Freie Universität Berlin

Computerorientierte Mathematik I

Runden, Maschinengenauigkeit und Gleitkommaarithmetik

Wiederholung: Rationale Zahlen

$$\mathbb{Q} = \left\{ \frac{a}{b} \,\middle|\, a, b \in \mathbb{Z},\ b \neq 0 \right\}$$

Problem:

$x \cdot b = a$ hat i.a. keine Lösung $x \in \mathbb{Z}$.

Ausweg:

Abschluss von $\mathbb{Z}$ unter Division!

Äquivalenzklassen von Paaren (a, b), $a, b \in \mathbb{Z}$, $b \neq 0$

Wiederholung: q-adischer Bruch

Definition: Die Darstellung

$$z_n \cdots z_0, z_{-1} \cdots z_{-m}\, q = \sum_{i=-m}^{n} z_i\, q^i\,,$$

mit $n, m, q \in \mathbb{N}$, $q > 1$ und $z_i \in \{0, \ldots, q-1\}$, heißt q-adischer Bruch.

Beispiel: Dezimalbruch $\rightarrow$ Dualbruch

$$1,375 = 1,011_2$$

Wiederholung: Reelle Zahlen

Darstellung: unendliche Dezimalbrüche

$$\mathbb{R} = \{z_n \cdots z_0, z_{-1}z_{-2}\cdots \,|\, z_i = 0, \ldots, 9\}$$

Satz: $\mathbb{R}$ ist nicht abzählbar!

Folgerung: Es gibt keine Zifferndarstellung von $\mathbb{R}$!

Folgerung: Numerisches Rechnen mit reellen Zahlen ist nicht möglich!

Wiederholung: Absoluter und relativer Fehler

Definition: Es sei $\tilde{x}$ eine Approximation von $x \in \mathbb{R}$. Dann heißt der Abstand

$$|x - \tilde{x}|$$

absoluter Fehler.

Definition: Es sei $\tilde{x}$ eine Approximation von $x \in \mathbb{R}$, $x \neq 0$. Dann heißt das Verhältnis

$$\frac{|x - \tilde{x}|}{|x|}$$

relativer Fehler.

Wiederholung: Festkommazahlen

Definition: Seien $n, m \in \mathbb{N}$ fest gewählt. Dann ist

$$z_{n-1} \cdots z_0, z_{-1} \cdots z_{-m}\, q = \sum_{i=-m}^{n-1} z_i q^i \,,$$

mit $z_i \in \{0, \ldots, q-1\}$, eine Festkommazahl mit $\ell = n + m$ Stellen.

Nicht optimal!

Wiederholung: Gleitkommazahlen

Definition: Jede in der Form

$$\tilde{x} = (-1)^s \cdot a \cdot q^e$$

mit Vorzeichenbit $s \in \{0, 1\}$, Exponent $e \in \mathbb{Z}$ und Mantisse $a = 0$ oder

$$a = 0, a_1 \cdots a_\ell = \sum_{i=1}^{\ell} a_i q^{-i}, \quad a_i \in \{0, \ldots, q-1\},\ a_1 \neq 0$$

darstellbare Zahl $\tilde{x}$ heißt Gleitkommazahl mit Mantissenlänge $\ell \in \mathbb{N},\ \ell \geq 1$. Die Darstellung heißt normalisierte Gleitkommadarstellung. Die Menge aller dieser Zahlen heißt $\mathbb{G}(q, \ell)$.

Runden

Normalisierte Darstellung:

$$x = a^* q^e\,, \quad e \in \mathbb{Z}\,, \quad q^{-1} \le a^* < q$$

Unendlicher q-adischer Bruch:

$$a^* = 0, a_1 a_2 \cdots a_\ell a_{\ell+1} \cdots = \sum_{k=1}^{\infty} a_k q^{-k}\,, \quad a_k \in \{0, \ldots, q-1\}$$

Runden:

$$a = \sum_{k=1}^{\ell} a_k q^{-k} + \begin{cases} 0\,, & \text{falls } a_{\ell+1} < \frac{1}{2} q \\ q^{-\ell}\,, & \text{falls } a_{\ell+1} \ge \frac{1}{2} q \end{cases}$$

Runden: Absoluter Fehler

Satz: Zu jedem $N \in \mathbb{N}$ gibt es ein $x \in \mathbb{R}$, so dass

$$|x - \mathrm{rd}(x)| \geq q^N \,.$$

Beweis. Wähle

$$x = 0, z_1 \cdots z_\ell z_{\ell+1} \cdot q^{\ell+1+N}$$

mit $z_1, z_{\ell+1} \neq 0$ und $N \in \mathbb{N}$. □

Der absolute Rundungsfehler kann beliebig groß werden!

Runden: Relativer Fehler

Satz: Es sei q eine gerade Zahl. Dann gilt

$$\frac{|x - \mathrm{rd}(x)|}{|x|} \leq \frac{1}{2} q^{1-\ell} =: \mathrm{eps}(q, \ell).$$

Die Zahl $\mathrm{eps}(q, \ell)$ heißt Maschinengenauigkeit.

Runden: Relativer Fehler

Beweisskizze. Sei O.B.D.A $x > 0$ und $a_{\ell+1} \geq \frac{q}{2}$ (aufrunden).

$$x = 0, a_1 \cdots a_\ell a_{\ell+1} \cdot q^e, \qquad a_1 \neq 0$$

$$\mathrm{rd}(x) = (0, a_1 \cdots a_\ell + \delta) \cdot q^e, \quad \delta = q^{-\ell}$$

$$\frac{|x - \mathrm{rd}(x)|}{|x|} \leq \frac{(\delta - a_{\ell+1} \cdot q^{-\ell-1}) \cdot q^e}{q^{-1} \cdot q^e} \leq \frac{q^{-\ell} - \frac{q}{2} \cdot q^{-\ell-1}}{q^{-1}} = \frac{1}{2} q^{1-\ell}$$

□

(Mit Zahlen Rechnen, Satz 1.11)

Maschinengenauigkeit

Der relative Rundungsfehler ist durch die Maschinengenauigkeit beschränkt!

$$\text{Mantissenlänge } \ell \quad \Leftrightarrow \quad \ell \text{ gültige Stellen}$$

$$\Leftrightarrow \quad \text{eps}(q, \ell) = \tfrac{1}{2} q^{1-\ell}$$

Praktische Realisierung (IEEE 754 Standard)

$$\boxed{s \,|\, e_{M-1} \cdots e_0 \,|\, a_2 \cdots a_\ell} \hat{=} (-1)^s a \cdot 2^e$$

Exponent	Mantisse	Bedeutung
$111\ldots111$	$000\ldots000$	$+/-\infty$
$111\ldots111$	$\neq 000\ldots000$	NaN, „Not a Number" (keine Zahl)
$000\ldots000$	$000\ldots000$	$+/-0.0$, Null
$000\ldots000$	$\neq 000\ldots000$	Denormalisierte Fließkommazahl
$000\ldots001$ bis $111\ldots110$	beliebig	Normalisierte Fließkommazahl

$$\frac{0}{0} \mapsto NaN\,, \quad \pm\frac{1}{0} \mapsto \pm\infty$$

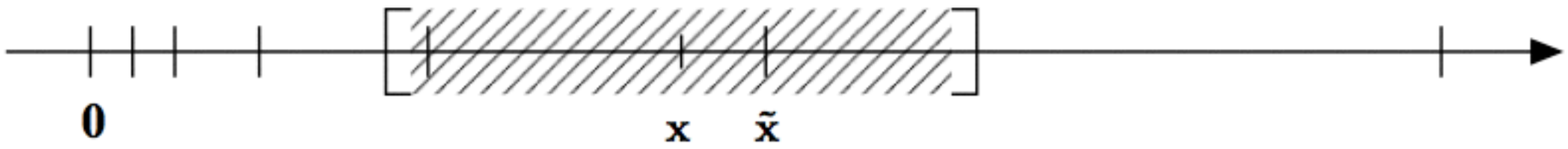
Praktische Realisierung (IEEE 754 Standard)

	Float	Double
Länge in Bits	32	64
Exponent		
Bits	8	11
Bias	127	1023
$e_{\min}$	-126	-1022
$e_{\max}$	128	1024
Mantisse		
Bits	23	52
normalisierte Gleitkommazahl (Exponent $= 00\ldots01$ bis $11\ldots10$)		
$x_{\min}$	$1,2 \cdot 10^{-38}$	$2,2 \cdot 10^{-308}$
$x_{\max}$	$3,4 \cdot 10^{+38}$	$1,8 \cdot 10^{+308}$
Maschinengenauigkeit *eps*	$6,0 \cdot 10^{-8}$	$1,1 \cdot 10^{-16}$
denormalisierte Gleitkommazahl (Exponent $= 00\ldots00$, Mantisse $\neq 00\ldots00$)		
$x_{\min}$	$1,4 \cdot 10^{-45}$	$4,9 \cdot 10^{-324}$

Zahlenmengen statt Zahlen

Menge aller Approximationen $\tilde{x}$ von x auf ℓ gültige Stellen im q-System:

$$\mathrm{rd}(x) \in \{\tilde{x} \in \mathbb{R} \mid \tilde{x} = x(1+\epsilon)\,,\ |\epsilon| \leq \mathrm{eps}(q,\ell)\}$$



Menge aller $x \in \mathbb{R}$, die auf $\tilde{x} = \mathrm{rd}(x) \in \mathbb{G}(q,\ell)$ gerundet werden:

$$R(\tilde{x}) = \{x \in \mathbb{R} \mid \tilde{x} = \mathrm{rd}(x)\}$$

Zahlenmengen statt Zahlen

Satz: Es sei q eine gerade Zahl und

$$\tilde{x} = a \cdot q^e \in \mathbb{G}(q, \ell)\,, \quad q^{-1} < a_0, a_1 \cdots a_\ell \leq 1\,.$$

Dann gilt $R(\tilde{x}) = [\alpha(\tilde{x}), \beta(\tilde{x}))$ mit

$$\begin{aligned} \alpha(\tilde{x}) &= \tilde{x} - q^{e-1} \cdot \mathrm{eps}(q, \ell)\,, \\ \beta(\tilde{x}) &= \tilde{x} + q^{e-1+a_0} \cdot \mathrm{eps}(q, \ell)\,. \end{aligned}$$

(Mit Zahlen Rechnen, Satz 1.12)

Zahlenmengen statt Zahlen

Folgerung: Die Abfrage `if` $\tilde{\mathtt{x}} == \tilde{\mathtt{y}}$, mit $\tilde{x}, \tilde{y} \in \mathbb{G}(q, \ell)$, ist sinnlos!

$$\tilde{x} = \tilde{y} \;\not\Rightarrow\; x = y\,, \quad \tilde{x} = \mathrm{rd}(x)\,,\ \tilde{y} = \mathrm{rd}(y)$$

Umgekehrt gilt

$$x = y \;\Rightarrow\; \mathrm{rd}(x) = \mathrm{rd}(y)$$

Gleichheitsabfragen von Gleitkommazahlen vermeiden!

Zahlenmengen statt Zahlen

Satz: Es gilt

$$\mathrm{rd}(x) = \mathrm{rd}(y) \Rightarrow |x - y| \leq (|x| + |y|) \cdot \mathrm{eps}(q, \ell)\,.$$

Umgekehrt gibt es zu jedem $\epsilon > 0$ reelle Zahlen x, y mit den Eigenschaften

$$|x - y| \leq \epsilon \quad \text{und} \quad \mathrm{rd}(x) \neq \mathrm{rd}(y)\,.$$

(Mit Zahlen Rechnen, Satz 1.13)

Algebraische Eigenschaften

Grundrechenarten führen aus $\mathbb{G} = \mathbb{G}(q, \ell)$ hinaus:

$$\tilde{x}, \tilde{y} \in \mathbb{G} \;\not\Rightarrow\; \tilde{x} + \tilde{y} \in \mathbb{G}\,.$$

Gleitkommaarithmetik: $\{+, -, \cdot, /\} \to \{\tilde{+}, \tilde{-}, \tilde{\cdot}, \tilde{/}\}$

$$\begin{aligned}
\tilde{x} \tilde{+} \tilde{y} &= \mathrm{rd}\,(\tilde{x} + \tilde{y}) \\
\tilde{x} \tilde{-} \tilde{y} &= \mathrm{rd}\,(\tilde{x} - \tilde{y}) \\
\tilde{x} \tilde{\cdot} \tilde{y} &= \mathrm{rd}\,(\tilde{x} \cdot \tilde{y}) \\
\tilde{x} \tilde{/} \tilde{y} &= \mathrm{rd}\,(\tilde{x} / \tilde{y})\ ,\ \tilde{y} \neq 0
\end{aligned}$$

Computer können nicht rechnen!

Die Gleitkommazahlen mit Gleitkommaarithmetik sind kein Körper!

Addition und Multiplikation sind

- nicht assoziativ
- nicht distributiv

I.a. kein Inverses bzgl. der Multiplikation

Computer können nicht rechnen!

Beispiel: Äquivalente Umformungen in $\mathbb{R}$ sind in Gleitkommaarithmetik nicht äquivalent!

Keine binomische Formel:

$$(a \tilde{+} b) \tilde{\cdot} (a \tilde{+} b) \neq a \tilde{\cdot} a \tilde{+} 2 \tilde{\cdot} a \tilde{\cdot} b \tilde{+} b \tilde{\cdot} b$$

Kein Assoziativgesetz:

$$(a \tilde{\cdot} a \tilde{+} 2 \tilde{\cdot} a \tilde{\cdot} b) \tilde{+} b \tilde{\cdot} b \neq a \tilde{\cdot} a \tilde{+} (2 \tilde{\cdot} a \tilde{\cdot} b \tilde{+} b \tilde{\cdot} b)$$

Computer können nicht rechnen!

“It is hard, but it’s harder to ignore it!”

—Cat Stevens

Kondition: Auswirkung von Eingabefehlern auf das Ergebnis