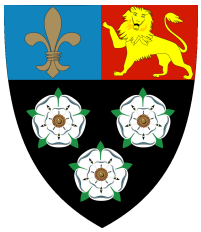
On the Approximation of Spectra of Linear Hilbert Space Operators

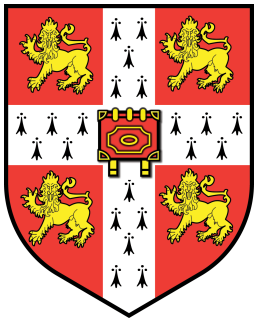
Anders C. Hansen

King's College

University of Cambridge

25 April 2008





This dissertation is submitted for the degree of Doctor of Philosophy.

Abstract

Title: On the Approximation of Spectra of Linear Hilbert Space Operators

Author: Anders C. Hansen

The main topic of this thesis is how to approximate and compute spectra of linear operators on separable Hilbert spaces. We consider several approaches including the finite section method, an infinite-dimensional version of the QR algorithm, as well as pseudospectral techniques. Several new theorems about convergence of the finite section method (and variants of it) for self-adjoint problems are obtained together with a rigorous analysis of the infinite-dimensional QR algorithm for normal operators. To attack (and solve) the long standing general computational spectral problem we look to the pseudospectral theory and introduce the complexity index. A generalization of the pseudospectrum is introduced, namely, the n-pseudospectrum. This set behaves very much like the original pseudospectrum, but has the advantage that it approximates the spectrum well for large n. The complexity index is a tool for indicating how complex or difficult it may be to approximate spectra of operators belonging to a certain class. We establish bounds on the complexity index and discuss some open problems regarding this new mathematical entity.

As the approximation framework also gives rise to several computational methods, we analyze and discuss implementation techniques for algorithms that can be derived from the theoretical model. In particular, we develop algorithms that can compute spectra of arbitrary bounded operators on separable Hilbert spaces, and the exposition is followed by several numerical examples. The thesis also contains a thorough discussion on how to implement the QR algorithm in infinite dimensions. This is supported by numerical computations. These examples reveal several surprisingly nice features of the infinite-dimensional QR algorithm, and this leaves a number of open problems that we debate. We also include a chapter on how the infinite-dimensional QR algorithm can be improved, in particular, how to speed it up. This approach is based on adapting the techniques used in finite dimensions to an infinite-dimensional setting.

Acknowledgements

Firstly, I would like to thank my advisor Arieh Iserles for sharing his excellent and very deep mathematical insight, for his encouragement, for always keeping his office door open and for the interesting conversations at the weekly Saturday lunches at Tanh-Binh. Having had Arieh as my advisor has been like having a cool dad who lets you stay out late, but still makes sure that you will stay out of trouble.

Secondly, there are four mathematicians that have been indispensable in my career so far, namely, (in order of appearance) Syvert Nørsett (NTNU), John Strain (Berkeley), Don Sarason (Berkeley) and Erik Bédos (Univ. of Oslo). I must thank Syvert for essentially starting my career by believing in me as an undergraduate and for helping me coming to Berkeley, no Syvert, no PhD for me. Equally indispensable has John been as my advisor at Berkeley, and I owe him many thanks for taking me under his wing. I must thank Professor Sarason (as we always called him) for teaching me most of the analysis I learned at Berkeley. He is not only a fantastic mathematician, but simply the best math educator I have experienced. My greatest thanks also goes to Erik for always keeping his office open when I am in Oslo and for countless math discussions.

Thirdly, I must thank six superb mathematicians that I admire greatly and that have been a great source of inspiration. They have influenced the results in this thesis directly and indirectly through discussions and criticism both from a pure and applied mathematical point of view. I would therefore like to thank (in alphabetical order): Bill Arveson (Berkeley), Brian Davies (King's College London), Percy Deift (NYU), Olavi Nevanlinna (Helsinki Univ. of Technology), Barry Simon (Caltech) and Nick Trefethen (Oxford).

Preface

The question I would like to address in this thesis is how to compute, or approximate, spectra of arbitrary linear operators on Hilbert spaces. When confronted with this challenge one does not only meet mathematical obstacles, but one is also faced with the task of balancing between pure and applied mathematics. As most of the interesting operators in mathematical physics act on infinite-dimensional Hilbert spaces, one must leave the classical theory of matrix computations and analysis and enter the more pure discipline of functional analysis and operator theory. However, as actual computations and design of algorithms are also important parts of this thesis, one cannot abandon the theory of matrix computations, but rather mix the two disciplines in the best possible way. This is a nontrivial task as the audience in the two different areas may have different interests and emphasis, and more importantly, different mathematical backgrounds and opinions. As Peter Lax said in his mathematical talk on the occasion of his acceptance of the Abel prize :"...and the relationship between the two disciplines (pure and applied mathematics) is delicate."

In order to please both pure and applied mathematicians I have chosen to give the two different communities what they appreciate, namely, mathematics written in their own language. The thesis is therefore organized in two parts; Part I-Theory and Part II-Applications. Part I contains the results that are intended for cross-disciplinary mathematical journals (pure and applied mathematics) and is written in a language expected for journals such as Journal of the American Mathematical Society or Communications on Pure and Applied Mathematics. Knowledge of functional analysis at graduate level is assumed. In Part II the emphasis is on mathematical results in application, where the design and execution of algorithms are the main topics. These results are intended for applied mathematical journals such as IMA Journal of Numerical Analysis or Proceedings of the Royal Society A. In this part graduate level functional analysis is not assumed, however, the reader is expected to know graduate level numerical linear algebra and numerical analysis.

I must emphasize that this does not mean that all the theory is presented in Part I, whereas the numerical examples are shown in Part II. The second part of the thesis contains theorems and proofs as well as theoretical mathematical discussions, but the emphasis is on mathematics in applications and theorems are often motivated by the desire to analyze algorithms. Even though the material in both parts is intimately connected, both Part I and Part II are self contained and can be read independently of the other.

This dissertation is the result of my own work and includes nothing which is the outcome of work done in collaboration except where specifically indicated in the text. Every lemma, proposition or theorem followed by a proof is original. There are quotes

from other work, however, in those cases there are references to either the original paper or book and no proofs are displayed. The thesis is based on a series of articles, in particular:

Part I-Theory

- On the approximation of spectra of linear operators on Hilbert spaces (Han08).
- On the complexity index, the n-pseudospectrum and construction of spectra of linear operators (Han11).

Part II-Applications

- Infinite-dimensional numerical linear algebra; theory and applications (Han10).
- The infinite-dimensional QR algorithm (Hanb).
- Hessenberg reduction and the infinite-dimensional QR algorithm (Hana).

Table of Contents

Part I

Theory

Chapter 1

Introduction

The main theme of Part I is how to construct and approximate spectra of arbitrary closed operators on separable Hilbert spaces. This task is a fascinating pure mathematical problem, but it is strongly motivated by applications. The reason is as follows. After the triumph of quantum mechanics, operator and spectral theory became indispensable mathematical disciplines in order to support quantum theory and also to secure its mathematical foundations. There is a vast literature on how to analyze spectra of linear operators and the field is still very much active.

So far, so good, the only problem is that the theoretical physicist may not only want theorems about structural properties of spectra, one may actually want to determine the spectra completely. When faced with this problem the mathematician may first recall that even if the dimension of the Hilbert space is finite, this is not trivial. One quickly realizes that, due to Abel's contribution on the unsolvability of the quintic using radicals, one is doomed to fail if one tries to construct the spectrum in terms of finitely many arithmetic operations and radicals of the matrix elements of the operator. However, in finite dimensions, there is a vast theory on how to obtain sequences of sets, whose construction only require finitely many arithmetic operations and radicals of the matrix elements, such that the sequence converges to the spectrum of the desired operator. Thus, at least in finite dimensions, one can construct the spectrum, and this construction automatically yields a method for approximating the spectrum. Even though this may be difficult in practice, one has a mathematical theory that guarantees that up to an arbitrarily small error, one can determine the spectra of operators on finite dimensional Hilbert spaces.

There is no automatic extension from the finite dimensional case, and the problem is therefore; what can be done in infinite dimensions? Moreover, how does one handle the case of an unbounded operator? Keeping the Schrödinger and Dirac operators in mind, one realizes that the unbounded cases may be the most important ones. We must emphasize that quite a lot is known about how to approximate spectra of Schrödinger and Dirac operators, but, as far as we know, even in the self-adjoint case, one still only knows how to deal with special cases, and current methods lack generality. To illustrate the present situation on how how to construct and approximate spectra of arbitrary operators we have chosen quotes from two of the leading authorities in operator and spectral theory. Bill Arveson points out that "Unfortunately, there is a dearth of literature on this basic problem, and so far as we have been able to tell, there are no proven techniques"(Arv94a). Also, Brian Davies (Dav05) has expressed his concern due to the following example. Let

$A_\epsilon : l^2(\mathbb{Z}) \to l^2(\mathbb{Z})$ be defined by

$$(A_\epsilon f)(n) = \begin{cases} \epsilon f(n+1) & n = 0 \\ f(n+1) & n \neq 0. \end{cases}$$

Now for $\epsilon \neq 0$ we have $\sigma(A_\epsilon) = \{z : |z| = 1\}$ but for $\epsilon = 0$ then $\sigma(A_0) = \{z : |z| \leq 1\}$. Davies argues as follows: "If ϵ is a very small constructively defined real number and one is not able to determine whether or not $\epsilon = 0$, then the spectrum of A_ϵ cannot be computed even approximately even though A_ϵ is well-defined constructively. This implies that there exist straightforward bounded operators whose spectrum will probably never be determined."

We will emphasize that these quotes are concerned with the general problem, and if one has more structure available e.g. self-adjointness, then much more can be said. However, during the last two decades the importance of non-normal operators and their spectra has become increasingly evident. In particular, the growing interest in non-Hermitian quantum mechanics (HN97), (HN96), non-self-adjoint differential operators (Dav02), (DSZ04) and in general non-normal phenomena (TC04), (TE05) has made non-self-adjoint operators and pseudospectral theory indispensable. This emphasizes the importance of the general problem and poses a slightly philosophical problem, namely, could there be operators whose spectra we can never determine. If such operators are indispensable in areas of mathematical physics it may lead to serious restrictions to our possible understanding of some physical systems.

In Chapter 2 we will discuss two well known approaches, namely, the finite section method and the infinite dimensional QR-algorithm. There is a vast literature on the finite section method, and we will only give reference to a selection of the written work on this topic (Arv91), (Arv93b), (Arv93a), (Arv94a), (Arv94b), (Bro06), (Bro07a). Our approach is very much inspired by the work of Arveson and Brown.

The infinite dimensional QR algorithm occurred first in the paper "Toda Flows with Infinitely Many Variables" (DLT85) by Deift, Li and Tomei. The Toda Flow is normally associated with tridiagonal self-adjoint matrices, and therefore, the emphasis in (DLT85) is on self-adjoint problems. In Chapter 2 we will show some new results regarding the convergence of the infinite dimensional QR algorithm for normal operators. Also, it should be pointed out that even though the results presented here are very close in spirit to the theory presented in (DLT85), the mathematical approaches deviate substantially.

Chapter 3 is devoted to the general non-normal spectral problem, and the most important tool used is the Complexity Index. The Complexity Index is meant as a device for classification of spectral problems. To be more explicit one would like to determine how difficult it is to approximate spectra of operators in a certain class, e.g. it may be easier to approximate spectra of compact operators than non-compact, and the Complexity Index is suited for these issues.

The last chapter in Part I is devoted to Szegő-type theorems (Sze20) on convergence of densities, where we start by extending some of the results by Arveson in (Arv93a), (Arv94a), (Arv94b) to unbounded operators. These results are exclusive to self-adjoint operators, but by using some very intriguing developments by Haagerup and Shultz (HS07) we are able to introduce some non-normal Szegő-type theorems. The crucial ingredient in this framework is the Brown measure (Bro86).

1.1 Background and Notation

We will in this section review some basic definition and introduce the notation used in the dissertation. Throughout the thesis $\mathcal{H}$ will always denote a separable Hilbert space, $\mathcal{B}(\mathcal{H})$ the set of bounded linear operators, $\mathcal{C}(\mathcal{H})$ the set of densely defined closed linear operators and $\mathcal{SA}(\mathcal{H})$ the set of self-adjoint operators on $\mathcal{H}$. For $T \in \mathcal{C}(\mathcal{H})$ the domain of T will be denoted by $\mathcal{D}(T)$ and the spectrum by $\sigma(T)$. Also, if $T - z$ is invertible, for $z \in \mathbb{C}$, we use the notation $R(z, T) = (T - z)^{-1}$. We will denote orthonormal basis elements of $\mathcal{H}$ by e_j, and if $\{e_j\}_{j\in\mathbb{N}}$ is a basis and $\xi \in \mathcal{H}$ then $\xi_j = \langle \xi, e_j \rangle$. The word basis will always refer to an orthonormal basis. If $\mathcal{H}$ is a finite-dimensional Hilbert space with a basis $\{e_j\}$ then $LT_{\text{pos}}(\mathcal{H})$ will denote the set of lower triangular matrices (with respect to $\{e_j\}$) with positive elements on the diagonal. The closure of a set $\Omega \in \mathbb{C}$ will be denoted by $\overline{\Omega}$ or $\text{cl}(\Omega)$. Throughout the thesis we will only consider operators T such that $\sigma(T) \neq \mathbb{C}$ and $\sigma(T) \neq \emptyset$, hence this assumption will not be specified in any of the upcoming theorems.

Convergence of sets in the complex plane will be quite crucial in our analysis and hence we need the Hausdorff metric as defined by the following.

Definition 1.1.1. *(i) For a set $\Sigma \subset \mathbb{C}$ and $\delta > 0$ we will let $\omega_\delta(\Sigma)$ denote the δ-neighborhood of Σ (i.e. the union of all δ-balls centered at points of Σ).*

(ii) Given two sets $\Sigma, \Lambda \subset \mathbb{C}$ we say that Σ is δ-contained in Λ if $\Sigma \subset \omega_\delta(\Lambda)$.

(iii) Given two compact sets $\Sigma, \Lambda \subset \mathbb{C}$ their Hausdorff distance is

$$d_H(\Sigma, \Lambda) = \max\{\sup_{\lambda\in\Sigma} d(\lambda, \Lambda), \sup_{\lambda\in\Lambda} d(\lambda, \Sigma)\}$$

where $d(\lambda, \Lambda) = \inf_{\rho\in\Lambda} |\rho - \lambda|$.

If $\{\Lambda_n\}_{n\in\mathbb{N}}$ is a sequence of compact subsets of $\mathbb{C}$ and $\Lambda \subset \mathbb{C}$ is compact such that $d_H(\Lambda_n, \Lambda) \to 0$ as $n \to \infty$ we may use the notation $\Lambda_n \longrightarrow \Lambda$.

As for the convergence of operators we follow the notation in (Kat95). Let $E \subset \mathcal{B}$ and $F \subset \mathcal{B}$ be closed subspaces of a Banach space $\mathcal{B}$. Define

$$\delta(E, F) = \sup_{\substack{x\in E \\ \|x\|=1}} \inf_{y\in F} \|x - y\|$$

and

$$\hat{\delta}(E, F) = \max[\delta(E, F), \delta(F, E)].$$

If A and B are two closed operators, with domains $\mathcal{D}(A)$ and $\mathcal{D}(B)$, their graphs

$$G(A) = \{(\xi, \eta) \in \mathcal{H} \times \mathcal{H} : \xi \in \mathcal{D}(A), \eta = A\xi\} \tag{1.1.1}$$

and $G(B)$ are closed subspaces of $\mathcal{H} \times \mathcal{H}$. We can therefore define (with a slight abuse of notation) the distance between A and B by

$$\hat{\delta}(A, B) = \hat{\delta}(G(A), G(B)).$$

If $\{T_n\}_{n\in\mathbb{N}}$ is a sequence of closed operators converging in the distance suggested above to a closed operator T then we may sometimes use the notation

$$T_n \xrightarrow{\hat{\delta}} T, \qquad n \to \infty.$$

Note that $\hat{\delta}$ is not a metric. To define a metric on $\mathcal{C}(\mathcal{H})$ there are several possibilities. We will discuss two approaches here that will be useful later on in the paper. For closed operators A and B define

$$d(A,B) = \max\left[\sup_{\xi\in S_A} \operatorname{dist}(\xi, S_B), \sup_{\xi\in S_B} \operatorname{dist}(\xi, S_A)\right],$$

where $\operatorname{dist}(\xi, S_A) = \inf_{\eta\in S_A} \|\xi - \eta\|$ and S_A and S_B are the unit spheres of $G(A)$ and $G(B)$, respectively. As shown in (Kat95) d is a metric on $\mathcal{C}(\mathcal{H})$ with the property that

$$\hat{\delta}(A,B) \leq d(A,B) \leq 2\hat{\delta}(A,B).$$

A more practical metric for our purpose is the one suggested in (CL63). The definition is as follows

$$p(A,B) = [\|R_A - R_B\|^2 + \|R_{A^*} - R_{B^*}\|^2 + 2\|AR_A - BR_B\|^2]^{1/2},$$

where $R_A = (1 + A^*A)^{-1}$. For our purposes it is important to link p to $\hat{\delta}$ and that follows from the fact, as proved in (CL63), that p and d are equivalent as metrics on $\mathcal{C}(\mathcal{H})$. In particular we have

$$d(A,B) \leq \sqrt{2}p(A,B) \leq 2d(A,B).$$

The following fact will be useful in the later developments.

Theorem 1.1.2. *((Kat95) p.204) Let $T, S \in \mathcal{C}(\mathcal{H})$ and $A \in \mathcal{B}(\mathcal{H})$. Then*

$$\hat{\delta}(S + A, T + A) \leq 2(1 + \|A\|^2)\hat{\delta}(S,T).$$

Chapter 2

Finite Sections and Infinite QR

This chapter follows up on the ideas initiated by Arveson in (Arv94a) and (Arv91), (Arv93b), (Arv93a), (Arv94b) on how to approximate spectra of linear operators on separable Hilbert spaces using the finite section method. We extend several of the theorems initiated by Arveson and also introduce some new variants of the finite section method that are beneficial for some special structured problems. These new ideas come from well known techniques in matrix analysis and we show how to extend these approaches to infinite dimensions. We also investigate the method introduced by Deift, Li and Tomei in (DLT85), namely, the infinite-dimensional QR algorithm. The work done here was developed independently of the work in (DLT85), and the author is indebted to Percy Deift for pointing out the connection. However, Deift et al. should be credited for being the first to discover the surprisingly useful infinite-dimensional version of the QR algorithm. Our techniques involve normal operators whereas the theory in (DLT85) focuses on the self-adjoint case, and thus our frameworks deviate substantially.

2.1 Quasidiagonality and the Finite-Section Method

The finite-section method for approximating the spectrum of bounded self-adjoint operators on Hilbert spaces is a well-known technique and has been studied in several articles and monographs (Arv94a), (Bro07a), (BS99), (HRS01). The approach is to first find a sequence of finite rank projections $\{P_n\}$ such that $P_{n+1} \geq P_n$ and $P_n \to I$ strongly, and then use known techniques to find the spectrum of the compression $A_n = P_n A P_n$.

The most obvious approach is to use some orthonormal basis $\{e_n\}$ for the Hilbert space $\mathcal{H}$ and then let P_n be the projection onto $\mathrm{sp}\{e_1, \ldots e_n\}$. Given a self-adjoint $A \in \mathcal{B}(\mathcal{H})$ and $\{e_n\}$ we may consider the associate infinite matrix (a_{ij})

$$a_{ij} = \langle Ae_j, e_i \rangle, \quad i, j = 1, 2, \ldots.$$

In this case the compression becomes $A_n \in \mathcal{B}(\mathcal{H}_n)$, where $\mathcal{H}_n = P_n\mathcal{H}$, $A_n = P_n A \lceil_{\mathcal{H}_n}$, where the matrix with respect to $\{e_1, \ldots, e_n\}$ is

$$A_n = \begin{pmatrix} a_{11} & a_{12} & \ldots & a_{1n} \\ a_{21} & a_{22} & \ldots & a_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ a_{n1} & a_{n2} & \ldots & a_{nn} \end{pmatrix}.$$

The operator-theoretical question is to analyze how the spectrum $\sigma(P_nA\lceil_{P_n\mathcal{H}})$ evolves as $n \to \infty$.

Definition 2.1.1. *Given a sequence $\{A_n\} \subset \mathcal{B}(\mathcal{H})$, define*

$$\Lambda = \{\lambda \in \mathbb{R} : \exists \lambda_n \in \sigma(A_n), \lambda_n \to \lambda\}.$$

Also, for every set S of real numbers let $N_n(S)$ (and $\tilde{N}_n(S)$) denote the number of eigenvalues counting multiplicity (and not counting multiplicity respectively) of A_n which belong to S.

Definition 2.1.2. *(i) A point $\lambda \in \mathbb{R}$ is called essential if, for every open set $U \subset \mathbb{R}$ containing λ, we have*

$$\lim_{n\to\infty} N_n(U) = \infty.$$

The set of essential points is denoted Λ_e

(ii) $\lambda \in \mathbb{R}$ is called transient if there is an open set $U \subset \mathbb{R}$ containing λ such that

$$\sup_{n\geq 1} N_n(U) < \infty.$$

Theorem 2.1.3. *(Arveson)(Arv94a) Let $A \in \mathcal{B}(\mathcal{H})$ and let $\{P_n\}$ be a sequence of projections converging strongly to the identity such that $P_{n+1} \geq P_n$. Define $A_n = P_nA\lceil_{P_n\mathcal{H}}$ and let Λ and Λ_e be as in definitions 2.1.1 and 2.1.2. Then $\sigma(A) \subset \Lambda$ and $\sigma_e(A) \subset \Lambda_e$.*

Definition 2.1.4. *(i) A filtration of $\mathcal{H}$ is a sequence $\mathcal{F} = \{\mathcal{H}_1, \mathcal{H}_2, \ldots\}$ of finite dimensional subspaces of $\mathcal{H}$ such that $\mathcal{H}_n \subset \mathcal{H}_{n+1}$ and*

$$\overline{\bigcup_{n=1}^{\infty} \mathcal{H}_n} = \mathcal{H}$$

(ii) Let $\mathcal{F} = \{\mathcal{H}_n\}$ be a filtration of $\mathcal{H}$ and let P_n be the projection onto $\mathcal{H}_n$. The degree of an operator $A \in \mathcal{B}(\mathcal{H})$ is defined by

$$deg(A) = \sup_{n\geq 1} \operatorname{rank}(P_nA - AP_n).$$

Arveson gave in (Arv94a), (Arv94b) a complete theory of the finite-section method applied to operators of finite degree, which is an abstraction of band-limited infinite matrices. We will not discuss that theory here, but refer the reader to the original articles. We will however present the following theorem, which is a special case of Theorem 3.8 in (Arv94a), to give the reader an impression of what one can expect to get when using the finite-section method.

Theorem 2.1.5. *(Arveson)(Arv94a) Let $A \in \mathcal{B}(\mathcal{H})$ be self-adjoint and*

$$\mathcal{F} = \{\mathcal{H}_1, \mathcal{H}_2, \ldots\}$$

be a filtration with corresponding projections $\{P_n\}$. Define $A_n = P_nA\lceil_{P_n\mathcal{H}}$ and let Λ and Λ_e be as in definitions 2.1.1 and 2.1.2. Suppose that A has finite degree with respect to $\mathcal{F}$. Then

(i) $\sigma_e(A) = \Lambda_e$

(ii) Every point of Λ *is either transient or essential.*

In this section we will investigate how the finite section method can be applied to quasi-diagonal operators. First we recall some basic definitions as well as some well know results.

Definition 2.1.6. *An operator* A *on a separable Hilbert space is diagonal if there exists a complete orthonormal set of eigenvectors of* A*.*

Definition 2.1.7. *An operator* A *on a separable Hilbert space is quasi-diagonal if there exists an increasing sequence* $\{P_n\}$ *of finite rank projections such that* $P_n\mathcal{H} \subset \mathcal{D}(A)$*,* $P_n \to I$*, strongly, and* $\|P_nA - AP_n\| \to 0$*. The sequence* $\{P_n\}$ *is said to quasi-diagonalize* A*.*

Before the next definition we need to recall that an unbounded operator A is said to commute with the bounded operator T if

$$TA \subset AT.$$

This means that whenever $\xi \in \mathcal{D}(A)$, then $T\xi$ also belongs to $\mathcal{D}(A)$ and $AT\xi = TA\xi$.

Definition 2.1.8. *An operator* A *on a separable Hilbert space is said to be block diagonal with respect to an increasing sequence* $\{P_n\}$ *of finite-dimensional projections converging strongly to* I *if* A *commutes with* $P_{n+1} - P_n$ *for all* n*.*

Note that if A is self-adjoint and $P_n\mathcal{H} \subset \mathcal{D}(A)$ then Definition 2.1.8 is equivalent to each of the assertions

(i) P_n commutes with A for every n.

(ii) $AP_n\mathcal{H} \subset P_n\mathcal{H}$.

The following theorem assures us the existence of a vast set of quasi-diagonal operators.

Theorem 2.1.9. *(Weyl,von Neumann, Berg)(Ber71) Let* A *be a (not necessarily bounded) normal operator on the separable Hilbert space* $\mathcal{H}$*. Then for* $\epsilon > 0$ *there exist a diagonal operator* D *and a compact operator* C *such that* $\|C\| < \epsilon$ *and* $A = D + C$*.*

Corollary 2.1.10. *Every normal operator is quasi-diagonal.*

We will need a couple of basic lemmas.

Lemma 2.1.11. *(Davies, Plum)(DP04) Let* $A \in \mathcal{B}(\mathcal{H})$ *be self-adjoint,* P *be a projection and* $\epsilon > 0$ *such that* $\|PAP - AP\| \leq \epsilon$*. If* $\lambda \in \sigma(PAP)$ *then* $(\lambda - \epsilon, \lambda + \epsilon) \cap \sigma(A) \neq \emptyset$*.*

Lemma 2.1.12. *Let* $A \in \mathcal{B}(\mathcal{H})$ *be self-adjoint and compact. Let* $\{P_n\}$ *be a sequence of finite-dimensional projections such that* $P_n \to I$ *strongly. Then* $P_nAP_n \to A$ *in norm.*

Proof. Since $P_n^\perp = I - P_n$ is a sequence of projections tending strongly to zero, $\|AP_n^\perp\| \to 0$. Since $P_n^\perp A$ is the adjoint of $AP_n^\perp$, its norm tends to zero as well, so that

$$\|A - P_nAP_n\| = \|P_n^\perp A + P_nAP_n^\perp\| \leq \|P_n^\perp A\| + \|AP_n^\perp\| \longrightarrow 0, \qquad n \to \infty.$$

□

Lemma 2.1.13. *Let A be a self-adjoint (not necessarily bounded) operator on a separable Hillbert space $\mathcal{H}$ with domain $\mathcal{D}(A)$ and a quasidiagonalizing sequence $\{P_n\}$. Then $A = D + C$ where D is self-adjoint with domain $\mathcal{D}(D) = \mathcal{D}(A)$ and block diagonal with respect to some subsequence $\{P_{n_k}\}$. Also, C is compact and self-adjoint.*

Proof. To see this we can extend Halmos' proof in (Hal70) to unbounded operators. Now, by possibly passing to a subsequence, we may assume that $\sum_n \|P_n A - AP_n\| < \infty$. The fact that $P_n \geq P_{n-1}$ assures us that $P_n - P_{n-1}$ is a projection. Thus, we may decompose $\mathcal{H} = \bigoplus_{n=1}^{\infty}(P_{n+1} - P_n)\mathcal{H}$ and define D on

$$\mathcal{D}(D) = \mathrm{sp}\{\xi \in \mathcal{H} : \xi \in (P_{n+1} - P_n)\mathcal{H}\}$$

in the following way. If $\xi \in (P_{n+1} - P_n)\mathcal{H}$ then $D\xi = (P_{n+1} - P_n)A(P_{n+1} - P_n)\xi$. Now D is densely defined, with $\mathcal{D}(D) \subset \mathcal{D}(A)$, and obviously (by definition) block diagonal with respect to $\{P_n\}$. Define the operator C on $\mathcal{D}(C) = \mathcal{D}(D)$ by $C = A - D$. We will show that C is compact on $\mathcal{H}$. Indeed, by letting

$$C_n = P_{n+1}(AP_n - P_nA)P_n - P_n(AP_n - P_nA)P_{n+1}$$

we can form the operator $\tilde{C} = \sum_n C_n$ since $\|C_n\| \leq 2\|AP_n - P_nA\|$ and $\sum_n \|P_nA - AP_n\| < \infty$, hence the previous sum is norm convergent. Also, since C_n is finite dimensional and therefore compact it follows that $\tilde{C}$ is compact. A straightforward calculation shows that $\tilde{C} = C$ on $\mathcal{D}(C)$ which is dense, thus we can extend C to $\tilde{C}$ on $\mathcal{H}$. It is easy to see that C_n is self-adjoint since A is self-adjoint and hence C is self-adjoint. Let $\tilde{D} = A - C$. Then $\mathcal{D}(\tilde{D}) = \mathcal{D}(A)$ and $\tilde{D}$ is a self-adjoint extension of D. Also, since $\tilde{D}$ is an extension of D (which is block diagonal with respect to $\{P_n\}$) it follows that D is block diagonal with respect to $\{P_n\}$. □

Theorem 2.1.14. *Let A be a self-adjoint operator (not necessary bounded) on the separable Hilbert space $\mathcal{H}$ and let $\{P_n\}$ be a sequence of projections that quasi-diagonalizes A. If $K \subset \mathbb{R}$ is a compact set such that $\sigma(A) \cap K \neq \emptyset$, then*

$$\sigma(P_nA\lceil_{P_n\mathcal{H}}) \cap K \longrightarrow \sigma(A) \cap K, \quad n \to \infty$$

in the Hausdorff distance.

Proof. To prove the assertion we need to establish the following; given $\delta > 0$ then

$$\sigma(P_nA\lceil_{P_n\mathcal{H}}) \cap K \subset \omega_\delta(\sigma(A) \cap K)$$

and

$$\omega_\delta(\sigma(P_nA\lceil_{P_n\mathcal{H}}) \cap K) \supset \sigma(A) \cap K$$

for all sufficiently large n. The second inclusion follows by Theorem VIII.24 ((RS72), p. 290) if we can show that $P_nAP_n \to A$ in the strong resolvent sense. By Theorem VIII.25 ((RS72), p. 292) it suffices to show that $P_nAP_n\xi \to A\xi$ for $\xi \in \mathcal{D}(A)$, which is a common core for $\{P_nAP_n\}$ and A, and this is easily seen. To see the first inclusion note that it will follow if we can show that

$$\sigma(P_{n_k}A\lceil_{P_{n_k}\mathcal{H}}) \cap K \subset \omega_{\delta/2}(\sigma(A) \cap K) \tag{2.1.1}$$

when k is large, for some subsequence $\{P_{n_k}\}$. Indeed, if that is the case we only need to show that

$$\sigma(P_m A\lceil_{P_m\mathcal{H}}) \subset \omega_{\delta/2}(\sigma(P_{n_k} A\lceil_{P_{n_k}\mathcal{H}}))$$

for large m and n_k where $m \leq n_k$. Now this is indeed the case because we may assume, by appealing to Lemma 2.1.13 and possibly passing to another subsequence, that A is block diagonal with respect to $\{P_{n_k}\}$. Thus,

$$\|P_m P_{n_k} A P_{n_k} P_m - P_{n_k} A P_{n_k} P_m\| = \|P_m A P_m - A P_m\| \longrightarrow 0, \quad m \to \infty,$$

by assumption, and hence the desired inclusion follows by appealing to Lemma 2.1.11.

Now we return to the task of showing (2.1.1). Note that by the spectral mapping theorem, the spectra $\sigma(P_n A\lceil_{P_n\mathcal{H}})$ and $\sigma(A)$ are the images of $\sigma((P_n(A+i)\lceil_{P_n\mathcal{H}})^{-1})$ and $\sigma((A+i)^{-1})$, respectively, under the mapping $f(x) = 1/x - i$. Note that

$$f^{-1}(\sigma(P_n A\lceil_{P_n\mathcal{H}}) \cap K), \quad f^{-1}(\overline{\omega_\delta((\sigma(A) \cap K))})$$

are both compact and neither contain zero. Thus, by the continuity of f on $\mathbb{C} \setminus \{0\}$ and again the spectral mapping theorem, the assertion follows if we can prove that

$$\sigma((P_n(A+i)\lceil_{P_n\mathcal{H}})^{-1}) \subset \omega_\delta(\sigma((A+i)^{-1})) \tag{2.1.2}$$

for arbitrary $\delta > 0$ and large n. By Lemma 2.1.13 we have that $A = D + C$ where D is self-adjoint and block diagonal with respect to some subsequence $\{P_{n_k}\}$ and C is compact and self-adjoint. To simplify the notation we use the initial indexes for the subsequence. We first observe that

$$(D + P_n C P_n + i)^{-1} \to (D + C + i)^{-1} \tag{2.1.3}$$

in norm. Indeed, an easy manipulation gives us

$$\begin{aligned}\|(D+C+i)^{-1} - (D + P_n C P_n + i)^{-1}\| \\ \leq \|(D + C + i)^{-1}\| \|C - P_n C P_n\| \|(D + P_n C P_n + i)^{-1}\|,\end{aligned}$$

where $\|(D+P_nCP_n+i)^{-1}\|$ is bounded by the spectral mapping theorem since $C - P_n C P_n$ is self-adjoint. Since, by Lemma 2.1.12, $\|C - P_n C P_n\| \to 0$, (2.1.3) follows. The normality of $(D + C + i)^{-1}$ and $(D + P_n C P_n + i)^{-1}$ assures us that for any $\delta > 0$ we have

$$\sigma((D + P_n C P_n + i)^{-1}) \subset \omega_\delta(\sigma((D + C + i)^{-1}))$$

for sufficiently large n. Hence, to finish the proof we have to show that

$$\sigma((P_n(A+i)\lceil_{P_n\mathcal{H}})^{-1}) \subset \sigma((D + P_n C P_n + i)^{-1}).$$

In fact we have

$$\sigma((D + P_n C P_n + i)^{-1}) = \sigma((P_n(A+i)\lceil_{P_n\mathcal{H}})^{-1}) \cup \sigma(((D+i)\lceil_{P_n^\perp\mathcal{H}})^{-1}).$$

Indeed,

$$(D + P_n C P_n + i) = ((D + P_n C P_n + i)\lceil_{P_n\mathcal{H}}) \oplus (D+i)\lceil_{P_n^\perp\mathcal{H}}.$$

So

$$\begin{aligned}(D + P_n C P_n + i)^{-1} &= ((D + P_n C P_n + i)\lceil_{P_n\mathcal{H}})^{-1} \oplus ((D+i)\lceil_{P_n^\perp\mathcal{H}})^{-1} \\ &= (P_n(A+i)\lceil_{P_n\mathcal{H}})^{-1} \oplus ((D+i)\lceil_{P_n^\perp\mathcal{H}})^{-1},\end{aligned}$$

implying the assertion. □

As for the convergence of eigenvectors of the finite-section method, very little has been investigated, however we have the following:

Proposition 2.1.15. *Let $\{A_n\}$ be a sequence of self-adjoint bounded operators on $\mathcal{H}$ such that $A_n \to A$ strongly. Then if $\{\lambda_n\}$ is a sequence of eigenvalues of A_n such that $\lambda_n \to \lambda \in \sigma(A)$, and if $\{\xi_n\}$ is a sequence of unit eigenvectors corresponding to $\{\lambda_n\}$, such that $\{\xi_n\}$ does not converge weakly to zero, then there is a subsequence $\{\xi_{n_k}\}$ such that $\xi_{n_k} \xrightarrow{w} \xi$ where $A\xi = \lambda\xi$*

Proof. Since $\{\xi_n\}$ does not converge weakly to zero and by weak compactness of the unit ball in $\mathcal{H}$ we can find a weakly convergent subsequence such that $\xi_{n_k} \xrightarrow{w} \xi \neq 0$. To see that $A\xi = \lambda\xi$ we observe that this will follow if we can show that $\lambda_{n_k}\xi \xrightarrow{w} A\xi$. But the latter follows easily if we can show that $\lambda_{n_k}\xi_{n_k} - \lambda_{n_k}\xi \xrightarrow{w} 0$, $A_{n_k}\xi - A\xi \xrightarrow{w} 0$ and $A_{n_k}\xi - A_{n_k}\xi_{n_k} \xrightarrow{w} 0$. The first two are obvious and the last follows from the fact that for $\eta \in \mathcal{H}$ we have

$$\begin{aligned}\langle A_{n_k}(\xi - \xi_{n_k}), \eta\rangle &= \langle \xi - \xi_{n_k}, A_{n_k}\eta\rangle \\ &= \langle \xi - \xi_{n_k}, A\eta\rangle + \langle \xi - \xi_{n_k}, (A_{n_k} - A)\eta\rangle \longrightarrow 0,\end{aligned}$$

as $k \to \infty$. □

2.2 Divide and conquer

The divide-and-conquer technique has its origin in finite-dimensional matrix analysis. The idea was originally to divide the problem into smaller problems for simplicity reasons, a concept we will not discuss here. Since the crucial assumption for the procedure is that the operator acts on a finite dimensional space, we can not use it directly and we will not discuss its details here, but refer the reader to (Cup81). However, one can use the concept of the method to improve the results of Theorem 2.1.5 for tridiagonal infinite matrices. How to reduce the original spectral problem to a spectral problem for tridiagonal operators is discussed in section 2.4.

Definition 2.2.1. *Let $A \in \mathcal{B}(\mathcal{H})$ and $\{e_j\}$ be an orthonormal basis for $\mathcal{H}$. A is said to be tridiagonal with respect to $\{e_j\}$ if $\langle Ae_j, e_i\rangle = 0$ for $|i - j| \geq 2$.*

Let $A \in \mathcal{B}(\mathcal{H})$ be self-adjoint and $\{e_j\}$ be an orthonormal basis for $\mathcal{H}$. Suppose that A is tridiagonal with respect to $\{e_j\}$ and suppose that $a_{ij} = \langle Ae_j, e_i\rangle$ for $i, j = 1, 2, \ldots$ is real. It is easy to see that this is no restriction. Let P_n be the projection onto $\mathrm{sp}\{e_1, \ldots, e_n\}$. In the finite-section method one decomposes A into

$$A = P_nAP_n \oplus P_n^{\perp}AP_n^{\perp} + T, \qquad T \in \mathcal{B}(\mathcal{H}),$$

and then computes the spectrum of P_nAP_n. The idea of the divide-and-conquer approach is to decompose A into

$$A = A_{1,n} \oplus A_{2,n} + \beta\eta \otimes \eta, \qquad \eta \in \mathcal{H},$$

where $A_{1,n} \in \mathcal{B}(P_n\mathcal{H})$, $A_{2,n} \in \mathcal{B}(P_n^{\perp}\mathcal{H})$, $\eta = e_n + e_{n+1}$ and then compute $\sigma(A_{1,n})$. It is easy to see that the divide and conquer technique is very close to the finite-section method

i.e. we have $\langle P_n A P_n e_j, e_i\rangle = \langle A_{1,n} e_j, e_i\rangle$ for all i, j except for $i = j = n$. The goal is to improve the results in Theorem 2.1.5.

In finite dimensions one has the following theorem (Cup81) which gave us the idea to a more general theorem in infinite dimensions.

Theorem 2.2.2. *(Cuppen) Let D be a diagonal (real) matrix,*

$$D = diag(d_1, \ldots, d_n)$$

where $n \geq 2$ and $d_1 < d_2 < \ldots < d_n$. Let $\eta \in \mathbb{R}^n$ with $\eta_i \neq 0$ for $i = 1, \ldots, n$ and $\beta > 0$ be a scalar. Then the eigenvalues $\{\lambda_1, \ldots, \lambda_n\}$ of the matrix $D + \beta\eta \otimes \eta$ satisfy

$$d_1 < \lambda_1 < d_2 < \lambda_2 < \ldots < d_n < \lambda_n < d_n + \beta\|\eta\|^2.$$

Some of the techniques in the proof of the next theorem are inspired by the proof of Theorem 2.2.2 which can be found in (Cup81). Before we can state and prove the main theorem we need to introduce the concept of Householder reflections in an infinite-dimensional setting.

Definition 2.2.3. *A Householder reflection is an operator $S \in \mathcal{B}(\mathcal{H})$ of the form*

$$S = I - \frac{2}{\|\xi\|^2}\xi \otimes \bar{\xi}, \qquad \xi \in \mathcal{H}.$$

In the case where $\mathcal{H} = \mathcal{H}_1 \oplus \mathcal{H}_2$ and I_i is the identity on $\mathcal{H}_i$ then

$$U = I_1 \oplus \left(I_2 - \frac{2}{\|\xi\|^2}\xi \otimes \bar{\xi}\right) \qquad \xi \in \mathcal{H}_2.$$

will be called a Householder transformation.

A straightforward calculation shows that $S^* = S^{-1} = S$ and thus also $U^* = U^{-1} = U$. An important property of the operator S is that if $\{e_j\}$ is an orthonormal basis for $\mathcal{H}$ and $\eta \in \mathcal{H}$ then one can choose $\xi \in \mathcal{H}$ such that

$$\langle S\eta, e_j\rangle = \langle (I - \frac{2}{\|\xi\|^2}\xi \otimes \bar{\xi})\eta, e_j\rangle = 0, \qquad j \neq 1.$$

Indeed, if $\eta_1 = \langle \eta, e_1\rangle \neq 0$ one may choose $\xi = \eta \pm \|\eta\|\zeta$, where $\zeta = \eta_1/|\eta_1|e_1$ and if if $\eta_1 = 0$ choose $\xi = \eta \pm \|\eta\|e_1$, The verification of the assertion is a straightforward calculation.

Theorem 2.2.4. *Let $A_{1,n}$ be defined as above and let $\{d_j\}_{j=1}^k = \sigma(A_{1,n})$ be arranged such that $d_j < d_{j+1}$.*

(i) If $d_l, d_{l+1} \notin \sigma(A)$, for some $l < k$, then there is a $\lambda \in \sigma(A)$ such that $d_l < \lambda < d_{l+1}$.

(ii) If $d_j \in \sigma(A_{1,n})$ has multiplicity $m \geq 2$ then $d_j \in \sigma(A)$ and d_j is an eigenvalue. Also, $m_{A_{1,n}}(d_j) \leq m_A(d_j) + 1$, where $m_{A_{1,n}}(d_j)$ and $m_A(d_j)$ denote the multiplicity of d_j as an element of $\sigma(A_{1,n})$ and $\sigma(A)$ respectively.

Proof. We will start with (i). Suppose that $d_l, d_{l+1} \notin \sigma(A)$. We will show that $\sigma(A) \cap (d_l, d_{l+1}) \neq \emptyset$. We argue as follows. Let $\epsilon > 0$, $I_a = (-a, a]$ be an interval containing $\sigma(A_{2,n})$ and let g be a step function on I_a of the form $g = \sum_{j=1}^{m} \chi_{(a_j,b_j]}$ such that $\sup_{x\in I_a} |x - g(x)| < \epsilon$. Let $\tilde{A}_{2,n} = g(A_{2,n})$. Then $\sigma(\tilde{A}_{2,n})$ contains only isolated eigenvalues and $\|\tilde{A}_{2,n} - A_{2,n}\| < \epsilon$. Also, let

$$\tilde{A} = A_{1,n} \oplus \tilde{A}_{2,n} + \beta\eta \otimes \eta.$$

Then $\tilde{A}$ is self-adjoint and $\|\tilde{A} - A\| < \epsilon$ so

$$d_H(\sigma(\tilde{A}), \sigma(A)) < \epsilon$$

where d_H denotes the Hausdorff metric. Also, by choosing ϵ small enough we have $d_l, d_{l+1} \notin \sigma(\tilde{A})$. Note that, since ϵ is arbitrary and $\sigma(A)$ is closed, the assertion that $\sigma(A) \cap (d_l, d_{l+1}) \neq \emptyset$ will follow if we can show that $\sigma(\tilde{A}) \cap (d_l, d_{l+1}) \neq \emptyset$.

Let P_n be the projection onto $\mathrm{sp}\{e_j\}_{j=1}^{n}$. Now, choose a unitary operator Q_1 on $P_n\mathcal{H}$ such that $Q_1 A_{1,n} Q^* = D_1$ where D_1 is diagonal with respect to $\{e_j\}_{j=1}^{n}$. Since $\sigma(\tilde{A}_{2,n})$ contains only finitely many eigenvalues we may choose a unitary Q_2 on $\mathrm{ran}P_n^{\perp}$ such that $Q_2 \tilde{A}_{2,n} Q_2^* = D_2$ is diagonal with respect to $\{e_j\}_{j=n+1}^{\infty}$. Thus,

$$(Q_1 \oplus Q_2)(A_{1,n} \oplus \tilde{A}_{2,n} + \beta\eta \otimes \eta)(Q_1^* \oplus Q_2^*) = D_1 \oplus D_2 + \beta\xi \otimes \bar{\xi},$$

where a straightforward calculation shows that $\xi = Q_1 e_n \oplus Q_2 e_{n+1}$. Let $D = D_1 \oplus D_2$.

Claim1: *There exists a unitary operator U and an integer N such that*

$$\langle U\xi, e_i \rangle = 0$$

for $i \geq N+1$ and $\langle U\xi, e_i \rangle \neq 0$ for $i \leq N$, and also that UDU^* is diagonal with respect to $\{e_j\}$. Note that the claim will follow if we can show that there is a unitary operator V such that $\langle V\xi, e_j \rangle \neq 0$ only for finitely many js and that $VDV^* = D$. Indeed, if we have such a V then we can find a unitary operator $\tilde{V}$ that permutes $\{e_j\}$ such that $U = \tilde{V}V$ is the desired unitary operator mentioned above.

To construct V we first note that, since D is diagonal with respect to $\{e_j\}$, the spectral projections $\chi_\lambda(D)$, $\lambda \in \sigma(D)$ are also diagonal with respect to $\{e_j\}$. Note that

$$D = \bigoplus_{\lambda\in\sigma(D)} \lambda\chi_\lambda(D).$$

We will use this decomposition to construct V. Let

$$i_\lambda = \inf\{j : \chi_\lambda(D)e_j \neq 0\}.$$

If $\chi_\lambda(D)\xi = 0$ let $V_\lambda = I$ on $\chi_\lambda(D)\mathcal{H}$. If not, choose a Householder reflection on $\chi_\lambda(D)\mathcal{H}$,

$$S = I - \frac{2}{\|\zeta\|^2}\zeta \otimes \bar{\zeta}, \qquad \zeta \in \chi_\lambda(D)\mathcal{H},$$

such that

$$\langle S\chi_\lambda(D)\xi, e_{i_\lambda} \rangle \neq 0 \quad \text{and} \quad \langle S\chi_\lambda(D)\xi, e_i \rangle = 0, \quad i \geq i_\lambda + 1. \tag{2.2.1}$$

Let $V_\lambda = S$. The fact that $\chi_\lambda(D)$ for $\lambda \in \sigma(D)$ is diagonal with respect to $\{e_j\}$ gives $V_\lambda \chi_\lambda(D) V_\lambda^* = \chi_\lambda(D)$. Letting

$$V = \bigoplus_{\lambda \in \sigma(D)} V_\lambda \tag{2.2.2}$$

we get $VDV^* = D$ and thus we have constructed the desired unitary operator V whose existence we asserted. As argued above, this yields existence of the unitary operator U asserted in Claim1. Let $N = \max\{j : \langle U\xi, e_j\rangle \neq 0\}$, let P_N be the projection onto $\mathrm{sp}\{e_j\}_{j=1}^N$ and $\tilde{D} = UDU^*$.

Claim2: *If $\lambda \in \sigma(P_N \tilde{D}\lceil_{P_N\mathcal{H}})$ then λ has multiplicity one.* We argue by contradiction. Suppose that $\lambda \in \sigma(P_N \tilde{D}\lceil_{P_N\mathcal{H}})$ has multiplicity greater than one. Then $\langle \tilde{D}e_p, e_p\rangle = \langle \tilde{D}e_q, e_q\rangle = \lambda$ for some $p, q \leq N$. Also, $\langle U\xi, e_p\rangle \neq 0$ and $\langle U\xi, e_q\rangle \neq 0$. Thus, it follows from the construction of U that $\langle De_{\tilde{p}}, e_{\tilde{p}}\rangle = \langle De_{\tilde{q}}, e_{\tilde{q}}\rangle = \lambda$ for some integers $\tilde{p}$ and $\tilde{q}$, and hence $e_{\tilde{p}}, e_{\tilde{q}} \in \mathrm{ran}\chi_\lambda(D)$. Also $\langle V\xi, e_{\tilde{p}}\rangle \neq 0$ and $\langle V\xi, e_{\tilde{q}}\rangle \neq 0$ and thus it follows that

$$\langle V_\lambda \chi_\lambda(D)\xi, e_j\rangle = \langle \bigoplus_{\lambda \in \sigma(D)} V_\lambda \xi, e_j\rangle \neq 0, \qquad j = \tilde{p}, \tilde{q},$$

and this contradicts (2.2.1). Armed with the results from Claim1 and Claim2 we can now continue with the proof.

Let $\zeta = U\xi$. We then have

$$U(D + \beta\xi \otimes \bar{\xi})U^* = (P_N \tilde{D} P_N + \beta P_N \zeta \otimes \overline{P_N\zeta})\lceil_{P_N\mathcal{H}} \oplus P_N^\perp \tilde{D}\lceil_{P_N^\perp\mathcal{H}},$$

since $P_N^\perp(\zeta \otimes \bar{\zeta}) = (\zeta \otimes \bar{\zeta})P_N^\perp = 0$. So, with a slight abuse of notation we will denote $P_N\zeta$ just by ζ. Note that

$$\sigma(\tilde{A}) = \sigma((P_N \tilde{D} P_N + \beta\zeta \otimes \bar{\zeta})\lceil_{P_N\mathcal{H}}) \cup \sigma(P_N^\perp \tilde{D}\lceil_{P_N^\perp\mathcal{H}}) \tag{2.2.3}$$

and hence our primary goal to prove that $\sigma(\tilde{A}) \cap (d_l, d_{l+1}) \neq \emptyset$ has been reduced to showing that

$$\sigma((P_N \tilde{D} P_N + \beta\zeta \otimes \bar{\zeta})\lceil_{P_N^\perp\mathcal{H}}) \cap (d_l, d_{l+1}) \neq \emptyset. \tag{2.2.4}$$

Before continuing with that task note that

$$d_l, d_{l+1} \in \sigma(P_N \tilde{D}\lceil_{P_N\mathcal{H}}). \tag{2.2.5}$$

Indeed, it is true, by the construction of $\tilde{D}$, that $d_l, d_{l+1} \in \sigma(\tilde{D})$. But by (2.2.3) it follows that $\sigma(P_N^\perp \tilde{D} P_N^\perp) \subset \sigma(\tilde{A})$ and since $d_l, d_{l+1} \notin \sigma(\tilde{A})$ the assertion follows. This observation will be useful later in the proof.

Now returning to the task of showing (2.2.4), let $\hat{D} = P_N \tilde{D}\lceil_{P_N\mathcal{H}}$ and then let λ be an eigenvalue of $\hat{D} + \beta\zeta \otimes \bar{\zeta}$ with corresponding nonzero eigenvector η. Here $\zeta \otimes \bar{\zeta}$ denotes, with a slight abuse of notation, the operator $(\zeta \otimes \bar{\zeta})\lceil_{P_N\mathcal{H}}$. Then we have

$$(\hat{D} + \beta\zeta \otimes \bar{\zeta})\eta = \lambda\eta \quad \text{so} \quad (\hat{D} - \lambda I)\eta = -\beta\langle\eta, \zeta\rangle\zeta. \tag{2.2.6}$$

Note that $\hat{D} - \lambda I$ is nonsingular. Indeed, had it been singular, we would have had $\lambda = \hat{d}_i$ for some $i \leq N$, where $\{\hat{d}_j\}_{j=1}^N = \sigma(\hat{D})$. Hence, by (2.2.6), we have

$$\langle(\hat{D} - \lambda I)\eta, e_i\rangle = -\beta\langle\eta, \zeta\rangle\langle\zeta, e_i\rangle = 0.$$

But, since $\zeta = U\xi$ and by Claim1, it is true that $\langle \zeta, e_i \rangle \neq 0$, so $\langle \eta, \zeta \rangle = 0$. Thus, by (2.2.6), it follows that $(\hat{D} - \lambda I)\eta = 0$, so $\langle (\hat{D} - \lambda)\eta, e_j \rangle = 0$ for $j \leq N$. Note that, by Claim2, $\sigma(\hat{D})$ contains only eigenvalues with multiplicity one, thus we have $\lambda = \hat{d}_i$ only for one such i. Thus, $\langle \eta, e_j \rangle = 0$ for $j \neq i$, so

$$\langle \eta, \zeta \rangle = \langle \zeta, e_i \rangle \langle \eta, e_i \rangle = 0.$$

But we have assumed that $\eta \neq 0$ so $\langle \eta, e_i \rangle \neq 0$ and therefore $\langle \zeta, e_i \rangle = 0$, a contradiction. We therefore deduce that $\hat{D} - \lambda I$ is nonsingular and $\langle \eta, \zeta \rangle \neq 0$. Thus, by (2.2.6), it follows that

$$\eta = -\beta \langle \eta, \zeta \rangle (\hat{D} - \lambda I)^{-1} \zeta$$

and

$$\langle \eta, \zeta \rangle (1 + \beta \langle (D - \lambda I)^{-1} \zeta, \zeta \rangle) = \langle \eta, \zeta \rangle f(\lambda) = 0,$$

where

$$f(\lambda) = 1 + \beta \sum_{j=1}^{N} \frac{|\zeta_j|^2}{\hat{d}_j - \lambda}, \qquad \zeta_j = \langle \zeta, e_j \rangle.$$

Since $\langle \eta, \zeta \rangle \neq 0$ it follows that $f(\lambda) = 0$. Note that, by (2.2.5), it is true that $d_l, d_{l+1} \in \{\hat{d}_j\}_{j=1}^{N}$ and so by the properties of f it follows that there is at least one

$$\lambda \in \sigma(\hat{D} + \beta \zeta \otimes \bar{\zeta})$$

such that $d_l < \lambda < d_{l+1}$, proving (2.2.4).

To show (ii) we need to prove that if $\sigma(A_{1,n})$ has an eigenvalue d with multiplicity $m > 1$ then $d \in \sigma(A)$ and $m_{A_{1,n}}(d) \leq m_A(d) + 1$. To prove that we proceed as in the proof of (i). Let P_n be the projection onto $\mathrm{sp}\{e_j\}_{j=1}^{n}$. Now, choose a unitary operator Q_1 on $P_n\mathcal{H}$ such that $Q_1 A_{1,n} Q_1^* = D_1$ where D_1 is diagonal with respect to $\{e_j\}_{j=1}^{n}$ so that

$$\begin{aligned}(Q_1 \oplus I_2)(A_{1,n} \oplus A_{2,n} + \beta \eta \otimes \eta)(Q_1^* \oplus I_2) \\ = D_1 \oplus A_{2,n} + \beta(\zeta \oplus e_{n+1}) \otimes (\bar{\zeta} \oplus e_{n+1}),\end{aligned}$$

where I_2 is the identity on $P_n^{\perp}\mathcal{H}$ and $\zeta = Q_1 e_n$. For any set S let $\#S$ denote the number of elements in S. Note that the assertion will follow if we can show that there is a unitary operator V on $P_n\mathcal{H}$, such that $V D_1 V^* = D_1$, and that

$$\#\{e_j : \langle \chi_d(D_1) V \zeta, e_j \rangle \neq 0,\ 1 \leq j \leq n\} \leq 1. \tag{2.2.7}$$

Indeed, if so is true, we have that

$$D_1 \oplus A_{2,n} + \beta(\zeta \oplus e_{n+1}) \otimes (\bar{\zeta} \oplus e_{n+1})$$

is unitarily equivalent to

$$B = D_1 \oplus A_{2,n} + \beta(V\zeta \oplus e_{n+1}) \otimes (\overline{V\zeta} \oplus e_{n+1}),$$

and $\Lambda = \{e_j : \langle V\zeta, e_j \rangle = 0\}$ are all eigenvectors of B. Also, the eigenvalue corresponding to the set

$$\tilde{\Lambda} = \{e_j \in \Lambda : \chi_d(D_1) e_j \neq 0\}$$

is d. Thus, by (2.2.7), we get the following estimate

$$\begin{aligned} m_A(d) &\geq \#\tilde{\Lambda} \\ &\geq \dim(\mathrm{ran}\chi_d(D_1)) - \#\{e_j : \langle \chi_d(D_1)V\zeta, e_j\rangle \neq 0,\ 1 \leq j \leq n\} \\ &\geq m_{A_{1,n}}(d) - 1, \end{aligned}$$

and this proves the assertion. The existence of V follows by exactly the same construction as done in the proof of Claim1 in the proof (i) by using Householder reflections. □

Note that the following theorem is similar to Theorem 2.3 and Theorem 3.8 in (Arv94a) and the proof requires similar techniques. Since the divide-and-conquer method is different form the finite-section method, we cannot use the theorems in (Arv94a) directly. However, one should note that the following theorem gives much stronger estimates on the behavior of the false eigenvalues that may occur.

Theorem 2.2.5. *Let $\{A_{1,n}\}$ be the sequence obtained from A as in Theorem 2.2.4 (recall also definitions 2.1.1 and 2.1.2).*

(i) $\sigma(A) \subset \Lambda$.

(ii) Let $a \in \sigma_e(A)^c$. Then a is transient.

(iii) If $U \subset \mathbb{R}$ is an open interval such that $U \cap \sigma(A) = \emptyset$ then $N_n(U) \leq 1$. If $U \cap \sigma(A)$ contains only one point then $\widetilde{N}_n(U) \leq 3$.

(iv) Let λ be an isolated eigenvalue of A with multiplicity m. If $U \subset \mathbb{R}$ is an open interval containing λ such that $U \setminus \{\lambda\} \cap \sigma(A) = \emptyset$ then $\widetilde{N}_n(U) \leq m + 3$.

(v) $\sigma_e(A) = \Lambda_e$,

(vi) Every point of $\mathbb{R}$ is either transient or essential.

Proof. Now, (i) follows from the fact that $A_{1,n} \to A$ strongly (see Theorem VIII.24 in (RS72), p. 290), which is easy to see. Also, (iii) follows immediately by Theorem 2.2.4 and (ii) follows by (iii) and (iv). Indeed, assuming (iv) we only have to show that if $a \in \sigma(A)^c$ then a is transient and this follows from (iii). Hence, we only have to prove (iv). Let λ be an isolated eigenvalue of A with multiplicity m. If $U \subset \mathbb{R}$ is an open interval containing λ such that $U \setminus \{\lambda\} \cap \sigma(A) = \emptyset$ then, by (iii), we have $\widetilde{N}_n(U) \leq 3$. But, by Theorem 2.2.4, we can have $\widetilde{N}_n(U) \leq 3$ and $N_n(U) > 3$ only if $\lambda \in \sigma(A_{1,n})$. Also, by Theorem 2.2.4, $m_{A_{1,n}}(\lambda) \leq m + 1$, and this yields the assertion.

To get (v) and (vi) we only have to show that $\sigma_e(A) \subset \Lambda_e$. Indeed, by (ii), we have $\sigma_e(A)^c \subset \Lambda_e^c$, so if $\sigma_e(A) \subset \Lambda_e$ then (v) follows. But then $\mathbb{R} \setminus \Lambda_e = \mathbb{R} \setminus \sigma(A)_e$ and the left hand side of the equality is, by (ii), contained in the set of transient points, thus we obtain (vi).

To show that $\sigma_e(A) \subset \Lambda_e$ we will show that $\Lambda_e^c \subset \sigma_e(A)^c$. Let $\lambda \in \Lambda_e^c$. We will show that $\lambda \in \sigma_e(A)^c$. Note that, by the definition of the essential spectrum, this follows if we can show that there is an operator $T \in \mathcal{B}(\mathcal{H})$ such that $T(A - \lambda I) = (A - \lambda I)T = I + C$, where C is compact.

Since $\lambda \in \Lambda_e^c$ there is a subsequence $\{n_k\} \subset \mathbb{N}$, an $\epsilon > 0$, and an integer K such that for $\Omega = (\lambda - \epsilon, \lambda + \epsilon)$ then $N_{n_k}(\Omega) \leq K$. Let P_k be the projection onto $sp\{e_j\}_{j=1}^{n_k}$ and

$E_k = \chi_\Omega(A_{1,n_k})$. Then A_{1,n_k}, P_k and E_k all commute, so we can let $B_k = (A_{1,n_k} - \lambda I)|_{\mathcal{H}_k}$ where $\mathcal{H}_k = \mathrm{ran}(P_k E_k^\perp)$. Note that B_k must be invertible with $\|B_k^{-1}\| \leq \epsilon^{-1}$. Since $P_k E_k^\perp = P_k - E_k$, we deduce that

$$(A_{1,n_k} - \lambda I)B_k^{-1}(P_k - E_k) = B_k^{-1}(P_k - E_k)(A_{1,n_k} - \lambda I) = P_k - E_k. \tag{2.2.8}$$

Since $\{B_k^{-1}\}$ is bounded and norm closed, while bounded sets of $\mathcal{B}(\mathcal{H})$ are weakly sequentially compact, we may assume, by possibly passing to a new subsequence that

$$\mathrm{WOT}\lim_{k\to\infty} B_k^{-1}(P_k - E_k) = T \in \mathcal{B}(\mathcal{H}), \quad \mathrm{WOT}\lim_{k\to\infty} E_k = C \in \mathcal{B}(\mathcal{H}).$$

The fact that $A_{1,n} \longrightarrow A$ strongly together with the uniform boundedness of $B_k^{-1}(P_k - E_k)$ allow us to take weak limits in (2.2.8) and we get $T(A - \lambda I) = (A - \lambda I)T = I + C$.

Note that C is compact, in fact it is trace class. For $\dim E_k \leq K$ so $\mathrm{trace}(E_k) \leq K$ and $\{H \in \mathcal{B}(\mathcal{H}) : \mathrm{trace}(H) \leq K\}$ is weakly closed. □

Corollary 2.2.6. *Let $\lambda \in \sigma_e(A)$ be an isolated eigenvalue. Then $\lambda \in \sigma(A_{1,n})$ for all sufficiently large n. Moreover, $m_n(\lambda) \longrightarrow \infty$, where $m_n(\lambda)$ is the multiplicity of λ as an element of $\sigma(A_{1,n})$.*

Proof. Since, by Theorem 2.2.5, $\sigma_e(A) = \Lambda_e$, for any open neighborhood U around λ we have $N_n(U) \to \infty$. Let U be an open interval containing λ such that $(U \setminus \{\lambda\}) \cap \sigma(A) = \emptyset$. Then, by Theorem 2.2.4, $U \cap \sigma(A_{1,n})$ cannot contain more that three distinct points, and since $N_n(U) \to \infty$ it follows that $A_{1,n}$ must have eigenvalues in U with multiplicity larger than two. Using Theorem 2.2.4 again it follows that $\lambda \in \sigma(A_{1,n})$ for all sufficiently large n. The last assertion of the corollary follows by similar reasoning. □

2.3 Detecting false eigenvalues

Let $A \in \mathcal{B}(\mathcal{H})$ be self-adjoint. The fact that both the finite-section method and the divide and conquer method may produce points that are not in the spectrum of A poses the question; can one detect false eigenvalues? The phenomenon of false eigenvalues is well known and is often referred to as spectral pollution.

Let $\lambda \in \mathbb{R}$. The easiest way to determine whether $\lambda \in \sigma(A)$ is to estimate

$$\mathrm{dist}(\lambda, \sigma(A)) = \inf_{\xi \in \mathcal{H}, \|\xi\|=1} \langle (A - \lambda)^2 \xi, \xi \rangle.$$

Let $\{P_n\}$ be an increasing sequence of finite-dimensional projections converging strongly to the identity. Let $\gamma(\lambda) = \mathrm{dist}(\lambda, \sigma(A))$ and

$$\gamma_n(\lambda) = \inf_{\xi \in P_n\mathcal{H}, \|\xi\|=1} \langle (A - \lambda)^2 \xi, \xi \rangle.$$

It is easy to show that γ and γ_n are Lipchitz continuous with Lipchitz constant bounded by one. This implies that $\gamma_n \to \gamma$ locally uniformly and hence one can use $\gamma_n(\lambda)$ as an approximation to $\mathrm{dist}(\lambda, \sigma(A))$. Obtaining $\gamma_n(\lambda)$ is done by finding the smallest eigenvalue of a self-adjoint (finite rank) matrix. In fact γ_n can be used alone to estimate $\sigma(A)$ and that has been investigated in (DP04). However, it seems that a combination of the finite-section method or the divide-and-conquer method, accompanied by estimates as in the previous sections and in (Arv94a), with some computed values of γ_n will give more efficient computational algorithms, especially for detecting isolated eigenvalues.

2.4 Tridiagonalization

In the previous section the crucial assumption was that the operator was tridiagonal with respect to some basis. We will in this section show how we can reduce the general problem to a tridiagonal one. In the finite-dimensional case every self-adjoint matrix is tridiagonalizable. This is not the case in infinite dimensions, however, it is well known that if a self-adjoint operator $A \in \mathcal{B}(\mathcal{H})$ has a cyclic vector ξ then A is tridiagonal with respect to the basis $\{e_j\}$ constructed by using the Gram-Schmidt procedure to $\{A^n\xi\}_{n=0}^{\infty}$. The problem is that our operator may not have a cyclic vector, however the following lemma is well known.

Lemma 2.4.1. *Let $A \in \mathcal{B}(\mathcal{H})$ and let $\mathcal{A}$ be the complex algebra generated by A, A^* and the identity. Then there is a (finite or infinite) sequence of nonzero $\mathcal{A}$-invariant subspaces $\mathcal{H}_1, \mathcal{H}_2 \ldots$ such that:*

(i) $\mathcal{H} = \mathcal{H}_1 \oplus \mathcal{H}_2 \oplus \cdots$

(ii) Each $\mathcal{H}_n$ contains a cyclic vector ξ_n for $\mathcal{A}$: $\mathcal{H}_n = \overline{\mathcal{A}\xi_n}$, $n = 1, 2, \ldots$.

Thus, if we knew the decomposition above we could decompose our operator A into $A = H_1 \oplus H_2 \oplus \cdots$ where H_n would have a cyclic vector and hence be tridiagonalizable. Also, we would have $\sigma(A) = \overline{\bigcup_j \sigma(H_j)}$. The problem is: how do we compute H_n? This is what we will discuss in this section.

Definition 2.4.2. *Let $A \in \mathcal{B}(\mathcal{H})$ and let $\{e_j\}$ be an orthonormal basis for $\mathcal{H}$. A is said to be Hessenberg with respect to $\{e_j\}$ if $\langle Ae_j, e_i\rangle = 0$ for $i \geq j + 2$.*

Theorem 2.4.3. *Let A be a bounded operator on a separable Hilbert space $\mathcal{H}$ and let $\{e_j\}$ be an orthonormal basis for $\mathcal{H}$. Then there exists an isometry V such that $V^*AV = H$ where H is Hessenberg with respect to $\{e_j\}$. Moreover $V = \underset{n\to\infty}{SOT\text{-}lim}\, V_n$ where $V_n = U_1 \cdots U_n$ and U_j is a Householder transformation. Also, the projection $P = VV^*$ satisfies $PAP = AP$.*

Proof. We will obtain H as the strong limit of a sequence $\{V_n^*AV_n\}$ where $V_n = U_1 \cdots U_n$ is a unitary operator and U_j is a Householder transformation. The procedure is as follows: Let P_n be the projection onto $\mathrm{sp}\{e_1, \ldots, e_n\}$. Suppose that we have the n elements in the sequence and that the n-th element is an operator $H_n = V_n^*AV_n$ that with respect to $\mathcal{H} = P_n\mathcal{H} \oplus P_n^{\perp}\mathcal{H}$ has the form

$$H_n = \begin{pmatrix} \tilde{H}_n & B_n \\ C_n & N_n \end{pmatrix}, \quad \tilde{H}_n = P_nH_nP_n,\ B_n = P_nH_nP_n^{\perp},\ C_n = P_n^{\perp}H_nP_n,$$

where $N_n = P_n^{\perp}H_nP_n^{\perp}$, $\tilde{H}_n$ is Hessenberg and $C_ne_j = 0$ for $j < n$. Let $\zeta = C_ne_n$. Choose $\xi \in P_n^{\perp}\mathcal{H}$ such that the Householder reflection $S \in \mathcal{B}(P_n^{\perp}\mathcal{H})$ defined by

$$S = I - \frac{2}{\|\xi\|^2}\xi \otimes \bar{\xi}, \quad \text{and} \quad U_n = P_n \oplus S, \tag{2.4.1}$$

gives $S\zeta = \{\tilde{\zeta}_1, 0, 0, \ldots\}$, and let $H_{n+1} = U_nH_nU_n$. Hence,

$$H_{n+1} = U_nH_nU_n = \begin{pmatrix} \tilde{H}_n & B_nS \\ SC_n & SN_nS \end{pmatrix} = \begin{pmatrix} \tilde{H}_{n+1} & B_{n+1} \\ C_{n+1} & N_{n+1} \end{pmatrix}, \tag{2.4.2}$$

where the last matrix is understood to be with respect to the decomposition $\mathcal{H} = P_{n+1}\mathcal{H} \oplus P_{n+1}^{\perp}\mathcal{H}$. Note that, by the choice of S, it is true that $\tilde{H}_{n+1}$ is Hessenberg and $C_{n+1}e_j = 0$ for $j < n+1$. Defining $H_1 = A$ and letting $V_n = U_1 \cdots U_n$ we have completed the construction of the sequence $\{V_n^* A V_n\}$.

Note that $H_n = V_n^* A V_n$ is bounded, since V_n is unitary (since U_j is unitary). And since a closed ball in $\mathcal{B}(\mathcal{H})$ is weakly sequentially compact, there is an $H \in \mathcal{B}(\mathcal{H})$ and a subsequence $\{H_{n_k}\}$ such that $H_{n_k} \xrightarrow{\text{WOT}} H$. But by (2.4.2) it is clear that for any j we have $H_n e_j = H_m e_j$ for sufficiently large m and n. It follows that SOT-$\lim_n H_n = H$. Also, by(2.4.2) H is Hessenberg. By similar reasoning, using the previous compactness argument (since V_n is bounded) and the fact that, by (2.4.1), $V_n e_j = V_m e_j$ for any j and m and n sufficiently large, we deduce that there exists a $V \in \mathcal{B}(\mathcal{H})$ such that

$$\underset{n\to\infty}{\text{SOT-lim}}\ V_n = V, \quad \underset{n\to\infty}{\text{WOT-lim}}\ V_n^* = V^*.$$

Since V is the strong limit of a sequence of unitary operators, it follows that V is an isometry. We claim that $V^*AV = H$. Indeed, since multiplication is jointly continuous in the strong operator topology on bounded sets and $H_n = V_n^* A V_n$ so $V_n H_n = A V_n$ we have $AV = VH$. Since V is an isometry the assertion follows. Note that $PAP = AP$ also follows since $PAP = VV^*AVV^* = VHV^* = AP$. □

Corollary 2.4.4. *Suppose that the assumptions in Theorem 2.4.3 are true, and suppose also that A is self-adjoint. Then there exists an isometry V such that $V^*AV = H$ where H is tridiagonal with respect to $\{e_j\}$. Moreover $V = \underset{n\to\infty}{\text{SOT-lim}}\ V_n$ where $V_n = U_1 \cdots U_n$ and U_j is a Householder transformation. Also, the projection $P = VV^*$ satisfies $PA = AP$.*

Proof. Follows immediately from the previous theorem. □

In the case where A is self-adjoint, by the previous corollary we have that $PA = AP$, where $P = VV^*$. Now, the "part" of A, namely $P^{\perp}A$, that we do not capture with the construction in the proof of Theorem 2.4.3 can be computed by the already constructed operators i.e. we have

$$P^{\perp}A = A - VHV^*.$$

Thus, we may apply Theorem 2.4.3 again to $P^{\perp}A$. And, of course this can be applied recursively. In other words; consider $V_1^* A V_1 = H_1$, where H_1 is tridiagonal w.r.t $\{e_j\}$. Let $P_1 = V_1 V_1^*$. Then $P_1 A = A P_1$ and $P_1^{\perp} A = A - V_1^* H_1 V_1$. Let $H_2 = V_2^* P_1^{\perp} A V_2$. In general we have

$$H_{n+1} = V_{n+1}^*(A - V_1 H_1 V_1^* - \cdots - V_n H_n V_n^*)V_{n+1}.$$

Using the previous construction we can actually recover the whole spectrum of A. More precisely we have the following:

Theorem 2.4.5. *Let A be self-adjoint and let*

$$H_{n+1} = V_{n+1}^*(A - V_1 H_1 V_1^* - \cdots - V_n H_n V_n^*)V_{n+1}$$

be defined as above. Then

$$\sigma(A) = \overline{\bigcup_{n\in\mathbb{N}} \sigma(H_n)}.$$

Proposition 2.4.6. *Let $\{P_j\}$ be a sequence of projections described above i.e. $P_j = V_j^* V_j$. Then $\mathrm{sp}\{e_1, \dots, e_n\} \subset \mathrm{ran}(P_m)$ for $m \geq n$.*

Proof. The proof is an easy induction using the fact that $e_1 \in \mathrm{ran}(P_1)$, which follows by the construction of V_1. □

Proof. Proof of Theorem 2.4.5 Let $P_j = V_j^* V_j$ and recall that by the construction of H_n we have

$$H_n = V_n^* P_{n-1}^\perp \cdots P_1^\perp A V_n, \tag{2.4.3}$$

where we have defined recursively

$$P_{n-1}^\perp \cdots P_1^\perp A = A - V_1 H_1 V_1^* - \cdots - V_{n-1} H_{n-1} V_{n-1}^*,$$

and by Corollary 2.4.4 it follows that

$$P_n P_{n-1}^\perp \cdots P_1^\perp A = P_{n-1}^\perp \cdots P_1^\perp A P_n. \tag{2.4.4}$$

Note that $\sigma(H_n) = \sigma(P_{n-1}^\perp \cdots P_1^\perp A\lceil_{P_n\mathcal{H}})$. Indeed, by Corollary 2.4.4, V_n is an isometry onto $P_n\mathcal{H}$, thus $\{V_n e_j\}$ is a basis for $P_n\mathcal{H}$, so for

$$\tilde{A} = (P_{n-1}^\perp \cdots P_1^\perp A)\lceil_{P_n\mathcal{H}}$$

it follows, by (2.4.3), that

$$\langle \tilde{A} V_n e_j, V_n e_i \rangle = \langle P_{n-1}^\perp \cdots P_1^\perp A V_n e_j, V_n e_i \rangle = \langle H_n e_j, e_i \rangle,$$

yielding that $\sigma(H_n) = \sigma(P_{n-1}^\perp \cdots P_1^\perp A\lceil_{P_n\mathcal{H}})$. Let us define the projection

$$E_n = P_n \wedge P_{n-1}^\perp \wedge \cdots \wedge P_1^\perp, \quad E_1 = P_1,$$

and note that $E_j \perp E_i$ for $i \neq j$. Now the theorem will follow if we can show that $AE_n = E_n A$,

$$A = \bigoplus_{n\in\mathbb{N}} E_n A$$

and

$$P_n P_{n-1}^\perp \cdots P_1^\perp A = E_n A.$$

We will start with the former assertion (this is immediate for $n = 1$ by Corollary 2.4.4). Indeed, if $\xi \in \mathrm{ran}(E_n)$ for $n \geq 2$ then, by Corollary 2.4.4,

$$\begin{aligned} A\xi &= A P_1^\perp \cdots P_{n-1}^\perp P_n \xi = P_n P_{n-1}^\perp \cdots P_1^\perp A\xi = P_{n-1}^\perp \cdots P_1^\perp A P_n \xi \\ &= P_{n-2}^\perp \cdots P_1^\perp A P_{n-1}^\perp P_n \xi = \cdots \text{etc.} \end{aligned} \tag{2.4.5}$$

Thus, it follows that $A\,\mathrm{ran}(E_n) \subset \mathrm{ran}(E_n)$. Since A is self-adjoint we have that $AE_n = E_n A$. We can now show that $A = E_1 A \oplus E_2 A \oplus \cdots$. First, an easy induction demonstrates that for any $n \in \mathbb{N}$ we have

$$A = E_1 A \oplus \cdots \oplus E_n A \oplus P_n^\perp \cdots P_1^\perp A.$$

Note that, by Proposition 2.4.6 and (2.4.4), it follows that $P_n^\perp \cdots P_1^\perp A e_j = 0$ for $j \leq n$ thus $Ae_n = (E_1 A \oplus \cdots \oplus E_n A)e_n$. Also, $E_{n+1} A e_j = 0$ for $j \leq n$. This gives us that if $T = E_1 A \oplus E_2 \oplus \cdots$. Then

$$Te_n = E_1 A \oplus \cdots \oplus E_n A e_n = Ae_n$$

yielding the assertion.

Finally, we will show that $P_n P_{n-1}^\perp \cdots P_1^\perp A = E_n A$. Note that in (2.4.5) we have also shown that $P_n P_{n-1}^\perp \cdots P_1^\perp A\xi = A\xi$ when $\xi \in \mathrm{ran}(E_n)$. So, to show that $P_n P_{n-1}^\perp \cdots P_1^\perp A = E_n A$, we only have to show that $P_n P_{n-1}^\perp \cdots P_1^\perp A\eta = 0$ when $\eta \in \mathrm{ran}(E_n^\perp)$. But, by the definition of E_n we have $\eta \in \bigcup_{j=1}^{n-1} P_j\mathcal{H} \cup P_n^\perp \mathcal{H}$ and an easy application of Corollary 2.4.4 gives

$$P_n P_{n-1}^\perp \cdots P_1^\perp A = P_n P_{n-2}^\perp \cdots P_1^\perp A P_{n-1}^\perp = P_n P_{n-1}^\perp P_{n-3}^\perp \cdots P_1^\perp A P_{n-2}^\perp = \cdots \text{etc},$$

which combined with (2.4.5) results in $P_n P_{n-1}^\perp \cdots P_1^\perp A\eta = 0$. □

2.5 The QR algorithm

The crucial assumption in the previous sections has been self-adjointness of the operator. Even when detecting false eigenvalues the tools we use rely heavily on self-adjointness. When we do not have self-adjointness the finite-section method may fail dramatically, the shift operator being a well known example. In fact the finite section method can behave extremely badly as the following theorem shows. First we need to recall a definition.

Definition 2.5.1. *Let A be a bounded operator on a Hilbert space $\mathcal{H}$. Then the numerical range of A is defined as*

$$W(A) = \{\langle A\xi, \xi\rangle : \|\xi\| = 1\},$$

and the essential numerical range is defined as

$$W_e(A) = \bigcap_{K \text{ compact}} \overline{W(A+K)}$$

Theorem 2.5.2. *(Pokrzywa)(Pok79) Let $A \in \mathcal{B}(\mathcal{H})$ and $\{P_n\}$ be a sequence of finite-dimensional projections converging strongly to the identity. Suppose that $S \subset W_e(A)$ then there exists a sequence $\{Q_n\}$ of finite-dimensional projections such that $P_n < Q_n$ (so $Q_n \to I$) strongly) and*

$$d_H(\sigma(A_n) \cup S, \sigma(\tilde{A}_n)) \to 0, \quad n \to \infty,$$

where

$$A_n = P_n A\lceil_{P_n\mathcal{H}}, \qquad \tilde{A}_n = Q_n A\lceil_{Q_n\mathcal{H}}$$

and d_H denotes the Hausdorff metric.

What Theorem 2.5.2 says is that if the essential range of a bounded operator A contains more than just elements from the spectrum, the finite section method may produce spectral pollution. As there is no restriction on the set S in Theorem 2.5.2 (e.g. S could be

isolated points or open sets), there is no hope that the finite section method can give any information about either the essential spectrum or isolated eigenvalues.

The next question is therefore; is there an alternative to the finite-section method in the case where the operator is not self-adjoint? Another important question is; can one find eigenvectors? These are the issues we will address when introducing the QR algorithm in infinite dimensions.

2.5.1 The QR decomposition

The QR algorithm is the standard tool for finding eigenvalues and eigenvectors in finite dimensions. We will discuss the method in detail, but first we need to extend the well known QR decomposition in finite dimensions to infinite dimensions.

Theorem 2.5.3. *Let A be a bounded operator on a separable Hilbert space $\mathcal{H}$ and let $\{e_j\}$ be an orthonormal basis for $\mathcal{H}$. Then there exists an isometry Q such that $A = QR$ where R is upper triangular with respect to $\{e_j\}$. Moreover*

$$Q = \underset{n\to\infty}{SOT\text{-}lim}\ V_n$$

where $V_n = U_1 \cdots U_n$ and U_j is a Householder transformation.

Proof. We will obtain R as the weak limit of a sequence $\{V_n^* A\}$ where V_n is unitary and the unitary operator is $Q = \text{SOT-lim}_{n\to\infty} V_n$. The procedure is as follows: Let P_n be the projection onto $\{e_1, \ldots, e_n\}$ and suppose that we have the n elements in the sequence and that the n-th element is an operator $R_n = V_n^* A$ such that, with respect to the decomposition $\mathcal{H} = P_n\mathcal{H} \oplus P_n^\perp\mathcal{H}$, we have

$$R_n = \begin{pmatrix} \tilde{R}_n & B_n \\ C_n & N_n \end{pmatrix}, \quad \tilde{R}_n = P_n R_n P_n,\ B_n = P_n R_n P_n^\perp,\ C_n = P_n^\perp R_n P_n,$$

where $N_n = P_n^\perp R_n P_n^\perp$ and $\tilde{R}_n$ is upper triangular and $Ce_j = 0$ for $j \le n-1$. Let $\zeta = Ce_n$. Choose $\xi \in P_n^\perp\mathcal{H}$ and define the Householder reflection $S \in \mathcal{B}(P_n^\perp\mathcal{H})$,

$$S = I - \frac{2}{\|\xi\|^2}\xi \otimes \bar{\xi}, \quad \text{and} \quad U_n = P_n \oplus S, \tag{2.5.1}$$

such that $S\zeta = \{\tilde{\zeta}_1, 0, 0, \ldots\}$. Finally let $R_{n+1} = U_n R_n$. Hence,

$$R_{n+1} = U_n R_n = \begin{pmatrix} \tilde{R}_n & B_n \\ SC_n & SN_n \end{pmatrix} = \begin{pmatrix} \tilde{R}_{n+1} & B_{n+1} \\ C_{n+1} & N_{n+1} \end{pmatrix}, \tag{2.5.2}$$

where the last matrix is understood to be with respect to the decomposition $\mathcal{H} = P_{n+1}\mathcal{H} \oplus P_{n+1}^\perp\mathcal{H}$. Note that, by the choice of S it is true that $\tilde{R}_{n+1}$ is upper triangular and $C_{n+1}e_j = 0$ for $j \le n$. Defining $R_1 = A$ and letting $V_n = U_1 \ldots U_n$, we have completed the construction of the sequence $\{V_n^* A\}$.

Note that $R_n = V_n^* A$ is bounded, since V_n is unitary (since U_j is unitary). And since a closed ball in $\mathcal{B}(\mathcal{H})$ is weakly sequentially compact, there is an $R \in \mathcal{B}(\mathcal{H})$ and a subsequence $\{R_{n_k}\}$ such that $R_{n_k} \xrightarrow{\text{WOT}} R$. But by (2.5.2) it is clear that for any integer j we have $P_j R_n P_j = P_j R_m P_j$ for sufficiently large n and m. Hence $\text{WOT-lim}_n R_n = R$.

Now, by (2.5.2) R is upper triangular with respect to $\{e_j\}$ and also $Re_j = R_n e_j$ for large n, thus SOT-$\lim_n R_n = R$. By similar reasoning, using the previous compactness argument (since V_n is bounded) and the fact that, by (2.5.1), for any integer j we have $V_n e_j = V_m e_j$ for sufficiently large m and n, it follows that there is a $V \in \mathcal{B}(\mathcal{H})$ such that $V_n \xrightarrow{\text{SOT}} V$ and, being a strong limit of unitary operators; V is an isometry. Let $Q = V$. Therefore, $A = QR$ since $A = V_n R_n$ and multiplication is jointly strongly continuous on bounded sets. □

2.5.2 The QR algorithm

Let $A \in \mathcal{B}(\mathcal{H})$ be invertible and let $\{e_j\}$ be an orthonormal basis for $\mathcal{H}$. By Theorem 2.5.3 we have $A = QR$, where Q is unitary and R is upper triangular with respect to $\{e_j\}$. Consider the following construction of unitary operators $\{\hat{Q}_k\}$ and upper triangular (w.r.t. $\{e_j\}$) operators $\{\hat{R}_k\}$. Let $A = Q_1 R_1$ be a QR decomposition of A and define $A_2 = R_1 Q_1$. Then QR factorize $A_2 = Q_2 R_2$ and define $A_3 = R_2 Q_2$. The recursive procedure becomes

$$A_{m-1} = Q_m R_m, \quad A_m = R_m Q_m. \tag{2.5.3}$$

Now define

$$\hat{Q}_m = Q_1 Q_2 \dots Q_m, \quad \hat{R}_m = R_m R_{m-1} \dots R_1. \tag{2.5.4}$$

Definition 2.5.4. *Let $A \in \mathcal{B}(\mathcal{H})$ be invertible and let $\{e_j\}$ be an orthonormal basis for $\mathcal{H}$. Sequences $\{\hat{Q}_j\}$ and $\{\hat{R}_j\}$ constructed as in (2.5.3) and (2.5.4) will be called a Q-sequence and an R-sequence of A with respect to $\{e_j\}$.*

The following observation will be useful in the later developments. From the construction in (2.5.3) and (2.5.4) we get

$$A = Q_1 R_1 = \hat{Q}_1 \hat{R}_1,$$

$$A^2 = Q_1 R_1 Q_1 R_1 = Q_1 Q_2 R_2 R_1 = \hat{Q}_2 \hat{R}_2,$$

$$A^3 = Q_1 R_1 Q_1 R_1 Q_1 R_1 = Q_1 Q_2 R_2 Q_2 R_2 R_1 = Q_1 Q_2 Q_3 R_3 R_2 R_1 = \hat{Q}_3 \hat{R}_3.$$

An easy induction gives us that

$$A^m = \hat{Q}_m \hat{R}_m.$$

Note that $\hat{R}_m$ must be upper triangular with respect to $\{e_j\}$ since R_j, $j \leq m$ is upper triangular with respect to $\{e_j\}$. Also, by invertibility of A, $\langle Re_i, e_i \rangle \neq 0$. From this it follows immediately that

$$\text{sp}\{A^m e_j\}_{j=1}^N = \text{sp}\{\hat{Q}_m e_j\}_{j=1}^N, \quad N \in \mathbb{N}. \tag{2.5.5}$$

In finite dimensions we have the following theorem:

Theorem 2.5.5. *Let $A \in \mathbb{C}^{N \times N}$ be a normal matrix with eigenvalues satisfying $|\lambda_1| > \dots > |\lambda_N|$. Let $\{\hat{Q}_m\}$ be a Q-sequence of unitary operators. Then $\hat{Q}_m A \hat{Q}_m^* \to D$, as $m \to \infty$, where D is diagonal.*

We will prove an analogue of this theorem in infinite dimensions, but first we need to state some presumably well-known results.

2.5.3 The distance and angle between subspaces

We follow the notation in (Kat95). Let $M \subset \mathcal{B}$ and $N \subset \mathcal{B}$ be closed subspaces of a Banach space $\mathcal{B}$. Define

$$\delta(M,N) = \sup_{\substack{x\in M\\ \|x\|=1}} \inf_{y\in M} \|x-y\|$$

and

$$\hat{\delta}(M,N) = \max[\delta(M,N), \delta(M,N)].$$

Given subspaces M and $\{M_k\}$ such that $\hat{\delta}(M_k, M) \to 0$ as $k \to \infty$, we will sometimes use the notation

$$M_k \xrightarrow{\hat{\delta}} M, \qquad k \to \infty.$$

If we replace $\mathcal{B}$ with a Hilbert space H we can express δ and $\hat{\delta}$ conveniently in terms of projections and operator norms. In particular, if E and F are the projections onto subspaces $M \subset \mathcal{H}$ and $N \subset \mathcal{H}$ respectively then

$$\delta(M,N) = \sup_{\substack{x\in M\\ \|x\|=1}} \inf_{y\in N} \|x-y\| = \sup_{\substack{x\in M\\ \|x\|=1}} \|F^{\perp}x\| = \|F^{\perp}E\|.$$

Since the operator $E - F = F^{\perp}E - FE^{\perp}$ is essentially the direct sum of operators $F^{\perp}E \oplus (-FE^{\perp})$, its norm is $\hat{\delta}(M,N)$, i.e.

$$\hat{\delta}(M,N) = \max(\|F^{\perp}E\|, \|E^{\perp}F\|) = \max(\|F^{\perp}E\|, \|FE^{\perp}\|) = \|E-F\|. \tag{2.5.6}$$

These observations come in handy in the proof of the next proposition.

Proposition 2.5.6. *Let $\{A_n\}$ be a sequence of N-dimensional subspaces of a Hilbert space $\mathcal{H}$ and let $B \subset \mathcal{H}$ be an N-dimensional subspace. If $\delta(A_n, B) \to 0$ or $\delta(B, A_n) \to 0$ then $\hat{\delta}(A_n, B) \to 0$.*

Proof. Suppose that $\delta(A_n, B) \to 0$. Let E_n and F be the projections onto onto A_n and B respectively. We need to show that $\|E_n - F\| \to 0$ as $n \to \infty$. Now E_n and F are N-dimensional projections such that $\|E_n^{\perp}F\| \to 0$ as $n \to \infty$. Thus, in view of (2.5.6), it suffices to show that $\|F^{\perp}E_n\| \to 0$. For the proof, note that

$$\|F - FE_nF\| = \|FE_n^{\perp}F\| \le \|E_n^{\perp}F\| \longrightarrow 0, \qquad n \to \infty.$$

Since $FE_n^{\perp}F$ can be viewed as a sequence of positive contractions acting on the finite dimensional space $F\mathcal{H}$, it follows that $\operatorname{trace}(F - FE_nF) \to 0$. Hence

$$\begin{aligned}\|F^{\perp}E_n\|^2 &= \|E_n - E_nFE_n\| \le \operatorname{trace}(E_n - E_nFE_n)\\ &= N - \operatorname{trace}(E_nFE_n) = N - \operatorname{trace}(FE_nF)\\ &= \operatorname{trace}(F - FE_nF) \to 0, \qquad n \to \infty.\end{aligned}$$

The proof that if $\delta(B, A_n) \to 0$ then $\hat{\delta}(A_n, B) \to 0$ is similar to the previous argument. □

Proposition 2.5.7. *Let $E = E_1 \oplus \ldots \oplus E_M$ where the E_js are finite-dimensional subspaces of a Hilbert space $\mathcal{H}$. Let $F_k = E_{1,k} + \ldots + E_{M,k}$ where $\hat{\delta}(E_{j,k}, E_j) \to 0$ as $k \to \infty$. Then $F_k \xrightarrow{\hat{\delta}} E$.*

Proof. Note that for projections P and Q on a Hilbert space where $\|P - Q\| < 1$ implies that $\dim P = \dim Q$. So writing E_j for the projection onto the space E_j etc., the hypothesis $\|E_{j,k} - E_j\| = \hat{\delta}(E_{j,k}, E_j) \to 0$ implies that $\dim E_{j,k} = \dim E_j$ for large k. The assertion now follows by Proposition 2.5.6 and the fact that

$$\delta(E, F_k) \leq \sum_{j=1}^{M} \|E_j - E_{j,k}\| \longrightarrow 0, \qquad k \to \infty.$$

□

Theorem 2.5.8. *Let $A \in \mathcal{B}(\mathcal{H})$ be an invertible normal operator. Suppose that $\sigma(A) = \omega \cup \Omega$ is a disjoint union such that $\omega = \{\lambda_i\}_{i=1}^{N}$ and the λ_is are isolated eigenvalues of finite multiplicity satisfying $|\lambda_1| > \ldots > |\lambda_N|$. Suppose further that $\sup\{|\gamma| : \gamma \in \Omega\} < |\lambda_N|$. Let $\{\xi_i\}_{i=1}^{M}$ be a collection of linearly independent vectors in $\mathcal{H}$ such that $\{\chi_\omega(A)\xi_i\}_{i=1}^{M}$ are linearly independent. The following observations are true.*

(i) There exists an M-dimensional subspace $B \subset \mathrm{ran}\chi_\omega(A)$ such that

$$\mathrm{sp}\{A^k \xi_i\}_{i=1}^{M} \xrightarrow{\hat{\delta}} B, \quad k \to \infty.$$

(ii) If

$$\mathrm{sp}\{A^k \xi_i\}_{i=1}^{M-1} \xrightarrow{\hat{\delta}} D \subset \mathcal{H}, \quad k \to \infty,$$

where D is an $(M-1)$-dimensional subspace, then

$$\mathrm{sp}\{A^k \xi_i\}_{i=1}^{M} \xrightarrow{\hat{\delta}} D \oplus \mathrm{sp}\{\xi\}, \quad k \to \infty,$$

where $\xi \in \mathrm{ran}\chi_\omega(A)$ is an eigenvector of A.

Proof. We will first prove (i). Consider the following construction of B: Let $\tilde{\lambda}_1 \in \{\lambda_i\}_{i=1}^{N}$ be the largest (in absolute value) element such that

$$\{\chi_{\tilde{\lambda}_1}(A)\xi_i\}_{i=1}^{M} \neq \{0\}.$$

If $\{\chi_{\tilde{\lambda}_1}(A)\xi_i\}_{i=1}^{M}$ are linearly independent let $B = \{\chi_{\tilde{\lambda}_1}(A)\xi_i\}_{i=1}^{M}$. If not, then $\{\chi_{\tilde{\lambda}_1}(A)\xi_i\}_{i=1}^{M}$ are linearly dependent spanning a space of dimension $k_1 < M$. By taking linear combinations of elements in $\{\xi_i\}_{i=1}^{M}$ we can find a new basis $\{\xi_{1,i}\}_{i=1}^{M}$ for $\mathrm{sp}\{\xi_i\}_{i=1}^{M}$ such that $\mathrm{sp}\{\chi_{\tilde{\lambda}_1}(A)\xi_{1,i}\}_{i=1}^{k_1} = \mathrm{sp}\{\chi_{\tilde{\lambda}_1}(A)\xi_i\}_{i=1}^{M}$ and $\chi_{\tilde{\lambda}_1}(A)\xi_{1,i} = 0$, for $k_1 + 1 \leq i \leq M$. Let $\tilde{\lambda}_2 \in \{\lambda_i\}_{i=1}^{N} \setminus \{\tilde{\lambda}_1\}$ be the largest element such that $\{\chi_{\tilde{\lambda}_2}(A)\xi_{1,i}\}_{i=k_1+1}^{M} \neq \{0\}$. If $\{\chi_{\tilde{\lambda}_2}(A)\xi_{1,i}\}_{i=k_1+1}^{M}$ are linearly independent let

$$B = \mathrm{sp}\{\chi_{\tilde{\lambda}_1}(A)\xi_{1,i}\}_{i=1}^{k_1} \oplus \mathrm{sp}\{\chi_{\tilde{\lambda}_2}(A)\xi_{1,i}\}_{i=k_1+1}^{M}.$$

If $\{\chi_{\tilde{\lambda}_2}(A)\xi_{1,i}\}_{i=k_1+1}^{M}$ are linearly dependent, spanning a space of dimension k_2, we proceed exactly as in the previous step. Repeating this process until $\{\chi_{\tilde{\lambda}_{n+1}}(A)\xi_{n,i}\}_{i=k_n+1}^{M}$ is linearly independent (note that this is possible by the assumption that $\{\chi_\omega(A)\xi_i\}_{i=1}^{M}$ are linearly independent) we get

$$B = \bigoplus_{j=1}^{n} \mathrm{sp}\{\chi_{\tilde{\lambda}_j}(A)\xi_{j,i}\}_{i=k_{j-1}+1}^{k_j} \oplus \mathrm{sp}\{\chi_{\tilde{\lambda}_{n+1}}(A)\xi_{n,i}\}_{i=k_n+1}^{M}, \quad n \leq N-1,$$

where $k_0 = 0$. We claim that $\mathrm{sp}\{A^k\xi_i\}_{i=1}^{M} \xrightarrow{\hat{\delta}} B$ as $k \to \infty$. Since

$$\dim(\mathrm{sp}\{A^k\xi_i\}_{i=1}^{M}) = M = \dim(B),$$

(recall that A is invertible) and

$$\mathrm{sp}\{A^k\xi_i\}_{i=1}^{M} = \sum_{j=1}^{n} \mathrm{sp}\{A^k\xi_{j,i}\}_{i=k_{j-1}+1}^{k_j} + \mathrm{sp}\{A^k\xi_{n,i}\}_{i=k_n+1}^{M}$$

by Proposition 2.5.7, we only have to demonstrate that

$$\mathrm{sp}\{A^k\xi_{j,i}\}_{i=k_{j-1}+1}^{k_j} \xrightarrow{\hat{\delta}} \mathrm{sp}\{\chi_{\tilde{\lambda}_j}(A)\xi_{j,i}\}_{i=k_{j-1}+1}^{k_j}, \quad k \to \infty, \quad j \leq n, \tag{2.5.7}$$

and

$$\mathrm{sp}\{A^k\xi_{n,i}\}_{i=k_n+1}^{M} \xrightarrow{\hat{\delta}} \mathrm{sp}\{\chi_{\tilde{\lambda}_{n+1}}(A)\xi_{n,i}\}_{i=k_n+1}^{M}. \tag{2.5.8}$$

To prove (2.5.7), by Proposition 2.5.6, we only need to show that

$$\begin{gathered} \sup_{\substack{\zeta\in E \\ \|\zeta\|=1}} \inf_{\eta\in E_k} \|\zeta-\eta\| = \delta(E,E_k) \longrightarrow 0, \quad k\to\infty, \\ E_k = \mathrm{sp}\{A^k\xi_{j,i}\}_{i=k_{j-1}+1}^{k_j},\ E = \mathrm{sp}\{\chi_{\tilde{\lambda}_j}(A)\xi_{j,i}\}_{i=k_{j-1}+1}^{k_j}, \end{gathered} \tag{2.5.9}$$

since $\dim E = \dim E_k$. It is easy to see that (2.5.9) will follow if for any sequence $\{\zeta_k\} \subset E$ of unit vectors there exists a sequence $\{\eta_k\}$, where $\eta_k \in E_k$, such that $\|\zeta_k - \eta_k\| \to 0$. To show this, note that by compactness of the unit ball in E we can assume, possibly passing to a subsequence, that $\zeta_k \to \zeta$. Thus, the task is reduced to showing that we can find $\{\eta_k\}$ such that $\|\zeta - \eta_k\| \to 0$. Now, $\zeta = \sum_i \alpha_i \chi_{\tilde{\lambda}_j}(A)\xi_{j,i}$, for some complex numbers $\{\alpha_i\}$, and we claim that the right choice of $\{\eta_k\}$ is

$$\eta_k = \sum_i \alpha_i A^k \xi_{j,i}/\tilde{\lambda}_j^k.$$

Indeed, by the previous construction, $\xi_{j,i} \perp \mathrm{ran}\chi_{\tilde{\lambda}_l}(A)$ for $l > j$. Thus,

$$\xi_{j,i} = (\chi_{\tilde{\lambda}_j}(A) + \chi_\theta(A))\xi_{j,i}, \qquad \theta = \{\lambda \in \sigma(A) : |\lambda| < |\tilde{\lambda}_j|\}.$$

This gives $A^k\xi_{j,i} = \tilde{\lambda}_j^k \chi_{\tilde{\lambda}_j}(A)\xi_{j,i} + A^k\chi_\theta(A)\xi_{j,i}$. Now, by the assumption on $\sigma(A)$, we have

$$\rho = \sup\{|z| : z \in \theta\} < |\tilde{\lambda}_j|.$$

Thus, since

$$\|A^k\chi_\theta(A)\xi_{j,i}\|/|\tilde{\lambda}_j^k| < |\rho/\tilde{\lambda}_j|^k\|\chi_\theta(A)\xi_{j,i}\|,$$

we have

$$A^k\xi_{j,i}/\tilde{\lambda}_j^k = (\tilde{\lambda}_j^k\chi_{\tilde{\lambda}_j}(A)\xi_{j,i} + A^k\chi_\theta(A)\xi_{j,i})/\tilde{\lambda}_j^k \longrightarrow \chi_{\tilde{\lambda}_j}(A)\xi_{j,i}, \quad k \to \infty,$$

which yields our claim. Now (2.5.8) follows by a similar argument.

To show (ii), note that, by the argument in the proof of (i) and our assumption, we have

$$\begin{aligned}\mathrm{sp}\{A^k\xi_i\}_{i=1}^{M-1} \xrightarrow{\hat{\delta}} D = \bigoplus_{j=1}^{n} \mathrm{sp}\{\chi_{\tilde{\lambda}_j}(A)\xi_{j,i}\}_{i=k_{j-1}+1}^{k_j} \\ \oplus\, \mathrm{sp}\{\chi_{\tilde{\lambda}_{n+1}}(A)\xi_{n,i}\}_{i=k_n+1}^{M-1}, \quad k \to \infty,\end{aligned} \tag{2.5.10}$$

for $n \leq N-2$, where $k_0 = 0$, $\{\tilde{\lambda}_j\}$ and $\{\xi_{j,i}\}$ are constructed as in the proof of (i). Now, there are two possibilities:

(1) There exists $\lambda \in \Lambda = \omega \setminus \{\tilde{\lambda}_j\}_{j=1}^{n+1}$ such that $\chi_\lambda(A)\xi_M \neq 0$.

(2) We have that $\chi_\Lambda(A)\xi_M = 0$.

Starting with Case 1 we may argue as in the proof of (i) to deduce that

$$\begin{aligned}\mathrm{sp}\{A^k\xi_i\}_{i=1}^{M} \xrightarrow{\hat{\delta}} \bigoplus_{j=1}^{n} \mathrm{sp}\{\chi_{\tilde{\lambda}_j}(A)\xi_{j,i}\}_{i=k_{j-1}+1}^{k_j} \\ \oplus\, \mathrm{sp}\{\chi_{\tilde{\lambda}_{n+1}}(A)\xi_{n,i}\}_{i=k_n+1}^{M-1} \oplus \mathrm{sp}\{\chi_{\tilde{\lambda}_{n+2}}(A)\xi_M\}, \quad k \to \infty,\end{aligned}$$

where $\tilde{\lambda}_{n+2} \in \omega \setminus \{\tilde{\lambda}_j\}_{j=1}^{n+1}$ is the largest element such that $\chi_{\tilde{\lambda}_{n+2}}(A)\xi_M \neq 0$, (note that the existence of $\tilde{\lambda}_{n+2}$ is guaranteed by the assumption that $\{\chi_\omega(A)\xi_i\}_{i=1}^{M}$ are linearly independent) and this yields the assertion.

Note that Case 2 has two subcases, namely,

(I) $\chi_\Lambda(A)\xi_M = 0$, but $\{\chi_{\tilde{\lambda}_{n+1}}(A)\xi_{n+1,i}\}_{i=k_n+1}^{M-1}$ and $\chi_{\tilde{\lambda}_{n+1}}(A)\xi_M$ are linearly independent.

(II) $\chi_\Lambda(A)\xi_M = 0$ and $\{\chi_{\tilde{\lambda}_{n+1}}(A)\xi_{n+1,i}\}_{i=k_n+1}^{M-1}$ and $\chi_{\tilde{\lambda}_{n+1}}(A)\xi_M$ are linearly dependent, but there exists a $\tilde{\lambda}_l$, the largest eigenvalue in $\{\tilde{\lambda}_j\}_{j=1}^{n+1}$ such that $\{\chi_{\tilde{\lambda}_l}(A)\xi_{l,i}\}_{i=k_{l-1}+1}^{k_l}$ and $\chi_{\tilde{\lambda}_l}(A)\xi_M$ are linearly independent.

Note that we cannot have $\chi_\Lambda(A)\xi_M = 0$ and also have that

$$\{\chi_{\tilde{\lambda}_j}(A)\xi_{j,i}\}_{i=k_{j-1}+1}^{k_j} \quad \text{and} \quad \chi_{\tilde{\lambda}_j}(A)\xi_M, \quad j \leq n,$$

are linearly dependent as well as $\{\chi_{\tilde{\lambda}_{n+1}}(A)\xi_{n,i}\}_{i=k_n+1}^{M-1}$ and $\chi_{\tilde{\lambda}_{n+1}}(A)\xi_M$ are linearly dependent at the same time because that would violate the linear independence assumption on $\{\chi_\omega(A)\xi_i\}_{i=1}^{M}$.

To prove (II) we may argue as in the proof of (i) and deduce that

$$\mathrm{sp}\{A^k\xi_{l,i}\}_{i=k_{l-1}+1}^{k_l} \xrightarrow{\hat{\delta}} \mathrm{sp}\{\chi_{\tilde{\lambda}_l}(A)\xi_{l,i}\}_{i=k_{l-1}+1}^{k_l}, \quad k\to\infty$$

and

$$\begin{aligned}\mathrm{sp}\{A^k\xi_{l,i}\}_{i=k_{l-1}+1}^{k_l}&+\mathrm{sp}\{A^k\chi_\Gamma(A)\xi_M\}\\ &\xrightarrow{\hat{\delta}} \mathrm{sp}\{\chi_{\tilde{\lambda}_l}(A)\xi_{l,i}\}_{i=k_{l-1}+1}^{k_l}+\mathrm{sp}\{\chi_{\tilde{\lambda}_l}(A)\xi_M\}, \quad k\to\infty\end{aligned}$$

where $\Gamma=\omega\setminus\{\tilde{\lambda}_j\}_{j=1}^{l-1}$. Thus, using (2.5.10), it is easy to see that this gives

$$\begin{aligned}\mathrm{sp}\{A^k\xi_i\}_{i=1}^M \xrightarrow{\hat{\delta}} &\bigoplus_{j=1}^{l-1}\mathrm{sp}\{\chi_{\tilde{\lambda}_j}(A)\xi_{j,i}\}_{i=k_{j-1}+1}^{k_j}\\ &\oplus\Big(\mathrm{sp}\{\chi_{\tilde{\lambda}_l}(A)\xi_{l,i}\}_{i=k_{l-1}+1}^{k_l}+\mathrm{sp}\{\chi_{\tilde{\lambda}_l}(A)\xi_M\Big)\}\\ &\bigoplus_{j=l+1}^{n}\mathrm{sp}\{\chi_{\tilde{\lambda}_j}(A)\xi_{j,i}\}_{i=k_{j-1}+1}^{k_j}\oplus\mathrm{sp}\{\chi_{\tilde{\lambda}_{n+1}}(A)\xi_{n,i}\}_{i=k_n+1}^{M-1}.\end{aligned}$$

Thus, letting P be the projection onto $\mathrm{sp}\{\chi_{\tilde{\lambda}_l}(A)\xi_{l,i}\}_{i=k_{l-1}+1}^{k_l}$, it follows that

$$\begin{aligned}\mathrm{sp}\{A^k\xi_i\}_{i=1}^M \xrightarrow{\hat{\delta}} &\bigoplus_{j=1}^{l-1}\mathrm{sp}\{\chi_{\tilde{\lambda}_j}(A)\xi_{j,i}\}_{i=k_{j-1}+1}^{k_j}\\ &\oplus\mathrm{sp}\{\chi_{\tilde{\lambda}_l}(A)\xi_{l,i}\}_{i=k_{l-1}+1}^{k_l}\oplus P^\perp\mathrm{sp}\{\chi_{\tilde{\lambda}_l}(A)\xi_M\}\\ &\bigoplus_{j=l+1}^{n}\mathrm{sp}\{\chi_{\tilde{\lambda}_j}(A)\xi_{j,i}\}_{i=k_{j-1}+1}^{k_j}\oplus\mathrm{sp}\{\chi_{\tilde{\lambda}_{n+1}}(A)\xi_{n,i}\}_{i=k_n+1}^{M-1}.\end{aligned}$$

Now Case (I) follows by similar reasoning. □

Theorem 2.5.9. *Let $A\in\mathcal{B}(\mathcal{H})$ be an invertible normal operator and let $\{e_j\}$ be an orthonormal basis for $\mathcal{H}$. Let $\{Q_k\}$ and $\{R_k\}$ be a Q- and R-sequences of A with respect to $\{e_j\}$. Suppose also that $\sigma(A)=\omega\cup\Omega$ such that $\omega\cap\Omega=\emptyset$ and $\omega=\{\lambda_i\}_{i=1}^N$, where the λ_is are isolated eigenvalues with finite multiplicity satisfying $|\lambda_1|>\ldots>|\lambda_N|$. Suppose further that $\sup\{|\theta|:\theta\in\Omega\}<|\lambda_N|$. Then there is a subset $\{\hat{e}_j\}_{j=1}^M\subset\{e_j\}$ such that $\mathrm{sp}\{Q_k\hat{e}_j\}\to\mathrm{sp}\{\hat{q}_j\}$ where $\hat{q}_j$ is an eigenvector of A and $M=\dim(\mathrm{ran}\chi_\omega(A))$. Moreover, $\mathrm{sp}\{\hat{q}_j\}_{j=1}^M=\mathrm{ran}\chi_\omega(A)$. Also, if $e_j\notin\{\hat{e}_j\}_{j=1}^M$, then $\chi_\omega(A)Q_ke_j\to 0$.*

The theorem will be proven in several steps. First we need a definition.

Definition 2.5.10. *Suppose that the hypotheses in Theorem 2.5.9 are true and let K be the smallest integer such that $\dim(\mathrm{sp}\{\chi_\omega(A)e_j\}_{j=1}^K)=M$. Define*

$$\Lambda_\omega=\{e_j:\chi_\omega(A)e_j\neq 0, j\leq K\}\quad \Lambda_\Omega=\{e_j:\chi_\omega(A)e_j=0, j\leq K\}$$

and $\tilde{\Lambda}_\omega=\{e_j\in\Lambda_\omega:\chi_\omega(A)e_j\in\mathrm{sp}\{\chi_\omega(A)e_i\}_{i=1}^{j-1}\}$.

The decomposition of A into

$$A = \left(\sum_{j=1}^{M} \lambda_j \, \xi_j \otimes \bar{\xi}_j \right) \oplus \chi_\Omega(A) A, \quad \lambda_j \in \omega.$$

where $\{\xi_j\}_{j=1}^m$ is an orthonormal set of eigenvectors of A as well as the following two technical lemmas will be useful in the proof.

Lemma 2.5.11. *Let $\{\hat{e}_1, \ldots, \hat{e}_M\} = \Lambda_\omega \setminus \tilde{\Lambda}_\omega$. If $e_m \in \Lambda_\Omega \cup \tilde{\Lambda}_\omega$, then*

$$\mathrm{sp}\{\chi_\omega(A) q_{k,j}\}_{j=1}^m = \mathrm{sp}\{\chi_\omega(A) \hat{q}_{k,j}\}_{j=1}^{s(m)}, \quad q_{k,j} = Q_k e_j, \quad \hat{q}_{k,j} = Q_k \hat{e}_j,$$

where $s(m)$ is the largest integer such that $\{\hat{e}_j\}_{j=1}^{s(m)} \subset \{e_j\}_{j=1}^m$.

Proof. We will show this by induction on the set $\{\tilde{e}_1, \ldots, \tilde{e}_p\} = \Lambda_\Omega \cup \tilde{\Lambda}_\omega$. Consider $\tilde{e}_\mu \in \{\tilde{e}_1, \ldots, \tilde{e}_p\}$. Then $\tilde{e}_\mu = e_{\tilde{m}}$ for some integer $\tilde{m}$. Suppose that $\mathrm{sp}\{\chi_\omega(A) q_{k,j}\}_{j=1}^{\tilde{m}} = \mathrm{sp}\{\chi_\omega(A) \hat{q}_{k,j}\}_{j=1}^{s(\tilde{m})}$. We will show that

$$\mathrm{sp}\{\chi_\omega(A) q_{k,j}\}_{j=1}^m = \mathrm{sp}\{\chi_\omega(A) \hat{q}_{k,j}\}_{j=1}^{s(m)},$$

where $e_m = \tilde{e}_{\mu+1}$.

First, note that $\mathrm{sp}\{\chi_\omega(A) q_{k,j}\}_{j=1}^{m-1} = \mathrm{sp}\{\chi_\omega(A) \hat{q}_{k,j}\}_{j=1}^{s(m)}$ follows from the induction hypothesis. Indeed, let β be the largest integer such that $\beta < m$ and $e_\beta \in \Lambda_\omega \setminus \tilde{\Lambda}_\omega$ i.e. if $\hat{e}_t = e_\beta$ then $t = s(m)$. Observe that since $e_{\tilde{m}} = \tilde{e}_\mu$ and $e_m = \tilde{e}_{\mu+1}$, it follows that if $\tilde{m} < \alpha < m$ then $e_\alpha \in \Lambda_\omega \setminus \tilde{\Lambda}_\omega$. So if $\beta < m-1$ then there is no $e_\alpha \in \Lambda_\omega \setminus \tilde{\Lambda}_\omega$ such that $\tilde{m} < \alpha < m$. Thus, $\tilde{m} = m-1$ and so $t = s(m) = s(\tilde{m})$, yielding the assertion.

If $\beta = m-1$ then for every e_j where $\tilde{m} < j \leq m-1$ we have $e_j \in \Lambda_\omega \setminus \tilde{\Lambda}_\omega$. So $e_{\tilde{m}+\nu} = \hat{e}_{s(\tilde{m})+\nu}$ for $\tilde{m}+\nu \leq m-1$ and $\nu \geq 1$, hence, $q_{k,\tilde{m}+\nu} = \hat{q}_{k,s(\tilde{m})+\nu}$ for $\tilde{m}+\nu \leq m-1$. Also, $e_{m-1} = \hat{e}_{s(m)}$ so $q_{k,m-1} = \hat{q}_{k,s(m)}$. Thus,

$$\begin{aligned} \mathrm{sp}\{\chi_\omega(A) q_{k,j}\}_{j=1}^{m-1} &= \mathrm{sp}\{\chi_\omega(A) q_{k,j}\}_{j=1}^{\tilde{m}} + \mathrm{sp}\{\chi_\omega(A) q_{k,j}\}_{j=\tilde{m}+1}^{m-1} \\ &= \mathrm{sp}\{\chi_\omega(A) q_{k,j}\}_{j=1}^{\tilde{m}} + \mathrm{sp}\{\chi_\omega(A) \hat{q}_{k,j}\}_{j=s(\tilde{m})+1}^{s(m)}, \end{aligned}$$

and by recalling the induction hypothesis this yields the assertion. Thus, we only need to prove that $\chi_\omega(A) q_{k,m} \in \mathrm{sp}\{\chi_\omega(A) q_{k,j}\}_{j=1}^{m-1}$. To show this, note that

$$\chi_\omega(A) A^k e_m = \sum_{j=1}^{m} r_{k,j} \chi_\omega(A) q_{k,j}, \quad r_{k,j} = \langle R_k e_m, e_j \rangle.$$

Note further that, since A is invertible, we have $r_{k,m} \neq 0$. In the case $e_m \in \Lambda_\Omega$ we have $\chi_\omega(A) A^k e_m = 0$. So, since $r_{k,m} \neq 0$, it follows that $\chi_\omega(A) q_{k,m}$ is a linear combination of elements in $\mathrm{sp}\{\chi_\omega(A) q_{k,j}\}_{j=1}^{m-1}$. In the case $e_m \in \tilde{\Lambda}_\omega$ note that, by again using the fact that $\chi_\omega(A) A^k e_m = \sum_{j=1}^m r_{k,j} \chi_\omega(A) q_{k,j}$ and $r_{k,m} \neq 0$, we only have to show that $\chi_\omega(A) A^k e_m \in \mathrm{sp}\{\chi_\omega(A) q_{k,j}\}_{j=1}^{m-1}$. Now, this is indeed the case. Since $e_m \in \tilde{\Lambda}_\omega$ we have that $\chi_\omega(A) e_m \in \mathrm{sp}\{\chi_\omega(A) e_j\}_{j=1}^{m-1}$. Thus, since A is invertible

$$\chi_\omega(A) A^k e_m \in \mathrm{sp}\{\chi_\omega(A) A^k e_j\}_{j=1}^{m-1}.$$

Also, observe that, by (2.5.5),

$$\mathrm{sp}\{A^k e_j\}_{j=1}^{m-1} = \mathrm{sp}\{q_{k,j}\}_{j=1}^{m-1}.$$

Hence,

$$\mathrm{sp}\{\chi_\omega(A)A^k e_j\}_{j=1}^{m-1} = \mathrm{sp}\{\chi_\omega(A)q_{k,j}\}_{j=1}^{m-1},$$

and this yields the assertion.

The initial induction step follows from a similar argument and we are done. □

Lemma 2.5.12. *Let* $\{\hat{e}_1, \ldots, \hat{e}_M\} = \Lambda_\omega \setminus \tilde{\Lambda}_\omega$. *Suppose that* $\mathrm{sp}\{\hat{q}_{k,j}\} \to \mathrm{sp}\{\hat{q}_j\}$ *for all* $j \le \mu$ *for some* $\mu < M$, *where* $\hat{q}_{k,j} = Q_k \hat{e}_j$ *and* $\hat{q}_j$ *is an eigenvector of* $\sum_{j=1}^M \lambda_j \, \xi_j \otimes \bar{\xi}_j$. *Let* $e_l = \hat{e}_{\mu+1}$. *If* $e_m \in \Lambda_\Omega \cup \tilde{\Lambda}_\omega$, *where* $m < l$ *then*

$$\chi_\omega(A)q_{k,m} \to 0, \qquad k \to \infty, \quad q_{k,m} = Q_k e_m.$$

Proof. Arguing by contradiction, suppose that $\chi_\omega(A)q_{k,m} \nrightarrow 0$. Since $\chi_\omega(A)$ has finite rank we may assume that $\chi_\omega(A)q_{k,m} \to q$. Note that by using the assumptions stated and the fact that Q_k is unitary (since A is invertible) it is straightforward to show that

$$\mathrm{sp}\{\chi_\omega(A)\hat{q}_{k,j}\}_{j=1}^{\mu} \xrightarrow{\hat{\delta}} \mathrm{sp}\{\chi_\omega(A)\hat{q}_j\}_{j=1}^{\mu}, \qquad k \to \infty.$$

Also, by using the notation and results from Lemma 2.5.11 we have that $s(m) = \mu$ and

$$\mathrm{sp}\{\chi_\omega(A)q_{k,j}\}_{j=1}^{m} = \mathrm{sp}\{\chi_\omega(A)\hat{q}_{k,j}\}_{j=1}^{s(m)},$$

and thus it follows that

$$q \in \mathrm{sp}\{\chi_\omega(A)\hat{q}_j\}_{j=1}^{\mu}.$$

Now

$$|\langle \chi_\omega(A)q_{k,m}, \hat{q}_{k,j}\rangle| \to |\langle \chi_\omega(A)q, \hat{q}_j\rangle|, \quad k \to \infty, \quad j \le \mu.$$

Also, observe that

$$\langle \chi_\omega(A)q_{k,m}, \hat{q}_{k,j}\rangle \to 0, \qquad k \to \infty, \quad j \le \mu.$$

Indeed, this is true by the facts that $q_{k,m} \perp \hat{q}_{k,j}$ and $\langle \chi_\Omega(A)q_{k,m}, \hat{q}_{k,j}\rangle \to 0$ for all $j \le \mu$, where the latter follows since $\mathrm{sp}\{\hat{q}_{k,j}\} \to \mathrm{sp}\{\hat{q}_j\}$ and $\chi_\Omega(A)\hat{q}_j = 0$. Hence, it follows that $\langle \chi_\omega(A)q, \hat{q}_j\rangle = 0$ for $j \le \mu$. So since $q \in \mathrm{sp}\{\chi_\omega(A)\hat{q}_j\}_{j=1}^{\mu}$, we have that $q = 0$, and we have reached the contradiction. □

Proof. Proof of Theorem 2.5.9 Let $\{\hat{e}_1, \ldots, \hat{e}_M\} = \Lambda_\omega \setminus \tilde{\Lambda}_\omega$. We claim that this is the desired subset of $\{e_j\}$ described in the theorem, i.e. we claim that for $\hat{e}_j \in \Lambda_\omega \setminus \tilde{\Lambda}_\omega$ it is true that $\mathrm{sp}\{\hat{q}_{k,j}\} \to \mathrm{sp}\{\hat{q}_j\}$, where $\hat{q}_{k,j} = Q_k \hat{e}_j$ and $\hat{q}_j$ is an eigenvector of $\sum_{j=1}^M \lambda_j \, \xi_j \otimes \bar{\xi}_j$. We will prove this by induction.

Suppose that $\mathrm{sp}\{\hat{q}_{k,j}\} \to \mathrm{sp}\{\hat{q}_j\}$ for $j \le \mu$. Suppose also that

$$\mathrm{sp}\{A^k \hat{e}_i\}_{i=1}^{\mu} \xrightarrow{\hat{\delta}} \mathrm{sp}\{\hat{q}_i\}_{i=1}^{\mu}, \qquad k \to \infty. \tag{2.5.11}$$

We will show that $\mathrm{sp}\{\hat{q}_{k,\mu+1}\} \to \mathrm{sp}\{\hat{q}_{\mu+1}\}$ and $\mathrm{sp}\{A^k\hat{e}_i\}_{i=1}^{\mu+1} \stackrel{\hat{\delta}}{\to} \mathrm{sp}\{\hat{q}_i\}_{i=1}^{\mu+1}$ where $\hat{q}_{\mu+1}$ is the desired eigenvector of $\sum_{j=1}^{M} \lambda_j \, \xi_j \otimes \bar{\xi}_j$. By using (2.5.11) and appealing to Theorem 2.5.8 it follows that

$$\mathrm{sp}\{A^k\hat{e}_i\}_{i=1}^{\mu+1} \stackrel{\hat{\delta}}{\longrightarrow} \mathrm{sp}\{\hat{q}_i\}_{i=1}^{\mu} \oplus \mathrm{sp}\{\xi\}, \quad \xi \in \mathrm{ran}\chi_\omega(A), \tag{2.5.12}$$

where ξ is an eigenvector of A. Hence, to prove the induction assertion we need to show that $\mathrm{sp}\{\hat{q}_{\mu+1,k}\} \to \mathrm{sp}\{\xi\}$.

Let $e_l = \hat{e}_{\mu+1}$. Note that $\hat{\delta}(\mathrm{sp}\{\hat{q}_i\}_{i=1}^{\mu} \oplus \mathrm{sp}\{\xi\}, \mathrm{sp}\{A^k\hat{e}_i\}_{i=1}^{\mu+1}) \to 0$ implies

$$\delta(\mathrm{sp}\{\hat{q}_i\}_{i=1}^{\mu} \oplus \mathrm{sp}\{\xi\}, \mathrm{sp}\{A^k e_i\}_{i=1}^{l}) \to 0,$$

since $\mathrm{sp}\{A^k\hat{e}_i\}_{i=1}^{\mu+1} \subset \mathrm{sp}\{A^k e_i\}_{i=1}^{l}$. Thus, it follows that

$$\begin{aligned} &\delta(\mathrm{sp}\{\hat{q}_i\}_{i=1}^{\mu} \oplus \mathrm{sp}\{\xi\}, \mathrm{sp}\{q_{k,i}\}_{i=1}^{l}) \\ &\qquad = \delta(\mathrm{sp}\{\hat{q}_i\}_{i=1}^{\mu} \oplus \mathrm{sp}\{\xi\}, \mathrm{sp}\{A^k e_i\}_{i=1}^{l}) \longrightarrow 0, \quad k \to \infty, \end{aligned} \tag{2.5.13}$$

since A is invertible, $A^k = Q_k R_k$ and R_k is upper triangular with respect to $\{e_j\}$. We will use this to prove that $\mathrm{sp}\{\hat{q}_{\mu+1,k}\} = \mathrm{sp}\{q_{l,k}\} \to \mathrm{sp}\{\xi\}$. Note that this, by Proposition 2.5.6, is equivalent to proving $\delta(\mathrm{sp}\{\xi\}, \mathrm{sp}\{q_{l,k}\}) \to 0$, which we henceforth do. Note also that

$$\sup_{\substack{\zeta \in \mathrm{sp}\{\xi\} \\ \|\zeta\|=1}} \inf_{\eta \in \mathrm{sp}\{q_{l,k}\}} \|\zeta - \eta\| = \delta(\mathrm{sp}\{\xi\}, \mathrm{sp}\{q_{l,k}\}),$$

thus the latter assertion follows if we can show that for any sequence $\{\zeta_k\}$ of unit vectors in $\mathrm{sp}\{\xi\}$ there exists a sequence $\{\eta_k\}$ of vectors in $\mathrm{sp}\{q_{l,k}\}$ such that $\|\zeta_k - \eta_k\| \to 0$. We will demonstrate this. It is easy to see that we can, without loss of generality, assume that $\zeta_k = \zeta$ where $\zeta \in \mathrm{sp}\{\xi\}$ is a unit vector. Let $\epsilon > 0$. By (2.5.13) we can find $\tilde{\eta}_k \in \mathrm{sp}\{q_{i,k}\}_{i=1}^{l}$ such that $\|\zeta - \tilde{\eta}_k\| < \epsilon/2$ for sufficiently large k. Now, $\tilde{\eta}_k = \sum_{i=1}^{l} \alpha_{i,k} q_{i,k}$ where $\sum_{i=1}^{l} |\alpha_{i,k}|^2 = 1$. So

$$\begin{aligned} \|\zeta - \tilde{\eta}_k\|^2 &= \|\zeta - \alpha_{l,k}q_{l,k}\|^2 - 2\mathrm{Re}\langle \zeta - \alpha_{l,k}q_{l,k}, \sum_{i=1}^{l-1} \alpha_{i,k}q_{i,k}\rangle + \sum_{i=1}^{l-1} |\alpha_{i,k}|^2 \\ &= \|\zeta - \alpha_{l,k}q_{l,k}\|^2 - 2\mathrm{Re}\langle \zeta, \sum_{i=1}^{l-1} \alpha_{i,k}q_{i,k}\rangle + \sum_{i=1}^{l-1} |\alpha_{i,k}|^2. \end{aligned}$$

Now $\zeta \perp \hat{q}_i$ for $i \leq \mu$ and also $\zeta \in \mathrm{ran}\chi_\omega(A)$. These observations together with the induction hypothesis $\mathrm{sp}\{\hat{q}_{k,i}\} \to \mathrm{sp}\{\hat{q}_i\}$ for $i \leq \mu$ and the fact that, by Lemma 2.5.12, if $e_m \in \Lambda_\Omega \cup \tilde{\Lambda}_\omega$, where $m < l$ then $\chi_\omega(A)q_{k,m} \to 0$, imply that $\langle \zeta, \sum_{i=1}^{l-1} \alpha_{i,k}q_{i,k}\rangle$ becomes arbitrarily small for large k. Thus for sufficiently large k we have

$$\|\zeta - \alpha_{l,k}q_{l,k}\|^2 + \sum_{i=1}^{l-1} |\alpha_{i,k}|^2 < \epsilon^2.$$

By choosing $\eta_k = \alpha_{l,k}q_{k,l} \in \mathrm{sp}\{q_{k,l}\}$, we have proved the assertion and hence the induction hypothesis. The initial step is straightforward.

We are left with two things to prove. Firstly we demonstrate that $\mathrm{sp}\{\hat{q}_j\}_{j=1}^M = \mathrm{sp}\{\xi_j\}_{j=1}^M$. It is easily seen, from orthonormality of $\{\hat{q}_{k,i}\}_{i=1}^M$, that $\{\hat{q}_i\}_{i=1}^M$ are all orthonormal. Hence, since they are eigenvectors of $\sum_{j=1}^M \lambda_j \xi_j \otimes \bar{\xi}_j$ it follows that $\mathrm{sp}\{\hat{q}_j\}_{j=1}^M = \mathrm{sp}\{\xi_j\}_{j=1}^M = \mathrm{ran}\chi_\omega(A)$. Finally, we need to show that $e_j \notin \{\hat{e}_j\}_{j=1}^M$, then $\chi_\omega(A)Q_k e_j \to 0$, and this follows easily from Lemma 2.5.12. □

As mentioned in the beginning of this chapter, the infinite dimensional QR algorithm occurred first in the paper "Toda Flows with Infinitely Many Variables" (DLT85) by Deift, Li and Tomei. Theorem 2.5.9 is related to Theorem 1 in section 4 of (DLT85), however, the techniques used in (DLT85) deviate quite substantially from the framework used in this paper. This is natural since one considers only self-adjoint operators in (DLT85). Further connections between our results and (DLT85) are currently being investigated.

Chapter 3

The Complexity Index

The previous chapter considered self-adjoint and normal cases, while in this chapter we will be focusing on non-normal operators. In the non-normal case very little has been done and even the monumental "Spectra and Pseudospectra" by Trefethen and Embree (TE05) leaves the question on how to approximate spectra of arbitrary non-normal operators open. Obviously, special cases have been considered e.g. several types of non-self-adjoint Schrödinger operators have been investigated in (TE05) and one has been able to successfully determine their spectra and pseudospectra via approximation techniques. However, these techniques are not suited for generality.

Now returning to the main question, namely, can one determine or compute spectra of arbitrary operators, we need to be more precise regarding the mathematical meaning. Given a closed operator T on a separable Hilbert space $\mathcal{H}$ with domain $\mathcal{D}(T)$, we suppose that $\{e_j\}_{j\in\mathbb{N}}$ is a basis for $\mathcal{H}$ such that $\text{span}\{e_j\}_{j\in\mathbb{N}} \subset \mathcal{D}(T)$, and thus we can form the matrix elements $x_{ij} = \langle Te_j, e_i\rangle$. Is it possible to recover the spectrum of T through a construction only using arithmetic operations and radicals of the matrix elements? (Much more precise definitions of this will be discussed in Section 3.1.) This obviously has to be a construction that involves some limit operation, but in finite dimensions this is certainly possible. For a finite-dimensional matrix one may deduce that all its spectral information can be revealed through a construction using only arithmetic operations and radicals of the matrix elements. More precisely, for a matrix $\{a_{ij}\}_{ij\leq N}$ one can form $\{\Omega_n\}_{n\in\mathbb{N}}$, where $\Omega_n \subset \mathbb{C}$ can be constructed using only finitely many arithmetic operations and radicals of the matrix elements $\{a_{ij}\}_{ij\leq N}$, and $\Omega_n \to \sigma(\{a_{ij}\}_{ij\leq N})$ in the Hausdorff metric as $n \to \infty$. For a compact operator C we may let P_m be the projection onto $\text{span}\{e_j\}_{j\leq m}$ and observe that $\sigma(P_m C\lceil_{P_m\mathcal{H}}) \to \sigma(C)$ in the Hausdorff metric as $m \to \infty$. Thus, as we are now faced with a finite dimensional problem that we can solve (at least as sketched above), we may deduce that, yes, we can construct and determine the spectrum of a compact operator using only its matrix elements. The question is: can this be done in general?

Another issue is the following. Supposing that one is able to construct the spectrum of a class of operators as suggested above, it would be interesting to determine if such a construction would be optimal in some sense. Now, suppose that one is interested in applying such a construction in applications, such a tool for determining the optimality would be useful. It turns out that the Complexity Index is a convenient tool for determining how difficult it is to construct or approximate spectra of a certain class of operators. For selected papers related to this topic we refer to (DLT85)(DVV94)(Sha00)(Sze20) (BCN01).

3.1 Defining the Complexity Index

Recall the example from Chapter 1 that causes some headache when considering the general non-normal problem. In particular, let $A_\epsilon : l^2(\mathbb{Z}) \to l^2(\mathbb{Z})$ be defined by

$$(A_\epsilon f)(n) = \begin{cases} \epsilon f(n+1) & n = 0 \\ f(n+1) & n \neq 0. \end{cases}$$

Now for $\epsilon \neq 0$ we have $\sigma(A_\epsilon) = \{z : |z| = 1\}$ but for $\epsilon = 0$ then $\sigma(A_0) = \{z : |z| \leq 1\}$. Davies argues as follows in (Dav05): "If ϵ is a very small constructively defined real number and one is not able to determine whether or not $\epsilon = 0$, then the spectrum of A_ϵ cannot be computed even approximately even though A_ϵ is well-defined constructively. This implies that there exists straightforward bounded operators whose spectrum will probably never be determined."

A numerical analyst may express the same concern. One can argue that if one should do a computation of the spectrum on a computer, the fact that the arithmetic operations carried out are not exact may lead to the result that one gets the true solution to a slightly perturbed problem. As suggested in the previous example this could be disastrous. The problem above does not occur (in the bounded case) if we are considering the pseudospectrum.

Definition 3.1.1. *Let T be a closed operator on a Hilbert space $\mathcal{H}$ such that $\sigma(T) \neq \mathbb{C}$, and let $\epsilon > 0$. The ϵ-pseudospectrum of T is defined as the set*

$$\sigma_\epsilon(T) = \sigma(T) \cup \{z \notin \sigma(T) : \|(z - T)^{-1}\| > \epsilon^{-1}\}.$$

The reason is that the pseudospectrum varies continuously with the operator T if T is bounded (we will be more specific regarding the continuity below.) One may argue that the pseudospectrum may give a lot of information about the operator and one should therefore estimate that instead, however, we are interested in getting a complete spectral understanding of the operator and will therefore estimate both the spectrum and the pseudospectrum. We would thus like to introduce a set which has the continuity property of the pseudospectrum but approximates the spectrum. For this we introduce the n-pseudospectrum.

Definition 3.1.2. *Let T be a closed operator on a Hilbert space $\mathcal{H}$ such that $\sigma(T) \neq \mathbb{C}$, and let $n \in \mathbb{Z}_+$ and $\epsilon > 0$. The (n, ϵ)-pseudospectrum of T is defined as the set*

$$\sigma_{n,\epsilon}(T) = \sigma(T) \cup \{z \notin \sigma(T) : \|R(z, T)^{2^n}\|^{1/2^n} > \epsilon^{-1}\}.$$

As we will see in Section 3.3, the n-pseudospectrum has all the nice continuity properties that the pseudospectrum has, but it also approximates the spectrum arbitrary well for large n.

We will in this section give the precise definition of what kind of approximating constructions for the spectrum we will be using. The motivation for such definitions are discussed in Example 3.1.5.

Definition 3.1.3. *Let $\mathcal{H}$ be a Hilbert space spanned by $\{e_j\}_{j\in\mathbb{N}}$ and let*

$$\Upsilon = \{T \in \mathcal{C}(\mathcal{H}) : \operatorname{span}\{e_j\}_{n\in\mathbb{N}} \subset \mathcal{D}(T)\}. \tag{3.1.1}$$

Let $\Delta \subset \Upsilon$ and $\Xi : \Delta \to \Omega$, where Ω denotes the collection of closed subsets of $\mathbb{C}$. Let

$$\Pi_\Delta = \{\{x_{ij}\}_{i,j\in\mathbb{N}} : \exists T \in \Delta,\, x_{ij} = \langle Te_j, e_i\rangle\}.$$

A set of estimating functions of order k for Ξ is a family of functions

$$\Gamma_{n_1} : \Pi_\Delta \to \Omega, \Gamma_{n_1,n_2} : \Pi_\Delta \to \Omega, \ldots, \Gamma_{n_1,\ldots,n_{k-1}} : \Pi_\Delta \to \Omega,$$

$$\Gamma_{n_1,\ldots,n_k} : \{\{x_{ij}\}_{i,j\le N(n_1,\ldots,n_k)} : \{x_{ij}\}_{i,j\in\mathbb{N}} \in \Pi_\Delta\} \to \Omega,$$

where $N(n_1,\ldots,n_k) < \infty$ depends on $n_1,\ldots,n_k$, with the following properties:

(i) The evaluation of $\Gamma_{n_1,\ldots,n_k}(\{x_{ij}\})$ requires only finitely many arithmetic operations and radicals of the elements $\{x_{ij}\}_{i,j\le N(n_1,\ldots,n_k)}$.

(ii) Also, we have the following relation between the limits

$$\begin{aligned}
\Xi(T) &= \lim_{n_1\to\infty} \Gamma_{n_1}(\{x_{ij}\}),\\
\Gamma_{n_1}(\{x_{ij}\}) &= \lim_{n_2\to\infty} \Gamma_{n_1,n_2}(\{x_{ij}\}),\\
&\vdots\\
\Gamma_{n_1,\ldots,n_{k-1}}(\{x_{ij}\}) &= \lim_{n_k\to\infty} \Gamma_{n_1,\ldots,n_k}(\{x_{ij}\}).
\end{aligned}$$

The limit is defined as follows, for $\omega \in \Omega$ then $\omega = \lim_{n\to\infty}\omega_n$ if and only if, for any compact ball K such that $\omega \cap K^o \neq \emptyset$ we have $d_H(\omega\cap K, \omega_n \cap K) \to 0$, when $n \to \infty$.

Definition 3.1.4. *Let $\mathcal{H}$ be a Hilbert space spanned by $\{e_j\}_{j\in\mathbb{N}}$, define Υ as in (3.1.1), and let $\Delta \subset \Upsilon$. A set valued function*

$$\Xi : \Delta \subset \mathcal{C}(\mathcal{H}) \to \Omega$$

is said to have complexity index k if k is the smallest integer for which there exists a set of estimating functions of order k for Ξ. Also, Ξ is said to have infinite complexity index if no set of estimating functions exists. If there is a function

$$\Gamma : \{\{x_{ij}\} : \exists T \in \Delta,\, x_{ij} = \langle Te_j, e_i\rangle\} \to \Omega$$

such that $\Gamma(\{x_{ij}\}) = \Xi(T)$, and the evaluation of $\Gamma(\{x_{ij}\})$ requires only finitely many arithmetic operations and radicals of a finite subset of $\{x_{ij}\}$, then Ξ is said to have complexity index zero. The complexity index of a function Ξ will be denoted by $C_{\text{ind}}(\Xi)$.

Example 3.1.5. Let $\mathcal{H}$ be a Hilbert space with basis $\{e_j\}$, $\Delta = \mathcal{B}(\mathcal{H})$ and $\Xi(T) = \sigma(T)$ for $T \in \mathcal{B}(\mathcal{H})$. Suppose that $\dim(\mathcal{H}) \le 4$. Then Ξ must have complexity index zero, since one can obviously express the eigenvalues of T using finitely many arithmetic operations and radicals of the matrix elements $x_{ij} = \langle Te_j, e_i\rangle$.

For $\dim(\mathcal{H}) \ge 5$ then obviously $C_{\text{ind}}(\Xi) > 0$, by the much celebrated theory of Abel on the unsolvability of the quintic using radicals.

Now, what about compact operators? Suppose for a moment that we can show that $C_{\text{ind}}(\Xi) = 1$ if $\dim(\mathcal{H}) < \infty$. (We consider this as a problem in matrix analysis and shall not discuss it any further, nor will any of the upcoming theorems rely on such a result.)

A standard way of determining the spectrum of a compact operator T is to let P_n be the projection onto $\text{span}\{e_j\}_{j\leq n}$ and compute the spectrum of $P_nA\lceil_{P_n\mathcal{H}}$. This approach is justified since $\sigma(P_nA\lceil_{P_n\mathcal{H}}) \to \sigma(T)$ as $n \to \infty$. By the assumption on the complexity index in finite dimensions, it follows that if Δ denotes the set of compact operators then $C_{\text{ind}}(\Xi) \leq 2$.

The reasoning in the example does not say anything about what the complexity index of spectra of compact operators is, it only suggest that the standard way of approximating spectra of such operators will normally make use of a construction requiring two limits. We will in this article discuss only upper bounds on the complexity index, as we consider that the most important question to solve first, since as of today there is no general approach to estimate the spectrum of an arbitrary bounded operator. Now, after having established upper bounds, an important problem to solve would be to actually determine the complexity index of spectra of subclasses of operators. These questions are left for future work.

3.2 The Main Theorems

The main theorems in this chapter state that indeed it is possible to estimate spectra and pseudospectra of all bounded operators given the matrix elements. For the unbounded case this is also possible if one also has access to the matrix elements of the adjoint. In this case the choice of bases is not arbitrary. We would like to emphasize that even though determining spectra and pseudospectra is the mathematical goal, another set that may be of practical interest is $\omega_\delta(\sigma(T))$ (the δ-neighborhood) for $T \in \mathcal{C}(\mathcal{H})$ and $\delta > 0$. The reason is that $\sigma(T)$ may contain parts that have Lebesgue measure zero, and therefore may be quite hard to detect. An easier alternative may then be $\omega_\delta(\sigma(T))$, although mathematically this set reveals less information about the operator.

Definition 3.2.1. *Let $\{e_n\}_{n\in\mathbb{N}}$ be a basis for the Hilbert space $\mathcal{H}$. By a weighted shift on $\mathcal{H}$ we mean an operator $W \in \mathcal{C}(\mathcal{H})$ with $\mathcal{D}(W) \supset \text{span}\{e_n\}_{n\in\mathbb{N}}$ with the property that there is a sequence of complex numbers $\{\alpha_j\}_{j\in\mathbb{N}}$ and an integer k such that for $\xi \in \mathcal{D}(W)$ we have $(W\xi)_j = \alpha_j\xi_{k+j}$. The set of weighted shifts on $\mathcal{H}$ (with respect to $\{e_n\}_{n\in\mathbb{N}}$) will be denoted by $WS(\mathcal{H})$.*

Theorem 3.2.2. *Let $\{e_j\}_{j\in\mathbb{N}}$ be a basis for the Hilbert space $\mathcal{H}$ and let*

$$\Delta = \{T \in \mathcal{C}(\mathcal{H}) : T = W + A,\ W \in WS(\mathcal{H}),\ A \in \mathcal{B}(\mathcal{H}) \\ \cap \{T \in \mathcal{C}(\mathcal{H}) : \|R(T,\cdot)^{2^n}\|^{1/2^n} \text{ is never constant for any } n\}.$$

Define, for $n \in \mathbb{Z}_+, \epsilon > 0$, the set valued functions $\Xi_1, \Xi_2, \Xi_3 : \Delta \to \Omega$ is defined by $\Xi_1(T) = \overline{\sigma_{n,\epsilon}(T)}$, $\Xi_2(T) = \overline{\omega_\epsilon(\sigma(T))}$ and $\Xi_3(T) = \sigma(T)$. Then

$$C_{ind}(\Xi_1) \leq 3, \qquad C_{ind}(\Xi_2) \leq 4, \qquad C_{ind}(\Xi_3) \leq 4.$$

Also, if $\Delta = \mathcal{B}(\mathcal{H})$ then $C_{ind}(\Xi_1) \leq 2$, $C_{ind}(\Xi_2) \leq 3$ and $C_{ind}(\Xi_3) \leq 3$.

Theorem 3.2.3. *Let $\{e_j\}_{j\in\mathbb{N}}$ be a basis for the Hilbert space $\mathcal{H}$, P_m be the projection onto $\text{span}\{e_j\}_{j=1}^m$ and d be some positive integer. Let $\Delta \subset \mathcal{C}(\mathcal{H})$ have the following properties: For $T \in \Delta$ we have*

(i) $\bigcup_m P_m\mathcal{H} \subset \mathcal{D}(T)$, $\bigcup_m P_m\mathcal{H} \subset \mathcal{D}(T^*)$.

(ii) $\langle Te_{j+l}, e_j\rangle = \langle Te_j, e_{j+l}\rangle = 0$, *for* $l > d$.

(iii) $TP_m\xi \to T\xi$, $T^*P_m\eta \to T^*\eta$, *as* $m \to \infty$ *for* $\xi \in \mathcal{D}(T)$ *and* $\eta \in \mathcal{D}(T^*)$.

Let $\epsilon > 0$ *and* $n \in \mathbb{Z}_+$ *and* $\Xi_1, \Xi_2, \Xi_3 : \Delta \to \Omega$ *be defined by* $\Xi_1(T) = \overline{\sigma_{n,\epsilon}(T)}$, $\Xi_2(T) = \overline{\omega_\epsilon(\sigma(T))}$ *and* $\Xi_3(T) = \sigma(T)$. *Then*

$$C_{ind}(\Xi_1) = 1, \qquad C_{ind}(\Xi_2) \leq 2, \qquad C_{ind}(\Xi_3) \leq 2.$$

Theorem 3.2.4. *Let* $\{e_j\}_{j\in\mathbb{N}}$ *and* $\{\tilde{e}_j\}_{j\in\mathbb{N}}$ *be bases for the Hilbert space* $\mathcal{H}$ *and let*

$$\tilde{\Delta} = \{T \in \mathcal{C}(\mathcal{H}\oplus\mathcal{H}) : T = T_1 \oplus T_2, T_1, T_2 \in \mathcal{C}(\mathcal{H}), T_1^* = T_2\}$$
$$\Delta = \{T \in \tilde{\Delta} : \operatorname{span}\{e_j\}_{j\in\mathbb{N}} \textit{ is a core for } T_1, \operatorname{span}\{\tilde{e}_j\} \textit{ is a core for } T_2\}.$$

Let $\epsilon > 0$, $\Xi_1 : \Delta \to \Omega$ *and* $\Xi_2 : \Delta \to \Omega$ *be defined by* $\Xi_1(T) = \overline{\sigma_\epsilon(T_1)}$ *and* $\Xi_2(T) = \sigma(T_1)$. *Then* $C_{\mathrm{ind}}(\Xi_1) \leq 2$ *and* $C_{\mathrm{ind}}(\Xi_2) \leq 3$.

Corollary 3.2.5. *Let* $\{e_j\}_{j\in\mathbb{N}}$ *be a basis for the Hilbert space* $\mathcal{H}$ *and let*

$$\Delta = \{A \in \mathcal{SA}(\mathcal{H}) : \operatorname{span}\{e_j\}_{j\in\mathbb{N}} \textit{ is a core for } A\}.$$

Let $\epsilon > 0$ *and* $\Xi_1, \Xi_2 : \Delta \to \Omega$ *be defined by* $\Xi_1(T) = \sigma(T)$ *and* $\Xi_2(T) = \overline{\omega_\epsilon(\sigma(T))}$. *Then* $C_{\mathrm{ind}}(\Xi_1) \leq 3$ *and* $C_{\mathrm{ind}}(\Xi_2) \leq 2$.

Remark 3.2.6. What Theorem 3.2.4 essentially says is that given the matrix elements of the operator and its adjoint, where the matrix elements come from a reasonable choice of bases, one can estimate the pseudospectra and the spectrum. Also, estimating the pseudospectrum of an unbounded operator is on the same level of difficulty as estimating the spectrum of a compact operator.

3.3 Properties of the n-pseudospectra of Bounded Operators

We will prove some of the properties of the n-pseudospectrum, but before doing that we need a couple of propositions and theorems that will come in handy.

Proposition 3.3.1. *Let* $\gamma : \mathbb{C} \to [0,\infty)$ *be continuous and let* $\{\gamma_k\}_{k\in\mathbb{N}}$ *be a sequence of functions such that* $\gamma_k : \mathbb{C} \to [0,\infty)$ *and* $\gamma_k \to \gamma$ *locally uniformly. Suppose that one of the two following properties are satisfied.*

(i) $\gamma_k \to \gamma$ *monotonically from above.*

(ii) For $\epsilon > 0$, *then* $\mathrm{cl}(\{z : \gamma(z) < \epsilon\}) = \{z : \gamma(z) \leq \epsilon\}$.

Then for any compact ball K *such that* $\{z : \gamma(z) < \epsilon\} \cap K^o \neq \emptyset$ *it follows that*

$$\mathrm{cl}(\{z : \gamma_k(z) < \epsilon\}) \cap K \longrightarrow \mathrm{cl}(\{z : \gamma(z) < \epsilon\}) \cap K, \qquad k \to \infty.$$

Proof. Let $\epsilon > 0$. We first claim that, in each case, for any $\nu > 0$ there exists an $\alpha > 0$ such that

$$\omega_\nu(\mathrm{cl}(\{z : \gamma(z) < \epsilon - \alpha\}) \cap K) \supset \mathrm{cl}(\{z : \gamma(z) < \epsilon\}) \cap K. \tag{3.3.1}$$

Arguing by contradiction and supposing the latter statement is false we deduce that there must be a sequence $\{\zeta_\alpha\} \subset \mathrm{cl}(\{z : \gamma(z) < \epsilon\}) \cap K$ such that $\zeta_\alpha \notin \omega_\nu(\mathrm{cl}(\{z : \gamma(z) < \epsilon - \alpha\}) \cap K)$. By compactness, we may assume without loss of generality that $\zeta_\alpha \to \zeta$ as $\alpha \to 0$. By continuity we have that $\gamma(\zeta_\alpha) \to \gamma(\zeta)$ and since $\zeta_\alpha \notin \omega_\nu(\mathrm{cl}(\{z : \gamma(z) < \epsilon - \alpha\}) \cap K)$ it follows that $\gamma(\zeta) = \epsilon$. Note that we must have

$$\zeta \in \bigcap_{\alpha > 0} \mathbb{C} \setminus \omega_\nu(\mathrm{cl}(\{z : \gamma(z) < \epsilon - \alpha\}) \cap K). \tag{3.3.2}$$

But there is a $\xi \in \{z : \gamma(z) < \epsilon\} \cap K$ such that $|\xi - \zeta| < \nu$. Now let $\alpha_1 = \gamma(\zeta) - \gamma(\xi)$. Then $\gamma(\xi) = \epsilon - \alpha_1$ and hence $\zeta \in \omega_\nu(\{z : \gamma(z) < \epsilon - \alpha_2\})$, for some $\alpha_2 < \alpha_1$ contradicting (3.3.2). We are now ready to prove the proposition, which will follow if we can show that for any $\nu > 0$ we have

$$\mathrm{cl}(\{z : \gamma(z) < \epsilon\}) \cap K \subset \omega_\nu(\mathrm{cl}(\{z : \gamma_k(z) < \epsilon\}) \cap K)$$

and $\omega_\nu(\mathrm{cl}(\{z : \gamma(z) < \epsilon\}) \cap K) \supset \mathrm{cl}(\{z : \gamma_k(z) < \epsilon\}) \cap K$, for all sufficiently large k.

Note that the first inclusion follows by using the claim in the first part of the proof and the locally uniform convergence of γ_k. Indeed, by the locally uniform convergence it follows that, for any $\alpha > 0$, we have

$$\mathrm{cl}(\{z : \gamma_k(z) < \epsilon\}) \cap K \supset \mathrm{cl}(\{z : \gamma(z) < \epsilon - \alpha\}) \cap K$$

for large k, thus by appealing to (3.3.1), we obtain the desired inclusion. To see the second inclusion, we first assume (i). Then $\{z : \gamma_k(z) < \epsilon\} \subset \{z : \gamma(z) < \epsilon\}$ and hence the inclusion follows. As for the second case we assume (ii). By arguing by contradiction, we suppose the statement is false and deduce that there is a sequence $\{z_k\}$ such that

$$z_k \in \mathrm{cl}(\{z : \gamma_k(z) < \epsilon\}) \cap K$$

and $z_k \notin \omega_\nu(\mathrm{cl}(\{z : \gamma(z) < \epsilon\}) \cap K)$. By compactness we may assume that $z_k \to z$ and then (by (ii)) $\gamma(z) > \epsilon$ which contradicts the fact that $\gamma_k(z_k) \to \gamma(z)$ which follows by continuity of γ and the local uniform convergence of $\{\gamma_k\}$. □

Theorem 3.3.2. *(Sha08) Let Ω be an open subset of $\mathbb{C}$, X be a Banach space and Y be a uniformly convex Banach space. Suppose $A : \Omega \to \mathcal{B}(X, Y)$ is an analytic operator valued function such that $A'(z)$ is invertible for all $z \in \Omega$. If $\|A(z)\| \leq M$ for all $z \in \Omega$ then $\|A(z)\| < M$ for all $z \in \Omega$.*

Before we continue let us define some functions that will be crucial throughout the paper.

Definition 3.3.3. *Let $\{P_m\}$ be an increasing sequence of projections converging strongly to the identity. Define, for $n \in \mathbb{Z}_+$ and $m \in \mathbb{N}$, the function $\Phi_{n,m} : \mathcal{B}(\mathcal{H}) \times \mathbb{C} \to \mathbb{R}$ by*

$$\Phi_{n,m}(S, z) = \min \left\{ \lambda^{1/2^{n+1}} : \lambda \in \sigma \left(P_m((S - z)^*)^{2^n} (S - z)^{2^n} \Big\lceil_{P_m \mathcal{H}} \right) \right\}.$$

Define also

$$\Phi_n(S,z) = \lim_{m\to\infty} \Phi_{n,m}(S,z),$$

and for $T \in \mathcal{B}(\mathcal{H})$

$$\gamma_n(z) = \min[\Phi_n(T,z), \Phi_n(T^*,\bar{z})]. \tag{3.3.3}$$

Theorem 3.3.4. *Let* $T \in \mathcal{B}(\mathcal{H})$, γ_n *be defined as in (3.3.3) and* $\epsilon > 0$. *Then the following is true*

(i) $\sigma_{n+1,\epsilon}(T) \subset \sigma_{n,\epsilon}(T)$.

(ii) $\sigma_{n,\epsilon}(T) = \{z \in \mathbb{C} : \gamma_n(z) < \epsilon\}$.

(iii) $\mathrm{cl}(\{z : \gamma_n(z) < \epsilon\}) = \{z : \gamma_n(z) \leq \epsilon\}$.

(iv) Let $\omega_\epsilon(\sigma(T))$ *denote the* ϵ*-neighborhood around* $\sigma(T)$. *Then*

$$d_H(\overline{\sigma_{n,\epsilon}(T)}, \overline{\omega_\epsilon(\sigma(T))}) \longrightarrow 0, \qquad n \to \infty.$$

(v) If $\{T_k\} \subset \mathcal{B}(\mathcal{H})$ *and* $T_k \to T$ *in norm, it follows that*

$$d_H(\overline{\sigma_{n,\epsilon}(T_k)}, \overline{\sigma_{n,\epsilon}(T)}) \longrightarrow 0, \qquad k \to \infty.$$

Proof. Now (i) follows by the definition of $\sigma_{n,\epsilon}(T)$ and the fact that

$$\begin{aligned} 1/\|R(z,T)^{2^{n+1}}\|^{1/2^{n+1}} &\geq 1/(\|R(z,T)^{2^n}\|^{1/2^{n+1}}\|R(z,T)^{2^n}\|^{1/2^{n+1}}) \\ &= 1/\|R(z,T)^{2^n}\|^{1/2^n}. \end{aligned} \tag{3.3.4}$$

To prove (ii) we have to show that $\gamma_n(z) = 1/\|R(z,T)^{2^n}\|^{1/2^n}$ when $z \notin \sigma(T)$ and that $\gamma_n(z) = 0$ when $z \in \sigma(T)$. The former is clear, so to see the latter we need to show that when $z \in \sigma(T)$ then either $|(T-z)^{2^n}|$ or $|((T-z)^{2^n})^*|$ is not invertible. To see that, we need to consider three cases: (1) $(T-z)^{2^n}$ is not one to one, (2) $(T-z)^{2^n}$ is not onto, but the range of $(T-z)^{2^n}$ is dense in $\mathcal{H}$ or (3) $(T-z)^{2^n}$ is not onto and $\overline{\mathrm{ran}((T-z)^{2^n}} \neq \mathcal{H}$.

Case (1): Now, by the polar decomposition, we have $(T-z)^{2^n} = U|(T-z)^{2^n}|$ where U is a partial isometry, and it is easy to see that $|(T-z)^{2^n}|$ is not invertible when $(T-z)^{2^n}$ is not one to one.

Case (2): Recall that U is unitary if and only if $((T-z)^{2^n})^*$ is one to one. Thus, since $\overline{\mathrm{ran}((T-z)^{2^n})} = \mathcal{H}$ and $\ker(((T-z)^{2^n})^*) = \mathrm{ran}((T-z)^{2^n})^\perp$, we have that U must be unitary. But that implies that $|(T-z)^{2^n}|$ cannot be invertible since $(T-z)^{2^n}$ is not invertible.

Case (3): If $\overline{\mathrm{ran}((T-z)^{2^n}} \neq \mathcal{H}$ it follows that $\ker(((T-z)^{2^n})^*)$ is nonzero, and since

$$((T-z)^{2^n})^* = U^*|((T-z)^{2^n})^*|$$

we may argue as in Case (1) to deduce that $|((T-z)^{2^n})^*|$ is not invertible and this proves the claim.

To see (iii) we argue by contradiction and assume that $\mathrm{cl}(\{z : \gamma_n(z) < \epsilon\}) = \{z : \gamma_n(z) \leq \epsilon\}$. is false. Then there exists a $\tilde{z} \in \sigma(T)^c$ such that $\gamma_n(\tilde{z}) = \epsilon$ and also a neighborhood θ around $\tilde{z}$ such that $\gamma_n(z) \geq \epsilon$ for $z \in \theta$. Now, for $z \in \theta$, it follows that

$1/\gamma_n(z) = \|R(z,T)^{2^n}\|^{1/2^n}$. Thus, $\|R(\tilde{z},T)^{2^n}\| = 1/\epsilon^{2^n}$ and $\|R(\tilde{z},T)^{2^n}\| \leq 1/\epsilon^{2^n}$ for $z \in \theta$. But $z \mapsto R(z,T)^{2^n}$ is obviously holomorphic and $\frac{d}{dz}R(z,T)^{2^n}$ is easily seen to be invertible for all $z \in \theta$. Thus, by Theorem 3.3.2, it follows that $\|R(\tilde{z},T)^{2^n}\| < 1/\epsilon^{2^n}$ for all $z \in \theta$, contradicting $\|R(\tilde{z},T)^{2^n}\| = 1/\epsilon^{2^n}$.

It is easy to see that to prove (iv) it suffices to show that $\gamma_n \to \gamma$ locally uniformly, where

$$\gamma(z) = \text{dist}(z, \sigma(T)).$$

To see the latter, let $\delta > 0$ and let ω_δ denote the open δ-neighborhood around $\sigma(T)$. Let also Ω be a compact set such $\sigma(T) \subset \Omega^o$ and $\Omega_\delta = \Omega \setminus \omega_\delta$. Note that for $z \in \Omega \setminus \sigma(T)$ we have

$$\gamma(z) = 1/\rho(R(z,T)),$$

where $\rho(R(z,T))$ denotes the spectral radius of $R(z,T)$, and also by (3.3.4) it follows that $\gamma_{n+1}(z) \geq \gamma_n(z)$. Thus, by the continuity of γ and γ_n together with the spectral radius formula we may appeal to Dini's Theorem to deduce that $\gamma_n \to \gamma$ locally uniformly on Ω_δ. By choosing n large enough we can guarantee that $|\gamma_n(z) - \gamma(z)| \leq \delta$ when $z \in \Omega_\delta$. Also, since $\gamma_n(z) \leq \gamma(z)$ for $z \in \Omega \setminus \sigma(T)$ and $\gamma(z) = \text{dist}(z, \sigma(T)) \leq \delta$ for $z \in \omega_\delta$ we have that $|\gamma_n(z) - \text{dist}(z, \sigma(T))| \leq \delta$ when $z \in \Omega \setminus \sigma(T)$ Since, by (ii), it is true that $\gamma_n(z) = \text{dist}(z, \sigma(T)) = \gamma(z) = 0$ when $z \in \sigma(T)$ we are done with (iv).

To see that (v) is true let $\gamma_{n,k}(z) = \min[\Phi_n(T_k, z), \Phi_n(T_k^*, \bar{z})]$. Then, by (ii), $\sigma_{n,\epsilon}(T_k) = \{z \in \mathbb{C} : \gamma_{n,k}(z) < \epsilon\}$. Also, since T is bounded and $T_k \to T$ in norm, there is a compact set $K \subset \mathbb{C}$ containing both $\sigma_{n,\epsilon}(T)$ and $\sigma_{n,\epsilon}(T_k)$. Thus, by appealing to (iii) and Proposition 3.3.1 we conclude that to prove (v) we only need to show that $\gamma_{n,k} \to \gamma_n$ locally uniformly as $k \to$. It suffices to show that $\gamma_{n,k}^{2^{n+1}} \to \gamma_n^{2^{n+1}}$ locally uniformly. Now

$$\begin{aligned} &|\Phi_n(T_k, z)^{2^{n+1}} - \Phi_n(T, z)^{2^{n+1}}| \\ &\qquad \leq d_H\left(\sigma\left(((T_k - z)^*)^{2^n}(T_k - z)^{2^n}\right), \sigma\left(((T - z)^*)^{2^n}(T - z)^{2^n}\right)\right) \qquad (3.3.5) \\ &\qquad \leq \|((T_k - z)^*)^{2^n}(T_k - z)^{2^n} - ((T - z)^*)^{2^n}(T - z)^{2^n}\| \longrightarrow 0, \end{aligned}$$

locally uniformly as $k \to \infty$. Similar estimate holds for $|\Phi_n(T_k^*, \bar{z})^{2^{n+1}} - \Phi_n(T^*, \bar{z})^{2^{n+1}}|$ and this yields the assertion. □

Remark 3.3.5. The advantage of the (n, ϵ)-pseudospectrum is that in addition to the continuity property stated above, we now have two parameters n and ϵ to tweak in order to estimate the spectrum. It is quite easy to construct examples (even 2-by-2 matrices) of operators $\{T_n\}$ for which $\sigma_{1,\epsilon}(T_n) \subset \sigma_{\epsilon/10^n}(T_n)$. And of course, in the self-adjoint case it would not make sense to take $n > 0$ as $\sigma_{n,\epsilon}(A) = \sigma_\epsilon(A)$ for self-adjoint A.

3.4 Properties of the n-pseudospectra of Unbounded Operators

The theory of n-pseudospectra for unbounded operators has a lot in common with the theory of n-pseudospectra for bounded operators, however, there is a major difference; the n-pseudospectrum of an unbounded operator can "jump". We will be more specific about this below.

Theorem 3.4.1. *Let $T \in \mathcal{C}(\mathcal{H})$, $n \in \mathbb{Z}_+$, $\epsilon > 0$ and let $K \subset \mathbb{C}$ be a compact ball such that $\sigma_\epsilon(T) \cap K^o \neq \emptyset$. Then the following is true*

(i) $\sigma_{n+1,\epsilon}(T) \subset \sigma_{n,\epsilon}(T)$.

(ii) Let $\omega_\epsilon(\sigma(T))$ denote the ϵ neighborhood around $\sigma(T)$. Then

$$d_H(\overline{\sigma_{n,\epsilon}(T)} \cap K, \overline{\omega_\epsilon(\sigma(T))} \cap K) \longrightarrow 0, \quad n \to \infty.$$

Proof. Follows by almost identical arguments as in the proof of Theorem 3.3.4. □

The difference between the bounded and the unbounded case is that if $T \in \mathcal{C}(\mathcal{H})$, $z \in \mathbb{C}$ and we define

$$\gamma_n(z) = \begin{cases} 0 & z \in \sigma(T) \\ \frac{1}{\|R(z,T)^{2^n}\|^{1/2^n}} & z \in \sigma(T)^c, \end{cases} \tag{3.4.1}$$

then we might have that $\mathrm{cl}(\{z : \gamma_n(z) < \epsilon\}) \neq \{z : \gamma_n(z) \leq \epsilon\}$. The reason is that there exists unbounded operators for where the norm of the resolvent is constant on an open set in $\mathbb{C}$ (Sha08). However, we have the following.

Theorem 3.4.2. *Let $T \in \mathcal{C}(\mathcal{H})$ and let γ_n be defined as in (3.4.1). Suppose that $\|R(\cdot,T)^{2^n}\|$ can never be constant on an open set, then* $\mathrm{cl}(\{z : \gamma_n(z) < \epsilon\}) = \{z : \gamma_n(z) \leq \epsilon\}$.

Proof. Follows by arguing similar to the argument in the proof of Theorem 3.3.4 (iii). □

Theorem 3.4.3. *Let $T \in \mathcal{C}(\mathcal{H})$ with domain $\mathcal{D}(T)$ and let $\{T_k\} \subset \mathcal{C}(\mathcal{H})$ be a sequence such that $T_k \xrightarrow{\hat{\delta}} T$. Define, for $z \in \mathbb{C}$*

$$\zeta(z) = \begin{cases} 0 & z \in \sigma(T) \\ \frac{1}{\|R(z,T)\|} & z \in \sigma(T)^c, \end{cases} \quad \zeta_k(z) = \begin{cases} 0 & z \in \sigma(T_k) \\ \frac{1}{\|R(z,T_k)\|} & z \in \sigma(T_k)^c, \end{cases} \quad k \in \mathbb{N}.$$

(i) If $z \in K$, where K is compact, it follows that there is a $C_K > 0$ depending on K such that

$$|\zeta(z)^2 - \zeta_k(z)^2| \leq C_K(1 + |z|^2)\hat{\delta}(T_k, T)$$

for sufficiently large k.

(ii) Suppose that $\|R(\cdot,T)^{2^n}\|$ can never be constant on an open set. Then if $K \subset \mathbb{C}$ is a compact ball such that $K^o \cap \sigma_{n,\epsilon}(T) \neq \emptyset$, then

$$d_H(\overline{\sigma_{n,\epsilon}(T_k)} \cap K, \overline{\sigma_{n,\epsilon}(T)} \cap K) \longrightarrow 0, \quad k \to \infty, \quad \epsilon > 0.$$

Proof. To show (i) we first claim that

$$\begin{aligned} \zeta(z) = \min\big[&\inf\{\sqrt{\lambda} : \lambda \in \sigma((T-z)^*(T-z))\}, \\ &\inf\{\sqrt{\lambda} : \lambda \in \sigma((T-z)(T-z)^*)\}\big] \\ \zeta_k(z) = \min\big[&\inf\{\sqrt{\lambda} : \lambda \in \sigma((T_k-z)^*(T_k-z))\}, \\ &\inf\{\sqrt{\lambda} : \lambda \in \sigma((T_k-z)(T_k-z)^*)\}\big]. \end{aligned} \tag{3.4.2}$$

We will show this for ζ, and the argument is identical for ζ_k. Indeed, for $z \notin \sigma(T)$ this is quite straightforward and hence we are left to show that either $|T-z|$ or $|(T-z)^*|$ is not invertible for $z \in \sigma(T)$. This is essentially the same argument as in Theorem 3.3.4, but we include it for completeness and to make sure that the same conclusions can be drawn using the polar decomposition of unbounded operators. We need to consider three cases. (1), $(T-z)$ is not one to one, (2), $(T-z)$ is not onto, but the range of $(T-z)$ is dense in $\mathcal{H}$ or (3), $(T-z)$ is not onto and $\overline{\operatorname{ran}((T-z))} \neq \mathcal{H}$.

Case (1): Now, by the polar decomposition, we have $(T-z) = U|(T-z)|$ where U is a partial isometry. Note that $\ker(T-z) = \ker(|T-z|)$ and $|T-z|$ is not invertible.

Case (2): Note that $(T-z)^*$ is one to one if and only if U is unitary and so U must be unitary since $\overline{\operatorname{ran}((T-z)} = \mathcal{H}$ and $\ker((T-z)^*) = \operatorname{ran}(T-z)^\perp$. But that implies that $|(T-z)|$ cannot be invertible since $(T-z)$ is not invertible.

Case (3): If $\overline{\operatorname{ran}((T-z)} \neq \mathcal{H}$ it follows that $\ker((T-z)^*)$ is nonzero, and since $(T-z)^* = U^*|(T-z)^*|$ we may argue as in Case (1) to deduce that $|(T-z)^*|$ is not invertible, and thus we have shown (3.4.2).

Note that by the spectral mapping theorem we have that

$$\sigma((T-z)^*(T-z)) = \psi(\sigma(R_{(T-z)}), \quad \sigma((T-z)(T-z)^*) = \psi(\sigma(R_{(T-z)^*}))$$

where $\psi(x) = 1/x - 1$ (recall that $R_{(T-z)}$ is short for $(1 + (T-z)^*(T-z))^{-1}$). Now let $\zeta^2(z) = \zeta(z)^2$ and $\zeta_k^2(z) = \zeta_k(z)^2$. Then it follows that

$$\begin{aligned}\zeta^2(z) &= \min\left(\inf\{\psi(\lambda) : \lambda \in \sigma(R_{(T-z)})\}, \inf\{\psi(\lambda) : \lambda \in \sigma(R_{(T-z)^*})\}\right) \\ &= \min\left(\psi(\|R_{(T-z)}\|), \psi(\|R_{(T-z)^*}\|)\right),\end{aligned}$$

by self-adjointness of $(T-z)^*(T-z)$ and $(T-z)(T-z)^*$. Similarly,

$$\zeta_k^2(z) = \min\left(\{\psi(\|R_{(T_k-z)}\|), \psi(\|R_{(T_k-z)^*}\|)\right).$$

Recall from the definition of p and Theorem 1.1.2 that for $z \in \mathbb{C}$ we have

$$\begin{aligned}\|R_{T_k-z} - R_{T-z}\|^2 + \|R_{(T_k-z)^*} - R_{(T-z)^*}\|^2 &\leq p(T_k - z, T-z)^2 \\ &\leq 8\hat{\delta}(T_k - z, T-z)^2 \\ &\leq 24(1+|z|^2)^2\hat{\delta}(T_k, T)^2.\end{aligned} \tag{3.4.3}$$

Also, since K is compact, there is a $\delta > 0$ such that

$$0 \notin \Omega = \omega_\delta(\{\psi^{-1} \circ \zeta^2(z) : z \ \in K\}),$$

where $\omega_\delta(\{\psi^{-1} \circ \zeta^2(z) : z \in K\})$ denotes the δ-neighborhood around $\{\psi^{-1} \circ \zeta^2(z) : z \in K\}$, and by (3.4.3) it follows that

$$\{\psi^{-1} \circ \zeta_k^2(z) : z \in K\} \subset \omega_\delta(\{\psi^{-1} \circ \zeta^2(z) : z \in K\})$$

for sufficiently large k. Let C be the Lipschitz constant of $\psi\lceil_\Omega$. Then if $z \in \mathbb{C} \setminus \sigma(T)$ we have that $\psi(\|R_{(T-z)}\|) = \psi(\|R_{(T-z)^*}\|)$ so by (3.4.3)

$$|\zeta(z)^2 - \tilde{\zeta}_k^2(z)| \leq C\sqrt{24}(1+|z|^2)\hat{\delta}(T_k, T). \tag{3.4.4}$$

If $z \in \sigma(T)$ then at least one of $\|R_{(T-z)}\|$ and $\|R_{(T-z)^*}\|$ is equal to one. Now, suppose that $\|R_{(T-z)}\| = 1$. If $\zeta_k^2(z) = \psi(\|R_{(T_k-z)}\|)$ then (3.4.4) follows, so suppose that $\zeta_k^2(z) = \psi(\|R_{(T_k-z)^*}\|)$ then $\|R_{(T_k-z)^*}\| > \|R_{(T_k-z)}\|$ so

$$|\zeta(z)^2 - \zeta_k^2(z)| \leq C(1 - \|R_{(T_k-z)^*}\|) \leq C(\|R_{(T-z)}\| - \|R_{(T_k-z)}\|),$$

and hence (3.4.4) follows by (3.4.3). Similar reasoning gives the same result for

$$\zeta_k^2(z) = \psi(\|R_{(T_k-z)^*}\|)$$

and $\|R_{(T-z)^*}\| = 1$ and we deduce that (3.4.4) holds for all $z \in K$.

To show that

$$d_H(\sigma_{n,\epsilon}(T_k) \cap K, \sigma_{n,\epsilon}(T) \cap K) \longrightarrow 0, \quad k \to \infty, \quad \epsilon > 0$$

in order to deduce (ii), we will deviate substantially from the techniques used in the proof of Theorem 3.3.4 (v). Before getting to the argument note that, since for any $z_0 \in \mathcal{C}$ we have

$$\sigma_{n,\epsilon}(T + z_0) = \{z + z_0 : z \in \sigma_{n,\epsilon}(T)\},$$

we may assume that T is invertible. For $m \in \mathbb{N}$ consider the operator T^m defined inductively on

$$\mathcal{D}(T^m) = \{\xi : \xi \in \mathcal{D}(T^{m-1}), T^{m-1}\xi \in \mathcal{D}(T)\},$$

by $T^m\xi = T(T^{m-1}\xi)$. Then T^m is a closed operator (DS88). Also, since T is invertible and T is densely defined, T^{-1} has dense range and so has T^{-m} which yields that T^m is densely defined. Note also that since $\mathcal{D}(T^m) \subset \mathcal{D}(T^{m-1})$ it follows that $p(T)$ is closed and densely defined for any polynomial p and $\mathcal{D}(p(T)) = \mathcal{D}(T^d)$ where d is the degree of the polynomial p. Thus for any $z \in \mathbb{C}$ we can define the adjoint $((T-z)^m)^*$. We can now continue with the argument. The reasoning above allows us to define

$$\begin{aligned}\gamma_{n,k}(z) = \min\big[&\inf\{\lambda^{1/2^n} : \lambda \in \sigma(|(T_k - z)^{2^n}|)\}, \\ &\inf\{\lambda^{1/2^n} : \lambda \in \sigma(|((T_k - z)^*)^{2^n}|)\}\big].\end{aligned}$$

Appealing to Proposition 3.3.1 and Theorem 3.4.2 (and recalling the assumption in (ii)), it suffices to show that $\gamma_{n,k} \to \gamma_n$ locally uniformly, where

$$\gamma_n(z) = \begin{cases} 0 & z \in \sigma(T) \\ \frac{1}{\|R(z,T)^{2^n}\|^{1/2^n}} & z \in \sigma(T)^c. \end{cases}$$

ClaimI: We claim that $\gamma_{n,k} \to \gamma_n$ locally uniformly on $\sigma(T)$. To see that, note that for $z \in \sigma(T)$ then, by the spectral mapping theorem for polynomials of unbounded operators (DS88), $(T-z)^{2^n}$ is not invertible. Hence, by reasoning similar to what we did in the proof of (i), either

$$\begin{aligned}\Big(\inf_{\|\xi\|=1,\, \xi \in \mathcal{D}(T^{2^n})} \|(T-z)^{2^n}\xi\|\Big)^{1/2^n} = 0, \\ \text{or} \quad \Big(\inf_{\|\xi\|=1,\, \xi \in \mathcal{D}((T^{2^n})^*)} \|((T-z)^{2^n})^*\xi\|\Big)^{1/2^n} = 0,\end{aligned} \tag{3.4.5}$$

(or both are equal to zero). Suppose that the first part of (3.4.5) is true. Then, for $\delta > 0$, we can find for any $z_0 \in \sigma(T) \cap K$ a vector $\xi_{z_0} \in \mathcal{D}(T^{2^n})$ such that $\|(T - z_0)^{2^n} \xi_{z_0}\|^{1/2^n} \leq \delta/3$. Recall that, for any $m \in \mathbb{N}$ we have $\hat{\delta}(T_k^m, T^m) = \hat{\delta}(T_k^{-m}, T^{-m})$ and that $R(T_k^m) \to R(T^m)$ if and only if $\hat{\delta}(T_k^m, T^m) \to 0$, and since $R(T_k) \to R(T)$ so $R(T_k)^m \to R(T)^m$ we get that $\hat{\delta}(T_k^m, T^m) \to 0$. Hence, by the definition of $\hat{\delta}$, it follows that

$$\sup_{\substack{\xi \in \mathcal{D}(T^m) \\ \|\xi\| + \|T^m \xi\| = 1}} \inf_{\eta \in \mathcal{D}(T_k^m)} \|\xi - \eta\| + \|T^m \xi - T_k^m \eta\| \longrightarrow 0, \quad k \to \infty.$$

Thus, there exists a sequence of unit vectors $\{\eta_{z_0,k}\}$ in $\mathcal{D}(T_k^m)$ such that $\eta_{z_0,k} \to \xi_{z_0}$ and $T_k^m \eta_{z_0,k} \to T^m \xi_{z_0}$ as $k \to \infty$. Now, since for any integer r we have $T_k^{-r} \to T^{-r}$ in norm, it follows that

$$T_k^l \eta_{z_0,k} = T_k^{-(m-l)} T_k^m \eta_{z_0,k} \longrightarrow T^{-(m-l)} T^m \xi_{z_0} = T^l \xi_{z_0}, \qquad k \to \infty.$$

for all $l \leq m$. In particular, it is true that $z \mapsto (T_k - z)^{2^n} \eta_{z_0,k} \to z \mapsto (T - z)^{2^n} \xi_{z_0}$ locally uniformly as $k \to \infty$. Note that $z \mapsto \|(T - z)^{2^n} \xi_{z_0}\|$ is continuous. Thus, there is a neighborhood Θ_{z_0} around z_0 such that $\|(T - z)^{2^n} \xi_{z_0}\| \leq \frac{2}{3}\delta$ for $z \in \Theta_{z_0}$ and hence $\|(T_k - z)^{2^n} \eta_{z_0,k}\| \leq \delta$ for $z \in \Theta_{z_0}$ and sufficiently large k. Covering $\sigma(T) \cap K$ with finitely many neighborhoods $\{\Theta_{z_j}\}_{j=1}^M$, of the type just described, for some $\{z_j\}_{j=1}^M \subset \sigma(T) \cap K$ and some $M \in \mathbb{N}$, we deduce that there are sequences $\{\eta_{z_j,k}\}$ and an integer k_0 such that

$$\max_{j \leq M} \sup_{z \in \Theta_{z_j}} \|(T_k - z)^{2^n} \eta_{z_j,k}\| \leq \delta^{2^n}, \qquad k \geq k_0.$$

And hence it follows that that for $z \in \bigcup_{z_j} \Theta_{z_j}$

$$\inf\{\lambda^{1/2^n} : \lambda \in \sigma(|(T_k - z)^{2^n}|)\} = (\inf_{\|\xi\|=1,\, \xi \in \mathcal{H}} \|(T_k - z)^{2^n} \xi\|)^{1/2^n} \leq \delta, \quad k \geq k_0.$$

Similar reasoning holds for the second part of (3.4.5) and hence we deduce that $\gamma_{n,k} \to \gamma_n$ locally uniformly on $\sigma(T)$.

Note that we have actually proved more than what we claimed, namely that if $\delta > 0$, $z_0 \in \partial\sigma(T)$ and ω is a neighborhood around z_0 such that $\gamma_n(z) \leq \delta/2$ for $z \in \omega$, then

$$\gamma_{n,k}(z) \leq \delta, \qquad z \in \omega, \qquad k \geq K, \tag{3.4.6}$$

for some K.

ClaimII: We claim that $\gamma_{n,k} \to \gamma_n$ locally uniformly on $\mathbb{C} \backslash \sigma(T)$. Note that $z \mapsto R(z, T)$ is analytic on $\mathbb{C} \setminus \sigma(T)$ and also, since

$$T_k \xrightarrow{\hat{\delta}} T$$

and $\sigma(T)^c \neq \emptyset$, it follows that if $B_r(a)$ is an open disc with center $a \in \mathbb{C}$, radius r and $B_r(a) \subset \mathbb{C} \backslash \omega_\nu(\sigma(T))$ for some $\nu > 0$ (recall that $\omega_\nu(\Omega)$ denotes the ν-neighborhood around $\Omega \subset \mathbb{C}$), then $R(z, T_k)$ exist and is bounded on a neighborhood of $\overline{B_r(a)}$ for sufficiently large k and hence $z \mapsto R(z, T_k)$ is analytic there. Now,

$$R(z, T_k) \longrightarrow R(z, T), \qquad k \to \infty, \quad z \in B_r(a) \tag{3.4.7}$$

pointwise. Let $f_k(z) = R(z, T_k)$ then, by Cauchy's formula, we have for $z \in B_r(a)$

$$\begin{aligned}\|f_k(a) - f_k(z)\| &\leq \frac{1}{2\pi}\left\|\int_{\partial B_r(a)} \frac{f_k(\omega)(a-z)}{(\omega-a)(\omega-z)}\, d\omega\right\| \\ &\leq \frac{4M}{R}|a-z|,\end{aligned}$$

where M is the bound on f_k on $\overline{B_r(a)}$. Hence, $\{f_k\}$ is locally uniformly Lipschitz and therefore the convergence in (3.4.7) must be locally uniform. Using the reasoning above, the fact that we have $\gamma_{n,k}(z) = 1/\|R(z, T_k)^{2^n}\|^{1/2^n}$ for $z \in \mathbb{C} \setminus \omega_\nu(\sigma(T))$ and sufficiently large k, and the reasoning leading to (3.4.6), then ClaimII easily follows.

By adding Claim I and Claim II we deduce (ii). □

3.5 Proofs of the Main Theorems

We are now ready to prove the main theorems, but before we do that we need a couple of preliminary results.

Proposition 3.5.1. *$T \in \mathcal{B}(\mathcal{H})$ and $\{P_m\}$ is an increasing sequence of finite rank projections converging strongly to the identity. Let $\Phi_{n,m}$ be as in Definition 3.3.3. Define, for $k \in \mathbb{N}$, the functions $\gamma_{n,m}, \gamma_{n,m,k} : \mathbb{C} \to \mathbb{R}$ by*

$$\begin{aligned}\gamma_{n,m}(z) &= \min[\Phi_{n,m}(T, z), \Phi_{n,m}(T^*, \bar{z})], \\ \gamma_{n,m,k}(z) &= \min[\Phi_{n,m}(P_k T P_k, z), \Phi_{n,m}(P_k T^* P_k, \bar{z})],\end{aligned} \tag{3.5.1}$$

and let γ_n be defined as in (3.3.3). Then $\gamma_{n,m} \to \gamma_n$ as $m \to \infty$ and $\gamma_{n,m,k} \to \gamma_{n,m}$ as $k \to \infty$ locally uniformly. The convergence $\gamma_{n,m} \to \gamma_n$ is monotonically from above.

Proof. To see that $\gamma_{n,m} \to \gamma_n$ monotonically from above and locally uniformly as $m \to \infty$, define $\gamma_n^1(z) = \Phi_n(T, z)$, $\gamma_n^2(z) = \Phi_n(T^*, \bar{z})$, $\gamma_{n,m}^1(z) = \Phi_{n,m}(T, z)$ and $\gamma_{n,m}^2(z) = \Phi_{n,m}(T^*, \bar{z})$, where Φ_n and $\Phi_{n,m}$ are defined as in Definition 3.3.3. It follows, by the definition of $\gamma_{n,m}$, that to prove the claim it suffices to show that $\gamma_{n,m}^1 \to \gamma_n^1$ and $\gamma_{n,m}^2 \to \gamma_n^2$ monotonically from above and locally uniformly as $m \to \infty$. Now, $\gamma_{n,m}^j$ is obviously continuous as well as γ_n^j and also, since $P_{n+1} \geq P_n$ and $P_n \to I$, we have that $\gamma_{n,m+1}^j(z) \leq \gamma_{n,m}^j(z)$ and $\lim_{m\to\infty} \gamma_{n,m}^j(z) = \gamma_n^j(z)$ for $z \in \mathbb{C}$. Thus, by appealing to Dini's Theorem, we deduce that $\gamma_{n,m}^j \to \gamma_n^j$ locally uniformly.

To see that $\gamma_{n,m,k} \to \gamma_{n,m}$ as $k \to \infty$, locally uniformly we argue as follows. Using self-adjointness of

$$\begin{aligned}T_m(z) &= P_m((T-z)^*)^{2^n}(T-z)^{2^n}\Big\rceil_{P_m\mathcal{H}} \\ T_{m,k}(z) &= P_m((P_k(T-z)P_k)^*)^{2^n}(P_k(T-z)P_k)^{2^n}\Big\rceil_{P_m\mathcal{H}}, \\ \widetilde{T}_m(z) &= P_m(T-z)^{2^n}((T-z)^*)^{2^n}\Big\rceil_{P_m\mathcal{H}} \\ \widetilde{T}_{m,k}(z) &= P_m(P_k(T-z)P_k)^{2^n}((P_k(T-z)P_k)^*)^{2^n}\Big\rceil_{P_m\mathcal{H}}\end{aligned} \tag{3.5.2}$$

and the fact that for self-adjoint $A, B \in \mathcal{B}(\mathcal{H})$ we have $d_H(\sigma(A), \sigma(B)) \leq \|A - B\|$ it suffices to show that $T_{m,k}(z) \to T_m(z)$ and $\widetilde{T}_{m,k}(z) \to \widetilde{T}_m(z)$, as $k \to \infty$, uniformly for all z in a compact set. To see that we observe that

$$\underset{k\to\infty}{\text{SOT-lim}} P_k(T - z)P_k = T - z, \qquad \underset{k\to\infty}{\text{SOT-lim}} (P_k(T - z)P_k)^* = (T - z)^*,$$

so since multiplication is strongly continuous on bounded sets and the fact P_m has finite rank it follows that the strong convergence implies norm convergence and we deduce that $T_{m,k} \to T_m$ and $\widetilde{T}_{m,k} \to \widetilde{T}_m$ pointwise as $k \to \infty$.

A closer examination shows that the operator valued functions $z \mapsto T_{m,k}(z)$ and $z \mapsto \widetilde{T}_{m,k}(z)$ are Lipschitz continuous on compact sets with a uniformly bounded Lipschitz constant, thus the convergence asserted is locally uniform. □

Theorem 3.5.2. *(Tre04) Let $\mathcal{H}_1$ and $\mathcal{H}_2$ be Hilbert spaces and let $H^\infty_{\mathcal{H}_1\to\mathcal{H}_2}$ denote the set of all bounded analytic function on the open unit disk $\mathbb{D}$ whose values are in $\mathcal{B}(\mathcal{H}_1, \mathcal{H}_2)$. Let $F \in H^\infty_{\mathcal{H}_1\to\mathcal{H}_2}$ and suppose that there is a $\delta > 0$ such that $F^*(z)F(z) \geq \delta I$ for all $z \in \mathbb{D}$. If there is a constant operator $A \in \mathcal{B}(\mathcal{H}_1, \mathcal{H}_2)$ such that*

$$\sup_{z\in\mathbb{D}} \|A - F^*(z)F(z)\|_1 < \infty,$$

where $\|\cdot\|_1$ denotes the trace-norm, then there is a $G \in H^\infty_{\mathcal{H}_2\to\mathcal{H}_1}$ such that $G(z)F(z) = I$ for all $z \in \mathbb{D}$.

Theorem 3.5.3. *(Sha08) Let Ω_0 be a connected open subset of $\mathbb{C}$ and Z a Banach space. Suppose that $F : \Omega_0 \to Z$ is an analytic vector valued function, $\|F(z)\| \leq M$ for all z in an open subset $\Omega \subset \Omega_0$, and $\|F(z_0)\| < M$ for some $z_0 \in \Omega_0$. Then $\|F(z)\| < M$ for all $z \in \Omega$.*

We are now ready to prove the main theorems.

Theorem 3.5.4. *Let $\{e_j\}_{j\in\mathbb{N}}$ be a basis for the Hilbert space $\mathcal{H}$ and let*

$$\begin{aligned}\Delta = \{&T \in \mathcal{C}(\mathcal{H}) : T = W + A,\ W \in WS(\mathcal{H}),\ A \in \mathcal{B}(\mathcal{H})\\ &\cap \{T \in \mathcal{C}(\mathcal{H}) : \|R(T, \cdot)^{2^n}\|^{1/2^n} \text{ is never constant for any } n\}.\end{aligned} \tag{3.5.3}$$

Define, for $n \in \mathbb{N}, \epsilon > 0$, the set valued functions $\Xi_1, \Xi_2, \Xi_3 : \Delta \to \Omega$ by $\Xi_1(T) = \overline{\sigma_{n,\epsilon}(T)}$, $\Xi_2(T) = \overline{\omega_\epsilon(\sigma(T))}$ and $\Xi_3(T) = \sigma(T)$. Then

$$C_{ind}(\Xi_1) \leq 3, \qquad C_{ind}(\Xi_2) \leq 4, \qquad C_{ind}(\Xi_3) \leq 4.$$

Also, if $\Delta = \mathcal{B}(\mathcal{H})$ then $C_{ind}(\Xi_1) \leq 2$, $C_{ind}(\Xi_2) \leq 3$ and $C_{ind}(\Xi_3) \leq 3$.

Proof. Note that if $T \in \Delta$ it follows that, for a compact ball $K \subset \mathbb{C}$ with K^o intersecting $\overline{\sigma_{n,\epsilon}(T)}$ or $\sigma(T)$ we have

$$\sigma(T) \cap K = \lim_{\epsilon\to 0} \overline{\sigma_{n,\epsilon}(T)} \cap K, \qquad \overline{\omega_\epsilon(\sigma(T))} \cap K = \lim_{n\to\infty} \overline{\sigma_{n,\epsilon}(T)} \cap K,$$

(the first assertion is obvious and the second follows from Theorem 3.4.1) thus, it suffices to show, in both cases, the bound on $C_{\text{ind}}(\Xi_1)$. We will first show that if $\Delta = \mathcal{B}(\mathcal{H})$ then

$C_{\text{ind}}(\Xi_1) \leq 2$, and then use this to show that if Δ is defined as in (3.5.3) then $C_{\text{ind}}(\Xi_1) \leq 3$. Let P_n be the projection onto $\text{span}\{e_1, \ldots, e_n\}$ and $x_{ij} = \langle Te_j, e_i\rangle$ for $T \in \mathcal{B}(\mathcal{H})$. Also, define the set

$$\Theta_k = \{z \in \mathbb{C} : \Re z, \Im z = r\delta, r \in \mathbb{Z}, |r| \leq k\}, \qquad \delta = \sqrt{\frac{1}{k}}, \tag{3.5.4}$$

and define the set of estimating functions Γ_{n_1,n_2} and Γ_{n_1} in the following way. Let

$$\begin{aligned}
\Gamma_{n_1,n_2}(\{x_{ij}\}) &= \{z \in \Theta_{n_2} : \nexists\, L \in LT_{\text{pos}}(P_{n_1}\mathcal{H}), T_{\epsilon,n_1,n_2}(z) = LL^*\} \\
&\qquad \cup \{z \in \Theta_{n_2} : \nexists\, L \in LT_{\text{pos}}(P_{n_1}\mathcal{H}), \widetilde{T}_{\epsilon,n_1,n_2}(z) = LL^*\}, \\
\Gamma_{n_1}(\{x_{ij}\}) &= \{z \in \mathbb{C} : (-\infty, 0] \cap \sigma(T_{\epsilon,n_1}(z)) \neq \emptyset\} \\
&\qquad \cup \{z \in \mathbb{C} : (-\infty, 0] \cap \sigma(\widetilde{T}_{\epsilon,n_1}(z)) \neq \emptyset\},
\end{aligned} \tag{3.5.5}$$

where $LT_{\text{pos}}(P_m\mathcal{H})$ denotes the set of lower triangular matrices in $\mathcal{B}(P_m\mathcal{H})$ (with respect to $\{e_j\}$) with strictly positive diagonal elements and

$$\begin{aligned}
T_{\epsilon,n_1,n_2}(z) &= T_{n_1,n_2}(z) - \epsilon^{2^{n+1}} I, \\
\widetilde{T}_{\epsilon,n_1,n_2}(z) &= \widetilde{T}_{n_1,n_2}(z) - \epsilon^{2^{n+1}} I, \\
T_{\epsilon,n_1}(z) &= T_{n_1}(z) - \epsilon^{2^{n+1}} I, \\
T_{\epsilon,n_1}(z) &= T_{n_1}(z) - \epsilon^{2^{n+1}} I,
\end{aligned} \tag{3.5.6}$$

where T_{n_1,n_2}, $\widetilde{T}_{n_1,n_2}$, T_{n_1} and $\widetilde{T}_{n_1}$ are defined as in (3.5.2). Note that, clearly, from the definition, Γ_{n_1,n_2} depends only on $\{x_{ij}\}_{i,j\leq n_2}$. We claim that $\Gamma_{n_1,n_2}(\{x_{ij}\})$ can be evaluated using only finitely many arithmetic operations and radicals of elements in $\{x_{ij}\}_{i,j\leq n_2}$. Indeed, $T_{\epsilon,n_1,n_2}(z)$ and $\widetilde{T}_{\epsilon,n_1,n_2}(z)$ are both in $\mathcal{B}(P_{n_1}\mathcal{H})$. Also, $a_{ij} = \langle T_{\epsilon,n_1,n_2}(z)e_j, e_i\rangle$ and $\tilde{a}_{ij} = \langle \widetilde{T}_{\epsilon,n_1,n_2}(z)e_j, e_i\rangle$, for $i, j \leq n_1$, are, by the definition of $T_{\epsilon,n_1,n_2}(z)$ and $\widetilde{T}_{\epsilon,n_1,n_2}(z)$, polynomials in $\{x_{ij}\}_{i,j\leq n_2}$. Since the existence of $L \in LT_{\text{pos}}(P_{n_1})$ such that $T_{\epsilon,n_1,n_2}(z) = LL^*$ can be determined using finitely many arithmetic operations and radicals of $\{a_{ij}\}_{i,j\leq n_1}$ (this is known as the Cholesky decomposition), similar reasoning holds for $\widetilde{T}_{\epsilon,n_1,n_2}(z)$ and the fact that Θ_{n_2} is finite, the assertion follows.

Step I: We will show that for any compact ball $K \subset \mathbb{C}$ such that $\Gamma_{n_1,n_2}(\{x_{ij}\}) \cap K^o \neq \emptyset$, then

$$d_H(\Gamma_{n_1,n_2}(\{x_{ij}\}) \cap K, \Gamma_{n_1}(\{x_{ij}\}) \cap K) \longrightarrow 0, \quad n_2 \to \infty.$$

Note that since $d_H(\Theta_{n_2} \cap K, K) \to 0$, as $n_2 \to \infty$, and by the observations that for $n_2 \geq n_1$ we have

$$\begin{aligned}
&\{z \in \mathbb{C} : \nexists\, L \in LT_{\text{pos}}(P_{n_1}\mathcal{H}), T_{\epsilon,n_1,n_2}(z) = LL^*\} \\
&\qquad \cup \{z \in \mathbb{C} : \nexists\, L \in LT_{\text{pos}}(P_{n_1}\mathcal{H}), \widetilde{T}_{\epsilon,n_1,n_2}(z) = LL^*\} \\
&= \{z \in \mathbb{C} : (-\infty, 0] \cap \sigma(T_{\epsilon,n_1,n_2}(z)) \neq \emptyset\} \cup \{z \in \mathbb{C} : (-\infty, 0] \cap \sigma(\widetilde{T}_{\epsilon,n_1,n_2}(z)) \neq \emptyset\} \\
&= \{z \in \mathbb{C} : \gamma_{n,n_1,n_2}(z) \leq \epsilon\},
\end{aligned}$$

where γ_{n,n_1,n_2} is defined in (3.5.1), and

$$\begin{aligned}
\{z \in \mathbb{C} : \gamma_{n,n_1}(z) \leq \epsilon\} &= \{z \in \mathbb{C} : (-\infty, 0] \cap \sigma(T_{\epsilon,n_1}(z)) \neq \emptyset\} \\
&\qquad \cup \{z \in \mathbb{C} : (-\infty, 0] \cap \in \sigma(\widetilde{T}_{\epsilon,n_1}(z)) \neq \emptyset\}
\end{aligned} \tag{3.5.7}$$

where γ_{n,n_1} is defined in (3.5.1), the assertion will follow if we can demonstrate that

$$d_H(\{z \in \mathbb{C} : \gamma_{n,n_1,n_2}(z) \leq \epsilon\} \cap K, \{z \in \mathbb{C} : \gamma_{n,n_1}(z) \leq \epsilon\} \cap K) \longrightarrow 0, \tag{3.5.8}$$

as $n_2 \to \infty$. Now, by Proposition 3.5.1 it follows that $\gamma_{n,n_1,n_2} \to \gamma_{n,n_1}$ locally uniformly hence, by Proposition 3.3.1, (3.5.8) will follow if we can prove the following.

Claim: We claim that

$$\mathrm{cl}(\{z \in \mathbb{C} : \gamma_{n,n_1}(z) < \epsilon\}) = \{z \in \mathbb{C} : \gamma_{n,n_1}(z) \leq \epsilon\}. \tag{3.5.9}$$

Now, letting ζ_{1,n_1} and ζ_{2,n_1} be defined by $\zeta_{1,n_1}(z) = \Phi_{n,n_1}(T, z)$ and $\zeta_{2,n_1}(z) = \Phi_{n,n_1}(T^*, \bar{z})$, where Φ_{n,n_1} is defined as in Definition 3.3.3. Then $\gamma_{n,n_1} = \min[\zeta_{1,n_1}, \zeta_{2,n_1}]$. Thus, (3.5.9) will follow if we can show that

$$\mathrm{cl}(\{z \in \mathbb{C} : \zeta_{j,n_1}(z) < \epsilon\}) = \{z \in \mathbb{C} : \zeta_{j,n_1}(z) \leq \epsilon\}, \qquad j = 1, 2. \tag{3.5.10}$$

We will demonstrate the latter, but before we do so we need to establish some facts about the set of points where ζ_{1,n_1} does not vanish. Let

$$\Omega = \{z \in \mathbb{C} : \zeta_{1,n_1}(z) \neq 0\},$$

then Ω is obviously open and we claim that $\mathbb{C} \setminus \Omega$ is finite. To see that we argue by contradiction and suppose that ζ_{1,n_1} vanishes at infinitely many points. If that was the case we would have

$$\inf_{\|\xi\|=1, \xi \in \mathcal{H}} \|(T - z)^{2^n} P_{n_1} \xi\| = 0, \tag{3.5.11}$$

for infinitely many zs. This is indeed impossible because, since P_{n_1} has finite rank, there is a finite dimensional subspace $\mathcal{H}_1 \subset \mathcal{H}$ such that $\mathrm{ran}(T - z)^{2^n - 1} P_{n_1} \subset \mathcal{H}_1$ for all $z \in \mathbb{C}$. Thus , if E is the projection onto $\mathcal{H}_1$ then, by (3.5.11), $\inf_{\eta \in \mathcal{H}_1} \|(ETE - zE)\eta\| = 0$ for infinitely many zs. But the infimum in the equation above is actually attained since $\mathcal{H}_1$ is finite dimensional and hence the finite rank operator ETE must have infinitely many eigenvalues and this is impossible. Armed with this fact we return to the task of showing (3.5.10). To do this for $j = 1$ we argue by contradiction and suppose that there is a

$$z_0 \notin \mathrm{cl}(\{z \in \mathbb{C} : \zeta_{1,n_1}(z) < \epsilon\}) \tag{3.5.12}$$

such that $\zeta_{1,n_1}(z_0) = \epsilon$. This implies that there is a neighborhood θ around z_0 such that $\zeta_{1,n_1}(z) \geq \epsilon$ for $z \in \theta$. We will now demonstrate that this is impossible. First note that by the definition of ζ_{1,n_1} we can make $\zeta_{1,n_1}(z)$ arbitrary large for large $|z|$. In particular, we can find an open set $\tilde{\theta} \subset \Omega$ such that $\zeta_{1,n_1}(z) > \epsilon$ for $z \in \tilde{\theta}$. Now choose a simply connected open set $\Omega_0 \subset \Omega$ such that $\theta \cup \tilde{\theta} \subset \Omega_0$ and ζ_{1,n_1} does not vanish on $\mathrm{cl}(\Omega_0)$. Note that this is possible by the fact that $\mathbb{C} \setminus \Omega$ is finite. Now define, for $z \in \Omega_0$, the operator

$$F(z) : P_{n_1}\mathcal{H} \to \mathcal{H}, \qquad F(z) = (T - z)^{2^n} P_{n_1}.$$

Now, obviously F is holomorphic. Note that, by continuity of ζ_{1,n_1} and the choice of Ω_0, there is a $\delta > 0$ such that

$$\inf_{z \in \Omega_0} \zeta_{1,n_1}(z) \geq \delta.$$

By possibly composing F with a holomorphic function we may assume that $\Omega_0 = \mathbb{D}$, the open disk with radius one centered at the origin. Hence we get that $F \in H^\infty_{P_{n_1}\mathcal{H}\to\mathcal{H}}$ and $F^*(z)F(z) \geq \delta I$, for all $z \in \mathbb{D}$, where I is the identity on $P_{n_1}\mathcal{H}$. Obviously, since P_{n_1} is a finite rank projection, it follows that

$$\sup_{z\in\mathbb{D}} \|F^*(z)F(z)\|_1 < \infty,$$

where $\|\cdot\|_1$ denotes the trace norm. Thus, we may appeal to Theorem 3.5.2 and deduce that there is a $G \in H^\infty_{\mathcal{H}\to P_{n_1}\mathcal{H}}$ such that $G(z)F(z) = I$ for all $z \in \mathbb{D}$. Again, by possibly composing with another holomorphic function (and with a slight abuse of notation) we have a holomorphic function G on Ω_0 such that $G(z) : \mathcal{H} \to P_{n_1}\mathcal{H}$ and

$$1/\zeta_{1,n_1}(z) = 1/(\inf_{\xi\in P_{n_1}\mathcal{H},\|\xi\|=1} \|F(z)\xi\|) = \|G(z)\|, \qquad z \in \Omega_0.$$

Then, by the reasoning above, it follows that $\|G(z)\| \leq 1/\epsilon$ for $z \in \theta$ and $\|G(z)\| < 1/\epsilon$ for $z \in \tilde{\theta}$. This implies, by Theorem 3.5.3, that $\|G(z)\| < 1/\epsilon$ for $z \in \theta$, but $\|G(z_0)\| = 1/\epsilon$ and $z_0 \in \theta$ (recall (3.5.12)) and we have finally reached the desired contradiction. By a similar argument one can show (3.5.10) for $j = 2$ and hence we are done with step I.

Step II: We will show that for any compact ball $K \subset \mathbb{C}$ such that $\overline{\sigma_{n,\epsilon}(T)} \cap K^o \neq \emptyset$, then

$$d_H(\Gamma_{n_1}(\{x_{ij}\}) \cap K, \overline{\sigma_{n,\epsilon}(T)} \cap K) \longrightarrow 0, \quad n_1 \to \infty.$$

But, by (3.5.7) and Theorem 3.3.4 (ii), this will follow if

$$d_H(\{z \in \mathbb{C} : \gamma_{n,n_1}(z) \leq \epsilon\} \cap K, \{z \in \mathbb{C} : \gamma_n(z) \leq \epsilon\} \cap K) \longrightarrow 0, \quad n_1 \to \infty,$$

where γ_n is defined in (3.3.3), and by Theorem 3.3.4 (iii) and Proposition 3.3.1 this is true if $\gamma_{n,n_1} \to \gamma_n$ locally uniformly, which in fact was established in Proposition 3.5.1. Now, adding Step I and Step II together we have shown that $C_{\text{ind}}(\Xi_1) \leq 2$ for $\Xi_1 : \Delta \to \Omega$ when $\Delta = \mathcal{B}(\mathcal{H})$, and we will now use this to establish the assertion of the theorem.

Step III: We will now show that if Δ is defined as in (3.5.3) then $C_{\text{ind}}(\Xi_1) \leq 3$. Suppose that we have $T = W + A$, where W is a weighted shift and A is bounded. Letting $x_{ij} = \langle Te_j, e_i\rangle$ we will define the set of estimating functions $\Gamma_{n_1,\ldots,n_3}, \ldots, \Gamma_{n_1}$ in the following way. Now, for $\xi \in \mathcal{H}$ we may without loss of generality assume that $(W\xi)_j = x_{j,j+k}\xi_j$ for some integer k. Define a new set $\{\tilde{x}_{ij}(n)\}$, depending on an integer n, in the following way: $\tilde{x}_{j,j+k}(n) = n$ if $|x_{j,j+k}| > n$ and $\tilde{x}_{ij}(n) = x_{ij}$ elsewhere. Note that $\{\tilde{x}_{ij}(n)\}$ gives rise to a bounded operator S_n whose matrix elements are $\{\tilde{x}_{ij}(n)\}$. Thus we can define

$$\Gamma_{n_1,\ldots,n_3}(\{x_{ij}\}) = \Gamma_{n_2,n_3}(\{\tilde{x}_{ij}(n_1)\}),$$

where Γ_{n_2,n_3} is defined as in (3.5.5). If we let $\Gamma_{n_1}(\{x_{ij}\}) = \Xi_1(S_{n_1})$, and since we have shown above that Γ_{n_2,n_3} and Γ_{n_2} is a set of estimating functions for $\Xi_1 : \mathcal{B}(\mathcal{H}) \to \Omega$, it follows that $\Gamma_{n_1,\ldots,n_3}, \ldots, \Gamma_{n_1}$ is a set of estimating functions for Ξ_1 if we can show that

$$\lim_{n_1\to\infty} \Xi_1(S_{n_1}) = \Xi_1(T).$$

Note that, by Theorem 3.4.3 (and assumption), the latter will follow if we can show that

$$S_n \xrightarrow{\hat{\delta}} T, \qquad n \to \infty. \tag{3.5.13}$$

Define the operator W_n by $(W_n\xi)_j = \tilde{x}_{j,j+k}(n)\xi_j$ for $\xi \in \mathcal{H}$. Then $S_n = W_n + A$. Thus, by Theorem 1.1.2, (3.5.13) will follow if we can show that $\delta(W_n, W) \to 0$ and $\delta(W, W_n) \to 0$ as $n \to \infty$. To show the former we need to demonstrate that

$$\sup_{\varphi \in G(W_n), \|\varphi\| \leq 1} \inf_{\psi \in G(W)} \|\varphi - \psi\| \longrightarrow 0, \qquad n \to \infty,$$

where $G(W)$ denotes the graph of W as defined in (1.1.1). Let $\varphi \in G(W_n)$ such that $\|\varphi\| \leq 1$. Then there is a $\xi \in \mathcal{H}$ such that $\varphi = (\xi, W_n\xi)$ and $\|W_n\xi\| + \|\xi\| \leq 1$. Now, choose $\eta \in \mathcal{D}(W)$ in the following way:

$$\eta_j = \begin{cases} \xi_j & \text{if} \quad \tilde{x}_{j,j+k}(n) = x_{j,j+k} \\ \frac{\tilde{x}_{j,j+k}(n)}{x_{j,j+k}}\xi_j & \text{if} \quad \tilde{x}_{j,j+k}(n) \neq x_{j,j+k}. \end{cases}$$

Let also $\Theta = \{j \in \mathbb{N} : \eta_j = \xi_j\}$ and $\theta = \{j \in \mathbb{N} : \eta_j \neq \xi_j\}$. Then,

$$\begin{aligned} &\|\xi - \eta\| + \|W_n\xi - W\eta\| \\ &\qquad = \sum_{j \in \Theta} |\xi_j - \eta_j|^2 + \sum_{j \in \theta} |\xi_j - \eta_j|^2 \\ &\qquad\quad + \sum_{j \in \Theta} |\tilde{x}_{j,j+k}(n)\xi_j - x_{j,j+k}\eta_j|^2 + \sum_{j \in \theta} |\tilde{x}_{j,j+k}(n)\xi_j - x_{j,j+k}\eta_j|^2 \\ &\qquad = \sum_{j \in \theta} |\xi_j - \eta_j|^2 + \sum_{j \in \theta} |\tilde{x}_{j,j+k}(n)\xi_j - x_{j,j+k}\eta_j|^2. \end{aligned}$$

Now $\sum_{j \in \theta} |\tilde{x}_{j,j+k}(n)|^2 |\xi_j|^2 \leq 1$ and $\tilde{x}_{j,j+k}(n) = n$ for $j \in \theta$ so $\sum_{j \in \theta} |\xi_{j+k}|^2 \leq 1/n^2$. So by the fact that $|\tilde{x}_{j,j+k}(n)/x_{j+k}| \leq 1$ and the choice of η it follows that

$$\sum_{j \in \theta} |\xi_j - \eta_j|^2 \leq 4/n^2.$$

Also, $\sum_{j \in \theta} |\tilde{x}_{j,j+k}(n)\xi_j - x_{j,j+k}\eta_j|^2 = 0$, by the choice of η, and thus $\|\xi - \eta\| + \|W_n\xi - W\eta\| \leq 2/n$. Hence $\inf_{\psi \in G(W)} \|\varphi - \psi\| \leq 2/n$ and so since φ was arbitrary we have

$$\sup_{\varphi \in G(A_N), \|\varphi\| \leq 1} \inf_{\psi \in G(A)} \|\varphi - \psi\| \leq 2/n \longrightarrow 0, \qquad n \to \infty.$$

The fact that $\delta(W, W_n) \to 0$ as $n \to \infty$ follows by similar reasoning. □

Remark 3.5.5. The assumption that $\|R(T, \cdot)^{2^n}\|^{1/2^n}$ is never constant for any n will be satisfied e.g. if $\mathbb{C} \setminus \sigma(T)$ is connected and the numerical range of T is contained in a sector of the complex plane.

Theorem 3.5.6. *Let $\{e_j\}_{j \in \mathbb{N}}$ and $\{\tilde{e}_j\}_{j \in \mathbb{N}}$ be bases for the Hilbert space $\mathcal{H}$ and let*

$$\tilde{\Delta} = \{T \in \mathcal{C}(\mathcal{H} \oplus \mathcal{H}) : T = T_1 \oplus T_2, T_1, T_2 \in \mathcal{C}(\mathcal{H}), T_1^* = T_2\}$$
$$\Delta = \{T \in \tilde{\Delta} : \operatorname{span}\{e_j\}_{j \in \mathbb{N}} \text{ is a core for } T_1, \operatorname{span}\{\tilde{e}_j\} \text{ is a core for } T_2\}.$$

Let $\epsilon > 0$, $\Xi_1 : \Delta \to \Omega$ and $\Xi_2 : \Delta \to \Omega$ be defined by $\Xi_1(T) = \overline{\sigma_\epsilon(T_1)}$ and $\Xi_2(T) = \sigma(T_1)$. Then $C_{\text{ind}}(\Xi_1) \leq 2$ and $C_{\text{ind}}(\Xi_2) \leq 3$.

Proof. Arguing as in the proof of Theorem 3.5.4, it suffices to show that $C_{\text{ind}}(\Xi_1) \leq 2$. Let P_m and $\widetilde{P}_m$ be the projections onto $\text{span}\{e_j\}_{j=1}^m$ and $\text{span}\{\tilde{e}_j\}_{j=1}^m$ respectively and define

$$S_m : \Delta \times \mathbb{C} \to \mathcal{B}(P_m\mathcal{H}, \mathcal{H}), \qquad \widetilde{S}_m : \Delta \times \mathbb{C} \to \mathcal{B}(\widetilde{P}_m\mathcal{H}, \mathcal{H})$$

by

$$S_m(T, z) = (TE_1 - z)P_m, \qquad \widetilde{S}_m(T, z) = (TE_2 - \bar{z})\widetilde{P}_m,$$

where $E_1 : \mathcal{H} \oplus \mathcal{H} \to \mathcal{H}$ and $E_2 : \mathcal{H} \oplus \mathcal{H} \to \mathcal{H}$ are the projections onto the first and second component, respectively. Also, define

$$S_{m,k} : \Delta \times \mathbb{C} \to \mathcal{B}(P_m\mathcal{H}, \mathcal{H}), \qquad \widetilde{S}_{m,k} : \Delta \times \mathbb{C} \to \mathcal{B}(\widetilde{P}_m\mathcal{H}, \mathcal{H})$$

by

$$S_{m,k}(T, z) = (P_kTE_1P_k - z)P_m \qquad \widetilde{S}_m(T, z) = (\widetilde{P}_kTE_2\widetilde{P}_k - \bar{z})\widetilde{P}_m.$$

Now, for $T \in \Delta$, let $\{x_{ij}\}$ be some ordering of the matrix elements

$$\{\langle T_1e_j, e_i\rangle\} \cup \{\langle T_2\tilde{e}_j, \tilde{e}_i\rangle\}_{i,j\in\mathbb{N}},$$

and define the estimating functions Γ_{n_1,n_2} and Γ_{n_1} by

$$\begin{aligned}
\Gamma_{n_1,n_2}(\{x_{ij}\}) &= \{z \in \Theta_{n_2} : \nexists\, L \in LT_{\text{pos}}(P_{n_1}\mathcal{H}), T_{\epsilon,n_1,n_2}(z) = LL^*\} \\
&\quad \cup \{z \in \Theta_{n_2} : \nexists\, L \in LT_{\text{pos}}(\widetilde{P}_{n_1}\mathcal{H}), \widetilde{T}_{\epsilon,n_1,n_2}(z) = LL^*\}, \\
\Gamma_{n_1}(\{x_{ij}\}) &= \{z \in \mathbb{C} : (-\infty, 0] \cap \sigma(T_{\epsilon,n_1}(z)) \neq 0\} \\
&\quad \cup \{z \in \mathbb{C} : (-\infty, 0] \cap \sigma(\widetilde{T}_{\epsilon,n_1}(z)) \neq \emptyset\},
\end{aligned}$$

where Θ_{n_2} is defined as in (3.5.4) and

$$T_{\epsilon,n_1,n_2}(z) = S_{n_1,n_2}(z)^*S_{n_1,n_2}(z) - \epsilon^2 I, \quad \widetilde{T}_{\epsilon,n_1,n_2}(z) = \widetilde{S}_{n_1,n_2}(z)^*\widetilde{S}_{n_1,n_2}(z) - \epsilon^2 I$$

and $T_{\epsilon,n_1}(z) = S_{n_1}(z)^*S_{n_1}(z) - \epsilon^2 I$, $\widetilde{T}_{\epsilon,n_1}(z) = \widetilde{S}_{n_1}(z)^*\widetilde{S}_{n_1}(z) - \epsilon^2 I$. As argued in the proof of Theorem 3.2.2, Γ_{n_1,n_2} depends on only finitely many elements in $\{x_{ij}\}$, and its evaluation requires finitely many arithmetic operations and radicals of the matrix elements $\{x_{ij}\}$. We are now ready to prove:

Step I. We will show that

$$\Gamma_{n_1}(\{x_{ij}\}) = \lim_{n_2\to\infty} \Gamma_{n_1,n_2}(\{x_{ij}\}).$$

Before we can do that, we must establish a couple of facts first. Now, let $\Phi_m : \Delta \times \mathbb{C} \to \mathbb{R}$, $\widetilde{\Phi}_m : \Delta \times \mathbb{C} \to \mathbb{R}$, $\Phi_{m,k} : \Delta \times \mathbb{C} \to \mathbb{R}$ and $\widetilde{\Phi}_{m,k} : \Delta \times \mathbb{C} \to \mathbb{R}$ be defined by

$$\begin{aligned}
\Phi_m(T, z) &= \min\{\sqrt{\lambda} : \lambda \in \sigma(S_m(T, z)^*S_m(T, z))\}, \\
\widetilde{\Phi}_m(T, z) &= \min\{\sqrt{\lambda} : \lambda \in \sigma(\widetilde{S}_m(T, z)^*\widetilde{S}_m(T, z))\}, \\
\Phi_{m,k}(T, z) &= \min\{\sqrt{\lambda} : \lambda \in \sigma(S_{m,k}(T, z)^*S_{m,k}(T, z))\}, \\
\widetilde{\Phi}_{m,k}(T, z) &= \min\{\sqrt{\lambda} : \lambda \in \sigma(\widetilde{S}_{m,k}(T, z)^*\widetilde{S}_{m,k}(T, z))\}.
\end{aligned}$$

Claim: We claim that

$$\{z \in \mathbb{C} : \Phi_m(T, z) \leq \epsilon\} = \mathrm{cl}(\{z \in \mathbb{C} : \Phi_m(T, z) < \epsilon\}). \tag{3.5.14}$$

Indeed, this is the case, and the proof is almost identical to the argument used in the proof of Theorem 3.2.2. Let

$$\Omega = \{z \in \mathbb{C} : \Phi_m(T, z) \neq 0\},$$

then Ω is obviously open and we claim that $\mathbb{C} \setminus \Omega$ is finite. To see that, we argue by contradiction and suppose that $\Phi_m(T, \cdot)$ vanishes at infinitely many points. If that was the case we would have

$$\inf_{\|\xi\|=1, \xi \in \mathcal{H}} \|(T_1 - z)P_m\xi\| = 0 \tag{3.5.15}$$

for infinitely many zs. But the infimum in (3.5.15) is attained since P_m has finite rank, so this implies that the operator $P_m T_1 \lceil_{P_m\mathcal{H}}$ has infinitely many eigenvalues. This is, of course, impossible since P_m has finite rank. Armed with this fact we return to the task of showing (3.5.14). Observe that since P_m has finite rank we can make $\inf_{\|\xi\|=1, \xi \in \mathcal{H}} \|(T_1 - z)P_m\xi\|$ arbitrary large for large $|z|$, and in particular, $\Phi_m(T, \cdot)$ can be made arbitrary large as long as $|z|$ is large. Using this we may argue exactly as in the proof of Theorem 3.2.2 and deduce that if there is a

$$z_0 \notin \mathrm{cl}(\{z \in \mathbb{C} : \Phi_m(T, z) < \epsilon\})$$

such that $\Phi_m(T, z_0) = \epsilon$ then there is an open connected set $\Omega_0 \subset \Omega$ containing z_0 and an operator valued holomorphic function G on Ω_0 such that we have $G(z) : \mathcal{H} \to P_m\mathcal{H}$,

$$1/\Phi_m(T, z) = \|G(z)\|, \qquad z \in \Omega_0,$$

and $\|G(z_1)\| < 1/\epsilon$ for some $z_1 \in \Omega_0$. By the assumption on z_0, there is a neighborhood θ around z_0 such that

$$\|G(z)\| \leq 1/\epsilon, \qquad z \in \theta$$

and since $\|G(z_1)\| < 1/\epsilon$ it follows, by Theorem 3.5.3, that $\|G(z)\| < 1/\epsilon$ for all $z \in \theta$. But $\|G(z_0)\| = 1/\epsilon$ and this is a contradiction.

Note that similar reasoning gives that

$$\{z \in \mathbb{C} : \widetilde{\Phi}_m(T, z) \leq \epsilon\} = \mathrm{cl}(\{z \in \mathbb{C} : \widetilde{\Phi}_m(T, z) < \epsilon\}). \tag{3.5.16}$$

So, by observing that

$$\begin{aligned} \Gamma_{n_1,n_2}(\{x_{ij}\}) &= \{z \in \Theta_{n_2} : \min[\Phi_{n_1,n_2}(T, z), \widetilde{\Phi}_{n_1,n_2}(T, z)] \leq \epsilon\}, \\ \Gamma_{n_1}(\{x_{ij}\}) &= \{z \in \mathbb{C} : \min[\Phi_{n_1}(T, z), \widetilde{\Phi}_{n_1}(T, z)] \leq \epsilon\} \end{aligned} \tag{3.5.17}$$

it suffices to show, by Proposition 3.3.1 that

$$\min[\Phi_{n_1,n_2}(T, z), \widetilde{\Phi}_{n_1,n_2}(T, z)] \to \min[\Phi_{n_1}(T, z), \widetilde{\Phi}_{n_1}(T, z)]$$

locally uniformly as $n_2 \to \infty$, which again will follow if we can show that the mappings

$$\begin{aligned} z \mapsto \langle S_{n_1,n_2}(T, z)^* S_{n_1,n_2}(T, z)e_j, e_i\rangle &\longrightarrow z \mapsto \langle S_{n_1}(T, z)^* S_{n_1}(T, z)e_j, e_i\rangle \\ z \mapsto \langle \widetilde{S}_{n_1,n_2}(T, z)^* \widetilde{S}_{n_1,n_2}(T, z)\tilde{e}_j, \tilde{e}_i\rangle &\longrightarrow z \mapsto \langle \widetilde{S}_{n_1}(T, z)^* \widetilde{S}_{n_1}(T, z)\tilde{e}_j, \tilde{e}_i\rangle \end{aligned} \tag{3.5.18}$$

locally uniformly as $n_2 \to \infty$, where $e_j, e_i \in P_{n_1}\mathcal{H}$ and $\tilde{e}_j, \tilde{e}_i \in \widetilde{P}_{n_1}\mathcal{H}$. Note that for $k \geq m$ we have $\langle S_{n_1,n_2}(T,z)^* S_{n_1,n_2}(T,z)e_j, e_i\rangle = \langle P_{n_2}(T-z)e_j, P_{n_2}(T-z)e_i\rangle$, yielding the first part of (3.5.18), and similar reasoning yields the second part.

Step II: We will show that

$$\lim_{n_1\to\infty} \Gamma_{n_1}(\{x_{ij}\}) = \overline{\sigma_\epsilon(T_1)}. \tag{3.5.19}$$

To do that we will first demonstrate the following;

$$\gamma_1(z) = \lim_{n_1\to\infty} \Phi_{n_1}(T,z), \qquad \gamma_2(z) = \lim_{n_1\to\infty} \widetilde{\Phi}_{n_1}(T,z)$$

exist, the convergence is monotonically from above and locally uniform and

$$\sigma_\epsilon(T_1) = \{z \in \mathbb{C} : \min[\gamma_1(z), \gamma_2(z)] < \epsilon\}. \tag{3.5.20}$$

Now, note that

$$\Phi_{n_1}(T,z) = \min_{\xi \in P_{n_1}\mathcal{H}} \|(T_1 - z)\xi\|, \qquad \widetilde{\Phi}_{n_1}(T,z) = \min_{\xi \in \widetilde{P}_{n_1}\mathcal{H}} \|(T_1 - z)^*\xi\|.$$

So, by the assumption that $\operatorname{span}\{e_j\}_{j\in\mathbb{N}}$ is a core for T_1 and $\operatorname{span}\{\tilde{e}_j\}_{j\in\mathbb{N}}$ is a core for T_2, it follows that the limits exist and that

$$\gamma_1(z) = \inf\{\lambda : \lambda \in \sigma(|(T_1 - z)|\}, \qquad \gamma_2(z) = \inf\{\lambda : \lambda \in \sigma(|(T_1 - z)^*|)\}.$$

By Dini's theorem it follows that the convergence is as asserted. Using this fact and by arguing as in the proof of Theorem 3.4.3 we get (3.5.20). The previous reasoning implies that $\min[\Phi_{n_1}(T,z), \widetilde{\Phi}_{n_1}(T,z)] \to \min[\gamma_1(z), \gamma_2(z)]$ monotonically from above and locally uniformly as $n_1 \to \infty$. So, by Proposition 3.3.4 and (3.5.20), it follows that, for compact ball K such that $\overline{\sigma_\epsilon(T_1)} \cap K^o \neq \emptyset$, we have

$$\operatorname{cl}(\{z \in \mathbb{C} : \min[\Phi_{n_1}(T,z), \widetilde{\Phi}_{n_1}(T,z)] < \epsilon\}) \cap K \longrightarrow \overline{\sigma_\epsilon(T_1)} \cap K,$$

as $n_1 \to \infty$. But by (3.5.14),(3.5.17) and (3.5.16) it follows that

$$\Gamma_{n_1}(\{x_{ij}\}) = \operatorname{cl}(\{z \in \mathbb{C} : \min[\Phi_{n_1}(T,z), \widetilde{\Phi}_{n_1}(T,z)] < \epsilon\}),$$

and hence (3.5.19) follows. □

Corollary 3.5.7. *Let $\{e_j\}_{j\in\mathbb{N}}$ be a basis for the Hilbert space $\mathcal{H}$ and let*

$$\Delta = \{A \in \mathcal{SA}(\mathcal{H}) : \operatorname{span}\{e_j\}_{j\in\mathbb{N}} \textit{ is a core for } A\}.$$

Let $\epsilon > 0$ and $\Xi_1, \Xi_2 : \Delta \to \Omega$ be defined by $\Xi_1(T) = \sigma(T)$ and $\Xi_2(T) = \overline{\omega_\epsilon(\sigma(T))}$. Then $C_{\text{ind}}(\Xi_1) \leq 3$ and $C_{\text{ind}}(\Xi_2) \leq 2$.

Theorem 3.5.8. *Let $\{e_j\}_{j\in\mathbb{N}}$ be a basis for the Hilbert space $\mathcal{H}$, P_m be the projection onto $\operatorname{span}\{e_j\}_{j=1}^m$ and d be some positive integer. Let $\Delta \subset \mathcal{C}(\mathcal{H})$ have the following properties: For $T \in \Delta$ we have*

(i) $\bigcup_m P_m\mathcal{H} \subset \mathcal{D}(T)$, $\bigcup_m P_m\mathcal{H} \subset \mathcal{D}(T^*)$.

(ii) $\langle Te_{j+l}, e_j\rangle = \langle Te_j, e_{j+l}\rangle = 0$, *for* $l > d$.

(iii) $TP_m\xi \to T\xi$, $T^*P_m\eta \to T^*\eta$, *as* $m \to \infty$ *for* $\xi \in \mathcal{D}(T)$ *and* $\eta \in \mathcal{D}(T^*)$.

Let $\epsilon > 0$ *and* $n \in \mathbb{Z}_+$ *and* $\Xi_1, \Xi_2, \Xi_3 : \Delta \to \Omega$ *be defined by* $\Xi_1(T) = \overline{\sigma_{n,\epsilon}(T)}$, $\Xi_2(T) = \overline{\omega_\epsilon(\sigma(T))}$ *and* $\Xi_3(T) = \sigma(T)$. *Then*

$$C_{ind}(\Xi_1) = 1, \qquad C_{ind}(\Xi_2) \leq 2, \qquad C_{ind}(\Xi_3) \leq 2.$$

Proof. As in the proof of Theorem 3.2.2 it suffices to demonstrate that $C_{\text{ind}}(\Xi_1) = 1$. Now, obviously we have $C_{\text{ind}}(\Xi) > 0$, so it suffices to show that $C_{\text{ind}}(\Xi) \leq 1$. We follow the proof of Theorem 3.2.2 closely. Let P_n be the projection onto $\text{span}\{e_1, \ldots, e_n\}$ and $x_{ij} = \langle Te_j, e_i\rangle$ for $T \in \Delta$. For $k \in \mathbb{N}$ define T^k inductively by $T^k\xi = T(T^{k-1}\xi)$ on

$$\mathcal{D}(T^k) = \{\xi : \xi \in \mathcal{D}(T^{k-1}), T^{k-1}\xi \in \mathcal{D}(T)\},$$

and define $\mathcal{D}((T^*)^k)$ similarly. Then it is easy to see that $\bigcup_m P_m\mathcal{H} \subset \mathcal{D}(T^k)$, so T^k is densely defined. The fact that T^k is closed is well known (DS88)(p. 603), and it follows (by a straightforward argument using the assumptions (ii) and (iii)) that $\bigcup_m P_m\mathcal{H}$ is a core for T^k. Similarly, we get that $(T^*)^k$ is closed and densely defined and that $\bigcup_m P_m\mathcal{H} \subset \mathcal{D}((T^*)^k)$ is a core for $(T^*)^k$. Using this, it is easy to see that we can, for integers m, k, define $T_{\epsilon,m,k}(z) = T_{m,k}(z) - \epsilon^{2^{n+1}}I$ and $\widetilde{T}_{\epsilon,m,k}(z) = \widetilde{T}_{m,k}(z) - \epsilon^{2^{n+1}}I$, where $T_{m,k}(z)$ and $\widetilde{T}_{m,k}(z)$ are defined in (3.5.2). Let, for $k \in \mathbb{N}$, Θ_k be defined as in (3.5.4) and

$$\begin{aligned}\Psi_k = \{z \in \mathbb{C} : \nexists\, L \in LT_{\text{pos}}(P_k\mathcal{H}), T_{\epsilon,k,2^nd+k}(z) = LL^*\} \\ \cup \{z \in \mathbb{C} : \nexists\, L \in LT_{\text{pos}}(P_k\mathcal{H}), \widetilde{T}_{\epsilon,k,2^nd+k}(z) = LL^*\},\end{aligned} \tag{3.5.21}$$

where $LT_{\text{pos}}(P_m\mathcal{H})$ denotes the set of lower triangular matrices in $P_m\mathcal{H}$ (with respect to $\{e_j\}$) with strictly positive diagonal elements. Now, define Γ_k by

$$\Gamma_k(\{x_{ij}\}) = \Psi_k \cap \Theta_k.$$

By the same reasoning as in the proof of Theorem 3.2.2, it follows that Γ_{n_1,n_2} depends only on finitely many of the x_{ij}s and requires only finitely many arithmetic operations and radicals of $\{x_{ij}\}$ for its evaluation. Now, to show that

$$\Xi(T) = \lim_{k\to\infty} \Gamma_k(\{x_{ij}\}),$$

we need to show that for any compact ball K such that $\overline{\sigma_{n,\epsilon}(T)} \cap K^o \neq \emptyset$ then

$$d_H(\overline{\sigma_{n,\epsilon}(T)} \cap K, \Gamma_k(\{x_{ij}\}) \cap K) \longrightarrow 0, \qquad k \to \infty.$$

But, since obviously $d_H(\Theta_k \cap K, K) \to 0$ as $k \to \infty$ it suffices to show that

$$d_H(\Psi_k \cap K, \overline{\sigma_{n,\epsilon}(T)} \cap K) \longrightarrow 0. \tag{3.5.22}$$

To prove that, note that by the reasoning in the beginning of the proof we may define $\Phi_{n,m} : \Delta \times \mathbb{C} \to \mathbb{R}$ by

$$\Phi_{n,m}(S, z) = \min\left\{\lambda^{1/2^{n+1}} : \lambda \in \sigma\left(P_m((S-z)^*)^{2^n}(S-z)^{2^n}\Big\lceil_{P_m\mathcal{H}}\right)\right\}.$$

Let $\gamma_{n,k} = \min[\Phi_{n,k}(T,\cdot), \Phi_{n,k}(T^*,\cdot)]$ and $\gamma_{n,k,m} = \min[\Phi_{n,k}(P_mTP_m,\cdot), \Phi_{n,k}(P_mT^*P_m,\cdot)]$. Before we can continue with the proof of (3.5.22) we need the following fact.

ClaimI: We claim that $\Psi_k = \{z \in \mathbb{C} : \gamma_{n,k}(z) \leq \epsilon\}$. To deduce the claim it suffices to show that

$$\gamma_{n,k}(z) = \gamma_{n,k,2^nd+k}(z), \qquad z \in \mathbb{C}, \tag{3.5.23}$$

and why becomes clear after we make the observation that we have

$$\begin{aligned}\Psi_k &= \{z \in \mathbb{C} : (-\infty, 0] \cap \sigma(T_{\epsilon,k,2^nd+k}(z)) \neq \emptyset\} \\ &\qquad \cup \{z \in \mathbb{C} : (-\infty, 0] \cap \sigma(\widetilde{T}_{\epsilon,k,2^nd+k}(z)) \neq \emptyset\} \\ &= \{z \in \mathbb{C} : \gamma_{n,k,2^nd+k}(z) \leq \epsilon\}.\end{aligned}$$

Now (3.5.23) will follow if we can prove that

$$\begin{aligned}&\langle((T-z)^*)^{2^n}(T-z)^{2^n}\xi, \eta\rangle \\ &\qquad = \langle(P_{2^nd+k}(T-z)P_{2^nd+k})^*)^{2^n}(P_{2^nd+k}(T-z)P_{2^nd+k})^{2^n}\xi, \eta\rangle. \\ &\langle(T-z)^{2^n}((T-z)^*)^{2^n}\xi, \eta\rangle \\ &\qquad = \langle(P_{2^nd+k}(T-z)P_{2^nd+k})^{2^n}(P_{2^nd+k}(T-z)P_{2^nd+k})^*)^{2^n}\xi, \eta\rangle,\end{aligned}$$

for $\xi, \eta \in P_k\mathcal{H}$. To show the latter it is easy to see that it suffices to show that

$$\begin{aligned}(P_{2^nd+k}TP_{2^nd+k})^l\xi &= T^l\xi, \qquad \xi \in P_k\mathcal{H}, \quad l \leq 2^n, \\ (P_{2^nd+k}T^*P_{2^nd+k})^l\xi &= T^l\xi, \qquad \xi \in P_k\mathcal{H}, \quad l \leq 2^n.\end{aligned} \tag{3.5.24}$$

To show the first part of (3.5.24), let $\mu \in \mathbb{N}$ such that $\mu > d$, and note that, by assumption, we can write $T\restriction_{\bigcup_m P_m\mathcal{H}}$ as (with a slight abuse of notation)

$$T = P_\mu TP_\mu + P_\mu^\perp TP_\mu^\perp + \sum_{j=-d}^{d-1} \zeta_j \otimes e_{\mu-j},$$

where $\zeta_j \in (P_{\mu+d} - P_{\mu-d})\mathcal{H}$. Now this gives us that, for $l \in \mathbb{N}$,

$$\begin{aligned}T^l = (P_\mu TP_\mu)^l &+ \text{terms of the form} \\ &(P_\mu^\perp TP_\mu^\perp + \sum_{j=-d}^{d-1} \zeta_j \otimes e_{\mu-j})^{p_1} \times (P_\mu TP_\mu)^{q_1} \\ &\times (P_\mu^\perp TP_\mu^\perp + \sum_{j=-d}^{d-1} \zeta_j \otimes e_{\mu-j})^{p_2} \times (P_\mu TP_\mu)^{q_2} \times \cdots \\ &\times (P_\mu^\perp TP_\mu^\perp + \sum_{j=-d}^{d-1} \zeta_j \otimes e_{\mu-j})^{p_t} \times (P_\mu TP_\mu)^{q_t},\end{aligned}$$

where $q_i \leq l-1$ and $p_i \leq l$. Note that since $T \in \Delta$ (using assumption (ii)) it is straightforward to show that

$$\langle(P_\mu TP_\mu)^q e_r, e_j\rangle = 0, \qquad r \leq k, \quad j > qd + k,$$

for any integer q. Hence,

$$(P^{\perp}_{2^n d+k} T P^{\perp}_{2^n d+k} + \sum_{j=-d}^{d-1} \zeta_j \otimes e_{2^n d+k-j})^p \times (P_{2^n d+k} T P_{2^n d+k})^q e_r = 0, \tag{3.5.25}$$

for $r \leq k$, $q \leq 2^n - 1$ and $p \leq 2^n$ yielding the first part of (3.5.24). The second part of (3.5.24) follows by similar reasoning.

Armed with ClaimI we have reduced the problem to showing that if K is a compact ball and K^o intersects $\overline{\sigma_{n,\epsilon}(T)}$, then

$$\lim_{k\to\infty} \{z \in \mathbb{C} : \gamma_{n,k}(z) \leq \epsilon\} \cap K = \overline{\sigma_{n,\epsilon}(T)} \cap K. \tag{3.5.26}$$

Now, the fact that $T \in \Delta$ and the reasoning in the beginning of the proof allows us to define

$$\gamma_n(z) = \min \Big[\inf \Big\{ \lambda^{2^{n+1}} : \lambda \in \sigma\Big(|(T-z)^{2^n}| \Big) \Big\}, \\ \inf \Big\{ \lambda^{2^{n+1}} : \lambda \in \sigma\Big(|((T-z)^*)^{2^n}| \Big) \Big\} \Big].$$

Note that, by arguing similarly as in the proof of (ii) and (iii) in Theorem 3.3.4, we deduce that $\sigma_{n,\epsilon}(T) = \{z \in \mathbb{C} : \gamma_n(z) < \epsilon\}$. By arguing as in Proposition 3.5.1, using the fact that $\bigcup_m P_m \mathcal{H}$ is a core for T^k and $(T^*)^k$ we deduce that $\gamma_{n,k} \to \gamma_n$ locally uniformly and monotonically from above. By arguing as in the proof of Theorem 3.5.6 we deduce that

$$\mathrm{cl}(\{z \in \mathbb{C} : \gamma_{n,k}(z) < \epsilon\}) = \{z \in \mathbb{C} : \gamma_{n,k}(z) \leq \epsilon\}.$$

Thus, using Proposition 3.3.1 we conclude that (3.5.26) is true, and we are done. □

3.6 Other Types of Pseudospectra

The disadvantage of the n-pseudospectrum is that even though one can estimate the spectrum by taking n very large, n may have to be too large for practical purposes. Thus, since we only have the estimate for $T \in \mathcal{C}(\mathcal{H}), \epsilon > 0$ that $\sigma(T) \subset \sigma_{n,\epsilon}(T)$, it is important to get a "lower" bound on $\sigma(T)$ i.e. we want to find a set $\Omega \subset \mathbb{C}$ such that $\Omega \subset \sigma(T)$. A candidate for this is described in the following.

Definition 3.6.1. *Let $T \in \mathcal{B}(\mathcal{H})$ and Φ_0 be defined as in Definition 3.3.3. Let $\zeta_1(z) = \Phi_0(T, z)$ $\zeta_2(z) = \Phi_0(T^*, \bar{z})$. Now let $\epsilon > 0$ and define the ϵ-residual pseudospectrum to be the set*

$$\sigma_{\mathrm{res},\epsilon}(T) = \{z : \zeta_1(z) > \epsilon,\, \zeta_2(z) = 0\}$$

and the adjoint ϵ-residual pseudospectrum to be the set

$$\sigma_{\mathrm{res}^*,\epsilon}(T) = \{z : \zeta_1(z) = 0,\, \zeta_2(z) > \epsilon\}.$$

Theorem 3.6.2. *Let $T \in \mathcal{B}(\mathcal{H})$ and let $\{T_k\} \subset \mathcal{B}(\mathcal{H})$ such that $T_k \to T$ in norm, as $k \to \infty$. Then for $\epsilon > 0$ we have the following,*

(i) $\sigma(T) \supset \bigcup_{\epsilon>0} \sigma_{\mathrm{res},\epsilon}(T) \cup \sigma_{\mathrm{res}^*,\epsilon}(T)$

(ii) $\mathrm{cl}(\{z \in \mathbb{C} : \zeta_1(z) < \epsilon\}) = \{z \in \mathbb{C} : \zeta_1(z) \leq \epsilon\}$

(iii) $\mathrm{cl}(\{z \in \mathbb{C} : \zeta_2(z) < \epsilon\}) = \{z \in \mathbb{C} : \zeta_2(z) \leq \epsilon\}$

(iv) For any compact ball $K \subset \mathbb{C}$ *such that* $\mathrm{cl}(\sigma_{\mathrm{res},\epsilon}(T)) \cap K^o \neq \emptyset$ *it follows that*

$$d_H(\mathrm{cl}(\sigma_{\mathrm{res},\epsilon}(T_k)) \cap K, \mathrm{cl}(\sigma_{\mathrm{res},\epsilon}(T)) \cap K) \longrightarrow 0, \qquad k \to \infty.$$

(v) For any compact ball $K \subset \mathbb{C}$ *such that* $\sigma_{\mathrm{res}^*,\epsilon}(T) \cap K^o \neq \emptyset$ *it follows that*

$$d_H(\mathrm{cl}(\sigma_{\mathrm{res}^*,\epsilon}(T_k)) \cap K, \mathrm{cl}(\sigma_{\mathrm{res}^*,\epsilon}(T) \cap K)) \longrightarrow 0, \qquad k \to \infty.$$

Proof. Note that (i) follows by arguing as in the proof of Theorem 3.3.4, so we will not be repeating that reasoning here. Now, we will show (ii), namely, that

$$\{z \in \mathbb{C} : \zeta_1(z) \leq \epsilon\} = \mathrm{cl}(\{z \in \mathbb{C} : \zeta_1(z) < \epsilon\}). \tag{3.6.1}$$

We argue by contradiction. Suppose that there is a $z_0 \in \mathbb{C} \setminus \mathrm{cl}(\{z \in \mathbb{C} : \zeta_1(z) < \epsilon\})$ such that $\zeta_1(z_0) = \epsilon$. Then, there is a neighborhood ω around z_0 such that $\zeta_1(z) \geq \epsilon$ for $z \in \omega$. We claim that this is impossible. Indeed, let φ be defined on ω by $\varphi(z) = 1/\zeta_1(z)$. Now

$$\varphi(z) = 1/\inf_{\|\xi\|=1, \xi \in \mathcal{H}} \|(T - z)\xi\|,$$

so $T - z$ is bounded from below by ϵ for $z \in \omega$. Let $\mathcal{H}_1 = \mathrm{ran}(T - z_0)$ and let $\widetilde{\mathcal{H}}$ be an infinite dimensional Hilbert space. Choose an isomorphism $V : \widetilde{\mathcal{H}} \to \mathcal{H}_1^{\perp} \oplus \widetilde{\mathcal{H}}$, and define the following operator

$$\widetilde{T}_c = (T - z_0) \oplus cV : \mathcal{H} \oplus \widetilde{\mathcal{H}} \to \mathcal{H}_1 \oplus \mathcal{H}_1^{\perp} \oplus \widetilde{\mathcal{H}}, \qquad c \in \mathbb{R}.$$

Note that $\widetilde{T}_c$ is invertible and for sufficiently large c we have $\varphi(z_0) = 1/\inf_{\|\xi\|=1, \xi \in \mathcal{H}} \|\widetilde{T}\xi\|$. Moreover, for z sufficiently close to z_0 it follows that

$$\varphi(z) = 1/\inf_{\|\xi\|=1, \xi \in \mathcal{H}} \|\widetilde{T}_c - (z_0 - z)\xi\|.$$

Let $G(z)$ be the inverse of $\widetilde{T}_c - (z_0 - z)$ for z in a neighborhood $\tilde{\omega}$ around z_0. Then $\varphi(z) = \|G(z)\|$. Now $\varphi(z_0) = 1/\epsilon$ and $\varphi(z) \leq 1/\epsilon$ for $z \in \tilde{\omega}$. But, clearly, $G'(z)$ is invertible for all $z \in \tilde{\omega}$ so by Theorem 3.3.2 it follows that $\|G(z)\| < 1/\epsilon$ for $z \in \tilde{\omega}$, contradicting $\varphi(z_0) = 1/\epsilon$ and we have shown (3.6.1). To show (iii) one argues almost exactly as in the proof of (ii).

We will now prove (iv). Firstly, to see the fact that $d_H(\sigma_{\mathrm{res},\epsilon}(T_k) \cap K, \sigma_{\mathrm{res},\epsilon}(T) \cap K) \to 0$, as $k \to \infty$, define Φ_0 as in Definition 3.3.3 and let $\zeta_{1,k}(z) = \Phi_0(T_k, z)$. Note that $\zeta_{1,k} \to \zeta_1$ locally uniformly as $k \to \infty$, by reasoning as in (3.3.5). Secondly, note that, for $\delta \in (0, \epsilon)$, we have

$$\mathrm{cl}(\{z \in \mathbb{C} : \zeta_1(z) > \epsilon,\, \zeta_2(z) \leq \delta\}) = \mathrm{cl}(\{z \in \mathbb{C} : \zeta_1(z) > \epsilon,\, \zeta_2(z) = 0\}).$$

So if we define $\zeta_{2,k}(z) = \Phi_0(T_k^*, \bar{z})$, it suffices to show that

$$d_H(\mathrm{cl}(\{z \in \mathbb{C} : \zeta_{1,k}(z) > \epsilon\}) \cap K, \mathrm{cl}(\{z \in \mathbb{C} : \zeta_1(z) > \epsilon\}) \cap K) \longrightarrow 0, \quad k \to \infty \tag{3.6.2}$$

and, by (ii), that $d_H(\{z \in \mathbb{C} : \zeta_{2,k}(z) \leq \delta\} \cap K, \{z \in \mathbb{C} : \zeta_2(z) \leq \delta\} \cap K) \to 0$ as $k \to \infty$. The latter follows from arguing similarly to the proof of Theorem 3.2.2, and hence we will concentrate on the former. Now, it is easy to see, by the definition of the Hausdorff metric and (ii), that (3.6.2) follows if we can show that

$$d_H(\{z \in \mathbb{C} : \zeta_{1,k}(z) \leq \epsilon\}, \{z \in \mathbb{C} : \zeta_1(z) \leq \epsilon\}) \longrightarrow 0, \quad k \to \infty,$$

but the latter follows by the locally uniform convergence of $\{\zeta_{1,k}\}$ and Proposition 3.3.1. Also, (v) follows by similar reasoning, and we are done. □

Theorem 3.6.3. *Let $\{e_j\}_{j\in\mathbb{N}}$ be a basis for $\mathcal{H}$ and define $\Xi_1, \Xi_2 : \mathcal{B}(\mathcal{H}) \to \Omega$, for $\epsilon > 0$, by $\Xi_1(T) = \mathrm{cl}(\sigma_{\mathrm{res},\epsilon}(T))$ and $\Xi_2(T) = \mathrm{cl}(\sigma_{\mathrm{res}^*,\epsilon}(T))$. Then $C_{\mathrm{ind}}(\Xi_1) \leq 2$ and $C_{\mathrm{ind}}(\Xi_2) \leq 2$.*

Proof. To show that $C_{\mathrm{ind}}(\Xi_1) \leq 2$ let Θ_k be defined as in (3.5.4) and define the estimating functions Γ_{n_1,n_2} and Γ_{n_1} in the following way. Define P_n to be the projection onto $\mathrm{span}\{e_1, \ldots, e_n\}$, choose $\delta \in (0, \epsilon)$ and define

$$\begin{aligned}
\Gamma_{n_1,n_2}(\{x_{ij}\}) &= \{z \in \Theta_{n_2} : \exists\, L \in LT_{\mathrm{pos}}(P_{n_1}\mathcal{H}), T_{\epsilon,n_1,n_2}(z) = LL^*\} \\
&\qquad \cap \{z \in \Theta_{n_2} : \nexists\, L \in LT_{\mathrm{pos}}(P_{n_1}\mathcal{H}), \widetilde{T}_{\delta,n_1,n_2}(z) = LL^*\}, \\
\Gamma_{n_1}(\{x_{ij}\}) &= \mathrm{cl}(\{z \in \mathbb{C} : (-\infty, 0] \cap \sigma(T_{\epsilon,n_1}(z)) = \emptyset\}) \\
&\qquad \cap \{z \in \mathbb{C} : (-\infty, 0] \cap \sigma(\widetilde{T}_{\delta,n_1}(z)) \neq \emptyset\},
\end{aligned}$$

where T_{ϵ,n_1,n_2}, $\widetilde{T}_{\delta,n_1,n_2}$, T_{ϵ,n_1} and $\widetilde{T}_{\delta,n_1}$ as defined as in (3.5.6). As the rest of the proof is just epsilon away from the proof of Theorem 3.2.2 we will just sketch the ideas. By letting $\zeta_{1,n_1}(z) = \Phi_{0,n_1}(T, z)$, $\zeta_{2,n_1}(z) = \Phi_{0,n_1}(T^*, \bar{z})$ and

$$\zeta_{1,n_1,n_2}(z) = \Phi_{0,n_1}(P_{n_2}TP_{n_2}, z), \quad \zeta_{2,n_1,n_2}(z) = \Phi_{0,n_1}(P_{n_2}T^*P_{n_2}, \bar{z}),$$

where Φ_0 is defined as in Definition 3.3.3, one observes that

$$\begin{aligned}
&\{z \in \Theta_{n_2} : \zeta_{1,n_1,n_2}(z) > \epsilon, \zeta_{2,n_1,n_2}(z) \leq \delta\} \\
&\qquad = \{z \in \mathbb{C} : \exists\, L \in LT_{\mathrm{pos}}(P_{n_1}\mathcal{H}), T_{\epsilon,n_1,n_2}(z) = LL^*\} \\
&\qquad\quad \cap \{z \in \mathbb{C} : \nexists\, L \in LT_{\mathrm{pos}}(P_{n_1}\mathcal{H}), \widetilde{T}_{\delta,n_1,n_2}(z) = LL^*\},
\end{aligned} \tag{3.6.3}$$

and

$$\Gamma_{n_1}(\{x_{ij}\}) = \mathrm{cl}(\{z : \zeta_{1,n_1}(z) > \epsilon, \zeta_{2,n_1}(z) \leq \delta\}).$$

Now, let ζ_1 and ζ_2 be defined as in Definition 3.6.1. By using (ii) in Theorem 3.6.2 and reasoning as in the proof of Theorem 3.2.2 (StepI and StepII) using arguments similar to the last part of the proof of Theorem 3.6.2 one deduces that, for compact ball $K \subset \mathbb{C}$ with K^o intersecting the appropriate sets,

$$\begin{aligned}
\mathrm{cl}(\{z \in \mathbb{C} : \zeta_{1,n_1}(z) > \epsilon\}) \cap K &\longrightarrow \mathrm{cl}(\{z \in \mathbb{C} : \zeta_1(z) > \epsilon\}) \cap K \\
\{z \in \mathbb{C} : \zeta_{2,n_1}(z) \leq \delta\} \cap K &\longrightarrow \{z \in \mathbb{C} : \zeta_2(z) \leq \delta\} \cap K, \quad n_1 \to \infty,
\end{aligned}$$

$$\begin{aligned}
\{z \in \Theta_{n_2} : \zeta_{1,n_1,n_2}(z) > \epsilon\} \cap K &\longrightarrow \mathrm{cl}(\{z \in \mathbb{C} : \zeta_{1,n_1}(z) > \epsilon\}) \cap K \\
\{z \in \Theta_{n_2} : \zeta_{2,n_1,n_2}(z) \leq \delta\} \cap K &\longrightarrow \{z \in \mathbb{C} : \zeta_{2,n_1}(z) \leq \delta\} \cap K, \quad n_2 \to \infty,
\end{aligned}$$

hence

$$\mathrm{cl}(\{z \in \mathbb{C} : \zeta_{1,n_1}(z) > \epsilon, \zeta_{2,n_1}(z) \leq \delta\}) \cap K \longrightarrow \mathrm{cl}(\{z : \zeta_1(z) > \epsilon, \zeta_2(z) \leq \delta\}) \cap K$$

as $n_1 \to \infty$, and

$$\{z \in \Theta_{n_2} : \zeta_{1,n_1,n_2}(z) > \epsilon, \zeta_{2,n_1,n_2}(z) \leq \delta\} \cap K \\ \longrightarrow \mathrm{cl}(\{z \in \mathbb{C} : \zeta_{1,n_1}(z) > \epsilon, \zeta_{2,n_1,n_2}(z) \leq \delta\}) \cap K$$

as $n_2 \to \infty$. But

$$\mathrm{cl}(\{z : \zeta_1(z) > \epsilon, \zeta_2(z) \leq \delta\}) = \mathrm{cl}(\{z : \zeta_1(z) > \epsilon, \zeta_2(z) = 0\}) = \mathrm{cl}(\sigma_{\mathrm{res},\epsilon}(T)),$$

and hence we have shown that $C_{\mathrm{ind}}(\Xi_1) \leq 2$. The fact that $C_{\mathrm{ind}}(\Xi_2) \leq 2$ follows by similar reasoning. □

3.7 Applications to Schrödinger and Dirac Operators

Non-Hermitian quantum mechanics has been an increasingly popular field in the last decades (TE05). As the importance of non-hermitian operators in physics has been established, the spectral theory of such operators has been given a substantial amount of attention (Dav99), (Dav02), (DK04). Since the spectral theory of non-hermitian operators is very different from the self-adjoint case, very little is known in general, and the same is true for the theory of approximating spectra. In fact it is an open problem how to approximate the spectrum and the pseudospectrum of an arbitrary Schrödinger operator. In this section we will show how to use the theory from the previous sections to get some insight on how to estimate spectra and pseudospectra of non-hermitian Schrödinger and Dirac operators with bounded potential. Let

$$P_j = -i\frac{\partial}{dx_j} \qquad Q_j = \text{multiplication by } x_j$$

with their appropriate domains in $\mathcal{H} = L^2(\mathbb{R}^d)$. Let $v \in L^\infty(\mathbb{R}^d)$ be a complex valued, continuous function, and define the Schrödinger operator

$$H = \frac{1}{2}\sum_{1 \leq j \leq d} P_j^2 + v(Q_1, \ldots, Q_d), \quad \mathcal{D}(H) = W_{2,2}(\mathbb{R}^d),$$

where $W_{2,2}(\mathbb{R}^d)$ is the Sobolev space of functions whose second derivative (in the distributional sense) is square integrable.

Similarly we can define the Dirac operator. Let $\mathcal{H} = \bigoplus_{k=1}^4 L^2(\mathbb{R}^3)$ and define (formally) $\tilde{P}_j$ on $\mathcal{H}$ by

$$\tilde{P}_j = \bigoplus_{k=1}^4 P_j, \qquad P_j = -i\frac{\partial}{dx_j}, \qquad j = 1, 2, 3,$$

where P_j is formally defined on $L^2(\mathbb{R}^3)$. Let

$$H_0 = \sum_{j=1}^3 \alpha_j \tilde{P}_j + \beta,$$

where α_j and β are 4-by-4 matrices satisfying the commutation relation

$$\alpha_j\alpha_k + \alpha_k\alpha_j = 2\delta_{jk}I, \qquad j,k = 1,2,3,4, \qquad \alpha_4 = \beta. \tag{3.7.1}$$

Then it is well known that H_0 is self-adjoint on $\bigoplus_{k=1}^4 W_{2,1}(\mathbb{R}^3)$ where

$$W_{2,1}(\mathbb{R}^3) = \{f \in L^2(\mathbb{R}^3) : \mathcal{F}f \in L_1^2(\mathbb{R}^3)\}$$

and $L_1^2(\mathbb{R}^3) = \{f \in L^2(\mathbb{R}^3) : (1+|\cdot|^2)^{1/2}f \in L^2(\mathbb{R}^3)\}$. Let $v \in L^\infty(\mathbb{R}^d)$ and define the Dirac operator

$$H_D = H_0 + \bigoplus_{k=1}^4 v(Q_1,Q_2,Q_3), \qquad \mathcal{D}(H) = \bigoplus_{k=1}^4 W_{2,1}(\mathbb{R}^3).$$

Note that H is closed since v is bounded. It is easy to see that

$$H^* = \frac{1}{2}\sum_{1\le j\le d} P_j^2 + \bar{v}(Q_1,\dots,Q_d), \quad \mathcal{D}(H^*) = W_{2,2}(\mathbb{R}^d)$$

and

$$H_D^* = H_0 + \bigoplus_{k=1}^4 \bar{v}(Q_1,Q_2,Q_3), \qquad \mathcal{D}(H_D^*) = \bigoplus_{k=1}^4 W_{2,1}(\mathbb{R}^3).$$

Thus, in order to estimate the pseudospectra of H and H_D, we may follow the ideas in the proof of Theorem 3.5.6. We will give a description of this for H and note that the procedure is exactly the same for H_D. Choose an orthonormal basis $\{\varphi_j\}$ for $W_{2,2}(\mathbb{R}^d)$ and let P_n be the projection onto $\operatorname{span}\{\varphi_j\}_{j=1}^n$. Now let $\{x_{ij}\}$ be defined by $x_{ij} = \langle H\varphi_j, \varphi_i\rangle$ and note that if we let $\tilde{x}_{ij} = \langle H^*\varphi_j, \varphi_i\rangle$ then $\tilde{x}_{ij} = \bar{x}_{ji}$. This allows us to define the set of estimating functions in the following way. Let $\epsilon > 0$ and define

$$\begin{aligned}\Gamma_{n_1,n_2}(\{x_{ij}\}) = \{z \in \Theta_{n_2} : \nexists\, L \in LT_{\mathrm{pos}}(P_{n_1}\mathcal{H}), T_{\epsilon,n_1,n_2}(z) = LL^*\}\\ \cup \{z \in \Theta_{n_2} : \nexists\, L \in LT_{\mathrm{pos}}(P_{n_1}\mathcal{H}), \widetilde{T}_{\epsilon,n_1,n_2}(z) = LL^*\}\end{aligned}$$

and

$$\Gamma_{n_1}(\{x_{ij}\}) = \{z \in \mathbb{C} : (-\infty,0] \cap \sigma(T_{\epsilon,n_1}(z)) \neq \emptyset\} \cup \{z \in \mathbb{C} : (-\infty,0] \cap \sigma(\widetilde{T}_{\epsilon,n_1}(z)) \neq \emptyset\},$$

where where Θ_{n_2} is defined as in (3.5.4) and

$$T_{\epsilon,n_1,n_2}(z) = S_{n_1}(P_{n_2}HP_{n_2},z)^*S_{n_1}(P_{n_2}HP_{n_2},z) - \epsilon^2 I,$$

$$\widetilde{T}_{\epsilon,n_1,n_2}(z) = S_{n_1}(P_{n_2}H^*P_{n_2},z)^*S_{n_1}(P_{n_2}H^*P_{n_2},z) - \epsilon^2 I$$

and $T_{\epsilon,n_1}(z) = S_{n_1}(H,z)^*S_{n_1}(H,z) - \epsilon^2 I$, $\widetilde{T}_{\epsilon,n_1}(z) = \widetilde{S}_{n_1}(H^*,z)^*\widetilde{S}_{n_1}(H^*z) - \epsilon^2 I$, where $S_m : \Delta \times \mathbb{C} \to \mathcal{B}(P_m\mathcal{H},\mathcal{H})$ is defined by $S_m(T,z) = (T-z)P_m$ and Δ denotes the set of closed operators having $W_{2,2}(\mathbb{R}^d)$ as their domain. Arguing as in the proof of Theorem 3.5.6 one deduces that

$$\sigma_\epsilon(H) = \lim_{n_1\to\infty} \Gamma_{n_1}(\{x_{ij}\}), \qquad \Gamma_{n_1}(\{x_{ij}\}) = \lim_{n_2\to\infty} \Gamma_{n_1,n_2}(\{x_{ij}\}).$$

Hence we get the following corollaries to Theorem 3.5.6.

Corollary 3.7.1. *Let $\{\varphi_j\}_{j\in\mathbb{N}}$ be a (not necessarily orthogonal) basis for $W_{2,2}(\mathbb{R}^d)$ that is orthogonal in $L^2(\mathbb{R}^d)$ and let Δ denote the set of Schrödinger operators on $L^2(\mathbb{R}^d)$ with potential function in $L^\infty(\mathbb{R}^d)$. Let $\epsilon > 0$, $\Xi_1 : \Delta \to \Omega$ and $\Xi_2 : \Delta \to \Omega$ be defined by $\Xi_1(H) = \overline{\sigma_\epsilon(H)}$ and $\Xi_2(H) = \sigma(H)$. Then $C_{\mathrm{ind}}(\Xi_1) \leq 2$ and $C_{\mathrm{ind}}(\Xi_2) \leq 3$.*

Corollary 3.7.2. *Let $\{\varphi_j\}_{j\in\mathbb{N}}$ be a (not necessarily orthogonal) basis for $\bigoplus_{k=1}^4 W_{2,1}(\mathbb{R}^3)$ that is orthogonal in $\bigoplus_{k=1}^4 L^2(\mathbb{R}^3)$, and let Δ denote the set of Dirac operators on the Hilbert space $\bigoplus_{k=1}^4 L^2(\mathbb{R}^3)$ with bounded potential function. Let $\epsilon > 0$, $\Xi_1 : \Delta \to \Omega$ and $\Xi_2 : \Delta \to \Omega$ be defined by $\Xi_1(H_D) = \overline{\sigma_\epsilon(H_D)}$ and $\Xi_2(T) = \sigma(H_D)$. Then $C_{\mathrm{ind}}(\Xi_1) \leq 2$ and $C_{\mathrm{ind}}(\Xi_2) \leq 3$.*

Remark 3.7.3. As the proof of Theorem 3.5.6, and hence also the proofs of Corollaries 3.7.1 and 3.7.2, are constructive, we have a constructive way of recovering spectra and pseudospectra of a large class of important operators in mathematical physics and hence the previous results may have impact in applications.

Chapter 4

Convergence of Densities

We finish Part I by extending some of the results in (Arv94a) from bounded to unbounded operators and also to non-normal operators. In this section we change the point of view from single operators to algebras of operators. Let us recall some basics and useful facts.

By a state τ on a C^*-algebra $\mathcal{A}$ with identity we mean a positive linear functional on the positive elements of $\mathcal{A}$ such that $\tau(I) = 1$ (I denoting the identity). The state τ is tracial if $\tau(BB^*) = \tau(B^*B)$ for all positive $B \in \mathcal{A}$ and faithful if $B = 0$ when $\tau(B) = 0$.

Let $\mathcal{A} \subset \mathcal{B}(\mathcal{H})$ be a C^*-algebra with a unique tracial state. Then a self-adjoint operator $A \in \mathcal{A}$ determines a natural probability measure μ_A on $\mathbb{R}$ by

$$\int_{\mathbb{R}} f(x)\, d\mu_A(x) = \tau(f(A)), \quad f \in C_0(\mathbb{R}).$$

Also, if τ is faithful then $\mathrm{supp}(\mu_A) = \sigma(A)$ and one refers to μ_A as the spectral distribution. As we have seen above, we can approximate the spectrum of A by using the techniques demonstrated in Chapter 3. We now turn the attention to the task of approximating μ.

4.1 The Self-Adjoint Case

If $\mathcal{A} \subset \mathcal{B}(\mathcal{H})$ is a C^*-algebra with a unique, faithful tracial state and $A \in \mathcal{A}$, then $\mathrm{supp}(\mu_A) = \sigma(A)$. Thus, if $\{A_n\}$ is a sequence of self-adjoint elements in $\mathcal{A}$ converging in some sense to a self-adjoint element $A \in \mathcal{A}$ and we are interested in determining the behavior of $\sigma(A_n)$ as $n \to \infty$, the behavior of μ_{A_n} is of great interest. In particular, we consider under which conditions can we guarantee that

$$\int_{-\infty}^{\infty} f(x) d\mu_{A_n}(x) \longrightarrow \int_{-\infty}^{\infty} f(x) d\mu_A(x),$$

for all $f \in C_0(\mathbb{R})$.

As our goal is to extend some of the theorems in (Arv94a) from bounded to unbounded operators, the C^*-algebra framework sketched above must be modified slightly. Since collections of unbounded operators can never form a C^*-algebra we have to look at C^*-algebras affiliated with unbounded operators.

Definition 4.1.1. *Let A be a self-adjoint, unbounded operator on $\mathcal{H}$. The operator A is affiliated with the C^*-algebra $\mathcal{A}$ if and only if $\mathcal{A} \supset \{f(A) : f \in C_0(\mathbb{R})\}$.*

We will also be needing some preliminary theory.

Definition 4.1.2. *(i) A filtration of $\mathcal{H}$ is a sequence $\mathcal{F} = \{\mathcal{H}_1, \mathcal{H}_2, \ldots\}$ of finite dimensional subspaces of $\mathcal{H}$ such that $\mathcal{H}_n \subset \mathcal{H}_{n+1}$ and*

$$\overline{\bigcup_{n=1}^{\infty} \mathcal{H}_n} = \mathcal{H}.$$

(ii) Let $\mathcal{F} = \{\mathcal{H}_n\}$ be a filtration of $\mathcal{H}$ and let P_n be the projection onto $\mathcal{H}_n$. The degree of an operator $A \in \mathcal{B}(\mathcal{H})$ is defined by

$$deg(A) = \sup_{n \geq 1} \operatorname{rank}(P_n A - A P_n).$$

Definition 4.1.3. *Let $\mathcal{A} \subset \mathcal{B}(\mathcal{H})$ be a C^*-algebra. An $\mathcal{A}$-filtration is a filtration of $\mathcal{H}$ such that the $*$-subalgebra of all finite degree operators in $\mathcal{A}$ is norm dense in $\mathcal{A}$.*

Proposition 4.1.4. *(Arveson) Let $\mathcal{A} \subset \mathcal{B}(\mathcal{H})$ be a C^*-algebra with a unique tracial state τ and suppose that $\{\mathcal{H}_n\}$ is an $\mathcal{A}$-filtration. Let τ_n be the state of $\mathcal{A}$ defined by*

$$\tau_n(A) = \frac{1}{d_n} \operatorname{trace}(P_n A), \qquad d_n = \dim(\mathcal{H}_n).$$

Then

$$\tau_n(A) \to \tau(A), \qquad \text{for all} \quad A \in \mathcal{A}.$$

Proposition 4.1.5. *(Arveson) Let $\mathcal{F} = \{\mathcal{H}_1, \mathcal{H}_2, \ldots\}$ be a filtration of $\mathcal{H}$, let P_n be the projection onto $\mathcal{H}_n$ and let $A_1, A_2, \ldots, A_p$ be a finite set of operators in $\mathcal{B}(\mathcal{H})$. Then for every $n = 1, 2, \ldots$ we have*

$$\operatorname{trace}|P_n A_1 A_2 \ldots A_p P_n - P_n A_1 P_n A_2 P_n \ldots P_n A_p P_n| \leq \|A_1\| \ldots \|A_p\| \sum_{k=1}^{p} \deg A_k.$$

Now, suppose that $\mathcal{A} \subset \mathcal{B}(\mathcal{H})$ is a C^*-algebra with a unique tracial state τ and $\{P_n\}$ is an increasing sequence of finite rank projections on $\mathcal{H}$ converging strongly to the identity. Define the tracial state

$$\tau_n(B) = \frac{1}{d_n} \operatorname{trace}(P_n B), \qquad d_n = \dim(P_n \mathcal{H}), \quad B \in \mathcal{B}(\mathcal{H}).$$

Now τ_n restricts to the normalized trace on $P_n \mathcal{B}(\mathcal{H}) P_n$ and, similar to τ, induces a measure $\mu_{P_n A \lceil_{P_n \mathcal{H}}}$ on $\mathbb{R}$ such that

$$\int_{\mathbb{R}} f(x)\, d\mu_{P_n A \lceil_{P_n \mathcal{H}}}(x) = \tau_n(f(P_n A \lceil_{P_n \mathcal{H}})), \quad f \in C_0(\mathbb{R}). \tag{4.1.1}$$

The question is then: what is the relationship between $\mu_{P_n A \lceil_{P_n \mathcal{H}}}$ and μ_A. In particular, under which assumptions (if any) can one guarantee that

$$\mu_{P_n A \lceil_{P_n \mathcal{H}}} \overset{\text{weak}^*}{\longrightarrow} \mu_A, \quad n \to \infty.$$

This has been investigated in (Arv94a)(Béd97)(Han08). In particular using Proposition 4.1.4 and Proposition 4.1.5 Arveson showed that

Theorem 4.1.6. *(Arveson)(Arv94a) Let $\mathcal{A} \subset \mathcal{B}(\mathcal{H})$ be a C^*-algebra and let $\mathcal{F} = \{\mathcal{H}_n\}$ be an $\mathcal{A}$-filtration. For a self-adjoint operator $A \in \mathcal{A}$ denote the spectral distribution by μ_A and let $\mu_{P_nA\lceil_{P_n\mathcal{H}}}$ be defined as in (4.1.1). Then*

$$\mu_{P_nA\lceil_{P_n\mathcal{H}}} \overset{\text{weak}^*}{\longrightarrow} \mu_A, \qquad n \to \infty.$$

The next theorem will be crucial in the sequel and replaces Proposition 4.1.5 in our framework, which deviates from Arveson's theory in order to include unbounded operators. Firstly, some notation. We let trace denote the trace on the set of trace class operators and $\|\cdot\|_2$ denote the Hilbert-Schmidt norm. Let also W^2_∞ denote the Sobolev space of measurable functions on $\mathbb{R}$ with second derivative (in the distributional sense) being L^∞.

Theorem 4.1.7. *(Laptev, Safarov)(LS96) Let A be a self-adjoint, unbounded operator on $\mathcal{H}$ and let P be projection such that PA is a Hilbert-Schmidt operator. Then for any $\psi \in W^2_\infty$ we have that*

$$|\mathrm{tr}(P\psi(A)P - P\psi(PAP)P)| \leq \|\psi''\|_\infty \|PA(I-P)\|_2^2.$$

Note also that, if $\mathcal{A} \subset \mathcal{B}(\mathcal{H})$ is a C^*-algebra with a unique tracial state, the result discussed in the introduction to this chapter extends to unbounded operators, namely, if A is self-adjoint and affiliated with $\mathcal{A}$ then

$$\int_{\mathbb{R}} f(x)\, d\mu_A(x) = \tau(f(A)), \quad f \in C_0(\mathbb{R}),$$

where μ_A is a probability measure on $\mathbb{R}$. The next theorem is an extension of Theorem 4.1.6 (which is Theorem 4.5 in (Arv94a)) to unbounded operators.

Theorem 4.1.8. *Let A be a self-adjoint, unbounded operator with domain $\mathcal{D}(A)$ and let $\mathcal{A}$ be a C^*-algebra with a unique tracial state τ. Suppose that $\{\mathcal{H}_n\}$ is an $\mathcal{A}$-filtration, where $\mathcal{H}_n \subset \mathcal{D}(A)$, and that $\mathcal{A}$ is affiliated with A. Let $d_n = \dim(\mathcal{H}_n)$ and $\lambda_1, \lambda_2, \ldots, \lambda_{d_n}$ be the eigenvalues of $A_n = P_nA\lceil_{\mathcal{H}_n}$, repeated according to multiplicity. Suppose that one of the following is true.*

(i) $\|P_nA(I-P_n)\|_2/\sqrt{d_n} \to 0$, *as* $n \to \infty$.

(ii) $A = D + C$, *where D commutes with P_n and $C \in \tilde{\mathcal{A}} \subset \mathcal{B}(\mathcal{H})$ and $\tilde{\mathcal{A}}$ is a C^*-algebra such that $\{\mathcal{H}_n\}$ is also an $\tilde{\mathcal{A}}$-filtration.*

Then for every $f \in C_0(\mathbb{R})$,

$$\lim_{n\to\infty} \frac{1}{d_n}(f(\lambda_1) + f(\lambda_2) + \ldots + f(\lambda_{d_n})) = \int_{\mathbb{R}} f(x)\, d\mu_A(x),$$

where μ_A denotes the Borel measure induced by τ.

Proof. Define

$$\tau_n(T) = \frac{1}{d_n}\mathrm{trace}(P_nT), \quad T \in \mathcal{A}.$$

Since τ_n restricts to the normalized trace on $P_n\mathcal{B}(\mathcal{H})P_n$ and since, by Proposition 4.1.4

$$\tau_n(B) \longrightarrow \tau(B), \qquad n \to \infty, \quad B \in \mathcal{A}$$

it follows that, in both cases (i) and (ii), it suffices to show that

$$\tau_n(f(A)) - \tau_n(f(P_nAP_n)) \to 0, \quad n \to \infty. \tag{4.1.2}$$

To show this for (i), note that we can approximate f in the L^∞ norm by elements from W^2_∞. Combining that fact with the observation that the linear functional

$$f \mapsto \tau_n(f(A)) - \tau_n(f(P_nAP_n))$$

has norm less than two, we reduce the problem to showing (4.1.2) when $f \in W^2_\infty$. Now, by Theorem 4.1.7,

$$\begin{aligned}|\tau_n(f(A)) - \tau_n(f(P_nAP_n))| &= \frac{1}{d_n}|\mathrm{trace}(P_nf(A)P_n) - \mathrm{trace}(P_nf(P_nAP_n)P_n)| \\ &\leq \frac{1}{2d_n}\|f''\|_\infty \|P_nA(I-P_n)\|_2^2,\end{aligned}$$

where the right hand side of the inequality tends to zero by assumption.

To prove the theorem when (ii) is assumed, note that, by the Stone-Weierstrass theorem, polynomials in $(x+i)^{-1}$ and $(x-1)^{-1}$ are dense in $C_0(\mathbb{R})$. Thus, by arguing as above, we can assume that $f(x) = (x+i)^{-k}(x-i)^{-l}$ for some positive integers k, l. It is not too hard to show that $(D+C\pm i)^{-1} - (D+B\pm i)^{-1}$ is small when $\|C-B\|$ is small and $B \in \mathcal{B}(\mathcal{H})$ is self-adjoint. Thus, for $\epsilon > 0$ we have

$$\|f(P_n(D+C)P_n) - f(P_n(D+B)P_n)\| \leq \epsilon, \quad \|f(D+C) - f(D+B)\| \leq \epsilon,$$

for $B \in \tilde{\mathcal{A}}$ and when $\|C-B\|$ is sufficiently small. Hence, since τ_n is uniformly bounded, we can assume that C has finite degree. Arguing as above we get

$$\begin{aligned}|\tau_n(f(A)) - \tau_n(f(P_nAP_n))| &\leq \frac{1}{2d_n}\|f''\|_\infty \|P_n(D+C)(I-P_n)\|_2^2 \\ &\leq \frac{1}{2d_n}\|f''\|_\infty \deg(C)\|C\|^2,\end{aligned}$$

and this yields the assertion. The proof of the fact that $\|P_nC(I-P_n)\|_2^2 \leq \deg(C)\|C\|^2$ can be found in the proof of Lemma 3.6 in (Arv94a). □

4.2 The Non-Normal Case and the Brown Measure

Our next goal is to prove an analogue of Theorem 4.1.6 for non-normal operators. But as there is no spectral distribution for non-normal operators we first need to introduce the Brown measure. Let $\mathcal{M}$ be a finite von Neumann algebra of operators on $\mathcal{H}$ with a faithful, normal tracial state τ. Let $T \in \mathcal{M}$, then the Fuglede-Kadison determinant $\Delta(T)$ (FK52) is defined as

$$\Delta(T) = \exp\left(\int_0^\infty \log t \, d\mu_{|T|}(t)\right),$$

where

$$\mu_{|T|}(\omega) = \tau(E_{|T|}(\omega)), \qquad \omega \in \mathrm{Borel}(\mathbb{R}),$$

and $E_{|T|}$ denotes the spectral projection measure corresponding to $|T|$. Now define

$$f(z) = \log(\Delta(T - z)), \qquad z \in \mathbb{C}. \tag{4.2.1}$$

It can be shown (HS07) that f is subharmonic and therefore gives rise to a measure (see Section 3 in (HK76))

$$d\mu_T = \frac{1}{2\pi}\nabla^2 f \, dm,$$

where m denotes the Lebesgue measure on $\mathbb{R}^2$ and $\nabla^2 f$ is understood to be in the distributional sense i.e. $\int \varphi \, d\mu_T = \int f \nabla^2 \varphi \, dm$, for $\varphi \in C_c^\infty(\mathbb{R}^2)$. The measure μ_T satisfies $\text{supp}(\mu_T) \subset \sigma(T)$ and is often referred to as Brown's spectral distribution measure. Now the inclusion $\text{supp}(\mu_T) \subset \sigma(T)$ can be proper, but (by Remark 4.4 in (Bro86)) if $\lambda \in \sigma(T)$ is isolated then $\mu_T(\{\lambda\}) \neq 0$. Thus, knowing μ_T would be a nice tool for locating isolated eigenvalues of T.

Note that if $\mathcal{M}$ is normal, then $\mu_T = \tau \circ E_T$, and also, if $\mathcal{M} = M_n(\mathbb{C})$ for some $n \in \mathbb{N}$ then the Fuglede-Kadison determinant and the Brown measure is defined for $T \in \mathcal{M}$ and

$$\Delta(T) = |\text{det}T|^{\frac{1}{n}}, \qquad \mu_T = \frac{1}{n}\sum_{j=1}^{n} \delta_{\lambda_j},$$

where δ_{λ_j} denotes the point measure at λ_j and $\lambda_1, \ldots, \lambda_n$ are the eigenvalues of T, repeated according to multiplicity.

Our approach is to extend Arveson's ideas regarding approximating the spectral distribution of self-adjoint operators to Browns spectral distribution. Let $\mathcal{F}$ be a filtration with corresponding projections $\{P_n\}$, and define the tracial state

$$\tau_n(B) = \frac{1}{d_n}\text{trace}(P_n B), \quad B \in \mathcal{B}(\mathcal{H}), \quad d_n = \dim(P_n\mathcal{H}).$$

In order to approximate f defined in (4.2.1), it could be tempting to define, for $z \in \mathbb{C}$ and $T \in \mathcal{B}(\mathcal{H})$, a measure by

$$\mu_{|P_n(T-z)P_n|}(\omega) = \tau_n(E_{|P_n(T-z)P_n|}(\omega)), \qquad \omega \in \text{Borel}(\mathbb{R}),$$

but knowing how bad the spectrum of P_nTP_n may approximate $\sigma(T)$ when T is non-self-adjoint we abandon that idea immediately and instead define the measure $\mu_{T,z,n}$ by

$$\mu_{T,z,n}(\omega) = \tau_n(E_{P_n(T-z)^*(T-z)\lceil_{P_n\mathcal{H}}}(\omega)), \quad \omega \in \text{Borel}(\mathbb{R}). \tag{4.2.2}$$

Using this measure we obtain the following results.

Theorem 4.2.1. *Let $\mathcal{M}$ be a finite Von Neumann algebra with a unique, faithful, normal tracial state τ. Suppose that $\mathcal{A} \subset \mathcal{M}$ is a C^*-alebra and that $\{\mathcal{H}_n\}$ is an $\mathcal{A}$-filtration with corresponding projections $\{P_n\}$. Define the tracial state τ_n by*

$$\tau_n(B) = \frac{1}{d_n}\text{trace}(P_n B), \quad B \in \mathcal{M}, \quad d_n = \dim(P_n\mathcal{H}).$$

For $z \in \mathbb{C}$ and $T \in \mathcal{A}$, define the measure $\mu_{T,z,n}$ as in (4.2.2). Let

$$f_n(z) = \frac{1}{2}\int_0^\infty \log t \, d\mu_{T,z,n}(t)$$

and $d\,\nu_n = \frac{1}{2\pi}\nabla^2 f_n d\,m$, *where* m *is Lebesgue measure on* $\mathbb{R}^2$. *Then* ν_n *defines a positive Borel measure on* $\mathbb{R}^2$ *satisfying* $\nu_n(\mathbb{C}) \leq 1$. *Moreover, there exists a positive Borel measure* ν *on* $\mathbb{R}^2$ *with* $\mathrm{supp}(\nu) \subset \sigma(T)$ *and a subsequence* $\{\nu_{n_k}\}$ *such that*

$$\nu_{n_k} \overset{\text{weak}^*}{\longrightarrow} \nu, \qquad k \to \infty.$$

Theorem 4.2.2. *Suppose the assumptions in Theorem 4.2.1 are true and that* $T \in \mathcal{A}$.

(i) Then, if $\rho : \mathbb{C} \to \mathbb{C}$ *defined by*

$$\rho(z) = \begin{cases} \log(1/\|(T-z)^{-1}\|) & z \in \mathbb{C} \setminus \sigma(T) \\ -\infty & z \in \sigma(T) \end{cases}$$

is locally integrable, the measure ν *from Theorem 4.2.1 is equal to the Brown measure* μ_T, *and*

$$\nu_n \overset{\text{weak}^*}{\longrightarrow} \mu_T, \qquad n \to \infty,$$

where ν_n *is defined as in Theorem 4.2.1.*

(ii) Suppose that $\omega \subset \mathbb{C}$ *is an open set such that* $\omega \cap \sigma(T) = \{\lambda_1, \ldots \lambda_k\}$, *where* λ_j *is an isolated eigenvalue. Suppose also that there is an* $\alpha > 0$ *such that*

$$\inf_{z \in \partial D(\lambda_j, r)} 1/(\|(T-z)^{-1}\|) \geq r^\alpha$$

for all sufficiently small r, *where* $D(\lambda_j, r)$ *denotes the disk with center* λ_j *and radius* r. *Then*

$$\nu_n \lceil_\omega \overset{\text{weak}^*}{\longrightarrow} \mu_T \lceil_\omega, \qquad n \to \infty.$$

If one actually wanted to use the Brown measure μ_T to estimate the position of the isolated eigenvalues one is faced with the task of evaluating an integral of the form

$$\int_{\mathbb{R}^2} f \nabla^2 \varphi \, dm. \qquad \varphi \in C_c^\infty. \tag{4.2.3}$$

As we may not know f explicitly this may not be possible. However, an approximation may help us. Now suppose that we have established that $\nu_n \to \mu_T$ (weak*) as $n \to \infty$, where ν_n is defined as in Theorem 4.2.1, we could approximate (4.2.3) by

$$\int_{\mathbb{R}^2} f_n \nabla^2 \varphi \, dm, = \int_{\mathbb{R}^2} \frac{1}{2n} \mathrm{trace}(\log(P_n(T-z)^*(T-z)P_n)\nabla^2\varphi(z)\, dm(z). \tag{4.2.4}$$

Note that as P_n has finite rank, it may be possible to evaluate f_n on a discrete set of points in $\mathbb{R}^2$ and use quadrature formulae to approximate (4.2.4).

Proof. (Proof of Theorem 4.2.1) The proof will be done in several steps.

Step I. We first need to show that ν_n indeed is a positive Borel measure. To prove that, it suffices, by Lemma 3.6 and Section 3.5 in (HK76), to demonstrate that f_n is subharmonic. To do that, let $\epsilon > 0$ and define

$$g_{n,\epsilon}(z) = \frac{1}{2}\tau_n(\log(P_n(T-z)^*(T-z)P_n + \epsilon I)).$$

We claim that $g_{n,\epsilon}$ is subharmonic. The method we use here is quite close to the techniques used in (HS07). Note that $g_{n,\epsilon}$ is infinitely smooth. Indeed, since

$$z \mapsto P_n(T-z)^*(T-z)P_n + \epsilon I$$

is obviously infinitely smooth and so is log on $\{z : \mathfrak{Re} z \geq \epsilon\}$ so

$$z \mapsto \operatorname{trace}\left(\log(P_n(T-z)^*(T-z)P_n + \epsilon I)\lceil_{P_n\mathcal{H}}\right)$$

is infinitely smooth, thus $g_{n,\epsilon}$ is infinitely smooth. Thus, we need to show that $\nabla^2 g_{n,\epsilon} = 0$. This we will do using brute force computations. Using the standard notation

$$\frac{\partial}{\partial\lambda} = \frac{1}{2}\left(\frac{\partial}{\partial\lambda_1} - i\frac{\partial}{\partial\lambda_2}\right) \quad \text{and} \quad \frac{\partial}{\partial\bar{\lambda}} = \frac{1}{2}\left(\frac{\partial}{\partial\lambda_1} + i\frac{\partial}{\partial\lambda_2}\right)$$

and letting $z = \lambda_1 + i\lambda_2$ we have

$$\nabla^2 g_{n,\epsilon} = \left(\frac{\partial^2}{\partial\lambda_1^2} + \frac{\partial^2}{\partial\lambda_2^2}\right) g_{n,\epsilon} = 4\frac{\partial^2}{\partial\bar{\lambda}\partial\lambda} g_{n,\epsilon}.$$

Let $\varphi(z) = P_n(T-z)^*(T-z)P_n + \epsilon I$. By the definition of the derivative, linearity and boundedness of τ_n we have that

$$\frac{\partial^2 g_{n,\epsilon}}{\partial\bar{\lambda}\partial\lambda} = \frac{1}{2}\frac{\partial^2\tau_n(\log\circ\varphi)}{\partial\bar{\lambda}\partial\lambda} = \frac{1}{2}\tau_n\left(\frac{\partial^2\log\circ\varphi}{\partial\bar{\lambda}\partial\lambda}\right)$$

so it is straightforward to show that

$$\begin{aligned}\frac{\partial^2 g_{n,\epsilon}}{\partial\bar{\lambda}\partial\lambda} &= \frac{1}{2}\tau_n\left(-\varphi^{-1}\frac{\partial\varphi}{\partial\bar{\lambda}}\varphi^{-1}\frac{\partial\varphi}{\partial\lambda} + \varphi^{-1}\frac{\partial^2\varphi}{\partial\bar{\lambda}\partial\lambda}\right)\\ &= \frac{1}{2}\tau_n\left(\varphi^{-1/2}\left(-\frac{\partial\varphi}{\partial\bar{\lambda}}\varphi^{-1}\frac{\partial\varphi}{\partial\lambda} + \frac{\partial^2\varphi}{\partial\bar{\lambda}\partial\lambda}\right)\varphi^{-1/2}\right).\end{aligned}$$

Thus, it suffices to show that $-\frac{\partial\varphi}{\partial\bar{\lambda}}\varphi^{-1}\frac{\partial\varphi}{\partial\lambda} + \frac{\partial^2\varphi}{\partial\bar{\lambda}\partial\lambda}$ is positive. Now,

$$\frac{\partial\varphi}{\partial\lambda} = -P_n(T-z)^*P_n, \qquad \frac{\partial\varphi}{\partial\bar{\lambda}} = -P_n(T-z)P_n, \qquad \frac{\partial^2\varphi}{\partial\bar{\lambda}\partial\lambda} = P_n.$$

Thus, we can compute

$$\begin{aligned}&-\frac{\partial\varphi}{\partial\bar{\lambda}}\varphi^{-1}\frac{\partial\varphi}{\partial\lambda} + \frac{\partial^2\varphi}{\partial\hat{\lambda}\partial\lambda}\\ &\quad= -P_n(T-z)P_n(P_n(T-z)^*(T-z)P_n + \epsilon I)^{-1}P_n(T-z)^*P_n + P_n\\ &\quad= -P_nB(B^*B + \epsilon I)^{-1}B^*P_n + P_n, \qquad B = (T-z)P_n\\ &\quad= -P_n((BB^* + \epsilon I)^{-1}BB^* + I)P_n\\ &\quad= -P_n(-\epsilon(BB^* + \epsilon I)^{-1})P_n\\ &\quad= \epsilon P_n((T-z)P_n(T-z)^* + \epsilon I)^{-1}P_n,\end{aligned} \tag{4.2.5}$$

which is clearly positive. Observe also that

$$f_n(z) = \frac{1}{2}\tau_n(\log(P_n(T-z)^*(T-z)P_n)) = \frac{1}{2}\int_0^\infty \log t \, d\mu_{T,z,n}(t)$$

and

$$g_{n,\epsilon}(z) = \frac{1}{2}\int_0^\infty \log(t+\epsilon)\, d\mu_{T,z,n}(t).$$

In particular $g_{n,\epsilon}$ decreases pointwise to f_n as $\epsilon \to 0$. Thus, f_n must be subharmonic or identically $-\infty$. But $f_n(z) > -\infty$ for $z \notin \sigma(T)$, and thus f_n must be subharmonic.

Step II. We will now show that $\nu_n(\mathbb{C}) \leq 1$ for all n. Define

$$\psi_R(z) = \begin{cases} \log R & |z| \leq 1 \\ \log(\frac{R}{|z|}) & 1 < |z| < R \\ 0 & |z| \geq R. \end{cases}$$

Then, since $\frac{1}{\log R}\psi_R$ increases monotonically to 1, it follows by monotone convergence that

$$\nu_n(\mathbb{C}) = \lim_{R\to\infty} \int_{\mathbb{C}} \frac{1}{\log R}\psi_R \, d\nu_n.$$

Now, by Lemma 2.12 in (HS07) it is true that

$$\int_{\mathbb{C}} \frac{1}{\log R}\psi_R \, d\nu_n = \frac{1}{\log R}\left(\frac{1}{2\pi}\left(\int_0^{2\pi} f_n(Re^{i\theta})\, d\theta - \int_0^{2\pi} f_n(e^{i\theta})\, d\theta\right)\right).$$

Thus, it suffices to show that $\lim_{R\to\infty} \frac{1}{2\pi \log R}(\int_0^{2\pi} f_n(Re^{i\theta})\, d\theta \leq 1$. Now,

$$\begin{aligned} \frac{1}{2\pi \log R}\left(\int_0^{2\pi} f_n(Re^{i\theta})\right) d\theta &= \frac{1}{4\pi \log R}\left(\int_0^{2\pi} \tau_n(\log(|P_n(T-Re^{i\theta})^*(T-Re^{i\theta})P_n|))\right) d\theta \\ &\leq \frac{1}{2\log R}\|\tau_n\| \log\left(\sup_{\theta\in[0,2\pi]} \||P_n(T-Re^{i\theta})^*(T-Re^{i\theta})P_n|\|\right) \\ &\leq \frac{1}{2\log R}\log((\|T\|+R)^2) \longrightarrow 1, \quad R\to\infty. \end{aligned}$$

Step III. The existence of ν now follows from the weak* compactness of the unit ball of $C_0(\mathbb{C})^*$ since we have proved in Step II that $\{\nu_n\}$ is uniformly bounded as elements in $C_0(\mathbb{C})^*$.

We are left with the task of proving that

$$\operatorname{supp}(\nu) \subset \sigma(T), \tag{4.2.6}$$

and this will be done in Step IV and V.

Step IV. We will show that $f_n(z) \to f(z)$ when $z \notin \sigma(T)$ and f is defined in (4.2.1). To prove that we need to demonstrate that

$$\lim_{n\to\infty} \frac{1}{2}\int_0^\infty \log t \, d\mu_{T,z,n}(t) = \int_0^\infty \log t \, d\mu_{|(T-z)|}(t), \quad z \notin \sigma(T). \tag{4.2.7}$$

Before we can prove (4.2.7) we need the following observation. Note that since $z \notin \sigma(T)$ then there is an $\epsilon > 0$ and $M < \infty$ such that

$$\sigma(|T-z|^2) \subset [\epsilon, M], \qquad \sigma(P_n(T-z)^*(T-z)\lceil_{P_n\mathcal{H}}) \subset [\epsilon, M]. \tag{4.2.8}$$

Indeed, letting

$$\epsilon = (\inf_{\|\xi\|=1, \xi\in\mathcal{H}} \langle (T-z)^*(T-z)\xi, \xi\rangle)^{1/2}$$

and

$$\epsilon_n = (\inf_{\|\xi\|=1, \xi\in\mathcal{H}} \langle (P_n(T-z)^*(T-z)P_n\xi, \xi\rangle)^{1/2}$$

then $\sigma(|T-z|) \subset [\epsilon, \infty)$ and $\sigma(P_n(T-z)^*(T-z)P_n) \subset [\epsilon_n, \infty)$ so

$$\begin{aligned}\mu_{|T-z|}([0,\epsilon)) &= \tau(E_{|T-z|}([0,\epsilon))) = 0\\ \mu_{T,z,n}([0,\epsilon)) &= \tau_n(E_{|P_n(T-z)^*(T-z)P_n}([0,\epsilon_n))) = 0,\end{aligned}$$

since $(E_{|T-z|}([0,\epsilon)) = E_{P_n(T-z)^*(T-z)P_n}([0,\epsilon_n)) = 0$. Also,

$$\begin{aligned}\epsilon_n &= (\inf_{\|\xi\|=1, \xi\in\mathcal{H}} \langle (P_n(T-z)^*(T-z)P_n\xi, \xi\rangle)^{1/2}\\ &= (\inf_{\|\xi\|=1, \xi\in\mathcal{H}_n} \langle (T-z)^*(T-z)\xi, \xi\rangle)^{1/2}\\ &\geq (\inf_{\|\xi\|=1, \xi\in\mathcal{H}} \langle (T-z)^*(T-z)\xi, \xi\rangle)^{1/2}\\ &= \epsilon.\end{aligned}$$

Thus, since

$$\epsilon = (\inf_{\|\xi\|=1, \xi\in\mathcal{H}} \langle (T-z)^*(T-z)\xi, \xi\rangle)^{1/2} = 1/\|(T-z)^{-1}\| > 0$$

and T is bounded then (4.2.8) follows. We can now return to the task of proving (4.2.7). Now, using (4.2.8), we have that

$$\begin{aligned}f_n(z) &= \frac{1}{2}\int_0^\infty \log t \, d\mu_{T,z,n}(t) = \tau_n(\chi_{[\epsilon,M]} \log \circ g(P_n(T-z)^*(T-z)\lceil_{P_n\mathcal{H}}))\\ f(z) &= \int_0^\infty \log t \, d\mu_{|T-z|}(t) = \tau(\chi_{[\epsilon,M]} \log \circ g((T-z)^*(T-z))),\end{aligned}$$

where $g(t) = \sqrt{t}$, $t \in [0, \infty)$. Thus, we are left with the task of showing that

$$\lim_{n\to\infty} \tau_n((\chi_{[\epsilon,M]} \log \circ g(P_n(T-z)^*(T-z)\lceil_{P_n\mathcal{H}})) = \tau((\chi_{[\epsilon,M]} \log \circ g)((T-z)^*(T-z))).$$

But, by the uniqueness of τ and Proposition 4.1.4 we have that

$$\lim_{n\to\infty} \tau_n(B) = \tau(B), \qquad B \in \mathcal{A},$$

thus our problem is reduced to showing

$$\begin{aligned}\lim_{n\to\infty} |\tau_n((\chi_{[\epsilon,M]} \log \circ\, g((T-z)^*(T-z)))&\\ - \tau_n((\chi_{[\epsilon,M]} \log \circ g)(P_n(T-z)^*(T-z)P_n))| &= 0.\end{aligned} \tag{4.2.9}$$

Thus, by the fact that the norm of the linear functionals

$$f \in C[\epsilon, M] \mapsto \tau_n(f((T-z)^*(T-z))) \\ - \tau_n(f((P_n(T-z)^*(T-z)P_n)^*(P_n(T-z)^*(T-z)P_n)))$$

is bounded by 2, the Stone-Weierstrass Theorem, (4.2.8) and linearity of τ_n it is true that (4.2.9) follows if we can show that

$$\lim_{n\to\infty} |\tau_n(((T-z)^*(T-z))^p) - \tau_n(((P_n(T-z)P_n)^*(P_n(T-z)P_n))^p)| = 0$$

for $p = 1, 2, \ldots$. Also, since the sequence of p-linear forms

$$B_n(T_1, T_2, \ldots, T_{2p}) = \tau_n(T_1T_2\cdots T_{2n}) - \tau_n(P_nT_1P_nT_2P_n\cdots P_nT_{2n}), \quad T_j \in \mathcal{A}$$

is uniformly bounded (by 2) we may assume that T and T^* have finite degree. By Proposition 4.1.5 we have that

$$|\tau_n(((T-z)^*(T-z))^p) - \tau_n(((P_n(T-z)P_n)^*(P_n(T-z)P_n))^p)| \\ \leq \|T-z\|^p\|(T-z)^*\|^p \frac{1}{d_n} p(\deg(T) + \deg(T^*)) \longrightarrow 0, \quad n \to \infty,$$

where $d_n = \dim(\mathcal{H}_n)$, and thus we have shown Step IV.

Step V. We claim that

$$\int_{\mathbb{R}^2} f_n \nabla^2\varphi \, dm \longrightarrow \int_{\mathbb{R}^2} f\nabla^2\varphi \, dm, \quad n \to \infty, \quad \varphi \in C_c^\infty, \tag{4.2.10}$$

when $\operatorname{supp}(\varphi) \subset \mathbb{C} \setminus \sigma(T)$. Let $\delta > 0$ and

$$\Omega_\delta = \{z \in \mathbb{C} : \operatorname{dist}(z, \sigma(T)) \leq \delta\}.$$

We claim that there is a constant $C > -\infty$ such that

$$\inf\{f_n(z) : z \in \mathbb{C} \setminus \Omega_\delta\} \geq C. \tag{4.2.11}$$

Indeed, this is the case. Firstly, observe that for $z \notin \sigma(T)$ it follows that

$$f_n(z) \geq \frac{1}{2}\int_0^1 \log t \, d\mu_{T,z,n}(t),$$

thus (4.2.11) will follow if we can show that there is an $\epsilon > 0$ such that

$$\operatorname{supp}(\mu_{T,z,n}) \subset [\epsilon, \infty) \quad \text{for all} \quad z \in \mathbb{C} \setminus \Omega_\delta.$$

Secondly, note that

$$\inf\{1/\|(T-z)^{-1}\| : z \in \mathbb{C} \setminus \Omega_\delta\} > 0.$$

So let

$$\epsilon = \inf_{z\in\mathbb{C}\setminus\Omega_\delta} \Big(\inf_{\|\xi\|=1, \xi\in\mathcal{H}} \langle (T-z)^*(T-z)\xi, \xi\rangle \Big)^{1/2} = \inf\{1/\|(T-z)^{-1}\| : z \in \mathbb{C} \setminus \Omega_\delta\}.$$

Then, as argued in Step IV, we have that

$$\mu_{T,z,n}([0,\epsilon)) = \tau_n(E_{P_n(T-z)^*(T-z)P_n}([0,\epsilon_n))) = 0,$$

since $\sigma(|P_n(T-z)^*(T-z)P_n|) \subset [\epsilon_n, \infty)$, where

$$\epsilon_n = \inf_{z\in\mathbb{C}\setminus\Omega_\delta}(\inf_{\|\xi\|=1,\xi\in\mathcal{H}}\langle(P_n(T-z)^*(T-z)P_n)\xi,\xi\rangle)^{1/2} \geq \epsilon$$

Pick $\delta > 0$ so small that $\mathrm{supp}(\varphi) \subset \mathbb{C}\setminus\Omega_\delta$. Let

$$g(z) = \begin{cases} \inf\{f_n(z) : z \in \mathbb{C}\setminus\Omega_\delta\} & z \in \mathbb{C}\setminus\Omega_\delta \\ 0 & z \in \Omega_\delta. \end{cases}$$

Then, by the reasoning above, g is integrable and dominates $\{f_n\}$ from below. Hence, (4.2.10) follows by Step IV and dominated convergence.

Note that (4.2.6) follows from Step V and the fact that $\mathrm{supp}(\mu_T) \subset \sigma(T)$, and thus we have proved the theorem. □

Proof. (Proof of Theorem 4.2.2) To prove (i) we need to show that

$$\int_{\mathbb{R}^2} f_n\nabla^2\varphi\, dm \longrightarrow \int_{\mathbb{R}^2} f\nabla^2\varphi\, dm, \quad n\to\infty, \quad \varphi \in C_c^\infty, \tag{4.2.12}$$

where f is defined in (4.2.1). Now, for $z \notin \sigma(T)$ we have

$$\begin{aligned}
f_n(z) &\geq \inf_{n\in\mathbb{N}} \tau_n(\log(P_n(T-z)^*(T-z)\lceil_{P_n\mathcal{H}})) \\
&= \frac{1}{d_n}\sum_j^{d_n} \lambda_j(\log(P_n(T-z)^*(T-z)\lceil_{P_n\mathcal{H}})) \\
&= \frac{1}{d_n}\sum_j^{d_n} \log(\lambda_j(P_n(T-z)^*(T-z)\lceil_{P_n\mathcal{H}})) \\
&\geq \frac{1}{d_n}\sum_j^{d_n} \log(\min_{j\in\{1,\ldots,d_n\}}\{\lambda_j(P_n(T-z)^*(T-z)\lceil_{P_n\mathcal{H}}\})) \\
&= \log((\inf_{\|\xi\|=1,\xi\in\mathcal{H}}\langle P_n(T-z)^*(T-z)P_n\xi,\xi\rangle)^{1/2}) \\
&\geq \log((\inf_{\|\xi\|=1,\xi\in\mathcal{H}}\langle (T-z)^*(T-z)\xi,\xi\rangle)^{1/2}) \\
&= \log(1/\|(T-z)^{-1}\|),
\end{aligned}$$

where $d_n = \dim(\mathcal{H}_n)$ and $\lambda_j(B)$ denotes the j-th eigenvalue of $B \in \mathcal{B}(\mathcal{H}_n)$ according to some ordering, where the eigenvalues of B are repeated according to multiplicity (obviously, the ordering is irrelevant in this context). Hence, f_n is dominated from below by ρ and since ρ is integrable, (4.2.12) follows by dominated convergence.

Now (ii) follows by noting that $z \mapsto \log(|z|^\alpha)$ is locally integrable and arguing as in the proof of (i) using dominated convergence. □

Part II

Applications

Chapter 5

Introduction

Mathematical scientists have been successfully computing eigenvalues and eigenvectors of linear operators since the 1950s. Such computations are a mainstay of the fields of acoustics, computational quantum chemistry through the Schrödinger operator and quantum mechanics. These are self-adjoint examples, but spectral analysis of non-self-adjoint operators is equally central to the stability calculations of fluid dynamics and non-hermitian quantum mechanics. The algorithms involved in applications like these are usually based on discretization of partial differential equations, and sometimes, though not always, they are accompanied by theorems guaranteeing convergence to the correct result as the discretization is refined.

A mathematician, however, may ask a broader question: what about the computation of spectra of arbitrary linear operators, not necessarily defined by derivatives and not necessarily consisting of just eigenvalues? In this generality much less has been done, even in the self-adjoint case, especially if one insists upon theorems guaranteeing convergence.

The purpose of these chapters is to shed light on this fundamental question in operator theory that has received some attention in the last decade (see (Arv91) (Arv93b), (Arv93a), (Arv94a), (Arv94b), (Bro06), (Bro07a) and (DP04), (Dav00), (Dav98), (Böt00), (HRS01), (LS04), (Bou06)(Bou07)), namely, how to compute the spectrum of a linear operator on an infinite dimensional, separable Hilbert space. The question is fundamental in the sense that our understanding of most physical phenomena in quantum mechanics, both relativistic and non-relativistic, depends on the understanding of the spectra of linear operators. However, to obtain complete understanding of such physical phenomena we not only need mathematical descriptions of the behavior of spectra of linear operators, we also need a mathematical theory on how to find explicit approximations to such spectra. If we compare our understanding of classical mechanics and quantum mechanics from computational point of view, there is only one restriction in the classical case, namely, computing power. In the classical case one needs to integrate a vector field on a manifold and there is a vast literature on how to prove rigorously that one can get arbitrarily close to the exact solution given a sufficiently efficient computer. In the quantum case much less is known, in fact it is a completely open question how to compute the spectrum of an arbitrary linear operator as pointed out in (Arv94b): "Unfortunately, there is a dearth of literature on this basic problem, and so far as we have been able to tell, there are no proven techniques." Since this observation was made, there have been new developments in the self-adjoint case (Dav00), but for the general non-self-adjoint case techniques for computing spectra

are not known. The lack of such techniques presents therefore a serious limitation of our possible understanding of quantum systems since non-self-adjoint operators are ubiquitous in quantum mechanics (HN96), (HN97).

In (Dav05) Davies questions whether one can actually compute the spectrum of a bounded operator on a Hilbert space. The example that Davies presents and that gives rise to the question is the following: Let $A_\epsilon : l^2(\mathbb{Z}) \to l^2(\mathbb{Z})$ be defined by

$$(A_\epsilon f)(n) = \begin{cases} \epsilon f(n+1) & n = 0 \\ f(n+1) & n \neq 0. \end{cases}$$

Now for $\epsilon \neq 0$ we have $\sigma(A_\epsilon) = \{z : |z| = 1\}$ but for $\epsilon = 0$ then $\sigma(A_0) = \{z : |z| \leq 1\}$. Davies argues as follows: "If ϵ is a very small constructively defined real number and one is not able to determine whether or not $\epsilon = 0$, then the spectrum of A_ϵ cannot be computed even approximately even though A_ϵ is well-defined constructively. This implies that there exist straightforward bounded operators whose spectrum will probably never be determined."

A numerical analyst may express the same concern. One can argue that if one should do a computation of the spectrum on a computer, the fact that the arithmetic operations carried out are not exact may lead to the outcome that one gets the true solution to a slightly perturbed problem. This type of analysis is often referred to as Backward Error Analysis in the numerical linear algebra literature. As suggested in the previous example, getting the answer to a slightly perturbed problem could be disastrous.

This poses a slightly philosophical question; is it impossible to compute spectra of arbitrary operators? And if so, does that mean that there are operators, whose spectral theory might be crucial for understanding physical phenomena, yet their spectra will never be determined? This would imply that there is a rather unpleasant barrier between what we can compute and what we want to compute. In Chapter 2 and Chapter 3 several new methods for estimating spectra and pseudospectra of operators were presented. Our goal in this part is to show that these results can be used for actual computations, and that, indeed, it is possible to compute spectra of arbitrary bounded operators on separable Hilbert spaces. We will emphasize the computational task and refer to Chapters 2 and 3 for justifications of the mathematical statements that will be presented.

Our theory is very much inspired by the pseudospectral theory that has emerged through the last two decades (TE05). The main reason is that to overcome the discontinuity problem suggested above, one is forced to consider the computation of a different set than the spectrum, even though estimating the spectrum may be the main goal. This is the main theme of Chapter 6 where we will see that variants of the pseudospectra, namely the n-pseudospectra, are excellent candidates for sets that approximate the spectrum well. Also, these sets do not behave discontinuously with the operator (we will be more specific about this later). We will in Chapter 6 also consider implementation details of algorithms that compute the n-pseudospectra and show how these can be used to compute spectra of arbitrary bounded linear operators.

In Chapter 7 we deviate from the pseudospectral theory and focus on the Infinite Dimensional QR algorithm (or the Infinite QR algorithm for short). The reason why the Infinite QR algorithm is a valuable supplement to the pseudospectral methods introduced in Chapter 6 is that spectral approximation methods based on pseudospectral

theory may struggle with very non-normal problems. The Infinite QR algorithm exhibits surprisingly good qualities when handling non-normal problems, and, although the Infinite QR algorithm can never reveal the whole spectrum of an operator, it turns out to be an indispensable tool.

Chapter 8 is a direct continuation of Chapter 7 and is devoted to the task of implementing the Hessenberg reduction of an infinite matrix. The Hessenberg reduction in infinite dimensions is motivated (as in finite dimensions) by the desire to speed up the QR algorithm. As expected, the Hessenberg reduction together with the Infinite QR algorithm cut the computational cost dramatically and therefore allow for more complicated problems. Several numerical examples follow at the end of that chapter.

5.1 Background and Notation

In this section we will briefly recall some basics from functional analysis and some notation. Throughout the thesis, $\mathcal{H}$ will always denote a separable Hilbert space and $\mathcal{B}(\mathcal{H})$ the set of bounded linear operators on $\mathcal{H}$. If $T \in \mathcal{B}(\mathcal{H})$ and $T - z$ is invertible, for $z \in \mathbb{C}$, we use the notation $R(z, T) = (T - z)^{-1}$. We will denote orthonormal basis elements of $\mathcal{H}$ by e_j, and if $\{e_j\}_{j\in\mathbb{N}}$ is a basis and $\xi \in \mathcal{H}$ then $\xi_j = \langle \xi, e_j \rangle$. The word basis will always refer to an orthonormal basis. If $T \in \mathcal{B}(\mathcal{H})$ then T is uniquely determined by its matrix elements $\langle Te_j, e_i \rangle$ and hence we will use the words bounded operator and infinite matrix interchangeably.

A couple of basic topological aspects of $\mathcal{B}(\mathcal{H})$ will be useful in the future developments ((KR97) gives a good overview of the ideas sketched here). Recall that a sequence $\{T_n\} \subset \mathcal{B}(\mathcal{H})$ converges to $T \in \mathcal{B}(\mathcal{H})$ in the strong operator topology, denoted by

$$\operatorname*{SOT\text{-}lim}_{n\to\infty} T_n = T,$$

if and only if $T_n\xi \to T\xi$ as $n \to \infty$ for all $\xi \in \mathcal{H}$. Also, $\{T_n\} \subset \mathcal{B}(\mathcal{H})$ converges to $T \in \mathcal{B}(\mathcal{H})$ in the weak operator topology, denoted by

$$\operatorname*{WOT\text{-}lim}_{n\to\infty} T_n = T,$$

if and only if $\langle T_n\xi, \eta \rangle \to \langle T\xi, \eta \rangle$ as $n \to \infty$ for all $\xi, \eta \in \mathcal{H}$. In connection with the weak operator topology, the following proposition will be useful in the future developments.

Proposition 5.1.1. *Let $\mathcal{H}$ be a Hilbert space. Then*

$$\{T \in \mathcal{B}(\mathcal{H}) : \|T\| \leq 1\}$$

is sequentially compact in the weak operator topology.

This proposition means that if $\{T_n\}$ is a bounded (in the operator norm) sequence in $\mathcal{B}(\mathcal{H})$ then there is an operator $T \in \mathcal{B}(\mathcal{H})$ and a subsequence $\{T_{n_k}\}$ such that

$$\operatorname*{WOT\text{-}lim}_{k\to\infty} T_{n_k} = T$$

Another part of basic operator theory is the functional calculus, namely, for a normal operator $T \in \mathcal{B}(\mathcal{H})$ and $f \in L^\infty(\mathbb{C})$ we can form the operator $f(T) \in \mathcal{B}(\mathcal{H})$. The functional calculus has several key features e.g. for $f, g \in L^\infty(\mathbb{C}$ we have

$$(fg)(T) = f(T)g(T), \qquad f(T)^* = \bar{f}(T).$$

This means that, if we let $\omega \subset \mathbb{C}$ and χ_ω denote the characteristic function on ω, then $\chi_\omega(T)$ must be a projection when T is normal.

The spectrum of $T \in \mathcal{B}(\mathcal{H})$ will be denoted by $\sigma(T)$, and $\sigma_d(T)$ denotes the set of isolated eigenvalues with finite multiplicity. In connection with the spectrum we need to recall some definitions.

Definition 5.1.2. *Let $T \in \mathcal{B}(\mathcal{H})$, then the essential spectrum is defined as*

$$\sigma_{\text{ess}}(T) = \bigcap_{K \text{ compact}} \sigma(T + K).$$

Definition 5.1.3. *Let T be a bounded operator on a Hilbert space $\mathcal{H}$. Then the numerical range of T is defined as*

$$W(T) = \{\langle T\xi, \xi\rangle : \|\xi\| = 1\},$$

and the essential numerical range is defined as

$$W_e(T) = \bigcap_{K \text{ compact}} \overline{W(T + K)}$$

Definition 5.1.4. *Let $T \in \mathcal{B}(\mathcal{H})$ then the essential spectral radius is defined as*

$$r_{\text{ess}}(T) = \sup\{|\lambda| : \lambda \in \sigma_{\text{ess}}(T)\}.$$

Definition 5.1.5. *Let T be a closed operator on a Hilbert space $\mathcal{H}$ such that $\sigma(T) \neq \mathbb{C}$, and let $\epsilon > 0$. The ϵ-pseudospectrum of T is defined as the set*

$$\sigma_\epsilon(T) = \sigma(T) \cup \{z \notin \sigma(T) : \|(z - T)^{-1}\| > \epsilon^{-1}\}.$$

Convergence of sets in the complex plane will be quite crucial in our analysis and hence we need the Hausdorff metric as defined by the following.

Definition 5.1.6. *(i) For a set $\Sigma \subset \mathbb{C}$ and $\delta > 0$ we will let $\omega_\delta(\Sigma)$ denote the δ-neighborhood of Σ (i.e. the union of all δ-balls centered at points of Σ).*

(ii) Given two sets $\Sigma, \Lambda \subset \mathbb{C}$ we say that Σ is δ-contained in Λ if $\Sigma \subset \omega_\delta(\Lambda)$.

(iii) Given two compact sets $\Sigma, \Lambda \subset \mathbb{C}$ their Hausdorff distance is

$$d_H(\Sigma, \Lambda) = \max\{\sup_{\lambda \in \Sigma} d(\lambda, \Lambda), \sup_{\lambda \in \Lambda} d(\lambda, \Sigma)\}$$

where $d(\lambda, \Lambda) = \inf_{\rho \in \Lambda} |\rho - \lambda|$.

If $\{\Lambda_n\}_{n \in \mathbb{N}}$ is a sequence of compact subsets of $\mathbb{C}$ and $\Lambda \subset \mathbb{C}$ is compact such that $d_H(\Lambda_n, \Lambda) \to 0$ as $n \to \infty$ we may use the notation $\Lambda_n \longrightarrow \Lambda$. The closure of a set $\Omega \subset \mathbb{C}$ will be denoted by $\overline{\Omega}$, however, when convenient, the notation $\text{cl}(\Omega)$ may be used.

The fact that arithmetic operations may not be carried out exactly on a computer is crucial in our analysis, and ϵ_{mach} will always denote the machine epsilon in the computer software used. The software of choice is MATLAB, and in that case $\epsilon_{\text{mach}} = 10^{-16}$.

Chapter 6

Pseudospectral Theory

Let $T \in \mathcal{B}(\mathcal{H})$, $\{e_j\}$ be a basis for $\mathcal{H}$, and suppose that we wish to compute the spectrum of T. As discussed at the beginning of Chapter 5, we are faced with the slightly unpleasant problem of computing something that may depend discontinuously on the matrix elements $\langle Te_j, e_i\rangle$. The fact that roundoff errors always will play a part in the computation may result in the fact that one gets the solution to a perturbed spectral problem, and this may be far from the solution of the original problem if the desired spectrum we wish to compute varies discontinuously with the operator. Thus, this seems like almost an impossible task to handle numerically.

The solution to the problem is to look to the pseudospectral theory. Note that if we were considering estimating the pseudospectrum instead of the spectrum, the problem suggested by the example in Chapter 5 would not occur. The reason is that the pseudospectrum varies continuously with the operator T if T is bounded (we will be more specific regarding the continuity below.) One may argue that the pseudospectrum may give a lot of information about the operator and one should therefore estimate it in place of the spectrum, however, we are interested in getting a complete spectral understanding of the operator and will therefore estimate both the spectrum and the pseudospectrum. We hence wish to introduce a set which has the continuity property of the pseudospectrum but approximates the spectrum, and this motivates our definition of the n-pseudospectrum. As we will see in Section 6.2, the n-pseudospectrum has all the nice continuity properties that the pseudospectrum has, but it also approximates the spectrum arbitrarily well for large n.

Before we continue with pseudospectral theory we would like to make a short detour via the finite section method and try to convince the reader that the finite section method is not a serious contender to the "method of the month" award among algorithms for the general computational spectral problem.

6.1 The Finite Section Method

Suppose that we have an operator $A \in \mathcal{B}(\mathcal{H})$ and that we know the matrix elements $a_{ij} = \langle Ae_j, e_i\rangle$ with respect to some basis $\{e_j\}$. The question is then how do we compute the spectrum and the pseudospectra of A using $\{a_{ij}\}$. A natural thought may be to reduce this to a finite-dimensional spectral problem by constructing (using $\{e_j\}$) a sequence of finite rank projections $\{P_m\}$ such that $P_{m+1} \geq P_m$ and $P_m \to I$ strongly, where I is

the identity, and then compute the spectrum and pseudospectra of $P_m A\lceil_{P_m\mathcal{H}}$. Typically P_m would be the projection onto $\text{span}\{e_1, \ldots, e_m\}$. This is often referred to as the finite section method in the literature. Now, this may work in some cases e.g. if the operator is compact or in the case of computing pseudospectra, if one is considering a Toeplitz operator. However, one must be very careful using the finite section method and it should not be used unless accompanied by a rigorous analysis that justifies the convergence

$$\sigma(P_m A\lceil_{P_m\mathcal{H}}) \longrightarrow \sigma(A), \qquad \sigma_\epsilon(P_m A\lceil_{P_m\mathcal{H}}) \longrightarrow \sigma_\epsilon(A), \qquad \epsilon > 0, \quad m \to \infty.$$

It is quite easy to find elementary counter examples to show that the finite section method can fail dramatically. Consider the shift operator defined by $Se_n = e_{n+1}$ on $l^2(\mathbb{N})$. This operator has the following matrix representation

$$S = \begin{pmatrix} 0 & 1 & 0 & 0 & \ldots \\ 0 & 0 & 1 & 0 & \ldots \\ 0 & 0 & 0 & 1 & \ldots \\ 0 & 0 & 0 & 0 & \ldots \\ \vdots & \vdots & \vdots & \vdots & \ddots \end{pmatrix}.$$

Thus, if P_m is the projection onto $\text{span}\{e_1, \ldots, e_m\}$, we would get that $\sigma((P_m S\lceil_{P_m\mathcal{H}}) = \{0\}$ for all m, but $\sigma(S)$ is the closed unit disc. To find examples where the finite section method fails when wanting to compute the pseudospectrum, one does not have to go very far away from the Toeplitz operators. The finite section method may have serious trouble finding the right pseudospectra of Laurent operators. Note that if we have a Laurent operator A_L given in its matrix representation with respect to the basis $\{e_j\}_{j=-\infty}^{\infty}$ and choose P_m to be the projection onto

$$\text{span}\{e_{-m}, \ldots, e_m\}$$

then $P_m A\lceil_{P_m\mathcal{H}}$ is a Toeplitz matrix. So, if A_T is the Toeplitz variant of A_L, meaning that it has the same matrix elements but is an operator on $l^2(\mathbb{N})$ instead of $l^2(\mathbb{Z})$, then

$$\sigma_\epsilon(P_m A_L\lceil_{P_m\mathcal{H}}) \longrightarrow \sigma_\epsilon(A_T), \qquad m \to \infty,$$

but we may have that

$$\sigma_\epsilon(A_L) \neq \sigma_\epsilon(A_T),$$

and in this case the finite section method will fail. This is visualized in the following example. Define the Laurent operator by

$$A_L = \begin{pmatrix} \ddots & \vdots & \vdots & \vdots & \vdots & \\ \ldots & 0 & 1 & 0 & 0 & \ldots \\ \ldots & 0 & 0 & 1 & 0 & \ldots \\ \ldots & 1-i & 0 & 0 & 1 & \ldots \\ \ldots & 0 & 1-i & 0 & 0 & \ldots \\ & \vdots & \vdots & \vdots & \vdots & \ddots \end{pmatrix},$$

then $\sigma_\epsilon(P_m A_L\lceil_{P_m\mathcal{H}})$ is far from $\sigma_\epsilon(A_L)$ as visualized in Figure 6.1 for $\epsilon = 0.1$.

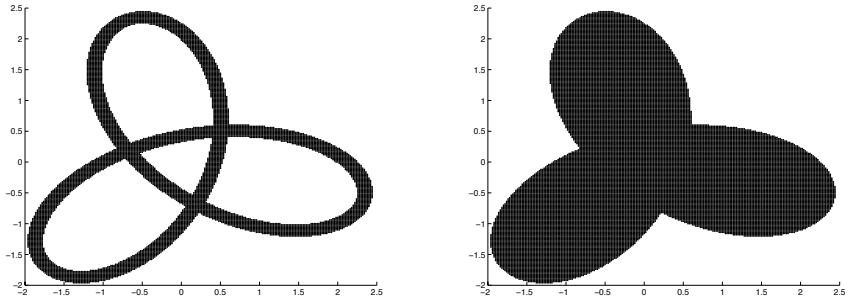


Figure 6.1: The first figure shows $\sigma_\epsilon(A_L)$ and the second figure shows $\sigma_\epsilon(P_m A_L\lceil_{P_m\mathcal{H}})$ for $\epsilon = 0.1$ and $m = 1000$.

6.2 The n-pseudospectrum

Given a closed operator T on $\mathcal{H}$, the motivation for the n-pseudospectrum is the desire to approximate the function

$$z \mapsto \operatorname{dist}(z, \sigma(T)),$$

in order to estimate $\sigma(T)$. A convenient formula for this is

$$\operatorname{dist}(z, \sigma(T)) = \frac{1}{\rho(R(z,T))},$$

where ρ denotes the spectral radius. Thus, in principle, we have reduced the problem of estimating the distance from z to $\sigma(T)$ to a problem of estimating the spectral radius of a bounded operator. Now, numerically that is a nontrivial task, but keeping in mind the spectral radius formula, namely,

$$\rho(A) = \lim_{n\to\infty} \|A^n\|^{1/n}, \qquad A \in \mathcal{B}(\mathcal{H}),$$

we can approximate the spectral radius by estimating the norm of powers of the operator. By choosing a subsequence of $\{\|A^n\|^{1/n}\}$, namely, $\{\|A^{2^n}\|^{1/2^n}\}$ we get a decreasing sequence

$$\|A^{2^n}\|^{1/2^n} \geq \|A^{2^{n+1}}\|^{1/2^{n+1}} \quad \text{and} \quad \rho(A) = \lim_{n\to\infty} \|A^{2^n}\|^{1/2^n}.$$

Hence, we have

$$1/\|R(z,T)^{2^n}\|^{1/2^n} \leq 1/\|R(z,T)^{2^{n+1}}\|^{1/2^{n+1}} \leq 1/\rho(R(z,T)) = \operatorname{dist}(z, \sigma(T))$$

and

$$\lim_{n\to\infty} \frac{1}{\|(R(z,T))^{2^n}\|^{1/2^n}} = \operatorname{dist}(z, \sigma(T)).$$

This gives the motivation for the following definition of the (n, ϵ)-pseudospectrum, or the n-pseudospectrum for short.

Definition 6.2.1. *Let T be a closed operator on a Hilbert space $\mathcal{H}$, and let $n \in \mathbb{Z}_+$ and $\epsilon > 0$. The (n, ϵ)-pseudospectrum of T is defined as the set*

$$\sigma_{n,\epsilon}(T) = \sigma(T) \cup \{z \notin \sigma(T) : \|R(z,T)^{2^n}\|^{1/2^n} > \epsilon^{-1}\}.$$

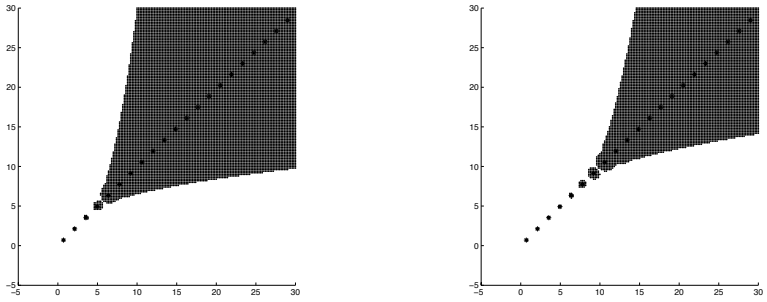


Figure 6.2: The figure shows $\sigma_\epsilon(H)$ (left) and $\sigma_{1,\epsilon}(H)$ (right) for $\epsilon = 0.1$.

As we see, the n-pseudospectrum is just a generalization of the pseudospectrum. By the analysis above, one can deduce that the n-pseudospectrum should be a better approximation to the spectrum than the pseudospectrum, and hopefully also share its nice continuity properties. In particular one should expect

$$\sigma_{n,\epsilon}(T) \supset \sigma_{n+1,\epsilon}(T),$$

and hope for

$$d_H(\overline{\sigma_{n,\epsilon}(T)}, \overline{\omega_\epsilon(\sigma(T))}) \longrightarrow 0, \quad n \to \infty,$$

where $\omega_\epsilon(\sigma(T))$ denotes the ϵ-neighborhood around $\sigma(T)$. A famous example in pseudospectral theory is the complex harmonic oscillator (DK04)

$$Hf(x) = -f''(x) + cx^2 f(x)$$

acting on $L^2(\mathbb{R})$. To visualize the difference between the pseudospectrum and the n-pseudospectrum we have computed the pseudospectrum and the 1-pseudospectrum for H when $c = i$ in Figure 6.2.

6.3 Properties of the n-pseudospectra of Bounded Operators

Theorem 6.3.1. *Let $T \in \mathcal{B}(\mathcal{H})$ and define for $z \in \mathbb{C}$ and $n \in \mathbb{Z}_+$.*

$$\begin{aligned}\gamma_n(z) = \min\Big[&\inf\{\lambda^{1/2^{n+1}} : \lambda \in \sigma\Big(((T-z)^*)^{2^n}(T-z)^{2^n}\Big)\}, \\ &\inf\{\lambda^{1/2^{n+1}} : \lambda \in \sigma\Big((T-z)^{2^n}((T-z)^*)^{2^n}\Big)\}\Big].\end{aligned} \tag{6.3.1}$$

Then the following is true.

(i) $\sigma_{n+1,\epsilon}(T) \subset \sigma_{n,\epsilon}(T)$.

(ii) $\sigma_{n,\epsilon}(T) = \{z \in \mathbb{C} : \gamma_n(z) < \epsilon\}$.

(iii) $\overline{\{z : \gamma_n(z) < \epsilon\}} = \{z : \gamma_n(z) \leq \epsilon\}$.

(iv) We have that

$$d_H(\overline{\sigma_{n,\epsilon}(T)}, \overline{\omega_\epsilon(\sigma(T))}) \longrightarrow 0, \qquad n \to \infty,$$

where $\omega_\epsilon(\sigma(T))$ denotes the ϵ-neighborhood around $\sigma(T)$.

(v) If $\{T_k\} \subset \mathcal{B}(\mathcal{H})$ and $T_k \to T$ in norm, it follows that

$$d_H(\sigma_{n,\epsilon}(T_k), \sigma_{n,\epsilon}(T)) \to 0, \quad k \to \infty.$$

Proof. This is Theorem 3.3.4 in Chapter 3 and a proof can be found there. □

Theorem 6.3.1 provides several important observations. Firstly, the fact that

$$d_H(\overline{\sigma_{n,\epsilon}(T)}, \overline{\omega_\epsilon(\sigma(T))}) \longrightarrow 0, \qquad n \to \infty$$

allows us to use the n-pseudospectrum as an approximation to the spectrum. Secondly, the problem of inexact arithmetic is solved by the fact that for each fixed n we have

$$d_H(\overline{\sigma_{n,\epsilon}(T_k)}, \overline{\sigma_{n,\epsilon}(T)}) \to 0, \quad k \to \infty,$$

when $T_k \to T$ in norm. Thus, in theory, we can get arbitrarily close to the spectrum by computing the n-pseudospectrum and still allow the computation to be in inexact arithmetic. Now, of course the ϵ_{mach} will have to decrease as n grows.

The function γ_n and the fact that $\overline{\sigma_{n,\epsilon}(T)} = \{z \in \mathbb{C} : \gamma_n(z) \leq \epsilon\}$ provide us with a tool for estimating the n-pseudospectrum. In fact, by recalling (6.3.1), we have now reduced the problem of finding the spectrum of a non-normal operator to a problem of finding the smallest element in the spectrum of a self-adjoint operator. In the following examples we will show some of the properties of the pseudospectra listed in Theorem 6.3.1

Example 6.3.2. To demonstrate the property $\sigma_{n+1,\epsilon}(T) \subset \sigma_{n,\epsilon}(T)$ of the pseudospectra we have chosen the following operator:

$$T = \begin{pmatrix} a_1 & b_1 & 0 & 0 & \dots \\ c_1 & a_2 & b_2 & 0 & \dots \\ 0 & c_2 & a_1 & b_3 & \dots \\ 0 & 0 & c_3 & a_2 & \dots \\ \vdots & \vdots & \vdots & \vdots & \ddots \end{pmatrix},$$

where $a_1 = 2$, $a_2 = 0.5$, $b_j = \frac{1+i2}{j^{1/6}}$ and $c_j = 1/j^{1/2}$. Now, T can be written as a sum of two operators where one is compact and the other one has only essential spectrum and thus A should have plenty of isolated eigenvalues. The four largest eigenvalues with corresponding n-pseudospectra are displayed in Figure 6.3.

Example 6.3.3. To visualize the property that if $\{T_k\} \subset \mathcal{B}(\mathcal{H})$ and $T_k \to T$ in norm, it follows that

$$d_H(\overline{\sigma_{n,\epsilon}(T_k)}, \overline{\sigma_{n,\epsilon}(T)}) \to 0, \quad k \to \infty,$$

a natural test object is the example by Davies introduced in Section 5. The discontinuity of the spectrum shown in that example was a strong motivation for the introduction of the n-pseudospectrum. Recall that we define $A_\delta : l^2(\mathbb{Z}) \to l^2(\mathbb{Z})$ by

$$(A_\delta f)(n) = \begin{cases} \delta f(n+1) & n = 0 \\ f(n+1) & n \neq 0, \end{cases}$$

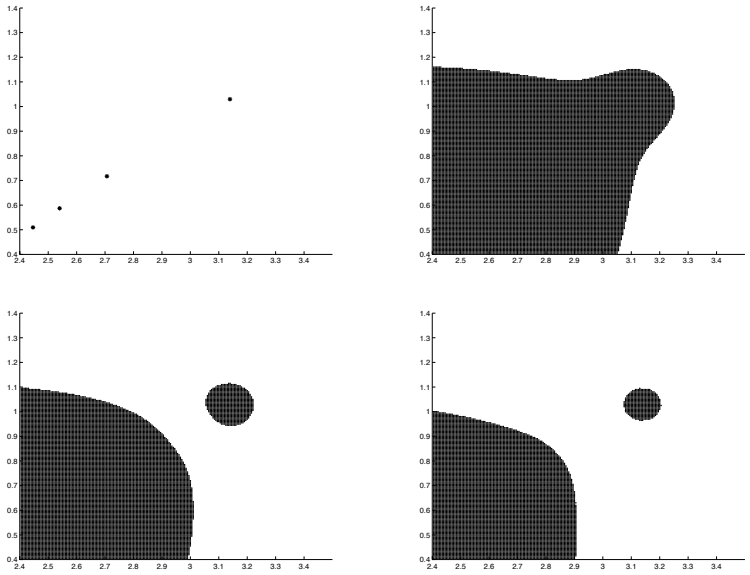


Figure 6.3: The first figure shows the first four eigenvalues, and the following figure shows $\sigma_\epsilon(T)$, $\sigma_{1,\epsilon}(T)$, $\sigma_{2,\epsilon}(T)$ for $\epsilon = 0.05$.

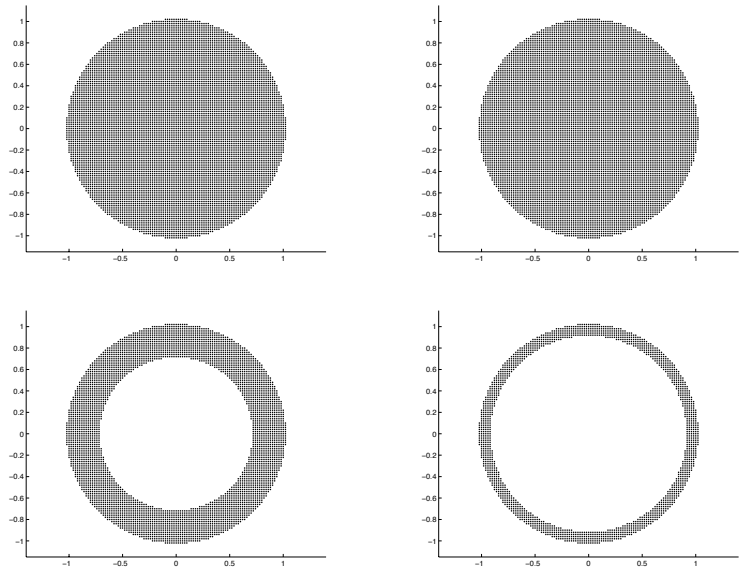


Figure 6.4: The figure shows $\sigma_{2,\epsilon}(A_0)$, $\sigma_{2,\epsilon}(A_{10^{-16}})$, $\sigma_{1,\epsilon}(A_{0.005})$, and $\sigma_{2,\epsilon}(A_{0.005})$, for $\epsilon = 0.025$.

and that for $\delta \neq 0$ we have $\sigma(A_\delta) = \{z : |z| = 1\}$ but for $\delta = 0$ then $\sigma(A_0) = \{z : |z| \leq 1\}$.

We have computed the n-pseudospectrum of A_0 for $\epsilon = 0.025$ and $n = 2$ which coincides with the closed ϵ-neighborhood of the unit disk. We have also computed the n-pseudospectrum of $A_{10^{-16}}$ to demonstrate that, at least up to the accuracy of the grid size we have chosen, the computed results are identical.

Now, if we actually wanted to compute the spectrum of $A_{10^{-16}}$ we would have to choose computer software with higher precision and also take n much larger. The ϵ_{mach} in MATLAB limits us to take $n \leq 2$ since our computation requires operations with $(\epsilon_{\text{mach}})^{1/2^{n+1}}$. Hence, since $(\epsilon_{\text{mach}})^{1/2^{3+1}} = 0.1$, we may experience that for $n = 3$, our computation will be accurate only up to the first decimal. However, we have visualized (in Figure 6.4) (iii) in Theorem 6.3.1 by computing the n-pseudospectra of $A_{0.005}$ for $n = 1, 2$, in which case the n-pseudospectra approximates the spectrum of $A_{0.005}$ quite well even for small values of n.

6.4 Computing the n-pseudospectrum

6.4.1 Designing the Algorithm

Numerically, a self-adjoint spectral problem is much easier to deal with than a non-self-adjoint problem, but we cannot attack the task of computing (6.3.1) as it is, since this is an infinite-dimensional problem. We therefore need to find an approximation to γ_n (as defined in (6.3.1)) that is suitable for computations. A natural choice seems to be to choose a sequence of finite rank projections $\{P_m\}$ such that $P_{m+1} \geq P_m$ and $P_m \to I$ strongly, e.g. we may choose a basis $\{e_j\}$ and let P_m be the projection onto $\text{span}\{e_1, \ldots, e_m\}$. Now we can try to approximate γ_n by the function

$$\begin{aligned}\gamma_{n,m}(z) = \min\Big[&\min\{\lambda^{1/2^{n+1}} : \lambda \in \sigma\left(P_m((T-z)^*)^{2^n}(T-z)^{2^n}\Big|_{P_m\mathcal{H}}\right)\}, \\ &\min\{\lambda^{1/2^{n+1}} : \lambda \in \sigma\left(P_m(T-z)^{2^n}((T-z)^*)^{2^n}\Big|_{P_m\mathcal{H}}\right)\}\Big],\end{aligned} \tag{6.4.1}$$

and if $\gamma_{n,m} \to \gamma_n$ in some sense we can hope that

$$\{z : \gamma_{n,m}(z) \leq \epsilon\} \longrightarrow \overline{\sigma_{n,\epsilon}(T)}, \quad m \to \infty.$$

In fact so is almost the case as the following theorem guarantees.

Theorem 6.4.1. *Let $T \in \mathcal{B}(\mathcal{H})$ and let $\{P_m\}$ is an increasing sequence of finite rank projections converging strongly to the identity such that $P_{m+1} \geq P_m$. Define $\gamma_{n,m}$ as in (6.4.1), then $\gamma_{n,m} \to \gamma_n$ locally uniformly as $m \to \infty$, and for a compact ball $K \subset \mathbb{C}$ such that $\sigma_{n,\epsilon}(T) \cap K^o \neq \emptyset$ we have*

$$\{z : \gamma_{n,m}(z) \leq \epsilon\} \cap K \longrightarrow \overline{\sigma_{n,\epsilon}(T)} \cap K, \quad m \to \infty,$$

where the convergence is understood to be in the Hausdorff metric.

Proof. The proof can be found in the proof of Theorem 3.5.4 in Chapter 3. □

Now, computing $\gamma_{n,m}$ still involves the products

$$((T-z)^*)^{2^n}(T-z)^{2^n}, \qquad (T-z)^{2^n}((T-z)^*)^{2^n},$$

which may be challenging to compute as T acts on an infinite dimensional space. However, there is a solution to this problem. Instead of computing the products $((T-z)^*)^{2^n}(T-z)^{2^n}$ and $(T-z)^{2^n}((T-z)^*)^{2^n}$ we will compute

$$(P_k(T-z)P_k)^*)^{2^n}(P_k(T-z)P_k)^{2^n}, \qquad (P_k(T-z)P_k)^{2^n}((P_k(T-z)P_k)^*)^{2^n},$$

where P_k is a finite rank projection as in Theorem 6.4.1. As P_k has finite rank this is feasible, in particular, we can define the function

$$\begin{aligned}&\gamma_{n,m,k}(z)\\ &= \min\Big(\min\{\lambda^{1/2^{n+1}} : \lambda \in \sigma\left(P_m((P_k(T-z)P_k)^*)^{2^n}(P_k(T-z)P_k)^{2^n}\Big|_{P_m\mathcal{H}}\right)\},\\ &\quad \min\{\lambda^{1/2^{n+1}} : \lambda \in \sigma\left(P_m(P_k(T-z)P_k)^{2^n}((P_k(T-z)P_k)^*)^{2^n}\Big|_{P_m\mathcal{H}}\right)\}\Big),\end{aligned} \tag{6.4.2}$$

and argue that $\gamma_{n,m,k} \to \gamma_{n,m}$ as $k \to \infty$. In fact we have:

Theorem 6.4.2. *Let $T \in \mathcal{B}(\mathcal{H})$ and let $\{P_m\}$ be an increasing sequence of finite rank projections converging strongly to the identity such that $P_{m+1} \geq P_m$. Define $\gamma_{n,m,k}$ as in (6.4.2), then $\gamma_{n,m,k} \to \gamma_{n,m}$ locally uniformly as $k \to \infty$, and for a compact ball $K \subset \mathbb{C}$ such that $\{z : \gamma_{n,m}(z) \leq \epsilon\} \cap K^o \neq \emptyset$ we have*

$$\{z : \gamma_{n,m,k}(z) \leq \epsilon\} \cap K \longrightarrow \{z : \gamma_{n,m}(z) \leq \epsilon\} \cap K, \quad k \to \infty,$$

where the convergence is understood to be in the Hausdorff metric.

Proof. The proof can be found in the proof of Theorem 3.5.4 in Chapter 3. □

For a full infinite matrix, $\gamma_{n,m,k}$ can be a tough challenge to compute, since there are two limit processes going on at the same time, namely $k \to \infty$ and $m \to \infty$. It is therefore important to take advantage of structured problems.

Definition 6.4.3. *Let $\{e_j\}_{j\in\mathbb{N}}$ be a basis for $\mathcal{H}$ and let $T \in \mathcal{B}(\mathcal{H})$. If*

$$\langle Te_{j+l}, e_j\rangle = \langle Te_j, e_{j+l}\rangle = 0, \qquad l > d,$$

then T is said to be banded with bandwidth d.

The following theorem is important for the computation of $\gamma_{n,m,k}$ when the infinite matrix is banded.

Theorem 6.4.4. *Let $T \in \mathcal{B}(\mathcal{H})$ and $\{e_j\}$ be a basis for $\mathcal{H}$. Let P_m be the projection onto*

$$\operatorname{span}\{e_1, \ldots, e_m\}.$$

Define $\gamma_{n,m}$ and $\gamma_{n,m,k}$ as in (6.4.1) and (6.4.2). Suppose that the matrix representation of T with respect to $\{e_j\}$ is banded with bandwidth d. Then, for $m > d$,

$$\gamma_{n,m}(z) = \gamma_{n,m,2^n d+m}(z), \qquad z \in \mathbb{C}.$$

Proof. The proof can be found in the proof of Theorem 3.5.8 in Chapter 3. □

6.4.2 The algorithm

As Theorem 6.4.1 suggest, the estimation of the n-pseudospectra can be done by computing values of $\gamma_{n,m,k}$ on a grid in the complex plane. Also, by Theorem 6.4.4, if the matrix T is banded with bandwidth d then

$$\gamma_{n,m}(z) = \gamma_{n,m,2^n d+m}(z), \qquad z \in \mathbb{C},$$

where d is the number of off diagonals. Thus, we are left with the task of computing

$$\min\left\{\lambda^{1/2^{n+1}} : \lambda \in \sigma\left(P_m((P_k(T-z)P_k)^*)^{2^n}(P_k(T-z)P_k)^{2^n}\Big\lceil_{P_m\mathcal{H}}\right)\right\} \tag{6.4.3}$$

and

$$\min\left\{\lambda^{1/2^{n+1}} : \lambda \in \sigma\left(P_m(P_k(T-z)P_k)^{2^n}((P_k(T-z)P_k)^*)^{2^n}\Big\lceil_{P_m\mathcal{H}}\right)\right\}. \tag{6.4.4}$$

As m becomes large, both (6.4.3) and (6.4.4) are difficult to compute since

$$\begin{aligned}&(P_m((P_k(T-z)P_k)^*)^{2^n}(P_k(T-z)P_k)^{2^n}\Big\lceil_{P_m\mathcal{H}}\\ &P_m(P_k(T-z)P_k)^{2^n}((P_k(T-z)P_k)^*)^{2^n}\Big\lceil_{P_m\mathcal{H}}\end{aligned} \tag{6.4.5}$$

may have many eigenvalues very close to zero, and standard numerical routines as MATLAB's `eigs` will have trouble detecting the smallest eigenvalue to sufficient precision (at least that is our experience). If one wants a contour plot of the n-pseudospectrum, there is no way around the previous problem, but if one only wants the n-pseudospectrum for one specific $\epsilon > 0$, it is unnecessary to compute the smallest eigenvalue in (6.4.3) and (6.4.4). In fact, since we are only interested in knowing whether $\gamma_{n,m,k}(z) \leq \epsilon$ for some complex z, we only need to check if the self-adjoint matrices

$$(P_m((P_k(T-z)P_k)^*)^{2^n}(P_k(T-z)P_k)^{2^n}\Big\lceil_{P_m\mathcal{H}} - \epsilon^{2^{n+1}}I$$

and

$$P_m(P_k(T-z)P_k)^{2^n}((P_k(T-z)P_k)^*)^{2^n}\Big\lceil_{P_m\mathcal{H}} - \epsilon^{2^{n+1}}I$$

are both positive definite. If they both are, then $z \notin \{z : \gamma_{n,m,k}(z) \leq \epsilon\}$.

6.4.3 The Cholesky Decomposition

It is well known that a self-adjoint matrix $A \in \mathbb{C}^{n\times n}$ is positive definite if and only if it has a Cholesky decomposition

$$A = GG^*,$$

where G is lower triangular with positive elements on the diagonal (GVL96). Thus, to determine whether A is positive definite or not, we need to find out if the decomposition $A = GG^*$ exists. This can be done in the following way. Let

$$A = \begin{pmatrix}\alpha & v^*\\ v & B\end{pmatrix} = \begin{pmatrix}\beta & 0\\ v/\beta & I_{n-1}\end{pmatrix}\begin{pmatrix}1 & 0\\ 0 & B - vv^*/\alpha\end{pmatrix}\begin{pmatrix}\beta & v^*/\beta\\ 0 & I_{n-1}\end{pmatrix}, \tag{6.4.6}$$

where $\alpha > 0$ if A is positive definite, so $\beta = \sqrt{\alpha}$. If $\alpha \leq 0$ we conclude that A is not positive definite and we are done. Now, $B - vv^*/\alpha$ is positive definite if and only if A is positive definite since it is a principal submatrix of U^*AU, where

$$U = \begin{pmatrix} 1 & -v^*/\alpha \\ 0 & I_{n-1} \end{pmatrix}.$$

If there is a Cholesky factorization $G_1 G_1^* = B - vv^*/\alpha$ then it follows from (6.4.6) that $A = GG^*$, where

$$G = \begin{pmatrix} \beta & 0 \\ v/\beta & G_1 \end{pmatrix}.$$

We can continue this argument with G_1 and do this recursively to obtain $\{G_1, G_2, \ldots G_{n-1}\}$. Thus, if all G_js turn out to be positive definite then A is positive definite, and if there is a G_j that is not positive definite then A cannot be positive definite. The standard algorithm for this requires $n^3/3$ flops. A neat tool for determining whether or not A is positive definite is MATLAB's `chol` routine that has an build in check for positive definiteness of matrices.

Suppose that T is a banded infinite matrix with bandwidth d, the following MATLAB program will plot the the following set

$$\{z : \gamma_{n,m,2^n d+m}(z) \leq \epsilon\} \cap K,$$

where K is a rectangle in $\mathbb{C}$ and $\gamma_{n,m,2^n d+m}$ is defined in (6.4.2).

Algorithm 6.4.1.

```
%Computes {z : gamma_n,m,2^nd+m(z) <= epsilon} \cap K,
%for an infinite matrix A with bandwith d=diag, where K is
%a rectangle with coordinates left, right, up, down.
%The size of the section of A must be 2^nd+m.

function s = n_pseu_chol(A,epsilon,left,right,down,up,grid_eps,n,diag)
r = (right-left)/grid_eps; e = grid_eps; si = size(A,2);
l = (up - down)/grid_eps;

for j=1:r
  for k=1:l
   z = left + j*grid_eps + (down + k*grid_eps)*i;
   B_1 = (((A-z*speye(si))^(2^n))')*(A-z*speye(si))^(2^n);
   B_2 = (A-z*speye(si))^(2^n)*((A-z*speye(si))^(2^n))';
   C_1 = B_1(1:si - (diag*(2^n)),1:si - (diag*(2^n)));
   C_2 = B_2(1:si - (diag*(2^n)),1:si - (diag*(2^n)));
   w = size(C,2); lambda = epsilon^(2^(n+1));
   [R,p_1] = chol(C_1 - lambda*speye(w));
   [R,p_2] = chol(C_2 - lambda*speye(w));
   if abs(max(p_1,p_2)) == 0
     else
     plot(z,'k.');
     hold on
   end
  end
end
```

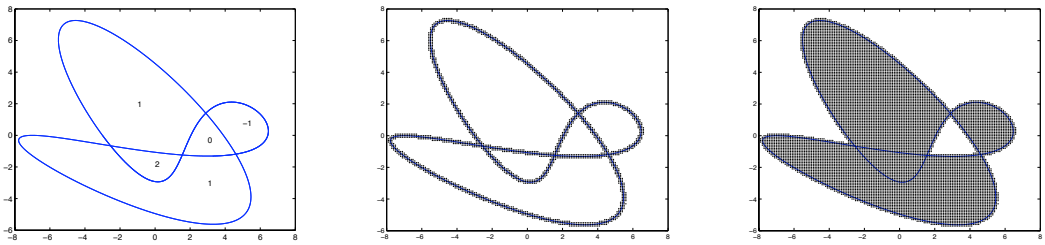


Figure 6.5: The first figure is the curve of the symbol f_1 with winding numbers, the second is the spectrum of the Laurent operator corresponding to f_1 computed with $\epsilon = 0.15$, $n = 2$, $m = 3000$ and grid-size being 0.1. The third figure is the spectrum of the Toeplitz operator corresponding to f_1 with the same numerical parameters as for the Laurent case.

6.4.4 Tests on Laurent and Toeplitz matrices

The spectral theory of Laurent and Toeplitz operators is very well understood, and they are therefore a natural choice when it comes to test objects for numerical algorithms. We briefly recall some of the basics from Laurent and Toeplitz operator theory. Given a Laurent operator A_L on $l^2(\mathbb{Z})$

$$A_L = \begin{pmatrix} \ddots & \vdots & \vdots & \vdots & \vdots & \\ \dots & a_0 & a_{-1} & a_{-2} & a_{-3} & \dots \\ \dots & a_1 & a_0 & a_{-1} & a_{-2} & \dots \\ \dots & a_2 & a_1 & a_0 & a_{-1} & \dots \\ \dots & a_3 & a_2 & a_1 & a_0 & \dots \\ & \vdots & \vdots & \vdots & \vdots & \ddots \end{pmatrix},$$

it is well known that A_L is a bounded operator if and only if there is a function $f \in L^\infty(\mathbb{T})$, where $\mathbb{T}$ denotes the circle, such that $\{a_n\}_{n=-\infty}^{\infty}$ is the sequence of Fourier coefficients of f, e.g.

$$a_n = \frac{1}{2\pi} \int f(e^{i\theta}) e^{-in\theta}\, d\theta, \quad n \in \mathbb{Z}.$$

Also, $\sigma(A_L) = \mathcal{R}(f)$, where $\mathcal{R}(f)$ denotes the essential range of f. For a Toeplitz operator A_T on $l^2(\mathbb{Z}_+)$, given by

$$A_T = \begin{pmatrix} a_0 & a_{-1} & a_{-2} & a_{-3} & \dots \\ a_1 & a_0 & a_{-1} & a_{-2} & \dots \\ a_2 & a_1 & a_0 & a_{-1} & \dots \\ a_3 & a_2 & a_1 & a_0 & \dots \\ \vdots & \vdots & \vdots & \vdots & \ddots \end{pmatrix},$$

we have a similar result, namely, A_T is bounded if and only if there is a function $f \in L^\infty(\mathbb{T})$ such that its Fourier coefficients are the sequence $\{a_n\}_{n\in\mathbb{Z}}$. The function f is called the symbol of the Laurent or Toeplitz operator.

As for the spectrum of A_T, note that $t \mapsto f(e^{it})$, $t \in [0, 2\pi]$ is a curve in $\mathbb{C}$, and hence we can assign a winding number to every point $z \in \mathbb{C}$ with respect to the curve. We

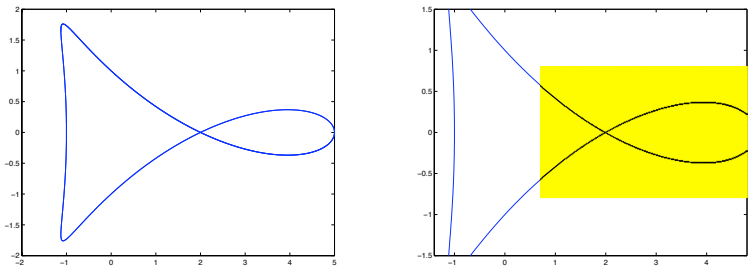


Figure 6.6: The first figure is the curve of the symbol f_2 and the second figure is the spectrum of the Laurent operator corresponding to f_2 computed (inside the rectangle) with $\epsilon = 0.1$, $n = 2$, $m = 3000$ and grid-size being 0.1.

then have that $\sigma(A_T)$ is equal to $\mathcal{R}(f)$ together with all complex numbers with non-zero winding number with respect to the curve.

In our examples (displayed in Figure 6.5 and Figure 6.6) we have chosen Laurent and Toeplitz operators with symbols

$$f_1(z) = 2z^{-3} - z^{-2} + 2iz^{-1} - 4z^2 - 2iz^3$$

and

$$f_2(z) = z^{-2} + z^{-1} + 1 + 2z,$$

where the corresponding winding numbers are displayed on the figures. Our numerical computation is done as suggested in Algorithm 6.4.1, where we check whether

$$\gamma_{n,m}(z) = \min\Big[\min\{\lambda^{1/2^{n+1}} : \lambda \in \sigma\left(P_m((T-z)^*)^{2^n}(T-z)^{2^n}\Big|_{P_m\mathcal{H}}\right)\},$$
$$\min\{\lambda^{1/2^{n+1}} : \lambda \in \sigma\left(P_m(T-z)^{2^n}((T-z)^*)^{2^n}\Big|_{P_m\mathcal{H}}\right)\}\Big]$$

is less than or equal to ϵ, for some $\epsilon > 0$, on a grid in the complex plane, where T here is either A_L or A_T. If $\gamma_{n,m}(z) \leq \epsilon$ the point z is assign a black color.

The choice of the projections is the natural one, namely, in the case of Laurent operators; P_m is the projection onto the span of $\{e_j\}_{j=-m}^{m}$ where $\{e_j\}_{j\in\mathbb{Z}}$ is the obvious basis for $l^2(\mathbb{Z})$ (i.e e_j has one on the j-th coordinate and zero elsewhere). For Toeplitz operators this is done similarly, but with P_m being the projection onto span$\{e_j\}_{j=1}^{m}$ and $\{e_j\}_{j\in\mathbb{N}}$ being the obvious basis for $l^2(\mathbb{N})$.

The computational costs of these figures are quite high due to large numbers of grid evaluations, and the computational time for some of them can typically take one night on a desktop computer. It is therefore difficult to get really accurate results. However, the computations done with the symbol f_2 in Figure 3 are done with a small grid-size to show accuracy.

6.5 Other Types of Pseudospectra

Even though the previous examples show very good results when computing spectra of Laurent and Toeplitz operators, one must be aware that the n-pseudospectrum can only

give an estimate on the position of the spectrum. The reason why the computations in the previous section are so close to the spectrum is simply because the n-pseudospectrum is close to the spectrum even for small n. This may of course not be the case in general, and we will now show how one can use the computations of $\gamma_{n,m}$ to determine subsets of the spectrum. The disadvantage of the n-pseudospectrum is that even though one can estimate the spectrum by taking n very large, n may have to be too large for practical purposes. Thus, since we only have the estimate for $T \in \mathcal{B}(\mathcal{H}), \epsilon > 0$ that $\sigma(T) \subset \sigma_{n,\epsilon}(T)$, it is important to get a "lower" bound on $\sigma(T)$ i.e. we want to find a set $\Omega \subset \mathbb{C}$ such that $\Omega \subset \sigma(T)$. A candidate for this is described in the following.

Definition 6.5.1. *Let $T \in \mathcal{B}(\mathcal{H})$ and define*

$$\zeta_1(z) = \min\left\{\lambda^{1/2} : \lambda \in \sigma\left((T-z)^*(T-z)\right)\right\},$$
$$\zeta_2(z) = \min\left\{\lambda^{1/2} : \lambda \in \sigma\left((T-z)(T-z)^*\right)\right\}.$$

Let $\epsilon > 0$ and define the ϵ-residual pseudospectrum to be the set

$$\sigma_{\mathrm{res},\epsilon}(T) = \{z : \zeta_1(z) > \epsilon,\, \zeta_2(z) = 0\},$$

and the adjoint ϵ-residual pseudospectrum to be the set

$$\sigma_{\mathrm{res}^*,\epsilon}(T) = \{z : \zeta_1(z) = 0,\, \zeta_2(z) > \epsilon\}.$$

Theorem 6.5.2. *Let $T \in \mathcal{B}(\mathcal{H})$ and let $\{T_k\} \subset \mathcal{B}(\mathcal{H})$ such that $T_k \to T$ in norm, as $k \to \infty$. Then for $\epsilon > 0$ we have the following,*

(i) $\sigma(T) \supset \bigcup_{\epsilon>0} \sigma_{\mathrm{res},\epsilon}(T) \cup \sigma_{\mathrm{res}^*,\epsilon}(T)$

(ii) $\mathrm{cl}(\{z \in \mathbb{C} : \zeta_1(z) < \epsilon\}) = \{z \in \mathbb{C} : \zeta_1(z) \leq \epsilon\}$

(iii) $\mathrm{cl}(\{z \in \mathbb{C} : \zeta_2(z) < \epsilon\}) = \{z \in \mathbb{C} : \zeta_2(z) \leq \epsilon\}$

(iv) For any compact ball $K \subset \mathbb{C}$ such that $\mathrm{cl}(\sigma_{\mathrm{res},\epsilon}(T)) \cap K^o \neq \emptyset$ it follows that

$$d_H(\mathrm{cl}(\sigma_{\mathrm{res},\epsilon}(T_k)) \cap K, \mathrm{cl}(\sigma_{\mathrm{res},\epsilon}(T)) \cap K) \longrightarrow 0, \qquad k \to \infty.$$

(v) For any compact ball $K \subset \mathbb{C}$ such that $\sigma_{\mathrm{res}^*,\epsilon}(T) \cap K^o \neq \emptyset$ it follows that

$$d_H(\mathrm{cl}(\sigma_{\mathrm{res}^*,\epsilon}(T_k)) \cap K, \mathrm{cl}(\sigma_{\mathrm{res}^*,\epsilon}(T) \cap K)) \longrightarrow 0, \qquad k \to \infty.$$

Proof. This is Theorem 3.6.2 in Chapter 3 and a proof can be found there. □

The previous theorem shows that the residual and the adjoint residual pseudospectra have similar continuity properties as the pseudospectra. Hence, these sets are suitable for computations. The approximations are very similar to the techniques we have used in the previous sections.

Theorem 6.5.3. *Let $T \in \mathcal{B}(\mathcal{H})$ and suppose that $\{P_m\}$ is a sequence of finite rank projections converging strongly to the identity such that $P_{m+1} \geq P_m$. Define*

$$\zeta_{1,m}(z) = \min\left\{\lambda^{1/2} : \lambda \in \sigma\left(P_m(T-z)^*(T-z)\Big|_{P_m\mathcal{H}}\right)\right\},$$
$$\zeta_{2,m}(z) = \min\left\{\lambda^{1/2} : \lambda \in \sigma\left(P_m(T-z)(T-z)^*\Big|_{P_m\mathcal{H}}\right)\right\}$$

and

$$\zeta_{1,m,k}(z) = \min\left\{\lambda^{1/2} : \lambda \in \sigma\left(P_m(P_k(T-z)P_k)^*(P_k(T-z)P_k)\Big|_{P_m\mathcal{H}}\right)\right\},$$
$$\zeta_{2,m,k}(z) = \min\left\{\lambda^{1/2} : \lambda \in \sigma\left(P_m(P_k(T-z)P_k)(P_k(T-z)P_k)^*\Big|_{P_m\mathcal{H}}\right)\right\}.$$

Let $\delta \in (0, \epsilon)$. Then we have the following.

(i) If K is a compact ball such that $\overline{\sigma_{\mathrm{res},\epsilon}(T)} \cap K^o \neq \emptyset$ then

$$\mathrm{cl}(\{z : \zeta_{1,m}(z) > \epsilon, \zeta_{2,m}(z) < \delta\}) \cap K \longrightarrow \overline{\sigma_{\mathrm{res},\epsilon}(T)} \cap K, \qquad m \to \infty.$$

(ii) If K is a compact ball such that $\overline{\sigma_{\mathrm{res}^*,\epsilon}(T)} \cap K^o \neq \emptyset$ then

$$\mathrm{cl}(\{z : \zeta_{1,m}(z) < \delta, \zeta_{2,m}(z) > \epsilon\}) \cap K \longrightarrow \overline{\sigma_{\mathrm{res}^*,\epsilon}(T)} \cap K, \qquad m \to \infty.$$

(iii) If K is a compact ball such that $\mathrm{cl}(\{z : \zeta_{1,m}(z) > \epsilon, \zeta_{2,m}(z) < \delta\}) \cap K^o \neq \emptyset$ then

$$\mathrm{cl}(\{z : \zeta_{1,m,k}(z) > \epsilon, \zeta_{2,m,k}(z) < \delta\}) \cap K \longrightarrow \mathrm{cl}(\{z : \zeta_{1,m}(z) > \epsilon, \zeta_{2,m}(z) < \delta\}) \cap K, \qquad k \to \infty.$$

(iv) If K is a compact ball such that $\mathrm{cl}(\{z : \zeta_{1,m}(z) < \epsilon, \zeta_{2,m}(z) > \epsilon\}) \cap K^o \neq \emptyset$ then

$$\mathrm{cl}(\{z : \zeta_{1,m,k}(z) < \delta, \zeta_{2,m,k}(z) > \epsilon\}) \cap K \longrightarrow \mathrm{cl}(\{z : \zeta_{1,m}(z) < \delta, \zeta_{2,m}(z) > \epsilon\}) \cap K, \qquad k \to \infty.$$

Proof. A proof of this theorem can be found in the proof of Theorem 3.6.3 in Chapter 3. □

We now have a computational tool for estimating the spectrum both from "above" and "below", meaning that for $T \in \mathcal{B}(\mathcal{H})$ we have

$$\sigma_{\mathrm{res},\epsilon}(T) \cup \sigma_{\mathrm{res}^*,\epsilon}(T) \subset \sigma(T) \subset \sigma_{n,\epsilon}(T).$$

Thus, it would be natural to compute, for $\epsilon > 0$ and $\delta \in (0, \epsilon)$, both

$$\{z : \zeta_{1,m,k}(z) > \epsilon, \zeta_{2,m,k}(z) \leq \delta\} \cup \{z : \zeta_{1,m,k}(z) \leq \delta, \zeta_{2,m,k}(z) > \epsilon\}$$

and

$$\{z : \gamma_{n,m,k}(z) \leq \epsilon\},$$

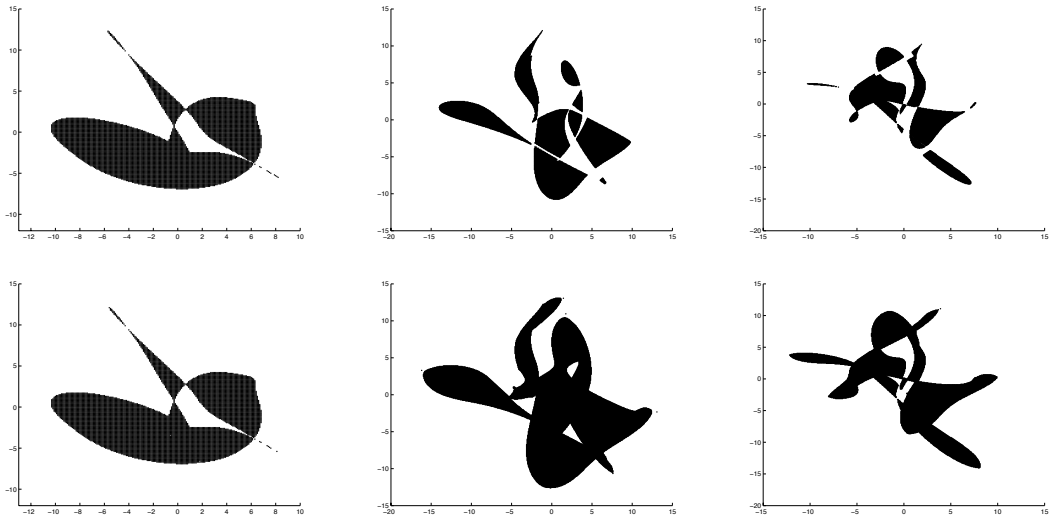


Figure 6.7: The figure shows on the first row $\Omega_{\epsilon,\epsilon-10^{-10},3000}(T_1)$ ($\epsilon = 0.01$), $\Omega_{\epsilon,\epsilon-10^{-10},3000}(T_2)$ ($\epsilon = 0.0005$), $\Omega_{\epsilon,\epsilon-10^{-10},3000}(T_4)$ ($\epsilon = 0.00001$) and on the second row $\{z : \gamma_{0,m}(z) \leq \epsilon\}$ ($m = 3000$) for the same operators and ϵs as on the first row.

where $\gamma_{n,m,k}$ is defined as in (6.4.1) to get an estimate for the spectrum. To simplify the notation we define

$$\Omega_{\epsilon,\delta,m}(T) = \{z : \zeta_{1,m}(z) > \epsilon, \zeta_{2,m}(z) \leq \delta\} \cup \{z : \zeta_{1,m}(z) \leq \delta, \zeta_{2,m}(z) > \epsilon\} \tag{6.5.1}$$

and

$$\Omega_{\epsilon,\delta,m,k}(T) = \{z : \zeta_{1,m,k}(z) > \epsilon, \zeta_{2,m,k}(z) \leq \delta\} \cup \{z : \zeta_{1,m,k}(z) \leq \delta, \zeta_{2,m,k}(z) > \epsilon\}. \tag{6.5.2}$$

Example 6.5.4. Given an infinite matrix T, we will in this example show how computations of

$$\sigma_{\mathrm{res},\epsilon}(T) \cup \sigma_{\mathrm{res}*,\epsilon}(T), \qquad \sigma_{n,\epsilon}(T)$$

can give quite good estimates on the position of the spectrum. As test objects we have chosen Toeplitz like operators, where we have kept much of the Toeplitz structure, but let some of the subdiagonals have alternating numbers instead of constants. As we are left with few (if any) mathematical tools to estimate the spectrum, we can only rely on the computed estimate, which in some cases seems quite acceptable. Consider the three infinite matrices

$$T_1 = \begin{pmatrix} 0 & a & b & c & 0 & 0 & \dots \\ d & 0 & a & b & c & 0 & \dots \\ f & e & 0 & a & b & c & \dots \\ g & f & d & 0 & a & b & \dots \\ 0 & g & f & e & 0 & a & \dots \\ 0 & 0 & g & f & d & 0 & \dots \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \ddots \end{pmatrix}, \quad T_2 = \begin{pmatrix} 0 & a & b & c & 0 & 0 & \dots \\ d & 0 & a & b & c & 0 & \dots \\ f & e & 0 & a & b & c & \dots \\ g & f & d & 0 & a & b & \dots \\ \phi_1 & g & f & e & 0 & a & \dots \\ 0 & \psi_1 & g & f & d & 0 & \dots \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \ddots \end{pmatrix},$$

and

$$T_3 = \begin{pmatrix} 0 & a & b & c & 0 & 0 & \dots \\ d & 0 & a & b & c & 0 & \dots \\ 0 & e & 0 & a & b & c & \dots \\ g & 0 & d & 0 & a & b & \dots \\ \phi_1 & g & 0 & e & 0 & a & \dots \\ 0 & \psi_1 & g & 0 & d & 0 & \dots \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \ddots \end{pmatrix},$$

where $a = 1 + 2i$, $b = -1$, $c = 5 + i$, $d = -2$, $e = 1 + 2i$, $f = -4$, $g = -1 - 2i$, $\phi_j = -2 + \frac{-5+15i}{j^{1/6}}$ and $\psi_j = 1 + 2i + \frac{5+15i}{j^{1/3}}$. Figure 6.7 shows computations of $\Omega_{\epsilon,\delta,m}(T_j)$ (where $\Omega_{\epsilon,\delta,m}(\cdot)$ is defined as in (6.5.2)) and

$$\{z : \gamma_{0,m}(z) \leq \epsilon\}$$

for T_j $(j = 1, 2, 3)$, $m = 3000$ and $\epsilon = 10^{-5}$, $\delta = \epsilon - 10^{-10}$. Since

$$\overline{\Omega_{\epsilon,\delta,m}(T)} \longrightarrow \overline{\sigma_{\mathrm{res},\epsilon}(T)} \cup \overline{\sigma_{\mathrm{res}*,\epsilon}(T)} \subset \sigma(T), \qquad m \to \infty$$

and

$$\{z : \gamma_{0,m}(z) \leq \epsilon\} \longrightarrow \overline{\sigma_\epsilon(T)}, \qquad m \to \infty,$$

it is reasonable to believe that the computation displays the following relation

$$\Omega_{\epsilon,\delta,m}(T) \subset \omega_\nu(\sigma(T)) \subset \{z : \gamma_{0,m}(z) \leq \epsilon\}, \qquad \nu > 0,$$

for some ν. As we tried this with several larger values of m up to $m = 10000$ without noticing any change, it suggests that ν is small in the experiment with T_1, where "small" here means relative to the resolution of the figures displayed.

6.6 Discrete Schrödinger Operators

6.6.1 The Non-self-adjoint Almost Mathieu Operator

An important operator in non-self-adjoint spectral theory is the non-self-adjoint harmonic oscillator H, defined by

$$Hf(x) = -f''(x) + cx^2 f(x),$$

acting on $L^2(\mathbb{R})$. One of the motivations for this operator was that one wanted to take a well known self-adjoint operator, alter it slightly so that it becomes non-self-adjoint, and then see how the spectral properties change. Indeed, the spectral properties of the non-self-adjoint harmonic oscillator are very different from the usual harmonic oscillator, as discussed in (DK04)(TE05). Our approach is to do the same with discrete Schrödinger operators.

The almost Mathieu operator on $l^2(\mathbb{Z})$ is known from the Ten Martini Problem, a problem that was initiated in 1981 by Kac and Simon and finally solved in 2003 by Puig (Pui04). The operator is defined as

$$(H_{b,\phi,\omega}x)_n = x_{n+1} + x_{n-1} + b\cos(2\pi\omega n + \phi)x_n, \qquad n \in \mathbb{Z},$$

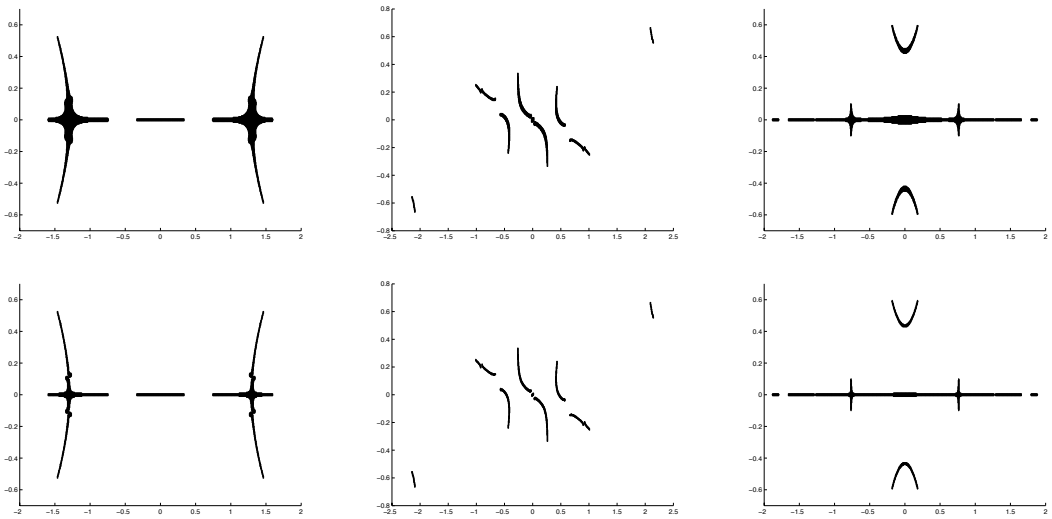


Figure 6.8: The first row shows $\sigma_\epsilon(H_{i,2,\pi/4})$, $\sigma_\epsilon(H_{1+i,2,\pi/4})$ and $\sigma_\epsilon(H_{i,2,\sqrt{5}})$ for $\epsilon = 0.005$. The second row shows $\sigma_{1,\epsilon}(H_{b,\phi,\omega})$ for the same values as in the first row.

where $\omega > 0$ is an irrational number, $\phi \in \mathbb{T} = \mathbb{R}/2\pi\mathbb{Z}$ and $b \in \mathbb{C}$. The usual almost Mathieu operator has $b \in \mathbb{R}$, so that $H_{b,\phi}$ is self-adjoint and the Ten Martini problem was to show that for real non-zero b then $\sigma(H_{b,\phi,\omega})$ is a cantor set.

We do not claim anything about the spectral properties of the non-self-adjoint almost Mathieu operator (NSAM operator), but we rather use it as an example of an operator where we before did not have computational tools at hand to handle the problem of numerically estimating the spectrum. Arveson gave a complete theory in (Arv94a) on how to handle the computational aspects of the spectral theory of the self-adjoint almost Mathieu operator. However, self-adjointness is crucial in Arvesons theory and therefore not suitable for our problems. But with the techniques suggested in the earlier sections of this chapter we can get numerical approximations to the spectra of these non-self-adjoint Schrödinger operators. In Figure 6.8 we have computed pseudospectra and 1-pseudospectra of the NSAM operator for different values of b and ω.

6.6.2 Random Non-self-adjoint Schrödinger operators

In this section we will consider the non-self-adjoint Anderson model that has several applications in physics (Dav01)(TE05). More specifically we will consider the operator H on $l^2(\mathbb{Z})$ defined by

$$(Hx)_n = e^{-g}x_{n-1} + e^{g}x_{n+1} + V_n x_n,$$

where $g > 0$ and V is a random real valued potential taking values from an interval [-B,B] according to some probability measure μ on [-B,B]. Faced with the problem of computing the spectrum or pseudospectra of H we immediately discard the finite section method due to Davies' analysis of the problem in (Dav01). In particular Davies pointed out that one must resist the temptation of projecting down to $l^2(-N, N)$ and impose boundary conditions because the spectrum of H on $l^2(-N, N)$ as $N \to \infty$ may have little to do with

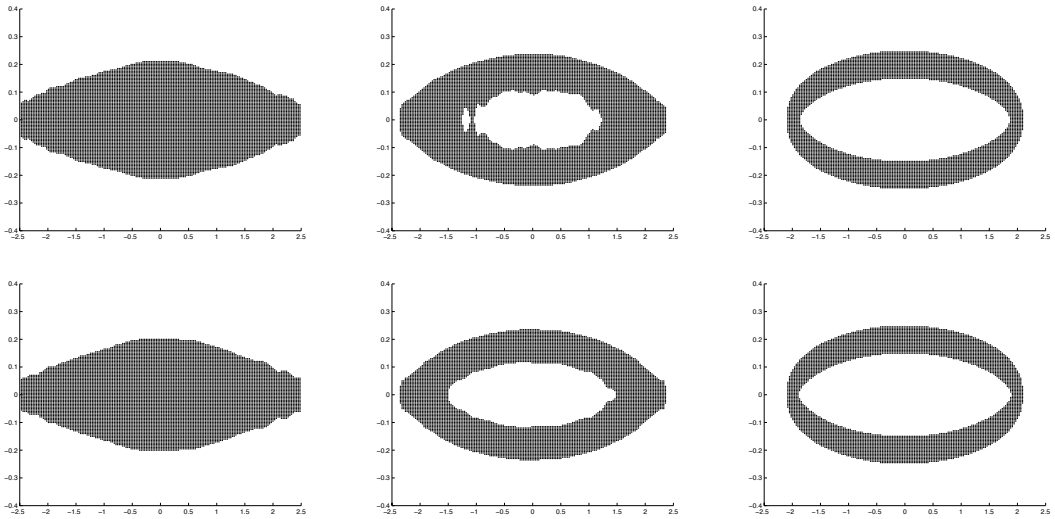


Figure 6.9: The first row shows $\{z : \gamma_{0,m}(z) \leq \epsilon\}$ for $B = 1, 0.5, 0.1$ and the second row shows $\{z : \gamma_{0,m}(z) \leq \epsilon\}$ for the same values of B. In both examples $m = 20000$ and $\epsilon = 0.05$.

the spectrum of H on $l^2(\mathbb{Z})$ regardless of the choice of boundary conditions. Having left the finite section approach, the method of computing the function $\gamma_{n,m}$ comes in handy as this will at least help us estimating the n-pseudospectrum of H.

Davies showed in (Dav01) that if $B < e^g - e^{-g}$ then

$$\sigma(H) \cap \{z \in \mathbb{C} : |z| < r\} = \emptyset,$$

where $r = e^g - e^{-g} - B > 0$ Also, if $B \geq e^g + e^{-g}$ then

$$\sigma(H) = E + [-B, B], \quad \text{a.s.},$$

where $E = \{e^{g+i\theta} + e^{-g-i\theta} : \theta \in [0, 2\pi]\}$ and a.s. means almost surely. But the question is what happens when $e^g - e^{-g} < B < e^g + e^{-g}$? The purpose of our numerical computation is not to investigate this seriously, but rather demonstrate that we now have a tool that could be used if one wants to use numerical simulations to get some insight on this problem. We have tested this only for one sample of H, and of course, since H is a random operator, one has to compute $\sigma(H)$ sufficiently many times to get a reliable result. However, this is just an example, and we postpone more serious computations for future work.

In this example we have estimated $\sigma_\epsilon(H)$ and $\sigma_{1,\epsilon}(H)$, where V is uniformly distributed, by computing

$$\{z : \gamma_{0,m}(z) \leq \epsilon\}, \quad \{z : \gamma_{1,m}(z) \leq \epsilon\}$$

for $g = 0.1$ (so that $e^{g+i\theta} + e^{-g-i\theta} = 2.01$ and $e^{g+i\theta} - e^{-g-i\theta} = 0.2003$), $\epsilon = 0.05$, $m = 20000$ and $B = 1, 0.5, 0.1$. An interesting observation is that for $B = 0.5$ the computation in Figure 6.9 suggests that the spectrum has a hole.

Chapter 7

The Infinite-Dimensional QR algorithm

In numerical linear algebra the QR algorithm (Wil65) (Par65) (PK69) (Wat82) represents the state of the art in the computation of the algebraic eigenvalue problem. The algorithm is exceedingly powerful, and is now the core of, for example, the MATLAB command `eig`. Now, all is well with the QR algorithm as long as the matrix we are working with has finitely many entries, but the problem is that not all matrices may have this nice property. In fact, most of the operators in mathematical physics act on an infinite dimensional space and hence a matrix representation of such an operator results in an infinite matrix. The problem of computing the spectrum of an infinite matrix may be very different from the finite dimensional case, however, it is not less important, as the set of infinite dimensional problems contains e.g. Schrödinger spectral problems and other problems related to quantum mechanics. In Chapter 6 several new methods for computing spectra of infinite matrices were introduced. The theoretical framework in Chapter 6 and Chapter 3 shows how one can compute the spectrum of an infinite matrix as long as it is a bounded operator on $l^2(\mathbb{N})$. Although this gives a way of computing spectra of arbitrary infinite matrices, there is a slight disadvantage, namely, the framework is based on pseudospectral theory, and isolated eigenvalues are therefore difficult to detect. Often one may not be able to determine the difference between an isolated eigenvalue or a small neighborhood of the eigenvalue. This will become increasingly difficult with non-normal operators. An alternative to the pseudospectral framework in Chapter 6 is the Infinite dimensional QR algorithm or just the Infinite QR algorithm as we will call it. The algorithm in infinite dimensions is very similar to the finite dimensional case, and the implementation is based on Householder transformations. Theoretically one can analyze the convergence of the algorithm as if it was carried out on an infinite computer (meaning that we allow storage and multiplication of infinite matrices), however, in practice there are tricks one can use to avoid this impossible assumption. In fact, as we will see later, there are no truncations or approximations done in the algorithm, and the information provided by the computation is (up to ϵ_{mach}) as if it was done with an infinite computer. Thus, one can conclude that, indeed, this is infinite dimensional numerical linear algebra.

The Infinite QR algorithm has existed as a pure mathematical concept for more than twenty years and it first appeared in the paper "Toda Flows with Infinitely Many Variables" (DLT85) in 1985. However, the framework presented here and in Chapter 2 was

developed independently and the author is indebted to Percy Deift for pointing out the connection to (DLT85). We will in this paper concentrate on the applied and computational properties of the Infinite QR algorithm and refer to Chapter 2 and (DLT85) for a more theoretical exposition.

7.1 Pollution in the Finite Section Method

Let us recall the basic ideas of the finite section method as discussed in Chapter 6. Suppose that we have an operator $A \in \mathcal{B}(\mathcal{H})$ and that we know the matrix elements $a_{ij} = \langle Ae_j, e_i \rangle$ with respect to some basis $\{e_j\}$. A quite common idea is to reduce this to a finite-dimensional spectral problem. One constructs (using $\{e_j\}$) a sequence of finite rank projections $\{P_m\}$ such that $P_{m+1} \geq P_m$ and $P_m \to I$ strongly, where I is the identity, and then compute the spectrum and pseudospectra of $P_m A\lceil_{P_m\mathcal{H}}$. Typically, P_m would be the projection onto $\text{span}\{e_1, \ldots, e_m\}$, but other choices are also possible. This is often referred to as the finite section method in the literature. It is well known that this may work in some cases e.g. if the operator is compact, or in some cases when the operator is self-adjoint. However, one must be very careful using the finite section method, and it should not be used unless accompanied by a rigorous analysis that justifies the convergence

$$d_H(\sigma(P_m A\lceil_{P_m\mathcal{H}}), \sigma(A)) \longrightarrow 0, \qquad m \to \infty.$$

As discussed before in Chapter 6, it is quite easy to find elementary counter examples to show that the finite section method can fail dramatically. Consider the shift operator $Se_n = e_{n+1}$ on $l^2(\mathbb{N})$. This operator has the following matrix representation

$$S = \begin{pmatrix} 0 & 1 & 0 & 0 & \ldots \\ 0 & 0 & 1 & 0 & \ldots \\ 0 & 0 & 0 & 1 & \ldots \\ 0 & 0 & 0 & 0 & \ldots \\ \vdots & \vdots & \vdots & \vdots & \ddots \end{pmatrix}. \tag{7.1.1}$$

Thus, if P_m is the projection onto $\text{span}\{e_1, \ldots, e_m\}$, we would get that $\sigma((P_m S\lceil_{P_m\mathcal{H}}) = \{0\}$ for all m, but $\sigma(S)$ is the closed unit disc. This example showed that we can have $\sigma(P_m T\lceil_{P_m\mathcal{H}}) \subset \sigma(T)$, where the inclusion is proper for all m, and $T \in \mathcal{B}(\mathcal{H})$. But we can also have that $\sigma(P_m T\lceil_{P_m\mathcal{H}}) \nsubseteq \sigma(T)$, as the following example shows. Let

$$A = \begin{pmatrix} 0 & 0 & 0 & 0 & 0 & 0 & \ldots \\ 1 & 0 & a_{23} & a_{24} & 0 & 0 & \ldots \\ 0 & 1 & 0 & 0 & 0 & 0 & \ldots \\ 0 & 0 & 1 & 0 & a_{45} & a_{46} & \ldots \\ 0 & 0 & 0 & 1 & 0 & 0 & \ldots \\ 0 & 0 & 0 & 0 & 1 & 0 & \ldots \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \ddots \end{pmatrix},$$

where $a_{2j,j+1} = 1$. Also, $a_{2j,2j+2} = -i$ if j is prime and $a_{2j,2j+2} = 0$ otherwise. If we let P_m be the projection onto $\text{span}\{e_1, \ldots, e_m\}$ and compute $\sigma(P_m A\lceil_{P_m\mathcal{H}})$ then the phenomenon "spectral pollution" occurs, namely, that $\sigma(P_m A\lceil_{P_m\mathcal{H}})$ contains elements

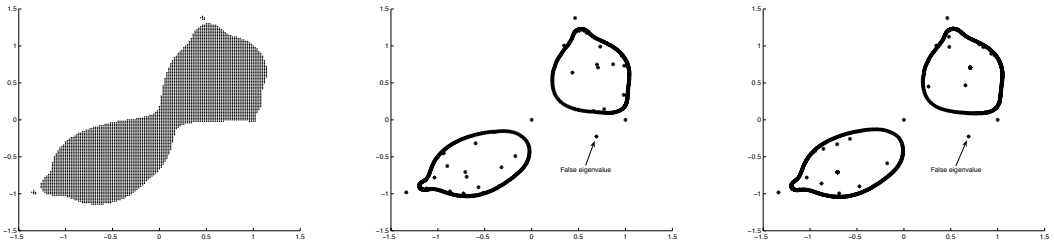


Figure 7.1: The first figure shows $\sigma_\epsilon(A)$ for $\epsilon = 0.02$ and the next figures shows $\sigma(P_m A\lceil_{P_m\mathcal{H}})$ for $m = 700$ and $m = 900$ with the false eigenvalue.

that have nothing to do with $\sigma(A)$. This is visualized in Figure 7.1. The phenomenon described in this example is often referred to as spectral pollution and is well known. As the following theorem suggest, the "pollution" can be arbitrarily bad.

Theorem 7.1.1. *(Pokrzywa)(Pok79) Let $A \in \mathcal{B}(\mathcal{H})$ and $\{P_n\}$ be a sequence of finite-dimensional projections converging strongly to the identity. Suppose that $S \subset W_e(A)$. Then there exists a sequence $\{Q_n\}$ of finite-dimensional projections such that $P_n < Q_n$ (so $Q_n \to I$ strongly) and*

$$d_H(\sigma(A_n) \cup S, \sigma(\tilde{A}_n)) \to 0, \quad n \to \infty,$$

where

$$A_n = P_n A\lceil_{P_n\mathcal{H}}, \qquad \tilde{A}_n = Q_n A\lceil_{Q_n\mathcal{H}}$$

and d_H denotes the Hausdorff metric.

However, life is not completely dark for the finite section method as the following result shows. (Also, in the self-adjoint case the finite section method sometimes perform quite well (Arv94a) (Bro07b) (Han08)).

Theorem 7.1.2. *Let $T \in \mathcal{B}(\mathcal{H})$ and $\{P_n\}$ be a sequence of finite-dimensional projections converging strongly to the identity. If $\lambda \notin W_e(T)$ then $\lambda \in \sigma(T)$ if and only if*

$$d(\lambda, \sigma(P_n T\lceil_{P_n\mathcal{H}})) \longrightarrow 0, \qquad n \to \infty.$$

However, if we want to use the finite section method and rely on Theorem 7.1.2 we must know $W_e(T)$, and that may be unpleasant to compute. Or we could hope that $\sigma_e(T) = W_e(T)$ and try to estimate $r_e(T)$, but that may also be a highly non-trivial problem. Now, it is known that if an operator T is hypo-normal ($T^*T - TT^* \geq 0$) then

$$\text{conv}(\sigma_e(T)) = W_e(T),$$

where $\text{conv}(\sigma_e(T))$ denotes the convex hull of $\sigma_e(T)$. But what if we have a "very non-normal" operator? Another problem we may encounter using the finite section method is that even though $\sigma_d(T)$ may be recovered, one may get a very misleading picture of the rest of the spectrum. Such problems are illustrated in the following examples. Unless otherwise stated P_m will denote the projection onto $\text{span}\{e_1, \ldots, e_m\}$. Let

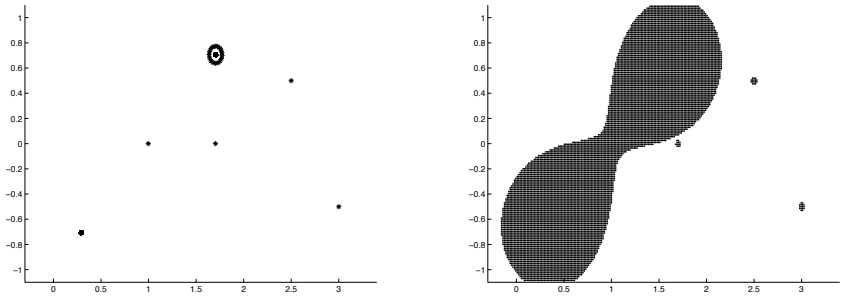


Figure 7.2: The first figure shows $\sigma(P_m A\lceil_{P_m\mathcal{H}})$ for $m = 500$ and the second figure shows $\sigma_\epsilon(A)$ for $\epsilon = 0.02$.

$$A = \begin{pmatrix} 2.5+0.5i & 0 & 0 & 0 & 0 & 0 & 0 & \dots \\ 1 & 3-0.5i & 0 & 0 & 0 & 0 & 0 & \dots \\ 0 & 1 & 1.7 & 0.05 & 0 & 0 & 0 & \dots \\ 0 & 0 & 0.05 & 1 & 0 & 0 & 0 & \dots \\ 0 & 0 & 0 & 1 & 1 & a_{56} & 0 & \dots \\ 0 & 0 & 0 & 0 & 1 & 1 & 0 & \dots \\ 0 & 0 & 0 & 0 & 0 & 1 & 1 & \dots \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \ddots \end{pmatrix},$$

where $a_{2j-1,2j} = i$ for $j \geq 3$. Now, A can be written as a compact perturbation of the operator

$$I + S^* + B, \qquad B = 0 \oplus \tilde{B} \quad \text{on} \quad P_m\mathcal{H} \oplus P_m^\perp\mathcal{H}, \quad m = 4,$$

where S is the shift operator (recall (7.1.1)) and

$$\langle \tilde{B}\tilde{e}_{2j}, \tilde{e}_{2j-1}\rangle = i, \qquad \langle \tilde{B}\tilde{e}_j, \tilde{e}_i\rangle = 0 \quad \text{elsewhere}, \quad i, j \in \mathbb{N},$$

where $\{\tilde{e}_j\}$ is the basis for $P_m^\perp\mathcal{H}$. Thus, A should have some isolated eigenvalues of finite multiplicity as well as some essential spectrum. Using the techniques introduced in (Han10) we can compute the pseudospectrum of A (Figure 7.2) to reveal that at least some of the isolated eigenvalues produced by the finite section method appear to be correct. Although, without the support from the picture of the pseudospectrum, the information from the finite section method would have been useless since one does not know the essential numerical range of A.

Another example of misleading information from the finite section method is the follwing. Let

$$T = \begin{pmatrix} 2.5+0.5i & 0 & 0 & 0 & 0 & 0 & 0 & \dots \\ 1 & 3-0.5i & 0 & 0 & 0 & 0 & 0 & \dots \\ 0 & 1 & 1.7 & 0.05 & 0 & 0 & 0 & \dots \\ 0 & 0 & 0.05 & t_4 & 0 & 0 & 0 & \dots \\ 0 & 0 & 0 & 0 & t_5 & 0 & 0 & \dots \\ 0 & 0 & 0 & 0 & 1 & t_6 & 0 & \dots \\ 0 & 0 & 0 & 0 & 0 & 1 & t_7 & \dots \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \ddots \end{pmatrix},$$

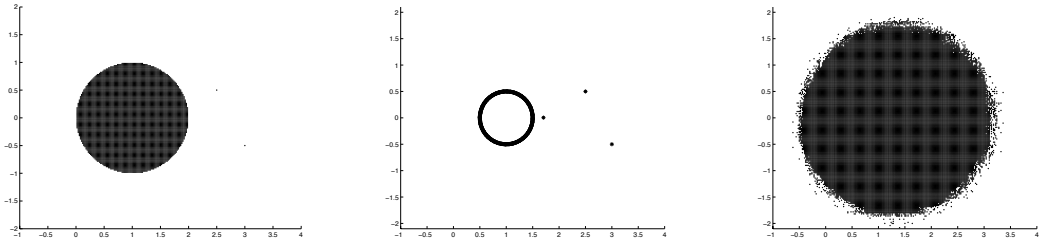


Figure 7.3: The first figure shows $\sigma(T)$, the second $\sigma(P_m T\lceil_{P_m\mathcal{H}})$ and $\sigma(P_m\tilde{T}\lceil_{P_m\mathcal{H}})$ (they are identical) and the third shows $\sigma_\epsilon(\tilde{T})$ (or rather a desperate try) for $\epsilon = 5\times 10^{-8}$.

where $t_j = 1 + 0.5(\sin j, \cos j)$. Using techniques from (Han10) we can compute the spectrum of T, and this is shown in Figure 7.3. Now, using the finite section method on T, we observe that, even though all the points that are displayed (Figure 7.3) are actually in the spectrum of T, the result is far from correct. To complicate the task slightly we perturb T by introducing a large number on the subdiagonal. In particular let

$$\tilde{T} = \begin{pmatrix} 2.5+0.5i & 0 & 0 & 0 & 0 & 0 & 0 & \dots \\ 1 & 3-0.5i & 0 & 0 & 0 & 0 & 0 & \dots \\ 0 & 1 & 1.7 & 0.05 & 0 & 0 & 0 & \dots \\ 0 & 0 & 0.05 & t_4 & 0 & 0 & 0 & \dots \\ 0 & 0 & 0 & 5\times 10^5 & t_5 & 0 & 0 & \dots \\ 0 & 0 & 0 & 0 & 1 & t_6 & 0 & \dots \\ 0 & 0 & 0 & 0 & 0 & 1 & t_7 & \dots \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \ddots \end{pmatrix}.$$

If we should use the pseudospectrum as a guide (as done in Figure 7.2) we see that this is quite difficult (Figure 7.3) due to the fact that $\tilde{T}$ is highly non-normal. The figure has in fact very little to do with the pseudospectrum as the roundoff error interferes in the computation. However, our point is that, in this case, very little can be said about the spectrum of $\tilde{T}$ if we should only rely on Figure 7.3. In Section 7.5.1 we will see that the Infinite QR algorithm gives quite satisfactory results on these examples.

7.2 The QR decomposition

The QR decomposition is the core of the QR algorithm. If $A \in \mathbb{C}^{n\times n}$, one may apply the Gram-Schmidt procedure to the columns of A and store these columns in a matrix Q and this gives us the QR decomposition

$$A = QR, \tag{7.2.1}$$

where Q is a unitary matrix and R upper triangular. It is therefore no surprise that a QR decomposition should exist in the infinite dimensional case, however, we need more than just the existence. A key ingredient in the QR algorithm is the Householder transformation, and for computational reasons one uses Householder transformations when computing the QR decomposition. It is therefore crucial that we can adopt these tools in the infinite dimensional setting. Our goal is therefore to extend the construction of the

QR decomposition, via Householder transformations, to infinite matrices. And moreover, to find a way so that one can implement the procedure on a computer.

7.2.1 Householder Reflections

Before we can state and prove the main theorem we need to introduce the concept of Householder reflections in an infinite-dimensional setting.

Definition 7.2.1. *A Householder reflection is an operator $S \in \mathcal{B}(\mathcal{H})$ of the form*

$$S = I - \frac{2}{\|\xi\|^2}\xi \otimes \bar{\xi}, \qquad \xi \in \mathcal{H}. \tag{7.2.2}$$

In the case where $\mathcal{H} = \mathcal{H}_1 \oplus \mathcal{H}_2$ and I_i is the identity on $\mathcal{H}_i$ then

$$U = I_1 \oplus \left(I_2 - \frac{2}{\|\xi\|^2}\xi \otimes \bar{\xi} \right) \qquad \xi \in \mathcal{H}_2,$$

will be called a Householder transformation.

A straightforward calculation shows that $S^* = S^{-1} = S$ and thus also $U^* = U^{-1} = U$. An important property of the operator S is that if $\{e_j\}$ is an orthonormal basis for $\mathcal{H}$ and $\eta \in \mathcal{H}$ then one can choose $\xi \in \mathcal{H}$ such that

$$\langle S\eta, e_j \rangle = \langle (I - \frac{2}{\|\xi\|^2}\xi \otimes \bar{\xi})\eta, e_j \rangle = 0, \qquad j \neq 1.$$

Indeed, if $\eta_1 = \langle \eta, e_1 \rangle \neq 0$ one may choose $\xi = \eta \pm \|\eta\|\zeta$, where $\zeta = \eta_1/|\eta_1|e_1$ and if if $\eta_1 = 0$ choose $\xi = \eta \pm \|\eta\|e_1$, The verification of the assertion is a straightforward calculation. Note that the notation

$$\xi \otimes \bar{\xi}, \qquad \xi \in \mathcal{H},$$

deviates from what one is used to in finite dimensions. Recall that in finite dimensions a Householder reflection is often expressed as

$$I - \frac{2}{\|x\|^2} x\bar{x}^T, \qquad x \in \mathbb{C}^n,$$

however, in a coordinate-free Hilbert space $\mathcal{H}$ the notation $\xi\bar{\xi}^T$ does not make any sense for $\xi \in \mathcal{H}$. One therefore uses the correct notation as in (7.2.2) and recalls that

$$\xi \otimes \bar{\xi}(\eta) = \langle \xi, \eta \rangle \xi, \qquad \eta \in \mathcal{H}.$$

7.2.2 Constructing the QR decomposition

We will now explain how to construct the QR decomposition in infinite dimensions. The approach is for now on only mathematical and we will in Section 7.4 discuss how to use certain tricks so that this procedure is actually implementable on a computer. First, we recall how this is done in finite dimensions. Let $A \in \mathbb{C}^{n\times n}$. We may construct R in (7.2.1) by multiplying A from the left by Householder transformations $U_1, \ldots, U_n$ such that

$$U_n \cdots U_1 A = R,$$

and R is upper triangular. Now, the fact that every Householder reflection is a unitary matrix makes

$$A = U_1 \cdots U_n R = QR$$

the desired composition. In infinite dimensions we will do exactly the same, however, we are faced with the challenge of making sense of an infinite product of Householder transformations. We will explain this more carefully. Let $A \in \mathcal{B}(l^2(\mathbb{N}))$. We will now obtain R as a limit of the infinite matrices $U_n \cdots U_1 A$, where U_j is a Householder reflection e.g.

$$R = \lim_{n\to\infty} U_n \cdots U_1 A,$$

where the limit is understood to be in an appropriate topology on $\mathcal{B}(l^2(\mathbb{N}))$. We will be more specific about this topology later. First, let us concentrate on how to construct elements $U_n \cdots U_1 A$ and worry about the limit later. Let P_n be the projection onto $\{e_1, \ldots, e_n\}$ and suppose that we have the n elements in the sequence and that the n-th element is an operator $R_n = U_n \cdots U_1 A$ such that, with respect to the decomposition $\mathcal{H} = P_n\mathcal{H} \oplus P_n^\perp \mathcal{H}$, (here we use $\mathcal{H} = l^2(\mathbb{N})$ to simplify notation) we have

$$R_n = \begin{pmatrix} \tilde{R}_n & B_n \\ C_n & N_n \end{pmatrix}, \quad \tilde{R}_n = P_n R_n P_n,\ B_n = P_n R_n P_n^\perp,\ C_n = P_n^\perp R_n P_n,$$

where $N_n = P_n^\perp R_n P_n^\perp$ and $\tilde{R}$ is upper triangular and $Ce_j = 0$ for $j \leq n-1$. Let $\zeta = Ce_n$. Choose $\xi \in P_n^\perp \mathcal{H}$ and define the Householder reflection $S \in \mathcal{B}(P_n^\perp \mathcal{H})$,

$$S = I - \frac{2}{\|\xi\|^2} \xi \otimes \bar{\xi}, \quad \text{and} \quad U_{n+1} = P_n \oplus S, \tag{7.2.3}$$

such that $S\zeta = \{\tilde{\zeta}_1, 0, 0, \ldots\}$. Finally let $R_{n+1} = U_{n+1} R_n$. Hence,

$$R_{n+1} = U_{n+1} R_n = \begin{pmatrix} \tilde{R}_n & B_n \\ SC_n & SN_n \end{pmatrix} = \begin{pmatrix} \tilde{R}_{n+1} & B_{n+1} \\ C_{n+1} & N_{n+1} \end{pmatrix}, \tag{7.2.4}$$

where the last matrix is understood to be with respect to the decomposition $\mathcal{H} = P_{n+1}\mathcal{H} \oplus P_{n+1}^\perp \mathcal{H}$. Note that, by the choice of S, it is true that $\tilde{R}_{n+1}$ is upper triangular and $C_{n+1} e_j = 0$ for $j \leq n$. How to choose the initial U_1 follows from similar reasoning. Thus, we have constructed the sequence $\{R_n\}$, and we can now turn the attention to finding a candidate R such that

$$R = \lim_{n\to\infty} R_n,$$

in some apropriate topology on $\mathcal{B}(\mathcal{H})$, where R is upper triangular. Note that $\|R_n\|$ is uniformly bounded since U_j is unitary. And since a closed ball in $\mathcal{B}(\mathcal{H})$ is weakly sequentially compact, there is an $R \in \mathcal{B}(\mathcal{H})$ and a subsequence $\{R_{n_k}\}$ such that

$$R_{n_k} \xrightarrow{\text{WOT}} R, \qquad k \to \infty.$$

But by (7.2.4) it is clear that for any integer j we have $P_j R_n P_j = P_j R_m P_j$ for sufficiently large n and m. Hence

$$\underset{n\to\infty}{\text{WOT-lim}}\ R_n = R.$$

Now, by (7.2.4) R is upper triangular with respect to $\{e_j\}$ and also $Re_j = R_n e_j$ for large n, thus

$$\operatorname*{SOT-lim}_{n\to\infty} R_n = R.$$

Hence, we have constructed the upper triangular infinite matrix R, and we now return to the task of forming Q. Let

$$V_n = U_1 \cdots U_n$$

By similar reasoning, using the previous compactness argument (since V_n is bounded) and the fact that, by (7.2.3), for any integer j we have $V_n e_j = V_m e_j$ for sufficiently large m and n, it follows that there is a $V \in \mathcal{B}(\mathcal{H})$ such that

$$V_n \xrightarrow{\text{SOT}} V, \qquad n \to \infty.$$

And, being a strong limit of unitary operators; V is an isometry. Let $Q = V$. Then, $A = QR$ since $A = V_n R_n$ and multiplication is jointly strongly continuous on bounded sets. We have just proved the following theorem.

Theorem 7.2.2. *Let A be a bounded operator on a separable Hilbert space $\mathcal{H}$ and let $\{e_j\}$ be an orthonormal basis for $\mathcal{H}$. Then there exist an isometry Q such that $A = QR$ where R is upper triangular with respect to $\{e_j\}$. Moreover*

$$Q = \operatorname*{SOT-lim}_{n\to\infty} V_n$$

where $V_n = U_1 \cdots U_n$ and U_j is a Householder transformation.

7.3 The QR algorithm

Let $A \in \mathcal{B}(\mathcal{H})$ be invertible and let $\{e_j\}$ be an orthonormal basis for $\mathcal{H}$. By Theorem 7.2.2 we have $A = QR$, where Q is unitary and R is upper triangular with respect to $\{e_j\}$. Consider the following construction of unitary operators $\{\hat{Q}_k\}$ and upper triangular (w.r.t. $\{e_j\}$) operators $\{\hat{R}_k\}$. Let $A = Q_1 R_1$ be a QR decomposition of A and define $A_2 = R_1 Q_1$. Then QR factorize $A_2 = Q_2 R_2$ and define $A_3 = R_2 Q_2$. The recursive procedure becomes

$$A_{m-1} = Q_m R_m, \quad A_m = R_m Q_m. \tag{7.3.1}$$

Now define

$$\hat{Q}_m = Q_1 Q_2 \dots Q_m, \quad \hat{R}_m = R_m R_{m-1} \dots R_1. \tag{7.3.2}$$

The recursive procedure (7.3.1) is known as the QR algorithm.

Definition 7.3.1. *Let $A \in \mathcal{B}(\mathcal{H})$ be invertible and let $\{e_j\}$ be an orthonormal basis for $\mathcal{H}$. The sequences $\{\hat{Q}_j\}$ and $\{\hat{R}_j\}$ constructed as in (7.3.1) and (7.3.2) will be called a Q-sequence and an R-sequence of A with respect to $\{e_j\}$.*

The following observation will be useful in the later developments. From the construction in (7.3.1) and (7.3.2) we get

$$A = Q_1 R_1 = \hat{Q}_1 \hat{R}_1,$$

$$A^2 = Q_1R_1Q_1R_1 = Q_1Q_2R_2R_1 = \hat{Q}_2\hat{R}_2,$$
$$A^3 = Q_1R_1Q_1R_1Q_1R_1 = Q_1Q_2R_2Q_2R_2R_1 = Q_1Q_2Q_3R_3R_2R_1 = \hat{Q}_3\hat{R}_3.$$

An easy induction gives us that

$$A^m = \hat{Q}_m\hat{R}_m.$$

Note that $\hat{R}_m$ must be upper triangular with respect to $\{e_j\}$ since R_j, $j \leq m$ is upper triangular with respect to $\{e_j\}$. Also, by invertibility of A it is true that $\langle Re_i, e_i\rangle \neq 0$. From this it follows immediately that

$$\text{span}\{A^m e_j\}_{j=1}^N = sp\{\hat{Q}_m e_j\}_{j=1}^N, \quad N \in \mathbb{N}. \tag{7.3.3}$$

In finite dimensions we have the following theorem:

Theorem 7.3.2. *Let $A \in \mathbb{C}^{N\times N}$ be a normal matrix with eigenvalues satisfying $|\lambda_1| > \ldots > |\lambda_N|$. Let $\{\hat{Q}_m\}$ be a Q-sequence of unitary operators. Then $\hat{Q}_m A\hat{Q}_m^* \to D$, as $m \to \infty$, where D is diagonal.*

We will show that an analogue of this theorem is true in infinite dimensions. But before we do that let us recall some basic Banach space geometry. We follow the notation in (Kat95). Let $E \subset \mathcal{B}$ and $F \subset \mathcal{B}$ be closed subspaces of a Banach space $\mathcal{B}$. Define

$$\delta(E,F) = \sup_{\substack{x\in E\\ \|x\|=1}} \inf_{y\in F} \|x-y\|$$

and

$$\hat{\delta}(E,F) = \max[\delta(E,F), \delta(F,E)].$$

Given subspaces E and $\{E_k\}$ such that $\hat{\delta}(E_k, E) \to 0$ as $k \to \infty$ we will sometimes use the notation

$$E_k \xrightarrow{\hat{\delta}} E, \qquad k \to \infty.$$

The reason why we introduce such convergence is the following. If we should prove an infinite dimensional analogue of Theorem 7.3.2, we obviously have to abandon any determinant based theory as in (Par65). Now, the ideas in (PP73) are much better suited for use in infinite dimensions as that theory is based on convergence of subspaces, a concept that is certainly not exclusive to finite dimensional spaces. Thus, having decided on using the ideas in (PP73) as an inspiration, the best way of approaching the Infinite QR algorithm is probably to think of it as an advanced power method. This is emphasized in the following theorem.

Theorem 7.3.3. *Let $A \in \mathcal{B}(\mathcal{H})$ be an invertible normal operator. Suppose that $\sigma(A) = \omega \cup \Omega$ is a disjoint union such that $\omega = \{\lambda_i\}_{i=1}^N$ and the λ_is are isolated eigenvalues of finite multiplicity satisfying $|\lambda_1| > \ldots > |\lambda_N|$. Suppose further that $\sup\{|\gamma| : \gamma \in \Omega\} < |\lambda_N|$. Let $\{\xi_i\}_{i=1}^M$ be a collection of linearly independent vectors in $\mathcal{H}$ such that $\{\chi_\omega(A)\xi_i\}_{i=1}^M$ are linearly independent. The following observations are true.*

(i) There exists an M-dimensional subspace $B \subset \text{ran}\chi_\omega(A)$ such that

$$\text{span}\{A^k\xi_i\}_{i=1}^M \xrightarrow{\hat{\delta}} B, \quad k \to \infty.$$

(ii) If

$$\operatorname{span}\{A^k\xi_i\}_{i=1}^{M-1} \xrightarrow{\hat{\delta}} D \subset \mathcal{H}, \quad k \to \infty,$$

where D is an $(M-1)$-dimensional subspace, then

$$\operatorname{span}\{A^k\xi_i\}_{i=1}^{M} \xrightarrow{\hat{\delta}} D \oplus \operatorname{span}\{\xi\}, \quad k \to \infty,$$

where $\xi \in \operatorname{ran}\chi_\omega(A)$ is an eigenvector of A.

Note that the previous theorem gives us some control on the behavior of $\operatorname{span}\{A^k\xi_i\}_{i=1}^{M}$ as $k \to \infty$, where ξ_i is as in Theorem 7.3.3. It is essentially that result combined with (7.3.3) (and some more analysis) that lead to the following theorem. The theory and proofs of both Theorem 7.3.3 and Theorem 7.3.4 can be found in Chapter 2.

Theorem 7.3.4. *Let $A \in \mathcal{B}(\mathcal{H})$ be an invertible normal operator and let $\{e_j\}$ be an orthonormal basis for $\mathcal{H}$. Let $\{Q_k\}$ and $\{R_k\}$ be a Q- and R-sequences of A with respect to $\{e_j\}$. Suppose also that $\sigma(A) = \omega \cup \Omega$ such that $\omega \cap \Omega = \emptyset$ and $\omega = \{\lambda_i\}_{i=1}^{N}$, where the λ_is are isolated eigenvalues with finite multiplicity satisfying $|\lambda_1| > \ldots > |\lambda_N|$. Suppose further that $\sup\{|\theta| : \theta \in \Omega\} < |\lambda_N|$. Then there is a subset $\{\hat{e}_j\}_{j=1}^{M} \subset \{e_j\}$ such that $\operatorname{span}\{Q_k\hat{e}_j\} \to \operatorname{span}\{\hat{q}_j\}$ where $\hat{q}_j$ is an eigenvector of A and $M = \dim(\operatorname{ran}\chi_\omega(A))$. Moreover, $\operatorname{span}\{\hat{q}_j\}_{j=1}^{M} = \operatorname{ran}\chi_\omega(A)$. Also, if $e_j \notin \{\hat{e}_j\}_{j=1}^{M}$, then $\chi_\omega(A)Q_ke_j \to 0$.*

Remark 7.3.5. The previous theorem is a little pessimistic regarding convergence properties for the Infinite QR algorithm. The fact that one can only find a subset $\{\hat{e}_j\}_{j=1}^{M} \subset \{e_j\}$ seems a little worrying since one immediately thinks that there may be examples where $\hat{e}_M = e_K$ and K is extremely large. We have to admit that such examples may exists, but they are rarely encountered in practice. In fact as we will see in Section 7.5.1 that the Infinite QR algorithm performs surprisingly well. Also, the rather strict (and somewhat unpleasant) assumption of normality required in Theorem 7.3.4 seem not to have any effect in practice. And from numerical examples (in Section 7.5.1) it seems that the Infinite QR algorithm works very well on non-normal problems, which are known to be numerically difficult (TE05). In fact that is where the algorithm really has it strength. We thus conclude that there is much more to be investigated form a theoretical point of view to justify the numerical results.

7.4 Implementing the Infinite QR algorithm

The previous sections have given a theoretical justification for why the infinite QR algorithm may work, but we are faced with the possibly unpleasant problem, namely, how to compute with infinite data structures on a computer. Fortunately there is a way to overcome such a problem. The key is to impose some structural requirements on the infinite matrix.

Definition 7.4.1. *Let T be an infinite matrix acting boundedly on $l^2(\mathbb{N})$ with basis $\{e_j\}$. We say that T has k subdiagonals if $\langle Te_j, e_i\rangle = 0$ when $i > j + k$.*

Theorem 7.4.2. *Let $A \in \mathcal{B}(l^2(\mathbb{N}))$ have k subdiagonals and let A_n be the n-th element in the QR iteration, e.g. $A_n = Q_n \cdots Q_1 A Q_1^* \cdots Q_n^*$, where*

$$Q_j = \underset{l\to\infty}{SOT\text{-}lim}\ U_l^j \cdots U_1^j$$

and U_l^j is a Householder transformation defined as in (7.2.3). Then A_n has k subdiagonals.

Proof. Straightforward. □

Theorem 7.4.3. *Let $A \in \mathcal{B}(l^2(\mathbb{N}))$ have k subdiagonals and let A_n be the n-th element in the QR iteration, i.e. $A_n = Q_n \cdots Q_1 A Q_1^* \cdots Q_n^*$, where*

$$Q_j = \underset{l\to\infty}{SOT\text{-}lim}\ U_l^j \cdots U_1^j$$

and U_l^j is a Householder transformation defined as in (7.2.3). Let P_m be the projection onto $\operatorname{span}\{e_1, \ldots, e_m\}$. *Then*

$$\begin{aligned} P_m A_n P_m =& P_m U_m^n \cdots U_1^n U_{k+m}^{n-1} \cdots U_1^{n-1} \cdots U_{(n-2)k+m}^2 \cdots U_1^2 U_{(n-1)k+m}^1 \cdots U_1^1 A \\ &\times U_1^1 \cdots U_{(n-1)k+m}^1 U_1^2 \cdots U_{(n-2)k+m}^2 \cdots U_1^{n-1} \cdots U_{k+m}^{n-1} U_1^n \cdots U_m^n P_m. \end{aligned} \tag{7.4.1}$$

and

$$\begin{aligned} P_m A_n P_m = P_m U_m^n \cdots U_1^n U_{k+m}^{n-1} \cdots U_1^{n-1} \cdots U_{(n-2)k+m}^2 \cdots U_1^2 U_{(n-1)k+m}^1 \cdots U_1^1 \\ \times P_{nk+m} A P_{nk+m} U_1^1 \cdots U_{(n-1)k+m}^1 U_1^2 \cdots U_{(n-2)k+m}^2 \cdots U_1^{n-1} \cdots U_{k+m}^{n-1} U_1^n \cdots U_m^n P_m. \end{aligned} \tag{7.4.2}$$

Proof. Let $\{e_j^i\}$ be the basis of $P_i^\perp \mathcal{H}$ such that $e_j^i = e_{j+i}$. Note that, by the assumption that A has k subdiagonals, it follows that each U_l^j is of the form

$$U_l^j = I_{l,1} \oplus \left(I_{l,2} - \frac{2}{\|\xi\|^2} \xi \otimes \bar{\xi} \right) \qquad \xi \in P_{l-1}^\perp \mathcal{H},$$

where $\langle \xi, e_j^l \rangle = 0$ for $j > k$ and $I_{l,1}$ is the identity on $P_{l-1}\mathcal{H}$ and $I_{l,2}$ is the identity on $P_{l-1}^\perp \mathcal{H}$. For $l = 1$ then $I_{0,1} = P_0 = 0$. This observation yields the following, namely,

$$\begin{aligned} P_r U_l^j &= U_l^j P_r, \quad r \geq l + k, \\ P_r U_m^j &= P_r, \quad m > r. \end{aligned} \tag{7.4.3}$$

First note that

$$P_m A_n P_m = P_m P_{(n-1)k+m} \cdots P_{k+m} P_m Q_n \cdots Q_1 A Q_1^* \cdots Q_n^* P_m P_{k+m} \cdots P_{(n-1)k+m} P_m$$

and, since multiplication is strongly continuous on bounded sets and by (7.4.3) then

$$\begin{aligned} P_m A_n P_m &= \text{SOT-lim}_l\, P_m P_{(n-1)k+m} \cdots P_{k+m} P_m U_l^n \cdots U_1^n Q_{n-1} \cdots Q_1 A \\ &\qquad \times Q_1^* \cdots Q_{n-1}^* (\text{SOT-lim}_l\, U_1^n \cdots U_l^n P_{(n-1)k+m} \cdots P_{k+m} P_m) \\ &= P_m U_m^n \cdots U_1^n P_{(n-1)k+m} \cdots P_{k+m} Q_{n-1} \cdots Q_1 A \\ &\qquad \times Q_1^* \cdots Q_{n-1}^* P_{k+m} \cdots P_{(n-1)k+m} U_1^n \cdots U_m^n P_m. \end{aligned} \tag{7.4.4}$$

This type or reasoning may, of course, be repeated and, thus, by using (7.4.4), we obtain

$$\begin{aligned}
&P_m A_n P_m \\
&= P_m U_m^n \cdots U_1^n (\text{SOT-lim}_l P_{(n-1)k+m} \cdots P_{k+m} U_l^{n-1} \cdots U_1^{n-1}) Q_{n-2} \cdots Q_1 A \\
&\qquad \times Q_1^* \cdots Q_{n-2}^* (\text{SOT-lim}_l U_1^{n-1} \cdots U_l^{n-1} P_{k+m} \cdots P_{(n-1)k+m}) U_1^n \cdots U_m^n P_m \\
&= P_m U_m^n \cdots U_1^n P_{k+m} U_{k+m}^{n-1} \cdots U_1^{n-1} P_{(n-1)k+m} \cdots P_{2k+m} Q_{n-3} \cdots Q_1 A \\
&\qquad \times Q_1^* \cdots Q_{n-3}^* P_{2k+m} \cdots P_{(n-1)k+m} U_1^{n-1} \cdots U_{k+m}^{n-1} P_{k+m} U_1^n \cdots U_m^n P_m \\
&= P_m U_m^n \cdots U_1^n U_{k+m}^{n-1} \cdots U_1^{n-1} P_{(n-1)k+m} \cdots P_{2k+m} Q_{n-3} \cdots Q_1 A \\
&\qquad \times Q_1^* \cdots Q_{n-3}^* P_{2k+m} \cdots P_{(n-1)k+m} U_1^{n-1} \cdots U_{k+m}^{n-1} U_1^n \cdots U_m^n P_m.
\end{aligned}$$

Repeating the same ideas $n-3$ more times eventually leads to

$$\begin{aligned}
&P_m U_m^n \cdots U_1^n U_{k+m}^{n-1} \cdots U_1^{n-1} P_{(n-1)k+m} \cdots P_{2k+m} Q_{n-3} \cdots Q_1 A \\
&\qquad \times Q_1^* \cdots Q_{n-3}^* P_{2k+m} \cdots P_{(n-1)k+m} U_1^{n-1} \cdots U_{k+m}^{n-1} U_1^n \cdots U_m^n P_m \\
&= P_m U_m^n \cdots U_1^n U_{k+m}^{n-1} \cdots U_{k+m}^{n-1} \cdots U_{(n-2)k+m}^2 \cdots U_1^2 U_{(n-1)k+m}^1 \cdots U_1^1 A \\
&\qquad \times U_1^1 \cdots U_{(n-1)k+m}^1 U_1^2 \cdots U_{(n-2)k+m}^2 \cdots U_1^{n-1} \cdots U_{k+m}^{n-1} U_1^n \cdots U_m^n P_m,
\end{aligned}$$

and this yields (7.4.1). Now, (7.4.2) follows from (7.4.1) and (7.4.3). □

Thus, if the infinite matrix A has k subdiagonals, this result allows us to actually implement the infinite QR algorithm because each U_l^j only affects finitely many columns or rows of A if multiplied either on the left or the right. In computer science it is often referred to as "Lazy evaluation" when one computes with infinite data structures, but defers the use of the information until needed, and hence solves the problem of infinite storage. The author is indebted to Nick Trefethen for pointing out this connection to the Infinite QR algorithm. A simple implementation is displayed in Algorithm 7.4.1.

Algorithm 7.4.1.

```
% The Infinite_QR(A,n,k,m) takes a section P_{nk+m}AP_{nk+m}
% of an infinite matrix A with k subdiagonals, performs n iterations
% of the infinite dimensional QR algorithm and returns
% J = P_mQ_nAQ*_nP_m.

function J = Infinite_QR(A,n,k,m)
d = size(A,2);
 for j=1:n
   A = Inf_QR(A,d-j*k,k);   % The output in each loop is actually
 end                        % U_(d-j*k)...U_1A_(j-1)U_1...U_(d-j*k)
J = A(1:m,1:m);             % if A_j is the j-th term in the QR iteration.
```

Algorithm 7.4.2.

```
% Inf_QR(A,n,k) takes a matrix A with k subdiagonals and performs
% multiplication by n Householder transformation from the left and
% right, i.e. B = U_n...U_1AU_1...U_n.

function B = Inf_QR(A,n,k)
B = A; d = size(A,1);
 for j = 1:n
    u = House(A(j:j+k,j));
```

```
    A(j:j+k,j:d) = A(j:j+k,j:d) - 2*u*(u'*A(j:j+k,j:d));
    B(j:j+k,1:d) = B(j:j+k,1:d) - 2*u*(u'*B(j:j+k,1:d));
    B(1:d,j:j+k) = B(1:d,j:j+k) - 2*(B(1:d,j:j+k)*u)*u';
 end
```

Algorithm 7.4.3.

```
% House(x) takes a vector x and creates a vector u
% such that (I + u*u')x = ce_1 where c is some complex
% number (depending on x) and e_1 = [1,0...].

function u = House(x)
v = x;
if v(1) == 0
   v(1) = v(1) + norm(v);          %This is the classical way
else                               %of creating Householder reflections
v(1) = x(1) + sign(x(1))*norm(x);  %as in finite dimensions.
end
u = v/norm(v);
```

7.5 Testing the Infinite QR algorithm

In this section we will see that the QR algorithm performs much better than what we have been able to prove in Theorem 7.3.4. In fact, the normality required in Theorem 7.3.4 is not needed at all to get satisfactory results. This section is meant to visualize some of the rather unexpected behavior of the Infinite QR algorithm and to give numerical support to suggest results that we have not proved rigorously. Throughout this section Q_n will denote the n-th "Q" matrix in the iteration i.e. if A is the initial matrix then $A_n = Q_n A Q_n^*$ represents the n-th element in the iteration.

To illustrate the infinite QR algorithm we have tested it on an infinite matrix T of the form $T = I + C$, where C is compact. This should give plenty of isolated eigenvalues and is therefore a good testing candidate. More examples will follow in the next sections. Now T is of the following form

$$T = \begin{pmatrix} 1 & t_{12} & 0 & 0 & 0 & 0 & \dots \\ 0 & 1 & t_{23} & 0 & 0 & 0 & \dots \\ t_{31} & 0 & 1 & t_{34} & 0 & 0 & \dots \\ 0 & t_{42} & 0 & 1 & t_{45} & 0 & \dots \\ 0 & 0 & t_{53} & 0 & 1 & t_{56} & \dots \\ 0 & 0 & 0 & t_{64} & 0 & 1 & \dots \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \ddots \end{pmatrix},$$

where

$$t_{2j-1,2j} = i - (1+2i)/j,\ t_{2j,2j+1} = (13-i)/\sqrt{j},\ t_{2j+1,2j-1} = -1,\ t_{2j+2,2j} = 1/j.$$

The plots shown in Figure 7.4 shows the elements of

$$P_m Q_n T Q_n^* \lceil_{P_m \mathcal{H}}, \qquad m = 30, n = 1500,$$

that are larger than 10^{-15} and 10^{-5}.

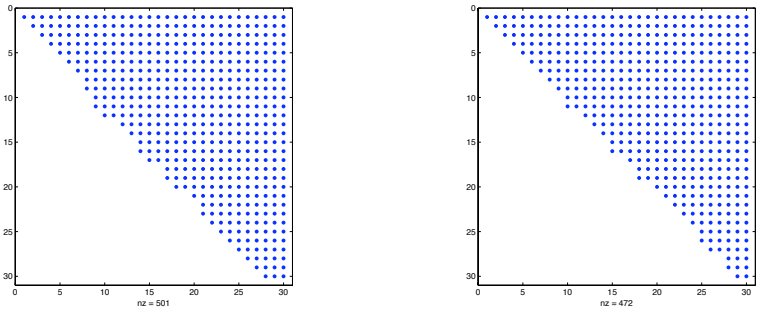


Figure 7.4: The figure shows the elements of $P_m Q_n T Q_n^* \lceil_{P_m\mathcal{H}}$ ($m = 30$, $n = 1500$) that are greater than 10^{-15} and 10^{-5}, respectively.

7.5.1 The Magical Result

We will in this section show some rather spectacular and so far mathematically unexplained features of the Infinite QR algorithm. So far the theoretical results cover only normal operators, but in practice the Infinite QR algorithm works very well for non-normal problems. Now, if an infinite matrix T has m eigenvalues with multiplicity one then, under some extra assumptions, one would expect that the eigenvalues will appear on the diagonal of

$$P_m Q_k T Q_k^* \lceil_{P_m\mathcal{H}}$$

as $k \to \infty$ i.e. we would expect that

$$\sigma(P_m Q_k T Q_k^* \lceil_{P_m\mathcal{H}}) \longrightarrow \sigma_d(T), \qquad k \to \infty.$$

But what happens with

$$\sigma(P_n Q_k T Q_k^* \lceil_{P_n\mathcal{H}})$$

when $n > m$ as k becomes large? That is the theme of the next examples. Let us go back to one of the examples in Section 7.1, where we computed the spectrum of the finite section $P_m A \lceil_{P_m\mathcal{H}}$ of the infinite matrix A, where

$$A = \begin{pmatrix} 2.5+0.5i & 0 & 0 & 0 & 0 & 0 & 0 & \dots \\ 1 & 3-0.5i & 0 & 0 & 0 & 0 & 0 & \dots \\ 0 & 1 & 1.7 & 0.05 & 0 & 0 & 0 & \dots \\ 0 & 0 & 0.05 & a_4 & 0 & 0 & 0 & \dots \\ 0 & 0 & 0 & 0 & a_5 & 0 & 0 & \dots \\ 0 & 0 & 0 & 0 & 1 & a_6 & 0 & \dots \\ 0 & 0 & 0 & 0 & 0 & 1 & a_7 & \dots \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \ddots \end{pmatrix}.$$

Recall that the problem with the finite section method in that case was that we do not know anything about $W_e(A)$ and hence could not deduce if the eigenvalues produced by the finite section method were correct. Also, recall the slightly misleading circle that appeared that had nothing to do with the boundary of the essential spectrum (see Figure

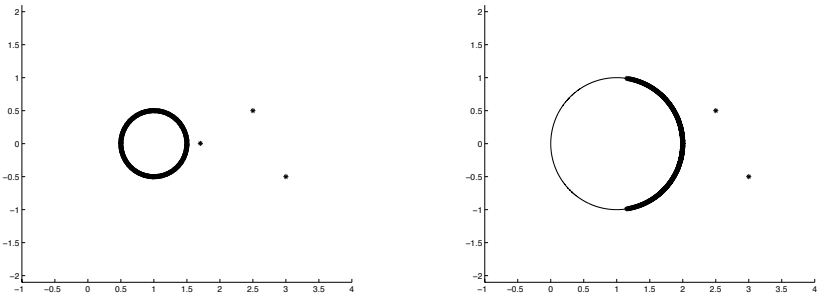


Figure 7.5: The first figure shows $\sigma(P_m A\lceil_{P_m\mathcal{H}})$ and the second figure shows $\sigma(P_m Q_n A Q_n^*\lceil_{P_m\mathcal{H}})$ (the fat line, the thin line is just to visualize the circle) for $m = 500$ and $n = 1500$.

7.5). We have plotted

$$\sigma(P_m A\lceil_{P_m\mathcal{H}}), \qquad \sigma(P_m Q_n A Q_n^*\lceil_{P_m\mathcal{H}}), \quad m = 500,\, n = 1500,$$

in Figure 7.5 to see the difference. Note that $\sigma(P_m Q_n A Q_n^*\lceil_{P_m\mathcal{H}})$ actually reveals part of the correct boundary of the essential spectrum. We have also tested the Infinite QR algorithm on the perturbed version of A from Section 7.1, namely $\tilde{A}$, where

$$\tilde{A} = \begin{pmatrix} 2.5+0.5i & 0 & 0 & 0 & 0 & 0 & 0 & \dots \\ 1 & 3-0.5i & 0 & 0 & 0 & 0 & 0 & \dots \\ 0 & 1 & 1.7 & 0.05 & 0 & 0 & 0 & \dots \\ 0 & 0 & 0.05 & a_4 & 0 & 0 & 0 & \dots \\ 0 & 0 & 0 & 5\times 10^5 & a_5 & 0 & 0 & \dots \\ 0 & 0 & 0 & 0 & 1 & a_6 & 0 & \dots \\ 0 & 0 & 0 & 0 & 0 & 1 & a_7 & \dots \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \ddots \end{pmatrix}.$$

Now, $\sigma_{\text{ess}}(A) = \sigma_{\text{ess}}(\tilde{A})$, but due to the fact that $\tilde{A}$ is highly non-normal it is impossible (at least with the ϵ_{mach} in MATLAB) to use the pseudospectral techniques in Chapter 6 to compute the spectrum of $\tilde{A}$. However, the infinite QR algorithm seems to be able to pick up parts of the boundary of the essential spectrum, as visualized in the first picture of Figure 7.6. Also, as it seems that the Infinite QR algorithm is able to pick up the extreme part of the spectrum, we were tempted to see what would happen when the Infinite QR algorithm is applied to a shift of $\tilde{A}$? In particular, one would expect that

$$\sigma(P_m Q_n(\tilde{A} - 2I)Q_n^*\lceil_{P_m\mathcal{H}})$$

should contain the left part of the circle for appropriately chosen m and n. This is visualized in Figure 7.6.

Finally, We have tested the Infinite QR algorithm on the infinite matrices from Section

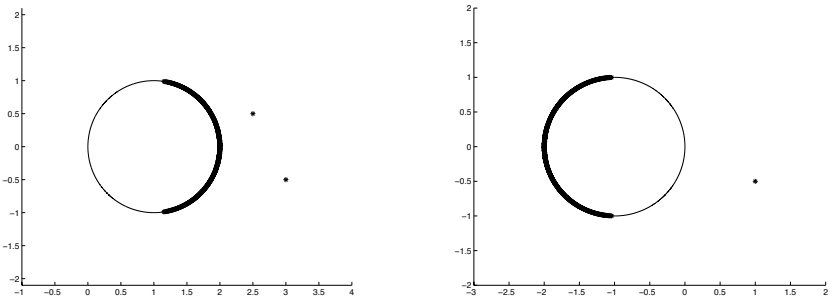


Figure 7.6: The first figure shows $\sigma(P_m Q_n \tilde{A} Q_n^* \lceil_{P_m\mathcal{H}})$ (the fat dots) and the second figure shows $\sigma(P_m Q_n (\tilde{A} - 2I) Q_n^* \lceil_{P_m\mathcal{H}})$, for $n = 1500$, $m = 500$.

7.1, namely

$$T = \begin{pmatrix} 0 & 0 & 0 & 0 & 0 & 0 & \dots \\ 1 & 0 & a_{23} & a_{24} & 0 & 0 & \dots \\ 0 & 1 & 0 & 0 & 0 & 0 & \dots \\ 0 & 0 & 1 & 0 & a_{45} & a_{46} & \dots \\ 0 & 0 & 0 & 1 & 0 & 0 & \dots \\ 0 & 0 & 0 & 0 & 1 & 0 & \dots \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \ddots \end{pmatrix}, \qquad \begin{cases} a_{2j,2j+2} = -i & j \text{ is prime} \\ a_{2j,2j+2} = 0 & \text{otherwise,} \end{cases} \qquad a_{2j,j+1} = 1,$$

and

$$\tilde{T} = \begin{pmatrix} 2.5 + 0.5i & 0 & 0 & 0 & 0 & 0 & 0 & \dots \\ 1 & 3 - 0.5i & 0 & 0 & 0 & 0 & 0 & \dots \\ 0 & 1 & 1.7 & 0.05 & 0 & 0 & 0 & \dots \\ 0 & 0 & 0.05 & 1 & 0 & 0 & 0 & \dots \\ 0 & 0 & 0 & 1 & 1 & a_{56} & 0 & \dots \\ 0 & 0 & 0 & 0 & 1 & 1 & 0 & \dots \\ 0 & 0 & 0 & 0 & 0 & 1 & 1 & \dots \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \ddots \end{pmatrix}, \qquad a_{2j-1,2j} = i, \quad j \geq 3.$$

We have used techniques from Chapter 6 to compute the spectra of A and $\tilde{A}$ and also run the Infinite QR algorithm with 1500 iteration. In both cases we see the same phenomenon, namely, that if one takes a finite section after running the Infinite QR algorithm, then a part of the boundary of the essential spectrum also occurs. This is visualized in Figure 7.7. Note that the part of the boundary that is captured is the extreme part (meaning the points with largest modulus). It seems that after running the Infinite QR algorithm the spectral information from the largest isolated eigenvalues and the largest approximate point spectrum gets "squeeced up". Although we can not explain this phenomenon, it is not completely counter intuitive, as this is what normally happens in finite dimensions.

Remark 7.5.1. As the numerical examples suggest, the Infinite QR algorithm seems to be able to detect extreme parts of the boundary of the essential spectrum. This immediately

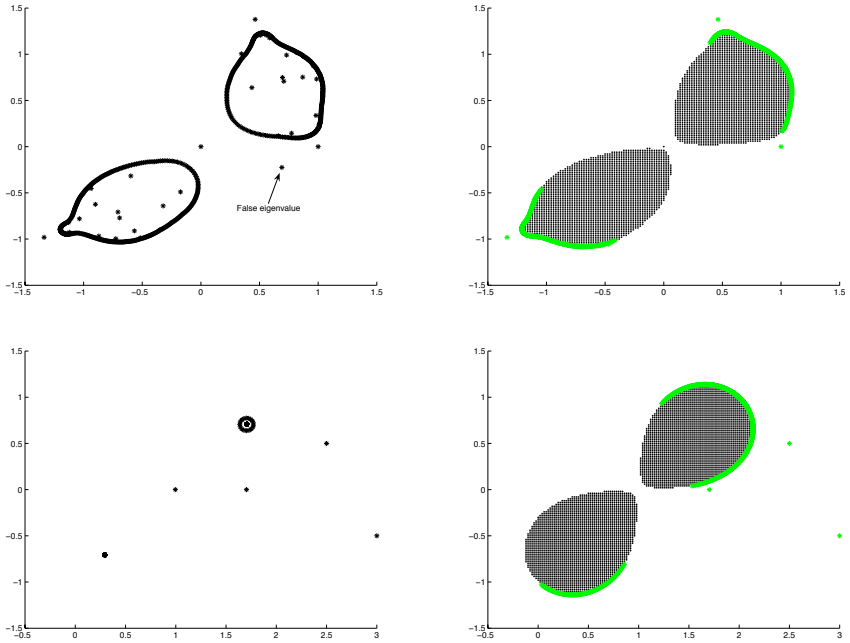


Figure 7.7: The left figures show $\sigma(P_m T\lceil_{P_m\mathcal{H}})(top)$ and $\sigma(P_m\tilde{T}\lceil_{P_m\mathcal{H}})$ (bottom) for $m = 500$. The right figures (the dark plot) shows $\sigma(T)$ (top), $\sigma(\tilde{T})$ (bottom) and (the light plot) $\sigma(P_m Q_n \tilde{T} Q_n^*\lceil_{P_m\mathcal{H}})$ (top), $\sigma(P_m Q_n \tilde{T} Q_n^*\lceil_{P_m\mathcal{H}})$ (bottom) for $n = 900$ and $m = 600$.

leads to the question of shifting as done in the computation visualized in Figure 7.6. This approach is under current investigation and is also the theme of Chapter 8.

Chapter 8

The Hessenberg Reduction

After having established the Infinite QR algorithm, the first question that comes to mind is how do we improve the algorithm, and, in particular, how can we speed it up? In finite dimensions there are two approaches that are very common, namely, Hessenberg reduction and shifting techniques. In fact the combination of these are actually the core in the MATLAB command `eig`. We will in this chapter explore some of these ideas and see how it may be possible to generalize these finite-dimensional techniques to infinite dimensions. As we will see, the Hessenberg reduction is a great success and cuts the computational costs dramatically. Unfortunately, shifting strategies may not give us any advantage regarding speed. There is a natural reason why this approach cannot work in infinite dimension, and this will be explained later. However, shifting strategies represent a great improvement on the QR algorithm, but for different reasons than speed. In particular, using shifts we can actually compute a larger part of the spectrum. There is no analogy to this phenomenon in finite dimensions, and it must be admitted that the quite satisfactory results using shifts in the infinite-dimensional case is quite surprising and not yet mathematically justified.

8.1 Constructing the Hessenberg Reduction

In finite-dimensional matrix analysis, a matrix is said to be upper Hessenberg if it has only one nonzero leading subdiagonal. In infinite dimensions we adopt the the same idea and hence consider the following definition.

Definition 8.1.1. *Let $A \in \mathcal{B}(l^2(\mathbb{N}))$. Suppose that A satisfies*

$$\langle Ae_j, e_i \rangle = 0, \qquad i > j + 1,$$

then A is said to be in upper Hessenberg form. If $A \in \mathcal{B}(l^2(\mathbb{N}))$ and A^* is in upper Hessenberg form, then A is said to be in lower Hessenberg form.

Throughout the chapter we will use a slightly incorrect terminology and just say that a matrix is upper or lower Hessenberg.

When faced with the finite dimensional eigenvalue problem, one normally reduces the matrix to upper Hessenberg form and then applies the QR algorithm. This is because the vast amount of zeros in the Hessenberg form simplifies the QR algorithm and speeds it

up. The desire to speed up the Infinite QR algorithm is the motivation for introducing the Hessenberg reduction in infinite dimensions.

In finite dimensions, a matrix is always unitary equivalent to another matrix that is upper Hessenberg, and the procedure to find this matrix is often referred to as the Hessenberg reduction. It turns out that the finite dimensional technique can be modified slightly to fit into the infinite dimensional framework. In fact, the Hessenberg reduction technique in infinite dimensions is quite similar to the finite dimensional approach. One needs to multiply the infinite matrix by unitary operators on each side to introduce zeros under the first subdiagonal. As usual these unitary operators are, as in the finite dimensional setting, Householder transformations, which are chosen in the usual way. The procedure can be visualized as follows.

$$U_1 \begin{bmatrix} \times & \times & \times & \times & \times & \times & \times & \dots \\ \times & \times & \times & \times & \times & \times & \times & \dots \\ \times & \times & \times & \times & \times & \times & \times & \dots \\ 0 & \times & \times & \times & \times & \times & \times & \dots \\ 0 & 0 & \times & \times & \times & \times & \times & \dots \\ 0 & 0 & 0 & \times & \times & \times & \times & \dots \\ 0 & 0 & 0 & 0 & \times & \times & \times & \dots \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \ddots \end{bmatrix} U_1 = \begin{bmatrix} \times & \times & \times & \times & \times & \times & \times & \dots \\ \times & \times & \times & \times & \times & \times & \times & \dots \\ 0 & \times & \times & \times & \times & \times & \times & \dots \\ 0 & \times & \times & \times & \times & \times & \times & \dots \\ 0 & \times & \times & \times & \times & \times & \times & \dots \\ 0 & 0 & 0 & \times & \times & \times & \times & \dots \\ 0 & 0 & 0 & 0 & \times & \times & \times & \dots \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \ddots \end{bmatrix}$$

$$U_2 \begin{bmatrix} \times & \times & \times & \times & \times & \times & \times & \dots \\ \times & \times & \times & \times & \times & \times & \times & \dots \\ 0 & \times & \times & \times & \times & \times & \times & \dots \\ 0 & \times & \times & \times & \times & \times & \times & \dots \\ 0 & \times & \times & \times & \times & \times & \times & \dots \\ 0 & 0 & 0 & \times & \times & \times & \times & \dots \\ 0 & 0 & 0 & 0 & \times & \times & \times & \dots \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \ddots \end{bmatrix} U_2 = \begin{bmatrix} \times & \times & \times & \times & \times & \times & \times & \dots \\ \times & \times & \times & \times & \times & \times & \times & \dots \\ 0 & \times & \times & \times & \times & \times & \times & \dots \\ 0 & 0 & \times & \times & \times & \times & \times & \dots \\ 0 & 0 & \times & \times & \times & \times & \times & \dots \\ 0 & 0 & \times & \times & \times & \times & \times & \dots \\ 0 & 0 & \times & \times & \times & \times & \times & \dots \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \ddots \end{bmatrix}$$

$$U_3 \begin{bmatrix} \times & \times & \times & \times & \times & \times & \times & \dots \\ \times & \times & \times & \times & \times & \times & \times & \dots \\ 0 & \times & \times & \times & \times & \times & \times & \dots \\ 0 & 0 & \times & \times & \times & \times & \times & \dots \\ 0 & 0 & \times & \times & \times & \times & \times & \dots \\ 0 & 0 & \times & \times & \times & \times & \times & \dots \\ 0 & 0 & \times & \times & \times & \times & \times & \dots \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \ddots \end{bmatrix} U_3 = \begin{bmatrix} \times & \times & \times & \times & \times & \times & \times & \dots \\ \times & \times & \times & \times & \times & \times & \times & \dots \\ 0 & \times & \times & \times & \times & \times & \times & \dots \\ 0 & 0 & \times & \times & \times & \times & \times & \dots \\ 0 & 0 & 0 & \times & \times & \times & \times & \dots \\ 0 & 0 & 0 & \times & \times & \times & \times & \dots \\ 0 & 0 & 0 & \times & \times & \times & \times & \dots \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \ddots \end{bmatrix}$$

Now, in finite dimensions (say, the dimension of the Hilbert space is n) this procedure terminates and one is left with a matrix $H = U_{n-2} \cdots U_1 A U_1 \cdots U_{n-2}$ in an upper Hessenberg form, where $A \in \mathbb{C}^{n \times n}$. In infinite dimensions we are faced with the problem that the procedure described above will not terminate. In particular, (assume now that $A \in \mathcal{B}(l^2(\mathbb{N}))$) we may construct

$$H_n = U_n \cdots U_1 A U_1 \cdots U_n, \qquad n \in \mathbb{N}, \tag{8.1.1}$$

as suggested above, but H_n will never be upper Hessenberg. Thus, the only way to overcome this problem is to interpret the construction (8.1.1) as a sequence leading to a limit, i.e. we are searching for an element $H \in \mathcal{B}(l^2(\mathbb{N}))$ which is upper Hessenberg and satisfies

$$H = \lim_{n\to\infty} H_n$$

in some appropriate topology on $\mathcal{B}(l^2(\mathbb{N}))$. In particular, we will obtain H as the strong limit of a sequence $\{V_n^* A V_n\}$ where $V_n = U_1 \cdots U_n$ is a unitary operator and U_j is a Householder transformation. The procedure is as follows: (Note that the construction is very similar to what we did in Chapter 7 regarding the construction of the QR decomposition, however, as the reader may notice, there are some fundamental differences.) Let P_n be the projection onto $\mathrm{sp}\{e_1, \ldots, e_n\}$. Suppose that we have the n elements in the sequence and that the n-th element is an operator

$$H_n = V_n^* A V_n$$

that with respect to $\mathcal{H} = P_n\mathcal{H} \oplus P_n^\perp\mathcal{H}$ (here we use $\mathcal{H} = l^2(\mathbb{N})$ to simplify notation) has the form

$$H_n = \begin{pmatrix} \tilde{H}_n & B_n \\ C_n & N_n \end{pmatrix}, \quad \tilde{H}_n = P_n H_n P_n,\ B_n = P_n H_n P_n^\perp,\ C_n = P_n^\perp H_n P_n,$$

where $N_n = P_n^\perp H_n P_n^\perp$, $\tilde{H}_n$ is upper Hessenberg and $C_n e_j = 0$ for $j < n$. Let $\zeta = C_n e_n$. Choose $\xi \in P_n^\perp\mathcal{H}$ such that the Householder reflection $S \in \mathcal{B}(P_n^\perp\mathcal{H})$ defined by

$$S = I - \frac{2}{\|\xi\|^2}\xi \otimes \bar{\xi}, \quad \text{and} \quad U_n = P_n \oplus S, \tag{8.1.2}$$

gives $S\zeta = \{\tilde{\zeta}_1, 0, 0, \ldots\}$, and let $H_{n+1} = U_n H_n U_n$. Hence,

$$H_{n+1} = U_n H_n U_n = \begin{pmatrix} \tilde{H}_n & B_n S \\ S C_n & S N_n S \end{pmatrix} = \begin{pmatrix} \tilde{H}_{n+1} & B_{n+1} \\ C_{n+1} & N_{n+1} \end{pmatrix}, \tag{8.1.3}$$

where the last matrix is understood to be with respect to the decomposition

$$\mathcal{H} = P_{n+1}\mathcal{H} \oplus P_{n+1}^\perp\mathcal{H}.$$

Note that, by the choice of S, it is true that $\tilde{H}_{n+1}$ is upper Hessenberg and $C_{n+1}e_j = 0$ for $j < n+1$. Defining $H_1 = A$ and letting $V_n = U_1 \cdots U_n$ we have completed the construction of the sequence $\{V_n^* A V_n\}$.

Note that $H_n = V_n^* A V_n$ is bounded, since V_n is unitary (since U_j is unitary). And since a closed ball in $\mathcal{B}(\mathcal{H})$ is weakly sequentially compact, there is an $H \in \mathcal{B}(\mathcal{H})$ and a subsequence $\{H_{n_k}\}$ such that

$$H_{n_k} \xrightarrow{\text{WOT}} H, \qquad k \to \infty.$$

But by (8.1.3) it is clear that for any j we have $H_n e_j = H_m e_j$ for sufficiently large m and n. It follows that

$$\underset{n\to\infty}{\text{SOT-lim}}\ H_n = H.$$

Also, by (8.1.3), H is upper Hessenberg. By similar reasoning, using the previous compactness argument (since V_n is bounded) and the fact that, by (8.1.2), $V_n e_j = V_m e_j$ for any j and m and n sufficiently large, we deduce that there exists a $V \in \mathcal{B}(\mathcal{H})$ such that

$$\operatorname*{SOT-lim}_{n\to\infty} V_n = V, \quad \operatorname*{WOT-lim}_{n\to\infty} V_n^* = V^*. \tag{8.1.4}$$

Note that in finite dimensions then V would always be unitary. This is not the case in infinite dimensions, and this is probably the most important difference when going from finite to infinite dimensions. As we will see later, V is an isometry, a concept we recall in the following definition.

Definition 8.1.2. *Let $T \in \mathcal{B}(\mathcal{H})$. If T satisfies*

$$\|T\xi\| = \|\xi\|, \qquad \xi \in \mathcal{H},$$

then T is said to be isometric (or an isometry).

In finite dimensions, an isometry is also unitary. This is not the case in infinite dimensions, however, we have the following. If $T \in \mathcal{B}(\mathcal{H})$ is an isomtery then

$$T^*T = I, \qquad TT^* = P,$$

where P is the orthogonal projection onto $\operatorname{ran}(T)$. (Note that the range of an isometry is always closed.)

Now returning to the Hessenberg reduction in infinite dimensions, let V be as in (8.1.4). Since V is the strong limit of a sequence of unitary operators, it follows that V is an isometry. We claim that

$$V^*AV = H.$$

Indeed, since multiplication is jointly continuous in the strong operator topology on bounded sets we have $AV = VH$ and since V is an isometry the assertion follows. Note also that the range of V is invariant under A. In particular we have

$$PAP = AP, \qquad P = VV^*,$$

and this is easily seen by a direct computation as follows,

$$PAP = VV^*AVV^* = VHV^* = AP.$$

The previous reasonings lead to the following theorem.

Theorem 8.1.3. *Let A be a bounded operator on a separable Hilbert space $\mathcal{H}$ and let $\{e_j\}$ be an orthonormal basis for $\mathcal{H}$. Then there exists an isometry V such that $V^*AV = H$ where H is upper Hessenberg with respect to $\{e_j\}$. Moreover $V = \operatorname*{SOT-lim}_{n\to\infty} V_n$, where $V_n = U_1 \cdots U_n$ and U_j is a Householder transformation. Also, the projection $P = VV^*$ satisfies $PAP = AP$.*

8.2 Implementing the Hessenberg Reduction

As we have pointed out before, there is a crucial difference between the result in Theorem 8.1.3 and the finite dimensional counterpart, namely, that in Theorem 8.1.3, V is not unitary but rather an isometry. This fact has dramatical consequences from a spectral theoretical perspective. Now, the fact that for $P = VV^*$ we have $PAP = AP$ means that if we compute $\sigma(H)$ we actually get $\sigma(A\lceil_{Pl^2(\mathbb{N})})$, and, of course, we may have that

$$\sigma(A) \neq \sigma(A\lceil_{Pl^2(\mathbb{N})}).$$

However, the Infinite QR algorithm is designed to get information about the boundary of the spectrum, and we always have

$$\partial\sigma(A\lceil_{Pl^2(\mathbb{N})}) \subset \partial\sigma(A). \tag{8.2.1}$$

But the fact that the inclusion in (8.2.1) may be proper is actually not going to harm us at all. The idea is the following: Recall from Theorem 7.4.3 that if $A \in \mathcal{B}(l^2(\mathbb{N}))$ and A_n is the n-th output of the Infinite QR algorithm then

$$\begin{aligned} P_m A_n P_m = {} & P_m U_m^n \cdots U_1^n \cdots U_{(n-2)k+m}^2 \cdots U_1^2 U_{(n-1)k+m}^1 \cdots U_1^1 \\ & \times P_{nk+m} A P_{nk+m} U_1^1 \cdots U_{(n-1)k+m}^1 U_1^2 \cdots U_{(n-2)k+m}^2 \cdots U_1^n \cdots U_m^n P_m. \end{aligned} \tag{8.2.2}$$

Thus, one uses only information from the section

$$P_{nk+m} A P_{nk+m}$$

of A to compute $P_m A_n P_m$. The idea is therefore to transform A into a new matrix $\widetilde{H}$ (via unitary transformations) such that $P_{nk+m}\tilde{H}P_{nk+m}$ is upper Hessenberg. Then $\widetilde{H}$ and A have the same spectrum and we may replace A with $\widetilde{H}$ in (8.2.2). Now the construction of $\widetilde{H}$ can be done by the construction suggested in the argument leading up to Theorem 8.1.3. In particular, if

$$H_j = U_j \cdots U_1 A U_1 \cdots U_j,$$

is constructed as in (8.1.1) and (8.1.3), then $P_m H_j P_m$ is upper Hessenberg when $j \geq m$. Moreover, we have the following theorem.

Theorem 8.2.1. *Let $A \in \mathcal{B}(l^2(\mathbb{N}))$ have k subdiagonals and let A_n be the n-th element in the Hessenbeg reduction, i.e.*

$$H_n = U_n \cdots U_1 A U_1 \cdots U_n,$$

where U_j is a Householder transformation defined as in (8.1.2). Let P_m be the projection onto $\operatorname{span}\{e_1, \ldots, e_m\}$. *If $m > nk + 1$ then*

$$P_m H_n P_m = U_n \cdots U_1 P_m A P_m U_1 \cdots U_n, \tag{8.2.3}$$

and

$$P_n H_n P_n = P_n U_n \cdots U_1 P_{nk+2} A P_{nk+2} U_1 \cdots U_n P_n. \tag{8.2.4}$$

Proof. For $1 \leq j \leq n$, let

$$H_j = U_j \cdots U_1 A U_1 \cdots U_j,$$

It is easy to see that H_j has $k + j(k-1)$ subdiagonals. Thus, each U_j is of the form

$$U_j = I_{j,1} \oplus \left(I_{j,2} - \frac{2}{\|\xi\|^2} \xi \otimes \bar{\xi} \right) \qquad \xi \in P_j^{\perp} \mathcal{H},$$

where $I_{j,1}$ is the identity on $P_j\mathcal{H}$, $I_{j,2}$ is the identity on $P_j^{\perp}\mathcal{H}$ and

$$\langle \xi, e_l^j \rangle = 0, \qquad l > jk + 1,$$

where $\{e_l^j\}_{l \in \mathbb{N}}$ is the natural basis for $P_j^{\perp}\mathcal{H}$ inherited from the basis $\{e_k\}$, and this yields (8.2.3). Note that (8.2.4) follows immediately from (8.2.3). □

Thus, if A has only k subdiagonals, we only need information from the finite section $P_{nk+2}AP_{nk+2}$ of A in order to compute the section $P_nH_nP_n$ of H_n. The following algorithm displays an easy implementation of the result in the previous theorem. Note that this algorithm is almost like the finite dimensional case except that the output is a finite section of the infinite Hessenberg reduction, rather than a Hessenberg reduction of a finite section of the infinite matrix.

Algorithm 8.2.1.

```
%Computes a section of the Hessenberg reduction of an
%infinite matrix with number of subdiagonals = k.

function J = Inf_Hessen(A,k)
B = A;
d = size(A,1);
m = (d-1)/k - 1;
for j = 1:m
    u = House(B(j+1:j*k + 1,j));
    B(j+1:j*k + 1,1:d) = B(j+1:j*k + 1,1:d) - 2*u*(u'*B(j+1:j*k + 1,1:d));
    B(1:d,j+1:j*k + 1) = B(1:d,j+1:j*k + 1) - 2*(B(1:d,j+1:j*k + 1)*u)*u';
end
J = B(1:m,1:m);
```

Algorithm 8.2.2.

```
% House(x) takes a vector x and creates a vector u
% such that (I + u*u')x = ce_1 where c is some complex
% number (depending on x) and e_1 = [1,0...].

function u = House(x)
v = x;
if v(1) == 0
   v(1) = v(1) + norm(v);          %This is the classical way
else                               %of creating Householder reflections
v(1) = x(1) + sign(x(1))*norm(x);  %as in finite dimensions.
end
u = v/norm(v);
```

Remark 8.2.2. We would like to emphasize that the input of Algorithm 8.2.1 is not an infinite matrix, but rather a section of an infinite matrix, where the choice of such a section is justified by Theorem 8.2.1, e.g. if the size of the section P_lAP_l of the infinite matrix A (with k-sub diagonals) is $l = nk + 2$, and P_lAP_l is put into Algorithm 8.2.1, then the output of Algorithm 8.2.1 is an n-by-n section $P_nH_nP_n$ of the n-th term H_n of the Hessenberg reduction of A.

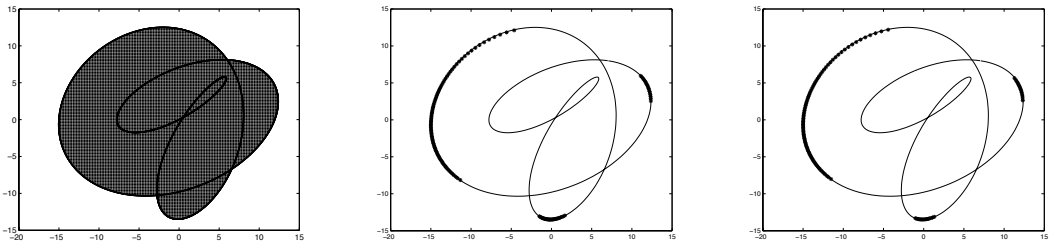


Figure 8.1: The first figure shows $\sigma(A)$ whereas the second and the third shows the output of the Infinite QR algorithm without and with the Hessenberg reduction (the fat dots), respectively. The thin line is the image of the symbol from (8.3.1).

8.3 Numerical Examples

Throughout this section with numerical examples we will, if $A \in \mathcal{B}(l^2(\mathbb{N}))$, let

$$A_n = Q_n A Q_n^*, \qquad H_n = U_n \cdots U_1 A U_1 \cdots U_n$$

denote the n-th iteration in the Infinite QR algorithm and the n-th term in the Hessenberg reduction, described in (8.1.2), respectively. Also, P_m will as usual always denote the projection onto $\text{span}\{e_1, \ldots, e_m\}$.

8.3.1 Comparison

The purpose of these examples is to compare the classical Infinite QR algorithm without Hessenberg reduction with the Infinite QR algorithm with Hessenberg reduction. Let

$$A = \begin{pmatrix} 0 & 2i & -2i & 2 & 0 & 0 & \ldots \\ 0 & 0 & 2i & -2i & 2 & 0 & \ldots \\ 5 & 0 & 0 & 2i & -2i & 2 & \ldots \\ 8i & 5 & 0 & 0 & 2i & -2i & \ldots \\ 0 & 8i & 5 & 0 & 0 & 2i & \ldots \\ 0 & 0 & 8i & 5 & 0 & 0 & \ldots \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \ddots \end{pmatrix}.$$

Now, A is a Toeplitz operator with symbol

$$f(z) = 8iz^{-3} + 5z^{-2} + 2iz - 2iz^2 + 2z^3, \tag{8.3.1}$$

and in this case the spectrum of A is well known and is visualized in the first picture of Figure 8.1. In the second and third picture we have plotted (in fat dots)

$$\sigma(P_m Q_n A Q_n^* \lceil_{P_m l^2(\mathbb{N})}), \qquad \sigma(P_m Q_n H_k Q_n^* \lceil_{P_m l^2(\mathbb{N})}),$$

respectively (where $m = 120$, $n = 750$, $k = 760$), together with the image (the thin line) of the symbol f. Note that up to the resolution of the picture the output is equivalent, however, the computational cost of the third picture is roughly half the computational cost (in CPU-time) of the second picture.

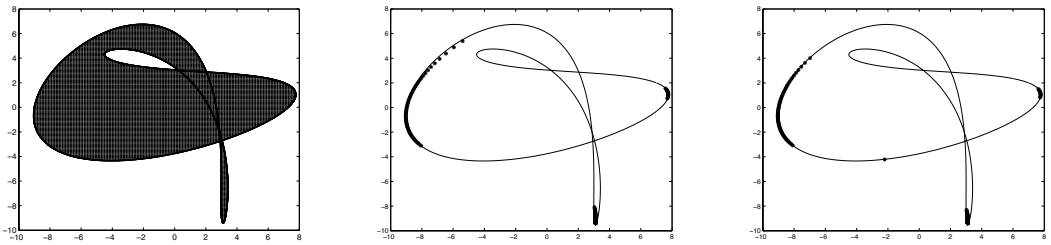


Figure 8.2: The first figure shows $\sigma(B)$ whereas the second and the third shows the output of the Infinite QR algorithm without and with the Hessenberg reduction (the fat dots), respectively. The thin line is the image of the symbol from (8.3.2).

Let

$$B = \begin{pmatrix} 0 & 2i & -2i & 2 & 0 & 0 & \dots \\ 0 & 0 & 2i & -2i & 2 & 0 & \dots \\ 5 & 0 & 0 & 2i & -2i & 2 & \dots \\ 2i & 5 & 0 & 0 & 2i & -2i & \dots \\ 0 & 2i & 5 & 0 & 0 & 2i & \dots \\ 0 & 0 & 2i & 5 & 0 & 0 & \dots \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \ddots \end{pmatrix},$$

Now, B is also a Toeplitz operator and has symbol

$$g(z) = 2iz^{-3} + 5z^{-2} + 2iz - 2iz^2 + 2z^3. \tag{8.3.2}$$

Figure 8.2 shows $\sigma(B)$ (first picture) and

$$\sigma(P_m Q_n B Q_n^* \lceil_{P_m l^2(\mathbb{N})}), \qquad \sigma(P_m Q_n H_k Q_n^* \lceil_{P_m l^2(\mathbb{N})}),$$

respectively (where $H_k = U_k \cdots U_1 B U_1 \cdots U_k$ $m = 120$, $n = 750$, $k = 880$), together with the image of the symbol of B (thin line). Note that in this case the Infinite QR algorithm with the Hessenberg reduction performs slightly better than the Infinite QR algorithm without the Hessenberg reduction.

8.3.2 Shifting Strategies

A common technique to speed up the QR algorithm in finite dimensions is to use shifts of the matrix as the algorithms proceed. The idea is roughly as follows. If one has a guess of a possible eigenvalue λ of the matrix A then one may use the QR algorithm on $A - \lambda I$ instead. The reason is that, in general when using the QR algorithm, the eigenvalues with the largest and smallest modulus will need the least amount of iterations to get the desired accuracy. One may say that one gets faster convergence to the largest and smallest (in modulus) eigenvalues. Since the smallest eigenvalue of $A - \lambda I$ should be close to zero (if λ was a good guess), by the heuristic argument above, it may be more efficient (in order to get the correct $\tilde{\lambda}$ for which λ was an approximation to) to use the QR algorithm on $A - \lambda I$ rather than A. Usually, the shift is updated as the algorithm proceeds.

A natural question to ask is whether this approach can be used in the infinite-dimensional case in order to speed up the convergence. Unfortunately, the answer is no. In the

infinite-dimensional case it is only the "largest" (in modulus) part of the spectrum that is dominating and not the smallest. However, one may use shifting strategies in the infinite-dimensional case, not in order to increase efficiency, but in order to get more spectral information from the operator.

The convergence of the Infinite QR algorithm is still a mystery. However, as the numerical results suggest, one is able to capture the largest (in modulus) part of the spectrum of the operator (the word "largest" here is used heuristically and due to the lack of theoretical support we are forced to use vague vocabulary, however, we believe the reader will understand the outline of the ideas from the examples) Now, of course, by shifting and rotating the operator one shifts and rotates the spectrum as well. Hence, in this way one should be able to capture other parts of the spectrum with the Infinite QR algorithm. This is best visualized by an example.

Let

$$A = \begin{pmatrix} 3.5 & 0 & 0 & 0 & 0 & 0 & \dots \\ -2 & -3.5 & 0 & 0 & 0 & 0 & \dots \\ 0.5i & -2 & 4i & 0 & 0 & 0 & \dots \\ 1 & 0.5i & -2 & -4i & 0 & 0 & \dots \\ 0 & 1 & 0.5i & -2 & 0 & 0 & \dots \\ 0 & 0 & 1 & 0.5i & -2 & 0 & \dots \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \ddots \end{pmatrix}.$$

Then A is a compact perturbation of a Toeplitz operator with symbol

$$f(z) = z^{-3} + 0.5z^{-2} - 2z^{-1}, \tag{8.3.3}$$

and the spectrum of A is visualized in Figure 8.3. If one runs the Infinite QR algorithm on A one gets the second figure in Figure 8.3. Here we have plotted the output of the Infinite QR algorithm with fat dots and the image of the circle under the symbol f as the fine line. Note that the "largest" part of the spectrum (in modulus) corresponds to the lower part of the spectrum and is picked up by the algorithm. It is clear from the picture (and the heuristically exposition in the introduction to this section) that by shifting the operator to the left, using the Infinite QR algorithm and then map the output back according to the shifting, one could hope to get the left part of the spectrum. This is visualized in the second picture in Figure 8.3, where we have kept the output from the first computation as well. The next two pictures in Figure 8.3 are made by using the shifting strategy suggested to the right (third picture) and upwards (fourth picture). In both cases the output from the previous computation has been kept, i.e. Figure 8.3 shows the following sets

$$\begin{aligned} &\sigma(A), \\ &\sigma(P_m Q_n A Q_n^* \lceil_{P_m l(\mathbb{N})}), \\ &\sigma(P_m Q_n A Q_n^* \lceil_{P_m l(\mathbb{N})}) \cup \{\lambda + \omega_1 : \lambda \in \sigma(P_m Q_n (A - \omega_1 I) Q_n^* \lceil_{P_m l(\mathbb{N})})\}, \\ &\sigma(P_m Q_n A Q_n^* \lceil_{P_m l(\mathbb{N})}) \cup \bigcup_{j=1}^{2} \{\lambda + \omega_j : \lambda \in \sigma(P_m Q_n (A - \omega_j I) Q_n^* \lceil_{P_m l(\mathbb{N})})\}, \\ &\sigma(P_m Q_n A Q_n^* \lceil_{P_m l(\mathbb{N})}) \cup \bigcup_{j=1}^{3} \{\lambda + \omega_j : \lambda \in \sigma(P_m Q_n (A - \omega_j I) Q_n^* \lceil_{P_m l(\mathbb{N})})\}, \end{aligned} \tag{8.3.4}$$

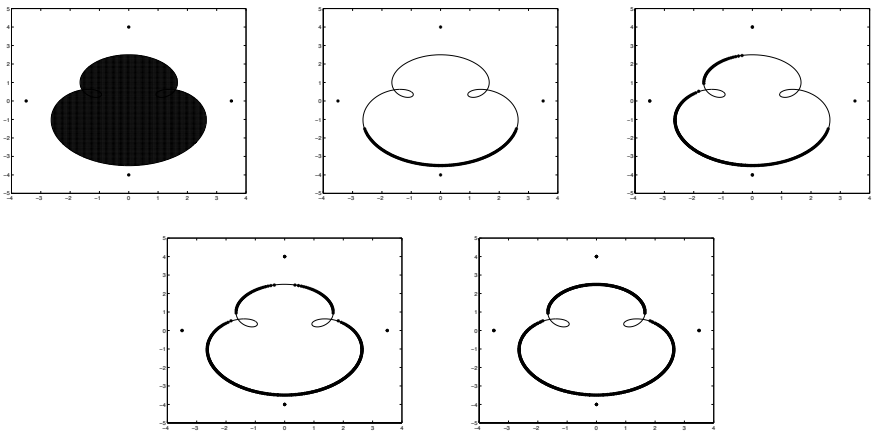


Figure 8.3: The figure shows the sets from (8.3.4) (the fat dots) together with the image of the symbol (8.3.3)(thin line).

where $\omega_1 = 5$, $\omega_2 = -5i$ and $\omega_3 = -5$. Also, $m = 135$ and $n = 660$.

The ideas in the previous example can be summarized in the following "Rotate and Shift"-algorithm.

Algorithm 8.3.1.
```
%Takes an infinite matrix A of upper Hessenberg form
%and computes
%      spec(P_mQ_1(exp(k*i*2*pi/n)*A + distance*I)Q*_1P_m) - distance
%for k = 1,...,n, where m = section_size.

function V = Rotate(A,n,section_size,distance)
s = size(A,2);
for k=1:n
    J = exp(k*i*2*pi/n)*A;
    J = J + distance*eye(s);
    B = Infinite_QR(J,s-section_size,1);
    f = eig(B);
    h = size(f,1);
    f = f - distance*ones(h,1);
    f = exp(-k*i*m)*f;
    plot(f,'k*');
    hold on
end
```

Remark 8.3.1. We would like to emphasize that the input in Algorithm 8.3.1 is actually not an infinite matrix, but a section of the matrix according to Theorem 7.4.3. In particular, if $A \in \mathcal{B}(l^2(\mathbb{N}))$ is an infinite matrix and the section $P_s A P_s$ is put into Algorithm 8.3.1 together with the variable section size $= m$, then the number of iterations used in the Infinite QR algorithm in Algorithm 8.3.1 is equal to $s - m$.

Note that one of the great strengths of the Infinite QR algorithm is that it is robust with respect to perturbations, and this is why the Infinite QR algorithm is an important supplement to the methods introduced in Chapter 6. If we introduce a perturbation of A,

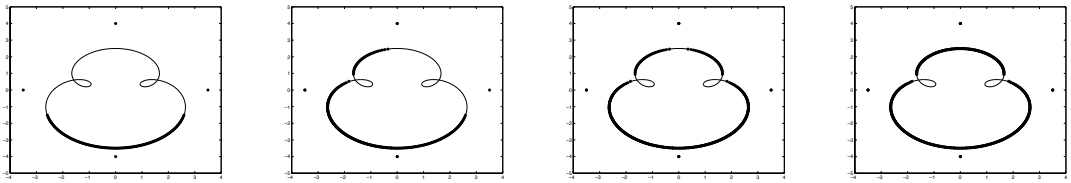


Figure 8.4: The figure shows the sets from (8.3.4) with A replaced by $\tilde{A}$ (the fat dots) together with the image of the symbol (8.3.3)(thin line).

namely, the highly non-normal matrix

$$\tilde{A} = \begin{pmatrix} 3.5 & 0 & 0 & 0 & 0 & 0 & \dots \\ 10^8 & -3.5 & 0 & 0 & 0 & 0 & \dots \\ 0.5i & -2 & 4i & 0 & 0 & 0 & \dots \\ 1 & 0.5i & -2 & -4i & 0 & 0 & \dots \\ 0 & 1 & 0.5i & -2 & 0 & 0 & \dots \\ 0 & 0 & 1 & 0.5i & -2 & 0 & \dots \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \ddots \end{pmatrix},$$

the shifting strategy and the Infinite QR algorithm work surprisingly well. In Figure 8.4 we have repeated the numerical experiment visualized in Figure 8.3, but replaced A with $\tilde{A}$.

8.3.3 Shifting Strategies and Hessenberg Reduction

It is obvious, from the previous examples, that the success of the rotate and shift technique presented above will depend on the geometry of the spectrum of the operator one is considering. A star shaped spectrum is obviously going to be hard (most likely impossible) to detect whereas shapes close to a circle may be much easier. In other words, the closer the the shape of the spectrum is to a convex set, the better.

Now, the problem with the rotate and shift strategy is that for each rotation one has to run the Infinite QR algorithm again, and this is computationally expensive when the infinite matrix has many non-zero subdiagonals. To combat this obstacle the idea is that if $A \in \mathcal{B}(l^2(\mathbb{N}))$ has more than one non-zero sub-diagonal, then one computes the semi upper Hessenberg matrix

$$H_k = U_k \cdots U_1 A U_1 \cdots U_k,$$

namely, the k-th term in the Hessenberg reduction, described in (8.1.2). This infinite matrix has the same spectrum as A, but because $P_k H_k P_k$ is upper Hessenberg it is much more suitable for computations. Hence, H_k is computed only once, and then the rotate and shift technique is used on H_k rather than A. Obviously k has to be chosen according to n where n denotes the number of iterations used in the Infinite QR algorithm. We will explore the strategy explained above in the following examples.

In this example we will demonstrate the rotate and shift technique on an operator with

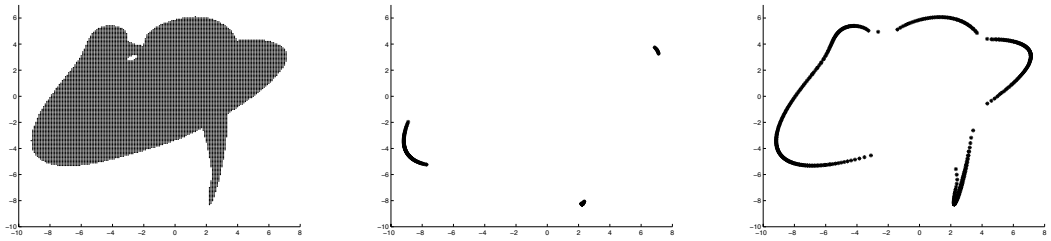


Figure 8.5: The first figure shows $\sigma(A)$, the second figure shows the output of the Infinite QR algorithm, the third show the output of the Infinite QR algorithm together with Hessenberg reduction and the shift and rotate techniqe.

a slightly more complicated shape of its spectrum. Let

$$A = \begin{pmatrix} 0 & 2i & -2i & 2 & 0 & 0 & 0 & \dots \\ 0 & 0 & 2i & i & 2 & 0 & 0 & \dots \\ 5 & 0 & 0 & 2i & -2i & 2 & 0 & \dots \\ 2i & 5 & 0 & 0 & 2i & i & 2 & \dots \\ 0 & 2i & 5 & 0 & 0 & 2i & -2i & \dots \\ 0 & 0 & 2i & 5 & 0 & 0 & 2i & \dots \\ 0 & 0 & 0 & 2i & 5 & 0 & 0 & \dots \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \ddots \end{pmatrix}.$$

In this case A is not a usual Toeplitz operator, however, the spectrum of A can be computed using the methods in Chapter 6, and this is visualized in the first picture of Figure 8.5. The second picture in Figure 8.5 shows

$$\sigma(P_m Q_n A Q_n^* \lceil_{P_m l^2(\mathbb{N})}), \qquad m = 120,\, n = 950,$$

and the third picture shows the output of Algorithm 8.3.1 applied to

$$H_k = U_k \cdots U_1 A U_1 \cdots U_k, \qquad k = 960,$$

with four rotations i.e. the following set is visualized

$$\bigcup_{j=1}^{4} \{e^{-\frac{2\pi i}{4}j}\lambda - \omega : \lambda \in \sigma(P_m Q_n (e^{\frac{2\pi i}{4}j} H_k + \omega I) Q_n^* \lceil_{P_m l(\mathbb{N})})\},$$

where $m = 120$, $\omega = 7$ and $n = 950$.

In the next example we use exactly the same strategy as in the previous example, however we leave the "Toeplitz like"-type of operators and also recall how poorly the

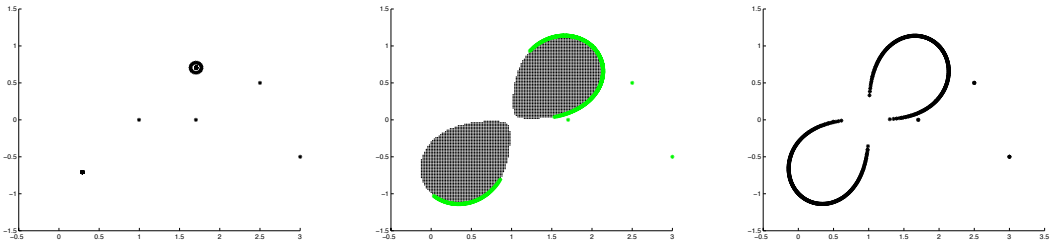


Figure 8.6: The first picture shows the output of the finite section method, the second picture shows $\sigma(B)$ (the dark plot) together with the output of the Infinite QR algorithm applied to B (light plot) and the third picture shows the output of the rotate and shift algorithm applied to B.

finite section may perform on non-self-adjoint problems. Let

$$B = \begin{pmatrix} 2.5+0.5i & 0 & 0 & 0 & 0 & 0 & 0 & \dots \\ 1 & 3-0.5i & 0 & 0 & 0 & 0 & 0 & \dots \\ 0 & 1 & 1.7 & 0.05 & 0 & 0 & 0 & \dots \\ 0 & 0 & 0.05 & 1 & 0 & 0 & 0 & \dots \\ 0 & 0 & 0 & 1 & 1 & a_{56} & 0 & \dots \\ 0 & 0 & 0 & 0 & 1 & 1 & 0 & \dots \\ 0 & 0 & 0 & 0 & 0 & 1 & 1 & \dots \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \ddots \end{pmatrix},$$

where $a_{2j-1,2j} = i$ for $j \geq 3$. In the first picture of Figure 8.6 we have shown $\sigma(P_m B\lceil_{P_m l^2(\mathbb{N})})$, where $m = 500$, to recall the rather poor performance of the finite section method. The second picture shows the spectrum of B (the dark color) together with $\sigma((P_m Q_n B Q_n^*\lceil_{P_m l^2(\mathbb{N})})$ (light color), where $m = 600$, $n = 900$. The third picture shows the output of Algorithm 8.3.1 applied to

$$H_k = U_k \cdots U_1 B U_1 \cdots U_k, \qquad k = 1400,$$

with four rotations i.e. the following set is visualized

$$\bigcup_{j=1}^{4} \{e^{-\frac{2\pi i}{4}j}\lambda - \omega : \lambda \in \sigma(P_m Q_n (e^{\frac{2\pi i}{4}j} H_k + \omega I) Q_n^*\lceil_{P_m l(\mathbb{N})})\},$$

where $m = 200$, $\omega = 7$ and $n = 1400$.

As we see from the numerical examples, the Infinite QR algorithm performs very well on non-normal problems, and one is able to get large parts of the boundary of the essential spectrum. If we should try to predict something about what we might expect to prove, a guess would be that one should be able to recover

$$\partial \mathrm{conv}(\sigma_e(T)) \cap \sigma(T), \qquad T \in \mathcal{B}(\mathcal{H})$$

with the Infinite QR algorithm, where T is assumed to have k subdiagonals with respect to some basis and $\mathrm{conv}(\sigma_e(T))$ denotes the convex hull of the essential spectrum.

Closing Remarks

The main result in this thesis is the development of methods that allow approximations and computations of spectra and pseudospectra of a large class of operators, including the whole $\mathcal{B}(\mathcal{H})$ and large parts of $\mathcal{C}(\mathcal{H})$. However, there are several important unanswered questions related to both theory and applications.

Concluding Remarks on Theory

The main question that is left open is the following: What is the complexity index of the spectrum when one allows operators from the whole $\mathcal{C}(\mathcal{H})$? The first thing we need to determine is whether or not it is greater than one. There is absolutely nothing that suggest that it should be one, but we must not rule out the possibility. However, if it turns out to be one, this would have a dramatic impact in applications. It would essentially mean that, from a complexity point of view, it is just as easy to approximate the spectrum of a finite dimensional matrix as it is to approximate the spectrum of an arbitrary closed operator on a separable Hilbert space. This is slightly counter intuitive, however, this is yet to be proved or disproved.

The previously suggested general problem is of course important, but with the establishment of the complexity index, the whole theory of classifying computational spectral problems in terms of their complexity emerges. We have shown that the complexity index for the spectrum, when considering self-adjoint operators, is less than or equal to three, however, we believe strongly that it is strictly less than three.

The theoretical exposition of the QR algorithm in this thesis is probably just scratching the surface. Note that the numerical examples suggest that much more than what we have rigorously shown here is true. In our theoretical framework on the QR algorithm we only consider normal operators. This is a natural extension of the work by Deift et al., however, it is in the non-normal case that the QR algorithm really shows its strength, and thus theoretical tools for proving convergence in this case are absolutely crucial.

Concluding Remarks on Applications

The main task left open regarding application is how to improve the algorithms, in particular, how to speed them up. As the reader may have observed, the algorithms suggested in Part-II are (although robust) quite simple. One would therefore think that there is room for vast improvements. Also, as the main goal for this thesis has been generality, special cases have not been given priority. As we now have reliable general methods, it would be important to take advantage of additional structure (typically other structural properties

than self-adjointness). This has been done in the case of banded infinite matrices, but what about particular types of operators in mathematical physics?

Some applications to Schrödinger and Dirac operators have been discussed, but only implemented in the discrete case. It would therefore be interesting to see if the methods suggested here could be used in non-hermitian quantum mechanics. The framework indicated in this dissertation is based on computing spectra from the matrix elements rather than from discretizations of differential operators. This means that one must find reliable ways of computing the matrix elements. In the case of operators in quantum mechanics, one must compute inner products of elements in $L^2(\mathbb{R}^n)$, and hence, one should probably join forces with the numerical integration community in order to pursue this project successfully.

Bibliography

[Arv91] William Arveson, *Discretized CCR algebras*, J. Operator Theory **26** (1991), no. 2, 225–239. MR MR1225515 (94f:46069)

[Arv93a] ———, *Improper filtrations for C^*-algebras: spectra of unilateral tridiagonal operators*, Acta Sci. Math. (Szeged) **57** (1993), no. 1-4, 11–24. MR MR1243265 (94i:46071)

[Arv93b] ———, *Noncommutative spheres and numerical quantum mechanics*, Operator algebras, mathematical physics, and low-dimensional topology (Istanbul, 1991), Res. Notes Math., vol. 5, A K Peters, Wellesley, MA, 1993, pp. 1–10. MR MR1259055

[Arv94a] ———, *C^*-algebras and numerical linear algebra*, J. Funct. Anal. **122** (1994), no. 2, 333–360. MR MR1276162 (95i:46083)

[Arv94b] ———, *The role of C^*-algebras in infinite-dimensional numerical linear algebra*, C^*-algebras: 1943–1993 (San Antonio, TX, 1993), Contemp. Math., vol. 167, Amer. Math. Soc., Providence, RI, 1994, pp. 114–129. MR MR1292012 (95i:46084)

[BCN01] A. Böttcher, A. V. Chithra, and M. N. N. Namboodiri, *Approximation of approximation numbers by truncation*, Integral Equations Operator Theory **39** (2001), no. 4, 387–395. MR MR1829276 (2002b:47035)

[Béd97] Erik Bédos, *On Følner nets, Szegő's theorem and other eigenvalue distribution theorems*, Exposition. Math. **15** (1997), no. 3, 193–228. MR MR1458766 (98h:47065a)

[Ber71] I. David Berg, *An extension of the Weyl-von Neumann theorem to normal operators*, Trans. Amer. Math. Soc. **160** (1971), 365–371. MR MR0283610 (44 #840)

[Böt00] Albrecht Böttcher, *C^*-algebras in numerical analysis*, Irish Math. Soc. Bull. (2000), no. 45, 57–133. MR MR1832325 (2002b:46116)

[Bou06] Lyonell Boulton, *Limiting set of second order spectra*, Math. Comp. **75** (2006), no. 255, 1367–1382 (electronic). MR MR2219033 (2007h:47049)

[Bou07] ———, *Non-variational approximation of discrete eigenvalues of self-adjoint operators*, IMA J. Numer. Anal. **27** (2007), no. 1, 102–121. MR MR2289273 (2007m:47034)

[Bro86] L. G. Brown, *Lidskiĭ's theorem in the type* II *case*, Geometric methods in operator algebras (Kyoto, 1983), Pitman Res. Notes Math. Ser., vol. 123, Longman Sci. Tech., Harlow, 1986, pp. 1–35. MR MR866489 (88d:47024)

[Bro06] Nathanial P. Brown, *AF embeddings and the numerical computation of spectra in irrational rotation algebras*, Numer. Funct. Anal. Optim. **27** (2006), no. 5-6, 517–528. MR MR2246575

[Bro07a] ———, *Quasi-diagonality and the finite section method*, Math. Comp. **76** (2007), no. 257, 339–360.

[Bro07b] ———, *Quasi-diagonality and the finite section method*, Math. Comp. **76** (2007), no. 257, 339–360 (electronic). MR MR2261025 (2007h:65057)

[BS99] Albrecht Böttcher and Bernd Silbermann, *Introduction to large truncated Toeplitz matrices*, Universitext, Springer-Verlag, New York, 1999. MR MR1724795 (2001b:47043)

[CL63] H. O. Cordes and J. P. Labrousse, *The invariance of the index in the metric space of closed operators*, J. Math. Mech. **12** (1963), 693–719. MR MR0162142 (28 #5341)

[Cup81] J. J. M. Cuppen, *A divide and conquer method for the symmetric tridiagonal eigenproblem*, Numer. Math. **36** (1980/81), no. 2, 177–195. MR MR611491 (82d:65038)

[Dav98] E. B. Davies, *Spectral enclosures and complex resonances for general self-adjoint operators*, LMS J. Comput. Math. **1** (1998), 42–74 (electronic). MR MR1635727 (2000e:47043)

[Dav99] ———, *Semi-classical states for non-self-adjoint Schrödinger operators*, Comm. Math. Phys. **200** (1999), no. 1, 35–41. MR MR1671904 (99m:34197)

[Dav00] ———, *A hierarchical method for obtaining eigenvalue enclosures*, Math. Comp. **69** (2000), no. 232, 1435–1455. MR MR1710648 (2001a:34148)

[Dav01] ———, *Spectral properties of random non-self-adjoint matrices and operators*, R. Soc. Lond. Proc. Ser. A Math. Phys. Eng. Sci. **457** (2001), no. 2005, 191–206. MR MR1843941 (2002f:47082)

[Dav02] ———, *Non-self-adjoint differential operators*, Bull. London Math. Soc. **34** (2002), no. 5, 513–532. MR MR1912874 (2003c:47081)

[Dav05] ———, *A defence of mathematical pluralism*, Philos. Math. (3) **13** (2005), no. 3, 252–276. MR MR2192174 (2006m:00005)

[DK04] E. B. Davies and A. B. J. Kuijlaars, *Spectral asymptotics of the non-self-adjoint harmonic oscillator*, J. London Math. Soc. (2) **70** (2004), no. 2, 420–426. MR MR2078902 (2005e:34266)

[DLT85] P. Deift, L. C. Li, and C. Tomei, *Toda flows with infinitely many variables*, J. Funct. Anal. **64** (1985), no. 3, 358–402. MR MR813206 (87a:58076)

[DP04] E. B. Davies and M. Plum, *Spectral pollution*, IMA J. Numer. Anal. **24** (2004), no. 3, 417–438. MR MR2068830 (2005c:47027b)

[DS88] Nelson Dunford and Jacob T. Schwartz, *Linear operators. Part I*, Wiley Classics Library, John Wiley & Sons Inc., New York, 1988, General theory, With the assistance of William G. Bade and Robert G. Bartle, Reprint of the 1958 original, A Wiley-Interscience Publication. MR MR1009162 (90g:47001a)

[DSZ04] Nils Dencker, Johannes Sjöstrand, and Maciej Zworski, *Pseudospectra of semi-classical (pseudo-) differential operators*, Comm. Pure Appl. Math. **57** (2004), no. 3, 384–415. MR MR2020109 (2004k:35432)

[DVV94] Trond Digernes, V. S. Varadarajan, and S. R. S. Varadhan, *Finite approximations to quantum systems*, Rev. Math. Phys. **6** (1994), no. 4, 621–648. MR MR1290691 (96e:81028)

[FK52] Bent Fuglede and Richard V. Kadison, *Determinant theory in finite factors*, Ann. of Math. (2) **55** (1952), 520–530. MR MR0052696 (14,660a)

[GVL96] Gene H. Golub and Charles F. Van Loan, *Matrix computations*, third ed., Johns Hopkins Studies in the Mathematical Sciences, Johns Hopkins University Press, Baltimore, MD, 1996. MR MR1417720 (97g:65006)

[Hal70] P. R. Halmos, *Ten problems in Hilbert space*, Bull. Amer. Math. Soc. **76** (1970), 887–933. MR MR0270173 (42 #5066)

[Hana] A.C. Hansen, *Hessenberg reduction and the infinite-dimensional qr algorithm*, Preprint.

[Hanb] ———, *The infinite-dimensional qr algorithm*, Submitted.

[Han08] Anders C. Hansen, *On the approximation of spectra of linear operators on Hilbert spaces*, J. Funct. Anal. **254** (2008), no. 8, 2092–2126. MR 2402104 (2009c:47004)

[Han10] ———, *Infinite-dimensional numerical linear algebra: theory and applications*, R. Soc. Lond. Proc. Ser. A Math. Phys. Eng. Sci. **466** (2010), no. 2124, 3539–3559.

[Han11] ———, *On the solvability complexity index, the n-pseudospectrum and approximations of spectra of operators*, J. Amer. Math. Soc. **24** (2011), no. 1, 81–124. MR 2726600

[HK76] W. K. Hayman and P. B. Kennedy, *Subharmonic functions. Vol. I*, Academic Press [Harcourt Brace Jovanovich Publishers], London, 1976, London Mathematical Society Monographs, No. 9. MR MR0460672 (57 #665)

[HN96] Naomichi Hatano and David R. Nelson, *Localization transitions in non-hermitian quantum mechanics*, Phys. Rev. Lett. **77** (1996), no. 3, 570–573.

[HN97] ———, *Vortex pinning and non-hermitian quantum mechanics*, Phys. Rev. B **56** (1997), no. 14, 8651–8673.

[HRS01] Roland Hagen, Steffen Roch, and Bernd Silbermann, *C^*-algebras and numerical analysis*, Monographs and Textbooks in Pure and Applied Mathematics, vol. 236, Marcel Dekker Inc., New York, 2001. MR MR1792428 (2002g:46133)

[HS07] Uffe Haagerup and Hanne Schultz, *Brown measures of unbounded operators affiliated with a finite von Neumann algebra*, Math. Scand. **100** (2007), no. 2, 209–263. MR MR2339369

[Kat95] Tosio Kato, *Perturbation theory for linear operators*, Classics in Mathematics, Springer-Verlag, Berlin, 1995, Reprint of the 1980 edition. MR MR1335452 (96a:47025)

[KR97] Richard V. Kadison and John R. Ringrose, *Fundamentals of the theory of operator algebras. Vol. I*, Graduate Studies in Mathematics, vol. 15, American Mathematical Society, Providence, RI, 1997, Elementary theory, Reprint of the 1983 original. MR MR1468229 (98f:46001a)

[LS96] A. Laptev and Yu. Safarov, *Szegő type limit theorems*, J. Funct. Anal. **138** (1996), no. 2, 544–559. MR MR1395969 (97k:58166)

[LS04] Michael Levitin and Eugene Shargorodsky, *Spectral pollution and second-order relative spectra for self-adjoint operators*, IMA J. Numer. Anal. **24** (2004), no. 3, 393–416. MR MR2068829 (2005c:47027a)

[Par65] Beresford Parlett, *Convergence of the QR algorithm*, Numer. Math. **7** (1965), 187–193. MR MR0176600 (31 #872)

[PK69] B. N. Parlett and W. Kahan, *On the convergence of a practical* QR *algorithm. (With discussion)*, Information Processing 68 (Proc. IFIP Congress, Edinburgh, 1968), Vol. 1: Mathematics, Software, North-Holland, Amsterdam, 1969, pp. 114–118. MR MR0255035 (40 #8242)

[Pok79] Andrzej Pokrzywa, *Method of orthogonal projections and approximation of the spectrum of a bounded operator*, Studia Math. **65** (1979), no. 1, 21–29. MR MR554538 (81m:47022)

[PP73] B. N. Parlett and W. G. Poole, Jr., *A geometric theory for the* QR, LU *and power iterations*, SIAM J. Numer. Anal. **10** (1973), 389–412, Collection of articles dedicated to the memory of George E. Forsythe. MR MR0336979 (49 #1752)

[Pui04] Joaquim Puig, *Cantor spectrum for the almost Mathieu operator*, Comm. Math. Phys. **244** (2004), no. 2, 297–309. MR MR2031032 (2004k:11129)

[RS72] Michael Reed and Barry Simon, *Methods of modern mathematical physics. I. Functional analysis*, Academic Press, New York, 1972. MR MR0493419 (58 #12429a)

[Sha00] Eugene Shargorodsky, *Geometry of higher order relative spectra and projection methods*, J. Operator Theory **44** (2000), no. 1, 43–62. MR MR1774693 (2001f:47004)

[Sha08] E. Shargorodsky, *On the level sets of the resolvent norm of a linear operator*, Bull. Lond. Math. Soc. **40** (2008), no. 3, 493–504. MR 2418805 (2009d:47004)

[Sze20] G. Szegő, *Beiträge zur Theorie der Toeplitzschen Formen*, Math. Z. **6** (1920), no. 3-4, 167–202. MR MR1544404

[TC04] Lloyd N. Trefethen and S. J. Chapman, *Wave packet pseudomodes of twisted Toeplitz matrices*, Comm. Pure Appl. Math. **57** (2004), no. 9, 1233–1264. MR MR2059680 (2005c:47035)

[TE05] Lloyd N. Trefethen and Mark Embree, *Spectra and pseudospectra*, Princeton University Press, Princeton, NJ, 2005, The behavior of nonnormal matrices and operators. MR MR2155029 (2006d:15001)

[Tre04] Sergei Treil, *An operator Corona theorem*, Indiana Univ. Math. J. **53** (2004), no. 6, 1763–1780. MR MR2106344 (2005j:30067)

[Wat82] David S. Watkins, *Understanding the* QR *algorithm*, SIAM Rev. **24** (1982), no. 4, 427–440. MR MR678561 (84g:65047)

[Wil65] J. H. Wilkinson, *Convergence of the* LR*,* QR*, and related algorithms*, Comput. J. **8** (1965), 77–84. MR MR0183108 (32 #590)