19.06.2019
PD Dr. Peer Kunstmann
M.Sc. Michael Ullmann

Karlsruhe Institute of Technology

Numerische Methoden

Lösungsvorschläge zum 5. Übungsblatt

Aufgabe 1 (Induzierte Matrix p-Normen)

Zu $p \in [1, \infty]$ setzen wir die Vektornorm

$$\begin{cases} \|v\|_p := \sqrt[p]{\sum_{i=1}^n |v_i|^p} = \left(\sum_{i=1}^n |v_i|^p\right)^{\frac{1}{p}}, & \text{falls } p < \infty, \\ \|v\|_\infty := \max_{i=1,\dots,n} |v_i|, & \text{falls } p = \infty \end{cases}$$

für Vektoren $v = (v_1, \dots, v_n) \in \mathbb{C}^n$. Zu einer Matrix $A \in \mathbb{C}^{n\times n}$ und einer Vektorraumnorm $\|\cdot\|$ auf $\mathbb{C}^n$ definieren wir die durch $\|\cdot\|$-induzierte Matrixnorm $\|\cdot\|_*$ durch

$$\|A\|_* := \sup_{v\in\mathbb{C}^n\setminus\{0\}} \frac{\|Av\|}{\|v\|} = \max_{v\in\mathbb{C}^n:\, \|v\|=1} \|Av\|.$$

Wir wollen nun für die drei Fälle, dass $p = 1$, $p = 2$ und $p = \infty$ ist, genauer betrachten. Zeigen Sie, dass für alle Matrizen $A = (a_{ij})_{i,j=1,\dots,n} \in \mathbb{C}^{n\times n}$ die Gleichheiten

$$\begin{aligned} \|A\|_1 &= \max_{j=1,\dots,n} \sum_{i=1}^n |a_{ij}| \quad (\text{"Spaltensummennorm"}), \\ \|A\|_2 &= \sqrt{\lambda_{\max}\left(A^H A\right)} = \left(\lambda_{\max}\left(A^H A\right)\right)^{\frac{1}{2}} \quad (\text{"Spektralnorm"}), \\ \|A\|_\infty &= \max_{i=1,\dots,n} \sum_{j=1}^n |a_{ij}| \quad (\text{"Zeilensummennorm"}), \end{aligned}$$

dabei bezeichnen wir mit $\lambda_{\max}(B)$ für eine postiv semi-definite Matrix $B \in \mathbb{C}^{n\times n}$ den maximalen Eigenwert. Die Normen auf der linken Seite sind stets die von der jeweiligen p-Vektorraumnorm induzierten Matrixnorm.

Lösung von Aufgabe 1

Sei $A = (a_{ij})_{i,j=1,\dots,n} \in \mathbb{C}^{n\times n}$ eine Matrix.

Zu Zeigen: $\|A\|_1 = \max_{j=1,\dots,n} \sum_{i=1}^n |a_{ij}|$.

Es gilt laut der Definition der Vektornorm $\|\cdot\|_1$ und der Dreiecks-Ungleichung

$$\begin{aligned} \|Av\|_1 &= \sum_{i=1}^n |(Av)_i|^1 = \sum_{i=1}^n \left|\sum_{j=1}^n a_{ij} v_j\right| \le \sum_{i=1}^n \sum_{j=1}^n |a_{ij} v_j| \\ &= \sum_{j=1}^n \sum_{i=1}^n |a_{ij}|\,|v_j| = \sum_{j=1}^n \left[|v_j| \sum_{i=1}^n |a_{ij}|\right] \le \sum_{j=1}^n \left[|v_j| \max_{k=1,\dots,n} \sum_{i=1}^n |a_{ik}|\right] = \max_{k=1,\dots,n} \sum_{i=1}^n |a_{ik}| \sum_{j=1}^n |v_j|^1 \\ &= \max_{j=1,\dots,n} \sum_{i=1}^n |a_{ij}|\, \|v\|_1 \end{aligned}$$

für alle Vektoren $v = (v_1, \dots, v_n) \in \mathbb{C}^n$. Daraus folgt nun

$$\|A\|_1 = \max_{v\in\mathbb{C}^n:\, \|v\|_1=1} \|Av\|_1 \le \max_{v\in\mathbb{C}^n:\, \|v\|_1=1} \left[\max_{j=1,\dots,n} \sum_{i=1}^n |a_{ij}|\, \|v\|_1\right] = \max_{v\in\mathbb{C}^n:\, \|v\|_1=1} \left[\max_{j=1,\dots,n} \sum_{i=1}^n |a_{ij}|\right] = \max_{j=1,\dots,n} \sum_{i=1}^n |a_{ij}|.$$

Wähle nun einen Index $j_0 \in \{1, \ldots, n\}$ so, dass

$$\max_{j=1,\ldots,n} \sum_{i=1}^{n} |a_{ij}| = \sum_{i=1}^{n} |a_{ij_0}|$$

gilt. Setze nun $v^* := e_{j_0} \in \mathbb{C}^n$ ($j_0$te Einheitsvektor im $\mathbb{C}^n$), d.h.

$$(v^*)_j = \begin{cases} 1 \text{ für } j = j_0 \\ 0 \text{ für } j \in \{1, \ldots, n\} \setminus \{j_0\}. \end{cases}$$

Damit gilt:

$$\begin{aligned} \|Av^*\|_1 &= \sum_{i=1}^{n} |(Av^*)_i|^1 = \sum_{i=1}^{n} \left| \sum_{j=1}^{n} a_{ij} v_j^* \right| = \sum_{i=1}^{n} \left| a_{ij_0} v_{j_0}^* + \sum_{j=1,\ j\neq j_0}^{n} a_{ij} v_j^* \right| \\ &= \sum_{i=1}^{n} \left| a_{ij_0} + \sum_{j=1,\ j\neq j_0}^{n} 0 \right| = \sum_{i=1}^{n} |a_{ij_0}| = \max_{j=1,\ldots,n} \sum_{i=1}^{n} |a_{ij}|. \end{aligned}$$

Weiter ist

$$\|v^*\|_1 = \sum_{j=1}^{n} \left|v_j^*\right|^1 = \left|v_{j_0}^*\right| + \sum_{j=1,\ j\neq j_0}^{n} \left|v_j^*\right| = 1 + \sum_{j=1,\ j\neq j_0}^{n} 0 = 1.$$

Also folgt:

$$\|A\|_1 = \max_{v\in\mathbb{C}^n:\ \|v\|_1=1} \|Av\|_1 \geq \|Av^*\|_1 = \max_{j=1,\ldots,n} \sum_{i=1}^{n} |a_{ij}|.$$

Zusammengefasst haben wir

$$\max_{j=1,\ldots,n} \sum_{i=1}^{n} |a_{ij}| \leq \|A\|_1 \leq \max_{j=1,\ldots,n} \sum_{i=1}^{n} |a_{ij}|,$$

was uns die Gleichheit

$$\|A\|_1 = \max_{j=1,\ldots,n} \sum_{i=1}^{n} |a_{ij}|$$

liefert. □

Zu Zeigen: $\|A\|_2 = \sqrt{\lambda_{\max}(A^H A)}$.
Die Matrix $A^H A$ ist positiv semi-definit, d.h. für alle Eigenwerte $\lambda \in \mathbb{C}$ der Matrix $A^H A$ gilt

$$\lambda \in [0, \infty)$$

und die Matrix $A^H A$ ist diagonalisierbar mit der Diagonalmatrix

$$D = \operatorname{diag}(\lambda_1, \ldots, \lambda_n) \in \mathbb{C}^{n\times n} \text{ mit Eigenwerten } 0 \leq \lambda_1 \leq \lambda_2 \leq \ldots \leq \lambda_n \text{ der Matrix } A^H A,$$

sowie einer unitären Matrix $U \in \mathbb{C}^{n\times n}$ (d.h. $UU^H = U^H U = I_n$ bzw. $U^{-1} = U^H$) mit

$$A^H A = UDU^H.$$

Nach der Definition der Vektornorm $\|\cdot\|_2$ und dem euklidischen Skalarprodukt gilt:

$$\begin{aligned} \|Av\|_2^2 &= \sum_{i=1}^{n} |(Av)_i|^2 = \langle Av, Av\rangle_2 = (Av)^H (Av) = v^H A^H Av = v^H UDU^H v = \left(DU^H v\right)\left(U^H v\right) \\ &= \left\langle DU^H v, U^H v\right\rangle_2 = \langle Dy, y\rangle_2 = \sum_{i=1}^{n} \overline{(Dy)}_i \cdot y_i = \sum_{i=1}^{n} \overline{\lambda}_i \overline{y}_i \cdot y_i = \sum_{i=1}^{n} \lambda_i |y_i|^2 \\ &\leq \sum_{i=1}^{n} \lambda_n |y_i|^2 = \lambda_n \sum_{i=1}^{n} |y_i|^2 = \lambda_n \langle y, y\rangle_2 = \lambda_n \left\langle U^H v, U^H v\right\rangle_2 = \lambda_n \left(U^H v\right)\left(U^H v\right) = \lambda_n v^H UU^H v = \lambda_n v^H I_n v \\ &= \lambda_n v^H v = \lambda_n \langle v, v\rangle_2 = \lambda_n \|v\|_2^2 = \lambda_{\max}\left(A^H A\right)\|v\|_2^2 \end{aligned}$$

für alle Vektoren $v = (v_1, \ldots, v_n) \in \mathbb{C}^n$ mit $y := U^H v$. Demnach gilt für alle Vektoren $v \in \mathbb{C}^n$:

$$\|Av\|_2 \leq \sqrt{\lambda_{\max}(A^H A)\|v\|_2^2} = \sqrt{\lambda_{\max}(A^H A)}\,\|v\|_2.$$

Daraus folgt nun nach der Definition der Matrixnorm $\|\cdot\|_2$:

$$\|A\|_2 = \max_{v\in\mathbb{C}^n:\ \|v\|_2=1} \|Av\|_2 \leq \max_{v\in\mathbb{C}^n:\ \|v\|_2=1} \left[\sqrt{\lambda_{\max}\left(A^HA\right)}\,\|v\|_2\right] = \max_{v\in\mathbb{C}^n:\ \|v\|_2=1} \sqrt{\lambda_{\max}\left(A^HA\right)} = \sqrt{\lambda_{\max}\left(A^HA\right)}.$$

Wähle nun $v^* \in \mathbb{C}^n\backslash\{0\}$ als einen Eigenvektor zum Eigenwert $\lambda_n = \lambda_{\max}\left(A^HA\right)$ mit $\|v^*\|_2 = 1$. Dann erhalten wir

$$\begin{aligned}\|Av^*\|_2^2 &= \langle Av^*, Av^*\rangle_2 = (Av^*)^H (Av^*) = (v^*)^H A^HAv^* = (v^*)^H \left(A^HAv^*\right) = (v^*)^H \lambda_n v^* \\ &= \lambda_n (v^*)^H v^* = \lambda_n \langle v^*, v^*\rangle_2 = \lambda_{\max}\left(A^HA\right)\|v^*\|_2^2 = \lambda_{\max}\left(A^HA\right).\end{aligned}$$

Also gilt:

$$\|A\|_2 = \max_{v\in\mathbb{C}^n:\ \|v\|_2=1} \|Av\|_2 \geq \|Av^*\|_2 = \lambda_{\max}\left(A^HA\right),$$

d.h. wir haben

$$\lambda_{\max}\left(A^HA\right) \leq \|A\|_2 \leq \lambda_{\max}\left(A^HA\right)$$

und somit auch

$$\|A\|_2 = \lambda_{\max}\left(A^HA\right)$$

gezeigt. □

Zu Zeigen: $\|A\|_\infty = \max\limits_{i=1,\dots,n} \sum\limits_{j=1}^n |a_{ij}|$.

Es gilt laut der Definition der Vektornorm $\|\cdot\|_\infty$ und der Dreiecks-Ungleichung

$$\begin{aligned}\|Av\|_\infty &= \max_{i=1,\dots,n} |(Av)_i| = \max_{i=1,\dots,n} \left|\sum_{j=1}^n a_{ij}v_j\right| \leq \max_{i=1,\dots,n} \sum_{j=1}^n |a_{ij}v_j| = \max_{i=1,\dots,n} \sum_{j=1}^n |a_{ij}|\,|v_j| \\ &\leq \max_{i=1,\dots,n} \sum_{j=1}^n \left[|a_{ij}| \max_{k=1,\dots,n} |v_k|\right] = \max_{i=1,\dots,n} \sum_{j=1}^n |a_{ij}|\,\|v\|_\infty\end{aligned}$$

für alle Vektoren $v \in \mathbb{C}^n$. Damit folgt nun:

$$\|A\|_\infty = \max_{v\in\mathbb{C}^n:\ \|v\|_\infty=1} \|Av\|_\infty \leq \max_{v\in\mathbb{C}^n:\ \|v\|_\infty=1} \left[\max_{i=1,\dots,n} \sum_{j=1}^n |a_{ij}|\,\|v\|_\infty\right] = \max_{v\in\mathbb{C}^n:\ \|v\|_\infty=1} \max_{i=1,\dots,n} \sum_{j=1}^n |a_{ij}| = \max_{i=1,\dots,n} \sum_{j=1}^n |a_{ij}|.$$

Wähle nun einen Index $i_0 \in \{1,\dots,n\}$ mit

$$\max_{i=1,\dots,n} \sum_{j=1}^n |a_{ij}| = \sum_{j=1}^n |a_{i_0j}|.$$

Setze dann den Vektor

$$v^* := \begin{pmatrix} \operatorname{sign}(a_{i_01}) \\ \vdots \\ \operatorname{sign}(a_{i_0n}) \end{pmatrix} \in \mathbb{C}^n,$$

wobei $\operatorname{sign}\colon \mathbb{R} \to \{-1, 0, 1\}$ die Vorzeichenfunktion ist, welche gegeben ist durch

$$\operatorname{sign}(x) = \begin{cases} -1 \text{ für } x < 0 \\ 0 \text{ für } x = 0 \\ 1 \text{ für } x > 0 \end{cases} \qquad \text{für } x \in \mathbb{R}.$$

Es gilt insbesondere $x\operatorname{sign}(x) = |x|$ für alle $x \in \mathbb{R}$. Ist $A = 0$, so ist $a_{ij} = 0$ für alle $i, j = 1,\dots,n$ und es folgt:

$$\|A\|_\infty = 0 = \sum_{j=1}^n |a_{ij}| \text{ für alle } i = 1,\dots,n, \text{ d.h. } \|A\|_\infty = \max_{i=1,\dots,n} \sum_{j=1}^n |a_{ij}| = 0.$$

Sei also nun $A \neq 0$, dann muss es einen Index $j_0 \in \{1,\dots,n\}$ geben mit

$$a_{i_0j_0} \neq 0.$$

Also gilt

$$\|v^*\|_\infty = \max_{j=1,\dots,n} \left|v_j^*\right| = \max_{j=1,\dots,n} |\operatorname{sign}(a_{i_0j})| = |\operatorname{sign}(a_{i_0j_0})| = 1.$$

Weiter ist

$$\begin{aligned}\|Av^*\|_\infty &= \max_{i=1,\dots,n} |(Av^*)_i| = \max_{i=1,\dots,n} \left|\sum_{j=1}^n a_{ij} v_j^*\right| = \max_{i=1,\dots,n} \left|\sum_{j=1}^n a_{ij} \operatorname{sign}(a_{i_0 j})\right| = \left|\sum_{j=1}^n a_{i_0 j} \operatorname{sign}(a_{i_0 j})\right| = \left|\sum_{j=1}^n |a_{i_0 j}|\right| \\ &= \sum_{j=1}^n |a_{i_0 j}| = \max_{i=1,\dots,n} \sum_{j=1}^n |a_{ij}|\,.\end{aligned}$$

Daraus folgt nun

$$\|A\|_\infty = \max_{v\in\mathbb{C}^n:\ \|v\|_\infty=1} \|Av\|_\infty \geq \|Av^*\|_\infty = \max_{i=1,\dots,n} \sum_{j=1}^n |a_{ij}|\,,$$

und wir bekommen

$$\max_{i=1,\dots,n} \sum_{j=1}^n |a_{ij}| \leq \|A\|_\infty \leq \max_{i=1,\dots,n} \sum_{j=1}^n |a_{ij}|\,,$$

sowie

$$\|A\|_\infty = \max_{i=1,\dots,n} \sum_{j=1}^n |a_{ij}|\,.$$

□

Bemerkung: Kurz zur Gleichheit der induzierten Matrixnorm in der Aufgabenstellung. Die Funktion $f\colon \mathbb{C}^n \to [0,\infty)$ gegeben durch

$$f(v) = \|Av\| \text{ für } v \in \mathbb{C}^n$$

ist Lipschitz-stetig, denn es gilt für alle Vektoren $u, v \in \mathbb{C}^n$ laut der umgekehrten Dreiecks-Ungleichung:

$$|f(u) - f(v)| = |\|Au\| - \|Av\|| \leq \|Au - Av\| = \|A(u-v)\| \leq \|A\|_* \|u - v\|\,,$$

da die induzierte Matrixnorm $\|\cdot\|_*$ verträglich mit der Vektornorm $\|\cdot\|$ ist. Insbesondere ist die Funktion f stetig, also existiert das Maximum der Funktion auf kompakten Mengen wie es die Menge

$$S^{n-1} := \{v \in \mathbb{C}^n\colon\ \|v\| = 1\} \subseteq \mathbb{C}^n$$

ist. Dann folgt:

$$\begin{aligned}\|A\|_* &= \sup_{v\in\mathbb{C}^n\setminus\{0\}} \frac{\|Av\|}{\|v\|} = \sup_{v\in\mathbb{C}^n\setminus\{0\}} \left\|A\frac{v}{\|v\|}\right\| \\ &= \sup_{u\in\mathbb{C}^n:\ \|u\|=1} \|Au\| = \max_{v\in\mathbb{C}^n:\ \|v\|=1} \|Av\|\,.\end{aligned}$$

□

Aufgabe 2 (Zu induzierten Matrixnormen)

(a) Sei $\|\cdot\|_*$ eine induzierte Matrixnorm auf $\mathbb{C}^{n\times n}$. Zeigen Sie, dass für die Einheitsmatrix $I_n \in \mathbb{C}^{n\times n}$ stets gilt:

$$\|I_n\|_* = 1.$$

(b) Finden Sie eine Matrixnorm $\|\cdot\|_*$ auf $\mathbb{C}^{n\times n}$ so, dass für die Einheitsmatrix $I_n \in \mathbb{C}^{n\times n}$

$$\|I_n\|_* \neq 1$$

gilt.

(c) Seien $\|\cdot\|_*$ eine durch $\|\cdot\|$ induzierte Matrixnorm auf $\mathbb{C}^{n\times n}$ und $A \in \mathbb{C}^{n\times n}$ eine reguläre Matrix. Zeigen Sie:

$$\|A\|_* = \frac{1}{\min\limits_{v\in\mathbb{C}^n\colon \|v\|=1} \|A^{-1}v\|}.$$

Lösung von Aufgabe 2

(a) Es gilt für die Einheitsmatrix $I_n \in \mathbb{C}^{n\times n}$ laut der Definition der induzierten Matrixnorm (siehe auch Aufgabe 1.):

$$\|I_n\|_* = \max_{v\in\mathbb{C}^n\colon \|v\|=1} \|I_n v\| = \max_{v\in\mathbb{C}^n\colon \|v\|=1} \|v\| = \max_{v\in\mathbb{C}^n\colon \|v\|=1} 1 = 1.$$

□

(b) Ein Beispiel ist z.B. die sogenannte Frobenius-Norm, welche gegeben ist durch

$$\|A\|_F := \sqrt{\mathrm{Spur}\,(A^H A)} = \sqrt{\sum_{i=1}^{n}\sum_{j=1}^{n} |a_{ij}|^2}$$

für Matrizen $A \in \mathbb{C}^{n\times n}$, wobei der Spuroperator $\mathrm{Spur}\colon \mathbb{C}^{n\times n} \to \mathbb{C}$ definiert ist durch

$$\mathrm{Spur}(B) = \sum_{i=1}^{n} b_{ii} \text{ für Matrizen } B = (b_{ij})_{i,j=1,\dots,n} \in \mathbb{C}^{n\times n}.$$

Nun gilt für die Einheitsmatrix $I_n \in \mathbb{C}^{n\times n}$:

$$\|I_n\|_F = \sqrt{\mathrm{Spur}\,(I_n^H I_n)} = \sqrt{\mathrm{Spur}\,(I_n)} = \sqrt{\sum_{i=1}^{n} 1} = \sqrt{n} \neq 1 \text{ für } n \geq 2.$$

□

(c) Es gilt laut der Definition der induzierten Matrixnorm (siehe auch Aufgabe 1.):

$$\begin{aligned}\|A\|_* &= \sup_{v\in\mathbb{C}^n\colon v\neq 0} \frac{\|Av\|}{\|v\|} = \sup_{w\in\mathbb{C}^n\colon w\neq 0} \frac{\|w\|}{\|A^{-1}w\|} = \sup_{w\in\mathbb{C}^n\colon w\neq 0} \frac{1}{\frac{\|A^{-1}w\|}{\|w\|}} \\ &= \sup_{w\in\mathbb{C}^n\colon w\neq 0} \frac{1}{\left\|A^{-1}\frac{w}{\|w\|}\right\|} = \max_{v\in\mathbb{C}^n\colon \|v\|=1} \frac{1}{\|A^{-1}v\|} = \frac{1}{\min\limits_{v\in\mathbb{C}^n\colon \|v\|=1} \|A^{-1}v\|},\end{aligned}$$

da für beliebige beschränkte, nicht-leere Mengen $\emptyset \neq M_1, M_2 \subseteq (0,\infty)$ mit

$$M_2 := \left\{\frac{1}{x}\colon x \in M_1\right\}$$

gilt:

$$\sup M_1 = \frac{1}{\inf M_2}.$$

□

Aufgabe 3 (Satz von Neumann)

Seien $S \in \mathbb{C}^{n\times n}$ eine Matrix und $I_n \in \mathbb{C}^{n\times n}$ die Einheitsmatrix in $\mathbb{C}^{n\times n}$, sowie $\|\cdot\|_*$ eine submultiplikative Matrixnorm auf $\mathbb{C}^{n\times n}$, d.h. $\|\cdot\|_*$ ist eine Norm auf $\mathbb{C}^{n\times n}$ mit der Eigenschaft

$$\|AB\|_* \le \|A\|_* \|B\|_* \text{ für alle Matrizen } A, B \in \mathbb{C}^{n\times n}.$$

Zeigen Sie, dass falls $\|S\|_* < 1$ ist, ist die Matrix $I_n + S \in \mathbb{C}^{n\times n}$ regulär und die inverse Matrix $(I_n + S)^{-1}$ hat die folgende Form und erfüllt die nachfolgende Abschätzung

$$(I_n + S)^{-1} = \sum_{k=0}^{\infty} (-1)^k S^k,$$
$$\left\|(I_n + S)^{-1}\right\|_* \le \frac{\|I_n\|_*}{1 - \|S\|_*}.$$

Dabei setzen wir $A^0 := I_n$ für jede Matrix $A \in \mathbb{C}^{n\times n}$. (**Hinweis**: Erinnern Sie sich an die geometrische Reihe.)

Lösung von Aufgabe 3

Wir bemerken, dass

$$\lim_{m\to\infty} S^m = 0$$

gilt, denn es ist

$$0 \le \lim_{m\to\infty} \|S^m - 0\|_* = \lim_{m\to\infty} \|S^m\|_* \le \lim_{m\to\infty} \|S\|_*^m = 0,$$

wegen der Submultiplikativität der Matrixnorm $\|\cdot\|_*$ und da die Norm $\|S\|_* < 1$ ist. Setze

$$T_m := \sum_{k=0}^{m} (-1)^k S^k,\ m \in \mathbb{N}_0,\ T := \lim_{m\to\infty} T_m = \lim_{m\to\infty} \sum_{k=0}^{m} (-1)^k S^k = \sum_{k=0}^{\infty} (-1)^k S^k,$$

wobei die Reihe existiert, weil die Matrixfolge $(T_m)_{m\in\mathbb{N}_0}$ eine Cauchy-Folge in $\mathbb{C}^{n\times n}$ bzgl. der Matrixnorm $\|\cdot\|_*$ ist, denn wir haben für $l, m \in \mathbb{N}$ mit $l \le m$ laut der Dreiecks-Ungleichung und der Submultiplikativität der Matrixnorm $\|\cdot\|_*$:

$$\|T_m - T_l\|_* = \left\|\sum_{k=0}^{m} (-1)^k S^k - \sum_{k=0}^{m} (-1)^k S^k\right\|_* = \left\|\sum_{k=l}^{m} (-1)^k S^k\right\|_* \le \sum_{k=l}^{m} \left\|(-1)^k S^k\right\|_*$$
$$= \sum_{k=l}^{m} \left\|S^k\right\|_* \le \sum_{k=1}^{m} \|S\|_*^k \ \to 0 \text{ für } l, m \to \infty,$$

wegen der Konvergenz der Reihe

$$\sum_{k=0}^{\infty} \|S\|_*^k = \frac{1}{1 - \|S\|_*}$$

nach der geometrischen Reihe, da $\|S\|_* < 1$ ist. Cauchy-Folgen in $\mathbb{C}^{n\times n}$ sind konvergent, damit existiert die Matrix T. Weiter gilt:

$$ST_m = S\sum_{k=0}^{m} (-1)^k S^k = \sum_{k=0}^{m} (-1)^k S^{k+1} = \sum_{k=0}^{m} (-1)^k S^k S = T_m S \text{ für alle } m \in \mathbb{N}_0.$$

Damit sehen wir ein, dass

$$T_m (I_n + S) = T_m + T_m S = \sum_{k=0}^{m} (-1)^k S^k + \sum_{k=0}^{m} (-1)^k S^k S = \sum_{k=0}^{m} (-1)^k S^k - \sum_{k=0}^{m} (-1)^{k+1} S^{k+1}$$
$$= I_n + \sum_{k=1}^{m} (-1)^k S^k + \sum_{k=1}^{m} (-1)^k S^k - (-1)^{m+1} S^{m+1} = I_n - (-1)^{m+1} S^{m+1} \ \to I_n \text{ für } m \to \infty$$

gilt. Weiter ist daher auch

$$(I_n + S) T_m = I_n T_m + S T_m = T_m + T_m S = T_m (I_n + S) \ \to I_n \text{ für } m \to \infty.$$

Also erhalten wir

$$(I_n + S) T = \lim_{m\to\infty} (I_n + S) T_m = I_n = \lim_{m\to\infty} T_m (I_n + S) = T (I_n + S),$$

d.h.

$$(I_n + S)^{-1} = T = \sum_{k=0}^{\infty} (-1)^k S^k.$$

Weiter erhalten wir die Abschätzung nach der geometrischen Reihe, da die Matrixnorm $\|S\|_* < 1$ ist,

$$\begin{aligned}\|T\|_* &= \left\| \sum_{k=0}^{\infty} (-1)^k S^k \right\|_* \leq \sum_{k=0}^{\infty} \left\| (-1)^k S^k \right\|_* = \sum_{k=0}^{\infty} \left\| S^k \right\|_* = \left\| S^0 \right\|_* + \sum_{k=1}^{\infty} \left\| S^k \right\|_* = \|I_n\|_* + \sum_{k=1}^{\infty} \left\| I_n S^k \right\|_* \\ &\leq \|I_n\|_* + \sum_{k=1}^{\infty} \|I_n\|_* \left\| S^k \right\|_* \leq \|I_n\|_* + \|I_n\|_* \sum_{k=1}^{\infty} \|S\|_*^k = \|I_n\|_* \left(1 + \sum_{k=1}^{\infty} \|S\|_*^k \right) = \|I_n\|_* \sum_{k=0}^{\infty} \|S\|_*^k = \frac{\|I_n\|_*}{1 - \|S\|_*}.\end{aligned}$$

□

Aufgabe 4 (Fehlerabschätzung und Kondition)

Prüfen Sie auch in allen drei folgenden Teilaufgaben nach, dass die Bedingungen für die jeweiligen Sätze erfüllt sind.

(a) Gegeben seien die Matrix

$$A = \begin{pmatrix} 4 & 1 \\ 3 & 1 \end{pmatrix} \text{ und der Vektor } b = \begin{pmatrix} 1 \\ 3 \end{pmatrix}.$$

Weiter erfüllt die Störmatrix $\Delta A \in \mathbb{R}^{2\times 2}$ und der Störvektor Δb die Abschätzungen

$$\|\Delta A\|_2 \leq \frac{1}{20} \text{ und } \|\Delta b\|_2 \leq 10^{-3}.$$

Wie groß ist maximal der relative Fehler (in der Spektralnorm $\|\cdot\|_2$) der Lösung des linearen Gleichungssystems

$$(A + \Delta A)(x + \Delta x) = b + \Delta b?$$

(b) Gegeben seien die Matrix

$$A = \begin{pmatrix} 1 & 1 \\ 0 & 2 \end{pmatrix} \text{ und der Vektor } b = \begin{pmatrix} 1 \\ 1 \end{pmatrix}.$$

Weiter seien die Störungen gegeben durch $\Delta A = 0 \in \mathbb{R}^{2\times 2}$ und

$$\Delta b = \begin{pmatrix} \varepsilon \\ 0 \end{pmatrix}$$

für ein $\varepsilon > 0$. Wie klein muss die Störung Δb sein (d.h. wie muss $\varepsilon > 0$ gewählt werden) um einen absoluten bzw. relativen Fehler in der Zeilensummennorm $\|\cdot\|_\infty$ von höchstens 10^{-3} zu gewährleisten?

(c) Gegeben seien die Matrix

$$A = \begin{pmatrix} 1 & 1 \\ 0 & 2 \end{pmatrix} \text{ und der Vektor } b = \begin{pmatrix} 1 \\ 1 \end{pmatrix}.$$

Weiter seien die Störungen gegeben durch $\Delta b = 0 \in \mathbb{R}^2$ und

$$\Delta A = \frac{\varepsilon}{2} \begin{pmatrix} 1 & 0 \\ 1 & 0 \end{pmatrix}$$

für ein $\varepsilon > 0$. Wie klein muss die Störung ΔA sein (d.h. wie muss $\varepsilon > 0$ gewählt werden) um einen relativen Fehler in der Spaltensummennorm $\|\cdot\|_1$ von höchstens 10^{-3} zu gewährleisten?

Lösung von Aufgabe 4

(a) Wir berechnen alles, was wir benötigen:

$$\begin{aligned}
\det(A) &= \det \begin{pmatrix} 4 & 1 \\ 3 & 1 \end{pmatrix} = 4 \cdot 1 - 3 \cdot 1 = 4 - 3 = 1 \neq 0, \\
A^{-1} &= \begin{pmatrix} 4 & 1 \\ 3 & 1 \end{pmatrix}^{-1} = \frac{1}{1} \begin{pmatrix} 1 & -1 \\ -3 & 4 \end{pmatrix} = \begin{pmatrix} 1 & -1 \\ -3 & 4 \end{pmatrix} \text{ nach der Cramerschen Regel}, \\
A^H &= \overline{A}^T = A^T = \begin{pmatrix} 4 & 1 \\ 3 & 1 \end{pmatrix}^T = \begin{pmatrix} 4 & 3 \\ 1 & 1 \end{pmatrix}, \\
\left(A^{-1}\right)^H &= \overline{A^{-1}}^T = \left(A^{-1}\right)^T = \begin{pmatrix} 1 & -3 \\ -1 & 4 \end{pmatrix}, \\
A^H A &= \begin{pmatrix} 4 & 3 \\ 1 & 1 \end{pmatrix} \begin{pmatrix} 4 & 1 \\ 3 & 1 \end{pmatrix} = \begin{pmatrix} 25 & 7 \\ 7 & 2 \end{pmatrix}, \\
\left(A^{-1}\right)^H A^{-1} &= \begin{pmatrix} 1 & -3 \\ -1 & 4 \end{pmatrix} \begin{pmatrix} 1 & -1 \\ -3 & 4 \end{pmatrix} = \begin{pmatrix} 10 & -13 \\ -13 & 17 \end{pmatrix}, \\
\|b\|_2 &= \sqrt{|1|^2 + |3|^2} = \sqrt{1+9} = \sqrt{10}, \\
\|I_2\|_2 &= 1 \text{ laut Aufgabe 2.(a).}
\end{aligned}$$

Weiter lauten die charakteristischen Polynome der Matrizen $A^H A$ und $\left(A^{-1}\right)^H A^{-1}$ gerade

$$p_{A^H A}(\lambda) = \det\left(A^H A - \lambda I_2\right) = \det \begin{pmatrix} 25-\lambda & 7 \\ 7 & 2-\lambda \end{pmatrix} = (25-\lambda)(2-\lambda) - 7 \cdot 7 = 50 - 25\lambda - 2\lambda + \lambda^2 - 49$$

$$= \lambda^2 - 27\lambda + 1,$$

$$p_{(A^{-1})^H A^{-1}}(\lambda) = \det\left(\left(A^{-1}\right)^H A^{-1} - \lambda I_2\right) = \det\begin{pmatrix} 10-\lambda & -13 \\ -13 & 17-\lambda \end{pmatrix} = (10-\lambda)(17-\lambda) - (-13)\cdot(-13)$$

$$= 170 - 10\lambda - 17\lambda + \lambda^2 - 169 = \lambda^2 - 27\lambda + 1 = p_{A^H A}(\lambda)$$

für alle $\lambda \in \mathbb{C}$, also haben die beiden charakteristischen Polynome $p_{A^H A}$ und $p_{(A^{-1})^H A^1}$ insbesondere auch dieselben Nullstellen $\lambda_1, \lambda_2 \in \mathbb{C}$. Es gilt z.B. laut der Mitternachtsformel:

$$\lambda_{1,2} = \frac{27 \pm \sqrt{(-27)^2 - 4}}{2} = \frac{27 \pm \sqrt{729-4}}{2} = \frac{27 \pm \sqrt{725}}{2} = \frac{27 \pm \sqrt{25\cdot 29}}{2} = \frac{27 \pm 5\sqrt{29}}{2}.$$

Damit lautet nun

$$\lambda_{\max}\left(A^H A\right) = \lambda_{\max}\left(\left(A^{-1}\right)^H A^{-1}\right) = \max\left\{\frac{27-5\sqrt{29}}{2}, \frac{27+5\sqrt{29}}{2}\right\} = \frac{27+5\sqrt{29}}{2}.$$

Wir erhalten so nun die letzten fehlenden Größen laut Aufgabe 1.:

$$\|A\|_2 = \sqrt{\lambda_{\max}\left(A^H A\right)} = \sqrt{\frac{27+5\sqrt{29}}{2}},$$

$$\left\|A^{-1}\right\|_2 = \sqrt{\lambda_{\max}\left(\left(A^{-1}\right)^H A^{-1}\right)} = \sqrt{\frac{27+5\sqrt{29}}{2}},$$

$$\kappa_2(A) := \operatorname{cond}_2(A) := \operatorname{cond}_{\|\cdot\|_2}(A) = \|A\|_2 \left\|A^{-1}\right\|_2 = \sqrt{\frac{27+5\sqrt{29}}{2}}\sqrt{\frac{27+5\sqrt{29}}{2}} = \frac{27+5\sqrt{29}}{2}.$$

Um die Fehlerabschätzung anwenden zu dürfen, benötigen wir, dass die Matrix A invertierbar ist, was oben gezeigt wurde (Determinante ist ungleich null) und die Bedingung $\left\|A^{-1}\right\|_2 \|\Delta A\|_2 < 1$ muss erfüllt sein, was wegen

$$\left\|A^{-1}\right\|_2 \|\Delta A\|_2 = \sqrt{\frac{27+5\sqrt{29}}{2}}\,\|\Delta A\|_2 \le \sqrt{\frac{27+5\sqrt{36}}{2}}\cdot\frac{1}{20} = \sqrt{\frac{27+5\cdot 6}{2}}\cdot\frac{1}{20} = \sqrt{\frac{27+30}{2}}\cdot\frac{1}{20}$$

$$= \sqrt{\frac{57}{2}}\cdot\frac{1}{20} \le \sqrt{\frac{60}{2}}\cdot\frac{1}{20} = \sqrt{30}\cdot\frac{1}{20} \le \frac{\sqrt{36}}{20} = \frac{6}{20} = \frac{3}{10} < 1$$

der Fall ist. Demnach gilt laut der bewiesenen Fehlerabschätzung aus der Vorlesung für den relativen Fehler $\frac{\|\Delta x\|_2}{\|x\|_2}$ des gestörten linearen Gleichungssystems

$$(A + \Delta A)(x + \Delta x) = b + \Delta b$$

gerade:

$$\frac{\|\Delta x\|_2}{\|x\|_2} \le \frac{\|I_2\|_2 \cdot \kappa_2(A)}{1 - \kappa_2(A)\cdot\frac{\|\Delta A\|_2}{\|A\|_2}}\left(\frac{\|\Delta A\|_2}{\|A\|_2} + \frac{\|\Delta b\|_2}{\|b\|_2}\right)$$

$$\le \frac{\frac{27+5\sqrt{29}}{2}}{1 - \|A^{-1}\|_2\|\Delta A\|_2}\left(\frac{\frac{1}{20}}{\sqrt{\frac{27+5\sqrt{29}}{2}}} + \frac{10^{-3}}{\sqrt{10}}\right)$$

$$\le \frac{\frac{27+5\sqrt{29}}{2}}{1 - \frac{3}{10}}\left(\frac{\sqrt{2}}{20\sqrt{27+5\sqrt{29}}} + 10^{-\frac{7}{2}}\right)$$

$$= \frac{\frac{27+5\sqrt{29}}{2}}{\frac{7}{10}}\left(\frac{\sqrt{2}}{20\sqrt{27+5\sqrt{29}}} + 10^{-\frac{7}{2}}\right)$$

$$= \frac{5\left(27+5\sqrt{29}\right)}{7}\left(\frac{\sqrt{2}}{20\sqrt{27+5\sqrt{29}}} + 10^{-\frac{7}{2}}\right)$$

$$= \frac{135+25\sqrt{29}}{7}\left(\frac{\sqrt{2}}{20\sqrt{27+5\sqrt{29}}} + 10^{-\frac{7}{2}}\right)$$

□

(b) Wir berechnen alles, was wir benötigen:

$$\det(A) = \begin{pmatrix} 1 & 1 \\ 0 & 2 \end{pmatrix} = 1\cdot 2 - 0\cdot 1 = 2 - 0 = 2 \neq 0,$$

$$A^{-1} = \begin{pmatrix} 1 & 1 \\ 0 & 2 \end{pmatrix}^{-1} = \frac{1}{2}\begin{pmatrix} 2 & -1 \\ 0 & 1 \end{pmatrix} \text{ nach der Cramerschen Regel,}$$
$$\|b\|_\infty = \max_{i=1,2} |b_i| = \max\{1,\ 1\} = 1,$$
$$\|I_2\|_\infty = 1 \text{ laut Aufgabe 2.(a),}$$
$$\|A\|_\infty = \max_{i=1,2} (|A_{i1}| + |A_{i2}|) = \max\{1+1,\ 0+2\} = \max\{2,\ 2\} = 2 \text{ laut Aufgabe 1.,}$$
$$\left\|A^{-1}\right\|_\infty = \max_{i=1,2} \left(\left|\left(A^{-1}\right)_{i1}\right| + \left|\left(A^{-1}\right)_{i2}\right|\right) = \max\left\{\frac{1}{2}(2+1),\ \frac{1}{2}(0+1)\right\} = \max\left\{\frac{3}{2},\ \frac{1}{2}\right\} = \frac{3}{2} \text{ laut Aufgabe 1.,}$$
$$\kappa_\infty(A) := \operatorname{cond}_\infty(A) := \operatorname{cond}_{\|\cdot\|_\infty}(A) = \|A\|_\infty \left\|A^{-1}\right\|_\infty = 2 \cdot \frac{3}{2} = 3,$$
$$\|\Delta A\|_\infty = \|0\|_\infty = 0,$$
$$\|\Delta b\|_\infty = \max_{i=1,2} |(\Delta b)_i| = \max\{\varepsilon,\ 0\} = \varepsilon.$$

Um die Fehlerabschätzung anwenden zu dürfen, benötigen wir, dass die Matrix A invertierbar ist, was oben gezeigt wurde (Determinante ist ungleich null) und die Bedingung $\left\|A^{-1}\right\|_\infty \|\Delta A\|_\infty < 1$ muss erfüllt sein, was wegen

$$\left\|A^{-1}\right\|_\infty \|\Delta A\|_\infty = \frac{3}{2} \cdot 0 = 0 < 1$$

der Fall ist. Demnach gilt laut der bewiesenen Fehlerabschätzung aus der Vorlesung für den relativen Fehler $\frac{\|\Delta x\|_\infty}{\|x\|_\infty}$ des gestörten linearen Gleichungssystems

$$A(x + \Delta x) = (A + \Delta A)(x + \Delta x) = b + \Delta b$$

gerade:

$$\begin{aligned} \frac{\|\Delta x\|_\infty}{\|x\|_\infty} &\leq \frac{\|I_2\|_\infty \cdot \kappa_\infty(A)}{1 - \kappa_\infty(A) \cdot \frac{\|\Delta A\|_\infty}{\|A\|_\infty}} \left(\frac{\|\Delta A\|_\infty}{\|A\|_\infty} + \frac{\|\Delta b\|_\infty}{\|b\|_\infty}\right) \\ &\leq \frac{3}{1 - \left\|A^{-1}\right\|_\infty \|\Delta A\|_\infty} \left(\frac{0}{2} + \frac{\varepsilon}{1}\right) = \frac{3}{1-0}\varepsilon = 3\varepsilon \leq 10^{-3} = \frac{1}{1000} \\ \Leftrightarrow\ 0 < \varepsilon &\leq \frac{1}{3 \cdot 1000} = \frac{1}{3000}. \end{aligned}$$

Demnach muss die Störung Δb in der Norm $\|\cdot\|_\infty$ kleiner als $\frac{1}{3000}$ gewählt werden um einen relativen Fehler von höchstens 10^{-3} zu erzeugen.
Für den absoluten Fehler $\|\Delta x\|_\infty$ haben wir die Abschätzung aus Vorlesung:

$$\begin{aligned} \|\Delta x\|_\infty &\leq \left\|A^{-1}\right\|_\infty \|\Delta b\|_\infty = \frac{3}{2}\varepsilon \leq 10^{-3} = \frac{1}{1000} \\ \Leftrightarrow\ 0 < \varepsilon &\leq \frac{2}{3 \cdot 1000} = \frac{1}{1500}. \end{aligned}$$

Demnach muss die Störung Δb in der Norm $\|\cdot\|_\infty$ kleiner als $\frac{1}{1500}$ gewählt werden um einen absoluten Fehler von höchstens 10^{-3} zu erzeugen.

□

(c) Wir berechnen alles, was wir benötigen:

$$\det(A) = \det\begin{pmatrix} 1 & 1 \\ 0 & 2 \end{pmatrix} = 1 \cdot 2 - 0 \cdot 1 = 2 - 0 = 2 \neq 0,$$
$$A^{-1} = \begin{pmatrix} 1 & 1 \\ 0 & 2 \end{pmatrix}^{-1} = \frac{1}{2}\begin{pmatrix} 2 & -1 \\ 0 & 1 \end{pmatrix} \text{ nach der Cramerschen Regel,}$$
$$\|b\|_1 = 1 + 1 = 2,$$
$$\|I_2\|_1 = 1 \text{ laut Aufgabe 2.(a),}$$
$$\|A\|_1 = \max_{j=1,2} (|A_{1j}| + |A_{2j}|) = \max\{1+0,\ 1+2\} = \max\{1,\ 3\} = 3 \text{ laut Aufgabe 1.,}$$
$$\left\|A^{-1}\right\|_1 = \max_{j=1,2} \left(\left|\left(A^{-1}\right)_{1j}\right| + \left|\left(A^{-1}\right)_{2j}\right|\right) = \max\left\{\frac{1}{2}(2+0),\ \frac{1}{2}(1+1)\right\} = \max\{1,\ 1\} = 1 \text{ laut Aufgabe 1.,}$$
$$\kappa_1(A) := \operatorname{cond}_1(A) := \operatorname{cond}_{\|\cdot\|_1}(A) = \|A\|_1 \left\|A^{-1}\right\|_1 = 3 \cdot 1 = 3,$$
$$\|\Delta A\|_1 = \max_{j=1,2} \left(\left|(\Delta A)_{1j}\right| + \left|(\Delta A)_{2j}\right|\right) = \max\left\{\frac{\varepsilon}{2}(1+1),\ \frac{\varepsilon}{2}(0+0)\right\} = \max\{\varepsilon,\ 0\} = \varepsilon \text{ laut Aufgabe 1.,}$$

$$\|\Delta b\|_1 = \|0\|_1 = 0.$$

Um die Fehlerabschätzung anwenden zu dürfen, benötigen wir, dass die Matrix A invertierbar ist, was oben gezeigt wurde (Determinante ist ungleich null) und die Bedingung $\left\|A^{-1}\right\|_1 \|\Delta A\|_1 < 1$ muss erfüllt sein, also muss gelten:

$$\left\|A^{-1}\right\|_1 \|\Delta A\|_1 = 1 \cdot \varepsilon = \varepsilon < 1.$$

Demnach gilt laut der bewiesenen Fehlerabschätzung aus der Vorlesung für den relativen Fehler $\frac{\|\Delta x\|_1}{\|x\|_1}$ des gestörten linearen Gleichungssystems

$$(A + \Delta A)(x + \Delta x) = b + \Delta b = b$$

gerade:

$$\begin{aligned}
\frac{\|\Delta x\|_1}{\|x\|_1} &\leq \frac{\|I_2\|_1 \cdot \kappa_1(A)}{1 - \kappa_1(A) \cdot \frac{\|\Delta A\|_1}{\|A\|_1}} \left(\frac{\|\Delta A\|_1}{\|A\|_1} + \frac{\|\Delta b\|_1}{\|b\|_1} \right) \\
&\leq \frac{3}{1 - \|A^{-1}\|_1 \|\Delta A\|_1} \left(\frac{\varepsilon}{3} + \frac{0}{2} \right) \\
&= \frac{3}{1-\varepsilon} \cdot \frac{\varepsilon}{3} = \frac{\varepsilon}{1-\varepsilon} \leq 10^{-3} = \frac{1}{1000} \\
\Leftrightarrow\ \varepsilon &\leq \frac{1-\varepsilon}{1000} = \frac{1}{1000} - \frac{\varepsilon}{1000} \\
\Leftrightarrow\ \frac{1001}{1000}\varepsilon &= \varepsilon \left(1 + \frac{1}{1000} \right) = \varepsilon + \frac{\varepsilon}{1000} \leq \frac{1}{1000} \\
\Leftrightarrow\ 0 < \varepsilon &\leq \frac{1}{1000} \cdot \frac{1000}{1001} = \frac{1}{1001} < 1
\end{aligned}$$

Demnach muss die Störung ΔA in der Norm $\|\cdot\|_1$ kleiner als $\frac{1}{1001}$ gewählt werden um einen relativen Fehler von höchstens 10^{-3} zu erzeugen. □