Darstellung rationaler und reeller Zahlen Vorlesung vom 30.10.15

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.

Eindeutigkeit durch $0,\overline{9}$ statt 1.

Praktische Realisierung: Dynamische Ziffernzahl. Aufwand pro Addition problemabhängig. (Hauptnenner, Kürzen).

Reelle Zahlen:

Reelle Zahlen als unendliche q-adische Brüche.

Satz: $\mathbb{R}$ ist nicht abzählbar. Folgerung: Es gibt keine Zifferndarstellung von $\mathbb{R}$.

Konsequenz: Numerisches Rechnen mit reellen Zahlen ist nicht möglich!

Festkommazahlen:

Absoluter und relativer Fehler. Beispiele.

Definition von Festkommazahlen und Gleitkommazahlen. Beispiele.

Festkommazahlen

$$z_{n-1}\, z_{n-2} \cdots z_0, z_{-1} \cdots z_{-m} = \sum_{i=-m}^{n-1} z_i\, q^i, \qquad z_i \in \{0, \dots, q-1\}\ .$$

$\ell = m + n$ Stellen verfügbar; $\quad n,\ m \in \mathbb{N}$ fest gewählt.

Beispiel: $q = 10$, $\ell = 4$, $n = 3$, $m = 1$

$x = 0{,}123$, Runden: $\tilde{x} = 0{,}1$ $\quad$ relativer Fehler: $|x - \tilde{x}|/|x| \approx 0.2$

$x = 123$, exakt darstellbar: $\tilde{x} = 123$ $\quad$ relativer Fehler: $|x - \tilde{x}|/|x| = 0$

Folgerung:
Im Sinne einer optimalen Stellenausnutzung n, m variabel halten!

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* $a = 0$ oder

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

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:

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

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:

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

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: $\tilde{x} = \mathrm{rd}(x) := a q^e$

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

Fehlerabschätzung: Absoluter Fehler

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

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

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:

Wähle $x = 0, a_1 \cdots a_\ell \, a_{\ell+1} \cdot q^{\ell+1+N}$ mit a_1, $a_{\ell+1} \neq 0$ und $N \in \mathbb{N}$.

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:

Wähle $x = 0, a_1 \cdots a_\ell \, a_{\ell+1} \cdot q^{\ell+1+N}$ mit a_1, $a_{\ell+1} \neq 0$ und $N \in \mathbb{N}$.

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

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

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

Beweisskizze: O.B.D.A. $x > 0$. Aufrunden: $a_{\ell+1} \geq \frac{1}{2} q$

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.

Beweisskizze: O.B.D.A. $x > 0$. Aufrunden: $a_{\ell+1} \geq \frac{1}{2}q$

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

Fehlerabschätzung: Relativer Fehler

Satz: Es sei q eine gerade Zahl. Dann gilt

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

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

Beweisskizze: O.B.D.A. $x > 0$. Aufrunden: $a_{\ell+1} \ge \tfrac{1}{2} q$

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

Fehlerabschätzung: Relativer Fehler

Satz: Es sei q eine gerade Zahl. Dann gilt

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

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

Beweisskizze: O.B.D.A. $x > 0$. Aufrunden: $a_{\ell+1} \geq \frac{1}{2} q$

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

$$\frac{|x - \mathrm{rd}(x)|}{|x|}$$

Fehlerabschätzung: Relativer Fehler

Satz: Es sei q eine gerade Zahl. Dann gilt

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

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

Beweisskizze: O.B.D.A. $x > 0$. Aufrunden: $a_{\ell+1} \geq \frac{1}{2} q$

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

Fehlerabschätzung: Relativer Fehler

Satz: Es sei q eine gerade Zahl. Dann gilt

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

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

Beweisskizze: O.B.D.A. $x > 0$. Aufrunden: $a_{\ell+1} \geq \frac{1}{2}q$

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

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.

Beweisskizze: O.B.D.A. $x > 0$. Aufrunden: $a_{\ell+1} \geq \frac{1}{2}q$

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

Maschinengenauigkeit

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

Mantissenlänge $\ell \iff \ell$ gültige Stellen $\iff eps(q,\ell) = \frac{1}{2}q^{-(\ell-1)}$

Praktische Realisierung

endlicher Exponenten-Bereich:

$$e \in \{e_{\min}, e_{\min}+1, \ldots, e_{\max}-1, e_{\max}\}$$

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
Vorzeichen s		
Bits	1	1
Exponent e		
Bits	8	11
Mantisse a		
Bits	23	52
Maschinengenauigkeit eps	$6,0 \cdot 10^{-8}$	$1,1 \cdot 10^{-16}$
$e_{\min}$	- 126	- 1022
$e_{\max}$	128	1024
$x_{\min}$	$1,2 \cdot 10^{-38}$	$2,2 \cdot 10^{-308}$
$x_{\max}$	$3,4 \cdot 10^{+38}$	$1,8 \cdot 10^{+308}$

Zahlenmengen statt Zahlen

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

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

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

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

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

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

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

Satz: Es sei q eine gerade Zahl und

$$\tilde{x} = aq^e \in \mathbb{G}(q,\ell), \quad q^{-1} < 0, a_1 \cdots a_\ell \leq 1\ .$$

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

$$\alpha(\tilde{x}) = \tilde{x} - q^{e-1} eps\ ,\ \ \beta(\tilde{x}) = \tilde{x} + q^{e-1} eps\ .$$

Gleichheits-Abfragen von Gleitkomma-Zahlen

Folgerung: Die Abfrage `if x̃ == ỹ` mit $\tilde{x},\ \tilde{y} \in \mathbb{G}(q,\ell)$ 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

Folgerung: Die Abfrage `if x̃ == ỹ` mit $\tilde{x},\ \tilde{y} \in \mathbb{G}(q, \ell)$ 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

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

Gleichheits-Abfragen von Gleitkomma-Zahlen

Folgerung: Die Abfrage `if x̃ == ỹ` mit $\tilde{x}, \ \tilde{y} \in \mathbb{G}(q, \ell)$ 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

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

Gleichheits-Abfragen von Gleitkomma-Zahlen verboten!

Praxisbeispiel

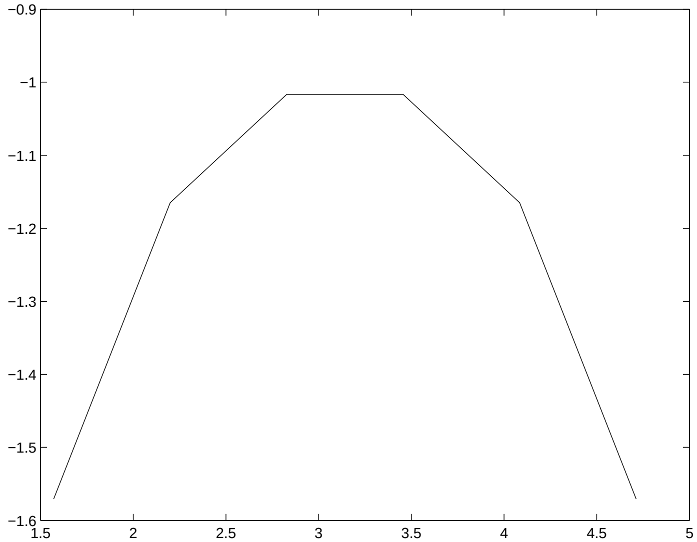
Aufgabe: Plotten Sie

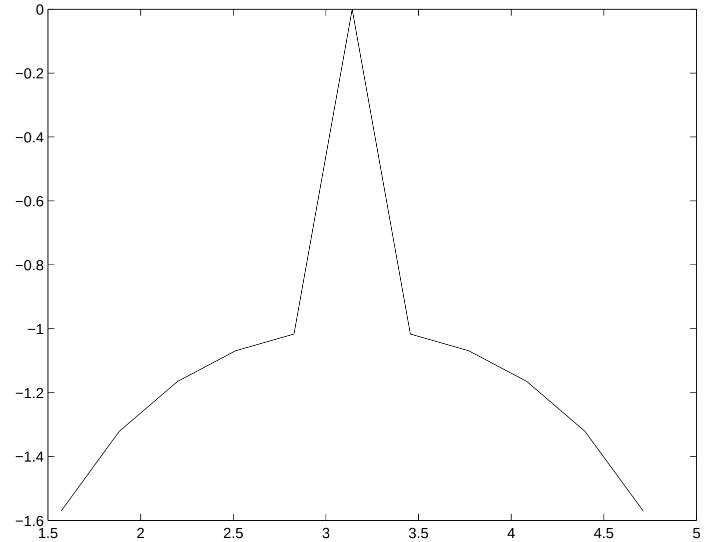
$$f(x) = \begin{cases} \dfrac{x-\pi}{\sin(x)} & \text{falls} \quad \sin(x) \neq 0 \\ -1 & \text{falls} \quad \sin(x) = 0 \end{cases} \qquad x \in [\tfrac{1}{2}\pi, \tfrac{2}{3}\pi]$$

```
function tumbplot(n)

h=pi/n;
for i=1:n+1
    x(i) = pi/2 + (i-1)*h;
    if (sin(x(i))==0) y(i) = -1;
    else y(i) = (x(i)-pi)/sin(x(i)); end;
end;
plot(x,y);
```

Was ist passiert?





Das Ergebnis von TumbPlot für $n = 5$ und $n = 10$.

siehe Abschnitt 5.3.6 im Skript

Gleichheits-Abfragen von Gleitkomma-Zahlen

Folgerung: Die Abfrage `if x̃ == ỹ` mit $\tilde{x},\ \tilde{y} \in \mathbb{G}(q,\ell)$ 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

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

Gleichheits-Abfragen von Gleitkomma-Zahlen verboten!

Algebraische Eigenschaften

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

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

Gleitkommaarithmetik:

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

Algebraische Eigenschaften

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

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

Gleitkommaarithmetik:

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

Der Hammer:

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

$\tilde{+}$, $\tilde{*}$ nicht assoziativ, nicht distributiv, i.a. kein Inverses bzgl. $\tilde{*}$

Computer können nicht rechnen!

Äquivalente Umformungen in $\mathbb{R}$ sind in Gleitkommaarithmetik nicht äquivalent.

Computer können nicht rechnen!

Äquivalente Umformungen in $\mathbb{R}$ sind in Gleitkommaarithmetik nicht äquivalent.

Beispiele:

keine binomische Formel:

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

Computer können nicht rechnen!

Äquivalente Umformungen in $\mathbb{R}$ sind in Gleitkommaarithmetik nicht äquivalent.

Beispiele:

keine binomische Formel:

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

kein Assoziativgesetz:

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

Computer können nicht rechnen!

Äquivalente Umformungen in $\mathbb{R}$ sind in Gleitkommaarithmetik nicht äquivalent.

Beispiele:

keine binomische Formel:

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

kein Assoziativgesetz:

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

It is hard, but it's harder to ignore it Cat Stevens

Ausblick: Kondition

Auswirkung von Eingabefehlern auf das Ergebnis