1 Grundlagen der Numerik

1.1 Gleitpunkt-Arithmetik

Es gibt nur endlich viele Zahlen auf dem Computer.

Gleitpunktzahl: $\widehat{x} = \sigma M B^E$

σ:	Vorzeichen		
B:	Basis (feste Zahl >1);	$B = 2:$	Dualzahl
		$B = 10:$	Dezimalzahl
M:	Mantisse		
E:	Exponent		

Für die Mantisse M gilt:

$$M = \sum_{j=1}^{l} m_{-j} B^{-j}$$

$0 \leq m_{-j} \leq B-1$	:	Ziffern der Mantisse
l	:	Mantissenlänge (fest)
$m_{-1} \neq 0$ für $\widehat{x} \neq 0$	:	Normierung

Für den Exponenten E gilt:

$$E = \tau \sum_{j=1}^{k} e_j B^{k-j}$$

τ	:	Vorzeichen
$0 \leq e_j \leq B-1$	:	Ziffern des Exponenten
k	:	definierter Exponentenbereich

Schreibweise mit Ziffern: $\widehat{x} = \sigma m_{-1} m_{-2} \ldots m_{-l} (\tau e_1 \ldots e_k)$

Beispiele: $B = 10,\ l = 5,\ k = 2$, d.h. maximaler Exponent: 99

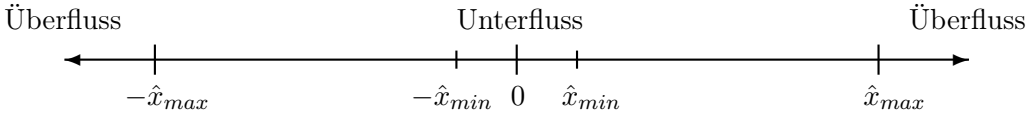
a) $x = -14.315 \quad \longrightarrow \quad \widehat{x} = -14315(02)$

b) $x = 0.00937 \quad \longrightarrow \quad \widehat{x} = 93700(-02)$

Definition: $\mathbb{M}(B;\ l,\ k)$ = Menge der *Maschinenzahlen.*

Es gilt:

- $\mathbb{M}$ ist endlich.
- $\mathbb{M}$ besitzt eine grösste Zahl $\widehat{x}_{max}$ und eine kleinste positive Zahl $\widehat{x}_{min}$.
- Die Zahlen von $\mathbb{M}$ sind nicht gleichabständig auf der Zahlengeraden verteilt.

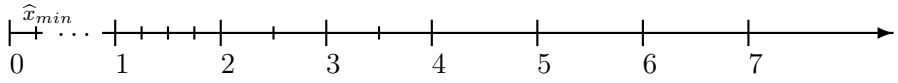


Beispiel: $\mathbb{M}(2;\ 3,\ 2)$

verfügbare Mantissen	verfügbare Exponenten
$000 = 0$	$00 = 0$
$100 = \frac{1}{2}$	$\pm 01 = \pm 1$
$101 = \frac{5}{8}$	$\pm 10 = \pm 2$
$110 = \frac{3}{4}$	$\pm 11 = \pm 3$
$111 = \frac{7}{8}$	

$\widehat{x}_{max} = \frac{7}{8} \cdot 2^3 = 7, \quad \widehat{x}_{min} = \frac{1}{2} \cdot 2^{-3} = 2^{-4} = \frac{1}{16}$

Die Zahlen von $\mathbb{M}(2;3,2)$ auf der Zahlengeraden:



Integerdarstellung $\mathbb{I}(B;r)$

$$i = \sigma N,\ N = \sum_{j=1}^{r} n_j B^{r-j} \text{ mit den Ziffern: } 0 \leq n_j \leq B-1\ ,$$

wobei $r = l + k + 1$.

Reduktionsabbildung

Um Zahlen aus $\mathbb{R}$ in $\mathbb{M}$ approximativ darstellen zu können, braucht es die Reduktionsabbildung

$$\rho: \begin{array}{ccc} \mathbb{R} & \longmapsto & \mathbb{M} \\ x & \longmapsto & \widehat{x} \end{array}$$

Definition: Sei $x \in \mathbb{R}$ fest; $\widehat{x} = \rho(x)$ ist die zu x nächstgelegene Zahl in $\mathbb{M}$; d.h. $\rho(x) \in \mathbb{M}$ so, dass $|x - \rho(x)|$ = minimal (*Runden*).

Definition: Der *relative Reduktionsfehler* ist somit gegeben als:

$$\left|\frac{x-\rho(x)}{x}\right| \leq \frac{\frac{1}{2}B^{E-l}}{B^{E-1}} = \frac{1}{2}B^{1-l} =: eps \;\; (Maschinengenauigkeit)$$

Es gilt: Für alle $\alpha \in \mathbb{R},\ 0 < \alpha \leq eps,$ ist $\rho(1+\alpha) = 1$.

Bemerkung: Der negative Zehneralgorithmus der Maschinengenauigkeit *eps* sagt aus, wieviele Ziffern einer reellen Zahl man ungefähr auf dem Computer (im Dezimalsystem ausgedrückt) richtig darstellen kann; z.B. für $eps \cong 6 \cdot 10^{-8}$ sind es 7 Ziffern.

Der sogenannte *IEEE* Standard (32 BIT), der auf den meisten PC- und Mac-Computern realisiert ist, hat die folgenden Charakterisierungen:

B	l	k	$\widehat{x}_{min}$	$\widehat{x}_{max}$	E_{min}	E_{max}	eps
2	24	7	$\sim \frac{3}{2} \cdot 10^{-39}$	$\sim \frac{3}{2} \cdot 10^{-38}$	-128	127	$\sim 6 \cdot 10^{-8}$

Pseudoarithmetik

$\mathbb{M}$ ist endlich. Daraus folgt, dass $\mathbb{M}$ im Gegensatz zu $\mathbb{R}$ gegenüber elementaren Operationen nicht abgeschlossen ist.

Im Allgemeinen gilt: $\widehat{x} + \widehat{y} \notin \mathbb{M},\ \widehat{x} * \widehat{y} \notin \mathbb{M},\ \widehat{x}/\widehat{y} \notin \mathbb{M}$

Beispiel: $\mathbb{M}(10;3,2)$:

$$\begin{aligned} 9.22 + 1.22 &= 10.44 \quad \notin \mathbb{M} \\ 922(01) + 122(01) &= 104(02) \quad \in \mathbb{M} \\ 9.22 * 1.22 &= 11.2484 \quad \notin \mathbb{M} \end{aligned}$$

$\widehat{x} * \widehat{y}$ hat doppelte Länge, $\widehat{x}/\widehat{y}$ geht nicht auf.

Die Arithmetik in $\mathbb{R}$ wird ersetzt durch die Pseudoarithmetik in $\mathbb{M}$:

$$\begin{aligned} \widehat{x} + \widehat{y} &= \rho(\widehat{x} + \widehat{y}) \\ \widehat{x} * \widehat{y} &= \rho(\widehat{x} * \widehat{y}) \\ \widehat{x}/\widehat{y} &= \rho(\widehat{x}/\widehat{y}) \end{aligned}$$

Bemerkung: Bei Überfluss muss die Rechnung abgebrochen werden.

Rechengesetze in $\mathbb{R}$*:*

- Assoziativ- und das Distributivgesetze gelten nicht in $\mathbb{M}$.
- Das Kommutativgesetz bezüglich Addition und Multiplikation gilt in $\mathbb{M}$.

Beispiel: Assoziativgesetz bezüglich Addition $\mathbb{M}(10;3,\cdot)$:

$$x = 9.22,\ y = 1.22,\ z = 0.242\ :\ x + y + z = 10.682$$

$$\begin{array}{rclcrclcl} \widehat{x}+\widehat{y} & = & \rho(10.44) & = & 10.4, & (\widehat{x}+\widehat{y})+\widehat{z} & = & \rho(10.642) & = & 10.6 \\ \widehat{y}+\widehat{z} & = & \rho(1.462) & = & 1.46, & \widehat{x}+(\widehat{y}+\widehat{z}) & = & \rho(10.69) & = & 10.7 \text{ (besser)} \end{array}$$

Folgerungen:

- Lange Summen von kleinen zu grossen Summanden summieren!
- In $\mathbb{M}$ hat $1+\widehat{x}=1$ viele Lösungen, nämlich alle $\widehat{x}$ mit $|\widehat{x}| < eps$.

1.2 Fehlerfortpflanzung

Der Fehler, der bei der Reduktionsabbildung entsteht, kann im Laufe eines Algorithmus vergrössert werden. Wir betrachten zuerst nur den Fehler nach *einer* elementaren (exakten) Operation.

Sei $\widehat{x}$ der fehlerbehaftete Wert und x der exakte Wert:

Absoluter Fehler: $\Delta x = \widehat{x} - x$

Relativer Fehler: $\delta x = \Delta x / x$ (für $x \neq 0$)

Nur bei der Addition (Subtraktion) kann der Fehler, und zwar der relative, bei einer exakten Operation viel grösser sein als der ursprüngliche Fehler.

Beweis:

i) *Absolute Fehler*:

$$\Delta(x \pm y) = (\widehat{x} \pm \widehat{y}) - (x \pm y) = \Delta x \pm \Delta y$$

$$\begin{aligned} \Delta(x * y) &= \widehat{x}\widehat{y} - xy = (x + \Delta x) + (y + \Delta y) - xy \\ &= x\Delta y + y\Delta x + \Delta x \Delta y \cong x\Delta y + y\Delta x = xy(\delta x + \delta y) \end{aligned}$$

$$\begin{aligned} \Delta(x/y) &= \frac{\widehat{x}}{\widehat{y}} - \frac{x}{y} = \frac{x+\Delta x}{y+\Delta y} \cdot \frac{y}{y} - \frac{x}{y} \cdot \frac{y+\Delta y}{y+\Delta y} = \frac{y\Delta x - x\Delta y}{y(y+\Delta y)} \\ &= \frac{x\Delta x}{(y+\Delta y)x} - \frac{x}{y} \cdot \frac{\Delta y}{y+\Delta y} \cong \frac{x}{y}\delta x - \frac{x}{y}\delta y = \frac{x}{y}(\delta x - \delta y) \end{aligned}$$

ii) Relative Fehler:

$$\begin{aligned}
\delta(x \pm y) &= \frac{\Delta(x \pm y)}{x \pm y} = \frac{x}{x \pm y} \cdot \frac{\Delta x}{x} \pm \frac{y}{x \pm y} \cdot \frac{\Delta y}{y} = \frac{x}{x \pm y}\delta x + \frac{y}{x \pm y}\delta y \\
\delta(x * y) &= \frac{\Delta(x * y)}{xy} \cong \delta x + \delta y \\
\delta(x/y) &= \frac{\Delta(x/y)}{x/y} \cong \delta x - \delta y
\end{aligned}$$

□

Folgerungen:

- Bei Addition (Subtraktion) addiert (subtrahiert) sich der absolute Fehler.
- Bei Multiplikation (Division) addiert (subtrahiert) sich der relative Fehler.
- Falls

  $$|x \pm y| << \begin{cases} |x| \\ |y| \end{cases}$$

  kann

  $$|\delta(x \pm y)| >> |\delta x| + |\delta y| \ .$$

  Dieses Phänomen nennt man *Auslöschung*:

Beispiele:

1. $\mathbb{M}(10; 4, \cdot)$; $x = \pi$, $y = \frac{22}{7}$, $x - y = -1.265\ldots \cdot 10^{-3}$

   $\widehat{x} - \widehat{y} = 3.142 - 3.143 = -1.000 \cdot 10^{-2}$ (nichsignifikante Nullen in der Mantisse);

   absoluter Fehler $\cong 2.6 \cdot 10^{-4}$; relativer Fehler $\cong -2 \cdot 10^{-}1$

   d.h. nur mehr eine richtige Ziffer im Resultat.

2. $\mathbb{M}(10; 3, \cdot)$, $x = 701$, $y = 700$; $\underbrace{x^2}_{\notin \mathbb{M}} - y^2 = 491401 - 490000 = 1401.$

   $$\rho(x^2) - \rho(y^2) = 4.91 \cdot 10^5 - 4.90 \cdot 10^5 = 1.00 \cdot 10^3;$$
   $$|\Delta| = 401, |\delta| \cong 3 \cdot 10^{-1} \ .$$

   Besser: $x^2 - y^2 = (x + y)(x - y)$

   $$\left.\begin{aligned} \widehat{x} + \widehat{y} &= 1.40 \cdot 10^3 \\ \widehat{x} - \widehat{y} &= 1.00 \end{aligned}\right\} \Rightarrow \ |\Delta| = 1; \ \ \delta \cong 1 \cdot 10^{-3};$$

Bemerkung:

- Subtraktion fast gleich grosser Zahlen vermeiden!
- Falls nicht möglich: "lokal" doppelte Genauigkeit!
- Es gilt: Falls der relative Fehler (im Dezimalsystem ausgedrückt) von der Grössenordnung 10^{-k} ist, dann sind ungefähr k Ziffern des Resultats richtig.

1.3 Kondition eines Problems

Wir betrachten jetzt *viele* elemtentare Operationen nacheinander. Genauer: Wir betrachte $y = H(x)$ (Input x, Resultat y, exakte Rechenoperationen H), wobei H eine genügend glatte Funktion sei.

Absoluter Fehler von $H(x)$:

$$\begin{aligned} \Delta H(x) = H(\widehat{x}) - H(x) &= H(x + \Delta x) - H(x) \\ &= H(x) + H'(x)\Delta x + \ldots \text{(Taylorentwicklung)} \end{aligned}$$

Also ist:

$$\Delta H(x) \cong H'(x)\Delta x$$

Relativer Fehler von $H(x)$:

$$\delta H(x) = \frac{\Delta H(x)}{H(x)} \cong \frac{xH'(x)\Delta x}{H(x)x} \text{ (für } H(x) \neq 0 \text{ und } x \neq 0) \,.$$

Also ist:

$$|\delta H| \cong \underbrace{\left|\frac{xH'(x)}{H(x)}\right|}_{=:\kappa_H} |\delta x| \,.$$

Definition: Die *Konditionszahl* von H, κ_H, ist der Verstärkungsfaktor des relativen Fehlers von x durch H.

Bemerkung: Falls κ_H gross ist, hat das Problem eine schlechte Kondition.

Beispiel: Die quadratische Funktion $x^2 - 2ax + 1 = 0,\ a > 1$, hat die beiden Nullstellen:

$$\begin{aligned} x_1 &= a + \sqrt{a^2 - 1} &> 1 \\ x_2 &= a - \sqrt{a^2 - 1} &< 1 \end{aligned}$$

Wir betrachten die kleinere Nullstelle $x_2 = H(a) = a - \sqrt{a^2 - 1}$:

$$\kappa_H = \left|\frac{aH'(a)}{H(a)}\right| = \left|\frac{-a}{\sqrt{a^2 - 1}}\right|$$

1. $a >> 1 :\ \kappa_H \cong 1$, d.h. das Problem ist gut konditioniert.

   Aber: $H(a) = a - \sqrt{a^2 - 1}$ ist ein schlechter Algorithmus (Auslöschung!).

   Guter Algorithmus:

   $$H(a) = \frac{1}{a + \sqrt{a^2 - 1}} \qquad \text{(keine Auslöschung!)}$$

2. $a \cong 1 :\ \kappa_H >> 1$, d.h. das Problem ist schlecht konditioniert; es muss mit höherer Genauigkeit gerechnet werden.

Kondition einer Matrix bzw. eines linearen Gleichungssystems

Wir betrachten

$$Ax = b;\ A\ n{\times}n \text{ Matrix, regulär; } b \neq 0\ .$$

Dieses lineare Gleichungssystem hat die Lösung

$$x = A^{-1}b \neq 0\ .$$

Annahme: b sei 'fehlerhaft', d.h. statt b habe man $\widehat{b}$ mit einem 'kleinen' (in der Norm) absoluten Fehler: $\Delta b := \widehat{b} - b$.

Begründung für diese Annahme:

- Darstellung der $b_j \in \mathbb{R}$ auf dem Computer (d.h. in $\mathbb{M}$).
- ungenaue Daten.

Wichtiger ist der relative Fehler

$$\delta b := \frac{1}{||b||}\Delta b\ .$$

Fragestellung:
Das lineare Gleichungssystem $A\widehat{x} = \widehat{b}$ hat die 'exakte' Lösung $\widehat{x} = A^{-1}b$. Wie ist der Zusammenhang zwischen δb und $\delta x := \frac{1}{||x||}\Delta x$?

Es gilt:

$$\begin{aligned}
\Delta x &= \widehat{x} - x = A^{-1}\widehat{b} - A^{-1}b = A^{-1}(\widehat{b} - b) = A^{-1}\Delta b \\
||\delta x|| &= \frac{||b|| \cdot ||A^{-1}\Delta b|| \cdot ||\Delta b||}{||x|| \cdot ||\Delta b|| \cdot ||b||} \overset{b=Ax}{=} \underbrace{\frac{||Ax||}{||x||}}_{\leq ||A||} \cdot \underbrace{\frac{||A^{-1}\Delta b||}{||\Delta b||}}_{\leq ||A^{-1}||} \cdot ||\delta b|| \\
||\delta x|| &\leq ||A|| \cdot ||A^{-1}|| \cdot ||\delta b|| \qquad (' =' \text{ ist möglich})
\end{aligned}$$

Folgerung: Der Fehler in b 'kann' in der Lösung um den Faktor $\kappa(A)$ verstärkt werden.

Definition: Die Grösse $\kappa(A) := ||A|| \cdot ||A^{-1}||$ heisst *Kondition der Matrix A.*

Es gilt:

$$1 = ||I_n|| = ||A \cdot A^{-1}|| \leq ||A|| \cdot ||A^{-1}|| = \kappa(A).$$

Falls auch A 'fehlerhaft' ist ($\widehat{A}\widehat{x} = \widehat{b}$), kann man zeigen:

$$||\delta x|| \leq \frac{\kappa(A)}{1 - \kappa(A)||\delta A||}\left(||\delta A|| + ||\delta b||\right)\ .$$

Bemerkung: Für $\kappa(A)||\delta A|| << 1$ ist $\kappa(A)$ ungefähr gleich der Kondition des Problems $Ax = b$.

In der 2-Norm gilt für $\kappa(A)$:

$$\kappa(A) = \frac{\sqrt{\mu_{max}}}{\sqrt{\mu_{min}}}, \quad \text{wobei} \quad \left.\begin{array}{rcl} \mu_{min} & := & \text{kleinster} \\ \mu_{max} & := & \text{grösster} \end{array}\right\} \text{Eigenwert von } A^T A.$$

Für den Fall dass $A = A^T$ gilt für $\kappa(A)$:

$$\kappa(A) = \frac{|\lambda_{max}|}{|\lambda_{min}|}, \quad \text{wobei} \quad \left.\begin{array}{rcl} \lambda_{min} & := & \text{kleinster} \\ \lambda_{max} & := & \text{grösster} \end{array}\right\} \text{Eigenwert von } A.$$

Zur *Erinnerung aus der linearen Algebra*:

$$||A||_2 := \sup_{x \neq 0} \left\{ \frac{||A||_2}{||x||_2} \right\} = \sup_{||x||_2 = 1} \{||Ax||_2\} \ .$$

Da $A^T A$ symmetrisch ist, existiert eine orthogonale Matrix T, so dass

$$T^T (A^T A) T = D = \left\{ \begin{array}{ccc} \mu_1 & & 0 \\ & \ddots & \\ 0 & & \mu_n \end{array} \right\} ,$$

wobei μ_i die Eigenwerte von $A^T A$ sind.

Wir betrachten die orthogonale Transformation $x = Ty$ und

$$\begin{aligned} 0 \leq ||A||_2^2 &= \sup_{||x||_2=1} \{{||Ax||_2}^2\} = \sup_{||y||_2=1} \{{||ATy||_2}^2\} \\ &= \sup_{||y||_2=1} \{y^T \underbrace{T^T A^T A T}_{=D} y\} \\ &= \sup_{||y||_2=1} \{\sum_1^n \mu_i y_i^2\} = \max_i \{\mu_i\} \ . \end{aligned}$$

Folgerungen:

- $||A||_2 = \sqrt{\mu_{max}} = \sqrt{\text{maximaler Eigenwert von } A^T A}$.
- Falls A orthogonal ist, ist $||A||_2 = 1$.
- Falls $A^T = A$, ist $||A||_2 = |\lambda_{max}|$, wobei λ_{max} der betragsmässig grösste Eigenwert von A ist.
- Für A regulär ist $||A^{-1}||_2 = \frac{1}{\sqrt{\mu_{min}}}$, wobei μ_{min} der minimale Eigenwert von $A^T A$ ist.
- Falls $A^T = A^{-1}$, ist $||A^{-1}||_2 = \frac{1}{|\lambda_{min}|}$, wobei λ_{min} der betragsmässig kleinste Eigenwert von A ist.

Beispiele:

1. $A = \begin{pmatrix} 2 & 1 & 1 \\ 1 & 2 & 1 \\ 1 & 1 & 2 \end{pmatrix}$, Eigenwerte $\lambda_{1,2,3} = 1, 1, 4$;
   Also ist $\kappa(A) = 4$, und es folgt, dass $Ax = b$ 'gut konditioniert' ist.

2. $A = \begin{pmatrix} 168 & 113 \\ 113 & 76 \end{pmatrix}$, Eigenwerte $\lambda_{1,2} = 244.004\ldots, -0.004\ldots$.
   Also ist $\kappa(A) \cong 5.95 \cdot 10^4$, und es folgt, dass $Ax = b$ (relativ) 'schlecht konditioniert' ist.

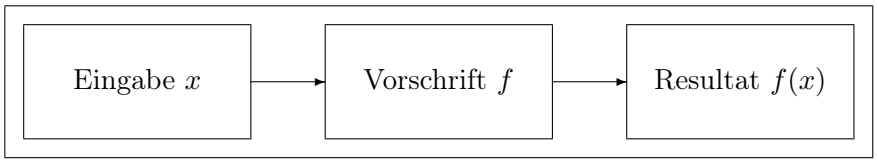
Bemerkungen:

- Auf einem Computer mit d-stelliger Arithmetik gilt: Hat A Kondition $\cong 10^k$ so wird typischerweise die Lösung von $\widehat{A}\widehat{x} = \widehat{b}$ nur etwa $(d - k)$ richtige Ziffern aufweisen.
- Beim Lösen eines linearen Gleichungssystems auf dem Computer sollte man sich immer auch eine Schätzung der Kondition der Koeffizientenmatrix beschaffen.

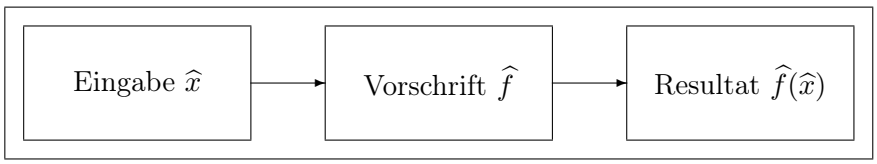
Übersicht über die Fehleranalyse

Gegeben sei das Problem f

Ideal:



Real:



Absolute Fehler:

Eingabefehler: $\widehat{x} - x$ (verursacht durch 'falsche Daten', Pseudoarithmetik)

Totaler Fehler: $\widehat{f}(\widehat{x}) - f(x) = \widehat{f}(\widehat{x}) - f(\widehat{x}) + f(\widehat{x}) - f(x)$

Berechnungsfehler: $\widehat{f}(\widehat{x}) - f(\widehat{x})$

Transportierter Eingabefehler: $f(\widehat{x}) - f(x)$

Der Berechungsfehler setzt sich zusammen aus dem *Diskretisationsfehler* (Approximation von f durch $\widehat{f}$) plus dem *Rundungsfehler* (durch Pseudoarithmetik).

Also gilt:

$$\begin{aligned}\text{Totaler Fehler} \quad &= \quad \text{Diskretisationsfehler} + \text{transportierter Eingabefehler} \\ &\quad + \text{Rundundungsfehler}\end{aligned}$$

Relative Fehler: $\dfrac{\text{absolute Fehler}}{\text{exakter Wert}}$

Bemerkung: Wir haben bis jetzt den transportierten Eingabefehler (Kondition eines Problems) und den Rundungsfehler (Pseudoarithmetik) betrachtet. Wir werden im Folgenden noch viele Beispiele von Diskretisationsfehlern kennenlernen, Rundungsfehler aber nicht mehr weiter betrachten.