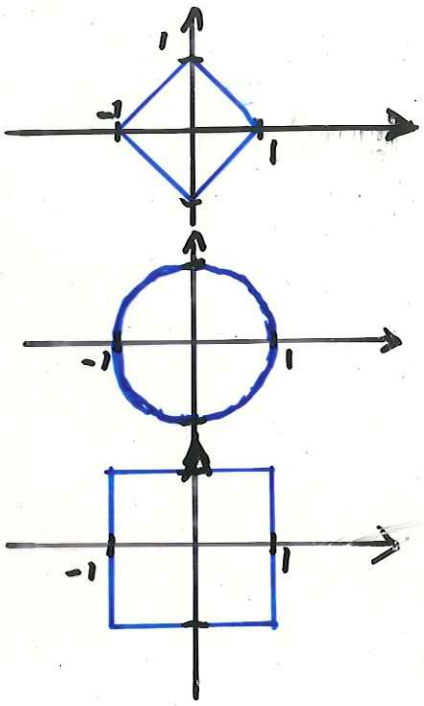
Vektor-Normen : $x \in \mathbb{R}^n$

$\{x \in \mathbb{R}^2 : \|x\| \leq 1\}$



$$\|x\|_1 = \sum_{i=1}^{n} |x_i|$$

$$\|x\|_2 = \left(\sum_{i=1}^{n} |x_i|^2\right)^{1/2}$$

$$\|x\|_\infty = \max_{1 \leq i \leq n} |x_i|$$

Matrix-Normen : $A \in \mathbb{R}^{n \times n}$ mit Einträgen a_{ij}

$$\|A\|_1 = \max_{j=1,\dots,n} \sum_{i=1}^{n} |a_{ij}|$$ → max. Spalten-summe

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

λ_{max} : max EW von $A^T A$

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

Landau Symbol $\mathcal{O}$

Betrachte zwei Funktionen $g, h : \mathbb{R} \to \mathbb{R}$. Wir verwenden die Notation

$$g(x) = \mathcal{O}(h(x)) \qquad (x \to x_0)$$

wenn es Konstanten $C > 0$ und $\delta > 0$ gibt, so daß

$$|g(x)| \leq C\,|h(x)| \qquad \forall x \text{ mit } |x - x_0| < \delta$$

gilt.

* Anschauliche Bedeutung

g wächst nicht wesentlich schneller als h

* Mathematische Definition

$$0 \leq \limsup_{x \to x_0} \left| \frac{g(x)}{h(x)} \right| < \infty$$

Beispiele

I. $g(x) = \mathcal{O}(1)$ ..... g wird durch Konstanten Wert beschränkt

II. $g(x) = \mathcal{O}(x)$ ..... g wächst ungefähr auf das Doppelte, wenn sich x verdoppelt

III. $g(x) = \mathcal{O}(x^2)$ ..... g wächst ungefähr auf das Vierfache, wenn sich x verdoppelt

IV. $g(x) = \mathcal{O}(2^x)$ ..... g wächst ungefähr auf das Doppelte, wenn sich x um 1 erhöht.

B2.2

Taylorentwicklung 1. Ordnung für $f: \mathbb{R}^n \to \mathbb{R}$

$$f(\tilde{x}) \doteq f(x) + \nabla f(x)^T (\tilde{x} - x)$$

$$= f(x) + \sum_{j=1}^{n} \frac{\partial f(x)}{\partial x_j} (\tilde{x}_j - x_j)$$

$$\Rightarrow f(\tilde{x}) - f(x) \doteq \sum_{j=1}^{n} \frac{\partial f(x)}{\partial x_j} (\tilde{x}_j - x_j) \qquad \left| \cdot \frac{1}{f(x)} \right.$$

$$\Rightarrow \frac{f(\tilde{x}) - f(x)}{f(x)} \doteq \sum_{j=1}^{n} \frac{\partial f(x)}{\partial x_j} \cdot \frac{1}{f(x)} (\tilde{x}_j - x_j) \cdot \frac{x_j}{x_j}$$

$$\Rightarrow \underbrace{\frac{f(\tilde{x}) - f(x)}{f(x)}}_{\text{rel. Fehler in der Ausgabe } \delta_y} \doteq \sum_{j=1}^{n} \underbrace{\frac{\partial f(x)}{\partial x_j} \cdot \frac{x_j}{f(x)}}_{\text{Fehler-verstärkung } \phi_j(x)} \cdot \underbrace{\frac{(\tilde{x}_j - x_j)}{x_j}}_{\text{rel. Fehler in der Eingabe in Eintrag } x_j}$$

B2.2

Wir haben

$$\left| \underbrace{\frac{f(\tilde{x}) - f(x)}{f(x)}}_{\delta_y} \right| \doteq \left| \sum_{j=1}^{n} \phi_j(x) \underbrace{\frac{\tilde{x}_j - x_j}{x_j}}_{\delta_{x_j}} \right|$$

$$\Rightarrow \quad |\delta_y| \doteq \left| \sum_{j=1}^{n} \phi_j(x) \, \delta_{x_j} \right|$$

$$\leq \sum_{j=1}^{n} |\phi_j(x)| \cdot |\delta x_j|$$

$$= |\phi_1(x)| \cdot |\delta x_1| + |\phi_2(x)| \cdot |\delta x_2| + \ldots + |\phi_n(x)| \cdot |\delta x_n|$$

$$\leq \max_{j=1,\ldots n} |\phi_j(x)| \cdot |\delta x_1| + \max_{j=1,\ldots n} |\phi_j(x)| \cdot |\delta x_2| + \ldots$$

$$\ldots + \max_{j=1,\ldots n} |\phi_j(x)| \cdot |\delta x_n|$$

$$= \underbrace{\max_{j=1,\ldots n} |\phi_j(x)|}_{} \cdot \underbrace{\sum_{j=1}^{n} |\delta x_j|}_{}$$

$$|\delta_y| \leq \underbrace{\kappa_{rel}^{\infty}(x)}_{} \cdot \underbrace{\| \delta x \|_1}_{}$$

rel. Eingabefehler in die 1-Norm

B2.2

Andere Möglichkeit:

$$|\delta y| \leq \sum_{j=1}^{n} |\phi_j(x)| \cdot |\delta x_j|$$

$$= |\phi_1(x)| \cdot |\delta x_1| + |\phi_2(x)| \cdot |\delta x_2| + \ldots + |\phi_n(x)| \cdot |\delta x_n|$$

$$\leq |\phi_1(x)| \cdot \max_{j=1,\ldots n} |\delta x_j| + |\phi_2(x)| \cdot \max_{j=1,\ldots n} |\delta x_j| + \ldots$$

$$\ldots + |\phi_n(x)| \cdot \max_{j=1,\ldots n} |\delta x_j|$$

$$= \underbrace{\sum_{j=1}^{n} |\phi_j(x)|}_{\kappa^1_{rel}(x)} \cdot \underbrace{\max_{j=1,\ldots,n} |\delta x_j|}_{\|\delta x\|_\infty}$$

$$|\delta y| \leq \kappa^1_{rel}(x) \cdot \|\delta x\|_\infty$$

$\|\delta x\|_\infty$: rel. Eingabefehler in der ∞-Norm

Kondition: $f: X \mapsto Y$

I) $f: \mathbb{R} \mapsto \mathbb{R}$ (Eingabe: Skalar, Ausgabe: Skalar)

$$\underbrace{\left|\frac{f(\tilde{x})-f(x)}{f(x)}\right|}_{\delta_y} \doteq \kappa_{rel}(x) \underbrace{\left|\frac{\tilde{x}-x}{x}\right|}_{\delta_x}$$

mit $\kappa_{rel}(x) \doteq \left| f'(x) \frac{x}{f(x)} \right|$

II) $f: \mathbb{R}^n \to \mathbb{R}$ (Eingabe: Vektor, Ausgabe: Skalar)

$x = (x_1, x_2, \ldots, x_n) \in \mathbb{R}^n$

$$\left|\frac{f(\tilde{x})-f(x)}{f(x)}\right| \lesssim \kappa_{rel}(x) \sum_{j=1}^{n} \left|\frac{\tilde{x}_j - x_j}{x_j}\right|$$

mit $\kappa_{rel}(x) = \kappa_{rel}^{\infty}(x) = \max_j |\phi_j(x)|$ und $\phi_j(x) = \frac{\partial f(x)}{\partial x_j} \cdot \frac{x_j}{f(x)}$

III) $f: \mathbb{R}^n \to \mathbb{R}^n$ (Eingabe: Vektor, Ausgabe: Vektor)

linear: $y = f(x) = A^{-1}x$, $A \in \mathbb{R}^{n \times n}$

$$\frac{\|f(\tilde{x})-f(x)\|}{\|f(x)\|} \leq \underbrace{\|A\| \cdot \|A^{-1}\|}_{} \frac{\|\tilde{x}-x\|}{\|x\|}$$

$\kappa(A)$: Konditionszahl der Matrix A

Rundungsfehler und Gleitpunktarithmetik

Normalisierte Gleitpunktdarstellung: $x \in \mathbb{R}$

$$x = \pm 0.d_1 d_2 \dots d_m \cdot b^e$$

$$= \pm \left(\sum_{j=1}^{m} d_j b^{-j} \right) \cdot b^e$$

- $b \in \mathbb{N} \setminus \{1\}$ Basis
- Exponent e : $r \le e \le R$
- Mantisse $q = \pm 0.d_1 d_2 \dots d_m$
- Mantissenlänge m
- Normalisierung: $d_1 \neq 0$

$\Rightarrow$ Maschinenzahlen $M(b, m, r, R)$

- betragsmäßig kleinste ($\neq 0$) Zahl: $x_{MIN} = b^{r-1}$
- betragsmäßig größte Zahl: $x_{MAX} = (1 - b^{-m}) b^R$

Reduktionsabbildung

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

N2.4

(Relative) Maschinengenauigkeit eps:

$$eps = \frac{b^{1-m}}{2}$$

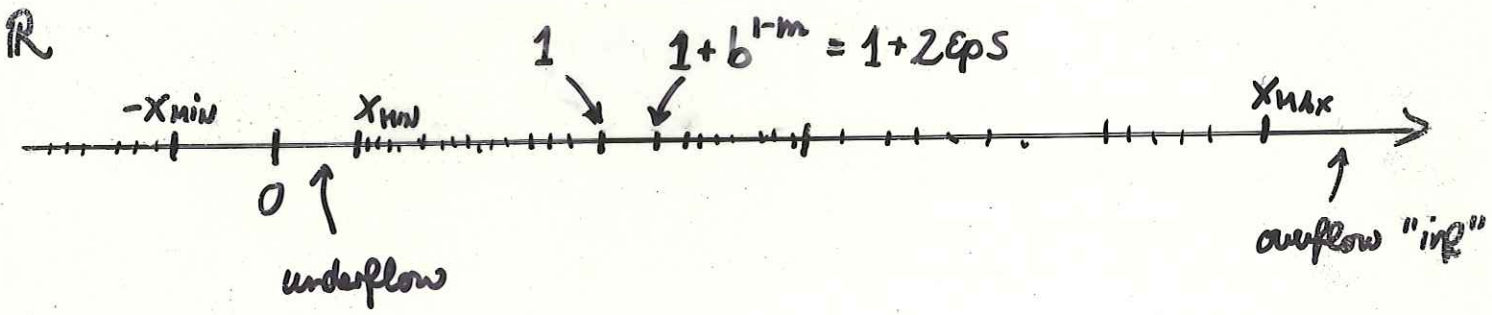
(Relativer) Rundungsfehler ε:

$$|\varepsilon| = \left| \frac{fl(x) - x}{x} \right| \leq \frac{b^{1-m}}{2} = eps$$

Für die Reduktionsabbildung gilt:

$$fl(x) = x(1+\varepsilon)$$

wobei Rundungsfehler $|\varepsilon| \leq eps$



Intervall (b=2)

$[1,2]$: $1, 1+b^{1-m}, 1+2\cdot b^{1-m}, \ldots, 2 \hat{=} 1, 1+2eps, 1+4eps, \ldots, 2$

$[2,4]$: $2, 2+2b^{1-m}, 2+4\cdot b^{1-m}, \ldots, 4$

$\vdots$

$[2^i, 2^{i+1}]$: "$[1,2]$"$\cdot 2^i$

Anmerkung zu "relativer Maschinengenauigkeit"

Relative Maschinengenauigkeit

$$eps = \frac{b^{1-m}}{2}$$

eps entspricht der halben Distanz von 1 zur nächstgrößeren exakt darstellbaren Zahl nach 1, d.h. die exakt darstellbaren Maschinenzahlen in $M(b,m,r,R)$ sind

$$1 \;,\; 1 + 2\,eps \;,\; 1 + 4\,eps, \;\ldots\;, 2$$

oder $\quad 1 \;,\; 1 + b^{1-m} \;,\; 1 + 2\,b^{1-m}, \;\ldots\;, 2$

ACHTUNG:

- Wenn nicht gerundet sondern abgeschnitten wird, ist die relative Maschinengenauigkeit

$$eps_c = b^{1-m} \quad \text{("chopping")}$$

- In manchen Büchern (und in MATLAB) ist die relative Maschinengenauigkeit definiert als

$$\overline{eps} = b^{1-m}$$

und damit genau dem Abstand von 1 zur nächstgrößeren Maschinenzahl. Um Verwirrung zu vermeiden, definieren wir dieses "MATLAB"-eps als $\overline{eps}$ mit "Strich".

FLOATGUI

Parameter (default): entspricht Notation D&R

$e_{min} = -4 \longrightarrow r = e_{min} + 1 = -3$

$e_{max} = 2 \longrightarrow R = e_{max} + 1 = 3$

$t = 3 \longrightarrow m = t + 1 = 4$

Damit ergibt sich:

- relative Maschinengenauigkeit

$$\overline{eps} = b^{1-m} = 2^{-3} = \frac{1}{8}$$

- betragsmäßig kleinste Zahl

$$x_{min} = b^{r-1} = 2^{-4} = \frac{1}{16}$$

- betragsmäßig größte Zahl

$$x_{max} = (1 - b^{-m})\, b^{R} = (1 - 2^{-4})\, 2^{3} = 8 - \frac{1}{2}$$

IEEE Standard

ACHTUNG: effektive Mantissenlänge um 1 größer als angegeben, da $d_1 = 1$ nicht gespeichert wird.

I) Single precision (32 bit)

Matlab

$$eps = \frac{2^{(1-24)}}{2} \approx 5.96 \cdot 10^{-8}$$

$$\overline{eps} = 2^{(1-24)} \approx 1.19 \cdot 10^{-7}$$ » eps ('single')

$$x_{MIN} = 2^{-126} \approx 1.175 \cdot 10^{-38}$$ » realmin ('single')

$$x_{MAX} = 2^{127}(2 - 2^{-23}) \approx 3.402 \cdot 10^{38}$$ » realmax ('single')

II) double precision (64 bit)

$$eps = \frac{2^{(1-53)}}{2} \approx 1,11 \cdot 10^{-16}$$

$$\overline{eps} = 2^{(1-53)} \approx 2.22 \cdot 10^{-16}$$ » eps

$$x_{MIN} = 2^{-1022} \approx 2.225 \cdot 10^{-308}$$ » realmin

$$x_{MAX} = 2^{1023}(2 - 2^{-52}) \approx 1.798 \cdot 10^{308}$$ » realmax

Gleitpunktoperationen $\textcircled{\nabla}$

Für $\nabla \in \{+, -, \times, \div\}$ gilt

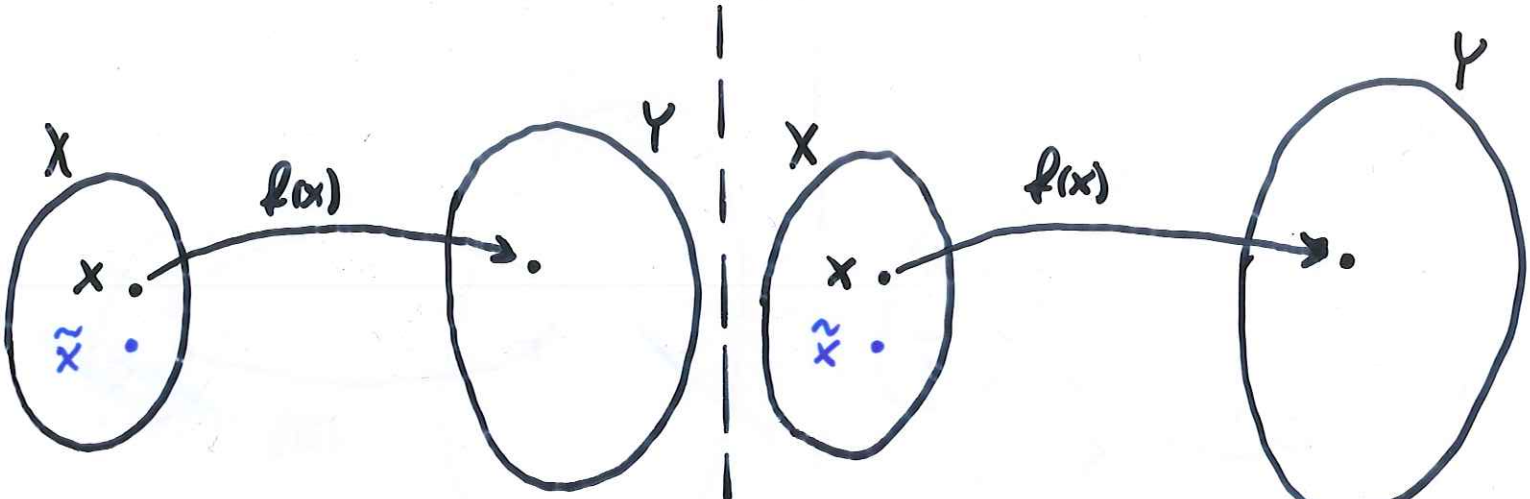
$$x \textcircled{\nabla} y = fl(x \nabla y) \quad \text{für } x, y \in \mathbb{M}(b, m, r, R)$$

$$\Rightarrow x \textcircled{\nabla} y = (x \nabla y)(1 + \varepsilon) \quad \text{für } x, y \in \mathbb{M}(b, m, r, R)$$

Aufpassen:

- Assoziativgesetz nicht mehr gültig
- Distributivgesetz nicht mehr gültig
- Gefahr von Auslöschung bei $\nabla \in \{+, -\}$

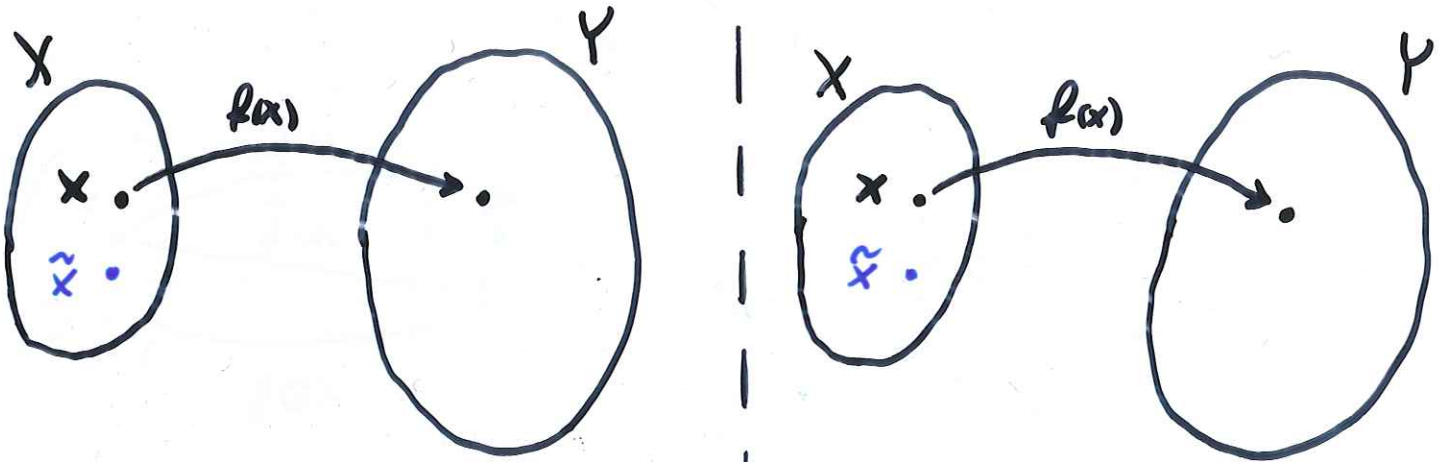
Skizze Kondition



gut konditioniert

schlecht konditioniert

Skizze Stabilität



rückwärts stabil & gut konditioniert

rückwärts stabil & schlecht konditioniert