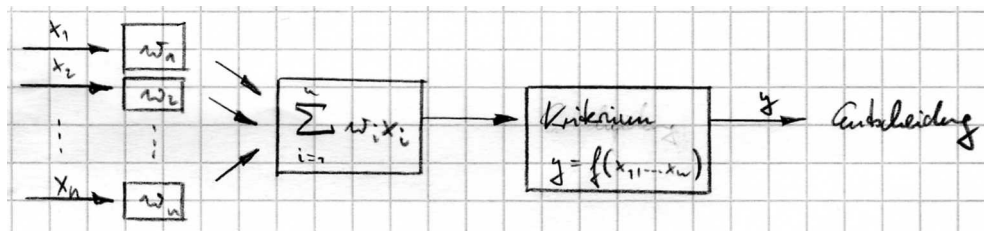


Neuronales Netz (Ein-Schicht-Modell, linear)



Verhalten des Netzes bestimmt durch *Gewichte* w_1, \dots, w_n

Training des Netzes Wähle w_1, \dots, w_n so, dass eine große Zahl von Tests mit vorgegebenen Eingangsdaten $(x_1^{(j)}, \dots, x_n^{(j)})^\top$ und bekannten Resultaten $y^{(j)}$ möglichst gut wiedergegeben wird:

$$\sum_{i=1}^n x_i^{(j)} w_i \approx y^{(j)}, \quad (j = 1, \dots, m)$$

↪ überbestimmtes lineares Gleichungssystem, Bestimmung von (w_1, \dots, w_n) als Kleinste-Quadrate-Lösung

praktisch Lösung der Normalgleichungen oder Lösung mittels QR-Zerlegung.

3 Rechnerarithmetik und Rundungsfehler

3.1 Gleitpunktarithmetik

Bemerkung 3.1 (Gleitpunktzahlen)

a) Ganzzahlige Datentypen (INTEGER) mit exakter Arithmetik

$$-\text{MaxInt} - 1, \dots, -1, 0, 1, \dots, \text{MaxInt},$$

z. B. für Indizes in Laufanweisungen.

b) *Normalisierte Gleitpunktdarstellung* (engl.: floating point numbers) zur Darstellung reeller Zahlen

$$F := \{ y : y = \pm m * \beta^{e-t} \} \subset \mathbb{R}$$

mit β ... Basis (meist 2, 8 oder 16),

t ... Mantissenlänge,

e ... Exponent, $e_{\min} \leq e \leq e_{\max}$

m ... Mantisse (ganzzahlig), $m = 0$ oder $\beta^{t-1} \leq m < \beta^t$

Schreibweise $y = \pm \beta^e * [0.d_1 d_2 \dots d_t]_\beta := \pm \beta^e \left(\frac{d_1}{\beta} + \frac{d_2}{\beta^2} + \dots + \frac{d_t}{\beta^t} \right)$

mit Mantisse $m = d_1 d_2 \dots d_t$ oder $m = 0$.

3. Rechnerarithmetik und Rundungsfehler



Beispiel Zahlendarstellung in Matlab

```
>> format long e                                % Datenausgabe mit vielen Dezimalstellen
>> 1                                             % Exakte Darstellung ganzer Zahlen
ans = 1
>> 1 - 1                                         % Exakte Arithmetik für ganze Zahlen
ans = 0
>> 1 - 1 + 1.0e-15                              % Beim Rechnen mit reellen Zahlen können
ans = 1.000000000000000e-015                    % Rundungsfehler auftreten, müssen aber nicht.
>> 1 + 1.0e-15 - 1
ans = 1.110223024625157e-015                    % Reihenfolge der Rechenschritte ist wesentlich
>> 1 + 1.0e-8 - 1
ans = 9.999999939225290e-009                    % Groessenordnung der Rundungsfehler: ca. 1.0e-16
>> sqrt(2)^2 - 2
ans = 4.440892098500626e-016                    % Groessenordnung der Rundungsfehler: ca. 1.0e-16
>> factorial(170)
ans = 7.257415615307994e+306                    % Darstellbarer Zahlenbereich nach oben beschaenkt
>> factorial(171)
ans = Inf                                         % Zahl 171! uebersteigt darstellbaren Zahlenbereich
```



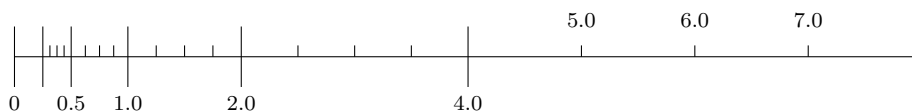
Martin-Luther-Universität Halle-Wittenberg, FB Mathematik und Informatik
Martin Arnold: Numerische Mathematik für Fachrichtung Informatik und Lehramt (WiS 2005/06)

Abbildung 3.1: Rechnen in Gleitpunktarithmetik: Beispiel Matlab.

Beispiel $\beta = 2$, $t = 3$, $e_{\min} = -1$, $e_{\max} = 3$

$$F \cap [0, \infty) = \left\{ 0, \overset{e=-1}{[0.100]_2}, \overset{e=-1}{[0.101]_2}, \overset{e=-1}{[0.110]_2}, \overset{e=0}{[0.111]_2}, \overset{e=0}{[0.100]_2}, \dots \right. \\ \left. 0, 0.25, 0.3125, 0.3750, 0.4375, 0.5, 0.625, 0.750, 0.875, \right. \\ \left. 1.0, 1.25, 1.50, 1.75, 2.0, 2.5, 3.0, 3.5, 4.0, 5.0, 6.0, 7.0 \right\} \\ \dots \overset{e=3}{[0.111]_2}$$

Beachte: Gleitpunktzahlen sind auf der reellen Achse *nicht* gleichverteilt.



IEEE-Standard 754 [1985] Binäre Gleitpunktarithmetik, Quasi-Standard

einfache Genauigkeit (single precision)

4 Byte, $\beta = 2$, $t = 23$, $e_{\min} = -126$, $e_{\max} = 127$

Zahlenbereich: $[1.2_E - 38, 3.4_E + 38]$

doppelte Genauigkeit (double precision)

8 Byte, $\beta = 2$, $t = 52$, $e_{\min} = -1022$, $e_{\max} = 1023$

Zahlenbereich: $[2.2_E - 308, 1.8_E + 308]$

Abstand zweier positiver Gleitpunktzahlen x, \tilde{x} :

Maschinenepsilon $\text{eps} = \beta^{1-t}$... kleinste Maschinenzahl, die zu $1 = [0.10 \dots 0]_\beta * \beta^1$ addiert einen von 1 verschiedenen Wert ergibt

Sind x, \tilde{x} unmittelbar benachbart, so gilt

$$\frac{1}{\beta} \text{eps} |x| \leq |x - \tilde{x}| \leq \text{eps} |x|.$$

Bemerkung 3.2 (Rundung, Rundungsfehler)

a) Sei G die Menge aller y wie in Bemerkung 3.1, jedoch für beliebiges $e \in \mathbb{Z}$, und $\text{fl} : \mathbb{R} \rightarrow G$ eine Abbildung mit

$$|x - \text{fl}(x)| = \min_{\tilde{x} \in G} |x - \tilde{x}|, \quad (x \in \mathbb{R}).$$

Der Übergang $x \mapsto \text{fl}(x)$ heißt *runden*. fl ist nicht eindeutig, praktisch meist: *Gerade-Zahl-Regel*, d. h., für $\tilde{x}_1, \tilde{x}_2 \in G$ mit $\tilde{x}_1 \neq \tilde{x}_2$ und

$$|x - \tilde{x}_1| = |x - \tilde{x}_2| = \min_{\tilde{x} \in G} |x - \tilde{x}|$$

wählt man fl so, dass d_t geradzahlig.

b) praktisch $\text{fl}(x) \stackrel{!}{\in} F$

Exponentenüberlauf (engl.: overflow): $|\text{fl}(x)| > \max \{ |y| : y \in F \}$

Exponentenunterlauf (engl.: underflow): $0 < |\text{fl}(x)| < \min \{ |y| : y \in F, y \neq 0 \}$

c) Zu jedem $x \in \mathbb{R}$ mit $\text{fl}(x) \in F$ ist

$\text{fl}(x) = x(1 + \delta)$ mit einem δ mit $|\delta| < \varepsilon$

und

$\text{fl}(x) = x/(1 + \bar{\delta})$ mit einem $\bar{\delta}$ mit $|\bar{\delta}| \leq \varepsilon$,

wobei $\varepsilon := \frac{1}{2}\beta^{1-t}$ die *Maschinengenauigkeit* (engl.: *unit round-off*) bezeichnet:

$\varepsilon \approx 5.96_{\text{E}} - 8$ (single), $\varepsilon \approx 1.11_{\text{E}} - 16$ (double).

Begründung

Für $x > 0$ ist $x = \mu * \beta^{e-t}$ mit einem $\mu \in [\beta^{t-1}, \beta^t - 1]$ und $e \in [e_{\min}, e_{\max}]$. Unmittelbar benachbarte Gleitpunktzahlen: $\underline{\mu} * \beta^{e-t}$, $\bar{\mu} * \beta^{e-t}$ mit $\underline{\mu} \leq \mu \leq \bar{\mu}$. Es gilt

$$\begin{aligned} |x - \text{fl}(x)| &= \min \{ |\mu - \underline{\mu}|, |\mu - \bar{\mu}| \} * \beta^{e-t} \\ &\leq \frac{1}{2} |\bar{\mu} - \underline{\mu}| * \beta^{e-t} \leq \frac{1}{2} * \beta^{e-t} = x \cdot \frac{1}{2\mu} \leq x \cdot \varepsilon \end{aligned}$$

Bemerkung 3.3 (Absoluter und relativer Fehler)

a) Der *absolute Fehler* einer Größe mit Soll-Wert $\bar{\xi}$ und Ist-Wert ξ ist

$$\delta\xi := |\xi - \bar{\xi}|,$$

für $\bar{\xi} \neq 0$ ist der zugehörige *relative Fehler* $f_{\text{rel}}(\bar{\xi}) := \frac{|\xi - \bar{\xi}|}{|\bar{\xi}|}$.

b) Der in Bemerkung 3.2c) betrachtete Rundungsfehler erfüllt

$$f_{\text{rel}}(x) = \frac{|\mathbf{fl}(x) - x|}{|x|} \leq \varepsilon.$$

Bemerkung 3.4 (Gleitpunktarithmetik)

Grundrechenarten $\text{op} \in \{+, -, *, /\}$, $\text{op} : \mathbb{R} \times \mathbb{R} \rightarrow \mathbb{R}$

Problem F nicht abgeschlossen bez. op

Anforderung („Standardmodell“)

$$\left. \begin{aligned} x \widetilde{\text{op}} y &= (x \text{ op } y)(1 + \delta) \\ x \overline{\text{op}} y &= \frac{x \text{ op } y}{1 + \bar{\delta}} \end{aligned} \right\}, \quad (x \in F),$$

mit $\delta = \delta(x, y; \text{op})$, $\bar{\delta} = \bar{\delta}(x, y; \text{op})$, $|\delta| \leq \varepsilon$, $|\bar{\delta}| \leq \varepsilon$.

praktisch $x \widetilde{\text{op}} y$ „unendlich genau“ (praktisch: „mit größerer Mantissenlänge“) auswerten, anschließend runden auf nächstgelegene Gleitpunktzahl (Gerade-Zahl-Regel).

Alternativen

- Runden auf nächst kleinere bzw. nächst größere Maschinenzahl \rightsquigarrow Intervallarithmetik
- „Abschneiden“ überzähliger Ziffern („chopping“)

Beispiel $\beta = 2$, $t = 3$, $x = \frac{7}{4} = [0.111]_2 * 2^1$, $y = \frac{3}{8} = [0.110]_2 * 2^{-1}$

$$\begin{aligned} x + y &= [0.111]_2 * 2^1 + [0.110]_2 * 2^{-1} = [11.100]_2 * 2^{-1} + [0.110]_2 * 2^{-1} \\ &= [100.010]_2 * 2^{-1} = [0.100010]_2 * 2^2 \end{aligned}$$

$$x \widetilde{+} y = [0.100]_2 * 2^2 = 2.00$$

absoluter Fehler: $|2.00 - \frac{17}{8}| = \frac{1}{8}$

relativer Fehler: $\frac{1}{8} : \frac{17}{8} \approx 6\%$

Maschinengenauigkeit: $\varepsilon = 2^{-3} = \frac{1}{8} = 12.5\%$

Bemerkung 3.5 (Rundungsfehleranalyse: Beispiel Addition)

geg.: $a, b, c \in F$

ges.: $s = a + b + c$ in Gleitpunktarithmetik

$$\underline{\tilde{s} := (a \tilde{+} b) \tilde{+} c}$$

$$\begin{aligned} \tilde{s} &= ((a \tilde{+} b) + c)(1 + \varepsilon_2) = ((a + b)(1 + \varepsilon_1) + c)(1 + \varepsilon_2) \\ &= s + (a + b)\varepsilon_1 + (a + b + c)\varepsilon_2 + (a + b)\varepsilon_1\varepsilon_2 \doteq s + (a + b)\varepsilon_1 + s \cdot \varepsilon_2 \end{aligned}$$

mit $|\varepsilon_1|, |\varepsilon_2| \leq \varepsilon$. Terme höherer Ordnung werden vernachlässigt („ \doteq “).

$$f_{\text{rel}}(s) \doteq \left| \varepsilon_2 + \frac{a + b}{a + b + c} \varepsilon_1 \right| \leq \left(1 + \left| \frac{a + b}{a + b + c} \right| \right) \varepsilon$$

Beachte Fehler in Zwischenergebnissen (ε_1) können verstärkt werden, ebenso auch Fehler in Ausgangsdaten.

kritisch $|a + b + c| \ll |a + b|$

$$\underline{\tilde{\tilde{s}} := a \tilde{+} (b \tilde{+} c)}$$

$$\tilde{\tilde{s}} \doteq s + (b + c)\varepsilon_3 + s \cdot \varepsilon_4 \quad \text{mit} \quad |\varepsilon_3|, |\varepsilon_4| \leq \varepsilon,$$

$$f_{\text{rel}}(s) \doteq \left| \varepsilon_4 + \frac{b + c}{a + b + c} \varepsilon_3 \right| \leq \left(1 + \left| \frac{b + c}{a + b + c} \right| \right) \varepsilon$$

Beachte Gleitpunktoperationen sind in der Regel weder assoziativ noch kommutativ.

Beispiel 3.6 (Addition in Gleitpunktarithmetik)

geg.: $\beta = 2, t = 3$

$$a = [0.111]_2 * 2^0 = \frac{7}{8}, \quad b = -[0.110]_2 * 2^0 = -\frac{6}{8}, \quad c = [0.110]_2 * 2^{-2} = \frac{3}{16} \in F$$

$$\begin{aligned} \tilde{s} &= ([0.111]_2 * 2^0 \simeq [0.110]_2 * 2^0) \tilde{+} [0.110]_2 * 2^{-2} \\ &= [0.001]_2 * 2^0 \tilde{+} [0.110]_2 * 2^{-2} = [0.100]_2 * 2^{-2} \tilde{+} [0.110]_2 * 2^{-2} \\ &= [1.01]_2 * 2^{-2} = [0.101]_2 * 2^{-1} = \frac{5}{16} = s, \quad \text{exaktes Ergebnis} \end{aligned}$$

$$\begin{aligned} \tilde{\tilde{s}} &= [0.111]_2 * 2^0 \tilde{+} (-[0.110]_2 * 2^0 \simeq [0.110]_2 * 2^{-2}) \\ &= [0.111]_2 * 2^0 \simeq [0.100]_2 * 2^0 = [0.011]_2 * 2^0 = [0.110]_2 * 2^{-1} = \frac{3}{8} \end{aligned}$$

Ergebnis: $|\tilde{\tilde{s}} - s| = \frac{1}{16}$, relativer Fehler 20%

Relativer Fehler in $b \tilde{+} c$: $\frac{1}{8} = 12.5\%$

Verstärkung im Endergebnis $\tilde{\tilde{s}}$ wegen $\left| \frac{b + c}{a + b + c} \right| = \frac{9}{5} = 1.8$

Bemerkung 3.7 (Auslöschung)

Problem Subtraktion annähernd gleich großer Zahlen in Gleitpunktarithmetik

geg.: $a, b \in \mathbb{R}$, ges.: $a - b$

$$\tilde{a} := \text{fl}(a) = a(1 + \varepsilon_a), \quad \tilde{b} := \text{fl}(b) = b(1 + \varepsilon_b) \quad \text{mit } |\varepsilon_a|, |\varepsilon_b| \leq \varepsilon$$

$$\begin{aligned} \text{fl}(a - b) &= \tilde{a} - \tilde{b} = (\tilde{a} - \tilde{b})(1 + \varepsilon_-) \quad \text{mit } |\varepsilon_-| \leq \varepsilon \\ &= a - b + \varepsilon_- \cdot (a - b) + \varepsilon_a \cdot a - \varepsilon_b \cdot b \end{aligned}$$

$$f_{\text{rel}}(a - b) = \left| \frac{a}{a - b} \varepsilon_a - \frac{b}{a - b} \varepsilon_b + \varepsilon_- \right| \leq \left(1 + \frac{|a| + |b|}{|a - b|} \right) \varepsilon$$

Relative Fehler $\varepsilon_a, \varepsilon_b$ in den Ausgangsdaten können drastisch verstärkt werden, falls $|a - b| \ll |a|, |b|$, insbesondere für $a \approx b$.

Beispiel $\beta = 2, a = \frac{3}{5}, b = \frac{4}{7}$

$$t = 5 \quad \tilde{a} = [0.10011]_2 * 2^0, \quad \tilde{b} = [0.10010]_2 * 2^0$$

$$\varepsilon_a \approx 0.010, \quad \varepsilon_b \approx 0.016, \quad \varepsilon \approx 0.031$$

$$\tilde{a} - \tilde{b} = [0.10000]_2 * 2^{-4} = \frac{1}{32}$$

$$\text{absoluter Fehler: } \frac{1}{32} - \frac{1}{35}, \quad \text{relativer Fehler: } 8.6\%$$

$$t = 3 \quad \tilde{a} = \tilde{b} = [0.101]_2 * 2^0$$

$$\varepsilon_a \approx 0.042, \quad \varepsilon_b \approx 0.094, \quad \varepsilon = 0.125$$

$$\tilde{a} - \tilde{b} = 0, \quad \text{relativer Fehler: } 100\%$$

Führende Ziffern $[0.1001 \dots]_2$ in a und b sind gleich und werden bei Subtraktion „ausgelöscht“ \Rightarrow „Auslöschung“.

Faustregel Vermeide – falls möglich – die Subtraktion annähernd gleich großer Zahlen in numerischen Algorithmen.

Strategien und Tricks

a) Unvermeidbare Subtraktionen annähernd gleich großer Zahlen möglichst an den Anfang des Algorithmus stellen.

b) Konjugierte Wurzelausdrücke

$$\begin{aligned} \bullet \quad \sqrt{1+x} - \sqrt{1-x} &= \frac{(\sqrt{1+x} - \sqrt{1-x})(\sqrt{1+x} + \sqrt{1-x})}{\sqrt{1+x} + \sqrt{1-x}} \\ &= \frac{2x}{\sqrt{1+x} + \sqrt{1-x}}, \quad |x| \ll 1 \end{aligned}$$

Bemerkung 3.7c): (Vermeidbare) Auslöschung

Beispiel:

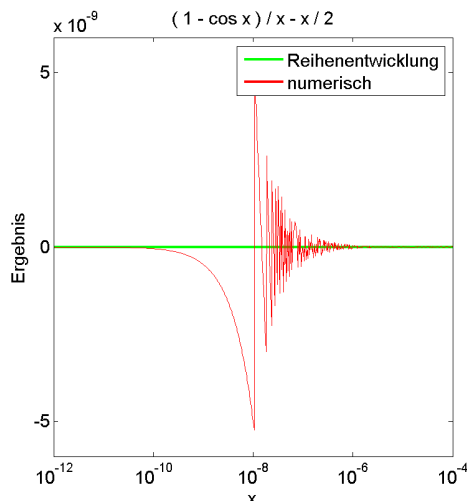
$$f(x) = \frac{1 - \cos x}{x}$$



see
numstab.m

```
% -> evaluate data
x = logspace (-12, -4, 801);
f = (1-cos(x))./x - x/2;
res = x/2 .* ( -x.^2/12 + ...
             x.^4/360 - x.^6/(360*7*8) );

% -> plot
semilogx ( x, res, 'g', x, f, 'r' );
xlabel ( 'x' );
ylabel ( 'Ergebnis' );
title ( '( 1 - cos x ) / x - x / 2' );
legend ( 'Reihenentwicklung', 'numerisch' );
```



Martin-Luther-Universität Halle-Wittenberg, FB Mathematik und Informatik
Martin Arnold: Numerische Mathematik für Fachrichtung Informatik und Lehramt (WiS 2005/06)

Abbildung 3.2: Analytische Umformungen zur Vermeidung von Auslöschung.

- $x^2 + px + q = 0 \Rightarrow x_{1,2} = -\frac{p}{2} \pm \sqrt{\frac{p^2}{4} - q}$

Auslöschung für $|q| \ll 1$, deshalb für $p \neq 0$

$$x_1 := -\frac{p}{2} - \operatorname{sgn}(p) \sqrt{\frac{p^2}{4} - q} \quad \text{mit} \quad \operatorname{sgn}(p) := \begin{cases} 1 & \text{für } p > 0, \\ 0 & \text{für } p = 0, \\ -1 & \text{für } p < 0, \end{cases}$$

$$x_2 := \frac{q}{x_1} \quad (\text{VIETAScher Wurzelsatz})$$

c) Analytische Umformungen, z. B. Reihenentwicklungen (vgl. Abb. 3.2)

$$\frac{1 - \cos x}{x} = \frac{1}{x} \left(1 - \left(1 - \frac{x^2}{2} + \frac{x^4}{24} \mp \dots \right) \right) = \frac{x}{2} \left(1 - \frac{x^2}{12} \pm \dots \right)$$

Fehler der Approximation $\frac{1 - \cos x}{x} \approx \frac{x}{2}$ betragsmäßig beschränkt durch $\frac{|x|}{2} \cdot \frac{x^2}{12}$

(Reihenrest alternierender Reihen, Satz von Leibniz)

Beispiel 3.9: Vektornormen

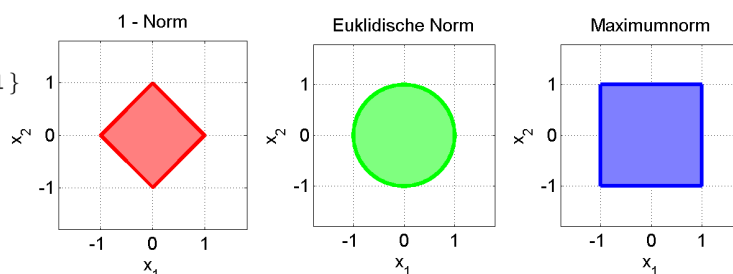
$$\|x\|_2 := \left(\sum_{i=1}^n x_i^2 \right)^{1/2} \quad \dots \quad \text{Euklidische Vektornorm}$$

$$\|x\|_1 := \sum_{i=1}^n |x_i| \quad \dots \quad \text{1-Norm}$$

$$\|x\|_\infty := \max_{i=1, \dots, n} |x_i| \quad \dots \quad \text{Maximumnorm, } \infty\text{-Norm}$$

Kugeln

$$\{x : \|x\| \leq 1\}$$



Martin-Luther-Universität Halle-Wittenberg, FB Mathematik und Informatik
Martin Arnold: Numerische Mathematik für Fachrichtung Informatik und Lehramt (WiS 2005/06)

Abbildung 3.3: Einheitskugeln im \mathbb{R}^2 .

3.2 Vektor- und Matrixnormen

Definition 3.8 (Vektornorm)

Eine Abbildung $\|\cdot\| : \mathbb{R}^n \rightarrow \mathbb{R}$ heißt Vektornorm auf \mathbb{R}^n , falls

1. $\|x\| \geq 0$, ($x \in \mathbb{R}^n$) und ($\|x\| = 0 \Leftrightarrow x = 0$) (Positivität),
2. $\|\alpha x\| = |\alpha| \|x\|$, ($\alpha \in \mathbb{R}$, $x \in \mathbb{R}^n$) (Homogenität),
3. $\|x + y\| \leq \|x\| + \|y\|$, ($x, y \in \mathbb{R}^n$) (Dreiecksungleichung).

Beispiel 3.9 (Vektornorm)

$$\text{a) } \|x\|_2 := \sqrt{\sum_{i=1}^n x_i^2} \quad \dots \quad \text{Euklidische Vektornorm}$$

$$\|x\|_1 := \sum_{i=1}^n |x_i| \quad \dots \quad \text{1-Norm}$$

$$\|x\|_\infty := \max_{i=1, \dots, n} |x_i| \quad \dots \quad \text{Maximumnorm, } \infty\text{-Norm}$$

b) Kugeln im \mathbb{R}^n : $\{x : \|x\| \leq 1\}$, vgl. Abb. 3.3.

Bemerkung 3.10 (Eigenschaften von Vektornormen)

a) Jedes Skalarprodukt $\langle \cdot, \cdot \rangle$ in \mathbb{R}^n erzeugt eine Vektornorm in \mathbb{R}^n :

$$\|x\| := \sqrt{\langle x, x \rangle}$$

mit $|\langle x, y \rangle| \leq \|x\| \cdot \|y\|$, ($x, y \in \mathbb{R}^n$) ... Cauchy-Schwarzsche Ungleichung.

b) Auf \mathbb{R}^n sind sämtliche Vektornormen *äquivalent*, d. h., zu beliebig vorgegebenen Vektornormen $\|\cdot\|_p, \|\cdot\|_q$ gibt es Konstanten $\underline{c}, \bar{c} > 0$ mit

$$\underline{c}\|x\|_q \leq \|x\|_p \leq \bar{c}\|x\|_q, \quad (x \in \mathbb{R}^n).$$

Definition 3.11 (Matrixnorm)

a) Eine Abbildung $\|\cdot\| : \mathbb{R}^{m \times n} \rightarrow \mathbb{R}$ heißt Matrixnorm, falls

1. $\|A\| \geq 0$, ($A \in \mathbb{R}^{m \times n}$) und ($\|A\| = 0 \Leftrightarrow A = 0$) (Positivität),
2. $\|\alpha A\| = |\alpha| \cdot \|A\|$, ($\alpha \in \mathbb{R}, A \in \mathbb{R}^{m \times n}$) (Homogenität),
3. $\|A + B\| \leq \|A\| + \|B\|$, ($A, B \in \mathbb{R}^{m \times n}$) (Dreiecksungleichung).

b) Eine Matrixnorm $\|\cdot\|$ heißt submultiplikativ, falls

$$\|AB\| \leq \|A\| \cdot \|B\|, \quad (A \in \mathbb{R}^{m \times n}, B \in \mathbb{R}^{n \times p}).$$

c) Eine submultiplikative Matrixnorm $\|\cdot\|$ heißt verträglich (auch: konsistent) mit einer vorgegebenen Vektornorm $\|\cdot\|$, falls

$$\|Ax\| \leq \|A\| \cdot \|x\|, \quad (A \in \mathbb{R}^{m \times n}, x \in \mathbb{R}^n).$$

Beispiel 3.12 (Frobeniusnorm)

$$\|A\|_F := \sqrt{\sum_{i=1}^m \sum_{j=1}^n a_{ij}^2} \quad \dots \text{ Frobeniusnorm}$$

- Submultiplikative Matrixnorm, verträglich mit $\|\cdot\|_2$:

$$\|Ax\|_2^2 = \sum_{i=1}^m \left(\sum_{j=1}^n a_{ij} x_j \right)^2 \leq \sum_{i=1}^m \left(\sum_{j=1}^n a_{ij}^2 \right) \left(\sum_{j=1}^n x_j^2 \right) = \|A\|_F^2 \cdot \|x\|_2^2$$

- $\|I_n\| = \sqrt{n}$

Satz 3.13 (Zugeordnete Matrixnorm)

Zu einer vorgegebenen Vektornorm $\|\cdot\|$ wird durch

$$A \mapsto \|A\| := \sup_{x \neq 0} \frac{\|Ax\|}{\|x\|} = \sup_{\|x\|=1} \|Ax\|$$

eine submultiplikative, mit $\|\cdot\|$ verträgliche Matrixnorm definiert, die der Vektornorm $\|\cdot\|$ zugeordnete Matrixnorm. Es gilt $\|I_n\| = 1$.

Beweis: vgl. Huckle/Schneider, Anhang B.2, z. B.

$$\|I_n\| = \sup_{x \neq 0} \frac{\|I_n x\|}{\|x\|} = 1. \quad \blacksquare$$

Beispiel 3.14 (Zugeordnete Matrixnorm)

a) Zeilensummennorm

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

ist $\|x\|_\infty$ zugeordnet, denn

- $\|Ax\|_\infty = \max_{i=1, \dots, m} \left| \sum_{j=1}^n a_{ij} x_j \right| \leq \max_{i=1, \dots, m} \sum_{j=1}^n |a_{ij}| |x_j| \leq \|A\|_\infty \cdot \|x\|_\infty$,
- zu einem $i_0 \in \{1, \dots, m\}$ mit $\sum_{j=1}^n |a_{i_0, j}| = \max_{i=1, \dots, m} \sum_{j=1}^n |a_{ij}|$ wählt man $x \in \mathbb{R}^n$ so, dass $\|x\|_\infty = 1$ und $x_j = 1$, falls $a_{i_0, j} > 0$, $x_j = -1$, falls $a_{i_0, j} < 0$

$$\Rightarrow \|Ax\|_\infty \geq \sum_{j=1}^n |a_{i_0, j} x_j| = \sum_{j=1}^n |a_{i_0, j}| = \|A\|_\infty.$$

b) Spaltensummennorm

$$\|A\|_1 := \max_{j=1, \dots, n} \sum_{i=1}^m |a_{ij}| \quad \text{ist } \|x\|_1 \text{ zugeordnet.}$$

c) Spektralnorm

$$\|A\|_2 := \max_{i=1, \dots, n} \sqrt{\lambda_i(A^\top A)}$$

Für orthogonale Matrizen $U \in \mathbb{R}^{n \times n}$, also $U^\top U = I_n$, ist $\lambda_i(U^\top U) = 1$ und $\|U\|_2 = 1$.

3.3 Kondition und Stabilität

Bemerkung 3.15 (Eingabefehler und Fehler im Ergebnis)

analytisch Eingabe $x \rightarrow$ Algorithmus / Berechnungsvorschrift
 \rightarrow Resultat $f(x)$

numerisch Fehler im Resultat entstehen durch

- Eingabefehler
- Fehler im Algorithmus

Numerische Eingabe $x \in F$ repräsentiert

Eingabemenge $E = \{ \tilde{x} \in \mathbb{R} : \text{fl}(\tilde{x}) = x \}$

Resultatmenge $R = f(E) := \{ f(\tilde{x}) : \tilde{x} \in E \}$



„**Kondition**“ Maß für das Verhältnis von R zu E

Ziel Fehler in der Berechnungsvorschrift sollen Menge R nicht deutlich vergrößern

Fehler im Ergebnis $y = f(x)$:

$$\delta_y := f(x + \delta_x) - f(x) \approx f'(x) \delta_x$$

Relativer Fehler:

$$\frac{\|\delta_y\|}{\|y\|} \approx \frac{\|f'(x) \delta_x\|}{\|y\|} \leq \frac{\|x\| \|f'(x)\|}{\|y\|} \cdot \frac{\|\delta_x\|}{\|x\|}$$

Definition 3.16 (Konditionszahl)

Zu einem Problem $x \mapsto f(x)$ heißt

$$\text{cond}_x := \frac{\|x\| \|f'(x)\|}{\|f(x)\|}$$

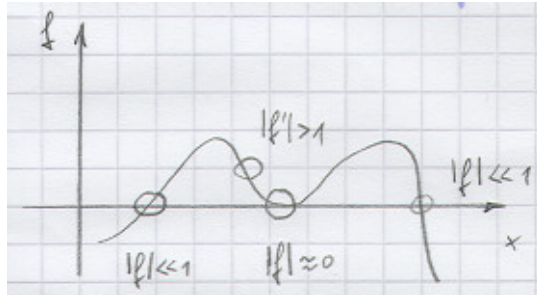
Konditionszahl. Das Problem ist gut konditioniert, wenn cond_x klein ist, und schlecht konditioniert für große Konditionszahlen cond_x .

Beispiel 3.17 (Kondition)

Exponentialfunktion $x \mapsto e^x$, $\text{cond}_x = \left| \frac{x e^x}{e^x} \right| = |x|$, gut konditioniert für $|x| \lesssim 1$.

Logarithmus $x \mapsto \ln x$, $\text{cond}_x = \left| \frac{x \cdot \frac{1}{x}}{\ln x} \right| = \frac{1}{|\ln x|}$, sehr schlecht konditioniert für $x \approx 1$.

Gute / schlechte Kondition



Polynomnullstellen häufig schlecht konditioniertes Problem

Beispiel: $\pi(t) = t^4 - 8t^3 + 24t^2 - 32t + 15.999\,999\,99 = (t - 2)^4 - 10^{-8}$

$$t_{1,2} = 2 \pm 0.01, \quad t_{3,4} = 2 \pm 0.01i$$

Relativer Fehler bei Darstellung von 15.999 999 99 durch 16.0: $\varepsilon_x = 6.25 \cdot 10^{-10}$, z. B. bei Maschinengenauigkeit $\varepsilon = 8 \cdot 10^{-10} \Rightarrow t_{1,2,3,4} = 2.0$

$$\text{cond}_{t_{1,2}} = \frac{0.01/2.01}{6.25 \cdot 10^{-10}} \approx 8.0 \cdot 10^6$$

Bemerkung 3.18 (Berechnungsvorschrift)

Zum mathematischen Problem $x \mapsto f(x)$ sei die Abbildung $x \mapsto \tilde{f}(x)$ gegeben zur Berechnung von $f(x)$ in Gleitpunktarithmetik (u. a. auch Reihenfolge der Rechenoperationen festgelegt) \rightsquigarrow *Berechnungsvorschrift*

Beispiel $f(x) = 1 - \sqrt{1 - x^2}$, für $|x| \ll 1$ gut konditioniert, $\text{cond}_x \approx 2$.

Berechnungsvorschrift 1: $\tilde{f}(x) := 1 - \left(\sqrt{1 - (x^2)}\right)$

Berechnungsvorschrift 2: $\tilde{f}(x) := \frac{(x^2)}{\left(1 + \left(\sqrt{1 - (x^2)}\right)\right)}$

Definition 3.19 (Numerische Stabilität)

Zu einem gut konditionierten Problem $x \mapsto f(x)$ heißt eine Berechnungsvorschrift $x \mapsto \tilde{f}(x)$ numerisch stabil, wenn die relativen Eingabefehler durch die Berechnungsvorschrift nicht vergrößert werden, und numerisch instabil sonst.

Bemerkung 3.20 (Lineare Gleichungssysteme: Kondition und Stabilität)

a) Betrachte zu gegebener regulärer Matrix $A \in \mathbb{R}^{n \times n}$ und gegebenem $b \in \mathbb{R}^n$ die Lösung des linearen Gleichungssystems $Ax = b$ als Abbildung $b \mapsto x := A^{-1}b$ mit gestörten Eingangsdaten b und exakten Matrizen $A, A^{-1} \Rightarrow$

$$\begin{aligned} A(x + \delta_x) &= b + \delta_b, & Ax &= b \\ \|\delta_x\| &= \|A^{-1}\delta_b\| \leq \|A^{-1}\| \|\delta_b\| \\ \|b\| &= \|Ax\| \leq \|A\| \|x\| \end{aligned}$$

Ergebnis $\frac{\|\delta_x\|}{\|x\|} \leq \|A\| \cdot \|A^{-1}\| \cdot \frac{\|\delta_b\|}{\|b\|}$

- Empfindlichkeit der Lösung gegenüber Störungen in den Eingangsdaten wird beschrieben durch die *Konditionszahl* $\text{cond}(A) := \|A\| \cdot \|A^{-1}\|$.
- Ergebnis lässt sich übertragen auf Empfindlichkeit gegenüber Störungen in A .
- Wegen $1 \leq \|I_n\| = \|A \cdot A^{-1}\| \leq \|A\| \cdot \|A^{-1}\|$ gilt stets $\text{cond}(A) \geq 1$.

Gut konditioniert: $\text{cond}(A) \lesssim 10^3$

Schlecht konditioniert: $\text{cond}(A) \gg 10^6$

Beispiel 1

$$A = \begin{pmatrix} 1 & 0 \\ 0 & \varepsilon \end{pmatrix}, \quad A^{-1} = \begin{pmatrix} 1 & 0 \\ 0 & 1/\varepsilon \end{pmatrix} \quad \text{mit } 0 < \varepsilon \ll 1$$

$\Rightarrow \text{cond}_2(A) = 1 \cdot \frac{1}{\varepsilon} = \frac{1}{\varepsilon} \rightarrow \infty$ für $\varepsilon \rightarrow 0$, Fehlerverstärkung um Faktor $1/\varepsilon$ möglich

Beispiel 2

Hilbert-Matrizen $H^{(n)} = (h_{ij}^{(n)})_{i,j} \in \mathbb{R}^{n \times n}$ mit $h_{ij}^{(n)} := 1/(i + j - 1)$, $(i, j = 1, \dots, n)$.

Zu $b^{(n)} \in \mathbb{R}^n$, $b^{(n)} = (b_i^{(n)})_i$ mit $b_i^{(n)} := \sum_{j=1}^n \frac{1}{i + j - 1}$ ist die Lösung $x^{(n)}$ des linearen

Gleichungssystems $H^{(n)}x^{(n)} = b^{(n)}$ gegeben durch $x^{(n)} = (1, 1, \dots, 1)^\top$.

Fehler der mit Matlab ($\varepsilon = 1.1_E - 16$) berechneten Lösung $\tilde{x}^{(n)}$:

n	$\ \tilde{x}^{(n)} - x^{(n)}\ _2$	$\text{cond}_2(H^{(n)})$
2	9.0e-16	1.9e+01
4	4.6e-13	1.6e+04
6	3.5e-10	1.5e+07
8	1.3e-08	1.5e+10
10	3.0e-04	1.6e+13
12	1.6e+00	1.7e+16

b) Für orthogonale Matrizen $Q \in \mathbb{R}^{n \times n}$ ist $Q^{-1} = Q^\top$ und $\|Q\|_2 = \|Q^\top\|_2 = 1$

$\Rightarrow \text{cond}_2(Q) = 1$. Operationen mit orthogonalen Matrizen lassen die Kondition einer

Matrix unverändert: $A = QR \Rightarrow \text{cond}_2(R) = \text{cond}_2(A)$.

c) Realisierung des Gauß-Algorithmus in Gleitpunktarithmetik:

Fehlerschranke hängt linear ab von $\max_{i,k} |l_{ik}|$.

Spaltenpivotisierung: $|l_{ik}| \leq 1 \rightsquigarrow$ kleine Fehlerschranke

Numerische Stabilität: numerische Lösung \tilde{x} erfüllt

$$(A + \delta_A)\tilde{x} = b$$

mit

$$\frac{\|\delta_A\|_\infty}{\|A\|_\infty} \leq 8n^3 \cdot \frac{\max_{i,j,k} |a_{ij}^{(k)}|}{\max_{i,j} |a_{ij}|} \varepsilon.$$