mathbf{D}=\begin{pmatrix}\langle P_-\frac{\partial\varphi_{1,3}}{\partial n}\big|_\Gamma K_-^{-1}P_-\frac{\partial\varphi_{1,3}}{\partial n}\big|_\Gamma\rangle & \langle P_-\frac{\partial\varphi_{1,3}}{\partial n}\big|_\Gamma K_-^{-1}P_-\frac{\partial\varphi_{3,1}}{\partial n}\big|_\Gamma\rangle\\ \langle P_-\frac{\partial\varphi_{3,1}}{\partial n}\big|_\Gamma K_-^{-1}P_-\frac{\partial\varphi_{1,2}}{\partial n}\big|_\Gamma\rangle & \langle P_-\frac{\partial\varphi_{3,1}}{\partial n}\big|_\Gamma K_-^{-1}P_-\frac{\partial\varphi_{3,1}}{\partial n}\big|_\Gamma\rangle\end{pmatrix},$$

and thus neglecting Q in the case of thin networks. Hence, in the first-order approximation, the perturbed eigenvalues of the intermediate Hamiltonian remain the same: $\lambda_5^Q=\lambda_6^Q=10$. Due to reflection symmetry of the junction there are two eigenfunctions of the intermediate Hamiltonian which correspond to the eigenvalue multiplicity 2 obtained based on the Theorem 3.1. The corresponding eigenfunctions and the projections of the normal currents onto E_+ are respectively symmetric and anti-symmetric:

$$\vec{\psi}_{a,s}=P_+\frac{\partial\varphi_{a,s}}{\partial n}|P_+\frac{\partial\varphi_{a,s}}{\partial n}|^{-1}$$

$$\vec{\psi}_a = \alpha_a \frac{1}{\sqrt{2}} \begin{pmatrix} 1 \\ 0 \\ -1 \end{pmatrix} = \alpha_a \; e_a, \; \vec{\psi}_s = \alpha_s \frac{1}{\sqrt{2+\gamma^2}} \begin{pmatrix} 1 \\ \gamma \\ 1 \end{pmatrix} = \alpha_s \; e_s,$$

and the orthogonal complement in E_+ is spanned by the vector

$$\frac{\gamma}{\sqrt{4+2\gamma^2}} \begin{pmatrix} 1 \\ -2/\gamma \\ 1 \end{pmatrix} =: \frac{1}{\sqrt{2+\beta^2}} \begin{pmatrix} 1 \\ \beta \\ 1 \end{pmatrix},$$

with $\beta = -2/\gamma$. In the first-order approximation $\mathcal{K}$ can be substituted by the contribution from the nearest eigenvalue $\lambda_4 = 8$, and hence $\mathcal{K}_{++} = k = 0$.

The intermediate DN-map is represented, due to Theorem 3.1 by the formula

$$\mathcal{M} = \frac{\alpha_a^2 P_a + \alpha_s^2 P_s}{\lambda - \lambda_2^Q} + \mathcal{K}_{++} \approx \frac{\alpha_a^2 P_a + \alpha_s^2 P_s}{\lambda - 10},$$

where $P_a = e_a\rangle \langle e_a, \; P_s = e_s\rangle \langle e_s$. Then denoting by $P^\perp := P_+ - [P_a + P_s] =: P_+ - P_Q$, we represent the scattering matrix of the symmetric junction as

$$S = P^\perp + \frac{ipP_Q - \frac{\alpha_a^2 P_a + \alpha_s^2 P_s}{\lambda - \lambda_2^Q}}{ipP_Q + \frac{\alpha_a^2 P_a + \alpha_s^2 P_s}{\lambda - \lambda_2^Q}}.$$

Here the role of the scalar Blaschke factor Θ_1 in (57) is played by the 2×2 matrix

$$\Theta = \frac{ip[\lambda - \lambda_2^Q]P_Q - [\alpha_a^2 P_a + \alpha_s^2 P_s]}{ip[\lambda - \lambda_2^Q]P_Q + [\alpha_a^2 P_a + \alpha_s^2 P_s]}.$$

The matrix Θ for low temperature is close to $-P_Q$ on the corresponding small essential spectral interval Δ_T centered at λ_2^Q. Then the scattering matrix is represented as

$$S = P_Q^\perp - P_Q,$$

with the pair of complementary projections $P_Q^\perp$ and P_Q, $\dim P_Q^\perp = 1$, $\dim P_Q = 2$. Scattering Ansatz on the model graph

$$\Psi = e^{ipP_+ x} e + e^{-ipP_+ x} S e$$

satisfies the following boundary condition at the vertex $x = 0$:

$$\Psi(0) = (I + S)e = 2P_Q^\perp e, \; \Psi'(0) = ipP_+(I - S)e = 2ipP_Q e.$$

Taking into account that $\dim P_Q^\perp = 1, \; \dim P_Q = 2$, we can re-write the previous formulae as boundary conditions imposed on the Ansatz:

$$P_Q^\perp \vec{\psi}'(0) = 0, \;\; \vec{\psi}(0) \text{ is parallel to } P_Q^\perp e \text{ or}$$

$$\langle e^\perp, \psi'(0)\rangle = 0, \; \frac{\psi_1(0)}{e_1^\perp} = \frac{\psi_2(0)}{e_2^\perp} = \frac{\psi_3(0)}{e_3^\perp}. \tag{71}$$

where $P_Q^\perp =: e^\perp\rangle \langle e^\perp, \; e^\perp = (e_1^\perp, e_2^\perp, e_3^\perp) = (2+\beta^2)^{-1/2}(1, \beta, 1)$. This condition coincides with the original Datta-Das Sarma boundary condition, see (1,2). Our analysis reveals the meaning of the phenomenological parameter β.

5. A solvable model of a thin junction

Generally, a Schrödinger operator with non-constant coefficients or one in a non-standard domain rarely admits spectral analysis in explicit form. For qualitative analysis of quantum systems the Schrödinger operator often is substituted by a solvable model, constructed by the von Neumann operator extension technique, [48], see for instance [8, 15, 6] and an extended list of references in [5]. In particular, the substitution of the network by a proper one-dimensional graph, with special boundary conditions at the vertices, looks like a convenient tool for the qualitative analysis of the Schrödinger equation on the network. Unfortunately the estimation of the error caused by the substitution of the network by the corresponding graph is difficult. Shrinking of a "fattened graph" Ω_δ to one-dimensional graph was studied in numerous papers, see for instance [37, 38].

The authors considered a compact network Ω_δ constructed of the vertex domains $\Omega_{\rm in}$, with the diameter proportional to δ^α, $0 < \alpha < 1$ and several finite leads ω^m, width δ, joining them to each other. In [67] they developed, based on [75, 68], a variational technique for description of the asymptotic behavior of the discrete spectrum of the Schrödinger operator on the quantum network Ω_δ of various grades α of thinness. It appeared that the (discrete) spectrum of the Laplacian on the compact shrinking "fattened graph" Ω_δ tends to the spectrum of the Laplacian on the corresponding one-dimensional graph but with different boundary conditions at vertices depending on the speed of shrinking: the Kirchhoff boundary conditions at the nodes, in the case of "small protrusion" $1/2 < \alpha < 1$, or the homogeneous Dirichlet boundary conditions, in case of " large protrusion", $0 < \alpha < 1/2$, see [38], Theorems 1, 2, 3.

In this section we consider a thin quantum networks with small protrusion $\alpha = 1$, assuming diam $\delta \ll$ diam $\Omega_{\rm in}$. We will construct a *quantitatively consistent* solvable model of the quantum network, in the form of a star-graph with a vertex supplied with inner space and appropriate vertex Hamiltonian. The scattering matrix of the properly fitted model serves as a local approximation – on a certain "essential" spectral interval Δ – of the scattering matrix of the original network. In contrast to the quoted above results for compact networks, where the wave-functions are obtained based on the variational approach, we use Dirichlet-to-Neumann map $\mathcal{DN}^\Lambda$ of an intermediate Hamiltonian L_Λ, to derive an explicit formula for the scattered waves on the original network, see (22). The scattering matrix of the star-graph model is obtained via replacement of $\mathcal{M}$ in (22) by the corresponding rational approximation on Δ, based on Theorem 3.1. This defines all parameters of the model in terms of spectral characteristics of L_Λ. In course of construction and fitting of the solvable star-graph model we also define the energy-dependent boundary conditions at the vertex for the Schrödinger equation on the graph. In the simplest case when only one resonance eigenvalue λ_0 of the intermediate Hamiltonian is present on Δ, this condition depends linearly on the spectral parameter, and is parametrized by coordinates $\Re k_0$, $\Im k_0$ of the corresponding resonance.

Note that in [34] an algorithm for construction of the scattering matrix of the quantum graph of the scattering matrices of star-shaped elements is described and, in [26] a convenient formula for the scattering matrix of the star-graph in terms of the boundary parameters at the vertex is suggested. For extended discussion of properties of star graphs see [30, 31].

The star-graph solvable model of the thin junction will be constructed as a finite-dimensional perturbation of an orthogonal sum of the non-perturbed Hamiltonian l_Λ in the open channels and a finite matrix A acting in the *inner space* of the vertex. We will choose the parameters of the model such that the model scattering matrix coincides with the few-pole approximation $\mathbf{S}_\Delta$ of the complete scattering matrix $\mathbf{S}$. Then the constructed model will be automatically *fitted* (i.e., quantitatively consistent). We assume that the spectral variable is scaled such that original the Schrödinger equation on the wires is just $-\Delta u + V_\delta u = \lambda u$. The scattering Ansatz of the model in open channels satisfies on the wires ω the same equation as the scattering Ansatz on the original network

$$l_\Lambda U := -\frac{d^2U^\omega}{dx^2} + \sum_{s,m} \frac{\pi^2\, s^2}{\delta^2}\, p_s^m U + V_\delta U := \lambda\ ,\ \lambda \in \Delta,\ U = U^\omega = \left(u^1, u^2, \dots u^M\right). \tag{72}$$

We assume that the Schrödinger equation on the quantum well $\Omega_{\text{int}} = \Omega_0$, with the same spectral parameter, is represented as

$$L_0 u := -\Delta u + Vu = \lambda u,$$

with a corresponding effective mass μ_0. We assume that the intermediate relative DN-map, $P_+ \frac{\partial u}{\partial n}\Big|_\Gamma =: \mathcal{DN}u$, with respect to Γ, is calculated and the corresponding rational approximation is selected. Then the construction of the vertex part of the model will be done with a major change of the original intermediate Hamiltonian.

For thin networks an auxiliary spectral interval Δ is selected inside $\Delta_1 = [\frac{\pi^2}{\delta^2}, 4\frac{\pi^2}{\delta^2}]$, and hence does not overlap with the continuous spectrum of the intermediate Hamiltonian L_Λ. Only a finite number N of eigenvalues of the intermediate operator are situated on Δ. Then substitution of $\mathcal{M}$ on Δ by the rational approximation $\mathcal{M}_\Delta$ may cause only a minor and controllable error. Now we will prove that there exist a finite-dimensional perturbation of the operator $l_\Lambda \oplus A$ such that the scattering matrix of the perturbed operator coincides with S_Δ. The perturbation will be constructed via operator restriction-extension procedure applied to the orthogonal sum $l_\Lambda \oplus A$, where A is an $N \times N$ Hermitian matrix : $E_A \to E_A$, $\dim E_A = N$. The parameters of the model will be properly selected to fit the spectral data of the intermediate Hamiltonian on the original quantum network, within the auxiliary spectral interval Δ.

Assume that the positive matrix A is defined by its spectral decomposition

$$A = \sum_r \alpha_r^2 P_r.$$

Here $\alpha_r^2 > 0$ are eigenvalues of A, and $P_r = \nu_r \rangle \langle \nu_r$ are the corresponding orthogonal spectral projections. The eigenvalues and the boundary parameters β of the model, see below (82), will be defined later, based on comparison of the scattering matrix of the model with the essential scattering matrix

$$S_\Delta = [iK_+ + \mathcal{D}\mathcal{N}_\Delta]^{-1}[iK_+ - \mathcal{D}\mathcal{N}_\Delta]. \tag{73}$$

Consider restrictions of both l_Λ and A to symmetric operators on the corresponding domains. The restriction of $l_\Lambda\big|_{D_0^l} = l_0$ is defined on functions vanishing near $x = 0$. Then the adjoint operator l_0^+ is defined on $W_2^2(E_+, R_+)$, and the boundary form of it is calculated via integration by parts:

$$\mathcal{J}_l(U, V) = \langle l_0^+ U, V\rangle - \langle U, l_0^+ V\rangle = \langle U'(0), V(0)\rangle - \langle U(0), V'(0)\rangle, \tag{74}$$

where $U(0), V(0) \in E_+$ and the derivatives are taken in the outgoing direction on Γ with respect to Ω_{in}.

Restriction of the matrix A is equivalent to selection of the deficiency subspace for the given value i of the spectral parameter. Choose a generating subspace N_i, $\overline{\bigvee_{k>0} A^k N_{-i}} = E_A$ such that $\frac{A+iI}{A-iI} N_i \cap N_i = 0$, $\dim N_i = d$, set $D_0^A = (A - iI)^{-1}(E_A \ominus N_i)$ and define the restriction of the inner Hamiltonian as $A \to A_0 = A\big|_{D_0^A}$. We develop the extension procedure for general N_i and fit it later based on spectral data of the intermediate operator, see Theorem 4.1, 4.2. In our construction $N_i \subset E_A$ will play a role of the deficiency subspace at the spectral point i, $\dim N_i = d$, $2d \le N$ and the dual deficiency subspace is $N_{-i} = \frac{A+iI}{A-iI} N_i$. The domain of the restricted operator A_0 is not dense in E_A, because A is bounded. Nevertheless, since the deficiency subspaces $N_{\pm i}$ do not overlap, the extension procedure for the orthogonal sum $l_0 \oplus A_0$ can be developed. We will do it here with use of the symplectic formalism, see for instance [54]. In this case the "formal adjoint" operator for A_0 is defined on the defect $N_i + N_{-i} := \mathcal{N}$ by the von Neumann formula : $A_0^+ e \pm i\, e = 0$ for $e \in N_{\pm i}$. Then the extension is constructed, see Lemmas 3.1–3.4 below, via restriction of the formal adjoint onto a certain plane in the defect where the boundary form vanishes (a "Lagrangian plane"). According to the classical von Neumann construction all Lagrangian planes are parametrized by isometries $V : N_i \to N_i$ in the form

$$\mathcal{T}_V = (I - V)\, N_i.$$

In case when the deficiency subspaces do not overlap, the corresponding isometry is admissible, and, according to [35] there exist a selfadjoint extension A_V of the restricted operator A_0. We construct this extension based on the following

Lemma 5.1. *The Lagrangian plane $\mathcal{T}_V$ in the defect forms a non-zero angle with the domain D_0^A of the restricted operator A_0.*

Proof. Indeed, if A_V is the extension, then on the $\mathcal{T}_V$ it coincides with the restriction of the formal adjoint, and on the domain D_0^A it coincides with A_0. Then

assuming that $\mathcal{T}_V$ and D_0^A overlap, we obtain, for some $f^\perp \perp N_i,\ \nu \in N_i$

$$\frac{1}{A-iI} f^\perp = \nu - V\nu.$$

Applying $A_V - iI$ to both parts of this equation, we obtain

$$f^\perp = -2i\nu,$$

hence $f^\perp = -2i\nu = 0$. □

It follows from the lemma that, once the extension is constructed on the Lagrangian plane, the whole construction of the extended operator can be accomplished in the form of a direct sum of the closure of the restricted operator and the extended operator on the Lagrangian plane.

Note that the operator extension procedure may be developed without assumption of non-overlapping, see [35]. In particular, the case $\dim E_A = 1$, which is not formally covered by the above procedure, was analyzed in [71] independently of [35]. The relevant formulae for the scattering matrix and scattered waves remain true and may be verified by the direct calculation. We will use this fact in Section 7 below.

We will use hereafter notations and some facts concerning the symplectic operator extension procedure, see Appendix and references therein. Choose an orthonormal basis in $N_i : \{f_s\}$, $s = 1, 2, \ldots, d$, as a set of deficiency vectors of the restricted operator A_0. Then the vectors $\hat{f}_s = \frac{A+iI}{A-iI} f_s$ form an orthonormal basis in the dual deficiency subspace N_{-i}. Under the above non-overlapping condition one can use the formal adjoint operator A_0^+ defined on the defect $N_i + N_{-i} = \mathcal{N}$:

$$u = \sum_{s=1}^{d} [x_s\, f_s + \hat{x}_s\, \hat{f}_s] \in \mathcal{N}, \tag{75}$$

by the von Neumann formula, see [4],

$$A_0^+ u = \sum_{s=1}^{d} [-i\ x_s\, f_s + i\ \hat{x}_s\, \hat{f}_s]. \tag{76}$$

In order to use the symplectic version of the operator-extension techniques we introduce in the defect a new basis $W_s^\pm$, on which the formal adjoint A_0^+ is correctly defined due to the above non-overlapping condition:

$$W_s^+ = \frac{f_s + \hat{f}_s}{2} = \frac{A}{A-iI} f_s,\ W_s^- = \frac{f_s - \hat{f}_s}{2i} = -\frac{I}{A-iI} f_s,$$

$$A_0^+ W_s^+ = W_s^-,\ A_0^+ W_s^- = -W_s^+.$$

It is convenient to represent elements $u \in \mathcal{N}$ via the new basis as

$$u = \sum_{s=1}^{d} [\xi_s^+\, W_s^+ + \xi_s^-\, W_s^-]. \tag{77}$$

Then, using notations $\sum_{s=1}^{d}\xi_{s,\pm}\,e_s := \vec{\xi}_{\pm}$ we re-write the above von Neumann formula as

$$u=\frac{A}{A-iI}\vec{\xi}^{u}_{+}-\frac{1}{A-iI}\vec{\xi}^{u}_{-},\quad A_0^{+}u=-\frac{1}{A-iI}\vec{\xi}^{u}_{+}-\frac{A}{A-iI}\vec{\xi}^{u}_{-} \tag{78}$$

The following formula of integration by parts for abstract operators was proved in ([54]):

Lemma 5.2. *Consider the elements u, v from the domain of the (formal) adjoint operator A_0^{+}:*

$$u=\frac{A}{A-iI}\vec{\xi}^{u}_{+}-\frac{1}{A-iI}\vec{\xi}^{u}_{-},\ v=\frac{A}{A-iI}\vec{\xi}^{v}_{+}-\frac{1}{A-iI}\vec{\xi}^{v}_{-}$$

with coordinates $\vec{\xi}^{u}_{\pm}$, $\vec{\xi}^{v}_{\pm}$:

$$\vec{\xi}^{u}_{\pm}=\sum_{s=1}^{d}\xi^{u}_{s,\pm}f_{s,i}\in N_i,\ \vec{\xi}^{v}_{\pm}=\sum_{s=1}^{d}\xi^{v}_{s,\pm}f_{s}\in N_i.$$

Then, the boundary form of the formal adjoint operator is equal to

$$\mathcal{J}_A(u,v)=\langle A_0^{+}u,v\rangle-\langle u,A_0^{+}v\rangle=\langle\vec{\xi}^{u}_{+},\vec{\xi}^{v}_{-}\rangle_N-\langle\vec{\xi}^{u}_{-},\vec{\xi}^{v}_{+}\rangle_N. \tag{79}$$

One can see that the coordinates $\vec{\xi}^{u}_{\pm}$, $\vec{\xi}^{v}_{\pm}$ of the elements u, v play the role of the boundary values $\{U'(0),\,U(0),\,V'(0),\,V(0)\}$. We will call them *symplectic coordinates* of the element u, v. The next statement proved in [54] is the core detail of the fundamental Krein formula [36], for generalized resolvents of symmetric operators. In our situation, it is used in course of calculation of the scattering matrix.

Lemma 5.3. *The vector-valued function of the spectral parameter*

$$u(\lambda)=\frac{A+iI}{A-\lambda I}\ \vec{\xi}^{u}_{+}:=u_0+\frac{A}{A-iI}\vec{\xi}^{u}_{+}-\frac{1}{A-iI}\vec{\xi}^{u}_{-}, \tag{80}$$

satisfies the adjoint equation $[A_0^{+}-\lambda I]u=0$, and the symplectic coordinates $\vec{\xi}^{u}_{\pm}\in N_i$ of it are connected by the formula

$$\vec{\xi}^{u}_{-}=-P_{N_i}\frac{I+\lambda A}{A-\lambda}\vec{\xi}^{u}_{+} \tag{81}$$

Proof. See in Appendix, Subsection 9.1 or in [54]. □

Introduce the map

$$P_{N_i}\frac{I+\lambda A}{A-\lambda I}P_{N_i}=:-\mathcal{M}:N_i\to N_i.$$

The matrix function $\mathcal{M}=P_{N_i}AP_{N_i}-P_{N_i}\frac{I+A^2}{A-\lambda I}P_{N_i}$ has a negative imaginary part in the upper half-plane $\Im\lambda>0$ and serves an abstract analog of the celebrated Weyl-Titchmarsh function. The operator function $\mathcal{M}$ exists almost everywhere on the real axis λ, and has a finite number of simple poles at the eigenvalues α_r^2 of

A. This function plays an important role in description of spectral properties of selfadjoint extensions of symmetric operators, see [36, 23].

We construct a solvable model of the quantum network as a selfadjoint extension of the orthogonal sum $l_0 \oplus A_0$. We consider the orthogonal sum of the corresponding adjoint l_0^+ and the formal adjoint: $l_0^+ \oplus A_0^+$, and calculate the corresponding boundary form $\mathbf{J}(\mathbf{U}, \mathbf{V}) := \mathcal{J}(U, V) + \mathcal{J}(u, v)$ on elements $(U, u) := \mathbf{U}$ from the orthogonal sum of the corresponding spaces. The selfadjoint extensions of the operator $l_0 \oplus A_0$ are obtained, based on restrictions of the adjoint operator $\mathbf{A}_0^+ = l_0^+ \oplus A_0^+$ onto Lagrangian planes of the form $\mathbf{J}(\mathbf{U}, \mathbf{V})$. These planes may be defined by the boundary conditions connecting symplectic coordinates $U'(0)$, $U(0)$, $\vec{\xi}_+^u$, $\vec{\xi}_-^u$ of components of corresponding elements in the deficiency subspaces. For instance, one may select a finite-dimensional operator $\beta : E_+ \oplus N_i \to E_+ \oplus N_i$ and define the Lagrangian plane $\mathbf{L}_\beta$ by the boundary condition

$$\begin{pmatrix} U'(0) \\ \vec{\xi}_+ \end{pmatrix} = \begin{pmatrix} \beta_{00} & \beta_{01} \\ \beta_{01}^+ & 0 \end{pmatrix} \begin{pmatrix} U(0) \\ \vec{\xi}_- \end{pmatrix}. \tag{82}$$

The extension defined by (82) on the Lagrangian plane is continued onto the whole space $L_2(E_+, R_+) \oplus E_A$ by forming the direct sum with the closure of the restricted operator A_0, see [35]. This construction gives a selfadjoint extension $\mathbf{A}_\beta$ of $l_0 \oplus A_0$ in $L_2(E_+, R_+) \oplus E_A$, defined by the boundary condition (82).

The absolutely continuous spectrum of the operator $\mathbf{A}_\beta$ coincides with the spectrum of the exterior part of the model, and hence it coincides with the spectrum of the trivial component l_Λ of the split operator $\mathcal{L}_\Lambda$ (in the open channels). The corresponding eigenfunctions of $\mathbf{A}_\beta$ on the first spectral band $\Delta_1 \supset \Lambda$ can be found, see [5], via substitution into the above boundary condition for the column, combined of the Scattering Ansatz in the open channels with (80), and, in the outer space, with $K_+ = \sqrt{\lambda - V_\delta - \pi^2 \delta^{-2}}$:

$$\Psi = \begin{pmatrix} e^{iK_+ x}\nu + e^{-i\,K_+ x}\mathbf{S}\nu \\ \frac{A+iI}{A-\lambda I}\vec{\xi}_+^u \end{pmatrix}, \tag{83}$$

with $\beta_{10} = \beta_{01}^+$. It gives the linear equation for the scattering matrix:

$$\begin{pmatrix} iK_+\,(\nu - S\nu) \\ \vec{\xi}_+ \end{pmatrix} = \begin{pmatrix} \beta_{00} & \beta_{01} \\ \beta_{10} & 0 \end{pmatrix} \begin{pmatrix} \nu + S\nu \\ \mathcal{M}\vec{\xi}_+ \end{pmatrix}.$$

Solving this equation we obtain the scattered waves and the scattering matrix:

Lemma 5.4. *The scattering matrix for the constructed extension is an analytic function of the spectral parameter λ:*

$$\mathbf{S}(\lambda) = \frac{iK_+ - [\beta_{00} + \beta_{01}\mathcal{M}\beta_{10}]}{iK_+ + [\beta_{00} + \beta_{01}\mathcal{M}\beta_{10}]}, \tag{84}$$

with the denominator of the fraction preceding the numerator. The coordinate $\vec{\xi}_+$ *of the inner component of the scattered wave* (83) *is defined as*

$$\vec{\xi}_+ = \beta_{10}\frac{2ip}{ip + [\beta_{00} + \beta_{01}\mathcal{M}\beta_{10}]},$$

with $p = \sqrt{\lambda - V_\delta + \pi^2\,\delta^{-2}}$.

6. Fitting of the solvable model

It remains to choose the eigenvalues of A, the subspace N_i and the matrix parameter β, such that the operator-function $[\beta_{00} + \beta_{01}\mathcal{M}\beta_{10}]$ acting in E_+ coincides with the essential DN-map $\mathcal{DN}_\Delta^\Lambda$ of the intermediate Hamiltonian. Denote by Q_s the spectral projection corresponding to the eigenvalue k_s^2 of A, framed by the projections P_i onto the deficiency subspace N_i

$$Q_s = P_i P_s P_i.$$

Then the above expression takes the form:

$$[\beta_{00} + \beta_{01}\mathcal{M}\beta_{10}] = \left[\beta_{00} + \sum_{r=1}^{N_T}\alpha_r^2\beta_{01}Q_{sr}\beta_{10}\right] - \sum_r \frac{1+\alpha_r^4}{\alpha_r^2 - \lambda}\beta_{01}Q_r\beta_{10}. \tag{85}$$

We will define the boundary parameters $\beta_{10},\ \beta_{01} = \beta_{10}^+$ later, but once they are defined, we choose β_{00} such that the first summand in the right side of (85) coincides with k_M $\beta_{00} + \sum_r \alpha_r^2\beta_{01}Q_r\beta_{10} = -k_M$. Then the scattering matrix takes the form:

$$\mathbf{S}(k) = \frac{iK_+ - k_M + \sum_{r=1}^N \frac{1+\alpha_r^4}{\alpha_r^2-\lambda}\beta_{01}Q_r\beta_{10}}{iK_+ + k_M - \sum_{r=1}^N \frac{1+\alpha_r^4}{\alpha_r^2-\lambda}\beta_{01}Q_r\beta_{10}}, \tag{86}$$

which coincides with the essential scattering matrix if and only if the corresponding Krein function

$$k_M - \sum_{r=1}^N \frac{1+\alpha_r^4}{\alpha_r^2-\lambda}\beta_{01}Q_r\beta_{10} \tag{87}$$

coincides with the essential part $\mathcal{DN}_\Delta^\Lambda$ of the DN-map of the intermediate Hamiltonian on the essential spectral interval Δ_T:

$$\mathcal{DN}^\Lambda \approx k(\lambda) + \sum_{r=1}^N \frac{P_+\frac{\partial\varphi_r}{\partial n}\rangle\langle P_+\frac{\partial\varphi_r}{\partial n}}{\lambda_r - \lambda}. \tag{88}$$

Summarizing these results we obtain the following conditional statement for the extension constructed based on the boundary condition (82) in case when $N_i \cap N_{-i} = 0$ or $\dim E_A = 1$:

Theorem 6.1. *The constructed operator* $\mathbf{A}_\beta$ *is a solvable model of the Quantum network on the essential interval* Δ*, if and only if the dimension of the space*

E_A coincides with the number N of eigenvalues of the intermediate operator on $\Delta \subset [\lambda_{\max}, \lambda_{\min}]$, the eigenvalues α_r^2 of the inner Hamiltonian $A = \sum_{r=1}^N \alpha_r^2 \nu_r \rangle \langle \nu_r$ coincide with eigenvalues of the intermediate operator on Δ, there exists a deficiency subspace N_i of the inner Hamiltonian such that $N_i \cap \frac{A+iI}{A-iI} N_i = 0$ and the operator $\beta_{01} : N_i \to E_+$ such that for the orthonormal basis $\{e_s\}_{s=1}^N$ of eigenvectors of A in E_A

$$P_+ \frac{\partial \Psi_r}{\partial n} = [1 + \alpha_r^4]^{1/2} \beta_{01} P_{N_i} \nu_r,\ r = 1, 2, \ldots, N. \tag{89}$$

Eliminating the inner variables, we can reduce the model to the Schrödinger equation with the constant potential on open channels, and appropriate boundary conditions on the bottom sections:

$$\left.\frac{dU^\omega}{dx}\right|_\Gamma = \left[k_M - \sum_{r=1}^N \frac{P_+ \frac{\partial \Psi_r}{\partial n} \rangle \langle P_+ \frac{\partial \Psi_r}{\partial n}}{\lambda_r - \lambda} U^\omega \right] \Bigg|_\Gamma . \tag{90}$$

Unfortunately, this straightforward construction does not fulfil basic requirements of quantum mechanics, and hence we proceed via construction a selfadjoint operator in $[L_2(0, \infty) \times E_+] \oplus E_A)$.

Dr. M. Harmer suggested an important strengthening of the previous conditional statement, by proving a general theorem of existence of the subspace N_i and the projection P_{N_i} which satisfy the condition of Theorem 4.1. The proof we provide below only slightly differs from the original proof in [27]: we added an explicit formulae for β_{01}, P_{N_i} in terms of the corresponding Gram matrix.

Denote by L_Λ^Δ the restriction of the intermediate operator L_Λ onto the invariant subspace $E_\Delta = E_A$ corresponding to the part $\sigma_\Delta = \{\lambda_1, \lambda_2, \ldots, \lambda_N\}$ of its spectrum on the essential interval Δ, and consider the linear map

$$\sum_s [1 + \alpha_s^4]^{-1/2} P_+ \left.\frac{\partial \varphi_s}{\partial n}\right|_\Gamma \langle *, \varphi_s \rangle := \Phi_\Delta \tag{91}$$

from E_Δ to E_+, $\dim E_+ = \mathbf{n}$.

Theorem 6.2 (M. Harmer). *The map Φ_Δ defines a one-to one correspondence between two d-dimensional subspaces, $2d < N$:*

$$\Phi_\Delta^+ \Phi_\Delta E_\Delta := N_\Delta \subset E_\Delta \ \text{ and } \ \Phi_\Delta \Phi_\Delta^+ E_+ := E_+^\Delta \subset E_+$$

If the subspace N_Δ is a generating subspace of L_Λ^Δ and

$$N_\Delta \cap \left(L_\Lambda^\Delta - iI\right)^{-1} \left(L_\Lambda^\Delta + iI\right) N_\Delta = 0, \tag{92}$$

then there exist a unique pair of the boundary operator $\beta_{01} : E_A \to E_+$ and the subspace $N_i \subset E_A$, which satisfy the condition of the previous theorem.

Remark 2. This theorem gives an interpretation of the solvable model described in Theorem 4.1, in terms of the intermediate Hamiltonian via selection of the inner

Hamiltonian A as a part L_Λ^Δ of L_Λ in the invariant subspace corresponding to the essential spectral interval Δ. The subspace $N_\Delta \subset E_\Delta$ plays the role of the deficiency subspace N_i of the inner Hamiltonian and $(L_\Lambda^\Delta - iI)^{-1}(L_\Lambda^\Delta + iI)N_\Delta$ plays the role of the dual subspace N_{-i}.

Proof. The map Φ_Δ is represented by the $\mathbf{n} \times N$ matrix Φ of columns ϕ_s with respect to the orthogonal basis of cse-functions $\{e_t\}$ in E_+. The condition (89) is equivalent to the representation of the operator Φ_Δ in form $\beta_{01}\ P_{N_i}$, where β_{01} is a bounded operator acting from E_A into E_+ and P_{N_i} is an orthogonal projection in E_A onto the deficiency subspace N_i. We will construct both β_{01}, P_{N_i} from the data encoded in Φ_Δ.

The non-negative Gram operator $\Phi_\Delta\ \Phi_\Delta^+$ in E_+ has the spectral representation

$$\Phi_\Delta\ \Phi_\Delta^+ = U^+ D\ U.$$

The non-negative diagonal matrix D is invertible on the orthogonal complement $\hat{E}$ of the corresponding null-space $\hat{E}_0$. We denote the restriction D onto $\hat{E}$ by $\hat{D}$. One can assume that the subspace $\hat{E}$ belongs to some extended space $\hat{E} \oplus \hat{E}_0$ which contains E_A, and the operator U^+ acts from $\hat{E} \oplus \hat{E}_0$ onto E_+ as an isometry. The operator $U^+\hat{D}^{1/2}$ coincides with Φ_Δ. Hence the operator Φ_Δ is presented as a product $\beta\hat{P}$, with $\beta = \beta_{01} = U^+\hat{D}^{1/2} : \hat{E} \to E_+$ and $\hat{P} = P_{\hat{E}} := P_{N_i} \subset E_A$, $\dim N_i = \dim \hat{E} = d$ and coincides with the dimension of the resonance entrance subspace of the intermediate operator. Up to some non-essential isometry we may assume that $E_A = E_\Delta, A = L_\Lambda^\Delta,\ N_i = N_\Delta$. The condition (92) guarantees that $N_i \cup N_{-i} = 0$. □

In the case when only one resonance eigenvalue α_0^2 of the intermediate operator sits on the essential spectral band, the obtained model scattering matrix

$$\mathbf{S}(p) = \frac{iK_+ - k_M + \frac{1+\alpha_0^4}{\alpha_0^2-\lambda}\beta_{01}Q_0\beta_{10}}{iK_+ + k_M - \frac{1+\alpha_0^4}{\alpha_0^2-\lambda}\beta_{01}Q_0\beta_{10}} \tag{93}$$

is a single-pole approximation of the scattering matrix of the network. The condition of the above theorem is obviously fulfilled for the single-pole approximation, when $P_+\frac{\partial\varphi_0}{\partial n} \neq 0$, $d = 1$, $N = 1$, and β_0 is a one-dimensional operator mapping the one-dimensional subspace N_i onto the resonance entrance subspace in E_+ spanned by $P_+\frac{\partial\varphi_0}{\partial n}$. For thin or shrinking networks one can estimate, (see [7] and more details in [43, 44]) the deviation of the single-pole and/or few-poles approximations from the exact scattering matrix on the network, in terms of the ratio $d/\operatorname{diam}\Omega_{\text{in}}$.

We postpone the discussion of the non-stationary scattering matrix for QN to forthcoming publications. But we notice here that the local wave operators (see [9]) and the corresponding scattering matrix on the essential spectral band can be defined for the pair $(\mathcal{L}, \mathbf{A}_\beta)$.

7. A solvable model as a jump-start in the analytic perturbation procedure

Recall that the exact scattering matrix was approximated by the essential or approximate scattering matrix. In this section we consider this phenomenon from the point of view of complex analysis, for the simplest star-shaped network constructed of a single model quantum well with one semi-infinite wire attached to it.

Consider a thin quantum network constructed of a quantum well Ω_{in} and a single quantum wire of width δ attached to it, $\delta/\operatorname{diam}\Omega_{\text{in}} \ll 1$. Assume that the Fermi level is situated on the first spectral band in the wire, which has multiplicity 1. Without loss of generality we may assume that the component of the corresponding solvable model in the open channel is presented by the Schrödinger equation with $2\mu^{\|} = 2\mu^{\perp} = I$, $V^{\omega} = V$, $K_+ = p = \sqrt{\lambda - \frac{\pi^2}{\delta^2} - V}$ and one-dimensional subspace E_+:

$$-u'' = p^2 u,\ 0 < x < \infty. \tag{94}$$

Assume that the model Hamiltonian $\mathbf{A}_\beta$ is constructed as suggested in the previous section based on the "inner Hamiltonian" A, the differential operator l_Λ and the boundary parameters which are reduced to the coupling constant $\beta_{01} := \beta$. Hereafter we will use the re-normalized eigenvalues $\alpha_s^2 - \frac{\pi^2}{\delta^2} - V := k_s^2 > 0$. Introducing that notation into the Krein function, and submitting the boundary parameter to the condition $\beta_{00} + \sum_{s=1}^{N} \alpha_s^2 \beta_{01}\, q_s\, \beta_{10} = 0$ we obtain the corresponding few-pole scattering matrix as a function of the wave-number p, with physically meaningful limit behavior at infinity $\mathbf{S}^\beta(p) \to I$:

$$\mathbf{S}^\beta(p) = \frac{ip - k - \beta^2 \sum_s \frac{1+\alpha_s^4}{p^2 - k_s^2} q_s}{ip + k + \beta^2 \sum_s \frac{1+\alpha_s^4}{p^2 - k_s^2} q_s}. \tag{95}$$

Here $q_s = |\langle e, e_s\rangle|^2$. Zeros $p_s(\beta)$ of the Scattering matrix (95) – the resonances – sit in the upper half-plane $\Im p > 0$ and approach the points $\pm k_s$, $k_{-s} := -k_s$, when $\beta \to 0$.

Assume that the resonance eigenvalue $\alpha_0^2 = k_0^2 + \delta^{-2}\pi^2 + V$ is situated close to the scaled Fermi-level Λ and the coupling constant $\beta := \beta_{01}$ is relatively small, see below. Separating the resonance term in the numerator and denominator of (95)

$$\left[ip - \beta^2 \frac{1+\alpha_0^4}{k_0^2 - p^2} q_0\right] + \left[k + \beta^2 \sum_{s\neq 0} \frac{1+\alpha_s^4}{p^2 - k_s^2} q_s\right],$$

and multiplying by $(ip)^{-1}$ one can see that $-\frac{\beta^2}{ip}\left[k + \sum_{s\neq 0} \frac{1+\alpha_s^4}{p^2-k_s^2} q_s\right]$ plays the role of the small parameter. The resonance $k_0(\beta)$ originated from the eigenvalue k_0^2 of the operator A (more precisely: from the point $+k_0$) can be obtained as a solution

$p = k_0(\beta)$ of the equation

$$p = k_0 - \frac{\beta^2(1+\alpha_0^4)q_0}{(p+k_0)\left(ip + k + \beta^2 \sum_{s\neq 0} \frac{1+\alpha_s^4}{p^2-k_s^2} q_s\right)}. \tag{96}$$

Another resonance originated from the point $-k_0$ corresponds to the same eigenvalue, and it sits at the symmetric point $-\bar{k}_0(\beta)$ with respect to the imaginary axis. Remaining resonances $k_s(\beta)$, $s \neq 0$, can be found from similar equations. All functions $k_s(\beta)$ are analytic functions of β, k in small neighborhoods of $(0, \pm k_s)$. They sit in the upper half-plane symmetrically with respect to the imaginary axis $k_{-s}(\beta) = -\bar{k}_s(\beta)$. The scattering matrix (95) is unitary on the real axis k and has poles at the complex-conjugate points $\bar{k}_s(\beta)$ in the lower half-plane, and hence it is presented by the finite Blaschke product which tends to 1 when $|k| \to \infty$:

$$\mathbf{S}^\beta(p) = \prod_s \frac{p - k_s(\beta)}{p - \bar{k}_s(\beta)}. \tag{97}$$

The outer component of the scattered wave is presented as

$$\Psi_0^\beta = e^{-ipx} + \mathbf{S}^\beta(k) e^{ipx}. \tag{98}$$

It fulfills appropriate boundary condition at the place of contact with the model quantum dot. The inner component of the scattered wave can be obtained from Lemma 3.4.

We explore the model scattering problem for small values of β. Though the resonances depend analytically on β, neither the scattering matrix (95,97) nor the scattered wave depend analytically of (β, p) on the product of a small neighborhood of the origin in β-plane and small neighborhoods of $\pm k_0$ in p-plane. The analyticity is lost due to presence of the points $\pm k_0$ where the resonances are created at $\beta = 0$: both $k_0(\beta)$ and $\bar{k}_0(\beta)$ approach the same point k_0 when $\beta \to 0$. The corresponding "resonance" factor of the scattering matrix

$$\mathbf{S}_0^\beta(p) = \frac{[p - k_0(\beta)]\,[p + \bar{k}_0(\beta)]}{[p - \bar{k}_0(\beta)]\,[p + k_0(\beta)]}, \tag{99}$$

is non-analytic on $(\Omega_\beta \times \Omega_{k_0} \times \Omega_{-k_0})$, though $k_\beta = \Re k_\beta + i\Im k_\beta$ is analytic function of β due to (96). But the complementary factor of the scattering matrix

$$\mathbf{S}_\beta^0(p) = \prod_{s\neq 0} \frac{p - k_s(\beta)}{p - \bar{k}_s(\beta)}. \tag{100}$$

is analytic on $(\Omega_\beta \times \Omega_{k_0} \times \Omega_{-k_0})$ and can be expanded into the power series over β^m, $m = 0, 1, 2, \ldots$. Assume now that the function $k_0(\beta)$ is known. Then the following statement is true:

Theorem 7.1. *There exists a one-dimensional perturbation $\mathbf{A}_0^\beta$ of the operator*

$$l_0 u = -u'', \; u\big|_0 = 0$$

with a non-trivial inner component, such that the scattering matrix of the pair $(\mathbf{A}_0^\beta, l_0)$ *coincides with* $-\mathbf{S}_0^\beta(p)$*:*

$$\mathbf{S}(A_0^\beta, l_0) = -\mathbf{S}_0^\beta\ (l_0)$$

Then the scattering matrix $\mathbf{S}(\mathbf{A}_\beta, \mathbf{A}_0^\beta)$ *of the complementary pair* $(\mathbf{A}_\beta, \mathbf{A}_0^\beta)$ *is equal to the complementary factor* $-\mathbf{S}_\beta^0(p)$*:*

$$\mathbf{S}(\mathbf{A}_\beta, \mathbf{A}_0^\beta) = -\mathbf{S}_\beta^0(A_0^\beta),$$

such that after the change of the variable we obtain the chain rule for the scattering matrices

$$\mathbf{S}^\beta(l_0) = \mathbf{S}_0^\beta(l_0)\ \mathbf{S}_\beta^0(l_0)$$

or

$$\mathbf{S}^\beta(p) = \mathbf{S}_0^\beta(p)\ \mathbf{S}_\beta^0(p).$$

The complementary factor is an analytic function of $(\beta,\ k)$ *on the product* $(\Omega_\beta \times \Omega_{k_0} \times \Omega_{-k_0})$ *of a small neighborhood of the origin in* β*-plane and a small neighborhood of the pair* $(k_0,\ -k_0)$ *in the p-plane.*

A *proof* can be obtained as a combination of [60, 2]. □

Corollary 7.1. *The exterior component of the scattered wave of the operator* $\mathbf{A}_0^\beta$ *presented by the Ansatz* (98) *with* $\mathbf{S}_\beta$ *taken in the form* (97) *is not analytic with respect to the coupling constant* $\beta_{01} := \beta$ *near the origin. The non-analyticity of the scattered wave is caused by the presence of the non-analytic factor* $-\mathbf{S}_0^\beta$ *in the scattering matrix. From Theorem* 4.1*, we interpret this factor as the scattering matrix for the pair* $\left(\mathbf{A}_0^\beta,\ l_0\right)$*. The complementary factor* $-\mathbf{S}_\beta^0$ *is analytic with respect to the coupling constant* β*. It can be interpreted as the scattering matrix for the pair* $\left(\mathbf{A}^\beta, \mathbf{A}_0^\beta\right)$*. Summarizing our observation we suggest, for our example, the following two-steps modification of the analytic perturbation procedure on continuous spectrum:*

a) *First step is the construction of the solvable model and calculation of the corresponding (non-analytic with respect to the coupling constant* β *at the origin) scattering matrix. This is "the jump-start" of the analytic perturbation procedure.*
b) *Second step is the calculation of the analytic factor of the scattering matrix of the model by the standard analytic perturbation procedure. The analytic factor is interpreted as the scattering matrix between the constructed solvable model and the perturbed operator* $\mathbf{A}^\beta$*.*

The obtained connection between resonances and analytic perturbation series on the continuous spectrum recalls the connection between small denominators in celestial mechanics and divergence of perturbation series, observed by H. Poincaré, see [63]. More historical comments about intermediate Hamiltonian and the jump-start may be found in [56], where similar modification of the analytic perturbation procedure for the Friedrichs model is suggested.

Remark 3. Note that recovering *exact* information on the resonance $k_0(\beta)$ and on the corresponding residue for the perturbed operator $\mathbf{A}_\beta$, which we need to develop the "jump-start" procedure, may be a tricky problem almost equivalent to the original spectral problem. On the other hand, if the *approximate* resonance factor $\mathbf{S}_0^\beta$ is used instead the exact factor, then the division of the scattering matrix through $\mathbf{S}_0^\beta$ would not eliminate singularity, hence the complementary factor of the scattering matrix would be still non-analytic at the origin and hence could not be obtained via analytic perturbation procedure.

8. Appendix: Symplectic operator extension procedure

John von Neumann in 1933 has found conditions which guarantee existence of a selfadjoint extension of given unbounded symmetric operator, and suggested a procedure of construction of the extension, see symplectic version in [54]. For given symmetric operator $\mathcal{A}_0$ defined on D_0 in the Hilbert space H, see [4] and given complex value λ, $\Im\lambda \neq 0$ of the spectral parameter:

Definition 8.1. *Define the* **deficiency subspaces**

$$N_\lambda := H \ominus \overline{[\mathcal{A}_0 - \lambda I]\ D_0},$$

$$N_{\bar\lambda} := H \ominus \overline{\left[\mathcal{A}_0 - \bar\lambda I\right]\ D_0}.$$

The dimension of N_λ, $N_{\bar\lambda}$ is constant on the whole upper and lower spectral half-plane $\Im\lambda > 0$, $\Im\lambda < 0$ respectively.

Definition 8.2. *Introduce the deficiency index* $(\dim N_\lambda, \dim N_{\bar\lambda}) := (n_+, n_-)$ *of the operator* $\mathcal{A}_0$.

J. von Neumann proved that

Theorem 8.1. *The Hermitian operator* $\mathcal{A}_0$ *has a selfadjoint extension if and only if* $n_+ = n_-$.

The idea of construction of the extension is based on the following theorems von Neumann, see for instance [4]:

Theorem 8.2. *The domain of the adjoint operator is represented as a direct sum of the domain* $D_{\bar A_0}$ *of the closure and the deficiency subspaces, in particular:*

$$D_{A_0^+} = D_{\bar A_0} + N_i + N_{-i}.$$

The deficiency subspaces of the densely-defined operator are the eigenspaces of the adjoint operator:

$$\mathcal{A}_0^+ e_i = -ie_i,\ e_i \in N_i,\ \mathcal{A}_0^+ e_{-i} = ie_{-i},\ e_{-i} \in N_{-i}.$$

Theorem 8.3. *If $\mathcal{A}_0$ is an Hermitian operator with deficiency indices (n_+, n_-), $n_- = n_+$ and V is an isometry $V : N_i \to N_{-i}$. Then the isometry V defines a selfadjoint extension $\mathcal{A}_V$ of $\mathcal{A}_0$, acting on the domain*

$$D_{A_V} = D_{\bar{A}_0} + \{e_i + Ve_i,\ e_i \in N_i\}$$

as a restriction of $\mathcal{A}_0^+$ onto D_{A_V}:

$$A_V : u_0 + e_i + Ve_i \to \bar{\mathcal{A}}_0 u_0 - ie_i + iVe_i.$$

J. von Neumann reduced the construction of the extension of the symmetric operator $\mathcal{A}_0$ to an equivalent problem of construction of an extension of the corresponding isometrical operator – the Caley transform of $\mathcal{A}_0$. It is much more convenient, for differential operators, to construct the extensions based on so-called *boundary form*.

Example 5. Symplectic Extension procedure for the differential operator. Consider the second-order differential operator

$$L_0 u = -\frac{d^2 u}{dx^2},$$

defined on all square integrable functions, $u \in L_2(0, \infty)$, with square-integrable derivatives of the first and second order and vanishing near the origin. This operator is symmetric and its adjoint L_0^+ is defined by the same differential expression on all square integrable functions with square integrable derivatives of the first and second order and no boundary condition at the origin. This operator is not symmetric: its boundary form

$$\mathcal{J}(u, v) = \langle L_0^+ u, v\rangle - \langle u, L_0^+ v\rangle = u'(0)\bar{v}(0) - u(0)\bar{v}'(0),\ u, v \in D_{L_0^+}$$

is generally non equal to zero for $u, v \in D_{L_0^+}$. But it vanishes on a "Lagrangian plane" $\mathcal{P}_\gamma \subset D_{L_0^+}$ defined by the boundary condition

$$u'(0) = \gamma u(0),\ \gamma = \bar{\gamma}.$$

The restriction L_γ of the L_0^+ onto the Lagrangian plane $\mathcal{P}_\gamma$ is a selfadjoint operator in $L_2(0, \infty)$: it is symmetric, and the inverse of it $(L_\gamma - \lambda I)^{-1}$, at each complex spectral point λ, exists and is defined on the whole space $L_2(0, \infty)$.

The operator extension procedure used above for the differential operator, can be applied to general symmetric operators and serves a convenient alternative for construction of solvable models of orthogonal sums of differential operators and finite matrices. We call the abstract analog of the extension procedure the *symplectic extension procedure*. Let A be a selfadjoint operator in a finite-dimensional Hilbert space E, $\dim E = d$, and $N_i := N$ is a subspace of E, $\dim N = n < d/2$, which does not overlap with $\frac{A+iI}{A-iI} N_i := N_{-i}$:

$$N_i \cap N_{-i} = \{0\}.$$

Define the operator A_0 as a restriction of A onto $D_0 := \frac{I}{A-iI} E \ominus N$. This operator is symmetric, and the subspaces $N_{\pm i}$ play roles of its deficiency subspaces. The

operator can A_0 can be extended to the selfadjoint operator $A_\Gamma \supset A_0$ via symplectic extension procedure involving the corresponding boundary form: selecting a basis $\{e_s^+\}_{s=1}^n := g_s \in N_i$, we consider the dual basis $\left\{\frac{A+iI}{A-iI}g_s = g_s^-\right\}_{s=1}^n \in N_i$. Introduce, following [54], another basis in the defect $N = N_i + N_{-i}$

$$W_s^+ = \frac{1}{2}\left[g_s + \frac{A+iI}{A-iI}g_s\right],\ W_s^- = \frac{1}{2i}\left[s_s - \frac{A+iI}{A-iI}g_s\right].$$

Due to $A_0^+ g_s + ig_s = 0,\ [A_0^+ - iI]\frac{A+iI}{A-iI}g_s = 0$ we have,

$$A_0^+ W_s^+ = W_s^-,\ A_0^+ W_s^- = -W_s^+.$$

Following [24] we will use the representation of elements from the domain of the adjoint operator by the expansion on the new basis:

$$u = u_0 + \sum_s \xi_+^s W_s^+ + \xi_-^s W_s^-,$$

with $u_0 \in D(A_0)$ and symplectic coordinates $\xi_\pm^s$.

We also introduce the *boundary vectors* of elements from $D(A_0^+)$

$$\vec{\xi}_\pm := \sum_s \xi_\pm^s g_s \in N_i,$$

$$u = u_0 + \frac{A}{A-iI}\vec{\xi_+^u} - \frac{I}{A-iI}\vec{\xi_-^u} := u_0 + n^u,\ u_0 \in D(A_0)\, n^u \in N.$$

Define the formal adjoint operator A_0^+ on the defect $\mathbf{N} = N_i + N_{-i}$ as:

$$A_0^+ e_+ = -ie_+, \text{ for } e_+ \in N_i,\ A_0^+ e_- = ie_-, \text{ for } e_+ \in N_{-i},$$

$$A_0^+(e_+ + e_-) = -ie_+ + ie_-.$$

Then we have:

$$A_0^+ W_s^+ = W_s^-,\ A_0^+ W_s^- = -W_s^+.$$

Following [54], we will use the representation of elements from the domain of the adjoint operator by the expansion on the new basis:

$$u = u_0 + \sum_s \xi_+^s W_s^+ + \xi_-^s W_s^-,$$

with $u_0 \in D(A_0)$ and symplectic coordinates $\xi_\pm^s$. We also introduce the *boundary vectors* of elements from $D(A_0^+)$

$$\vec{\xi}_\pm := \sum_s \xi_\pm^s g_s \in N_i,$$

$$u = u_0 + \frac{A}{A-iI}\vec{\xi_+^u} - \frac{I}{A-iI}\vec{\xi_-^u} := u_0 + n^u,\ u_0 \in D(A_0)\, n^u \in N.$$

Then the boundary form of A_0^+ is calculated as

$$\langle A_0^+ u, v\rangle - \langle u, A_0^+ v\rangle := \mathcal{J}(u,v) = \langle \vec{\xi_+^u}, \vec{\xi_-^v}\rangle - \langle \vec{\xi_-^u}, \vec{\xi_+^v}\rangle$$

8.1. Operator Extensions: Krein formula

Theorem: **Krein formula.** *Consider a closed symmetric operator A_0 in the Hilbert space $\mathcal{H}$, obtained via restriction of the selfadjoint operator A onto the dense domain $D(A_0)$, with finite-dimensional deficiency subspaces $N_{\mp i}$, $P_{N_i} := P_+$, $\dim N_i = \dim N_{-i}$. Then the resolvent of the selfadjoint extension A_M defined by the boundary conditions*

$$\vec{\xi}_+ = M\vec{\xi}_- \tag{101}$$

is represented, at regular points of A_M, by the formula:

$$(A_M - \lambda I)^{-1} = \frac{I}{A-\lambda I} - \frac{A+iI}{A-\lambda I}P_+ M \frac{I}{I + P_+ \frac{I+\lambda A}{A-\lambda I} P_+ M} P_+ \frac{A-iI}{A-\lambda I}$$

Proof. For the convenience of the reader we provide below the sketch of the proof of the Krein formula via symplectic operator extension procedure. Solution of the homogeneous equation $(A^+ - \lambda I)\, u = f$ is reduced to finding $u_0, \vec{\xi}_\pm$ from the equation

$$(A-\lambda I)u_0 - \frac{I+\lambda A}{A-iI}\vec{\xi}^u_+ - \frac{A-\lambda I}{A-iI}\vec{\xi}^u_- = f. \tag{102}$$

Applying to this expression the operator $\frac{A-iI}{A-\lambda I}$, due to $(A-iI)u_0 \perp N_i$, we obtain

$$\vec{\xi}_- = -\frac{I}{I + \frac{I+\lambda A}{A-\lambda I}P_+} P_+ \frac{A-iI}{A-\lambda I} f.$$

Then, from the above equation (102) and from the boundary condition (101), we derive:

$$u_0 = \frac{1}{A-iI}\left[\frac{I+\lambda A}{A-\lambda I}\vec{\xi}_+ + \vec{\xi}_-\right] + \frac{I}{A-\lambda I} f,$$

and

$$\begin{aligned} u &= u_0 + \frac{A}{A-iI}\vec{\xi}_+ - \frac{I}{A-iI}\vec{\xi}_- \\ &= \frac{I}{A-\lambda I} f - \frac{A+iI}{A-\lambda I} P_+ M \frac{I}{I + P_+ \frac{I+\lambda A}{A-\lambda I} P_+ M} P_+ \frac{A-iI}{A-\lambda I} f. \end{aligned}$$

□

Acknowledgement

The author acknowledges support from the Russian Academy of Sciences, Grant RFBR 03-01-00090. The author is grateful to V. Katsnelson for important references and very interesting materials provided, and to M. Harmer for deep remarks concerning the fitting of the star-graph model.

References

[1] V. Adamjan, D. Arov, *On a class of scattering operators and characteristic operator-functions of contractions.* (Russian) Dokl. Akad. Nauk SSSR 160 (1965) pp. 9–12.

[2] V. Adamyan, B. Pavlov, *Local Scattering problem and a Solvable model of Quantum Network.* In: Operator Theory: Advances and Applications, Vol. 198, Birkhäuser Verlag, Basel/Switzerland, (2009) pp. 1–10.

[3] V. Adamyan, B. Pavlov, A. Yafyasov, *Modified Krein Formula and analytic perturbation procedure for scattering on arbitrary junction.* International Newton Institute report series NI07016, Cambridge, 18 April 2007, 33p. To be published in the proceedings of M.G. Krein memorial conference, Odessa, April 2007.

[4] N.I. Akhiezer, I.M. Glazman, *Theory of Linear Operators in Hilbert Space*, (Frederick Ungar, Publ., New York, vol. 1, 1966) (translated from Russian by M. Nestel).

[5] S. Albeverio, P. Kurasov, *Singular Perturbations of Differential Operators*, London Math. Society Lecture Note Series 271. Cambridge University Press (2000).

[6] S. Albeverio, F. Gesztesy, R. Hoegh-Krohn, H. Holden, *Solvable models in quantum mechanics.* Springer-Verlag, New York, 1988.

[7] N. Bagraev, A. Mikhailova, B. Pavlov, L. Prokhorov, A. Yafyasov, *Parameter regime of a resonance quantum switch.* In: Phys. Rev. B, 71, 165308 (2005), pp. 1–16.

[8] F.A. Berezin, L.D. Faddeev, *A remark on Schrödinger equation with a singular potential* Dokl. AN SSSR, **137** (1961) pp. 1011–1014.

[9] M. Birman, *A local test for the existence of wave operators.* (Russian) Izv. Akad. Nauk SSSR Ser. Mat. 32 1968 914–942.

[10] V. Bogevolnov, A. Mikhailova, B. Pavlov, A. Yafyasov, *About Scattering on the Ring* In: "Operator Theory: Advances and Applications", Vol. 124 (Israel Gohberg Anniversary Volume), Ed. A. Dijksma, A.M. Kaashoek, A.C.M. Ran, Birkhäuser, Basel (2001) pp. 155–187.

[11] J. Bruening, B. Pavlov, *On calculation of Kirchhoff constants of Helmholtz resonator.* International Newton Institute report series NI07060-AGA, Cambridge, 04 September 2007, 40p.

[12] R. Courant, D. Hilbert, *Methods of mathematical physics.* Vol. II. *Partial differential equations.* Reprint of the 1962 original. Wiley Classics Library. A Wiley-Interscience Publication. John Wiley & Sons, Inc., New York (1989). xxii + 830 pp.

[13] S. Datta, *Electronic Transport in Mesoscopic systems.* Cambridge University Press, Cambridge (1995).

[14] S. Datta and B. Das Sarma, *Electronic analog of the electro-optic modulator.* Appl. Phys. Lett. **56**, 7 (1990) pp. 665–667.

[15] Yu.N. Demkov, V.N. Ostrovskij, *Zero-range potentials and their applications in Atomic Physics*, Plenum Press, NY-London, (1988).

[16] P. Exner, P. Ŝeba, *A new type of quantum interference transistor.* Phys. Lett. A 129:8,9, 477 (1988).

[17] P. Exner, O. Post, *Convergence of graph-like thin manifolds* J. Geom. Phys. **54**,1, (2005) pp. 77–115.

[18] M. Faddeev, B. Pavlov, *Scattering by resonator with the small opening.* Proc. LOMI, v126 (1983). (English Translation J. of Sov. Math. v. 27, 2527 (1984).

[19] E. Fermi, *Sul motto dei neutroni nelle sostance idrogenate* (in Italian) Richerca Scientifica **7** p. 13 (1936).

[20] F. Gesztesy, B.Simon, *Inverse spectral analysis with partial information on the potential. I. The case of an a.c. component in the spectrum.* Papers honouring the 60th birthday of Klaus Hepp and of Walter Hunziker, Part II (Zürich, 1995). Helv. Phys. Acta 70, no. 1-2, (1997) pp 66–71.

[21] F. Gesztezy, Y. Latushkin, M. Mitrea and M. Zinchenko, *Non-selfadjoint operators, infinite determinants and some applications*, Russian Journal of Mathematical Physics, **12**, 443–71 (2005).

[22] F. Gesztezy, M. Mitrea and M. Zinchenko, *On Dirichlet-to-Neumann maps and some applications to modified Fredholm determinants preprint*, (2006).

[23] V.I. Gorbachuk, M.L. Gorbachuk, *Boundary value problems for operator differential equations*. Translated and revised from the 1984 Russian original. Mathematics and its Applications (Soviet Series), 48. Kluwer Academic Publishers Group, Dordrecht, 1991. xii + 347

[24] D. Gramotnev, D. Pile, *Double resonant extremely asymmetrical scattering of electromagnetic waves in non-uniform periodic arrays* In: Opt. Quant. Electronics., **32**, (2000) pp. 1097–1124.

[25] D. Grieser, *Spectra of graph neighborhoods and scattering*. Proc. Lond. Math. Soc. (3) 97, no. 3 (2008) pp. 718–752.

[26] M. Harmer, Hermitian symplectic geometry and extension theory. *Journal of Physics A: Mathematical and General*, **33** (2000) pp. 9193–9203.

[27] M. Harmer, *Fitting parameters for a Solvable Model of a Quantum Network* The University of Auckland, Department of Mathematics report series 514 (2004), 8 p.

[28] M. Harmer, B. Pavlov, A. Yafyasov, *Boundary condition at the junction*, in: Journal of Computational Electronics, **6** (2007) pp. 153–157.

[29] T. Kato, *Perturbation theory for linear operators* Springer Verlag, Berlin-Heidelberg-NY, second edition (1976).

[30] J.P. Keating, J. Marlof, B. Winn, *Value distribution of the eigenfunctions and spectral determinants of quantum star-graphs* Communication of Mathematical Physics, **241**, 2-3 (2003) pp. 421–452.

[31] J.P. Keating, B. Winn, *No quantum ergodicity for star graphs* Communication of Mathematical Physics, **250**, 2 (2004) pp. 219–285.

[32] G.R. Kirchhoff. *Gesammelte Abhandlungen* Publ. Leipzig: Barth, 1882, 641p.

[33] J. Brüning, B. Pavlov, On calculation of Kirchhoff constants for Helmholtz resonator International Newton Institute, report series NI07060-AGA, Cambridge, 04 September, 2007, 38 p.

[34] V. Kostrykin and R. Schrader, *Kirchhoff's rule for quantum wires*. J. Phys. A: Math. Gen. **32**, 595 (1999).

[35] M.A. Krasnosel'skij, *On selfadjoint extensions of Hermitian Operators* (in Russian) Ukrainskij Mat. Journal **1**, 21 (1949).

[36] M.G. Krein, *Concerning the resolvents of an Hermitian operator with deficiency index* (m, m), Doklady AN USSR, **52** (1946) pp. 651–654.

[37] P. Kuchment, H. Zeng, *Convergence of Spectra of mesoscopic Systems Collapsing onto Graph* Journal of Mathematical Analysis and Applications, **258**(2001) pp. 671–700.

[38] P. Kuchment, *Graph models for waves in thin structures* Waves in Periodic and Random Media, **12**, 1 (2002) R 1–R 24.

[39] S. Lall, P. Krysl, J. Marsden, *Structure-preserving model reduction for mechanical systems* In: Complexity and nonlinearity in physical systems (Tucson, AZ, 2001), Phys. D **184**, 1-4 (2003) pp. 304–318.

[40] Lax, Peter D., Phillips, Ralph, S. *Scattering theory.* Second edition. With appendices by Cathleen S. Morawetz and Georg Schmidt. Pure and Applied Mathematics, 26. Academic Press, Inc., Boston, MA, (1989) xii + 309 pp.

[41] M.S. Livshits, *Method of nonselfadjoint operators in the theory of wave guides* In: Radio Engineering and Electronic Physics. Publ. by American Institute of Electrical Engineers, **1** (1962) pp. 260–275.

[42] O. Madelung, *Introduction to solid-state theory.* Translated from German by B.C. Taylor. Springer Series in Solid-State Sciences, 2. Springer-Verlag, Berlin, New York (1978).

[43] A. Mikhailova, B. Pavlov, L. Prokhorov, *Modeling of quantum networks*' arXiv math-ph: 031238, 2004, 69 p.

[44] A. Mikhailova, B. Pavlov, L. Prokhorov, *Intermediate Hamiltonian via Glazman splitting and analytic perturbation for meromorphic matrix-functions.* In: Mathematische Nachrichten, **280**, 12, (2007) pp. 1376–1416.

[45] A. Mikhailova, B. Pavlov, *Remark on the compensation of singularities in Krein's formula* In: *Operator Theory: Advances and Applications* Vol. 186, Proceedings of OTAMP06, Lund. Editors: S. Naboko, P. Kurasov. pp. 325–337.

[46] R. Mittra, S. Lee, *Analytical techniques in the theory of guided waves* The Macmillan Company, NY, Collier-Macmillan Limited, London, 1971.

[47] L. Ko, R. Mittra, *A new approach based on a combination of integral equation and asymptotic techniques for solving electromagnetic scattering problems* IEEE Trans. Antennas and Propagation AP-25, no. 2, (1977) pp. 187–197.

[48] J. von Neumann, *Mathematical foundations of quantum mechanics* Twelfth printing. Princeton Landmarks in Mathematics. Princeton Paperbacks. Princeton University Press, Princeton, NJ, (1996).

[49] R.G. Newton, *Scattering theory of waves and particles*Newton, Reprint of the 1982 second edition [Springer, New York; MR0666397 (84f:81001)], with list of errata prepared for this edition by the author. Dover Publications, Inc., Mineola, NY, 2002.

[50] N. Nikol'skii, S. Khrushchev, *A functional model and some problems of the spectral theory of functions* (Russian) Translated in Proc. Steklov Inst. Math. 1988, no. 3, 101–214. Mathematical physics and complex analysis (Russian). Trudy Mat. Inst. Steklov. 176 (1987), 97–210, 327.

[51] B. Sz.-Nagy, C., Foias,*Harmonic analysis of operators on Hilbert space.* Translated from the French and revised North-Holland Publishing Co., Amsterdam-London; American Elsevier Publishing Co., Inc., New York; Akadémiai Kiadó, Budapest 1970 xiii + 389 pp.

[52] S. Novikov, *Schrödinger operators on graphs and symplectic geometry* In: The Arnoldfest (Toronto, ON, 1997), Ed.: Fields Inst. Commun., 24, Amer. Math. Soc., Providence, RI, (1999) pp. 397–413.

[53] B. Pavlov, *On one-dimensional scattering of plane waves on an arbitrary potential*, Teor. i Mat. Fiz.,v. 16, N1, 1973, pp. 105–119.

[54] B. Pavlov, *The theory of extensions and explicitly solvable models* (in Russian) Uspekhi Mat. Nauk, **42**, (1987) pp. 99–131.

[55] B. Pavlov, *S-Matrix and Dirichlet-to-Neumann Operators* In: *Encyclopedia of Scattering*, ed. R. Pike, P. Sabatier, Academic Press, Harcourt Science and Tech. Company (2001) pp. 1678–1688.

[56] B. Pavlov, I. Antoniou, *Jump-start in analytic perturbation procedure for Friedrichs model.* In J. Phys. A: Math. Gen. 38 (2005) pp. 4811–4823.

[57] B. Pavlov, V. Kruglov, *Operator Extension technique for resonance scattering of neutrons by nuclei.* In: Hadronic Journal **28** (2005) pp. 259–268.

[58] B. Pavlov, V. Kruglov, *Symplectic operator-extension technique and zero-range quantum models* In: New Zealand mathematical Journal 34,2 (2005) pp. 125–142.

[59] B. Pavlov, A. Yafyasov, *Standing waves and resonance transport mechanism in quantum networks* With A. Yafyasov, Surface Science 601 (2007), pp. 2712–2716.

[60] B. Pavlov, *A star-graph model via operator extension* Mathematical Proceedings of the Cambridge Philosophical Society, Volume 142, Issue 02, March 2007, pp. 365–384.

[61] B. Pavlov, T. Rudakova, V. Ryzhii, I. Semenikhin, *Plasma waves in two-dimensional electron channels: propagation and trapped modes.* Russian Journal of mathematical Physics, **14**, 4 (2004) pp. 465–487.

[62] L. Petrova, B. Pavlov, *Tectonic plate under a localized boundary stress: fitting of a zero-range solvable model.* Journal of Physics A, **41** (2008) 085206 (15 pp.).

[63] H. Poincaré, *Methodes nouvelles de la mécanique celeste* Vol. 1 (1892), Second edition: Dover, New York (1957).

[64] C. Presilla, J. Sjostrand, *Transport properties in resonance tunnelling heterostructures* In: J. Math. Phys. **37**, 10 (1996), pp. 4816–4844.

[65] I. Prigogine, *Irreversibility as a Symmetry-breaking Process* In: Nature, **246**, 9 (1973).

[66] Lord Rayleigh, *The theory of Helmholtz resonator* Proc. Royal Soc. London **92** (1916) pp. 265–275.

[67] J. Rubinstein, M. Shatzman, *Variational approach on multiply connected thin strips I : Basic estimates and convergence of the Laplacian spectrum* Arch. Ration. Mech. Analysis **160**, 4, 271 (2001).

[68] M. Schatzman, *On the eigenvalues of the Laplace operator on a thin set with Neumann boundary conditions* Applicable Anal. **61**, 293 (1996).

[69] I.A. Shelykh, N.G. Galkin, and N.T. Bagraev, *Quantum splitter controlled by Rashba spin-orbit coupling. Phys. Rev.B* **72**, 235316 (2005).

[70] J.H. Schenker, M. Aizenman, *The creation of spectral gaps by graph decoration* Letters of Mathematical Physics, **53**, 3, (2000) pp. 253–262.

[71] J. Shirokov, *Strongly singular potentials in three-dimensional Quantum Mechanics* (In Russian) Teor. Mat. Fiz. **42** 1 (1980) pp. 45–49.

[72] J. Splettstoesser, M. Governale, and U. Zülicke, *Persistent current in ballistic mesoscopic rings with Rashba spin-orbit coupling. Phys. Rev. B*, **68**:165341, (2003).

[73] P. Streda, P. Seba, *Antisymmetric spin filtering in one-dimensional electron systems via uniform spin-orbit coupling* Phys. Rev. Letters **90**, 256601 (2003).

[74] J. Sylvester, G. Uhlmann, *The Dirichlet to Neumann map and applications.* In: *Proceedings of the Conference "Inverse problems in partial differential equations* (*Arcata,* 1989)*"*, SIAM, Philadelphia, 101 (1990).

[75] A. Wentzel, M. Freidlin, *Reaction-diffusion equations with randomly perturbed boundary conditions* Annals of Probability **20**, 2 (1992) pp. 963–986.

[76] E.P. Wigner, *On a class of analytic functions from the quantum theory of collisions* Annals of mathematics, **2**, N. 53, 36 (1951).

[77] H.Q. Xu, *Diode and transistor behaviour of three-terminal ballistic junctions* Applied Phys. Letters **80**, 853 (2002).

B. Pavlov
V.A. Fock Institute for Physics of St.-Petersburg University,
Petrodvorets, 198905, Russia.
e-mail: pavlovenator@gmail.com

Operator Theory:
Advances and Applications, Vol. 197, 337–373

Integral Equations in the Theory of Levy Processes

Lev Sakhnovich

The paper is dedicated to the memory of the outstanding mathematician M. Livshitz. I am happy and proud that I was his pupil.

Abstract. In the article we consider the Levy processes and the corresponding semigroups. We represent the generators of these semigroups in convolution forms. Using the obtained convolution form and the theory of integral equations we investigate the properties of a wide class of Levy processes (potential, quasi-potential, the probability of the Levy process remaining within the given domain,long time behavior). We analyze in detail a number of concrete examples of the Levy processes (the stable processes, the variance damped Levy processes, the variance gamma processes, the normal Gaussian process, the Meixner process, the compound Poisson process.)

Mathematics Subject Classification (2000). Primary 60G51; Secondary 60J45, 60G17, 45A05.

Keywords. Semigroup, generator, convolution form, potential, quasi-potential, sectorial operators, long time behavior.

Introduction

In the famous article by M. Kac [11] a number of examples demonstrate the interconnection between the probability theory and the theory of integral and differential equations. In particular in article [11] Cauchy process was considered. Later M. Kac method was used both for symmetric stable processes [30], [19] and non-symmetric stable processes [20]–[22]. In the present article with the help of M. Kac's idea [11] and the theory of integral equations with the difference kernels [22] we investigate a wide class of Levy processes. We note that within the last ten years the Levy processes have got a number of new important applications, particularly to financial problems. We consider separately the examples of Levy processes which are used in the financial mathematics.

Now we shall formulate the main results of the article.

1. The Levy process X_t defines a strongly continuous semigroup P_t (see [23]). The generator L of the semigroup P_t is a pseudo-differential operator. We show that for a broad class of the Levy processes the generator L can be represented in a convolution type form (Section 2):

$$Lf = \frac{d}{dx} S \frac{d}{dx} f, \tag{0.1}$$

where the operator S is defined by the relation

$$Sf = \frac{1}{2} Af + \int_{-\infty}^{\infty} k(y-x) f(y) dy. \tag{0.2}$$

Such a representation opens a possibility to use the theory of integral equations with difference kernels [22].

2. We introduce the important in our theory notion of the truncated generator L_Δ which coincides with L on the bounded system of segments Δ. we define the quasi-potential B by the relation $-L_\Delta B f = f$ (Sections 3 and 4). Under our assumptions the operator

$$Bf = \int_\Delta \Phi(x,y) f(y) dy \tag{0.3}$$

is compact. It means that the operator B has a discrete spectrum λ_j $(j = 1, 2, \dots)$, $\lambda_j \to 0$. Hence the corresponding truncated generator L_Δ has a discrete spectrum too. Using the results of the theory of the integral equations with the difference kernels we represent the quasi-potential B in the explicit form.

3. The probability $p(t, \Delta)$ of the Levy process remaining within the given domain Δ (ruin problem) is investigated in Section 5 of this paper. M. Kac [11] had obtained the first results of this type for symmetric stable processes. Later we transposed these results for the non-symmetric stable processes [20], [22]. In this paper we show that integro-differential equation of M. Kac type is true for all the Levy processes, which have a continuous density. M.Kac writes [11]: "We are led here to integro-differential equations which offer formidable analytic difficulties and which we were able to solve only in very few cases." Kac was able to overcome these difficulties only for Cauchy processes. H. Widom [30] solved these equations for symmetric stable processes. Both symmetric and non-symmetric stable processes were investigated in our works [19]–[22]. Now we develop these results and transfer them on the wide class of the Levy processes.

4. In Sections 6–8 we investigate the structure and the properties of the quasi-potential B. In particular we prove that the operator B is compact and the following important inequality:

$$\Phi(x,y) \geq 0 \tag{0.4}$$

is true. From inequality (0.4) and Krein-Rutman theorem [13] we deduce that the operator B has a positive eigenvalue λ_1 not less in modulus than every other eigenvalue of B.

5. Section 9 contains formulas and estimations for the probability $p(t,\Delta)$ that a sample of the process X_τ remains inside the given domain Δ when $0\leq\tau\leq t$. Under certain conditions we have obtained the asymptotic formula

$$p(t,\Delta)=e^{-t/\lambda_1}[c_1+o(1)],\quad t\to\infty. \tag{0.5}$$

6. In Sections 10–12 we separately consider the class of the stable processes which is a special case of the Levy processes. We use the notation $p(t,a)=p(t,\Delta)$ when $\Delta=[-a,a]$. We consider the important case when a depends on t and $a(t)\to\infty$ when $t\to\infty$. We compare the obtained results with well-known results (the iterated logarithm law, the results for the first hitting time, the results for the most visited sites problems). Further we introduce the notation $p(t,-b,a)=p(t,\Delta)$ when $\Delta=[-b,a]$. In case of the Wiener process we found the asymptotic behavior of $p(t,-b,a)$ when $b\to\infty$. It is easy to see that $p(t,-\infty,a)$ coincides with the formula for the first hitting time.

7. We analyze in detail a number of concrete examples of the Levy processes which are used in the financial mathematics (stable processes, the variance damped Levy processes, the variance gamma processes, the normal Gaussian process, the Meixner process, compound Poisson process).

1. Main notions

Let us consider the Levy processes X_t on R. If $P(X_0=0)=1$ then Levy-Khinchine formula gives (see[4], [23])

$$\mu(z,t)=E\{\exp[izX_t]\}=\exp[-t\lambda(z)],\quad t\geq 0, \tag{1.1}$$

where

$$\lambda(z)=\frac{1}{2}Az^2-i\gamma z-\int_{-\infty}^{\infty}(e^{ixz}-1-ixz1_{|x|<1})d\nu(x). \tag{1.2}$$

Here $A\geq 0$, $\gamma=\overline{\gamma}$, and $\nu(x)$ is a monotonically increasing function satisfying the conditions

$$\int_{-\infty}^{\infty}\frac{x^2}{1+x^2}d\nu(x)<\infty. \tag{1.3}$$

By $P_t(x_0,\Delta)$ we denote the probability $P(X_t\in\Delta)$ when $P(X_0=x_0)=1$ and $\Delta\in R$. The transition operator P_t is defined by the formula

$$P_tf(x)=\int_{-\infty}^{\infty}P_t(x,dy)f(y). \tag{1.4}$$

Let C_0 be the Banach space of continuous functions $f(x)$, satisfying the condition $\lim f(x)=0$, $|x|\to\infty$ with the norm $||f||=\sup_x|f(x)|$. We denote by C_0^n the set of $f(x)\in C_0$ such that $f^{(k)}(x)\in C_0$, $(1\leq k\leq n)$. It is known that [23]

$$P_tf\in C_0, \tag{1.5}$$

if $f(x)\in C_0^2$.

Now we formulate the following important result (see [4], [23]).

Theorem 1.1. *The family of the operators* $P_t \quad (t \geq 0)$ *defined by the Levy process* X_t *is a strongly continuous semigroup on* C_0 *with the norm* $||P_t|| = 1$. *Let* L *be its infinitesimal generator. Then*

$$Lf = \frac{1}{2}A\frac{d^2f}{dx^2} + \gamma\frac{df}{dx} + \int_{-\infty}^{\infty}(f(x+y) - f(x) - y\frac{df}{dx}1_{|y|<1})d\nu(y), \tag{1.6}$$

where $f \in C_0^2$.

2. Convolution type form of infinitesimal generator

1. In this section we prove that under some conditions the infinitesimal generator L can be represented in the special convolution type form

$$Lf = \frac{d}{dx}S\frac{d}{dx}f, \tag{2.1}$$

where the operator S is defined by the relation

$$Sf = \frac{1}{2}Af + \int_{-\infty}^{\infty}k(y-x)f(y)dy. \tag{2.2}$$

We assume that for arbitrary $M(0 < M < \infty)$ the inequality

$$\int_{-M}^{M}|k(t)|dt < \infty \tag{2.3}$$

is true. The representation of L in form (2.1) is convenient as the operator L is expressed with the help of the classic differential and convolution operators.

By C_s we denote the set of functions $f(x) \in C_0$ which have the following property

For every $f(x) \in C_s$ there exist such m and M $(0 < m < M < \infty)$ that

$$f(x) = 0, \quad x \notin [-M, -m] \bigcup [m, M]. \tag{2.4}$$

Lemma 2.1. *Let the following conditions be fulfilled.*

1. *The function* $\nu(x)$ *is monotonically increasing and*

$$\int_{-\infty}^{\infty}\frac{x^2}{1+x^2}d\nu(x) < \infty. \tag{2.5}$$

2. *For arbitrary* $M \quad (0 < M < \infty)$ *we have*

$$\int_{-M}^{M}|\nu(x)|dx < \infty, \quad \int_{-M}^{M}|x|d\nu(x) < \infty. \tag{2.6}$$

Then the expression

$$J = \int_{-\infty}^{\infty}[f(y+x) - f(x)]d\nu(y) \tag{2.7}$$

can be represented in the convolution type form

$$J = \frac{d}{dx}\int_{-\infty}^{\infty} f'(y)k(y-x)dy \tag{2.8}$$

where $f(x)\in C_0^2, \quad k(x) = \int_0^x \nu(y)dy.$

Proof. Let us introduce the following notations

$$J_1 = \frac{d}{dx}\int_{-\infty}^{x} f'(y)k(y-x)dy, \quad f(x)\in C_s, \tag{2.9}$$

$$J_2 = \frac{d}{dx}\int_{x}^{\infty} f'(y)k(y-x)dy \quad f(x)\in C_s. \tag{2.10}$$

Using (2.9) we have

$$J_1 = -\frac{d}{dx}\int_{-M}^{x} [f(y) - f(x) + f(x)]k'(y-x)dy. \tag{2.11}$$

From (2.9) and (2.11) we deduce the relation

$$J_1 = f(x)k'(-M-x) + \int_{-M}^{x} [f(y) - f(x)]k''(y-x)dy. \tag{2.12}$$

When $M\to\infty$ we obtain the equality

$$J_1 = \int_{-\infty}^{0} [f(y+x) - f(x)]k''(y)dy. \tag{2.13}$$

In the same way we deduce the relation

$$J_2 = \int_{0}^{\infty} [f(y+x) - f(x)]k''(y)dy. \tag{2.14}$$

Relation (2.8) follows directly from formulas (2.13), (2.14) and the equality $J = J_1 + J_2$. The lemma is proved. □

Lemma 2.2. *Let the following conditions be fulfilled.*

1. *The function* $\nu(x)$ *satisfies condition* 1 *of Lemma* 2.1.
2. *For arbitrary* $M \quad (0 < M < \infty)$ *we have*

$$\int_{-M}^{M} |k(x)|dx < \infty, \quad \int_{-M}^{M} |x\nu(x)|dx < \infty, \tag{2.15}$$

where

$$k'(x) = \nu(x), \quad x\neq 0. \tag{2.16}$$

Then the equality

$$J = \int_{-\infty}^{\infty} [f(y+x) - f(x) - y\frac{df(x)}{dx}1_{D(y)}]d\nu(y) + \Gamma f'(x), \tag{2.17}$$

is true, where $\Gamma = \overline{\Gamma}$ *and* $f(x)\in C_s$.

Proof. From (2.9) we obtain the relation

$$J_1 = f'(x)\gamma_1 - \int_{x-1}^{x} [f'(y) - f'(x)]k'(y-x)dy - \int_{-M}^{x-1} f'(y)k'(y-x)dy, \quad (2.18)$$

where $\gamma_1 = k(-1)$. We introduce the notations

$$P_1(x,y) = f(y) - f(x) - (y-x)f'(x), \quad P_2(x,y) = f(y) - f(x). \quad (2.19)$$

Integrating by parts (2.18) and passing to the limit when $M \to \infty$ we deduce that

$$J_1 = f'(x)\gamma_2 + \int_{x-1}^{x} P_1(x,y)k''(y-x)dy + \int_{-M}^{x-1} P_2(x,y)k''(y-x)dy, \quad (2.20)$$

where $\gamma_2 = k(-1) - k'(-1)$. It follows from (2.19) and (2.20) that

$$J_1 = \int_{-\infty}^{x} \left[f(y+x) - f(x) - y\frac{df(x)}{dx}1_{D(y)} \right] d\nu(y) + \gamma_2 f'(x). \quad (2.21)$$

In the same way it can be proved that

$$J_2 = \int_{x}^{\infty} [f(y+x) - f(x) - y\frac{df(x)}{dx}1_{D(y)}]d\nu(y) + \gamma_3 f'(x), \quad (2.22)$$

where $\gamma_3 = -k(1) + k'(1)$. The relation (2.17) follows directly from (2.21) and (2.22). Here $\Gamma = \gamma_2 + \gamma_3$. The lemma is proved. □

Remark 2.1. The operator $L_0 f = \frac{d}{dx} f$ can be represented in form (2.1), (2.2), where

$$S_0 f = \int_{-\infty}^{\infty} p_0(x-y)f(y)dy, \quad (2.23)$$

$$p_0(x) = \frac{1}{2}\,\mathrm{sign}(x). \quad (2.24)$$

From Lemmas 2.1, 2.2 and Remark 2.1 we deduce the following assertion.

Theorem 2.1. *Let the conditions of either Lemma* 2.1 *or Lemma* 2.2 *be fulfilled. Then the corresponding operator L has a convolution type form* (2.1), (2.2).

Proposition 2.1. *The generator L of the Levy process X_t admits the convolution type representation* (2.1), (2.2) *if there exist such $C > 0$ and $0 < \alpha < 2$, $\alpha \neq 1$ that*

$$\nu'(y) \leq C|y|^{-\alpha-1}. \quad (2.25)$$

Proof. The function $\nu(y)$ has the form

$$\nu(y) = \int_{-\infty}^{y} \nu'(t)dt1_{y<0} - \int_{y}^{\infty} \nu'(t)dt1_{y>0}. \quad (2.26)$$

First we shall consider the case, when $1 < \alpha < 2$, and introduce the function

$$k_0(y) = \int_{-\infty}^{y} (y-t)\nu'(t)dt1_{y<0} - \int_{y}^{\infty} (y-t)\nu'(t)dt1_{y>0}. \quad (2.27)$$

We obtain the relation

$$k(y) = k_0(y) + (\gamma - \Gamma)\mathrm{p}_0(y), \quad 1 < \alpha < 2, \quad (2.28)$$

where $p_0(y)$ and $k_0(y)$ are defined by (2.24) and (2.27) respectively. The constant Γ is defined by the relation:

$$\Gamma = k_0(-1) - k_0'(-1) - k_0(-1) + k_0'(1), \quad 1 < \alpha < 2. \tag{2.29}$$

It follows from (2.25)–(2.27) that the conditions of Lemma 2.2 are fulfilled. Hence the proposition is true when $1 < \alpha < 2$. Let us consider the case when $0 < \alpha < 1$. As in the previous case the function $\nu(x)$ is defined by relation (2.26). We introduce the functions

$$k_0(y) = \int_{-\infty}^{y} \nu'(t)dt\, y + \int_{y}^{0} \nu'(t)tdt, \quad y < 0, \tag{2.30}$$

$$k_0(y) = -\int_{y}^{\infty} \nu'(t)dt\, y - \int_{0}^{y} \nu'(t)tdt, \quad y > 0, \tag{2.31}$$

and

$$k(y) = k_0(y) + \gamma \mathrm{p}_0(y) \quad 0 < \alpha < 1. \tag{2.32}$$

In view of (2.25) and (2.30), (2.31) the conditions of Lemma 2.1 are fulfilled. Hence the proposition is proved. □

Corollary 2.1. *If condition* (2.25) *is fulfilled then*

$$k_0(y) \geq 0, \quad -\infty < y < \infty, \quad 1 < \alpha < 2, \tag{2.33}$$

$$k_0(y) \leq 0, \quad -\infty < y < \infty, \quad 0 < \alpha < 1. \tag{2.34}$$

Let us consider the important case when $\alpha = 1$.

Proposition 2.2. *The generator L of the Levy process X_t admits the convolution type representation* (2.1), (2.2) *if there exist such $C > 0$ and $m > 0$ that*

$$\nu'(y) \leq C|y|^{-2}e^{-m|y|}. \tag{2.35}$$

Proof. Using formulas (2.26)–(2.29) we see that the conditions of Lemma 2.2 are fulfilled. The proposition is proved. □

Example 2.1. The stable processes. For the stable processes we have $A = 0, \quad \gamma = \overline{\gamma}$ and

$$\nu'(y) = |y|^{-\alpha-1}(C_1 1_{y<0} + C_2 1_{y>0}), \tag{2.36}$$

where $C_1 > 0 \quad C_2 > 0$. Hence the function $\nu(y)$ has the form

$$\nu(y) = \frac{1}{\alpha}|y|^{-\alpha}(C_1 1_{y<0} - C_2 1_{y>0}). \tag{2.37}$$

Let us introduce the functions

$$k_0(y) = \frac{1}{\alpha(\alpha-1)}|y|^{1-\alpha}(C_1 1_{y<0} + C_2 1_{y>0}), \tag{2.38}$$

where $0 < \alpha < 2, \quad \alpha \neq 1$. When $\alpha = 1$ we have

$$k_0(y) = -\log|y|\,(C_1 1_{y<0} + C_2 1_{y>0}). \tag{2.39}$$

It means that the conditions of Theorem 2.1 are fulfilled. Hence the generator L for the stable processes admits the convolution type representation (2.1), (2.2).

Proposition 2.3. *The kernel $k(y)$ of the operator S in representation* (2.1) *for the stable processes has form* (2.28)*, when $1 \leq \alpha < 2$, and has form* (2.32) *when $0 < \alpha < 1$.*

Example 2.2. The variance damped Levy processes. For the variance damped Levy processes we have $A = 0, \quad \gamma = \overline{\gamma}$ and

$$\nu'(y) = C_1 e^{-\lambda_1 |y|} |y|^{-\alpha-1} 1_{y<0} + C_2 e^{-\lambda_2 |y|} y^{-\alpha-1} 1_{y>0}, \tag{2.40}$$

where $C_1 > 0 \quad C_2 > 0, \quad \lambda_1 > 0, \quad \lambda_2 > 0 \quad 0 < \alpha < 2$. It follows from (2.40) that the conditions of Proposition 2.1 are fulfilled when $\alpha \neq 1$. If $\alpha = 1$ the conditions of Proposition 2.2 are fulfilled. Hence the generator L for the variance damped Levy processes admits the convolution type representation (2.1), (2.2) and the kernel $k(y)$ is defined by formulas (2.27), (2.28), when $1 \leq \alpha < 2$, and by formula (2.32) when $0 < \alpha < 1$.

Example 2.3. The variance Gamma process. For the variance Gamma process we have $A = 0, \quad \gamma = \overline{\gamma}$ and

$$\nu'(y) = C_1 e^{-G|y|} |y|^{-1} 1_{y<0} + C_2 e^{-M|y|} y^{-1} 1_{y>0}, \tag{2.41}$$

where $C_1 > 0 \quad C_2 > 0, \quad G > 0, \quad M > 0$. It follows from (2.41) that the conditions of Proposition 2.2 are fulfilled and the generator L of variance Gamma process admits the convolution type representation (2.1), (2.2). The kernel $k(y)$ is defined by relations (2.30) and (2.31).

Example 2.4. The normal inverse Gaussian process. In the case of the normal inverse Gaussian process we have $A = 0, \quad \gamma = \overline{\gamma}$ and

$$\nu'(y) = C e^{\beta y} K_1(|y|) |y|^{-1}, \quad C > 0, \quad -1 \leq \beta \leq 1, \tag{2.42}$$

where $K_\lambda(x)$ denotes the modified Bessel function of the third kind with the index λ. Using equalities

$$|K_1(|x|)| \leq M e^{-|x|}/|x|, \quad M > 0, \quad 0 < x_0 \leq |x|, \tag{2.43}$$

$$|K_1(|x|)x| \leq M, \quad 0 \leq |x| \leq x_0 \tag{2.44}$$

we see that the conditions of Proposition 2.2 are fulfilled. Hence the corresponding generator L admits the convolution type representation (2.1), (2.2) and the kernel $k(y)$ is defined by relations (2.30) and (2.31).

Example 2.5. The Meixner process. For the Meixner process we have

$$\nu'(y) = C \frac{\exp \beta x}{x \sinh \pi x}, \tag{2.45}$$

where $C > 0, \quad -\pi < \beta < \pi$. The conditions of Proposition 2.2 are fulfilled. Hence the corresponding generator L admits the convolution type representation (2.1), (2.2) and the kernel $k(y)$ is defined by relations (2.30), (2.31).

Remark 2.2. Examples 2.1–2.5 are used in the finance problems [24].

Example 2.6. Compound Poisson process. We consider the case when $A = 0$, $\gamma = 0$ and

$$M = \int_{-\infty}^{\infty} \nu'(y)dy < \infty. \tag{2.46}$$

Using formulas (2.1) and (2.2) we deduce that the corresponding generator L has the following convolution form

$$Lf = -Mf(x) + \int_{-\infty}^{\infty} \nu'(y-x)f(y)dy. \tag{2.47}$$

3. Potential

The operator

$$Qf = \int_0^{\infty} (P_t f)dt \tag{3.1}$$

is called *potential* of the semigroup P_t. The generator L and the potential Q are (in general) unbounded operators. Therefore the operators L and Q are defined not in the whole space $L^2(-\infty, \infty)$ but only in the subsets D_L and D_Q respectively. We use the following property of the potential Q (see[23]).

Proposition 3.1. *If* $f = Q\,g$, $(g \in D_Q)$ *then* $f \in D_L$ *and*

$$-Lf = g. \tag{3.2}$$

Example 3.1. Compound Poisson process. Let the generator L has form (2.47) where

$$M = \int_{-\infty}^{\infty} \nu'(x)dx < \infty, \quad \int_{-\infty}^{\infty} [\nu'(x)]^2 dx < \infty. \tag{3.3}$$

We introduce the functions

$$K(u) = -\frac{1}{M\sqrt{2\pi}} \int_{-\infty}^{\infty} \nu'(x)e^{-iux}dx, \tag{3.4}$$

$$N(u) = \frac{K(u)}{1 - \sqrt{2\pi}K(u)}. \tag{3.5}$$

Let us note that

$$|K(u)| < \frac{1}{\sqrt{2\pi}}, \quad u \neq 0; \quad K(0) = -\frac{1}{\sqrt{2\pi}}. \tag{3.6}$$

It means that $N(u) \in L^2(-\infty, \infty)$. Hence the function

$$n(x) = -\frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} N(u)e^{-iux}du \tag{3.7}$$

belongs to $L^2(-\infty, \infty)$ as well. It follows from (2.47), (3.2)and (3.7) that the corresponding potential Q has the form (see [23], Ch. 11)

$$Qf = \frac{1}{M}[f(x) + \int_{-\infty}^{\infty} f(y)n(x-y)dy]. \tag{3.8}$$

Proposition 3.2. *Let conditions* (3.3) *be fulfilled. Then the operators* L *and* Q *are bounded in the space* $L^2(-\infty,\infty)$.

Now we shall give an example when the kernel $n(x)$ can be written in an explicit form.

Example 3.2. We consider the case when

$$\nu'(x) = e^{-|x|}, \quad -\infty < x < \infty. \tag{3.9}$$

In this case $M = 2$ and the operator L takes the form

$$Lf = -2f(x) + \int_{-\infty}^{\infty} f(y)e^{-|x-y|}dy. \tag{3.10}$$

Formulas (3.4)–(3.7) imply that

$$Qf = \frac{1}{2}f(x) - \frac{1}{4\sqrt{2}}\int_{-\infty}^{\infty} f(y)e^{-|x-y|\sqrt{2}}dy. \tag{3.11}$$

4. Truncated generators and quasi-potentials

Let us denote by Δ the set of segments $[a_k, b_k]$ such that

$$a_1 < b_1 < a_2 < b_2 < \cdots < a_n < b_n, \quad 1{\leq}k{\leq}n.$$

By C_Δ we denote the set of functions $g(x)$ on $L^2(\Delta)$ such that

$$g(a_k) = g(b_k) = g'(a_k) = g'(b_k) = 0, \quad 1{\leq}k{\leq}n, \quad g''(x){\in}L^p(\Delta), \quad p > 1. \tag{4.1}$$

We introduce the operator P_Δ by relation $P_\Delta f(x) = f(x)$ if $x{\in}\Delta$ and $P_\Delta f(x) = 0$ if $x{\notin}\Delta$.

Definition 4.1. The operator

$$L_\Delta = P_\Delta L P_\Delta \tag{4.2}$$

is called a *truncated generator*.

Definition 4.2. The operator B with the definition domain dense in $L^p(\Delta)$ is called *a quasi-potential* if the functions $f = Bg$ belong to definition domain of L_Δ and

$$-L_\Delta f = g. \tag{4.3}$$

It follows from (4.3) that

$$-P_\Delta L f = g, \quad (f = Bg). \tag{4.4}$$

Remark 4.1. In a number of cases (see the next section) we need relation (4.4). In these cases we can use the quasi-potential B, which is often simpler than the corresponding potential Q.

Remark 4.2. The operators of type (4.2) are investigated in the book ([22], Ch. 2).

From relation (4.3) we deduce that

$$Bg{\neq}0, \quad if \quad g{\neq}0. \tag{4.5}$$

Definition 4.3. We call the operator B a *regular* if the following conditions are fulfilled:

1. The operator B is compact and has the form

$$Bf = \int_{\Delta} \Phi(x,y)f(y)dy, \quad f(y)\in L^p(\Delta), \quad p\geq 1, \tag{4.6}$$

   where the function $\Phi(x,y)$ can have a discontinuity only when $x = y$.
2. There exists a function $\phi(x)$ such that

$$|\Phi(x,y)|\leq\phi(x-y), \tag{4.7}$$

$$\int_{-R}^{R} \phi(x)dx < \infty \quad \text{if} \quad 0 < R < \infty. \tag{4.8}$$

3. $$\Phi(x,y) \geq 0, \quad x,y\in\Delta, \tag{4.9}$$
$$\Phi(a_k,y) = \Phi(b_k,y) = 0, \quad 1\leq k\leq n. \tag{4.10}$$

4. Relation (4.5) is valid.

Remark 4.3. In view of condition (4.7) the regular operator B is bounded in the spaces $L^p(\Delta)$, $1\leq p\leq\infty$ (see [22], p. 24).

Remark 4.4. If the quasi-potential B is regular, then the corresponding truncated generator L_Δ has a discrete spectrum.

Further we prove that for a broad class of Levy processes the corresponding quasi-potentials B are regular.

Example 4.1. We consider the case when

$$\phi(x) = M|x|^{-\varkappa}, \quad 0 < \varkappa < 1. \tag{4.11}$$

Proposition 4.1. *Let condition* (4.11) *be true and let the corresponding regular operator B have an eigenfunction $f(x)$ with an eigenvalue $\lambda\neq 0$. Then the function $f(x)$ is continuous.*

Proof. According to Definition 4.3 there exists an integer $N(\varkappa)$ such that the kernel $\Phi_N(x,t)$ of the operator

$$B^N f = \int_{\Delta} \Phi_N(x,y)f(y)dy, \quad f(y)\in L^p(\Delta) \tag{4.12}$$

is continuous. Hence the function $f(x)$ is continuous. The proposition is proved. □

5. The probability of the Levy process remaining within the given domain

In many theoretical and applied problems it is important to estimate the quantity

$$p(t,\Delta) = P(X_\tau \in \Delta; 0 \le \tau \le t), \tag{5.1}$$

i.e., the probability that a sample of the process X_τ remains inside Δ for $0 \le \tau \le t$ (*ruin problem*).

The integro-differential equations corresponding to the stable processes were derived by Kac [11] (symmetric case) and in our works (non-symmetric case, see [20]–[22]). Now we get rid of the requirement for the process to be stable and consider the Levy process X_t with the continuous density $\rho(x,t)$. In view of (1.1) we have

$$\rho(x,t) = \frac{1}{2\pi}\int_{-\infty}^{\infty} e^{-ixz}\mu(z,t)dz, \quad t > 0. \tag{5.2}$$

We introduce the sequence of functions

$$Q_{n+1}(x,t) = \int_0^t \int_{-\infty}^{\infty} Q_0(x-\xi, t-\tau)V(\xi)Q_n(\xi,\tau)d\xi d\tau, \tag{5.3}$$

where the function $V(x)$ is defined by relations $V(x) = 1$ when $x \notin \Delta$ and $V(x) = 0$ when $x \in \Delta$. We use the notation

$$Q_0(x,t) = \rho(x,t). \tag{5.4}$$

For Levy processes the following relation

$$Q_0(x,t) = \int_{-\infty}^{\infty} Q_0(x-\xi, t-\tau)Q_0(\xi,\tau)d\xi \tag{5.5}$$

is true. Using (5.3) and (5.5) we have

$$0 \le Q_n(x,t) \le t^n Q_0(x,t)/n!. \tag{5.6}$$

Hence the series

$$Q(x,t,u) = \sum_{n-0}^{\infty}(-1)^n u^n Q_n(x,t) \tag{5.7}$$

converges. The probabilistic meaning of $Q(x,t,u)$ is defined by the relation (see [12], Ch. 4):

$$E\{\exp[-u\int_0^t V(X_\tau)d\tau], c_1 < X_t < c_2\} = \int_{c_1}^{c_2} Q(x,t,u)dx. \tag{5.8}$$

The inequality $V(x) \ge 0$ and relation (5.8) imply that the function $Q(x,t,u)$ monotonically decreases with respect to the variable "u" and the formulas

$$0 \le Q(x,t,u) \le Q(x,t,0) = Q_0(x,t) = \rho(x,t) \tag{5.9}$$

are true. In view of (5.2) and (5.9) the Laplace transform

$$\psi(x,s,u) = \int_0^{\infty} e^{-st}Q(x,t,u)dt, \quad s > 0 \tag{5.10}$$

has meaning. According to (5.3) the function $Q(x,t,u)$ is the solution of the equation

$$Q(x,t,u)+u\int_0^t\int_{-\infty}^{\infty}\rho(x-\xi,t-\tau)V(\xi)Q(\xi,\tau,u)d\xi d\tau=\rho(x,t). \tag{5.11}$$

Taking from both parts of (5.11) the Laplace transform and bearing in mind (5.10) we obtain

$$\psi(x,s,u)+u\int_{-\infty}^{\infty}V(\xi)R(x-\xi,s)\psi(\xi,s,u)d\xi=R(x,s), \tag{5.12}$$

where

$$R(x,s)=\int_0^{\infty}e^{-st}\rho(x,t)dt. \tag{5.13}$$

Multiplying both parts of relation (5.12) by $\exp(ixp)$ and integrating them with respect to $x\quad(-\infty<x<\infty)$ we have

$$\int_{-\infty}^{\infty}\psi(x,s,u)e^{ixp}[s+\lambda(p)+uV(x)]dx=1. \tag{5.14}$$

Here we use relations (1.1), (5.2) and (5.13). Now we introduce the function

$$h(p)=\frac{1}{2\pi}\int_{\Delta}e^{-ixp}f(x)dx, \tag{5.15}$$

where the function $f(x)$ belongs to C_{Δ}. Multiplying both parts of (5.14) by $h(p)$ and integrating them with respect to $p\quad(-\infty<p<\infty)$ we deduce the equality

$$\int_{-\infty}^{\infty}\int_{-\infty}^{\infty}\psi(x,s,u)e^{ixp}[s+\lambda(p)]h(p)dxdp=f(0). \tag{5.16}$$

We have used the relations

$$V(x)f(x)=0,\quad -\infty<x<\infty, \tag{5.17}$$

$$\frac{1}{2\pi}\lim\int_{-N}^{N}\int_{\Delta}e^{-ixp}f(x)dxdp=f(0),\quad N\to\infty. \tag{5.18}$$

Since the function $Q(x,t,u)$ monotonically decreases with respect to "u" this is also true for the function $\psi(x,s,u)$ according to (5.10). Hence there exists the limit

$$\psi(x,s)=\lim\psi(x,s,u),\quad u\to\infty, \tag{5.19}$$

where

$$\psi(x,s)=0,\quad x\notin\Delta. \tag{5.20}$$

The probabilistic meaning of $\psi(x,s)$ follows from the equality

$$\int_0^{\infty}e^{-st}p(t,\Delta)dt=\int_{\Delta}\psi(x,s)dx. \tag{5.21}$$

Using the properties of the Fourier transform and conditions (5.19), (5.20) we deduce from (5.16) the following assertion.

Theorem 5.1. *Let the considered Levy process have a continuous density. Then the relation*

$$((sI - L_\Delta)f, \psi(x,s))_\Delta = f(0) \tag{5.22}$$

is true.

Remark 5.1. For the symmetric stable processes relation (5.22) was deduced by M. Kac [11] and for the non-symmetric stable processes it was deduced in our works [20]–[22].

Remark 5.2. As it is known the stable processes, the variance damped Levy processes, the variance gamma processes, the normal inverse Gaussian process, the Meixner process have continuous densities (see([24], [31]).

Remark 5.3. We have obtained the formula (5.21) for Laplace transform of $p(t,\Delta)$ in terms of $\psi(x,s)$. The double Laplace transform of $p(t,\Delta)$ was obtained by G. Baxter and M.D. Donsker [3] for the case when $\Delta = (-\infty, a]$.

We express the important function $\psi(x,s)$ with the help of the quasi-potential B.

Theorem 5.2. *Let the considered Levy process have the continuous density and let the quasi-potential B be regular. Then in the space $L^p(\Delta)$ $(p>1)$ there is one and only one function*

$$\psi(x,s) = (I + sB^\star)^{-1}\Phi(0,x), \quad 0 \leq s < s_0, \tag{5.23}$$

which satisfies relation (5.22).

Proof. In view of (4.4) we have

$$-BL_\Delta f = f, \quad f \in C_\Delta. \tag{5.24}$$

Relations (5.23) and (5.24) imply that

$$((sI - L_\Delta)f, \psi(x,s))_\Delta = -((I+sB)L_\Delta f, \psi)_\Delta = -(L_\Delta f, \Phi(0,x))_\Delta. \tag{5.25}$$

Since $\Phi(0,x) = B^\star\delta(x)$, ($\delta(x)$ is the Dirac function) then according to (5.23) and (5.25) relation (5.22) is true.

Let us suppose that in $L(\Delta)$ there is another function $\psi_1(x,s)$ satisfying (5.22). Then the equality

$$((sI - L_\Delta)f, \phi(x,s))_\Delta = 0, \quad \phi = \psi - \psi_1 \tag{5.26}$$

is valid. We write relation (5.26) in the form

$$(L_\Delta f, (I + sB^\star)\phi)_\Delta = 0. \tag{5.27}$$

Due to (4.4) the range of L_Δ is dense in $L^p(\Delta)$. Hence in view of (5.27) we have $\phi = 0$. The theorem is proved. □

The analytical apparatus for the construction and investigation of the function $\psi(x,s)$ is based on relation (5.22) and properties of the quasi-potential B. In the following three sections we shall investigate the properties of the operator B.

6. Non-negativity of the kernel $\Phi(x, y)$

In this section we deduce the following important property of the kernel $\Phi(x, y)$.

Proposition 6.1. *Let the density $\rho(x, t)$ of Levy process X_t be continuous ($t > 0$) and let the corresponding quasi-potential B satisfy conditions* (4.6)–(4.8) *of Definition* 4.3. *Then the kernel $\Phi(x, y)$ is non-negative, i.e.,*

$$\Phi(x, y) \geq 0. \tag{6.1}$$

Proof. In view of (5.9) and (5.10) we have $\psi(x, s, u) \geq 0$. Relation (5.19) implies that $\psi(x, s) \geq 0$. Now it follows from (5.23) that

$$\Phi(0, x) = \psi(x, 0) \geq 0. \tag{6.2}$$

Let us consider the domains Δ_1 and Δ_2 which are connected by relation $\Delta_2 = \Delta_1 + \delta$. We denote the corresponding truncated generators by L_{Δ_1} and L_{Δ_2}, the corresponding quasi-potentials by B_1 and B_2 and the corresponding kernels by $\Phi_1(x, y)$ and $\Phi_2(x, y)$. We introduce the unitary operator

$$Uf = f(x - \delta), \tag{6.3}$$

which maps the space $L^2(\Delta_2)$ onto $L^2(\Delta_1)$. At the beginning we suppose that the conditions of Theorem 2.1 are fulfilled. Using formulas (2.1) and (2.2) we deduce that

$$L_{\Delta_2} = U^{-1} L_{\Delta_1} U. \tag{6.4}$$

Hence the equality

$$B_2 = U^{-1} B_1 U \tag{6.5}$$

is valid. The last equality can be written in the terms of the kernels

$$\Phi_2(x, y) = \Phi_1(x + \delta, y + \delta). \tag{6.6}$$

According to (6.2) and (6.6) we have

$$\Phi_1(\delta, y + \delta) \geq 0. \tag{6.7}$$

As δ is an arbitrary real number, relation (6.1) follows directly from (6.7). We remark that an arbitrary generator L can be approximated by the operators of form (2.1) (see[23], Ch. 2). Hence the proposition is proved. □

In view of (4.1), (4.5) and relation $Bf \in C_\Delta$ the following assertion is true.

Proposition 6.2. *Let the quasi-potential B satisfy the conditions of Proposition* 6.1. *Then the equalities*

$$\Phi(a_k, y) = \Phi(b_k, y) = 0 \quad 1 \leq k \leq n \tag{6.8}$$

are valid.

7. Sectorial operators

1. We introduce the following notions.

Definition 7.1. The bounded operator B in the space $L^2(\Delta)$ is called *sectorial* if

$$(Bf,f)\neq 0, \quad f\neq 0 \tag{7.1}$$

and

$$-\frac{\pi}{2}\beta\leq \arg(Bf,f)\leq \frac{\pi}{2}\beta, \quad 0<\beta\leq 1. \tag{7.2}$$

It is easy to see that the following assertions are true.

Proposition 7.1. *Let the operator B be sectorial. Then the operator $(I+sB)^{-1}$ is bounded when $s\geq 0$.*

Proposition 7.2. *Let the conditions of Theorem* 5.2 *be fulfilled. If the operator B is sectorial, then formula* (5.23) *is valid for all $s\geq 0$.*

In the present section we deduce the conditions under which the quasi-potential B is sectorial. Let us consider the case when

$$\int_x^\infty y d\nu(y)<\infty, \quad (x>0), \tag{7.3}$$

$$\int_{-\infty}^x |y| d\nu(y)<\infty, \quad (x<0). \tag{7.4}$$

The corresponding kernel $k(x)$ of the operator S (see (2.2)) has the form

$$k(x)=\int_x^\infty (y-x)d\nu(y)d<\infty, \quad (x>0), \tag{7.5}$$

$$k(x)=\int_{-\infty}^x (x-y)d\nu(y)<\infty, \quad (x<0). \tag{7.6}$$

We obtain the following statement.

Proposition 7.3. *Let conditions* (7.3) *and* (7.4) *be fulfilled. Then the kernel $k(x)$ is monotone on the half-axis $(-\infty,0)$ and on the half-axis $(0,\infty)$.*

We shall use the following Pringsheim's result.

Theorem 7.1. (see [25], Ch. 1.) *Let $f(t)$ be non-increasing function over $(0,\infty)$ and integrable on any finite interval $(0,\ell)$. If $f(t)\to 0$ when $t\to\infty$, then for any positive x we have*

$$\frac{1}{2}[f(x+0)+f(x-0)]=\frac{2}{\pi}\int_{+0}^\infty \cos xu\left[\int_0^\infty f(t)\cos tudt\right]du, \tag{7.7}$$

$$\frac{1}{2}[f(x+0)+f(x-0)]=\frac{2}{\pi}\int_0^\infty \sin xu\left[\int_0^\infty f(t)\sin tudt\right]du. \tag{7.8}$$

It follows from (7.3)–(7.6) that

$$k(x)\to 0 \quad and \quad k'(x)\to 0, \quad when \quad x\to\pm\infty. \tag{7.9}$$

We suppose in addition that

$$xk(x)\to 0 \quad and \quad x^2k'(x)\to 0, \quad when \quad x\to\pm 0. \tag{7.10}$$

Using the integration by parts we deduce the assertion.

Proposition 7.4. *Let conditions* (7.3), (7.4) *and* (7.9), (7.10) *be fulfilled. Then the relation*

$$\int_{-\infty}^{\infty} k(t)\cos xtdt = \int_{-\infty}^{\infty} \frac{1-\cos xt}{x^2} d\nu(t) \tag{7.11}$$

is true.

Relation (7.11) implies that

$$\int_{-\infty}^{\infty} k(t)\cos xtdt > 0. \tag{7.12}$$

It follows from Proposition 7.3, Theorem 7.1 and relations (7.9), (7.10) that the kernel $k(x)$ of the operator S admits the representation

$$k(x) = \int_{-\infty}^{\infty} m(t)e^{ixt}dt. \tag{7.13}$$

In view of (7.12) we have

$$\mathrm{Re}[m(u)] > 0. \tag{7.14}$$

Due to (7.13) and (7.14) the relation

$$(Sf, f) = \int_{-\infty}^{\infty} m(u)|\int_{\Delta} f(t)e^{iut}dt|^2 du \tag{7.15}$$

is valid. Hence we have

$$-\frac{\pi}{2}\le \arg(Sf, f)\le\frac{\pi}{2}, \quad f(t)\in L^2(\Delta). \tag{7.16}$$

Proposition 7.5. *Let conditions* (7.3), (7.4) *and* (7.9), (7.10) *be fulfilled. Then the corresponding operator B is sectorial.*

Proof. Let the function $g(x)$ satisfies conditions (4.1). Then the relation

$$(-Lg, g) = (Sg', g') \tag{7.17}$$

holds. Equalities (4.3) and (7.17) imply that

$$(f, Bf) = (Sg', g'), \quad g = Bf. \tag{7.18}$$

Inequality (7.1) follows from relations (7.14) and (7.18). Relations (7.16) and (7.18) imply (7.2) with $\beta = 1$. The proposition is proved. □

Remark 7.1. The variance damped processes (Example 2.2.), the normal inverse Gaussian process (Example 2.4.), the Meixner process (Example 2.5.) satisfy the conditions of Proposition 7.5. Hence the corresponding operators B are sectorial.

2. Now we introduce the notion of the strongly sectorial operators.

Definition 7.2. The sectorial operator B is called *a strongly sectorial* if for some $\beta < 1$ relation (7.2) is valid.

Proposition 7.6. *Let the following conditions be fulfilled.*

1. *Relations* (7.3), (7.4) *and* (7.9), (7.10) *are valid.*
2. *For some* $m > 0$ *the inequality*

$$\frac{m}{|x|^2} \leq \nu'(x), \quad |x| \leq 1 \tag{7.19}$$

is true.

3. $$\int_{-\infty}^{\infty} k(t)dt < \infty. \tag{7.20}$$

Then the corresponding operator B *is strongly sectorial.*

Proof. As it is known (see [25], Ch. 1) the inequality

$$\left| \int_{-\infty}^{\infty} k(t) \sin xt dt \right| \leq \frac{M}{|x|}, \quad M > 0, \quad |t| \geq 1 \tag{7.21}$$

is valid. From formulas (7.11) and (7.19) we conclude that

$$\int_{-\infty}^{\infty} k(t) \cos xt dt \geq \int_{-1/x}^{1/x} \nu'(t) \frac{1 - \cos xt}{x^2} dt \geq \frac{N}{|x|}, \quad N > 0, \quad |x| \geq 1. \tag{7.22}$$

It follows from (7.21) and (7.22) that

$$-\frac{\pi}{2}\beta \leq \arg(Sf, f) \leq \frac{\pi}{2}\beta, \quad 0 < \beta < 1. \tag{7.23}$$

Hence according to (7.18) the corresponding operator B is strongly sectorial. The proposition is proved. □

Remark 7.2. The variance damped processes (Example 2.2, $\alpha \geq 1$), the normal inverse Gaussian process (Example 2.4.), the Meixner process (Example 2.5.) satisfy the conditions of Proposition 7.6. Hence the corresponding operators B are strongly sectorial.

Proposition 7.7. *Let conditions* (7.3), (7.4) *and* (7.9), (7.10) *be fulfilled. If the operator* S *has the form*

$$Sf = Af + \int_{\Delta} k(x - t) f(t) dt, \quad A > 0. \tag{7.24}$$

Then the corresponding operator B *is strongly sectorial.*

Proof. It is easy to see that for some $\beta < 1$ relation (7.23) is true. According to relation (7.18) the corresponding operator B is strongly sectorial. □

8. Quasi-potential B, structure and properties

Let us begin with the symmetric segment $\Delta = [-c, c]$.

Theorem 8.1. (*see* [22], *p.* 140) *Let the following conditions be fulfilled.*

1. *There exist the functions* $N_k(x) \in L^p(-c, c), \quad p > 1$ *which satisfy the equations*
$$SN_k = x^{k-1}, \quad k = 1, 2. \tag{8.1}$$
2. $$r = \int_{-c}^{c} N_1(x)dx \neq 0. \tag{8.2}$$

Then the corresponding operator B *has the form*

$$Bf = \int_{-c}^{c} \Phi(x, y, c) f(y)dy \tag{8.3}$$

where

$$\Phi(x, y, c) = \frac{1}{2} \int_{x+y}^{2c-|x-y|} q[(s + x - y)/2, (s - x + y)/2]ds, \tag{8.4}$$

$$q(x, y) = [N_1(-y)N_2(x) - N_2(-y)N_1(x)]/r. \tag{8.5}$$

It follows from (8.4) and (8.5) that

$$\Phi(\pm c, y) = \Phi(x, \pm c) = 0. \tag{8.6}$$

Here we use the following relation

$$\begin{aligned} &q[(s + x - y)/2, (s - x + y)/2] \\ &= [N_1((x - y - s)/2)N_2((s + x - y)/2) - N_2((x - y - s)/2)N_2((s + x - y)/2)]/r. \end{aligned} \tag{8.7}$$

Thus

$$q[(s + x - y)/2, (s - x + y)/2] = -q[(-s + x - y)/2, (-s - x + y)/2]. \tag{8.8}$$

From formulas (8.4) and (8.5) we deduce the following statement.

Proposition 8.1. *Let the conditions of Theorem* 8.1 *be fulfilled. There exists a function* $\phi(x)$ *such that*

$$|\Phi(x, y, c)| \leq \phi(x - y), \tag{8.9}$$

$$\int_{-R}^{R} \phi(x)dx < \infty \quad \text{if} \quad 0 < R < \infty. \tag{8.10}$$

Proof. Relation (8.4) can be written in the form

$$\Phi(x, y, c) = \int_{x}^{c+(x-y-|x-y|)/2} q(t, t - x + y)dt. \tag{8.11}$$

By relations

$$N_k(x) = 0, \quad x \notin [-c, c], \quad k = 1, 2 \tag{8.12}$$

we extend the functions $N_k(x)$ from the segment $[-c,c]$ to the segment $[-2c,2c]$. It follows from (8.11) and (8.12) that inequality (8.9) is valid if

$$\phi(x)=\int_{-c}^{c}[|N_1(t)N_2(t-x)|+|N_2(t)N_1(t-x)|]dt/|r|. \tag{8.13}$$

Equality (8.13) imply that $\phi(x)\in L^p[-2c,2c]$. The proposition is proved. □

It follows from Proposition 8.1 that the operator B is bounded in all the spaces $L^p(-c,c)$, $p\geq 1$. We shall prove that the operator B is compact.

Proposition 8.2. *Let the conditions of Theorem* 8.1 *be fulfilled. Then the operator* B *is compact in all the spaces* $L^p(-c,c)$, $p\geq 1$.

Proof. Let us consider the operator $B^\star$ in the space $L^q(-c,c)$, $1/p+1/q=1$. Using relation (8.3) we have

$$B^\star f_n=\int_{-c}^{c}\Phi(y,x,c)f_n(y)dy \tag{8.14}$$

where the functions $f_n(x)\to 0$ in the weak sense. Relation (8.14) can be represented in the following form

$$B^\star f_n=\int_{-c}^{c}f_n(y)\int_{y}^{c+(y-x-|x-y|)/2}q(t,t-y+x)dtdy. \tag{8.15}$$

By interchanging the order of the integration in (8.15) we see that $||B^\star f_n||\to 0$, i.e., the operator $B^\star$ is compact. Hence the operator B is compact too. The proposition is proved. □

Using formulas (8.5) and (8.11) we obtain the assertion.

Proposition 8.3. *Let the conditions of Theorem* 8.1 *be fulfilled. If the functions* $N_1(x)$ *and* $N_2(x)$ *can have a discontinuity only when* $x=\pm c$ *then the function* $\Phi(x,y,c)$ *can have a discontinuity only when* $x=y$.

Corollary 8.1. *Let the conditions of Proposition* 8.3 *be fulfilled. Then the eigenvectors of the corresponding operator* B *are continuous.*

Remark 8.1. In view of (6.4) and (6.5) Proposition 8.1 is valid not only in the case of the symmetric segment $[-c,c]$ but in the general case $[-a,b]$ too.

9. Long time behavior

1. In order to investigate the asymptotic behavior of $p(t,\Delta)$ when $t\to\infty$, we use the non-negativity of the kernel $\Phi(x,y)$. We apply the following Krein-Rutman theorem (see [13], Section 6).

Theorem 9.1. *If a linear compact operator* B *leaving invariant a cone* K, *has a point of the spectrum different from zero, then it has a positive eigenvalue* λ_1 *not*

less in modulus than any other eigenvalues λ_k, $(k > 1)$. *To this eigenvalue* λ_1 *corresponds at least one eigenvector* $g_1 \in K, (Bg_1 = \lambda_1 g_1)$ *of the operator* B *and at least one eigenvector* $h_1 \in K^\star, (B^\star h_1 = \lambda_1 h_1)$ *of the operator* $B^\star$.

We remark that in our case the cone K consists of non-negative functions $f(x) \in L^p(\Delta)$. Hence we have

$$g_1(x) \geq 0, \quad h_1(x) \geq 0. \tag{9.1}$$

We introduce the following normalizing condition

$$(g_1, h_1) = \int_\Delta g_1(x) h_1(x) dx = 1. \tag{9.2}$$

Let the interval Δ_1 and the point x_0 be such that

$$x_0 \in \Delta_1 \in \Delta. \tag{9.3}$$

Together with quantity $p(t, \Delta)$ we consider the expression

$$p(x_0, \Delta_1, t, \Delta) = P((X_\tau \in \Delta) \bigcap (X_t \in \Delta_1), 0 \leq \tau \leq t), \tag{9.4}$$

where $x_0 = X_0$. If the relations $x_0 = 0, \quad \Delta_1 = \Delta$ are true, then $p(x_0, \Delta_1, t, \Delta) = p(t, \Delta)$. In this section we investigate the asymptotic behavior of $p(x_0, \Delta_1, t, \Delta)$ and $p(t, \Delta)$ when $t \to \infty$.

Theorem 9.2. *Let the considered Levy process have the continuous density and let the corresponding quasi-potential* B *be regular and strongly sectorial. And let the operator* B *have a point of the spectrum different from zero. Then the asymptotic equality*

$$p(t, \Delta) = e^{-t/\lambda_1}[q(t) + o(1)], \quad t \to +\infty \tag{9.5}$$

is true. The function $q(t)$ *has the form*

$$q(t) = c_1 + \sum_{k=2}^{m} c_k e^{it\nu_k} \geq 0, \tag{9.6}$$

where ν_k *are real.*

Proof. The spectrum $(\lambda_k,\ k > 1)$ of the operator B is situated in the sector

$$-\frac{\pi}{2}\beta \leq \arg z \leq \frac{\pi}{2}\beta, \quad 0 \leq \beta < 1, \quad |z| \leq \lambda_1. \tag{9.7}$$

We introduce the domain D_ϵ:

$$-\frac{\pi}{2}(\beta + \epsilon) \leq \arg z \leq \frac{\pi}{2}(\beta + \epsilon), \quad |z - (1/2)\lambda_1| < (1/2)(\lambda_1 - r), \tag{9.8}$$

where $0 < \epsilon < 1 - \beta), \quad r < \lambda_1$. If z belongs to the domain D_ϵ then the relation

$$\mathrm{Re}(1/z) > 1/\lambda_1 \tag{9.9}$$

holds. As the operator B is compact, only a finite number of eigenvalues $\lambda_k, \quad 1 < k \leq m$ of this operator does not belong to the domain D_ϵ. We denote the boundary

of domain D_ϵ by Γ_ϵ. Without loss of generality we may assume that the points of spectrum $\lambda_k \neq 0$ do not belong to Γ_ϵ. Taking into account the equality

$$(\Phi(0,x), g_1(x)) = \lambda_1 g_1(0), \tag{9.10}$$

we deduce from formulas (5.21) and (5.23) the relation

$$p(t,\Delta) = \sum_{k=1}^{m} \sum_{j=0}^{n_k} e^{-t/\lambda_k} t^j c_{k,j} + J, \tag{9.11}$$

where n_k is the index of the eigenvalue λ_k,

$$J = -\frac{1}{2i\pi} \int_\Gamma \frac{1}{z} e^{-t/z} ((B^\star - zI)^{-1} \Phi(0,x), 1) dz. \tag{9.12}$$

We note that

$$n_1 = 1. \tag{9.13}$$

Indeed, if $n_1 > 1$ then there exists such a function f_1 that

$$Bf_1 = \lambda_1 f_1 + g_1. \tag{9.14}$$

In this case the relations

$$(Bf_1, h_1) = \lambda_1 (f_1, h_1) + (g_1, h_1) = \lambda_1 (f_1, h_1) \tag{9.15}$$

are true. Hence $(g_1, h_1) = 0$. The last relation contradicts condition (9.2). It proves equality (9.13).

Relation (8.9) implies that

$$\Phi(0,x) \in L^p(\Delta). \tag{9.16}$$

We denote by $W(B)$ the numerical range of B. The closure of the convex hull of $W(B)$ is situated in the sector (9.7). Hence the estimation

$$||(B^\star - zI)^{-1}||_p \leq M/|z|, \quad z \in \Gamma_\epsilon \tag{9.17}$$

is true (see [26] for the Hilbert case $p = 2$ and [16], [28] for the Banach space $p \geq 1$). By $||B||_p$ we denote the norm of the operator B in the space $L^p(\Delta)$.

It follows from estimation (9.17) that the integral J exists.

Among the numbers λ_k we choose the ones for which $\mathrm{Re}(1/\lambda_k)$, $(1 \leq k \leq m)$ has the smallest value δ. Among the obtained numbers we choose μ_k, $(1 \leq k \leq \ell)$ the indexes n_k of which have the largest value n. We deduce from (9.10)–(9.12) that

$$p(t,\Delta) = e^{-t\delta} t^n [\sum_{k=1}^{\ell} e^{-t/\mu_k} c_k + o(1)], \quad t \to \infty. \tag{9.18}$$

We note that the function

$$Q(t) = \sum_{k=1}^{\ell} e^{it\mathrm{Im}(\mu_k^{-1})} c_k \tag{9.19}$$

is almost periodic (see [14]). Hence in view of (9.18) and the inequality $p(t,\Delta) > 0$, $t \geq 0$ the following relation

$$Q(t) \geq 0, \quad -\infty < t < \infty \tag{9.20}$$

is valid.

First we assume that at least one of the inequalities

$$\delta < \lambda_1^{-1}, \quad n > 1 \tag{9.21}$$

is true. Using (9.21) and the inequality

$$\lambda_1 \geq \lambda_k, \quad k = 2, 3, \ldots \tag{9.22}$$

we have

$$\mathrm{Im}\mu_j^{-1} \neq 0, \quad 1 \leq j \leq \ell. \tag{9.23}$$

It follows from (9.19) that

$$c_j = \lim \frac{1}{2T} \int_{-T}^{T} Q(t) e^{-it(\mathrm{Im}\mu_j^{-1})} dt, \quad T \to \infty. \tag{9.24}$$

In view of (9.20) the relations

$$|c_j| \leq \lim \frac{1}{2T} \int_{-T}^{T} Q(t)dt = 0, \quad T \to \infty, \tag{9.25}$$

are valid, i.e., $c_j = 0, \quad 1 \leq j \leq \ell$. This means that relations (9.21) are not true. Hence the equalities

$$\delta = \lambda_1^{-1}, \quad n = 1 \tag{9.26}$$

are true. From (9.18) and (9.19) we get the asymptotic equality

$$p(t,\Delta) = e^{-t/\lambda_1}[q(t) + o(1)] \quad t \to \infty, \tag{9.27}$$

where the function $q(t)$ is defined by relation (9.6) and

$$c_k = \overline{g_k(0)} \int_{\Delta} h_k(x)dx, \quad \nu_k = \mathrm{Im}(\mu^{-1}). \tag{9.28}$$

Here $g_k(x)$ are the eigenfunctions of the operator B corresponding to the eigenvalues λ_k, and $h_k(x)$ are the eigenfunctions of the operator $B^\star$ corresponding to the eigenvalues $\overline{\lambda_k}$. The following conditions are fulfilled

$$(g_k, h_k) = \int_{\Delta} \overline{g_k(x)} h_k(x)dx = 1, \tag{9.29}$$

$$(g_k, h_\ell) = \int_{\Delta} \overline{g_k(x)} h_\ell(x)dx = 0, \quad k \neq \ell. \tag{9.30}$$

Using the almost periodicity of the function $q(t)$ we deduce from (9.27) the inequality

$$q(t) \geq 0. \tag{9.31}$$

The theorem is proved. □

Corollary 9.1. *Let the conditions of Theorem* 9.2 *be fulfilled. Then all the eigenvalues* λ_j *of* B *belong to the disk*

$$|z-(1/2)\lambda_1|\leq(1/2)\lambda_1. \tag{9.32}$$

All the eigenvalues λ_j *of* B *which belong to the boundary of disc* (9.32) *have the indexes* $n_j=1$.

Remark 9.1. The exponential decay of the transition probability $P_t(x,B)$ was proved by P. Tuominen and R.L. Tweedie [29]. Theorem 9.2. gives the exponential decay of $p(t,\Delta)$. These two results are independent.

Using formula (9.11) we obtain the following assertion.

Corollary 9.2. *Let the considered Levy process have the continuous density and let the corresponding quasi-potential* B *be regular and strongly sectorial. And let the operator* B *have no points of the spectrum different from zero. Then the equality*

$$\lim[p(t,\Delta)e^{t/\lambda}]=0, \quad t\rightarrow+\infty \tag{9.33}$$

is true for any $\lambda>0$.

2. Now we find the conditions under which the operator B has a point of the spectrum different from zero.

We represent the corresponding operator B in the form $B=B_1+iB_2$ where the operators B_1 and B_2 are self-adjoint. We assume that $B_1\in\Sigma_p$, i.e.,

$$\sum_1^\infty |s_n|^{-p}<\infty, \tag{9.34}$$

where s_n are eigenvalues of the operator B_1 and $p>1$. As operator B is sectorial, then

$$B_1\geq 0. \tag{9.35}$$

Theorem 9.3. *Let the considered Levy process have the continuous density and let the corresponding quasi-potential* B *be regular and strongly sectorial. If* $B_1\in\Sigma_p$, $p>1$ *and*

$$1/p>\beta, \tag{9.36}$$

then the operator B *has a point of the spectrum different from zero.*

Proof. It follows from estimation (9.17) that

$$||(I-zB)^{-1}||_p\leq M, \quad |\arg z|\geq\beta+\epsilon. \tag{9.37}$$

Let us suppose that the formulated assertion is not true, i.e., the operator B has no points of the spectrum different from zero. We set

$$A(r,B)=\sup||(I-re^{i\theta}B)^{-1}||, \quad 0\leq\theta\leq 2\pi. \tag{9.38}$$

It follows (see [9]) from condition $B_1\in\Sigma_p$ that $B_2\in\Sigma_p$ and

$$\log A(r,B)=O(r^p). \tag{9.39}$$

According to Phragmen-Lindelöf theorem and relations (9.36)–(9.39) we have

$$||(I - zB)^{-1}|| \leq M. \tag{9.40}$$

The last relation is possible only when $B = 0$. But in our case $B \neq 0$. The obtained contradiction proves the theorem. □

Proposition 9.1. *Let the kernel of $\Phi(x, y)$ of the corresponding operator B be bounded. If this operator B is strongly sectorial, then it has a point of the spectrum different from zero.*

Proof. As in Theorem 9.3 we suppose that the operator B has no points of the spectrum different from zero. Using the boundedness of the kernel $\Phi(x, y)$ we obtain the inequality

$$TrB_1 < \infty. \tag{9.41}$$

It follows from relations (9.35) and (9.41) that (see the triangular model of M. Livshits [15]) $\rho = 1$. Since $\beta < 1$ all the conditions of Theorem 9.3 are fulfilled. Hence the proposition is proved. □

3. Now we shall consider the important case when

$$\mathrm{rank}\lambda_1 = 1. \tag{9.42}$$

Theorem 9.4. *Let the conditions of Theorem* 9.2 *be fulfilled. In the case* (9.42) *the following relation*

$$p(t, \Delta) = e^{-t/\lambda_1}[c_1 + o(1)], \quad t \to +\infty \tag{9.43}$$

is true.

Proof. In view of (9.31) we have

$$\lim \frac{1}{T} \int_0^T q(t)dt \geq \left| \lim \frac{1}{T} \int_0^T q(t) e^{-it(\mathrm{Im}\mu_j^{-1})} dt \right|, \quad T \to \infty, \tag{9.44}$$

i.e.,

$$g_1(0) \int_\Delta h_1(x)dx \geq |\overline{g_j(0)} \int_\Delta h_j(x)dx|. \tag{9.45}$$

In the same way we can prove that

$$g_1(x_0) \int_{\Delta_1} h_1(x)dx \geq |\overline{g_j(x_0)} \int_{\Delta_1} h_j(x)dx|, \tag{9.46}$$

where

$$x_0 \in \Delta_1 \in \Delta. \tag{9.47}$$

It follows from (9.46) that

$$g_1(x_0)h_1(x) \geq |\overline{g_j(x_0)}h_j(x)|. \tag{9.48}$$

We introduce the normalization condition

$$g_1(x_0) = g_j(x_0). \tag{9.49}$$

Due to (9.46) and (9.48) the inequalities

$$\int_{\Delta_1} h_1(x)dx \geq |\int_{\Delta_1} h_j(x)dx|, \tag{9.50}$$

$$h_1(x) \geq |h_j(x)| \tag{9.51}$$

are true. The equality sign in (9.50) and (9.51) can be only if

$$h_j(x) = |h_j(x)|e^{i\alpha}. \tag{9.52}$$

It is possible only in the case when $j = 1$. Hence there exists such a point x_1 that

$$h_1(x_1) > |h_j(x_1)|. \tag{9.53}$$

Thus we have

$$1 = \int_{\Delta_1} g_1(x)h_1(x)dx > \int_{\Delta_1} \overline{g_j(x)}h_j(x)dx = 1, \tag{9.54}$$

where $x_1 \in \Delta_1$. The received contradiction (9.54) means that $j = 1$. Now the assertion of the theorem follows directly from (9.5). □

Corollary 9.3. *Let conditions of Theorem* 9.2 *be fulfilled. If* $\mathrm{rank}\lambda_1 = 1$ *and* $x_0 \in \Delta_1 \in \Delta$ *then the asymptotic equality*

$$p(x_0, \Delta_1, t, \Delta) = e^{-t/\lambda_1} g_1(x_0) \int_{\Delta_1} h_1(x)dx[1 + o(1)], \quad t \to +\infty \tag{9.55}$$

is true.

The following Krein-Rutman theorem [13] gives the sufficient conditions when relation (9.42) is valid.

Theorem 9.5. *Suppose that the non-negative kernel* $\Phi(x, y)$ *satisfies the condition*

$$\int_{\Delta}\int_{\Delta} |\Phi(x, y)^2|dxdy < \infty \tag{9.56}$$

and has the following property: for each $\epsilon > 0$ *there exists an integer* $N = N(\epsilon)$ *such that the kernel* $\Phi_N(x, y)$ *of the operator* B^N *takes the value zero on a set of points of measure not greater than* ϵ. *Then*

$$\mathrm{rank}\lambda_1 = 1; \quad \lambda_1 > \lambda_k, \quad k = 2, 3, \ldots \tag{9.57}$$

It is easy to see that the following assertion is valid.

Proposition 9.2. *Let the inequality*

$$\Phi(x, y) > 0, \tag{9.58}$$

be true, when $x \neq a_k$, $x \neq b_k$, $y \neq a_k$, $y \neq b_k$. *Then*

$$g_1(x) > 0, \tag{9.59}$$

when $x \neq a_k$, $\quad x \neq b_k$.

4. Let us consider separately the case when the operator B is regular and

$$k(x) = k(-x). \tag{9.60}$$

The corresponding operator S is self-adjoint. Hence the operator B is self-adjoint and strongly sectorial. In this case equality (9.11) can be written in the form

$$p(t, \Delta) = \sum_{k=1}^{\infty} e^{-t/\lambda_k} g_k(0) \int_{\Delta} g_k(x) dx. \tag{9.61}$$

10. Stable processes, main notions

1. Let $X_1, X_2, \dots$ be mutually independent random variables with the same law of distribution $F(x)$. The distribution $F(x)$ is called *strictly stable* if the random variable

$$X = (X_1 + X_2 + \cdots + X_n)/n^{1/\alpha} \tag{10.1}$$

is also distributed according to the law $F(x)$. The number α $(0 < \alpha \leq 2)$ is called a *characteristic exponent* of the distribution. The homogeneous process $X(\tau)$ $(X(0) = 0)$ with independent increments is called a stable process if

$$E[\exp(i\xi X(\tau))] = \exp\left\{-\tau|\xi|^{\alpha}\left[1 - i\beta(\operatorname{sign}\xi)\left(\tan\frac{\pi\alpha}{2}\right)\right]\right\}, \tag{10.2}$$

where $0 < \alpha < 2, \alpha \neq 1, -1 \leq \beta \leq 1, \quad \tau > 0$. When $\alpha = 1$ we have

$$E[\exp(i\xi X(\tau))] = \exp\left\{-\tau|\xi|\left[1 + \frac{2i\beta}{\pi}(\operatorname{sign}\xi)(\log|\xi|)\right]\right\}, \tag{10.3}$$

where $-1 \leq \beta \leq 1, \quad \tau > 0$. The stable processes are a natural generalization of the Wiener processes. In many theoretical and applied problems it is important to estimate the value

$$p_{\alpha}(t, a) = P(\sup|X(\tau)| < a, \quad 0 \leq \tau \leq t). \tag{10.4}$$

For the stable processes Theorem 9.1. was proved earlier (see [19]–[22]). The value of $p_{\alpha}(t, a)$ decreases very quickly by the exponential law when $t \to \infty$. This fact prompted the idea to consider the case when the value of a depends on t and $a(t) \to \infty, \quad t \to \infty$. In this paper we deduce the conditions under which one of the following three cases is realized:

1) $\lim p_{\alpha}(t, a(t)) = 1$, $t \to \infty$.
2) $\lim p_{\alpha}(t, a(t)) = 0$, $t \to \infty$.
3) $\lim p_{\alpha}(t, a(t)) = p_{\infty}$, $0 < p_{\infty} \leq 1$, $t \to \infty$.

We also investigate the situation when $t \to 0$.

Remark 10.1. In the famous work by M. Kac [11] the connection of the theory of stable processes and the theory of integral equations was shown. M. Kac considered in detail only the case $\alpha = 1, \quad \beta = 0$. The case $0 < \alpha < 2$, $\beta = 0$ was later studied by H. Widom [30]. As to the general case $0 < \alpha < 2$, $-1 \leq \beta \leq 1$ it was investigated in our works [19]–[22]. In all the mentioned works the parameter a was fixed. Further we consider the important case when a depends on t and $a(t) \to \infty$, $t \to \infty$.

11. Stable processes, quasi-potential

1. In this section we formulate some results from our paper [20] (see also [22], Ch. 7). Here $\psi_\alpha(x,s,a)$ is defined by the relation

$$\psi_\alpha(x,s,a)=(I+sB_\alpha^\star)^{-1}\Phi_\alpha(0,x,a). \tag{11.1}$$

The quasi-potential B_α and its kernel $\Phi_\alpha(x,y,a)$ will be written later in the explicit form.

Further we consider three cases.

Case 1. $0<\alpha<2$, $\alpha\neq1$, $-1<\beta<1$.

Case 2. $1<\alpha<2$, $\beta=\pm1$.

Case 3. $\alpha=1$, $\beta=0$.

Now we introduce the operators

$$B_\alpha f=\int_{-a}^{a}\Phi_\alpha(x,y,a)f(y)dy \tag{11.2}$$

acting in the space $L^2(-a,a)$.

In *Case* 1 the kernel $\Phi_\alpha(x,y,a)$ has the following form (see[20], [22])

$$\Phi_\alpha(x,y,a)=C_\alpha(2a)^{\mu-1}\int_{a|x-y|}^{a^2-xy}[z^2-a^2(x-y)^2]^{-\rho}[z-a(x-y)]^{2\rho-\mu}dz, \tag{11.3}$$

where the constants μ,ρ, and C_α are defined by the relations $\mu=2-\alpha$,

$$\sin\pi\rho=\frac{1-\beta}{1+\beta}\sin\pi(\mu-\rho),\quad 0<\mu-\rho<1, \tag{11.4}$$

$$C_\alpha=\frac{\sin\pi\rho}{(\sin\pi\alpha/2)(1-\beta)\Gamma(1-\rho)\Gamma(1+\rho-\mu)}. \tag{11.5}$$

Here $\Gamma(z)$ is Euler's gamma function. We remark that the constants μ,ρ, and C_α do not depend on parameter a.

In *Case* 2 when $\beta=1$ the relation [20], [22]

$$\Phi_\alpha(x,y,a)=\frac{(\cos\pi\alpha/2)}{(2a)^{\alpha-1}\Gamma(\alpha)}\{[a(|x-y|+y-x)]^{\alpha-1}-(a-x)^{\alpha-1}(a+y)^{\alpha-1}\} \tag{11.6}$$

holds. In *Case* 2 when $\beta=-1$ we have [20], [22]

$$\Phi_\alpha(x,y,a)=\frac{(\cos\pi\alpha/2)}{(2a)^{\alpha-1}\Gamma(\alpha)}\{[a(|x-y|+x-y)]^{\alpha-1}-(a+x)^{\alpha-1}(a-y)^{\alpha-1}\}. \tag{11.7}$$

Finally, in *Case* 3 according to M. Kac [11] the equality

$$\Phi_1(x,y,a)=\frac{1}{4}\log\frac{[a^2-xy+\sqrt{(a^2-x^2)(a^2-y^2)}]}{[a^2-xy-\sqrt{(a^2-x^2)(a^2-y^2)}]} \tag{11.8}$$

is valid.

The important assertion (see [22], Ch. 7) follows from formulas (11.2)–(11.8):

Proposition 11.1. *Let one of the following conditions be fulfilled:*

I. $0 < \alpha < 2$, $\alpha \neq 1$, $-1 < \beta < 1$.
II. $1 < \alpha < 2$, $\beta = \pm 1$.
III. $\alpha = 1$, $\beta = 0$.

Then the corresponding operator B_α is regular and strongly sectorial.

2. Let us introduce the notation

$$p_\alpha(t, -b, a) = P(-b < X(\tau) < a, \quad 0 \leq \tau \leq t) \tag{11.9}$$

where $a > 0, \quad b > 0$. We consider in short the case when the parameter b is not necessary equal to a. As in the case $(-a, a)$ we have the relation

$$\int_0^\infty e^{-su} p_\alpha(u, , -b, a) du = \int_{-b}^{a} \psi_\alpha(x, s, -b, a) dx. \tag{11.10}$$

Here $\psi_\alpha(x, s, -b, a)$ is defined by the relation

$$\psi_\alpha(x, s, -b, a) = (I + sB_\alpha^\star)^{-1} \Phi_\alpha(0, x, -b, a). \tag{11.11}$$

Now the operator B_α has the form

$$B_\alpha f = \int_{-b}^{a} \Phi_\alpha(x, y, -b, a) f(y) dy \tag{11.12}$$

and acts in the space $L^2(-b, a)$. The kernel $\Phi_\alpha(x, y, -b, a)$ is connected with $\Phi_\alpha(x, y, a)$ (see (11.3) and (11.7)) by the formula

$$\Phi_\alpha(x, y, -b, a) = \Phi_\alpha \left(x + \frac{b-a}{2}, y + \frac{b-a}{2}, \frac{a+b}{2} \right). \tag{11.13}$$

In this way we have reduced the non-symmetric case $(-b, a)$ to the symmetric one $(-\frac{a+b}{2}, \frac{a+b}{2})$. Let us consider separately the case $0 < \alpha < 2$, $\beta = 0$. In this case the operator B_α is self-adjoint.We denote by λ_j, $(j = 1, 2, \dots)$ the eigenvalues of B_α and by $g_j(x)$ the corresponding real normalized eigenfunctions. Then we can write the new formula (see [20]) for $p_\alpha(t, -b, a)$ which is different from (9.11):

$$p_\alpha(t, -b, a) = \sum_{j=1}^{\infty} g_j(0) \int_{-b}^{a} g_j(x) dx e^{-t\mu_j}, \tag{11.14}$$

where $\mu_j = 1/\lambda_j$.

12. On sample functions behavior of stable processes

From the scaling property of the stable processes we deduce the relations

$$p_\alpha(t, a) = p_\alpha \left(\frac{t}{a^\alpha}, 1 \right), \tag{12.1}$$

$$\lambda_k(a, \alpha) = a^\alpha \lambda_k(1, \alpha). \tag{12.2}$$

We introduce the notations

$$\begin{aligned} \lambda_\alpha(1) = \lambda_\alpha, \qquad & p_\alpha(t,1) = p_\alpha(t), \\ g_\alpha(x,1) = g_\alpha(x), \quad & h_\alpha(x,1) = h_\alpha(x). \end{aligned} \tag{12.3}$$

Using relations (12.1), (12.2) and notations (12.3) we can rewrite Theorem 9.1 in the following way.

Theorem 12.1. *Let one of the following conditions be fulfilled:*

I. $0 < \alpha < 2$, $\alpha \neq 1$, $-1 < \beta < 1$.
II. $1 < \alpha < 2$, $\beta = \pm 1$.
III. $\alpha = 1$, $\beta = 0$.

Then the asymptotic equality

$$p_\alpha(t,a) = e^{-t/[a^\alpha \lambda_\alpha]} g_\alpha(0) \int_{-1}^{1} h_\alpha(x)dx[1 + o(1)], \quad t \to \infty. \tag{12.4}$$

holds.

Proof. The corresponding operator B_α is regular and strongly sectorial (see Proposition 11.1). The stable processes have the continuous density (see [31]). So all conditions of Theorem 9.1 are fulfilled. It proves the theorem. □

Remark 12.1. The operator B_α is self-adjoint when $\beta = 0$. In this case $h_\alpha = g_\alpha$.

Remark 12.2. The value λ_α characterizes how fast $p_\alpha(t,a)$ converges to zero when $t \to \infty$. The two-sided estimation for λ_α when $\beta = 0$ is given in [17] (see also [22], p. 150).

3. Now we consider the case when the parameter a depends on t. From Theorem 12.1. we deduce the assertions.

Corollary 12.1. *Let one of conditions* I–III *of Theorem* 12.1 *be fulfilled and*

$$\frac{t}{a^\alpha(t)} \to \infty, \quad t \to \infty. \tag{12.5}$$

Then the following equalities

$$1) \qquad p_\alpha(t, a(t)) = e^{-t/[a^\alpha(t)\lambda_\alpha]} g_\alpha(0) \int_{-1}^{1} h_\alpha(x)dx[1 + o(1)], \quad t \to \infty. \tag{12.6}$$

$$2) \qquad \lim p_\alpha(t,a) = 0, \quad t \to \infty. \tag{12.7}$$

$$3) \qquad \lim P[\sup |X(\tau)| > a(t)] = 1, \quad 0 \le \tau \le t, \quad t \to \infty \tag{12.8}$$

are true.

Corollary 12.2. *Let one of conditions* I–III *of Theorem* 12.1 *be fulfilled and*

$$\frac{t}{[a(t)]^\alpha} \to 0, \quad t \to 0. \tag{12.9}$$

Then the following equalities

1) $$\lim p_\alpha(t, a(t)) = 1, \quad t \to 0. \tag{12.10}$$

2) $$\lim P[\sup |X(\tau)| > a(t)] = 0 \quad 0 \le \tau \le t, \quad t \to 0 \tag{12.11}$$

are true.

Corollary 12.2 follows from (12.1) and the relation

$$\lim p_\alpha(t) = 1, \quad t \to 0. \tag{12.12}$$

Corollary 12.3. *Let one of conditions* I–III *of Theorem* 12.1 *be fulfilled and*

$$\frac{t}{[a(t)]^\alpha} \to T, \quad 0 < T < \infty, \quad t \to \infty. \tag{12.13}$$

Then the following equality is true:

$$\lim p_\alpha(t, a(t)) = p_\alpha(T), \quad t \to \infty. \tag{12.14}$$

Corollary 12.3 follows from (12.1).

13. Wiener process

1. We consider separately the important special case when $\alpha = 2$ (Wiener process). In this case the kernel $\Phi_2(x, t, -b, a)$ of the operator B_2 coincides with the Green's function (see [4], [11]) of the equation

$$-\frac{1}{2}\frac{d^2y}{dx^2} = f(x), \quad -b \le x \le a \tag{13.1}$$

with the boundary conditions

$$y(-b) = y(a) = 0, \quad b > 0, \quad a > 0. \tag{13.2}$$

It is easy to see that

$$\Phi_2(x, t, -b, a) = \frac{2}{a+b} \begin{cases} (t+b)(a-x) & -b \le t \le x \le a \\ (a-t)(b+x) & -b \le x \le t \le a. \end{cases} \tag{13.3}$$

Equality (12.1) is also true when $\alpha = 2$ and when $b = a$, i.e.,

$$p_2(t, a) = p_2(t/a^2, 1). \tag{13.4}$$

The eigenvalues of problem (13.1), (13.2) have the form

$$\mu_n = \left(\frac{n\pi}{a+b}\right)^2 /2, \quad n = 1, 2, 3 \ldots \tag{13.5}$$

The corresponding normalized eigenfunctions are defined by the equality

$$g_n(x) = \sqrt{\frac{2}{a+b}} \sin\left[\left(\frac{n\pi}{a+b}\right)(x+b)\right]. \tag{13.6}$$

Using formulas (13.5) and (13.6) we have

$$p_2(t,-b,a)=\sum_{m=0}^{\infty}\frac{4}{(2m+1)\pi}\sin\frac{(2m+1)b\pi}{a+b}e^{-t(\frac{(2m+1)\pi}{a+b})^2/2}. \tag{13.7}$$

Remark 13.1. If $b=a=1$ then relation (13.7)takes the form

$$p_2(t)=\sum_{m=0}^{\infty}(-1)^m\frac{2}{\pi(m+1/2)}e^{-t[\pi(m+1/2)]^2/2}. \tag{13.8}$$

Series (13.8) satisfies the conditions of Leibniz theorem. It means that $p_2(t,a)$ can be calculated with a given precision when the parameters t and a are fixed.

From (13.4) and (13.8) we deduce that

$$p_2(t,a)=\frac{4}{\pi}e^{-t\pi^2/8[a(t)]^2}(1+o(1)), \tag{13.9}$$

where $t/[a(t)]^2\to\infty$.

Proposition 13.1. *Theorem* 12.1 *and Corollaries* 12.1–12.3 *are true in the case when* $\alpha=2$ *too.*

Remark 13.2. From the probabilistic point of view it is easy to see that the function $p_2(t),(t>0)$ is monotonic decreasing and

$$0<p_2(t)\leq 1;\quad \lim p_2(t)=1,\quad t\to 0. \tag{13.10}$$

2. Now we shall describe the behavior of $p(t,-b,a)$ when $b\to\infty$. To do it we consider

$$\frac{d}{dt}p_2(t,-b,a)=-\frac{2\pi}{(a+b)^2}\sum_{m=0}^{\infty}(2m+1)\sin\frac{(2m+1)b\pi}{a+b}e^{-t(\frac{2m+1}{a+b}\pi)^2/2}. \tag{13.11}$$

We use the following Poisson result (see[7]).

Theorem 13.1. *If the function* $F(x)$ *satisfies the inequalities*

$$\int_0^{\infty}|F(x)|dx<\infty,\quad \int_0^{\infty}|F'(x)|dx<\infty \tag{13.12}$$

then the equality

$$\sum_{m=0}^{\infty}F(m)=\frac{1}{2}F(0)+\int_0^{\infty}F(x)dx+2\sum_{m=1}^{\infty}\int_0^{\infty}F(x)\cos 2\pi mxdx \tag{13.13}$$

is true.

Thus in case (13.11) we have

$$F(x)=G(x)-G(2x), \tag{13.14}$$

where

$$G(x)=-\frac{2\pi}{(a+b)^2}x\sin\frac{xb\pi}{a+b}e^{-t(\frac{x}{a+b}\pi)^2/2}. \tag{13.15}$$

It is easy to see that conditions (13.12) are fulfilled and

$$F(0)=0, \quad \int_0^\infty F(x)dx=\frac{1}{2}\int_0^\infty G(x)dx. \tag{13.16}$$

Using (13.15) and (13.16) we deduce the equality

$$\int_0^\infty F(x)dx=-\frac{1}{\pi t}\int_0^\infty ue^{-u^2/2}\sin\frac{ua}{\sqrt{t}}du, \tag{13.17}$$

where $u=\frac{x\pi}{a+b}\sqrt{t}$. Now we use the following relation from the sine transformation theory (see[25]).

$$\int_0^\infty ue^{-u^2/2}\sin(xu)du=\sqrt{\frac{\pi}{2}}xe^{-u^2/2}. \tag{13.18}$$

In view of (13.17) and (13.18) the equality

$$\int_0^\infty F(x)dx=-\frac{a}{\sqrt{2\pi}}t^{-3/2}e^{-a^2/2t} \tag{13.19}$$

is true. Now we calculate the integrals

$$J_m=2\int_0^\infty G(2x)\cos 2\pi mxdx, \quad I_m=2\int_0^\infty G(x)\cos 2\pi mxdx. \tag{13.20}$$

Using again formula (13.18) we have

$$J_m=-\sqrt{2/\pi}t^{-3/2}[A_me^{-A_m^2/2t}-B_me^{-B_m^2/2t}], \tag{13.21}$$

where $A_m=2m(a+b)+a, \quad B_m=2m(a+b)-a$. In the same way we obtain that

$$I_m=-\sqrt{1/2\pi}t^{-3/2}[C_me^{-C_m^2/2t}-D_me^{-D_m^2/2t}], \tag{13.22}$$

where $C_m=m(a+b)+a$, $D_m=m(a+b)-a$. From relation (13.7) and equality

$$\left[\int_{c/\sqrt{t}}^{d/\sqrt{t}}e^{-u^2/2}du\right]'=-\frac{1}{2}t^{-3/2}(de^{-d^2/2t}-ce^{-c^2/2t}) \tag{13.23}$$

we obtain the following representation of $p_2(t,-b,a)$:

$$p_2(t,-b,a)=1-\sqrt{2/\pi}\int_{a/\sqrt{t}}^\infty e^{-u^2/2t}du+q_\alpha(t,-b,a), \tag{13.24}$$

where

$$q_2(t,-b,a)=\sqrt{2/\pi}\sum_{m=1}^\infty[2\int_{B_m/\sqrt{t}}^{A_m/\sqrt{t}}e^{-u^2/2t}du-\int_{D_m/\sqrt{t}}^{C_m/\sqrt{t}}e^{-u^2/2t}du]. \tag{13.25}$$

So we have deduced two formulas (13.7) and (13.24) for $p_2(t,-b,a)$. Formula (13.7) is useful when t is big and the parameters a and b are fixed.

Proposition 13.2. *In the case of the Wiener process* ($\alpha=2$) *the asymptotic equality*

$$p_2(t,-b,a)=\frac{4}{\pi}\sin\frac{a\pi}{a+b}e^{-t(\pi)^2/[2(a+b)^2]}[1+o(1)], \quad t\to\infty \tag{13.26}$$

holds.

Formula (13.24) is useful when b is big and parameters a and t are fixed.

Proposition 13.3. *In the case of the Wiener process* ($\alpha = 2$) *the asymptotic equality*

$$p_2(t, -b, a) \tag{13.27}$$
$$= 1 - \sqrt{2/\pi} \int_{a/\sqrt{t}}^{\infty} e^{-u^2/2t} du - \sqrt{2/\pi} \int_{b/\sqrt{t}}^{(b+2a)/\sqrt{t}} e^{-u^2/2t} du[1 + o(1)],$$

where $b \to \infty$, *is valid.*

The well-known formula (see [8]) for *the first hitting time*

$$p_2(t, -\infty, a) = 1 - \sqrt{2/\pi} \int_{a/\sqrt{t}}^{\infty} e^{-u^2/2t} du \tag{13.28}$$

follows directly from (13.27).

14. Iterated logarithm law, most visited sites and first hitting time

It is interesting to compare our results (Theorem 9.1, Corollaries 12.1–12.3 and Proposition 13.1–13.3) with the well-known results mentioned in the title of the section.

1. We begin with the famous Khinchine theorem (see[4]) about the iterated logarithm law.

Theorem 14.1. *Let* $X(t)$ *be stable process* ($0 < \alpha < 2$). *Then almost surely* (*a.s.*) *that*

$$\lim_{t \to \infty} \frac{\sup |X(t)|}{(t \log t)^{1/\alpha} |\log |\log t||^{(1/\alpha)+\epsilon}} = \begin{cases} 0 & \epsilon > 0 \ \ a.s. \\ \infty & \epsilon = 0 \ \ a.s. \end{cases} \tag{14.1}$$

We introduce the random process

$$U(t) = \sup_{0 \le \tau \le t} |X(\tau)|. \tag{14.2}$$

From Corollaries 12.1–12.3 and Proposition 13.1 we deduce the assertion.

Theorem 14.2. *Let one of conditions* I–III *of Theorem* 12.1 *be fulfilled or let* $\alpha = 2$ *and*

$$b(t) \to \infty, \quad t \to \infty. \tag{14.3}$$

Then

$$b(t) U(t)/t^{1/\alpha} \to \infty \quad (P), \quad U(t)/[b(t) t^{1/\alpha}] \to 0 \quad (P). \tag{14.4}$$

(It is denoted by symbol (P), that the convergence is in probability.)

In particular we have:

$$[(\log^{\epsilon} t) U(t)]/t^{1/\alpha} \to \infty \quad (P), \quad U(t)/[(\log^{\epsilon} t) t^{1/\alpha}] \to 0 \quad (P), \tag{14.5}$$

when $\epsilon > 0$ and $t \to \infty$.

We see that our approach and the classical one have some similar points (estimation of $|X(\tau)|$), but these approaches are essentially different. We consider

the behavior of $|X(\tau)|$ on the interval $(0,t)$, and in the classical case $|X(\tau)|$ is considered on the interval (t,∞).

2. We denote by $V(t)$ the most visited site of stable process X up to time t (see [1]). We formulate the following result (see [1] and references therein).

Theorem 14.3. *Let* $1<\alpha<2, \quad \beta=0, \quad \gamma>9/(\alpha-1)$. *Then the relation*

$$\lim \frac{(\log t)^{\gamma}}{t^{1/\alpha}}|V(t)|=\infty, \quad t\to\infty \quad (a.s.) \tag{14.6}$$

is true.

To this important result we add the following estimation.

Theorem 14.4. *Let one of the conditions* I–III *of Theorem* 12.1 *be fulfilled or let* $\alpha=2$ *and*

$$b(t)\to\infty, \quad t\to\infty. \tag{14.7}$$

Then

$$|V(t)|/[b(t)t^{1/\alpha}]\to 0 \quad (P). \tag{14.8}$$

The formulated theorem follows directly from the inequality $U(t)\geq|V(t)|$.

In particular we have

$$|V(t)|/[(\log^{\epsilon} t)t^{1/\alpha}]\to 0 \quad (P) \tag{14.9}$$

when $\epsilon>0$ and $t\to\infty$.

Conjecture 14.1. *Let conditions of Theorem* 14.4 *be fulfilled. Then*

$$\lim_{t\to\infty}[V(t)/U(t)]=1. \tag{14.10}$$

3. The first hitting time T_a is defined by the formula

$$T_a=\inf_{t\geq 0}(X(t)\geq a). \tag{14.11}$$

It is obvious that

$$P(T_a>t)=P[\sup_{0\leq\tau\leq t} X(\tau)<a]. \tag{14.12}$$

We have

$$P(T_a>t)\geq P[-b<X(\tau)<a, 0\leq\tau\leq t]=p_{\alpha}(t,-b,a). \tag{14.13}$$

So our formulas for $p(t,-b,a)$ estimate $P(T_a>t)$ from below. It is easy to see that

$$p(t,-b,a)\to P(T_a>t), \quad b\to+\infty. \tag{14.14}$$

Remark 14.1. Our results can be interpreted in terms of the first hitting time $T_{[-b,a]}$ one of the barriers either $-b$ or a (ruin problem). Namely, we have

$$P(T_{[-b,a]}>t)=p(t,-b,a). \tag{14.15}$$

The distribution of the first hitting time for the Levy processes is an open problem.

Remark 14.2. B.A. Rogozin in his interesting work [18] established the law of the overshoot distribution for the stable processes when the existing interval is fixed.

References

[1] R.F. Bass, N. Eisenbaum, and Z. Shi, *The Most Visited Sites of Symmetric Stable Processes*. Probability Theory and Related Fields **116** (2000), 391–404.

[2] F.F. Bonsall and J. Duncan, *Numerical Ranges*. 1–49. MAA Studies in Mathematics **21** (ed. Bartle R.G.), 1980.

[3] G. Baxter and M.D. Donsker, *On the Distribution of the Supremum Functional for Processes with Stationary Independent Increments*. Trans. Amer. Math. Soc. **8** (1957), 73–87.

[4] J. Bertoin, *Levy Processes*. University Press, Cambridge, 1996.

[5] Z. Chuangyi, *Almost Periodic Type Functions and Ergodicity*, Kluwer, Beijing, New York, 2003.

[6] K.L. Chung, *Green, Brown and Probability*. World Scientific, 2002.

[7] M.A. Evgrafov, *Asymptotic Estimates and Entire Functions, Gordon and Breach*. New York, 1961.

[8] W. Feller, *An Introduction to Probability Theory and its Applications*. J. Wiley and Sons, 1971.

[9] I. Gohberg and M.G. Krein, *Introduction to the theory of linear nonselfadjont operators*. Amer. Math. Soc., Providence, R.I., 1969.

[10] K. Ito, *On Stochastic Differential Equations*. Memoirs Amer. Math. Soc. **4**, 1951.

[11] M. Kac, *On some Connections Between Probability Theory and Differential and Integral Equations*. Proc. Sec. Berkeley Symp. Math. Stat. and Prob., Berkeley, 189–215, 1951.

[12] M. Kac, *Probability and Related Topics in Physical Sciences*. Amer. Math. Soc., Colorado, 1957.

[13] M.G. Krein and M.A. Rutman, *Linear Operators Leaving Invariant a Cone in a Banach Space*. Amer. Math. Soc. Translation **26**, 1950.

[14] B.M. Levitan, *Some Questions of the Theory of Almost Periodic Functions*. Amer. Math. Soc., Translation **28**, 1950.

[15] M.S. Livshits, *Operators, Oscillations, Waves, Open Systems*. Amer. Math. Soc., Providence, R.I., 1973.

[16] A. Pietsch, *Eigenvalues and s-Numbers*. Cambridge University Press, 1987.

[17] S.M. Pozin and L.A. Sakhnovich, *Two-sided Estimation of the Smallest Eigenvalue of an Operator Characterizing Stable Processes*. Theory Prob. Appl. **36**:2 (1991), 385–388.

[18] B.A. Rogozin, *The distribution of the first hit for stable and asymptotically stable walks on an interval*. Theory Probab. Appl. **17** (1972), 332–338.

[19] L.A. Sakhnovich, *Abel Integral Equations in the theory of Stable Processes*. Ukr. Math. Journ. **36**:2 (1984), 193–197.

[20] L.A. Sakhnovich, *Integral Equations in the theory of Stable Processes*. St. Petersburg Math. J. **4**:4 (1993), 819–829.

[21] L.A. Sakhnovich, *The Principle of Imperceptibility of the Boundary in the Theory of Stable Processes*. St. Petersburg Math. J. **6**:6 (1995), 1219–1228.

[22] L.A. Sakhnovich, *Integral Equations with Difference Kernels.* Operator Theory **84**, Birkhäuser, 1996.

[23] K. Sato, *Levy Processes and Infinitely Divisible Distributions.* University Press, Cambridge, 1999.

[24] W. Schoutens, *Levy Processes in Finance.* Wiley series in Probability and Statistics, 2003.

[25] E.C. Titchmarsh, *Introduction to the Theory of Fourier Integrals.* Oxford, 1937.

[26] M. Stone, *Linear Transformation in Hilbert space.* New York, 1932.

[27] M. Thomas and O. Barndorff, (ed.), *Levy Processes; Theory and Applications.* Birkhäuser, 2001.

[28] L.N. Trefethen and M. Embree, *Spectra and Pseudospectra.* Princeton University Press, 2005.

[29] P. Tuominen and R.L. Tweedie, *Exponential Decay and Ergodicity of General Markov Processes and their Discrete Skeletons.* Adv. in Appl. Probab. **11** (1979), 784–803.

[30] H. Widom, *Stable Processes and Integral Equations.* Trans. Amer. Math. Soc. **98** (1961), 430–449.

[31] V.M. Zolotarev, *One-dimensional stable distribution.* Amer. Math. Soc. Providence, 1986.

Lev Sakhnovich
99 Cove ave.
Milford , CT., 06461, USA
e-mail: lsakhnovich@gmail.com

Operator Theory:
Advances and Applications, Vol. 197, 375–389

β-spread of Sets in Metric Spaces and Critical Values of Smooth Functions

Yosef Yomdin

To the memory of Moshe Livsic

Abstract. This paper provides new geometric restrictions on the set of critical values of differentiable functions. The classical and widely used condition on critical values of differentiable mappings is given by the Morse-Sard theorem ([20, 38, 39]): if the mapping is C^k-smooth, with k sufficiently big, then the set of its critical values has the Lebesgue measure (or, more precisely, the Hausdorff measure of an appropriate dimension) zero.

In a work of the author since 1981 ([43]–[46], [50], and others) which was strongly inspired by questions, remarks, and constructive criticism of Moshe Livsic, it was shown that in fact the critical values of any differentiable mapping satisfy geometric restrictions much stronger than just the property to be of measure zero. These restrictions are given in terms of the metric entropy and they turn out to be pretty close to a complete characterization of the possible sets of critical values. Still a gap between the necessary and sufficient conditions remained.

In the present paper we close (partially) this gap, introducing into consideration of critical values a certain geometric invariant (we call it β-spread) which was previously studied in quite different relations. We show that in many important cases β-spread provides a complete characterization of critical values of differentiable functions.

Mathematics Subject Classification (2000). 58K05; 57R45; 31B15.

Keywords. Critical values, Sard theorem, metric entropy, spanning trees.

1. Introduction

This paper provides new geometric restrictions on the set of critical values of differentiable functions and mappings. Let us remind the reader that the classical and widely used condition on critical values of differentiable mappings is given by the Morse-Sard theorem ([20, 38, 39]): if the mapping $f : M^n \to \mathbb{R}$ is C^k-smooth, with k greater than or equal to the dimension n of the source, then the set of its critical values has the Lebesgue measure zero.

Since around 1981 the author has started investigation of the geometry of smooth mappings ([43]–[46], [50], and others). This work was strongly inspired by numerous discussions with Moshe Livsic, by his questions, remarks, and constructive criticism.

In particular, it was shown that in fact the critical values of any differentiable mapping satisfy geometric restrictions much stronger than just the property to be of measure zero. These restrictions are given in terms of metric entropy which is the logarithm of the minimal number of balls of radius ϵ, covering the space X (see, in particular, [1, 14, 15], [17]–[19], [25]–[31], [40]–[42] and references there). Metric entropy restriction turns out to be pretty close to a complete characterization of critical values of smooth functions. Since then many additional results have been obtained in this direction (see [2]–[6], [23, 24, 26], [32]–[36]). Still a gap between the necessary and sufficient conditions remains.

In the present paper we close (partially) this gap, introducing into consideration of critical values a certain natural geometric invariant that we call β-spread. Similar invariants have been intensively studied earlier, but in quite different problems ([14, 21, 29, 30]. See, in particular, [29] where a short proof is given of the main results on the β-spread we use in this paper).

Some initial results on the β-spread of critical values were obtained in [44]. It was shown there that in some important cases β-spread provides a complete characterization of critical values of differentiable functions. However, the preprint [44] was never published. In the present paper we prove strongly extended results of [44] concerning critical values. In [47] we plan to present extended versions of some geometric results of [44], relating them, in particular, with "capacities" and "transfinite diameters" as appear in Potential Theory (see [16, 37]).

β-spread naturally appears also in discrete versions of "Remez-type" inequalities for polynomials, as well as in geometry of semi-algebraic sets (see [48, 49]).

Let us mention also a possible relation of the β-spread with a "spread-like" invariant introduced in [22] as well as with the "irregularities of geometric distributions" (see [7]–[13]).

Now let us describe the problem treated in the present paper in somewhat more detail. First we return to the metric entropy of critical values. The main result of [43] is the upper bound for the ϵ-entropy of the set of "near-critical" values of a C^k-mapping $f : M \to \mathbb{R}^m$ of an n-dimensional compact smooth manifold M. This result implies, in a special case of critical values $\Delta(f)$ of a function $f : M \to \mathbb{R}$, the following strengthening of the Morse-Sard theorem:

$$\dim_e(\Delta(f)) \leq \frac{n}{k},$$

where $\dim_\epsilon$ is the so-called entropy (or "box") dimension (see Section 2 below). This bound gives a necessary condition for a given set to be the set of critical values, which is much stronger than the condition just to be of measure zero. The following example from [43] illustrates the flavor of the new restrictions on critical values:

Corollary 5.5, [43]. *The set* $\{1, 1/2^a, 1/3^a, \dots, 1/n^a, \dots, 0\}$ *cannot be the set of critical values of a* k *times continuously differentiable function on an* n*-dimensional compact smooth manifold, if* $k > n(a+1)$.

Thus the results of [43] (see also [50]) show that actual geometric restrictions on critical values of general differentiable mappings are strong enough to imply non-trivial consequences, for instance, for geodesics on compact Riemannian manifold, for critical values of complex analytic functions, etc.

The following important question then arises: *what is a necessary and sufficient condition for a given compact set* $A \subset \mathbb{R}^m$ *to be the set* $\Delta(f)$ *of critical values of some* C^k*-smooth mapping* $f : M \to \mathbb{R}^m$ *of an* n*-dimensional compact smooth manifold* M*?*

Finiteness of the β-spread for $\beta = \frac{n}{k}$ in many cases provides an answer. Conjecturally, this is the necessary and sufficient condition we look for.

Roughly, the β-spread $V_\beta(A)$ of a subset A in a metric space X, for $\beta > 0$, is the supremum with respect to all finite subsets of A of the sum of βth degrees of the lengths of edges in the shortest tree, connecting the points of a given finite subset of A. This invariant turns out to be a close neighbor of the ϵ-entropy of A, but sometimes it provides a more accurate geometric description of the set. It is also closely related to the notions of capacity and transfinite diameter which appears in Potential Theory (see, for example, [16, 37]). We plan to cover these relations in [47]. In the present paper we state most of the required properties of the β-spread without proofs, referring the reader to [44, 29], and to the references there. In particular, in [44] detailed (elementary but sometimes pretty involved) proofs of all the results on β-spread used in the present paper are given. Here we concentrate instead on the description of the critical values of smooth functions.

The following are the main results of this paper: in Theorem 3.1 we give (somewhat overlapping with [6]) a complete description of the sets of critical values of C^k functions of one variable in terms of their $\frac{1}{k}$-spread.

Next we prove in Theorem 3.2 the finiteness of $V_{\frac{n}{k}}(\Delta(f)$ in the following two cases: for C^k-functions $f : M \to \mathbb{R}$ for any k, and for mappings $f : M \to \mathbb{R}^m$, $m > 1$, assuming $k \leq 3$. In Theorem 3.3 we extend this last result to any k but only for the "strongly critical" values $\Delta_l(f)$ (attained at the points where all the derivatives of f up to the order $[\frac{k-1}{2}]$ vanish).

Finally, in Theorem 3.4 we give in a general case an explicit upper bound for the β-spread $V_\beta(\Delta(f)$ of critical values of a C^k-smooth mapping f, for each $\beta > \frac{n}{k}$. However, this bound blows up as $\beta \to \frac{n}{k}$.

The role of M. Livsic in this work has been mentioned above. The author would also like to thank (after 27 years) Y. Kannai, H. Furstenberg and Y. Katznelson for useful discussions, and SFB-40, Universität Bonn where the first version of this work ([44]) has been written, for its kind hospitality. We would like to thank V. Katsnelson and M. Sodin for useful recent discussions and for providing new references.

2. β-spread, metric entropy and Hausdorff measure

In this section we give all the necessary definitions and properties of β-spread, referring the reader to [29] and references there, and to [44, 47] for most of the proofs. We also recall the definitions and some basic properties of metric entropy and Hausdorff measure.

2.1. Definition and some properties of β-spread

Let G_p be the set of all connected nonoriented trees with p vertices. We write $(i,j) \in g$, for $g \in G_p$, if the vertices i and j are connected by the edge in g.

Definition 2.1. *Let X be a metric space, $\beta > 0$. For each $x_1, \dots, x_p \in X$ and $g \in G_p$ let $\rho_\beta(g, x_1, \dots, x_p) = \sum_{(i,j)\in g} d(x_i, x_j)^\beta$, where d is a distance in X. Define $\rho_\beta(x_1, \dots, x_p)$ as $\inf_{g\in G_p} \rho_\beta(g, x_1, \dots, x_p)$. The tree g on which the infimum is achieved is called the β-minimal spanning tree. Now let $A \subset X$. We define the β-spread of A, $V_\beta(A)$, by*

$$V_\beta(A) = \sup_{p, x_1, \dots, x_p \in A} \rho_b(x_1, \dots, x_p).$$

For $x_1, \dots, x_p \in X$, $\rho_\beta(x_1, \dots, x_p)$ is called a β-weight of the minimal spanning tree g on $(x_1, \dots, x_p)$. Notice that the 1-minimal tree is also minimal for any β (see [29] and references there).

Under a different name β-spread for subsets of a real line has been studied by Besicovich and Taylor in [14]. A notion of β-weight have appeared earlier in geometric combinatorics and in fractal geometry. Compare [29, 30], [17]–[19] and references there. However, we are not aware of any appearance of the general notion of β-spread in metric spaces, as defined above.

Let us also notice that as a function of β the spread $V_\beta(A)$ is a kind of a zeta-function. For $A = \{0, 1, 3, 6, 10, 15, \dots, \frac{1}{2}n(n+1), \dots\}$ the spread $V_\beta(A)$ is exactly the Riemann ζ-function $\zeta(-\beta)$, while for $A = \{\alpha_0, \alpha_1, \dots, \alpha_n, \dots\}$ with $\alpha_0 = 0$, $\alpha_1 = 1, \dots, \alpha_n = \sum_{i=1}^n \frac{1}{i}$, $n \geq 1$, we have $V_\beta(A) = \zeta(\beta)$. So it may be a good idea to substitute into $V_\beta(A)$ complex values of β. See [30] for a detailed treatment of fractal geometry from this point of view.

Now we state some simple properties of the β-spread: let X be a metric space, $A, A_1, A_2, \dots \subset X$, $\beta > 0$.

Proposition 2.1.

1. *If $A_1 \subset A_2$ then $V_\beta(A_1) \leq V_\beta(A_2)$.*
2. *If $A_1 \subset A_2 \subset A_3 \subset \cdots$, and $A = \cup A_i$, then $V_\beta(A) = \lim V_\beta(A_i)$.*
3. *$V_\beta(A) = V_\beta(\bar{A})$, where $\bar{A}$ is the closure of A.*
4. *$V_\beta(A_1 \cup A_2) \leq V_\beta(A_1) + V_\beta(A_2) + [d(A_1, A_2)]^\beta$. Here the distance $d(A_1, A_2)$ of two subsets $A_1, A_2 \subset X$ is defined by $d(A_1, A_2) = \inf_{x\in A_1, y\in A_2} d(x, y)$.*
5. *For $A \subset X$ let $\operatorname{diam}(A) = \sup_{x,y\in A} d(x, y)$. If*

$$d(A_1, A_2) \geq \max(\operatorname{diam} A_1, \operatorname{diam} A_2)$$

then $V_\beta(A_1 \cup A_2) = V_\beta(A_1) + V_\beta(A_2) + d(A_1, A_2)^\beta$.

6. *Let* $\beta_1 > \beta_2 > 0$. *Then* $V_{\beta_1}(A) \leq [\text{diam } A]^{\beta_1-\beta_2} V_{\beta_2}(A)$.
7. *Let* X, Y *be metric spaces and let* $A \subset X$, $B \subset Y$. *If there exists an epimorphism* $\psi : A \to B$ *with* $d(\psi(x), \psi(y)) \leq K \cdot d(x,y)^N$, *for any* $x, y \in A$, $N > 0$, *then for each* $\beta > 0$, $V_\beta(B) \leq K^\beta V_{N\beta}(A)$.

Let us give two simple examples.

1. $A = [a, b]$. Here $V_\beta(A) = \infty$, $\beta < 1$ and $V_\beta(A) = (b-a)^\beta$, $\beta > 1$.
2. $A = \{1, 1/2^a, 1/3^a, \ldots, 1/n^a, \ldots, 0\}$. Here $V_\beta(A) = \infty$, $\beta \leq \frac{1}{a+1}$, and $V_\beta(A) = \sum_{n=1}^{\infty} [(\frac{1}{n})^a - (\frac{1}{n+1})^a]^\beta < \infty$ for $\beta > \frac{1}{a+1}$.

To further clarify the geometric meaning of β-spread, we compare it with some other geometric characteristics of subsets in a metric space.

2.2. Metric entropy

The notion of metric entropy can be traced at least to the work of Minkowski ("Minkowski content" – see [25, 15, 30]). In the study of functional classes and in Dynamics this notion was introduced by Kolmogorov ([27, 28]). The presentation below follows these last works.

Definition 2.2. *We define* $M(\epsilon, A)$ *as the minimal number of sets of diameter* $\leq \epsilon$ *covering* A. $N(\epsilon, A)$ *is defined as the maximal cardinality of an* $\frac{\epsilon}{2}$*-separated set* S *in* A, (*a subset* S *in* A *is said to be* ϵ*-separated if every two its distinct points are at a distance greater than* ϵ *from each other*). *Now* $H_\epsilon(A) = \log_2 M(\epsilon, A)$ and $C_\epsilon(A) = \log_2 N(\epsilon, A)$ *are called the* ϵ*-entropy and the* ϵ*-capacity of* A, *respectively.*

For each $\epsilon > 0$ the quantities $M(\epsilon, A)$, $N(\epsilon, A)$, $H_\epsilon(A)$ and $C_\epsilon(A)$ are related by the following inequalities ([28], Theorem IV):

$$N(2\epsilon, A) \leq M(\epsilon, A) \leq N(\epsilon, A),$$
$$C_{2\epsilon}(A) \leq H_\epsilon(A) \quad \leq C_\epsilon(A).$$

In the examples above we have: for $A = [a, b]$ the covering number $M(\epsilon, A) \asymp \frac{b-a}{\epsilon}$, and for $A = \{1, 1/2^a, 1/3^a, \ldots, 1/n^a, \ldots, 0\}$ the covering number $M(\epsilon, A) \asymp (\frac{1}{\epsilon})^{\frac{1}{a+1}}$.

Comparing the definition of β-spread with that of $N(\epsilon, A)$, we obtain immediately:

Proposition 2.2. *For each* $\beta > 0$,

$$V_\beta(A) \geq \sup_{\epsilon > 0} \epsilon^\beta (N(\epsilon, A) - 1).$$

To obtain an upper estimate for the β-spread in terms of ϵ-entropy is more difficult. First we consider the following function which is in some sense an inverse function of $N(\epsilon, A)$:

Definition 2.3. *For $x_1, \dots, x_p \in X$, let $\nu(x_1, \dots, x_p) = \min_{i \neq j} d(x_i, x_j)$. For $A \subset X$ define $\eta_A(p)$ for any natural $p \geq 2$ by*

$$\eta_A(p) = \sup_{x_1, \dots, x_p \in A} \nu(x_1, \dots, x_p).$$

The following theorem relates $\eta_A(p)$ and $V_\beta(A)$:

Theorem 2.1. *For any $\beta > 0$*

$$\sup_{p \geq 2}(p-1)\eta_A^\beta(p) \;\leq\; V_\beta(A) \;\leq\; \sum_{j=2}^{\infty} \eta_A^\beta(j).$$

The proof of Theorem 2.1 is given in [44, 47].

The following inequalities relate $\eta_A(p)$ with the metric entropy of A:

Lemma 2.1. *For each $\epsilon > 0$ and $p = M(\epsilon, A) + 1$ we have $\eta_A(p) \leq \epsilon$.*

Proof: We cover A by $M(\epsilon, A)$ sets of diameter ϵ. For any p points $x_1, \dots, x_p \in X$ at least two belong to the same ϵ-ball and hence the distance between these two points is at most ϵ. Therefore $\nu(x_1, \dots, x_p) \leq \epsilon$. This completes the proof.

The function $M(\epsilon, A)$ is decreasing in ϵ. Let us define the inverse function $\epsilon_A(p)$ as the smallest ϵ for which $M(\epsilon, A) \leq p$. We have, in particular, $M(\epsilon_A(p), A) \leq p$.

Lemma 2.2. *For each p we have $\eta_A(p) \leq \epsilon_A(p)$.*

Proof: The function $\eta_A(p)$ is decreasing in p. Hence $\eta_A(p) \leq \eta_A(M(\epsilon_A(p), A)) \leq \epsilon_A(p)$ by Lemma 2.1.

Finally, the following proposition provides an upper bound on $V_\beta(A)$ in terms of the metric entropy of A:

Proposition 2.3. *For each $\beta > 0$ we have $V_\beta(A) \leq \sum_{p=2}^{\infty} \epsilon_A^\beta(p)$.*

Proof: It follows directly from Theorem 2.1 and Lemma 2.2.

2.3. Hausdorff measure

We can compare V_β also with the β-Hausdorff measure, which is defined as

$$S_\beta(A) = \lim_{\alpha \to 0} S_\beta^\alpha(A),$$

where $S_\beta^\alpha(A)$ is the lower bound of all sums of the form $\sum_{i=1}^{\infty} r_i^\beta$, $r_i \leq \alpha$ and $A \subset \cup_{i=1}^{\infty} A_i$, with the diam $A_i \leq r_i$. (See, e.g., [3].) We have the following easy inequality between the β-spread and the Hausdorff measure:

Proposition 2.4. *For all $\beta > 0$, $S_\beta(A) \leq V_\beta(A)$.*

2.4. Entropy, Hausdorff and V-dimensions

We recall here the notions of the entropy and the Hausdorff dimensions, and define in the same way the V-dimension.

Definition 2.4. *Let $A \subset X$ be a bounded subset.*

1. $\dim_e A = \inf\{\beta,\ \exists K$, *such that for each* $\epsilon > 0$, $N(\epsilon, A) \leq K(\frac{1}{\epsilon})^\beta\}$ *is called the entropy dimension of A.*
2. $\dim_H A = \inf\{\beta,\ S_\beta(A) < \infty\}$ *is called the Hausdorff dimension of A.*
3. $\dim_V A$ *is defined as* $\inf\{\beta,\ V_\beta(A) < \infty\}$.

The notion of the entropy dimension appears in fractal geometry under many different names, in particular: "Minkowski dimension" – probably, the most justified historically, – "capacity dimension", "box dimension".

The notion of V-dimension was introduced in [44]. Essentially the same definition, under the name "MST (minimal spanning tree) dimension", is given in [29].

The following result was conjectured by H. Furstenberg, who also gave a proof of it (around 1982, unpublished):

Theorem 2.2. *For any bounded $A \subset X$, $\dim_V A = \dim_e A \leq \dim_H A$.*

A proof of this result can be found in [44, 29].

2.5. β-spread for subsets in $\mathbb{R}^n$

Coincidence of dimensions $\dim_V$ and $\dim_e$ stated in Theorem 2.2 above shows that for any bounded $A \subset \mathbb{R}^n$, $\dim_V(A) \leq n$. In fact, we have the following more precise result:

Theorem 2.3. *There is a constant K_n depending only on the dimension n such that for any $A \subset \mathbb{R}^n$,*

$$V_n(A) \leq K_n[\mathrm{diam}(A)]^n.$$

The proof of this theorem has a long story, starting at least in [21]. Another proof based essentially on the same ideas as in [21] was given independently by the author in [44]. See [29] for a nice and short proof of Theorem 2.3, as well as for a discussion of some history of the subject.

The opposite inequality is much easier:

Theorem 2.4. *Let $A \subset \mathbb{R}^n$ be open. Then $V_n(A) \geq m(A)$.*

For subsets of a real line V_β can be computed explicitly. Since $V_\beta(A) = V_\beta(\bar{A})$ we can assume A to be closed.

Theorem 2.5. *Let $A \subset \mathbb{R}$ be a closed bounded set, $a = \inf\ A$, $b = \sup\ A$, and $A = [a, b] \setminus \cup_{i=1}^\infty V_i$, where V_i are disjoint open intervals. Denote the length of V_i by α_i. Then*

1. *$V_\beta(A) = (b-a)^\beta$ for $\beta \geq 1$.*
2. *For $\beta < 1$, $V_\beta(A) = \infty$ if $\sum_{i=1}^\infty \alpha_i < b - a$.*
3. *If $\sum_{i=1}^\infty \alpha_i = b - a$, then $V_\beta(A) = \sum_{i=1}^\infty \alpha_i^\beta$ for $\beta < 1$.*

For our applications of β-spread to critical values we need also the following results, concerning the mappings of subsets in $\mathbb{R}$ with a bounded β-spread:

Proposition 2.5. *Let $A \subset \mathbb{R}$ be a bounded subset, $0 < \beta < 1$, and let $V_\beta(A) < \infty$. Then there exists a homeomorphism $\psi : \mathbb{R} \to \mathbb{R}$, such that for any $x_1, x_2 \in A$*

$$|\psi(x_1) - \psi(x_2)| \ \geq \ \omega(|x_1 - x_2|) \cdot |x_1 - x_2|^\beta,$$

where $\omega(\xi) \to \infty$ as $\xi \to 0$.

Proposition 2.6. *Let $A \subset R$ be a bounded set, $0 < \beta < 1$, and let $V_\beta(A) < \infty$. Then there exists a bounded set $A' \subset \mathbb{R}$ and a homeomorphism $\phi : \mathbb{R} \to \mathbb{R}$, $\phi(A') = A$, such that for any $x_1, x_2 \in A'$*

$$|\phi(x_1) - \phi(x_2)| \leq \gamma(|x_1 - x_2|)|x_1 - x_2|^{1/\beta},$$

where $\gamma(\xi) \to 0$ as $\xi \to 0$.

Similarly, one proves the following:

Proposition 2.7. *Let $A \subset \mathbb{R}$ be a bounded subset, such that $V_\beta(A) < \infty$ for all $\beta > 0$. Then there exists a bounded subset $A' \subset \mathbb{R}$ and a homeomorphism $\phi : \mathbb{R} \to \mathbb{R}$, $\phi(A') = A$, such that for each $N > 0$, $|\phi(x_1) - \phi(x_2)| \leq |x_1 - x_2|^N$, for any $x_1,\ x_2 \in A'$ with $|x_1 - x_2|$ sufficiently small.*

The proof of these three propositions is given in [44].

3. β-spread of critical values

Let $f : M^n \to \mathbb{R}$ be a C^k-smooth function on a smooth n-dimensional manifold M, $k \geq 1$. We denote by $\Delta(f)$ the set of critical values of f, $\Delta(f) = f(\Sigma(f)) \subset \mathbb{R}$, where $\Sigma(f) = \{x \in M, df(x) = 0\}$.

The Morse-Sard theorem (see [38, 39], [20], Theorem 3.4.3) gives an upper bound for the Hausdorff dimension of $\Delta(f)$:

$$\dim_h \Delta(f) \leq \frac{n}{k}. \tag{3.1}$$

In [43, 50] it is shown, that the entropy dimension is a more adequate and stronger notion in study of critical values. In particular, the following strengthening of (3.1) is true ([43], Theorem 5.4): assuming that M is compact,

$$\dim_e \Delta(f) \leq \frac{n}{k}. \tag{3.2}$$

Moreover, the necessary condition (3.2) for a given set to be of the form $\Delta(f)$ turns out to be "almost sufficient": let us say that a bounded set $A \subset \mathbb{R}$ has a property $P(n,k)$ if there is a compact n-dimensional manifold M and a C^k-smooth function $f : M \to \mathbb{R}$, such that $A \subset \Delta(f)$. Then we have ([43], Theorem 5.6):

$(*)$ If a bounded set $A \subset \mathbb{R}$ has a property $P(n,k)$, then $\dim_e A \leq \frac{n}{k}$. If $\dim_e A < \frac{n}{k}$, then A has a property $P(n,k)$.

However, to give a necessary and sufficient condition for a given compact set to be the set of critical values of a function of a given smoothness on a compact manifold of a given dimension, one has to analyze the metric properties of this set in more detail.

Conjecture. A compact set $A \subset \mathbb{R}$ is of a form $A = \Delta(f)$ for some C^k-smooth $f : M^n \to \mathbb{R}$, $k > n$, M compact, if and only if $V_{n/k}(A) < \infty$.

In the present paper we check this conjecture for functions of one variable and also prove the necessity of the condition $V_{n/k}(A) < \infty$ for functions $f : M^n \to \mathbb{R}$ with arbitrary n (and also for mappings $f : M^n \to \mathbb{R}^p$, under the restriction $k \leq 3$).

Theorem 3.1 (Compare [6]**).** *A compact set $A \subset \mathbb{R}$ is the set of all critical values of a k times continuously differentiable function $f : [0,1] \to \mathbb{R}$, $k > 1$, if and only if $V_{1/k}(A) < \infty$.*

A is the set of all critical values of an infinitely differentiable function $f : [0,1] \to \mathbb{R}$ if and only if $\dim_e A = 0$.

Proof: **1. Necessity**. We use the following lemma, which can be proved easily by successive applications of the mean value theorem:

Lemma 3.1. *Let $f : [a,b] \to \mathbb{R}$ be a k times differentiable function. If f has in $[a,b]$ more than $k-2$ critical points, then*

$$\max_{x\in[a,b]} f(x) - \min_{x\in[a,b]} f(x) \leq |f^{(k)}(c)|(b-a)^k,$$

for some $c \in [a,b]$.

Remark. Generalization of this lemma to functions of several variables is obtained in [43], Theorem 3.11. For $k \leq 3$ this generalization follows easily from Lemma 3.2 below. On the other hand, Lemma 3.1 is directly applicable to any couple of "strongly critical" points, where at least $[\frac{k}{2}]$ of the successive derivatives vanish – see Lemma 3.3 below.

Now let $f : [0,1] \to \mathbb{R}$ be a k times continuously differentiable function. Denote by M the $\max_{c\in[a,b]} |f^{(k)}(c)|$.

For any $y_1, \ldots, y_p \in \Delta(f)$, $p \geq k-1$, let $x_1, \ldots, x_p \in [0,1]$ be the critical points of f, such that $f(x_i) = y_i$, $i = 1, \ldots, p$. Reordering this set, if necessary, we can assume that $x_1 < x_2 < \cdots < x_p$.

Consider intervals $e_1 = [a, x_{k-1}]$, $e_2 = [x_{k-1}, x_{2k-3}], \ldots, e_s = [x_r, b]$, where $s = (\frac{p-1}{k-2})$, $r = (k-2)(s-1)+1$, for $k \geq 3$, and $s = r = p-1$ for $k = 2$. Each of these intervals contains exactly $k-1$ points x_i ($2 = k$ points for $k = 2$), except the last one, which contains not less than $k-1$ and not more than $2k-4$ points x_i.

Denote by d_j the length of e_j, $j = 1, \ldots, s$. For any x_m, $x_n \in e_j$, $|y_n - y_n| = |f(x_m) - f(x_n)| \leq M d_j^k$, by Lemma 3.1.

Now the chain $g = \{(1,2),(2,3),\dots,(p-1,p)\}$ is a tree on $y_1,\dots,y_p$ (possibly, non-minimal), and thus we have:

$$\rho_{1/k}(y_1,\dots,y_p) \le \rho_{1/k}(\gamma, y_1,\dots,y_p) = \sum_{i=1}^{p-1} |y_{i+1}-y_i|^{1/k}$$

$$= \sum_{j=1}^{s} \sum_{[x_i,x_{i+1}]\in e_j} |y_{i+1}-y_i|^{1/k} \le (2k-4)M^{1/k}\sum_{j=1}^{s} d_j = M^{1/k}(2k-4)(b-a).$$

Taking supremum in the left-hand side, we obtain

$$V_{1/k}(\Delta(f)) \le M^{1/k}(2k-4)(b-a) < \infty.$$

For infinitely differentiable f we have $V_{1/k}(\Delta(f)) < \infty$ for all k, and hence $\dim_e \Delta(f) = 0$.

2. Sufficiency. Let us fix some infinitely differentiable function $u : [0,1] \to \mathbb{R}$ with the following properties:

1. $u(t) > 0 \quad \text{for} \quad t \in (0,1)$.
2. $\int_0^1 u(t)dt = 1$.
3. $u(0) = u'(0) = \dots = u(1) = u'(1) = \dots = 0$.

Denote by N_q the $\max_{t\in[0,1]}|u^{(q-1)}(t)|$.

Now let $A \subset \mathbb{R}$ be a compact set with $V_{1/k}(A) < \infty$. By Proposition 2.6 there exists a compact set $A' \subset \mathbb{R}$ and a homeomorphism $\phi : \mathbb{R} \to \mathbb{R}$ with $\phi(A') = A$ and $|\phi(x_1) - \phi(x_2)| \le \gamma(|x_1 - x_2|)|x_1 - x_2|^k$, where $\gamma(\xi) \to 0$ as $\xi \to 0$.

Let $a = \inf\ A'$, $b = \sup\ A'$, and let $A' = [a,b] \setminus \bigcup_{i=1} U_i$, where $U_i = (a_i, b_i)$ are disjoint intervals. Let β_i be the length of U_i.

Denote also by α_i the length of $V_i = \phi(U_i) = (\phi(a_i), \phi(b_i))$. We have $\alpha_i \le \gamma(\beta_i)\beta_i^k$ for all i.

Define h on $[a,b]$ by

$$h(x) = 0,\ x \in A',\ h(x) = \frac{\alpha_i}{\beta_i}\ u\left(\frac{x-a_i}{\beta_i}\right),\ x \in U_i,$$

and let $f(x) = \phi(a) + \int_a^x h(t)dt$.

We shall prove that $f \in C^k$ on $[a,b]$ and $\Delta(f) = A$.

Clearly, $f \in C^\infty$ on $[a,b] \setminus A'$. To prove that $f \in C^k$ on all $[a,b]$ it is sufficient to show that $\lim_{y\to x, y\in[a,b]\setminus A'}\ f^{(q)}(y) = 0$ for any $q,\ 1 \le q \le k$, and $x \in A'$.

Now if $x \in A'$ is not a point of condensation of intervals U_i, this follows from the property 3 of u. If $x \in A'$ is a condensation point of U_i, we use the following estimate:

$$|f^{(q)}(y)| = |h^{(q-1)}(y)| \le (\frac{1}{\beta_i})^q N_q \gamma(\beta_i)\beta_i^k \le N_q \cdot \gamma(\beta_i)\beta_i^{k-q} \quad for \quad y \in U_i.$$

Since for intervals converging to $x \in A'$ their lengths tend to zero and since $\gamma(\xi) \to 0$ as $\xi \to 0$, we obtain that $\lim_{y\to x} f^{(q)}(y) = 0, \quad 1 \le q \le k$.

Thus we proved that $f \in C^k[a,b]$ and also that all the derivatives of f up to order k vanish on A'. Because of the property 1 of u we have in fact $\Sigma(f) = A'$.

To prove that $\Delta(f) = f(\Sigma(f)) = A$, we show that $f(x) = \phi(x)$ for each $x \in A'$. Indeed, $A = \phi(A') = [\phi(a), \phi(b)] \setminus \cup_{i=1}^{\infty} V_i$ and $\sum_{i=1}^{\infty} \alpha_i = \phi(b) - \phi(a)$. Hence for each $x \in A'$, $\phi(x) = \phi(a) + \Sigma_{U_i \subset [a,x]} \alpha_i$. But

$$f(x) = \phi(a) + \int_a^x h(t)dt = \phi(a) + \sum_{U_i \subset [a,x]} \frac{\alpha_i}{\beta_i} \int_{U_i} u(\frac{t - a_i}{\beta_i})dt$$

$$= \phi(a) + \sum_{U_i \subset [a,x]} \alpha_i \int_0^1 u(t)dt = \phi(a) + \sum_{U_i \subset [a,x]} \alpha_i = \phi(x),$$

by property 2 of u.

The proof of sufficiency for the case of infinitely differentiable functions is the same, but instead of Proposition 2.6 we use Proposition 2.7. This completes the proof.

In the case of several variables we consider the mappings $f : M^n \to \mathbb{R}^p$ and the critical points of a rank zero: $\Sigma(f) = \{x \in M, df(x) = 0\}$ and, as above, $\Delta(f) = f(\Sigma(f)) \subset \mathbb{R}^p$.

Theorem 3.2. *Let $f : M^n \to \mathbb{R}^p$ be a k times differentiable mapping, $k = 1, 2, 3$, M compact. Then $V_{n/k}(\Delta(f)) < \infty$. Also for $p = 1$ and any k, $V_{n/k}(\Delta(f)) < \infty$.*

Proof: Without loss of generality we can assume that $M = B_1^n$, the ball of a radius 1 in $\mathbb{R}^n$. Denote by D_k the $\max_{y \in B_1^n} \|d^k f(y)\|$.

Lemma 3.2. *For $k = 1, 2, 3$, and for any two critical points $x_1, x_2 \in \Sigma(f)$ we have $\|f(x_1) - f(x_2)\| \leq p \cdot D_k \|x_1 - x_2\|^k$.*

Proof: We apply Lemma 4.2 to the restriction of f to the straight line, passing through x_1 and x_2 in $\mathbb{R}^n$. The two critical points of this restriction, which are required in Lemma 4.2 (for $k \leq 3$), are x_1 and x_2 themselves. This completes the proof.

Now by Lemma 3.2 the mapping $f : \Sigma(f) \to \Delta(f)$ satisfies the conditions of p. 7 of Proposition 2.1. Since $\Sigma(f) \subset B_1^n$ we have $V_n(\Sigma(f)) < \infty$ by Theorem 2.3. Hence by p. 7 of Proposition 2.1 we obtain

$$V_{n/k}(\Delta(f)) \leq (pD_k)^{n/k} V_n(\Sigma(f)) < \infty.$$

For $p = 1$, in a special case $n = 1$, the required result was proved in Theorem 3.1. For functions of several variables it follows from a deep result of L.D. Ivanov ([26], Chapter VI, Section 3, Theorem 2).

Now let us define $\Sigma_q(f)$ to be the set of all $x \in M$, where all the derivatives of f up to the order q vanish, and $\Delta_q(f) = f(\Sigma_q(f))$.

Theorem 3.3. *Let $f : M^n \to \mathbb{R}^p$ be a k times differentiable mapping, M compact, and $q \geq [\frac{k}{2}]$. Then $V_{n/k}(\Delta_q(f)) < \infty$.*

Proof: We use the following lemma:

Lemma 3.3. *For $q \geq [\frac{k}{2}]$, and for any two "strongly critical" points $x_1, x_2 \in \Sigma_q(f)$, $\|f(x_1) - f(x_2)\| \leq p \cdot D_k \|x_1 - x_2\|^k$.*

Proof: We apply Lemma 3.1 to the restriction of f to the straight line, passing through x_1 and x_2 in $\mathbb{R}^n$. The two critical points x_1 and x_2 of this restriction, counted with their multiplicities, provide $k-1$ critical points which are required in Lemma 3.1. This completes the proof of the lemma. The rest of the proof of Theorem 3.3 is the same as in Theorem 3.2.

Finally we give an upper bound for the β-spread of critical values for the case of $C^{k+\alpha}$ smooth mappings. Let $f : B^n_r \to \mathbb{R}^p$ be a k times continuously differentiable mapping, B^n_r being a closed ball of radius r in $\mathbb{R}^n$. We assume that the k derivative $f^{(k)}$ satisfies the Hölder condition $\|f^{(k)}(x) - f^{(k)}(y)\| \leq K\|x-y\|^\alpha$. We denote $k+\alpha$ by l and put $R_l(f) = \frac{K}{k!} r^l$. As above, denote by $\Sigma(f) = \{x \in B^n_r, df(x) = 0\}$ the set of critical points of f of rank zero, $\Delta(f) = f(\Sigma(f))$.

Theorem 3.4. *For each $\beta = (1+\delta)\frac{l}{n}$, $\delta > 0$, we have*

$$V_\beta(\Delta(f)) \leq R^\beta_l(f) C^{1+\delta} \zeta(1+\delta).$$

Here $C = C(l, m, n)$ is the constant depending only on l, m, n and $\zeta(x)$ is the ζ-function $\zeta(x) = \sum_{j=1}^\infty (\frac{1}{j})^x$.

Proof: Let us recall that $M(\epsilon, A)$ denotes the covering number of A, while $\epsilon_A(p)$ is defined as the smallest ϵ for which $M(\epsilon, A) \leq p$. We use the inequality of Proposition 2.3: for each A and for each $\beta > 0$ we have $V_\beta(A) \leq \sum_{p=2}^\infty \epsilon^\beta_A(p)$. Hence $V_\beta(\Delta) \leq \sum_{p=2}^\infty \epsilon^\beta_\Delta(p)$.

Now we use the result of [50], Theorem 9.2 (see also [43], Theorem 1.1):

$$M(\epsilon, \Delta) \leq C = C(l, m, n)$$

for $\epsilon \geq R_l(f)$ and

$$M(\epsilon, \Delta) \leq C \left(\frac{R_l(f)}{\epsilon} \right)^{\frac{n}{l}}$$

for $\epsilon \leq R_l(f)$.

This implies

$$\epsilon_\Delta(p) \leq R_l(f), \quad p \leq C, \quad \text{and} \quad \epsilon_\Delta(p) \leq R_l(f) \left(\frac{C}{p} \right)^{\frac{l}{n}}, \quad p \geq C.$$

Therefore, for each $\beta = (1+\delta)\frac{l}{n}$, $\delta > 0$, we have

$$V_\beta(\Delta) \leq \sum_{p=2}^\infty \epsilon^\beta_\Delta(p) \leq C R^\beta_l(f) + R^\beta_l(f) C^{1+\delta} \sum_{p=C+1}^\infty (\frac{1}{p})^{1+\delta} \leq R^\beta_l(f) C^{1+\delta} \zeta(1+\delta).$$

This completes the proof of Theorem 3.4.

References

[1] V.I. Bakhtin, Properties of the entropy and the Hausdorff dimensions, *Vestnik Mosk. Gos. Univ.*, Ser. 1, no. 5, 1982, 25–29.

[2] S. Bates, On the image size of singular maps. I. *Proc. Amer. Math. Soc.* **114** (1992), no. 3, 699–705.

[3] S. Bates, On the image size of singular maps. II. *Duke Math. J.* **68** (1992), no. 3, 463–476.

[4] S. Bates, Toward a precise smoothness hypothesis in Sard's theorem. Proc. Amer. Math. Soc. 117 (1993), no. 1, 279–283.

[5] S. Bates, C. Moreira, De nouvelles perspectives sur le théorème de Morse-Sard. (French) [Some new perspectives on the Morse-Sard theorem] *C.R. Acad. Sci. Paris Sér. I Math.* **332** (2001), no. 1, 13–17.

[6] S. Bates, A. Norton, On sets of critical values in the real line. Duke Math. J. 83 (1996), no. 2, 399–413.

[7] J. Beck, J. Spencer, Unit distances, *J. Combin. Theory* Ser. A **37** (1984), no. 3, 231–238.

[8] J. Beck, Uniformity and irregularity, *Proceedings of the International Congress of Mathematicians*, Vol. 1, 2 (Berkeley, Calif., 1986), 1400–1407, Amer. Math. Soc., Providence, RI, 1987.

[9] J. Beck, Irregularities of distribution. I, *Acta Math.* **159** (1987), no. 1-2, 1–49.

[10] J. Beck, Irregularities of distribution. II, *Proc. London Math. Soc.* (3) *56* (1988), no. 1, 1–50.

[11] J. Beck, W. Chen, Irregularities of point distribution relative to half-planes I, *Mathematika* **40** (1993), no. 1, 102–126.

[12] J. Beck, W. Chen, Irregularities of point distribution relative to convex polygons II, *Mathematika* **40** (1993), no. 1, 127–136.

[13] J. Beck, W. Chen, Irregularities of point distribution relative to convex polygons III, *J. London Math. Soc.* (2) **56** (1997), no. 2, 222–230.

[14] A.S. Besicovitch, SA. J. Taylor, on the complementary intervals of a linear closed set of zero Lebesgue measure, *J. London Math. Soc.* **29** (1954), 449–459.

[15] G. Bouligand, Ensembles impropres et nombre dimensionnel, *Bull. Sci. Math.* (2) **52** (1928), 320–344 and 361–376.

[16] L. Carleson, Selected Problems on Exceptional Sets, Van Nostrand Math. Studies, **13**, 1967.

[17] K. Falconer, Techniques in fractal geometry, John Wiley and Sons, Ltd., Chichester, 1997. xviii+256 pp.

[18] K. Falconer, Fractal geometry. Mathematical foundations and applications, Second edition. John Wiley and Sons, Inc., Hoboken, NJ, 2003. xxviii+337 pp.

[19] K. Falconer, M. Jarvenpaa, Packing dimensions of sections of sets, *Math. Proc. Cambridge Philos. Soc.* **125** (1999), no. 1, 89–104.

[20] H. Federer, Geometric Measure Theory, Die Grundlehren der Math. Wissenschafften, Vol. 153, Springer-Verlag, 1969.

[21] L.T. Fejes-Toth, Über einen geometrischen Satz, *Mathematische Zeitschrift* **46** (1940), 83–85.

[22] J. Goodman, R. Pollack, B. Sturmfels, The intrinsic spread of a configuration in R^d. *J. Amer. Math. Soc.* **3** (1990), no. 3, 639–651.

[23] J. Harrison, A. Norton, Geometric integration on fractal curves in the plane. *Indiana Univ. Math. J.* **40** (1991), no. 2, 567–594.

[24] J. Harrison, A. Norton, The Gauss-Green theorem for fractal boundaries. Duke Math. J. 67 (1992), no. 3, 575–588.

[25] H. Minkowski, Theorie der konvexen Körper, insbesondere Begründung ihres Oberflächenbegriffs, in: Gesammelte Abhandlungen von Hermann Minkowski (part II, Chapter XXV), Chelsea, New York, 1967, 131–229.

[26] L.D. Ivanov, Variazii mnozestv i funkzii, Moscow, Nauka 1975.

[27] A.N. Kolmogorov, Asymptotic characteristics of some completely bounded metric spaces, *Dokl. Akad. Nauk USSR* **108**, 1956, 585–589.

[28] A.N. Kolmogorov, V.M. Tihomirov, ϵ-entropy and ϵ-capacity of sets in functional spaces, *Uspehi Mat. Nauk* **2** (86), 14, 1959, 3–86. English transl., Amer. Math. Soc. Transl. (2) **17**, 1961, 277–364.

[29] G. Kozma, Z. Lotker, G. Stupp, The minimal spanning tree and the upper box dimension, *Proc. Amer. Math. Soc.* **134** (2006), no. 4, 1183–1187.

[30] M. Lapidus, M. van Frankenhuysen, Fractal Geometry and Number Theory, Birkhäuser, 2000.

[31] G.G. Lorentz, Metric entropy and approximation, *Bull. Amer. Math. Soc.* **72**, 1966, 903–937.

[32] S.C. Milne, Peano curves and smoothness of functions, *Adv. in Math.* **35**, no.2, 1980, 129–157.

[33] A. Norton, A critical set with nonnull image has large Hausdorff dimension, *Trans. Amer. Math. Soc.* **296** (1986), no. 1, 367–376.

[34] A. Norton, A $C^{1\times\infty}$ function with an interval of critical values *Indiana Univ. Math. J.* **40** (1991), no. 4, 1483–1488.

[35] A. Norton, The Zygmund Morse-Sard theorem, *J. Geom. Anal.* **4** (1994), no. 3, 403–424.

[36] A. Norton, C. Pugh, Critical sets in the plane, *Michigan Math. J.* **38** (1991), no. 3, 441–459.

[37] R. Rumely, Capacity Theory on Algebraic Curves, *Lecture Notes in Mathematics*, **1378**, Springer-Verlag, 1989. i+437 pp.

[38] A. Sard, The measure of the critical values of differentiable maps, *Bull. Amer. Math. Soc.* **48**, 1942, 883–890.

[39] A. Sard, Hausdorff measure of critical images on Banach manifolds, *Amer. J. of Math.*, **87**, 1965, 158–174.

[40] C. Tricot, Formule intégrale pour la dimension fractale dans le plan. (French) [Integral formulation for the fractal dimension in the plane] *C.R. Acad. Sci. Paris Sér. I Math.* **333** (2001), no. 3, 169–172.

[41] C. Tricot, Curves and fractal dimension. With a foreword by Michel Mendes France. Translated from the 1993 French original. Springer-Verlag, New York, 1995. xiv+323 pp.

[42] A.G. Vitushkin, On representation of functions by means of superpositions and related topics, *L'Enseignement Math.* XXIII, 1977, 255–320.

[43] Y. Yomdin, The geometry of critical and near-critical values of differentiable mappings. *Math. Ann.* **264**, (1983), n. 4. 495–515.

[44] Y. Yomdin, β-spread of sets in metric spaces and critical values of smooth functions, Preprint, MPI Bonn, (1983).

[45] Y. Yomdin, Global bounds for the Betti numbers of regular fibers of differentiable mappings. *Topology* **24** , (1985), n. 2, 145–152.

[46] Y. Yomdin, Sard's theorem and its improved versions in numerical analysis. *Proc. A.M.S. Conf., Colorado, 1988. Lectures in Applied Math.* **29**, (1990), 701–706.

[47] Y. Yomdin, β-spread of sets in metric spaces and capacity, in preparation.

[48] Y. Yomdin, Discrete subsets of semialgebraic sets, in preparation.

[49] Y. Yomdin, Remez-type inequalities for polynomials bounded on discrete sets, submitted.

[50] Y. Yomdin, G. Comte, Tame Geometry with Applications in Smooth Analysis, *Lecture Notes in Mathematics*, **1834**, Springer, Berlin, Heidelberg, New York, 2004.

Yosef Yomdin
Department of Mathematics
The Weizmann Institute
Rehovot 76100, Israel
e-mail: yosef.yomdin@weizmann.ac.il