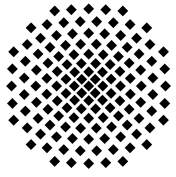
# Numerische und stochastische Grundlagen

Institut für Parallele und Verteilte Systeme
Universität Stuttgart

Wintersemester 2011/2012

Universität Stuttgart

# Fehler

## Fehlerarten

**Modellfehler** Zusammenhang zwischen $y$ und $x$ unbekannt.
Vermutung/Modell: Es gilt $y = f(x)$.

**Eingangsfehler** Datenfehler (unscharfe Messung), ...

$$x - \tilde{x}$$

**Verfahrensfehler** Approximative Lösung des Ausgangsproblems

$$f(\tilde{x}) - \hat{f}(\tilde{x})$$

**Rundungsfehler** endlicher Darstellungsbereich im Computer

$$\hat{f}(\tilde{x}) - \tilde{f}(\tilde{x})$$

Im Folgenden nur Eingangs- und Rundungsfehler, d.h. $\hat{f} = f$.

# Fehler

## Aufgabe

Es sei $f : V \to W$, auszuwerten, $\tilde{f}$ sei der dazu verwendete im Rechner realisierte Algorithmus und $\tilde{x} \in V$ die gestörte Eingabe zu $x \in V$.

## Notation

absoluter Fehler $|\Delta x| := |x - \tilde{x}|$, relativer Fehler $|\delta x| := \frac{|\Delta x|}{|x|}$

Es gilt offensichtlich $\tilde{x} = x(1 - \delta x)$

## Ziel

Abschätzung des Gesamtfehlers (durch Eingabe- und Rundungsfehler)

$$|f(x) - \tilde{f}(\tilde{x})|$$

# Kondition & Konditionszahlen eines Problems

$$\underbrace{\|f(x) - \tilde{f}(\tilde{x})\|}_{\text{Gesamtfehler}} = \|f(x) - f(\tilde{x}) + f(\tilde{x}) - \tilde{f}(\tilde{x})\|$$

$$\leq \underbrace{\|f(x) - f(\tilde{x})\|}_{\text{Kondition des Problems}} + \underbrace{\|f(\tilde{x}) - \tilde{f}(\tilde{x})\|}_{\text{Stabilität des Verfahrens}}$$

## Kondition des Problems

UNVERMEIDBAR

Ein ***Problem*** heißt ***gut konditioniert***, falls kleine Störungen in den Eingangsdaten kleine Änderungen der Ergebnisse hervorrufen. Anderenfalls heißt das Problem ***schlecht konditioniert***.

## Konditionszahlen

Verhältnis von Fehler im Resultat $y = f(x)$ zu Fehler in der Eingabe $x$.

$$\kappa_{\text{abs}} := \frac{|\Delta y|}{|\Delta x|}, \qquad \kappa := \kappa_{\text{rel}} := \frac{|\delta y|}{|\delta x|}$$

# Stabilität eines Algorithmus

$$\underbrace{\|f(x) - \tilde{f}(\tilde{x})\|}_{\text{Gesamtfehler}} = \|f(x) - f(\tilde{x}) + f(\tilde{x}) - \tilde{f}(\tilde{x})\|$$

$$\leq \underbrace{\|f(x) - f(\tilde{x})\|}_{\text{Kondition des Problems}} + \underbrace{\|f(\tilde{x}) - \tilde{f}(\tilde{x})\|}_{\text{Stabilität des Verfahrens}}$$

## Stabilität des Verfahrens

nicht einheitlich

1. Falls die durch den Algorithmus erzeugten Fehler in der Größenordnung des durch die Kondition des Problems unvermeidbaren Fehler bleiben, heißt ein ***Verfahren stabil***.
2. Falls Approximationsfehler durch den Algorithmus nicht übermäßig verstärkt werden, heißt ein ***Verfahren stabil***.
3. Ein ***Verfahren*** heißt ***stabil***, falls für jedes $x$ ein benachbartes $\tilde{x}$ existiert, so dass $f(\tilde{x}) - \tilde{f}(x)$ klein ist.

Anderenfalls heißt das Verfahren ***instabil***.

## Rundungsfehler

Stabile Verfahren sind relativ robust gegenüber Rundungsfehlern

# Kondition & Stabilität

$$\underbrace{\|f(x) - \tilde{f}(\tilde{x})\|}_{\text{Gesamtfehler}} = \|f(x) - f(\tilde{x}) + f(\tilde{x}) - \tilde{f}(\tilde{x})\|$$

$$\leq \underbrace{\|f(x) - f(\tilde{x})\|}_{\text{Kondition des Problems}} + \underbrace{\|f(\tilde{x}) - \tilde{f}(\tilde{x})\|}_{\text{Stabilität des Verfahrens}}$$

## Unvermeidbare und Vermeidbare Fehlerverstärkung

Die Kondition eines Problems beschreibt die ***unvermeidbare*** Fehlerverstärkung, d.h. unabhängig vom gewählten Algorithmus.
Verschiedene Algorithmen können zu unterschiedlichen Fehlerverstärkungen führen, d.h. prinzipiell ***vermeidbare*** Fehlerverstärkung.

## Faustregel

Anwendung eines stabilen Algorithmus auf ein gut konditioniertes Problem liefert gutes Ergebnis.
Falls der Algorithmus instabil oder das Problem schlecht konditioniert ist, ist das Ergebnis zu hinterfragen!

# Beispiele - Konditionszahl

Betrachte eine Funktion $f \in \mathcal{C}^1(\mathbb{R})$ und den absoluten Eingabefehler $\Delta x = x - \tilde{x}$. Dann gilt

$$\frac{f(x + \Delta x) - f(x)}{\Delta x} \approx f'(x)$$

mit $y = f(x)$ gilt also

$$\Delta y = (y + \Delta y) - y = f(x + \Delta x) - f(x) \approx f'(x)\, \Delta x$$

und wir erhalten

$$\kappa_{\text{abs}}(x) = \frac{|\Delta y|}{|\Delta x|} \approx |f'(x)|, \qquad \kappa(x) \approx \left|\frac{x\, f'(x)}{f(x)}\right|.$$

# Beispiele - Konditionszahl

Betrachte eine Funktion $f \in \mathcal{C}^1(\mathbb{R}^N, \mathbb{R})$ und den absoluten Eingabefehler $\Delta x = x - \tilde{x}$. Dann gilt

$$f(x + \Delta x) - f(x) \approx \nabla f(x) \cdot \Delta x = \sum_{d=1}^{N} \frac{\partial f(x)}{\partial x_d} \Delta x_d$$

mit $y = f(x)$ gilt also

$$\Delta y = (y + \Delta y) - y = f(x + \Delta x) - f(x) \approx \nabla f(x)\, \Delta x$$

und wir erhalten

$$\kappa_{\text{abs}} = \frac{|\Delta y|}{\|\Delta x\|} \approx \|\nabla f(x)\|.$$

norm-abhängige Definitionen der Konditionszahlen, z.B.

$$\kappa_{\text{abs}}^{\infty}(x) := \max_{d=1,\ldots,N} \left| \frac{\partial f(x)}{\partial x_d} \right|$$

Die relative Konditionszahl wird analog definiert über die Terme

$$\left| \frac{\partial f(x)}{\partial x_d} \frac{x_d}{f(x)} \right|$$

# Beispiele - Kondition

## Multiplikation

$$f : \mathbb{R}^2 \to \mathbb{R}, \quad f(x_1, x_2) = x_1 \cdot x_2, \quad \frac{\partial f(x)}{\partial x_1} = x_2, \quad \frac{\partial f(x)}{\partial x_2} = x_1$$

Relative Kondition:

$$\kappa^{\infty}(x) = \max\{|\frac{\partial f(x)}{\partial x_d} \frac{x_d}{f(x)}|\} = |\frac{x_1 \cdot x_2}{x_1 \cdot x_2}| = 1$$

## Addition

$$f : \mathbb{R}^2 \to \mathbb{R}, \quad f(x_1, x_2) = x_1 + x_2, \quad \frac{\partial f(x)}{\partial x_d} = 1$$

Relative Kondition:

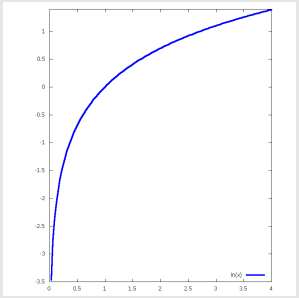
$$\kappa^{\infty}(x) = \max\{|\frac{\partial f(x)}{\partial x_d} \frac{x_d}{f(x)}|\} = |\frac{x_d}{x_1 + x_2}|$$

# Beispiele - Kondition

## Logarithmus

$$\ln : \mathbb{R}^+ \to \mathbb{R}, \quad \ln(x)' = \frac{1}{x}$$

$$\kappa_{\ln} \to \infty \text{ für } x \to 0$$



## Kondition (spezielle) Nullstellensuche

Bestimme die ***größte*** Nullstelle $x^2 + 2px - q = 0$ mit $p \gg q > 0$.
Diese ist gegeben durch $N(p,q) := -p + \sqrt{p^2+q}$.

$$\begin{aligned}
\Delta N(p,q) &= \frac{\partial N(p,q)}{\partial p}\Delta p + \frac{\partial N(p,q)}{\partial q}\Delta q \\
\delta N(p,q) &= \frac{1}{N(p,q)}\left(\frac{\partial N(p,q)}{\partial p} p\delta p + \frac{\partial N(p,q)}{\partial q} q\delta q\right) \\
&= \underbrace{-\frac{p}{\sqrt{p^2+q}}}_{|\cdot|<1}\delta p + \underbrace{\frac{p+\sqrt{p^2+q}}{2\sqrt{p^2+q}}}_{|\cdot|<1}\delta q
\end{aligned}$$

gut konditioniertes Problem

# Beispiel - Stabilität

Zur Berechnung von $N(p,q) := -p + \sqrt{p^2+q}$ setze

$$s := p^2, \quad t := s+q, \quad u := \sqrt{t}, \quad v := -p-u.$$

Definiere zwei Algorithmen zur Auswertung von $N(p,q)$

$$\tilde{N}_1(p,u) := -p+u, \quad \tilde{N}_2(q,u) := -\frac{q}{v} = -\frac{q}{-p-u}$$

Wurzelsatz von Vieta: $q = x_1 x_2$

## Stabilität der Algorithmen (bzgl. $\delta u$)

$$\delta\tilde{N}_1(u) = \frac{u}{-p+u}\delta u = \underbrace{\frac{1}{q}(p\sqrt{p^2+q}+p^2+q)}_{|\cdot|>\frac{2p^2}{q}\gg 1}\delta u$$

$$\delta\tilde{N}_2(u) = -\frac{\sqrt{p^2+q}}{\underbrace{p+\sqrt{p^2+q}}_{|\cdot|<1}}\delta v$$

# Zahldarstellung

## Zahlen

**Ganze Zahlen $\mathbb{Z}$:** Diskrete Menge, unendlich abzählbar

**Reelle Zahlen $\mathbb{R}$:** Kontinuum, unendlich überabzählbar

## Stellenwertsysteme

Wähle $\beta \in \mathbb{N}, \beta \geq 2$ die ***Basis***, und die ***Ziffern*** $m_i \in \{0, \ldots, \beta - 1\}$

$$x = \pm(\ldots m_n \beta^n + \ldots + m_1 \beta + m_0 + m_{-1}\beta^{-1} + \ldots m_{-k}\beta^{-k} + \ldots)$$

Zahldarstellung in Ziffernschreibweise

$$x = (\ldots m_{-n} \ldots m_{-1} m_0 m_1 \ldots m_{-k} \ldots)_\beta$$

# Zahldarstellung

## Festkommazahlen

Gegebene Stellenzahl $n \in \mathbb{N}$ und Vorkommastellenzahl $k \leq n$

$$x = \operatorname{sgn}(x) \sum_{i=k-n}^{k-1} m_i \beta^i, \qquad \text{für feste } n, k \in \mathbb{B}$$

- Feste Auflösung $\beta^{k-n}$ ($k = n$ ganzahlige Werte)
- Fester Zahlbereich z.B. $[-999.9999, 999.9999]$
- Häufiger Überlauf, Ergebnis häufig außerhalb des Wertebereichs
- Absolute Genauigkeit $x - \tilde{x}$ uniform
- Relative Genauigkeit $\frac{x-\tilde{x}}{x}$ nicht uniform

Physikalische Konstanten

- Plancksches Wirkungsquantum $h = 6.62606896\ldots \cdot 10^{-34} Js$
- Avogadro-Zahl $N_A = 6.02214179\ldots \cdot 10^{23}\,\mathrm{mol}^{-1}$

# Relative Genauigkeit bei Festkommazahlen

Betrachte Ergebnis $x$ einer Operation, das zwischen den kleinsten positiven Zahlen $z_{\min}$ und $2z_{\min}$

$$x = \frac{z_{\min} + 2z_{\min}}{2} = \frac{3}{2} z_{\min}$$

liegt. Dies ist nicht darstellbar und es muß gerundet werden (zu $z_{\min}$ oder $2z_{\min}$). Für den relativen Fehler ergibt sich

$$\left|\frac{x - z_{\min}}{x}\right| = \left|\frac{x - 2z_{\min}}{x}\right| = \left|\frac{\frac{1}{2} z_{\min}}{\frac{3}{2} z_{\min}}\right| = \frac{1}{3} = 33\%.$$

## Normalisierte Gleitkommazahlen

Zahlen der Gestalt

$$x = \pm m \cdot \beta^{\pm e}, \quad m = \pm \sum_{i=1}^{n} m_i \beta^{-i}, \quad e = \pm \sum_{j=0}^{s-1} e_j \beta^j$$

heißen ***normalisierte Gleitkommazahlen*** mit der ***Mantisse*** $m$, dem ***Exponenten*** $e$ und der ***Basis*** $\beta$. Der Raum der normalisierten Gleitkommazahlen mit $n$ Stellen für die Mantisse, und $s$ Stellen für den Exponenten zur Basis $\beta$ wird mit $\mathbb{F}(\beta, n, s)$ bezeichnet. Für die Eindeutigkeit der Darstellung für $x \neq 0$ ist die ***Normalisierung***

$$m = 0.m_1 m_2 m_3 \ldots m_n$$

mit $m_1 \neq 0$ bzw. $m_1 \in \{1, \ldots, \beta - 1\}$ bzw. $\beta^{-1} \leq |m| < 1$ notwendig.

# Eigenschaften

- Mit der Normierung $m_1 \neq 0$ gilt

$$\beta^{-1} \leq |m_1|\beta^{-1} \leq |\pm \sum_{i=1}^{n} m_i \beta^{-i}| = |m| < 1$$

- Maximaler Exponent ($\beta - 1$ ist maximale Ziffer):

$$e_{\max} = \pm \sum_{j=0}^{s-1} e_j^{\max} \beta^j = \pm \sum_{j=0}^{s-1} (\beta - 1)\beta^j = \beta^s - 1$$

- Maximales/Minimales Element $x_{\max\,/\,\min} = \pm(1 - \beta^{-n})\beta^{\beta^s - 1}$
- Kleinstes positives Element $x_{\min +} = \beta^{-\beta^s}$
- Maximaler relativer Abstand $\rho = \beta^{1-n}$ (Resolution)

# Rundung

$$\mathbb{F}(\beta, n, s) \subsetneq \mathbb{R}, \quad \text{round} : \mathbb{R} \to \mathbb{F}(\beta, n, s)$$

- Zu jedem $x \in \mathbb{R}$ exisitiert jeweils genau ein

  $$\text{floor}(x) := \max\{f \in \mathbb{F} \,|\, f \leq x\}, \quad \text{ceil}(x) := \min\{f \in \mathbb{F} \,|\, f \geq x\}.$$

- Forderungen an Rundungsvorschrift:

  Surjektivität: $\forall f \in \mathbb{F} \quad \exists x \in \mathbb{R} \quad \text{round}(x) = f$

  Idempotenz: $\text{round}(\text{round}(x)) = \text{round}(x)$

  Monotonie: $x \leq y \Rightarrow \text{round}(x) \leq \text{round}(y)$

- Rundungsfehler:

  $$\text{absolut } \Delta x = \text{round}(x) - x, \quad \text{relativ } \delta x = \frac{\Delta x}{x} \text{ für } x \neq 0$$

  $$\text{round}(x) = x \cdot (1 + \delta x) \quad \text{für alle } x \in \mathbb{R} \setminus \{0\}$$

# Rundungsarten

**Abrunden** (round to $-\infty$)

$$\text{round}_-(x) := \text{floor}(x) := \max\{f \in \mathbb{F} \,|\, f \leq x\}$$

**Aufrunden** (round to $+\infty$)

$$\text{round}_+(x) := \text{ceil}(x) := \min\{f \in \mathbb{F} \,|\, f \geq x\}$$

**Abhacken** (round to 0)

$$\text{round}_0(x) := \begin{cases} \text{ceil}(x) & x \geq 0 \\ \text{floor}(x) & x < 0 \end{cases}$$

**Korrektes Runden** (round to nearest)

$$\text{round}(x) := \begin{cases} \text{ceil}(x) & x > \frac{1}{2}(\text{ceil}(x) + \text{floor}(x)) \\ \text{floor}(x) & x < \frac{1}{2}(\text{ceil}(x) + \text{floor}(x)) \\ \text{specialcase} & x = \frac{1}{2}(\text{ceil}(x) + \text{floor}(x)) \end{cases}$$

# Schranken für den Rundungsfehler

## Abhacken

$$
\begin{aligned}
|\delta_0 x| &= \left|\frac{x - \mathrm{round}_0(x)}{x}\right| \\
&= \left|\frac{1}{x}\beta^e(0.m_1 m_2 \ldots m_n m_{n+1} m_{n_2} \ldots - 0.m_1 m_2 \ldots m_n)\right| \\
&= \left|\frac{\beta^e}{x}(0.00 \ldots m_{n+1} m_{n_2} \ldots)\right| \\
&= \left|\beta^{e-n} \underbrace{\frac{1}{x}(0.m_{n+1} m_{n_2} \ldots)}_{\beta^{-1} \le |m| < 1, x \ge \beta^{e-1}}\right| < \frac{\beta^{e-n}}{\beta^{e-1}} = \beta^{1-n} = \rho
\end{aligned}
$$

## Korrektes Runden & Maschinengenauigkeit

Für den relativen Rundungsfehler gilt

$$|\delta x| < \frac{1}{2}\beta^{1-n} = \frac{1}{2}\rho =: \epsilon_{\text{machine}}$$

und somit gibt es ein $\epsilon$ mit $|\epsilon| < \epsilon_{\text{machine}}$ mit

$$\mathrm{round}(x) = x(1+\epsilon), \quad \epsilon = \delta x = \frac{\mathrm{round}(x) - x}{x}$$

# ANSI/IEEE-754

Definiert mehrere Genauigkeitsstufen zur Basis $\beta = 2$: single, single-extended, double, double-extended

- double: 53 Bit für Mantisse (mit Vorzeichen), 11 Bit für Exponent
- Exponent wird ohne Vorzeichen gespeichert, z.B. $e = c - 1023$ mit $c \in \{1, 2, \ldots, 2046\}$ für double.
- hidden bit: Es gilt, außer für $x = 0$, $m_1 = 1$. Falls $x = 0$ gesondert codiert wird, muss man $m_1$ ***nicht*** speichern.
- Definiert obige vier Rundungsarten und fordert $\oplus, \ominus, \odot, \oslash$ werden korrekt gerundet (ebenso die Quadratwurzel).

# Beispiel

## Dramatische Konsequenzen

Im Zweiten Golfkrieg verfehlte am 25. 2. 1991 eine amerikanische Patriot-Rakete in Saudi-Arabien eine nahende irakische Scud-Rakete. Die Scud-Rakete schlug in eine Kaserne ein, wobei 28 US-Soldaten ums Leben kamen.

- Die interne Uhr der Patriot-Rakete speichert die seit dem Hochfahren des Systems verstrichene Zeit in Zehntelsekunden (24-Bit-Register).
- Da eine Zehntelsekunde im Binärsystem nicht exakt darstellbar ist, wurden nur die ersten 24 Stellen verwendet und ein daraus resultierender Rundungsfehler begangen:

$$0.1\,s \;=\; \left(0.000\overline{1100}\right)_2 s \;\approx\; 0.000110011001100110011001100\,s\,,$$

$$\text{Fehler} \;\approx 9.5 \cdot 10^{-8}\,s\,.$$

- Nach dem letzten Einschalten war das System nicht heruntergefahren worden.
- Nach 100 Betriebsstunden akkumulierte sich der Rundungsfehler zu

$$100 \cdot 60 \cdot\; 60 \cdot 10 \cdot 9.5 \cdot 10^{-8}\ \text{Zehntelsekunden} \;\approx\; 0.34\ \text{Sekunden}\,.$$

- In dieser Zeit legte die Scud-Rakete rund 570 Meter zurück und konnte damit von den Sensoren der Patriot-Rakete nicht mehr aufgespürt werden.

# Gleitkommaarithmetik: $+, -, \cdot, /$ vs. $\oplus, \ominus, \odot, \oslash$

## Problem

$\mathbb{F}(\beta, n, s)$ ist ***nicht abgeschlossen*** bzgl. der Operationen

$$+, -, \cdot, /$$

Beispielsweise: $x_{\max} + x_{\min +}$, $\beta^n + \beta^{-n}$

## Lösung

Definiere neue arithmetische Operationen $\oplus, \ominus, \odot, \oslash$

$$a \otimes b := \text{round}(a \times b) \text{ für } \times \in \{+, -, \cdot, /\} \text{ und } a, b \in \mathbb{F}(\beta, n, s)$$

Kommutativ: $a \oplus b = b \oplus a$, $a \odot b = b \odot a$

Assoziativ: $(a \oplus b) \oplus c \neq a \oplus (b \oplus c)$

Distributiv: $(a \oplus b) \odot c \neq (a \odot c) \oplus (b \odot c)$

# Eigenschaften

## Assoziativgesetz

Addiere $2^{20}$, $2^4$, $2^7$, $-2^3$ und $-2^{20}$. Mit 8 Binärstellen gilt

$$\begin{aligned}
(((2^{20} \oplus -2^{20}) \oplus 2^4) \oplus -2^3) \oplus 2^7 &= 136 \\
2^{20} \oplus (-2^{20} \oplus (2^4 \oplus (-2^3 \oplus 2^7))) &= 0 \\
(2^{20} \oplus (-2^{20} \oplus 2^4)) \oplus (-2^3 \oplus 2^7) &= 120 \\
(2^{20} \oplus ((-2^{20} \oplus 2^4) \oplus -2^3)) \oplus 2^7 &= 128
\end{aligned}$$

## Distributivgesetz

Mit 5 Dezimalstellen gilt

$$\begin{aligned}
(4.2832 - 4.2821) \cdot 5.7632 &= 0.00633952 \\
(4.2832 \ominus 4.2821) \odot 5.7632 &= 0.00633952 \\
(4.2832 \odot 5.7632) \ominus (4.2821 \odot 5.7632) &= (24.685 \ominus 24.679) \\
&= 0.006000
\end{aligned}$$

## Konsequenz

Mathematisch äquivalent Algorithmen zur Auswertung rationaler Ausdrücke können im Rechner ***völlig*** unterschiedliche Ergebnisse

# Zahldarstellungsfehler - Ein Beispiel

The Explosion of the Ariane 5 On June 4, 1996 an unmanned Ariane 5 rocket launched by the European Space Agency exploded just forty seconds after its lift-off from Kourou, French Guiana. Ariane explosion The rocket was on its first voyage, after a decade of development costing 7 billion. The destroyed rocket and its cargo were valued at 500 million. A board of inquiry investigated the causes of the explosion and in two weeks issued a report. It turned out that the cause of the failure was a software error in the inertial reference system. Specifically a 64 bit floating point number relating to the horizontal velocity of the rocket with respect to the platform was converted to a 16 bit signed integer. The number was larger than 32,767, the largest integer storeable in a 16 bit signed integer, and thus the conversion failed.

# Kummulierte Fehler - Ein Beispiel

## Question

How did the Vancouver Stock Exchange index gain 574.081 points while the stock prices were unchanged?

The stock index was calculated to four decimal places, but truncated (not rounded) to three. It was recomputed with each trade, some 3000 each day. The result was a loss of an index point a day, or 20 points a month. On Friday, November 25, 1983, the index stood at 524.811. After incorporating three weeks of work for consultants from Toronto and California computing the proper corrections for 22 months of compounded error, the index began Monday morning at 1098.892, up 574.081. (Toronto Star, 29 November 1983)

# Fehleranalyse

## Rundungsfehler

Die Gleitpunktarithmetik stellt eine wesentliche Fehlerquelle in numerischen Programmen dar.

## Fehlerfortpflanzung

Betrachte gerundete Größen $x$ und $y$ mit $|\delta x|, |\delta y| \ll 1$ und bestimme bei ***exakter*** Arithmetik den fortgepflanzten Fehler

$$\Delta(x \circ y) = (x + \Delta x) \circ (y + \Delta y) - x \circ y$$

für $\circ \in \{+, -, \cdot, /\}$. Es gilt für $\pm$

$$\frac{\Delta(x \pm y)}{x \pm y} = \frac{x}{x \pm y} \cdot \frac{\Delta x}{x} \pm \frac{y}{x \pm y} \cdot \frac{\Delta y}{y}$$

und in erster Näherung

$$\frac{\Delta(x \cdot y)}{x \cdot y} \approx \frac{\Delta x}{x} + \frac{\Delta y}{y}, \qquad \frac{\Delta(x/y)}{x/y} \approx \frac{\Delta x}{x} - \frac{\Delta y}{y}$$

# Auslöschung

## Definition

Falls $|x \pm y| \ll \max\{|x|, |y|\}$ treten große Verstärkung des relativen Fehlers bei $\pm$ auf.

Führende identische Ziffern löschen sich bei Subtraktion zwei Zahlen gleichen Vorzeichens aus, d.h. führende Nicht-Nullen verschwinden. Die Zahl relevanter Ziffern kann dabei drastisch abnehmen. Auslöschung wahrscheinlich, wenn beide Zahlen betragsmäßig ähnlich groß sind.

# Auslöschung: Beispiele

- Subtrahiere 4444.4444 von 4444.5555. Beide Zahlen haben 8 gültige Stellen, das Ergebnis jedoch nurmehr 4!
- Subtrahiere 999999 von einer Million. Wir nehmen eine Störung von $\pm 1$ bei beiden Zahlen an und erhalten neben dem exakten Ergebnis 1 die exakt berechnete Differenz der gestörten Zahlen:

$$(1000000 + 1) - (999999 - 1) \;=\; 3\,.$$

Somit ergibt sich als relativer Fehler

$$\frac{\Delta(x-y)}{x-y} \;=\; \frac{3-1}{1} \;=\; 2\,,$$

obwohl die relativen Störungen der Eingabedaten nur von der Größenordnung $O(10^{-6})$ sind.

# Rundungsfehleranalyse

## Vorwärtsanalyse

kaum durchführbar

Interpretiere das berechnete Ergebnis als gestörtes exaktes Ergebnis. Verfolge den Fehler von Schritt zu Schritt und schätze den akkumulierten Fehler für die Teilergebnisse ab.

$$\frac{|f(\tilde{x}) - \tilde{f}(\tilde{x})|}{|f(\tilde{x})|} \leq C_V \kappa \varepsilon$$

Für kleines $C_V$ heißt das Verfahren vorwärts-stabil.

## Rückwärtsanalyse

leichter durchführbar

Interpretiere das berechnete Ergebnis als exaktes Ergebnis zu gestörten Eingangsdaten $x + \Delta x$.

$$|\frac{\Delta x}{x}| \leq C_R \varepsilon$$

Für kleines $C_R$ heißt das Verfahren rückwärts-stabil.

Rückwärts-Stabilität impliziert Vorwärts-Stabilität.
Beides in der Regel nur qualitativ sinnvoll, der wahre Fehler wird überschätzt.

# Stabilitätsanalyse: Addition

## Modellproblem

$$f(x, y) = x + y$$

## Algorithmus

$$\tilde{f}(x, y) = x \oplus y$$

## Vorwärtsanalyse

$$\tilde{f}(x, y) = x \oplus y = (x + y)(1 + \delta(x + y))$$

## Rückwärtsanalyse

$$\tilde{f}(x, y) = x(1+\delta x)+y(1+\delta y) = f(x(1+\delta x), y(1+\delta y))$$

# Stabilitätsanalyse: Nullstellensuche