2 Computerarithmetik

2.1 Gleitkommazahlen

Mathematische Modelle beschreiben Phänomene quantitativ mittels unendlicher Systeme von Zahlen. Beispiele sind die rationalen Zahlen $\mathbb{Q}$ (abzählbar unendlich) sowie die reellen Zahlen $\mathbb{R}$ (überabzählbar unendlich).

Auf einem Computer stehen bei alphanumerischen Rechnungen immer nur endlich viele, sogenannte **Gleitkommazahlen** zur Verfügung, die wir generisch mit $\mathbb{F}$ (für "floating point numbers") bezeichnen, und die je nach Hersteller und Compiler variieren. Es gilt immer

$$\mathbb{F} \subset \mathbb{Q} \subset \mathbb{R}, \quad |\mathbb{F}| < \infty \,. \tag{2.1.1}$$

Definition 2.1.1 *Gleitkommazahlen sind die Teilmenge $\mathbb{F}$ von $\mathbb{R}$ von Zahlen der Form*

$$x = (-1)^s \cdot (0.a_1 a_2, \ldots, a_t) \cdot \beta^e = (-1)^s \cdot m \cdot \beta^{e-t} \,, \tag{2.1.2}$$

wobei

- $\beta \in \mathbb{N}, \beta \geq 2$ *die Basis der Gleitkommazahl $x \in \mathbb{F}$ ist,*
- $t \in \mathbb{N}$ *die Anzahl der erlaubten signifikanten Stellen a_i von $x \in \mathbb{F}$ ist, mit*

$$0 \leq a_i \leq \beta - 1 \,, \tag{2.1.3}$$

- $m = a_1 a_2 a_3, \ldots, a_t$ *eine ganze Zahl, die sogenannte* ***Mantisse****, ist, mit*

$$0 \leq m \leq \beta^t - 1 \,, \tag{2.1.4}$$

- e *eine ganze Zahl, der* ***Exponent*** *von $x \in \mathbb{F}$ in (2.1.2) ist; er variiert in einem endlichen Intervall, d.h.*

$$L \leq e \leq U \tag{2.1.5}$$

  mit $L < 0$, $U > 0$ ganz, und

- s *das Vorzeichen von $x \in \mathbb{F}$ ist.*

Falls N Speicherpositionen für $x \in \mathbb{F}$ zur Verfügung stehen, gilt die Aufteilung

$s \quad \rightarrow$ eine Position

$m \quad \rightarrow$ t Positionen

$e \quad \rightarrow$ $N - t - 1$ Positionen.

Bemerkung 2.1.2 $x \in \mathbb{F}$ in (2.1.2) ist auch gegeben durch

$$x = (-1)^s \, \beta^e \Big(\frac{a_1}{\beta} + \frac{a_2}{\beta^2} + \cdots + \frac{a_t}{\beta^t} \Big) \,. \tag{2.1.6}$$

Bemerkung 2.1.3 Darstellung (2.1.2) ist nicht eindeutig - um Eindeutigkeit zu erhalten, nehmen wir immer an:

$$a_1 \neq 0\,. \tag{2.1.7}$$

a_1 heisst **führende Stelle**. Dann gilt

$$0 < \beta^{t-1} \leq m \leq \beta^t - 1\,. \tag{2.1.8}$$

Insbesondere ist also dann $x = 0$ nicht in $\mathbb{F}$. Deshalb treffen wir

Konvention 2.1.4 *Die Menge aller $x \in \mathbb{F}$ der Form (2.1.2) ist*

$$\mathbb{F}(\beta, t, L, U) = \{0\} \cup \left\{ x \in \mathbb{R} : x = (-1)^s \beta^e \sum_{i=1}^{t} a_i\, \beta^{-i} \right\}, \tag{2.1.9}$$

die Menge der Gleitkommazahlen mit t signifikanten Stellen, Basis $\beta \geq 2$, Ziffern $0 \leq a_i \leq \beta - 1$ und Exponentenbereich (L, U) mit $L \leq e \leq U$, mit (2.1.7).

Bemerkung 2.1.5 Es gilt

$$x \in \mathbb{F}(\beta, t, L, U) \Longrightarrow -x \in \mathbb{F}(\beta, t, L, U)\,, \tag{2.1.10}$$

$$x_{\min}(\mathbb{F}) := \beta^{L-1} \leq |x| \leq \beta^U (1 - \beta^{-1}) =: x_{\max}(\mathbb{F})\,, \tag{2.1.11}$$

$$|\mathbb{F}(\beta, t, L, U)| = 2(\beta - 1)\,\beta^{t-1}(U - L + 1) + 1\,. \tag{2.1.12}$$

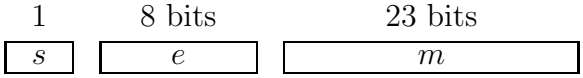
Bemerkung 2.1.6 (Nicht normalisierte Gleitkommazahlen $\mathbb{F}_D$)

Aus (2.1.11) folgt für $x \in \mathbb{R}$ mit $0 < |x| < x_{\min}$, dass $x \notin \mathbb{F}$. Dies kann behoben werden durch Aufgeben von $a_1 \neq 0$ **nur für diese** x. Damit erhält man x der Form (2.1.6) mit $1 \leq m \leq \beta^{t-1} - 1$, $x \in (-\beta^{L-1}, \beta^{L-1})$. Damit ist immer noch die Darstellung (2.1.9) eindeutig und die Menge aller solcher x der Form (2.1.9) heisst $\mathbb{F}_D \supset \mathbb{F}$. Es gilt

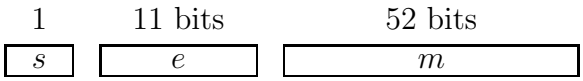
$$0 < x_{min}(\mathbb{F}_D) = \min\left\{ |x| \neq 0 :\ x \in \mathbb{F}_D(\beta, t, L, U) \right\} = \beta^{L-t} < x_{min}(\mathbb{F})\,. \tag{2.1.13}$$

Auf den meisten Rechnern hat man einfach und doppelt genaue Zahlen. Für **binäre Gleitkommazahlen** ($\beta = 2$) ist

$N = 32$ für einfach genaue Zahlen wie folgt verteilt:



$N = 64$ für doppelt genaue Zahlen:



Bemerkung 2.1.7 (**IEEE/IEC Standard**)

Die Gleitkommadarstellung wurde 1985 durch das "Institute of Electronics and Electrical Engineers" (IEEE) entwickelt und 1989 durch die "International Electronical Commission (IEC) als Standard IEC 559 angenommen. Es gilt:

	β	t	L	U
IEEE single	2	24	-125	128
IEEE double	2	53	-1021	1024

und, für die Ausnahmewerte $0, \pm\infty$:

Wert	Exponent	Mantisse
± 0	$L-1$	0
$\pm\infty$	$U+1$	0
NaN	$U+1$	$\neq 0$

2.2 Runden mit Maschinenepsilon

Zwei Zahlen $x, y \in \mathbb{F}$, $x \neq y$, können nicht beliebig nahe zueinander liegen. Es gilt für

$$0 \neq x \in \mathbb{F}: \ \beta^{-1}\,\varepsilon_M\,|x| \leq \min\left\{|x-y|: \ y \in \mathbb{F}\backslash\{0\}\right\} \leq \varepsilon_M|x|\,, \tag{2.2.1}$$

mit dem "Maschinenepsilon" ε_M.

Definition 2.2.1 *(Maschinenepsilon ε_M)*

Die kleinste Zahl $0 < \varepsilon_M \in \mathbb{F}$ mit $1 + \varepsilon_M > 1$ in $\mathbb{F}$ heisst ***Maschinenepsilon****; es gilt*

$$\varepsilon_M = \beta^{1-t}\,. \tag{2.2.2}$$

Das Maschinenepsilon erfüllt $\varepsilon_M = \min\left\{|1-y| : y \in \mathbb{F}\backslash\{1\}\right\}$.

Bemerkung 2.2.2 *Beachte, dass das Maschinenepsilon ε_M* nicht *gleich $x_{min}(\mathbb{F})$, der kleinsten Nichtnullzahl in $\mathbb{F}$, ist. Es gilt vielmehr in der Regel:*

$$0 < x_{min}(\mathbb{F}) << \varepsilon_M(\mathbb{F}).$$

Es stellt sich die Frage, wie man ε_M auf einen gegebenen Rechner oder Compiler findet, wenn z.B. keine Dokumentation verfügbar ist. Der folgende MATLAB code findet für MATLAB double precision den Wert $\varepsilon_M = +2.2204 \cdot 10^{-16}$:

```
e=1; while(1+e>1) e=e/2; end; 2*e
```

Fig. 1: MATLAB code zur Bestimmung von ε_M in MATLAB

Beachte, dass Operationen zwischen $x, y \in \mathbb{F}$ nicht Ergebnisse in $\mathbb{F}$ liefern müssen; es gilt

$$x, y \in \mathbb{F} \text{ impliziert nicht } x \circ y \in \mathbb{F}\,.$$

Hier steht $\circ$ für eine generische Operation, $\circ \in \{+, -, *, /\}$, $\quad \circ : \mathbb{R} \times \mathbb{R} \to \mathbb{R}$.

Abhilfe schafft hier die **Rundung** von $x \circ y$.

Definition 2.2.3 *(Rundung)*

*Sei $\mathbb{F}(\beta, t, L, U)$ eine Menge von Gleitkommazahlen. Die **Rundung** $f\ell$ ist eine Abbildung $f\ell$: $\mathbb{R} \to \mathbb{F}$ definiert für $x \in \mathbb{R}$ in der **normalisierten Positionsdarstellung***

$$0 \neq x = (-1)^s\, \beta^e \sum_{j=1}^{\infty} a_j\, \beta^{-j} \in \mathbb{R}$$

mit Exponent $L \leq e \leq U$ durch

$$f\ell(x) := (-1)^s\, (0.a_1 a_2 \ldots \widetilde{a}_t)\, \beta^e\,, \tag{2.2.3}$$

wobei

$$\widetilde{a}_t := \begin{cases} a_t & \text{für} \quad a_{t+1} < \beta/2\,, \\ a_t + 1 & \text{für} \quad a_{t+1} \geq \beta/2\,. \end{cases}$$

Proposition 2.2.4

$x \in \mathbb{F} \Longrightarrow f\ell(x) = x,$

$x, y \in \mathbb{R} \wedge x \leq y \Longrightarrow f\ell(x) \leq f\ell(y).$

Bemerkung 2.2.5 (Abschneiden) Alternativ zur Rundung kann man auch Abschneiden. Dann ist $f\ell$ definiert wie in (2.2.3), mit $\widetilde{a}_t = a_t$.

Bemerkung 2.2.6 (Überlauf/Unterlauf)

(2.2.3) gilt nur für $x \in \mathbb{R}$ mit Exponent $e \in [L, U]$. Für $x \in (-\infty, -x_{\max}) \cup (x_{\max}, \infty)$ ist $f\ell(x)$ in (2.2.3) nicht definiert.

Sei $x, y \in \mathbb{F}$ und $z = x \circ y \in \mathbb{R}$. Falls

$$|z| = |x \circ y| > x_{\max}(\mathbb{F}) := \max\{|x| : x \in \mathbb{F}\}$$

sprechen wir von **Überlauf**, für

$$|z| = |x \circ y| < x_{\min}(\mathbb{F}) := \min\{|x| : 0 \neq x \in \mathbb{F}\}$$

von **Unterlauf**.

Theorem 2.2.7 *Sei $\mathbb{F} = \mathbb{F}(\beta, t, L, U) \subset \mathbb{R}$ ein Gleitkommazahlsystem und $z \in \mathbb{R}$ gegeben im Bereich von $\mathbb{F}$, d.h. mit*

$$x_{\min}(\mathbb{F}) \le |z| \le x_{\max}(\mathbb{F})\,. \tag{2.2.4}$$

Dann gibt es Zahlen $\delta_i \in \mathbb{R}$ mit

$$f\ell(z) = z(1+\delta_1), \quad |\delta_i| < u := \frac{1}{2}\,\beta^{1-t}, \;\; i = 1, 2, \tag{2.2.5}$$

$$f\ell(z) = z/(1+\delta_2)\,. \tag{2.2.6}$$

Beweis: (2.2.5): Wegen (2.1.10) sei ohne Beschränkung der Allgemeinheit $z > 0$. Dann ist

$$z = m\beta^{e-t}, \;\; \beta^{t-1} \le m < \beta^t - 1\,.$$

Also ist

$$\mathbb{F} \ni z_- := \lfloor m \rfloor\, \beta^{e-t} \le z \le \lceil m \rceil\, \beta^{e-t} =: z_+ \in \mathbb{F}\,,$$

d.h. z liegt zwischen den Gleitkommazahlen $z_-, z_+ \in \mathbb{F}$. Also ist $f\ell(z) \in \{z_-, z_+\}$ und

$$|f\ell(z) - z| \le \frac{|z_+ - z_-|}{2} \le \frac{\beta^{e-t}}{2}\,.$$

Daher folgt

$$\frac{|f\ell(z) - z|}{|z|} \le \frac{\frac{1}{2}\, m\beta^{e-t}}{m\,\beta^{e-t}} = \frac{1}{2}\,\beta^{1-t} =: \mathrm{u}\,.$$

Hier gilt Gleichheit nur dann, wenn $m = \beta^{t-1}$. Dann aber ist $z = f\ell(z) \in \mathbb{F}$, deshalb gilt $|\delta| < \mathrm{u}$. (2.2.6) beweist man analog. □

Bemerkung 2.2.8 Die Zahl u in (2.2.5) erfüllt wegen (2.2.2)

$$\mathrm{u} = \frac{1}{2}\,\beta^{1-t} = \frac{1}{2}\,\varepsilon_M\,. \tag{2.2.7}$$

Sie heisst **Rundungseinheit** (Unit Roundoff) der Gleitkommaarithmetik $\mathbb{F}(\beta, t, L, U)$.

Beispiele für die Werte der Maschinenarithmetik sowie von u enthält die folgende Tabelle:

Machine and arithmetic	β	t	L	U	u
Cray-1 single	2	48	-8192	8191	4×10^{-15}
Cray-1 double	2	96	-8192	8191	1×10^{-29}
DEC VAX G format, double	2	53	-1023	1023	1×10^{-16}
DEC VAX D format, double	2	56	-127	127	1×10^{-17}
HP 28 and 48G calculators	10	12	-499	499	5×10^{-12}
IBM 3090 single	16	6	-64	63	5×10^{-7}
IBM 3090 double	16	14	-64	63	1×10^{-16}
IBM 3090 extended	16	28	-64	63	2×10^{-33}
IEEE single	2	24	-125	128	6×10^{-8}
IEEE double	2	53	-1021	1024	1×10^{-16}
IEEE extended (typical)	2	64	-16381	16384	5×10^{-20}

Tab. 2.1: Floating point arithmetic parameters

2.3 Gleitkommaoperationen

Wir sehen, dass $\circ \in \{+, -, *, /\}$ aus $\mathbb{F}$ hinausführt: $\mathbb{F} \circ \mathbb{F} \notin \mathbb{F}$ im Allgemeinen. Um in $\mathbb{F}$ zu bleiben, holt man das Ergebnis von $\mathbb{F} \circ \mathbb{F}$ wieder zurück nach $\mathbb{F}$ durch Anwendung der Rundung $f\ell$. Die Verknüpfung von $\circ$ mit Rundung führt auf die sogenannten Maschinenoperationen, die wie folgt definiert sind. Der grösseren Allgemeinheit wegen lassen wir gleich Argumente $x, y \in \mathbb{R}$ zu.

Definition 2.3.1 *(Maschinenoperationen)*

Für $x, y \in \mathbb{R}$ mit $x \circ y$ im Bereich von $\mathbb{F}$ heisst

$$x \ \boxed{\circ}\ y := f\ell\big(f\ell(x) \circ f\ell(y)\big) \in \mathbb{F} \tag{2.3.8}$$

Maschinenoperation zu $\circ$.

Für die Analyse von Algorithmen benutzen wir wegen (2.2.5), (2.2.6) das sogenannte **Standardmodell des Rundungfehlers:** für jede Maschinenoperation gilt

$$x \ \boxed{\circ}\ y = (x \circ y)(1 + \delta), \ \ |\delta| \leq u, \ \ \circ = +, -, *, / \,. \tag{2.3.9}$$

Bemerkung 2.3.2 (Petaflop)

Die schnellsten Grossrechner führen bis zu 10^{15} Maschinenoperationen $\boxed{\circ}$ / Sekunde aus. In MATLAB double precision ist $\mathrm{u} = \frac{1}{2}\varepsilon_M = \frac{1}{2}2^{-53} \approx 10^{-16}$, so dass Akkumulation von δ's in (2.3.9) schnell die Grösse 1 ergibt, **falls** (im schlechtesten Fall) bei jeder Operation $\boxed{\circ}$ der maximale Fehler $\delta = \mathrm{u}$ realisiert wird.