Numerische Verfahren

(für Studierende des Studienganges Bauingenieurwesen und Umwelttechnik)

http://www.tu-harburg.de/~matjz/work/lectures/numver/
http://www.tu-harburg.de/~matjz/work/exercises/numver/

Jens-Peter M. Zemke
Technische Universität Hamburg-Harburg
Institut für Numerische Simulation
zemke@tu-harburg.de
http://www.tu-harburg.de/~matjz/

Numerische Verfahren

Jens-Peter M. Zemke
zemke@tu-harburg.de

Institut für Numerische Simulation
Technische Universität Hamburg-Harburg

01.04.2008

TUHH
Technische Universität Hamburg-Harburg

Übersicht

Einleitung

Die Aufgabe der Numerischen Mathematik ist es, Algorithmen (d.h. Rechenvorschriften) für die näherungsweise numerische Lösung mathematischer Probleme der

Einleitung

Die Aufgabe der Numerischen Mathematik ist es, Algorithmen (d.h. Rechenvorschriften) für die näherungsweise numerische Lösung mathematischer Probleme der

- Naturwissenschaften,

Einleitung

Die Aufgabe der Numerischen Mathematik ist es, Algorithmen (d.h. Rechenvorschriften) für die näherungsweise numerische Lösung mathematischer Probleme der

- Naturwissenschaften,
- Technik,

Einleitung

Die Aufgabe der Numerischen Mathematik ist es, Algorithmen (d.h. Rechenvorschriften) für die näherungsweise numerische Lösung mathematischer Probleme der

- Naturwissenschaften,
- Technik,
- Ökonomie

Einleitung

Die Aufgabe der Numerischen Mathematik ist es, Algorithmen (d.h. Rechenvorschriften) für die näherungsweise numerische Lösung mathematischer Probleme der

- Naturwissenschaften,
- Technik,
- Ökonomie
- u.s.w.

bereitzustellen und zu diskutieren.

Einleitung

Die Aufgabe der Numerischen Mathematik ist es, Algorithmen (d.h. Rechenvorschriften) für die näherungsweise numerische Lösung mathematischer Probleme der

- Naturwissenschaften,
- Technik,
- Ökonomie
- u.s.w.

bereitzustellen und zu diskutieren.

Gesichtspunkte bei der Bewertung eines Algorithmus (und beim Vergleich von Algorithmen) sind der Aufwand (z.B. die Anzahl der Operationen), der Speicherplatzbedarf und eine Fehleranalyse.

Einleitung

Die Aufgabe der Numerischen Mathematik ist es, Algorithmen (d.h. Rechenvorschriften) für die näherungsweise numerische Lösung mathematischer Probleme der

- Naturwissenschaften,
- Technik,
- Ökonomie
- u.s.w.

bereitzustellen und zu diskutieren.

Gesichtspunkte bei der Bewertung eines Algorithmus (und beim Vergleich von Algorithmen) sind der Aufwand (z.B. die Anzahl der Operationen), der Speicherplatzbedarf und eine Fehleranalyse.

Einleitung

Die Aufgabe der Numerischen Mathematik ist es, Algorithmen (d.h. Rechenvorschriften) für die näherungsweise numerische Lösung mathematischer Probleme der

- Naturwissenschaften,
- Technik,
- Ökonomie
- u.s.w.

bereitzustellen und zu diskutieren.

Gesichtspunkte bei der Bewertung eines Algorithmus (und beim Vergleich von Algorithmen) sind der Aufwand (z.B. die Anzahl der Operationen), der Speicherplatzbedarf und eine Fehleranalyse.

Einleitung

Die Aufgabe der Numerischen Mathematik ist es, Algorithmen (d.h. Rechenvorschriften) für die näherungsweise numerische Lösung mathematischer Probleme der

- Naturwissenschaften,
- Technik,
- Ökonomie
- u.s.w.

bereitzustellen und zu diskutieren.

Gesichtspunkte bei der Bewertung eines Algorithmus (und beim Vergleich von Algorithmen) sind der Aufwand (z.B. die Anzahl der Operationen), der Speicherplatzbedarf und eine Fehleranalyse.

Man unterscheidet drei Typen von Fehlern nach ihren Quellen.

Einleitung

Die drei Typen von Fehlern nach Quellen sind:

Einleitung

Die drei Typen von Fehlern nach Quellen sind:

Datenfehler: Die Eingangsdaten einer Aufgabe können fehlerhaft sein, wenn sie etwa aus vorhergehenden Rechnungen, physikalischen Messungen oder empirischen Untersuchungen stammen.

Einleitung

Die drei Typen von Fehlern nach Quellen sind:

Datenfehler: Die Eingangsdaten einer Aufgabe können fehlerhaft sein, wenn sie etwa aus vorhergehenden Rechnungen, physikalischen Messungen oder empirischen Untersuchungen stammen.

Verfahrensfehler: Dies sind Fehler, die dadurch entstehen, dass man ein Problem diskretisiert (z.B. eine Differentialgleichung durch eine Differenzengleichung ersetzt) oder ein Iterationsverfahren nach endlich vielen Schritten abbricht.

Einleitung

Die drei Typen von Fehlern nach Quellen sind:

Datenfehler: Die Eingangsdaten einer Aufgabe können fehlerhaft sein, wenn sie etwa aus vorhergehenden Rechnungen, physikalischen Messungen oder empirischen Untersuchungen stammen.

Verfahrensfehler: Dies sind Fehler, die dadurch entstehen, dass man ein Problem diskretisiert (z.B. eine Differentialgleichung durch eine Differenzengleichung ersetzt) oder ein Iterationsverfahren nach endlich vielen Schritten abbricht.

Rundungsfehler: Bei der Ausführung der Rechenoperationen auf einer Rechenanlage entstehen Fehler, da das Ergebnis (aber auch schon alle Zwischenergebnisse) nur im Rahmen eines begrenzten Zahlbereichs (sog. Maschinenzahlen) dargestellt werden kann, also gerundet werden muss.

Einleitung

Die Frage, wie sich Datenfehler auf die Lösung einer Aufgabe auswirken, nennt man das Konditionsproblem der Aufgabe.

Einleitung

Die Frage, wie sich Datenfehler auf die Lösung einer Aufgabe auswirken, nennt man das Konditionsproblem der Aufgabe.

Bewirken kleine Eingangsfehler auch nur kleine Ergebnisfehler, so nennt man das Problem gut konditioniert, anderenfalls schlecht konditioniert.

Einleitung

Die Frage, wie sich Datenfehler auf die Lösung einer Aufgabe auswirken, nennt man das Konditionsproblem der Aufgabe.

Bewirken kleine Eingangsfehler auch nur kleine Ergebnisfehler, so nennt man das Problem gut konditioniert, anderenfalls schlecht konditioniert.

Einleitung

Die Frage, wie sich Datenfehler auf die Lösung einer Aufgabe auswirken, nennt man das Konditionsproblem der Aufgabe.

Bewirken kleine Eingangsfehler auch nur kleine Ergebnisfehler, so nennt man das Problem gut konditioniert, anderenfalls schlecht konditioniert.

Die Kondition eines Problems hängt nicht nur von der Aufgabenstellung,

Einleitung

Die Frage, wie sich Datenfehler auf die Lösung einer Aufgabe auswirken, nennt man das Konditionsproblem der Aufgabe.

Bewirken kleine Eingangsfehler auch nur kleine Ergebnisfehler, so nennt man das Problem gut konditioniert, anderenfalls schlecht konditioniert.

Die Kondition eines Problems hängt nicht nur von der Aufgabenstellung,

z.B. „Lösung eines linearen Gleichungssystems“,

Einleitung

Die Frage, wie sich Datenfehler auf die Lösung einer Aufgabe auswirken, nennt man das Konditionsproblem der Aufgabe.

Bewirken kleine Eingangsfehler auch nur kleine Ergebnisfehler, so nennt man das Problem gut konditioniert, anderenfalls schlecht konditioniert.

Die Kondition eines Problems hängt nicht nur von der Aufgabenstellung,

z.B. „Lösung eines linearen Gleichungssystems“,

sondern auch von den Eingangsdaten,

Einleitung

Die Frage, wie sich Datenfehler auf die Lösung einer Aufgabe auswirken, nennt man das Konditionsproblem der Aufgabe.

Bewirken kleine Eingangsfehler auch nur kleine Ergebnisfehler, so nennt man das Problem gut konditioniert, anderenfalls schlecht konditioniert.

Die Kondition eines Problems hängt nicht nur von der Aufgabenstellung,

z.B. „Lösung eines linearen Gleichungssystems“,

sondern auch von den Eingangsdaten,

z.B. den Koeffizienten der gegebenen Matrix

ab.

Einleitung

Beispiel 1.1

Das lineare Gleichungssystem

$$\begin{pmatrix} 1 & a \\ 0 & 1 \end{pmatrix} \begin{pmatrix} x \\ y \end{pmatrix} = \begin{pmatrix} 1 \\ 0 \end{pmatrix}, \quad a \in \mathbb{R} \tag{1.1}$$

besitzt die Lösung

$$x_0 = 1, \quad y_0 = 0.$$

Einleitung

Beispiel 1.1

Das lineare Gleichungssystem

$$\begin{pmatrix} 1 & a \\ 0 & 1 \end{pmatrix} \begin{pmatrix} x \\ y \end{pmatrix} = \begin{pmatrix} 1 \\ 0 \end{pmatrix}, \quad a \in \mathbb{R} \tag{1.1}$$

besitzt die Lösung

$$x_0 = 1, \quad y_0 = 0.$$

Das gestörte Gleichungssystem

$$\begin{pmatrix} 1 & a \\ 0 & 1 \end{pmatrix} \begin{pmatrix} x \\ y \end{pmatrix} = \begin{pmatrix} 1 \\ \delta \end{pmatrix}$$

hat die Lösung

$$x_\delta = 1 - \delta a, \quad y_\delta = \delta.$$

Einleitung

Beispiel 1.1 (Fortsetzung)

Damit gilt

$$\begin{pmatrix} x_0 \\ y_0 \end{pmatrix} - \begin{pmatrix} x_\delta \\ y_\delta \end{pmatrix} = \delta \begin{pmatrix} -a \\ 1 \end{pmatrix}.$$

Einleitung

Beispiel 1.1 (Fortsetzung)

Damit gilt

$$\begin{pmatrix} x_0 \\ y_0 \end{pmatrix} - \begin{pmatrix} x_\delta \\ y_\delta \end{pmatrix} = \delta \begin{pmatrix} -a \\ 1 \end{pmatrix}.$$

Änderungen der rechten Seite

$$\begin{pmatrix} 1 \\ 0 \end{pmatrix} \rightarrow \begin{pmatrix} 1 \\ \delta \end{pmatrix}$$

in der zweiten Komponente

Einleitung

Beispiel 1.1 (Fortsetzung)

Damit gilt

$$\begin{pmatrix} x_0 \\ y_0 \end{pmatrix} - \begin{pmatrix} x_\delta \\ y_\delta \end{pmatrix} = \delta \begin{pmatrix} -a \\ 1 \end{pmatrix}.$$

Änderungen der rechten Seite

$$\begin{pmatrix} 1 \\ 0 \end{pmatrix} \to \begin{pmatrix} 1 \\ \delta \end{pmatrix}$$

in der zweiten Komponente werden also mit dem Faktor

$$\sqrt{1 + a^2}$$

(bzgl. der Euklidischen Norm) verstärkt.

Einleitung

Beispiel 1.1 (Fortsetzung)

Damit gilt

$$\begin{pmatrix} x_0 \\ y_0 \end{pmatrix} - \begin{pmatrix} x_\delta \\ y_\delta \end{pmatrix} = \delta \begin{pmatrix} -a \\ 1 \end{pmatrix}.$$

Änderungen der rechten Seite

$$\begin{pmatrix} 1 \\ 0 \end{pmatrix} \to \begin{pmatrix} 1 \\ \delta \end{pmatrix}$$

in der zweiten Komponente werden also mit dem Faktor

$$\sqrt{1 + a^2}$$

(bzgl. der Euklidischen Norm) verstärkt. Damit ist das Problem für kleine $|a|$ gut und für große $|a|$ schlecht konditioniert. □

Einleitung

Ein numerisches Verfahren heißt gut konditioniert (numerisch stabil), wenn die gelieferte Lösung eines gegebenen Problems die exakte Lösung eines Problems ist, das aus dem ursprünglichen Problem durch geringe Änderung der Eingangsdaten hervorgeht.

Einleitung

Ein numerisches Verfahren heißt gut konditioniert (numerisch stabil), wenn die gelieferte Lösung eines gegebenen Problems die exakte Lösung eines Problems ist, das aus dem ursprünglichen Problem durch geringe Änderung der Eingangsdaten hervorgeht.

Anderenfalls heißt das numerische Verfahren schlecht konditioniert (oder numerisch instabil).

Einleitung

Ein numerisches Verfahren heißt gut konditioniert (numerisch stabil), wenn die gelieferte Lösung eines gegebenen Problems die exakte Lösung eines Problems ist, das aus dem ursprünglichen Problem durch geringe Änderung der Eingangsdaten hervorgeht.

Anderenfalls heißt das numerische Verfahren schlecht konditioniert (oder numerisch instabil).

Beispiel 1.2

Das Integral

$$\int_0^1 \frac{x^{20}}{x+10}\,dx$$

kann auf folgende Weise berechnet werden:

Einleitung

Beispiel 1.2 (Fortsetzung)

Für

$$y_n := \int_0^1 \frac{x^n}{x+10}\,dx$$

gilt

Einleitung

Beispiel 1.2 (Fortsetzung)

Für

$$y_n := \int_0^1 \frac{x^n}{x+10}\,dx$$

gilt

$$y_n + 10y_{n-1} = \int_0^1 \frac{x^n + 10x^{n-1}}{x+10}\,dx = \int_0^1 x^{n-1}\,dx = \frac{1}{n}, \tag{1.2}$$

$$y_0 = \int_0^1 \frac{dx}{x+10} = \ln(1.1).$$

Einleitung

Beispiel 1.2 (Fortsetzung)

Wertet man die Differenzenformel

$$y_n = \frac{1}{n} - 10y_{n-1}$$

für $n = 1, \ldots, 20$ aus, so erhält man die zweite Spalte der Tabelle 1.1.

Einleitung

Beispiel 1.2 (Fortsetzung)

Wertet man die Differenzenformel

$$y_n = \frac{1}{n} - 10y_{n-1}$$

für $n = 1, \ldots, 20$ aus, so erhält man die zweite Spalte der Tabelle 1.1. Obwohl das Problem, das Integral zu berechnen, gut konditioniert ist, erhält man ein unbrauchbares Resultat. Das Verfahren ist also instabil.

Einleitung

Beispiel 1.2 (Fortsetzung)

Wertet man die Differenzenformel

$$y_n = \frac{1}{n} - 10y_{n-1}$$

für $n = 1, \ldots, 20$ aus, so erhält man die zweite Spalte der Tabelle 1.1. Obwohl das Problem, das Integral zu berechnen, gut konditioniert ist, erhält man ein unbrauchbares Resultat. Das Verfahren ist also instabil.

Löst man (1.2) nach y_{n-1} auf,

$$y_{n-1} = 0.1 \left(\frac{1}{n} - y_n \right),$$

Einleitung

Beispiel 1.2 (Fortsetzung)

Wertet man die Differenzenformel

$$y_n = \frac{1}{n} - 10y_{n-1}$$

für $n = 1, \ldots, 20$ aus, so erhält man die zweite Spalte der Tabelle 1.1. Obwohl das Problem, das Integral zu berechnen, gut konditioniert ist, erhält man ein unbrauchbares Resultat. Das Verfahren ist also instabil.

Löst man (1.2) nach y_{n-1} auf,

$$y_{n-1} = 0.1\left(\frac{1}{n} - y_n\right),$$

und startet man mit der groben Näherung $y_{30} = 0$, so erhält man $y_{20}, \ldots, y_0$ mit einer Genauigkeit von wenigstens 10 gültigen Stellen, siehe hierzu die dritte Spalte von Tabelle 1.1. □

Einleitung, Tabelle 1.1

n	vorwärts	rückwärts
0	$9.53101798043249E-0002$	$9.53101798043249E-0002$
1	$4.68982019567514E-0002$	$4.68982019567514E-0002$
2	$3.10179804324860E-0002$	$3.10179804324860E-0002$
3	$2.31535290084733E-0002$	$2.31535290084733E-0002$
4	$1.84647099152673E-0002$	$1.84647099152671E-0002$
5	$1.53529008473272E-0002$	$1.53529008473289E-0002$
⋮	⋮	⋮
12	$7.03899326661614E-0003$	$7.03897613105546E-0003$
13	$6.53314425691553E-0003$	$6.53331561252233E-0003$
14	$6.09712885941609E-0003$	$6.09541530334817E-0003$
15	$5.69537807250574E-0003$	$5.71251363318499E-0003$
16	$5.54621927494255E-0003$	$5.37486366815006E-0003$
17	$3.36133666233920E-0003$	$5.07489273026414E-0003$
18	$2.19421889321635E-0002$	$4.80662825291418E-0003$
19	$-1.66790310374267E-0001$	$4.56529641822660E-0003$
20	$1.71790310374267E+0000$	$4.34703581773402E-0003$
21		$4.14868944170746E-0003$
22		$3.96765103747089E-0003$
23		$3.80175049485636E-0003$
24		$3.64916171810310E-0003$
25		$3.50838281896903E-0003$
⋮	⋮	⋮

Zahlendarstellung

Üblicherweise stellt man Zahlen im Dezimalsystem dar, d.h. eine reelle Zahl x wird durch die Koeffizienten α_i der Dezimaldarstellung von x festgelegt:

$$x = \pm \left(\alpha_n \cdot 10^n + \alpha_{n-1} \cdot 10^{n-1} + \cdots + \alpha_0 \cdot 10^0 + \alpha_{-1} \cdot 10^{-1} + \cdots \right)$$

$$\text{mit} \quad \alpha_i \in \{0, 1, \ldots, 9\}.$$

Abgekürzt schreibt man dafür auch

$$\pm \alpha_n \alpha_{n-1} \ldots \alpha_0 . \alpha_{-1} \alpha_{-2} \ldots .$$

Zahlendarstellung

Aus technischen Gründen arbeiten digitale Rechenanlagen im Dualsystem (zur Basis 2) oder im Hexadezimalsystem (zur Basis 16).

Wir bleiben der Anschauung halber bei der Basis 10.

Für die interne Darstellung einer Zahl in einem Rechner steht nur eine feste Anzahl t (=Wortlänge) von Dezimalstellen zur Verfügung.

Diese Wortlänge wird auf zwei Arten zur Darstellung einer Zahl benutzt:

- **Festpunktdarstellung**
- **Gleitpunktdarstellung**

Festpunktdarstellung

Bei der **Festpunktdarstellung** sind n_1 und n_2, die Zahl der Stellen vor und nach dem Dezimalpunkt, festgelegt:

Beispiel ($t = 8$, $n_1 = 3$, $n_2 = 5$)

$$\begin{array}{rcl|l} 30.411 & \rightarrow & 030 & 41100 \\ 0.0023 & \rightarrow & 000 & 00230. \end{array}$$

□

Wegen des verhältnismäßig kleinen Bereichs darstellbarer Zahlen wird mit Festpunktzahlen nur dann gearbeitet, wenn keine zu großen Unterschiede in der Größenordnung der auftretenden Zahlen bestehen (vor allem im kaufmännisch-organisatorischen Bereich: Stückzahlen: $n_2 = 0$, Preise: $n_2 = 2$.)

Gleitpunktdarstellung

Schreibt man x in der **Gleitpunktdarstellung**, so liegt die Mantissenlänge $t = n_1 + n_2$ fest; die Lage des Dezimalpunktes wird durch einen Exponenten markiert:

$$x = \pm \left(\alpha_{n_1-1} \cdot 10^{n_1-1} + \alpha_{n_1-2} \cdot 10^{n_1-2} + \cdots + \alpha_0 \cdot 10^0 + \right.$$
$$\left. \cdots + \alpha_{-1} \cdot 10^{-1} + \alpha_{-n_2} \cdot 10^{-n_2} \right).$$

$$= \pm \left(\alpha_{n_1-1} \cdot 10^{-1} + \alpha_{n_1-2} \cdot 10^{-2} + \right.$$
$$\left. \cdots + \alpha_{-n_2} \cdot 10^{-(n_1+n_2)} \right) \cdot 10^{n_1}$$

$$= \pm 0. \underbrace{\alpha_{n_1-1}\alpha_{n_1-2} \ldots \alpha_{-n_2}}_{:= \text{Mantisse}} \cdot 10^{n_1}, \qquad n_1 := \text{Exponent}$$

Gleitpunktdarstellung

Beispiel ($t = 4$)

$$0.0023 \rightarrow 0.0023_{10}0 \quad \text{oder} \quad 0.2300_{10} - 2. \qquad \square$$

Die Gleitpunktdarstellung einer Zahl ist i.A. nicht eindeutig. Sie heißt normalisiert, falls $x = 0$ oder für die erste Ziffer $\alpha_1 \neq 0$ gilt. Normalisierte Gleitpunktzahlen ungleich Null sind eindeutig. Daher betrachten wir von nun an nur noch normalisierte Gleitpunktzahlen.

Rundungsfehler und Gleitpunktrechnung

Die Menge $\mathbb{F}$ (engl. *floating point numbers*) der in einer Maschine darstellbaren Zahlen ist endlich (die Mantissenlänge t ist endlich, und für die Darstellung des Exponenten stehen auch nur $e < \infty$ viele Stellen zur Verfügung).
Für ein gegebenes $x \in \mathbb{R}$ bezeichnen wir mit $\mathrm{fl}(x) \in \mathbb{F}$ eine Maschinenzahl, durch die x am besten approximiert wird, d.h.

$$|\mathrm{fl}(x) - x| \leqslant |a - x| \quad \text{für alle} \quad a \in \mathbb{F}.$$

Diese Vorschrift ist noch nicht eindeutig (wird 0.5 auf- oder abgerundet?). Wir setzen fest:
Sei $x \in \mathbb{R}$ gegeben mit der normalisierten Gleitpunktdarstellung

$$x = \pm 0.\alpha_1\alpha_2 \ldots \alpha_t\alpha_{t+1} \cdot \cdots \cdot 10^n,$$

Rundungsfehler und Gleitpunktrechnung

dann wird x durch die folgende Vorschrift gerundet:

$$\mathrm{fl}(x) = \begin{cases} \pm\ 0.\alpha_1\alpha_2\ldots\alpha_t \cdot 10^n & , \quad \text{falls} \quad 0 \leqslant \alpha_{t+1} \leqslant 4 \\ \pm(0.\alpha_1\alpha_2\ldots\alpha_t + 10^{-t}) \cdot 10^n & , \quad \text{falls} \quad 5 \leqslant \alpha_{t+1} \leqslant 9. \end{cases}$$

Für den absoluten Fehler gilt

$$|x - \mathrm{fl}(x)| \leqslant \frac{1}{2} \cdot 10^{n-t},$$

und für den relativen Fehler

$$\left|\frac{x - \mathrm{fl}(x)}{x}\right| \leqslant \frac{1}{2} \cdot 10^{n-t} 10^{-n+1} = 5 \cdot 10^{-t} \qquad (\alpha_1 \neq 0)\,.$$

Mit der Abkürzung $\mathbf{u} = 5 \cdot 10^{-t}$ (Maschinengenauigkeit) gilt also

$$\mathrm{fl}(x) = (1+\varepsilon)x, \quad |\varepsilon| \leqslant \mathbf{u}. \tag{1.3}$$

Rundungsfehler und Gleitpunktrechnung

$\mathrm{fl}(x)$ ist nicht stets eine Maschinenzahl, da nicht beliebig große oder kleine Zahlen dargestellt werden können:

Beispiel ($t = 4$, $e = 2$)

Exponentenüberlauf:

$$\mathrm{fl}(0.99997_{10}99) = 0.1000_{10}100 \notin \mathbb{F},$$

Exponentenunterlauf:

$$\mathrm{fl}(0.01234_{10} - 99) = 0.1234_{10} - 100 \notin \mathbb{F}.$$ □

Setzt man im zweiten Fall $\mathrm{fl}(0.01234_{10} - 99) = 0$ oder $0.0123_{10} - 99$, so gilt zwar $\mathrm{fl}(\,\cdot\,) \in \mathbb{F}$, aber es ist nicht mehr (1.3) erfüllt. Da bei den heutigen Anlagen e genügend groß ist, tritt Exponentenüberlauf oder -unterlauf nur sehr selten auf. Wir nehmen daher für das Weitere $e = \infty$ an, so dass bei der Rundung (1.3) gilt.

Rundungsfehler und Gleitpunktrechnung

Offensichtlich gilt

$$x, y \in \mathbb{F} \qquad \not\Rightarrow \qquad x \pm y, x \cdot y, x/y \in \mathbb{F}.$$

Statt der Operationen $+, -, \cdot, /$ sind daher auf dem Rechner als Ersatz die **Gleitpunktoperationen** $\boxplus, \boxminus, \boxdot, \boxslash$ realisiert, die man mit Hilfe von fl so beschreiben kann ($x, y \in \mathbb{F}$):

$$x \circledcirc y := \mathrm{fl}(x \circ y) \quad \forall \circ \in \{+, -, \cdot, /\}.$$

(In der Maschine wird die Operation „exakt“ ausgeführt, danach wird gerundet). Wegen (1.3) gilt

$$x \circledcirc y = (x \circ y)(1 + \varepsilon), \quad |\varepsilon| \leqslant \mathbf{u},$$

für jede der Operationen $\circ \in \{+, -, \cdot, /\}$. (Der relative Fehler hat also die Größenordnung der Maschinengenauigkeit).

Rundungsfehler und Gleitpunktrechnung

Waren aber x und y keine Maschinenzahlen, so wird zunächst gerundet und dann $\mathrm{fl}(x) \boxdot \mathrm{fl}(y)$ berechnet.
Hierfür gilt wegen $\mathrm{fl}(x) = (1+\varepsilon_x)x$, $\mathrm{fl}(y) = (1+\varepsilon_y)y$

$$\frac{\mathrm{fl}(x) + \mathrm{fl}(y) - (x+y)}{x+y} = \frac{x}{x+y}\,\varepsilon_x + \frac{y}{x+y}\,\varepsilon_y.$$

Haben also bei der Addition die Summanden entgegengesetztes Vorzeichen, den gleichen Exponenten und gleiche führende Ziffern der Mantisse, so ist $x+y$ klein gegen x und gegen y und der relative Fehler wird verstärkt. (Man spricht dann von Auslöschung).

Rundungsfehler und Gleitpunktrechnung

Beispiel ($t = 6$, $x = 1234.567$, $y = -1234.60$)

Es gilt

$$\left|\frac{(\mathrm{fl}(x) + \mathrm{fl}(y)) - (x + y)}{x + y}\right| = \left|\frac{-0.03 - (-0.033)}{-0.033}\right| = \frac{1}{11},$$

aber

$$\varepsilon_x = \left|\frac{\mathrm{fl}(x) - x}{x}\right| = \frac{0.003}{1234.567} \approx 2.5 \cdot 10^{-6}, \quad \varepsilon_y = 0.$$

□

Die Operationen $\cdot$ (und $/$) sind wegen

$$\frac{\mathrm{fl}(x) \cdot \mathrm{fl}(y) - x \cdot y}{x \cdot y} = \varepsilon_x + \varepsilon_y + \varepsilon_x \cdot \varepsilon_y \approx \varepsilon_x + \varepsilon_y$$

für die Fehlerfortpflanzung in einer Rechnung unkritisch.