Lecture Notes on Spectra and Pseudospectra of Matrices and Operators

Arne Jensen
Department of Mathematical Sciences
Aalborg University

Abstract

We give a short introduction to the pseudospectra of matrices and operators. We also review a number of results concerning matrices and bounded linear operators on a Hilbert space, and in particular results related to spectra. A few applications of the results are discussed.

Contents

1 Introduction

We give an introduction to the pseudospectra of matrices and operators, and give a few applications. Since these notes are intended for a wide audience, some elementary concepts are reviewed. We also note that one can understand the main points concerning pseudospectra already in the finite dimensional case. So the reader not familiar with operators on a separable Hilbert space can assume that the space is finite dimensional.

Let us briefly outline the contents of these lecture notes. In Section 2 we recall some results from linear algebra, mainly to fix notation, and to recall some results that may not be included in standard courses on linear algebra. In Section 4 we state some results from the theory of bounded operators on a Hilbert space. We have decided to limit the exposition to the case of bounded operators. If some readers are unfamiliar with these results, they can always assume that the Hilbert space is finite dimensional. In Section 5 we finally define the pseudospectra and give a number of results concerning equivalent definitions and simple properties. Section 6 is devoted to some simple examples of pseudospectra. Section 7 contains a few results on perturbation theory for eigenvalues. We also give an application to the location of pseudospectra. In Section 8 we give some examples of applications to continuous time linear systems, and in Section 9 we give some applications to linear discrete time systems. Section 10 contains further matrix examples.

The general reference to results on spectra and pseudospectra is the book [TE05]. There are also many results on pseudospectra in the book [Dav07].

A number of exercises have been included in the text. The reader should try to solve these. The reader should also experiment on the computer using either Maple or MATLAB, or preferably both.

2 Results from linear algebra

In this section we recall some results from linear algebra that are needed later on. We assume that the readers can find most of the results in their own textbooks on linear algebra. For some of the less familiar results we provide references. My own favorite books dealing with linear algebra are [Str06] and [Kat95, Chapters I and II]. The first book is elementary, whereas the second book is a research monograph. It contains in the first two chapters a complete treatment of the eigenvalue problem and perturbation of eigenvalues, in the finite dimensional case, and is the definitive reference for these results.

We should note that Section 4 also contains a number of definitions and results that

are important for matrices. The results in this section are mainly those that do not generalize in an easy manner to infinite dimensions.

To unify the notation we denote a finite dimensional vector space over the complex numbers by $\mathcal{H}$. Usually we identify it with a coordinate space $\mathbf{C}^n$. The linear operators on $\mathcal{H}$ are denoted by $\mathcal{B}(\mathcal{H})$ and are usually identified with the $n \times n$ matrices over $\mathbf{C}$. We deal exclusively with vector spaces over the complex numbers, since we are interested in spectral theory.

The spectrum of a linear operator $A \in \mathcal{B}(\mathcal{H})$ is denoted by $\sigma(A)$, and consists of the eigenvalues of A. The eigenvalues are the roots of the characteristic polynomial $p(\lambda) = \det(A - \lambda I)$. Here I denotes the identity operator. Assume $\lambda_0 \in \sigma(A)$. The multiplicity of λ_0 as a root of $p(\lambda)$ is called the *algebraic* multiplicity of λ_0, and is denoted by $m_a(\lambda_0)$. The dimension of the eigenspace

$$m_g(\lambda_0) = \dim\{u \in \mathcal{H} \mid Au = \lambda_0 u\} \tag{2.1}$$

is called the *geometric* multiplicity of λ_0. We have $m_g(\lambda_0) \leq m_a(\lambda_0)$ for each eigenvalue.

We recall the following definition and theorem. We state the result in the matrix case.

Definition 2.1. *Let A be a complex $n \times n$ matrix. A is said to be* diagonalizable*, if there exist a diagonal matrix D and an invertible matrix V such that*

$$A = VDV^{-1}. \tag{2.2}$$

The columns in V are eigenvectors of A. The following result states that a matrix is diagonalizable, if and only if it has 'enough' linearly independent eigenvectors.

Theorem 2.2. *Let A be a complex $n \times n$ matrix. Let $\sigma(A) = \{\lambda_1, \lambda_2, \ldots, \lambda_m\}$, $\lambda_i \neq \lambda_j$, $i \neq j$. A is diagonalizable, if and only if $m_g(\lambda_1) + \ldots + m_g(\lambda_m) = n$.*

As a consequence of this result, A is diagonalizable, if and only if we have $m_g(\lambda_j) = m_a(\lambda_j)$ for $j = 1, 2, \ldots, m$. Conversely, if there exists a j such that $m_g(\lambda_j) < m_a(\lambda_j)$, then A is *not diagonalizable*.

Not all linear operators on a finite dimensional vector space are diagonalizable. For example the matrix

$$N = \begin{bmatrix} 0 & 1 \\ 0 & 0 \end{bmatrix}$$

has zero as the only eigenvalue, with $m_a(0) = 2$ and $m_g(0) = 1$. This matrix is nilpotent, with $N^2 = 0$.

A general result states that all non-diagonalizable operators on a finite dimensional vector space have a nontrivial nilpotent component. This is the so-called *Jordan canonical form* of $A \in \mathcal{B}(\mathcal{H})$. We recall the result, using the operator language. A proof can be found in [Kat95, Chapter I §5]. It is based on complex analysis and reduces the problem to partial fraction decomposition. An elementary linear algebra based proof can be found in [Str06, Appendix B].

Let $A \in \mathcal{B}(\mathcal{H})$, with $\sigma(A) = \{\lambda_1, \lambda_2, \ldots, \lambda_m\}$, $\lambda_i \neq \lambda_j$, $i \neq j$. The resolvent is given by

$$R_A(z) = (A - zI)^{-1}, \quad z \in \mathbf{C} \setminus \sigma(A). \tag{2.3}$$

Let λ_k be one of the eigenvalues, and let Γ_k denote a small circle enclosing λ_k, and the other eigenvalues lying outside this circle. The Riesz projection for this eigenvalue is given by

$$P_k = -\frac{1}{2\pi i} \int_{\Gamma_k} R_A(z) dz. \tag{2.4}$$

These projections have the following properties for $k, l = 1, 2, \ldots, m$.

$$P_k P_l = \delta_{kl} P_k, \quad \sum_{k=1}^{m} P_k = I, \quad P_k A = A P_k. \tag{2.5}$$

Here δ_{kl} denotes the Kronecker delta, viz.

$$\delta_{kl} = \begin{cases} 1 & \text{if } k = l, \\ 0 & \text{if } k \neq l. \end{cases}$$

We have $m_a(\lambda_k) = \operatorname{rank} P_k$. One can show that $AP_k = \lambda_k P_k + N_k$, where N_k is nilpotent, with $N_k^{m_a(\lambda_k)} = 0$. Define

$$S = \sum_{k=1}^{m} \lambda_k P_k, \quad N = \sum_{k=1}^{m} N_k.$$

Theorem 2.3 (Jordan canonical form)**.** *Let S and N be the operators defined above. Then S is diagonalizable and N is nilpotent. They satisfy $SN = NS$. We have*

$$A = S + N. \tag{2.6}$$

If S' is diagonalizable, N' nilpotent, $S'N' = N'S'$, and $A = S' + N'$, then $S' = S$ and $N' = N$, i.e. uniqueness holds.

The matrix version of this result will be presented and discussed in Section 3.

The definition of the pseudospectrum to be given below depends on the choice of a norm on $\mathcal{H}$. Let $\mathcal{H} = \mathbf{C}^n$. One family of norms often used are the p-norms. They are given by

$$\|u\|_p = \Big(\sum_{k=1}^{n} |u_k|^p \Big)^{1/p}, \quad 1 \leq p < \infty, \tag{2.7}$$

$$\|u\|_\infty = \max_{1 \leq k \leq n} |u_k|. \tag{2.8}$$

The $\|u\|_2$ is the only norm in the family coming from an inner product, and is the usual Euclidean norm. These norms are equivalent in the sense that they give the same topology on $\mathcal{H}$. Equivalence of the norms $\|\cdot\|$ and $\|\cdot\|'$ means that there exist constants c and C, such that

$$c\|u\| \leq \|u\|' \leq C\|u\| \quad \text{for all } u \in \mathcal{H}.$$

These constants usually depend on the dimension of $\mathcal{H}$.

Exercise 2.4. Find constants that show that the three norms $\|\cdot\|_1$, $\|\cdot\|_2$ and $\|\cdot\|_\infty$ on $\mathbf{C}^n$ are equivalent. How do they depend on the dimension?

We will now assume that $\mathcal{H}$ is equipped with an inner product, denoted by $\langle\cdot,\cdot\rangle$. Usually we identify with $\mathbf{C}^n$, and take

$$\langle u, v\rangle = \sum_{k=1}^{n} \overline{u_k} v_k.$$

Note that our inner product is linear in the *second* variable. We assume that the reader is familiar with the concepts of orthogonality and orthonormal bases. We also assume that the reader is familiar with orthogonal projections.

Convention. In the sequel we will assume that the norm $\|\cdot\|$ is the one coming from this inner product, i.e.

$$\|u\| = \|u\|_2 = \sqrt{\langle u, u\rangle}.$$

Given the inner product, the adjoint to $A \in \mathcal{B}(\mathcal{H})$ is the unique linear operator A^* satisfying $\langle u, Av\rangle = \langle A^*u, v\rangle$ for all $u, v \in \mathcal{H}$. We can now state the spectral theorem.

Definition 2.5. *An operator A on an inner product space $\mathcal{H}$ is said to be normal, if $A^*A = AA^*$. An operator with $A = A^*$ is called a self-adjoint operator.*

Theorem 2.6 (Spectral Theorem)**.** *Assume that A is normal. We write $\sigma(A) = \{\lambda_1, \lambda_2, \ldots, \lambda_m\}$, $\lambda_i \neq \lambda_j$, $i \neq j$. Then there exist orthogonal projections P_k, $k = 1, 2, \ldots, m$, satisfying*

$$P_k P_l = \delta_{kl} P_k, \quad \sum_{k=1}^{m} P_k = I, \quad P_k A = A P_k,$$

such that

$$A = \sum_{k=1}^{m} \lambda_k P_k.$$

Comparing the spectral theorem and the Jordan canonical form, then we see that for a normal operator the nilpotent part is identically zero, and that the projections can be chosen to be orthogonal.

The spectral theorem is often stated as the existence of a unitary transform U diagonalizing a matrix A. If $A = UDU^{-1}$, then the columns in U constitute an orthonormal basis for $\mathcal{H}$ consisting of eigenvectors for A. Further results concerning such similarity transforms will be found in Section 3.

When $\mathcal{H}$ is an inner product space, we can define the singular values of A.

Definition 2.7. *Let $A \in \mathcal{B}(\mathcal{H})$. The singular values of A are the (non-negative) square roots of the eigenvalues of A^*A.*

The operator norm is given by $\|A\| = \sup_{\|u\|=1}\|Au\|$. We have that $\|A\| = s_{\max}(A)$, the largest singular value of A. This follows from the fact that $\|A^*A\| = \|A\|^2$ and the spectral theorem. If A is invertible, then $\|A^{-1}\| = (s_{\min}(A))^{-1}$. Here $s_{\min}(A)$ denotes the smallest singular value of A.

Exercise 2.8. Prove the statements above concerning the connections between operator norms and singular values.

The condition number of an invertible matrix is defined as

$$\operatorname{cond}(A) = \|A\| \cdot \|A^{-1}\|. \tag{2.9}$$

It follows that

$$\operatorname{cond}(A) = \frac{s_{\max}(A)}{s_{\min}(A)}.$$

The singular values give techniques for computing norm and condition number numerically, since eigenvalues of self-adjoint matrices can be computed efficiently and numerically stably, usually by iteration methods.

In practical computations a number of different norms on matrices are used. Thus when computing the norm of a matrix in for example MATLAB or Maple, one should be careful to get the right norm. In particular, one should remember that the default call of norm in MATLAB gives the operator norm in the $\|\cdot\|_2$-sense, whereas in Maple it gives the operator norm in the $\|\cdot\|_\infty$-sense.

Let us briefly recall the terminology used in MATLAB. Let $X = [x_{kl}]$ be an $n \times n$ matrix. The command `norm`(X) computes the largest singular value of X and is thus equal to the operator norm of X (with the norm $\|\cdot\|_2$). We have

$$\texttt{norm}(X,1) = \max\{\sum_{k=1}^{n} |x_{kl}| \mid l = 1,\dots,n\},$$

and

$$\texttt{norm}(X,\texttt{inf}) = \max\{\sum_{l=1}^{n} |x_{kl}| \mid k = 1,\dots,n\}.$$

Note the interchange of the role of rows and columns in the two definitions. One should note that `norm`$(X,1)$ is the operator norm, if $\mathbf{C}^n$ is equipped with $\|\cdot\|_1$, and `norm`$(X,$`inf`$)$ is the operator norm, if $\mathbf{C}^n$ is equipped with $\|\cdot\|_\infty$. Thus for consistency one can also use the call `norm`$(X,2)$ to compute `norm`(X).

Finally there is the Frobenius norm. It is defined as

$$\texttt{norm}(X,\texttt{'fro'}) = \sqrt{\sum_{k=1}^{n}\sum_{l=1}^{n} |x_{kl}|^2}.$$

Thus this is the $\|\cdot\|_2$ norm of X considered as a vector in $\mathbf{C}^{n^2}$.

The same norms can be computed in Maple using the command `Norm` from the `LinearAlgebra` package, see the help pages in Maple, and remember that the default is different from the one in MATLAB, as mentioned above.

3 Some matrix results. Similarity transforms

In this section we supplement the discussion in the previous section, focusing on an $n \times n$ matrix A with complex entries. The following concept is important.

Definition 3.1. *Let A, B, and S be $n \times n$ matrices. Assume that S is invertible. If $B = S^{-1}AS$, then the matrices A and B are said to be similar. S is called a similarity transform.*

Note that without some kind of normalization a similarity transform is never unique. If S is a similarity transform implementing the similarity $B = S^{-1}AS$, then cS for any $c \in \mathbf{C}$, $c \neq 0$, is also a similarity transform implementing the same similarity.

Assume that λ is an eigenvalue of A with an eigenvector v, then λ is an eigenvalue of B, and $S^{-1}v$ a corresponding eigenvector. Thus the two matrices A and B have the same eigenvalues with the same geometric multiplicities.

Thus if A is a linear operator on a finite dimensional vector space $\mathcal{H}$, and we fix a basis in $\mathcal{H}$, we get a matrix A representing this linear operator. Since one basis is mapped onto another basis by an invertible matrix S, any two matrix representations of an operator are similar. The point of these observations is that the eigenvalues of A are independent of the choice of basis and hence matrix representation, but the eigenvectors are *not independent* of the choice of basis.

If A is normal, then there exists an orthonormal basis consisting of eigenvectors. If we take U to be the matrix whose columns are these eigenvectors, then this matrix is unitary. If A is any matrix representation of A, then $\Lambda = U^*AU$ is a diagonal matrix with the eigenvalues on the diagonal. This is often the form in which the spectral theorem (Theorem 2.6) is given in elementary linear algebra texts.

Let us see what happens, if a matrix A is diagonalizable, but not normal. Then we can find an invertible matrix V, such that

$$\Lambda = V^{-1}AV, \tag{3.1}$$

and the columns still consist of eigenvectors of A, see also Theorem 2.2. Now since A is not normal, the eigenvectors of the matrix A may be a very ill conditioned basis of $\mathcal{H}$, whereas the eigenvectors of the matrix Λ form an orthonormal basis, viz. the canonical basis in $\mathbf{C}^n$. The kind of problem that is encountered can be understood by computing the condition number $\mathrm{cond}(V)$.

Let us now give an example, using the Toeplitz matrix from Section 10.1. We recall a few details here, for the reader's convenience. A is the $n \times n$ Toeplitz matrix with the

following structure.

$$A = \begin{bmatrix} 0 & 1 & 0 & \cdots & 0 & 0 \\ \frac{1}{4} & 0 & 1 & \cdots & 0 & 0 \\ 0 & \frac{1}{4} & 0 & \cdots & 0 & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & 0 & 0 & \cdots & 0 & 1 \\ 0 & 0 & 0 & \cdots & \frac{1}{4} & 0 \end{bmatrix}. \tag{3.2}$$

Let Q denote the diagonal $n \times n$ matrix with entries $2, 4, 8, \ldots, 2^n$ on the diagonal. Then one can verify that

$$QAQ^{-1} = B, \tag{3.3}$$

where

$$B = \begin{bmatrix} 0 & \frac{1}{2} & 0 & \cdots & 0 & 0 \\ \frac{1}{2} & 0 & \frac{1}{2} & \cdots & 0 & 0 \\ 0 & \frac{1}{2} & 0 & \cdots & 0 & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & 0 & 0 & \cdots & 0 & \frac{1}{2} \\ 0 & 0 & 0 & \cdots & \frac{1}{2} & 0 \end{bmatrix}. \tag{3.4}$$

The matrix B is symmetric, and its eigenvalues can be found to be

$$\lambda_k = \cos\left(\frac{k\pi}{n+1}\right), \quad k = 1, \ldots, n. \tag{3.5}$$

Thus this matrix can be diagonalized using a unitary matrix U. Therefore the original matrix A is diagonalized by $V = Q^{-1}U$, using the conventions in (3.1). Since multiplication by a unitary matrix leaves the condition number unchanged, we have $\mathrm{cond}(V) = \mathrm{cond}(Q)$. The condition number of Q given above is $\mathrm{cond}(Q) = 2^{n-1}$. Thus for $n = 25$ the condition number $\mathrm{cond}(V)$ is approximately $1.6777\,10^7$, for $n = 50$ it is $5.6295\,10^{14}$, and for $n = 100$ it is $6.3383\,10^{29}$. From the explicit expression it is clear that it grows exponentially with n.

Exercise 3.2. Verify all the statements above concerning the matrix A given in (3.2). Try to find the diagonalizing matrix V by direct numerical computation, compute its condition number, and compare with the exact values given above, for $n = 25, 50, 100$. What are your conclusions?

Let v_j denote the j^{th} eigenvector of A. Then $e_j = V^{-1}v_j$ is just the j^{th} canonical basis vector in $\mathbf{C}^n$, i.e. the vector with a one in entry j and all other entries equal to zero. A consequence of the large condition number of the matrix V is reflected in the fact that the basis consisting of the v_j vectors is a poor basis for $\mathbf{C}^n$.

Exercise 3.3. Verify the above statement by plotting the 25 eigenvectors. You can use either Maple or MATLAB. Note that all vectors are large for small indices and very small for large indices.

Now let us recall one of the important results, which is valid for all matrices. It is what is usually called Schur's Lemma.

Theorem 3.4 (Schur's Lemma)**.** *Let A be an $n \times n$ matrix. Then there exists a unitary matrix U such that $U^{-1}AU = A_{\text{upper}}$, where A_{upper} is an upper triangular matrix.*

We return to the Jordan canonical form given in Theorem 2.3. We present the matrix form of this result. Given an arbitrary $n \times n$ matrix A, there exist an invertible matrix V and a matrix J with a particular structure, such that

$$J = V^{-1}AV. \tag{3.6}$$

Let us describe the structure of V and J in some detail. Assume that λ_j is an eigenvalue of A. Recall that $m_a(\lambda_j)$ denotes the algebraic multiplicity of the eigenvalue, and $m_g(\lambda_j)$ denotes its geometric multiplicity, i.e. the number of linearly independent eigenvectors. Then there exist an $n \times m_a(\lambda_j)$ matrix V_j and an $m_a(\lambda_j) \times m_a(\lambda_j)$ matrix J_j, such that

$$AV_j = V_j J_j. \tag{3.7}$$

The matrix V_j has linearly independent columns, and the matrix J_j is a block diagonal matrix, i.e. $J_j = \text{diag}(J_{j,1}, \ldots, J_{j,m_g(\lambda_j)})$. Each block has the structure

$$J_{j,\ell} = \begin{bmatrix} \lambda_j & 1 & 0 & \cdots & 0 & 0 \\ 0 & \lambda_j & 1 & \cdots & 0 & 0 \\ 0 & 0 & \lambda_j & \cdots & 0 & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & 0 & 0 & \cdots & \lambda_j & 1 \\ 0 & 0 & 0 & \cdots & 0 & \lambda_j \end{bmatrix}, \quad \ell = 1, 2, \ldots, m_g(\lambda_j). \tag{3.8}$$

The number of rows and columns in each block depends on the particular matrix A. The sum of the row dimensions (and column dimensions) must equal $m_a(\lambda_j)$ in order to get a matrix J_j as described above. Since we have $m_g(\lambda_j)$ blocks, the total number of ones above the diagonal is exactly $m_a(\lambda_j) - m_g(\lambda_j)$. The columns of V_j consist of what is sometimes called generalized eigenvectors of A corresponding to the eigenvalue λ_j. This means that the subspace spanned by the columns of V_j, denoted by $\mathcal{V}_j$, can be described as

$$\mathcal{V}_j = \{v \mid (A - \lambda_j I)^k v = 0 \quad \text{for some k}\}. \tag{3.9}$$

Now the Jordan form (3.6) follows by forming the matrix as the columns in V_1, followed by the columns in V_2 and so on. The matrix J has the block diagonal structure $J = \text{diag}(J_1, \ldots, J_m)$, where m is the number of distinct eigenvalues of A.

A few examples may clarify the above definitions. Consider first the matrix with just one eigenvalue.

$$J = \begin{bmatrix} 3 & 0 & 0 & 0 \\ 0 & 3 & 0 & 0 \\ 0 & 0 & 3 & 1 \\ 0 & 0 & 0 & 3 \end{bmatrix}.$$

For this particular matrix $m_a(3) = 4$ and $m_g(3) = 3$. We have $J = J_1$ and $J_1 = \text{diag}(J_{1,1}, J_1, 2, J_{1,3})$, where

$$J_{1,1} = \begin{bmatrix} 3 \end{bmatrix}, \quad J_{1,2} = \begin{bmatrix} 3 \end{bmatrix}, \quad \text{and} \quad J_{1,3} = \begin{bmatrix} 3 & 1 \\ 0 & 3 \end{bmatrix}.$$

As another example we take the Jordan matrix

$$J = \begin{bmatrix} 2 & 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 2 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 4 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 4 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 4 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 6 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 6 \end{bmatrix}.$$

This matrix has the eigenvalues $2, 4, 6$. Eigenvalue 2 has algebraic multiplicity 2 and geometric multiplicity 1. Eigenvalue 4 has algebraic multiplicity 3 and geometric multiplicity 2. For eigenvalue 6 the algebraic and geometric multiplicities are both 2.

We have in this case $J = \text{diag}(J_1, J_2, J_3)$, where

$$J_1 = \begin{bmatrix} 2 & 1 \\ 0 & 2 \end{bmatrix}, \quad \begin{bmatrix} 4 & 0 & 0 \\ 0 & 4 & 1 \\ 0 & 0 & 4 \end{bmatrix}, \quad \text{and} \quad J_3 = \begin{bmatrix} 6 & 0 \\ 0 & 6 \end{bmatrix}.$$

We have $J_1 = J_{1,1}$, $J_2 = \text{diag}(J_{2,1}, J_{2,2})$ and $J_3 = \text{diag}(J_{3,1}, J_{3,2})$, where

$$J_{1,1} = \begin{bmatrix} 2 & 1 \\ 0 & 2 \end{bmatrix}, \quad J_{2,1} = \begin{bmatrix} 4 \end{bmatrix}, \quad J_{2,2} = \begin{bmatrix} 4 & 1 \\ 0 & 4 \end{bmatrix}, \quad J_{3,1} = \begin{bmatrix} 6 \end{bmatrix}, \quad \text{and} \quad J_{3,2} = \begin{bmatrix} 6 \end{bmatrix}.$$

Comparing the Jordan form and the result from Schur's Lemma (Theorem 3.4) we see that we can get a transformation of a given matrix A into an upper triangular matrix using a unitary transform (which of course has condition number 1), and we can also get a transformation into the canonical Jordan form, where the transformed matrix is sparse (at most bidiagonal) and highly structured. But the transformation matrix may have a very large condition number, as shown by the example above.

4 Results from operator theory

In this section we state some results from operator theory. We have decided not to discuss unbounded operators, and we have also decided to focus on Hilbert spaces. Most of the results on pseudospectra are valid for unbounded operators on Hilbert and Banach spaces. Even if you main interest is the finite dimensional results, you will need

the concepts and definitions from this section to read the following section. In reading it you can safely assume that all Hilbert spaces are finite dimensional.

Let $\mathcal{H}$ be a Hilbert space (always with the complex numbers as the scalars). The inner product is denoted by $\langle \cdot, \cdot \rangle$, and the norm by $\|u\| = \sqrt{\langle u, u \rangle}$. As in the finite dimensional case our inner product is linear in the *second* variable.

We will not review the concepts of orthogonality and orthonormal basis. Neither will we review the Riesz representation theorem, nor the properties of orthogonal projections. We refer the reader to any of the numerous introductions to functional analysis. Our own favorite is [RS80], and we will sometimes refer to it for results we need. Another favorite is [Kat95].

We denote the bounded operators on a Hilbert space $\mathcal{H}$ by $\mathcal{B}(\mathcal{H})$, as in the finite dimensional case. This space is a Banach space, equipped with the operator norm $\|A\| = \sup_{\|u\|=1} \|Au\|$. The adjoint of $A \in \mathcal{B}(\mathcal{H})$ is the unique bounded operator A^* satisfying $\langle v, Au \rangle = \langle A^* v, u \rangle$. We have $\|A^*\| = \|A\|$ and $\|A^* A\| = \|A\|^2$.

We recall that the spectrum $\sigma(A)$ consists of those $z \in \mathbf{C}$, for which $A - zI$ has no bounded inverse. The spectrum of an operator $A \in \mathcal{B}(\mathcal{H})$ is always non-empty. The resolvent

$$R_A(z) = (A - zI)^{-1}, \quad z \notin \sigma(A),$$

is an analytic function with values in $\mathcal{B}(\mathcal{H})$. The spectrum of $A \in \mathcal{B}(\mathcal{H})$ is a compact subset of the complex plane, which means that it is bounded and closed. For future reference, we recall that $\Omega \subseteq \mathbf{C}$ is compact, if and only if it is bounded and closed. That Ω is bounded means there is an $R > 0$, such that $\Omega \subseteq \{z \mid |z| \leq R\}$. That Ω is closed means that for any convergent sequence $z_n \in \Omega$ we have $\lim_{n\to\infty} z_n \in \Omega$. There are two very simple results on the resolvent that are important.

Proposition 4.1 (First Resolvent Equation)**.** *Let $A \in \mathcal{B}(\mathcal{H})$ and let $z_1, z_2 \notin \sigma(A)$. Then*

$$R_A(z_2) - R_A(z_1) = (z_2 - z_1) R_A(z_1) R_A(z_2) = (z_2 - z_1) R_A(z_2) R_A(z_1).$$

Exercise 4.2. Prove this result.

Proposition 4.3 (Second Resolvent Equation)**.** *Let $A, B \in \mathcal{B}(\mathcal{H})$, and let $C = B - A$. Assume that $z \notin \sigma(A) \cup \sigma(B)$. Then we have*

$$R_B(z) - R_A(z) = -R_A(z) C R_B(z) = -R_B(z) C R_A(z).$$

If $I + R_A(z)C$ is invertible, then we have

$$R_B(z) = (I + R_A(z)C)^{-1} R_A(z).$$

Exercise 4.4. Prove this result.

We now recall the definition of the spectral radius.

Definition 4.5. *Let $A \in \mathcal{B}(\mathcal{H})$. The* spectral radius *of A is defined by*

$$\rho(A) = \sup\{|z| \mid z \in \sigma(A)\}.$$

Theorem 4.6. *Let $A \in \mathcal{B}(\mathcal{H})$. Then*

$$\rho(A) = \lim_{n\to\infty} \|A^n\|^{1/n} = \inf_{n\geq 1} \|A^n\|^{1/n}.$$

For all A we have that $\rho(A) \leq \|A\|$. If A is normal, then $\rho(A) = \|A\|$.

Proof. See for example [RS80, Theorem VI.6]. □

We also need the *numerical range* of a linear operator. This is usually not a topic in introductory courses on operator theory, but it plays an important role later. The numerical range of A is sometimes called the *field of values* of A.

Definition 4.7. *Let $A \in \mathcal{B}(\mathcal{H})$. The numerical range of A is the set*

$$W(A) = \{\langle u, Au\rangle \mid \|u\| = 1\}. \tag{4.1}$$

Note that the condition in the definition is $\|u\| = 1$ and not $\|u\| \leq 1$.

Theorem 4.8 (Toeplitz-Hausdorff)**.** *The numerical range $W(A)$ is always a convex set. If $\mathcal{H}$ is finite dimensional, then $W(A)$ is a compact set.*

Proof. The convexity is non-trivial to prove. See for example [Kat95]. Assume $\mathcal{H}$ finite dimensional. Since $u \mapsto \langle u, Au\rangle$ is continuous and $\{u \in \mathcal{H} \mid \|u\| = 1\}$ is compact in this case, the compactness of $W(A)$ follows. □

Exercise 4.9. Let $\mathcal{H} = \mathbf{C}^2$ and let A be a 2×2 matrix. Show that $W(A)$ is the union of an ellipse and its interior (including the degenerate case, when it is a line segment or a point).

Comment: This exercise is elementary in the sense that it requires only the definitions and analytic geometry in the plane, but it is not easy. One strategy is to separate into the cases

(i) A has one eigenvalue,

and

(ii) A has two different eigenvalues.

In case (i) one can reduce to a matrix

$$\begin{bmatrix} 0 & \alpha \\ 0 & 0 \end{bmatrix},$$

and in case (ii) to a matrix

$$\begin{bmatrix} 1 & \alpha \\ 0 & 0 \end{bmatrix}.$$

Here $\alpha \in \mathbf{C}$. The reduction is by translation and scaling. Even with this reduction the case (ii) is not easy.

In analogy with the spectral radius we define the numerical radius as follows.

Definition 4.10. *Let $A \in \mathcal{B}(\mathcal{H})$. The numerical radius of A is given by*

$$\mu(A) = \sup\{|z| \mid z \in W(A)\}.$$

If $\Omega \subset \mathbf{C}$ is a subset of the complex plane, then we denote the closure of this set by $\mathrm{cl}(\Omega)$. We recall that $z \in \mathrm{cl}(\Omega)$, if and only if there is a convergent sequence $z_n \in \Omega$, such that $z = \lim_{n\to\infty} z_n$.

Proposition 4.11. *Let $A \in \mathcal{B}(\mathcal{H})$. Then $\sigma(A) \subseteq \mathrm{cl}(W(A))$.*

Proof. We refer to for example [Kat95] for the proof. □

Let us note that in the finite dimensional case we have $\sigma(A) \subseteq W(A)$, since $W(A)$ is closed. Since $W(A)$ is convex, we have $\mathrm{conv}(\sigma(A)) \subseteq W(A)$. Here $\mathrm{conv}(\Omega)$ denotes the smallest closed convex set in the plane containing $\Omega \subset \mathbf{C}$. It is called the *convex hull* of Ω.

We note the following general result:

Proposition 4.12. *Let $A \in \mathcal{B}(\mathcal{H})$. If A is normal, then $W(A) = \mathrm{conv}(\sigma(A))$.*

Proof. We refer to for example [Kat95] for the proof. □

There is a result on the numerical range which shows that in the infinite dimensional case the numerical range behaves nicely under approximation.

Theorem 4.13. *Let $\mathcal{H}$ be an infinite dimensional Hilbert space, and let $A \in \mathcal{B}(\mathcal{H})$ be a bounded operator. Let $\mathcal{H}_n$, $n = 1, 2, 3, \ldots$ be a sequence of closed subspaces of $\mathcal{H}$, such that $\mathcal{H}_n \subsetneqq \mathcal{H}_{n+1}$, and such that $\bigcup_{n=1}^{\infty} \mathcal{H}_n$ is dense in cH. Let P_n denote the orthogonal projection onto $\mathcal{H}_n$, and let $A_n = P_n A P_n$, considered as an operator on $\mathcal{H}_n$, i.e. the restriction of the operator A to the space $\mathcal{H}_n$. Then we have the following results.*

(i) *For $n = 1, 2, 3, \ldots$ we have $\sigma(A_n) \subseteq cl(W(A_n)) \subseteq cl(W(A))$.*

(ii) *For $n = 1, 2, 3, \ldots$ we have $cl(W(A_n)) \subseteq cl(W(A_{n+1}))$.*

(iii) *We have $cl(W(A)) = cl(\bigcup_{n=1}^{\infty} W(A_n))$.*

Proof. The first inclusion in (i) is a restatement of Proposition 4.11. The second inclusion follows from

$$W(A_n) = \{\langle u, Au\rangle \mid u \in \mathcal{H}_n,\ \|u\| = 1\} \subseteq \{\langle u, Au\rangle \mid u \in \mathcal{H},\ \|u\| = 1\} = W(A)$$

by taking closure. The result (ii) is proved in the same way. Concerning the result (iii), then we note that since $\bigcup_{n=1}^{\infty} \mathcal{H}_n$ is dense in $\mathcal{H}$, we have $u = \lim_{n\to\infty} P_n u$ for all $u \in \mathcal{H}$. Thus we can use

$$\lim_{n\to\infty} \frac{\langle P_n u, AP_n u\rangle}{\|P_n u\|^2} = \frac{\langle u, Au\rangle}{\|u\|^2}$$

to get the result (iii). □

A typical application of this result is to numerically find a good approximation to the numerical range of an operator on an infinite dimensional Hilbert space, by taking as the sequence $\mathcal{H}_n$ a sequence of finite dimensional subspaces.

We have decided not to state the spectral theorem for bounded normal operators in an infinite dimensional Hilbert space. The definition of a normal operator is still that $A^*A = AA^*$. See textbooks on operator theory and functional analysis.

We need to have a general functional calculus available. We will briefly introduce the *Dunford calculus*. This calculus is also called the holomorfic functional calculus, see [Dav07, page 27]. Let $A \in \mathcal{B}(\mathcal{H})$ and let $\Omega \subseteq \mathbf{C}$ be a connected open set, such that $\sigma(A) \subset \Omega$. Let $f\colon \Omega \to \mathbf{C}$ be a holomorphic function. Let Γ be a simple closed contour in Ω containing $\sigma(A)$ in its interior. Then we define

$$f(A) = \frac{-1}{2\pi i} \int_\Gamma f(z) R_A(z) dz. \tag{4.2}$$

(We freely use the Riemann integral of continuous functions with values in a Banach space.)

It is possible to generalize by allowing sets Ω that are not connected and closed contours with several components, but we do not assume that the reader is familiar with this aspect of complex analysis. Thus we will only consider connected sets Ω and simple closed contours in the definition of the Dunford calculus.

The functional calculus name is justified by the properties $(\alpha f + \beta g)(A) = \alpha f(A) + \beta g(A)$ and $(fg)(A) = f(A)g(A)$ for f and g holomorphic functions satisfying the above conditions. Here α and β are complex numbers. We also have $f(A)^* = \overline{f}(A)$.

In some cases there is a different way to define functions of a bounded operator, using a power series. If $A \in \mathcal{B}(\mathcal{H})$, and if f has a power series expansion around zero with radius of convergence $\rho > \rho(A)$, viz.

$$f(z) = \sum_{k=0}^{\infty} c_k z^k, \quad |z| < \rho,$$

(the series is absolutely and uniformly convergent for $|z| \leq \rho' < \rho$), then we can define

$$f(A) = \sum_{k=0}^{\infty} c_k A^k.$$

The series is norm convergent in $\mathcal{B}(\mathcal{H})$. This definition, and the one using the Dunford calculus, give the same $f(A)$, when both are applicable.

Exercise 4.14. Carry out the details in the power series definition.

One often used consequence is the so-called Neumann series (the operator version of the geometric series).

Proposition 4.15. *Let $A \in \mathcal{B}(\mathcal{H})$ with $\|A\| < 1$. Then $I - A$ is invertible and*

$$(I - A)^{-1} = \sum_{k=0}^{\infty} A^k,$$

where the series is norm convergent. We have

$$\|(I - A)^{-1}\| \leq \frac{1}{1 - \|A\|}.$$

Exercise 4.16. Prove this result.

Exercise 4.17. Let $A \in \mathcal{B}(\mathcal{H})$. Use Proposition 4.15 to show that for $|z| > \|A\|$ we have

$$R_A(z) = -\sum_{n=0}^{\infty} z^{-n-1} A^n. \tag{4.3}$$

One consequence of Proposition 4.15 is the stability of invertibility for a bounded operator. We state the result as follows.

Proposition 4.18. *Assume that $A, B \in \mathcal{B}(\mathcal{H})$, such that A is invertible. If $\|B\| < \|A^{-1}\|^{-1}$, then $A + B$ is invertible. We have*

$$\|(A + B)^{-1} - A^{-1}\| \leq \frac{\|B\|\|A^{-1}\|}{1 - \|B\|\|A^{-1}\|}.$$

Proof. Write $A + B = A(I + A^{-1}B)$. The assumption implies $\|A^{-1}B\| < 1$ and the results follow from Proposition 4.15. □

Another function often used in the functional calculus is the exponential function. Since the power series for $\exp(z)$ has infinite radius of convergence, we can define $\exp(A)$ by

$$\exp(A) = \sum_{k=0}^{\infty} \frac{1}{k!} A^k.$$

This definition is valid for all $A \in \mathcal{B}(\mathcal{H})$. If we consider the initial value problem

$$\begin{aligned} \frac{du}{dt}(t) &= Au(t), \\ u(0) &= u_0, \end{aligned}$$

where $u \colon \mathbf{R} \to \mathcal{H}$ is a continuously differentiable function, then the solution is given by

$$u(t) = \exp(tA)u_0.$$

This result is probably familiar in the finite dimensional case, from the theory of linear systems of ordinary differential equations, but it is valid also in this operator theory context.

Exercise 4.19. Prove that for any $A \in \mathcal{B}(\mathcal{H})$ we have

$$\frac{d}{dt} \exp(tA) = A \exp(tA),$$

where the derivative is taken in operator norm sense.

5 Pseudospectra

We now come to the definition of the pseudospectra. We will consider an operator $A \in \mathcal{B}(\mathcal{H})$. Unless stated explicitly, the definitions and results are valid for both the finite dimensional and the infinite dimensional Hilbert spaces $\mathcal{H}$. As mentioned in the introduction, most definitions and results are also valid for closed operators on a Banach space.

For a normal operator on a finite dimensional $\mathcal{H}$ we have the spectral theorem as stated in Theorem 2.6, and in this case the eigenvalues and associated eigenprojections give a valid 'picture' of the operator. But for non-normal operators this is not the case.

Let us look at the simple problem of solving an operator equation $Au - zu = v$, where we assume that $z \notin \sigma(A)$. We want solutions that are stable under small perturbations of the right hand side v and/or the operator A. Consider first $Au' - zu' = v'$ with $\|v - v'\| < \varepsilon$. Then $\|u - u'\| < \varepsilon \|(A - zI)^{-1}\|$. Now the point is that the norm of the resolvent $\|(A - zI)^{-1}\|$ can be large, even when z in not very close to the spectrum $\sigma(A)$. Thus what we need is that ε is sufficiently small, compared to $\|(A - zI)^{-1}\|$.

Consider next a small perturbation of A. Let $B \in \mathcal{B}(\mathcal{H})$ with $\|B\| < \varepsilon$. We compare the solutions to $Au - zu = v$ and $(A + B)u' - zu' = v$. We have

$$u - u' = \big((A - zI)^{-1} - (A + B - zI)^{-1}\big)v.$$

Using the second resolvent equation (see Proposition 4.3), we can rewrite this expression as

$$u - u' = (A - zI)^{-1}B\big(I + (A - zI)^{-1}B\big)^{-1}(A - zI)^{-1}v,$$

provided $\|(A - zI)^{-1}B\| \leq \varepsilon\|(A - zI)^{-1}\| < 1$. Using the Neumann series (see Proposition 4.18) we get the estimate

$$\|u - u'\| \leq \frac{\varepsilon\|(A - zI)^{-1}\|}{1 - \varepsilon\|(A - zI)^{-1}\|}\|(A - zI)^{-1}\|\|v\|.$$

Thus again a good estimate requires that $\varepsilon\|(A - zI)^{-1}\|$ is small.

We will now simplify our notation by using the resolvent notation, as in Section 4, i.e. $R_A(z) = (A - zI)^{-1}$.

Definition 5.1. *Let $A \in \mathcal{B}(\mathcal{H})$ and $\varepsilon > 0$. The ε-pseudospectrum of A is given by*

$$\sigma_\varepsilon(A) = \sigma(A) \cup \{z \in \mathbf{C} \setminus \sigma(A) \mid \|R_A(z)\| > \varepsilon^{-1}\}. \tag{5.1}$$

The following theorem gives two important aspects of the pseudospectra. As a consequence of this theorem one can use either condition (ii) or condition (iii) as alternate definitions of the pseudospectrum.

Theorem 5.2. *Let $A \in \mathcal{B}(\mathcal{H})$ and $\varepsilon > 0$. Then the following three statements are equivalent.*

(i) $z \in \sigma_\varepsilon(A)$.

(ii) *There exists* $B \in \mathcal{B}(\mathcal{H})$ *with* $\|B\| < \varepsilon$ *such that* $z \in \sigma(A+B)$.

(iii) $z \in \sigma(A)$ *or there exists* $v \in \mathcal{H}$ *with* $\|v\| = 1$ *such that* $\|(A - zI)v\| < \varepsilon$.

Proof. Let us first show that (i) implies (iii). Assume $z \in \sigma_\varepsilon(A)$ and $z \notin \sigma(A)$. Then we can find $u \in \mathcal{H}$ such that $\|R_A(z)u\| > \varepsilon^{-1}\|u\|$. Let $v = R_A(z)u$. Then $\|(A - zI)v\| < \varepsilon\|v\|$, and (iii) follows by normalizing v.

Next we show that (iii) implies (ii). If $z \in \sigma(A)$, we can take $B = 0$. Thus assume $z \notin \sigma(A)$. Let $v \in \mathcal{H}$ with $\|v\| = 1$ and $\|(A - zI)v\| < \varepsilon$. Define a rank one operator B by

$$Bu = -\langle v, u\rangle(A - zI)v.$$

Then $\|B\| < \varepsilon$, and $(A - zI + B)v = 0$, such that z is an eigenvalue of $A + B$.

Finally let us show that (ii) implies (i). Here we use proof by contradiction. Assume that (ii) holds and furthermore that $z \notin \sigma(A)$ and $\|R_A(z)\| \leq \varepsilon^{-1}$. We have

$$A + B - zI = (I + BR_A(z))(A - zI).$$

Now our assumptions imply that $\|BR_A(z)\| < \varepsilon \cdot \varepsilon^{-1} = 1$, thus $(I + BR_A(z))$ is invertible, see Proposition 4.15. Since $(A - zI)$ is invertible, too, it follows that $A + B - zI$ is invertible, contradicting $z \in \sigma(A + B)$. □

The result (iii) is sometimes formulated using the following terminology.

Definition 5.3. *Let* $A \in \mathcal{B}(\mathcal{H})$, $\varepsilon > 0$, $z \in \mathbf{C}$, *and* $u \in \mathcal{H}$ *with* $\|u\| = 1$. *If* $\|(A - zI)u\| < \varepsilon$, *then* z *is called an* ε*-pseudoeigenvalue for* A *and* u *is called a corresponding* ε*-pseudoeigenvector.*

In the finite dimensional case we have the following result, which follows immediately from the discussion of singular values in Section 2.

Theorem 5.4. *Assume that* $\mathcal{H}$ *is finite dimensional and* $A \in \mathcal{B}(\mathcal{H})$. *Let* $\varepsilon > 0$. *Then* $z \in \sigma_\varepsilon(A)$, *if and only if* $s_{\min}(A - zI) < \varepsilon$.

Since the singular values of a matrix can be computed numerically, this result provides a method for plotting the pseudospectra of a given matrix. One chooses a finite grid of points in the complex plane, and evaluates $s_{\min}(A - zI)$ at each point. Plotting level curves for these points provides a picture of the pseudospectra of A.

Let us now state some simple properties of the pseudospectra. We use the notation

$$D_\delta = \{z \in \mathbf{C} \mid |z| < \delta\}.$$

Proposition 5.5. *Let* $A \in \mathcal{B}(\mathcal{H})$. *Each* $\sigma_\varepsilon(A)$ *is a bounded open subset of* $\mathbf{C}$. *We have* $\sigma_{\varepsilon_1}(A) \subset \sigma_{\varepsilon_2}(A)$ *for* $0 < \varepsilon_1 < \varepsilon_2$. *Furthermore,* $\cap_{\varepsilon>0}\sigma_\varepsilon(A) = \sigma(A)$. *For* $\delta > 0$ *we have* $D_\delta + \sigma_\varepsilon(A) \subseteq \sigma_{\varepsilon+\delta}(A)$.

Proof. The results are easy consequences of the definition and Theorem 5.2. □

Exercise 5.6. Give the details of this proof.

Concerning the relation between the pseudospectra of A and A^* we have the following result. We use the notation $\overline{\Omega} = \{\overline{z} \mid z \in \Omega\}$ for a subset $\Omega \subseteq \mathbf{C}$.

Proposition 5.7. *Let $A \in \mathcal{B}(\mathcal{H})$. Then for $\varepsilon > 0$ we have $\sigma_\varepsilon(A^*) = \overline{\sigma_\varepsilon(A)}$.*

Proof. We recall that $\sigma(A^*) = \overline{\sigma(A)}$. Furthermore, if $z \notin \sigma(A)$, then $\|(A^* - \overline{z}I)^{-1}\| = \|(A - zI)^{-1}\|$. □

We have the following result.

Proposition 5.8. *Let $A \in \mathcal{B}(\mathcal{H})$ and assume that $V \in \mathcal{B}(\mathcal{H})$ is invertible. Let $\kappa = \mathrm{cond}(V)$, see (2.9) for the definition. Let $B = V^{-1}AV$. Then*

$$\sigma(B) = \sigma(A), \tag{5.2}$$

and for $\varepsilon > 0$ we have

$$\sigma_{\varepsilon/\kappa}(A) \subseteq \sigma_\varepsilon(B) \subseteq \sigma_{\kappa\varepsilon}(A). \tag{5.3}$$

Proof. We have $R_B(z) = V^{-1}R_A(z)V$ for $z \notin \sigma(A)$, which implies the first result. Then we get $\|R_B(z)\| \leq \kappa\|R_A(z)\|$ and $\|R_A(z)\| \leq \kappa\|R_B(z)\|$, which imply the second result. □

We give some further results on the location of the pseudospectra. We start with the following general result. Although the result is well known, we include the proof. For a subset $\Omega \subset \mathbf{C}$ we set as usual

$$\mathrm{dist}(z, \Omega) = \inf\{|\zeta - z| \mid \zeta \in \Omega\},$$

and note that if Ω is compact, then the infimum is attained for some point in Ω.

Proposition 5.9. *Let $A \in \mathcal{B}(\mathcal{H})$. Then for $z \notin \sigma(A)$ we have*

$$\|R_A(z)\| \geq \frac{1}{\mathrm{dist}(z, \sigma(A))}. \tag{5.4}$$

If A is normal, then we have

$$\|R_A(z)\| = \frac{1}{\mathrm{dist}(z, \sigma(A))}. \tag{5.5}$$

Proof. Let $z \notin \sigma(A)$ and take $\zeta_0 \in \sigma(A)$ such that $|z - \zeta_0| = \mathrm{dist}(z, \sigma(A))$. Assume $\|R_A(z)\| < (\mathrm{dist}(z, \sigma(A)))^{-1}$. Write $(A - \zeta_0 I) = (A - zI)(I + (z - \zeta_0)R_A(z))$. Due to our assumptions both factors on the right hand side are invertible, leading to a contradiction. This proves the first result. The second result is a consequence of the spectral

theorem. Let us give some details in the case where $\mathcal{H}$ is finite dimensional. The Spectral Theorem, Theorem 2.6, gives for a normal operator A that

$$(A - zI)^{-1} = \sum_{k=1}^{m} \frac{1}{\lambda_k - z} P_k.$$

Assume $u \in \mathcal{H}$ with $\|u\| = 1$. The properties of the spectral projections imply that we have

$$\|(A - zI)^{-1}u\|^2 = \sum_{k=1}^{m} \frac{1}{|\lambda_k - z|^2} \|P_k u\|^2 \leq \max_{k=1\ldots m} \frac{1}{|\lambda_k - z|^2} \sum_{j=1}^{m} \|P_j u\|^2 = \frac{1}{\operatorname{dist}(z, \sigma(A))^2}.$$

This proves the result in the finite dimensional case. □

Corollary 5.10. *Let $A \in \mathcal{B}(\mathcal{H})$ and $\varepsilon > 0$. Then*

$$\{z \mid \operatorname{dist}(z, \sigma(A)) < \varepsilon\} \subseteq \sigma_\varepsilon(A). \tag{5.6}$$

If A is normal, then

$$\sigma_\varepsilon(A) = \{z \mid \operatorname{dist}(z, \sigma(A)) < \varepsilon\}. \tag{5.7}$$

We have the following result, where we get an inclusion in the other direction.

Theorem 5.11 (Bauer–Fike)**.** *Let A be an $N \times N$ matrix, which is diagonalizable, such that $A = V\Lambda V^{-1}$, where Λ is a diagonal matrix. Then for $\varepsilon > 0$ we have*

$$\{z \mid \operatorname{dist}(\sigma(A), z) < \varepsilon\} \subseteq \sigma_\varepsilon(A) \subseteq \{z \mid \operatorname{dist}(\sigma(A), z) < \kappa\varepsilon\}, \tag{5.8}$$

where $\kappa = \operatorname{cond}(V)$.

Proof. The first inclusion is the result (5.6). The second inclusion follows from

$$\|(A - zI)^{-1}\| = \|V(\Lambda - zI)^{-1}V^{-1}\| \leq \kappa \|(\Lambda - zI)^{-1}\| = \frac{\kappa}{\operatorname{dist}(\sigma(A), z)},$$

since the diagonal matrix Λ is normal, such that we can use (5.5). □

The result Theorem 5.2(ii) shows that if $\sigma_\varepsilon(A)$ is much larger than $\sigma(A)$, then small perturbations can move eigenvalues very far. See for example Figure 15. So it is important to know whether the pseudospectra are sensitive to small perturbations. If they were, they would be of little value. Fortunately this is not the case. We have the following result.

Theorem 5.12. *Let $A \in \mathcal{B}(\mathcal{H})$ and $\varepsilon > 0$ be given. Let $E \in \mathcal{B}(\mathcal{H})$ with $\|E\| < \varepsilon$. Then we have*

$$\sigma_{\varepsilon - \|E\|}(A) \subseteq \sigma_\varepsilon(A + E) \subseteq \sigma_{\varepsilon + \|E\|}(A). \tag{5.9}$$

Proof. Let $z \in \sigma_{\varepsilon-\|E\|}(A)$. By Theorem 5.2(ii) we can find $F \in \mathcal{B}(\mathcal{H})$ with $\|F\| < \varepsilon - \|E\|$, such that

$$z \in \sigma(A+F) = \sigma((A+E)+(F-E)).$$

Now $\|F-E\| \leq \|F\| + \|E\| < \varepsilon$, so Theorem 5.2(ii) implies $z \in \sigma_\varepsilon(A+E)$. The other inclusion is proved in the same way. □

Exercise 5.13. Prove the second inclusion in (5.9).

There is one nontrivial fact concerning the pseudospectra, which we cannot discuss in detail, since it requires a substantial knowledge of nontrivial results in analysis and partial differential equations.

To state the result we remind the reader of the definition of connected components of an open subset of the complex plane. The connected components are the largest connected open subsets of a given open set in the complex plane. The decomposition into connected components is unique.

Theorem 5.14. *Let $\mathcal{H}$ be finite dimensional, of dimension n. Let $A \in \mathcal{B}(\mathcal{H})$. Let $\varepsilon > 0$ be arbitrary. Then $\sigma_\varepsilon(A)$ is non-empty, open, and bounded. It has at most n connected components, and each connected component contains at least one eigenvalue of A.*

The key ingredient in the proof of this result is the fact that the function $f\colon z \mapsto \|R_A(z)\|$ has no local maxima. This is a nontrivial result, which comes from the fact that this function is what is called subharmonic. For results on subharmonic functions we refer the reader to [Con78, Chapter X, §3.2]. We warn the reader that the function f may have local minima, and we will actually give an explicit example later.

Exercise 5.15. For $A \in \mathcal{B}(\mathcal{H})$ prove the following two results:

1. For any $c \in \mathbf{C}$ and $\varepsilon > 0$ we have $\sigma_\varepsilon(A + cI) = c + \sigma_\varepsilon(A)$.
2. For any $c \in \mathbf{C}$, $c \neq 0$, and $\varepsilon > 0$ we have $\sigma_{|c|\varepsilon}(cA) = c\sigma_\varepsilon(A)$.

6 Examples I

In this section we give some examples of pseudospectra of matrices. The computations are performed using MATLAB with the toolbox `EigTool`. We only mention a few features of each example, and encourage the readers to experiment on their own with the possibilities in this toolbox. In this section we show the figures generated using `EigTool` and comment on various features seen in these figures.

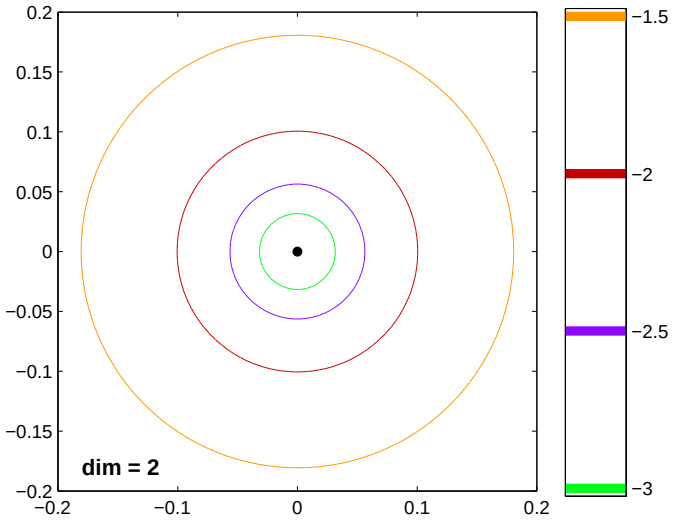

Figure 1: Pseudospectra of A

6.1 Example 1

The 2×2 matrix A is given by

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

This is of course the simplest non-normal matrix. The spectrum is $\sigma(A) = \{0\}$. In this case the norm of the resolvent can be calculated explicitly. The result is

$$\|R_A(z)\| = \frac{\sqrt{2}}{\sqrt{1 + 2|z|^2 - \sqrt{1 + 4|z|^2}}}.$$

Thus for z close to zero the behavior is

$$\|R_A(z)\| \approx \frac{1}{\sqrt{2}|z|^2}.$$

The pseudospectra from `EigTool` are shown in Figure 1. the values of ε are $10^{-1.5}$, 10^{-2}, $10^{-2.5}$, and 10^{-3}. You can read off these exponents from the scale on the right hand side in Figure 1. In subsequent examples we will not mention the range of ε explicitly.

Exercise 6.1. Verify the results on the resolvent norm and its behavior for small z given in this example. Do the exact values and the numerical values agree reasonably well?

Exercise 6.2. We modify the example by considering

$$A_c = \begin{bmatrix} 0 & c \\ 0 & 0 \end{bmatrix}, \quad c \neq 0.$$

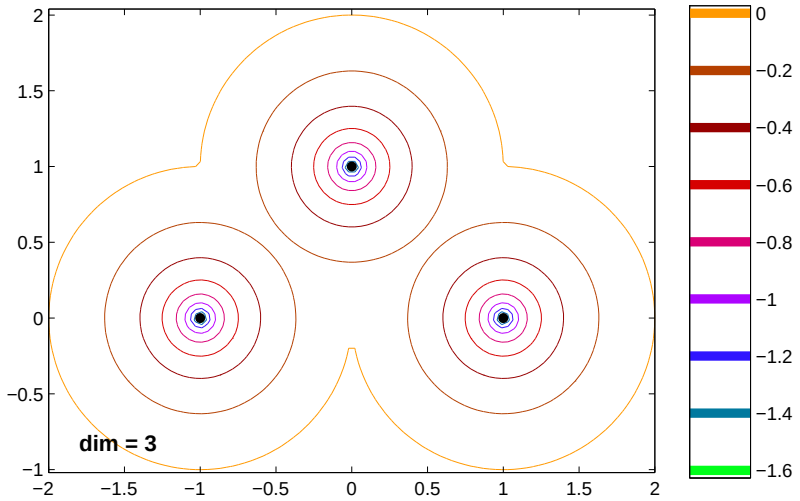

Figure 2: Pseudospectra of B

Do some computer experiments finding the pseudospectra for both $|c|$ small and $|c|$ large. You can take $c > 0$ without loss of generality. Also analyze what happens to the pseudospectra as a function of c, for a fixed ε, using the definitions and Exercise 5.15

6.2 Example 2

We now take a normal matrix, for simplicity a diagonal matrix. We take

$$B = \begin{bmatrix} 1 & 0 & 0 \\ 0 & -1 & 0 \\ 0 & 0 & i \end{bmatrix}.$$

The spectrum is $\sigma(B) = \{1, -1, i\}$. Some pseudospectra are shown in Figure 2. It is evident from the figure that the pseudospectra for each ε considered is the union of three disks centered at the three eigenvalues.

6.3 Example 3

For this example we take the following matrix

$$C = \begin{bmatrix} 1 & 1 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 0 & 0 \end{bmatrix}.$$

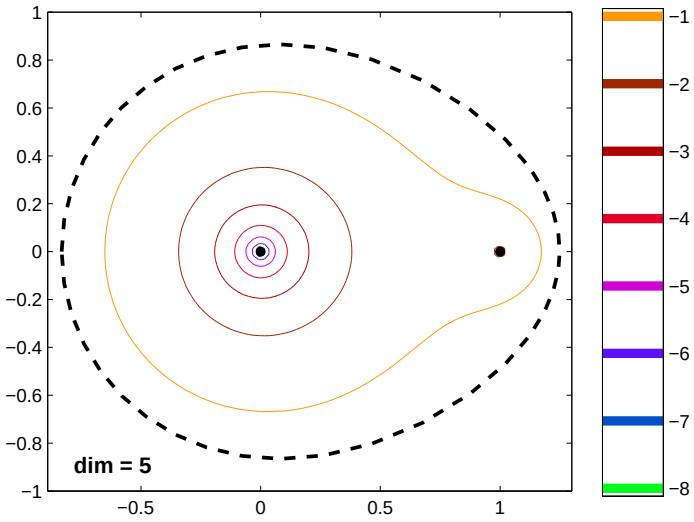

Figure 3: Pseudospectra of C. The boundary of the numerical range is plotted as a dashed curve

We have $\sigma(C) = \{1, 0\}$. Using the notation from Section 2 for algebraic and geometric multiplicity, then we have $m_a(1) = m_g(1) = 1$, $m_a(0) = 4$, $m_g(0) = 1$. Some pseudospectra are shown in Figure 3. It is evident from the figure that the resolvent norm $\|R_C(z)\|$ is much larger at comparable distances from 0 than from 1. On this plot we have shown the boundary of the numerical range of C as a dashed curve.

Note that the matrix C is not in the Jordan canonical form. Let us also consider the corresponding Jordan canonical form. Let us denote it by J. We have $J = Q^{-1}CQ$, where

$$J = \begin{bmatrix} 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 \end{bmatrix} \quad \text{and} \quad Q = \begin{bmatrix} -1 & -1 & -1 & -1 & 1 \\ 1 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 \end{bmatrix}.$$

The pseudospectra of J are shown in Figure 4 in full on the left hand side, and enlarged around 1 in the right hand part. The numerical range is also plotted, as in Figure 3. Comparing the two figures one sees how much closer one has to get to eigenvalue 1 for the Jordan form, before the resolvent norm starts growing. This is a consequence of the size of the condition number of Q. We have

$$\text{cond}(Q) = 3 + 2\sqrt{2} \approx 5.828427125.$$

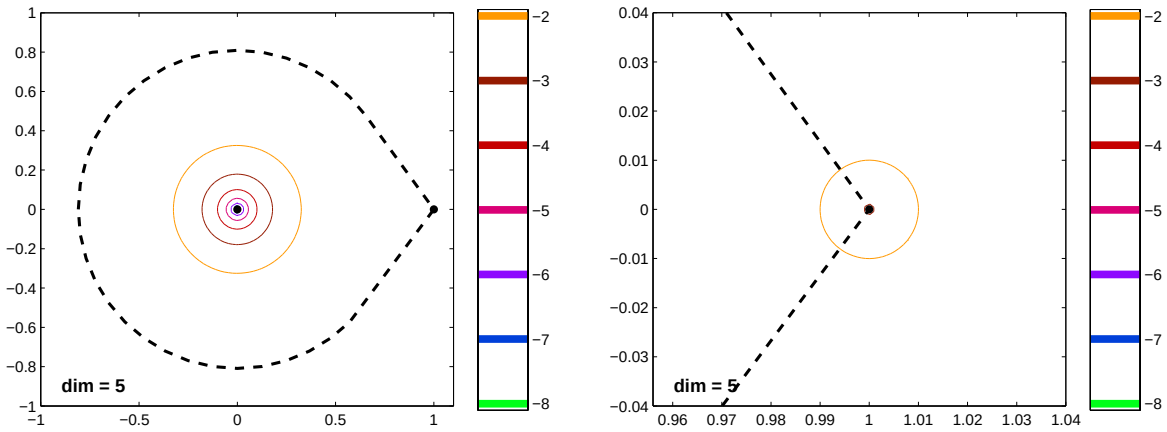

Figure 4: Left hand part: Pseudospectra of J, the Jordan canonical form of C. Right hand part: Enlarged around eigenvalue 1. The boundary of the numerical range is plotted as a dashed curve in both parts

6.4 Example 4

We will give another very simple example. This time we take a rank one projection, which is not normal, i.e. a non-orthogonal projection. Let us start with a general setup. Let $\mathcal{H}$ be a Hilbert space. Let P be a rank one projection, which is not normal. Then there exists a pair $a, b \in \mathcal{H}$ of *linearly independent* vectors, such that

$$Pu = \frac{1}{\langle b, a\rangle}\langle b, u\rangle a. \tag{6.1}$$

This projection is in the two-dimensional case often described as the projection onto the line determined by a in the direction determined by b. It is straightforward to verify that

$$P^*u = \frac{1}{\langle a, b\rangle}\langle a, u\rangle b. \tag{6.2}$$

One can check that $P^*P = PP^*$, if and only if $b = \nu a$ for some $\nu \in \mathbf{C}$, $\nu \neq 0$, i.e. the two vectors are linearly dependent. Thus the projection considered here is never normal.

Since P is a projection, we have $\sigma(P) = \{0, 1\}$, and the eigenvalue 1 has multiplicity 1, whereas the eigenvalue 0 has multiplicity equal to the dimension of $\mathcal{H}$ minus one, in the finite dimensional case, and infinite multiplicity in the infinite dimensional case. In all cases the resolvent can be found explicitly. It is given as

$$(P - zI)^{-1} = \frac{1}{1 - z}P + \frac{1}{0 - z}(I - P), \quad \text{for all } z \in \mathbf{C} \setminus \{0, 1\}. \tag{6.3}$$

Exercise 6.3. Verify all the statements above, including (6.2) and (6.3).

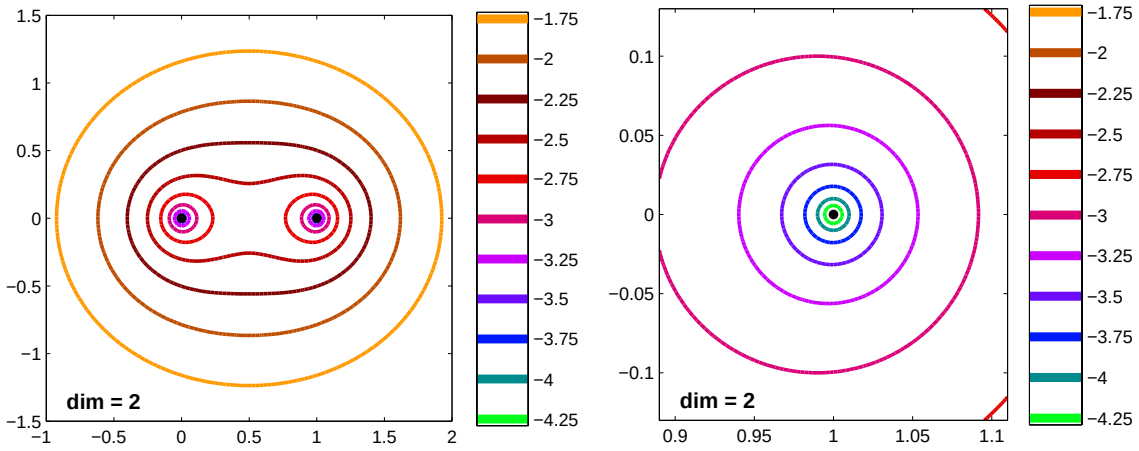

Figure 5: Left hand part: Pseudospectra of A from (6.4), with $\langle b, a\rangle = 10^{-2}$. Right hand part: Enlarged around eigenvalue 1.

Now we will consider the case $\dim \mathcal{H} = 2$, in which case $\{a, b\}$ is a basis for $\mathcal{H}$. The matrix of P in this basis is given as

$$A = \begin{bmatrix} 1 & \dfrac{1}{\langle b, a\rangle} \\ 0 & 0 \end{bmatrix}. \tag{6.4}$$

Thus if $\langle b, a\rangle$ is very small, i.e. the two vectors are almost orthogonal, the off-diagonal entry is very large. This effect can be seen in the pseudospectra. We have taken the matrix A in (6.4) and plotted some of the pseudospectra for $\langle b, a\rangle = 10^{-2}$ and $\langle b, a\rangle = 10^{-3}$ in Figure 5 and Figure 6, respectively. From the right hand part of Figure 6 one sees that for $\varepsilon = 10^{-4.5}$ the radius of the blue circle is approximately 0.03, which shows a large deviation from the behavior in the normal case, where the radius would equal ε.

Exercise 6.4. Do some numerical experiments with matrices of the form (6.4).

If one constructs an orthonormal basis from the basis $\{a, b\}$, the picture changes very little in this case. Let us carry out the details. Assume for definiteness that $\|a\| = 1$ and $\|b\| = 1$. Take as the first basis vector $e_1 = a$ and as the second basis vector $e_2 = \beta(b - \langle a, b\rangle a)$, where $\beta = \|b - \langle a, b\rangle a\|^{-1}$, using the usual Gram-Schmidt procedure. Computing the matrix of P relative to the basis $\{e_1, e_2\}$ yields the following result

$$B = \begin{bmatrix} 1 & \alpha \\ 0 & 0 \end{bmatrix}, \quad \text{where} \quad \alpha = \beta\left(\frac{1}{\langle b, a\rangle} - \langle a, b\rangle\right). \tag{6.5}$$

We have the estimate

$$\|b - \langle a, b\rangle a\| \leq 2,$$

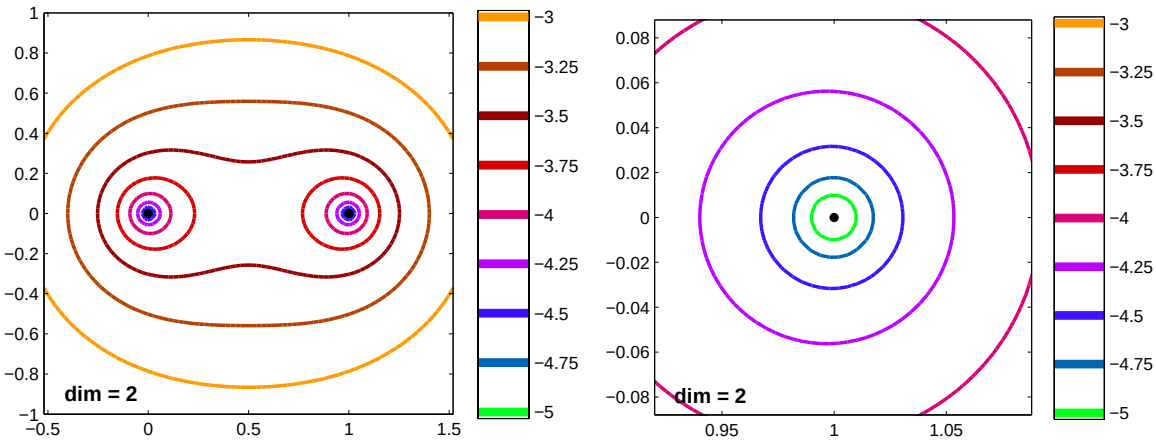

Figure 6: Left hand part: Pseudospectra of A from (6.4), with $\langle b, a\rangle = 10^{-3}$. Right hand part: Enlarged around eigenvalue 1.

since a and b both have norm one, and also the estimate

$$\|b - \langle a, b\rangle a\| \geq \|b\| - \|\langle a, b\rangle a\| = 1 - |\langle b, a\rangle|.$$

Thus we have the estimates

$$\frac{1}{2} \leq \beta \leq \frac{1}{1 - |\langle b, a\rangle|}.$$

Thus if $|\langle b, a\rangle|$ is small, compared to ε, the pseudospectra of the matrices A and B will be almost the same, see Theorem 5.12.

Finally let us diagonalize the matrix A. We have $A = V\Lambda V^{-1}$, where

$$V = \begin{bmatrix} 1 & -\dfrac{1}{\langle b, a\rangle} \\ 0 & 1 \end{bmatrix} \quad \text{and} \quad \Lambda = \begin{bmatrix} 1 & 0 \\ 0 & 0 \end{bmatrix}.$$

The condition number of V can be found explicitly. We have

$$\operatorname{cond}(V) = \frac{1 + 2|\langle b, a\rangle|^2 + \sqrt{1 + 4|\langle b, a\rangle|^2}}{2|\langle b, a\rangle|^2}.$$

Thus we see that the matrix V has large condition number, if $|\langle b, a\rangle|$ is small.

Exercise 6.5. Carry out the computations leading to V and $\operatorname{cond}(V)$ above.

Exercise 6.6. Compare through numerical experiments the pseudospectra obtained with `EigTool` to the unions of disks obtained from Theorem 5.11.

7 Perturbation Theory

We will give some computations from perturbation theory to see how eigenvalues may move far due to small perturbations, when a matrix is non-normal. We will not be completely rigorous, since this requires a substantial machinery. The complete theory for perturbation of eigenvalues on finite dimensional spaces can be found in [Kat95, Chapter I and II]. We should caution the reader that this is not an easily accessible theory. A fair amount of complex analysis is needed to understand the rigorous results.

First we consider the following set-up. Let A be an $n \times n$ matrix. Assume that λ_j is a simple eigenvalue of A with corresponding eigenvector v_j, i.e. $Av_j = \lambda_j v_j$ and $v_j \neq 0$. An eigenvalue λ_j is called simple, if $m_a(\lambda_j) = m_g(\lambda_j) = 1$, or equivalently, if λ_j is a simple zero of the characteristic polynomial $\det(A - zI)$.

Let V be another $n \times n$ matrix. We can assume $\|V\| = 1$. Consider a family of matrices

$$A(g) = A + gV.$$

Now one can use the complex version of the implicit function theorem to conclude that $\det(A(g) - zI) = 0$ for sufficiently small g has a unique solution $\lambda_j(g)$, with $\lambda_j(0) = \lambda_j$, which then is a simple eigenvalue of $A(g)$. We write $A(g)v_j(g) = \lambda_j(g)v_j(g)$. One can show that both the eigenvalue and eigenfunction are analytic functions of g for g small. Thus we have power series expansions

$$\lambda_j(g) = \lambda_j + g\lambda_j^1 + g^2\lambda_j^2 + \cdots, \tag{7.1}$$

$$v_j(g) = v_j + gv_j^1 + g^2 v_j^2 + \cdots. \tag{7.2}$$

Insert these expansions into the equation $(A + gV)v(g) = \lambda(g)v(g)$ and equate the coefficients of the powers of g. The result is for the coefficients to g^j, $j = 0, 1, 2$ as follows:

$$Av_j = \lambda_j v_j, \tag{7.3}$$

$$Av_j^1 + Vv_j = \lambda_j v_j^1 + \lambda_j^1 v_j, \tag{7.4}$$

$$Av_j^2 + Vv_j^1 = \lambda_j v_j^2 + \lambda_j^1 v_j^1 + \lambda_j^2 v_j. \tag{7.5}$$

The equation (7.3) is just the given eigenvalue equation. We rewrite (7.4) as

$$(A - \lambda_j I)v_j^1 = (\lambda_j^1 I - V)v_j. \tag{7.6}$$

Our first goal is to find an expression for λ_j^1. For this purpose we need another vector. We have that $\overline{\lambda_j}$ is an eigenvalue of the adjoint A^*. Let $u_j \neq 0$ be an eigenvector, i.e. $A^* u_j = \overline{\lambda_j} u_j$. Now take inner product between u_j and the left hand side of (7.6), and compute as follows (remember that our inner product is linear in the second variable

and conjugate linear in the first variable):

$$\begin{aligned}\langle u_j, (A-\lambda_j)v_j^1\rangle &= \langle A^*u_j, v_j^1\rangle - \lambda_j\langle u_j, v_j^1\rangle \\ &= \langle \overline{\lambda_j}u_j, v_j^1\rangle - \lambda_j\langle u_j, v_j^1\rangle \\ &= \lambda_j\langle u_j, v_j^1\rangle - \lambda_j\langle u_j, v_j^1\rangle = 0.\end{aligned}$$

Using this result, we get from (7.6), assuming $\langle u_j, v_j\rangle \neq 0$,

$$\lambda_j^1 = \frac{\langle u_j, Vv_j\rangle}{\langle u_j, v_j\rangle}. \tag{7.7}$$

If A is normal, then we can take $u_j = v_j$. This is evident in the special case of a selfadjoint A and follows from the property $AA^* = A^*A$ in the general case. Thus in the normal case the effect of the perturbation V is determined by its size (which we here have normalized to $\|V\| = 1$) and its mapping properties relative to v_j.

If A is not normal, then λ_j^1 can become very large, if $\langle u_j, Vv_j\rangle \approx 1$ and $\langle u_j, v_j\rangle$ close to zero. Note that if actually $\langle u_j, v_j\rangle = 0$, then the derivation above of λ_j^1 is not valid.

In order to find also the first order change in the eigenvector using simple arguments we need to make an assumption on A. We assume that *all eigenvalues of A are simple*. This means that A has n distinct eigenvalues $\lambda_1, \lambda_2, \ldots, \lambda_n$. Corresponding eigenvectors are denoted by $v_1, v_2, \ldots, v_n$. The eigenvalues of A^* are $\overline{\lambda_1}, \overline{\lambda_2}, \ldots, \overline{\lambda_n}$, and the corresponding eigenvectors are denoted by $u_1, u_2, \ldots, u_n$. Both the $\{v_j\}_{j=1,\ldots,n}$ and the $\{u_j\}_{j=1,\ldots,n}$ are bases for $\mathbf{C}^n$. They have a special property. Assume that $k \neq j$, such that $\lambda_k \neq \lambda_j$. We compute as follows.

$$\begin{aligned}0 &= \langle u_k, Av_j\rangle - \langle u_k, Av_j\rangle = \langle A^*u_k, v_j\rangle - \langle u_k, Av_j\rangle \\ &= \langle \overline{\lambda_k}u_k, v_j\rangle - \langle u_k, \lambda_j v_j\rangle \\ &= (\lambda_k - \lambda_j)\langle u_k, v_j\rangle.\end{aligned}$$

We conclude that $\langle u_k, v_j\rangle = 0$ for all $j, k = 1, 2, \ldots, n$ with $k \neq j$. Furthermore, we must have $\langle u_j, v_j\rangle \neq 0$ for all $j = 1, 2, \ldots, n$. This is seen as follows. Suppose $\langle u_j, v_j\rangle = 0$ for some j. Then we have $\langle u_k, v_j\rangle = 0$ for all $k = 1, 2, \ldots, n$. Thus v_j is orthogonal to all vectors in the basis $\{u_j\}_{j=1,\ldots,n}$, which implies $v_j = 0$, a contradiction. We will choose to normalize by the condition $\langle u_j, v_j\rangle = 1$ for all $j = 1, 2, \ldots, n$. Let us introduce some terminology:

Definition 7.1. *Let $\{v_j\}_{j=1,\ldots,n}$ and $\{u_j\}_{j=1,\ldots,n}$ be two bases for $\mathbf{C}^n$. They are called a pair of biorthogonal bases, if they satisfy*

$$\langle u_k, v_j\rangle = \delta_{jk} \quad \textit{for all } j, k = 1, 2, \ldots, n.$$

Here δ_{jk} denotes the Kronecker delta.

There is a simple formula for the coefficients of any vector relative to each of these bases.

Proposition 7.2. *Let Let $\{v_j\}_{j=1,\dots,n}$ and $\{u_j\}_{j=1,\dots,n}$ be a pair of biorthogonal bases for $\mathbf{C}^n$, and let $x \in \mathbf{C}^n$. Then we have*

$$x = \sum_{k=1}^{n} \langle u_k, x\rangle v_k, \tag{7.8}$$

$$x = \sum_{j=1}^{n} \langle v_j, x\rangle u_j. \tag{7.9}$$

Exercise 7.3. Prove this proposition.

Now we come back to the computation of the term v_j^1 in (7.2). We use the above result and the representation in Proposition 7.2

$$v_j^1 = \sum_{k=1}^{n} \langle u_k, v_j^1\rangle v_k.$$

Thus we must try to compute $\langle u_k, v_j^1\rangle$. Assume first that $k \neq j$. Then we can use the equation (7.6). Take inner product with u_k on both sides to get

$$\langle u_k, (A - \lambda_j I)v_j^1\rangle = \lambda_j^1 \langle u_k, v_j\rangle - \langle u_k, Vv_j\rangle.$$

Using $\langle u_k, v_j\rangle = 0$ and $A^* u_k = \overline{\lambda_k} u_k$ we get

$$(\lambda_k - \lambda_j)\langle u_k, v_j^1\rangle = -\langle u_k, Vv_j\rangle.$$

Since $\lambda_k - \lambda_j \neq 0$, we have

$$\langle u_k, v_j^1\rangle = \frac{\langle u_k, Vv_j\rangle}{\lambda_j - \lambda_k}.$$

We cannot determine the value of $\langle u_j, v_j^1\rangle$ from the equations we have. This just reflects the fact that (7.6) is an inhomogeneous linear equation with v_j^1 as the unknown, and the solution is only determined up to a vector in the kernel (null space) of the coefficient matrix $A - \lambda_j I$. We will need an additional condition. So for the moment we have found

$$v_j^1 = cv_j + \sum_{\substack{k=1 \\ k\neq j}}^{n} \frac{\langle u_k, Vv_j\rangle}{\lambda_j - \lambda_k} v_k.$$

Inserting this expression into (7.2) and taking inner product with u_j we find that

$$\langle u_j, v_j(g)\rangle = \langle u_j, v_j\rangle + cg\langle u_j, v_j\rangle + \mathcal{O}(g^2) = 1 + cg + \mathcal{O}(g^2).$$

The requirement one imposes is that for computation of a first order term we must have $\langle u_j, v_j(g)\rangle = 1 + \mathcal{O}(g^2)$, leading to $c = 0$.

So the final result is that

$$v_j(g) = v_j + g\sum_{\substack{k=1\\k\neq j}}^{n} \frac{\langle u_k, Vv_j\rangle}{\lambda_j - \lambda_k} v_k.$$

We see that if the eigenvalues are closely spaced, then the contribution from the second term can be large.

We now give an application of the result (7.7) to pseudospectra of matrices. Since we have assumed $\|V\| = 1$, we get from (7.7) the estimate

$$|\lambda_j^1| \leq \frac{\|u_j\|\|v_j\|}{|\langle u_j, v_j\rangle|} = \kappa(\lambda_j). \tag{7.10}$$

The number $\kappa(\lambda_j)$ is called the condition number of the eigenvalue λ_j. It follows from the Cauchy-Schwarz inequality that we always have $\kappa(\lambda_j) \geq 1$. As before we let $D_\delta = \{z \in \mathbf{C} \,|\, |z| < \delta\}$.

Theorem 7.4. *Let A be an $n\times n$ matrix. Assume that the eigenvalues of A, λ_j, $j = 1,\ldots,n$ all are simple. Let $\varepsilon > 0$. then we have*

$$\sigma_\varepsilon(A) \subseteq \bigcup_{j=1}^{n} (\lambda_j + D_{\varepsilon\kappa(\lambda_j)+O(\varepsilon^2)}). \tag{7.11}$$

Proof. It follows from (7.1) and (7.10) that we have

$$|\lambda_j - \lambda_j(g)| \leq \kappa(\lambda_j)|g| + O(g^2),$$

for all perturbations V with $\|V\| = 1$, since we assume that all eigenvalues are simple. It also follows from the arguments given above that $\kappa(\lambda_j)$ always is finite. The result then follows from Theorem 5.2(ii). □

The theorem shows that for ε small the pseudospectra look like a union of disks with radius $\varepsilon\kappa(\lambda_j)$ and centered at λ_j.

7.1 Examples

We will give two very simple examples illustrating the above results (and some of their limitations). As the first example we take the matrix (6.4).

$$A = \begin{bmatrix} 1 & \alpha \\ 0 & 0 \end{bmatrix}, \quad \text{where} \quad \alpha = \frac{1}{\langle b, a\rangle}.$$

Here we have introduced the parameter α to simplify the notation. The eigenvalues are $\lambda_1 = 1$ and $\lambda_2 = 0$. The eigenvectors of A and A^* must be chosen such that $\langle u_j, v_k \rangle = \delta_{jk}$. We get

$$v_1 = \begin{bmatrix} 1 \\ 0 \end{bmatrix}, \quad v_2 = \begin{bmatrix} -\alpha \\ 1 \end{bmatrix}, \quad u_1 = \begin{bmatrix} 1 \\ \overline{\alpha} \end{bmatrix}, \quad u_2 = \begin{bmatrix} 0 \\ 1 \end{bmatrix}.$$

We choose as the perturbation the matrix

$$V = \begin{bmatrix} 0 & 1 \\ 1 & 0 \end{bmatrix}.$$

This matrix satisfies $\|V\| = 1$. Now it is straightforward to compute the first order corrections to the eigenvalues and eigenvectors. We get

$$\lambda_1^1 = \alpha \quad \text{and} \quad \lambda_2^1 = -\alpha.$$

Exercise 7.5. Verify the above computations.

Exercise 7.6. Compute the first order corrections to the two eigenvectors.

This example is so simple that one can carry out the exact determinations of the eigenvalues. The result is (for g sufficiently small)

$$\lambda_1(g) = \tfrac{1}{2} + \tfrac{1}{2}\sqrt{1 + 4(g\alpha + g^2)},$$
$$\lambda_2(g) = \tfrac{1}{2} - \tfrac{1}{2}\sqrt{1 + 4(g\alpha + g^2)}.$$

From these expressions one can find higher order corrections. Using a computer algebra program one can find for example

$$\lambda_1(g) = 1 + \alpha g + (1 - \alpha^2)g^2 + (-2\alpha + 2\alpha^3)g^3 + \mathcal{O}(g^4), \tag{7.12}$$
$$\lambda_2(g) = 0 - \alpha g - (1 - \alpha^2)g^2 - (-2\alpha + 2\alpha^3)g^3 + \mathcal{O}(g^4). \tag{7.13}$$

We recall from the discussion in Section 6.4 that we mainly consider the case where α is large. It is evident that for large α the coefficients to the powers of g grow quite rapidly. Thus to get any approximation at all we need to have $g\alpha$ small. A numerical example, based on the exact value and the three approximations to $\lambda_1(g)$ above, is shown in Figure 7. We have taken $\alpha = 10^3$. Note that for $g \geq 5.5\,10^{-4}$ the second and third order approximations are *worse* than the first order approximation.

The result from Theorem 7.4 can be applied to this matrix. One has to have ε to be quite small, before one can begin to see the disks. It is easy to compute the eigenvalue condition numbers. We have

$$\kappa(\lambda_1) = \kappa(\lambda_2) = \sqrt{1 + |\alpha|^2}.$$

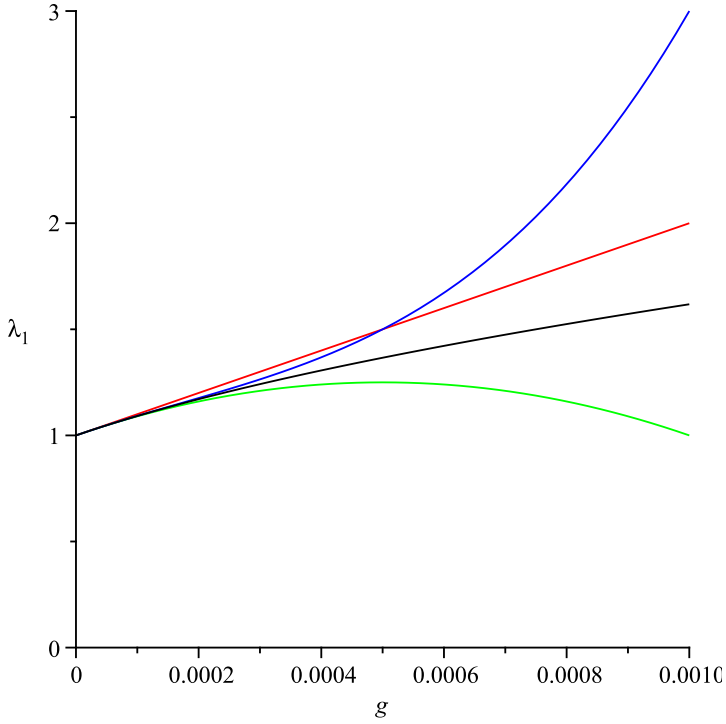

Figure 7: Black: $\lambda_1(g)$. Red: First order approximation. Green: Second order approximation. Blue: Third order approximation.

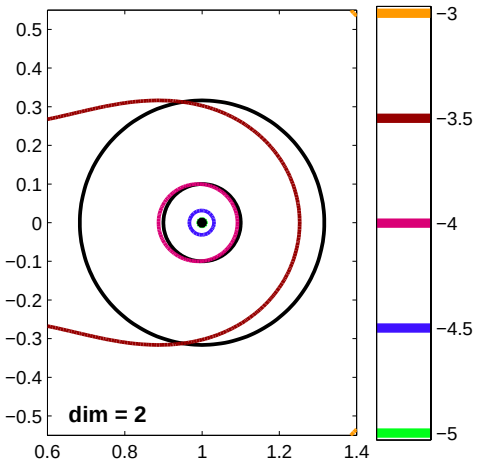

Figure 8: Pseudospectra for the 2×2 matrix A with $\alpha = 10^3$. Black circles come from Theorem 7.4 for $\varepsilon = 10^{-4}$ and $\varepsilon = 10^{-3.5}$

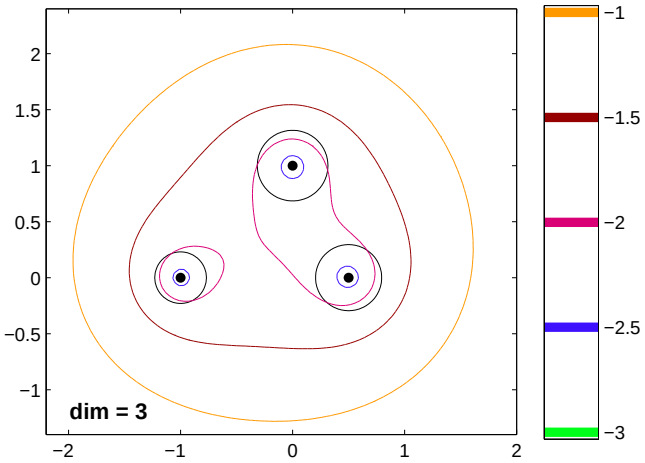

Figure 9: Pseudospectra for the 3×3 matrix S. Black circles are the ones from Theorem 7.4 corresponding to $\varepsilon = 10^{-2}$

Thus in our example with $\alpha = 10^3$ we have $\kappa(\lambda_1) = \kappa(\lambda_2) \approx 10^3$. We repeat the some of the computations shown in Figure 6. In Figure 8 we have plotted a few pseudospectra. We have also plotted in black two circles from Theorem 7.4, corresponding to $\varepsilon = 10^{-4}$ and $\varepsilon = 10^{-3.5}$. For the smaller value of ε the circle and the pseudospectrum boundary almost coincide, whereas for the larger value there are substantial discrepancies.

Let us illustrate Theorem 7.4 with another example. Take the matrix

$$S = \begin{bmatrix} -1 & 6 & 0 \\ 0 & i & 8 \\ 0 & 0 & \frac{1}{2} \end{bmatrix}.$$

The eigenvalues are $\lambda_1 = -1$, $\lambda_2 = i$, and $\lambda_3 = \frac{1}{2}$. The three eigenvalue condition numbers can be computed exactly in Maple or approximately in MATLAB. The results for the approximate values are

$$\kappa(\lambda_1) \approx 23.0, \quad \kappa(\lambda_2) \approx 31.5, \quad \text{and} \quad \kappa(\lambda_3) \approx 29.5.$$

In Figure 9 we have plotted some pseudospectra for this matrix. The three circles corresponding to $\varepsilon = 10^{-2}$ (magenta curves) from Theorem 7.4 (without the error term) are plotted in black. It shows that the results from Theorem 7.4 have to be used with some caution.

8 Applications of pseudospectra I

We now give some applications of the pseudospectra. In some cases we can treat both finite dimensional and infinite dimensional $\mathcal{H}$. In other cases it is technically too demanding to treat the general case, so we treat only finite dimensional $\mathcal{H}$. We first state the results, and then we give the proofs.

Let us return to the problem considered briefly above. We consider the initial value problem

$$\frac{du}{dt}(t) = Au(t), \tag{8.1}$$

$$u(0) = u_0, \tag{8.2}$$

where $u\colon \mathbf{R} \to \mathcal{H}$ is a continuously differentiable function. Then the solution is given by

$$u(t) = \exp(tA)u_0. \tag{8.3}$$

In the finite dimensional case the following stability result is well known, see any introductory text on ordinary differential equations.

Proposition 8.1. *Let A be an $n \times n$ matrix. Assume that all eigenvalues λ of A satisfy $\operatorname{Re}\lambda < 0$. Then we have*

$$\lim_{t\to\infty} \|e^{tA}\| = 0.$$

The result shows that 0 is an asymptotically stable solution to (8.1).

Now if A is not normal, then the solution can become very large, before it starts to decay. Our goal is to show that you can use pseudospectra to quantify these qualitative statements.

We start with an example.

Example 8.2. We consider the two matrices

$$A = \begin{bmatrix} -1 & 1 & 0 \\ 0 & -1 & 1 \\ 0 & 0 & -1 \end{bmatrix}, \qquad B = \begin{bmatrix} -1 & 1 & 5 \\ 0 & -1 & 1 \\ 0 & 0 & -1 \end{bmatrix}.$$

We note that they both have -1 as their only eigenvalue, and that both matrices are not normal. We plot the operator norms $\|e^{tA}\|$ and $\|e^{tB}\|$ as functions of t in Figure 10. The question is which curve belongs to which matrix? We will return to this question below, and also explain the meaning of the green line segments on the figure.

To obtain results on the transient behavior of the solution to an initial value problem as given in (8.1) and (8.2), we need a number of definitions.

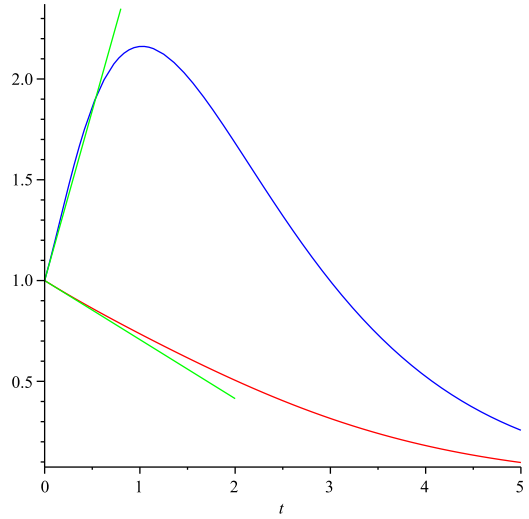

Figure 10: Plot of $\|e^{tA}\|$ and $\|e^{tB}\|$

Definition 8.3. *We define the following quantities. Let $A \in \mathcal{B}(\mathcal{H})$.*

$$\alpha(A) = \sup\{\operatorname{Re} z \mid z \in \sigma(A)\}, \tag{8.4}$$

$$\alpha_\varepsilon(A) = \sup\{\operatorname{Re} z \mid z \in \sigma_\varepsilon(A)\}, \tag{8.5}$$

$$\omega(A) = \sup\{\operatorname{Re} z \mid z \in W(A)\}. \tag{8.6}$$

$\alpha(A)$ is called the spectral abscissa of A, $\alpha_\varepsilon(A)$ is called the pseudospectral abscissa of A, and $\omega(A)$ is called the numerical abscissa of A.

Briefly stated, the value of $\omega(A)$ determines the initial behavior of $\|e^{tA}\|$, while $\alpha(A)$ determines the long time behavior. We state some precise results.

Theorem 8.4. *Let $A \in \mathcal{B}(\mathcal{H})$. Then we have*

$$\alpha(A) = \lim_{t\to\infty} \frac{1}{t} \log\|e^{tA}\|. \tag{8.7}$$

Furthermore, we also have

$$\|e^{tA}\| \geq e^{t\alpha(A)} \quad \text{for all } t \geq 0. \tag{8.8}$$

The estimate (8.8) tells us that the norm can never decay faster than $e^{t\alpha(A)}$, while the limit result (8.7) tells us that for t large the norm behaves precisely as $e^{t\alpha(A)}$.

Concerning the initial behavior, then we have the following result.

Theorem 8.5.

$$\omega(A) = \frac{d}{dt}\,\|e^{tA}\|\,\big|_{t=0} = \lim_{t\downarrow 0} \frac{1}{t} \log\|e^{tA}\|. \tag{8.9}$$

We also have

$$\|e^{tA}\| \leq e^{t\omega(A)} \quad \text{for all } t \geq 0. \tag{8.10}$$

The estimate (8.10) tells us that the norm can never grow faster than $e^{t\omega(A)}$, while the result (8.9) tells us that initially the solution actually grows that fast.

Now we see what kind of information the pseudospectra can provide. These results are a little more complicated to state, and to use.

Theorem 8.6. *For all $\varepsilon > 0$ we have*

$$\sup_{t\geq 0}\|e^{tA}\| \geq \frac{\alpha_\varepsilon(A)}{\varepsilon}. \tag{8.11}$$

The estimate (8.11) tells us that there will be values at some time $t > 0$ at least as large as $\frac{\alpha_\varepsilon(A)}{\varepsilon}$.

Definition 8.7. *Let $A \in \mathcal{B}(\mathcal{H})$. The Kreiss constant is given by*

$$\mathcal{K}(A) = \sup_{\varepsilon>0} \frac{\alpha_\varepsilon(A)}{\varepsilon}. \tag{8.12}$$

Then we have

Corollary 8.8.

$$\sup_{t\geq 0}\|e^{tA}\| \geq \mathcal{K}(A).$$

In the matrix case we can also get an upper bound.

Theorem 8.9. *If A is an $n \times n$ matrix, then we have*

$$\|e^{tA}\| \leq en\mathcal{K}(A).$$

It is also possible to get an estimate valid for a finite time interval, but the estimate is somewhat complicated. Here is a result of that type.

Theorem 8.10. *Let $a = \operatorname{Re} z$. Let $K = \operatorname{Re} z\|(A - zI)^{-1}\|$. Then for $\tau > 0$ we have*

$$\sup_{0\leq t\leq\tau} \|e^{tA}\| \geq e^{a\tau}\left(1 + \frac{e^{a\tau} - 1}{K}\right)^{-1}$$

Example 8.2 continued. In view of the results stated above, let us see how we can answer the question posed in Example 8.2. We take the two matrices A and B, and then use `EigTool` to find the pseudospectra. We also find the numerical range, and compute the numerical abscissa, $\omega(A)$ and $\omega(B)$. The results are plotted in Figure 11. The numerical results are

$$\omega(A) = -0.292893 \quad \text{and} \quad \omega(B) = 1.68614.$$

Thus using Theorem 8.5 it is clear that the upper (red) curve in Figure 10 is of $\|e^{tB}\|$, and the lower (blue) curve shows $\|e^{tB}\|$. We have plotted the two tangent line segments from Theorem 8.5 in green in Figure 10.

In `EigTool` the estimates in Theorem 8.10 can be computed and plotted. In Figure 12 we have plotted the result for the matrix B. The green curve plots the estimate from Theorem 8.10 as a function of τ. The lower estimate (8.8) is plotted as a black dashed curve.

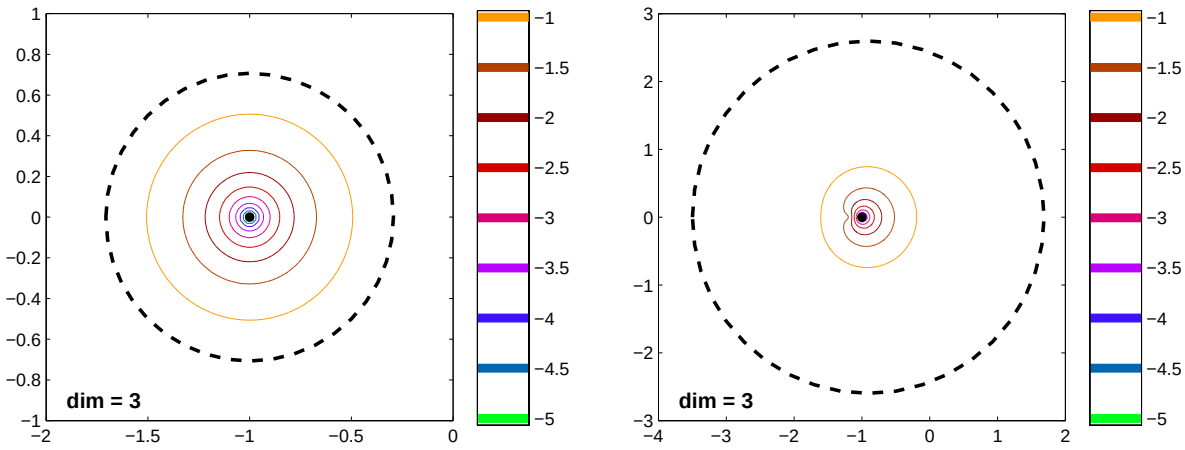

Figure 11: Pseudospectra of A (left hand plot) and B (right hand plot) from Example 8.2. The black dashed curves show the boundaries of the numerical ranges. Note that the scales on the two figures are different

8.1 Proofs

The proofs use the Dunford calculus as defined in (4.2). We start with a general result. In some cases we refer to the literature, since the proofs are somewhat complicated or require substantial preparation.

Proposition 8.11. *Let $A \in \mathcal{B}(\mathcal{H})$. Then we have $\|e^{tA}\| \le e^{t\|A\|}$ for all $t \ge 0$. Assume that we have an estimate*

$$\|e^{tA}\| \le Me^{t\omega} \quad \text{for } t \ge 0, \tag{8.13}$$

where $M \ge 1$ and $\omega \in \mathbf{R}$. Then all $z \in \mathbf{C}$ with $\operatorname{Re} z > \omega$ belong to the resolvent set of A, and we have the formula

$$(A - zI)^{-1} = -\int_0^\infty e^{-tz}e^{tA}dt. \tag{8.14}$$

If Γ is a simple closed contour with $\sigma(A)$ in its interior, then

$$e^{tA} = \frac{-1}{2\pi i}\int_\Gamma e^{tz}(A - zI)^{-1}dz. \tag{8.15}$$

Proof. The estimate $\|e^{tA}\| \le e^{t\|A\|}$ follows from the power series for the exponential function.

To prove (8.14) one first notices that the assumptions imply that

$$\|e^{-tz}e^{tA}\| = \|e^{t(A-zI)}\| \le Me^{-t(\operatorname{Re} z-\omega)}.$$

Thus the integral in (8.14) is defined. Then we use the result

$$e^{t(A-zI)} = (A - zI)^{-1}\frac{d}{dt}e^{t(A-zI)}.$$

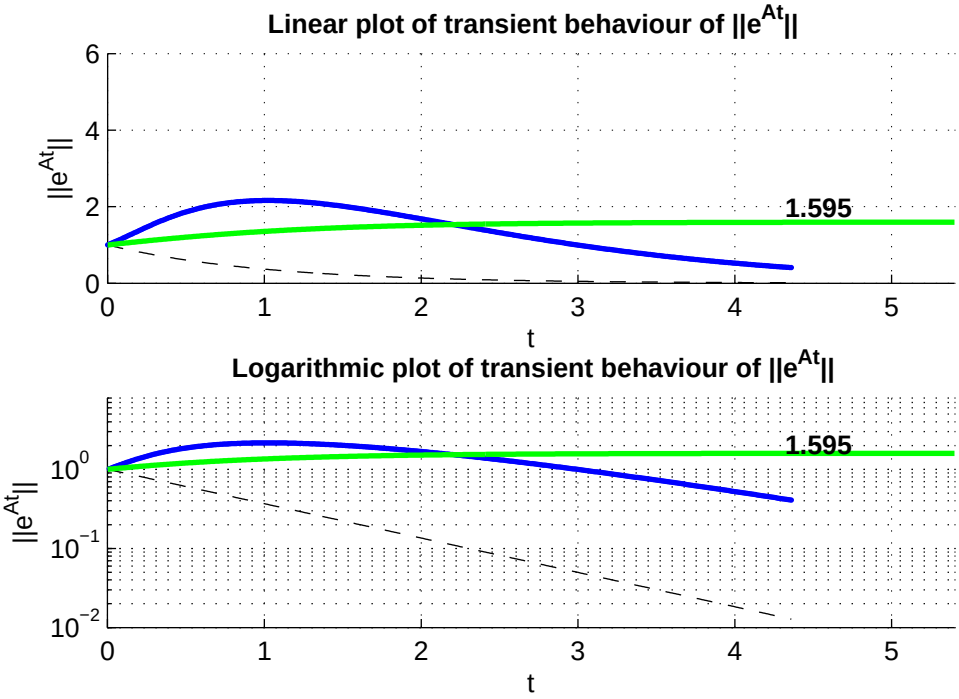

Figure 12: Transient plots for B using `EigTool`. The green curve is the lower estimate obtained from Theorem 8.10, plotted as a function of τ. The dashed black curve is the lower bound from (8.8). Note that `EigTool` uses the notation A for any matrix plotted using this option

Further details are omitted. The formula (8.15) is just the Dunford calculus definition of e^{tA}. Now one has to know that this definition is the same as the power series definition. Details are omitted. Detailed proofs can be found in [Dav07]. □

The transform in (8.14) is the Laplace transform. The inversion formula is (8.15). Next we use this result to get an estimate of e^{tA} involving pseudospectra.

Proposition 8.12. *Let $A \in \mathcal{B}(\mathcal{H})$ and let $\varepsilon > 0$ be fixed. Let Γ_ε be a simple closed contour with $\sigma_\varepsilon(A)$ in its interior. Denote the arc length of Γ_ε by L_ε. Then for all $t \geq 0$ we have the estimate*

$$\|e^{tA}\| \leq \frac{L_\varepsilon e^{t\alpha_\varepsilon(A)}}{2\pi\varepsilon}. \tag{8.16}$$

Proof. The estimate (8.16) follows from (8.15) and the definition of $\alpha_\varepsilon(A)$, see (8.5). □

Proof of Theorem 8.4. We begin with the proof of (8.8). The proof is by contradiction. Assume there exists $\tau > 0$ such that

$$\|e^{\tau A}\| < e^{\tau\alpha(A)}.$$

To simplify the argument we will also assume that $\|e^{tA}\| \leq 1$ for all $t \geq 0$. This can always be obtained by replacing A with $A - \|A\|I$. Define β and ν by

$$\|e^{\tau A}\| = \nu = e^{\tau\beta}.$$

Note that $\nu < e^{\tau\alpha(A)}$ implies $\beta < \alpha(A)$. Now we have

$$\|e^{tA}\| \leq 1 \quad \text{for } 0 \leq t < \tau.$$

For $\tau \leq t < 2\tau$ we have

$$\|e^{tA}\| = \|e^{(\tau+(t-\tau))A}\| = \|e^{\tau A}e^{(t-\tau)A}\| \leq \nu.$$

Continuing this argument we find that

$$\|e^{tA}\| \leq \nu^2 \quad \text{for } 2\tau \leq t < 3\tau,$$

and in general

$$\|e^{tA}\| \leq \nu^n \quad \text{for } n\tau \leq t < (n+1)\tau.$$

Thus for $n\tau \leq t < (n+1)\tau$ we get

$$\|e^{tA}\| \leq e^{n\tau\beta} = e^{t\beta}e^{s\beta} \leq Me^{t\beta}.$$

Here $0 \leq s < 1$ and $M = \sup_{0\leq s<1} e^{s\beta}$.

Thus we have shown that $\|e^{tA}\| \leq Me^{t\beta}$ for all $t \geq 0$. Now we use Proposition 8.11 to conclude that all $z \in \mathbf{C}$ with $\operatorname{Re} z > \beta$ belong to the resolvent set of A. But this result contradicts the definition of $\alpha(A)$, since $\beta < \alpha(A)$.

Now let us consider (8.7). The estimate (8.8) immediately implies that

$$\liminf_{t\to\infty} t^{-1}\log\|e^{tA}\| \geq \alpha(A).$$

Let $\varepsilon > 0$ be arbitrary. From (8.16) we conclude that

$$\limsup_{t\to\infty} t^{-1}\log\|e^{tA}\| \leq \alpha_\varepsilon(A).$$

From Proposition 5.5 we get that $\lim_{\varepsilon\downarrow 0}\alpha_\varepsilon(A) = \alpha(A)$. This remark concludes the proof of Theorem 8.4. □

Proof of Theorem 8.5. The proof is somewhat demanding. We have decided not to include it. We refer to [TE05, §17]. The result is also a consequence of the so-called Lumer-Phillips theorem in semigroup theory, see [Dav07, Theorem 8.3.4]. Note that in order to get from this theorem to Theorem 8.5 one has to do a number of computations. □

Proof of Theorem 8.6. We can assume that $\alpha_\varepsilon(A) > 0$, since otherwise there is nothing to prove. Assume that for a z with $\operatorname{Re} z > 0$ there exists a constant $K > 1$ such that

$$\|(A - zI)^{-1}\| = \frac{K}{\operatorname{Re} z}.$$

Then we can conclude that

$$\sup_{t\geq 0}\|e^{tA}\| \geq K.$$

To get this result let $M = \sup_{t\geq 0}\|e^{tA}\|$. Then we use (8.14) to get

$$\frac{K}{\operatorname{Re} z} = \|(A - zI)^{-1}\| = \|\int_0^\infty e^{-zt}e^{tA}dt\| \leq M\int_0^\infty |e^{-zt}|dt = \frac{M}{\operatorname{Re} z},$$

which implies the result stated above.

Now we use this result to prove Theorem 8.6. Choose a z in the right half plane such that $\operatorname{Re} z = \alpha_\varepsilon(A)$. Then for this z we have

$$\|(A - zI)^{-1}\| = \frac{1}{\varepsilon} = \frac{K}{\operatorname{Re} z}, \quad \text{where} \quad K = \frac{\alpha_\varepsilon(A)}{\varepsilon}.$$

Now (8.11) follows from the first half of the proof. □

Proof of Theorem 8.9. This result is highly non-trivial. See [TE05, §18]. □

Proof of Theorem 8.10. The proof can be found in [TE05, §15]. □

9 Applications of pseudospectra II

Instead of the continuous time problem (8.1), (8.2) one can also consider the discrete time problem. Thus one considers the problem

$$u_{n+1} = Au_n, \tag{9.1}$$

$$u_0 = \phi. \tag{9.2}$$

with the solution

$$u_n = A^n\phi. \tag{9.3}$$

Results concerning the behavior for both small, intermediate, and large values of n are similar to those for the continuous problem, but still somewhat different. We start with some definitions. Proofs are given at the end of this section.

Definition 9.1. *Let $A \in \mathcal{B}(\mathcal{H})$. We define the following quantities.*

$$\rho(A) = \sup\{|z| \mid z \in \sigma(A)\}, \tag{9.4}$$

$$\rho_\varepsilon(A) = \sup\{|z| \mid z \in \sigma_\varepsilon(A)\}, \tag{9.5}$$

$$\mu(A) = \sup\{|z| \mid z \in W(A)\}. \tag{9.6}$$

$\rho(A)$ is called the spectral radius of A, $\rho_\varepsilon(A)$ is called the pseudo-spectral radius of A, and $\mu(A)$ is called the numerical radius of A.

Let us start with some upper bounds on A^n.

Theorem 9.2. *Assume $A \in \mathcal{B}(\mathcal{H})$. Then we have the following results.*
(i) *For $n \geq 0$ we have*

$$\|A^n\| \leq \|A\|^n. \tag{9.7}$$

(ii) *For any $\varepsilon > 0$ and all $n \geq 0$ we have*

$$\|A^n\| \leq \frac{(\rho_\varepsilon(A))^{n+1}}{\varepsilon}. \tag{9.8}$$

We have the following result concerning the behavior for large n.

Theorem 9.3. *Let $A \in \mathcal{B}(\mathcal{H})$. We have*

$$\rho(A) = \lim_{n\to\infty} \|A^n\|^{1/n}. \tag{9.9}$$

We also have

$$\|A^n\| \geq (\rho(A))^n \quad \text{for all } n \geq 0. \tag{9.10}$$

Thus the estimate (9.10) shows that the norm cannot decrease faster than $(\rho(A))^n$, and the result (9.9) shows that for sufficiently large n it behaves that way.

The initial behavior is governed by the numerical radius. We have the following result.

Theorem 9.4. *Let $A \in \mathcal{B}(\mathcal{H})$. Then we have*

$$\|A^n\| \leq 2(\mu(A))^n \quad \text{for all } n \geq 0. \tag{9.11}$$

Thus the numerical radius determines how fast the norm of the powers of A can grow. Due to the factor 2 this result is not as good as one would like. But it is the best result obtainable.

The intermediate behavior is governed by the pseudo-spectra. The simplest bound is the following.

Theorem 9.5. *Let $A \in \mathcal{B}(\mathcal{H})$. Then we have for all $\varepsilon > 0$*

$$\sup_{n\geq 0} \|A^n\| \geq \frac{\rho_\varepsilon(A) - 1}{\varepsilon}. \tag{9.12}$$

Obviously, this bound only yields information, if ε and A satisfy that $\rho_\varepsilon(A) > 1$. One can do a scaling to get the following result.

Corollary 9.6. *Let $\gamma > 0$. Then for all $\varepsilon > 0$ we have*

$$\sup_{n\geq 0} \|\gamma^{-n} A^n\| \geq \frac{\rho_\varepsilon(A) - \gamma}{\varepsilon}. \tag{9.13}$$

There is a number of other estimates, which we will not state here. We refer to [TE05].

9.1 Proofs

In this section we prove some of the results stated above. We start with the following result.

Proposition 9.7. *Let $A \in \mathcal{B}(\mathcal{H})$. Assume that for some $M \geq 1$ and $\gamma > 0$ we have*

$$\|A^n\| \leq M\gamma^n \quad \text{for } n \geq 0.$$

Then any $z \in \mathbf{C}$ with $|z| > \gamma$ belongs to the resolvent set of A, and we have the representation

$$(A - zI)^{-1} = \sum_{k=0}^{\infty} \frac{-1}{z^{k+1}} A^k. \tag{9.14}$$

Let Γ be a simple closed contour with $\sigma(A)$ in its interior. Then we have

$$A^k = \frac{-1}{2\pi i} \int_\Gamma z^k (A - zI)^{-1} dz. \tag{9.15}$$

Proof. The first result is a simple application of the geometric series. The second result states a consequence of the Dunford calculus. □

Proof of Theorem 9.2. The estimate (9.7) is trivial. The estimate (9.8) follows from (9.15), if one takes as Γ the circle centered at the origin with radius $\rho_\varepsilon(A)$. □

Proof of Theorem 9.3. The results stated are standard results from functional analysis, and proofs can be found in most introductions to functional analysis or operator theory. See for example [RS80, Theorem VI.6]. □

Proof of Theorem 9.4. This is not an easy result to obtain. We refer to [TE05] for the proof. □

Proof of Theorem 9.5. We start with the following result. Assume that

$$\|(A - zI)^{-1}\| = \frac{K}{|z| - 1} \quad \text{for some } K > 1 \text{ and } z \text{ with } |z| = r > 1.$$

Then we can conclude that

$$\sup_{n \geq 0} \|A^n\| \geq rK - r + 1 > K. \tag{9.16}$$

To obtain this result, let $M = \sup_{n \geq 0} \|A^n\|$. We can without loss of generality assume that this quantity is finite. Using the assumption and (9.14) we compute as follows.

$$\frac{rK}{r-1} = r\|(A - zI)^{-1}\| \leq 1 + M \sum_{k=1}^{\infty} r^{-k} = 1 + \frac{M}{r-1}.$$

Solving for M yields the first estimate in (9.16). The second estimate follows from $rK - r + 1 = K + (K-1)(r-1) > K$.

Now apply this result to prove (9.12). We fix $\varepsilon > 0$ and choose a z such that $|z| = \rho_\varepsilon(A)$. Solve for K in

$$\|(A - zI)^{-1}\| = \frac{1}{\varepsilon} = \frac{K}{|z| - 1}$$

to get

$$K = \frac{\rho_\varepsilon(A) - 1}{\varepsilon},$$

and the result follows from (9.16). □

10 Examples II

In this section we give a number of matrix examples illustrating the concepts introduced in the previous sections. Several of the examples are from the book [TE05]. The figures below have been generated using either the MATLAB toolbox `EigTool` or Maple.

10.1 A Toeplitz matrix

A matrix with the following structure is called a Toeplitz matrix: $A = [a_{ij}]$ with $a_{ij} = b_{i-j}$. Thus it looks like

$$A = \begin{bmatrix} b_0 & b_{-1} & b_{-2} & \cdots & b_{1-N} \\ b_1 & b_0 & b_{-1} & \cdots & b_{2-N} \\ b_2 & b_1 & b_0 & \cdots & b_{3-N} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ b_{N-1} & b_{N-2} & b_{N-3} & \cdots & b_0 \end{bmatrix}. \tag{10.1}$$

The entries are constant along lines parallel with the main diagonal.

A lot is known about Toeplitz matrices and their infinite dimensional analogues, the Toeplitz operators. We will use a few Toeplitz matrices as examples.

We start with the following example. A is an $N \times N$ Toeplitz matrix with the following structure.

$$A = \begin{bmatrix} 0 & 1 & 0 & \cdots & 0 & 0 \\ \frac{1}{4} & 0 & 1 & \cdots & 0 & 0 \\ 0 & \frac{1}{4} & 0 & \cdots & 0 & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & 0 & 0 & \cdots & 0 & 1 \\ 0 & 0 & 0 & \cdots & \frac{1}{4} & 0 \end{bmatrix}. \tag{10.2}$$

Let Q denote the diagonal $N \times N$ matrix with entries $2, 4, 8, \ldots, 2^N$ on the diagonal. Then one can verify that

$$QAQ^{-1} = B, \tag{10.3}$$

where

$$B = \begin{bmatrix} 0 & \frac{1}{2} & 0 & \cdots & 0 & 0 \\ \frac{1}{2} & 0 & \frac{1}{2} & \cdots & 0 & 0 \\ 0 & \frac{1}{2} & 0 & \cdots & 0 & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & 0 & 0 & \cdots & 0 & \frac{1}{2} \\ 0 & 0 & 0 & \cdots & \frac{1}{2} & 0 \end{bmatrix}. \tag{10.4}$$

The matrix B is symmetric, and its eigenvalues can be found to be

$$\lambda_k = \cos\left(\frac{k\pi}{N+1}\right), \quad k = 1, \ldots, N. \tag{10.5}$$

Due to (10.3) the spectrum of A is the same. Thus A is a non-symmetric matrix with a real spectrum. We will now look at its pseudospectra. For $N = 64$ the pseudospectra and the eigenvalues are shown in Figure 13. This figure also shows the boundary of the numerical range. The numerical abscissa is $\omega(A) = 1.24854$.

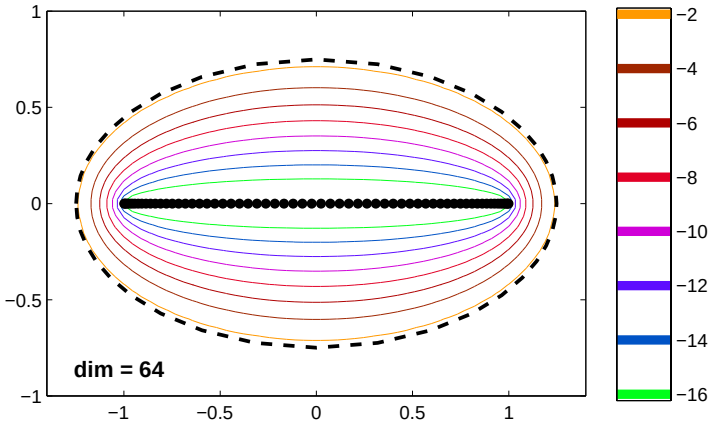

Figure 13: Pseudospectra and eigenvalues of A. The black dashed curve is the boundary of the numerical range

If one considers the infinite dimensional analogue of this matrix, then one can show that the spectrum is the ellipse $\{z^{-1} + \frac{1}{4}z \,|\, |z| = 1\}$, and its interior.

Let us now illustrate the result in Theorem 5.2(ii). We take the matrix A and add a random matrix E of norm $\|E\| = 10^{-3}$. We then plot the eigenvalues of this matrix. We repeat this procedure 1000 times. This generates Figure 14. In Figure 15 the same procedure is repeated for $\|E\| = 10^{-8}$ Notice that most of the eigenvalues move far from the spectrum, even for the perturbations of size 10^{-8}.

It is remarkable that for all finite values of N the spectrum of the matrix is on the real line, thus not in any sense approximating the spectrum of the infinite dimensional matrix. In other words, from computations of the spectrum of the truncated infinite matrix one gets no impression of where the spectrum of the infinite matrix is located. But the pseudospectra give a good approximation to the infinite matrix spectrum, since it fills out the ellipse approximately. So one thing one can learn from this example is that one can never be sure of the location of the spectrum of an infinite matrix, based on computations with truncated finite matrices.

One can show that most of the eigenvalues in Figure 14 lie close to the ellipse, which is the image of $|z| = 10^{-3/64}$ under the map $f(z)$ defined above.

One may also compare the numerical results above with Proposition 5.8. The condition number of Q can be found exactly, since Q is diagonal. We get $\mathrm{cond}(Q) = 2^{N-1}$. Thus for $N = 64$ one has

$$\mathrm{cond}(Q) = 2^{63} = 9223372036854775808 \approx 9.223372037 \cdot 10^{18}.$$

Let us now look at some of the results from Section 9, applied to the matrices A and B. Since A and B have the same spectrum, they have the same spectral radius. Using

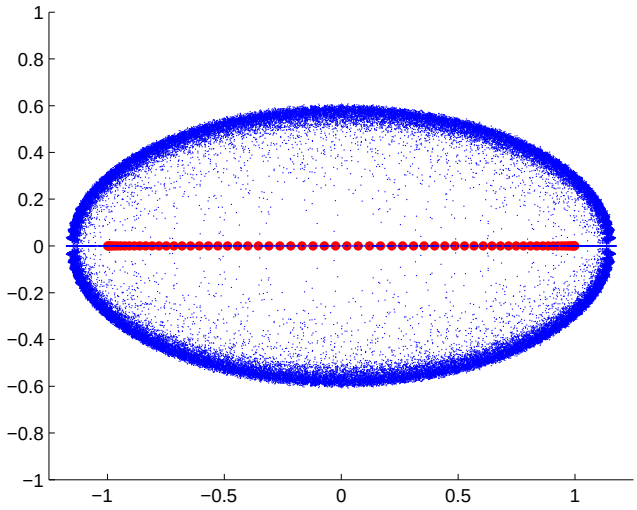

Figure 14: Eigenvalues of A plotted in red. Blue points are eigenvalues of $A + E$, with E random and of norm 10^{-3}, for 1000 choices of E

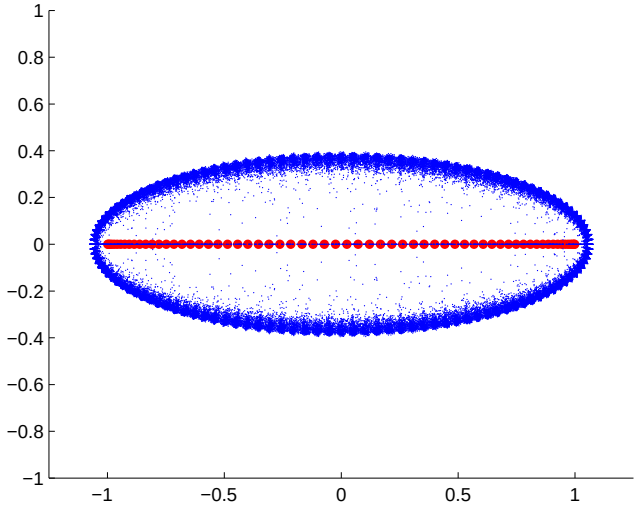

Figure 15: Eigenvalues of A plotted in red. Blue points are eigenvalues of $A + E$, with E random and of norm 10^{-8}, for 1000 choices of E

(10.5), we get

$$\rho(A) = \rho(B) = \cos(\frac{\pi}{N+1})$$

in general. Note that this number is strictly less than one. Since B is symmetric, the norm equals the spectral radius, and we have

$$\|B^k\| = \rho(B^k) = (\rho(B))^k < 1 \quad \text{for all } k \geq 1. \tag{10.6}$$

Using (10.3), we have

$$\|A^k\| = \|Q^{-1}B^kQ\| \leq \|Q\| \cdot \|B^k\| \cdot \|Q^{-1}\| < \text{cond}(Q), \tag{10.7}$$

which is large for $N = 64$. The results (10.7) and (10.6) imply that

$$\lim_{k\to\infty}\|A^k\| = 0 \quad \text{and} \quad \lim_{k\to\infty}\|B^k\| = 0.$$

But the initial and intermediate behavior are both quite different. This is illustrated in Figure 16. Here we have plotted the norms $\|A^k\|$ and $\|B^k\|$ for $k = 0, \ldots, 1600$. The norms $\|B^k\|$ decay monotonically, whereas the norms $\|A^k\|$ initially grow exponentially. Only for k larger than around 1000 can one see the decay setting in. The initial growth is like $\omega(A)^k$. Here $\omega(A) \approx 1.25$. In Figure 17 we have carried the computations further to the value $k = 10^4$. Now one sees the decay also in $\|A^k\|$.

Exercise 10.1. Try to verify some of the statements and results in this example.

Exercise 10.2. Try to modify the matrix in this example to a circulant matrix, i.e. a matrix where each row is obtained from the previous one by a shift. Take for example

$$A = \begin{bmatrix} 0 & 1 & 0 & \cdots & 0 & \frac{1}{4} \\ \frac{1}{4} & 0 & 1 & \cdots & 0 & 0 \\ 0 & \frac{1}{4} & 0 & \cdots & 0 & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & 0 & 0 & \cdots & 0 & 1 \\ 1 & 0 & 0 & \cdots & \frac{1}{4} & 0 \end{bmatrix}. \tag{10.8}$$

Compute spectrum and pseudospectra using `EigTool`. Discuss the results you find. What do the results tell you about a circulant matrix? Can you prove this property of a circulant matrix?

Exercise 10.3. If you are familiar with functional analysis and Fourier analysis, you realize that the infinite Toeplitz matrix is a convolution with the sequence b_k. The circulant matrix from the previous exercise is a convolution acting on N-periodic sequences. Thus the spectra of both can be found using Fourier analysis.

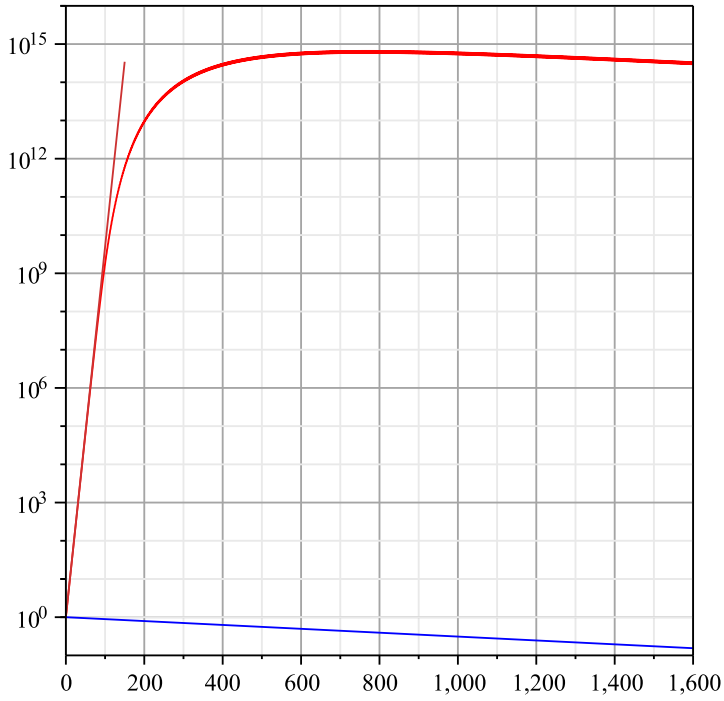

Figure 16: Plots of the powers $\|A^k\|$ (red) and $\|B^k\|$ (blue) up to $k = 1600$. Note that the vertical scale is logarithmic. The straight line is a plot of 1.25^k

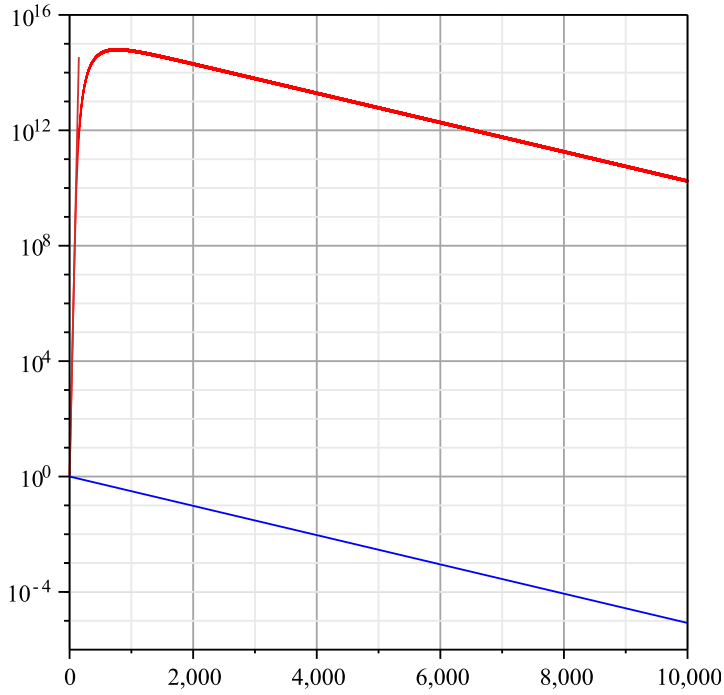

Figure 17: Plots of the powers $\|A^k\|$ (red) and $\|B^k\|$ (blue) up to $k = 10000$. Note that the vertical scale is logarithmic. The straight line is a plot of 1.25^k

10.2 Differentiation matrices

Many problems to be solved numerically involve differentiation, for example solving a differential equation numerically. There are various ways of doing this. Given a sequence of points $\mathbf{x} = \{x_0, x_1, \ldots, x_N\}$ and function values $\mathbf{u} = \{u_0, u_1, \ldots, u_N\}$, where $u_j = f(x_j)$, i.e. sampled values of a function. Then one would like to find approximately the values of the derivative of f at the sample points, $f'(x_j)$.

One such technique is the finite difference method. For equally spaced sample points with $x_j - x_{j-1} = h$ one can approximate the first derivative by

$$f'(x_j) \approx \frac{u_j - u_{j-1}}{h} \quad \text{or} \quad f'(x_j) \approx \frac{u_{j+1} - u_j}{h}.$$

The second derivative can be approximated by the following expression

$$f''(x_j) \approx \frac{u_{j+1} + u_{j-1} - 2u_j}{h^2}.$$

If f is sufficiently smooth, then Taylor's theorem implies that the error in the approximation to the first derivative above is of the order $O(h)$ (meaning that it is bounded by a constant depending on the second derivative of f multiplied by h). The symmetric difference used in approximating the second derivative is better in the sense that the error is $O(h^2)$, with a constant depending on the fourth derivative of f. Since differentiation is linear, the map from the points $\mathbf{u}$ to the approximate derivatives is a linear one,

$$\mathbf{w} = D_N \mathbf{u},$$

where D_N is an $(N+1) \times (N+1)$ differentiation matrix.

Exercise 10.4. Use Taylor's theorem to verify the order of approximation statements above.

Exercise 10.5. In using the second derivative approximation above one will need the values u_{-1} and u_{N+1} to approximate the derivatives at the end points. Assume that these values always are equal to zero (Dirichlet boundary condition). Then the matrix implementing the second derivative is a Toeplitz matrix. Find it. If one instead assumes a periodic boundary condition, which means $u_{-1} = u_N$ and $u_{N+1} = u_0$, the matrix becomes a circulant matrix. Find also this matrix.

The finite difference method requires a large number of sample points to give a good approximation to the derivative, i.e. a small error. If one is willing to use an irregular grid for sampling, then there is a class of efficient methods called spectral differentiation methods. A particular case is the Chebyshev differentiation method. We fix the interval to be $[-1, 1]$ (this can always be obtained by a simple change of variables). The Chebyshev points are given by

$$x_j = \cos(\frac{j\pi}{N}), \quad j = 0, 1, 2, \ldots, N. \tag{10.9}$$

Note that these points cluster at the boundary points -1 and 1 for N large. One can visualize the points as the partition of the unit arc from 0 to π in N equal arcs, and the division points projected onto the interval $[-1,1]$, see Figure 18.

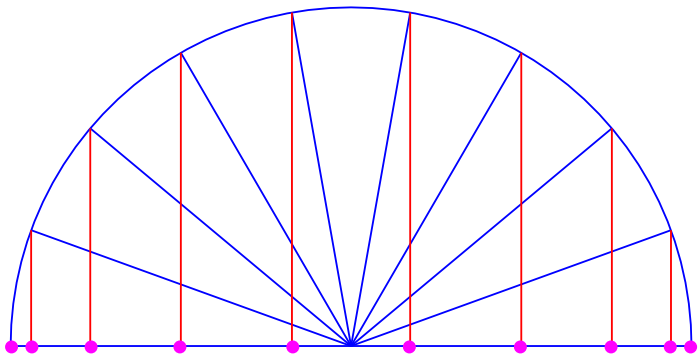

Figure 18: The Chebyshev points for $N = 9$

Note that we start the indexing from zero. In implementations vectors and matrices usually have to be indexed with the index sequence starting from one. This gives some extra, fairly trivial, bookkeeping to be taken care of.

Assume that we have grid points given by (10.9) and a sequence $\mathbf{u} = \{u_0, u_1, \ldots, u_N\}$. We compute the derivative sequence $\mathbf{w}$ as follows. Let p be the unique polynomial of degree N or less, which interpolates the points (x_j, u_j), $j = 0, \ldots, N$. This means that the polynomial satisfies

$$p(x_j) = u_j, \quad j = 0, \ldots, N.$$

Then we compute the derivative sequence as

$$w_j = p'(x_j), \quad j = 0, \ldots, N.$$

Exercise 10.6. Prove the existence and uniqueness of the interpolating polynomial of degree less than or equal to N for a sequence of $N+1$ points, as stated above.

Since differentiation is linear, there is a matrix D_N, the Chebyshev differentiation matrix, such that $\mathbf{w} = D_N\mathbf{u}$. We will not go through the derivation of the formula for D_N. The results can be found in [Tre00]. We state the result for reference. We have for the off-diagonal elements

$$(D_N)_{ij} = \frac{c_i}{c_j}\frac{(-1)^{i+j}}{x_i - x_j}, \quad i \neq j, \quad i, j = 0, 1, 2, \ldots, N. \tag{10.10}$$

Here $c_0 = c_N = 2$, and $c_i = 1$, $i = 1, \ldots, N-1$. The diagonal entries are determined by the requirement that the sum of each row is zero.

The last condition can easily be understood. If we take the sequence $u_j = 1$, $j = 0, 1, \ldots, N$, then the interpolating polynomial is the constant one $p(x) = 1$. Its derivative is of course zero. Thus for this sequence $\mathbf{u}$ we have

$$D_N \mathbf{u} = \mathbf{0},$$

or written for each entry

$$(D_N \mathbf{u})_i = \sum_{j=0}^{N} (D_N)_{ij} = 0.$$

Here are the first three Chebyshev differentiation matrices. For $N = 1$ we have $x_0 = 1$ and $x_1 = -1$, and

$$D_1 = \begin{bmatrix} \frac{1}{2} & -\frac{1}{2} \\ \frac{1}{2} & -\frac{1}{2} \end{bmatrix}.$$

For $N = 2$ we have $x_0 = 1$, $x_1 = 0$, and $x_2 = -1$, and

$$D_2 = \begin{bmatrix} \frac{3}{2} & -2 & \frac{1}{2} \\ \frac{1}{2} & 0 & -\frac{1}{2} \\ -\frac{1}{2} & 2 & -\frac{3}{2} \end{bmatrix}.$$

For $N = 3$ we have $x_0 = 1$, $x_1 = \frac{1}{2}$, $x_2 = -\frac{1}{2}$, and $x_3 = -1$. The matrix is given by

$$D_3 = \begin{bmatrix} \frac{19}{6} & -4 & \frac{4}{3} & -\frac{1}{2} \\ 1 & -\frac{1}{3} & -1 & \frac{1}{3} \\ -\frac{1}{3} & 1 & \frac{1}{3} & -1 \\ \frac{1}{2} & -\frac{4}{3} & 4 & -\frac{19}{6} \end{bmatrix}.$$

One can easily check that none of the three matrices is normal. We have the estimate

$$\|D_N\| > \frac{N^2}{3}.$$

The upper left hand corner in D_N can be shown to be $(2N^2 + 1)/6$, which establishes this result. See [Tre00].

We also have that all matrices D_N are nilpotent. More precisely, we have

$$(D_N)^{N+1} = 0.$$

This result is a simple consequence of the differentiation method, and the uniqueness of the interpolating polynomial. For any $\mathbf{u}$ we have that $(D_N)^{N+1}\mathbf{u}$ is the vector obtained by interpolating with respect to the samples $\mathbf{u}$ and then differentiate the interpolating polynomial $N + 1$ times. Since the polynomial is of degree at most N, this derivative is zero, irrespective of $\mathbf{u}$, which proves the result.

Exercise 10.7. Give the details in the argument leading to the conclusion that $(D_N)^{N+1} = 0$. Note that the uniqueness of the interpolating polynomial is essential in this argument.

Now if one tries to verify the nilpotency numerically, one rapidly runs into trouble. The same happens, if one wants to look at the behavior of $\|(D_N)^k\|$ as a function of k. The behavior using floating point computations is not at all the one that the mathematical result predicts, even for fairly small N.

We will now try to illustrate this and other phenomena. Maple is very convenient for making experiments, when the expected result depends on the computational precision, since we can vary this parameter, in contrast to MATLAB, where there is little control of the precision available to the user.

We take a small D_N, with $N = 5$. Thus we expect that $\|(D_5)^k\|$ will be zero for $k \geq 6$. Let us see how this works out numerically. We compute with three different precisions, which in Maple is given by the variable `Digits`. We take the values 8, 16, and 32. The results are shown in Figure 19. The same computations for $N = 10$ with the same values for the variable `Digits` are shown in Figure 20.

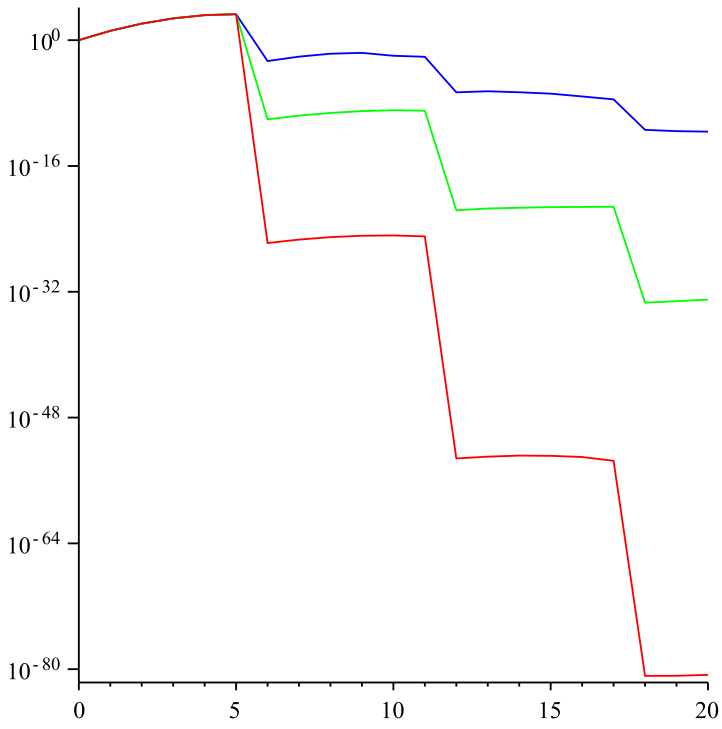

Figure 19: Plots of the powers $\|(D_5)^k\|$ for $k = 0, 1, \ldots, 20$, computed with three levels of precision, given by the Maple variable `Digits`. Blue curve: `Digits`=8, green curve: `Digits`=16, and red curve: `Digits`=32. Note that the vertical scale is logarithmic.

For $N = 5$ we see from Figure 19 that the norm decreases substantially for $k = 6$. The following iterations show the effects of the rounding errors and how the results are im-

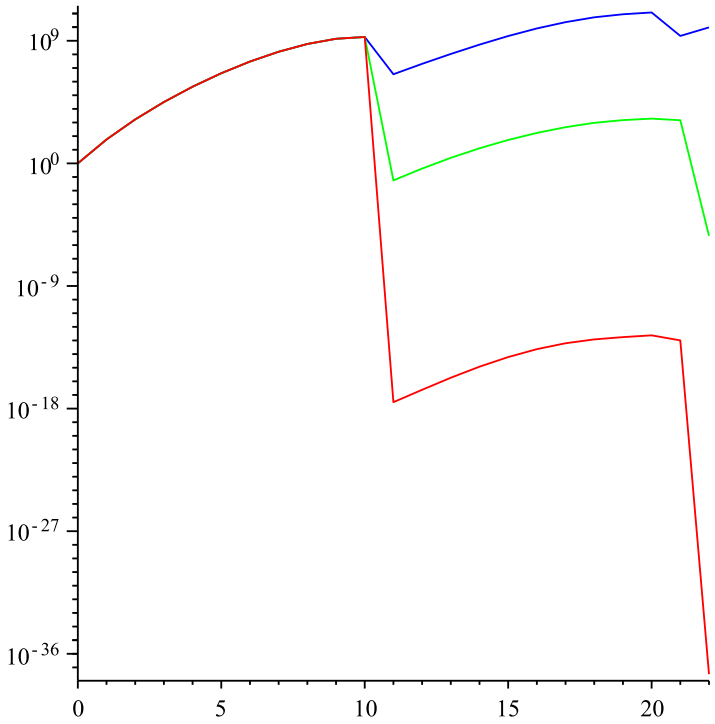

Figure 20: Plots of the powers $\|(D_{10})^k\|$ for $k = 0, 1, \ldots, 22$, computed with three levels of precision, given by the Maple variable `Digits`. Blue curve: `Digits`=8, green curve: `Digits`=16, and red curve: `Digits`=32. Note that the vertical scale is logarithmic.

proved by increasing the precision. The same comments apply to Figure 20. If we carry the computations to larger values for k in the case `Digits`=8 for D_{10}, we see exponential growth of the norms $\|D_{10})^k\|$. Recall that the spectrum of D_N always is $\sigma(D_N) = \{0\}$, since it is nilpotent. Thus the the numerical behavior is quite different from the mathematical one given in Theorem 9.3. The computations are shown in Figure 21.

Now let us see what we can learn about the matrices D_N by using pseudospectra. Due to the norm growth $\|D_N\| \asymp N^2$ it is preferable to use a scaled version of the matrices when computing the pseudospectra. Thus we take

$$\widetilde{D}_N = N^{-2} D_N$$

in our computations. The pseudospectra as computed in `EigTool` are shown in Figure 22, for the matrices $\widetilde{D}_{10}$ and $\widetilde{D}_{30}$. If we rescale by the N^2 factor, we see that the resolvent norm is large far from the spectrum $\sigma(D_N) = \{0\}$. Numerically the computed eigenvalues are not close to zero. We have $\rho(\widetilde{D}_{10}) \approx 2.865 \cdot 10^{-3}$ and $\rho(\widetilde{D}_{30}) \approx 1.061 \cdot 10^{-2}$.

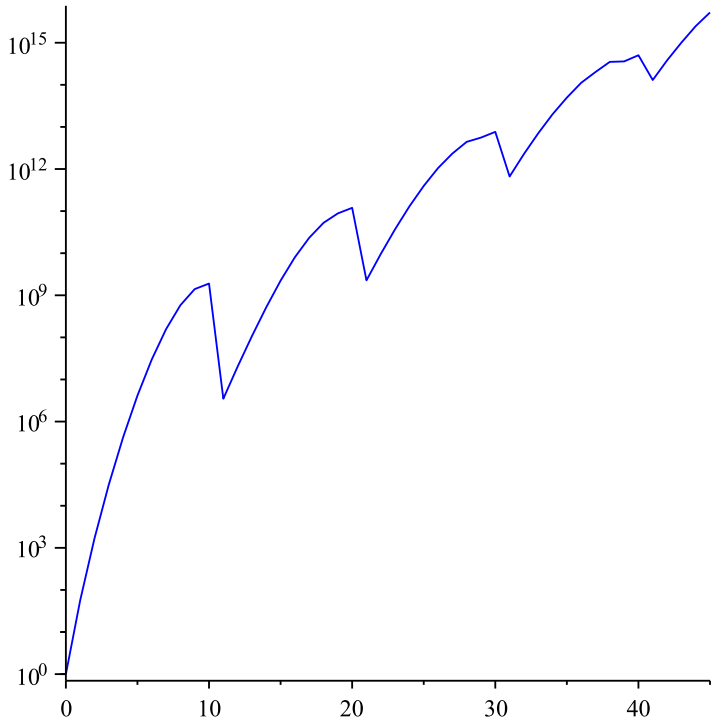

Figure 21: Plot of the powers $\|(D_{10})^k\|$ for $k = 0, 1, \ldots, 45$, computed with `Digits`=8. Note that the vertical scale is logarithmic.

11 Some infinite dimensional examples

In this section we will discuss some infinite dimensional examples. A full understanding of this section requires some knowledge of functional analysis and operator theory, see for example the books [Dav07, Kat95, Lax02, RS80]. We will not give all the technical details, in particular concerning the examples involving unbounded operators.

11.1 An infinite Toeplitz matrix

In this section we discuss an infinite Toeplitz matrix and its approximation by finite matrices. The Hilbert space we use is

$$\mathcal{H} = \ell^2(\mathbf{Z}) = \{\mathbf{x} = (x_n)_{n\in\mathbf{Z}} \mid \sum_{n=-\infty}^{\infty} |x_n|^2 < \infty\}$$

with the inner product given by

$$\langle \mathbf{y}, \mathbf{x} \rangle = \sum_{n=-\infty}^{\infty} \overline{y}_n x_n.$$

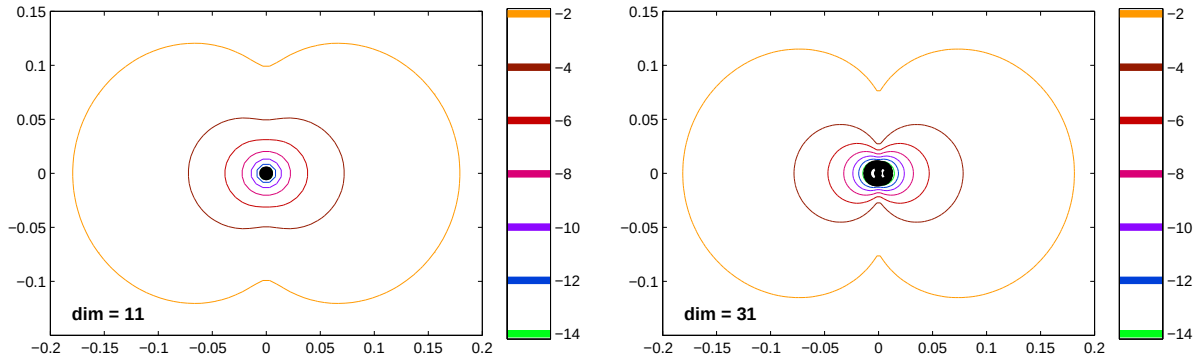

Figure 22: Pseudospectra of $\widetilde{D}_N$ for $N = 10$ (left hand part) and $N = 30$ (right hand part)

The canonical basis for $\ell^2(\mathbf{Z})$ is given by $\{\mathbf{e}_j\}_{j\in\mathbf{Z}}$, where

$$(\mathbf{e}_j)_n = \delta_{jn}.$$

Thus $\mathbf{e}_j$ is the doubly infinite sequence with the j^{th} entry equal to 1 and all other entries equal to zero.

We also need the Fourier transform $\mathcal{F}\colon \ell^2(\mathbf{Z}) \to L^2([-\pi,\pi])$ and its inverse. They are given by

$$(\mathcal{F}\mathbf{x})(\omega) = \frac{1}{\sqrt{2\pi}} \sum_{n=-\infty}^{\infty} x_k e^{ik\omega}, \tag{11.1}$$

$$(\mathcal{F}^{-1}f)_k = \frac{1}{\sqrt{2\pi}} \int_{-\pi}^{\pi} e^{-ik\omega} f(\omega)\, d\omega. \tag{11.2}$$

Note that another Fourier transform is defined in the next section, using the same symbol $\mathcal{F}$.

The sequence $\mathbf{a}$ is defined by

$$\mathbf{a} = 2\mathbf{e}_1 + \mathbf{e}_5,$$

such that

$$a_n = \begin{cases} 2 & \text{for } n = 1, \\ 1 & \text{for } n = 5, \\ 0 & \text{for } n \in \mathbf{Z} \setminus \{1,5\}. \end{cases}$$

We then define the operator A by $A\mathbf{x} = \mathbf{a} * \mathbf{x}$, where $*$ denotes convolution, such that

$$(A\mathbf{x})_n = \sum_{k=-\infty}^{\infty} a_{n-k}x_k = 2x_{n-1} + x_{n-5}, \quad n \in \mathbf{Z}. \tag{11.3}$$

The operator A is clearly a bounded operator on $\mathcal{H}$. Using the Fourier transform we find that

$$(\mathcal{F}A\mathcal{F}^{-1}f)(\omega) = (2e^{i\omega} + e^{5i\omega})f(\omega).$$

Thus A is unitarily equvalent with a multiplication operator. This implies that A is a normal operator, and also that the spectrum is given by

$$\sigma(A) = \{2e^{i\omega} + e^{5i\omega} \mid \omega \in [-\pi, \pi]\}. \tag{11.4}$$

The spectrum is shown in Figure 23.

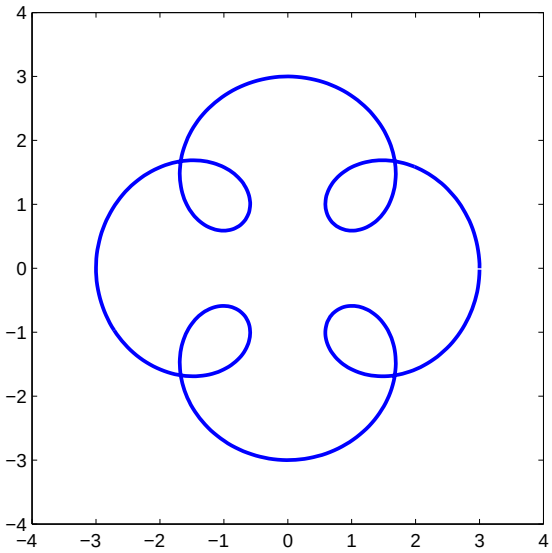

Figure 23: Spectrum of the operator A defined in (11.3).

The matrix of A with respect to the canonical basis is easily found from (11.3). Write the matrix as $[a_{jk}]$. Then $a_{j,j+1} = 2$, $a_{j,j+5} = 1$, and all other entries are zero. Thus it looks like

$$\begin{bmatrix}
\ddots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \ddots \\
\dots & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & \dots \\
\dots & 2 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & \dots \\
\dots & 0 & 2 & 0 & 0 & 0 & 0 & 0 & 0 & \dots \\
\dots & 0 & 0 & 2 & 0 & 0 & 0 & 0 & 0 & \dots \\
\dots & 0 & 0 & 0 & 2 & 0 & 0 & 0 & 0 & \dots \\
\dots & 1 & 0 & 0 & 0 & 2 & 0 & 0 & 0 & \dots \\
\dots & 0 & 1 & 0 & 0 & 0 & 2 & 0 & 0 & \dots \\
\dots & 0 & 0 & 1 & 0 & 0 & 0 & 2 & 0 & \dots \\
\ddots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \ddots
\end{bmatrix}.$$

This matrix is called a banded Toeplitz matrix. We define the subspace

$$\mathcal{H}_n = \{\mathbf{x} \in \ell^2(\mathbf{Z}) \mid x_k = 0 \text{ for all } |k| > n\}$$

and denote by A_n the matrix of the restriction of A to this subspace, again with respect to the canonical basis. Thus A_n is the truncated matrix. Actually we can consider any finite truncation of the matrix. Due to the banded structure of the matrix of A all these matrices will be finite banded Toeplitz matrices.

The numerical range of A is given by

$$W(A) = \text{conv}(\sigma(A)),$$

since A is normal, see Proposition 4.12. Then Theorem 4.13 shows that the numerical range of A can be approximated numerically using the numerical ranges of the truncated matrices A_n

The spectra of A_n do in no sense approximate the spectrum of A. This is easily seen since each A_n is lower triangular with zeroes on the main diagonal. Thus $(A_n)^n = 0$ for all n such that for all n we have $\sigma(A_n) = \{0\}$.Clearly the point zero and the spectrum of A are very different.

Let us now look at the pseudospectra. They approximate the spectrum in some sense, however it is not as good an approximation as is seen in other examples. In Figure 24 we show some pseudospectra in the case $n = 20$ (a 41×42 matrix), and in Figure 25 the computations are repeated for $n = 200$. The spectrum of the infinite matrix is shown in both figures, as a blue curve. The boundary of the numerical range of the finite matrix is shown as the dashed black curve. We note that in the case $n = 20$ the approximation to the convex hull of the spectrum of A is fairly good, but improves considerably for $n = 200$. For $n = 200$ we are getting close to the limits of what one can compute with `EigTool`. Note that the resolvent norm is larger than 10^{60} quite far from zero. Also note how far out the contour for $\varepsilon = 10^{10}$, when one goes from $n = 20$ to $n = 200$.

Exercise 11.1. Repeat the computations leading to Figure 24 and Figure 25. Try other values of n. Note that you have to modify several of the parameters in `EigTool` to get the figures shown here.

11.2 Advection-diffusion operator

Let $\mathcal{H} = L^2(\mathbf{R})$. This space is defined as

$$L^2(\mathbf{R}) = \{u : \mathbf{R} \to \mathbf{C} \mid \int_{-\infty}^{\infty} |u(x)|^2 dx < \infty\}.$$

The inner product is given by

$$\langle u, v \rangle = \int_{-\infty}^{\infty} \overline{u(x)} v(x) dx.$$

There are some technical matters concerning the integral, which has to be the Lebesgue integral, and identification of functions that differ on a set of Lebesgue measure zero, which we omit. The space $L^2(\mathbf{R})$ is a Hilbert space.

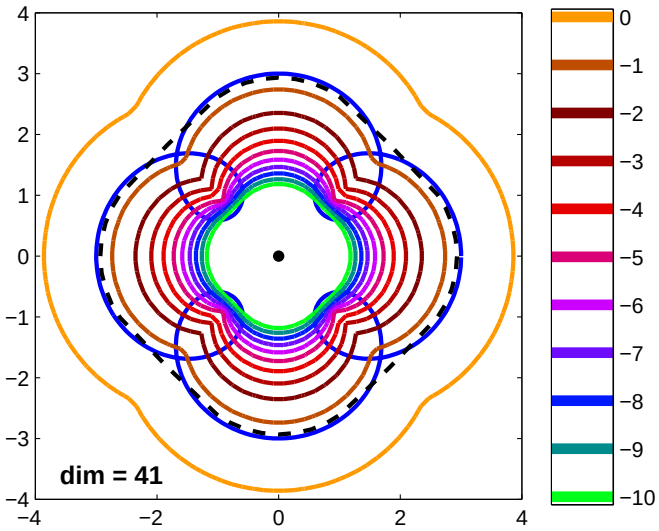

Figure 24: Plot of the pseudospectra of the Toeplitz matrix A_{20}. The black dashed curve is the boundary of the numerical range of A_{20}. The blue curve is the spectrum of the infinite dimensional Toeplitz matrix A, see Figure 23.

Let us recall the definition of the Fourier transform $\mathcal{F}\colon \mathcal{H} \to \mathcal{H}$ and its inverse. The Fourier transform and its inverse are given by

$$(\mathcal{F}u)(\xi) = \hat{u}(\xi) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} u(x) e^{-ix\xi} dx, \tag{11.5}$$

$$(\mathcal{F}^{-1}v)(x) = \check{v}(x) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} v(\xi) e^{ix\xi} d\xi. \tag{11.6}$$

Again, there are some technical details convergence of these integrals that we omit. The operator $\mathcal{F}$ is unitary, due to the choice of the constant in front of the integral.

We let D denote the differentiation operator, such that $Du(x) = u'(x)$. For $u \in \mathcal{H}$ the differentiation is in the sense of distributions. The operator we want to consider is given by

$$A^\eta = \eta D^2 + D, \quad \eta > 0. \tag{11.7}$$

It is defined on a dense subset of $\mathcal{H}$, consisting of all functions $u\mathcal{H}$ such that $Du, D^2u \in \mathcal{H}$. Let us explain a little about this operator. It occurs in several different places in mathematical physics. It is related to the description of advection-diffusion, which is sometimes also called convection-diffusion. The related time-dependent partial differential equation is

$$\frac{\partial f}{\partial t} = \eta \frac{\partial^2 f}{\partial x^2} + \frac{\partial f}{\partial x}.$$

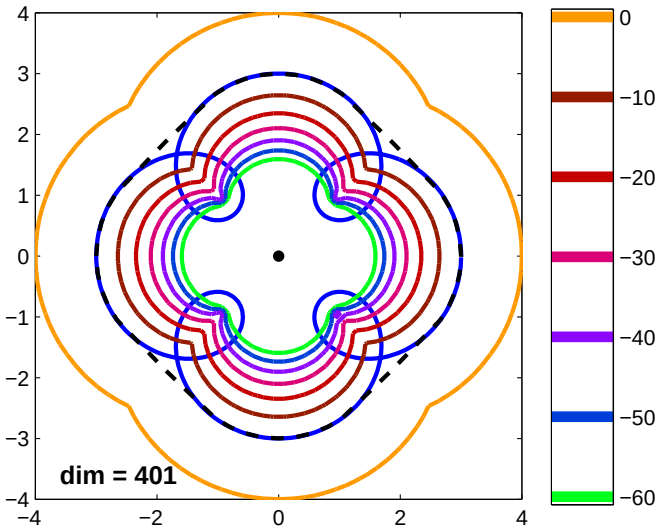

Figure 25: Plot of the pseudospectra of the Toeplitz matrix A_{200}. The black dashed curve is the boundary of the numerical range of A_{200}. The blue curve is the spectrum of the infinite dimensional Toeplitz matrix A, see Figure 23.

The parameter η is the diffusion strength. If η is small, it is the drift term (advection) that determines the behavior. The equation above can be solved by considering $f(t,x)$ as a function of t with values in $\mathcal{H}$. Writing $\phi(t) = f(t,\cdot)$ the equation can be thought of as the equation

$$\frac{d}{dt}\phi(t) = A^\eta \phi(t)$$

with the (formal) solution

$$\phi(t) = e^{tA^\eta}\phi(0).$$

All this can be made rigorous. See the references given above.

Thus it is clear that to understand the behavior of solutions to this partial differential equation a good understanding of the operator A^η is desirable, including its spectrum, its pseudospectra, and its numerical range.

To approximate numerically one could consider first a reduction to a finite interval, and then a discretization of the operator on the finite interval, using spectral methods, [Tre00]. We will carry out a number of steps in that direction and end up with some numerical experiments based on the Chebyshev differentiation matrix from Section 10.2.

The spectrum of A is easy to determine using the Fourier transform. We have that $\mathcal{F}A^\eta\mathcal{F}^{-1}v(\xi) = (-\eta\xi^2 + i\xi)v(\xi)$. Thus we have

$$\sigma(A^\eta) = \{-\eta\xi^2 + i\xi \mid \xi \in \mathbf{R}\}, \tag{11.8}$$

which is a parabola. This result also implies that A is a normal operator. The numerical range of a normal operator is the convex hull of its spectrum, as mentioned after Proposition 4.11, a result that also holds for unbounded normal operators. Thus we have the result

$$W(A^\eta) = \{x + iy \mid x \leq -\eta y^2\}, \tag{11.9}$$

i.e. the parabola and its interior.

Now let us reduce to a problem on a finite interval. We take as our space $\mathcal{H}^a = L^2([-a,a])$, $a > 0$. The operator A^η is restricted to this space with Dirichlet boundary conditions. This means that we consider A^η restricted to the dense set $C_0^\infty((-a,a))$ (the set of smooth functions with support in the open interval $(-a,a)$) and then take what is called the closure of this operator. The closure is denoted by $A^{\eta,a}$. Informally stated the functions in the domain of $A^{\eta,a}$ must satisfy the two boundary conditions $u(-a) = 0$ and $u(a) = 0$. Details can be found in [Dav07, Lax02].

Now we determine the spectrum of $A^{\eta,a}$. We will use a change of dependent variable to do this. Let M be the operator of multiplication by $\exp(-x/(2\eta))$. It is a bounded operator on $\mathcal{H}^a$ wit the inverse equal to multiplication by $\exp(-x/(2\eta))$. Thus we have

$$M^{-1}A^{\eta,a}M = \eta D^2 - \frac{1}{4\eta}I = T^\eta$$

by a straightforward calculation. The spectrum is preserved under a similarity transform. We solve the eigenvalue problem for T^η. Written explicitly as an ordinary differential equation problem, we have to solve the problem

$$\eta u''(x) - \frac{1}{4\eta}u(x) = \lambda u(x), \quad u(-a) = u(a) = 0.$$

The solution of this problem is completely elementary. One finds a sequence of eigenvalues

$$\lambda_k = -\frac{1}{4\eta} - \frac{\eta\pi^2 k^2}{4a^2}, \quad k = 1,2,3,\ldots, \tag{11.10}$$

and corresponding normalized eigenfunctions

$$v_k(x) = \begin{cases} \dfrac{1}{\sqrt{a}}\cos(\pi k x/(2a)), & \text{for } k \text{ odd}, \\ \dfrac{1}{\sqrt{a}}\sin(\pi k x/(2a)), & \text{for } k \text{ even}. \end{cases}$$

Using Fourier analysis one can verify that the functions $\{v_k\}_{k\in\mathbf{N}}$ form an orthonormal basis of $\mathcal{H}_a$. Thus we have shown that

$$\sigma(A^{\eta,a}) = \left\{ -\frac{1}{4\eta} - \frac{\eta\pi^2 k^2}{4a^2} \;\middle|\; k = 1,2,3,\ldots \right\} \tag{11.11}$$

The corresponding eigenfunctions (not normalized) are given by

$$v_k(x) = \begin{cases} e^{-x/(2\eta)} \cos(\pi k x/(2a)), & \text{for } k \text{ odd}, \\ e^{-x/(2\eta)} \sin(\pi k x/(2a)), & \text{for } k \text{ even}. \end{cases} \tag{11.12}$$

Exercise 11.2. Carry out the details in the computations leading to (11.11) and (11.12).

Now the remarkable fact is that the spectra $\sigma(A^{\eta,a})$ do not converge or even approximate the spectrum of $\sigma(A^\eta)$ for large a. The spectra are all a sequence of points on the negative real axis, whereas the spectrum of A^η is given by the parabola in (11.8).

However, the numerical ranges of $A^{\eta,a}$ do approximate the numerical range of A^η given in (11.9), due to Theorem 4.13.

We now look at what we can say about the pseudospectra in general. We have the following result, where we for simplicity will take $\eta = 1$. Let us introduce the notation

$$\mathcal{P} = \{x + iy \mid x < -y^2\}.$$

Now given $\varepsilon > 0$ and $\lambda \in \mathcal{P}$, there exists $a_0 > 0$, such that $\lambda \in \sigma_\varepsilon(A^{1,a})$. The constant a_0 depends on ε and λ. Thus the pseudospectra 'fill up' the interior of the parabola (11.8) for $\eta = 1$, i.e. the spectrum of A^1.

To prove this result we will use the characterization of the pseudospectra given in Theorem 5.2(iii). Let r_1 and r_2 denote the roots of the polynomial $z^2 + z - \lambda$, labelled in such a manner that $\operatorname{Re}(r_2 - r_1) \le 0$. It is a tedious elementary exercise to verify that for $\lambda \in \mathcal{P}$ we have $\operatorname{Re} r_1 < 0$ and $\operatorname{Re} r_2 < 0$.

Choose a small $\delta > 0$ and a function $\chi \in C^\infty(\mathbf{R})$, such that $0 \le \chi(x) \le 1$ for all $x \in \mathbf{R}$, and

$$\chi(x) = \begin{cases} 1 & \text{for } x < -\delta, \\ 0 & \text{for } x \ge 0. \end{cases}$$

We now define

$$\begin{aligned} \phi_a(x) &= e^{r_1(x+a)} - e^{r_2(x+a)}, \\ \widetilde{\psi}_a(x) &= \chi(x-a)\phi_a(x), \\ \psi_a(x) &= \frac{1}{c_a}\widetilde{\psi}_a(x), \end{aligned}$$

where

$$c_a = \Big(\int_{-a}^{a} |\widetilde{\psi}_a(x)|^2 dx \Big)^{1/2}.$$

With these definitions we have

$$\psi_a(-a) = \psi_a(a) = 0 \quad \text{and} \quad \|\psi_a\|_{\mathcal{H}^a} = 1.$$

This means that ψ_a is a smooth function satisfying the boundary conditions, so it belongs to the domain of $A^{1,a}$. Furthermore it is normalized.

A computation shows that we have

$$\|A^{1,a}\psi_a - \lambda\psi_a\| \le ce^{\operatorname{Re} r_1(2a-\delta)}. \tag{11.13}$$

Thus we can determine an $a_0 > 0$ such that for $a > a_0$ we have $ce^{\operatorname{Re} r_1(2a-\delta)} < \varepsilon$, which verifies Theorem 5.2(iii).

Since the estimate (11.13) is not quite straightforward, we will give some of the details. First one notices that due to $\operatorname{Re} r_1 < 0$, $\operatorname{Re} r_2 < 0$, and the choice of ϕ_a, we get that

$$0 < \int_{-\infty}^{\infty} |\widetilde{\psi}_a(x)|^2 dx < \infty.$$

Thus we can determine $a_1 > 0$ and $0 < C_1 < C_2 < \infty$ such that for $a > a_1$ we have

$$C_1 \le c_a \le C_2.$$

Next we use that r_1 and r_2 are the roots in the polynomial $z^2 + z - \lambda$ to get that $A^{1,a}\phi_a = \lambda\phi_a$. This leads to the result

$$A^{1,a}\widetilde{\psi}_a(x) = \lambda\widetilde{\psi}_a(x) + \chi'(x-a)\phi_a(x) + 2\chi'(x-a)\phi_a'(x) + \chi''(x-a)\phi_a(x).$$

By construction the functions $\chi'(x-a)$ and $\chi''(x-a)$ are nonzero only for $a-\delta \le x \le a$. Therefore we get an estimate

$$|\phi_a(x)| \le 2e^{\operatorname{Re} r_1(2a-\delta)}, \quad x \in [a-\delta, a].$$

which leads to

$$|A^{1,a}\widetilde{\psi}_a(x) - \lambda\widetilde{\psi}_a(x)| \le ce^{\operatorname{Re} r_1(2a-\delta)}, \quad x \in [a-\delta, a],$$

and then

$$|A^{1,a}\psi_a(x) - \lambda\psi_a(x)| \le \frac{c}{C_1}e^{\operatorname{Re} r_1(2a-\delta)}, \quad x \in [a-\delta, a].$$

This estimate implies (11.13).

We will now discuss how to approximate $A^{\eta,a}$ by a matrix, by using a discretization. We here use the spectral method discussed in Section 10.2, the Chebyshev differentiation matrix. As formulated in Section 10.2 this method is modelled on the differentiation operator on the interval $[-1,1]$. Thus we should carry out a change of variables to get from the interval $[-a,a]$ to the interval $[-1,1]$. The map

$$\Phi\colon L^2([-a,a]) \to L^2([-1,1]), \quad (\Phi u)(x) = \sqrt{a}u(ax)$$

is unitary and performs the change of variables. We have

$$\Phi A^{\eta,a}\Phi^{-1} = \frac{1}{a}A^{\eta/a,1},$$

as can be seen by a straightforward computation. Thus large a results can be obtained by choosing η sufficiently small, and then scaling the results. Using the result from Exercise 5.15 and the fact that Φ is unitary, we get that

$$\sigma_\varepsilon(A^{\eta,a}) = \frac{1}{a}\sigma_{\varepsilon/a}(A^{\eta/a,1}).$$

Given the Chebyshev differentiation matrix (10.10), we can impose the Dirichlet boundary condition by deleting the first row. the first column, the last row, and the last column, i.e. in MATLAB notation we take the matrix $D(2:N,2:N)$ as the discretized differentiation operator. Note that for a given N the matrix from (10.10) is an $(N+1)\times(N+1)$ matrix, such that after deletion we end up with an $(N-1)\times(N-1)$ matrix.

In figure Figure 26 we show the result of a computation using `EigTool`, in the case where $N = 40$ in (10.10) and $\eta = 1/30$. The black dashed curve is the parabola given in (11.8). The results of similar computations for $N = 100$ and $\eta = 1/30$ are shown in Figure 27, with the right hand part enlarged in Figure 28.

These numerical results again show that the pseudospectra may give good information on the location of the spectrum of the infinite model.

Exercise 11.3. Verify the computations given above.

Exercise 11.4. Carry out further numerical experiments for various values of η.

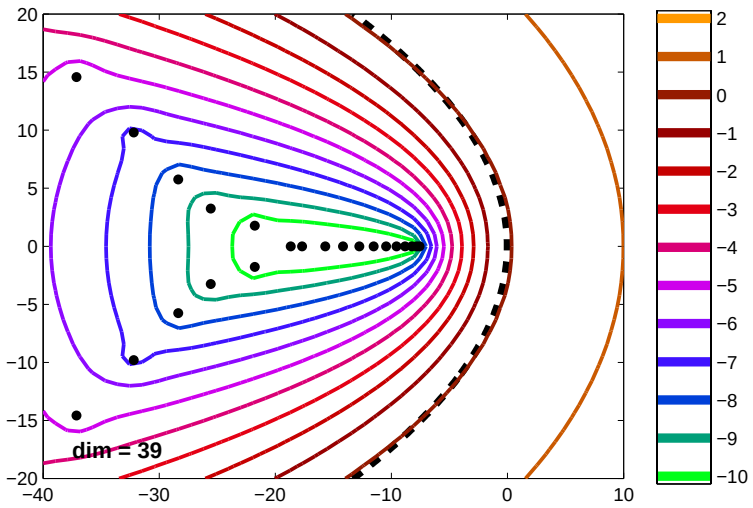

Figure 26: Plot of the pseudospectra of the discretized $A^{\eta,1}$ for $\eta = 1/30$ and $N = 40$ in (10.10). The black dashed curve is the parabola (11.8).

These numerical results again show that the pseudospectra may give good information on the location of the spectrum of the infinite model.

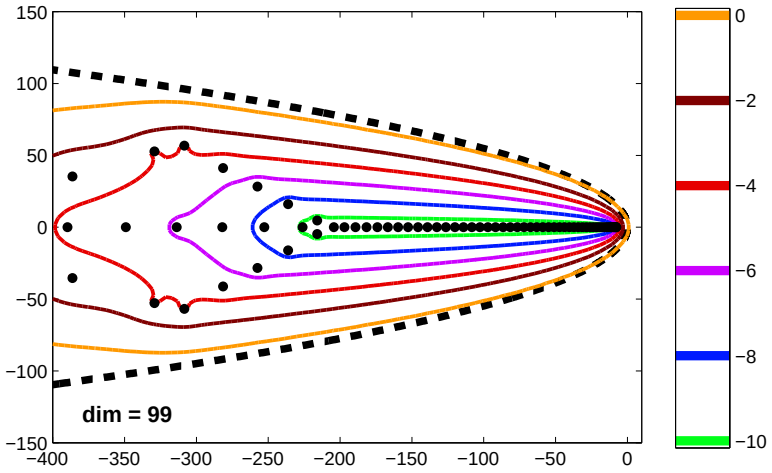

Figure 27: Plot of the pseudospectra of the discretized $A^{\eta,1}$ for $\eta = 1/30$ and $N = 100$ in (10.10). The black dashed curve is the parabola (11.8).

Exercise 11.5. Verify the computations given above.

Exercise 11.6. Carry out further numerical experiments for various values of η.

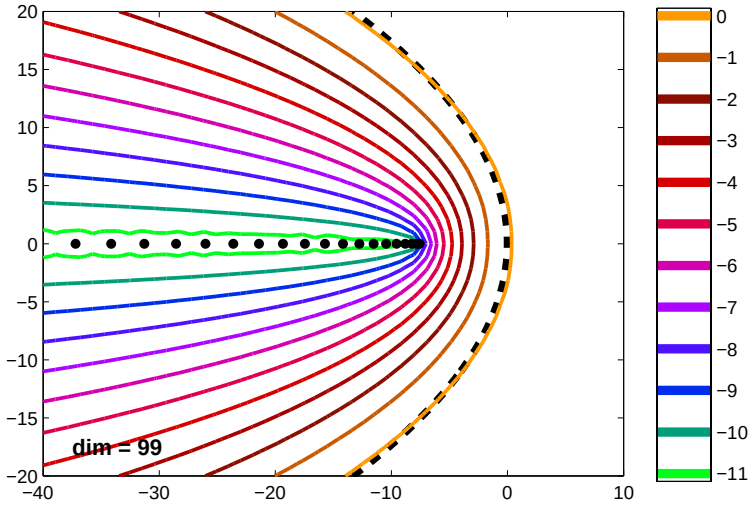

Figure 28: Enlargement of the right hand part of Figure 27

References

[Con78] John B. Conway, *Functions of one complex variable*, second ed., Graduate Texts in Mathematics, vol. 11, Springer-Verlag, New York, 1978.

[Dav07] E. Brian Davies, *Linear Operators and their Spectra*, Cambridge University Press, Cambridge, UK, 2007.

[Kat95] Tosio Kato, *Perturbation Theory for Linear Operators*, Classics in Mathematics, Springer-Verlag, Berlin, 1995, Reprint of the 1980 edition.

[Lax02] Peter D. Lax, *Functional analysis*, Pure and Applied Mathematics (New York), Wiley-Interscience [John Wiley & Sons], New York, 2002.

[RS80] Michael Reed and Barry Simon, *Methods of Modern Mathematical Physics I: Functional Analysis*, Academic Press, New York, 1980.

[Str06] Gilbert Strang, *Linear Algebra and Its Applications*, fourth ed., Thomson Brooks/Cole, Belmont CA, USA, 2006.

[TE05] Lloyd N. Trefethen and Mark Embree, *Spectra and Pseudospectra*, Princeton University Press, Princeton and Oxford, 2005.

[Tre00] Lloyd N. Trefethen, *Spectral Methods in* MATLAB, SIAM, Philadelphia, PA, USA, 2000.