R. Herzog, G. Müller, M. Hashemi, S. Meggendorfer
Universität Heidelberg

Übung 02

Ausgabedatum: 29. April 2022
Abgabedatum: 10. Mai 2022

Hausaufgabe 1. (Absolute und relative Kondition) 4 Punkte

Bestimmen Sie die absoluten und relativen (partiellen) Konditionszahlen der folgenden Funktionen (in Abhängigkeit der Argumente) an den Punkten im Definitionsbereich, in denen die Funktionen differenzierbar sind. Erklären Sie, wo die Auswertung der Funktion (absolut bzw. relativ) am sensibelsten auf Abweichungen im Argument reagiert.

(*i*) $f\colon \mathbb{R}_{\geqslant 0} \to \mathbb{R},\ f(x) = \sqrt{x}$

(*ii*) $f\colon \mathbb{R} \to \mathbb{R},\ f(x) = \sin(x)$

(*iii*) $f\colon \mathbb{R}^n \to \mathbb{R},\ f(x) = \|x\|_2$

(*iv*) $f\colon \mathbb{R}^n \to \mathbb{R},\ f(x) = \|x\|_1$

Lösung.

(*i*) Die Funktion $f\colon \mathbb{R}_{\geqslant 0} \to \mathbb{R},\ f(x) = \sqrt{x}$ ist auf $\mathbb{R}^n \setminus \{0\}$ differenzierbar. Da wir $x = 0$ also nicht untersuchen müssen und f keine weiteren Nullstellen besitzt sind alle Konditionszahlen auf dem verbleibenden Bereich definiert.

Da die Funktion reelle Zahlen in reelle Zahlen abbildet gibt es nur je eine absolute und relative Konditionszahl. Da die Operatornormen hier mit den Beträgen zusammenfallen ergeben sich aus den Beträgen der partiellen absoluten bzw. relativen Konditionszahl die absolute bzw. relative Konditionszahl.

Wir haben also für $x \neq 0$

$$K_{11}(x) = f'(x) = \frac{1}{2\sqrt{x}}, \qquad K(x) = |K_{11}(x)| = K_{11}(x) = \frac{1}{2\sqrt{x}}$$
$$k_{11}(x) = K_{11}(x)\frac{x}{f(x)} = \frac{1}{2\sqrt{x}}\frac{x}{\sqrt{x}} = \frac{1}{2}, \qquad k(x) = |k_{11}(x)| = k_{11}(x) = \frac{1}{2}$$

Je kleiner der Wert des auszuwertenden Punkts absolut ist, desdo stärker werden also absolute Abweichungen sich absolut auf das Ergebnis auswirken (die Wurzel wird steiler bei 0).

Relative Abweichungen werden sich unabhängig von dem auszuwertenden Punkt auf relative Abweichungen im Ergebnis auswirken. Das liegt daran, dass relative Abweichungen für Punkte die eine schlechtere absolute Kondition haben, absolut kleiner sind – diese Effekte heben sich also auf. (1 Punkt)

(*ii*) Die Funktion $f\colon \mathbb{R} \to \mathbb{R}$, $f(x) = \sin(x)$ ist überall differenzierbar aber hat Nullstellen bei $x \in \{k\pi \mid k \in \mathbb{Z}\}$, hier und in $x = 0$ können wir also die relativen Konditionszahlen nicht mit unserer Definition untersuchen.

Da die Funktion reelle Zahlen in reelle Zahlen abbildet gibt es nur je eine absolute und relative Konditionszahl. Da die Operatornormen hier mit den Beträgen zusammenfallen ergeben sich aus den Beträgen der partiellen absoluten bzw. relativen Konditionszahl die absolute bzw. relative Konditionszahl.

Wir haben also

$$K_{11}(x) = f'(x) = \cos(x), \qquad K(x) = |K_{11}(x)| = |\cos(x)|,$$
$$k_{11}(x) = K_{11}(x)\frac{x}{f(x)} = x\frac{\cos(x)}{\sin(x)} = \frac{x}{\tan(x)}, \qquad k(x) = |k_{11}(x)| = \left|\frac{x}{\tan(x)}\right|, \quad x \notin \{k\pi \mid k \in \mathbb{Z}\}$$

Absolute Abweichungen wirken sich hier absolut am stärksten auf das Ergebnis aus, wenn der auszuwertende Punkt in der Nähe des Nulldurchgangs vom sin liegt, da der hier am steilsten ist.

(1 Punkt)

(*iii*) Die Funktion $f\colon \mathbb{R}^n \to \mathbb{R}$, $f(x) = \|x\|_2$ ist überall in $\mathbb{R}^n \setminus \{0\}$ differenzierbar. Da wir $x = 0$ also nicht untersuchen müssen und f keine weiteren Nullstellen besitzt (Definitheit der Norm) sind alle Konditionszahlen auf dem verbleibenden Bereich definiert.

Wir haben also für $x \neq 0$

$$(K_{1j}(x)) = f'(x) = \frac{1}{\|x\|_2}(x_1, \dots, x_n), \qquad K(x) = \|(K_{1j}(x))\|_{2\to 2} = \sqrt{\sum_{j=1}^{n} \frac{x_j^2}{\|x\|_2^2}} = 1,$$

$$(k_{1j}(x)) = \left(K_{1j}(x)\frac{x_j}{f(x)}\right) = \left(\frac{x_j^2}{\|x\|_2^2}\right), \qquad k(x) = \|(k_{1j}(x))\|_{\infty\to\infty} = \sum_{j=1}^{n} \frac{x_j^2}{\|x\|_2^2} = 1$$

Wir sehen, dass die absolute und relative Konditionszahlen unabhängig vom auszuwertenden Punkt den Wert 1 annehmen. Damit ist das Problem gut konditioniert. (1 Punkt)

(*iv*) Die Funktion $f\colon \mathbb{R}^n \to \mathbb{R}$, $f(x) = \|x\|_1$ ist immer dann in $x \in \mathbb{R}^n$ differenzierbar, wenn $x_j \neq 0$, $j = 1, \dots, n$. Dort besitzt sie auch keine Nullstellen (Definitheit der Norm), daher sind alle Konditionszahlen auf dem verbleibenden Bereich definiert.

Wir haben also für x mit $x_j \neq 0$, $j = 1, \dots, n$

$$(K_{1j}(x)) = f'(x) = (\operatorname{sgn}(x_j)), \qquad K(x) = \|(K_{1j}(x))\|_{2\to 2} = \sqrt{\sum_{j=1}^{n} \operatorname{sgn}^2(x_j)} = \sqrt{n},$$

$$(k_{1j}(x)) = \left(K_{1j}(x)\frac{x_j}{f(x)}\right) = \left(\frac{\operatorname{sgn}(x_j)x_j}{\|x\|_1}\right) = \frac{|x_j|}{\|x\|_1}, \qquad k(x) = \|(k_{1j}(x))\|_{\infty\to\infty} = \sum_{j=1}^{n} \frac{|x_j|}{\|x\|_1} = 1$$

Wie schon bei der letzten Teilaufgabe ist die Kondition hier unabhängig von dem auszuwertenden Punkt. Die auftretenden Konstanten kennen wir aus den Normäquivalenzaussagen. (1 Punkt)

Hausaufgabe 2. (Darstellungen der abs. Konditionszahl (siehe Bemerkung 3.5 im Skript) 6 Punkte

Es sei $F\colon \mathbb{R}^n \to \mathbb{R}^m$ und

$$LS(x) := \lim_{\varepsilon \searrow 0} \sup_{\substack{\|\Delta x\| < \varepsilon \\ \Delta x \neq 0}} \frac{\|F(x+\Delta x) - F(x)\|_2}{\|\Delta x\|_2} \in [0, \infty].$$

(*i*) Zeigen Sie:

(a) Ist F an x differenzierbar, dann ist $K(x) = LS(x)$.

(b) Ist F in einer Umgebung von x Lipschitz-stetig mit Konstante $L > 0$, dann ist $LS(x) \leqslant L$.

(ii) Untersuchen sie den Term $LS(x)$ an den Punkten, an denen die Abbildungen nicht differenzierbar sind.

(a) $f\colon \mathbb{R}_{\geqslant 0} \to \mathbb{R}$, $f(x) = \sqrt{x}$

(b) $f\colon \mathbb{R}^n \to \mathbb{R}$, $f(x) = \|x\|_2$

(c) $f\colon \mathbb{R}^n \to \mathbb{R}$, $f(x) = \|x\|_1$

Lösung.

(i) (a) Wir nehmen an, dass F an der Stelle x differenzierbar ist. Es sei $c > 0$ beliebig. Aufgrund der Definition der Differenzierbarkeit existiert ein $\varepsilon(c) > 0$ zu c, sodass

$$\|F(x+\Delta x) - F(x) - F'(x)\,\Delta x\|_2 \leqslant c\,\|\Delta x\|_2 \quad \text{für alle } \Delta x \text{ mit } \|\Delta x\| < \varepsilon(c).$$

Dabei ist $\|\cdot\|$ eine beliebige, aber fest gewählte Norm auf $\mathbb{R}^n$. Daraus folgt weiter mit der inversen Dreiecksungleichung ($|\|a\| - \|b\|| \leqslant \|a-b\|$), dass

$$|\|F(x+\Delta x) - F(x)\|_2 - \|F'(x)\,\Delta x\|_2| \leqslant c\,\|\Delta x\|_2$$

für alle Δx mit $\|\Delta x\| < \varepsilon(c)$. Teilt man durch $\|\Delta x\|$, dann erhält man

$$\left| \frac{\|F(x+\Delta x) - F(x)\|_2}{\|\Delta x\|_2} - \frac{\|F'(x)\,\Delta x\|_2}{\|\Delta x\|_2} \right| \leqslant c,$$

für alle Δx mit $\|\Delta x\| < \varepsilon(c)$ mit $\Delta x \neq 0$. Daraus folgern wir, dass die Differenz der termweise genommenen Suprema über alle $\Delta x \neq 0$ mit $\|\Delta x\| < \varepsilon$ für ein beliebiges $\varepsilon \in (0, \varepsilon(c))$ die gleiche Bedingung erfüllen, also

$$\left| \sup_{\substack{\|\Delta x\| < \varepsilon \\ \Delta x \neq 0}} \frac{\|F(x+\Delta x) - F(x)\|_2}{\|\Delta x\|_2} - \sup_{\substack{\|\Delta x\| < \varepsilon \\ \Delta x \neq 0}} \frac{\|F'(x)\,\Delta x\|_2}{\|\Delta x\|_2} \right| \leqslant c, \tag{0.1}$$

für beliebiges $\varepsilon \in (0, \varepsilon(c))$. (“Wenn zwei Funktionen innerhalb eines c-Schlauches voneinander liegen, können ihre Suprema nicht weiter auseinander liegen als c.”) Der zweite Term in (0.1) entspricht hierbeit der Operatornorm. Es folgt also, dass für das beliebige $c > 0$ ein $\varepsilon(c) > 0$ existiert, so dass

$$\left| \sup_{\substack{\|\Delta x\| < \varepsilon \\ \Delta x \neq 0}} \frac{\|F(x+\Delta x) - F(x)\|_2}{\|\Delta x\|_2} - \|F'(x)\|_{2\to 2} \right| \leqslant c$$

für alle $\varepsilon \in (0, \varepsilon(c))$. Das zeigt die Aussage. (3 Punkte)

(b) Ist F in einer Umgebung $B_\varepsilon^{\|\cdot\|}(x)$ von x Lipschitz stetig mit Konstante $L > 0$, dann ist

$$LS(x) := \lim_{\varepsilon \searrow 0} \sup_{\substack{\|\Delta x\| < \varepsilon \\ \Delta x \neq 0}} \frac{\|F(x+\Delta x) - F(x)\|_2}{\|\Delta x\|_2} \leqslant \lim_{\varepsilon \searrow 0} \sup_{\substack{\|\Delta x\| < \varepsilon \\ \Delta x \neq 0}} \frac{L\|\Delta x\|_2}{\|\Delta x\|_2} = L$$

und die "Konditionszahl" damit durch die (lokale) Lipschitzkonstante beschränkt. (1 Punkt)

(*ii*) (a) Die Wurzelfunktion ist, wie wir wissen, nur in $x = 0$ nicht differenzierbar und auch nicht lokal um 0 Lipschitz stetig. Für den lim sup-Term gilt:

$$LS(x) = \lim_{\varepsilon \searrow 0} \sup_{0 < \Delta x < \varepsilon} \frac{|\sqrt{\Delta x}|}{|\Delta x|} = \lim_{\varepsilon \searrow 0} \underbrace{\sup_{0 < \Delta x < \varepsilon} \frac{1}{|\sqrt{\Delta x}|}}_{=\infty} = \infty$$

Dass dieser unbeschränkt ist ist natürlich nur dadurch möglich, dass die Wurzelfunktion in $x = 0$ nicht Lipschitz stetig ist. (1 Punkt)

(b) Die 2-Norm ist nur in $x = 0$ nicht differenzierbar. Allerdings ist jede Norm auf Grund der inversen Dreiecksungleichung (bezüglich sich selbst) Lipschitz-stetig mit Konstante 1, damit wissen wir schonmal, dass $LS(x)$ in diesem Fall durch 1 beschränkt ist. Tatsächlich ist

$$LS(x) = \lim_{\varepsilon \searrow 0} \sup_{\substack{\|\Delta x\| < \varepsilon \\ \Delta x \neq 0}} \frac{\|\Delta x\|_2}{\|\Delta x\|_2} = 1$$

(c) Die 1-Norm ist immer dann in $x \in \mathbb{R}^n$ nicht differenzierbar, wenn ein $i \in \{1, \dots, n\}$ mit $x_i = 0$ existiert. Allerdings ist jede Norm auf Grund der inversen Dreiecksungleichung (bezüglich sich selbst) Lipschitz-stetig mit Konstante 1. Auf Grund der Konstanten in den Normäquivalenzabschätzungen in (2.1) des Skripts wissen wir also, dass die 1-Norm bezüglich der 2-Norm $\sqrt{n}$-Lipschitz-stetig ist. Damit wissen wir schonmal, dass $LS(x)$ in diesem Fall durch $\sqrt{n}$ beschränkt ist. Tatsächlich ist

$$LS(x) = \lim_{\varepsilon \searrow 0} \sup_{\substack{\|\Delta x\| < \varepsilon \\ \Delta x \neq 0}} \frac{\|\Delta x\|_1}{\|\Delta x\|_2} = \sqrt{n},$$

denn $\|x\|_1 = \sqrt{n}\|x\|_2$ entlang der Raumdiagonalen, also z. B. immer wenn $x = \frac{\varepsilon}{2}\mathbf{1}/\|\mathbf{1}\|$. (1 Punkt)

Hausaufgabe 3.

(Kondition und lineare Gleichungssysteme) 5 Punkte

Die Lage eines ebenen Objekts im dreidimensionalen Raum mit Raumkoordinaten $(x, y, z)^\mathsf{T} \in \mathbb{R}^3$ soll über drei Distanzsensoren bestimmt werden. Die Sensoren sind an den Stellen

$$p_1 = (0.9, 0, 0)^\mathsf{T}, \qquad p_2 = (1, 1, 0)^\mathsf{T} \qquad \text{und} \qquad p_3 = (1, -0.5, 0)^\mathsf{T}$$

angebracht und messen den vertikalen Abstand des Objekts zu den Sensoren. Die Sensoren liefert also drei Punkte mit den Koordinaten $(0.9, 0, z_1)^\mathsf{T}$, $(1, 1, z_2)^\mathsf{T}$ und $(1, -0.5, z_3)^\mathsf{T}$, die bekanntermaßen in der Ebene liegen. Aus diesen Messungen soll die Lage des Objekts mit Hilfe der Parameter $a = (a_1, a_2, a_3)^\mathsf{T} \in \mathbb{R}^3$ über die Ebenengleichung

$$a_1 x + a_2 y + a_3 = z$$

rekonstruiert und beschrieben werden.

Für das Objekt, das wir in diesem Beispiel untersuchen wollen, würden die Sensoren ohne Messfehler die Messwerte $z_{\text{Mess}} = (3.9, 6, 3)$ liefern.

(*i*) Bestimmen Sie die Ebenenparameter a_{Mess} zu den Messwerten z_{Mess} exakt.

(*ii*) Bestimmen Sie den größten relativen Fehler $\|\Delta a\|_2 / \|a_{\text{mess}}\|_2$ in den Ebenenparametern, der bei einer relativen Messungenauigkeit des Sensors an p_1 von höchstens 10% auftreten kann.

(*iii*) Wenn Sie den ersten Sensor von p_1 nach $(0, 0, 0)^\mathsf{T}$ verschieben, müsste dieser Sensor Ihnen für das gleiche Objekt den Wert $z_1 = 3$ liefern. Zeigen Sie ohne Berechnung des absoluten Fehlers $\|\Delta a\|_2$, dass diese Verschiebung gegenüber der Situation in Aussage (*ii*) zu einem besseren relativen Fehler in den Ebenenparametern führt, wenn die Messungenauigkeit des verschobenen Sensors weiterhin höchstens 10% beträgt.

(*iv*) Zeigen Sie, dass die auftretende Matrix der Konfiguration aus Aussage (*iii*) besser konditioniert ist als die Matrix der Ursprungskonfiguration. Erklären Sie geometrisch, warum das so ist.

Lösung.

Kurze Erklärung zur Ebenengleichung: Eine allgemeine Beschreibung einer (zweidimensionalen Hyper-)Ebene $E \subseteq \mathbb{R}^3$ hat die Form

$$E = \left\{ \begin{pmatrix} x \\ y \\ z \end{pmatrix} \in \mathbb{R}^3 \;\middle|\; \hat{a}^\mathsf{T} \begin{pmatrix} x \\ y \\ z \end{pmatrix} = \hat{b} \right\}$$

für einen Normalenvektor $\hat{a} \in \mathbb{R}^3$, der auf die Ebene rechtwinklig steht und sie damit orientiert, und einen "offset" $\hat{b} \in \mathbb{R}$, der bestimmt, wie weit die Ebene aus dem Urpsrung ausgerückt wird.

Unsere Aufgabe ist nicht lösbar, wenn die Ebene vertikal ausgerichtet ist, also genau dann, wenn $\hat{a}_3 = 0$ ist. Wir gehen also von jetzt an davon aus, dass $\hat{a}_3 \neq 0$. Wir können also die Ebenengleichung durch $-\hat{a}_3$ teilen und die Koeffizienten umdefinieren um die oben erwähnte Form der Ebenengleichung

$$\underbrace{-\frac{\hat{a}_1}{\hat{a}_3}}_{a_1 :=} x \underbrace{-\frac{\hat{a}_2}{\hat{a}_3}}_{a_2 :=} y - z = \underbrace{\frac{\hat{b}}{\hat{a}_3}}_{-a_3 :=} .$$

zu erhalten. Diese wählen wir, weil hier bei der Rekonstruktion der Ebenenparameter die Messwerte z direkt als rechte Seite eines linearen Gleichungssystems auftauchen.

(i) Die Ebenenparameter erhalten wir direkt aus dem linearen Gleichungssystem, dass sich aus den Messwerttripeln und den Ebenenparametern zu

$$\begin{pmatrix} 0.9 & 0 & 1 \\ 1 & 1 & 1 \\ 1 & -0.5 & 1 \end{pmatrix} \begin{pmatrix} a_1 \\ a_2 \\ a_3 \end{pmatrix} = \begin{pmatrix} 3.9 \\ 6 \\ 3 \end{pmatrix}$$

ergibt. Die Matrix ist invertierbar mit der Inversen

$$\begin{pmatrix} 0.9 & 0 & 1 \\ 1 & 1 & 1 \\ 1 & -0.5 & 1 \end{pmatrix}^{-1} = \begin{pmatrix} -10 & \frac{10}{3} & \frac{20}{3} \\ 0 & \frac{2}{3} & -\frac{2}{3} \\ 10 & -3 & -6 \end{pmatrix}$$

und die eindeutige Lösung ergibt sich zu $a = (1, 2, 3)^\mathsf{T}$. (1 Punkt)

(ii) Eine Störung $\Delta z = (\alpha, 0, 0)^\mathsf{T}$ mit $\alpha \in [-0.39, 0.39]$ in den Messungen liefert in den Ebenenparametern die Abweichung

$$\Delta a = \begin{pmatrix} 0.9 & 0 & 1 \\ 1 & 1 & 1 \\ 1 & -0.5 & 1 \end{pmatrix}^{-1} \begin{pmatrix} \alpha \\ 0 \\ 0 \end{pmatrix} = \begin{pmatrix} -10 & \frac{10}{3} & \frac{20}{3} \\ 0 & \frac{2}{3} & -\frac{2}{3} \\ 10 & -3 & -6 \end{pmatrix} \begin{pmatrix} \alpha \\ 0 \\ 0 \end{pmatrix} = \alpha \begin{pmatrix} -10 \\ 0 \\ 10 \end{pmatrix}$$

und damit den relativen Fehler

$$\frac{\|\Delta a\|_2}{\|a\|_2} = \alpha \sqrt{\frac{200}{14}} = \alpha \sqrt{\frac{100}{7}}$$

was für $\alpha = 0.39$ im Rahmen der 10% Messgenauigkeit des Sensors maximal wird. Der Wert ist dann

$$\frac{\|\Delta a\|_2}{\|a\|_2} = \frac{39}{\sqrt{700}} \approx 1.474$$

(2 Punkte)

(*iii*) Durch den verschobenen Sensor ändert sich die Matrix im linearen Gleichungssystem, das jetzt

$$\begin{pmatrix} 0 & 0 & 1 \\ 1 & 1 & 1 \\ 1 & -0.5 & 1 \end{pmatrix} \begin{pmatrix} a_1 \\ a_2 \\ a_3 \end{pmatrix} = \begin{pmatrix} 3 \\ 6 \\ 3 \end{pmatrix}$$

ist und natürlich weiterhin durch $a = (1, 2, 3)^\intercal$ gelöst wird.

Mit der Abschätzung für relative Fehler beim Lösen linearer Gleichungssysteme (Gleichung (3.21) im Skript) können wir den relativen Fehler ohne Berechnung der absoluten Abweichung über die Matrixkondition abschätzen.

Mit den Abschätzungen zur Normäquivalenz aus Gleichung (2.1) im Skript erhalten wir, dass die Kondition (bzgl. der $2 \to 2$-Operatornorm) der Matrix A durch

$$\kappa(A) := \|A\|_{2\to 2} \, \|A^{-1}\|_{2\to 2} \leq 3\|A\|_{\infty\to\infty} \, \|A^{-1}\|_{\infty\to\infty} = 3 * 3 * 2 = 18$$

beschränkt ist.

Mit Gleichung (3.21) können wir nun abschätzen, dass für Störungen $\Delta z = (\alpha, 0, 0)$ mit $\alpha \in (-0.3, 0.3)$ die verschobene Konfiguration die relative Fehlerabschätzung

$$\frac{\|\Delta a\|_2}{\|a\|_2} \leqslant \kappa(A) \frac{\|\Delta z\|_2}{\|z\|_2} \leqslant 18 * \alpha \frac{1}{\sqrt{54}} \leqslant 0.3 * \frac{18}{\sqrt{54}} \approx 0.7348 < 1.474$$

erfüllt und damit garantiert geeigneter für unser Objekt ist.

(*iv*) Dass die verschobene Konfiguration auch allgemein weniger starke relative Fehler produziert ist intuitiv einleuchtend, denn die Punkte in der Urprungskonfiguration liegen fast auf einer Linie. Kleine (relative) Fehler in den Messwerten führen also zu großen Änderungen in der Orientierung der Ebene. Setzt man einen Sensor weiter von der Verbindungslinie der anderen beiden Punkte weg und produziert so ein fast gleichseitiges Dreieck, dann führt eine kleine Abweichung in den Messdaten zu weniger Änderung in der Orientierung der Ebene. Dieser Effekt ist ganz ähnlich zu dem, dass ein Tisch auf 3 weit auseinanderliegenden Beinen stabiler steht als auf 3 Beinen, die fast auf einer Linie am Tisch befestigt sind.

Mathematisch sieht man das sofort an der Kondition der Matrizen der jeweiligen Konfigurationen.

So ist

$$\left\|\begin{pmatrix}0 & 0 & 1\\ 1 & 1 & 1\\ 1 & -0.5 & 1\end{pmatrix}\right\|_{\infty\to\infty} = 3, \qquad \left\|\begin{pmatrix}-1 & \frac{1}{3} & \frac{2}{3}\\ 0 & \frac{2}{3} & -\frac{2}{3}\\ 1 & 0 & 0\end{pmatrix}\right\|_{\infty\to\infty} = 2$$

$$\left\|\begin{pmatrix}0.9 & 0 & 1\\ 1 & 1 & 1\\ 1 & -0.5 & 1\end{pmatrix}\right\|_{\infty\to\infty} = 3, \qquad \left\|\begin{pmatrix}-10 & \frac{10}{3} & \frac{20}{3}\\ 0 & \frac{2}{3} & -\frac{2}{3}\\ 10 & -3 & -6\end{pmatrix}\right\|_{\infty\to\infty} = 20$$

und damit

$$\left\|\begin{pmatrix}0 & 0 & 1\\ 1 & 1 & 1\\ 1 & -0.5 & 1\end{pmatrix}\right\|_{\infty\to\infty} \left\|\begin{pmatrix}-1 & \frac{1}{3} & \frac{2}{3}\\ 0 & \frac{2}{3} & -\frac{2}{3}\\ 1 & 0 & 0\end{pmatrix}\right\|_{\infty\to\infty} = 6$$

$$\left\|\begin{pmatrix}0.9 & 0 & 1\\ 1 & 1 & 1\\ 1 & -0.5 & 1\end{pmatrix}\right\|_{\infty\to\infty} \left\|\begin{pmatrix}-10 & \frac{10}{3} & \frac{20}{3}\\ 0 & \frac{2}{3} & -\frac{2}{3}\\ 10 & -3 & -6\end{pmatrix}\right\|_{\infty\to\infty} = 60$$

Beachte: Diese Ausdrücke sind die Konditionszahlen der Matrizen zu den jeweiligen Konfiguration bzgl. der $\infty \to \infty$ -Operatornorm. Wieder mit den Konstanten der Normabschätzungen erhält man damit, dass daher

$$\kappa\left(\begin{pmatrix}0 & 0 & 1\\ 1 & 1 & 1\\ 1 & -0.5 & 1\end{pmatrix}\right) \leq \frac{9}{10}\kappa\left(\begin{pmatrix}0.9 & 0 & 1\\ 1 & 1 & 1\\ 1 & -0.5 & 1\end{pmatrix}\right)$$

und damit die bessere Kondition der verschobenen Konfiguration.

(2 Punkte)

Hausaufgabe 4. (Kondition zusammengesetzter Funktionen) 3 Punkte

Gegeben seien die Funktionen

$$g\colon \mathbb{R} \to \mathbb{R}^2 \qquad \text{und} \qquad h\colon \mathbb{R}^2 \to \mathbb{R}$$
$$g(x) = \begin{pmatrix}e^x\\ 1\end{pmatrix} \qquad \text{und} \qquad h(x) = e^{x_2}.$$

Bestimmen Sie die absoluten und die relativen Konditionszahlen der Funktionen g und h und der zusammengesetzten Funktion $f := (h \circ g)$. Erklären Sie, warum die Abschätzungen aus Gleichungen (3.25) und (3.27) im Skript für dieses Beispiel zu einer ungenauen Fehleranalyse führen können.

Lösung.

Die relative Kondition können wir nach unserer Definition nur für Nicht-Nullstellen der Funktionen $f, g, h \neq 0$ und für Nicht-Nullargumente bestimmen. Für das gegebene Beispiel schränken wir uns also auf Nicht-Nullargumente ein. Dann sind

$$
\begin{aligned}
&g\colon \mathbb{R} \to \mathbb{R}^2, && h\colon \mathbb{R}^2 \to \mathbb{R} && \text{und} \quad f\colon \mathbb{R} \to \mathbb{R},\\
&g(x) = \begin{pmatrix} e^x \\ 1 \end{pmatrix}, && h(x) = e^{x_2} && \text{und} \quad f(x) = e,\\
&g'(x) = (K^g_{ij}(x)) = \begin{pmatrix} e^x \\ 0 \end{pmatrix}, && h'(x) = (K^h_{ij}(x)) = (0, e^{x_2}) && \text{und} \quad f'(x) = K^f(x) = 0,\\
&(k^g_{i1}(x)) = \begin{pmatrix} xK^g_{i1}(x) \\ \hline g_i(x) \end{pmatrix} = \begin{pmatrix} x \\ 0 \end{pmatrix}, && (k^h_{1j}(x)) = \left(K^h_{1j}(x) \frac{x_j}{h(x)} \right) = (0, x_2) && \text{und} \quad (k^f_{ij}(x)) = 0,\\
&K^g(x) = \|K^h_{i1}(x)\|_{2\to 2} = e^x, && K^h(x) = \|K^h_{1j}(x)\|_{2\to 2} = e^{x_2} && \text{und} \quad K^f(x) = 0,\\
&k^g(x) = \|k^h_{i1}(x)\|_{\infty\to\infty} = |x|, && k^h(x) = \|k^h_{1j}(x)\|_{2\to 2} = |x_2| && \text{und} \quad k^f(x) = 0.
\end{aligned}
$$

Wie wir sehen hat die Funktion g nur einen Eingang und einen sensiblen (nichtkonstanten) und einen unsensiblen (konstanten) Ausgang. Die Funktion h hat einen sensiblen und einen unsensiblen Eingang und nur einen Ausgang. Da die Hintereinanderausführung von h und g die beiden unsensiblen (konstanten) Eingänge miteinander verknüpft ergibt sich die konstante Funktion f, die gegenüber Störungen in den Argumenten natürlich vollständig unsensibel reagiert - Fehler setzen sich nicht in die Ergebnisse fort, der absolute wie der relative Fehler im Ergebnis für f sind also immer 0.

Die Abschätzungen aus Gleichungen (3.25) und (3.27) im Skript sehen diesen Zusammenhang nicht. Wie im Skript erwähnt handelt es sich hier um worst-case Abschätzungen, die davon ausgehen müssen, dass evtl. die beiden sensiblen Ein- bzw. Ausgänge miteinander verknüpft werden und die zusammengesetzte Funktion f für große x eine starke Fehlerverstärkung liefert. Die Abschätzung liefert hier also eine absolute bzw. relative Fehlerskalierung mit (insbesondere für große x) deutlicher Überschätzung von

$$0 = K^f(x) \leqslant K^h(g(x))K^g(x) = e^{1+x} \qquad \text{und} \qquad 0 = k^f(x) \leqslant k^h(g(x))k^g(x) = |x|.$$

(3 Punkte)

Für die Abgabe Ihrer Lösungen zu diesem Übungsblatt verwenden Sie bitte die dafür vorgesehene Abgabefunktion in Moodle.