Pseudospectra

and

the numerical solution of differential equations

Mavroudis Ioannidis

Master's Thesis

Supervisor: Claus Führer

Examiner: Tony Stillfjord

LUND
UNIVERSITY

Faculty of Science
Department of Mathematics
Sweden
September 2021

Abstract

Eigenvalue analysis has been a key tool to science and engineering for several decades. Eigenvalues can predict the behaviour of many mathematical systems of equations but alone they cannot fully explain phenomena such as stability or stiffness. Together, eigenvalues and pseudospectra can give a better understanding of several phenomena such as instability in nonnormal matrices or operators.

In this thesis, the basic concepts of pseudospectra are utilized to assist in understanding how pseudospectra can better explain the stability of PDE discretizations, the stability of the method of lines, the stiffness of ODEs and the GKS-stability of boundary conditions.

It has been based on the book of Lloyd N. Trefethen and Mark Embree, "Spectra and Pseudospectra The Behavior of Nonnormal Matrices and Operators" [Trefethen and Embree, 2005]. Despite that, it tries to explain in a more analytical manner certain points of the book. Using also other references, we attempt to clarify some more aspects of pseudospectra. The code for the figures has been based on Lloyd N. Trefethen [Trefethen, 1999] and is presented in the Appendix. For more on MATLAB codes for solving problems using spectra and pseudospectra, see [Trefethen, 2000].

Keywords: Pseudospectra, numerical solutions of differential equations, spectral differentiation matrices, Lax-stability, stability of MOL, stiffness, GKS-stability

Acknowledgements

I would like to thank my thesis advisor, Claus Führer, for his guidance. He not only helped me with choosing the thesis topic, but he also allowed this paper to be my work. He gave me valuable support whenever I needed it.

I would also like to acknowledge the examiner, Tony Stillfjord, for his insightful comments on this thesis.

Finally, I would like to thank my friends for their help and support throughout the process of writing this thesis.

Thank you.

Contents

List of Figures

Introduction

In different fields of applied mathematics, eigenvalue analysis has been an important part. This is true when the matrices are normal or close to normal. Eigenvalues alone fail to explain the behaviour of matrices when the matrices are far from normal. In this thesis, we introduce and explain how pseudospectra can fill in the gap that occurs in nonnormal cases.

In Chapter 1, we present an example where eigenvalues fail to predict the stability of a numerical method for solving an ordinary differential equation. We define the terms *pseudospectrum, pseudoeigenvalue, pseudoeigenvector* and give an example on pseudospectra.

Differential equations are the most common type of mathematical model in fields of quantitative study and especially in science and engineering. The most common categories differential equations are classified into are the *ordinary differential equations* (ODE) and the *partial differential equations* (PDE).

The numerical solution of differential equations is implicitly connected to matrices. This is due to the fact that discretizing a differential equation (for example on a time-space grid), a linear operator appears, which of course can be expressed in a form of a matrix. Nonetheless, in many cases, the matrices or operators are not close to normal.

Because of the limitations of the eigenvalues to explain not expected phenomena such as stability and stiffness, we use the pseudospectra to better explain these phenomena. In Chapter 2, we explain, through a series of examples, how the pseudospectra can give a different perspective in concepts such as stability and stiffness.

Although computation of pseudospectra can be applied on small matrices fast, there are some difficulties when the matrices are large. In Chapter 3, we introduce ideas to speed up the computation of pseudospectra.

Chapter 1

Pseudospectra

1.1 Spectra of matrices

Eigenvalues are one of the most important tools in mathematics. Given an $N \times N$ matrix A with complex coefficients, a nonzero complex column vector v and a complex number λ, v is an *eigenvector* of A and λ its corresponding *eigenvalue*, if $Av = \lambda v$. The spectrum of A is the set of all its eigenvalues, denoted by $\sigma(A)$. The spectrum is also defined as the set of $z \in \mathbb{C}$ where the *resolvent* matrix $(z - A)^{-1}$ does not exist. We use $z - A$ as shorthand for $zI - A$, where I is the identity.

For matrix A, let Λ be the diagonal matrix of its eigenvalues and V the matrix of all its corresponding eigenvectors. Assuming all eigenvalues are distinct, all eigenvectors are linearly independent and V is nonsingular. Then, we can *diagonalize* matrix A, $AV = V\Lambda \Rightarrow A = V\Lambda V^{-1}$.

An important property that a matrix may have is normality. The reason is that the matrix V that diagonalizes a normal matrix may become unitary after normalization. Thus, normal matrices have good behaviour in many cases.

Definition. *A square matrix A is normal if $AA^* = A^*A$.*

A normal matrix has a complete set of orthogonal eigenvectors. If every eigenvector is normalized, the matrix V becomes *unitary* (or *orthogonal* in the real case) with $V^{-1} = V^*$, where V^* is the conjugate transpose of V and $\|V\|_2 = \|V^{-1}\|_2 = 1$. Note that a Hermitian matrix, $A = A^*$, is a special case of a normal matrix.

Suppose now that V is invertible and $\|\cdot\|$ a matrix norm induced by a vector norm. The *condition number* of V with respect to the norm $\|\cdot\|$ is defined as $\kappa(V) = \|V\|\|V^{-1}\|$. Note that $\kappa(V) \geq \|VV^{-1}\| = \|I\| = 1$. In the case of a normal matrix A, if we assume that V is the matrix with the eigenvectors of A as columns and $\|\cdot\| = \|\cdot\|_2$, then V and V^{-1} are unitary, $V^* = V^{-1}$ and $\kappa(V) = \|V\|\|V^{-1}\| = 1$. ($\|V\| = \|V^{-1}\| = 1$ because a unitary matrix preserves the Eucleidian norm, i.e $\|Vx\| = \|x\|$ for every vector x). On the other hand a matrix A such that $\kappa(V) >> 1$, is considered to be 'far from normal'. So, the condition number is somewhat a measure of the distance of the matrix from normality.

Eigenvalues give insight into how a system behaves, but this does not hold always when a 'far from normal' matrix appears. We will now state a case where an eigenvalue stability condition fails. It is from an article by J. L. M. Van Dorsselaer, J. F. B. M. Kraaijevanger and M. N. Spijker [Van Dorsselaer et al., 1994].

Consider a large system of ordinary differential equations of the form $U'(t) = AU(t) + b(t)$ with $A \in \mathbb{C}^{N \times N}$ (A is t-independent) and initial conditions $U(0) = u_0$. Applying a Runge-Kutta numerical method with step size equal to $h > 0$, we arrive at a discrete process $u_n = \phi(hA)u_{n-1} + b_n$ to approximate U, i.e. u_n approximates $U(nh)$. The rational function ϕ depends on the particular Runge-

Kutta method under consideration, but not on the given A, b or the initial conditions. We define the method as *stable* if a small perturbation equal to v_0 on the initial conditions implies a small perturbation on v_n for every $n \geq 1$. For the rest of the section $\| \cdot \|$ stands for $\| \cdot \|_2$.

First, note that $v_{n+1} = \phi(hA)v_n$, therefore $v_n = \phi(hA)^n v_0$, $n \geq 1$. The eigenvalue condition for stability states that if $\sigma(hA) \subseteq int(S)$, where S is the set of $z \in \mathbb{C}$ such that $\phi(|z|) \leq 1$, then $\|\phi(hA)^n\|$ are uniformly bounded by a constant c_0. (S is called the *stability region* of the method). Indeed, if $\sigma(hA) \subseteq int(S)$, then using the spectral mapping theorem, one has $\sigma(\phi(hA)) = \phi(\sigma(hA)) \subseteq \phi(int(S)) \subseteq D$, hence $lim_{n\to\infty}\|\phi(hA)^n\| = 0$.

Despite the fact this result seems to be satisfactory, c_0 is not always small enough for the process to be considered as stable in practice. If the matrix composed of the generalised eigenvectors of hA has a large condition number, then c_0 may be huge, and this is something that really appears in practical problems and not a seldom pathological situation.

1.2 Pseudospectra of matrices

We now come to the term *pseudospectrum.* We give more than one definitions and then prove their equivalence.

Definition (First definition of pseudospectra)**.** *Let A be an $N \times N$, complex square matrix and $\varepsilon > 0$.The set of all $z \in \mathbb{C}$ such that $\| (z - A)^{-1} \| > \varepsilon^{-1}$, denoted (temporarily, for purposes of clearness) by $\sigma_\varepsilon^1(A)$ is the $\varepsilon-$pseudospectrum of A, where we consider that if $z - A$ is not invertible (i.e. $z \in \sigma(A)$), then $\|(z - A)^{-1}\| = +\infty$.*

For normal matrices, the norm of the resolvent is large when z is around an eigenvalue of A. For matrices with a very large condition number, the resolvent may be large even when z is far from any eigenvalue of A. Another definition of the $\varepsilon -$ *pseudospectrum* is related to perturbation theory.

Definition (Second definition of pseudospectra)**.** *The $\varepsilon -$ pseudospectrum of A is the set $\sigma_\varepsilon^2(A)$ of $z \in \mathbb{C}$ such that $z \in \sigma(A + E)$ for some N-dimensional, complex, square matrix E with $\|E\| < \varepsilon$.*

Definition (Third definition of pseudospectra)**.** *The $\varepsilon -$ pseudospectrum of A is the set $\sigma_\varepsilon^3(A)$ of $z \in \mathbb{C}$ such that $\| (z - A) v\| < \varepsilon$ for some $v \in \mathbb{C}^N$ with $\|v\| = 1$.*

Theorem (Equivalence of the definitions of pseudospectra)**.** *For any matrix $A \in \mathbb{C}^{N\times N}$, the three definitions of $\varepsilon -$ pseudospectra are equivalent, therefore we will use the notation $\sigma_\varepsilon(A)$ from now on.*

Proof. $\sigma^2_\varepsilon(A) \subseteq \sigma^3_\varepsilon(A)$

Consider $z \in \sigma^2_\varepsilon(A)$. Suppose that $(A+E)v = zv$ for some $E \in \mathbb{C}^{N\times N}$ with $\|E\| < \varepsilon$ and some nonzero $v \in \mathbb{C}^N$ and $\|v\| = 1$. Then $\|\,(z-A)\,v\| = \|Ev\| < \varepsilon$.

$\sigma^3_\varepsilon(A) \subseteq \sigma^1_\varepsilon(A)$

Consider $z \in \sigma^3_\varepsilon(A)$. If $z \in \sigma(A)$, then $\|(z-A)^{-1}\| = +\infty > \varepsilon^{-1}$. If $z \notin \sigma(A)$, then suppose $(z-A)\,v = su$ for some $v, u \in \mathbb{C}^N$ with $\|v\| = \|u\| = 1$ and $s < \varepsilon$. Then $(z-A)^{-1}\,u = s^{-1}v$, so $\|\,(z-A)^{-1}\,\| \geq s^{-1} > \varepsilon^{-1}$.

$\sigma^1_\varepsilon(A) \subseteq \sigma^2_\varepsilon(A)$

Consider $z \in \sigma^1_\varepsilon(A)$. If $z \in \sigma(A)$, then $z \in \sigma(A+E)$ with E the N-dimensional zero matrix. If $z \notin \sigma(A)$ and $\|\,(z-A)^{-1}\,\| > \varepsilon^{-1}$, then $(z-A)^{-1}\,u = s^{-1}v$ and consequently $zv - Av = su$ for some $v, u \in \mathbb{C}^N$ with $\|v\| = \|u\| = 1$ and $s < \varepsilon$. To establish that $z \in \sigma(A+E)$ it is enough to show that there exists a matrix $E \in \mathbb{C}^{N\times N}$ with $\|E\| = s$ and $Ev = su$. Then v is an eigenvector of $A+E$ with eigenvalue z. E can be taken to be a rank-1 matrix of the form $E = suw^*$ for some $w \in \mathbb{C}^N$ with $w^*v = 1$. If $\|\cdot\|$ is the 2-norm, this is evident by taking $w = v$. In the case of an arbitrary norm $\|\cdot\|$, Hahn-Banach theorem guarantees the existence of a bounded linear functional L on $\mathbb{C}^N$ with $\|Lv\| = 1$ and $\|L\| = 1$. Left multiplication by w corresponds to the operation of L. □

From these definitions follows that the pseudospectra are nested sets. $\sigma_{\varepsilon_1}(A) \subseteq \sigma_{\varepsilon_2}(A), 0 < \varepsilon_1 \leq \varepsilon_2$ and that the intersection of all pseudospectra is the spectrum, $\bigcap_{\varepsilon>0} \sigma_\varepsilon(A) = \sigma(A)$.

The elements of $\varepsilon -$ *pseudospectrum* are called the $\varepsilon -$ *pseudoeigenvalues* z and v is a corresponding $\varepsilon -$ *pseudoeigenvector* to z.

In the case of a complex Hilbert space, we equip it with the inner product $(u, v) = v * \bar{u}$ and $\|v\| = \|v\|_2 = \sqrt{v * \bar{v}}$. The norm of a matrix is its largest singular value and the norm of its inverse is the inverse of its smallest singular value. $\| (z - A)^{-1} \| = [s_{min} (z - A)]^{-1}$, where $s_{min} (z - A)$ is the smallest singular value of $z - A$. This leads to a fourth and last definition of $\varepsilon -$*pseudospectra*.

Definition (Fourth definition of pseudospectra)**.** *For* $\| \cdot \| = \| \cdot \|_2$*,* $\sigma_\varepsilon(A)$ *is the set of* $z \in \mathbb{C}$ *such that* $s_{min} (z - A) < \varepsilon$*.*

This definition is equivalent to the previous three in the complex Hilbert space equipped with the inner product.

We now define the condition number of an eigenvalue of a matrix. Suppose $A \in \mathbb{C}^{N \times N}$ is a matrix with N distinct eigenvalues. This implies the existence of left and right eigenvectors determined up to scaling. $\mathbf{u}_j^* A = \lambda_j \mathbf{u}_j^*$, $A\mathbf{v}_j = \lambda_j \mathbf{v}_j$ for $j = 1, \ldots, N$. The condition number of λ_j is defined as $\kappa(\lambda_j) = \frac{\|\mathbf{u}_j\| \|\mathbf{v}_j\|}{|\mathbf{u}_j^* \mathbf{v}_j|}$. The condition number is 1, when $\|\mathbf{u}_j\| \|\mathbf{v}_j\| = |\mathbf{u}_j^* \mathbf{v}_j|$, i.e., when $\mathbf{u}_j$ and $\mathbf{v}_j$ are collinear (for equality to hold in Cauchy-Schwarz inequality). This is always true when A is a normal matrix since left and right eigenvectors can be taken to be the same.

Are pseudospectra affected under unitary similarity transformations? Simply note that $\| (z - UAU^*)^{-1} \| = \| [U (z - A) U^*]^{-1} \| = \| U (z - A)^{-1} U^* \| = \| (z - A)^{-1} \|, \forall z \in \mathbb{C}$. Thus, the resolvent norm is invariant to the unitary similarity matrix U. That implies, $\sigma_\varepsilon(A) = \sigma_\varepsilon(UAU^*), \forall \varepsilon \geq 0$.

An important characterization of normality is the following: A matrix $A \in \mathbb{C}^{N \times N}$ is normal if and only if it has a complete set of orthogonal eigenvectors, that is, if it is unitarily diagonalizable, $A = U \Lambda U^*$, where U is unitary and Λ is a diagonal matrix of eigenvalues.

We will now prove a theorem that geometrically connects the spectrum with the $\varepsilon - pseudospectra$. Consider the Minkowski sum $\sigma(A) + B_\varepsilon = \{z \in \mathbb{C} : z = z_1 + z_2, z_1 \in \sigma(A), z_2 \in B_\varepsilon\} = \{z \in \mathbb{C} : \text{dist}(z, \sigma(A)) < \varepsilon\}$, where $B_\varepsilon = \{z \in \mathbb{C} : |z| < \varepsilon\}$. We will use the following observation: For a normal matrix, all eigenvalues have condition number 1; equivalently the resolvent norm satisfies $\| \, (z - A)^{-1} \, \| = \frac{1}{dist\big(z, \sigma(A)\big)}$, where $dist\left(z, \sigma(A)\right)$ is the distance of point to a set in the complex plane, i.e. $dist\left(z, \sigma(A)\right) = inf_{\zeta \in \sigma(A)} dist(z, \zeta)$.

Theorem. *For any $A \in \mathbb{C}^{N \times N}$, $\sigma_\varepsilon(A) \supseteq \sigma(A) + B_\varepsilon, \forall \varepsilon > 0$*
If $\| \cdot \| = \| \cdot \|_2$, then, A is normal if and only if $\sigma_\varepsilon(A) = \sigma(A) + B_\varepsilon, \forall \varepsilon > 0$.

Proof. If z is an eigenvalue of A, then $z + \delta$ is an eigenvalue of $A + \delta$ for any $\delta \in \mathbb{C}$; since $\|\delta \mathbf{I}\| = |\delta|$, this establishes $\sigma_\varepsilon(A) \supseteq \sigma(A) + B_\varepsilon, \forall \varepsilon > 0$.

If A is normal, it can be assumed without loss of generality to be diagonal without any effect on norms if $\| \cdot \| = \| \cdot \|_2$, with diagonal elements a_{ij} equal to the eigenvalues λ_j. In this case the resolvent is also diagonal which implies that it satisfies $\| \, (z - A)^{-1} \, \| = \frac{1}{dist\big(z, \sigma(A)\big)}$ and $\| \, (z - A)^{-1} \, \| > \varepsilon^{-1}$ implies that this is equivalent to $\sigma_\varepsilon(A) \subseteq \sigma(A) + B_\varepsilon, \forall \varepsilon > 0$. Combining the above, we have that if A is normal, then $\sigma_\varepsilon(A) = \sigma(A) + B_\varepsilon$.

For the converse, $\sigma_\varepsilon(A) = \sigma(A) + B_\varepsilon, \forall \varepsilon > 0$ implies that each eigenvalue of A has condition number 1. If $\| \cdot \| = \| \cdot \|_2$, an eigenvalue having condition number 1 means (as mentioned before) that $\mathbf{u}_j$ and $\mathbf{v}_j$ are collinear, therefore each right eigenvector of A is also a left eigenvector. This implies that A and A^* have the same eigenvectors. Therefore A is normal. □

The following important theorem states that the condition number measures how much 'larger' than the spectrum an ε-pseudospectrum is possible to be, as a function of ε of course.

Theorem (Bauer-Fike)**.** *Suppose $A \in \mathbb{C}^{N \times N}$ is diagonalizable, $A = V \Lambda V^{-1}$, where the columns of $\boldsymbol{V}$ are the eigenvectors of $\boldsymbol{A}$. Then for each $\varepsilon > 0$, with*

$\| \cdot \| = \| \cdot \|_2$, $\sigma(A) + B_\varepsilon \subseteq \sigma_\varepsilon(A) \subseteq \sigma(A) + B_{\varepsilon\kappa(V)}$.

Proof. The first inclusion was established in the previous theorem. For the second, $(z - \mathbf{A})^{-1} = \left(z - \mathbf{V}\Lambda\mathbf{V}^{-1}\right)^{-1} = [\mathbf{V}\,(z-\Lambda)\,V^{-1}]^{-1} = \mathbf{V}\,(z-\Lambda)^{-1}\,V^{-1}$ which implies $\|\,(z-\mathbf{A})^{-1}\,\|_2 \leq \kappa(\mathbf{V})\|\,(z-\Lambda)^{-1}\,\|_2 = \frac{\kappa(\mathbf{V})}{\mathrm{dist}(z,\sigma(\mathbf{A}))}$. The first definition of pseudospectra, $\|\,(z-\mathbf{A})^{-1}\,\| > \varepsilon^{-1}$, leads to dist $\big(z, \sigma(\mathbf{A})\big) < \varepsilon\kappa(\mathbf{V})$, which completes the proof. □

For purposes of completeness, we mention here some basic properties of pseudospectra [Trefethen and Embree, 2005].

Theorem (Properties of pseudospectra)**.** *Let $A \in \mathbb{C}^{N\times N}$ and $\varepsilon > 0$ be arbitrary.*

1. *$\sigma_\varepsilon(A)$ is nonempty, open and bounded with at most N connected components, each containing one or more eigenvalues of A.*
2. *If $\| \cdot \| = \| \cdot \|_2$, then $\sigma_\varepsilon(A^*) = \overline{\sigma_\varepsilon(A)}$.*
3. *If $\| \cdot \| = \| \cdot \|_2$ and $A_1 \oplus A_2 = \begin{pmatrix} A_1 & 0 \\ 0 & A_2 \end{pmatrix}$, then $\sigma_\varepsilon(A_1 \oplus A_2) = \sigma_\varepsilon(A_1) \cup \sigma_\varepsilon(A_2)$.*
4. *For any $c \in \mathbb{C}$, $\sigma_\varepsilon(A + c) = c + \sigma_\varepsilon(A)$.*
5. *For any nonzero $c \in \mathbb{C}$, $\sigma_{|c|\varepsilon}(cA) = c\sigma_\varepsilon(A)$.*

1.3 Pseudospectrum of a Toeplitz matrix

Our goal now is to present an example where we check the sensitivity of the eigenvalues under perturbations of the original matrix. We start with a simple definition.

A *Toeplitz matrix* is a matrix of the form

$$A = \begin{pmatrix} a_0 & a_{-1} & & \dots & a_{1-N} \\ a_1 & a_0 & a_{-1} & & \vdots \\ & \ddots & \ddots & \ddots & \\ \vdots & & & a_0 & a_{-1} \\ a_{N-1} & \dots & & a_1 & a_0 \end{pmatrix}, a_{1-N}, \dots, a_0, \dots, a_{N-1} \in \mathbb{C}.$$

The function $f(z) = \sum_k a_k z^k$ is called the *symbol* of the Toeplitz matrix A. In [Ekström and Serra-Capizzano, 2018] the eigenvalues of a Toeplitz matrix of the form

$$A = \begin{pmatrix} a_0 & 0 & \dots & 0 & a_{-\omega} & & \\ 0 & a_0 & \ddots & \ddots & \ddots & \ddots & \\ \vdots & \ddots & \ddots & \ddots & \ddots & \ddots & a_{-\omega} \\ 0 & \ddots & \ddots & \ddots & \ddots & \ddots & 0 \\ a_\omega & \ddots & \ddots & \ddots & \ddots & \ddots & \vdots \\ & \ddots & \ddots & \ddots & \ddots & a_0 & 0 \\ & & a_\omega & 0 & \dots & 0 & a_0 \end{pmatrix},$$

are $\lambda_k(A) = a_0 + 2\sqrt{a_\omega a_{-\omega}} cos\omega\frac{k\pi}{N+1}$, $1 \leq k \leq N$.

$$\text{Let } A = \begin{pmatrix} 0 & 1 & & & \\ \frac{1}{4} & 0 & 1 & & \\ & \ddots & \ddots & \ddots & \\ & & \frac{1}{4} & 0 & 1 \\ & & & \frac{1}{4} & 0 \end{pmatrix}$$

The eigenvalues of A are $\lambda_k(A) = cos\frac{k\pi}{N+1}$, $1 \leq k \leq N$. To test the sensitivity of the eigenvalues of A we consider a perturbation of A, $A + E$, where E is a random matrix with $\|E\| = \varepsilon$. The image of the circle $|z| = \varepsilon^{\frac{1}{N}}$ under the *symbol* of A, $f(z) = z^{-1} + \frac{1}{4}z$, is an ellipse, the interior of which is approximated by the ε-pseudospectrum of A, when ε increases. Without going into details, we mention that this is due to the fact that if z lies in the interior of the image of the unit circle, then $\|(z - A)^{-1}\|$ grows exponentially as N grows, whereas it is uniformly bounded for z outside this curve.

Although the eigenvalues of A are on the real axis, the eigenvalues of $A + E$ move close to the ellipse of the ε-pseudospectra of A as seen in Figure 1.1. The sensitivity of the eigenvalues becomes clearer as N increases.

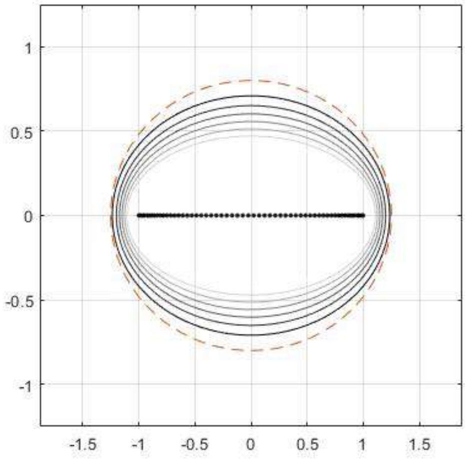


Figure 1.1: Boundaries of pseudospectra $\sigma_\varepsilon(A), \varepsilon = 10^{-2}, 10^{-3}, \ldots, 10^{-8}$, for the matrix A of dimension N = 64. The dashed ellipse is the image of the unit circle under the symbol $f(z) = z^{-1} + \frac{1}{4}z$.

For a particular example, consider E to be the matrix with ε at the bottom left cell, and 0 everywhere else (probably $\|E\| = \varepsilon$). If we symmetrize the matrix A by the diagonal similarity transformation $DAD^{-1} = S$ with $D = diag(2, 2^2, \ldots, 2^N)$, matrix S has the same eigenvalues as A.

$$S = \begin{pmatrix} 0 & \frac{1}{2} & & & \\ \frac{1}{2} & 0 & \frac{1}{2} & & \\ & \ddots & \ddots & \ddots & \\ & & \frac{1}{2} & 0 & \frac{1}{2} \\ & & & \frac{1}{2} & 0 \end{pmatrix}$$

After applying the same similarity transformation to $A + E$,
we obtain $D\,(A + E)\,D^{-1}$

$$D\,(A+E)\,D^{-1} = \begin{pmatrix} 0 & \frac{1}{2} & & & \\ \frac{1}{2} & 0 & \frac{1}{2} & & \\ & \ddots & \ddots & \ddots & \\ & & \frac{1}{2} & 0 & \frac{1}{2} \\ 2^{N-1}\varepsilon & & & \frac{1}{2} & 0 \end{pmatrix}$$

As A and S have the same eigenvalues, it is now clear that we have two matrices, on the one hand $A+E$ and on the other hand $D\,(A+E)\,D^{-1} = S + DED^{-1}$, whose spectra match in the same time that their difference grows exponentially with N.

Chapter 2

Numerical solutions of differential equations

2.1 Differentiation matrices and their pseudospectra

In scientific computing, the derivative of a function is often approximated (in some grid, after discretization) by a *differentiation matrix* which multiplies a vector of data. When the grid is not periodic, the differentiation matrix is usually nonnormal and its nonnormality grows exponentially as the number of grid points increases. Nonnormality of a matrix A is measured by $v(A) = \left(\|A\|_F^2 - \sum_j |\lambda_j|^2\right)^{1/2}$, where $\|\cdot\|_F$ is the Frobenius norm (see [Higham, 2020]).

In this section, we work on Chebyshev and Legendre spectral differentiation matrices where the eigenvalues are sensitive to perturbations. Such behaviour of nonnormal matrices affects numerical stability.

To demonstrate such an application, we consider the example of *Chebyshev spectral differentiation* on the interval $[-1,1]$ with $N+1$ *Chebyshev points* $x_j = cos\left(j\pi/N\right),\ j = 0,1,\dots,N$. The *spectral differentiation method* uses a polynomial p to interpolate a given function u on the grid x_j and differentiate the polynomial to define a discrete derivative $\mathbf{w} = (w_j)_{0\leq j\leq N}$. We use the notation $\mathbf{x} = (x_j)_{0\leq j\leq N}$, $u_j = u(x_j)$ and $\mathbf{u} = (u_j)_{0\leq j\leq N}$

- Let p be the unique polynomial of degree at most N with $p(x_j) = u_j$, $0 \leq j \leq N$.

- Set $w_j = p'(x_j)$ and w_j approximates $u'(x_j)$.

Since the differential operator is linear, instead of constructing the polynomial explicitly, we use an $(N+1)\times(N+1)$ matrix $\mathbf{D}_N$, $\mathbf{w} = \mathbf{D}_N\mathbf{u}$. Spectral methods manipulate these matrices explicitly to solve problems of ordinary or partial differential equations with no boundary conditions.

On small grids one can even compute manually p and use it to demonstrate $\mathbf{D}_N$, but of course this makes no sense in practice. For a trivial example, with $N = 2$, $x_0 = 1$, $x_1 = 0$, $x_2 = -1$, and using divided differences we easily find $p(x) = u_0 + (u_0 - u_1)(x-1) + \frac{u_2-2u_1+u_0}{2}x(x-1)$, therefore $p'(x) = u_0 - u_1 + (u_2 - 2u_1 + u_0)x - \frac{u_2-2u_1+u_0}{2} = \frac{u_0-u_2}{2} + (u_2 - 2u_1 + u_0)x$, and after calculating w_0, w_1 and w_2, we arrive at $\mathbf{D}_2 = \begin{pmatrix} \frac{3}{2} & -2 & \frac{1}{2} \\ \frac{1}{2} & 0 & -\frac{1}{2} \\ -\frac{1}{2} & 2 & -\frac{3}{2} \end{pmatrix}$.

In practise, matrix $\mathbf{D}_N$ is determined analytically by the following formula [Trefethen and Embree, 2005]: The off-diagonal entries of $\mathbf{D}_N$ are $\mathbf{D}_{ij} = \frac{c_i}{c_j}\frac{(-1)^{i+j}}{(x_i-x_j)}$, $i \neq j$, $i,j = 0,\dots,N$, where $c_0 = c_N = 2$ and $c_i = 1$ otherwise. The diagonal entries are defined by the condition that each row sums to zero.

For instance, $\mathbf{D}_3 = \begin{pmatrix} \frac{19}{6} & -4 & \frac{4}{3} & -\frac{1}{2} \\ 1 & -\frac{1}{3} & -1 & \frac{1}{3} \\ -\frac{1}{3} & 1 & \frac{1}{3} & -1 \\ \frac{1}{2} & -\frac{4}{3} & 4 & -\frac{19}{6} \end{pmatrix}$.

The following theorem states the nonnormality of these matrices.

Theorem. *For any N,* $\|\boldsymbol{D}_N\| > N^2/3$ *but* $(\boldsymbol{D}_N)^{N+1} = 0$.

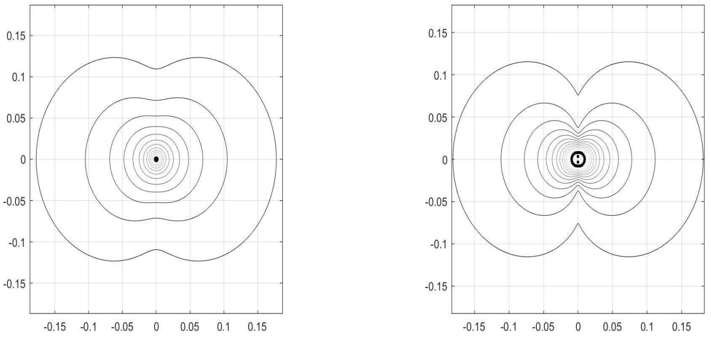
Proof. The upper left corner of $\mathbf{D}_N$ is $(2N^2+1)/6$ which is greater than $N^2/3$. The same must be true of $\|\mathbf{D}_N\|$.

To see this, consider the vectror $u \in \mathbb{R}^{N+1}$ with 1 as its first entry and zeros everywhere else. As $\|u\| = 1$, it follows that $\|\mathbf{D}_N u\| \leq sup_{v \in \mathbb{R}^{N+1}, \|v\|=1} \|\mathbf{D}_N v\| = \|\mathbf{D}_N\|$. But $\mathbf{D}_N u$ contains a number greater than $N^2/3$ as its first entry and zeros everywhere else, so $\|\mathbf{D}_N u\| > N^2/3$.

We now prove the nilpotency. The definition of $\mathbf{D}_N$ implies that for any vector $\mathbf{u}$, $\mathbf{D}_N \mathbf{u}$ is the vector containing the values of the derivative of p evaluated at the grid $\mathbf{x}$, where p is the polynomial of degree at most N that interpolates the function u at the grid. Left-multiplying by $\mathbf{D}_N$ N more times gives the values of the $(n+1) - th$ derivative of p at the grid. Therefore, the result will always be zero regardless of $\mathbf{u}$. Thus $(\mathbf{D}_N)^{N+1}$ is the zero matrix. □

The cleanest results are obtained when we consider the normalized matrices $\mathbf{A}_N = N^{-2}\mathbf{D}_N$. Increasing N, the 'nonnormality' of the matrices increases. In figures 2.1a and 2.1b, one may see the growth of the lack of normality as N grows: $\mathbf{A}_N$ is nilpotent, so $\sigma(\mathbf{A}_N) = \{0\}$, meaning that its $\varepsilon - pseudospectrum$ should be something like a disc of radius ε if $\mathbf{A}_N$ was close to normal, but this seems to be far from truth as N increases.

For N large enough, the ε-pseudospectra of $\mathbf{A}_N$ appear converging to a figure at which $\| (x\mathbf{I} - \mathbf{A}_N)^{-1} \|$ grows approximately in proportion to $1.8^{\frac{1}{x}}$, as $x \to 0$.



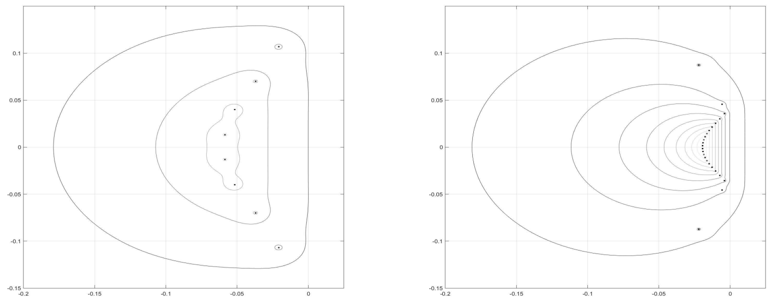
(a) $N = 8$

(b) $N = 24$

Figure 2.1: ε-pseudospectra of Chebyshev differentiation matrix $\mathbf{A}_N$, with $\varepsilon = 10^{-2}, 10^{-4}, \ldots$. Of course, as in all pseudospectra images, the external curve corresponds to the boundary of the ε-pseudospectrum with the largest ε, i.e. 10^{-2}. Obviously, as ε shrinks to zero, the pseudospectrum shrinks to the spectrum, something expected.

We now adjust $\mathbf{D}_N$ for applications with boundary conditions, a situation that appears more often in practice. Assume that the condition $u_0 = 0$ is imposed at the point $x = 1$. This condition is imposed by deleting the first row and the first column of the matrix $\mathbf{D}_N$ creating an $N \times N$ matrix $\tilde{\mathbf{D}}_N$ and respectively $\tilde{\mathbf{A}}_N = N^{-2}\tilde{\mathbf{D}}_N$. The matrix no longer has any eigenvalue equal to zero. The eigenvalues and the pseudospectra for different dimensions of the matrix $\tilde{\mathbf{A}}_N$ are shown in Figures 2.2a and 2.2b. The norms of $\mathbf{A}_N$ and $\tilde{\mathbf{A}}_N$, as $N \to \infty$, converge to 0.5498 and 0.0886, respectively.

The objective of spectral differentiation is to provide a 'spectrally accurate' approximation of exact differentiation by fast decreasing the errors of approximation. In the example that we have mentioned, we saw a 'spectrally accurate pseudospectrum': The right edge of each $\varepsilon - \textit{pseudospectrum}$ is a vertical line segment in the complex plane, something that also holds true for the differen-



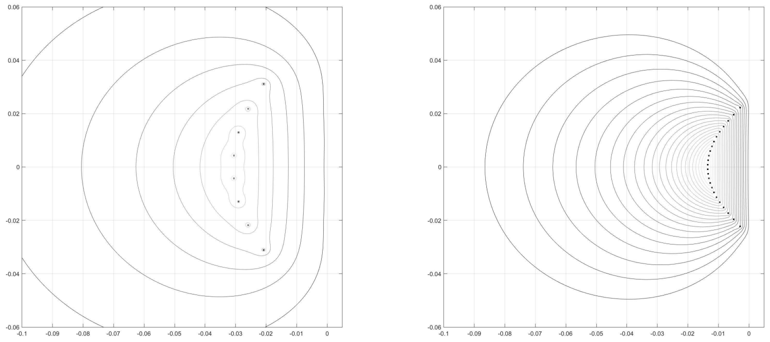
(a) $N = 8$ (b) $N = 24$

Figure 2.2: ε-pseudospectra of Chebyshev differentiation matrix $\tilde{\mathbf{A}}_N$, ε spanning over the same range as in figure 2.1

tial operator $\frac{d}{dx}$ on $[-1, 1]$ with the aforementioned boundary condition. On the contrary, as one may see in both figures, the eigenvalues do not present such a behaviour.

Repeating the example with *Legendre* instead of *Chebyshev* points, consisting of $x = 1$ and the zeros of the N-degree Legendre polynomial $P_N(x)$, results in shrinking the eigenvalues of the matrices $\tilde{\mathbf{D}}_N$ (see figure 2.3, where the large magnitude eigenvalues disappear). One could assume that such a discretization would allow an increase in stable time step sizes for time-dependent PDEs, a great advantage for applications. But this was proven not to be the truth: The pseudospectra of the matrices $\tilde{\mathbf{A}}_N$ near the origin, are almost the same as those for the Chebyshev points (compare figures 2.3 and 2.2) and the same vertical lines appear.

It was proven that it is the pseudospectra and not the spectra that determine the spectral accuracy. This explains why the Legendre grids do not permit increased time steps.



(a) $N = 8$ (b) $N = 24$

Figure 2.3: ε-pseudospectra of Legendre differentiation matrix $\tilde{\mathbf{A}}_N$, for various ε.

2.2 Discretization of the advection equation and Lax-stability

In time-dependent partial differential equations some finite-difference discretizations are stable and others are unstable, giving useless results. In this section we introduce the phenomenon of Lax-stability of such discretizations. We consider the advection equation $u_t = u_x$, $x \in (-1, 1)$, $t \geq 0$ with initial data

$$u\left(x,0\right) = \begin{cases} cos^2\left(\pi\left(x - \frac{1}{4}\right)\right), & \left|x - \frac{1}{4}\right| \leq \frac{1}{2} \\ 0, & otherwise \end{cases}$$

boundary data $u(1,t) = 0$ for all t and no boundary data for $x = -1$.

Assuming the problem is approximated numerically on a regular $\Delta x - \Delta t$ grid, we use the notation u_j^n for the discrete approximation at $x = -1 + j\Delta x$, $t = n\Delta t$, and also write $\mathbf{u}^n = (u_j^n)_{0 \leq j \leq \frac{2}{\Delta x}}$. Now, using centred differences in x, $\frac{\partial u}{\partial x}(j\Delta x, n\Delta t) \approx (\mathbf{D}\mathbf{u}^n)_j$ where $\mathbf{D}$ is the tridiagonal matrix such that $(\mathbf{D}\mathbf{u}^n)_j =$

$\frac{u_{j+1}^n - u_{j-1}^n}{2\Delta x}$ and the third-order Adams-Bashforth formula in t, $u_j^{n+1} = u_j^n + \Delta t \mathbf{D}\left(\frac{23}{12}u_j^n - \frac{16}{12}u_j^{n-1} + \frac{5}{12}u_j^{n-2}\right)$, we test the stability of the method.

We consider $N = 60$, $\Delta x = 2/N$, $u_N^n = 0$, $u_0^n = u_1^n$ and initial values u_j^0, u_j^1, u_j^2, $(1 \leq j \leq N-1)$ taken from the exact solution $u(x,t) = u(x+t,0)$. It can be proven that the numerical solution is stable if and only if $\frac{\Delta t}{\Delta x}$ is less than about 0.724.

Now consider the same example on a grid of N *Legendre points* $(x_j)_{0 \leq j \leq N-1}$. (Remember that these are the roots of the *N-th degree Legendre polynomial*). We interpolate $u_j^n, 0 \leq j \leq N-1$ at these points and the boundary value $u_N^n = 0$ for all t by a polynomial p_N of degree N, i.e. $p_N(x_j) = u_j^n$ for $0 \leq j \leq N-1$ and $p_N(1) = 0$ (the polynomial is well defined since all together the points are n+1). The approximate spatial derivative is the derivative of the polynomial.

In this example, for a finite t, the numerical solution shows a good behaviour near the boundary $x = 1$, whereas, near the other boundary $x = -1$, there is a terrible instability and the smooth image of a wave moving left as time passes (the exact solution of the PDE with the aforementioned initial conditions) is destroyed. Why does this happen?

The Adams-Bashforth formula can be written in the following form, using $3N \times 3N$ block matrices:

$$S = \begin{pmatrix} \mathbf{I} & \mathbf{0} & \mathbf{0} \\ \mathbf{I} & \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \mathbf{I} & \mathbf{0} \end{pmatrix} + \frac{\Delta t}{12} \begin{pmatrix} 23\mathbf{D} & -16\mathbf{D} & 5\mathbf{D} \\ \mathbf{0} & \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & \mathbf{0} \end{pmatrix},$$

where $\mathbf{D}$ is the Legendre differentiation matrix. Matrix S maps $(\mathbf{u}^n, \mathbf{u}^{n-1}, \mathbf{u}^{n-2})^T$ to $(\mathbf{u}^{n+1}, \mathbf{u}^n, \mathbf{u}^{n-1})^T$. This formula corresponds to $\mathbf{u}^{n+1} = (\mathbf{I} + \frac{23\Delta t}{12}\mathbf{D})\mathbf{u}^n - \frac{4\Delta t}{3}\mathbf{D}\mathbf{u}^{n-1} + \frac{5\Delta t}{12}\mathbf{D}\mathbf{u}^{n-2}$.

Stability is now investigated by the norms of powers of S. This is what we call *Lax-stability*: An iterative numerical method for solving a PDE is defined as *Lax-stable* if the norm of the matrix involved is uniformly bounded as $\Delta x, \Delta t$ converge to 0.

An important theorem that finds lower bounds for the norms of the powers of the iteration matrix is the following (see [Trefethen and Embree, 2005]):

Theorem. *Assume that A is a matrix or bounded operator and there exists a constant $K > 1$ such that $\|(z - A)^{-1}\| = K/(|z| - 1)$ for some z of radius $|z| = r > 1$. Then, $sup_{k \geq 0} \|A^k\| > K$.*

We now return to our example with the Legendre grid. The method is proven to be Lax-stable only if $\Delta t = \mathcal{O}\left(N^{-2}\right)$ as $N \rightarrow \infty$ and will converge to the exact solution if there are no rounding errors. That it will indeed converge is a consequence of the famous *Lax-equivalence theorem*, which states that if we have a consistent finite difference method for a well-posed linear initial value problem, then the method converges to the exact solution if and only if it is Lax-stable.

On the contrary, if $\Delta t = \mathcal{O}\left(N^{-1}\right)$ as $N \rightarrow \infty$, although the eigenvalues remain in the unit circle, the scheme is no longer Lax-stable. To see this, we focus on Figure 2.4. According to the previous theorem, for $z = -1.1$, since $\|(z - S)^{-1}\| \approx 10^6 = \frac{10^5}{|z|-1}$, we have $sup_{k \geq 0} \|S^k\| > 10^5$. Therefore we do not have Lax-stability, so we do not expect convergence.

Discretization for large time simulations is not possible. In the example, it seemed that it was possible because we considered a constant-coefficient linear problem with no rounding errors. If perturbations of any kind occur, the instability is no longer transient (i.e. present only for short time simulations and disappearing as t grows), but global.

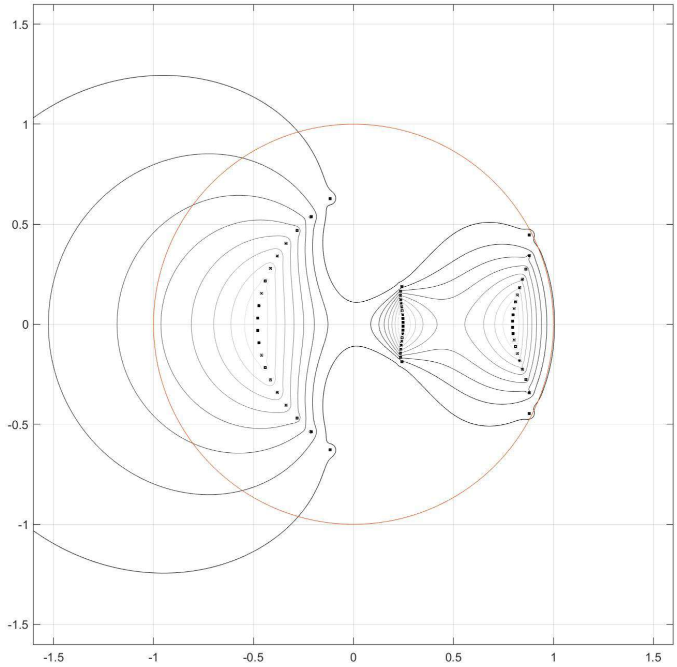


Figure 2.4: Pseudospectra of the matrix S with $N = 20$, $\Delta t = 0.4N^{-1}$ with $\varepsilon = 10^{-2}, 10^{-4}, \ldots$. Observe that the boundary of a pseudospectrum with $\varepsilon = 10^{-6}$ intersects the negative real semiaxis at a point z of radius $r \approx 1.2$.

2.3 Stability of various discretizations of the advection equation with the method of lines

For the discretization of a time-dependent PDE, we use the *method of lines*. We can discretize the PDE first with respect to its spatial variables and then discretize the resulting system of coupled ODEs in time.

We consider the advection equation $u_t = u_x$ on $[-\pi, \pi]$ with periodic boundary conditions and initial data $u(x, 0) = f(x)$. Applying the method of lines, we discretize the spatial variable by an 'upwind' approximation and the time variable by the forward Euler formula. So, we have, on the grid $x = -\pi + j\Delta x$, $t = n\Delta t$, $\frac{\partial u}{\partial x} \approx \frac{u_{j+1} - u_j}{\Delta x}$, and the problem is deduced to the following system of ODEs: $\frac{du_j(t)}{dt} = \frac{u_{j+1}(t) - u_j(t)}{\Delta x}$, $n \in \mathbb{N}$. Finally we arrive at $u_j^{n+1} = u_j^n + \frac{u_{j+1}^n - u_j^n}{\Delta x}\Delta t$. This can be of course wrirren down as $\mathbf{u}^{n+1} = \mathbf{A}\mathbf{u}^n$, where

$$\mathbf{A} = \begin{pmatrix} 1-\sigma & \sigma & & & & \\ & 1-\sigma & \sigma & & & \\ & & \ddots & \ddots & & \\ & & & 1-\sigma & \sigma \\ & & & & 1-\sigma \end{pmatrix}, \sigma = \frac{\Delta t}{\Delta x}.$$

For a normal or nearly normal spatial discretization operator, the discretization is stable if all eigenvalues of this operator lie in the stability region of the time discretization operator.

In the example, the stability region of the forward Euler formula is the disc with centre at -1 and radius 1 (an elementary fact from Numerical Anaysis). The eigenvalues of the spatial discretization operator lie in the circle of center $-\frac{\Delta t}{\Delta x}$ with radius $\frac{\Delta t}{\Delta x}$. To see this, consider the Fourier modes $u_j^n = \lambda^n \epsilon^{ikj\Delta x}$, where k are arbitrary wave numbers and i the imaginary unit. Since the operator that performs the spatial discretization is $\frac{u_{j+1}^n - u_j^n}{\Delta x}$, this implies $\frac{e^{ik(j+1)\Delta x} - e^{ikj\Delta x}}{\Delta x} = \lambda e^{ikj\Delta x}$ and after dividing by $e^{ikj\Delta x}$ we arrive at $\lambda = \frac{e^{ik\Delta x} - 1}{\Delta x}$. Now, multiplying with Δt to adjust the time scale, we have the scaled eigenvalues $\lambda = \frac{\Delta t}{\Delta x}(e^{ik\Delta x} - 1)$. Note that the operator is normal, so pseudospectra are not necessary here for the stability to be checked. Therefore, the discretization is expected to be stable if $\Delta t \leq \Delta x$.

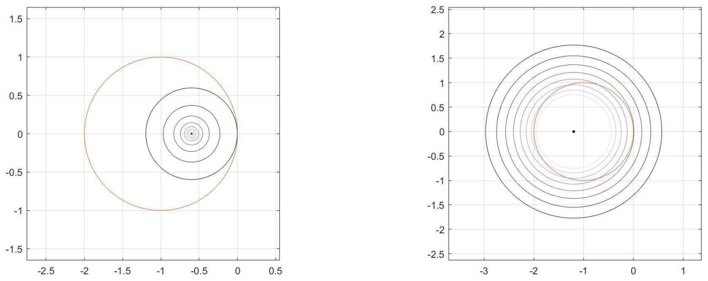
Considering another example of the advection equation without periodic boundary conditions and different discretizations for x and t, we discretize the spatial variable by a 'centred' approximation which gives imaginary eigenvalues and the time variable by the third-order Adams-Bashforth formula which has a different stability region than the forward Euler formula. Here, the eigenvalues of the spatial discretization operator are in the stability region if $\frac{\Delta t}{\Delta x} \leq 0.724$, making the computation stable. The discretization matrix here is not normal but it is close to normal, so spectral analysis is sufficient for stability predictions.

On the other hand, if we want to study, for instance, the stability of the Legendre spectral discretization, the discretization operator is far from normal, hence the need to examine the $\varepsilon - pseudospectra$ arrises. Here, although the eigenvalues of the example are in the stability region, the pseudospectra protrude outside, therefore, we do not expect stability of the method.

The corresponding stability condition now is:

The discretization is stable if the ε-pseudospectra of the spatial discretization operator lie within a distance of $\mathcal{O}(\varepsilon)$ from the stability region of the time discretization operator, when $\varepsilon \to 0$.

Consider another example of the advection equation $u_t = u_x$ on $[-1, 1]$, with boundary conditions $u\,(1, t) = 0$ and initial data. For the matrix $\mathbf{A}$ (see previous page), its eigenvalues for $\sigma < 2$ satisfy $|\lambda| < 1$. Nevertheless, the norms of the powers of $\mathbf{A}$ for $1 < \sigma < 2$ will grow exponentially before decaying, so we do not expect stability. The $\varepsilon - pseudospectra$ confirms the results as they extend outside the stability region for $1 < \sigma < 2$. The discretization is stable only for $\sigma \leq 1$ as we see in Figures 2.5a and 2.5b.



(a) $\sigma = 0.6$

(b) $\sigma = 1.2$

Figure 2.5: Pseudospectra of matrix $\mathbf{A}$ (blue), along with the stability region of the time discretization operator (pink). Observe that, in (a), the distance between the boundary of ε-pseudospectra and the boundary of the stability region converges fast to zero. On the contrary, in (b), this does not happen, indicating instalility.

2.4 The concept of stiffness in ODEs

A concept that needs to be explained when one is dealing with numerical solutions of ordinary differential equations is that of *stiffness*. There is no precise definition of stiffness. See, e.g. [Söderlind et al., 2015], for a historical recursion on the attempts for defining a problem as stiff. Some commonly-recognisable characteristics of a stiff problem are:

1. The problem contains widely varying time scales.
2. Stability is more of a constraint on the time step than accuracy.
3. Explicit methods do not work.

A problem is stiff when it includes some terms that make the solution manifest a transient behaviour. To explain the relationship between these statements and the concepts of spectra and pseudospectra we work on an example. We consider the initial value problem $u'(t) = -100\,\big(u(t) - \cos(\mathrm{t})\big) - \sin(\mathrm{t})$, $u(0) = 1$ with exact solution $u(t) = \cos(\mathrm{t})$. The constant -100 has a major effect on the solution for other initial data. Comparing the solutions computed for a given time with the second-order Adams-Bashforth (AB2) and the backward differentiation (BD2) formulas for different time steps Δt, we observed that BD2 behaved much better than AB2. BD2 converged smoothly for any time step, AB2 generated such behaviour only for small enough Δt. Although AB2 converged to the correct solution, it required much more computation steps than BD2. That is one of the reasons stiffness is of importance to the numerical solution of ODEs. For other equations, the AB formulas would be preferable since they are explicit and do not require an iterative solution at each time step.

We can test whether a formula has a time step constraint and what that constraint is if we apply the formula to the linear test equation $w' = \lambda w$. In our example, if $u(t)$ is any solution to the problem $u'(t) = -100\,\big(u(t) - \cos(\mathrm{t})\big) - \sin(\mathrm{t})$, we set $w(t) = u(t) - \cos(\mathrm{t})$ and the problem becomes $w' = \lambda w$, $\lambda = -100$.

When we apply the AB2 formula to the linear test equation, we get the characteristic polynomial of the recurrence formula. The roots of the characteristic polynomial $p(z) = z^2 - (\frac{3}{2}\lambda\Delta t + 1)z + \frac{1}{2}\lambda\Delta t$ are obviously real and the smaller of them is $z_1 = \frac{3}{4}\lambda\Delta t + \frac{1}{2} - \frac{1}{4}\sqrt{9(\lambda\Delta t)^2 + 4\lambda\Delta t + 4}$. So, $z_1 < -1$ if and only if $\lambda\Delta t < -1$, which implies that for stability to be achieved one should choose $\Delta t \leq \frac{1}{\lambda} = \frac{1}{100}$, otherwise the formula will amplify any truncation errors.

When we test the BD2 formula the same way, we find that all the roots of its characteristic polynomial are stable, regardless of the choice of Δt.

The method of lines discretizes time-dependent partial differential equations reducing them to a system of ordinary differential equations. In the general case of a system of ordinary differential equations, $\mathbf{u}' = \mathbf{f}(\mathbf{u}, t)$ where $\mathbf{u}(t)$ is an N-vector for each t and $\mathbf{f}$ is in general nonlinear, we can do an eigenvalue analysis through a process of four steps.

To determine whether the problem of computing a particular solution $\mathbf{u}^*(t)$ near a particular time t^* is stiff we follow the four steps below:

1. The first step is to discretize the time-dependent PDE turning it into a system of ODEs.

2. The second step is to *linearize* the system of equations.
   We set $\mathbf{u}(t) = \mathbf{u}^*(t) + \mathbf{w}(t)$ assuming $\mathbf{w}(t)$ is small, making stability and stiffness depend on $\mathbf{w}(t)$. If $\mathbf{f}$ is differentiable with respect to the components of $\mathbf{u}$, $\mathbf{A}(t)$ is the *Jacobian matrix* of $\mathbf{f}$ with respect to $\mathbf{u}$, we have $\mathbf{f}(\mathbf{u}, t) = \mathbf{f}(\mathbf{u}^*, t) + \mathbf{A}(t)\mathbf{w}(t) + o\left(\|\mathbf{w}\|\right)$. Since $\mathbf{u}' = \mathbf{f}(\mathbf{u}, t)$, we obtain $\mathbf{w}'(t) = \mathbf{u}'(t) - \mathbf{u}'^*(t) = f(\mathbf{u}, t) - f(\mathbf{u}^*, t) = \mathbf{A}(t)\mathbf{w}(t) + o\left(\|\mathbf{w}\|\right) \approx \mathbf{A}(t)\mathbf{w}(t)$ for small $\mathbf{w}(t)$. Changing variables leads to $\mathbf{u}' = \mathbf{A}(t)\mathbf{u}(t)$.

3. The third step is to *freeze coefficients*.
   Stability and stiffness appear at some times t^*. Setting $\mathbf{A} = \mathbf{A}(t^*)$ at a fixed time t^*, we obtain $\mathbf{u}' = \mathbf{A}\mathbf{u}$.

4. Finally, if $\mathbf{A}$ is diagonalizable, we *diagonalize* it and we get a set of N scalar, linear, constant-coefficient model problems $u' = \lambda u$.

We can view the rough conditions of stability and stiffness through the eigenvalue analysis.

1. *Eigenvalue characterization of stability.* A numerical ODE formula is stable for computing $\mathbf{u}^*(t)$ near t^* if Δt is small enough that for each eigenvalue λ of $\mathbf{A}(t^*)$, $\lambda\Delta t$ lies inside or close $\left(\mathcal{O}\left(\Delta t\right)\right)$ to the stability region.

2. *Eigenvalue characterization of stiffness.* An ODE is stiff for the solution $\mathbf{u}^*(t)$ near t^* if the largest eigenvalue modulus $|\lambda|$ of $\mathbf{A}(t^*)$, is much greater

than $(\mathbf{u}^*)'(t)$. In our example, $cos'(t) \leq 1 << 100 = |\lambda| = \big|d(\lambda w)/dw)\big|$, so the problem is stiff.

On the contrary, for many ODEs, the rate of change of $\mathbf{u}(t)$ results from $\mathbf{A}(t)$. The *stiffness ratio* for a solution of an ODE may be defined as the ratio of the absolute value of the eigenvalues of $\mathbf{A}$, or their real parts. For example, if the eigenvalues of a Hermitian matrix range from -10^6 to -1, the stiffness ratio is 10^6 and the problem would be highly stiff.

Attempting to characterize stiffness by eigenvalues or stiffness ratios cannot always be correct because stiffness is a transient phenomenon and eigenvalues are not always linked to matrix behaviour. To address this problem, we will view stability and stiffness through pseudospectral analysis through an example. Assume the linear constant-coefficient matrix equation $\mathbf{u}' = \mathbf{A}\mathbf{u}$, $\mathbf{u}(0) = \mathbf{u}_0$, where $\mathbf{u}_0$ is an N-vector and $\mathbf{A}$ is an $N \times N$ triangular matrix.

Let $\mathbf{u}_0 = \begin{pmatrix} 1 & 1 & \dots & 1 \end{pmatrix}^T$, $\mathbf{A} = \begin{pmatrix} -1 & & & \\ -2 & -1 & & \\ \vdots & \ddots & \ddots & \\ -2 & \dots & -2 & -1 \end{pmatrix}$.

Matrix $\mathbf{A}$ has the single eigenvalue -1. In Figure 2.6 the pseudospectra are in the left-half plane and extend far beyond -1. As expected, the AB2 formula explodes for time steps greater than $\Delta t = \mathcal{O}\left(10^{-1}\right)$. We can now express the pseudospectral view of stability and stiffness.

1. *Pseudospectral characterization of stability.* A numerical ODE formula is stable for computing $\mathbf{u}^*(t)$ near t^* if Δt is small enough that the ε-pseudospectra of $\Delta t\mathbf{A}(t^*)$ lie within a distance $\mathcal{O}\left(\varepsilon + \Delta t\right)$ of the stability region.

2. *Pseudospectral characterization of stiffness.* An ODE is stiff for the so-

lution of $\mathbf{u}^*(t)$ near t^* if the pseudospectra of $\mathbf{A}(t^*)$ extend far into the left-half plane as compared with $(\mathbf{u}^*)'(t)$.

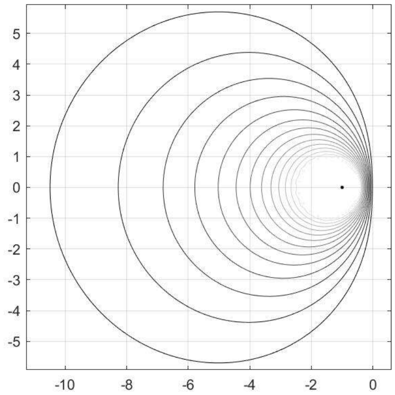


Figure 2.6: Eigenvalue and ε-pseudospectra of the 40 × 40 matrix A.

Throughout this section, we have been working on ODEs, and we have not mentioned the solution of PDEs, where stiffness plays an important role. Nevertheless, for many time-dependent PDEs, a discretization in space leads to a stiff system of ODEs where one can use the statements of stability and stiffness to address such problems.

2.5 GKS-stability of boundary conditions

So far, we have worked on numerical solutions of differential equations with potentially explosive behaviour, such as ones with terrible instability arrising when time step size is not small enough. A different kind of stability theory, one that examines instability when the boundary conditions for finite-difference discretizations of linear hyperbolic PDEs are badly chosen, is the *GKS-stability* theory, named after Gustafson, Kreiss and Sundstroem who developed it in early 1970s. As we will see, this theory can be worded in terms of the group velocities of waves propagating in dispersion on the finite-difference grid that we have chosen for the discretization of the PDE.

We consider the linear, scalar, hyperbolic, constant-coefficient, one-dimensional, initial boundary value problem $u_t = u_x$, $u\left(x, 0\right) = u_0(x)$ for $0 < x < 1$, $u\left(1, t\right) = 0$ for $t > 0$, where u_0 is the initial data. The analytic solution is a wave propagating leftward at speed 1 until it is absorbed in the boundary:

$$u\left(x,t\right) = \begin{cases} u_0(x+t) & \text{for } x+t<1 \\ 0 & \text{for } x+t \geq 1 \end{cases}$$

We discretize the problem by setting $\Delta x = 1/N$, where N is a positive integer, $\Delta t = \sigma \Delta x$ for $\sigma < 1$ and compute the approximations $v_j^n \approx u\left(j\Delta x, n\Delta t\right)$. We use the following formula, known as Crank-Nicolson formula:

$$\frac{v_j^{n+1} - v_j^n}{\Delta t} = \frac{\frac{1}{2}(v_{j+1}^n - v_{j-1}^n)}{2\Delta x} + \frac{\frac{1}{2}(v_{j+1}^{n+1} - v_{j-1}^{n+1})}{2\Delta x} = \frac{v_{j+1}^n - v_{j-1}^n}{4\Delta x} + \frac{v_{j+1}^{n+1} - v_{j-1}^{n+1}}{4\Delta x}.$$

It is not the ideal formula for this equation, nevertheless it is more convenient for the example.

To complete the numerical method, we need to define v_0^{n+1}. For the example, we consider the extrapolation from the interior. We examine two cases (a) $v_0^{n+1} = v_1^n$, (b) $v_0^{n+1} = v_2^n$.

The procedure of moving from step n to $n+1$ is linear, therefore a matrix $\mathbf{A}$ exists such that:

$$\left(v_0^{n+1}, v_1^{n+1}, \ldots, v_{N-1}^{n+1}\right)^T = \mathbf{A}\left(v_0^n, v_1^n, \ldots, v_{N-1}^n\right)^T.$$

A necessary and sufficient condition for convergence to the solution as $N \to \infty$ is the numerical method to be stable. That is $\|\mathbf{A}^n\| \leq C$ for all $n \leq N$ where C is a constant independent of N according to the *Lax Equivalence Theorem.*

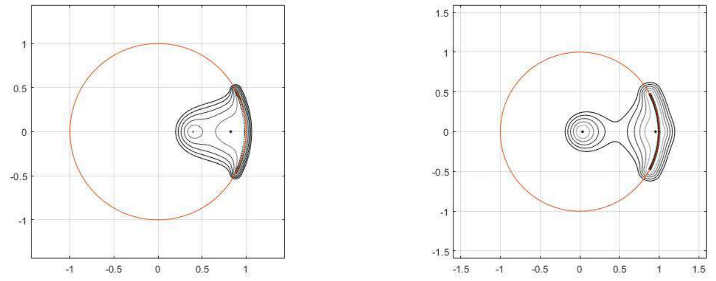
As one may see in Figures 2.7a and 2.7b, the eigenvalues of matrices $\mathbf{A}$ are in close positions for both choices of boundary conditions, so spectra analysis does not reveal any difference in stability of the two choises. Nevertheless, the ε-pseudospectra analysis of the matrices $\mathbf{A}$ for the two cases gives an insight into their stability. The plot of the ε-pseudospectra of the matrix $\mathbf{A}$ for case (b) in Figure 2.7b reveals a bulge near $z = 1$, which is not present in case (a) in Figure 2.7a, indicating instability if $v_0^{n+1} = v_2^n$. On the other hand, ε-pseudospectra in Figure 2.7a extend past $z = 1$, but the distance of their boundary from $z = 1$ is a linear function of ε, indicating stability.

The plot of the norms of $\mathbf{A}^n$ for the two boundary conditions confirms the instability in case (b) (Figure 2.8).

To visualise instability working on certain initial data, consider

$$u_0(x) = e^{-100\left(x-1/2\right)^2}$$

These data represent a Gaussian with value $u_0(1) \approx 1.39 \times 10^{-11}$ at the right boundary, so we can approximately assume that it also satisfies the boundary



(a) $v_0^{n+1} = v_1^n$ (b) $v_0^{n+1} = v_2^n$

Figure 2.7: Eigenvalue and ε-pseudospectra of the 60 × 60 matrix A.

condition $u_0(1) = 0$. If we plot the solutions for $0 \leq x \leq 1$ and $0 \leq t \leq 1$ for the two boundary conditions, the instability in case (b) is reaffirmed. In case (a), the plot is a wave propagating leftward with velocity -1 which dies out once it hits the left-hand boundary, i.e. the correct analytical solution. The instability of case (b) takes the form of a saw-toothed reflected wave of similar amplitude travelling rightward with group velocity $+1$. In the example, the unstable wave is $v_j^n = (-1)^j$. Refinement of the mesh could not solve the problem.

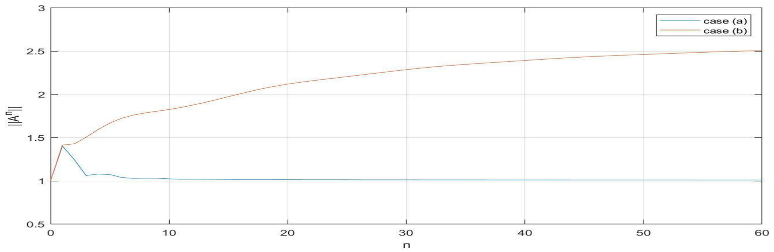


Figure 2.8: Norms $\|A^n\|$ for the two boundary conditions, confirming instability in case (b).

Reconsidering the example but with the leap frog approximation, writen out as

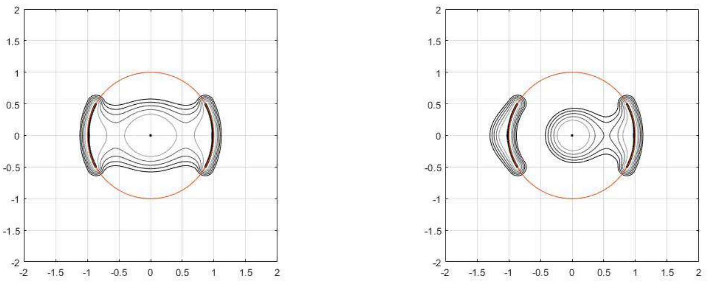
$$\frac{v_j^{n+1} - v_j^{n-1}}{2\Delta t} = \frac{v_{j+1}^n - v_{j-1}^n}{2\Delta x},$$

the procedure is equivalent to multiplication by a $2N \times 2N$ matrix:

$$\left(v_0^{n+1}, \ldots, v_{N-1}^{n+1}, v_0^n, \ldots, v_{N-1}^n\right)^T = \mathbf{A}\left(v_0^n, \ldots, v_{N-1}^n, v_0^{n-1}, \ldots, v_{N-1}^{n-1}\right)^T.$$

The boundary conditions are: (a) $v_0^{n+1} = v_1^n$, (b) $v_0^{n+1} = v_1^{n+1}$.

The ε-pseudospectra of the matrix $\mathbf{A}$ reveals a bulge in case (b) at $z = -1$ but not in case (a). The plot of the norms of $\|\mathbf{A}^n\|$ for the two boundary conditions confirms the instability in case (b). In this example, there is also a reflected wave $v_j^n = (-1)^n$, travelling rightward at group velocity $+1$, therefore a departure from the analytical solution, affirming instability.



(a) $v_0^{n+1} = v_1^n$

(b) $v_0^{n+1} = v_1^{n+1}$

Figure 2.9: Eigenvalue and ε-pseudospectra of the 60×60 matrix A.

The concept of this section is not as simple as it may appear. The example that we considered showed that instability is connected with a bulge of pseudospectra near the edge of the unit circle, but the problem under consideration had many simplifying assumptions: it was linear, scalar, constant-coefficient and one-dimensional. In fact, a bulge in the pseudospectra guarantees instability, however, the lack of a bulge guarantees stability only under certain circumstances but not in general.

In their classical work [Gustafsson et al., 1972], Gustafson, Kreiss and Sundstroem, gave a, rather complicated, definition of this kind of stability (now called GKS-stability) and proved that it is equivalent to the absence of *right-going waves*.

In particular, GKS-instable methods have the following features:

1. a bulge in the pseudospectra near a point z_0 on the unit circle.
2. growth of the norms $\|\mathbf{A}^n\|$ at a non-negligible rate.
3. existence of a wave that propagates from the boundary to the interior.

Chapter 3

Computation of pseudospectra

3.1 Basic pseudospectra computation

The pseudospectra of a given matrix or operator are non-empty sets in the complex plane. Since pseudospectra are norm-dependent, we have to choose a norm prior to constructing an algorithm for plotting them. We use the 2-norm (which is derived from the Eucleidian inner product).

With this choise in mind, remember the fourth definition of pseudospectra as the set of $z \in \mathbb{C}$ such that $s_{min}\left(z - A\right) < \varepsilon$, and a method for pseudospectra computation and hence plotting arrises immediately: The algorithm finds the set of singular values of the matrix $z - A$ for each point z on a grid in the complex plane (this is possible when the dimension N of A is small enough). Following that, it checks the smallest singular value and if it is smaller than ε

it lies within the ε-pseudospecrtrum of A, otherwise it lies outside. Finally, the algorithm sends the results to a contour plotter.

Producing pseudospectra in this plain way is sufficient only for small matrices that appear when we discretize problems in small grids, i.e. ones with large time or space steps. For large matrices, the speed of the method and appearance of the contour plots are unsatisfactory. An improvement of the method is necessary.

3.2 Computational speed improvement

The first thing that would improve the speed of the calculation of the method is to avoid regions of the complex plane where the resolvent norm is small and not interesting.

A second idea to speed up the computation of pseudospectra is triangularization followed by inverse Lanczos iteration. Our goal is to compute the smallest singular value of the resolvent $z - A$ and not the set of all singular values. To achieve that, we first apply a Schur decomposition on A, i.e. we factorize $A = UTU^*$, where T is upper triangular and U is unitary. Since, for the 2-norm, $\sigma_\varepsilon(A) = \sigma_\varepsilon(T)$, (see section 1.2), we just have to compute the smallest singular value of the upper triangular matrix $z - T$, which is the square root of the smallest eigenvalue of $(z-T)^*(z-T)$. This is equal to the smallest positive eigenvalue of the block matrix

$$\begin{pmatrix} 0 & z-T \\ (z-T)^* & 0 \end{pmatrix}$$

Adverse iteration is a method for finding this eigenvalue: We apply the so-called *power method* which generates values that converge to the largest eigenvalue of $((z-T)^*(z-T))^{-1}$ through iterations. The desired eigenvalue is then simply the inverse. Lanczos iteration linearly combines these iterations to speed up convergence, not asymptotically, but, in practical terms, significantly.

Nonetheless, for matrices with very large dimension ($N >> 1000$), complexity remains a serious problem (for example Schur decomposition is not that simple) and the aforementioned methods still fail, so other techniques are demanded.

A way out in this case is to reduce the dimension of the matrix by orthogonal projection on a subspace of smaller dimension. There are several choises of subspaces for the matrix to be projected on. This method does not find the exact ε-pseudospectrum, but for most applications, the results are satisfactory. The method gets rid of insignificant eigenvalues, such as ones with no physical meaning or ones that have appeared as a result of discretization.

For example, we may project an $N \times N$ matrix A onto the subspace $\mathcal{U} \in \mathbb{C}^N$ associated with the meaningful eigenvalues of the matrix (this of course requires that we have first determined the regions of the complex plane that are of interest for the particular problem). Using an $N \times p$ matrix $\mathbf{U}$ with colunms from $\mathcal{U}$, the projected matrix is $\mathbf{U}^*\mathbf{AU}$ and it is $p \times p$ - dimensional. It can be proven that $\sigma_\varepsilon(\mathbf{U}^*\mathbf{AU}) \subseteq \sigma_\varepsilon(\mathbf{A})$. We can choose $\mathcal{U}$ (i.e. the 'meaningful eigenvalues') in such a way that p is small enough for computations to be fast and in the same time the ε-pseudospectrum of $\mathbf{U}^*\mathbf{AU}$ to be a good enough approximation of $\sigma_\varepsilon(\mathbf{A})$.

3.3 Other ideas of increasing the speed of the computation

Alternative ideas to speeding the computation of the method include:

- The Lanczos iteration can be improved by selective re-orthogonalization and Chebyshev acceleration. See, e.g. [Braconnier and Higham, 1996] for details.

- The use of multiple processors could further speed up the process of computing pseudospectra. This is achievable, since the computations at each point of the grid are independent from the computation at all other points. Each processor needs not communicate or synchronize with the others until the computation is finished: it just performs computations regarding the points of the grid that it has been set as responsible for.

- The use of Krylov subspace iteration is a technique that could speed up the process of computing pseudospectra. It is an alternative for orthogonal projection. Given an $N \times N$ matrix A, a vector x and $p \leq N$, define the *Krylov subspace* $\mathcal{U}$ as the subspace of $\mathbb{C}^N$ generated by $x, Ax, A^2x, \ldots, A^{p-1}x$. Constructing an $N \times p$ matrix U with its columns forming an orthonormal basis of $\mathcal{U}$, project A as U^*AU and the eigenvalues of the latter converge to the eigenvalues of A as p increases to N. These approximations then give us approximations for the ε-pseudospectra of A.

- A completely different way of producing plots of pseudospectra instead of using a contour plotter would be to trace the boundary curves directly. This way the boundary curves can be determined to great accuracy and with fewer evaluations of the singular values of the resolvent since no grid is involved. How can we do this? Remember that, starting from the first definition of pseudospectra, the contour of an ε-pseudospectrum is the set of $z \in \mathbb{C}$, such that $\|(z - A)^{-1}\| = \varepsilon^{-1}$. Therefore, if we stick to the 2-norm, it is just a level curve of the function $\mathbb{C} \xrightarrow{f_A} \mathbb{R}$, $f_A(z) = \|(z - A)^{-1}\| = \frac{1}{s_{min}(z-A)}$. So, the main idea is the following: Starting from a point z_0 at the desired level curve, we go forward to a point z_1 performing a small step on the complex plane perpendicular to the gradient of f_A. Repeating the procedure again and again, one arrives at the contour plot.

- Sometimes we may be interested only for finding bounds for a pseudospectrum rather than the pseudospectrum itself. Explicit bounds exist for various categories of matrices (see [Gong et al., 2016]).

Chapter 4

Conclusions

To sum up the thesis, we saw that spectra analysis is not always sufficient for the stability of numerical methods to be guaranteed when dealing with differential equations. The matrices that appear as differentiation operators when trying to discretize such an equation are often 'far from normal', and this is exactly the situation where eigenvalues do not reveal enough information on the behaviour of the matrices.

After setting up the basic definitions, we focused mainly (but not exclusively) on the advection equation, because it is simple enough for technical difficulties to be avoided and in the same time it is sufficient for the usefulness of pseudospectra analysis to be observed. This way, we compared the stability of different discretizations, as well as different time steps and different boundary conditions. Important concepts that were also explained throughout the text include lax-stability, GKS-stability and stiffness.

Furthermore we briefly described algorithms for pseudospectra computation and, since the simpler of them are hopelessly slow, we mentioned some methods for speed improvement.

Despite the fact that pseudospectra gave us the opportunity to have very good approximations on how the norm of iteration matrices behaves (at least much better than spectra alone), at the present time questions on pseudospectra limitations remain open. For example, it is not known whether pseudospectra determine the behavior of the norms of nonderogatory matrices, i.e. matrices with the property that their eigenvalues are associated each with just a single Jordan block.

Appendix

Code File: Ps.m

The following code was used for the Figures to be created. It is based on Schur decomposition for plotting eigenvalues. It uses projection onto a subspace and inverse Lanczos iteration to find the minimum singular value of the matrices $z - A$ for z on a preselected grid on the complex plane and finally plots the level lines of the function $\frac{1}{s_{min}(z-A)}$.

```
%Basic code for 2-norm pseudospectra.
%2-norm pseudospectra are computed
%with the fourth definition of pseudospectra


% Set up grid for contour plot:
  npts = 1000;                                            % Grid Resolution
  s = .8*norm(A,1);          % Edges of the plot as function of the norm of A
  xmin = -s; xmax = s; ymin = -s; ymax = s;                          % Axes
  x = xmin:(xmax-xmin)/(npts-1):xmax;
  y = ymin:(ymax-ymin)/(npts-1):ymax;
  [xx,yy] = meshgrid(x,y); zz = xx + sqrt(-1)*yy;
```

```
% Compute Schur form and find eigenvalues as the diagonal entries of T:
  [U,T] = schur(A);
  if isreal(A), [U,T] = rsf2csf(U,T); end, T = triu(T); eigA = diag(T);
% plot eigenvalues on the complex plane
  hold off, plot(real(eigA),imag(eigA),'.','markersize',15), hold on
  axis([xmin xmax ymin ymax]), axis square, grid on, drawnow

% Reorder Schur decomposition and compress to interesting subspace.
% We are not interested in finding eigenvalues with
% large negative real part:
  select = find(real(eigA)>-250);                     % Subspace Selection
  n = length(select);
  for i = 1:n
    for k = select(i)-1:-1:i
      G([2 1],[2 1]) = planerot([T(k,k+1) T(k,k)-T(k+1,k+1)]')';
      J = k:k+1; T(:,J) = T(:,J)*G; T(J,:) = G'*T(J,:);
    end
  end
  T = triu(T(1:n,1:n)); I = eye(n);

% Compute resolvent norms by inverse Lanczos iteration:
  %initialization of minimum singular value, for every z on the grid
  sigmin = Inf*ones(length(y),length(x));
  %start of the loop over grid points
  for i = 1:length(y)
    if isreal(A) & (ymax==-ymin) & (i>length(y)/2);
      sigmin(i,:) = sigmin(length(y)+1-i,:);
    else
      for j = 1:length(x)
        z = zz(i,j); T1 = z*I-T; T2 = T1';
    if real(z)<100                             % Grid Points
     sigold = 0; qold = zeros(n,1); beta = 0; H = [];
     q = randn(n,1) + sqrt(-1)*randn(n,1); q = q/norm(q);
     for k = 1:99;
      v = T1\(T2\q) - beta*qold;
      alpha = real(q'*v); v = v - alpha*q;
      beta = norm(v); qold = q; q = v/beta;
      H(k+1,k) = beta; H(k,k+1) = beta; H(k,k) = alpha;
      if (alpha>1e100), sig = alpha; else sig = max(eig(H(1:k,1:k)));
      end
      if (abs(sigold/sig-1)<.001) | (sig<3 & k>2) break, end
      sigold = sig;
     end
     sigmin(i,j) = 1/sqrt(sig);
    end
   end
  end
 end
% Level lines, i.e. the boundaries of various epsilon-pseudospectra
% 10^{-20} is added for log(0) to be avoided if sigmin=0
 contour(x,y,log10(sigmin+1e-20),-8:-1);
```

Bibliography

[Braconnier and Higham, 1996] Braconnier, T. and Higham, N. J. (1996). Computing the field of values and pseudospectra using the Lanczos method with continuation. *Bit Numer Math*, 36:422–440.

[Ekström and Serra-Capizzano, 2018] Ekström, S.-E. and Serra-Capizzano, S. (2018). Eigenvalues and eigenvectors of banded toeplitz matrices and the related symbols. *Numerical Linear Algebra with Applications*, 25(5):e2137.

[Gong et al., 2016] Gong, F., Meyerson, O., Meza, J., Stoiciu, M., and Ward, A. (2016). Explicit bounds for the pseudospectra of various classes of matrices and operators. *Involve, a Journal of Mathematics*, 9(3):517–540.

[Gustafsson et al., 1972] Gustafsson, B., Kreiss, H., and Sundstroem, A. (1972). Stability theory of difference approximations for mixed initial boundary value problems. *Math. Comp.*, 26:649–686.

[Higham, 2020] Higham, N. (2020). What is a (non)normal matrix? https://nhigham.com/2020/11/24/what-is-a-nonnormal-matrix, Last accessed 2 September 2021.

[Söderlind et al., 2015] Söderlind, G., Jay, L., and Calvo, M. (2015). Stiffness 1952–2012: Sixty years in search of a definition. *Bit Numer Math*, 55:531–558.

[Trefethen, 1999] Trefethen, L. N. (1999). Computation of pseudospectra. *Acta Numerica*, 8:247–295.

[Trefethen, 2000] Trefethen, L. N. (2000). *Spectral Methods in MATLAB*, chapter 9. Oxford University, Oxford, England.

[Trefethen and Embree, 2005] Trefethen, L. N. and Embree, M. (2005). *Spectra and Pseudospectra The Behavior of Nonnormal Matrices and Operators*, chapter 1, 7. Princeton University Press, 41 William Street, Princeton, New Jersey 08540.

[Van Dorsselaer et al., 1994] Van Dorsselaer, J. L. M., Kraaijevanger, J. F. B. M., and Spijker, M. N. (1994). About stability estimates and resolvent conditions. *Numerical Analysis and Mathematical Modelling*, 29(1):215–225.