Rechnet mein Taschenrechner richtig?

René Lamour

Humboldt-Universität zu Berlin
Institut für Mathematik

Lange Nacht der Wissenschaften 2011


SIEMENS

Taschenrechner lösen unsere täglichen kleinen Rechenaufgaben.

Bekommt jeder (egal auf welchen Weg) immer das Gleiche heraus?

Ist das (Taschenrechner-) Ergebnis überhaupt richtig?

Sehen wir uns ein Beispiel an!

Taschenrechner lösen unsere täglichen kleinen Rechenaufgaben.

Bekommt jeder (egal auf welchen Weg) immer das Gleiche heraus?

Ist das (Taschenrechner-) Ergebnis überhaupt richtig?

Sehen wir uns ein Beispiel an!

Taschenrechner lösen unsere täglichen kleinen Rechenaufgaben.

Bekommt jeder (egal auf welchen Weg) immer das Gleiche heraus?

Ist das (Taschenrechner-) Ergebnis überhaupt richtig?

Sehen wir uns ein Beispiel an!

Taschenrechner lösen unsere täglichen kleinen Rechenaufgaben.

Bekommt jeder (egal auf welchen Weg) immer das Gleiche heraus?

Ist das (Taschenrechner-) Ergebnis überhaupt richtig?

Sehen wir uns ein Beispiel an!

$$\frac{1}{\sqrt{a+b}-\sqrt{a}}$$

$$\frac{1}{\sqrt{a+b}-\sqrt{a}} = \frac{\sqrt{a+b}+\sqrt{a}}{(\sqrt{a+b}-\sqrt{a})(\sqrt{a+b}+\sqrt{a})}$$

$$\frac{1}{\sqrt{a+b}-\sqrt{a}} = \frac{\sqrt{a+b}+\sqrt{a}}{(\sqrt{a+b}-\sqrt{a})(\sqrt{a+b}+\sqrt{a})}$$

$$\frac{1}{\sqrt{a+b}-\sqrt{a}} = \frac{\sqrt{a+b}+\sqrt{a}}{(\sqrt{a+b}-\sqrt{a})(\sqrt{a+b}+\sqrt{a})}$$

$$= \frac{\sqrt{a+b}+\sqrt{a}}{b}$$

$$\frac{1}{\sqrt{a+b}-\sqrt{a}} = \frac{\sqrt{a+b}+\sqrt{a}}{(\sqrt{a+b}-\sqrt{a})(\sqrt{a+b}+\sqrt{a})}$$

$$= \frac{\sqrt{a+b}+\sqrt{a}}{b}$$

Wir benutzen hier bekannte mathematische Regeln beim Rechnen mit Zahlen

wie Kommutativität, Distributivität und Assoziativität

Rechnen wir mit Zahlen und setzen für $a = 10^5$ und $b = 10^{-4}$ ein.

$$\frac{1}{\sqrt{a+b}-\sqrt{a}} = \frac{\sqrt{a+b}+\sqrt{a}}{b}$$

Rechnen wir mit Zahlen und setzen für $a = 10^5$ und $b = 10^{-4}$ ein.

$$\frac{1}{\sqrt{a+b}-\sqrt{a}} = \frac{\sqrt{a+b}+\sqrt{a}}{b}$$

$$6329113.924 =$$

Rechnen wir mit Zahlen und setzen für $a = 10^5$ und $b = 10^{-4}$ ein.

$$\frac{1}{\sqrt{a+b}-\sqrt{a}} = \frac{\sqrt{a+b}+\sqrt{a}}{b}$$

$$6329113.924 = 6324555.322$$

Rechnen wir mit Zahlen und setzen für $a = 10^5$ und $b = 10^{-4}$ ein.

$$\frac{1}{\sqrt{a+b} - \sqrt{a}} = \frac{\sqrt{a+b} + \sqrt{a}}{b}$$

$$6329113.924 \stackrel{?}{=} 6324555.322$$

Rechnen wir mit Zahlen und setzen für $a = 10^5$ und $b = 10^{-4}$ ein.

$$\frac{1}{\sqrt{a+b}-\sqrt{a}} = \frac{\sqrt{a+b}+\sqrt{a}}{b}$$

$$6329113.924 \stackrel{?}{=} 6324555.322$$

Das ist ein Unterschied von 4558.60208 bei mathematisch identischen Ausdrücken!
Die vierte Stelle ist falsch bei 10-stelliger Anzeige

Was ist anders bei der Computerrechnung?

Wir verwenden keine reellen Zahlen mehr.

Für Festkommazahlen bedeutet Mantissenlänge die Anzahl der mitgeführten gültigen Ziffern:

$$12300.$$
$$1.23$$
$$0.000123$$

Der Computer verwendet normalisierte Gleitkommazahlen.

Beispiel: $\pm 0.222029388 \cdot 10^{\pm 9}$

Hier entspricht die Mantissenlänge der Anzahl der Ziffern nach dem Komma.

Was ist anders bei der Computerrechnung?

Wir verwenden keine reellen Zahlen mehr.

Für Festkommazahlen bedeutet Mantissenlänge die Anzahl der mitgeführten gültigen Ziffern:

$$12300.$$
$$1.23$$
$$0.000123$$

Der Computer verwendet normalisierte Gleitkommazahlen.

Beispiel: $\pm 0.222029388 \cdot 10^{\pm 9}$

Hier entspricht die Mantissenlänge der Anzahl der Ziffern nach dem Komma.

Was ist anders bei der Computerrechnung?

Wir verwenden keine reellen Zahlen mehr.

Für Festkommazahlen bedeutet Mantissenlänge die Anzahl der mitgeführten gültigen Ziffern:

$$12300.$$
$$1.23$$
$$0.000123$$

Der Computer verwendet normalisierte Gleitkommazahlen.

Beispiel: $\pm 0.222029388 \cdot 10^{\pm 9}$

Hier entspricht die Mantissenlänge der Anzahl der Ziffern nach dem Komma.

Was ist anders bei der Computerrechnung?

Wir verwenden keine reellen Zahlen mehr.

Für Festkommazahlen bedeutet Mantissenlänge die Anzahl der mitgeführten gültigen Ziffern:

$$12300.$$
$$1.23$$
$$0.000123$$

Der Computer verwendet normalisierte Gleitkommazahlen.

Beispiel: $\pm 0.222029388 \cdot 10^{\pm 9}$

Hier entspricht die Mantissenlänge der Anzahl der Ziffern nach dem Komma.

Was ist anders bei der Computerrechnung?

Wir verwenden keine reellen Zahlen mehr.

Für Festkommazahlen bedeutet Mantissenlänge die Anzahl der mitgeführten gültigen Ziffern:

$$12300.$$
$$1.23$$
$$0.000123$$

Der Computer verwendet normalisierte Gleitkommazahlen.

Beispiel: $\pm 0.222029388 \cdot 10^{\pm 9}$

Hier entspricht die Mantissenlänge der Anzahl der Ziffern nach dem Komma.

Was ist anders bei der Computerrechnung?

Wir verwenden keine reellen Zahlen mehr.

Für Festkommazahlen bedeutet Mantissenlänge die Anzahl der mitgeführten gültigen Ziffern:

$$12300.$$
$$1.23$$
$$0.000123$$

Der Computer verwendet normalisierte Gleitkommazahlen.

Beispiel: $\pm 0.222029388 \cdot 10^{\pm 9}$

Hier entspricht die Mantissenlänge der Anzahl der Ziffern nach dem Komma.

Was ist anders bei der Computerrechnung?

Wir verwenden keine reellen Zahlen mehr.

Für Festkommazahlen bedeutet Mantissenlänge die Anzahl der mitgeführten gültigen Ziffern:

$$12300.$$
$$1.23$$
$$0.000123$$

Der Computer verwendet normalisierte Gleitkommazahlen.

Beispiel: $\pm 0.222029388 \cdot 10^{\pm 9}$

Hier entspricht die Mantissenlänge der Anzahl der Ziffern nach dem Komma.

Mit welcher Mantissenlänge rechnet mein Taschenrechner?

Mit welcher Mantissenlänge rechnet mein Taschenrechner?
Wir testen, wann

$$(1+\varepsilon) = 1$$

mit $1 \gg \varepsilon > 0$.

Mit welcher Mantissenlänge rechnet mein Taschenrechner?
Wir testen, wann

$$(1+\varepsilon) = 1$$

mit $1 \gg \varepsilon > 0$.

$$1.\overbrace{0000000000000|00000}^{\text{Mantissenlänge}-1\text{ Stellen}}|$$

Mit welcher Mantissenlänge rechnet mein Taschenrechner?
Wir testen, wann

$$(1+\varepsilon)=1$$

mit $1 \gg \varepsilon > 0$.

$$\begin{array}{rl}
 & 1.\overbrace{0000000000000|00000|}^{\text{Mantissenlänge}-1\text{ Stellen}} \\
+ & 0.0000001 \qquad\qquad\quad |
\end{array}$$

Mit welcher Mantissenlänge rechnet mein Taschenrechner?
Wir testen, wann

$$(1+\varepsilon) = 1$$

mit $1 \gg \varepsilon > 0$.

$$\begin{array}{r} 1.\overbrace{0000000000000|00000|}^{\text{Mantissenlänge}-1\ \text{Stellen}} \\ +\,0.0000000000000|01| \end{array}$$

Mit welcher Mantissenlänge rechnet mein Taschenrechner?
Wir testen, wann

$$(1+\varepsilon) = 1$$

mit $1 \gg \varepsilon > 0$.

$$\begin{array}{r} 1.\overbrace{0000000000000|00000|}^{\text{Mantissenlänge}-1\text{ Stellen}} \\ +\,0.0000000000000|00000|1 \end{array}$$

Mit welcher Mantissenlänge rechnet mein Taschenrechner?
Wir testen, wann

$$(1+\varepsilon) = 1$$

mit $1 \gg \varepsilon > 0$.

$$\begin{aligned} &1.\overbrace{0000000000000|00000|}^{\text{Mantissenlänge}-1\text{ Stellen}} \\ +\,&0.0000000000000|00000|1 = 10^{-\text{Mantissenlänge}} \end{aligned}$$

Mit welcher Mantissenlänge rechnet mein Taschenrechner?
Wir testen, wann

$$(1+\varepsilon) = 1$$

mit $1 \gg \varepsilon > 0$.

$$\begin{aligned} &1.\overbrace{0000000000000|00000|}^{\text{Mantissenlänge}-1\text{ Stellen}} \\ +\,&0.0000000000000|00000|1 = 10^{-\text{Mantissenlänge}} \end{aligned}$$

Wir testen $(1+10^{-s}) - 1 = 0$?

Mit welcher Mantissenlänge rechnet mein Taschenrechner?
Wir testen, wann

$$(1+\varepsilon)=1$$

mit $1 \gg \varepsilon > 0$.

$$\begin{aligned} &1.\overbrace{00000000000000|00000|}^{\text{Mantissenlänge}-1\text{ Stellen}} \\ +&0.00000000000000|00000|1 = 10^{-\text{Mantissenlänge}} \end{aligned}$$

Wir testen $(1+10^{-s})-1=0$?
Mein Taschenrechner hat 10 angezeigte Stellen, aber 11 Mantissenstellen.

Es gelten nicht: Assoziativgesetz und Distributivgesetz

$$(\varepsilon + 1) - 1 \neq \varepsilon + (1 - 1)$$

Kommt daher das schlechte Ergebnis des Beispiels?

Durch notwendige Rundung werden Fehler gemacht. Wie wirken sich diese Fehler auf ein Berechnungsergebnis aus?

Betrachten wir die Berechnung eines Funktionswertes y einer Funktion f mit

$$y = f(d), \quad f \in C^1(\mathbb{R}^n, \mathbb{R}), \quad d \in \mathbb{R}^n.$$

Wir wollen die Abhängigkeit des relativen Fehlers von y vom relativen Fehler von d untersuchen.

Es gelten nicht: Assoziativgesetz und Distributivgesetz

$$(\varepsilon + 1) - 1 \neq \varepsilon + (1 - 1)$$

Kommt daher das schlechte Ergebnis des Beispiels?

Durch notwendige Rundung werden Fehler gemacht. Wie wirken sich diese Fehler auf ein Berechnungsergebnis aus?

Betrachten wir die Berechnung eines Funktionswertes y einer Funktion f mit

$$y = f(d), \quad f \in C^1(\mathbb{R}^n, \mathbb{R}), \quad d \in \mathbb{R}^n.$$

Wir wollen die Abhängigkeit des relativen Fehlers von y vom relativen Fehler von d untersuchen.

Es gelten nicht: Assoziativgesetz und Distributivgesetz

$$(\varepsilon + 1) - 1 \neq \varepsilon + (1 - 1)$$

Kommt daher das schlechte Ergebnis des Beispiels?

Durch notwendige Rundung werden Fehler gemacht. Wie wirken sich diese Fehler auf ein Berechnungsergebnis aus?

Betrachten wir die Berechnung eines Funktionswertes y einer Funktion f mit

$$y = f(d), \quad f \in C^1(\mathbb{R}^n, \mathbb{R}), \quad d \in \mathbb{R}^n.$$

Wir wollen die Abhängigkeit des relativen Fehlers von y vom relativen Fehler von d untersuchen.

Es gelten nicht: Assoziativgesetz und Distributivgesetz

$$(\varepsilon + 1) - 1 \neq \varepsilon + (1 - 1)$$

Kommt daher das schlechte Ergebnis des Beispiels?

Durch notwendige Rundung werden Fehler gemacht. Wie wirken sich diese Fehler auf ein Berechnungsergebnis aus?

Betrachten wir die Berechnung eines Funktionswertes y einer Funktion f mit

$$y = f(d), \quad f \in C^1(\mathbb{R}^n, \mathbb{R}), \quad d \in \mathbb{R}^n.$$

Wir wollen die Abhängigkeit des relativen Fehlers von y vom relativen Fehler von d untersuchen.

Es gelten nicht: Assoziativgesetz und Distributivgesetz

$$(\varepsilon + 1) - 1 \neq \varepsilon + (1 - 1)$$

Kommt daher das schlechte Ergebnis des Beispiels?

Durch notwendige Rundung werden Fehler gemacht. Wie wirken sich diese Fehler auf ein Berechnungsergebnis aus?

Betrachten wir die Berechnung eines Funktionswertes y einer Funktion f mit

$$y = f(d), \quad f \in C^1(\mathbb{R}^n, \mathbb{R}), \quad d \in \mathbb{R}^n.$$

Wir wollen die Abhängigkeit des relativen Fehlers von y vom relativen Fehler von d untersuchen.

Es gelten nicht: Assoziativgesetz und Distributivgesetz

$$(\varepsilon + 1) - 1 \neq \varepsilon + (1 - 1)$$

Kommt daher das schlechte Ergebnis des Beispiels?

Durch notwendige Rundung werden Fehler gemacht. Wie wirken sich diese Fehler auf ein Berechnungsergebnis aus?

Betrachten wir die Berechnung eines Funktionswertes y einer Funktion f mit

$$y = f(d), \quad f \in C^1(\mathbb{R}^n, \mathbb{R}), \quad d \in \mathbb{R}^n.$$

Wir wollen die Abhängigkeit des relativen Fehlers von y vom relativen Fehler von d untersuchen.

Wir betrachten für festes $d \in \mathbb{R}^n$ und einen festen Fehlervektor $\triangle d \in \mathbb{R}^n$ die skalare Funktion F mit hinreichend glattem f

$$F(s) := f(d + s\triangle d).$$

Für F gilt der Mittelwertsatz und daher

$$F(1) - F(0) = f(d + \triangle d) - f(d) = F'(\theta)(1 - 0).$$

$$F'(\theta) = f'(d + \theta \triangle d) = \begin{pmatrix} \frac{\partial f(d+\theta\triangle d)}{\partial d_1} & \dots & \frac{\partial f(d+\theta\triangle d)}{\partial d_n} \end{pmatrix} \begin{pmatrix} \triangle d_1 \\ \vdots \\ \triangle d_n \end{pmatrix}.$$

Wir betrachten für festes $d \in \mathbb{R}^n$ und einen festen Fehlervektor $\triangle d \in \mathbb{R}^n$ die skalare Funktion F mit hinreichend glattem f

$$F(s) := f(d + s\triangle d).$$

Für F gilt der Mittelwertsatz und daher

$$F(1) - F(0) = f(d + \triangle d) - f(d) = F'(\theta)(1 - 0).$$

$$F'(\theta) = f'(d + \theta \triangle d) = \begin{pmatrix} \frac{\partial f(d+\theta\triangle d)}{\partial d_1} & \dots & \frac{\partial f(d+\theta\triangle d)}{\partial d_n} \end{pmatrix} \begin{pmatrix} \triangle d_1 \\ \vdots \\ \triangle d_n \end{pmatrix}.$$

Wir betrachten für festes $d \in \mathbb{R}^n$ und einen festen Fehlervektor $\triangle d \in \mathbb{R}^n$ die skalare Funktion F mit hinreichend glattem f

$$F(s) := f(d + s\triangle d).$$

Für F gilt der Mittelwertsatz und daher

$$F(1) - F(0) = f(d + \triangle d) - f(d) = F'(\theta)(1 - 0).$$

$$F'(\theta) = f'(d + \theta \triangle d) = \begin{pmatrix} \frac{\partial f(d+\theta\triangle d)}{\partial d_1} & \dots & \frac{\partial f(d+\theta\triangle d)}{\partial d_n} \end{pmatrix} \begin{pmatrix} \triangle d_1 \\ \vdots \\ \triangle d_n \end{pmatrix}.$$

Wir betrachten für festes $d \in \mathbb{R}^n$ und einen festen Fehlervektor $\triangle d \in \mathbb{R}^n$ die skalare Funktion F mit hinreichend glattem f

$$F(s) := f(d + s\triangle d).$$

Für F gilt der Mittelwertsatz und daher

$$F(1) - F(0) = f(d + \triangle d) - f(d) = F'(\theta)(1 - 0).$$

$$F'(\theta) = f'(d + \theta \triangle d) = \begin{pmatrix} \frac{\partial f(d+\theta\triangle d)}{\partial d_1} & \dots & \frac{\partial f(d+\theta\triangle d)}{\partial d_n} \end{pmatrix} \begin{pmatrix} \triangle d_1 \\ \vdots \\ \triangle d_n \end{pmatrix}.$$

$$\triangle y = f(d + \triangle d) - f(d)$$

$$\begin{aligned}\triangle y &= f(d + \triangle d) - f(d)\\ &= f'(d + \theta \triangle d)\end{aligned}$$

$$\begin{aligned} \triangle y &= f(d + \triangle d) - f(d) \\ &= f'(d + \theta \triangle d) \\ &= \begin{pmatrix} \frac{\partial f}{\partial d_1} & \cdots & \frac{\partial f}{\partial d_n} \end{pmatrix} \begin{pmatrix} \triangle d_1 \\ \vdots \\ \triangle d_n \end{pmatrix} \end{aligned}$$

$$\begin{aligned}
\triangle y &= f(d + \triangle d) - f(d) \\
&= f'(d + \theta \triangle d) \\
&= \begin{pmatrix} \frac{\partial f}{\partial d_1} & \cdots & \frac{\partial f}{\partial d_n} \end{pmatrix} \begin{pmatrix} \triangle d_1 \\ \vdots \\ \triangle d_n \end{pmatrix} \\
&= \sum_{i=1}^{n} \frac{\partial f}{\partial d_i} \triangle d_i
\end{aligned}$$

$$\begin{aligned}
\triangle y &= f(d + \triangle d) - f(d) \\
&= f'(d + \theta \triangle d) \\
&= \begin{pmatrix} \frac{\partial f}{\partial d_1} & \cdots & \frac{\partial f}{\partial d_n} \end{pmatrix} \begin{pmatrix} \triangle d_1 \\ \vdots \\ \triangle d_n \end{pmatrix} \\
&= \sum_{i=1}^{n} \frac{\partial f}{\partial d_i} \triangle d_i = \sum_{i=1}^{n} d_i \frac{\partial f}{\partial d_i} \frac{\triangle d_i}{d_i}
\end{aligned}$$

$$
\begin{aligned}
\triangle y &= f(d + \triangle d) - f(d) \\
&= f'(d + \theta \triangle d) \\
&= \begin{pmatrix} \frac{\partial f}{\partial d_1} & \cdots & \frac{\partial f}{\partial d_n} \end{pmatrix} \begin{pmatrix} \triangle d_1 \\ \vdots \\ \triangle d_n \end{pmatrix} \\
&= \sum_{i=1}^{n} \frac{\partial f}{\partial d_i} \triangle d_i = \sum_{i=1}^{n} {\color{green} d_i} \frac{\partial f}{\partial d_i} \frac{\triangle d_i}{\color{green} d_i}
\end{aligned}
$$

Wegen $y = f(d)$ ist für $y \neq 0$

$$\begin{aligned}\triangle y &= f(d+\triangle d) - f(d)\\ &= f'(d+\theta \triangle d)\\ &= \sum_{i=1}^{n} \frac{\partial f}{\partial d_i} \triangle d_i = \sum_{i=1}^{n} d_i \frac{\partial f}{\partial d_i} \frac{\triangle d_i}{d_i}\end{aligned}$$

Wegen $y = f(d)$ ist für $y \neq 0$

$$\frac{\triangle y}{y} = \sum_{i=1}^{n} \frac{d_i}{f(d)} \frac{\partial f}{\partial d_i} \frac{\triangle d_i}{d_i} = \begin{pmatrix} \frac{d_1}{f(d)} \frac{\partial f}{\partial d_1} & \cdots & \frac{d_n}{f(d)} \frac{\partial f}{\partial d_n} \end{pmatrix} \begin{pmatrix} \frac{\triangle d_1}{d_1} \\ \vdots \\ \frac{\triangle d_n}{d_n} \end{pmatrix}$$

$$
\begin{aligned}
\triangle y &= f(d + \triangle d) - f(d) \\
&= f'(d + \theta \triangle d) \\
&= \sum_{i=1}^{n} \frac{\partial f}{\partial d_i} \triangle d_i = \sum_{i=1}^{n} d_i \frac{\partial f}{\partial d_i} \frac{\triangle d_i}{d_i}
\end{aligned}
$$

Wegen $y = f(d)$ ist für $y \neq 0$

$$
\frac{\triangle y}{y} = \sum_{i=1}^{n} \frac{d_i}{f(d)} \frac{\partial f}{\partial d_i} \frac{\triangle d_i}{d_i} = \begin{pmatrix} \frac{d_1}{f(d)} \frac{\partial f}{\partial d_1} & \cdots & \frac{d_n}{f(d)} \frac{\partial f}{\partial d_n} \end{pmatrix} \begin{pmatrix} \frac{\triangle d_1}{d_1} \\ \vdots \\ \frac{\triangle d_n}{d_n} \end{pmatrix}
$$

$$\begin{aligned}\triangle y &= f(d + \triangle d) - f(d)\\ &= f'(d + \theta \triangle d)\\ &= \sum_{i=1}^{n} \frac{\partial f}{\partial d_i} \triangle d_i = \sum_{i=1}^{n} d_i \frac{\partial f}{\partial d_i} \frac{\triangle d_i}{d_i}\end{aligned}$$

Wegen $y = f(d)$ ist für $y \neq 0$

$$\frac{\triangle y}{y} = \sum_{i=1}^{n} \frac{d_i}{f(d)} \frac{\partial f}{\partial d_i} \frac{\triangle d_i}{d_i} = \begin{pmatrix} \frac{d_1}{f(d)} \frac{\partial f}{\partial d_1} & \cdots & \frac{d_n}{f(d)} \frac{\partial f}{\partial d_n} \end{pmatrix} \begin{pmatrix} \frac{\triangle d_1}{d_1} \\ \vdots \\ \frac{\triangle d_n}{d_n} \end{pmatrix}$$

$$\frac{|\triangle y|}{|y|} \leqslant \underbrace{\left\| \begin{pmatrix} \frac{d_1}{f(d)} \frac{\partial f}{\partial d_1} & \cdots & \frac{d_n}{f(d)} \frac{\partial f}{\partial d_n} \end{pmatrix} \right\|}_{:=K \text{ - relative Kondition}} \left\| \begin{pmatrix} \frac{\triangle d_1}{d_1} \\ \vdots \\ \frac{\triangle d_n}{d_n} \end{pmatrix} \right\|$$

Wir wählen die Unendlichnorm

$$\|v\|_\infty := \max_i |v_i|.$$

Die induzierte Matrixnorm ist die Zeilensummennorm

$$\|A\|_\infty := \max_i \sum_{j=1}^{n} |a_{ij}|.$$

Grundrechenarten: $+ \; - \; * \; /$

Grundrechenarten: $+ \; - \; * \; /$ zwei Operanden $y = f(d_1, d_2)$

Grundrechenarten: $+ \; - \; * \; /$ zwei Operanden $y = f(d_1, d_2)$

$$|\frac{\triangle y}{y}| \leqslant K(d_1, d_2, \triangle d_1, \triangle d_2) \max(|\frac{\triangle d_1}{d_1}|, |\frac{\triangle d_2}{d_2}|)$$

Grundrechenarten: $+ \ - \ * \ /$ zwei Operanden $y = f(d_1, d_2)$

$$|\frac{\triangle y}{y}| \leqslant K(d_1, d_2, \triangle d_1, \triangle d_2) \max(|\frac{\triangle d_1}{d_1}|, |\frac{\triangle d_2}{d_2}|)$$

$$K = \left|\frac{d_1}{f(d_1,d_2)} \frac{\partial f}{\partial d_1}(d_1+\triangle d_1, d_2+\triangle d_2)\right| + \left|\frac{d_2}{f(d_1,d_2)} \frac{\partial f}{\partial d_2}(d_1+\triangle d_1, d_2+\triangle d_2)\right|$$

$$K = \left| \frac{d_1}{f(d_1,d_2)} \frac{\partial f}{\partial d_1}(d_1+\triangle d_1, d_2+\triangle d_2) \right| + \left| \frac{d_2}{f(d_1,d_2)} \frac{\partial f}{\partial d_2}(d_1+\triangle d_1, d_2+\triangle d_2) \right|$$

Addition: $f(d) := d_1 + d_2, \quad d_1, d_2 > 0,$

$$K = \left| \frac{d_1}{f(d_1,d_2)} \frac{\partial f}{\partial d_1}(d_1+\triangle d_1, d_2+\triangle d_2) \right| + \left| \frac{d_2}{f(d_1,d_2)} \frac{\partial f}{\partial d_2}(d_1+\triangle d_1, d_2+\triangle d_2) \right|$$

Addition: $f(d) := d_1 + d_2, \quad d_1, d_2 > 0, \quad \frac{\partial f}{\partial d_i} = 1$

$$K = \left| \frac{d_1}{f(d_1,d_2)} \frac{\partial f}{\partial d_1}(d_1+\triangle d_1, d_2+\triangle d_2) \right| + \left| \frac{d_2}{f(d_1,d_2)} \frac{\partial f}{\partial d_2}(d_1+\triangle d_1, d_2+\triangle d_2) \right|$$

Addition: $f(d) := d_1 + d_2, \quad d_1, d_2 > 0, \quad \frac{\partial f}{\partial d_i} = 1$

$K = \frac{d_1}{d_1+d_2} + \frac{d_2}{d_1+d_2} = 1.$

$$K = \left| \frac{d_1}{f(d_1,d_2)} \frac{\partial f}{\partial d_1}(d_1+\triangle d_1, d_2+\triangle d_2) \right| + \left| \frac{d_2}{f(d_1,d_2)} \frac{\partial f}{\partial d_2}(d_1+\triangle d_1, d_2+\triangle d_2) \right|$$

Addition: $f(d) := d_1 + d_2, \quad d_1, d_2 > 0, \quad \frac{\partial f}{\partial d_i} = 1$

$K = \frac{d_1}{d_1+d_2} + \frac{d_2}{d_1+d_2} = 1.$

Multiplikation: $f(d) := d_1 * d_2,\ d_1, d_2 > 0$

$$K = \left| \frac{d_1}{f(d_1,d_2)} \frac{\partial f}{\partial d_1}(d_1+\triangle d_1, d_2+\triangle d_2) \right| + \left| \frac{d_2}{f(d_1,d_2)} \frac{\partial f}{\partial d_2}(d_1+\triangle d_1, d_2+\triangle d_2) \right|$$

Addition: $f(d) := d_1 + d_2, \quad d_1, d_2 > 0, \quad \frac{\partial f}{\partial d_i} = 1$

$K = \frac{d_1}{d_1+d_2} + \frac{d_2}{d_1+d_2} = 1.$

Multiplikation: $f(d) := d_1 * d_2,\ d_1, d_2 > 0\ , \frac{\partial f}{\partial d_1} = d_2,\ \frac{\partial f}{\partial d_2} = d_1$

$$K = \left| \frac{d_1}{f(d_1,d_2)} \frac{\partial f}{\partial d_1}(d_1+\triangle d_1, d_2+\triangle d_2) \right| + \left| \frac{d_2}{f(d_1,d_2)} \frac{\partial f}{\partial d_2}(d_1+\triangle d_1, d_2+\triangle d_2) \right|$$

Addition: $f(d) := d_1 + d_2, \quad d_1, d_2 > 0, \quad \frac{\partial f}{\partial d_i} = 1$

$K = \frac{d_1}{d_1+d_2} + \frac{d_2}{d_1+d_2} = 1.$

Multiplikation: $f(d) := d_1 * d_2,\ d_1, d_2 > 0\ , \frac{\partial f}{\partial d_1} = d_2,\ \frac{\partial f}{\partial d_2} = d_1$

$K = \frac{d_1}{d_1 * d_2} * d_2 + \frac{d_2}{d_1 * d_2} * d_1 = 2.$

$$K = \left| \frac{d_1}{f(d_1,d_2)} \frac{\partial f}{\partial d_1}(d_1+\triangle d_1, d_2+\triangle d_2) \right| + \left| \frac{d_2}{f(d_1,d_2)} \frac{\partial f}{\partial d_2}(d_1+\triangle d_1, d_2+\triangle d_2) \right|$$

Addition: $f(d) := d_1 + d_2, \quad d_1, d_2 > 0, \quad \frac{\partial f}{\partial d_i} = 1$

$K = \frac{d_1}{d_1+d_2} + \frac{d_2}{d_1+d_2} = 1.$

Multiplikation: $f(d) := d_1 * d_2,\ d_1, d_2 > 0\ , \frac{\partial f}{\partial d_1} = d_2,\ \frac{\partial f}{\partial d_2} = d_1$

$K = \frac{d_1}{d_1 * d_2} * d_2 + \frac{d_2}{d_1 * d_2} * d_1 = 2.$

Division: $f(d) := \frac{d_1}{d_2}, \quad d_1, d_2 > 0, \quad \frac{\partial f}{\partial d_1} = \frac{1}{d_2},\ \frac{\partial f}{\partial d_2} = -\frac{d_1}{d_2^2}$

$K = \frac{d_1}{\frac{d_1}{d_2}} * \frac{1}{d_2} + \frac{d_2}{\frac{d_1}{d_2}} * \frac{d_1}{d_2^2} = 2.$

$$K = \left| \frac{d_1}{f(d_1,d_2)} \frac{\partial f}{\partial d_1}(d_1+\triangle d_1, d_2+\triangle d_2) \right| + \left| \frac{d_2}{f(d_1,d_2)} \frac{\partial f}{\partial d_2}(d_1+\triangle d_1, d_2+\triangle d_2) \right|$$

Addition: $f(d) := d_1 + d_2, \quad d_1, d_2 > 0, \quad \frac{\partial f}{\partial d_i} = 1$

$$K = \frac{d_1}{d_1+d_2} + \frac{d_2}{d_1+d_2} = 1.$$

Multiplikation: $f(d) := d_1 * d_2,\ d_1, d_2 > 0\ , \frac{\partial f}{\partial d_1} = d_2,\ \frac{\partial f}{\partial d_2} = d_1$

$$K = \frac{d_1}{d_1 * d_2} * d_2 + \frac{d_2}{d_1 * d_2} * d_1 = 2.$$

Division: $f(d) := \frac{d_1}{d_2}, \quad d_1, d_2 > 0, \quad \frac{\partial f}{\partial d_1} = \frac{1}{d_2},\ \frac{\partial f}{\partial d_2} = -\frac{d_1}{d_2^2}$

$$K = \frac{d_1}{\frac{d_1}{d_2}} * \frac{1}{d_2} + \frac{d_2}{\frac{d_1}{d_2}} * \frac{d_1}{d_2^2} = 2.$$

Subtraktion: $f(d) := d_1 - d_2, \quad d_1 \geq d_2 \geq 0$

$$K = \frac{d_1}{d_1 - d_2} + \frac{d_2}{d_1 - d_2} = \frac{d_1 + d_2}{d_1 - d_2}$$

Subtraktion: $f(d) := d_1 - d_2, \quad d_1 \geq d_2 \geq 0$

$K = \frac{d_1}{d_1 - d_2} + \frac{d_2}{d_1 - d_2} = \frac{d_1 + d_2}{d_1 - d_2} \to \infty$ für $d_2 \to d_1$

Dieser Effekt wird Auslöschung genannt, weil zwei etwa gleich große Gleikommazahlen viele gleiche Ziffern haben, die sich bei der Differenzbildung auslöschen.

Dieser Effekt wird Auslöschung genannt, weil zwei etwa gleich große Gleikommazahlen viele gleiche Ziffern haben, die sich bei der Differenzbildung auslöschen.

Die Auslöschung ist hauptverantwortlich für extrem ungenaue Rechnerergebnisse, aber sie ist manchmal schwer zu entdecken.

Dieser Effekt wird Auslöschung genannt, weil zwei etwa gleich große Gleikommazahlen viele gleiche Ziffern haben, die sich bei der Differenzbildung auslöschen.

Die Auslöschung ist hauptverantwortlich für extrem ungenaue Rechnerergebnisse, aber sie ist manchmal schwer zu entdecken.

$$\frac{1}{\sqrt{a+b}-\sqrt{a}} = \frac{\sqrt{a+b}+\sqrt{a}}{b}$$

Beispiel

S.M. Rump. *Algorithms for Verified Inclusions - Theory and Practice.* In R.E. Moore, editor, Reliability in Computing, Volume 19 of Perspectives in Computing, pages 109-126. Academic Press, 1988.

Beispiel

Es ist zu berechnen

$$y = 333.75b^6 + a^2(11a^2b^2 - b^6 - 121b^4 - 2) + 5.5b^8 + a/(2b)$$

für $a = 77617.0$ und $b = 33096.0$.

Beispiel

Es ist zu berechnen

$$y = 333.75b^6 + a^2(11a^2b^2 - b^6 - 121b^4 - 2) + 5.5b^8 + a/(2b)$$

für $a = 77617.0$ und $b = 33096.0$.

Rump gab für unterschiedliche Mantissenlängen folgende Ergebnisse an:

Beispiel

Es ist zu berechnen

$$y = 333.75b^6 + a^2(11a^2b^2 - b^6 - 121b^4 - 2) + 5.5b^8 + a/(2b)$$

für $a = 77617.0$ und $b = 33096.0$.

Rump gab für unterschiedliche Mantissenlängen folgende Ergebnisse an:

IBM S/370, FORTRAN

float	$y = +1.172603\ldots$
double	$y = +1.1726039400531\ldots$
extended	$y = +1.172603940053178\ldots$

Neue Rechnung

$$y = 333.75b^6 + a^2(11a^2b^2 - b^6 - 121b^4 - 2) + 5.5b^8 + a/(2b)$$

$$\begin{aligned}
b_2 &= b * b \\
b_4 &= b_2 * b_2 \\
b_6 &= b_2 * b_4 \\
b_8 &= b_4 * b_4 \\
a_2 &= a * a \\
y &= 333.75b_6 + a_2(11a_2 * b_2 - b_6 - 121b_4 - 2) + 5.5b_8 + a/(2b)
\end{aligned}$$

Neue Rechnung

$$y = 333.75b^6 + a^2(11a^2b^2 - b^6 - 121b^4 - 2) + 5.5b^8 + a/(2b)$$

$$\begin{aligned}
b_2 &= b * b\\
b_4 &= b_2 * b_2\\
b_6 &= b_2 * b_4\\
b_8 &= b_4 * b_4\\
a_2 &= a * a\\
y &= 333.75b_6 + a_2(11a_2 * b_2 - b_6 - 121b_4 - 2) + 5.5b_8 + a/(2b)
\end{aligned}$$

Ergebnis:

Typ	y
float	-6.338253e+29

$$y = 333.75 b_6 + a_2 (11 a_2 * b_2 - b_6 - 121 b_4 - 2) + 5.5 b_8 + a/(2b)$$

Andere Rechnung

$$\begin{aligned}\bar{y} = 333.75{*}b*b*b*b*b*b + a*a*(11*a*a*b*b \\ - b*b*b*b*b*b - 121*b*b*b*b - 2) \\ + 5.5*b*b*b*b*b*b*b*b + a/(2b)\end{aligned}$$

Ergebnis:

Typ	y
float	-6.338253e+29

$$y = 333.75b_6 + a_2(11a_2 * b_2 - b_6 - 121b_4 - 2) + 5.5b_8 + a/(2b)$$

Andere Rechnung

$$\begin{aligned}\bar{y} = 333.75{*}b * b * b * b * b * b + a * a * (&11 * a * a * b * b \\ &- b * b * b * b * b * b - 121 * b * b * b * b - 2) \\ &+ 5.5 * b * b * b * b * b * b * b * b + a/(2b)\end{aligned}$$

Ergebnis:

Typ	y	$\bar{y}$
float	-6.338253e+29	6.338253e+29

$$y = 333.75b_6 + a_2(11a_2 * b_2 - b_6 - 121b_4 - 2) + 5.5b_8 + a/(2b)$$

Andere Rechnung

$$\begin{aligned}\bar{y} = 333.75{*}b * b * b * b * b * b + a * a * (11 * a * a * b * b \\ - b * b * b * b * b * b - 121 * b * b * b * b - 2) \\ + 5.5 * b * b * b * b * b * b * b * b + a/(2b)\end{aligned}$$

Ergebnis:

Typ	y	$\bar{y}$
float	-6.338253e+29	6.338253e+29
double	-1.180591620717411e+21	

$$y = 333.75 b_6 + a_2 (11 a_2 * b_2 - b_6 - 121 b_4 - 2) + 5.5 b_8 + a/(2b)$$

Andere Rechnung

$$\begin{aligned} \bar{y} = 333.75{*}b * b * b * b * b * b + a * a * (11 * a * a * b * b \\ - b * b * b * b * b * b - 121 * b * b * b * b - 2) \\ + 5.5 * b * b * b * b * b * b * b * b + a/(2b) \end{aligned}$$

Ergebnis:

Typ	y	$\bar{y}$
float	-6.338253e+29	6.338253e+29
double	-1.180591620717411e+21	1.172603940053178

$$y = 333.75b_6 + a_2(11a_2 * b_2 - b_6 - 121b_4 - 2) + 5.5b_8 + a/(2b)$$

Andere Rechnung

$$\begin{aligned}\bar{y} = 333.75{*}b * b * b * b * b * b + a * a * (11 * a * a * b * b \\ - b * b * b * b * b * b - 121 * b * b * b * b - 2) \\ + 5.5 * b * b * b * b * b * b * b * b + a/(2b)\end{aligned}$$

Ergebnis:

Typ	y	$\bar{y}$
float	-6.338253e+29	6.338253e+29
double	-1.180591620717411e+21	1.172603940053178
34 Stellen	-1998.827396059946821368...	dito

$$y = 333.75 b_6 + a_2(11 a_2 * b_2 - b_6 - 121 b_4 - 2) + 5.5 b_8 + a/(2b)$$

Andere Rechnung

$$\begin{aligned}\bar{y} = 333.75{*}b * b * b * b * b * b + a * a * (11 * a * a * b * b \\ - b * b * b * b * b * b - 121 * b * b * b * b - 2) \\ + 5.5 * b * b * b * b * b * b * b * b + a/(2b)\end{aligned}$$

Ergebnis:

Typ	y	$\bar{y}$
float	-6.338253e+29	6.338253e+29
double	-1.1805916207174113e+21	1.1726039400531787
34 Stellen	-1998.827396059946821368...	dito
ab 37 Stellen	-0.827396059946821368...	dito

$$y = 333.75b_6 + a_2(11a_2 * b_2 - b_6 - 121b_4 - 2) + 5.5b_8 + a/(2b)$$

Andere Rechnung

$$\begin{aligned}\bar{y} = 333.75{*}b * b * b * b * b * b + a * a * (11 * a * a * b * b \\ - b * b * b * b * b * b - 121 * b * b * b * b - 2) \\ + 5.5 * b * b * b * b * b * b * b * b + a/(2b)\end{aligned}$$

Ergebnis:

Typ	y	$\bar{y}$
float	-6.338253e+29	6.338253e+29
double	-1.180591620717411e+21	1.172603940053179
34 Stellen	-1998.827396059946821368...	dito
ab 37 Stellen	-0.827396059946821368...	dito

Das Rump'sche Beispiel läßt sich umformen zu:

$$y = (5.5 * 77617^8 - 2) - (5.5 * 77617^8 - \frac{77617}{2*33096})$$

Das Rump'sche Beispiel läßt sich umformen zu:

$y = (5.5 * 77617^8 - 2) - (5.5 * 77617^8 - \frac{77617}{2*33096})$

Dabei ist $5.5 * 77617^8 = 7.2 \ldots e{+}39$

Das Rump'sche Beispiel läßt sich umformen zu:

$y = (5.5 * 77617^8 - 2) - (5.5 * 77617^8 - \frac{77617}{2*33096})$

Dabei ist $5.5 * 77617^8 = 7.2 \ldots e{+}39$

und $\frac{77617}{2*33096} - 2 = -0.827396059946821$3

Das Rump'sche Beispiel läßt sich umformen zu:

$y = (5.5 * 77617^8 - 2) - (5.5 * 77617^8 - \frac{77617}{2*33096})$

Dabei ist $5.5 * 77617^8 = 7.2 \ldots e{+}39$

und $\frac{77617}{2*33096} - 2 = -0.8273960599468213$

Kommt eine solche Auslöschungsituation in der Praxis vor?

Das Rump'sche Beispiel läßt sich umformen zu:

$y = (5.5 * 77617^8 - 2) - (5.5 * 77617^8 - \frac{77617}{2*33096})$

Dabei ist $5.5 * 77617^8 = 7.2 \ldots e{+}39$

und $\frac{77617}{2*33096} - 2 = -0.827396059946821$

Kommt eine solche Auslöschungssituation in der Praxis vor?

Murphys Gesetz lautet:
Whatever can go wrong, will go wrong!

Seien Sie wachsam,

überprüfen Sie die Ergebnisse Ihres Taschenrechners,

misstrauen Sie dem Vorzeichen!

Seien Sie wachsam,

überprüfen Sie die Ergebnisse Ihres Taschenrechners,

misstrauen Sie dem Vorzeichen!

Weiterhin viel Vergnügen und vielen Dank

für Ihre Aufmerksamkeit!