Bemerkung 2.7

Lineare Abb. zwischen endlich-dim. Vektorräumen können durch Matrizen beschrieben.

Ziel: Charakterisiere das Abb. verhalten von $\mathcal{L}$

Def 2.8 (Operatornorm)

$$\mathcal{L} : (X, \| \ \|_X) \longrightarrow (Y, \| \ \|_Y)$$

$$\|\mathcal{L}\| := \sup_{\|x\|_X = 1} \|\mathcal{L}(x)\|_Y$$

$$= \sup_{x \neq 0} \frac{\| \mathcal{L}(x) \|_Y}{\| x \|_X} =$$

$$\sup_{\|x\|_X \leq 1} \| \mathcal{L}(x) \|_Y$$

"Maß" für die Verformung der "Einheitskugel unter der Abb. $\mathcal{L}$.

Bemerkung 2.9

Ein linearer Operator ist beschränkt genau dann wenn stetig.

$$\|\mathcal{L}(x_2) - \mathcal{L}(x_1)\|_Y =$$

$$\|\mathcal{L}(x_2 - x_1)\|_Y \leq \|\mathcal{L}\| \cdot$$

$$\|x_2 - x_1\|_X \quad \text{(Lipschitz-stetig)}$$

$\mathcal{L}$ beschränkt, falls $\|\mathcal{L}\|$ endlich

Für $X = \mathbb{R}^n$, $Y = \mathbb{R}^m$ und die $\mathcal{L}$-repräsentierende Matrix $B \in \mathbb{R}^{m \times n}$ berechnet man:

$$\|B\|_\infty = \max_{i=1,\dots,m} \sum_{k=1}^{n} |b_{ik}|$$

(Zeilensummennorm)

$$\|B\|_1 = \max_{i=1,\dots,n} \sum_{k=1}^{m} |b_{ki}|$$

(Spaltensummennorm)

$A \in \mathbb{R}^{n \times n}$

$$\|A\|_2 = \sqrt{\lambda_{max}(A^T A)}$$

(Spektralnorm)

<u>Relative Kondition:</u>

Ziel: schätze $\dfrac{\|f(\tilde{x}) - f(x)\|_Y}{\|f(x)\|_Y}$

durch $\dfrac{\|\tilde{x} - x\|}{\|x\|_X}$

Satz 2.10 (rel. Konditionszahl)

Für injektives $\mathcal{L}$ gilt

$$\frac{\|\mathcal{L}(\bar{x}) - \mathcal{L}(x)\|_Y}{\|\mathcal{L}(x)\|_Y} \leq \kappa(\mathcal{L}) \cdot \frac{\|\bar{x} - x\|_X}{\|x\|_X}$$

mit $\kappa(\mathcal{L}) = \dfrac{\sup\limits_{\|x\|_X = 1} \|\mathcal{L}(x)\|_Y}{\inf\limits_{\|x\|_X = 1} \|\mathcal{L}(x)\|_Y}$

Ist $\mathcal{L}$ bijektiv gilt

$$\kappa(\mathcal{L}) = \|\mathcal{L}\| \cdot \|\mathcal{L}^{-1}\|$$

<u>Beispiel 2.11</u> (Kondition einer Basis)

Sei V n-dimensionaler Vektorraum und $\underline{\Phi} = \{\phi_1, \dots, \phi_n\}$ eine Basis von V. (z.B. Polynomraum)

$$\mathcal{L} : \mathbb{R}^n \longrightarrow V, \quad \mathcal{L}(a) := \sum_{j=1}^{n} a_j \phi_j$$

Da Φ Basis ist Ψ bijektiv.

$$\kappa(\Phi) := \kappa(\Psi)$$

(Konditionszahl der Basis)

<u>Bemerkung</u> <u>2.12</u>

a) $\kappa(\Psi)$ hängt von der Wahl der Normen ab.

b) Falls Ψ beschränkt, ist $\kappa(\Psi)$ auch für injektives (und nicht surjektives Ψ) definiert.

da $\inf\limits_{\|x\|_X = 1} \|\mathcal{L}(x)\|_Y \neq 0$, denn

$\mathcal{L}(x) = 0$ gilt <u>nur</u> für $x = 0$.

<u>Anschauliche Bedeutung</u>:

$\mathcal{K}(\mathcal{L})$: Verhältnis max. Dehnung zu max. Stauchung der Einheitskugel unter der Abb. $\mathcal{L}$ gemessen in der Bildnorm $\| \; \|_Y$

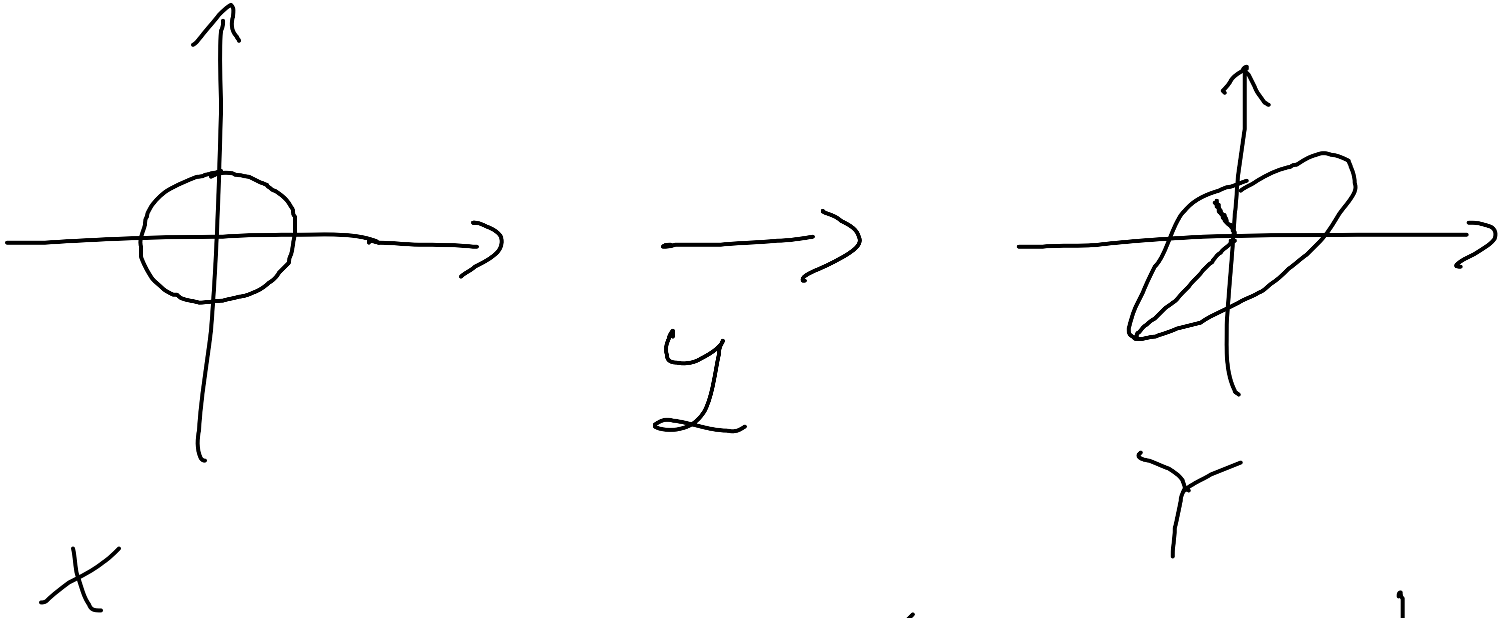

Abb. 2.13 (Kondition)

Beispiel 2.14 (Matrixkondition)

$A \in \mathbb{R}^{n \times n}$ invertierbar

$$\kappa_p(A) := \|A\|_p \, \|A^{-1}\|_p \quad (2.14)$$

$p = 1, \ldots, \infty$

Schnittpunkt von Geraden in der Ebene (s. Bsp. 2.1 b)

$$\underbrace{\begin{pmatrix} 3 & 1.001 \\ 6 & 1.997 \end{pmatrix}}_{A} \underbrace{\begin{pmatrix} x_1 \\ x_2 \end{pmatrix}}_{x} = \underbrace{\begin{pmatrix} 1,999 \\ 4,003 \end{pmatrix}}_{b}$$

Lsg. ist $x = \begin{pmatrix} 1 \\ -1 \end{pmatrix}$. Betrachte

„Datenstörung" in b:

$$\bar{b} := \begin{pmatrix} 2.002 \\ 4 \end{pmatrix} \longrightarrow \bar{x} := A^{-1} \bar{b}$$

Man berechnet

$$A^{-1} = -\frac{1}{0.015} \begin{pmatrix} 1.997 & -1.001 \\ -6 & 3 \end{pmatrix}$$

und $\bar{x} = \begin{pmatrix} 0.4004 \\ 0.8 \end{pmatrix}$

in der $\| \ \|_\infty$ - Norm gilt:

$$\frac{\|\tilde{b} - b\|_\infty}{\|b\|_\infty} \approx 7.5 \cdot 10^{-4}$$

(Datenstörung)

$$\frac{\|\tilde{x} - x\|_\infty}{\|x\|_\infty} = 1.8$$

(Resultatstörung)

sowie $\kappa(A) = \|A\|_\infty \cdot \|A^{-1}\|_\infty = 4738.2$

b) Ist $X = \mathbb{R}^n$, $Y = \mathbb{R}$ und ein nichtlineares $f : X \to Y$, das differenzierbar ist, gilt:

$$\left| \frac{f(\bar{x}) - f(x)}{f(x)} \right| \leq \kappa_\infty(x) \cdot \underbrace{\sum_{j=1}^{n} \left| \frac{\bar{x}_j - x_j}{x_j} \right|}_{\left\| \frac{\bar{x} - x}{x} \right\|_1}$$

$$\kappa_\infty(x) = \max_{j=1,\dots,n} \left| \frac{\partial f}{\partial x_j} \cdot \frac{x_j}{f(x_j)} \right|$$

(das sieht man mit Hilfe von Taylorentwicklung 1. Ordnung)

Kondition der Addition:

$$f: \mathbb{R}^2 \longrightarrow \mathbb{R} \qquad x = \begin{pmatrix} x_1 \\ x_2 \end{pmatrix} \longmapsto f(x) := x_1 + x_2$$

$$\frac{\partial f}{\partial x_j} \cdot \frac{\partial x_j}{f(x)} = \frac{x_j}{x_1 + x_2}, \quad j = 1, 2$$

$$\kappa_{\infty}(x) = \max\left\{ \left| \frac{x_1}{x_1+x_2} \right|, \left| \frac{x_2}{x_1+x_2} \right| \right\} \qquad (2.15)$$

Haben x_1, x_2 gleiches Vorzeichen:

$$\kappa_{\infty}(x) \leq 1$$

Falls $x_1 \approx -x_2$, so kann

$\kappa_{\infty}(x)$ sehr groß werden!

(s. Auslöschung in Kap. 3)

Kap. 3 Numerische Zahldarstellung und Gleitpunktarithmetik

Darstellung einer reellen Zahl

$$x = \pm \left(\sum_{j=1}^{\infty} d_j \, b^{-j} \right) \cdot b^{e} \quad (3.1)$$

wähle e so, dass $d_1 \neq 0$

Rechnerdarstellung ist endlich
(Maschinenzahlen)

$\longrightarrow$ normalisierte Gleitpunktdarstellung

$$x = f \cdot b^{e}$$

b : Grundzahl (Basis) des gewählten Zahlensystems (z.B. dezimal, binär)

$r \leq e \leq R$ (Exponent)

$f : \pm 0.d_1 d_2 \ldots d_m$

$d_j \in \{0, 1, \ldots, b-1\}$

f : Mantisse

m : Mantissenlänge

Rechner verarbeitet reelle Zahlen über Reduktionsabbildung

$$fl : \mathbb{R} \to M(b, m, r, R)$$

(Maschinenzahlen)

Rundung:

$$fl(x) := \pm \begin{cases} \left(\sum_{j=1}^{m} d_j b^{j}\right) \cdot b^e & d_{m+1} < \frac{b}{2} \\ \left(\sum_{j=1}^{m} d_j b^{j} + b^{-m}\right) \cdot b^e & d_{m+1} \geq \frac{b}{2} \end{cases}$$

$\to$ kleinstmögliche Zahl

$$x_{min} = 0.1000\ldots \cdot b^r = b^{r-1}$$

größtmögliche Zahl

$$x_{max} = 0.aa\ldots a \cdot b^R = (1 - b^{-m}) b^R$$

$$a = b - 1$$

$$|x| < x_{min} \to fl(x) = 0$$

$$|x| > x_{max} \to fl(x) = \infty \text{ (overflow)}$$

Rundungsfehler

$$|fl(x) - x| \leq \frac{b^{-m}}{2} b^e \quad \text{(absolut)}$$

$$(3.2) \quad \left| \frac{fl(x) - x}{x} \right| \leq \frac{\frac{b^{-m}}{2} \cdot b^e}{b^{-1} \cdot b^e} = \frac{b^{1-m}}{2}$$

$$eps := \frac{b^{1-m}}{2} \quad \text{(Maschinengenauigkeit) (relativ)}$$

$$\mathrm{eps} = \inf \{ \delta > 0 \mid fl(1+\delta) > 1 \}$$

für $|\varepsilon| \leq \mathrm{eps}$ gilt

$$\boxed{fl(x) = x(1+\varepsilon)} \qquad (3.3)$$

§ 3.1 Gleitpunktarithmetik

Algorithmus: „Folge arithmetischer Operationen". Verknüpfung von Maschinenzahlen liefert <u>nicht</u> notwendig wieder eine Maschinenzahl.

Beispiel: $b = 10$, $m = 3$

$$0.346 \cdot 10^2 + 0.785 \cdot 10^2 =$$

$$0.1131 \cdot 10^3 \neq 0.113 \cdot 10^3$$

Ersetze die üblichen arithmet. Operationen durch Gleitpunktoperationen (Pseudoarithmetik). $\circledcirc\!\!\!\nabla$

$$\nabla \in \{+, -, \times, \div\}.$$

Führe über Mantissenlänge hinaus weitere Stellen mit, genaue Rechnung nach Exponentenausgleich, Normalisierung, Rundung, so dass

$$x \circledtriangledown y = fl(x \triangledown y) \quad (3.4)$$

wg. (3.3) Annahme:

$$x \circledtriangledown y = (x \triangledown y)(1 + \varepsilon)$$

Eigenschaften:

- (3.4) nicht mehr gültig für eine Sequenz von Operationen

- Assoziativität und Distributivität gehen verloren

<u>Bsp.</u> $b = 10,\ m = 3$

Maschinenzahlen $x = 0.659 \cdot 10^4$

$y = 0.100 \cdot 10^1$

$z = 0.400 \cdot 10^1$

$(\in M(b = 10,\ m = 3))$

$(x + y) + z = (y + z) + x = 6595$ (exakt)

$x \oplus y = 0.659 \cdot 10^4$

$(x \oplus y) \oplus z = 0.659 \cdot 10^4$

aber:

$$y \oplus z = 0.500 \cdot 10^1$$

$$(y \oplus z) \oplus x = 0.660 \cdot 10^4 = fl(x+y+z)$$

(3.5): „lokaler" Fehler einer „Operation ist im Rahmen von eps

Frage: Fehlerpropagation im Algorithmus
($\to$ Stabilität des Algorithmus)

Algorithmus heißt stabil, falls die Fehler in der „Größenordnung" der unvermeidbaren Fehler liegen (→ Kondition des Problems)

<u>Beispiel 3.6</u> (Auslöschung) $b = 10$, $m = 3$

$x = 0.73563 \quad y = 0.73441 \quad eps = \frac{1}{2} \cdot 10^{-2}$

$x - y = 0.00122$

$\bar{x} = fl(x) = 0.736 \quad \bar{y} = fl(y) = 0.734$

$$|\delta_x| = 0.500 \cdot 10^{-3}$$

$$|\delta_y| = 0.560 \cdot 10^{-3}$$

$$\text{Fehler} \quad \left| \frac{(\bar{x}-\bar{y}) - (x-y)}{x-y} \right|$$

$$= \left| \frac{0.002 - 0.00122}{0.00122} \right| = 0.64$$

$$\gg \delta_x, \delta_y$$

Stimmen die führenden t Ziffern bei Subtraktion zweier Zahlen überein, steigt der Fehler im Resultat mit dem Faktor b^t

$\Rightarrow$ Subtraktion „ähnlich" großer Zahlen ist <u>instabil</u>!