IPP-Report

Max-Planck-Institut für Plasmaphysik

R. Meyer-Spasche, S. Huber

Spektren und Pseudospektren – Theorie, Numerik und Anwendungen

Spectra and Pseudospectra – Theory, Numerics and Applications

IPP 5/120
April 2008

Vorlesungsmitschrift: Spektren und Pseudospektren

Vorlesung: Rita Meyer-Spasche, rim@ipp.mpg.de;
Ausarbeitung: Sebastian Huber*, sebastian-huber@web.de.
Fassung vom 10. April 2008

Technische Universität München

Zusammenfassung

Abstract This text gives a brief and concise introduction to the classical perturbation theory for eigenvalues of matrices and to the presently developing theory of pseudo spectra of matrices. It is based on a 2-hour course at TUM for students of mathematics, techno-mathematics and physics after the Vordiplom or Bachelor exam, resp.. The course was given in winter terms, both in 2001/2002 and in 2004/2005.

Since Lyapunov's pioneering work in the 19th century, stability investigations involve the computation of eigenvalues resp. spectra. It is known for a long time that unavoidable perturbations of mathematical models result in perturbations of eigenvalues and thus might affect the stability results. Until recently, this was not taken into account in a systematic way, but only on an individual basis.

Part 1 of this text deals with the classical estimates of how much the eigenvalues change if the elements of a matrix are perturbed. Part 2 deals with pseudospectra for quadratic and rectangular matrices, i.e. with the developing new theory of how to take perturbations into account in a more systematic way. Part 3 touches some applications to stability of flows within the framework of dynamical systems, and Part 4 lists the exercises given to the students.

*attended the course in winter term 2004/2005 as a student, not a member of IPP

Inhaltsverzeichnis

1 Gestörte Matrizen und Eigenpaare: Klassische Abschätzungen

1.1 Grundbegriffe (Wiederholung)

Es sei $A = (a_{ij}) \in \mathbb{C}^{n\times n}$. Das *charakteristische Polynom* ist definiert durch $p(z) := \det(zI - A)$. Es hat den Grad $\deg p = n$ und n Nullstellen $p(\lambda_i) = 0$ in $\mathbb{C}$, $i = 1, \ldots, n$. Die Nullstellen λ_i von p sind die *Eigenwerte* von A. Die Menge

$$\lambda(A) := \{\lambda_1, \ldots, \lambda_n\} = \{\lambda \in \mathbb{C} : \det(\lambda I - A) = 0\}$$

heißt *Spektrum von A*. Für jeden Eigenwert λ_i gibt es mindestens einen Vektor $x_i \neq 0$ mit $\|x_i\| = 1$, so dass $Ax_i = \lambda_i x_i$. Die Vielfachheit von λ_i als Nullstelle von $p(z)$ heißt *algebraische Vielfachheit* des Eigenwerts λ_i. Für die Determinante von A gilt $\det(A) = \lambda_1 \cdot \ldots \cdot \lambda_n$ und für die Spur gilt $\mathrm{tr}(A) = \sum_{i=1}^n a_{ii} = \sum_{i=1}^n \lambda_i$. Zu jedem Eigenwert λ gibt es einen *rechten Eigenvektor* $x \in \mathbb{C}^n\backslash\{0\}$ mit $Ax = \lambda x$ und einen *linken Eigenvektor* $y \in \mathbb{C}^n\backslash\{0\}$ mit $y^H A = \lambda y^H$. Es gilt $(y^H A)^H = (\lambda y^H)^H = \bar{\lambda} y = A^H y$ und damit

$$A = A^H \implies \lambda(A) \subset \mathbb{R}.$$

Ist $A \in \mathbb{R}^{n\times n}$, so gilt zudem

$$\lambda \in \lambda(A) \iff \bar{\lambda} \in \lambda(A).$$

Es sei nun A als Endomorphismus $A : \mathbb{C}^n \mapsto \mathbb{C}^n$ aufgefasst. Das *Bild von A* ist definiert durch $\mathrm{range}\, A := \{y \in \mathbb{C}^n : \exists x \in \mathbb{C}^n \text{ mit } Ax = y\}$. Es sei $S \subset \mathbb{C}^n$ ein linearer Raum. Falls $A(S) \subset S$, dann heißt S *invariant unter A*. Für linear unabhängige Eigenvektoren $Ax_i = \lambda x_i$ mit $i = 1, \ldots, q \leq n$ definiere $S_\lambda := \mathrm{lin}\,\{x_1, \ldots, x_q\}$. Es gilt $\dim S_\lambda = q \leq n$. Ist S_λ maximaler invarianter Unterraum zum Eigenwert λ, so heißt $q = \dim S_\lambda$ *geometrische Vielfachheit* des Eigenwerts λ. Sind die geometrische und algebraische Vielfachheit voneinander verschieden, so hat der Eigenwert λ einen *Defekt*. Wenn A einen defekten Eigenwert hat, so ist A *defekt*.

1.2 Beispiele

1.2.1. Beispiel.

$$A = \begin{pmatrix} \lambda_1 & 1 \\ 0 & \lambda_2 \end{pmatrix}$$

Für $\lambda_1 \neq \lambda_2$ haben λ_1 und λ_2 die algebraische und geometrische Vielfachheit 1.

1.2.2. Beispiel.

$$B = \begin{pmatrix} \lambda_1 & 1 \\ 0 & \lambda_1 \end{pmatrix}$$

Das ist ein 2×2-Jordanbock, also ist λ_1 algebraisch 2-facher und geometrisch 1-facher Eigenwert. B hat einen Defekt.

1.2.3. Beispiel.

$$M_{01}(\epsilon) = \begin{pmatrix} 0 & 1 & & & \\ & 0 & 1 & & \\ & & \ddots & \ddots & \\ & & & 0 & 1 \\ \epsilon & & & & 0 \end{pmatrix}_{n\times n}$$

$\epsilon = 0$**:** $M_{01}(0)$ ist ein $n \times n$-Jordanblock, also ist $\lambda = 0$ algebraisch n-facher und geometrisch 1-facher Eigenwert. $M_{01}(0)$ ist singulär.

$\epsilon > 0$**:** Für alle Eigenwerte λ muss gelten

$$p(\lambda) = \det(\lambda I - M_{01}(\epsilon)) = \lambda^n + (-1)^{n+1}(-\epsilon)(-1)^{n-1} = 0 \iff \lambda^n = \epsilon.$$

Also sind die n-ten komplexen Wurzeln aus ϵ die Eigenwerte und es gilt für $j = 1, \ldots, n$

$$\lambda_j = \left|\sqrt[n]{\epsilon}\right| \; e^{j\frac{2\pi i}{n}}.$$

Ist beispielsweise $\epsilon = 10^{-n}$ und $n = 10$, so ist $|\lambda_j| = 10^{-1}$. Die unsymmetrische Störung mit $\epsilon = 10^{-10}$ bewirkt eine symmetrische Störung des Eigenwertes $\lambda = 0$ mit dem Betrag 10^{-1}.

Wir wollen untersuchen:

- Wie stark kann eine Störung der Elemente der Matrix der Größe $|\epsilon|$ die Eigenpaare stören? Welche Abschätzungen der Störung gibt es?
- Wie stark hängt die resultierende Störung von der Art der Störung und der Struktur der Matrix ab?

1.3 Gerschgorinkreise

1.3.1. Satz (Gerschgorin 1931). *Es sei $A \in \mathbb{C}^{n\times n}$. Dann gilt*

$$\lambda(A) \subset \bigcup_{i=1}^{n} B_i$$

mit den Gerschgorinkreisen

$$B_i := \left\{ z \in \mathbb{C} : |z - a_{ii}| \leq \sum_{\substack{j=1 \\ j\neq i}}^{n} |a_{ij}| \right\}.$$

Beweis. Es sei $x = (x_1, \ldots, x_n)^T$ ein Eigenvektor zu einem Eigenwert λ von A. Wähle $r \in \{1, \ldots, n\}$ so, dass $\max_i |x_i| = |x_r|$ und normiere so, dass $x_r = 1$. Damit gilt $|x_i| \leq 1$ für alle $i \in \{1, \ldots, n\}$. Betrachte die r-te Gleichung in $Ax = \lambda x$:

$$\sum_{j=1}^{n} a_{rj} x_j = \lambda x_r = \lambda$$

und damit

$$|\lambda - a_{rr}| = \left| \sum_{\substack{j=1 \\ j \neq r}}^{n} a_{rj} x_j \right| \leq \sum_{\substack{j=1 \\ j \neq r}}^{n} |a_{rj} x_j| \leq \sum_{\substack{j=1 \\ j \neq r}}^{n} |a_{rj}| .$$

Also ist $\lambda \in B_r$. Mit den selben Überlegungen für die anderen Eigenwerte folgt die Behauptung. □

1.3.2. Bemerkung. Es seien $X^{-1}AX = D + F$ mit $D = \operatorname{diag}(d_1, \ldots, d_n)$ und $f_{ii} = 0$. Dann gilt $AX = X(D + F)$ und mit Aufgabe 1 folgt $\lambda(A) = \lambda(D + F)$. Also gilt

$$\lambda(A) \subset \bigcup_{i=1}^{n} \widetilde{B}_i$$

mit

$$\widetilde{B}_i := \left\{ z \in \mathbb{C} : |z - d_i| \leq \sum_{j=1}^{n} |f_{ij}| \right\}.$$

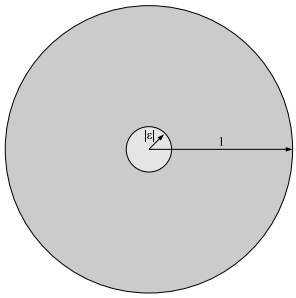
1.3.3. Bemerkung. Es gilt $\lambda(A) = \lambda(A^T)$. Man kann also Satz 1.3.1 auf Zeilen- oder Spaltensummen anwenden, um die Radien der Gerschgorinkreise zu erhalten.

1.3.4. Beispiel. Es sei $A = \operatorname{diag}(d_1, \ldots, d_n)$. Bei Diagonalmatrizen reduzieren sich die Gerschgorinkreise auf Punkte. Wird ein Element von A von der Größe ϵ gestört und liegt kein zweites Diagonalelement in dem ϵ-Gerschgorinkreis, so ändert sich höchstens ein Eigenwert von A um höchstens ϵ.

1.3.5. Beispiel.

$$M_{01}(\epsilon) = \begin{pmatrix} 0 & 1 & & & \\ & 0 & 1 & & \\ & & \ddots & \ddots & \\ & & & 0 & 1 \\ \epsilon & & & & 0 \end{pmatrix}_{n \times n}$$

$\epsilon \neq 0$**:** Bei diesem Beispiel überschneiden sich alle Kreise:



Bisher wissen wir nur: Alle Eigenwerte liegen in der Vereinigung aller Gerschgorinkreise. Angenommen, die Gerschgorinkreise überschneiden sich nicht. Liegt dann in jedem Kreis genau ein Eigenwert?

1.3.6. Satz (Gerschgorin 1931). *Es seien* $A \in \mathbb{C}^{n\times n}$, $\bigcup_{i=1}^{s} B_i$ *zusammenhängend und* $(\bigcup_{i=1}^{s} B_i) \cap (\bigcup_{i=s+1}^{n} B_i) = \emptyset$. *Dann liegen genau* s *Eigenwerte von* A *in* $\bigcup_{i=1}^{s} B_i$.

Beweis. Die Eigenwerte einer quadratischen Matrix hängen stetig von den Elementen der Matrix ab [Ka82, S. 123]. Es seien

$$r_i := \sum_{\substack{j=1 \\ j\neq i}}^{n} |a_{ij}|$$

für $i = 1, \ldots, n$ die Radien der Gerschgorinkreise, $D := \operatorname{diag}(a_{11}, \ldots, a_{nn})$, $C := A - D$ und $F(t) := D + tC$ mit $t \in [0, 1]$. Es gilt $F(0) = D$ und $F(1) = A$. Es sei $p(\lambda, t)$ das charakteristische Polynom von $F(t)$. Die Koeffizienten von p und die Nullstellen $\lambda_i(t)$ hängen stetig von t ab. Es gilt

$$\left(\bigcup_{i=1}^{s} B(a_{ii}, r_i t)\right) \cap \left(\bigcup_{i=s+1}^{n} B(a_{ii}, r_i t)\right) = \emptyset$$

für $t = 1$ und also für alle $t \in [0, 1]$. Für $t = 0$ enthält $\bigcup_{i=1}^{s} B(a_{ii}, 0)$ genau s Eigenwerte von $F(0)$. Hieraus folgt aus Stetigkeitsgründen, dass für $0 \leq t \leq 1$ die Menge $\bigcup_{i=1}^{s} B(a_{ii}, r_i t)$ genau s Eigenwerte enthält. Also auch für $F(1) = A$. □

1.3.7. Bemerkung. Es sei $A \in \mathbb{C}^{n\times n}$. Da ähnliche Matrizen die gleichen Eigenwerte haben, kann man A mit einer Diagonalmatrix $D = \operatorname{diag}(d_1, \ldots, d_n)$ skalieren, ohne dass sich die Eigenwerte ändern: $D^{-1}AD$. Damit ergibt sich

$$|\lambda - a_{ii}| \leq \sum_{\substack{j=1 \\ j\neq i}}^{n} \left|\frac{a_{ij} d_j}{d_i}\right|.$$

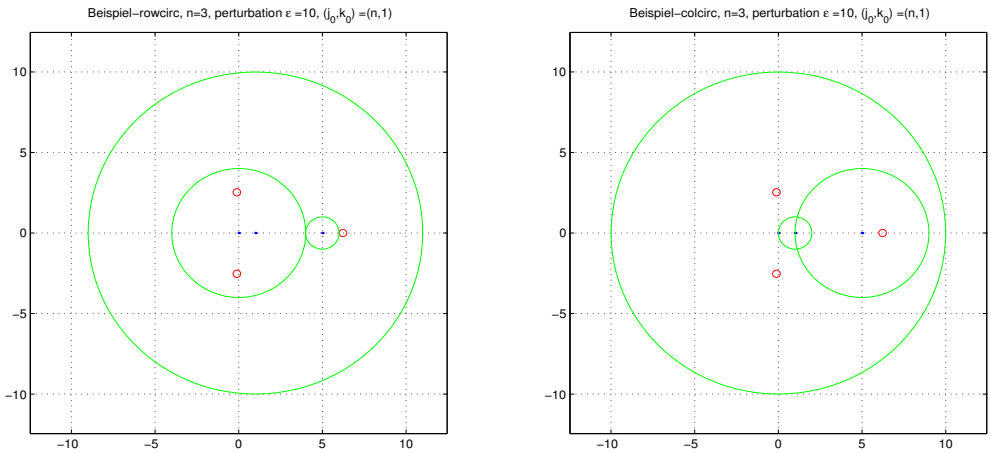


Abbildung 1: Eigenwerte (rot) und Gerschgorinkreise (grün) von Matrix M aus Bemerkung 1.3.8, mit Zeilenradien (links) und Spaltenradien (rechts).

1.3.8. Bemerkung. Liegt innerhalb einer Zusammenhangskomponente in jedem der Kreise jeweils ein Eigenwert? Nein: Die Matrix

$$M = \begin{pmatrix} 0 & 4 & 0 \\ 0 & 5 & 1 \\ 10 & 0 & 1 \end{pmatrix}$$

hat die Eigenwerte

$$\begin{pmatrix} \lambda_1 \\ \lambda_2 \\ \lambda_3 \end{pmatrix} \approx \begin{pmatrix} -0.1142 + 2.5316i \\ -0.1142 - 2.5316i \\ 6.2284 \end{pmatrix}$$

mit dem zeilenweisen Gerschgoringebiet

$$B(0,4) \cup B(5,1) \cup B(1,10).$$

Es liegt kein Eigenwert in $B(5,1)$; oder spaltenweise:

$$B(0,10) \cup B(5,4) \cup B(1,1).$$

Es liegt kein Eigenwert in $B(1,1)$. Die zeilenweisen und spaltenweisen Gerschgoringebiete sind in Abbildung 1 zu sehen, zusammen mit den drei Eigenwerten der Matrix M.

1.4 Der Satz von Bauer-Fike

Ziel dieses Abschnittes ist Satz 1.4.4. Vorbereitungen: Es sei $x \in \mathbb{C}^n$. Die *p-Norm* ist für $p \geq 1$ definiert durch

$$\|x\|_p := \sqrt[p]{\sum_{j=1}^{n} |x_j|^p}.$$

Es sei $M \in \mathbb{C}^{n\times n}$, dann ist die zu $\|\cdot\|_p$ gehörige *Operatornorm* definiert durch

$$\|M\|_p := \sup_{x\neq 0} \frac{\|Mx\|_p}{\|x\|_p} = \sup_{\|x\|_p = 1} \|Mx\|_p .$$

Es gilt

$$\|Mx\|_p \leq \|M\|_p \, \|x\|_p .$$

Für $A \in \mathbb{C}^{m\times n}$, $B \in \mathbb{C}^{n\times q}$ und $AB \in \mathbb{C}^{m\times q}$ gilt

$$\|AB\|_p \leq \|A\|_p \, \|B\|_p .$$

Die *Konditionszahl* bezüglich dem Lösen linearer Gleichungssysteme $Mx = b$ ist für reguläres M definiert durch

$$\mathcal{K}_p(M) := \|M\|_p \left\|M^{-1}\right\|_p .$$

Die Abbildung $\Phi : \mathbb{C}^{n\times n} \mapsto \mathbb{R}^{2n\times 2n}$, $A + iB \mapsto \begin{pmatrix} A & -B \\ B & A \end{pmatrix}$ ist eine Einbettung des $\mathbb{C}^{n\times n}$ in den $\mathbb{R}^{2n\times 2n}$. Sie hat folgende Eigenschaften:

$$\begin{aligned} \text{unitär} &\implies \text{orthogonal} \\ \text{hermitesch} &\implies \text{symmetrisch} \\ \text{positiv definit} &\implies \text{positiv definit} \\ (A+iB)(x+iy) = b+ic &\implies \begin{pmatrix} A & -B \\ B & A \end{pmatrix} \begin{pmatrix} x \\ y \end{pmatrix} = \begin{pmatrix} b \\ c \end{pmatrix} . \end{aligned}$$

Wir untersuchen nun die Auswirkung von Störungen in A auf $\|A^{-1}\|_p$.

1.4.1. Lemma. *Es sei $F \in \mathbb{C}^{n\times n}$ und $\|F\|_p < 1$ für ein $p \geq 1$. Dann ist $I - F$ regulär. Es gilt*

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

und

$$\left\|(I-F)^{-1}\right\|_p \leq \frac{1}{1-\|F\|_p}.$$

Beweis. Angenommen $I - F$ ist singulär. Dann existiert ein $x \neq 0$ mit $(I-F)x = 0$ und damit $Fx = x$. Es ist

$$0 \neq \|x\|_p = \|Fx\|_p \leq \|F\|_p \, \|x\|_p .$$

Damit folgt $\|F\|_p \geq 1$ im Widerspruch zur Voraussetzung, also ist $I - F$ regulär. Es gilt

$$\left(\sum_{k=0}^{N} F^k\right)(I-F) = I - F^{N+1}$$

und damit

$$\left(\lim_{N\to\infty} \sum_{k=0}^{N} F^k\right)(I-F) = I - \lim_{N\to\infty} F^{N+1}.$$

Mit

$$\left\|F^k\right\|_p \leq \|F\|_p^k \xrightarrow{k\to\infty} 0$$

folgt $F^k \xrightarrow{k\to\infty} 0$ und damit $\left(\sum_{k=0}^{\infty} F^k\right)(I-F) = I$, also $(I-F)^{-1} = \sum_{k=0}^{\infty} F^k$. Daraus folgt

$$\left\|(I-F)^{-1}\right\|_p = \left\|\sum_{k=0}^{\infty} F^k\right\|_p \leq \sum_{k=0}^{\infty} \left\|F^k\right\|_p \leq \sum_{k=0}^{\infty} \|F\|_p^k = \frac{1}{1-\|F\|_p}.$$

□

1.4.2. Satz. *Es seien A regulär und $\|A^{-1}E\|_p < 1$. Dann ist $A+E$ regulär und*

$$\left\|(A+E)^{-1} - A^{-1}\right\|_p \leq \|E\|_p \left\|A^{-1}\right\|_p^2 \frac{1}{1-\|A^{-1}E\|_p}.$$

Beweis. A ist regulär, also existiert $F := -A^{-1}E$ und es ist

$$A+E = A(I+A^{-1}E) = A(I-F).$$

Es gilt nach Voraussetzung $\|F\|_p < 1$. Mit Lemma 1.4.1 folgt $I-F$ und also $A+E$ ist regulär und $\left\|(I-F)^{-1}\right\|_p \leq \frac{1}{1-\|F\|_p}$. Zusammen mit $(A+E)^{-1} = (I-F)^{-1}A^{-1}$ folgt

$$\left\|(A+E)^{-1}\right\|_p \leq \frac{\|A^{-1}\|_p}{1-\|F\|_p}.$$

Weiter gilt

$$-A^{-1}E = I - I - A^{-1}E = I - A^{-1}(A+E)$$

und damit

$$-A^{-1}E\,(A+E)^{-1} = (A+E)^{-1} - A^{-1}.$$

Also

$$\left\|(A+E)^{-1} - A^{-1}\right\|_p = \left\|-A^{-1}E\,(A+E)^{-1}\right\|_p \leq \left\|A^{-1}\right\|_p \|E\|_p \left\|(A+E)^{-1}\right\|_p.$$

□

1.4.3. Bemerkung. Ist $A \approx I$, so bewirkt eine Störung von $\mathcal{O}(\epsilon)$ in A eine Störung von $\mathcal{O}(\epsilon)$ in A^{-1}. Ist dagegen $\det(A) \approx 0$, so ist $\|A^{-1}\|_p$ sehr groß und A^{-1} kann sehr viel empfindlicher bzgl. Störungen sein.

Wir untersuchen nun die Auswirkungen von Störungen in A und b auf $\|x\|_p$ in $Ax = b$.

Gegeben sei folgendes parametrisches System $(A+\epsilon F)x(\epsilon) = b + \epsilon f$ mit $x(0) = x$, $\epsilon \in \mathbb{R}$, $A, F \in \mathbb{C}^{n\times n}$ und $x, b, f \in \mathbb{C}^n$. Ist A regulär, so folgt $x(\epsilon) = A^{-1}(b + \epsilon f - \epsilon F x(\epsilon))$.

Es gilt $\dot{x}(0) = A^{-1}(f - Fx(0) - 0 \cdot F\dot{x}(0)) = A^{-1}(f - Fx(0))$. Eine Taylorentwicklung von $x(\epsilon)$ um 0 ergibt $x(\epsilon) = x(0) + \epsilon\dot{x}(0) + \mathcal{O}(\epsilon^2)$. Damit folgt

$$\begin{aligned}
\frac{\|x(\epsilon) - x\|_p}{\|x\|_p} &= \frac{\|x(\epsilon) - x(0)\|_p}{\|x(0)\|_p} \\
&\leq |\epsilon| \frac{\|A^{-1}(f - Fx(0))\|_p}{\|x(0)\|_p} + \mathcal{O}(\epsilon^2) \\
&\leq |\epsilon| \left\|A^{-1}\right\|_p \left(\frac{\|f\|_p}{\|x(0)\|_p} + \|F\|_p \right) + \mathcal{O}(\epsilon^2) \\
&\leq |\epsilon| \left\|A^{-1}\right\|_p \left(\frac{\|f\|_p}{\|b\|_p} \|A\|_p + \|F\|_p \right) + \mathcal{O}(\epsilon^2) \\
&\leq \underbrace{\|A\|_p \left\|A^{-1}\right\|_p}_{\mathcal{K}_p(A)} \left(\underbrace{|\epsilon| \frac{\|f\|_p}{\|b\|_p}}_{\rho_b} + \underbrace{|\epsilon| \frac{\|F\|_p}{\|A\|_p}}_{\rho_A} \right) + \mathcal{O}(\epsilon^2)
\end{aligned}$$

mit der Konditionszahl $\mathcal{K}_p(A)$ von A, dem relativen Fehler ρ_b von b und dem relativen Fehler ρ_A von A. Also gilt für den relativen Fehler ρ_x von x

$$\rho_x \leq \mathcal{K}_p(A)(\rho_A + \rho_b)$$

für genügend kleine ϵ.

1.4.4. Satz (Bauer-Fike 1960). *Es sei μ Eigenwert von $A + E \in \mathbb{C}^{n\times n}$ und $X^{-1}AX = D$ mit $D = \operatorname{diag}(\lambda_1, \ldots, \lambda_n)$. Dann gilt*

$$\min_{\lambda \in \lambda(A)} |\lambda - \mu| \leq \mathcal{K}_p(X) \|E\|_p$$

für jede p-Norm auf $\mathbb{C}^{n\times n}$.

Beweis. Für $\mu \in \lambda(A)$ ist nichts zu beweisen. Es sei also $\mu \notin \lambda(A)$. Dann ist $A + E - \mu I$ singulär und $D - \mu I$ regulär. Damit ist $X^{-1}(A + E - \mu I)X$ singulär und folglich auch

$$\underbrace{X^{-1}AX}_{D} + X^{-1}EX - \underbrace{X^{-1}\mu I X}_{\mu I} = (D - \mu I)(I + (D - \mu I)^{-1} X^{-1}EX).$$

Also ist $I + (D - \mu I)^{-1} X^{-1}EX$ singulär und es folgt mit Lemma 1.4.1

$$1 \leq \left\|(D - \mu I)^{-1} X^{-1}EX\right\|_p \leq \left\|(D - \mu I)^{-1}\right\|_p \|E\|_p \underbrace{\left\|X^{-1}\right\|_p \|X\|_p}_{=\mathcal{K}_p(X)}.$$

Es ist $(D - \mu I)^{-1} = \operatorname{diag}\left((\lambda_1 - \mu)^{-1}, \ldots, (\lambda_n - \mu)^{-1}\right)$ und damit

$$\left\|(D - \mu I)^{-1}\right\|_p \leq \max_i \frac{1}{|\lambda_i - \mu|} = \frac{1}{\min_i |\lambda_i - \mu|}.$$

Damit ist alles gezeigt. □

1.5 Abschätzung mit der Schur-Zerlegung

1.5.1. Lemma. *Die Matrix $T \in \mathbb{C}^{n\times n}$ sei eine rechte obere Blockdreiecksmatrix mit*

$$T = \begin{pmatrix} T_{11} & T_{12} \\ 0 & T_{22} \end{pmatrix}.$$

Dann gilt $\lambda(T) = \lambda(T_{11}) \cup \lambda(T_{22})$.

Beweis. Für einen Eigenvektor x zum Eigenwert λ von T gilt

$$Tx = \begin{pmatrix} T_{11} & T_{12} \\ 0 & T_{22} \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \end{pmatrix} = \lambda \begin{pmatrix} x_1 \\ x_2 \end{pmatrix}.$$

Also $T_{22}x_2 = \lambda x_2$. Gilt $x_2 \neq 0$, so folgt $\lambda \in \lambda(T_{22})$. Ist andererseits $x_2 = 0$, so folgt $\lambda \in \lambda(T_{11})$. Damit gilt $\lambda(T) \subset \lambda(T_{11}) \cup \lambda(T_{22})$. Beide Mengen haben n Elemente, müssen also gleich sein. □

1.5.2. Lemma. *Es seien $A \in \mathbb{C}^{n\times n}$, $B \in \mathbb{C}^{p\times p}$ und $X \in \mathbb{C}^{n\times p}$ mit $AX = XB$,* $\operatorname{rank} X = p$ *und $p \leq n$. Dann existiert ein $Q \in \mathbb{C}^{n\times n}$, $Q^HQ = I$ mit*

$$Q^HAQ = T = \begin{pmatrix} T_{11} & T_{12} \\ 0 & T_{22} \end{pmatrix}$$

und $\lambda(T_{11}) = \lambda(A) \cap \lambda(B)$.

Beweis. Es sei $X = Q \begin{pmatrix} R \\ 0 \end{pmatrix}$ eine QR-Zerlegung von X mit $Q \in \mathbb{C}^{n\times n}$, $Q^HQ = I$ und einer oberen Dreiecksmatrix $R \in \mathbb{C}^{p\times p}$. Es ist $\operatorname{rank} R = \operatorname{rank} X = p$ und es gilt

$$AQ \begin{pmatrix} R \\ 0 \end{pmatrix} = Q \begin{pmatrix} R \\ 0 \end{pmatrix} B \iff Q^HAQ \begin{pmatrix} R \\ 0 \end{pmatrix} = \begin{pmatrix} R \\ 0 \end{pmatrix} B.$$

Es sei

$$\begin{pmatrix} T_{11} & T_{12} \\ T_{21} & T_{22} \end{pmatrix} \begin{pmatrix} R \\ 0 \end{pmatrix} = \begin{pmatrix} R \\ 0 \end{pmatrix} B.$$

Damit folgt $T_{11}R = RB$ und somit $\lambda(T_{11}) = \lambda(B)$. Aus Aufgabe 1 folgt $\lambda(B) \subset \lambda(A)$. Also gilt $\lambda(T_{11}) = \lambda(A) \cap \lambda(B)$. Aus $T_{21}R = 0$ folgt $T_{21} = 0$. □

1.5.3. Bemerkung. Anwendung für $p \geq 1$: Falls ein invarianter Unterraum von A bekannt ist, so kann man mit Lemma 1.5.2 die Matrix A unitär in Blockdreiecksform transformieren. Iterieren dieses Vorgehens führt auf die Schur-Zerlegung von A.

1.5.4. Satz (Schur 1909). *Für $A \in \mathbb{C}^{n\times n}$ existiert ein unitäres $Q \in \mathbb{C}^{n\times n}$ mit $Q^HAQ = T = D + N$, wobei $D = \operatorname{diag}(\lambda_1, \ldots, \lambda_n)$ und $N \in \mathbb{C}^{n\times n}$ strikt in oberer Dreiecksform ist. Q kann so gewählt werden, dass die λ_i in beliebiger Reihenfolge in der Diagonalen erscheinen.*

Beweis. Der Satz stimmt für $n = 1$. Angenommen er stimmt für $n-1$ und $Ax = \lambda x$ mit $x \neq 0$. Aus Lemma 1.5.2 mit $B = (\lambda)_{1\times 1}$ folgt, dass ein U existiert mit $U^H U = I$ und $U^H A U = \begin{pmatrix} \lambda & w^H \\ 0 & C \end{pmatrix}$. Weiter existiert V mit $V^H C V$ in oberer Dreiecksform und $Q := U \begin{pmatrix} 1 & 0 \\ 0 & V \end{pmatrix}$. Damit ist $Q^H A Q$ in oberer Dreiecksform. □

1.5.5. Bemerkung. Die Spalten q_k von $Q = (q_1, \ldots, q_n)$ heißen *Schur-Vektoren.* Für die Spalten von $AQ = Q(D + N)$ gilt

$$Aq_k = \lambda_k q_k + \sum_{i=1}^{k-1} n_{ik} q_i$$

für $k = 1, \ldots, n$. Weiterhin gilt:

- Die q_k sind nicht unbedingt invariant unter A, aber $S_k := \operatorname{lin}\{q_1, \ldots, q_k\}$ ist invariant, $k = 1, \ldots, n$. Sei $Q_k := (q_1, \ldots, q_k)$. Dann folgt $\lambda(Q_k^H A Q_k) = \{\lambda_1, \ldots, \lambda_k\}$. Man kann die Reihenfolge der Eigenwerte beliebig wählen: Zu jeder Menge von k Eigenwerten von A gibt es also mindestens einen k-dimensionalen invarianten Unterraum.
- Die Spalte q_k ist invariant und Eigenvektor von A genau dann, wenn $n_k = 0$ (k-te Spalte von N).

Unter welchen Bedingungen ist $n_k = 0$ für alle $k \in \{1, \ldots, n\}$? Wie wir gleich sehen werden, genau dann, wenn A normal ist.

1.5.6. Lemma. *$A \in \mathbb{C}^{n\times n}$ erfüllt $AA^H = A^H A$ genau dann, wenn ein $Q \in \mathbb{C}^{n\times n}$ existiert mit $Q^H Q = I$, so dass $Q^H A Q = D = \operatorname{diag}(\lambda_1, \ldots, \lambda_n)$.*

Beweis. Es sei $Q^H A Q = D$, dann gilt

$$Q^H A A^H Q = Q^H A Q Q^H A^H Q = D D^H = D^H D = Q^H A^H Q Q^H A Q = Q^H A^H A Q.$$

Also $AA^H = A^H A$.

Ist andererseits $AA^H = A^H A$, so existieren nach Satz 1.5.4 Matrizen $Q, T \in \mathbb{C}^{n\times n}$ mit $Q^H Q = I$ und $Q^H A Q = T = D + N$. Aus $AA^H = A^H A$ folgt $T^H T = T T^H$. Es ist

$$\begin{aligned} T^H T &= (D+N)^H (D+N) \\ &= (D^H + N^H)(D+N) \\ &= D^H D + N^H D + D^H N + N^H N \end{aligned}$$

und

$$\begin{aligned} T T^H &= (D+N)(D+N)^H \\ &= (D+N)(D^H + N^H) \\ &= D D^H + N D^H + D N^H + N N^H. \end{aligned}$$

Für

$$N = \begin{pmatrix} 0 & a \\ 0 & 0 \end{pmatrix}_{2\times 2} \quad \text{und} \quad D = \begin{pmatrix} \lambda_1 & 0 \\ 0 & \lambda_2 \end{pmatrix}_{2\times 2}$$

gilt:

$$N^H D = \begin{pmatrix} 0 & 0 \\ \bar{a}\lambda_1 & 0 \end{pmatrix} \qquad D^H N = \begin{pmatrix} 0 & \bar{\lambda_1} a \\ 0 & 0 \end{pmatrix} \qquad N^H N = \begin{pmatrix} 0 & 0 \\ 0 & \bar{a}a \end{pmatrix}$$
$$ND^H = \begin{pmatrix} 0 & a\bar{\lambda_2} \\ 0 & 0 \end{pmatrix} \qquad DN^H = \begin{pmatrix} 0 & 0 \\ \lambda_2\bar{a} & 0 \end{pmatrix} \qquad NN^H = \begin{pmatrix} a\bar{a} & 0 \\ 0 & 0 \end{pmatrix}.$$

Für $n = 2$ kann $T^H T = TT^H$ also nur gelten für $N = 0$. Vollständige Induktion nach n liefert $N = 0$ für beliebiges n. □

1.5.7. Definition. Es sei $C \in \mathbb{C}^{m\times n}$, dann ist die *Frobeniusnorm* definiert durch

$$\|C\|_F := \sqrt{\sum_{i=1}^{m}\sum_{j=1}^{n} |c_{ij}|^2}.$$

Es sei $Q^H AQ = T = \operatorname{diag}(\lambda_1, \ldots, \lambda_n) + N$, dann ist

$$\|N\|_F^2 = \|A\|_F^2 - \sum_{i=1}^{n} |\lambda_i|^2$$

die *Abweichung von A von der Normalität.*

1.5.8. Satz. *Es sei $Q^H AQ = D + N$ eine Schur-Zerlegung von $A \in \mathbb{C}^{n\times n}$ mit $Q^H Q = I$, $D = \operatorname{diag}(\lambda_1, \ldots, \lambda_n)$, N eine echte obere Dreiecksmatrix, $q \in \mathbb{N}$ die kleinste natürliche Zahl mit $|N|^q = (|n_{ij}|)^q = 0$, $E \in \mathbb{C}^{n\times n}$ und $\mu \in \lambda(A + E)$. Dann gilt*

$$\min_{\lambda\in\lambda(A)} |\lambda - \mu| \le \max\left\{\theta, \sqrt[q]{\theta}\right\}$$

mit

$$\theta := \|E\|_2 \sum_{k=0}^{q-1} \|N\|_2^k.$$

Beweis. Für $\mu \in \lambda(A)$ ist die Aussage trivial. Es sei also $\mu \notin \lambda(A)$. Setze

$$\delta := \min_{\lambda\in\lambda(A)} |\lambda - \mu| = \frac{1}{\left\|(\mu I - D)^{-1}\right\|_2} \neq 0.$$

Aus $\mu \in \lambda(A + E)$ folgt, dass $(\mu I - A) - E$ singulär ist. Also ist auch $I - (\mu I - A)^{-1} E$ singulär. Mit Lemma 1.4.1 gilt

$$\begin{aligned} 1 &\le \left\|(\mu I - A)^{-1} E\right\|_2 \\ &\le \left\|(\mu I - A)^{-1}\right\|_2 \|E\|_2 \\ &= \left\|Q^H (\mu I - D - N)^{-1} Q\right\|_2 \|E\|_2 \\ &\le \left\|(\mu I - D - N)^{-1}\right\|_2 \|E\|_2 . \end{aligned} \tag{1}$$

Es ist

$$\begin{aligned}(\mu I - D - N)^{-1} &= ((\mu I - D)(I - (\mu I - D)^{-1} N)^{-1} \\ &= \left(I - \underbrace{(\mu I - D)^{-1} N}_{=:F} \right)^{-1} (\mu I - D)^{-1}. \end{aligned} \tag{2}$$

Nun gehen wir analog zum Beweis von Lemma 1.4.1 vor. Da $|N|^q$ und $(\mu I - D)^{-1}$ diagnonal ist, folgt $((\mu I - D)^{-1} N)^q = 0$. Man braucht $\|F\|_2 < 1$ nicht, weil F nilpotent ist. Damit ist die Summe endlich:

$$\left(\sum_{k=0}^{q-1} F^k \right) (I - F) = I - F^q = I.$$

Also ist

$$(I - F)^{-1} = \sum_{k=0}^{q-1} ((\mu I - D)^{-1} N)^k.$$

Mit (2) folgt

$$\left\| (\mu I - D - N)^{-1} \right\|_2 = \left\| (\mu I - D)^{-1} \sum_{k=0}^{q-1} ((\mu I - D)^{-1} N)^k \right\|_2 \leq \frac{1}{\delta} \sum_{k=0}^{q-1} \left(\frac{\|N\|_2}{\delta} \right)^k.$$

Um weiter abzuschätzen, müssen wir nun zwei Fälle unterscheiden. Für $\delta > 1$ gilt

$$\left\| (\mu I - D - N)^{-1} \right\|_2 \leq \frac{1}{\delta} \sum_{k=0}^{q-1} \|N\|_2^k$$

und mit (1) folgt

$$1 \leq \frac{1}{\delta} \|E\|_2 \sum_{k=0}^{q-1} \|N\|_2^k, \qquad \text{also } \delta \leq \theta.$$

Für $\delta \leq 1$ folgt

$$\left\| (\mu I - D - N)^{-1} \right\|_2 \leq \frac{1}{\delta^q} \sum_{k=0}^{q-1} \|N\|_2^k$$

und mit (1) folgt $\delta \leq \sqrt[q]{\theta}$. Also gilt insgesamt

$$\min_{\lambda \in \lambda(A)} |\lambda - \mu| \leq \max \left\{ \theta, \sqrt[q]{\theta} \right\}.$$

□

1.5.9. Bemerkung. Vergleich der Abschätzungen für $A + E$ von Satz 1.4.4 mit

$$X^{-1} A X = D \;\text{ und }\; \delta \leq \mathcal{K}_p(X) \|E\|_p, \quad p \geq 1,$$

und von Satz 1.5.8 mit

$$Q^H A Q = D + N, \quad \theta = \|E\|_2 \sum_{k=0}^{q-1} \|N\|_2^k \quad \text{und} \quad \delta \leq \max\left\{\theta, \sqrt[q]{\theta}\right\}.$$

Die Schur-Schranke ist allgemeiner anwendbar. Für normale Matrizen liefern beide Sätze $\delta \leq \|E\|_p$ für $p = 2$, aber Satz 1.4.4 erlaubt auch andere p-Normen, $p \geq 1$.

Welche Abschätzung ist besser, wenn beide möglich sind? Vermutlich mal die eine, mal die andere.

Die Schwäche beider Abschätzungen: Sie werfen alle Eigenwerte einer Matrix in einen Topf. Der schlechtest-konditionierte bestimmt die Schranke. Dies ist sehr ungünstig, wenn eine Matrix sehr unterschiedlich empfindliche Eigenwerte hat, die in disjunkten Gerschgoringebieten liegen. Im nächsten Abschnitt sehen wir uns an, wie die Kondition einzelner Eigenwerte definiert werden kann.

1.5.10. Beispiel.

$$\begin{aligned} B(10^{-3}, 10^{-3}) &= \begin{pmatrix} 1 & 2 & 3 \\ 0 & 4 & 5 \\ 10^{-3} & 0 & 4 + 10^{-3} \end{pmatrix} \\ \mu_1 &\approx 1.0001 \\ \mu_2 &\approx 3.9427 \\ \mu_3 &\approx 4.0582 \\ \text{Bauer-Fike-Schranke} &\approx 0.1147 \\ \text{Schur-Schranke} &\approx (0.0422)^{1/3} \approx 0.3481 \end{aligned}$$

1.6 Die Kondition einzelner Eigenwerte

1.6.1. Lemma. *Die Eigenwerte von A hängen stetig von den Elementen von A ab.*

Beweisidee: Die Koeffizienten des charakteristischen Polynoms $p(z) = \det(zI - A)$ hängen stetig von den Matrixelementen ab. Die Nullstellen von p hängen stetig von den Koeffizienten von p ab. Details findet man z.B. in den Büchern von Kato, siehe [Ka82, S. 123].

1.6.2. Beispiel. Es sei $A = M_{01}(\epsilon) \in \mathbb{R}^{n \times n}$, die Matrix aus Beispiel 1.3.5. Dann gilt für jeden Eigenwert λ von A und n fest: $\lambda^n = \epsilon$ und $\lambda(\epsilon) = \epsilon^{\frac{1}{n}}$. Für $\epsilon \neq 0$ ist dies differenzierbar, und wir erhalten auf der reellen Achse

$$\frac{\mathrm{d}\lambda}{\mathrm{d}\epsilon} = \frac{1}{n}\epsilon^{\frac{1}{n}-1} \xrightarrow{\epsilon \to 0} \infty \quad \text{für} \quad n \geq 2.$$

Also scheint eine Konditionszahl von ∞ für einen mehrfachen Eigenwert angemessen. Wenn es Matrizen E gibt, $\|E\|$ klein und $A + E$ hat einen mehrfachen Eigenwert im Jordankästchen, so sollten die dadurch betroffenen Eigenwerte von A sehr große Konditionszahlen haben.

1.6.3. Definition. Sei $A \in \mathbb{C}^{n\times n}$ und für Vektoren $x, y \in \mathbb{C}^n$ mit $x^H x = y^H y = 1$ gelte $Ax = \lambda x$ und $y^H A = \lambda y^H$. Wir definieren

$$s(\lambda) := \left| y^H x \right| .$$

In Aufgabe 3 war zu zeigen:

- $s(\lambda)$ ist genau dann für einen Eigenwert λ von A eindeutig bestimmt, wenn A (bis auf Normierung) zu λ genau einen rechten und einen linken Eigenvektor hat.
- Wenn $X^{-1}AX = \operatorname{diag}(\lambda_1, \ldots, \lambda_n)$ mit $\lambda_i \neq \lambda_j$ für $i \neq j$ gilt und die Spalten $x_j = Xe_j$ von X sämtlich $|x_j^H x_j| = 1$ erfüllen, so ist $y_j^H := e_j^H X^{-1} \left(\left\| e_j^H X^{-1} \right\|_2 \right)^{-1}$ eindeutig bestimmt, $s(\lambda_j) = \left| y_j^H x_j \right| \neq 0,\ j = 1, \ldots, n$, und es gilt

$$\mathcal{K}_F(X)^2 = n \sum_{j=1}^{n} \left(\frac{1}{s(\lambda_j)} \right)^2 ,$$

mit $\mathcal{K}_F(X) = \|X\|_F \|X^{-1}\|_F$.

Im folgenden wird sich zeigen, dass $\mathcal{K}(\lambda) := \frac{1}{s(\lambda)}$ sich als Konditionszahl des Eigenwertes λ eignet. Dabei wird immer stillschweigend angenommen: Gibt es zum Eigenwert λ mehrere linear unabhängige Eigenvektoren, so ist $s(\lambda)$ stets für ein zusammengehöriges Paar x, y^H zu bilden.

1.6.4. Satz. *Es seien λ ein einfacher Eigenwert von A mit $Ax = \lambda x$, $y^H A = \lambda y^H$, $x^H x = y^H y = 1$ und μ der zu λ korrespondierende Eigenwert von $A + E$. Dann gilt*

$$\mu - \lambda = \frac{y^H E x}{y^H x} + \mathcal{O}(\|E\|^2)$$

und

$$|\mu - \lambda| \leq \frac{\|E\|}{|y^H x|} + \mathcal{O}(\|E\|^2).$$

Wir nennen $\mathcal{K}(\lambda) := \frac{1}{s(\lambda)} = \frac{1}{|y^H x|}$ die Konditionszahl des Eigenwertes λ.

Beweis. Es gilt

$$\begin{aligned} (A+E)(x+\delta x) &= \mu(x+\delta x) \\ Ax + Ex + A\delta x + E\delta x &= \mu x + \mu\delta x \\ Ex + A\delta x + E\delta x &= (\mu-\lambda)x + (\mu-\lambda)\delta x + \lambda\delta x \\ Ex + A\delta x &= (\mu-\lambda)x + \lambda\delta x, \end{aligned}$$

mit Vernachlässigung der sehr viel kleineren Terme 2. Ordnung, $E\delta x$ und $(\mu-\lambda)\delta x$. Dann gilt weiter

$$\begin{aligned} y^H E x + y^H A\delta x &= (\mu-\lambda) y^H x + \lambda y^H \delta x, \\ y^H E x &= (\mu-\lambda) y^H x. \end{aligned}$$

Unter der Annahme, dass δx und $\mu - \lambda$ stetig von E abhängen und also durch die Norm von E abgeschätzt werden können, gilt bei Wiederhinzunahme der Terme 2. Ordnung

$$\begin{aligned} \mu - \lambda &= \frac{y^H E x}{y^H x} + \mathcal{O}(\|E\|^2), \\ |\mu - \lambda| &\leq \frac{\|E\|}{|y^H x|} + \mathcal{O}(\|E\|^2) \end{aligned}$$

für jede Norm mit $\|Ax\| \leq \|A\| \, \|x\|$. □

1.6.5. Korollar. *Es sei $A \in \mathbb{C}^{n \times n}$ normal oder reell symmetrisch. Dann gilt*

$$|\mu - \lambda| \leq \|E\| + \mathcal{O}(\|E\|^2).$$

Beweis. Ist $A^H A = A A^H$, so existiert ein Q mit $Q^H Q = I$ und $Q^H A Q = D$. Q enthält die rechten und Q^H die linken Eigenvektoren von A mit $\left|y^H x\right| = 1$ für jeden Eigenwert.

Ist A reell symmetrisch, so ist D reell symmetrisch und Q kann reell orthogonal gewählt werden. Wieder gilt $\left|y^H x\right| = 1$ für jeden Eigenwert. □

1.6.6. Bemerkung. Diese Abschätzung ist scharf. Unter $|\lambda - \mu| \leq \|E\|_2$ ist nicht drunter zu kommen: ist z.B. A diagonal und $E = \operatorname{diag}(\epsilon, 0, \ldots, 0)$, gilt das Gleichheitszeichen.

Ist $A \in \mathbb{R}^{n \times n}$ ein Jordanblock, so ist $x = e_1$ und $y^H = e_n^H$. Es folgt $s(\lambda) = y^H x = 0$ für $n \geq 2$. Dies passt zur Überlegung für $\frac{\mathrm{d}\lambda}{\mathrm{d}\epsilon}$ mit $\epsilon \to 0$. Andererseits ist $s(\lambda) = 1$ für jeden Eigenwert jeder normalen Matrix, und das ist der maximale Wert. Insgesamt erhalten wir also:

$$1 \leq \mathcal{K}(\lambda) = \frac{1}{s(\lambda)} \leq \infty.$$

1.6.7. Satz (Bauer-Fike-Demmel). *Es sei $A \in \mathbb{C}^{n \times n}$ diagonalähnlich mit $X^{-1} A X = D = \operatorname{diag}(\lambda_1, \ldots, \lambda_n)$, $A x_i = \lambda_i x_i$, $y_i^H A = \lambda_i y_i^H$ und $\|x_i\|_2 = \left\|y_i^H\right\|_2 = 1$. Dann gilt*

$$\mu_i \in B\left(\lambda_i, n \frac{\|E\|_2}{|y_i^H x_i|}\right)$$

für die Eigenwerte μ_i von $A + E$.

Beweis. Das kann man mit Hilfe von Gerschgorinkreisen zeigen, siehe [Dem97, S. 150]. □

1.6.8. Beispiel. Fortführung von Beispiel 1.5.10. Es seien

$$A = \begin{pmatrix} 1 & 2 & 3 \\ 0 & 4 & 5 \\ 0 & 0 & 4.001 \end{pmatrix}$$

und

$$E = \begin{pmatrix} 0 & 0 & 0 \\ 0 & 0 & 0 \\ 10^{-3} & 0 & 0 \end{pmatrix}.$$

Dann gilt

$$\begin{aligned} \mu_1 &\approx 1.0001 \in B(1, 3 \cdot 0.0012) \\ \mu_2 &\approx 3.9427 \in B(4, 3 \cdot 6.0093) \\ \mu_3 &\approx 4.0582 \in B(4.001, 3 \cdot 6.0092) \end{aligned}$$

für $A + E$. Wendet man den Satz anders herum an, d.h. das ungestörte $A + E$ wird mit $-E$ gestört, so gilt

$$\begin{aligned} \lambda_1 &= 1 \in B(1.0001, 3 \cdot 0.0012) \\ \lambda_2 &= 4 \in B(3.9427, 3 \cdot 0.0521) \\ \lambda_3 &= 4.001 \in B(4.0582, 3 \cdot 0.0520). \end{aligned}$$

Es gibt also fast keinen Unterschied für λ_1 und μ_1. Für λ_2, λ_3, μ_2 und μ_3 gibt es jedoch große Unterschiede. Das ist nicht weiter verwunderlich, weil $A + E$ besser konditioniert ist als A, und somit bessere Schranken zu erwarten sind. Alle hier auftretenden Zahlen (außer den exakt bekannten Eigenwerten $\lambda_1, \lambda_2, \lambda_3$) wurden mit Matlab berechnet.

Ein Vergleich der zwei Sätze liefert:

$$\begin{aligned} \text{Satz 1.4.4} \quad &\Longrightarrow \quad \min_{\lambda} |\mu - \lambda| \leq \mathcal{K}_p(X) \|E\|_p, \ p \geq 1; \\ \text{Satz 1.6.7} \quad &\Longrightarrow \quad |\mu_j - \lambda_j| \leq n \|E\|_2 \frac{1}{s(\lambda_j)} = \mathcal{K}(\lambda_j)\, n \|E\|_2 . \end{aligned}$$

1.7 Die Rolle mehrfacher Eigenwerte

Wir haben gesehen, dass ein mehrfacher, defekter Eigenwert die Konditionszahl ∞ hat und das angemessen ist. Das heißt aber nicht, dass er nicht berechnet werden kann. Es gibt keinen Verlust von einer festen Anzahl an Stellen, sondern den Verlust eines festen Bruchteils an gültigen Stellen. Es sei ϵ das Maschinenepsilon, dh die größte in der Maschine darstellbare positive Zahl ϵ mit $1 + \epsilon = 1$. Bei der Matrix

$$\begin{pmatrix} 1 & 1 \\ \epsilon & 1 \end{pmatrix}$$

geht der Eigenwert 1 in $1 \pm \sqrt{\epsilon}$ über, was einem Verlust von ungefähr der Hälfte der Stellen entspricht. Bei

$$M_{11}(\epsilon) = \begin{pmatrix} 1 & 1 & & & \\ & 1 & 1 & & \\ & & \ddots & \ddots & \\ & & & 1 & 1 \\ \epsilon & & & & 1 \end{pmatrix}_{n \times n}$$

hat man einen n-fachen Eigenwert und nur noch etwa den n-ten Teil an gültigen Stellen.

Geometrische Eigenschaft der Konditionszahl bei Matrixinversion: Die Konditionszahl $\mathcal{K}(A) = \|A\| \|A^{-1}\|$ ist ein Maß für den Abstand von A zu einer singulären Matrix, d.h. zu einer Matrix mit $\mathcal{K}(A+E) = \infty$. Genauer:

1.7.1. Satz. *Es sei $A \in \mathbb{C}^{n\times n}$ regulär. Dann gilt*

$$\min\left\{\frac{\|E\|_2}{\|A\|_2} : A+E \text{ ist singulär}\right\} = \frac{1}{\mathcal{K}_2(A)}.$$

Beweis. Siehe [Dem97, S. 33f]. □

Ähnliches gilt bei Eigenwerten:

1.7.2. Satz. *Es sei λ ein einfacher Eigenwert von A, mit $\|x\|_2 = \|y\|_2 = 1$ und $\mathcal{K}(\lambda) = \frac{1}{|y^H x|}$. Dann gibt es ein $E \in \mathbb{C}^{n\times n}$, so dass $A+E$ einen mehrfachen Eigenwert hat, und es ist*

$$\frac{\|E\|_2}{\|A\|_2} \leq \frac{1}{\sqrt{\mathcal{K}(\lambda)^2 - 1}}.$$

Ist $\mathcal{K}(\lambda) \gg 1$, d.h. ist λ sehr schlecht konditioniert, so ist

$$\frac{1}{\sqrt{\mathcal{K}(\lambda)^2 - 1}} \approx \frac{1}{\mathcal{K}(\lambda)}.$$

Beweis. Ohne Einschränkung ist A in Schur-Form. Nach Satz 1.5.4 existiert ein $Q \in \mathbb{C}^{n\times n}$ mit $Q^H Q = I$, $Q^H A Q = T = D + N$, $d_1 = \lambda$. Sind x, y Eigenvektoren von A zu λ, so sind $Q^H x$ und $Q^H y$ Eigenvektoren von T zu λ und es gilt $(Q^H y)^H (Q^H x) = y^H Q^H Q x = y^H x$. Es gibt also keine Konditionsverschlechterung bei Übergang zur Schur-Form. Also sei

$$A = \begin{pmatrix} \lambda & A_{12} \\ 0 & A_{22} \end{pmatrix},$$

$x = e_1$, $\tilde{y} = (1,\, A_{12}(\lambda I - A_{22})^{-1})^H$ und $y = \frac{\tilde{y}}{\|\tilde{y}\|_2}$. Dann gilt

$$\mathcal{K}(\lambda) = \frac{1}{|y^H x|} = \frac{\|\tilde{y}\|_2}{\|\tilde{y}^H x\|_2} = \|\tilde{y}\|_2 = \sqrt{1 + \left\|A_{12}(\lambda I - A_{22})^{-1}\right\|_2^2}$$

und weiter

$$\sqrt{\mathcal{K}(\lambda)^2 - 1} = \left\|A_{12}(\lambda I - A_{22})^{-1}\right\|_2 \leq \|A_{12}\|_2 \left\|(\lambda I - A_{22})^{-1}\right\|_2 \leq \frac{\|A_{12}\|_2}{\sigma_{\min}(\lambda I - A_{22})}.$$

Aus der Definition von $\sigma_{\min}$ folgt: Es existiert ein E mit $\|E\|_2 = \sigma_{\min}(\lambda I - A_{22})$ und $A_{22} + E - \lambda I$ ist singulär. Also ist λ ist Eigenwert von $A_{22} + E$ und damit hat

$$\tilde{A} = \begin{pmatrix} \lambda & A_{12} \\ 0 & A_{22} + E \end{pmatrix}$$

einen doppelten Eigenwert λ mit $\|E\|_2 = \sigma_{\min}(\lambda I - A_{22}) \leq \frac{\|A\|_2}{\sqrt{\mathcal{K}(\lambda)^2 - 1}}$. □

1.7.3. Satz. *Es sei A diagonalisierbar mit Eigenwerten λ_i und Eigenvektoren x_i und y_i^H für die gilt $\|x_i\|_2 = \|y_i\|_2 = 1$. Es sei $X^{-1}AX = D = \operatorname{diag}(\lambda_1, \ldots, \lambda_n)$. Dann gilt*

$$\max_i \frac{1}{|y_i^H x_i|} \leq \|X\|_2 \left\|X^{-1}\right\|_2 .$$

Gilt $X = (x_1, \cdots, x_n)$, so gilt

$$\|X\|_2 \left\|X^{-1}\right\|_2 \leq n \max_i \frac{1}{|y_i^H x_i|}.$$

Beweis. Demmel, Paper von 1983 (Buch S. 153). □

1.7.4. Bemerkung. Außerdem gilt:

$$\|X\|_2 \leq \|X\|_F \leq \sqrt{n}\, \|X\|_2$$

und

$$\|X\|_F^2 \left\|X^{-1}\right\|_F^2 = n \sum_{j=1}^{n} \left(\frac{1}{|y_i^H x_i|} \right)^2 .$$

Die Kondition bezüglich der Eigenwertberechnung und die Kondition bezüglich der Inversion und dem Lösen von Gleichungssystemen hängen eng zusammen.

1.8 Singulärwerte

Für $A \in \mathbb{C}^{n\times n}$, $Ax = \lambda x$, $y^H A = \lambda y^H$ sind $\operatorname{lin}\{x\}$, $\operatorname{lin}\{y^H\}$ invariant unter A. Für $B \in \mathbb{C}^{m\times n}$, $m > n$, gibt es keine invarianten Unterräume, aber es existieren $u, v \neq 0$ mit $Av = \sigma u$, $A^H u = \sigma v$ und $\sigma \geq 0$. σ ist Singulärwert und u, v sind Singulärvektoren von A.

1.8.1. Satz. *Es sei $A \in \mathbb{C}^{m\times n}$, $m \geq n$. Dann gibt es ein $U \in \mathbb{C}^{m\times n}$ mit $U^H U = I$ und ein $V \in \mathbb{C}^{n\times n}$ mit $V^H V = I$, so dass gilt*

$$A = USV^H$$

mit $S = \operatorname{diag}(\sigma_1, \ldots, \sigma_n)$ und $\sigma_1 \geq \sigma_2 \geq \cdots \geq \sigma_n \geq 0$.

Die Spalten $u_1, \ldots, u_n$ von U heißen linke Singulärvektoren. *Die Spalten $v_1, \ldots, v_n$ von V heißen* rechte Singulärvektoren. *Die σ_i heißen Singulärwerte. Das ist die schlanke Singulärwertzerlegung (SVD, Singular Value Decomposition).*

Beweis. Demmel, S. 110. □

1.8.2. Satz. *Es sei $A = USV^H$ die Singulärwertzerlegung von $A \in \mathbb{C}^{m\times n}$, $m \geq n$.*

1. *Falls $A = A^H = QDQ^H$, $QQ^H = I$, so gilt $\sigma_i = |\lambda_i|$, $u_i = q_i$ und $v_i = \operatorname{sign}(\lambda_i) u_i$ mit* $\operatorname{sign}(0) := 1$.

2. *$A^H A \in \mathbb{C}^{n\times n}$ hat die Eigenwerte σ_i^2: es gilt $\lambda(A^H A) = \{\sigma_i^2\}, i = 1, \ldots, n$.*

3. *Für $AA^H \in \mathbb{C}^{m\times m}$ gilt $\lambda(AA^H) = \{\sigma_1^2, \ldots, \sigma_n^2, 0, \ldots, 0\}$.*

4. *Für $A \in \mathrm{GL}(\mathbb{C}, n)$ und $\|A\|_2 = \sigma_1$ gilt $\|A^{-1}\|_2 = \frac{1}{\sigma_n}$ und $\mathcal{K}_2(A) = \frac{\sigma_1}{\sigma_n} = \frac{\sigma_{\max}}{\sigma_{\min}}$.*

5. *Aus $\sigma_1 \geq \sigma_2 \geq \cdots \geq \sigma_r > \sigma_{r+1} = 0 = \cdots = 0$ folgt $\operatorname{rank} A = r$, $\operatorname{kern} A = \operatorname{lin}\{v_{r+1}, \ldots, v_n\}$ und $\operatorname{range} A = \operatorname{lin}\{u_1, \ldots, u_r\}$.*

6. *Es seien $V = (v_1, \ldots, v_n)$, $U = (u_1, \ldots, u_n)$ und $A = USV^H = \sum_{i=1}^n \sigma_i u_i v_i^H$. Dann ist die unter der Norm $\|\cdot\|_2$ dichteste Matrix M mit $\operatorname{rank} M = k < n$ gegeben durch $A_k := \sum_{i=1}^k \sigma_i u_i v_i^H$ und es ist $\|A - A_k\|_2 = \sigma_{k+1}$. Es gilt $A_k = US_kV^H$ mit $S_k = \operatorname{diag}(\sigma_1, \ldots, \sigma_k, 0, \ldots, 0)$.*

Beweis. Demmel, S. 111ff. □

2 Pseudospektren

2.1 Definitionen

Die Abfrage $f(x_0) - a = 0$ ist numerisch ungeschickt, besser ist $|f(x_0) - a| < \epsilon$. Nur, wie groß sollte ϵ gewählt werden? Das ist problem- und maschinenabhängig. Wie sieht das bei Eigenwerten aus? In Anwendungen werden Matrizen und ihre Eigenwerte auf vielfältige Weisen gestört. Wie wir gesehen haben, gibt es Abschätzungen für die Störungen der Eigenwerte bei bekannter Größe der Störungen der Matrixelemente spätestens seit Einführung der Gerschgorinkreise, 1931. Erst kürzlich ist daraus aber die Konsequenz gezogen worden, schon den Begriff 'Spektrum' in 'ϵ-Pseudospektrum' zu verallgemeinern. Es wird quasi

$$|\det(\lambda I - A)| < \epsilon \quad \text{statt} \quad \det(\lambda I - A) = 0$$

betrachtet, jedoch in Formulierungen, die für die Anwendung auf gegebene Matrizen bequemer zu handhaben sind.

2.1.1. Definition. Es seien $A \in \mathbb{C}^{n\times n}$ und $\epsilon \geq 0$. Die Menge

$$\lambda_\epsilon(A) := \left\{ z \in \mathbb{C} : z \in \lambda(A + E) \text{ für ein } E \in \mathbb{C}^{n\times n} \text{ mit } \|E\|_2 \leq \epsilon \right\}$$

heißt *ϵ-Pseudospektrum von A* in der Norm $\|\cdot\|_2$.

2.1.2. Bemerkung. Aus $\epsilon_1 < \epsilon_2$ folgt $\lambda_{\epsilon_1}(A) \subset \lambda_{\epsilon_2}(A)$. Es gilt $\lambda_{\epsilon=0}(A) = \lambda(A) \subset \lambda_\epsilon(A)$ für alle $\epsilon > 0$. Also sind Pseudospektren eine Verallgemeinerung von Spektren.

Ist diese Definition sinnvoll gemacht, d.h. hat sie die nötige Invarianz? Es gibt mehrere Arten, Spektren zu definieren. Führen die verschiedenen Definitionen in allen Fällen jeweils auf dasselbe ϵ-Pseudospektrum, d.h. bleibt die Äquivalenz bei der Verallgemeinerung immer erhalten? Zunächst sehen wir uns mehrere Definitionen 'ϵ-Pseudospektrum' an. Die Äquivalenz wird dann Satz 2.1.9 liefern.

2.1.3. Definition. Es seien $A \in \mathbb{C}^{n\times n}$ und $\epsilon \geq 0$. Definiere

$$\lambda_\epsilon(A) := \left\{ z \in \mathbb{C} : \left\|(zI - A)^{-1}\right\|_2 \geq \frac{1}{\epsilon} \right\}.$$

2.1.4. Bemerkung. Für $\epsilon = 0$ gilt

$$\begin{aligned} \lambda_0(A) &= \left\{ z \in \mathbb{C} : \left\|(zI - A)^{-1}\right\|_2 > \frac{1}{\eta} \text{ für jedes } \eta > 0 \right\} \\ &= \{ z \in \mathbb{C} : (zI - A) \text{ ist singulär } \} = \lambda(A). \end{aligned}$$

$R(z) = (zI - A)^{-1}$ ist die Resolvente des linearen Operators A in z (Funktionalanalysis). Definition 2.1.3 ist auch noch in unendlich dimensionalen Banach- und Hilberträumen eine brauchbare Definition.

2.1.5. Definition. Es seien $A \in \mathbb{C}^{n\times n}$ und $\epsilon \geq 0$. Definiere

$$\lambda_\epsilon(A) := \{z \in \mathbb{C} : \|(zI - A)x\|_2 \leq \epsilon \text{ für ein } x \in \mathbb{C}^n \text{ mit } \|x\|_2 = 1\}.$$

Der Wert z heißt ϵ-*Pseudoeigenwert von* A und der Vektor x heißt ϵ-*Pseudoeigenvektor von* A in der Norm $\|\cdot\|_2$.

2.1.6. Bemerkung. Im Prinzip könnten in diesen drei Definitionen auch andere miteinander verträgliche Vektor- und Matrix-Normen genommen werden als das Paar {euklidische Norm, Spektralnorm}. Momentan ist aber kein Vorteil so einer Verallgemeinerung sichtbar. Wir beschränken uns deshalb hier auf die Betrachtung mit diesem Normen-Paar.

2.1.7. Bemerkung. Pseudospektren sind zwar Norm-abhängig, liefern dafür aber auch mehr Information als Spektren, da sie zusätzlich noch etwas über die Störanfälligkeit des Spektrums sagen, die in dieser Norm gemessen wird.

2.1.8. Definition. Es seien $A \in \mathbb{C}^{n\times n}$ und $\epsilon \geq 0$. Definiere

$$\lambda_\epsilon(A) := \{z \in \mathbb{C} : \sigma_{\min}(zI - A) \leq \epsilon\}$$

wobei $\sigma_{\min}$ der kleinste Singulärwert von $zI - A$ ist.

2.1.9. Satz. *Für Matrizen $A \in \mathbb{C}^{n\times n}$ sind die in den vier Definitionen 2.1.1, 2.1.3, 2.1.5 und 2.1.8 definierten Mengen gleich, also sind die vier Definitionen äquivalent.*

Beweis. Ist $z \in \lambda(A)$ oder $\epsilon = 0$, so gilt die Behauptung. Es sei also nun jeweils $z \in \lambda_\epsilon(A)\backslash\lambda(A)$ und $\epsilon > 0$. Damit existiert $(zI - A)^{-1}$.

2.1.1 $\Longrightarrow$ 2.1.5 : Es sei nun $z \in \lambda_\epsilon(A)$ nach 2.1.1. Dann existiert ein E mit $\|E\|_2 \leq \epsilon$, so dass $z \in \lambda(A + E)$. Also existiert ein $x \in \mathbb{C}^n$ mit $\|x\|_2 = 1$, so dass $(A + E)x = zx$. Also $(zI - A)x = Ex$ und $\|Ex\|_2 \leq \|E\|_2 \leq \epsilon$. Also $z \in \lambda_\epsilon(A)$ nach 2.1.5.

2.1.5 $\Longrightarrow$ 2.1.3 : Es sei nun $z \in \lambda_\epsilon(A)$ nach 2.1.5. Dann existiert ein x mit $\|x\|_2 = 1$ und $\|(zI - A)x\|_2 \leq \epsilon$. Es gibt $y \in \mathbb{C}^n$ mit $\|y\|_2 = 1$ und $0 < \sigma \leq \epsilon$, so dass $(zI-A)x = \sigma y$ gilt. Dann folgt $(zI - A)^{-1} y = \sigma^{-1}x$, also $\left\|(zI - A)^{-1}\right\|_2 = \sigma^{-1} \geq \epsilon^{-1}$.

2.1.3 $\Longrightarrow$ 2.1.1 : Es sei nun $z \in \lambda_\epsilon(A)$ nach 2.1.3. Also ist $\left\|(zI - A)^{-1}\right\|_2 \geq \epsilon^{-1}$. Dann existieren $x, y \in \mathbb{C}^n$ mit $(zI - A)^{-1} y = \sigma^{-1}x$, $\|y\|_2 = \|x\|_2 = 1$ und $\sigma^{-1} \geq \epsilon^{-1}$. Also $(zI - A)x = \sigma y$. Wir brauchen ein $E \in \mathbb{C}^{n\times n}$ mit $\|E\|_2 \leq \epsilon$, $(A+E)x = zx$ und $\|x\|_2 = 1$. Definiere $E := \sigma y w^H$ mit $w \in \mathbb{C}^n$ und $w^H w = 1$. Bei der Norm $\|\cdot\|_2$ wähle $w = x$. Damit folgt $\|E\|_2 = |\sigma| \leq \epsilon$ und $(A+E)x = (A+\sigma y x^H)x = Ax + \sigma y = zIx$. Also $z \in \lambda(A+E)$.

2.1.8 $\Longleftrightarrow$ 2.1.3 : Es sei nun $z \in \lambda_\epsilon(A)$ nach 2.1.3. Dann gilt $\left\|(zI - A)^{-1}\right\|_2 \geq \epsilon^{-1}$. Es gilt $\left\|(zI - A)^{-1}\right\|_2 = (\sigma_{min}(zI - A))^{-1} \geq \epsilon^{-1}$ genau dann, wenn $\sigma_{\min}(zI - A) \leq \epsilon$.

Also sind alle vier Definitionen äquivalent und damit ist der Satz bewiesen. □

2.2 Beispiele (Aufgabe 6)

Zunächst grundsätzliche Beobachtungen: Bei reellen, reell gestörten Matrizen liegt mit λ auch $\bar{\lambda}$ in $\lambda(A+E)$, also ist das Pseudospektrum symmetrisch zur reellen Achse. Bei imaginären, imaginär gestörten Matrizen liegt mit λ auch $-\bar{\lambda}$ in $\lambda(A+E)$, also ist das Pseudospektrum symmetrisch zur imaginären Achse. Bei reellen, beliebig komplex gestörten Matrizen braucht das ϵ-Pseudospektrum für $\epsilon \neq 0$ keine Symmetrie zu haben. Geht die Störung $||E|| \leq \epsilon$ gegen 0, so nimmt der Durchmesser von $\lambda_\epsilon(A)$ ab. Es gilt

$$\lim_{\epsilon \to 0} \lambda_\epsilon(A) = \lambda(A).$$

Bei Matrix (2), `a = diag(one(4,1),1)`, also der 5×5-Matrix $a = M_{01}(0)$, fallen im reellen Pseudospektrum die radialen Strahlen mit zehner-Symmetrie auf: diese entsprechen im wesentlichen den Eigenwerten von $a+E = M_{01}(\mu)$ und $a-E = M_{01}(-\mu)$, $0 < \mu \leq \epsilon$, die sich um einen Faktor i voneinander unterscheiden.

Auch die Kurven im rellen Pseudospektrum von Matrix (5) und somit Matrix (6) lassen sich sehr schnell identifizieren durch isolierte Betrachtung der dominierenden positiven und negativen Störungen, d.h. durch Ersetzung der drei Nullen durch μ und/oder $-\mu$.

Wir wollen noch Satz 1.5.8 anwenden. Es seien

$$A = \begin{pmatrix} 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 0 \end{pmatrix} = D + N \quad \text{und} \quad E \quad \text{zufallsgeneriert}, \quad \|E\|_2 \leq \epsilon.$$

Es gilt $N^p = 0$ mit $p \geq 4$ und $\|N\|_2^k = 1$. Setze

$$\theta = \|E\|_2 \sum_{k=0}^{p-1} \|N\|_2^k = 4\epsilon,$$

dann gilt nach Satz 1.5.8

$$\min_{\lambda \in \lambda(A)} |\lambda - \mu| \leq \max\left\{\theta, \sqrt[p]{\theta}\right\}.$$

Da $\lambda(A) = \{0\}$, können wir aus den Schur-Schranken für die einzelnen Eigenwerte in diesem Fall eine Schranke für das gesamte Pseudospektrum $\lambda_\epsilon(A)$ bekommen. Das gibt beispielsweise

ϵ	$=$	$1 \cdot 10^{-12}$	$100 \cdot 10^{-12}$	$10^4 \cdot 10^{-12}$
$\sqrt[p]{\theta}$	$\approx$	$1.42 \cdot 10^{-3}$	$4.5 \cdot 10^{-3}$	$14.2 \cdot 10^{-3}$

Die Zahlen wurden jeweils nach oben gerundet. Die 4.Wurzel wächst langsamer als die 2.Wurzel und die n-te Wurzel für $n \geq 5$ noch langsamer. Die Bauer-Fike-Schranke kann man nicht anwenden, da A nicht diagonalisierbar ist.

2.3 Eigenschaften von Pseudospektren

Wie können wir Pseudospektren weiter charakterisieren?

2.3.1. Lemma. *Pseudospektren sind invariant unter unitären Ähnlichkeitstransformationen.*

Beweis. Es sei $Q \in \mathbb{C}^{n\times n}$ mit $Q^HQ = I$. Dann gilt

$$\left(zI - QAQ^H\right)^{-1} = \left(Q(zI - A)Q^H\right)^{-1} = Q\left(zI - A\right)^{-1}Q^H$$

und damit

$$\left\|\left(zI - QAQ^H\right)^{-1}\right\|_2 = \left\|\left(zI - A\right)^{-1}\right\|_2.$$

Also $\lambda_\epsilon(QAQ^H) = \lambda_\epsilon(A)$. □

2.3.2. Bemerkung. Ist A normal mit $QAQ^H = D$, so ist das ϵ-Pseudospektrum von A gegeben durch die ϵ-Gerschgorinkreise um die Eigenwerte von A. Es gilt $\lambda_\epsilon(A) = \bigcup_{i=1}^n B(\lambda_i, \rho_\epsilon)$ mit $0 \le \rho_\epsilon \le \epsilon$.

2.3.3. Satz. *Es sei $A \in \mathbb{C}^{n\times n}$. Dann gilt*

$$\lambda(A) + \delta_\epsilon \subset \lambda_\epsilon(A), \qquad \epsilon \ge 0 \text{ beliebig},$$

mit $\delta_\epsilon := \{z \in \mathbb{C} : |z| \le \epsilon\}$ und $\lambda(A) + \delta_\epsilon := \{z \in \mathbb{C} : z = z_1 + z_2,\ z_1 \in \lambda(A),\ z_2 \in \delta_\epsilon\}$. Wenn A normal ist, gilt $\lambda(A) + \delta_\epsilon = \lambda_\epsilon(A)$.

Beweis. Es sei $\lambda \in \lambda(A)$. Dann gilt $\lambda + z \in \lambda(zI + A)$ mit $z \in \delta_\epsilon$. Setze $E := zI$. Es gilt $\|E\| = \|zI\| = |z| \le \epsilon$ und damit $\lambda + z \in \lambda_\epsilon(A)$.

Ist A normal, so gilt $\lambda_\epsilon(A) = \lambda_\epsilon(QAQ^H)$ mit $Q^HQ = I$. Ohne Einschränkung sei nun $A = \operatorname{diag}(\lambda_1, \ldots, \lambda_n)$. Für $\epsilon = 0$ oder $z \in \lambda(A)$ ist die Aussage trivial. Es sei also $\epsilon > 0$ und $z \in \lambda_\epsilon(A)\backslash\lambda(A)$. Die Resolvente von A in z ist diagonal, also gilt

$$\left\|\left(zI - A\right)^{-1}\right\|_2 = \frac{1}{\min_{\lambda\in\lambda(A)}|z - \lambda|} = \frac{1}{\operatorname{dist}(z, \lambda(A))}.$$

Weiter gilt $\left\|\left(zI - A\right)^{-1}\right\|_2 \ge \epsilon^{-1}$ und damit $\operatorname{dist}(z, \lambda(A)) \le \epsilon$. Also $z \in \lambda(A) + \delta_\epsilon$. □

Aufgabe: Wenn $\lambda(A) + \delta_\epsilon = \lambda_\epsilon(A)$ gilt, folgt dann, dass A normal ist?

2.3.4. Bemerkung. Es sei

$$\mathcal{N}_\epsilon(M) := \{z \in \mathbb{C} : \operatorname{dist}(z, M) < \epsilon\}$$

die ϵ-Umgebung der Menge M in $\mathbb{C}$. Aus Satz 2.3.3 folgt $\overline{\mathcal{N}_\epsilon(\lambda(A))} \subset \lambda_\epsilon(A)$. Das ϵ-Pseudospektrum von A enthält also für alle $\epsilon > 0$ eine ϵ-Umgebung des Spektrums von A. Für normale Matrizen ist das ϵ-Pseudospektrum gleich dem Abschluß der ϵ-Umgebung.

2.3.5. Satz (Bauer-Fike). *Es sei $A \in \mathbb{C}^{n\times n}$ mit $X^{-1}AX = D = \operatorname{diag}(\lambda_1, \ldots, \lambda_n)$ für ein $X \in \mathbb{C}^{n\times n}$. Dann gilt*

$$\lambda(A) + \delta_\epsilon \subset \lambda_\epsilon(A) \subset \lambda(A) + \delta_{\epsilon\mathcal{K}(X)}.$$

Beweis. Das folgt aus Satz 2.3.3 und Satz 1.4.4. □

2.3.6. Satz. *Es seien $A \in \mathbb{C}^{n\times n}$, $z \in \mathbb{C}$ und $\epsilon \geq 0$. Pseudospektren haben dann folgende Eigenschaften (1. mit 5. wie Spektren):*

1. *$\lambda_\epsilon(A)$ ist nicht leer, kompakt und hat höchstens n Zusammenhangskomponenten.*
2. *$\lambda_\epsilon(A^H) = (\lambda_\epsilon(A))^H = \{\bar{z} \in \mathbb{C} : z \in \lambda_\epsilon(A)\}$.*
3. *$\lambda_\epsilon(A + zI) = z + \lambda_\epsilon(A)$.*
4. *$\lambda_{\epsilon|z|}(zA) = z\lambda_\epsilon(A)$.*
5. *Seien $A_i \in \mathbb{C}^{n_i\times n_i}$, $i = 1, 2$. Definiere $A_1 \oplus A_2 := \begin{pmatrix} A_1 & 0 \\ 0 & A_2 \end{pmatrix}$.*
   Dann gilt $\lambda_\epsilon(A_1) \cup \lambda_\epsilon(A_2) = \lambda_\epsilon(A_1 \oplus A_2)$.
6. *Es sei $A_{11} \in \mathbb{C}^{m\times m}$ mit $m < n$ und es gebe A_{12}, A_{21} und A_{22}, so dass*

$$A = \begin{pmatrix} A_{11} & A_{12} \\ A_{21} & A_{22} \end{pmatrix}.$$

   Dann gilt

$$\lambda_\epsilon(A) \subset \lambda_{\tau(\epsilon)}(A_{11}) \cup \lambda_{\tau(\epsilon)}(A_{22})$$

   mit

$$\tau(\epsilon) = (\epsilon + \|A_{21}\|_2)\sqrt{1 + \frac{\|A_{12}\|_2}{\epsilon + \|A_{21}\|_2}}.$$

Beweis.

1. Es ist $\emptyset \neq \lambda(A) \subset \lambda_\epsilon(A)$.
   $\lambda_\epsilon(A)$ ist beschränkt: Für $\mu \in \lambda_\epsilon(A)$, $\epsilon \geq 0$, existieren $E \in \mathbb{C}^{n\times n}, \|E\|_2 \leq \epsilon$ und $x \in \mathbb{C}^n, \|x\|_2 = 1$, mit $(A + E)x = \mu x$ und

$$|\mu| = \|\mu x\|_2 = \|Ax + Ex\|_2 \leq \|A\|_2 \|x\|_2 + \|E\|_2 \|x\|_2 \leq \|A\|_2 + \|E\|_2 \leq \|A\|_2 + \epsilon.$$

   $\lambda_\epsilon(A)$ ist abgeschlossen: Nach Lemma 1.6.1 hängen die Eigenwerte von A stetig von den Elementen von A ab, also auch die gestörten Eigenwerte von den Elementen der Störungen. Die Menge der Störungen,

$$\{E \in \mathbb{C}^{n\times n} : \|E\|_2 \leq \epsilon\},$$

ist beschränkt und abgeschlossen. $\lambda_\epsilon(A)$ ist also stetiges Bild einer kompakten Menge und daher abgeschlossen.
$\lambda_\epsilon(A)$ hat höchstens n Zusammenhangskomponenten: $\mu \in \lambda(A+E) \subset \lambda_\epsilon(A)$ hängt stetig von den Elementen von E ab und $E = 0$ ist eine mögliche Störung. Also gibt es einen stetigen Weg von einem der Eigenwerte von A zu μ. Da A n Eigenwerte hat, gibt es höchstens n Zusammenhangskomponenten von $\lambda_\epsilon(A)$.

2. Das folgt aus $\|A\|_2 = \left\|A^H\right\|_2$ für alle $A \in \mathbb{C}^{n\times n}$.

3. Es gilt
$$(A + zI + E)x = \mu x \iff (A+E)x = (\mu - z)x$$
und
$$\mu - z \in \lambda_\epsilon(A) \iff \mu \in z + \lambda_\epsilon(A).$$

4. Für $z = 0$ ist die Aussage trivial. Es sei also $z \neq 0$. Gilt $\mu \in \lambda_{\epsilon|z|}(zA)$, so existiert ein $E \in \mathbb{C}^{n\times n}$ mit $\|E\| \leq \epsilon\,|z|$ und $\mu \in \lambda(zA+E)$. Dann gibt es ein $x \in \mathbb{C}^n$ mit $\|x\| = 1$ und $(zA+E)x = \mu x$. Also gilt $(A + \frac{1}{z}E)x = \frac{\mu}{z}x$ und $\left\|\frac{1}{z}E\right\| \leq \epsilon$. Daraus folgt $\frac{\mu}{z} \in \lambda_\epsilon(A)$. Die Umkehrung folgt analog.

5. Es seien $\mu \in \lambda_\epsilon(A_1) \cup \lambda_\epsilon(A_2)$ und ohne Einschränkung $\mu \in \lambda_\epsilon(A_1)$. Also existieren ein $E \in \mathbb{C}^{n_1\times n_1}$ mit $\|E\| \leq \epsilon$ und ein $x \in \mathbb{C}^{n_1}$ mit $\|x\| = 1$, so dass $(A_1+E)x = \mu x$ gilt. Setze
$$E_\oplus := \begin{pmatrix} E & 0 \\ 0 & 0 \end{pmatrix} \quad \text{und} \quad x_\oplus := \begin{pmatrix} x \\ 0 \end{pmatrix}.$$
Damit gilt $(A_1 \oplus A_2 + E_\oplus)x_\oplus = \mu x_\oplus$, $\|x_\oplus\| = 1$ und $\|E_\oplus\| \leq \epsilon$. Also $\mu \in \lambda_\epsilon(A_1 \oplus A_2)$.
Die Umkehrung $\lambda_\epsilon(A_1 \oplus A_2) \subset \lambda_\epsilon(A_1) \cup \lambda_\epsilon(A_2)$ folgt aus Punkt 6. mit $A_{12} = A_{21} = 0$ und $\tau(\epsilon) = \epsilon$.

6. Zum Beweis siehe den Beweis von Theorem 3.2 in der Arbeit von Grammont und Largillier [GL02].

□

Besondere Beachtung verdient auch der Spezialfall $A_{21} = 0$ von Aussage 6 und sein Vergleich mit Lemma 1.5.1.

2.4 Berechnung von Pseudospektren

Grundsätzlich kann jede Definition von $\lambda_\epsilon(A)$ zur Berechnung von Pseudospektren benutzt werden. Dabei kann jeder Algorithmus zur Berechnung von Eigenwerten verwendet werden. Effizienz?

In der Praxis sind bisher vor allem drei Familien von Algorithmen erprobt worden [gate]: Bestimmung der Spektren zufällig gestörter Matrizen, Gitteralgorithmen und Fortsetzungsverfahren.

Programm von Demmel (Aufgabe 6). Es gibt einen schnellen Überblick über die Empfindlichkeit einer Matrix gegen Störungen, aber es ist nur für kleine bis mittlere Matrizen praktikabel. Die Überlagerung mit Zufallsmatrizen $E \in \mathbb{C}^{n\times n}$ mit $\|E\| \leq \epsilon$ und die Berechnung von $\lambda(A+E)$ ist gut für die Abschätzung des Einflusses von Rundungsfehlern, wenn man davon ausgeht, dass Rundungsfehler zufallsverteilt sind. Diskretisierungsfehler bei Differentialgleichungen und Fehler bei mathematischer Modellbildung sind von anderer Art. Diese sind problemabhängig, systematisch und anwendungsspezifisch. Deshalb ist eine Überlagerung mit Zufallsmatrizen zur Abschätzung solcher Fehler nicht besonders geeignet, aber besser als gar nichts.

Gitteralgorithmen. Das ist eine grosse Gruppe von Verfahren, die heute als Standard gelten. Prinzip: Wähle einen Bereich $B \subset \mathbb{C}$ der komplexen Ebene. Lege ein Gitter G über B und rechne $\sigma_{\min}(zI - A)$ für jeden Gitterpunkt $z \in G$ aus. Dann Höhenlinien plotten.

- Auswahl des Gebietes B:
  - Vorauswahl mit Gerschgorinkreisen.
  - Punktewolken à la Aufgabe 6.
  - Zunächst sehr grobes Gitter benutzen.
  - Theoretische Kennntnisse aus Anwendungsproblemen.
- Grobes Gitter systematisch dort verfeinern wo es nötig ist, wenn ein Überblick vorhanden ist.
- $\sigma_{\min}(zI-A)$ für alle Gitterpunkte berechnen. Definition 2.1.8 scheint die beste Wahl zu sein, weil da untiäre Transformationen benutzt werden. Der Aufwand ist dabei $\mathcal{O}(n^3)$ für jeden Gitterpunkt, es ist also eine Heidenarbeit, wenn man es naiv angeht.

In den letzten 5 Jahren gab es einige Fortschritte zur Effizienzsteigerung (Stand 2004). Einige Prinzipien zur Effizienzsteigerung:

- Nur auf einem Teil von A arbeiten, wenn nur einige Eigenwerte von A von Interesse sind. Besonders bei grossen Matrizen sind häufig nur sehr wenige Eigenwerte von Interesse. Man braucht einen invarianten Unterraum $U \in \mathbb{C}^{n\times k}$ mit $k \ll n$ und $u_i^H u_j = \delta_{ij}$, so dass AU deutlich kleiner als A ist, aber die interessanten Eigenwerte enthält. Die Arbeit U zu finden ist meistens vernachlässigbar gegenüber anschließend gewonnener Ersparnis. Ist U perfekt invariant, so ist das Spektrum davon unbeeinflußt. Das Pseudospektrum wird mehr oder weniger davon berührt. Der Fehler hängt dabei stark von der Kondition der beteiligten Eigenwerte ab, insbesondere der weggelassenen. Ist U nur näherungsweise invariant, so werden das Spektrum und Pseudospektrum deutlicher beeinflußt. Einige Abschätzungen des durch das Abspalten gemachten Fehlers findet man in [GL02]. Das Abspalten von Unterräumen ist auch bei nicht-linearen Problemen üblich und erfordert da dann besonders viel Geschick.

- Vor Berechnung der Singulärwerte $\sigma_{\min}(zI - A)$ die Matrix A ein einziges mal transformieren: Hessenbergform, vollständige oder partielle Schur-Form.
- Die beiden vorherigen Punkte kombinieren: Gewünschte Eigenwerte gemäß Transformation oben in die Matrix bringen und dahinter abschneiden.
- Zur Berechnung von

  $$\sigma_{\min}(zI - A) = \sigma_{\min}(zI - Q^H AQ) = \sigma_{\min}(zI - T)$$

  inverse Iteration benutzen, eventuell unter Ausnutzung der Schur-Form T von A. Es sei $x_0 \in \mathbb{C}^n$ gegeben, dann sind alle weiteren Vektoren gegeben durch

  $$\begin{aligned} (zI - T)^H x_{k+\frac{1}{2}} &= x_k \\ (zI - T)\tilde{x}_{k+1} &= x_{k+\frac{1}{2}} \\ x_{k+1} &= \frac{\tilde{x}_{k+1}}{\|\tilde{x}_{k+1}\|}. \end{aligned}$$

  Es konvergiert $\|\tilde{x}_k\|^{-1}$ gegen $\sigma_{\min}^2$. Die Transformation auf Schur-Form kostet $\mathcal{O}(n^3)$ und die Berechnung der Singulärwerte liegt im Allgemeinen bei $\mathcal{O}(n^3)$. Die inverse Iteration ist bei Ausnutzung der Dreiecksform von T mit $\mathcal{O}(n^2)$ billiger und lohnt sich, wenn man viele Gitterpunkte hat. Man muss vorher allerdings T berechnen.
- Bei wirklich großen Matrizen (z.B. $n = 10^5$) wäre eine Schur-Zerlegung zu viel Aufwand: Anoldi-Iteration und Krylov-Unterraummethoden für das Auffinden von (fast) invarianten Unterräumen sind da effizenter (Y. Saad). Anoldi um 1950: Punkt 9.4 Golub, van Loan
- Bei Berechnung von Spektren: Implicitly restarted Anoldi. Entsprechend Anwenden für Pseudospektren.

Wie gross ist der Fehler bei der Approximation des Pseudospektrums? Das ist nicht wirklich bekannt und es gibt noch viele offene Fragen.

Software:

- `ARPACK`: Fortran-Code zur Berechnung von (einigen) Eigenwerten von grossen Matrizen [LSY98]; in MATLAB verfügbar unter dem Namen EIGS ('Some EIGenvalues');
- Matlab: `eigs` und `svds` für große Matrizen;
- `Eigtool`: Programmpaket (Werkzeug) für die komfortable Berechnung und grafische Darstellung von (Teilen von) Pseudospektren im Rahmen von Matlab. Basiert auf Matlab-EIGS und also auf ARPACK; ist Update des älteren `Pseudospectra GUI`; Autor: Tom Wright, Oxford. Zugang und weitere Informationen über den Pseudospectra Gateway [gate].
- weitere Informationen im Internet: Pseudospectra Gateway [gate].

Fortsetzungsverfahren. Die Fortsetzungsverfahren basieren auf der Idee, dass die Berechnung im Gitterpunkt z effizienter ist, wenn man das Ergebnis im Nachbarpunkt schon kennt und damit die Iteration startet [Lu97]. Sie sind in Kombination mit der inversen Iteration leider häufig nicht sehr erfolgreich und auch noch mit anderen Problemen behaftet (siehe auch: A Brief Survey of Computational Methods, Dec. 2004 [gate]).

2.5 Eigenwerte und Pseudospektren von rechteckigen Matrizen

Für $A \in \mathbb{C}^{m\times n}$ mit $m \geq n$ kann es streng genommen keinen invarianten Unterraum geben, weil für $A : \mathbb{C}^n \mapsto \mathbb{C}^m$ Urbild- und Bildraum nicht übereinstimmen. Man kann aber verallgemeineren:

$$Ax = \lambda \tilde{I} x \quad \text{mit} \quad \tilde{I} = \begin{pmatrix} I \\ 0 \end{pmatrix} \in \mathbb{C}^{m\times n}.$$

Aus

$$\begin{pmatrix} A_1 \\ A_2 \end{pmatrix} x = \lambda \begin{pmatrix} I \\ 0 \end{pmatrix} x$$

folgt $A_1 x = \lambda x$ und $A_2 x = 0$. Ist $A_2 x \neq 0$, so ist x kein Eigenvektor und λ kein Eigenwert. Viele rechteckige Matrizen haben keine Eigenwerte. Und wenn doch, dann hängen die Eigenwerte nicht stetig von den Elementen von A ab. Aber: $A_2 x \neq 0$, $\|A_2 x\| = \epsilon_0 \implies \lambda \in \lambda_\epsilon(A)$ für $\epsilon \geq \epsilon_0$. Möglich: $\lambda_\epsilon(A) = \emptyset$ für $\epsilon < \epsilon_0$. Die Elemente von $\lambda_\epsilon(A)$ hängen für $\epsilon > 0$ stetig von den Elementen von A ab, wenn Pseudospektren für rechteckige Matrizen so definiert werden, wie es die Definition über die Singulärwerte nahelegt. Genauer gilt:

2.5.1. Definition. Für $B = (b_1, \ldots, b_n) \in \mathbb{C}^{m\times n}$ ist $B^+ \in \mathbb{C}^{n\times m}$ die *Moore-Penrose Inverse*, wenn gilt:

1. $y \in \operatorname{lin}\{b_1, \ldots, b_n\} \implies x = B^+ y$ löst $Bx = y$,
2. $BB^+B = B$,
3. $B^+BB^+ = B^+$,
4. $(BB^+)^H = BB^+$,
5. $(B^+B)^H = B^+B$.

Existiert B^+ nicht, setze $\|B^+\| := \infty$.

2.5.2. Satz. *Seien $A \in \mathbb{C}^{m\times n}$ mit $m \geq n$, $\tilde{I} = \begin{pmatrix} I \\ 0 \end{pmatrix} \in \mathbb{C}^{m\times n}$ und $\epsilon \geq 0$. Dann sind die folgenden Definitionen äquivalent:*

- $\lambda_\epsilon(A) := \left\{ z \in \mathbb{C} : \left\| \left(z\tilde{I} - A \right) x \right\|_2 \leq \epsilon \text{ für ein } x \in \mathbb{C}^n \text{ mit } \|x\|_2 = 1 \right\}$,

- $\lambda_\epsilon(A) := \{z \in \mathbb{C} : z \in \lambda(A+E) \textit{ für ein } E \textit{ mit } \|E\|_2 \leq \epsilon\}$,
- $\lambda_\epsilon(A) := \left\{ z \in \mathbb{C} : \left\| \left(z\tilde{I} - A \right)^+ \right\|_2 \geq \epsilon^{-1} \right\}$,
- $\lambda_\epsilon(A) := \left\{ z \in \mathbb{C} : \sigma_{\min}(z\tilde{I} - A) \leq \epsilon \right\}$.

Im Fall $m = n$ wird die Moore-Penrose Inverse zur regulären Inversen, und die Definitionen in diesem Satz fallen mit den entsprechenden Definitionen für quadratische Matrizen zusammen.

Beweis. Der Beweis kann in etwa so geführt werden wie der Beweis von Satz 2.1.9, vgl [WT01b]. □

2.5.3. Beispiel. Die Matrix

$$A = \begin{pmatrix} 1 & 10 & 10 \\ 0 & 2.1 & 4.2 \\ 0 & 0.1 & 0.2 \\ 0 & 0.1 & 0.2 \end{pmatrix}$$

hat $x_1 = (10, -2, 1)^T$ als Eigenvektor zum Eigenwert $\lambda_1 = 0$ und $x_2 = (1, 0, 0)^T$ als Eigenvektor zum Eigenwert $\lambda_2 = 1$.

2.5.4. Beispiel. Die Matrix

$$B = \begin{pmatrix} 1 & 0 & 0 \\ 0 & 0 & 1 \\ 0 & -1 & 0 \\ 0.1 & 0.2 & 0.2 \end{pmatrix}$$

hat keine Eigenwerte und -vektoren. Es gibt aber den ϵ-Pseudoeigenwert 1 für $\epsilon \geq 0.1$ und die ϵ-Pseudoeigenwerte $\pm i$ für $\epsilon \geq 2.9 * 10^{-2}$ mit ϵ-Pseudoeigenvektoren $v_\pm = (0, 0.1, \pm 0.1i)^T$.

3 Anwendungen

3.1 Stabilität in dynamischen Systemen

3.1.1. Definition. Es sei $D \subset \mathbb{R}^n$ offen, $n \geq 1$, und $f : D \mapsto \mathbb{R}^n$ eine glatte Abbildung (zweimal stetig differenzierbar). Die Bewegungsgleichung

$$\dot{u} = f(u) \quad \text{mit} \quad u(0) = u_0 \tag{3}$$

definiert auf D ein *dynamisches System*, wenn sie für jedes $u_0 \in D$ genau eine Lösung hat, die für alle $t \in [0, \infty)$ existiert und in D bleibt. Die Lösung wird mit $u(t; u_0)$ bezeichnet.

3.1.2. Bemerkung. Dieses System ist

- *autonom*, weil f nicht von t abhängt,
- *deterministisch*, weil es immer genau eine Lösung hat und
- *kontinuierlich*, weil es durch eine Differentialgleichung beschrieben wird.

3.1.3. Definition. Die Menge

$$\gamma(u_0) := \{u(t; u_0) \in D : t \in [0, \infty)\}$$

heißt *Trajektorie (Bahn, Orbit)* des Systems durch u_0. Die Abbildung

$$\Phi_t : D \mapsto \mathbb{R}^n, \; u_0 \mapsto u(t; u_0)$$

heißt *Fluß des Systems.*

3.1.4. Beispiel. Gegeben seien $A = \begin{pmatrix} a & c \\ 0 & b \end{pmatrix} \in \mathbb{R}^{2\times 2}$ und ein Anfangswert $u_0 \in \mathbb{R}^2$. Gesucht ist $u(t; u_0)$ mit

$$\dot{u} = Au, \qquad \begin{pmatrix} \dot{u_1} \\ \dot{u_2} \end{pmatrix} = \begin{pmatrix} a & c \\ 0 & b \end{pmatrix} \begin{pmatrix} u_1 \\ u_2 \end{pmatrix}, \quad \text{und} \quad u(0) = u_0.$$

Aus der Theorie linearer Systeme folgt:
$c \neq 0$ und $a \neq b$:

$$u_1(t) = \alpha e^{at} + \beta e^{bt}$$
$$u_2(t) = \frac{b-a}{c} \beta e^{bt}$$

$c \neq 0$ und $a = b$:

$$u_1(t) = (\alpha + \beta t) e^{at}$$
$$u_2(t) = \frac{\beta}{c} e^{at}$$

$c = 0$:

$$u_1(t) = \alpha e^{at}$$
$$u_2(t) = \beta e^{bt},$$

wobei die Koeffizienten α und β jeweils aus der Anfangsbedingung $u(0) = u_0$ zu bestimmen sind.

Wie verhält sich $u(t; u_0)$ für $t \to \infty$?
Für $\operatorname{Re} a > 0$ und $\alpha \neq 0$ gilt

$$u_1(t; u_0) \to \infty \quad \text{und} \quad \|u(t; u_0)\| \to \infty.$$

Für $\operatorname{Re} a < 0$ und $\operatorname{Re} b < 0$ gilt

$$u(t; u_0) \to 0.$$

Was gilt für $c \neq 0$ und $\operatorname{Re} a = \operatorname{Re} b < 0$? Falls u_0 und damit α und β geeignet gewählt sind, wächst die Lösung zunächst an und fällt erst dann ab. Auch im Fall $\operatorname{Re} a \neq \operatorname{Re} b < 0$ und $c \neq 0$ kann es solche Lösungen geben, vgl. Aufgabe 11.

Angenommen ein Experiment ist stochastisch gestört und die Störung wirft das Experiment immer wieder auf eine Anfangsbedingung zurück, die ein Ansteigen der Lösung verlangt. Gilt dann immer noch $u(t; u_0) \to 0$? In der Fluidmechanik (nicht-linear) sind Beispiele bekannt, in denen ein gestörtes Experiment sich ganz anders verhält als ungestört.

In der Theorie der dynamischen Systeme ist es wichtig herauszufinden, wie sich die Lösungen eines gegebenen Systems für alle Anfangswerte im gegebenen Definitionsbereich D und für alle Zeiten verhalten. Es gibt zwei Arten der graphischen Darstellung: Phasendiagramm und Zeitdiagramm. In Abbildung 2 ist die Lösung von

$$\dot{u}_1 = u_2, \qquad \dot{u}_2 = 1 + (u_1 - 2)u_2 - u_1^2, \tag{4}$$

mit Anfangswert $u_0 = (0, 4)^T$ auf beide Arten dargestellt.

Der einfachste Fall zeitlichen Verhaltens ist $\lim_{t\to\infty} u(t; u_0) = \bar{u}$, $|\bar{u}| < \infty$, für alle $u_0 \in D$. Ein anderes Beispiel ist $u(t; 0) = \sin t$. Das motiviert folgende Definition:

3.1.5. Definition. Die Menge

$$\omega(u_0) := \{v \in \mathbb{R}^n : u(t_k; u_0) \to v \text{ für eine Folge } (t_k) \text{ mit } t_k \to \infty \text{ für } k \to \infty\}$$

heißt *ω-Limesmenge* des dynamischen Systems *zum Anfangswert* u_0.

Die Limesmenge von $\lim_{t\to\infty} u(t; u_0) = \bar{u} < \infty$ ist $\omega(u_0) = \{\bar{u}\}$ und die von $u(t; 0) = \sin t$ ist $\omega(0) = [-1, 1]$.

3.1.6. Definition. Gilt $f(\bar{u}) = 0$ für ein $\bar{u} \in D$, so heißt $\bar{u}$ *Gleichgewichtspunkt* oder *stationäres Gleichgewicht* des dynamisches Systems.

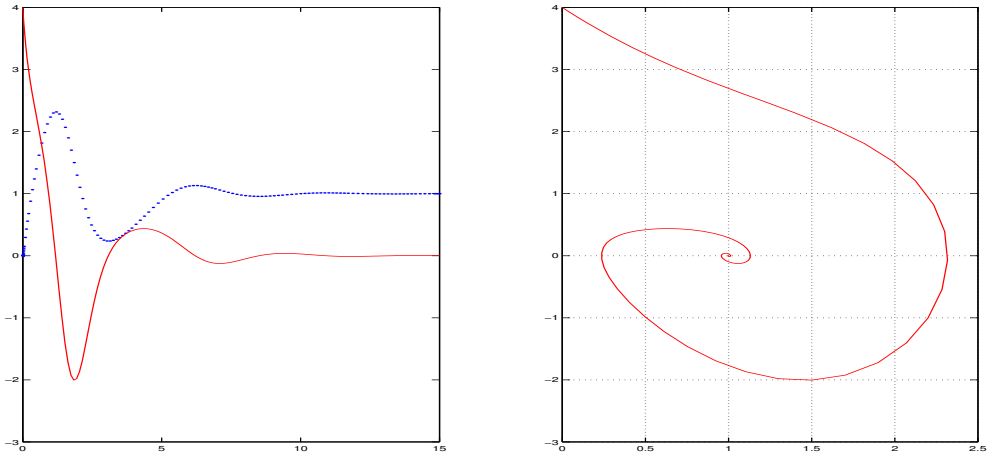


Abbildung 2: Lösung von Gleichung (4) mit Anfangswert $u_0 = (0,4)^T$. Links $u_i(t;u_0)$, $i = 1,2$, gegen t, $0 \le t \le 15$ (Zeitdiagramm), rechts $u_2(t;u_0)$ gegen $u_1(t;u_0)$, mit t als Bahn-Parameter (Phasendiagramm).

Für $\dot{u} = f(u) = Au$ ist $\bar{u} = 0$ ein Gleichgewichtspunkt. Ist die Matrix A regulär, so ist $\bar{u} = 0$ der einzige Gleichgewichtspunkt. Interessanter sind nicht-lineare f. Wichtige Fragen sind dann: Wie verhält sich das System in der Nähe von Gleichgewichtspunkten? Wie stark können benachbarte Bahnen auseinanderlaufen? Wie stark können Störungen eine Trajektorie verändern?

Für den folgenden Satz betrachten wir vorübergehend nicht-autonome f, weil dieser Satz in der Numerik eine wichtige Rolle spielt.

3.1.7. Satz. *Es seien $G \subset \mathbb{R}^{n+1}$ offen und $f : \bar{G} \mapsto \mathbb{R}^n$ stetig, beschränkt und f erfülle in G eine Lipschitzbedingung*

$$\|f(u,t) - f(v,t)\| \le L \|u - v\|$$

mit $L > 0$. Es seien $u_1, u_2 : (a,b) \mapsto G$ zwei stückweise stetig differenzierbare Näherungslösungen von

$$\dot{u} = f(t,u) \quad \textit{mit} \quad u(t_0) = u_0,$$

für die

$$\|u_i(t_0) - u_0\| \le \delta_i$$

und

$$\|\dot{u}_i(t) - f(t,u_i)\| \le \epsilon_i, \qquad i = 1,2,$$

für alle $t \in (a,b)$ gilt. Dann gilt auf (a,b)

$$\|u_1(t) - u_2(t)\| \le (\delta_1 + \delta_2) e^{L|t-t_0|} + \frac{\epsilon_1 + \epsilon_2}{L}(e^{L|t-t_0|} - 1).$$

Beweis. Variationen dieses Satzes und deren Beweise findet man in sehr vielen Büchern, siehe z.B. [HNW93, Theorem 10.2, p. 58] oder [Num3, Satz 2.2 und Lemma 2.3 ff]. □

Der Satz gibt Abschätzungen bei

- Störungen in den Anfangswerten (stetige Abhängigkeit von Anfangswerten),
- numerischen Fehlern bei Anwendung von Differenzenverfahren,
- Abhängigkeit von f von Parametern,
- nur näherungsweise bekanntem f (schwierige mathematische Modellierung) und
- stochastisch gestörtem f.

Ein Beispiel für das Auseinanderlaufen von Trajektorien sind chaotische Lösungen. Beim Werfen einer Münze z.B. gilt in guter Näherung die Newtonsche Mechanik. Es gibt also zu jeder Anfangsbedingung genau eine Lösung. Zwei exakt gleiche Würfe bewirken also exakt dasselbe Ergebnis. Doch hängen die Lösungen in vielen Fällen so empfindlich von der Anfangsbedingung ab, dass kleine Abweichungen große Folgen haben und es keinem Menschen gelingt, mehrfach den gleichen Wurf zu machen.

Trajektorien können auch zusammenlaufen ($t \mapsto -t$), das ist im Satz nicht berücksichtigt. Beispiel: $\dot{u} = \lambda u$, $\lambda > 0 \implies \lambda = L$, $\lambda < 0 \implies -\lambda = L$.

3.1.8. Definition. Die Menge $M \subset D$ heißt *invariante Menge* des dynamischen Systems (3), wenn

$$u(t; u_0) \in M \text{ für alle } u_0 \in M \text{ und alle } t \in [0, \infty) \text{ gilt .}$$

Wenn $\bar{u}$ ein Gleichgewichtspunkt ist, dann gilt $u(t; \bar{u}) = \bar{u}$ für alle $t \in [0, \infty)$. Also ist $M = \omega(\bar{u}) = \{\bar{u}\}$ eine invariante Menge.

3.1.9. Definition. Die kompakte invariante Menge $M \subset \mathbb{R}^n$ heißt

- *stabil*, wenn es zu jeder Umgebung U von M eine Umgebung V von M gibt, so dass $u(t; u_0) \in U$ für alle $u_0 \in V$ und alle $t \in [0, \infty)$ gilt;
- *asymptotisch stabil*, wenn M stabil ist und es zusätzlich eine Umgebung $\tilde{V}$ von M gibt, so dass $\operatorname{dist}(u(t; u_0), M) \to 0$ für $t \to \infty$ und für alle $u_0 \in \tilde{V}$;
- *instabil*, wenn sie nicht stabil ist.

Beispiele: Massenpunkte im Potentialgebirge; Strömungen zwischen rotierenden Zylindern [Dy, Me99]; Gleichgewichte des in Aufgabe 12 zu untersuchenden Systems.

Wir betrachten noch einmal $\dot{u} = Au$, $A \in \mathbb{R}^{n \times n}$ regulär. Der Gleichgewichtspunkt $\bar{u} = 0$ ist asymptotisch stabil, falls $\operatorname{Re} \lambda_i < 0$ für alle $i \in \{1, \ldots, n\}$ und instabil, falls $\operatorname{Re} \lambda_i > 0$ für ein $i \in \{1, \ldots, n\}$. Diese Beobachtung soll nun verallgemeinert werden.

Wir betrachten nun parameterabhängige Systeme. Es seien $D \subset \mathbb{R}^n$ offen und $f : D \times \mathbb{R} \mapsto \mathbb{R}^n$ eine glatte Abbildung und

$$\dot{u} = f(u, \mu) \quad \text{mit} \quad u(0) = u_0 \in D. \tag{5}$$

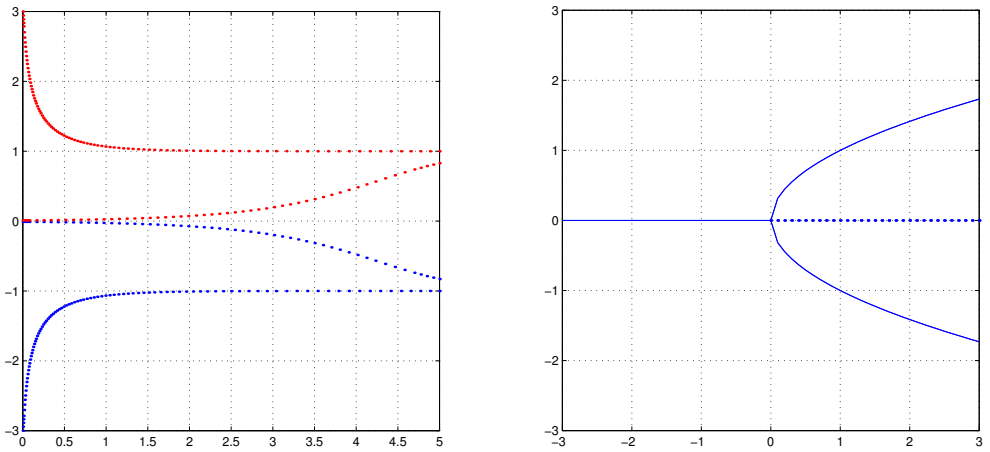


Abbildung 3: (Beispiel 3.1.12) Links: Lösung von $\dot{u} = u(\mu - u^2)$ für $\mu = 1$ und Anfangswerte $u_0 = -3, -0.01, 0.01, 3$; rechts: Gleichgewichte $\bar{u}(\mu)$ gegen μ für $\mu \in [-3, 3]$.

Angenommen, für jedes feste $\mu \in \mathbb{R}$ ist dadurch ein dynamisches System auf D definiert, d.h. für jedes $u_0 \in D$, μ fest, existiert genau eine Lösung $u(t; u_0, \mu)$, die für alle $t \in [0, \infty)$ existiert und in D bleibt.

3.1.10. Satz (Lyapunov 1892). *Sei $\bar{u} \in \mathbb{R}^n$ ein Gleichgewichtspunkt von $\dot{u} = f(u, \mu_0)$, d.h. es gelte $f(\bar{u}, \mu_0) = 0$. Dann gilt:*

1. *$\bar{u}$ ist asymptotisch stabil, wenn $\operatorname{Re}\lambda < 0$ für alle Eigenwerte λ von $D_u f(\bar{u}, \mu_0)$ gilt.*
2. *$\bar{u}$ ist instabil, wenn $\operatorname{Re}\lambda > 0$ für einen Eigenwert λ von $D_u f(\bar{u}, \mu_0)$ gilt.*

Hierbei bezeichnet $D_u f(\bar{u}, \mu_0)$ die Jacobimatrix von f in $\bar{u}$ bei festem Parameterwert μ_0.

Beweis. siehe z.B. [StH96, Theorem 2.3.5, p.130]. □

Stabilität (Definition 3.1.9) ist eine Eigenschaft der Lösung $\bar{u}$ von $\dot{u} = f(u, \mu_0)$, aber es reicht *in vielen Fällen* aus, das lineare Problem $\dot{v} = D_u f(\bar{u}, \mu_0)v$ zu untersuchen. Bei gewöhnlichen Systemen gilt ein noch stärkerer Satz, der *Satz von Hartman-Grobman*. Allgemein (bei PDEs und Systemen von PDEs und in der Physik) ist es üblich, von der Gültigkeit des *Prinzip der linearisierten Stabilität* zu sprechen.

3.1.11. Beispiel. $\dot{u} = u(1-u) = f(u)$: $\bar{u}_1 = 0$, $\bar{u}_2 = 1$.
$f'(u) = 1 - 2u$, $f'(0) = 1 > 0 \implies \bar{u}_1 = 0$ ist instabil; $f'(1) = -1 < 0 \implies \bar{u}_2 = 1$ ist stabil. Das Intervall $M = [0, 1]$ ist invariante Menge: $u_0 \in M \implies u(t; u_0) \in M$. M ist instabil: $u_0 < 0 \implies u(t; u_0) \to -\infty$. Es gibt also keine Umgebung von M, s.d. Trajektorien in der Nähe bleiben.

3.1.12. Beispiel. $\dot{u} = u(\mu - u^2) = f(u, \mu)$: $\bar{u}_1 = 0$ ist Gleichgewicht für alle $\mu \in \mathbb{R}$; $\bar{u}_{2,3} = \pm\sqrt{\mu}$ sind Gleichgewichte für $\mu \geq 0$;

$D_u f(u,\mu) = \mu - 3u^2$, $D_u f(0;\mu) = \mu$, $D_u f(\pm\sqrt{\mu};\mu) = \mu - 3\mu$,
$\bar{u}_1 = 0$ ist stabil für $\mu < 0$, instabil für $\mu > 0$;
$\bar{u}_{2,3} = \pm\sqrt{\mu}$ existieren reell für $\mu > 0$ und sind dann stabil. $\bar{u}_1$ und $\bar{u}_{2,3}$ tauschen für $\mu = 0$ ihre Stabilität aus. Es gilt hier das *Prinzip vom Austausch der Stabilität.*

Beim Taylorproblem (Strömungen zwischen rotierenden Zylindern) [Dy, Me99]: statt Polynom stationäre Navier-Stokes-Gleichungen (μ ist hier die Reynoldszahl, alle anderen Parameter halten wir fest). Beim Übergang der Couette-Strömung in Taylorwirbel wechselt der Realteil eines Eigenwertes der Jacobimatrix das Vorzeichen. Der dadurch erzeugte Austausch von Stabilität wurde zunächst im Experiment beobachtet, später auch mathematisch bewiesen.

Es gibt viele andere Arten, wie Gleichgewichte ihre Stabilität verlieren können.

Geschichtlich: Vor Lyapunov: Stabilitätsüberlegungen waren qualitativ, physikalisch heuristisch. Kritische Parameterwerte waren praktisch nur im Experiment feststellbar, nicht vorhersagend ausrechenbar.

Lyapunov hat 1892 eine mathematische Theorie zur Stabilität in der Mechanik von Massenpunkten entwickelt. Die Anwendung dieser Theorie auf Planetenbewegungen in der Himmelsmechanik war sehr erfolgreich (Poincaré). Die ersten Anwendungen auf Probleme der Strömungsmechanik waren Flops.

Navier-Stokes-Gleichungen (ca. 1850): Es seien $u(x,t) : \mathbb{R}^3 \times \mathbb{R} \mapsto \mathbb{R}^3$ die makroskopische Geschwindigkeit, $p(x,t) : \mathbb{R}^3 \times \mathbb{R} \mapsto \mathbb{R}$ der hydrostatische Druck, $\rho(x,t) : \mathbb{R}^3 \times \mathbb{R} \mapsto \mathbb{R}$ die Dichte des Fluids und ν die kinematische Viskosität

$$\begin{aligned} \frac{\partial u}{\partial t} + (u \cdot \nabla)u &= -\nabla p + \nu \nabla^2 u \\ \frac{\partial \rho}{\partial t} + \nabla \cdot (\rho u) &= 0 \end{aligned}$$

Spezialfälle:
1. $\nu = 0$ für ideale Gase, Eulersche Gleichungen (Euler 1741). Physikalisch betrachtet, ist $\nu = 0$ manchmal eine gute Näherung. Bei Raumtemperatur: $\nu \approx 10^{-2}\frac{\mathrm{cm}^2}{\mathrm{s}}$ für Luft, aber $\nu \approx 15 \cdot 10^{-2}\frac{\mathrm{cm}^2}{\mathrm{s}}$ für Wasser. Mathemathisch betrachtet, ist es ein ungeheurer Eingriff, $\nu = 0$ zu setzen: von den höchsten Ableitungen hängt ab, welche Rand- und Anfangsbedingungen gesetzt werden sollten. Bei deren Streichung bekommen die Gleichungen mathematisch einen ganz anderen Charakter.

Auf ähnliche Weise wurde in der Plasmaphysik zunächst 'ideale Magnetohydrodynamik' (elektrische Leitfähigkeit κ unendlich gut, $1/\kappa = 0$) betrachtet und erst später 'resistive Magnetohydrodynamik' ($1/\kappa \neq 0$). Unterschiede im Stabilitätsverhalten von Plasmen für $1/\kappa \neq 0$ und für $1/\kappa = 0$ konnten mit Hilfe von Pseudospektren verstanden werden [BRK94].
2. $\nu \neq 0$, $\rho = \mathrm{const}$. In diesem Fall reduziert sich die zweite Gleichung auf $\mathrm{div}\, u = 0$. Um eine von konkreten Abmessungen unabhängige *dimensionslose* Darstellung zu bekommen, wird die Reynoldszahl R eingeführt: $R = L * V/\nu$, L charakteristische Länge,

V charakteristische Geschwindigkeit. Stationäre Gleichgewichte sind durch $\partial u/\partial t = 0$ gekennzeichnet. Für mehrere Konfigurationen gibt es Gleichgewichte, die auch bzgl ihrer räumlichen Abhängigkeit sehr einfach sind und deshalb seit langem als Lösungen der Navier-Stokes-Gleichungen explizit bekannt. Beispiele:
Ebene Couetteströmung. Zwei unendlich ausgedehnte Platten bewegen sich mit konstanter Geschwindigkeit in entgegengesetzter Richtung. Die Flüssigkeit dazwischen hat ein lineares Geschwindigkeitsprofil $\bar{u}(x)$, das leicht explizit angebbar ist. Der Ansatz $u(x,t) = \bar{u}(x) + U(x,t)$ in den Navier-Stokes-Gleichungen und Vernachlässigung der nicht-linearen Terme führt auf ein lineares System für die Störung $U(x,t)$ (Physiker-Art, die Jacobimatrix des Systems zu bestimmen; ist praktisch für die Handhabung großer Systeme). Stabilitätsanalyse für dieses System unter Annahme der Gültigkeit des Prinzips der linearisierten Stabilität: die ebene Couette-Strömung verhält sich wie $\bar{u} = 0$ in Aufgaben 11 und 12. Es ist $\mathrm{Re}\lambda(\bar{u}) < 0$ für alle Eigenwerte, aber das ϵ-Pseudospektrum der Jacobimatrix reicht für ziemlich kleine ϵ schon in die rechte Halbebene: es reichen sehr kleine Störungen des Experiments, um die Grundströmung zu destabilisieren.
Rohr-Poiseuille-Strömung Strömung mit parabolischem Geschwindigkeitsprofil in einem Zylinder. Auch hier ist die Grundströmung explizit bekannt und gemäß dem Prinzip der linearisierten Stabilität für diejenigen Parameterwerte stabil, für die im Experiment Instabilität beobachtet wurde. Die im Experiment beobachtete kritische Reynoldszahl ließ sich also nicht durch Verallgemeinerung der Lyapunov'schen Stabilitätstheorie herleiten.

An diesen beiden Problemen haben sich Orr, Sommerfeld, von Mises, Hopf, Harrison und andere versucht, ohne Erfolg. Darauf Orr (1907): 'Es wird niemals gelingen, das Instabilwerden von Strömungen mathematisch zu modellieren!' Weniger als 20 Jahre später gelang G.I. Taylor der Durchbruch [Tay]: er untersuchte stattdessen die Stabilität der
Couette-Strömung zwischen rotierenden Zylindern. Diese Grundströmung ist ebenfalls explizit bekannt. Für die Couette-Strömung gelten aber sowohl das Prinzip der linearisierten Stabilität als auch das Prinzip vom Austausch der Stabilität. Spätere Untersuchungen haben gezeigt, daß der Navier-Stokes-Operator nicht selbst-adjungiert ist und ϵ-Pseudospektren schon für kleine ϵ stark von den Spektren abweichen können [TRD93]. Noch 1972 hat Yih vermeintlich bewiesen, dass alle Eigenwerte der Jacobimatrix für die Couette-Strömung für gewisse Parameterwerte reell sind. Erst 1984 wurde numerisch gezeigt, dass das falsch ist [DR, Me99].

4 Übungsaufgaben

19. Okt. 2004

Spektren und Pseudo-Spektren – Theorie, Numerik und Anwendungen (Meyer-Spasche)

Übungsblatt 1

Aufg. 1: Es gelte

$$AX = XB, \qquad A \in \mathbb{C}^{n\times n},\ B \in \mathbb{C}^{k\times k},\ X \in \mathbb{C}^{n\times k},\ k \le n.$$

Zeigen Sie:
a) Die Bildmenge

$$ran(X) := \{y \in \mathbb{C}^n :\ \text{ex } z \in \mathbb{C}^k \text{ mit } y = Xz\}$$

ist invariant unter A.
b) Wenn X vollen Spaltenrang hat, dann gilt $\lambda(B) \subset \lambda(A)$, wobei

$$\lambda(C) := \{\lambda \in \mathbb{C} :\ \det(\lambda I - C) = 0\}$$

das Spektrum der Matrix C bezeichnet.
c) Falls $n = k$ und X nicht-singulär ist, dann gilt $\lambda(B) = \lambda(A)$.

26. Okt. 2004

Spektren und Pseudo-Spektren – Theorie, Numerik und Anwendungen (Meyer-Spasche)

Übungsblatt 2

Aufg. 2: Seien $M_{01} := (m_{jk}) \in \mathbb{C}^{n\times n}$ und $E_{j_0,k_0}(\varepsilon) := (e_{j,k}(\varepsilon)) \in \mathbb{C}^{n\times n}$ mit

$$m_{jk} := \begin{cases} 1 & fallsk = j+1, \quad j = 1,\ldots,n-1, \\ 0 & sonst, \end{cases}$$

und

$$e_{jk}(\varepsilon) := \begin{cases} \varepsilon & falls(j,k) = (j_0,k_0), \\ 0 & sonst. \end{cases}$$

a) Berechnen und plotten Sie mit Matlab das Spektrum von $M_{01} + E_{j_0,k_0}(\varepsilon)$ und vergleichen Sie es mit dem Spektrum von M_{01} für $n = 10$, $(j_0,k_0) = (n,1)$ und $\varepsilon = 1.e^{-10}$.
b) Verändern Sie n, ε und (j_0,k_0) und beobachten Sie, wie sich die Abweichung ändert. Für welche Indizes (j_0,k_0) ist die Abweichung bei festem ε und n *minimal*, für welchen Index *maximal*? Beweis?
c) Ermitteln Sie die Gerschgorin-Kreise von $M_{01} + E_{j_0,k_0}(\varepsilon)$. Was können Sie aus ihrer Kenntnis schließen?

2. Nov. 04

Spektren und Pseudo-Spektren – Theorie, Numerik und Anwendungen (Meyer-Spasche)

Übungsblatt 3

Aufg. 3: Sei $A \in \mathbb{C}^{n\times n}$ und für Vektoren $x, y \in \mathbb{C}^n$ mit $x^H x = y^H y = 1$ gelte $Ax = \lambda x$ und $y^H A = \lambda y^H$. Wir definieren

$$s(\lambda) := |y^H x|.$$

a) Geben Sie für $n = 2$ eine hinreichende Bedingung dafür, daß $s(\lambda)$ für jeden Eigenwert von A eindeutig bestimmt ist. Zeigen Sie, daß $s(\lambda)$ nicht eindeutig bestimmt ist, wenn diese Bedingung verletzt ist.
b) Zeigen Sie: Wenn $X^{-1}AX = diag(\lambda_1, \ldots, \lambda_n)$ mit $\lambda_i \neq \lambda_j$ für $i \neq j$ gilt und die Spalten x_j von X sämtlich $|x_j^H x_j| = 1$ erfüllen, so ist $s(\lambda_j) \neq 0,\ j = 1, \ldots, n$, und es gilt

$$\kappa_F(X)^2 = n \sum_{j=1}^{n} \left(\frac{1}{s(\lambda_j)} \right)^2,$$

mit $\kappa_F(X) := ||X||_F ||X^{-1}||_F$ und $||C||_F := \left(\sum_{i=1}^{m} \sum_{j=1}^{n} |c_{ij}|^2 \right)^{1/2}$ für $C \in \mathbb{C}^{m\times n}$.
c) Betrachten Sie speziell die Matrix $B(\varepsilon_1, \varepsilon_2) = A(\varepsilon_1) + E_{j_0,k_0}(\varepsilon_2)$ für $\varepsilon_1 = \varepsilon_2 = 10^{-3}$, $(j_0, k_0) = (3, 1)$ und $E_{j_0,k_0}(\varepsilon_2)$ wie in Aufg.2, und

$$A = \begin{pmatrix} 1 & 2 & 3 \\ 0 & 4 & 5 \\ 0 & 0 & 4+\varepsilon_1 \end{pmatrix}.$$

Bestimmen Sie mit Hilfe von Matlab die Eigenwerte λ_j von $B(\varepsilon_1, \varepsilon_2)$, $s(\lambda_j)$ und $||E_{j_0,k_0}(\varepsilon_2)||_2 / s(\lambda_j)$ für $j = 1, 2, 3$.

9. Nov. 04

Spektren und Pseudo-Spektren – Theorie, Numerik und Anwendungen (Meyer-Spasche)

Übungsblatt 4

Aufg. 4: Lesen Sie die folgenden Abschnitte und denken Sie darüber nach:
a) J.H. Wilkinson, *Rundungsfehler*, p. 177f: 'Ein Beispiel mit schlecht konditionierten Eigenwerten' ;
b) C. Moler, *Numerical Computing with MATLAB*, Kapitel 10, Abschnitt 10.6: Eigenvalue Sensitivity and Accuracy, p. 14f

16. Nov. 04

Spektren und Pseudo-Spektren – Theorie, Numerik und Anwendungen (Meyer-Spasche)

Übungsblatt 5

Aufg. 5: Betrachten Sie noch einmal die Matrizen A und $A+E$ aus Aufg. 3c,

$$A = \begin{pmatrix} 1 & 2 & 3 \\ 0 & 4 & 5 \\ 0 & 0 & 4.001 \end{pmatrix} \qquad \text{und} \qquad E = \begin{pmatrix} 0 & 0 & 0 \\ 0 & 0 & 0 \\ 0.001 & 0 & 0 \end{pmatrix}.$$

Laut Golub, van Loan [GvL96, p.322] ist $\lambda(A+E) \approx \{1.0001, 4.0582, 3.9427\}$.
a) Zeigen Sie: A ist diagonal-ähnlich.
b) Berechnen Sie die Bauer-Fike-Schranke und die Schur-Schranke für die gestörten Eigenwerte.
c) Wenn Punkt b) mit Hilfe von Matlab (Funktionen `eig, diag, norm, cond, schur`) durchgeführt wird: welches Format ist für den Output angemessener: `format short` oder `format short e`?

23. Nov. 04

Spektren und Pseudo-Spektren – Theorie, Numerik und Anwendungen (Meyer-Spasche)

Übungsblatt 6

Aufg. 6: Besorgen Sie sich über
`http://www.cs.berkeley.edu/~demmel/ma221/Matlab/eigscat.m`
oder über
`http://www.ipp.mpg.de/~rim/ws0405ueb.html`
das MATLAB-program `eigscat.m` . Bearbeiten Sie damit Question 4.14 aus dem Buch: *J. Demmel, Applied Numerical Linear Algebra, SIAM, Philadelphia 1997,* d.h. plotten Sie wie dort beschrieben jeweils das Spektrum der ursprünglich gegebenen Matrizen zusammen mit den Spektren von vielen reell und komplex gestörten Matrizen. Die gestörten Matrizen sollen mit dem Zufallsgenerator erzeugt werden.

Vergleichen Sie die erzeugten Punktemengen mit den durch Sätze 1, 3, 5 und 7 gegebenen Schranken. Was ist der Unterschied zwischen reellen und komplexen Störungen? Wie verändern sich die durch die gestörten Eigenwerte überdeckten Mengen für err $\to 0$?

Anlage: Kopie von S. 190 aus dem im Text zitierten Buch von Demmel.

30. November 2004

Spektren und Pseudo-Spektren – Theorie, Numerik und Anwendungen (Meyer-Spasche)

Übungsblatt 7

Aufg. 7: $A \in \mathbb{C}^{n\times n}$ habe die Eigenwerte $\lambda_1, \ldots, \lambda_n$ mit $\lambda_i \neq \lambda_j$ für $i \neq j$. Sei $\lambda(A) \subset D \subset \mathbb{C}$ und $f : D \to \mathbb{C}$ sei darstellbar als $f(z) = \sum_{i=-\infty}^{\infty} a_i z^i$. Weiter sei $Q^H A Q = T$ die Schur-Form von A (also Q unitär und T in oberer Dreiecksform).
a. Zeigen Sie, daß $f(A) = Q f(T) Q^H$ gilt, also die Berechnung von $f(A)$ über die Berechnung von $f(T)$ ausgeführt werden kann.
b. Zeigen Sie, daß $(f(T))_{ii} = f(T_{ii})$ gilt, also die Diagonale von $f(T)$ aus der Diagonalen von T berechnet werden kann.
c. Zeigen Sie, daß $Tf(T) = f(T)T$ gilt.
d. Benutzen Sie c. um zu zeigen, daß die i-te Oberdiagonale von $f(T)$ aus der $(i-1)$-sten Oberdiagonalen und den darunterliegenden Oberdiagonalen berechnet werden kann.
Punkte b, c und d können für eine Rekursionsformel für $f(T)$ benutzt werden: beginnend mit der Diagonalen, kann die erste, 2. Oberdiagonale berechnet werden, usw.

Aufg. 8 *Spektralabbildungssatz (spectral mapping theorem):* Sei $A \in \mathbb{C}^{n\times n}$. Wenden Sie Aufg. 7 auf die Schur-Form von A an um zu zeigen, daß

$$\lambda(f(A)) = f(\lambda(A)) = \{f(\lambda_i),\ \lambda_i \in \lambda(A),\ i = 1, \ldots, n\}.$$

(Der Beweis könnte auch unter Benutzung der Jordanschen Normalform geführt werden, siehe Demmel, p. 177, p. 188).

7. Dezember 2004

Spektren und Pseudo-Spektren – Theorie, Numerik und Anwendungen (Meyer-Spasche)

Übungsblatt 8

Aufg. 9: Für $A \in \mathbb{K}^{n\times n}$, $\mathbb{K} = \mathbb{R}$ oder $\mathbb{K} = \mathbb{C}$, und $\varepsilon \geq 0$ definieren wir als ε-*Pseudospektrum von* A

$$\lambda_\varepsilon(A) := \{z \in \mathbb{C} : z \in \lambda(A+E),\ E \in \mathbb{K}^{n\times n},\ ||E|| \leq \varepsilon\}.$$

So wie die Konditionszahlen von A, hängt auch das ε-Pseudospektrum von der verwendeten Norm ab.
Zeigen Sie für $A \in \mathbb{K}^{n\times n}$, $\varepsilon \geq 0$ und die 2-Norm:
a. $\lambda_\varepsilon(A)$ ist eine nicht-leere, abgeschlossene und beschränkte Menge, die höchstens n Zusammenhangskomponenten hat;
b. $\lambda_\varepsilon(A^H) = \{\bar{z} \in \mathbb{C} : z \in \lambda_\varepsilon(A)\}$;
c. Für jedes $c \in \mathbb{K}$ gilt $\lambda_\varepsilon(A + cI) = c + \lambda_\varepsilon(A)$;
d. Für jedes $c \in \mathbb{K}$ gilt $\lambda_{|c|\varepsilon}(cA) = c\lambda_\varepsilon(A)$;
e. Gilt $\lambda_\varepsilon(A_1 \oplus A_2) = \lambda_\varepsilon(A_1) \cup \lambda_\varepsilon(A_2)$,
mit $A_1 \oplus A_2 := \begin{pmatrix} A_1 0 \\ 0 A_2 \end{pmatrix}$ für $A_i \in \mathbb{K}^{n_i \times n_i}, i = 1, 2$?

21. Dezember 2004

Spektren und Pseudo-Spektren – Theorie, Numerik und Anwendungen (Meyer-Spasche)

Übungsblatt 9

Aufg. 10: Besuchen Sie den 'Pseudospectra Gateway'
`http://web.comlab.ox.ac.uk/projects/pseudospectra/`
und ergänzen Sie dort Ihr Wissen über Pseudospektren,
insbesondere zu dem Punkt 'software'.
Sammeln Sie eigene Erfahrungen mit einem der dort angebotenen Programme.

11. Januar 2005

Spektren und Pseudo-Spektren – Theorie, Numerik und Anwendungen (Meyer-Spasche)

Übungsblatt 10

Aufg. 11: Gegeben seien die Matrizen

$$A_1 := \begin{pmatrix} -1 & 1 \\ 0 & -1 \end{pmatrix} \qquad \text{und} \qquad A_2 := \begin{pmatrix} -1 & 5 \\ 0 & -2 \end{pmatrix}.$$

a. Plotten Sie die Höhenlinien der Pseudospektren $\lambda_\varepsilon(A_i),\ i = 1, 2,\ \varepsilon = 0.05(0.1)0.65$, entsprechend Definition 2.2 (Def. 2.1.3 in der Vorlesungsmitschrift) für beide Matrizen, in derselben Skala.
b. Lösen Sie die Dgl-systeme

$$\dot{u} = A_i\ u, \qquad u(0) = u_0, \quad i = 1, 2,$$

für Anfangswerte in einer Umgebung des Nullpunktes, insbesondere für $u_0 = (1, 5)^T$ und $u_0 = (5, 5)^T$. Wie verhalten sich die Lösungen für kleine Zeiten $t \leq 1$, wie für $t \to \infty$?

18.1.2005

Spektren und Pseudo-Spektren – Theorie, Numerik und Anwendungen (Meyer-Spasche)

Übungsblatt 11

Aufg. 12: Betrachten Sie im euklidischen $\mathbb{R}^2$ das dynamische System

$$\dot{u} = Au + ||u||Bu, \quad \text{mit } A := \begin{pmatrix} -R^{-1} & 1. \\ 0 & -2R^{-1} \end{pmatrix}, \quad B := \begin{pmatrix} 0 & -1 \\ 1 & 0 \end{pmatrix}.$$

Dies ist ein vereinfachtes mathematisches Modell für das Instabilwerden der ebenen Couetteströmung. Dabei steht R für die Reynoldszahl und $u \equiv 0$ für die Couetteströmung.

$u(t; u_0, R)$ beschreibt die gestörte Strömung mit Anfangswert u_0 und Parameterwert R. Wählen Sie $R = 25$ und plotten Sie $\ln ||u(t; u_0, R)||$ gegen t für $0 \leq t \leq T = 200$ für folgende Anfangswerte: $u_0 = (0, c)^H$, c so daß $||u_0|| = 10^{-n}$, $n = 7 : -1 : 2$. Qualitativ ändert sich das Verhalten der Lösungen bei $||u_0|| = 4.22 * 10^{-4}$. Wählen Sie deshalb zusätzlich noch $||u_0|| = 4 * 10^{-4}$ und $||u_0|| = 5 * 10^{-4}$.

25. Januar 2005

Spektren und Pseudo-Spektren – Theorie, Numerik und Anwendungen (Meyer-Spasche)

Übungsblatt 12

Aufg. 13: Betrachten Sie das dynamische System

$$\begin{aligned} \dot{x} &= (1-r)x - \frac{1}{2}y(1-\frac{x}{r}), \\ \dot{y} &= (1-r)y + \frac{1}{2}x(1-\frac{x}{r}), \qquad r = \sqrt{x^2+y^2} \end{aligned}$$

in kartesischen Koordinaten x, y oder auch in Polarkoordinaten r, φ:

$$\begin{aligned} \dot{r} &= r(1-r), \\ \dot{\varphi} &= \sin^2(\varphi/2). \end{aligned}$$

Finden Sie alle Gleichgewichtspunkte und invarianten Mengen und bestimmen Sie ihre Stabilität. Plotten Sie ein Phasendiagram mit MATLAB (Empfehlung: Anfangswerte sollten $0 < x_0 \leq 1.5;\ 0 < y_0 \leq 1$ erfüllen).
Dies ist ein Beispiel für: 'attraktiv' und 'asymptotisch stabil' sind nur bei linearen Systemen gleichwertig.

1. Februar 2005

Spektren und Pseudo-Spektren – Theorie, Numerik und Anwendungen (Meyer-Spasche)

Übungsblatt 13

Aufg. 14: Lesen Sie den als Kopie angehängten Text: *Barry A. Cipra: 'Are Eigenvalues Overvalued?' SIAM NEWS, Jan. 1995.* Dies ist ein Bericht von Barry Cipra (mathematician and writer) über einen Vortrag von Nick Trefethen über Pseudospektren, geschrieben für eine breite Leserschaft von Mathematikern.

Literatur

[Dem97] **J.W. Demmel (1997):** *Applied Numerical Linear Algebra*, **SIAM, Philadelphia, 1997**

[gate] Embree M, Trefethen LN: 'Pseudospectra Gateway': Zugang zu Webseiten mit Literatur, Software und Neuigkeiten über Pseudpspektren
`http://web.comlab.ox.ac.uk/projects/pseudospectra/`

[GL02] L. Grammont and A. Largillier: On ϵ-spectra and stability radii, J. Comp. Appl. Math **147** (2002) 453-469

[GvL96] **G.H. Golub, C.F. van Loan:** *Matrix Computations*, **3rd ed., The Johns Hopkins U Press, 1996**

[Ka82] T. Kato (1982):*A Short Introduction to Perturbation Theory for Linear Operators*, Springer Verlg New York, Heidelberg, Berlin

[LSY98] R.B. Lehoucq, D.C. Sorensen,C. Yang: *ARPACK User's Guide*, SIAM, Philadelphia, 1998

[Lu97] S.H. Lui: Computation of pseudospectra by continuation. SIAM J Sci Comp **18** (1997), 565 - 573

[Mol04] C. Moler: *Numerical Computing with MATLAB*, SIAM, Philadelphia 2004, and
`http://www.mathworks.com/moler/`

[SB.] Stoer, Bulirsch: *Numerische Mathematik*, Springer-Verlag,

[StH96] A.M. Stuart, A.R. Humphries (1996): *Dynamical Systems and Numerical Analysis* Cambridge University Press

[Tr99] **L.N. Trefethen: Spectra and Pseudospectra** pp 217 – 250 in: *The Graduate Student's Guide to Numerical Analysis '98*, M. Ainsworth et al. (eds.) Springer Verlag Berlin 1999

[TrE05] L.N. Trefethen, M. Embree (2005): *Spectra and Pseudospectra, The Behavior of Nonnormal Matrices and Operators* Princeton Univ. Press, Princeton, N.J.

[Wilk65] J.H. Wilkinson: *The Algebraic Eigenvalue Problem* Clarendon Press, Oxford 1965

[Wilk69] J.H. Wilkinson: *Rundungsfehler* Springer-Verlag, Berlin 1969

[WT01] T.G. Wright, L.N. Trefethen: Large-scale computation of pseudospectra using ARPACK and EIGS, SIAM J Sci Comp **23**, 591 - 605

[WT01b] T.G. Wright, L.N. Trefethen: Eigenvalues and Pseudospectra of Rectangular Matrices, Report NA-01/13, Oxford U 2001

Dynamische Systeme und Fluiddynamik:

[BRK94] D. Borba, K.S. Riedel, W. Kerner, G.T.A. Huysmans, M. Ottaviani, and P. J. Schmid: The pseudospectrum of the resistive magnetohydrodynamics operator: Resolving the resistive Alfvén paradox Phys. Plasmas **1** (1994) 3151-3160.

[DKS93] J.L.M. van Dorsselaer, J.f.B.M. Kraaijevanger, M.N. Spijker: Linear stability analysis in the numerical solution of initial value problems. Acta Numerica **2** (1993), p. 199-237

[DR] P.G. Drazin, W.H. Reid: *Hydrodynamic Stability.* Cambridge University Press 1981

[Dy] *An Album of Fluid Motion.* M. van Dyke, ed. The Parabolic Press, Stanford 1982

[HNW93] E. Hairer, S.P. Norsett, G. Wanner 1998: *Solving Ordinary Differential Equations, Part I: Nonstiff Problems* Springer Verlag Berlin

[Me99] R. Meyer-Spasche: *Pattern Formation in Viscous Flows.* ISNM 128, Birkhäuser Verlag 1999

[Num3] B. Simeon: *Numerik gewöhnlicher Differentialgleichungen*, Skriptum zur Vorlesung im WS 2003/04, TUM

[Tay] G.I. Taylor: Stability of a viscous liquid contained between two rotating cylinders. Phil. Trans. Roy. Soc. A 123(1923), 289-343; also contained in: G.I. Taylor's collected work

[TRD93] L.N. Trefethen, A.E. Trefethen, S.C. Reddy, T.A. Driscoll: Hydrodynamic stability without eigenvalues, Science **261** (1993), 578 – 584