COMPUTERORIENTIERTE MATHEMATIK I

Christof Schütte
Wintersemester 2016/17

Computer-orientierte Mathematik

3. Vorlesung - Christof Schuette

11.11.16

Memo: Rationale und reelle Zahlen

Rationale Zahlen:

Rationale Zahlen als Brüche ganzer Zahlen.
q-adische Brüche, periodische q-adische Brüche. Beispiele.
Satz: Jede rationale Zahl ist als periodischer q-adischer Bruch darstellbar.
Keine Eindeutigkeit! Beispiele.
Praktische Realisierung:
Dynamische Ziffernzahl. Aufwand pro Addition problemabhängig.

Reelle Zahlen:

Reelle Zahlen als unendliche q-adische Brüche.
Abzählbarkeit und Zifferndarstellung.
Satz: $\mathbb{R}$ ist nicht abzählbar.
Folgerung: Es gibt keine Zifferndarstellung von $\mathbb{R}$.
Konsequenz: Numerisches Rechnen mit reellen Zahlen ist nicht möglich!

Memo: Gleitkommazahlen $\mathbb{G}(\ell, q)$

Definition: (Gleitkommazahlen) Jede in der Form

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

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

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

oder $a = 0$ darstellbare Zahl $\tilde{x}$ heißt Gleitkommazahl mit *Mantissenlänge* $\ell \in \mathbb{N}$, $\ell \geq 1$.

Die Menge all dieser Zahlen heißt $\mathbb{G}(q, \ell)$.

Die Darstellung (1) heißt normalisierte Gleitkommadarstellung.

Approximation durch Runden

normalisierte Darstellung: einer reellen Zahl $0 < x \in \mathbb{R}$:

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

Approximation durch Runden

normalisierte Darstellung: einer reellen Zahl $0 < x \in \mathbb{R}$:

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

unendlicher q-adischer Bruch: (Achtung: Eindeutigkeit)

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

Approximation durch Runden

normalisierte Darstellung: einer reellen Zahl $0 < x \in \mathbb{R}$:

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

unendlicher q-adischer Bruch: (Achtung: Eindeutigkeit)

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

Runden: $\tilde{x} = \mathrm{rd}(x) := aq^e$

$$a = \sum_{i=1}^{\ell} a_i q^{-i} + \begin{cases} 0 & \text{falls } a_{\ell+1} < \frac{1}{2}q \\ q^{-\ell} & \text{falls } a_{\ell+1} \geq \frac{1}{2}q \end{cases}$$

Fehlerabschätzung: 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:
Die Behauptung folgt durch Wahl von

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

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

Der absolute Rundungsfehler kann beliebig groß werden.

Fehlerabschätzung: Relativer Fehler

Satz: Es sei q eine gerade Zahl. Dann gilt

$$\frac{|x - \mathrm{rd}(x)|}{|x|} \leq \tfrac{1}{2} q^{-(\ell-1)} =: eps(q, \ell) \qquad \forall x \in \mathbb{R},\ x \neq 0\ .$$

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

Fehlerabschätzung: Relativer Fehler

Satz: Es sei q eine gerade Zahl. Dann gilt

$$\frac{|x - \mathrm{rd}(x)|}{|x|} \leq \tfrac{1}{2} q^{-(\ell-1)} =: eps(q,\ell) \qquad \forall x \in \mathbb{R},\ x \neq 0\ .$$

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

Beweis: O.B.d.A.: $x > 0$. Annahme: $a_{\ell+1} > q/2$

$$x \;=\; 0,a_1\ a_2 \cdots a_\ell\ a_{\ell+1} \ldots \cdot q^e, \quad a_1 \neq 0$$

Fehlerabschätzung: Relativer Fehler

Satz: Es sei q eine gerade Zahl. Dann gilt

$$\frac{|x - \mathrm{rd}(x)|}{|x|} \leq \tfrac{1}{2} q^{-(\ell-1)} =: eps(q,\ell) \qquad \forall x \in \mathbb{R},\ x \neq 0\ .$$

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

Beweis: O.B.d.A.: $x > 0$. Annahme: $a_{\ell+1} > q/2$

$$\begin{aligned} x &= 0,a_1\ a_2 \cdots a_\ell\ a_{\ell+1} \ldots \cdot q^e, \quad a_1 \neq 0 \\ \mathrm{rd}(x) &= (0,a_1\ a_2 \cdots a_\ell + \delta) \cdot q^e \\ \delta &= q^{-\ell} \end{aligned}$$

Fehlerabschätzung: Relativer Fehler

Satz: Es sei q eine gerade Zahl. Dann gilt

$$\frac{|x - \mathrm{rd}(x)|}{|x|} \leq \tfrac{1}{2} q^{-(\ell-1)} =: eps(q,\ell) \qquad \forall x \in \mathbb{R},\ x \neq 0\ .$$

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

Beweis: O.B.d.A.: $x > 0$. Annahme: $a_{\ell+1} > q/2$

$$\begin{aligned}
x &= 0,a_1\ a_2 \cdots a_\ell\ a_{\ell+1} \ldots \cdot q^e, \quad a_1 \neq 0 \\
\mathrm{rd}(x) &= (0,a_1\ a_2 \cdots a_\ell + \delta) \cdot q^e \\
\delta &= q^{-\ell} \\
\frac{|x - \mathrm{rd}(x)|}{|x|} &\leq \frac{(\delta - a_{\ell+1} q^{-(\ell+1)}) \cdot q^e}{q^{-1} \cdot q^e} \leq \frac{\delta - \frac{1}{2} q \cdot q^{-(\ell+1)}}{q^{-1}} = \tfrac{1}{2} q \cdot q^{-\ell}
\end{aligned}$$

Fehlerabschätzung: Relativer Fehler

Der relative Rundungsfehler ist durch $eps(q, \ell)$ beschränkt.

Praktische Realisierung

endlicher Exponenten-Bereich:

$$e \in [e_{\min}, e_{\max}] \cap \mathbb{Z}$$

endlicher Zahlen-Bereich:

$$x_{\min} := q^{e_{\min}-1} \leq |\tilde{x}| \leq (1 - q^{-\ell}) q^{e_{\max}} =: x_{\max}$$

- $x < x_{\min}$: `underflow` oder $x = 0$
- $x > x_{\max}$: `overflow` oder $x = NaN$

IEEE 754- Standard

	Float	Double
Länge in Bits	32	64
Exponent		
Bits	8	11
e_{min}	-126	-1022
e_{max}	128	1024
Mantisse		
Bits	23	52
normalisierte Gleitkommazahl		
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*	$6,0 \cdot 10^{-8}$	$1,1 \cdot 10^{-16}$

Zahlenmengen statt Zahlen

$$\frac{|x - \mathrm{rd}(x)|}{|x|} \leq eps(q, \ell)$$

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

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

Zahlenmengen statt Zahlen

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

Äquivalenzrelation:

$$x \equiv y \quad \Leftrightarrow \quad \mathrm{rd}(x) = \mathrm{rd}(y)$$

Äquivalenzklassen

$$[\tilde{x}] = \{x \in \mathbb{R} \mid \tilde{x} = \mathrm{rd}(x)\} \, .$$

Zahlenmengen statt Zahlen

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

Äquivalenzrelation:

$$x \equiv y \quad \Leftrightarrow \quad \mathrm{rd}(x) = \mathrm{rd}(y)$$

Äquivalenzklassen

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

Satz: Es sei

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

Dann gilt $[\tilde{x}] = [\ \tilde{x} - q^{e-1} eps,\ \tilde{x} + q^{e-1} eps\ )$

Algebraische Eigenschaften

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

$$\tilde{x}, \tilde{y} \in \mathbb{G} \quad \not\Rightarrow \quad \tilde{x} + \tilde{y} \in \mathbb{G}\,, \qquad \text{analog: } -, \cdot, /$$

Algebraische Eigenschaften

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

$$\tilde{x}, \tilde{y} \in \mathbb{G} \quad \not\Rightarrow \quad \tilde{x} + \tilde{y} \in \mathbb{G}\,, \qquad \text{analog: } -, \cdot, /$$

Gleitkommaarithmetik:

$$\tilde{x} \tilde{+} \tilde{y} = \mathrm{rd}(\tilde{x} + \tilde{y}), \qquad \tilde{x} \tilde{-} \tilde{y} = \mathrm{rd}(\tilde{x} - \tilde{y}),$$

$$\tilde{x} \tilde{*} \tilde{y} = \mathrm{rd}(\tilde{x}\tilde{y}), \qquad \tilde{x} \tilde{:} \tilde{y} = \mathrm{rd}(\tilde{x} : \tilde{y}),\ \tilde{y} \neq 0$$

Algebraische Eigenschaften

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

$$\tilde{x}, \tilde{y} \in \mathbb{G} \quad \not\Rightarrow \quad \tilde{x} + \tilde{y} \in \mathbb{G}\,, \qquad \text{analog: } -, \cdot, /$$

Gleitkommaarithmetik:

$$\tilde{x} \tilde{+} \tilde{y} = \mathrm{rd}(\tilde{x} + \tilde{y}), \qquad \tilde{x} \tilde{-} \tilde{y} = \mathrm{rd}(\tilde{x} - \tilde{y}),$$

$$\tilde{x} \tilde{*} \tilde{y} = \mathrm{rd}(\tilde{x}\tilde{y}), \qquad \tilde{x} \tilde{:} \tilde{y} = \mathrm{rd}(\tilde{x} : \tilde{y}),\ \tilde{y} \neq 0$$

Die Gleitkommazahlen mit Gleitkommaarithmetik sind <u>kein</u> Körper.

Das Kernproblem

Übliche Umformungen sind nicht mehr äquivalent.

Beispiel: binomische Formel

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

sogar

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

Gleichheits-Abfragen von Gleitkomma-Zahlen

Bemerkung:

Die Abfrage `if x̃ == ỹ` ist sinnlos!

$$\tilde{x} = \tilde{y} \quad \not\Longrightarrow \quad x = y, \qquad \tilde{x} = \mathrm{rd}(x),\ \tilde{y} = \mathrm{rd}(y)$$

Gleichheits-Abfragen von Gleitkomma-Zahlen

Bemerkung:

Die Abfrage `if x̃ == ỹ` ist sinnlos!

$$\tilde{x} = \tilde{y} \quad \not\Longrightarrow \quad x = y, \qquad \tilde{x} = \mathrm{rd}(x),\ \tilde{y} = \mathrm{rd}(y)$$

umgekehrt: $x = y \Longrightarrow \mathrm{rd}(x) = \mathrm{rd}(y)$ aber

$$x = a + b, y = x \quad \not\Longrightarrow \quad \tilde{x} = \tilde{y}, \qquad \tilde{x} = \mathrm{rd}(a) \tilde{+} \mathrm{rd}(b)$$

Gleichheits-Abfragen von Gleitkomma-Zahlen

Bemerkung:

Die Abfrage `if x̃ == ỹ` ist sinnlos!

$$\tilde{x} = \tilde{y} \quad \not\Longrightarrow \quad x = y, \qquad \tilde{x} = \mathrm{rd}(x),\ \tilde{y} = \mathrm{rd}(y)$$

umgekehrt: $x = y \Longrightarrow \mathrm{rd}(x) = \mathrm{rd}(y)$ aber

$$x = a + b, y = x \quad \not\Longrightarrow \quad \tilde{x} = \tilde{y}, \qquad \tilde{x} = \mathrm{rd}(a) \tilde{+} \mathrm{rd}(b)$$

Gleichheits-Abfragen von Gleitkomma-Zahlen vermeiden!

Ausblick: Auswertung eines Polynoms mit Matlab

Berechne das Polynom $f(x) = x^3 + 12a^2x - 6ax^2 - 8a^3$
mit $a = 4\,999\,999$ an der Stelle $x_0 = 10\,000\,000$.

```
» a = 4999999;
» x = 10000000;
» f = x^3 + 12 * a^2 * x - 6 * a * x^2 - 8 * a^3

f =
   393216
```

Ausblick: Auswertung eines Polynoms mit Matlab

Berechne das Polynom $f(x) = x^3 + 12a^2x - 6ax^2 - 8a^3$
mit $a = 4\,999\,999$ an der Stelle $x_0 = 10\,000\,000$.

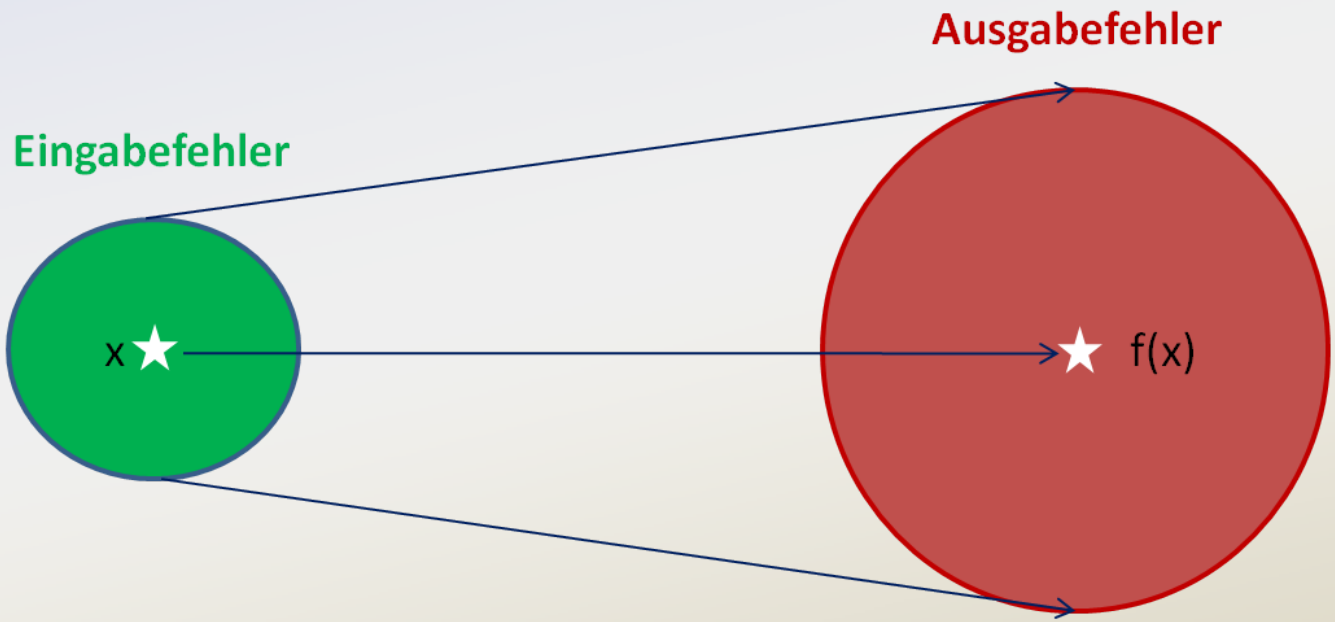
```
» a = 4999999;
» x = 10000000;
» f = x^3 + 12 * a^2 * x - 6 * a * x^2 - 8 * a^3

f =
   393216

» f = (x - 2 * a)^3
f =
   8
```

Was ist hier schief gelaufen?

Auswirkung von Eingabefehlern auf das Ergebnis



Relative Kondition der Multiplikation

Gegeben: $x, y \in \mathbb{R}$, x, $y \neq 0$.

Approximationen mit relativem Fehler ε:

$$\tilde{x} = x(1+\varepsilon_x), \quad \tilde{y} = y(1+\varepsilon_y), \qquad \varepsilon = \max\{|\varepsilon_x|, |\varepsilon_y|\}$$

Relative Kondition der Multiplikation

Gegeben: $x, y \in \mathbb{R}$, x, $y \neq 0$.
Approximationen mit relativem Fehler ε:

$$\tilde{x} = x(1+\varepsilon_x), \quad \tilde{y} = y(1+\varepsilon_y), \qquad \varepsilon = \max\{|\varepsilon_x|, |\varepsilon_y|\}$$

Satz: Es gilt

$$\frac{|(x \cdot y) - (\tilde{x} \cdot \tilde{y})|}{|x \cdot y|} \leq 2\,\varepsilon + \varepsilon^2 \,.$$

Relative Kondition der Multiplikation

Gegeben: $x, y \in \mathbb{R}$, x, $y \neq 0$.
Approximationen mit relativem Fehler ε:

$$\tilde{x} = x(1+\varepsilon_x), \quad \tilde{y} = y(1+\varepsilon_y), \qquad \varepsilon = \max\{|\varepsilon_x|, |\varepsilon_y|\}$$

Satz: Es gilt

$$\frac{|(x \cdot y) - (\tilde{x} \cdot \tilde{y})|}{|x \cdot y|} \leq 2\,\varepsilon + \varepsilon^2 \,.$$

Dominierender Fehleranteil: 2ε

Relative Kondition der Multiplikation

Gegeben: $x, y \in \mathbb{R}$, $x,\ y \neq 0$.
Approximationen mit relativem Fehler ε:

$$\tilde{x} = x(1+\varepsilon_x), \quad \tilde{y} = y(1+\varepsilon_y), \qquad \varepsilon = \max\{|\varepsilon_x|, |\varepsilon_y|\}$$

Satz: Es gilt

$$\frac{|(x \cdot y) - (\tilde{x} \cdot \tilde{y})|}{|x \cdot y|} \leq 2\,\varepsilon + \varepsilon^2\,.$$

Dominierender Fehleranteil: 2ε
Die relative Kondition ist der Verstärkungsfaktor κ von ε: $\kappa = 2$

Relative Kondition der Multiplikation

Gegeben: $x, y \in \mathbb{R}$, $x,\ y \neq 0$.
Approximationen mit relativem Fehler ε:

$$\tilde{x} = x(1+\varepsilon_x), \quad \tilde{y} = y(1+\varepsilon_y), \qquad \varepsilon = \max\{|\varepsilon_x|, |\varepsilon_y|\}$$

Satz: Es gilt

$$\frac{|(x \cdot y) - (\tilde{x} \cdot \tilde{y})|}{|x \cdot y|} \leq 2\,\varepsilon + \varepsilon^2\,.$$

Dominierender Fehleranteil: 2ε
Die relative Kondition ist der Verstärkungsfaktor κ von ε: $\kappa = 2$
Vernachlässigung des *Terms höherer Ordnung* ε^2

Definition (Landau-Symbol o)
Sei $f : I \to \mathbb{R}$ mit $I = (-a, a)$ eine Funktion. Wir verabreden die Schreibweise

$$\lim_{\varepsilon \to 0} \frac{f(\varepsilon)}{\varepsilon} = 0 \quad \Longleftrightarrow \quad f(\varepsilon) = o(\varepsilon) \quad (\text{für } \varepsilon \to 0)\,.$$

Definition (Landau-Symbol o)
Sei $f : I \to \mathbb{R}$ mit $I = (-a, a)$ eine Funktion. Wir verabreden die Schreibweise

$$\lim_{\varepsilon \to 0} \frac{f(\varepsilon)}{\varepsilon} = 0 \quad \Longleftrightarrow \quad f(\varepsilon) = o(\varepsilon) \quad (\text{für } \varepsilon \to 0)\,.$$

Beispiele:

$\varepsilon^2 = o(\varepsilon)$

Definition (Landau-Symbol o)
Sei $f : I \to \mathbb{R}$ mit $I = (-a, a)$ eine Funktion. Wir verabreden die Schreibweise

$$\lim_{\varepsilon \to 0} \frac{f(\varepsilon)}{\varepsilon} = 0 \quad \Longleftrightarrow \quad f(\varepsilon) = o(\varepsilon) \quad (\text{für } \varepsilon \to 0)\,.$$

Beispiele:

$$\varepsilon^2 = o(\varepsilon) \qquad \varepsilon\sqrt{\varepsilon} + \varepsilon \sum_{i=1}^{28} (\sin(\varepsilon))^i = o(\varepsilon), \qquad \ldots$$