Bergische Universität Wuppertal

The Computation of Spectral Supersets of Linear Operators in Hilbert Spaces

Dissertation
zur Erlangung des akademischen Grades
Doktor der Naturwissenschaften (Dr. rer. nat.)
an der
Fakultät für Mathematik und Naturwissenschaften
Fachgruppe Mathematik und Informatik

vorgelegt von
Lukas Alexander Vorberg

betreut durch
Prof. Birgit Jacob

Wuppertal, März 2024

Prüfungstermin:	14. Februar 2024
Gutachter:	Prof. Birgit Jacob Prof. Marco Marletta
Prüfungskommission:	Prof. Birgit Jacob Prof. Bálint Farkas Prof. Andreas Frommer Dr. Christian Wyss

Acknowledgments

I would like to express my deepest gratitude to the following individuals, without whom this thesis would not have been possible:

First and foremost, I am immensely thankful to my supervisor, *Birgit Jacob*. Her constructive support, insightful guidance, and ability to foster a very harmonious working environment within our group were invaluable throughout this journey.

My profound gratitude is also extended to my second supervisor, *Christian Wyss*, whose support and extensive knowledge, coupled with a remarkable intuition, significantly enriched this work in many areas.

Special thanks go to *Marco Marletta*, who not only provided me with the opportunity for a research stay at Cardiff University but also shared inspiring insights that contributed to the development of the algorithm for computing the quadratic numerical range.

I am also very grateful to *Sabine Bögli*, who generously offered me the chance for a research stay at Durham University. Her contributions and ideas regarding the relationship between the pseudospectra of block operator matrices and their Schur complements were instrumental.

Sincere appreciation is expressed to *Felix Schwenninger*, who not only provided expert guidance when co-supervising my master's thesis but also offered me the opportunity for a research stay at the University of Twente, which led to my first published article [28].

My thankfulness is extended to *Jochen Glück* and *Julian Hölz* for their valuable input, which played a pivotal role in achieving the results related to the concentration phenomenon for random sampling of QNR points.

I further give thanks to *Andreas Frommer* and *Karsten Kahl* for engaging and fruitful discussions that have substantially broadened my insight.

To my fellow PhD students *Annika*, *Julian*, *Mehmet*, *Merlin*, *Nathanael* and *René*, I am grateful not only for the academic exchange but also for the cherished friendships we have forged over the years. Our shared experiences, whether in the office or during leisure activities like roundnet and bouldering, have greatly enriched this journey.

My heartfelt thanks go to my girlfriend, *Johanna*, whose presence has brought boundless joy and vitality to my free time. Her support and companionship have been a wellspring of strength, allowing me to recharge and approach my work with renewed vigor.

Mein tiefster Dank gilt meinen Eltern, *Susanne* und *Stefan*, für ihre unermüdliche Unterstützung und Förderung auf jedem Schritt dieses Weges. Meinen Großeltern, *Gerda*, *Willi*, *Lieselotte* und *Karl Friedrich*, möchte ich für die Lebenserfahrungen und Werte, die sie an mich weitergegeben haben, sowie für die vielen schönen gemeinsamen Erlebnisse danken. Meiner Schwester *Laura* und meinem Schwager *Maximilian* danke ich für die tiefe freundschaftliche Verbundenheit und die Gewissheit, immer füreinander da zu sein.

Undoubtedly, there are many others who also deserve acknowledgment at this point, even if they are not explicitly mentioned here.

Thank you all for being an integral part of this significant chapter in my journey through life. Your contributions, whether large or small, have left an indelible mark on this work, and I am profoundly grateful.

Contents

Symbols

Symbol	Meaning	Page
$\mathcal{X}, \mathcal{X}_1, \mathcal{X}_2, \mathcal{Y}$	normed complex Banach spaces	
$\|\cdot\|$, $\|\cdot\|_{\mathcal{X}}$	norm (on $\mathcal{X}$)	
$\mathcal{X}'$	dual space of $\mathcal{X}$	
$\mathcal{H}, \mathcal{H}_1, \mathcal{H}_2,$ $\mathcal{V}, \mathcal{V}_1, \mathcal{W}$	normed complex Hilbert spaces	
$\langle\cdot,\cdot\rangle$, $\langle\cdot,\cdot\rangle_{\mathcal{H}}$	inner product (on $\mathcal{H}$)	
$\mathcal{U}_n, \mathcal{U}_{1,n}, \mathcal{U}_{2,n}$	complex spaces of dimension $n < \infty$	
$\mathcal{L}(\mathcal{X}, \mathcal{Y})$	space of linear and bounded operators mapping from $\mathcal{X}$ to $\mathcal{Y}$	
$\mathcal{L}(\mathcal{X})$	$\mathcal{L}(\mathcal{X}, \mathcal{X})$	
$\mathcal{D}(A)$	domain of the operator A	
$\mathcal{R}(A)$	$\{Ax \,\vert\, x \in \mathcal{D}(A)\}$ range of the operator A	
$\mathcal{G}(A)$	$\{(x, Ax) \,\vert\, x \in \mathcal{D}(A)\}$ graph of the operator A	
$\mathcal{B}(S)$	Borel σ-algebra on S	
$\sigma, \sigma_1, \sigma_2$	normalized surface measures	
$\mathbb{E}M$	expected value of M	
$\mathcal{U}(d)$	unitary group of degree d	
μ	normalized Haar measure on $\mathcal{U}(d)$	
$\Re\lambda$	real part of $\lambda \in \mathbb{C}$	
$\Im\lambda$	imaginary part of $\lambda \in \mathbb{C}$	
$\mathbb{C}_+$	$\{\lambda \in \mathbb{C} \,\vert\, \Re\lambda > 0\}$ open right half plane in $\mathbb{C}$	
$\mathbb{C}_-$	$\{\lambda \in \mathbb{C} \,\vert\, \Re\lambda < 0\}$ open left half plane in $\mathbb{C}$	
$\mathrm{B}_\varepsilon(x_0)$	$\{x \in \mathcal{X} \,\vert\, \|x - x_0\|_{\mathcal{X}} < \varepsilon\}$ ball with radius ε and center $x_0 \in \mathcal{X}$	

Symbol	Meaning	Page
∂S	$\{s \in \overline{S} \mid \forall \varepsilon > 0 \ \exists t \notin S$ such that $t \in \mathrm{B}_\varepsilon(s)\}$ boundary of the set S	
S^{-1}	$\{s^{-1} \mid s \in S \setminus \{0\}\}$ inverse of the set S	
S^*	$\{\overline{s} \mid s \in S\}$ complex conjugate of the set $S \subset \mathbb{C}$	
$\operatorname{dist}(\lambda, S)$	$\inf_{s\in S} \lVert\lambda - s\rVert$ distance of the point λ to the set S	
$\operatorname{dist}(K, L)$	asymmetric distance between the sets K and L	9
$\mathrm{d_H}(K, L)$	Hausdorff-distance between the sets K and L	9
$\varrho(A)$	resolvent set of the operator A	1
$R(A, \lambda)$	resolvent operator $(A - \lambda I)^{-1}$	1
$\sigma(A)$	spectrum of the operator A	1
$r(A)$	spectral radius of the operator A	3
$\sigma_\mathrm{p}(A)$	point spectrum of the operator A	10
$\sigma_\mathrm{c}(A)$	continuous spectrum of the operator A	10
$\sigma_\mathrm{r}(A)$	residual spectrum of the operator A	10
$\sigma_\mathrm{app}(A)$	approximate point spectrum of the operator A	11
$\sigma_\varepsilon(A)$	ε-pseudospectrum of the operator A	13
$W(A)$	numerical range of the operator A	16
$w(A)$	$\sup_{\lambda\in W(A)} \lvert\lambda\rvert$ numerical radius of the operator A	
$W^2(\mathscr{A})$	quadratic numerical range of the block operator matrix $\mathscr{A}$	21
$S_{\mathcal{H}_i}$	$\{x \in \mathcal{H}_i \mid \lVert x\rVert = 1\}$ unit sphere in $\mathcal{H}_i$	21
$\langle\cdot,\cdot\rangle_\mathrm{F}$	Frobenius inner product	59
Sec_ω	$\{z \in \mathbb{C} \mid z \neq 0$ and $\lvert\arg z\rvert < \omega\}$ sector	76
$S(\lambda)$	Schur complement	5,76
a_x	$\langle Ax, x\rangle$ with $x \in S_{\mathcal{H}_1}$	85
$b_{y,x}$	$\langle By, x\rangle$ with $(x, y) \in S_{\mathcal{H}_1} \times S_{\mathcal{H}_2}$	85
$c_{x,y}$	$\langle Cx, y\rangle$ with $(x, y) \in S_{\mathcal{H}_1} \times S_{\mathcal{H}_2}$	85
d_y	$\langle Dy, y\rangle$ with $y \in S_{\mathcal{H}_2}$	85
$M_{x,y}$	$\begin{bmatrix} a_x & b_{y,x} \\ c_{x,y} & d_y \end{bmatrix} \in \mathbb{C}^{2\times 2}$	21,85
a_φ	$\langle A\varphi, \varphi\rangle$ mapping from interval to $\mathbb{C}$	86
$b_{\psi,\varphi}$	$\langle B\psi, \varphi\rangle$ mapping from interval to $\mathbb{C}$	86
$c_{\varphi,\psi}$	$\langle C\varphi, \psi\rangle$ mapping from interval to $\mathbb{C}$	86
d_ψ	$\langle D\psi, \psi\rangle$ mapping from interval to $\mathbb{C}$	86

Symbol	**Meaning**	**Page**
$M_{\varphi,\psi}$	$\begin{bmatrix} a_{\varphi} & b_{\psi,\varphi} \\ c_{\varphi,\psi} & d_{\psi} \end{bmatrix}$ mapping from interval to $\mathbb{C}^{2\times 2}$	86
f_{α,λ_0}	objective function	89
$\lambda_{x,y}^{(\alpha)}$	specific eigenvalue of $M_{x,y}$	89

Introduction

The consideration of the spectra of operators is an elegant means to transform an intricate and abstract object like an operator in a Banach space into a subset of the complex plane which is much easier to grasp and can be handled more intuitively. Theoretical significance abounds in this concept, as the analysis of the spectrum yields valuable insights into the properties of the corresponding operator. This leads to a wide range of applications across diverse domains where spectral theory has had far-reaching impacts and remains a dynamic field of ongoing research.

However, the exact determination of an operator's spectrum, whether through analytical or computational methods, is only possible in rare cases. Moreover, the susceptibility of the spectrum to perturbations poses a substantial challenge, introducing uncertainties from model variations, approximations, and computational errors.

In this thesis, we will explore various approaches aimed to address these challenges. Our primary tools are spectral supersets which provide more opportunities for numerical computation strategies and / or exhibit greater resilience with regard to perturbations. These supersets navigate a fine line, as they still need to be precise enough to offer insights into key properties of the operator under consideration.

One way to construct a perturbation-resistant superset of the spectrum of an operator A is to consider the union of the spectra of slightly perturbed versions of A. The resulting set is called the ε-pseudospectrum defined as

$$\sigma_\varepsilon(A) = \bigcup_{\|P\|<\varepsilon} \sigma(A+P)$$

for some $\varepsilon > 0$. See [51] for an in-depth treatment of pseudospectra, their applications and many examples. The numerical computation of $\sigma_\varepsilon(A)$ has been intensively studied in the matrix case. However, for infinite-dimensional operators the usual procedure is to compute the pseudospectra of finite-dimensional approximations but so far convergence properties remain unproven in the general case.

Our approach, considering Hilbert spaces, is based on the idea to establish an enclosure of the pseudospectrum using the well-known numerical range

$$W(A) = \{\langle Ax, x\rangle \mid x \in \mathcal{D}(A),\ \|x\| = 1\},$$

see [23] for an overview. More precisely, we utilize the numerical ranges of the inverses of shifted versions of A and first show that

$$\sigma_\varepsilon(A) \subset \bigcap_{s \in S} \left[\left(\mathrm{B}_{\delta_s}(W((A-s)^{-1})) \right)^{-1} + s \right] \tag{1}$$

for $\varepsilon > 0$ and $S \subset \varrho(A)$. Furthermore, we prove that, depending on the choice of S, this enclosure can be remarkably precise in the sense that it is contained in the closure of the pseudospectrum under optimal selection of shifts.

To obtain a numerically computable superset of $\sigma_\varepsilon(A)$, we refine our approach further. We demonstrate that it suffices to calculate the numerical ranges of approximating matrices, i.e.

$$\sigma_\varepsilon(A) \subset \bigcap_{s \in S} \left[\left(\mathrm{B}_{\delta_s}(W((A_n-s)^{-1})) \right)^{-1} + s \right] \tag{2}$$

for n sufficiently large, where $(A_n)_n$ is a sequence of matrices which approximates the operator A strongly and S is a finite subset of $\varrho(A)$. At this point, we can leverage the existence of highly effective algorithms in the literature for computing numerical ranges of matrices.

The introduction of a decomposition $\mathcal{H} = \mathcal{H}_1 \oplus \mathcal{H}_2$ of a Hilbert space $\mathcal{H}$ gives the opportunity to take the concept of the numerical range one step further. In this scenario, every bounded operator $\mathscr{A}\colon \mathcal{H} \to \mathcal{H}$ can be expressed as a block operator matrix of the form

$$\mathscr{A} = \begin{bmatrix} A & B \\ C & D \end{bmatrix}$$

where $A\colon \mathcal{H}_1 \to \mathcal{H}_1$, $B\colon \mathcal{H}_2 \to \mathcal{H}_1$, $C\colon \mathcal{H}_1 \to \mathcal{H}_2$, $D\colon \mathcal{H}_2 \to \mathcal{H}_2$ are bounded operators. With this at hand, the quadratic numerical range (QNR) is defined by

$$W^2(\mathscr{A}) = \bigcup_{x \in S_{\mathcal{H}_1}, y \in S_{\mathcal{H}_2}} \sigma\left(\begin{bmatrix} \langle Ax, x\rangle & \langle By, x\rangle \\ \langle Cx, y\rangle & \langle Dy, y\rangle \end{bmatrix} \right),$$

where $S_{\mathcal{H}_i} = \{x \in \mathcal{H}_i \mid \|x\| = 1\}$, $i = 1, 2$. The monograph [52] provides a detailed overview of many properties of the QNR and in [45] and [46], approximation schemes for unbounded operators are established and convergence theorems are proven relating the QNR of an operator to the QNR of its finite-dimensional discretizations.

However, to the best of our knowledge, there are no effective algorithms for numerical computation available even for the matrix case prior to this work. The prevailing method relies on random vector sampling, which is not only very expensive but also tends to yield bad results, particularly for higher-dimensional matrices.

We present a novel computational method that achieves superior results in less time. The key innovation lies in the strategic selection of the utilized vector pairs $(x, y) \in S_{\mathcal{H}_1} \times S_{\mathcal{H}_2}$, resulting in a point cloud that represents a high-quality approximation of the image of the actual QNR, even with a relatively small set of vectors. This selection process places particular emphasis on those (x, y)

that correspond to points near the boundary of the QNR. At the heart of this methodology is the maximization of a purposefully tailored objective function. This enables us to seek the boundary of the quadratic numerical range when moving from an interior point in a specified direction. With the application of speedup techniques, we demonstrate through various examples that the resulting algorithm outperforms the random vector method in both quality of the produced images and computational speed.

This thesis unfolds as follows: Chapter 1 lays the groundwork with known results about the spectral theory for unbounded operators. We present the contents in a manner tailored to meet the requirements of the subsequent chapters. Alongside explanations of the fundamental principles, we place particular emphasis on the theory of spectral perturbations, certain spectral subsets, and the spectra of compact operators.

Moving forward to Chapter 2, we delve deeper into the three aforementioned spectral supersets that underlie the entirety of this thesis. For each of these sets, we start by outlining their basic properties and especially their relation to the spectrum. Subsequently, we provide an overview of the current state-of-the-art techniques for their numerical computation, summarizing the key principles of the algorithms available in the existing literature. Each of the three sections ends with a consistent example, enabling us to make meaningful comparisons between the spectral supersets.

In Chapter 3, we expand upon the content presented in the article [18]. This chapter details the process of obtaining the enclosure (1) and specifies how it can be designed in an optimal way. Additionally, we explore various approximation schemes enabling the derivation of a computable enclosure (2) for the pseudospectrum of an infinite-dimensional operator that is expressed in terms of finite-dimensional matrices. We commence with a more general strong approximation scheme, progressing to a uniform approximation scheme that provides an additional estimate for the starting index n_0 at which (2) holds on bounded sets. The requirements for both of these schemes are satisfied for example by finite element discretizations as we show. In subsequent sections, we study two classes of structured block operator matrices and explain how to derive strong approximations for them. The obtained results are applied to the advection-diffusion operator, the Hain-Lüst operator and a Stokes-type operator and then discussed on the basis of the plotted enclosures. Afterwards, in the concluding section of this chapter, we explore the relationship between the pseudospectra of a third class of structured block operator matrices and the pseudospectra of their Schur complements.

Moving on to the final Chapter 4, which builds upon the article [30], we focus on the development of the novel algorithm for computing the quadratic numerical range of matrices. Beginning with foundational analytical considerations, we establish conditions under which segments of the QNR can be expressed through explicit formulas. We also delve into the analysis of curves and their derivatives within the QNR. Subsequently, we outline our approach for seeking the boundary through the maximization of an objective function and describe how this function can be chosen and why it needs to be adapted to the size and shape of the QNR in order to deal with non-convex regions. We support our method by proving that this adaptation can in in fact result in the obtainment of a given boundary point up to a small error whenever a weak regularity condition is satisfied. The resulting algorithm at this point enables us to progress towards the boundary of the quadratic numerical range after specifying an arbitrary point in the interior and an arbitrary search direction. Building upon this foundation, we then develop an algorithm that computes an approximation of the QNR of any given matrix while automatically adapting the objective function in the process. Additionally, we introduce a speedup technique that significantly reduces computational effort. Multiple examples illustrate the high efficacy of this approach, particularly in comparison to the random vector sampling method. We conclude this chapter with a section devoted to the question why random vector sampling yields suboptimal results for higher-dimensional matrices. We show that the probability that a point in the QNR, determined by the random vector method, lies outside of a small neighborhood of the expected value decreases exponentially with an increase in the dimension of the matrix if its norm remains constant.

Chapter 1

The Spectra of Operators

Knowledge about the eigenvalues of a matrix or more generally about the spectrum of a linear operator is a powerful tool in analysis and numerics and is of great importance in many different areas. These range from more abstract problems in functional analysis and control theory or stability analysis of linear dynamical systems to applied settings in mechanics, chemistry, physics, economics and ecology.

In this chapter, we will outline the basis of the theory surrounding the spectra of operators. Our focus is on the aspects that will be most relevant to the scope of this thesis, like the influence of perturbations, certain spectral subsets and the special case of compact operators. For more information we refer to [15, 32, 47, 51, 56] where this topic is treated on a much broader scale.

1.1 The Spectra of Unbounded Operators

In linear algebra, a point $\lambda \in \mathbb{C}$ is said to be an eigenvalue of a square matrix M if $M - \lambda I$ is not injective (or equivalently not surjective) where I is the identity matrix. In this section we give a brief overview on how the concept of eigenvalues can be extended to the concept of spectra of potentially unbounded linear operators $A \colon \mathcal{D}(A) \subset \mathcal{X} \to \mathcal{X}$ acting on a potentially infinite dimensional complex Banach space $\mathcal{X}$.

Let I denote the identity on $\mathcal{X}$.

Definition 1.1.1. a) The *resolvent set* $\varrho(A)$ is the set of all $\lambda \in \mathbb{C}$ for which $A - \lambda I$ is bijective and the *resolvent operator*

$$R(A, \lambda) := (A - \lambda I)^{-1}$$

is bounded.

b) The *spectrum* of A is defined by

$$\sigma(A) = \mathbb{C} \setminus \varrho(A),$$

i.e. it is the set of all $\lambda \in \mathbb{C}$ for which $A - \lambda I$ does not have a bounded inverse.

In the following we will abbreviate an expression like $A - \lambda I$ by $A - \lambda$ whenever its meaning is clear from the context.

It is immediate from the definition that

$$\sigma(\alpha + \beta A) = \alpha + \beta\sigma(A) \tag{1.1}$$

for $\alpha, \beta \in \mathbb{C}$.

The concept of spectra is of interest only, if the operator under consideration is closed, i.e. if its graph $\mathcal{G}(A) := \{(x, Ax) \mid x \in \mathcal{D}(A)\}$ is a closed subset of $\mathcal{X} \times \mathcal{X}$. This is because if A is not closed, we always have $\sigma(A) = \mathbb{C}$ due to the fact that in this case none of the $A - \lambda$ with $\lambda \in \mathbb{C}$ is closed either and if $A - \lambda$ is invertible, we have

$$\mathcal{G}\big((A - \lambda)^{-1}\big) = \{((A - \lambda)x, x) \mid x \in \mathcal{D}(A)\}$$

and thus $(A - \lambda)^{-1}$ is not closed too and therefore not bounded. On the other hand, if A is closed, the boundedness of $(A - \lambda)^{-1}$ follows directly from the surjectivity of $A - \lambda$ by use of the closed graph theorem.

Theorem 1.1.2. *Let $\lambda \in \varrho(A)$. Then the following assertions hold:*

a) $R(A, \lambda) - R(A, \mu) = (\lambda - \mu)R(A, \lambda)R(A, \mu)$ *for all* $\mu \in \varrho(A)$*;*

b) *The function* $R(A, \cdot)\colon \varrho(A) \to \mathcal{L}(\mathcal{X})$*,* $\lambda \mapsto R(A, \lambda)$*, is analytic;*

c) $\|R(A, \lambda)\| \geq \frac{1}{\operatorname{dist}(\lambda, \sigma(A))}$.

Proof. a) From the definition of the resolvent we see that

$$x = (A - \lambda)R(A, \lambda)x = R(A, \lambda)(A - \lambda)x$$

holds for every $x \in \mathcal{D}(A)$. In other words, A and $R(A, \lambda)$ commute on the domain of A. Furthermore, we have

$$\begin{aligned} R(A, \lambda) &= R(A, \lambda)(AR(A, \mu) - \mu R(A, \mu)), \\ R(A, \mu) &= (AR(A, \lambda) - \lambda R(A, \lambda))R(A, \mu) \end{aligned}$$

also by definition of the resolvent. Subtracting these two equations and using the commutation property just mentioned yields the desired resolvent equation.

b), c) Let $\mu \in \mathbb{C}$ with $|\lambda - \mu| \leq \frac{\delta}{\|R(A,\lambda)\|}$ for some $\delta \in (0, 1)$. We will show that $\mu \in \varrho(A)$ and

$$R_\mu := \sum_{n=0}^{\infty} (\mu - \lambda)^n R(A, \lambda)^{n+1} \tag{1.2}$$

coincides with $R(A, \mu)$. For $x \in \mathcal{X}$ we have

$$\begin{aligned} \|(\mu - \lambda)^n R(A, \lambda)^{n+1} x\| &\leq \frac{\delta^n}{\|R(A, \lambda)\|^n} \|R(A, \lambda)\|^{n+1} \|x\| \\ &= \delta^n \|R(A, \lambda)\| \|x\| \end{aligned}$$

and thus the series (1.2) converges with $\|R_\mu\| \le \frac{\|R(A,\lambda)\|}{1-\delta}$. By using a) we have

$$(A-\mu)R(A,\lambda) = (\lambda-\mu)R(A,\lambda) + I.$$

This can be used to obtain

$$\begin{aligned}(A-\mu)R_\mu &= -\sum_{n=0}^{\infty}(\mu-\lambda)^{n+1}R(A,\lambda)^{n+1} + \sum_{n=0}^{\infty}(\mu-\lambda)^{n}R(A,\lambda)^{n}\\ &= I\end{aligned}$$

and similarly

$$R_\mu(A-\mu)x = x$$

for all $x \in \mathcal{D}(A)$, which concludes the proof. ❑

Corollary 1.1.3. *$\sigma(A)$ is a closed subset of $\mathbb{C}$.*

Proof. $\varrho(A)$ is open because by Theorem 1.1.2 c) we know that for every $\lambda \in \varrho(A)$ the open ball around λ with radius $1/\|R(A,\lambda)\|$ is contained in $\varrho(A)$. ❑

Theorem 1.1.4. *If A is bounded, then $\sigma(A)$ is compact. More precisely we have*

$$\max\big\{|\lambda| \,\big|\, \lambda \in \sigma(A)\big\} = \lim_{n\to\infty}\big\|A^n\big\|^{1/n} = \inf_{n\in\mathbb{N}}\big\|A^n\big\|^{1/n} \le \|A\|.$$

$r(A) := \inf_{n\in\mathbb{N}}\big\|A^n\big\|^{1/n}$ is called the spectral radius *of A.*

The proof of Theorem 1.1.4 requires the following lemma about sequences in $\mathbb{R}$.

Lemma 1.1.5. *Let $(a_n)_n \subset \mathbb{R}$ be a sequence that satisfies $0 \le a_{n+m} \le a_n a_m$ for all $m, n \in \mathbb{N}$. Then*

$$\lim_{n\to\infty}\sqrt[n]{a_n} = \inf_{n\in\mathbb{N}}\sqrt[n]{a_n} =: a.$$

Proof. Let $\varepsilon > 0$ and choose $N \in \mathbb{N}$ such that $\sqrt[N]{a_N} < a + \varepsilon$. Let $b := \max\{a_1, \dots, a_N\}$ and write $n \in \mathbb{N}$ in the form $n = kN + r$ with $k \in \mathbb{N}$ and $1 \le r \le N$. Then we have

$$\begin{aligned}\sqrt[n]{a_n} = a_{kN+r}^{1/n} &\le \big(a_N^k a_r\big)^{1/n} \le (a+\varepsilon)^{kN/n} b^{1/n}\\ &= (a+\varepsilon)(a+\varepsilon)^{-r/n} b^{1/n} \le a + 2\varepsilon\end{aligned}$$

for n large enough. ❑

Proof of Theorem 1.1.4. By choosing $a_n = \big\|A^n\big\|$ in Lemma 1.1.5 we have that $\big\|A^n\big\|^{1/n}$ converges to $r(A) = \inf_{n\in\mathbb{N}}\|A^n\|^{1/n}$ for $n \to \infty$. Let $|\lambda| > r(A)$. Then

$$\limsup_{n\to\infty}\left\|\left(\frac{A}{\lambda}\right)^n\right\|^{1/n} = \lim_{n\to\infty}\frac{\big\|A^n\big\|^{1/n}}{|\lambda|} = \frac{r(A)}{|\lambda|} < 1$$

ensures the convergence of $R_\lambda := -\lambda^{-1}\sum_{n=0}^{\infty}\left(\frac{A}{\lambda}\right)^n$. Moreover,

$$(A-\lambda)R_\lambda = -\sum_{n=0}^{\infty}\lambda^{-(n+1)}A^{n+1} + \sum_{n=0}^{\infty}\lambda^{-n}A^n = I$$

and similarly $R_\lambda(A-\lambda) = I$. Thus, $\lambda \in \varrho(A)$, $R_\lambda = R(A,\lambda)$ and $\sigma(A) \subset \overline{\mathrm{B}_{r(A)}(0)}$. Hence, with Corollary 1.1.3, $\sigma(A)$ is compact and $r_0 := \max\left\{|\lambda| \,\middle|\, \lambda \in \sigma(A)\right\}$ exists with $r_0 \leq r(A)$.

It remains to show $r_0 \geq r(A)$. Let therefore $|\mu| > r_0$. For an arbitrary $l \in \mathcal{L}(\mathcal{X})'$ we consider the function

$$f_l(\lambda) := l\left(R(A,\lambda)\right)$$

which is analytic on $\left\{\lambda \in \mathbb{C} \,\middle|\, |\lambda| > r_0\right\}$ by Theorem 1.1.2 b). As we have already seen, this function can be represented as

$$f_l(\lambda) = -\sum_{n=0}^{\infty} l(A^n)\lambda^{-(n+1)} \tag{1.3}$$

on $\left\{\lambda \in \mathbb{C} \,\middle|\, |\lambda| > r(A)\right\}$. From [11, Theorem V.1.11] we know that the series representation of a complex-analytic function is unique and so the equality (1.3) also holds on the larger set $\left\{\lambda \in \mathbb{C} \,\middle|\, |\lambda| > r_0\right\}$ and in particular in μ. Due to the convergence we conclude

$$\lim_{n\to\infty} l\left(A^n\mu^{-(n+1)}\right) = 0$$

which makes $\left(A^n\mu^{-(n+1)}\right)$ a weak null-sequence since $l \in \mathcal{L}(\mathcal{X})'$ was chosen arbitrarily. Hence, $\left(A^n\mu^{-(n+1)}\right)$ is bounded therefore there exists a $K > 0$ such that

$$\left\|A^n\right\|^{1/n} \leq K^{1/n}|\mu|^{(n+1)/n} \to |\mu|.$$

Thus, $|\mu| \geq r(A)$ for every $|\mu| > r_0$ which implies $r_0 \geq r(A)$. ❑

Theorem 1.1.6. *Let $\mathcal{X}$ be a Hilbert space with inner product $\langle\cdot,\cdot\rangle$ and $A \in \mathcal{L}(\mathcal{X})$ normal, i.e. $A^*A = AA^*$. Then we have*

$$r(A) = \|A\|.$$

Proof. We have $\|A^*A\| \leq \|A^*\|\|A\| = \|A\|^2$ and

$$\|Ax\|^2 = \langle Ax, Ax\rangle = \langle A^*Ax, x\rangle \leq \|A^*Ax\|$$

for every $x \in \mathcal{X}$ with $\|x\| = 1$. Hence, $\|A^*A\| = \|A\|^2$ and thus also $\left\|(A^*A)^2\right\| = \|A^*A\|^2$. With this at hand we see

$$\left\|A^2\right\|^2 = \left\|\left(A^2\right)^* A^2\right\| = \left\|(A^*A)^2\right\| = \|A^*A\|^2 = \left(\|A\|^2\right)^2,$$

i.e. $\left\|A^2\right\| = \|A\|^2$, and because A^n is normal for every $n \in \mathbb{N}$ as well we have $\|A\|^{2^k} = \left\|A^{2^k}\right\|$ for all $k \in \mathbb{N}$. Hence,

$$r(A) = \lim_{n\to\infty}\left\|A^n\right\|^{1/n} = \lim_{k\to\infty}\left\|A^{2^k}\right\|^{1/2^k} = \|A\|.$$ ❑

Lemma 1.1.7. *Let A be closed and $\lambda \in \varrho(A)$. Then we have*

$$\sigma(R(A,\lambda)) \setminus \{0\} = (\sigma(A) - \lambda)^{-1}.$$

Proof. For a $\mu \in \mathbb{C} \setminus \{0\}$ we have

$$R(A,\lambda) - \mu = -\big(A - (\lambda + \tfrac{1}{\mu})\big)\mu R(A,\lambda).$$

Note that $R(A,\lambda)\colon \mathcal{X} \to \mathcal{D}(A)$ is bijective and therefore $R(A,\lambda) - \mu\colon \mathcal{X} \to \mathcal{X}$ is bijective if and only if $A - (\lambda + \frac{1}{\mu})\colon \mathcal{D}(A) \to \mathcal{X}$ is bijective. Hence, we have $\mu \in \varrho(R(A,\lambda))$ if and only if $\lambda + \frac{1}{\mu} \in \varrho(A)$ if and only if $\mu = (\nu - \lambda)^{-1}$ for some $\nu \in \varrho(A)$. ❑

Corollary 1.1.8. *Let $\mathcal{X}$ be a Hilbert space, $A \in \mathcal{L}(\mathcal{X})$ normal and $\lambda \in \varrho(A)$. Then we have*

$$\|R(A,\lambda)\| = \frac{1}{\operatorname{dist}(\lambda, \sigma(A))}.$$

Proof. Since $R(A,\lambda)$ is normal as well this is a direct consequence of Theorem 1.1.6 and Lemma 1.1.7 because

$$\begin{aligned}\|R(A,\lambda)\| &= \max\big\{|\mu| \,\big|\, \mu \in \sigma(R(A,\lambda))\big\}\\ &= \max\left\{\frac{1}{|\mu - \lambda|} \,\middle|\, \mu \in \sigma(A)\right\}\\ &= \frac{1}{\operatorname{dist}(\lambda, \sigma(A))}.\end{aligned}$$ ❑

We conclude this section with the consideration of block operator matrices. The following results can be found in [52] and we also include the proofs here for convenience of the reader. Let $\mathcal{X}_1$ and $\mathcal{X}_2$ be Banach spaces such that $\mathcal{X} = \mathcal{X}_1 \times \mathcal{X}_2$.

Definition 1.1.9. a) An operator $\mathscr{A}\colon \mathcal{D}(\mathscr{A}) \subset \mathcal{X} \to \mathcal{X}$ that can be written in the form

$$\mathscr{A} = \begin{bmatrix} A & B \\ C & D \end{bmatrix}$$

with linear operators $A\colon \mathcal{D}(A) \subset \mathcal{X}_1 \to \mathcal{X}_1$, $B\colon \mathcal{D}(B) \subset \mathcal{X}_2 \to \mathcal{X}_1$, $C\colon \mathcal{D}(C) \subset \mathcal{X}_1 \to \mathcal{X}_2$ and $D\colon \mathcal{D}(D) \subset \mathcal{X}_2 \to \mathcal{X}_2$ such that

$$\mathcal{D}(\mathscr{A}) = \big(\mathcal{D}(A) \cap \mathcal{D}(C)\big) \times \big(\mathcal{D}(B) \cap \mathcal{D}(D)\big)$$

is called *block operator matrix*;

b) Let $\mathscr{A}$ be a block operator matrix. Then the operator function S defined by

$$S(\lambda) = D - \lambda - C(A - \lambda)^{-1}B \quad \text{for} \quad \lambda \in \varrho(A)$$

is called *Schur complement* of $\mathscr{A}$;

c) $\sigma(S) := \{\lambda \in \varrho(A) \,|\, 0 \in \sigma(S(\lambda))\}$;

d) $\varrho(S) := \{\lambda \in \varrho(A) \,|\, 0 \in \varrho(S(\lambda))\}$.

Such a block operator matrix does not need to be closed or closable even if we assume all of its entries A, B, C and D to be closed. It is therefore natural to ask under which conditions $\mathscr{A}$ is closed or closable and how this closure can be obtained.

Theorem 1.1.10. *Let $\mathscr{A}$ be a block operator matrix and assume that C is closable, $\mathcal{D}(A) \subset \mathcal{D}(C)$, $\varrho(A) \neq \varnothing$, $\mathcal{D}(B)$ is a dense subset of $\mathcal{X}_2$ and that for some (and hence for all) $\lambda \in \varrho(A)$, the operator $(A-\lambda)^{-1}B$ is bounded on $\mathcal{D}(B)$. Then $\mathscr{A}$ is closable (closed, respectively) if and only if $S(\lambda)$ is closable (closed, respectively) for some (and hence for all) $\lambda \in \varrho(A)$. In this case, the closure $\overline{\mathscr{A}}$ is given by*

$$\overline{\mathscr{A}} = \lambda + \begin{bmatrix} I & 0 \\ C(A-\lambda)^{-1} & I \end{bmatrix} \begin{bmatrix} A-\lambda & 0 \\ 0 & \overline{S(\lambda)} \end{bmatrix} \begin{bmatrix} I & \overline{(A-\lambda)^{-1}B} \\ 0 & I \end{bmatrix} \tag{1.4}$$

independently of $\lambda \in \varrho(A)$ where

$$\mathcal{D}(\overline{\mathscr{A}}) = \left\{ \begin{bmatrix} x \\ y \end{bmatrix} \in \mathcal{X}_1 \times \mathcal{X}_2 \,\middle|\, x + \overline{(A-\lambda)^{-1}B}y \in \mathcal{D}(A),\ y \in \mathcal{D}(\overline{S(\lambda)}) \right\}.$$

Proof. We start by noting that the assumptions on $(A-\lambda)^{-1}B$ and $S(\lambda)$ do not depend on the choice of $\lambda \in \varrho(A)$. This is because from the resolvent identity Theorem 1.1.2 a) we get that for $\lambda_0 \in \varrho(A)$ the differences

$$\begin{gathered} (A-\lambda_0)^{-1}B - (A-\lambda)^{-1}B = (\lambda_0 - \lambda)(A-\lambda_0)^{-1}(A-\lambda)^{-1}B, \\ S(\lambda_0) - S(\lambda) = -(\lambda_0 - \lambda)(I + C(A-\lambda_0)^{-1}(A-\lambda)^{-1}B) \end{gathered}$$

are bounded. Next, $(\mathscr{A} - \lambda)$ can be written in the form

$$\mathscr{A} - \lambda = \begin{bmatrix} I & 0 \\ C(A-\lambda)^{-1} & I \end{bmatrix} \begin{bmatrix} A-\lambda & 0 \\ 0 & S(\lambda) \end{bmatrix} \begin{bmatrix} I & (A-\lambda)^{-1}B \\ 0 & I \end{bmatrix} \tag{1.5}$$

where $(A-\lambda)^{-1}B = \overline{(A-\lambda)^{-1}B}|_{\mathcal{D}(B)}$ can be replaced by $\overline{(A-\lambda)^{-1}B}$ because for the domain of the middle factor we have

$$\mathcal{D}(A) \times \mathcal{D}(S(\lambda)) = \mathcal{D}(A) \times \big(\mathcal{D}(B) \cap \mathcal{D}(D)\big) \subset \mathcal{D}(A) \times \mathcal{D}(B).$$

With this, the first and last factor in (1.5) are bounded and boundedly invertible in $\mathcal{X}_1 \times \mathcal{X}_2$. Therefore, $\mathscr{A} - \lambda$ is closable (closed, respectively) if and only if the middle factor is. Now since $\varrho(A) \neq \varnothing$ implies that A is closed, we have that $\mathscr{A} - \lambda$ is closable (closed, respectively) if and only if $S(\lambda)$ is and the closure of $\mathscr{A} - \lambda$ is then given by taking the closure of the middle factor in (1.5). Lastly, the independence of $\overline{\mathscr{A}}$ on λ implies the independence of the right hand side of (1.4) on λ. ❑

Corollary 1.1.11. *Under the assumptions of Theorem 1.1.10 we have*

$$\sigma(\overline{\mathscr{A}}) \setminus \sigma(A) = \sigma(\overline{S})$$

and, for $\lambda \in \varrho(\overline{S}) = \varrho(\overline{\mathscr{A}}) \cap \varrho(A) \subset \varrho(\overline{\mathscr{A}})$,

$$\begin{aligned}
&(\overline{\mathscr{A}} - \lambda)^{-1} \\
&\quad = \begin{bmatrix} I & -\overline{(A-\lambda)^{-1}B} \\ 0 & I \end{bmatrix} \begin{bmatrix} (A-\lambda)^{-1} & 0 \\ 0 & \overline{S(\lambda)}^{-1} \end{bmatrix} \begin{bmatrix} I & 0 \\ -C(A-\lambda)^{-1} & I \end{bmatrix} \\
&\quad = \begin{bmatrix} \big((\overline{\mathscr{A}}-\lambda)^{-1}\big)_1 & \big((\overline{\mathscr{A}}-\lambda)^{-1}\big)_2 \\ \big((\overline{\mathscr{A}}-\lambda)^{-1}\big)_3 & \big((\overline{\mathscr{A}}-\lambda)^{-1}\big)_4 \end{bmatrix},
\end{aligned}$$

where

$$\begin{aligned}
\big((\overline{\mathscr{A}}-\lambda)^{-1}\big)_1 &= (A-\lambda)^{-1} + \overline{(A-\lambda)^{-1}B}\,\overline{S(\lambda)}^{-1} C(A-\lambda)^{-1}, \\
\big((\overline{\mathscr{A}}-\lambda)^{-1}\big)_2 &= -\overline{(A-\lambda)^{-1}B}\,\overline{S(\lambda)}^{-1}, \\
\big((\overline{\mathscr{A}}-\lambda)^{-1}\big)_3 &= -\overline{S(\lambda)}^{-1} C(A-\lambda)^{-1}, \\
\big((\overline{\mathscr{A}}-\lambda)^{-1}\big)_4 &= \overline{S(\lambda)}^{-1}.
\end{aligned}$$

Proof. This is a direct consequence of Theorem 1.1.10 because the first and last factor in the right hand side of (1.4) are bounded and boundedly invertible in $\mathcal{X}_1 \times \mathcal{X}_2$. ❑

In [52], results like Theorem 1.1.10 and Corollary 1.1.11 are also given for differently structured block operator matrices but the versions stated here suffice for the scope of this thesis.

1.2 The Spectra of Perturbed Operators

Let us come back to the consideration of general linear operators $A\colon \mathcal{D}(A) \subset \mathcal{X} \to \mathcal{X}$ and study the influence of perturbations of A on the spectrum $\sigma(A)$.

Theorem 1.2.1. *Let* $\lambda \in \varrho(A)$. *Then the following assertions hold:*

a) $\lambda \in \varrho(A+P)$ *for any* $P \in \mathcal{L}(\mathcal{X})$ *that satisfies* $\|P\| < \frac{1}{\|R(A,\lambda)\|}$;

b) *Conversely, for any* $\varepsilon > \frac{1}{\|R(A,\lambda)\|}$, *there exists a* $P \in \mathcal{L}(\mathcal{X})$ *with* $\|P\| < \varepsilon$ *such that* $(A+P)x = \lambda x$ *for some* $x \in \mathcal{D}(A)$, $x \neq 0$, *i.e.* $\lambda \in \sigma(A+P)$.

Proof. a) Let $P \in \mathcal{L}(\mathcal{X})$ with $\|P\| < \frac{1}{\|R(A,\lambda)\|}$ and rewrite

$$A + P - \lambda = (I + P(A-\lambda)^{-1})(A-\lambda).$$

Here, $\|-P(A-\lambda)^{-1}\| \leq \|P\|\|(A-\lambda)^{-1}\| < 1$ implies the invertibility of $I+P(A-\lambda)^{-1}$ by a Neumann-series argument. Hence, $A+P-\lambda$ is invertible.

b) Let $\varepsilon > \frac{1}{\|R(A,\lambda)\|}$. Then $\|\varepsilon R(A,\lambda)\| > 1$ and thus there exist $u \in \mathcal{D}(A)$ and $v \in \mathcal{X}$ with $\|u\| > 1$ and $\|v\| = 1$ such that $\varepsilon R(A,\lambda)v = u$ or equivalently

$$R(A,\lambda)\frac{\varepsilon v}{\|u\|} = \frac{u}{\|u\|}.$$

This proves the existence of $x \in \mathcal{D}(A)$ and $y \in \mathcal{X}$ with $\|x\| = 1$ and $\|y\| < \varepsilon$ such that $y = (A-\lambda)x$. Now take $S := \operatorname{span}\{x\}$ and consider the functional $P_S \in S'$ that maps an element of S, i.e. a scalar multiple μx of x, to μ. We have $\|P_S\| = 1$ and by the Hahn-Banach theorem, P_S can be extended to a functional $\widetilde{P}$ on $\mathcal{X}$ with the same norm and $\widetilde{P}|_S = P_S$. Multiplying the scalar value of $\widetilde{P}$ to $-y$ results in an operator $P \in \mathcal{L}(\mathcal{X})$ with $\|P\| = \|y\| < \varepsilon$ such that $(A+P)x = \lambda x$. ❑

The results of Theorem 1.2.1 are also referred to as the upper semicontinuity of the spectrum. Roughly speaking, it says that a small perturbation of A can only result in a small enlargement of $\sigma(A)$. However, the spectrum can be lower simidiscontinuous in general as the following example from [32, p. 210] shows. Here, the spectrum of A shrinks suddenly when it is subject to a small perturbation.

Example 1.2.2. Let $\mathcal{X} = \ell^p(\mathbb{Z})$, $p \le 1 \le \infty$, and let $\{x_n\}_{n\in\mathbb{Z}}$ be the canonical basis on $\mathcal{X}$, i.e. $x_n = (\delta_{nj})_j$. Define $A \in \mathcal{L}(\mathcal{X})$ via $Ax_0 = 0$ and $Ax_n = x_{n-1}$ for $n \neq 0$. Then $\|A\| = 1$ and thus $\sigma(A)$ is a subset of the unit disk in $\mathbb{C}$ by Theorem 1.1.4. Due to the fact that for any $\lambda \in \mathbb{C}$ with $|\lambda| < 1$ the vector $u := \sum_{n=0}^{\infty} \lambda^n x_n$ satisfies $(A-\lambda)u = 0$ and because $\sigma(A)$ is closed, we have equality, namely

$$\sigma(A) = \big\{\lambda \in \mathbb{C} \,\big|\, |\lambda| \le 1\big\}.$$

Now let $\varepsilon > 0$ and consider the perturbation $P \in \mathcal{L}(\mathcal{X})$ with $\|P\| = \varepsilon$ defined via $Px_0 = \varepsilon x_{-1}$ and $Px_n = 0$ for $n \neq 0$. Then we have $(A+P)^m x_n = \varepsilon x_{n-m}$ for $0 \le n < m$ and $(A+P)^m x_n = x_{n-m}$ for $n \in \mathbb{Z}\setminus[0,m[$, which implies $\|(A+P)^m\| = \max\{1,\varepsilon\}$ for all $m \in \mathbb{N}$. Hence, $r(A+P) = \lim_{m\to\infty}\|(A+P)^m\|^{1/m} = 1$ and thus

$$\sigma(A+P) \subset \big\{\lambda \in \mathbb{C} \,\big|\, |\lambda| \le 1\big\} \tag{1.6}$$

as before. However, $0 \in \varrho(A+P)$, because $A+P$ is invertible with $(A+P)^{-1}x_{-1} = \varepsilon^{-1}x_0$ and $(A+P)^{-1}x_{n-1} = x_n$ for $n \neq 0$. Moreover, we have $\|(A+P)^{-1}\| = \max\{1,\varepsilon^{-1}\}$ for all $m \in \mathbb{N}$, which implies $r\big((A+P)^{-1}\big) = 1$ as above. Thus, $\sigma\big((A+P)^{-1}\big) \subset \big\{\lambda \in \mathbb{C} \,\big|\, |\lambda| \le 1\big\}$ and by Lemma 1.1.7 this and (1.6) yield

$$\sigma(A+P) = \big\{\lambda \in \mathbb{C} \,\big|\, |\lambda| = 1\big\}.$$

Hence, an arbitrarily small perturbation of A can result in the shrinkage of the spectrum from the whole unit disk to the unit circle.

In the matrix case however, this can not happen as the following result shows. Before we we state it, we have to clarify the notion of the distance between two sets in $\mathbb{C}$.

Definition 1.2.3. For two nonempty sets $K, L \subset \mathbb{C}$ we define the *distance* $\operatorname{dist}(K, L)$ via

$$\operatorname{dist}(K, L) = \sup_{k \in K} \left(\inf_{l \in L} \|k - l\| \right)$$

and the *Hausdorff-distance* $\mathrm{d_H}(K, L)$ via

$$\mathrm{d_H}(K, L) = \max \left\{ \operatorname{dist}(K, L), \operatorname{dist}(L, K) \right\}.$$

Note, that $K \subset B_\varepsilon(L)$ whenever $\operatorname{dist}(K, L) < \varepsilon$ and $\operatorname{dist}(K, L) \leq \varepsilon$ whenever $K \subset B_\varepsilon(L)$.

Theorem 1.2.4. *Let $\mathcal{X}$ be finite dimensional. Then $\sigma(A)$ is a continuous function of A in the sense that*

$$\lim_{\|P\| \to 0} \mathrm{d_H}\big(\sigma(A + P), \sigma(A)\big) = 0.$$

Proof. This is [32, Theorem II.5.14]. ❑

However, for some A, $\sigma(A)$ can still be very sensitive with regard to perturbations even though there is a continuous dependence.

Example 1.2.5. Consider the $n \times n$ matrix

$$A = \begin{bmatrix} 1 & 1 & 0 & \dots & 0 \\ 0 & \ddots & \ddots & \ddots & \vdots \\ \vdots & \ddots & \ddots & \ddots & 0 \\ \vdots & & \ddots & \ddots & 1 \\ 0 & \dots & \dots & 0 & 1 \end{bmatrix}$$

with its only eigenvalue $\lambda = 1$ and consider the perturbed version

$$A + P = \begin{bmatrix} 1 & 1 & 0 & \dots & 0 \\ 0 & \ddots & \ddots & \ddots & \vdots \\ \vdots & \ddots & \ddots & \ddots & 0 \\ 0 & & \ddots & \ddots & 1 \\ \varepsilon & 0 & \dots & 0 & 1 \end{bmatrix}$$

with some $\varepsilon > 0$. Then, for $x := \begin{bmatrix} \varepsilon^{-\frac{n-1}{n}} & \varepsilon^{-\frac{n-2}{n}} & \dots & 1 \end{bmatrix}^\mathsf{T}$ we have

$$(A + P)x = \left(1 + \varepsilon^{\frac{1}{n}}\right) x.$$

Therefore, a perturbation of order $\mathcal{O}(\varepsilon)$ produced a perturbation of an eigenvalue of order $\mathcal{O}\big(\varepsilon^{\frac{1}{n}}\big)$.

1.3 Spectral Subsets

The spectrum can be decomposed into different subsets, depending on the reason why $A - \lambda$ is not boundedly invertible.

Definition 1.3.1. a) The *point spectrum* $\sigma_{\mathrm{p}}(A)$ consists of all $\lambda \in \mathbb{C}$ for which $A - \lambda$ is not injective. The elements in $\sigma_{\mathrm{p}}(A)$ are called *eigenvalues* of A.

b) The *continuous spectrum* $\sigma_{\mathrm{c}}(A)$ consists of all $\lambda \in \mathbb{C}$ for which $A - \lambda$ is injective, not surjective and has dense range.

c) The *residual spectrum* $\sigma_{\mathrm{r}}(A)$ consists of all $\lambda \in \mathbb{C}$ for which $A - \lambda$ is injective, not surjective and does not have dense range.

If A is closed, we have $\sigma(A) = \sigma_{\mathrm{p}}(A) \cup \sigma_{\mathrm{c}}(A) \cup \sigma_{\mathrm{r}}(A)$, and if $\dim \mathcal{X} < \infty$, we have $\sigma(A) = \sigma_{\mathrm{p}}(A)$.

Lemma 1.3.2. *Let A be closed and densely defined. Then we have*

$$\sigma_{\mathrm{r}}(A) = \sigma_{\mathrm{p}}(A'),$$

where A' denotes the adjoint of A.

Proof. Let $\lambda \in \mathbb{C}$. By the Hahn-Banach theorem we know that

$$\overline{(A - \lambda)\mathcal{D}(A)} \neq \mathcal{X}$$

if and only if there exists a continuous linear functional $0 \neq x' \in \mathcal{X}'$ with

$$x'((A - \lambda)x) = 0 \tag{1.7}$$

for all $x \in \mathcal{D}(A)$. This is equivalent to $x'(Ax) = \lambda x'(x)$ for all $x \in \mathcal{D}(A)$ which means that $x' \in \mathcal{D}(A')$ and $(A' - \lambda)x' = 0$. ❑

Remark 1.3.3. If $A\colon \mathcal{D}(A) \subset \mathcal{H} \to \mathcal{H}$ is a closed and densely defined operator on a Hilbert space $\mathcal{H}$ with inner product $\langle \cdot, \cdot \rangle$, the result of Lemma 1.3.2 can also be expressed in terms of the Hilbert space adjoint A^* of A, namely

$$\sigma_{\mathrm{r}}(A) = (\sigma_{\mathrm{p}}(A^*))^*.$$

The complex conjugation on the right hand side of this equation stems from the fact that in this case (1.7) can be rewritten as

$$\langle (A - \lambda)x, y \rangle = 0$$

with some $y \in \mathcal{H}$ by the Fréchet-Riesz theorem. Due to the antilinearity of the inner product in the second argument we now obtain $(A^* - \overline{\lambda})y = 0$.

Let $\lambda \in \mathbb{C}$. If there exists a nonzero $x \in \mathcal{D}(A)$ (called *eigenvector*) with $(A - \lambda)x = 0$ we have $\lambda \in \sigma_{\mathrm{p}}(A)$. If we weaken this condition and assume instead that there exists no constant $c > 0$ such that $\|(A - \lambda)x\| \geq c\|x\|$ for all $x \in \mathcal{D}(A)$ then λ is an element of another – potentially larger – subset of $\sigma(A)$ that will be defined next.

Definition 1.3.4. The *approximate point spectrum* $\sigma_{\mathrm{app}}(A)$ is defined as the set of all $\lambda \in \mathbb{C}$ for which there exists a sequence $(x_n)_n \subset \mathcal{D}(A)$ with $\|x_n\| = 1$ for all $n \in \mathbb{N}$ such that $(A - \lambda)x_n$ converges to 0 as $n \to \infty$. The elements in $\sigma_{\mathrm{app}}(A)$ are called *approximate eigenvalues* of A and the corresponding sequences $(x_n)_n$ are the *approximate eigenvectors.*

The spectrum, point spectrum and approximate point spectrum are related by $\sigma_{\mathrm{p}}(A) \subset \sigma_{\mathrm{app}}(A) \subset \sigma(A)$.

Lemma 1.3.5. *We have*

$$\partial\sigma(A) \subset \sigma_{\mathrm{app}}(A),$$

where $\partial\sigma(A)$ denotes the boundary of $\sigma(A)$.

Proof. Let $\lambda \in \partial\sigma(A)$. Then there exists a sequence $(\lambda_n)_n \subset \varrho(A)$ with

$$\lim_{n\to\infty} \lambda_n = \lambda.$$

By Theorem 1.1.2 we have

$$\lim_{n\to\infty} \|R(A, \lambda_n)\| = \infty$$

and therefore there exists a sequence $(y_n)_n \subset \mathcal{X}$ with $\|y_n\| = 1$ for all $n \in \mathbb{N}$ such that

$$\lim_{n\to\infty} \|R(A, \lambda_n)y_n\| = \infty.$$

Let $x_n := \frac{1}{\|R(A,\lambda_n)y_n\|} R(A, \lambda_n)y_n$. Then $\|x_n\| = 1$, $x_n \in D(A)$ for all $n \in \mathbb{N}$ and

$$\lim_{n\to\infty} (A - \lambda)x_n = \lim_{n\to\infty} \left(\frac{1}{\|R(A, \lambda_n)y_n\|} y_n + (\lambda_n - \lambda)x_n \right) = 0.$$

Hence, $\lambda \in \sigma_{\mathrm{app}}(A)$. ❑

Lemma 1.3.6. *Let A be a closed operator. Then the following assertions hold:*

a) $\sigma_{\mathrm{app}}(A) = \sigma_{\mathrm{p}}(A) \cup \{\lambda \in \mathbb{C} \,|\, (A - \lambda)\mathcal{D}(A)$ *is not closed in* $\mathcal{X}\}$*;*

b) $\sigma(A) = \sigma_{\mathrm{app}}(A) \cup \sigma_{\mathrm{r}}(A)$.

Proof. Let $\lambda \notin \sigma_{\mathrm{app}}(A)$. Then there exists a constant $c > 0$ such that

$$\|(A - \lambda)x\| \geq c\|x\| \tag{1.8}$$

for all $x \in \mathcal{D}(A)$. Let $(y_n)_n = ((A - \lambda)x_n)_n$ be a sequence in $(A - \lambda)\mathcal{D}(A)$ with $\lim_{n\to\infty} y_n = y \in \mathcal{X}$. Then the inequality (1.8) implies that $(x_n)_n$ is a Cauchy sequence and therefore we have $\lim_{n\to\infty} x_n = x$ for some $x \in \mathcal{X}$. By the closedness of A we obtain $x \in \mathcal{D}(A)$ and $y = (A - \lambda)x$ and therefore $(A - \lambda)\mathcal{D}(A)$ is closed.

If on the other hand, $(A - \lambda)\mathcal{D}(A)$ is closed and $(A - \lambda)$ is injective, then $(A - \lambda)^{-1}$ exists on $(A - \lambda)\mathcal{D}(A)$ and is closed because A is. Hence, $(A - \lambda)^{-1}$ is bounded by the closed graph theorem and therefore

$$\|x\| = \|(A - \lambda)^{-1}(A - \lambda)x\| \leq c\|(A - \lambda)x\|$$

holds for all $x \in \mathcal{D}(A)$ and thus we have $\lambda \notin \sigma_{\mathrm{app}}(A)$. This proves a) which implies b). ❑

1.4 The Spectra of Compact Operators

A linear operator $A\colon \mathcal{X} \to \mathcal{X}$ is called *compact*, if one of the following equivalent conditions is satisfied:

a) A maps every bounded set to a relatively compact set;

b) A maps the closed unit ball to a relatively compact set;

c) For every bounded sequence $(x_n)_n$ in $\mathcal{X}$, the sequence $(Tx_n)_n$ contains a converging subsequence.

Theorem 1.4.1 (Riesz, Schauder)**.** *Let A be compact. Then the following assertions hold:*

a) *The possibly empty set $\sigma(A) \setminus \{0\}$ contains at most countably many eigenvalues λ_j of A;*

b) *If $\sigma(A)$ is infinite, then $\lambda_j \to 0$ as $j \to \infty$.*

Proof. See [47, Theorem 2.10]. ❑

Corollary 1.4.2. *Let A be a closed operator with compact resolvent. Then the following assertions hold:*

a) *$\sigma(A)$ is either empty or $\sigma(A) = \sigma_{\mathrm{p}}(A)$ contains at most countably many eigenvalues λ_j;*

b) *If $\sigma(A)$ is infinite, then $|\lambda_j| \to \infty$ as $j \to \infty$.*

Proof. Let $\mu \in \varrho(A)$. From Lemma 1.1.7 we obtain

$$\sigma(R(A,\mu)) \setminus \{0\} = (\sigma(A) - \mu)^{-1}.$$

Moreover, by replacing 'bijective' with 'injective' in the proof of this lemma, we also get the equality

$$\sigma_{\mathrm{p}}(R(A,\mu)) \setminus \{0\} = (\sigma_{\mathrm{p}}(A) - \mu)^{-1}.$$

From Theorem 1.4.1 we know that $\sigma(R(A,\mu)) \setminus \{0\}$ contains at most countably many eigenvalues μ_j of $R(A,\mu)$ that converge to zero if there are infinitely many. Hence, we have

$$(\sigma(A) - \mu)^{-1} = \sigma(R(A,\mu)) \setminus \{0\} = \sigma_{\mathrm{p}}(R(A,\mu)) \setminus \{0\} = (\sigma_{\mathrm{p}}(A) - \mu)^{-1}$$

These facts imply assertions a) and b), where $\lambda_j = \mu + \mu_j^{-1}$. ❑

Chapter 2

Spectral Supersets

The explicit computation of the whole spectrum of a linear operator by analytical or numerical techniques is only possible in rare cases. Moreover, the spectrum is in general quite sensitive with respect to small perturbations of the operator. Therefore, one is interested in supersets of the spectrum that are easier to compute and that are also robust under perturbations.

In this chapter, we recall what is known about these desired properties when it comes to the ε-pseudospectrum, the numerical range and the quadratic numerical range. For each of these we start by giving the definitions and basic properties.

2.1 The ε-Pseudospectrum

Let $\mathcal{X}$ be a complex Banach space and $A\colon \mathcal{D}(A) \subset \mathcal{X} \to \mathcal{X}$ a linear operator. As we have already seen in Section 1.2, the spectrum of A tends to be quite sensitive with respect to small perturbations of the operator. It is therefore natural to consider the union of the spectra of versions of A that have been slightly perturbed.

Definition 2.1.1. The *ε-pseudospectrum* of A is defined as

$$\sigma_\varepsilon(A) = \bigcup_{\|P\|<\varepsilon} \sigma(A+P).$$

The notion of the ε-pseudospectrum has been independently introduced by Landau [35], Varah [54], Godunov [33], Trefethen [49] and Hinrichsen and Pritchard [25]. Besides the fact that the pseudospectrum is robust under perturbations, it is also suitable to determine the transient growth behavior of linear dynamic models in finite time, which may be far from the asymptotic behavior. For an overview on the pseudospectrum and its applications we refer the reader to [51] and [14].

Scaling and shifting of the operator A translates to the pseudospectrum in the following way: We have

$$\sigma_\varepsilon(\alpha + \beta A) = \alpha + \beta \sigma_{\frac{\varepsilon}{|\beta|}}(A)$$

for $\alpha, \beta \in \mathbb{C}$ because of (1.1).

From Theorem 1.1.2 c), we have

$$\|R(A,\lambda)\| \geq \frac{1}{\operatorname{dist}(\lambda, \sigma(A))}$$

for all $\lambda \in \varrho(A)$. Thus, $\|R(A,\lambda)\|$ approaches ∞ when λ approaches the spectrum. We will therefore introduce the following notational convention that will tacitly be used throughout this thesis.

$$\boxed{\text{If } \lambda \in \sigma(A), \text{ we write } \|R(A,\lambda)\| = \infty.}$$

Theorem 2.1.2. *The ε-pseudospectrum can be equivalently defined by*

$$\sigma_\varepsilon(A) = \left\{ \lambda \in \mathbb{C} \,\middle|\, \|R(A,\lambda)\| > \frac{1}{\varepsilon} \right\}.$$

Proof. This is a consequence of Theorem 1.2.1. Let $\lambda \in \bigcup_{\|P\|<\varepsilon} \sigma(A+P)$. Then there exists a $P \in \mathcal{L}(\mathcal{X})$ with $\|P\| < \varepsilon$ such that

$$\lambda \in \sigma(A+P). \tag{2.1}$$

Now suppose $\|R(A,\lambda)\| \leq \frac{1}{\varepsilon}$, i.e. $\frac{1}{\|R(A,\lambda)\|} \geq \varepsilon$. Then Theorem 1.2.1 a) implies $\lambda \in \varrho(A+P)$, which contradicts (2.1). The other inclusion follows directly from Theorem 1.2.1 b). ❑

For two sets $K, L \subset \mathbb{C}$ we define

$$K + L = \left\{ k + l \,\middle|\, k \in K,\ l \in L \right\}.$$

Theorem 2.1.3. *The following spectral inclusion properties hold:*

a) *In general, we have*

$$\sigma_\varepsilon(A) \supset \sigma(A) + \mathrm{B}_\varepsilon(0);$$

b) *If $\mathcal{X}$ is a Hilbert space and $A \in \mathcal{L}(\mathcal{X})$ a normal operator, then*

$$\sigma_\varepsilon(A) = \sigma(A) + \mathrm{B}_\varepsilon(0).$$

Proof. a) $\sigma_\varepsilon(A) \supset \sigma(A)$ is obvious. Let $\lambda \in \varrho(A)$ with $\operatorname{dist}(\lambda, \sigma(A)) < \varepsilon$. Then Theorem 1.1.2 c) yields

$$\|R(A,\lambda)\| \geq \frac{1}{\operatorname{dist}(\lambda, \sigma(A))} > \frac{1}{\varepsilon}.$$

Hence, $\lambda \in \sigma_\varepsilon(A)$ by Theorem 2.1.2.

b) Let $\lambda \in \sigma_\varepsilon(A) \setminus \sigma(A)$. Then

$$\frac{1}{\operatorname{dist}(\lambda, \sigma(A))} = \|R(A,\lambda)\| > \frac{1}{\varepsilon}$$

by Corollary 1.1.8 and Theorem 2.1.2. ❑

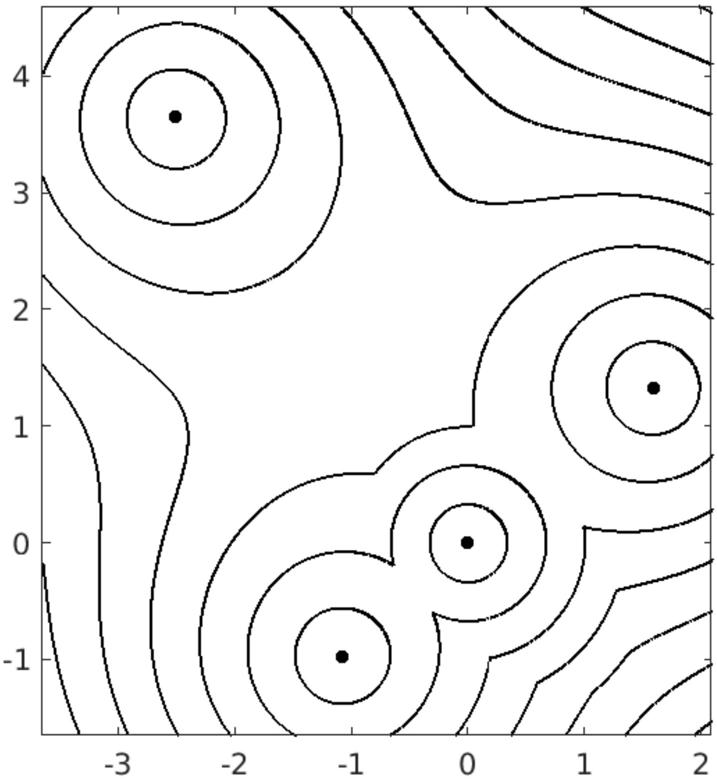


Figure 2.1: $\sigma_\varepsilon(A)$ from Example 2.1.4 computed with `EigTool`

The numerical computation of the pseudospectrum of a matrix has been intensively studied in the literature. Most algorithms use simple grid-based methods, where one computes the smallest singular value of $A - z$ at the points z of a grid, or path-following methods, see the survey [50] or the overview at [14]. Both methods face several challenges. The main problem of grid-based methods is first to find a suitable region in the complex plane and then to perform the computation on a usually very large number of grid points. The main difficulty of path-following algorithms is to find a starting point, that is, a point on the boundary of the pseudospectrum. Moreover, as the pseudospectrum may be disconnected it is difficult to find every component. However, there are several speedup techniques available, see [50], which are essential for applications.

The usual procedure to compute the pseudospectrum of a linear operator on an infinite-dimensional Hilbert space is to approximate it by matrices and then to calculate the pseudospectrum of one of the approximating matrices. In [51, Chapter 43], spectral methods are used for the approximation, but no convergence properties of the pseudospectrum under discretization are proved. So far, only few results are available concerning the relations between the pseudospectra of the discretized operators and those of the infinite-dimensional operator. Convergence properties of the pseudospectrum under discretization have been studied for the linearized Navier-Stokes equation [19], for band-dominated bounded operators [40] and for Toeplitz operators [7]. Bögli and Siegl [4, 6] prove local and global convergence of the pseudospectra of a sequence of linear operators which converge in a generalized resolvent sense. In [10], Colbrook et al. introduced a way of computing pseudospectra of operators that can be written as an infinite-dimensional matrix. Here, the operator A gets truncated and the algorithm produces subsets of the pseudospectrum that converge to $\sigma_\varepsilon(A)$ with a form of error control as the size of the truncation increases. Further, Wolff [57] shows some abstract convergence results for the approximate point spectrum of a linear operator using the pseudospectra of the approximations.

Example 2.1.4. Let us consider the matrix

$$A = \begin{bmatrix} 0 & 0 & 0 & 2\mathrm{i} \\ 0 & 0 & 0 & 0 \\ 0 & 0 & -1+2\mathrm{i} & 2-2\mathrm{i} \\ 3 & 0 & 2-2\mathrm{i} & -1+2\mathrm{i} \end{bmatrix}.$$

Figure 2.1 depicts the boundaries of the sets $\sigma_{\frac{1}{3}}(A), \dots, \sigma_{\frac{7}{3}}(A)$ computed with the `EigTool` Matlab package from [13]. Here, the black dots represent the eigenvalues of A. As we see, the level sets differ from $\sigma(A) + \mathrm{B}_\varepsilon(0)$ and if ε is small enough, $\sigma_\varepsilon(A)$ consists of several disconnected components.

2.2 The Numerical Range

Another well-established and thoroughly studied superset is the numerical range. It is always convex which comes with both advantages and disadvantages as we will see in the following short overview on its fundamental properties and in comparison to other spectral supersets later on. More details can be found in [22, 23, 27, 32].

Let $\mathcal{H}$ be a Hilbert space over $\mathbb{C}$ and let $\langle\cdot,\cdot\rangle\colon \mathcal{H}\times\mathcal{H}\to\mathbb{C}$ denote the inner product on $\mathcal{H}$. We will consider a linear operator $A\colon \mathcal{D}(A)\subset\mathcal{H}\to\mathcal{H}$.

Definition 2.2.1. The *numerical range* of the operator A is defined by

$$W(A) = \{\langle Ax, x\rangle \mid x \in \mathcal{D}(A),\ \|x\| = 1\}.$$

In the literature, this set is sometimes also referred to as the *field of values* of A.

It is immediate form the definition that

$$W(\alpha + \beta A) = \alpha + \beta W(A) \tag{2.2}$$

for $\alpha, \beta \in \mathbb{C}$. If the operator A is bounded, then $W(A)$ is bounded as well because of

$$|\langle Ax, x\rangle| \leq \|A\|$$

for all $x \in \mathcal{H}$ with $\|x\| = 1$ by the Cauchy-Schwarz inequality. If in addition $\dim \mathcal{H} < \infty$, then $W(A)$ is compact since it is the image of a compact set under a continuous mapping.

Theorem 2.2.2 (Toeplitz-Hausdorff Theorem)**.** *The numerical range $W(A)$ is a convex set.*

Proof. Let λ and κ be elements of $W(A)$. We will show that the line segment between λ and κ is also included in $W(A)$. Due to the linearity of the numerical range (2.2) we can assume without loss of generality that $\lambda = 0$ and $\kappa = 1$, so let $x, y \in \mathcal{D}(A)$ with $\|x\| = \|y\| = 1$ such that $\langle Ax, x\rangle = 0$ and $\langle Ay, y\rangle = 1$. Our goal is to find coefficients $\alpha, \beta \in \mathbb{R}$ such that for $z = \alpha x + \beta y \in \mathcal{D}(A)$ the system

$$\begin{cases} \|z\|^2 = \alpha^2 + \beta^2 + 2\alpha\beta\Re\langle x, y\rangle = 1, \\ \langle Az, z\rangle = \beta^2 + \alpha\beta(\langle Ax, y\rangle + \langle Ay, x\rangle) = r \end{cases} \tag{2.3}$$

with $0 < r < 1$ is satisfied. (2.3) is clearly solvable whenever $B := \langle Ax, y\rangle + \langle Ay, x\rangle$ is real. In the other case, if $B \in \mathbb{C} \setminus \mathbb{R}$, we can multiply x by an appropriate scalar factor in order to obtain a version of (2.3) with a real B. To this end consider $\widetilde{x} = \mu x$ where $\mu = a + ib$ satisfies the system

$$\begin{cases} |\mu|^2 = a^2 + b^2 = 1, \\ \Im B(\widetilde{x}) = a\Im B(x) + b\Re(\langle Ax, y\rangle - \langle Ay, x\rangle) = 0 \end{cases}$$

which is clearly solvable. ❑

Theorem 2.2.3. *The following spectral inclusion properties hold:*

a) $\sigma_{\mathrm{p}}(A) \subset W(A)$*;*

b) $\sigma_{\mathrm{app}}(A) \subset \overline{W(A)}$*;*

c) *If A is bounded, we have $\sigma(A) \subset \overline{W(A)}$;*

d) *If A is closed and has compact resolvent, we have $\sigma(A) \subset W(A)$.*

Proof. a) Let $\lambda \in \sigma_{\mathrm{p}}(A)$ with corresponding eigenvector $x \in \mathcal{D}(A)$ such that $\|x\| = 1$. Then $\langle Ax, x\rangle = \lambda$.

b) Let $\lambda \in \sigma_{\mathrm{app}}(A)$. Then there exists a sequence of approximate eigenvectors $(x_n)_n \subset \mathcal{D}(A)$ such that $\lim_{n\to\infty}\|(A - \lambda)x_n\| = 0$. By the Cauchy-Schwarz inequality we obtain

$$\begin{aligned} \langle Ax_n, x_n\rangle - \lambda &= \langle (A - \lambda)x_n, x_n\rangle \\ &\leq \|(A - \lambda)x_n\| \end{aligned}$$

and therefore

$$\lim_{n\to\infty} \langle Ax_n, x_n\rangle = \lambda.$$

This yields $\lambda \in \overline{W(A)}$.

c) From Theorem 1.1.4 we know that $\sigma(A)$ is bounded and from Lemma 1.3.5 we have $\partial\sigma(A) \subset \sigma_{\text{app}}(A)$. This together with b) proves the assertion.

d) This follows immediately from a) and Corollary 1.4.2. ❑

Theorem 2.2.4. *If $A \in \mathcal{L}(\mathcal{H})$ is a normal operator, then the closure of the numerical range coincides with the convex hull of the spectrum, i.e.*

$$\overline{W(A)} = \operatorname{conv}(\sigma(A)).$$

Proof. $\overline{W(A)}$ coincides with $\operatorname{conv}(\sigma(A))$ if and only if every closed half plane in $\mathbb{C}$ containing $\sigma(A)$ also contains $W(A)$. Since rotating and shifting A rotates and shifts $\sigma(A)$ and $W(A)$ accordingly, it suffices to consider the case where

$$\sigma(A) \subset \{\lambda \in \mathbb{C} \,|\, \Re\lambda \le 0\}. \tag{2.4}$$

Suppose, $W(A) \not\subset \{\lambda \in \mathbb{C} \,|\, \Re\lambda \le 0\}$, i.e. there exists $a+ib \in W(A)$ with $a > 0$. Let $x \in \mathcal{D}(A)$, $\|x\| = 1$, with $\langle Ax, x\rangle = a+ib$ and take $y \in \mathcal{H}$ such that $Ax = (a+ib)x+y$ and $\langle x, y\rangle = 0$.
For any $\lambda \in \mathbb{R}$ with $\lambda > 0$ we have $\lambda \in \varrho(A)$ by (2.4) and

$$\operatorname{dist}(\lambda, \sigma(A)) = \frac{1}{\|R(A,\lambda)\|} \le \|(A-\lambda)x\|$$

by Corollary 1.1.8. This implies

$$\begin{aligned}\lambda^2 \le \operatorname{dist}(\lambda, \sigma(A))^2 &\le \|(a-\lambda+ib)x+y\|^2 \\ &= (a-\lambda)^2 + b^2 + \|y\|^2.\end{aligned}$$

Hence, $2a\lambda \le a^2 + b^2 + \|y\|^2$ with $a > 0$ for every $\lambda \in \mathbb{R}$ with $\lambda > 0$. This is impossible. ❑

Let us now consider a bounded linear operator $A\colon \mathcal{H} \to \mathcal{H}$ where $\dim \mathcal{H} < \infty$, i.e. the case in which $\mathcal{H}$ is isomorphic to $\mathbb{C}^n$ for some $n \in \mathbb{N}$ and A has a matrix representation. Over the years, several approaches for the numerical computation of $W(A)$ in the matrix case have been developed and they are based on at least one of the two following core ideas. Note, that the convexity of the numerical range is exploited in all of them.

The first method has been introduced by Johnson in [31] and is based on the following observation: If $0 \le \theta \le 2\pi$ and x_θ is a unit eigenvector associated to the largest eigenvalue of the hermitian matrix $\frac{1}{2}\left(e^{i\theta}A + e^{-i\theta}A^*\right)$, then $\langle Ax_\theta, x_\theta\rangle \in \partial W(A)$. An approximation to $W(A)$ is then obtained by choosing a mesh $\theta_j = (j-1)\frac{2\pi}{k}$, $j = 1, \dots, k$, and the computation of the boundary points $p_{\theta_j} = \langle Ax_{\theta_j}, x_{\theta_j}\rangle$

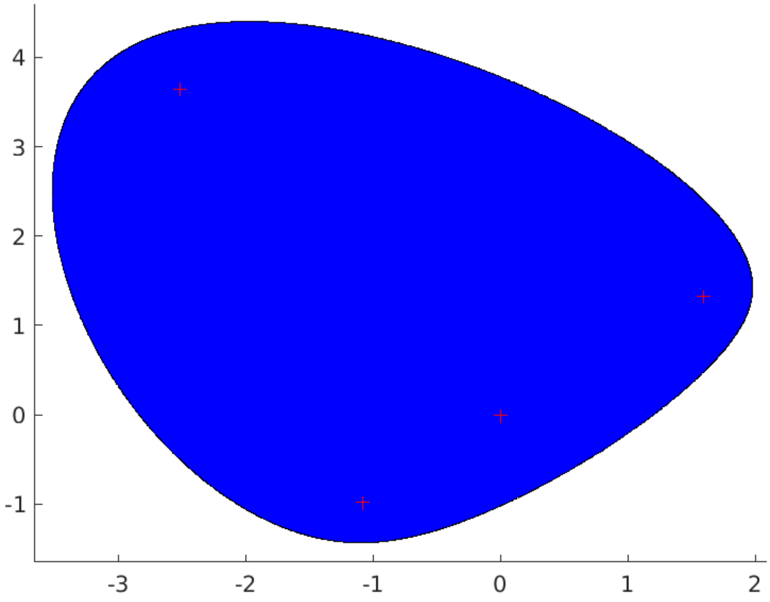


Figure 2.2: The numerical range and the eigenvalues of A from Example 2.2.5

such that the union of line segments that join p_{θ_j} to $p_{\theta_{j+1}}$, $j = 1, \ldots, k$, can be plotted. In [12], Cowen and Harel paid particular attention to the case in which a flat segment occurs in the boundary of $W(A)$ and in [8], Braconnier and Higham applied a specific implementation of the Lanczos method to optimize the required eigenvalue solves for the hermitian matrices. Loisel and Maxwell developed an algorithm for the computation of $\partial W(A)$ that is based on tracking the dominant eigenpair of the Hermitian part of $\mathrm{e}^{\mathrm{i}t}A$ by solving an ordinary differential equation, see [41].

The second method for the computation of the numerical range is based on the fact that $W(A)$ is always an ellipse if $\dim \mathcal{H} = 2$, see [23, Lemma 1.1-1]. For a bounded operator A, Theorem 2.2.2 can also be deduced from this property by showing that $W(A)$ is equal to the union of the numerical ranges of all two-dimensional compressions of A, see [23, Theorem 1.1-2]. In [42], Marcus and Pesce showed that for matrices it is sufficient to consider the compressions derived from pairs of real orthonormal vectors, i.e.

$$W(A) = \bigcup_{x,y} W(A_{xy})$$

where

$$A_{xy} = \begin{bmatrix} \langle Ax, x \rangle & \langle Ay, x \rangle \\ \langle Ax, y \rangle & \langle Ay, y \rangle \end{bmatrix}$$

with x and y varying over all pairs of real orthonormal vectors. They utilized this result to develop an algorithm for the computation of the numerical range by generating a random set of real orthonormal vector pairs and computing the union of the sets $W(A_{xy})$. Bebiano et al. improved this approach by using suitably chosen vectors which generate boundary points of $W(A)$, see [2], and Uhlig made use of Johnson's algorithm to select x and y such that each ellipse is more likely to constrain the boundary, see [53].

Example 2.2.5. Let us consider the matrix

$$A = \begin{bmatrix} 0 & 0 & 0 & 2\mathrm{i} \\ 0 & 0 & 0 & 0 \\ 0 & 0 & -1+2\mathrm{i} & 2-2\mathrm{i} \\ 3 & 0 & 2-2\mathrm{i} & -1+2\mathrm{i} \end{bmatrix}$$

from Example 2.1.4 again. Figure 2.2 depicts the numerical range $W(A)$ as a blue set and the eigenvalues of A as red crosses. The computation of the numerical range was done in Matlab using Johnson's algorithm.

2.3 The Quadratic Numerical Range

In [37], Langer and Tretter introduced the quadratic numerical range (QNR) as a new concept to enclose the spectrum of a block operator matrix in a Hilbert space. The QNR is a subset of the numerical range that is not necessarily convex and consists of at most two connected components which need not be convex either. See [36] and the monograph [52], where many more properties are proven as well. For applications of the QNR, we refer to [17], [29], [34] and [39] where the superset is exploited for Krylov type methods, damped systems, spectral perturbation results and the location of zeros of polynomials. In [45] and [46] approximation schemes for possibly unbounded operators are established and convergence theorems are proven relating the QNR of an operator to the QNR of its finite-dimensional discretizations.

This section is devoted to state the definition and basic properties of the quadratic numerical range. Let therefore $\mathcal{H}$ be a Hilbert space over $\mathbb{C}$, let $\langle\cdot,\cdot\rangle\colon \mathcal{H}\times\mathcal{H}\to\mathbb{C}$ be the inner product on $\mathcal{H}$ and consider an arbitrary but fixed decomposition of $\mathcal{H}$ denoted by $\mathcal{H}_1\oplus\mathcal{H}_2=\mathcal{H}$. Furthermore, let $\mathscr{A}\colon\mathcal{H}\to\mathcal{H}$ be a bounded linear operator that will henceforth be written in the block operator matrix form

$$\mathscr{A} = \begin{bmatrix} A & B \\ C & D \end{bmatrix}$$

where $A\colon\mathcal{H}_1\to\mathcal{H}_1$, $B\colon\mathcal{H}_2\to\mathcal{H}_1$, $C\colon\mathcal{H}_1\to\mathcal{H}_2$, $D\colon\mathcal{H}_2\to\mathcal{H}_2$ are bounded operators. Note, that every bounded operator on $\mathcal{H}$ can be written in such a form once a decomposition $\mathcal{H}=\mathcal{H}_1\oplus\mathcal{H}_2$ is chosen.

Definition 2.3.1. The *quadratic numerical range (QNR)* is given by

$$W^2(\mathscr{A}) = \bigcup_{x \in S_{\mathcal{H}_1}, y \in S_{\mathcal{H}_2}} \sigma\left(\begin{bmatrix} \langle Ax, x\rangle & \langle By, x\rangle \\ \langle Cx, y\rangle & \langle Dy, y\rangle \end{bmatrix}\right),$$

where $S_{\mathcal{H}_i} = \{x \in \mathcal{H}_i \,|\, \|x\| = 1\}$, $i = 1, 2$.

In other words, the QNR consists of the solutions λ of the quadratic equations

$$\lambda^2 - (\langle Ax, x\rangle + \langle Dy, y\rangle)\lambda + \langle Ax, x\rangle\langle Dy, y\rangle - \langle By, x\rangle\langle Cx, y\rangle = 0 \tag{2.5}$$

with $(x, y) \in S_{\mathcal{H}_1} \times S_{\mathcal{H}_2}$.

Just as for the numerical range and the spectrum itself, we have

$$W^2(\alpha + \beta\mathscr{A}) = \alpha + \beta W^2(\mathscr{A})$$

for $\alpha, \beta \in \mathbb{C}$.

In order to shorten the notation we will henceforth use the abbreviations

$$M_{x,y} := \begin{bmatrix} \langle Ax, x\rangle & \langle By, x\rangle \\ \langle Cx, y\rangle & \langle Dy, y\rangle \end{bmatrix} \in \mathbb{C}^{2\times 2}$$

for $(x, y) \in S_{\mathcal{H}_1} \times S_{\mathcal{H}_2}$. This matrix can also be written in the form $M_{x,y} = P\mathscr{A}|_{\operatorname{ran} P}$, where P is the orthogonal projection to the two-dimensional subspace of $\mathcal{H}_1 \oplus \mathcal{H}_2$ spanned by $\begin{bmatrix} x \\ 0 \end{bmatrix}$ and $\begin{bmatrix} 0 \\ y \end{bmatrix}$. Hence, we have

$$W^2(\mathscr{A}) \subset \{\lambda \in \mathbb{C} \,|\, |\lambda| \le \|\mathscr{A}\|\}. \tag{2.6}$$

The eigenvalues of a matrix depend continuously on its entries, see Theorem 1.2.4, and therefore the mapping

$$S_{\mathcal{H}_1} \times S_{\mathcal{H}_2} \ni (x, y) \mapsto \sigma(M_{x,y}) \tag{2.7}$$

is continuous as well. Thus, if $\dim \mathcal{H} < \infty$, $W^2(\mathscr{A})$ is compact since it is the image of a compact set under a continuous mapping. Moreover, the continuity of (2.7) explains the fact, that the QNR consists of at most two connected components.

Theorem 2.3.2. *We have*

$$W^2(\mathscr{A}) \subset W(\mathscr{A}).$$

Proof. Let $\lambda \in W^2(\mathscr{A})$. Then there exist $(x, y) \in S_{\mathcal{H}_1} \times S_{\mathcal{H}_2}$ such that $\lambda \in \sigma(M_{x,y})$, i.e. there exists a vector $\begin{bmatrix} v_1 \\ v_2 \end{bmatrix} \in \mathbb{C}^2$ with $|v_1|^2 + |v_2|^2 = 1$ such that

$$\begin{bmatrix} \langle Ax, x\rangle & \langle By, x\rangle \\ \langle Cx, y\rangle & \langle Dy, y\rangle \end{bmatrix} \begin{bmatrix} v_1 \\ v_2 \end{bmatrix} = \lambda \begin{bmatrix} v_1 \\ v_2 \end{bmatrix}.$$

Taking the scalar product with $\begin{bmatrix} v_1 \\ v_2 \end{bmatrix}$ yields

$$\left\langle \begin{bmatrix} A & B \\ C & D \end{bmatrix} \begin{bmatrix} v_1 x \\ v_2 y \end{bmatrix}, \begin{bmatrix} v_1 x \\ v_2 y \end{bmatrix} \right\rangle = \lambda.$$

This implies $\lambda \in W(\mathscr{A})$ because $\left\| \begin{bmatrix} v_1 x \\ v_2 y \end{bmatrix} \right\| = 1.$ ❑

Theorem 2.3.3. *The following spectral inclusion properties hold:*

a) $\sigma_{\mathrm{p}}(\mathscr{A}) \subset W^2(\mathscr{A})$;

b) $\sigma(\mathscr{A}) \subset \overline{W^2(\mathscr{A})}$.

The proof of Theorem 2.3.3 requires the following lemma about 2 × 2 matrices.

Lemma 2.3.4. *Let $M \in \mathbb{C}^{2\times 2}$. If there exists a vector $x \in \mathbb{C}^2$ such that*

$$\|x\| = 1 \qquad \textit{and} \qquad \|Mx\| < \varepsilon \tag{2.8}$$

for some $\varepsilon > 0$, then $\operatorname{dist}(0, \sigma(M)) \leq \sqrt{\|M\|\varepsilon}$.

Proof. We only have to consider the case in which $0 \notin \sigma(M)$. Then we can transform (2.8) into an estimate for the norm of M^{-1} via

$$\begin{aligned} \|M^{-1}\| &= \sup_{\|y\|=1} \|M^{-1}y\| > \left\| M^{-1} \frac{Mx}{\|Mx\|} \right\| \\ &= \left\| \frac{x}{\|Mx\|} \right\| = \frac{\|x\|}{\|Mx\|} > \frac{1}{\varepsilon}. \end{aligned} \tag{2.9}$$

Moreover, because M^{-1} can be written as

$$M^{-1} = \frac{1}{\det M} \left(U^{\mathsf{T}} M U\right)^{\mathsf{T}}$$

with the unitary matrix $U := \begin{bmatrix} 0 & 1 \\ -1 & 0 \end{bmatrix}$ and because taking the transpose of a matrix does not change the norm, we obtain

$$\|M^{-1}\| = \frac{\|M\|}{|\det M|} = \frac{\|M\|}{|\lambda_1 \lambda_2|}$$

where λ_1 and λ_2 are the eigenvalues of M. Combining this with (2.9) yields

$$\min\{|\lambda_1|, |\lambda_2|\} \leq \sqrt{\|M\|\varepsilon}.$$ ❑

Proof of Theorem 2.3.3. a) Let $\lambda \in \sigma_{\mathrm{p}}(\mathscr{A})$. Then there exists an eigenvector $(x, y) \in \mathcal{H}_1 \oplus \mathcal{H}_2$ such that

$$\begin{aligned} Ax + By &= \lambda x, \\ Cx + Dy &= \lambda y. \end{aligned}$$

By choosing $(\hat{x}, \hat{y}) \in S_{\mathcal{H}_1} \times S_{\mathcal{H}_2}$ such that $x = \|x\|\hat{x}$ and $y = \|y\|\hat{y}$ (with $\hat{x} \in S_{\mathcal{H}_1}$ arbitrary if e.g. $x = 0$) this gives us

$$\begin{aligned} \langle Ax, \hat{x}\rangle + \langle By, \hat{x}\rangle &= \lambda\langle x, \hat{x}\rangle, \\ \langle Cx, \hat{y}\rangle + \langle Dy, \hat{y}\rangle &= \lambda\langle y, \hat{y}\rangle, \end{aligned}$$

which can be rewritten as

$$M_{\hat{x},\hat{y}} \begin{bmatrix} \|x\| \\ \|y\| \end{bmatrix} = \lambda \begin{bmatrix} \|x\| \\ \|y\| \end{bmatrix}.$$

Hence, $\lambda \in \sigma_{\mathrm{p}}(M_{\hat{x},\hat{y}}) \subset W^2(\mathscr{A})$.

b) By Lemma 1.3.6 b) we have $\sigma(\mathscr{A}) = \sigma_{\mathrm{app}}(\mathscr{A}) \cup \sigma_{\mathrm{r}}(\mathscr{A})$ and we split the proof into two cases according to this decomposition.
Let $\lambda \in \sigma_{\mathrm{app}}(\mathscr{A})$. Then there exists a sequence $(x_n, y_n)_n \subset \mathcal{H}_1 \oplus \mathcal{H}_2$ with $\|x_n\|^2 + \|y_n\|^2 = 1$ for all $n \in \mathbb{N}$ such that

$$\lim_{n\to\infty} \left\| (\mathscr{A} - \lambda) \begin{bmatrix} x_n \\ y_n \end{bmatrix} \right\| = 0.$$

Choosing $(\hat{x}_n, \hat{y}_n) \in S_{\mathcal{H}_1} \times S_{\mathcal{H}_2}$ as in part a) of the proof, we obtain

$$\lim_{n\to\infty} \left(M_{\hat{x}_n,\hat{y}_n} - \lambda \right) \begin{bmatrix} \|x_n\| \\ \|y_n\| \end{bmatrix} = 0.$$

This together with $\|M_{\hat{x}_n,\hat{y}_n}\| \leq \|\mathscr{A}\|$ for all $n \in \mathbb{N}$ (c.f. (2.6) and the reasoning above) yields

$$\lim_{n\to\infty} \operatorname{dist}(\lambda, \sigma(M_{\hat{x}_n,\hat{y}_n})) = 0$$

by Lemma 2.3.4, i.e. $\lambda \in \overline{W^2(\mathscr{A})}$
Let now $\lambda \in \sigma_{\mathrm{r}}(\mathscr{A})$. Then by Remark 1.3.3 we have $\overline{\lambda} \in \sigma_{\mathrm{p}}(\mathscr{A}^*)$ and therefore $\overline{\lambda} \in W^2(\mathscr{A}^*)$ by part a). $W^2(\mathscr{A}^*)$ coincides with $W^2(\mathscr{A})^*$, because all the coefficients in (2.5) are complex conjugated. Hence, $\lambda \in W^2(\mathscr{A})$. ❑

Example 2.3.5. Let us consider the matrix

$$\mathscr{A} = \left[\begin{array}{cc|cc} 0 & 0 & 0 & 2\mathrm{i} \\ 0 & 0 & 0 & 0 \\ \hline 0 & 0 & -1+2\mathrm{i} & 2-2\mathrm{i} \\ 3 & 0 & 2-2\mathrm{i} & -1+2\mathrm{i} \end{array}\right]$$

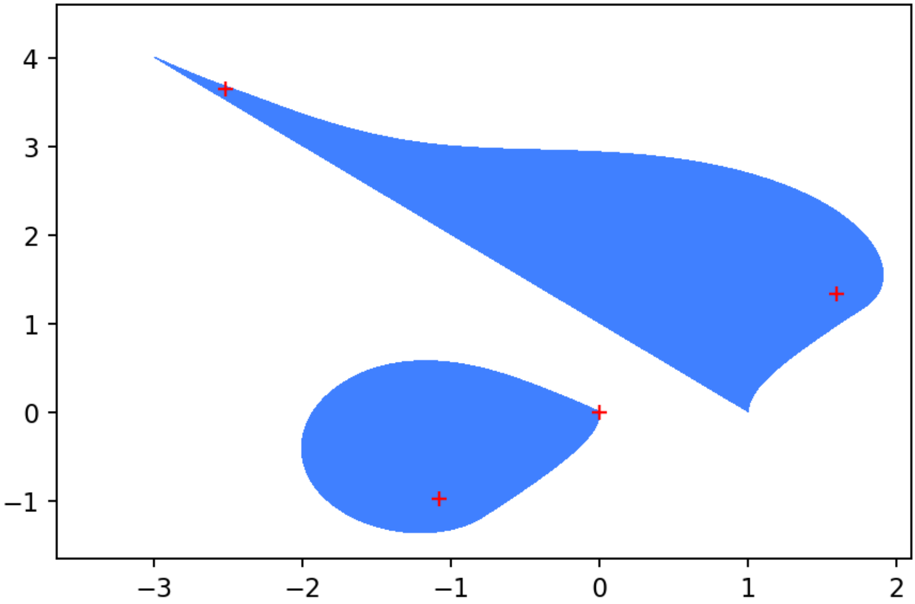


Figure 2.3: The QNR and the eigenvalues of $\mathscr{A}$ from Example 2.3.5

from Examples 2.1.4 and 2.2.5 again with the decomposition $\mathbb{C}^4 = \mathbb{C}^2 \oplus \mathbb{C}^2$. Figure 2.3 depicts the quadratic numerical range $W^2(\mathscr{A})$ as a blue set and the eigenvalues of $\mathscr{A}$ as red crosses. The computation of the QNR was done in Python by use of the algorithm introduced in Chapter 4.

Chapter 3

Pseudospectrum Enclosures by Discretization

As described in Section 2.1, the ε-pseudospectrum of an operator is defined as the union of the spectra of slightly perturbed versions of the operator. This superset of the spectrum is robust under perturbations and preserves valuable information. Unfortunately, it is hard to compute in the infinite-dimensional case, which is why we are interested in a computable yet tight enclosure.

A simple method to enclose the pseudospectrum is in terms of the numerical range. More precisely, under an additional weak assumption, the ε-pseudospectrum is contained in an ε-neighborhood of the numerical range of the operator, see Remark 3.1.9. While this superset is easy to compute for matrices, it can not distinguish disconnected components of the pseudospectrum as the numerical range is convex.

In this chapter, we propose a new method to enclose the pseudospectrum via the numerical range of the inverse of the matrix or linear operator. More precisely, for a linear operator A on a Hilbert space and $\varepsilon > 0$ we show

$$\sigma_\varepsilon(A) \subset \bigcap_{s\in S} \left[\left(\mathrm{B}_{\delta_s}(W((A-s)^{-1})) \right)^{-1} + s \right], \tag{3.1}$$

see Theorem 3.1.3. Here, S is a suitable subset of the resolvent set of A and the inverse of a subset $B \subset \mathbb{C}$ is defined via $B^{-1} := \left\{ b^{-1} \,\middle|\, b \in B \setminus \{0\} \right\}$. This inclusion holds for matrices as well as for linear operators on Hilbert spaces. Further, we show that the enclosure of the pseudospectrum in (3.1) becomes optimal in the sense that it is contained in the closure of the pseudospectrum if the set S is chosen optimally, see Theorem 3.1.6. The idea to study the numerical range of the inverses stems from the fact that the spectrum of a matrix can be expressed in terms of inverses of shifted matrices [26].

We will then refine the enclosure (3.1) of the pseudospectrum of linear operators further and show that it is sufficient to calculate the numerical ranges of approximating matrices. More precisely, we show in Theorem 3.2.6 that

$$\sigma_\varepsilon(A) \subset \bigcap_{s\in S} \left[\left(\mathrm{B}_{\delta_s}(W((A_n - s)^{-1}))\right)^{-1} + s\right] \tag{3.2}$$

if n is sufficiently large. Here $(A_n)_n$ is a sequence of matrices which approximates the operator A strongly. We refer to Section 3.2 for the precise definition of strong approximation. If we even have a uniform approximation of the operator A, then we are able to prove an estimate for the index n such that (3.2) holds in intersections with compact subsets of the complex plane, see Section 3.3. In Section 3.4 we show that finite element discretizations of elliptic partial differential operators yield uniform approximations. Further, as an example of a strong approximation we study in Sections 3.5 and 3.6 two classes of structured block operator matrices. Subsequently, in Section 3.7, we apply our obtained results to the advection-diffusion operator, the Hain-Lüst operator and a Stokes-type operator by plotting and discussing the computed supersets.

From a numerical point of view this new method faces similar challenges as grid-based methods as a suitable set S of points has to be found and then the numerical ranges of a large number of matrices have to be computed. However, this new method has the advantage that it enables us to enclose the pseudospectrum of an infinite-dimensional operator by a set which is expressed by the approximating matrices.

In the final section we investigate the relation of the pseudospectrum of a block operator matrix to the pseudospectrum of its Schur complement.

This chapter is an extended version of the article [18]. Here, some of the results have been improved, newly added or are accompanied by more detailed explanations.

Let $\mathcal{H}$ be a Hilbert space with inner product $\langle\cdot,\cdot\rangle$. If not explicitly stated different we assume that $A\colon \mathcal{D}(A) \subset \mathcal{H} \to \mathcal{H}$ is a closed and densely defined linear operator throughout this chapter.

3.1 Pseudospectrum Enclosures using the Numerical Range

In this section we present the basic idea of considering numerical ranges of shifted inverses of an operator in order to obtain an enclosure of its pseudospectrum.

The central idea is the following: If $\lambda \in \mathbb{C}$ is such that $1/\lambda$ has a certain positive distance δ to the numerical range of the inverse operator A^{-1}, then this yields an estimate of the form

$$\|(A-\lambda)x\| \geq \varepsilon\|x\|, \qquad x \in \mathcal{D}(A),$$

with some constant $\varepsilon > 0$. This will in turn be used to show $\lambda \in \varrho(A)$ with $\|(A-\lambda)^{-1}\| \leq \frac{1}{\varepsilon}$, i.e. $\lambda \notin \sigma_\varepsilon(A)$. We make this explicit with the next proposition:

Proposition 3.1.1. *Suppose that* $0 \in \varrho(A)$*. Then for every* $0 < \varepsilon < \frac{1}{\|A^{-1}\|}$ *and* $\delta \geq \frac{\|A^{-1}\|^2\varepsilon}{1-\|A^{-1}\|\varepsilon}$ *we have*

$$\sigma_\varepsilon(A) \subset \left(\mathrm{B}_\delta(W(A^{-1}))\right)^{-1}.$$

Proof. The inclusion is trivial for $\delta = \infty$, so we can assume $\delta < \infty$. Let us denote $U := \left(\mathrm{B}_\delta(W(A^{-1}))\right)^{-1} \subsetneqq \mathbb{C}$. As a first step we show that

$$\|(A-\lambda)x\| \geq \varepsilon \quad \text{for all} \quad \lambda \in \mathbb{C} \setminus U,\, x \in \mathcal{D}(A),\, \|x\| = 1. \tag{3.3}$$

So let $\lambda \in \mathbb{C} \setminus U$. We consider two cases. First suppose that $|\lambda| > \frac{1}{\|A^{-1}\|} - \varepsilon$. Then $\lambda \neq 0$, $\lambda^{-1} \notin \mathrm{B}_\delta(W(A^{-1}))$ and hence $\operatorname{dist}(\lambda^{-1}, W(A^{-1})) \geq \delta$. For $x \in \mathcal{D}(A)$, $\|x\| = 1$ we find

$$\delta \leq |\lambda^{-1} - \langle A^{-1}x, x\rangle| = |\langle(\lambda^{-1} - A^{-1})x, x\rangle| \leq \|(\lambda^{-1} - A^{-1})x\|.$$

Consequently,

$$\begin{aligned}\|(A-\lambda)x\| = |\lambda|\|A(\lambda^{-1} - A^{-1})x\| &\geq \frac{|\lambda|}{\|A^{-1}\|}\|(\lambda^{-1} - A^{-1})x\| \\ &> \frac{\delta}{\|A^{-1}\|}\left(\frac{1}{\|A^{-1}\|} - \varepsilon\right) = \frac{\delta(1 - \|A^{-1}\|\varepsilon)}{\|A^{-1}\|^2} \geq \varepsilon.\end{aligned}$$

In the other case if $|\lambda| \leq \frac{1}{\|A^{-1}\|} - \varepsilon$ then $|\lambda|\|A^{-1}\| \leq 1 - \|A^{-1}\|\varepsilon$ and hence $I - \lambda A^{-1}$ is invertible by a Neumann series argument with $\|(I - \lambda A^{-1})^{-1}\| \leq \frac{1}{\|A^{-1}\|\varepsilon}$. For $x \in \mathcal{D}(A)$ with $\|x\| = 1$ this implies

$$\|(A-\lambda)x\| = \|A(I - \lambda A^{-1})x\| \geq \frac{1}{\|A^{-1}\|\|(I - \lambda A^{-1})^{-1}\|} \geq \varepsilon.$$

We have thus shown (3.3). In particular, $\lambda \in \mathbb{C} \setminus U$ implies $\lambda \notin \sigma_{\mathrm{app}}(A)$, i.e.

$$\sigma_{\mathrm{app}}(A) \cap \mathbb{C} \setminus U = \varnothing. \tag{3.4}$$

By Theorem 2.2.2 and because A^{-1} is bounded, the set $\mathrm{B}_\delta(W(A^{-1}))$ is convex and bounded. Therefore, $\mathbb{C} \setminus \mathrm{B}_\delta(W(A^{-1}))$ is connected and hence also

$$\mathbb{C} \setminus U = \left(\mathbb{C} \setminus \mathrm{B}_\delta(W(A^{-1}))\right)^{-1}.$$

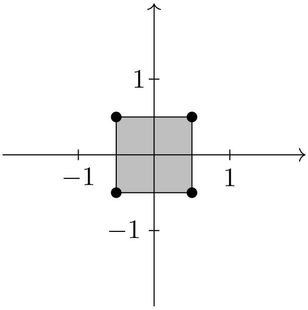
On the other hand, the boundedness of $\mathrm{B}_\delta(W(A^{-1}))$ implies that a neighborhood around 0 belongs to $\mathbb{C} \setminus U$. Consequently, the set $\mathbb{C} \setminus U$ is connected and satisfies $0 \in \varrho(A) \cap \mathbb{C} \setminus U$. Using (3.4) and the fact that $\partial\sigma(A) \subset \sigma_{\mathrm{app}}(A)$ (see Lemma 1.3.5), we conclude that

$$\mathbb{C} \setminus U \subset \varrho(A).$$

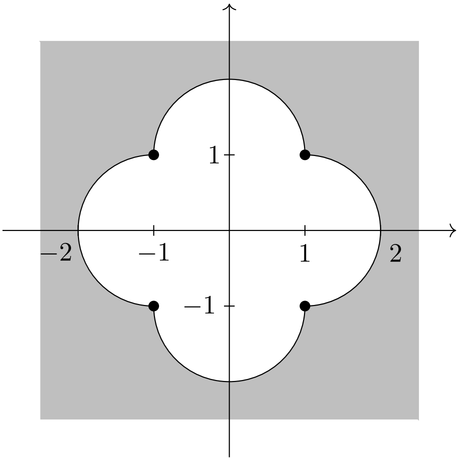
Now (3.3) implies that if $\lambda \in \mathbb{C} \setminus U$ then $\|(A-\lambda)^{-1}\| \leq \frac{1}{\varepsilon}$ and therefore we obtain $\lambda \notin \sigma_\varepsilon(A)$. ❑

The following simple example demonstrates that the δ-neighborhood around the numerical range is actually needed to obtain an enclosure of the pseudospectrum.

Example 3.1.2. Let $A = \mathrm{diag}(-1+i, -1-i, 1+i, 1-i) \in \mathbb{C}^{4\times4}$. Then $A^{-1} = \frac{1}{2}\mathrm{diag}(-1-i, -1+i, 1-i, 1+i)$ is normal and because of the compactness of the numerical range in finite dimensions and Theorem 2.2.4, its numerical range is simply the convex hull of its eigenvalues. Thus, $W(A^{-1})$ is the following square:



Then, using the fact that $z \mapsto \frac{1}{z}$ is a Möbius transformation, we obtain for $W(A^{-1})^{-1}$ the following curve plus its exterior:



We see that $W(A^{-1})^{-1}$ touches the spectrum of A. This is of course clear: if an eigenvalue $1/\lambda$ of A^{-1} is on the boundary of $W(A^{-1})$, then the eigenvalue λ of A is on the boundary of $W(A^{-1})^{-1}$. In particular in this example we do not have $\sigma_\varepsilon(A) \subset W(A^{-1})^{-1}$ for any $\varepsilon > 0$ since $\sigma_\varepsilon(A)$ contains discs with radius ε around the eigenvalues, see Theorem 2.1.3.

Applying Proposition 3.1.1 to the shifted operator $A - s$ and then taking the intersection over a suitable set of shifts, we obtain our first main result on an enclosure of the pseudospectrum:

Theorem 3.1.3. *Consider a set* $S \subset \varrho(A)$ *such that*

$$M := \sup_{s \in S} \|(A-s)^{-1}\| < \infty.$$

Then for $0 < \varepsilon < \frac{1}{M}$ *we get the inclusion*

$$\sigma_\varepsilon(A) \subset \bigcap_{s \in S} \left[\left(\mathrm{B}_{\delta_s}(W((A-s)^{-1})) \right)^{-1} + s \right], \tag{3.5}$$

where $\delta_s \geq \frac{\|(A-s)^{-1}\|^2 \varepsilon}{1 - \|(A-s)^{-1}\|\varepsilon}$.

Proof. For every $s \in S$ we can apply Proposition 3.1.1 to the operator $A - s$ and obtain

$$\sigma_\varepsilon(A) - s = \sigma_\varepsilon(A-s) \subset \left(\mathrm{B}_{\delta_s}(W((A-s)^{-1})) \right)^{-1}. \qquad \square$$

Proposition 3.1.4. *For* $s \in \varrho(A)$ *and* $0 < \delta < \infty$ *we have that*

$$\overline{\mathrm{B}_{\rho_s}(s)} \cap \left[\left(\mathrm{B}_\delta(W((A-s)^{-1})) \right)^{-1} + s \right] = \varnothing$$

where $\rho_s = \frac{1}{w((A-s)^{-1}) + \delta} \geq \frac{1}{\|(A-s)^{-1}\| + \delta}$. *Here,*

$$w((A-s)^{-1}) := \sup_{\|x\|=1} |\langle (A-s)^{-1}x, x \rangle|$$

denotes the numerical radius.

Proof. Let $s \in \varrho(A)$ and $t \in \left(\mathrm{B}_\delta(W((A-s)^{-1})) \right)^{-1} + s$. Then

$$\frac{1}{t-s} \in \mathrm{B}_\delta(W((A-s)^{-1}))$$

and we can estimate

$$\frac{1}{|t-s|} < \delta + \sup_{\|x\|=1} |\langle (A-s)^{-1}x, x \rangle| = \delta + w((A-s)^{-1}) = \frac{1}{\rho_s}.$$

This implies $|t - s| > \rho_s$ and therefore $t \notin \overline{\mathrm{B}_{\rho_s}(s)}$. $\square$

The sets $\left(\mathrm{B}_{\delta_s}(W((A-s)^{-1})) \right)^{-1}$ in Theorem 3.1.3 are unbounded whenever

$$\delta_s \geq \inf_{\|x\|=1} |\langle (A-s)^{-1}x, x \rangle| =: m((A-s)^{-1})$$

and this will always be the case if

$$\varepsilon \geq \frac{m((A-s)^{-1})}{\|(A-s)^{-1}\|(\|(A-s)^{-1}\| + m((A-s)^{-1}))}$$

due to $\delta_s \geq \frac{\|(A-s)^{-1}\|^2 \varepsilon}{1 - \|(A-s)^{-1}\|\varepsilon}$. This phenomenon is very common as we will see in the next and other examples.

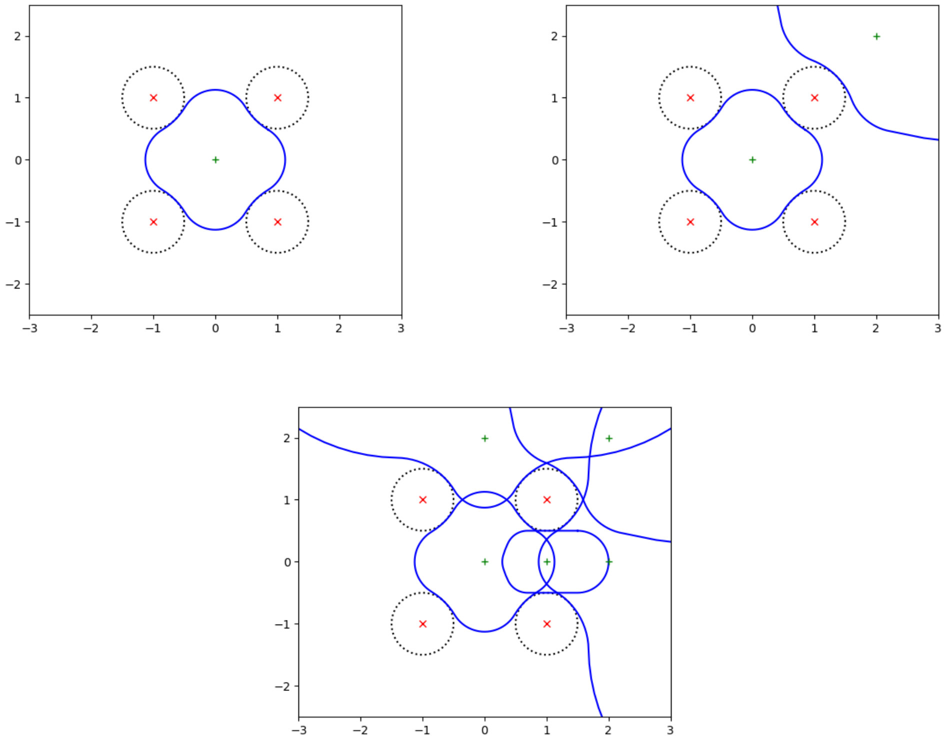


Figure 3.1: Improvement of pseudospectrum enclosure by shifts

Example 3.1.5. Let us consider the matrix

$$A = \operatorname{diag}(-1+i, -1-i, 1+i, 1-i) \in \mathbb{C}^{4\times 4}$$

from Example 3.1.2 again. Figure 3.1 depicts the eigenvalues of A as red crosses, the boundary of $\sigma_\varepsilon(A)$ with $\varepsilon = 1/2$ as dotted black lines, the shifts S as green crosses and the boundaries of the sets $\left(\mathrm{B}_{\delta_s}(W((A-s)^{-1}))\right)^{-1}$ with $\delta_s = \frac{\|(A-s)^{-1}\|^2\varepsilon}{1-\|(A-s)^{-1}\|\varepsilon}$ as blue lines. In the top left, we have $S = \{0\}$, i.e. we are in the case of Proposition 3.1.1 without any shifts. As we can see, this only enables us to exclude a neighborhood of 0 from the pseudospectrum. However, the introduction of shifts in the top right and bottom of the figure causes a drastic improvement of the enclosure.

The following theorem shows that the enclosure of the pseudospectrum in Theorem 3.1.3 becomes optimal if the shifts are chosen optimally. In particular, the enclosure can be a superset of $\sigma_\varepsilon(A)$ and a subset of $\overline{\sigma_\varepsilon(A)}$.

Theorem 3.1.6. *Let $\varepsilon > 0$ and $S_\gamma := \left\{ s \in \varrho(A) \,\middle|\, \|(A-s)^{-1}\| = \frac{1}{\varepsilon+\gamma} \right\}$ for $\gamma > 0$. Let further $\frac{\|(A-s)^{-1}\|^2\varepsilon}{1-\|(A-s)^{-1}\|\varepsilon} \leq \delta_s < \infty$. Then the following assertions hold:*

a)
$$\begin{aligned}\sigma_\varepsilon(A) &\subset \bigcap_{\gamma>0}\bigcap_{s\in S_\gamma}\left[\left(\mathrm{B}_{\delta_s}(W((A-s)^{-1}))\right)^{-1}+s\right]\\ &\subset \left\{\lambda\in\mathbb{C}\,\middle|\, \|(A-\lambda)^{-1}\|\geq\frac{1}{\varepsilon}\right\};\end{aligned}$$

b) *If ε^{-1} is not a global minimum of the norm of the resolvent of A, we have*

$$\left\{\lambda\in\mathbb{C}\,\middle|\, \|(A-\lambda)^{-1}\|\geq\frac{1}{\varepsilon}\right\}=\overline{\sigma_\varepsilon(A)};$$

c) *If ε^{-1} is not a global minimum of the norm of the resolvent of A, we have*

$$\overline{\sigma_\varepsilon(A)}=\bigcap_{\gamma>0}\bigcap_{s\in S_\gamma}\left[\left(\overline{\mathrm{B}_{\delta_s}(W((A-s)^{-1}))}\right)^{-1}+s\right];$$

d) *If A is normal with compact resolvent and $L>0$, there exists an $\varepsilon_0>0$ (depending on L) such that for all $\varepsilon<\varepsilon_0$ we have*

$$\sigma_\varepsilon(A)\cap\overline{\mathrm{B}_L(0)}=\bigcap_{\gamma>0}\bigcap_{s\in S_\gamma}\left[\left(\mathrm{B}_{\widehat{\delta_s}}(W((A-s)^{-1}))\right)^{-1}+s\right]\cap\overline{\mathrm{B}_L(0)},$$

where $\widehat{\delta_s}=\frac{\|(A-s)^{-1}\|^2\varepsilon}{1-\|(A-s)^{-1}\|\varepsilon}$.

Proof. a) The first inclusion follows from Theorem 3.1.3. In order to prove the second inclusion first note that

$$S_\gamma\cap\bigcap_{s\in S_\gamma}\left[\left(\mathrm{B}_{\delta_s}(W((A-s)^{-1}))\right)^{-1}+s\right]=\varnothing$$

for every $\gamma>0$ by Proposition 3.1.4. Hence,

$$\begin{aligned}\bigcap_{\gamma>0}\bigcap_{s\in S_\gamma}\left[\left(\mathrm{B}_{\delta_s}(W((A-s)^{-1}))\right)^{-1}+s\right]&\subset\bigcap_{\gamma>0}\mathbb{C}\setminus S_\gamma\\ &=\mathbb{C}\setminus\bigcup_{\gamma>0}S_\gamma=\mathbb{C}\setminus\left\{s\in\varrho(A)\,\middle|\,\|(A-s)^{-1}\|<\frac{1}{\varepsilon}\right\}\\ &=\left\{s\in\mathbb{C}\,\middle|\,\|(A-s)^{-1}\|\geq\frac{1}{\varepsilon}\right\}.\end{aligned}$$

b) From [6, Theorem 3.2] we have that the norm of the resolvent can only be constant on an open subset of $\varrho(A)$ at its minimum. Since by assumption ε^{-1} is not this minimum, we obtain the equality

$$\left\{\lambda\in\mathbb{C}\,\middle|\, \|(A-\lambda)^{-1}\|\geq\frac{1}{\varepsilon}\right\}=\overline{\sigma_\varepsilon(A)}.$$

c) Taking the closure in Theorem 3.1.3 yields

$$\overline{\sigma_\varepsilon(A)} \subset \bigcap_{\gamma>0} \bigcap_{s\in S_\gamma} \left[\left(\overline{\mathrm{B}_{\delta_s}(W((A-s)^{-1}))}\right)^{-1} + s\right].$$

The other inclusion can be shown as in part a) since we also have

$$S_\gamma \cap \bigcap_{s\in S_\gamma} \left[\left(\overline{\mathrm{B}_{\delta_s}(W((A-s)^{-1}))}\right)^{-1} + s\right] = \varnothing$$

for every $\gamma > 0$ as a consequence of Proposition 3.1.4.

d) By a) it suffices to show that $\lambda \in \varrho(A) \cap \overline{\mathrm{B}_L(0)}$, $\|(A-\lambda)^{-1}\| = \frac{1}{\varepsilon}$ implies $\lambda \notin \left(\mathrm{B}_{\widehat{\delta}_s}(W((A-s)^{-1}))\right)^{-1} + s$ for some $\gamma > 0$ and $s \in S_\gamma$. Let

$$\varepsilon_1 = \frac{1}{2}\min\left\{\operatorname{dist}(\mu, \sigma(A)\setminus\{\mu\}) \,\middle|\, \mu \in \sigma(A)\cap\overline{\mathrm{B}_L(0)}\right\}.$$

Since A has compact resolvent, Corollary 1.4.2 yields that the minimum exists and is positive. With

$$\varepsilon_0 = \frac{1}{2}\min\left\{\operatorname{dist}(\mu, \sigma(A)\setminus\{\mu\}) \,\middle|\, \mu \in \sigma(A)\cap\overline{\mathrm{B}_{L+3\varepsilon_1}(0)}\right\} \tag{3.6}$$

we then have $0 < \varepsilon_0 \le \varepsilon_1$. Let now $\varepsilon < \varepsilon_0$ and $\lambda \in \varrho(A)\cap\overline{\mathrm{B}_L(0)}$ with $\|(A-\lambda)^{-1}\| = \frac{1}{\varepsilon}$. Since A is normal, we get $\operatorname{dist}(\lambda,\sigma(A)) = \varepsilon$ by Corollary 1.1.8 and hence there exists a $\mu \in \sigma(A)$ such that $|\lambda-\mu| = \varepsilon$. In particular, we have $\mu \in \mathrm{B}_{L+\varepsilon_1}(0)$. Choose now $\gamma \in (0, \varepsilon_0-\varepsilon)$, i.e. $\varepsilon < \varepsilon+\gamma < \varepsilon_0$, and set

$$s = \mu + \frac{\varepsilon+\gamma}{\varepsilon}(\lambda-\mu).$$

Then $s \in \mathrm{B}_{\varepsilon_0}(\mu)$ and

$$\operatorname{dist}(s,\sigma(A)) = |\mu - s| = \varepsilon+\gamma.$$

Indeed, if $\mu' \in \sigma(A)\cap\overline{\mathrm{B}_{L+3\varepsilon_1}(0)}$ with $\mu \neq \mu'$, then $\mathrm{B}_{\varepsilon_0}(\mu)\cap\mathrm{B}_{\varepsilon_0}(\mu') = \varnothing$ and hence $|\mu'-s| > \varepsilon_0$. If $\mu' \in \sigma(A)$ and $|\mu'| > L+3\varepsilon_1$, then $\operatorname{dist}(\mu', \mathrm{B}_{\varepsilon_0}(\mu)) > \varepsilon_1$ since $\mathrm{B}_{\varepsilon_0}(\mu) \subset \mathrm{B}_{L+\varepsilon_1+\varepsilon_0}(0)$ and thus $|\mu'-s| > \varepsilon_1 \ge \varepsilon_0$. Due to $|\mu - s| < \varepsilon_0$ we therefore obtain $\operatorname{dist}(s,\sigma(A)) = |\mu-s|$ and because A is normal we can conclude

$$\|(A-s)^{-1}\| = \frac{1}{\varepsilon+\gamma}$$

by Corollary 1.1.8 again, i.e. $s \in S_\gamma$. Since

$$\begin{aligned} \frac{1}{\widehat{\delta}_s + \|(A-s)^{-1}\|} &= \left(\frac{\|(A-s)^{-1}\|}{1-\|(A-s)^{-1}\|\varepsilon}\right)^{-1} \\ &= \frac{1}{\|(A-s)^{-1}\|} - \varepsilon \\ &= \gamma, \end{aligned} \tag{3.7}$$

Proposition 3.1.4 implies

$$\overline{\mathrm{B}_\gamma(s)} \cap \Big[\big(\mathrm{B}_{\widehat{\delta_s}}(W((A-s)^{-1}))\big)^{-1} + s\Big] = \varnothing.$$

By our choice of s we have $\lambda \in \overline{\mathrm{B}_\gamma(s)}$ and thus

$$\lambda \notin \big(\mathrm{B}_{\widehat{\delta_s}}(W((A-s)^{-1}))\big)^{-1} + s. \qquad ❑$$

Remark 3.1.7. i) Note that ε_0 in part d) depends on L. For instance, if we consider an operator A with

$$\sigma(A) = \left\{ \mu_n = \sum_{k=1}^{n} \frac{1}{k} \;\middle|\; n \in \mathbb{N} \right\}$$

we have $\lim_{n\to\infty} |\mu_n - \mu_{n+1}| = 0$, $\lim_{n\to\infty} \mu_n = \infty$ and from (3.6) we obtain $\varepsilon_0 \to 0$ for $L \to \infty$.

ii) The cutoff with the large ball $\overline{\mathrm{B}_L(0)}$ in part d) is not needed in the matrix case (i.e. $\dim \mathcal{H} < \infty$), or if the eigenvalues of A satisfy a uniform gap condition. On the other hand, the equality in d) will typically not hold for all $\varepsilon > 0$, i.e. the restriction $\varepsilon < \varepsilon_0$ is needed, even in the matrix case. This is illustrated with the next (counter-)example.

Example 3.1.8. Let the normal matrix A be given by

$$A = \begin{pmatrix} 1 & 0 \\ 0 & -1 \end{pmatrix}$$

and consider $\varepsilon = 1$. Then $\sigma_\varepsilon(A) = \mathrm{B}_1(1) \cup \mathrm{B}_1(-1)$, see Corollary 1.1.8, and in particular $0 \notin \sigma_\varepsilon(A)$. See Figure 3.2 for the pseudospectrum with an enclosure. We will show now that $0 \in \big(\mathrm{B}_{\widehat{\delta_s}}(W((A-s)^{-1}))\big)^{-1} + s$ for all $s \in S_\gamma$, $\gamma > 0$, where $\widehat{\delta_s} = \frac{\|(A-s)^{-1}\|^2\varepsilon}{1-\|(A-s)^{-1}\|\varepsilon}$. Hence,

$$\sigma_\varepsilon(A) \subsetneqq \bigcap_{\gamma>0} \bigcap_{s\in S_\gamma} \Big[\big(\mathrm{B}_{\widehat{\delta_s}}(W((A-s)^{-1}))\big)^{-1} + s\Big]$$

in this case. First observe that for $s \in S_\gamma$, i.e. $\|(A-s)^{-1}\| = \frac{1}{\varepsilon+\gamma}$, we have $\frac{1}{\widehat{\delta_s}+\|(A-s)^{-1}\|} = \gamma$, see (3.7). This implies

$$\widehat{\delta_s} = \frac{1}{\gamma} - \|(A-s)^{-1}\| = \frac{1}{\gamma} - \frac{1}{\varepsilon+\gamma} = \frac{\varepsilon}{\gamma(\varepsilon+\gamma)} = \frac{1}{\gamma(1+\gamma)}$$

since $\varepsilon = 1$. We also have

$$(A-s)^{-1} = \begin{pmatrix} (1-s)^{-1} & 0 \\ 0 & (-1-s)^{-1} \end{pmatrix}$$

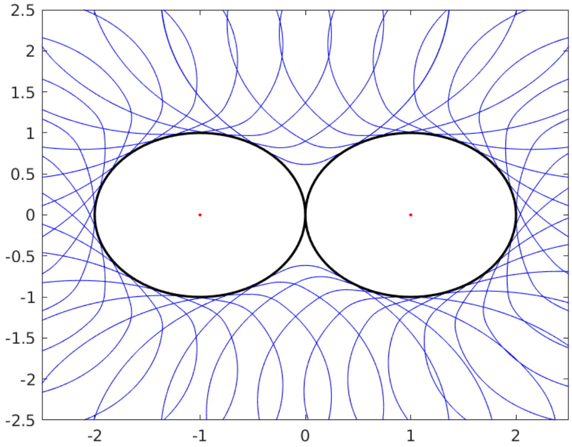


Figure 3.2: Exemplary enclosure of the 1-pseudospectrum of A from Example 3.1.8. The blue lines depict the boundaries of the sets $\big(\mathrm{B}_{\widehat{\delta}_s}(W((A-s)^{-1}))\big)^{-1}+s$ for some s in an S_γ.

and hence

$$W((A-s)^{-1}) = \big\{ r(1-s)^{-1} + (1-r)(-1-s)^{-1} \,\big|\, r \in [0,1] \big\}.$$

Due to A being normal, S_γ is the boundary of the $(1+\gamma)$-neighborhood of $\{-1,1\}$. Thus by taking $s_0 \in S_\gamma$ with $\Re s_0 = 0$ we have

$$|s|^2 \geq |s_0|^2 = (1+\gamma)^2 - 1^2 = \gamma^2 + 2\gamma$$

and hence $|s| > \gamma$, see Figure 3.3. From

$$|(\pm 1 - s)^{-1} - (-s^{-1})| = \left| \frac{1}{\pm 1 - s} + \frac{1}{s} \right| = \frac{1}{|s||\pm 1 - s|} \leq \frac{1}{|s|(1+\gamma)}$$

we get

$$\begin{aligned} &\operatorname{dist}\big(-s^{-1}, W((A-s)^{-1})\big) \\ &\quad \leq \min_{r\in[0,1]} r\left|(1-s)^{-1} - (-s^{-1})\right| + (1-r)\left|(-1-s)^{-1} - (-s)^{-1}\right| \\ &\quad \leq \frac{1}{|s|(1+\gamma)} < \frac{1}{\gamma(1+\gamma)} = \widehat{\delta}_s. \end{aligned}$$

This shows that $-s^{-1} \in \mathrm{B}_{\widehat{\delta}_s}(W((A-s)^{-1}))$ and therefore

$$0 \in \big(\mathrm{B}_{\widehat{\delta}_s}(W((A-s)^{-1}))\big)^{-1} + s.$$

Remark 3.1.9. Note that under the assumption $\sigma(A) \subset \overline{W(A)}$ (which holds for example if A has a compact resolvent by Theorem 2.2.3) it is known (see e.g.

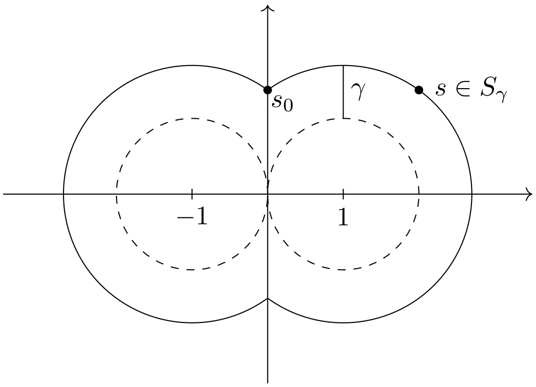


Figure 3.3: 1-pseudospectrum of A from Example 3.1.8 with S_γ

[51] for the matrix case) that the pseudospectrum can also be enclosed by an ε-neighborhood of the numerical range, namely

$$\sigma_\varepsilon(A) \subset \mathrm{B}_\varepsilon(W(A)). \tag{3.8}$$

Indeed, for $\lambda \in \sigma_\varepsilon(A) \setminus \sigma(A)$ we have $\|(A-\lambda)^{-1}\| > \frac{1}{\varepsilon}$ and therefore

$$\|(A-\lambda)x\| < \varepsilon \qquad \text{for all } x \in \mathcal{D}(A), \|x\| = 1.$$

This implies

$$|\langle Ax, x\rangle - \lambda| = |\langle (A-\lambda)x, x\rangle| \le \|(A-\lambda)x\| < \varepsilon$$

for $x \in \mathcal{D}(A)$, $\|x\| = 1$. See Remark 3.7.3 for a comparison of the enclosure (3.8) with our method (3.5).

3.2 A Strong Approximation Scheme

In this section we consider finite-dimensional approximations A_n to the full operator A. Our aim is to prove a version of Theorem 3.1.3 which provides a pseudospectrum enclosure for the full operator A in terms of numerical ranges of the approximating matrices A_n; this will allow us to compute the enclosure by numerical methods.

We suppose that $0 \in \varrho(A)$ and consider a sequence of approximations A_n of the operator A of the following form:

a) $\mathcal{U}_n \subset \mathcal{H}$, $n \in \mathbb{N}$, are finite-dimensional subspaces of the Hilbert space $\mathcal{H}$;

b) $P_n \in \mathcal{L}(\mathcal{H})$ are projections (not necessarily orthogonal) onto $\mathcal{U}_n$, i.e. $\mathcal{R}(P_n) = \mathcal{U}_n$, such that

$$\lim_{n\to\infty} P_n x = x \qquad \text{for all} \qquad x \in \mathcal{H}; \tag{3.9}$$

c) $A_n \in \mathcal{L}(\mathcal{U}_n)$ are invertible such that

$$\lim_{n\to\infty} A_n^{-1} P_n x = A^{-1} x \qquad \text{for all} \qquad x \in \mathcal{H}. \tag{3.10}$$

In this case we say that the family $(P_n, A_n)_{n\in\mathbb{N}}$ *approximates A strongly.* Note that (3.9) implies that $\bigcup_{n\in\mathbb{N}} \mathcal{U}_n$ is dense in $\mathcal{H}$ and that $\sup_{n\in\mathbb{N}} \|P_n\| < \infty$ by the uniform boundedness principle.

Lemma 3.2.1. *Let $\mathcal{U}_n$, P_n be such that* (3.9) *holds and let $A_n \in \mathcal{L}(\mathcal{U}_n)$ be invertible. Then the following assertions are equivalent:*

a) $\lim_{n\to\infty} A_n^{-1} P_n x = A^{-1} x$ *for all $x \in \mathcal{H}$, i.e.* (3.10) *holds;*

b) $\sup_{n\in\mathbb{N}} \|A_n^{-1}\|_{\mathcal{L}(\mathcal{U}_n)} < \infty$ *and for all $x \in \mathcal{D}(A)$ there exists a sequence $(x_n)_n$ with $x_n \in \mathcal{U}_n$ such that*

$$\lim_{n\to\infty} x_n = x, \qquad \lim_{n\to\infty} A_n x_n = Ax.$$

Proof. a) $\Rightarrow$ b). The uniform boundedness principle yields

$$\sup_{n\in\mathbb{N}} \|A_n^{-1} P_n\|_{\mathcal{L}(\mathcal{H})} < \infty.$$

Since $\|A_n^{-1} u\| = \|A_n^{-1} P_n u\| \le \|A_n^{-1} P_n\|_{\mathcal{L}(\mathcal{H})} \|u\|$ for all $u \in \mathcal{U}_n$, this shows the first part. For the second, let $x \in \mathcal{D}(A)$ and set $y = Ax$ and $x_n = A_n^{-1} P_n y$. Then $x_n \to A^{-1} y = x$ and $A_n x_n = P_n y \to y = Ax$ as $n \to \infty$.

b)$\Rightarrow$ a). Let $y \in \mathcal{H}$. Set $x = A^{-1} y$ and choose $x_n \in \mathcal{U}_n$ according to b). Then

$$A_n^{-1} P_n y = A_n^{-1} P_n A x = A_n^{-1} (P_n A x - A_n x_n) + x_n.$$

Since both $P_n A x \to Ax$ and $A_n x_n \to Ax$ as $n \to \infty$ and $\|A_n^{-1}\|$ is uniformly bounded, we obtain a). ❑

The following lemma shows that if A is approximated by A_n strongly, then $A - \lambda$ is approximated by $A_n - \lambda$ strongly too, provided $\|(A_n - \lambda)^{-1}\|$ is uniformly bounded in n.

Lemma 3.2.2. *Suppose that $(P_n, A_n)_{n\in\mathbb{N}}$ approximates A strongly. If $\lambda \in \varrho(A)$ is such that $\lambda \in \varrho(A_n)$ for all $n \in \mathbb{N}$ and $\sup_{n\in\mathbb{N}} \|(A_n - \lambda)^{-1}\| < \infty$, then*

$$\lim_{n\to\infty} (A_n - \lambda)^{-1} P_n x = (A - \lambda)^{-1} x \qquad \textit{for all} \qquad x \in \mathcal{H}.$$

Proof. This follows immediately from Lemma 3.2.1 since

$$\lim_{n\to\infty} A_n x_n = Ax \quad \iff \quad \lim_{n\to\infty} (A_n - \lambda) x_n = (A - \lambda) x$$

whenever $\lim_{n\to\infty} x_n = x$. ❑

Remark 3.2.3. In the literature there is a variety of notions describing the approximation of a linear operator. Two notions that are close to our definition of a strong approximation scheme are *generalized strong resolvent convergence*, considered in [3, 4, 55], and *discrete-stable convergence*, see [9]. There are however subtle differences between these two notions and our setting: First, we do *not* assume that $P_n(\mathcal{D}(A)) \subset \mathcal{D}(A_n)$. Second, in Lemma 3.2.1 b) we do not have the convergence of A_nP_nx to Ax, which would be the case for discrete-stable convergence. Up to these differences, the results of Lemmas 3.2.1 and 3.2.2 are well known in the literature, see [3, Lemma 1.2.2, Theorem 1.2.9] and [9, Lemma 3.16].

We now prove a convergence result for the numerical range of the inverse operator under strong approximations.

Lemma 3.2.4. *Suppose that $(P_n, A_n)_{n\in\mathbb{N}}$ approximates A strongly. Then the following assertions hold:*

a) *For every $x \in \mathcal{H}$ with $\|x\| = 1$ there exists a sequence $(y_n)_n$ with $y_n \in \mathcal{U}_n$ and $\|y_n\| = 1$ for all $n \in \mathbb{N}$ such that*

$$\lim_{n\to\infty} \langle A_n^{-1}y_n, y_n\rangle = \langle A^{-1}x, x\rangle;$$

b) *For all $\delta > 0$ there exists an $n_0 \in \mathbb{N}$ such that*

$$W(A^{-1}) \subset \mathrm{B}_\delta\left(W(A_n^{-1})\right) \quad \textit{for all} \quad n \geq n_0.$$

Proof. a) We set $y_n = P_nx/\|P_nx\|$. Note that y_n is well defined for almost all n since $\|P_nx\| \to \|x\| = 1$. We get $y_n \to x$ as $n \to \infty$ and

$$\begin{aligned}|\langle A^{-1}x, x\rangle - \langle A_n^{-1}y_n, y_n\rangle| &\leq |\langle A^{-1}x - A_n^{-1}P_nx, x\rangle| + |\langle A_n^{-1}P_nx, x - y_n\rangle| \\ &\quad + |\langle A_n^{-1}(P_nx - y_n), y_n\rangle| \\ &\leq \|A^{-1}x - A_n^{-1}P_nx\| + \|A_n^{-1}\|\|P_nx\|\|x - y_n\| \\ &\quad + \|A_n^{-1}\|\|P_nx - y_n\|,\end{aligned}$$

which yields the assertion.

b) Since $W(A^{-1})$ is bounded, it is precompact and hence there exist points $z_1, \dots, z_m \in W(A^{-1})$ such that

$$W(A^{-1}) \subset \bigcup_{j=1}^{m} \mathrm{B}_{\delta/2}(z_j).$$

For every $j \in \{1, \dots, m\}$ we have $z_j = \langle A^{-1}x_j, x_j\rangle$ with some $x_j \in \mathcal{H}$, $\|x_j\| = 1$, and by a) there exists $n_j \in \mathbb{N}$ such that for all $n \geq n_j$ there is a $y_j \in \mathcal{U}_n$ with $\|y_j\| = 1$ such that

$$\left|\langle A^{-1}x_j, x_j\rangle - \langle A_n^{-1}y_j, y_j\rangle\right| < \frac{\delta}{2}.$$

Hence,

$$W(A^{-1}) \subset \bigcup_{j=1}^{m} \mathrm{B}_\delta\left(\langle A_n^{-1}y_j, y_j\rangle\right) \subset \mathrm{B}_\delta\left(W(A_n^{-1})\right)$$

for all $n \geq n_0 := \max\{n_1, \ldots, n_m\}$. ❑

The previous lemma allows us easily to prove an approximation version of the basic enclosure result Proposition 3.1.1.

Proposition 3.2.5. *Suppose that $(P_n, A_n)_{n\in\mathbb{N}}$ approximates A strongly. For $0 < \varepsilon < \frac{1}{\|A^{-1}\|}$ and $\delta > \frac{\|A^{-1}\|^2\varepsilon}{1-\|A^{-1}\|\varepsilon}$ there exists an $n_0 \in \mathbb{N}$ such that*

$$\sigma_\varepsilon(A) \subset \left(\mathrm{B}_\delta(W(A_n^{-1}))\right)^{-1} \quad \textit{for all} \quad n \geq n_0.$$

Proof. By Proposition 3.1.1 we have

$$\sigma_\varepsilon(A) \subset \left(\mathrm{B}_{\widehat{\delta}}(W(A^{-1}))\right)^{-1}$$

where $\widehat{\delta} = \frac{\|A^{-1}\|^2\varepsilon}{1-\|A^{-1}\|\varepsilon}$. Since $\delta - \widehat{\delta} > 0$, Lemma 3.2.4 yields a constant $n_0 \in \mathbb{N}$ such that

$$W(A^{-1}) \subset \mathrm{B}_{\delta-\widehat{\delta}}\left(W(A_n^{-1})\right) \quad \text{for all} \quad n \geq n_0.$$

Consequently, $\mathrm{B}_{\widehat{\delta}}(W(A^{-1})) \subset \mathrm{B}_\delta(W(A_n^{-1}))$ for $n \geq n_0$ and the proof is complete. ❑

Combining the previous proposition with shifts of the operator, we get our second main result. It is analogous to Theorem 3.1.3, but provides an enclosure of the pseudospectrum of the infinite-dimensional operator in terms of numerical ranges of the approximating matrices.

Theorem 3.2.6. *Suppose that $(P_n, A_n)_{n\in\mathbb{N}}$ approximates A strongly. Let the shifts $s_1, \ldots, s_m \in \varrho(A)$ be such that*

$$\sup_{n\in\mathbb{N}} \|(A_n - s_j)^{-1}\| < \infty \quad \textit{for all} \quad j = 1, \ldots, m.$$

Let $0 < \varepsilon < \frac{1}{\max_{j=1,\ldots,m}\|(A-s_j)^{-1}\|}$ and $\delta_j > \frac{\|(A-s_j)^{-1}\|^2\varepsilon}{1-\|(A-s_j)^{-1}\|\varepsilon}$ for all $j = 1, \ldots, m$. Then there exists an $n_0 \in \mathbb{N}$ such that

$$\sigma_\varepsilon(A) \subset \bigcap_{j=1}^{m} \left[\left(\mathrm{B}_{\delta_j}(W((A_n - s_j)^{-1}))\right)^{-1} + s_j\right] \quad \textit{for all} \quad n \geq n_0.$$

Proof. In view of Lemma 3.2.2, Proposition 3.2.5 can be applied to every $A - s_j$. Hence, there exists an $n_j \in \mathbb{N}$ such that

$$\sigma_\varepsilon(A - s_j) \subset \left(\mathrm{B}_{\delta_j}(W((A_n - s_j)^{-1}))\right)^{-1} \quad \text{for all} \quad n \geq n_j.$$

Since $\sigma_\varepsilon(A) = \sigma_\varepsilon(A - s_j) + s_j$, the claim follows with

$$n_0 := \max\{n_1, \ldots, n_m\}.$$ ❑

3.3 A Uniform Approximation Scheme

In this section we pose additional assumptions on the approximations A_n of the infinite-dimensional operator A, that will allow us to estimate the starting index n_0 for which the pseudospectrum enclosures from Proposition 3.2.5 and Theorem 3.2.6 hold on bounded sets.

Throughout this section we assume that A has a compact resolvent, $0 \in \varrho(A)$ and that $\mathcal{D}(A) \subset \mathcal{W} \subset \mathcal{H}$ where the Hilbert space $\mathcal{W}$ is continuously and densely embedded into $\mathcal{H}$. The closed graph theorem then implies $A^{-1} \in \mathcal{L}(\mathcal{H}, \mathcal{W})$. Further, we suppose that there is a sequence of approximations of the operator A in the following sense:

a) $\mathcal{U}_n \subset \mathcal{H}$, $n \in \mathbb{N}$, are finite-dimensional subspaces of $\mathcal{H}$;

b) There exist projections $P_n \in \mathcal{L}(\mathcal{H})$ onto $\mathcal{U}_n$, $n \in \mathbb{N}$, not necessarily orthogonal, with $\sup_{n\in\mathbb{N}} \|P_n\| < \infty$ and $\|(I - P_n)|_{\mathcal{W}}\|_{\mathcal{L}(\mathcal{W},\mathcal{H})} \to 0$ as $n \to \infty$;

c) There exist invertible operators $A_n \in \mathcal{L}(\mathcal{U}_n)$, $n \in \mathbb{N}$, such that $\|A^{-1} - A_n^{-1} P_n\| \to 0$ as $n \to \infty$.

We say that $(P_n, A_n)_{n\in\mathbb{N}}$ *approximates A uniformly.* For $\|(I - P_n)|_{\mathcal{W}}\|_{\mathcal{L}(\mathcal{W},\mathcal{H})}$ we will write abbreviatory $\|I - P_n\|_{\mathcal{L}(\mathcal{W},\mathcal{H})}$.

Remark 3.3.1. i) Property c) already implies that A has compact resolvent: Indeed, A^{-1} is the uniform limit of the finite rank operators $A_n^{-1} P_n$ and hence compact.

ii) If $(P_n, A_n)_{n\in\mathbb{N}}$ approximates A uniformly, then also strongly. Note here that from b) we first obtain $P_n x \to x$ for $x \in \mathcal{W}$, which can then be extended to all $x \in \mathcal{H}$ by the density of $\mathcal{W}$ in $\mathcal{H}$ and the uniform boundedness of the P_n. One particular consequence of the strong approximation is

$$\sup_{n\in\mathbb{N}} \|A_n^{-1}\| < \infty,$$

see Lemma 3.2.1.

iii) Property c) amounts to the convergence of A_n to A in *generalized norm resolvent sense*, see [3, 4, 55] for this notion. Note however that our setting has the additional assumption that $P_n \to I$ *uniformly* in $\mathcal{L}(\mathcal{W}, \mathcal{H})$ where $\mathcal{D}(A) \subset \mathcal{W} \subset \mathcal{H}$. For generalized norm resolvent convergence this is not the case, but it will be a crucial element in the following proofs.

In order to obtain improved enclosures of the pseudospectrum under a uniform approximation scheme, that is, additional estimates of the starting index n_0 for which the pseudospectrum enclosures from Proposition 3.2.5 and Theorem 3.2.6 hold on bounded sets, we refine the results from Section 3.1 in terms of certain subsets of the full numerical range of A^{-1}. For $d > 0$ we define

$$W(A^{-1}, d) = \left\{ \langle A^{-1}x, x\rangle \,\middle|\, \|x\| = 1,\ x \in \mathcal{W},\ \|x\|_{\mathcal{W}} \le d \right\}. \tag{3.11}$$

Clearly, we have $W(A^{-1},d) \subset W(A^{-1})$. Moreover, since $\mathcal{W}$ is dense in $\mathcal{H}$ we get

$$\overline{\bigcup_{d>0} W(A^{-1},d)} = \overline{W(A^{-1})}. \tag{3.12}$$

Proposition 3.3.2. *Let $L > 0$ and $d = L\|A^{-1}\|_{\mathcal{L}(\mathcal{H},\mathcal{W})}$. Then the following assertions hold:*

a) $\sigma(A) \cap \overline{\mathrm{B}_L(0)} \subset W(A^{-1},d)^{-1}$*;*

b) *If in addition* $0 < \varepsilon < \frac{1}{\|A^{-1}\|}$*,* $L > \varepsilon$ *and* $\delta \geq \frac{\|A^{-1}\|^2\varepsilon}{1-\|A^{-1}\|\varepsilon}$*, we have*

$$\sigma_\varepsilon(A) \cap \overline{\mathrm{B}_{L-\varepsilon}(0)} \subset \left(\mathrm{B}_\delta(W(A^{-1},d))\right)^{-1}.$$

Proof. a) Let $\lambda \in \sigma(A)$ with $|\lambda| \leq L$. Since A has compact resolvent, there exists an $x \in \mathcal{D}(A)$ with $\|x\| = 1$ and $Ax = \lambda x$ by Corollary 1.4.2. This implies

$$\frac{1}{|\lambda|}\|x\|_{\mathcal{W}} = \|A^{-1}x\|_{\mathcal{W}} \leq \|A^{-1}\|_{\mathcal{L}(\mathcal{H},\mathcal{W})}\|x\| = \|A^{-1}\|_{\mathcal{L}(\mathcal{H},\mathcal{W})}$$

and thus we obtain

$$\|x\|_{\mathcal{W}} \leq \|A^{-1}\|_{\mathcal{L}(\mathcal{H},\mathcal{W})}|\lambda| \leq L\|A^{-1}\|_{\mathcal{L}(\mathcal{H},\mathcal{W})} = d.$$

Consequently, $\lambda^{-1} = \langle A^{-1}x, x\rangle \in W(A^{-1},d)$.

b) The proof is similar to the one of Proposition 3.1.1. We set

$$U = \left(\mathrm{B}_\delta(W(A^{-1},d))\right)^{-1}$$

and first show

$$\|(A-\lambda)x\| \geq \varepsilon \quad \text{for all} \quad \lambda \in \overline{\mathrm{B}_{L-\varepsilon}(0)} \setminus U \text{ and } x \in \mathcal{D}(A) \text{ with } \|x\| = 1. \tag{3.13}$$

Let therefore $\lambda \in \overline{\mathrm{B}_{L-\varepsilon}(0)} \setminus U$ and $x \in \mathcal{D}(A)$ with $\|x\| = 1$. We consider three cases. Suppose first that $|\lambda| > \frac{1}{\|A^{-1}\|} - \varepsilon$ and $\|x\|_{\mathcal{W}} \leq d$. Then $\lambda \neq 0$ and from $\lambda \notin U$ we obtain $\operatorname{dist}(\lambda^{-1}, W(A^{-1},d)) \geq \delta$, which implies

$$\delta \leq |\lambda^{-1} - \langle A^{-1}x,x\rangle| = |\langle(\lambda^{-1} - A^{-1})x, x\rangle| \leq \|(\lambda^{-1} - A^{-1})x\|.$$

Thus,

$$\begin{aligned}\|(A-\lambda)x\| = |\lambda|\|A(\lambda^{-1} - A^{-1})x\| &\geq \frac{|\lambda|}{\|A^{-1}\|}\|(\lambda^{-1} - A^{-1})x\| \\ &> \frac{\delta}{\|A^{-1}\|}\left(\frac{1}{\|A^{-1}\|} - \varepsilon\right) = \frac{\delta(1-\|A^{-1}\|\varepsilon)}{\|A^{-1}\|^2} \geq \varepsilon.\end{aligned}$$

In the second case assume $\|x\|_{\mathcal{W}} \geq d$. Then

$$d \leq \|x\|_{\mathcal{W}} \leq \|A^{-1}\|_{\mathcal{L}(\mathcal{H},\mathcal{W})}\|Ax\|,$$

which in view of $\lambda \in \overline{\mathrm{B}_{L-\varepsilon}(0)}$ implies

$$\|(A-\lambda)x\| \geq \|Ax\| - |\lambda| \geq \frac{d}{\|A^{-1}\|_{\mathcal{L}(\mathcal{H},\mathcal{W})}} - |\lambda| = L - |\lambda| \geq \varepsilon.$$

Finally, if $|\lambda| \leq \frac{1}{\delta+\|A^{-1}\|}$, the same reasoning as in the proof of Proposition 3.1.1 can be applied. Indeed, $|\lambda|\|A^{-1}\| \leq 1-\|A^{-1}\|\varepsilon$ yields that $I-\lambda A^{-1}$ is invertible by a Neumann series argument with $\|(I-\lambda A^{-1})^{-1}\| \leq \frac{1}{\|A^{-1}\|\varepsilon}$. For $x \in \mathcal{D}(A)$ with $\|x\| = 1$ this implies

$$\|(A-\lambda)x\| = \|A(I-\lambda A^{-1})x\| \geq \frac{1}{\|A^{-1}\|\|(I-\lambda A^{-1})^{-1}\|} \geq \varepsilon.$$

Hence, we have $\|(A-\lambda)x\| \geq \varepsilon$ once again and therefore (3.13) is proven. Now, since A has a compact resolvent, (3.13) yields that $\lambda \in \overline{\mathrm{B}_{L-\varepsilon}(0)} \setminus U$ implies $\lambda \in \varrho(A)$ with $\|(A-\lambda)^{-1}\| \leq \frac{1}{\varepsilon}$ by use of Corollary 1.4.2. Consequently, $\sigma_\varepsilon(A) \cap \overline{\mathrm{B}_{L-\varepsilon}(0)} \subset U$. ❑

From Proposition 3.3.2 we get again a shifted version:

Theorem 3.3.3. *Let $S \subset \varrho(A)$ be such that*

$$M_0 := \sup_{s\in S} \|(A-s)^{-1}\| < \infty \quad \textit{and} \quad M_1 := \sup_{s\in S} \|(A-s)^{-1}\|_{\mathcal{L}(\mathcal{H},\mathcal{W})} < \infty.$$

For $0 < \varepsilon < \frac{1}{M_0}$, $L > \varepsilon$, $d = LM_1$ and $\delta_s \geq \frac{\|(A-s)^{-1}\|^2\varepsilon}{1-\|(A-s)^{-1}\|\varepsilon}$ we get the inclusion

$$\sigma_\varepsilon(A) \cap \bigcap_{s\in S} \overline{\mathrm{B}_{L-\varepsilon}(s)} \subset \bigcap_{s\in S} \left[\left(\mathrm{B}_{\delta_s}(W((A-s)^{-1},d))\right)^{-1} + s\right].$$

Proof. Apply Proposition 3.3.2 b) to $A-s$ for all $s \in S$ and note that

$$\lambda \in \sigma_\varepsilon(A-s) \cap \overline{\mathrm{B}_{L-\varepsilon}(0)}$$

if and only if

$$\lambda + s \in \sigma_\varepsilon(A) \cap \overline{\mathrm{B}_{L-\varepsilon}(s)}.$$ ❑

Remark 3.3.4. By the continuity of the embedding $\mathcal{W} \hookrightarrow \mathcal{H}$, the condition $M_1 < \infty$ already implies $M_0 < \infty$.

For a uniform approximation scheme, the numerical range of A^{-1} can now be approximated with explicit control on the starting index n_0:

Lemma 3.3.5. *Suppose that $(P_n, A_n)_{n\in\mathbb{N}}$ approximates A uniformly and set*

$$C_0 = \sup_{n\in\mathbb{N}} \left(\|A_n^{-1}\|\|P_n\| + 6\|A_n^{-1}\|\|P_n\|^2\right). \tag{3.14}$$

Then the following assertions hold:

a) *If $d > 0$, $0 < \delta \le \frac{C_0}{2}$ and $n_0 \in \mathbb{N}$ are such that for every $n \ge n_0$*

$$\|A^{-1} - A_n^{-1}P_n\| + dC_0\|I - P_n\|_{\mathcal{L}(\mathcal{W},\mathcal{H})} < \delta,$$

we have

$$W(A^{-1}, d) \subset \mathrm{B}_\delta(W(A_n^{-1})) \quad \textit{for all} \quad n \ge n_0;$$

b) *If $\delta > 0$ and $n_0 \in \mathbb{N}$ are such that for every $n \ge n_0$ we have $\|A^{-1} - A_n^{-1}P_n\| < \delta$, then*

$$W(A_n^{-1}) \subset \mathrm{B}_\delta(W(A^{-1})) \quad \textit{for all} \quad n \ge n_0.$$

Proof. Let $x \in \mathcal{W}$ with $\|x\| = 1$ and $\|x\|_\mathcal{W} \le d$. Then we obtain

$$\begin{aligned}
&|\langle A^{-1}x, x\rangle - \langle A_n^{-1}P_nx, P_nx\rangle| \\
&\qquad \le |\langle A^{-1}x - A_n^{-1}P_nx, x\rangle| + |\langle A_n^{-1}P_nx, x - P_nx\rangle| \\
&\qquad \le \|A^{-1} - A_n^{-1}P_n\|\|x\|^2 + \|A_n^{-1}\|\|P_n\|\|x\|\|I - P_n\|_{\mathcal{L}(\mathcal{W},\mathcal{H})}\|x\|_\mathcal{W} \\
&\qquad \le \|A^{-1} - A_n^{-1}P_n\| + d\|A_n^{-1}\|\|P_n\|\|I - P_n\|_{\mathcal{L}(\mathcal{W},\mathcal{H})}
\end{aligned}$$

as well as

$$\begin{aligned}
|1 - \|P_nx\|| \le \|x - P_nx\| &\le \|I - P_n\|_{\mathcal{L}(\mathcal{W},\mathcal{H})}\|x\|_\mathcal{W} \\
&\le d\|I - P_n\|_{\mathcal{L}(\mathcal{W},\mathcal{H})}.
\end{aligned}$$

Let $n \ge n_0$. Then

$$\begin{aligned}
|1 - \|P_nx\|| \le d\|I - P_n\|_{\mathcal{L}(\mathcal{W},\mathcal{H})} &< \frac{\delta}{C_0} \\
&\le \frac{1}{2}
\end{aligned}$$

and hence $\|P_nx\| \ge \frac{1}{2}$. Let $x_n = \frac{P_nx}{\|P_nx\|}$. Then $\|x_n\| = 1$ and

$$\begin{aligned}
\left|1 - \frac{1}{\|P_nx\|^2}\right| = \left|\frac{\|P_nx\|^2 - 1}{\|P_nx\|^2}\right| &= \frac{(\|P_nx\| + 1)|\|P_nx\| - 1|}{\|P_nx\|^2} \\
&= \left(\frac{1}{\|P_nx\|} + \frac{1}{\|P_nx\|^2}\right)|1 - \|P_nx\|| \\
&\le 6|1 - \|P_nx\|| \\
&\le 6d\|I - P_n\|_{\mathcal{L}(\mathcal{W},\mathcal{H})}.
\end{aligned}$$

This implies

$$\begin{aligned}
|\langle A_n^{-1}P_nx, P_nx\rangle - \langle A_n^{-1}x_n, x_n\rangle| &= \left|\langle A_n^{-1}P_nx, P_nx\rangle - \frac{\langle A_n^{-1}P_nx, P_nx\rangle}{\|P_nx\|^2}\right| \\
&= \left|1 - \frac{1}{\|P_nx\|^2}\right| |\langle A_n^{-1}P_nx, P_nx\rangle| \\
&\le 6d\|I - P_n\|_{\mathcal{L}(\mathcal{W},\mathcal{H})}\|A_n^{-1}\|\|P_n\|^2,
\end{aligned}$$

and thus for $n \geq n_0$ we arrive at

$$\begin{aligned}
&|\langle A^{-1}x, x\rangle - \langle A_n^{-1}x_n, x_n\rangle| \\
&\qquad \leq \|A^{-1} - A_n^{-1}P_n\| + d\|I - P_n\|_{\mathcal{L}(\mathcal{W},\mathcal{H})}(\|A_n^{-1}\|\|P_n\| + 6\|A_n^{-1}\|\|P_n\|^2) \\
&\qquad \leq \|A^{-1} - A_n^{-1}P_n\| + dC_0\|I - P_n\|_{\mathcal{L}(\mathcal{W},\mathcal{H})} \\
&\qquad < \delta.
\end{aligned}$$

This yields $\langle A^{-1}x, x\rangle \in \mathrm{B}_\delta(W(A_n^{-1}))$ if $n \geq n_0$ and proves a).

In order to show part b), let $x \in \mathcal{U}_n$ with $\|x\| = 1$. As $x = P_n x$ we have

$$\begin{aligned}
|\langle A_n^{-1}x, x\rangle - \langle A^{-1}x, x\rangle| \leq \|A_n^{-1}x - A^{-1}x\|\|x\| &= \|A_n^{-1}P_n x - A^{-1}x\| \\
&\leq \|A^{-1} - A_n^{-1}P_n\|.
\end{aligned}$$

Thus, $\langle A_n^{-1}x, x\rangle \in \mathrm{B}_\delta(W(A^{-1}))$ for $n \geq n_0$. ❑

Corollary 3.3.6. *If $(P_n, A_n)_{n\in\mathbb{N}}$ approximates A uniformly, then*

$$\overline{W(A^{-1})} = \left\{\lambda \in \mathbb{C} \,\middle|\, \exists(\lambda_n)_{n\in\mathbb{N}} \text{ with } \lambda_n \in W(A_n^{-1}) \text{ and } \lim_{n\to\infty} \lambda_n = \lambda\right\}$$

or, equivalently,

$$\overline{W(A^{-1})} = \bigcap_{m\in\mathbb{N}} \overline{\bigcup_{n\geq m} W(A_n^{-1})}.$$

Proof. We first show the inclusion "$\supset$". Let $(\lambda_n)_{n\in\mathbb{N}}$ be a convergent sequence in $\mathbb{C}$ with $\lambda_n \in W(A_n^{-1})$ and define $\lambda = \lim_{n\to\infty} \lambda_n$. Let $\delta > 0$ be arbitrary. Lemma 3.3.5 b) implies that there exists an $n_0 \in \mathbb{N}$ such that $\lambda_n \in \mathrm{B}_\delta(W(A^{-1}))$ for every $n \geq n_0$. This implies $\lambda \in \mathrm{B}_\delta(W(A^{-1}))$ for every $\delta > 0$, and thus $\lambda \in \overline{W(A^{-1})}$.

Conversely, let $\lambda \in W(A^{-1}, d)$ for some $d > 0$. Using Lemma 3.3.5 a), we can construct a sequence $(\lambda_n)_{n\in\mathbb{N}}$ in $\mathbb{C}$ with $\lambda_n \in W(A_n^{-1})$ and $\lambda = \lim_{n\to\infty} \lambda_n$. The statement now follows from (3.12). ❑

The last result shows that $\overline{W(A^{-1})}$ can be represented as the pointwise limit of the finite-dimensional numerical ranges $W(A_n^{-1})$. Lemma 3.3.5 even yields a uniform approximation, but this is asymmetric, since one inclusion only holds for the restricted numerical range $W(A^{-1}, d)$. A more symmetric result is discussed in the next remark:

Remark 3.3.7. If $\mathcal{U}_n \subset \mathcal{W}$ for some $n \in \mathbb{N}$ then, due to the fact that the space $\mathcal{U}_n$ is finite-dimensional,

$$d_n := \sup_{x\in\mathcal{U}_n} \frac{\|x\|_\mathcal{W}}{\|x\|} < \infty.$$

Using the same reasoning as in the proof of Lemma 3.3.5 b), we then obtain

$$W(A_n^{-1}) \subset \mathrm{B}_\delta(W(A^{-1}, d_n))$$

if $\|A^{-1} - A_n^{-1}P_n\| < \delta$.

Note however that for finite element discretization schemes the condition $\mathcal{U}_n \subset \mathcal{W}$ will usually *not* be fulfilled. In our examples for instance, $\mathcal{U}_n$ are piecewise linear finite elements while $\mathcal{W} \subset H^2(\Omega)$ is a second order Sobolev space, and thus $\mathcal{U}_n \not\subset \mathcal{W}$.

Under a uniform approximation scheme the pseudospectrum can be approximated as follows.

Proposition 3.3.8. *Suppose that $(P_n, A_n)_{n\in\mathbb{N}}$ approximates A uniformly. Let $r > 0$, $0 < \varepsilon < \frac{1}{\|A^{-1}\|}$ and*

$$\frac{\|A^{-1}\|^2\varepsilon}{1-\|A^{-1}\|\varepsilon} < \delta \leq \frac{\|A^{-1}\|^2\varepsilon}{1-\|A^{-1}\|\varepsilon} + \frac{7}{2}\|A^{-1}\|.$$

If we choose $n_0 \in \mathbb{N}$ such that for every $n \geq n_0$

$$\|A^{-1} - A_n^{-1}P_n\| + (r+\varepsilon)\|A^{-1}\|_{\mathcal{L}(\mathcal{H},\mathcal{W})}C_0\|I - P_n\|_{\mathcal{L}(\mathcal{W},\mathcal{H})} < \delta - \frac{\|A^{-1}\|^2\varepsilon}{1-\|A^{-1}\|\varepsilon},$$

where C_0 is defined in (3.14), *we obtain*

$$\sigma_\varepsilon(A) \cap \overline{\mathrm{B}_r(0)} \subset \big(\mathrm{B}_\delta(W(A_n^{-1}))\big)^{-1} \quad \textit{for all} \quad n \geq n_0.$$

Proof. Let $\widehat{\delta} = \frac{\|A^{-1}\|^2\varepsilon}{1-\|A^{-1}\|\varepsilon}$, $L = r + \varepsilon$ and $d = L\|A^{-1}\|_{\mathcal{L}(\mathcal{H},\mathcal{W})}$. Proposition 3.3.2 implies

$$\sigma_\varepsilon(A) \cap \overline{\mathrm{B}_r(0)} \subset (\mathrm{B}_{\widehat{\delta}}(W(A^{-1}, d)))^{-1}.$$

Next, note that

$$\begin{aligned}\delta - \widehat{\delta} &\leq \frac{7}{2}\|A^{-1}\| = \lim_{n\to\infty}\frac{7}{2}\|A_n^{-1}P_n\| \\ &\leq \frac{1}{2}\limsup_{n\to\infty}\big(\|A_n^{-1}\|\|P_n\| + 6\|A_n^{-1}\|\|P_n\|^2\big) \\ &\leq \frac{C_0}{2},\end{aligned}$$

because P_n is a projection. We can therefore apply Lemma 3.3.5 with δ replaced by $\delta - \widehat{\delta}$ and n_0 chosen as stated above and obtain

$$W(A^{-1}, d) \subset \mathrm{B}_{\delta-\widehat{\delta}}(W(A_n^{-1})) \quad \text{for} \quad n \geq n_0$$

and hence the assertion. ❑

3.4 Finite Element Discretization

As an example for a uniform approximation scheme defined in Section 3.3 we now consider finite element discretizations. We use the standard textbook approach via form methods, which can be found e.g. in [1, 48].

Let $\mathcal{V}$ and $\mathcal{H}$ be Hilbert spaces with $\mathcal{V} \subset \mathcal{H}$ densely and continuously embedded. In particular, there is a constant $c > 0$ such that

$$\|x\| \leq c\|x\|_{\mathcal{V}} \quad \text{for all} \quad x \in \mathcal{V}. \tag{3.15}$$

Moreover, we consider a bounded and coercive sesqui-linear form $a\colon \mathcal{V} \times \mathcal{V} \to \mathbb{C}$, that is, there exist constants $M, \gamma > 0$ such that

$$|a(x,y)| \leq M\|x\|_{\mathcal{V}}\|y\|_{\mathcal{V}} \qquad \text{and} \qquad \Re a(x,x) \geq \gamma\|x\|_{\mathcal{V}}^2 \tag{3.16}$$

for all $x, y \in \mathcal{V}$. Let $A\colon \mathcal{D}(A) \subset \mathcal{H} \to \mathcal{H}$ be the operator associated with a, which is given by

$$\begin{aligned} \mathcal{D}(A) &= \big\{x \in \mathcal{V} \,\big|\, \exists c_x > 0 : |a(x,y)| \leq c_x\|y\| \text{ for } y \in \mathcal{V}\big\}, \\ \langle Ax, y\rangle &= a(x,y) \text{ for all } x \in D(A) \text{ and } y \in V. \end{aligned}$$

Another way of phrasing it is that $x \in \mathcal{D}(A)$ if and only if $x \in \mathcal{V}$ and there exists an $f \in \mathcal{H}$ with $a(x,y) = \langle f, y\rangle$ for every $y \in \mathcal{V}$. In this case we have $Ax = f$. Then A is a densely defined, closed operator with $0 \in \varrho(A)$ and $\|A^{-1}\| \leq \frac{c^2}{\gamma}$, where $c > 0$ is the constant from (3.15).

Let $(\mathcal{U}_n)_{n\in\mathbb{N}}$ be a sequence of finite-dimensional subspaces of $\mathcal{V}$ which are nested, that is $\mathcal{U}_n \subset \mathcal{U}_{n+1}$. We denote by $a_n = a|_{\mathcal{U}_n}$ the restriction of a from $\mathcal{V}$ to $\mathcal{U}_n$. The form a_n is again bounded and coercive with the same constants M and γ. Let $A_n \in \mathcal{L}(\mathcal{U}_n)$ be the operator associated with a_n, i.e.

$$a_n(x,y) = \langle A_n x, y\rangle \quad \text{for} \quad x, y \in \mathcal{U}_n.$$

Then again $0 \in \varrho(A_n)$ and $\|A_n^{-1}\| \leq \frac{c^2}{\gamma}$. Let $P_n \in \mathcal{L}(\mathcal{H})$ be the orthogonal projection onto $\mathcal{U}_n$. Thus, $\|P_n\| = 1$ and $A_n = P_n A_{n+1}|_{\mathcal{U}_n}$, that is, A_n is a compression of A_{n+1}.

To obtain a uniform approximation scheme, we now consider an additional Hilbert space $\mathcal{W}$ which is densely and continuously embedded into $\mathcal{H}$ such that $\mathcal{D}(A) \subset \mathcal{W} \subset \mathcal{V}$. We assume that there exists a sequence of operators $Q_n \in \mathcal{L}(\mathcal{W}, \mathcal{V})$ with $\mathcal{R}(Q_n) \subset \mathcal{U}_n$ and

$$\lim_{n\to\infty} \|I - Q_n\|_{\mathcal{L}(\mathcal{W},\mathcal{V})} = 0. \tag{3.17}$$

Lemma 3.4.1. *For all $n \in \mathbb{N}$ the estimates*

$$\begin{aligned} \|I - P_n\|_{\mathcal{L}(\mathcal{W},\mathcal{H})} &\leq c\|I - Q_n\|_{\mathcal{L}(\mathcal{W},\mathcal{V})}, \\ \|A^{-1} - A_n^{-1}P_n\| &\leq \frac{cM}{\gamma}\|A^{-1}\|_{\mathcal{L}(\mathcal{H},\mathcal{W})}\|I - Q_n\|_{\mathcal{L}(\mathcal{W},\mathcal{V})} \end{aligned}$$

hold. In particular, the family $(P_n, A_n)_{n\in\mathbb{N}}$ approximates A uniformly.

Proof. For $w \in \mathcal{W}$ we calculate

$$\begin{aligned} \|w - P_n w\| = \inf_{u\in\mathcal{U}_n} \|w - u\| \leq \|w - Q_n w\| &\leq c\|w - Q_n w\|_{\mathcal{V}} \\ &\leq c\|I - Q_n\|_{\mathcal{L}(\mathcal{W},\mathcal{V})}\|w\|_{\mathcal{W}}, \end{aligned}$$

which shows the first assertion. Moreover, for $f \in \mathcal{H}$ we set $x = A^{-1}f$ and $x_n = A_n^{-1}P_nf$ and obtain

$$\begin{aligned} a(x, y) &= \langle Ax, y\rangle = \langle f, y\rangle \quad \text{for all} \quad y \in \mathcal{V}, \\ a(x_n, u) &= \langle A_nx_n, u\rangle = \langle P_nf, u\rangle = \langle f, u\rangle \quad \text{for all} \quad u \in \mathcal{U}_n. \end{aligned}$$

Using the Lemma of Cea [48, Theorem VII.5.A], we find

$$\begin{aligned} \|A^{-1}f - A_n^{-1}P_nf\| = \|x - x_n\| \le c\|x - x_n\|_{\mathcal{V}} &\le \frac{cM}{\gamma} \inf_{u\in\mathcal{U}_n} \|x - u\|_{\mathcal{V}} \\ &\le \frac{cM}{\gamma}\|x - Q_nx\|_{\mathcal{V}} \le \frac{cM}{\gamma}\|I - Q_n\|_{\mathcal{L}(\mathcal{W},\mathcal{V})}\|x\|_{\mathcal{W}} \\ &\le \frac{cM}{\gamma}\|I - Q_n\|_{\mathcal{L}(\mathcal{W},\mathcal{V})}\|A^{-1}\|_{\mathcal{L}(\mathcal{H},\mathcal{W})}\|f\|, \end{aligned}$$

which implies the second assertion. ❑

Theorem 3.4.2. *Let A be the operator associated with the coercive form a and let A_n, Q_n be as above. Moreover, let $r > 0$, $0 < \varepsilon < \frac{1}{\|A^{-1}\|}$ and*

$$\frac{\|A^{-1}\|^2\varepsilon}{1 - \|A^{-1}\|\varepsilon} < \delta \le \frac{\|A^{-1}\|^2\varepsilon}{1 - \|A^{-1}\|\varepsilon} + \frac{7}{2}\|A^{-1}\|.$$

If $n_0 \in \mathbb{N}$ is such that for every $n \ge n_0$

$$\|I - Q_n\|_{\mathcal{L}(\mathcal{W},\mathcal{V})} < \frac{\delta - \frac{\|A^{-1}\|^2\varepsilon}{1-\|A^{-1}\|\varepsilon}}{c\|A^{-1}\|_{\mathcal{L}(\mathcal{H},\mathcal{W})}\left(\frac{M}{\gamma} + (r + \varepsilon)\frac{7c^2}{\gamma}\right)},$$

then

$$\sigma_\varepsilon(A) \cap \overline{\mathrm{B}_r(0)} \subset \left(\mathrm{B}_\delta(W(A_n^{-1}))\right)^{-1} \quad \textit{for all} \quad n \ge n_0.$$

Proof. We check that the conditions of Proposition 3.3.8 are satisfied: Using Lemma 3.4.1, we estimate for $n \ge n_0$ and with C_0 from (3.14),

$$\begin{aligned} &\|A^{-1} - A_n^{-1}P_n\| + (r + \varepsilon)\|A^{-1}\|_{\mathcal{L}(\mathcal{H},\mathcal{W})}C_0\|I - P_n\|_{\mathcal{L}(\mathcal{W},\mathcal{H})} \\ &\qquad\qquad \le c\|A^{-1}\|_{\mathcal{L}(\mathcal{H},\mathcal{W})}\|I - Q_n\|_{\mathcal{L}(\mathcal{W},\mathcal{V})}\left(\frac{M}{\gamma} + (r + \varepsilon)\frac{7c^2}{\gamma}\right) \\ &\qquad\qquad < \delta - \frac{\|A^{-1}\|^2\varepsilon}{1 - \|A^{-1}\|\varepsilon}. \end{aligned}$$

❑

Example 3.4.3 (Finite element discretization of elliptic partial differential operators)**.** Let $\Omega \subset \mathbb{R}^2$ be a bounded, open, convex domain with polygonal boundary Γ and $\Gamma_D \subset \Gamma$ a union of polygons of Γ. Let

$$\mathcal{V} := H_0^1(\Omega),$$

equipped with the H^1-norm. On $\mathcal{V}$ we consider the sesqui-linear form

$$a(u,v) = \int_\Omega \left(\sum_{i,j=1}^{2} a_{ij} u_{x_i} \overline{v}_{x_j} + \sum_{i=1}^{2} b_i u_{x_i} \overline{v} + cu\overline{v} \right) dx, \tag{3.18}$$

where $a_{ij} \in C^{0,1}(\overline{\Omega})$ and $b_i, c \in L^\infty(\Omega)$. We suppose that a is coercive and uniformly elliptic. Let $\{\mathcal{T}_n\}_{n\in\mathbb{N}}$ be a family of nested, admissible and quasi-uniform triangulations of Ω satisfying $\sup_{T\in\mathcal{T}_n} \operatorname{diam}(T) \le \frac{1}{n}$. Furthermore, let

$$\mathcal{W} := H^2(\Omega) \cap H_0^1(\Omega),$$

be equipped with the H^2-norm, and

$$\mathcal{U}_n := \left\{ u \in C^0(\overline{\Omega}) \,\middle|\, u|_T \in \mathbb{P}_1(T), T \in \mathcal{T}_n, u|_\Gamma = 0 \right\} \quad \text{for} \quad n \in \mathbb{N}.$$

Here, $\mathbb{P}_1(T)$ denotes the set of polynomials of degree 1 on the triangle T. We get $\mathcal{U}_n \subset \mathcal{V}$. Moreover, the operator A associated with a is given by

$$\begin{aligned} Au &= -\sum_{i,j=1}^{2} \partial_{x_j}(a_{ij} u_{x_i}) + \sum_{i=1}^{2} b_i u_{x_i} + cu, \\ \mathcal{D}(A) &= \mathcal{W}. \end{aligned}$$

For the proof of $\mathcal{D}(A) = \mathcal{W}$ we refer to [21, Theorem 3.2.1.2 and §2.4.2].

By the Sobolev embedding theorem we have $H^2(\Omega) \hookrightarrow C^0(\overline{\Omega})$. For $u \in \mathcal{W}$ we define $Q_n u$ as the unique element of $\mathcal{U}_n$ satisfying $(Q_n u)(x) = u(x)$ for every vertex of the triangulation $\mathcal{T}_n$. Then $Q_n \in \mathcal{L}(\mathcal{W},\mathcal{V})$ with $\mathcal{R}(Q_n) \subset \mathcal{U}_n$. Moreover, [1, Theorem 9.27] implies that there is a constant $K > 0$ such that

$$\|I - Q_n\|_{\mathcal{L}(\mathcal{W},\mathcal{V})} \le \frac{K}{n}, \qquad n \in \mathbb{N}.$$

We conclude that Theorem 3.4.2 can be applied in this example with $n_0 \in \mathbb{N}$ chosen such that

$$n_0 > \frac{Kc\|A^{-1}\|_{\mathcal{L}(\mathcal{H},\mathcal{W})} \left(\frac{M}{\gamma} + (r+\varepsilon)\frac{7c^2}{\gamma} \right)}{\delta - 2\|A^{-1}\|^2 \varepsilon}.$$

Note, that in Example 3.4.3 we can also consider Ω to be an open interval in $\mathbb{R}$. All results continue to hold in an analogous way.

3.5 Discretization of Structured Block Operator Matrices I

In this section we investigate discretizations of a certain kind of block operator matrices. Let $\mathcal{H}_1$ and $\mathcal{H}_2$ be Hilbert spaces with inner products $\langle \cdot, \cdot \rangle_{\mathcal{H}_1}$ and $\langle \cdot, \cdot \rangle_{\mathcal{H}_2}$

respectively such that $\mathcal{H} = \mathcal{H}_1 \times \mathcal{H}_2$ and consider block operator matrices of the form

$$\mathscr{A} = \begin{bmatrix} A & B \\ C & D \end{bmatrix},$$

where A is a closed, densely defined linear operator $A\colon \mathcal{D}(A) \subset \mathcal{H}_1 \to \mathcal{H}_1$, $B \in \mathcal{L}(\mathcal{H}_2, \mathcal{H}_1)$, $C \in \mathcal{L}(\mathcal{H}_1, \mathcal{H}_2)$ and $D \in \mathcal{L}(\mathcal{H}_2)$. Then the block operator matrix $\mathscr{A}$ is a closed, densely defined operator on the product space $\mathcal{H}_1 \times \mathcal{H}_2$ with domain $\mathcal{D}(\mathscr{A}) = \mathcal{D}(A) \times \mathcal{H}_2$. Additionally, we assume that $0 \in \varrho(A)$, $0 \in \varrho(D)$ and that both A and $-D$ are *uniformly accretive*, i.e. there exist constants $\gamma_A, \gamma_D > 0$ such that

$$\Re\langle Ax, x\rangle_{\mathcal{H}_1} \geq \gamma_A \|x\|^2_{\mathcal{H}_1} \quad \text{for all} \quad x \in \mathcal{D}(A), \tag{3.19}$$

$$\Re\langle Dy, y\rangle_{\mathcal{H}_2} \leq -\gamma_D \|y\|^2_{\mathcal{H}_2} \quad \text{for all} \quad y \in \mathcal{H}_2. \tag{3.20}$$

Moreover, we either suppose

$$\frac{1}{4}(\|B^*\| + \|C\|)^2 < \gamma_A \gamma_D \tag{3.21}$$

or

$$C = B^*. \tag{3.22}$$

Note, that (3.21) implies the existence of an $\eta > 0$ such that

$$\gamma_{A_\eta} := \gamma_A - \frac{1}{2\eta}(\|B^*\| + \|C\|) > 0,$$
$$\gamma_{D_\eta} := \gamma_D - \frac{\eta}{2}(\|B^*\| + \|C\|) > 0.$$

In the next lemma we show that under the above assumptions there is a gap in the spectrum of $\mathscr{A}$ along the imaginary axis, and we also prove an estimate for the norm of the resolvent. Similar results were obtained in [37, 38] under the additional assumption that A is sectorial and, in [38], without the condition that B and D are bounded. However, no corresponding resolvent estimates were shown. We remark that the boundedness of D is not essential in Lemma 3.5.1 but will be used thereafter.

Lemma 3.5.1. a) *If* (3.21) *holds, we have*

$$\left\{\lambda \in \mathbb{C} \,\middle|\, -\gamma_{D_\eta} < \Re\lambda < \gamma_{A_\eta}\right\} \subset \varrho(\mathscr{A})$$

and

$$\|(\mathscr{A} - \lambda)^{-1}\| \leq \frac{1}{\min\left\{\gamma_{A_\eta} - \Re\lambda, \gamma_{D_\eta} + \Re\lambda\right\}}$$

if $-\gamma_{D_\eta} < \Re\lambda < \gamma_{A_\eta}$.

b) *If* (3.22) *holds, we have* $\left\{\lambda \in \mathbb{C} \,\middle|\, -\gamma_D < \Re\lambda < \gamma_A\right\} \subset \varrho(\mathscr{A})$ *and*

$$\|(\mathscr{A} - \lambda)^{-1}\| \leq \frac{1}{\min\{\gamma_A - \Re\lambda, \gamma_D + \Re\lambda\}}$$

if $-\gamma_D < \Re\lambda < \gamma_A$.

Proof. Consider the block operator matrix

$$J = \begin{bmatrix} I & 0 \\ 0 & -I \end{bmatrix}.$$

a) For $\lambda \in U := \{\lambda \in \mathbb{C} \mid -\gamma_{D_\eta} < \Re\lambda < \gamma_{A_\eta}\}$, $x \in \mathcal{D}(A)$ and $y \in \mathcal{H}_2$ we estimate

$$\begin{aligned}
&\Re \left\langle J(\mathscr{A} - \lambda) \begin{bmatrix} x \\ y \end{bmatrix}, \begin{bmatrix} x \\ y \end{bmatrix} \right\rangle \\
&\qquad = \Re\big(\langle (A-\lambda)x, x\rangle_{\mathcal{H}_1} + \langle y, B^*x\rangle_{\mathcal{H}_2} - \langle Cx, y\rangle_{\mathcal{H}_2} - \langle (D-\lambda)y, y\rangle_{\mathcal{H}_2}\big) \\
&\qquad \geq (\gamma_A - \Re\lambda)\|x\|_{\mathcal{H}_1}^2 - \|y\|_{\mathcal{H}_2}\|B^*x\|_{\mathcal{H}_2} - \|Cx\|_{\mathcal{H}_2}\|y\|_{\mathcal{H}_2} \\
&\qquad\quad + (\gamma_D + \Re\lambda)\|y\|_{\mathcal{H}_2}^2 \\
&\qquad \geq (\gamma_A - \Re\lambda)\|x\|_{\mathcal{H}_1}^2 - (\|B^*\| + \|C\|)\|x\|_{\mathcal{H}_1}\|y\|_{\mathcal{H}_2} \\
&\qquad\quad + (\gamma_D + \Re\lambda)\|y\|_{\mathcal{H}_2}^2 \\
&\qquad \geq (\gamma_A - \Re\lambda)\|x\|_{\mathcal{H}_1}^2 - (\|B^*\| + \|C\|)\left(\frac{1}{2\eta}\|x\|_{\mathcal{H}_1}^2 + \frac{\eta}{2}\|y\|_{\mathcal{H}_2}^2\right) \\
&\qquad\quad + (\gamma_D + \Re\lambda)\|y\|_{\mathcal{H}_2}^2 \\
&\qquad \geq c_\lambda \left\| \begin{bmatrix} x \\ y \end{bmatrix} \right\|^2,
\end{aligned}$$

where $c_\lambda = \min\{\gamma_{A_\eta} - \Re\lambda, \gamma_{D_\eta} + \Re\lambda\}$.

b) For $\lambda \in U := \{\lambda \in \mathbb{C} \mid -\gamma_D < \Re\lambda < \gamma_A\}$, $x \in \mathcal{D}(A)$ and $y \in \mathcal{H}_2$ we estimate

$$\begin{aligned}
&\Re \left\langle J(\mathscr{A} - \lambda) \begin{bmatrix} x \\ y \end{bmatrix}, \begin{bmatrix} x \\ y \end{bmatrix} \right\rangle \\
&\qquad = \Re\big(\langle (A-\lambda)x, x\rangle_{\mathcal{H}_1} + \langle By, x\rangle_{\mathcal{H}_1} - \langle B^*x, y\rangle_{\mathcal{H}_2} - \langle (D-\lambda)y, y\rangle_{\mathcal{H}_2}\big) \\
&\qquad = \Re\langle (A-\lambda)x, x\rangle_{\mathcal{H}_1} - \Re\langle (D-\lambda)y, y\rangle_{\mathcal{H}_2} \\
&\qquad \geq (\gamma_A - \Re\lambda)\|x\|_{\mathcal{H}_1}^2 + (\gamma_D + \Re\lambda)\|y\|_{\mathcal{H}_2}^2 \\
&\qquad \geq c_\lambda \left\| \begin{bmatrix} x \\ y \end{bmatrix} \right\|^2,
\end{aligned}$$

where $c_\lambda = \min\{\gamma_A - \Re\lambda, \gamma_D + \Re\lambda\}$.

In both cases, it follows that

$$\left\| J(\mathscr{A} - \lambda) \begin{bmatrix} x \\ y \end{bmatrix} \right\| \left\| \begin{bmatrix} x \\ y \end{bmatrix} \right\| \geq \left| \left\langle J(\mathscr{A} - \lambda) \begin{bmatrix} x \\ y \end{bmatrix}, \begin{bmatrix} x \\ y \end{bmatrix} \right\rangle \right| \geq c_\lambda \left\| \begin{bmatrix} x \\ y \end{bmatrix} \right\|^2$$

for all $\begin{bmatrix} x \\ y \end{bmatrix} \in \mathcal{D}(\mathscr{A})$ and therefore

$$\left\| (\mathscr{A} - \lambda) \begin{bmatrix} x \\ y \end{bmatrix} \right\| \geq c_\lambda \left\| \begin{bmatrix} x \\ y \end{bmatrix} \right\| \quad \text{for all} \quad \begin{bmatrix} x \\ y \end{bmatrix} \in \mathcal{D}(\mathscr{A}), \tag{3.23}$$

because $\|Jw\| = \|w\|$ for all $w \in \mathcal{H}_1 \times \mathcal{H}_2$. In particular, $\lambda \notin \sigma_{\mathrm{app}}(\mathscr{A})$, i.e. $U \cap \sigma_{\mathrm{app}}(\mathscr{A}) = \varnothing$. The adjoint of $\mathscr{A}$ is the block operator matrix

$$\mathscr{A}^* = \begin{bmatrix} A^* & B \\ B^* & D^* \end{bmatrix},$$

which also satisfies the assumptions of this lemma. Indeed, (3.20) obviously also holds for D^*. Moreover, the uniform accretivity (3.19) of A together with $0 \in \varrho(A)$ imply that $A - \gamma_A$ is *m-accretive*, see [32, §V.3.10]. This in turn yields that $A^* - \gamma_A$ is m-accretive too and hence

$$\Re\langle A^*x, x\rangle_{\mathcal{H}_1} \geq \gamma_A \|x\|^2_{\mathcal{H}_1} \quad \text{for all} \quad x \in \mathcal{D}(A^*).$$

It follows that (3.23) also holds for $\mathscr{A}^*$. In particular, $\ker \mathscr{A}^* = \{0\}$ or, equivalently, $\mathcal{R}(\mathscr{A}) \subset \mathcal{H}_1 \times \mathcal{H}_2$ is dense. On the other hand, (3.23) implies that $\ker \mathscr{A} = \{0\}$ and that $\mathcal{R}(\mathscr{A})$ is closed. Consequently, $\mathcal{R}(\mathscr{A}) = \mathcal{H}_1 \times \mathcal{H}_2$ and therefore $0 \in \varrho(\mathscr{A})$. Using $\partial\sigma(\mathscr{A}) \subset \sigma_{\mathrm{app}}(\mathscr{A})$ and the connectedness of the set U, we obtain $U \subset \varrho(\mathscr{A})$. Now, (3.23) implies $\|(\mathscr{A} - \lambda)^{-1}\| \leq 1/c_\lambda$ for all $\lambda \in U$. ❑

We consider approximations $\mathscr{A}_n$ of $\mathscr{A}$ of the form

$$\mathscr{A}_n = \begin{bmatrix} A_n & B_n \\ C_n & D_n \end{bmatrix},$$

where

a) $(P_{1,n}, A_n)_{n\in\mathbb{N}}$ is a family which approximates A strongly in the sense of Section 3.2;

b) all projections $P_{1,n}$ are orthogonal and all A_n are uniformly accretive with the same constant γ_A as in (3.19);

c) $\mathcal{U}_{2,n} \subset \mathcal{H}_2$, $n \in \mathbb{N}$, are finite-dimensional subspaces of $\mathcal{H}_2$ and $P_{2,n} \colon \mathcal{H}_2 \to \mathcal{U}_{2,n}$ are orthogonal projections onto $\mathcal{U}_{2,n}$;

d) $B_n = P_{1,n}B|_{\mathcal{U}_{2,n}}$, $C_n = P_{2,n}C|_{\mathcal{U}_{1,n}}$ and $D_n = P_{2,n}D|_{\mathcal{U}_{2,n}}$, where $\mathcal{U}_{1,n} = \mathcal{R}(P_{1,n})$.

Lemma 3.5.2. *The following assertions hold:*

a) *Either* $\{\lambda \in \mathbb{C} \mid -\gamma_{D_\eta} < \Re\lambda < \gamma_{A_\eta}\} \subset \varrho(\mathscr{A}_n)$ *and*

$$\|(\mathscr{A}_n - \lambda)^{-1}\| \leq \frac{1}{\min\{\gamma_{A_\eta} - \Re\lambda, \gamma_{D_\eta} + \Re\lambda\}} \quad \textit{for} \quad -\gamma_{D_\eta} < \Re\lambda < \gamma_{A_\eta},$$

if (3.21) *holds, or* $\{\lambda \in \mathbb{C} \mid -\gamma_D < \Re\lambda < \gamma_A\} \subset \varrho(\mathscr{A}_n)$ *and*

$$\|(\mathscr{A}_n - \lambda)^{-1}\| \leq \frac{1}{\min\{\gamma_A - \Re\lambda, \gamma_D + \Re\lambda\}} \quad \textit{for} \quad -\gamma_D < \Re\lambda < \gamma_A,$$

if (3.22) *holds.*

b) $(\mathcal{P}_n, \mathscr{A}_n)_{n\in\mathbb{N}}$ *approximates* $\mathscr{A}$ *strongly where* $\mathcal{P}_n = \operatorname{diag}(P_{1,n}, P_{2,n})$.

Proof. a) From

$$\langle D_n y, y\rangle_{\mathcal{H}_2} = \langle P_{2,n} D y, y\rangle_{\mathcal{H}_2} = \langle D y, y\rangle_{\mathcal{H}_2} \quad \text{for all} \quad y \in \mathcal{U}_{2,n}$$

it follows that $-D_n$ is uniformly accretive with constant γ_D from (3.20). Consequently, Lemma 3.5.1 can be applied to $\mathscr{A}_n$.

b) In view of a) and Lemma 3.2.1 it suffices to show that for all $\begin{bmatrix} x \\ y \end{bmatrix} \in \mathcal{D}(A)\times\mathcal{H}_2$ there exist $\begin{bmatrix} x_n \\ y_n \end{bmatrix} \in \mathcal{U}_{1,n} \times \mathcal{U}_{2,n}$ such that

$$\lim_{n\to\infty} \begin{bmatrix} x_n \\ y_n \end{bmatrix} = \begin{bmatrix} x \\ y \end{bmatrix} \quad \text{and} \quad \lim_{n\to\infty} \mathscr{A}_n \begin{bmatrix} x_n \\ y_n \end{bmatrix} = \mathscr{A} \begin{bmatrix} x \\ y \end{bmatrix}. \tag{3.24}$$

Let therefore $\begin{bmatrix} x \\ y \end{bmatrix} \in \mathcal{D}(A) \times \mathcal{H}_2$. From Lemma 3.2.1 we get $x_n \in \mathcal{U}_{1,n}$ with $x_n \to x$ and $A_n x_n \to Ax$ as $n \to \infty$. Set $y_n = P_{2,n} y$. Then $y_n \to y$ and

$$\begin{aligned} \|D_n y_n - Dy\|_{\mathcal{H}_2} &\le \|P_{2,n}(Dy_n - Dy)\|_{\mathcal{H}_2} + \|P_{2,n} Dy - Dy\|_{\mathcal{H}_2} \\ &\le \|Dy_n - Dy\|_{\mathcal{H}_2} + \|P_{2,n} Dy - Dy\|_{\mathcal{H}_2} \to 0, \\ &\quad n \to \infty, \end{aligned}$$

i.e. $D_n y_n \to Dy$. The proof of $B_n y_n \to By$ and $C_n x_n \to Cx$ is the same. Hence, we have shown (3.24). ❑

Theorem 3.5.3. *Let either* $s_1, \dots, s_m \in \{\lambda \in \mathbb{C} \mid -\gamma_{D_\eta} < \Re\lambda < \gamma_{A_\eta}\}$, $0 < \varepsilon < \min_{j=1,\dots,m}\big(\min\{\gamma_{A_\eta} - \Re s_j, \gamma_{D_\eta} + \Re s_j\}\big)$ *and*

$$\delta_j > \frac{\varepsilon}{\min\{\gamma_{A_\eta} - \Re s_j, \gamma_{D_\eta} + \Re s_j\}^2 - \varepsilon \min\{\gamma_{A_\eta} - \Re s_j, \gamma_{D_\eta} + \Re s_j\}}$$

for $j = 1, \dots, m$, *if* (3.21) *holds, or* $s_1, \dots, s_m \in \{\lambda \in \mathbb{C} \mid -\gamma_D < \Re\lambda < \gamma_A\}$, $0 < \varepsilon < \min_{j=1,\dots,m}(\min\{\gamma_A - \Re s_j, \gamma_D + \Re s_j\})$ *and*

$$\delta_j > \frac{\varepsilon}{\min\{\gamma_A - \Re s_j, \gamma_D + \Re s_j\}^2 - \varepsilon \min\{\gamma_A - \Re s_j, \gamma_D + \Re s_j\}}$$

for $j = 1, \dots, m$, *if* (3.22) *holds.*

Then there exists an $n_0 \in \mathbb{N}$ *such that*

$$\sigma_\varepsilon(\mathscr{A}) \subset \bigcap_{j=1}^{m} \Big[\big(\mathrm{B}_{\delta_j}(W((\mathscr{A}_n - s_j)^{-1}))\big)^{-1} + s_j\Big] \quad \textit{for all} \quad n \ge n_0.$$

Proof. Lemma 3.5.1 yields

$$\|(\mathscr{A} - s_j)^{-1}\| < \frac{1}{\varepsilon}$$

and Lemma 3.5.2 implies

$$\|(\mathscr{A}_n - s_j)^{-1}\| < \frac{1}{\varepsilon}.$$

for $j = 1, \dots, m$. Hence, the assertion follows from Theorem 3.2.6. ❑

Remark 3.5.4. Suppose that A is the operator associated with a bounded coercive sesqui-linear form a on a densely and continuously embedded Hilbert space $\mathcal{V}_1 \subset \mathcal{H}_1$ and that $\mathcal{U}_{1,n}$, $\mathcal{W}$, $P_{1,n} \in \mathcal{L}(\mathcal{H}_1)$ and $A_n \in \mathcal{L}(\mathcal{U}_{1,n})$ are chosen as in Section 3.4. Then $(P_{1,n}, A_n)$ approximates A uniformly, and hence also strongly, see Remark 3.3.1. Moreover, the coercivity of a implies that A and all A_n are uniformly accretive with constant $\gamma_A = \gamma$ from (3.16). Hence, all assumptions of this section are fulfilled in this case.

Furthermore, this setting would also be covered by the assumptions of the next section.

Example 3.5.5. As an example for a block operator matrix that fits into the framework of this section we take a look at the Hain-Lüst operator. See [45] and [46] for results on the approximation of the quadratic numerical range of such a block operator. The Hain-Lüst operator is defined by

$$\mathscr{A} = \begin{bmatrix} A & B \\ B^* & D \end{bmatrix} = \begin{bmatrix} -\frac{1}{100}\frac{\mathrm{d}^2}{\mathrm{d}\xi^2} + 2 & I \\ I & 2\mathrm{e}^{2\pi\mathrm{i}\cdot} - 3 \end{bmatrix}$$

on the Hilbert space $\mathcal{H} := L^2(0,1) \times L^2(0,1)$ with

$$\mathcal{D}(A) = \{u \in H^2(0,1) \,|\, u(0) = u(1) = 0\},$$
$$\mathcal{D}(B) = \mathcal{D}(D) = L^2(0,1)$$

and $\mathcal{D}(\mathscr{A}) = \mathcal{D}(A) \times \mathcal{D}(D)$. Here, all the assumptions stated in the beginning of this section are fulfilled because we have $0 \in \varrho(A) \cap \varrho(D)$ and A and $-D$ are uniformly accretive with $\gamma_A = 2 + \frac{1}{400}$ and $\gamma_D = 1$ because of the Poincaré-Friedrichs inequality, see [56, p. 231], and the fact that $\Re\big(2\mathrm{e}^{2\pi\mathrm{i}\xi} - 3\big) \le -1$ for all $\xi \in [0,1]$. Thus, we have the spectrum free strip $\big\{\lambda \in \mathbb{C} \,\big|\, -1 < \Re\lambda < 2 + \frac{1}{400}\big\} \subset \varrho(\mathscr{A})$ and a corresponding bound of the norm of the resolvent by Lemma 3.5.1. Confer Example 3.7.2, where we construct finite element discretization matrices for this operator and plot a superset of the pseudospectrum that gives us the ability to distinguish disconnected components of the spectrum, in particular by displaying a portion of the spectrum free strip.

3.6 Discretization of Structured Block Operator Matrices II

In this section we will consider a different class of structured block operator matrices that – when compared to the setting in Section 3.5 – is more general in the sense that it also allows for unbounded off-diagonal entries and more restrictive in the sense that it requires the operator in the top left to be associated to a sesqui-linear form.

Let therefore $\mathcal{H}_1, \mathcal{H}_2$ be Hilbert spaces with inner products $\langle \cdot, \cdot \rangle_{\mathcal{H}_1}$ and $\langle \cdot, \cdot \rangle_{\mathcal{H}_2}$ respectively such that $\mathcal{H} = \mathcal{H}_1 \times \mathcal{H}_2$ and consider a block operator matrix $\mathscr{A} \colon \mathcal{D}(\mathscr{A}) \subset \mathcal{H} \to \mathcal{H}$ of the form

$$\mathscr{A} = \begin{bmatrix} A & B \\ C & D \end{bmatrix}$$

with $\mathcal{D}(\mathscr{A}) = \mathcal{D}(A) \times \mathcal{D}(B)$ such that the following assumptions are fulfilled:

a) There is a densely and continuously embedded Hilbert space $\mathcal{V}_1 \subset \mathcal{H}_1$, i.e. there exists a $c > 0$ such that

$$\|x\|_{\mathcal{H}_1} \leq c\|x\|_{\mathcal{V}_1} \quad \text{for all} \quad x \in \mathcal{V}_1,$$

and $a \colon \mathcal{V}_1 \times \mathcal{V}_1 \to \mathbb{C}$ is a bounded and coercive sesqui-linear form, i.e. there exist constants $M, \gamma_A > 0$ such that

$$|a(x,u)| \leq M\|x\|_{\mathcal{V}_1}\|u\|_{\mathcal{V}_1} \qquad \text{and} \qquad \Re a(x,x) \geq \gamma_A \|x\|^2_{\mathcal{V}_1}$$

for all $x, u \in \mathcal{V}_1$ such that the operator $A \colon \mathcal{D}(A) \subset \mathcal{H}_1 \to \mathcal{H}_1$ is associated with a. This means, that $x \in \mathcal{D}(A)$ if and only if $x \in \mathcal{V}_1$ and there exists an $f \in \mathcal{H}_1$ with $a(x,u) = \langle f, u \rangle_{\mathcal{H}_1}$ for every $u \in \mathcal{V}_1$. In this case we have $Ax = f$. Then A is a densely defined closed operator with $0 \in \varrho(A)$;

b) $B \colon \mathcal{D}(B) \subset \mathcal{H}_2 \to \mathcal{H}_1$ is a linear and densely defined operator such that $\mathcal{V}_1 \subset \mathcal{D}(B^*)$ and there exists a constant $k_{B^*} > 0$ such that $\|B^* u\|_{\mathcal{H}_2} \leq k_{B^*}\|u\|_{\mathcal{V}_1}$ for all $u \in \mathcal{V}_1$;

c) $C \colon \mathcal{D}(C) \subset \mathcal{H}_1 \to \mathcal{H}_2$ is a linear and densely defined operator such that $\mathcal{V}_1 \subset \mathcal{D}(C)$ and there exists a constant $k_C > 0$ such that $\|Cx\|_{\mathcal{H}_2} \leq k_C\|x\|_{\mathcal{V}_1}$ for all $x \in \mathcal{V}_1$;

d) $D \in \mathcal{L}(\mathcal{H}_2)$ and $-D$ is uniformly accretive, i.e. there exists a constant $\gamma_D > 0$ such that $\Re\langle Dy, y \rangle_{\mathcal{H}_2} \leq -\gamma_D \|y\|^2_{\mathcal{H}_2}$ for every $y \in \mathcal{H}_2$;

e) Either

$$\frac{1}{4}(k_{B^*} + k_C)^2 < \gamma_A \gamma_D \tag{3.25}$$

or

$$C = B^*. \tag{3.26}$$

Note, that (3.25) implies the existence of an $\eta > 0$ such that

$$\begin{aligned} \gamma_{A_\eta} &:= \gamma_A - \frac{1}{2\eta}(k_{B^*} + k_C) > 0, \\ \gamma_{D_\eta} &:= \gamma_D - \frac{\eta}{2}(k_{B^*} + k_C) > 0. \end{aligned}$$

The given assumptions do not guarantee that $\mathscr{A}$ is a closed operator. In order to obtain an inclusion like

$$\sigma_\varepsilon(\overline{\mathscr{A}}) \subset \bigcap_{s\in S} \Big[\big(B_{\delta_s}(W((\mathscr{A}_n - s)^{-1}))\big)^{-1} + s \Big]$$

for a suitable set S and sufficiently large $n \in \mathbb{N}$ we therefore have to make sure that $\mathscr{A}$ is closable and we have to find suitable approximation matrices $\mathscr{A}_n$. We will accomplish that by introducing the following form:

Definition 3.6.1. Let $\mathcal{V} := \mathcal{V}_1 \times \mathcal{H}_2$ and $\widetilde{a}\colon \mathcal{V} \times \mathcal{V} \to \mathbb{C}$,

$$\widetilde{a}\left(\begin{bmatrix} x \\ y \end{bmatrix}, \begin{bmatrix} u \\ v \end{bmatrix} \right) := a(x,u) + \langle y, B^* u\rangle_{\mathcal{H}_2} - \langle Cx, v\rangle_{\mathcal{H}_2} - \langle Dy, v\rangle_{\mathcal{H}_2}.$$

Lemma 3.6.2. *$\widetilde{a}$ is a bounded and coercive sesqui-linear form.*

Proof. We have

$$\begin{aligned}\left|\widetilde{a}\left(\begin{bmatrix} x \\ y \end{bmatrix}, \begin{bmatrix} u \\ v \end{bmatrix} \right)\right| &\leq M\|x\|_{\mathcal{V}_1}\|u\|_{\mathcal{V}_1} + \|y\|_{\mathcal{H}_2}\|B^*u\|_{\mathcal{H}_2} \\ &\quad + \|Cx\|_{\mathcal{H}_2}\|v\|_{\mathcal{H}_2} + \|Dy\|_{\mathcal{H}_2}\|v\|_{\mathcal{H}_2} \\ &\leq M\|x\|_{\mathcal{V}_1}\|u\|_{\mathcal{V}_1} + k_{B^*}\|y\|_{\mathcal{H}_2}\|u\|_{\mathcal{V}_1} \\ &\quad + k_C\|x\|_{\mathcal{V}_1}\|v\|_{\mathcal{H}_2} + \|D\|\|y\|_{\mathcal{H}_2}\|v\|_{\mathcal{H}_2} \\ &\leq \frac{\widetilde{M}}{2}\left(\|x\|_{\mathcal{V}_1} + \|y\|_{\mathcal{H}_2}\right)\left(\|u\|_{\mathcal{V}_1} + \|v\|_{\mathcal{H}_2}\right) \\ &\leq \widetilde{M}\left\|\begin{bmatrix} x \\ y \end{bmatrix}\right\|_{\mathcal{V}}\left\|\begin{bmatrix} u \\ v \end{bmatrix}\right\|_{\mathcal{V}}\end{aligned}$$

with $\widetilde{M} = 2\max\big\{M, k_{B^*}, k_C, \|D\|\big\} > 0$.
For the coercivity, we consider two cases. If (3.25) holds, we obtain

$$\begin{aligned}\Re\widetilde{a}\left(\begin{bmatrix} x \\ y \end{bmatrix}, \begin{bmatrix} x \\ y \end{bmatrix} \right) &\geq \gamma_A\|x\|_{\mathcal{V}_1}^2 + \Re(\langle y, B^*x\rangle_{\mathcal{H}_2} - \langle Cx, y\rangle_{\mathcal{H}_2}) + \gamma_D\|y\|_{\mathcal{H}_2}^2 \\ &\geq \gamma_A\|x\|_{\mathcal{V}_1}^2 - \|y\|_{\mathcal{H}_2}\|B^*x\|_{\mathcal{H}_2} - \|Cx\|_{\mathcal{H}_2}\|y\|_{\mathcal{H}_2} + \gamma_D\|y\|_{\mathcal{H}_2}^2 \\ &\geq \gamma_A\|x\|_{\mathcal{V}_1}^2 - (k_{B^*} + k_C)\|y\|_{\mathcal{H}_2}\|x\|_{\mathcal{V}_1} + \gamma_D\|y\|_{\mathcal{H}_2}^2 \\ &\geq \gamma_A\|x\|_{\mathcal{V}_1}^2 - (k_{B^*} + k_C)\left(\frac{\eta}{2}\|y\|_{\mathcal{H}_2}^2 + \frac{1}{2\eta}\|x\|_{\mathcal{V}_1}^2\right) + \gamma_D\|y\|_{\mathcal{H}_2}^2 \\ &= \left(\gamma_A - \frac{1}{2\eta}(k_{B^*} + k_C)\right)\|x\|_{\mathcal{V}_1}^2 \\ &\quad + \left(\gamma_D - \frac{\eta}{2}(k_{B^*} + k_C)\right)\|y\|_{\mathcal{H}_2}^2 \qquad (3.27) \\ &\geq \widetilde{\gamma}\left\|\begin{bmatrix} x \\ y \end{bmatrix}\right\|_{\mathcal{V}}^2\end{aligned}$$

with $\widetilde{\gamma} = \min\left\{\gamma_{A_\eta}, \gamma_{D_\eta}\right\} > 0$. In the other case, if (3.26) holds, we have

$$\begin{aligned}\Re\widetilde{a}\left(\begin{bmatrix}x\\y\end{bmatrix},\begin{bmatrix}x\\y\end{bmatrix}\right) &\geq \gamma_A\|x\|^2_{\mathcal{V}_1} + \Re(\langle y, B^*x\rangle_{\mathcal{H}_2} - \langle Cx, y\rangle_{\mathcal{H}_2}) + \gamma_D\|y\|^2_{\mathcal{H}_2} \\ &= \gamma_A\|x\|^2_{\mathcal{V}_1} + \gamma_D\|y\|^2_{\mathcal{H}_2} \\ &\geq \widetilde{\gamma}\left\|\begin{bmatrix}x\\y\end{bmatrix}\right\|^2_{\mathcal{V}}\end{aligned} \tag{3.28}$$

with $\widetilde{\gamma} = \min\{\gamma_A, \gamma_D\} > 0$. ❑

Denote by $J\widetilde{\mathscr{A}}$ the operator associated to $\widetilde{a}$, where

$$J = \begin{bmatrix} I & 0 \\ 0 & -I\end{bmatrix}.$$

Then $J\widetilde{\mathscr{A}}$ is closed with $0 \in \varrho(J\widetilde{\mathscr{A}})$. For $\begin{bmatrix}x\\y\end{bmatrix} \in \mathcal{D}(\mathscr{A}) = \mathcal{D}(A)\times\mathcal{D}(B) \subset \mathcal{V}$ and $\begin{bmatrix}u\\v\end{bmatrix} \in \mathcal{V}$ we have

$$\begin{aligned}\widetilde{a}\left(\begin{bmatrix}x\\y\end{bmatrix},\begin{bmatrix}u\\v\end{bmatrix}\right) &= a(x,u) + \langle y, B^*u\rangle_{\mathcal{H}_2} - \langle Cx, v\rangle_{\mathcal{H}_2} - \langle Dy, v\rangle_{\mathcal{H}_2} \\ &= \langle Ax + By, u\rangle_{\mathcal{H}_1} + \langle -Cx - Dy, v\rangle_{\mathcal{H}_2} \\ &= \left\langle J\mathscr{A}\begin{bmatrix}x\\y\end{bmatrix},\begin{bmatrix}u\\v\end{bmatrix}\right\rangle.\end{aligned}$$

Hence, $J\mathscr{A} \subset J\widetilde{\mathscr{A}}$ holds and thus $\mathscr{A}$ is closable with $\overline{\mathscr{A}} \subset \widetilde{\mathscr{A}}$.

We also obtain the existence of a spectrum free strip similar to Lemma 3.5.1:

Lemma 3.6.3. a) *If* (3.25) *holds, we have*

$$\left\{\lambda\in\mathbb{C}\,\middle|\,-\gamma_{D_\eta} < \Re\lambda < c^{-2}\gamma_{A_\eta}\right\} \subset \varrho(\widetilde{\mathscr{A}})$$

and

$$\|(\widetilde{\mathscr{A}}-\lambda)^{-1}\| \leq \frac{1}{\min\left\{c^{-2}\gamma_{A_\eta} - \Re\lambda, \gamma_{D_\eta} + \Re\lambda\right\}}$$

if $-\gamma_{D_\eta} < \Re\lambda < c^{-2}\gamma_{A_\eta}$.

b) *If* (3.26) *holds, we have* $\left\{\lambda\in\mathbb{C}\,\middle|\, -\gamma_D < \Re\lambda < c^{-2}\gamma_A\right\} \subset \varrho(\widetilde{\mathscr{A}})$ *and*

$$\|(\widetilde{\mathscr{A}}-\lambda)^{-1}\| \leq \frac{1}{\min\{c^{-2}\gamma_A - \Re\lambda, \gamma_D + \Re\lambda\}}$$

if $-\gamma_D < \Re\lambda < c^{-2}\gamma_A$.

Proof. Let $\lambda \in U := \{\lambda \in \mathbb{C} \mid -\gamma_{D_\eta} < \Re\lambda < c^{-2}\gamma_{A_\eta}\}$ and $\begin{bmatrix} x \\ y \end{bmatrix} \in \mathcal{D}(\widetilde{\mathscr{A}})$. Just as in (3.27) we obtain

$$\Re\left\langle J(\widetilde{\mathscr{A}} - \lambda)\begin{bmatrix} x \\ y \end{bmatrix}, \begin{bmatrix} x \\ y \end{bmatrix}\right\rangle \geq c_\lambda \left\|\begin{bmatrix} x \\ y \end{bmatrix}\right\|^2,$$

where $c_\lambda = \min\{c^{-2}\gamma_{A_\eta} - \Re\lambda, \gamma_{D_\eta} + \Re\lambda\}$ and with the same arguments that led to (3.23) we therefore have

$$\left\|(\widetilde{\mathscr{A}} - \lambda)\begin{bmatrix} x \\ y \end{bmatrix}\right\| \geq c_\lambda \left\|\begin{bmatrix} x \\ y \end{bmatrix}\right\| \quad \text{for all} \quad \begin{bmatrix} x \\ y \end{bmatrix} \in \mathcal{D}(\widetilde{\mathscr{A}}). \tag{3.29}$$

In particular, $\lambda \notin \sigma_{\mathrm{app}}(\widetilde{\mathscr{A}})$, i.e. $U \cap \sigma_{\mathrm{app}}(\widetilde{\mathscr{A}}) = \varnothing$. Because of $0 \in \varrho(\widetilde{\mathscr{A}})$ we can use $\partial\sigma(\widetilde{\mathscr{A}}) \subset \sigma_{\mathrm{app}}(\widetilde{\mathscr{A}})$ and the connectedness of the set U to obtain $U \subset \varrho(\widetilde{\mathscr{A}})$. Now, (3.29) implies $\|(\widetilde{\mathscr{A}} - \lambda)^{-1}\| \leq 1/c_\lambda$ for all $\lambda \in U$ which proves a). For b), the same line of reasoning can be applied with (3.28) instead of (3.27). ❑

Theorem 3.6.4. *Let $\mathcal{U}_{1,n} \subset \mathcal{V}_1$, $n \in \mathbb{N}$, be a finite dimensional subspace for which there exists a mapping $Q_n\colon \mathcal{V}_1 \to \mathcal{U}_{1,n}$ with*

$$\lim_{n\to\infty} \|x - Q_n x\|_{\mathcal{V}_1} = 0 \quad \textit{for all} \quad x \in \mathcal{V}_1$$

and let $\mathcal{U}_{2,n} \subset \mathcal{H}_2$, $n \in \mathbb{N}$, be a finite dimensional subspace for which there exists a projection $P_{2,n} \in \mathcal{L}(\mathcal{H}_2)$ with $\mathcal{R}(P_{2,n}) = \mathcal{U}_{2,n}$ and

$$\lim_{n\to\infty} \|y - P_{2,n} y\|_{\mathcal{H}_2} = 0 \quad \textit{for all} \quad y \in \mathcal{H}_2.$$

Let further $P_{1,n} \in \mathcal{L}(\mathcal{H}_1)$ be a projection with $\mathcal{R}(P_{1,n}) = \mathcal{U}_{1,n}$ and

$$\lim_{n\to\infty} \|x - P_{1,n} x\|_{\mathcal{H}_1} = 0 \quad \textit{for all} \quad x \in \mathcal{H}_1$$

and denote by $J\widetilde{\mathscr{A}_n}$ the operator associated to the restricted form

$$\widetilde{a}|_{\mathcal{U}_{1,n}\times\mathcal{U}_{2,n}}\colon (\mathcal{U}_{1,n} \times \mathcal{U}_{2,n}) \times (\mathcal{U}_{1,n} \times \mathcal{U}_{2,n}) \to \mathbb{C}.$$

Then $\left(\begin{bmatrix} P_{1,n} & 0 \\ 0 & P_{2,n} \end{bmatrix}, \widetilde{\mathscr{A}_n}\right)_{n\in\mathbb{N}}$ is a strong approximation of $\widetilde{\mathscr{A}}$ in the sense of Section 3.2.

Proof. Let us define

$$R_n = \begin{bmatrix} Q_n & 0 \\ 0 & P_{2,n} \end{bmatrix}.$$

For $\begin{bmatrix} f \\ g \end{bmatrix} \in \mathcal{H}$ we set

$$\begin{bmatrix} x \\ y \end{bmatrix} = \widetilde{\mathscr{A}}^{-1}\begin{bmatrix} f \\ g \end{bmatrix} \quad \text{and} \quad \begin{bmatrix} x_n \\ y_n \end{bmatrix} = \widetilde{\mathscr{A}_n}^{-1}\begin{bmatrix} P_{1,n} & 0 \\ 0 & P_{2,n} \end{bmatrix}\begin{bmatrix} f \\ g \end{bmatrix}.$$

Then we have

$$\widetilde{a}\left(\begin{bmatrix} x \\ y \end{bmatrix}, \begin{bmatrix} u \\ v \end{bmatrix} \right) = \left\langle J\widetilde{\mathscr{A}} \begin{bmatrix} x \\ y \end{bmatrix}, \begin{bmatrix} u \\ v \end{bmatrix} \right\rangle = \left\langle \begin{bmatrix} f \\ -g \end{bmatrix}, \begin{bmatrix} u \\ v \end{bmatrix} \right\rangle$$

for all $\begin{bmatrix} u \\ v \end{bmatrix} \in \mathcal{V}$ and

$$\begin{aligned} \widetilde{a}\left(\begin{bmatrix} x_n \\ y_n \end{bmatrix}, \begin{bmatrix} u_n \\ v_n \end{bmatrix} \right) &= \left\langle J\widetilde{\mathscr{A}_n} \begin{bmatrix} x_n \\ y_n \end{bmatrix}, \begin{bmatrix} u_n \\ v_n \end{bmatrix} \right\rangle = \left\langle J \begin{bmatrix} P_{1,n} & 0 \\ 0 & P_{2,n} \end{bmatrix} \begin{bmatrix} f \\ g \end{bmatrix}, \begin{bmatrix} u_n \\ v_n \end{bmatrix} \right\rangle \\ &= \left\langle \begin{bmatrix} f \\ -g \end{bmatrix}, \begin{bmatrix} u_n \\ v_n \end{bmatrix} \right\rangle \end{aligned}$$

for all $\begin{bmatrix} u_n \\ v_n \end{bmatrix} \in \mathcal{U}_{1,n} \times \mathcal{U}_{2,n}$. With this at hand we can now make use of the lemma of Cea [48, Theorem VII.5.A] in order to estimate

$$\begin{aligned} \left\| \widetilde{\mathscr{A}}^{-1} \begin{bmatrix} f \\ g \end{bmatrix} - \widetilde{\mathscr{A}_n}^{-1} \begin{bmatrix} P_{1,n} & 0 \\ 0 & P_{2,n} \end{bmatrix} \begin{bmatrix} f \\ g \end{bmatrix} \right\|_{\mathcal{H}} &= \left\| \begin{bmatrix} x \\ y \end{bmatrix} - \begin{bmatrix} x_n \\ y_n \end{bmatrix} \right\|_{\mathcal{H}} \\ &\leq \max\{c, 1\} \left\| \begin{bmatrix} x \\ y \end{bmatrix} - \begin{bmatrix} x_n \\ y_n \end{bmatrix} \right\|_{\mathcal{V}} \\ &\leq \max\{c, 1\} \frac{\widetilde{M}}{\widetilde{\gamma}} \inf_{\left[\begin{smallmatrix} u_n \\ v_n \end{smallmatrix}\right] \in \mathcal{U}_{1,n} \times \mathcal{U}_{2,n}} \left\| \begin{bmatrix} x \\ y \end{bmatrix} - \begin{bmatrix} u_n \\ v_n \end{bmatrix} \right\|_{\mathcal{V}} \\ &\leq \max\{c, 1\} \frac{\widetilde{M}}{\widetilde{\gamma}} \left\| \begin{bmatrix} x \\ y \end{bmatrix} - R_n \begin{bmatrix} x \\ y \end{bmatrix} \right\|_{\mathcal{V}} \\ &= \max\{c, 1\} \frac{\widetilde{M}}{\widetilde{\gamma}} \left(\|x - Q_n x\|_{\mathcal{V}_1}^2 + \|y - P_{2,n} y\|_{\mathcal{H}_2}^2 \right)^{\frac{1}{2}} \to 0 \end{aligned}$$

for $n \to \infty$. ❑

Now Theorem 3.2.6 can be applied here in order to obtain the inclusion

$$\sigma_\varepsilon(\widetilde{\mathscr{A}}) \subset \bigcap_{s \in S} \left[\left(\mathrm{B}_{\delta_s}(W((\widetilde{\mathscr{A}_n} - s)^{-1})) \right)^{-1} + s \right]$$

for a suitable set S.

In the following we are interested in finding sufficient conditions under which the equality $\widetilde{\mathscr{A}} = \overline{\mathscr{A}}$ holds.

Remark 3.6.5. A^* is densely defined with $\mathcal{D}(A^*) \subset \mathcal{V}_1$, because A^* is associated to the adjoint form a^* of a defined by $a^*(u, x) = \overline{a(x, u)}$ for $x, u \in \mathcal{V}_1$, see [32, Theorem 6.2.5, p. 323]. This is true because if we denote the operator associated to a^* by A^{ad}, we have

$$\langle Ax, u \rangle_{\mathcal{H}_1} = a(x, u) = \overline{a^*(u, x)} = \overline{\langle A^{\mathrm{ad}} u, x \rangle_{\mathcal{H}_1}} = \langle x, A^{\mathrm{ad}} u \rangle_{\mathcal{H}_1}$$

for all $x \in \mathcal{D}(A)$ and $u \in \mathcal{D}(A^{\mathrm{ad}})$. This implies $A^{\mathrm{ad}} \subset A^*$. Since we have $0 \in \varrho(A)$, we also have $0 \in \varrho(A^*)$ and because of $0 \in \varrho(A^{\mathrm{ad}})$ we can therefore conclude $A^{\mathrm{ad}} = A^*$.

Lemma 3.6.6. *If $\mathcal{D}(C^*)$ is a dense subset of $\mathcal{H}_2$, we have*

$$\overline{\mathscr{A}} \begin{bmatrix} x \\ y \end{bmatrix} = \begin{bmatrix} A(x + \overline{A^{-1}B}y) \\ \overline{C}x + Dy \end{bmatrix}$$

for all $\begin{bmatrix} x \\ y \end{bmatrix} \in \mathcal{D}(\overline{\mathscr{A}})$ *where*

$$\mathcal{D}(\overline{\mathscr{A}}) = \left\{ \begin{bmatrix} x \\ y \end{bmatrix} \in \mathcal{H} \,\middle|\, x + \overline{A^{-1}B}y \in \mathcal{D}(A),\ y \in \mathcal{D}(\overline{CA^{-1}B}) \right\}.$$

Proof. As seen in Remark 3.6.5, we have $\mathcal{D}(A^*) \subset \mathcal{V}_1 \subset \mathcal{D}(B^*)$ and therefore the inclusion

$$A^{-1}B \subset (B^*A^{-*})^* \in \mathcal{L}(\mathcal{H}_2, \mathcal{H}_1)$$

holds, so $A^{-1}B$ is bounded on $\mathcal{D}(B)$. The closure of $A^{-1}B$ is then given by $\overline{A^{-1}B} = (B^*A^{-*})^*$. Furthermore, the denseness of $\mathcal{D}(C^*)$ is equivalent to C being closable, so we can conclude that $CA^{-1}B$ is also closable. The assertion then follows from Theorem 1.1.10 via

$$\begin{aligned} \overline{\mathscr{A}} \begin{bmatrix} x \\ y \end{bmatrix} &= \begin{bmatrix} I & 0 \\ CA^{-1} & I \end{bmatrix} \begin{bmatrix} A & 0 \\ 0 & D - \overline{CA^{-1}B} \end{bmatrix} \begin{bmatrix} I & \overline{A^{-1}B} \\ 0 & I \end{bmatrix} \begin{bmatrix} x \\ y \end{bmatrix} \\ &= \begin{bmatrix} A(x + \overline{A^{-1}B}y) \\ C(x + \overline{A^{-1}B}y) + Dy - \overline{CA^{-1}B}y \end{bmatrix} \end{aligned}$$

for all $\begin{bmatrix} x \\ y \end{bmatrix} \in \left\{ \begin{bmatrix} x \\ y \end{bmatrix} \in \mathcal{H} \,\middle|\, x + \overline{A^{-1}B}y \in \mathcal{D}(A),\ y \in \mathcal{D}(\overline{CA^{-1}B}) \right\}$ and

$$\overline{CA^{-1}B} \subset \overline{C}\,\overline{A^{-1}B}.$$ ❑

Lemma 3.6.7. *If $\mathcal{D}(C^*)$ is a dense subset of $\mathcal{H}_2$, we have*

$$\mathscr{A}^* \begin{bmatrix} u \\ v \end{bmatrix} = \begin{bmatrix} A^*(u + \overline{A^{-*}C^*}v) \\ B^*u + D^*v \end{bmatrix}$$

for all $\begin{bmatrix} u \\ v \end{bmatrix} \in \mathcal{D}(\mathscr{A}^*)$ *where*

$$\mathcal{D}(\mathscr{A}^*) = \left\{ \begin{bmatrix} u \\ v \end{bmatrix} \in \mathcal{H} \,\middle|\, u \in \mathcal{D}(B^*),\ u + \overline{A^{-*}C^*}v \in \mathcal{D}(A^*) \right\}.$$

Proof. We have $\mathcal{D}(A) \subset \mathcal{V}_1 \subset \mathcal{D}(C)$ and therefore the inclusion

$$A^{-*}C^* \subset (CA^{-1})^* \in \mathcal{L}(\mathcal{H}_2, \mathcal{H}_1)$$

holds, so $A^{-*}C^*$ is bounded on $\mathcal{D}(C^*)$. The closure of $A^{-*}C^*$ is then given by $\overline{A^{-*}C^*} = (CA^{-1})^*$.

Let $\begin{bmatrix} u \\ v \end{bmatrix} \in \mathcal{D}(\mathscr{A}^*)$ be such that

$$\left\langle \mathscr{A} \begin{bmatrix} x \\ y \end{bmatrix}, \begin{bmatrix} u \\ v \end{bmatrix} \right\rangle = \left\langle \begin{bmatrix} x \\ y \end{bmatrix}, \begin{bmatrix} w \\ z \end{bmatrix} \right\rangle$$

for all $\begin{bmatrix} x \\ y \end{bmatrix} \in \mathcal{D}(\mathscr{A})$. If we then choose x to be 0, we obtain

$$\langle By, u \rangle_{\mathcal{H}_1} = \langle y, z - D^*v \rangle_{\mathcal{H}_2}.$$

Hence, $u \in \mathcal{D}(B^*)$ and $z = B^*u + D^*v$. If on the other hand we choose y to be 0, we obtain

$$\langle Ax, u \rangle_{\mathcal{H}_1} + \langle CA^{-1}Ax, v \rangle_{\mathcal{H}_2} = \langle x, w \rangle_{\mathcal{H}_1}$$

or equivalently

$$\langle Ax, u + \overline{A^{-*}C^*}v \rangle_{\mathcal{H}_1} = \langle x, w \rangle_{\mathcal{H}_1}.$$

Hence, $u + \overline{A^{-*}C^*}v \in \mathcal{D}(A^*)$ and $w = A^*(u + \overline{A^{-*}C^*}v)$. Since for any

$$\begin{bmatrix} u \\ v \end{bmatrix} \in \left\{ \begin{bmatrix} u \\ v \end{bmatrix} \in \mathcal{H} \,\middle|\, u \in \mathcal{D}(B^*),\ u + \overline{A^{-*}C^*}v \in \mathcal{D}(A^*) \right\}$$

we do also obtain the equation

$$\left\langle \mathscr{A} \begin{bmatrix} x \\ y \end{bmatrix}, \begin{bmatrix} u \\ v \end{bmatrix} \right\rangle = \left\langle \begin{bmatrix} x \\ y \end{bmatrix}, \begin{bmatrix} A^*(u + \overline{A^{-*}C^*}v) \\ B^*u + D^*v \end{bmatrix} \right\rangle$$

for every $\begin{bmatrix} x \\ y \end{bmatrix} \in \mathcal{D}(\mathscr{A})$, the proof is complete. ❑

Definition 3.6.8. Let $n \in \mathbb{N}$. We define the *Frobenius inner product* $\langle \cdot, \cdot \rangle_{\mathrm{F}}$ on $\mathcal{H}_2^{n \times n}$ via

$$\langle K, L \rangle_{\mathrm{F}} = \sum_{i=1}^{n} \sum_{j=1}^{n} \langle K_{i,j}, L_{i,j} \rangle_{\mathcal{H}_2}$$

for all $K, L \in \mathcal{H}_2^{n \times n}$.

Theorem 3.6.9. *Suppose that one of the following conditions is fulfilled:*

a) $\mathcal{D}(C^*)$ *and* $\mathcal{R}(C|_{\mathcal{D}(A)})$ *are dense in* $\mathcal{H}_2$, $a(x,u) = \alpha\langle Cx, B^*u\rangle_{\mathcal{H}_2}$ *holds for an* $\alpha \in \mathbb{C}\setminus\{0\}$ *and all* $x, u \in \mathcal{V}_1$ *and* B *is closed;*

b) $\mathcal{R}(B^*|_{\mathcal{D}(A^*)})$ *is dense in* $\mathcal{H}_2$ *and* $A^* = \alpha C^* B^*$ *holds for an* $\alpha \in \mathbb{C}\setminus\{0\}$ *with* $\frac{1}{\alpha} \notin \sigma_{\mathrm{p}}(D^*)$*;*

c) $\mathcal{H}_1 = \mathcal{H}_2^n$, $\mathcal{V}_1 = \mathcal{V}_{1,1}\times\cdots\times\mathcal{V}_{1,n} \subset \mathcal{D}(C^*)^n \cap \mathcal{D}(B)^n$ *and* $\bigcap_{i=1}^n \mathcal{V}_{1,i}$ *is dense in* $\mathcal{H}_2$ *for some* $n \in \mathbb{N}$*. Furthermore,* $a(x,u) = \alpha\langle \boldsymbol{C}^*x, \boldsymbol{B}u\rangle_{\mathrm{F}}$ *holds for an* $\alpha \in \mathbb{C}\setminus\{0\}$ *and all* $x, u \in \mathcal{V}_1$ *where*

$$\boldsymbol{B}\colon \mathcal{V}_1 \subset \mathcal{H}_1 \to \mathcal{H}_2^{n\times n}, \qquad \boldsymbol{B}u := \begin{bmatrix} Bu_1 & \dots & Bu_n \end{bmatrix},$$

is closable and

$$\boldsymbol{C}\colon \mathcal{D}(\boldsymbol{C}) \subset \mathcal{H}_2^{n\times n} \to \mathcal{H}_1$$

such that $\mathcal{V}_1 \subset \mathcal{D}(\boldsymbol{C}^*)$ *and* $\mathcal{R}(\boldsymbol{C}^*|_{D(A)})$ *is dense in* $\mathcal{H}_2^{n\times n}$*. Moreover,* $B^*u = \beta\sum_{i=1}^n (Bu_i)_i$ *for all* $u \in \mathcal{V}_1$ *and some* $\beta \in \mathbb{C}$ *with* $|\beta| = 1$ *and* $Cx = \gamma\sum_{i=1}^n (\boldsymbol{C}^*x)_{i,i}$ *for all* $x \in \mathcal{V}_1$ *and some* $\gamma \in \mathbb{C}$*.*

Then we have

$$\widetilde{\mathscr{A}} = \overline{\mathscr{A}}.$$

Remark 3.6.10. Note that the assumptions on a in a) and c) imply $A = \alpha BC|_{\mathcal{D}(A)}$, $A^* = \overline{\alpha}C^*B^*|_{\mathcal{D}(A^*)}$ and $A = \alpha\boldsymbol{B}^*\boldsymbol{C}^*|_{\mathcal{D}(A)}$, $A^* = \overline{\alpha}\overline{\boldsymbol{C}}\boldsymbol{B}$ respectively.

Proof of Theorem 3.6.9. a) In this setting we can state the operator $\widetilde{\mathscr{A}}$ explicitly. Let $\begin{bmatrix} x \\ y \end{bmatrix} \in \mathcal{D}(\widetilde{\mathscr{A}})$ and $\begin{bmatrix} f \\ g \end{bmatrix} \in \mathcal{H}$ be such that

$$\widetilde{a}\left(\begin{bmatrix} x \\ y \end{bmatrix}, \begin{bmatrix} u \\ v \end{bmatrix}\right) = \left\langle \begin{bmatrix} f \\ g \end{bmatrix}, \begin{bmatrix} u \\ v \end{bmatrix}\right\rangle \quad \text{for all} \quad \begin{bmatrix} u \\ v \end{bmatrix} \in \mathcal{V}.$$

Then we have

$$\widetilde{a}\left(\begin{bmatrix} x \\ y \end{bmatrix}, \begin{bmatrix} 0 \\ v \end{bmatrix}\right) = \langle -Cx - Dy, v\rangle_{\mathcal{H}_2} = \langle g, v\rangle_{\mathcal{H}_2}$$

for all $v \in \mathcal{H}_2$ which implies $g = -Cx - Dy$ and we also have

$$\widetilde{a}\left(\begin{bmatrix} x \\ y \end{bmatrix}, \begin{bmatrix} u \\ 0 \end{bmatrix}\right) = \langle \alpha Cx + y, B^*u\rangle_{\mathcal{H}_2} = \langle f, u\rangle_{\mathcal{H}_1}$$

for all $u \in \mathcal{V}_1$ which implies $\alpha Cx + y \in \mathcal{D}(B)$ and $f = B(\alpha Cx + y)$. On the other hand, if we choose $\begin{bmatrix} x \\ y \end{bmatrix} \in \mathcal{V}$ such that $\alpha Cx + y \in \mathcal{D}(B)$, we obtain that

$$\begin{aligned} \widetilde{a}\left(\begin{bmatrix} x \\ y \end{bmatrix}, \begin{bmatrix} u \\ v \end{bmatrix}\right) &= a(x,u) + \langle y, B^*u\rangle_{\mathcal{H}_2} - \langle Cx, v\rangle_{\mathcal{H}_2} - \langle Dy, v\rangle_{\mathcal{H}_2} \\ &= \langle \alpha Cx + y, B^*u\rangle_{\mathcal{H}_2} - \langle Cx + Dy, v\rangle_{\mathcal{H}_2} \\ &= \langle B(\alpha Cx + y), u\rangle_{\mathcal{H}_1} - \langle Cx + Dy, v\rangle_{\mathcal{H}_2} \end{aligned}$$

holds for all $\begin{bmatrix} u \\ v \end{bmatrix} \in \mathcal{V}$. This shows

$$J\widetilde{\mathscr{A}}\begin{bmatrix} x \\ y \end{bmatrix} = \begin{bmatrix} B(\alpha Cx + y) \\ -Cx - Dy \end{bmatrix}$$

for all $\begin{bmatrix} x \\ y \end{bmatrix} \in \mathcal{D}(\widetilde{\mathscr{A}}) = \left\{ \begin{bmatrix} x \\ y \end{bmatrix} \in \mathcal{V} \,\middle|\, \alpha Cx + y \in \mathcal{D}(B) \right\}$.
Due to $D(C^*)$ being dense in $\mathcal{H}_2$, $\overline{\mathscr{A}}$ is given as in Lemma 3.6.6 and since $\widetilde{\mathscr{A}}$ is also a closed extension of $\mathscr{A}$ we have $\overline{\mathscr{A}} \subset \widetilde{\mathscr{A}}$. By the same line of reasoning as in the beginning of the proof of Lemma 3.6.6 we see that $\overline{A^{-1}B} = (B^*A^{-*})^* \in \mathcal{L}(\mathcal{H}_2, \mathcal{H}_1)$. In addition, we have $A = \alpha BC|_{\mathcal{D}(A)}$ and thus, using $D(A^*) \subset D(B^*)$ from Remark 3.6.5,

$$\frac{1}{\alpha} I|_{\mathcal{D}(A)} = A^{-1}BC|_{\mathcal{D}(A)} \subset (C^*B^*A^{-*})^* \in \mathcal{L}(\mathcal{H}_1)$$

holds which implies $\frac{1}{\alpha} I = \overline{A^{-1}BC} = (C^*B^*A^{-*})^*$ and because of $\overline{A^{-1}B}C \subset (C^*B^*A^{-*})^*$ we can therefore deduce $\overline{A^{-1}B}C = \frac{1}{\alpha} I|_{\mathcal{D}(C)}$. Combining this with $B \subset A\overline{A^{-1}B}$ we see that for $x \in \mathcal{V}_1$ and $y \in \mathcal{H}_2$, $\alpha Cx + y \in \mathcal{D}(B)$ implies $x + \overline{A^{-1}B}y \in \mathcal{D}(A)$. This is because from $\alpha Cx + y \in \mathcal{D}(A\overline{A^{-1}B})$ we get $\overline{A^{-1}B}(\alpha Cx + y) \in \mathcal{D}(A)$ and the fact that $\overline{A^{-1}B}$ is defined on $\mathcal{H}_2$ yields $x + \overline{A^{-1}B}y \in \mathcal{D}(A)$. The same arguments explain

$$B(\alpha Cx + y) = A(x + \overline{A^{-1}B}y) \quad \text{for all} \quad \begin{bmatrix} x \\ y \end{bmatrix} \in \mathcal{D}(\widetilde{\mathscr{A}}).$$

It remains to show that $y \in \mathcal{D}(\overline{CA^{-1}B})$. Since A is invertible and $A = \alpha BC|_{\mathcal{D}(A)}$, we have that $C|_{\mathcal{D}(A)}$ is injective and therefore

$$CA^{-1}B|_{\mathcal{R}(C|_{\mathcal{D}(A)})} = CA^{-1}BCC|_{\mathcal{D}(A)}^{-1}|_{\mathcal{R}(C|_{\mathcal{D}(A)})} = \frac{1}{\alpha} I|_{\mathcal{R}(C|_{\mathcal{D}(A)})}.$$

Now the assumption that $\mathcal{R}(C|_{\mathcal{D}(A)})$ is dense in $\mathcal{H}_2$ implies that $\overline{CA^{-1}B} = \frac{1}{\alpha} I$, i.e. $\mathcal{D}(\overline{CA^{-1}B}) = \mathcal{H}_2$. Thus, the inclusion $\widetilde{\mathscr{A}} \subset \overline{\mathscr{A}}$ also holds true.

b) Because of $A^* = \alpha C^*B^*|_{\mathcal{D}(A^*)}$ and $\mathcal{D}(A^*) \subset \mathcal{D}(C^*B^*)$, the inclusion

$$\mathcal{R}(B^*|_{\mathcal{D}(A^*)}) \subset \mathcal{D}(C^*)$$

holds. Therefore, $\mathcal{D}(C^*)$ is dense in $\mathcal{H}_2$ as well, so $\overline{\mathscr{A}}$ and $\mathscr{A}^*$ are given as in Lemma 3.6.6 and Lemma 3.6.7 respectively. We will now show that $0 \in \varrho(\overline{\mathscr{A}})$. To this end, first note that since $\widetilde{\mathscr{A}}$ is a closed extension of $\mathscr{A}$, we have $\overline{\mathscr{A}} \subset \widetilde{\mathscr{A}}$. Hence, we can use that the inequality

$$\Re\left\langle \overline{\mathscr{A}} \begin{bmatrix} x \\ y \end{bmatrix}, \begin{bmatrix} x \\ y \end{bmatrix} \right\rangle \geq \widetilde{\gamma} \left\| \begin{bmatrix} x \\ y \end{bmatrix} \right\|_{\mathcal{V}}^2 \geq \frac{1}{\max\{c, 1\}} \widetilde{\gamma} \left\| \begin{bmatrix} x \\ y \end{bmatrix} \right\|_{\mathcal{H}}^2$$

holds for all $\begin{bmatrix} x \\ y \end{bmatrix} \in \mathcal{D}(\overline{\mathscr{A}})$ and some $\widetilde{\gamma} > 0$ by Lemma 3.6.2 to see that $\overline{\mathscr{A}}$ is injective with closed range. In order to obtain $0 \in \varrho(\overline{\mathscr{A}})$ it is therefore sufficient to prove $\ker \mathscr{A}^* = \{0\}$. So let $\begin{bmatrix} u \\ v \end{bmatrix} \in \mathcal{D}(\mathscr{A}^*)$ be such that $\mathscr{A}^* \begin{bmatrix} u \\ v \end{bmatrix} = 0$. This leads to the two equations

$$A^*(u + \overline{A^{-*}C^*}v) = 0, \tag{3.30}$$

$$B^*u + D^*v = 0. \tag{3.31}$$

From (3.30) we obtain $u = -\overline{A^{-*}C^*}v$, because $0 \in \varrho(A^*)$ and thus (3.31) gives us

$$-B^*\overline{A^{-*}C^*}v + D^*v = 0. \tag{3.32}$$

By the same line of reasoning as in the beginning of the proof of Lemma 3.6.7 we see that $\overline{A^{-*}C^*} \in \mathcal{L}(\mathcal{H}_2, \mathcal{H}_1)$ and since B^* is closed, so is $B^*\overline{A^{-*}C^*}$. Therefore, $B^*A^{-*}C^*$ is closable and $\overline{B^*A^{-*}C^*} \subset B^*\overline{A^{-*}C^*}$. A^* is invertible and by assumption we have $A^* = \alpha C^*B^*|_{\mathcal{D}(A^*)}$, so $B^*|_{\mathcal{D}(A^*)}$ is injective and $A^{-*}C^*B^*|_{\mathcal{D}(A^*)} = \frac{1}{\alpha}I|_{\mathcal{D}(A^*)}$ holds. This implies

$$\begin{aligned} B^*A^{-*}C^*|_{\mathcal{R}(B^*|_{\mathcal{D}(A^*)})} &= B^*A^{-*}C^*B^*B^*|^{-1}_{\mathcal{D}(A^*)}|_{\mathcal{R}(B^*|_{\mathcal{D}(A^*)})} \\ &= \frac{1}{\alpha}I|_{\mathcal{R}(B^*|_{\mathcal{D}(A^*)})}. \end{aligned}$$

Hence, we obtain $B^*\overline{A^{-*}C^*} = \overline{B^*A^{-*}C^*} = \frac{1}{\alpha}I$ by using that $\mathcal{R}(B^*|_{\mathcal{D}(A^*)})$ is dense in $\mathcal{H}_2$. With this at hand, (3.32) simplifies to

$$\left(D^* - \frac{1}{\alpha}I\right)v = 0$$

and since we have $\frac{1}{\alpha} \notin \sigma_{\mathrm{p}}(D^*)$ by assumption this implies $v = 0$ and therefore also $u = 0$, i.e. $\ker \mathscr{A}^* = \{0\}$. Due to $\mathcal{R}(\overline{\mathscr{A}})$ being closed we can now conclude that $\mathcal{R}(\overline{\mathscr{A}}) = \mathcal{H}$ holds, so we have $0 \in \varrho(\overline{\mathscr{A}})$. From $\overline{\mathscr{A}} \subset \widetilde{\mathscr{A}}$ and $0 \in \varrho(\overline{\mathscr{A}}) \cap \varrho(\widetilde{\mathscr{A}})$ we then get the asserted equality $\overline{\mathscr{A}} = \widetilde{\mathscr{A}}$.

c) Here we can state the operator $\widetilde{\mathscr{A}}$ explicitly again. Let $\begin{bmatrix} x \\ y \end{bmatrix} \in \mathcal{D}(\widetilde{\mathscr{A}})$ and $\begin{bmatrix} f \\ g \end{bmatrix} \in \mathcal{H}$ be such that

$$\widetilde{a}\left(\begin{bmatrix} x \\ y \end{bmatrix}, \begin{bmatrix} u \\ v \end{bmatrix} \right) = \left\langle \begin{bmatrix} f \\ g \end{bmatrix}, \begin{bmatrix} u \\ v \end{bmatrix} \right\rangle \quad \text{for all} \quad \begin{bmatrix} u \\ v \end{bmatrix} \in \mathcal{V}.$$

Then we have

$$\widetilde{a}\left(\begin{bmatrix} x \\ y \end{bmatrix}, \begin{bmatrix} 0 \\ v \end{bmatrix} \right) = \langle -Cx - Dy, v \rangle_{\mathcal{H}_2} = \langle g, v \rangle_{\mathcal{H}_2}$$

for all $v \in \mathcal{H}_2$ which implies $g = -Cx - Dy$ and we also have

$$
\begin{aligned}
\widetilde{a}\left(\begin{bmatrix} x \\ y \end{bmatrix}, \begin{bmatrix} u \\ 0 \end{bmatrix} \right) &= \langle \alpha \boldsymbol{C}^* x, \boldsymbol{B} u \rangle_{\mathrm{F}} + \langle y, B^* u \rangle_{\mathcal{H}_2} \\
&= \langle \alpha \boldsymbol{C}^* x, \boldsymbol{B} u \rangle_{\mathrm{F}} + \left\langle y, \beta \sum_{i=1}^{n} (B u_i)_i \right\rangle_{\mathcal{H}_2} \\
&= \langle \alpha \boldsymbol{C}^* x, \boldsymbol{B} u \rangle_{\mathrm{F}} + \big\langle \overline{\beta} \operatorname{diag}(y, \dots, y), \boldsymbol{B} u \big\rangle_{\mathrm{F}} \\
&= \langle f, u \rangle
\end{aligned}
$$

for all $u \in \mathcal{H}_1$ which implies $\alpha \boldsymbol{C}^* x + \overline{\beta} \operatorname{diag}(y, \dots, y) \in \mathcal{D}(\boldsymbol{B}^*)$ and

$$
f = \boldsymbol{B}^* \big(\alpha \boldsymbol{C}^* x + \overline{\beta} \operatorname{diag}(y, \dots, y) \big).
$$

On the other hand, if we choose $\begin{bmatrix} x \\ y \end{bmatrix} \in \mathcal{V}$ such that

$$
\alpha \boldsymbol{C}^* x + \overline{\beta} \operatorname{diag}(y, \dots, y) \in \mathcal{D}(\boldsymbol{B}^*),
$$

we obtain that

$$
\begin{aligned}
\widetilde{a}\left(\begin{bmatrix} x \\ y \end{bmatrix}, \begin{bmatrix} u \\ v \end{bmatrix} \right) &= a(x, u) + \langle y, B^* u \rangle_{\mathcal{H}_2} - \langle Cx, v \rangle_{\mathcal{H}_2} - \langle Dy, v \rangle_{\mathcal{H}_2} \\
&= \langle \alpha \boldsymbol{C}^* x, \boldsymbol{B} u \rangle_{\mathrm{F}} + \big\langle \overline{\beta} \operatorname{diag}(y, \dots, y), \boldsymbol{B} u \big\rangle_{\mathrm{F}} \\
&\quad - \langle Cx + Dy, v \rangle_{\mathcal{H}_2} \\
&= \big\langle \boldsymbol{B}^* \big(\alpha \boldsymbol{C}^* x + \overline{\beta} \operatorname{diag}(y, \dots, y) \big), u \big\rangle_{\mathcal{H}_1} \\
&\quad - \langle Cx + Dy, v \rangle_{\mathcal{H}_2}
\end{aligned}
$$

holds for all $\begin{bmatrix} u \\ v \end{bmatrix} \in \mathcal{V}$. This shows

$$
J\widetilde{\mathscr{A}} \begin{bmatrix} x \\ y \end{bmatrix} = \begin{bmatrix} \boldsymbol{B}^* \big(\alpha \boldsymbol{C}^* x + \overline{\beta} \operatorname{diag}(y, \dots, y) \big) \\ -Cx - Dy \end{bmatrix}
$$

for all $\begin{bmatrix} x \\ y \end{bmatrix} \in \mathcal{D}(\widetilde{\mathscr{A}}) = \left\{ \begin{bmatrix} x \\ y \end{bmatrix} \in \mathcal{V} \,\middle|\, \alpha \boldsymbol{C}^* x + \overline{\beta} \operatorname{diag}(y, \dots, y) \in \mathcal{D}(\boldsymbol{B}^*) \right\}$. Due to $D(C^*)$ being dense in $\mathcal{H}_2$, $\overline{\mathscr{A}}$ is given as in Lemma 3.6.6 and since $\widetilde{\mathscr{A}}$ is also a closed extension of $\mathscr{A}$ we have $\overline{\mathscr{A}} \subset \widetilde{\mathscr{A}}$. Because of our assumption on a we have $A^* = \overline{\alpha}\overline{\boldsymbol{C}}\boldsymbol{B}$ and thus $\mathcal{D}(A^*) \subset \mathcal{D}(\boldsymbol{B})$. We therefore obtain the inclusion

$$
A^{-1} \boldsymbol{B}^* \subset (\boldsymbol{B} A^{-*})^* \in \mathcal{L}(\mathcal{H}_2^{n \times n}, \mathcal{H}_1),
$$

so $A^{-1} \boldsymbol{B}^*$ is bounded on $\mathcal{D}(\boldsymbol{B}^*)$. The closure of $A^{-1} \boldsymbol{B}^*$ is then given by $\overline{A^{-1} \boldsymbol{B}^*} = (\boldsymbol{B} A^{-*})^*$. In addition, we have $A = \alpha \boldsymbol{B}^* \boldsymbol{C}^* |_{\mathcal{D}(A)}$ and thus

$$
\frac{1}{\alpha} I |_{\mathcal{D}(A)} = A^{-1} \boldsymbol{B}^* \boldsymbol{C}^* |_{\mathcal{D}(A)} \subset (\boldsymbol{C} \boldsymbol{B} A^{-*})^* \in \mathcal{L}(\mathcal{H}_1)
$$

holds which implies $\frac{1}{\alpha}I = \overline{A^{-1}\boldsymbol{B}^*\boldsymbol{C}^*} = (\boldsymbol{CB}A^{-*})^*$ and because of $\overline{A^{-1}\boldsymbol{B}^*\boldsymbol{C}^*} \subset (\boldsymbol{CB}A^{-*})^*$ we can therefore deduce

$$\overline{\overline{A^{-1}\boldsymbol{B}^*}\boldsymbol{C}^*} = \frac{1}{\alpha}I|_{\mathcal{D}(\boldsymbol{C}^*)}. \tag{3.33}$$

Let $K \in \mathcal{H}_2^{n\times n}$ be such that each column $K_i \in \mathcal{H}_2^n = \mathcal{H}_1$, $i = 1, \dots, n$, is an element of $\mathcal{D}(B^*)$. Then for $u \in \mathcal{D}(\boldsymbol{B})$ we obtain

$$\begin{aligned}\langle \boldsymbol{B}u, K\rangle_{\mathrm{F}} &= \left\langle \begin{bmatrix} Bu_1 & \dots & Bu_n\end{bmatrix}, K\right\rangle_{\mathrm{F}} \\ &= \sum_{i=1}^n \langle Bu_i, K_i\rangle_{\mathcal{H}_1} = \sum_{i=1}^n \langle u_i, B^*K_i\rangle_{\mathcal{H}_2} \\ &= \left\langle \begin{bmatrix} u_1 \\ \vdots \\ u_n\end{bmatrix}, \begin{bmatrix} B^*K_1 \\ \vdots \\ B^*K_n\end{bmatrix}\right\rangle_{\mathcal{H}_1}.\end{aligned}$$

This implies $\mathcal{D}(B^*)^n \subset \mathcal{D}(\boldsymbol{B}^*)$ and for $K \in \mathcal{D}(B^*)^n$ we have

$$\boldsymbol{B}^*K = \begin{bmatrix} B^*K_1 \\ \vdots \\ B^*K_n\end{bmatrix}.$$

For $y \in \mathcal{H}_2$ define Y_j as the vector in $\mathcal{H}_1 = \mathcal{H}_2^n$ with y in the j-th component and 0 everywhere else. Then

$$\mathcal{V}_2 := \left\{ y \in \mathcal{H}_2 \,\middle|\, Y_j \in \mathcal{V}_1 \text{ for all } j = 1, \dots, n\right\} = \bigcap_{i=1}^n \mathcal{V}_{1,i}$$

is a dense subset of $\mathcal{H}_2$. Let now $y \in \mathcal{V}_2$. Then

$$B^*Y_j = \beta \sum_{i=1}^n (BY_{j,i})_i = \beta(By)_j$$

for all $j = 1, \dots, n$. For $Y := \begin{bmatrix} Y_1 & \dots & Y_n\end{bmatrix} = \operatorname{diag}(y, \dots, y)$ we thus obtain $Y \in \mathcal{D}(B^*)^n$ and $\boldsymbol{B}^*\operatorname{diag}(y, \dots, y) = \beta By$ which implies

$$A^{-1}\boldsymbol{B}^*\,\overline{\beta}\operatorname{diag}(y, \dots, y) = |\beta|^2 A^{-1}By = A^{-1}By.$$

Since $\mathcal{V}_2$ is dense in $\mathcal{H}_2$, $\overline{A^{-1}\boldsymbol{B}^*} \in \mathcal{L}(\mathcal{H}_2^{n\times n}, \mathcal{H}_1)$ and $\overline{A^{-1}B} \in \mathcal{L}(\mathcal{H}_2, \mathcal{H}_1)$ by the same line of reasoning as in the beginning of the proof of Lemma 3.6.6, we conclude that

$$\overline{A^{-1}\boldsymbol{B}^*}\,\overline{\beta}\operatorname{diag}(y, \dots, y) = \overline{A^{-1}B}y$$

holds for every $y \in \mathcal{H}_2$. Combining this with $\boldsymbol{B}^* \subset A\overline{A^{-1}\boldsymbol{B}^*}$ and (3.33) we see that for $x \in \mathcal{V}_1$ and $y \in \mathcal{H}_2$, $\alpha\boldsymbol{C}^*x + \overline{\beta}\operatorname{diag}(y, \dots, y) \in \mathcal{D}(\boldsymbol{B}^*)$

implies $x + \overline{A^{-1}B}y \in \mathcal{D}(A)$. This is because from $\alpha \boldsymbol{C}^* x + \overline{\beta}\operatorname{diag}(y, \dots, y) \in \mathcal{D}(A\overline{A^{-1}\boldsymbol{B}^*})$ we get

$$\overline{A^{-1}\boldsymbol{B}^*}\big(\alpha \boldsymbol{C}^* x + \overline{\beta}\operatorname{diag}(y, \dots, y)\big) = x + \overline{A^{-1}B}y \in \mathcal{D}(A).$$

The same arguments explain

$$\boldsymbol{B}^*\big(\alpha \boldsymbol{C}^* x + \overline{\beta}\operatorname{diag}(y, \dots, y)\big) = A(x + \overline{A^{-1}B}y)$$

for all $\begin{bmatrix} x \\ y \end{bmatrix} \in \mathcal{D}(\widetilde{\mathscr{A}})$. It remains to show that $y \in \mathcal{D}(\overline{CA^{-1}B})$. Since A is invertible and $A = \alpha \boldsymbol{B}^* \boldsymbol{C}^*|_{\mathcal{D}(A)}$, we have that $\boldsymbol{C}^*|_{\mathcal{D}(A)}$ is injective and therefore

$$\begin{aligned}
\boldsymbol{C}^* A^{-1} \boldsymbol{B}^*|_{\mathcal{R}(\boldsymbol{C}^*|_{\mathcal{D}(A)})} &= \boldsymbol{C}^* A^{-1} \boldsymbol{B}^* \boldsymbol{C}^* \boldsymbol{C}^*|_{\mathcal{D}(A)}^{-1}|_{\mathcal{R}(\boldsymbol{C}^*|_{\mathcal{D}(A)})} \\
&= \frac{1}{\alpha} I|_{\mathcal{R}(\boldsymbol{C}^*|_{\mathcal{D}(A)})}.
\end{aligned}$$

Now the assumption that $\mathcal{R}(\boldsymbol{C}^*|_{\mathcal{D}(A)})$ is dense in $\mathcal{H}_2$ implies that $\overline{\boldsymbol{C}^* A^{-1} \boldsymbol{B}^*} = \frac{1}{\alpha} I$. By our assumption on C we thus obtain for $y \in \mathcal{V}_2$ that

$$\begin{aligned}
CA^{-1}By = \gamma \sum_{i=1}^{n} \big(\boldsymbol{C}^* A^{-1} By\big)_{i,i} &= \frac{\gamma}{\beta} \sum_{i=1}^{n} (\boldsymbol{C}^* A^{-1} \boldsymbol{B}^* \operatorname{diag}(y, \dots, y))_{i,i} \\
&= \frac{\gamma}{\alpha\beta} \sum_{i=1}^{n} (\operatorname{diag}(y, \dots, y))_{i,i} \\
&= \frac{n\gamma}{\alpha\beta} y.
\end{aligned}$$

We conclude $CA^{-1}B|_{\mathcal{V}_2} = \frac{n\gamma}{\alpha\beta} I|_{\mathcal{V}_2}$ and due to $\mathcal{V}_2$ being dense in $\mathcal{H}_2$ we also have $\overline{CA^{-1}B} = \frac{n\gamma}{\alpha\beta} I$, i.e. $\mathcal{D}(\overline{CA^{-1}B}) = \mathcal{H}_2$. Thus, the inclusion $\widetilde{\mathscr{A}} \subset \overline{\mathscr{A}}$ also holds true. ❑

Remark 3.6.11. We see from the proof that the assumptions in a) that $\mathcal{R}(C|_{\mathcal{D}(A)})$ is dense in $\mathcal{H}_2$ and in c) that $Cx = \gamma \sum_{i=1}^{n} (\boldsymbol{C}^* x)_{i,i}$ for all $x \in \mathcal{V}_1$ and some $\gamma \in \mathbb{C}$ and that $\mathcal{R}(\boldsymbol{C}^*|_{D(A)})$ is dense in $\mathcal{H}_2^{n \times n}$ are only required to show that $y \in \mathcal{D}(\overline{CA^{-1}B})$. Hence, in case they are not fulfilled, we still obtain the inclusions $\overline{\mathscr{A}} \subset \widetilde{\mathscr{A}} \subset \widehat{\mathscr{A}}$ where

$$\widehat{\mathscr{A}} \begin{bmatrix} x \\ y \end{bmatrix} = \begin{bmatrix} A(x + \overline{A^{-1}B}y) \\ Cx + Dy \end{bmatrix}$$
$$\text{for all} \quad \begin{bmatrix} x \\ y \end{bmatrix} \in \mathcal{D}(\widehat{\mathscr{A}}) := \left\{ \begin{bmatrix} x \\ y \end{bmatrix} \in \mathcal{V} \,\middle|\, x + \overline{A^{-1}B}y \in \mathcal{D}(A) \right\}.$$

The following Example 3.6.12 is inspired by [5, Example 4.16] and deals with a Stokes-type block operator matrix that fits into the framework of this section. In the subsequent Example 3.6.13, this operator is then considered on a two-dimensional spacial domain.

Example 3.6.12. Let $\mathcal{H}_1 = \mathcal{H}_2 = L^2(0,1)$ and choose $\alpha_1, \alpha_2 \in \mathbb{C}\setminus\{0\}$ and $\gamma_D > 0$ such that either

$$\frac{5}{4}(|\alpha_1| + |\alpha_2|)^2 < \gamma_D \qquad \text{or} \qquad \alpha_2 = -\overline{\alpha}_1. \tag{3.34}$$

With this, consider the Stokes-type block operator matrix

$$\mathscr{A} = \begin{bmatrix} A & B \\ C & D \end{bmatrix} := \begin{bmatrix} -\frac{\mathrm{d}^2}{\mathrm{d}\xi^2} & \alpha_1 \frac{\mathrm{d}}{\mathrm{d}\xi} \\ \alpha_2 \frac{\mathrm{d}}{\mathrm{d}\xi} & -\gamma_D + \mathrm{i} \end{bmatrix}$$

where

$$\mathcal{D}(\mathscr{A}) = \left\{ x \in H^2(0,1) \,\middle|\, x(0) = \tfrac{\mathrm{d}}{\mathrm{d}\xi}x(1) = 0 \right\} \times \left\{ y \in H^1(0,1) \,\middle|\, y(1) = 0 \right\}.$$

Here, all the assumptions made in the beginning of this section are fulfilled:

a) $A\colon \mathcal{D}(A) \subset \mathcal{H}_1 \to \mathcal{H}_1$, $A = -\frac{\mathrm{d}^2}{\mathrm{d}\xi^2}$, with

$$\mathcal{D}(A) = \left\{ x \in H^2(0,1) \,\middle|\, x(0) = \tfrac{\mathrm{d}}{\mathrm{d}\xi}x(1) = 0 \right\}$$

is associated to the sesqui-linear form

$$a\colon \mathcal{V}_1 \times \mathcal{V}_1 \to \mathbb{C}, \qquad a(x,u) = \left\langle \tfrac{\mathrm{d}}{\mathrm{d}\xi}x, \tfrac{\mathrm{d}}{\mathrm{d}\xi}u \right\rangle_{\mathcal{H}_1},$$

where $\mathcal{V}_1 = \{ x \in H^1(0,1) \,|\, x(0) = 0 \}$ and we have

$$|a(x,u)| \le \|x\|_{\mathcal{V}_1} \|u\|_{\mathcal{V}_1} \qquad \text{and} \qquad \Re a(x,x) \ge \gamma_A \|x\|_{\mathcal{V}_1}^2$$

for all $x, u \in \mathcal{V}_1$ with $\gamma_A = \frac{1}{5}$ by the Poincaré-Friedrichs inequality, see [56, p. 231];

b) $B\colon \mathcal{D}(B) \subset \mathcal{H}_2 \to \mathcal{H}_1$, $B = \alpha_1 \frac{\mathrm{d}}{\mathrm{d}\xi}$, with

$$\mathcal{D}(B) = \{ y \in H^1(0,1) \,|\, y(1) = 0 \}$$

is densely defined, $\mathcal{D}(B^*) = \mathcal{V}_1$ and $\|B^*u\|_{\mathcal{H}} \le |\alpha_1| \|u\|_{\mathcal{V}_1}$ for all $u \in \mathcal{V}_1$;

c) $C\colon \mathcal{D}(C) \subset \mathcal{H}_1 \to \mathcal{H}_2$, $C = \alpha_2 \frac{\mathrm{d}}{\mathrm{d}\xi}$, with

$$\mathcal{D}(C) = \{ x \in H^1(0,1) \,|\, x(0) = 0 \} = \mathcal{V}_1$$

is densely defined and $\|Cx\|_{\mathcal{H}} \le |\alpha_2| \|x\|_{\mathcal{V}_1}$ for all $x \in \mathcal{V}_1$;

d) $D \in \mathcal{L}(\mathcal{H}_2)$, $y \mapsto (-\gamma_D + \mathrm{i})y$, satisfies $\Re\langle Dy, y\rangle_{\mathcal{H}_2} \le -\gamma_D \|y\|_{\mathcal{H}_2}^2$ for every $y \in \mathcal{H}_2$;

e) With $k_{B^*} = |\alpha_1|$ and $k_C = |\alpha_2|$ we either have $\frac{1}{4}(k_{B^*} + k_C)^2 < \gamma_A \gamma_D$ or $C = B^*$ by (3.34).

Moreover, $\mathcal{D}(C^*) = \mathcal{R}(C|_{\mathcal{D}(A)}) = \{v \in H^1(0,1) \,|\, v(1) = 0\}$ is dense in $L^2(0,1)$, B is closed and $a(x,u) = \alpha \langle Cx, B^*u\rangle$ holds for all $x, u \in \mathcal{V}_1$ where $\alpha = -\frac{1}{\alpha_1\alpha_2}$. Hence, Theorem 3.6.9 a) can be applied and we obtain $\widetilde{\mathscr{A}} = \overline{\mathscr{A}}$ with $\overline{\mathscr{A}}$ as in Lemma 3.6.6.

Additionally, the set $\mathcal{R}(B^*|_{\mathcal{D}(A^*)}) = \{u \in H^1(0,1) \,|\, u(1) = 0\} = \mathcal{D}(C^*)$ is dense in $L^2(0,1)$, we have $A^* = \overline{\alpha} C^* B^*$ and if we further assume $-\overline{\alpha}_1\overline{\alpha}_2 \neq -\gamma_D - \mathrm{i}$, we ensure $\frac{1}{\overline{\alpha}} \notin \sigma_{\mathrm{p}}(D^*)$. In this case, Theorem 3.6.9 b) can be applied as well.

Furthermore, Lemma 3.6.3 provides the existence of a spectrum free strip with a corresponding bound for the norm of the resolvent.

Example 3.6.13. Let $\Omega \subset \mathbb{R}^2$ be a bounded open domain with piecewise C^1-boundary contained in $[-s,s]^2$ for an $s > 0$, $\mathcal{H}_1 = L^2(\Omega)^2$, $\mathcal{H}_2 = L^2(\Omega)$ and choose $\alpha_1, \alpha_2 \in \mathbb{C} \setminus \{0\}$ and $\gamma_D > 0$ such that

$$(4s^2 + 1)(|\alpha_1| + |\alpha_2|)^2 < \gamma_D. \tag{3.35}$$

With this, consider the Stokes-type block operator matrix

$$\mathscr{A} = \begin{bmatrix} A & B \\ C & D \end{bmatrix} := \begin{bmatrix} -\boldsymbol{\Delta} & \alpha_1 \operatorname{grad} \\ \alpha_2 \operatorname{div} & -\gamma_D + \mathrm{i} \end{bmatrix}$$

where $\mathcal{D}(\mathscr{A}) = H^2(\Omega)^2 \cap H_0^1(\Omega)^2 \times H_0^1(\Omega)$. Here, all the assumptions made in the beginning of this section are fulfilled:

a) $A\colon \mathcal{D}(A) \subset \mathcal{H}_1 \to \mathcal{H}_1$, $A = -\boldsymbol{\Delta}$, is the vector valued Laplace operator with $\mathcal{D}(A) = H^2(\Omega)^2 \cap H_0^1(\Omega)^2$. A is associated to the sesqui-linear form

$$a\colon \mathcal{V}_1 \times \mathcal{V}_1 \to \mathbb{C}, \qquad a(x,u) = \langle \mathbf{grad}\, x, \mathbf{grad}\, u \rangle_{\mathrm{F}},$$

where $\mathcal{V}_1 = H_0^1(\Omega)^2$ and we have

$$|a(x,u)| \leq \|x\|_{\mathcal{V}_1} \|u\|_{\mathcal{V}_1} \qquad \text{and} \qquad \Re a(x,x) \geq \gamma_A \|x\|_{\mathcal{V}_1}^2$$

for all $x, u \in \mathcal{V}_1$ with $\gamma_A = \frac{1}{4s^2+1}$ by the Poincaré-Friedrichs inequality, see [56, p. 231];

b) $B\colon \mathcal{D}(B) \subset \mathcal{H}_2 \to \mathcal{H}_1$, $B = \alpha_1 \operatorname{grad}$, with $\mathcal{D}(B) = H_0^1(\Omega)$ is densely defined, $\mathcal{V}_1 \subset H^1(\Omega)^2 \subset \mathcal{D}(B^*) = \{u \in L^2(\Omega)^2 \,|\, \operatorname{div} u \in L^2(\Omega)\}$ and $\|B^*u\|_{\mathcal{H}_2} \leq 2|\alpha_1| \|u\|_{\mathcal{V}_1}$ for all $u \in \mathcal{V}_1$;

c) $C\colon \mathcal{D}(C) \subset \mathcal{H}_1 \to \mathcal{H}_2$, $C = \alpha_2 \operatorname{div}$, with $\mathcal{V}_1 \subset \mathcal{D}(C) = H^1(\Omega)^2$ is densely defined and $\|Cx\|_{\mathcal{H}_2} \leq 2|\alpha_2| \|x\|_{\mathcal{V}_1}$ for all $x \in \mathcal{V}_1$;

d) $D \in \mathcal{L}(\mathcal{H}_2)$, $y \mapsto (-\gamma_D + \mathrm{i})y$, satisfies $\Re\langle Dy, y\rangle_{\mathcal{H}_2} \leq -\gamma_D \|y\|_{\mathcal{H}_2}^2$ for every $y \in \mathcal{H}_2$;

e) With $k_{B^*} = 2|\alpha_1|$ and $k_C = 2|\alpha_2|$ we have $\frac{1}{4}(k_{B^*} + k_C)^2 < \gamma_A \gamma_D$ by (3.35).

Here, C^* is given by $-\overline{\alpha}_2 \operatorname{grad}$ and via integration by parts we see that our regularity assumptions on $\partial\Omega$ yield $\mathcal{D}(C^*) = H_0^1(\Omega)$ which is dense in $L^2(\Omega)$. Moreover, $\mathcal{H}_1 = \mathcal{H}_2^2$, $\mathcal{V}_1 = \mathcal{D}(C^*)^2 = \mathcal{D}(B)^2$ and with

$$\boldsymbol{C}\colon \mathcal{D}(C)^2 \subset \mathcal{H}_2^{2\times 2} \to \mathcal{H}_1, \quad \boldsymbol{C}K = \begin{bmatrix} CK_1 \\ CK_2 \end{bmatrix},$$

$a(x,u) = -\frac{1}{\overline{\alpha}_1\overline{\alpha}_2}\langle \boldsymbol{C}^*x, \boldsymbol{B}u\rangle$, $B^*u = -\frac{\overline{\alpha}_1}{\alpha_1}\big((Bu_1)_1 + (Bu_2)_2\big)$ and $Cx = \frac{1}{\alpha_2}\big((\boldsymbol{C}^*x)_{1,1} + (\boldsymbol{C}^*x)_{2,2}\big)$ hold for all $x, u \in \mathcal{V}_1$. However, $\mathcal{R}(\boldsymbol{C}^*|_{\mathcal{D}(A)})$ is orthogonal to

$$\left\{K \in H^1(\Omega)^{2\times 2} \,\middle|\, \begin{bmatrix} \operatorname{div} K_1 \\ \operatorname{div} K_2 \end{bmatrix} = 0\right\}$$

and therefore not dense in $\mathcal{H}_2^{2\times 2}$. Hence, by Remark 3.6.11 we have $\overline{\mathscr{A}} \subset \widetilde{\mathscr{A}} \subset \widehat{\mathscr{A}}$ with $\overline{\mathscr{A}}$ as in Lemma 3.6.6.

Furthermore, Lemma 3.6.3 provides the existence of a spectrum free strip with a corresponding bound for the norm of the resolvent.

3.7 Numerical Examples

In order to exemplify the previously developed theory we take a look at the results of numerical computations. We investigate the steps that were involved in the discretization of a given operator and describe a visualization of supersets of the pseudospectrum that was created by using Matlab.

Example 3.7.1. Let us consider the the advection-diffusion operator $A\colon \mathcal{D}(A) \subset L^2(0,1) \to L^2(0,1)$ defined by

$$A = \eta \frac{\mathrm{d}^2}{\mathrm{d}\xi^2} + \frac{\mathrm{d}}{\mathrm{d}\xi}$$

with $\mathcal{D}(A) = \big\{x \in H^2(0,1) \,\big|\, x(0) = x(1) = 0\big\}$, which has also been examined in [51, pp. 115]. For $x \in \mathcal{D}(A)$ and $u \in C^\infty(0,1)$ we have

$$\begin{aligned} \langle Ax, u\rangle &= \int_0^1 \left(\eta \frac{\mathrm{d}^2}{\mathrm{d}\xi^2} x(\xi) + \frac{\mathrm{d}}{\mathrm{d}\xi} x(\xi)\right) \overline{u(\xi)}\,\mathrm{d}\xi \\ &= \int_0^1 \frac{\mathrm{d}}{\mathrm{d}\xi} x(\xi)\overline{u(\xi)} - \eta \frac{\mathrm{d}}{\mathrm{d}\xi} x(\xi) \frac{\mathrm{d}}{\mathrm{d}\xi}\overline{u(\xi)}\,\mathrm{d}\xi \\ &=: a(x,u). \end{aligned} \tag{3.36}$$

Let $\big\{\mathcal{T}_{\frac{1}{n}}\big\}_{n\in\mathbb{N}}$ be the family of decompositions of the interval $(0,1)$ where every subinterval $T \in \mathcal{T}_{\frac{1}{n}}$ is of length $\frac{1}{n}$ and set

$$\mathcal{U}_n = \left\{x \in C(0,1) \,\middle|\, x|_T \in \mathbb{P}_1(T), T \in \mathcal{T}_{\frac{1}{n}}, x(0) = x(1) = 0\right\}$$

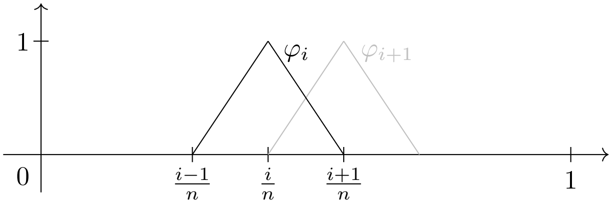


Figure 3.4: 'Hat' function φ_i

for $n \in \mathbb{N}$. Here, $\mathbb{P}_1(T)$ denotes the set of polynomials of degree 1 on the subinterval T. The piecewise linear 'hat' functions

$$\varphi_i = \begin{cases} n\xi - i + 1, & \xi \in (\frac{i-1}{n}, \frac{i}{n}), \\ i + 1 - n\xi, & \xi \in (\frac{i}{n}, \frac{i+1}{n}), \\ 0, & \text{else}, \end{cases}$$

for $i \in \{1, \dots, n-1\}$ form a basis of $\mathcal{U}_n$, cf. Figure 3.4. Evaluating (3.36) at these basis functions, the finite-element discretization matrices A_n of A are given by

$$A_n = \Big((a(\varphi_i, \varphi_j))_{i,j} \cdot (\langle \varphi_i, \varphi_j \rangle)_{i,j}^{-1} \Big)^{\mathsf{T}}.$$

With the choice of $\eta = 0.015$, Figure 3.5 shows the eigenvalues of A_n for $n = 40$ (red) and the sets

$$\big(\mathrm{B}_{\delta_j}(W((A_n - s_j)^{-1}))\big)^{-1} + s_j$$

(blue) for a number of shifts $s_1, \dots, s_m$ where $\delta_j = 1.1 \frac{\|(A_n - s_j)^{-1}\|^2 \varepsilon}{1 - \|(A_n - s_j)^{-1}\| \varepsilon}$ and $\varepsilon \approx 16$.

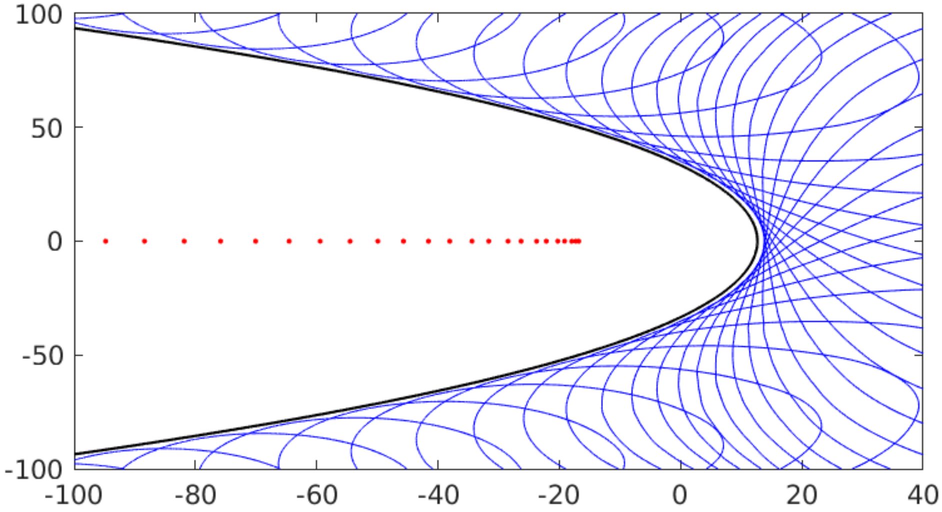


Figure 3.5: Pseudospectrum approximation for the advection-diffusion operator

The shifts are located at a certain distance to the expected pseudospectrum so as to obtain a relatively small superset thereof. The black line corresponds to the boundary of $\sigma_\varepsilon(A_n)$ computed by `EigTool`, see [13]. This demonstrates the result of Theorem 3.2.6 which actually yields an enclosure for the pseudospectrum of the operator A while the black line only shows the boundary of the pseudospectrum of the approximation matrix A_n.

Example 3.7.2. In this example we will reconsider the Hain-Lüst operator from Example 3.5.5 which fits into the framework of section 3.5. We recall that the Hain-Lüst operator is defined by

$$\mathscr{A} = \begin{bmatrix} A & B \\ B^* & D \end{bmatrix} = \begin{bmatrix} -\frac{1}{100}\frac{\mathrm{d}^2}{\mathrm{d}\xi^2} + 2 & I \\ I & 2\mathrm{e}^{2\pi\mathrm{i}\cdot} - 3 \end{bmatrix}$$

on the Hilbert space $\mathcal{H} := L^2(0,1) \times L^2(0,1)$ with

$$\begin{aligned} \mathcal{D}(A) &= \big\{ u \in H^2(0,1) \,\big|\, u(0) = u(1) = 0 \big\}, \\ \mathcal{D}(B) &= \mathcal{D}(D) = L^2(0,1) \end{aligned}$$

and $\mathcal{D}(\mathscr{A}) = \mathcal{D}(A) \times \mathcal{D}(D)$. Hence, for $\begin{bmatrix} x \\ y \end{bmatrix} \in \mathcal{D}(\mathscr{A})$ and $\begin{bmatrix} u \\ v \end{bmatrix} \in C^\infty(0,1) \times C^\infty(0,1)$ with $u(0) = u(1) = v(0) = v(1) = 0$ we have

$$\begin{aligned} \left\langle \mathscr{A} \begin{bmatrix} x \\ y \end{bmatrix}, \begin{bmatrix} u \\ v \end{bmatrix} \right\rangle &= \int_0^1 \left(\left(-\frac{1}{100}\frac{\mathrm{d}^2}{\mathrm{d}\xi^2} + 2 \right) x(\xi) + y(\xi) \right) \overline{u(\xi)} \,\mathrm{d}\xi \\ &\quad + \int_0^1 \big(x(\xi) + \big(2\mathrm{e}^{2\pi\mathrm{i}\xi} - 3\big)\, y(\xi) \big) \overline{v(\xi)} \,\mathrm{d}\xi \\ &= \int_0^1 \frac{1}{100} \frac{\mathrm{d}}{\mathrm{d}\xi} x(\xi) \frac{\mathrm{d}}{\mathrm{d}\xi} \overline{u(\xi)} + \big(2x(\xi) + y(\xi)\big) \overline{u(\xi)} \,\mathrm{d}\xi \\ &\quad + \int_0^1 \big(x(\xi) + \big(2\mathrm{e}^{2\pi\mathrm{i}\xi} - 3\big)\, y(\xi) \big) \overline{v(\xi)} \,\mathrm{d}\xi \qquad (3.37) \\ &=: a\left(\begin{bmatrix} x \\ y \end{bmatrix}, \begin{bmatrix} u \\ v \end{bmatrix} \right). \end{aligned}$$

As in the previous example, let $\{\mathcal{T}_{\frac{1}{n}}\}_{n\in\mathbb{N}}$ be the family of decompositions of the interval $(0,1)$ where every subinterval $T \in \mathcal{T}_{\frac{1}{n}}$ is of length $\frac{1}{n}$ and set

$$\mathcal{U}_{1,n} = \mathcal{U}_{2,n} = \Big\{ u \in C(0,1) \,\Big|\, u|_T \in \mathbb{P}_1(T), T \in \mathcal{T}_{\frac{1}{n}}, u(0) = u(1) = 0 \Big\}$$

for $n \in \mathbb{N}$. Here, $\mathbb{P}_1(T)$ denotes the set of polynomials of degree 1 on the subinterval T. The piecewise linear 'hat' functions

$$\widetilde{\varphi}_i = \begin{cases} nx - i + 1, & x \in (\frac{i-1}{n}, \frac{i}{n}), \\ i + 1 - nx, & x \in (\frac{i}{n}, \frac{i+1}{n}), \\ 0, & \text{else}, \end{cases}$$

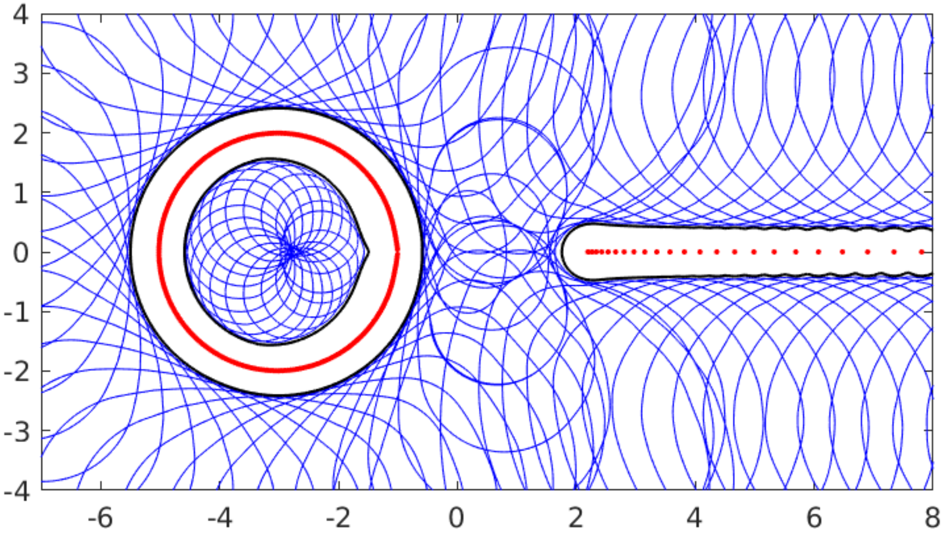


Figure 3.6: Pseudospectrum approximation for the Hain-Lüst block operator matrix

for $i \in \{1, \dots, n-1\}$ form a basis of $\mathcal{U}_{j,n}$, $j = 1, 2$, and therefore the functions

$$\varphi_i = \begin{cases} (\widetilde{\varphi}_i, 0), & i \leq n-1, \\ (0, \widetilde{\varphi}_{i-n+1}), & i > n-1, \end{cases}$$

for $i \in \{1, \dots, 2(n-1)\}$ form a basis of $\mathcal{U}_{1,n} \times \mathcal{U}_{2,n}$. Evaluating (3.37) on these basis functions, the finite-element discretization matrices $\mathscr{A}_n$ of $\mathscr{A}$ are given by

$$\mathscr{A}_n = \Big(\big(a(\varphi_i, \varphi_j)\big)_{i,j} \cdot \big(\langle \varphi_i, \varphi_j \rangle\big)_{i,j}^{-1} \Big)^{\mathsf{T}}.$$

Due to Lemma 3.5.2, Theorem 3.2.6 can be applied here. In order to illustrate the inclusion specified therein the boundaries of the sets

$$\big(\mathrm{B}_{\delta_j}(W((\mathscr{A}_n - s_j)^{-1}))\big)^{-1} + s_j$$

(blue) are depicted in Figure 3.6 for shifts $s_1, \dots, s_m \in \varrho(\mathscr{A})$. The choice of the shifts was determined by the expected shape of the pseudospectrum aiming to obtain a relatively small superset thereof. They are located on two circles around -3 with radii greater and smaller than 2 and on lines parallel to the real axis in the right half plane. Here, $n = 600$, $\delta_j = 1.1 \frac{\|(\mathscr{A}_n - s_j)^{-1}\|^2 \varepsilon}{1 - \|(\mathscr{A}_n - s_j)^{-1}\| \varepsilon}$ and $\varepsilon \approx 0.4$. The red dots are the eigenvalues of $\mathscr{A}_n$ while the black lines correspond to the boundaries of the pseudospectrum of the approximation matrix $\sigma_\varepsilon(\mathscr{A}_n)$ computed by `EigTool`, see [13]. Note that according to Theorem 3.2.6 the intersection of the blue areas form an enclosure of the pseudospectrum of the actual operator $\sigma_\varepsilon(\mathscr{A})$, while the black lines only give the information for the discretized operator. Furthermore, a large portion of the spectrum free strip $\big\{\lambda \in \mathbb{C} \,\big|\, -1 < \Re\lambda < 2 + \frac{1}{400}\big\} \subset \varrho(\mathscr{A})$ mentioned in Example 3.5.5 becomes visible.

Remark 3.7.3. As already mentioned in Remark 3.1.9, we also have the enclosure

$$\sigma_\varepsilon(A) \subset \mathrm{B}_\varepsilon(W(A))$$

for operators A with a compact resolvent.

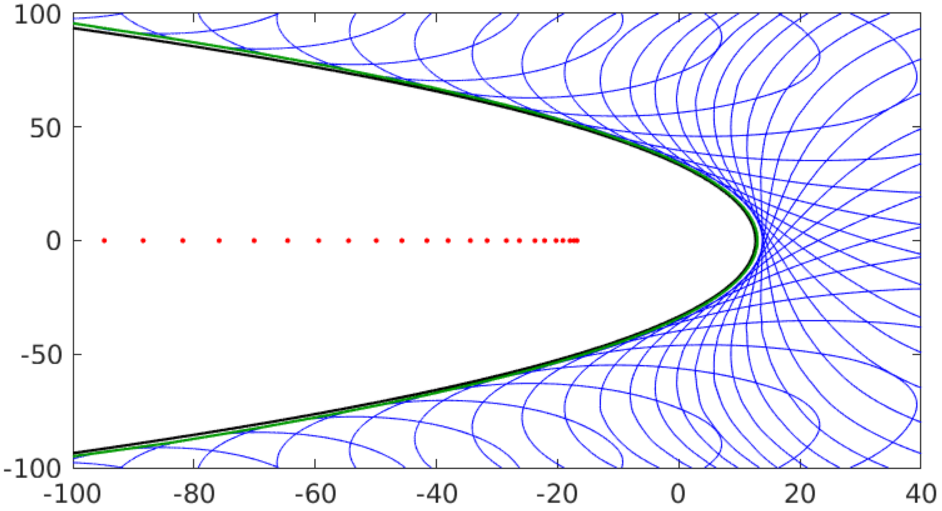


Figure 3.7: ε-neighborhood of the numerical range of the advection-diffusion operator

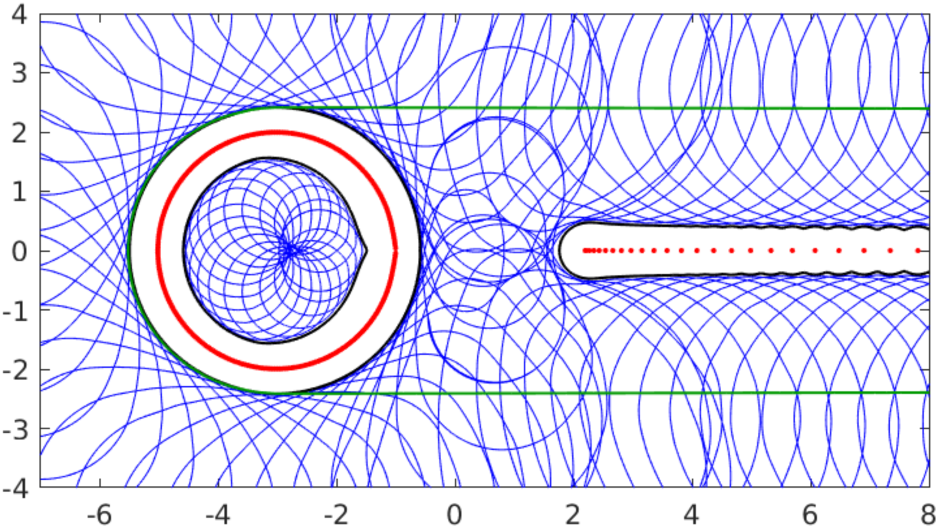


Figure 3.8: ε-neighborhood of the numerical range of the Hain-Lüst block operator matrix

Note that, because both sides of the enclosure are in terms of the same operator A, this only yields an enclosure for the discretized operator when applied numerically, not the full operator. So let us take a look at the discretizations of the advection-diffusion operator (Figure 3.7) and the Hain-Lüst operator (Figure 3.8) again. Here, the ε-neighborhoods of the numerical ranges are depicted by green lines. As you can see, this approach leads to a very similar result in case of the advection-diffusion operator (where the pseudospectrum is convex), while it fails to distinguish disconnected components of the pseudospectrum in case of the Hain-Lüst operator.

Example 3.7.4. Let us reconsider the Stokes-type block operator matrix from Example 3.6.12 that fits into the framework of Section 3.6. Here, $\mathcal{H}_1 = \mathcal{H}_2 = L^2(0,1)$, $\mathcal{V}_1 = \big\{x \in H^1(0,1) \,\big|\, x(0) = 0\big\}$ and

$$\mathscr{A} = \begin{bmatrix} A & B \\ C & D \end{bmatrix} := \begin{bmatrix} -\frac{\mathrm{d}^2}{\mathrm{d}\xi^2} & \frac{1}{100}\frac{\mathrm{d}}{\mathrm{d}\xi} \\ \frac{1}{100}\frac{\mathrm{d}}{\mathrm{d}\xi} & -1+\mathrm{i} \end{bmatrix}$$

with

$$\mathcal{D}(\mathscr{A}) = \Big\{x \in H^2(0,1) \,\Big|\, x(0) = \tfrac{\mathrm{d}}{\mathrm{d}\xi}x(1) = 0\Big\} \times \big\{y \in H^1(0,1) \,\big|\, y(1) = 0\big\}.$$

As we have already seen, $\widetilde{\mathscr{A}} = \overline{\mathscr{A}}$ holds in this case where $\overline{\mathscr{A}}$ is given as in Lemma 3.6.6.
Let further $\big\{\mathcal{T}_{\frac{1}{n}}\big\}_{n\in\mathbb{N}}$ be the family of subintervals of $(0,1)$ of length $\frac{1}{n}$ and set

$$\mathcal{U}_{1,n} = \mathcal{U}_{2,n} = \Big\{x \in C(0,1) \,\Big|\, x|_T \in \mathbb{P}_1(T),\ T \in \mathcal{T}_{\frac{1}{n}},\ x(0) = 0\Big\} \subset \mathcal{V}_1$$

for $n \in \mathbb{N}$ where $\mathbb{P}_1(T)$ denotes the set of polynomials of degree 1 on the subinterval T. We will now show that every $x \in \mathcal{V}_1$ can be approximated by a function in $\mathcal{U}_{1,n}$ in the H^1-norm. Let therefore $x \in \mathcal{V}_1$ and $\varepsilon > 0$. Since $C^\infty(0,1)$ is dense in $L^2(0,1)$, there exists an $f \in C^\infty(0,1)$ such that

$$\Big\|\tfrac{\mathrm{d}}{\mathrm{d}\xi}x - \tfrac{\mathrm{d}}{\mathrm{d}\xi}f\Big\|_{L^2} < \frac{\varepsilon}{2}.$$

$\frac{\mathrm{d}}{\mathrm{d}\xi}f \in C^\infty(0,1)$ on the other hand is uniformly continuous on $[0,1]$ and thus there exists a $\delta > 0$ such that for all $\xi_1, \xi_2 \in [0,1]$ with $|\xi_1 - \xi_2| < \delta$ we have

$$\Big|\tfrac{\mathrm{d}}{\mathrm{d}\xi}f(\xi_1) - \tfrac{\mathrm{d}}{\mathrm{d}\xi}f(\xi_2)\Big| < \frac{\varepsilon}{2}.$$

Now take $n \in \mathbb{N}$ such that $\frac{1}{n} < \delta$ and choose a piecewise constant function v_n that takes an arbitrary value of $\frac{\mathrm{d}}{\mathrm{d}\xi}f$ in every $T \in \mathcal{T}_{\frac{1}{n}}$. Then

$$\Big\|\tfrac{\mathrm{d}}{\mathrm{d}\xi}f - v_n\Big\|_\infty < \frac{\varepsilon}{2}$$

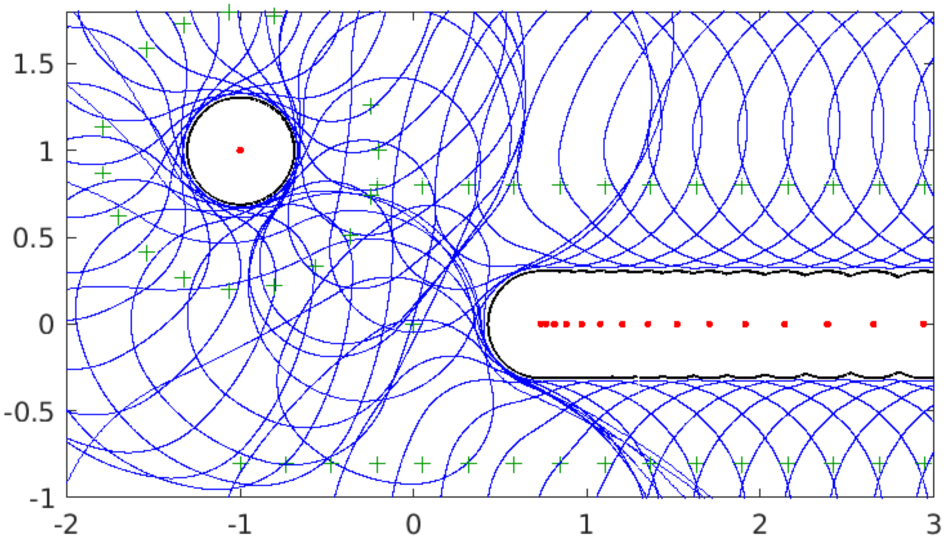


Figure 3.9: Pseudospectrum approximation for the Stokes-type block operator matrix

and hence

$$\begin{aligned}\left\|\tfrac{\mathrm{d}}{\mathrm{d}\xi}x - v_n\right\|_{L^2} &\leq \left\|\tfrac{\mathrm{d}}{\mathrm{d}\xi}x - \tfrac{\mathrm{d}}{\mathrm{d}\xi}f\right\|_{L^2} + \left\|\tfrac{\mathrm{d}}{\mathrm{d}\xi}f - v_n\right\|_{L^2} \\ &\leq \left\|\tfrac{\mathrm{d}}{\mathrm{d}\xi}x - \tfrac{\mathrm{d}}{\mathrm{d}\xi}f\right\|_{L^2} + \left\|\tfrac{\mathrm{d}}{\mathrm{d}\xi}f - v_n\right\|_{\infty} \\ &< \frac{\varepsilon}{2} + \frac{\varepsilon}{2} = \varepsilon.\end{aligned}$$

With this, defining u_n via

$$u_n(\xi) := \int_0^{\xi} v_n(s)\,\mathrm{d}s \quad \text{for all} \quad \xi \in [0,1],$$

we obtain $u_n \in \mathcal{U}_{1,n}$ and

$$\left\|\tfrac{\mathrm{d}}{\mathrm{d}\xi}x - \tfrac{\mathrm{d}}{\mathrm{d}\xi}u_n\right\|_{L^2} < \varepsilon.$$

It therefore remains to verify

$$\|x - u_n\|_{L^2} < \varepsilon,$$

which holds true because by using $x(0) = u_n(0) = 0$ we have

$$\begin{aligned}\|x - u_n\|_{L^2}^2 &= \int_0^1 |x(s) - u_n(s)|^2\,\mathrm{d}s \\ &= \int_0^1 \left|\int_0^s \tfrac{\mathrm{d}}{\mathrm{d}\xi}x(t)\,\mathrm{d}t - \int_0^s v_n(t)\,\mathrm{d}t\right|^2 \mathrm{d}s\end{aligned}$$

$$
\begin{aligned}
&\leq \int_0^1 \left(\int_0^s \left| \tfrac{\mathrm{d}}{\mathrm{d}\xi} x(t) - v_n(t) \right| \mathrm{d}t \right)^2 \mathrm{d}s \\
&\leq \left\| \tfrac{\mathrm{d}}{\mathrm{d}\xi} x - v_n \right\|_{L^1}^2 \leq \left\| \tfrac{\mathrm{d}}{\mathrm{d}\xi} x - v_n \right\|_{L^2}^2 \\
&< \varepsilon^2.
\end{aligned}
$$

Therefore, all the assumptions of Theorem 3.6.4 are fulfilled and thus Theorem 3.2.6 can be applied in order to obtain

$$
\sigma_\varepsilon(\widetilde{\mathscr{A}}) \subset \bigcap_{s \in S} \left[\left(\mathrm{B}_{\delta_s}(W((\widetilde{\mathscr{A}_n} - s)^{-1})) \right)^{-1} + s \right]
$$

for a suitable set S. The strong approximation matrices $\widetilde{\mathscr{A}_n}$ are again obtained via finite-element discretization by first constructing the basis 'hat' functions

$$
\widetilde{\varphi}_i = \begin{cases} nx - i + 1, & x \in (\frac{i-1}{n}, \frac{i}{n}), \\ i + 1 - nx, & x \in (\frac{i}{n}, \frac{i+1}{n}), \\ 0, & \text{else}, \end{cases}
$$

for $i \in \{1, \ldots, n-1\}$ that form a basis of $\mathcal{U}_{j,n}$, $j = 1, 2$, and

$$
\varphi_i = \begin{cases} (\widetilde{\varphi}_i, 0), & i \leq n-1, \\ (0, \widetilde{\varphi}_{i-n+1}), & i > n-1, \end{cases}
$$

for $i \in \{1, \ldots, 2(n-1)\}$ that form a basis of $\mathcal{U}_{1,n} \times \mathcal{U}_{2,n}$ and by then computing

$$
\widetilde{\mathscr{A}_n} = \left((\widetilde{a}(\varphi_i, \varphi_j))_{i,j} \cdot (\langle \varphi_i, \varphi_j \rangle)_{i,j}^{-1} \right)^{\mathsf{T}}.
$$

Figure 3.9 depicts the eigenvalues of $\widetilde{\mathscr{A}_n}$ for $n = 180$ (red) and the sets

$$
\left(\mathrm{B}_{\delta_j}(W((A_n - s_j)^{-1})) \right)^{-1} + s_j
$$

(blue) for a number of shifts $s_1, \ldots, s_m$ where $\delta_j = 1.1 \frac{\|(A_n - s_j)^{-1}\|^2 \varepsilon}{1 - \|(A_n - s_j)^{-1}\| \varepsilon}$ and $\varepsilon \approx 0.3$. The choice of the shifts was determined by the expected shape of the pseudospectrum aiming to obtain a relatively small superset thereof. Here, the positions of the chosen shifts are illustrated by green crosses that are located at the origin, on a circle with radius $4/5$ around $-1 + \mathrm{i}$ as well as above and below the real axis with a distance of $4/5$ to $\{z \in \mathbb{C} \mid \Im z = 0\}$. The black lines correspond to the boundary of $\sigma_\varepsilon(\widetilde{\mathscr{A}_n})$ computed by `EigTool`, see [13]. This demonstrates the result of Theorem 3.2.6 which actually yields an enclosure for the pseudospectrum of the operator $\widetilde{\mathscr{A}} = \overline{\mathscr{A}}$ while the black line only shows the boundary of the pseudospectrum of the approximation matrix $\widetilde{\mathscr{A}_n}$. Moreover, the plot shows a part of the spectrum free strip $\left\{ \lambda \in \mathbb{C} \,\middle|\, -1 < \Re\lambda < \frac{1}{5} \right\} \subset \varrho(\overline{\mathscr{A}})$ from Lemma 3.6.3.

3.8 Pseudospectra of Schur Complements

In this section we will relate the pseudospectra of a certain class of block operator matrices to the pseudospectra of their Schur complements. Before we can state this relation and the required assumptions precisely, we first have to introduce a couple of definitions. Let therefore $\mathcal{X}_1$ be a Banach space such that $\mathcal{X} = \mathcal{X}_1 \times \mathcal{X}_1$. In the following, all appearing matrix norms are the operator norm induced by the vector 2-norm.

Definition 3.8.1. Let $0 \le \omega < \pi$. A linear operator $T\colon \mathcal{D}(T) \subset \mathcal{X}_1 \to \mathcal{X}_1$ is called *sectorial* of angle ω, if $\sigma(T) \subset \overline{\mathrm{Sec}_\omega}$ and

$$\sup\left\{\|\lambda(T-\lambda)^{-1}\| \,\middle|\, \lambda \notin \overline{\mathrm{Sec}_{\omega'}}\right\} < \infty$$

for all $\omega < \omega' < \pi$. Here, $\mathrm{Sec}_\omega \coloneqq \{z \in \mathbb{C} \,|\, z \neq 0 \text{ and } |\arg z| < \omega\}$ for $0 < \omega < \pi$ and $\mathrm{Sec}_0 \coloneqq (0,\infty)$.

For a thorough introduction to sectorial operators we refer to [24].

Throughout this section, we will consider specific structured block operator matrices $\mathscr{A}\colon \mathcal{D}(\mathscr{A}) \subset \mathcal{X} \to \mathcal{X}$ of the form

$$\mathscr{A} = \begin{bmatrix} T^2 & \beta T \\ \gamma T & \delta \end{bmatrix},$$

where $T\colon \mathcal{D}(T) \subset \mathcal{X}_1 \to \mathcal{X}_1$ is a densely defined sectorial operator of some angle $0 \le \omega < \pi$ with $\varrho(T^2) \neq \varnothing$, $\beta, \gamma, \delta \in \mathbb{C}$ and $\mathcal{D}(\mathscr{A}) = \mathcal{D}(T^2) \times \mathcal{D}(T)$.

Recall from Definition 1.1.9 that the Schur complement of $\mathscr{A}$ is given by

$$S(\lambda) = \delta - \lambda - \beta\gamma T(T^2-\lambda)^{-1}T \quad \text{for} \quad \lambda \in \varrho(T^2)$$

and that its spectrum and resolvent are defined by

$$\begin{aligned} \sigma(S) &= \left\{\lambda \in \varrho(T^2) \,\middle|\, 0 \in \sigma(S(\lambda))\right\} \\ \text{and} \quad \varrho(S) &= \left\{\lambda \in \varrho(T^2) \,\middle|\, 0 \in \varrho(S(\lambda))\right\} \end{aligned}$$

respectively. Similarly, we now define the pseudospectrum of the operator function S by

$$\sigma_\varepsilon(S) = \left\{\lambda \in \varrho(T^2) \,\middle|\, 0 \in \sigma_\varepsilon(S(\lambda))\right\} \quad \text{for} \quad \varepsilon > 0.$$

We have that

$$(T^2-\lambda)^{-1}T \subset T(T^2-\lambda)^{-1} \in \mathcal{L}(\mathcal{X}_1).$$

Hence, $(T^2-\lambda)^{-1}T$ is bounded on $D(T)$ and the closure is given by

$$\overline{(T^2-\lambda)^{-1}T} = T(T^2-\lambda)^{-1}$$

since T is densely defined. We also obtain

$$T(T^2-\lambda)^{-1}T \subset T^2(T^2-\lambda)^{-1} \in \mathcal{L}(\mathcal{X}_1)$$

and thus

$$\overline{S(\lambda)} = \delta - \lambda - \beta\gamma T^2 (T^2 - \lambda)^{-1}. \tag{3.38}$$

With Theorem 1.1.10 we conclude that for an arbitrary $\lambda \in \varrho(T^2)$ the closure $\overline{\mathscr{A}}$ is given by

$$\begin{aligned}
&(\overline{\mathscr{A}} - \lambda) \begin{bmatrix} x \\ y \end{bmatrix} \\
&= \begin{bmatrix} I & 0 \\ \gamma T(T^2-\lambda)^{-1} & I \end{bmatrix} \begin{bmatrix} T^2 - \lambda & 0 \\ 0 & \overline{S(\lambda)} \end{bmatrix} \begin{bmatrix} I & \beta\overline{(T^2-\lambda)^{-1}T} \\ 0 & I \end{bmatrix} \begin{bmatrix} x \\ y \end{bmatrix} \\
&= \begin{bmatrix} I & 0 \\ \gamma T(T^2-\lambda)^{-1} & I \end{bmatrix} \begin{bmatrix} T^2 - \lambda & 0 \\ 0 & \overline{S(\lambda)} \end{bmatrix} \begin{bmatrix} I & \beta T(T^2-\lambda)^{-1} \\ 0 & I \end{bmatrix} \begin{bmatrix} x \\ y \end{bmatrix} \\
&= \begin{bmatrix} I & 0 \\ \gamma T(T^2-\lambda)^{-1} & I \end{bmatrix} \begin{bmatrix} (T^2-\lambda)(x + \beta T(T^2-\lambda)^{-1}y) \\ \overline{S(\lambda)}y \end{bmatrix} \\
&= \begin{bmatrix} (T^2-\lambda)(x + \beta T(T^2-\lambda)^{-1}y) \\ \gamma Tx + \beta\gamma T^2(T^2-\lambda)^{-1}y + \overline{S(\lambda)}y \end{bmatrix} \\
&= \begin{bmatrix} (T^2-\lambda)(x + \beta T(T^2-\lambda)^{-1}y) \\ \gamma Tx + (\delta - \lambda)y \end{bmatrix}
\end{aligned}$$

for all $\begin{bmatrix} x \\ y \end{bmatrix} \in \mathcal{D}(\overline{\mathscr{A}})$ where

$$\mathcal{D}(\overline{\mathscr{A}}) = \left\{ \begin{bmatrix} x \\ y \end{bmatrix} \in \mathcal{X} \,\middle|\, x + \beta T(T^2-\lambda)^{-1}y \in \mathcal{D}(T^2) \right\}.$$

Note, that $\beta T(T^2-\lambda)^{-1}y \in \mathcal{D}(T)$ is always true which yields $x \in \mathcal{D}(T)$ whenever $\begin{bmatrix} x \\ y \end{bmatrix} \in \mathcal{D}(\overline{\mathscr{A}})$. From Corollary 1.1.11 we have that

$$\sigma(\overline{\mathscr{A}}) \setminus \sigma(T^2) = \sigma(\overline{S}) \tag{3.39}$$

and, for $\lambda \in \varrho(\overline{S}) = \varrho(\overline{\mathscr{A}}) \cap \varrho(T^2) \subset \varrho(\overline{\mathscr{A}})$,

$$(\overline{\mathscr{A}} - \lambda)^{-1} = \begin{bmatrix} \left((\overline{\mathscr{A}} - \lambda)^{-1}\right)_1 & \left((\overline{\mathscr{A}} - \lambda)^{-1}\right)_2 \\ \left((\overline{\mathscr{A}} - \lambda)^{-1}\right)_3 & \left((\overline{\mathscr{A}} - \lambda)^{-1}\right)_4 \end{bmatrix}, \tag{3.40}$$

where

$$\begin{aligned}
\left((\overline{\mathscr{A}} - \lambda)^{-1}\right)_1 &= (T^2-\lambda)^{-1} + \beta\gamma T(T^2-\lambda)^{-1}\overline{S(\lambda)}^{-1}T(T^2-\lambda)^{-1}, \\
\left((\overline{\mathscr{A}} - \lambda)^{-1}\right)_2 &= -\beta T(T^2-\lambda)^{-1}\overline{S(\lambda)}^{-1}, \\
\left((\overline{\mathscr{A}} - \lambda)^{-1}\right)_3 &= -\gamma\overline{S(\lambda)}^{-1}T(T^2-\lambda)^{-1}, \\
\left((\overline{\mathscr{A}} - \lambda)^{-1}\right)_4 &= \overline{S(\lambda)}^{-1}.
\end{aligned} \tag{3.41}$$

Theorem 3.8.2. *Let $\varepsilon > 0$. Then we have:*

a) *The inclusion*

$$\sigma_\varepsilon(\overline{S}) \subset \sigma_\varepsilon(\overline{\mathscr{A}}) \setminus \sigma(T^2)$$

holds;

b) *For $L > 0$ the inclusion*

$$\big(\mathrm{B}_\varepsilon(\sigma(T^2)) \setminus (\mathrm{B}_\varepsilon(\delta) \cup \sigma(T^2))\big) \cap \mathrm{B}_L(0) \subset \big(\sigma_{\varepsilon_L}(\overline{\mathscr{A}}) \setminus \sigma(T^2)\big) \cap \mathrm{B}_L(0)$$

holds with $\varepsilon_L := \varepsilon + |\beta\gamma| \left(1 + \frac{1}{\varepsilon} L\right)$;

c) *If we further assume that $0 \in \sigma(T)$, the inclusion*

$$\mathrm{B}_\varepsilon(\delta) \setminus \sigma(T^2) \subset \sigma_\varepsilon(\overline{\mathscr{A}}) \setminus \sigma(T^2)$$

holds.

Remark 3.8.3. Recall from Theorem 2.1.3 that if T^2 is a normal operator on a Hilbert space, we have $\mathrm{B}_\varepsilon(\sigma(T^2)) = \sigma_\varepsilon(T^2)$.

Proof of Theorem 3.8.2. a) For $\lambda \in \sigma_\varepsilon(\overline{S})$ there exists a $P \in \mathcal{L}(\mathcal{X}_1)$ with $\|P\| < \varepsilon$ such that

$$0 \in \sigma\left(\delta + P - \lambda - \beta\gamma T^2 (T^2 - \lambda)^{-1}\right).$$

Here, $\delta + P - \lambda - \beta\gamma T^2(T^2 - \lambda)^{-1}$ can be interpreted as the closure of the Schur complement of the operator $\begin{bmatrix} T^2 & \beta T \\ \gamma T & \delta + P \end{bmatrix}$ which yields

$$\lambda \in \sigma\left(\overline{\begin{bmatrix} T^2 & \beta T \\ \gamma T & \delta + P \end{bmatrix}}\right) \setminus \sigma(T^2)$$

by using (3.39) and thus $\sigma_\varepsilon(\overline{S}) \subset \sigma_\varepsilon(\overline{\mathscr{A}}) \setminus \sigma(T^2)$.

b) Let now $\lambda \in \varrho(\overline{S}) \cap \big(\mathrm{B}_\varepsilon(\sigma(T^2)) \setminus \mathrm{B}_\varepsilon(\delta)\big) \cap \mathrm{B}_L(0) \subset \varrho(\overline{\mathscr{A}})$. From (3.40) we have that

$$\begin{aligned} \left\|(\overline{\mathscr{A}} - \lambda)^{-1}\right\| &\geq \left\|\left((\overline{\mathscr{A}} - \lambda)^{-1}\right)_1\right\| \\ &\geq \max\left\{|\mu| \,\middle|\, \mu \in \sigma\left(\left((\overline{\mathscr{A}} - \lambda)^{-1}\right)_1\right)\right\} \end{aligned}$$

by Theorem 1.1.4. Therefore, using (3.41) and applying the spectral mapping theorem for sectorial operators, [24, Theorem 2.7.8], yields

$$\begin{aligned}\big\|(\overline{\mathscr{A}}-\lambda)^{-1}\big\| &\geq \sup_{\mu\in\sigma(T)}\Big|(\mu^2-\lambda)^{-1}+\beta\gamma\mu^2(\mu^2-\lambda)^{-2} \\ &\qquad\qquad\cdot\big(\delta-\lambda-\beta\gamma\mu^2(\mu^2-\lambda)^{-1}\big)^{-1}\Big| \\ &= \sup_{\mu\in\sigma(T)}\left|\frac{\delta-\lambda-\beta\gamma\mu^2(\mu^2-\lambda)^{-1}+\beta\gamma\mu^2(\mu^2-\lambda)^{-1}}{(\mu^2-\lambda)\big(\delta-\lambda-\beta\gamma\mu^2(\mu^2-\lambda)^{-1}\big)}\right| \\ &= \sup_{\mu\in\sigma(T)}\left|\frac{\delta-\lambda}{(\mu^2-\lambda)(\delta-\lambda)-\beta\gamma\mu^2}\right| \\ &= \sup_{\mu\in\sigma(T)}\frac{1}{\big|\mu^2-\lambda-\frac{\beta\gamma\mu^2}{\delta-\lambda}\big|} \\ &= \sup_{\mu\in\sigma(T)}\frac{1}{\big|\mu^2-\lambda-\frac{\beta\gamma}{\delta-\lambda}(\mu^2-\lambda)-\frac{\beta\gamma}{\delta-\lambda}\lambda\big|} \\ &> \frac{1}{\varepsilon+|\beta\gamma|\left(1+\frac{1}{\varepsilon}L\right)}\end{aligned}$$

Hence, in combination with (3.39) we obtain

$$\big(\mathrm{B}_\varepsilon(\sigma(T^2))\setminus(\mathrm{B}_\varepsilon(\delta)\cup\sigma(T^2))\big)\cap\mathrm{B}_L(0)\subset\big(\sigma_{\varepsilon_L}(\overline{\mathscr{A}})\setminus\sigma(T^2)\big)\cap\mathrm{B}_L(0).$$

c) It remains to consider the case in which $\lambda\in\varrho(\overline{S})\cap\mathrm{B}_\varepsilon(\delta)\subset\varrho(\overline{\mathscr{A}})$ and $0\in\sigma(T)$. Here, we estimate

$$\begin{aligned}\big\|(\overline{\mathscr{A}}-\lambda)^{-1}\big\| &\geq \big\|\big((\overline{\mathscr{A}}-\lambda)^{-1}\big)_4\big\| \geq \max\left\{|\mu|\,\Big|\,\mu\in\sigma\left(\overline{S(\lambda)}^{-1}\right)\right\} \\ &\geq \sup_{\mu\in\sigma(T)}\frac{1}{\big|\delta-\lambda-\frac{\beta\gamma\mu^2}{\mu^2-\lambda}\big|} > \sup_{\mu\in\sigma(T)}\frac{1}{\varepsilon+\big|\frac{\beta\gamma\mu^2}{\mu^2-\lambda}\big|} \\ &= \frac{1}{\varepsilon}\end{aligned}$$

by using the same reasoning as before. This implies

$$\mathrm{B}_\varepsilon(\delta)\setminus\sigma(T^2)\subset\sigma_\varepsilon(\overline{\mathscr{A}})\setminus\sigma(T^2)$$

by utilizing (3.39) and the proof is complete. ❑

Theorem 3.8.4. *Let $\mathcal{X}_1$ be a Hilbert space, $\varepsilon>0$ and further assume that there exists an $L>0$ such that $T(T^2-\lambda)^{-1}$ is a normal operator for every $\lambda\in\big(\varrho(\overline{\mathscr{A}})\cap\mathrm{B}_L(0)\big)\setminus\sigma_\varepsilon(T^2)$. Then the inclusion*

$$\big(\sigma_\varepsilon(\overline{\mathscr{A}})\setminus\sigma(T^2)\big)\cap\mathrm{B}_L(0)\subset\big(\sigma_{\varepsilon_L}(\overline{S})\cup\big[\sigma_\varepsilon(T^2)\setminus\sigma(T^2)\big]\big)\cap\mathrm{B}_L(0)$$

holds with

$$\begin{aligned}\varepsilon_L \coloneqq \varepsilon \left(1 + |\beta| \frac{\sqrt{\varepsilon + L}}{\varepsilon}\right) \left(1 + |\gamma| \frac{\sqrt{\varepsilon + L}}{\varepsilon}\right) \\ \cdot \left(\frac{|\delta - \beta\gamma| + L}{\varepsilon} + \frac{|\beta\gamma| L}{\varepsilon^2} + 1\right).\end{aligned}$$

Proof. Let $\lambda \in \big(\varrho(\overline{\mathscr{A}}) \cap \mathrm{B}_L(0)\big) \setminus \sigma_\varepsilon(T^2) \subset \varrho(\overline{S}) \cap \varrho(T^2)$. Looking at Corollary 1.1.11, $(\overline{\mathscr{A}} - \lambda)^{-1}$ can also be written in the form

$$\begin{bmatrix} I & -\beta T(T^2-\lambda)^{-1} \\ 0 & I \end{bmatrix} \begin{bmatrix} (T^2-\lambda)^{-1} & 0 \\ 0 & \overline{S(\lambda)}^{-1} \end{bmatrix} \begin{bmatrix} I & 0 \\ -\gamma T(T^2-\lambda)^{-1} & I \end{bmatrix}.$$

This allows us to estimate

$$\begin{aligned}\|(\overline{\mathscr{A}} - \lambda)^{-1}\| \le \big(1 + |\beta| \|T(T^2-\lambda)^{-1}\|\big) \left(\|(T^2-\lambda)^{-1}\| + \|\overline{S(\lambda)}^{-1}\|\right) \\ \cdot \big(1 + |\gamma| \|T(T^2-\lambda)^{-1}\|\big)\end{aligned}$$

Here, the normality of $T(T^2-\lambda)^{-1}$ makes Theorem 1.1.6 applicable and the spectral mapping theorem for sectorial operators, [24, Theorem 2.7.8], can be used to deduce

$$\begin{aligned}\|T(T^2-\lambda)^{-1}\| &= \sup_{\mu \in \sigma(T)} \left|\frac{\mu}{\mu^2 - \lambda}\right| = \sup_{\mu \in \sigma(T)} \left(\frac{|\mu^2|}{|\mu^2-\lambda|^2}\right)^{\frac{1}{2}} \\ &\le \sup_{\mu \in \sigma(T)} \left(\frac{1}{\varepsilon} \frac{|\mu^2 - \lambda + \lambda|}{|\mu^2-\lambda|}\right)^{\frac{1}{2}} < \left(\frac{1}{\varepsilon}\left(1 + \frac{L}{\varepsilon}\right)\right)^{\frac{1}{2}} \\ &= \frac{\sqrt{\varepsilon + L}}{\varepsilon}. \end{aligned} \tag{3.42}$$

Using (3.38), we obtain

$$\begin{aligned}&\|(T^2-\lambda)^{-1}\| + \|\overline{S(\lambda)}^{-1}\| \\ &\qquad \le \|(T^2-\lambda)^{-1}\overline{S(\lambda)}\| \|\overline{S(\lambda)}^{-1}\| + \|\overline{S(\lambda)}^{-1}\| \\ &\qquad = \big(\|(T^2-\lambda)^{-1}\big(\delta - \lambda - \beta\gamma T^2(T^2-\lambda)^{-1}\big)\| + 1\big) \|\overline{S(\lambda)}^{-1}\|\end{aligned}$$

where we can estimate

$$\begin{aligned}\|(T^2-\lambda)^{-1}\big(\delta - \lambda - \beta\gamma T^2(T^2-\lambda)^{-1}\big)\| &\le \frac{1}{\varepsilon} \|\delta - \lambda - \beta\gamma + \beta\gamma\lambda(T^2-\lambda)^{-1}\| \\ &\le \frac{1}{\varepsilon}\left(|\delta - \beta\gamma| + |\lambda| + \frac{|\beta\gamma\lambda|}{\varepsilon}\right) \\ &< \frac{|\delta - \beta\gamma| + L}{\varepsilon} + \frac{|\beta\gamma| L}{\varepsilon^2}.\end{aligned}$$

Combining this with (3.42) yields

$$\begin{aligned}\big\|(\overline{\mathscr{A}}-\lambda)^{-1}\big\| < \left(1+|\beta|\frac{\sqrt{\varepsilon+L}}{\varepsilon}\right)\left(1+|\gamma|\frac{\sqrt{\varepsilon+L}}{\varepsilon}\right)\\ \cdot\left(\frac{|\delta-\beta\gamma|+L}{\varepsilon}+\frac{|\beta\gamma|L}{\varepsilon^2}+1\right)\big\|\overline{S(\lambda)}^{-1}\big\|\end{aligned}$$

If we additionally assume that $\lambda \in \sigma_\varepsilon(\overline{\mathscr{A}})$, we therefore have

$$\big\|\overline{S(\lambda)}^{-1}\big\| > \frac{1}{\varepsilon_L}. \qquad ❑$$

In combination with Theorem 3.8.2, we have shown the following chain of inclusions:

Corollary 3.8.5. *Let $\mathcal{X}_1$ be a Hilbert space, $\varepsilon > 0$ and further assume that $0 \in \sigma(T)$ and that there exists an $L > 0$ such that $T(T^2-\lambda)^{-1}$ is a normal operator for every $\lambda \in \big(\varrho(\overline{\mathscr{A}}) \cap \mathrm{B}_L(0)\big) \setminus \sigma_\varepsilon(T^2)$. Then the chain of inclusions*

$$\begin{aligned}\big(\sigma_\varepsilon(\overline{S}) \cup \big[\big(\mathrm{B}_\varepsilon(\sigma(T^2)) \cup \mathrm{B}_\varepsilon(\delta)\big) \setminus \sigma(T^2)\big]\big) \cap \mathrm{B}_L(0)\\ \subset \big(\sigma_{\varepsilon_L}(\overline{\mathscr{A}}) \setminus \sigma(T^2)\big) \cap \mathrm{B}_L(0)\\ \subset \big(\sigma_{\widehat{\varepsilon_L}}(\overline{S}) \cup \big[\sigma_\varepsilon(T^2) \setminus \sigma(T^2)\big]\big) \cap \mathrm{B}_L(0)\end{aligned}$$

holds with $\varepsilon_L := \varepsilon + |\beta\gamma|\left(1+\frac{1}{\varepsilon}L\right)$ and

$$\begin{aligned}\widehat{\varepsilon_L} := \varepsilon_L\left(1+|\beta|\frac{\sqrt{\varepsilon+L}}{\varepsilon}\right)\left(1+|\gamma|\frac{\sqrt{\varepsilon+L}}{\varepsilon}\right)\\ \cdot\left(\frac{|\delta-\beta\gamma|+L}{\varepsilon}+\frac{|\beta\gamma|L}{\varepsilon^2}+1\right).\end{aligned}$$

Example 3.8.6. Let $\mathcal{Y}$ be a Banach space, $p \in [1,\infty)$, $\mathcal{X}_1 = L^p(\mathbb{R},\mathcal{Y})$ and consider the operator $T\colon \mathcal{D}(T) \subset L^p(\mathbb{R},\mathcal{Y}) \to L^p(\mathbb{R},\mathcal{Y})$ with

$$Tx = \frac{\mathrm{d}}{\mathrm{d}\xi}x$$

for all $x \in \mathcal{D}(T) = W^{1,p}(\mathbb{R},\mathcal{Y})$. From [24, Theorem 8.4.1] we have that T is densely defined and sectorial of angle $\pi/2$ with $\sigma(T) = \mathrm{i}\mathbb{R}$. In particular, $0 \in \sigma(T)$ and via the spectral mapping theorem for sectorial operators, [24, Theorem 2.7.8], $\varrho(T^2) \neq \varnothing$. Hence, Theorem 3.8.2 can be applied to the block operator matrix

$$\begin{gathered}\mathscr{A}\colon \mathcal{D}(\mathscr{A}) \subset L^p(\mathbb{R},\mathcal{Y}) \times L^p(\mathbb{R},\mathcal{Y}) \to L^p(\mathbb{R},\mathcal{Y}) \times L^p(\mathbb{R},\mathcal{Y}),\\ \mathscr{A} = \begin{bmatrix}\frac{\mathrm{d}^2}{\mathrm{d}\xi^2} & \beta\frac{\mathrm{d}}{\mathrm{d}\xi}\\ \gamma\frac{\mathrm{d}}{\mathrm{d}\xi} & \delta\end{bmatrix},\end{gathered}$$

with $\beta,\gamma,\delta \in \mathbb{C}$ and $\mathcal{D}(\mathscr{A}) = W^{2,p}(\mathbb{R},\mathcal{Y}) \times W^{1,p}(\mathbb{R},\mathcal{Y})$.

Let us now consider the case in which $p = 2$, i.e. where $\mathcal{X}_1$ is a Hilbert space. Here, T^2 is self-adjoint and in particular normal which yields $\mathrm{B}_\varepsilon(\sigma(T^2)) = \sigma_\varepsilon(T^2)$ by Theorem 2.1.3. Moreover, T is skew-adjoint and $(T^2-\lambda)^{-1}$ is normal for every $\lambda \in \varrho(T^2)$ because $T^2 - \lambda$ is. Thus, as T and $(T^2-\lambda)^{-1}$ commute, so do T^* and $(T^2-\lambda)^{-1}$ and T and $(T^2-\lambda)^{-*}$. These facts imply

$$\begin{aligned} T(T^2-\lambda)^{-1}\big(T(T^2-\lambda)^{-1}\big)^* &\supset T(T^2-\lambda)^{-1}(T^2-\lambda)^{-*}T^* \\ &= (T^2-\lambda)^{-*}T^*T(T^2-\lambda)^{-1}|_{\mathcal{D}(T^*)} \\ &= -(T^2-\lambda)^{-*}T^2(T^2-\lambda)^{-1}|_{\mathcal{D}(T^*)} \end{aligned}$$

and therefore

$$\begin{aligned} T(T^2-\lambda)^{-1}\big(T(T^2-\lambda)^{-1}\big)^* &= \overline{T(T^2-\lambda)^{-1}\,(T(T^2-\lambda)^{-1})^*} \\ &\supset \overline{-(T^2-\lambda)^{-*}T^2(T^2-\lambda)^{-1}|_{\mathcal{D}(T^*)}} \\ &= -(T^2-\lambda)^{-*}T^2(T^2-\lambda)^{-1} \in \mathcal{L}(\mathcal{X}_1) \end{aligned}$$

because $\mathcal{D}(T^*) = \mathcal{D}(T)$ is dense in $\mathcal{X}_1$. Here, the fact that the right hand side is everywhere defined yields equality. The commutativity of $T(T^2-\lambda)^{-1}$ and $\big(T(T^2-\lambda)^{-1}\big)^*$ is now obtained by the observation

$$\big(T(T^2-\lambda)^{-1}\big)^*\,T(T^2-\lambda)^{-1} \supset -(T^2-\lambda)^{-*}T^2(T^2-\lambda)^{-1} \in \mathcal{L}(\mathcal{X}_1)$$

where the fact that the right hand side is everywhere defined yields equality again. We have thus shown that $T(T^2-\lambda)^{-1}$ is a normal operator for every $\lambda \in \varrho(T^2)$. Hence, Theorem 3.8.4 and Corollary 3.8.5 can be applied to this example with an arbitrary $L > 0$.

Chapter 4

Computing the Quadratic Numerical Range

The development of effective algorithms for the computation of the quadratic numerical range (QNR) of a matrix $\mathscr{A}$ faces several challenges and ideas approved for the numerical range $W(\mathscr{A})$ can not be adapted straightforwardly. This is in particular true because in contrast to the numerical range, the QNR does not need to be convex in general, see for instance Example 2.3.5 or Section 4.4. See also Section 2.2 for an overview on the different techniques utilized in numerical range computation which all essentially rely on convexity.

So far, the only method for computing the quadratic numerical range is based on random vector sampling, see [16] for a Matlab implementation. This method however comes with various disadvantages. We show that especially for higher dimensional matrices the computed points will very likely cluster in a small subset of each component making convergence very slow and computationally expensive. More precisely, we show that the probability of a point in the QNR generated by the random vector sampling method to be outside of a small neighborhood of the expected value decays exponentially with an increase of the dimension of the matrix when its norm stays constant.

We will present and exemplify of a novel algorithm for the computation of the quadratic numerical range of a matrix that yields much better results in less time compared to the random vector sampling method. This new approach is more deterministic and based on the idea of seeking the boundary by maximization of an objective function. Multiple examples illustrate the efficacy of this algorithm by comparing it side to side to the random vector sampling approach.

This chapter is based on the article [30] and its contents are organized as follows: Section 4.1 provides an overview of the fundamental components that constitute the new algorithm. In Section 4.2 we examine curves in the quadratic numerical range and develop useful tools regarding their differentiation. Section 4.3 contains the algorithm for the computation of the QNR alongside explanations for the chosen procedure. This algorithm is then exemplified in Section 4.4 where it is compared to the random vector sampling method. In Section 4.5, a bound on the probability for the random vector sampling method to produce a point exceeding a neighborhood of the expectation value in dependence on norm and size of the matrix is given.

Throughout this chapter we consider matrices $\mathscr{A} \in \mathbb{C}^{n\times n}$ with a block decomposition of the form

$$\mathscr{A} = \begin{bmatrix} A & B \\ C & D \end{bmatrix}$$

where $A\colon \mathcal{H}_1 \to \mathcal{H}_1$, $B\colon \mathcal{H}_2 \to \mathcal{H}_1$, $C\colon \mathcal{H}_1 \to \mathcal{H}_2$, $D\colon \mathcal{H}_2 \to \mathcal{H}_2$ and $\mathcal{H}_1 \oplus \mathcal{H}_2 := \mathbb{C}^{n_1} \oplus \mathbb{C}^{n_2} = \mathbb{C}^n$. Note that every matrix can be written in such a form once a decomposition $\mathbb{C}^n = \mathcal{H}_1 \oplus \mathcal{H}_2$ is chosen. We denote the scalar product on $\mathbb{C}^{n_1}$ and $\mathbb{C}^{n_2}$ by $\langle\cdot,\cdot\rangle$ and $\|\cdot\|$ denotes either the 2-norm of a vector or the operator norm of a matrix induced by the 2-norm.

4.1 A New Computational Technique

The algorithm that is described in this chapter enables us to compute a very precise approximation of the quadratic numerical range of a given matrix. Its effectiveness stems from its innovative ability to detect points that are either on the boundary or at least very close to the boundary of the QNR.

At the core of this detection process is a specially designed objective function, the maximization of which involves calculating its steepest ascent gradient. This step necessitates the examination of differentiable curves within the QNR and a meaningful distinction of the two points within the quadratic numerical range that are associated with the same vector pair in the unit spheres.

As we will see, the objective function features a penalty term that allows for the detection of points on highly non-convex parts of the boundary. Proposition 4.3.1 ensures that a relatively weak condition on a given boundary point is sufficient for it to be detectable by the algorithm up to a small error.

At this point, we are able to progress towards the boundary of the quadratic numerical range after specifying an arbitrary starting point in the interior and an arbitrary search direction. In order to develop an efficient algorithm that builds upon this procedure, a sensible choice of starting points from the cloud of previously computed QNR points is crucial.

The utilized selection process employs a "box approach", that not only ensures a uniform distribution of the starting points but also prioritizes those that are presumably already relatively close to the boundary. With this strategy, the computational cost is significantly reduced.

Furthermore, the iterative nature of the process that allows for the detection of boundary points coupled with a random choice of search directions leads to a filling of the interior over time and results in a cloud of points that resembles a sharply contoured and connected image of each QNR component.

4.2 Curves within the Quadratic Numerical Range

Let us start by recalling that the quadratic numerical range (QNR) is defined by

$$W^2(\mathscr{A}) = \bigcup_{x\in S_{\mathcal{H}_1}, y\in S_{\mathcal{H}_2}} \sigma\left(\begin{bmatrix}\langle Ax,x\rangle & \langle By,x\rangle\\ \langle Cx,y\rangle & \langle Dy,y\rangle\end{bmatrix}\right),$$

where $S_{\mathcal{H}_i} = \{x \in \mathcal{H}_i \mid \|x\| = 1\}$, $i = 1, 2$. In other words, the QNR consists of the solutions λ of the quadratic equations

$$\lambda^2 - (\langle Ax,x\rangle + \langle Dy,y\rangle)\lambda + \langle Ax,x\rangle\langle Dy,y\rangle - \langle By,x\rangle\langle Cx,y\rangle = 0 \tag{4.1}$$

with $(x, y) \in S_{\mathcal{H}_1} \times S_{\mathcal{H}_2}$.

Just as in the introductory Section 2.3 we will shorten our notation in the following way:

$$M_{x,y} := \begin{bmatrix}a_x & b_{y,x}\\ c_{x,y} & d_y\end{bmatrix} := \begin{bmatrix}\langle Ax,x\rangle & \langle By,x\rangle\\ \langle Cx,y\rangle & \langle Dy,y\rangle\end{bmatrix} \in \mathbb{C}^{2\times 2} \tag{4.2}$$

for $(x, y) \in S_{\mathcal{H}_1} \times S_{\mathcal{H}_2}$.

Proposition 4.2.1. *Let $(x_0, y_0) \in S_{\mathcal{H}_1} \times S_{\mathcal{H}_2}$ be such that $\sigma(M_{x_0,y_0})$ consists of two simple eigenvalues. Then there exists a neighborhood $U \subset S_{\mathcal{H}_1} \times S_{\mathcal{H}_2}$ of (x_0, y_0) such that $\bigcup_{(x,y)\in U} \sigma(M_{x,y})$ consists of two disconnected components W_1 and W_2 that can be separated by a straight line and there exists $\theta_0 \in [0, 2\pi]$ and a branch of the complex root $\sqrt{\cdot}\colon G \to \mathbb{C}$ with $G = \mathbb{C} \setminus \{r\mathrm{e}^{\mathrm{i}\theta_0} \mid r \geq 0\}$ such that*

$$W_1 = \left\{ \frac{a_x + d_y}{2} + \sqrt{\left(\frac{a_x - d_y}{2}\right)^2 + b_{y,x}c_{x,y}} \;\middle|\; (x,y) \in U \right\} \tag{4.3}$$

and

$$W_2 = \left\{ \frac{a_x + d_y}{2} - \sqrt{\left(\frac{a_x - d_y}{2}\right)^2 + b_{y,x}c_{x,y}} \;\middle|\; (x,y) \in U \right\}. \tag{4.4}$$

Proof. From Theorem 1.2.4 we have that the eigenvalues of a matrix depend continuously on its entries. Therefore, there exists a neighborhood $U \subset S_{\mathcal{H}_1} \times S_{\mathcal{H}_2}$ of (x_0, y_0) such that $\bigcup_{(x,y)\in U} \sigma(M_{x,y})$ consists of two disconnected components W_1 and W_2 that can be separated by a straight line. Without loss of generality, we assume that W_1 and W_2 are separated by the imaginary axis because considering the shifted and rotated matrix $\mathrm{e}^{\mathrm{i}\theta}(\mathscr{A} + zI)$ for some $\theta \in [0, 2\pi]$ and $z \in \mathbb{C}$ would lead to the computation of the eigenvalues of

$$\begin{bmatrix}\tilde{a}_x & \tilde{b}_{y,x}\\ \tilde{c}_{x,y} & \tilde{d}_y\end{bmatrix} = \mathrm{e}^{\mathrm{i}\theta}\begin{bmatrix}a_x + z & b_{y,x}\\ c_{x,y} & d_y + z\end{bmatrix},$$

so the radicant would be given by

$$\left(\frac{\tilde{a}_x - \tilde{d}_y}{2}\right)^2 + \tilde{b}_{y,x}\tilde{c}_{x,y} = \mathrm{e}^{2\mathrm{i}\theta}\left(\left(\frac{a_x - d_y}{2}\right)^2 + b_{y,x}c_{x,y}\right)$$

and thus $\tilde{G} = \mathrm{e}^{2\mathrm{i}\theta}G$.
So let us assume $W_1 \subset \mathbb{C}_+$ and $W_2 \subset \mathbb{C}_-$ and let $\lambda_1 \in W_1$ and $\lambda_2 \in W_2$ be eigenvalues of $M_{x,y}$ for given $(x, y) \in U$, i.e. $\Re\lambda_2 < 0 < \Re\lambda_1$ and λ_1 and λ_2 are solutions of $(a_x - \lambda)(d_y - \lambda) - b_{y,x}c_{x,y} = 0$ which is equivalent to

$$\left(\lambda - \frac{a_x + d_y}{2}\right)^2 = \left(\frac{a_x - d_y}{2}\right)^2 + b_{y,x}c_{x,y}.$$

Then there is a solution $q \in \mathbb{C}$ of $q^2 = \left(\frac{a_x - d_y}{2}\right)^2 + b_{y,x}c_{x,y}$ such that $\lambda_1 = \frac{a_x+d_y}{2} + q$ and $\lambda_2 = \frac{a_x+d_y}{2} - q$. It follows

$$0 < \Re(\lambda_1 - \lambda_2) = 2\Re q$$

and we conclude that $q \in \mathbb{C}_+$ and thus $q^2 \in \mathbb{C}\setminus\mathbb{R}_{\le 0}$. So by defining $G := \mathbb{C}\setminus\mathbb{R}_{\le 0} = \mathbb{C}\setminus\{r\mathrm{e}^{\mathrm{i}\pi} \mid r \ge 0\}$ and $\sqrt{\cdot}\colon G \to \mathbb{C}$ as the principal branch of the complex root with $\Re\sqrt{z} > 0$ for all $z \in G$ we obtain

$$\lambda_1 = \frac{a_x + d_y}{2} + \sqrt{\left(\frac{a_x - d_y}{2}\right)^2 + b_{y,x}c_{x,y}}$$

and

$$\lambda_2 = \frac{a_x + d_y}{2} - \sqrt{\left(\frac{a_x - d_y}{2}\right)^2 + b_{y,x}c_{x,y}}. \qquad ❑$$

Remark 4.2.2. Note, that the assumption in Proposition 4.2.1 on $\sigma(M_{x_0,y_0})$ to consist of two simple eigenvalues is fulfilled for every $(x_0, y_0) \in S_{\mathcal{H}_1} \times S_{\mathcal{H}_2}$ if $W^2(\mathscr{A})$ consists of two disconnected components. Furthermore, we have $U = S_{\mathcal{H}_1} \times S_{\mathcal{H}_2}$ if the two components of $W^2(\mathscr{A})$ can be separated by a straight line. In this case, the formulas in (4.3) and (4.4) can be used to match each of the two eigenvalues of a matrix $M_{x,y}$ to a specific component.

In the following we will consider curves in the QNR, i.e. continuous mappings $\lambda_{\varphi,\psi}$ from an interval I into $W^2(\mathscr{A})$ which are generated from continuous curves $\varphi\colon I \to S_{\mathcal{H}_1}$ and $\psi\colon I \to S_{\mathcal{H}_2}$ such that $\lambda_{\varphi,\psi}(t)$ solves (4.1) with $\varphi(t)$ in place of x and $\psi(t)$ in place of y for all $t \in I$. We are interested in the derivative of such a curve in the QNR and in order to shorten the notation in the upcoming formulas we will henceforth and similarly to (4.2) use the abbreviations

$$M_{\varphi,\psi} := \begin{bmatrix} a_\varphi & b_{\psi,\varphi} \\ c_{\varphi,\psi} & d_\psi \end{bmatrix} := \begin{bmatrix} \langle A\varphi, \varphi\rangle & \langle B\psi, \varphi\rangle \\ \langle C\varphi, \psi\rangle & \langle D\psi, \psi\rangle \end{bmatrix} : I \to \mathbb{C}^{2\times 2}$$

for curves $(\varphi, \psi)\colon I \to S_{\mathcal{H}_1} \times S_{\mathcal{H}_2}$.

Theorem 4.2.3. *Let $(x_0, y_0) \in S_{\mathcal{H}_1} \times S_{\mathcal{H}_2}$ be such that $\sigma(M_{x_0,y_0})$ consists of two simple eigenvalues. Let $t_1 > 0$ and $(\varphi, \psi)\colon [0, t_1] \to S_{\mathcal{H}_1} \times S_{\mathcal{H}_2}$, $t \mapsto (\varphi(t), \psi(t))$, with $(\varphi(0), \psi(0)) = (x_0, y_0)$ be a differentiable curve in $S_{\mathcal{H}_1} \times S_{\mathcal{H}_2}$. Then there exists a $0 < t_0 \leq t_1$ such that $\sigma(M_{\varphi,\psi})\colon [0, t_0] \to \mathbb{C}^2$ consists of two differentiable curves. Denote by $\lambda_{\varphi,\psi}\colon [0, t_0] \to \mathbb{C}$ one of these two curves. Then*

$$\frac{\mathrm{d}}{\mathrm{d}t}\lambda_{\varphi,\psi} = \left\langle S(\varphi, \psi, \lambda_{\varphi,\psi}) \begin{bmatrix} \varphi \\ \psi \end{bmatrix}, \begin{bmatrix} \dot{\varphi} \\ \dot{\psi} \end{bmatrix} \right\rangle + \left\langle S(\varphi, \psi, \lambda_{\varphi,\psi}) \begin{bmatrix} \dot{\varphi} \\ \dot{\psi} \end{bmatrix}, \begin{bmatrix} \varphi \\ \psi \end{bmatrix} \right\rangle$$

with

$$S(\varphi, \psi, \lambda_{\varphi,\psi}) = \frac{1}{2\lambda_{\varphi,\psi} - a_\varphi - d_\psi} \begin{bmatrix} (\lambda_{\varphi,\psi} - d_\psi)A & c_{\varphi,\psi}B \\ b_{\psi,\varphi}C & (\lambda_{\varphi,\psi} - a_\varphi)D \end{bmatrix}.$$

Proof. From Proposition 4.2.1 we have that there exists a neighborhood $U \subset S_{\mathcal{H}_1} \times S_{\mathcal{H}_2}$ of (x_0, y_0) such that $\bigcup_{(x,y)\in U} \sigma(M_{x,y})$ consists of two disconnected components W_1 and W_2 that can be described in a differentiable dependence on the $(x, y) \in U$ via the formulas in (4.3) and (4.4). We therefore obtain that $\sigma(M_{\varphi,\psi})\colon [0, t_0] \to \mathbb{C}^2$ consists of two differentiable curves by choosing $t_0 > 0$ such that $(\varphi(t), \psi(t)) \in U$ for all $t \in [0, t_0]$.
An eigenvalue $\lambda_{\varphi,\psi}(t)$ of $M_{\varphi,\psi}(t)$, $t \in [0, t_0]$, satisfies

$$\big(\lambda_{\varphi,\psi}(t) - a_\varphi(t)\big)\big(\lambda_{\varphi,\psi}(t) - d_\psi(t)\big) - b_{\psi,\varphi}(t)c_{\varphi,\psi}(t) = 0,$$

so that upon differentiation we get

$$\begin{aligned} \left(\frac{\mathrm{d}}{\mathrm{d}t}\lambda_{\varphi,\psi} - \frac{\mathrm{d}}{\mathrm{d}t}a_\varphi\right)(\lambda_{\varphi,\psi} - d_\psi) + (\lambda_{\varphi,\psi} - a_\varphi)\left(\frac{\mathrm{d}}{\mathrm{d}t}\lambda_{\varphi,\psi} - \frac{\mathrm{d}}{\mathrm{d}t}d_\psi\right) \\ -\frac{\mathrm{d}}{\mathrm{d}t}b_{\psi,\varphi}c_{\varphi,\psi} - b_{\psi,\varphi}\frac{\mathrm{d}}{\mathrm{d}t}c_{\varphi,\psi} = 0. \end{aligned} \tag{4.5}$$

Herein

$$\begin{aligned} \frac{\mathrm{d}}{\mathrm{d}t}a_\varphi &= \langle A\dot{\varphi}, \varphi\rangle + \langle A\varphi, \dot{\varphi}\rangle \\ \frac{\mathrm{d}}{\mathrm{d}t}b_{\psi,\varphi} &= \langle B\dot{\psi}, \varphi\rangle + \langle B\psi, \dot{\varphi}\rangle \\ \frac{\mathrm{d}}{\mathrm{d}t}c_{\varphi,\psi} &= \langle C\dot{\varphi}, \psi\rangle + \langle C\varphi, \dot{\psi}\rangle \\ \frac{\mathrm{d}}{\mathrm{d}t}d_\psi &= \langle D\dot{\psi}, \psi\rangle + \langle D\psi, \dot{\psi}\rangle, \end{aligned}$$

which transforms (4.5) into

$$\begin{aligned} &(2\lambda_{\varphi,\psi} - a_\varphi - d_\psi)\frac{\mathrm{d}}{\mathrm{d}t}\lambda_{\varphi,\psi} \\ &\quad= \left\langle \begin{bmatrix} (\lambda_{\varphi,\psi} - d_\psi)A & c_{\varphi,\psi}B \\ b_{\psi,\varphi}C & (\lambda_{\varphi,\psi} - a_\varphi)D \end{bmatrix} \begin{bmatrix} \varphi \\ \psi \end{bmatrix}, \begin{bmatrix} \dot{\varphi} \\ \dot{\psi} \end{bmatrix} \right\rangle \\ &\qquad+ \left\langle \begin{bmatrix} (\lambda_{\varphi,\psi} - d_\psi)A & c_{\varphi,\psi}B \\ b_{\psi,\varphi}C & (\lambda_{\varphi,\psi} - a_\varphi)D \end{bmatrix} \begin{bmatrix} \dot{\varphi} \\ \dot{\psi} \end{bmatrix}, \begin{bmatrix} \varphi \\ \psi \end{bmatrix} \right\rangle. \end{aligned} \tag{4.6}$$

Now the fact that $\bigcup_{t\in[0,t_0]}\sigma(M_{\varphi,\psi}(t))$ consists of two disconnected components implies that $\lambda_{\varphi,\psi}(t) \neq \frac{a_\varphi+d_\psi}{2}(t)$ for all $t \in [0,t_0]$ because the sum of the eigenvalues of $M_{\varphi,\psi}(t)$ is equal to the sum of its diagonal entries. This allows us to divide (4.6) by $2\lambda_{\varphi,\psi} - a_\varphi - d_\psi$, yielding the desired formula for the derivative of $\lambda_{\varphi,\psi}$. ❑

4.3 An Algorithm for Computing the QNR

Our goal is to develop an algorithm for the computation of the quadratic numerical range that does not only rely on random vector sampling. This means, that we want to make a choice on the utilized vectors $(x,y) \in S_{\mathcal{H}_1} \times S_{\mathcal{H}_2}$ such that the image resulting from the point cloud of eigenvalues of the matrices $M_{x,y}$ is a very good approximation of the image of the actual QNR even for a small number of vectors. We will therefore place particular emphasis on those $(x,y) \in S_{\mathcal{H}_1} \times S_{\mathcal{H}_2}$ that correspond to boundary points of $W^2(\mathscr{A})$.

4.3.1 Seeking the Boundary

Starting at a given point $\lambda_0 \in W^2(\mathscr{A})$ with corresponding $(x_0,y_0) \in S_{\mathcal{H}_1} \times S_{\mathcal{H}_2}$ we wish to gradually compute a sequence $(x_n,y_n)_{n\in\mathbb{N}} \subset S_{\mathcal{H}_1} \times S_{\mathcal{H}_2}$ such that the associated $(\lambda_n)_{n\in\mathbb{N}}$ in the quadratic numerical range converge towards the boundary. In order to do so we first have to declare a direction in which we want to approach the boundary, so in the following, we will therefore start by focusing on moving parallel to the positive real axis since every other direction can be easily reduced to this case by a rotation of the matrix $\mathscr{A}$.

If we would leave it at aiming for $\Re\lambda_{n+1} \geq \Re\lambda_n$ for every $n \in \mathbb{N}$ however, we could face the problem of missing out on points on concave parts of the boundary, cf. Figure 4.1, where starting from λ_0 an algorithm that only focuses on maximization of the real part would eventually either move toward b or toward c but has no reason to stop at a. We overcome this problem by seeking a sequence $(\lambda_n)_{n\in\mathbb{N}}$ that

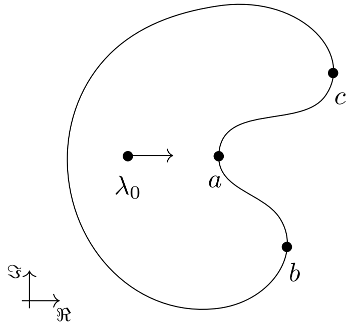


Figure 4.1: The boundary of the QNR might have concave parts

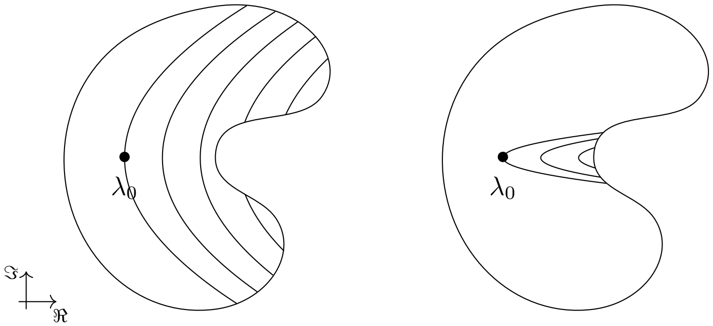


Figure 4.2: Level sets of f_{α,λ_0} for p small (left) and p large (right)

satisfies

$$\Re\lambda_{n+1} - p\big(\Im(\lambda_{n+1} - \lambda_0)\big)^2 \geq \Re\lambda_n - p\big(\Im(\lambda_n - \lambda_0)\big)^2$$

with a given penalty constant $p > 0$ for all $n \in \mathbb{N}$. More precisely, we will consider the objective function $f_{\alpha,\lambda_0} \colon S_{\mathcal{H}_1} \times S_{\mathcal{H}_2} \to \mathbb{R}$, given by

$$f_{\alpha,\lambda_0}(x,y) = \Re\lambda_{x,y}^{(\alpha)} - p\big(\Im(\lambda_{x,y}^{(\alpha)} - \lambda_0)\big)^2 \tag{4.7}$$

for some $\alpha \in [0, 2\pi[$, and aim to construct a sequence $(x_n, y_n)_{n\in\mathbb{N}} \subset S_{\mathcal{H}_1} \times S_{\mathcal{H}_2}$ such that $f_{\alpha,\lambda_0}(x_n, y_n)$ increases with n. Here and from now on, $\lambda_{x,y}^{(\alpha)}$ specifically denotes the one of the two eigenvalues $\lambda_{x,y}^{(\alpha)}$ and $\tilde{\lambda}_{x,y}^{(\alpha)}$ of $M_{x,y}$ such that

$$\begin{aligned} & \Re(\mathrm{e}^{\mathrm{i}\alpha}\lambda_{x,y}^{(\alpha)}) > \Re(\mathrm{e}^{\mathrm{i}\alpha}\tilde{\lambda}_{x,y}^{(\alpha)}), && \text{if} \quad \Re(\mathrm{e}^{\mathrm{i}\alpha}\lambda_{x,y}^{(\alpha)}) \neq \Re(\mathrm{e}^{\mathrm{i}\alpha}\tilde{\lambda}_{x,y}^{(\alpha)}), \\ \text{or} \quad & \Im(\mathrm{e}^{\mathrm{i}\alpha}\lambda_{x,y}^{(\alpha)}) \geq \Im(\mathrm{e}^{\mathrm{i}\alpha}\tilde{\lambda}_{x,y}^{(\alpha)}), && \text{if} \quad \Re(\mathrm{e}^{\mathrm{i}\alpha}\lambda_{x,y}^{(\alpha)}) = \Re(\mathrm{e}^{\mathrm{i}\alpha}\tilde{\lambda}_{x,y}^{(\alpha)}). \end{aligned}$$

Note, that if $W^2(\mathscr{A})$ consists of two disconnected components that can be separated by a straight line, α can be chosen such that each component is either the set of all $\lambda_{x,y}^{(\alpha)}$ or the set of all $\tilde{\lambda}_{x,y}^{(\alpha)} = \lambda_{x,y}^{(\alpha+\pi)}$ with $(x,y) \in S_{\mathcal{H}_1} \times S_{\mathcal{H}_2}$.

Figure 4.2 illustrates the effect of the penalty constant p on the level sets of f_{α,λ_0} for different choices of p, showing that a larger p leads to a significant narrowing of the search direction. The picture also indicates, that in practice p has to be chosen in dependence on the size and shape of the QNR.

This dependence will be specified in the next result, which can be interpreted as follows: If a boundary point λ_∂ with the same imaginary part as λ_0 satisfies the additional condition, that there exists an $r > 0$ such that for every $0 < \varepsilon < r$ there exists a $\delta > 0$ such that

$$\Big\{z \in \mathbb{C} \,\Big|\, \frac{\varepsilon}{2} < \Re(z - \lambda_\partial) < \varepsilon, |\Im(z - \lambda_\partial)| < \delta\Big\} \cap W^2(\mathscr{A}) = \varnothing \tag{4.8}$$

holds, our strategy of creating a sequence in $S_{\mathcal{H}_1} \times S_{\mathcal{H}_2}$ for which the value of the objective function increases can in fact result in the obtainment of λ_∂ up to a small error if p is chosen large enough. Figure 4.3 illustrates condition (4.8).

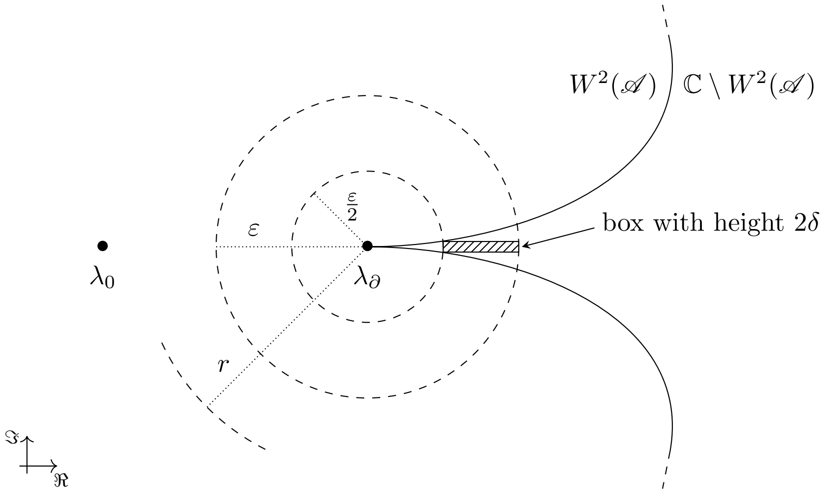


Figure 4.3: λ_∂ satisfies Condition (4.8)

Proposition 4.3.1. *Let $\lambda_0 \in W^2(\mathscr{A})$ and $\lambda_\partial \in \partial W^2(\mathscr{A})$ with $\Im\lambda_\partial = \Im\lambda_0$ and suppose that λ_∂ satisfies Condition (4.8) for some $r > 0$. Then for every $\varepsilon > 0$ there exist $p > 0$ and $(x_\mathrm{m}, y_\mathrm{m}) \in S_{\mathcal{H}_1} \times S_{\mathcal{H}_2}$ such that for all but up to one $\alpha \in [0, \pi[$ at least one of the objective functions f_{α,λ_0} or $f_{\alpha+\pi,\lambda_0}$ given by (4.7) has a local maximum in $(x_\mathrm{m}, y_\mathrm{m})$ and $\lambda^{(\alpha)}_{x_\mathrm{m},y_\mathrm{m}} \in \mathrm{B}_\varepsilon(\lambda_\partial) \cap \partial W^2(\mathscr{A})$ or $\lambda^{(\alpha+\pi)}_{x_\mathrm{m},y_\mathrm{m}} \in \mathrm{B}_\varepsilon(\lambda_\partial) \cap \partial W^2(\mathscr{A})$ respectively.*

Proof. Let $\varepsilon > 0$. Without loss of generality, we assume that $\lambda_\partial = 0$ and therefore also $\Im\lambda_0 = 0$ by applying a shift to $\mathscr{A}$. Moreover, we will assume that $\varepsilon < r$, where $r > 0$ is the constant for which (4.8) holds. Hence, there exists a $\delta > 0$ such that

$$\left\{ z \in \mathbb{C} \,\middle|\, \frac{\varepsilon}{2} < \Re z < \varepsilon, |\Im z| < \delta \right\} \cap W^2(\mathscr{A}) = \varnothing \tag{4.9}$$

and for which we assume that $\delta < \frac{\sqrt{3}}{2}\varepsilon$.

Let us define the set

$$K := \left\{ z \in \mathbb{C} \,\middle|\, \left(\Re z > \frac{\varepsilon}{2} \wedge |\Im z| < \delta \right) \vee \Re z > \varepsilon \right\}$$

and consider the function

$$F \colon \mathbb{C} \to \mathbb{R}, \quad F(z) := \Re z - p(\Im z)^2.$$

By choosing $p > \frac{\varepsilon}{\delta^2} > \frac{4}{3\varepsilon}$, we have $(\Im z)^2 < \delta^2 < \frac{3}{4}\varepsilon^2$ if $F(z) \geq 0$ and $\Re z \leq \varepsilon$ and obtain

$$\left\{ z \in \mathbb{C} \setminus K \,\middle|\, F(z) \geq 0 \right\} \subset \mathrm{B}_\varepsilon(0).$$

Hence, the restriction of the continuous function F to $\mathbb{C} \setminus K$ has a local maximum in $\mathrm{B}_\varepsilon(0)$.

This can be used in the context of the quadratic numerical range because due to (4.9) we have

$$W^2(\mathscr{A}) \cap \mathrm{B}_\varepsilon(0) \subset \mathbb{C} \setminus K$$

and we also know that $\{z \in \mathbb{C} \setminus K \,|\, F(z) \geq 0\} \cap W^2(\mathscr{A})$ is non-empty because of $0 = \lambda_\partial \in W^2(\mathscr{A})$ and $F(0) = 0$. Therefore, the restriction of F to the closed set $W^2(\mathscr{A})$ has a local maximum at some $\lambda \in W^2(\mathscr{A}) \cap \mathrm{B}_\varepsilon(0)$ and there exist $(x_\mathrm{m}, y_\mathrm{m}) \in S_{\mathcal{H}_1} \times S_{\mathcal{H}_2}$ such that λ is an eigenvalue of $M_{x_\mathrm{m}, y_\mathrm{m}}$.

Furthermore, $\lambda \in \partial W^2(\mathscr{A})$ holds because if we assume otherwise, there exists a $\gamma_\max > 0$ such that $\lambda + \gamma \in W^2(\mathscr{A})$ and $F(\lambda + \gamma) = F(\lambda) + \gamma > F(\lambda)$ for every $\gamma \in]0, \gamma_\max[$ which is a contradiction to λ being a local maximum.

From Theorem 1.2.4 we know that the eigenvalues of a matrix depend continuously on its entries, so if λ is the only eigenvalue of $M_{x_\mathrm{m}, y_\mathrm{m}}$, there exists a neighborhood U of $(x_\mathrm{m}, y_\mathrm{m})$ such that $\lambda_{x,y}^{(\alpha)} \in \mathrm{B}_\varepsilon(0)$ and $\lambda_{x,y}^{(\alpha+\pi)} \in \mathrm{B}_\varepsilon(0)$ for all $(x, y) \in U$ and all $\alpha \in [0, \pi[$. Hence, both f_{α,λ_0} and $f_{\alpha+\pi,\lambda_0}$ have a local maximum in $(x_\mathrm{m}, y_\mathrm{m})$ and

$$\lambda_{x_\mathrm{m}, y_\mathrm{m}}^{(\alpha)} = \lambda_{x_\mathrm{m}, y_\mathrm{m}}^{(\alpha+\pi)} = \lambda \in \mathrm{B}_\varepsilon(0) \cap \partial W^2(\mathscr{A})$$

for all $\alpha \in [0, \pi[$.

In the other case, if λ is one of two distinct eigenvalues of $M_{x_\mathrm{m}, y_\mathrm{m}}$, we choose $\alpha \in [0, \pi[$ such that $\Re(\mathrm{e}^{\mathrm{i}\alpha} \lambda_{x_\mathrm{m}, y_\mathrm{m}}^{(\alpha)}) \neq \Re(\mathrm{e}^{\mathrm{i}\alpha} \lambda_{x_\mathrm{m}, y_\mathrm{m}}^{(\alpha+\pi)})$ and again by continuity we will find a neighborhood U of $(x_\mathrm{m}, y_\mathrm{m})$ such that $\Re\lambda_1 \neq \Re\lambda_2$ for all $\lambda_1 \in \left\{\mathrm{e}^{\mathrm{i}\alpha} \lambda_{x,y}^{(\alpha)} \,\middle|\, (x, y) \in U\right\}$ and $\lambda_2 \in \left\{\mathrm{e}^{\mathrm{i}\alpha} \lambda_{x,y}^{(\alpha+\pi)} \,\middle|\, (x, y) \in U\right\}$ and either $\lambda_{x,y}^{(\alpha)} \in \mathrm{B}_\varepsilon(0)$ for all $(x, y) \in U$ if $\lambda = \lambda_{x_\mathrm{m}, y_\mathrm{m}}^{(\alpha)}$ or $\lambda_{x,y}^{(\alpha+\pi)} \in \mathrm{B}_\varepsilon(0)$ for all $(x, y) \in U$ if $\lambda = \lambda_{x_\mathrm{m}, y_\mathrm{m}}^{(\alpha+\pi)}$. Hence, either f_{α,λ_0} or $f_{\alpha+\pi,\lambda_0}$ has a local maximum in $(x_\mathrm{m}, y_\mathrm{m})$ and $\lambda_{x_\mathrm{m}, y_\mathrm{m}}^{(\alpha)} = \lambda \in \mathrm{B}_\varepsilon(0) \cap \partial W^2(\mathscr{A})$ or $\lambda_{x_\mathrm{m}, y_\mathrm{m}}^{(\alpha+\pi)} = \lambda \in \mathrm{B}_\varepsilon(0) \cap \partial W^2(\mathscr{A})$ respectively. ❑

Let us now fix an $\alpha \in [0, 2\pi[$. As explained above, we are looking for a sequence $(x_n, y_n)_{n \in \mathbb{N}} \subset S_{\mathcal{H}_1} \times S_{\mathcal{H}_2}$ such that $f_{\alpha,\lambda_0}(x_n, y_n)$ increases with n. Let us say we arrived at (x_n, y_n) so far, so our goal is to find (x_{n+1}, y_{n+1}) such that $f_{\alpha,\lambda_0}(x_{n+1}, y_{n+1}) \geq f_{\alpha,\lambda_0}(x_n, y_n)$.

As a first step, we will therefore identify the steepest ascent gradient of f_{α,λ_0} in (x_n, y_n). Considering a differentiable curve

$$(\varphi, \psi) \colon [0, t_1] \to S_{\mathcal{H}_1} \times S_{\mathcal{H}_2}, \quad t \mapsto (\varphi(t), \psi(t)),$$

with $(\varphi(0), \psi(0)) = (x_n, y_n)$ and assuming that $\sigma(M_{x_n, y_n})$ consists of two simple eigenvalues λ_n and $\tilde{\lambda}_n$ with $\Re(\mathrm{e}^{\mathrm{i}\alpha} \lambda_n) > \Re(\mathrm{e}^{\mathrm{i}\alpha} \tilde{\lambda}_n)$, we know by Theorem 4.2.3 that there exists a $0 < t_0 \leq t_1$ and a differentiable curve $\lambda_{\varphi,\psi} \colon [0, t_0] \to W^2(\mathscr{A})$ such that

$$f_{\alpha,\lambda_0}\big(\varphi(t), \psi(t)\big) = \Re\big(\lambda_{\varphi,\psi}(t)\big) - p\big(\Im\big(\lambda_{\varphi,\psi}(t) - \lambda_0\big)\big)^2$$

for all $t \in [0, t_0]$. If we take a look at $\frac{\mathrm{d}}{\mathrm{d}t} f_{\alpha,\lambda_0}\big(\varphi(\cdot), \psi(\cdot)\big)$ at the point $t = 0$, we see that again by Theorem 4.2.3

$$\begin{aligned}
&\frac{\mathrm{d}}{\mathrm{d}t} f_{\alpha,\lambda_0}(\varphi(0),\psi(0)) \\
&\quad= \Re\left(\frac{\mathrm{d}}{\mathrm{d}t}\lambda_{\varphi,\psi}(0)\right) - 2p\Im(\lambda_n-\lambda_0)\Im\left(\frac{\mathrm{d}}{\mathrm{d}t}\lambda_{\varphi,\psi}(0)\right) \\
&\quad= \Re\left(\left\langle S(x_n,y_n,\lambda_n)\begin{bmatrix}x_n\\y_n\end{bmatrix}, \begin{bmatrix}\dot\varphi(0)\\\dot\psi(0)\end{bmatrix}\right\rangle + \left\langle S(x_n,y_n,\lambda_n)\begin{bmatrix}\dot\varphi(0)\\\dot\psi(0)\end{bmatrix}, \begin{bmatrix}x_n\\y_n\end{bmatrix}\right\rangle\right) \\
&\qquad - 2p\Im(\lambda_n-\lambda_0)\Im\left(\left\langle S(x_n,y_n,\lambda_n)\begin{bmatrix}x_n\\y_n\end{bmatrix}, \begin{bmatrix}\dot\varphi(0)\\\dot\psi(0)\end{bmatrix}\right\rangle\right. \\
&\qquad\qquad\qquad\qquad\qquad\qquad \left. + \left\langle S(x_n,y_n,\lambda_n)\begin{bmatrix}\dot\varphi(0)\\\dot\psi(0)\end{bmatrix}, \begin{bmatrix}x_n\\y_n\end{bmatrix}\right\rangle\right) \\
&\quad= \Re\left\langle T_+(x_n,y_n,\lambda_n)\begin{bmatrix}x_n\\y_n\end{bmatrix}, \begin{bmatrix}\dot\varphi(0)\\\dot\psi(0)\end{bmatrix}\right\rangle \\
&\qquad - 2p\Im(\lambda_n-\lambda_0)\Im\left\langle T_-(x_n,y_n,\lambda_n)\begin{bmatrix}x_n\\y_n\end{bmatrix}, \begin{bmatrix}\dot\varphi(0)\\\dot\psi(0)\end{bmatrix}\right\rangle \\
&\quad= \Re\left\langle T_+(x_n,y_n,\lambda_n) + 2p\Im(\lambda_n-\lambda_0)\mathrm{i}T_-(x_n,y_n,\lambda_n)\begin{bmatrix}x_n\\y_n\end{bmatrix}, \begin{bmatrix}\dot\varphi(0)\\\dot\psi(0)\end{bmatrix}\right\rangle \\
&\quad= \Re\left\langle T(x_n,y_n,\lambda_n,\lambda_0)\begin{bmatrix}x_n\\y_n\end{bmatrix}, \begin{bmatrix}\dot\varphi(0)\\\dot\psi(0)\end{bmatrix}\right\rangle
\end{aligned}$$

where

$$\begin{aligned}
T_+(x_n,y_n,\lambda_n) &:= S(x_n,y_n,\lambda_n) + S^*(x_n,y_n,\lambda_n), \\
T_-(x_n,y_n,\lambda_n) &:= S(x_n,y_n,\lambda_n) - S^*(x_n,y_n,\lambda_n)
\end{aligned}$$

and

$$T(x_n,y_n,\lambda_n,\lambda_0) := T_+(x_n,y_n,\lambda_n) + 2p\Im(\lambda_n-\lambda_0)\mathrm{i}T_-(x_n,y_n,\lambda_n).$$

Let us denote the tangential space of $S_{\mathcal{H}_1}$ in x_n with regard to the real part of the inner product of $\mathcal{H}_1$ by $T_{x_n}S_{\mathcal{H}_1} := \{u \in \mathcal{H}_1 \mid \Re\langle x_n, u\rangle = 0\}$ and analogously $T_{y_n}S_{\mathcal{H}_2} := \{v \in \mathcal{H}_2 \mid \Re\langle y_n, v\rangle = 0\}$. Then, for $(u,v) \in T_{x_n}S_{\mathcal{H}_1} \times T_{y_n}S_{\mathcal{H}_2}$ with $\|u\| = \|v\| = 1$, we will consider the curves $\varphi\colon [0,2\pi] \to S_{\mathcal{H}_1}$ and $\psi\colon [0,2\pi] \to S_{\mathcal{H}_2}$ defined by

$$\varphi(t) = \cos(t)x_n + \sin(t)u \qquad \text{and} \qquad \psi(t) = \cos(t)y_n + \sin(t)v, \tag{4.10}$$

which satisfy $\|\varphi(t)\|^2 = \cos^2(t) + 2\Re\langle\cos(t)x_n, \sin(t)u\rangle + \sin^2(t) = 1$, $\varphi(0) = x_n$ and $\dot\varphi(0) = u$ as well as $\|\psi(t)\|^2 = 1$, $\psi(0) = y_n$ and $\dot\psi(0) = v$.

Therefore, our problem can be simplified to finding a vector $(u,v) \in T_{x_n}S_{\mathcal{H}_1} \times T_{y_n}S_{\mathcal{H}_2}$ with $\|u\| = \|v\| = 1$ for which the term

$$\Re\left\langle T(x_n,y_n,\lambda_n,\lambda_0)\begin{bmatrix}x_n\\y_n\end{bmatrix}, \begin{bmatrix}u\\v\end{bmatrix}\right\rangle$$

is maximized. This vector is given by the normalized orthogonal projection of

$$\begin{bmatrix} w \\ z \end{bmatrix} := T(x_n, y_n, \lambda_n, \lambda_0) \begin{bmatrix} x_n \\ y_n \end{bmatrix}$$

onto $T_{x_n} S_{\mathcal{H}_1} \times T_{y_n} S_{\mathcal{H}_2}$, i.e.

$$\begin{bmatrix} \tilde{u} \\ \tilde{v} \end{bmatrix} = \begin{bmatrix} w - \Re\langle w, x_n \rangle x_n \\ z - \Re\langle z, y_n \rangle y_n \end{bmatrix}, \qquad \begin{bmatrix} u \\ v \end{bmatrix} = \begin{bmatrix} \tilde{u}/\|\tilde{u}\| \\ \tilde{v}/\|\tilde{v}\| \end{bmatrix}.$$

With u and v at hand, we will now, in a second step, search for $(s,t) \in [0, 2\pi] \times [0, 2\pi]$ such that $f_{\alpha,\lambda_0}\big(\varphi(s), \psi(t)\big)$ is maximal with φ and ψ as in (4.10)

Algorithm 4.3.1: Seeking the Boundary

```
Input: $A$, $B$, $C$, $D$, $x_0$, $y_0$, $\lambda_0$, $p$, $\alpha$, and $i_{\max}$
function $f(s,t,A,B,C,D,x,y,u,v,p,\alpha,\lambda_0)$
    $x = \cos(s)x + \sin(s)u$
    $y = \cos(t)y + \sin(t)v$
    return $\Re\lambda^{(\alpha)}_{x,y} - p\Im\big(\lambda^{(\alpha)}_{x,y} - \lambda_0\big)^2$
function find_boundary$(A,B,C,D,x,y,\lambda_0,p,\alpha)$
    if $2\lambda^{(\alpha)}_{x,y} \neq a_x + d_y$ then
        $S = \frac{1}{2\lambda^{(\alpha)}_{x,y} - a_x - d_y} \begin{bmatrix} (\lambda^{(\alpha)}_{x,y} - d_y)A & c_{x,y}B \\ b_{y,x}C & (\lambda^{(\alpha)}_{x,y} - a_x)D \end{bmatrix}$
        $T_+ = S + S^*$
        $T_- = S - S^*$
        $T = T_+ + 2p\Im\big(\lambda^{(\alpha)}_{x,y} - \lambda_0\big)\mathrm{i}T_-$
        $\left[\begin{smallmatrix} w \\ z \end{smallmatrix}\right] = T\left[\begin{smallmatrix} x \\ y \end{smallmatrix}\right]$
        if $w \neq 0$ and $z \neq 0$ then
            $u = \frac{w - \Re\langle w,x\rangle x}{\|w - \Re\langle w,x\rangle x\|}$
            $v = \frac{z - \Re\langle z,y\rangle y}{\|z - \Re\langle z,y\rangle y\|}$
            Determine $(s,t) \in [0,2\pi] \times [0,2\pi]$ such that
              $f(s,t,A,B,C,D,x,y,u,v,p,\alpha,\lambda_0)$ is maximal
            $x = \cos(s)x + \sin(s)u$
            $y = \cos(t)y + \sin(t)v$
    return $(x,y)$
$(x[0],y[0]) = $ find_boundary$(A,B,C,D,x_0,y_0,\lambda_0,p,\alpha)$
for $i = 1,\dots,i_{\max} - 1$
    $(x[i],y[i]) = $ find_boundary$(A,B,C,D,x[i-1],y[i-1],\lambda_0,p,\alpha)$
    if $(x[i],y[i]) == (x[i-1],y[i-1])$ then
        return $\{(x[0],y[0]),\dots,(x[i-1],y[i-1])\}$
return $(x,y)$
```

and we have effectively reduced our problem to a two-dimensional optimization. This yields a new pair of vectors $(x_{n+1}, y_{n+1}) := (\varphi(s), \psi(t)) \in S_{\mathcal{H}_1} \times S_{\mathcal{H}_2}$ with $f_{\alpha,\lambda_0}(x_{n+1}, y_{n+1}) \geq f_{\alpha,\lambda_0}(x_n, y_n)$.

Algorithm 4.3.1 summarizes in pseudocode how we proceed towards the boundary in direction of the positive real axis. It takes the matrix $\mathscr{A}$ in form of its decomposition parts A, B, C and D, a starting point $\lambda_0 \in W^2(\mathscr{A})$ with corresponding $(x_0, y_0) \in S_{\mathcal{H}_1} \times S_{\mathcal{H}_2}$, a penalty constant p, an angle α and a number of iterations $i_{\max}$ as its input and returns arrays $x \subset S_{\mathcal{H}_1}$ and $y \subset S_{\mathcal{H}_2}$. Note, that the algorithm does not only return the vectors associated to the point closest to the boundary after $i_{\max}$ iterations but an array of vectors that can be used to compute some points along the way. Later on, these points can be plotted as well in order to fill out the interior of the quadratic numerical range.

4.3.2 Box Approach

In order to formulate an algorithm which computes the quadratic numerical range of a given matrix $\mathscr{A}$ with increasing quality we proceed as follows:
Initially, we fix an $\alpha \in [0, \pi[$ for the objective function (4.7) and compute a few points within $W^2(\mathscr{A})$ by using the random vector sampling method, i.e. we randomly generate some $(x, y) \in S_{\mathcal{H}_1} \times S_{\mathcal{H}_2}$ and insert the eigenvalues $\lambda_{x,y}^{(\alpha)}$ of $M_{x,y}$ in an array W and the other eigenvalues $\lambda_{x,y}^{(\alpha+\pi)}$ of $M_{x,y}$ in a second array $\widetilde{W}$. Those points will serve as candidates for the starting points λ_0 of Algorithm 4.3.1, but in order to control their number and decrease the required iteration steps of Algorithm 4.3.1, we will preselect the starting points via a box approach such that the computational cost will be reduced.

We start by creating a rectangular grid of small equally sized boxes covering all the sampled points in W. Then, we pick one sample point from the interior of each box that is non-empty as a representative and determine all non-empty boxes that are not surrounded by other non-empty boxes. Now, only the representatives of those boxes will be used as a starting point λ_0. This ensures that the λ_0 will be evenly spread throughout the cloud of computed points even though they might cluster. Moreover, it allows us to restrict our choice of starting points to those that are presumably already close to the boundary, which leads to a higher chance of reaching the boundary within only a few iterations of Algorithm 4.3.1.

For each starting point, we will then select some randomly oriented but equally spread search directions, rotate the matrix $\mathscr{A}$ accordingly and proceed towards the boundary via Algorithm 4.3.1. This yields new vectors $(x, y) \in S_{\mathcal{H}_1} \times S_{\mathcal{H}_2}$ and we subsequently insert the new corresponding points $\lambda_{x,y}^{(\alpha)}$ into W. At this point, it has also shown to be advantageous to compute the other eigenvalues as well and insert them into $\widetilde{W}$.

Afterwards, we do the same for $\alpha + \pi$ in place of α and $\widetilde{W}$ interchanged with W by using a separate grid of boxes and obtain larger clouds of points W and $\widetilde{W}$ as a result. Now, the whole process can be repeated via the construction of new grids of boxes.

If we repeat this procedure over and over again, the number of starting points will eventually remain relatively constant. At this point, we increase the number of boxes, i.e. reduce their size, in order to increase the resolution of the box approach and therefore heighten our ability to distinguish potential starting points in the interior from potential starting points close to the boundary.

When it comes to the determination of the penalty constant p one has to balance two aspects: Larger penalty constants lead to higher accuracy in the given search direction, c.f. Proposition 4.3.1, while smaller penalty constants result in a faster convergence towards the boundary. We therefore start with a small penalty constant to cover a large area in the beginning and increase p over time while also making it dependent on the size of the cloud of points in the current iteration.

Algorithm 4.3.2 explains in pseudocode, how the starting points are selected and how p is chosen. It takes arrays $x \subset S_{\mathcal{H}_1}$, $y \subset S_{\mathcal{H}_2}$ and $W \subset W^2(\mathscr{A})$, the square root of the number of boxes $\mathscr{b}$ and the current iteration i as its input and returns arrays $x_0 \subset S_{\mathcal{H}_1}$, $y_0 \subset S_{\mathcal{H}_2}$ and $\lambda_0 \subset W^2(\mathscr{A})$ as well as the penalty constant p.

Algorithm 4.3.3 contains the full instructions for the computation of the numerical range. It takes the matrix $\mathscr{A}$ in form of its decomposition parts A, B, C and D, an angle α and the time the algorithm should run for $\tau_{\max}$ as its input and returns a cloud of points $(W, \widetilde{W}) \subset W^2(\mathscr{A})$. The square roots of the numbers of boxes $\mathscr{b}$ and $\widetilde{\mathscr{b}}$ will be increased in dependence of counters $\mathscr{c}$ and $\widetilde{\mathscr{c}}$ that keep track of the number of starting points.

```
Algorithm 4.3.2: Grid

Input: $x$, $y$, $W$, $\mathscr{b}$ and $i$
$x_0 = \{\}$, $y_0 = \{\}$, $\lambda_0 = \{\}$, $G = \text{zeros}(\mathscr{b}, \mathscr{b})$ and $I = \text{zeros}(\mathscr{b}, \mathscr{b})$
$\Re_{\max} = \max \Re W$, $\Im_{\max} = \max \Im W$, $\Re_{\min} = \min \Re W$ and $\Im_{\min} = \min \Im W$
$p = i / \max\{\Re_{\max} - \Re_{\min}, \Im_{\max} - \Im_{\min}\}$
$h_{\Re} = (\Re_{\max} - \Re_{\min})/\mathscr{b}$ and $h_{\Im} = (\Im_{\max} - \Im_{\min})/\mathscr{b}$
for $j = 0$ to the length of $W - 1$
    $k = \text{integer}((\Im_{\max} - \Im W[j])/h_{\Im})$ and $l = \text{integer}((\Re W[j] - \Re_{\min})/h_{\Re})$
    if $k == \mathscr{b}$ then $k -= 1$ and if $l == \mathscr{b}$ then $l -= 1$
    if $G[k][l] == 0$ then $G[k][l] = 1$ and $I[k][l] = j + 1$
for $k = 1, \ldots, \mathscr{b} - 2$
    for $l = 1, \ldots, \mathscr{b} - 2$
        if $G[k+m][l+n] = 1$ for all $m, n \in \{-1, 0, 1\}$ then $I[k][l] = 0$
$j = 0$
for $k = 0, \ldots, \mathscr{b} - 1$
    for $l = 0, \ldots, \mathscr{b} - 1$
        if $I[k][l] \neq 0$ then
            $x_0[j] = x[I[k][l] - 1]$, $y_0[j] = y[I[k][l] - 1]$, $\lambda_0[j] = W[I[k][l] - 1]$
              and $j$ += 1
return $(x_0, y_0, \lambda_0, p)$
```

Algorithm 4.3.3: Computing the Quadratic Numerical Range

```
Input: $A$, $B$, $C$, $D$, $\alpha$ and $\tau_{\max}$
$\tau$ = current time + $\tau_{\max}$ and timeflag = false
$\mathscr{A} = \left[\begin{smallmatrix} A & B \\ C & D \end{smallmatrix}\right]$, $W = \{\}$, $\widetilde{W} = \{\}$, $\mathscr{b} = 20$, $\widetilde{\mathscr{b}} = 20$, $\mathscr{c} = 0$ and $\widetilde{\mathscr{c}} = 0$
Generate arrays of random vectors $x \subset S_{\mathcal{H}_1}$ and $y \subset S_{\mathcal{H}_2}$
$n$ = length of $x$
for $i = 0$ to $n - 1$
    $W[i] = \lambda^{(\alpha)}_{x[i],y[i]}$ and $\widetilde{W}[i] = \lambda^{(\alpha+\pi)}_{x[i],y[i]}$
for $i = 0, \dots, \infty$
    for $j = 0, \pi$
        if $j == 0$ then $(x_0, y_0, \lambda_0, p) = \mathrm{Grid}(x, y, W, \mathscr{b}, i)$
        if $j == \pi$ then $(x_0, y_0, \lambda_0, p) = \mathrm{Grid}(x, y, \widetilde{W}, \widetilde{\mathscr{b}}, i)$
        for $k = 0$ to the length of $\lambda_0$ $-1$
            $\theta_0$ = random angle in $[0, 2\pi[$
            for $l = 0, \dots, 4$
                $\theta = \theta_0 + l\frac{2\pi}{5}$ and $\left[\begin{smallmatrix} A & B \\ C & D \end{smallmatrix}\right] = \mathrm{e}^{\mathrm{i}\theta}\mathscr{A}$
                $(\hat{x}, \hat{y})$ = Seeking_the_Boundary$(A, B, C, D, x_0[k],$
                                    $y_0[k], \mathrm{e}^{\mathrm{i}\theta}\lambda_0[k], p, \alpha + j - \theta, 2)$
                $\hat{n}$ = length of $\hat{x}$
                for $m = 0$ to $\hat{n} - 1$
                    $W[n+m] = \lambda^{(\alpha)}_{\hat{x}[m],\hat{y}[m]}$ and $\widetilde{W}[n+m] = \lambda^{(\alpha+\pi)}_{\hat{x}[m],\hat{y}[m]}$
                    $x[n+m] = \hat{x}[m]$ and $y[n+m] = \hat{y}[m]$
                if current time $> \tau$ then timeflag = true and break
                $n \mathrel{+}= \hat{n}$
            if timeflag == true then break
        if timeflag == true then break
        if $j == 0$ then
            if $1 \geq (\text{length of } \lambda_0)/\mathscr{c} > 0.99$ then $\mathscr{b} = \mathrm{integer}(\sqrt{2}\mathscr{b})$
            $\mathscr{c}$ = length of $\lambda_0$
        if $j == \pi$ then
            if $1 \geq (\text{length of } \lambda_0)/\widetilde{\mathscr{c}} > 0.99$ then $\widetilde{\mathscr{b}} = \mathrm{integer}(\sqrt{2}\widetilde{\mathscr{b}})$
            $\widetilde{\mathscr{c}}$ = length of $\lambda_0$
    if timeflag == true then break
return $(W, \widetilde{W})$
```

Example 4.3.2. Let us consider a small example to see how the starting points are selected from an existing cloud of points.

The top-left panel of Figure 4.4 displays a set of 100 points within one of the QNR components that have been computed via random vector sampling. The top-right panel provides a zoomed-in view, showing the same cloud overlaid by a

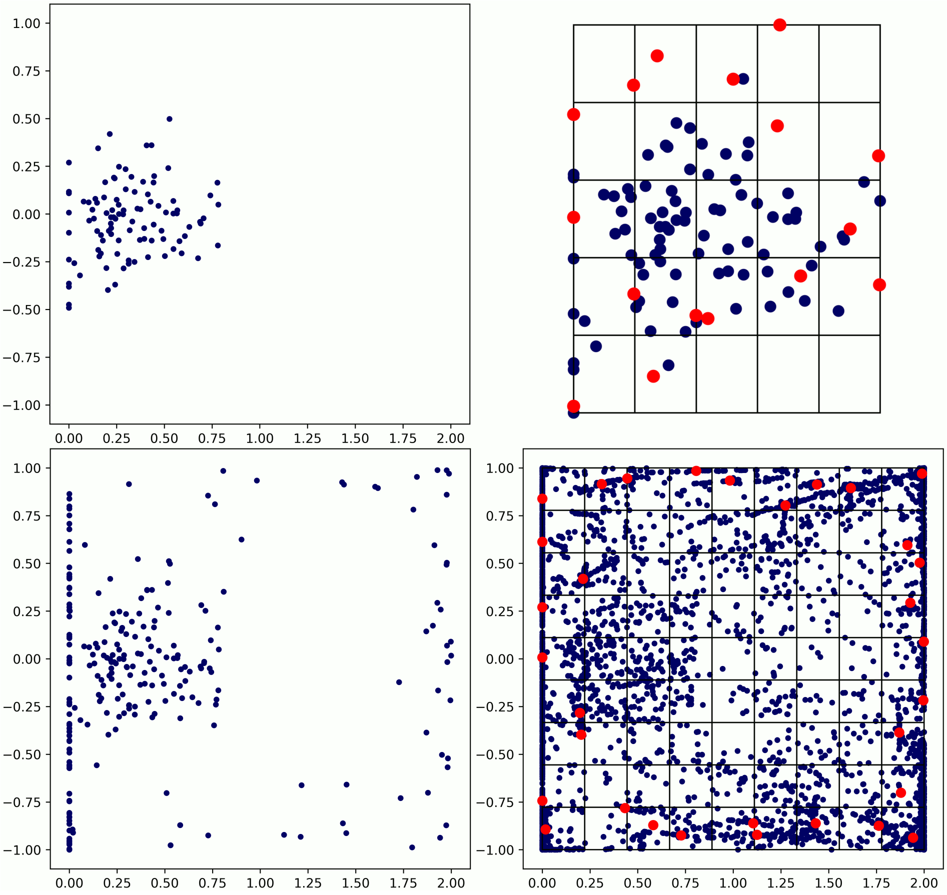


Figure 4.4: Starting point selection via grid of boxes

rectangular grid of boxes with $b = 5$ that precisely encompasses the points. Within each non-empty box that is not surrounded by other non-empty boxes, one point is selected and highlighted in red. These red points are now chosen as starting points for the algorithm's first iteration and five random search directions are determined for each of them. The bottom-left panel depicts the subsequent cloud of points after the iterative process 4.3.1 was executed for all of these directions, illustrating the algorithm's advancement toward the component's boundary. Note that in this simple example, the QNR component to be computed coincides with the square $[0, 2] \times [-\mathrm{i}, \mathrm{i}]$.

After 15 iterations, as shown in the bottom-right panel, the algorithm has effectively identified portions within the interior of the quadratic numerical range, strategically choosing starting points (red) that are relatively uniformly distributed and close to the boundary.

4.4 Examples

The following pictures are the result of a Python implementation of Algorithm 4.3.3. Here, the search for $(s,t) \in [0,2\pi] \times [0,2\pi]$ such that $f_{\alpha,\lambda_0}\big(\varphi(s),\psi(t)\big)$ is maximal with φ and ψ as in (4.10) is further reduced to a one-dimensional optimization, i.e. $s = t$, in order to speed this step up. This allows us to compute much more points in the QNR within the same amount of time and we obtain better pictures in the end.

For comparison, we also provide pictures computed via random vector sampling. To obtain a uniform distribution of vectors on the unit spheres, they are generated by sampling from a multivariate normal distribution with zero mean and identity covariance matrix, followed by normalization.

Example 4.4.1. Let us consider the matrix

$$\mathscr{A}_1 = \left[\begin{array}{ccccc|ccccc}
2 & \mathrm{i} & 0 & \dots & 0 & 1 & 3+\mathrm{i} & 0 & \dots & 0 \\
\mathrm{i} & \ddots & \ddots & \ddots & \vdots & 3+\mathrm{i} & \ddots & \ddots & \ddots & \vdots \\
0 & \ddots & \ddots & \ddots & 0 & 0 & \ddots & \ddots & \ddots & 0 \\
\vdots & \ddots & \ddots & \ddots & \mathrm{i} & \vdots & \ddots & \ddots & \ddots & 3+\mathrm{i} \\
0 & \dots & 0 & \mathrm{i} & 2 & 0 & \dots & 0 & 3+\mathrm{i} & 1 \\
\hline
1 & 3+\mathrm{i} & 0 & \dots & 0 & -2 & \mathrm{i} & 0 & \dots & 0 \\
3+\mathrm{i} & \ddots & \ddots & \ddots & \vdots & \mathrm{i} & \ddots & \ddots & \ddots & \vdots \\
0 & \ddots & \ddots & \ddots & 0 & 0 & \ddots & \ddots & \ddots & 0 \\
\vdots & \ddots & \ddots & \ddots & 3+\mathrm{i} & \vdots & \ddots & \ddots & \ddots & \mathrm{i} \\
0 & \dots & 0 & 3+\mathrm{i} & 1 & 0 & \dots & 0 & \mathrm{i} & -2
\end{array}\right],$$

where the blocks A, B, C and D are equally sized tridiagonal matrices, cf. [52, Example 1.1.5]. In Figure 4.5, the results of the algorithm are compared to the random vector sampling method when executed for the determination of the QNR of $\mathscr{A}_1$ with $\dim(\mathscr{A}_1) = 40$. In the top row, the execution time was one minute and in the bottom row 40 minutes while the plots in the left column are a result of the algorithm and the plots in the right column are a result of the random vector sampling method. Here, α was chosen to be zero such that the sets W (dark blue) and $\widetilde{W}$ (light blue) coincide with the two disconnected components of $W^2(\mathscr{A}_1)$. As we see, the algorithm is capable of detecting the rough shape of the quadratic numerical range already after a short period of time and refines the result very well afterwards while the random vector sampling method is only able to locate a small area of the QNR which gets slightly enlarged over time.

Figure 4.6 demonstrates the efficacy of the algorithm (left) even when the dimension of $\mathscr{A}_1$ is increased to 4000. Here, the superiority over the random vector sampling method (right) becomes even more visible. The computation time was two hours in both pictures.

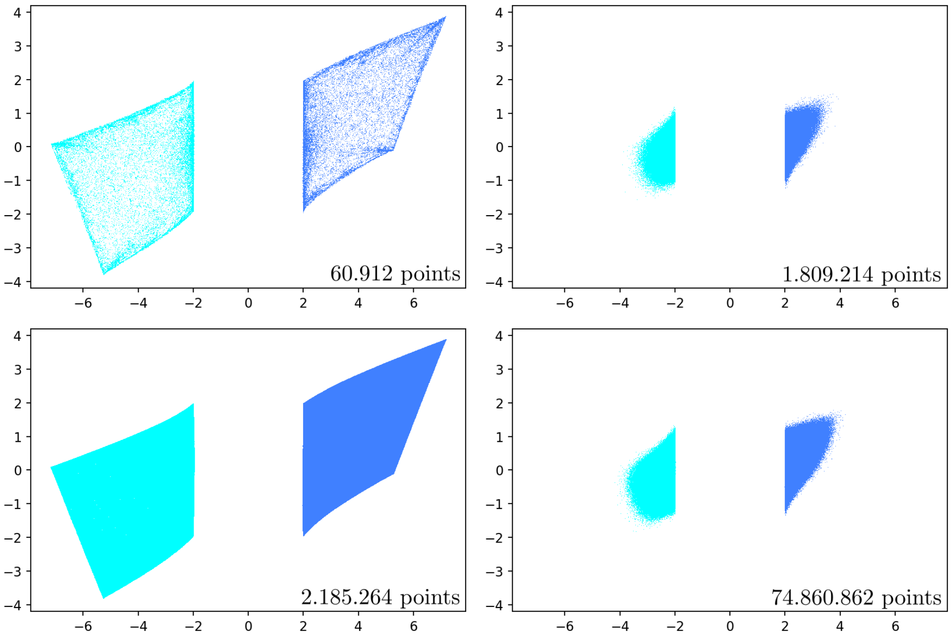


Figure 4.5: QNR of $\mathscr{A}_1$: Algorithm versus random vector sampling for different amounts of time

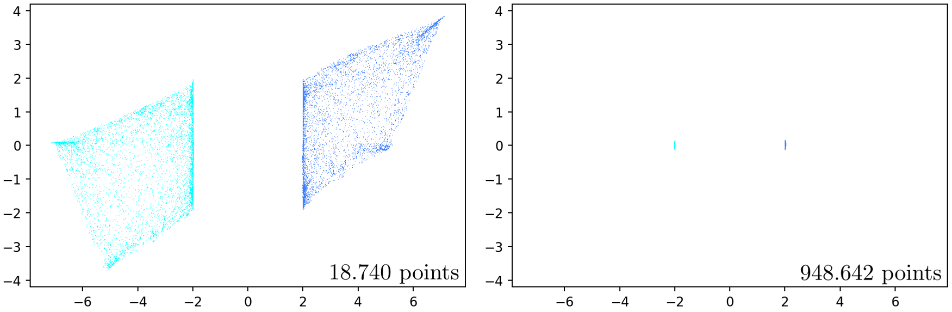


Figure 4.6: QNR of high dimensional $\mathscr{A}_1$: Algorithm versus random vector sampling

Example 4.4.2. Let us consider a smaller matrix like

$$\mathscr{A}_2 = \left[\begin{array}{cccc|cccc} 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 \\ \hline -2 & -1 & 0 & 0 & \mathrm{i} & 5\mathrm{i} & 0 & 0 \\ -1 & -2 & -1 & 0 & -5\mathrm{i} & \mathrm{i} & 5\mathrm{i} & 0 \\ 0 & -1 & -2 & -1 & 0 & -5\mathrm{i} & \mathrm{i} & 5\mathrm{i} \\ 0 & 0 & -1 & -2 & 0 & 0 & -5\mathrm{i} & \mathrm{i} \end{array}\right],$$

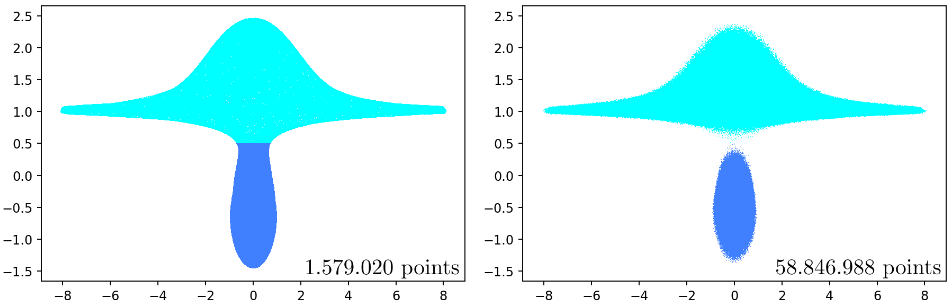


Figure 4.7: QNR of $\mathscr{A}_2$ with the algorithm (left) and random vector sampling (right)

cf. [52, Example 1.3.3]. Figure 4.7 shows the quadratic numerical range of $\mathscr{A}_2$ after executing the algorithm and the random vector sampling method with $\alpha = \pi/2$ for 30 minutes each. Although the random vector method generally covers a much bigger part of the quadratic numerical range of smaller matrices like this when compared to higher dimensional ones, here, it still fails to adequately depict some parts of the boundary and struggles to close the gap between the two components, which seem to be connected as the plot of the algorithm suggests. As we see, this is not compensated by the fact that only 1.579.020 points were computed with the algorithm while 58.846.988 points were sampled via the random vector method within the same amount of time.

Example 4.4.3. Let us consider the matrices

$$\mathscr{A}_3 = \left[\begin{array}{cc|cc} 0 & 0 & 0 & 1 \\ 0 & 1 & 2 & 3 \\ \hline 0 & -2 & -1 & 0 \\ -1 & -3 & 0 & 0 \end{array}\right],$$

cf. [52, Example 1.3.10], and

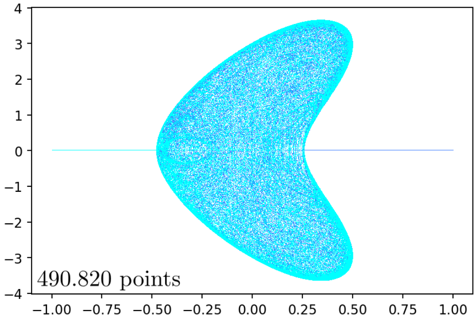


Figure 4.8: QNR of $\mathscr{A}_3$

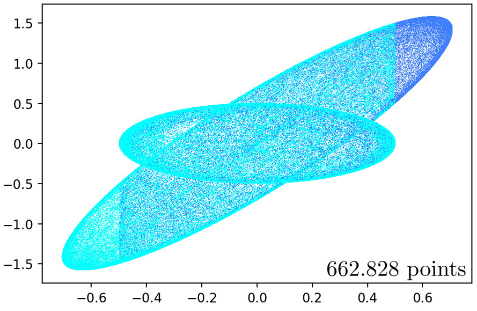


Figure 4.9: QNR of $\mathscr{A}_4$

$$\mathscr{A}_4 = \left[\begin{array}{ccc|cc} 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1+\mathrm{i} & 0 & 0 \\ 0 & 2\mathrm{i} & 0 & 0 & 0 \\ \hline 0 & 0 & 0 & 0 & 0 \\ -1 & 2 & -2 & \mathrm{i} & 0 \end{array}\right].$$

Figures 4.8 and 4.9 demonstrate how the result of the algorithm can look like if the QNR consists of only one connected component and α is arbitrarily chosen to be 0. They are the result of an execution of the algorithm for 15 minutes each.

4.5 Concentration Phenomenon for Random Sampling

As we see in the examples of Section 4.4 and especially in Figures 4.5 and 4.6 the points in the quadratic numerical range computed via the random vector sampling method are very unequally spread and cluster in a small subset of each component. In this section, we will examine this phenomenon and prove that the probability of a sampling point to fall outside of a small neighborhood of the expected value decays exponentially with an increase of the dimension of the matrix when its norm stays constant.

Proposition 4.5.1. *Consider the probability spaces $(S_{\mathcal{H}_i}, \mathcal{B}(S_{\mathcal{H}_i}), \sigma_i)$, $i = 1, 2$, where σ_i is the normalized surface measure. Let*

$$M\colon S_{\mathcal{H}_1} \times S_{\mathcal{H}_2} \to \mathbb{C}^{2\times 2}, \quad (x, y) \mapsto \begin{bmatrix} \langle Ax, x\rangle & \langle By, x\rangle \\ \langle Cx, y\rangle & \langle Dy, y\rangle \end{bmatrix}.$$

Then the expected value $\mathbb{E}M$ of M is given by

$$\mathbb{E}M = \begin{bmatrix} \frac{\operatorname{trace}(A)}{\dim(\mathcal{H}_1)} & 0 \\ 0 & \frac{\operatorname{trace}(D)}{\dim(\mathcal{H}_2)} \end{bmatrix}.$$

Proof. The expected value is

$$\begin{aligned} \mathbb{E}M &= \int_{S_{\mathcal{H}_1}\times S_{\mathcal{H}_2}} \begin{bmatrix} \langle Ax, x\rangle & \langle By, x\rangle \\ \langle Cx, y\rangle & \langle Dy, y\rangle \end{bmatrix} \mathrm{d}(\sigma_1 \times \sigma_2)(x, y) \\ &= \begin{bmatrix} \int_{S_{\mathcal{H}_1}} \langle Ax, x\rangle \,\mathrm{d}\sigma_1(x) & \int_{S_{\mathcal{H}_1}\times S_{\mathcal{H}_2}} \langle By, x\rangle \,\mathrm{d}(\sigma_1 \times \sigma_2)(x, y) \\ \int_{S_{\mathcal{H}_1}\times S_{\mathcal{H}_2}} \langle Cx, y\rangle \,\mathrm{d}(\sigma_1 \times \sigma_2)(x, y) & \int_{S_{\mathcal{H}_2}} \langle Dy, y\rangle \,\mathrm{d}\sigma_2(y) \end{bmatrix}. \end{aligned}$$

Here,

$$\begin{aligned} \int_{S_{\mathcal{H}_1}\times S_{\mathcal{H}_2}} \langle By, x\rangle \,\mathrm{d}(\sigma_1 \times \sigma_2)(x, y) &= \int_{S_{\mathcal{H}_2}} \left\langle By, \int_{S_{\mathcal{H}_1}} x \,\mathrm{d}\sigma_1(x) \right\rangle \mathrm{d}\sigma_2(y) \\ &= \int_{S_{\mathcal{H}_2}} \langle By, 0\rangle \,\mathrm{d}\sigma_2(y) \\ &= 0 \end{aligned}$$

by Fubini's theorem and via a similar argumentation we also obtain

$$\int_{S_{\mathcal{H}_1}\times S_{\mathcal{H}_2}} \langle Cx, y\rangle \,\mathrm{d}(\sigma_1 \times \sigma_2)(x,y) = 0.$$

Let d be the dimension of $\mathcal{H}_1$ and denote by $e_1, \ldots, e_d$ an orthonormal basis of $\mathcal{H}_1$. Then the trace of A is given by

$$\operatorname{trace}(A) = \sum_{k=1}^{d} \langle Ae_k, e_k\rangle = \sum_{k=1}^{d} \langle AUe_k, Ue_k\rangle,$$

where U is an arbitrary unitary matrix. Denoting the normalized Haar measure on the unitary group $\mathcal{U}(d)$ by μ we have by Fubini's theorem

$$\begin{aligned}
\operatorname{trace}(A) &= \int_{\mathcal{U}(d)} \sum_{k=1}^{d} \langle AUe_k, Ue_k\rangle \,\mathrm{d}\mu(U) \\
&= \int_{S_{\mathcal{H}_1}} \int_{\mathcal{U}(d)} \sum_{k=1}^{d} \langle AUe_k, Ue_k\rangle \,\mathrm{d}\mu(U)\,\mathrm{d}\sigma_1(x) \\
&= d \int_{S_{\mathcal{H}_1}} \int_{\mathcal{U}(d)} \langle AUx, Ux\rangle \,\mathrm{d}\mu(U)\,\mathrm{d}\sigma_1(x) \\
&= d \int_{\mathcal{U}(d)} \int_{S_{\mathcal{H}_1}} \langle AUx, Ux\rangle \,\mathrm{d}\sigma_1(x)\,\mathrm{d}\mu(U) \\
&= d \int_{S_{\mathcal{H}_1}} \langle Ax, x\rangle \,\mathrm{d}\sigma_1(x)
\end{aligned}$$

due to the invariance of the Haar measure and σ_1 under unitary transformations.

It follows analogously that

$$\operatorname{trace}(D) = \dim(\mathcal{H}_2) \int_{S_{\mathcal{H}_2}} \langle Dy, y\rangle \,\mathrm{d}\sigma_2(y),$$

which concludes the proof. ❑

Recall from Definition 1.2.3 that for two nonempty sets $K, L \subset \mathbb{C}$ we define the distance $\operatorname{dist}(K, L)$ via

$$\operatorname{dist}(K, L) = \sup_{k\in K}\left(\inf_{l\in L} \|k - l\|\right)$$

and the Hausdorff-distance $\mathrm{d_H}(K, L)$ via

$$\mathrm{d_H}(K, L) = \max\left\{\operatorname{dist}(K, L), \operatorname{dist}(L, K)\right\}.$$

Note, that $K \subset B_\varepsilon(L)$ whenever $\operatorname{dist}(K, L) < \varepsilon$ and $\operatorname{dist}(K, L) \le \varepsilon$ whenever $K \subset B_\varepsilon(L)$.

Lemma 4.5.2. *Let $M_1, M_2 \in \mathbb{C}^{2\times 2}$. Then*

$$\mathrm{d_H}\big(\sigma(M_1), \sigma(M_2)\big) \le \big((\|M_1\| + \|M_2\|)\|M_1 - M_2\|\big)^{\frac{1}{2}}.$$

Proof. For a $\lambda \in \varrho(M_1)$, [32, p. 28] yields that

$$\|(\lambda - M_1)^{-1}\| \le \frac{\|\lambda - M_1\|}{|\det(\lambda - M_1)|} \le \frac{\|\lambda - M_1\|}{\operatorname{dist}(\lambda, \sigma(M_1))^2}$$

or in other words

$$\operatorname{dist}(\lambda, \sigma(M_1)) \le \big(\|\lambda - M_1\| \|(\lambda - M_1)^{-1}\|^{-1}\big)^{\frac{1}{2}}.$$

If we further assume $\lambda \in \sigma(M_2)$, we have

$$\|\lambda - M_1\| \le \|M_2\| + \|M_1\|$$

on one hand and on the other hand we obtain

$$\|(\lambda - M_1)^{-1}\|^{-1} \le \|M_2 - M_1\|$$

because otherwise $\|(\lambda - M_1)^{-1}\|^{-1} > \|M_2 - M_1\|$ implies

$$\|(M_2 - M_1)(\lambda - M_1)^{-1}\| < 1$$

and therefore $I - (M_2 - M_1)(\lambda - M_1)^{-1} = (\lambda - M_2)(\lambda - M_1)^{-1}$ is invertible by a Neumann series argument. This yields $\lambda \in \varrho(M_2)$, which is a contradiction.

Hence, we have

$$\operatorname{dist}(\lambda, \sigma(M_1)) \le \big((\|M_1\| + \|M_2\|)\|M_1 - M_2\|\big)^{\frac{1}{2}}$$

for every $\lambda \in \sigma(M_2)$ and we analogously obtain

$$\operatorname{dist}(\lambda, \sigma(M_2)) \le \big((\|M_1\| + \|M_2\|)\|M_1 - M_2\|\big)^{\frac{1}{2}}$$

for every $\lambda \in \sigma(M_1)$. Thus,

$$\begin{aligned}\mathrm{d_H}\big(\sigma(M_1), \sigma(M_2)\big) &= \max\big\{\operatorname{dist}\big(\sigma(M_1), \sigma(M_2)\big), \operatorname{dist}\big(\sigma(M_2), \sigma(M_1)\big)\big\} \\ &= \max\left\{\sup_{\lambda\in\sigma(M_1)} \operatorname{dist}(\lambda, \sigma(M_2)), \sup_{\lambda\in\sigma(M_2)} \operatorname{dist}(\lambda, \sigma(M_1))\right\} \\ &\le \big((\|M_1\| + \|M_2\|)\|M_1 - M_2\|\big)^{\frac{1}{2}}.\end{aligned}$$

❑

Theorem 4.5.3. *Denote by S^{n-1} the $(n-1)$-dimensional sphere in $\mathbb{R}^n$ and let $f\colon S^{n-1} \to \mathbb{R}$ be a function with Lipschitz constant L. Then for all $\varepsilon > 0$ we have*

$$\sigma\left(\left|f(x) - \int_{S^{n-1}} f \,\mathrm{d}\sigma\right| > \varepsilon\right) \le 4\exp\left(-\frac{\delta\varepsilon^2 n}{L^2}\right)$$

where σ is the normalized surface measure on S^{n-1} and $\delta > 0$ an absolute constant.

Proof. This is [44, Corollary V.2]. ❑

Theorem 4.5.4. *Consider M as in Proposition 4.5.1. Then we have for all $\varepsilon > 0$*

$$\sigma_1 \times \sigma_2\big(\mathrm{d_H}(\sigma(M_{x,y}), \sigma(\mathbb{E}M)) > \varepsilon\big) \leq 32 \exp\left(-\beta \frac{\varepsilon^4 n_0}{\|\mathscr{A}\|^4}\right)$$

where $\beta > 0$ is an absolute constant and $n_0 = \min\{\dim(\mathcal{H}_1), \dim(\mathcal{H}_2)\}$.

Proof. We start by considering the function $\Re\langle A\cdot, \cdot\rangle \colon S_{\mathcal{H}_1} \to \mathbb{R}$ for which we have

$$|\Re\langle Ax, x\rangle - \Re\langle Ay, y\rangle| \leq 2\|A\|\|x - y\|, \quad x, y \in S_{\mathcal{H}_1}.$$

Thus, from Theorem 4.5.3 and Proposition 4.5.1, we obtain

$$\sigma_1\left(\left|\Re\langle Ax, x\rangle - \Re\frac{\operatorname{trace}(A)}{\dim(\mathcal{H}_1)}\right| > \varepsilon\right) \leq 4 \exp\left(-\delta\varepsilon^2 \frac{\dim(\mathcal{H}_1)}{4\|A\|^2}\right)$$

and analogously

$$\sigma_1\left(\left|\Im\langle Ax, x\rangle - \Im\frac{\operatorname{trace}(A)}{\dim(\mathcal{H}_1)}\right| > \varepsilon\right) \leq 4 \exp\left(-\delta\varepsilon^2 \frac{\dim(\mathcal{H}_1)}{4\|A\|^2}\right)$$

with an absolute constant $\delta > 0$. Combining both of these statements, we get

$$\begin{aligned}\sigma_1\left(\left|\langle Ax, x\rangle - \frac{\operatorname{trace}(A)}{\dim(\mathcal{H}_1)}\right| > \varepsilon\right) &\leq \sigma_1\left(\left|\Re\langle Ax, x\rangle - \Re\frac{\operatorname{trace}(A)}{\dim(\mathcal{H}_1)}\right| > \frac{\varepsilon}{\sqrt{2}}\right)\\ &\quad + \sigma_1\left(\left|\Im\langle Ax, x\rangle - \Im\frac{\operatorname{trace}(A)}{\dim(\mathcal{H}_1)}\right| > \frac{\varepsilon}{\sqrt{2}}\right) \qquad (4.11)\\ &\leq 8 \exp\left(-\delta\varepsilon^2 \frac{\dim(\mathcal{H}_1)}{8\|A\|^2}\right)\end{aligned}$$

and via the same argumentation with D in place of A we obtain

$$\sigma_2\left(\left|\langle Dy, y\rangle - \frac{\operatorname{trace}(D)}{\dim(\mathcal{H}_2)}\right| > \varepsilon\right) \leq 8 \exp\left(-\delta\varepsilon^2 \frac{\dim(\mathcal{H}_2)}{8\|D\|^2}\right). \qquad (4.12)$$

In order to find an estimate like this for $\sigma_1 \times \sigma_2(|\langle By, x\rangle| > \varepsilon)$ as well we first fix $y \in S_{\mathcal{H}_2}$ and consider the function $g_y \colon S_{\mathcal{H}_1} \to \mathbb{C}$, $x \mapsto \langle By, x\rangle$. For this we have

$$\sigma_1\left(|g_y(x)| > \varepsilon\right) \leq 8 \exp\left(-\delta\varepsilon^2 \frac{\dim(\mathcal{H}_1)}{2\|B\|^2}\right)$$

again by Theorem 4.5.3 and a similar argumentation as before because $\mathbb{E}g_y = 0$.

Considering $\Omega := \{(x,y) \in S_{\mathcal{H}_1} \times S_{\mathcal{H}_2} \mid |\langle By, x\rangle| > \varepsilon\}$, we then obtain

$$\begin{aligned}\sigma_1 \times \sigma_2(|\langle By, x\rangle| > \varepsilon) &= \int_{S_{\mathcal{H}_2}} \int_{S_{\mathcal{H}_1}} \mathbb{1}_\Omega(x,y)\, \mathrm{d}\sigma_1\, \mathrm{d}\sigma_2 \\ &= \int_{S_{\mathcal{H}_2}} \sigma_1\left(|g_y(x)| > \varepsilon\right)\, \mathrm{d}\sigma_2 \\ &\leq \int_{S_{\mathcal{H}_2}} 8\exp\left(-\delta\varepsilon^2 \frac{\dim(\mathcal{H}_1)}{2\|B\|^2}\right)\, \mathrm{d}\sigma_2 \\ &= 8\exp\left(-\delta\varepsilon^2 \frac{\dim(\mathcal{H}_1)}{2\|B\|^2}\right).\end{aligned} \tag{4.13}$$

The estimate

$$\sigma_1 \times \sigma_2(|\langle Cx, y\rangle| > \varepsilon) \leq 8\exp\left(-\delta\varepsilon^2 \frac{\dim(\mathcal{H}_2)}{2\|C\|^2}\right) \tag{4.14}$$

can be shown via the same arguments and we then combine (4.11), (4.12), (4.13) and (4.14) to obtain

$$\begin{aligned}&\sigma_1 \times \sigma_2(\|M_{x,y} - \mathbb{E}M\| > \varepsilon) \\ &\qquad \leq \sigma_1 \times \sigma_2(\|M_{x,y} - \mathbb{E}M\|_{\mathrm{F}} > \varepsilon) \\ &\qquad \leq 32\exp\left(-\delta\varepsilon^2 \frac{n_0}{8\max\{4\|A\|^2, \|B\|^2, \|C\|^2, 4\|D\|^2\}}\right) \\ &\qquad \leq 32\exp\left(-\delta\varepsilon^2 \frac{n_0}{32\|\mathscr{A}\|^2}\right),\end{aligned} \tag{4.15}$$

where $\|\cdot\|_{\mathrm{F}}$ denotes the Frobenius norm.

Next, we apply Lemma 4.5.2 to obtain

$$\mathrm{d_H}\big(\sigma(M_{x,y}), \sigma(\mathbb{E}M)\big) \leq \big((\|M_{x,y}\| + \|\mathbb{E}M\|)\|M_{x,y} - \mathbb{E}M\|\big)^{\frac{1}{2}},$$

where we have $\|M_{x,y}\| \leq \|\mathscr{A}\|$ because $M_{x,y} = P\mathscr{A}|_{\mathrm{ran}P}$, where P is the orthogonal projection to the two-dimensional subspace of $\mathcal{H}_1 \oplus \mathcal{H}_2$ spanned by $\begin{bmatrix} x \\ 0 \end{bmatrix}$ and $\begin{bmatrix} 0 \\ y \end{bmatrix}$. Moreover, $\|\mathbb{E}M\| \leq \|\mathbb{E}M\|_{\mathrm{F}} \leq \sqrt{2}\|\mathscr{A}\|$ because the trace of a matrix is the sum of its eigenvalues. Thus,

$$\mathrm{d_H}\big(\sigma(M_{x,y}), \sigma(\mathbb{E}M)\big) \leq \big((1+\sqrt{2})\|\mathscr{A}\|\|M_{x,y} - \mathbb{E}M\|\big)^{\frac{1}{2}}$$

and we conclude by using (4.15) that

$$\begin{aligned}&\sigma_1 \times \sigma_2\big(\mathrm{d_H}(\sigma(M_{x,y}), \sigma(\mathbb{E}M)) > \varepsilon\big) \\ &\qquad\qquad \leq \sigma_1 \times \sigma_2\left(\left((1+\sqrt{2})\|\mathscr{A}\|\|M_{x,y} - \mathbb{E}M\|\right)^{\frac{1}{2}} > \varepsilon\right)\end{aligned}$$

$$
\begin{aligned}
&= \sigma_1 \times \sigma_2 \left(\|M_{x,y} - \mathbb{E}M\| > \frac{\varepsilon^2}{(1+\sqrt{2})\|\mathscr{A}\|} \right) \\
&\leq 32 \exp\left(-\delta \frac{\varepsilon^4}{(3+2\sqrt{2})\|\mathscr{A}\|^2} \frac{n_0}{32\|\mathscr{A}\|^2} \right) \\
&= 32 \exp\left(-\beta \frac{\varepsilon^4 n_0}{\|\mathscr{A}\|^4} \right)
\end{aligned}
$$

with $\beta = \frac{\delta}{(96+64\sqrt{2})}$. ❑

Remark 4.5.5. Equation (4.11) can also be interpreted in the context of the numerical range and yields an estimate for the probability of a point in $W(\mathscr{A})$ that is computed via the random vector sampling method to fall outside of a small neighborhood of the expected value.

In [43], Martinsson and Tropp reformulated a result from [20] such that an exponential bound for the deviation of the estimation of the trace of a matrix A via $\langle Ax, x\rangle$ is obtained. The proof however relies on A to be a self-adjoint positive semi-definite matrix.

Example 4.5.6. Let us consider the matrix

$$
\mathscr{A}_5 = \left[\begin{array}{cccccc|cccccc}
2 & 0 & \cdots & \cdots & \cdots & 0 & 1 & 0 & \cdots & \cdots & \cdots & 0 \\
0 & \ddots & \ddots & & & \vdots & 0 & 0 & \ddots & & & \vdots \\
\vdots & \ddots & 2 & \ddots & & \vdots & \vdots & \ddots & \ddots & \ddots & & \vdots \\
\vdots & & \ddots & -2 & \ddots & \vdots & \vdots & & \ddots & \ddots & \ddots & \vdots \\
\vdots & & & \ddots & \ddots & 0 & \vdots & & & \ddots & \ddots & 0 \\
0 & \cdots & \cdots & \cdots & 0 & -2 & 0 & \cdots & \cdots & \cdots & 0 & 0 \\
\hline
0 & 0 & \cdots & \cdots & \cdots & 0 & 1+\mathrm{i} & 0 & \cdots & \cdots & \cdots & 0 \\
0 & \ddots & \ddots & & & \vdots & 0 & \ddots & \ddots & & & \vdots \\
\vdots & \ddots & \ddots & \ddots & & \vdots & \vdots & \ddots & 1+\mathrm{i} & \ddots & & \vdots \\
\vdots & & \ddots & \ddots & \ddots & \vdots & \vdots & & \ddots & 1-\mathrm{i} & \ddots & \vdots \\
\vdots & & & \ddots & 0 & 0 & \vdots & & & \ddots & \ddots & 0 \\
0 & \cdots & \cdots & \cdots & 0 & 1 & 0 & \cdots & \cdots & \cdots & 0 & 1-\mathrm{i}
\end{array} \right]
$$

with A, B, C and D equally sized. We have $\|\mathscr{A}_5\| \approx 2.36$ independent of its dimension. In Figure 4.10, each plot depicts 10.000.000 points in the QNR of a version of $\mathscr{A}_5$ with $\dim(\mathscr{A}_5) = 4^n$, $n = 1, \dots, 4$, that were generated via the random vector sampling method. The concentration phenomenon proven in Theorem 4.5.4 becomes clearly visible and with an increase of the dimension, the QNR seems to split into two disconnected components, which is not the case as Figure 4.11 shows. There, only 225.048 points have been computed, but they give a much more accurate picture of the QNR.

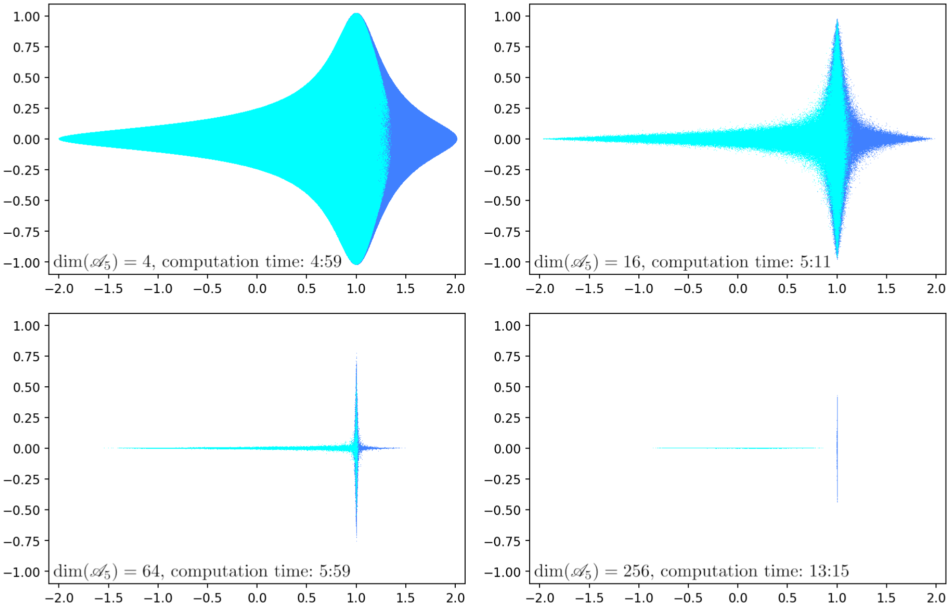


Figure 4.10: QNR of $\mathscr{A}_5$ computed with the random vector sampling method for different dimensions

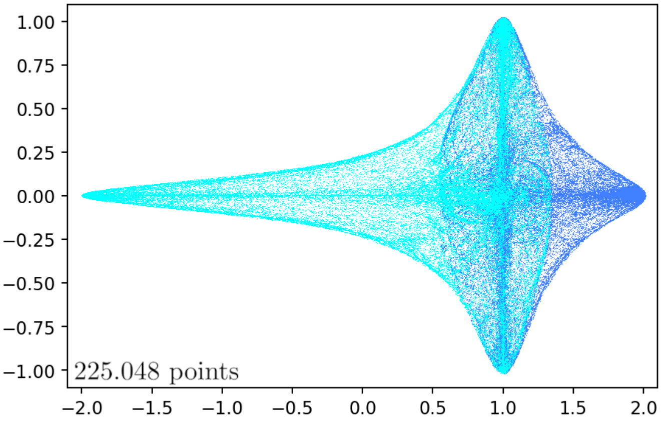


Figure 4.11: QNR of $\mathscr{A}_5$ with $\dim(\mathscr{A}_5) = 256$ computed with the algorithm in 13 minutes

Bibliography

[1] W. Arendt and K. Urban. *Partielle Differenzialgleichungen. Eine Einführung in analytische und numerische Methoden.* Berlin: Springer Spektrum, 2nd edition edition, 2018.

[2] N. Bebiano, J. a. da Providência, A. Nata, and J. a. P. da Providência. Revisiting the inverse field of values problem. *Electron. Trans. Numer. Anal.*, 42:1–12, 2014.

[3] S. Bögli. *Spectral approximation for linear operators and applications.* PhD thesis, University of Bern, 2014.

[4] S. Bögli. Local convergence of spectra and pseudospectra. *J. Spectr. Theory*, 8(3):1051–1098, 2018.

[5] S. Bögli and M. Marletta. Essential numerical ranges for linear operator pencils. *IMA J. Numer. Anal.*, 40(4):2256–2308, 2020.

[6] S. Bögli and P. Siegl. Remarks on the convergence of pseudospectra. *Integral Equations Oper. Theory*, 80(3):303–321, 2014.

[7] A. Böttcher and H. Wolf. Spectral approximation for Segal-Bargmann space Toeplitz operators. In *Linear operators (Warsaw, 1994)*, volume 38 of *Banach Center Publ.*, pages 25–48. Polish Acad. Sci. Inst. Math., Warsaw, 1997.

[8] T. Braconnier and N. J. Higham. Computing the field of values and pseudospectra using the Lanczos method with continuation. *BIT*, 36(3):422–440, 1996. International Linear Algebra Year (Toulouse, 1995).

[9] F. Chatelin. *Spectral approximation of linear operators.* Computer Science and Applied Mathematics. Academic Press, Inc. [Harcourt Brace Jovanovich, Publishers], New York, 1983.

[10] M. J. Colbrook, B. Roman, and A. C. Hansen. How to compute spectra with error control. *Phys. Rev. Lett.*, 122(25):250201, 6, 2019.

[11] J. B. Conway. *Functions of one complex variable*, volume 11 of *Graduate Texts in Mathematics.* Springer-Verlag, New York-Berlin, second edition, 1978.

[12] C. Cowen and E. Harel. An effective algorithm for computing the numerical range, 1995. URL: `https://www.math.iupui.edu/~ccowen/Downloads/33NumRange.pdf` [cited 2023-10-25].

[13] M. Embree and L. N. Trefethen. `EigTool`. URL: `https://www.cs.ox.ac.uk/pseudospectra/eigtool/` [cited 2023-10-25].

[14] M. Embree and L. N. Trefethen. Pseudospectra gateway. URL: `http://www.comlab.ox.ac.uk/pseudospectra` [cited 2023-10-25].

[15] K.-J. Engel and R. Nagel. *One-parameter semigroups for linear evolution equations*, volume 194 of *Graduate Texts in Mathematics*. Springer-Verlag, New York, 2000. With contributions by S. Brendle, M. Campiti, T. Hahn, G. Metafune, G. Nickel, D. Pallara, C. Perazzoli, A. Rhandi, S. Romanelli and R. Schnaubelt.

[16] M. Fazlollahi. Extension of quadratic numerical range of block operator matrices. *Int. J. Contemp. Math. Sci.*, 3(29-32):1529–1534, 2008.

[17] A. Frommer, B. Jacob, K. Kahl, C. Wyss, and I. Zwaan. Krylov type methods for linear systems exploiting properties of the quadratic numerical range. *Electron. Trans. Numer. Anal.*, 53:541–561, 2020.

[18] A. Frommer, B. Jacob, L. A. Vorberg, C. Wyss, and I. N. Zwaan. Pseudospectrum enclosures by discretization. *Integral Equations and Operator Theory*, 93:1–31, 2020.

[19] D. Gerecht, R. Rannacher, and W. Wollner. Computational aspects of pseudospectra in hydrodynamic stability analysis. *J. Math. Fluid Mech.*, 14(4):661–692, 2012.

[20] S. Gratton and D. Titley-Peloquin. Improved bounds for small-sample estimation. *SIAM J. Matrix Anal. Appl.*, 39(2):922–931, 2018.

[21] P. Grisvard. *Elliptic problems in nonsmooth domains*, volume 24 of *Monographs and Studies in Mathematics*. Pitman (Advanced Publishing Program), Boston, MA, 1985.

[22] K. Gustafson. The Toeplitz-Hausdorff theorem for linear operators. *Proc. Amer. Math. Soc.*, 25:203–204, 1970.

[23] K. E. Gustafson and D. K. M. Rao. *Numerical range.* Universitext. Springer-Verlag, New York, 1997. The field of values of linear operators and matrices.

[24] M. Haase. *The functional calculus for sectorial operators*, volume 169 of *Operator Theory: Advances and Applications*. Birkhäuser Verlag, Basel, 2006.

[25] D. Hinrichsen and A. J. Pritchard. On spectral variations under bounded real matrix perturbations. *Numer. Math.*, 60(4):509–524, 1992.

[26] M. E. Hochstenbach, D. A. Singer, and P. F. Zachlin. Eigenvalue inclusion regions from inverses of shifted matrices. *Linear Algebra Appl.*, 429(10):2481–2496, 2008.

[27] R. A. Horn and C. R. Johnson. *Topics in matrix analysis.* Cambridge University Press, Cambridge, 1994. Corrected reprint of the 1991 original.

[28] B. Jacob, F. L. Schwenninger, and L. A. Vorberg. Remarks on input-to-state stability of collocated systems with saturated feedback. *Math. Control Signals Systems*, 32(3):293–307, 2020.

[29] B. Jacob, C. Tretter, C. Trunk, and H. Vogt. Systems with strong damping and their spectra. *Math. Methods Appl. Sci.*, 41(16):6546–6573, 2018.

[30] B. Jacob, L. Vorberg, and C. Wyss. Computing the quadratic numerical range, 2023. `arXiv:2305.16079`.

[31] C. R. Johnson. Numerical determination of the field of values of a general complex matrix. *SIAM J. Numer. Anal.*, 15(3):595–602, 1978.

[32] T. Kato. *Perturbation theory for linear operators.* Berlin: Springer-Verlag, reprint of the corr. print. of the 2nd ed. 1980 edition, 1995.

[33] V. I. Kostin and S. I. Razzakov. Convergence of the orthogonal-power method of calculation of a spectrum. In *Numerical methods in linear algebra*, volume 6 of *Trudy Inst. Mat.*, pages 55–84, 207. "Nauka" Sibirsk. Otdel., Novosibirsk, 1985.

[34] V. Kostrykin, K. A. Makarov, and A. K. Motovilov. Perturbation of spectra and spectral subspaces. *Trans. Amer. Math. Soc.*, 359(1):77–89, 2007.

[35] H. J. Landau. On Szegö's eigenvalue distribution theorem and non-Hermitian kernels. *J. Analyse Math.*, 28:335–357, 1975.

[36] H. Langer, A. Markus, V. Matsaev, and C. Tretter. A new concept for block operator matrices: the quadratic numerical range. *Linear Algebra Appl.*, 330(1-3):89–112, 2001.

[37] H. Langer and C. Tretter. Spectral decomposition of some nonselfadjoint block operator matrices. *J. Operator Theory*, 39(2):339–359, 1998.

[38] H. Langer and C. Tretter. Diagonalization of certain block operator matrices and applications to Dirac operators. In *Operator theory and analysis (Amsterdam, 1997)*, volume 122 of *Oper. Theory Adv. Appl.*, pages 331–358. Birkhäuser, Basel, 2001.

[39] H. Linden. The quadratic numerical range and the location of zeros of polynomials. *SIAM Journal on Matrix Analysis and Applications*, 25(1):266–284, 2003.

[40] M. Lindner and T. Schmidt. Recycling Givens rotations for the efficient approximation of pseudospectra of band-dominated operators. *Oper. Matrices*, 11(4):1171–1196, 2017.

[41] S. Loisel and P. Maxwell. Path-following method to determine the field of values of a matrix with high accuracy. *SIAM J. Matrix Anal. Appl.*, 39(4):1726–1749, 2018.

[42] M. Marcus and C. Pesce. Computer generated numerical ranges and some resulting theorems. *Linear and Multilinear Algebra*, 20(2):121–157, 1987.

[43] P.-G. Martinsson and J. A. Tropp. Randomized numerical linear algebra: foundations and algorithms. *Acta Numer.*, 29:403–572, 2020.

[44] V. D. Milman and G. Schechtman. *Asymptotic theory of finite-dimensional normed spaces*, volume 1200 of *Lecture Notes in Mathematics.* Springer-Verlag, Berlin, 1986. With an appendix by M. Gromov.

[45] A. Muhammad and M. Marletta. Approximation of the quadratic numerical range of block operator matrices. *Integral Equations Oper. Theory*, 74(2):151–162, 2012.

[46] A. Muhammad and M. Marletta. A numerical investigation of the quadratic numerical range of Hain-Lüst operators. *Int. J. Comput. Math.*, 90(11):2431–2451, 2013.

[47] R. Schnaubelt. *Spectral Theory.* Lecture Notes, 2023. URL: `https://www.math.kit.edu/iana3/~schnaubelt/media/st-skript.pdf` [cited 2023-10-25].

[48] R. E. Showalter. *Hilbert space methods for partial differential equations.* Pitman, London-San Francisco, Calif.-Melbourne, 1977. Monographs and Studies in Mathematics, Vol. 1.

[49] L. N. Trefethen. Approximation theory and numerical linear algebra. In *Algorithms for approximation, II (Shrivenham, 1988)*, pages 336–360. Chapman and Hall, London, 1990.

[50] L. N. Trefethen. Computation of pseudospectra. In *Acta numerica, 1999*, volume 8 of *Acta Numer.*, pages 247–295. Cambridge Univ. Press, Cambridge, 1999.

[51] L. N. Trefethen and M. Embree. *Spectra and pseudospectra.* Princeton University Press, Princeton, NJ, 2005. The behavior of nonnormal matrices and operators.

[52] C. Tretter. *Spectral theory of block operator matrices and applications.* London: Imperial College Press, 2008.

[53] F. Uhlig. Faster and more accurate computation of the field of values boundary for n by n matrices. *Linear and Multilinear Algebra*, 62(5):554–567, 2014.

[54] J. M. Varah. On the separation of two matrices. *SIAM J. Numer. Anal.*, 16(2):216–222, 1979.

[55] J. Weidmann. *Lineare Operatoren in Hilberträumen. Teil I: Grundlagen.* Wiesbaden: B. G. Teubner, 2000.

[56] D. Werner. *Funktionalanalysis*. Springer-Lehrbuch. Springer Berlin Heidelberg, 2011.

[57] M. P. H. Wolff. Discrete approximation of unbounded operators and approximation of their spectra. *J. Approx. Theory*, 113(2):229–244, 2001.