BIRKHÄUSER

Operator Theory
Advances and Applications
Vol. 207

# A Glimpse at Hilbert Space Operators

## Paul R. Halmos in Memoriam

Sheldon Axler
Peter Rosenthal
Donald Sarason
Editors

BIRKHÄUSER

Operator Theory: Advances and Applications

Vol. 207

**Founded in 1979 by Israel Gohberg**

# A Glimpse at Hilbert Space Operators

## Paul R. Halmos in Memoriam

Sheldon Axler
Peter Rosenthal
Donald Sarason
Editors

Birkhäuser

2010 Mathematics Subject Classification: Primary 47A, 47B; secondary 46

Library of Congress Control Number: 2010927369

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 978-3-0346-0346-1

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

ISBN 978-3-0346-0346-1
e-ISBN 978-3-0346-0347-8

9 8 7 6 5 4 3 2 1

www.birkhauser.ch

# Contents

# Preface

Paul Richard Halmos, who lived a life of unbounded devotion to mathematics and to the mathematical community, died at the age of 90 on October 2, 2006. This volume is a memorial to Paul by operator theorists he inspired.

Paul's initial research, beginning with his 1938 Ph.D. thesis at the University of Illinois under Joseph Doob, was in probability, ergodic theory, and measure theory. A shift occurred in the 1950s when Paul's interest in foundations led him to invent a subject he termed algebraic logic, resulting in a succession of papers on that subject appearing between 1954 and 1961, and the book *Algebraic Logic*, published in 1962.

Paul's first two papers in pure operator theory appeared in 1950. After 1960 Paul's research focused on Hilbert space operators, a subject he viewed as encompassing finite-dimensional linear algebra.

Beyond his research, Paul contributed to mathematics and to its community in manifold ways: as a renowned expositor, as an innovative teacher, as a tireless editor, and through unstinting service to the American Mathematical Society and to the Mathematical Association of America. Much of Paul's influence flowed at a personal level. Paul had a genuine, uncalculating interest in people; he developed an enormous number of friendships over the years, both with mathematicians and with nonmathematicians. Many of his mathematical friends, including the editors of this volume, while absorbing abundant quantities of mathematics at Paul's knee, learned from his advice and his example what it means to be a mathematician.

The first section of this volume contains three tributes to Paul written on the occasion of his death. They elaborate on the brief remarks in the preceding paragraph, and are reproduced here with the kind permission of their authors and their original publishers. The last item in the first section reproduces the late George Piranian's Mathematical Review of Paul's article *How to write mathematics.* A list of Paul's publications comprises the next section, which is followed by a section of photographs of Paul and photographs taken by Paul.

The main and final section consists of a collection of expository articles by prominent operator theorists. From these articles, this generation of operator theorists and future generations will get a glimpse of many aspects of their subject, and of how Paul enriched and advanced it through his fundamental insights and prescient questions.

**Acknowledgement**

We thank the authors of the expository articles for the high quality of their expositions.

We thank Thomas Hempfling, Birkhäuser's Executive Editor for Mathematics, for his constant support and sound advice.

We thank Mary Jennings for volunteering her expertise in assembling the photo section. This was an arduous task that Mary carried out with painstaking care, offering helpful suggestions along the way. (Other photo acknowledgements are in the introduction to the photo section.)

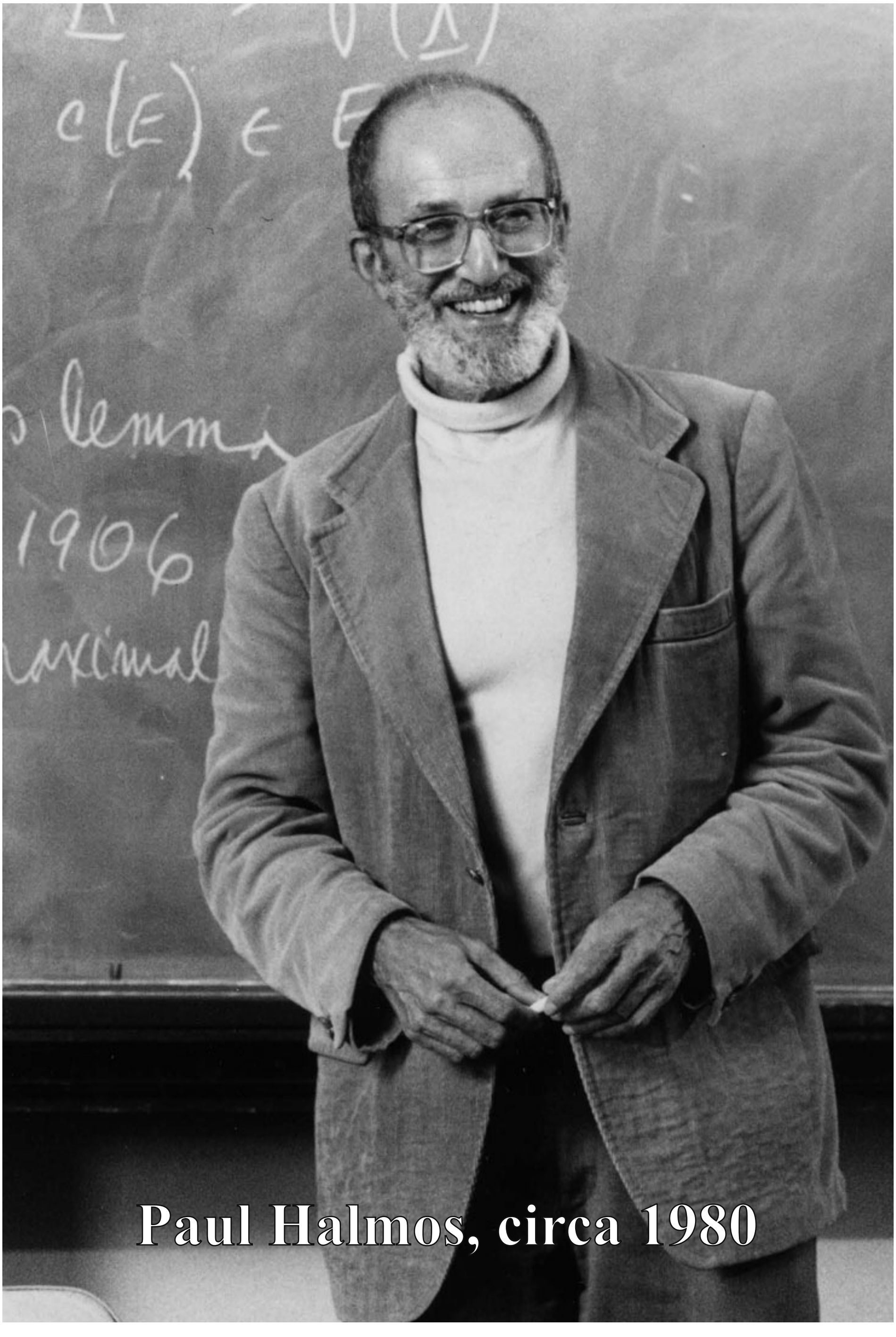
This volume was first proposed by the late Israel Gohberg, the founder and until his recent death the chief editor of the series in which it appears. With Israel's passing, operator theory has lost another of its giants.

Sheldon Axler, Peter Rosenthal, Donald Sarason

December 2009

# Part I

# Paul Halmos



Paul Halmos, circa 1980

Operator Theory:
Advances and Applications, Vol. 207, 3–10

# Paul Halmos – Expositor Par Excellence*

V.S. Sunder

**Abstract.** Paul Richard Halmos, one of the best expositors of mathematics – be it with pen on paper or with chalk on blackboard – passed away on October 2, 2006 after a brief period of illness. This article is an attempt to pay homage to him by recalling some of his contributions to mathematics.

**Mathematics Subject Classification (2000).** 47-xx.

**Keywords.** Halmos, Hilbert space, operator theory, ergodic theory.

## Introduction

Here is what Donald Sarason – arguably the most accomplished PhD student of Halmos – writes about his extraordinary teacher (in [1]):

"Halmos is renowned as an expositor. His writing is some thing he works hard at, thinks intensely about, and is fiercely proud of. (Witness: "How to write mathematics" (see [2]).) In his papers, he is not content merely to present proofs that are well organized and clearly expressed; he also suggests the thought processes that went into the construction of his proofs, pointing out the pitfalls he encountered and indicating helpful analogies. His writings clearly reveal his commitment as an educator. In fact, Halmos is instinctively a teacher, a quality discernible in all his mathematical activities, even the most casual ones.

Most of us, when we discover a new mathematical fact, how ever minor, are usually eager to tell someone about it, to display our cleverness. Halmos behaves differently: he will not tell you his discovery, he will ask you about it, and challenge you to find a proof. If you find a better proof than his, he will be delighted, because then you and he will have taught each other.

To me, Halmos embodies the ideal mixture of researcher and teacher. In him, each role is indistinguishable from the other. Perhaps that is the key to his remarkable influence."

* Reprinted with permission from the February 2007 issue of *Resonance*.

Many of his expository writings, elaborating on his views on diverse topics – writing, lecturing, and doing mathematics – are a 'must read' for every serious student of mathematics. Conveniently, many of them have been collected together in [2].

And here is what one finds in the web pages of the Mathematics Association of America (MAA):

"Professor Halmos was a famed author, editor, teacher, and speaker of distinction. Nearly all of his many books are still in print. His *Finite Dimensional Vector Spaces, Naive Set Theory, Measure Theory, Problems for Mathematicians Young and Old*, and *I Want to be a Mathematician* are classic books that reflect his clarity, conciseness, and color. He edited the American Mathematical Monthly from 1981–1985, and served for many years as one of the editors of the Springer-Verlag series *Undergraduate Texts in Mathematics* and *Graduate Texts in Mathematics*."

While Halmos will be the first to acknowledge that there were far more accomplished mathematicians around him, he would at the same time be the last to be apologetic about what he did. There was the famous story of how, as a young and very junior faculty member at the University of Chicago, he would not let himself be bullied by the very senior faculty member André Weil on a matter of faculty recruitment. His attitude – which functional analysts every where can do well to remember and take strength from – was that while algebraic geometry might be very important, the usefulness of operator theory should not be denied.

Even in his own area of specialisation, there were many mathematicians more powerful than he; but he 'had a nose' for what to ask and which notions to concentrate on. The rest of this article is devoted to trying to justify the assertion of the last line and describing some of the mathematics that Halmos was instrumental in creating. Also, the author has attempted to conform with Halmos' tenet that symbols should, whenever possible, be substituted by words, in order to assist the reader's assimilation of the material. An attempt to write a mathematical article subject to this constraint will convince the reader of the effort Halmos put into his writing!

## The invariant subspace problem

Although Halmos has done some work in probability theory (his PhD thesis was written under the guidance of the celebrated probabilist J.L. Doob), statistics (along with L.J. Savage, he proved an important result on sufficient statistics), ergodic theory and algebraic logic, his preferred area of research (where he eventually 'settled down') was undoubtedly operator theory, more specifically, the study of bounded operators on Hilbert space. Recall that a Hilbert space means a vector space over the field of complex numbers which is equipped with an inner product and is complete with respect to the norm arising from the inner product. Most of his research work revolved around the so-called invariant subspace problem, which asks: Does every bounded (= continuous) linear operator on a Hilbert space (of dimension at least 2) admit a non-trivial invariant subspace, meaning: Is there a closed subspace, other than the zero subspace and the whole space (the two extreme trivial ones[1]) which is mapped into itself by the operator? The answer is negative over the field of real numbers (any rotation in the plane yielding a counterexample), and is positive in the finite-dimensional complex case (thanks to complex matrices having complex eigenvalues).

The first progress towards the solution of this problem came when von Neumann showed that if an operator is compact (i.e., if it maps the unit ball into a compact set, or equivalently, if it is uniformly approximable on the unit ball by operators with finite-dimensional range), then it does indeed have a non-trivial invariant subspace. This was later shown, by Aronszajn and Smith, to continue to be true for compact operators over more general Banach, rather than just Hilbert, spaces.

Then Smith asked and Halmos publicized the question of whether an operator whose square is compact had invariant subspaces. It was subsequently shown by Bernstein and Robinson, using methods of 'non-standard analysis', that if some non-zero polynomial in an operator is compact, then it has invariant subspaces. Very shortly after, Halmos came up with an alternative proof of this result, using standard methods of operator theory.

## Quasitriangularity, quasidiagonality and the Weyl-von Neumann-Voiculescu theorem

Attempting to isolate the key idea in the proof of the Aronszajn-Smith theorem, Halmos identified the notion of *quasitriangular* operators. 'Triangular' operators – those which possess an upper triangular matrix with respect to some orthonormal basis – may also be described (since finite-dimensional operators are triangular) as those which admit an increasing sequence of finite-dimensional invariant subspaces whose union is dense in the Hilbert space. Halmos' definition of quasitriangularity amounts to weakening 'invariant' to 'asymptotically invariant' in the previous sentence. An entirely equivalent requirement, as it turns out, is that the operator is of the form 'triangular + compact'; and the search was on for invariant subspaces of quasitriangular operators. This was until a beautiful 'index-theoretic' characterisation of quasitriangularity was obtained by Apostol, Foias and Voiculescu, which had the unexpected consequence that if an operator or its adjoint is not quasitriangular, then it has a non-trivial invariant subspace.

There is a parallel story involving quasidiagonality, also starting with a definition by Halmos and ending with a spectacular theorem of Voiculescu. Recall that in finite-dimensional Hilbert spaces, according to the spectral theorem, selfadjoint operators have diagonal matrices with respect to some orthonormal basis, and two selfadjoint operators are unitarily equivalent precisely when they

[1] This is why the Hilbert space needs to be at least 2-dimensional.

have the same eigenvalues (i.e., diagonal entries in a diagonal form) which occur with the same multiplicity. Thus the spectrum of an operator (i.e., the set $^{(}T) = \{\lambda \in C : (T - \lambda)$ is not invertible$\}$) and the associated spectral multiplicity (the multiplicity of $\lambda$ is the dimension of the nullspace of $T - \lambda$) form a complete set of invariants for unitary equivalence in the class of self-adjoint operators.

In the infinite-dimensional case, Hermann Weyl proved that the so-called 'essential spectrum' of a self-adjoint operator is left unchanged when it is perturbed by a compact operator; (here, the 'essential spectrum' of a self-adjoint operator is the complement, in the spectrum, of 'isolated eigenvalues of finite multiplicity';) while von Neumann showed that the essential spectrum is a complete invariant for 'unitary equivalence modulo compact perturbation' in the class of self-adjoint operators. Thus, if one allows compact perturbations, spectral multiplicity is no longer relevant. It follows that self-adjoint operators are expressible in the form 'diagonal + compact'; von Neumann even proved the strengthening with 'compact' replaced by 'Hilbert-Schmidt'. (Recall that an operator $T$ is said to be a Hilbert-Schmidt operator if $\sum \|Te_n\|^2 \leq \infty$ for some (equivalently every) orthonormal basis $\{e_n\}$ of the Hilbert space.)

Halmos asked if both statements had valid counterparts for normal operators; specifically, does every normal operator admit a decomposition of the form (a) diagonal + Hilbert-Schmidt, and less stringently (b) diagonal + compact. Both questions were shown to have positive answers as a consequence of the brilliant 'noncommutative Weyl von Neumann theorem' due to Voiculescu (about representations of $C^*$-algebras, which specialises in the case of commutative $^*$-algebras to the desired statements about normal operators); however(a) had also been independently settled by I.D. Berg.

## Subnormal operators and unitary dilations

There were two other major contributions to operator theory by Halmos: subnormal operators and unitary dilations. Both were born of his unwavering belief that the secret about general operators lay in their relationship to normal operators. He defined a subnormal operator to be the restriction of a normal operator to an invariant subspace; the most striking example is the unilateral shift. (Recall that the bilateral shift is the clearly unitary, hence normal, operator on the bilateral sequence space $\ell^2(Z) = \{f : Z \to C : \sum_{n \in Z} |f(n)|^2 < \infty\}$ defined by the equation $(Wf)(n) = f(n-1)$. In the previous sentence, if we replace $Z$ by $Z_+$, the analogous equation defines the *unilateral shift* $U$ on the one-sided sequence space $\ell^2(Z_+)$, which is the prototypical isometric operator which is not unitary. It should be clear that $\ell^2(Z_+)$ may be naturally identified with a subspace of $\ell^2(Z)$ which is invariant under $W$, and that the restriction of $W$ to that subspace may be identified with $U$.) Halmos proved that a general subnormal operator exhibits many properties enjoyed by this first example. For instance, he showed that the normal extension of a subnormal operator is unique under a mild (and natural) minimality condition. (The minimal normal extension of the unilateral shift is the bilateral shift.) Halmos also established that the spectrum of a subnormal operator is obtained by 'filling in some holes' in the spectrum of its minimal normal extension. Many years later, Scott Brown fulfilled Halmos' hope by establishing the existence of non-trivial invariant subspaces of a subnormal operator.

More generally than extensions, Halmos also initiated the study of *dilations.* It is best to first digress briefly into operator matrices. The point is that if $T$ is an operator on $\mathcal{H}$, then any direct sum decomposition $\mathcal{H} = \mathcal{H}_1 \oplus \mathcal{H}_2$ leads to an identification

$$T = \begin{bmatrix} T_{11} & T_{12} \\ T_{21} & T_{22} \end{bmatrix},$$

where $T_{ij} : \mathcal{H}_j \to \mathcal{H}_i$; $1 \le i, j \le 2$ are operators which are uniquely determined by the requirement that if the canonical decomposition $\mathcal{H} \ni x = x_1 + x_2$, $x_i \in \mathcal{H}_i$ is written as $x = \left[\begin{smallmatrix} x_1 \\ x_2 \end{smallmatrix}\right]$, then

$$T \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} = \begin{bmatrix} T_{11} & T_{12} \\ T_{21} & T_{22} \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} = \begin{bmatrix} T_{11}x_1 + T_{12}x_2 \\ T_{21}x_1 + T_{22}x_2 \end{bmatrix}.$$

Thus, for instance, the orthogonal projection $P_1$ of $\mathcal{H}$ onto $\mathcal{H}_1$ is given by

$$P_1 = \begin{bmatrix} \mathrm{id}_{\mathcal{H}_1} & 0 \\ 0 & 0 \end{bmatrix}$$

and $T_{11} = P_1 T|_{\mathcal{H}_1}$. It is customary to call $T_{11}$ the *compression* of $T$ to $\mathcal{H}_1$ and to call $T$ a *dilation* of $T_{11}$. (Note that if, and only if, $T_{21} = 0$, then 'compression' and 'dilation' are nothing but 'restriction' and 'extension'.) Halmos wondered, but not for long, if every operator had a normal dilation; he proved that an operator has a unitary dilation if (and only if) it is a contraction (i.e., maps the unit ball of the Hilbert space into itself).

Subsequently, Sz.-Nagy showed that every contraction in fact has a 'power dilation': i.e., if $T$ is a contraction, then there is a unitary operator $U$ on some Hilbert space such that, simultaneously, $U^n$ is a dilation of $T^n$ for every $n \ge 0$. Halmos noticed that this established the equivalence of the following conditions:

- $T$ is a contraction,
- $T^n$ is a contraction, for each $n$,
- $\|p(T)\| \le \sup\{|f(z)| : z \in D$ (i.e., $|z| < 1)\}$,

and asked if the following conditions were equivalent:

- $T$ is similar to a contraction, i.e., there exists an invertible operator $S$ such that $S^{-1}TS$ is a contraction.
- $\sup_n \|T^n\| \le K$.
- $\|p(T)\| \le K \sup\{|f(z)| : z \in D\}$.

This question, as well as generalisations with $D$ replaced by more general domains in $C^n$, had to wait a few decades before they were solved by Gilles Pisier using 'completely bounded maps' and 'operator spaces' which did not even exist in Halmos' time!

His influence on *operator theory* may be gauged by the activity in this area during the period between his two expository papers (which are listed as his 'Technical papers' at the end of this article, and which may both be found in [1]).

He listed 10 open problems in the area in the first paper, and reviewed the progress made, in the second paper. While concluding the latter paper, he wrote:

> *I hope that despite its sins of omission, this survey conveyed the flavor and the extent of progress in the subject during the last decade.*

Likewise, I hope I have been able to convey something of the brilliance of the expositor in Halmos and the excitement and direction he brought to operator theory in the latter half of the last century.

## Ergodic theory

Although the above account mainly discusses Halmos' contributions to operator theory, undoubtedly due to limitations of the author's familiarity with the areas in which Halmos worked, it would be remiss on the author's part to not make at least passing mention of his contributions to ergodic theory.

He wrote the first English book on ergodic theory (the first book on the subject being Hopf's, in German). He made his influence felt in the field through the problems he popularised and the investigations he undertook. For instance, he gave a lot of publicity, through his book, to the question of whether a 'non-singular' transformation of a measure space – i.e., one which preserved the class of sets of measure zero – admitted an equivalent $\sigma$-finite measure which it preserved. This led to the negative answer by Ornstein, subsequent results in the area by Ito, Arnold, and others and culminated in the very satisfying results by Krieger on *orbit equivalence*. Other contributions of his include the consideration of topologies on the set of measure-preserving transformations of a measure space (influenced no doubt by the 'category-theoretic' results obtained by Oxtoby and Ulam about the ampleness of ergodic homeomorphisms among all homeomorphisms of a cube in $n$ dimensions) and initiating the search for square roots (and cube roots, etc.) of an ergodic transformation.

## A brief biography

A brief non-mathematical account of his life follows. For a more complete and eminently readable write-up which serves the same purpose (and much more attractively, with numerous quotes of Halmos which serve to almost bring him to life), the reader is advised to look at the web-site:

`http://scidiv.bcc.ctc.edu/Math/Halmos.html`

Halmos' life was far from 'routine' – starting in Hungary and quickly moving to America. The following paragraph from his autobiographical book [8] contains a very pithy summing up of his pre-America life:

> *My father, a widower, emigrated to America when I, his youngest son, was* 8 *years old. When he got established, he remarried, presented us with two step-sisters, and began to import us: first my two brothers, and later, almost immediately after he became a naturalized citizen, myself. In view of my father's citizenship I became an instant American the moment I arrived, at the age of* 13*.*

The automathography referred to in [8] contains other vignettes where we can see the problems/difficulties the young Halmos faced in coping with an alien language and culture and a periodically unfriendly 'goddam foreigner' attitude.

After a not particularly spectacular period of undergraduate study, he began by studying philosophy and mathematics, hoping to major in the former. Fortunately for the thousands who learnt linear algebra, measure theory and Hilbert space theory through his incomparable books, he fared poorly in the oral comprehensive exam for the masters' degree, and switched to mathematics as a major. It was only later, when he interacted with J.L. Doob that he seems to have become aware of the excitement and attraction of mathematics; and wrote a thesis on *Invariants of Certain Stochastic Transformations: The Mathematical Theory of Gambling Systems.*

After he finished his PhD in 1938, he "*typed* 120 *letters of application, and got two answers: both NO.*" "*The U of I took pity on me and kept me on as an instructor.*" In the middle of that year a fellow graduate student and friend (Warren Ambrose) of Halmos received a fellowship at the Institute for Advanced Study, Princeton. "*That made me mad. I wanted to go, too! So I resigned my teaching job, borrowed $* 1000 *from my father,* [*wrangled*] *an unsupported membership* (= *a seat in the library*) *at the Institute, and moved to Princeton.*"

There, he attended courses, including the one by John von Neumann (*Everybody called him Johnny*) on 'Rings of Operators'. Von Neumann's official assistant who was more interested in Topology, showed von Neumann the notes that Halmos was taking of the course, and Halmos became the official note-taker for the course and subsequently became von Neumann's official assistant. The next year, "*with no official pre-arrangement, I simply tacked up a card on the bulletin board in Fine Hall saying that I would offer a course called 'Elementary theory of matrices', and I proceeded to offer it.*" About a dozen students attended the course, some took notes and these notes were subsequently pruned into what became *Finite-dimensional vector spaces*; and Halmos' career and book-writing skills were off and running.

As a personal aside, this book was this author's first introduction to the charm of abstract mathematics, and which prompted him to go to graduate school at Indiana University to become Halmos' PhD student – his last one as it turned out. This author cannot begin to enumerate all the things he learnt from this supreme teacher, and will forever be in his debt.

As a final personal note, I should mention Virginia, his warm and hospitable wife since 1945. I remember going to their house for lunch once and finding Paul all alone at home; his grumbled explanation: “Ginger has gone cycling to the old folks home, to read to some people there, who are about 5 years younger than her!”. (She was past 70 then.) She still lives at Los Gatos, California. They never had children, but there were always a couple of cats in their house.

It seems appropriate to end with this quote from the man himself: “I’m not a religious man, but it’s almost like being in touch with God when you’re thinking about mathematics.”

## Suggested reading

[1] P.R. Halmos, Selecta – Research Mathematics, Springer-Verlag, New York, 1983.

[2] P.R. Halmos, Selecta – Expository Writing, Springer-Verlag, New York, 1983.

**Technical papers by Halmos**

[3] Ten Problems in Hilbert space, *Bulletin of the AMS*, Vol. 76, No. 5, pp. 887–993, 1970.

[4] Ten Years in Hilbert space, *Integral Eqs. and Operator Theory*, Vol. 2/4, pp. 529–564, 1979.

**Expository articles by Halmos**

[5] How to write Mathematics, *l’Enseignement mathématique*, Vol. XVI, No. 2, pp. 123–152, 1970.

[6] How to write Mathematics, *Notices of the AMS*, Vol. 23, No. 4, pp. 155–158, 1974.

[7] The Teaching of Problem Solving, *Amer. Math. Monthly*, Vol. 82, No. 5, pp. 466–470, 1975.

**Two non-technical books by Halmos**

[8] *I want to be a Mathematician – an Automathography*, Springer-Verlag, New York, 1985.

[9] *I have a Photographic Memory*, AMS, Providence, RI, 1987.

V.S. Sunder
The Institute of Mathematical Sciences
Chennai, 600 113, India
e-mail: sunder@imsc.res.in

Operator Theory:
Advances and Applications, Vol. 207, 11–25
© 2010 Springer Basel AG

# Paul Halmos: In His Own Words*

John Ewing

**Abstract.** Paul Halmos died on October 2, 2006, at the age of 90. After his death, many people wrote about his career and praised both his mathematical and his expository skills. Paul would have complained about that: He often said he could smell great mathematicians, and he himself was not one of them.

But he was wrong. He was a master of mathematics in multiple ways, and he influenced mathematicians and mathematical culture throughout his career. Unlike most other master mathematicians, Paul's legacy was not merely mathematics but rather advice and opinion about mathematical life—writing, publishing, speaking, research, or even thinking about mathematics. Paul wrote about each of these topics with an extraordinary mixture of conviction and humility. Mathematicians paid attention to what he wrote, and they often quoted it (and still do—"every talk ought to have one proof"). They disagreed and frequently wrote rebuttals. They passed along his wisdom to their students, who passed it along to theirs. Paul Halmos's writing affected the professional lives of nearly every mathematician in the latter half of the twentieth century, and it will continue to influence the profession for years to come.

How does one write about great writing? Explanations of great exposition always fall flat, like analyses of great poems or elucidations of famous paintings. Art is best exhibited, not explained.

And so here is a collection of excerpts from the writing of Paul Halmos, giving advice, offering opinions, or merely contemplating life as a mathematician—all in his own words. – J.E.

**Mathematics Subject Classification (2000).** 00A05, 00B10.

**Keywords.** Mathematical exposition.

## On writing

### Excerpts from: "How to write mathematics", *Enseign. Math.* (2) 16 (1970), 123–152.

... I think I can tell someone how to write, but I can't think who would want to listen. The ability to communicate effectively, the power to be intelligible,

* This article first appeared in *Notices of the AMS*, 54 (2007), 1136–1144. Reprinted with permission.

is congenital, I believe, or, in any event, it is so early acquired that by the time someone reads my wisdom on the subject he is likely to be invariant under it. To understand a syllogism is not something you can learn; you are either born with the ability or you are not. In the same way, effective exposition is not a teachable art; some can do it and some cannot. There is no usable recipe for good writing.

Then why go on? A small reason is the hope that what I said isn't quite right; and, anyway, I'd like a chance to try to do what perhaps cannot be done. A more practical reason is that in the other arts that require innate talent, even the gifted ones who are born with it are not usually born with full knowledge of all the tricks of the trade. A few essays such as this may serve to "remind" (in the sense of Plato) the ones who want to be and are destined to be the expositors of the future of the techniques found useful by the expositors of the past.

The basic problem in writing mathematics is the same as in writing biology, writing a novel, or writing directions for assembling a harpsichord: the problem is to communicate an idea. To do so, and to do it clearly, you must have something to say, and you must have someone to say it to, you must organize what you want to say, and you must arrange it in the order you want it said in, you must write it, rewrite it, and re-rewrite it several times, and you must be willing to think hard about and work hard on mechanical details such as diction, notation, and punctuation. That's all there is to it. . .

It might seem unnecessary to insist that in order to say something well you must have something to say, but it's no joke. Much bad writing, mathematical and otherwise, is caused by a violation of that first principle. Just as there are two ways for a sequence not to have a limit (no cluster points or too many), there are two ways for a piece of writing not to have a subject (no ideas or too many).

The first disease is the harder one to catch. It is hard to write many words about nothing, especially in mathematics, but it can be done, and the result is bound to be hard to read. There is a classic crank book by Carl Theodore Heisel [*The Circle Squared Beyond Refutation*, Heisel, Cleveland, 1934] that serves as an example. It is full of correctly spelled words strung together in grammatical sentences, but after three decades of looking at it every now and then I still cannot read two consecutive pages and make a one-paragraph abstract of what they say; the reason is, I think, that they don't say anything.

The second disease is very common: there are many books that violate the principle of having something to say by trying to say too many things. . .

The second principle of good writing is to write for someone. When you decide to write something, ask yourself who it is that you want to reach. Are you writing a diary note to be read by yourself only, a letter to a friend, a research announcement for specialists, or a textbook for undergraduates? The problems are much the same in any case; what varies is the amount of motivation you need to put in, the extent of informality you may allow yourself, the fussiness of the detail that is necessary, and the number of times things have to be repeated. All writing is influenced by the audience, but, given the audience, the author's problem is to communicate with it as best he can. . .

Everything I've said so far has to do with writing in the large, global sense; it is time to turn to the local aspects of the subject.

The English language can be a beautiful and powerful instrument for interesting, clear, and completely precise information, and I have faith that the same is true for French or Japanese or Russian. It is just as important for an expositor to familiarize himself with that instrument as for a surgeon to know his tools. Euclid can be explained in bad grammar and bad diction, and a vermiform appendix can be removed with a rusty pocket knife, but the victim, even if he is unconscious of the reason for his discomfort, would surely prefer better treatment than that...

My advice about the use of words can be summed up as follows. (1) Avoid technical terms, and especially the creation of new ones, whenever possible. (2) Think hard about the new ones that you must create; consult Roget; and make them as appropriate as possible. (3) Use the old ones correctly and consistently, but with a minimum of obtrusive pedantry...

Everything said about words, applies, mutatis mutandis, to the even smaller units of mathematical writing, the mathematical symbols. The best notation is no notation; whenever possible to avoid the use of a complicated alphabetic apparatus, avoid it. A good attitude to the preparation of written mathematical exposition is to pretend that it is spoken. Pretend that you are explaining the subject to a friend on a long walk in the woods, with no paper available; fall back on symbolism only when it is really necessary.

## On speaking

**Excerpts from: "How to talk mathematics", *Notices of AMS* 21 (1974), 155–158.**

What is the purpose of a public lecture? Answer: to attract and to inform. We like what we do, and we should like for others to like it too; and we believe that the subject's intrinsic qualities are good enough so that anyone who knows what they are cannot help being attracted to them. Hence, better answer: the purpose of a public lecture is to inform, but to do so in a manner that makes it possible for the audience to absorb the information. An attractive presentation with no content is worthless, to be sure, but a lump of indigestible information is worth no more...

*Less is more*, said the great architect Mies van der Rohe, and if all lecturers remember that adage, all audiences would be both wiser and happier.

Have you ever disliked a lecture because it was too elementary? I am sure that there *are* people who would answer yes to that question, but not many. Every time I have asked the question, the person who answered said no, and then looked a little surprised at hearing the answer. A public lecture should be simple and elementary; it should not be complicated and technical. If you believe and can act on this injunction ("be simple"), you can stop reading here; the rest of what I have to say is, in comparison, just a matter of minor detail.

To begin a public lecture to 500 people with "Consider a sheaf of germs of holomorphic functions..." (I have heard it happen) loses people and antagonizes them. If you mention the Künneth formula, it does no harm to say that, at least as far as Betti numbers go, it is just what happens when you multiply polynomials. If you mention functors, say that a typical example is the formation of the duals of vector spaces and the adjoints of linear transformations.

Be simple by being concrete. Listeners are prepared to accept unstated (but hinted) generalizations much more than they are able, on the spur of the moment, to decode a precisely stated abstraction and to re-invent the special cases that motivated it in the first place. Caution: being concrete should not lead to concentrating on the trees and missing the woods. In many parts of mathematics a generalization is simpler and more incisive than its special parent. (Examples: Artin's solution of Hilbert's 17th problem about definite forms via formally real fields; Gelfand's proof of Wiener's theorem about absolutely convergent Fourier series via Banach algebras.) In such cases there is always a concrete special case that is simpler than the seminal one and that illustrates the generalization with less fuss; the lecturer who knows his subject will explain the complicated special case, and the generalization, by discussing the simple cousin.

Some lecturers defend complications and technicalities by saying that that's what *their* subject is like, and there is nothing they can do about it. I am skeptical, and I am willing to go so far as to say that such statements indicate incomplete understanding of the subject and of its place in mathematics. Every subject, and even every small part of a subject, if it is identifiable, if it is big enough to give an hour talk on, has its simple aspects, and they, the simple aspects, the roots of the subject, the connections with more widely known and older parts of mathematics, are what a non-specialized audience needs to be told.

Many lecturers, especially those near the foot of the academic ladder, anxious to climb rapidly, feel under pressure to say something brand new—to impress their elders with their brilliance and profundity. Two comments: (1) the best way to do that is to make the talk simple, and (2) it doesn't really have to be done. It may be entirely appropriate to make the lecturer's recent research the focal point of the lecture, but it may also be entirely appropriate not to do so. An audience's evaluation of the merits of a talk is not proportional to the amount of original material included; the explanation of the speaker's latest theorem may fail to improve his chance of creating a good impression.

An oft-quoted compromise between trying to be intelligible and trying to seem deep is this advice: address the first quarter of your talk to your high-school chemistry teacher, the second to a graduate student, the third to an educated mathematician whose interests are different from yours, and the last to the specialists. I have done my duty by reporting the formula, but I'd fail in my duty if I didn't warn that there are many who do not agree with it. A good public lecture should be a work of art. It should be an architectural unit whose parts reinforce each other in conveying the maximum possible amount of information—not a campaign speech that offers something to everybody, and more likely than not, ends by pleasing nobody. **Make it simple, and you won't go wrong...**

**Excerpt from: *I Want to Be a Mathematician*, p. 401, Springer-Verlag, New York (1985).**

... As for working hard, I got my first hint of what that means when Carmichael told me how long it took him to prepare a fifty-minute invited address. Fifty hours, he said: an hour of work for each minute of the final presentation. When many years later, six of us wrote our "history" paper ("American mathematics from 1940..."), I calculated that my share of the work took about 150 hours; I shudder to think how many manhours the whole group put in. A few of my hours went toward preparing the lecture (as opposed to the paper). I talked it, the whole thing, out loud, and then, I talked it again, the whole thing, into a dictaphone. Then I listened to it, from beginning to end, six times—three times for spots that needed polishing (and which I polished before the next time), and three more times to get the timing right (and, in particular, to get the feel for the timing of each part.) Once all that was behind me, and I had prepared the transparencies, I talked the whole thing through one final rehearsal time (by myself—no audience). That's work...

## On exposition

**Excerpt from: Response from Paul Halmos on winning the Steele Prize for Exposition (1983).**

Not long ago I ran across a reference to a publication titled *A Method of Taking Votes on More Than Two Issues.* Do you know, or could you guess, who the author is? What about an article titled "On automorphisms of compact groups"? Who wrote that one? The answer to the first question is C.L. Dodgson, better known as Lewis Carroll, and the answer to the second question is Paul Halmos.

Lewis Carroll and I have in common that we both called ourselves mathematicians, that we both strove to do research, and that we both took very seriously our attempts to enlarge the known body of mathematical truths. To earn his living, Lewis Carroll was a teacher, and, just for fun, because he loved to tell stories, he wrote *Alice's Adventures in Wonerland.* To earn my living, I've been a teacher for almost fifty years, and, just for fun, because I love to organize and clarify, I wrote *Finite Dimensional Vector Spaces.* And what's the outcome? I doubt if as many as a dozen readers of these words have ever looked at either *A Method of Taking Votes...* or "*On automorphisms...*" but Lewis Carroll is immortal for the Alice stories, and I got the Steele Prize for exposition. I don't know what the Reverend Mr. C.L. Dodgson thought about his fame, but, as for me, I was brought up with the Puritan ethic: if something is fun, then you shouldn't get recognized and rewarded for doing it. As a result, while, to be sure, I am proud and happy, at the same time I can't help feeling just a little worried and guilty.

I enjoy studying, learning, coming to understand, and then explaining, but it doesn't follow that communicating what I know is always easy; it can be devilishly hard. To explain something you must know not only what to put in, but also what to leave out; you must know when to tell the whole truth and when to get the right idea across by telling a little white fib. The difficulty in exposition is not the style, the choice of words—it is the structure, the organization. The words are important, yes, but the arrangement of the material, the indication of the connections of its parts with each other and with other parts of mathematics, the proper emphasis that shows what's easy and what deserves to be treated with caution—these things are much more important...

## On publishing

**Excerpts from: "Four panel talks on publishing", *American Mathematical Monthly* 82 (1975), 14–17.**

... Let me remind you that most laws (with the exception only of the regulatory statutes that govern traffic and taxes) are negative. Consider, as an example, the Ten Commandments. When Moses came back from Mount Sinai, he told us what to be by telling us, eight out of ten times, what not to do. It may therefore be considered appropriate to say what not to publish. I warn you in advance that all the principles that I was able to distill from interviews and from introspection, and that I'll now tell you about, are a little false. Counterexamples can be found to each one—but as directional guides the principles still serve a useful purpose.

First, then, do not publish fruitless speculations: do not publish polemics and diatribes against a friend's error. Do not publish the detailed working out of a known principle. (Gauss discovered exactly which regular polygons are ruler-and-compass constructible, and he proved, in particular, that the one with 65537 sides—a Fermat prime—is constructible; please do not publish the details of the procedure. It's been tried.)

Do not publish in 1975 the case of dimension 2 of an interesting conjecture in algebraic geometry, one that you don't know how to settle in general, and then follow it by dimension 3 in 1976, dimension 4 in 1977, and so on, with dimension $k - 3$ in 197k. Do not, more generally, publish your failures: I tried to prove so-and-so; I couldn't; here it is—see?!

Adrian Albert used to say that a theory is worth studying if it has at least three distinct good hard examples. Do not therefore define and study a new class of functions, the ones that possess left upper bimeasurably approximate derivatives, unless you can, at the very least, fulfill the good graduate student's immediate request: show me some that do and show me some that don't.

A striking criterion for how to decide not to publish something was offered by my colleague John Conway. Suppose that you have just finished typing a paper.

Suppose now that I come to you, horns, cloven hooves, forked tail and all, and ask: if I gave you $1,000,000, would you tear the paper up and forget it? If you hesitate, your paper is lost—do not publish it. That's part of a more general rule: when in doubt, let the answer be no. . .

## On research

**Excerpt from: *I Want to Be a Mathematician*, pp. 321–322, Springer-Verlag, New York (1985).**

Can anyone tell anyone else how to do research, how to be creative, how to discover something new? Almost certainly not. I have been trying for a long time to learn mathematics, to understand it, to find the truth, to prove a theorem, to solve a problem—and now I am going to try to describe just how I went about it. The important part of the process is mental, and that is indescribable—but I can at least take a stab at the physical part.

Mathematics is not a deductive science—that's a cliché. When you try to prove a theorem, you don't just list the hypotheses, and then start to reason. What you do is trial and error, experimentation, guesswork. You want to find out what the facts are, and what you do is in that respect similar to what a laboratory technician does, but it is different in the degree of precision and information. Possibly philosophers would look on us mathematicians the same way we look on the technicians, if they dared.

I love to do research, I want to do research, I have to do research, and I hate to sit down and begin to do research—I always try to put it off just as long as I can.

It is important to me to have something big and external, not inside myself, that I can devote my life to. Gauss and Goya and Shakespeare and Paganini are excellent, their excellence gives me pleasure, and I admire and envy them. They were also dedicated human beings. Excellence is for the few but dedication is something everybody can have—and should have—and without it life is not worth living.

Despite my great emotional involvement in work, I just hate to start doing it; it's a battle and a wrench every time. Isn't there something I can (must?) do first? Shouldn't I sharpen my pencils, perhaps? In fact I never use pencils, but pencil sharpening has become the code phrase for anything that helps to postpone the pain of concentrated creative attention. It stands for reference searching in the library, systematizing old notes, or even preparing tomorrow's class lecture, with the excuse that once those things are out of the way I'll really be able to concentrate without interruption.

When Carmichael complained that as dean he didn't have more than 20 hours a week for research I marvelled, and I marvel still. During my productive years I probably averaged 20 hours of concentrated mathematical thinking a week, but much more than that was extremely rare. The rare exception came, two or three times in my life, when long ladders of thought were approaching their climax. Even though I never was dean of a graduate school, I seemed to have psychic energy for only three or four hours of work, "real work", each day; the rest of the time I wrote, taught, reviewed, conferred, refereed, lectured, edited, travelled, and generally sharpened pencils all the ways I could think of. Everybody who does research runs into fallow periods. During mine the other professional activities, down to and including teaching trigonometry, served as a sort of excuse for living. Yes, yes. I may not have proved any new theorems today, but at least I explained the law of sines pretty well, and I have earned my keep.

Why do mathematicians do research? There are several answers. The one I like best is that we are curious—we need to know. That is almost the same as "because we want to," and I accept that—that's a good answer too. There are, however, more answers, ones that are more practical.

## On teaching

**Excerpt from: "The problem of learning to teach", *American Mathematical Monthly* 82 (1975), 466–476.**

The best way to learn is to do; the worst way to teach is to talk.

About the latter: did you ever notice that some of the best teachers of the world are the worst lecturers? (I can prove that, but I'd rather not lose quite so many friends.) And, the other way around, did you ever notice that good lecturers are not necessarily good teachers? A good lecture is usually systematic, complete, precise—and dull; it is a bad teaching instrument. When given by such legendary outstanding speakers as Emil Artin and John von Neumann, even a lecture can be a useful tool—their charisma and enthusiasm come through enough to inspire the listener to go forth and do something—it looks like such fun. For most ordinary mortals, however, who are not so bad at lecturing as Wiener was—not so stimulating!—and not so good as Artin—and not so dramatic!—the lecture is an instrument of last resort for good teaching.

My test for what makes a good teacher is very simple: it is the pragmatic one of judging the performance by the product. If a teacher of graduate students consistently produces Ph.D.'s who are mathematicians and who create high-quality new mathematics, he is a good teacher. If a teacher of calculus consistently produces seniors who turn into outstanding graduate students of mathematics, or into leading engineers, biologists, or economists, he is a good teacher. If a teacher of third grade "new math" (or old) consistently produces outstanding calculus students, or grocery store check-out clerks, or carpenters, or automobile mechanics, he is a good teacher.

For a student of mathematics to hear someone talk about mathematics does hardly any more good than for a student of swimming to hear someone talk about swimming. You can't learn swimming techniques by having someone tell you where to put your arms and legs; and you can't learn to solve problems by having someone tell you to complete the square or to substitute $\sin u$ for $y$.

Can one learn mathematics by reading it? I am inclined to say no. Reading has an edge over listening because reading is more active—but not much. Reading with pencil and paper on the side is very much better—it is a big step in the right direction. The very best way to read a book, however, with, to be sure, pencil and paper on the side, is to keep the pencil busy on the paper and throw the book away.

Having stated this extreme position, I'll rescind it immediately. I know that it is extreme, and I don't really mean it—but I wanted to be very emphatic about not going along with the view that learning means going to lectures and reading books. If we had longer lives, and bigger brains, and enough dedicated expert teachers to have a student/teacher ratio of 1/1, I'd stick with the extreme views—but we don't. Books and lectures don't do a good job of transplanting the facts and techniques of the past into the bloodstream of the scientist of the future—but we must put up with a second best job in order to save time and money. But, and this is the text of my sermon today, if we rely on lectures and books only, we are doing our students and their students, a grave disservice...

**Excerpt from: "The heart of mathematics", *American Mathematical Monthly* 87 (1980), 519–524.**

... How can we, the teachers of today, use the problem literature? Our assigned task is to pass on the torch of mathematical knowledge to the technicians, engineers, scientists, humanists, teachers, and, not least, research mathematicians of tomorrow: do problems help?

Yes, they do. The major part of every meaningful life is the solution of problems; a considerable part of the professional life of technicians, engineers, scientists, etc., is the solution of mathematical problems. It is the duty of all teachers, and of teachers of mathematics in particular, to expose their students to problems much more than to facts. It is, perhaps, more satisfying to stride into a classroom and give a polished lecture on the Weierstrass $M$-test than to conduct a fumble-and-blunder session that *ends* in the question: "Is the boundedness assumption of the test necessary for its conclusion?" I maintain, however, that such a fumble session, intended to motivate the student to search for a counterexample, is infinitely more valuable.

I have taught courses whose entire content was problems solved by students (and then presented to the class). The number of theorems that the students in such a course were exposed to was approximately half the number that they could have been exposed to in a series of lectures. In a problem course, however, exposure means the acquiring of an intelligent questioning attitude and of some technique for plugging the leaks that proofs are likely to spring; in a lecture course, exposure sometimes means not much more than learning the name of a theorem, being intimidated by its complicated proof, and worrying about whether it would appear on the examination.

... Many teachers are concerned about the amount of material they must cover in a course. One cynic suggested a formula; since, he said, students on the average remember only about 40% of what you tell them, the thing to do is to cram into each course 250% of what you hope will stick. Glib as that is, it probably would not work.

Problem courses do work. Students who have taken my problem courses were often complimented by their subsequent teachers. The compliments were on their alert attitude, on their ability to get to the heart of the matter quickly, and on their intelligently searching questions that showed that they understood what was happening in class. All this happened on more than one level, in calculus, in linear algebra, in set theory, and, of course, in graduate courses on measure theory and functional analysis.

Why must we cover everything that we hope students will ultimately learn? Even if (to stay with an example already mentioned) we think that the Weierstrass $M$-test is supremely important, and that every mathematics student must know that it exists and must understand how to apply it—even then a course on the pertinent branch of analysis might be better for omitting it. Suppose that there are 40 such important topics that a student must be exposed to in a term. Does it follow that we must give 40 complete lectures and hope that they will all sink in? Might it not be better to give 20 of the topics just a ten-minute mention (the name, the statement, and an indication of one of the directions in which it can be applied), and to treat the other 20 in depth, by student-solved problems, student-constructed counterexamples, and student-discovered applications? I firmly believe that the latter method teaches more and teaches better. Some of the material doesn't get *covered* but a lot of it gets *discovered* (a telling old pun that deserves to be kept alive), and the method thereby opens doors whose very existence might never have been suspected behind a solidly built structure of settled facts. As for the Weierstrass $M$-test, or whatever was given short shrift in class—well, books and journals do exist, and students have been known to read them in a pinch...

## On mathematics

### Excerpt from: "Mathematics as a creative art", *American Scientist* 56 (1968), 375–389.

Do you know any mathematicians—and, if you do, do you know anything about what they do with their time? Most people don't. When I get into a conversation with the man next to me in a plane, and he tells me that he is something respectable like a doctor, lawyer, merchant or dean, I am tempted to say that I am in roofing and siding. If I tell him that I am a mathematician, his most likely reply will be that he himself could never balance his check book, and it must be fun to be a whiz at math. If my neighbor is an astronomer, a biologist, a chemist, or any other kind of natural or social scientist, I am, if anything, worse off—this man *thinks* he knows what a mathematician is, and he is probably wrong. He thinks that I spend my time (or should) converting different orders of magnitude, comparing binomial coefficients and powers of 2, or solving equations involving rates of reactions.

C.P. Snow points to and deplores the existence of two cultures; he worries about the physicist whose idea of modern literature is Dickens, and he chides the poet who cannot state the second law of thermodynamics. Mathematicians, in converse with well-meaning, intelligent, and educated laymen (do you mind if I refer to all nonmathematicians as laymen?) are much worse off than physicists in converse with poets. It saddens me that educated people don't even know that my subject exists. There is something that they call mathematics, but they neither know how the professionals use the word, nor can they conceive why anybody should do it. It is, to be sure, possible that an intelligent and otherwise educated person doesn't know that egyptology exists, or haematology, but all you have to tell him is that it does, and he will immediately understand in a rough general way why it should and he will have some empathy with the scholar of the subject who finds it interesting.

Usually when a mathematician lectures, he is a missionary. Whether he is talking over a cup of coffee with a collaborator, lecturing to a graduate class of specialists, teaching a reluctant group of freshman engineers, or addressing a general audience of laymen—he is still preaching and seeking to make converts. He will state theorems and he will discuss proofs and he will hope that when he is done his audience will know more mathematics than they did before. My aim today is different—I am not here to proselytize but to enlighten—I seek not converts but friends. I do not want to teach you what mathematics is, but only *that* it is.

I call my subject mathematics—that's what all my colleagues call it, all over the world—and there, quite possibly, is the beginning of confusion. The word covers two disciplines—many more, in reality, but two, at least two, in the same sense in which Snow speaks of two cultures. In order to have some words with which to refer to the ideas I want to discuss, I offer two temporary and ad hoc neologisms. Mathematics, as the work is customarily used, consists of at least two distinct subjects, and I propose to call them *mathology* and *mathophysics.* Roughly speaking, mathology is what is called pure mathematics, and mathophysics is called applied mathematics, but the qualifiers are not emotionally strong enough to disguise that they qualify the same noun. If the concatenation of syllables I chose here reminds you of other words, no great harm will be done; the rhymes alluded to are not completely accidental. I originally planned to entitle this lecture something like "Mathematics is an art," or "Mathematics is not a science," and "Mathematics is useless," but the more I thought about it the more I realized that I mean that "Mathology is an art," "Mathology is not a science," and "Mathology is useless." When I am through, I hope you will recognize that most of you have known about mathophysics before, only you were probably calling it mathematics; I hope that all of you will recognize the distinction between mathology and mathophysics; and I hope that some of you will be ready to embrace, or at least applaud, or at the very least, recognize mathology as a respectable human endeavor.

In the course of the lecture I'll have to use many analogies (literature, chess, painting), each imperfect by itself, but I hope that in their totality they will serve to delineate what I want delineated. Sometimes in the interest of economy of time, and sometimes doubtless unintentionally, I'll exaggerate; when I'm done, I'll be glad to rescind anything that was inaccurate or that gave offense in any way...

Mathematics is abstract thought, mathematics is pure logic, mathematics is creative art. All these statements are wrong, but they are all a little right, and they are all nearer the mark than "mathematics is numbers" or "mathematics is geometric shapes". For the professional pure mathematician, mathematics is the logical dovetailing of a carefully selected sparse set of assumptions with their surprising conclusions via a conceptually elegant proof. Simplicity, intricacy, and above all, logical analysis are the hallmark of mathematics.

The mathematician is interested in extreme cases—in this respect he is like the industrial experimenter who breaks lightbulbs, tears shirts, and bounces cars on ruts. How widely does a reasoning apply, he wants to know, and what happens when it doesn't? What happens when you weaken one of the assumptions, or under what conditions can you strengthen one of the conclusions? It is the perpetual asking of such questions that makes for broader understanding, better technique, and greater elasticity for future problems.

Mathematics—this may surprise or shock you some—is never deductive in its creation. The mathematician at work makes vague guesses, visualizes broad generalizations, and jumps to unwarranted conclusions. He arranges and rearranges his ideas, and he becomes convinced of their truth long before he can write down a logical proof. The conviction is not likely to come early—it usually comes after many attempts, many failures, many discouragements, many false starts. It often happens that months of work result in the proof that the method of attack they were based on cannot possibly work and the process of guessing, visualizing, and conclusion-jumping begins again. A reformulation is needed and—and this too may surprise you—more experimental work is needed. To be sure, by "experimental work" I do not mean test tubes and cyclotrons. I mean thought-experiments. When a mathematician wants to prove a theorem about an infinite-dimensional Hilbert space, he examines its finite-dimensional analogue, he looks in detail at the 2- and 3-dimensional cases, he often tries out a particular numerical case, and he hopes that he will gain thereby an insight that pure definition-juggling has not yielded. The deductive stage, writing the result down, and writing down its rigorous proof are relatively trivial once the real insight arrives; it is more like the draftsman's work, not the architect's...

The mathematical fraternity is a little like a self-perpetuating priesthood. The mathematicians of today train the mathematicians of tomorrow and, in effect, decide whom to admit to the priesthood. Most people do not find it easy to join—mathematical talent and genius are apparently exactly as rare as talent and genius in paint and music—but anyone can join, everyone is welcome. The rules are nowhere explicitly formulated, but they are intuitively felt by everyone in the profession. Mistakes are forgiven and so is obscure exposition—the indispensable requisite is mathematical insight. Sloppy thinking, verbosity without content, and polemic have no role, and—this is to me one of the most wonderful aspects of mathematics—they are much easier to spot than in the nonmathematical fields of human endeavor (much easier than, for instance, in literature among the arts, in art criticism among the humanities, and in your favorite abomination among the social sciences).

Although most of mathematical creation is done by one man at a desk, at a blackboard, or taking a walk, or, sometimes, by two men in conversation, mathematics is nevertheless a sociable science. The creator needs stimulation while he is creating and he needs an audience after he has created. Mathematics is a sociable science in the sense that I don't think it can be done by one man on a desert island (except for a very short time), but it is not a mob science, it is not a team science. A theorem is not a pyramid; inspiration has never been known to descend on a committee. A great theorem can no more be obtained by a "project" approach than a great painting: I don't think a team of little Gausses could have obtained the theorem about regular polygons under the leadership of a rear admiral anymore than a team of little Shakespeares could have written Hamlet under such conditions…

## On pure and applied

**Excerpt from: "Applied mathematics is bad mathematics", pp. 9–20, appearing in *Mathematics Tomorrow*, edited by Lynn Steen, Springer-Verlag, New York (1981).**

It isn't really (applied mathematics, that is, isn't really bad mathematics), but it's different.

Does that sound as if I had set out to capture your attention, and, having succeeded, decided forthwith to back down and become conciliatory? Nothing of the sort! The "conciliatory" sentence is controversial, believe it or not; lots of people argue, vehemently, that it (meaning applied mathematics) is not different at all, it's all the same as pure mathematics, and anybody who says otherwise is probably a reactionary establishmentarian and certainly wrong.

If you're not a professional mathematician, you may be astonished to learn that (according to some people) there are different kinds of mathematics, and that there is anything in the subject for anyone to get excited about. There are; and there is; and what follows is a fragment of what might be called the pertinent sociology of mathematics: what's the difference between pure and applied, how do mathematicians feel about the rift, and what's likely to happen to it in the centuries to come…

The pure and applied distinction is visible in the arts and in the humanities almost as clearly as in the sciences: witness Mozart versus military marches, Rubens versus medical illustrations, or Virgil's *Aeneid* versus Cicero's *Philippics.* Pure literature deals with abstractions such as love and war, and it tells about imaginary examples of them in emotionally stirring language. Pure mathematics deals with abstractions such as the multiplication of numbers and the congruence of triangles, and it reasons about Platonically idealized examples of them with intellectually convincing logic.

There is, to be sure, one sense of the word in which all literature is “applied”. Shakespeare’s sonnets have to do with the everyday world, and so does Tolstoy’s *War and Peace*, and so do Caesar’s commentaries on the wars he fought; all start from what human beings see and hear, and all speak of how human beings move and feel. In that same somewhat shallow sense all mathematics is applied. It all starts from sizes and shapes (whose study leads ultimately to algebra and geometry), and it reasons about how sizes and shapes change and interact (and such reasoning leads ultimately to the part of the subject that the professionals call analysis).

There can be no doubt that the fountainhead, the inspiration, of all literature is the physical and social universe we live in, and the same is true about mathematics. There is no doubt that the physical and social universe daily affects each musician, and painter, and writer, and mathematician, and that therefore a part at least of the raw material of the artist is the work of facts and motions, sights and sounds. Continual contact between the work and art is bound to change the latter, and perhaps even to improve it.

The ultimate goal of “applied literature”, and of applied mathematics, is action. A campaign speech is made so as to cause you to pull the third lever on a voting machine rather than the fourth. An aerodynamic equation is solved so as to cause a plane wing to lift its load fast enough to avoid complaints from the home owners near the airport. These examples are crude and obvious; there are subtler ones. If the biography of a candidate, a factually correct and honest biography, does not directly mention the forthcoming election, is it then pure literature? If a discussion of how mathematically idealized air flows around moving figures of various shapes, a logically rigorous and correct discussion, does not mention airplanes or airports, is it then pure mathematics? And what about the in-between cases: the biography that, without telling lies, is heavily prejudiced; and the treatise on aerodynamics that, without being demonstrably incorrect, uses cost-cutting rough approximations—are they pure or applied?. . .

To confuse the issue still more, pure mathematics can be practically useful and applied mathematics can be artistically elegant. Pure mathematicians, trying to understand involved logical and geometrical interrelations, discovered the theory of convex sets and the algebraic and topological study of various classes of functions. Almost as if by luck, convexity has become the main tool in linear programming (an indispensable part of modern economic and industrial practice), and functional analysis has become the main tool in quantum theory and particle physics. The physicist regards the applicability of von Neumann algebras (a part of functional analysis) to elementary particles as the only justification of the former; the mathematician regards the connections as the only interesting aspect of the latter. *De gustibus non disputandum est?*

Just as pure mathematics can be useful, applied mathematics can be more beautifully useless than is sometimes recognized. Applied mathematics is not engineering; the applied mathematician does not design airplanes or atomic bombs. Applied mathematics is an intellectual discipline, not a part of industrial technology. The ultimate goal of applied mathematics is action, to be sure, but, before that, applied mathematics is a part of theoretical science concerned with the general principles behind what makes planes fly and bombs explode...

The deepest assertion about the relation between pure and applied mathematics that needs examination is that it is symbiotic, in the sense that neither can survive without the other. Not only, as is universally admitted, does the applied need the pure, but, in order to keep from becoming inbred, sterile, meaningless, and dead, the pure needs the revitalization and the contact with reality that only the applied can provide...

## On being a mathematician

**Excerpt from: *I Want to Be a Mathematician*, p. 400, Springer-Verlag, New York (1985).**

It takes a long time to learn to live—by the time you learn your time is gone. I spent most of a lifetime trying to be a mathematician—and what did I learn? What does it take to be one? I think I know the answer: you have to be born right, you must continually strive to become perfect, you must love mathematics more than anything else.

Born right? Yes. To be a scholar of mathematics you must be born with talent, insight, concentration, taste, luck, drive, and the ability to visualize and guess. For teaching you must in addition understand what kinds of obstacles learners are likely to place before themselves, and you must have sympathy for your audience, dedicated selflessness, verbal ability, clear style, and expository skill. To be able, finally, to pull your weight in the profession with the essential clerical and administrative jobs, you must be responsible, conscientious, careful, and organized—it helps if you also have some qualities of leadership and charisma.

You can't be perfect, but if you don't try, you won't be good enough.

To be a mathematician you must love mathematics more than family, religion, money, comfort, pleasure, glory. I do not mean that you must love it to the exclusion of family, religion, and the rest, and I do not mean that if you do love it, you'll never have any doubts, you'll never be discouraged, you'll never be ready to chuck it all and take up gardening instead. Doubts and discouragements are part of life. Great mathematicians have doubts and get discouraged, but usually they can't stop doing mathematics anyway, and, when they do, they miss it very deeply...

John Ewing
Math for America
160 Fifth Ave, 8th fl
New York, NY 10010, USA
e-mail: `jewing@mathforamerica.org`

Operator Theory:
Advances and Applications, Vol. 207, 27–29
© 2010 Springer Basel AG

# Obituary: Paul Halmos, 1916–2006

Heydar Radjavi and Peter Rosenthal

Paul Halmos, one of the most influential mathematicians of the last half of the twentieth century, died at the age of ninety on October 2, 2006. Paul wrote "To be a mathematician you must love mathematics more than family, religion, money, comfort, pleasure, glory." Paul did love mathematics. He loved thinking about it, talking about it, giving lectures and writing articles and books. Paul also loved language, almost as much as he loved mathematics. That is why his books and expository articles are so wonderful. Paul took Hardy's famous dictum that "there is no permanent place in the world for ugly mathematics" very seriously: he reformulated and polished all the mathematics that he wrote and lectured about, and presented it in the most beautiful way.

Irving Kaplansky, the great Canadian mathematician who also died in 2006 (at the age of eighty-nine), wrote "Paul Halmos is the wittiest person I know." Many quotations from Paul's writing illustrating Kaplansky's statement can be found on the internet (just Google "Paul Halmos quotations"). Here are some:

- *You can't be perfect, but if you don't try, you won't be good enough.*
- *If you have to ask, you shouldn't even ask.*
- *Once the problem is solved, its repetitive application has as much to do with mathematics as the work of a Western Union messenger boy has to do with Marconi's genius.*
- *The criterion for quality is beauty, intricacy, neatness, elegance, satisfaction, appropriateness – all subjective, but somehow mysteriously shared by all.*
- *There is no Berlitz course for the language of mathematics; apparently the only way to learn it is to live with it for years.*
- *The recommendations I have been making are based partly on what I do, more on what I regret not having done, and most on what I wish others had done for me.*
- *Almost everybody's answer to "What to publish?" can be expressed in either one word – "less" – or two words – "good stuff".*

This obituary is reprinted with the kind permission of the Canadian Mathematical Society from Volume 39 issue 2 (March, 2007) of the CMS Notes.

Man-Duen Choi put together a number of titles of Halmos's writings to form a cute narrative – see "A Postscript" on page 799 of volume 103 (1996) of the American Mathematical Monthly.

Paul liked to be provocative. He wrote, for example, "The best way to learn is to do; the worst way to teach is to talk." He did follow this with "Having stated this extreme position, I'll rescind it immediately. I know that it is extreme, and I don't really mean it – but I wanted to be very emphatic about not going along with the view that learning means going to lectures and reading books." However, his explanation did not mollify some people who were very proud of their ability to lecture.

Perhaps Paul's most provocative comment in print (those who had the pleasure of participating in discussions with him heard even more provocative statements) was his title for an article published in 1981: "Applied Mathematics is Bad Mathematics." Although Paul began the article with "It isn't really (applied mathematics, that is, isn't really bad mathematics), but it's different," the title angered many applied mathematicians.

Paul made fundamental contributions to ergodic theory and measure theory (his classic books "Lectures on Ergodic Theory" and "Measure Theory", and many papers) and to algebraic logic (see "Algebraic Logic" and, with S. Givant, "Logic as Algebra"). His book "Naive Set Theory" is a beautiful exposition of axiomatic (Zermelo-Fraenkel) set theory, in spite of its "naive" title. But Paul's most important contributions to research in mathematics, at least from our prejudiced point of view, were to the theory of operators on Hilbert space.

Paul created and led a vigorous school of operator theory. He introduced central concepts such as unitary dilations, subnormal operators and quasitriangularizability, and proved the fundamental theorems about them. These, and other concepts he developed, became major subjects of research; there is now a large body of knowledge about each of these topics.

Paul had extraordinary ability to discover the central questions concerning a large number of different aspects of operator theory. In particular, his famous article "Ten problems in Hilbert space" shaped a great deal of subsequent research in operator theory and in $C^*$-algebras. As Berkeley mathematician Don Sarason (who competes with Errett Bishop for the title "Most-distinguished of Halmos' twenty-one Ph.D. students") wrote, in his introduction to Paul's selected works, "Halmos embodies the ideal mixture of researcher and teacher. In him, each role is indistinguishable from the other. Perhaps that is the key to his remarkable influence."

Paul wrote what he termed "an automathography", a fascinating book entitled "I Want to be a Mathematician." This is a mathematical autobiography, and contains much advice that is very useful to all mathematicians and to all those who aspire to be mathematicians. Towards the end of that book, Paul evaluates his career as follows: "I was, in I think decreasing order of quality, a writer, an editor, a teacher, and a research mathematician."

Paul's self-evaluation may be quite accurate, but it is important to understand how high a standard he was setting for himself. He was certainly as good a mathematical writer as ever existed. As an editor, Paul played a central role in developing several of the series of mathematics books published by Springer-Verlag, as well as in editing several journals. Virtually everyone who ever heard him lecture will testify that his lectures were maximally interesting, clear and inspiring. (Luckily, several videotapes of Paul's lectures can be purchased from the A.M.S. and the M.A.A.). Moreover, Paul's total contribution to research in mathematics is very impressive.

As Paul wrote, "it takes a long time to learn to live – by the time you learn your time is gone." However, one can learn much about living from others, and Paul taught many mathematicians a huge amount, about mathematics and about life. For instance, Paul wrote "I like to start every course I teach with a problem." Those who wish to follow Paul's example in this respect can use Paul's book "Problems for Mathematicians Young and Old"; it contains a surprising variety of beautiful problems from a variety of areas of mathematics.

Paul advised efficiency in all tasks: if a letter has to be answered, or a review has to be written for Mathematical Reviews, do it right away, rather than thinking for months "I'd better get to that." Good advice but, we have to confess, we failed to follow it in the writing of this obituary. It doesn't take a psychoanalyst to figure out why we failed: we wanted to postpone this last goodbye to Paul Halmos. But the time has now come. Goodbye, Paul; thanks very much for so much.

Heydar Radjavi
Department of Mathematics
University of Waterloo
Waterloo
Ontario N2L 3G1, Canada
e-mail: hradjavi@math.uwaterloo.ca

Peter Rosenthal
Department of Mathematics
University of Toronto
Toronto
Ontario M5S 3G3, Canada
e-mail: rosent@math.toronto.edu

Operator Theory:
Advances and Applications, Vol. 207, 31–31

# Mathematical Review of **"How to Write Mathematics"** [*] by P.R. Halmos

G. Piranian

Halmos delivers a vigorous piece of his mind on the craft of writing. His principal message concerns the spiral plan, whose wide adoption would certainly raise the standards of mathematical prose in our books and journals. Lest readers deceive themselves and believe that they need not go to the original paper, I abstain from describing the plan.

In addition to giving counsel on global strategy, Halmos points out many tactical devices by which authors can bring their manuscripts nearer to unobtrusive perfection. He expresses his displeasure over some barbarisms, for example, the hanging theorem, the thoughtless misuse of "given" and "any", and the sloppy construction "if . . . , then . . . if . . . ". He argues against the display of cumbersome symbolic messages that only a machine or a fool would decode, and he looks after many fine points, such as the hyphen in the compound adjective "measure-preserving". He even expresses his contempt for the pedantic copy-editor who in the preceding sentence would insist on putting the second quotation mark after the period.

This review is not a catalogue of what the paper offers. It merely serves notice that a mathematician with an eminently successful personal style has described his technique of writing, and that whoever pays heed will profit.

[*] *Enseignement Math.* (2) **16** 1970 123–152

Operator Theory:
Advances and Applications, Vol. 207, 33–40
© 2010 Springer Basel AG

# Publications of Paul R. Halmos

The publications are listed chronologically, articles and books separately. Not listed are translations of Paul's articles and books, of which there have been many (in French, German, Bulgarian, Russian, Czech, Polish, Finnish, Catalan).

## Research and Expository Articles

(1939-1) *On a necessary condition for the strong law of large numbers*. Ann. of Math. (2) 40 (1939), 800–804.

(1941-1) *Statistics, set functions, and spectra*. Rec. Math. [Mat. Sbornik] N.S. 9 (51) (1941), 241–248.

(1941-2) *The decomposition of measures*. Duke Math. J. 8 (1941), 386–392.

(1942-1) *The decomposition of measures*. II. Duke Math. J. 9 (1942), 43–47 (with W. Ambrose and S. Kakutani).

(1942-2) *Square roots of measure preserving transformations*. Amer. J. Math. 64 (1942), 153–166.

(1942-3) *On monothetic groups*. Proc. Nat. Acad. Sci. U.S.A. 28 (1942), 254–258 (with H. Samelson).

(1942-4) *Operator models in classical mechanics*. II. Ann. of Math. (2) 43 (1942), 332–350 (with J. von Neumann).

(1943-1) *On automorphisms of compact groups*. Bull. Amer. Math. Soc. 49 (1943), 619–624.

(1944-1) *Approximation theories for measure preserving transformations*. Trans. Amer. Math. Soc. 55 (1944), 1–18.

(1944-2) *Random alms*. Ann. Math. Statistics 15 (1944), 182–189.

(1944-3) *The foundations of probability*. Amer. Math. Monthly 51 (1944), 493–510.

(1944-4) *In general a measure preserving transformation is mixing*. Ann. of Math. (2) 45 (1944), 786–792.

(1944-5) *Comment on the real line*. Bull. Amer. Math. Soc. 50 (1944), 877–878.

(1946-1) *The theory of unbiased estimation.* Ann. Math. Statistics 17 (1946), 34–43.

(1946-2) *An ergodic theorem.* Proc. Nat. Acad. Sci. U.S.A. 32 (1946), 156–161.

(1947-1) *On the set of values of a finite measure.* Bull. Amer. Math. Soc. 53 (1947), 138–141.

(1947-2) *Invariant measures.* Ann. of Math. (2) 48 (1947), 735–754.

(1947-3) *Functions of integrable functions.* J. Indian Math. Soc. (N.S.) 11 (1947), 81–84.

(1948-1) *The range of a vector measure.* Bull. Amer. Math. Soc. 54 (1948), 416–421.

(1949-1) *On a theorem of Dieudonné.* Proc. Nat. Acad. Sci. U.S.A. 35 (1949), 38–42.

(1949-2) *A nonhomogeneous ergodic theorem.* Trans. Amer. Math. Soc. 66 (1949), 284–288.

(1949-3) *Application of the Radon–Nikodym theorem to the theory of sufficient statistics.* Ann. Math. Statistics 20 (1949), 225–241 (with L.J. Savage).

(1949-4) *Measurable transformations.* Bull. Amer. Math. Soc. 55 (1949), 1015–1034.

(1950-1) *The marriage problem.* Amer. J. Math. 72 (1950), 214–215 (with H.E. Vaughan).

(1950-2) *Commutativity and spectral properties of normal operators.* Acta Sci. Math. Szeged 12 (1950), 153–156.

(1950-3) *Normal dilations and extensions of operators.* Summa Brasil. Math. 2 (1950), 125–134.

(1952-1) *Commutators of operators.* Amer. J. Math. 74 (1952), 237–240.

(1952-2) *Some present problems on operators in Hilbert space* (Spanish). Symposium sobre algunos problemas matemáticos se estan estudiando en Latino América, Diciembre, 1951, pp. 9–14. Centro de Cooperación Científica de la Unesco para América Latina. Montevideo, Uruguay, 1952.

(1953-1) *Spectra and spectral manifolds.* Ann. Soc. Polon. Math. 25 (1952), 43–49 (1953).

(1953-2) *Square roots of operators.* Proc. Amer. Math. Soc. 4 (1953), 142–149 (with G. Lumer and J.J. Schäffer).

(1954-1) *Commutators of operators.* II. Amer. J. Math. 76 (1954), 191–198.

(1954-2) *Polyadic Boolean algebras.* Proc. Nat. Acad. Sci. U.S.A. 40 (1954), 296–301.

(1954-3) *Square roots of operators.* II. Proc. Amer. Math. Soc. 5 (1954), 589–595 (with G. Lumer).

(1956-1) *Algebraic logic. I. Monadic Boolean algebras.* Compositio Math. 12 (1956), 217–249.

(1956-2) *Predicates, terms, operations, and equality in polyadic Boolean algebras.* Proc. Nat. Acad. Sci. U.S.A. 42 (1956), 130–136.

(1956-3) *The basic concepts of algebraic logic.* Amer. Math. Monthly 63 (1956), 363–387.

(1956-4) *Algebraic logic. II. Homogeneous locally finite polyadic Boolean algebras of infinite degree.* Fund. Math. 43 (1956), 255–325.

(1956-5) *Algebraic logic. III. Predicates, terms, and operations in polyadic algebras.* Trans. Amer. Math. Soc. 83 (1956), 430–470.

(1957-1) *"Nicolas Bourbaki".* Scientific American, May 1957, 88–99.

(1957-2) *Algebraic logic. IV. Equality in polyadic algebras.* Trans. Amer. Math. Soc. 86 (1957), 1–27.

(1958-1) *Von Neumann on measure and ergodic theory.* Bull. Amer. Math. Soc. 64 (1958), 86–94.

(1958-2) *Products of symmetries.* Bull. Amer. Math. Soc. 64 (1958), 77–78 (with S. Kakutani).

(1959-1) *Free monadic algebras.* Proc. Amer. Math. Soc. 10 (1959), 219–227.

(1959-2) *The representation of monadic Boolean algebras.* Duke Math. J. 26 (1959), 447–454.

(1961-1) *Recent progress in ergodic theory.* Bull. Amer. Math. Soc. 67 (1961), 70–80.

(1961-2) *Injective and projective Boolean algebras.* Proc. Sympos. Pure Math., Vol. II (1961), 114–122, American Mathematical Society, Providence, RI.

(1961-3) *Shifts on Hilbert spaces.* J. Reine Angew. Math. 208 (1961), 102–112.

(1963-1) *What does the spectral theorem say?* Amer. Math. Monthly 70 (1963), 241–247.

(1963-2) *Partial isometries.* Pacific J. Math. 13 (1963), 285–296 (with J.E. McLaughlin).

(1963-3) *Algebraic properties of Toeplitz operators.* J. Reine Angew. Math. 213 (1963/64), 89–102 (with A. Brown).

(1963-4) *A glimpse into Hilbert space.* 1963 Lectures on Modern Mathematics, Vol. 1, pp. 1–22, Wiley, New York.

(1964-1) *On Foguel's answer to Nagy's question.* Proc. Amer. Math. Soc. 15 (1964), 791–793.

(1964-2) *Numerical ranges and normal dilations.* Acta Sci. Math. Szeged 25 (1964), 1–5.

(1965-1) *Cesàro operators.* Acta Sci. Math. Szeged 26 (1965), 125–137 (with A. Brown and A.L. Shields).

(1965-2) *Commutators of operators on Hilbert space.* Canad. J. Math. 17 (1965), 695–708 (with A. Brown and C. Pearcy).

(1966-1) *Invariant subspaces of polynomially compact operators.* Pacific J. Math. 16 (1966), 433–437.

(1968-1) *Permutations of sequences and the Schröder–Bernstein theorem.* Proc. Amer. Math. Soc. 19 (1968), 509–510.

(1968-2) *Irreducible operators.* Michigan Math. J. 15 (1968), 215–223.

(1968-3) *Quasitriangular operators.* Acta Sci. Math. Szeged. 29 (1968), 283–293.

(1968-4) *Mathematics as a creative art.* American Scientist 56 (1968), 375–389.

(1969-1) *Two subspaces.* Trans. Amer. Math. Soc. 144 (1969), 381–389.

(1969-2) *Powers of partial isometries.* J. Math. Mech. 19 (1969/70), 657–663 (with L.J. Wallen).

(1969-3) *Invariant subspaces.* Abstract Spaces and Approximation (Proc. Conf., Oberwolfach, 1968), pp. 26–30, Birkhäuser, Basel, 1969.

(1970-1) *Finite-dimensional Hilbert spaces.* Amer. Math. Monthly 77 (1970), 457–464.

(1970-2) *Capacity in Banach algebras.* Indiana Univ. Math. J. 20 (1970/71), 855–863.

(1970-3) *Ten problems in Hilbert space.* Bull. Amer. Math. Soc. 76 (1970), 887–933.

(1970-4) *How to write mathematics.* Enseignement Math. (2) 16 (1970), 123–152.

(1971-1) *Eigenvectors and adjoints.* Linear Algebra and Appl. 4 (1971), 11–15.

(1971-2) *Reflexive lattices of subspaces.* J. London Math. Soc. (2) 4 (1971), 257–263.

(1971-3) *Positive approximants of operators.* Indiana Univ. Math. J. 21 (1971/72), 951–960.

(1972-1) *Continuous functions of Hermitian operators.* Proc. Amer. Math. Soc. 31 (1972), 130–132.

(1972-2) *Products of shifts.* Duke Math. J. 39 (1972), 772–787.

(1973-1) *The legend of John von Neumann.* Amer. Math. Monthly 80 (1973), 382–394.

(1973-2) *Limits of shifts.* Acta Sci. Math. Szeged 34 (1973), 131–139.

(1974-1) *Spectral approximants of normal operators.* Proc. Edinburgh Math. Soc. (2) 19 (1974/75), 51–58.

(1974-2) *How to talk mathematics.* Notices Amer. Math. Soc. 21 (1974), 155–158.

(1975-1) *What to publish.* Amer. Math. Monthly 82 (1975), 14–17.

(1976-1) *Products of involutions.* Linear Algebra and Appl. 13 (1976), 157–162 (with W.H. Gustafson and H. Radjavi).

(1976-2) *American mathematics from* 1940 *to the day before yesterday.* Amer. Math. Monthly 83 (1976), 503–516 (with J.H. Ewing, W.H. Gustafson, S.H. Moolgavkar, S.H. Wheeler, W.P. Ziemer).

(1976-3) *Some unsolved problems of unknown depth about operators on Hilbert space.* Proc. Royal Soc. Edinburgh Sect. A 76 (1976/77), 67–76.

(1977-1) *Logic from A to G.* Math. Mag. 50 (1977), 5–11.

(1977-2) *The work of Frigyes Riesz* (Hungarian). Mat. Lapok 29 (1977/81), 13–20.

(1978-1) *Integral operators.* Lecture Notes in Math., Vol. 693, pp. 1–15, Springer, Berlin, 1978.

(1978-2) *Fourier series.* Amer. Math. Monthly 85 (1978), 33–34.

(1978-3) *Arithmetic progressions.* Amer. Math. Monthly 85 (1978), 95–96.

(1978-4) *Invariant subspaces.* Amer. Math. Monthly 85 (1978), 182.

(1978-5) *Schauder bases.* Amer. Math. Monthly 85 (1978), 256–257.

(1978-6) *The Serre conjecture.* Amer. Math. Monthly 85 (1978), 357–359 (with W.H. Gustafson and J.M. Zelmanowitz).

(1979-1) *Ten years in Hilbert space.* Integral Equations Operator Theory 2 (1979), 529–564.

(1980-1) *Finite-dimensional points of continuity of Lat.* Linear Algebra Appl. 31 (1980), 93–102 (with J.B. Conway).

(1980-2) *Limsups of Lats.* Indiana Univ. Math. J. 29 (1980), 293–311.

(1980-3) *The heart of mathematics.* Amer. Math. Monthly 87 (1980), 519–524.

(1980-4) *Does mathematics have elements*? Math. Intelligencer 3 (1980/81), 147–153.

(1981-1) *Applied mathematics is bad mathematics.* Mathematics Tomorrow, Springer–Verlag New York Inc., 1981, pp. 9–20.

(1982-1) *Quadratic interpolation.* J. Operator Theory 7 (1982), 303–305.

(1982-2) *Asymptotic Toeplitz operators.* Trans. Amer. Math. Soc. 273 (1982), 621–630 (with J. Barría).

(1982-3) *The thrills of abstraction.* Two-Year College Math. Journal 13 (1982), 242–251.

(1983-1) *BDF or the infinite principal axis theorem.* Notices Amer. Math. Soc. 30 (1983), 387–391.

(1983-2) *The work of F. Riesz. Functions, series, operators.* Vol. I, II (Budapest, 1980), Colloq. Math. Soc. János Bolyai 35, 37–48, North-Holland, Amsterdam, 1983.

(1984-1) *Weakly transitive matrices.* Illinois J. Math. 28 (1984), 370–378 (with J. Barría).

(1984-2) *Béla Szökefalvi–Nagy.* Anniversary volume on approximation theory and functional analysis (Oberwolfach, 1983), pp. 35–39, Internat. Schriftenreihe Numer. Math. 65, Birkhäuser, Basel, 1984.

(1984-3) *Subnormal operators and the subdiscrete topology.* Anniversary volume on approximation theory and functional analysis (Oberwolfach, 1983), pp. 49–65, Internat. Schriftenreihe Numer. Math. 65, Birkhäuser, Basel, 1984.

(1986-1) *I want to be a mathematician* (excerpts). Math. Intelligencer 8 (1986), 26–32.

(1988-1) *Fifty years of linear algebra: a personal reminiscence.* Visiting scholars' lectures – 1987 (Lubbock, TX), pp. 71–89, Texas Tech. Math. Series 15 (1988).

(1988-2) *Some books of auld lang syne.* A century of mathematics in America, Part I, pp. 131–174, Hist. Math. 1, Amer. Math. Soc., Providence, RI, 1988.

(1990-1) *Allen L. Shields.* Math. Intelligencer 12 (1990), 20 (with F.W. Gehring).

(1990-2) *Vector bases for two commuting matrices.* Linear and Multilinear Algebra 27 (1990), 147–157 (with J. Barría).

(1990-3) *Has progress in mathematics slowed down*? Amer. Math. Monthly 97 (1990), 561–588.

(1991-1) *Bad products of good matrices.* Linear and Multilinear Algebra 29 (1991), 1–20.

(1992-1) *Large intersections of large sets.* Amer. Math. Monthly 99 (1992), 307–313.

(1992-2) *The problem of positive contraction approximation.* Chinese J. Math. 20 (1992), 241–248.

(1993-1) *Postcards from Max.* Amer. Math. Monthly 100 (1993), 942–944.

(1995-1) *To count or to think, that is the question.* Nieuw Arch. Wisk. (4) 13 (1995), 61–76.

(2000-1) *An autobiography of polyadic algebras.* Log. J. IGPL 8 (2000), 383–392.

## Books

Paul R. Halmos. *Finite Dimensional Vector Spaces.* Annals of Mathematics Studies, no. 7. Princeton University Press, Princeton, NJ, 1942.

Paul R. Halmos. *Measure Theory.* D. Van Nostrand Company, Inc., New York, NY, 1950.

Paul R. Halmos. *Introduction to Hilbert Space and the Theory of Spectral Multiplicity.* Chelsea Publishing Company, New York, NY, 1951.

Paul R. Halmos. *Lectures on Ergodic Theory.* Publications of the Mathematical Society of Japan, no. 3, The Mathematical Society of Japan, 1956. Republished by Chelsea Publishing Company, New York, NY, 1960.

Paul R. Halmos. *Introduction to Hilbert Space and the Theory of Spectral Multiplicity. 2nd ed.* Chelsea Publishing Company, New York, NY, 1957. Reprinted in 1998 by AMS Chelsea Publishing, Providence, RI.

Paul R. Halmos. *Finite-Dimensional Vector Spaces. 2nd ed.* The University Series in Undergraduate Mathematics. D. Van Nostrand Company, Inc., Princeton–Toronto–New York–London, 1958. Reprinted by Springer–Verlag, New York–Heidelberg, 1974, in the series Undergraduate Texts in Mathematics.

Paul R. Halmos. *Naive Set Theory.* The University Series in Undergraduate Mathematics, D. Van Nostrand Company, Princeton–Toronto–London–New York, 1960. Reprinted by Springer–Verlag, New York–Heidelberg, 1974, in the series Undergraduate Texts in Mathematics.

Paul R. Halmos. *Algebraic Logic.* Chelsea Publishing Company, New York, NY, 1962.

Paul R. Halmos. *Lectures on Boolean Algebras.* Van Nostrand Mathematical Studies, no. 1, D. Van Nostrand Company, Inc., Princeton, NJ, 1963.

Paul R. Halmos. *A Hilbert Space Problem Book.* D. Van Nostrand Company, Inc., Princeton–Toronto–London, 1967.

Paul R. Halmos and V.S. Sunder. *Bounded Integral Operators on $L^2$ Spaces.* Ergebnisse der Mathematik und ihrer Grenzgebiete, 96. Springer–Verlag, Berlin–New York, 1978.

Paul R. Halmos. *A Hilbert Space Problem Book.* Second edition. Springer–Verlag, New York–Berlin, 1982.

Paul R. Halmos. *Selecta: Research Contributions.* Springer–Verlag, New York, 1983.

Paul R. Halmos. *Selecta: Expository Writing.* Springer–Verlag, New York, 1983.

Paul R. Halmos. *I want to be a mathematician. An automathography.* Springer–Verlag, New York, 1985. Published also by Mathematical Association of America, Washington, DC, 1985.

Paul R. Halmos. *I have a photographic memory.* American Mathematical Society, Providence, RI, 1987.

Paul R. Halmos. *Problems for mathematicians, young and old.* The Dolciani Mathematical Expositions, 12. Mathematical Association of America, Washington, DC, 1991.

Paul R. Halmos. *Linear algebra problem book.* The Dolciani Mathematical Expositions, 16. Mathematical Association of America, Washington, DC, 1995.

Steven Givant and Paul R. Halmos. *Logic as algebra.* The Dolciani Mathematical Expositions, 21. Mathematical Association of America, Washington, DC, 1995.

Steven Givant and Paul R. Halmos. *Introduction to Boolean algebras.* Undergraduate Texts in Mathematics, Springer, New York, 2009.

Operator Theory:
Advances and Applications, Vol. 207, 41–77

# Photos

**Part 1.** Paul through the years, pages 43–50

**Part 2.** Some operator theorists of the Halmos era, and other mathematicians, pages 51–77

Paul was a prodigious amateur photographer. He seems not to have been motivated by photography as art, but rather to have been impelled by a desire to preserve tangible mementos of his experiences.

In the preface to his book *I Have a Photographic Memory*, published in 1987, Paul estimates that, by then, he had taken over 16,000 snapshots, about 6000 of which are mathematically connected. Photography served in part as an adjunct to his professional activities. He took snapshots of the students in his classes to help him learn their names. He took photos of mathematicians visiting his department, and of mathematicians at departments he visited. Before visiting another department, he would look over any photos of local mathematicians in his collection, so he could greet them quickly by name. *I Have a Photographic Memory* contains over 600 of his photographs of mathematicians.

All of the photos in Part 1, and the bulk of those in Part 2, are from Paul's personal collection; those in the latter group were taken by Paul, except for the one on the last photo page, in which Paul is one of the subjects (and which was no doubt taken with his camera). All but three of the photos from Paul's collection appearing here are reproduced with the kind permission of Virginia Halmos. Photos of the following mathematicians were provided by their subjects: Jim Agler and Nicholas Young; Michael Dritschel; Nathan Feldman; John M$^c$Carthy; Carl Pearcy; Gilles Pisier; Mihai Putinar; Alexander Volberg. We are grateful to Jim Rovnyak for providing the photo of Marvin and Betsy Rosenblum. Three of the photographs by Paul, the one of David Lowdenslager, the joint one of Larry Brown, Ronald Douglas and Peter Fillmore, and the joint one of Mary R. Embry-Wardrop, Catherine L. Olsen and Pratibha Ghatage, are reproduced from *I Have a Photographic Memory*, © 1987, American Mathematical Society, with permission of the publisher.

The mathematicians pictured in Part 2 include the contributors to this volume, operator theorists and a few other mathematicians closely connected to Paul and/or to his work, and operator theorists mentioned prominently in our expository articles.

In assembling the photos in this volume, we were fortunate to have access to Paul's personal collection through Gerald Alexanderson, who at the time was the collection's temporary custodian in his role as its cataloguer. We deeply appreciate Jerry's invaluable assistance. Thanks also go to José Barría for his aid in this project. Paul's photo collection, along with his personal papers, are now in the Texas Archives in Austin.

The photos to follow, we hope, will help to evoke Paul's spirit.

A 1929 photo of Paul getting off the train in Basel

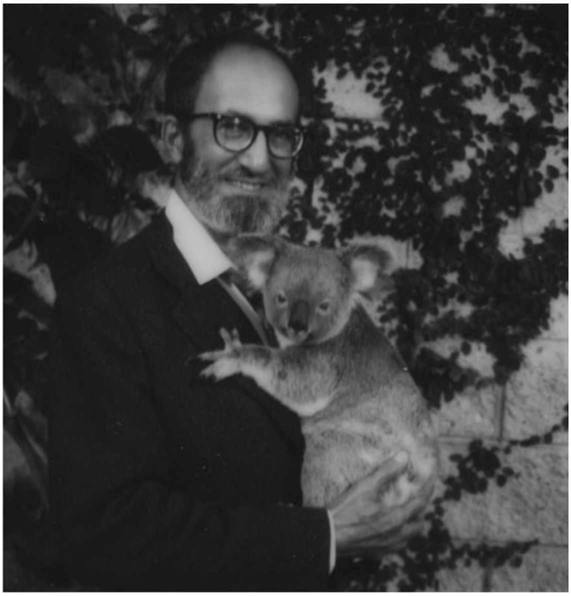
On board the steamship Bremen, 1931

Paul and Virginia Halmos, Cambridge, 1950

In a double breasted suit, 1951

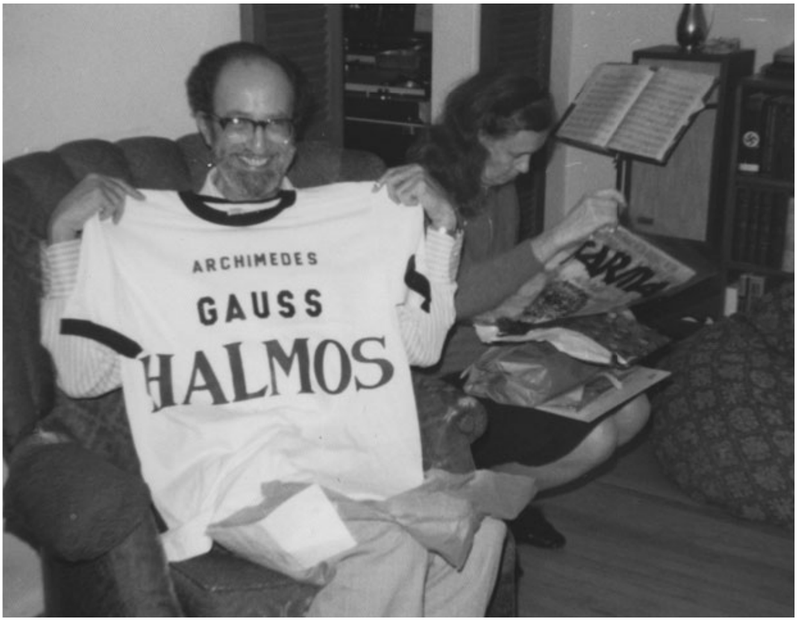
With “Bertie”, short for Bertrand Russell, as a puppy, 1967

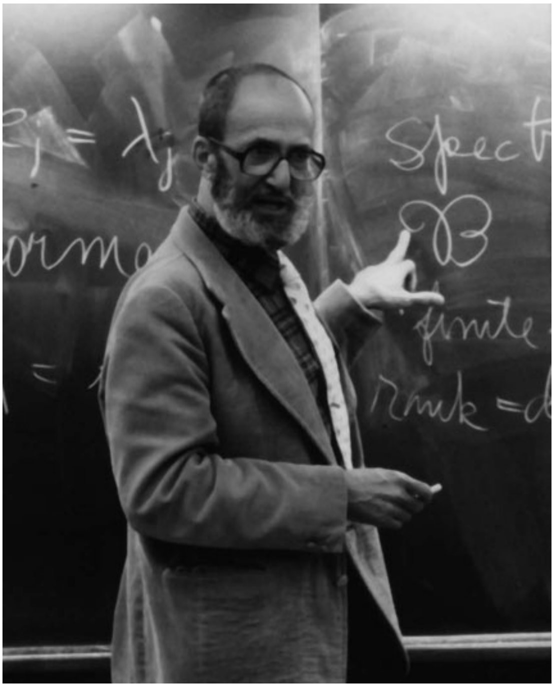
Eugene, 1969, lecturing



Brisbane, 1975

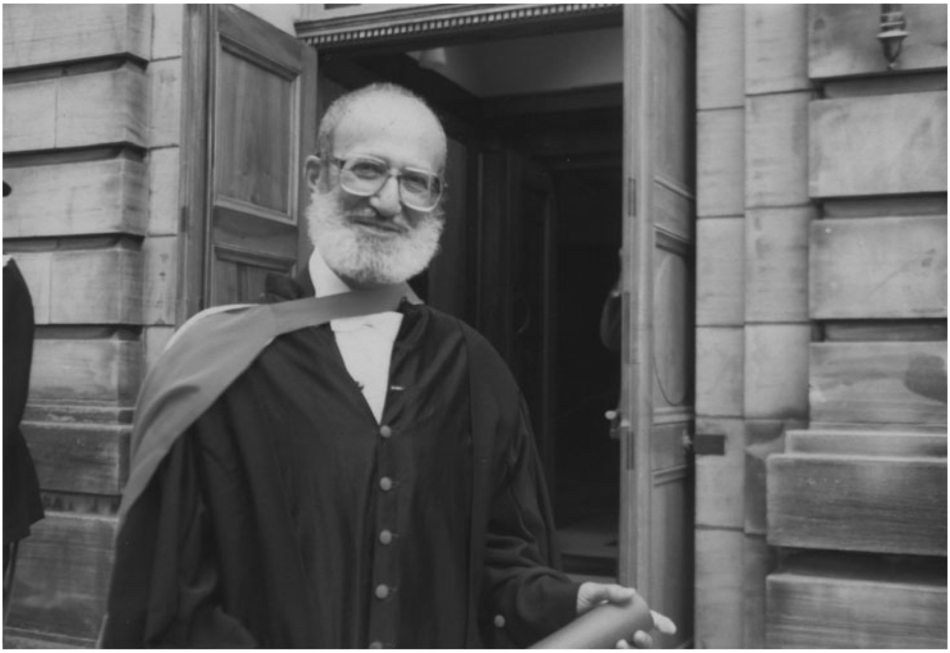
Paul and Virginia, Bloomington, 1976

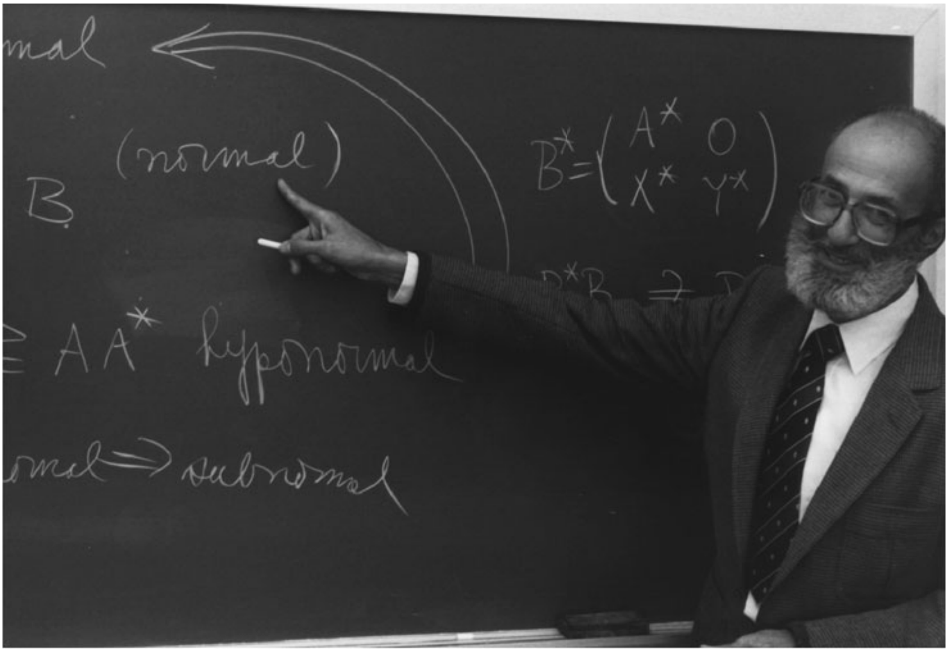




Circa 1980

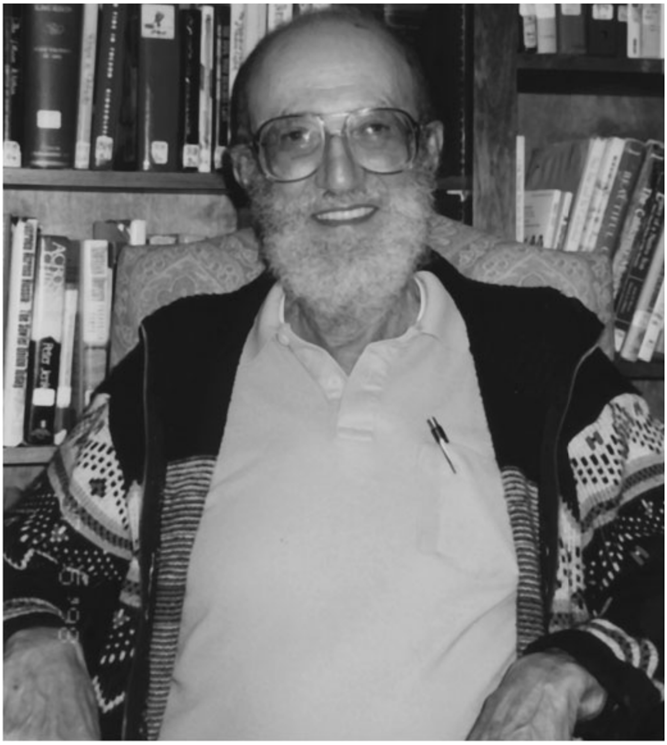
St. Andrews,
circa 1980



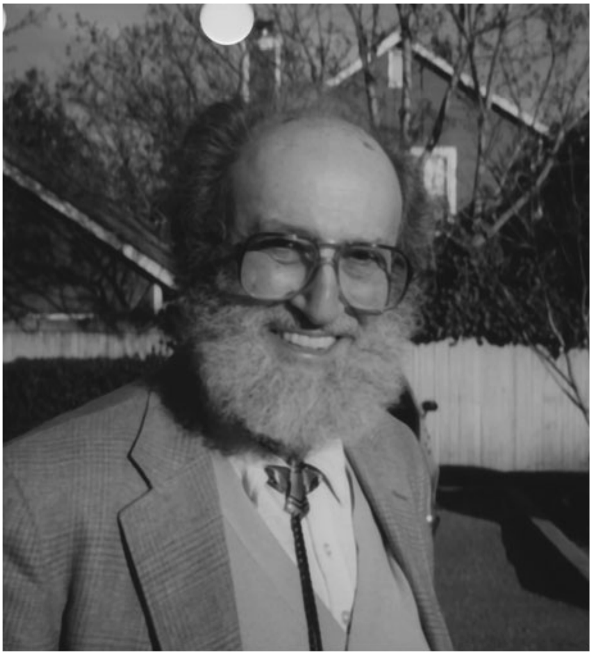


Circa 1980

With Max Zorn, 1983



Los Gatos,
1998



Los Gatos,
2000

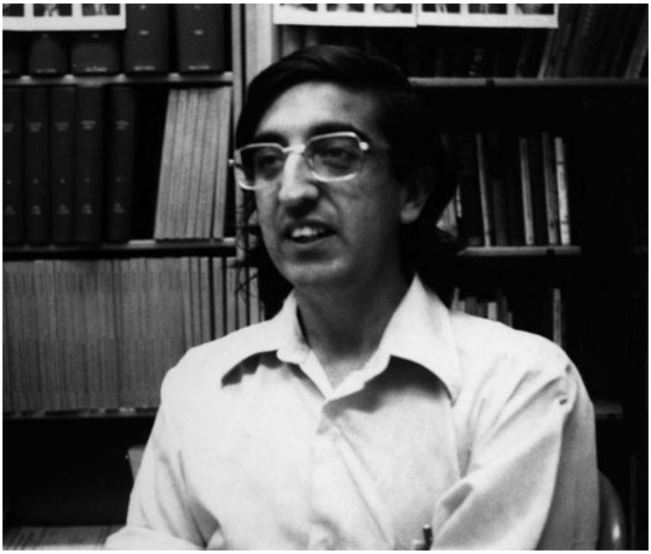
Jim Agler, right
and Nicholas Young,
San Diego,
2009

Tsuyoshi Ando,
Overwolfach,
1980

Constantin Apostol,
Bloomington,
1974

Bill Arveson,
Bloomington,
1970

Sheldon Axler, right, and Donald Sarason,
Kalamazoo, 1975



Jose Barria,
Bloomington,
1974

Hari Bercovici,
Bloomington,
1982

Charles Berger, 1977

Arlen Brown,
Bloomington,
1978

Kevin Clancey, left,
and Douglas Clark,
1974

The reknowned BDF
From right to left: Larry Brown, Ronald Douglas, Peter Fillmore

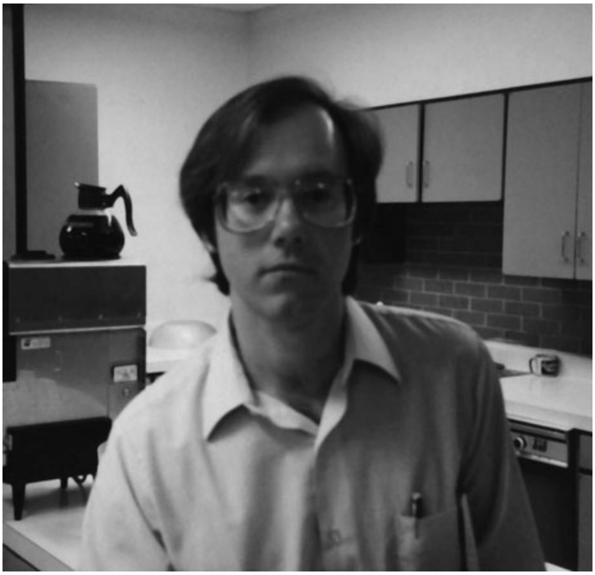
Scott Brown,
Berkeley,
1984

Lewis Coburn,
1965

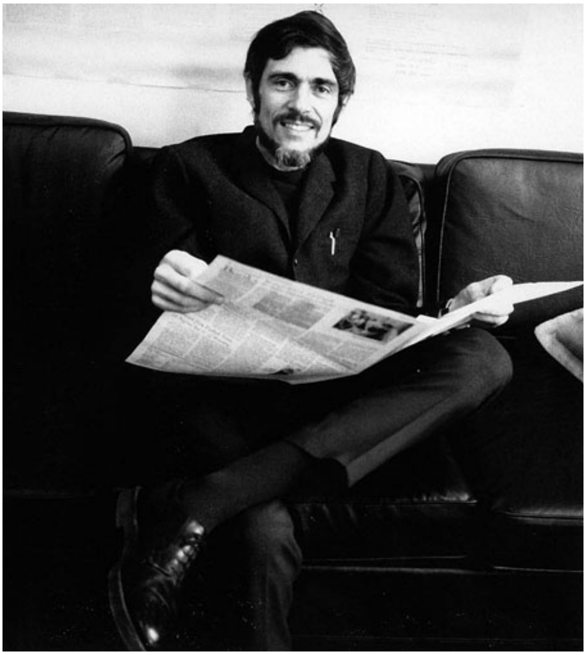
John B. Conway,
Bloomington,
1983

Carl Cowan,
San Antonio,
1976

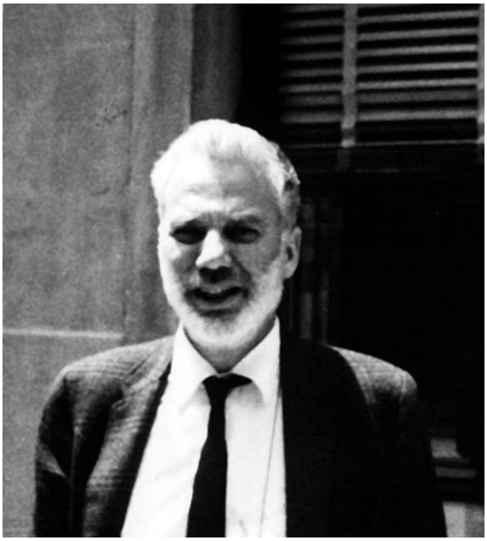
Raul Curto,
Bloomington,
1979



Ken Davidson,
Tempe,
1984



Chandler Davis,
Bloomington,
1968



Joseph Doob,
New York,
1969

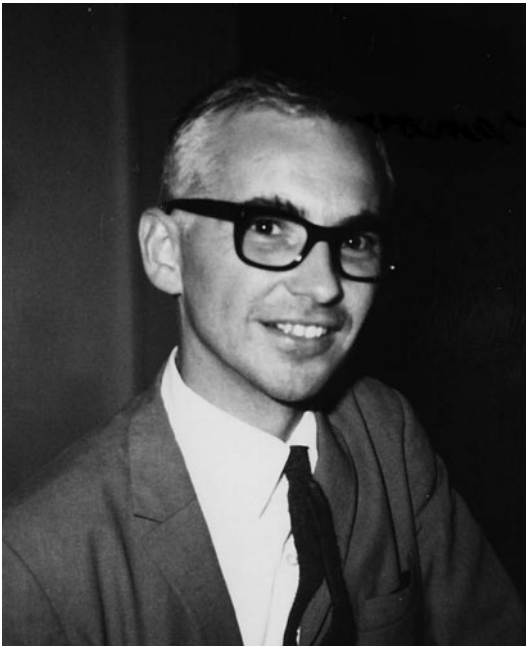
John Ewing

Michael Dritschel

Nathan Feldman

Bela Sz.-Nagy and Ciprian Foias,
1983

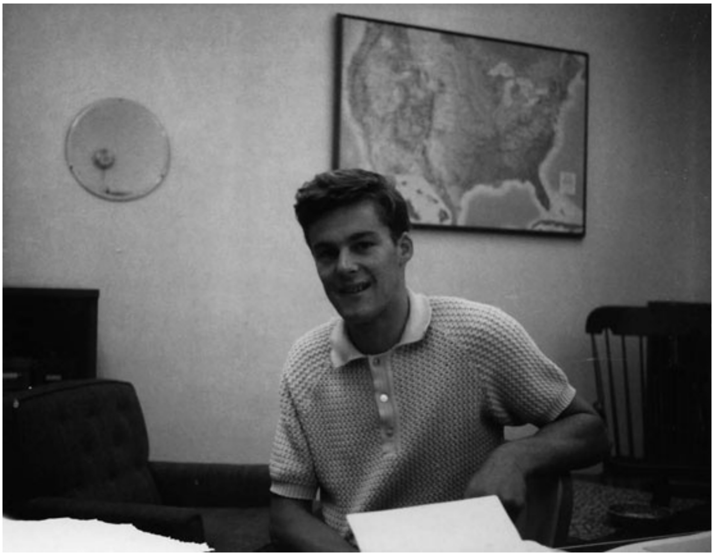
Israel Gohberg,
Bloomington,
1975



Henry Helson,
1969

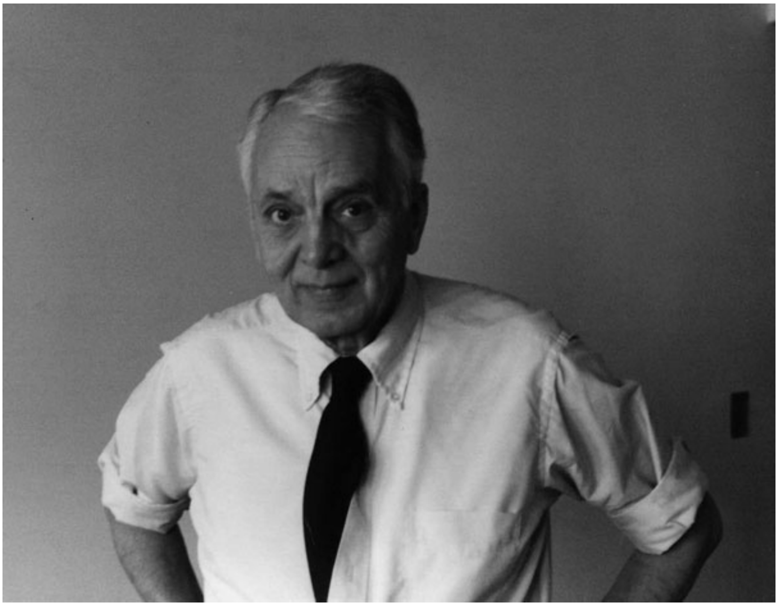
Don Hadwin, Fort Worth, 1990

Ken Harrison, Honolulu, 1969



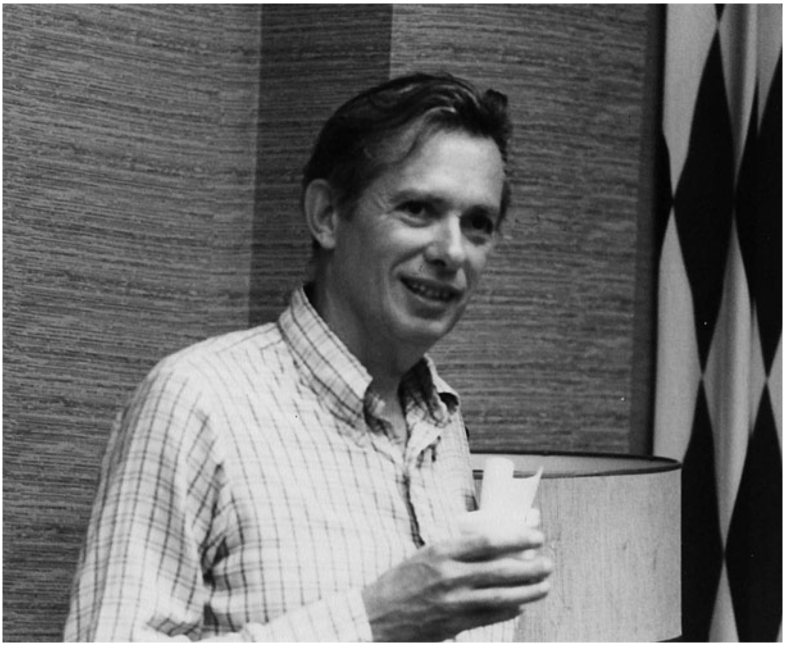
Domingo Herrero, Tempe, 1985

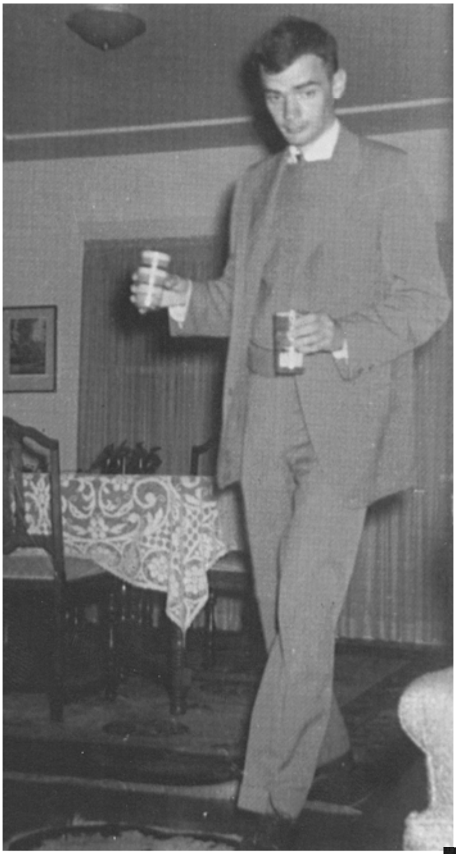
Tom Kriete, Charlottesville, 1986



Warren Ambrose

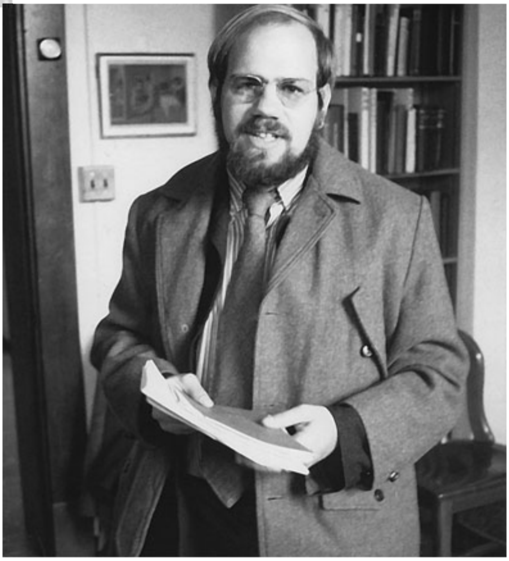
Errett Bishop





David Lowdenslager,
Berkeley,
1956

John M$^c$ Carthy



Paul Muhly,
Bloomington,
1969

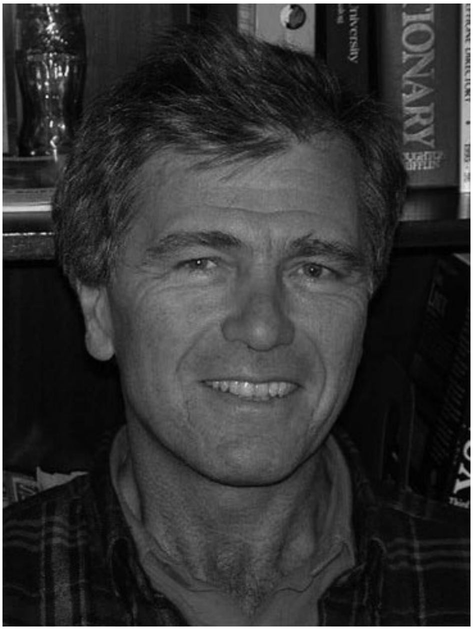
Zeev Nehari,
1968

Nikolai Nikolski,
Lancaster,
1984

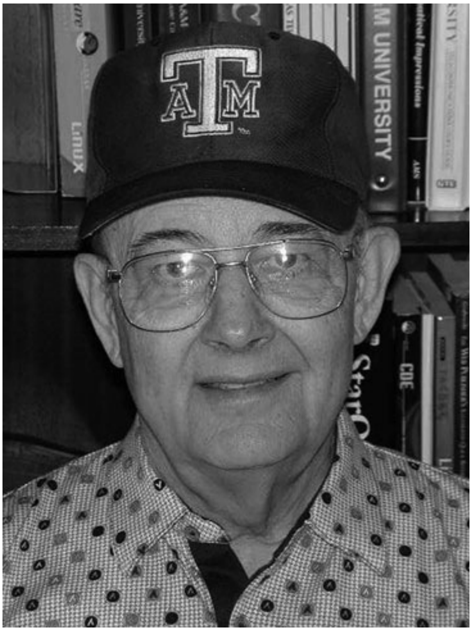
Eric Nordgren,
Amherst,
1969

Robert Olin,
1977

Vladimir Peller,
Budapest,
1980



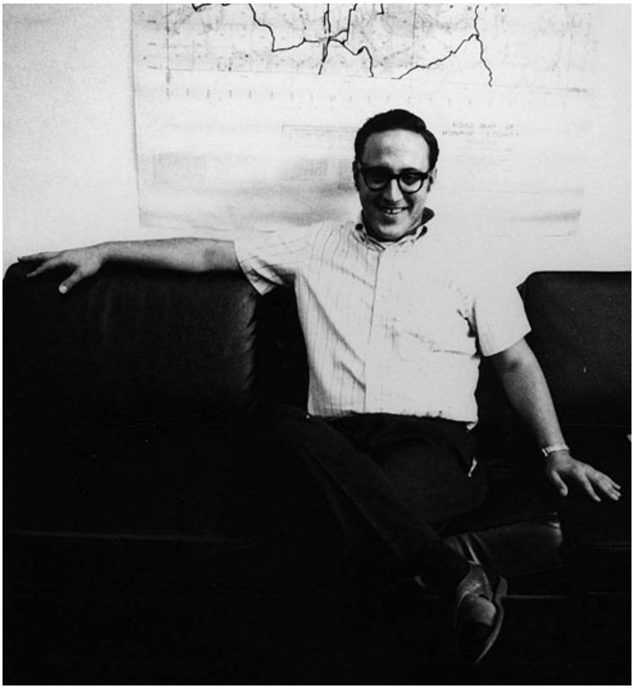
Gilles Pisier



Carl Pearcy

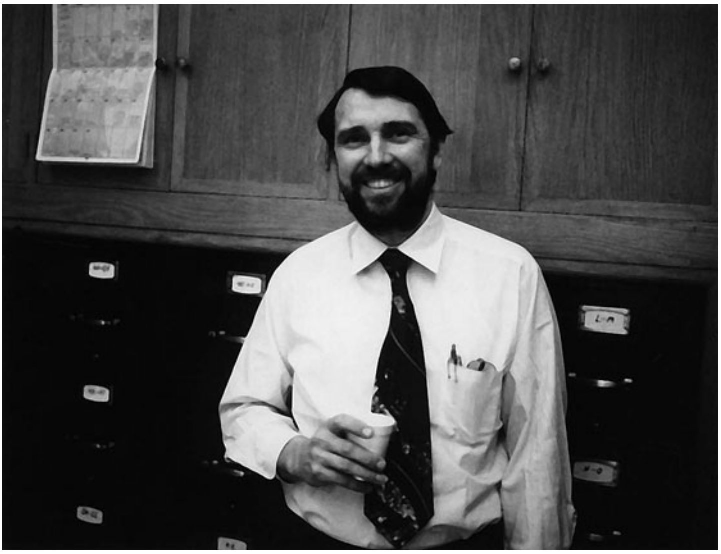
Mihai Putinar

Haydar Radjavi,
Bloomington,
1974

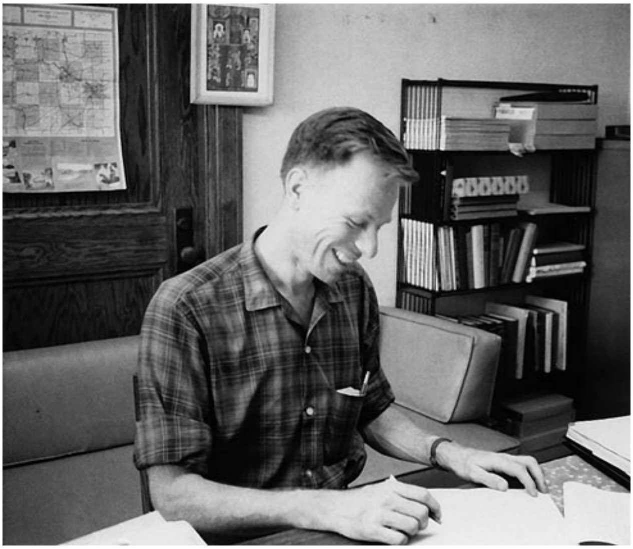


Peter Rosenthal,
1969

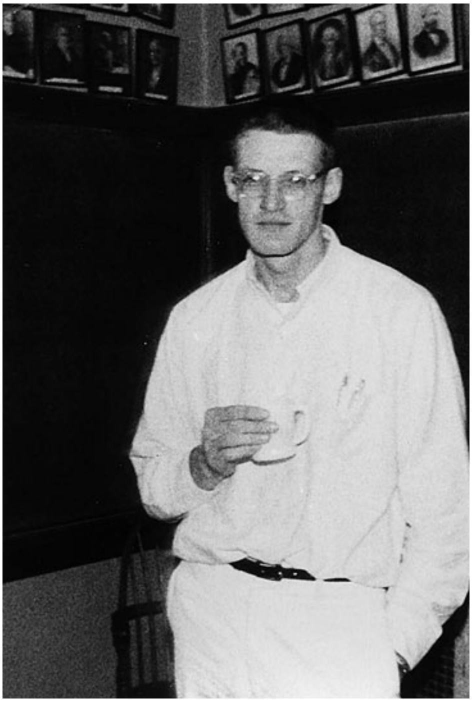
Marvin and Betsy Rosenblum



James Rovnyak



Allen Shields,
Ann Arbor,
1964



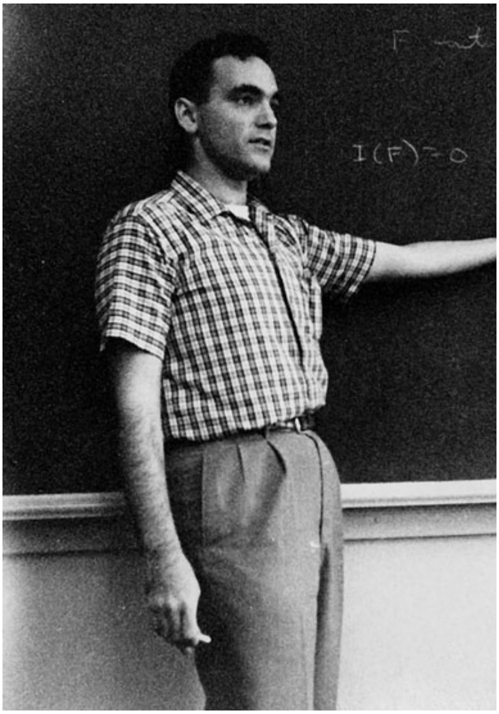
Joseph Stampfli,
1961

V.S. Sunder,
Bloomington,
1973

James Thomson

Alexander
Volberg

Dan-Vergil Voiculescu, 1968



Harold Widom, 1962

Four generations of operator theory.
Right to left,
Paul Halmos,
his student Donald Sarason,
Don's student Sheldon Axler,
Sheldon's student Pamela Gorkin,
Lancaster, 1984

Mary R. Embry-Wardrop, Catherine L. Olson, Pratibha Ghatage, Crawfordsville, Indiana, 1973

# Part II

# Articles

Operator Theory:
Advances and Applications, Vol. 207, 81–97

# What Can Hilbert Spaces Tell Us About Bounded Functions in the Bidisk?

Jim Agler and John E. McCarthy

*Dedicated to the memory of Paul R. Halmos*

**Abstract.** We discuss various theorems about bounded analytic functions on the bidisk that were proved using operator theory.

**Mathematics Subject Classification (2000).** 32-02, 32A70, 46E22.

**Keywords.** Bidisk, Hilbert space model, bounded analytic functions, $H^\infty$.

## 1. Introduction

Much of modern single operator theory (as opposed to the study of operator algebras) rests on a foundation of complex analysis. Every cyclic operator can be represented as multiplication by the independent variable on the completion of the polynomials with respect to some norm. The nicest way to have a norm is in $L^2(\mu)$, and then one is led to the study of subnormal operators, introduced by P.R. Halmos in [34]. The study of cyclic subnormal operators becomes the study of the spaces $P^2(\mu)$, the closure of the polynomials in $L^2(\mu)$, and the theory of these spaces relies on a blend of complex analysis and functional analysis; see J. Conway's book [27] for an exposition. Alternatively, one can start with a Hilbert space that is amenable to complex analysis, such as the Hardy space $H^2$, and study classes of operators on that space that have a good function theoretic representation, such as Toeplitz, Hankel or composition operators. All of these classes of operators have a rich theory, which depends heavily on function theory – for expositions, see, *e.g.*, [41, 43] and [29].

The traffic, of course, goes both ways. There are many questions in function theory that have either been answered or illuminated by an operator theory approach. The purpose of this article is to describe how operator theory has fared

Jim Agler was partially supported by National Science Foundation Grant DMS 0400826, and John McCarthy was partially supported by National Science Foundation Grant DMS 0501079.

when studying $H^\infty(\mathbb{D}^2)$, the algebra of bounded analytic functions on the bidisk $\mathbb{D}^2$. We focus on function theory results that were proved, originally at least, using operator theory.

For the topics in Sections 2 to 7, we shall first describe the situation on the disk $\mathbb{D}$, and then move on to the bidisk. The topics in Sections 8 to 9 do not really have analogues in one dimension. For simplicity, we shall stick to scalar-valued function theory, though many of the results have natural matrix-valued analogues.

We shall use the notation that points in the bidisk are called $\lambda$ or $\zeta$, and their coordinates will be given by superscripts: $\lambda = (\lambda^1, \lambda^2)$. We shall use $z$ and $w$ to denote the coordinate functions on $\mathbb{D}^2$. The closed unit ball of $H^\infty(\mathbb{D})$ will be written $H^\infty_1(\mathbb{D})$ and the closed unit ball of $H^\infty(\mathbb{D}^2)$ as $H^\infty_1(\mathbb{D}^2)$.

## 2. Realization formula

The realization formula is a way of associating isometries (or contractions) with functions in the ball of $H^\infty(\mathbb{D})$. In one dimension, it looks like the following; see, *e.g.*, [19] or [9] for a proof.

**Theorem 2.1.** *The function $\phi$ is in the closed unit ball of $H^\infty(\mathbb{D})$ if and only if there is a Hilbert space $\mathcal{H}$ and an isometry $V : \mathbb{C} \oplus \mathcal{H} \to \mathbb{C} \oplus \mathcal{H}$, such that, writing $V$ as*

$$V = \begin{array}{c c} & \begin{array}{cc} \mathbb{C} & \mathcal{H} \end{array} \\ \begin{array}{c} \mathbb{C} \\ \mathcal{H} \end{array} & \begin{pmatrix} A & B \\ C & D \end{pmatrix} \end{array}, \tag{2.2}$$

*one has*

$$\phi(\lambda) = A + \lambda B(I - \lambda D)^{-1}C. \tag{2.3}$$

This formula was generalized to the bidisk in [3]. It becomes

**Theorem 2.4.** *The function $\phi$ is in the closed unit ball of $H^\infty(\mathbb{D}^2)$ if and only if there are auxiliary Hilbert spaces $\mathcal{H}_1$ and $\mathcal{H}_2$ and an isometry*

$$V : \mathbb{C} \oplus \mathcal{H}_1 \oplus \mathcal{H}_2 \to \mathbb{C} \oplus \mathcal{H}_1 \oplus \mathcal{H}_2$$

*such that, if $\mathcal{H} := \mathcal{H}_1 \oplus \mathcal{H}_2$, $V$ is written as*

$$V = \begin{array}{c c} & \begin{array}{cc} \mathbb{C} & \mathcal{H} \end{array} \\ \begin{array}{c} \mathbb{C} \\ \mathcal{H} \end{array} & \begin{pmatrix} A & B \\ C & D \end{pmatrix} \end{array}, \tag{2.5}$$

*and $\mathcal{E}_\lambda = \lambda^1 I_{\mathcal{H}_1} \oplus \lambda^2 I_{\mathcal{H}_2}$, then*

$$\phi(\lambda) = A + B\mathcal{E}_\lambda(I_{\mathcal{H}} - D\mathcal{E}_\lambda)^{-1}C. \tag{2.6}$$

There is a natural generalization of (2.6) to functions of $d$ variables. One chooses $d$ Hilbert spaces $\mathcal{H}_1, \dots, \mathcal{H}_d$, lets $\mathcal{H} = \mathcal{H}_1 \oplus \cdots \oplus \mathcal{H}_d$, lets $\mathcal{E}_\lambda = \lambda^1 I_{\mathcal{H}_1} \oplus \cdots \oplus \lambda^d I_{\mathcal{H}_d}$, and then, for any isometry $V$ as in (2.5), let

$$\psi(\lambda) = A + B\mathcal{E}_\lambda(I_{\mathcal{H}} - D\mathcal{E}_\lambda)^{-1}C. \tag{2.7}$$

The set of all functions $\psi$ that are realizable in this way is exactly the Schur-Agler class, a class of analytic functions of $d$ variables that can also be defined as

$$\{\psi : \|\psi(T_1, \ldots, T_d)\| \leq 1 \quad \forall \text{ commuting contractive matrices } (T_1, \ldots, T_d)\}. \tag{2.8}$$

Von Neumann's inequality [58] is the assertion that for $d = 1$ the Schur-Agler class equals $H_1^\infty(\mathbb{D})$; Andô's inequality [16] is the analogous equality for $d = 2$. Once $d > 2$, the Schur-Agler class is a proper subset of the closed unit ball of $H^\infty(\mathbb{D}^d)$ [57, 30]. Many of the results in Sections 3 to 7 are true, with similar proofs, for the Schur-Agler class in higher dimensions (or rather the norm for which this is the unit ball)[1], but it is not known how to generalize them to $H^\infty(\mathbb{D}^d)$.

The usefulness of the realization formula stems primarily not from its ability to represent functions, but to produce functions with desired properties with the aid of a suitably chosen isometry $V$ (dubbed a *lurking isometry* by Joe Ball). An example of a lurking isometry argument is the proof of Pick's theorem in Section 3.

It is well known that equality occurs in the Schwarz lemma on the disk only for Möbius transformations. The Schwarz lemma on $\mathbb{D}^d$ reads as follows (see [47] for a proof).

**Proposition 2.9.** *If $f$ is in $H_1^\infty(\mathbb{D}^d)$, then*

$$\sum_{r=1}^{d}(1 - |\lambda^r|^2)\left|\frac{\partial f}{\partial \lambda^r}(\lambda)\right| \leq 1 - |f(\lambda)|^2. \tag{2.10}$$

G. Knese [38] proved that equality in (2.10) resulted in a curious form of the realization formula. (Notice that for $d \geq 3$, the hypothesis is that $f$ lie in $H_1^\infty(\mathbb{D}^d)$, but the conclusion means $f$ must be in the Schur-Agler class.)

**Theorem 2.11.** *Suppose $f \in H_1^\infty(\mathbb{D}^d)$ is a function of all d variables, and equality holds in* (2.10) *everywhere on $\mathbb{D}^d$. This occurs if and only if $f$ has a representation as in* (2.7) *where each space $\mathcal{H}_r$ is one-dimensional, and the unitary $V$ is symmetric* (*equal to its own transpose*).

Analyzing the realization formula, J.M. Anderson, M. Dritschel and J. Rovnyak were able to obtain the following higher derivative version of the Schwarz lemma [15]:

**Theorem 2.12.** *Let $f$ be in $H_1^\infty(\mathbb{D}^2)$, and $n_1, n_2$ non-negative integers with $n = n_1 + n_2$. Let $\lambda = (z, w)$ be in $\mathbb{D}^2$, with $|\lambda| = \max(|z|, |w|)$. Then*

$$\left|\frac{\partial^n f}{\partial^{n_1} z \partial^{n_2} w}\right| \leq (n-2)! \frac{1-|f(\lambda)|^2}{(1-|\lambda|)^{n-1}}\left[\frac{n_1^2 - n_1}{1-|z|^2} + \frac{2n_1 n_2}{\sqrt{1-|z|^2}\sqrt{1-|w|^2}} + \frac{n_2^2 - n_2}{1-|w|^2}\right].$$

[1] Specifically, Theorems 3.8, 4.3, 6.4 and (i)⇔(iii) of Theorem 7.7.

## 3. Pick problem

The Pick problem on the disk is to determine, given $N$ points $\lambda_1, \dots, \lambda_N$ in $\mathbb{D}$ and $N$ complex numbers $w_1, \dots, w_N$, whether there exists $\phi \in H_1^\infty(\mathbb{D})$ such that

$$\phi(\lambda_i) = w_i, \qquad i = 1, \dots, N.$$

G. Pick proved [44] that the answer is yes if and only if the $N$-by-$N$ matrix

$$\left( \frac{1 - w_i \bar{w}_j}{1 - \lambda_i \bar{\lambda}_j} \right) \tag{3.1}$$

is positive semi-definite.

D. Sarason realized in [49] that Pick's theorem can be proved by showing that operators that commute with the backward shift on an invariant subspace can be lifted with preservation of norm to operators that commute with it on all of $H^2$; this result was then generalized by B. Sz.-Nagy and C. Foiaş to the commutant lifting theorem [52]. Here is a proof of Pick's theorem using a lurking isometry. We shall let

$$k_\lambda^S(\zeta) \;=\; \frac{1}{1 - \bar{\lambda}\zeta}$$

denote the Szegő kernel.

*Proof.* (Necessity) If such a $\phi$ exists, then $I - M_\phi M_\phi^*$ is a positive operator on all of $H^2$. In particular, for any scalars $c_1, \dots, c_N$, we have

$$0 \;\le\; \langle (I - M_\phi M_\phi^*) \sum c_j k_{\lambda_j}^S, \sum c_i k_{\lambda_i}^S \rangle \;=\; \sum c_j \bar{c}_i \frac{1 - w_i \bar{w}_j}{1 - \lambda_i \bar{\lambda}_j}. \tag{3.2}$$

Therefore (3.1) is positive semi-definite.

(Sufficiency) If (3.1) is positive semi-definite of rank $M$, then one can find vectors $\{g_i\}_{i=1}^N$ in $\mathbb{C}^M$ such that

$$\frac{1 - w_i \bar{w}_j}{1 - \lambda_i \bar{\lambda}_j} \;=\; \langle g_i, g_j \rangle_{\mathbb{C}^M}. \tag{3.3}$$

We can rewrite (3.3) as

$$1 + \langle \lambda_i g_i, \lambda_j g_j \rangle_{\mathbb{C}^M} \;=\; w_i \bar{w}_j + \langle g_i, g_j \rangle_{\mathbb{C}^M}. \tag{3.4}$$

The lurking isometry $V : \mathbb{C} \oplus \mathbb{C}^M \to \mathbb{C} \oplus \mathbb{C}^M$ is defined by

$$V : \begin{pmatrix} 1 \\ \lambda_i g_i \end{pmatrix} \mapsto \begin{pmatrix} w_i \\ g_i \end{pmatrix}. \tag{3.5}$$

We extend linearly to the span of

$$\left\{ \begin{pmatrix} 1 \\ \lambda_i g_i \end{pmatrix} \;:\; i = 1, \dots, N \right\}, \tag{3.6}$$

and if this is not the whole space $\mathbb{C} \oplus \mathbb{C}^M$, we extend $V$ arbitrarily so that it remains isometric. Write $V$ as

$$V = \begin{array}{c|cc} & \mathbb{C} & \mathbb{C}^M \\ \hline \mathbb{C} & A & B \\ \mathbb{C}^M & C & D \end{array},$$

and define $\phi$ by

$$\phi(\lambda) = A + \lambda B(I - \lambda D)^{-1} C. \tag{3.7}$$

By the realization formula Theorem 2.1, $\phi$ is in $H_1^\infty(\mathbb{D})$. Moreover, as (3.5) implies that

$$A + B\lambda_i g_i = w_i \tag{3.8}$$

$$C + D\lambda_i g_i = g_i, \tag{3.9}$$

we get that

$$(I - \lambda_i D)^{-1} C = g_i,$$

and hence

$$\phi(\lambda_i) = A + \lambda_i B g_i = w_i,$$

so $\phi$ interpolates. □

(It is not hard to show that $\phi$ is actually a Blaschke product of degree $M$.)

A similar argument using Theorem 2.4 solves the Pick problem on the bidisk. The theorem was first proved, by a different method, in [2].

**Theorem 3.8.** *Given points $\lambda_1, \dots, \lambda_N$ in $\mathbb{D}^2$ and complex numbers $w_1, \dots, w_N$, there is a function $\phi \in H_1^\infty(\mathbb{D}^2)$ that maps each $\lambda_i$ to the corresponding $w_i$ if and only if there are positive semi-definite matrices $\Gamma^1$ and $\Gamma^2$ such that*

$$1 - w_i \bar{w}_j = (1 - \lambda_i^1 \bar{\lambda}_j^1)\Gamma_{ij}^1 + (1 - \lambda_i^2 \bar{\lambda}_j^2)\Gamma_{ij}^2. \tag{3.9}$$

On the polydisk, a necessary condition to solve the Pick problem analogous to (3.9) has recently been found by A. Grinshpan, D. Kaliuzhnyi-Verbovetskyi, V. Vinnikov and H. Woerdeman [33]. As of this writing, it is unknown if the condition is also sufficient, but we would conjecture that it is not.

**Theorem 3.10.** *Given points $\lambda_1, \dots, \lambda_N$ in $\mathbb{D}^d$ and complex numbers $w_1, \dots, w_N$, a necessary condition for there to be a function $\phi \in H_1^\infty(\mathbb{D}^d)$ that maps each $\lambda_i$ to the corresponding $w_i$ is: For every $1 \leq p < q \leq d$, there are positive semi-definite matrices $\Gamma^p$ and $\Gamma^q$ such that*

$$1 - w_i \bar{w}_j = \prod_{r \neq q} (1 - \lambda_i^r \bar{\lambda}_j^r)\Gamma_{ij}^q + \prod_{r \neq p} (1 - \lambda_i^r \bar{\lambda}_j^r)\Gamma_{ij}^p. \tag{3.11}$$

## 4. Nevanlinna problem

If the Pick matrix (3.1) is singular (*i.e.*, if $M < N$) then the solution is unique; otherwise it is not. R. Nevanlinna found a parametrization of all solutions in this latter case [40] (see also [20] for a more modern approach).

**Theorem 4.1.** *If* (3.1) *is invertible, there is a* 2*-by-*2 *contractive matrix-valued function*

$$G = \begin{pmatrix} G_{11} & G_{12} \\ G_{21} & G_{22} \end{pmatrix}$$

*such that the set of all solutions of the Pick problem is given by*

$$\{\phi = G_{11} + G_{12}\frac{\psi G_{21}}{1 - G_{22}\psi} \ : \ \psi \ \in \ H_1^\infty(\mathbb{D})\}.$$

On the bidisk, we shall discuss uniqueness in Section 8 below. Consider now the non-unique case. Let $\phi$ be in $H_1^\infty(\mathbb{D}^2)$, and so by Theorem 2.4 it has a representation as in (2.6). Define vector-valued functions $F_1, F_2$ by

$$\begin{array}{cc} & \mathbb{C} \\ \begin{array}{c}\mathcal{H}_1 \\ \mathcal{H}_2\end{array} & \begin{pmatrix} F_1(\lambda) \\ F_2(\lambda) \end{pmatrix} \end{array} \ := \ (I_{\mathcal{H}} - D\mathcal{E}_\lambda)^{-1}C. \tag{4.2}$$

For a given solvable Pick problem with a representation as (3.9), say that $\phi$ is affiliated with $(\Gamma^1, \Gamma^2)$ if, for some representation of $\phi$ and $F_1, F_2$ as in (4.2),

$$\begin{aligned} F_1(\lambda_i)^* F_1(\lambda_j) &= \Gamma^1_{ij} \\ F_2(\lambda_i)^* F_2(\lambda_j) &= \Gamma^2_{ij} \end{aligned}$$

for $i, j = 1, \ldots, N$. The situation is complicated by the fact that for a given $\phi$, the pairs $(\Gamma^1, \Gamma^2)$ with which it is affiliated *may or may not* be unique. J. Ball and T. Trent [21] proved:

**Theorem 4.3.** *Given a solvable Pick problem, with a representation as in* (3.9)*, there is a matrix-valued function* $G$

$$G \ = \ \begin{array}{cc} & \begin{array}{cc} \mathbb{C} & \mathbb{C}^M \end{array} \\ \begin{array}{c}\mathbb{C} \\ \mathbb{C}^M\end{array} & \begin{pmatrix} G_{11} & G_{12} \\ G_{21} & G_{22} \end{pmatrix} \end{array}$$

*in the closed unit ball of* $H^\infty(\mathbb{D}^2, B(\mathbb{C} \oplus \mathbb{C}^M, \mathbb{C} \oplus \mathbb{C}^M))$*, such that the function* $\phi$ *solves the Pick problem and is affiliated with* $(\Gamma^1, \Gamma^2)$ *if and only if it can be written as*

$$\phi \ = \ G_{11} + G_{12}\Psi(I - G_{22}\Psi)^{-1}G_{21} \tag{4.4}$$

*for some* $\Psi$ *in* $H_1^\infty(\mathbb{D}^2, B(\mathbb{C}^M, \mathbb{C}^M))$*.*

## 5. Takagi problem

The case where the Pick matrix (3.1) has some negative eigenvalues was first studied by T. Takagi [54], and later by many other authors [1, 42, 20]. See the book [19] for an account. The principal difference is that if one wishes to interpolate with a unimodular function (*i.e.*, a function that has modulus one on the unit circle $\mathbb{T}$), then one has to allow poles inside $\mathbb{D}$. A typical result is

**Theorem 5.1.** *Suppose the Pick matrix is invertible, and has $\pi$ positive eigenvalues and $\nu$ negative eigenvalues. Then there exists a meromorphic interpolating function $\phi$ that is unimodular, and is the quotient of a Blaschke product of degree $\pi$ by a Blaschke product of degree $\nu$.*

If $\Gamma$ is not invertible, the problem is degenerate. It turns out that there is a big difference between solving the problem of finding Blaschke products $f, g$ such that

$$f(\lambda_i) \ = \ w_i g(\lambda_i)$$

and the problem of solving

$$f(\lambda_i)/g(\lambda_i) \ = \ w_i.$$

(The difference occurs if $f$ and $g$ both vanish at some node $\lambda_i$; in the first problem the interpolation condition becomes vacuous, but in the second one needs a relation on the derivatives.) The first problem is more easily handled as the limit of non-degenerate problems; see the paper [23] for recent developments on this approach. The second version of the problem was been solved by H. Woracek using Pontryjagin spaces [59].

**Question 1.** *What is the right version of Theorem* 5.1 *on the bidisk?*

## 6. Interpolating sequences

Given a sequence $\{\lambda_i\}_{i=1}^{\infty}$ in the polydisk $\mathbb{D}^d$, we say it is interpolating for $H^\infty(\mathbb{D}^d)$ if, for any bounded sequence $\{w_i\}_{i=1}^{\infty}$, there is a function $\phi$ in $H^\infty(\mathbb{D}^d)$ satisfying $\phi(\lambda_i) = w_i$. L. Carleson characterized interpolating sequences on $\mathbb{D}$ in [24].

Before stating his theorem, let us introduce some definitions. A *kernel* on $\mathbb{D}^d$ is a positive semi-definite function $k : \mathbb{D}^d \times \mathbb{D}^d \to \mathbb{C}$, *i.e.*, a function such that for any choice of $\lambda_1, \dots, \lambda_N$ in $\mathbb{D}^d$ and any complex numbers $a_1, \dots, a_N$, we have

$$\sum a_i \bar{a}_j k(\lambda_i, \lambda_j) \ \geq \ 0.$$

Given any kernel $k$ on $\mathbb{D}^d$, a sequence $\{\lambda_i\}_{i=1}^{\infty}$ has an associated Grammian $G^k$, where

$$[G^k]_{ij} \ = \ \frac{k(\lambda_i, \lambda_j)}{\sqrt{k(\lambda_i, \lambda_i)\, k(\lambda_j, \lambda_j)}}.$$

We think of $G^k$ as an infinite matrix, representing an operator on $\ell^2$ (that is not necessarily bounded). When $k$ is the Szegő kernel on $\mathbb{D}^d$,

$$k^S(\zeta,\lambda) \;=\; \frac{1}{(1-\zeta^1\bar{\lambda}^1)(1-\zeta^2\bar{\lambda}^2)\cdots(1-\zeta^d\bar{\lambda}^d)}, \tag{6.1}$$

we call the associated Grammian the *Szegő Grammian.* The Szegő kernel is the reproducing kernel for the Hardy space $H^2(\mathbb{D}^d) = P^2(m)$, where $m$ is $d$-dimensional Lebesgue measure on the distinguished boundary $\mathbb{T}^d$ of $\mathbb{D}^d$.

An analogue of the pseudo-hyperbolic metric on the polydisk is the *Gleason distance*, defined by

$$\rho(\zeta,\lambda) \;:=\; \sup\{|\phi(\zeta)| : \|\phi\|_{H^\infty(\mathbb{D}^d)} \leq 1, \phi(\lambda) = 0\}.$$

We shall call a sequence $\{\lambda_i\}_{i=1}^\infty$ *weakly separated* if there exists $\varepsilon > 0$ such that, for all $i \neq j$, the Gleason distance $\rho(\lambda_i,\lambda_j) \geq \varepsilon$. We call the sequence *strongly separated* if there exists $\varepsilon > 0$ such that, for all $i$, there is a function $\phi_i$ in $H_1^\infty(\mathbb{D})$ such that

$$\phi_i(\lambda_j) \;=\; \begin{cases} \varepsilon, & j = i \\ 0, & j \neq i \end{cases}$$

In $\mathbb{D}$, a straightforward argument using Blaschke products shows that a sequence is strongly separated if and only if

$$\prod_{j\neq i} \rho(\lambda_i,\lambda_j) \geq \varepsilon \qquad \forall\ i.$$

We can now state Carleson's theorem. Let us note that he proved it using function theoretic methods, but later H. Shapiro and A. Shields [51] found a Hilbert space approach, which has proved to be more easily generalized, *e.g.*, to characterizing interpolating sequences in the multiplier algebra of the Dirichlet space [39].

**Theorem 6.2.** *On the unit disk, the following are equivalent:*

(1) *There exists $\varepsilon > 0$ such that*

$$\prod_{j\neq i} \rho(\lambda_i,\lambda_j) \geq \varepsilon \qquad \forall\ i.$$

(2) *The sequence $\{\lambda_i\}_{i=1}^\infty$ is an interpolating sequence for $H^\infty(\mathbb{D})$.*

(3) *The sequence $\{\lambda_i\}_{i=1}^\infty$ is weakly separated and the associated Szegő Grammian is a bounded operator on $\ell^2$.*

In 1987 B. Berndtsson, S.-Y. Chang and K.-C. Lin proved the following theorem [22]:

**Theorem 6.3.** *Let $d \geq 2$. Consider the three statements*

(1) *There exists $\varepsilon > 0$ such that*

$$\prod_{j\neq i} \rho(\lambda_i,\lambda_j) \geq \varepsilon \qquad \forall i.$$

(2) *The sequence* $\{\lambda_i\}_{i=1}^{\infty}$ *is an interpolating sequence for* $H^\infty(\mathbb{D}^d)$.

(3) *The sequence* $\{\lambda_i\}_{i=1}^{\infty}$ *is weakly separated and the associated Szegő Grammian is a bounded operator on* $\ell^2$.

*Then* (1) *implies* (2) *and* (2) *implies* (3). *Moreover the converses of these implications are false.*

We call the kernel $k$ on $\mathbb{D}^d$ *admissible* if, for each $1 \leq r \leq d$, the function

$$(1 - \zeta^r \bar{\lambda}^r) k(\zeta, \lambda)$$

is positive semidefinite. (This is the same as saying that multiplication by each coordinate function on the Hilbert function space with reproducing kernel $k$ is a contraction.)

On the unit disk, all admissible kernels are in some sense compressions of the Szegő kernel, and so to prove theorems about $H^\infty(\mathbb{D})$ one can often just use the fact that it is the multiplier algebra of $H^2$. On the bidisk, there is no single dominant kernel, and one must look at a huge family of them. That is the key idea needed in Theorems 2.4 and 3.8, and it allows a different generalization of Theorem 6.2, which was proved in [8]. (If this paragraph seems cryptic, there is a more detailed exposition of this point of view in [9].)

For the following theorem, let $\{e_i\}_{i=1}^{\infty}$ be an orthonormal basis for $\ell^2$.

**Theorem 6.4.** *Let* $\{\lambda_i\}_{i=1}^{\infty}$ *be a sequence in* $\mathbb{D}^2$. *The following are equivalent:*

(i) $\{\lambda_i\}_{i=1}^{\infty}$ *is an interpolating sequence for* $H^\infty(\mathbb{D}^2)$.

(ii) *The following two conditions hold.*

(a) *For all admissible kernels* $k$, *their normalized Grammians are uniformly bounded:*

$$G^k \ \leq \ MI$$

*for some positive constant* $M$.

(b) *For all admissible kernels* $k$, *their normalized Grammians are uniformly bounded below:*

$$NG^k \ \geq \ I$$

*for some positive constant* $N$.

(iii) *The sequence* $\{\lambda_i\}_{i=1}^{\infty}$ *is strongly separated and condition* (a) *alone holds.*

(iv) *Condition* (b) *alone holds.*

*Moreover, Condition* (a) *is equivalent to both* (a′) *and* (a″)*:*

(a′) *There exists a constant* $M$ *and positive semi-definite infinite matrices* $\Gamma^1$ *and* $\Gamma^2$ *such that*

$$M\delta_{ij} - 1 \ = \ \Gamma^1_{ij}(1 - \bar{\lambda}^1_i \lambda^1_j) + \Gamma^2_{ij}(1 - \bar{\lambda}^2_i \lambda^2_j).$$

(a″) *There exists a function* $\Phi$ *in* $H^\infty(\mathbb{D}^2, B(\ell^2, \mathbb{C}))$ *of norm at most* $\sqrt{M}$ *such that* $\Phi(\lambda_i)e_i = 1$.

*Condition* (b) *is equivalent to both* (b′) *and* (b″)*:*

(b′) *There exists a constant $N$ and positive semi-definite infinite matrices $\Delta^1$ and $\Delta^2$ such that*

$$N - \delta_{ij} \;=\; \Delta^1_{ij}(1 - \bar{\lambda}^1_i \lambda^1_j) + \Delta^2_{ij}(1 - \bar{\lambda}^2_i \lambda^2_j).$$

(b″) *There exists a function $\Psi$ in $H^\infty(\mathbb{D}^2, B(\mathbb{C}, \ell^2))$ of norm at most $\sqrt{N}$ such that $\Psi(\lambda_i) = e_i$.*

Neither Theorem 6.3 nor 6.4 are fully satisfactory. For example, the following is still an unsolved problem:

**Question 2.** *If a sequence on $\mathbb{D}^2$ is strongly separated, is it an interpolating sequence?*

## 7. Corona problem

The corona problem on a domain $\Omega$ asks whether, whenever one is given $\phi_1, \dots, \phi_N$ in $H^\infty(\Omega)$ satisfying

$$\sum_{i=1}^N |\phi_i(\lambda)|^2 \;\geq\; \varepsilon > 0, \tag{7.1}$$

there always exist $\psi_1, \dots, \psi_N$ in $H^\infty(\Omega)$ satisfying

$$\sum_{i=1}^N \phi_i \psi_i = 1. \tag{7.2}$$

If the answer is affirmative, the domain is said to have no corona.

Carleson proved that the disk has no corona in [25]. The most striking example of our ignorance about the bidisk is that the answer there is still unknown.

**Question 3.** *Is the corona theorem true for $\mathbb{D}^2$?*

The best result known is due to T. Trent [55], who proved that a solution can be found with the $\psi_i$'s in a specific Orlicz space $\exp(L^{1/3})$, which is contained in $\cap_{p<\infty} H^p(m)$.

There is a version of the corona theorem, the Toeplitz-corona theorem, proved at various levels of generality by several authors [17], [53], [50], [46]. We use $k^S$ as in (6.1) (with $d = 1$).

**Theorem 7.3.** *Let $\phi_1, \dots, \phi_N$ be in $H^\infty(\mathbb{D})$ and $\delta > 0$. Then the following are equivalent:*

(i) *The function*

$$\left[\sum_{i=1}^N \phi_i(\zeta)\overline{\phi_i(\lambda)} - \delta\right] k^S(\zeta, \lambda) \tag{7.4}$$

*is positive semi-definite on $\mathbb{D} \times \mathbb{D}$.*

(ii) *The multipliers $M_{\phi_i}$ on $H^2$ satisfy the inequality*

$$\sum_{i=1}^{N} M_{\phi_i} M_{\phi_i}^* \geq \delta I. \tag{7.5}$$

(iii) *There exist functions $\psi_1, \ldots, \psi_N$ in $H^\infty(\mathbb{D})$ such that*

$$\sum_{i=1}^{N} \psi_i \phi_i = 1 \quad \text{and} \quad \sup_{\lambda \in \mathbb{D}} \left[ \sum_{i=1}^{N} |\psi_i(\lambda)|^2 \right] \leq \frac{1}{\delta}. \tag{7.6}$$

The Toeplitz-corona theorem is often considered a weak version of the corona theorem, because the proof is easier and the hypothesis (7.5) is more stringent than (7.1). It does, however, have a stronger conclusion: condition (iii) gives the exact best bound for the norm of the $\psi_i$'s, whereas the corona theorem asserts that if (7.1) holds, then (iii) holds for *some* $\delta > 0$. (Moreover, in practice, checking the hypothesis (7.5) is an eigenvalue problem, and so quite feasible with polynomial data. Checking (7.1) is a minimization problem over a function on the disk that one would expect to have many local minima, even if the $\phi_i$'s are polynomials of fairly low degree.)

The Toeplitz-corona theorem does generalize to the bidisk, but again it is not enough to check (7.4) for a single kernel (or (7.5) on a single Hilbert function space), but rather one must find a uniform lower bound that works for all admissible kernels. For details see [21, 4].

**Theorem 7.7.** *Let $\phi_1, \ldots, \phi_N$ be in $H^\infty(\mathbb{D}^2)$ and $\delta > 0$. Then the following are equivalent:*

(i) *The function*

$$\left[ \sum_{i=1}^{N} \phi_i(\zeta) \overline{\phi_i(\lambda)} - \delta \right] k(\zeta, \lambda) \tag{7.8}$$

*is positive semi-definite for all admissible kernels $k$.*

(ii) *For every measure $\mu$ on $\mathbb{T}^2$, the multipliers $M_{\phi_i}$ on $P^2(\mu)$ satisfy the inequality*

$$\sum_{i=1}^{N} M_{\phi_i} M_{\phi_i}^* \geq \delta I. \tag{7.9}$$

(iii) *There exist functions $\psi_1, \ldots, \psi_N$ in $H^\infty(\mathbb{D}^2)$ such that*

$$\sum_{i=1}^{N} \psi_i \phi_i = 1 \qquad \text{and} \qquad \sup_{\lambda \in \mathbb{D}^2} \left[ \sum_{i=1}^{N} |\psi_i(\lambda)|^2 \right] \leq \frac{1}{\delta}. \tag{7.10}$$

Although Theorem 7.7 seems to depend on the specific properties of the bidisk, (indeed, using Theorem 2.4 one can prove the equivalence of (i) and (iii) in the Schur-Agler norm on the polydisk), there is a remarkable generalization by E. Amar that applies not only to the polydisk, but to any smooth convex domain [14].

**Theorem 7.11.** *Let $\Omega$ be a bounded convex domain in $\mathbb{C}^d$ containing the origin, and assume that either $\Omega$ is $\mathbb{D}^d$ or its boundary is smooth. Let $X$ be $\mathbb{T}^d$ in the former case, the boundary of $\Omega$ in the latter. Let $\phi_1, \dots, \phi_N$ be in $H^\infty(\Omega)$ and $\delta > 0$. Then the following are equivalent:*

(i) *There exist functions $\psi_1, \dots, \psi_N$ in $H^\infty(\mathbb{D})$ such that*

$$\sum_{i=1}^N \psi_i \phi_i = 1 \qquad \text{and} \qquad \sup_{\lambda \in \Omega} \left[ \sum_{i=1}^N |\psi_i(\lambda)|^2 \right] \leq \frac{1}{\delta}.$$

(ii) *For every measure $\mu$ on $X$, the multipliers $M_{\phi_i}$ on $P^2(\mu)$ satisfy the inequality*

$$\sum_{i=1}^N M_{\phi_i} M_{\phi_i}^* \geq \delta I. \tag{7.12}$$

(iii) *For every measure $\mu$ on $X$, and every $f$ in $P^2(\mu)$, there exist functions $\psi_1, \dots, \psi_N$ in $P^2(\mu)$ such that*

$$\sum_{i=1}^N \psi_i \phi_i = f \qquad \text{and} \qquad \sum_{i=1}^N \|\psi_i\|^2 \leq \frac{1}{\delta} \|f\|^2, \tag{7.13}$$

*where the norms on both sides of* (7.13) *are in $P^2(\mu)$.*

In [56], T. Trent and B. Wick have shown that in Amar's theorem it is sufficient to consider measures $\mu$ that are absolutely continuous and whose derivatives are bounded away from zero.

## 8. Distinguished and toral varieties

A Pick problem is called *extremal* if it is solvable with a function of norm 1, but not with anything smaller. In one dimension, this forces the solution to be unique. (In the notation of Section 3, this corresponds to the Pick matrix (3.1) being singular, the vectors in (3.6) spanning $\mathbb{C}^{1+M}$, and the unique solution being (3.7).) On the bidisk, problems can be extremal in either one or two dimensions. For example, consider the problems: $w_1 = 0, w_2 = 1/2$, $\lambda_1 = (0,0)$ and $\lambda_2$ either $(1/2, 0)$ or $(1/2, 1/2)$. The first problem has the unique solution $z$; the latter problem has a unique solution on the one-dimensional set $\{z = w\}$, but is not unique off this set. Indeed, Theorem 4.3 says in this case that the general solution is given by

$$\phi(z,w) \;=\; tz + (1-t)w + t(1-t)(z-w)^2 \frac{\Psi}{1 - [(1-t)z + tw]\Psi},$$

where $\Psi$ is any function in $H_1^\infty(\mathbb{D}^2)$ and $t$ is any number in $[0,1]$.

If an extremal Pick problem on $\mathbb{D}^2$ does not have a solution that is unique on all $\mathbb{D}^2$, then the set on which it is unique must be a variety[2] (the zero set of a polynomial). But this is not an arbitrary variety – it has special properties.

[2] We use the word variety where algebraic geometers would say algebraic set – *i.e.*, we do not require that a variety be irreducible.

Let $\mathbb{E}$ be the exterior of the closed disk, $\mathbb{C} \setminus \overline{\mathbb{D}}$. Say a variety $V$ in $\mathbb{C}^2$ is *toral* if every irreducible component intersects $\mathbb{T}^2$ in an infinite set, and say it is *distinguished* if

$$V \subset \mathbb{D}^2 \cup \mathbb{T}^2 \cup \mathbb{E}^2.$$

Distinguished varieties first appeared implicitly in the paper [48] by W. Rudin, and later in the operator theoretic context of sharpening Andô's inequality for matrices [11]; they turn out to be intimately connected to function theory on $\mathbb{D}^2$ (see Theorem 9.2, for example). Toral varieties are related to inner functions [12] and to symmetry of a variety with respect to the torus [13]. The uniqueness variety was partially described in [11, 12]:

**Theorem 8.1.** *The uniqueness set for an extremal Pick problem on $\mathbb{D}^2$ is either all of $\mathbb{D}^2$ or a toral variety. In the latter case, it contains a distinguished variety.*

It is perhaps the case that the uniqueness set is all of $\mathbb{D}^2$ whenever the data is in some sense "generic" (see, *e.g.*, [7]), but how is that made precise?

**Question 4.** *When is the uniqueness set all of $\mathbb{D}^2$?*

Distinguished varieties have a determinantal representation. The following theorem was proved in [11], and, more constructively, in [36].

**Theorem 8.2.** *A variety $V$ is a distinguished variety if and only if there is a pure matrix-valued rational inner function $\Psi$ on the disk such that*

$$V \cap \mathbb{D}^2 \ = \ \{(z,w) \ \in \ \mathbb{D}^2 : \det(\Psi(z) - wI) = 0\}.$$

Another way to picture distinguished varieties is by taking the Cayley transform of both variables; then they become varieties in $\mathbb{C}^2$ with the property that when one coordinate is real, so is the other.

## 9. Extension property

If $G$ is a subset of $\mathbb{D}^2$, we shall say that a function $f$ defined on $G$ is holomorphic if, for every point $P$ in $G$, there is an open ball $B(P, \varepsilon)$ in $\mathbb{D}^2$ and a holomorphic function on the ball whose restriction to $G$ is $f$. Given such a holomorphic function $f$, one can ask whether there is a single function $F$ on $\mathbb{D}^2$ that extends it, and, if so, whether $F$ can be chosen with additional properties.

H. Cartan proved that if $G$ is a subvariety, then a global extension $F$ always exists [26] (indeed he proved this on any pseudo-convex domain, the bidisk being just a special case). If $f$ is bounded, one can ask whether one can find an extension $F$ with the same sup-norm. If $G$ is an analytic retract of $\mathbb{D}^2$, *i.e.*, there is an analytic map $r : \mathbb{D}^2 \to G$ that is the identity on $G$, then $F = f \circ r$ will work. (All retracts of $\mathbb{D}^2$ are either singletons, embedded disks, or the whole bidisk [47].) It turns out that extending without increasing the norm is only possible for retracts [10].

**Theorem 9.1.** *Let $G \subseteq \mathbb{D}^2$ and assume that $G$ is relatively polynomially convex (i.e., $G^\wedge \cap \mathbb{D}^2 = G$ where $G^\wedge$ denotes the polynomially convex hull of $G$). If every polynomial $f$ on $G$ has an extension to a function $F$ in $H^\infty(\mathbb{D}^2)$ of the same norm, then $G$ is a retract.*

Let us remark that although the theorem can be proved without using operator theory, it was discovered by studying pairs of commuting operators having $G$ as a spectral set.

One can also ask if a bounded function $f$ has a bounded extension $F$, but with a perhaps greater norm. G. Henkin and P. Polyakov proved that this can always be done if $G$ is a subvariety of the polydisk that exits transversely [35]. In the case of a distinguished variety, Knese showed how to bound the size of the extension even when there are singularities on $\mathbb{T}^2$ [36, 37] (he also gives a construction of the function $C$ below):

**Theorem 9.2.** *Let $V$ be a distinguished variety. Then there is a rational function $C(z)$ with no poles in $\mathbb{D}$ such that, for every polynomial $f(z,w)$ there is a rational function $F$ which agrees with $f$ on $V \cap \mathbb{D}^2$ and satisfies the estimate*

$$|F(z,w)| \ \leq \ |C(z)| \sup_{(z,w) \ \in \ V} |f(z)|.$$

*If $V$ has no singularities on $\mathbb{T}^2$, then $C$ can be taken to be a constant.*

## 10. Conclusion

Paul R. Halmos contributed in many ways to the development of operator theory. The purpose of this article is to show that recasting many known results about $H^\infty(\mathbb{D})$ in terms of operator theory has been extremely fruitful in understanding $H^\infty(\mathbb{D}^2)$. So far, however, it has not helped very much in understanding $H^\infty(\mathbb{B}_2)$, where $\mathbb{B}_2$ is the ball in $\mathbb{C}^2$. There is another kernel on the ball,

$$k(\zeta,\lambda) \ = \ \frac{1}{1-\zeta^1\bar{\lambda}^1-\zeta^2\bar{\lambda}^2},$$

introduced by S. Drury [31], and operator theory has been very effective in studying this kernel [5, 6, 18, 28, 32, 45].

**Question 5.** *What is the correct Pick theorem on $H^\infty(\mathbb{B}_2)$?*

## References

[1] V.M. Adamian, D.Z. Arov, and M.G. Kreĭn. Analytic properties of Schmidt pairs for a Hankel operator and the generalized Schur-Takagi problem. *Math. USSR. Sb.*, 15:31–73, 1971.

[2] J. Agler. Some interpolation theorems of Nevanlinna-Pick type. Preprint, 1988.

[3] J. Agler. On the representation of certain holomorphic functions defined on a polydisc. In *Operator Theory: Advances and Applications, Vol.* 48, pages 47–66. Birkhäuser, Basel, 1990.

[4] J. Agler and J.E. M$^{c}$Carthy. Nevanlinna-Pick interpolation on the bidisk. *J. Reine Angew. Math.*, 506:191–204, 1999.

[5] J. Agler and J.E. M$^{c}$Carthy. Complete Nevanlinna-Pick kernels. *J. Funct. Anal.*, 175(1):111–124, 2000.

[6] J. Agler and J.E. M$^{c}$Carthy. Nevanlinna-Pick kernels and localization. In A. Gheondea, R.N. Gologan, and D. Timotin, editors, *Proceedings of* 17*th International Conference on Operator Theory at Timisoara, 1998*, pages 1–20. Theta Foundation, Bucharest, 2000.

[7] J. Agler and J.E. M$^{c}$Carthy. The three point Pick problem on the bidisk. *New York Journal of Mathematics*, 6:227–236, 2000.

[8] J. Agler and J.E. M$^{c}$Carthy. Interpolating sequences on the bidisk. *International J. Math.*, 12(9):1103–1114, 2001.

[9] J. Agler and J.E. M$^{c}$Carthy. *Pick Interpolation and Hilbert Function Spaces.* American Mathematical Society, Providence, 2002.

[10] J. Agler and J.E. M$^{c}$Carthy. Norm preserving extensions of holomorphic functions from subvarieties of the bidisk. *Ann. of Math.*, 157(1):289–312, 2003.

[11] J. Agler and J.E. M$^{c}$Carthy. Distinguished varieties. *Acta Math.*, 194:133–153, 2005.

[12] J. Agler, J.E. M$^{c}$Carthy, and M. Stankus. Toral algebraic sets and function theory on polydisks. *J. Geom. Anal.*, 16(4):551–562, 2006.

[13] J. Agler, J.E. M$^{c}$Carthy, and M. Stankus. Geometry near the torus of zero-sets of holomorphic functions. *New York J. Math.*, 14:517–538, 2008.

[14] E. Amar. On the Toeplitz-corona problem. *Publ. Mat.*, 47(2):489–496, 2003.

[15] J.M. Anderson, M. Dritschel, and J. Rovnyak. Shwarz-Pick inequalities for the Schur-Agler class on the polydisk and unit ball. *Comput. Methods Funct. Theory*, 8:339–361, 2008.

[16] T. Andô. On a pair of commutative contractions. *Acta Sci. Math.* (*Szeged*), 24:88–90, 1963.

[17] W.B. Arveson. Interpolation problems in nest algebras. *J. Funct. Anal.*, 20:208–233, 1975.

[18] W.B. Arveson. Subalgebras of C*-algebras III: Multivariable operator theory. *Acta Math.*, 181:159–228, 1998.

[19] J.A. Ball, I. Gohberg, and L. Rodman. *Interpolation of rational matrix functions.* Birkhäuser, Basel, 1990.

[20] J.A. Ball and J.W. Helton. A Beurling-Lax theorem for the Lie group $U(m,n)$ which contains most classical interpolation theory. *Integral Equations and Operator Theory*, 9:107–142, 1983.

[21] J.A. Ball and T.T. Trent. Unitary colligations, reproducing kernel Hilbert spaces, and Nevanlinna-Pick interpolation in several variables. *J. Funct. Anal.*, 197:1–61, 1998.

[22] B. Berndtsson, S.-Y. Chang, and K.-C. Lin. Interpolating sequences in the polydisk. *Trans. Amer. Math. Soc.*, 302:161–169, 1987.

[23] V. Bolotnikov, A. Kheifets, and L. Rodman. Nevanlinna-Pick interpolation: Pick matrices have bounded number of negative eigenvalues. *Proc. Amer. Math. Soc.*, 132:769–780, 2003.

[24] L. Carleson. An interpolation problem for bounded analytic functions. *Amer. J. Math.*, 80:921–930, 1958.

[25] L. Carleson. Interpolations by bounded analytic functions and the corona problem. *Ann. of Math.*, 76:547–559, 1962.

[26] H. Cartan. *Séminaire Henri Cartan* 1951/2. W.A. Benjamin, New York, 1967.

[27] J.B. Conway. *The Theory of Subnormal Operators.* American Mathematical Society, Providence, 1991.

[28] S. Costea, E.T. Sawyer, and B.D. Wick. The corona theorem for the Drury-Arveson Hardy space and other holomorphic Besov-Sobolev spaces on the unit ball in $\mathbb{C}^n$. http://front.math.ucdavis.edu/0811.0627, to appear.

[29] C.C. Cowen and B.D. MacCluer. *Composition operators on spaces of analytic functions.* CRC Press, Boca Raton, 1995.

[30] M.J. Crabb and A.M. Davie. Von Neumann's inequality for Hilbert space operators. *Bull. London Math. Soc.*, 7:49–50, 1975.

[31] S.W. Drury. A generalization of von Neumann's inequality to the complex ball. *Proc. Amer. Math. Soc.*, 68:300–304, 1978.

[32] D. Greene, S. Richter, and C. Sundberg. The structure of inner multipliers on spaces with complete Nevanlinna Pick kernels. *J. Funct. Anal.*, 194:311–331, 2002.

[33] A. Grinshpan, D. Kaliuzhnyi-Verbovetskyi, V. Vinnikov, and H. Woerdeman. Classes of tuples of commuting contractions satisfying the multivariable von Neumann inequality. *J. Funct. Anal.* 256 (2009), no. 9, 3035–3054.

[34] P.R. Halmos. Normal dilations and extensions of operators. *Summa Brasil. Math.*, 2:125–134, 1950.

[35] G.M. Henkin and P.L. Polyakov. Prolongement des fonctions holomorphes bornées d'une sous-variété du polydisque. *Comptes Rendus Acad. Sci. Paris Sér. I Math.*, 298(10):221–224, 1984.

[36] G. Knese. Polynomials defining distinguished varieties. To appear in *Trans. Amer. Math. Soc.*

[37] G. Knese. Polynomials with no zeros on the bidisk. To appear in *Analysis and PDE.*

[38] G. Knese. A Schwarz lemma on the polydisk. *Proc. Amer. Math. Soc.*, 135:2759–2768, 2007.

[39] D. Marshall and C. Sundberg. Interpolating sequences for the multipliers of the Dirichlet space. Preprint; see http://www.math.washington.edu/∼marshall/preprints/preprints.html, 1994.

[40] R. Nevanlinna. Über beschränkte Funktionen. *Ann. Acad. Sci. Fenn. Ser. A*, 32(7):7–75, 1929.

[41] N.K. Nikol'skiĭ. *Operators, functions and systems: An easy reading.* AMS, Providence, 2002.

[42] A.A. Nudelman. On a new type of moment problem. *Dokl. Akad. Nauk. SSSR.*, 233:5:792–795, 1977.

[43] V.V. Peller. *Hankel operators and their applications.* Springer, New York, 2002.

[44] G. Pick. Über die Beschränkungen analytischer Funktionen, welche durch vorgegebene Funktionswerte bewirkt werden. *Math. Ann.*, 77:7–23, 1916.

[45] G. Popescu. Von Neumann inequality for $(B(\mathcal{H})^n)_1$. *Math. Scand.*, 68:292–304, 1991.

[46] M. Rosenblum. A corona theorem for countably many functions. *Integral Equations and Operator Theory*, 3(1):125–137, 1980.

[47] W. Rudin. *Function Theory in Polydiscs.* Benjamin, New York, 1969.

[48] W. Rudin. Pairs of inner functions on finite Riemann surfaces. *Trans. Amer. Math. Soc.*, 140:423–434, 1969.

[49] D. Sarason. Generalized interpolation in $H^\infty$. *Trans. Amer. Math. Soc.*, 127:179–203, 1967.

[50] C.F. Schubert. The corona theorem as an operator theorem. *Proc. Amer. Math. Soc.*, 69:73–76, 1978.

[51] H.S. Shapiro and A.L. Shields. On some interpolation problems for analytic functions. *Amer. J. Math.*, 83:513–532, 1961.

[52] B. Szokefalvi-Nagy and C. Foiaş. Commutants de certains opérateurs. *Acta Sci. Math. (Szeged)*, 29:1–17, 1968.

[53] B. Szokefalvi-Nagy and C. Foiaş. On contractions similar to isometries and Toeplitz operators. *Ann. Acad. Sci. Fenn. Ser. AI Math.*, 2:553–564, 1976.

[54] T. Takagi. On an algebraic problem related to an analytic theorem of Carathéodory and Fejer. *Japan J. Math.*, 1:83–93, 1929.

[55] T.T. Trent. A vector-valued $H^p$ corona theorem on the polydisk. *Integral Equations and Operator Theory*, 56:129–149, 2006.

[56] T.T. Trent and B.D. Wick. Toeplitz corona theorems for the polydisk and the unit ball. http://front.math.ucdavis.edu/0806.3428, to appear.

[57] N.Th. Varopoulos. On an inequality of von Neumann and an application of the metric theory of tensor products to operators theory. *J. Funct. Anal.*, 16:83–100, 1974.

[58] J. von Neumann. Eine Spektraltheorie für allgemeine Operatoren eines unitären Raumes. *Math. Nachr.*, 4:258–281, 1951.

[59] H. Woracek. An operator theoretic approach to degenerated Nevanlinna-Pick interpolation. *Math. Nachr.*, 176:335–350, 1995.

Jim Agler
U.C. San Diego
La Jolla, CA 92093, USA

John E. M$^{\text{c}}$Carthy
Washington University
St. Louis, MO 63130, USA
e-mail: `mccarthy@math.wustl.edu`

Operator Theory:
Advances and Applications, Vol. 207, 99–123
© 2010 Springer Basel AG

# Dilation Theory Yesterday and Today

William Arveson

**Abstract.** Paul Halmos' work in dilation theory began with a question and its answer: Which operators on a Hilbert space $H$ can be extended to normal operators on a larger Hilbert space $K \supseteq H$? The answer is interesting and subtle.

The idea of representing operator-theoretic structures in terms of conceptually simpler structures acting on larger Hilbert spaces has become a central one in the development of operator theory and, more generally, noncommutative analysis. The work continues today. This article summarizes some of these diverse results and their history.

**Mathematics Subject Classification (2000).** 46L07.

**Keywords.** Dilation theory, completely positive linear maps.

## 1. Preface

What follows is a brief account of the development of dilation theory that highlights Halmos' contribution to the circle of ideas. The treatment is not comprehensive. I have chosen topics that have interested me over the years, while perhaps neglecting others. In order of appearance, the cast includes dilation theory for subnormal operators, operator-valued measures and contractions, connections with the emerging theory of operator spaces, the role of extensions in dilation theory, commuting sets of operators, and semigroups of completely positive maps. I have put Stinespring's theorem at the center of it, but barely mention the model theory of Sz.-Nagy and Foias or its application to systems theory.

After reflection on the common underpinnings of these results, it seemed a good idea to feature the role of Banach $*$-algebras in their proofs, and I have done that. An appendix is included that summarizes what is needed. Finally, I have tried to make the subject accessible to students by keeping the prerequisites to a minimum; but of course familiarity with the basic theory of operators on Hilbert spaces and $C^*$-algebras is necessary.

## 2. Origins

Hilbert spaces are important because positive definite functions give rise to inner products on vector spaces – whose completions are Hilbert spaces – and positive definite functions are found in every corner of mathematics and mathematical physics. This association of a Hilbert space with a positive definite function involves a construction, and like all constructions that begin with objects in one category and generate objects in another category, it is best understood when viewed as a *functor*. We begin by discussing the properties of this functor in some detail since, while here they are simple and elementary, similar properties will re-emerge later in other contexts.

Let $X$ be a set and let

$$u : X \times X \to \mathbb{C}$$

be a complex-valued function of two variables that is *positive definite* in the sense that for every $n = 1, 2, \dots,$ every $x_1, \dots, x_n \in X$ and every set $\lambda_1, \dots, \lambda_n$ of complex numbers, one has

$$\sum_{k,j=1}^{n} u(x_k, x_j)\lambda_j \bar{\lambda}_k \geq 0. \tag{2.1}$$

Notice that if $f : X \to H$ is a function from $X$ to a Hilbert space $H$ with inner product $\langle \cdot, \cdot \rangle$, then the function $u : X \times X \to \mathbb{C}$ defined by

$$u(x, y) = \langle f(x), f(y) \rangle, \qquad x, y \in X \tag{2.2}$$

is positive definite. By passing to a subspace of $H$ if necessary, one can obviously arrange that $H$ is the closed linear span of the set of vectors $f(X)$ in the range of $f$, and in that case the function $f : X \to H$ is said to be *minimal* (for the positive definite function $u$). Let us agree to say that two Hilbert space-valued functions $f_1 : X \to H_1$ and $f_2 : X \to H_2$ are *isomorphic* if there is a unitary operator $U : H_1 \to H_2$ such that

$$U(f_1(x)) = f_2(x), \qquad x \in X.$$

A simple argument shows that all minimal functions for $u$ are isomorphic.

For any positive definite function $u : X \times X \to \mathbb{C}$, a self-map $\phi : X \to X$ may or may not preserve the values of $u$ in the sense that

$$u(\phi(x), \phi(y)) = u(x, y), \qquad x, y \in X;$$

but when this formula does hold, one would expect that $\phi$ should acquire a Hilbert space interpretation. In order to discuss that, let us think of Hilbert spaces as the objects of a category whose morphisms are isometries; thus, a homomorphism from $H_1$ to $H_2$ is a linear isometry $U \in \mathcal{B}(H_1, H_2)$. Positive definite functions are also the objects of a category, in which a homomorphism from $u_1 : X_1 \times X_1 \to \mathbb{C}$ to $u_2 : X_2 \times X_2 \to \mathbb{C}$ is a function $\phi : X_1 \to X_2$ that preserves the positive structure in the sense that

$$u_2(\phi(x), \phi(y)) = u_1(x, y), \qquad x, y \in X_1. \tag{2.3}$$

Given a positive definite function $u : X \times X \to \mathbb{C}$, one can construct a Hilbert space $H(u)$ and a function $f : X \to H(u)$ as follows. Consider the vector space $\mathbb{C}X$ of all complex-valued functions $\xi : X \to \mathbb{C}$ with the property that $\xi(x) = 0$ for all but a finite number of $x \in X$. We can define a sesquilinear form $\langle \cdot, \cdot \rangle$ on $\mathbb{C}X$ by way of

$$\langle \xi, \eta \rangle = \sum_{x,y \in X} u(x,y)\xi(x)\bar{\eta}(y), \qquad \xi, \eta \in \mathbb{C}X,$$

and one finds that $\langle \cdot, \cdot \rangle$ is positive semidefinite because of the hypothesis on $u$. An application of the Schwarz inequality shows that the set

$$N = \{\xi \in \mathbb{C}X : \langle \xi, \xi \rangle = 0\}$$

is in fact a linear subspace of $\mathbb{C}X$, so this sesquilinear form can be promoted naturally to an inner product on the quotient $\mathbb{C}X/N$. The completion of the inner product space $\mathbb{C}X/N$ is a Hilbert space $H(u)$, and we can define the sought-after function $f : X \to H(u)$ as follows:

$$f(x) = \delta_x + N, \qquad x \in X, \tag{2.4}$$

where $\delta_x$ is the characteristic function of the singleton $\{x\}$. By construction, $u(x,y) = \langle f(x), f(y) \rangle$. Note too that this function $f$ is *minimal* for $u$. While there are many (mutually isomorphic) minimal functions for $u$, we fix attention on the minimal function (2.4) that has been constructed.

Given two positive definite functions $u_k : X_k \times X_k \to \mathbb{C}$, $k = 1, 2$, choose a homomorphism from $u_1$ to $u_2$, namely a function $\phi : X_1 \to X_2$ that satisfies (2.3). Notice that while the two functions $f_k : X_k \to H(u_k)$

$$f_1(x) = \delta_x + N_1, \quad f_2(y) = \delta_y + N_2, \qquad x \in X_1, \quad y \in X_2$$

need not be injective, we do have the relations

$$\langle f_2(\phi(x)), f_2(\phi(y)) \rangle_{H(u_2)} = u_2(\phi(x), \phi(y)) = u_1(x,y) = \langle f_1(x), f_1(y) \rangle_{H(u_1)},$$

holding for all $x, y \in X_1$. Since $H(u_1)$ is spanned by $f_1(X_1)$, a familiar and elementary argument (that we omit) shows that there is a unique linear isometry $U_\phi : H(u_1) \to H(u_2)$ such that

$$U_\phi(f_1(x)) = f_2(\phi(x)), \qquad x \in X_1. \tag{2.5}$$

At this point, it is straightforward to verify that the expected composition formulas $U_{\phi_1} U_{\phi_2} = U_{\phi_1 \circ \phi_2}$ hold in general, and we conclude:

**Proposition 2.1.** *The construction* (2.4) *gives rise to a covariant functor* $(u, \phi) \to (H(u), U_\phi)$ *from the category of positive definite functions on sets to the category of complex Hilbert spaces.*

It is significant that if $X$ is a topological space and $u : X \times X \to \mathbb{C}$ is a *continuous* positive definite function, then the associated map $f : X \to H(u)$ of (2.4) is also continuous. Indeed, this is immediate from (2.2):

$$\|f(x) - f(y)\|^2 = u(x,x) + u(y,y) - u(x,y) - u(y,x), \qquad x, y \in X.$$

The functorial nature of Proposition 2.1 pays immediate dividends:

*Remark* 2.2 (Automorphisms). Every positive definite function

$$u : X \times X \to \mathbb{C}$$

has an associated group of internal symmetries, namely the group $G_u$ of all bijections $\phi : X \to X$ that preserve $u$ in the sense that

$$u(\phi(x), \phi(y)) = u(x, y), \qquad x, y \in X.$$

Notice that Proposition 2.1 implies that this group of symmetries has a natural unitary representation $U : G_u \to \mathcal{B}(H(u))$ associated with it. Indeed, for every $\phi \in G_u$, the unitary operator $U_\phi \in \mathcal{B}(H(u))$ is defined uniquely by

$$U_\phi(f(x)) = f(\phi(x)), \qquad x \in X.$$

The properties of this unitary representation of the automorphism group of $u$ often reflect important features of the environment that produced $u$.

**Examples:** There are many examples of positive definite functions; some of the more popular are reproducing kernels associated with domains in $\mathbb{C}^n$. Here is another example that is important for quantum physics and happens to be one of my favorites. Let $Z$ be a (finite- or infinite-dimensional) Hilbert space and consider the positive definite function $u : Z \times Z \to \mathbb{C}$ defined by

$$u(z, w) = e^{\langle z, w \rangle}, \qquad z, w \in Z.$$

We will write the Hilbert space $H(u)$ defined by the construction of Proposition 2.1 as $e^Z$, since it can be identified as the symmetric Fock space over the one-particle space $Z$. We will not make that identification here, but we do write the natural function (2.4) from $Z$ to $e^Z$ as $f(z) = e^z$, $z \in Z$.

One finds that the automorphism group of Remark 2.2 is the full unitary group $\mathcal{U}(Z)$ of $Z$. Hence the functorial nature of the preceding construction leads immediately to a (strongly continuous) unitary representation $\Gamma$ of the unitary group $\mathcal{U}(Z)$ on the Hilbert space $e^Z$. In explicit terms, for $U \in \mathcal{U}(Z)$, $\Gamma(U)$ is the unique unitary operator on $e^Z$ that satisfies

$$\Gamma(U)(e^z) = e^{Uz}, \qquad U \in \mathcal{U}(Z), \quad z \in Z.$$

The map $\Gamma$ is called *second quantization* in the physics literature. It has the property that for every one-parameter unitary group $\{U_t : t \in \mathbb{R}\}$ acting on $Z$, there is a corresponding one-parameter unitary group $\{\Gamma(U_t) : t \in \mathbb{R}\}$ that acts on the "first quantized" Hilbert space $e^Z$. Equivalently, for every self-adjoint operator $A$ on $Z$, there is a corresponding "second quantized" self-adjoint operator $d\Gamma(A)$ on $e^Z$ that is uniquely defined by the formula

$$e^{itd\Gamma(A))} = \Gamma(e^{itA}), \qquad t \in \mathbb{R},$$

as one sees by applying Stone's theorem which characterizes the generators of strongly continuous one-parameter unitary groups.

Finally, one can exploit the functorial nature of this construction further to obtain a natural representation of the canonical commutation relations on $e^Z$, but we will not pursue that here.

## 3. Positive linear maps on commutative $*$-algebras

The results of Sections 4 and 5 on subnormal operators, positive operator-valued measures and the dilation theory of contractions can all be based on a single dilation theorem for positive linear maps of commutative Banach $*$-algebras. That commutative theorem has a direct commutative proof. But since we require a more general noncommutative dilation theorem in Section 6 that contains it as a special case, we avoid repetition by merely stating the commutative result in this section. What we want to emphasize here is the unexpected appearance of complete positivity even in this commutative context, and the functorial nature of dilation theorems of this kind.

A Banach $*$-algebra is a Banach algebra $\mathcal{A}$ that is endowed with an isometric involution – an antilinear mapping $a \mapsto a^*$ of $\mathcal{A}$ into itself that satisfies $a^{**} = a$, $(ab)^* = b^*a^*$ and $\|a^*\| = \|a\|$. In this section we will be concerned with Banach $*$-algebras that are *commutative*, and which have a multiplicative unit $\mathbf{1}$ that satisfies $\|\mathbf{1}\| = 1$. The basic properties of Banach $*$-algebras and their connections with $C^*$-algebras are summarized in the appendix.

An operator-valued linear map $\phi : \mathcal{A} \to \mathcal{B}(H)$ of a Banach $*$-algebra is said to be *positive* if $\phi(a^*a) \geq 0$ for every $a \in \mathcal{A}$. The most important fact about operator-valued positive linear maps of commutative algebras is something of a miracle. It asserts that a positive linear map $\phi : \mathcal{A} \to \mathcal{B}(H)$ of a commutative Banach $*$-algebra $\mathcal{A}$ is *completely* positive in the following sense: For every $n$-tuple $a_1, \dots, a_n$ of elements of $\mathcal{A}$, the $n \times n$ operator matrix $(\phi(a_i^* a_j))$ is positive in the natural sense that for every $n$-tuple of vectors $\xi_1, \dots, \xi_n \in H$, one has

$$\sum_{i,j=1}^{n} \langle \phi(a_i^* a_j)\xi_j, \xi_i \rangle \geq 0. \tag{3.1}$$

Notice that the hypothesis $\phi(a^*a) \geq 0$ is the content of these inequalities for the special case $n = 1$. This result for commutative $C^*$-algebras $\mathcal{A}$ is due to Stinespring (see Theorem 4 of [Sti55]), and the proof of (3.1) can be based on that result combined with the properties of the completion map $\iota : \mathcal{A} \to C^*(\mathcal{A})$ that carries a commutative Banach $*$-algebra $\mathcal{A}$ to its enveloping $C^*$-algebra $C^*(\mathcal{A}) \cong C(X)$ (see Remark A.3 of the appendix).

The notion of complete positivity properly belongs to the noncommutative world. We will return to it in Section 6 where we will prove a general result (Theorem 6.1) which, when combined with (3.1), implies the following assertion about positive linear maps of commutative $*$-algebras.

**Scholium A:** *Let $\mathcal{A}$ be a commutative Banach $*$-algebra with unit and let $H$ be a Hilbert space. For every operator-valued linear map $\phi : \mathcal{A} \to \mathcal{B}(H)$ satisfying*

*$\phi(a^*a) \geq 0$ for all $a \in \mathcal{A}$, there is a pair $(V, \pi)$ consisting of a representation $\pi : \mathcal{A} \to \mathcal{B}(K)$ of $\mathcal{A}$ on another Hilbert space $K$ and a linear operator $V \in \mathcal{B}(H, K)$ such that*

$$\phi(a) = V^*\pi(a)V, \qquad a \in \mathcal{A}. \tag{3.2}$$

*Moreover, $\phi$ is necessarily bounded, its norm is given by*

$$\sup_{\|a\|\leq 1} \|\phi(a)\| = \|\phi(\mathbf{1})\| = \|V\|^2, \tag{3.3}$$

*and $V$ can be taken to be an isometry when $\phi(\mathbf{1}) = \mathbf{1}$.*

*Remark* 3.1 (Minimality and uniqueness of dilation pairs). Fix $\mathcal{A}$ as above. By a *dilation pair* for $\mathcal{A}$ we mean a pair $(V, \pi)$ consisting of a representation $\pi : \mathcal{A} \to \mathcal{B}(K)$ and a bounded linear map $V : H \to K$ from some other Hilbert space $H$ into the space $K$ on which $\pi$ acts. A dilation pair $(V, \pi)$ is said to be *minimal* if the set of vectors $\{\pi(a)V\xi : a \in \mathcal{A},\ \xi \in H\}$ has $K$ as its closed linear span. By replacing $K$ with an appropriate subspace and $\pi$ with an appropriate subrepresentation, we can obviously replace every such pair with a minimal one. Moreover, the representation associated with a minimal pair must be nondegenerate, and therefore $\pi(\mathbf{1}) = \mathbf{1}_K$.

Note that every dilation pair $(V, \pi)$ gives rise to a positive linear map $\phi : \mathcal{A} \to \mathcal{B}(H)$ that is defined by the formula (3.2), and we say that $(V, \pi)$ is a dilation pair *for* $\phi$. A positive map $\phi$ has many dilation pairs associated with it, but the minimal ones are equivalent in the following sense: If $(V_1, \pi_1)$ and $(V_2, \pi_2)$ are two *minimal* dilation pairs for $\phi$ then there is a unique unitary operator $W : K_1 \to K_2$ such that

$$WV_1 = V_2, \quad \text{and} \quad W\pi_1(a) = \pi_2(a)W, \qquad a \in \mathcal{A}. \tag{3.4}$$

The proof amounts to little more than checking inner products on the two generating sets $\pi_1(\mathcal{A})V_1H \subseteq K_1$ and $\pi_2(\mathcal{A})V_2H \subseteq K_2$ and noting that

$$\begin{aligned}\langle \pi_2(a)V_2\xi, \pi_2(b)V_2\eta\rangle &= \langle \pi_2(b^*a)V_2\xi, V_2\eta\rangle = \langle \phi(b^*a)\xi, \eta\rangle \\ &= \langle \pi_1(a)V_1\xi, \pi_1(b)V_1\eta\rangle,\end{aligned}$$

for $a, b \in \mathcal{A}$ and $\xi, \eta \in H$.

Finally, note that in cases where $\phi(\mathbf{1}) = \mathbf{1}$, the operator $V$ of a minimal pair $(V, \pi)$ is an isometry, so by making an obvious identification we can replace $(V, \pi)$ with an equivalent one in which $V$ is the inclusion map of $H$ into a larger Hilbert space $\iota : H \subseteq K$ and $\pi$ is a representation of $\mathcal{A}$ on $K$. After these identifications we find that $V^* = P_H$, and (3.2) reduces to the more traditional assertion

$$\phi(a) = P_H\pi(a) \restriction_H, \qquad a \in \mathcal{A}. \tag{3.5}$$

*Remark* 3.2 (Functoriality). It is a worthwhile exercise to think carefully about what a dilation actually *is*, and the way to do that is to think in categorical terms. Fix a commutative Banach $*$-algebra $\mathcal{A}$ with unit $\mathbf{1}$. Operator-valued positive linear maps of $\mathcal{A}$ are the objects of a category, in which a homomorphism from $\phi_1 : \mathcal{A} \to \mathcal{B}(H_1)$ to $\phi_2 : \mathcal{A} \to \mathcal{B}(H_2)$ is defined as a unitary operator $U : H_1 \to H_2$ satisfying $U\phi_1(a) = \phi_2(a)U$ for all $a \in \mathcal{A}$; equivalently, $U$ should implement a unitary equivalence of positive linear maps of $\mathcal{A}$. Thus the positive linear maps of $\mathcal{A}$ can be viewed as a groupoid – a category in which every arrow is invertible.

There is a corresponding groupoid whose objects are minimal dilation pairs $(V, \pi)$. Homomorphisms of dilation pairs $(V_1, \pi_1) \to (V_2, \pi_2)$ (here $\pi_j$ is a representation of $\mathcal{A}$ on $K_j$ and $V_j$ is an operator in $\mathcal{B}(H_j, K_j)$) are defined as unitary operators $W : K_1 \to K_2$ that satisfy

$$W\pi_1(a) = \pi_2(a)W, \qquad a \in \mathcal{A}, \quad \text{and} \quad WV_1 = V_2.$$

The "set" of all dilation pairs for a fixed positive linear map $\phi : \mathcal{A} \to \mathcal{B}(H)$ is a subgroupoid, and we have already seen in Remark 3.1 that its elements are all isomorphic. But here we are mainly concerned with how the dilation functor treats arrows between different positive linear maps.

A functor is the end product of a *construction.* In order to describe how the dilation functor acts on arrows, we need more information than the statement of Scholium A contains, namely the following: *There is a construction which starts with a positive linear map $\phi : \mathcal{A} \to \mathcal{B}(H)$ and generates a particular dilation pair $(V, \pi)_\phi$ from that data.* Scholium A asserts that such dilation pairs exist for every $\phi$, but since the proof is missing, we have not seen the construction. Later on, however, we will show how to construct a particular dilation pair $(V, \pi)_\phi$ from a completely positive map $\phi$ when we prove Stinespring's theorem in Section 6. That construction is analogous to the construction underlying (2.4), which exhibits an explicit function $f : X \to H(u)$ that arises from the construction of the Hilbert space $H(u)$, starting with a positive definite function $u$. In order to continue the current discussion, we ask the reader to assume the result of the construction of Theorem 6.1, namely that we are somehow given a *particular* dilation pair $(V, \pi)_\phi$ for every positive linear map $\phi : \mathcal{A} \to \mathcal{B}(H)$.

That puts us in position to describe how the dilation functor acts on arrows. Given two positive linear maps $\phi_j : \mathcal{A} \to \mathcal{B}(H_j)$, $j = 1, 2$, let $U : H_1 \to H_2$ be a unitary operator satisfying $U\phi_1(a) = \phi_2(a)U$ for $a \in \mathcal{A}$. Let $(V_1, \pi_1)$ and $(V_2, \pi_2)$ be the dilation pairs that have been constructed from $\phi_1$ and $\phi_2$ respectively. Notice that since $U^*\phi_2(a)U = \phi_1(a)$ for $a \in \mathcal{A}$, it follows that $(V_2U, \pi_2)$ is a second minimal dilation pair for $\phi_1$. By (3.4), there is a unique unitary operator $\tilde{U} : K_1 \to K_2$ that satisfies

$$\tilde{U}V_1 = V_2U, \quad \text{and} \quad \tilde{U}\pi_1(a) = \pi_2(a)\tilde{U}, \qquad a \in \mathcal{A}.$$

One can now check that the association $\phi, U \to (V, \pi)_\phi, \tilde{U}$ defines a covariant functor from the groupoid of operator-valued positive linear maps of $\mathcal{A}$ to the groupoid of minimal dilation pairs for $\mathcal{A}$.

## 4. Subnormality

An operator $A$ on a Hilbert space $H$ is said to be *subnormal* if it can be extended to a normal operator on a larger Hilbert space. More precisely, there should exist a normal operator $B$ acting on a Hilbert space $K \supseteq H$ that leaves $H$ invariant and restricts to $A$ on $H$. Halmos' paper [Hal50] introduced the concept, and grew out of his observation that a subnormal operator $A \in \mathcal{B}(H)$ must satisfy the following system of peculiar inequalities:

$$\sum_{i,j=0}^{n} \langle A^i \xi_j, A^j \xi_i \rangle \geq 0, \qquad \forall\, \xi_0, \xi_1, \dots, \xi_n \in H, \quad n = 0, 1, 2, \dots. \tag{4.1}$$

It is an instructive exercise with inequalities involving $2 \times 2$ operator matrices to show that the case $n = 1$ of (4.1) is equivalent to the single operator inequality $A^*A \geq AA^*$, a property called *hyponormality* today. Subnormal operators are certainly hyponormal, but the converse is false even for weighted shifts (see Problem 160 of [Hal67]). Halmos showed that the full set of inequalities (4.1) – together with a second system of necessary inequalities that we do not reproduce here – implies that $A$ is subnormal. Several years later, his student J. Bram proved that the second system of inequalities follows from the first [Bra55], and simpler proofs of that fact based on semigroup considerations emerged later [Szf77]. Hence the system of inequalities (4.1) is by itself necessary and sufficient for subnormality.

It is not hard to reformulate Halmos' notion of subnormality (for single operators) in a more general way that applies to several operators. Let $\Sigma$ be a commutative semigroup (written additively) that contains a neutral element 0. By a *representation* of $\Sigma$ we mean an operator-valued function $s \in \Sigma \mapsto A(s) \in \mathcal{B}(H)$ satisfying $A(s+t) = A(s)A(t)$ and $A(0) = \mathbf{1}$. Notice that we make no assumption on the norms $\|A(s)\|$ as $s$ varies over $\Sigma$. For example, a commuting set $A_1, \dots, A_d$ of operators on a Hilbert space $H$ gives rise to a representation of the $d$-dimensional additive semigroup

$$\Sigma = \{(n_1, \dots, n_d) \in \mathbb{Z}^d : n_1 \geq 0, \dots, n_d \geq 0\}$$

by way of

$$A(n_1, \dots, n_d) = A_1^{n_1} \cdots A_d^{n_d}, \qquad (n_1, \dots, n_d) \in \Sigma.$$

In general, a representation $A : \Sigma \to \mathcal{B}(H)$ is said to be *subnormal* if there is a Hilbert space $K \supseteq H$ and a representation $B : \Sigma \to \mathcal{B}(K)$ consisting of normal operators such that each $B(s)$ leaves $H$ invariant and

$$B(s) \restriction_H = A(s), \qquad s \in \Sigma.$$

We now apply Scholium A to prove a general statement about commutative operator semigroups that contains the Halmos-Bram characterization of subnormal operators, as well as higher-dimensional variations of it that apply to semigroups generated by a finite or even infinite number of mutually commuting operators.

**Theorem 4.1.** *Let $\Sigma$ be a commutative semigroup with 0. A representation $A : \Sigma \to \mathcal{B}(H)$ is subnormal iff for every $n \geq 1$, every $s_1, \dots, s_n \in \Sigma$ and every $\xi_1, \dots, \xi_n \in H$, one has*

$$\sum_{i,j=1}^{n} \langle A(s_i)\xi_j, A(s_j)\xi_i \rangle \geq 0. \tag{4.2}$$

*Proof.* The proof that the system of inequalities (4.2) is necessary for subnormality is straightforward, and we omit it. Here we outline a proof of the converse, describing all essential steps in the construction but leaving routine calculations for the reader. We shall make use of the hypothesis (4.2) in the following form: For every function $s \in \Sigma \mapsto \xi(s) \in H$ such that $\xi(s)$ vanishes for all but a finite number of $s \in \Sigma$, one has

$$\sum_{s,t\in\Sigma} \langle A(s)\xi(t), A(t)\xi(s)\rangle \geq 0. \tag{4.3}$$

We first construct an appropriate commutative Banach $*$-algebra. Note that the direct sum of semigroups $\Sigma \oplus \Sigma$ is a commutative semigroup with zero element $(0,0)$, but unlike $\Sigma$ it has a natural involution $x \mapsto x^*$ defined by $(s,t)^* = (t,s)$, $s,t \in \Sigma$. We will also make use of a weight function $w : \Sigma \oplus \Sigma \to [1,\infty)$ defined as follows:

$$w(s,t) = \max(\|A(s)\| \cdot \|A(t)\|, 1), \qquad s,t \in \Sigma.$$

Using $\|A(s+t)\| = \|A(s)A(t)\| \leq \|A(s)\| \cdot \|A(t)\|$, one finds that

$$1 \leq w(x+y) \leq w(x)w(y), \quad w(x^*) = w(x), \qquad x,y \in \Sigma \oplus \Sigma.$$

Note too that $w((0,0)) = 1$ because $A(0) = \mathbf{1}$. Consider the Banach space $\mathcal{A}$ of all functions $f : \Sigma \oplus \Sigma \to \mathbb{C}$ having finite weighted $\ell^1$-norm

$$\|f\| = \sum_{x\in\Sigma\oplus\Sigma} |f(x)| \cdot w(x) < \infty. \tag{4.4}$$

Since $w \geq 1$, the norm on $\mathcal{A}$ dominates the ordinary $\ell^1$ norm, so that every function in $\mathcal{A}$ belongs to $\ell^1(\Sigma \oplus \Sigma)$. Ordinary convolution of functions defined on commutative semigroups

$$(f * g)(z) = \sum_{\{x,y\in\Sigma\oplus\Sigma:\ x+y=z\}} f(x)g(y), \qquad z \in \Sigma \oplus \Sigma$$

defines an associative commutative multiplication in $\ell^1(\Sigma \oplus \Sigma)$, and it is easy to check that the above properties of the weight function $w$ imply that with respect to convolution and the involution $f^*(s,t) = \bar{f}(t,s)$, $\mathcal{A}$ becomes a commutative Banach $*$-algebra with normalized unit $\delta_{(0,0)}$.

We now use the semigroup $A(\cdot)$ to construct a linear map $\phi : \mathcal{A} \to \mathcal{B}(H)$:

$$\phi(f) = \sum_{(s,t)\in\Sigma\oplus\Sigma} f(s,t)A(s)^*A(t).$$

Note that $\|\phi(f)\| \leq \|f\|$ because of the definition of the norm of $f$ in terms of the weight function $w$. Obviously $\phi(\delta_{(s,t)}) = A(s)^*A(t)$ for all $s,t \in \Sigma$, and in particular $\phi(\delta_{(0,0)}) = \mathbf{1}$. It is also obvious that $\phi(f^*) = \phi(f)^*$ for $f \in \mathcal{A}$.

What is most important for us is that $\phi$ is a *positive* linear map, namely for every $f \in \mathcal{A}$ and every vector $\xi \in H$

$$\langle \phi((f^*) * f)\xi, \xi\rangle \geq 0. \tag{4.5}$$

To deduce this from (4.3), note that since $\phi : \mathcal{A} \to \mathcal{B}(H)$ is a bounded linear map and every function in $\mathcal{A}$ can be norm-approximated by functions which are finitely nonzero, it suffices to verify (4.5) for functions $f : \Sigma \oplus \Sigma \to \mathbb{C}$ such that $f(x) = 0$ for all but a finite number of $x \in \Sigma \oplus \Sigma$. But for two finitely supported functions $f, g \in \mathcal{A}$ and any function $H : \Sigma \oplus \Sigma \to \mathbb{C}$, the definition of convolution implies that $f * g$ is finitely supported, and

$$\sum_{z \in \Sigma \oplus \Sigma} (f * g)(z) H(z) = \sum_{x,y \in \Sigma \oplus \Sigma} f(x) g(y) H(x + y).$$

Fixing $\xi \in H$ and taking $H(s,t) = \langle A(s)^* A(t)\xi, \xi\rangle = \langle A(t)\xi, A(s)\xi\rangle$, we conclude from the preceding formula that

$$\begin{aligned}\langle \phi(f * g)\xi, \xi\rangle &= \sum_{s,t,u,v \in \Sigma} f(s,t) g(u,v) \langle A(t+v)\xi, A(s+u)\xi\rangle \\ &= \sum_{s,t,u,v \in \Sigma} f(s,t) g(u,v) \langle A(t)A(v)\xi, A(u)A(s)\xi\rangle.\end{aligned}$$

Thus we can write

$$\begin{aligned}\langle \phi(f^* * f)\xi, \xi\rangle &= \sum_{s,t,u,v \in \Sigma} \bar{f}(t,s) f(u,v) \langle A(t)A(v)\xi, A(u)A(s)\xi\rangle \\ &= \sum_{t,u \in \Sigma} \langle A(t)(\sum_{v \in \Sigma} f(u,v)A(v)\xi), A(u)(\sum_{s \in \Sigma} f(t,s)A(s)\xi)\rangle \\ &= \sum_{t,u \in \Sigma} \langle A(t)\xi(u), A(u)\xi(t)\rangle,\end{aligned}$$

where $t \in \Sigma \mapsto \xi(t) \in H$ is the vector function

$$\xi(t) = \sum_{s \in \Sigma} f(t,s) A(s)\xi, \qquad t \in \Sigma.$$

Notice that the rearrangements of summations carried out in the preceding formula are legitimate because all sums are finite, and in fact the vector function $t \mapsto \xi(t)$ is itself finitely nonzero. (4.5) now follows from (4.3).

At this point, we can apply Scholium A to find a Hilbert space $K$ which contains $H$ and a $*$*-representation* $\pi : \mathcal{A} \to \mathcal{B}(K)$ such that

$$P_H \pi(f) \restriction_H = \phi(f), \qquad f \in \mathcal{A}.$$

Hence the map $x \mapsto \pi(\delta_x)$ is a $*$-preserving representation of the $*$-semigroup $\Sigma \oplus \Sigma$, which can be further decomposed by way of $\pi(\delta_{(s,t)}) = B(s)^* B(t)$, where $B : \Sigma \to \mathcal{B}(K)$ is the representation $B(t) = \pi(\delta(0,t))$. Since the commutative semigroup of operators $\{\pi(\delta_x) : x \in \Sigma \oplus \Sigma\}$ is closed under the $*$-operation, $B(\Sigma)$ is a semigroup of mutually commuting normal operators. After taking $s = 0$ in the formulas $P_H B(s)^* B(t) \restriction_H = A(s)^* A(t)$, one finds that $A(t)$ is the compression of $B(t)$ to $H$. Moreover, since for every $t \in \Sigma$

$$P_H B(t)^* B(t) \restriction_H = A(t)^* A(t) = P_H B(t)^* P_H B(t) \restriction_H,$$

we have $P_H B(t)^*(\mathbf{1} - P_H)B(t)P_H = 0$. Thus we have shown that $H$ is invariant under $B(t)$ and the restriction of $B(t)$ to $H$ is $A(t)$. □

*Remark* 4.2 (Minimality and functoriality). Let $\Sigma$ be a commutative semigroup with zero. A normal extension $s \in \Sigma \mapsto B(s) \in \mathcal{B}(K)$ of a representation $s \in \Sigma \mapsto A(s) \in \mathcal{B}(H)$ on a Hilbert space $K \supseteq H$ is said to be *minimal* if the set of vectors $\{B(t)^*\xi : t \in \Sigma,\ \xi \in H\}$ has $K$ as its closed linear span. This corresponds to the notion of minimality described in Section 6. The considerations of Remark 3.1 imply that all minimal dilations are equivalent, and we can speak unambiguously of *the* minimal normal extension of $A$. A similar comment applies to the functorial nature of the map which carries subnormal representations of $\Sigma$ to their minimal normal extensions.

*Remark* 4.3 (Norms and flexibility). It is a fact that the minimal normal extension $B$ of $A$ satisfies $\|B(t)\| = \|A(t)\|$ for $t \in \Sigma$. The inequality $\geq$ is obvious since $A(t)$ is the restriction of $B(t)$ to an invariant subspace. However, if one attempts to use the obvious norm estimate for representations of Banach $*$-algebras (see the appendix for more detail) to establish the opposite inequality, one finds that the above construction gives only

$$\|B(t)\| = \|\pi(\delta_{(0,t)})\| \leq \|\delta_{(0,t)}\| = w(0,t) = \max(\|A(t)\|, 1),$$

which is not good enough when $\|A(t)\| < 1$. On the other hand, we can use the flexibility in the possible norms of $\mathcal{A}$ to obtain the correct estimate as follows. For each $\epsilon > 0$, define a new weight function $w_\epsilon$ on $\Sigma \oplus \Sigma$ by

$$w_\epsilon(s,t) = \max(\|A(s)\| \cdot \|A(t)\|, \epsilon), \qquad s,t \in \Sigma.$$

If one uses $w_\epsilon$ in place of $w$ in the definition (4.4) of the norm on $\mathcal{A}$, one obtains another commutative Banach $*$-algebra which serves equally well as the original to construct the minimal normal extension $B$ of $A$, and it has the additional feature that $\|B(t)\| \leq \max(\|A(t)\|, \epsilon)$ for $t \in \Sigma$. Since $\epsilon$ can be arbitrarily small, the desired estimate $\|B(t)\| \leq \|A(t)\|$ follows. In particular, for every $t \in \Sigma$ we have $A(t) = 0 \iff B(t) = 0$.

## 5. Commutative dilation theory

Dilation theory began with two papers of Naimark, written and published somehow during the darkest period of world war II: [Nai43a], [Nai43b]. Naimark's theorem asserts that a countably additive measure $E : \mathcal{F} \to \mathcal{B}(H)$ defined on a $\sigma$-algebra $\mathcal{F}$ of subsets of a set $X$ that takes values in the set of positive operators on a Hilbert space $H$ and satisfies $E(X) = \mathbf{1}$ can be expressed in the form

$$E(S) = P_H Q(S) \restriction_H, \qquad S \in \mathcal{F},$$

where $K$ is a Hilbert space containing $H$ and $Q : \mathcal{F} \to \mathcal{B}(K)$ is a spectral measure. A version of Naimark's theorem (for regular Borel measures on topological spaces) can be found on p. 50 of [Pau02]. Positive operator-valued measures $E$ have become fashionable in quantum physics and quantum information theory, where they go by the unpronounceable acronym POVM. It is interesting that the Wikipedia page for projection operator-valued measures (`http://en.wikipedia.org/wiki/POVM`) contains more information about Naimark's famous theorem than the Wikipedia page for Naimark himself (`http://en.wikipedia.org/wiki/Mark_Naimark`).

In his subnormality paper [Hal50], Halmos showed that every contraction $A \in \mathcal{B}(H)$ has a unitary dilation in the sense that there is a unitary operator $U$ acting on a larger Hilbert space $K \supseteq H$ such that

$$A = P_H U \restriction_H .$$

Sz.-Nagy extended that in a most significant way [SN53] by showing that every contraction has a unitary *power* dilation, and the latter result ultimately became the cornerstone for an effective model theory for Hilbert space contractions [SNF70]. Today, these results belong to the toolkit of every operator theorist, and can be found in many textbooks. In this section we merely state Sz.-Nagy's theorem and sketch a proof that is in the spirit of the preceding discussion.

**Theorem 5.1.** *Let $A \in \mathcal{B}(H)$ be an operator satisfying $\|A\| \leq 1$. Then there is a unitary operator $U$ acting on a Hilbert space $K$ containing $H$ such that*

$$A^n = P_H U^n \restriction_H, \qquad n = 0, 1, 2, \dots. \tag{5.1}$$

*If $U$ is* minimal *in the sense that $K$ is the closed linear span of $\cup_{n\in\mathbb{Z}} U^n H$, then it is uniquely determined up to a natural unitary equivalence.*

*Sketch of proof.* Consider the commutative Banach $*$-algebra $\mathcal{A} = \ell^1(\mathbb{Z})$, with multiplication and involution given by

$$(f * g)(n) = \sum_{k=-\infty}^{+\infty} f(k)g(n-k), \quad f^*(n) = \bar{f}(-n), \qquad n \in \mathbb{Z},$$

and normalized unit $\mathbf{1} = \delta_0$. Define $A(n) = A^n$ if $n \geq 0$ and $A(n) = A^{*|n|}$ if $n < 0$. Since $\|A(n)\| \leq 1$ for every $n$, we can define a linear map $\phi : \mathcal{A} \to \mathcal{B}(H)$ in the obvious way

$$\phi(f) = \sum_{n=-\infty}^{+\infty} f(n)A(n), \qquad f \in \mathcal{A}.$$

It is obvious that $\|\phi(f)\| \leq \|f\|$, $f \in \mathcal{A}$, but not at all obvious that $\phi$ is a positive linear map. However, there is a standard method for showing that for every $\xi \in H$, the sequence of complex numbers $a_n = \langle A(n)\xi, \xi\rangle$, $n \in \mathbb{Z}$, is of positive type in the sense that for every finitely nonzero sequence of complex numbers $\lambda_n$, $n \in \mathbb{Z}$, one has $\sum_{n\in\mathbb{Z}} a_{n-m}\lambda_n\bar{\lambda}_m \geq 0$; for example, see p. 36 of [Pau02]. By approximating $f \in \mathcal{A}$ in the norm of $\mathcal{A}$ with finitely nonzero functions and using

$$\langle \phi((f^*) * f)\xi, \xi\rangle = \sum_{m,n=-\infty}^{+\infty} \langle A(n-m)\xi, \xi\rangle f(n)\bar{f}(m),$$

it follows that $\langle \phi(f)\xi, \xi\rangle \geq 0$, and we may conclude that $\phi$ is a positive linear map of $\mathcal{A}$ to $\mathcal{B}(H)$ satisfying $\phi(\delta_0) = \mathbf{1}$.

Scholium A implies that there is a $*$-representation $\pi : \mathcal{A} \to \mathcal{B}(K)$ of $\mathcal{A}$ on a larger Hilbert space $K$ such that $\pi(f)$ compresses to $\phi(f)$ for $f \in \mathcal{A}$. Finally, since the enveloping $C^*$-algebra of $\mathcal{A} = \ell^1(\mathbb{Z})$ is the commutative $C^*$-algebra $C(\mathbb{T})$, the representation $\pi$ promotes to a representation $\tilde{\pi} : C(\mathbb{T}) \to \mathcal{B}(K)$ (see the appendix). Taking $z \in C(\mathbb{T})$ to be the coordinate variable, we obtain a unitary operator $U \in \mathcal{B}(K)$ by way of $U = \tilde{\pi}(z)$, and formula (5.1) follows. We omit the proof of the last sentence. □

No operator theorist can resist repeating the elegant proof of von Neumann's inequality that flows from Theorem 5.1. von Neumann's inequality [vN51] asserts that for every operator $A \in \mathcal{B}(H)$ satisfying $\|A\| \leq 1$, one has

$$\|f(A)\| \leq \sup_{|z|\leq 1} |f(z)| \tag{5.2}$$

for every polynomial $f(z) = a_0 + a_1 z + \cdots + a_n z^n$. von Neumann's original proof was difficult, involving calculations with Möbius transformations and Blaschke products. Letting $U \in \mathcal{B}(K)$ be a unitary power dilation of $A$ satisfying (5.1), one has $f(A) = P_H f(U) \restriction_H$, for every polynomial $f$, hence

$$\|f(A)\| \leq \|f(U)\| = \sup_{z\in\sigma(U)} |f(z)| \leq \sup_{|z|=1} |f(z)|.$$

## 6. Completely positivity and Stinespring's theorem

While one can argue that the GNS construction for states of $C^*$-algebras is a dilation theorem, it is probably best thought of as an application of the general method of associating a Hilbert space with a positive definite function as described in Section 2. Dilation theory *proper* went noncommutative in 1955 with the publication of a theorem of Stinespring [Sti55]. Stinespring told me that his original motivation was simply to find a common generalization of Naimark's commutative result that a positive operator-valued measure can be dilated to a spectral measure and the GNS construction for states of (noncommutative) $C^*$-algebras. The theorem that emerged went well beyond that, and today has become a pillar upon which significant parts of operator theory and operator algebras rest. The fundamental new idea underlying the result was that of a completely positive linear map.

The notion of positive linear functional or positive linear map is best thought of in a purely *algebraic* way. More specifically, let $\mathcal{A}$ be a $*$-algebra, namely a complex algebra endowed with an antilinear mapping $a \mapsto a^*$ satisfying $(ab)^* = b^*a^*$ and $a^{**} = a$ for all $a, b \in \mathcal{A}$. An operator-valued linear map $\phi : \mathcal{A} \to \mathcal{B}(H)$ (and in particular a complex-valued linear functional $\phi : \mathcal{A} \to \mathbb{C}$) is called *positive* if it satisfies

$$\phi(a^*a) \geq 0, \qquad a \in \mathcal{A}. \tag{6.1}$$

One can promote this notion of positivity to matrix algebras over $\mathcal{A}$. For every $n = 1, 2, \dots,$ the algebra $M_n(\mathcal{A})$ of $n \times n$ matrices over $\mathcal{A}$ has a natural involution, in which the adjoint of an $n \times n$ matrix is defined as the transposed matrix of adjoints $(a_{ij})^* = (a_{ji}^*)$, $1 \leq i, j \leq n$, $a_{ij} \in \mathcal{A}$. Fixing $n \geq 1$, a linear map $\phi : \mathcal{A} \to \mathcal{B}(H)$ induces a linear map $\phi_n$ from $M_n(\mathcal{A})$ to $n \times n$ operator matrices $(\phi(a_{ij}))$ which, after making the obvious identifications, can be viewed as a linear map of $M_n(\mathcal{A})$ to operators on the direct sum of $n$ copies of $H$. The original map $\phi$ is called *completely positive* if each $\phi_n$ is a positive linear map. More explicitly, complete positivity at level $n$ requires that (6.1) should hold for $n \times n$ matrices: For every $n \times n$ matrix $A = (a_{ij})$ with entries in $\mathcal{A}$ and every $n$-tuple of vectors $\xi_1, \dots, \xi_n \in H$, the $n \times n$ matrix $B = (b_{ij})$ defined by $B = A^*A$ should satisfy

$$\sum_{i,j=1}^{n} \langle \phi(b_{ij})\xi_j, \xi_i \rangle = \sum_{i,j,k=1}^{n} \langle \phi(a_{ki}^* a_{kj})\xi_j, \xi_i \rangle \geq 0.$$

Note that this system of inequalities reduces to a somewhat simpler-looking system of inequalities (3.1) that we have already seen in Section 3.

If $\mathcal{A}$ happens to be a $C^*$-algebra, then the elements $x \in \mathcal{A}$ that can be represented in the form $x = a^*a$ for some $a \in \mathcal{A}$ are precisely the self-adjoint operators $x$ having nonnegative spectra. Since $M_n(\mathcal{A})$ is also a $C^*$-algebra in a unique way for every $n \geq 1$, completely positive linear maps of $C^*$-algebras have a very useful spectral characterization: they should map self-adjoint $n \times n$ operator matrices with nonnegative spectra to self-adjoint operators with nonnegative spectra. Unfortunately, this spectral characterization breaks down completely for positive linear maps of more general Banach $*$-algebras, and in that more general context one must always refer back to positivity as it is expressed in (6.1).

Stinespring's original result was formulated in terms of operator maps defined on $C^*$-algebras. We want to reformulate it somewhat into the more flexible context of linear maps of Banach $*$-algebras.

**Theorem 6.1.** *Let $\mathcal{A}$ be a Banach $*$-algebra with normalized unit and let $H$ be a Hilbert space. For every completely positive linear map $\phi : \mathcal{A} \to \mathcal{B}(H)$ there is a representation $\pi : \mathcal{A} \to \mathcal{B}(K)$ of $\mathcal{A}$ on another Hilbert space $K$ and a bounded linear map $V : H \to K$ such that*

$$\phi(a) = V^*\pi(a)V, \qquad a \in \mathcal{A}. \tag{6.2}$$

*Moreover, the norm of the linking operator $V$ is given by $\|V\|^2 = \|\phi(\mathbf{1})\|$.*

We have omitted the statement and straightforward proof of the converse, namely that every linear map $\phi : \mathcal{A} \to \mathcal{B}(H)$ of the form (6.2) must be completely positive, in order to properly emphasize the *construction* of the dilation from the basic properties of a completely positive map.

*Proof.* The underlying construction is identical with the original [Sti55], but a particular estimate requires care in the context of Banach $*$-algebras, and we will make that explicit. Consider the tensor product of complex vector spaces $\mathcal{A} \otimes H$, and the sesquilinear form $\langle \cdot, \cdot \rangle$ defined on it by setting

$$\left\langle \sum_{j=1}^{m} a_j \otimes \xi_j, \sum_{k=1}^{n} b_k \otimes \eta_k \right\rangle = \sum_{j,k=1}^{m,n} \langle \phi(b_k^* a_j)\xi_j, \eta_k \rangle.$$

The fact that $\phi$ is completely positive implies that $\langle \zeta, \zeta \rangle \geq 0$ for every $\zeta \in \mathcal{A} \otimes H$. Letting $\mathcal{N} = \{\zeta \in \mathcal{A} \otimes H : \langle \zeta, \zeta \rangle = 0\}$, the Schwarz inequality implies that $\mathcal{N}$ is a linear subspace and that the sesquilinear form can be promoted to an inner product on the quotient $K_0 = (\mathcal{A} \otimes H)/\mathcal{N}$. Let $K$ be the completion of the resulting inner product space.

Each $a \in \mathcal{A}$ gives rise to a left multiplication operator $\pi(a)$ acting on $\mathcal{A} \otimes H$, defined uniquely by $\pi(a)(b \otimes \xi) = ab \otimes \xi$ for $b \in \mathcal{A}$ and $\xi \in H$. The critical estimate that we require is

$$\langle \pi(a)\zeta, \pi(a)\zeta \rangle \leq \|a\|^2 \langle \zeta, \zeta \rangle, \qquad a \in \mathcal{A}, \quad \zeta \in \mathcal{A} \otimes H, \tag{6.3}$$

and it is proved as follows. Writing $\zeta = a_1 \otimes \xi_1 + \cdots + a_n \otimes \xi_n$, we find that

$$\begin{aligned} \langle \pi(a)\zeta, \pi(a)\zeta \rangle &= \sum_{j,k=1}^{n} \langle ab_j \otimes \xi_j, ab_k \otimes \xi_k \rangle = \sum_{j,k=1}^{n} \langle \phi(b_k^* a^* a b_j)\xi_j, \xi_k \rangle \\ &= \sum_{j,k=1}^{n} \langle a^* a b_j \otimes \xi_j, b_k \otimes \xi_k \rangle = \langle \pi(a^* a)\zeta, \zeta \rangle. \end{aligned}$$

This formula implies that the linear functional $\rho(a) = \langle \pi(a)\zeta, \zeta \rangle$ satisfies $\rho(a^* a) = \langle \pi(a)\zeta, \pi(a)\zeta \rangle \geq 0$. Proposition A.1 of the appendix implies

$$\rho(a^* a) \leq \rho(\mathbf{1})\|a^* a\| \leq \|\zeta\|^2 \|a\|^2,$$

and (6.3) follows.

It is obvious that $\pi(ab) = \pi(a)\pi(b)$ and that $\pi(\mathbf{1})$ is the identity operator. Moreover, as in the argument above, we have $\langle \pi(a)\eta, \zeta \rangle = \langle \eta, \pi(a^*)\zeta \rangle$ for all $a \in \mathcal{A}$ and $\eta, \zeta \in \mathcal{A} \otimes H$. Finally, (6.3) implies that $\pi(a)\mathcal{N} \subseteq \mathcal{N}$, so that each operator $\pi(a)$, $a \in \mathcal{A}$, promotes naturally to a linear operator on the quotient $K_0 = (\mathcal{A} \otimes H)/\mathcal{N}$. Together with (6.3), these formulas imply that $\pi$ gives rise to a $*$ representation of $\mathcal{A}$ as bounded operators on $K_0$ which extends uniquely to a representation of $\mathcal{A}$ on the completion $K$ of $K_0$, which we denote by the same letter $\pi$.

It remains only to discuss the connecting operator $V$, which is defined by $V\xi = \mathbf{1} \otimes \xi + \mathcal{N}$, $\xi \in H$. One finds that $\pi(a)V\xi = a \otimes \xi + \mathcal{N}$, from which it follows that

$$\langle \pi(a)V\xi, V\eta \rangle = \langle a \otimes \xi + \mathcal{N}, \mathbf{1} \otimes \eta + \mathcal{N} \rangle = \langle \phi(a)\xi, \eta \rangle, \qquad \xi, \eta \in H.$$

Taking $a = \mathbf{1}$, we infer that $\|V\|^2 = \|V^* V\| = \|\phi(\mathbf{1})\|$, and at that point the preceding formula implies $\phi(a) = V^* \pi(a) V$, $a \in \mathcal{A}$. □

## 7. Operator spaces, operator systems and extensions

In this section we discuss the basic features of operator spaces and their matrix hierarchies, giving only the briefest of overviews. The interested reader is referred to one of the monographs [BLM04], [ER00], [Pau86] for more about this developing area of noncommutative analysis.

Complex Banach spaces are the objects of a category whose maps are contractions – linear operators of norm $\leq 1$. The isomorphisms of this category are surjective isometries. A *function space* is a norm-closed linear subspace of some $C(X)$ – the space of (complex-valued) continuous functions on a compact Hausdorff space $X$, endowed with the sup norm. All students of analysis know that every Banach space $E$ is isometrically isomorphic to a function space. Indeed, the Hahn-Banach theorem implies that the natural map $\iota : E \to E''$ of $E$ into its double dual has the stated property after one views elements if $\iota(E)$ as continuous functions on the weak*-compact unit ball $X$ of $E'$. In this way the study of Banach spaces can be reduced to the study of function spaces, and that fact is occasionally useful.

An *operator space* is a norm-closed linear subspace $\mathcal{E}$ of the algebra $\mathcal{B}(H)$ of all bounded operators on a Hilbert space $H$. Such an $\mathcal{E}$ is itself a Banach space, and is therefore isometrically isomorphic to a function space. However, the key fact about operator spaces is that they determine an entire hierarchy of operator spaces, one for every $n = 1, 2, \dots$. Indeed, for every $n$, the space $M_n(\mathcal{E})$ of all $n \times n$ matrices over $\mathcal{E}$ is naturally an operator subspace of $\mathcal{B}(n \cdot H)$, $n \cdot H$ denoting the direct sum of $n$ copies of $H$. Most significantly, a linear map of operator spaces $\phi : \mathcal{E}_1 \to \mathcal{E}_2$ determines a sequence of linear maps $\phi_n : M_n(\mathcal{E}_1) \to M_n(\mathcal{E}_2)$, where $\phi_n$ is the linear map obtained by applying $\phi$ element-by-element to an $n \times n$ matrix over $\mathcal{E}_1$. One says that $\phi$ is a *complete isometry* or a *complete contraction* if every $\phi_n$ is, respectively, an isometry or a contraction. There is a corresponding notion of *complete boundedness* that will not concern us here.

Operator spaces can be viewed as the objects of a category whose maps are *complete* contractions. The isomorphisms of this category are complete isometries, and one is led to seek properties of operator spaces that are invariant under this refined notion of isomorphism. Like Shiva, a given Banach space acquires many inequivalent likenesses when it is realized concretely as an operator space. That is because in operator space theory one pays attention to what happens at every level of the hierarchy. The result is a significant and fundamentally noncommutative refinement of classical Banach space theory.

For example, since an operator space $\mathcal{E} \subseteq \mathcal{B}(H)$ is an "ordinary" Banach space, it can be represented as a function space $\iota : \mathcal{E} \to C(X)$ as in the opening paragraphs of this section. If we form the hierarchy of $C^*$-algebras $M_n(C(X))$, $n = 1, 2, \dots$, then we obtain a sequence of embeddings

$$\iota_n : M_n(\mathcal{E}) \to M_n(C(X)), \qquad n = 1, 2, \dots.$$

Note that the $C^*$-algebra $M_n(C(X))$ is basically the $C^*$-algebra of all matrix-valued continuous functions $F : X \to M_n(\mathbb{C})$, with norm

$$\|F\| = \sup_{x \in X} \|F(x)\|, \quad F \in M_n(C(X)).$$

While the map $\iota$ is surely an isometry at the first level $n = 1$, it may or may not be a complete isometry; indeed for the more interesting examples of operator spaces it is not. Ultimately, the difference between Banach spaces and operator spaces can be traced to the noncommutativity of operator multiplication, and for that reason some analysts like to think of operator space theory as the "quantized" reformulation of functional analysis.

Finally, one can think of operator spaces somewhat more flexibly as norm-closed linear subspaces $\mathcal{E}$ of unital (or even nonunital) $C^*$-algebras $\mathcal{A}$. That is because the hierarchy of $C^*$-algebras $M_n(\mathcal{A})$ is well defined independently of any particular faithful realization $\mathcal{A}$ as a $C^*$-subalgebra of $\mathcal{B}(H)$.

One can import the notion of *order* into the theory of operator spaces in a natural way. A *function system* is a function space $E \subseteq C(X)$ with the property that $E$ is closed under complex conjugation and contains the constants. One sometimes assumes that $E$ separates points of $X$ but here we do not. Correspondingly, an *operator system* is a self-adjoint operator space $\mathcal{E} \subseteq \mathcal{B}(H)$ that contains the identity operator $\mathbf{1}$. The natural notion of order between self-adjoint operators, namely $A \leq B \iff B - A$ is a positive operator, has meaning in any operator system $\mathcal{E}$, and in fact every operator *system* is linearly spanned by its positive operators. Every member $M_n(\mathcal{E})$ of the matrix hierarchy over an operator system $\mathcal{E}$ is an operator system, so that it makes sense to speak of completely positive maps from one operator system to another.

Krein's version of the Hahn-Banach theorem implies that a positive linear functional defined on an operator system $\mathcal{E}$ in a $C^*$-algebra $\mathcal{A}$ can be extended to a positive linear functional on all of $\mathcal{A}$. It is significant that this extension theorem fails in general for operator-valued positive linear maps. Fortunately, the following result of [Arv69] provides an effective noncommutative counterpart of Krein's order-theoretic Hahn-Banach theorem:

**Theorem 7.1.** *Let $\mathcal{E} \subseteq \mathcal{A}$ be an operator system in a unital $C^*$-algebra. Then every operator-valued completely positive linear map $\phi : \mathcal{E} \to \mathcal{B}(H)$ can be extended to a completely positive linear map of $\mathcal{A}$ into $\mathcal{B}(H)$.*

There is a variant of 7.1 that looks more like the original Hahn-Banach theorem. Let $\mathcal{E} \subseteq \mathcal{A}$ be an operator space in a $C^*$-algebra $\mathcal{A}$. Then every operator-valued *complete* contraction $\phi : \mathcal{E} \to \mathcal{B}(H)$ can be extended to a completely contractive linear map of $\mathcal{A}$ to $\mathcal{B}(H)$. While the latter extension theorem emerged more than a decade after Theorem 7.1 (with a different and longer proof [Wit81], [Wit84]), Vern Paulsen discovered a simple device that enables one to deduce it readily from the earlier result. That construction begins with an operator space $\mathcal{E} \subseteq \mathcal{A}$ and generates an associated operator *system* $\tilde{\mathcal{E}}$ in the $2 \times 2$ matrix algebra

$M_2(\mathcal{A})$ over $\mathcal{A}$ as follows:

$$\tilde{\mathcal{E}} = \left\{ \begin{pmatrix} \lambda \cdot \mathbf{1} & A \\ B^* & \lambda \cdot \mathbf{1} \end{pmatrix} : A, B \in \mathcal{E},\ \lambda \in \mathbb{C} \right\}.$$

Given a completely contractive linear map $\phi : \mathcal{E} \to \mathcal{B}(H)$, one can define a linear map $\tilde{\phi} : \tilde{\mathcal{E}} \to \mathcal{B}(H \oplus H)$ in a natural way

$$\tilde{\phi}\left( \begin{pmatrix} \lambda \cdot \mathbf{1} & A \\ B^* & \lambda \cdot \mathbf{1} \end{pmatrix} \right) = \begin{pmatrix} \lambda \cdot \mathbf{1} & \phi(A) \\ \phi(B)^* & \lambda \cdot \mathbf{1} \end{pmatrix},$$

and it is not hard to see that $\tilde{\phi}$ is completely positive (I have reformulated the construction in a minor but equivalent way for simplicity; see Lemma 8.1 of [Pau02] for the original). By Theorem 7.1, $\tilde{\phi}$ extends to a completely positive linear map of $M_2(\mathcal{A})$ to $\mathcal{B}(H \oplus H)$, and the behavior of that extension on the upper right corner is a completely contractive extension of $\phi$.

## 8. Spectral sets and higher-dimensional operator theory

Some aspects of commutative operator theory can be properly understood only when placed in the noncommutative context of the matrix hierarchies of the preceding section. In this section we describe the phenomenon in concrete terms, referring the reader to the literature for technical details.

Let $A \in \mathcal{B}(H)$ be a Hilbert space operator. If $f$ is a rational function of a single complex variable that has no poles on the spectrum of $A$, then there is an obvious way to define an operator $f(A) \in \mathcal{B}(H)$. Now fix a compact subset $X \subseteq \mathbb{C}$ of the plane that contains the spectrum of $A$. The algebra $R(X)$ of all rational functions whose poles lie in the complement of $X$ forms a unital subalgebra of $C(X)$, and this functional calculus defines a unit-preserving homomorphism $f \mapsto f(A)$ of $R(X)$ into $\mathcal{B}(H)$. One says that $X$ is a *spectral set* for $A$ if this homomorphism has norm 1:

$$\|f(A)\| \le \sup_{z \in X} |f(z)|, \qquad f \in R(X). \tag{8.1}$$

Von Neumann's inequality (5.2) asserts that the closed unit disk is a spectral set for every contraction $A \in \mathcal{B}(H)$; indeed, that property is characteristic of contractions. While there is no corresponding characterization of the operators that have a more general set $X$ as a spectral set, we are still free to consider the class of operators that *do* have $X$ as a spectral set and ask if there is a generalization of Theorem 5.1 that would apply to them. Specifically, given an operator $A \in \mathcal{B}(H)$ that has $X$ as a spectral set, is there a normal operator $N$ acting on a larger Hilbert space $K \supseteq H$ such that the spectrum of $N$ is contained in the boundary $\partial X$ of $X$ and

$$f(A) = P_H f(N) \restriction_H, \qquad f \in R(X)? \tag{8.2}$$

A result of Foias implies that the answer is yes if the complement of $X$ is connected, but it is no in general. The reason the answer is no in general is that the hypothesis (8.1) is not strong enough; and that phenomenon originates in the *noncommutative* world. To see how the hypothesis must be strengthened, let $N$ be a normal operator with spectrum in $\partial X$. The functional calculus for normal operators gives rise to a representation $\pi : C(\partial X) \to \mathcal{B}(K)$, $\pi(f) = f(N)$, $f \in C(\partial X)$. It is easy to see that representations of $C^*$-algebras are completely positive and completely contractive linear maps, hence if the formula (8.2) holds then the map $f \in R(X) \mapsto f(A)$ must be not only be a contraction, it must be a *complete* contraction.

Let us examine the latter assertion in more detail. Fix $n = 1, 2, \dots$ and let $M_n(R(X))$ be the algebra of all $n \times n$ matrices with entries in $R(X)$. One can view an element of $M_n(R(X))$ as a matrix-valued rational function

$$F : z \in X \mapsto F(z) = (f_{ij}(z)) \in M_n(\mathbb{C}),$$

whose component functions belong to $R(X)$. Notice that we can apply such a matrix-valued function to an operator $A$ that has spectrum in $X$ to obtain an $n \times n$ matrix of operators – or equivalently an operator $F(A) = (f_{ij}(A))$ in $\mathcal{B}(n \cdot H)$. The map $F \in M_n(R(X)) \mapsto F(A) \in \mathcal{B}(n \cdot H)$ is a unit-preserving homomorphism of complex algebras. $X$ is said to be a *complete* spectral set for an operator $A \in \mathcal{B}(H)$ if it contains the spectrum of $A$ and satisfies

$$\|F(A)\| \leq \sup_{z \in X} \|F(z)\|, \qquad F \in M_n(R(X)), \quad n = 1, 2, \dots. \tag{8.3}$$

Now if there is a normal operator $N$ with spectrum in $\partial X$ that relates to $A$ as in (8.2), then for every $n = 1, 2, \dots$ and every $F \in M_n(R(X))$,

$$\|F(A)\| \leq \|F(N)\| \leq \sup_{z \in \partial X} \|F(z)\| = \sup_{z \in X} \|F(z)\|,$$

and we conclude that $X$ must be a *complete* spectral set for $A$.

The following result from [Arv72] implies that complete spectral sets suffice for the existence of normal dilations. It depends in an essential way on the extension theorem (Theorem 7.1) for completely positive maps.

**Theorem 8.1.** *Let $A \in \mathcal{B}(H)$ be an operator that has a compact set $X \subseteq \mathbb{C}$ as a complete spectral set. Then there is a normal operator $N$ on a Hilbert space $K \supseteq H$ having spectrum in $\partial X$ such that*

$$f(A) = P_H f(N) \restriction_H, \qquad f \in R(X).$$

The unitary power dilation of a contraction is unique up to natural equivalence. That reflects a property of the unit circle $\mathbb{T}$: Every positive linear map $\phi : C(\mathbb{T}) \to \mathcal{B}(H)$ is uniquely determined by its values on the nonnegative powers $1, z, z^2, \dots$ of the coordinate variable $z$. In general, however, positive linear maps of $C(X)$ are not uniquely determined by their values on subalgebras of $C(X)$, with the result that there is no uniqueness assertion to complement the existence assertion of Theorem 8.1 for the dilation theory of complete spectral sets.

On the other hand, there is a "many operators" generalization of Theorem 8.1 that applies to completely contractive unit-preserving homomorphisms of arbitrary function algebras $A \subseteq C(X)$ that act on compact Hausdorff spaces $X$, in which $\partial X$ is replaced by the Silov boundary of $X$ relative to $A$. The details can be found in Theorem 1.2.2 of [Arv72] and its corollary.

## 9. Completely positive maps and endomorphisms

In recent years, certain problems arising in mathematical physics and quantum information theory have led researchers to seek a different kind of dilation theory, one that applies to semigroups of completely positive linear maps that act on von Neumann algebras. In this section, we describe the simplest of these dilation theorems as it applies to the simplest semigroups acting on the simplest of von Neumann algebras. A fuller accounting of these developments together with references to other sources can be found in Chapter 8 of the monograph [Arv03].

Let $\phi : \mathcal{B}(H) \to \mathcal{B}(H)$ be a unit-preserving completely positive (UCP) map which is *normal* in the sense that for every normal state $\rho$ of $\mathcal{B}(H)$, the composition $\rho \circ \phi$ is also a normal state. One can think of the semigroup

$$\{\phi^n : n = 0, 1, 2, \dots\}$$

as representing the discrete time evolution of an irreversible quantum system. What we seek is another Hilbert space $K$ together with a normal $*$-endomorphism $\alpha : \mathcal{B}(K) \to \mathcal{B}(K)$ that is in some sense a "power dilation" of $\phi$. There are a number of ways one can make that vague idea precise, but only one of them is completely effective. It is described as follows.

Let $K \supseteq H$ be a Hilbert space that contains $H$ and suppose we are given a normal $*$-endomorphism $\alpha : \mathcal{B}(K) \to \mathcal{B}(K)$ that satisfies $\alpha(\mathbf{1}) = \mathbf{1}$. We write the projection $P_H$ of $K$ on $H$ simply as $P$, and we identify $\mathcal{B}(H)$ with the corner $P\mathcal{B}(K)P \subseteq \mathcal{B}(K)$. $\alpha$ is said to be a *dilation* of $\phi$ if

$$\phi^n(A) = P\alpha^n(A)P, \qquad A \in \mathcal{B}(H) = P\mathcal{B}(K)P, \quad n = 0, 1, 2, \dots. \tag{9.1}$$

Since $\phi$ is a unit-preserving map of $\mathcal{B}(H)$, $P = \phi(P) = P\alpha(P)P$, so that $\alpha(P) \geq P$. Hence we obtain an increasing sequence of projections

$$P \leq \alpha(P) \leq \alpha^2(P) \leq \cdots. \tag{9.2}$$

The limit projection $P_\infty = \lim_n \alpha^n(P)$ satisfies $\alpha(P_\infty) = P_\infty$, hence the compression of $\alpha$ to the larger corner $P_\infty \mathcal{B}(K) P_\infty \cong \mathcal{B}(P_\infty K)$ of $\mathcal{B}(K)$ is a unital $*$-endomorphism that is itself a dilation of $\phi$. By cutting down if necessary we can assume that the configuration is *proper* in the sense that

$$\lim_{n\to\infty} \alpha^n(P) = \mathbf{1}_K, \tag{9.3}$$

and in that case the endomorphism $\alpha$ is said to be a *proper* dilation of $\phi$. We have refrained from using the term *minimal* to describe this situation because in the context of semigroups of completely positive maps, the notion of *minimal* dilation is a more subtle one that requires a stronger hypothesis. That hypothesis is discussed in Remark 9.3 below.

*Remark* 9.1 (Stinespring's theorem is not enough). It is by no means obvious that dilations should exist. One might attempt to construct a dilation of the semigroup generated by a single UCP map $\phi : \mathcal{B}(H) \to \mathcal{B}(H)$ by applying Stinespring's theorem to the individual terms of the sequence of powers $\phi^n$, $n = 0, 1, 2, \dots$, and then somehow putting the pieces together to obtain the dilating endomorphism $\alpha$. Indeed, Stinespring's theorem provides us with a Hilbert space $K_n \supseteq H$ and a representation $\pi_n : \mathcal{B}(H) \to \mathcal{B}(K_n)$ for every $n \geq 0$ such that

$$\phi^n(A) = P_H \pi_n(A) \restriction_H, \qquad A \in \mathcal{B}(H), \quad n = 0, 1, 2, \dots.$$

However, while these formulas certainly inherit a relation to each other by virtue of the semigroup formula $\phi^{m+n} = \phi^m \circ \phi^n$, $m, n \geq 0$, if one attempts to exploit these relationships one finds that the relation between $\pi_m$, $\pi_n$ and $\pi_{m+n}$ is extremely awkward. Actually, there is no apparent way to assemble the von Neumann algebras $\pi_n(\mathcal{B}(H))$ into a single von Neumann algebra that plays the role of $\mathcal{B}(K)$, on which one can define a single endomorphism $\alpha$ that converts these formulas into the single formula (9.1). Briefly put, *Stinespring's theorem does not apply to semigroups.*

These observations suggest that the problem of constructing dilations in this context calls for an entirely new method, and it does. The proper result for normal UCP maps acting on $\mathcal{B}(H)$ was discovered by Bhat and Parthasarathy [BP94], building on earlier work of Parthasarathy [Par91] that was set in the context of quantum probability theory. The result was later extended by Bhat to semigroups of completely positive maps that act on arbitrary von Neumann algebras [Bha99]. The construction of the dilation has been reformulated in various ways; the one I like is in Chapter 8 of [Arv03] (also see [Arv02]). Another approach, due to Muhly and Solel [MS02], is based on correspondences over von Neumann algebras. The history of earlier approaches to this kind of dilation theory is summarized in the notes of Chapter 8 of [Arv03].

We now state the appropriate result for $\mathcal{B}(H)$ without proof:

**Theorem 9.2.** *For every normal UCP map $\phi : \mathcal{B}(H) \to \mathcal{B}(H)$, there is a Hilbert space $K \supseteq H$ and a normal $*$-endomorphism $\alpha : \mathcal{B}(H) \to \mathcal{B}(H)$ satisfying $\alpha(\mathbf{1}) = \mathbf{1}$ that is a proper dilation of $\phi$ as in* (9.1).

*Remark* 9.3 (Minimality and uniqueness). The notion of minimality for a dilation $\alpha : \mathcal{B}(K) \to \mathcal{B}(K)$ of a UCP map $\phi : \mathcal{B}(H) \to \mathcal{B}(H)$ is described as follows. Again, we identify $\mathcal{B}(H)$ with the corner $P\mathcal{B}(K)P$. We have already pointed out that the projections $\alpha^n(P)$ increase with $n$. However, the sequence of (nonunital) von Neumann subalgebras $\alpha^n(\mathcal{B}(H))$, $n = 0, 1, 2, \dots$, neither increases nor decreases with $n$, and that behavior requires care. The proper notion of minimality in this context is that the set of all vectors in $K$ of the form

$$\alpha^{n_1}(A_1)\alpha^{n_2}(A_2)\cdots\alpha^{n_k}(A_k)\xi, \tag{9.4}$$

where $k = 1, 2, \dots$, $n_k = 0, 1, 2, \dots$, $A_k \in \mathcal{B}(H)$, and $\xi \in H$, should have $K$ as their closed linear span. Equivalently, the smallest subspace of $K$ that contains $H$

and is invariant under the set of operators

$$\mathcal{B}(H) \cup \alpha(\mathcal{B}(H)) \cup \alpha^2(\mathcal{B}(H)) \cup \cdots$$

should be all of $K$. It is a fact that every minimal dilation is proper, but the converse is false. It is also true that every proper dilation can be reduced in a natural way to a minimal one, and finally, that any two minimal dilations of the semigroup $\{\phi^n : n \geq 0\}$ are isomorphic in a natural sense.

We also point out that there is a corresponding dilation theory for one-parameter semigroups of UCP maps. These facts are discussed at length in Chapter 8 of [Arv03].

## Appendix: Brief on Banach $*$-algebras

Banach $*$-algebras (defined at the beginning of Section 3) are useful because they are flexible – it is usually a simple matter to define a Banach $*$-algebra with the properties one needs. More importantly, it is far easier to define states and representations of Banach $*$-algebras than it is for the more rigid category of $C^*$-algebras. For example, we made use of the technique in the proof of Theorem 4.1 and the estimate of Remark 4.3.

On the other hand, it is obviously desirable to have $C^*$-algebraic tools available for carrying out analysis. Fortunately one can have it both ways, because every Banach $*$-algebra $\mathcal{A}$ is associated with a unique enveloping $C^*$-algebra $C^*(\mathcal{A})$ which has the "same" representation theory and the "same" state space as $\mathcal{A}$. In this Appendix we briefly describe the properties of this useful functor $\mathcal{A} \to C^*(\mathcal{A})$ for the category of Banach $*$-algebras that have a normalized unit $\mathbf{1}$. There are similar results (including Proposition A.1 below) for many nonunital Banach $*$-algebras – including the group algebras of locally compact groups – provided that they have appropriate approximate units. A comprehensive treatment can be found in [Dix64].

The fundamental fact on which these results are based is the following (see Proposition 4.7.1 of the text [Arv01] for a proof):

**Proposition** A.1. *Every positive linear functional $\rho$ on a unital Banach $*$-algebra $\mathcal{A}$ is bounded, and in fact $\|\rho\| = \rho(\mathbf{1})$.*

What we actually use here is the following consequence, which is proved by applying Proposition A.1 to functionals of the form $\rho(a) = \langle \phi(a)\xi, \xi \rangle$:

**Corollary** A.2. *Every operator-valued positive linear map $\phi : \mathcal{A} \to \mathcal{B}(H)$ is bounded, and $\|\phi\| = \|\phi(\mathbf{1})\|$.*

By a *representation* of a Banach $*$-algebra $\mathcal{A}$ we mean a $*$-preserving homomorphism $\pi : \mathcal{A} \to \mathcal{B}(H)$ of $\mathcal{A}$ into the $*$-algebra of operators on a Hilbert space. It is useful to assume the representation is nondegenerate in the sense that $\pi(\mathbf{1}) = \mathbf{1}$; if that is not the case, it can be arranged by passing to the subrepresentation defined on the subspace $H_0 = \pi(\mathbf{1})H$. Representations of Banach $*$-algebras arise from positive linear functionals (by way of the GNS construction which makes use of Proposition A.1) or from completely positive linear maps (by a variation of Theorem 6.1, by making use of Corollary A.2).

While we have made no hypothesis on the norms $\|\pi(a)\|$ associated with a representation $\pi$, it follows immediately from Proposition A.1 that every representation of $\mathcal{A}$ has norm 1. Indeed, for every unit vector $\xi \in H$ and $a \in \mathcal{A}$, $\rho(a) = \langle \pi(a)\xi, \xi \rangle$ defines a positive linear functional on $\mathcal{A}$ with $\rho(\mathbf{1}) = 1$, so that

$$\|\pi(a)\xi\|^2 = \langle \pi(a)^*\pi(a)\xi, \xi \rangle = \langle \pi(a^*a)\xi, \xi \rangle = \rho(a^*a) \leq \|a^*a\| \leq \|a\|^2,$$

and $\|\pi(a)\| \leq \|a\|$ follows. It is an instructive exercise to find a direct proof of the inequality $\|\pi(a)\| \leq \|a\|$ that does not make use of Proposition A.1.

*Remark* A.3 (Enveloping $C^*$-algebra of a Banach $*$-algebra). Consider the seminorm $\|\cdot\|_1$ defined on $\mathcal{A}$ by

$$\|a\|_1 = \sup_\pi \|\pi(a)\|, \qquad a \in \mathcal{A},$$

the supremum taken over a "all" representations of $\mathcal{A}$. Since the representations of $\mathcal{A}$ do not form a set, the quotes simply refer to an obvious way of choosing sufficiently many representatives from unitary equivalence classes of representations so that every representation is unitarily equivalent to a direct sum of the representative ones. It is clear that $\|a^*a\|_1 = \|a\|_1^2$. Indeed, $\|\cdot\|_1$ is a $C^*$-seminorm, and the completion of $\mathcal{A}/\{x \in \mathcal{A} : \|x\|_1 = 0\}$ is a $C^*$-algebra $C^*(\mathcal{A})$, called the enveloping $C^*$-algebra of $\mathcal{A}$. The natural completion map

$$\iota : \mathcal{A} \to C^*(\mathcal{A}) \tag{A.1}$$

is a $*$-homomorphism having dense range and norm 1. This completion (A.1) has the following universal property: For every representation $\pi : \mathcal{A} \to \mathcal{B}(H)$ there is a unique representation $\tilde{\pi} : C^*(\mathcal{A}) \to \mathcal{B}(H)$ such that $\tilde{\pi} \circ \iota = \pi$. The map $\pi \to \tilde{\pi}$ is in fact a bijection. Indeed, Proposition A.1 is equivalent to the assertion that there is a bijection between the set of positive linear functionals $\rho$ on $\mathcal{A}$ and the set of positive linear functionals $\tilde{\rho}$ on its enveloping $C^*$-algebra, defined by a similar formula $\tilde{\rho} \circ \iota = \rho$.

One should keep in mind that the completion map (A.1) can have a nontrivial kernel in general, but for many important examples it is injective. For example, it is injective in the case of group algebras – the Banach $*$-algebras $L^1(G)$ associated with locally compact groups $G$. When $G$ is commutative, the enveloping $C^*$-algebra of $L^1(G)$ is the $C^*$-algebra $C_\infty(\hat{G})$ of continuous functions that vanish at $\infty$ on the character group $\hat{G}$ of $G$.

## References

[Arv69] W. Arveson. Subalgebras of $C^*$-algebras. *Acta Math.*, 123:141–224, 1969.

[Arv72] W. Arveson. Subalgebras of $C^*$-algebras II. *Acta Math.*, 128:271–308, 1972.

[Arv01] W. Arveson. *A Short Course on Spectral Theory*, volume 209 of *Graduate Texts in Mathematics*. Springer-Verlag, New York, 2001.

[Arv02] W. Arveson. Generators of noncommutative dynamics. *Erg. Th. Dyn. Syst.*, 22:1017–1030, 2002. arXiv:math.OA/0201137.

[Arv03] W. Arveson. *Noncommutative Dynamics and $E$-semigroups*. Monographs in Mathematics. Springer-Verlag, New York, 2003.

[Bha99] B.V.R. Bhat. Minimal dilations of quantum dynamical semigroups to semigroups of endomorphisms of $C^*$-algebras. *J. Ramanujan Math. Soc.*, 14(2):109–124, 1999.

[BLM04] D. Blecher and C. Le Merdy. *Operator algebras and their modules*, volume 30 of *LMS Monographs*. Clarendon Press, Oxford, 2004.

[BP94] B.V.R. Bhat and K.R. Parthasarathy. Kolmogorov's existence theorem for Markov processes in $C^*$-algebras. *Proc. Indian Acad. Sci.* (*Math. Sci.*), 104:253–262, 1994.

[Bra55] J. Bram. Subnormal operators. *Duke Math. J.*, 22(1):75–94, 1955.

[Dix64] J. Dixmier. *Les $C^*$-algèbres et leurs Représentations*. Gauthier–Villars, Paris, 1964.

[ER00] E.G. Effros and Z.-J. Ruan. *Operator Spaces*. Oxford University Press, Oxford, 2000.

[Hal50] P.R. Halmos. Normal dilations and extensions of operators. *Summa Brasil.*, 2:125–134, 1950.

[Hal67] P. Halmos. *A Hilbert space problem book*. Van Nostrand, Princeton, 1967.

[MS02] P. Muhly and B. Solel. Quantum Markov processes (correspondences and dilations). *Int. J. Math.*, 13(8):863–906, 2002.

[Nai43a] M.A. Naimark. On the representation of additive operator set functions. *C.R.* (*Dokl.*) *Acad. Sci. URSS*, 41:359–361, 1943.

[Nai43b] M.A. Naimark. Positive definite operator functions on a commutative group. *Bull.* (*Izv.*) *Acad. Sci. URSS* (*Ser. Math.*), 7:237–244, 1943.

[Par91] K.R. Parthasarathy. *An introduction to quantum stochastic calculus*, volume I. Birkhäuser Verlag, Basel, 1991.

[Pau86] V. Paulsen. *Completely bounded maps and dilations*. Wiley, New York, 1986.

[Pau02] V. Paulsen. *Completely bounded maps and operator algebras*. Cambridge University Press, Cambridge, UK, 2002.

[SN53] B. Sz.-Nagy. Sur les contractions de l'espace de Hilbert. *Acta Sci. Math. Szeged*, 15:87–92, 1953.

[SNF70] B. Sz.-Nagy and C. Foias. *Harmonic analysis of operators on Hilbert space*. American Elsevier, New York, 1970.

[Sti55] W.F. Stinespring. Positive functions on $C^*$-algebras. *Proc. Amer. Math. Soc.*, 6:211–216, 1955.

[Szf77] F.H. Szfraniec. Dilations on involution semigroups. *Proc. Amer. Math. Soc.*, 66(1):30–32, 1977.

[vN51] J. von Neumann. Eine Spektraltheorie für allgemeine Operatoren eines unitären Raumes. *Math. Nachr.*, 4:258–281, 1951.

[Wit81] G. Wittstock. Ein operatorwertiger Hahn-Banach Satz. *J. Funct. Anal.*, 40:127–150, 1981.

[Wit84] G. Wittstock. On matrix order and convexity. In *Functional Analysis: Surveys and recent results*, volume 90 of *Math. Studies*, pages 175–188. North-Holland, 1984.

William Arveson
Department of Mathematics
University of California
Berkeley, CA 94720, USA
e-mail: `arveson@math.berkeley.edu`

Operator Theory:
Advances and Applications, Vol. 207, 125–133

# Toeplitz Operators

Sheldon Axler

**Abstract.** This article discusses Paul Halmos's crucial work on Toeplitz operators and the consequences of that work.

**Mathematics Subject Classification (2000).** 47B35.

**Keywords.** Toeplitz operator, Paul Halmos.

A *Toeplitz matrix* is a matrix that is constant on each line parallel to the main diagonal. Thus a Toeplitz matrix looks like this:

$$\begin{matrix} a_0 & a_{-1} & a_{-2} & a_{-3} & \cdots \\ a_1 & a_0 & a_{-1} & a_{-2} & \ddots \\ a_2 & a_1 & a_0 & a_{-1} & \ddots \\ a_3 & a_2 & a_1 & a_0 & \ddots \\ \vdots & \ddots & \ddots & \ddots & \ddots \end{matrix}$$

In this article, Toeplitz matrices have infinitely many rows and columns, indexed by the nonnegative integers, and the entries of the matrix are complex numbers. Thus a Toeplitz matrix is determined by a two-sided sequence $(a_n)_{n=-\infty}^{\infty}$ of complex numbers, with the entry in row $j$, column $k$ (for $j, k \geq 0$) of the Toeplitz matrix equal to $a_{j-k}$.

We can think of the Toeplitz matrix above as acting on the usual Hilbert space $\ell^2$ of square-summable sequences of complex numbers, equipped with its standard orthonormal basis. The question then arises of characterizing the two-sided sequences $(a_n)_{n=-\infty}^{\infty}$ of complex numbers such that the corresponding Toeplitz matrix is the matrix of a bounded operator on $\ell^2$. The answer to this question points toward the fascinating connection between Toeplitz operators and complex function theory.

This paper is an extension and modification of the author's article Paul Halmos and Toeplitz Operators, which was published in *Paul Halmos: Celebrating* 50 *Years of Mathematics*, Springer, 1991, edited by John H. Ewing and F.W. Gehring.

Let $D$ denote the open unit disk in the complex plane and let $\sigma$ denote the usual arc length measure on the unit circle $\partial D$, normalized so that $\sigma(\partial D) = 1$. For $f \in L^1(\partial D, \sigma)$ and $n$ an integer, the $n^{\text{th}}$ Fourier coefficient of $f$, denoted $\hat{f}(n)$, is defined by

$$\hat{f}(n) = \int_{\partial D} f(z)\overline{z^n}\, d\sigma(z).$$

The characterization of the Toeplitz matrices that represent bounded operators on $\ell^2$ is now given by the following result.

**Theorem 1.** *The Toeplitz matrix corresponding to a two-sided sequence* $(a_n)_{n=-\infty}^{\infty}$ *of complex numbers is the matrix of a bounded operator on* $\ell^2$ *if and only if there exists a function* $f \in L^\infty(\partial D, \sigma)$ *such that* $a_n = \hat{f}(n)$ *for every integer* $n$.

The result above first seems to have appeared in print in the Appendix of a 1954 paper by Hartman and Wintner [16], although several decades earlier Otto Toeplitz had proved the result in the special case of symmetric Toeplitz matrices (meaning that $a_{-n} = \overline{a_n}$ for each integer $n$). One direction of the result above is an easy consequence of adopting the right viewpoint. Specifically, the Hardy space $H^2$ is defined to be the closed linear span in $L^2(\partial D, \sigma)$ of $\{z^n : n \geq 0\}$. For $f \in L^\infty(\partial D, \sigma)$, the Toeplitz operator with symbol $f$, denoted $T_f$, is the operator on $H^2$ defined by

$$T_f h = P(fh),$$

where $P$ denotes the orthogonal projection of $L^2(\partial D, \sigma)$ onto $H^2$. Clearly $T_f$ is a bounded operator on $H^2$. The matrix of $T_f$ with respect to the orthonormal basis $(z^n)_{n=0}^{\infty}$ is the Toeplitz matrix corresponding to the two-sided sequence $(\hat{f}(n))_{n=-\infty}^{\infty}$, thus proving one direction of Theorem 1.

## Products of Toeplitz operators

Paul Halmos's first paper on Toeplitz operators was a joint effort with Arlen Brown published in 1964 [5]. The Brown/Halmos paper set the tone for much of the later work on Toeplitz operators. Some of the results in the paper now seem easy, perhaps because in 1967 Halmos incorporated them into the chapter on Toeplitz operators in his marvellous and unique *A Hilbert Space Problem Book* [11], from which several generations of operator theorists have learned the tools of the trade. Multiple papers have been published in the 1960s, 1970s, 1980s, 1990s, and the 2000s that extend and generalize results that first appeared in the Brown/Halmos paper. Although it is probably the most widely cited paper ever written on Toeplitz operators, Halmos records in his automathography ([12], pages 319–321) that this paper was rejected by the first journal to which it was submitted before being accepted by the prestigious Crelle's journal.

The Brown/Halmos paper emphasized the difficulties flowing from the observation that the linear map $f \mapsto T_f$ is not multiplicative. Specifically, $T_f T_g$ is rarely equal to $T_{fg}$. Brown and Halmos discovered precisely when $T_f T_g = T_{fg}$. To state their result, first we recall that the Hardy space $H^\infty$ is defined to be the set of functions $f$ in $L^\infty(\partial D, \sigma)$ such that $\hat{f}(n) = 0$ for every $n < 0$. Note that a function $f \in L^\infty(\partial D, \sigma)$ is in $H^\infty$ if and only if the matrix of $T_f$ is a lower-triangular matrix. Similarly, the matrix of $T_f$ is an upper-triangular matrix if and only if $\bar{f} \in H^\infty$. The Brown/Halmos paper gives the following characterization of which Toeplitz operators multiply well.

**Theorem 2.** *Suppose $f, g \in L^\infty(\partial D, \sigma)$. Then $T_f T_g = T_{fg}$ if and only if either $\bar{f}$ or $g$ is in $H^\infty$.*

As a consequence of the result above, the Brown/Halmos paper shows that there are no zero divisors among the set of Toeplitz operators:

**Theorem 3.** *If $f, g \in L^\infty(\partial D, \sigma)$ and $T_f T_g = 0$, then either $f = 0$ or $g = 0$.*

The theorem above naturally leads to the following question:

**Question 1.** *Suppose $f_1, f_2, \dots, f_n \in L^\infty(\partial D, \sigma)$ and*

$$T_{f_1} T_{f_2} \dots T_{f_n} = 0.$$

*Must some $f_j = 0$?*

Halmos did not put the question above in print, but I heard him raise and popularize it at a number of conferences. The Brown/Halmos paper shows that the question above has answer "yes" if $n = 2$. Several people extended the result to $n = 3$, but after that progress was painfully slow. In 1996 Kun Yu Guo [9] proved that the question above has an affirmative answer if $n = 5$. In 2000 Caixing Gu [8] extended the positive result to the case where $n = 6$. Recently Alexandru Aleman and Dragan Vukotić [1] completely solved the problem, cleverly showing that the question above has an affirmative answer for all values of $n$.

## The spectrum of a Toeplitz operator

Recall that the spectrum of a linear operator $T$ is the set of complex numbers $\lambda$ such that $T - \lambda I$ is not invertible; here $I$ denotes the identity operator. The Brown/Halmos paper contains the following result, which was the starting point for later deep work about the spectrum of a Toeplitz operator.

**Theorem 4.** *The spectrum of a Toeplitz operator cannot consist of exactly two points.*

In the best Halmosian tradition, the Brown/Halmos paper suggests an open problem as a yes/no question:

**Question 2.** *Can the spectrum of a Toeplitz operator consist of exactly three points?*

A bit later, in [10] (which was written after the Brown/Halmos paper although published slightly earlier) Halmos asked the following bolder question.

**Question 3.** *Does every Toeplitz operator have a connected spectrum?*

This has always struck me as an audacious question, considering what was known at the time. The answer was known to be "yes" when the symbols are required to be either real valued or in $H^\infty$, but these are extremely special and unrepresentative cases. For the general complex-valued function, even the possibility that the spectrum could consist of exactly three points had not been eliminated when Halmos posed the question above.

Nevertheless, Harold Widom [19] soon answered Halmos's question by proving the following theorem (the essential spectrum of an operator $T$ is the set of complex numbers $\lambda$ such that $T - \lambda I$ is not invertible modulo the compact operators).

**Theorem 5.** *Every Toeplitz operator has a connected spectrum and a connected essential spectrum.*

Ron Douglas [7] has written that Widom's proof of the theorem above is unsatisfactory because "the proof gives us no hint as to why the result is true", but no alternative proof has been found.

## Subnormal Toeplitz operators

Recall that a linear operator $T$ is called normal if it commutes with its adjoint ($T^*T = TT^*$). The Brown/Halmos paper gave the following characterization of the normal Toeplitz operators.

**Theorem 6.** *Suppose $f \in L^\infty(\partial D, \sigma)$. Then $T_f$ is normal if and only if there is a real-valued function $g \in L^\infty(\partial D, \sigma)$ and complex constants $a, b$ such that $f = ag+b$.*

One direction of the theorem above is easy because if $g$ is a real-valued function in $L^\infty(\partial D, \sigma)$ then $T_g$ is self-adjoint, which implies that $aT_g+bI$ is normal for all complex constants $a, b$.

A Toeplitz operator is called analytic if its symbol is in $H^\infty$. The reason for this terminology is that the Fourier series of a function $f \in L^1(\partial D, \sigma)$, which is the formal sum

$$\sum_{n=-\infty}^{\infty} \hat{f}(n) z^n,$$

is the Taylor series expansion

$$\sum_{n=0}^{\infty} \hat{f}(n) z^n$$

of an analytic function on the unit disk if $\hat{f}(n) = 0$ for all $n < 0$.

An operator $S$ on a Hilbert space $H$ is called subnormal if there is a Hilbert space $K$ containing $H$ and a normal operator $T$ on $K$ such that $T|_K = S$. For example, if $f \in H^\infty$, then the Toeplitz operator $T_f$ is subnormal, as can be seen by considering the Hilbert space $L^2(\partial D, \sigma)$ and the normal operator of multiplication by $f$ on $L^2(\partial D, \sigma)$. Thus every analytic Toeplitz operator is subnormal.

All normal Toeplitz operators and all analytic Toeplitz operators are subnormal. These two classes of Toeplitz operators were the only known examples of subnormal Toeplitz operators in 1970 when Paul Halmos gave a famous series of lectures [14] in which he posed ten open problems in operator theory. One of Halmos's ten questions asked if there were any other examples:

**Question 4.** *Is every subnormal Toeplitz operator either normal or analytic?*

In 1979 Halmos described [15] what had happened to the ten problems in the years since they had been posed. The problem about Toeplitz operators was still unsolved, but Halmos's question had stimulated good work on the problem. Several papers had been written with partial results providing strong evidence that the question above had an affirmative answer.

In the spring of 1983 I believed that the time was right for a breakthrough on this problem, so I organized a seminar at Michigan State University to focus on this problem. We went through every paper on this topic, including a first draft of a manuscript by Shun-Hua Sun. Sun claimed to have proved that no nonanalytic Toeplitz operator can lie in a certain important subclass of the subnormal operators. There was a uncorrectable error in the proof (and the result is false), but Sun had introduced clever new ideas to the subject. His proof worked for all but a single family of operators, and thus this particular family was an excellent candidate for a counter-example that no one expected to exist.

The Spring quarter ended and I sent a copy of my seminar notes to Carl Cowen at Purdue University. When I returned from a long trip abroad, I found a letter from Cowen, who had amazingly answered Halmos's question (negatively!) by proving that each operator in the suspicious family singled out by Sun's work is a subnormal Toeplitz operator that is neither normal nor analytic. Here is what Cowen had proved, where we are abusing notation and thinking of $f$, which starts out as a function on $D$, as also a function on $\partial D$ (just extend $f$ by continuity to $\partial D$):

**Theorem 7.** *Suppose $b \in (0,1)$. Let $f$ be a one-to-one analytic mapping of the unit disk onto the ellipse with vertices $\frac{1}{1+b}$, $\frac{-1}{1+b}$, $\frac{i}{1-b}$, and $\frac{-i}{1-b}$. Then the Toeplitz operator with symbol $f + b\bar{f}$ is subnormal but is neither normal nor analytic.*

I told my PhD student John Long about Cowen's wonderful result, although I did not show Long the proof. Within a week, Long came back to me with a beautiful and deep proof that was shorter and more natural than Cowen's. Because there was now no reason to publish Cowen's original proof, Cowen and Long decided to publish Long's proof in a joint paper [6]. Thus the contributions to that paper are as follows: Cowen first proved the result and provided the crucial knowledge of the correct answer, including the idea of using ellipses; the proof in the paper is due to Long. At no time did the two authors actually work together.

## The symbol map

Paul Halmos's second major paper on Toeplitz operators was a joint effort with José Barría that was published in 1982 [4]. The main object of investigation is the Toeplitz algebra $\mathcal{T}$, which is defined to be the norm-closed algebra generated by all the Toeplitz operators on $H^2$. The most important tool in the study of $\mathcal{T}$ is what is called the symbol map $\varphi$, as described in the next theorem.

**Theorem 8.** *There exists a unique multiplicative linear map $\varphi : \mathcal{T} \to L^\infty(\partial D, \sigma)$ such that $\varphi(T_f) = f$ for every $f \in L^\infty(\partial D, \sigma)$.*

The surprising point here is the existence of a multiplicative map on $\mathcal{T}$ such that $\varphi(T_f) = f$ for every $f \in L^\infty(\partial D, \sigma)$. Thus

$$\varphi(T_f T_g) = \varphi(T_f)\varphi(T_g) = fg$$

for all $f, g \in L^\infty(\partial D, \sigma)$. The symbol map $\varphi$ was discovered and exploited by Douglas ([7], Chapter 7).

The symbol map $\varphi$ was a magical and mysterious homomorphism to me until I read the Barría/Halmos paper, where the authors actually construct $\varphi$ (as opposed to Douglas's more abstract proof).

Here is how the Barría/Halmos paper constructs $\varphi$: The authors prove that if $S \in \mathcal{T}$, then $S$ is an asymptotic Toeplitz operator in the sense that in the matrix of $S$, the limit along each line parallel to the main diagonal exists. Consider a Toeplitz matrix in which each line parallel to the main diagonal contains the limit of the corresponding line from the matrix of $S$. The nature of the construction ensures that this Toeplitz matrix represents a bounded operator and thus is the matrix of $T_f$ for some $f \in L^\infty(\partial D, \sigma)$. Starting with $S \in \mathcal{T}$, we have now obtained a function $f \in L^\infty(\partial D, \sigma)$. Define $\varphi(S)$ to be $f$. Then $\varphi$ is the symbol map whose existence is guaranteed by Theorem 8.

A more formal statement of the Barría/Halmos result is given below. Here we are using $\varphi$ as in Theorem 8. Thus the point here is that we can actually construct the symbol map $\varphi$.

**Theorem 9.** *Suppose $S \in \mathcal{T}$ and the matrix of $S$ with respect to the standard basis of $H^2$ is $(b_{j,k})_{j,k=0}^{\infty}$. Then for each integer $n$, the limit (as $j \to \infty$) of $b_{n+j,j}$ exists. Let*

$$a_n = \lim_{j\to\infty} b_{n+j,j}$$

*and let*

$$f = \sum_{n=-\infty}^{\infty} a_n z^n,$$

*where the infinite sum converges in the norm of $L^2(\partial D, \sigma)$. Then $f \in L^\infty(\partial D, \sigma)$ and $\varphi(S) = f$.*

The Barría/Halmos construction of $\varphi$ is completely different in spirit and technique from Douglas's existence proof. I knew Douglas's proof well—an idea that I got from reading it was a key ingredient in my first published paper [2]. But until the Barría/Halmos paper came along, I never guessed that $\varphi$ could be explicitly constructed or that so much additional insight could be squeezed from a new approach.

## Compact semi-commutators

An operator of the form $T_fT_g - T_{fg}$ is called a semi-commutator. As discussed earlier, the Brown/Halmos paper gave a necessary and sufficient condition on functions $f, g \in L^\infty(\partial D, \sigma)$ for the semi-commutator $T_fT_g - T_{fg}$ to equal 0. One of the fruitful strands of generalization stemming from this result involves asking for $T_fT_g - T_{fg}$ to be small in some sense. In this context, the most useful way an operator can be small is to be compact.

In 1978 Sun-Yung Alice Chang, Don Sarason, and I published a paper [3] giving a sufficient condition on functions $f, g \in L^\infty(\partial D, \sigma)$ for $T_fT_g - T_{fg}$ to be compact. This condition included all previously known sufficient conditions. To describe this condition, for $g \in L^\infty(\partial D, \sigma)$ let $H^\infty[g]$ denote the smallest norm-closed subalgebra of $L^\infty(\partial D, \sigma)$ containing $H^\infty$ and $g$. The Axler/Chang/Sarason paper showed that if $f, g \in L^\infty(\partial D, \sigma)$ and

$$H^\infty[\bar{f}] \cap H^\infty[g] \subset H^\infty + C(\partial D),$$

then $T_fT_g - T_{fg}$ is compact.

We could prove that the condition above was necessary as well as sufficient if we put some additional hypotheses on $f$ and $g$. We conjectured that the condition above was necessary without the additional hypotheses, but we were unable to prove so.

A brilliant proof verifying the conjecture was published by Alexander Volberg [18] in 1982. Combining Volberg's result of the necessity with the previous result of the sufficiency gives the following theorem.

**Theorem 10.** *Suppose $f, g \in L^\infty(\partial D, \sigma)$. Then $T_fT_g - T_{fg}$ is compact if and only if $H^\infty[\bar{f}] \cap H^\infty[g] \subset H^\infty + C(\partial D)$.*

A key step in Volberg's proof of the necessity uses the following specific case of a theorem about interpolation of operators that had been proved 26 years earlier by Elias Stein ([17], Theorem 2).

**Theorem 11.** *Let $d\mu$ be a positive measure on a set $X$, and let $v$ and $w$ be positive measurable functions on $X$. Suppose $S$ is a linear operator on both $L^2(vd\mu)$ and $L^2(wd\mu)$, with norms $\|S\|_v$ and $\|S\|_w$, respectively. If $\|S\|_{\sqrt{vw}}$ denotes the norm of $S$ on $L^2(\sqrt{vw}d\mu)$, then*

$$\|S\|_{\sqrt{vw}} \le \sqrt{\|S\|_v \, \|S\|_w}.$$

When I received a preprint of Volberg's paper in Spring 1981 I told Paul Halmos about the special interpolation result that it used. Within a few days

Halmos surprised me by producing a clean Hilbert space proof of the interpolation result above that Volberg had needed. Halmos's proof (for the special case of Theorem 11) was much nicer than Stein's original proof. With his typical efficiency, Halmos put his inspiration into publishable form quickly and submitted the paper to the journal to which Volberg has submitted his article. I ended up as the referee for both papers. It was an unusual pleasure to see how a tool used in one paper had led to an improved proof of the tool. Halmos's short and delightful paper [13] containing his proof of the interpolation result was published in the same issue of the *Journal of Operator Theory* as Volberg's paper.

## Remembering Paul Halmos

I would like to close with a few words about my personal debt to Paul Halmos. Paul is my mathematical grandfather. His articles and books have been an important part of my mathematical heritage. I first met Paul for a few seconds when I was a graduate student, and then for a few minutes when I gave my first conference talk right after receiving my PhD. Four years later I got to know Paul well when I spent a year's leave at Indiana University. Later when Paul became Editor of the *American Mathematical Monthly*, he selected me as one of the Associate Editors. Still later Paul and I spent several years working together as members of the Editorial Board for the Springer series Graduate Texts in Mathematics, Undergraduate Texts in Mathematics, and Universitext.

Paul is one of the three people who showed me how to be a mathematician (the other two are my wonderful thesis advisor Don Sarason, who was Paul's student, and Allen Shields). Watching Paul, I saw how an expert proved a theorem, gave a talk, wrote a paper, composed a referee's report, edited a journal, and edited a book series. I'm extremely lucky to have had such an extraordinary model.

## References

[1] Alexandru Aleman and Dragan Vukotić, Zero products of Toeplitz operators, *Duke Math. J.* 148 (2009), 373–403.

[2] Sheldon Axler, Factorization of $L^\infty$ functions, *Ann. Math.* 106 (1977), 567–572.

[3] Sheldon Axler, Sun-Yung A. Chang, and Donald Sarason, Products of Toeplitz operators, *Integral Equations Operator Theory* 1 1978, 285–309.

[4] José Barría and P.R. Halmos, Asymptotic Toeplitz operators, *Trans. Am. Math. Soc.* 273 (1982), 621–630.

[5] Arlen Brown and P.R. Halmos, Algebraic properties of Toeplitz operators, *J. Reine Angew. Math.* 213 (1964), 89–102.

[6] Carl C. Cowen and John J. Long, Some subnormal Toeplitz operators, *J. Reine Angew. Math.* 351 (1984), 216–220.

[7] Ronald G. Douglas, *Banach Algebra Techniques in Operator Theory*, Academic Press, 1972; second edition published by Springer, 1998.

[8] Caixing Gu, Products of several Toeplitz operators, *J. Funct. Anal.* 171 (2000), 483–527.

[9] Kun Yu Guo, A problem on products of Toeplitz operators, *Proc. Amer. Math. Soc.* 124 (1996), 869–871.

[10] P.R. Halmos, A glimpse into Hilbert space, *Lectures on Modern Mathematics*, Vol. I, edited by T.L. Saaty, Wiley, 1963, 1–22.

[11] P.R. Halmos, *A Hilbert Space Problem Book*, Van Nostrand, 1967; second edition published by Springer, 1982.

[12] P.R. Halmos, *I Want to Be a Mathematician*, Springer, 1985.

[13] P.R. Halmos, Quadratic interpolation, *J. Operator Theory* 7 (1982), 303–305.

[14] P.R. Halmos, Ten problems in Hilbert space, *Bull. Am. Math. Soc.* 76 (1970), 887–993.

[15] P.R. Halmos, Ten years in Hilbert space, *Integral Equations Operator Theory* 2 (1979), 529–564.

[16] Philip Hartman and Aurel Wintner, The spectra of Toeplitz's matrices, *Amer. J. Math.* 76 (1954), 867–882.

[17] Elias M. Stein, Interpolation of linear operators, *Trans. Am. Math. Soc.* 83 (1956), 482–492.

[18] A.L. Volberg, Two remarks concerning the theorem of S. Axler, S.-Y.A. Chang and D. Sarason, *J. Operator Theory* 7 (1982), 209–218.

[19] Harold Widom, On the spectrum of a Toeplitz operator, *Pacific J. Math.* 14 (1964), 365–375.

Sheldon Axler
Mathematics Department
San Francisco State University
San Francisco, CA 94132, USA
e-mail: `axler@sfsu.edu`

Operator Theory:
Advances and Applications, Vol. 207, 135–176

# Dual Algebras and Invariant Subspaces

Hari Bercovici

*In memory of Paul Halmos, who knew how to ask the right questions*

**Abstract.** We will discuss methods for proving invariant space results which were first introduced by Scott Brown for subnormal operators. These methods are now related with the idea of a dual algebra.

**Mathematics Subject Classification (2000).** Primary 47L45; Secondary 47A15, 47A45, 47B20, 47B48.

**Keywords.** Dual algebra, invariant subspace, contraction, subnormal operator, hyper-reflexivity, dominating set, functional calculus.

## 1. Introduction

Consider a complex Banach space $\mathfrak{H}$ of infinite dimension, and the algebra $\mathcal{B}(\mathfrak{H})$ of bounded linear operators acting on $\mathfrak{H}$. A (closed, linear) subspace $\mathfrak{M} \subset \mathfrak{H}$ is said to be *invariant* for an operator $T \in \mathcal{B}(\mathfrak{H})$ if $T\mathfrak{M} \subset \mathfrak{M}$. We will be concerned with the existence of *proper* invariant subspaces $\mathfrak{M}$, i.e., spaces different from $\{0\}$ and $\mathfrak{H}$. We denote by $\mathfrak{H}^*$ the dual of $\mathfrak{H}$. Given a vector $x \in \mathfrak{H}$ and a functional $\varphi \in \mathfrak{H}^*$, we define the continuous linear functional $x \otimes \varphi \in \mathcal{B}(\mathfrak{H})^*$ by setting

$$(x \otimes \varphi)(T) = \varphi(Tx), \quad T \in \mathcal{B}(\mathfrak{H}).$$

Fix now $T \in \mathcal{B}(\mathfrak{H})$, and denote by $\mathcal{P}_T$ the unital subalgebra of $\mathcal{B}(\mathfrak{H})$ generated by $T$. Thus, $\mathcal{P}_T$ consists of all operators of the form $p(T)$ with $p \in \mathbb{C}[z]$ a polynomial. The linear functional on $\mathbb{C}[z]$ defined by $p \mapsto (x \otimes \varphi)(p(T))$ will be denoted $x \otimes_T \varphi$.

**Lemma 1.1.** *The operator $T \in \mathcal{B}(\mathfrak{H})$ has a nontrivial invariant subspace if and only if there exist $x \in \mathfrak{H} \setminus \{0\}$ and $\varphi \in \mathfrak{H}^* \setminus \{0\}$ such that $x \otimes_T \varphi = 0$.*

*Proof.* The equality $\varphi(p(T)x) = 0$ simply means that $\varphi$ is zero on the cyclic space for $T$ generated by $x$. This space is not zero if $x \neq 0$, and it is not $\mathfrak{H}$ if $\varphi \neq 0$. □

The author was supported in part by a grant from the National Science Foundation.

This result has been of limited applicability because there is no obvious way to insure that the factors $\varphi$ and $x$ are not zero when $x \otimes_T \varphi = 0$. There are however some nonzero functionals $\psi$ on $\mathbb{C}[z]$ with the property that an equality of the form $x \otimes_T \varphi = \psi$ also implies the existence of nontrivial invariant subspaces. For fixed $\lambda \in \mathbb{C}$, denote by $e_\lambda : \mathbb{C}[z] \to \mathbb{C}$ the functional of evaluation at $\lambda$:

$$e_\lambda(p) = p(\lambda), \quad p \in \mathbb{C}[z].$$

**Lemma 1.2.** *Assume that $x \in \mathfrak{H}$, $\varphi \in \mathfrak{H}^*$, and $\lambda \in \mathbb{C}$ are such that $x \otimes_T \varphi = e_\lambda$, and denote by $\mathfrak{M}$ the cyclic space for $T$ generated by $x$. Then $\overline{(T - \lambda)\mathfrak{M}} \neq \mathfrak{M}$; in particular $T$ has nontrivial invariant subspaces.*

*Proof.* We have $\varphi|\mathfrak{M} \neq 0$ because $\varphi(x) = (x \otimes_T \varphi)(1) = 1$, and $\varphi|\overline{(T - \lambda)\mathfrak{M}} = 0$ since $\varphi((T - \lambda)^n x) = (x \otimes_T \varphi)((z - \lambda)^n) = e_\lambda((z - \lambda)^n) = 0$ for $n \geq 1$. □

Not all operators with a nontrivial invariant subspace satisfy the hypotheses of the preceding lemma. For instance, a compact quasi-nilpotent operator with trivial kernel has no invariant subspaces $\mathfrak{M}$ such that $\overline{(T - \lambda)\mathfrak{M}} \neq \mathfrak{M}$ for some $\lambda \in \mathbb{C}$. There are however many operators to which the result can be applied. This idea was pioneered by Scott Brown, who showed in [34] that all subnormal operators (in the sense defined by Halmos [66]) on a Hilbert space have nontrivial invariant subspaces. Since then, the technique has been extended to cover many classes of operators on Hilbert and Banach spaces, and it is our goal to give an exposition of the main methods developed for this purpose.

Let us note at this point that an invariant subspace for $T$ is also invariant for every operator in the weak operator closure $\mathcal{W}_T$ of $\mathcal{P}_T$. In case $\mathfrak{H}$ is a Hilbert space, $\mathcal{W}_T$ is also closed in the weak* topology given by identifying $\mathcal{B}(\mathfrak{H})$ with the dual of the trace class. Algebras closed in this topology are now known as *dual algebras*. The functionals $x \otimes \varphi$ are always continuous in the weak* topology, and the sharpest factorization results show that for some dual algebras $\mathcal{A}$, every weak* continuous functional on $\mathcal{A}$ is of the form $(x \otimes \varphi)|\mathcal{A}$. Indeed, one of the features of these developments is that one can prove the existence of a factorization $e_\lambda = x \otimes_T \varphi$ only by showing that a much larger class of functionals can be factored in this way. These factorization theorems do in fact yield further insight into the structure of the invariant subspace lattice. In particular, they demonstrate the difficulty of one of Halmos's famous ten problems [67]. Namely, Problem 3 asks whether an operator $T$ must have nontrivial invariant subspaces if $T^2$ or $T^{-1}$ do. Let $T$ be a normal diagonal operator whose eigenvalues $\Lambda = \{\lambda_n\}_{n=1}^\infty$ satisfy $|\lambda_j| > 1$, and the nontangential cluster set of $\Lambda$ is the unit circle. The set $\Lambda$ can be chosen so that $T$ is reductive, while Corollary 7.2 implies that $T^{-1}$ has a vast supply of nonreducing subspaces. Thus the techniques of this paper, applied to $T^{-1}$, will never yield one of the invariant subspaces of $T$.

These notes are organized as follows. In Section 2 we present an open mapping theorem for bilinear maps. This result is an abstract form of the central argument of [34], and it is formulated so as to be applicable to operators on Banach spaces. The proofs in the case of Hilbert space, which also yield additional information, are presented in full. The adaptation to Banach spaces requires almost no effort. Section 3 contains a hyper-reflexivity theorem for weak operator algebras on a Hilbert space. This result follows from the open mapping theorem. In Sections 4 and 5, we show that these abstract results apply in very concrete situations. Section 4 deals with contractions on a Hilbert space, whose spectra are dominating in the unit disk. The existence of invariant subspaces for such operators was first proved in [38]. Section 5 demonstrates the hyper-reflexivity of an algebra which commutes with two isometries with orthogonal ranges. An example is provided by the free semigroup algebra, whose hyper-reflexivity was first proved in [56]. Section 6 contains a second (approximate) factorization technique. The main result is an abstract form of the central result in [36]. This result is needed in Section 7, where the fundamental result of [39] is proved: a contraction on Hilbert space has invariant subspaces if its spectrum contains the unit circle. Here, as well as in Section 4, the Sz.-Nagy–Foias functional model for contractions plays a central role. In Section 8 we present two Banach space techniques which are useful in going beyond Hilbert space. The elementary but ingenious Theorems 8.1 and 8.2 are due to Zenger [101]. Section 9 illustrates the use of the open mapping result of Proposition 2.5 in proving that certain contractions on Banach spaces have nontrivial invariant subspaces if they have dominating spectra. Here the existence of an analogue of the Sz.-Nagy–Foias functional calculus must be postulated. The best result in this direction is due to Ambrozie and Müller [8], who showed the existence of invariant subspaces when the spectrum contains the unit circle. The arguments we present contain many of the ideas in the proof of this general result. In Section 10 we present an invariant subspace result for operators with large localizable spectra. The use of local spectral theory in this context was first suggested by Apostol [13], and one of the first significant results was Brown's theorem [35] concerning hyponormal operators with large spectra. Our presentation follows [63], with some simplifications due to the fact that we restrict ourselves to localizable spectra in the unit disk. Finally, in Section 11 we provide some historical perspective, and a rough guide to the sizeable literature, some of which is collected in the references.

## 2. An open mapping theorem for bilinear maps

S. Brown's result [34] relies on an open mapping theorem for bilinear maps which can be proved in a fairly abstract setting. Fix two Hilbert spaces $\mathfrak{H}, \mathfrak{K}$, a Banach space $\mathfrak{X}$, and a continuous bilinear map $F : \mathfrak{H} \times \mathfrak{K} \to \mathfrak{X}$. The scalar product in a Hilbert space will be denoted $\langle \cdot, \cdot \rangle$. We will denote by $S_F$ the collection of those vectors $x \in \mathfrak{X}$ with the following property: Given $\varepsilon > 0$ and finitely many vectors $h_1, h_2, \ldots, h_n \in \mathfrak{H}$ and $k_1, k_2, \ldots, k_n \in \mathfrak{K}$, there exist $u \in \mathfrak{H}$ and $v \in \mathfrak{K}$ such that

(a) $\|u\|, \|v\| < 1$,
(b) $\|x - F(u, v)\| < \varepsilon$,
(c) $|\langle u, h_\ell \rangle| + |\langle v, k_\ell \rangle| < \varepsilon$ for $\ell = 1, 2, \ldots, n$, and
(d) $\|F(u, k_\ell)\| + \|F(h_\ell, v)\| < \varepsilon$ for $\ell = 1, 2, \ldots, n$.

It is obvious that the set $S_F$ is closed and balanced (i.e., $\lambda x \in S_F$ if $x \in S_F$ and $|\lambda| \leq 1$). It is also clear that condition (a) could be replaced by the weaker condition

(a′) $\|u\|, \|v\| < 1 + \varepsilon$.

**Lemma 2.1.** *The set $S_F$ is convex.*

*Proof.* Consider vectors $x_1, x_2 \in S_F$, $t \in (0,1)$, and fix $\varepsilon > 0$ and $h_1, h_2, \dots, h_n \in \mathfrak{H}$, $k_1, k_2, \dots, k_n \in \mathfrak{K}$. By definition, there exist $u_i \in \mathfrak{H}$ and $v_i \in \mathfrak{K}$ satisfying conditions (a–d) with $x_i, u_i, v_i$ in place of $x, u, v$, $i = 1, 2$. Choosing $u_2, v_2$ after $u_1, v_1$, we can also require that

$$|\langle u_1, u_2\rangle| + |\langle v_1, v_2\rangle| + \|F(u_1, v_2)\| + \|F(u_2, v_1)\| < \varepsilon.$$

Let us set then

$$u = t^{1/2}u_1 + (1-t)^{1/2}u_2, \quad v = t^{1/2}v_1 + (1-t)^{1/2}v_2.$$

The conditions on $u_i, v_i$ imply immediately that

$$\|u\|^2 \leq t\|u_1\|^2 + (1-t)\|u_2\|^2 + 2[t(1-t)]^{1/2}\varepsilon,$$

so that $\|u\| < 1$ if $\varepsilon$ is sufficiently small, and similarly $\|v\| < 1$. Next, setting $x = tx_1 + (1-t)x_2$, we have

$$\begin{aligned}\|x - F(u,v)\| \leq t\|x_1 - F(u_1, v_1)\| &+ (1-t)\|x_2 - F(u_2, v_2)\| \\ &+ \|F(u_1, v_2)\| + \|F(u_2, v_1)\| < 2\varepsilon.\end{aligned}$$

Clearly $|\langle u, h_\ell\rangle| + |\langle v, k_\ell\rangle| < 2\varepsilon$ and $\|F(u, k_\ell)\| + \|F(h_\ell, v)\| < 2\varepsilon$ for $\ell = 1, 2, \dots, n$. Since $\varepsilon > 0$ is arbitrary, we conclude that $x \in S_F$. □

**Theorem 2.2.** *Assume that $S_F$ contains the ball of radius $\rho^2 > 0$ centered at $0 \in \mathfrak{X}$. Then the map $F$ is open. More precisely, for every $x \in \mathfrak{X}$, $u \in \mathfrak{H}$, and $v \in \mathfrak{K}$ with $\|x - F(u,v)\| < r^2$ for some $r > 0$, there exist $h \in \mathfrak{H}$ and $k \in \mathfrak{K}$ such that $\|h\|, \|k\| < r/\rho$ and $x = F(u+h, v+k)$. Given $\varepsilon > 0$ and vectors $h_i \in \mathfrak{H}$, $k_i \in \mathfrak{K}$, $1 \leq i \leq n$, we can also require that*

$$|\langle h_i, h\rangle| + |\langle k_i, k\rangle| + \|F(h_i, k)\| + \|F(h, k_i)\| < \varepsilon$$

*for all $i$.*

*Proof.* The hypothesis implies that, for every $\alpha > 0$, $\alpha S_F$ contains the ball of radius $\alpha\rho^2$ centered at $0 \in \mathfrak{X}$. We proceed as in the proof of Banach's open mapping theorem. Let $x, u, v$ be as in the statement, and fix a positive number $\varepsilon < 1/2$. We can then find inductively vectors $u_n \in \mathfrak{H}$ and $v_n \in \mathfrak{K}$ such that $u_0 = 0, v_0 = 0$, and the following conditions are satisfied for $n \geq 1$:

(i) $\|u_n\|, \|v_n\| \leq (1/\rho)\|x - F(u + \sum_{\ell=0}^{n-1} u_\ell, v + \sum_{\ell=0}^{n-1} v_\ell)\|^{1/2}$,

(ii) $\|x - F(u + \sum_{\ell=0}^{n-1} u_\ell, v + \sum_{\ell=0}^{n-1} v_\ell) - F(u_n, v_n)\| < \varepsilon^{2n+3}$, and

(iii) $\|F(u_n, v + \sum_{\ell=0}^{n-1} v_\ell)\| + \|F(u + \sum_{\ell=0}^{n-1} u_\ell, v_n)\| < \varepsilon^{2n+3}$.

With the notation $h_n = \sum_{\ell=0}^{n} u_\ell$, $k_n = \sum_{\ell=0}^{n} v_\ell$, we have

$$\begin{aligned}\|x - F(u + h_n, v + k_n)\| &\le \|x - F(u + h_{n-1}, v + k_{n-1}) - F(u_n, v_n)\| \\ &\quad + \|F(u_n, v + k_{n-1})\| + \|F(u + h_{n-1}, v_n)\| \\ &< 2\varepsilon^{2n+3} < \varepsilon^{2(n+1)}\end{aligned}$$

for $n \ge 1$, and condition (ii) implies that $\|u_n\| \le \varepsilon^n/\rho$ for $n \ge 2$. Thus

$$\sum_{n=1}^{\infty} \|u_n\| \le \frac{1}{\rho}\|x - F(u, v)\|^{1/2} + \frac{1}{\rho}\sum_{n=2}^{\infty} \varepsilon^n < \frac{r}{\rho}$$

if $\varepsilon$ is small enough. Similarly, $\sum_{n=1}^{\infty} \|v_n\| < r/\rho$, and hence the vectors $h = \sum_{n=1}^{\infty} u_n$, $k = \sum_{n=1}^{\infty} v_n$ satisfy the conclusion of the theorem. The final assertion is easily verified. One must merely require that

$$|\langle h_i, u_\ell\rangle| + |\langle k_i, v_\ell\rangle| + \|F(h_i, v_\ell)\| + \|F(u_\ell, k_i)\| < 2^{-\ell}\varepsilon$$

for $\ell \ge 1$. □

It is worth noting that the preceding proof uses the continuity of $F$ only to deduce that $F(u + h, v + k) = x$ from the fact that $h_n \to h$, $k_n \to k$, and $F(u + h_n, v + k_n) \to x$. In other words, one could have started with a partially defined bilinear map whose graph is closed.

The conclusion of this open mapping theorem can be strengthened considerably by a simple inflation device. For a natural number $N$, denote by $\ell^2_N(\mathfrak{H})$ the space $\mathfrak{H}^N$ endowed with the usual direct sum Hilbert space norm, and denote by $\ell^1_{N^2}(\mathfrak{X})$ the space of arrays $[x_{ij}]_{i,j=1}^N$ of vectors in $\mathfrak{X}$ endowed with the norm

$$\|[x_{ij}]_{i,j=1}^N\| = \sum_{i,j=1}^{N} \|x_{ij}\|.$$

Given a continuous bilinear map $F : \mathfrak{H} \times \mathfrak{K} \to \mathfrak{X}$, the map $F_N : \ell^2_N(\mathfrak{H}) \times \ell^2_N(\mathfrak{K}) \to \ell^1_{N^2}(\mathfrak{X})$ defined by

$$F_N(h, k) = [F(h_i, k_j)]_{i,j=1}^N, \quad h = (h_i)_{i=1}^N \in \ell^2_N(\mathfrak{H}), k = (k_j)_{j=1}^N \in \ell^2_N(\mathfrak{K}),$$

is also continuous. Using vectors with only one nonzero component, it is easy to verify that elements of the form $[\delta_{ii_0}\delta_{jj_0}x]_{i,j=1}^N$ $(1 \le i_0, j_0 \le N)$ belong to $S_{F_N}$ provided that $x \in S_F$. If $F$ satisfies the hypothesis of Theorem 2.2, it follows now that so does $F_N$, and this yields the following result.

**Corollary 2.3.** *Assume that $S_F$ contains the ball of radius $\rho^2 > 0$ centered at $0 \in \mathfrak{X}$. If $x_{ij} \in \mathfrak{X}$, $u_i \in \mathfrak{H}$, $v_j \in \mathfrak{K}$, $i, j = 1, 2, \ldots, N$, satisfy the inequality*

$$\sum_{i,j=1}^{N} \|x_{ij} - F(u_i, v_j)\| < r^2$$

*for some $r > 0$, then there exist $h_i \in \mathfrak{H}, k_j \in \mathfrak{K}$, $i, j = 1, \dots, N$ such that*

$$\sum_{i=1}^{N} \|h_i\|^2 < \frac{r^2}{\rho^2}, \quad \sum_{j=1}^{N} \|k_j\|^2 < \frac{r^2}{\rho^2},$$

*and $x_{ij} = F(u_i + h_i, v_j + k_j)$ for $i, j = 1, 2, \dots, N$. Given $\varepsilon > 0$ and vectors $a_i \in \mathfrak{H}$, $b_i \in \mathfrak{K}$, $i = 1, 2, \dots, n$, we can also require that*

$$\sum_{i=1}^{N} \sum_{j=1}^{n} (|\langle h_i, a_j \rangle| + |\langle k_i, b_j \rangle| + \|F(h_i, b_j)\| + \|F(a_j, k_i)\|) < \varepsilon.$$

Replacing $N$ by $\aleph_0$ produces a possibly discontinuous map $F_{\aleph_0}$. Nevertheless, we can prove a factorization theorem even in this case. Another approach to this result would be to observe that $F_{\aleph_0}$ can be defined as a closed bilinear map (see the remark following Theorem 2.2).

**Proposition 2.4.** *Under the hypotheses of Theorem 2.2, for any family of vectors $x_{ij} \in \mathfrak{X}$, $i, j = 1, 2, \dots$, there exist sequences $h_i \in \mathfrak{H}$ and $k_j \in \mathfrak{K}$ such that $F(h_i, k_j) = x_{ij}$ for all $i, j \geq 1$.*

*Proof.* The system of equations $F(h_i, k_j) = x_{ij}$ is equivalent to $F(h_i', k_j') = \alpha_i \beta_j x_{ij}$ if $\alpha_i, \beta_j$ are nonzero scalars; just substitute $h_i' = \alpha_i h_i$, $k_j' = \beta_j k_j$. Thus there is no loss of generality in assuming that $\|x_{ij}\| < 2^{-2(i+j+2)}$ for $i, j \geq 1$. For each $N \geq 0$, we construct sequences $h_i(N) \in \mathfrak{H}$, $k_j(N) \in \mathfrak{K}$ such that

(1) $h_i(N) = 0$, $k_i(N) = 0$ for $i > N$,
(2) $F(h_i(N), k_j(N)) = x_{ij}$ for $1 \leq i, j \leq N$, and
(3) $\|h_i(N) - h_i(N-1)\| < 2^{-N}/\rho$, $\|k_i(N) - k_i(N-1)\| < 2^{-N}/\rho$ for $i \geq 1$.

When $N = 0$ we have $h_i(N) = 0$ and $k_i(N) = 0$ for all $i$. If $h_i(N-1), k_j(N-1)$ have already been constructed the existence of $h_i(N)$ and $k_j(N)$ follows from Corollary 2.3 because

$$\sum_{i,j=1}^{N} \|x_{ij} - F(h_i(N-1), k_j(N-1))\| \leq \sum_{i=1}^{\infty} (\|x_{iN}\| + \|x_{Ni}\|) < 2^{-2N}.$$

The sequences $(h_i(N))_{N=1}^{\infty}$, $(k_j(N))_{N=1}^{\infty}$ are obviously Cauchy, and their limits $h_i, k_j$ satisfy the requirements of the proposition. □

The results above can be extended in some form to the case of Banach spaces $\mathfrak{H}, \mathfrak{K}$. Consider a closed bilinear map $F : D(F) \subset \mathfrak{H} \times \mathfrak{K} \to \mathfrak{X}$, where the domain $D(F)$ contains $\mathfrak{H}_0 \times \mathfrak{K}_0$; the linear spaces $\mathfrak{H}_0 \subset \mathfrak{H}$ and $\mathfrak{K}_0 \subset \mathfrak{K}$ need not be either closed or dense. Denote by $\widetilde{S}_F$ the collection of those vectors $x \in \mathfrak{X}$ with the following property: Given $\varepsilon > 0$ and finitely many vectors $h_1, h_2, \dots, h_n \in \mathfrak{H}_0$ and $k_1, k_2, \dots, k_n \in \mathfrak{K}_0$, there exist $u \in \mathfrak{H}_0$ and $v \in \mathfrak{K}_0$ such that

($\widetilde{\text{a}}$) $\|u\|, \|v\| < 1$,
($\widetilde{\text{b}}$) $\|x - F(u, v)\| < \varepsilon$, and
($\widetilde{\text{d}}$) $\|F(u, k_\ell)\| + \|F(h_\ell, v)\| < \varepsilon$ for $\ell = 1, 2, \dots, n$.

This set is still closed and balanced, but we cannot prove its convexity. However, Theorem 2.2 does not use convexity, and is therefore applicable if $\widetilde{S}_F$ contains the ball of radius $\rho^2$ centered at the origin. To treat systems of equations, we consider the map $F_N : D(F_N) \subset \ell^\infty_N(\mathfrak{H}) \times \ell^\infty_N(\mathfrak{K}) \to \ell^\infty_{N^2}(\mathfrak{X})$, where $D(F_N)$ consists of those pairs $(h,k) \in \ell^\infty_N(\mathfrak{H}) \times \ell^\infty_N(\mathfrak{K})$ with the property that $(h_i, k_j) \in D(F)$ for $i,j = 1,2,\dots,N$. If $\widetilde{S}_F$ contains the ball of radius $\rho$ centered at the origin, it is easy to see that $\widetilde{S}_{F_N}$ contains the ball of radius $(\rho/N)^2$ centered at the origin. We obtain the following result.

**Proposition 2.5.** *With the above notation, assume that $\widetilde{S}_F$ contains the ball of radius $\rho^2$ centered at $0 \in \mathfrak{X}$. If $x_{ij} \in \mathfrak{X}$, $u_i \in \mathfrak{H}$, $v_i \in \mathfrak{K}$, satisfy*

$$\|x_{ij} - F(u_i, v_j)\| < r^2, \quad 1 \le i,j \le N,$$

*for some $r > 0$, then there exits $h_i \in \mathfrak{H}$, $k_j \in \mathfrak{K}$ such that $\|h_i\|, \|k_j\| < N(r/\rho)$, and $F(u_i + h_i, v_j + k_j) = x_{ij}$ for $i,j = 1,2,\dots,N$. Moreover, given arbitrary $x_{ij} \in \mathfrak{X}$ for $i,j \ge 1$, there exists sequences $h_i \in \mathfrak{H}$ and $k_j \in \mathfrak{K}$ such that $F(h_i, k_j) = x_{ij}$ for all $i,j \ge 1$.*

## 3. Hyper-reflexivity and dilations

Consider now a Hilbert space $\mathfrak{H}$, and a weak* closed unital subalgebra $\mathcal{W} \subset \mathcal{B}(\mathfrak{H})$. Then $\mathcal{W}$ can be identified with the dual of the space $\mathcal{W}_*$ of all weak* continuous linear functionals on $\mathcal{W}$. The Riesz representation theorem allows us to identify the dual of $\mathfrak{H}$ with $\mathfrak{H}$; we will write $h^* \in \mathfrak{H}^*$ for the functional associated with a vector $h \in \mathfrak{H}$. We are interested in the bilinear map $F : \mathfrak{H} \times \mathfrak{H}^* \to \mathcal{W}_*$ defined by $F(h, k^*) = [h \otimes k^*]_{\mathcal{W}}$, where $[h \otimes k^*]_{\mathcal{W}} = (h \otimes k^*)|\mathcal{W}$ for $h,k \in \mathfrak{H}$. Observe that $F$ is continuous; in fact

$$|\langle h,k\rangle| \le \|[h \otimes k^*]_{\mathcal{W}}\| \le \|h\|\|k\|, \quad h,k \in \mathfrak{H}.$$

For an arbitrary operator $T \in \mathcal{B}(\mathfrak{H})$, we denote by

$$d(T) = \inf\{\|T - S\| : S \in \mathcal{W}\}$$

the distance from $T$ to $\mathcal{W}$, and we set

$$r(T) = \sup\{\|(I-P)TP\| : P \in \operatorname{Lat}(\mathcal{W})\},$$

where $\operatorname{Lat}(\mathcal{W})$ denotes the collection of orthogonal projections $P \in \mathcal{B}(\mathfrak{H})$ which are invariant for $\mathcal{W}$ (i.e., $(I-P)SP = 0$ for every $S \in \mathcal{W}$). Recall that $\mathcal{W}$ is said to be *reflexive* if $r(T) = 0$ implies that $T \in \mathcal{W}$. The algebra $\mathcal{W}$ is said to be *hyper-reflexive* if there exists a constant $c > 0$ such that $d(T) \le cr(T)$ for every $T \in \mathcal{B}(\mathfrak{H})$; the smallest such $c$ is called the *hyper-reflexivity constant* of $\mathcal{W}$. Since $r(T)$ can be calculated using cyclic invariant projections (i.e., projections onto spaces of the form $\overline{\mathcal{W}x}$), it is easy to see that

$$r(T) = \sup\{|\langle Th,k\rangle| : [h \otimes k^*]_{\mathcal{W}} = 0, \|h\|, \|k\| \le 1\}.$$

The number $d(T)$ has a similar description.

**Lemma 3.1.** *If $\mathcal{W}$ is closed in the weak operator topology, then*

$$d(T) = \sup\left\{\left|\sum_{j=1}^{n}\langle Th_j, k_j\rangle\right| : n \geq 1, \sum_{j=1}^{n}\|h_j\|\|k_j\| \leq 1, \sum_{j=1}^{n}[h_j \otimes k_j^*]_{\mathcal{W}} = 0\right\}$$

*for every $T \in \mathcal{B}(\mathfrak{H})$.*

*Proof.* The quotient space $\mathcal{B}(\mathfrak{H})/\mathcal{W}$ can be identified isometrically with the dual of ${}^{\perp}\mathcal{W} = \{\varphi \in \mathcal{B}(\mathfrak{H})_* : \varphi|\mathcal{W} = 0\}$. The assumption that $\mathcal{W}$ is closed in the weak operator topology implies that the finite rank elements in ${}^{\perp}\mathcal{W}$ are dense in ${}^{\perp}\mathcal{W}$. Thus we have

$$d(T) = \|T + \mathcal{W}\| = \sup\{\varphi(T) : T \in {}^{\perp}\mathcal{W}, \varphi \text{ of finite rank}, \|\varphi\| \leq 1\}.$$

The lemma follows because a functional of finite rank $\varphi \in \mathcal{B}(\mathfrak{H})_*$ can be written as $\varphi = \sum_{j=1}^{n} h_j \otimes k_j^*$ with $\sum_{j=1}^{n}\|h_j\|\|k_j\| = \|\varphi\|$. □

When Theorem 2.2 applies to the map $F$, it follows that $\mathcal{W}$ has many invariant subspaces. In fact, we can prove that $\mathcal{W}$ is hyper-reflexive. To simplify statements, let us say that $\mathcal{W}$ has *property* $(\mathbb{X}_\rho)$ if $S_F$ contains the ball of radius $\rho^2$ centered at $0 \in \mathcal{W}^*$. We will also say that $\mathcal{W}$ has property $(\mathbb{A}_1(r))$ if every $\varphi \in \mathcal{W}_*$ with $\|\varphi\| < 1/r$ can be written as $\varphi = [h \otimes k^*]_{\mathcal{W}}$ for some vectors $h, k$ with norm $< 1$. Property $(\mathbb{A}_1(r))$ is only possible for $r \geq 1$. Theorem 2.2 shows that property $(\mathbb{X}_\rho)$ implies property $(\mathbb{A}_1(\rho^{-2}))$.

**Lemma 3.2.** *If $\mathcal{W}$ has property $(\mathbb{A}_1(\rho^{-2}))$ then it is closed in the weak operator topology.*

*Proof.* Assume to the contrary that there exists $T \notin \mathcal{W}$, $\|T\| = 1$, which is in the weak operator closure of $\mathcal{W}$. By the Hahn-Banach theorem, there exists $\varphi \in {}^{\perp}\mathcal{W}$ such that $\varphi(T) = 1$. The functional $\varphi$ can be written as $\sum_{n=1}^{\infty} h_n \otimes k_n^*$ with $\sum_{n=1}^{\infty}\|h_n\|\|k_n\| < \infty$. Choose $N$ so that $\sum_{n=N}^{\infty}\|h_n\|\|k_n\| < \rho^2/4$, and use property $(\mathbb{A}_1(\rho^{-2}))$ to find vectors $h, k$ of norm $< 1/2$ such that $[h \otimes k]_{\mathcal{W}} = \sum_{n=N}^{\infty}[h_n \otimes k_n^*]_{\mathcal{W}}$. In other words, the functionals

$$\varphi_1 = h \otimes k^* + \sum_{n=1}^{N-1} h_n \otimes k_n^* \quad \text{and} \quad \varphi_2 = -h \otimes k^* + \sum_{n=N}^{\infty} h_n \otimes k_n^*$$

belong to ${}^{\perp}\mathcal{W}$. The functional $\varphi_1$ is weak operator continuous so that $\varphi_1(T) = 0$. We conclude that

$$1 = \varphi(T) = \varphi_1(T) + \varphi_2(T) = \varphi_2(T) \leq \|\varphi_2\|,$$

and this is not possible because $\|\varphi_2\| < 1/2$. □

**Theorem 3.3.** *Assume that $\mathcal{W} \subset \mathcal{B}(\mathfrak{H})$ is a weak* closed unital algebra with property $(\mathbb{X}_\rho)$. Then $\mathcal{W}$ is hyper-reflexive with constant $\leq c = 1 + 2\rho^{-2}$.*

*Proof.* Lemma 3.2 implies that $\mathcal{W}$ is closed in the weak operator topology. Fix $T \in \mathcal{B}(\mathfrak{H})$ and vectors $h_j, k_j \in \mathfrak{H}$, $1 \le j \le n$, such that

$$\sum_{j=1}^{n} [h_j \otimes k_j^*]_{\mathcal{W}} = 0.$$

By Lemma 3.1, it will suffice to show that

$$\left| \sum_{j=1}^{n} \langle Th_j, k_j \rangle \right| \le cr(T) \sum_{j=1}^{n} \|h_j\| \|k_j\|. \tag{3.1}$$

We will assume without loss of generality that $\|h_j\| = \|k_j\|$ for all $j$.

Assume for the moment that $[h_i \otimes k_j^*]_{\mathcal{W}} = 0$ and

$$\langle h_i, h_j \rangle = \langle k_i, k_j \rangle = \langle Th_i, k_j \rangle = 0$$

for $i \ne j$. Then the vectors $h = \sum_{j=1}^{n} h_j$, $k = \sum_{j=1}^{n} k_j$ satisfy $[h \otimes k^*] = 0$, $\|h\|\|k\| = \sum_{j=1}^{n} \|h_j\| \|k_j\|$, and

$$\left| \sum_{j=1}^{n} \langle Th_j, k_j \rangle \right| = |\langle Th, k \rangle| \le r(T) \|h\| \|k\|$$

by the definition of $r(T)$. The idea of the proof is to replace the numbers $\langle Th_j, k_j \rangle$ by $\langle Ta_j, b_j \rangle$ in such a way that the new vectors $a_j, b_j$ satisfy (at least approximately) these orthogonality requirements. The details are as follows. Fix $\varepsilon > 0$, and set $\delta = \varepsilon^2 \rho^2 / 2n^2$. Apply first Theorem 2.2 $n$ times to obtain vectors $a_j, b_j \in \mathfrak{H}$, $1 \le j \le n$ such that

(i) $\|a_j\|, \|b_j\| < \rho^{-1} \|h_j\| + \varepsilon$ for $1 \le j \le n$,
(ii) $[a_j \otimes b_j^*]_{\mathcal{W}} = [h_j \otimes k_j^*]_{\mathcal{W}}$ for $1 \le j \le n$,
(iii) $\|[h_i \otimes b_j^*]_{\mathcal{W}}\| + \|[a_i \otimes k_j^*]_{\mathcal{W}}\| < \delta$ for $1 \le i, j \le n$, and
(iv) $|\langle Th_i, b_j \rangle| + |\langle Ta_i, k_j \rangle| < \varepsilon$ for $1 \le i, j \le n$.

By choosing these vectors inductively, we can also require that

(v) $\|a_i \otimes b_j^*\| + |\langle a_i, b_j \rangle| + |\langle Ta_i, b_j \rangle| < \delta$ for $i \ne j$.

Apply next Corollary 2.3 for $N = 2n$ with $h_1, h_2, \ldots, h_n, a_1, a_2, \ldots, a_n$ (resp. $k_1, k_2, \ldots, k_n, b_1, b_2, \ldots, b_n$) in place of $u_i$ (resp. $v_i$), and

$$x_{ij} = \begin{cases} [h_i \otimes k_j^*]_{\mathcal{W}} & \text{for } 1 \le i, j \le n \\ [h_{j-n} \otimes k_{j-n}^*]_{\mathcal{W}} & \text{for } n < j \le 2n \\ 0 & \text{for } i \le n < j \text{ or } j \le n < i \text{ or } n < i \ne j \le 2n. \end{cases}$$

By (iii) and (v),

$$\sum_{i,j=1}^{2n} \|x_{ij} - [u_i \otimes v_j^*]_{\mathcal{W}}\| < 2n^2 \delta = \rho^2 \varepsilon^2$$

and we deduce the existence of vectors $h_j', k_j', a_j', b_j' \in \mathfrak{H}$, $1 \le j \le n$, satisfying

(vi) $\|x_j' - x_j\|, \|y_j' - y_j\|, \|a_j' - a_j\|, \|b_j' - b_j\| < \varepsilon$ for $1 \le j \le n$,

(vii) $[h'_i \otimes h'^*_j]_{\mathcal{W}} = [h_i \otimes k^*_j]_{\mathcal{W}}$ for $1 \le i, j \le n$,
(viii) $[a'_j \otimes b'^*_j]_{\mathcal{W}} = [h_j \otimes k^*_j]_{\mathcal{W}}$ for $1 \le j \le n$,
(ix) $[h'_i \otimes b'^*_j]_{\mathcal{W}} = [a'_i \otimes k'^*_j]_{\mathcal{W}} = 0$ for $1 \le i, j \le n$, and $[a'_i \otimes b'^*_j]_{\mathcal{W}} = 0$ for $i \ne j$.

In order to simplify the formulas, we will write $O(\varepsilon)$ for a quantity which is bounded by $C\varepsilon$, where $C$ depends only on $n, T, h_j, k_j$ and $\rho$. Inequalities (iv), (v) and (vi) imply

(x) $\langle Th'_i, b'_j\rangle = O(\varepsilon)$, $\langle Ta'_i, k_j\rangle = O(\varepsilon)$ for $1 \le i, j \le n$,
(xi) $\langle a'_i, b'_j\rangle = O(\varepsilon)$, $\langle Ta'_j, b'_j\rangle = O(\varepsilon)$ for $i \ne j$, and
(xii) $\langle Th_j, k_j\rangle - \langle Th'_j, k'_j\rangle = O(\varepsilon)$ for $1 \le j \le n$.

Observe now that (vii), (viii) and (ix) imply $[(h'_j - a'_j) \otimes (k'_j + b'_j)^*]_{\mathcal{W}} = 0$, so that

$$|\langle T(h'_j - a'_j), k'_j + b'_j\rangle| \le r(T)\|h'_j - a'_j\|\|k'_j + b'_j\|.$$

Expanding the scalar product in the left-hand side yields

$$|\langle Th'_j, k'_j\rangle - \langle Ta'_j, b'_j\rangle| \le r(T)\|h'_j - a'_j\|\|k'_j + b'_j\| + |\langle Th'_j, b'_j\rangle| + |\langle Ta'_j, k'_j\rangle|,$$

so that

(xiii) $|\langle Th'_j, k'_j\rangle - \langle Ta'_j, b'_j\rangle| \le r(T)\|h'_j - a'_j\|\|k'_j + b'_j\| + O(\varepsilon)$, $1 \le j \le n$,

by (x). To estimate the norms above, use (vi) and (ix) to obtain

$$\begin{aligned}\|h'_j - a'_j\|^2 &= \|h'_j\|^2 + \|a'_j\|^2 + O(\varepsilon)\\ &\le (\|h_j\| + \varepsilon)^2 + (\|a_j\| + \varepsilon)^2 + O(\varepsilon)\\ &\le (1 + \rho^{-2})\|h_j\|^2 + O(\varepsilon),\end{aligned}$$

and analogous estimates for $k'_j + v'_j$ yield

(xiv) $\|h'_j - a'_j\|\|k'_j + v'_j\| \le (1 + \rho^{-2})\|h_j\|^2 + O(\varepsilon)$ for $1 \le j \le n$.

Using (vii) and (ix) we see that the vectors $a' = \sum_{j=1}^n a'_j$, $b' = \sum_{j=1}^n b'_j$ satisfy

$$[a' \otimes b'^*]_{\mathcal{W}} = \sum_{j=1}^n [a'_j \otimes b'^*_j]_{\mathcal{W}} = \sum_{j=1}^n [h_i \otimes k^*_j]_{\mathcal{W}} = 0,$$

and therefore $|\langle Ta', b'\rangle| \le r(T)\|a'\|\|b'\|$. Using (xi) we obtain

(xv) $|\sum_{j=0}^n \langle Ta'_j, b'_j\rangle| \le r(T)\|a'\|\|b'\| + O(\varepsilon)$.

The norms on the right are estimated using (vi) and (xi)

$$\begin{aligned}\|a'\|^2 = \sum_{j=1}^n \|a'_j\|^2 + O(\varepsilon) &= \sum_{j=1}^n \|a_j\|^2 + O(\varepsilon)\\ &\le \rho^{-2}\sum_{j=1}^n \|h_j\|^2 + O(\varepsilon) \le \rho^{-2} + O(\varepsilon),\end{aligned}$$

so that (xv) yields $|\sum_{j=0}^{n}\langle Ta'_j, b'_j\rangle| \le \rho^{-2}r(T) + O(\varepsilon)$. Combining this inequality with (xii) and (xiii) we obtain

$$\left|\sum_{j=1}^{n}\langle Th_j, k_j\rangle\right| \le cr(T)\sum_{j=1}^{n}\|h_j\|\|k_j\| + O(\varepsilon),$$

and (3.1) is obtained by letting $\varepsilon$ tend to zero. □

There are several invariant subspaces appearing implicitly in the above proof: most instances of vectors $u, v$ such that $[u \otimes v]_{\mathcal{W}} = 0$ yield nontrivial subspaces. To obtain invariant subspaces with more specific properties, assume that there exist weak* continuous characters (i.e., algebra homomorphisms) $\varphi : \mathcal{W} \to \mathbb{C}$. Recall [88] that a subspace $\mathfrak{M} \subset \mathfrak{H}$ is *semi-invariant* for $\mathcal{W}$ if the map $\Phi : \mathcal{W} \to \mathcal{B}(\mathfrak{M})$ defined by compression

$$\Phi(T) = P_{\mathfrak{M}}T|\mathfrak{M}$$

is multiplicative. Alternatively, $\mathfrak{M} = \mathfrak{M}' \ominus \mathfrak{M}''$, where $\mathfrak{M}'' \subset \mathfrak{M}'$ are invariant subspaces for $\mathcal{W}$.

**Proposition 3.4.** *Under the assumptions of Theorem* 3.3, *let* $(\varphi_j)_{j=1}^{\infty} \subset \mathcal{W}_*$ *be a sequence of characters. There exist a semi-invariant subspace* $\mathfrak{M}$ *for* $\mathcal{W}$ *and an orthonormal basis* $(u_j)_{j=1}^{\infty}$ *in* $\mathfrak{M}$ *such that*

$$P_{\mathfrak{M}}Tu_j = \varphi_j(T)u_j, \quad j \ge 1.$$

*Proof.* Proposition 2.4 implies the existence of vectors $u_{jn}, v_{jn} \in \mathfrak{H}$ such that

(a) $[u_{jn} \otimes v_{jn}^*]_{\mathcal{W}} = \varphi_j$ for $j, n \ge 1$, and
(b) $[u_{jn} \otimes v_{j'n'}^*]_{\mathcal{W}} = 0$ if $j \ne j'$ or $n \ne n'$.

Denote by $\mathfrak{M}'$ the closed linear span of the spaces $\mathcal{W}u_{jn}$, $j, n \ge 1$, let $\mathfrak{M}'' \subset \mathfrak{M}'$ be the closed linear span of the spaces $(\ker \varphi_j)u_{jn}$, $j, n \ge 1$, and set $\mathfrak{M}_0 = \mathfrak{M}' \ominus \mathfrak{M}''$. With the notation $w_{jn} = P_{\mathfrak{M}_0}u_{jn}$, we have

$$P_{\mathfrak{M}_0}(T - \varphi_j(T))w_{jn} = 0, \quad j, n \ge 1,$$

and the vectors $\{w_{jn} : j, n \ge 1\}$ are linearly independent. Indeed, relations (a) and (b) show that the vectors $v_{jn}$ are orthogonal to $\mathfrak{M}''$, and therefore

$$\langle w_{jn}, v_{j'n'}\rangle = \langle u_{jn}, v_{j'n'}\rangle = [u_{nj} \otimes v_{nj}^*]_{\mathcal{W}}(I) = \delta_{n,n'}\delta_{j,j'}.$$

To conclude the proof, choose now an orthonormal sequence $u_j$ such that $u_j$ is in the linear span of $\{w_{jn} : n \ge 1\}$, and denote by $\mathfrak{M}$ the subspace generated by $\{u_j : j \ge 1\}$. The proposition follows because $\mathfrak{M}$ is invariant for all the operators $P_{\mathfrak{M}_0}T|\mathfrak{M}_0$, $T \in \mathcal{W}$, and hence it is semi-invariant for $\mathcal{W}$. □

## 4. Dominating spectrum

It is time to show that the results of the preceding sections apply in concrete situations. We start with examples in which the dual algebra $\mathcal{W} \subset \mathcal{B}(\mathfrak{H})$ is commutative. Thus, assume that $T \in \mathcal{B}(\mathfrak{H})$ is a contraction, i.e., $\|T\| \leq 1$. It is well known [92] that there exists an orthogonal decomposition $T = T_0 \oplus T_1$ such that $T_1$ is unitary and $T_0$ has no unitary direct summand. When searching for invariant subspaces it makes good sense to restrict ourselves to the case $T = T_0$, in which case $T$ is said to be *completely non-unitary.* In this case, Sz.-Nagy has shown that there exists a unitary operator $U$ on a space $\mathfrak{K} \supset \mathfrak{H}$ such that $\mathfrak{H}$ is semi-invariant for $U$, $T = P_{\mathfrak{H}}U|\mathfrak{H}$, and the spectral measure of $U$ is absolutely continuous relative to arclength measure $m$ on the unit circle $\mathbb{T} \subset \mathbb{C}$. Denote by $H^\infty$ the Banach algebra of bounded analytic functions on the unit disk $\mathbb{D}$. The elements $f \in H^\infty$ can be regarded alternatively as measurable functions in $L^\infty(\mathbb{T})$, and therefore the spectral functional calculus $f(U)$ makes sense for such functions. This leads naturally to the definition of the *Sz.-Nagy–Foias functional calculus* given by

$$f(T) = P_{\mathfrak{H}} f(U)|\mathfrak{H}, \quad f \in H^\infty.$$

When regarded as a subspace of $L^\infty$, the algebra $H^\infty$ is closed in the weak* topology given by duality with $L^1(\mathbb{T})$. The map $f \to f(T)$ is then a contractive algebra homomorphisms, and it is also continuous when $H^\infty$ and $\mathcal{B}(\mathfrak{H})$ are given their weak* topologies. In particular, given $h, k \in \mathfrak{H}$ the functional $h \otimes_T k^*$ defined by

$$(h \otimes_T k^*)(f) = \langle f(T)h, k\rangle, \quad f \in H^\infty,$$

belongs to the predual $H^\infty_*$. Given $\lambda \in \mathbb{D}$, the evaluation functional $e_\lambda : f \to f(\lambda)$ belongs to $H^\infty_*$.

**Lemma 4.1.** *Assume that $T$ is a completely nonunitary contraction, $\lambda \in \mathbb{D}$, and $(h_n)_{n=1}^\infty$ is an orthonormal sequence in $\mathfrak{H}$ such that $\lim_{n\to\infty} \|(T - \lambda)h_n\| = 0$ or $\lim_{n\to\infty} \|(T^* - \bar{\lambda})h_n\| = 0$. Then*

$$\lim_{n\to\infty} \|e_\lambda - h_n \otimes_T h_n^*\| = 0$$

*and*

$$\lim_{n\to\infty} (\|h_n \otimes_T k^*\| + \|k \otimes_T h_n^*\|) = 0$$

*for every $k \in \mathfrak{H}$.*

*Proof.* It suffices to treat the case of an approximate eigenvalue for $T$. Replacing $T$ by $(T - \lambda)(I - \lambda T^*)^{-1}$ we may assume that $\lambda = 0$. The Hahn-Banach theorem implies the existence of functions $f_n \in H^\infty$ with $\|f_n\|_\infty = 1$ such that

$$\|e_0 - h_n \otimes_T h_n^*\| = f_n(0) - \langle f_n(T)h_n, h_n\rangle = \langle (f_n(0) - f_n(T))h_n, h_n\rangle.$$

Writing $f_n(0) - f_n(z) = zg_n(z)$ with $g_n \in H^\infty$, we have $\|g_n\| \leq 2$, and thus

$$\|e_0 - h_n \otimes_T h_n^*\| = \langle g_n(T)Th_n, h_n\rangle \leq 2\|Th_n\| \to 0$$

as $n \to \infty$. Similarly, $\|h_n \otimes_T k^*\| = \langle u_n(T)h_n, k\rangle$ for some $u_n \in H^\infty$, $\|u_n\| = 1$. Writing again $u_n(z) = u_n(0) + zv_n(z)$, we obtain

$$\|h_n \otimes_T k^*\| = u_n(0)\langle h_n, k\rangle + \langle v_n(T)Th_n, h_n\rangle \leq |\langle h_n, k\rangle| + 2\|Th_n\|,$$

and the right-hand side converges to zero because the $h_n$ are orthonormal. The study of $\|k \otimes_T h_n^*\|$ requires a closer look at the unitary dilation $U \in \mathcal{B}(\mathfrak{K})$. Denote by $\mathfrak{K}_-$ the closed linear span of $U^{*n}\mathfrak{H}$, and set $V = U^*|\mathfrak{K}_-$. The space $\mathfrak{H}$ is invariant for $V^*$, and $V^*|\mathfrak{H} = T$. Consider the von Neumann–Wold decomposition of the isometry $V$ as $V = S \oplus W$ on $\mathfrak{K}_- = \mathfrak{A} \oplus \mathfrak{B}$, with $S$ a shift and $W$ unitary. Writing $h_n = a_n \oplus b_n$ with $a_n \in \mathfrak{A}, b_n \in \mathfrak{B}$, we have

$$\|b_n\| = \|W^*b_n\| \leq \|(S \oplus W)^*(h_n)\| = \|Th_n\| \to 0$$

as $n \to \infty$, and similarly $\lim_{n\to\infty} \|S^*a_n\| = 0$. Write now $k = \alpha \oplus \beta$ with $\alpha \in \mathfrak{A}, \beta \in \mathfrak{B}$, and choose functions $u_n \in H^\infty$ with unit norm such that

$$\|k \otimes_T h_n^*\| = \langle u_n(T)h_n, k\rangle = \langle u_n(S^*)a_n, \alpha\rangle + \langle u_n(W^*)b_n, \beta\rangle.$$

The last term on the right-hand side clearly tends to zero as $n \to \infty$. It remains to show that $\langle u_n(S^*)a_n, \alpha\rangle \to 0$ as well. If we write $a_n = a_n' + a_n''$ with $a_n' = (I - SS^*)a_n$, we have $\|a_n''\| \to 0$ while $a_n'$ tends to zero weakly. Thus

$$\langle u_n(S^*)a_n, \alpha\rangle = \langle u_n(S^*)a_n', \alpha\rangle + \langle u_n(S^*)a_n'', \alpha\rangle = u_n(0)\langle a_n', \alpha\rangle + \langle u_n(S^*)a_n'', \alpha\rangle$$

tends to zero as $n \to \infty$. □

The preceding lemma shows that $e_\lambda$ belongs to the set $S_F$ for the map $F : (h, k^*) \mapsto h \otimes_T k^*$ provided that $\lambda$ is in the essential spectrum $\sigma_e(T)$ of $T$.

A subset $\Lambda \subset \mathbb{D}$ is said to be *dominating* if

$$\sup\{|f(\lambda)| : \lambda \in \Lambda\} = \|f\|_\infty$$

for every $f \in H^\infty$. An annulus with outer circle $\mathbb{T}$ is dominating, and it is easy to construct dominating sets which have no accumulation points in $\mathbb{D}$. A geometric characterization of dominating sets is given in [33]: $\Lambda$ is dominating if and only if almost every $\zeta \in \mathbb{T}$ is a nontangential limit point of $\Lambda$. The Hahn-Banach theorem implies that $\Lambda$ is dominating if and only if the closed, convex balanced hull of $\{e_\lambda : \lambda \in \Lambda\}$ is the unit ball of $H_*^\infty$.

**Theorem 4.2.** *Let $T \in \mathcal{H}(\mathfrak{H})$ be a completely nonunitary contraction such that $\sigma_e(T) \cap \mathbb{D}$ is dominating.*

1. *The functional calculus $f \mapsto f(T)$ is an isometry of $H^\infty$ onto the weak operator closed algebra $\mathcal{W}_T$ generated by $T$.*
2. *The algebra $\mathcal{W}_T$ has property $(\mathbb{X}_1)$. In particular, $\mathcal{W}_T$ is hyper-reflexive with constant $\leq 3$.*

*Proof.* For each $\lambda \in \sigma_e(T)$, there exists an orthonormal sequence $(x_n)_{n=1}^\infty$ such that $\lim_{n\to\infty} \|e_\lambda - x_n \otimes_T x_n^*\| = 0$. For every function $f \in H^\infty$ we have therefore

$$|f(\lambda)| = |e_\lambda(f)| = \lim_{n\to\infty} |\langle f(T)x_n, x_n\rangle| \leq \|f(T)\|.$$

Thus

$$\|f\|_\infty = \sup\{|f(\lambda)| : \lambda \in \sigma_e(T)\} \le \|f(T)\| \le \|f\|_\infty,$$

so that indeed $f \mapsto f(T)$ is isometric. Let us set for the moment $\mathcal{A} = \{u(T) : u \in H^\infty\}$. Since the functional calculus is continuous relative to the weak* topologies, it follows that

$$\mathcal{A} \cap \{X \in \mathcal{B}(\mathfrak{H}) : \|X\| \le 1\} = \{f(T) : \|f\|_\infty \le 1\}$$

is weak* compact. We conclude from the Kreĭn–Smul′jan theorem that $\mathcal{A}$ is weak* closed, and that the map $f \mapsto f(T)$ is a weak* homeomorphism from $H^\infty$ to $\mathcal{A}$. The predual of this homeomorphism carries the functional $[h \otimes k^*]_{\mathcal{A}}$ to $h \otimes_T k^*$, and thus Lemma 4.1 shows that $\mathcal{A}$ has property $(\mathbb{X}_1)$. In particular, $\mathcal{A}$ has property $(\mathbb{A}_1(1))$, and hence it is weak operator closed by Lemma 3.2. Thus we have in fact $\mathcal{A} = \mathcal{W}_T$, and the theorem follows. □

**Corollary 4.3.** *Assume that $T \in \mathcal{B}(\mathfrak{H})$ is a contraction such that $\sigma(T) \cap \mathbb{D}$ is dominating. Then $T$ has nontrivial invariant subspaces.*

*Proof.* If $T$ is not completely nonunitary, its unitary part provides an invariant subspace. If there is $\lambda \in \sigma(T) \setminus \sigma_e(T)$ then the kernel or the range of $\lambda - T$ is a nontrivial invariant subspace. Thus, we may assume that $T$ is completely nonunitary and $\sigma(T) = \sigma_e(T)$, in which case the preceding theorem applies. □

The existence of invariant subspaces for subnormal operators follows from this corollary. Indeed, the results of [89] allow one to reduce to the case of a subnormal contraction $T$ whose spectrum is $\overline{\mathbb{D}}$.

It is possible to prove property $(\mathbb{X}_1)$ for larger classes of contractions. Recall that a contraction $T \in \mathcal{B}(\mathfrak{H})$ is said to be of class $C_{0\cdot}$ if

$$\lim_{n\to\infty} \|T^n h\| = 0$$

for every $h \in \mathfrak{H}$. If $T^*$ is of class $C_{0\cdot}$ then $T$ is said to be of class $C_{\cdot 0}$, and $T \in C_{00}$ if $T$ is both of class $C_{0\cdot}$ and of class $C_{\cdot 0}$.

**Lemma 4.4.** *Assume that $T \in \mathcal{B}(\mathfrak{H})$ is a contraction of class $C_{0\cdot}$ (resp. $C_{\cdot 0}$), and $(x_n)_{n=1}^\infty \subset \mathfrak{H}$ converges weakly to zero. Then $\lim_{n\to\infty} \|y \otimes_T x_n\| = 0$ (resp., $\lim_{n\to\infty} \|x_n \otimes_T y\| = 0$) for every $y \in \mathfrak{H}$.*

*Proof.* By symmetry, it suffices to consider the sequence $x_n \otimes_T y$. Since the powers of $T^*$ tend to zero, it follows [92] that there exists a unilateral shift $S$ on a Hilbert space $\mathfrak{K} \supset \mathfrak{H}$ such that $S^*\mathfrak{H} \subset \mathfrak{H}$ and $S^*|\mathfrak{H} = T^*$. We have $x \otimes_T y^* = x \otimes_S y^*$ for $x, y \in \mathfrak{H}$, so it will suffice to show that $\lim_{n\to\infty} x_n \otimes_S y^* = 0$ for every $y \in \mathfrak{K}$. Moreover, since the sequence $x_n$ is necessarily bounded, it suffices to prove this for a dense sequence of vectors $y \in \mathfrak{K}$. The linear space $\bigcup_{k=1}^\infty \ker S^{*k}$ is dense in $\mathfrak{K}$, so we may assume that $y \in \ker S^{*k}$ for some $k \ge 1$. Choose functions $f_n \in H^\infty$ of unit norm such that $\|x_n \otimes_T y\| = \langle f_n(S)x_n, y\rangle$. For fixed $k$, we can write

$$f_n(z) = a_n(0) + a_n(1)z + \cdots + a_n(k-1)z^{k-1} + z^k g_n(z),$$

with $g_n \in H^\infty$. Therefore

$$\|x_n \otimes_S y\| = \langle f_n(S)x_n, y\rangle = \sum_{j=0}^{k-1} a_n(j)\langle x_n, S^{*j}y\rangle,$$

and the desired conclusion follows by letting $n \to \infty$ in this formula. □

We will apply this result to weighted shifts. Given a bounded sequence $w = (w_n)_{n=0}^\infty$ of numbers in $(0,1)$, and a Hilbert space $\mathfrak{H}$ with an orthonormal basis $(e_n)_{n=0}^\infty$, the weighted shift $S_w$ is determined by its action on the basis: $We_n = w_n e_{n+1}$ for $n \geq 0$. Clearly $S_w$ is a contraction, and it is of class $C_{00}$ if and only if $\sum_{n=0}^\infty (1 - w_n) = \infty$.

**Proposition 4.5.** *Let $T = S_w$ be a weighted shift of class $C_{00}$ such that $w_n \leq w_{n+1} \to 1$ as $n \to \infty$. Then the conclusions of Theorem 4.2 hold for $T$.*

*Proof.* Given $\lambda \in \mathbb{D}$, it is easy to verify that the spaces $\mathfrak{H}_n = (\lambda - T)^n\mathfrak{H}$ are closed and strictly decreasing. Choose unit vectors $x_n \in \mathfrak{H}_n \ominus \mathfrak{H}_{n+1}$, and note that $x_n \otimes_T x_n = e_\lambda$ and $x_n$ converges weakly to zero. We can now proceed as in the proof of Theorem 4.2 by virtue of Lemma 4.4. □

Among the operators covered by this proposition is the Bergman shift with weights $w_n = (n+1)^{1/2}/(n+2)^{1/2}$, $n \geq 0$.

Proposition 3.4 applies whenever the conclusions of Theorem 4.2 hold. We deduce that operators $T$ covered by the above results have semi-invariant subspaces $\mathfrak{M}$ such that the compression $P_{\mathfrak{M}}T|\mathfrak{M}$ is a diagonal operator with arbitrarily given eigenvalues in $\mathbb{D}$. When all of these eigenvalues are zero and $\mathfrak{M} = \mathfrak{M}' \ominus \mathfrak{M}''$, with $\mathfrak{M}', \mathfrak{M}''$ invariant, the space $\mathfrak{M}' \ominus \overline{T\mathfrak{M}'} \supset \mathfrak{M}$ is infinite-dimensional, and any space $\mathfrak{N}$ satisfying $\mathfrak{M}'' \subset \mathfrak{N} \subset \mathfrak{M}'$ is invariant for $T$. This reveals a shockingly large invariant subspace lattice, especially in the case of weighted shifts.

## 5. A noncommutative example

In this section we will assume that $\mathcal{W} \subset \mathcal{B}(\mathfrak{H})$ is a unital, weak operator closed algebra with the property that its commutant

$$\mathcal{W}' = \{X \in \mathcal{B}(\mathfrak{H}) : TX = XT \text{ for all } T \in \mathcal{W}\}$$

contains two isometric operators $U_1, U_2$ with orthogonal ranges, i.e., $U_1^* U_2 = 0$. This situation arises, for instance, when $\mathcal{W}$ is the algebra generated by all left creation operators on the full Fock space associated with a Hilbert space of dimension at least 2.

**Lemma 5.1.** *Given vectors $h_1, h_2, \ldots, h_p \in \mathfrak{H}$ and $\varepsilon > 0$, there exists an isometry $W \in \mathcal{W}'$ such that $W\mathfrak{H} \subset U_1\mathfrak{H}$ and $\|W^* h_j\| < \varepsilon$ for $j = 1, 2, \ldots, p$.*

*Proof.* We construct inductively isometries $W_n$ such that $W_1 = U_1$, $W_{n+1} \in \{W_nU_1, W_nU_2\}$, and $\sum_{j=1}^{p} \|W_{n+1}^* h_j\|^2 \leq (1/2)\sum_{j=1}^{p} \|W_n^* h_j\|^2$. This is possible because $W_nU_1$ and $W_nU_2$ have orthogonal ranges, and therefore

$$\sum_{j=1}^{p} \|(W_nU_1)^* h_j\|^2 + \sum_{j=1}^{p} \|(W_nU_2)^* h_j\|^2 \leq \sum_{j=1}^{p} \|W_n^* h_j\|^2.$$

The lemma is satisfied by $W = W_n$ if $n$ is sufficiently large. □

We recall an inequality which comes from the usual estimate of the $L^1$ norm of the Dirichlet kernel.

**Lemma 5.2.** *There exist constants $c, d > 0$ with the following property. For every function $u(z) = \sum_{\ell=0}^{\infty} u_n z^n$ in $H^\infty$, $|\sum_{\ell=1}^{n} u_n| \leq c \log n + d$ for all $n \geq 1$.*

*Proof.* We have $\sum_{\ell=1}^{n} u_n = (1/2\pi)\int_0^{2\pi} u(e^{it})v_n(t)\,dt$, where $v_n(t) = \sum_{\ell=1}^{n} \cos(\ell t)$. The lemma follows because $\|v\|_1 \leq c \log n + d$ (cf. Section II.12 in [102]). □

**Proposition 5.3.** *Fix vectors $x, y, h_1, h_2, \ldots, h_p \in \mathfrak{H}$ and $\varepsilon > 0$. There exist vectors $x', y' \in \mathfrak{H}$ with the following properties:*

1. $\|x'\| = \|x\|$, $\|y'\| = \|y\|$,
2. $[x' \otimes y'^*]_{\mathcal{W}} = [x \otimes y^*]_{\mathcal{W}}$, *and*
3. $\|[h_j \otimes y'^*]_{\mathcal{W}}\| + \|[x' \otimes h_j^*]_{\mathcal{W}}\| < \varepsilon$ *for* $j = 1, 2, \ldots, p$.

*Proof.* We may assume without loss of generality that $\|x\| = \|y\| = 1$. Lemma 5.1 allows us to choose isometries $W, V \in \mathcal{W}'$ such that $W^*V = 0$ and $\|W^*h_j\| < \varepsilon/3$ for $j = 1, 2, \ldots, p$. If we denote $u_k = W^kVx, v_k = W^kVy$, it is easy to verify that $[u_k \otimes v_\ell^*]_{\mathcal{W}} = \delta_{k\ell}[x \otimes y^*]_{\mathcal{W}}$ and $\langle u_k, u_\ell\rangle = \delta_{k\ell}\|x\|^2$. For instance, if $k < \ell$ and $T \in \mathcal{W}$,

$$\langle Tu_k, v_\ell\rangle = \langle W^kVTx, W^\ell Vy\rangle = \langle W^{k-\ell-1}Tx, W^*Vy\rangle = 0.$$

Thus the vectors $x_n = n^{-1/2}\sum_{k=1}^{n} W^kVx$ and $y_n = n^{-1/2}\sum_{k=1}^{n} W^kVy$ satisfy $\|x_n\| = \|x\|, \|y_n\| = \|y\|$, and $[x_n \otimes y_n^*]_{\mathcal{W}} = [x \otimes y^*]_{\mathcal{W}}$. It will therefore suffice to show that condition (iii) is satisfied by $x' = x_n$ and $y' = y_n$ provided that $n$ is sufficiently large. Since $x_n \in W\mathfrak{H}$, we also have $Tx_n \in W\mathfrak{H}$ for $T \in \mathcal{W}$, and therefore

$$\begin{aligned}|[x_n \otimes h_j^*]_{\mathcal{W}}(T)| &= \langle Tx_n, h_j\rangle = \langle W^*WTx_n, h_j\rangle \\ &= \langle W^*Tx_n, W^*h_j\rangle \leq \|T\|\|x_n\|\|W^*h_j\|.\end{aligned}$$

We conclude that $\|[x_n \otimes h_j^*]_{\mathcal{W}}\| < \varepsilon/3$ by the choice of $W$. In order to estimate $[h_j \otimes y_n^*]_{\mathcal{W}}$, we write $h_j = a_j + b_j$ with $a_j \in \ker W^*$ and $\|b_j\| < \varepsilon/3$, so that $\|[h_j \otimes y_n^*]_{\mathcal{W}}\| \leq \|[a_j \otimes y_n^*]_{\mathcal{W}}\| + \varepsilon/3$ for $j = 1, 2, \ldots, p$. To conclude the proof, it suffices to prove that $\lim_{n\to\infty} \|[a \otimes y_n^*]_{\mathcal{W}}\| = 0$ for every $a \in \ker W^*$. To do this, consider an operator $T \in \mathcal{W}$, and consider the analytic function $f(\lambda) = \sum_{k=0}^{\infty}\langle Ta, v_k\rangle$, $\lambda \in \mathbb{D}$. We claim that $\|f\|_\infty \le \|T\|\|a\|\|y\|$. Indeed, we have

$$\begin{aligned}\frac{f(\lambda)}{1-|\lambda|^2} &= \left(\sum_{k=0}^{\infty}\bar{\lambda}^k\lambda^k\right)\left(\sum_{m=0}^{\infty}\lambda^m\langle Ta, W^mVy\rangle\right)\\ &= \sum_{k,m=1}^{\infty}\bar{\lambda}^k\lambda^{k+m}\langle W^kTa, W^{m+k}Vy\rangle\\ &= \sum_{k,\ell=1}^{\infty}\bar{\lambda}^k\lambda^\ell\langle W^kTa, W^\ell Vy\rangle = \left\langle T\sum_{k=1}^{\infty}\bar{\lambda}^kW^ka, \sum_{\ell=1}^{\infty}\bar{\lambda}^\ell W^\ell Vy\right\rangle,\end{aligned}$$

where we have used the fact that $\langle W^kTa, W^\ell Vy\rangle = \langle W^{k-\ell}Ta, Vy\rangle = 0$ for $k > \ell$. Since the vectors $(W^ka)_{k=0}^{\infty}$ are mutually orthogonal, we have $\|\sum_{k=1}^{\infty}\bar{\lambda}^kW^ka\| = (1-|\lambda|^2)^{-1/2}\|a\|$, and similarly $\|\sum_{\ell=1}^{\infty}\bar{\lambda}^\ell W^\ell Vy\| = (1-|\lambda|^2)^{-1/2}\|y\|$. The preceding identity implies the desired estimate $|f(\lambda)| \le \|T\|\|A\|\|y\|$. We can now estimate

$$\begin{aligned}|[a\otimes y_n^*]_{\mathcal{W}}(T)| = |\langle Ta, y_n\rangle| &= \frac{1}{n^{1/2}}\left|\sum_{k=1}^{n}\langle Ta, u_k\rangle\right|\\ &\le \frac{1}{n^{1/2}}(c\log n + d)\|T\|\|a\|\|y\|,\end{aligned}$$

and this implies that $\|[a\otimes y_n^*]_{\mathcal{W}}\| \le n^{-1/2}(c\log n + d)\|a\|\|y\| \to 0$ as $n\to\infty$. □

The following result follows immediately.

**Corollary 5.4.** *Let $\mathcal{W}$ be a unital, weak operator closed subalgebra of $\mathcal{B}(\mathfrak{H})$ such that $\mathcal{W}$ commutes with two isometries with orthogonal ranges. Then $\mathcal{W}$ has property $(\mathbb{X}_1)$, and therefore $\mathcal{W}$ is hyper-reflexive with constant $\le 3$.*

## 6. Approximate factorization

There is a second factorization technique which is appropriate for some situations in which strong vanishing results such as Lemma 4.4 are not available. The main ingredient has again an abstract version which we now present.

Consider a separable Hilbert space $\mathfrak{D}$ and a separable, diffuse probability space $(\Omega, \mathcal{F}, \mu)$, i.e., $L^2(\mu)$ is separable and $\mu$ has no atoms. We denote by $L^2(\mu, \mathfrak{D})$ the space of measurable, square integrable functions $f : \Omega \to \mathfrak{D}$. Given two functions $f, g \in L^2(\mu, \mathfrak{D})$, we define the function $f\cdot g^* \in L^1(\mu)$ by setting $(f\cdot g^*)(\omega) = \langle f(\omega), g(\omega)\rangle$ for $\omega\in\Omega$, where the scalar product is calculated in $\mathfrak{D}$; note that

$$\langle f, g\rangle = \int_\Omega (f\cdot g^*)(\omega)\,d\mu(\omega), \quad f, g \in L^2(\mu, \mathfrak{D}).$$

A subspace $\mathfrak{H} \subset L^2(\mu, \mathfrak{D})$ will be said to be *localizable* if for every $\sigma \in \mathcal{F}$ with $\mu(\sigma) > 0$, and for every $\varepsilon > 0$, there exists $f \in \mathfrak{H}$ such that $\|\chi_{\Omega\setminus\sigma} f\| < \varepsilon \|\chi_\sigma f\|$. Here $\chi_\sigma$ denotes, as usual, the characteristic function of $\sigma$.

**Lemma 6.1.** *Assume that $\mathfrak{H} \subset L^2(\mu, \mathfrak{D})$ is a localizable space, $\sigma \in \mathcal{F}$ has positive measure, and $h_1, h_2, \dots, h_p \in L^2(\mu, \mathfrak{D})$. For every $\varepsilon > 0$ there exists $f \in \mathfrak{H}$ such that $\|\chi_{\Omega\setminus\sigma} f\| < \varepsilon \|\chi_\sigma f\|$ and $\langle f, h_j \rangle = 0$ for $j = 1, 2, \dots, p$.*

*Proof.* We may assume that the vectors $h_j$ belong to $\mathfrak{H}$ and are orthonormal. Consider a number $\eta > 0$ and subsets $\sigma_{n+1} \subset \sigma_n \subset \sigma$ such that $\lim_{n\to\infty} \mu(\sigma_n) = 0$. By definition, there exist vectors $f_n \in \mathfrak{H}$ such that $\|\chi_{\Omega\setminus\sigma_n} f_n\| < \eta \|\chi_{\sigma_n} f_n\|$. Replacing $f_n$ by $f_n/\|f_n\|$, we may assume that $\|f_n\| = 1$. Observe that for any $h \in L^2(\mu, \mathfrak{D})$ we have

$$|\langle f_n, h\rangle| \le |\langle \chi_{\Omega\setminus\sigma_n} f_n, h\rangle| + |\langle f_n, \chi_{\sigma_n} h\rangle| \le \eta\|h\| + \|\chi_{\sigma_n} h\|,$$

so that we will have $|\langle f_n, h_j\rangle| \le 2\eta$ for large $n$. For such a value of $n$, define $f = f_n - \sum_{j=1}^{p} \langle f_n, h_j\rangle h_j$, so that $\|f - f_n\| \le 2p^{1/2}\eta$. Thus

$$\|\chi_{\Omega\setminus\sigma} f\| \le \|\chi_{\Omega\setminus\sigma_n} f_n\| + 2p^{1/2}\eta \le (1 + 2p^{1/2})\eta,$$

and hence $\|\chi_\sigma f\| \ge 1 - (1 + 2p^{1/2})\eta$. If we choose $\eta$ so that

$$\frac{(1 + 2p^{1/2})\eta}{1 - (1 + 2p^{1/2})\eta} < \varepsilon,$$

the function $f$ satisfies the requirements of the lemma. □

We will use a slight variation of Lemma 6.1.

**Lemma 6.2.** *Assume that $\mathfrak{H} \subset L^2(\mu, \mathfrak{D})$ is localizable, $\mathfrak{H}' \subset \mathfrak{H}$ is a dense linear manifold, and $\varepsilon > 0$. Given a function $f \in L^\infty(\mu)$ such that $0 \le f \le 1$ and $\|f\|_\infty > 1 - \varepsilon$, there exists $x \in \mathfrak{H}'$ such that $|\langle x, h_j\rangle| = 0$ for $j = 1, 2, \dots, p$, and*

$$\|(1 - f)^{1/2} x\|^2 < \frac{\varepsilon}{1 - \varepsilon} \|f^{1/2} x\|^2.$$

*Proof.* It suffices to consider the case $\mathfrak{H}' = \mathfrak{H}$ since the linear manifold $\{x \in \mathfrak{H}' : \langle x, h_j\rangle = 0, 1 \le j \le p\}$ is also dense in $\{x \in \mathfrak{H} : \langle x, h_j\rangle = 0, 1 \le j \le p\}$. The set $\sigma = \{\omega : f(\omega) > 1 - \alpha\}$ has positive measure for some $\alpha < \varepsilon$. Fix $\beta > 0$ such that

$$\frac{\alpha + \beta^2}{1 - \alpha} < \frac{\varepsilon}{1 - \varepsilon},$$

and choose $x \in \mathfrak{H}$ such that $\|\chi_{\Omega\setminus\sigma} x\| < \beta \|\chi_\sigma x\|$ and $\langle x, h_j\rangle = 0$ for all $j$. Observe that

$$\|f^{1/2} x\|^2 \ge \int_\sigma f(\omega) \|x(\omega)\|^2 \, d\mu(\omega) \ge (1 - \alpha)\|\chi_\sigma x\|^2,$$

and therefore

$$\begin{aligned}\|(1-f)^{1/2}x\|^2 &= \int_\sigma (1-f(\omega))\|x(\omega)\|^2\,d\mu(\omega) + \int_{\Omega\setminus\sigma} (1-f(\omega))\|x(\omega)\|^2\,d\mu(\omega)\\ &\le (\alpha+\beta^2)\|\chi_\sigma x\|^2 \le \frac{\alpha+\beta^2}{1-\alpha}\|f^{1/2}x\|^2 < \frac{\varepsilon}{1-\varepsilon}\|f^{1/2}x\|^2,\end{aligned}$$

thus yielding the desired inequality. □

The main result of this section is as follows.

**Theorem 6.3.** *Let $\mathfrak{H} \subset L^2(\mu,\mathfrak{D})$ be a localizable space. Given $\varepsilon > 0$, a function $f \in L^1(\mu)$, and functions $h_1, h_2, \dots, h_p \in L^2(\mu,\mathfrak{D})$, there exist $x, y \in \mathfrak{H}$ such that*

(1) $\|f - x\cdot y^*\|_1 < \varepsilon$,
(2) $\|x\|, \|y\| \le \|f\|_1^{1/2}$,
(3) $\langle x, h_j\rangle = \langle y, h_j\rangle = 0$ *for* $j = 1, 2, \dots, p$, *and*
(4) $x = y$ *if* $f \ge 0$.

Assume that $\nu$ is another probability measure on $\Omega$, mutually absolutely continuous relative to $\mu$. In other words, $d\nu = \rho\,d\mu$ with $\rho \in L^1(\mu)$ a strictly positive function of norm one. One can then replace the space $\mathfrak{H}$ with

$$\mathfrak{H}' = \{\rho^{-1/2}h : h \in \mathfrak{H}\} \subset L^2(\nu).$$

It is easy to see that the conclusion of Theorem 6.3 is equivalent to the corresponding statement for the space $\mathfrak{H}'$. This allows us to make an additional assumption about $\mathfrak{H}$. Namely, choose an orthonormal basis $(e_n)_{n=1}^\infty$ in $\mathfrak{H}$, and define

$$\rho = \frac{1}{2} + \sum_{n=1}^\infty \frac{1}{2^{n+1}} e_n \cdot e_n^*.$$

Then the functions $\rho^{-1/2}e_n$ are bounded. For the remainder of this section, $\mathfrak{H}$ is assumed to be a localizable space where the bounded functions are dense, and $h_1, h_2, \dots, h_p \in L^2(\mu,\mathfrak{D})$.

The essential point in the proof is the approximation of functions of the form $f = \chi_\sigma$ by products of the form $x\cdot x^*$, with $x \in \mathfrak{H}$. This is achieved when $\|x(\omega)\|$ is close to $\chi_\sigma(\omega)$ on a set with large measure. Given $\sigma \in \mathcal{F}$, finitely many vectors $h_1, h_2, \dots, h_p \in L^2(\mu,\mathfrak{D})$, and $\eta, \delta \in (0,1)$, we denote by $\mathfrak{S}(\sigma; h_1, h_2, \dots, h_p; \eta; \delta)$ the collection of those functions $x \in \mathfrak{H}$ which can be written as $x = g + b$, where $g, b \in L^2(\mu,\mathfrak{D})$ such that:

(i) $\|g(\omega)\| \le \chi_\sigma(\omega)$ $\mu$-almost everywhere,
(ii) $\|b\| \le \eta\|g\|$,
(iii) $|\langle g, h_j\rangle| < \delta\|x\|$ for
(iv) $\langle x, h_j\rangle = 0$ for $j = 1, 2, \dots, p$. $j = 1, 2, \dots, p$, and

We can now show that the set $\mathfrak{S}(\sigma; h_1, h_2, \dots, h_p; \eta; \delta)$ has elements of fairly large norm.

**Proposition 6.4.** *We have* $\sup\{\|x\| : x \in \mathfrak{S}(\sigma; h_1, h_2, \dots, h_p; \eta; \delta)\} > 2^{-3}\eta\mu(\sigma)^{1/2}$ *provided that* $\eta < 1/2$.

*Proof.* We set $\mathfrak{S} = \mathfrak{S}(\sigma; h_1, h_2, \dots, h_p; \eta; \delta)$, $\gamma\mu(\sigma)^{1/2} = \sup\{\|x\| : x \in \mathfrak{S}\}$, and assume to the contrary that $\gamma \le 2^{-3}\eta$. There exist elements $x_n = g_n + b_n \in \mathfrak{S}$ such that $\lim_{n\to\infty} \|x_n\| = \gamma\mu(\sigma)^{1/2}$. We have

$$\|g_n\| \le \frac{\|x_n\|}{1-\eta} \le \frac{1}{4}\eta\mu(\sigma)^{1/2}.$$

With the notation $\sigma_n = \{\omega \in \sigma : \|g_n(\omega)\| < 1/2\}$, we have

$$\mu(\sigma \setminus \sigma_n) \le 4\int_{\sigma\setminus\sigma_n} \|g_n(\omega)\|^2 \, d\mu(\omega) \le 4\|g_n\|^2 \le \frac{1}{4}\eta^2\mu(\sigma),$$

so that $\mu(\sigma_n) \ge (1 - 2^{-2}\eta^2)\mu(\sigma)$. Passing to a subsequence, we may assume that

(a) $x_n$ converges weakly to a vector$x \in \mathfrak{H}$,
(b) $\chi_{\sigma_n} g_n$ converges weakly to $u \in L^2(\mu, \mathfrak{D})$,
(c) $\chi_{\Omega\setminus\sigma_n} b_n$ converges weakly to $v \in L^2(\mu, \mathfrak{D})$, and
(d) $\chi_{\sigma_n}$ converges weak* in $L^\infty$ to $f$, $0 \le f \le \chi_\sigma$.

We have $\int_\sigma f \, d\mu = \lim_{n\to\infty} \mu(\sigma_n) \ge (1 - 2^{-2}\eta^2)\mu(\sigma)$, and therefore $\|f\|_\infty > 1 - \eta^2/2$. Lemma 6.2 yields a bounded function $z \in \mathfrak{H}$ such that

(e) $\langle z, u\rangle = \langle z, v\rangle = \langle z, x\rangle = \langle z, h_j\rangle = \langle z, fh_j\rangle = 0$ for $j = 1, 2, \dots, p$, and
(f) $\|(1-f)^{1/2}z\|^2 < \eta^2\|f^{1/2}z\|^2$.

Dividing $z$ by a sufficiently large constant, we may also assume that

(g) $\|z(\omega)\| \le 1/2$ for almost all $\omega$.

We will obtain a contradiction by showing that, for large $n$, the vector $x_n' = x_n + z$ belongs to $\mathfrak{S}$ and $\|x_n'\| > \gamma\mu(\sigma)^{1/2}$. This last inequality is verified because

$$\lim_{n\to\infty} \|x_n'\|^2 = \gamma^2\mu(\sigma) + \|z\|^2 > \gamma^2\mu(\sigma)$$

by (a) and (e). Condition (iv) in the definition of $\mathfrak{S}$ is satisfied by (e). To verify the remaining properties, we set $g_n' = g_n + \chi_{\sigma_n} z$ and $b_n' = b_n + \chi_{\Omega\setminus\sigma_n} z$, and observe that condition (i) is satisfied by (g) and the definition of $\sigma_n$. Next we calculate

$$\begin{aligned}\|b_n'\|^2 - \eta^2\|g_n'\|^2 &= \|b_n\|^2 - \eta^2\|g_n\|^2 + \|\chi_{\Omega\setminus\sigma_n} z\|^2 - \eta^2\|\chi_{\sigma_n} x\|^2 \\ &\quad + 2\Re[\langle b_n, \chi_{\Omega\setminus\sigma_n} z\rangle - \eta^2\langle g_n, \chi_{\sigma_n} x\rangle] \\ &\le \|\chi_{\Omega\setminus\sigma_n} z\|^2 - \eta^2\|\chi_{\sigma_n} x\|^2 + 2\Re[\langle b_n, \chi_{\Omega\setminus\sigma_n} z\rangle - \eta^2\langle g_n, \chi_{\sigma_n} x\rangle],\end{aligned}$$

where we used property (ii) for $x_n$. Using (b–f) we obtain

$$\limsup_{n\to\infty}(\|b_n'\|^2 - \eta^2\|g_n'\|^2) \le \|(1-f)^{1/2}z\|^2 - \eta^2\|f^{1/2}z\|^2 < 0,$$

so that $x_n'$ also satisfies (ii) eventually. Finally,

$$\begin{aligned}|\langle g_n', h_j\rangle| &\le |\langle g_n, h_j\rangle| + |\langle \chi_{\sigma_n} z, h_j\rangle| \le \delta\|x_n\| + |\langle \chi_{\sigma_n} z, h_j\rangle| \\ &\le \delta\gamma\mu(\sigma)^{1/2} + |\langle \chi_{\sigma_n} z, h_j\rangle,\end{aligned}$$

so that

$$\limsup_{n\to\infty} |\langle g_n', h_j\rangle| \le \delta\gamma\mu(\sigma)^{1/2},$$

and we see that condition (iii) is satisfied for large $n$. Thus $x_n' \in \mathfrak{S}$, as claimed. □

Proposition 6.4 can be improved considerably.

**Proposition 6.5.** *We have* $\sup\{\|x\| : x \in \mathfrak{S}(\sigma; h_1, h_2, \dots, h_p; \eta; \delta) \ge \mu(\sigma)^{1/2}$ *for all* $\delta, \eta \in (0, 1)$.

*Proof.* The set $\mathfrak{S}$ is smaller if $\delta, \eta$ are smaller, so it suffices to show that $\sup\{\|x\| : x \in \mathfrak{S}\} \ge (1-\eta)\mu(\sigma)^{1/2}$ for $\eta < 1/2$. Suppose to the contrary that the supremum is $\alpha(1-\eta)\mu(\sigma)^{1/2}$ for some $\alpha < 1$. Choose elements $x_n = g_n + b_n \in \mathfrak{S}$ such that $\lim_{n\to\infty} \|x_n\| = \alpha(1-\eta)\mu(\sigma)^{1/2}$, and observe that $\|g_n\| \le \alpha\mu(\sigma)^{1/2}$. Denoting $\sigma_n = \{\omega \in \sigma : \|g(\omega)\| < \alpha^{1/2}\}$, we have

$$\mu(\sigma \setminus \sigma_n) \le \frac{1}{\alpha} \int_{\sigma\setminus\sigma_n} \|g(\omega)\|^2 \, d\mu(\omega) \le \alpha\mu(\sigma),$$

so that $\mu(\sigma_n) \ge (1-\alpha)\mu(\sigma)$. Choose now positive numbers $\delta_n \to 0$, and vectors $z_n \in \mathfrak{S}_n = \mathfrak{S}(\sigma_n; b_n, g_n, h_1, h_2, \dots, h_p; \eta, \delta_n)$ such that

$$\|z_n\| \ge 2^{-3}\eta\mu(\sigma_n)^{1/2}.$$

This is possible by Proposition 6.4. The orthogonality of $z_n$ and $x_n$ implies that the vectors $x_n' = x_n + (1-\alpha^{1/2})z_n$ have norm greater than $\alpha(1-\eta)\mu(\sigma)^{1/2}$ for large $n$, and we will obtain a contradiction by showing that $x_n' \in \mathfrak{S}$ for large $n$. To do this, note that (iv) is satisfied. Write $z_n = \gamma_n + \beta_n$ as required by the definition of the sets $\mathfrak{S}_n$, and set $g_n' = g_n + (1-\alpha^{1/2})\gamma_n$, $b_n' = b_n + (1-\alpha^{1/2})\beta_n$, so that $x_n' = b_n' + g_n'$. Clearly (i) is satisfied by the definition of $\sigma_n$. Next, observe as in the proof of the preceding proposition that

$$\|b_n'\|^2 - \eta^2\|g_n'\|^2 \le \|b_n\|^2 - \eta^2\|g_n\|^2 + 2(1-\alpha^{1/2})\Re[\langle g_n, \gamma_n\rangle + \langle b_n, \beta_n\rangle.$$

Since

$$|\langle b_n, \beta_n\rangle| = |\langle g_n, \gamma_n\rangle| \le \delta_n\|z_n\| \le \delta_n(1+\eta)\mu(\sigma)^{1/2},$$

we deduce that (ii) is satisfied if $\delta_n$ is chosen such that

$$\delta_n(1+\eta)\mu(\sigma)^1/2 < \eta^2\|g_n\|^2 - \|b_n\|^2.$$

Similarly, (iii) is satisfied provided $\delta_n$ is sufficiently small. □

We can now prove Theorem 6.3.

*Proof.* Assuming that $f \ne 0$, choose pairwise disjoint sets $\sigma_1, \sigma_2, \dots, \sigma_n$ with positive measure, and scalars $\gamma_1, \gamma_2, \dots, \gamma_n$ such that $\|f - \sum_{i=1}^n \gamma_i\chi_{\sigma_i}\|_1 < \varepsilon/2$ and $\sum_{i=1}^n |\gamma_i|\mu(\sigma_i) < \|f\|_1$. Fix $\delta, \eta > 0$, and use Proposition 6.5 to find inductively vectors $z_i \in \mathfrak{S}_i = \mathfrak{S}(\sigma_i; h_1, \dots, h_p, z_1, \dots, z_{i-1}; \eta; \delta)$ such that $\|z_i\| \ge (1-\eta)\mu(\sigma_i)^{1/2}$ for $i = 1, 2, \dots, n$. Factor $\gamma_i = \alpha_i\overline{\beta_i}$ with $|\alpha_i| = |\beta_i| = |\gamma_i|^{1/2}$; if $\gamma_i > 0$, choose $\alpha_i = \beta_i > 0$. We claim that the vectors $x = \sum_{i=1}^n \alpha_i z_i$ and $y = \sum_{i=1}^n \beta_i z_i$ satisfy the requirements of the theorem provided that $\delta, \eta$ are small enough. Conditions (3) and (4) are trivially verified. By orthogonality,

$$\|x\|^2 = \|y\|^2 = \sum_{i=1}^{n} |\gamma_i| \|z_i\|^2 \leq \sum_{i=1}^{n} |\gamma_i| (\|g_i\| + \|b_i\|)^2 \leq (1+\eta) \sum_{i=1}^{n} |\gamma_i| \mu(\sigma_i),$$

so that (2) is satisfied when $\eta$ is sufficiently small. Next observe that $g_i \cdot g_j^* = 0$ when $i \neq j$ because $\sigma_i \cap \sigma_j = \varnothing$. Thus

$$\begin{aligned} \|f - x \cdot y^*\|_1 &< \frac{\varepsilon}{2} + \left\| \sum_{i=1}^{n} \gamma_i \chi_{\sigma_i} - x \cdot y^* \right\|_1 \\ &\leq \frac{\varepsilon}{2} + \sum_{i=1}^{n} |\gamma_i| \|\chi_{\sigma_i} - g_i \cdot g_i^*\|_1 + \sum_{i,j=1}^{n} |\alpha_i| |\beta_j| (2\|g_i\| \|b_j\| + \|b_i\| \|b_j\|), \end{aligned}$$

and it suffices to show that each term in these sums can be made arbitrarily small. This is obvious for the terms containing a factor $\|b_i\|$, and

$$\|\chi_{\sigma_i} - g_i \cdot g_i^*\|_1 = \mu(\sigma_i) - \|g_i\|^2 \leq \mu(\sigma_i) - \frac{\|x_i\|^2}{(1+\eta)^2} \leq \mu(\sigma_i) - \left(\frac{1-\eta}{1+\eta}\right)^2 \mu(\sigma_i),$$

which is $< 4\eta$. The theorem is proved. □

It should be noted that the approximate factorization in Theorem 6.3 cannot generally be replaced by exact factorization $f = x \cdot y^*$ with $x, y \in \mathfrak{H}$. An easy example is provided by the Hardy space $H^2$, for which the F. and M. Riesz theorem implies that the nonzero products $x\bar{y}$ cannot vanish on a set of positive measure. The factorable functions $f \in L^1$ are precisely those for which $\log |f|$ is integrable. However, exact factorization is possible if one of the vectors is allowed to be in $L^2(\mu, \mathfrak{D})$. This follows from the following result.

**Theorem 6.6.** *Assume that $\mathfrak{H} \subset L^2(\mu, \mathfrak{D})$ is localizable. Then for every nonnegative function $h \in L^1(\mu)$ and every $\varepsilon > 0$ there exists a vector $x \in \mathfrak{H}$ such that $\|x(\omega)\|^2 \geq h(\omega)$ almost everywhere, and $\|x\|^2 < \|h\|_1 + \varepsilon$.*

*Proof.* Fix a nonnegative function $h \in L^1(\mu)$, and observe that the conclusion of the theorem is true with $x = 0$ if $\|h\|_1 = 0$. We assume therefore that $\|h\|_1 \neq 0$. Upon replacing $h$ by $\mu(\sigma)h/\|h\|_1$ we may actually restrict ourselves to the case when $\|h\|_1 = \mu(\sigma)$, where $\sigma = \{\omega : h(\omega) \neq 0\}$. Define $\rho = h + \chi_{\Omega \setminus \sigma}$, so that $\|\rho\|_1 = 1$. Replace $d\mu$ by $d\mu' = \rho\, d\mu$ and $\mathfrak{H}$ by $\mathfrak{H}' = \rho^{-1/2}\mathfrak{H}$. Performing this substitution, we can assume from the beginning that $f = \chi_\sigma$.

Fix a number $\alpha \in (0,1)$, and set $\delta = \alpha^2/4$. Theorem 6.3 implies the existence of $x_1 \in \mathfrak{H}$ such that $\|(1+\alpha)\chi_\sigma - x_1 \cdot x_1\|_1 < \delta$. If we set

$$\sigma_1 = \{\omega \in \sigma : \|x_1(\omega)\|^2 \leq 1 + \alpha/2\}$$

then $\mu(\sigma_1) \leq 2\delta/\alpha = \alpha/2$. Observe also that we have

$$\|x_1\|^2 \leq \|h\|_1 + \alpha + \delta \leq \|h\|_1 + 2\alpha = 1 + 2\alpha.$$

We will construct by induction vectors $x_n$ such that

(a) $\mu(\sigma_n) \leq \alpha/2^n$, where $\sigma_n = \{\omega \in \sigma : \|x_n(\omega)\|^2 \leq 1 + \alpha/2^n\}$, and
(b) $\|x_{n+1} - x_n\| \leq (\alpha/2^{n-4})^{1/2}$.

Assume that $x_n$ has been constructed, and define $g_n \in L^1(\mu)$ by $g_n = 9\chi_{\sigma_n}$. By (a), we have $\|g_n\|_1 \leq 9\alpha/2^n$. Let $\delta_n$ be a small positive number, subject to certain conditions to be specified shortly (in fact $\delta_n = \alpha^3/10^n$ will satisfy all the requirements). Theorem 6.3 implies the existence of $y_n \in \mathfrak{H}$ such that $\|g_n - y_n \cdot y_n\|_1 < \delta_n$. Observe that

$$\|y_n\|^2 \leq \|g_n\|_1 + \delta_n \leq 9\alpha/2^n + \delta_n,$$

so that $\|y_n\| \leq (\alpha/2^{n-4})^{1/2}$ if $\delta_n$ is chosen sufficiently small. Define $x_{n+1} = x_n + y_n$, and note that condition (b) is satisfied. To complete the inductive process we must show that (a) is satisfied with $n+1$ in place of $n$, provided that $\delta_n$ is chosen sufficiently small. Consider a point $\omega \in \sigma$ such that $|g_n(\omega) - \|y_n(\omega)\|^2| < (\alpha/2^{n+3})^2$. If $\omega \notin \sigma_n$ this means that $\|y_n(\omega)\| < \alpha/2^{n+3}$. If $\|x_n(\omega)\| > 2$ then certainly $\|x_{n+1}(\omega)\| \geq 3/2$ and $\|x_{n+1}(\omega)\|^2 \geq 1 + \alpha/2^{n+1}$. If $\|x_n(\omega)\| \leq 2$ then

$$\begin{aligned}\|x_{n+1}(\omega)\|^2 &\geq (\|x_n(\omega)\| - \|y_n(\omega)\|)^2\\ &\geq \|x_n(\omega)\|^2 - 2\|x_n(\omega)\|\|y_n(\omega)\|\\ &\geq 1 + \frac{\alpha}{2^n} - 4\frac{\alpha}{2^{n+3}} = 1 + \frac{\alpha}{2^{n+1}}.\end{aligned}$$

On the other hand, if $\omega \in \sigma_n$, then $\|y_n(\omega)\|^2 \geq 9 - (\alpha/2^{n+3})^2 \geq 8$ and $\|x_n(\omega)\|^2 \leq 2$. Therefore

$$\begin{aligned}\|x_{n+1}(\omega)\|^2 &\geq (\|y_n(\omega)\| - \|x_n(\omega)\|)^2\\ &\geq (2\sqrt{2} - \sqrt{2})^2 = 2 \geq 1 + \frac{\alpha}{2^{n+1}}.\end{aligned}$$

We conclude that

$$\sigma_{n+1} \subset \left\{\omega : |g_n(\omega) - \|y_n(\omega)\|^2| \geq \left(\frac{\alpha}{2^{n+3}}\right)^2\right\},$$

and therefore $\mu(\sigma_{n+1}) \leq \delta_n(2^{n+3}/\alpha)^2$. It is easy to choose now $\delta_n$ in order to satisfy (a).

Denote by $x$ the limit of the sequence $\{x_n\}_{n=1}^\infty$. Since $\sum_n \mu(\sigma_n) < \infty$, it follows that $\|x_n(\omega)\|^2 \geq h(\omega)$ almost everywhere. Moreover,

$$\|x\| \leq \|x_1\| + \sum_{n=1}^\infty \|x_{n+1} - x_n\| \leq (\|h\|_1 + 2\alpha)^{1/2} + \sum_{n=1}^\infty \left(\frac{\alpha}{2^{n-4}}\right)^{1/2},$$

so that $\|x\|^2 < \|h\|_1 + \varepsilon$ for sufficiently small $\alpha$. The theorem follows. □

The preceding result is useful when dealing with algebras consisting of subnormal operators.

## 7. Contractions with isometric functional calculus

We return now to the study of contraction operators $T$ on a separable Hilbert space $\mathfrak{H}$. We have noted earlier that the Sz.-Nagy–Foias functional calculus is defined when $T$ is completely nonunitary. This calculus is also defined when $T$ is unitary and its spectral measure is absolutely continuous relative to arclength measure on $\mathbb{T}$. We will say that $T$ is *absolutely continuous* when its unitary summand has absolutely continuous spectral measure.

In terms of unitary dilations, absolute continuity is equivalent to the existence of a a unitary $U \in \mathcal{B}(\mathfrak{K})$, where $\mathfrak{K} \supset \mathfrak{H}$, $\mathfrak{H}$ is semi-invariant for $U$, and $T = P_{\mathfrak{H}}U|\mathfrak{H}$. Up to unitary equivalence, $\mathfrak{K}$ may be assumed to be contained in $L^2(m, \mathfrak{D})$, where $m$ denotes normalized arclength measure on $\mathbb{T}$, and $U$ is multiplication by $z$: $(Uf)(z) = zf(z)$ for $f \in \mathfrak{K}$ and $z \in \mathbb{T}$. Given $x, y \in \mathfrak{H}$, the function $x \cdot y^* \in L^1(m)$ does not depend on the particular unitary equivalence used. Indeed, its Fourier coefficients are

$$\widehat{x \cdot y^*}(n) = \int_{\mathbb{T}} \bar{z}^n (x \cdot y^*)(z)\, dm(z) = \begin{cases} \langle T^{*n}x, y\rangle & \text{for } n > 0, \\ \langle T^{-n}x, y\rangle & \text{for } n \geq 0. \end{cases}$$

It is also clear that

$$(x \otimes_T y^*)(u) = \int_{\mathbb{T}} u(z)(x \cdot y^*)(z)\, dm(z)$$

for all $u \in H^\infty$.

We define the class $\mathbb{A}$ to consist of those absolutely continuous contractions such that $\|f(T)\| = \|f\|_\infty$ for every $f \in H^\infty$.

**Proposition 7.1.** *If $T \in \mathbb{A}$, then the space $\mathfrak{H}$ is localizable when viewed as a subspace of $L^2(m, \mathfrak{D})$.*

*Proof.* Fix $\varepsilon > 0$, a set $\sigma \subset \mathbb{T}$ with positive measure, and choose $\delta > 0$. There exists a function $u \in H^\infty$ such that $|u| = \chi_\sigma + \delta\chi_{\mathbb{T}\setminus\sigma}$ almost everywhere. By hypothesis, there exists a unit vector $x \in \mathfrak{H}$ such that $\|u(T)x\|^2 > 1 - \delta$, and therefore $\|ux\|_2^2 = \|u(U)x\|^2 \geq \|u(T)x\|^2 > 1 - \delta$ as well. Thus

$$\begin{aligned} 1 - \delta &= \|\chi_\sigma x\|_2^2 + \|\chi_{\mathbb{T}\setminus\sigma}x\|_2^2 - \delta \\ &< \|ux\|_2^2 = \|\chi_\sigma x\|_2^2 + \delta^2\|\chi_{\mathbb{T}\setminus\sigma}x\|_2^2, \end{aligned}$$

from which we infer $\|\chi_\sigma x\|^2 > (1 - 2\delta)/(1 - \delta)$. It follows that $\|\chi_{\mathbb{T}\setminus\sigma}x\| < \varepsilon\|\chi_\sigma x\|$ if $\delta$ is sufficiently small. □

**Corollary 7.2.** *If $T \in \mathbb{A} \cap C_{00}$ then $\mathcal{W}_T$ has property $(\mathbb{X}_1)$. In particular, $\mathcal{W}_T$ has property $(\mathbb{A}_1(1))$ and is hyper-reflexive with constant $\leq 3$.*

*Proof.* For every $f \in L^1(m)$, Theorem 6.3 implies the existence of orthogonal sequences $x_n, y_n \in \mathfrak{H}$ such that $\|x_n\|, \|y_n\| \leq \|f\|_1$ and $\lim_{n\to\infty} \|f - x_n \cdot y_n^*\|_1 = 0$. The corollary follows then from Lemma 4.4 because all weak* continuous functionals on $H^\infty$ are given by functions in $L^1$. □

Corollary 7.2 is not true for arbitrary $T \in \mathbb{A}$, as shown for instance by a shift of multiplicity one. However, we have the following result.

**Theorem 7.3.** *For every $T \in \mathbb{A}$, the algebra $\mathcal{W}_T$ has property $(\mathbb{A}_1(1))$.*

**Corollary 7.4.** *Any contraction $T \in \mathcal{B}(\mathfrak{H})$ satisfying $\sigma(T) \supset \mathbb{T}$ has nontrivial invariant subspaces.*

The corollary follows from a theorem of Apostol [10] which implies that, provided $\sigma(T) \supset \mathbb{T}$, we either have $T \in \mathbb{A}$, or $T$ has nontrivial hyperinvariant subspaces.

We will provide the proof of Theorem 7.3 under the additional hypothesis that $T$ is of class $C_{0\cdot}$; note that this case suffices for the proof of Corollary 7.4.

Given $f \in L^1$, we denote by $[f]$ the weak* continuous functional on $H^\infty$ determined by $f$, i.e., $[f](u) = \int_{\mathbb{T}} u(z)f(z)\,dm(z)$ for $u \in H^\infty$. We have $[f] = 0$ if and only if $\int_{\mathbb{T}} z^n f(z)\,dm(z) = 0$ for $n \geq 0$. With this notation, we have $x \otimes_T y^* = [x \cdot y^*]$ for $x, y \in \mathfrak{H}$.

For the remainder of this section, we will view $\mathfrak{H}$ as a subspace of $L^2(m, \mathfrak{D})$, and we denote by $\mathfrak{K}_+$ the smallest invariant subspace for $U$ containing $\mathfrak{H}$. The operator $U_+ = U|\mathfrak{K}_+$ satisfies $U_+^*\mathfrak{H} \subset \mathfrak{H}$ and $T^* = U_+^*|\mathfrak{H}$. It follows that $[x \cdot y^*] = [P_{\mathfrak{H}} x \cdot y^*]$ for $x \in \mathfrak{K}_+$ and $y \in \mathfrak{H}$. Indeed,

$$\langle U_+^n x, y\rangle = \langle x, U_+^{*n} y\rangle = \langle x, T^{*n} y\rangle = \langle P_{\mathfrak{H}}, T^{*n} y\rangle = \langle T^n P_{\mathfrak{H}}, y\rangle \quad \text{for } n \geq 0.$$

We will need the Wold decomposition of $U_+$. Thus, write $\mathfrak{K}_+ = \mathfrak{M} \oplus \mathfrak{R}$ such that $U_+|\mathfrak{M}$ is a unilateral shift, and $U_+|\mathfrak{R}$ is a unitary operator. Given vectors $m, m' \in \mathfrak{M}$ and $r, r' \in \mathfrak{R}$, we have

$$(m + r) \cdot (m' + r')^* = m \cdot m'^* + r \cdot r'^*.$$

The assumption that $T \in C_{0\cdot}$ implies that $\lim_{n\to\infty} \|[x \cdot y_n^*]\| = 0$ if $x, y_n \in \mathfrak{H}$ and $y_n$ tends to zero weakly. Similarly, $U_+|\mathfrak{M} \in C_{\cdot 0}$, and thus $\lim_{n\to\infty} \|[m_n \cdot m^*]\| = 0$ if $m, m_n \in \mathfrak{M}$ and $m_n$ tends to zero weakly.

The essential step in the proof of Theorem 7.3 is an approximation argument which can be repeated as in the proof of the open mapping theorem.

**Proposition 7.5.** *Let $\alpha, \varepsilon > 0$, $x, y \in \mathfrak{H}$, and $f \in L^1$ satisfy $\|[f] - [x \cdot y^*]\| < \alpha$. There exist $x_1, y_1 \in \mathfrak{H}$ with the following properties:*

(1) $\|[f] - [x_1 \cdot y_1^*]\| < \varepsilon$,
(2) $\|y - y_1\| < 3\alpha^{1/2}$, *and*
(3) $\|x_1\| < \|x\| + 3\alpha^{1/2}$.

*Proof.* Fix a positive number $\delta$, and denote $\beta = 1 + \|x\| + \|y\|$. By Theorem 6.3, there exist orthogonal sequences $(x_n)_{n=2}^\infty$ and $(y_n)_{n=2}^\infty$ in $\mathfrak{H}$ such that $\|x_n\|, \|y_n\| < \alpha^{1/2}$ and

$$\|[f] - [x \cdot y^*] - [x_n \cdot y_n^*]\| < \delta.$$

Using the observations preceding the statement, we see that $\|[x \cdot y_n^*]\| < \delta$ and $\|[P_{\mathfrak{M}} x_n \cdot y]\| < \delta$ for large $n$. Fix such a value $n$, and define $x' = P_{\mathfrak{H}}(x + P_{\mathfrak{M}} x_n)$ and $y' = y + y_n$. We have then

(i) $\|x-x'\|, \|y-y'\| < \alpha^{1/2}$ and
(ii) $\|[f]-[x'\cdot y'^*]-[P_{\mathfrak{R}}x_n\cdot(P_{\mathfrak{R}}y_n)^*]\| < 3\delta\beta$

Define next the measurable set

$$\sigma=\{z\in\mathbb{T}:\|(P_{\mathfrak{R}}y_n)(z)\|\le\|(P_{\mathfrak{R}}y')(z)\|\},$$

and consider a function $\psi\in H^\infty$ such that $|\psi|=\delta\chi_\sigma+2\chi_{\mathbb{T}\setminus\sigma}$ a.e. Choose an integer $N$ so that $\|U_+^{*N}P_{\mathfrak{M}}y_n\|<\delta$, and define

$$y_1=y'+T^{*N}\psi(T^*)y_n.$$

Clearly,

$$\|y-y_1\|\le\|y-y'\|+\|y'-y_1\|<\alpha^{1/2}+2\|y_n\|<3\alpha^{1/2},$$

so that (2) is verified. The choice of $\sigma$, and the fact that $P_{\mathfrak{R}}(u(T^*)y_n)(z)=u(\bar{z})(P_{\mathfrak{R}}y_n)(z)$, imply that

(iii) $\|(P_{\mathfrak{R}}y_1)(z)\|\ge(1-\delta)\max\{\|(P_{\mathfrak{R}}y')(z)\|,\|(P_{\mathfrak{R}}y_n)(z)\|\}$ a.e.

We can therefore find a measurable function $g$ on $\mathbb{T}$ such that

$$g(z)\|(P_{\mathfrak{R}}y_1)(z)\|^2=(x'\cdot(P_{\mathfrak{R}}y')^*)(z)+(P_{\mathfrak{R}}x_n\cdot(P_{\mathfrak{R}}y_n)^*)(z)$$

almost everywhere, and $g(z)=0$ when $(P_{\mathfrak{R}}y_1)(z)=0$. The inequality (iii) easily implies

$$|g(z)|\|(P_{\mathfrak{R}}y_1)(z)\|\le\frac{1}{1-\delta}(\|(P_{\mathfrak{R}}x')(z)\|+\|(P_{\mathfrak{R}}x_n)(z)\|)$$

almost everywhere. Thus the function $x_2\in\mathfrak{R}$ defined by $x''(z)=g(z)(P_{\mathfrak{R}}y_1)(z)$ satisfies

$$\|x''\|\le\frac{1}{1-\delta}(\|P_{\mathfrak{R}}x'\|+\|x_n\|)<\frac{1}{1-\delta}(\|P_{\mathfrak{R}}x'\|+\alpha^{1/2}),$$

and

(iv) $x''\cdot y_1=x'\cdot(P_{\mathfrak{R}}y')^*+P_{\mathfrak{R}}x_n\cdot(P_{\mathfrak{R}}y_n)^*$.

We define now $x_1=P_{\mathfrak{H}}(P_{\mathfrak{M}}x'+x'')$, and observe that

$$\begin{aligned}\|x_1\|^2&\le\|P_{\mathfrak{M}}x'\|^2+\|x''\|^2<\|P_{\mathfrak{M}}x'\|^2+\frac{1}{(1-\delta)^2}(\|P_{\mathfrak{R}}x'\|+\alpha^{1/2}))^2\\&\le\frac{1}{(1-\delta)^2}(\|x'\|+\alpha^{1/2})^2,\end{aligned}$$

which implies (3) if $\delta$ is sufficiently small. Finally, we use (iv) to calculate

$$[x_1\cdot y_1^*]=[(P_{\mathfrak{M}}x'+x'')\cdot y_1^*]=[(P_{\mathfrak{M}}x')\cdot y_1^*]+[x'\cdot(P_{\mathfrak{R}}y')^*]+[P_{\mathfrak{R}}x_n\cdot(P_{\mathfrak{R}}y_n)^*].$$

Using the definition of $y_1$, we see that

$$\|[(P_{\mathfrak{M}}x')\cdot y_1^*]-[(P_{\mathfrak{M}}x')\cdot y'^*]\|=\|[(P_{\mathfrak{M}}x')\cdot(P_{\mathfrak{M}}(T^{*N}\psi(T^*)y_n)^*]\|<\delta(\beta+\alpha^{1/2}).$$

These two relations, combined with (ii), imply $\|[f]-[x_1\cdot y_1^*]\|<4\delta(\beta+\alpha^{1/2})$ so that condition (1) is also satisfied if $\delta$ is small enough. □

We can now complete the proof of Theorem 7.3 in the $C_0.$-case.

*Proof.* Fix $f \in L^1$ with $\|f\|_1 < 1$, and $0 < \delta < 1 - \|f\|_1$. Proposition 7.5 allows us to construct sequences of vectors $x_n, y_n \in \mathfrak{H}$ such that

(a) $\|[f] - [x_n \cdot y_n^*]\|_1 < \delta^{4n}$ for $n \geq 1$,
(b) $\|x_1\|, \|y_1\| \leq (1-\delta)^{1/2}$,
(c) $\|y_n - y_{n-1}\| < 3\delta^{2n}$ for $n \geq 2$, and
(d) $\|x_n\| < \|x_{n-1}\| + 3\delta^{2n}$ for $n \geq 2$.

The sequence $y_n$ is Cauchy and its limit $y$ satisfies $\|y\| < (1-\delta)^{1/2} + 3\sum_{n=1}^{\infty} \delta^{2n}$, while some subsequence of the $x_n$ converges weakly to a vector $x$ satisfying the same norm estimate. Clearly $[f] = [x \cdot y^*]$ and $\|x\|, \|y\| < 1$ if $\delta$ is sufficiently small. □

The same argument easily yields the fact that the map $(x, y) \mapsto x \otimes_T y^*$ is open, not just at the origin.

## 8. Banach space geometry

In attempting to extend to Banach spaces the techniques developed in the context of Hilbert space, one encounters two main difficulties. The first one is the fact that the sets $\widetilde{S}_F$ considered in Section 2 are generally not convex. The second difficulty is the lack of a functional calculus. Thus, if $\mathfrak{H}$ is a Banach space and $T \in \mathcal{B}(\mathfrak{H})$ is a contraction, we cannot generally construct a calculus with functions in $H^\infty$, even when $T \in C_{00}$ in the appropriate sense. There are two ways to deal with the issue of functional calculus. One can assume that an $H^\infty$ functional calculus exists. There are many interesting examples in which this assumption does hold. A second way out is to use a local version of the $H^\infty$ functional calculus, and this also covers numerous examples.

In this section we will deal with the abstract preliminaries needed in an appropriate substitute for the argument of Lemma 2.1.

**Theorem 8.1.** *Let $\mathfrak{H}$ be a Banach space of dimension $n < \infty$, let $(e_j)_{j=1}^n$ be a basis in $\mathfrak{H}$, and let $(\varphi_j)_{j=1}^n \subset \mathfrak{H}^*$ be the dual basis, i.e., $\varphi_i(x_j) = \delta_{ij}$. Given positive numbers $(\gamma_j)_{j=1}^n$ such that $\sum_{j=1}^n \gamma_j = 1$, there exist $\alpha = (\alpha_j)_{j=1}^n, \beta = (\beta_j)_{j=1}^n \in \mathbb{C}^n$ such that*

(1) $\|\sum_{j=1}^n \alpha_j e_j\| = \|\sum_{j=1}^n \beta_j \varphi_j\| = 1$, *and*
(2) $\alpha_j \beta_j = \gamma_j$ *for* $j = 1, 2, \ldots, n$.

*Proof.* Assume that $f : \mathfrak{H} \to \mathbb{R}$ is a function such that $f|\{h : \|h\| \leq 1\}$ attains its maximum at a point $x$ such that $\|x\| = 1$, and $f$ is differentiable at $x$. In this case, the linear functional $f'(x) : \mathfrak{H} \to \mathbb{R}$ attains its norm at $x$. Indeed, for any vector $y \in \mathfrak{H}$ with $\|y\| \leq 1$, we have

$$f'(x)(x-y) = \lim_{t \downarrow 0} \frac{F(x) - F(x + t(y-x))}{t} \geq 0$$

since $\|x+t(y-x)\| = \|(1-t)x+ty\| \leq 1$ for $t \in (0,1)$. We apply this observation to the function $f : \mathfrak{H} \to [0,+\infty)$ defined by

$$f(h) = \prod_{j=1}^{n} |h_j|^{\gamma_j}, \quad h = \sum_{j=1}^{n} h_j e_j \in \mathfrak{H}.$$

Assume that $f$ attains its maximum on the unit ball at a point $x = \sum_{j=1}^{n} \alpha_j e_j$. Obviously $\|x\| = 1$, and none of the coordinates $\alpha_j$ are zero. We set then $\beta_j = \gamma_j/\alpha_j$, $\varphi = \sum_{j=1}^{n} \beta_j \varphi_j$, and conclude the proof by showing that $\Re\varphi$ is a positive multiple of $f'(x)$. Indeed, writing $\alpha_j = \xi_j + i\eta_j$,

$$f'(x)e_j = \lim_{t \downarrow 0} \frac{f(x+te_j) - f(x)}{t} = f(x)\frac{\gamma_j \xi_j}{\xi_j^2 + \eta_j^2} = f(x)\Re\varphi(e_j),$$

and similarly $f'(x)(ie_j) = f(x)\Re\varphi(ie_j)$. Thus $\varphi$ attains its norm at $x$, and clearly $\varphi(x) = \sum_{j=1}^{n} \gamma_j = 1$. □

The Hahn-Banach theorem allows us to write this result in an equivalent form.

**Theorem 8.2.** *Let $\mathfrak{H}$ be a Banach space, and let $(e_j)_{j=1}^{n} \subset \mathfrak{H}$ be linearly independent. Given positive numbers $(\gamma_j)_{j=1}^{n}$ such that $\sum_{j=1}^{n} \gamma_j = 1$, there exist $\alpha = (\alpha_j)_{j=1}^{n} \in \mathbb{C}^n$ and $\varphi \in \mathfrak{H}^*$ such that*

(1) $\|\sum_{j=1}^{n} \alpha_j e_j\| = \|\varphi\| = 1$, *and*
(2) $\varphi(\alpha_j e_j) = \gamma_j$ *for* $j = 1, 2, \ldots, n$.

It will be important to control the size of the constants $\alpha_j$ in the above theorem. Given two subspaces $\mathfrak{M}', \mathfrak{M}'' \subset \mathfrak{H}$ we say that $\mathfrak{M}''$ is *$\varepsilon$-orthogonal* to $\mathfrak{M}'$ for some $\varepsilon > 0$ if $\|h' - h''\| \geq (1-\varepsilon)\|h'\|$ for all $h' \in \mathfrak{M}'$ and $h' \in \mathfrak{M}''$.

**Proposition 8.3.** *For every $\varepsilon > 0$ and every finite-dimensional subspace $\mathfrak{M}$ of a Banach space $\mathfrak{H}$, there exists a finite codimensional subspace $\mathfrak{N} \subset \mathfrak{H}$ which is $\varepsilon$-orthogonal to $\mathfrak{M}$.*

*Proof.* Fix $\varepsilon > 0$, and for each unit vector $h \in \mathfrak{M}$ choose a functional $\varphi_h \in \mathfrak{H}^*$ such that $\|\varphi_h\| = \varphi_h(h) = 1$. By compactness, there exist unit vectors $h_1, h_2, \ldots, h_n \in \mathfrak{M}$ such that the sets $\{x \in \mathfrak{M} : |\varphi_{h_j}(x)| > 1 - \varepsilon\}$ cover the unit sphere of $\mathfrak{M}$. Denote $\mathfrak{N} = \bigcap_{j=1}^{n} \ker \varphi_{h_j}$, and pick vectors $h \in \mathfrak{M}$, $k \in \mathfrak{N}$ such that $\|h\| = 1$. We have $\|h - k\| \geq |\varphi_{h_j}(h-k)| = |\varphi_{h_j}(h)|$, and this number is greater than $1 - \varepsilon$ for some $j$. The proposition follows. □

The relation of $\varepsilon$-orthogonality is not symmetric. Note however that the $\varepsilon$-orthogonality of $\mathfrak{M}$ and $\mathfrak{N}$ is equivalent to the requirement that the projection $P : \mathfrak{M} + \mathfrak{N} \to \mathfrak{M}$ with kernel $\mathfrak{N}$ has norm $< 1/(1-\varepsilon)$, and this implies that the projection $Q : \mathfrak{M} + \mathfrak{N} \to \mathfrak{N}$ with kernel $\mathfrak{M}$ has norm $< 1 + 1/(1-\varepsilon)$. Choosing, for instance, $\varepsilon = 1/2$, both of these projections will have norm $< 3$.

**Corollary 8.4.** *Assume that the spaces $\mathfrak{M}, \mathfrak{N} \subset \mathfrak{H}$ are $\varepsilon$ orthogonal, and $\varphi_0 \in \mathfrak{N}^*$. There exists $\varphi \in \mathfrak{H}^*$ such that $\varphi|\mathfrak{N} = \varphi_0$, $\varphi|\mathfrak{M} = 0$, and*

$$\|\varphi\| \leq \left[1 + \frac{1}{1-\varepsilon}\right] \|\varphi_0\|.$$

## 9. Dominating spectrum in Banach spaces

In this section we will deal with an operator $T \in \mathcal{B}(\mathfrak{H})$, with $\mathfrak{H}$ a Banach space, under the assumption that $T$ admits an $H^\infty$ functional calculus. Thus, we assume that there exists a unital algebra homomorphism $\Psi : H^\infty \to \mathcal{B}(\mathfrak{H})$ such that $\Psi(\mathrm{Id}_{\mathbb{D}}) = T$, $\|\Psi(f)\| \leq \|f\|_\infty$ for $f \in H^\infty$, and the map $f \mapsto \varphi(\Psi(f)x)$ is weak* continuous for every $x \in \mathfrak{H}$ and $\varphi \in \mathfrak{H}^*$. This weak* continuous functional will now be denoted $x \otimes_T \varphi$, in agreement with our earlier usage. We will also write $f(T)$ in place $\Psi(f)$, and this agrees with the Riesz–Dunford functional calculus in case $f$ is analytic in a neighborhood of $\overline{\mathbb{D}}$. We will denote by $T^* \in \mathcal{B}(\mathfrak{H}^*)$ the dual of $T$, i.e., $T^*\varphi = \varphi \circ T$ for $\varphi \in \mathfrak{H}^*$. Recall that $e_\lambda$ is the functional defined by $e_\lambda(f) = f(\lambda)$ for $f \in H^\infty$.

**Lemma 9.1.** *Assume that $\varepsilon > 0$, $\lambda \in \mathbb{D}$, $x \in \mathfrak{H}$, and $\varphi \in \mathfrak{H}^*$ are such that*

$$\min\left\{\frac{\|(T-\lambda)x\|}{\|x\|}, \frac{\|(T^*-\lambda)\varphi\|}{\|\varphi\|}\right\} < \varepsilon.$$

*Then we have $\|x \otimes_T \varphi - \varphi(x)e_\lambda)\| < 2\varepsilon\|x\|\|\varphi\|/(1-|\lambda|)$.*

*Proof.* Assume that $\|(T-\lambda)x\| < \varepsilon$. Every $f \in H^\infty$ can be written as $f(z) = f(\lambda) + (z-\lambda)g(z)$ with $g \in H^\infty$ and $\|g\| \leq 2/(1-|\lambda|)$. Therefore

$$|(x \otimes_T \varphi - \varphi(x)e_\lambda)(f)| = |\varphi(g(T)(T-\lambda)x)| \leq \frac{2\varepsilon\|x\|\|\varphi\|}{1-|\lambda|}\|f\|,$$

thus yielding the desired estimate. The case where $\|(T^*-\lambda)\varphi\| < \varepsilon\|\varphi\|$ is treated similarly. □

In order to approximate convex combinations $\sum_{j=1}^n \gamma_j e_{\lambda_j}$ by tensors $x \otimes_T \varphi$, we use Theorem 8.1. Given distinct points $\lambda_1, \lambda_2, \dots, \lambda_n \in \mathbb{D}$, we write

$$\delta(\lambda_1, \lambda_2, \dots, \lambda_n) = \min_{1 \leq j \leq n} \prod_{k \neq j} \left|\frac{\lambda_j - \lambda_k}{1 - \overline{\lambda_j}\lambda_k}\right|.$$

It is easy to see that any $n$ vectors $x_j$ such that $\|(T-\lambda_j)x_j\| < \eta\|x_j\|$ must be linearly independent if $\eta$ is sufficiently small. Indeed, consider the functions (Blaschke products) $B_j \in H^\infty$ defined by

$$B_j(\lambda) = \prod_{k \neq j} \frac{\lambda_j - \lambda}{1 - \overline{\lambda_j}\lambda}.$$

Observe that the vector $x=\sum_{k=1}^n x_k$ satisfies

$$\begin{aligned}\|x\| \geq \|B_j(T)x\| &\geq \|B_j(\lambda_j)x_j\| - \sum_{k=1}^n \|[B_j(T)-B_j(\lambda_k)]x_k\| \\ &\geq \delta(\lambda_1,\lambda_2,\dots,\lambda_n)\|x_j\| - \sum_{k=1}^n \frac{2\eta}{1-|\lambda_j|}\|x_k\|,\end{aligned}$$

so that

$$\delta(\lambda_1,\lambda_2,\dots,\lambda_n)\|x_j\| \leq \|x\| + \sum_{k=1}^n \frac{2\eta}{1-|\lambda_j|}\|x_k\|.$$

Adding these inequalities we obtain

$$\left[\delta(\lambda_1,\lambda_2,\dots,\lambda_n) - n\sum_{j=1}^n \frac{2\eta}{1-|\lambda_j|}\right]\sum_{k=1}^n \|x_k\| \leq n\left\|\sum_{k=1}^n x_k\right\|. \tag{9.1}$$

Consider now a point $\lambda\in\mathbb{D}$ in the left spectrum $\sigma_\ell(T)$ of $T$, i.e., $\lambda-T$ does not have a left inverse. If $\lambda$ is not an eigenvalue of $T$, then $\lambda$ is in the *left essential spectrum* $\sigma_{\ell e}(T)$. More precisely, for every $\varepsilon>0$, and every finite codimensional subspace $\mathfrak{N}\subset\mathfrak{H}$, there exists a unit vector $x\in\mathfrak{N}$ such that $\|(\lambda-T)x\|<\varepsilon$.

**Lemma 9.2.** *Assume that $\lambda_1,\lambda_2,\dots,\lambda_n\in\sigma_\ell(T)\cap\mathbb{D}$ are distinct points, and $\varepsilon>0$. Given positive numbers $\gamma_j$ such that $\sum_{j=1}^n\gamma_j=1$, there exist unit vectors $x\in\mathfrak{H}$ and $\varphi\in\mathfrak{H}^*$ such that $\|x\otimes_T\varphi-\sum_{j=1}^n\gamma_j e_{\lambda_j}\|<\varepsilon$.*

*Proof.* Choose $y_j\in\mathfrak{H}$ such that $\|(T-\lambda_j)y_j\|<\eta\|y_j\|$ for some $\eta>0$. Theorem 8.2 implies the existence of a unit vector of the form $x=\sum_{j=1}^n x_j$, where $x_j=\alpha_j y_j$, and of a unit vector $\varphi\in\mathfrak{H}^*$ such that $\varphi(x_j)=\gamma_j$ for all $j$. By Lemma 9.1,

$$\left\|x\otimes_T\varphi-\sum_{j=1}^n\gamma_j e_{\lambda_j}\right\| \leq \sum_{j=1}^n \frac{2\eta\|x_j\|}{1-|\lambda_j|},$$

and (9.1) shows that the conclusion is satisfied if $\eta$ is sufficiently small. □

In order to obtain a result similar to Lemma 4.1, the vectors $y_j$ in the above proof must be chosen more carefully.

**Lemma 9.3.** *Let $h_1,h_2,\dots,h_N\in\mathfrak{H}$ be such that $\{f(T)h_j : f\in H^\infty,\|f\|\leq 1\}$ is totally bounded for $1\leq j\leq N$. Assume that $\lambda_1,\lambda_2,\dots,\lambda_n\in\sigma_{\ell e}(T)\cap\mathbb{D}$ are distinct points, and $\varepsilon>0$. Given positive numbers $\gamma_j$ such that $\sum_{j=1}^n\gamma_j=1$, and functionals $\varphi_1,\varphi_2,\dots,\varphi_N\in\mathfrak{H}^*$, there exist unit vectors $x\in\mathfrak{H}$ and $\varphi\in\mathfrak{H}^*$ such that $\|x\|\|\varphi\|<3$, $\|x\otimes_T\varphi-\sum_{j=1}^n\gamma_j e_{\lambda_j}\|<\varepsilon$, and*

$$\sum_{j=1}^N(\|x_j\otimes_T\varphi\|+\|x\otimes_T\varphi_j\|)<\varepsilon.$$

*Proof.* Fix a positive number $\eta$. The hypothesis implies the existence of a finite-dimensional space $\mathfrak{M} \subset \mathfrak{H}$ such that $\operatorname{dist}(f(T)h_j, \mathfrak{M}) < \eta$ for every $j = 1, 2, \ldots, N$, and every $f \in H^\infty$ with $\|f\|_\infty \leq 1$. By Proposition 8.3, there exists a finite codimensional space $\mathfrak{N}_0$ which is $(1/2)$-orthogonal to $\mathfrak{M}$. The finite codimensional space

$$\mathfrak{N} = \mathfrak{N}_0 \cap \bigcap_{k=1}^{N} \ker \varphi_k$$

is also $(1/2)$-orthogonal to $\mathfrak{M}$. Choose now vectors $y_j \in \mathfrak{N}$ such that $\|(T-\lambda_j)y_j\| < \eta\|y_j\|$, and use Theorem 8.2 to find a unit vector $x = \sum_{j=1}^{n} x_j$, $x_j = \alpha_j y_j$, and a functional of norm one $\varphi_0 \in \mathfrak{N}^*$ such that $\varphi_0(x_j) = \gamma_j$ for $j = 1, 2, \ldots, n$. Corollary 8.4 yields a functional $\varphi \in \mathfrak{H}^*$ such that $\|\varphi\| \leq 3$, $\varphi(x_j))|\mathfrak{N} = \varphi_0$, and $\varphi|\mathfrak{M} = 0$. The calculation in the proof of Lemma 9.2 yields now

$$\left\| x \otimes_T \varphi - \sum_{j=1}^{n} \gamma_j e_{\lambda_j} \right\| \leq \sum_{j=1}^{n} \frac{6\eta\|x_j\|}{1-|\lambda_j|},$$

and this number is $< \varepsilon$ for small $\eta$ by (9.1). A similar calculation, using the relations $\varphi_k(x_j) = 0$, yields

$$\|x \otimes_T \varphi_k\| \leq \sum_{j=1}^{n} \frac{2\eta\|x_j\|\|\varphi_k\|}{1-|\lambda_j|},$$

and these numbers will be $< \varepsilon/2N$ for small $\eta$. Finally, the condition $\varphi|\mathfrak{M} = 0$ implies

$$\|h_k \otimes_T \varphi\| = \sup_{\|f\|_\infty \leq 1} |\varphi(f(T)h_k)| \leq \|\varphi\|\eta \leq 3\eta,$$

and this number is also $< \varepsilon/N$ if $\eta$ is sufficiently small. □

We can now conclude that the map $S : \mathfrak{H} \times \mathfrak{H}^* \to H^\infty_*$, $F(h, \varphi) = h \otimes_T \varphi$, satisfies the hypothesis of Proposition 2.5 provided that $\sigma_{\ell e}(T)$ is dominating.

**Proposition 9.4.** *Assume that $\sigma_{\ell e}(T)$ is dominating, and the set*

$$\{f(T)h : f \in H^\infty, \|f\|_\infty \leq 1\}$$

*is totally bounded for every $h \in \mathfrak{H}$. Then the set $\widetilde{S}_F$ contains the ball of radius $1/48$ centered at $0 \in H^\infty_*$.*

*Proof.* Consider points $\lambda_j \in \sigma_{\ell e}(T)$ and scalars $t_j$ such that $\sum_{j=1}^{n} |t_j| \leq 1/48$. It will suffice to show that $\psi = \sum_{j=1}^{n} t_j e_{\lambda_j} \in \widetilde{S}_F$. We can write $\psi = \sum_{\ell=1}^{4} i^\ell s_\ell \psi_\ell$, where $0 \leq s_\ell \leq 1/48$, and each $\psi_\ell$ is a convex combination of the functionals $e_{\lambda_j}$. Lemma 9.3 yields vectors $x_\ell \in \mathfrak{H}$ and $\varphi_\ell \in \mathfrak{H}^*$ such that $\|x_\ell\|, \|\varphi_\ell\| \leq 3^{1/2}$, $\|\psi_\ell - x_\ell \otimes_T \varphi_\ell\|$ is arbitrarily small, and the functionals $x_\ell \otimes_T \varphi_{\ell'}$ have arbitrarily small norm for $\ell \neq \ell'$. The vectors $x = \sum_{\ell=1}^{4} i^\ell s_\ell^{1/2} x_\ell$ and $\varphi = \sum_{\ell=1}^{4} s_\ell^{1/2} \varphi_\ell$ have norm at most one, and $\|\psi - x \otimes_T \varphi\|$ is as small as desired. Choosing the appropriate vectors $x_j, \varphi_j$, we can also require that a finite number of functionals of the form $h \otimes_T \varphi$ and $x \otimes \varphi'$ have arbitrarily small norm. □

The total boundedness in the hypothesis follows if the map $f \mapsto f(T)h$ is continuous from the weak* topology to the norm topology of $\mathfrak{H}$. This amounts to requiring that $T \in C_{0\cdot}$. Proposition 2.5 yields the following result.

**Corollary 9.5.** *Assume that $T \in C_{0\cdot}$ and $\sigma_\ell(T) \cap \mathbb{D}$ is dominating. Then $T$ has nontrivial invariant subspaces.*

The techniques in this section can be refined to yield the following result of Ambrozie and Müller [8].

**Theorem 9.6.** *If $\sigma(T) \supset \mathbb{T}$ then $T$ has nontrivial invariant subspaces.*

One important observation is that we only need to have sufficiently many vectors such thus $\|(T-\lambda)x\| < (1-|\lambda|)^2\|x\|$. Results of Apostol [12] show that if such vectors are not abundant, then $T$ has hyper-invariant subspaces.

## 10. Localizable spectrum

Assume again that $T$ is an operator on a Banach space $\mathfrak{H}$. Given $x \in \mathfrak{H}$ and $\varphi \in \mathfrak{H}^*$, it may happen that there exists a functional $\psi \in H^\infty_*$ such that $\psi(f) = \varphi(f(T)x)$ whenever $f$ is a polynomial. When this situation occurs, we will write $\psi = x \otimes_T \varphi$. The bilinear map $S(x,\varphi) = x \otimes_T \varphi$ is only partially defined in general.

There may be vectors $x \in \mathfrak{H}$ for which $x \otimes_T \varphi$ is always defined. Assume for instance that $|\mu| < 1$, and *the local spectrum of* $x$ is contained in $D(\mu,\delta) = \{\lambda : |\mu - \lambda| < \delta\} \subset \mathbb{D}$. This simply means that there exists a continuous function $u : \mathbb{C} \setminus D(\mu,\delta) \to \mathfrak{H}$, analytic on $u : \mathbb{C} \setminus \overline{D}(\mu,\delta)$, such that

$$(\lambda - T)u(\lambda) = x, \quad x \notin D(\mu,\delta).$$

Such a function obviously satisfies $u(\lambda) = (\lambda - T)^{-1}x$ outside $\sigma(T)$. Assume that $R > 0$ is so large that $\sigma(T) \subset D(0,R)$. For any polynomial$f$ we have

$$f(T)x = \frac{1}{2\pi i}\int_{|\lambda|=R} f(\lambda)(\lambda - T)^{-1}x\,d\lambda = \frac{1}{2\pi i}\int_{|\lambda-\mu|=\delta} f(\lambda)u(\lambda)\,d\lambda$$

by Cauchy's theorem. Thus we can use the formula

$$f(T)x = \frac{1}{2\pi i}\int_{|\lambda-\mu|=\delta'} f(\lambda)u(\lambda)\,d\lambda$$

to define $f(T)x$ for every $f \in H^\infty$ (note though that $f(T)$ itself is not defined). The map $f \mapsto f(T)x$ is weak* to norm continuous.

Let us denote by $\mathfrak{X}_{\mu,\delta}$ the set of continuous functions $u : \mathbb{C} \setminus D(\mu,\delta) \to \mathfrak{H}$, analytic on $\mathbb{C} \setminus \overline{D}(\mu,\delta)$, such that $(\lambda - T)u(\lambda)$ is a constant function. Obviously $\mathfrak{X}_{\mu,\delta}$ is a linear space, and it becomes a Banach space with the norm

$$\|u\|_{\mu,\delta} = \sup\{|u(\lambda)| : |\lambda - \mu| \geq \delta\} = \sup\{|u(\lambda)| : |\lambda - \mu| = \delta\}.$$

If $u \in \mathfrak{X}_{\mu,\delta}$, the function $\lambda \mapsto Tu(\lambda)$ also belongs to $\mathfrak{X}_{\mu,\delta}$, and it will be denoted $T_{\mu,\delta}u$. Obviously $T_{\mu,\delta} \in \mathcal{B}(\mathfrak{X}_{\mu,\delta})$ and $\|T_{\mu,\delta}\| \leq \|T\|$.

**Lemma 10.1.** *We have $\sigma(T_{\mu,\delta}) \subset \overline{D}(\mu,\delta)$.*

*Proof.* Assume that $\alpha \in \mathbb{C} \setminus \overline{D}(\mu,\delta)$ and $u \in \mathfrak{X}_{\mu,\delta}$. Define a function $v$ by

$$v(\lambda) = \begin{cases} -u'(\alpha) & \text{if } \lambda = \alpha, \\ \frac{u(\lambda)-u(\alpha)}{\alpha-\lambda} & \text{if } \lambda \neq \alpha. \end{cases}$$

Since $u \in \mathfrak{X}_{\mu,\delta}$, we have

$$(\lambda - T)v(\lambda) = \frac{(\alpha - T)u(\alpha) - (\lambda - T)u(\alpha)}{\alpha - \lambda} = u(\alpha)$$

for $\lambda \neq \alpha$, and therefore $v \in \mathfrak{X}_{\mu,\delta}$ as well. We also have

$$(\alpha - T)v(\lambda) = (\alpha - \lambda)v(\lambda) + u(\alpha) = u(\lambda),$$

showing that $(\alpha - T_{\mu,\delta})v = u$, and thus $\alpha - T_{\mu,\delta}$ is onto. On the other hand, an equation of the form $(\alpha - T_{\mu,\delta})u = 0$ implies

$$0 = (\alpha - T)u(\alpha) = (\lambda - T)u(\lambda),$$

so that $u(\lambda) = 0$ for $|\lambda| > \|T\|$, and therefore $u = 0$ by analytic continuation. Thus $\alpha - T_{\mu,\delta}$ is one-to-one as well. □

**Lemma 10.2.** *Assume that $T$ has no eigenvalues, and the space $\mathfrak{X}_{\mu,\delta}$ is nonzero for every $\delta > 0$. Then $\sigma_{\ell e}(T_{\mu,\delta}) \cap D(\mu,\delta) \neq \varnothing$ for all $\delta > 0$.*

*Proof.* The assumption that $T$ has no eigenvalues implies that $T_{\mu,\delta}$ does not have any eigenvalues either. Assume, to get a contradiction, that $\sigma_{\ell e}(T_{\mu,\delta}) \cap D(\mu,\delta) = \varnothing$ for some $\delta > 0$, and fix $\eta \in (0,\delta)$. This assumption implies that $\sigma_{\ell e}(T_{\mu,\delta}|\mathfrak{K}) \cap D(\mu,\delta) = \varnothing$ for every invariant subspace $\mathfrak{K}$ of $T_{\mu,\delta}$. Consider in particular the closure $\mathfrak{K}$ of $\mathfrak{X}_{\mu,\eta}$ in $\mathfrak{X}_{\mu,\delta}$. For $\eta < |\lambda - \mu| < \delta$ the operator $(\lambda - T_{\mu,\delta})|\mathfrak{K}$ has dense range (because $\lambda \notin \sigma(T_{\mu,\eta})$), and therefore it is invertible since $\lambda \notin \sigma_{\ell e}(T_{\mu,\delta})$. We conclude that the closed set $\sigma(T_{\mu,\delta}|\mathfrak{K}) \cap D(\mu,\delta)$ is empty because otherwise its boundary points would belong to $\sigma_{\ell e}(T_{\mu,\delta})$. With this preparation, consider an arbitrary element $u \in \mathfrak{X}_{\mu,\eta}$ and define a function $v : \mathbb{C} \to \mathfrak{X}_{\mu,\delta}$ by setting

$$v(\lambda) = \begin{cases} (\lambda - T_{\mu,\eta})^{-1}u & \text{for } \lambda \in \mathbb{C} \setminus \overline{D}(\mu,\eta), \\ (\lambda - T_{\mu,\delta})^{-1}u & \text{for } \lambda \in D(\mu,\delta). \end{cases}$$

The two definitions must agree on $D(\mu,\delta) \setminus \overline{D}(\mu,\eta)$, and the function $v$ is bounded and entire. Thus $v \equiv \lim_{\lambda\to\infty} v(\lambda) = 0$, and this implies that $u = 0$ as well. We conclude that $\mathfrak{X}_{\mu,\eta} = \{0\}$, contrary to the hypothesis. □

The following result is an analogue of Lemma 9.1.

**Lemma 10.3.** *Let $x \in \mathfrak{H}$ have local spectrum contained in $D(\mu,\delta)$ with $|\mu|+|\delta|<1$, and let $u \in \mathfrak{X}_{\mu,\delta}$ satisfy $(\lambda - T)u(\lambda) = x$. We have*

$$\|x \otimes_T \varphi - \varphi(x)e_\nu\| \leq \frac{2\|\varphi\|}{1-|\nu|}\|(\nu - T_{\mu,\delta})u\|_{\mu,\delta}$$

*for every $\varphi \in \mathfrak{H}^*$ and $\nu \in D(\mu,\delta)$.*

*Proof.* Given $f \in H^\infty$, write $f(\lambda) - f(\nu) = (\lambda - \nu)g(\lambda)$ with $g \in H^\infty$, where $\|g\|_\infty \leq 2/(1-|\nu|)$. The usual properties of the Riesz–Dunford functional calculus imply

$$\varphi(f(T)x) - \varphi(x)f(\nu) = \varphi((f(T) - f(\nu))x) = \frac{1}{2\pi i}\int_{|\lambda-\mu|=\delta} g(\lambda)(T-\nu)u(\lambda)\,d\lambda,$$

and the desired estimate follows because $(T-\nu)u(\lambda) = ((T_{\mu,\delta} - \nu)u)(\lambda)$. □

A point $\mu \in \mathbb{D}$ will be said to be in the *localizable left spectrum* $\sigma_\ell^{\text{loc}}(T)$ of $T$ if $\mathfrak{X}_{\mu,\delta} \neq \{0\}$ for every $\delta > 0$. If the space $\mathfrak{X}_{\mu,\delta}$ is infinite dimensional for every $\delta > 0$, we say that $\mu$ is in the *localizable left essential spectrum* $\sigma_{\ell e}^{\text{loc}}(T)$ of $T$. The difference $\sigma_\ell^{\text{loc}}(T) \setminus \sigma_{\ell e}^{\text{loc}}(T)$ consists of isolated eigenvalues of finite multiplicity.

The following result will allow us to make effective use of Lemma 10.3.

**Lemma 10.4.** *Assume that $T$ has no eigenvalues, $\mu \in \sigma_{\ell e}^{\text{loc}}(T)$, $\mathfrak{N} \subset \mathfrak{H}$ is a finite codimensional space, and $\delta > 0$. There exist $\nu \in D(\mu,\delta)$, unit vectors $y_n \in \mathfrak{N}$ and functions $u_n \in \mathfrak{X}_{\mu,\delta}$ such that $y_n = (\lambda - T)u_n(\lambda)$ and $\lim_{n\to\infty}\|(\nu - T_{\mu,\lambda})u_n\| = 0$.*

*Proof.* Lemma 10.2 implies the existence of $\nu \in \sigma_{\ell e}(T_{\mu,\delta}) \cap D(\mu,\delta)$. Select unit vectors $v_n \in \mathfrak{X}_{\mu,\delta}$ such that $\lim_{n\to\infty}\|(\nu - T_{\mu,\delta})v_n\| = 0$. The vectors $x_n = (\lambda - T)v_n(\lambda)$ can be assumed to belong to $\mathfrak{N}$, and a subsequence of $y_n = x_n/\|x_n\|$ will satisfy the requirements of the lemma provided that $\|x_n\|$ does not tend to zero. We have

$$x_n - (\lambda - \nu)v_n(\lambda) = (\nu - T)v_n(\lambda),$$

and $|\lambda - \nu|$ is bounded below on $\partial D(\mu,\delta)$. It follows that

$$\lim_{n\to\infty}\left\|\frac{1}{\lambda-\nu}x_n - v_n(\lambda)\right\| = 0$$

uniformly on $\partial D(\mu,\delta)$, and this cannot happen if $\|x_n\| \to 0$. The lemma follows. □

For the remainder of this section we will denote by $\mathfrak{H}_0$ the linear manifold generated by vectors $x$ with local spectrum contained in some $\overline{D}(\mu,\delta)$ with $|\mu| + |\delta| < 1$. We will consider the map $S : (x,\varphi) \mapsto x \otimes_T \varphi$, whose domain contains $\mathfrak{H}_0 \times \mathfrak{H}^*$. The set $\widetilde{S}_F$ is as defined in Section 2, with $\mathfrak{K}_0 = \mathfrak{H}^*$.

**Lemma 10.5.** *Given distinct points $\lambda_1, \lambda_2, \ldots, \lambda_n \in \sigma_{\ell e}^{\text{loc}}(T) \cap \mathbb{D}$, and positive numbers $\gamma_j$ such that $\sum_{j=1}^n \gamma_j = 1$, the functional $\psi = (1/3)\sum_{j=1}^n \gamma_j e_{\lambda_j}$ belongs to $\widetilde{S}_F$.*

*Proof.* Fix $h_1, h_2, \dots, h_N \in \mathfrak{H}_0$, $\varphi_1, \varphi_2, \dots, \varphi_N \in \mathfrak{H}^*$ and a positive number $\eta$. There is a finite-dimensional space $\mathfrak{M} \subset \mathfrak{H}$ such that $\operatorname{dist}(f(T)h_j, \mathfrak{M}) < \eta$ for every $j$, and every $f \in H^\infty$ with $\|f\|_\infty \le 1$. Find, as before, a finite codimensional space $\mathfrak{N}$ which is $(1/2)$-orthogonal to $\mathfrak{M}$ and $\mathfrak{N} \subset \ker \varphi_j$ for each $j$. Construct inductively vectors $y_j \in \mathfrak{H}$, functions $u_j \in \mathfrak{X}_{\mu_j,\eta}$, points $\nu_j \in D(\mu_j, \eta)$, finite-dimensional spaces $\mathfrak{M} = \mathfrak{M}_1 \subset \cdots \subset \mathfrak{M}_n$, and finite codimensional spaces $\mathfrak{N} = \mathfrak{N}_1 \supset \cdots \supset \mathfrak{N}_n$ such that

1. $y_j$ is a unit vector in $\mathfrak{N}_j$,
2. $(\lambda - T)u_j(\lambda) = y_j$,
3. $\|(\nu_j - T_{\mu_j,\eta})u_j\| < \eta$,
4. $y_i \in \mathfrak{M}_j$ for $i < j$, and
5. $\mathfrak{N}_j$ is $\eta$-orthogonal to $\mathfrak{M}_j$.

This construction is possible by Proposition 8.3 and the preceding lemmas. Theorem 8.2 implies the existence of a unit vector $x = \sum_{j=1}^n x_j$, where $x_j = \alpha_j y_j$, and of a functional of unit norm $\varphi_0 \in \mathfrak{N}^*$ such that $\varphi_0(x_j) = \gamma_j$ for $j = 1, 2, \dots, n$. We extend then $\varphi_0$ to $\varphi \in \mathfrak{H}^*$ such that $\varphi|\mathfrak{M} = 0$ and $\|\varphi\| \le 3$. The various $\eta$-orthogonalities will imply that $\|x_j\| \le 3$ for all $j$ if $\eta$ is sufficiently small. Lemma 10.3 implies then the inequality

$$\|x_j \otimes_T \varphi - \gamma_j e_{\nu_j}\| \le \frac{6\eta}{1 - |\nu_j|},$$

and therefore

$$\left\| x \otimes_T \varphi - \sum_{j=1}^n \gamma_j e_{\mu_j} \right\| \le \sum_{j=1}^n \frac{6\eta}{1 - |\nu_j|} + \sum_{j=1}^n \gamma_j \|e_{\mu_j} - e_{\nu_j}\|.$$

The right-hand side can be made arbitrarily small for an appropriate choice of $\eta$. Another use of Lemma 10.3 yields

$$\|x_j \otimes_T \varphi_k\| \le \frac{6\eta\|\varphi_k\|}{1 - |\nu_j|},$$

and these quantities can be made arbitrarily small. Finally, the choice of $\mathfrak{M}$ and the equality $\varphi|\mathfrak{M} = 0$ allows us to conclude that $\|h_k \otimes_T \varphi\|$ is arbitrarily small, as in the proof of Lemma 9.3. The lemma follows. □

The proof of Proposition 9.4 can now be repeated to yield the following result.

**Proposition 10.6.** *Assume that $\sigma_{\ell e}^{\mathrm{loc}}(T) \cap \mathbb{D}$ is a dominating set. Then $\widetilde{S}_F$ contains the ball of radius $1/48$ centered at $0 \in H^\infty_*$.*

**Corollary 10.7.** *Assume that $\sigma_{\ell}^{\mathrm{loc}}(T) \cap \mathbb{D}$ is a dominating set. Then $T$ has nontrivial invariant subspaces.*

## 11. Notes

### Section 2

It is well known that surjective, continuous bilinear maps need not be open at the origin. Theorem 2.2 appears in [34] in a concrete situation. A second appearance is in [38]. The possibility of solving systems of equations of the form $F(h_i, k_j) = x_{ij}$ was first noted in [27] and [23]. A general study of the open mapping properties of bilinear maps was initiated in [29], and one form of the abstract results in this section appears there.

### Section 3

Theorem 3.3 was proved in [22]. See also [28] for a related result. The use of systems in finding a variety of invariant subspaces is exploited in [27, 23, 14] and [15].

### Section 4

Operators with dominating essential spectrum, also known as (BCP) operators, were introduced in [38], where it is shown that they have nontrivial invariant subspaces. Their reflexivity was proved later in [18]. See also [65] for a related hyper-reflexivity result. A detailed exposition of factorization and invariant subspace results for contractions is given in [26].

The invariant subspace theorem for subnormal operators is deduced in [34] by reducing the problem (using the results of [89]) to pure cyclic subnormal contractions with dominating essential spectrum, and proving the factorization theorem for this extremely restricted class. Factorization was proved for arbitrary subnormal operators in [79]. Later, it was proved [24] that the algebra generated by a subnormal operator has property $(\mathbb{A}_1(1))$. A very brief proof of the existence of invariant subspaces for subnormal operators was given by Thomson [95], who later found a very deep structure theorem for subnormals [96].

For the case of the Bergman shift or, more generally, multiplication operators on Bergman spaces, many of the invariant subspace results obtained in this section were also proved by function theoretical methods. Indeed, the progress in understanding Bergman spaces has been quite remarkable, and we only quote some of the work done in this area [5, 6, 7, 68, 69, 70, 73].

### Section 5

The first result on hyper-reflexivity is due to Arveson [17], who proved that nest algebras have hyper-reflexivity constant 1. Additional examples were provided in [54] and [56], where the free semigroup algebras (with one generator in [54]) are shown to be hyper-reflexive. These results, along with theorem 3.3, were used in [72] to show that all semigroup algebras are hyper-reflexive. The best constant in Theorem 3.3 is not known, but it must be $> 1$. Indeed, the algebras with hyper-reflexivity constant 1 are very special, as shown in [55]. Hyper-reflexivity, as well as reflexivity, makes sense for weak operator closed linear spaces [76], and indeed theorem 3.3 applies in this generality.

### Section 6

The main idea in the proof of Theorem 6.3 appears in [36]. The current abstract form was developed in [19], [30] and [32]. Theorem 6.6 is from [21], where it is used, along with results from [71], to show that the algebra generated by any family of commuting isometries is reflexive. Further applications of these results appear in [85]. Exact factorization theorems in the case of contractions have been pursued, for instance, in [44, 45] and [48].

### Section 7

The existence of invariant subspaces for contractions whose spectra contain the unit circle was proved in [39]; see also [36] for the $C_{00}$ case. Theorem 7.3 was proved in [20] and [46].

### Section 8

Proposition 8.3 is Lemma III.1.1 in [91]. Theorems 8.1 and 8.2 are from [101]. These results proved to be very useful in the study of numerical ranges of Banach space operators, and were not known very well to Hilbert space operator theorists. A particular form of these results was rediscovered in [10], and a Hilbert space formulation was rediscovered in [98]. Eschmeier [58] used these results in their original form.

### Section 9

Banach space extensions of the results of [38] were already pursued in [10]; see also [59] for extensions to domains in higher dimensions. Our approach in this section is closest to that of Ambrozie and Müller [8], where Theorem 9.6 is proved.

### Section 10

Local spectral theory in this context appeared first in [13]. The possibility of applying this technique to hyponormal operators [35] arose because Putinar [86] proved that hyponormal operators are restrictions of generalized scalar operators, as defined in [52]. This yields vectors with small local spectrum. Extensions to Banach space operators were pursued by several authors, but Corollary 10.7 (when formulated for general domains, as in [63]) subsumes these earlier results. Note that this provides yet another proof that subnormal operators have invariant subspaces.

### Extensions to more general domains

Operators on Hilbert space, whose spectra are dominating in a multiply connected domain, were first studied in [51]. This was pursued by many authors in the Hilbert space [81, 49] and Banach space settings [100]. In particular, [100] extends the results of [8] to such domains. There are also substantial advances in the case of $n$-tuples of commuting operators, for which one needs to consider domains in $\mathbb{C}^n$ [59, 60, 61, 57, 64, 74]. A recent result of Ambrozie and Müller [9] is that $n$-tuples of contractions on a Hilbert space, whose Taylor spectra are dominating in the unit polydisk, have nontrivial invariant subspaces. As in the case $n = 1$, the existence of a good functional calculus (polynomial boundedness) must be assumed.

## References

[1] J. Agler, *An invariant subspace theorem*, J. Funct. Anal. 38 (1980), 315–323.

[2] J. Agler and J.E. McCarthy, *Operators that dominate normal operators*, J. Operator Theory 40 (1998), 385–407.

[3] E. Albrecht and B. Chevreau, *Invariant subspaces for $l^p$-operators having Bishop's property ($\beta$) on a large part of their spectrum*, J. Operator Theory 18 (1987), 339–372.

[4] E. Albrecht and J. Eschmeier, *Analytic functional models and local spectral theory*, Proc. London Math. Soc. (3) 75 (1997), 323–348.

[5] A. Aleman, H. Hedenmalm and S. Richter, *Recent progress and open problems in the Bergman space*, Oper. Theory Adv. Appl. 156 (2005), 27–59.

[6] A. Aleman, S. Richter and C. Sundberg, *Beurling's theorem for the Bergman space*, Acta Math. 177 (1996), 275–310.

[7] ———, *Analytic contractions, nontangential limits, and the index of invariant subspaces*, Trans. Amer. Math. Soc. 359 (2007), 3369–3407.

[8] C. Ambrozie and V. Müller, *Invariant subspaces for polynomially bounded operators*, J. Funct. Anal. 213 (2004), 321–345.

[9] ———, *Invariant subspaces for $n$-tuples of contractions with dominating Taylor spectrum*, J. Operator Theory 61 (2009), 63–73.

[10] C. Apostol, *Spectral decompositions and functional calculus*, Rev. Roumaine Math. Pures Appl. 13 (1968), 1481–1528.

[11] ———, *Ultraweakly closed operator algebras*, J. Operator Theory 2 (1979), 49-61.

[12] ———, *Functional calculus and invariant subspaces*, J. Operator Theory 4 (1980), 159–190.

[13] ———, *The spectral flavour of Scott Brown's techniques*, J. Operator Theory 6 (1981), 3–12.

[14] C. Apostol, H. Bercovici, C. Foias and C. Pearcy, *Invariant subspaces, dilation theory, and the structure of the predual of a dual algebra. I*, J. Funct. Anal. 63 (1985), 369–404.

[15] ———, *Invariant subspaces, dilation theory, and the structure of the predual of a dual algebra. II*, Indiana Univ. Math. J. 34 (1985), 845–855.

[16] A. Arias and G. Popescu, *Factorization and reflexivity on Fock spaces*, Integral Equations Operator theory 23 (1995), 268–286.

[17] W.B. Arveson, *Interpolation problems in nest algebras*, J. Funct. Anal. 20 (1975), 208–233.

[18] H. Bercovici, C. Foias, J. Langsam and C. Pearcy, *(BCP)-operators are reflexive*, Michigan Math. J. 29 (1982), 371–379.

[19] H. Bercovici, *Factorization theorems for integrable functions*, Analysis at Urbana. II (ed. E.R. Berkson et al.), Cambridge University Press, Cambridge, 1989.

[20] ———, *Factorization theorems and the structure of operators on Hilbert space*, Ann. of Math. (2) 128 (1988), 399-413.

[21] ———, *A factorization theorem with applications to invariant subspaces and the reflexivity of isometries*, Math. Research Letters, 1 (1994), 511–518.

[22] ———, *Hyper-reflexivity and the factorization of linear functionals*, J. Funct Anal. 158 (1998), 242–252.

[23] H. Bercovici, B. Chevreau, C. Foias and C. Pearcy, *Dilation theory and systems of simultaneous equations in the predual of an operator algebra. II*, Math. Z. 187 (1984), 97–103.

[24] H. Bercovici and J.B. Conway, *A note on the algebra generated by a subnormal operator*, Operator Theory: Advances and Applications 32 (1988), 53–56.

[25] H. Bercovici, C. Foias, J. Langsam and C. Pearcy, *(BCP)-operators are reflexive*, Michigan Math. J. 29 (1982), 371–379.

[26] H. Bercovici, C. Foias. and C. Pearcy, *Dual algebras with applications to invariant subspaces and dilation theory*, CBMS Regional Conf. Ser. in Math., No. 56, Amer. Math. Soc, Providence, R.I., 1985.

[27] ———, *Dilation theory and systems of simultaneous equations in the predual of an operator algebra. I*, Michigan Math. J. 30 (1983), 335–354.

[28] ———, *On the reflexivity of algebras and linear spaces of operators*, Michigan Math. J. 33 (1986), 119–126.

[29] ———, *Two Banach space methods and dual operator algebras*, J. Funct. Anal. 78 (1988), 306–345.

[30] H. Bercovici and W.S. Li, *A near-factorization theorem for integrable functions*, Integral Equations Operator Theory 17 (1993), 440–442.

[31] H. Bercovici, V. Paulsen and C. Hernandez, *Universal compressions of representations of $H^\infty(G)$*, Math. Ann. 281 (1988), 177–191.

[32] H. Bercovici and B. Prunaru, *An improved factorization theorem with applications to subnormal operators*, Acta Sci. Math. (Szeged) 63 (1997), 647–655.

[33] L. Brown, A. Shields and K. Zeller, *On absolutely convergent exponential sums*, Trans. Amer. Math. Soc. 96 (1960), 162–183.

[34] S.W. Brown, *Some invariant subspaces for subnormal operators*, Integral Equations Operator Theory 1 (1978), 310–333.

[35] ———, *Hyponormal operators with thick spectra have invariant subspaces*, Ann. of Math. (2) 125 (1987), 93–103.

[36] ———, *Contractions with spectral boundary*, Integral Equations Operator Theory 11 (1988), 49–63.

[37] S.W. Brown and B. Chevreau, *Toute contraction à calcul fonctionnel isométrique est réflexive*, C. R. Acad. Sci. Paris Sér. I Math. 307 (1988), 185–188.

[38] S.W. Brown, B. Chevreau and C. Pearcy, *Contractions with rich spectrum have invariant subspaces*, J. Operator Theory 1 (1979), 123–136.

[39] ———, *On the structure of contraction operators. II*, J. Funct. Anal. 76 (1988), 30–55.

[40] S.W. Brown and E. Ko, *Operators of Putinar type*, Oper. Theory Adv. Appl. 104 (1998), 49–57.

[41] G. Cassier, *Un exemple d'opérateur pour lequel les topologies faible et ultrafaible ne coïncident pas sur l'algèbre duale*, J. Operator Theory 16 (1986), 325–333.

[42] ———, *Algèbres duales uniformes d'opérateurs sur l'espace de Hilbert*, Studia Math. 95 (1989), 17–32.

[43] G. Cassier, I. Chalendar and B. Chevreau, *A mapping theorem for the boundary set $X_T$ of an absolutely continuous contraction $T$*, J. Operator Theory 50 (2003), 331–343.

[44] I. Chalendar and J. Esterle, *$L^1$-factorization for $C_{00}$-contractions with isometric functional calculus*, J. Funct. Anal. 154 (1998), 174–194.

[45] I. Chalendar, J.R. Partington and R.C. Smith, *$L^1$ factorizations, moment problems and invariant subspaces*, Studia Math. 167 (2005), 183–194.

[46] B. Chevreau, *Sur les contractions à calcul fonctionnel isométrique. II*, J. Operator Theory 20 (1988), 269–293.

[47] B. Chevreau, G. Exner and C. Pearcy, *On the structure of contraction operators. III*, Michigan Math. J. 36 (1989), 29–62.

[48] ———, *Boundary sets for a contraction*, J. Operator Theory 34 (1995), 347–380.

[49] B. Chevreau and W.S. Li, *On certain representations of $H^\infty(G)$ and the reflexivity of associated operator algebras*, J. Funct. Anal. 128 (1995), 341–373.

[50] B. Chevreau and C. Pearcy, *On the structure of contraction operators. I*, J. Funct. Anal. 76 (1988), 1–29.

[51] ———, *On Sheung's theorem in the theory of dual operator algebras*, Oper. Theory Adv. Appl. 28 (1988), 43–49.

[52] I. Colojoară and C. Foiaş, *Theory of generalized spectral operators*, Gordon and Breach, New York, 1968.

[53] J.B. Conway and M. Ptak, *The harmonic functional calculus and hyperreflexivity*, Pacific J. Math. 204 (2002), 19–29.

[54] K.R. Davidson, *The distance to the analytic Toeplitz operators*, Illinois J. Math. 31 (1987), 265–273.

[55] K.R. Davidson and R. Levene, *1-hyperreflexivity and complete hyperreflexivity*, J. Funct. Anal. 235 (2006), 666–701.

[56] K.R. Davidson and D.R. Pitts, *Invariant subspaces and hyper-reflexivity for free semigroup algebras*, Proc. London Math. Soc. (3) 78 (1999), 401–430.

[57] M. Didas, *Invariant subspaces for commuting pairs with normal boundary dilation and dominating Taylor spectrum*, J. Operator Theory 54 (2005), 169–187.

[58] J. Eschmeier, *Operators with rich invariant subspace lattices*, J. Reine Angew. Math. 396 (1989), 41–69.

[59] ———, *Representations of $H^\infty(G)$ and invariant subspaces*, Math. Ann. 298 (1994), 167–186.

[60] ———, *Algebras of subnormal operators on the unit ball*, J. Operator Theory 42 (1999), 37–76.

[61] ———, *On the reflexivity of multivariable isometries*, Proc. Amer. Math. Soc. 134 (2006), 1783–1789

[62] J. Eschmeier and B. Prunaru, *Invariant subspaces for operators with Bishop's property $(\beta)$ and thick spectrum*, J. Funct. Anal. 94 (1990), 196–222.

[63] ———, *Invariant subspaces and localizable spectrum*, Integral Equations Operator Theory 42 (2002), 461–471.

[64] G. Exner, Y.S. Jo and I.B. Jung, *Representations of $H^\infty(\mathbb{D}^N)$*, J. Operator Theory 45 (2001), 233–249.

[65] D.W. Hadwin, *Compressions, graphs, and hyperreflexivity*, J. Funct. Anal. 145 (1997), 1–23.

[66] P.R. Halmos, *Normal dilations and extensions of operators*, Summa Brasil. Math. 2 (1950). 125–134.

[67] ———,*Ten problems in Hilbert space*, Bull. Amer. Math. Soc. 76 (1970), 887–933.

[68] H. Hedenmalm, *An invariant subspace of the Bergman space having the codimension two property*, J. Reine Angew. Math. 443 (1993), 1–9.

[69] H. Hedenmalm, B. Korenblum and K. Zhu, *Beurling type invariant subspaces of the Bergman spaces*, J. London Math. Soc. (2) 53 (1996), 601–614.

[70] H. Hedenmalm, S. Richter and K. Seip, *Interpolating sequences and invariant subspaces of given index in the Bergman spaces*, J. Reine Angew. Math. 477 (1996), 13–30.

[71] K. Horák, and V. Müller, *On commuting isometries*, Czechoslovak Math. J. 43(118) (1993), 373–382.

[72] F. Jaëck and S.C. Power, *Hyper-reflexivity of free semigroupoid algebras*, Proc. Amer. Math. Soc. 134 (2006), 2027–2035.

[73] B. Korenblum and K. Zhu, *Complemented invariant subspaces in Bergman spaces*, J. London Math. Soc. (2) 71 (2005), 467–480.

[74] M. Kosiek and A. Octavio, *On common invariant subspaces for commuting contractions with rich spectrum*, Indiana Univ. Math. J. 53 (2004), 823–844.

[75] W.S. Li, *On polynomially bounded operators. I*, Houston J. Math. 18 (1992), 73–96.

[76] A.I. Loginov and V.S. Shulman, *Hereditary and intermediate reflexivity of $W^*$-algebras*, Akad. Nauk SSSR Ser. Mat. 39 (1975), 1260–1273.

[77] M. Marsalli, *Systems of equations in the predual of a von Neumann algebra*, Proc. Amer. Math. Soc. 111 (1991), 517–522.

[78] V. Müller and M. Ptak, *Hyperreflexivity of finite-dimensional subspaces*, J. Funct. Anal. 218 (2005), 395–408.

[79] R.F. Olin and J.E. Thomson, *Algebras of subnormal operators*, J. Funct. Anal. 37 (1980), 271–301.

[80] ———, *Algebras generated by a subnormal operator*, Trans. Amer. Math. Soc. 271 (1982), 299–311.

[81] V. Pata, K.X. Zheng and A. Zucchi, *On the reflexivity of operator algebras with isometric functional calculus*, J. London Math. Soc. (2) 61 (2000), 604–616.

[82] ———, *Cellular-indecomposable subnormal operators. III*, Integral Equations Operator Theory 29 (1997), 116–121.

[83] G. Popescu, *A generalization of Beurling's theorem and a class of reflexive algebras*, J. Operator Theory 41 (1999), 391–420.

[84] B. Prunaru, *A structure theorem for singly generated dual uniform algebras*, Indiana Univ. Math. J. 43 (1994), 729–736.

[85] ———, *Approximate factorization in generalized Hardy spaces*, Integral Equations Operator Theory 61 (2008), 121–145.

[86] M. Putinar, *Hyponormal operators are subscalar*, J. Operator Theory 12 (1984), 385–395.

[87] G.F. Robel, *On the structure of* $(BCP)$*-operators and related algebras, I*, J. Operator Theory 12 (1984), 23–45.

[88] D. Sarason, *Invariant subspaces and unstarred operator algebras*, Pacific J. Math. 17 (1966), 511–517.

[89] ———, *Weak-star density of polynomials*, J. Reine Angew. Math. 252 (1972), 1–15.

[90] J. Sheung, *On the preduals of certain operator algebras*, Ph.D. Thesis, Univ. Hawaii, 1983.

[91] I. Singer, *Bases in Banach spaces. II.* Editura Academiei, Bucharest, 1981.

[92] B. Sz.-Nagy and C. Foias, *Harmonic analysis of operators on Hilbert space*, North-Holland, Amsterdam, 1970.

[93] J.G. Stampfli, *An extension of Scott Brown's invariant subspace theorem:* $K$*-spectral sets*, J. Operator Theory 3 (1980), 3–21.

[94] P. Sullivan, *Dilations and subnormal operators with rich spectrum*, J. Operator Theory 29 (1993), 29–42.

[95] J.E. Thomson, *Invariant subspaces for algebras of subnormal operators*, Proc. Amer. Math. Soc. 96 (1986), 462–464.

[96] ———, *Approximation in the mean by polynomials*, Ann. of Math. (2) 133 (1991), 477–507.

[97] J. Wermer, *On invariant subspaces of normal operators*, Proc. Amer. Math. Soc. 3 (1952), 270–277.

[98] D.J. Westwood, *On* $C_{00}$*-contractions with dominating spectrum*, J. Funct. Anal. 66 (1986), 96–104.

[99] ———, *Weak operator and weak*$^*$ *topologies on singly generated algebras*, J. Operator Theory 15 (1986), 267–280.

[100] O. Yavuz, *Invariant subspaces for Banach space operators with a multiply connected spectrum*, Integral Equations Operator Theory 58 (2007), 433–446.

[101] C. Zenger, *On convexity properties of the Bauer field of values of a matrix*, Numer. Math. 12 (1968), 96–105.

[102] A. Zygmund, *Trigonometric series*, 2nd ed., Cambridge University Press, Cambridge, 1968.

Hari Bercovici<br>
Department of Mathematics
Indiana University
Bloomington, IN 47405, USA
e-mail: bercovic@indiana.edu

Operator Theory:
Advances and Applications, Vol. 207, 177–194

# The State of Subnormal Operators

John B. Conway and Nathan S. Feldman

**Abstract.** In this paper we present the highlights of the theory of subnormal operators that was initiated by Paul Halmos in 1950. This culminates in Thomson's Theorem on bounded point evaluations where several applications are presented. Throughout the paper are some open problems.

**Mathematics Subject Classification (2000).** Primary 47B20; Secondary 47A15.

**Keywords.** Subnormal operator, hyponormal operator, hypercyclic operator, Thomson's Theorem, trace estimate, square root, invariant subspace, finite rank self-commutator.

## 1. Introduction

A prominent feature of Paul Halmos's mathematical life was his influence on other mathematicians. Many know of his students and their robust mathematical careers, but another aspect was his ability to pull from an example or an isolated result just the right question or concept and by so doing tweak and excite the curiosity of others, thus launching a series of papers that frequently started a new area of investigation. Though by no means a unique example of this, the theory of subnormal operators is a prime instance of this phenomenon. A *subnormal operator* on a Hilbert space $\mathcal{H}$ is one such that there is a larger Hilbert space $\mathcal{K}$ that contains $\mathcal{H}$ and a normal operator $N$ on $\mathcal{K}$ such that: (a) $N\mathcal{H} \subseteq \mathcal{H}$; (b) $Nh = Sh$ for all $h$ in $\mathcal{H}$.

In 1950 Paul Halmos introduced this concept at the same time that he defined hyponormal operators [29]. His inspiration was an examination of the unilateral shift, from which he abstracted two properties, one analytic (having a normal extension) and one algebraic ($S^*S \geq SS^*$). One must marvel at his focusing on these two properties, each of which is satisfied by an enormous collection of examples, even though he himself did not realize at the time how extensive these theories were and how rich the class of examples. This constitutes a high compliment to his great sense of mathematical taste and his instinct to isolate what must be a good idea.

In his paper he gave a characterization of subnormal operators in terms of a positivity condition and explored some basic facts. Though Paul wrote other papers where he discussed particular subnormal and hyponormal operators, he never again wrote one devoted to either topic, even though he had PhD students who did.

The first of Paul's students to write on subnormal operators was Joseph Bram [7]; in fact this was the first paper devoted solely to subnormal operators. In this work Bram improved Halmos's characterization of subnormality and established many basic properties of the operators. In fact the paper is still worth reading and contains nuggets that have attracted attention and been more thoroughly explored as well as a few that deserve a second look. One particular fundamental result that will be used repeatedly in this paper is the characterization of cyclic subnormal operators. Let $\mu$ be a compactly supported regular Borel measure on the complex plane, $\mathbb{C}$, and define $P^2(\mu)$ to be the closure in $L^2(\mu)$ of the analytic polynomials. That is, $P^2(\mu)$ is the closed linear span of $\{z^n : n \geq 0\}$. Now define the operator $S_\mu$ on $P^2(\mu)$ by

$$S_\mu f(z) = zf(z).$$

This is a subnormal operator since the operator $N_\mu$ defined on $L^2(\mu)$ by

$$N_\mu f(z) = zf(z)$$

is a normal extension. Also note that $S_\mu$ is *cyclic*; that is, there is a vector $e$ such that $P^2(\mu) = \bigvee\{S_\mu^n e : n \geq 0\}$. In this case we can take for $e$ the constant function 1. Bram showed that if $S$ is any subnormal operator on $\mathcal{H}$ that has a cyclic vector $e$, then there is a measure $\mu$ and a unitary $U : \mathcal{H} \to P^2(\mu)$ such that $U^* S_\mu U = S$ and $Ue = 1$. The proof of this is not difficult (see [14], page 51), but it sets the stage for an intimate interaction between the theory of analytic functions and subnormal operators. The polynomials are of course analytic and it was observed that in every example of such a measure $\mu$ the space $P^2(\mu)$ consisted of the direct sum of an $L^2$-space and a Hilbert space of analytic functions. In fact it was later established (see §5 below) that this is always the case.

Another of Paul's students, Errett Bishop, proved a basic fact about subnormal operators as a consequence of a general theory he developed in his thesis [6]: the set of subnormal operators is closed in the strong operator topology. (The *strong operator topology*, or SOT, on $\mathcal{B}(\mathcal{H})$ is the one defined by the seminorms $p_h(T) = \|Th\|$ for all $h$ in $\mathcal{H}$.) Using the Halmos-Bram criterion for subnormality, it is rather straightforward to show that the SOT limit of a sequence of subnormal operators is again subnormal. A crucial fact in showing this is that an SOT convergent sequence of operators is norm bounded. However the SOT is not metrizable so that it is necessary to consider the limits of nets. In this case the net may not be bounded, complicating the argument. A direct proof, avoiding the machinery that Bishop developed in his thesis, was found much later in [16]; where it was shown that an operator is subnormal if and only if it is an SOT limit of a sequence of normal operators.

These three papers are the entirety of the direct contributions of Halmos and his students to subnormal operators. This seems all the more remarkable given that the subject has grown to the extent that it has.

## 2. Fundamentals of subnormal operators

In this section we will review some basic facts concerning subnormal operators. No proofs will be give, but rather the results will be cited from [13] or [14] where the proofs can be found.

All Hilbert spaces are separable and over the complex numbers and $\mathcal{B}(\mathcal{H})$ denotes the algebra of all bounded operators from $\mathcal{H}$ into itself. It is well known that $\mathcal{B}(\mathcal{H})$ is the Banach space dual of $\mathcal{B}_1(\mathcal{H})$, the space of trace-class operators furnished with the trace norm. (See [14], page 8.) Here for every $B$ in $\mathcal{B}(\mathcal{H})$ the corresponding element $F_B$ of $\mathcal{B}_1(\mathcal{H})^*$ is given by

$$F_B(A) = \operatorname{tr}(AB)$$

for every trace class operator $A$. As a consequence of this, $\mathcal{B}(\mathcal{H})$ has a natural weak$^*$ topology. We mention that a similar pairing results in the fact that $\mathcal{B}_1(\mathcal{H})$ is the Banach space dual of the algebra of compact operators on $\mathcal{H}$.

Define a *dual algebra* to be any Banach algebra with identity that is the dual of a Banach space. So $\mathcal{B}(\mathcal{H})$ is a dual algebra as is any weak$^*$ closed subalgebra of $\mathcal{B}(\mathcal{H})$ that contains the identity operator. In particular, every von Neumann algebra is a dual algebra. Also note that for any $\sigma$-finite measure space its $L^\infty$ space is also a dual algebra. If $\mathcal{A}$ and $\mathcal{B}$ are two dual algebras, call a homomorphism $\rho : \mathcal{A} \to \mathcal{B}$ a *dual algebra homomorphism* if it is continuous when both $\mathcal{A}$ and $\mathcal{B}$ have their weak$^*$ topologies. Similarly we define a *dual algebra isomorphism.*

If $S$ is a subnormal operator on $\mathcal{H}$ and $N$ is a normal extension operating on $\mathcal{K}$, then for any other normal operator $M$ on a space $\mathcal{L}$ the normal operator $N \oplus M$ is also a normal extension. So the normal extension fails to be unique. However it is possible to obtain a minimal extension and this is unique up to unitary equivalence. In fact, define $N$ to be a *minimal normal extension* of $S$ provided there is no proper reducing subspace for $N$ that contains $\mathcal{H}$. With this definition it is not hard to show the following result.

**2.1. Proposition.** *If $S$ is a subnormal operator on $\mathcal{H}$ and $N$ is a normal extension on $\mathcal{K}$, then $N$ is a minimal normal extension if and only if $\mathcal{K}$ is the closed linear span of*

**2.2** $$\{N^{*k}h : h \in \mathcal{H} \text{ and } k \geq 0\}.$$

*Moreover any two minimal normal extensions are unitarily equivalent.*

For a proof see [14], pages 38–39. In light of the uniqueness part of the preceding result we are justified in speaking of *the* minimal normal extension.

This result implies that for a compactly supported measure $\mu$ on $\mathbb{C}$, $N_\mu$ is the minimal normal extension of $S_\mu$. In fact the set appearing in (2.2) contains $\{\bar{z}^k z^n : k, n \geq 0\}$, whose linear span is dense in $L^2(\mu)$ by the Stone-Weierstrass Theorem.

Another basic notion is that of purity. A subnormal operator $S$ is *pure* provided there is no reducing subspace $\mathcal{M}$ for $S$ such that $S|\mathcal{M}$ is normal. For example, if $m$ is normalized arc length measure on the unit circle, then $P^2(m)$ is the classical Hardy space $H^2$ and $S_m$ is the unilateral shift, which is pure. On the other hand if $\mu = m + \delta_2$, where $\delta_2$ is the unit point mass at 2, then $P^2(\mu) = H^2 \oplus \mathbb{C}$ and $S_\mu$ is not pure. Of course for certain measures $\mu$, such as Lebesgue measure on the unit interval, $P^2(\mu) = L^2(\mu)$ and $S_\mu$ is the antithesis of pure.

**2.3. Proposition.** *If $S$ is a subnormal operator, then there is a reducing subspace $\mathcal{H}_0$ for $S$ such that $S|\mathcal{H}_0$ is normal and $S|\mathcal{H}_0^\perp$ is pure.*

For the proof see [14], page 38. This allows us to split off the pure part of any subnormal operator, treat it separately, then reassemble to get the original operator and see if what was proved for the pure part can be extended to the original subnormal operator. The persistent reader will see that this frequently is the approach used in obtaining the major results on subnormal operators.

The idea here is that the presence of a non-trivial normal summand in a subnormal operator introduces too much symmetry. Under most situations in mathematics we enjoy it when symmetry is present, but with subnormal operators the presence of a normal summand introduces a complication when we explore their structure.

If $N$ is a normal operator on $\mathcal{K}$ and $N = \int z \, dE(z)$ is its spectral decomposition given by the Spectral Theorem ([13], page 263), then there is a scalar-valued measure $\mu$ that has the same sets of measure zero as the spectral measure $E$. One possibility is to take $\mu(\Delta) = \sum_n 2^{-n}\langle E(\Delta)h_n, h_n\rangle$, where $\{h_n\}$ is a countable dense subset of the unit ball of $\mathcal{K}$. This measure is far from unique since any measure that is mutually absolutely continuous with it can also serve. Nevertheless we will refer to the scalar-valued spectral measure for $N$.

The role of the scalar-valued spectral measure for a normal operator is to correctly delineate its functional calculus. If $\mu$ is the scalar-valued spectral measure for $N$, then for every $\phi$ in $L^\infty(\mu)$ we can define the operator $\phi(N) = \int \phi \, dE$ and we have the following result.

**2.4. Theorem.** *The map $\rho : L^\infty(\mu) \to \mathcal{B}(\mathcal{K})$ defined by $\rho(\phi) = \phi(N)$ is an isometric *-isomorphism of $L^\infty(\mu)$ onto $W^*(N)$, the von Neumann algebra generated by $N$, and it is also a dual-algebra isomorphism.*

See [13], page 288.

If $S$ is a subnormal operator, the *scalar-valued spectral measure* for $S$ is any scalar-valued spectral measure for its minimal normal extension. In the next section we will see how to develop a functional calculus for a subnormal operator that extends the Riesz functional calculus and is important in obtaining some of the deeper results for subnormal operators.

We end this section by introducing a very important collection of examples of subnormal operators, the Bergman operators. If $G$ is a bounded open set in $\mathbb{C}$, denote by $L^2_a(G)$ the Hilbert space of all analytic functions defined on $G$ that are square integrable with respect to area measure on $G$. If the operator $S$ is defined on $L^2_a(G)$ as multiplication by the independent variable, then $S$ is subnormal since it has the normal extension $N_\mu$, where $\mu = \text{Area}\,|G$. In fact this is easily seen to be the minimal normal extension of $S$. This operator is called the *Bergman operator* for $G$. When $G = \mathbb{D}$ this is the *Bergman shift.* These operators constitute one of the most important collections of examples of subnormal operators. Understanding them would lead us far in understanding subnormal operators, not that results about Bergman operators can be used to derive results about the arbitrary subnormal operator but rather the difficulties encountered in understanding Bergman operators seem to parallel the difficulties in understanding arbitrary subnormal operators.

A rudimentary exploration of Bergman operators can be found in various parts of [14] but more comprehensive treatments are in [21] and [32].

## 3. The functional calculus

For any operator $S$ and a function $f$ analytic in a neighborhood of the spectrum of $S$, it is well known that the Riesz functional calculus defines an operator $f(S)$. (See [13] or any reference in functional analysis.) If $\mathrm{Hol}(\sigma(S))$ denotes the algebra of all functions that are analytic in a neighborhood of $\sigma(S)$, the spectrum of $S$, then a standard result is that the map $f \to f(S)$ defines an algebraic homomorphism into $\mathcal{B}(\mathcal{H})$ and the Spectral Mapping Theorem is valid: $\sigma(f(S)) = f(\sigma(S))$. Using this functional calculus it is possible to derive results about general operators, though, as is to be expected, at this level of generality what can be learned is limited.

When $S$ is a subnormal operator there is the bonus that each $f(S)$ is also subnormal and, hence, its norm equals its spectral radius. Thus the Riesz functional calculus becomes an isometric isomorphism. Note that though a hyponormal operator has its norm equal to its spectral radius, it does not follow that $f(S)$ is hyponormal even though $S$ may be. Thus the functional calculus marks a fork in the road where the development of the two types of operators diverge. Indeed it is the functional calculus that facilitates a deep study of subnormal operators, including the proof of the existence of a rich collection of invariant subspaces.

There is another way to define this functional calculus, which, by the uniqueness of the Riesz functional calculus, is equivalent for functions in $\mathrm{Hol}(\sigma(S))$, but which allows us to use a larger class of functions. For the remainder of this section let $S$ be the subnormal operator acting on $\mathcal{H}$, and let $N$ and $\mu$ be its minimal normal extension and scalar-valued spectral measure. Recall Theorem 2.4 and let $\mathcal{R}(S) = \{\phi \in L^\infty(\mu) : \phi(N)\mathcal{H} \subseteq \mathcal{H}\}$. This set of functions is called the *restriction algebra* for $S$. It is easy to prove the following. (See [14], page 85.)

**3.1. Proposition.** *If $S, N, \mu$ and $\mathcal{R}(S)$ are as above, then the following hold.*

(a) *$\mathcal{R}(S)$ is a weak* closed subalgebra of $L^\infty(\mu)$.*

(b) *If $\rho : \mathcal{R}(S) \to \mathcal{B}(\mathcal{H})$ is defined by $\rho(\phi) = \phi(N)|\mathcal{H}$, then $\rho$ is an isometric dual algebra isomorphism onto its image.*

This enables us to legitimately define $\phi(S) \equiv \phi(N)|\mathcal{H}$ and have a functional calculus that extends the Riesz functional calculus. At the level of complete generality, identifying the range of this isomorphism remains an unsolved problem, though it is clearly a subalgebra of the commutant of $S$. In some cases, it is easy.

**3.2. Proposition.** *If $S$ is the cyclic subnormal operator $S_\mu$, then $\mathcal{R}(S_\mu) = P^2(\mu) \cap L^\infty(\mu)$ and $\{\phi(S_\mu) : \phi \in P^2(\mu) \cap L^\infty(\mu)\}$ is the commutant of $S_\mu$.*

For a proof see [14], page 86. Using the work of James Thomson on bounded point evaluations (see Theorem 5.1 below) it is possible to characterize, in a limited fashion, the functions in $P^2(\mu) \cap L^\infty(\mu)$ and this enables the solution of many problems regarding cyclic subnormal operators.

There is a smaller class of functions than the restriction algebra whose structure we understand and which we can use to derive many results for subnormal operators, including the existence of invariant subspaces. We consider $P^\infty(\mu)$, the weak* closed subalgebra of $L^\infty(\mu)$ generated by the analytic polynomials. It is clear that $P^\infty(\mu) \subseteq \mathcal{R}(S)$ so it makes sense to examine the functional calculus $\phi \to \phi(S)$ defined for $\phi$ in $P^\infty(\mu)$. This was done extensively in [17], where the principal tool was the work of Don Sarason [39] characterizing $P^\infty(\mu)$ as the algebra of bounded analytic functions on a certain open set. A special case of Sarason's Theorem is the following.

**3.3. Sarason's Theorem.** *If $S$ is a pure subnormal operator with scalar-valued spectral measure $\mu$ and $K_\mu$ is the closure of the set of $\lambda$ in the plane such that $p \to p(\lambda)$ extends to a weak* continuous linear functional on $P^\infty(\mu)$; then $K_\mu$ contains the support of $\mu$ and the identity map on polynomials extends to an isometric dual algebra isomorphism of $P^\infty(\mu)$ onto $H^\infty(\operatorname{int} K_\mu)$.*

A proof of Sarason's Theorem can also be found in [14], page 301. This enables us to refine and make more precise Proposition 3.1 in the case of $P^\infty(\mu)$.

**3.4. Proposition.** ([17]) *If $S_\mu$ is pure and $K_\mu$ is as above, then there is an isometric dual algebra isomorphism $\phi \to \phi(S_\mu)$ of $H^\infty(\operatorname{int} K_\mu)$ onto $P^\infty(S_\mu)$.*

Also see [14], page 86 for a proof.

With Sarason's Theorem in mind we see that the functional calculus for subnormal operators shows that there is an intimate connection with analytic function theory. This relationship will become even more intimate when we state Thomson's Theorem on bounded point evaluations. Using Sarason's Theorem and the functional calculus it is possible to derive several structure theorems for subnormal operators. Here are two.

**3.5. Theorem.** [17] *If $S$ is a pure subnormal operator with scalar-valued spectral measure $\mu$ and $\phi \in H^\infty(\operatorname{int} K_\mu)$ such that $\phi$ is not constant on any component of $\operatorname{int} K_\mu$, then the minimal normal extension of $\phi(S)$ is $\phi(N)$.*

A proof of this result can also be found in [14], page 317.

**3.6. Theorem.** [17] *A Spectral Mapping Theorem is valid for this functional calculus: if $S$ is a pure subnormal operator with scalar-valued spectral measure $\mu$ and $\phi \in H^\infty(\operatorname{int} K_\mu)$, then $\sigma(\phi(S)) = \operatorname{cl}[\phi(\sigma(S) \cap \operatorname{int} K_\mu)]$.*

A proof of this result can also be found in [14], page 326.

Note that several variations on this theme are possible. For example, these results can be extended to the weak* closure in $L^\infty(\mu)$ of the rational functions with poles off $\sigma(S)$ and where the essential spectrum replaces the spectrum of $S$. These are due to Dudziak[20] and can also be found in [14].

## 4. Invariant subspaces

Recall that if $S$ is an operator on the Hilbert space $\mathcal{H}$ and $\mathcal{M}$ is a closed subspace of $\mathcal{H}$, then $\mathcal{M}$ is said to be an *invariant subspace* for $S$ if $S\mathcal{M} \subseteq \mathcal{M}$. Let $\operatorname{Lat} S$ denote the collection of all invariant subspaces for $S$; $\mathcal{M}$ is *non-trivial* if $\mathcal{M} \neq (0)$ and $\mathcal{M} \neq \mathcal{H}$.

**4.1. Theorem.** (Brown [8]) *Every subnormal operator has a non-trivial invariant subspace.*

Scott Brown proved this theorem by using what has since been called "The Scott Brown" technique, a method of factoring certain weak* continuous linear functionals. The strategy is as follows. To prove the existence of invariant subspaces it suffices to assume the subnormal operator is cyclic; thus we may assume that $S = S_\mu$ operating on $P^2(\mu)$ for some compactly supported measure $\mu$ on the complex plane. To avoid trivialities, we can assume that $S_\mu$ is pure. From the preceding section $P^\infty(S_\mu)$ is precisely $\{M_\phi : \phi \in P^\infty(\mu)\}$. Let $\lambda \in \operatorname{int} K_\mu$ so that $p \to p(\lambda)$ defined on the set of polynomials extends to a weak* continuous linear functional $\phi \to \phi(\lambda)$ defined on $P^\infty(\mu)$. Assuming that $S_\mu$ has no invariant subspaces, which implies certain spectral conditions such as $\sigma(S_\mu) = \sigma_{ap}(S_\mu)$, Brown then shows that there are functions $f, g$ in $P^2(\mu)$ such that $\phi(\lambda) = \int \phi f \bar{g} d\mu$ for all $\phi$ in $P^\infty(\mu)$. (This step is the hard one and the first use of the Scott Brown technique.) Now let $\mathcal{M}$ be the closure in $P^2(\mu)$ of $\{(z-\lambda)pf : p \text{ is a polynomial}\}$; that is, $\mathcal{M}$ is the closed linear span of $\{(S_\mu - \lambda){S_\mu}^n f : n \geq 0\}$. Clearly $\mathcal{M}$ is invariant for $S_\mu$; it remains to show that $\mathcal{M}$ is non-trivial. First observe that $(z-\lambda)f \in \mathcal{M}$ and this function is not 0. In fact if $0 = (z-\lambda)f = (S_\mu - \lambda)f$, then $\lambda$ is an eigenvalue for $S_\mu$ and it follows that the space $\mathbb{C}f$ is a one-dimensional reducing subspace for $S$. (See [14], Proposition II.4.4.) This violates the purity of $S_\mu$. Hence $\mathcal{M} \neq (0)$. Second it follows by the factorization that $g \perp \mathcal{M}$ and $g \neq 0$, so $\mathcal{M} \neq P^2(\mu)$. Therefore $\mathcal{M}$ is the sought after non-trivial invariant subspace.

Thomson [41] discovered a proof of this factorization in the context of $P^2(\mu)$ by using additional function theory and this produced a far shorter proof of the invariant subspace result. Thankfully this occurred several years after Brown's result and allowed time for the Scott Brown technique to be studied for its own sake. A whole sequence of papers appeared devoted to the factorization of weak* continuous linear functionals on dual algebras. These factorization results were extended to collections of linear functionals and then applied to other classes of operators.

Brown himself used this technique to prove that certain hyponormal operators have invariant subspaces. Recall that for a compact subset of the plane, $C(K)$ denotes the algebra of all continuous functions from $K$ into $\mathbb{C}$ while $R(K)$ is the closure in $C(K)$ of all rational functions with poles off $K$.

**4.2. Theorem.** (Brown [9]) *If $S$ is a hyponormal operator and $R(\sigma(S)) \neq C(\sigma(S))$, then $S$ has a non-trivial invariant subspace.*

In particular it follows that when $S$ is hyponormal and its spectrum has non-empty interior, then $S$ has a non-trivial invariant subspace.

In a far reaching development, Apostol, Bercovici, Foias, and Pearcy [3] carried the technique further and applied it to a wide collection of contractions. An excellent place to see a self-contained treatment of the Scott Brown technique and its extensions is the lecture notes [4]. Though there is never a linear order of generality in these matters, a candidate for the most far reaching result that was proved using the Scott Brown technique and that applies to non-hyponormal operators is the result of Brown, Chevreau, and Pearcy [10] that any contraction whose spectrum contains the unit circle has a non-trivial invariant subspace.

A surprising development in this circle of ideas was the 1985 result in [3] that the lattice of invariant subspaces of the Bergman shift is as complicated as can be imagined. A general result of this paper is that under certain assumptions on a contraction $T$ and certain scalars $\lambda$ in the spectrum of $T$, there are invariant subspaces $\mathcal{M}$ and $\mathcal{N}$ of $T$ such that $\mathcal{N} \subseteq \mathcal{M}$, $\mathcal{M} \ominus \mathcal{N}$ is infinite dimensional, and the compression of $T$ to the orthogonal difference $\mathcal{M} \ominus \mathcal{N}$ is $\lambda$ times the identity operator. Note that this means that if $\mathcal{L}$ is any closed linear subspace with $\mathcal{N} \subseteq \mathcal{L} \subseteq \mathcal{M}$ then $\mathcal{L} \in \operatorname{Lat} T$. Thus in $\operatorname{Lat} T$ there are copies of the lattice of all subspaces of an infinite-dimensional Hilbert space. The remarkable thing is that it can be verified that the Bergman shift satisfies the required conditions with the scalar $\lambda = 0$.

Thus if $S$ is the Bergman shift operating as multiplication by $z$ on $L^2_a(\mathbb{D})$, then there are invariant subspaces $\mathcal{M}$ for $S$ such that $\mathcal{M} \ominus z\mathcal{M}$ has any dimension possible. So from an abstract result came a fact about function theory not previously known. In fact it remained the case that there were no known examples of such invariant subspaces until 1993 when Hedenmalm [31] found an example of an invariant subspace for $S$ with the property that $\mathcal{M} \ominus z\mathcal{M}$ has dimension 2. Earlier, in 1991, Hedenmalm [30] proved the existence of contractive zero divisors within the Bergman space $L^2_a(\mathbb{D})$ for any Bergman space zero-set. That is, if $Z$ is a zero set for the Bergman space, then Hedenmalm proved that there exists a function $G$ in $L^2_a(\mathbb{D})$ such that for any function $f$ in $L^2_a(\mathbb{D})$ with $f|Z = 0$, we have $f/G \in L^2_a(\mathbb{D})$ and $\|f/G\|_{L^2_a(\mathbb{D})} \leq \|f\|_{L^2_a(\mathbb{D})}$. Thus, the function $G$ plays a similar roll for the Bergman space as the Blaschke products do for the Hardy space. Since these results of Hedenmalm, there has been a flurry of papers and new results about invariant subspaces for the Bergman shift. In particular in 1993, Seip [40] characterized the interpolating and sampling sequences in the Bergman space. This allowed Hedenmalm, Richter, and Seip [33] to give a function theoretic construction of invariant subspaces $\mathcal{M}$ for the Bergman shift such that $\mathcal{M} \ominus z\mathcal{M}$ has arbitrary dimension. Some of the work on Bergman spaces has been extended to weighted Bergman spaces and even to $P^2(\mu)$ spaces. The following theorem is a beautiful result of Aleman, Richter, and Sundberg [2] from 2007, which can also be stated for more general Hilbert spaces of analytic functions.

**4.3. Theorem.** *If $S_\mu = M_z$ on $P^2(\mu)$ is pure and the set of analytic bounded point evaluations for $P^2(\mu)$ is the open unit disk, then the following are equivalent.*

(a) $\mu(\partial\mathbb{D}) > 0$.

(b) *There is a measurable set $E \subseteq \partial\mathbb{D}$ with $\mu(E) > 0$ such that every function $f \in P^2(\mu)$ has non-tangential limits a.e. on $E$.*

(c) *For every invariant subspace $\mathcal{M}$ of $S_\mu$,* $\dim(\mathcal{M} \ominus z\mathcal{M}) = 1$.

Two sources of further reading on this topic are the books [32] and [21].

**4.4. Theorem.** (Olin and Thomson [37]) *Every subnormal operator is reflexive.*

Their method of proof exploits the underlying analytic structure of subnormal operators and generalizes the factorization theorem that Brown used in order to get the existence of invariant subspaces. (Also see [14], page 363 for a proof.) Using the results they obtained to prove the reflexivity of subnormal operators, Olin and Thomson also showed that for a subnormal operator $S$ the weakly closed algebra it generates is the same as its weak* closed algebra $P^\infty(S)$ (also see [14], Theorem VII.5.2); so $P^\infty(S) = \operatorname{Alg}\operatorname{Lat} S$.

A different and somewhat elementary proof of the reflexivity of subnormal operators was given by McCarthy [35]. He uses the routine observation that the proof can be reduced to the cyclic case. This enables the application of Thomson's Theorem on bounded point evaluations discussed in the next section though he still requires the Olin and Thomson result on factorization of weak* continuous linear functionals on $P^\infty(S)$.

Recall that a subspace is *hyperinvariant* for an operator $S$ if it is invariant for every operator that commutes with $S$. The following basic question remains open for subnormal operators.

**4.5. Problem.** *Does every subnormal operator have a hyperinvariant subspace?*

## 5. Bounded point evaluations

When Scott Brown [8] proved in 1978 that subnormal operators have invariant subspaces, he sidestepped the issue of determining whether or not bounded point evaluations exist. It had been known for some time that a bounded point evaluation for a cyclic subnormal operator gives rise to a natural invariant subspace for that operator; in fact, even a hyperinvariant subspace. While Brown's proof of the existence of invariant subspaces for subnormals introduced many new ideas and has proven fruitful in other areas, the important question of the existence of bounded point evaluations remained open until 1991 when James Thomson [42] proved not only that bounded point evaluations exist, but they exist in abundance. Moreover he was able to give a beautiful structure theory for cyclic subnormal operators.

A *bounded point evaluation* for $P^2(\mu)$ is a complex number $\lambda$ such that the densely defined linear functional on $P^2(\mu)$ given by $p \mapsto p(\lambda)$ for polynomials $p$ extends to be a continuous linear functional on $P^2(\mu)$, that is, there is a number $M > 0$ such that

$$|p(\lambda)| \le M\|p\|_2$$

for all polynomials $p$.

It follows that if $f \in P^2(\mu)$ and $\{p_n\}$ is a sequence of polynomials such that $p_n \to f$ in $L^2(\mu)$, then from the above inequality, we see that the sequence $\{p_n(\lambda)\}$ is a Cauchy sequence of complex numbers, and hence must converge to a complex number denoted by $\hat{f}(\lambda)$. So, when $\lambda$ is a bounded point evaluation, the map $f \mapsto \hat{f}(\lambda)$ is a well-defined continuous linear functional on $P^2(\mu)$; thus by the Riesz Representation Theorem must be represented by a function $K_\lambda \in P^2(\mu)$ as follows:

$$\hat{f}(\lambda) = \langle f, K_\lambda \rangle$$

for all $f \in P^2(\mu)$. We will let bpe$(\mu)$ denote the set of all bounded point evaluations for $P^2(\mu)$. The functions $\{K_\lambda : \lambda \in \text{bpe}(\mu)\}$ are called the reproducing kernels for $P^2(\mu)$. It is easy to see that $\lambda \in \text{bpe}(\mu)$ if and only if $\overline{\lambda}$ is an eigenvalue for $S_\mu^*$. From an operator theory point of view, scalar multiples of the reproducing kernels are precisely the eigenvectors for $S_\mu^*$. Thus a bounded point evaluation exists for $P^2(\mu)$ precisely when $S_\mu^*$ has an eigenvalue. In this case, the corresponding eigenspace for $\overline{\lambda}$ is the space spanned by the reproducing kernel, $[K_\lambda]$, and is a hyperinvariant subspace for $S_\mu^*$. Hence its orthogonal complement, $[K_\lambda]^\perp$, is a hyperinvariant subspace for $S_\mu$. The orthogonal complement is precisely

$$Z_\lambda = [K_\lambda]^\perp = \{f \in P^2(\mu) : \hat{f}(\lambda) = 0\}.$$

Since $\lambda$ is a bounded point evaluation it follows that $1 \notin Z_\lambda$ but $(z - \lambda) \in Z_\lambda$, so $Z_\lambda$ is non-trivial.

For each $f \in P^2(\mu)$, $\hat{f}$ defines a function from bpe$(\mu)$ into $\mathbb{C}$. Furthermore it is easy to see that $\hat{f} = f$ $\mu$-a.e. on the set of bounded point evaluations. A bounded point evaluation $\lambda$ for $P^2(\mu)$ is an *analytic bounded point evaluation* if there is an $\epsilon > 0$ such that $B(\lambda, \epsilon) \subseteq \text{bpe}(\mu)$ and for any $f \in P^2(\mu)$ the function $\hat{f}$ is analytic in a neighborhood of $\lambda$. According to this definition, the neighborhood on which the function $\hat{f}$ is analytic may depend on the function $f$. However, an easy application of the Baire Category Theorem shows that in fact there is a single neighborhood $U$ of $\lambda$ such that $U \subseteq \text{bpe}(\mu)$ and for every $f \in P^2(\mu)$ the function $\hat{f}$ is analytic on $U$. We will let abpe$(\mu)$ denote the set of all analytic bounded point evaluations for $P^2(\mu)$. The set abpe$(\mu)$ is an open subset of bpe$(\mu)$ and for every $f \in P^2(\mu)$, $\hat{f}$ is an analytic function on abpe$(\mu)$. Thus all functions in $P^2(\mu)$

extend to be analytic on abpe($\mu$). Trent [43] showed that in fact abpe($\mu$) is a dense open subset of bpe($\mu$) and that abpe($\mu$) $= \sigma(S_\mu) \setminus \sigma_{ap}(S_\mu)$.

In June of 1989, Thomson began circulating preprints of an amazing result [42] that has come to be known as 'Thomson's Theorem'.

**5.1. Thomson's Theorem.** *If $\mu$ is a compactly supported positive regular Borel measure on $\mathbb{C}$ and $S_\mu = M_z$ on $P^2(\mu)$ is a pure subnormal operator, then the following hold.*

(a) *$P^2(\mu)$ has bounded point evaluations, in fact, the closure of the set of bounded point evaluations contains the support of $\mu$.*

(b) *The reproducing kernels have dense linear span in $P^2(\mu)$.*

(c) *The set of bounded point evaluations is the same as the set of analytic bounded point evaluations.*

(d) *If $G$ is the open set of analytic bounded point evaluations for $P^2(\mu)$, then $G$ is simply connected and the map $f \mapsto \hat{f}$ from $P^2(\mu) \cap L^\infty(\mu) \to H^\infty(G)$ is an isometric dual isomorphism.*

One of the beautiful things about Thomson's Theorem is that the only real hypothesis is that the operator $S_\mu$ is pure. This is a very simple and natural operator theoretic assumption, and yet from this Thomson proves some very deep results about $P^2(\mu)$. From a function theoretic point of view, the fact that $S_\mu$ is pure means that $P^2(\mu)$ does not have an $L^2$-summand, or equivalently, that there is a function $g \in L^2(\mu)$ such that $|g| > 0$ $\mu$-a.e. and $g \perp P^2(\mu)$; this follows from a result of Chaumat, see [14], page 246.

Some basic questions about the operator $S_\mu$ have immediate answers from Thomson's Theorem and the answers all involve the set of bounded point evaluations. For example if $S_\mu$ is pure and $G = \text{bpe}(\mu)$, then the spectrum, approximate point spectrum, and essential spectrum can easily be computed: $\sigma(S_\mu) = \operatorname{cl} G$, $\sigma_{ap}(S_\mu) = \sigma_e(S_\mu) = \partial G$. Also $S_\mu$ is irreducible if and only if $G$ is connected. Moreover the commutant of $S_\mu$, which consists of the multiplication operators $f(S_\mu)$ for $f \in P^2(\mu) \cap L^\infty(\mu)$, can naturally be identified with $H^\infty(G)$.

Below are a few other consequences of Thomson's Theorem for cyclic subnormal operators that are more substantial.

**5.2. Theorem.** *If $S_\mu = M_z$ on $P^2(\mu)$ is pure and $G = bpe(\mu)$, then the following hold.*

(a) *$S_\mu$ has a trace class self-commutator and* $\operatorname{tr}[S_\mu^*, S_\mu] = \frac{1}{\pi} Area(G)$.

(b) *If $f \in P^2(\mu) \cap L^\infty(\mu)$, then $\sigma(f(S_\mu)) = \operatorname{cl} f(G)$, where $f(S_\mu)$ is the operator of multiplication by $f$ on $P^2(\mu)$.*

(c) *$S_\mu$ has a square root if and only if $0 \notin G$.*

(d) *$S_\mu^*$ is hypercyclic if and only if every component of $G$ intersects the unit circle.*

(e) *$S_\mu$ has a finite rank self-commutator if and only if $G$ has only finitely many components, $G$ is a quadrature domain, and the measures $\mu$ and $\omega_G + \sum_{k=1}^n \delta_{a_k}$, the sum of harmonic measure on $G$ plus a finite sum of point masses at points $\{a_k\}_{k=1}^n$ in $G$, are mutually absolutely continuous.*

*Trace Estimates.* Recall that the self-commutator of an operator $T$ is the self-adjoint operator $[T^*, T] = T^*T - TT^*$. An operator $T$ is normal if its self-commutator is equal to zero, it is essentially normal if its self-commutator is compact, and it is hyponormal if its self-commutator is positive semi-definite. For any operator $T$, $[T^*, T]$ is trace class exactly when $[T^*, T]$ is compact and its (real) eigenvalues, which necessarily converge to zero, form an absolutely convergent series. In that case the trace of $[T^*, T]$, denoted by $\operatorname{tr}[T^*, T]$, is equal to the sum of its eigenvalues.

The Berger-Shaw Theorem [5] implies that every rationally cyclic hyponormal operator $T$ has a trace class self-commutator and gives an estimate for the trace of $[T^*, T]$; namely $\operatorname{tr}[T^*, T] \leq \frac{1}{\pi}\text{Area}[\sigma(T)]$.

Since subnormal operators are also hyponormal, the Berger-Shaw Theorem implies that a cyclic subnormal operator $S_\mu$ does have a trace class self-commutator and that $\operatorname{tr}[S_\mu^*, S_\mu] \leq \frac{1}{\pi}\text{Area}[\sigma(S_\mu)]$. In 1997, Feldman [23] using Thomson's Theorem and some techniques from the Berger-Shaw Theorem [5] was able to give another proof that a pure cyclic subnormal operator $S_\mu$ has a trace class self-commutator and was able to compute the trace of the self-commutator precisely as $\frac{1}{\pi}\text{Area}(G)$.

The following problem would be a natural generalization of the above result to non-cyclic subnormal operators. See Feldman [23] for more details and some examples surrounding this problem.

**5.3. Problem.** *If $S$ is a pure subnormal operator such that $R(\sigma_e(S)) = C(\sigma_e(S))$, and ind$(S - \lambda I) = -1$ for all $\lambda \in G = \sigma(S) \setminus \sigma_e(S)$, then is $\operatorname{tr}[S^*, S] = \frac{1}{\pi}\text{Area}[G]$?*

*The commutant.* It is easy to identify the commutant of $S_\mu$ as the multiplication operators $f(S_\mu)$ on $P^2(\mu)$ by functions $f \in P^2(\mu) \cap L^\infty(\mu)$. However, Thomson's Theorem helps us to identify the Banach algebra $P^2(\mu) \cap L^\infty(\mu)$ as the space of all bounded analytic functions on $H^\infty(G)$. This allows us to find the spectrum of $f(S_\mu)$, it's norm, and so forth.

*Square roots:* The above identification of the commutant of $S_\mu$ also answers immediately the square root problem for cyclic subnormal operators. The operator $S_\mu$ will have a square root exactly when the analytic function $z$ has a square root in $H^\infty(G)$. Since $G$ is simply connected, this will be precisely when $0 \notin G$.

We see here that it is important to be able to say which analytic functions on $G$ belong to $P^2(\mu)$. For this problem it suffices to assume that the set of analytic bounded point evaluations is the open unit disk. It is clear that $P^2(\mu) \subseteq \text{Hol}(\mathbb{D}) \cap L^2(\mu|\mathbb{D})$. It can also be shown using Thomson's Theorem that $N^+(\mathbb{D}) \cap L^2(\mu) \subseteq P^2(\mu)$ where $N^+(\mathbb{D})$ is the Smirnov-Nevanlinna class of functions that are quotients of bounded analytic functions $g/h$ with $h$ an outer function. In general the following problem remains open.

**5.4. Problem.** *Which analytic functions on the disk belong to $P^2(\mu)$ when* $\text{bpe}(\mu) = \mathbb{D}$*?*

It is not known in general which subnormal operators have square roots. However, since we now understand which cyclic subnormal operators have square roots, it is natural to ask if we can use that knowledge to say something about square roots of the arbitrary subnormal operator. This approach is part of a general approach to the study of subnormal operators.

A *part* of an operator $S$ is any operator of the form $S|\mathcal{M}$ where $\mathcal{M}$ is an invariant subspace for $S$. A *cyclic part* is one where $S|\mathcal{M}$ is a cyclic operator. It is reasonable to ask if a subnormal operator has a given property when all its cyclic parts have this property. For a property P one might say that a subnormal operator has P locally if all its cyclic parts have P. In this terminology the question is whether having P locally implies that the operator has it globally. A useful tool in this regard is the deep result of Olin and Thomson (see [14], page 361) on the existence of full analytic subspaces for subnormal operators. Using this result and Thomson's Theorem it is rather straightforward to prove the following.

**5.5. Theorem.** *If $S$ is a subnormal operator, then $S$ has a square root in $P^\infty(S)$ if and only if every cyclic part of $S$ has a square root.*

**5.6. Problem.** *Characterize the subnormal operators having a square root.*

*Cyclicity.* In 1999, Feldman [24] proved that the adjoint of every pure subnormal operator is cyclic. So even though most pure subnormal operators are not cyclic, their adjoints are always cyclic! This is rather surprising, especially since the class of pure subnormal operators is closed under direct sums. This suggests that the adjoints of subnormal operators might possess some stronger forms of cyclicity, such as hypercyclicity. An operator $T$ is *hypercyclic* on a Hilbert space $\mathcal{H}$ if there is a vector $h \in \mathcal{H}$ such that the orbit of $h$ under $T$, $\{T^n h : n \geq 0\}$, is dense in $\mathcal{H}$. A necessary condition for an operator to be hypercyclic is that every component of its spectrum must intersect the unit circle (see [34]).

Using Thomson's Theorem and techniques developed by Godefroy and Shapiro [28], it follows that for a pure cyclic subnormal operator $S_\mu$, $S_\mu^*$ is hypercyclic if and only if every component of $\mathrm{bpe}(\mu)$ intersects the unit circle.

In [26] Feldman, Miller, and Miller classified the hyponormal operators whose adjoints are hypercyclic as those operators $S$ such that every "part of the spectrum" of $S$ intersects both the inside and outside of the unit circle. A "part of the spectrum" of $S$ is a compact set of the form $\sigma(S|\mathcal{M})$ where $\mathcal{M}$ is an invariant subspace of $S$.

**5.7. Theorem.** *If $S$ is a hyponormal operator, then $S^*$ is hypercyclic if and only if for every invariant subspace $\mathcal{M}$ of $S$ we have $\sigma(S|\mathcal{M}) \cap \{z : |z| < 1\} \neq \emptyset$ and $\sigma(S|\mathcal{M}) \cap \{z : |z| > 1\} \neq \emptyset$.*

An operator is *weakly hypercyclic* if there is a vector whose orbit is weakly dense in the underlying Hilbert space. The following problem remains unsolved.

**5.8. Problem.** *Characterize the pure cyclic subnormal operators $S$ such that $S^*$ is weakly hypercyclic.*

As mentioned above, if $S$ is a pure subnormal operator, then $S^*$ is cyclic. A commuting tuple of operators $S = (S_1, S_2, \dots, S_n)$ is said to be a subnormal tuple if each operator is subnormal and they have commuting normal extensions. The tuple $S$ is pure if there is no common reducing subspace on which the tuple is a normal tuple. Finally a tuple is cyclic if there is a vector $h$ such that the smallest closed invariant subspace for the tuple that contains $h$ is the whole space.

It is possible that for a pure subnormal tuple each operator is not pure. For example if $A$ and $B$ are two pure subnormal operators with $\|A\| < 1$ and $\|B\| < 1$ and $S_1 = A \oplus I$ and $S_2 = I \oplus B$, then $S = (S_1, S_2)$ is a pure subnormal tuple and yet neither $S_1$ nor $S_2$ is pure. Hence the following question seems natural.

**5.9. Problem.** *If $S = (S_1, S_2, S_3, \dots, S_n)$ is a pure subnormal tuple, is the adjoint $S^* = (S_1^*, S_2^*, \dots, S_n^*)$ cyclic?*

Naturally if one of the operators $S_i$ is pure, then the answer is yes because $S_i^*$ will be cyclic, but otherwise the problem remains open.

*Finite Rank Self-Commutators.* Subnormal operators with finite rank self-commutators have been studied by several authors, including Xia [45, 46, 47], Olin, Thomson, and Trent [38], McCarthy and Yang [36], Yakubovich [44], and more. Such operators are very special and are closely related to the unilateral shift and rational functions of the shift.

If $S_m$ denotes the unilateral shift, $S_m = M_z$ on $H^2(\mathbb{D}) = P^2(m)$ where $m$ is normalized Lebesgue measure on the circle, then the self-commutator of $S_m$ is the rank one projection onto the constant functions. It is well known and not difficult to show that if $S$ is any pure subnormal operator that has rank-one self-commutator, then $S$ is simply the translate of a multiple of the unilateral shift, that is, $S = f(S_m)$ where $f(z) = az + b$. More generally, if $f(z)$ is any rational function with poles off the closed unit disk, then the subnormal operator $S = f(S_m)$ will have a finite rank self-commutator. Finally, a computation shows that a finite-dimensional extension of an operator with finite rank self-commutator also has finite rank self-commutator. It is natural to try to describe all subnormal operators with finite rank self-commutators, and in particular the cyclic ones. The following theorem was first proven by Olin, Thomson and Trent [38] in an unpublished manuscript. It also follows from the work of Xia [45, 46, 47] and McCarthy and Yang [36].

**5.10. Theorem.** *If $S_\mu = M_z$ on $P^2(\mu)$ is pure and $G = \operatorname{bpe}(\mu)$, then $S_\mu$ has a finite rank self-commutator if and only if $G$ has only finitely many components, $G$ is a quadrature domain, and the measures $\mu$ and $\omega_G + \sum_{k=1}^n \delta_{a_k}$, the sum of harmonic measure on $G$ plus a finite sum of point masses at points $\{a_k\}_{k=1}^n$ in $G$, are mutually absolutely continuous.*

So $S_\mu$ has finite rank self-commutator when $\mu$ is equivalent to harmonic measure plus a finite sum of point masses and the set $G$ of analytic bounded point evaluations is a very special type of set, namely a quadrature domain. A domain $G$ is a *quadrature domain* if there is a meromorphic function $S(z)$ defined on $G$ that is continuous up to the boundary of $G$ and $S(z) = \overline{z}$ for all $z \in \partial G$. If $G$ is simply connected, then $G$ is a quadrature domain if and only if the Riemann map from the open unit disk onto $G$ is a rational function.

Notice that if we assume that $S_\mu$ is irreducible and thus $G$ is connected, the classification above takes on the following form.

**5.11. Theorem.** *An irreducible cyclic subnormal operator $S_\mu$ has finite rank self-commutator if and only if $S_\mu$ is a finite-dimensional extension of $f(S_m)$, where $f$ is a univalent rational function with poles off the closed unit disk and $S_m$ is the unilateral shift.*

In [25], Feldman studies cyclic and non-cyclic subnormal operators that are arbitrary extensions of certain functions of the unilateral shift; such operators have self-commutators whose ranges are not dense in the whole space. Also, McCarthy and Yang [36] give a nice proof of the theorem of Olin, Thomson, and Trent, making use of Thomson's theorem on the existence of bounded point evaluations and also characterize the rationally cyclic subnormal operators with finite rank self-commutators in essentially the same manner, except in the rationally cyclic case the set of bounded point evaluations need not be simply connected. Finally, Yakubovich [44] has characterized all irreducible subnormal operators with finite rank self-commutators using Riemann surface analogues of quadrature domains.

*Extensions of Thomson's Theorem.* One natural extension of Thomson's Theorem is to consider bounded point evaluations for rationally cyclic subnormal operators. That is, let $R^2(K,\mu)$ denote the closure of $\operatorname{Rat}(K)$, the set of rational functions with poles off $K$, in $L^2(\mu)$. It has been known since 1975 that there are compact sets $K$ with no interior such that the space $R^2(K,\text{Area})$ has no bounded point evaluations (see Fernstrom [27]). However, after Thomson's Theorem appeared, Conway and Elias [15] were able to use his results and his techniques to show that if $S = M_z$ on $R^2(K,\mu)$ is pure, then the set of analytic bounded point evaluations for $R^2(K,\mu)$ is dense in the interior of the spectrum of $S$. Nevertheless, rationally cyclic subnormal operators are more complicated creatures then cyclic subnormal operators. For instance, one may construct a compact set $K$ from a Swiss cheese set by adding some "bubbles" to the holes and create a rationally cyclic subnormal operator $S = M_z$ on $R^2(K,\mu)$ so that $R^2(K,\mu)$ has analytic bounded point evaluations which form an open dense subset of the spectrum of $S$ and yet the reproducing kernels are not dense in $R^2(K,\mu)$. So certain parts of Thomson's Theorem are not true in the rationally cyclic case.

In [22] Feldman proposes a natural generalization of the idea of bounded point evaluation that may be considered for any subnormal operator. First consider the rationally cyclic case. For a positive regular Borel measure $\nu$ on a compact set $K$, $\nu$ is called an *interpolating measure* for $S = M_z$ on $R^2(K,\mu)$ if the densely defined map $A : \operatorname{Rat}(K) \to L^2(\nu)$ defined by $A(f) = f$ extends to be a surjective bounded linear operator $A : R^2(K,\mu) \to L^2(\nu)$. When $\nu = \delta_\lambda$ is the unit point mass measure at $\lambda$, then $\nu$ is an interpolating measure precisely when $\lambda$ is a bounded point evaluation. Furthermore, one needs the map $A$ to be surjective, so that one can naturally construct invariant subspaces from $A$. For example, the kernel of $A$ would then be a non-trivial invariant subspace for $S$.

**5.12. Problem.** *Does $M_z$ on $R^2(K,\mu)$ have an interpolating measure?*

In general if $S$ is any subnormal operator on a Hilbert space $\mathcal{H}$, then a measure $\nu$ is an *interpolating measure* for $S$ if there is a surjective linear map $A : \mathcal{H} \to L^2(\nu)$ such that $AS = M_zA$. Feldman has shown that the interpolating measures for cyclic subnormal operators all arise from the bounded point evaluations. However, there are interpolating measures for $M_z$ on the Hardy space of the slit disk that are supported on the slit. Furthermore, Feldman has shown that operators, such as the dual of the Bergman shift, have a "complete set" of interpolating measures. The following question remains unanswered.

**5.13. Problem.** *Does every subnormal operator have an interpolating measure?*

## 6. Conclusion

Thus we see that from Halmos's definition of a subnormal operator an entire theory has arisen that has generated other areas of operator theory such as dual algebras and has influenced a new line of inquiry in an established area of function theory, namely Bergman spaces. This is remarkable and something that must be noted whenever Paul Halmos is discussed.

## References

[1] A. Aleman, S. Richter, and C. Sundberg, *Beurling's Theorem for the Bergman space*, Acta Math. 177 (1996), no. 2, 275–310.

[2] A. Aleman, S. Richter, C. Sundberg, *Analytic contractions, nontangential limits, and the index of invariant subspaces*, Trans. Amer. Math. Soc. 359 (2007), no. 7, 3369–3407.

[3] C. Apostol, H. Bercovici, C. Foias, and C. Pearcy *Invariant subspaces, dilation theory, and the structure of the predual of a dual algebra. I*, J. Funct. Anal. 63 (1985) 369–404.

[4] H. Bercovici, C. Foias, and C. Pearcy, *Dual algebras with applications to invariant subspaces and dilation theory*, CBMS notes 56, Amer. Math. Soc., Providence, 1985.

[5] C.A. Berger and B.I. Shaw, *Selfcommutators of multicyclic hyponormal operators are trace class*, Bull. Amer. Math. Soc. 79 (1973) 1193–1199.

[6] E. Bishop, *Spectral theory for operators on a Banach space*, Trans. Amer. Math. Soc. 86 (1957) 414–445.

[7] J. Bram, *Subnormal operators*, Duke Math. J., 22 (1955) 75–94.

[8] S.W. Brown, *Some invariant subspaces for subnormal operators*, Integral Equations Operator Theory, 1 (1978) 310–333.

[9] S.W. Brown, *Hyponormal operators with thick spectra have invariant subspaces*, Ann. Math., 125 (1987) 93–103.

[10] S.W. Brown, B. Chevreau, and C. Pearcy, *On the structure of contraction operators II,* J. Funct. Anal. 76 (1988) 30–55.

[11] R.W. Carey and J.D. Pincus, *Mosaics, principal functions, and mean motion in von Neumann algebras*, Acta Math. 138 (1977) 153–218.

[12] J.B. Conway, *The dual of a subnormal operator*, J. Operator Theory 5 (1981) 195–211.

[13] J.B. Conway, *A Course in Functional Analysis*, Second Edition, Springer-Verlag, New York (1990).

[14] J.B. Conway, *The Theory of Subnormal Operators*, Amer. Math. Soc., Providence, (1991).

[15] J.B. Conway and N. Elias, *Analytic bounded point evaluations for spaces of rational functions*, J. Functional Analysis 117 (1993) 1–24.

[16] J.B. Conway and D.W. Hadwin, *Strong limits of normal operators*, Glasgow Mathematical J., 24 (1983) 93–96.

[17] J.B. Conway and R.F. Olin, *A functional calculus for subnormal operators,II*, Memoirs Amer. Math. Soc. 184 (1977).

[18] M.J. Cowen and R.G. Douglas, *Complex geometry and operator theory*, Acta Math. 141 (1978) 187–261.

[19] S.J. Dilworth and V.G. Troitsky, *Spectrum of a weakly hypercyclic operator meets the unit circle*, Trends in Banach spaces and operator theory (Memphis, TN, 2001), 67–69, Contemp. Math., 321, Amer. Math. Soc., Providence, RI, 2003.

[20] J.J. Dudziak, *Spectral mapping theorems for subnormal operators*, J. Funct. Analysis, 56 (1984) 360–387.

[21] P. Duren and A. Schuster, *Bergman spaces*, Amer. Math. Soc., Providence (2004).

[22] N.S. Feldman, *Interpolating Measures for Subnormal Operators* [Abstracts from the mini-workshop held August 14–20, 2005.] Oberwolfach Reports 2 (2005), no. 3, 2043–2089.

[23] N.S. Feldman, *The Berger-Shaw Theorem for Cyclic Subnormal Operators*, Indiana Univ. Math. J., 46 No. 3, (1997), p. 741–751.

[24] N.S. Feldman, *Pure subnormal operators have cyclic adjoints*, J. Funct. Anal. 162 (1999) 379–399.

[25] N.S. Feldman, *Subnormal operators, self-commutators, and pseudocontinuations*, Integral Equations Operator Theory 37 (2000), no. 4, 402–422.

[26] N.S. Feldman, V. Miller, and L. Miller, *Hypercyclic and Supercyclic Cohyponormal Operators*, Acta Sci. Math. (Szeged) 68 (2002), no. 1-2, 303–328.

[27] C. Fernstrom, *Bounded point evaluations and approximation in $L^p$ by analytic functions*, Spaces of analytic functions (Sem. Functional Anal. and Function Theory, Kristiansand, 1975), pp. 65–68. Lecture Notes in Math., Vol. 512, Springer, Berlin, 1976.

[28] G. Godefroy and J.H. Shapiro, *Operators with dense invariant cyclic manifolds*, J. Funct. Anal., 98 (1991), 229–269.

[29] P.R. Halmos, *Normal dilations and extensions of operators*, Summa Bras. Math., 2 (1950) 125–134.

[30] H. Hedenmalm, *A factorization theorem for square area-integrable analytic functions*, J. Reine Angew. Math. 422 (1991), 45–68.

[31] H. Hedenmalm, *An invariant subspace of the Bergman space having the codimension two property*, J. Reine Angew. Math., 443 (1993) 1–9.

[32] H. Hedenmalm, B. Korenblum, and K. Zhu, *Theory of Bergman spaces*, Springer-Verlag, New York (2000).

[33] H. Hedenmalm, S. Richter, K. Seip, *Interpolating sequences and invariant subspaces of given index in the Bergman spaces*, J. Reine Angew. Math. 477 (1996), 13–30.

[34] C. Kitai, *Invariant closed sets for linear operators*, Dissertation, Univ. of Toronto, 1982.

[35] John E. McCarthy, *Reflexivity of subnormal operators*, Pacific J. Math. 161 (1993) 359–370.

[36] J.E. McCarthy and L. Yang, *Subnormal operators and quadrature domains*, Adv. Math. 127 (1997), no. 1, 52–72.

[37] R.F. Olin and J.E. Thomson, *Algebras of subnormal operators*, J. Funct. Analysis 37 (1980) 271–301.

[38] R. Olin, J.E. Thomson, and T. Trent, *Subnormal operators with finite rank self-commutators*, unpublished manuscript.

[39] D. Sarason, *Weak-star density of polynomials*, *J. Reine Angew. Math.* 252 (1972) 1–15.

[40] K. Seip, *Beurling type density theorems in the unit disk*, Invent. Math. 113 (1993), no. 1, 21–39.

[41] J.E. Thomson, *Invariant subspaces for algebras of subnormal operators*, Proc. Amer. Math. Soc. 96 (1986) 462–464.

[42] J.E. Thomson, *Approximation in the mean by polynomials*, Ann. of Math. (2) 133 (1991), no. 3, 477–507.

[43] T.T. Trent, $H^2(\mu)$ *spaces and bounded point evaluations*, Pacific J. Math. 80 (1979), no. 1, 279–292.

[44] D.V. Yakubovich, *Subnormal operators of finite type. II. Structure theorems*, Rev. Mat. Iberoamericana 14 (1998), no. 3, 623–681.

[45] Daoxing Xia, *The analytic model of a subnormal operator*, Integral Equations Operator Theory 10 (1987) 258–289.

[46] D. Xia, *Analytic theory of subnormal operators*, Integral Equations Operator Theory 10 (1987), no. 6, 880–903.

[47] D. Xia, *Errata: Analytic theory of subnormal operators* [*Integral Equations Operator Theory* 10 (1987)*, no.* 6*,* 880–903] Integral Equations Operator Theory 12 (1989), no. 6, 898–899.

John B. Conway
George Washington University
Washington, DC 20052, USA
e-mail: conway@gwu.edu

Nathan S. Feldman
Washington & Lee University
Lexington, VA 24450, USA
e-mail: feldmanN@wlu.edu

Operator Theory:
Advances and Applications, Vol. 207, 195–207

# Polynomially Hyponormal Operators

Raúl Curto and Mihai Putinar

*To the memory of Paul R. Halmos*

**Abstract.** A survey of the theory of $k$-hyponormal operators starts with the construction of a polynomially hyponormal operator which is not subnormal. This is achieved via a natural dictionary between positive functionals on specific convex cones of polynomials and linear bounded operators acting on a Hilbert space, with a distinguished cyclic vector. The class of unilateral weighted shifts provides an optimal framework for studying $k$-hyponormality. Non-trivial links with the theory of Toeplitz operators on Hardy space are also exposed in detail. A good selection of intriguing open problems, with precise references to prior works and partial solutions, is offered.

**Mathematics Subject Classification (2000).** Primary 47B20; Secondary 47B35, 47B37, 46A55, 30E05.

**Keywords.** Hyponormal operator, weighted shift, moment problem, convex cone, Toeplitz operator.

## 1. Hyponormal operators

Let $\mathcal{H}$ be a separable complex Hilbert space. A linear operator $S$ acting on $\mathcal{H}$ is called *subnormal* if there exists a linear bounded extension of it to a larger Hilbert space, which is normal. Denoting by $\mathcal{K}$ this larger space and $P = P^{\mathcal{K}}_{\mathcal{H}}$ the orthogonal projection onto $\mathcal{H}$, the above definition can be translated into the identity

$$S = PN|_{\mathcal{H}} = N|_{\mathcal{H}},$$

with $N$ a normal operator acting on $\mathcal{K}$. The spectral theorem asserts that the normality condition

$$[N^*, N] := N^*N - NN^* = 0$$

This work was partially supported by NSF grants.

implies the existence of a standard functional model for $N$, specifically described as an orthogonal direct sum of multipliers

$$(Nh)(z) = zh(z), \quad h \in L^2(\mu),$$

where $\mu$ is a positive Borel measure, compactly supported in the complex plane $\mathbb{C}$. In turn, the canonical example of a subnormal operator is (a direct sum of) multipliers

$$(Sf)(z) = zf(z), \quad f \in P^2(\mu),$$

where $P^2(\mu)$ stands for the closure of polynomials in the Lebesgue space $L^2(\mu)$.

The self-commutator of a subnormal operator is non-negative:

$$\begin{aligned}\langle [S^*, S]f, f\rangle &= \|Sf\|^2 - \|S^*f\|^2 = \|Nf\|^2 - \|PN^*f\|^2 \\ &= \|N^*f\|^2 - \|PN^*f\|^2 \geq 0.\end{aligned}$$

It was Paul Halmos [36] who isolated the class of *hyponormal* operators, as those linear bounded operators $T \in \mathcal{L}(\mathcal{H})$ which satisfy $[T^*, T] \geq 0$; it was soon discovered that not all hyponormal operators are subnormal. Much later it was revealed that a typical hyponormal operator model departs quite sharply from the above-mentioned multipliers. More precisely, such a model is provided by one-dimensional singular integrals, of the form

$$(T\phi)(x) = x\phi(x) + ia(x)\phi(x) - b(x)(\text{p.v.}) \int_{-M}^{M} \frac{b(t)\phi(t)dt}{t-x},$$

where $M > 0$, $a, b \in L^\infty[-M, M]$ are real-valued functions and $\phi \in L^2[-M, M; dt]$. The reader will easily verify that

$$(T^*\phi)(x) = x\phi(x) - ia(x)\phi(x) + b(x)(\text{p.v.}) \int_{-M}^{M} \frac{b(t)\phi(t)dt}{t-x},$$

hence

$$[T^*, T]\phi = 2b(x) \int_{-M}^{M} b(t)\phi(t)dt,$$

and consequently

$$\langle [T^*, T]\phi, \phi\rangle = 2\left| \int_{-M}^{M} b(t)\phi(t)dt \right|^2 \geq 0.$$

The question of understanding better the gap between subnormal and hyponormal operators was raised by Halmos; cf. his Hilbert space problem book [37]. In this direction, the following technical problem has naturally appeared: if $S$ is a subnormal operator and $p$ is a polynomial, then it is clear from the definition that $p(S)$ is also subnormal. One can see using simple examples of Toeplitz operators that, in general, $T^2$ is not hyponormal if $T$ is hyponormal. What happens if $p(T)$ is hyponormal for all polynomials $p$? Is $T$ in this case subnormal? About 15 years ago we were able to provide a counterexample.

**Theorem 1.1 ([32]).** *There exists a polynomially hyponormal operator which is not subnormal.*

Much more is known today. A whole scale of intermediate classes of operators, a real jungle, was discovered during the last decade. Their intricate structure is discussed in Sections 3 and 4 of this note.

## 2. Linear operators as positive functionals

The main idea behind the proof of Theorem 1.1 is very basic, and it proved to be, by its numerous applications, more important than the result itself.

Let $A \in \mathcal{L}(\mathcal{H})$ be a bounded self-adjoint operator with cyclic vector $\xi$; that is, the linear span of the vectors $A^k\xi$ $(k \geq 0)$, is dense in $\mathcal{H}$. Denote $M := \|A\|$ and let $\Sigma^2$ denote the convex cone of all sums of squares of moduli of complex-valued polynomials, in the real variable $x$. If $p \in \Sigma^2 + (M^2 - x^2)\Sigma^2$, that is

$$p(x) = \sum_i |q_i(x)|^2 + \sum_j (M^2 - x^2)|r_j(x)|^2,$$

with $q_i, r_j \in \mathbb{C}[x]$, then

$$\langle p(A)\xi, \xi\rangle = \sum_i \|q_i(A)\xi\|^2 + \sum_j (M^2\|r_j(A)\xi\|^2 - \|Ar_j(A)\xi\|^2) \geq 0.$$

Since every non-negative polynomial $p$ on the interval $[-M, M]$ belongs to $\Sigma^2 + (M^2 - x^2)\Sigma^2$, the Riesz representation theorem implies the existence of a positive measure $\sigma$, supported in $[-M, M]$, so that

$$\langle p(A)\xi, \xi\rangle = \int p(t)d\sigma(t).$$

From here, a routine path leads us to the full spectral theorem; see for details [50]. An intermediate step in the above reasoning is important for our story, namely

**Proposition 2.1.** *There exists a canonical bijection between contractive self-adjoint operators $A$ with a distinguished cyclic vector $\xi$ and linear functionals $L \in \mathbb{C}[x]'$ which are non-negative on the cone $\Sigma^2 + (1 - x^2)\Sigma^2$. The correspondence is established by the compressed functional calculus map*

$$L(p) = \langle p(A)\xi, \xi\rangle, \quad p \in \mathbb{C}[x].$$

The reader would be tempted to generalize the above proposition to an arbitrary tuple of commuting self-adjoint operators. Although the result is the same, the proof requires a much more subtle Positivstellensatz (that is, a standard decomposition of a positive polynomial, on the polydisk in this case, into a weighted sum of squares). Here is the correspondence.

**Proposition 2.2.** *There exists a canonical bijection between commuting $d$-tuples of contractive self-adjoint operators $A_1, \dots, A_d$ with a distinguished common cyclic vector $\xi$ and linear functionals $L \in \mathbb{C}[x_1, \dots, x_d]'$ which are non-negative on the* *cone* $\Sigma^2 + (1 - x_1^2)\Sigma^2 + \cdots + (1 - x_d^2)\Sigma^2$. *The correspondence is established by the compressed functional calculus map*

$$L(p) := \langle p(A)\xi, \xi\rangle, \qquad p \in \mathbb{C}[x_1, \ldots, x_d].$$

The Positivstellensatz alluded to above (proved by the second named author in 1994 [49]) asserts that a strictly positive polynomial $p$ on the hypercube $[-1,1] \times [-1,1] \times \cdots \times [-1,1] \subset \mathbb{R}^d$ belongs to $\Sigma^2 + (1 - x_1^2)\Sigma^2 + \cdots + (1 - x_d^2)\Sigma^2$. The survey [40] contains ample remarks on the links between the spectral theorem, Positivestellensätze in real algebra, optimization and applications to control theory.

In order to bring the classes of close-to-normal operators into the picture, we need a non-commutative calculus, applied to an operator and its adjoint. The idea goes back to the quasi-nilpotent equivalence relation introduced by I. Colojoara and C. Foiaş [8], and the hereditary functional calculus cast into a formal definition by J. Agler [2]. Let $T \in \mathcal{L}(\mathcal{H})$ and let $z$ denote the complex variable in $\mathbb{C}$. For every monomial we define the *hereditary functional calculus* by

$$z^m \overline{z}^n (T, T^*) := T^{*n} T^m,$$

that is, we place all powers of $T^*$ to the left of the powers of $T$. It is clear that some weak positivity of this functional calculus map is persistent for all operators $T$. More specifically, if $p \in \mathbb{C}[z]$ then

$$p(z)\overline{p(z)}(T, T^*) = p(T)^* p(T) \geq 0.$$

Aiming at a correspondence between positive functionals and operators as in the self-adjoint case, we define $\Sigma_a^2$ to be the convex cone generated in the algebra $\mathbb{C}[z, \overline{z}]$ by $|p(z)|^2$, $p \in \mathbb{C}[z]$. On the other hand, we denote as above by $\Sigma^2$ the convex cone of all sums of squares of moduli of polynomials, that is, polynomials of the form $|p(z, \overline{z})|^2$.

The main terms of the dictionary are contained in the following result, going back to the works of J. Agler [2], S. McCullough and V. Paulsen [47], and the authors [32].

**Theorem 2.3.**

a) *There exists a bijective correspondence between linear contractive operators* $T \in \mathcal{L}(\mathcal{H})$ *with a distinguished cyclic vector* $\xi$ *and linear functionals* $L \in \mathbb{C}[z, \overline{z}]'$ *which are non-negative on the convex cone* $(1 - |z|^2)\Sigma_a^2 + \Sigma_a^2$, *established by the hereditary calculus*

$$L(p) := \langle p(T, T^*)\xi, \xi\rangle, \quad p \in \mathbb{C}[z, \overline{z}].$$

b) *The operator* $T$ *is subnormal if and only if* $L$ *is non-negative on* $\Sigma^2$.

c) *The operator* $T$ *is hyponormal if and only if*

$$L(|r + \overline{z}s|^2) \geq 0, \quad r, s \in \mathbb{C}[z].$$

d) *The operator* $T$ *is polynomially hyponormal if and only if*

$$L(|r + \overline{q}s|^2) \geq 0, \quad q, r, s \in \mathbb{C}[z]. \tag{2.1}$$

The proof of assertion b) is based on the celebrated Bram-Halmos criterion for subnormality [37]. The proofs of a), c) and d) are simple manipulations of the definitions.

The above interpretation of subnormality and polynomial hyponormality invites the study of the filtration given by the condition

$$L(|p|^2) \geq 0 \text{ for all } p(z,\bar{z}) \equiv \sum_{j=0}^{k} \bar{z}^j p_j(z) \; (p_j \in \mathbb{C}[z]), \tag{2.2}$$

which defines the so-called *k-hyponormal* operators. In [32] we proved that polynomial hyponormality does not imply 2-hyponormality, and therefore does not imply subnormality either.

The proof of Theorem 1.1 consists in the construction of a linear functional which separates the convex cones associated to subnormal, respectively polynomially hyponormal operators; see [32] for details. The pioneering work of G. Cassier [5] contains an explicit construction of the same kind. The existence of the separating functional is known in the locally convex space theory community as the Kakutani-Eidelheit Lemma, and it is nowadays popular among the customers of multivariate moment problems (cf. [40]).

## 3. *k*-hyponormality for unilateral weighted shifts

Given a bounded sequence $\equiv \{\alpha_n\}_{n=0}^{\infty}$ of positive numbers, the unilateral weighted shift $W_\alpha$ acts on $\ell^2(\mathbb{Z}_+)$ by $W_\alpha e_n := \alpha_n e_{n+1}$ $(n \geq 0)$. Within this class of operators, the condition (2.2) acquires a rather simple form:

$$W_\alpha \text{ is } k\text{-hyponormal } \iff H_n := (\gamma_{n+i+j})_{i,j=0}^{k} \geq 0 \text{ (all } n \geq 0),$$

where $\gamma_0 := 1$ and $\gamma_{p+1} := \alpha_p^2 \gamma_p$ $(p \geq 0)$ [14]. Thus, detecting $k$-hyponormality amounts to checking the positivity of a sequence of $(k+1) \times (k+1)$ Hankel matrices. With this characterization at hand, it is possible to distinguish between $k$-hyponormality and $(k+1)$-hyponormality for every $k \geq 1$. Moreover, by combining the main result in [32] with the work in [47], we know that there exists a polynomially hyponormal unilateral weighted shift $W_\alpha$ which is not subnormal; however, it remains an open problem to find a specific weight sequence $\alpha$ with that property.

While $k$-hyponormality of weighted shifts admits a simple characterization, the same cannot be said of polynomial hyponormality. When one adds the condition deg $q \leq k$ to (2.1), we obtain the notion of weak $k$-hyponormality. We thus have a staircase leading up from hyponormality to subnormality, passing through 2-hyponormality, 3-hyponormality, and so on. A second staircase starts at hyponormality, goes up to quadratic hyponormality, to cubic hyponormality, and eventually reaches polynomial hyponormality. How these two staircases intertwine is not well understood. A number of papers have been written describing the links for specific families of weighted shifts, e.g., those with recursively generated tails and those obtained by restricting the Bergman shift to suitable invariant subspaces [14], [15], [16], [17], [18], [19], [20], [21], [22], [29], [30], [43], [44]; the overall problem, however, remains largely unsolved.

One first step is to study the precise connections between quadratic hyponormality and 2-hyponormality. While there are several results that establish quantitative differences between these two notions, there are two qualitative results that stand out. The first one has to do with a propagation phenomenon valid for the class of 2-hyponormal weighted shifts.

**Theorem 3.1.**

(i) *If $\alpha_0 = \alpha_1$ and $W_\alpha$ is 2-hyponormal, then $W_\alpha = \alpha_0 U_+$, that is, $W_\alpha$ is a multiple of the* (*unweighted*) *unilateral shift* [14] (*for a related propagation result, see* [6]);

(ii) *The set $Q := \{(x, y) \in \mathbb{R}^2_+ : W_{1,(1,x,y)^\wedge}$ is quadratically hyponormal$\}$ contains a closed convex set with nonempty interior* [19].

*Thus, there exist many nontrivial quadratically hyponormal weighted shifts with two equal weights.*

The second result entails completions of weighted shifts. J. Stampfli showed in [51] that given three initial weights $\alpha_0 < \alpha_1 < \alpha_2$, it is always possible to find new weights $\alpha_3, \alpha_4, \ldots$ such that $W_\alpha$ is subnormal; that is, $W_\alpha$ is a subnormal *completion* of the initial segment of weights. In [16] and [17], the first named author and L. Fialkow obtained the so-called Subnormal Completion Criterion, a concrete test that determines when a collection of initial weights $\alpha_0, \ldots, \alpha_m$ admits a subnormal completion $W_\alpha$. On the other hand, quadratically hyponormal completions require different tools, as discovered in [19].

**Theorem 3.2.** *Let $\alpha_0 < \alpha_1 < \alpha_2 < \alpha_3$ be a given collection of positive weights.*

(i) *There always exist weights $\alpha_4, \alpha_5, \ldots$ such that $W_\alpha$ is quadratically hyponormal* [28].

(ii) *There exists weights $\alpha_4, \alpha_5, \ldots$ such that $W_\alpha$ is 2-hyponormal if and only if*

$$H(2) := \begin{pmatrix} \gamma_0 & \gamma_1 & \gamma_2 \\ \gamma_1 & \gamma_2 & \gamma_3 \\ \gamma_2 & \gamma_3 & \gamma_4 \end{pmatrix} \geq 0 \quad \textit{and} \quad \begin{pmatrix} \gamma_3 \\ \gamma_4 \end{pmatrix} \in \operatorname{Ran} \begin{pmatrix} \gamma_1 & \gamma_2 \\ \gamma_2 & \gamma_3 \end{pmatrix} \quad [16].$$

In a slightly different direction, attempts have been made to characterize, for specific families of weighted shifts, the weight sequences that give rise to subnormal weighted shifts. We recall a well-known characterization of subnormality for weighted shifts due to C. Berger and independently established by R. Gellar and L.J. Wallen: $W_\alpha$ is subnormal if and only if $\gamma_n = \int t^n \, d\mu(t)$, where $\mu$ is a probability Borel measure supported in the interval $[0, \|W_\alpha\|^2]$ [9, III.8.16]. The measure $\mu$ is finitely atomic if and only if there exist scalars $\varphi_0, \ldots, \varphi_{r-1}$ such that $\gamma_{n+r} = \varphi_0 \gamma_n + \cdots + \varphi_{r-1} \gamma_{n+r-1}$ (all $n \geq 0$) [16]; we call such shifts *recursively generated.* The positivity conditions in Theorem 3.2(ii) ensure the existence of a recursively generated subnormal (or, equivalently, 2-hyponormal) completion.

In an effort to unravel how $k$-hyponormality and weak $k$-hyponormality are interrelated, researchers have looked at weighted shifts whose first few weights are unrestricted but whose tails are subnormal and recursively generated [4], [7], [34], [35], [41], [42]. A special case involves shifts whose weight sequences are of the form $x, \alpha_0, \alpha_1, \dots$, with $W_\alpha$ subnormal. Then $W_{x,\alpha}$ is subnormal if and only if $\frac{1}{t} \in L^1(\mu)$ and $x^2 \leq (\|\frac{1}{t}\|_{L^1(\mu)})^{-1}$ [14]. Thus, the subnormality of a weighted shift can be maintained if one alters the first weight slightly. The following result states that this is the only possible finite rank perturbation that preserves subnormality; quadratic hyponormality, on the other hand, is a lot more stable.

**Theorem 3.3 ([29]).**

(i) *Let $W_\alpha$ be subnormal and let $F$ ($\neq cP_{\langle e_0 \rangle}$) be a nonzero finite rank operator. Then $W_\alpha + F$ is not subnormal.*

(ii) *Let $\alpha$ be a strictly increasing weight sequence, and assume that $W_\alpha$ is 2-hyponormal. Then $W_{\alpha'}$ remains quadratically hyponormal for all $\alpha'$ such that $\alpha' - \alpha$ is a small nonzero finite rank perturbation.*

On a related matter, taking $W_\alpha$ as the restriction of the Bergman shift to the invariant subspace generated by $\{e_2, e_3, \dots\}$, it is possible to find a range of values for $x > 0$ such that $W_{x,\alpha}$ is quartically hyponormal but not 3-hyponormal [22]. On the other hand, when the weights are given by $\alpha_n := \sqrt{\frac{an+b}{cn+d}}$ $(n \geq 0)$, with $a, b, c, d \geq 0$ and $ad - bc > 0$, it was shown in [31] that $W_\alpha$ is always subnormal.

In many respects, 2-hyponormality behaves much like subnormality, particularly within the classes of unilateral weighted shifts and of Toeplitz operators on $H^2(\mathbb{T})$; for instance, a 2-hyponormal operator always leaves the kernel of its self-commutator invariant [26, Lemma 2.2]. The results in [7], [13], [14], [16], [17], [23], [20], [29] and [42] all seem to indicate the existence of a model theory for 2-hyponormal operators, with building blocks given by weighted shifts with recursive subnormal tails and Toeplitz operators with special trigonometric symbols. In [26] the beginnings of such a theory are outlined, including a connection to Agler's abstract model theory [3] – see [26, Section 5]. The proposed model theory involves a new notion, that of *weakly subnormal* operator $T$, characterized by an extension $\hat{T} \in \mathcal{L}(\mathcal{K})$ such that $\hat{T}^*\hat{T}f = \hat{T}\hat{T}^*f$ (all $f \in \mathcal{H}$); we refer to $\hat{T}$ as a *partially normal extension* of $T$.

At the level of weighted shifts, it was proved in [26, Theorem 3.1] that if $\alpha$ is strictly increasing then $W_\alpha$ is weakly subnormal precisely when $\limsup u_{n+1}/u_n < \infty$, where $u_n := \alpha_n^2 - \alpha_{n-1}^2$. This characterization allows one to show that every 2-hyponormal weighted shift is automatically weakly subnormal [26, Theorem 1.2] and that the class of weakly subnormal shifts is strictly larger than the class of 2-hyponormal shifts [26, Example 3.7]; however, there exist quadratically hyponormal weighted shifts which are not weakly subnormal [26, Example 5.5]. Moreover, it was shown in [26] that if $W_\alpha$ is 2-hyponormal, then the sequence of quotients $u_{n+1}/u_{n+2}$ is bounded, and bounded away from zero; in particular, the sequence $\{u_n\}$ is eventually decreasing. On the other hand, if $T$ is 2-hyponormal or weakly subnormal, with rank-one self-commutator, then $T$ is subnormal; if, in addition, $T$ is pure, then $T$ is unitarily equivalent to a linear function of $U_+$. Weak subnormality can also be used to characterize $k$-hyponormality, as follows: an operator $T$ is $(k+1)$-hyponormal if and only if $T$ is weakly subnormal and admits a partially normal extension $\hat{T}$ which is $k$-hyponormal [21].

All of the previous results encourage us to consider the following

**Problem 3.4.** *Develop a model theory for* 2*-hyponormality, parallel to subnormal operator theory.*

While it is easy to see that the class of 2-hyponormal contractions forms a *family* (in the sense of J. Agler), it is an open problem whether the same is true of weakly subnormal contractions [26, Question 6.5]. M. Dritschel and S. McCullough found in [33] a sufficient condition for a 2-hyponormal contraction to be extremal.

As we have mentioned before, nontrivial 2-hyponormal weighted shifts are closely related to recursively generated subnormal shifts, i.e., those shifts whose Berger measures are finitely atomic. In [20] a study of extensions of recursively generated weight sequences was done. Given a recursively generated weight sequence $(0 < \alpha_0 < \cdots < \alpha_k)$, and an $n$-step extension $\alpha : x_n, \ldots, x_1, (\alpha_0, \ldots, \alpha_k)^\wedge$, it was established that

$$W_\alpha \text{ is subnormal } \iff \begin{cases} W_\alpha \text{ is } ([\frac{k+1}{2}]+1)\text{-hyponormal} & (n=1) \\ W_\alpha \text{ is } ([\frac{k+1}{2}]+2)\text{-hyponormal} & (n>1) \end{cases}.$$

In particular, the subnormality of an extension is *independent* of its length if the length is bigger than 1. As a consequence, if $\alpha(x)$ is a canonical rank-one perturbation of the recursive weight sequence $\alpha$, then subnormality and $k$-hyponormality for $W_{\alpha(x)}$ eventually coincide! This means that the subnormality of $W_{\alpha(x)}$ can be detected after finitely many steps. Conversely, if $k$- and $(k+1)$-hyponormality for $W_{\alpha(x)}$ coincide then $\alpha(x)$ must be recursively generated, i.e., $W_{\alpha(x)}$ is a recursive subnormal.

## 4. The case of Toeplitz operators

Recall Paul Halmos's Problem 5 (cf. [38], [39]): Is every subnormal Toeplitz operator either normal or analytic? As we know, this was answered in the negative by C. Cowen and J. Long [12]. It is then natural to ask: Which Toeplitz operators are subnormal? We recall the following result.

**Theorem 4.1 ([1]).** *If*

(i) $T_\varphi$ *is hyponormal;*

(ii) $\varphi$ *or* $\bar{\varphi}$ *is of bounded type (i.e.,* $\varphi$ *or* $\bar{\varphi}$ *is a quotient of two analytic functions);*

(iii) $\ker[T_\varphi^*, T_\varphi]$ *is invariant for* $T_\varphi$*,*

*then* $T_\varphi$ *is normal or analytic.*

(We mention in passing a recent result of S.H. Lee and W.Y. Lee [46]: if $T \in \mathcal{L}(\mathcal{H})$ is a pure hyponormal operator, if $\ker[T^*, T]$ is invariant for $T$, and if

$[T^*,T]$ is rank-two, then $T$ is either a subnormal operator or Putinar's matricial model of rank two.)

Since $\ker[T^*,T]$ is invariant under $T$ for every subnormal operator $T$, Theorem 4.1 answers Problem 5 affirmatively when $\varphi$ or $\bar{\varphi}$ is of bounded type. Also, every hyponormal Toeplitz operator which is unitarily equivalent to a weighted shift must be subnormal [52], [10], a fact used in

**Theorem 4.2 ([12], [10]).** *Let $0<\alpha<1$ and let $\psi$ be a conformal map of the unit disk onto the interior of the ellipse with vertices $\pm(1+\alpha)i$ and passing through $\pm(1-\alpha)$. If $\varphi:=(1-\alpha^2)^{-1}(\psi+\alpha\bar{\psi})$, then $T_\varphi$ is a weighted shift with weight sequence $\alpha_n=(1-\alpha^{2n+2})^{-\frac{1}{2}}$. Therefore, $T_\varphi$ is subnormal but neither normal nor analytic.*

These results show that subnormality for weighted shifts and for Toeplitz operators are conceptually quite different. One then tries to answer the following

**Problem 4.3.** *Characterize subnormality of Toeplitz operators in terms of their symbols.*

Since subnormality is equivalent to $k$-hyponormality for every $k \geq 1$ (this is the Bram-Halmos Criterion), one possible strategy is to first characterize $k$-hyponormality, and then use it to characterize subnormality. As a first step, we pose the following

**Problem 4.4.** *Characterize 2-hyponormality for Toeplitz operators.*

As usual, the Toeplitz operator $T_\varphi$ on $H^2(\mathbb{T})$ with symbol $\varphi \in L^\infty(\mathbb{T})$ is given by $T_\varphi g := P(\varphi g)$, where $P$ denotes the orthogonal projection from $L^2(\mathbf{T})$ to $H^2(\mathbf{T})$. In [25, Chapter 3] the following question was considered:

**Problem 4.5.** *Is every 2-hyponormal Toeplitz operator $T_\varphi$subnormal ?*

For the case of trigonometric symbol, one has

**Theorem 4.6 ([25]).** *Every trigonometric Toeplitz operator whose square is hyponormal must be normal or analytic; in particular, every 2-hyponormal trigonometric Toeplitz operator is subnormal.*

Theorem 4.6 shows that there is a big gap between hyponormality and quadratic hyponormality for Toeplitz operators. For example, if $\varphi(z) \equiv \sum_{n=-m}^{N} a_n z^n$ $(0 < m < N)$ is such that $T_\varphi$ is hyponormal, then by Theorem 4.6, $T_\varphi$ is never quadratically hyponormal, since $T_\varphi$ is neither analytic nor normal. One can extend Theorem 4.6. First we observe

**Proposition 4.7 ([26]).** *If $T \in \mathcal{L}(\mathcal{H})$ is 2-hyponormal then $T(\ker[T^*,T]) \subseteq \ker[T^*,T]$.*

**Corollary 4.8.** *If $T_\varphi$ is 2-hyponormal and if $\varphi$ or $\bar{\varphi}$ is of bounded type then $T_\varphi$ is normal or analytic, so that $T_\varphi$ is subnormal.*

**Theorem 4.9 ([27, Theorem 8]).** *If the symbol $\varphi$ is almost analytic (i.e., $z^n\varphi$ analytic for some positive $n$), but not analytic, and if $T_\varphi$ is 2-hyponormal, then $\varphi$ must be a trigonometric polynomial.*

In [27, Lemma 9] it was shown that if $T_\varphi$ is 2-hyponormal and $\varphi = q\bar{\varphi}$, where $q$ is a finite Blaschke product, then $T_\varphi$ is normal or analytic. Moreover, [27, Theorem 10] states that when $\log|\varphi|$ is not integrable, a 2-hyponormal Toeplitz operator $T_\varphi$ with nonzero finite rank self-commutator must be analytic.

One also has

**Theorem 4.10 (cf. [27]).** *If $T_\varphi$ is 2-hyponormal and if $\varphi$ or $\bar{\varphi}$ is of bounded type (i.e., $\varphi$ or $\bar{\varphi}$ is a quotient of two analytic functions), then $T_\varphi$ is normal or analytic, so that $T_\varphi$ is subnormal.*

The following problem arises naturally:

**Problem 4.11.** *If $T_\varphi$ is a 2-hyponormal Toeplitz operator with nonzero finite rank self-commutator, does it follow that $T_\varphi$ is analytic? If so, is $\varphi$ a linear function of a finite Blaschke product?*

A partial positive answer to Problem 4.11 was given in [27, Theorem 10]. In view of Cowen and Long's counterexample [12], it is worth turning attention to hyponormality of Toeplitz operators, which has been studied extensively. An elegant theorem of C. Cowen [11] characterizes the hyponormality of a Toeplitz operator $T_\varphi$ on $H^2(\mathbf{T})$ by properties of the symbol $\varphi \in L^\infty(\mathbf{T})$. The variant of Cowen's theorem [11] that was first proposed in [48] has been most helpful. We conclude this section with a result that extends the work of Cowen and Long to 2-hyponormality and quadratic hyponormality.

**Theorem 4.12 ([23, Theorem 6]).** *Let $0 < \alpha < 1$ and let $\psi$ be the conformal map of the unit disk onto the interior of the ellipse with vertices $\pm(1+\alpha)i$ and passing through $\pm(1-\alpha)$. Let $\varphi = \psi + \lambda\bar{\psi}$ and let $T_\varphi$ be the corresponding Toeplitz operator on $H^2$. Then*

(i) *$T_\varphi$ is hyponormal if and only if $\lambda$ is in the closed unit disk $|\lambda| \leq 1$.*

(ii) *$T_\varphi$ is subnormal if and only if $\lambda = \alpha$ or $\lambda$ is in the circle $\left|\lambda - \frac{\alpha(1-\alpha^{2k})}{1-\alpha^{2k+2}}\right| = \frac{\alpha^k(1-\alpha^2)}{1-\alpha^{2k+2}}$ for $k = 0, 1, 2, \ldots$. (Observe that the case $\lambda = \alpha$ is part of the main result in [12].)*

(iii) *$T_\varphi$ is 2-hyponormal if and only if $\lambda$ is in the unit circle $|\lambda| = 1$ or in the closed disk $\left|\lambda - \frac{\alpha}{1+\alpha^2}\right| \leq \frac{\alpha}{1+\alpha^2}$.*

(iv) ([45]) *$T_\varphi$ is 2-hyponormal if and only if $T_\varphi$ is quadratically hyponormal.*

## References

[1] M.B. Abrahamse, Subnormal Toeplitz operators and functions of bounded type, *Duke Math. J.* 43 (1976), 597–604.

[2] J. Agler, Hypercontractions and subnormality, *J. Operator Th.* 13 (1985), no. 2, 203–217.

[3] J. Agler, An abstract approach to model theory, in *Surveys of Recent Results in Operator Theory*, vol. II (J.B. Conway and B.B. Morrel, eds.), Pitman Res. Notes Math. Ser., vol. 192, Longman Sci. Tech., Harlow, 1988, pp. 1–23.

[4] J.Y. Bae, G. Exner and I.B. Jung, Criteria for positively quadratically hyponormal weighted shifts, *Proc. Amer. Math. Soc.* 130 (2002), 3287–3294

[5] G. Cassier, Problème des moments sur un compact de $\mathbb{R}^n$ et décomposition de polynômes à plusieurs variables*, J. Funct. Anal.* 58 (1984), no. 3, 254–266.

[6] Y.B. Choi, A propagation of quadratically hyponormal weighted shifts. *Bull. Korean Math. Soc.* 37 (2000), 347–352.

[7] Y.B. Choi, J.K. Han and W.Y. Lee, One-step extension of the Bergman shift, *Proc. Amer. Math. Soc.* 128(2000), 3639–3646.

[8] I. Colojoara and C. Foiaş, *Theory of Generalized Spectral Operators*, Mathematics and its Applications, vol. 9, Gordon and Breach, Science Publishers, New York-London-Paris, 1968.

[9] J.B. Conway, *The Theory of Subnormal Operators*, Mathematical Surveys and Monographs, vol. 36, American Mathematical Society, Providence, RI, 1991.

[10] C. Cowen, More subnormal Toeplitz operators, *J. Reine Angew. Math.* 367(1986), 215–219.

[11] C. Cowen, Hyponormal and subnormal Toeplitz operators, in *Surveys of Some Recent Results in Operator Theory,* Vol. I (J.B. Conway and B.B. Morrel, eds.), Pitman Res. Notes in Math., vol. 171, Longman Publ. Co. 1988, pp. 155–167.

[12] C. Cowen and J. Long, Some subnormal Toeplitz operators, *J. Reine Angew. Math.* 351(1984), 216–220.

[13] R. Curto, Joint hyponormality: A bridge between hyponormality and subnormality, *Proc. Symposia Pure Math.* 51(1990), 69–91.

[14] R. Curto, Quadratically hyponormal weighted shifts, *Integral Equations Operator Theory* 13(1990), 49–66.

[15] R. Curto, Polynomially hyponormal operators on Hilbert space, in *Proceedings of ELAM VII*, *Revista Unión Mat. Arg.* 37(1991), 29–56.

[16] R. Curto and L. Fialkow, Recursively generated weighted shifts and the subnormal completion problem, *Integral Equations Operator Theory* 17(1993), 202–246.

[17] R. Curto and L. Fialkow, Recursively generated weighted shifts and the subnormal completion problem, II, *Integral Equations Operator Theory*, 18(1994), 369–426.

[18] R. Curto, I.S. Hwang and W.Y. Lee, Weak subnormality of operators, *Arch. Math.* 79(2002), 360–371.

[19] R. Curto and I.B. Jung, Quadratically hyponormal shifts with two equal weights, *Integral Equations Operator Theory* 37(2000), 208–231.

[20] R. Curto, I.B. Jung and W.Y. Lee, Extensions and extremality of recursively generated weighted shifts, *Proc. Amer. Math. Soc.* 130(2002), 565–576.

[21] R. Curto, I.B. Jung and S.S. Park, A characterization of $k$-hyponormality via weak subnormality, *J. Math. Anal. Appl.* 279(2003), 556–568.

[22] R. Curto and S.H. Lee, Quartically hyponormal weighted shifts need not be 3-hyponormal, *J. Math. Anal. Appl.* 314 (2006), 455–463.

[23] R. Curto, S.H. Lee and W.Y. Lee, Subnormality and 2-hyponormality for Toeplitz operators, *Integral Equations Operator Theory* 44(2002), 138–148.

[24] R. Curto, S.H. Lee and W.Y. Lee, A new criterion for $k$-hyponormality via weak subnormality, *Proc. Amer. Math. Soc.* 133(2005), 1805–1816.

[25] R. Curto and W.Y. Lee, Joint hyponormality of Toeplitz pairs, *Memoirs Amer. Math. Soc.* 150, no. 712, Amer. Math. Soc., Providence, 2001.

[26] R. Curto and W.Y. Lee, Towards a model theory for 2-hyponormal operators, *Integral Equations Operator Theory* 44(2002), 290–315.

[27] R. Curto and W.Y. Lee, Subnormality and $k$-hyponormality of Toeplitz operators: A brief survey and open questions, in *Operator Theory and Banach Algebras*, The Theta Foundation, Bucharest, 2003; pp. 73–81.

[28] R. Curto and W.Y. Lee, Solution of the quadratically hyponormal completion problem, *Proc. Amer. Math. Soc.* 131(2003), 2479–2489.

[29] R. Curto and W.Y. Lee, $k$-hyponormality of finite rank perturbations of unilateral weighted shifts, *Trans. Amer. Math. Soc.* 357(2005), 4719–4737.

[30] R. Curto and S.S. Park, $k$-hyponormality of powers of weighted shifts via Schur products, *Proc. Amer. Math. Soc.* 131(2003), 2761–2769.

[31] R. Curto, Y.T. Poon and J. Yoon, Subnormality of Bergman-like weighted shifts, *J. Math. Anal. Appl.* 308(2005), 334–342.

[32] R. Curto and M. Putinar, Nearly subnormal operators and moment problems, *J. Funct. Anal.* 115 (1993), no. 2, 480–497.

[33] M.A. Dritschel and S. McCullough, Model theory for hyponormal contractions, *Integral Equations Operator Theory* 36(2000), 182–192.

[34] G. Exner, I.B. Jung and D. Park, Some quadratically hyponormal weighted shifts, *Integral Equations Operator Theory* 60 (2008), 13–36.

[35] G. Exner, I.B. Jung and S.S. Park, Weakly $n$-hyponormal weighted shifts and their examples, *Integral Equations Operator Theory* 54 (2006), 215–233.

[36] P.R. Halmos, Normal dilations and extensions of operators, *Summa Brasil. Math.* 2, (1950). 125–134.

[37] P.R. Halmos, *A Hilbert Space Problem Book*, Second edition, Graduate Texts in Mathematics, vol. 19, Springer-Verlag, New York-Berlin, 1982.

[38] P.E. Halmos, Ten problems in Hilbert space, *Bull. Amer. Math. Soc.* 76(1970), 887–933.

[39] P.R. Halmos, Ten years in Hilbert space, *Integral Equations Operator Theory* 2(1979), 529–564.

[40] J.W. Helton and M. Putinar, Positive polynomials in scalar and matrix variables, the spectral theorem, and optimization, in *Operator Theory, Structured Matrices, and Dilations*, pp. 229–306, Theta Ser. Adv. Math. 7, Theta, Bucharest, 2007.

[41] T. Hoover, I.B. Jung, and A. Lambert, Moment sequences and backward extensions of subnormal weighted shifts, *J. Austral. Math. Soc.* 73 (2002), 27–36.

[42] I. Jung and C. Li, A formula for $k$-hyponormality of backstep extensions of subnormal weighted shifts, *Proc. Amer. Math. Soc.* 129 (2001), 2343–2351.

[43] I. Jung and S. Park, Quadratically hyponormal weighted shifts and their examples, *Integral Equations Operator Theory* 36(2000), 480–498.

[44] I. Jung and S. Park, Cubically hyponormal weighted shifts and their examples, *J. Math. Anal. Appl.* 247 (2000), 557–569.

[45] S.H. Lee and W.Y. Lee, Quadratic hyponormality and 2-hyponormality for Toeplitz operators, *Integral Equations Operator Theory* 54 (2006), 597–602.

[46] S.H. Lee and W.Y. Lee, Hyponormal operators with rank-two self-commutators, *J. Math. Anal. Appl.* 351(2009), 616–626.

[47] S. McCullough and V. Paulsen, A note on joint hyponormality, *Proc. Amer. Math. Soc.* 107(1989), 187–195.

[48] T. Nakazi and K. Takahashi, Hyponormal Toeplitz operators and extremal problems of Hardy spaces, *Trans. Amer. Math. Soc.* 338 (1993), 753–767.

[49] M. Putinar, Positive polynomials on compact semi-algebraic sets, *Indiana Univ. Math.* J. 42 (1993), 969–984.

[50] F. Riesz and B. Sz.-Nagy, *Functional Analysis*, Dover, New York, 1990.

[51] J. Stampfli, Which weighted shifts are subnormal? *Pacific J. Math.* 17(1966), 367–379.

[52] S. Sun, Bergman shift is not unitarily equivalent to a Toeplitz operator, *Kexue Tongbao* (English Ed.) 28 (1983), 1027–1030.

Raúl Curto
Department of Mathematics
University of Iowa
Iowa City, IA 52246, USA
e-mail: rcurto@math.uiowa.edu

Mihai Putinar
Department of Mathematics
University of California
Santa Barbara, CA 93106, USA
e-mail: mputinar@math.ucsb.edu

Operator Theory:
Advances and Applications, Vol. 207, 209–222
© 2010 Springer Basel AG

# Essentially Normal Operators

Kenneth R. Davidson

**Abstract.** This is a survey of essentially normal operators and related developments. There is an overview of Weyl–von Neumann theorems about expressing normal operators as diagonal plus compact operators. Then we consider the Brown–Douglas–Fillmore theorem classifying essentially normal operators. Finally we discuss almost commuting matrices, and how they were used to obtain two other proofs of the BDF theorem.

**Mathematics Subject Classification (2000).** 47-02,47B15,46L80.

**Keywords.** Essentially normal operator, compact perturbation, normal, diagonal, almost commuting matrices, extensions of C*-algebras.

## 1. Introduction

Problem 4 of Halmos's *Ten problems in Hilbert space* [**26**] asked whether every normal operator is the sum of a diagonal operator and a compact operator. I believe that this was the first of the ten problems to be solved. Indeed two solutions were independently produced by Berg [**6**] and Sikonia [**39**] almost immediately after dissemination of the question. But that is only the beginning of the story, as, like many of Paul's problems, the answer to the question is just a small step in the bigger picture.

The subsequent discussion in Halmos's article goes in several directions. In particular, he discusses operators of the form normal plus compact. It is not at all clear, *a priori*, if this set is even norm closed. It turned out that the characterization of this class by other invariants is very interesting.

It is immediately clear that if $T$ is normal plus compact, then $T^*T - TT^*$ is compact. Operators with this latter property are called *essentially normal*. Not all essentially normal operators are normal plus compact. For example, the unilateral shift $S$ acts on a basis $\{e_n : n \geq 0\}$ by $Se_n = e_{n+1}$. It evidently satisfies

$$S^*S - SS^* = I - SS^* = e_0e_0^*,$$

Partially supported by an NSERC grant.

which is the rank one projection onto $\mathbb{C}e_0$. Thus the unilateral shift is essentially normal. However it has non-zero Fredholm index:

$$\operatorname{ind} S = \dim\ker S - \dim\ker S^* = -1.$$

For a normal operator $N$, one has $\ker N = \ker N^*$. Thus if $N$ is Fredholm, then $\operatorname{ind} N = 0$. Index is invariant under compact perturbations, so the same persists for normal plus compact operators.

Brown, Douglas and Fillmore [**9**] raised the question of classifying essentially normal operators. The answer took them from a naive question in operator theory to the employment of new techniques from algebraic topology in the study of C*-algebras. They provided a striking answer to the question of which essentially normal operators are normal plus compact. They are precisely those essentially normal operators with the property that $\operatorname{ind}(T - \lambda I) = 0$ for every $\lambda \notin \sigma_e(T)$, namely the index is zero whenever it is defined. This implies, in particular, that the set of normal plus compact operators is norm closed.

Had they stopped there, BDF might have remained 'just' a tour de force that solved an interesting question in operator theory. However they recognized that their methods had deeper implications about the connection between topology and operator algebras. They defined an invariant $\operatorname{Ext}(\mathfrak{A})$ for any C*-algebra $\mathfrak{A}$, and determined nice functorial properties of this object in the case of separable, commutative C*-algebras. They showed that Ext has a natural pairing with the topological K-theory of Atiyah [**4**] which makes Ext a K-homology theory. This opened up a whole new world for C*-algebraists, and a new game was afoot.

At roughly the same time, George Elliott [**18**] introduced a complete algebraic invariant for AF C*-algebras. These C*-algebras, introduced by Bratteli [**8**], are defined by the property that they are the closure of an (increasing) union of finite-dimensional sub-algebras. It was soon recognized [**19**] that this new invariant is the $K_0$ functor from ring theory. A very striking converse to Elliott's theorem was found by Effros, Handelmann and Shen [**17**] which characterized those groups which arise as the $K_0$ group of an AF algebra.

The upshot was that two very different results almost simultaneously seeded the subject of C*-algebras with two new topological tools that provide interesting new invariants, namely Ext and $K_0$. These results created tremendous excitement, and launched a program which continues to this day to classify amenable C*-algebras. It revitalized the subject, and has led to a sophisticated set of tools which describe and distinguish many new algebras. It is fair to say that the renaissance of C*-algebras was due to these two developments. Indeed, not only are they related by the spirit of K-theory, they are in fact two sides of the same coin. Kasparov [**30**] introduced his bivariant KK-theory shortly afterwards which incorporates the two theories into one.

It is not my intention to survey the vast literature in C*-algebras which has developed as a consequence of the introduction of K-theory. I mention it to highlight the fallout of the pursuit of a natural problem in operator theory by three very insightful investigators. I will limit the balance of this article to the original operator theory questions, which have a lot of interest in their own right.

## 2. Weyl–von Neumann theorems

In 1909, Hermann Weyl [**46**] proved that every self-adjoint bounded operator $A$ on a *separable* Hilbert space can be written as $A = D + K$ where $D$ is a diagonal operator with respect to some orthonormal basis and $K$ is compact. Hilbert's spectral theorem for Hermitian operators says, as formulated by Halmos [**27**], that every Hermitian operator on Hilbert space is unitarily equivalent to a multiplication operator $M_\varphi$ on $L^2(\mu)$, where $\varphi$ is a bounded real-valued measurable function. The starting point of a proof of Weyl's theorem is the observation that if $f$ is any function in $L^2(\mu)$ supported on $\varphi^{-1}\big([t, t+\varepsilon)\big)$, then $f$ is almost an eigenvector in the sense that $\|M_\varphi f - tf\| < \varepsilon\|f\|$. One can carefully extract an orthonormal basis for $L^2(\mu)$ consisting of functions with increasingly narrow support.

To make this more precise, suppose for convenience that $A$ has a cyclic vector. Then the spectral theorem produces a measure $\mu$ on the spectrum $\sigma(A) \subset \mathbb{R}$ so that $A$ is identified with $M_x$ on $L^2(\mu)$. Let $P_n$ be the span of the characteristic functions of diadic intervals of length $2^{-n}$. Then the previous observation can be used to show that $\|P_nA - AP_n\| < 2^{-n}$. So a routine calculation shows that $A$ is approximated within $2^{1-n}$ by the operator

$$D_n = P_nAP_n + \sum_{k\geq n}(P_{k+1} - P_k)A(P_{k+1} - P_k),$$

and $A - D_n$ is compact. The operator $D_n$ is a direct sum of finite rank self-adjoint operators, and so is diagonalizable—providing the desired approximant.

Weyl observed that if $A$ and $B$ are two Hermitian operators such that $A - B$ is compact, then the limit points of the spectrum of $A$ and $B$ must be the same. We now interpret this by saying that the essential spectra are equal, $\sigma_e(A) = \sigma_e(B)$, where $\sigma_e(\cdot)$ denotes the spectrum of the image in the Calkin algebra $\mathcal{B}(\mathcal{H})/\mathfrak{K}$. Von Neumann [**45**] established the converse: if two Hermitian operators have the same essential spectrum, then they are unitarily equivalent modulo a compact perturbation.

Halmos's questions asks for an extension to normal operators. It seems to require a new approach, because the trick of compressing a Hermitian operator to the range of an almost commuting projection $P_{k+1} - P_k$ yields a Hermitian matrix, but the same argument fails for normal operators. Also the spectrum is now a subset of the plane. David Berg [**6**] nevertheless answered Halmos's question by adapting this method. Sikonia gave a similar proof at the same time in his doctoral thesis (see [**39**]). Other proofs came quickly afterwards (for example Halmos [**27**]).

A proof that does the job simultaneously for a countable family of commuting Hermitian operators $\{A_i\}$ works by building a single Hermitian operator $A$ so that $\mathrm{C}^*(A)$ contains every $A_i$. To accomplish this, first observe that the spectral theorem shows that $\mathrm{C}^*(\{A_i\})$ is contained in a commutative $\mathrm{C}^*$-algebra spanned by a countable family $\{E_j\}$ of commuting projections. Consider the Hermitian operator $A = \sum_{j\ge 1} 3^{-j} E_j$. It is easy to see that

$$\tfrac{1}{3}E_1 \le E_1 A \le \tfrac{1}{2}E_1 \quad \text{and} \quad 0 \le E_1^\perp A \le \tfrac{1}{6}E_1^\perp .$$

Thus it follows that the spectrum of $A$ is contained in $[0,1/6]\cup[1/3,1/2]$, and that the spectral projection for $[1/3,1/2]$ is $E_1$. Similarly, each projection $E_j$ belongs to $\mathrm{C}^*(A)$. It follows that each $A_i$ is a continuous function of $A$. Diagonalizing $A$ modulo compacts then does the same for each $A_i$, although one cannot control the norm of the compact perturbations for more than a finite number at a time.

At this point, we introduce a few definitions to aid in the discussion.

**Definition 2.1.** Let $\mathfrak{J}$ be a normed ideal of $\mathcal{B}(\mathcal{H})$. Two operators $A$ and $B$ are *unitarily equivalent modulo* $\mathfrak{J}$ if there is a unitary $U$ so that $A - UBU^* \in \mathfrak{J}$. We say that $A$ and $B$ are *approximately unitarily equivalent modulo* $\mathfrak{J}$ if there is a sequence of unitaries $U_k$ so that $A - U_k B U_k^* \in \mathfrak{J}$ and

$$\lim_{k\to\infty} \|A - U_k B U_k^*\|_{\mathfrak{J}} = 0.$$

We write $A \sim_{\mathfrak{J}} B$. When $\mathfrak{J} = \mathcal{B}(\mathcal{H})$, we simply say that $A$ and $B$ are *approximately unitarily equivalent* and write $A \sim_a B$.

A careful look at Weyl's proof shows that the perturbation will lie in certain smaller ideals, the Schatten classes $\mathfrak{S}_p$ with norm $\|K\|_p = \mathrm{Tr}(|K|^p)^{1/p}$, provided that $p > 1$; and this norm can also be made arbitrarily small. Kuroda [**32**] improved on this to show that one can obtain a small perturbation in any unitarily invariant ideal $\mathfrak{J}$ that strictly contains the trace class operators. So any Hermitian operator is approximately unitarily equivalent modulo $\mathfrak{J}$ to a diagonalizable operator. Berg's proof works in a similar way via a process of dividing up the plane, and actually yields small perturbations in $\mathfrak{S}_p$ for $p > 2$. Likewise, for a commuting $n$-tuple, one can obtain small perturbations in $\mathfrak{S}_p$ for $p > n$. It is natural to ask whether this is sharp. Elementary examples show that one cannot obtain perturbations in $\mathfrak{S}_p$ when $p < n$.

In the case of $n = 1$, there is an obstruction found by Kato [**31**] and Rosenblum [**37**]. If the spectral measure of $A$ is not singular with respect to the Lebesgue measure, then there is no trace class perturbation which is diagonal. In particular, $M_x$ on $L^2(0,1)$ is such an operator. For $n \ge 2$, Voiculescu [**41, 42**] showed that every commuting $n$-tuple of Hermitian operators is approximately unitarily equivalent to a diagonalizable $n$-tuple modulo the Schatten class $\mathfrak{S}_n$. (See [**14**] for an elementary argument.) Moreover, Voiculescu identified a somewhat smaller ideal $\mathfrak{S}_n^-$ which provides an obstruction when the $n$-tuple has a spectral measure that is not singular with respect to Lebesgue measure on $\mathbb{R}^n$. Bercovici and Voiculescu [**5**] strengthened this to the analogue of Kuroda's theorem, showing that if a unitarily invariant ideal is not included in $\mathfrak{S}_n^-$, then a small perturbation to a diagonal operator is possible.

The ideas involved in Voiculescu's work mentioned above build on a very important theorem of his that preceded these results, and had a direct bearing on the work of Brown, Douglas and Fillmore and on subsequent developments in C*-algebras. This is known as Voiculescu's Weyl–von Neumann Theorem [**40**]. Rather than state it in full generality, we concentrate on some of its major corollaries. The definition of approximate unitary equivalence can readily be extended to two maps from a C*-algebra $\mathfrak{A}$ into $\mathcal{B}(\mathcal{H})$: say $\rho \sim_a \sigma$ for two maps $\rho$ and $\sigma$ if there is a sequence of unitary operators $U_n$ such that

$$\lim_{n\to\infty} \|\rho(a)U_n - U_n\sigma(a)\| = 0 \quad \text{for all} \quad a \in \mathfrak{A}.$$

One similarly defines approximate unitary equivalence relative to an ideal.

Voiculescu showed that if $\mathfrak{A}$ is a separable C*-algebra acting on $\mathcal{H}$ and $\rho$ is a $*$-representation which annihilates $\mathfrak{A}\cap\mathfrak{K}$, then $\mathrm{id} \sim_{\mathfrak{K}} \mathrm{id}\oplus\rho$, where id is the identity representation of $\mathfrak{A}$. Let $\pi$ denote the quotient map of $\mathcal{B}(\mathcal{H})$ onto the Calkin algebra $\mathcal{B}(\mathcal{H})/\mathfrak{K}$. Voiculescu's result extends to show for two representations $\rho_1$ and $\rho_2$, one has $\rho_1 \sim_a \rho_2$ if and only if $\rho_1 \sim_{\mathfrak{K}} \rho_2$. In particular, this holds provided that

$$\ker\rho_1 = \ker\pi\rho_1 = \ker\rho_2 = \ker\pi\rho_2.$$

If $N$ is normal, then it may have countably many eigenvalues of finite multiplicity which do not lie in the essential spectrum. However they must be asymptotically close to the essential spectrum. One can peel off a finite diagonalizable summand, and make a small compact perturbation on the remainder to move the other eigenvalues into $\sigma_e(N)$. One now has a normal operator $N'$ with $\sigma(N') = \sigma_e(N')$. If one applies Voiculescu's Theorem to $\mathrm{C}^*(N')$, one recovers the Weyl–von Neumann–Berg Theorem.

Another consequence of this theorem was a solution to Problem 8 of Halmos's ten problems, which asked whether the reducible operators are dense in $\mathcal{B}(\mathcal{H})$. One merely takes any representation $\rho$ of $\mathrm{C}^*(T)$ which factors through $\mathrm{C}^*(T) + \mathfrak{K}/\mathfrak{K}$ to obtain $T \sim_a T \oplus \rho(T)$. Moreover, one can show that $\mathrm{id} \sim_{\mathfrak{K}} \sigma$ where $\sigma$ is a countable direct sum of irreducible representations.

The implications of Voiculescu's theorem for essentially normal operators will be considered in the next section. We mention here that a very insightful treatment of Voiculescu's theorem is contained in Arveson's paper [**3**]. In particular, it provides a strengthening of the results for normal operators. Hadwin [**25**] contains a further refinement which shows that $\rho_1 \sim_{\mathfrak{K}} \rho_2$ if and only if $\operatorname{rank}\rho_1(a) = \operatorname{rank}\rho_2(a)$ for all $a \in \mathfrak{A}$. All of these ideas are treated in Chapter 2 of [**15**].

## 3. Essentially normal operators

We return to the problem of classifying essentially normal operators. Let $T$ be essentially normal. Then $t = \pi(T)$ is a normal element of the Calkin algebra. So $\mathrm{C}^*(t) \simeq \mathrm{C}(X)$ where $X = \sigma(t) = \sigma_e(T)$. This determines a $*$-monomorphism $\tau$ of $\mathrm{C}(X)$ into $\mathcal{B}(\mathcal{H})/\mathfrak{K}$ determined by $\tau(z) = t$. Evidently $\tau$ determines $T$ up to a compact perturbation. Two essentially normal operators $T_1$ and $T_2$ are unitarily equivalent modulo $\mathfrak{K}$ if and only if $\sigma_e(T_1) = \sigma_e(T_2) =: X$ and the associated monomorphisms $\tau_1$ and $\tau_2$ of $\mathrm{C}(X)$ are strongly unitarily equivalent, meaning that there is a unitary $U$ so that $\mathrm{ad}\,\pi(U)\tau_1 = \tau_2$. (The weak version would allow equivalence by a unitary in the Calkin algebra. This turns out to be equivalent for commutative C*-algebras.) A monomorphism $\tau$ is associated to an *extension* of the compact operators. Let $\mathfrak{E} = \pi^{-1}(\tau(\mathrm{C}(X)) = \mathrm{C}^*(T) + \mathfrak{K}$. Then $\tau^{-1}\pi$ is a $*$-homomorphism of the C*-algebra $\mathfrak{E}$ onto $\mathrm{C}(X)$ with kernel $\mathfrak{K}$. We obtain the short exact sequence

$$0 \to \mathfrak{K} \to \mathfrak{E} \to \mathrm{C}(X) \to 0.$$

For example, the unilateral shift $S$ is unitarily equivalent to the Toeplitz operator of multiplication by $z$ on $H^2$. So one readily sees that $\mathrm{C}^*(S)$ is unitarily equivalent to the Toeplitz C*-algebra

$$\mathfrak{T}(\mathbb{T}) = \{T_f + K : f \in \mathrm{C}(\mathbb{T}) \text{ and } K \in \mathfrak{K}\}.$$

The map $\tau_1(f) = \pi(T_f)$ is a monomorphism of $\mathrm{C}(\mathbb{T})$ into the Calkin algebra. It is not hard to use the Fredholm index argument to show that this extension does not split; i.e., there is no $*$-monomorphism of $\mathrm{C}(\mathbb{T})$ into $\mathfrak{T}$ taking $z$ to a compact perturbation of $T_z$. Therefore this extension is not equivalent to the representation of $\mathrm{C}(\mathbb{T})$ on $L^2(\mathbb{T})$ given by $\tau_2(f) = \pi(M_f)$ where $M_f$ is the multiplication operator.

Turning the problem around, Brown, Douglas and Fillmore consider the class of all $*$-monomorphisms of $\mathrm{C}(X)$ into $\mathcal{B}(\mathcal{H})/\mathfrak{K}$ for any compact metric space $X$; or equivalently they consider all extensions of $\mathfrak{K}$ by $\mathrm{C}(X)$. Two extensions are called equivalent if the corresponding monomorphisms are strongly unitarily equivalent. The collection of all equivalence classes of extensions of $\mathrm{C}(X)$ is denoted by $\mathrm{Ext}(X)$. One can turn this into a commutative semigroup by defining $[\tau_1]+[\tau_2] = [\tau_1 \oplus \tau_2]$, which uses the fact that we can identify the direct sum $\mathcal{H}\oplus\mathcal{H}$ of two separable Hilbert spaces with the original space $\mathcal{H}$.

More generally, one can define $\mathrm{Ext}(\mathfrak{A})$ for any C*-algebra. The theory works best if one sticks to separable C*-algebras. Among these, things work out particularly well when $\mathfrak{A}$ is nuclear.

The Weyl–von Neumann–Berg Theorem is exactly what is needed to show that $\mathrm{Ext}(X)$ has a zero element. An extension is *trivial* if it splits, i.e., there is a $*$-monomorphism $\sigma$ of $\mathrm{C}(X)$ into $\mathcal{B}(\mathcal{H})$ so that $\tau = \pi\sigma$. The generators of $\sigma(\mathrm{C}(X))$ can be perturbed by compact operators to commuting diagonal operators. The converse of von Neumann is adapted to show that any two trivial extensions are equivalent. An elementary argument can be used to construct approximate eigenvectors. Repeated application yields $\tau \sim_{\mathfrak{K}} \tau\oplus\sigma$, where $\sigma$ is a trivial extension. So the equivalence class of all trivial extensions forms a zero element for $\mathrm{Ext}(X)$.

In fact, $\mathrm{Ext}(X)$ is a group. Brown, Douglas and Fillmore gave a complicated proof, which required a number of topological lemmas. The proof was significantly simplified by Arveson **[2]** by pointing out the crucial role of completely positive maps. The map $\tau$ may be lifted to a completely positive unital map $\sigma$ into $\mathcal{B}(\mathcal{H})$, meaning that $\tau = \pi\sigma$. Then the Naimark dilation theorem dilates this map to a

$*$-representation $\rho$ of $\mathrm{C}(X)$ on a larger space $\mathcal{K} \supset \mathcal{H}$; say

$$\rho(f) = \begin{bmatrix} \tau(f) & \rho_{12}(f) \\ \rho_{21}(f) & \rho_{22}(f) \end{bmatrix}.$$

Since the range of $\tau$ commutes modulo compacts, it is not hard to see that the ranges of $\rho_{12}$ and $\rho_{21}$ consist of compact operators. It follows that $\pi\rho_{22}$ is a $*$-homomorphism of $\mathrm{C}(X)$. The map $\pi\rho_{22} \oplus \tau_0$, where $\tau_0$ is any trivial extension, yields an inverse.

More generally, Choi and Effros [**12**] showed that $\mathrm{Ext}(\mathfrak{A})$ is a group whenever $\mathfrak{A}$ is a separable nuclear C*-algebra. The argument uses nuclearity and the structure of completely positive maps to accomplish the lifting. The dilation follows from Stinespring's theorem for completely positive maps. Voiculescu's Theorem provides the zero element consisting of the class of trivial extensions. (See Arveson [**3**] where all of this is put together nicely.) When $\mathfrak{A}$ is not nuclear, Anderson [**1**] showed that $\mathrm{Ext}(\mathfrak{A})$ is generally not a group. Recently Haagerup and Thorbjornsen [**24**] have shown that Ext of the reduced C*-algebra of the free group $\mathbb{F}_2$ is not a group.

Next we observe that Ext is a covariant functor from the category of compact metric spaces with continuous maps into the category of abelian groups. Suppose that $p : X \to Y$ is a continuous map between compact metric spaces, and $\tau$ is an extension of $\mathrm{C}(X)$. Build an extension of $\mathrm{C}(Y)$ by fixing a trivial extension $\sigma_0$ in $\mathrm{Ext}(Y)$ and defining

$$\sigma(f) = \tau(f \circ p) \oplus \sigma_0(f) \quad \text{for all} \quad f \in \mathrm{C}(Y).$$

So far, we have seen little topology, although the original BDF proof used more topological methods to establish these facts. Now we discuss some of those aspects which are important for developing Ext as a homology theory. If $A$ is a closed subset of $X$, $j$ is the inclusion map and $p$ is the quotient map of $X$ onto $X/A$, then

$$\mathrm{Ext}(A) \xrightarrow{j_*} \mathrm{Ext}(X) \xrightarrow{p_*} \mathrm{Ext}(X/A)$$

is exact. Ext also behaves well with respect to projective limits of spaces. If $X_n$ are compact metric spaces and $p_n : X_{n+1} \to X_n$ for $n \geq 1$, define $X = \operatorname{proj\,lim} X_n$ to be the subset of $\prod_{n\geq 1} X_n$ consisting of sequences $(x_n)$ such that $p_n(x_{n+1}) = x_n$ for all $n \geq 1$. There are canonical maps $q_n : X \to X_n$ so that $q_n = p_n q_{n+1}$. One can likewise define $\operatorname{proj\,lim} \mathrm{Ext}(X_n)$. Since $q_{n*}$ defines a compatible sequence of homomorphisms of $\mathrm{Ext}(X)$ into $\mathrm{Ext}(X_n)$, one obtains a natural map

$$\kappa : \mathrm{Ext}(\operatorname{proj\,lim} X_n) \longrightarrow \operatorname{proj\,lim} \mathrm{Ext}(X_n).$$

The key fact is that this map is always surjective. Moreover, it is an isomorphism when each $X_n$ is a finite set. This latter fact follows by noting that when each $X_n$ is finite, $X$ is totally disconnected. From our discussion of the Weyl–von Neumann Theorem, $\mathrm{C}(X)$ is generated by a single self-adjoint element, and every extension is diagonalizable and hence trivial. So $\mathrm{Ext}(X) = \{0\}$.

In [**10**], it is shown that any covariant functor from compact metric spaces into abelian groups satisfying the properties established in the previous paragraph is a homotopy invariant. That is, if $f$ and $g$ are homotopic maps from $X$ to $Y$, then $f_* = g_*$. In particular, if $X$ is contractible, then $\operatorname{Ext}(X) = \{0\}$.

There is a pairing between $\operatorname{Ext}(X)$ and $K^1(X)$ which yields a map into the integers based on Fredholm index.

First consider the group $\pi^1(X) = \mathrm{C}(X)^{-1}/\mathrm{C}(X)_0^{-1}$. An element $[f]$ of $\pi^1(X)$ is a homotopy class of invertible functions on $X$. Thus if $[\tau] \in \operatorname{Ext}(X)$, one can compute $\operatorname{ind}\tau(f)$ independent of the choice of representatives. Moreover, this determines a homomorphism $\gamma[\tau]$ of $\pi^1(X)$ into $\mathbb{Z}$. Hence we have defined a homomorphism

$$\gamma : \operatorname{Ext}(X) \to \operatorname{Hom}(\pi^1(X), \mathbb{Z}).$$

Similarly, by considering the induced monomorphisms of $\mathfrak{M}_n(\mathrm{C}(X))$ into $\mathcal{B}(\mathcal{H}^{(n)})/\mathfrak{K}$, we can define an analogous map

$$\gamma^n : \operatorname{Ext}(X) \to \operatorname{Hom}(\mathrm{GL}_n(X)/\mathrm{GL}_n(X)_0, \mathbb{Z}).$$

The direct limit of the sequence of groups $\mathrm{GL}_n(X)/\mathrm{GL}_n(X)_0$ is called $K^1(X)$. The inductive limit of $\gamma^n$ is a homomorphism $\gamma^\infty : \operatorname{Ext}(X) \to \operatorname{Hom}(K^1(X), \mathbb{Z})$. This is the *index map*, and is the first pairing of Ext with $K$-theory.

When $X$ is a planar set, an elementary argument shows that $\pi^1(X)$ is a free abelian group with generators $[z - \lambda_i]$ where one chooses a point $\lambda_i$ in each bounded component of $\mathbb{C} \setminus X$. An extension $\tau$ is determined by the essentially normal element $t = \tau(z) \in \mathcal{B}(\mathcal{H})/\mathfrak{K}$. The index map is given by

$$\gamma([\tau])([z - \lambda_i]) = \operatorname{ind}(t - \lambda_i).$$

The Brown–Douglas–Fillmore Theorem classifies essentially normal operators by showing that when $X$ is planar, the map $\gamma$ is an isomorphism. The hard part of the proof is a lemma that shows that when $\gamma[\tau] = 0$, one can cut the spectrum in half by a straight line, and split $\tau$ as the sum of two elements coming from Ext of the two halves. Repeated bisection eventually eliminates all of the holes.

An immediate consequence is the fact that an essentially normal operator $T$ is normal plus compact if and only if $\operatorname{ind}(T - \lambda I) = 0$ whenever $\lambda \notin \sigma_e(T)$. More generally, two essentially normal operators $T_1$ and $T_2$ are unitarily equivalent modulo $\mathfrak{K}$ if and only if $\sigma_e(T_1) = \sigma_e(T_2) =: X$ and

$$\operatorname{ind}(T_1 - \lambda I) = \operatorname{ind}(T_2 - \lambda I) \text{ for all } \lambda \in \mathbb{C} \setminus X.$$

Another immediate corollary is that the set of normal plus compact operators is norm closed, since the essentially normal operators are closed, the set of Fredholm operators is open, and index is continuous.

More information on the BDF theory with an emphasis on the K-theoretical aspects is contained in the monograph [**16**] by Douglas. Most of the results mentioned here are treated in Chapter 9 of [**15**].

## 4. Almost commuting matrices

Peter Rosenthal [**38**] asked whether nearly commuting matrices are close to commuting. To make sense of this question, one says that $A$ and $B$ nearly commute if $\|AB - BA\|$ is small, while close to commuting means that there are *commuting* matrices $A'$ and $B'$ with $\|A - A'\|$ and $\|B - B'\|$ both small. He makes it clear that to be an interesting problem, one must obtain estimates independent of the dimension of the space on which the matrices act. We may also limit $A$ and $B$ to have norm at most one. Peter recalled, in a private communication, that he discussed this problem with Paul Halmos when he was his student at the University of Michigan. The most interesting case, and the hardest, occurs when the matrices are all required to be Hermitian. Halmos mentions the problem specifically in this form in [**28**].

This ‘finite-dimensional’ problem is closely linked to the Brown–Douglas–Fillmore Theorem, as we shall see. I put finite dimensional in quotes because problems about matrices which ask for quantitative answers independent of dimension are really infinite-dimensional problems, and can generally be stated in terms of the compact operators rather than matrices of arbitrary size.

If $A$ and $B$ are Hermitian matrices in $\mathfrak{M}_n$, then $T = A + iB$ is a matrix satisfying

$$[T^*, T] = T^*T - TT^* = 2i(AB - BA).$$

So if $A$ and $B$ almost commute, then $T$ is almost normal; and they are close to commuting if and only if $T$ is close to a normal matrix.

Halmos [**26**] defines an operator $T$ to be *quasidiagonal* if there is a sequence $P_n$ of finite rank projections increasing to the identity so that $\|P_nT - TP_n\|$ goes to 0. The quasidiagonal operators form a closed set which is also closed under compact perturbations. Normal operators are quasidiagonal, and thus so are normal plus compact operators. Fredholm quasidiagonal operators have index 0.

Now suppose that $T$ is an essentially normal operator which is quasidiagonal. Then one can construct a sequence of projections $P_n$ increasing to the identity so that $\sum_{n\geq 1} \|P_nT - TP_n\|$ is small. A small compact perturbation of $T$ yields the operator $\sum_{n\geq 1} \oplus T_n$ where

$$T_n = (P_n - P_{n-1})T(P_n - P_{n-1})|_{(P_n - P_{n-1})\mathcal{H}}.$$

The essentially normal property means that $\lim_n \|[T_n^*, T_n]\| = 0$. So a positive solution to the nearly commuting problem would show that $T$ can be perturbed by a block diagonal compact operator to a direct sum of normal operators. This provides a direct link to the BDF theorem.

If one fixes the dimension $n$ and limits $A$ and $B$ to the (compact) unit ball, then a compactness argument establishes the existence of a function $\delta(\varepsilon, n)$ such that if $A$ and $B$ are in the unit ball of $\mathfrak{M}_n$ and $\|AB - BA\| < \delta(\varepsilon, n)$, then $A$ and $B$ are within $\varepsilon$ of a commuting pair. (See [**35**].) For this reason, the problem is much less interesting for fixed $n$. Pearcy and Shields [**36**] obtain the explicit estimate $\delta(\varepsilon, n) = 2\varepsilon^2/n$ when $A$ and $A'$ are Hermitian but $B$ is arbitrary.

After these initial results, a variety of counterexamples were found to various versions of the problem. Voiculescu [**43**] used very deep methods to establish the existence of triples $A_n, B_n, C_n$ of norm one Hermitian $n \times n$ matrices which asymptotically commute but are bounded away from commuting Hermitian triples. An explicit and somewhat stronger example due to the author [**13**] provides matrices $A_n = A_n^*$ and normal matrices $B_n$ in the unit ball of $\mathfrak{M}_{n^2+1}$ with $\|A_nB_n - B_nA_n\| = n^{-2}$, but bounded away from commuting pairs $A'_n$ and $B'_n$ with $A'_n$ Hermitian but $B'_n$ are arbitrary. Voiculescu [**44**] also constructs asymptotically commuting unitary matrices which are bounded away from commuting unitaries. Exel and Loring [**20**] provide a very slick example in which the pairs of unitaries are actually bounded away from arbitrary commuting pairs. Finally, we mention a paper of Choi [**11**] who also found pairs of arbitrary matrices which asymptotically commute but are bounded away from commuting pairs.

We sketch the Exel–Loring example [**20**]. Let $U_n$ be the cyclic shift on a basis $e_1, \dots, e_n$ and let $V_n$ be the diagonal unitary with eigenvalues $\omega^j$, $1 \le j \le n$, where $\omega = e^{2\pi i/n}$. Then $U_nV_nU_n^{-1}V_n^{-1} = \omega I_n$. In particular,

$$\|U_nV_n - V_nU_n\| = |1 - \omega| = 2\sin\pi/n.$$

Now if $A_n$ and $B_n$ are commuting matrices within $1/3$ of $U_n$ and $V_n$, define $A_s = (1-s)U_n + sA_n$ and $B_s = (1-s)V_n + sB_n$. One can check that for $0 \le s, t \le 1$,

$$\gamma(s,t) = \det\big((1-t)I_n + tA_sB_sA_s^{-1}B_s^{-1}\big)$$

is never 0; and $\gamma(s,0) = \gamma(s,1) = 1$. For fixed $s$ and $0 \le t \le 1$, $\gamma(s,\cdot)$ determines a closed loop in $\mathbb{C} \setminus \{0\}$. When $s = 0$, it reduces to the loop $(1 - t + t\omega)^n$, which has winding number 1. But at $s = 1$, it is the constant loop 1, which has winding number 0. This establishes a contradiction.

With all this negative evidence, one might suspect that the Hermitian pair question would also have a negative answer. However, the examples all use some kind of topological obstruction, which Loring [**34**] and Loring–Exel [**21**] make precise. The case of a pair of Hermitian matrices is different.

This author [**13**] provided a partial answer to the Hermitian case by proving an absorption result. If $T \in \mathfrak{M}_n$ is an arbitrary matrix, there is a normal matrix $N$ in $\mathfrak{M}_n$ with $\|N\| \le \|T\|$ and a normal matrix $N'$ in $\mathfrak{M}_{2n}$ so that

$$\|T \oplus N - N'\| \le 75\|T^*T - TT^*\|^{1/2}.$$

Thus if $T$ has a small commutator, one obtains a normal matrix close to $T \oplus N$. While this does not answer the question exactly, it can take advantage of an approximate normal summand of $T$—and in the case of an essentially normal operator $T$, such a summand is available with spectrum equal to $\sigma_e(T)$. The real problem is that the normal matrix $N$ may have too much spectrum, in some sense.

This approach was pursued in a paper by Berg and the author [**7**] in order to provide an operator theoretic proof of the BDF Theorem. The key was to establish a variant of the absorption theorem for the annulus. Specifically, if $T$ is an invertible operator with $\|T\| \le R$ and $\|T^{-1}\| \le 1$, then there are normal operators $N$ and $N'$

satisfying the same bounds (so the spectrum lies in the annulus $\{z : 1 \le |z| \le R\}$) such that

$$\|T \oplus N - N'\| \le 95\|T^*T - TT^*\|^{1/2}.$$

We established this by showing that the polar decomposition of $T$ almost commutes, and using the normal summand to provide room for the perturbation. Combining this with an elementary extraction of approximate eigenvectors allows one to show that, with $T$ as above, there is a normal operator $N$ with spectrum in the annulus so that $\|T - N\| \le 100\|T^*T - TT^*\|^{1/2}$.

The second important step of our proof of BDF is to establish that if $T$ is essentially normal and $\operatorname{ind}(T - \lambda I) = 0$ for $\lambda \not\in \sigma_e(T)$, then $T$ is quasidiagonal. Since the set of quasidiagonal operators is closed, it suffices to work with a small perturbation. Now $T \sim_a T \oplus N$ where $N$ is a diagonal normal operator with $\sigma(N) = \sigma_e(N) = \sigma_e(T)$. We fatten up the spectrum of $N$ with a small perturbation so that it is a nice domain with finitely many smooth holes. An approximation technique replaces $T \oplus N$ by a finite direct sum of operators with topologically annular spectra. Essentially one cuts the spectrum into annular regions without cutting through any holes. Then the Riesz functional calculus and the case for the annulus do the job. This results in a proof of the BDF theorem for essentially normal operators (i.e., planar $X$).

The almost commuting matrix question was finally solved by Huaxin Lin [**33**]. This paper is a tour de force. It starts with an idea of Voiculescu's. Suppose that there is a counterexample, namely asymptotically commuting $n \times n$ Hermitian matrices $A_n$ and $B_n$ of norm 1 which are bounded away from commuting pairs. Let $T_n = A_n + iB_n$. Then $T = T_1 \oplus T_2 \oplus \cdots$ is a block diagonal, essentially normal operator which is not a *block diagonal* compact perturbation of a normal operator. One should consider $T$ as an element of the von Neumann algebra $\mathfrak{M} := \prod \mathfrak{M}_n$ with commutator $T^*T - TT^*$ lying in the ideal $\mathfrak{J} = \sum \mathfrak{M}_n$ of sequences which converge to 0. The image $t = T + \mathfrak{J}$ is a normal element of the quotient algebra. Lin succeeds in proving that $t$ can be approximated by a normal element having finite spectrum. Then it is an easy matter to lift the spectral projections to projections in $\mathfrak{M}$, and so approximate $T$ by a normal element in $\mathfrak{M}$, which yields a good normal approximation to all but finitely many $T_n$. Unfortunately this is an extremely difficult proof.

Lin's Theorem was made much more accessible by Friis and Rordam [**22**], who provide a short, slick and elementary proof. They begin with the same setup. Observe that by the spectral theorem, every self-adjoint element of any von Neumann algebra can be approximated arbitrarily well be self-adjoint elements with finite spectrum. This property, called *real rank zero* (RR0), passes to quotients like $\mathfrak{M}/\mathfrak{J}$. In $\mathfrak{M}_n$, the invertible matrices are dense. If you prove this by modifying the positive part of the polar decomposition, the estimates are independent of dimension. Thus the argument can be readily extended to show that the invertible elements are dense in $\mathfrak{M}$. This property, called *topological stable rank one* (tsr1), also passes to quotients. Moreover, we observe that a normal element can perturbed to an invertible *normal* operator.

Now let $t$ be a normal element of $\mathfrak{M}/\mathfrak{J}$, and fix $\varepsilon > 0$. Cover the spectrum $\sigma(t)$ with a grid of lines spaced $\varepsilon$ apart horizontally and vertically. Use the tsr1 property to make a small perturbation which is normal and has a hole in the spectrum inside each square of the grid. Then use the continuous functional calculus to obtain another perturbation to a normal element nearby that has spectrum contained in the grid.

The RR0 property says that self-adjoint elements can be approximated by self-adjoint elements with finite spectrum. In particular, if $a = a^*$ has $\sigma(a) = [0, 1]$, this property allows us to find $b = b^*$ with $\|a-b\| < \varepsilon$ such that $b-1/2$ is invertible. A modification of this idea works on each line segment of the grid. So another small perturbation yields a normal operator with spectrum contained in the grid minus the mid-point of each line segment. A further use of the functional calculus collapses the remaining components of the spectrum to the lattice points of the grid. This produces the desired normal approximation with finite spectrum.

In their sequel **[23]**, Friis and Rordam use similar methods in the Calkin algebra provided that the index data is trivial. They establish quasidiagonality for essentially normal operators with zero index data, and thus provide a third proof of the BDF theorem.

Finally we mention that a very recent paper of Hastings **[29]** provides a constructive proof that almost commuting Hermitian matrices are close to commuting, with explicit estimates. This is a welcome addition since the soft proof provides no norm estimates at all. It is still an open question whether a perturbation of size $O(\|T^*T - TT^*\|^{1/2})$ is possible as in the case of the absorption results.

## References

[1] J. Anderson, *A C*-algebra A for which* Ext($A$) *is not a group*, Ann. Math. 107 (1978), 455–458.

[2] W. Arveson, *A note on essentially normal operators*, Proc. Royal Irish Acad. **74** (1974), 143–146.

[3] W. Arveson, *Notes in extensions of C*-algebras*, Duke Math. J. **44** (1977), 329–355.

[4] M. Atiyah, *K-theory*, W.A. Benjamin Inc., New York, 1967.

[5] H. Bercovici and D. Voiculescu, *The analogue of Kuroda's theorem for n-tuples*, The Gohberg anniversary collection, Vol. II (Calgary, AB, 1988), 57–60, Oper. Theory Adv. Appl. 41, Birkhäuser, Basel, 1989.

[6] I.D. Berg, *An extension of the Weyl-von Neumann theorem to normal operators*, Trans. Amer. Math. Soc. 160 (1971), 365–371.

[7] I.D. Berg and K. Davidson, *Almost commuting matrices and a quantitative version of the Brown-Douglas-Fillmore theorem*, Acta Math. 166 (1991), 121–161.

[8] Bratteli, O., *Inductive limits of finite-dimensional C*-algebras*, Trans. Amer. Math. Soc. **171** (1972), 195–234.

[9] L. Brown, R. Douglas and P. Fillmore, *Unitary equivalence modulo the compact operators and extensions of C*-algebras*, Proc. conference on Operator theory, Halifax, NS, Lect. Notes Math. 3445, Springer Verlag, Berlin, 1973.

[10] L. Brown, R. Douglas and P. Fillmore, *Extensions of C*-algebras and K-homology*, Ann. Math. 105 (1977), 265–324.

[11] M.D. Choi, *Almost commuting matrices need not be nearly commuting*, Proc. Amer. Math. Soc. 102 (1988), 529–533.

[12] M.D. Choi and E. Effros, *The completely positive lifting problem for C*-algebras*, Ann. Math. **104** (1976), 585–609.

[13] K. Davidson, *Almost commuting Hermitian matrices*, Math. Scand. 56 (1985), 222–240.

[14] K. Davidson, *Normal operators are diagonal plus Hilbert-Schmidt*, J. Operator Theory **20** (1988) 241–250.

[15] K. Davidson, *C*-Algebras by Example*, Fields Institute Monograph Series **6**, American Mathematical Society, Providence, RI, 1996.

[16] R. Douglas, *C*-algebra extensions and K-homology*, Annals Math. Studies 95, Princeton University Press, Princeton, N.J., 1980.

[17] E. Effros, D. Handelman and C. Shen, *Dimension groups and their affine transformations*, Amer. J. Math. **102** (1980), 385–402.

[18] G. Elliott, *On the classification of inductive limits of sequences of semi-simple finite-dimensional algebras*, J. Algebra **38** (1976), 29–44.

[19] G. Elliott, *On totally ordered groups and* $K_0$ , Proc. Ring Theory conf., Waterloo, D. Handelman and J. Lawrence (eds.), Lect. Notes Math. **734** (1978), 1–49, Springer–Verlag, New York, 1978.

[20] R. Exel and T. Loring, *Almost commuting unitary matrices*, Proc. Amer. Math. Soc. 106 (1989), 913–915.

[21] R. Exel and T. Loring, *Invariants of almost commuting unitaries*, J. Funct. Anal. 95 (1991), 364–376.

[22] P. Friis and M. Rordam, *Almost commuting self-adjoint matrices—a short proof of Huaxin's theorem*, J. reine angew. Math. 479 (1996), 121–131.

[23] P. Friis and M. Rordam, *Approximation with normal operators with finite spectrum, and an elementary proof of the Brown–Douglas–Fillmore theorem*, Pacific J. Math. 199 (2001), 347–366.

[24] U. Haagerup and S. Thorbjörnsen, *A new application of random matrices:* $\mathrm{Ext}(\mathrm{C}^*_{\mathrm{red}}(\mathbb{F}_2))$ *is not a group*, Ann. Math. 162 (2005), 711–775.

[25] D. Hadwin, *An operator-valued spectrum*, Indiana Univ. Math. J. **26** (1977), 329–340.

[26] P.R. Halmos, *Ten problems in Hilbert space*, Bull. Amer. Math. Soc. 76 (1970), 887–933.

[27] P.R. Halmos, *What does the spectral theorem say?*, Amer. Math. Monthly 70 (1963), 241–247.

[28] P.R. Halmos, *Some unsolved problems of unknown depth about operators on Hilbert space*, Proc. Roy. Soc. Edinburgh Sect. A 76 (1976/77), 67–76.

[29] M. Hastings, *Making almost commuting matrices commute*, Comm. Math. Phys. 291 (2009), no. 2, 321–345.

[30] G. Kasparov, *The operator K-functor and extensions of C*-algebras* (*Russian*), Izv. Akad. Nauk SSSR Ser. Mat. 44, (1980), 571–636.

[31] T. Kato, *Perturbation of continuous spectra by trace class operators*, Proc. Japan Acad. 33 (1957), 260–264.

[32] S. Kuroda, *On a theorem of Weyl–von Neumann*, Proc. Japan Acad. 34 (1958), 11–15.

[33] H. Lin, *Almost commuting Hermitian matrices and applications*, Fields Institute Comm. 13 (1997), 193–233.

[34] T. Loring, *K-theory and asymptotically commuting matrices*, Canad. J. Math. 40 (1988), 197–216.

[35] W. Luxemburg and R. Taylor, *Almost commuting matrices are near commuting matrices*, Indag. Math. 32 (1970), 96–98.

[36] C. Pearcy and A. Shields, *Almost commuting matrices*, J. Funct. Anal. 33 (1979), 332–338.

[37] M. Rosenblum, *Perturbation of the continuous spectrum and unitary equivalence*, Pacific J. Math. 7 (1957), 997–1010.

[38] P. Rosenthal, *Research Problems: Are Almost Commuting Matrices Near Commuting Matrices?*, Amer. Math. Monthly 76 (1969), 925–926.

[39] W. Sikonia, *The von Neumann converse of Weyl's theorem*, Indiana Univ. Math. J. 21 (1971/1972), 121–124.

[40] D. Voiculescu, *A non-commutative Weyl–von Neumann Theorem*, Rev. Roum. Pures Appl. **21** (1976), 97–113.

[41] D. Voiculescu, *Some results on norm-ideal perturbations of Hilbert space operators*, J. Operator Theory 2 (1979), 3–37.

[42] D. Voiculescu, *Some results on norm-ideal perturbations of Hilbert space operators, II*, J. Operator Theory 5 (1981), 77–100.

[43] D. Vioculescu, *Remarks on the singular extension in the C*-algebra of the Heisenberg group*, J. Operator Theory 5 (1981), 147–170.

[44] D. Vioculescu, *Asymptotically commuting finite rank unitaries without commuting approximants*, Acta Sci. Math. (Szeged) 45 (1983), 429–431.

[45] J. von Neumann, *Charakterisierung des Spektrums eines Integraloperators*, Actualités Sci. Indust. 229, Hermann, Paris, 1935.

[46] H. Weyl, *Über beschränkte quadratische Formen deren Differenz vollstetig ist*, Rend. Circ. Mat. Palermo 27 (1909), 373–392.

Kenneth R. Davidson
Pure Math. Dept.
University of Waterloo
Waterloo
ON N2L–3G1, Canada
e-mail: `krdavids@uwaterloo.ca`

Operator Theory:
Advances and Applications, Vol. 207, 223–254

# The Operator Fejér-Riesz Theorem

Michael A. Dritschel and James Rovnyak

*To the memory of Paul Richard Halmos*

**Abstract.** The Fejér-Riesz theorem has inspired numerous generalizations in one and several variables, and for matrix- and operator-valued functions. This paper is a survey of some old and recent topics that center around Rosenblum's operator generalization of the classical Fejér-Riesz theorem.

**Mathematics Subject Classification (2000).** Primary 47A68; Secondary 60G25, 47A56, 47B35, 42A05, 32A70, 30E99.

**Keywords.** Trigonometric polynomial, Fejér-Riesz theorem, spectral factorization, Schur complement, noncommutative polynomial, Toeplitz operator, shift operator.

## 1. Introduction

The classical Fejér-Riesz factorization theorem gives the form of a nonnegative trigonometric polynomial on the real line, or, equivalently, a Laurent polynomial that is nonnegative on the unit circle. For the statement, we write $\mathbb{D} = \{z\colon |z| < 1\}$ and $\mathbb{T} = \{\zeta\colon |\zeta| = 1\}$ for the open unit disk and unit circle in the complex plane.

**Fejér-Riesz Theorem.** *A Laurent polynomial $q(z) = \sum_{k=-m}^{m} q_k z^k$ which has complex coefficients and satisfies $q(\zeta) \geq 0$ for all $\zeta \in \mathbb{T}$ can be written*

$$q(\zeta) = |p(\zeta)|^2, \qquad \zeta \in \mathbb{T},$$

*for some polynomial $p(z) = p_0 + p_1 z + \cdots + p_m z^m$, and $p(z)$ can be chosen to have no zeros in $\mathbb{D}$.*

The original sources are Fejér [22] and Riesz [47]. The proof is elementary and consists in showing that the roots of $q(z)$ occur in pairs $z_j$ and $1/\bar{z}_j$ with $|z_j| \geq 1$. Then the required polynomial $p(z)$ is the product of the factors $z - z_j$ adjusted by a suitable multiplicative constant $c$. Details appear in many places; see, e.g., [28, p. 20], [34, p. 235], or [60, p. 26].

The Fejér-Riesz theorem arises naturally in spectral theory, the theory of orthogonal polynomials, prediction theory, moment problems, and systems and control theory. Applications often require generalizations to functions more general than Laurent polynomials, and, more than that, to functions whose values are matrices or operators on a Hilbert space. The spectral factorization problem is to write a given nonnegative matrix- or operator-valued function $F$ on the unit circle in the form $F = G^*G$ where $G$ has an analytic extension to the unit disk (in a suitably interpreted sense). The focal point of our survey is the special case of a Laurent polynomial with operator coefficients.

The operator Fejér-Riesz theorem (Theorem 2.1) obtains a conclusion similar to the classical result for a Laurent polynomial whose coefficients are Hilbert space operators: if $Q_j$, $j = -m, \dots, m$, are Hilbert space operators such that

$$Q(\zeta) = \sum_{k=-m}^{m} Q_k \zeta^k \geq 0, \qquad \zeta \in \mathbb{T}, \tag{1.1}$$

then there is a polynomial $P(z) = P_0 + P_1 z + \cdots + P_m z^m$ with operator coefficients such that

$$Q(\zeta) = P(\zeta)^* P(\zeta), \qquad \zeta \in \mathbb{T}. \tag{1.2}$$

This was first proved in full generality in 1968 by Marvin Rosenblum [49]. The proof uses Toeplitz operators and a method of Lowdenslager, and it is a fine example of operator theory in the spirit of Paul Halmos. Rosenblum's proof is reproduced in §2.

Part of the fascination of the operator Fejér-Riesz theorem is that it can be stated in a purely algebraic way. The hypothesis (1.1) on $Q(z)$ is equivalent to the statement that an associated Toeplitz matrix is nonnegative. The conclusion (1.2) is equivalent to $2m + 1$ nonlinear equations whose unknowns are the coefficients $P_0, P_1, \dots, P_m$ of $P(z)$. Can it be that this system of equations can be solved by an algebraic procedure? The answer is, yes, and this is a recent development. The iterative procedure uses the notion of a Schur complement and is outlined in §3.

There is a surprising connection between Rosenblum's proof of the operator Fejér-Riesz theorem and spectral factorization. The problem of spectral factorization is formulated precisely in §4, using Hardy class notions. A scalar prototype is Szegő's theorem (Theorem 4.1) on the representation of a positive integrable and log-integrable function $w$ on the unit circle in the form $|h|^2$ for some $H^2$ function $h$. The operator and matrix counterparts of Szegő's theorem, Theorems 4.5 and 4.7, have been known for many years and go back to fundamental work in the 1940s and 1950s which was motivated by applications in prediction theory (see the historical notes at the end of §4). We present a proof that is new to the authors and we suspect not widely known. It is based on Theorem 4.3, which traces its origins to Rosenblum's implementation of the Lowdenslager method. In §4 we also state without proof some special results that hold in the matrix case.

The method of Schur complements points the way to an approach to multivariable factorization problems, which is the subject of §5. Even in the scalar case, the obvious first ideas for multivariable generalizations of the Fejér-Riesz theorem are false by well-known examples. Part of the problem has to do with what one might think are natural restrictions on degrees. In fact, the restrictions on degrees are not so natural after all. When they are removed, we can prove a result, Theorem 5.1, that can be viewed as a generalization of the operator Fejér-Riesz theorem in the strictly positive case. We also look at the problem of outer factorization, at least in some restricted settings.

In recent years there has been increasing interest in noncommutative function theory, especially in the context of functions of freely noncommuting variables. In §6 we consider noncommutative analogues of the $d$-torus, and corresponding notions of nonnegative trigonometric polynomials. In the freely noncommutative setting, there is a very nice version of the Fejér-Riesz theorem (Theorem 6.1). In a somewhat more general noncommutative setting, which also happens to cover the commutative case as well, we have a version of Theorem 5.1 for strictly positive polynomials (Theorem 6.2).

Our survey does not aim for completeness in any area. In particular, our bibliography represents only a selection from the literature. The authors regret and apologize for omissions.

## 2. The operator Fejér-Riesz theorem

In this section we give the proof of the operator Fejér-Riesz theorem by Rosenblum [49]. The general theorem had precursors. A finite-dimensional version was given by Rosenblatt [48], an infinite-dimensional special case by Gohberg [26].

We follow standard conventions for Hilbert spaces and operators. If $A$ is an operator, $A^*$ is its adjoint. Norms of vectors and operators are written $\|\cdot\|$. Except where noted, no assumption is made on the dimension of a Hilbert space, and nonseparable Hilbert spaces are allowed.

**Theorem 2.1 (Operator Fejér-Riesz Theorem).** *Let $Q(z) = \sum_{k=-m}^{m} Q_k z^k$ be a Laurent polynomial with coefficients in $\mathfrak{L}(\mathfrak{G})$ for some Hilbert space $\mathfrak{G}$. If $Q(\zeta) \geq 0$ for all $\zeta \in \mathbb{T}$, then*

$$Q(\zeta) = P(\zeta)^* P(\zeta), \qquad \zeta \in \mathbb{T}, \tag{2.1}$$

*for some polynomial $P(z) = P_0 + P_1 z + \cdots + P_m z^m$ with coefficients in $\mathfrak{L}(\mathfrak{G})$. The polynomial $P(z)$ can be chosen to be outer.*

The definition of an outer polynomial will be given later; in the scalar case, a polynomial is outer if and only if it has no zeros in $\mathbb{D}$.

The proof uses (unilateral) shift and Toeplitz operators (see [11] and [29]). By a **shift operator** here we mean an isometry $S$ on a Hilbert space $\mathfrak{H}$ such that the unitary component of $S$ in its Wold decomposition is trivial. With natural identifications, we can write $\mathfrak{H} = \mathfrak{G} \oplus \mathfrak{G} \oplus \cdots$ for some Hilbert space $\mathfrak{G}$ and

$$S(h_0, h_1, \dots) = (0, h_0, h_1, \dots)$$

when the elements of $\mathfrak{H}$ are written in sequence form. Suppose that such a shift $S$ is chosen and fixed. If $T, A \in \mathfrak{L}(\mathfrak{H})$, we say that $T$ is **Toeplitz** if $S^*TS = T$, and that $A$ is **analytic** if $AS = SA$. An analytic operator $A$ is said to be **outer** if $\overline{\operatorname{ran}} A$ is a subspace of $\mathfrak{H}$ of the form $\mathfrak{F} \oplus \mathfrak{F} \oplus \cdots$ for some closed subspace $\mathfrak{F}$ of $\mathfrak{G}$.

As block operator matrices, Toeplitz and analytic operators have the forms

$$T = \begin{pmatrix} T_0 & T_{-1} & T_{-2} & \cdots \\ T_1 & T_0 & T_{-1} & \ddots \\ T_2 & T_1 & T_0 & \ddots \\ \vdots & \ddots & \ddots & \ddots \end{pmatrix}, \qquad A = \begin{pmatrix} A_0 & 0 & 0 & \cdots \\ A_1 & A_0 & 0 & \ddots \\ A_2 & A_1 & A_0 & \ddots \\ \vdots & \ddots & \ddots & \ddots \end{pmatrix}. \tag{2.2}$$

Here

$$T_j = \begin{cases} E_0^* S^{*j} T E_0 | \mathfrak{G}, & j \geq 0, \\ E_0^* T S^{|j|} E_0 | \mathfrak{G}, & j < 0, \end{cases} \tag{2.3}$$

where $E_0 g = (g, 0, 0, \dots)$ is the natural embedding of $\mathfrak{G}$ into $\mathfrak{H}$. For examples, consider Laurent and analytic polynomials $Q(z) = \sum_{k=-m}^{m} Q_k z^k$ and $P(z) = P_0 + P_1 z + \cdots + P_m z^m$ with coefficients in $\mathfrak{L}(\mathfrak{G})$. Set $Q_j = 0$ for $|j| > m$ and $P_j = 0$ for $j > m$. Then the formulas

$$T_Q = \begin{pmatrix} Q_0 & Q_{-1} & Q_{-2} & \cdots \\ Q_1 & Q_0 & Q_{-1} & \ddots \\ Q_2 & Q_1 & Q_0 & \ddots \\ \vdots & \ddots & \ddots & \ddots \end{pmatrix}, \qquad T_P = \begin{pmatrix} P_0 & 0 & 0 & \cdots \\ P_1 & P_0 & 0 & \ddots \\ P_2 & P_1 & P_0 & \ddots \\ \vdots & \ddots & \ddots & \ddots \end{pmatrix} \tag{2.4}$$

define bounded operators on $\mathfrak{H}$. Boundedness follows from the identity

$$\int_{\mathbb{T}} \langle Q(\zeta) f(\zeta), g(\zeta) \rangle_{\mathfrak{G}} \, d\sigma(\zeta) = \sum_{k,j=0}^{\infty} \langle Q_{j-k} f_k, g_j \rangle_{\mathfrak{G}}, \tag{2.5}$$

where $\sigma$ is normalized Lebesgue measure on $\mathbb{T}$ and $f(\zeta) = f_0 + f_1 \zeta + f_2 \zeta^2 + \cdots$ and $g(\zeta) = g_0 + g_1 \zeta + g_2 \zeta^2 + \cdots$ have coefficients in $\mathfrak{G}$, all but finitely many of which are zero. The operator $T_Q$ is Toeplitz, and $T_P$ is analytic. Moreover,

- $Q(\zeta) \geq 0$ for all $\zeta \in \mathbb{T}$ if and only if $T_Q \geq 0$;
- $Q(\zeta) = P(\zeta)^* P(\zeta)$ for all $\zeta \in \mathbb{T}$ if and only if $T_Q = T_P^* T_P$.

**Definition 2.2.** We say that the polynomial $P(z)$ is **outer** if the analytic Toeplitz operator $A = T_P$ is outer.

In view of the example (2.4), the main problem is to write a given nonnegative Toeplitz operator $T$ in the form $T = A^*A$, where $A$ is analytic. We also want to know that if $T = T_Q$ for a Laurent polynomial $Q$, then we can choose $A = T_P$ for an outer analytic polynomial $P$ of the same degree. Lemmas 2.3 and 2.4 reduce the problem to showing that a certain isometry is a shift operator.

**Lemma 2.3 (Lowdenslager's Criterion).** *Let $\mathfrak{H}$ be a Hilbert space, and let $S \in \mathfrak{L}(\mathfrak{H})$ be a shift operator. Let $T \in \mathfrak{L}(\mathfrak{H})$ be Toeplitz relative to $S$ as defined above, and suppose that $T \geq 0$. Let $\mathfrak{H}_T$ be the closure of the range of $T^{1/2}$ in the inner product of $\mathfrak{H}$. Then there is an isometry $S_T$ mapping $\mathfrak{H}_T$ into itself such that*

$$S_T T^{1/2} f = T^{1/2} S f, \qquad f \in \mathfrak{H}.$$

*In order that $T = A^*A$ for some analytic operator $A \in \mathfrak{L}(\mathfrak{H})$, it is necessary and sufficient that $S_T$ is a shift operator. In this case, $A$ can be chosen to be outer.*

*Proof.* The existence of the isometry $S_T$ follows from the identity $S^*TS = T$, which implies that $T^{1/2}Sf$ and $T^{1/2}f$ have the same norms for any $f \in \mathfrak{H}$.

If $S_T$ is a shift operator, we can view $\mathfrak{H}_T$ as a direct sum $\mathfrak{H}_T = \mathfrak{G}_T \oplus \mathfrak{G}_T \oplus \cdots$ with $S_T(h_0, h_1, \dots) = (0, h_0, h_1, \dots)$. Here $\dim \mathfrak{G}_T \leq \dim \mathfrak{G}$. To see this, notice that a short argument shows that $T^{1/2}S_T^*$ and $S^*T^{1/2}$ agree on $\mathfrak{H}_T$, and therefore $T^{1/2}(\ker S_T^*) \subseteq \ker S^*$. The dimension inequality then follows because $T^{1/2}$ is one-to-one on the closure of its range. Therefore we may choose an isometry $V$ from $\mathfrak{G}_T$ into $\mathfrak{G}$. Define an isometry $W$ on $\mathfrak{H}_T$ into $\mathfrak{H}$ by

$$W(h_0, h_1, \dots) = (Vh_0, Vh_1, \dots).$$

Define $A \in \mathfrak{L}(\mathfrak{H})$ by mapping $\mathfrak{H}$ into $\mathfrak{H}_T$ via $T^{1/2}$ and then $\mathfrak{H}_T$ into $\mathfrak{H}$ via $W$:

$$Af = WT^{1/2}f, \qquad f \in \mathfrak{H}.$$

Straightforward arguments show that $A$ is analytic, outer, and $T = A^*A$.

Conversely, suppose that $T = A^*A$ where $A \in \mathfrak{L}(\mathfrak{H})$ is analytic. Define an isometry $W$ on $\mathfrak{H}_T$ into $\mathfrak{H}$ by $WT^{1/2}f = Af$, $f \in \mathfrak{H}$. Then $WS_T = SW$, and hence ${S_T^*}^n = W^*{S^*}^nW$ for all $n \geq 1$. Since the powers of $S^*$ tend strongly to zero, so do the powers of $S_T^*$, and therefore $S_T$ is a shift operator. □

**Lemma 2.4.** *In Lemma 2.3, let $T = T_Q$ be given by (2.4) for a Laurent polynomial $Q(z)$ of degree $m$. If $T = A^*A$ where $A \in \mathfrak{L}(\mathfrak{H})$ is analytic and outer, then $A = T_P$ for some outer analytic polynomial $P(z)$ of degree $m$.*

*Proof.* Let $Q(z) = \sum_{k=-m}^{m} Q_k z^k$. Recall that $Q_j = 0$ for $|j| > m$. By (2.3) applied to $A$, what we must show is that ${S^*}^j AE_0 = 0$ for all $j > m$. It is sufficient to show that ${S^*}^{m+1}AE_0 = 0$. By (2.3) applied to $T$, since $T = A^*A$ and $A$ is analytic,

$$E_0^* A^* {S^*}^j AE_0 = E_0^* {S^*}^j TE_0 = Q_j = 0, \qquad j > m.$$

It follows that $\operatorname{ran} {S^*}^{m+1}AE_0 \perp \operatorname{ran} AS^kE_0$ for all $k \geq 0$, and therefore

$$\operatorname{ran} {S^*}^{m+1}AE_0 \perp \operatorname{ran} A. \tag{2.6}$$

Since $A$ is outer, $\overline{\operatorname{ran}}\, A$ reduces $S$, and so $\operatorname{ran} {S^*}^{m+1}AE_0 \subseteq \overline{\operatorname{ran}}\, A$. Therefore ${S^*}^{m+1}AE_0 = 0$ by (2.6), and the result follows. □

The proof of the operator Fejér-Riesz theorem is now easily completed.

*Proof of Theorem* 2.1. Define $T = T_Q$ as in (2.4). Lemmas 2.3 and 2.4 reduce the problem to showing that the isometry $S_T$ is a shift operator. It is sufficient to show that $\|{S_T^*}^n f\| \to 0$ for every $f$ in $\mathfrak{H}_T$.

**Claim:** If $f = T^{1/2}h$ where $h \in \mathfrak{H}$ has the form $h = (h_0, \dots, h_r, 0, \dots)$, then ${S_T^*}^n f = 0$ for all sufficiently large $n$.

For if $u \in \mathfrak{H}$ and $n$ is any positive integer, then

$$\left\langle {S_T^*}^n f, T^{1/2}u \right\rangle_{\mathfrak{H}_T} = \left\langle f, S_T^n T^{1/2}u \right\rangle_{\mathfrak{H}_T} = \left\langle T^{1/2}h, T^{1/2}S^n u \right\rangle_{\mathfrak{H}} = \langle Th, S^n u\rangle_{\mathfrak{H}}.$$

By the definition of $T = T_Q$, $Th$ has only a finite number of nonzero entries (depending on $m$ and $r$), and the first $n$ entries of $S^n u$ are zero (irrespective of $u$). The claim follows from the arbitrariness of $u$.

In view of the claim, $\|{S_T^*}^n f\| \to 0$ for a dense set of vectors in $\mathfrak{H}_T$, and hence by approximation this holds for all $f$ in $\mathfrak{H}_T$. Thus $S_T$ is a shift operator, and the result follows. □

A more general result is proved in the original version of Theorem 2.1 in [49]. There it is only required that $Q(z)g$ is a Laurent polynomial for a dense set of $g$ in $\mathfrak{G}$ (the degrees of these polynomials can be unbounded). We have omitted an accompanying uniqueness statement: the outer polynomial $P(z)$ in Theorem 2.1 can be chosen such that $P(0) \geq 0$, and then it is unique. See [2] and [50].

## 3. Method of Schur complements

We outline now a completely different proof of the operator Fejér-Riesz theorem. The proof is due to Dritschel and Woerdeman [19] and is based on the notion of a Schur complement. The procedure constructs the outer polynomial $P(z) = P_0 + P_1 z + \cdots + P_m z^m$ one coefficient at a time. A somewhat different use of Schur complements in the operator Fejér-Riesz theorem appears in Dritschel [18]. The method in [18] plays a role in the multivariable theory, which is taken up in §5.

We shall explain the main steps of the construction assuming the validity of two lemmas. Full details are given in [19] and also in the forthcoming book [3] by Bakonyi and Woerdeman. The authors thank Mihaly Bakonyi and Hugo Woerdeman for advance copies of key parts of [3], which has been helpful for our exposition. The book [3] includes many additional results not discussed here.

**Definition 3.1.** Let $\mathfrak{H}$ be a Hilbert space. Suppose $T \in \mathfrak{L}(\mathfrak{H})$, $T \geq 0$. Let $\mathfrak{K}$ be a closed subspace of $\mathfrak{H}$, and let $P_{\mathfrak{K}} \in \mathfrak{L}(\mathfrak{H}, \mathfrak{K})$ be orthogonal projection of $\mathfrak{H}$ onto $\mathfrak{K}$. Then (see Appendix, Lemma A.2) there is a unique operator $S \in \mathfrak{L}(\mathfrak{K})$, $S \geq 0$, such that

(i) $T - P_{\mathfrak{K}}^* S P_{\mathfrak{K}} \geq 0$;
(ii) if $\widetilde{S} \in \mathfrak{L}(\mathfrak{K})$, $\widetilde{S} \geq 0$, and $T - P_{\mathfrak{K}}^* \widetilde{S} P_{\mathfrak{K}} \geq 0$, then $\widetilde{S} \leq S$.

We write $S = S(T, \mathfrak{K})$ and call $S$ the **Schur complement of** $T$ **supported on** $\mathfrak{K}$.

Schur complements satisfy an inheritance property, namely, if $\mathfrak{K}_- \subseteq \mathfrak{K}_+ \subseteq \mathfrak{H}$, then $S(T,\mathfrak{K}_-) = S(S(T,\mathfrak{K}_+),\mathfrak{K}_-)$. If $T$ is specified in matrix form,

$$T = \begin{pmatrix} A & B^* \\ B & C \end{pmatrix} : \mathfrak{K} \oplus \mathfrak{K}^\perp \to \mathfrak{K} \oplus \mathfrak{K}^\perp,$$

then $S = S(T,\mathfrak{K})$ is the largest nonnegative operator in $\mathfrak{L}(\mathfrak{K})$ such that

$$\begin{pmatrix} A - S & B^* \\ B & C \end{pmatrix} \geq 0.$$

The condition $T \geq 0$ is equivalent to the existence of a contraction $G \in \mathfrak{L}(\mathfrak{K},\mathfrak{K}^\perp)$ such that $B = C^{\frac12} G A^{\frac12}$ (Appendix, Lemma A.1). In this case, $G$ can be chosen so that it maps $\overline{\operatorname{ran}}\, A$ into $\overline{\operatorname{ran}}\, C$ and is zero on the orthogonal complement of $\overline{\operatorname{ran}}\, A$, and then

$$S = A^{\frac12}(I - G^*G)A^{\frac12}.$$

When $C$ is invertible, this reduces to the familiar formula $S = A - B^*C^{-1}B$.

**Lemma 3.2.** *Let $M \in \mathfrak{L}(\mathfrak{H})$, $M \geq 0$, and suppose that*

$$M = \begin{pmatrix} A & B^* \\ B & C \end{pmatrix} : \mathfrak{K} \oplus \mathfrak{K}^\perp \to \mathfrak{K} \oplus \mathfrak{K}^\perp$$

*for some closed subspace $\mathfrak{K}$ of $\mathfrak{H}$.*
(1) *If $S(M,\mathfrak{K}) = P^*P$ and $C = R^*R$ for some $P \in \mathfrak{L}(\mathfrak{K})$ and $R \in \mathfrak{L}(\mathfrak{K}^\perp)$, then there is a unique $X \in \mathfrak{L}(\mathfrak{K},\mathfrak{K}^\perp)$ such that*

$$M = \begin{pmatrix} P^* & X^* \\ 0 & R^* \end{pmatrix}\begin{pmatrix} P & 0 \\ X & R \end{pmatrix} \qquad \text{and} \qquad \operatorname{ran} X \subseteq \overline{\operatorname{ran}}\, R. \tag{3.1}$$

(2) *Conversely, if* (3.1) *holds for some operators $P, X, R$, then $S(M,\mathfrak{K}) = P^*P$.*

We omit the proof and refer the reader to [3] or [19] for details.

*Proof of Theorem* 2.1 *using Schur complements.* Let $Q(z) = \sum_{k=-m}^{m} Q_k z^k$ satisfy $Q(\zeta) \geq 0$ for all $\zeta \in \mathbb{T}$. We shall recursively construct the coefficients of an outer polynomial $P(z) = P_0 + P_1 z + \cdots + P_m z^m$ such that $Q(\zeta) = P(\zeta)^* P(\zeta)$, $\zeta \in \mathbb{T}$.

Write $\mathfrak{H} = \mathfrak{G} \oplus \mathfrak{G} \oplus \cdots$ and $\mathfrak{G}^n = \mathfrak{G} \oplus \cdots \oplus \mathfrak{G}$ with $n$ summands. As before, set $Q_k = 0$ for $|k| > m$, and define $T_Q \in \mathfrak{L}(\mathfrak{H})$ by

$$T_Q = \begin{pmatrix} Q_0 & Q_{-1} & Q_{-2} & \cdots \\ Q_1 & Q_0 & Q_{-1} & \ddots \\ Q_2 & Q_1 & Q_0 & \ddots \\ \vdots & \ddots & \ddots & \ddots \end{pmatrix}.$$

For each $k = 0, 1, 2, \ldots$, define

$$S(k) = S(T_Q, \mathfrak{G}^{k+1}),$$

which we interpret as the Schur complement of $T_Q$ on the first $k+1$ summands of $\mathfrak{H} = \mathfrak{G} \oplus \mathfrak{G} \oplus \cdots$. Thus $S(k)$ is a $(k+1) \times (k+1)$ block operator matrix satisfying

$$S(S(k), \mathfrak{G}^{j+1}) = S(j), \quad 0 \leq j < k < \infty, \tag{3.2}$$

by the inheritance property of Schur complements.

**Lemma 3.3.** *For each* $k = 0, 1, 2, \ldots$,

$$S(k+1) = \begin{pmatrix} Y_0 & \begin{pmatrix} Y_1 & \cdots & Y_{k+1} \end{pmatrix} \\ \begin{pmatrix} Y_1^* \\ \vdots \\ Y_{k+1}^* \end{pmatrix} & S(k) \end{pmatrix}$$

*for some operators* $Y_0, Y_1, \ldots, Y_{k+1}$ *in* $\mathfrak{L}(\mathfrak{G})$*. For* $k \geq m-1$,

$$\begin{pmatrix} Y_0 & Y_1 & \cdots & Y_{k+1} \end{pmatrix} = \begin{pmatrix} Q_0 & Q_{-1} & \cdots & Q_{-k-1} \end{pmatrix}.$$

Again see [3] or [19] for details. Granting Lemmas 3.2 and 3.3, we can proceed with the construction.

**Construction of $P_0, P_1$.** Choose $P_0 = S(0)^{\frac{1}{2}}$. Using Lemma 3.3, write

$$S(1) = \begin{pmatrix} Y_0 & Y_1 \\ Y_1^* & S(0) \end{pmatrix}.$$

In Lemma 3.2(1) take $M = S(1)$ and use the factorizations

$$S(S(1), \mathfrak{G}^1) \overset{(3.2)}{=} S(0) = P_0^* P_0 \qquad \text{and} \qquad S(0) = P_0^* P_0 .$$

Choose $P_1 = X$ where $X \in \mathfrak{L}(\mathfrak{G})$ is the operator produced by Lemma 3.2(1). Then

$$S(1) = \begin{pmatrix} P_0^* & P_1^* \\ 0 & P_0^* \end{pmatrix} \begin{pmatrix} P_0 & 0 \\ P_1 & P_0 \end{pmatrix} \qquad \text{and} \qquad \operatorname{ran} P_1 \subseteq \overline{\operatorname{ran}} \, P_0. \tag{3.3}$$

**Construction of $P_2$.** Next use Lemma 3.3 to write

$$S(2) = \begin{pmatrix} Y_0 & \begin{pmatrix} Y_1 & Y_2 \end{pmatrix} \\ \begin{pmatrix} Y_1^* \\ Y_2^* \end{pmatrix} & S(1) \end{pmatrix},$$

and apply Lemma 3.2(1) to $M = S(2)$ with the factorizations

$$S(S(2), \mathfrak{G}^1) \overset{(3.2)}{=} S(0) = P_0^* P_0,$$
$$S(1) = \begin{pmatrix} P_0^* & P_1^* \\ 0 & P_0^* \end{pmatrix} \begin{pmatrix} P_0 & 0 \\ P_1 & P_0 \end{pmatrix}.$$

This yields operators $X_1, X_2 \in \mathfrak{L}(\mathfrak{G})$ such that

$$S(2) = \begin{pmatrix} P_0^* & X_1^* & X_2^* \\ 0 & P_0^* & P_1^* \\ 0 & 0 & P_0^* \end{pmatrix} \begin{pmatrix} P_0 & 0 & 0 \\ X_1 & P_0 & 0 \\ X_2 & P_1 & P_0 \end{pmatrix}, \tag{3.4}$$

$$\operatorname{ran} \begin{pmatrix} X_1 \\ X_2 \end{pmatrix} \subseteq \overline{\operatorname{ran}} \begin{pmatrix} P_0 & 0 \\ P_1 & P_0 \end{pmatrix}. \tag{3.5}$$

In fact, $X_1 = P_1$. To see this, notice that we can rewrite (3.4) as

$$S(2) = \begin{pmatrix} \widetilde{P}^* & \widetilde{X}^* \\ 0 & \widetilde{R}^* \end{pmatrix} \begin{pmatrix} \widetilde{P} & 0 \\ \widetilde{X} & \widetilde{R} \end{pmatrix},$$

$$\widetilde{P} = \begin{pmatrix} P_0 & 0 \\ X_1 & P_0 \end{pmatrix}, \qquad \widetilde{X} = \begin{pmatrix} X_2 & P_1 \end{pmatrix}, \qquad \widetilde{R} = P_0.$$

By (3.3) and (3.5), $\operatorname{ran} P_1 \subseteq \overline{\operatorname{ran}}\, P_0$ and $\operatorname{ran} X_2 \subseteq \overline{\operatorname{ran}}\, P_0$, and therefore $\operatorname{ran} \widetilde{X} \subseteq \overline{\operatorname{ran}}\, P_0$. Hence by Lemma 3.2(2),

$$S(S(2), \mathfrak{G}^2) = \widetilde{P}^* \widetilde{P} = \begin{pmatrix} P_0^* & X_1^* \\ 0 & P_0^* \end{pmatrix} \begin{pmatrix} P_0 & 0 \\ X_1 & P_0. \end{pmatrix}. \tag{3.6}$$

Comparing this with

$$S(S(2), \mathfrak{G}^2) \overset{(3.2)}{=} S(1) \overset{(3.3)}{=} \begin{pmatrix} P_0^* & P_1^* \\ 0 & P_0^* \end{pmatrix} \begin{pmatrix} P_0 & 0 \\ P_1 & P_0 \end{pmatrix}, \tag{3.7}$$

we get $P_0^* P_1 = P_0^* X_1$. By (3.5), $\operatorname{ran} X_1 \subseteq \overline{\operatorname{ran}}\, P_0$, and therefore $X_1 = P_1$. Now choose $P_2 = X_2$ to obtain

$$S(2) = \begin{pmatrix} P_0^* & P_1^* & P_2^* \\ 0 & P_0^* & P_1^* \\ 0 & 0 & P_0^* \end{pmatrix} \begin{pmatrix} P_0 & 0 & 0 \\ P_1 & P_0 & 0 \\ P_2 & P_1 & P_0 \end{pmatrix}, \tag{3.8}$$

$$\operatorname{ran} \begin{pmatrix} P_1 \\ P_2 \end{pmatrix} \subseteq \overline{\operatorname{ran}} \begin{pmatrix} P_0 & 0 \\ P_1 & P_0 \end{pmatrix}. \tag{3.9}$$

**Inductive step.** We continue in the same way for all $k = 1, 2, 3, \dots$. At the $k$th stage, the procedure produces operators $P_0, \dots, P_k$ such that

$$S(k) = \begin{pmatrix} P_0^* & \cdots & P_k^* \\ & \ddots & \vdots \\ 0 & & P_0^* \end{pmatrix} \begin{pmatrix} P_0 & & 0 \\ \vdots & \ddots & \\ P_k & \cdots & P_0 \end{pmatrix}, \tag{3.10}$$

$$\operatorname{ran} \begin{pmatrix} P_1 \\ \vdots \\ P_k \end{pmatrix} \subseteq \overline{\operatorname{ran}} \begin{pmatrix} P_0 & & 0 \\ \vdots & \ddots & \\ P_k & \cdots & P_0 \end{pmatrix}. \tag{3.11}$$

By Lemma 3.3, in the case $k \geq m$,

$$S(k) = \begin{pmatrix} Q_0 & \begin{pmatrix} Q_{-1} & \cdots & Q_{-m} & 0 & \cdots & 0 \end{pmatrix} \\ \begin{pmatrix} Q_{-1}^* \\ \cdots \\ Q_{-m}^* \\ 0 \\ \vdots \\ 0 \end{pmatrix} & S(k-1) \end{pmatrix}. \tag{3.12}$$

The zeros appear here when $k > m$, and their presence leads to the conclusion that $P_k = 0$ for $k > m$. We set then

$$P(z) = P_0 + P_1 z + \cdots + P_m z^m.$$

Comparing (3.10) and (3.12) in the case $k = m$, we deduce $2m+1$ relations which are equivalent to the identity

$$Q(\zeta) = P(\zeta)^* P(\zeta), \qquad \zeta \in \mathbb{T}.$$

**Final step:** $P(z)$ **is outer.** Define $T_P$ as in (2.4). With natural identifications,

$$T_P = \begin{pmatrix} P_0 & \begin{pmatrix} 0 & 0 & \cdots & \end{pmatrix} \\ \begin{pmatrix} P_1 \\ P_2 \\ \vdots \end{pmatrix} & T_P \end{pmatrix}. \tag{3.13}$$

The relations (3.11), combined with the fact that $P_k = 0$ for all $k > m$, imply that

$$\operatorname{ran} \begin{pmatrix} P_1 \\ P_2 \\ \vdots \end{pmatrix} \subseteq \overline{\operatorname{ran}}\, T_P.$$

Hence for any $g \in \mathfrak{G}$, a sequence $f_n$ can be found such that

$$T_P f_n \to \begin{pmatrix} P_1 \\ P_2 \\ \vdots \end{pmatrix} g.$$

Then by (3.13),

$$T_P \begin{pmatrix} g \\ f_n \end{pmatrix} \to \begin{pmatrix} P_0 \\ 0 \end{pmatrix} g.$$

It follows that $\overline{\operatorname{ran}}\, T_P$ contains every vector $(P_0 g, 0, 0, \dots)$ with $g \in \mathfrak{L}(\mathfrak{G})$, and hence $\overline{\operatorname{ran}}\, T_P \supseteq \overline{\operatorname{ran}}\, P_0 \oplus \overline{\operatorname{ran}}\, P_0 \oplus \cdots$. The reverse inclusion holds because by (3.11), the ranges of $P_1, P_2, \dots$ are all contained in $\overline{\operatorname{ran}}\, P_0$. Thus $P(z)$ is outer. □

## 4. Spectral factorization

The problem of spectral factorization is to write a nonnegative operator-valued function $F$ on the unit circle in the form $F = G^*G$ where $G$ is analytic (in a sense made precise below). The terminology comes from prediction theory, where the nonnegative function $F$ plays the role of a spectral density for a multidimensional stationary stochastic process. The problem may be viewed as a generalization of a classical theorem of Szegő from Hardy class theory and the theory of orthogonal polynomials (see Hoffman [35, p. 56] and Szegő [62, Chapter X]).

We write $H^p$ and $L^p$ for the standard Hardy and Lebesgue spaces for the unit disk and unit circle. See Duren [20]. Recall that $\sigma$ is normalized Lebesgue measure on the unit circle $\mathbb{T}$.

**Theorem 4.1 (Szegő's Theorem).** *Let $w \in L^1$ satisfy $w \geq 0$ a.e. on $\mathbb{T}$ and*

$$\int_{\mathbb{T}} \log w(\zeta)\, d\sigma > -\infty.$$

*Then $w = |h|^2$ a.e. on $\mathbb{T}$ for some $h \in H^2$, and $h$ can be chosen to be an outer function.*

Operator and matrix generalizations of Szegő's theorem are stated in Theorems 4.5 and 4.7 below. Some vectorial function theory is needed to formulate these and other results. We assume familiarity with basic concepts but recall a few definitions. For details, see, e.g., [30, 61] and [50, Chapter 4].

In this section, $\mathfrak{G}$ denotes a separable Hilbert space. Functions $f$ and $F$ on the unit circle with values in $\mathfrak{G}$ and $\mathfrak{L}(\mathfrak{G})$, respectively, are called weakly measurable if $\langle f(\zeta), v\rangle$ and $\langle F(\zeta)u, v\rangle$ are measurable for all $u, v \in \mathfrak{G}$. Nontangential limits for analytic functions on the unit disk are taken in the strong (norm) topology for vector-valued functions, and in the strong operator topology for operator-valued functions. We fix notation as follows:

(i) We write $L^2_{\mathfrak{G}}$ and $L^\infty_{\mathfrak{L}(\mathfrak{G})}$ for the standard Lebesgue spaces of weakly measurable functions on the unit circle with values in $\mathfrak{G}$ and $\mathfrak{L}(\mathfrak{G})$.

(ii) Let $H^2_{\mathfrak{G}}$ and $H^\infty_{\mathfrak{L}(\mathfrak{G})}$ be the analogous Hardy classes of analytic functions on the unit disk. We identify elements of these spaces with their nontangential boundary functions, and so the spaces may alternatively be viewed as subspaces of $L^2_{\mathfrak{G}}$ and $L^\infty_{\mathfrak{L}(\mathfrak{G})}$.

(iii) Let $N^+_{\mathfrak{L}(\mathfrak{G})}$ be the space of all analytic functions $F$ on the unit disk such that $\varphi F$ belongs to $H^\infty_{\mathfrak{L}(\mathfrak{G})}$ for some bounded scalar outer function $\varphi$. The elements of $N^+_{\mathfrak{L}(\mathfrak{G})}$ are also identified with their nontangential boundary functions.

A function $F \in H^\infty_{\mathfrak{L}(\mathfrak{G})}$ is called **outer** if $FH^2_{\mathfrak{G}}$ is dense in $H^2_{\mathfrak{F}}$ for some closed subspace $\mathfrak{F}$ of $\mathfrak{G}$. A function $F \in N^+_{\mathfrak{L}(\mathfrak{G})}$ is **outer** if there is a bounded scalar outer function $\varphi$ such that $\varphi F \in H^\infty_{\mathfrak{L}(\mathfrak{G})}$ and $\varphi F$ is outer in the sense just defined. The definition of an outer function given here is consistent with the previously defined notion for polynomials in §2.

A function $A \in H^\infty_{\mathfrak{L}(\mathfrak{G})}$ is called **inner** if multiplication by $A$ on $\mathfrak{H}^2_{\mathfrak{G}}$ is a partial isometry. In this case, the initial space of multiplication by $A$ is a subspace of $\mathfrak{H}^2_{\mathfrak{G}}$ of the form $\mathfrak{H}^2_{\mathfrak{F}}$ where $\mathfrak{F}$ is a closed subspace of $\mathfrak{G}$. To prove this, notice that both the kernel of multiplication by $A$ and the set on which it is isometric are invariant under multiplication by $z$. Therefore the initial space of multiplication by $A$ is a reducing subspace for multiplication by $z$, and so it has the form $\mathfrak{H}^2_{\mathfrak{F}}$ where $\mathfrak{F}$ is a closed subspace of $\mathfrak{G}$ (see [29, p. 106] and [50, p. 96]).

Every $F \in H^\infty_{\mathfrak{L}(\mathfrak{G})}$ has an **inner-outer factorization** $F = AG$, where $A$ is an inner function and $G$ is an outer function. This factorization can be chosen such that the isometric set $H^2_{\mathfrak{F}}$ for multiplication by $A$ on $H^2_{\mathfrak{G}}$ coincides with the closure of the range of multiplication by $G$. The inner-outer factorization is extended in an obvious way to functions $F \in N^+_{\mathfrak{L}(\mathfrak{G})}$. Details are given, for example, in [50, Chapter 5].

The main problem of this section can now be interpreted more precisely:

**Factorization Problem.** *Given a nonnegative weakly measurable function $F$ on $\mathbb{T}$, find a function $G$ in $N^+_{\mathfrak{L}(\mathfrak{G})}$ such that $F = G^*G$ a.e. on $\mathbb{T}$. If such a function exists, we say that $F$ is* **factorable**.

If a factorization exists, the factor $G$ can be chosen to be outer by the inner-outer factorization. Moreover, an outer factor $G$ can be chosen such that $G(0) \geq 0$, and then it is unique [50, p. 101]. By the definition of $N^+_{\mathfrak{L}(\mathfrak{G})}$, a necessary condition for $F$ to be factorable is that

$$\int_{\mathbb{T}} \log^+ \|F(\zeta)\| \, d\sigma < \infty, \tag{4.1}$$

where $\log^+ x$ is zero or $\log x$ according as $0 \leq x \leq 1$ or $1 < x < \infty$, and so we only need consider functions which satisfy (4.1). In fact, in proofs we can usually reduce to the bounded case by considering $F/|\varphi|^2$ for a suitable scalar outer function $\varphi$.

The following result is another view of Lowdenslager's criterion, which we deduce from Lemma 2.3. A direct proof is given in [61, pp. 201–203].

**Lemma 4.2.** *Suppose $F \in L^\infty_{\mathfrak{L}(\mathfrak{G})}$ and $F \geq 0$ a.e. on $\mathbb{T}$. Let $\mathfrak{K}_F$ be the closure of $F^{\frac{1}{2}} H^2_{\mathfrak{G}}$ in $L^2_{\mathfrak{G}}$, and let $S_F$ be the isometry multiplication by $\zeta$ on $\mathfrak{K}_F$. Then $F$ is factorable if and only if $S_F$ is a shift operator, that is,*

$$\bigcap_{n=0}^{\infty} \zeta^n \, \overline{F^{\frac{1}{2}} H^2_{\mathfrak{G}}} = \{0\}. \tag{4.2}$$

*Proof.* In Lemma 2.3 take $\mathfrak{H} = H^2_{\mathfrak{G}}$ viewed as a subspace of $L^2_{\mathfrak{G}}$, and let $S$ be multiplication by $\zeta$ on $\mathfrak{H}$. Define $T \in \mathfrak{L}(\mathfrak{H})$ by $Tf = PFf$, $f \in \mathfrak{H}$, where $P$ is the projection from $L^2_{\mathfrak{G}}$ onto $H^2_{\mathfrak{G}}$. One sees easily that $T$ is a nonnegative Toeplitz operator, and so we can define $\mathfrak{H}_T$ and an isometry $S_T$ as in Lemma 2.3. In fact, $S_T$ is unitarily equivalent to $S_F$ via the natural isomorphism $W \colon \mathfrak{H}_T \to \mathfrak{K}_F$ such that $W(T^{\frac{1}{2}} f) = F^{\frac{1}{2}} f$ for every $f$ in $\mathfrak{H}$. Thus $S_F$ is a shift operator if and only if $S_T$ is a shift operator, and by Lemma 2.3 this is the same as saying that $T = A^*A$ where $A \in \mathfrak{L}(\mathfrak{H})$ is analytic, or equivalently $F$ is factorable [50, p. 110]. □

We obtain a very useful sufficient condition for factorability.

**Theorem 4.3.** *Suppose* $F \in L^\infty_{\mathfrak{L}(\mathfrak{G})}$ *and* $F \geq 0$ *a.e. For* $F$ *to be factorable, it is sufficient that there exists a function* $\psi$ *in* $L^\infty_{\mathfrak{L}(\mathfrak{G})}$ *such that*

(i) $\psi F \in H^\infty_{\mathfrak{L}(\mathfrak{G})}$*;*

(ii) *for all* $\zeta \in \mathbb{T}$ *except at most a set of measure zero,* $\psi(\zeta)\big|\overline{F(\zeta)\mathfrak{G}}$ *is one-to-one.*

*If these conditions are met and* $F = G^*G$ *a.e. with* $G$ *outer, then* $\psi G^* \in H^\infty_{\mathfrak{L}(\mathfrak{G})}$*.*

Theorem 4.3 appears in Rosenblum [49] with $\psi(\zeta) = \zeta^m$ (viewed as an operator-valued function). The case of an arbitrary inner function was proved and applied in a variety of ways by Rosenblum and Rovnyak [50, 51]. V. I. Matsaev first showed that more general functions $\psi$ can be used. Matsaev's result is evidently unpublished, but versions were given by D.Z. Arov [1, Lemma to Theorem 4] and A.S. Markus [41, Theorem 34.3 on p. 199]. Theorem 4.3 includes all of these versions.

We do not know if the conditions (i) and (ii) in Theorem 4.3 are necessary for factorability. It is not hard to see that they are necessary in the simple cases $\dim \mathfrak{G} = 1$ and $\dim \mathfrak{G} = 2$ (for the latter case, one can use [50, Example 1, p. 125]). The general case, however, is open.

*Proof of Theorem* 4.3. Let $F$ satisfy (i) and (ii). Define a subspace $\mathfrak{M}$ of $L^2_{\mathfrak{G}}$ by

$$\mathfrak{M} = \bigcap_{n=0}^{\infty} \zeta^n \, \overline{F^{\frac{1}{2}} H^2_{\mathfrak{G}}} = \bigcap_{n=0}^{\infty} \overline{\zeta^n F^{\frac{1}{2}} H^2_{\mathfrak{G}}}.$$

We show that $\mathfrak{M} = \{0\}$. By (i),

$$\psi F^{\frac{1}{2}} \mathfrak{M} = \psi F^{\frac{1}{2}} \bigcap_{n=0}^{\infty} \overline{\zeta^n F^{\frac{1}{2}} H^2_{\mathfrak{G}}} \subseteq \bigcap_{n=0}^{\infty} \overline{\zeta^n \psi F H^2_{\mathfrak{G}}} \subseteq \bigcap_{n=0}^{\infty} \zeta^n H^2_{\mathfrak{G}} = \{0\}. \tag{4.3}$$

Thus $\psi F^{\frac{1}{2}} \mathfrak{M} = \{0\}$. Now if $g \in \mathfrak{M}$, then $\psi F^{\frac{1}{2}} g = 0$ a.e. by (4.3). Hence $F^{\frac{1}{2}} g = 0$ a.e. by (ii). By the definition of $\mathfrak{M}$, $g \in \overline{F^{\frac{1}{2}} H^2_{\mathfrak{G}}}$, and standard arguments show from this that $g(\zeta) \in \overline{F(\zeta)^{\frac{1}{2}} \mathfrak{G}}$ a.e. Therefore $g = 0$ a.e. It follows that $\mathfrak{M} = \{0\}$, and so $F$ is factorable by Lemma 4.2.

Let $F = G^*G$ a.e. with $G$ outer. We prove that $\psi G^* \in H^\infty_{\mathfrak{L}(\mathfrak{G})}$ by showing that $\psi G^* H^2_{\mathfrak{G}} \subseteq H^2_{\mathfrak{G}}$. Since $G$ is outer, $\overline{G H^2_{\mathfrak{G}}} = H^2_{\mathfrak{F}}$ for some closed subspace $\mathfrak{F}$ of $\mathfrak{G}$. By (i),

$$\psi G^* (G H^2_{\mathfrak{G}}) = \psi F H^2_{\mathfrak{G}} \subseteq H^2_{\mathfrak{G}}.$$

Therefore $\psi G^* H^2_{\mathfrak{F}} \subseteq H^2_{\mathfrak{G}}$. Suppose $f \in H^2_{\mathfrak{G} \ominus \mathfrak{F}}$, and consider any $h \in L^2_{\mathfrak{G}}$. Then

$$\langle G^* f, h \rangle_{L^2_{\mathfrak{G}}} = \int_{\mathbb{T}} \langle f(\zeta), G(\zeta) h(\zeta) \rangle_{\mathfrak{G}} \, d\sigma = 0,$$

because $\operatorname{ran} G(\zeta) \subseteq \mathfrak{F}$ a.e. Thus $\psi G^* f = 0$ a.e. It follows that $\psi G^* H^2_{\mathfrak{G}} \subseteq H^2_{\mathfrak{G}}$, and therefore $\psi G^* \in H^\infty_{\mathfrak{L}(\mathfrak{G})}$. □

For a simple application of Theorem 4.3, suppose that $F$ is a Laurent polynomial of degree $m$, and choose $\psi$ to be $\zeta^m I$. In short order, this yields another proof of the operator Fejér-Riesz theorem (Theorem 2.1).

Another application is a theorem of Sarason [55, p. 198], which generalizes the factorization of a scalar-valued function in $H^1$ as a product of two functions in $H^2$ (see [35, p. 56]).

**Theorem 4.4.** *Every $G$ in $N^+_{\mathfrak{L}(\mathfrak{G})}$ can be written $G = G_1G_2$, where $G_1$ and $G_2$ belong to $N^+_{\mathfrak{L}(\mathfrak{G})}$ and*

$$G_2^*G_2 = [G^*G]^{1/2} \quad \textit{and} \quad G_1^*G_1 = G_2G_2^* \qquad \textit{a.e.}$$

*Proof.* Suppose first that $G \in H^\infty_{\mathfrak{L}(\mathfrak{G})}$. For each $\zeta \in \mathbb{T}$, write

$$G(\zeta) = U(\zeta)[G^*(\zeta)G(\zeta)]^{\frac{1}{2}},$$

where $U(\zeta)$ is a partial isometry with initial space $\overline{\operatorname{ran}}\,[G^*(\zeta)G(\zeta)]^{\frac{1}{2}}$. It can be shown that $U$ is weakly measurable. We apply Theorem 4.3 with $F = [G^*G]^{\frac{1}{2}}$ and $\psi = U$. Conditions (i) and (ii) of Theorem 4.3 are obviously satisfied, and so we obtain an outer function $G_2 \in H^\infty_{\mathfrak{L}(\mathfrak{G})}$ such that

$$G_2^*G_2 = [G^*G]^{1/2} \qquad \text{a.e.}$$

and $UG_2^* \in H^\infty_{\mathfrak{L}(\mathfrak{G})}$. Set $G_1 = UG_2^*$. By construction $G_1 \in H^\infty_{\mathfrak{L}(\mathfrak{G})}$,

$$G = U(G^*G)^{\frac{1}{2}} = (UG_2^*)G_2 = G_1G_2,$$

and $G_1^*G_1 = G_2U^*UG_2^* = G_2G_2^*$ a.e. The result follows when $G \in H^\infty_{\mathfrak{L}(\mathfrak{G})}$.

The general case follows on applying what has just been shown to $\varphi^2G$, where $\varphi$ is a scalar-valued outer function such that $\varphi^2G \in H^\infty_{\mathfrak{L}(\mathfrak{G})}$. □

The standard operator generalization of Szegő's theorem also follows from Theorem 4.3.

**Theorem 4.5.** *Let $F$ be a weakly measurable function on $\mathbb{T}$ having invertible values in $\mathfrak{L}(\mathfrak{G})$ satisfying $F \geq 0$ a.e. If*

$$\int_{\mathbb{T}} \log^+ \|F(\zeta)\| \, d\sigma < \infty \quad \textit{and} \quad \int_{\mathbb{T}} \log^+ \|F(\zeta)^{-1}\| \, d\sigma < \infty,$$

*then $F$ is factorable.*

*Proof.* Since $\log^+ \|F(\zeta)\|$ is integrable, we can choose a scalar outer function $\varphi_1$ such that

$$F_1 = F/|\varphi_1|^2 \in L^\infty_{\mathfrak{L}(\mathfrak{G})}.$$

Since $\log^+ \|F(\zeta)^{-1}\|$ is integrable, so is $\log^+ \|F_1(\zeta)^{-1}\|$. Hence there is a bounded scalar outer function $\varphi$ such that

$$\varphi F_1^{-1} \in L^\infty_{\mathfrak{L}(\mathfrak{G})}.$$

We apply Theorem 4.3 to $F_1$ with $\psi = \varphi F_1^{-1}$. Condition (i) is satisfied because $\psi F_1 = \varphi I$. Condition (ii) holds because the values of $\psi$ are invertible a.e. Thus $F_1$ is factorable, and hence so is $F$. $\square$

Theorem 4.3 has a half-plane version, the scalar inner case of which is given in [50, p. 117]. This has an application to the following generalization of Akhiezer's theorem on factoring entire functions [50, Chapter 6].

**Theorem 4.6.** *Let $F$ be an entire function of exponential type $\tau$, having values in $\mathfrak{L}(\mathfrak{G})$, such that $F(x) \geq 0$ for all real $x$ and*

$$\int_{-\infty}^{\infty} \frac{\log^+ \|F(t)\|}{1+t^2}\, dt < \infty.$$

*Then $F(x) = G(x)^* G(x)$ for all real $x$ where $G$ is an entire function with values in $\mathfrak{L}(\mathfrak{G})$ such that $\exp(-i\tau z/2)G$ is of exponential type $\tau/2$ and the restriction of $G$ to the upper half-plane is an outer function.*

**Matrix case**

We end this section by quoting a few results for matrix-valued functions. The matrix setting is more concrete, and one can do more. Statements often require invertibility assumptions. We give no details and leave it to the interested reader to consult other sources for further information.

Our previous definitions and results transfer in an obvious way to matrix-valued functions. For this we choose $\mathfrak{G} = \mathbb{C}^r$ for some positive integer $r$ and identify operators on $\mathbb{C}^r$ with $r \times r$ matrices. The operator norm of a matrix is denoted $\|\cdot\|$. We write $L^\infty_{r\times r}, H^\infty_{r\times r}$ in place of $L^\infty_{\mathfrak{L}(\mathfrak{G})}, H^\infty_{\mathfrak{L}(\mathfrak{G})}$ and $\|\cdot\|_\infty$ for the norms on these spaces.

Theorem 4.5 is more commonly stated in a different form for matrix-valued functions.

**Theorem 4.7.** *Suppose that $F$ is an $r \times r$ measurable matrix-valued function having invertible values on $\mathbb{T}$ such that $F \geq 0$ a.e. and $\log^+ \|F\|$ is integrable. Then $F$ is factorable if and only if $\log\det F$ is integrable.*

Recall that when $F$ is factorable, there is a unique outer $G$ such that $F = G^*G$ and $G(0) \geq 0$. It makes sense to inquire about the continuity properties of the mapping $\Phi\colon F \to G$ with respect to various norms. For example, see Jacob and Partington [37]. We cite one recent result in this area.

**Theorem 4.8 (Barclay [5]).** *Let $F, F_n$, $n = 1, 2, \ldots$, be $r \times r$ measurable matrix-valued functions on $\mathbb{T}$ having invertible values a.e. and integrable norms. Suppose that $F = G^*G$ and $F_n = G_n^* G_n$, where $G, G_n$ are $r \times r$ matrix-valued outer functions such that $G(0) \geq 0$ and $G_n(0) \geq 0$, $n = 1, 2, \ldots$. Then*

$$\lim_{n\to\infty} \int_{\mathbb{T}} \|G(\zeta) - G_n(\zeta)\|^2\, d\sigma = 0$$

*if and only if*

(i) $\lim_{n\to\infty} \int_{\mathbb{T}} \|F(\zeta) - F_n(\zeta)\| \, d\sigma = 0$, *and*

(ii) *the family of functions* $\{\log\det F_n\}_{n=0}^{\infty}$ *is uniformly integrable.*

A family of functions $\{\varphi_\alpha\}_{\alpha\in A} \subseteq L^1$ is **uniformly integrable** if for every $\varepsilon > 0$ there is a $\delta > 0$ such that $\int_E |\varphi_\alpha| \, d\sigma < \varepsilon$ for all $\alpha \in A$ whenever $\sigma(E) < \delta$. See [5] for additional references and similar results in other norms.

A theorem of Bourgain [9] characterizes all functions on the unit circle which are products $\bar{h}g$ with $g, h \in H^\infty$: *A function* $f \in L^\infty$ *has the form* $f = \bar{h}g$ *where* $g, h \in H^\infty$ *if and only if* $\log|f|$ *is integrable.* This resolves a problem of Douglas and Rudin [17]. The problem is more delicate than spectral factorization; when $|f| = 1$ a.e., the factorization cannot be achieved in general with inner functions. Bourgain's theorem was recently generalized to matrix-valued functions.

**Theorem 4.9 (Barclay [4, 6]).** *Suppose* $F \in L^\infty_{r\times r}$ *has invertible values a.e. Then* $F$ *has the form* $F = H^*G$ *a.e. for some* $G, H$ *in* $H^\infty_{r\times r}$ *if and only if* $\log|\det F|$ *is integrable. In this case, for every* $\varepsilon > 0$ *such a factorization can be found with*

$$\|G\|_\infty \|H\|_\infty < \|F\|_\infty + \varepsilon.$$

The proof of Theorem 4.9 in [6] is long and technical. In fact, Barclay proves an $L^p$-version of this result for all $p$, $1 \le p \le \infty$.

Another type of generalization is factorization with indices. We quote one result to illustrate this notion.

**Theorem 4.10.** *Let* $F$ *be an* $r \times r$ *matrix-valued function with rational entries. Assume that* $F$ *has no poles on* $\mathbb{T}$ *and that* $\det F(\zeta) \neq 0$ *for all* $\zeta$ *in* $\mathbb{T}$. *Then there exist integers* $\varkappa_1 \le \varkappa_2 \le \cdots \le \varkappa_r$ *such that*

$$F(z) = F_-(z)\operatorname{diag}\{z^{\varkappa_1}, \dots, z^{\varkappa_r}\}F_+(z),$$

*where* $F_\pm$ *are* $r \times r$ *matrix-valued functions with rational entries such that*

(i) $F_+(z)$ *has no poles for* $|z| \le 1$ *and* $\det F_+(z) \neq 0$ *for* $|z| \le 1$;

(ii) $F_-(z)$ *has no poles for* $|z| \ge 1$ *including* $z = \infty$ *and* $\det F_-(z) \neq 0$ *for* $|z| \ge 1$ *including* $z = \infty$.

The case in which $F$ is nonnegative on $\mathbb{T}$ can be handled using the operator Fejér-Riesz theorem (the indices are all zero in this case). The general case is given in Gohberg, Goldberg, and Kaashoek [27, pp. 236–239]. This is a large subject that includes, for example, general theories of factorization in Bart, Gohberg, Kaashoek, and Ran [7] and Clancey and Gohberg [13].

### Historical remarks

Historical accounts of spectral factorization appear in [2, 30, 50, 52, 61]. Briefly, the problem of factoring nonnegative matrix-valued functions on the unit circle rose to prominence in the prediction theory of multivariate stationary stochastic processes. The first results of this theory were announced by Zasuhin [65] without complete proofs; proofs were supplied by M.G. Kreĭn in lectures. Modern accounts of prediction theory and matrix generalizations of Szegő's theorem are based on fundamental papers of Helson and Lowdenslager [31, 32], and Wiener and Masani [63, 64]. The general case of Theorem 4.5 is due to Devinatz [15]; other proofs are given in [16, 30, 50]. For an engineering view and computational methods, see [38, Chapter 8] and [56].

The original source for Lowdenslager's Criterion (Lemmas 2.3 and 4.2) is [40]; an error in [40] was corrected by Douglas [16]. There is a generalization, given by Sz.-Nagy and Foias [61, pp. 201–203], in which the isometry may have a nontrivial unitary component and the shift component yields a maximal factorable summand. Lowdenslager's Criterion is used in the construction of canonical models of operators by de Branges [10]. See also Constantinescu [14] for an adaptation to Toeplitz kernels and additional references.

## 5. Multivariable theory

It is natural to wonder to what extent the results for one variable carry over to several variables. Various interpretations of "several variables" are possible. The most straightforward is to consider Laurent polynomials in complex variables $z_1, \dots, z_d$ that are nonnegative on the $d$-torus $\mathbb{T}^d$. The method of Schur complements in §3 suggests an approach to the factorization problem for such polynomials. Care is needed, however, since the first conjectures for a multivariable Fejér-Riesz theorem that might come to mind are false, as explained below. Multivariable generalizations of the Fejér-Riesz theorem are thus necessarily weaker than the one-variable result. One difficulty has to do with degrees, and if the condition on degrees is relaxed, there is a neat result in the strictly positive case (Theorem 5.1).

By a Laurent polynomial in $z = (z_1, \dots, z_d)$ we understand an expression

$$Q(z) = \sum_{k_1=-m_1}^{m_1} \cdots \sum_{k_d=-m_d}^{m_d} Q_{k_1,\dots,k_d} z_1^{k_1} \cdots z_d^{k_d}. \tag{5.1}$$

We assume that the coefficients belong to $\mathfrak{L}(\mathfrak{G})$, where $\mathfrak{G}$ is a Hilbert space. With obvious interpretations, the scalar case is included. By an analytic polynomial with coefficients in $\mathfrak{L}(\mathfrak{G})$ we mean an analogous expression, of the form

$$P(z) = \sum_{k_1=0}^{m_1} \cdots \sum_{k_d=0}^{m_d} P_{k_1,\dots,k_d} z_1^{k_1} \cdots z_d^{k_d}. \tag{5.2}$$

The numbers $m_1, \dots, m_d$ in (5.1) and (5.2) are upper bounds for the degrees of the polynomials in $z_1, \dots, z_d$, which we define as the smallest values of $m_1, \dots, m_d$ that can be used in the representations (5.1) and (5.2).

Suppose that $Q(z)$ has the form (5.1) and satisfies $Q(\zeta) \ge 0$ for all $\zeta \in \mathbb{T}^d$, that is, for all $\zeta = (\zeta_1, \dots, \zeta_d)$ with $|\zeta_1| = \cdots = |\zeta_d| = 1$. Already in the scalar case, one cannot always find an analytic polynomial $P(z)$ such that $Q(\zeta) = P(\zeta)^* P(\zeta)$, $\zeta \in \mathbb{T}^d$. This was first explicitly shown by Lebow and Schreiber [39]. There are also difficulties in writing $Q(\zeta) = \sum_{j=1}^{r} P_j(\zeta)^* P_j(\zeta)$, $\zeta \in \mathbb{T}^d$, for some finite set of analytic polynomials, at least if one requires that the degrees of the analytic polynomials do not exceed those of $Q(z)$ as in the one-variable case (see Naftalevich and Schreiber [44], Rudin [53], and Sakhnovich [54, §3.6]). The example in [44] is based on a Cayley transform of a version of a real polynomial over $\mathbb{R}^2$ called Motzkin's polynomial, which was the first explicit example of a nonnegative polynomial in $\mathbb{R}^d$, $d > 1$, which is not a sum of squares of polynomials. What is not mentioned in these sources is that if we loosen the restriction on degrees, the polynomial in [44] can be written as a sum of squares (see [19]). Nevertheless, for three or more variables, very general results of Scheiderer [57] imply that there exist nonnegative, but not strictly positive, polynomials which cannot be expressed as such finite sums regardless of degrees.

**Theorem 5.1.** *Let $Q(z)$ be a Laurent polynomial in $z = (z_1, \dots, z_d)$ with coefficients in $\mathfrak{L}(\mathfrak{G})$ for some Hilbert space $\mathfrak{G}$. Suppose that there is a $\delta > 0$ such that $Q(\zeta) \geq \delta I$ for all $\zeta \in \mathbb{T}^d$. Then*

$$Q(\zeta) = \sum_{j=1}^{r} P_j(\zeta)^* P_j(\zeta), \qquad \zeta \in \mathbb{T}^d, \tag{5.3}$$

*for some analytic polynomials $P_1(z), \dots, P_r(z)$ in $z = (z_1, \dots, z_d)$ which have coefficients in $\mathfrak{L}(\mathfrak{G})$. Furthermore, for any fixed $k$, the representation* (5.3) *can be chosen such that the degree of each analytic polynomial in $z_k$ is no more than the degree of $Q(z)$ in $z_k$.*

The scalar case of Theorem 5.1 follows by a theorem of Schmüdgen [58], which states that strictly positive polynomials over compact semialgebraic sets in $\mathbb{R}^n$ (that is, sets which are expressible in terms of finitely many polynomial inequalities) can be written as weighted sums of squares, where the weights are the polynomials used to define the semialgebraic set (see also [12]); the proof is nonconstructive. On the other hand, the proof we sketch using Schur complements covers the operator-valued case, and it gives an algorithm for finding the solution. One can also give estimates for the degrees of the polynomials involved, though we have not stated these.

We prove Theorem 5.1 for the case $d = 2$, following Dritschel [18]. The general case is similar. The argument mimics the method of Schur complements, especially in its original form used in [18]. In place of Toeplitz matrices whose entries are operators, in the case of two variables we use Toeplitz matrices whose entries are themselves Toeplitz matrices. The fact that the first level Toeplitz blocks are infinite in size causes problems, and so we truncate these blocks to finite size. Then everything goes through, but instead of factoring the original polynomial $Q(z)$, the result is a factorization of polynomials $Q^{(N)}(z)$ that are close to $Q(z)$. When $Q(\zeta) \geq \delta I$ on $\mathbb{T}^d$ for some $\delta > 0$, there is enough wiggle room to factor $Q(z)$ itself. We isolate the main steps in a lemma.

**Lemma 5.2.** *Let*

$$Q(z) = \sum_{j=-m_1}^{m_1} \sum_{k=-m_2}^{m_2} Q_{jk} z_1^j z_2^k$$

*be a Laurent polynomial with coefficients in* $\mathfrak{L}(\mathfrak{G})$ *such that* $Q(\zeta) \geq 0$ *for all* $\zeta = (\zeta_1, \zeta_2)$ *in* $\mathbb{T}^2$*. Set*

$$Q^{(N)}(z) = \sum_{j=-m_1}^{m_1} \sum_{k=-m_2}^{m_2} \frac{N+1-|k|}{N+1} Q_{jk}\, z_1^j z_2^k.$$

*Then for each* $N \geq m_2$*, there are analytic polynomials*

$$F_\ell(z) = \sum_{j=0}^{m_1} \sum_{k=0}^{N} F_{jk}^{(\ell)}\, z_1^j z_2^k, \qquad \ell = 0, \dots, N, \tag{5.4}$$

*with coefficients in* $\mathfrak{L}(\mathfrak{G})$ *such that*

$$Q^{(N)}(\zeta) = \sum_{\ell=0}^{N} F_\ell(\zeta)^* F_\ell(\zeta), \qquad \zeta \in \mathbb{T}^2. \tag{5.5}$$

*Proof.* Write

$$Q(z) = \sum_{j=-m_1}^{m_1} \Big( \sum_{k=-m_2}^{m_2} Q_{jk} z_2^k \Big) z_1^j = \sum_{j=-m_1}^{m_1} R_j(z_2)\, z_1^j,$$

and extend all sums to run from $-\infty$ to $\infty$ by setting $Q_{jk} = 0$ and $R_j(z_2) = 0$ if $|j| > m_1$ or $|k| > m_2$. Introduce a Toeplitz matrix $T$ whose entries are the Toeplitz matrices $T_j$ corresponding to the Laurent polynomials $R_j(z_2)$, that is,

$$T = \begin{pmatrix} T_0 & T_{-1} & T_{-2} & \cdots \\ T_1 & T_0 & T_{-1} & \ddots \\ T_2 & T_1 & T_0 & \ddots \\ \vdots & \ddots & \ddots & \ddots \end{pmatrix}, \qquad T_j = \begin{pmatrix} Q_{j0} & Q_{j,-1} & Q_{j,-2} & \cdots \\ Q_{j1} & Q_{j0} & Q_{j,-1} & \ddots \\ Q_{j2} & Q_{j1} & Q_{j0} & \ddots \\ \vdots & \ddots & \ddots & \ddots \end{pmatrix},$$

$j = 0, \pm 1, \pm 2, \dots$. Notice that $T$ is finitely banded, since $T_j = 0$ for $|j| > m_1$. The identity (2.5) has the following generalization:

$$\langle Th, h \rangle = \sum_{p=0}^{\infty} \sum_{q=0}^{\infty} \langle T_{q-p} h_p, h_q \rangle = \int_{\mathbb{T}^2} \langle Q(\zeta) h(\zeta), h(\zeta) \rangle_{\mathfrak{G}}\, d\sigma_2(\zeta)\,.$$

Here $\zeta = (\zeta_1, \zeta_2)$ and $d\sigma_2(\zeta) = d\sigma(\zeta_1) d\sigma(\zeta_2)$. Also,

$$h(\zeta) = \sum_{p=0}^{\infty} \sum_{q=0}^{\infty} h_{pq}\, \zeta_1^p \zeta_2^q\,,$$

where the coefficients are vectors in $\mathfrak{G}$ and all but finitely many are zero, and

$$h = \begin{pmatrix} h_0 \\ h_1 \\ \vdots \end{pmatrix}, \qquad h_p = \begin{pmatrix} h_{p0} \\ h_{p1} \\ \vdots \end{pmatrix}, \qquad p = 0, 1, 2, \dots .$$

It follows that $T$ acts as a bounded operator on a suitable direct sum of copies of $\mathfrak{G}$. Since $Q(\zeta) \geq 0$ on $\mathbb{T}^2$, $T \geq 0$.

Fix $N \geq m_2$. Set

$$T' = \begin{pmatrix} T_0' & T_{-1}' & T_{-2}' & \cdots \\ T_1' & T_0' & T_{-1}' & \ddots \\ T_2' & T_1' & T_0' & \ddots \\ \vdots & \ddots & \ddots & \ddots \end{pmatrix},$$

where $T_j'$ is the upper $(N+1) \times (N+1)$ block of $T_j$ with a normalizing factor:

$$T_j' = \frac{1}{N+1} \begin{pmatrix} Q_{j0} & Q_{j,-1} & \cdots & Q_{j,-N} \\ Q_{j1} & Q_{j0} & \cdots & Q_{j,-N+1} \\ & & \cdots & \\ Q_{jN} & Q_{j,N-1} & \cdots & Q_{j0} \end{pmatrix}, \qquad j = 0, \pm 1, \pm 2, \dots .$$

Then $T'$ is the Toeplitz matrix corresponding to the Laurent polynomial

$$\Psi(w) = \sum_{j=-m_1}^{m_1} T_j' \, w^j .$$

Moreover, $T' \geq 0$ since it is a positive constant multiple of a compression of $T$. Thus $\Psi(w) \geq 0$ for $|w| = 1$. By the operator Fejér-Riesz theorem (Theorem 2.1),

$$\Psi(w) = \Phi(w)^* \Phi(w), \qquad |w| = 1, \tag{5.6}$$

for some analytic polynomial $\Phi(w) = \sum_{j=0}^{m_1} \Phi_j w^j$ whose coefficients are $(N+1) \times (N+1)$ matrices with entries in $\mathfrak{L}(\mathfrak{G})$. Write

$$\Phi_j = \begin{pmatrix} \Phi_{jN} & \Phi_{j,N-1} & \cdots & \Phi_{j0} \end{pmatrix},$$

where $\Phi_{jk}$ is the $k$th column in $\Phi_j$. Set

$$\widetilde{F}(z) = \sum_{j=0}^{m_1} \sum_{k=0}^{N} \Phi_{jk} \, z_1^j z_2^k .$$

The identity (5.6) is equivalent to $2m_1 + 1$ relations for the coefficients of $\Psi(w)$. The coefficients of $\Psi(w)$ are constant on diagonals, there being $N + 1 - k$ terms in the $k$th diagonal above the main diagonal, and similarly below. If these terms are summed, the result gives $2m_1 + 1$ relations equivalent to the identity

$$Q^{(N)}(\zeta) = \widetilde{F}(\zeta)^* \widetilde{F}(\zeta), \qquad \zeta \in \mathbb{T}^2. \tag{5.7}$$

We omit the calculation, which is straightforward but laborious. To convert (5.7) to the form (5.5), write

$$\Phi_{jk} = \begin{pmatrix} F_{jk}^{(0)} \\ F_{jk}^{(1)} \\ \vdots \\ F_{jk}^{(N)} \end{pmatrix}, \qquad j = 0, \dots, m_1 \text{ and } k = 0, \dots, N.$$

Then

$$\widetilde{F}(z) = \begin{pmatrix} F_0(z) \\ F_1(z) \\ \vdots \\ F_N(z) \end{pmatrix},$$

where $F_0(z), \dots, F_N(z)$ are given by (5.4), and so (5.7) takes the form (5.5). □

*Proof of Theorem* 5.1 *for the case $d = 2$.* Suppose $N \geq m_2$, and set

$$\widetilde{Q}(z) = \sum_{j=-m_1}^{m_1} \sum_{k=-m_2}^{m_2} \frac{N+1}{N+1-|k|} Q_{jk}\, z_1^j z_2^k.$$

The values of $\widetilde{Q}(z)$ are selfadjoint on $\mathbb{T}^2$, and $\widetilde{Q}(z) = Q(z) + S(z)$, where

$$S(z) = \sum_{j=-m_1}^{m_1} \sum_{k=-m_2}^{m_2} \frac{|k|}{N+1-|k|} Q_{jk}\, z_1^j z_2^k.$$

Now choose $N$ large enough that $\|S(\zeta)\| < \delta$, $\zeta \in \mathbb{T}^2$. Then $\widetilde{Q}(\zeta) \geq 0$ on $\mathbb{T}^2$, and the result follows on applying Lemma 5.2 to $\widetilde{Q}(z)$. □

Further details can be found in [18], and a variation on this method yielding good numerical results is given in Geronimo and Lai [23].

While, as we mentioned, there is in general little hope of finding a factorization of a positive trigonometric polynomial in two or more variables in terms of one or more analytic polynomials of the same degree, it happens that there are situations where the existence of such a factorization is important. In particular, Geronimo and Woerdeman consider this question in the context of the autoregressive filter problem [24, 25], with the first paper addressing the scalar case and the second the operator-valued case, both in two variables. They show that for scalar-valued polynomials in this setting there exists a factorization in terms of a single **stable** (so invertible in the bidisk $\mathbb{D}^2$) analytic polynomial of the same degree if and only if a full rank condition holds for certain submatrices of the associated Toeplitz matrix ([24, Theorem 1.1.3]). The condition for operator-valued polynomials is similar, but more complicated to state. We refer the reader to the original papers for details.

Stable scalar polynomials in one variable are by definition outer, so the Geronimo and Woerdeman results can be viewed as a statement about outer factorizations in two variables. In [19], a different notion of outerness is considered. As we saw in §3, in one variable outer factorizations can be extracted using Schur complements. The same Schur complement method in two or more variables gives rise to a version of "outer" factorization which in general does not agree with that coming from stable polynomials. In [19], this Schur complement version of outerness is used when considering outer factorizations for polynomials in two or more variables. As in the Geronimo and Woerdeman papers, it is required that the factorization be in terms of a single analytic polynomial of the same degree as the polynomial being factored. Then necessary and sufficient conditions for such an outer factorization under these constraints are found ([19, Theorem 4.1]).

The problem of spectral factorization can also be considered in the multivariable setting. Blower [8] has several results along these lines for bivariate matrix-valued functions, including a matrix analogue of Szegő's theorem similar to Theorem 4.7. His results are based on a two-variable matrix version of Theorem 5.1, and the arguments he gives coupled with Theorem 5.1 can be used to extend these results to polynomials in $d > 2$ variables as well.

## 6. Noncommutative factorization

We now present some noncommutative interpretations of the notion of "several variables," starting with the one most frequently considered, and for which there is an analogue of the Fejér-Riesz theorem. It is due to Scott McCullough and comes very close to the one-variable result. Further generalizations have been obtained by Helton, McCullough and Putinar in [33]. For a broad overview of the area, two nice survey articles have recently appeared by Helton and Putinar [34] and Schmüdgen [59] covering noncommutative real algebraic geometry, of which the noncommutative analogues of the Fejér-Riesz theorem are one aspect.

In keeping with the assumptions made in [42], all Hilbert spaces in this section are taken to be separable. Fix Hilbert spaces $\mathfrak{G}$ and $\mathfrak{H}$, and assume that $\mathfrak{H}$ is infinite dimensional.

Let $S$ be the free semigroup with generators $a_1, \dots, a_d$. Thus $S$ is the set of words

$$w = a_{j_1} \cdots a_{j_k}, \quad j_1, \dots, j_k \in \{1, \dots, d\}, \quad k = 0, 1, 2, \dots , \tag{6.1}$$

with the binary operation concatenation. The empty word is denoted $e$. The length of the word (6.1) is $|w| = k$ (so $|e| = 0$). Let $S_m$ be the set of all words (6.1) of length at most $m$. The cardinality of $S_m$ is $\ell_m = 1 + d + d^2 + \cdots + d^m$.

We extend $S$ to a free group $G$. We can think of the elements of $G$ as words in $a_1, \dots, a_d, a_1^{-1}, \dots, a_d^{-1}$, with two such words identified if one can be obtained from the other by cancelling adjacent terms of the form $a_j$ and $a_j^{-1}$. The binary operation in $G$ is also concatenation. Words in $G$ of the form $h = v^{-1}w$ with $v, w \in S$ play a special role and are called **hereditary**. Notice that a hereditary word $h$ has many representations $h = v^{-1}w$ with $v, w \in S$. Let $H_m$ be the set of hereditary words $h$ which have at least one representation in the form $h = v^{-1}w$ with $v, w \in S_m$.

We can now introduce the noncommutative analogues of Laurent and analytic polynomials. A hereditary polynomial is a formal expression

$$Q = \sum_{h \in H_m} h \otimes Q_h, \tag{6.2}$$

where $Q_h \in \mathfrak{L}(\mathfrak{G})$ for all $h$. Analytic polynomials are hereditary polynomials of the special form

$$P = \sum_{w \in S_m} w \otimes P_w, \tag{6.3}$$

where $P_w \in \mathfrak{L}(\mathfrak{G})$ for all $w$. The identity

$$Q = P^*P$$

is defined to mean that

$$Q_h = \sum_{\substack{v,w \in S_m \\ h = v^{-1}w}} P_v^* P_w, \qquad h \in H_d.$$

Next we give meaning to the expressions $Q(U)$ and $P(U)$ for hereditary and analytic polynomials (6.2) and (6.3) and any tuple $U = (U_1, \dots, U_d)$ of unitary operators on $\mathfrak{H}$. First define $U^w \in \mathfrak{L}(\mathfrak{H})$ for any $w \in S$ by writing $w$ in the form (6.1) and setting

$$U^w = U_{j_1} \cdots U_{j_k}.$$

By convention, $U^e = I$ is the identity operator on $\mathfrak{H}$. If $h \in \mathfrak{G}$ is a hereditary word, set

$$U^h = (U^v)^* U^w$$

for any representation $h = v^{-1}w$ with $v, w \in S$; this definition does not depend on the choice of representation. Finally, define $Q(U), P(U) \in \mathfrak{L}(\mathfrak{H} \otimes \mathfrak{G})$ by

$$Q(U) = \sum_{h \in H_m} U^h \otimes Q_h\,, \qquad P(U) = \sum_{w \in S_m} U^w \otimes P_w.$$

The reader is referred to, for example, Murphy [43, §6.3] for the construction of tensor products of Hilbert spaces and algebras, or Palmer, [45, §1.10] for a more detailed account.

**Theorem 6.1 (McCullough [42]).** *Let*

$$Q = \sum_{h \in H_m} h \otimes Q_h$$

*be a hereditary polynomial with coefficients in $\mathfrak{L}(\mathfrak{G})$ such that $Q(U) \geq 0$ for every tuple $U = (U_1, \dots, U_d)$ of unitary operators on $\mathfrak{H}$. Then for some $\ell \leq \ell_m$, there* *exist analytic polynomials*

$$P_j = \sum_{w \in S_m} w \otimes P_{j,w}, \qquad j = 1, \dots, \ell,$$

*with coefficients in* $\mathfrak{L}(\mathfrak{G})$ *such that*

$$Q = P_1^* P_1 + \cdots + P_\ell^* P_\ell.$$

*Moreover, for any tuple* $U = (U_1, \dots, U_d)$ *of unitary operators on* $\mathfrak{H}$*,*

$$Q(U) = P_1(U)^* P_1(U) + \cdots + P_\ell(U)^* P_\ell(U).$$

*In these statements, when* $\mathfrak{G}$ *is infinite dimensional, we can choose* $\ell = 1$*.*

As noted by McCullough, when $d = 1$, Theorem 6.1 gives a weaker version of Theorem 2.1. However, Theorem 2.1 can be deduced from this by a judicious use of Beurling's theorem and an inner-outer factorization.

McCullough's theorem uses one of many possible choices of noncommutative spaces on which some form of trigonometric polynomials can be defined. We place this, along with the commutative versions, within a general framework, which we now explain.

The complex scalar-valued trigonometric polynomials in $d$ variables form a unital $*$-algebra $\mathfrak{P}$, the involution taking $z^n$ to $z^{-n}$, where for $n = (n_1, \dots, n_d)$, $-n = (-n_1, \dots, -n_d)$. If instead the coefficients are in the algebra $\mathfrak{C} = \mathfrak{L}(\mathfrak{G})$ for some Hilbert space $\mathfrak{G}$, then the unital involutive algebra of trigonometric polynomials with coefficients in $\mathfrak{C}$ is $\mathfrak{P} \otimes \mathfrak{C}$. The unit is $1 \otimes 1$. A representation of $\mathfrak{P} \otimes \mathfrak{C}$ is a unital algebra $*$-homomorphism from $\mathfrak{P} \otimes \mathfrak{C}$ into $\mathfrak{L}(\mathfrak{H})$ for a Hilbert space $\mathfrak{H}$. The key thing here is that $z_1, \dots, z_d$ generate $\mathfrak{C}$, and so assuming we do not mess with the coefficient space, a representation $\pi$ is determined by specifying $\pi(z_k)$, $k = 1, \dots, d$.

First note that since $z_k^* z_k = 1$, $\pi(z_k)$ is isometric, and since $z_k^* = z_k^{-1}$, we then have that $\pi(z_k)$ is unitary. Assuming the variables commute, the $z_k$s generate a commutative group $G$ which we can identify with $\mathbb{Z}^d$ under addition, and the irreducible representations of commutative groups are one dimensional. This essentially follows from the spectral theory for normal operators (see, for example, Edwards [21, p. 718]). However, the one-dimensional representations are point evaluations on $\mathbb{T}^d$. Discrete groups with the discrete topology are examples of locally compact groups. Group representations of locally compact groups extend naturally to the algebraic group algebra, which in this case is $\mathfrak{P}$, and then on to the algebra $\mathfrak{P} \otimes \mathfrak{C}$ by tensoring with the identity representation of $\mathfrak{C}$. So a seemingly more complex way of stating that a commutative trigonometric polynomial $P$ in several variables is positive / strictly positive is to say that for each (topologically) irreducible unitary representation $\pi$ of $G$, the extension of $\pi$ to a unital $*$-representation of the algebra $\mathfrak{P} \otimes \mathfrak{C}$, also called $\pi$, has the property that $\pi(P) \geq 0$ / $\pi(P) > 0$. By the way, since $\mathbb{T}^d$ is compact, $\pi(P) > 0$ implies the existence of some $\epsilon > 0$ such that $\pi(P - \epsilon 1 \otimes 1) = \pi(P) - \epsilon 1 \geq 0$.

What is gained through this perspective is that we may now define noncommutative trigonometric polynomials over a finitely generated discrete (so locally compact) group $G$ in precisely the same manner. These are the elements of the algebraic group algebra $\mathfrak{P}$ generated by $G$; that is, formal complex linear combinations of elements of $G$ endowed with pointwise addition and a convolution product (see Palmer [45, Section 1.9]). Then a trigonometric polynomial in $\mathfrak{P} \otimes \mathfrak{C}$ is formally a finite sum over $G$ of the form $P = \sum_g g \otimes P_g$ where $P_g \in \mathfrak{C}$ for all $g$.

We also introduce an involution by setting $g^* = g^{-1}$ for $g \in G$. A trigonometric polynomial $P$ is **selfadjoint** if for all $g$, $P_{g^*} = P_g^*$. There is an order structure on selfadjoint elements defined by saying that a selfadjoint polynomial $P$ is **positive / strictly positive** if for every irreducible unital $*$-representation $\pi$ of $G$, the extension as above of $\pi$ to the algebra $\mathfrak{P} \otimes \mathfrak{C}$ (again called $\pi$), satisfies $\pi(P) \geq 0$ / $\pi(P) > 0$; where by $\pi(P) > 0$ we mean that there exists some $\epsilon > 0$ independent of $\pi$ such that $\pi(P - \epsilon(1 \otimes 1)) \geq 0$. Letting $\Omega$ represent the set of such irreducible representations, we can in a manner suggestive of the Gel'fand transform define $\hat{P}(\pi) = \pi(P)$, and in this way think of $\Omega$ as a sort of noncommutative space on which our polynomial is defined. The Gel'fand-Raĭkov theorem (see, for example, Palmer [46, Theorem 12.4.6]) ensures the existence of sufficiently many irreducible representations to separate $G$, so in particular, $\Omega \neq \emptyset$.

For a finitely generated discrete group $G$ with generators $\{a_1, \dots, a_d\}$, let $S$ be a fixed unital subsemigroup of $G$ containing the generators. The most interesting case is when $S$ is the subsemigroup generated by $e$ (the group identity) and $\{a_1, \dots, a_d\}$. As an example of this, if $G$ is the noncommutative free group in $d$ generators, then the unital subsemigroup generated by $\{a_1, \dots, a_d\}$ consists of group elements $w$ of the form $e$ (for the empty word) and those which are an arbitrary finite product of positive powers of the generators, as in (6.1).

We also need to address the issue of what should play the role of Laurent and analytic trigonometric polynomials in the noncommutative setting. The **hereditary** trigonometric polynomials are defined as those polynomials of the form $P = \sum_j w_{j1}^* w_{j2} \otimes P_j$, where $w_{j1}, w_{j2} \in S$. We think of these as the Laurent polynomials. Trigonometric polynomials over $S$ are referred to as **analytic** polynomials. The **square** of an analytic polynomial $Q$ is the hereditary trigonometric polynomial $Q^*Q$. Squares are easily seen to be positive. As a weak analogue of the Fejér-Riesz theorem, we prove a partial converse below.

We refer to those hereditary polynomials which are selfadjoint as real hereditary polynomials, and denote the set of such polynomials by $H$. While these polynomials do not form an algebra, they are clearly a vector space. Those which are finite sums of squares form a cone $C$ in $H$ (that is, $C$ is closed under sums and positive scalar multiplication). Any real polynomial is the sum of terms of the form $1 \otimes A$ or $w_2^* w_1 \otimes B + w_1^* w_2 \otimes B^*$, where $w_1, w_2 \in S$ and $A$ is selfadjoint. The first of these is obviously the difference of squares. Using $w^*w = 1$ for any $w \in G$, we also have

$$w_2^* w_1 \otimes B + w_1^* w_2 \otimes B^* = (w_1 \otimes B + w_2 \otimes 1)^*(w_1 \otimes B + w_2 \otimes 1) - 1 \otimes (1 + B^*B).$$

Hence $H = C - C$.

For $A, B \in \mathfrak{L}(\mathfrak{H})$ and $w_1, w_2 \in S$,

$$\begin{aligned} 0 &\leq (w_1 \otimes A + w_2 \otimes B)^*(w_1 \otimes A + w_2 \otimes B) \\ &\leq (w_1 \otimes A + w_2 \otimes B)^*(w_1 \otimes A + w_2 \otimes B) \\ &\qquad + (w_1 \otimes A - w_2 \otimes B)^*(w_1 \otimes A - w_2 \otimes B) \\ &= 2(1 \otimes A^*A + 1 \otimes B^*B) \\ &\leq (\|A\|^2 + \|B\|^2)(1 \otimes 1). \end{aligned}$$

Applying this iteratively, we see that for any $P \in H$, there is some constant $0 \leq \alpha < \infty$ such that $\alpha 1 \pm P \in C$. In other words, the cone $C$ is **archimedean**. In particular, $1 \otimes 1$ is in the algebraic interior of $C$, meaning that if $P \in H$, then there is some $0 < t_0 \leq 1$ such that for all $0 < t < t_0$, $t(1 \otimes 1) + (1 - t)P \in C$.

**Theorem 6.2.** *Let $G$ be a finitely generated discrete group, $P$ a strictly positive trigonometric polynomial over $G$ with coefficients in $\mathfrak{L}(\mathfrak{G})$. Then $P$ is a sum of squares of analytic polynomials.*

*Proof.* The proof uses a standard GNS construction and separation argument. Suppose that for some $\epsilon > 0$, $P - \epsilon(1 \otimes 1) \geq 0$ but that $P \notin C$. Since $C$ has nonempty algebraic interior, it follows from the Edelheit-Kakutani theorem[1] that there is a nonconstant linear functional $\lambda : H \to \mathbb{R}$ such that $\lambda(C) \geq 0$ and $\lambda(P) \leq 0$. Since $\lambda$ is nonzero, there is some $R \in H$ with $\lambda(R) > 0$, and so since the cone $C$ is archimedean, there exists $\alpha > 0$ such that $\alpha(1 \otimes 1) - R \in C$. From this we see that $\lambda(1 \otimes 1) > 0$, and so by scaling, we may assume $\lambda(1 \otimes 1) = 1$.

We next define a nontrivial scalar product on $H$ by setting

$$\langle w_1 \otimes A, w_2 \otimes B \rangle = \lambda(w_2^* w_1 \otimes B^* A)$$

and extending linearly to all of $H$. It is easily checked that this satisfies all of the properties of an inner product, except that $\langle w \otimes A, w \otimes A \rangle = 0$ may not necessarily imply that $w \otimes A = 0$. Even so, such scalar products satisfy the Cauchy-Schwarz inequality, and so $N = \{w \otimes A : \langle w \otimes A, w \otimes A \rangle = 0\}$ is a vector subspace of $H$. Therefore this scalar product induces an inner product on $H/N$, and the completion $\mathfrak{H}$ of $H/N$ with respect to the associated norm makes $H/N$ into a Hilbert space.

We next define a representation $\pi : H \to \mathfrak{L}(\mathfrak{H})$ by the left regular representation; that is, $\pi(P)[w \otimes A] = [P(w \otimes A)]$, where $[\,\cdot\,]$ indicates an equivalence class in $H/N$. Since $P \geq \epsilon(1 \otimes 1) \geq 0$ for some $\epsilon > 0$, $P - \epsilon/2(1 \otimes 1) > 0$. Suppose that $P \notin C$. Then

$$\lambda((P - \epsilon/2(1 \otimes 1)) + \epsilon/2(1 \otimes 1)) = \lambda(P - \epsilon/2(1 \otimes 1)) + \epsilon/2 \leq 0.$$

Hence

$$\langle \pi(P - \epsilon/2(1 \otimes 1))[1 \otimes 1], [1 \otimes 1] \rangle \leq -\epsilon/2,$$

[1] [36] Holmes, Corollary, §4B. *Let $A$ and $B$ be nonempty convex subsets of $X$, and assume the algebraic interior of $A$, $\operatorname{cor}(A)$ is nonempty. Then $A$ and $B$ can be separated if and only if $\operatorname{cor}(A) \cap B = \emptyset$.*

and so $\pi(P - \epsilon/2(1 \otimes 1)) \not\geq 0$. The representation $\pi$ obviously induces a unitary representation of $G$ via $\pi(a_i) = \pi(a_i \otimes 1)$, where $a_i$ is a generator of $G$. The (irreducible) representations of $G$ are in bijective correspondence with the essential unital $*$-representations of the group $C^*$-algebra $C^*(G)$ (Palmer, [46, Theorem 12.4.1]), which then restrict back to representations of $H$. Since unitary representations of $G$ are direct integrals of irreducible unitary representations (see, for example, Palmer, [46, p. 1386]), there is an irreducible unitary representation $\pi'$ of $G$ such that the corresponding representation of $H$ has the property that $\pi'(P - \epsilon/2(1 \otimes 1)) \not\geq 0$, giving a contradiction. □

The above could equally well have been derived using any $C^*$-algebra in the place of $\mathfrak{L}(\mathfrak{H})$. One could also further generalize to non-discrete locally compact groups, replacing the trigonometric polynomials by functions of compact support.

We obtain Theorem 5.1 as a corollary if we take $G$ to be the free group in $d$ commuting letters. On the other hand, if $G$ is the noncommutative free group on $d$ letters, it is again straightforward to specify the irreducible representations of $G$. These take the generators $(a_1, \dots, a_d)$ to irreducible $d$-tuples $(U_1, \dots, U_d)$ of (noncommuting) unitary operators, yielding a weak form of McCullough's theorem.

As mentioned earlier, it is known by results of Scheiderer [57] that when $G$ is the free group in $d$ *commuting* letters, $d \geq 3$, there are positive polynomials which cannot be expressed as sums of squares of analytic polynomials, so no statement along the lines of Theorem 6.1 can be true for trigonometric polynomials if it is to hold for all finitely generated discrete groups. Just what can be said in various special cases is still largely unexplored.

## Appendix: Schur complements

We prove the existence and uniqueness of Schur complements for Hilbert space operators as required in Definition 3.1.

**Lemma A.1.** *Let $T \in \mathfrak{L}(\mathfrak{H})$, where $\mathfrak{H}$ is a Hilbert space. Let $\mathfrak{K}$ be a closed subspace of $\mathfrak{H}$, and write*

$$T = \begin{pmatrix} A & B^* \\ B & C \end{pmatrix} : \mathfrak{K} \oplus \mathfrak{K}^\perp \to \mathfrak{K} \oplus \mathfrak{K}^\perp.$$

*Then $T \geq 0$ if and only if $A \geq 0$, $C \geq 0$, and $B = C^{\frac{1}{2}} G A^{\frac{1}{2}}$ for some contraction $G \in \mathfrak{L}(\mathfrak{K}, \mathfrak{K}^\perp)$. The operator $G$ can be chosen so that it maps $\overline{\operatorname{ran}} A$ into $\overline{\operatorname{ran}} C$ and is zero on the orthogonal complement of $\overline{\operatorname{ran}} A$, and then it is unique.*

*Proof.* If $B = C^{\frac{1}{2}} G A^{\frac{1}{2}}$ where $G \in \mathfrak{L}(\mathfrak{K}, \mathfrak{K}^\perp)$ is a contraction, then

$$T = \begin{pmatrix} A^{\frac{1}{2}} & 0 \\ C^{\frac{1}{2}} G & C^{\frac{1}{2}} (I - GG^*)^{\frac{1}{2}} \end{pmatrix} \begin{pmatrix} A^{\frac{1}{2}} & G^* C^{\frac{1}{2}} \\ 0 & (I - GG^*)^{\frac{1}{2}} C^{\frac{1}{2}} \end{pmatrix} \geq 0.$$

Conversely, if $T \geq 0$, it is trivial that $A \geq 0$ and $C \geq 0$. Set

$$N = T^{\frac{1}{2}} = \begin{pmatrix} N_1 \\ N_2 \end{pmatrix} : \mathfrak{H} \to \mathfrak{H} \oplus \mathfrak{K}.$$

Then $A = N_1N_1^*$ and $C = N_2N_2^*$, and so there exist partial isometries $V_1 \in \mathfrak{L}(\mathfrak{K}, \mathfrak{H})$ and $V_2 \in \mathfrak{L}(\mathfrak{K}^\perp, \mathfrak{H})$ with initial spaces $\overline{\operatorname{ran}} A$ and $\overline{\operatorname{ran}} C$ such that $N_1^* = V_1 A^{\frac{1}{2}}$ and $N_2^* = V_2 C^{\frac{1}{2}}$. Thus $B = N_2N_1^* = C^{\frac{1}{2}} G A^{\frac{1}{2}}$, where $G = V_2^* V_1$ is a contraction. By construction $G$ has the properties in the last statement, and clearly such an operator is unique. □

**Lemma** A.2. *Let $\mathfrak{H}$ be a Hilbert space, and suppose $T \in \mathfrak{L}(\mathfrak{H})$, $T \geq 0$. Let $\mathfrak{K}$ be a closed subspace of $\mathfrak{H}$, and write*

$$T = \begin{pmatrix} A & B^* \\ B & C \end{pmatrix} : \mathfrak{K} \oplus \mathfrak{K}^\perp \to \mathfrak{K} \oplus \mathfrak{K}^\perp.$$

*Then there is a largest operator $S \geq 0$ in $\mathfrak{L}(\mathfrak{K})$ such that*

$$\begin{pmatrix} A - S & B^* \\ B & C \end{pmatrix} \geq 0. \tag{A.1}$$

*It is given by $S = A^{\frac{1}{2}}(I - G^*G)A^{\frac{1}{2}}$, where $G \in \mathfrak{L}(\mathfrak{K}, \mathfrak{K}^\perp)$ is a contraction which maps $\overline{\operatorname{ran}} A$ into $\overline{\operatorname{ran}} C$ and is zero on the orthogonal complement of $\overline{\operatorname{ran}} A$.*

*Proof.* By Lemma A.1, we may define $S = A^{\frac{1}{2}}(I - G^*G)A^{\frac{1}{2}}$ with $G$ as in the last statement of the lemma. Then

$$\begin{pmatrix} A - S & B^* \\ B & C \end{pmatrix} = \begin{pmatrix} A^{\frac{1}{2}} G^* G A^{\frac{1}{2}} & A^{\frac{1}{2}} G^* C^{\frac{1}{2}} \\ C^{\frac{1}{2}} G A^{\frac{1}{2}} & C \end{pmatrix} = \begin{pmatrix} A^{\frac{1}{2}} G^* \\ C^{\frac{1}{2}} \end{pmatrix} \begin{pmatrix} G A^{\frac{1}{2}} & C^{\frac{1}{2}} \end{pmatrix} \geq 0.$$

Consider any $X \geq 0$ in $\mathfrak{L}(\mathfrak{K})$ such that

$$\begin{pmatrix} A - X & B^* \\ B & C \end{pmatrix} \geq 0.$$

Since $A \geq X \geq 0$, we can write $X = A^{\frac{1}{2}} K A^{\frac{1}{2}}$ where $K \in \mathfrak{L}(\mathfrak{K})$ and $0 \leq K \leq I$. We can choose $K$ so that it maps $\overline{\operatorname{ran}} A$ into itself and is zero on $(\overline{\operatorname{ran}} A)^\perp$. Then

$$\begin{aligned} \begin{pmatrix} A - X & B^* \\ B & C \end{pmatrix} &= \begin{pmatrix} A - A^{\frac{1}{2}} K A^{\frac{1}{2}} & A^{\frac{1}{2}} G^* C^{\frac{1}{2}} \\ C^{\frac{1}{2}} G A^{\frac{1}{2}} & C \end{pmatrix} \\ &= \begin{pmatrix} A^{\frac{1}{2}} & 0 \\ 0 & C^{\frac{1}{2}} \end{pmatrix} \begin{pmatrix} I - K & G^* \\ G & I \end{pmatrix} \begin{pmatrix} A^{\frac{1}{2}} & 0 \\ 0 & C^{\frac{1}{2}} \end{pmatrix}. \end{aligned}$$

By our choices $G$ and $K$, we deduce that

$$\begin{pmatrix} I - K & G^* \\ G & I \end{pmatrix} \geq 0.$$

By Lemma A.1, $G = G_1(I - K)^{\frac{1}{2}}$ where $G_1 \in \mathfrak{L}(\mathfrak{K}, \mathfrak{K}^\perp)$ is a contraction. Therefore $G^*G \leq I - K$, and so

$$X = A^{\frac{1}{2}} K A^{\frac{1}{2}} \leq A^{\frac{1}{2}}(I - G^*G)A^{\frac{1}{2}} = S.$$

This shows $S$ is maximal with respect to the property (A.1). □

## References

[1] D.Z. Arov, *Stable dissipative linear stationary dynamical scattering systems*, J. Operator Theory 2 (1979), no. 1, 95–126, English Transl. with appendices by the author and J. Rovnyak, Oper. Theory Adv. Appl., vol. 134, Birkhäuser Verlag, Basel, 2002, 99–136.

[2] M. Bakonyi and T. Constantinescu, *Schur's algorithm and several applications*, Pitman Research Notes in Mathematics Series, vol. 261, Longman Scientific & Technical, Harlow, 1992.

[3] M. Bakonyi and H.J. Woerdeman, *Matrix completions, moments, and sums of Hermitian squares*, book manuscript, in preparation, 2008.

[4] S. Barclay, *A solution to the Douglas-Rudin problem for matrix-valued functions*, Proc. London Math. Soc. (3), 99 (2009), no. 3, 757–786.

[5] ———, *Continuity of the spectral factorization mapping*, J. London Math. Soc. (2) 70 (2004), no. 3, 763–779.

[6] ———, *Banach spaces of analytic vector-valued functions*, Ph.D. thesis, University of Leeds, 2007.

[7] H. Bart, I. Gohberg, M.A. Kaashoek, and A.C.M. Ran, *Factorization of matrix and operator functions: the state space method*, Oper. Theory Adv. Appl., vol. 178, Birkhäuser Verlag, Basel, 2008.

[8] G. Blower, *On analytic factorization of positive Hermitian matrix functions over the bidisc*, Linear Algebra Appl. 295 (1999), no. 1-3, 149–158.

[9] J. Bourgain, *A problem of Douglas and Rudin on factorization*, Pacific J. Math. 121 (1986), no. 1, 47–50.

[10] L. de Branges, *The expansion theorem for Hilbert spaces of entire functions*, Entire Functions and Related Parts of Analysis (Proc. Sympos. Pure Math., La Jolla, Calif., 1966), Amer. Math. Soc., Providence, RI, 1968, pp. 79–148.

[11] A. Brown and P.R. Halmos, *Algebraic properties of Toeplitz operators*, J. Reine Angew. Math. 213 (1963/1964), 89–102.

[12] G. Cassier, *Problème des moments sur un compact de $\mathbf{R}^n$ et décomposition de polynômes à plusieurs variables*, J. Funct. Anal. 58 (1984), no. 3, 254–266.

[13] K.F. Clancey and I. Gohberg, *Factorization of matrix functions and singular integral operators*, Oper. Theory Adv. Appl., vol. 3, Birkhäuser Verlag, Basel, 1981.

[14] T. Constantinescu, *Factorization of positive-definite kernels*, Topics in operator theory: Ernst D. Hellinger memorial volume, Oper. Theory Adv. Appl., vol. 48, Birkhäuser Verlag, Basel, 1990, pp. 245–260.

[15] A. Devinatz, *The factorization of operator-valued functions*, Ann. of Math. (2) 73 (1961), 458–495.

[16] R.G. Douglas, *On factoring positive operator functions*, J. Math. Mech. 16 (1966), 119–126.

[17] R.G. Douglas and W. Rudin, *Approximation by inner functions*, Pacific J. Math. 31 (1969), 313–320.

[18] M.A. Dritschel, *On factorization of trigonometric polynomials*, Integral Equations Operator Theory 49 (2004), no. 1, 11–42.

[19] M.A. Dritschel and H.J. Woerdeman, *Outer factorizations in one and several variables*, Trans. Amer. Math. Soc. 357 (2005), no. 11, 4661–4679.

[20] P.L. Duren, *Theory of $H^p$ spaces*, Academic Press, New York, 1970; Dover reprint, Mineola, New York, 2000.

[21] R.E. Edwards, *Functional analysis. Theory and applications*, Holt, Rinehart and Winston, New York, 1965; Dover reprint, Mineola, New York, 1995.

[22] L. Fejér, *Über trigonometrische Polynome*, J. Reine Angew. Math. 146 (1916), 53–82.

[23] J.S. Geronimo and Ming-Jun Lai, *Factorization of multivariate positive Laurent polynomials*, J. Approx. Theory 139 (2006), no. 1-2, 327–345.

[24] J.S. Geronimo and H.J. Woerdeman, *Positive extensions, Fejér-Riesz factorization and autoregressive filters in two variables*, Ann. of Math. (2) 160 (2004), no. 3, 839–906.

[25] ______, *The operator-valued autoregressive filter problem and the suboptimal Nehari problem in two variables*, Integral Equations Operator Theory 53 (2005), no. 3, 343–361.

[26] I. Gohberg, *The factorization problem for operator functions*, Izv. Akad. Nauk SSSR Ser. Mat. 28 (1964), 1055–1082, Amer. Math. Soc. Transl. (2) **49** 130–161.

[27] I. Gohberg, S. Goldberg, and M.A. Kaashoek, *Classes of linear operators. Vol. I*, Oper. Theory Adv. Appl., vol. 49, Birkhäuser Verlag, Basel, 1990.

[28] U. Grenander and G. Szegő, *Toeplitz forms and their applications*, California Monographs in Mathematical Sciences, University of California Press, Berkeley, 1958.

[29] P.R. Halmos, *Shifts on Hilbert spaces*, J. Reine Angew. Math. 208 (1961), 102–112.

[30] H. Helson, *Lectures on invariant subspaces*, Academic Press, New York, 1964.

[31] H. Helson and D. Lowdenslager, *Prediction theory and Fourier series in several variables*, Acta Math. 99 (1958), 165–202.

[32] ______, *Prediction theory and Fourier series in several variables. II*, Acta Math. 106 (1961), 175–213.

[33] J.W. Helton, S.A. McCullough, and M. Putinar, *Matrix representations for positive noncommutative polynomials*, Positivity 10 (2006), no. 1, 145–163.

[34] J.W. Helton and M. Putinar, *Positive polynomials in scalar and matrix variables, the spectral theorem, and optimization*, Operator theory, structured matrices, and dilations, Theta Ser. Adv. Math., vol. 7, Theta, Bucharest, 2007, pp. 229–306.

[35] K. Hoffman, *Banach spaces of analytic functions*, Prentice-Hall Inc., Englewood Cliffs, N. J., 1962; Dover reprint, Mineola, New York, 1988.

[36] R.B. Holmes, *Geometric functional analysis and its applications*, Graduate Texts in Mathematics, No. 24, Springer-Verlag, New York, 1975.

[37] B. Jacob and J.R. Partington, *On the boundedness and continuity of the spectral factorization mapping*, SIAM J. Control Optim. 40 (2001), no. 1, 88–106.

[38] T. Kailath, A.H. Sayed, and B. Hassibi, *Linear estimation*, Prentice Hall, Englewood Cliffs, NJ, 1980.

[39] A. Lebow and M. Schreiber, *Polynomials over groups and a theorem of Fejér and Riesz*, Acta Sci. Math. (Szeged) 44 (1982), no. 3-4, 335–344 (1983).

[40] D. Lowdenslager, *On factoring matrix-valued functions*, Ann. of Math. (2) 78 (1963), 450–454.

[41] A.S. Markus, *Introduction to the spectral theory of polynomial operator pencils*, Translations of Mathematical Monographs, vol. 71, Amer. Math. Soc., Providence, RI, 1988.

[42] S. McCullough, *Factorization of operator-valued polynomials in several non-commuting variables*, Linear Algebra Appl. 326 (2001), no. 1-3, 193–203.

[43] G.J. Murphy, *$C^*$-algebras and operator theory*, Academic Press Inc., Boston, MA, 1990.

[44] A. Naftalevich and M. Schreiber, *Trigonometric polynomials and sums of squares*, Number theory (New York, 1983–84), Lecture Notes in Math., vol. 1135, Springer-Verlag, Berlin, 1985, pp. 225–238.

[45] T.W. Palmer, *Banach algebras and the general theory of $*$-algebras. Vol.* 1, Encyclopedia of Mathematics and its Applications, vol. 49, Cambridge University Press, Cambridge, 1994.

[46] ______, *Banach algebras and the general theory of $*$-algebras. Vol.* 2, Encyclopedia of Mathematics and its Applications, vol. 79, Cambridge University Press, Cambridge, 2001.

[47] F. Riesz, *Über ein Problem des Herrn Carathéodory*, J. Reine Angew. Math. 146 (1916), 83–87.

[48] M. Rosenblatt, *A multi-dimensional prediction problem*, Ark. Mat. 3 (1958), 407–424.

[49] M. Rosenblum, *Vectorial Toeplitz operators and the Fejér-Riesz theorem*, J. Math. Anal. Appl. 23 (1968), 139–147.

[50] M. Rosenblum and J. Rovnyak, *Hardy classes and operator theory*, Oxford University Press, New York, 1985; Dover reprint, Mineola, New York, 1997.

[51] ______, *The factorization problem for nonnegative operator-valued functions*, Bull. Amer. Math. Soc. 77 (1971), 287–318.

[52] Yu.A. Rozanov, *Stationary random processes*, Holden-Day Inc., San Francisco, Calif., 1967.

[53] W. Rudin, *The extension problem for positive-definite functions*, Illinois J. Math. 7 (1963), 532–539.

[54] L.A. Sakhnovich, *Interpolation theory and its applications*, Kluwer, Dordrecht, 1997.

[55] D. Sarason, *Generalized interpolation in $H^\infty$*, Trans. Amer. Math. Soc. 127 (1967), 179–203.

[56] A.H. Sayed and T. Kailath, *A survey of spectral factorization methods*, Numer. Linear Algebra Appl. 8 (2001), no. 6-7, 467–496, Numerical linear algebra techniques for control and signal processing.

[57] C. Scheiderer, *Sums of squares of regular functions on real algebraic varieties*, Trans. Amer. Math. Soc. 352 (2000), no. 3, 1039–1069.

[58] K. Schmüdgen, *The $K$-moment problem for compact semi-algebraic sets*, Math. Ann. 289 (1991), no. 2, 203–206.

[59] ———, *Noncommutative real algebraic geometry – some basic concepts and first ideas*, Emerging Applications of Algebraic Geometry, The IMA Volumes in Mathematics and its Applications, vol. 149, Springer-Verlag, Berlin, 2009, pp. 325–350.

[60] B. Simon, *Orthogonal polynomials on the unit circle. Part 1*, Amer. Math. Soc. Colloq. Publ., vol. 54, Amer. Math. Soc., Providence, RI, 2005.

[61] B. Sz.-Nagy and C. Foias, *Harmonic analysis of operators on Hilbert space*, North-Holland Publishing Co., Amsterdam, 1970.

[62] G. Szegő, *Orthogonal polynomials*, fourth ed., Amer. Math. Soc. Colloq. Publ., vol. 23, Amer. Math. Soc., Providence, RI, 1975.

[63] N. Wiener and P. Masani, *The prediction theory of multivariate stochastic processes. I. The regularity condition*, Acta Math. 98 (1957), 111–150.

[64] ———, *The prediction theory of multivariate stochastic processes. II. The linear predictor*, Acta Math. 99 (1958), 93–137.

[65] V. Zasuhin, *On the theory of multidimensional stationary random processes*, C. R. (Doklady) Acad. Sci. URSS (N.S.) 33 (1941), 435–437.

Michael A. Dritschel
School of Mathematics and Statistics
Herschel Building
University of Newcastle
Newcastle upon Tyne
NE1 7RU, UK
e-mail: `m.a.dritschel@ncl.ac.uk`

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

Operator Theory:
Advances and Applications, Vol. 207, 255–285

# A Halmos Doctrine and Shifts on Hilbert Space

Paul S. Muhly

*To the memory of Paul Halmos, with continuing respect, admiration and gratitude*

**Abstract.** This a survey of some recent work on noncommutative function theory related to tensor algebras that derives in part from Paul Halmos's paper, *Shifts on Hilbert space.*

**Mathematics Subject Classification (2000).** 46E22, 46E50, 46G20, 46H15, 46H25, 46K50, 46L08, 46L89.

**Keywords.** Shifts on Hilbert space, tensor algebra, matricial function theory, fully matricial set, $C^*$-correspondence, operator algebra.

## 1. Introduction

I took a reading course from Paul in the summer of 1966. He was putting the final touches on the first edition of his very influential *A Hilbert Space Problem Book* [26], and he told me we would use it as our starting point. I was honored (and frightened) that he trusted me with his draft copy (perhaps his only copy, I never knew). It was in two volumes – loosely bound, some parts were typed, but there were large inserts written by hand and taped here and there. It seemed very fragile – and personal. The book was divided into three parts: the first contained the problems, proper, along with narrative to motivate them. The second part, very short, consisted of hints. The third part gave solutions along with supplemental discussion. The first volume of the draft contained the first two parts; the second volume contained the third part. While I was thrilled that Paul lent me his copy, I was also dismayed that he only let me have the first volume. Nevertheless, the course was exhilarating.

Most details from our meetings are now long gone from my memory, but one stands out; it concerns what I like to call a Halmos Doctrine. I use the indefinite article because Paul's speech and writing were so definitive, so authoritative and so unequivocal, I am sure there are other assertions which also qualify as Halmos doctrines. Also, I suspect that there may not be unanimity about Paul's most important doctrine[1]. In any case, I don't want to leave the impression that Paul was doctrinaire. I don't think he was and this seems clear from his writings. He frequently expressed willingness to back off from points he made – points which he expressed in extreme forms in order to rivet the reader's attention or to prompt discussion. He welcomed argument and discussion. Indeed, he expected it. I learned this the hard way: In one of our meetings, I announced that a problem he had given me was trivial. He snapped: "You were supposed to make it nontrivial." I responded meekly: "I did what you told me to do." His withering retort: "Well screw you! [*sic*]" Certainly, at the time I found the reply withering, but on reflection I realized it was not intended as a parry to the insult that his problem was trivial. It wasn't personal at all. Rather, it expressed strongly his view that it is a mathematician's responsibility to explore what may appear trivial, seeking nuggets of truth that are in fact profound. I would add further that one should then polish these to the point that they seem once again trivial. Paul was very good at this. His paper, *Shifts on Hilbert space* [25], is a wonderful illustration.

The Halmos doctrine to which I am referring was presented to me something like this:

> If you want to study a problem about operators on infinite-dimensional Hilbert space, your first task is to formulate it in terms of operators on finite-dimensional spaces. Study it there before attacking the infinite.

It may seem naive and impossible to follow in any realistic manner, but if one takes this doctrine as strong advice – an admonition, even – to be fully aware of the finite-dimensional antecedents of any problem one might consider in infinite dimensions, then of course it is reasonable and fully consistent with best practices in the subject. A quintessential example of how the advice can be put into practice may be found in the introduction to Murray and von Neumann's Rings of Operators [47].

I took this Halmos doctrine seriously, and I still do. In particular, it has driven my thinking in the work I have been doing the last fifteen years or so with Baruch Solel on tensor algebras and what we call Hardy algebras. I won't try to recapitulate all of our work in this tribute to Paul, but I would like to describe how this doctrine and his paper, *Shifts on Hilbert space*, inspired parts of it.

## 2. Halmos's theorem

Recall Arne Beurling's theorem from 1949 [9], which asserts that *if $U_+$ denotes the operator of multiplication by the independent variable $z$ on the Hardy space $H^2(\mathbb{T})$, then a subspace $\mathcal{M}$ of $H^2(\mathbb{T})$ is invariant under $U_+$ if and only if there is an analytic function $\theta$ on the unit disc, whose boundary values have modulus* 1 *almost everywhere such that $\mathcal{M} = \theta H^2(\mathbb{T})$* – the set of all multiples $\theta\xi$, where $\xi$ runs over

[1] However, as far as I can tell only one has risen to any kind of "official" status. I did a Google search on "Halmos Doctrine" – with quotes included – and only one response was returned: "More is less and less is more."

$H^2(\mathbb{T})$. Beurling was inspired by problems in spectral synthesis and, in particular, by work of Norbert Wiener on translation invariant subspaces of various function spaces. His analysis rested on deep function theoretic properties of functions in the Hardy space and, in particular, on the structure of the functions $\theta$, which he called *inner functions*. Nevertheless, he noted that the geometry of Hilbert space also played a fundamental role in his analysis. In 1959, Peter Lax extended Beurling's analysis in two directions: He moved from the Hardy space of the circle to the Hardy space of the upper half-plane and he allowed his functions to take values in a finite-dimensional Hilbert space. He was motivated by problems in partial differential equations. To prove his theorem, he had to extend parts of the function theory that Beurling had used to the setting of vector-valued functions and, as a result, his analysis may seem complicated. However, Lax's generalization is easy to state: *Suppose $\mathcal{E}$ is a finite-dimensional Hilbert space and that $H^2(\mathbb{R}, \mathcal{E})$ denotes the Hardy space for the upper half-plane consisting of functions with values in $\mathcal{E}$. Then a subspace $\mathcal{M}$ of $H^2(\mathbb{R}, \mathcal{E})$ is invariant under multiplication by the functions $\chi_\lambda$, $\lambda \geq 0$, where $\chi_\lambda(t) = e^{i\lambda t}$, if and only if $\mathcal{M}$ is of the form $\Theta H^2(\mathbb{R}, \mathcal{F})$, where $\mathcal{F}$ is another Hilbert space, and where $\Theta$ is an operator-valued function, such that $\Theta(z)$ maps $\mathcal{F}$ into $\mathcal{E}$, that is bounded and analytic in the upper half-plane and which has boundary values almost everywhere* (*on* $\mathbb{R}$) *that are isometries mapping $\mathcal{F}$ into $\mathcal{E}$.* Two years later, Halmos published his *Shifts on Hilbert space* [25]. He returned to the single operator setting, but instead of studying multiplication by $z$ on the vector-valued Hardy space $H^2(\mathbb{T}, \mathcal{E})$, he Fourier transformed the picture for most of the discussion and focused on shifts on the sequence spaces $\ell^2(\mathbb{Z}_+, \mathcal{E})$. This change of perspective allowed his Hilbert spaces of "coefficients" to be of arbitrary dimension. Moreover, and more important, he distilled the Hilbert space aspects of Beurling's and Lax's arguments and effectively removed the function theory from their theorems. More accurately, he made clear the role operator theory played in their theorems and how it interacts with the function theory.

Halmos was not alone in trying to identify the precise role of Hilbert space techniques in the theory. Many others were actively pursuing the subject. Indeed, vectorial function theory began with an explosion in the late 50's and continues today as an important component of operator theory. I cannot rehearse here all the contributions by all the players. However, Halmos's paper had a terrific impact on that theory and still, today, is frequently cited. Indeed, on MathSciNet, which records citations since about 1997, as of February 9, 2009, it is the third most cited research paper that Halmos wrote, receiving 37 citations. That's a pretty good record in view of his estimate that the expected time to obsolescence for a mathematics research paper is five years [27].

Although nowadays Halmos's theorem and proof are well known to many, I want to begin by showing how short and simple the analysis is. It will be useful to have his arguments available for reference later.

Fix an auxiliary Hilbert space $\mathcal{E}$ and form the Hilbert space, $\ell^2(\mathbb{Z}_+, \mathcal{E})$, consisting of all norm-squared summable, $\mathcal{E}$-valued functions defined on the nonnegative integers, $\mathbb{Z}_+$. The *unilateral shift* (*of multiplicity equal to the dimension* *of* $\mathcal{E}$) is the operator $U_+$ defined on $\ell^2(\mathbb{Z}_+, \mathcal{E})$ by the formula

$$U_+\xi(n) = \xi(n-1),$$

$\xi \in \ell^2(\mathbb{Z}_+, \mathcal{E})$. It is easy to see that two shifts determined by two different coefficient Hilbert spaces are unitarily equivalent if and only if their multiplicities are the same. Halmos's extension of the theorems of Beurling and Lax is

**Theorem 2.1.** *Let $\mathcal{M}$ be a subspace of $\ell^2(\mathbb{Z}_+, \mathcal{E})$. Then $\mathcal{M}$ is invariant under $U_+$ if and only if there is a partial isometry $\Theta$ on $\ell^2(\mathbb{Z}_+, \mathcal{E})$ that commutes with $U_+$ such that $\mathcal{M}$ is the range of $\Theta$: $\mathcal{M} = \Theta\ell^2(\mathbb{Z}_+, \mathcal{E})$.*

Halmos called a partial isometry that commutes with $U_+$ a rigid Taylor function. Nowadays, such an operator is called an *inner operator*, or something similar, to recognize its connections with the inner functions of Beurling's theorem. The $\Theta$ is essentially uniquely determined by $\mathcal{M}$ in the sense that if $\Theta_1$ is another inner operator with the same range as $\Theta$, then there is a partial isometry $C_0$ on $\mathcal{E}$ such that $\Theta_1 = \Theta C$ where $C$ is defined by the formula $C\xi(n) = C_0(\xi(n))$ for all $\xi \in \ell^2(\mathbb{Z}_+, \mathcal{E})$ and all $n \in \mathbb{Z}_+$. Such a partial isometry is called a *constant* inner operator because in the Fourier transformed picture, it becomes a constant operator-valued function.

Halmos's proof of Theorem 2.1 was based upon the idea captured in the following definition.

**Definition 2.2.** *A subspace $\mathcal{F}$ of $\ell^2(\mathbb{Z}_+, \mathcal{E})$ is called a* wandering subspace *for $U_+$ provided $\mathcal{F}$ and $U_+\mathcal{F}$ are orthogonal.*

As an example, observe that the collection of all functions in $\ell^2(\mathbb{Z}_+, \mathcal{E})$ that are supported at one point of $\mathbb{Z}_+$, say 0, forms a wandering subspace, $\mathcal{E}_0$, which of course has the same dimension as $\mathcal{E}$. More generally, if $\mathcal{M}$ is a subspace of $\ell^2(\mathbb{Z}_+, \mathcal{E})$ that is invariant under $U_+$, then it is immediate that the subspace $\mathcal{F} := \mathcal{M} \ominus U_+\mathcal{M}$ is a wandering subspace since $U_+\mathcal{F}$ is contained in $U_+\mathcal{M}$ which is orthogonal to $\mathcal{F}$ by definition. Owing to the isometric nature of $U_+$, if $\mathcal{F}$ is a wandering subspace for $U_+$, then for all pairs of distinct non-negative integers $i$ and $j$, $U_+^i\mathcal{F}$ and $U_+^j\mathcal{F}$ are orthogonal. This means that the closed linear span $\bigvee_{n\geq 0}\{U_+^n\mathcal{F}\}$, which manifestly is invariant for $U_+$, is the orthogonal direct sum $\sum_{n\geq 0}^{\oplus} U_+^n\mathcal{F}$ of the subspaces $U_+^n\mathcal{F}$. Consequently, if $\tilde{\mathcal{F}}$ is a Hilbert space of the same dimension as $\mathcal{F}$ and if $\tilde{U}_+$ denotes the unilateral shift on $\ell^2(\mathbb{Z}_+, \tilde{\mathcal{F}})$, then the map $W$ from $\ell^2(\mathbb{Z}_+, \tilde{\mathcal{F}})$ to the orthogonal direct sum $\sum_{n\geq 0} U_+^n\mathcal{F}$, defined by the formula $W\xi := \sum_{n\geq 0} U_+^n(\xi(n))$, is a Hilbert space isomorphism from $\ell^2(\mathbb{Z}_+, \tilde{\mathcal{F}})$ onto$\sum_{n\geq 0} U_+^n\mathcal{F}$ that intertwines $\tilde{U}_+$ and $U_+$, i.e., $W\tilde{U}_+ = U_+W$. So, in order to prove Theorem 2.1, the objective becomes to prove the following two lemmas:

**Lemma 2.3.** *If $\mathcal{M}$ is an invariant subspace of $\ell^2(\mathbb{Z}_+, \mathcal{E})$ for $U_+$ and if $\mathcal{F} := M \ominus U_+\mathcal{M}$ is the wandering subspace determined by $\mathcal{M}$, then $\mathcal{M} = \bigvee_{n\geq 0}\{U_+^n\mathcal{F}\}$.*

**Lemma 2.4.** *If $\mathcal{F}$ is a wandering subspace of $\ell^2(\mathbb{Z}_+, \mathcal{E})$, then the Hilbert space dimension of $\mathcal{F}$ is at most the Hilbert space dimension of $\mathcal{E}$.*

Indeed, with the two lemmas in hand, the proof of Theorem 2.1 is almost immediate: Given an invariant subspace $\mathcal{M}$ of $\ell^2(\mathbb{Z}_+, \mathcal{E})$, Lemma 2.3 implies that $\mathcal{M} = \sum_{n\geq 0}^{\oplus} U_+^n \mathcal{F}$. Then, by Lemma 2.4 we may map $\mathcal{F}$ isometrically onto a subspace $\tilde{\mathcal{F}}$ of $\mathcal{E}$, say by an isometry $V_0$. The operator $\Theta$ on $\ell^2(\mathbb{Z}_+, \mathcal{E})$ defined by the formula

$$\Theta\xi := \sum_{n\geq 0} U_+^n V_0^*(\xi(n)), \tag{1}$$

$\xi \in \ell^2(\mathbb{Z}_+, \mathcal{E})$, then, is a partial isometry on $\ell^2(\mathbb{Z}_+, \mathcal{E})$ that commutes with $U_+$ and has range $\mathcal{M}$.

Lemma 2.3 is essentially Lemma 1 in [25]. The proof is a small variation of a proof of the Wold decomposition of an isometry: If $\xi \in \mathcal{M} \ominus \sum_{n\geq 0}^{\oplus} U_+^n \mathcal{F}$, then in particular $\xi$ lies in $\mathcal{M}$ and is orthogonal to $\mathcal{F}$, which by definition is $\mathcal{M} \ominus \mathcal{U}_+\mathcal{M}$. Thus, $\xi$ lies in $U_+\mathcal{M}$. Since $\xi$ is also orthogonal $U_+\mathcal{F} = U_+\mathcal{M} \ominus U_+^2\mathcal{M}$, $\xi$ also belongs to $U_+^2\mathcal{M}$. Continuing in this way, it follows that $\xi$ lies in $\bigcap_{n\geq 0} U_+^n \mathcal{M}$. But this space is $\{0\}$ since it is contained in $\bigcap_{n\geq 0} U_+^n \ell^2(\mathbb{Z}_+, \mathcal{E})$, which manifestly is the zero space.

Lemma 2.3 is Lemma 4 in [25]. The proof Halmos gives rests on the functional representation of elements in $\ell^2(\mathbb{Z}_+, \mathcal{E})$, i.e., the proof involves analyzing elements in $\ell^2(\mathbb{Z}_+, \mathcal{E})$ in terms of their Fourier transforms, which lie in the vector-valued Hardy space $H^2(\mathbb{T}, \mathcal{E})$. Israel Halperin seems to be the one who provided a proof that was on the same level as the rest of Halmos's arguments and was completely free of function theory (see the addendum to the proof of [66, Page 108].) It runs like this: Choose an orthonormal basis $\{u_n\}_{n\in I}$ for $\mathcal{F}$ and an orthonormal basis $\{v_m\}_{m\in J}$ for $\mathcal{E}_0$, which, recall, has the same dimension as $\mathcal{E}$. Then $\{U_+^k v_m\}_{k\geq 0, m\in J}$ is an orthonormal basis for $\ell^2(\mathbb{Z}_+, \mathcal{E})$. Consequently, the dimension of $\mathcal{F}$, $\dim \mathcal{F}$, is the sum

$$\sum_{n\in I} \|u_n\|^2 = \sum_{n\in I, m\in J, k\geq 0} |(u_n, U_+^k v_m)|^2 = \sum_{n\in I, m\in J, k\geq 0} |(U_+^{*k} u_n, v_m)|^2.$$

On the other hand, since $\mathcal{F}$ is a wandering subspace, the vectors $\{U_+^{*k} u_n\}_{k\geq 0, n\in I}$ form an orthogonal subset of $\ell^2(\mathbb{Z}_+, \mathcal{E})$ consisting of vectors of norm at most one. Consequently, the last sum is dominated by $\sum_{m\in J} \|v_m\|^2 = \dim \mathcal{E}$.

Of course now with Halmos's theorem, Theorem 2.1, before us, numerous questions arise. Two of the most pressing are: How to describe the commutant? How does the function theory from Beurling, Lax and others interact with that description? The answers are well known. The key is to Fourier transform $\ell^2(\mathbb{Z}_+, \mathcal{E})$ to $H^2(\mathbb{T}, \mathcal{E})$. When this is done, $U_+$ becomes multiplication by the independent variable $z$ and the commutant of $U_+$ is represented by the space of all (boundary values of) $B(\mathcal{E})$-valued functions that are bounded and analytic on the open unit disc $\mathbb{D}$. There is an enormous literature on this subject, but while it provides a lot of inspiration, very little of it extends to the setting I will discuss below without a lot of modification. Consequently, I will not venture into a discussion of it now. However, I do want to emphasize that my involvement with the function theory that I will describe in the final section is an outgrowth of efforts to generalize Halmos's theorem. Thus, from Hilbert space and Halmos's theorem, I found my way back to function theory.

## 3. $C^*$-correspondences, tensor algebras and $C^*$-envelopes

Much of my time has been spent pursuing Halmos's doctrine in the context of the question: How can the theory of finite-dimensional algebras inform the theory of not-necessarily-self-adjoint operator algebras? By an *operator algebra* I mean simply a norm closed subalgebra of $B(H)$, for some Hilbert space $H$. The work of Murray and von Neumann [47] made it clear they were inspired by the theory of semisimple algebras and the related theory of group representations. Indeed, their notation for a ring of operators (now called a von Neumann algebra) is $M$, suggesting that it should be viewed as a generalization of a matrix algebra, especially if the algebra is a factor. I have long believed that a non-self-adjoint operator algebra should be viewed as an analogue of a non-semisimple finite-dimensional algebra. This point of view may seem naive and that not much can be gained from it. Indeed, if one opens almost any book on ring theory and tries to formulate the basic elements of it in the context of operator algebras, one very quickly runs into serious difficulties. Very little of the finite-dimensional theory seems to transfer to the infinite-dimensional setting – especially the classification theory and representation theory of finite-dimensional algebras. More frustrating – for me, at least – has been the difficulty of transferring large collections of meaty examples – sufficiently large to capture the thrust of much of the representation theory of rings that was and continues to be so prominent in ring theory.

Of course I am not alone in the view that non-self-adjoint operator algebras should be viewed as infinite-dimensional analogues of non-semisimple finite-dimensional algebras and I do not mean to slight in any way the work of the many who have pursued the analogy. There is a very large literature on the subject. Operator theory has numerous antecedents that provide support for this analogy, but probably the first paper to take a truly algebraic perspective was the pioneering work of Richard Kadison and Isadore Singer [31]. My own perspective has been shaped to a very large extent by ideas I first learned – and continued to learn – from Ron Douglas. He has long promoted a function-theoretic/algebrogeometric point of view. A good source for this is his monograph written with Vern Paulsen [19].

To understand a bit more about what I was looking for, observe that one of the striking features of Murray and von Neumann's work is that they took great pains to separate intrinsic properties of von Neumann algebras from the properties that are artifacts of the way in which the algebras are represented on Hilbert space. They did not express themselves quite in this way, but it was a theme that ran throughout their work. This separation was made more prominent once the theory of $C^*$-algebras got under way. A $C^*$-algebra is defined as a Banach algebra with certain properties which allow it to be represented faithfully as a norm-closed self-adjoint algebra of operators on Hilbert space. One is then more freely able to study intrinsic properties of $C^*$-algebras separate from their representation theory. Of course, Sakai's theorem [62], which identifies a von Neumann algebra abstractly as a $C^*$-algebra which is a dual space, completed the separation of von Neumann algebras from the Hilbert spaces on which they might be found. Indeed, today the term *$W^*$-algebra* is used to refer to a $C^*$-algebra that is a dual space.

The first person to take up this issue for non-self-adjoint operator algebras was Bill Arveson in his seminal paper, *Subalgebras of $C^*$-algebras* [3]. Here, he proposed to show that each non-self-adjoint operator algebra $A$ can be encased in an essentially unique $C^*$-algebra, $C^*(A)$, called the *$C^*$-envelope* of $A$. He took a lot of inspiration from the theory of function algebras. Suppose that a unital operator algebra is represented unitally and completely[2] isometrically in a $C^*$-algebra $B$ in such a way that the image generates $B$. A (2-sided) ideal $J$ in $B$ is called a *boundary ideal* in $B$ for $A$ if the restriction of the quotient map $\pi : B \to B/J$ is completely isometric when restricted to the image of $A$ in $B$. A maximal boundary ideal in $B$ for $A$ is necessarily unique and is called the *Shilov boundary ideal* in $B$ for $A$. The quotient of $B$ by the Shilov boundary ideal is then unique up to $C^*$-isomorphism and is the $C^*$-envelope of $A$, $C^*(A)$. Thus $C^*(A)$ is a quotient of any $C^*$-algebra that is generated by a completely isometrically isomorphic copy of $A$. Unfortunately, Arveson was not able to prove the existence of the Shilov boundary ideal in general. That had to wait ten years for the work of Masamichi Hamana [28]. However, it was clear already in [3] that this was an important concept and showed promise for enabling one to study operator algebras independent of any place they might be represented. As a concrete example, consider the disc algebra $A(\mathbb{D})$, which consists of the continuous functions on the closed unit disc that are analytic on the interior. This algebra may be profitably studied as a subalgebra of at least three different $C^*$-algebras: The continuous functions on the closed disc $\overline{\mathbb{D}}$, the continuous functions on $\mathbb{T}$, the boundary of $\overline{\mathbb{D}}$, and the $C^*$-algebra generated by all Toeplitz operators with continuous symbols. Observe that $A(\mathbb{D})$ generates each of these $C^*$-algebras and that the continuous functions on $\mathbb{T}$, $C(\mathbb{T})$, is a quotient of each of the other two; it is the $C^*$-envelope of $A(\mathbb{D})$.

Arveson was also motivated by the dilation theory of Bela Sz.-Nagy [65] and by Forest Stinespring's analysis of positive maps on $C^*$-algebras [64]. He generalized both of these papers by showing that if $\rho$ is a completely contractive representation of $A$ on a Hilbert space $H$, then there is a $C^*$-representation of $C^*(A)$, $\pi$, on a Hilbert space $K$ and an isometry $V$ embedding $H$ into $K$ such that

$$\rho(a) = V^*\pi(a)V \tag{2}$$

for all $a \in A$.

[2] I will use the terminology from operator spaces freely. See [10, 20, 50, 53] for this very important theory. However, for those who are not familiar with the subject, it is safe, for the purposes of this survey, to omit the adverb "completely" from the discussion and think solely in terms of the adjective that "completely" modifies.

The triple $(K, \pi, V)$ is called *a $C^*$-dilation* of $\rho$. (It is not uniquely determined by $\rho$ in general.) To prove this dilation theorem, Arveson made an extensive study of completely positive maps, proving a Hahn-Banach-Krein type theorem for them that subsequently contributed enormously to the birth of the theory of operator spaces. Of special note for this discussion is the paper of David Blecher, Zhong-Jin Ruan and Allan Sinclair [11], in which the authors characterized operator algebras abstractly, without reference to any Hilbert spaces on which they might act. Their work was especially inspirational to me.

Sometime in 1995 I obtained a preprint of Mihai Pimsner's paper [52] which contained the key to what I was looking for. This paper was based on the notion of what has come to be known as a $C^*$-correspondence. A *$C^*$-correspondence from a $C^*$-algebra $A$ to a* $C^*$-algebra $B$ is a right Hilbert $B$-module $E$ that is endowed with a left $A$-module structure through a $C^*$-homomorphism $\varphi$ from $A$ into the algebra of all continuous adjointable operators on $E$, $\mathcal{L}(E)$. (I follow the notation and (most of) the terminology in Chris Lance's lovely monograph [34].) Many view $C^*$-correspondences as a natural generalization of $C^*$-homomorphisms. Indeed, $C^*$-algebras form the objects of a category whose morphisms are (isomorphism classes) of $C^*$-correspondences. The composition of a $C^*$-correspondence $E$ from $A$ to $B$ with a $C^*$-correspondence $F$ from $B$ to $C$ is their balanced tensor product $E \otimes_B F$. It is the completion of the algebraic balanced tensor product of $E$ and $F$ in the norm that comes from the $C$-valued inner product defined by the formula $\langle \xi_1 \otimes \eta_1, \xi_2 \otimes \eta_2 \rangle_C := \langle \eta_1, \varphi_B(\langle \xi_1, \xi_2 \rangle_B)\eta_2 \rangle_C$. The left and right actions of $A$ and $C$ on $E \otimes_B F$ are the obvious ones. In particular, a $C^*$-correspondence from $A$ to $A$, or simply a $C^*$-correspondence over $A$, can profitably be viewed as a generalized endomorphism of $A$.

Given a $C^*$-algebra $A$ and a $C^*$-correspondence $E$ over $A$, Pimsner constructed an algebra, denoted $\mathcal{O}(E)$, that may be viewed as a generalized Cuntz algebra and has come to be called a *Cuntz-Pimsner algebra*. Here is how[3]: Form the tensor powers of $E$ (all powers are balanced over $A$) and take their direct sum, obtaining the *Fock space* over $E$, $\mathcal{F}(E) := A \oplus E \oplus E^{\otimes 2} \oplus \cdots$. It is also a $C^*$-correspondence over $A$ in an obvious way (the left action is written $\varphi_\infty$). The *creation operators* $T_\xi$, $\xi \in E$, are then defined on $\mathcal{F}(E)$ by the equation $T_\xi \eta = \xi \otimes \eta$, $\eta \in \mathcal{F}(E)$. Each $T_\xi$ is a bounded, adjointable operator on $\mathcal{F}(E)$ and the $C^*$-subalgebra of $\mathcal{L}(\mathcal{F}(E))$ they generate, $\mathcal{T}(E)$, is called the *Toeplitz-Cuntz-Pimsner algebra* of $E$. Recall that $K(E)$ is the $C^*$-subalgebra of $\mathcal{L}(E)$ generated by the operators of the form $\xi \otimes \eta^*$, which are defined by the equation $\xi \otimes \eta^*(\zeta) := \xi\langle \eta, \zeta \rangle$, and set $J := \varphi^{-1}(K(E))$ – an ideal in $A$. Then Pimsner proves that $K(\mathcal{F}(E)J)$ is an ideal in $\mathcal{T}(E)$ [52, Theorem 3.13]. The quotient $\mathcal{T}(E)/K(\mathcal{F}(E)J)$ is defined to

[3] I shall assume in this discussion and throughout the rest of this paper that $\varphi$ is injective and essential. The latter condition means that $\varphi$ extends to a unital embedding of the multiplier algebra of into $\mathcal{L}(E)$. These restrictions can be relaxed, but at the cost of digressions that need not concern us. Also, all representations of $C^*$-algebras on Hilbert space will be assumed to be essential, i.e., non-degenerate.

be the Cuntz-Pimsner algebra $\mathcal{O}(E)$. This is not Pimsner's official definition, but his Theorem 3.13 shows that the two are equivalent.

The reason I thought Pimsner's work would be useful for the purpose of connecting finite-dimensional algebra to operator algebra goes back to an old observation of Hochschild [30] that asserts that *every* finite-dimensional algebra over an algebraically closed field is a quotient of a tensor algebra. Further, coupling Hochschild's theorem with an earlier theorem of Nesbitt and Scott [48], which in current language asserts that *every* finite-dimensional algebra is Morita equivalent to one which is commutative modulo its radical, one is led to try to represent each finite-dimensional algebra in terms of a finite directed graph, called the *quiver* of the algebra. When an algebra is commutative modulo its radical, the tensor algebra *is* the path algebra built from the quiver. The representation of finite-dimensional algebras in terms of quivers has grown into an active and deep area of algebra. (For a survey, see [24] by Gabriel, who introduced the term 'quiver' in [23].) When reflecting on graph $C^*$-algebras and Cuntz-Krieger algebras that have been the focus of much $C^*$-algebraic research in recent years, and with the appearance of Pimsner's paper [52], Baruch and I were led to formulate and to study the notion of the *tensor algebra of a $C^*$-correspondence*: If $E$ is a $C^*$-correspondence over the $C^*$-algebra $A$, then we defined the tensor algebra of $E$ to be the norm-closed subalgebra, $\mathcal{T}_+(E)$, of $\mathcal{T}(E)$ generated by $\varphi_\infty(A)$ and $\{T_\xi \mid \xi \in E\}$. We believed that these algebras offer a means to interpret finite-dimensional algebra, particularly the representation theory that derives from the theory of quivers, in the setting of operator algebras. Not only have our initial speculations led to interesting developments in the theory of operator algebras *per se*, they have also led to contacts with numerous subjects that are remote from our initial focus. The starting point was our discovery that the $C^*$-envelope of the tensor algebra of a $C^*$-correspondence *is* the Cuntz-Pimsner algebra of the correspondence [40, Theorem 6.4].[4]

When $A = E = \mathbb{C}$, $\mathcal{F}(E)$ is simply $\ell^2(\mathbb{Z}_+)$, and $\mathcal{T}_+(E)$ is the disc algebra $A(\mathbb{D})$, realized as matrices of analytic Toeplitz operators. Indeed, in this setting the operator $T_1$ is just the unilateral shift. The $C^*$-envelope, then, is $C(\mathbb{T})$, as I have noted. When $A = \mathbb{C}$, but $E = \mathbb{C}^d$, then $\mathcal{F}(E)$ is the full Fock space over $\mathbb{C}^d$, which of course is also a Hilbert space, and $\mathcal{T}_+(E)$ is the noncommutative disc algebra introduced by Gelu Popescu in [56]. Our theorem shows that its $C^*$-envelope is $\mathcal{O}(\mathbb{C}^d)$, which is the well-known Cuntz algebra, usually denoted $\mathcal{O}_d$.

The noncommutative disc algebra and its weak closure in its natural representation on the Fock space, called *the free semigroup algebra*, $\mathcal{L}_d$, have been a terrific source of inspiration for Baruch and me – and for many others, of course. The literature on these algebras is growing rapidly, and so I cannot recap or even cite all of it, but I do want to note that Gelu Popescu introduced the concept in

[4] Our theorem was proved in a context that is a bit less restrictive than the one described here. More recently, using technology developed in [40] and ideas of Takeshi Katsura [32], Elias Katsoulis and David Kribs [33] removed all special hypotheses on the correspondence $E$ and so one can now say, without qualification, that the $C^*$-envelope of $\mathcal{T}_+(E)$ is $\mathcal{O}(E)$.

[56], used the notation $F^\infty$ and called $F^\infty$ the Hardy algebra. Ken Davidson and David Pitts, who also made some of the pioneering contributions, use the name "free semigroup algebra" for something more general than $\mathcal{L}_d$. However, the term "free semigroup algebra" and the notation $\mathcal{L}_d$ seem now to be most commonly used in the literature.

When the $C^*$-algebra $A$ is commutative and finite dimensional and when $E$ is also finite dimensional, then $(E, A)$ may be described in terms of a graph or quiver [40, Example 2.9]. Indeed, one may think of $A$ as the space of functions on the finite set of vertices $G^0$ of a directed graph $G := (G^0, G^1, r, s)$. $G^0$ then labels the minimal projections in $A$, $\{p_v\}_{v\in G^0}$. For $u$ and $v$ in $G^0$, the complex vector space $p_vEp_u$ is finite dimensional and so the edge set, $G^1$, is defined to have $\dim(p_vEp_u)$ edges $e$ starting at $u$ and ending at $v$. One writes $r(e) = v$ and $s(e) = u$. (The functions $r$ and $s$ are to be read "range" and "source".) The correspondence, $E$, in turn, may be identified with the space of all complex-valued functions on $G^1$, which is a $C^*$-correspondence over $A$ via the left and right actions defined by the formula $a\xi b(e) = a(r(e))\xi(e)b(s(e))$, $e \in G^1$, and inner product defined by the formula $\langle \xi, \eta\rangle(u) := \sum_{s(e)=u} \overline{\xi(e)}\eta(e)$. Finally, $\mathcal{T}_+(E)$, is the norm completion of the quiver algebra built from $G$ [39, 40, 41] and the $C^*$-envelope is the $C^*$-algebra of $G$.

If $A$ is a $C^*$-algebra and if $\varphi : A \to A$ is a $C^*$-endomorphism, then $A$ becomes a $C^*$-correspondence over $A$, denoted ${}_\varphi A$, by letting the right action be the multiplication from $A$, by letting the inner product be the obvious one, $\langle \xi, \eta\rangle := \xi^*\eta$, and by letting the left action be given by $\varphi$. When $\varphi$ is an automorphism, the tensor algebra in this case is a so-called analytic crossed product. These first arose in the work of Kadison and Singer [31] (see their Theorem 2.2.1, in particular) and then in the work of Arveson [2] in which he showed that they constitute a complete set of conjugacy invariants for $\varphi$ (under the assumption that $A$ is an abelian von Neumann algebra realized as $L^\infty(m)$ and $\varphi$ is given by an ergodic $m$-preserving transformation). They have been of considerable interest subsequently. The extension to a general endomorphism was first considered by Justin Peters in [51]. If $\varphi$ is an automorphism, then the $C^*$-envelope of $\mathcal{T}_+({}_\varphi A)$ is the crossed product $A \rtimes_\varphi \mathbb{Z}$. If $\varphi$ is an endomorphism that is not an automorphism, then one must first extend $\varphi$ to an automorphism $\varphi_\infty$ of the inductive limit $A_\infty$ constructed from the inductive system induced by $A$ and $\varphi$ in the standard well-known fashion. The $C^*$-envelope of $\mathcal{T}_+({}_\varphi A)$ in this case, then, is $A_\infty \rtimes_{\varphi_\infty} \mathbb{Z}$.

An endomorphism is a special case of a completely positive map. With these, too, one can build $C^*$-correspondences. The simplest is the so-called GNS correspondence. If $A$ is a unital $C^*$-algebra and if $P : A \to A$ is a unital completely positive map, then the *GNS correspondence determined by* $P$ is the (Hausdorff) completion, $A \otimes_P A$, of the algebraic tensor product $A \otimes A$ in the pre-inner product $\langle \xi_1 \otimes \eta_1, \xi_2 \otimes \eta_2\rangle := \eta_1^* P(\langle \xi_1, \xi_2\rangle)\eta_2$. The actions of $A$ on $A \otimes_P A$ are given by the formula $a(\xi \otimes \eta)b := (a\xi) \otimes (\eta b)$. Observe that when $P$ is an endomorphism $A \otimes_P A$ is isomorphic to ${}_P A$ via the map $\xi \otimes \eta \to P(\xi)\eta$. I don't know a simple description of $\mathcal{O}(A \otimes_P A)$ for a general completely positive map $P$. However in [36], Alberto

Marrero and I showed that when $P$ is a completely positive unital map on the $n\times n$ matrices, $M_n(\mathbb{C})$, then $\mathcal{O}(A\otimes_P A)$ is strongly Morita equivalent to $\mathcal{O}_d$ where $d$ is the index of $P$ defined in [43].

## 4. Representations and dilations

Baruch's and my approach to understanding the $C^*$-envelope of $\mathcal{T}_+(E)$ was based on the (completely contractive) representations of $(E,A)$ and their dilations. Our starting point was the simple fact from algebra that there is a bijective correspondence between the representations of a tensor algebra and the bimodule representations of the bimodule from which the tensor algebra is constructed. So we began by proving that there is a bijective correspondence between the completely contractive representations of $\mathcal{T}_+(E)$ on a Hilbert space $H$ and pairs $(T,\sigma)$, where $\sigma : A \to B(H)$ is a $C^*$-representation and $T : E \to B(H)$ is a completely contractive bimodule map; i.e., $T(a\cdot\xi\cdot b) = \sigma(a)T(\xi)\sigma(b)$ [40, Theorem 3.10]. One direction is clear: Given a completely contractive representation $\rho$ of $\mathcal{T}_+(E)$ on $H$, setting $\sigma := \rho\circ\varphi_\infty$ and $T(\xi) := \rho(T_\xi)$, $\xi\in E$, gives the desired bimodule map. To go from $(T,\sigma)$ to a completely contractive representation of $\mathcal{T}_+(E)$, we proved in [40, Lemma 3.5] that a bimodule map $T$ is completely contractive if and only if the operator $\tilde{T} : E\otimes_\sigma H \to H$, defined by $\tilde{T}(\xi\otimes h) = T(\xi)h$, is a contraction and satisfies the equation

$$\tilde{T}\sigma^E\circ\varphi(\cdot) = \sigma(\cdot)\tilde{T}, \tag{3}$$

where $\sigma^E$ is Rieffel's induced representation of $\mathcal{L}(E)$ [61]. (Recall that $\sigma^E$ is defined by the formula $\sigma^E(T)(\xi\otimes h) = (T\xi)\otimes h$, $T\in\mathcal{L}(E)$, i.e., $\sigma^E(T)$ is just $T\otimes I$. It is useful to have both notations $\sigma^E(T)$ and $T\otimes I$.) The representation of $\mathcal{T}_+(E)$ determined by $(T,\sigma)$ is denoted $\sigma\times T$. Borrowing terminology from the theory of crossed products, we called such pairs (completely contractive) *covariant representations* of $(E,A)$ and we called $\sigma\times T$ the *integrated form* of $(T,\sigma)$.

In the special case when $A = E = \mathbb{C}$, $\sigma$ must represent $\mathbb{C}$ as scalar multiples of the identity operator on $H$. Also, $E\otimes_\sigma H$ may be canonically identified with $H$, and when this is done, $\sigma^E\circ\varphi$ may also be viewed as representing $\mathbb{C}$ as scalar multiples of the identity. With these identifications, $\widetilde{T}$ is simply an ordinary contraction operator on $H$. Thus, in this case, one captures the well-known fact that the (completely) contractive representations of the disc algebra on the Hilbert space $H$ are in bijective correspondence with the contraction operators on $H$.[5]

In the setting with $A = \mathbb{C}$ and $E = \mathbb{C}^d$, $A = \mathbb{C}$ still must be represented as scalar multiples of the identity and so in equation (3) $\sigma$ and $\sigma^E\circ\varphi$ may seem to be the same. But in fact, they are acting on different Hilbert spaces. The space $E\otimes_\sigma H$ should be viewed as $d$ copies of $H$ arranged in a column. When this is done, $\widetilde{T}$ is a so-called *row contraction,* i.e., $\widetilde{T} = (T_1, T_2, \dots, T_d)$, where each $T_i$ is an operator on $H$, with $\|\widetilde{T}\|^2 = \|\sum_i T_iT_i^*\| \le 1$.

[5] Of course, it is well known that every contractive representation of $A(\mathbb{D})$ is completely contractive, but from the perspective of the theory I am discussing here, that is a separate issue.

When $E = {}_{\varphi}A$, for some endomorphism $\varphi$ of $A$, then since $E$ is cyclic as a right $A$ module, $E \otimes_{\sigma} H$ is isomorphic to $H$ via the map $\xi \otimes h \to \sigma(\xi)h$. When the identification of $E \otimes_{\sigma} H$ and $H$ is made via this isomorphism, $\sigma^{E}$ is identified with $\sigma$, $\widetilde{T}$ is identified with the operator $T(1)$, and equation (3) becomes the more familiar covariance equation

$$\widetilde{T}\sigma \circ \varphi(\cdot) = \sigma(\cdot)\widetilde{T}.$$

To show that a completely contractive covariant representation $(T, \sigma)$ can be extended to a completely contractive representation of $\mathcal{T}_{+}(E)$, Baruch and I needed to show that $(T, \sigma)$, can be dilated to a covariant representation, $(V, \rho)$, where $\tilde{V}$ is an isometry [40, Theorem 3.3]. I won't rehearse the details here except to say that it was based on a careful analysis of Schäffer's matrix proof [63] of Sz.-Nagy's dilation theorem [65] and Popescu's generalization of it [54]. This dilation is unique, subject to a familiar minimality condition [40, Proposition 3.2]. The integrated form, $\rho \times V$, then, is a dilation of $\sigma \times T$. We called a covariant representation $(V, \rho)$, where $\widetilde{V}$ is an isometry, an *isometric covariant representation.* An isometric representation $(V, \rho)$ may be characterized by the property that $\rho \times V$ extends uniquely to a $C^*$-representation of the Toeplitz-Cuntz-Pimsner algebra $\mathcal{T}(E)$ [40, Theorem 2.12]. For this reason, isometric covariant representations of $(E, A)$ are called in the literature *Toeplitz representations.* (See [22], where this term was first used.) Thus every completely contractive representation of $\mathcal{T}_{+}(E)$ may be dilated to a $C^*$-representation of the Toeplitz algebra $\mathcal{T}(E)$.

Pimsner proved that isometric representations $(V, \rho)$ such that $\rho \times V$ passes to the quotient $\mathcal{O}(E)$ have the special property that $\rho^{(1)} \circ \varphi|_J = \rho|_J$, where $\rho^{(1)} : K(E) \to B(H_\rho)$ is defined by the formula

$$\rho^{(1)}(\xi \otimes \eta^*) := V(\xi) \otimes V(\eta)^* = \widetilde{V}(\xi \otimes \eta^*)\widetilde{V}^*, \tag{4}$$

$\xi \otimes \eta^* \in K(E)$ [52, Theorem 3.12]. Observe that in the case when $A = E = \mathbb{C}$, this condition on $V$ means, simply, that $\widetilde{V}$ is unitary. On the other hand, when one is dealing with a single isometry, say $W_+$, on Hilbert space $H$, then it is well known how to extend $W_+$ to a unitary operator $W$ on a Hilbert space $K$ containing $H$. Further, the extension is uniquely determined if one assumes that $W$ is minimal. However, in the setting discussed here, the problem of extending an isometric covariant representation $(\rho, V)$ to a bigger space to obtain a covariant representation with this property is more complicated. Also, the solution is not unique. The problem is that not only must $V$ be extended, but so must $\rho$, and the process for extending $\rho$ is based on the Hahn-Banach theorem. (See [40, 45] for further discussion of this matter.)

If $(T, \sigma)$ is an arbitrary completely contractive covariant representation of $E$, then one can form $\sigma^{(1)}$ using the same type of formula as in equation (4): $\sigma^{(1)}(\xi \otimes \eta^*) = \widetilde{T}(\xi \otimes \eta^*)\widetilde{T}^*$. In this more general case, however, $\sigma^{(1)}$ is only a completely positive map (see [41, Remark 5.2] and the surrounding discussion). If, however, $\sigma^{(1)} \circ \varphi|_J = \sigma|_J$, then $\widetilde{T}^*$ acts isometrically on the essential subspace of $\sigma|_J$, $\sigma(J)H$.

In this event, Baruch and I called $(T,\sigma)$ *coisometric.* When $\widetilde{T}^*$ is an isometry, i.e., when $\widetilde{T}$ is a co-isometry, then we called $(T,\sigma)$ *fully coisometric.* So, the covariant presentations $(T,\sigma)$ of $(E,A)$ such that $\sigma\times T$ gives a $C^*$-representation of $\mathcal{O}(E)$ are precisely the ones that are isometric and co-isometric. We have, however, resisted the temptation to call them unitary. The completely contractive representations $(T,\sigma)$ that are isometric and fully coisometric are precisely the ones that pass to a quotient of the Cuntz-Pimsner algebra called the Doplicher-Roberts algebra [21, Theorem 6,6].

In sum, then, every completely contractive covariant representation of $(E,A)$, $(T,\sigma)$, on a Hilbert space $H$, may be dilated to one, $(U,\rho)$, that is isometric and coisometric acting on a Hilbert space $K$ containing $H$. The integrated form, then, $\rho\times U$, is a $C^*$-representation of $\mathcal{O}(E)$ such that

$$\sigma\times T(F)=P\rho\times U(F)|_H,$$

for all $F\in\mathcal{T}_+(E)$. This equation is precisely equation (2), where the operator $V$ there is the imbedding of $H$ into $K$ so that $V^*$ becomes the projection $P$ of $K$ onto $H$.

## 5. Induced representations and Halmos's theorem

A representation $\pi=\rho\times V$ of $\mathcal{T}_+(E)$ is called *induced* if there is a representation $\sigma:A\to B(H_\sigma)$ such that $(V,\rho)$ is (unitarily equivalent to) the representation on $\mathcal{F}(E)\otimes_\sigma H_\sigma$ given by the formulas $\rho(a)=\varphi_\infty(a)\otimes I$ and $V(\xi)=T_\xi\otimes I$. In [41], Baruch and I showed that these are natural generalizations of pure isometries, i.e., shifts, and that every isometric representation of $\mathcal{T}_+(E)$ splits as the direct sum of an induced representation and a fully coisometric representation of the Cuntz-Pimsner algebra $\mathcal{O}(E)$ restricted to $\mathcal{T}_+(E)$. This decomposition is a natural generalization of the Wold decomposition of an isometry.

To see more clearly the connection with shifts and the Wold decomposition, consider the case when $A=E=\mathbb{C}$. In this setting, the Fock space $\mathcal{F}(E)$ is $\ell^2(\mathbb{Z}_+)$, $\varphi_\infty$ is just the representation of $A=\mathbb{C}$ in $B(\ell^2(\mathbb{Z}_+))$ as scalar multiples of the identity and $T_\xi=\xi T_1$, where $T_1$ is the unilateral shift, as before. Also, $\sigma$ must be the representation of $A=\mathbb{C}$ as scalar multiples of the identity in $B(H_\sigma)$. Consequently, $\rho$ is the representation of $A=\mathbb{C}$ as scalar multiples of the identity in $B(\ell^2(\mathbb{Z}_+)\otimes_\sigma H_\sigma)$ and $V(\xi)=\xi V(1)=\xi(T_1\otimes I)$. Of course $T_1\otimes I$ is the unilateral shift of multiplicity equal to the dimension of $H_\sigma$. It is not hard to see that two induced isometric representations of $\mathcal{T}_+(E)$ are unitarily equivalent if and only if the two representations of $A$ from which they are induced are unitarily equivalent. This fact is a manifest generalization of the fact that two shifts are unitarily equivalent if and only if their multiplicities are equal.

The key to Baruch's and my approach to the Wold decomposition theorem – and to other things, as well – is to analyze the intertwining equation (3) associated with a general completely contractive covariant representation $(T,\sigma)$ of $(E,A)$. The problem with $\widetilde{T}$ is that it acts from $E \otimes_\sigma H_\sigma$ to $H_\sigma$, and so one cannot form its powers. However, there is a way around this. Simply define $\widetilde{T}_n : E^{\otimes n} \otimes H_\sigma \to H_\sigma$ by the formula $\widetilde{T}_n(\xi_1 \otimes \xi_2 \otimes \cdots \otimes \xi_n \otimes h) := T(\xi_1)T(\xi_2)\cdots T(\xi_n)h$, $\xi_1 \otimes \xi_2 \otimes \cdots \otimes \xi_n \otimes h \in E^{\otimes n} \otimes H_\sigma$. Of course, $\widetilde{T}_1 = \widetilde{T}$. Then each $\widetilde{T}_n$ is a contraction operator and they are all related via the equation

$$\begin{aligned} \widetilde{T}_{n+m} &= \widetilde{T}_n(I_{E^{\otimes n}} \otimes \widetilde{T}_m) \\ &= \widetilde{T}_m(I_{E^{\otimes m}} \otimes \widetilde{T}_n). \end{aligned} \tag{5}$$

An alternate perspective we have found useful is to promote $\widetilde{T}$ to an operator matrix $\widetilde{\widetilde{T}}$ on $\mathcal{F}(E) \otimes H_\sigma$ via the formula

$$\widetilde{\widetilde{T}} = \begin{bmatrix} 0 & \widetilde{T} & & & \\ & 0 & I_E \otimes \widetilde{T} & & \\ & & 0 & I_{E^{\otimes 2}} \otimes \widetilde{T} & \\ & & & 0 & \ddots \\ & & & & \ddots \end{bmatrix}.$$

Then the formulas in equation (5) may be read from the matrix entries of the powers of $\widetilde{\widetilde{T}}$.

If $(V, \rho)$ is an isometric covariant representation on a Hilbert space $H$, then all the $\widetilde{V}_n$'s are isometries [41, Lemma 2.2] and they implement the powers of an endomorphism $\Phi$ of the *commutant* of $\rho(A)$, $\rho(A)'$, via the equation:

$$\Phi^n(x) = \widetilde{V}_n(I_{E^{\otimes n}} \otimes x)\widetilde{V}_n^*, \tag{6}$$

$x \in \rho(A)'$ [41, Lemma 2.3]. Observe that $\Phi$ is non-unital precisely when $\widetilde{V}$ is not surjective, i.e., precisely when $(V, \rho)$ is not fully coisometric. In this event, the projections $P_n := \Phi^n(I)$ form a decreasing family whose infimum $P_\infty := \bigwedge_{n \geq 0} P_n$ is the projection onto a subspace that reduces $(V, \rho)$. Further, the restriction $(V_f, \rho_f)$ of $(V, \rho)$ to $P_\infty H$ is isometric and fully coisometric [46, Theorem 2.9]. Consequently, $\rho_f \times V_f$ is a representation of $\mathcal{O}(E)$ on $P_\infty H$.

On the other hand, if $Q := I - \Phi(I)$, and if $Q_n := \Phi^n(Q)$, $n \geq 0$, so $Q = Q_0$, then the $Q_n$ satisfy the following properties:

1. Each $Q_n$ commutes with $\rho(A)$.
2. $Q_n Q_m = 0$, $n \neq m$.
3. $\sum_{n \geq 0}^{\oplus} Q_n = I - P_\infty$.
4. If $H_0 := Q_0 H$ and if $\sigma$ is the representation of $A$ on $H_0$ defined by restricting $\rho$ to $H_0$, then $\widetilde{V}_n$ is a Hilbert space isomorphism from $E^{\otimes n} \otimes H_0$ onto $H_0$ such that $\widetilde{V}_n(\varphi_n(a) \otimes I_{H_0}) = \sigma(a)\widetilde{V}_n$ for all $a \in A$, where $\varphi_n$ gives the left action of $A$ on $E^{\otimes n}$ via the formula $\varphi_n(a)(\xi_1 \otimes \xi_2 \otimes \cdots \otimes \xi_n) := (\varphi(a)\xi_1) \otimes \xi_2 \otimes \cdots \otimes \xi_n$.

Since $I$ lies in $\rho(A)'$, so does each $Q_n$ by [41, Lemma 2.3]. The orthogonality relations 2. are immediate from the fact that $\Phi$ is an endomorphism. Also, the "completeness" assertion, 3., follows from $P_\infty := \bigwedge_{n \geq 0} P_n$ because $I - P_{n+1} = \sum_{k=0}^{n} Q_k$.

Finally, 4. is an easy consequence of equation (3) applied inductively to the definition of $\widetilde{V}_n$. So, if we let $\widetilde{V}_0$ be the identity operator and set $U := \sum_{n\geq 0} \widetilde{V}_n$, then $U$ a Hilbert space isomorphism from $\mathcal{F}(E) \otimes_\sigma H_0$ onto $H \ominus P_\infty H$ that implements a unitary equivalence between the isometric covariant representation of $(E, A)$ induced by $\sigma$ and the restriction of $(V, \rho)$ to $(I - P_\infty)H$. This gives the analogue of the Wold decomposition theorem proved as [41, Theorem 2.9]: *If $(V, \rho)$ is an isometric covariant representation of $(E, A)$, then $(V, \rho)$ decomposes uniquely as the direct sum $(V, \rho) = (V_i, \rho_i) \oplus (V_f, \rho_f)$, where $(V_i, \rho_i)$ is an induced representation and $(V_f, \rho_f)$ is isometric and fully coisometric.*

Of course, one recognizes an analogy between the wandering subspaces of Halmos [25] and the $Q_n$. In particular property (3) is an exact analogue of [25, Lemma 1] (Lemma 2.3). However, on close inspection the analogy breaks down. If $W$ is an isometry on a Hilbert space $H$ and if $\mathcal{M} \subseteq H$ is a wandering subspace for $W$, i.e, if $\mathcal{M}$ and $W\mathcal{M}$ are orthogonal subspaces, then not only must the family $\{W^n\mathcal{M}\}_{n\geq 0}$ be a family of mutually orthogonal subspaces, they also have the same dimension. One might expect, therefore, that in the setting we are discussing, the representations obtained by restricting $\rho$ to the ranges of the $Q_n$ are all unitarily equivalent. However, this need not be the case. In fact, they may all be disjoint; i.e., there may be no nonzero operators intertwining them! (I will indicate why in moment.) Further, suppose $\rho$ is a representation of $\mathcal{T}_+(E)$ that is induced by a representation $\sigma$ of $A$ on $H_\sigma$, so $\rho(\mathcal{T}_+(E))$ acts on $\mathcal{F}(E) \otimes_\sigma H_\sigma$, and suppose $\mathcal{M}$ is a subspace of $\mathcal{F}(E) \otimes_\sigma H_\sigma$ that is invariant under $\rho(\mathcal{T}_+(E))$. Then while it is true that the restriction of $\rho(\mathcal{T}_+(E))$ to $\mathcal{M}$ yields a representation $\rho_1$ that is induced, say by the representation $\sigma_1$ of $A$ [41, Proposition 2.11], the relation between $\sigma$ and $\sigma_1$ can be complicated. Thus, it appears that all hope of generalizing Halmos's theorem to the setting I am describing is lost. However, things are not as bleak as they may seem. One needs to add an extra concept which is taken from ergodic theory: quasi-invariance.

First recall that two $C^*$-representations $\pi_1$ and $\pi_2$ of a $C^*$ -algebra $A$ are called *quasi-equivalent* in case some multiple of $\pi_1$ is unitarily equivalent to a multiple of $\pi_2$ [18, 5.3.1 and 5.3.2]. In the situation of this discussion, where $E$ is a $C^*$-correspondence over $A$, Baruch and I called a representation $\pi$ of $A$ on Hilbert space $H$ *quasi-invariant under* $E$ in case $\pi$ and the induced representation $\pi^{\mathcal{F}(E)} \circ \varphi_\infty$ are quasi-equivalent. Since $\pi$ is a subrepresentation of $\pi^{\mathcal{F}(E)} \circ \varphi_\infty$, $\pi$ is quasi-invariant if and only if $\pi^{\mathcal{F}(E)} \circ \varphi_\infty$ is unitarily equivalent to a multiple of $\pi$. The choice of terminology is justified by this example [41, Remark 4.6]: Suppose $A = C(X)$ for some compact Hausdorff space $X$, suppose $E$ is the correspondence determined by a homeomorphism $\tau$ of $X$, i.e., suppose $E = {}_\tau C(X)$, where we think of $\tau$ as inducing an automorphism of $C(X)$ in the usual fashion, and suppose that $\pi$ is the multiplication representation of $C(X)$ on $L^2(\mu)$ for some measure $\mu$ on $X$, $(\pi(f)\xi)(x) = f(x)\xi(x)$, $f \in C(X)$, $\xi \in L^2(\mu)$). Then $\pi$ is quasi-invariant in the sense just defined if and only if the measure $\mu$ is quasi-invariant in the usual sense, vis., the $\tau$-translate of any $\mu$-null set is a $\mu$-null set. At the other extreme, it is possible for the measures $\mu \circ \tau^n$ to be pairwise singular. In this event, the representation $\rho$ of $\mathcal{T}_+(E)$ on $\mathcal{F}(E) \otimes_\pi H_\pi$ that is induced by $\pi$ has the property that the representations of $A$ obtained by restricting $\rho \circ \varphi_\infty$ to $Q_n\mathcal{F}(E) \otimes_\pi H_\pi$ are pairwise disjoint.

Not only is "quasi-invariance" sufficient for formulating a version of Halmos's theorem in this context, it is necessary.

**Theorem 5.1.** [41, Theorem 4.7] *Suppose $\pi$ is a $C^*$-representation of $A$ on a Hilbert space $H_\pi$, suppose that $\rho$ is the representation of $\mathcal{T}_+(E)$ on $\mathcal{F}(E) \otimes_\pi H_\pi$ that is induced by $\pi$ and suppose that $\mathcal{M} \subseteq \mathcal{F}(E) \otimes_\pi H_\pi$ is a subspace that is invariant under $\rho(\mathcal{T}_+(E))$. If $\pi$ is quasi-invariant under $E$, then there is a family of partial isometries $\{\Theta_i\}_{i \in I}$ with orthogonal ranges in the commutant of $\rho(\mathcal{T}_+(E))$ such that $\mathcal{M} = \sum_{i \in I}^{\oplus} \Theta_i \mathcal{F}(E) \otimes_\pi H_\pi$. Conversely, if every subspace of $\mathcal{F}(E) \otimes_\pi H_\pi$ that is invariant under $\rho(\mathcal{T}_+(E))$ has this form, then $\pi$ is quasi-invariant under $E$.*

To see why this is so, let $P$ be the projection of $\mathcal{F}(E) \otimes_\pi H_\pi$ onto $\mathcal{M}$. Since $\mathcal{M}$ reduces $\rho \circ \varphi_\infty$, $P$ commutes with $\rho \circ \varphi_\infty(A)$, and so, therefore, does $Q := P - \Phi(P)$. Let $\mathcal{M}_0$ be the range of $Q$. Since the restriction of $\rho \circ \varphi_\infty$ to $\mathcal{M}_0$ is a subrepresentation of $\rho \circ \varphi_\infty$, it is quasi-equivalent to a subrepresentation of $\pi$. Consequently, from [18, 5.3.1] and the well-known structure of normal homomorphisms between von Neumann algebras, there is a family of partial isometries $\{\Theta_{0i}\}_{i \in I}$ with orthogonal ranges, mapping $H_\pi$ to $\mathcal{M}_0$, such that $\Theta_{0i}\pi(\cdot) = \rho \circ \varphi_\infty(\cdot)|_{\mathcal{M}_0}\Theta_{0i}$ and such that the range of $\sum_{i \in I} \Theta_{0i}\Theta_{0i}^*$ is $\mathcal{M}_0$. If $\Theta_i$ is defined to be $\sum_{k \geq 0} \Phi^k(\Theta_{0i})$, then an easy calculation shows: (i) Each $\Theta_i$ is a partial isometry in the commutant of $\rho(\mathcal{T}_+(E))$, (ii) the ranges of the $\Theta_i$ are mutually orthogonal, and (iii) and $\mathcal{M} = \sum_{i \in I}^{\oplus} \Theta_i \mathcal{F}(E) \otimes_\pi H_\pi$.

This part of the proof is clearly a variation of Halmos's arguments for Theorem [25, Theorem 3]. (See the justification for equation (1).) It is very similar to that given by Popescu [55, Theorem 2.2], which was based on his generalization of the Wold decomposition [54]. Davidson and Pitts made similar arguments for Theorem 2.1 in [16].

For the converse, observe that each space $\sum_{k \geq n} E^{\otimes k} \otimes H_\pi$ is invariant under $\rho(\mathcal{T}_+(E))$. If each of these has the indicated form, then it is not difficult to see that each $\Theta_i$ restricted to $H_\pi$, regarded as the 0th summand of $\mathcal{F}(E) \otimes_\pi H_\pi$, intertwines $\pi$ and the representation $\pi^{E^{\otimes n}} \circ \varphi_n$ of $A$ on $E^{\otimes n} \otimes_\pi H_\pi$. Since the ranges of the $\Theta_i$ sum to $E^{\otimes n} \otimes_\pi H_\pi$, it follows that each representation $\pi^{E^{\otimes n}} \circ \varphi_n$ is quasi-equivalent to a subrepresentation of $\pi$. Since $\pi$ is a subrepresentation of $\pi^{\mathcal{F}(E)} \circ \varphi_\infty$, the representation $\pi^{\mathcal{F}(E)} \circ \varphi_\infty$ is quasi-equivalent to $\pi$, i.e., $\pi$ is quasi-invariant under $E$.

Absent from the discussion so far is Halmos's fundamental lemma [25, Lemma 4] (i.e., Lemma 2.4), which asserts that if $U_+$ is a unilateral shift of multiplicity $n$, then every wandering subspace for $U_+$ has dimension at most $n$. That is what is necessary to conclude that the minimal number of partial isometries $\Theta_i$ necessary in the representation in Theorem 5.1 – $\mathcal{M} = \sum_{i \in I}^{\oplus} \Theta_i \mathcal{F}(E) \otimes_\pi H_\pi$ – is one. One may wonder, in particular, if Theorem 5.1 contains Halmos's theorem, Theorem 2.1, as a special case. That question has two answers, at least, one "yes", the other "no". The answer is "yes" in this sense: Once one knows that the dimension of $\mathcal{M}_0$ is at most the dimension of $H_\pi$, then one only needs one $\Theta_{0i}$. Further, I think it is fair to say that Halperin's elegant proof [66, Page 108] that the dimension of $\mathcal{M}_0$ must be no more than the dimension of $H_\pi$ is as simple as it can be. So, one can say that in the case when $A = E = \mathbb{C}$, Theorem 5.1 yields [25, Theorem 3]. On the other hand, what is really at issue here is the comparison theory of projections in the commutant of $\rho \circ \varphi_\infty$. This is a subject that I have taken up with various collaborators in the context non-self-adjoint crossed products, also called analytic crossed products. (However, we did not express ourselves in terms of correspondences.) The main result of [37, Theorem 3.3], for example, can be phrased like this: Suppose $A$ is a finite factor, and suppose $E = {}_\varphi A$ is the correspondence associated with an automorphism $\varphi$ of $A$. Suppose $\pi$ is a normal representation of $A$ such that $\pi(A)$ admits a separating and cyclic vector. Then the dimension of $\mathcal{M}_0$ relative to the commutant of $\rho \circ \varphi_\infty(A)$ is dominated by the dimension of $H_\pi$ relative to the commutant of $\rho \circ \varphi_\infty(A)$. Consequently, in this case, one $\Theta_{0i}$ suffices. The results of [38] show that one can get a similar sort of theorem without the assumption that $A$ is a factor, but then one must assume that $\varphi$ fixes the center elementwise. I don't know a general useful statement involving $A$, $E$ and $\pi$ that guarantees that $\mathcal{M}_0$ is dominated by $H_\pi$ in the commutant of $\rho \circ \varphi_\infty(A)$.

At the risk of sounding Panglossian, I think the possibility that more than one $\Theta_i$ may be necessary in Theorem 5.1 is terrific good news. The reason is this. Recall that Beurling's theorem and Halmos's generalization of it are the mainstays of the analogy promoted in operator theory that the disc algebra $A(\mathbb{D})$ and $H^\infty(\mathbb{D})$ should be regarded as replacements for the polynomials in one indeterminant, $\mathbb{C}[X]$. Of course, $A(\mathbb{D})$ and $H^\infty(\mathbb{D})$ are completions of $\mathbb{C}[X]$, but what makes the function theory in these algebras so compelling is the fact that each weak-$*$ closed ideal in $H^\infty(\mathbb{D})$ is principal; it is generated by an inner function. This is an easy consequence of Beurling's theorem. Halmos's theorem implies that the same is true for any left or right weak-$*$ closed ideal in $H^\infty(\mathbb{D}) \otimes \mathcal{M}_n(\mathbb{C})$. When $n = \infty$, one must replace $M_n(\mathbb{C})$ by $B(H)$ for an infinite-dimensional Hilbert space $H$ and one must understand $H^\infty(\mathbb{D}) \otimes B(H)$ as a completion of the algebraic tensor product. Further, the importance of $H^\infty(\mathbb{D}) \otimes M_n(\mathbb{C})$ resides also in the fact that it appears as the commutant of the unilateral shift of multiplicity $n$. The work of Popescu, Davidson and Pitts, and others make it clear that Popesecu's noncommutative disc algebra and the free semigroup algebras should be regarded as operator algebra versions of *free algebras.* Indeed, analogous to the one-variable setting they are completions of free algebras. Their generalizations of Halmos's Theorem, [55, Theorem 2.2] and [16, Theorem 2.1], show that the free semigroup algebras, $\mathcal{L}_d$, are *free ideal rings* in the sense that each weak-$*$ closed left or right ideal is free as an $\mathcal{L}_d$-module [12, 1.2]. Theorem 5.1 shows that the tensor algebras and Hardy algebras (to be discussed in a minute) should be regarded as generalized free ideal rings, too.

## 6. Duality and commutants

What is conspicuously missing in Theorem 5.1 and the theory discussed so far is a description of the commutant of an induced representation. In the setting of free semigroup algebras, Geulu Popescu showed that the commutant of $\mathcal{L}_d$ acting on $\mathcal{F}(\mathbb{C}^d)$ (to the left) is the weakly closed algebra $\mathcal{R}_d$ generated by the "right creation operators", i.e., the operators defined by the equation $R_\xi\eta := \eta \otimes \xi$, $\xi \in \mathbb{C}^d, \eta \in \mathcal{F}(\mathbb{C}^d)$ [57]. Further, he showed that the map $U$ defined on decomposable tensors in $\mathcal{F}(\mathbb{C}^d)$ by the formula

$$U(\xi_1 \otimes \xi_2 \otimes \cdots \otimes \xi_n) = \xi_n \otimes \xi_{n-1} \otimes \cdots \otimes \xi_1$$

extends to a unitary operator on $\mathcal{F}(\mathbb{C}^d)$ such that $UT_\xi U^* = R_\xi$, for all $\xi \in \mathbb{C}^d$. Thus $\mathcal{L}_d$ and $\mathcal{R}_d$ are isomorphic. Also, the commutant of an induced representation of $\mathcal{L}_d$, the image of which is $\mathcal{L}_d \otimes I_H$ for some Hilbert space $H$, is $\mathcal{R}_d \otimes B(H)$. These facts are used quite a bit in the free semigroup algebra literature. Note, however, that the formulas make very little sense at the level of general tensor algebras: $R_\xi$ isn't a module map and $U$ needn't be an isometry – indeed, $U$ isn't even well defined in some cases. Thus the commutant of an induced representation appears to be mysterious. Certainly, it was to Baruch and me when we wrote [41].

The breakthrough came to us when we were studying product systems of $W^*$-correspondences in [42]. We were inspired to study these by Arveson's work on $E_0$-semigroups (See [5] for a full treatment.) and by the belief that algebras generated by creation operators on product systems should give rise to analogues of Hardy space theory on the upper half-plane. While the function-theoretic aspiration is still largely unfulfilled, we found interesting connections with semigroups of completely positive maps and a host of other things, upon which I won't elaborate here. The breakthrough was that the commutant problem rests, ultimately, on how to interpret equation (3).

Recall that the point of equation (3) is this: the completely contractive representations of $\mathcal{T}_+(E)$ are determined by pairs $(\sigma, \widetilde{T})$ where $\sigma$ is a $C^*$-representation of $A$ on a Hilbert space $H$ and $\widetilde{T}$ is a contraction operator from $E \otimes_\sigma H$ to $H$ that intertwines $\sigma^E \circ \varphi$ and $\sigma$. But the intertwining space is a Banach space of operators and the contraction operators in it comprise its closed unit ball. More than that, the set of *adjoints* of the operators in this intertwining space is naturally a $W^*$-correspondence over the commutant of $\sigma(A)$, $\sigma(A)'$. We focus on this space of adjoints, $\{X \in B(H, E \otimes_\sigma H) \mid X\sigma(a) = \sigma^E \circ \varphi(a)X\}$, calling it the *$\sigma$-dual of $E$* and denoting it by $E^\sigma$. The bimodule structure on $E^\sigma$ is defined via the formulae

$$X \cdot a := Xa,$$

and

$$a \cdot X := (I_E \otimes a)X,$$

$a \in \sigma(A)'$ and $X \in E^\sigma$. Further, the $\sigma(A)'$-valued inner product is defined by the formula

$$\langle X, Y \rangle := X^*Y,$$

$X, Y \in E^\sigma$. With these operations, it is easily seen that $E^\sigma$ is a $C^*$-correspondence over $\sigma(A)'$. However, by virtue of being an intertwining space, $E^\sigma$ is weakly closed and therefore inherits a dual space structure that is compatible with the natural dual space structure on $\sigma(A)'$.

I won't go into detail here, but these dual space structures and compatibility relations are what are necessary for a $C^*$-correspondence over a $W^*$-algebra to be a $W^*$-correspondence. For further information see [42, Section 2], which is based primarily on [6] and [49]. So, when I use the term $W^*$-correspondence, it will suffice to think $C^*$-correspondence over a $W^*$-algebra that has additional structure that will not be of concern in the discussion. There are, however, a few minor adjustments one must make to some of the terms we have been discussing in the context of $C^*$-correspondences. Suppose $E$ is a $W^*$-correspondence over a $W^*$-algebra $A$; then the Fock space over $E$, $\mathcal{F}(E)$, is taken to be a *completion* of the Fock space of $E$ regarded simply as a $C^*$-correspondence. When this is done, $\mathcal{F}(E)$ is a $W^*$-correspondence and so the algebra $\mathcal{L}(\mathcal{F}(E))$ is a $W^*$-algebra [49, Proposition 3.10]. The *Hardy algebra of* $E$ is defined to be the closure of the tensor algebra $\mathcal{T}_+(E)$ in the weak-$*$ topology on $\mathcal{L}(\mathcal{F}(E))$ and is denoted $H^\infty(E)$. Observe that when $A = E = \mathbb{C}$, and when they are viewed as a $W^*$-algebra and a $W^*$-correspondence, respectively, then $\mathcal{F}(E) = \ell^2(\mathbb{Z}_+)$ and $H^\infty(E)$ is $H^\infty(\mathbb{D})$ realized as the algebra of all analytic Toeplitz operators.

To get a feeling for the $\sigma$-dual of a correspondence, consider first the simple case when $E = \mathbb{C}^d$ and let $\sigma$ be the representation of $\mathbb{C}$ on $\mathbb{C}^n$ in the only way possible: $\sigma(c) = cI_n$, where $I_n$ the $n \times n$ identity matrix. Then as I remarked earlier, $E \otimes_\sigma \mathbb{C}^n$ should be viewed as $d$ copies of $\mathbb{C}^n$ arranged in a column. Consequently, since the intertwining condition is trivial (operators are linear, by assumption), $E^\sigma$ is just the collection of all $d$-tuples of $n \times n$ matrices arranged in a column. This space is often denoted $\mathbf{C}_d(M_n(\mathbb{C}))$ and called *column d-space over* $M_n(\mathbb{C})$. More generally, one can compute duals of correspondences associated with directed graphs and it turns out that these are natural "graphs of matrix spaces indexed by the opposite graph" [44, Example 4.3]. Suppose $\varphi$ is an endomorphism of the $C^*$-algebra $A$, that $E$ is the $C^*$-correspondence ${}_\varphi A$, and that $\sigma$ is a representation of $A$ on the Hilbert space $H$. Suppose, for the sake of discussion, that $\sigma(A)$ admits a cyclic vector $\Omega$ and that the state determined by $\Omega$ is invariant under $\varphi$. Then the map $S$ defined by the equation, $S(\sigma(a)\Omega) = \sigma(\varphi(a))\Omega$, extends to an isometry defined on all of $H$ and $E^\sigma$ is naturally identified as $\{mS \mid m \in \sigma(\varphi(A))'\}$ [44, Example 4.6]. More generally, if $E = A \otimes_P A$ is the GNS correspondence determined by a unital completely positive map on the $C^*$-algebra $A$ and if $\sigma$ is a representation of $A$ on a Hilbert space $H$, then it is not difficult to see that $E^\sigma$ is the space of all operators from $H$ to $A \otimes_{\sigma\circ P} H$ that intertwine $\sigma$ and the Stinespring representation associated with $\sigma \circ P$ [44, Example 4.4]. (Recall that $A \otimes_{\sigma\circ P} H$ is the Hausdorff completion of the algebraic tensor product $A \otimes H$ in the pre-inner product $\langle a_1 \otimes h_1, a_2 \otimes h_2\rangle := \langle h_1, \sigma \circ P(a_1^* a_2) h_2\rangle$ and the Stinespring representation $\pi$ is given by the formula $\pi(a)(b \otimes h) = (ab) \otimes h$.)

Now $E^\sigma$ is a $W^*$-correspondence over $\sigma(A)'$, where $\sigma$ is a representation of $A$ on $H$, and so one can contemplate the identity representation $\iota$ of $\sigma(A)'$ on $H$ and one can form $(E^\sigma)^\iota$. This, of course, is a $W^*$-correspondence over $\sigma(A)''$, the weak closure of $\sigma(A)$. It is naturally isomorphic to a $W^*$-completion of $E$. More accurately, if $A$ were a $W^*$-algebra, if $E$ were a $W^*$-correspondence over $A$, and if $\sigma$ were a faithful normal representation of $A$ on $H$, then $(E^\sigma)^\iota$ would be naturally isomorphic to $E$ [44, Theorem 3.6]. This isomorphism, in turn, lies at the heart of the connection between duals and tensor products: If $E_1$ and $E_2$ are $W^*$-correspondences over a $W^*$-algebra $A$ and if $\sigma$ is a faithful normal representation of $A$ on a Hilbert space $H$, then $E_1^\sigma \otimes E_2^\sigma$ is isomorphic to $(E_2 \otimes E_1)^\sigma$ via the map

$$\eta_1 \otimes \eta_2 \mapsto (I_{E_2} \otimes \eta_1)\eta_2$$

by [44, Lemma 3.7]. Observe that the composition in this equation makes sense when properly interpreted: $\eta_2$ is a map from $H$ to $E_2 \otimes_\sigma H$ and $\eta_1$ is a map from $H$ to $E_1 \otimes H$. Consequently, $I_{E_2} \otimes \eta_1$ is a map from $E_2 \otimes_\sigma H$ and the composite, $(I_{E_2} \otimes \eta_1)\eta_2$, then, is a map from $H$ to $E_2 \otimes_{I_{E_1} \otimes \sigma} (E_1 \otimes_\sigma H)$. Since this space is naturally identified with $(E_2 \otimes E_1) \otimes_\sigma H$, the composition $(I_{E_2} \otimes \eta_1)\eta_2$ has the right properties.

The solution to the commutant mystery is that the commutant of the image of a tensor algebra under an induced representation is naturally isomorphic to an induced representation of the Hardy algebra of the dual correspondence. It is best to frame this entirely in terms of Hardy algebras. So, let $E$ be a $W^*$-correspondence over a $W^*$-algebra $A$ and let $\sigma$ be a faithful normal representation of $A$ on a Hilbert space $H$. Also form the dual $W^*$-correspondence $E^\sigma$ over $\sigma(A)'$ and let $\iota$ denote the identity representation of $\sigma(A)'$ on $H$. Finally, let $\lambda$ be the representation of $H^\infty(E)$ induced by $\sigma$ and let $\rho$ be the representation of $H^\infty(E^\sigma)$ induced by $\iota$.

**Theorem 6.1.** [44, Theorem 3.9] *Define $U$ from $\mathcal{F}(E^\sigma) \otimes_\iota H$ to $\mathcal{F}(E) \otimes_\sigma H$ by the formula*

$$U(\eta_1 \otimes \eta_2 \otimes \cdots \otimes \eta_n \otimes h) := (I_{E^{\otimes(n-1)}} \otimes \eta_1)(I_{E^{\otimes(n-2)}} \otimes \eta_2) \cdots (I_E \otimes \eta_{n-1})\eta_n h,$$

*for decomposable tensors $\eta_1 \otimes \eta_2 \otimes \cdots \otimes \eta_n \otimes h$ in $(E^\sigma)^{\otimes n} \otimes_\iota H \subseteq \mathcal{F}(E^\sigma) \otimes_\iota H$. Then $U$ is a Hilbert space isomorphism from $\mathcal{F}(E^\sigma) \otimes_\iota H$ onto $\mathcal{F}(E) \otimes_\sigma H$ with the property that $U\rho(H^\infty(E^\sigma))U^*$ is the commutant of $\lambda(H^\infty(E))$.*

## 7. Noncommutative function theory

The identification of the commutant of an induced representation and the structure of a $\sigma$-dual led Baruch and me to view elements of tensor and Hardy algebras as functions on “the unit disc” in the $\sigma$-dual. This, in turn, has led us to study a kind of noncommutative function theory that has interesting algebraic and analytic progenitors. I want to describe some of it. As I indicated above, once $\sigma : A \to B(H)$ is fixed, the points in the closed unit ball of $E^{\sigma *}$, $\overline{\mathbb{D}(E^{\sigma *})}$, label all the representations $\rho$ of the tensor algebra $\mathcal{T}_+(E)$ with the property that $\rho \circ \varphi_\infty = \sigma$.

This observation invites one to view elements of $\mathcal{T}_+(E)$ as $B(H)$-valued functions on $\overline{\mathbb{D}(E^{\sigma *})}$. For $F \in \mathcal{T}_+(E)$, the corresponding function will be denoted $\widehat{F}$, and its value at a point $\eta^* \in \overline{\mathbb{D}(E^{\sigma *})}$ is defined by the equation

$$\widehat{F}(\eta^*) := \sigma \times \eta^*(F). \tag{7}$$

This formula may seem obscure, unclear, and *ad hoc*, but in fact special cases of it are quite well known and very useful, e.g., the holomorphic functional calculus and the Sz.-Nagy–Foiaş functional calculus.

Suppose that $A = E = \mathbb{C}$ and that $\sigma$ is the one-dimensional representation of $\mathbb{C}$. Then it is immediate that $E^{\sigma *}$ may be identified with the complex plane, and so $\overline{\mathbb{D}(E^{\sigma *})}$ is just the closed unit disc. From the way $\mathcal{T}_+(E)$ is defined, an element $F$ in $\mathcal{T}_+(E)$ is an analytic Toeplitz operator and so its matrix with respect to the usual basis in $\mathcal{F}(E) = \ell^2(\mathbb{Z}_+)$ has the form

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

If $\eta^*$ is in the closed unit disc, then the value $\widehat{F}(\eta^*)$ in $B(H) = \mathbb{C}$ defined through equation (7) is $\sum_{k\geq 0} a_k(\eta^*)^k$. This series converges, of course, when $|\eta^*| < 1$. If $\eta^*$ is on the boundary, then the series is Abel summable. If $\sigma$ is the representation of $\mathbb{C}$ by scalar multiples of the identity on a Hilbert space $H$, finite or infinite dimensional, then $E^{\sigma *}$ is the full algebra of operators on $H$, $B(H)$, and if $\eta^* \in \overline{\mathbb{D}(E^{\sigma *})}$, then $\eta^*$ is a contraction operator. Again, equation (7) gives $\widehat{F}(\eta^*) = \sum_{k\geq 0} a_k(\eta^*)^k$, where the series converges in operator norm, if $\eta^*$ is in the open ball $\mathbb{D}(E^{\sigma *})$, while if $\|\eta^*\| = 1$, then, in general, the series is only Abel summable. I should note that the radius of convergence of the series representing $\widehat{F}$ is at least the reciprocal of the spectral radius of $\eta^*$, which is always $\geq 1$.

I want to digress momentarily to point out an important difference between $\mathcal{T}_+(E)$ and $H^\infty(E)$, in the case when $E$ is a $W^*$-correspondence over a $W^*$-algebra. If $F \in \mathcal{T}_+(E)$, then formula (7) makes sense at every point $\eta^*$ in $\overline{\mathbb{D}(E^{\sigma *})}$. On the other hand, if $F$ is merely assumed to be in $H^\infty(E)$, then $\widehat{F}(\eta^*)$ makes sense for all $\eta^*$ in the open ball $\mathbb{D}(E^{\sigma *})$ because $\sigma \times \eta^*$ extends from $\mathcal{T}_+(E)$ to an ultraweakly continuous, completely contractive representation of $H^\infty(E)$ in $B(H)$ in this case [44, Corollary 2.14]. However, $\sigma \times \eta^*$ need not extend from $\mathcal{T}_+(E)$ to $H^\infty(E)$ as an ultraweakly continuous representation when $\eta^*$ on the boundary of $\mathbb{D}(E^{\sigma *})$. In particular, when $A = E = \mathbb{C}$, then $H^\infty(E)$ is $H^\infty(\mathbb{D})$ and, for example, if $\eta^*$ has an eigenvalue of modulus 1, then $\sigma \times \eta^*$ will never have such an extension. On the other hand, if $\eta^*$ is a completely non-unitary contraction, then it is a key fact about the Sz.-Nagy–Foiaş functional calculus that $\sigma \times \eta^*$ admits an ultraweakly continuous extension to $H^\infty(E)$. (Of course, when $H$ is finite dimensional being completely non-unitary and having spectral radius less than one are one and the same thing.) One of the interesting questions in the subject is to find general conditions that allow $\sigma \times \eta^*$ to extend to an ultraweakly continuous representation of $H^\infty(E)$. A sufficient condition was isolated in the free semigroup setting by Popescu; he assumed that the representation was "completely non-coisometric" (see his [54, Proposition 2.9]). Baruch and I generalized his work in Section 7 of [44]. However, as far as I know, no necessary condition has been found. As one might imagine, the issue is closely related to the problem of formulating a good notion of absolute continuity for representations of Cuntz algebras and, more generally, of Cuntz-Pimsner algebras. (See [13, 17] for contributions to this "absolute continuity" problem.)

When $A = \mathbb{C}$ and $E = \mathbb{C}^d$ and $\sigma$ is a representation of $A$ on a Hilbert space $H$, then as mentioned above, $E^\sigma$ is column space over $B(H)$, $\mathbf{C}_d(B(H))$; $E^{\sigma *}$, then, is row space over $B(H)$, i.e., all $d$-tuples of operators in $B(H)$, $(Z_1, Z_2, \dots, Z_d)$, arranged in a row. This space is denoted $\mathbf{R}_d(B(H))$. Although $\mathbf{C}_d(B(H))$ and $\mathbf{R}_d(B(H))$ are isometrically isomorphic as Banach spaces, they are very different as operator spaces (See, e.g., [20].) The unit ball $\mathbb{D}(\mathbf{R}_d(B(H)))$ is the space of all (strict) row contractions, i.e., all $d$-tuples $\mathbf{Z}^* := (Z_1, Z_2, \dots, Z_d)$ such that $\| \sum_{i=1}^d Z_i Z_i^* \| < 1$. If $F \in \mathcal{L}_d$, then $F$ has a series expansion in creation operators. Specifically, let $\{e_i\}_{i=1}^d$ be an orthonormal basis for $\mathbb{C}^d$and let $S_i$ denote the creation operator $T_{e_i}$. If $w = i_1 i_2 \cdots i_k$ is a word in the free semigroup on $\{1, 2, \dots, d\}$, then one writes $S_w$ for the operator $S_{i_1} S_{i_2} \cdots S_{i_k}$. As Davidson and Pitts show in [16, Section 1], every $F \in \mathcal{L}_d$ can be written as

$$F = \sum_{w \in \mathbb{F}_d^+} c_w S_w,$$

where the sum ranges over the free semigroup on $d$ letters, $\mathbb{F}_d^+$, and the series is Cesaro summable to $F$ in the ultraweak operator topology. When equation (7) is interpreted in this setting, one arrives at the formula:

$$\widehat{F}(\mathbf{Z}^*) = \sum_{w \in \mathbb{F}_d^+} c_w Z_w, \tag{8}$$

where $Z_w = Z_{i_1} Z_{i_2} \cdots Z_{i_k}$, if $w = i_1 i_2 \cdots i_k$, and where the series converges uniformly on balls of radius less than 1. Thus, $\widehat{F}$ is a bona fide analytic $B(H)$-valued function. But what kind of function is it? How does one recognize such a function? How does one categorize the space of all such functions as $F$ runs over $\mathcal{L}_d$? A little reflection reveals lots of natural, but difficult questions.

Already in the first nontrivial situation, when $d = 1$ and $\dim H = 2$, these questions are not easy. As I just indicated, in this case, if $T$ is a $2 \times 2$ matrix, $\widehat{F}(T) = \sum a_k T^k$ – something that is very familiar. But one can think of $\widehat{F}$ in this way: The disc $\mathbb{D}(E^{\sigma *})$ in this setting is the collection of all $2 \times 2$ matrices of norm less than one – a classical domain in $\mathbb{C}^4$. So $\widehat{F}$ is really a $2 \times 2$ matrix of holomorphic functions in four variables. So, what distinguishes functions of the form $\widehat{F}$ from all the other holomorphic matrix-valued functions of four complex variables? In short:

When is a matrix of functions a function of matrices?

One solution for *polynomial* functions is inspired by work of Joe Taylor [68, p. 238 ff.] and may be expressed in this way: *Suppose $f$ is a polynomial mapping from $M_2(\mathbb{C})$ to $M_2(\mathbb{C})$, i.e., suppose $f$ is a $2 \times 2$ matrix of polynomials, each of which is a polynomial in 4 variables, viewed as a $2 \times 2$ matrix $z = \begin{bmatrix} z_1 & z_2 \\ z_3 & z_4 \end{bmatrix}$. Then there is a polynomial $p$ in one variable such that $f(z) = p(z)$ in the sense of the usual polynomial calculus of matrices if and only if $f(z)m = mf(w)$ for every triple of matrices $m$, $z$, and $w$ such that $zm = mw$.* The key to the proof is to fix $m$ and then to analyze the possibilities for $z$ and $w$ in terms of the Jordan canonical form of $m$.

Taylor was motivated to try to extend the functional calculus he had developed for commuting families of operators on a Banach space to the noncommutative realm. He realized that the polynomial matrix-valued functions that one encounters when viewing elements of $\mathbb{C}\langle X_1, X_2, \ldots, X_d\rangle$ as functions defined on the space of representations of $\mathbb{C}\langle X_1, X_2, \ldots, X_d\rangle$ can be organized in terms of what Dan Voiculescu has recently dubbed "fully matricial sets" and "fully matricial functions" [69, 70]. Voiculescu, in turn, came upon these notions through his work on free probability. Although the definitions may seem complicated, it is worthwhile to have a look at them.

**Definition 7.1.** *Let $G$ be a Banach space and for each $n$ let $\mathfrak{M}_n(G)$ denote the $n \times n$ matrices over $G$.*

1. *A* fully matricial $G$-set *is a sequence $\mathbf{\Omega} = \{\Omega_n\}_{n\geq 1}$ such that*
   (a) *$\Omega_n$ is a subset of $\mathfrak{M}_n(G)$ for each $n$.*
   (b) *$\Omega_{n+m} \cap (\mathfrak{M}_m(G) \oplus \mathfrak{M}_n(G)) = \Omega_m \oplus \Omega_n$, $m, n \geq 1$, where for $X \in \mathfrak{M}_n(G)$ and $Y \in \mathfrak{M}_m(G)$, one writes $X \oplus Y$ for the matrix $\begin{bmatrix} X & 0 \\ 0 & Y \end{bmatrix} \in \mathfrak{M}_{n+m}(G)$; $\mathfrak{M}_m(G) \oplus \mathfrak{M}_n(G)$ then denotes the set of all such $X \oplus Y$.*
   (c) *If $X \in \Omega_m$, $Y \in \Omega_n$ and if $S \in GL(m+n, \mathbb{C})$ is such that $Ad(S)(X \oplus Y) \in \Omega_{m+n}$, then there is an $S' \in GL(m, \mathbb{C})$ and an $S' \in GL(n, \mathbb{C})$ so that $Ad(S')(X) \in \Omega_m$ and $Ad(S'')(Y) \in \Omega_n$.*
2. *If $H$ is another Banach space, then a sequence $\mathbf{R} = \{R_n\}_{n\geq 1}$ of functions, with $R_n$ defined on $\Omega_n$, is called a* fully matricial $H$-valued function defined on $\mathbf{\Omega}$ *in case*
   (a) *$R_n$ maps $\Omega_n$ into $\mathfrak{M}_n(H)$;*
   (b) *if $X \oplus Y \in \Omega_m \oplus \Omega_n$, then $R_{m+n}(X \oplus Y) = R_m(X) \oplus R_n(Y)$; and*
   (c) *if $X \in \Omega_n$ and if $S$ is in a sufficiently small neighborhood of $I$ in $GL(n, \mathbb{C})$ so that $Ad(S) \otimes I_G(X)$ lies in $\Omega_n$, then $R_n(Ad(S) \otimes I_G(X)) = Ad(S) \otimes I_H(R_n(X))$.*

   $\mathbf{R}$ *is called* continuous, analytic, *etc. iff each $R_n$ is continuous, analytic, etc.*

Clearly, these notions connect naturally to operator space theory. To see their relevance for our discussion, consider first the case of the free algebra on $d$ generators, $\mathbb{C}\langle X_1, X_2, \ldots, X_d\rangle$. The $n$-dimensional representations of $\mathbb{C}\langle X_1, X_2, \ldots, X_d\rangle$ are completely determined by $d$-tuples of $n \times n$ matrices and every $d$-tuple gives a representation. So, if one makes the following choices: $G := \mathbb{C}^d$, $\Omega_n := \mathfrak{M}_n(G)$, and $H := \mathbb{C}$, then one finds that $\mathbf{\Omega} := \{\Omega_n\}_{n\geq 1}$ is a fully matricial $\mathbb{C}^d$-set – by default – and that an element $f \in \mathbb{C}\langle X_1, X_2, \ldots, X_d\rangle$ determines a fully matricial $\mathbb{C}$-valued function $\{R_n\}_{n\geq 1}$ on $\mathbf{\Omega}$ through the formula $R_n(T_1, T_2, \ldots, T_d) := f(T_1, T_2, \ldots, T_d)$. Conversely, it can be shown, using arguments from [68, Section 6], that if $\{R_n\}_{n\geq 1}$ is a fully matricial $\mathbb{C}$-valued function on $\Omega$ consisting of polynomial functions in the $d \cdot n^2$ variables that parameterize $\mathfrak{M}_n(G)$, then there is an $f$ in the free algebra $\mathbb{C}\langle X_1, X_2, \ldots, X_d\rangle$ such that $R_n(T_1, T_2, \ldots, T_d) = f(T_1, T_2, \ldots, T_d)$ for all $n$ and for all $d$-tuples of $n \times n$ matrices. Thus, in this case, one finds that the notions of fully matricial sets and functions provide a scaffolding upon which one can organize the (finite-dimensional) representation theory of $\mathbb{C}\langle X_1, X_2, \ldots, X_d\rangle$.

Returning to the discussion of tensor algebras and Hardy algebras, suppose $A$ is a $W^*$-algebra, suppose $E$ is a $W^*$-correspondence over $A$, and that $\sigma : A \to B(H)$ is a faithful normal representation. For each positive integer $n$ let $n\sigma$ denote the $n$-fold multiple of $\sigma$ acting on the direct sum of $n$ copies of $H$. Then a moment's thought directed toward the fact that the commutant of $n\sigma(A)$ is $\mathfrak{M}_n(\sigma(A)')$ reveals that the sequence $\{\mathbb{D}(E^{n\sigma *})\}_{n\geq 1}$ is an example of a fully matricial $E^{\sigma *}$-set. Further, if $F \in H^\infty(E)$ and if $\widehat{F_n}$ is the function on $\mathbb{D}(E^{n\sigma *})$ defined by equation (7), then the sequence $\{\widehat{F_n}\}_{n\geq 1}$ is a fully matricial $B(H)$-valued analytic function on $\{\mathbb{D}(E^{n\sigma *})\}_{n\geq 1}$. Although one is used to thinking of infinite-dimensional structures, when one thinks of $W^*$-algebras and $W^*$-correspondences, the finite-dimensional theory, i.e., the theory built from finite graphs or quivers, is very rich and yields fully matricial sets and functions that generalize those from the free algebra setting. Thus, it is clear that the theory I have been discussing fits nicely into this "fully matricial function theory". But how nicely? How can one identify explicitly the fully matricial $B(H)$-valued functions on $\{\mathbb{D}(E^{n\sigma *})\}_{n\geq 1}$ that come from elements $F$ in $H^\infty(E)$?

One can build analytic fully matricial $B(H)$-valued functions on $\{\mathbb{D}(E^{n\sigma *})\}_{n\geq 1}$ that don't come from elements in $H^\infty(E)$, but as of this writing, the only ones I know of all come from formal "power series" over $E$. By a formal power series, I simply mean an infinite sum $f \sim \sum_{n\geq 0} f_n$, where each $f_n$ lies in the $n$-fold tensor power of $E$, $E^{\otimes n}$. One then has the following assertion, extending [59, Theorem 1.1] in the case when $E = \mathbb{C}^d$, that is familiar from basic complex analysis: *Let $R$ satisfy $\frac{1}{R} = \overline{\lim}_n \|f_n\|^{\frac{1}{n}}$. Then for every normal representation $\sigma : A \to B(H_\sigma)$, and every $\zeta^* \in E^{\sigma *}$, with $\|\zeta^*\| < R$, the series $f(\zeta^*) = \sum_n f_n(\zeta^*)$ converges in the norm of $B(H_\sigma)$ where $f_n(\zeta^*) := \zeta^{*(n)}(f_n)$ and $\zeta^{*(n)}$ is defined by the formula $\zeta^{*(n)}(\xi_1 \otimes \xi_2 \otimes \cdots \otimes \xi_n)(h) := \zeta^*(\xi_1 \otimes \zeta^*(\xi_2 \otimes \cdots \otimes \zeta^*(\xi_n \otimes h)) \cdots)$, which in turn, is a special case of equation* (7)*. The convergence is uniform on balls of strictly smaller* *radius and the resulting function, denoted by $f^\sigma$, is holomorphic as a Banach space-valued map defined on the open ball $\mathbb{D}_R(E^\sigma)^* := \{\zeta^* \in E^{\sigma *} \mid \|\zeta^*\| < R\}$ mapping to $B(H_\sigma)$. Moreover, $\{f^{n\sigma}\}_{n\geq 1}$ is a fully matricial $B(H_\sigma)$-valued function on the fully matricial $E^{\sigma *}$-set, $\{\mathbb{D}_R(E^{n\sigma})^*\}_{n\geq 1}$.* Results of Arveson from [4], show that one can (almost) always construct a formal power series $f$ and a representation $\sigma$ such that $f^\sigma$ is bounded and analytic on $\mathbb{D}(E^\sigma)^*$ that is not of the form $\widehat{F}$ for any $F \in H^\infty(E)$.

Dmitry Kalyuzhnyĭ-Verbovetzkiĭ and Victor Vinnikov have been studying the issue in the case when $E = \mathbb{C}^d$ and have gotten a fairly precise characterization of fully matricial $B(H)$-valued analytic functions on $\{\mathbb{D}(E^{n\sigma *})\}_{n\geq 1}$. What is impressive is that they require very little *a priori* regularity on the functions. Indeed, Victor has told me that he views the definition of a fully matricial function as a non-commutative generalization of the Cauchy-Riemann equations. This is because the definition of a fully matricial function leads naturally to certain derivations and related "differential equations" (see Section 6 of Taylor's paper [68]). One can do similar things in the setting of $W^*$-correspondences more general than $\mathbb{C}^d$, but at this point, I do not know sharp results.

In another direction, fully matricial sets and fully matricial functions on them have recently arisen in the theory of matrix inequalities. Indeed, so-called dimension-free matrix inequalities are just another name for certain fully matricial sets. These, and the functions on them, have been studied by Bill Helton and his collaborators. It is impossible to do justice to this subject here or even to cite all the relevant literature. However, I want to call special attention to [29] in which connections with the theory of fully matricial sets and functions are explicitly made. Also, although he does not express himself in terms of these objects, Popescu's work in [58, 60] and elsewhere leads naturally to fully matricial domains and functions.

The contribution to the problem of characterizing the functions $\widehat{F}$, $F \in H^\infty(E)$, that Baruch and I made was recently published in [46]. It is a generalization of the Nevanlinna-Pick interpolation theorem and so should seem familiar to operator theorists with a function theoretic bent. A central role is played by the notion of a completely positive definite kernel that was introduced by Stephen Barreto, Raja Bhat, Volkmar Liebscher and Michael Skeide in [8].

Let $\Omega$ be a set and let $A$ and $B$ be $C^*$-algebras. A function $K$ on $\Omega \times \Omega$ with values in the continuous linear transformations from $A$ to $B$, $\mathcal{B}(A, B)$, is called *completely positive definite kernel on* $\Omega$ in case for each finite set of distinct points $\{\omega_1, \omega_2, \ldots, \omega_n\}$ in $\Omega$, for each choice of $n$ elements in $A$, $a_1, a_2, \ldots, a_n$, and for each choice of $n$ elements in $B$, $b_1, b_2, \ldots, b_n$, the inequality

$$\sum_{i,j=1}^{n} b_i^* K(\omega_i, \omega_j)[a_i^* a_j] b_j \geq 0$$

holds in $B$. In [8, Theorem 3.2.3], the authors prove that a completely positive definite kernel in this sense gives rise to a $C^*$-correspondence from $A$ to $B$ that may be viewed as a family of $B$-valued functions on $\Omega$ and they prove an analogue of the Kolmogoroff decomposition theorem that is familiar from the theory of reproducing kernel Hilbert spaces: There is a $C^*$-correspondence $\mathcal{E}$ from $A$ to $B$ and a mapping $\omega \to k_\omega$ from $\Omega$ to $\mathcal{E}$ such that

$$K(\omega, \omega')[a] = \langle k_\omega, ak_{\omega'}\rangle$$

for all $a \in A$. Further, $E$ is uniquely determined by $K$ and the condition that the bimodule spanned by $\{k_\omega\}_{\omega\in\Omega}$ is dense in $E$.

In [46], Baruch and I let $\Omega$ be $\mathbb{D}(E^{\sigma *})$, where $E$ is a $W^*$-correspondence over the $W^*$-algebra $A$ and $\sigma$ is a faithful normal representation of $A$ on a Hilbert space $H$. There are two types of kernels with which we are concerned. The first is a generalized Szegö kernel $K_S$ on $\mathbb{D}(E^{\sigma *})$ defined by the formula

$$K_S(\zeta^*, \omega^*) = (id - \theta_{\zeta,\omega})^{-1}$$

where $id$ denotes the identity map on $\sigma(A)'$ and where $\theta_{\zeta,\omega}(a) = \langle \zeta, a\omega\rangle$, $a \in \sigma(A)'$. This kernel is completely positive definite with values in the operators *on* $\sigma(A)'$. However, it is convenient to extend the scalars to $B(H)$. When this is done, the Kolmogoroff-type decomposition of $K_S$ is afforded by the Cauchy-like kernel inside the "augmented" Fock space $B(H)\otimes_{\sigma(A)'}\mathcal{F}(E^\sigma)\otimes_{\sigma(A)'}B(H)$, given by the formula

$$k_{\zeta^*} = \mathbf{1} \oplus \zeta \oplus \zeta^{\otimes 2} \oplus \zeta^{\otimes 3} \oplus \cdots ,$$

with $\mathbf{1}$ denoting the identity operator in $\sigma(A)'$ and $\zeta^{\otimes n}$ denoting the $n^{th}$ tensor power of $\zeta$ as an element of $E^\sigma$. The reproducing kernel correspondence associated with $K_S$, then, is the closed $B(H)$-bimodule generated by $\{k_{\zeta^*} \mid \zeta^* \in \mathbb{D}(E^{\sigma *})\}$ inside $B(H) \otimes_{\sigma(A)'} \mathcal{F}(E^\sigma) \otimes_{\sigma(A)'} B(H)$.

The second kernel is determined by $K_S$ and any $B(H)$-valued function $F$ defined on $\mathbb{D}(E^{\sigma *})$ through the formula

$$K_F(\zeta^*, \omega^*) := (id - Ad(F(\zeta^*), F(\omega^*)) \circ K_S(\zeta^*, \omega^*),$$

where $id$ is the identity map on $B(H)$, and where for any pair of operators $A$ and $B$ in $B(H)$, $Ad(A, B)$ is the map on $B(H)$ defined by the formula $Ad(A, B)(T) = ATB^*$. The "identification theorem" that Baruch and I prove is the following generalization of our Nevanlinna-Pick theorem [44, Theorem 5.3].

**Theorem 7.2.** [46, Theorems 3.1, 3.3 and 3.6] *Let $A$ be a $W^*$-algebra, let $E$ be a $W^*$-correspondence over $A$, and let $\sigma : A \to B(H)$ be a faithful normal representation. Then a function $\Phi : \mathbb{D}(E^{\sigma *}) \to B(H)$ is of the form $\Phi(\zeta^*) = \widehat{F}(\zeta^*)$ for some $F \in H^\infty(E)$ of norm at most one if and only if the kernel $K_F$ on $\mathbb{D}(E^\sigma)^*$ is completely positive definite.*

This result, thus, enables one to view the $\widehat{F}$, $F \in H^\infty(E)$, as the collection of all multipliers of a reproducing kernel correspondence consisting of $B(H)$-valued functions defined on $\mathbb{D}(E^{\sigma *})$. But that begs the question: How to describe these spaces and multipliers? In the setting when $A = E = \mathbb{C}$, of course, the multipliers are isometrically isomorphic to $H^\infty(\mathbb{D})$. Arveson was the first to observe a difference when $A = \mathbb{C}$ and $E = \mathbb{C}^d$, $d \geq 2$. In this case, when $\sigma$ is the one-dimensional representation, the reproducing kernel Hilbert correspondence is the *symmetric* Fock space realized as what some now call the Drury-Arveson space of holomorphic functions on the $d$-dimensional ball, $\mathbb{B}^d$. Arveson showed that not every bounded analytic function comes from a multiplier [4]. Because of the evident analogy between Theorem 7.2 and the classical analysis of Schur class functions on the disc, we have come to call functions of the form $\widehat{F}$ *Schur-class functions* or *Schur-Agler class functions.* The book of Jim Agler and John McCarthy [1] has been an important source of inspiration for us and the proof of Theorem 7.2 owes a lot to the exposition there. I also want to call attention to the paper by Joe Ball, Animikh Biswas, Quanlei Fang, and Sanne ter Horst [7]. They have an approach to reproducing kernel Hilbert correspondences that is also based on the work of Barreto et al., but leads to a formulation and picture that are somewhat different from ours.

At the simplest level, Theorem 7.2 has this corollary which describes when an operator-valued function is defined through the holomorphic functional calculus: *Suppose $H$ is a finite- or infinite-dimensional Hilbert space and let $\varphi$ be a function from the open unit ball in $B(H)$, $B(H)_1$, to $B(H)$. Then there is a scalar-valued, bounded analytic function $f$ on the disc $\mathbb{D}$, of sup-norm at most* 1*, such that $\varphi(T) = f(T)$ for all $T \in B(H)_1$, where $f(T)$ is defined through the usual holomorphic functional calculus, if and only if the function $K_\varphi$ from $B(H)_1 \times B(H)_1$ to the bounded linear transformations on $B(H)$ defined by the formula*

$$\begin{aligned} T \mapsto K_\varphi(Z,W)(T) : &= (id - Ad(\varphi(Z), \varphi(W))) \circ (id - Ad(Z^*, W^*))^{-1}(T) \\ &= \sum_{n \geq 0} Z^{*n} T W^n - \varphi(Z) (\sum_{n \geq 0} Z^{*n} T W^n) \varphi(W)^* \end{aligned}$$

*is a completely positive definite kernel.*

One of the principal applications of Theorem 7.2 in [46] is to identify the automorphisms of $H^\infty(E)$ – at least when $A$ is a factor or, more generally, when $A$ has an atomic center [46, Theorem 4.22]. (There are additional technical matters that I omit here.) It turns out that they are induced by analogues of fractional linear – or Möbius transformations. One might expect this, since our discs have the character of bounded symmetric domains – albeit in usually infinite-dimensional Banach spaces. However, not every Möbius transformation need be allowed. I will omit a formal statement, but I bring it up here because the key step in [46] for understanding an automorphism of $H^\infty(E)$ was to express it in terms of Schur-class functions. There are points of contact between this result and a recent preprint of Bill Helton, Igor Klep, Scott McCullough and Nick Slinglend [29] that I mentioned above in the context of matrix inequalities.

While the definitions of fully matricial sets and functions, as well as the function theory that goes with these objects, are formulated in the setting of infinite-dimensional spaces, it is clear that matrix inequality theory provides a rich environment in which to continue to follow this variant of Halmos's doctrine:

If you want to study fully matricial function theory in infinite-dimensional spaces, then you must understand the finite-dimensional situation first. The problem is – and this is good news, really – the finite-dimensional theory already is full of difficult problems of its own, and the surface has barely been scratched.

## References

[1] J. Agler and J. McCarthy, *Pick Interpolation and Hilbert Function Spaces*, Graduate Studies in Mathematics, vol. 44, Amer. Math. Soc., Providence (2002).

[2] Wm.B. Arveson, *Operator algebras and measure preserving automorphisms*, Acta Mathematica 118 (1967), 95–109.

[3] Wm.B. Arveson, *Subalgebras of $C^*$-algebras,* Acta Mathematica 123 (1969), 141–224.

[4] Wm.B. Arveson, *Subalgebras of $C^*$-algebras, III,* Acta Mathematica 181 (1998), 159–228.

[5] Wm.B. Arveson, *Noncommutative Dynamics and E-semigroups,* Springer-Verlag, New York, 2003.

[6] M. Baillet, Y. Denizeau and J.-F. Havet, *Indice d'une esperance conditionelle*, Compositio Math. 66 (1988), 199–236.

[7] J. Ball, A. Biswas, Q. Fang and S. ter Horst, *Multivariable generalizations of the Schur class: positive kernel characterization and transfer function realization*, Operator Theory: Advances and Applications, 187 (2008), 17–79.

[8] S. Barreto, B. Bhat, V. Liebscher and M. Skeide, *Type I product systems of Hilbert modules,* J. Functional Anal 212 (2004), 121–181.

[9] A. Beurling, *On two problems concerning linear transformations on Hilbert space*, Acta Math. 81 (1949), 239–255.

[10] D. Blecher and C. Le Merdy, *Operator algebras and their modules – an operator space approach*, London Mathematical Society Monographs. New Series, 30, Oxford Science Publications. The Clarendon Press, Oxford University Press, Oxford, 2004. x+387 pp.

[11] D. Blecher, Z.-J. Ruan, and A. Sinclair, *A characterization of operator algebras*, J. Functional Anal. 89 (1990), 188–201.

[12] P.M. Cohn, *Free Rings and Their Relations*, Academic Press, New York, 1985.

[13] K. Davidson, J. Li and D. Pitts, *Absolutely continuous representations and a Kaplansky density theorem for free semigroup algebras*, J. Funct. Anal. 224 (2005), 160–191.

[14] K. Davidson and D. Pitts, *The algebraic structure of non-commutative analytic Toeplitz algebras,* Math. Ann. 311 (1998), 275–303.

[15] K. Davidson and D. Pitts, *Nevanlinna-Pick interpolation for non-commutative analytic Toeplitz algebras*, Integral Equations Operator Theory 31 (1998), 321–337.

[16] K. Davidson and D. Pitts, *Invariant subspaces and hyper-reflexivity for free semigroup algebras,* Proc. London Math. Soc. (3) 78 (1999), 401–430.

[17] K. Davidson and D. Yang, *A note on absolute continuity in free semigroup algebras*, Houston J. Math. 34 (2008), 283–288.

[18] J. Dixmier, *$C^*$-algebras*, North Holland, 1977.

[19] R.G. Douglas and V. Paulsen, *Hilbert Modules Over Function Algebras*, Pitman Research Notes in Mathematics Series 217 John Wiley, 1989.

[20] E.G. Effros and Z.-J. Ruan, *Operator spaces*, London Mathematical Society Monographs. New Series, 23. The Clarendon Press, Oxford University Press, New York, 2000.

[21] N. Fowler, P. Muhly, and I. Raeburn, *Representations of Cuntz-Pimsner algebras,* Indiana U. Math. J., 52 (2002), 569–605.

[22] N. Fowler and I. Raeburn, *The Toeplitz algebra of a Hilbert bimodule*, Indiana U. Math. J., 48 (1999), 155–181.

[23] P. Gabriel, *Unzerlegbare Darstellungen I*, Manuscripta Math. 6 (1972), 71–103.

[24] P. Gabriel, *Representations of Finite-Dimensional Algebras,* Encyclopaedia of Mathematical Sciences, Vol. 73, Springer-Verlag, New York, 1992.

[25] P.R. Halmos, *Shifts on Hilbert spaces*, J. Reine Angew. Math. 208 (1961), 102–112.

[26] P.R. Halmos, *A Hilbert Space Problem Book*, D. Van Nostrand Col., Inc., Princeton, N.J. – Toronto, Ont.-London, 1967, xvii+365 (Second edition, Graduate Texts in Mathematics 19, Springer-Verlag, New York, 1982, xvii+369.)

[27] P.R. Halmos, *How to write mathematics*, Enseignement Math. (2) 16 (1970), 123–152.

[28] M. Hamana, *Injective envelopes of operator systems*, Publ. Res. Inst. Math. Sci. 15 (1979), 773–785.

[29] J.W. Helton, I. Klep, S. McCullough, and N. Slinglend, *Noncommutative ball maps*, J. Funct. Anal. 257 (2009), 47–87.

[30] G. Hochschild, *On the structure of algebras with nonzero radical*, Bull. Amer. Math. Soc. 53(1947), 369–377.

[31] R.V. Kadison and I. Singer, *Triangular operator algebras. Fundamentals and hyperreducible theory*, Amer. J. Math. 82 (1960), 227–259.

[32] T. Katsura, *On $C^*$-algebras associated with $C^*$-correspondences,* J. Funct. Anal. 217 (2004), 366–401.

[33] E. Katsoulis and D. Kribs, *Tensor algebras of $C^*$-correspondences and their $C^*$-envelopes,* J. Funct. Anal. 234 (2006), 226–233.

[34] E.C. Lance, *Hilbert $C^*$-modules,* London Math. Soc. Lect. Note Series 210, Cambridge Univ. Press, Cambridge, 1995.

[35] P. Lax, *Translation invariant spaces*, Acta Math. 101 (1959), 163–178.

[36] A. Marrero and P. Muhly, *Cuntz-Pimnser algebras, completely positive maps and Morita equivalence* Proc. Amer. Math. Soc. 134 (2006), 1133–1135.

[37] M. McAsey, P. Muhly and K.-S. Saito, *Nonselfadjoint crossed products* (*invariant subspaces and maximality*), Trans. Amer. Math. Soc. 248 (1979), 381–409.

[38] M. McAsey, P. Muhly and K.-S. Saito, *Nonselfadjoint crossed products. II,* J. Math. Soc. Japan 33 (1981), 485–495.

[39] P. Muhly, *A finite-dimensional introduction to operator algebra,* in *Operator Algebras and Applications,* A. Katavolos, ed., NATO ASI Series Vol. 495, Kluwer, Dordrecht, 1997, pp. 313–354.

[40] P. Muhly and B. Solel, *Tensor algebras over $C^*$-correspondences* (*Representations, dilations, and $C^*$-envelopes*), J. Functional Anal. 158 (1998), 389–457.

[41] P. Muhly and B. Solel, *Tensor algebras, induced representations, and the Wold decomposition,* Canadian J. Math. 51 (1999), 850–880.

[42] P. Muhly and B. Solel, *Quantum Markov Processes* (*Correspondences and Dilations*)*,* Int. J. Math. 13 (2002), 863–906.

[43] P. Muhly and B. Solel, *The curvature and index of completely positive maps,* Proc. London Math. Soc. (3) 87 (2003), 748–778.

[44] P. Muhly and B. Solel, *Hardy algebras, $W^*$-correspondences and interpolation theory,* Math. Annalen 330 (2004), 353–415.

[45] P. Muhly and B. Solel, *Extensions and Dilations for $C^*$-dynamical systems*, in *Operator Theory, Operator Algebras and Applications*, Contemporary Mathematics 414, Amer. Math. Soc. 2006, 375–382.

[46] P. Muhly and B. Solel, *Schur Class Operator Functions and Automorphisms of Hardy Algebras*, Documenta Math. 13 (2008), 365–411.

[47] F. Murray and J. von Neumann, *Rings of operators*, Ann. of Math. (2) 37 (1936), 116–229.

[48] C. Nesbitt and W. Scott, *Matrix algebras over algebraically closed fields,* Ann. Math. (2) 44 (1943), 147–160.

[49] W. Paschke, *Inner product modules over $C^*$-algebras*, Trans. Amer. Math. Soc. 182 (1973), 443–468.

[50] V. Paulsen, *Completely bounded maps and operator algebras*, Cambridge Studies in Advanced Mathematics, 78, Cambridge University Press, Cambridge, 2002. xii + 300 pp.

[51] J. Peters, *Semi-crossed products of $C^*$-algebras*, J. Funct. Anal. 59 (1984), 498–534.

[52] M. Pimsner, *A class of $C^*$-algebras generalizing both Cuntz-Krieger algebras and crossed products by $\mathbb{Z}$*, in Free Probability Theory, D. Voiculescu, Ed., Fields Institute Communications 12, 189–212, Amer. Math. Soc., Providence, 1997.

[53] G. Pisier, *Introduction to operator space theory*, London Mathematical Society Lecture Note Series, 294, Cambridge University Press, Cambridge, 2003. viii+478 pp.

[54] G. Popescu, *Isometric dilations for infinite sequences of noncommuting operators,* Trans. Amer. Math. Soc. 316 (1989), 523–536.

[55] G. Popescu, *Characteristic functions for infinite sequences of noncommuting operators,* J. Operator Theory 22 (1989), 51–71.

[56] G. Popescu, *von Neumann inequality for* $(B(H)^n)_1$, Math. Scand. 68 (1991), 292–304.

[57] G. Popescu, *Multi-analytic operators on Fock spaces*, Math. Ann. 303 (1995), 31–46.

[58] G. Popescu, *Operator theory on noncommutative varieties*, Indiana Univ. Math. J. 55, 2006, 389–442.

[59] G. Popescu, *Free holomorphic functions on the unit ball of* $B(H)^n$, J. Funct. Anal. 241 (2006), 268–333.

[60] G. Popescu, *Operator theory on noncommutative domains*, preprint (arXiv:math/0703062).

[61] M. Rieffel, *Induced representations of $C^*$-algebras,* Advances in Math. 13 (1974), 176–257.

[62] S. Sakai, *A characterization of $W^*$-algebras*, Pacific J. Math. 6 (1956), 763–773.

[63] J.J. Schäffer, *On unitary dilations of contractions*, Proc. Amer. Math. Soc. 6 (1955), 322.

[64] W.F. Stinespring, *Positive functions on $C^*$-algebras*, Proc. Amer. Math. Soc. 6 (1955), 211–216.

[65] B. Sz.-Nagy, *Sur les contractions de l'espace de Hilbert*, Acta Sci. Math. 15 (1953), 87–92.

[66] B. Sz.-Nagy and C. Foiaş, Sur les contractions de l'espace de Hilbert. V. Translations bilatérales. Acta Sci. Math. (Szeged) 23 (1962), 106–129.

[67] B. Sz.-Nagy and C. Foiaş, *Harmonic Analysis of Operators on Hilbert Space*, Elsevier, 1970.

[68] J.L. Taylor, *A general framework for a multi-operator functional calculus*, Advances in Math. 9 (1972), 183–252.

[69] D. Voiculescu, *Free Analysis Questions I: Duality Transform for the Coalgebra of* $\partial_{X:B}$, Int. Math. Res. Not. 2004, no. 16, 793–822.

[70] D. Voiculescu, *Free Analysis Questions II: The Grassmannian completion and the series expansions at the origin*, preprint arXiv: 0806.0361.

Paul S. Muhly
Department of Mathematics
University of Iowa
Iowa City, IA 52242, USA
e-mail: pmuhly@math.uiowa.edu

Operator Theory:
Advances and Applications, Vol. 207, 287–324

# The Behavior of Functions of Operators Under Perturbations

V.V. Peller

*To the memory of Paul Halmos*

**Abstract.** This is a survey article. We consider different problems in connection with the behavior of functions of operators under perturbations of operators. We deal with three classes of operators: unitary operators, self-adjoint operators, and contractions. We study operator Lipschitz and operator differentiable functions. We also study the behavior of functions under perturbations of an operator by an operator of Schatten–von Neumann class $\boldsymbol{S}_p$ and apply the results to the Livschits–Krein and Koplienko–Neidhardt trace formulae. We also include in this survey article recent unexpected results obtained in a joint paper with Aleksandrov on operator Hölder–Zygmund functions.

**Mathematics Subject Classification (2000).** 47A55, 47B10, 47B35, 47B15, 46E15.

**Keywords.** Perturbation theory, self-adjoint operator, unitary operator, contraction, Lipschitz class, operator Lipschitz functions, Hölder classes, Zygmund class, Besov classes, Hankel operators, double operator integrals, multiple operator integrals.

## 1. Introduction

This survey article is devoted to problems in perturbation theory that arise in an attempt to understand the behavior of the function $f(A)$ of an operator $A$ under perturbations of $A$.

Consider the following example of such problems. Suppose that $\varphi$ is a function on the real line $\mathbb{R}$, $A$ is a self-adjoint operator on Hilbert space. The spectral theorem for self-adjoint operators allows us to define the function $\varphi(A)$ of $A$.

The author is partially supported by NSF grant DMS 0700995.

Suppose that $K$ is a bounded self-adjoint operator. We can ask the question of when the function

$$K \mapsto \varphi(A+K) - \varphi(A) \tag{1.1}$$

is differentiable. We can consider differentiability in the sense of Gâteaux or in the sense of Fréchet and we can consider the problem for bounded self-adjoint operators $A$ or for arbitrary self-adjoint operators (i.e., not necessarily bounded).

It is obvious that for this map to be differentiable (in the sense of Fréchet) it is necessary that $\varphi$ is a differentiable function on $\mathbb{R}$. Functions, for which the map (1.1) is differentiable are called *operator differentiable.* This term needs a clarification: we can consider operator differentiable functions in the sense of Gâteaux or Fréchet and we can consider this property for bounded $A$ or arbitrary self-adjoint operators $A$. In [W] Widom asked the question: when are differentiable functions differentiable?

We also consider in this survey the problem of the existence of higher operator derivatives.

Another example of problems of perturbation theory we are going to consider in this survey is the problem to describe *operator Lipschitz functions*, i.e., functions $\varphi$ on $\mathbb{R}$, for which

$$\|\varphi(A) - \varphi(B)\| \le \text{const}\,\|A - B\| \tag{1.2}$$

for self-adjoint operators $A$ and $B$. Sometimes such functions are called uniformly operator Lipschitz. Here $A$ and $B$ are allowed to be unbounded provided the difference $A - B$ is bounded. If $\varphi$ is a function, for which (1.2) holds for bounded operators $A$ and $B$ with a constant that can depend on $\|A\|$ and $\|B\|$, then $\varphi$ is called *locally operator Lipschitz.* It is easy to see that if $\varphi$ is operator Lipschitz, then $\varphi$ must be a Lipschitz function, i.e.,

$$|\varphi(x) - \varphi(y)| \le \text{const}\,|x - y|, \quad x,\, y \in \mathbb{R}, \tag{1.3}$$

and if $\varphi$ is locally operator Lipschitz, then $\varphi$ is locally a Lipschitz function, i.e., (1.3) must hold on each bounded subset of $\mathbb{R}$.

We also consider in this survey the problem for which functions $\varphi$

$$\|\varphi(A) - \varphi(B)\| \le \text{const}\,\|A - B\|^{\alpha} \tag{1.4}$$

for self-adjoint operators $A$ and $B$. Here $0 < \alpha < 1$. If $\varphi$ satisfies (1.4), it is called *an operator Hölder function of order* $\alpha$. Again, it is obvious that for $\varphi$ to be operator Hölder of order $\alpha$ it is necessary that $\varphi$ belongs to the Hölder class $\Lambda_\alpha(\mathbb{R})$, i.e.,

$$|\varphi(x) - \varphi(y)| \le \text{const}\,|x - y|^{\alpha}. \tag{1.5}$$

We also consider functions $f$ on $\mathbb{R}$ for which

$$\|\varphi(A - K) - 2\varphi(A) + \varphi(A + K)\| \le \text{const}\,\|K\| \tag{1.6}$$

for selfadjoint operators $A$ and $K$. Functions $\varphi$ satisfying (1.6) are called *operator Zygmund functions.*

In this paper we also study the whole scale of operator Hölder–Zygmund classes.

Another group of problems we are going to consider is the behavior of functions of operators under perturbations of trace class (or other classes of operators). In particular, the problem to describe the class of functions $f$, for which

$$f(A+K)-f(A)\in \boldsymbol{S}_1 \quad \text{whenever} \quad K\in \boldsymbol{S}_1,$$

is very important in connection with the Lifshits–Krein trace formula. We use the notation $\boldsymbol{S}_p$ for Schatten–von Neumann classes.

We also consider problems of perturbation theory related to the Koplienko trace formula, which deals with Hilbert–Schmidt perturbations.

It is also important to study similar problems for unitary operators and functions on the unit circle $\mathbb{T}$ and for contractions and analytic functions in the unit disk $\mathbb{D}$.

The study of the problem of differentiability of functions of self-adjoint operators on Hilbert space was initiated By Daletskii and S.G. Krein in [DK]. They showed that for a function $\varphi$ on the real line $\mathbb{R}$ of class $C^2$ and for bounded self-adjoint operators $A$ and $K$ the function

$$t\mapsto \varphi(A+tK) \tag{1.7}$$

is differentiable in the operator norm and the derivative can be computed in terms of double operator integrals:

$$\frac{d}{dt}\varphi(A+tK)\Big|_{t=0}=\iint\limits_{\mathbb{R}\times\mathbb{R}}\frac{\varphi(x)-\varphi(y)}{x-y}\,dE_A(x)\,K\,dE_A(y), \tag{1.8}$$

where $E_A$ is the spectral measure of $A$. The expression on the right is a double operator integral. The beautiful theory of double operator integrals due to Birman and Solomyak was created later in [BS1], [BS2], and [BS4] (see also the survey article [BS6]). A brief introduction into the theory of double operator integrals will be given in § 2.

The condition $\varphi\in C^2$ was relaxed by Birman and Solomyak in [BS4]: they proved that the function (1.7) is differentiable and the Daletskii–Krein formula (1.8) holds under the condition that $\varphi$ is differentiable and the derivative $\varphi'$ satisfies a Hölder condition of order $\alpha$ for some $\alpha>0$. The approach of Birman and Solomyak is based on their formula

$$\varphi(A+K)-\varphi(A)=\iint\limits_{\mathbb{R}\times\mathbb{R}}\frac{\varphi(x)-\varphi(y)}{x-y}\,dE_{A+K}(x)\,B\,dE_A(y). \tag{1.9}$$

Actually, Birman and Solomyak showed in [BS4] that formula (1.9) is valid under the condition that the divided difference $\mathfrak{D}\varphi$,

$$(\mathfrak{D}\varphi)(x,y)=\frac{\varphi(x)-\varphi(y)}{x-y},$$

is a Schur multiplier of the space of all bounded linear operators (see § 2 for definitions).

Nevertheless, Farforovskaya proved in [Fa1] that the condition $\varphi \in C^1$ does not imply that $\varphi$ is operator Lipschitz, which implies that the condition $\varphi \in C^1$ is not sufficient for the differentiability of the map (1.7) (see also [Fa3] and [KA]).

A further improvement was obtained in [Pe2] and [Pe4]: it was shown that the function (1.7) is differentiable and (1.8) holds under the assumption that $\varphi$ belongs to the Besov space $B^1_{\infty 1}(\mathbb{R})$ (see § 4) and under the same assumption $\varphi$ must be uniformly operator Lipschitz. In the same paper [Pe2] a necessary condition was found: $\varphi$ must locally belong to the Besov space $B^1_1(\mathbb{R}) = B^1_{11}(\mathbb{R})$. This necessary condition also implies that the condition $\varphi \in C^1$ is not sufficient. Actually, in [Pe2] and [Pe4] a stronger necessary condition was also obtained; see §7 for further discussions. Finally, we mention another sufficient condition obtained in [ABF] which is slightly better than the condition $\varphi \in B^1_{\infty 1}(\mathbb{R})$, though I believe it is more convenient to work with Besov spaces. We refer the reader to Sections 6 and 7 of this survey for a detailed discussion.

After it had become clear that Lipschitz functions do not have to be operator Lipschitz, many mathematicians believed that Hölder functions of order $\alpha$, $0 < \alpha < 1$, are not necessarily operator Hölder functions of order $\alpha$. In [Fa1] the following upper estimate for self-adjoint operators $A$ and $B$ with spectra in an interval $[a, b]$ was obtained:

$$\|\varphi(A) - \varphi(B)\| \le \text{const}\, \|\varphi\|_{\Lambda_\alpha(\mathbb{R})} \left(\log\left(\frac{b-a}{\|A-B\|} + 1\right) + 1\right)^2 \|A - B\|^\alpha,$$

where $\varphi \in \Lambda_\alpha(\mathbb{R})$. A similar inequality was obtained in [FN] for arbitrary moduli of continuity.

*Surprisingly, it turns out that the logarithmic factor in the above inequality is unnecessary. In other words, for an arbitrary* $\alpha \in (0, 1)$, *a Hölder function of order* $\alpha$ *must be operator Hölder of order* $\alpha$. *Moreover, the same is true for Zygmund functions and for the whole scale of Hölder–Zygmund classes.* This has been proved recently in [AP2], see also [AP1]. We discuss the results of [AP2] in § 10.

The problem of the existence of higher-order derivatives of the function (1.7) was studied in [St] where it was shown that under certain assumptions on $\varphi$, the function (1.7) has a second derivative that can be expressed in terms of the following triple operator integral:

$$\frac{d^2}{dt^2}\varphi(A + tB)\Big|_{t=0} = 2 \iiint\limits_{\mathbb{R}\times\mathbb{R}\times\mathbb{R}} \left(\mathfrak{D}^2\varphi\right)(x, y, z)\, dE_A(x)\, B\, dE_A(y)\, B\, dE_A(z),$$

where $\mathfrak{D}^2\varphi$ stands for the divided difference of order 2 (see § 8 for the definition). To interpret triple operator integrals, repeated integration was used in [St] (see also the earlier paper [Pa], in which an attempt to define multiple operator integrals was given). However, the class of integrable functions in [Pa] and [St] was rather narrow and the assumption on $\varphi$ imposed in [St] for the existence of the second operator derivative was too restrictive. Similar results were also obtained in [St] for the $n$th derivative and multiple operator integrals.

In [Pe8] a new approach to multiple operator integrals was given. It is based on integral projective tensor products of $L^\infty$ spaces and gives a much broader class of integrable functions than under the approaches of [Pa] and [St]. It was shown in [Pe8] that under the assumption that $f$ belongs to the Besov space $B^n_{\infty 1}(\mathbb{R})$ the function (1.7) has $n$ derivatives and the $n$th derivative can be expressed in terms of a multiple operator integral. Similar results were also obtained in [Pe8] in the case of an unbounded self-adjoint operator $A$.

To study the problem of differentiability of functions of unitary operators, we should consider a Borel function $f$ on the unit circle $\mathbb{T}$ and the map

$$U \mapsto f(U),$$

where $U$ is a unitary operator on Hilbert space. If $U$ and $V$ are unitary operators and $V = e^{\mathrm{i}A}U$, where $A$ is a self-adjoint operator, we can consider the one-parameter family of unitary operators

$$e^{\mathrm{i}tA}U, \quad 0 \le t \le 1,$$

and study the question of the differentiability of the function

$$t \mapsto \big(e^{\mathrm{i}tA}U\big)$$

and the question of the existence of its higher derivatives. The results in the case of unitary operators are similar to the results for self-adjoint operators, see [BS4], [Pe2], [ABF], [Pe8].

Similar questions can be also considered for functions of contractions. It turns out that to study such problems for contractions, one can use double and multiple operator integrals with respect to semi-spectral measures. This approach was proposed in [Pe3] and [Pe9]. We discuss these issues in § 9.

In Sections 2 and 3 of this survey we give a brief introduction in double operator integrals and multiple operator integrals. In §4 we introduce the reader to Besov spaces.

In §5 we define Hankel operators and state the nuclearity criterion obtained in [Pe1]. It turns out that Hankel operators play an important role in perturbation theory and the nuclearity criterion is used in § 7 to obtain necessary conditions for operator differentiability and operator Lipschitzness.

The last 3 sections are devoted to perturbations of operators by operators from Schatten–von Neumann classes $\boldsymbol{S}_p$. In §11 we discuss the problem of classifying the functions $\varphi$ for which the Lifshits–Krein trace formulae are valid. This problem is closely related to the problem of classifying the functions $\varphi$ on $\mathbb{R}$ for which a trace class perturbation of a self-adjoint operator $A$ leads to a trace class perturbation of $\varphi(A)$. We present in § 11 sufficient conditions obtained in [Pe2] and [Pe4] that improve earlier results by M.G. Krein and Birman–Solomyak. We also discuss necessary conditions.

Section 12 deals with perturbations of class $\boldsymbol{S}_2$ and trace formulae by Koplienko and Neidhardt. We discuss the results of [Pe7] that improve earlier results by Koplienko and Neidhardt.

Finally, in the last section we present recent results of [AP3] (see also [AP1]) that concern the following problem. Suppose that $A$ and $B$ are self-adjoint operators such that $A-B \in \boldsymbol{S}_p$ and let $\varphi \in \Lambda_\alpha(\mathbb{R})$. What can we say about $\varphi(A)-\varphi(B)$? We also discuss a similar problem for higher-order differences. In addition to this we state very recent results obtained in [NP] and [PS].

It was certainly impossible to give proofs of all the results discussed in this survey. I tried to give proofs of certain key results that demonstrate principal ideas.

## 2. Double operator integrals

In this section we give a brief introduction in the theory of double operator integrals developed by Birman and Solomyak in [BS1], [BS2], and [BS4], see also their survey [BS6].

Let $(\mathcal{X}, E_1)$ and $(\mathcal{Y}, E_2)$ be spaces with spectral measures $E_1$ and $E_2$ on a Hilbert spaces $\mathcal{H}_1$ and $\mathcal{H}_2$. Let us first define double operator integrals

$$\int\limits_{\mathcal{X}}\int\limits_{\mathcal{Y}} \Phi(x,y)\,dE_1(x)\,Q\,dE_2(y), \tag{2.1}$$

for bounded measurable functions $\Phi$ and operators $Q : \mathcal{H}_2 \to \mathcal{H}_1$ of Hilbert–Schmidt class $\boldsymbol{S}_2$. Consider the set function $F$ whose values are orthogonal projections on the Hilbert space $\boldsymbol{S}_2(\mathcal{H}_2, \mathcal{H}_1)$ of Hilbert–Schmidt operators from $\mathcal{H}_2$ to $\mathcal{H}_1$, which is defined on measurable rectangles by

$$F(\Delta_1 \times \Delta_2)Q = E_1(\Delta_1)QE_2(\Delta_2), \quad Q \in \boldsymbol{S}_2(\mathcal{H}_2, \mathcal{H}_1),$$

$\Delta_1$ and $\Delta_2$ being measurable subsets of $\mathcal{X}$ and $\mathcal{Y}$. Note that left multiplication by $E_1(\Delta_1)$ obviously commutes with right multiplication by $E_2(\Delta_2)$.

It was shown in [BS5] that $F$ extends to a spectral measure on $\mathcal{X} \times \mathcal{Y}$. If $\Phi$ is a bounded measurable function on $\mathcal{X} \times \mathcal{Y}$, we define

$$\int\limits_{\mathcal{X}}\int\limits_{\mathcal{Y}} \Phi(x,y)\,dE_1(x)\,Q\,dE_2(y) = \left(\ \int\limits_{\mathcal{X}_1\times\mathcal{X}_2} \Phi\,dF\right)Q.$$

Clearly,

$$\left\| \int\limits_{\mathcal{X}}\int\limits_{\mathcal{Y}} \Phi(x,y)\,dE_1(x)\,Q\,dE_2(y)\right\|_{\boldsymbol{S}_2} \le \|\Phi\|_{L^\infty}\|Q\|_{\boldsymbol{S}_2}.$$

If the transformer

$$Q \mapsto \int\limits_{\mathcal{X}}\int\limits_{\mathcal{Y}} \Phi(x,y)\,dE_1(x)\,Q\,dE_2(y)$$

maps the trace class $\boldsymbol{S}_1$ into itself, we say that $\Phi$ is a *Schur multiplier of $\boldsymbol{S}_1$ associated with the spectral measures $E_1$ and $E_2$*. In this case the transformer

$$Q \mapsto \int\limits_{\mathcal{Y}}\int\limits_{\mathcal{X}} \Phi(x,y)\,dE_2(y)\,Q\,dE_1(x), \quad Q \in \boldsymbol{S}_2(\mathcal{H}_1,\mathcal{H}_2), \tag{2.2}$$

extends by duality to a bounded linear transformer on the space of bounded linear operators from $\mathcal{H}_1$ to $\mathcal{H}_2$ and we say that the function $\Psi$ on $\mathcal{X}_2 \times \mathcal{X}_1$ defined by

$$\Psi(y,x) = \Phi(x,y)$$

is *a Schur multiplier of the space of bounded linear operators associated with $E_2$ and $E_1$*. We denote the space of such Schur multipliers by $\mathfrak{M}(E_2,E_1)$

Birman in Solomyak obtained in [BS4] the following result:

**Theorem 2.1.** *Let $A$ be a self-adjoint operator* (*not necessarily bounded*) *and let $K$ be a bounded self-adjoint operator. Suppose that $\varphi$ is a continuously differentiable function on $\mathbb{R}$ such that the divided difference $\mathfrak{D}\varphi \in \mathfrak{M}(E_{A+K},E_A)$. Then*

$$\varphi(A+K) - \varphi(A) = \iint\limits_{\mathbb{R}\times\mathbb{R}} \frac{\varphi(x)-\varphi(y)}{x-y}\,dE_{A+K}(x)K\,dE_A(y) \tag{2.3}$$

*and*

$$\|\varphi(A+K)-\varphi(A)\| \le \mathrm{const}\,\|\mathfrak{D}\varphi\|_{\mathfrak{M}}\|K\|,$$

*where $\|\mathfrak{D}\varphi\|_{\mathfrak{M}}$ is the norm of $\mathfrak{D}\varphi$ in $\mathfrak{M}(E_{A+K},E_A)$.*

In the case when $K$ belongs to the Hilbert Schmidt class $\boldsymbol{S}_2$, the same result was established in [BS4] under weaker assumptions on $\varphi$:

**Theorem 2.2.** *Let $A$ be a self-adjoint operator* (*not necessarily bounded*) *and let $K$ be a self-adjoint operator of class $\boldsymbol{S}_2$. If $\varphi$ is a Lipschitz function on $\mathbb{R}$, then* (2.3) *holds,*

$$\varphi(A+K) - \varphi(A) \in \boldsymbol{S}_2,$$

*and*

$$\|\varphi(A+K)-\varphi(A)\|_{\boldsymbol{S}_2} \le \sup_{x\ne y}\frac{|\varphi(x)-\varphi(y)|}{|x-y|}\,\|A-B\|_{\boldsymbol{S}_2}.$$

Note that if $\varphi$ is not differentiable, $\mathfrak{D}\varphi$ is not defined on the diagonal of $\mathbb{R}\times\mathbb{R}$, but formula (2.3) still holds if we define $\mathfrak{D}\varphi$ to be zero on the diagonal.

Similar results also hold for functions on the unit circle and for unitary operators.

It is easy to see that if a function $\Phi$ on $\mathcal{X}\times\mathcal{Y}$ belongs to the *projective tensor product* $L^\infty(E_1)\hat{\otimes}L^\infty(E_2)$ of $L^\infty(E_1)$ and $L^\infty(E_2)$ (i.e., $\Phi$ admits a representation

$$\Phi(x,y) = \sum_{n\ge 0} f_n(x)g_n(y), \tag{2.4}$$

where $f_n \in L^\infty(E_1)$, $g_n \in L^\infty(E_2)$, and

$$\sum_{n\ge 0}\|f_n\|_{L^\infty}\|g_n\|_{L^\infty} < \infty), \tag{2.5}$$

then $\Phi \in \mathfrak{M}(E_1, E_2)$, i.e., $\Phi$ is a Schur multiplier of the space of bounded linear operators. For such functions $\Phi$ we have

$$\int\limits_{\mathcal{X}}\int\limits_{\mathcal{Y}} \Phi(x,y)\,dE_1(x)Q\,dE_2(y) = \sum_{n\ge 0}\left(\int\limits_{\mathcal{X}} f_n\,dE_1\right)Q\left(\int\limits_{\mathcal{Y}} g_n\,dE_2\right).$$

Note that if $\Phi$ belongs to the projective tensor product $L^\infty(E_1)\hat{\otimes}L^\infty(E_2)$, its norm in $L^\infty(E_1)\hat{\otimes}L^\infty(E_2)$ is, by definition, the infimum of the left-hand side of (2.5) over all representations (2.4).

One can define in the same way projective tensor products of other function spaces.

More generally, $\Phi$ is a Schur multiplier if $\Phi$ belongs to the *integral projective tensor product* $L^\infty(E_1)\hat{\otimes}_{\mathrm{i}}L^\infty(E_2)$ of $L^\infty(E_1)$ and $L^\infty(E_2)$, i.e., $\Phi$ admits a representation

$$\Phi(x,y) = \int_\Omega f(x,\omega)g(y,\omega)\,d\sigma(\omega), \tag{2.6}$$

where $(\Omega,\sigma)$ is a measure space, $f$ is a measurable function on $\mathcal{X}\times\Omega$, $g$ is a measurable function on $\mathcal{Y}\times\Omega$, and

$$\int_\Omega \|f(\cdot,\omega)\|_{L^\infty(E_1)}\|g(\cdot,\omega)\|_{L^\infty(E_2)}\,d\sigma(\omega) < \infty. \tag{2.7}$$

If $\Phi \in L^\infty(E_1)\hat{\otimes}_{\mathrm{i}}L^\infty(E_2)$, then

$$\begin{aligned}\int\limits_{\mathcal{X}}\int\limits_{\mathcal{Y}} &\Phi(x,y)\,dE_1(x)\,Q\,dE_2(y)\\ &= \int\limits_\Omega\left(\int\limits_{\mathcal{X}} f(x,\omega)\,dE_1(x)\right)Q\left(\int\limits_{\mathcal{Y}} g(y,\omega)\,dE_2(y)\right)d\sigma(\omega).\end{aligned}$$

Clearly, the function

$$\omega \mapsto \left(\int\limits_{\mathcal{X}} f(x,\omega)\,dE_1(x)\right)Q\left(\int\limits_{\mathcal{Y}} g(y,\omega)\,dE_2(y)\right)$$

is weakly measurable and

$$\int\limits_\Omega\left\|\left(\int\limits_{\mathcal{X}} f(x,\omega)\,dE_1(x)\right)T\left(\int\limits_{\mathcal{Y}} g(y,\omega)\,dE_2(y)\right)\right\|d\sigma(\omega) < \infty.$$

Moreover, it can easily be seen that such functions $\Phi$ are Schur multipliers of an arbitrary Schatten–von Neumann class $\boldsymbol{S}_p$ with $p \ge 1$.

It turns out that all Schur multipliers of the space of bounded linear operators can be obtained in this way.

More precisely, the following result holds (see [Pe2]):

**Theorem 2.3.** *Let $\Phi$ be a measurable function on $\mathscr{X}\times\mathscr{Y}$. The following are equivalent:*

(i) $\Phi\in\mathfrak{M}(E_1,E_2)$;

(ii) $\Phi\in L^\infty(E_1)\hat{\otimes}_{\rm i}L^\infty(E_2)$;

(iii) *there exist measurable functions $f$ on $\mathscr{X}\times\Omega$ and $g$ on $\mathscr{Y}\times\Omega$ such that* (2.6) *holds and*

$$\left\|\int_\Omega|f(\cdot,\omega)|^2\,d\sigma(\omega)\right\|_{L^\infty(E_1)}\left\|\int_\Omega|g(\cdot,\omega)|^2\,d\sigma(\omega)\right\|_{L^\infty(E_2)}<\infty. \tag{2.8}$$

Note that the implication (iii)⇒(ii) was established in [BS4]. Note also that the equivalence of (i) and (ii) is deduced from Grothendieck's theorem. In the case of matrix Schur multipliers (this corresponds to discrete spectral measures of multiplicity 1), the equivalence of (i) and (ii) was proved in [Be].

It is interesting to observe that if $f$ and $g$ satisfy (2.7), then they also satisfy (2.8), but the converse is false. However, if $\Phi$ admits a representation of the form (2.6) with $f$ and $g$ satisfying (2.8), then it also admits a (possibly different) representation of the form (2.6) with $f$ and $g$ satisfying (2.7).

Let us also mention one more observation by Birman and Solomyak, see [BS4]. Suppose that $\mu$ and $\nu$ are scalar measures on $\mathscr{X}$ and $\mathscr{Y}$ that are mutually absolutely continuous with $E_1$ and $E_2$. Let $\mathfrak{N}_{\mu,\nu}$ be the class of measurable functions $k$ on $\mathscr{X}\times\mathscr{Y}$ such that the integral operator from $L^2(\mu)$ to $L^2(\nu)$ with kernel function $k$ belongs to the trace class $\boldsymbol{S}_1$.

**Theorem 2.4.** *A measurable function $\Phi$ on $\mathscr{X}\times\mathscr{Y}$ is a Schur multiplier of $\boldsymbol{S}_1$ associated with $E_1$ and $E_2$ if and only if $\Phi$ is a multiplier of $\mathfrak{N}_{\mu,\nu}$, i.e.,*

$$k\in\mathfrak{N}_{\mu,\nu}\quad\Rightarrow\quad\Phi k\in\mathfrak{N}_{\mu,\nu}.$$

## 3. Multiple operator integrals

The equivalence of (i) and (ii) in the Theorem 2.3 suggests the idea explored in [Pe8] to define multiple operator integrals.

To simplify the notation, we consider here the case of triple operator integrals; the case of arbitrary multiple operator integrals can be treated in the same way.

Let $(\mathscr{X},E_1)$, $(\mathscr{Y},E_2)$, and $(\mathscr{Z},E_3)$ be spaces with spectral measures $E_1$, $E_2$, and $E_3$ on Hilbert spaces $\mathcal{H}_1$, $\mathcal{H}_2$, and $\mathcal{H}_3$. Suppose that $\Phi$ belongs to the integral projective tensor product $L^\infty(E_1)\hat{\otimes}_{\rm i}L^\infty(E_2)\hat{\otimes}_{\rm i}L^\infty(E_3)$, i.e., $\Phi$ admits a representation

$$\Phi(x,y,z)=\int_\Omega f(x,\omega)g(y,\omega)h(z,\omega)\,d\sigma(\omega), \tag{3.1}$$

where $(\Omega,\sigma)$ is a measure space, $f$ is a measurable function on $\mathcal{X}\times\Omega$, $g$ is a measurable function on $\mathcal{Y}\times\Omega$, $h$ is a measurable function on $\mathcal{Z}\times\Omega$, and

$$\int_\Omega \|f(\cdot,\omega)\|_{L^\infty(E)}\|g(\cdot,\omega)\|_{L^\infty(F)}\|h(\cdot,\omega)\|_{L^\infty(G)}\,d\sigma(\omega)<\infty.$$

Suppose now that $T_1$ is a bounded linear operator from $\mathcal{H}_2$ to $\mathcal{H}_1$ and $T_2$ is a bounded linear operator from $\mathcal{H}_3$ to $\mathcal{H}_2$.

For a function $\Phi$ in $L^\infty(E_1)\hat{\otimes}_{\rm i}L^\infty(E_2)\hat{\otimes}_{\rm i}L^\infty(E_3)$ of the form (3.1), we put

$$\int\limits_{\mathcal{X}}\int\limits_{\mathcal{Y}}\int\limits_{\mathcal{Z}}\Phi(x,y,z)\,dE_1(x)T_1\,dE_2(y)T_2\,dE_3(z) \tag{3.2}$$
$$\stackrel{\rm def}{=}\int\limits_{\Omega}\Big(\int\limits_{\mathcal{X}}f(x,\omega)\,dE_1(x)\Big)T_1\Big(\int\limits_{\mathcal{Y}}g(y,\omega)\,dE_2(y)\Big)T_2\Big(\int\limits_{\mathcal{Z}}h(z,\omega)\,dE_3(z)\Big)\,d\sigma(\omega).$$

The following lemma from [Pe8] (see also [ACDS] for a different proof) shows that the definition does not depend on the choice of a representation (3.1).

**Lemma 3.1.** *Suppose that $\Phi\in L^\infty(E_1)\hat{\otimes}_{\rm i}L^\infty(E_2)\hat{\otimes}_{\rm i}L^\infty(E_3)$. Then the right-hand side of* (3.2) *does not depend on the choice of a representation* (3.1)*.*

It is easy to see that the following inequality holds

$$\left\|\int\limits_{\mathcal{X}}\int\limits_{\mathcal{Y}}\int\limits_{\mathcal{Z}}\Phi(x,y,z)\,dE_1(x)T_1\,dE_2(y)T_2\,dE_3(z)\right\|\le\|\Phi\|_{L^\infty\hat{\otimes}_{\rm i}L^\infty\hat{\otimes}_{\rm i}L^\infty}\cdot\|T_1\|\cdot\|T_2\|.$$

In particular, the triple operator integral on the left-hand side of (3.2) can be defined if $\Phi$ belongs to the projective tensor product $L^\infty(E_1)\hat{\otimes}L^\infty(E_2)\hat{\otimes}L^\infty(E_3)$, i.e., $\Phi$ admits a representation

$$\Phi(x,y,z)=\sum_{n\ge1}f_n(x)g_n(y)h_n(z),$$

where $f_n\in L^\infty(E_1)$, $g_n\in L^\infty(E_2)$, $h_n\in L^\infty(E_3)$ and

$$\sum_{n\ge1}\|f_n\|_{L^\infty(E_1)}\|g_n\|_{L^\infty(E_2)}\|h_n\|_{L^\infty(E_3)}<\infty.$$

In a similar way one can define multiple operator integrals, see [Pe8].

Recall that earlier multiple operator integrals were considered in [Pa] and [St]. However, in those papers the class of functions $\Phi$ for which the left-hand side of (3.2) was defined is much narrower than in the definition given above.

Multiple operator integrals arise in connection with the problem of evaluating higher-order operator derivatives. It turns out that if $A$ is a self-adjoint operator on Hilbert space and $K$ is a bounded self-adjoint operator, then for sufficiently nice functions $\varphi$ on $\mathbb{R}$, the function

$$t\mapsto\varphi(A+tK) \tag{3.3}$$

has $n$ derivatives in the norm and the $n$th derivative can be expressed in terms of multiple operator integrals. We are going to consider this problem in § 8.

Note that recently in [JTT] Haagerup tensor products were used to define multiple operator integrals. However, it is not clear whether this can lead to a broader class of functions $\varphi$, for which the function $f$ has an $n$th derivative and the $n$th derivative can be expressed in terms of a multiple operator integral.

## 4. Besov spaces

The purpose of this section is to give a brief introduction to the Besov spaces that play an important role in problems of perturbation theory. We start with Besov spaces on the unit circle.

Let $1 \le p,\ q \le \infty$ and $s \in \mathbb{R}$. The Besov class $B^s_{pq}$ of functions (or distributions) on $\mathbb{T}$ can be defined in the following way. Let $w$ be an infinitely differentiable function on $\mathbb{R}$ such that

$$w \ge 0, \quad \operatorname{supp} w \subset \left[\frac{1}{2}, 2\right], \quad \text{and} \quad w(x) = 1 - w\left(\frac{x}{2}\right) \quad \text{for} \quad x \in [1,2]. \tag{4.1}$$

and $w$ is a linear function on the intervals $[1/2, 1]$ and $[1,2]$.

Consider the trigonometric polynomials $W_n$, and $W_n^\#$ defined by

$$W_n(z) = \sum_{k\in\mathbb{Z}} w\left(\frac{k}{2^n}\right) z^k, \quad n \ge 1,$$

$$W_0(z) = \bar{z} + 1 + z, \quad \text{and} \quad W_n^\#(z) = \overline{W_n(z)}, \quad n \ge 0.$$

Then for each distribution $\varphi$ on $\mathbb{T}$,

$$\varphi = \sum_{n\ge0} \varphi * W_n + \sum_{n\ge1} \varphi * W_n^\#.$$

The Besov class $B^s_{pq}$ consists of functions (in the case $s > 0$) or distributions $\varphi$ on $\mathbb{T}$ such that

$$\left\{\|2^{ns}\varphi * W_n\|_{L^p}\right\}_{n\ge0} \in \ell^q \quad \text{and} \quad \left\{\|2^{ns}\varphi * W_n^\#\|_{L^p}\right\}_{n\ge1} \in \ell^q \tag{4.2}$$

Besov classes admit many other descriptions. In particular, for $s > 0$, the space $B^s_{pq}$ admits the following characterization. A function $\varphi$ belongs to $B^s_{pq}$, $s > 0$, if and only if

$$\int_{\mathbb{T}} \frac{\|\Delta_\tau^n \varphi\|_{L^p}^q}{|1-\tau|^{1+sq}} d\boldsymbol{m}(\tau) < \infty \quad \text{for} \quad q < \infty$$

and

$$\sup_{\tau\ne1} \frac{\|\Delta_\tau^n \varphi\|_{L^p}}{|1-\tau|^s} < \infty \quad \text{for} \quad q = \infty, \tag{4.3}$$

where $\boldsymbol{m}$ is normalized Lebesgue measure on $\mathbb{T}$, $n$ is an integer greater than $s$, and $\Delta_\tau$ is the difference operator:

$$(\Delta_\tau\varphi)(\zeta) = \varphi(\tau\zeta) - \varphi(\zeta), \quad \zeta \in \mathbb{T}.$$

We use the notation $B^s_p$ for $B^s_{pp}$.

The spaces $\Lambda_\alpha \stackrel{\mathrm{def}}{=} B^\alpha_\infty$ form the *Hölder–Zygmund scale.* If $0 < \alpha < 1$, then $\varphi \in \Lambda_\alpha$ if and only if

$$|\varphi(\zeta) - \varphi(\tau)| \le \operatorname{const} |\zeta - \tau|^\alpha, \quad \zeta, \tau \in \mathbb{T},$$

while $f \in \Lambda_1$ if and only if

$$|\varphi(\zeta\tau) - 2\varphi(\zeta) + \varphi(\zeta\bar{\tau})| \le \operatorname{const} |1 - \tau|, \quad \zeta, \tau \in \mathbb{T}.$$

It follows from (4.3) that for $\alpha > 0$, $\varphi \in \Lambda_\alpha$ if and only if

$$|(\Delta^n_\tau \varphi)(\zeta)| \le \operatorname{const} |1 - \tau|^\alpha,$$

where $n$ is a positive integer such that $n > \alpha$.

It is easy to see from the definition of Besov classes that the Riesz projection $\mathbb{P}_+$,

$$\mathbb{P}_+\varphi = \sum\nolimits_{n \ge 0} \hat{\varphi}(n) z^n,$$

is bounded on $B^s_{pq}$ and functions in $\big(B^s_{pq}\big)_+ \stackrel{\mathrm{def}}{=} \mathbb{P}_+ B^s_{pq}$ admit a natural extension to analytic functions in the unit disk $\mathbb{D}$. It is well known that the functions in $\big(B^s_{pq}\big)_+$ admit the following description:

$$\varphi \in \big(B^s_{pq}\big)_+ \Leftrightarrow \int_0^1 (1-r)^{q(n-s)-1} \|\varphi^{(n)}_r\|^q_p \, dr < \infty, \quad q < \infty,$$

and

$$\varphi \in \big(B^s_{p\infty}\big)_+ \Leftrightarrow \sup_{0<r<1} (1-r)^{n-s} \|\varphi^{(n)}_r\|_p < \infty,$$

where $\varphi_r(\zeta) \stackrel{\mathrm{def}}{=} \varphi(r\zeta)$ and $n$ is a nonnegative integer greater than $s$.

Besov spaces play a significant role in many problems of operator theory and it is also important to consider Besov spaces $B^s_{pq}$ when $p$ or $q$ can be less than 1. Everything mentioned above also holds for arbitrary positive $p$ and $q$ provided $s > 1/p - 1$.

Let us proceed now to Besov spaces on the real line. We consider homogeneous Besov spaces $B^s_{pq}(\mathbb{R})$ of functions (distributions) on $\mathbb{R}$. We use the same function $w$ as in (4.1) and define the functions $W_n$ and $W^\#_n$ on $\mathbb{R}$ by

$$\mathcal{F}W_n(x) = w\left(\frac{x}{2^n}\right), \quad \mathcal{F}W^\#_n(x) = \mathcal{F}W_n(-x), \quad n \in \mathbb{Z},$$

where $\mathcal{F}$ is the *Fourier transform.* The Besov class $B^s_{pq}(\mathbb{R})$ consists of distributions $\varphi$ on $\mathbb{R}$ such that

$$\{\|2^{ns}\varphi * W_n\|_{L^p}\}_{n\in\mathbb{Z}} \in \ell^q(\mathbb{Z}) \quad \text{and} \quad \{\|2^{ns}\varphi * W^\#_n\|_{L^p}\}_{n\in\mathbb{Z}} \in \ell^q(\mathbb{Z}).$$

According to this definition, the space $B^s_{pq}(\mathbb{R})$ contains all polynomials. However, it is not necessary to include all polynomials. The definition of $B^2_{pq}(\mathbb{R})$ can be slightly modified in a natural way so that it contains no polynomials of degree greater than $s - 1/p$ (see [AP2]).

Besov spaces $B^s_{pq}(\mathbb{R})$ admit equivalent definitions that are similar to those discussed above in the case of Besov spaces of functions on $\mathbb{T}$. We refer the reader to [Pee] and [Pe6] for more detailed information on Besov spaces.

## 5. Nuclearity of Hankel operators

It turns out (see [Pe2]) that Hankel operators play an important role in our problems of perturbation theory. For a function $\varphi$ on the unit circle $\mathbb{T}$, the Hankel operator $H_\varphi$ on the Hardy class $H^2 \subset L^2$ is defined by

$$H_\varphi : H^2 \to H^2_- \stackrel{\text{def}}{=} L^2 \ominus H^2, \quad H_\varphi f \stackrel{\text{def}}{=} \mathbb{P}_- \varphi f,$$

where $\mathbb{P}_-$ is the orthogonal projection onto $H^2_-$. By Nehari's theorem,

$$\|H_\varphi\| = \operatorname{dist}_{L^\infty}(\varphi, H^\infty)$$

(see [Pe6], Ch. 1, § 1).

In this paper we need the following result that describes the Hankel operator of trace class $\boldsymbol{S}_1$.

**Theorem 5.1.** *$H_\varphi \in \boldsymbol{S}_1$ if and only if $\mathbb{P}_-\varphi \in B^1_1$.*

Theorem 5.1 was obtained in [Pe1], see also [Pe6], Ch. 6, § 1.

Consider now the following class of integral operators. Let $\varphi \in L^\infty$. The operator $\mathcal{C}_\varphi$ on $L^2$ is defined by

$$(\mathcal{C}_\varphi f)(\zeta) = \int_{\mathbb{T}} \frac{\varphi(\zeta) - \varphi(\tau)}{1 - \bar{\tau}\zeta} f(\tau)\, d\boldsymbol{m}(\tau).$$

The following result can be deduced from Theorem 5.1 (see [Pe6], Ch. 6, $\S\dot{7}$).

**Theorem 5.2.** *Let $\varphi \in L^\infty$. Then $\mathcal{C}_\varphi \in \boldsymbol{S}_1$ if and only if $\varphi \in B^1_1$.*

*Proof.* Indeed, it is easy to show that

$$\mathcal{C}_\varphi f = H_\varphi f_+ - H^*_{\bar{\varphi}} f_-,$$

where $f_- = \mathbb{P}_- f$ and $f_+ = f - f_-$. Theorem 5.2 follows now immediately from Theorem 5.1. □

## 6. Operator Lipschitz and operator differentiable functions. Sufficient conditions

In this section we discuss sufficient conditions for a function on the unit circle or on the real line to be operator Lipschitz or operator differentiable. We begin with unitary operators.

The following lemma gives us an estimate for the norm of $\mathfrak{D}\varphi$ in the projective tensor product $L^\infty \hat{\otimes} L^\infty$ in the case of trigonometric polynomials $\varphi$. It was obtained in [Pe2]. We give here a slightly modified proof given in [Pe8].

**Lemma 6.1.** *Let $\varphi$ be a trigonometric polynomial of degree $m$. Then*

$$\|\mathfrak{D}\varphi\|_{L^\infty \hat{\otimes} L^\infty} \le \operatorname{const} m \|\varphi\|_{L^\infty}. \tag{6.1}$$

*Proof.* First of all, it is evident that it suffices to consider the case when $m = 2^l$. Next, it suffices to prove the result for analytic polynomials $\varphi$ (i.e., linear combinations of $z^j$ with $j \ge 0$). Indeed, if (6.1) holds for analytic polynomials, then it obviously also holds for conjugate trigonometric polynomials. Let now $\varphi$ be an arbitrary trigonometric polynomial of degree $2^l$. We have

$$\varphi = \sum_{j=1}^{l} \varphi * W_j^{\#} + \sum_{j=0}^{l} \varphi * W_j$$

(see § 4). Applying (6.1) to each term of this expansion, we obtain

$$\begin{aligned}\|\mathfrak{D}\varphi\|_{L^\infty \hat{\otimes} L^\infty} &\le \sum_{j=1}^{l} \big\|\mathfrak{D}\big(\varphi * W_j^{\#}\big)\big\|_{L^\infty \hat{\otimes} L^\infty} + \sum_{j=0}^{l} \|\mathfrak{D}(\varphi * W_j)\|_{L^\infty \hat{\otimes} L^\infty} \\ &\le \operatorname{const}\left(\sum_{j=1}^{l} 2^j \big\|\varphi * W_j^{\#}\big\|_{L^\infty} + \sum_{j=0}^{l} 2^j \|\varphi * W_j\|_{L^\infty}\right) \\ &\le \operatorname{const} \sum_{j=0}^{l} 2^j \|\varphi\|_{L^\infty} \le \operatorname{const} 2^l \|\varphi\|_{L^\infty}.\end{aligned}$$

Assume now that $\varphi$ is an analytic polynomial of degree $m$. It is easy to see that

$$(\mathfrak{D}\varphi)(z_1, z_2) = \sum_{j,k \ge 0} \hat{\varphi}(j + k + 1) z_1^j z_2^k.$$

We have

$$\sum_{j,k \ge 0} \hat{\varphi}(j + k + 1) z_1^j z_2^k = \sum_{j,k \ge 0} \alpha_{jk} \hat{\varphi}(j + k + 1) z_1^j z_2^k + \sum_{j,k \ge 0} \beta_{jk} \hat{\varphi}(j + k + 1) z_1^j z_2^k,$$

where

$$\alpha_{jk} = \begin{cases} \frac{1}{2}, & j = k = 0, \\ \frac{j}{j+k}, & j + k \ne 0 \end{cases} \quad \text{and} \quad \beta_{jk} = \begin{cases} \frac{1}{2}, & j = k = 0, \\ \frac{k}{j+k}, & j + k \ne 0. \end{cases}$$

Clearly, it is sufficient to estimate

$$\left\| \sum_{j,k \ge 0} \alpha_{jk} \hat{\varphi}(j + k + 1) z_1^j z_2^k \right\|_{L^\infty \hat{\otimes} L^\infty}.$$

It is easy to see that

$$\sum_{j,k \ge 0} \alpha_{jk} \hat{\varphi}(j + k + 1) z_1^j z_2^k = \sum_{k \ge 0} \left( \Big( \big((S^*)^{k+1} \varphi\big) * \sum_{j \ge 0} \alpha_{jk} z^j \Big)(z_1) \right) z_2^k,$$

where $S^*$ is backward shift, i.e., $(S^*)^k \varphi = \mathbb{P}_+ \bar{z}^k \varphi$.

Thus

$$\left\|\sum_{j,k\ge0}\alpha_{jk}\hat\varphi(j+k+1)z_1^jz_2^k\right\|_{L^\infty\hat\otimes L^\infty}\le\sum_{k\ge0}\left\|\left((S^*)^{k+1}\varphi\right)*\sum_{j\ge0}\alpha_{jk}z^j\right\|_{L^\infty}.$$

Put

$$Q_k(z)=\sum_{i\ge k}\frac{i-k}{i}z^i,\quad k>0,\quad\text{and}\quad Q_0(z)=\frac12+\sum_{i\ge1}z^i.$$

Then it is easy to see that

$$\left\|\left((S^*)^{k+1}\varphi\right)*\sum_{j\ge0}\alpha_{jk}z^j\right\|_{L^\infty}=\|\psi*Q_k\|_{L^\infty},$$

where $\psi=S^*\varphi$, and so

$$\left\|\sum_{j,k\ge0}\alpha_{jk}\hat\varphi(j+k+1)z_1^jz_2^k\right\|_{L^\infty\hat\otimes L^\infty}\le\sum_{k\ge0}\|\psi*Q_k\|_{L^\infty}.$$

Consider the function $r$ on $\mathbb{R}$ defined by

$$r(x)=\begin{cases}1, & |x|\le1,\\ \frac{1}{|x|}, & |x|\ge1.\end{cases}$$

It is easy to see that the Fourier transform $\mathcal{F}r$ of $r$ belongs to $L^1(\mathbb{R})$. Define the functions $R_n$, $n\ge1$, on $\mathbb{T}$ by

$$R_k(\zeta)=\sum_{j\in\mathbb{Z}}r\left(\frac{j}{k}\right)\zeta^j.$$

An elementary estimate obtained in Lemma 4.3 of [Pe8] shows that

$$\|R_k\|_{L^1}\le\text{const}.$$

It is easy to see that for $f\in H^\infty$, we have

$$\|f*Q_k\|_{L^\infty}=\|f-f*R_k\|_{L^\infty}\le\|f\|_{L^\infty}+\|f*R_k\|_{L^\infty}\le\text{const}\,\|f\|_{L^\infty}.$$

Thus

$$\sum_{k\ge0}\|\psi*Q_k\|_{L^\infty}=\sum_{k=0}^m\|\psi*Q_k\|_{L^\infty}\le\text{const}\,m\|\psi\|_{L^\infty}\le\text{const}\,m\|\varphi\|_{L^\infty}.\qquad\square$$

The following result was obtained in [Pe2].

**Theorem 6.2.** *Let $\varphi\in B^1_{\infty1}$. Then $\mathfrak{D}\varphi\in C(\mathbb{T})\hat\otimes C(\mathbb{T})$ and*

$$\|\mathfrak{D}\varphi\|_{L^\infty\hat\otimes L^\infty}\le\text{const}\,\|\varphi\|_{B^1_{\infty1}}.$$

*Proof.* We have

$$\varphi = \sum_{j>0} \varphi * W_j^{\#} + \sum_{j\ge 0} \varphi * W_j.$$

By Lemma 6.1, each of the functions $\mathfrak{D}\big(\varphi * W_j^{\#}\big)$ and $\mathfrak{D}(\varphi * W_j)$ belongs to $C(\mathbb{T})\hat{\otimes}C(\mathbb{T})$ and

$$\begin{aligned}\sum_{j>0} \big\|\mathfrak{D}\big(\varphi * W_j^{\#}\big)\big\|_{L^\infty\hat{\otimes}L^\infty} &+ \sum_{j\ge 0} \|\mathfrak{D}(\varphi * W_j)\|_{L^\infty\hat{\otimes}L^\infty}\\ &\le \operatorname{const}\Big(\sum_{j>0} 2^j \big\|\varphi * W_j^{\#}\big\|_{L^\infty} + \sum_{j\ge 0} 2^j \|\varphi * W_j\|_{L^\infty}\Big)\\ &\le \operatorname{const}\|\varphi\|_{B^1_{\infty 1}}. \qquad \square\end{aligned}$$

It follows from Theorem 6.2 that for $\varphi \in B^1_{\infty 1}$, the divided difference $\mathfrak{D}\varphi$ belongs to the space $\mathfrak{M}(E,F)$ of Schur multipliers with respect to arbitrary Borel spectral measures $E$ and $F$ on $\mathbb{T}$ (see § 2). By the Birman–Solomyak formula for unitary operators, we have

$$\varphi(U) - \varphi(V) = \iint\limits_{\mathbb{T}\times\mathbb{T}} \frac{\varphi(\zeta)-\varphi(\tau)}{\zeta-\tau}\, dE_U(\zeta)\,(U-V)\,dE_V(\tau), \tag{6.2}$$

which implies the following result:

**Theorem 6.3.** *Let $\varphi \in B^1_{\infty 1}$. Then $\varphi$ is operator Lipschitz, i.e.,*

$$\|\varphi(U) - \varphi(V)\| \le \operatorname{const}\|U - V\|,$$

*for unitary operators $U$ and $V$ on Hilbert space.*

*Proof.* It follows from (6.2) that

$$\begin{aligned}\|\varphi(U) - \varphi(V)\| &\le \|\mathfrak{D}\varphi\|_{\mathfrak{M}(E_U,E_V)}\|U - V\|\\ &\le \|\mathfrak{D}\varphi\|_{L^\infty\hat{\otimes}L^\infty}\|U - V\| \le \operatorname{const}\|\varphi\|_{B^1_{\infty 1}}\|U - V\|. \qquad \square\end{aligned}$$

Let us now show that the condition $\varphi \in B^1_{\infty 1}$ also implies that $\varphi$ is operator differentiable.

**Theorem 6.4.** *Let $\varphi$ be a function on $\mathbb{T}$ of class $B^1_{\infty 1}$. If $A$ is a bounded self-adjoint operator and $U$ is a unitary operator, and $U_s \overset{\text{def}}{=} e^{\mathrm{i}sA}U$, then the function*

$$s \mapsto \varphi(U_s) \tag{6.3}$$

*is differentiable in the norm and*

$$\frac{d}{ds}\big(\varphi(U_s)\big)\Big|_{s=o} = \mathrm{i}\left(\iint \frac{\varphi(\zeta)-\varphi(\tau)}{\zeta-\tau}\, dE_U(\zeta)A\,dE_U(\tau)\right)U. \tag{6.4}$$

*Moreover, the map*

$$A \mapsto \varphi\big(e^{\mathrm{i}A}U\big) \tag{6.5}$$

*defined on the space of bounded self-adjoint operators is differentiable in the sense of Fréchet.*

*Proof.* Let us prove that the function (6.3) is norm differentiable and that formula (6.4) holds. By Theorem 6.2, there exist continuous functions $f_n$ and $g_n$ on $\mathbb{T}$ such that

$$(\mathfrak{D}\varphi)(\zeta,\tau)=\sum_{n\ge1}f_n(\zeta)g_n(\tau),\quad \zeta,\,\tau\in\mathbb{T},$$

and

$$\sum_{n\ge1}\|f_n\|_{L^\infty}\|g_n\|_{L^\infty}<\infty. \tag{6.6}$$

By the Birman–Solomyak formula (6.2),

$$\begin{aligned}\varphi(U_s)-\varphi(U)&=\iint\limits_{\mathbb{T}\times\mathbb{T}}(\mathfrak{D}\varphi)(\zeta,\tau)\,dE_{U_s}(\zeta)(U_s-U)\,dE_U(\tau)\\&=\sum_{n\ge1}f_n(U_s)(U_s-U)g_n(U).\end{aligned}$$

On the other hand,

$$\mathrm{i}\left(\iint\frac{\varphi(\zeta)-\varphi(\tau)}{\zeta-\tau}\,dE_U(\zeta)A\,dE_U(\tau)\right)U=\mathrm{i}\sum_{n\ge1}f_n(U)Ag_n(U)U.$$

We have

$$\begin{aligned}&\frac1s\big(\varphi(U_s)-\varphi(U)\big)-\mathrm{i}\sum_{n\ge1}f_n(U)Ag_n(U)U\\&\qquad=\sum_{n\ge1}\left(\frac1sf_n(U_s)(U_s-U)g_n(U)-\mathrm{i}f_n(U)Ag_n(U)U\right)\\&\qquad=\sum_{n\ge1}\left(\frac1sf_n(U_s)(U_s-U)g_n(U)-\frac1sf_n(U)(U_s-U)g_n(U)\right)\\&\qquad\quad+\sum_{n\ge1}\left(\frac1sf_n(U)(U_s-U)g_n(U)-\mathrm{i}f_n(U)Ag_n(U)U\right).\end{aligned}$$

Clearly,

$$\left\|\frac1s(U_s-U)\right\|\le\operatorname{const}.$$

Since, $f_n\in C(\mathbb{T})$, it is easy to see that

$$\lim_{s\to0}\|f_n(U_s)-f_n(U)\|=0.$$

It follows now easily from (6.6) that

$$\lim_{s\to0}\left\|\sum_{n\ge1}\left(\frac1sf_n(U_s)(U_s-U)g_n(U)-\frac1sf_n(U)(U_s-U)g_n(U)\right)\right\|=0.$$

On the other hand, it is easy to see that

$$\lim_{s\to\mathbf{0}}\left\|\frac{1}{s}(U_s-U)-\mathrm{i}AU\right\|=0$$

and again, it follows from (6.6) that

$$\lim_{s\to 0}\left\|\sum_{n\ge1}\left(\frac{1}{s}f_n(U)(U_s-U)g_n(U)-\mathrm{i}f_n(U)Ag_n(U)U\right)\right\|=0$$

which proves that the function (6.3) is norm differentiable and (6.4) holds.

One can easily see that the same reasoning also shows that the map (6.5) is differentiable in the sense of Fréchet. □

The above results of this section were obtained in [Pe2]. Recall that earlier Birman and Solomyak proved in [BS4] the same results for functions $\varphi$ whose derivatives belong to the Hölder class $\Lambda_\alpha$ with some $\alpha>0$.

In the case of differentiability in the Hilbert–Schmidt norm, the following result was proved by Birman and Solomyak in [BS4].

**Theorem 6.5.** *Let $\varphi\in C^1(\mathbb{T})$. If under the hypotheses of Theorem* 6.4 *the self-adjoint operator $A$ belongs to the Hilbert–Schmidt class $\boldsymbol{S}_2$, then formula* (6.4) *holds in the Hilbert–Schmidt norm.*

Let us state now similar results for (not necessarily bounded) self-adjoint operators. The following result shows that functions in $B^1_{\infty1}(\mathbb{R})$ are operator Lipschitz.

**Theorem 6.6.** *Let $\varphi$ be a function on $\mathbb{R}$ of class $B^1_{\infty1}(\mathbb{R})$ and let $A$ and $B$ be self-adjoint operators such that $A-B$ is bounded. Then the operator $\varphi(A)-\varphi(B)$ is bounded and*

$$\|\varphi(A)-\varphi(B)\|\le\operatorname{const}\|\varphi\|_{B^1_{\infty1}(\mathbb{R})}\|A-B\|.$$

**Theorem 6.7.** *Let $\varphi\in B^1_{\infty1}(\mathbb{R})$. Suppose that $A$ is a self-adjoint operator* (*not necessarily bounded*) *and $K$ is a bounded self-adjoint operator. Then the function*

$$t\mapsto f(A+tK)-f(A)$$

*is norm differentiable and*

$$\frac{d}{dt}f(A+tK)\Big|_{t=0}=\iint\limits_{\mathbb{R}\times\mathbb{R}}\frac{\varphi(x)-\varphi(y)}{x-y}\,dE_A(x)K\,dE_A(y).$$

*Moreover, the map*

$$K\mapsto f(A+K)-f(A)$$

*defined on the space of bounded self-adjoint operators is differentiable in the sense of Fréchet.*

We refer the reader to [Pe4] and [Pe8] for the proofs of Theorems 6.6 and 6.7.

## 7. Operator Lipschitz and operator differentiable functions. Necessary conditions

**Theorem 7.1.** *Let $\varphi$ be a continuously differentiable function on $\mathbb{T}$. If $\varphi$ is operator Lipschitz, then $\varphi \in B_1^1$.*

Note that the condition $\varphi \in B_1^1$ implies that

$$\sum_{n\ge 0} |\widehat{\varphi'}(2^n)| < \infty.$$

This follows easily from (4.2). On the other hand, it is well known that for an arbitrary sequence $\{c_n\}_{n\ge 0}$ in $\ell^2$, there exists $\varphi \in C^1(\mathbb{T})$ such that

$$\widehat{\varphi'}(2^n) = c_n, \quad n \ge 0.$$

Thus the condition $\varphi \in C^1(\mathbb{T})$ is not sufficient for $\varphi$ to be operator Lipschitz.

*Proof.* Let $U$ be multiplication by $z$ on $L^2$ (with respect to Lebesgue measure) and let $A$ be a self-adjoint operator on $L^2$ of class $\boldsymbol{S}_2$. Put $V_t = e^{\mathrm{i}tA}U$, $t \in \mathbb{R}$. It is easy to see that

$$\frac{1}{t}\|V_t - U\| \le \mathrm{const}\,\|A\|,$$

and since $\varphi$ is operator Lipschitz, we have

$$\left\|\frac{1}{t}\big(\varphi(V_t) - \varphi(U)\big)\right\| \le \mathrm{const}\,\|A\|.$$

By Theorem 6.5,

$$\lim_{t\to 0}\frac{1}{t}\big(\varphi(V_t) - \varphi(U)\big) = \mathrm{i}\left(\iint \frac{\varphi(\zeta)-\varphi(\tau)}{\zeta-\tau}\,dE_U(\zeta)A\,dE_U(\tau)\right)U$$

in the Hilbert–Schmidt norm. It follows that

$$\left\|\iint \frac{\varphi(\zeta)-\varphi(\tau)}{\zeta-\tau}\,dE_U(\zeta)A\,dE_U(\tau)\right\| \le \mathrm{const}\,\|A\|.$$

This means that the divided difference $\mathfrak{D}\varphi$ is a Schur multiplier in $\mathfrak{M}(E_U, E_U)$.

Consider now the class $\mathfrak{N}$ of kernel functions of trace class integral operators on $L^2$ with respect to Lebesgue measure. By Theorem 2.4,

$$k \in \mathfrak{N} \Rightarrow (\mathfrak{D}\varphi)k \in \mathfrak{N}.$$

Put now

$$k(\zeta,\tau) \overset{\mathrm{def}}{=} \tau.$$

Clearly, the integral operator with kernel function $k$ is a rank one operator. We have

$$\big((\mathfrak{D}\varphi)k\big)(\zeta,\tau) = \frac{\varphi(\zeta)-\varphi(\tau)}{1-\bar{\tau}\zeta}, \quad \zeta,\, \tau \in \mathbb{T}.$$

Thus $\mathcal{C}_\varphi \in \boldsymbol{S}_1$ and it follows now from Theorem 5.2 that $\varphi \in B_1^1$. □

**Remark.** It is easy to see that the reasoning given in the proof of Theorem 7.1 also gives the following result:

*Suppose that $\varphi \in C^1(\mathbb{T})$ and the divided difference $\mathfrak{D}\varphi$ is not a Schur multiplier in $\mathfrak{M}(E_U, E_U)$ (or, in other words, $\mathfrak{D}\varphi$ is not a multiplier of the class $\mathfrak{N}$ of kernel functions of trace class integral operators on $L^2(\boldsymbol{m})$). Then $\varphi$ is not operator Lipschitz.*

Theorem 7.1 was proved in [Pe2]. A more elaborate application of the nuclearity criterion for Hankel operators allowed the author to obtain in [Pe2] a stronger necessary condition. To state it, we introduce the class $\mathcal{L}$.

**Definition.** A bounded function $\varphi$ on $\mathbb{T}$ is said to belong to $\mathcal{L}$ if the Hankel operators $H_\varphi$ and $H_{\bar\varphi}$ map the Hardy class $H^1$ into the Besov space $B_1^1$, i.e.,

$$\mathbb{P}_-\varphi g \in B_1^1 \quad \text{and} \quad \mathbb{P}_-\bar\varphi f \in B_1^1,$$

whenever $f \in H^1$.

It is easy to see that $\mathcal{L} \subset B_1^1$.

The following result was obtained in [Pe2].

**Theorem 7.2.** *Let $\varphi$ be an operator Lipschitz function of class $C^1(\mathbb{T})$. Then $\varphi \in \mathcal{L}$.*

The proof of Theorem 7.2 is given in [Pe2]. It is based on the nuclearity criterion for Hankel operators, see Theorem 5.1. Actually, it is shown in [Pe2] that if $\varphi \in C^1(\mathbb{T}) \setminus \mathcal{L}$, then $\mathfrak{D}\varphi$ is not a multiplier of the class $\mathfrak{N}$.

S. Semmes observed (see the proof in [Pe5]) that $\varphi \in \mathcal{L}$ if and only if the measure

$$\|\operatorname{Hess}\varphi\|\, dx\, dy$$

is a Carleson measure in the unit disk $\mathbb{D}$, where $\operatorname{Hess}\varphi$ is the Hessian matrix of the harmonic extension of $\varphi$ to the unit disk.

M. Frazier observed that actually $\mathcal{L}$ is the Triebel–Lizorkin space $F_{\infty 1}^1$. Note that the definition of the Triebel–Lizorkin spaces $\dot F_{pq}^s$ on $\mathbb{R}^n$ for $p = \infty$ and $q > 1$ can be found in [T], § 5.1. A definition for all $q > 0$, which is equivalent to Triebel's definition when $q > 1$, was given by Frazier and Jawerth in [FrJ]. Their approach did not use harmonic extensions, but a straightforward exercise in comparing kernels shows that Frazier and Jawerth's definition of $\dot F_{\infty 1}^1$ is equivalent to the definition requiring $\|\operatorname{Hess}\varphi\| dxdy$ to be a Carleson measure on the upper half-space. Our space $\mathcal{L}$ is the analogue for the unit disc.

The condition $\varphi \in \mathcal{L}$ (and a fortiori the condition $\varphi \in B_1^1$) is also a necessary condition for the function $\varphi$ to be operator differentiable.

Similar results also hold in the case of functions of self-adjoint operators:

**Theorem 7.3.** *Let $\varphi$ be a continuously differentiable function on $\mathbb{R}$. If $\varphi$ is locally operator Lipschitz, then $\varphi$ belongs to $B_1^1(\mathbb{R})$ locally.*

Note that the latter property means that the restriction of $\varphi$ to any finite interval coincides with the restriction to this interval of a function of class $B_1^1(\mathbb{R})$.

**Theorem 7.4.** *Let $\varphi$ be a continuously differentiable function on $\mathbb{R}$. If $\varphi$ is operator Lipschitz, then $\varphi$ belongs to the class $\mathcal{L}(\mathbb{R})$.*

Note that the class $\mathcal{L}(\mathbb{R})$ can be defined by analogy with the class $\mathcal{L}$ of functions on $\mathbb{T}$. The same description in terms of Carleson measures also holds. Theorem 7.3 can be improved: if $\varphi$ is locally operator Lipschitz, then $\varphi$ must belong to $\mathcal{L}(\mathbb{R})$ locally.

Theorems 7.3 and 7.4 were proved in [Pe2] and [Pe4].

To conclude this section, we mention that the same necessary conditions also hold for operator differentiability.

## 8. Higher-order operator derivatives

For a function $\varphi$ on the circle the *divided differences $\mathfrak{D}^k\varphi$ of order $k$* are defined inductively as follows:

$$\mathfrak{D}^0\varphi \overset{\text{def}}{=} \varphi;$$

if $k \ge 1$, then in the case when $\lambda_1, \lambda_2, \dots, \lambda_{k+1}$ are distinct points in $\mathbb{T}$,

$$(\mathfrak{D}^k\varphi)(\lambda_1,\dots,\lambda_{k+1}) \overset{\text{def}}{=} \frac{(\mathfrak{D}^{k-1}\varphi)(\lambda_1,\dots,\lambda_{k-1},\lambda_k) - (\mathfrak{D}^{k-1}\varphi)(\lambda_1,\dots,\lambda_{k-1},\lambda_{k+1})}{\lambda_k - \lambda_{k+1}}$$

(the definition does not depend on the order of the variables). Clearly,

$$\mathfrak{D}\varphi = \mathfrak{D}^1\varphi.$$

If $\varphi \in C^k(\mathbb{T})$, then $\mathfrak{D}^k\varphi$ extends by continuity to a function defined for all points $\lambda_1, \lambda_2, \dots, \lambda_{k+1}$.

The following result was established in [Pe8].

**Theorem 8.1.** *Let $n \ge 1$ and let $\varphi$ be a function on $\mathbb{T}$ of class $B^n_{\infty 1}$. Then $\mathfrak{D}^n\varphi$ belongs to the projective tensor product $\underbrace{C(\mathbb{T})\hat{\otimes}\cdots\hat{\otimes}C(\mathbb{T})}_{n+1}$ and*

$$\|\mathfrak{D}^n\varphi\|_{L^\infty\hat{\otimes}\cdots\hat{\otimes}L^\infty} \le \operatorname{const}\|\varphi\|_{B^n_{\infty 1}}. \tag{8.1}$$

The constant on the right-hand side of (8.1) can depend on $n$.

As in the case of double operator integrals, the following lemma gives us a crucial estimate.

**Lemma 8.2.** *Let $n$ and $m$ be a positive integers and let $\varphi$ be a trigonometric polynomial of degree $m$. Then*

$$\|\mathfrak{D}^n\varphi\|_{L^\infty\hat{\otimes}\cdots\hat{\otimes}L^\infty} \le \operatorname{const} m^n\|\varphi\|_{L^\infty}. \tag{8.2}$$

Note that the constant on the right-hand side of (8.2) can depend on $n$, but does not depend on $m$.

The proof of Lemma 8.2 is based on the same ideas as the proof of Lemma 6.1. We refer the reader to [Pe8] for the proof of Lemma 8.2.

We deduce now Theorem 8.1 from Lemma 8.2.

*Proof of Theorem* 8.1. We have

$$\varphi=\sum_{j>0}\varphi*W_j^{\#}+\sum_{j\ge0}\varphi*W_j.$$

By Lemma 8.2,

$$\begin{aligned}\|\mathfrak{D}^n\varphi\|_{L^\infty\hat{\otimes}\cdots\hat{\otimes}L^\infty}&\le\sum_{j>0}\big\|\mathfrak{D}^n(\varphi*W_j^{\#})\big\|_{L^\infty\hat{\otimes}\cdots\hat{\otimes}L^\infty}+\sum_{j\ge0}\big\|\mathfrak{D}^n(\varphi*W_j)\big\|_{L^\infty\hat{\otimes}\cdots\hat{\otimes}L^\infty}\\&\le\operatorname{const}\Big(\sum_{j>0}2^{jn}\|\varphi*W_j^{\#}\|_{L^\infty}+\sum_{j\ge0}2^{jn}\|\varphi*W_j\|_{L^\infty}\Big)\\&\le\operatorname{const}\|\varphi\|_{B^n_{\infty1}}.\end{aligned}$$

□

Suppose now that $U$ is a unitary operator and $A$ is a bounded self-adjoint operator on Hilbert space. Consider the family of operators

$$U_t=e^{\mathrm{i}tA}U.\quad t\in\mathbb{R}.$$

The following result was proved in [Pe8].

**Theorem 8.3.** *Let $\varphi\in B^n_{\infty1}$. Then the function*

$$t\mapsto\varphi(U_t)$$

*has $n$ derivatives in the norm and*

$$\begin{aligned}&\frac{d^n}{dt^n}\big(\varphi(U_t)\big)\Big|_{s=0}\\&=\mathrm{i}^n n!\left(\underbrace{\int\cdots\int}_{n+1}(\mathfrak{D}^n\varphi)(\zeta_1,\dots,\zeta_{n+1})\,dE_U(\zeta_1)A\cdots A\,dE_U(\zeta_{n+1})\right)U^n.\end{aligned}$$

Similar results hold for self-adjoint operators. The following results can be found in [Pe8]:

**Theorem 8.4.** *Let $\varphi\in B^n_{\infty1}(\mathbb{R})$. Then $\mathfrak{D}^n\varphi$ belongs to the integral projective tensor product $\underbrace{L^\infty\hat{\otimes}\cdots\hat{\otimes}L^\infty}_{n+1}$.*

**Theorem 8.5.** *Suppose that $\varphi\in B^n_{\infty1}(\mathbb{R})\cap B^1_{\infty1}(\mathbb{R})$. Let $A$ be a self-adjoint operator and let $K$ be a bounded self-adjoint operator on Hilbert space. Then the function*

$$t\mapsto\varphi(A+tK)\tag{8.3}$$

*has an $n$th derivative in the norm and*

$$\begin{aligned}&\frac{d^n}{dt^n}\big(\varphi(A+tK)\big)\Big|_{s=0}\\&\quad=n!\underbrace{\int\cdots\int}_{n+1}(\mathfrak{D}^n\varphi)(x_1,\dots,x_{n+1})\,dE_U(x_1)A\cdots A\,dE_U(x_{n+1}).\end{aligned}\tag{8.4}$$

Note that under the hypotheses of Theorem 8.5, the function (8.3) has all the derivatives in the norm up to order $n$. However, in the case of unbounded self-adjoint operators it can happen that the $n$th derivative exists in the norm, but lower-order derivatives do not exist in the norm. For example, if $\varphi \in B^2_{\infty 1}(\mathbb{R})$, but $\varphi \notin B^1_{\infty 1}(\mathbb{R})$, then it can happen that the function (8.3) does not have the first derivative in the norm, but it is possible to interpret its second derivative so that the second derivative exists in the norm and can be computed by formula (8.4); see detailed comments in [Pe8].

Earlier sufficient conditions for the function (8.3) to have $n$ derivatives in the norm and satisfy (8.4) were found in [St]. However, the conditions found in [St] were much more restrictive.

## 9. The case of contractions

Let $T$ be a contraction (i.e., $\|T\| \le 1$) on Hilbert space. Von Neumann's inequality (see [SNF]) says that for an arbitrary analytic polynomial $\varphi$,

$$\|\varphi(T)\| \le \max_{|\zeta| \le 1} |\varphi(\zeta)|.$$

This allows one to define the functional calculus

$$\varphi \mapsto \varphi(T)$$

on the disk-algebra $C_A$.

In this case we consider the questions of the behavior of $\varphi(T)$ under perturbations of $T$. As in the case of unitary operators a function $\varphi \in C_A$ is called *operator Lipschitz* if

$$\|\varphi(T) - \varphi(R)\| \le \text{const}\, \|T - R\|$$

for arbitrary contractions $T$ and $R$. We also consider differentiability properties.

For contractions $T$ and $R$, we consider the one-parameter family of contractions

$$T_t = (1-t)T + tR, \quad \mathbf{0} \le t \le 1,$$

and we study differentiability properties of the map

$$t \mapsto \varphi(T_t) \tag{9.1}$$

for a given function $\varphi$ in $C_A$.

It was observed in [Pe3] that if $\varphi$ is an analytic function in $\big(B^1_{\infty 1}\big)_+$, then $\varphi$ is operator Lipschitz. To prove this, double operator integrals with respect to semi-spectral measures were used.

Recently in [KS] it was proved that if $\varphi \in C_A$, the the following are equivalent:

(i) $\|\varphi(U) - \varphi(V)\| \le \text{const}\, \|U - V\|$ for arbitrary unitary operators $U$ and $V$;

(ii) $\|\varphi(T) - \varphi(R)\| \le \text{const}\, \|T - R\|$ for arbitrary contractions $T$ and $R$.

In [Pe9] it was shown that the same condition $\big(B^1_{\infty 1}\big)_+$ implies that the function (9.1) is differentiable in the norm and the derivative can be expressed in terms of double operator integrals with respect to semi-spectral measures. It was also established in [Pe9] that under the condition $\varphi \in B^n_{\infty 1}$, the function (9.1) is $n$ times differentiable in the norm and the $n$th derivative can be expressed in terms of a multiple operator integral with respect to semi-spectral measures.

**Definition.** Let $\mathcal{H}$ be a Hilbert space and let $(\mathcal{X}, \mathcal{B})$ be a measurable space. A map $\mathcal{E}$ from $\mathcal{B}$ to the algebra $B(\mathcal{H})$ of all bounded operators on $\mathcal{H}$ is called a *semi-spectral measure* if

$$\mathcal{E}(\Delta) \ge \mathbf{0}, \quad \Delta \in \mathcal{B},$$
$$\mathcal{E}(\varnothing) = \mathbf{0} \quad \text{and} \quad \mathcal{E}(\mathcal{X}) = I,$$

and for a sequence $\{\Delta_j\}_{j\ge 1}$ of disjoint sets in $\mathcal{B}$,

$$\mathcal{E}\left(\bigcup_{j=1}^{\infty} \Delta_j\right) = \lim_{N\to\infty} \sum_{j=1}^{N} \mathcal{E}(\Delta_j) \quad \text{in the weak operator topology.}$$

If $\mathcal{K}$ is a Hilbert space, $(\mathcal{X}, \mathcal{B})$ is a measurable space, $E : \mathcal{B} \to B(\mathcal{K})$ is a spectral measure, and $\mathcal{H}$ is a subspace of $\mathcal{K}$, then it is easy to see that the map $\mathcal{E} : \mathcal{B} \to B(\mathcal{H})$ defined by

$$\mathcal{E}(\Delta) = P_{\mathcal{H}} E(\Delta) \big| \mathcal{H}, \quad \Delta \in \mathcal{B}, \tag{9.2}$$

is a semi-spectral measure. Here $P_{\mathcal{H}}$ stands for the orthogonal projection onto $\mathcal{H}$.

Naimark proved in [Na] (see also [SNF]) that all semi-spectral measures can be obtained in this way, i.e., a semi-spectral measure is always a *compression* of a spectral measure. A spectral measure $E$ satisfying (9.2) is called a *spectral dilation of the semi-spectral measure* $\mathcal{E}$.

A spectral dilation $E$ of a semi-spectral measure $\mathcal{E}$ is called *minimal* if

$$\mathcal{K} = \operatorname{clos\,span}\{E(\Delta)\mathcal{H} : \ \Delta \in \mathcal{B}\}.$$

It was shown in [MM] that if $E$ is a minimal spectral dilation of a semi-spectral measure $\mathcal{E}$, then $E$ and $\mathcal{E}$ are mutually absolutely continuous and all minimal spectral dilations of a semi-spectral measure are isomorphic in the natural sense.

If $\varphi$ is a bounded complex-valued measurable function on $\mathcal{X}$ and $\mathcal{E} : \mathcal{B} \to B(\mathcal{H})$ is a semi-spectral measure, then the integral

$$\int_{\mathcal{X}} f(x)\, d\mathcal{E}(x) \tag{9.3}$$

can be defined as

$$\int_{\mathcal{X}} f(x)\, d\mathcal{E}(x) = P_{\mathcal{H}} \left( \int_{\mathcal{X}} f(x)\, dE(x) \right) \Bigg| \mathcal{H}, \tag{9.4}$$

where $E$ is a spectral dilation of $\mathcal{E}$. It is easy to see that the right-hand side of (9.4) does not depend on the choice of a spectral dilation. The integral (9.3) can also be computed as the limit of sums

$$\sum f(x_\alpha)\mathcal{E}(\Delta_\alpha), \quad x_\alpha \in \Delta_\alpha,$$

over all finite measurable partitions $\{\Delta_\alpha\}_\alpha$ of $\mathcal{X}$.

If $T$ is a contraction on a Hilbert space $\mathcal{H}$, then by the Sz.-Nagy dilation theorem (see [SNF]), $T$ has a unitary dilation, i.e., there exist a Hilbert space $\mathcal{K}$ such that $\mathcal{H} \subset \mathcal{K}$ and a unitary operator $U$ on $\mathcal{K}$ such that

$$T^n = P_{\mathcal{H}} U^n \big| \mathcal{H}, \quad n \geq 0, \tag{9.5}$$

where $P_{\mathcal{H}}$ is the orthogonal projection onto $\mathcal{H}$. Let $E_U$ be the spectral measure of $U$. Consider the operator set function $\mathcal{E}$ defined on the Borel subsets of the unit circle $\mathbb{T}$ by

$$\mathcal{E}(\Delta) = P_{\mathcal{H}} E_U(\Delta) \big| \mathcal{H}, \quad \Delta \subset \mathbb{T}.$$

Then $\mathcal{E}$ is a semi-spectral measure. It follows immediately from (9.5) that

$$T^n = \int_{\mathbb{T}} \zeta^n \, d\mathcal{E}(\zeta) = P_{\mathcal{H}} \int_{\mathbb{T}} \zeta^n \, dE_U(\zeta) \Big| \mathcal{H}, \quad n \geq 0. \tag{9.6}$$

Such a semi-spectral measure $\mathcal{E}$ is called a *semi-spectral measure* of $\mathbb{T}$. Note that it is not unique. To have uniqueness, we can consider a minimal unitary dilation $U$ of $T$, which is unique up to an isomorphism (see [SNF]).

It follows easily from (9.6) that

$$\varphi(T) = P_{\mathcal{H}} \int_{\mathbb{T}} \varphi(\zeta) \, dE_U(\zeta) \Big| \mathcal{H}$$

for an arbitrary function $\varphi$ in the disk-algebra $C_A$.

In [Pe9] double operator integrals and multiple operator integrals with respect to semi-spectral measures were introduced.

Suppose that $(\mathcal{X}_1, \mathcal{B}_1)$ and $(\mathcal{X}_2, \mathcal{B}_2)$ are measurable spaces, and $\mathcal{E}_1 : \mathcal{B}_1 \to B(\mathcal{H}_1)$ and $\mathcal{E}_2 : \mathcal{B}_2 \to B(\mathcal{H}_2)$ are semi-spectral measures. Then double operator integrals

$$\iint\limits_{\mathcal{X}_1 \times \mathcal{X}_2} \Phi(x_1, x_2) \, d\mathcal{E}_1(x_1) Q \, d\mathcal{E}_2(X_2).$$

were defined in [Pe9] in the case when $Q \in \boldsymbol{S}_2$ and $\Phi$ is a bounded measurable function and in the case when $Q$ is a bounded linear operator and $\Phi$ belongs to the integral projective tensor product of the spaces $L^\infty(\mathcal{E}_1)$ and $L^\infty(\mathcal{E}_2)$.

Similarly, multiple operator integrals with respect to semi-spectral measures were defined in [Pe9] for functions that belong to the integral projective tensor product of the corresponding $L^\infty$ spaces.

Let us now state the results obtained in [Pe9].

For a contraction $T$ on Hilbert space, we denote by $\mathcal{E}_T$ a semi-spectral measure of $\mathbb{T}$. Recall that if $\varphi' \in C_A$, the function $\mathfrak{D}\varphi$ extends to the diagonal

$$\boldsymbol{\Delta} \stackrel{\text{def}}{=} \big\{(\zeta, \zeta) : \ \zeta \in \mathbb{T}\big\}$$

by continuity: $(\mathfrak{D}\varphi)(\zeta, \zeta) = \varphi'(\zeta)$, $\zeta \in \mathbb{T}$.

**Theorem 9.1.** *Let $\varphi \in \big(B^1_{\infty 1}\big)_+$. Then for contractions $T$ and $R$ on Hilbert space the following formula holds:*

$$\varphi(T) - \varphi(R) = \iint\limits_{\mathbb{T}\times\mathbb{T}} \frac{\varphi(\zeta) - \varphi(\tau)}{\zeta - \tau}\, d\mathcal{E}_T(\zeta)\,(T-R)\,d\mathcal{E}_R(\tau). \tag{9.7}$$

**Theorem 9.2.** *Let $\varphi$ be a function analytic in $\mathbb{D}$ such that $\varphi' \in C_A$. If $T$ and $R$ are contractions such that $T - R \in \boldsymbol{S}_2$, then formula* (9.7) *holds.*

**Remark.** Theorem 9.2 can be extended to the more general case when $\varphi' \in H^\infty$. In this case we can define $\mathfrak{D}\varphi$ to be zero on the diagonal $\boldsymbol{\Delta}$.

The following result is an immediate consequence of the above remark; it was obtained recently in [KS] by a completely different method.

**Corollary 9.3.** *Suppose that $\varphi$ is a function analytic in $\mathbb{D}$ such that $\varphi' \in H^\infty$. If $T$ and $R$ are contractions on Hilbert space such that $T - R \in \boldsymbol{S}_2$, then*

$$\varphi(T) - \varphi(R) \in \boldsymbol{S}_2$$

*and*

$$\|\varphi(R) - \varphi(T)\|_{\boldsymbol{S}_2} \le \|\varphi'\|_{H^\infty}\|T-R\|_{\boldsymbol{S}_2}.$$

We proceed now to the differentiability problem. Let $T$ and $R$ be contractions on Hilbert space and let $\varphi \in C_A$. We are interested in differentiability properties of the function (9.1).

Let $\mathcal{E}_t$ be a semi-spectral measure of $T_t$ on the unit circle $\mathbb{T}$, i.e.,

$$T_t^n = \int\limits_{\mathbb{T}} \zeta^n\, d\mathcal{E}_t(\zeta), \quad n \ge 0.$$

Put $\mathcal{E} \overset{\text{def}}{=} \mathcal{E}_0$.

The following results were established in [Pe9].

**Theorem 9.4.** *Suppose that $\varphi \in \big(B^1_{\infty 1}\big)_+$. Then the function* (9.1) *is differentiable in the norm and*

$$\frac{d}{ds}\varphi(T_s)\Big|_{s=t} = \iint\limits_{\mathbb{T}\times\mathbb{T}} \frac{\varphi(\zeta) - \varphi(\tau)}{\zeta - \tau}\, d\mathcal{E}_t(\zeta)\,(R-T)\,d\mathcal{E}_t(\tau).$$

Note that the same result holds in the case when $T - R \in \boldsymbol{S}_2$ and $\varphi$ is an analytic function in $\mathbb{D}$ such that $\varphi' \in C_A$. In this case the derivative exists in the Hilbert –Schmidt norm; see [Pe9].

We conclude this section with an analog of Theorem 8.5 for contractions.

**Theorem 9.5.** *Suppose that $\varphi \in \big(B^n_{\infty 1}\big)_+$. Then the function* (9.1) *has nth derivative in the norm*

$$\frac{d^n}{ds^n}\varphi(T_s)\Big|_{s=t} = n!\underbrace{\int\cdots\int}_{n+1}(\mathfrak{D}^n\varphi)(\zeta_1,\dots,\zeta_{n+1})\,d\mathcal{E}_t(\zeta_1)\,(R-T)\cdots(R-T)\,d\mathcal{E}_t(\zeta_{n+1}).$$

We refer the reader to [Pe9] for proofs.

## 10. Operator Hölder–Zygmund functions

As we have mentioned in the introduction, surprisingly, Hölder functions of order $\alpha$ must also be operator Hölder functions of order $\alpha$. The same is true for all spaces of the scale $\Lambda_\alpha$ of Hölder–Zygmund classes.

Recall that the results of this section were obtained in [AP2] (see also [AP1]).

Let us consider the case of unitary operators.

**Theorem 10.1.** *Let* $0 < \alpha < 1$ *and* $\varphi \in \Lambda_\alpha$. *If* $U$ *and* $V$ *are unitary operators on Hilbert space, then*

$$\|\varphi(U) - \varphi(V)\| \le \text{const}\,\|\varphi\|_{\Lambda_\alpha} \cdot \|U - V\|^\alpha.$$

Note that the constant on the right-hand side of the inequality depends on $\alpha$.

In the proof of Theorem 10.1 we are going to use the following norm on $\Lambda_\alpha$ (see §4):

$$\|f\|_{\Lambda_\alpha} = \sup_{n\ge 0} 2^{n\alpha}\|\varphi * W_n\|_{L^\infty} + \sup_{n>0} 2^{n\alpha}\|\varphi * W_n^\#\|_{L^\infty}.$$

*Proof of Theorem* 10.1. Let $\varphi \in \Lambda_\alpha$. We have

$$\varphi = \mathbb{P}_+\varphi + \mathbb{P}_-\varphi = \varphi_+ + \varphi_-.$$

We estimate $\|\varphi_+(U) - \varphi_+(V)\|$. The norm of $\varphi_-(U) - \varphi_-(V)$ can be estimated in the same way. Thus we assume that $\varphi = \varphi_+$. Let

$$\varphi_n \overset{\text{def}}{=} \varphi * W_n.$$

Then

$$\varphi = \sum_{n\ge 0} \varphi_n. \tag{10.1}$$

Clearly, we may assume that $U \ne V$. Let $N$ be the nonnegative integer such that

$$2^{-N} < \|U - V\| \le 2^{-N+1}. \tag{10.2}$$

We have

$$\varphi(U) - \varphi(V) = \sum_{n\le N}\big(\varphi_n(U) - \varphi_n(V)\big) + \sum_{n>N}\big(\varphi_n(U) - \varphi_n(V)\big).$$

By Lemma 6.1,

$$\begin{aligned}\Big\|\sum_{n\le N}\big(\varphi_n(U) - \varphi_n(V)\big)\Big\| &\le \sum_{n\le N}\|\varphi_n(U) - \varphi_n(V)\|\\ &\le \text{const}\sum_{n\le N} 2^n\|U - V\|\cdot\|\varphi_n\|_{L^\infty}\\ &\le \text{const}\,\|U - V\|\sum_{n\le N} 2^n 2^{-n\alpha}\|\varphi\|_{\Lambda_\alpha}\\ &\le \text{const}\,\|U - V\| 2^{N(1-\alpha)}\|\varphi\|_{\Lambda_\alpha} \le \text{const}\,\|U - V\|^\alpha\|\varphi\|_{\Lambda_\alpha},\end{aligned}$$

the last inequality being a consequence of (10.2).

On the other hand,

$$\left\|\sum_{n>N}\big(\varphi_n(U)-\varphi_n(V)\big)\right\| \le \sum_{n>N} 2\|\varphi_n\|_{L^\infty} \le \operatorname{const}\sum_{n>N} 2^{-n\alpha}\|\varphi\|_{\Lambda_\alpha}$$
$$\le \operatorname{const} 2^{-N\alpha}\|\varphi\|_{\Lambda_\alpha} \le \operatorname{const}\|U-V\|^\alpha\|\varphi\|_{\Lambda_\alpha}. \qquad \square$$

Consider now the case of an arbitrary positive $\alpha$.

**Theorem 10.2.** *Let $n$ be a positive integer, $0<\alpha<n$, and let $f\in\Lambda_\alpha$. Then for a unitary operator $U$ and a bounded self-adjoint operator $A$ on Hilbert space the following inequality holds:*

$$\left\|\sum_{k=0}^{n}(-1)^k\binom{n}{k}\varphi\big(e^{\mathrm{i}kA}U\big)\right\| \le \operatorname{const}\|\varphi\|_{\Lambda_\alpha}\|A\|^\alpha.$$

The proof of Theorem 10.2 given in [AP2] is based on multiple operator integrals and Lemma 8.2. Let me explain how we arrive at triple operator integrals in the case $n=2$. In this case the following formula holds:

$$\begin{aligned} &f(U_1)-2f(U_2)+f(U_3)\\ &\quad=2\iiint(\mathcal{D}^2f)(\zeta,\tau,\upsilon)\,dE_1(\zeta)(U_1-U_2)\,dE_2(\tau)(U_2-U_3)\,dE_3(\upsilon)\\ &\qquad+\iint(\mathcal{D}f)(\zeta,\tau)\,dE_1(\zeta)(U_1-2U_2+U_3)\,dE_3(\tau),\end{aligned}$$

where $U_1$, $U_2$, and $U_3$ are unitary operators and $f$ is a function on $\mathbb{T}$ such that the function $\mathcal{D}^2f$ belongs to the space $C(\mathbb{T})\hat{\oplus}_{\mathrm{i}}C(\mathbb{T})\hat{\oplus}_{\mathrm{i}}C(\mathbb{T})$. We refer the reader to [AP2] for the proofs.

Consider now more general classes of functions. Suppose that $\omega$ is a modulus of continuity, i.e., $\omega$ is a nonnegative continuous function on $[0,\infty)$ such that $\omega(0)=0$ and

$$\omega(x+y)\le\omega(x)+\omega(y),\quad x,\,y\ge0.$$

The class $\Lambda_\omega$ consists, by definition, of functions $\varphi$ such that

$$|\varphi(\zeta)-\varphi(\tau)|\le\operatorname{const}\omega(|\zeta-\tau|),\quad \zeta,\,\tau\in\mathbb{T}.$$

We put

$$\|\varphi\|_{\Lambda_\omega}\stackrel{\text{def}}{=}\sup_{\zeta\ne\tau}\frac{|\varphi(\zeta)-\varphi(\tau)|}{\omega(|\zeta-\tau|)}$$

Given a modulus of continuity $\omega$, we define

$$\omega^*(x)=x\int_x^\infty\frac{\omega(t)}{t^2}\,dt.$$

**Theorem 10.3.** *Let $\omega$ be a modulus of continuity and let $U$ and $V$ be unitary operators on Hilbert space. Then for a function $\varphi\in\Lambda_\omega$,*

$$\|\varphi(U)-\varphi(V)\|\le\operatorname{const}\|\varphi\|_{\Lambda_\omega}\omega^*\big(\|U-V\|\big).$$

Note that if $\omega$ is a modulus of continuity, for which

$$\omega^*(x) \le \operatorname{const}\omega(x),$$

then for unitary operators $U$ and $V$ the following inequality holds:

$$\|\varphi(U) - \varphi(V)\| \le \operatorname{const}\|\varphi\|_{\Lambda_\omega}\omega\bigl(\|U - V\|\bigr).$$

We refer the reader to [AP2] for an analog of Theorem 10.3 for higher-order moduli of continuity.

Finally, to conclude this section, I would like to mention that similar results also hold for self-adjoint operators and for contractions. In particular, the analog of Theorem 10.3 for self-adjoint operators improves the estimate obtained in [FN]. We refer the reader to [AP2] for detailed results.

Note that the case of functions of dissipative operators will be treated separately in [AP4].

## 11. Lifshits–Krein trace formulae

The spectral shift function for a trace class perturbation of a self-adjoint operator was introduced in a special case by I.M. Lifshitz [L] and in the general case by M.G. Krein [Kr1]. It was shown in [Kr1] that for a pair of self-adjoint (not necessarily bounded) operators $A$ and $B$ satisfying $B - A \in \boldsymbol{S}_1$, there exists a unique function $\xi \in L^1(\mathbb{R})$ such that

$$\operatorname{trace}\bigl(\varphi(B) - \varphi(A)\bigr) = \int_{\mathbb{R}} \varphi'(x)\xi(x)\,dx, \tag{11.1}$$

whenever $\varphi$ is a function on $\mathbb{R}$ such that the Fourier transform of $\varphi'$ is in $L^1(\mathbb{R})$. The function $\xi$ is called the *spectral shift function corresponding to the pair* $(A, B)$.

A similar result was obtained in [Kr2] for pairs of unitary operators $(U, V)$ with $V - U \in \boldsymbol{S}_1$. For each such pair there exists a function $\xi$ on the unit circle $\mathbb{T}$ of class $L^1(\mathbb{T})$ such that

$$\operatorname{trace}\bigl(\varphi(V) - \varphi(U)\bigr) = \int_{\mathbb{T}} \varphi'(\zeta)\xi(\zeta)\,d\boldsymbol{m}(\zeta), \tag{11.2}$$

whenever $\varphi'$ has absolutely convergent Fourier series. Such a function $\xi$ is unique modulo an additive constant and it is called a *spectral shift function corresponding to the pair* $(U, V)$. We refer the reader to the lectures of M.G. Krein [Kr3], in which the above results were discussed in detail (see also [BS3] and the survey article [BY]).

Note that the spectral shift function plays an important role in perturbation theory. We mention here the paper [BK], in which the following remarkable formula was found:

$$\det S(x) = e^{-2\pi \mathrm{i}\xi(x)},$$

where $S$ is the scattering matrix corresponding to the pair $(A, B)$.

It was shown later in [BS4] that formulae (11.1) and (11.2) hold under less restrictive assumptions on $\varphi$.

Note that the right-hand sides of (11.1) and (11.2) make sense for an arbitrary Lipschitz function $\varphi$. However, it turns out that the condition $\varphi \in \mathrm{Lip}$ (i.e., $\varphi$ is a Lipschitz function) does not imply that $\varphi(B) - \varphi(A)$ or $\varphi(V) - \varphi(U)$ belongs to $\boldsymbol{S}_1$. This is not even true for bounded $A$ and $B$ and continuously differentiable $\varphi$. The first such examples were given in [Fa2].

In this section we present results of [Pe2] and [Pe4] that give necessary conditions and sufficient conditions on the function $\varphi$ for trace formulae (11.1) and (11.2) to hold.

We start with the case of unitary operators. Recall that the class $\mathcal{L}$ of functions on $\mathbb{T}$ was defined in § 7.

**Theorem 11.1.** *Let $\varphi \in C^1(\mathbb{T})$. Suppose that $\varphi \notin \mathcal{L}$. Then there exist unitary operators $U$ and $V$ such that*

$$U - V \in \boldsymbol{S}_1,$$

*but*

$$\varphi(U) - \varphi(V) \notin \boldsymbol{S}_1.$$

**Corollary 11.2.** *Let $\varphi \in C^1(\mathbb{T}) \setminus B_1^1$. Then there exist unitary operators $U$ and $V$ such that*

$$U - V \in \boldsymbol{S}_1,$$

*but*

$$\varphi(U) - \varphi(V) \notin \boldsymbol{S}_1.$$

*Proof of Theorem* 11.1. As we have discussed in § 7, if $\varphi \notin \mathcal{L}$, then the divided difference $\mathfrak{D}\varphi$ is not a multiplier of the class $\mathfrak{N}$ of kernel functions of trace class integral operators on $L^2(\boldsymbol{m})$. Now, the same reasoning as in the proof of Theorem 7.1 allows us to construct sequences of unitary operators $\{U_n\}_{n\ge1}$ and $\{V_n\}_{n\ge1}$ such that

$$\lim_{n\to\infty} \|U_n - V_n\|_{\boldsymbol{S}_1} = 0$$

but

$$\lim_{n\to\infty} \frac{\|\varphi(U_n) - \varphi(V_n)\|_{\boldsymbol{S}_1}}{\|U_n - V_n\|_{\boldsymbol{S}_1}} = \infty.$$

It is easy to see now that we can select certain terms of these sequences with repetition (if necessary) and obtain sequences $\{\mathcal{U}_n\}_{n\ge1}$ and $\{\mathcal{V}_n\}_{n\ge1}$ of unitary operators such that

$$\sum_{n\ge1} \|\mathcal{U}_n - \mathcal{V}_n\|_{\boldsymbol{S}_1} < \infty,$$

but

$$\sum_{n\ge1} \|\varphi(\mathcal{U}_n) - \varphi(\mathcal{V}_n)\|_{\boldsymbol{S}_1} = \infty.$$

Now it remains to define unitary operators $U$ and $V$ by

$$U = \sum_{n\ge1} \oplus\,\mathcal{U}_n \quad \text{and} \quad V = \sum_{n\ge1} \oplus\,\mathcal{V}_n. \qquad \square$$

The following sufficient condition improves earlier results in [BS4].

**Theorem 11.3.** *Suppose that $\varphi \in B^1_{\infty 1}$. Let $U$ and $V$ be unitary operators such that $V - U \in \boldsymbol{S}_1$ and let $\xi$ be a spectral shift function corresponding to the pair $(U, V)$. Then*

$$\varphi(V) - \varphi(U) \in \boldsymbol{S}_1 \tag{11.3}$$

*and trace formula* (11.2) *holds.*

*Proof.* Let us first prove (11.3). By Theorem 6.2, $\mathfrak{D}\varphi \in C(\mathbb{T})\hat{\otimes}C(\mathbb{T})$ which implies that $\mathfrak{D}\varphi \in \mathfrak{M}(E_V, E_U)$. Thus by the Birman–Solomyak formula,

$$\varphi(V) - \varphi(U) = \iint\limits_{\mathbb{T}\times\mathbb{T}} \frac{\varphi(\zeta) - \varphi(\tau)}{\zeta - \tau}\, dE_V(\zeta)\,(U - V)\,dE_U(\tau).$$

It follows that $\varphi(V) - \varphi(U) \in \boldsymbol{S}_1$.

To prove that (11.2) holds, we recall that by the results of [Kr2], (11.2) holds for trigonometric polynomials. It suffices now to approximate $\varphi$ by trigonometric polynomials in the norm of $B^1_{\infty 1}$. □

Let us proceed to the case of self-adjoint operators. The following results were obtained in [Pe4].

**Theorem 11.4.** *Suppose that $\varphi$ is a continuously differentiable function on $\mathbb{R}$ such that $\varphi \notin \mathcal{L}(\mathbb{R})$. Then there exist self-adjoint operators $A$ and $B$ such that*

$$B - A \in \boldsymbol{S}_1,$$

*but*

$$\varphi(B) - \varphi(A) \notin \boldsymbol{S}_1.$$

In particular, Theorem 11.4 implies that the condition that $\varphi \in B^1_1(\mathbb{R})$ locally is a necessary condition for trace formula (11.1) to hold.

**Theorem 11.5.** *Suppose that $\varphi \in B^1_{\infty 1}(\mathbb{R})$. Let $A$ and $B$ be self-adjoint operators* (*not necessarily bounded*) *such that $B - A \in \boldsymbol{S}_1$ and let $\xi$ be the spectral shift function that corresponds to the pair $(A, B)$. Then $\varphi(B) - \varphi(A) \in \boldsymbol{S}_1$ and trace formula* (11.1) *holds.*

The proof of Theorem 11.5 is more complicated than the proof of Theorem 11.3, because nice functions are not dense in $B^1_{\infty 1}(\mathbb{R})$, and to prove (11.1) we have to use a weak approximation, see [Pe4].

## 12. Koplienko–Neidhardt trace formulae

In this section we consider trace formulae in the case of perturbations of Hilbert–Schmidt class $\boldsymbol{S}_2$.

Let $A$ and $B$ be self-adjoint operators such that $K \stackrel{\text{def}}{=} B - A \in \boldsymbol{S}_2$. In this case the operator $\varphi(B) - \varphi(A)$ does not have to be in $\boldsymbol{S}_1$ even for very nice functions $\varphi$.

The idea of Koplienko in [Ko] was to consider the operator

$$\varphi(B)-\varphi(A)-\frac{d}{ds}\Big(\varphi(A+sK)\Big)\Big|_{s=0}$$

and find a trace formula under certain assumptions on $\varphi$. It was shown in [Ko] that there exists a unique function $\eta\in L^1(\mathbb{R})$ such that

$$\operatorname{trace}\left(\varphi(B)-\varphi(A)-\frac{d}{ds}\Big(\varphi(A+sK)\Big)\Big|_{s=0}\right)=\int_{\mathbb{R}}\varphi''(x)\eta(x)\,dx \tag{12.1}$$

for rational functions $\varphi$ with poles off $\mathbb{R}$. The function $\eta$ is called the *generalized spectral shift function corresponding to the pair* $(A,B)$.

A similar problem for unitary operators was considered by Neidhardt in [Ne]. Let $U$ and $V$ be unitary operators such that $V-U\in\boldsymbol{S}_2$. Then $V=\exp(\mathrm{i}A)U$, where $A$ is a self-adjoint operator in $\boldsymbol{S}_2$. Put $U_s=e^{\mathrm{i}sA}U$, $s\in\mathbb{R}$. It was shown in [Ne] that there exists a function $\eta\in L^1(\mathbb{T})$ such that

$$\operatorname{trace}\left(\varphi(V)-\varphi(U)-\frac{d}{ds}\Big(\varphi(U_s)\Big)\Big|_{s=0}\right)=\int_{\mathbb{T}}\varphi''\eta\,d\boldsymbol{m}, \tag{12.2}$$

whenever $\varphi''$ has absolutely convergent Fourier series. Such a function $\eta$ is unique modulo a constant and it is called a *generalized spectral shift function corresponding to the pair* $(U,V)$.

We state in this section results of [Pe7] that guarantee the validity of trace formulae (12.1) and (12.2) under considerably less restrictive assumptions on $\varphi$.

**Theorem 12.1.** *Suppose that $U$ and $V=e^{\mathrm{i}A}U$ are unitary operators on Hilbert space such that $U-V\in\boldsymbol{S}_2$. Let $\varphi\in B^2_{\infty1}$. Then*

$$\varphi(V)-\varphi(U)-\frac{d}{ds}\Big(\varphi\big(e^{\mathrm{i}sA}U\big)\Big)\Big|_{s=0}\in\boldsymbol{S}_1$$

*and trace formula* (12.2) *holds.*

**Theorem 12.2.** *Suppose that $A$ and $B$ are self-adjoint operators* (*not necessarily bounded*) *on Hilbert space such that $K=B-A\in\boldsymbol{S}_2$. Let $\varphi\in B^2_{\infty1}(\mathbb{R})$. Then*

$$\varphi(B)-\varphi(A)-\frac{d}{ds}\Big(\varphi(A+sK)\Big)\in\boldsymbol{S}_1$$

*and trace formula* (12.1) *holds.*

## 13. Perturbations of class $\boldsymbol{S}_p$

In the final section of this survey article we consider the problem of the behavior of the function of an operator under perturbations by operators of Schatten–von Neumann class $\boldsymbol{S}_p$. In § 11 we have already considered the special case of perturbations of trace class. We have seen that the condition $\varphi\in\mathrm{Lip}$ (i.e., $\varphi$ is a Lipschitz function) does not guarantee that trace class perturbations of an operator lead to trace class changes of the function of the operator.

On the other hand, Theorem 2.2 shows that for a Lipschitz function $\varphi$ the condition $A-B\in\boldsymbol{S}_2$ implies that $\varphi(A)-\varphi(B)\in\boldsymbol{S}_2$.

In this section we discuss the results obtained in [Pe3] that deal with perturbations of class $\boldsymbol{S}_p$ with $p<1$. Then we state the results of [AP3] that discuss the behavior of functions of class $\Lambda_\alpha$ under perturbations by operators of Schatten–von Neumann classes $\boldsymbol{S}_p$. Finally, we mention recent results of [NP] and [PS].

In the case $p<1$ the following results were found in [Pe3]:

**Theorem 13.1.** *Let $0<p<1$ and let $\varphi\in B_{\infty p}^{1/p}$. Suppose that $U$ and $V$ are unitary operators such that $U-V\in\boldsymbol{S}_p$. Then $\varphi(U)-\varphi(V)\in\boldsymbol{S}_p$.*

**Theorem 13.2.** *Let $0<p<1$. Suppose that $\varphi$ is a continuously differentiable function on $\mathbb{T}$ such that $\varphi(U)-\varphi(V)\in\boldsymbol{S}_p$, whenever $U$ and $V$ are unitary operators such that $U-V\in\boldsymbol{S}_p$. Then $\varphi\in B_p^{1/p}$.*

As in the case $p=1$, Theorem 13.2 can be improved: under the hypotheses of Theorem 13.2, the Hankel operators $H_\varphi$ and $H_{\bar\varphi}$ must map the Hardy class $H^1$ into the Besov space $B_p^{1/p}$.

The same results also hold for contractions and analogs of these results can also be obtained for bounded self-adjoint operators (in the analog of Theorem 13.2 for self-adjoint operators the conclusion is that $\varphi$ belongs to $B_p^{1/p}$ locally).

We proceed now to the results of [AP3] (see also [AP1]) that describe the behavior of $\varphi(U)$ for functions of class $\Lambda_\alpha$ under perturbations of $U$ by operators of class $\boldsymbol{S}_p$.

**Definition.** Let $p>0$. We say that a compact operator $T$ belongs to the ideal $\boldsymbol{S}_{p,\infty}$ if its singular values $s_n(T)$ satisfies the estimate:

$$\|T\|_{\boldsymbol{S}_{p,\infty}}\stackrel{\text{def}}{=}\sup_{n\ge0}s_n(T)(1+n)^{1/p}<\infty.$$

Clearly,

$$\boldsymbol{S}_p\subset\boldsymbol{S}_{p,\infty}\subset\boldsymbol{S}_q$$

for any $q>p$. Note that $\|\cdot\|_{\boldsymbol{S}_{p,\infty}}$ is not a norm, though for $p>1$, the space $\boldsymbol{S}_{p,\infty}$ has a norm equivalent to $\|\cdot\|_{\boldsymbol{S}_{p,\infty}}$.

**Theorem 13.3.** *Let $p\ge1$, $0<\alpha<1$, and let $\varphi\in\Lambda_\alpha$. Suppose that $U$ and $V$ are unitary operators on Hilbert space such that $U-V\in\boldsymbol{S}_p$. Then*

$$\varphi(U)-\varphi(V)\in\boldsymbol{S}_{\frac{p}{\alpha},\infty}$$

*and*

$$\|\varphi(U)-\varphi(V)\|_{\boldsymbol{S}_{\frac{p}{\alpha},\infty}}\le\operatorname{const}\|f\|_{\Lambda_\alpha}\|B-A\|_{\boldsymbol{S}_p}^{\alpha}.$$

In the case when $p>1$ Theorem 13.3 can be improved by using interpolation arguments.

**Theorem 13.4.** *Let $p > 1$, $0 < \alpha < 1$, and let $\varphi \in \Lambda_\alpha$. Suppose that $U$ and $V$ are unitary operators on Hilbert space such that $U - V \in \boldsymbol{S}_p$. Then*

$$\varphi(U) - \varphi(V) \in \boldsymbol{S}_p$$

*and*

$$\|\varphi(U) - \varphi(V)\|_{\boldsymbol{S}_p} \le \operatorname{const} \|f\|_{\Lambda_\alpha} \|B - A\|^\alpha_{\boldsymbol{S}_p}.$$

Note that the constants in the above inequalities depend on $\alpha$.

Let us sketch the proof of Theorem 13.3. We refer the reader to [AP3] for a detailed proof.

As in the proof of Theorem 10.1, we assume that $\varphi \in \big(\Lambda_\alpha\big)_+$ and we consider the expansion (10.1). Put

$$Q_N = \sum_{n \le N} \big(\varphi_n(U) - \varphi_n(V)\big) \quad \text{and} \quad R_N = \sum_{n > N} \big(\varphi_n(U) - \varphi_n(V)\big).$$

Then

$$\|R_N\| \le 2 \sum_{n \ge N} \|\varphi_n\|_{L^\infty} \le \operatorname{const} 2^{-\alpha N} \|\varphi\|_{\Lambda_\alpha}.$$

It follows from Lemma 6.1 that

$$\|\varphi_n(U) - \varphi_n(V)\|_{\boldsymbol{S}_p} \le \operatorname{const} 2^n \|\varphi_n\|_{L^\infty} \|U - V\|_{\boldsymbol{S}_p}$$

which implies that

$$\|Q_N\|_{\boldsymbol{S}_p} \le \operatorname{const} 2^{(1-\alpha)N} \|\varphi\|_{\Lambda_\alpha} \|U - V\|_{\boldsymbol{S}_p}.$$

The proof can easily be completed on the basis of the following estimates:

$$s_n(Q_N) \le (1+n)^{-1/p} \|Q_N\|_{\boldsymbol{S}_p}$$

and

$$s_n\big(\varphi(U) - \varphi(V)\big) \le s_n(Q_N) + \|R_N\|. \qquad \square$$

Consider now the case of higher-order differences.

**Theorem 13.5.** *Let $0 < \alpha < n$ and $p \ge n$. Suppose that $U$ is a unitary operator and $A$ is a self-adjoint operator of class $\boldsymbol{S}_p$. Then*

$$\sum_{k=0}^n (-1)^k \binom{n}{k} \varphi\big(e^{\mathrm{i}kA} U\big) \in \boldsymbol{S}_{\frac{p}{\alpha},\infty}$$

*and*

$$\left\| \sum_{k=0}^n (-1)^k \binom{n}{k} \varphi\big(e^{\mathrm{i}kA} U\big) \right\|_{\boldsymbol{S}_{\frac{p}{\alpha},\infty}} \le \operatorname{const} \|f\|_{\Lambda_\alpha} \|A\|^\alpha_{\boldsymbol{S}_p}.$$

Again, if $p > n$, Theorem 13.5 can be improved by using interpolation arguments.

**Theorem 13.6.** *Let $0 < \alpha < n$ and $p > n$. Suppose that $U$ is a unitary operator and $A$ is a self-adjoint operator of class $\boldsymbol{S}_p$. Then*

$$\sum_{k=0}^{n}(-1)^k \binom{n}{k} \varphi\big(e^{\mathrm{i}kA}U\big) \in \boldsymbol{S}_{\frac{p}{\alpha}}$$

*and*

$$\left\|\sum_{k=0}^{n}(-1)^k \binom{n}{k} \varphi\big(e^{\mathrm{i}kA}U\big)\right\|_{\boldsymbol{S}_{\frac{p}{\alpha}}} \le \operatorname{const}\|f\|_{\Lambda_\alpha}\|A\|_{\boldsymbol{S}_p}^{\alpha}.$$

We refer the reader to [AP3] for the proofs of Theorems 13.5 and 13.6.

Note that similar results also hold for contractions and for self-adjoint operators. Analogs of these results for dissipative operators will be given in [AP4].

To conclude this section, we mention briefly recent results of [NP] and [PS].

The following results have been obtained in [NP]:

**Theorem 13.7.** *Let $f$ be a Lipschitz function on $\mathbb{R}$, and let $A$ and $B$ be (not necessarily bounded) self-adjoint operators such that* $\operatorname{rank}(A-B) = 1$. *Then $f(A)-f(B) \in \boldsymbol{S}_{1,\infty}$ and*

$$\|f(A) - f(B)\|_{\boldsymbol{S}_{1,\infty}} \le \operatorname{const}\|f\|_{\mathrm{Lip}}\|A - B\|.$$

This implies the following result (see[NP]):

**Theorem 13.8.** *Let $f$ be a Lipschitz function on $\mathbb{R}$, and let $A$ and $B$ be (not necessarily bounded) self-adjoint operators such that $A - B \in \boldsymbol{S}_1$. Then $f(A) - f(B) \in \boldsymbol{S}_\Omega$, i.e.,*

$$\sum_{j=0}^{n} s_n\big(f(A) - f(B)\big) \le \operatorname{const}\log(n+2)\|f\|_{\mathrm{Lip}}\|A - B\|_{\boldsymbol{S}_1}.$$

It is still unknown whether the assumptions that $f \in \mathrm{Lip}$ and $A - B \in \boldsymbol{S}_1$ imply that $f(A) - f(B) \in \boldsymbol{S}_{1,\infty}$. The results of [NP] imply that if $1 \le p < \infty$, $\varepsilon > 0$, $f \in \mathrm{Lip}$, and $A - B \in \boldsymbol{S}_p$, then $f(A) - f(B) \in \boldsymbol{S}_{p+\varepsilon}$.

In the case $1 < p < \infty$ the last result has been improved recently in [PS]:

**Theorem 13.9.** *Let $1 < p < \infty$, $f \in \mathrm{Lip}$, and let $A$ and $B$ be self-adjoint operators such that $A - B \in \boldsymbol{S}_p$. Then $f(A) - f(B) \in \boldsymbol{S}_p$*

## References

[AP1] A.B. Aleksandrov and V.V. Peller, *Functions of perturbed operators*, C.R. Acad. Sci. Paris, Sér. I 347 (2009), 483–488.

[AP2] A.B. Aleksandrov and V.V. Peller, *Operator Hölder–Zygmund functions*, to appear in Advances in Mathematics.

[AP3] A.B. Aleksandrov and V.V. Peller, *The behavior of functions of operators under perturbations of class $\boldsymbol{S}_p$*, J. Funct. Anal. 258 (2010), 3675–3724.

[AP4] A.B. Aleksandrov and V.V. Peller, *Functions of perturbed dissipative operators*, to appear.

[ABF] J. Arazy, T. Barton, and Y. Friedman, *Operator differentiable functions*, Int. Equat. Oper. Theory 13 (1990), 462–487.

[ACDS] N.A. Azamov, A.L. Carey, P.G. Dodds, and F.A. Sukochev, *Operator integrals, spectral shift and spectral flow*, Canad. J. Math. 61 (2009), 241–263.

[Be] G. Bennett, *Schur multipliers*, Duke Math. J. 44 (1977), 603–639.

[BK] M.S. Birman and M.G. Krein, *On the theory of wave operators and scattering operators*, Dokl. Akad. Nauk SSSR 144 (1962), 475–478.
English transl.: Sov. Math. Dokl. 3 (1962), 740–744.

[BS1] M.S. Birman and M.Z. Solomyak, *Double Stieltjes operator integrals*, Problems of Math. Phys., Leningrad. Univ. 1 (1966), 33–67 (Russian).
English transl.: Topics Math. Physics 1 (1967), 25–54, Consultants Bureau Plenum Publishing Corporation, New York.

[BS2] M.S. Birman and M.Z. Solomyak, *Double Stieltjes operator integrals. II*, Problems of Math. Phys., Leningrad. Univ. 2 (1967), 26–60 (Russian).
English transl.: Topics Math. Physics 2 (1968), 19–46, Consultants Bureau Plenum Publishing Corporation, New York.

[BS3] M.S. Birman and M.Z. Solomyak, *Remarks on the spectral shift function*, Zapiski Nauchn. Semin. LOMI 27 (1972), 33–46 (Russian).
English transl.: J. Soviet Math. 3 (1975), 408–419.

[BS4] M.S. Birman and M.Z. Solomyak, *Double Stieltjes operator integrals. III*, Problems of Math. Phys., Leningrad. Univ. 6 (1973), 27–53 (Russian).

[BS5] M.S. Birman and M.Z. Solomyak, *Tensor product of a finite number of spectral measures is always a spectral measure*, Integral Equations Operator Theory 24 (1996), 179–187.

[BS6] M.S. Birman and M.Z. Solomyak, *Double operator integrals in Hilbert space*, Int. Equat. Oper. Theory 47 (2003), 131–168.

[BY] M.S. Birman and D.R. Yafaev, *The spectral shift function. The papers of M.G. Kreĭn and their further development*, Algebra i Analiz 4 (1992), 1–44 (Russian).
English transl.: St. Petersburg Math. J. 4 (1993), 833–870.

[DK] Yu.L. Daletskii and S.G. Krein, *Integration and differentiation of functions of Hermitian operators and application to the theory of perturbations* (Russian), Trudy Sem. Functsion. Anal., Voronezh. Gos. Univ. 1 (1956), 81–105.

[Fa1] Yu.B. Farforovskaya, *The connection of the Kantorovich-Rubinshtein metric for spectral resolutions of selfadjoint operators with functions of operators*, Vestnik Leningrad. Univ. 19 (1968), 94–97. (Russian).

[Fa2] Yu.B. Farforovskaya, *An example of a Lipschitzian function of selfadjoint operators that yields a nonnuclear increase under a nuclear perturbation.* Zap. Nauchn. Sem. Leningrad. Otdel. Mat. Inst. Steklov. (LOMI) 30 (1972), 146–153 (Russian).

[Fa3] Yu.B. Farforovskaya, *An estimate of the norm of* $| f(B) - f(A) |$ *for selfadjoint operators* $A$ *and* $B$, Zap. Nauchn. Sem. Leningrad. Otdel. Mat. Inst. Steklov. (LOMI) 56 (1976), 143–162 (Russian).

[FN] Yu.B. Farforovskaya and L. Nikolskaya, *Modulus of continuity of operator functions*, Algebra i Analiz 20:3 (2008), 224–242.

[FrJ] M. Frazier and B. Jawerth, *A discrete transform and decompositions of distribution spaces*, J. Funct. Anal. 93 (1990), 34–170.

[JTT] K. Jushchenko, I.G. Todorov, and L. Turowska, *Multidimensional operator multipliers*, Trans. Amer. Math. Soc. 361 (2009), 4683–4720.

[KA] T. Kato, *Continuity of the map $S \mapsto | S |$ for linear operators,* Proc. Japan Acad. 49 (1973), 157–160.

[KS] E. Kissin and V.S. Shulman, *On fully operator Lipschitz functions*, J. Funct. Anal. 253 (2007), 711–728.

[Ko] L.S. Koplienko, *The trace formula for perturbations of nonnuclear type*, Sibirsk. Mat. Zh. 25:5 (1984), 62–71 (Russian).
English transl.: Sib. Math. J. 25 (1984), 735–743.

[Kr1] M.G. Krein, *On a trace formula in perturbation theory*, Mat. Sbornik 33 (1953), 597–626 (Russian).

[Kr2] M.G. Krein, *On perturbation determinants and a trace formula for unitary and self-adjoint operators*, Dokl. Akad. Nauk SSSR 144 (1962) 268–271 (Russian).
English transl. in: Topics in integral and differential equations and operator theory, Birkhäuser, Basel, 1983, 107–172.

[Kr3] M.G. Krein, *On some new investigations in the perturbation theory of self-adjoint operators*, in: The First Summer Math. School, Kiev, 1964, 103–187 (Russian).

[L] I.M. Lifshitz, *On a problem in perturbation theory connected with quantum statistics*, Uspekhi Mat. Nauk 7 (1952), 171–180 (Russian).

[MM] M.M. Malamud and S.M. Malamud, *Spectral theory of operator measures in a Hilbert space*, Algebra i Analiz 15:3 (2003), 1–77.
English transl.: St. Petersburg Math. J. 15:3 (2004), 323–373.

[Na] M.A. Naimark, *Spectral functions of symmetric operator*, Izvestia Akad. Nauk SSSR, Ser. Matem. 4:3 (1940), 277–318 (Russian).

[Ne] H. Neidhardt, *Spectral shift function and Hilbert–Schmidt perturbation: extensions of some work of L.S. Koplienko*, Math. Nachr. 138 (1988), 7–25.

[NP] F.L. Nazarov and V.V. Peller, *Lipschitz functions of perturbed operators*, C.R. Acad. Sci. Paris, Sér. I 347 (2009), 857–862.

[Pa] B.S. Pavlov, *On multiple operator integrals*, Problems of Math. Anal., No. 2: Linear Operators and Operator Equations (Russian), 99–122. Izdat. Leningrad. Univ., Leningrad, 1969.

[Pee] J. Peetre, *New thoughts on Besov spaces*, Duke Univ. Press., Durham, NC, 1976.

[Pe1] V.V. Peller, *Hankel operators of class $S_p$ and their applications* (*rational approximation, Gaussian processes, the problem of majorizing operators*), Mat. Sbornik, 113 (1980), 538–581.
English Transl. in Math. USSR Sbornik, 41 (1982), 443–479.

[Pe2] V.V. Peller, *Hankel operators in the theory of perturbations of unitary and self-adjoint operators*, Funktsional. Anal. i Prilozhen. 19:2 (1985), 37–51 (Russian).
English transl.: Funct. Anal. Appl. 19 (1985), 111–123.

[Pe3] V.V. Peller, *For which $f$ does $A - B \in S_p$ imply that $f(A) - f(B) \in S_p$?*, Operator Theory, Birkhäuser, 24 (1987), 289-294.

[Pe4] V.V. Peller *Hankel operators in the perturbation theory of unbounded self-adjoint operators.* Analysis and partial differential equations, 529–544, Lecture Notes in Pure and Appl. Math., 122, Dekker, New York, 1990.

[Pe5] V.V. Peller, *Functional calculus for a pair of almost commuting selfadjoint operators*, J. Funct. Anal., 112 (1993), 325–345.

[Pe6] V.V. Peller, *Hankel operators and their applications,* Springer-Verlag, New York, 2003.

[Pe7] V.V. Peller, *An extension of the Koplienko–Neidhardt trace formulae*, J. Funct. Anal. 221 (2005), 456–481.

[Pe8] V.V. Peller, *Multiple operator integrals and higher operator derivatives*, J. Funct. Anal. 233 (2006), 515–544.

[Pe9] V.V. Peller, *Differentiability of functions of contractions*, In: Linear and complex analysis, AMS Translations, Ser. 2 226 (2009), 109–131, AMS, Providence.

[PS] D. Potapov and F. Sukochev, *Operator-Lipschitz functions in Schatten–von Neumann classes*, arXiv:0904.4095, 2009.

[St] V.V. Sten'kin, *Multiple operator integrals*, Izv. Vyssh. Uchebn. Zaved. Matematika 4 (79) (1977), 102–115 (Russian).
English transl.: Soviet Math. (Iz. VUZ) 21:4 (1977), 88–99.

[SNF] B. Sz.-Nagy and C. Foiaş, *Harmonic analysis of operators on Hilbert space,* Akadémiai Kiadó, Budapest, 1970.

[T] H. Triebel, *Theory of function spaces,* Monographs in Mathematics, 78, Birkhäuser Verlag, Basel, 1983.

[W] H. Widom, *When are differentiable functions differentiable?*, In: Linear and Complex Analysis Problem Book, Lect. Notes Math. 1043 (1984), 184–188.

V.V. Peller
Mathematics Department
Michigan State University
East Lansing, Michigan 48824, USA
e-mail: peller@math.msu.edu

Operator Theory:
Advances and Applications, Vol. 207, 325–339

# The Halmos Similarity Problem

Gilles Pisier

**Abstract.** We describe the ideas that led to the formulation of the problem and its solution, as well as some related questions left open.

**Mathematics Subject Classification (2000).** Primary 47A20,47B35; Secondary 46L07,47L25.

**Keywords.** Similarity, polynomially bounded, completely bounded, Hankel operator, length.

The Halmos similarity problem is the sixth one in Halmos's list in [26]. The origin of the problem clearly goes back to Sz.-Nagy's works [59, 60] on power bounded operators. An operator $T\colon\ H \to H$ on a Hilbert space $H$ is called power bounded if $\sup_{n\geq 1} \|T^n\| < \infty$. When $\|T\| \leq 1, T$ is called a contraction. An operator $T_2\colon\ H \to H$ is called similar to an operator $T_1\colon\ H \to H$ if there is an invertible operator $\xi\colon\ H \to H$ such that $\xi T_1 \xi^{-1} = T_2$. Obviously any $T$ that is similar to a contraction must be power bounded.

Sz.-Nagy [59, 60] proved that the converse holds in several cases, e.g., when $T$ is compact. The converse also holds (as proved in 1960 by Rota [58]), if $\sigma(T)$ (the spectrum of $T$) is included in the open unit disc $D \subset \mathbb{C}$, or equivalently if the spectral radius of $T$ is $< 1$. Quite naturally, Sz.-Nagy asked in [60] whether the converse held in general, namely whether any power bounded $T$ had to be similar to a contraction. Sz.-Nagy's quest for a criterion for "similarity to a contraction" was almost surely motivated by his own work with F. Riesz in ergodic theory (see [57, p. 402]) where contractions play a key role.

In 1964, Foguel gave in [19] the first counterexample, i.e., power bounded but *not* similar to a contraction. Shortly after it circulated, Halmos published his version of it in "On Foguel's answer to Sz.-Nagy's question" [25]. Foguel's example is as follows: Let

$$T = \begin{pmatrix} S^* & Q \\ 0 & S \end{pmatrix} \in B(\ell_2 \oplus \ell_2),$$

where $S$ is the unilateral shift on $\ell_2$ (i.e., $Se_n = e_{n+1}$) and $Q$ is the orthogonal projection onto the span of $\{e_{n(k)}\}$ for a Hadamard lacunary sequence $\{n(k)\}$, i.e., such that $\inf_k\{n(k+1)/n(k)\} > 1$ (Foguel uses $n(k) = 4^k$). Then $T$ is power bounded but *not* similar to a contraction.

*Remark* 1. Other power bounded examples appeared later on, notably in [15, 9, 48]. Peller [48] asked whether, for any $c > 1$, any power bounded operator is similar to one with powers bounded by $c$. A negative answer is given in [28] (see [53] for related work). In analogy with Foguel's work, the similarity problem for continuous parameter semigroups of operators is considered in [40, 32].

It must have been when Foguel's paper appeared that polynomial boundedness came into the spotlight. An operator $T\colon\ H \to H$ is called polynomially bounded (in short p.b.) if there is a constant $C$ such that for any polynomial $P$ we have

$$\|P(T)\| \leq C\|P\|_\infty$$

where

$$\|P\|_\infty = \sup\{|P(z)| \mid |z| \leq 1\}.$$

Here of course, for any polynomial $P(z) = a_0 + a_1 z + \cdots + a_n z^n$ we set $P(T) = a_0 I + a_1 T + \cdots + a_n T^n$. We will denote by $C_{pb}(T)$ the smallest $C$ for which this holds (note $C_{pb}(T) \geq 1$ because of $P \equiv 1$). Obviously by spectral theory, a normal operator $T$ is p.b. iff $\|T\| \leq 1$ and then $C_{pb}(T) = 1$. But it is much less obvious that actually any $T$ with $\|T\| \leq 1$ (i.e., any "contraction") is p.b. with $C_{pb}(T) = 1$. This is a celebrated inequality due to von Neumann [38]: if $\|T\| \leq 1$ we have

$$\forall P \text{ polynomial} \qquad \|P(T)\| \leq \|P\|_\infty. \tag{1}$$

Incidentally, by [20], Hilbert spaces are the only Banach spaces satisfying this. From von Neumann's inequality, it is clear that any $T$ that is similar to a contraction, i.e., $T = \xi T_1 \xi^{-1}$ with $\|T_1\| \leq 1$, must be p.b. and (since $P(T) = \xi P(T_1)\xi^{-1}$) satisfies

$$\|P(T)\| \leq \|\xi\|\|\xi^{-1}\|\|P\|_\infty.$$

Equivalently, we have

$$C_{pb}(T) \leq \inf\{\|\xi\|\|\xi^{-1}\| \mid \|\xi T \xi^{-1}\| \leq 1\}. \tag{2}$$

Our guess is that, immediately after Foguel's example, the idea to replace power bounded by polynomially bounded must have already popped up. In 1968 Lebow [31] showed that Foguel's example was not p.b. It thus became conceivable that the "right" problem to ask was: Is every p.b. operator similar to a contraction? This was put in writing by Halmos in [26], but, presumably, had circulated verbally earlier than that (Lebow mentions the problem but without reference).

Actually, in (1) and (12) a stronger result called a "dilation theorem" holds (see [6, 46] for more on dilation theory):

**Theorem 2 (Sz.-Nagy's dilation).** *For any $T\colon\ H \to H$, $\|T\| \leq 1$ there is a larger Hilbert space $\widehat{H} \supset H$ and a unitary $\widehat{T}\colon\ \widehat{H} \to \widehat{H}$ such that*

$$\forall n \geq 1 \qquad T^n = P_H \widehat{T}^n{}_{|H}.$$

In that case $\widehat{T}$ is called a "strong dilation" of $T$. (One says that $\widehat{T}$ is a dilation if this merely holds for $n = 1$.)

We will discuss below (see (12)) the possible extension of this to several commuting operators.

In 1984, Paulsen [43] made a crucial step. Transplanting a result proved by Haagerup [24] for $C^*$-algebras from the self-adjoint to the non-self-adjoint case, he obtained the first (and in some sense the only!) criterion for similarity of a bounded operator $T\colon\ H \to H$ to a contraction: *complete polynomial boundedness.*

**Definition.** An operator $T\colon\ H \to H$ is called completely polynomially bounded (c.p.b. in short) if there is a constant $C$ such that for any $n$ and any matrix-valued polynomial $[P_{ij}(z)]_{1\le i,j\le n}$ we have

$$\|[P_{ij}(T)]\| \le C \sup\{\|[P_{ij}(z)]\|_{M_n} \mid z \in \mathbb{C}, |z| \le 1\}, \tag{3}$$

where the norm of the operator-valued matrix $[P_{ij}(T)]$ is computed on $H \oplus \cdots \oplus H$ ($n$ times) and $\|.\|_{M_n}$ is the usual norm of an $n \times n$ matrix with complex entries. We will (temporarily) denote by $C_{cpb}(T)$ the smallest $C$ such that this holds.

Equivalently, using the identifications $M_n = B(\ell_2^n)$ and $H \oplus \cdots \oplus H = \ell_2^n \otimes H$, we see that $C_{cpb}(T)$ is the smallest constant $C$ for which, for any $n$ and any finite sequence $a_k$ in $M_n = B(\ell_2^n)$, we have

$$\left\|\sum a_k \otimes T^k\right\|_{B(\ell_2^n \otimes H)} \le C \sup\nolimits_{z \in \bar{D}} \left\|\sum a_k z^k\right\|_{B(\ell_2^n)}.$$

This is easy to generalize with an arbitrary Hilbert space $K$ in place of $\ell_2^n$. Thus we have for any finite set of coefficients $a_k \in B(K)$

$$\left\|\sum a_k \otimes T^k\right\|_{B(K \otimes H)} \le C \sup_{z \in \bar{D}} \left\|\sum a_k z^k\right\|_{B(K)}. \tag{4}$$

**Theorem 3 (Paulsen's criterion).** *An operator $T\colon\ H \to H$ is similar to a contraction iff it is c.p.b. and moreover*

$$C_{cpb}(T) = \inf\{\|\xi\|\|\xi^{-1}\| \mid \|\xi T \xi^{-1}\| \le 1\}.$$

The "only if" part is easy: by Sz.-Nagy's dilation theorem for any contraction $T_1$ we can write

$$P_{ij}(T_1) = P_H P_{ij}(\widehat{T}_1)_{|H}$$

but from the spectral theory of unitary operators one deduces easily

$$\|[P_{ij}(\widehat{T}_1)]\| = \sup\{\|[P_{ij}(z)]\| \mid z \in \sigma(\widehat{T}_1)\} \le \sup\{\|[P_{ij}(z)]\| \mid z \in \mathbb{C}, |z| = 1\},$$

and hence $C_{cpb}(T_1) \le 1$ for any contraction $T_1$. But if $T = \xi^{-1} T_1 \xi$ then obviously

$$C_{cpb}(T) \le \|\xi\|\|\xi^{-1}\| C_{cpb}(T_1) \le \|\xi\|\|\xi^{-1}\|.$$

This yields

$$C_{cpb}(T) \le \inf\{\|\xi\|\|\xi^{-1}\| \mid \|\xi T \xi^{-1}\| \le 1\}.$$

The converse is more delicate. It uses the factorization of completely bounded maps for which we need some background.

We start by an elementary fact.

**Proposition.** *Let $A$ be any $C^*$-algebra. Then the algebra $M_n(A)$ of $n \times n$ matrices with entries in $A$ can be equipped with a unique $C^*$-algebra norm. When $A \subset B(H)$ (isometric $*$-homomorphism) the norm in $M_n(A)$ coincides with the norm induced by the space $B(H \oplus \cdots \oplus H)$ (where $H$ is repeated $n$ times).*

**Definition.** Let $A, B$ be $C^*$-algebras. Consider subspaces $E \subset A$ and $F \subset B$ and a map $u\colon\ E \to F$. We say that $u$ is completely bounded (in short c.b.) if $\|u\|_{cb} < \infty$, where

$$\|u\|_{cb} = \sup_{n \geq 1} \|u_n\colon\ M_n(E) \to M_n(F)\|,$$

where $M_n(E)$ (resp. $M_n(F)$) denotes the space of $n \times n$ matrices with entries in $E$, equipped with the norm induced by $M_n(A)$ (resp. $M_n(B)$) and where $u_n$ is the linear map defined by

$$u_n([a_{ij}]) = [u(a_{ij})].$$

We denote by $CB(E, F)$ the space of all c.b. maps $u\colon\ E \to F$.

Equipped with $\|\cdot\|_{cb}$, the space $CB(E, F)$ becomes a normed space (a Banach space if $F$ is closed in $B(K)$).

Let $D = \{z \in \mathbb{C} \mid |z| < 1\}$ and let $A(D)$ denote the disc algebra, i.e., the algebra of all analytic functions on $D$ that extend continuously to $\overline{D}$. Let $\mathcal{P} \subset A(D)$ be the dense subalgebra formed by the (analytic) polynomials. We equip $A(D)$ as usual with the norm

$$\forall f \in A(D) \qquad \|f\| = \sup_{z \in D} |f(z)| = \sup_{z \in \overline{D}} |f(z)| = \sup_{z \in \partial D} |f(z)|.$$

For any bounded $T\colon\ H \to H$, let us denote by $u_T\colon\ \mathcal{P} \to B(H)$ the homomorphism corresponding to the classical spectral calculus, i.e., we set

$$u_T(P) = P(T).$$

It is then obvious that

$$C_{pb}(T) = \|u_T\| \quad \text{and} \quad C_{cpb}(T) = \|u_T\|_{cb}.$$

Note that since the unit ball of $A(D)$ is the closed convex hull of the finite Blaschke products (see [23, p. 196]), $\|u_T\| = \sup\{\|B(T)\|\}$ where the supremum runs over all finite Blaschke products $B$.

The general form of Paulsen's result in [43] is as follows:

**Theorem 4.** *Let $\mathcal{A}$ be a unital subalgebra of a $C^*$-algebra $A$. Consider a unital homomorphism $u\colon\ \mathcal{A} \to B(H)$ and a constant $C$. The following are equivalent:*

(i) $\|u\|_{cb} \leq C$.

(ii) *There is an invertible $\xi\colon\ H \to H$ with $\|\xi\|\|\xi^{-1}\| \leq C$ such that the homomorphism $a \mapsto \xi u(a) \xi^{-1}$ satisfies $\|\xi u(\cdot)\xi^{-1}\|_{cb} = 1$.*

(iii) *There is an invertible $\xi\colon\ H \to H$ with $\|\xi\|\|\xi^{-1}\| \leq C$ a Hilbert space $\widehat{H} \supset H$ and a (necessarily contractive) $*$-homomorphism $\widehat{u}\colon\ A \to B(\widehat{H})$ such that*

$$\forall a \in \mathcal{A} \qquad u(a) = \xi^{-1} P_H \widehat{u}(a)_{|H} \xi.$$

This last statement resets similarity problems in a much broader algebraic framework. Already in [26], Halmos gave a slightly more "algebraic" description of Foguel's example. This direction was amplified by remarkable joint work by Foias and Williams that long remained unpublished (probably due to Williams's untimely death). Independently, Carlson and Clark (see [11, 12]) developed ideas inspired by the homology of Hilbert modules. They studied short exact sequences in the category of Hilbert modules over the disc algebra. Later on, they joined forces with Foias and published [10].

*Remark* 5. Let $u_1\colon\ \mathcal{A} \to B(H_1)$ and $u_2\colon\ \mathcal{A} \to B(H_2)$ be two (unital) homomorphisms and let $D\colon\ \mathcal{A} \to B(H_2, H_1)$ be a "derivation," i.e., we have

$$\forall a, b \in \mathcal{A} \qquad D(ab) = u_1(a)D(b) + D(a)u_2(b).$$

Then direct verification shows that

$$u(a) = \begin{pmatrix} u_1(a) & D(a) \\ 0 & u_2(a) \end{pmatrix} \tag{5}$$

is a (unital) homomorphism from $\mathcal{A}$ to $B(H_1 \oplus H_2)$. The derivation $D$ is called "inner" if there is $T$ in $B(H_2, H_1)$ such that

$$\forall a \in \mathcal{A} \qquad D(a) = u_1(a)T - Tu_2(a).$$

In that case, we have

$$u(a) = \begin{pmatrix} u_1(a) & D(a) \\ 0 & u_2(a) \end{pmatrix} = \xi \begin{pmatrix} u_1(a) & 0 \\ 0 & u_2(a) \end{pmatrix} \xi^{-1}$$

where $\xi = \left(\begin{smallmatrix} 1 & -T \\ 0 & 1 \end{smallmatrix}\right)$ and $\xi^{-1} = \left(\begin{smallmatrix} 1 & T \\ 0 & 1 \end{smallmatrix}\right)$. Thus, when $D$ is inner, $u$ is similar to $u_1 \oplus u_2$.

Carlson and Clark [11] relate this to the homology of Hilbert modules. Let $H = H_1 \oplus H_2$. Let $\mathcal{H}, \mathcal{H}_1, \mathcal{H}_2$ be respectively the Hilbert $\mathcal{A}$-modules associated to $H, H_1, H_2$ with the action of $\mathcal{A}$ corresponding respectively to $u, u_1, u_2$. Consider then the short exact sequence

$$0 \to \mathcal{H}_1 \to \mathcal{H} \to \mathcal{H}_2 \to 0.$$

Carlson and Clark [11] observe that this *splits* (that is $\mathcal{H} \simeq \mathcal{H}_1 \oplus \mathcal{H}_2$) iff $u$ is similar to $u_1 \oplus u_2$, meaning there is an invertible $\xi$ on $H_1 \oplus H_2$ such that $u = \xi \left(\begin{smallmatrix} u_1 & 0 \\ 0 & u_2 \end{smallmatrix}\right) \xi^{-1}$. See [16] for more on Hilbert modules.

Generally speaking, the homomorphisms of the form (5) have proved to be a very fruitful source of examples, as will be illustrated below.

Unaware of the unpublished Foias–Williams preprint, Peller [48] observed that Hankel operators led to a very nice class of examples for which the Halmos problem should be checked.

Let $S\colon\ \ell_2 \to \ell_2$ be the shift operator. Recall that $\Gamma\colon\ \ell_2 \to \ell_2$ is a Hankel operator iff $\Gamma S = S^*\Gamma$. Equivalently, the entries $\Gamma_{ij}$ depend only on $i + j$. Let

$$R_\Gamma = \begin{pmatrix} S^* & \Gamma \\ 0 & S \end{pmatrix} \in B(\ell_2 \oplus \ell_2).$$

Then for any polynomial $P$ we have

$$P(R_\Gamma) = \begin{pmatrix} P(S^*) & \Gamma P'(S) \\ 0 & P(S) \end{pmatrix}.$$

Indeed, this is easy to check for $P = z^n$ by induction on $n$. Note that the homomorphism $P \mapsto P(R_\Gamma)$ is of the general form appearing in Remark 5. Since $\|P(S)\| = \|P(S^*)\| = \|P\|_\infty$, we find that $R_\Gamma$ is p.b. iff there is a constant $\alpha$ such that

$$\forall P \text{ polynomial} \qquad \|\Gamma P'(S)\| \le \alpha \|P\|_\infty. \tag{6}$$

By the well-known Nehari theorem and Fefferman's $H^1$-BMO duality, it is known that $\Gamma$ is bounded iff there is a function $\varphi$ in BMO such that $\Gamma_{ij} = \widehat{\varphi}(i+j)$ $\forall i, j \ge 0$. Moreover, we may assume that $\widehat{\varphi}(k) = 0$ for all $k < 0$. Let $\mathrm{BMO}_a$ denote the subspace of such $\varphi$'s. Then the correspondence $\varphi \to \Gamma(\varphi)$ defined by $\Gamma(\varphi)_{ij} = \widehat{\varphi}(i+j)$ is a linear isomorphism from $\mathrm{BMO}_a$ onto the subspace of $B(\ell_2)$ formed of all the bounded Hankel operators.

Moreover, $\varphi \to \Gamma(\varphi)$ is a 2-sided "module map" in the following sense: For any $f$ in $H^\infty$ (in particular for any polynomial) we have

$$f(S^*)\Gamma(\varphi) = \Gamma(\varphi) f(S) = \Gamma(f\varphi).$$

Here $f(S)\colon\ \ell_2 \to \ell_2$ represents the (Toeplitz) operator of multiplication by $f$ on $H^2 \simeq \ell_2$. In particular if $\Gamma = \Gamma(\varphi)$, we have $\Gamma P'(S) = \Gamma(\varphi P')$. We will denote $R(\varphi) = R_{\Gamma(\varphi)}$. Peller showed that if $\varphi' \in$ BMO then $\Gamma = \Gamma(\varphi)$ satisfies (6) and hence $R(\varphi) = R_{\Gamma(\varphi)}$ is polynomially bounded. He asked whether this implies that $R(\varphi)$ is similar to a contraction. Bourgain [8] then proved that indeed it is so and, with Aleksandrov, Peller in the summer of 1995 finally showed that $R(\varphi)$ is p.b. iff it is similar to a contraction. This seemed to destroy all hopes to use $R(\varphi)$ for a counterexample. However, it turned out that the operator-valued (sometimes called "vector-valued") analogue of $R(\varphi)$ does lead to a counterexample. To describe this we first need some more terminology.

**Definition.** A function $M\colon\ \mathbb{N} \to \mathbb{C}$ is called a Paley multiplier if

$$\sup_n \sum_{2^n \le k < 2^{n+1}} |M(k)|^2 < \infty.$$

It is well known that this condition characterizes the Fourier multipliers bounded from $H^1$ to $H^2$ (or equivalently from $H^2$ to $\mathrm{BMO}_a$).

Let $(B_n)$ be a sequence of operators on $H$ for which there are positive constants $\beta_1, \beta_2$ such that

$$\forall x = (x_n)_n \in \ell_2 \quad \beta_1 \left(\sum |x_n|^2\right)^{1/2} \le \left\|\sum x_n B_n\right\| \le \beta_2 \left(\sum |x_n|^2\right)^{1/2}. \tag{7}$$

We will also need to assume that there is a constant $\gamma > 0$ such that for any finitely supported scalar sequence $x = (x_n)_n$ we have

$$\gamma \sum |x_n| \le \left\|\sum x_n \overline{B}_n \otimes B_n\right\|. \tag{8}$$

**Example:** Let $(B_n)$ be a "spin system," i.e., mutually anti-commuting self-adjoint unitaries. Then (7) holds with $\beta_1 = 1$ and $\beta_2 = \sqrt{2}$ (see, e.g., [55, p. 76]) and (8) holds with $\gamma = 1/2$. The latter does hold because $\{B_n \otimes \overline{B}_n\}$ form a *commuting* sequence of self-adjoint unitaries such that, if $f$ denotes the classical "vacuum state" on the $C^*$-algebra generated by $\{B_n\}$, $f \otimes f$ vanishes on any product of distinct terms from $\{B_n \otimes \overline{B}_n\}$. It follows that the spectral distribution of $\{B_n \otimes \overline{B}_n\}$ with respect to $f \otimes f$ coincides with that of the Rademacher functions on $[0,1]$ (i.e., independent ±1-valued functions). Therefore we have

$$(1/2)\sum |x_n| \le \sup\left\{\left|\sum x_n\varepsilon_n\right| \mid \varepsilon_n = \pm 1\right\} = \left\|\sum x_n B_n \otimes \overline{B}_n\right\|.$$

Consider any Hankel matrix $\Gamma$ acting on $\widehat{H} = \ell_2(H) = H \oplus H \oplus \cdots$. Now its entries $\Gamma_{ij}$ are in $B(H)$ but still depend only on $i+j$. Let $\widehat{S}\colon\ \widehat{H} \to \widehat{H}$ be the usual (multivariate) shift and let

$$R(\Gamma) = \begin{pmatrix} \widehat{S}^* & \Gamma \\ 0 & \widehat{S} \end{pmatrix} \in B(\widehat{H} \oplus \widehat{H}).$$

Then the $(1,2)$ entry of this $2 \times 2$ matrix is $R(\Gamma)_{12} = \Gamma \in B(\widehat{H})$.

The same calculation as in the scalar case yields

$$P(R(\Gamma)) = \begin{pmatrix} P(\widehat{S}^*) & \Gamma P'(\widehat{S}) \\ 0 & P(\widehat{S}) \end{pmatrix}.$$

Clearly $R(\Gamma)$ is polynomially bounded iff there is a constant $\alpha$ such that

$$\forall P \text{ polynomial} \qquad \|\Gamma P'(\widehat{S})\| \le \alpha \|P\|_\infty. \tag{9}$$

Note that $P(R(\Gamma))_{12} = \Gamma P'(\widehat{S}) \in B(\widehat{H})$ and if $P = \sum a_k z^k$; then $\Gamma P'(\widehat{S})_{01} = \sum_{k\ge 1} k\Gamma_{0k}a_k$. In particular, we note for further use that if $P(z) = z^n$, and hence $P'(z) = nz^{n-1}$ we have for any $n \ge 0$

$$[\ [R(\Gamma)^n]_{12}\ ]_{01} = n\Gamma_{0n}. \tag{10}$$

We now turn to the Hankel matrix $\Gamma^M$ acting on $\widehat{H} = \ell_2(H) = H \oplus H \oplus \cdots$ defined by:

$$\Gamma^M_{ij} = \frac{1}{i+j} M(i+j) B_{i+j} \quad \text{if} \quad i+j \ne 0$$

and $\Gamma^M_{oo} = 0$ (say). We have

$$P(R(\Gamma^M)) = \begin{pmatrix} P(\widehat{S}^*) & \Gamma^M P'(\widehat{S}) \\ 0 & P(\widehat{S}) \end{pmatrix}.$$

Thus $R(\Gamma^M)$ is polynomially bounded iff there is a constant $\alpha$ such that

$$\forall P \text{ polynomial} \qquad \|\Gamma^M P'(\widehat{S})\| \le \alpha \|P\|_\infty. \tag{11}$$

Actually, it turns out that, if $M$ is Paley, we even have a constant $\alpha$ such that $\|\Gamma^M P'(\widehat{S})\| \le \alpha \|P\|_{BMO}$ for all $P$ ([49]), and a fortiori (11) holds.

Then the main result of [49] can be summarized as follows:

**Theorem 6.**

(i) *Assuming* (7), $R(\Gamma^M)$ *is polynomially bounded iff* $M$ *is a Paley multiplier.*
(ii) *Assuming* (7) *and* (8), $R(\Gamma^M)$ *is similar to a contraction only if*

$$\sum |M(n)|^2 < \infty.$$

**Note:** The fact that, assuming (7), $\sum |M(n)|^2 < \infty$ implies that $R(\Gamma^M)$ is similar to a contraction was proved more recently by É. Ricard [56], refining [7]. Thus (ii) above is also a characterization if we assume both (7) and (8).

The original proof (see [49]) that $R(\Gamma^M)$ is p.b. when $M$ is Paley appeared rather difficult because it used the Brownian motion description of BMO. Soon afterward, Kislyakov [29] saw how to eliminate the use of any probabilistic argument. His proof is based on "real methods" involved with the classical space BMO. A similar proof was given by McCarthy [35]. Shortly after that, Davidson and Paulsen [14] found the "right" proof that $R(\Gamma^M)$ is p.b. Their proof uses only the $B(H)$-valued version of Nehari's theorem (due to Sarason–Page, see [39]) and hence may be considered as strictly "operator theoretic."

The proof of (ii) in Theorem 6, based on Theorem 3 (Paulsen's criterion), is much easier: if $T$ is completely polynomially bounded and $T \in B(\ell_2 \otimes H)$ then for any pair of unit vectors $h, k \in \ell_2$ the mapping taking a polynomial $P$ to $\langle P(T)h, k\rangle$ is clearly c.b. with c.b. norm $\le C_{cpb}(T)$. In particular if, say $T = [T_{ij}]$ with $T_{ij} \in B(H)$, the mapping $P \mapsto P(T)_{i(0)j(0)}$ is c.b. with c.b. norm $\le C_{cpb}(T)$ for any choice of $(i(0), j(0))$. More precisely, for any finite set of coefficients $a_k \in B(K)$, we clearly have

$$\left\|\sum a_k \otimes (T^k)_{i(0)j(0)}\right\|_{B(K\otimes H)} \le \left\|\sum a_k \otimes T^k\right\|_{B(K\otimes \ell_2 \otimes H)}.$$

But now if $T = R(\Gamma)$, then (4) and (10) show that

$$\left\|\sum a_k \otimes k\Gamma_{0k}\right\|_{B(K\otimes H)} \le C_{cpb}(T) \sup_{z\in D} \left\|\sum a_k z^k\right\|_{B(K)}.$$

Consider now $\Gamma = \Gamma^M$, $K = \overline{H}$ and $a_k = x_k \overline{B_k}$, with $x_k$ an arbitrary finite sequence of scalars. Then $k\Gamma^M_{0k} = M(k)B_k$ for all $k > 0$ so that (7) and (8) combined with this last inequality yield

$$\gamma \sum\nolimits_{k>0} |M(k)x_k| \le C\beta_2 \left(\sum |x_k|^2\right)^{1/2}$$

from which we deduce

$$\left(\sum\nolimits_{k>0} |M(k)|^2\right)^{1/2} \le C\beta_2\gamma^{-1}. \qquad \square$$

The most classical example of a Paley multiplier is the indicator function of a Hadamard lacunary sequence in $\mathbb{N}$, e.g., we can take

$$M(n) = \begin{cases} 1 & \text{if } n \in \{2^k \mid k \geq 0\} \\ 0 & \text{otherwise} \end{cases}.$$

This gives us a Paley multiplier with $\sum |m(n)|^2 = \infty$ and hence

**Corollary 7.** *There is a p.b. operator on $\widehat{H} \oplus \widehat{H} \simeq \ell_2$ that is not similar to a contraction.*

In [51] we exhibit a pair $T_1, T_2$ of operators on $\ell_2$ that are *each* similar to a contraction but there is no invertible $\xi$ such that *both* $\xi T_1 \xi^{-1}, \xi T_2 \xi^{-1}$ are contractions.

*Remark* 8. The operators of the form $R(\Gamma)$ (of Foias-Williams-Peller type) can be considered on the Bergman space. However, it was proved recently (see [4, 13]) that in that class the analogous counterexamples to the preceding two similarity problems do not exist (in fact even power bounded implies similarity to a contraction).

The history of the Halmos similarity problem is tied up with the quest for extensions of the von Neumann inequality (see, e.g., [18]). The extension to several variables is rather puzzling: In 1963 Ando [5] obtained a version for pairs $T_1, T_2$ of commuting contractions and polynomials $P$ in two variables:

$$\forall P \text{ polynomial} \qquad \|P(T_1, T_2)\| \leq \sup\{|P(z_1, z_2)| \mid |z_1| \leq 1, |z_2| \leq 1\} \tag{12}$$

but in 1974 Varopoulos [64] showed that this does *not* extend to triples $T_1, T_2, T_3$ of (mutually) commuting contractions. More precisely, let us examine the validity for triples $T_1, T_2, T_3$ of (mutually) commuting contractions and polynomials $P$ in three variables of the inequality:

$$\forall P \qquad \|P(T_1, T_2, T_3)\| \leq C \sup\{|P(z_1, z_2, z_3)| \mid |z_1| \leq 1, |z_2| \leq 1, |z_3| \leq 1\}. \tag{13}$$

We wish to emphasize that Varopoulos [64] only proves that this does not hold with $C = 1$. In other words, although unlikely, it is still not known whether (13) holds with a universal finite constant $C$!

Ando actually proved a dilation theorem extending Theorem 2 to commuting pairs:

**Theorem 9 (Ando's dilation).** *For any commuting pair $T_1, T_2$ on $H$ with $\|T_1\| \leq 1$, $\|T_2\| \leq 1$ there is $\widehat{H}$ containing $H$ and a commuting pair $(\widehat{T}_1, \widehat{T}_2)$ of unitaries on $\widehat{H}$ such that*

$$\forall n, k \geq 1 \qquad T_1^n T_2^k = P_H \widehat{T}_1^n \widehat{T}_2^k{}_{|H}.$$

Before Varopoulos's counterexample, Parrott [41] had already shown that Ando's dilation does not extend to commuting triples of contractions.

See [36, 37] and references there for work related to “polynomial boundedness” on higher-dimensional domains, in the context of Douglas’s Hilbert modules.

In a different direction, consider a general bounded open domain $\Omega \subset \mathbb{C}$ that is connected but possibly not simply connected. Let us denote by $A(\Omega)$ the closure of the set $R(\Omega)$ of rational (bounded) functions with poles off of $\bar{\Omega}$, in the uniform sup norm on $\Omega$. We are interested in contractive unital homomorphisms $u\colon\ A(\Omega) \to B(H)$. Given such a $u$, let $T$ be the image under $u$ of the function $f(z) = z$, so that $u(f) = f(T)$ for any $f$ in $R(\Omega)$. We say that $\bar{\Omega}$ is a spectral set for $T$ if $\|u\|$=1, that is if

$$\forall f \in R(\Omega) \qquad\qquad \|f(T)\| \leq \|f\|_{A(\Omega)} = \|f\|_{C(\partial\Omega)}.$$

For instance, the unit disc is a spectral set for any contraction $T$. In analogy with the unitary dilation of contractions, it was conjectured that if $\bar{\Omega}$ is a spectral set for $T$ then $T$ admits a normal dilation $N$ with spectrum in $\partial\Omega$. More precisely, it was conjectured that there is $\hat{H} \supset H$, and a normal operator $N \in B(\hat{H})$ with spectrum in $\partial\Omega$ such that

$$\forall f \in R(\Omega) \qquad\qquad f(T) = P_H f(N)_{|H}.$$

When this holds, one says that $T$ admits a normal $\partial\Omega$-dilation. By Arveson’s results (see [46]) this happens iff $u$ is completely contractive, i.e., $\|u\|_{cb} = 1$. Thus the conjecture amounted to showing that, for a unital homomorphism $u\colon\ A(\Omega) \to B(H)$, contractive automatically implies completely contractive. This longstanding conjecture was confirmed by Agler [1] for the annulus or doubly connected domains, but was disproved independently in [2] and [17] for certain triply connected domains (conformally equivalent to an annulus with a disk removed). The example $T$ in [2] is a $4 \times 4$ matrix but this involves some machine numerical computation (in the $2 \times 2$ case, the conjecture was verified for any $\Omega$ by various authors).

Consider a uniform algebra $A$, i.e., a unital subalgebra of the space $C(\Omega)$ of continuous functions on a compact set $\Omega$. We will also assume that $A$ separates the points of $\Omega$. We will say that $A$ is “Halmos” if every bounded unital homomorphism $u\colon\ A \to B(H)$ is completely bounded. By Theorem 4 this means that every such $u$ is similar to a completely contractive homomorphism. In [50] this property is investigated in the broader framework of (possibly non-commutative) operator algebras. In particular, by [50], a uniform algebra $A$ is Halmos iff there is an exponent $\alpha$ and a constant $K$ such that for all $u$ as above we have

$$\|u\|_{cb} \leq K\|u\|^{\alpha}. \tag{14}$$

Let $\alpha_{\min}(A)$ denote the infimum of the $\alpha$’s for which there is a $K$ such that (14) holds for all $u$ as above. If $A$ is Halmos, then (14) holds for $\alpha = \alpha_{\min}(A)$ (for some $K$) and moreover $\alpha_{\min}(A)$ is an integer. The value of this integer can be identified by the validity of certain factorizations for matrices with entries in $A$, for which we refer the reader to [50]. By [49], we know that the disc algebra $A(D)$ is not Halmos. From this it is immediate that the ball and polydisc algebra on $\mathbb{C}^n$ are not Halmos. Actually, no example is known of a Halmos uniform algebra except for the obvious case $A = C(\Omega)$, so we can formulate:

**Problem.** Are $C^*$-algebras the only Halmos uniform algebras?

In that direction, there are only partial results, as follows.

**Theorem 10 ([50]).** *If a Halmos uniform algebra $A$ satisfies $\alpha_{\min}(A) \leq 2$ then $A$ is a $C^*$-algebra.*

Consider a uniform algebra $A \subset C(\Omega)$, equipped with a probability measure $\mu$ on $\Omega$ that is multiplicative on $A$ (for example $f \to \int f \, d\mu = f(0)$ for the disc algebra with $\mu$ = normalized Lebesgue measure on $\partial D$). Following [30] we will now introduce the "martingale extension" of $A$. Consider the product space $(\Omega^\infty, \mu^\infty) = (\Omega, \mu)^{\mathbb{N}}$ equipped with its natural filtration $\{\mathcal{A}_n\}$ where $\mathcal{A}_n$ is the $\sigma$-algebra generated by the first $n$-coordinates. We will denote by $M_A$ the algebra of all continuous functions $f$ on $\Omega^\infty$ such that, for each $n$, $f = \mathbb{E}^{\mathcal{A}_n} f$ belongs to $A$ as a function of the $n$th variable when the preceding coordinates are fixed. We will need to assume that $A$ admits a multiplicative linear functional $\varphi$ (perhaps distinct from $\mu$) and a non-zero $\psi \in A^*$ such that

$$\forall f, g \in A \qquad\qquad \psi(fg) = \varphi(f)\psi(g) + \psi(f)\varphi(g).$$

**Theorem 11 ([30]).** *In the preceding situation, $M_A$ is not Halmos.*

The case $A = \mathbb{C}$ was treated in a preliminary (unpublished) version of [49] on which [30] is based, but the interest of the preceding statement is that it assumes no "concrete" structure on $A$ (except for the existence of $\mu$, which is a rather mild assumption) and this is precisely the difficulty in the above problem. We need to construct a non trivial (meaning bounded but not c.b.) unital homomorphism on $A$ "from scratch," and the preceding statement does something like that but unfortunately on $M_A$ instead of $A$.

*Remark* 12. Several important open questions remain in the non-separable case, concerning for instance the algebra $H^\infty$ of bounded analytic functions on $D$. It is not known whether any contractive unital homomorphism $u\colon H^\infty \to B(H)$ is completely contractive, i.e., such that $\|u\|_{cb} = 1$ (or even completely bounded). This problem is studied in [47]. Of course the answer is clearly yes if $u$ is weak$*$-continuous. By [22] (see also [47]) it suffices to show that $u_2 = Id \otimes u$ is contractive on $M_2(H^\infty)$.

*Remark* 13. Let $\mathcal{A}$ be a unital $C^*$-algebra and let $u\colon \mathcal{A} \to B(H)$ be a bounded unital homomorphism. We will say that $u$ is "unitarizable" (Kadison [27] used the term "orthogonalizable") if there is an invertible $\xi\colon H \to H$ such that $a \to \xi u(a) \xi^{-1}$ is a $*$-homomorphism on $\mathcal{A}$, equivalently if it maps unitaries to unitaries. In 1955, Kadison [27] already asked whether any bounded homomorphism $u$ on a $C^*$-algebra (with values in $B(H)$) is automatically unitarizable.

This longstanding (still open) problem can be viewed as the $C^*$-analogue of the Halmos similarity problem. It is known to be equivalent to the "derivation problem." The latter asks whether any bounded derivation $D\colon\ \mathcal{A} \to B(H)$ on a $C^*$-algebra $\mathcal{A} \subset B(H)$ is inner. We refer the reader to [54, 55, 46] for more information on this problem.

## References

[1] J. Agler. *Rational dilation on an annulus.* Ann. of Math. (2) 121 (1985), no. 3, 537–563.

[2] J. Agler, J. Harland and B.J. Raphael. *Classical function theory, operator dilation theory, and machine computation on multiply-connected domains.* Mem. Amer. Math. Soc. 191 (2008), no. 892, viii+159 pp.

[3] A.B. Aleksandrov and V. Peller, *Hankel Operators and Similarity to a Contraction.* Int. Math. Res. Not. 6 (1996), 263–275.

[4] A. Aleman and O. Constantin, *Hankel operators on Bergman spaces and similarity to contractions.* Int. Math. Res. Not. 35 (2004), 1785–1801.

[5] T. Ando, *On a Pair of Commutative Contractions.* Acta Sci. Math. 24 (1963), 88–90.

[6] W. Arveson. *Dilation theory yesterday and today.* Oper. Theory Adv. Appl. **207** (2010), 99–123 (in this volume).

[7] C. Badea and V.I. Paulsen, *Schur Multipliers and Operator-Valued Foguel–Hankel Operators.* Indiana Univ. Math. J. 50 (2001), no. 4, 1509–1522.

[8] J. Bourgain, *On the Similarity Problem for Polynomially Bounded Operators on Hilbert Space.* Israel J. Math. 54 (1986), 227–241.

[9] M. Bożejko, *Littlewood Functions, Hankel Multipliers and Power Bounded Operators on a Hilbert Space.* Colloquium Math. 51 (1987), 35–42.

[10] J. Carlson, D. Clark, C. Foias and J.P. Williams, *Projective Hilbert $\boldsymbol{A}(\boldsymbol{D})$-Modules.* New York J. Math. 1 (1994), 26–38, electronic.

[11] J.F. Carlson and D.N. Clark, *Projectivity and Extensions of Hilbert Modules Over $A(D^N)$.* Michigan Math. J. 44 (1997), 365–373.

[12] J.F. Carlson and D.N. Clark, *Cohomology and Extensions of Hilbert Modules.* J. Funct. Anal. 128 (1995), 278–306.

[13] O. Constantin and F. Jaëck, *A joint similarity problem on vector-valued Bergman spaces.* J. Funct. Anal. 256 (2009), 2768–2779.

[14] K.R. Davidson and V.I. Paulsen, *Polynomially bounded operators.* J. Reine Angew. Math. 487 (1997), 153–170.

[15] A.M. Davie, *Quotient Algebras of Uniform Algebras.* J. London Math. Soc. 7 (1973), 31–40.

[16] R.G. Douglas and V.I. Paulsen. Hilbert modules over function algebras. Pitman Research Notes in Mathematics Series, 217. John Wiley & Sons, Inc., New York, 1989.

[17] M.A. Dritschel and S. McCullough. *The failure of rational dilation on a triply connected domain.* J. Amer. Math. Soc. 18 (2005), 873–918.

[18] S. Drury, *Remarks on von Neumann's Inequality. Banach Spaces, Harmonic Analysis, and Probability Theory.* Proceedings, R. Blei and S. Sydney (eds.), Storrs 80/81, Springer Lecture Notes 995, 14–32.

[19] S. Foguel, *A Counterexample to a Problem of Sz.-Nagy.* Proc. Amer. Math. Soc. 15 (1964), 788–790.

[20] C. Foias, *Sur Certains Théorèmes de J. von Neumann Concernant les Ensembles Spectraux.* Acta Sci. Math. 18 (1957), 15–20.

[21] C. Foias and J.P. Williams, *On a Class of Polynomially Bounded Operators.* Preprint (unpublished, approximately 1976).

[22] C. Foias and I. Suciu. *On operator representation of logmodular algebras.* Bull. Acad. Polon. Sci. Sér. Sci. Math. Astronom. Phys. 16 (1968) 505–509.

[23] J.B. Garnett. *Bounded analytic functions.* Pure and Applied Mathematics, 96. Academic Press, Inc., New York-London, 1981.

[24] U. Haagerup, *Solution of the Similarity Problem for Cyclic Representations of $C^*$-Algebras.* Annals of Math. 118 (1983), 215–240.

[25] P. Halmos, *On Foguel's answer to Nagy's question.* Proc. Amer. Math. Soc. 15 1964 791–793.

[26] P. Halmos, *Ten Problems in Hilbert Space.* Bull. Amer. Math. Soc. 76 (1970), 887–933.

[27] R. Kadison. On the orthogonalization of operator representations. Amer. J. Math. 77 (1955), 600–620.

[28] N. Kalton and C. Le Merdy, *Solution of a Problem of Peller Concerning Similarity.* J. Operator Theory 47 (2002), no. 2, 379–387.

[29] S. Kislyakov, *Operators that are (dis)Similar to a Contraction: Pisier's Counterexample in Terms of Singular Integrals (Russian).* Zap. Nauchn. Zap. Nauchn. Sem. St.-Petersburg. Otdel. Mat. Inst. Steklov. (POMI) 247 (1997), Issled. po Linein. Oper. i Teor. Funkts. 25, 79–95, 300; translation in J. Math. Sci. (New York) 101 (2000), no. 3, 3093–3103.

[30] S. Kislyakov, *The similarity problem for some martingale uniform algebras (Russian).* Zap. Nauchn. Sem. St.-Petersburg. Otdel. Mat. Inst. Steklov. (POMI) 270 (2000), Issled. po Linein. Oper. i Teor. Funkts. 28, 90–102, 365; translation in J. Math. Sci. (N. Y.) 115 (2003), no. 2, 2141–2146.

[31] A. Lebow, *A Power Bounded Operator which is not Polynomially Bounded.* Mich. Math. J. 15 (1968), 397–399.

[32] C. Le Merdy, *The Similarity Problem for Bounded Analytic Semigroups on Hilbert Space.* Semigroup Forum 56 (1998), 205–224.

[33] B. Lotto, *von Neumann's Inequality for Commuting, Diagonalizable Contractions, I.* Proc. Amer. Math. Soc. 120 (1994), 889–895.

[34] B. Lotto and T. Steger, *von Neumann's Inequality for Commuting, Diagonalizable Contractions, II.* Proc. Amer. Math. Soc. 120 (1994), 897–901.

[35] J.E. McCarthy. On Pisier's Construction. arXiv:math/9603212.

[36] G. Misra and S. Sastry, *Completely Contractive Modules and Associated Extremal Problems.* J. Funct. Anal. 91 (1990), 213–220.

[37] G. Misra and S. Sastry, *Bounded Modules, Extremal Problems and a Curvature Inequality*. J. Funct. Anal. 88 (1990), 118–134.

[38] J. von Neumann, *Eine Spektraltheorie für allgemeine Operatoren eines unitären Raumes*. Math. Nachr. 4 (1951), 258–281.

[39] N. Nikolskii, *Treatise on the Shift Operator*. Springer Verlag, Berlin, 1986.

[40] E.W. Packel. *A semigroup analogue of Foguel's counterexample*. Proc. AMS 21 (1969) 240–244.

[41] S. Parrott, *Unitary Dilations for Commuting Contractions*. Pacific J. Math. 34 (1970), 481–490.

[42] S. Parrott, *On a Quotient Norm and the Sz.-Nagy–Foias Lifting Theorem*. J. Funct. Anal. 30 (1978), 311–328.

[43] V. Paulsen, *Every Completely Polynomially Bounded Operator is Similar to a Contraction*. J. Funct. Anal. 55 (1984), 1–17.

[44] V. Paulsen, *Completely Bounded Homomorphisms of Operator Algebras*. Proc. Amer. Math. Soc. 92 (1984), 225–228.

[45] V. Paulsen, *Completely Bounded Maps and Dilations*. Pitman Research Notes in Math. 146, Longman, Wiley, New York, 1986.

[46] V.I. Paulsen, Completely bounded maps and operator algebras. Cambridge Studies in Advanced Mathematics, 78. Cambridge University Press, Cambridge, 2002.

[47] V.I. Paulsen and M. Raghupathi. Representations of logmodular algebras. Preprint, 2008.

[48] V. Peller, *Estimates of Functions of Power Bounded Operators on Hilbert Space*. J. Oper. Theory 7 (1982), 341–372.

[49] G. Pisier, *A Polynomially Bounded Operator on Hilbert Space which is not Similar to a Contraction*. J. Amer. Math. Soc. 10 (1997), 351–369.

[50] G. Pisier, *The Similarity Degree of an Operator Algebra*. St. Petersburg Math. J. 10 (1999), 103–146.

[51] G. Pisier, *Joint Similarity Problems and the Generation of Operator Algebras with Bounded Length*. Integr. Equ. Op. Th. 31 (1998), 353–370.

[52] G. Pisier, *The Similarity Degree of an Operator Algebra II*. Math. Zeit. 234 (2000), 53–81.

[53] G. Pisier, *Multipliers of the Hardy Space $H^1$ and Power Bounded Operators*. Colloq. Math. 88 (2001), no. 1, 57–73.

[54] G. Pisier, *Similarity problems and completely bounded maps*. Second, Expanded Edition. Springer Lecture Notes 1618 (2001).

[55] G. Pisier, *Introduction to operator space theory*. London Mathematical Society Lecture Note Series, 294. Cambridge University Press, Cambridge, 2003.

[56] É. Ricard, *On a Question of Davidson and Paulsen*. J. Funct. Anal. 192 (2002), no. 1, 283–294.

[57] F. Riesz and B. Sz.-Nagy, *Leçons d'Analyse Fonctionnelle*. Gauthier–Villars, Paris, Akadémiai Kiadó, Budapest, 1965.

[58] G.C. Rota, *On Models for Linear Operators*. Comm. Pure Appl. Math. 13 (1960), 468–472.

[59] B. Sz.-Nagy, *On Uniformly Bounded Linear Transformations on Hilbert Space.* Acta Sci. Math. (Szeged) 11 (1946-48), 152–157.

[60] B. Sz.-Nagy, *Completely Continuous Operators with Uniformly Bounded Iterates.* Publ. Math. Inst. Hungarian Acad. Sci. 4 (1959), 89–92.

[61] B. Sz.-Nagy, *Sur les Contractions de l'espace de Hilbert.* Acta Sci. Math. 15 (1953), 87–92.

[62] B. Sz.-Nagy, *Spectral Sets and Normal Dilations of Operators.* Proc. Intern. Congr. Math. (Edinburgh, 1958), 412–422, Cambridge Univ. Press.

[63] B. Sz.-Nagy and C. Foias, *Harmonic Analysis of Operators on Hilbert Space.* Akadémiai Kiadó, Budapest, 1970.

[64] N. Varopoulos, *On an Inequality of von Neumann and an Application of the Metric Theory of Tensor Products to Operators Theory.* J. Funct. Anal. 16 (1974), 83–100.

Gilles Pisier
Texas A&M University
College Station, TX 77843, USA

*and*

Université Paris VI
4 Place Jussieu
Institut Math. Jussieu
Équipe d'Analyse Fonctionnelle, Case 186
F-75252 Paris Cedex 05, France
e-mail: pisier@math.tamu.edu

Operator Theory:
Advances and Applications, Vol. 207, 341–349
© 2010 Springer Basel AG

# Paul Halmos and Invariant Subspaces

Heydar Radjavi and Peter Rosenthal

**Abstract.** This paper consists of a discussion of the contributions that Paul Halmos made to the study of invariant subspaces of bounded linear operators on Hilbert space.

**Mathematics Subject Classification (2000).** 47A15.

**Keywords.** Invariant subspaces, bounded linear operators.

Paul Halmos said that problems are "the heart of mathematics" [35] and explained "The purpose of formulating the yes-or-no question is not only to elicit the answer; its main purpose is to point to an interesting area of ignorance." [32]

The invariant subspace problem was not originally formulated by Halmos. (Sometime in the nineteen-thirties, von Neumann showed that compact operators have invariant subspaces [2]). However, Paul was the person who made the problem widely known. Moreover, he proposed the study of some important special cases of the problem and suggested numerous related questions which pointed to very interesting areas of ignorance. As a result of the work of hundreds of mathematicians, much of the ignorance has been replaced by very interesting knowledge.

The purpose of this article is to outline some of the mathematics that developed from Paul's research and probing questions related to invariant subspaces.

We begin with the basic definitions. Throughout, we are concerned with bounded linear operators on a complex, separable, infinite-dimensional Hilbert space, which we generally denote by $\mathcal{H}$. By a *subspace* we mean a subset of $\mathcal{H}$ that is closed in the topological sense as well as with respect to the vector operations. A subspace is *invariant* under an operator if the operator maps the subspace into itself. That is, the subspace $\mathcal{M}$ is invariant under the operator $A$ if $Ax \in \mathcal{M}$ whenever $x \in \mathcal{M}$. The *trivial subspaces* are $\{0\}$ and $\mathcal{H}$; they are both invariant under all operators. The *invariant subspace problem* is the question: does every bounded linear operator on $\mathcal{H}$ have a nontrivial invariant subspace? In other words, if $A$ is a bounded linear operator on $\mathcal{H}$, must there exist a subspace $\mathcal{M}$ other than $\{0\}$ and $\mathcal{H}$ which is invariant under $A$? In spite of all the work concerning

We thank Jonathan Korman for his assistance in the preparation of this article.

this problem that Paul and many others generated, the answer is still unknown. However, a surprisingly large number of theorems have been established concerning related questions.

In particular, very interesting mathematics resulted from Paul's 1963 proposal [28] that two special cases of the invariant subspace problem be attempted.

Earlier, in 1954, Aronszajn and Smith[2] published a paper proving that every compact operator on a Banach space had a nontrivial invariant subspace (they acknowledged that von Neumann had proven the Hilbert-space case earlier, in unpublished work). Smith raised the question of whether an operator must have a nontrivial invariant subspace if its square is compact (it is easy to give examples of operators that are not compact but have compact squares). In the 1963 article[28], Paul publicized Smith's (unpublished) question.

This problem turned out to be extremely interesting. Abraham Robinson and his student Allen Bernstein [10] gave an affirmative answer. In fact, they proved that polynomially-compact operators (that is, operators $A$ such that $p(A)$ is compact for some nonzero polynomial $p$) have nontrivial invariant subspaces. Surprisingly, their proof was based on Robinson's theory of non-standard analysis, and their result still seems to be the most impressive theorem obtained by using that theory.

Paul read the pre-print of the Bernstein-Robinson paper and translated their argument into standard analysis [29]. He subsequently [31] abstracted the concept of quasitriangular operator that had been implicit in the work of Aronszajn-Smith[2] and Bernstein-Robinson [10] (and its generalization by Arverson-Feldman [6]). Paul defined an operator $A$ to be *quasitriangular* if there is an increasing sequence $\{E_n\}$ of finite rank projections such that $||AE_n - E_nAE_n|| \to 0$. It is easy to see that the range of a projection $P$ is invariant under the operator $A$ if and only if $AP = PAP$. Thus an operator is quasitriangular if there is an increasing sequence of finite-dimensional "almost-invariant" subspaces of the operator.

Since the essence of the concept of quasitriangularity had been used in establishing the existence of nontrivial invariant subspaces for compact operators and generalizations, Paul's question of whether quasitriangular operators had invariant subspaces appeared to be very reasonable. However, Apostol, Foias, and Voiculescu [1] proved the remarkable result that an operator whose adjoint does not have an eigenvector must be quasitriangular. Since the ortho-complement of a one-dimensional space spanned by an eigenvector of $A^*$ is invariant under $A$, it follows that every non-quasitriangular operator does have a nontrivial invariant subspace. Thus Paul's question of the existence of nontrivial invariant subspaces of quasitriangular operators is equivalent to the general invariant subspace problem.

Although the hypothesis of quasitriangularity had become superfluous, the quest continued for existence results using compactness. Arveson-Feldman[6] had shown that a quasinilpotent operator has a nontrivial invariant subspace if the uniformly-closed algebra generated by the operator contains a compact operator other than 0. A natural question was whether the hypothesis of quasinilpotence could be removed from the Arverson-Feldman result.

In lectures and informal discussions with many mathematicians, Paul raised several related questions. One was: must two compact operators that commute with each other have a common nontrivial invariant subspace? Another was his asking about a generalization of the previously-known results that seemed to be completely out of reach, the question of whether an operator had to have an nontrivial invariant subspace if it simply commuted with a compact operator other than 0. It was a huge surprise to Paul and everyone else who had thought about such problems when Victor Lomonosov [41] established an affirmative answer to this question. In fact, Lomonosov proved much more. In particular, he showed that if $K$ is a compact operator other than 0, there is a nontrivial subspace that is simultaneously invariant under all of the operators that commute with $K$. In addition, Lomonosov [41] proved that the operator $A$ has a nontrivial invariant subspace if it commutes with any operator $B$ that is not a multiple of the identity operator and that itself commutes with a compact operator other than 0. The scope of this latter result is still not clear; at first, it seemed possible that every operator satisfied its hypothesis. Although this was shown not to be the case [25], it remains barely conceivable that all the operators that do not satisfy the hypothesis either have eigenvectors or have adjoints that have eigenvectors, in which case Lomonosov's work would solve the invariant subspace problem.

H.M. Hilden found a beautiful and unbelievably simple proof of part of Lomonosov's theorem. Hilden's proof that there is a nontrivial invariant subspace that is simultaneously invariant under all the operators that commute with each given compact operator other than 0 uses only very elementary results of functional analysis and can be presented in a couple of pages (see [43]).

The other special case of the invariant subspace problem that Paul raised in his 1963 article concerned subnormal operators. Recall that an operator is *normal* if it commutes with its adjoint. Normal operators have lots of invariant subspaces; in particular, their spectral subspaces are invariant. Paul defined an operator to be *subnormal* [26][1] if it is the restriction of a normal operator to an invariant subspace. (Restrictions of normal operators to their spectral subspaces, or to any other of their reducing subspaces, are normal. Some normal operators have non-reducing invariant subspaces, and the restriction of a normal operator to such a subspace need not be normal.)

The problem of existence of nontrivial invariant subspaces for subnormal operators stimulated a great deal of research into the properties of subnormal operators. In fact, there were so many results obtained that John Conway wrote two books on the subject [17, 18]. The invariant subspace problem for subnormal operators was ultimately solved by Scott Brown [14]. His proof is a beautiful blend of techniques from operator theory and complex analysis. Brown's work was generalized in several directions. One impressive variant obtained by Brown-

[1] Actually, Paul used the terminology differently in his 1950 paper [26]. He subsequently [28] introduced the terminology as we have defined it and this has become standard.

Chevreau-Pearcy [16] established that every contraction whose spectrum contains the unit circle has a nontrivial invariant subspace.

Paul introduced another generalization of normal operators. He defined [26] an operator $A$ to be *hyponormal* if $A^*A - AA^*$ is a positive operator[2]. It is easy to see that every subnormal operator is hyponormal, so a natural question is: does every hyponormal operator have a nontrivial invariant subspace? This problem remains open, although Berger [9] showed that for every hyponormal operator $A$ there is a natural number $n$ such that $A^n$ has a nontrivial invariant subspace, and Brown [15] proved that hyponormal operators have nontrivial invariant subspaces if their spectra are substantial enough that the uniform closure of the rational functions on the spectrum does not contain every function that is continuous on the spectrum.

Paul was interested in many other aspects of invariant subspaces in addition to their existence. Let $\mathcal{S}$ be any collection of bounded linear operators on a given space $\mathcal{H}$. Paul noted that the collection of all subspaces (including $\{0\}$ and $\mathcal{H}$) that are simultaneously invariant under all the members of $\mathcal{S}$ is a complete lattice under inclusion, with the supremum of a family of subspaces being the closed linear span of the family and the infimum being the intersection of the members of the family. Paul therefore suggested the notation Lat $\mathcal{S}$ for the collection of all subspaces that are simultaneously invariant under all the operators in $\mathcal{S}$. Similarly, Paul noted that the collection of all operators that leave all of the subspaces in a given family $\mathcal{F}$ invariant forms an algebra; he used the notation Alg $\mathcal{F}$ to denote that algebra of operators. Paul suggested that an algebra $\mathcal{A}$ be called *reflexive* if $\mathcal{A} = \text{Alg Lat } \mathcal{A}$, and that a lattice $\mathcal{L}$ of subspaces be called *reflexive* if $\mathcal{L} = \text{Lat Alg } \mathcal{L}$ (see [33]).

We began thinking about reflexive algebras (but without that name and without Halmos's very useful notation) at about the same time as Paul did. Earlier, Sarason [50] had proven that every weakly-closed commutative unital algebra of normal operators is reflexive (though he didn't use that term). Also, Arveson [3] had shown that the only weakly-closed algebra of operators that contains a maximal abelian self-adjoint algebra and has no nontrivial invariant subspaces is the algebra of all operators.

It is clear, if you think about it for a minute, that Arveson's theorem can be reformulated as follows in terms of Halmos's notation: if $\mathcal{A}$ is a weakly-closed algebra that contains a maximal abelian self-adjoint algebra and $\text{Lat}\mathcal{A} = \{\{0\}, \mathcal{H}\}$, then $\mathcal{A}$ is reflexive. We thought about Arveson's theorem for much more than a minute and also spent a considerable amount of time studying Sarason's result. This led to our conjecturing and then proving that a weakly-closed algebra $\mathcal{A}$ is reflexive if it contains a maximal abelian self-adjoint algebra and $\text{Lat}\mathcal{A}$ is totally ordered [45]. A special case of this theorem that we were proud of then, and still like this many years later, is the fact that an operator $A$ on $L^2(0,1)$ is a weak limit of non-commutative polynomials in multiplication by $x$ and the Volterra operator

[2] Paul used different terminology in [26].

if and only if it leaves invariant all of the subspaces

$$\mathcal{M}_\lambda = \{f \in \mathcal{L}^2(0,1) \mid f = 0 \text{ a.e. on } [0,\lambda]\}.$$

It seemed possible that every weakly-closed algebra containing a maximal abelian self-adjoint algebra was reflexive; there were some other special cases that were established ( [46], [21]). However, Arveson [4] produced an example of such an algebra that was not reflexive. Moreover, Arveson [4] proved that a weakly-closed algebra which contains a maximal abelian self-adjoint algebra is reflexive if Lat$\mathcal{A}$ is generated as a lattice by a finite number of totally-ordered sublattices. Arveson's beautiful paper [4] contains a number of interesting concepts and numerous results about them. For another exposition of Arveson's work and some of its aftermath see [20]. There has been a great deal of subsequent work on CSL algebras (i.e., reflexive algebras such that the projections onto the invariant subspaces of the algebra all commute with each other) and other concepts originating in Arveson's paper. Even more generally, the study of reflexive algebras appears to have played an important role in stimulating investigation of many other kinds of nonselfadjoint algebras.

Consistent with Paul's definition, an individual operator $A$ is called *reflexive* if the weakly-closed algebra generated by $A$ and the identity is reflexive. For an individual operator $A$, it is customary to write Lat$A$ for Lat$\{A\}$. If an operator is reflexive, it has a large number of invariant subspaces in the sense that an operator $B$ that leaves every member of Lat$A$ invariant is in the weakly-closed algebra generated by $A$ and the identity operator. By Sarason's theorem [50], every normal operator is reflexive. In the same paper, Sarason proved that every analytic Toeplitz operator is reflexive. It has also been shown that isometries are reflexive [22]. A striking result is the strengthening of Scott Brown's theorem by Olin and Thompson [44] (also see [42]) to the result that subnormal operators are reflexive.

The reflexive lattices are, as Paul observed [33], those that can be written as Lat$\mathcal{S}$ for some collection $\mathcal{S}$ of operators. Arveson's fundamental paper [4] contains a number of results concerning reflexive lattices and there has also been some subsequent work (see [20]). The case when $\mathcal{S}$ consists of a single operator is of particular interest. An abstract lattice is said to be *attainable* if there is an operator $A$ on $\mathcal{H}$ such that the lattice is order-isomorphic to Lat$A$ [49].

In the early 1960's, Paul raised the question of which totally-ordered lattices are attainable. It is easy to see that every such lattice is order-isomorphic to a closed subset of $[0,1]$ with the usual order (just use the separability of $\mathcal{H}$). Donoghue [24] proved that $\omega + 1$ is attainable (where $\omega$ is the order-type of the natural numbers), and he [24], Dixmier [23] and Brodskii [12] each independently proved that the closed unit interval is attainable, attained by the Volterra integral operator. In response to Paul's questions, it was shown that $\omega + n$ and $[0,1] + n$, for $n$ any natural number, are attainable [49]. It was subsequently proven that $\omega + \omega + 1$ is attainable [39]. Davidson and Barría went on to show that the ordinal sum $\alpha + 1 + \beta^*$ is attainable for all countable ordinals $\alpha$ and $\beta$ [8], and Barría [7] showed that every ordinal sum of a finite number of natural numbers and copies of the closed unit interval, in any order, is attainable.

There is a large body of work concerning invariant subspaces of multiplication by the independent variable and other operators on the Hardy-Hilbert space $\mathcal{H}^2$ and on other spaces of analytic functions. This work was initiated by the beautiful paper of Beurling [11] in which he characterized the invariant spaces of the unilateral shift in terms of inner functions. Although Paul did not do a lot of work in this area, he did make some important contributions. He gave a Halmosian treatment (in the words of Don Sarason [37]) of higher multiplicity shifts [27] and, in joint work with Arlen Brown [13], gave a similarly incisive treatment of the basic properties of Toeplitz operators.

A subspace is said to be *reducing* for an operator if both it and its orthogonal complement are invariant under the operator. An operator is said to be *reducible* if it has a nontrivial reducing subspace; otherwise the operator is said to be *irreducible*. Paul [30] proved that the set of irreducible operators is uniformly dense in the algebra of all operators on $\mathcal{H}$ and then asked the corresponding question about the set of reducible operators. Although he obtained some partial results [30], he was unable to settle this. Paul's seemingly innocuous question remained open for eight years until Voiculescu [52] gave an affirmative answer. This paper of Voiculescu's that was stimulated by Paul's question introduced the concept of approximate equivalence of representations of $C^*$ algebras and powerful techniques for the study of that concept which provided a very fruitful approach to a number of problems in the theory of $C^*$-algebras.

We have described only some of the most direct consequences of Paul Halmos's interests in invariant subspaces. There is a huge body of additional work concerning the above and related topics. There are literally thousands of research papers that have some concern with some aspects of invariant subspaces and that use ideas that can explicitly or implicitly be traced back to Paul Halmos.

For further information, we recommend the following expository writings (the first two for the obvious reason of the authors' self-interest and the rest simply because they are excellent): [47, 48, 20, 51, 5, 40, 19] and, of course, any of Paul's own writings, especially [36, 37, 38, 34].

## References

[1] C. Apostol, C. Foias, and D. Voiculescu, *Some results on non-quasitriangular operators IV*, Rev. Roumaine Math. Pures Appl. 18 (1973), 487–514.

[2] N. Aronszajn and K.T. Smith,*Invariant subspaces of completely continuous operators*, Ann. of Math. 60 (1954), 345–350.

[3] W.B. Arveson, *A density theorem for operator algebras*, Duke Math. J. 34 (1967), 635–647.

[4] W.B. Arveson, *Operator algebras and invariant subspaces*, Ann. of Math. (2) 100 (1974), 433–532.

[5] W.B. Arveson *Ten lectures on operator algebras*, CBMS Regional Conference Series in Mathematics 55, A.M.S., Providence, 1984.

[6] W.B. Arveson and J. Feldman, *A note on invariant subspaces*, Michigan Math. J. 15 (1968), 60–64.

[7] J. Barría, *The invariant subspaces of a Volterra operator*, J. Op. Th. 6 (1981), 341–349.

[8] J. Barría and K.R. Davidson, *Unicellular operators*, Trans. Amer. Math. Soc. 284 (1984), 229–246.

[9] Charles A. Berger, *Sufficiently high powers of hyponormal operators have rationally invariant subspaces*, Int. Equat. Op. Th. 1 (1978), 444–447.

[10] A.R. Bernstein and A. Robinson, *Solution of an invariant subspace problem of K.T. Smith and P.R. Halmos*, Pacific J. Math, 16 (1966), 421–431.

[11] A. Beurling, *On two problems concerning linear transformations in Hilbert space*, Acta Math. 81 (1949), 239–255.

[12] A.S. Brodskii, *On a problem of I.M. Gelfand, Uspehi Mat. Nauk.*, 12 (1957), 129–132.

[13] Arlen Brown and P.R. Halmos, *Algebraic properties of Toeplitz operators*, J. Reine Angew. Math. 231 (1963), 89–102.

[14] Scott W. Brown, *Some invariant subspaces for subnormal operators*, Int. Equat. Oper. Th. 1 (1978), 310–333.

[15] Scott W. Brown, *Hyponormal operators with thick spectra have invariant subspaces*, Ann. of Math. (2) 125 (1987), 93–103.

[16] Scott W. Brown, Bernard Chevreau, and Carl Pearcy, *Contractions with rich spectrum have invariant subspaces*, J. Op. Th. 1 (1979), 123–136.

[17] John B. Conway, *Subnormal Operators*, Pitman, London, 1981.

[18] John B. Conway, *The Theory of Subnormal Operators*, Amer. Math. Soc. Surveys and Monographs *36*, Providence, 1991.

[19] John B. Conway, *A Course in Operator Theory*, GTM 21, Amer. Math. Soc., U.S.A., 1999.

[20] Kenneth R. Davidson, *Nest algebras*, Pitman Research Notes in Mathematics Series, Longman Scientific & Technical, Great Britain, 1988.

[21] Chandler Davis, Heydar Radjavi, and Peter Rosenthal, *On operator algebras and invariant subspaces*, Cann. J. Math. 21 (1969), 1178–1181.

[22] J.A. Deddens, *Every isometry is reflexive*, Proc. Amer. Math. Soc. 28 (1971), 509–512.

[23] J. Dixmier, *Les opérateurs permutables à l'opérateur integral*, Portugal. Math. Fas. 2 8 (1949), 73–84.

[24] W.F. Donoghue, *The lattice of invariant subspaces of a completely continuous quasi-nilpotent transformation*, Pacific J. Math.7 (1957), 1031–1035.

[25] D.W. Hadwin, E.A. Nordgren, Heydar Radjavi, and Peter Rosenthal, *An operator not satisfying Lomonosov's hypothesis*, J. Func. Anal. 38 (1980), 410–415.

[26] P.R. Halmos, *Normal dilations and extensions of operators*, Summa Bras. Math. II (1950), 125–134.

[27] P.R. Halmos, *Shifts on Hilbert spaces*, J. Reine Angew. Math. 208 (1961), 102–112.

[28] P.R. Halmos, *A Glimpse into Hilbert Space*, Lectures on Modern Mathematics, Vol. I, T. Saaty (Ed.), pp. 1–22, J. Wiley & Sons Inc. (1963).

[29] P.R. Halmos, *Invariant subspaces of polynomially compact operators*, Pacific J. Math. 16, (1966), 433–437.

[30] P.R. Halmos, *Irreducible operators*, Michigan Math. J. 15 (1968), 215–223.

[31] P.R. Halmos, *Quasitriangular operators*, Acta Scien. Math. XXIX (1968), 259–293.

[32] P.R. Halmos, *Ten Problems in Hilbert Space*, Bull. A.M.S. 76, (1970), 887–933.

[33] P.R. Halmos, *Reflexive lattices of subspaces*, J. London Math. Soc. (2) 4, (1971), 257–263.

[34] P.R. Halmos, *Ten years in Hilbert space*, Integ. Equat. Op. Th. 2/4 (1979), 529–563.

[35] P.R. Halmos, *The Heart of Mathematics*, American Mathematical Monthly 87 (1980), 519–524.

[36] P.R. Halmos, *A Hilbert Space Problem Book*, Second Edition, GTM 19, Springer-Verlag, New York, 1982.

[37] P.R. Halmos, *Selecta: Research Contributions*, Springer-Verlag, New York, 1983.

[38] P.R. Halmos, *Selecta: Expository Writing*, Springer-Verlag, New York, 1983.

[39] K.J. Harrison and W.E. Longstaff, *An invariant subspace lattice of order type* $\omega + \omega + 1$, Proc. Math. Soc. 79 (1980), 45–49.

[40] David R. Larson, *Triangularity in operator algebras*, Surveys of some recent results in operator theory, Vol. II, Pitman Res. Notes Math. 192, Longman, Harlow, 1988, pp. 121–188.

[41] V.J. Lomonosov, *Invariant subspaces for operators commuting with compact operators*, Functional Anal. and Appl. 7 (1973), 55–56.

[42] John E. McCarthy, *Reflexivity of subnormal operators*, Pacific J. Math. 161 (1993), 359–370.

[43] A.J. Michaels, *Hilden's simple proof of Lomonosov's invariant subspace theorem*, Adv. Math. 25 (1977), 56–58.

[44] Rober F. Olin and James E. Thomson, *Algebras of subnormal operators*, J. Func. Anal. 37 (1980), 271–301.

[45] Heydar Radjavi and Peter Rosenthal, *On invariant subspaces and reflexive algebras*, Amer. J. Math. 91 (1969), 683–692.

[46] Heydar Radjavi and Peter Rosenthal, *A sufficient condition that an operator algebra be self-adjoint*, Cann. J. Math. 23 (1971), 588–597.

[47] Heydar Radjavi and Peter Rosenthal, *Invariant Subspaces*, second edition, Dover, Mineola, N.Y., 2003.

[48] Heydar Radjavi and Peter Rosenthal, *Simultaneous triangularization*, Springer, New York, 2000.

[49] Peter Rosenthal, *Examples of invariant subspace lattices*, Duke Math. J. 37 (1970), 103–112.

[50] D.E. Sarason, *Invariant subspaces and unstarred operator algebras*, Pacific J. Math. 17 (1966), 511–517.

[51] D.E. Sarason, *Invariant subspaces*, Mathematical Surveys 13 (edited by C. Pearcy), A.M.S., Providence, 1974, pp. 1–47.

[52] Dan Voiculescu, *A non-commutative Weyl-von Neumann theorem*, Rev. Roumaine Math. Pures Appl. 21 (1976), 97–113.

Heydar Radjavi
Department of Mathematics
University of Waterloo
Waterloo
Ontario N2L 3G1, Canada
e-mail: `hradjavi@math.uwaterloo.ca`

Peter Rosenthal
Department of Mathematics
University of Toronto
Toronto
Ontario M5S 3G3, Canada
e-mail: `rosent@math.toronto.edu`

Operator Theory:
Advances and Applications, Vol. 207, 351–357

# Commutant Lifting

Donald Sarason

*In fond memory of Paul*

**Abstract.** This article is the story of how the author had the good fortune to be able to prove the primordial version of the commutant lifting theorem. The phrase "good fortune" is used advisedly. The story begins with the intersection of two lives, Paul's and the author's.

**Mathematics Subject Classification (2000).** Primary 47A20, 47A45; Secondary 47A15, 30D55.

**Keywords.** Commutant lifting, unitary dilation, operator models, Hardy spaces.

Paul Halmos's paper [6], one of his first two in pure operator theory, spawned three major developments: the theory of subnormal operators, the theory of hyponormal operators, and the theory of unitary dilations and operator models. Several of the articles in this volume, including this one, concern these developments; they illustrate how Paul's original ideas in [6] grew into major branches of operator theory. The present article belongs to the realm of unitary dilations.

The commutant lifting theorem of Béla Sz.-Nagy and Ciprian Foiaş is a centerpiece of the theory of unitary dilations, in large part because of its intimate connection with interpolation problems, including many arising in engineering. This article describes my own involvement with commutant lifting. It is a personal history that I hope conveys a picture of how research often gets done, in particular, how fortunate happenstance can play a decisive role.

My first piece of good fortune, as far as this story goes, was to be a mathematics graduate student at the University of Michigan when Paul Halmos arrived in Ann Arbor in the fall of 1961. At that time I had passed the Ph.D. oral exams but was unsure of my mathematical direction, except to feel it should be some kind of analysis. I had taken the basic functional analysis course the preceding fall but felt I lacked a good grasp of the subject. Paul was to teach the same course in fall 1961; I decided to sit in.

Prior to encountering Paul in person I was aware that he was well known and that he had written a book on measure theory; that was basically the extent of my knowledge. Paul's entry in the first class meeting, some 48 years ago, stands out in my memory. It was an electrifying moment – Paul had a commanding classroom presence. His course concentrated on Hilbert space, taught by his version of the Moore method, with an abundance of problems for the students to work on. Through the course I was encouraged to ask Paul to direct my dissertation, and he agreed. As a topic he suggested I look at invariant subspaces of normal operators.

Paul's basic idea in [6] is to use normal operators to gain insight into the structure of more general Hilbert space operators. The idea is a natural one; thanks to the spectral theorem, the structure of normal operators is well understood, at least at an abstract level.

Paul proved in [6] that every Hilbert space contraction has, in the terminology of the paper, a unitary dilation: given a contraction $T$ acting on a Hilbert space $H$, there is a Hilbert space $H'$ containing $H$ as a subspace, and a unitary operator $U$ on $H'$, such that one obtains the action of $T$ on a vector in $H$ by applying $U$ to the vector followed by the orthogonal projection of $H'$ onto $H$; in Paul's terminology, $T$ is the compression of $U$ to $H$. A few years after [6] appeared Béla Sz.-Nagy [23] improved Paul's result by showing that, with $H$ and $T$ as above, one can take the containing Hibert space $H'$ and the unitary operator $U$ on $H'$ in such a way that $T^n$ is the compression to $H$ of $U^n$ for every positive integer $n$. As was quickly recognized, Sz.-Nagy's improvement is a substantial one. For example, Sz.-Nagy's result has as a simple corollary the inequality of J. von Neumann: if $T$ is a Hilbert space contraction and $p$ is a polynomial, then the norm of $p(T)$ is bounded by the supremum norm of $p$ on the unit circle – an early success of the Halmos idea to use normal operators (in this case, unitary operators) to study more general operators. A dilation of the type Sz.-Nagy constructed was for a time referred to as a strong dilation, or as a power dilation; it is now just called a dilation.

After becoming Paul's student I was swept into several confluent mathematical currents. The Sz.-Nagy–Foiaş operator model theory, the creation of a remarkable collaboration spanning over 20 years, was in its relatively early stages, and was of course of great interest to Paul. Exciting new connections between abstract analysis and complex analysis were emerging, leading in particular to the subject of function algebras. These connections often involved Hardy spaces, which thus gained enhanced prominence. Kenneth Hoffman's book [10] embodied and propelled this ferment. (It is the only mathematics book I have studied nearly cover to cover.)

As my dissertation was slated to be about invariant subspaces of normal operators, I learned as much as I could about normal operators, picking up in the process the basics of vector-valued function theory. In the end, the dissertation focused on a particular normal operator whose analysis involved Hardy spaces in an annulus [17]. Simultaneously with working on my dissertation, I tried to understand the Sz.-Nagy–Foiaş theory.

I received my degree in the spring of 1963. From Ann Arbor I went to the Institute for Advanced Study, where I spent a year as an NSF postdoc before joining the Berkeley mathematics faculty in the fall of 1964 (just in time to witness the Free Speech Movement). I don't remember exactly when I started thinking about commutant lifting; most likely it happened at the Institute. Let me back up a little.

Every contraction has a unitary dilation, but not a unique one: given any unitary dilation, one can inflate it by tacking on a unitary direct summand. However, there is always a unitary dilation that is minimal, i.e., not producible by inflation of a smaller one, and this minimal unitary dilation is unique to within unitary equivalence. The simplest unitary operator that can be a minimal unitary dilation of an operator besides itself is the bilateral shift on $L^2$ of the unit circle, the operator $W$ on $L^2$ of multiplication by the coordinate function. In trying to understand the Sz.-Nagy–Foiaş theory better, I asked myself which operators (other than $W$ itself) can have $W$ as a minimal unitary dilation. Otherwise put, the question asks for a classification of those proper subspaces $K$ of $L^2$ with the property that $W$ is the minimal unitary dilation of its compression to $K$. Any such subspace, one can show, is either a nonreducing invariant subspace of $W$, or a nonreducing invariant subspace of $W^*$, or the orthogonal complement of a nonreducing invariant subspace of $W$ in a larger one. The invariant subspace structure of $W$ is given by the theorem of Arne Beurling [3] and its extension by Henry Helson and David Lowdenslager [7]. One concludes that the operators in question, besides $W$ itself, are, to within unitary equivalence, the unilateral shift $S$ (the restriction of $W$ to the Hardy space $H^2$), the adjoint $S^*$ of $S$, and the compressions of $S$ to the proper invariant subspaces of $S^*$.

By Beurling's theorem, the general proper, nontrivial, invariant subspace of $S$ is the subspace $uH^2$ with $u$ a nonconstant inner function. The corresponding orthogonal complement $K_u^2 = H^2 \ominus uH^2$ is the general proper, nontrivial, invariant subspace of $S^*$. The compression of $S$ to $K_u^2$ will be denoted by $S_u$. The operators $S_u$, along with $S$ and $S^*$, are the simplest Sz.-Nagy–Foiaş model operators.

Early in their program Sz.-Nagy and Foiaş defined an $H^\infty$ functional calculus for completely nonunitary contractions, among which are the operators $S_u$. For $\varphi$ a function in $H^\infty$, the operator $\varphi(S_u)$ is the compression to $K_u^2$ of the operator on $H^2$ of multiplication by $\varphi$. The operator $\varphi(S_u)$ depends only on the coset of $\varphi$ in the quotient algebra $H^\infty/uH^\infty$. One thereby gets an injection of $H^\infty/uH^\infty$ onto a certain operator algebra on $K_u^2$ whose members commute with $S_u$.

At some point, either when I was still in Ann Arbor or during my year at the Institute, I read James Moeller's paper [12], in which he determines the spectra of the operators $S_u$. Moeller's analysis shows that if the point $\lambda$ is not in the spectrum of $S_u$, the operator $(S_u - \lambda I)^{-1}$ is an $H^\infty$ function of $S_u$. This made me wonder whether every operator commuting with $S_u$ might not be an $H^\infty$ function of $S_u$.

In pondering this question, an obvious way for one to start is to look at the case where $u$ is a finite Blaschke product, in other words, where $K_u^2$ has finite dimension. In this case a positive answer lies near the surface, but more is true thanks to the solutions of the classical interpolation problems of Carathéodory–Fejér [4] and Nevanlinna–Pick [14], [16], to which I was led by my question. Those solutions tell you that if $u$ is a finite Blaschke product, then every operator on $K_u^2$ that commutes with $S_u$ is an $H^\infty$ function of $S_u$ for an $H^\infty$ function whose supremum norm equals the operator norm. Knowing this, one can generalize to the case of an infinite Blaschke product by means of a limit argument. Once I realized that, I was dead sure the same result holds for general $u$. But here I was stuck for quite a while; the behavior of inner functions that contain singular factors is subtler than that of those that do not. A step in the right direction, it seemed, would be to prove for general $u$ that the injection of $H^\infty/uH^\infty$ into $\mathcal{B}(K_u^2)$ (the algebra of operators on $K_u^2$) preserves norms, something the classical interpolation theory gives you for the Blaschke case. But on that I was stuck as well.

Good fortune accompanied me to Berkeley, where I became a colleague of Henry Helson. Some mathematicians, like me when I was younger, tend to keep to themselves the problems they are trying to solve; others, like Henry, are driven to talk with others about the problems, sharing their sometimes tentative ideas. In the Academic Year 1965–1966 Henry was working on a problem in prediction theory related to earlier work he had done with Gabor Szegö [9]. He would regularly drop by my office to discuss the problem. Back then prediction theory was a mystery to me, and I failed to understand very much of what Henry was saying. I did a lot of nodding, interrupted by an occasional comment or question. This was going on one Friday afternoon in fall 1965 when Henry brought up a proof he had found of Zeev Nehari's theorem on boundedness of Hankel operators [13]. Henry's proof, which is much slicker than the original one, uses ideas from his paper with Szegö, namely, a duality argument facilitated by a factorization result of Frigyes Riesz. (Riesz's result states that a nonzero function $f$ in the Hardy space $H^1$ can be factored as $f = f_1 f_2$, where $f_1$ and $f_2$ are in $H^2$, and $|f_1|^2 = |f_2|^2 = |f|$ almost everywhere on the unit circle.)

The following day it suddenly struck me that Henry's technique was exactly what I needed to show that the injection $H^\infty/uH^\infty \to \mathcal{B}(K_u^2)$ preserves norms in the general case, and also that it is weak-star-topology → weak-operator-topology continuous. Once I knew that I was able to combine it with what I already knew to prove the theorem I had long sought, which states: *Every operator $T$ on $K_u^2$ that commutes with $S_u$ equals $\varphi(S_u)$ for a function $\varphi$ in $H^\infty$ whose supremum norm equals $\|T\|$*. The proof was completed over the weekend. It uses some vector-valued function theory, including the vector generalization of Beurling's theorem due to Peter Lax [11], and ideas from an earlier paper of mine [18].

After proving the theorem I spent quite a while exploring some of its implications. I wrote up my results in the summer of 1966; the paper containing them [19] was published in May of 1967.

In [19] I was not brave enough to conjecture that my theorem generalizes to arbitrary unitary dilations (although I did prove a rather restrictive vector-valued generalization). It did not take long for Sz.-Nagy and Foiaş to produce the generalization, their famous commutant lifting theorem. Their paper [24] contains an informative discussion of the theorem.

Perhaps I should have looked more deeply into Hankel operators after Henry showed me his proof of Nehari's theorem, but I did not; my thoughts were elsewhere. Sometime in the 1970s Douglas Clark observed that my theorem is a fairly simple corollary of Nehari's. Clark did not publish his observation; it appears, though, in the notes for a course he gave at the University of Georgia. A bit later Nikolai Nikolski independently made the same observation. The derivation of my theorem from Nehari's can be found in Nikolski's book [15] (pp. 180ff.).

Following Sz.-Nagy and Foiaş's original proof of the commutant lifting theorem, several alternative proofs were found. My favorite, because it brings us back to Hankel operators, is due to Rodrigo Arocena [2].

In 1968 Vadim Adamyan, Damir Arov and Mark Kreĭn published the first [1] of a series of papers on Hankel operators. Originally they seemingly were unaware of Nehari's paper; a reference to Nehari was added to their paper in proof. Among other things, Adamyan–Arov–Kreĭn found a proof of Nehari's theorem along the same lines as the familiar operator theory approach to the Hamburger moment problem.

Nehari's theorem is a special case of the commutant lifting theorem. Arocena realized that the Adamyan–Arov–Kreĭn technique can be juiced up to give a proof of the full theorem. An exposition of Arocena's proof can be found in my article [21].

Ever since Sz.-Nagy and Foiaş proved their theorem, commutant lifting has played a central role in operator theory. A picture of the scope of the idea of commutant lifting, and of its engineering connections, can be found in the book of Foiaş and Arthur Frazho [5].

My own romance with commutant lifting seems to have come full circle. My recent paper [22] contains a version of commutant lifting for unbounded operators: If $T$ is a closed densely defined operator on $K_u^2$ that commutes with $S_u$, then $T = \varphi(S_u)$ for a function $\varphi$ in the Nevanlinna class ($\varphi = \psi/\chi$, where $\psi$ and $\chi$ are in $H^\infty$ and $\chi$ is not the zero function). Is there a general theorem in the theory of unitary dilations that contains this result?

Small footnote: I eventually understood enough about Henry's problem in prediction theory to contribute to its solution. The result is our joint paper [8] and my subsequent paper [20]. My work on [8] took place while Henry was on leave in France during the Academic Year 1966–1967, and our communication took place via airmail.

## References

[1] V.M. Adamyan, D.Z. Arov and M.G. Kreĭn, *Infinite Hankel matrices and generalized problems of Carathéodory–Fejér and I. Schur.* Funkcional Anal. Prilozhen. 2,4 (1968), 1–17; MR0636333 (58#30446).

[2] R. Arocena, *Unitary extensions of isometries and contractive intertwining dilations.* Operator Theory: Advances and Applications, Birkhäuser Verlag, Basel, 1989, vol. 41, pp. 13–23; MR1038328 (91c:47009).

[3] A. Beurling, *On two problems concerning linear transformations in Hilbert space.* Acta Math. 81 (1949), 239–255; MR0027954 (10,381e).

[4] C. Carathéodory and L. Fejér, *Über den Zusammenhang der Extremen von harmonischen Funktionen mit ihren Koeffizienten und über den Picard–Landau'schen Satz.* Rend. Circ. Mat. Palermo 32 (1911), 218–239.

[5] C. Foiaş and A. Frazho, *The commutant lifting approach to interpolation problems.* Operator Theory: Advances and Applications, Birkhäuser Verlag, Basel, 1990, vol. 44; MR1120546 (92k:47033).

[6] P.R. Halmos, *Normal dilations and extensions of operators.* Summa Brasil. Math. 2 (1950), 125–134; MR0044036 (13,259b).

[7] H. Helson and D. Lowdenslager, *Invariant subspaces.* Proc. Internat. Sympos. Linear Spaces (Jerusalem, 1960), 251–262, Jerusalem Academic Press, Jerusalem; Pergamon, Oxford, 1961; MR0157251 (28#487).

[8] H. Helson and D. Sarason, *Past and future.* Math. Scand. 21 (1967), 5–16; MR0236989 (38#5282).

[9] H. Helson and G. Szegö, *A problem in prediction theory.* Ann. Mat. Pura Appl. (4)51 (1960), 107–138; MR0121608 (22#12343).

[10] K. Hoffman, *Banach spaces of analytic functions.* Prentice–Hall Inc., Englewood Cliffs, NJ, 1962; MR0133008 (24#A2844). Reprinted by Dover Publications, Inc., New York, 1988; MR1102893 (92d:46066).

[11] P. Lax, *Translation invariant spaces.* Proc. Internat. Sympos. Linear Spaces (Jerusalem, 1960), 299–306, Jerusalem Academic Press, Jerusalem; Pergamon, Oxford, 1961; MR0140931 (25#4345).

[12] J.W. Moeller, *On the spectra of some translation invariant spaces.* J. Math. Anal. Appl. 4 (1962), 276–296; MR0150592 (27#588).

[13] Z. Nehari, *On bounded bilinear forms.* Ann. of Math. (2) 65 (1957), 153–162; MR0082945 (18,633f).

[14] R. Nevanlinna, *Über beschränkte Funktionen die in gegebenen Punktionen vorgeschriebene Werte annehmen.* Ann. Acad. Sci. Fenn. Ser. A 13 (1919), no. 1.

[15] N.K. Nikol'skiĭ, *Treatise on the shift operator.* Grundlehren der Mathematischen Wissenschaften, 273, Springer–Verlag, Berlin, 1986; MR0827223 (87i:47042).

[16] G. Pick, *Über die Beschränkungen analytischer Funktionen, welche durch vorgegebene Funktionswerte bewirkt werden.* Math. Ann. 77 (1916), 7–23.

[17] D. Sarason, *The $H^p$ spaces of an annulus.* Mem. Amer. Math. Soc. (1965), no. 56; MR0188824 (32#5236).

[18] D. Sarason, *Invariant subspaces and unstarred operator algebras.* Pacific J. Math. 17 (1966), 511–517; MR0192365 (33#590).

[19] D. Sarason, *Generalized interpolation in $H^\infty$.* Trans. Amer. Math. Soc. 127 (1967), 179–203; MR0208383 (34#8193).

[20] D. Sarason, *An addendum to "Past and Future".* Math. Scand. 30 (1972), 62–64; MR0385990 (52#6849).

[21] D. Sarason, *New Hilbert spaces from old.* Paul Halmos – Celebrating 50 Years of Mathematics, John H. Ewing and F.W. Gehring, editors, Springer–Verlag, New York, 1991, pp. 195–204.

[22] D. Sarason, *Unbounded operators commuting with restricted backward shifts.* Operators and Matrices, 2 (2008), no. 4, 583–601; MR2468883.

[23] B. Sz.-Nagy, *Sur les contractions de l'espace de Hilbert.* Acta Sci. Math. Szeged 15 (1953), 87–92; MR0058128 (15,326d).

[24] B. Sz.-Nagy and C. Foiaş, *The "Lifting Theorem" for intertwining operators and some new applications.* Indiana Univ. Math. J. 20 (1971), no. 10, 901–904; MR0417815 (54#5863).

Donald Sarason
Department of Mathematics
University of California
Berkeley, CA 94720-3840, USA
e-mail: sarason@math.berkeley.edu

Operator Theory:
Advances and Applications, Vol. 207, 359–362

# Double Cones are Intervals

V.S. Sunder

**Abstract.** The purpose of this short note is to point out the following fact and some of its pleasant 'consequences': the so-called double-cones in (4-dimensional) Minkowski space are nothing but the intervals $(A, B) = \{C \in H_2 : A < C < B\}$ in the space $H_2$ of $2 \times 2$ complex Hermitian matrices, where we write $X < Y$ if $(Y - X)$ is positive-definite.

**Mathematics Subject Classification (2000).** 47N50.

**Keywords.** Minkowski space, light cones, Hermitian matrices, positive-definite matrices.

## 1. Introduction

This short note is a result of two events:

(i) I was recently approached by the editors of a volume being brought out in the memory of Paul Halmos, and I certainly wanted to contribute some token of many fond memories of Paul; and

(ii) some time ago, I learned something with considerable pleasure which, I am sure, is just the kind of 'fun and games with matrices' that brought a gleam into Paul's eye.

## 2. The $(H_2, |\cdot|)$ model of Minkowski space

It must be stated that most of what follows is probably 'old hat' and the many things put down here are for the sake of setting up notation preparatory to justifying the assertion of the abstract.

In the language of the first pages of a physics text on relativity, Minkowski space is nothing but 4 $(= 1+3)$-dimensional real space $\mathbb{R}^4 = \{x = (x_0, x_1, x_2, x_3) : x_i \in \mathbb{R}\ \forall i\}$ equipped with the form defined by $q(x) = x_0^2 - x_1^2 - x_2^2 - x_3^2$. We shall prefer to work with a 'matricial' model (which might appeal more to an operator-theorist).

Thus we wish to consider the real Hilbert space

$$H_2 = \{\begin{pmatrix} a & z \\ \bar{z} & b \end{pmatrix} : a, b \in \mathbb{R}, z \in \mathbb{C}\}$$

of 2 × 2 complex Hermitian matrices, and observe that the assignment

$$\mathbb{R}^4 \ni x = (x_0, x_1, x_2, x_3) \overset{\phi}{\mapsto} X = \begin{pmatrix} x_0 + x_3 & x_1 - ix_2 \\ x_1 + ix_2 & x_0 - x_3 \end{pmatrix} \in H_2 \tag{2.1}$$

defines a (real-)linear isomorphism, and that we have the following identities:

$$q(x) = |X| \ , \ x_0 = \mathrm{tr}(X)$$

where $|X|$ denotes the determinant of the matrix $X$ and $\mathrm{tr}(X) = \frac{1}{2}\,\mathrm{Tr}(X)$ denotes the normalised trace (= the familiar matrix trace scaled so as to assign the value 1 to the identity matrix $I$).

It should be noted that the isomorphism $\phi$ of equation (2.1) can be alternatively written as

$$\phi(x) = \sum_{i=0}^{3} x_i \sigma_i$$

where $\sigma_0$ is the identity matrix, $I$ and $\sigma_1, \sigma_2, \sigma_3$ are the celebrated *Pauli matrices* and that $\{\sigma_0, \sigma_1, \sigma_2, \sigma_3\}$ is an orthonormal basis for $H_2$ with respect to the normalised Hilbert-Schmidt inner product given by $\langle X, Y\rangle = \mathrm{tr}(Y^* X)$, so $\phi$ is even a (real) unitary isomorphism.

Recall that the so-called *positive* and *negative light cones* (at the origin) are defined as

$$\begin{aligned} C^+(0) &= \{x \in \mathbb{R}^4 : x_0 > 0, q(x) > 0\} \\ C^-(0) &= \{x \in \mathbb{R}^4 : x_0 < 0, q(x) > 0\} \\ &= -C^+(0) \end{aligned}$$

while the 'positive light cone' and the 'negative light cone' at $x \in \mathbb{R}^4$ are defined as

$$\begin{aligned} C^+(x) &= \{y \in \mathbb{R}^4 : (y_0 - x_0) > 0, q(y - x) > 0\} \\ &= x + C^+(0) \\ C^-(x) &= \{y \in \mathbb{R}^4 : (y_0 - x_0) < 0, q(y - x) > 0\} \\ &= x - C^+(0) \ . \end{aligned}$$

Finally, a *double-cone* is a set of the form $D(x, y) = C^+(x) \cap C^-(y)$ (which is non-empty precisely when $y \in C^+(x)$).

The key observation for us is the following

Remark 2.1.

$$\phi(C^+(0)) = P \tag{2.2}$$

*where $P$ is the subset of $H_2$ consisting of* positive-definite *matrices.*

Reason: Any $X \in H_2$ has two real eigenvalues $\lambda^*(X)$ and $\lambda_*(X)$ satisfying $\lambda_*(X) \leq \lambda^*(X)$ and

$$\frac{1}{2}(\lambda_*(X) + \lambda^*(X)) = \operatorname{tr}(X),\ \lambda_*(X) \cdot \lambda^*(X) = |X|$$

so

$$X \in P \Leftrightarrow \lambda^*(X), \lambda_*(X) > 0 \Leftrightarrow \operatorname{tr}(X), |X| > 0.$$

Thus we do indeed find that

$$\begin{aligned} C^+(X) &= X + P \\ C^-(Y) &= Y - P \\ D(X,Y) = (X,Y) &= \{Z \in H_2 : X < Z < Y\}\ , \end{aligned}$$

where we write $A < C \Leftrightarrow C - A \in P$. (It should be emphasized that $A < C$ implies that $C - A$ is invertible, not merely positive semi-definite.) □

It should be observed that the relative compactness of these double cones is an easy corollary of the above remark. (I am told that this fact is of some physical significance.)

## 3. Some applications

This section derives some known facts using our Remark 2.1.

Proposition 3.1. *The collection* $\{D(x,y) : y - x \in C^+(0)\}$ (*resp.,* $\{(X,Y) : Y - X \in P\}$) *forms a base for the topology of* $\mathbb{R}^4$ (*resp.,* $H_2$).

*Proof.* Suppose $U$ is an open neighbourhood of a $Z \in H_2$. By definition, we can find $\epsilon > 0$ such that $W \in H_2, \|W - Z\| < 2\epsilon \Rightarrow W \in U$, where we write $\|A\| = \max\{\|Av\| : v \in \mathbb{C}^2, \|v\| = 1\}$. Then $X = Z - \epsilon I$, $Y = Z + \epsilon I$ satisfy $Z \in (X,Y) \subset U$. □

Thus the usual topology on $\mathbb{R}^4$ has a basis consisting of 'intervals'.

Recall next that the *causal complement* $O^\perp$ of a set $O \subset \mathbb{R}^4$ (resp., $H_2$) is defined by

$$O^\perp = \{z \in \mathbb{R}^4 : q(z - w) < 0\ \forall w \in O\}$$

(resp.,

$$O^\perp = \{Z \in H_2 : |Z - W| < 0\ \forall W \in O\}\ ).$$

Proposition 3.2.

$$(X,Y)^{\perp\perp} = (X,Y)\ \forall X < Y.$$

*Proof.* We shall find it convenient to write $P_0$ for the set of positive semi-definite matrices (i.e., $Z \in P_0 \Leftrightarrow \lambda_*(Z) \geq 0$).

We assert now that

$$\begin{aligned} Z \in (X,Y)^\perp \quad \Leftrightarrow \quad & \text{there exist unit vectors } v_1, v_2 \text{ such that} \\ & \langle Zv_1, v_1\rangle \leq \langle Xv_1, v_1\rangle \text{ and } \langle Zv_2, v_2\rangle \geq \langle Yv_2, v_2\rangle. \end{aligned} \tag{3.3}$$

Notice first that, by definition,

$$Z \in (X,Y)^{\perp} \Leftrightarrow |Z - W| < 0 \text{ whenever } W \in (X,Y)$$

Choose $\epsilon > 0$ such that $Y - X > \epsilon I$; so also $X + \epsilon I \in (X,Y)$ and $Y - \epsilon I \in (X,Y)$. If $Z \in (X,Y)^{\perp}$, we find that $|Z - (X + \epsilon I)| < 0$ and $|Z - (Y - \epsilon I)| < 0$. Now a $2 \times 2$ Hermitian matrix $C$ has negative determinant if and only if we can find unit vectors $v_1$ and $v_2$ such that $\langle Cv_1, v_1\rangle < 0 < \langle Cv_2, v_2\rangle$. Applying this to our situation, we can find unit vectors $v_1(\epsilon), v_2(\epsilon)$ such that $\langle Zv_1(\epsilon), v_1(\epsilon)\rangle < \langle Xv_1(\epsilon), v_1(\epsilon)\rangle + \epsilon$ and $\langle Zv_2(\epsilon), v_2(\epsilon)\rangle > \langle Yv_2(\epsilon), v_2(\epsilon)\rangle - \epsilon$. By compactness of the unit sphere in $\mathbb{C}^2$, we find, letting $\epsilon \downarrow 0$, that the condition (3.3) is indeed met.

Conversely, suppose condition (3.3) is met. If $W \in (X,Y)$, observe that $\langle Zv_1, v_1\rangle \leq \langle Xv_1, v_1\rangle < \langle Wv_1, v_1\rangle$ and $\langle Zv_2, v_2\rangle \geq \langle Yv_2, v_2\rangle > \langle Wv_2, v_2\rangle$ from which we may conclude that indeed $|Z - W| < 0$.

Since $O \subset O^{\perp\perp}\ \forall O$, we only need to prove that

$$W \notin (X,Y) \Rightarrow \exists Z \in (X,Y)^{\perp} \text{ such that } |W - Z| \geq 0 \ (\Rightarrow W \notin (X,Y)^{\perp\perp}) .$$

If $W \notin (X,Y)$, we can find orthonormal vectors $\{v_1, v_2\}$ such that either $\langle Wv_1, v_1\rangle \leq \langle Xv_1, v_1\rangle$ or $\langle Wv_2, v_2\rangle \geq \langle Yv_2, v_2\rangle$. Suppose the former holds. (The other case is settled in the same manner.) Define the operator $Z$ by $Z = W + N|v_2\rangle\langle v_2|$ for some $N$ chosen so large as to ensure that $\langle Zv_2, v_2\rangle > \langle Yv_2, v_2\rangle$. We then find that also $\langle Zv_1, v_1\rangle = \langle Wv_1, v_1\rangle \leq \langle Xv_1, v_1\rangle$, and so we may deduce from condition (3.3) that $Z \in (X,Y)^{\perp}$. Since the construction ensures that $(Z-W)v_1 = 0$, we find that $|Z - W| = 0$, and the proof of the proposition is complete. □

We close with a minimal bibliography; all the background for the mathematics here can be found in [PRH], while physics-related topics such as Minkowski metric, double cones, etc., are amply treated in [BAU]. In fact, it was trying to understand the proof given in [BAU] for what we call Proposition 3.2 which led to this whole exercise.

## References

[PRH] Paul R. Halmos, *Finite-dimensional vector spaces*, van Nostrand, Princeton, 1958.

[BAU] H. Baumgaertel, *Operatoralgebraic Methods in Quantum Field Theory. A Set of Lectures*, (Akademie Verlag, 1995).

V.S. Sunder
The Institute of Mathematical Sciences
Chennai 600113, India
e-mail: sunder@imsc.res.in

## Operator Theory: Advances and Applications (OT)

BIRKHÄUSER

## Operator Theory: Advances and Applications (OT)

Edited by
**Joseph A. Ball** (Blacksburg, VA, USA), **Harry Dym** (Rehovot, Israel), **Marinus A. Kaashoek** (Amsterdam, The Netherlands), **Heinz Langer** (Vienna, Austria), **Christiane Tretter** (Bern, Switzerland)

This series is devoted to the publication of current research in operator theory, with particular emphasis on applications to classical analysis and the theory of integral equations, as well as to numerical analysis, mathematical physics and mathematical methods in electrical engineering.

**OT 192: Gohberg, I. / Leiterer, J.**, Holomorphic Operator Functions of One Variable and Applications. Methods from Complex Analysis in Several Variables (2009).
ISBN 978-3-0346-0125-2

**OT 190/191: Adamyan, V. / Berezansky, Yu.M. / Gohberg, I. / Gorbachuk, M.L. / Gorbachuk, V. / Kochubei, A.N. / Langer, H. / Popov, G.** (eds.), Modern Analysis and Applications. (2009)
Volume 1 (OT 190): Operator Theory and Related Topics. ISBN 978-3-7643-9918-4
Volume 2 (OT 191): Differential Operators and Mechanics. ISBN 978-3-7643-9920-7
Set Volumes 1/2: ISBN 978-3-7643-9924-5

**OT 189: Rodino, L. / Wong, M.W.** (eds.), New Developments in Pseudo-Differential Operators (2008). ISBN 978-3-7643-8968-0

**OT 188: Behrndt, J. / Förster, K.-H. / Langer, H. / Trunk, C.** (eds.), Spectral Theory in Inner Product Spaces and Applications (2008).
ISBN 978-3-7643-8910-9

**OT 187: Ando, T. / Curto, R.E. / Jung, I.B. / Lee, W.Y.** (eds.), Recent Advances in Operator Theory and Applications (2008).
ISBN 978-3-7643-8892-8

**OT 186: Janas, J. / Kurasov, P. / Laptev, A. / Naboko, S. / Stolz, G.** (eds.), Methods of Spectral Analysis in Mathematical Physics (2008).
ISBN 978-3-7643-8754-9

**OT 185: Vasilevski, N.**, Commutative Algebras of Toeplitz Operators on the Bergman Space (2008).
ISBN 978-3-7643-8725-9

**OT 184: Qin, Y.**, Nonlinear Parabolic-Hyperbolic Coupled Systems and Their Attractors (2008). Subseries **A**dvances in **P**artial **D**ifferential **E**quations. ISBN 978-3-7643-8813-3

**OT 183: Nazaikinskii, V.E.. / Savin, A.Yu. / Sternin, B.Yu.**, Elliptic Theory and Noncommutative Geometry (2008). Subseries **A**dvances in **P**artial **D**ifferential **E**quations.
ISBN 978-3-7643-8774-7

**OT 182: van der Mee, C.**,
Exponentially Dichotomous Operators and Applications (2008). Subseries **L**inear **O**perators and **L**inear **S**ystems. ISBN 978-3-7643-8732-7

**OT 181: Bastos, A. / Gohberg, I. / Lebre, A.B. / Speck, F.-O.** (eds.), Operator Algebras, Operator Theory and Applications (2008).
**OT 180: Okada, S. / Ricker, W.J. / Sánchez Pérez, E.A.**, Optimal Domain and Integral Extension of Operators Acting in Function Spaces (2008). ISBN 978-3-7643-8647-4

**OT 179: Ball, J.A. / Eidelman, Y. / Helton, J.W. / Olshevsky, V. / Rovnyak, J.** (eds.),
Recent Advances in Matrix and Operator Theory (2007). ISBN 978-3-7643-8538-5

**OT 178: Bart, H. / Gohberg, I. / Kaashoek, M.A. / Ran, A.C.M.**,
Factorization of Matrix and Operator Functions: The State Space Method (2007).
ISBN 978-3-7643-8267-4

**OT 177: López-Gómez, J. / Mora-Corral, C.**, Algebraic Multiplicity of Eigenvalues of Linear Operators (2007). ISBN 978-3-7643-8400-5

**OT 176: Alpay, D. / Vinnikov, V.** (Eds.),
System Theory, the Schur Algorithm and Multidimensional Analysis (2007).
ISBN 978-3-7643-8136-3

**OT 175: Förster, K.-H. / Jonas, P. / Langer, H. / Trunk, C.** (Eds.), Operator Theory in Inner Product Spaces (2007).
ISBN 978-3-7643-8269-8