# 7. Rundungsfehler

## 7.1 Zahlendarstellung

Menge der Gleitkommazahlen $\mathbb{M} \subset \mathbb{Q}$ :

$$\mathbb{M} := \{ x \mid x = \sigma \cdot B^e \cdot m, \quad 0 \le m < B \}$$

mit

$\sigma = \pm 1$ : Vorzeichen

$B$ : Basis, meist $B=2$

Taschenrechner und Vorles'g auch $B=10$

$e$ : Exponent, $e_0 \le e \le e_1$

$m = \sum_{k=0}^{n-1} d_k B^{-k}$, $d_k \in \{0, 1, \cdots, B-1\}$ : Mantisse

$d_1 > 0$ : normalisiert, $d_1 = 0$ "denormal"

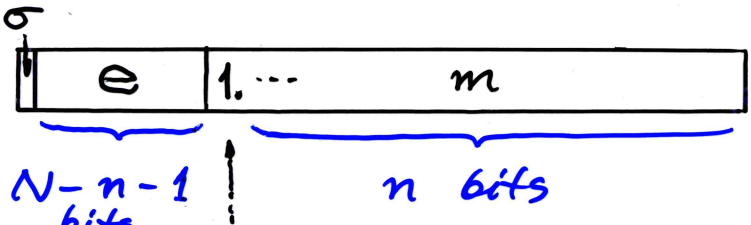
$n$ : Mantissenlänge

### Bemerkungen

— $\mathbb{M}$ ist endlich, enthält nur rationale Zahlen mit Nenner $2^l$ (für $B=2$).

— Gleitkommazahlen werden meist in Speicherbereiche fester Länge $N$ gepackt, z. B.

$B=2$ :




total $N$ bits

**Definition** : "Maschinenepsilon" :

$$\frac{1}{2}\,eps = \min_{1+\delta>1} \delta .$$

$B=2$: $eps = 2^{-n}$

$eps$ = eine Einheit an der letzten Stelle der Mantisse !

**Definition:** "Überflussgrenze":

realmax = grösste darstellbare Zahl
$= 2^{e_1+1}$ $(B=2)$

"Unterflussgrenze":

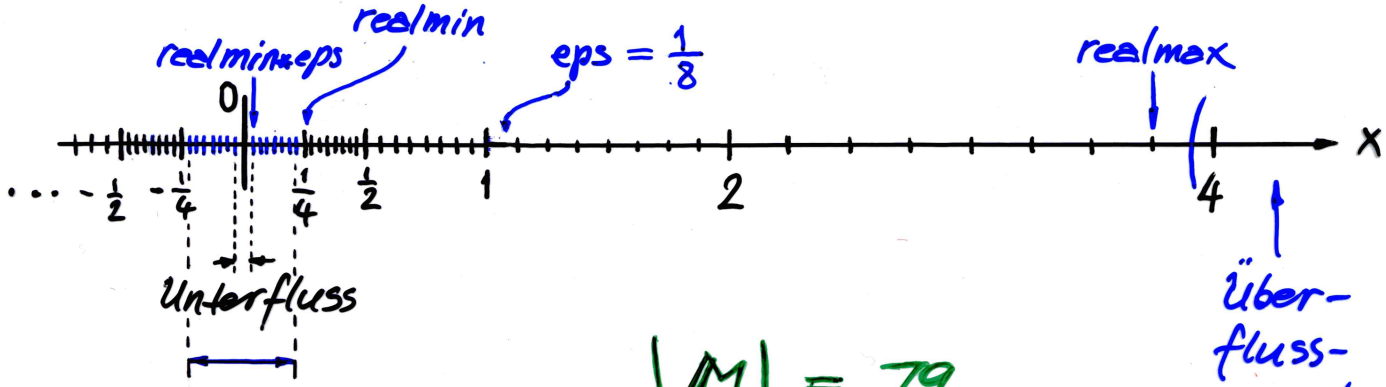
realmin = kleinste darstellbare normalisierte positive Zahl
$= 2^{e_0}$

eps, realmax, realmin sind Matlab-Befehle!

**Folgerung:** Die kleinste positive Zahl $\in \mathbb{M}$ (auch nicht-normalisiert) ist realmin * eps

## Beispiele

(i) $B=2$, $N=7$, $n=4$, $-2 \le e \le 1$




$|\mathbb{M}| = 79$

(ii) Matlab (= IEEE Standard)

$B=2$, $N=64$, $n=52$, $-1022 \le e \le 1023$ (11 bits)

$$eps = 2^{-52} = 2.2204 \cdot 10^{-16}$$

$$realmax = 2^{1024} = 1.7977 \cdot 10^{308}$$

$$realmin = 2^{-1022} = 2.2251 \cdot 10^{-308}$$

$$eps * realmin = 2^{-1074} = 4.9407 \cdot 10^{-324}$$

# Die Rundungsabbildung $\rho(x)$

Sei $x \in \mathbb{R}$

**Def.** $\rho(x) \in \mathbb{M}$ : eine der (höchstens 2) zu $x$ nächstgelegenen Gleitkommazahlen.

**Bemerkung:** In den Grenzfällen ($x$ genau zwischen 2 Zahlen $\in \mathbb{M}$) sollte gleich häufig auf- bzw. abgerundet werden.

**SATZ:** Ausserhalb des Ü/U'flussbereiches gilt

$$\left| \frac{\rho(x) - x}{x} \right| \leq eps$$

(Schranke für relativen Fehler bei Anwendung von $\rho$)

## Maschinenoperationen, $\oplus, \circledast, \ldots$

Erst exakt, dann $\rho$ anwenden, z.B.

**Def:** $a, b \in \mathbb{M}$. $a \oplus b := \rho(a+b)$ (z.B.)

**Bemerkung:** Maschinenoperationen erfüllen die bekannten Rechengesetze manchmal nicht.

**Bsp.** Basis $B = 10$, Mantissenlänge $n = 3$

$a = 8.02$
$b = 3.12$
$c = 0.444$

$s_1 := \rho(a+b) = 11.1$
$\rho(s_1 + c) = 11.5$

$\rho(s_2 + a) = 11.6$
$s_2 := \rho(c+b) = 3.56$

**Exakt:** $s = a+b+c = 11.584$, $\rho(s) = 11.6$

Reihenfolge der Terme in einer Summe kann relevant sein. Besser: von kleinen zu grossen Termen summieren.

# Über- und Unterflusseffekte

**Vorbemerkung:** Ist das Resultat einer Rechnung im Ü/Uflussbereich: nichts zu machen!

**Ziel:** Falls Endresultat $\in \mathbb{M}$, sollte die Rechnung funktionieren

**Beispiel:** Euklidische Norm in $\mathbb{R}^2$

$$\|x\| = \sqrt{x_1^2 + x_2^2} \quad \text{für Vektor } x = \begin{pmatrix} x_1 \\ x_2 \end{pmatrix}$$

Schlechter Algorithmus: eingeschränkt auf den Bereich $(\sqrt{realmin}, \sqrt{realmax})$!

Man kann es besser machen:

```
xl = |x1| ;   xs = |x2| ;
if xl < xs,  xl = |x2| ; xs = |x1| ; end;
||x|| =  xl * sqrt( 1 + (xs/xl)^2 ) ;
```

## 7.2. Fehlerfortpflanzung, Auslöschung

Sei $a \in \mathbb{R}$, $\tilde{a}$ eine Näherung für $a$

**Def:** absoluter Fehler: $\Delta a := \tilde{a} - a$

relativer Fehler: $\delta a := \frac{\Delta a}{a} = \frac{\tilde{a} - a}{a} = \frac{\tilde{a}}{a} - 1, \quad a \neq 0$

$$\Rightarrow \boxed{\tilde{a} = a + \Delta a = a(1 + \delta a)}$$

Fehlerfortpflanzung:

$$\Delta(a \pm b) = \Delta a \pm \Delta b$$

$$\delta(a \cdot b) \doteq \delta a + \delta b; \quad \delta\left(\frac{a}{b}\right) \doteq \delta a - \delta b, \quad b \neq 0$$

**Beweis:** $\tilde{a} = a(1+\delta a)$, $\tilde{b} = b(1+\delta b)$ $\Rightarrow$ $\tilde{a} \cdot \tilde{b} = a \cdot b (1 + \underbrace{\delta a + \delta b}_{\delta(a \cdot b)} + \underbrace{\delta a \cdot \delta b}_{\text{klein}})$

Realistisch: Nur Fehlerschranken bekannt:

$|\Delta a| \le \varepsilon_1, \; |\Delta b| \le \varepsilon_2 \Rightarrow |\Delta(a \pm b)| \le \varepsilon_1 + \varepsilon_2$

(folgt aus Dreiecksungleichung)

$|\delta a| \le \varepsilon_1, \; |\delta b| \le \varepsilon_2 \Rightarrow |\delta(a \cdot b)| \le \varepsilon_1 + \varepsilon_2$

$|\delta(\frac{a}{b})| \le \varepsilon_1 + \varepsilon_2$ !!

## Auslöschung

Subtraktion fast gleicher Zahlen

Beispiel: $a = \frac{22}{7}$, $b = \pi$ in $M$ mit $B=10$, $n=5$

$\Rightarrow eps = \frac{1}{2} B^{1-n} = 5 \cdot 10^{-5}$

$a = 3.1429 \quad \delta a = 1.36 \cdot 10^{-5} < eps$

$b = 3.1416 \quad \delta b = 0.23 \cdot 10^{-5} < eps$

$a - b = 0.0013000$

$\frac{22}{7} - \pi = 0.0012645$

Auffüllen mit Nullen in Basis B

$\delta(a-b) = 2.8 \cdot 10^{-2} \gg eps$

grosser relativer Fehler!

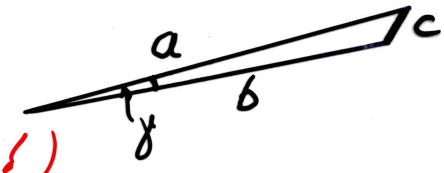
## Massnahmen

(i) Möglichst lange exakt rechnen

(ii) Mathematische Identitäten verwenden

z.B. $\sqrt{x} - \sqrt{y} = \frac{x-y}{\sqrt{x}+\sqrt{y}}$ oder $1 - \cos y = 2\sin^2(\frac{y}{2})$

Bsp: cos-Satz im Dreieck




$c = \sqrt{a^2 + b^2 - 2ab\cos\gamma}$ (Auslöschg!)

$= \sqrt{(a-b)^2 + 4ab\sin^2(\frac{\gamma}{2})}$ viel besser!

(iii) Taylorreihen einsetzen

Bsp: $\log(1+x) = x - \frac{x^2}{2} + \frac{x^3}{3} + \dots - \frac{(-x)^N}{N}$

Auslöschg für $x \to 0$! $\sinh(x) = \frac{e^x - e^{-x}}{2} = x + \frac{x^3}{3!} + \frac{x^5}{5!} + \dots$

Diagnose: Dieselbe Rechnung mit gestörten Daten wiederholen.

# 7.3. Numerische Differentiation

(durch Differenzenquotienten)

Z.B. gebraucht für Quasi-Newton-Verfahren

Sei $f$ an der Stelle $x$ differenzierbar, und sei ein numerischer Auswertungsalgorithmus für $f$ vorhanden:


$|\delta| \le \mu \cdot eps$, $\mu > 0$, nicht gross, z.B. $\mu \approx 3$

$\tilde{f} = f(x)\cdot(1+\delta)$

(*) $f(x) = \tilde{f} - \delta \cdot f(x) \Leftarrow \tilde{f}$

Approximation der Ableitung $f'(x)$ durch den Differenzenquotienten mit Schritt $h > 0$:

$$f'(x) = \frac{f(x+h) - f(x)}{h} - \frac{h}{2} f''(x) - \cdots$$

$\tilde{f}_1 = (1+\delta_1) f(x+h)$

$\tilde{f}_0 = (1+\delta_0) f(x)$

$$= \frac{1}{h}\left\{\tilde{f}_1 - \tilde{f}_0 - \delta_1 f(x+h) + \delta_0 f(x)\right\} - \frac{h}{2} f''(x) - \cdots$$

mit und (*)

Fehler:

$$err := \frac{1}{h}(\tilde{f}_1 - \tilde{f}_0) - f'(x) = \frac{\delta_1}{h} f(x+h) - \frac{\delta_0}{h} f(x) + \frac{h}{2} f''(x) + \cdots$$

Abschätzung mit Dreiecks-Ungleichung:

$$|err| = \left|\frac{1}{h}(\tilde{f}_1 - \tilde{f}_0) - f'(x)\right| \le \underbrace{\frac{2\mu \cdot eps}{h}|f(x)| + \frac{h}{2}|f''(x)|}$$

$\stackrel{!}{=} \min$ durch gute Wahl von $h$

Resultat: $h = c\sqrt{eps}$ mit $c = 2\sqrt{\mu |f(x)/f''(x)|}$

damit: $|err| \le \sqrt{eps} \cdot \sqrt{\mu |f(x)\cdot f''(x)|}$

Wähle den Differentiationsschritt $h$ in der Grössenordnung von $\sqrt{eps}$. $f'(x)$ hat dann auch die Genauigkeit $\sqrt{eps}$.