Vorlesung 3

Rechnen mit Fließkommazahlen

3.1 Fließkommaarithmetik

Nun möchte man mit Fließkommazahlen natürlich auch rechnen, d.h. wir benötigen zweistellige Operationen

$$\circledast \ : \ \mathbb{F} \times \mathbb{F} \to \mathbb{F} \qquad \circledast \in \{\oplus, \ominus, \odot, \oslash\}, \tag{3.1}$$

welche den Grundrechenarten $+, -, \cdot, /$ auf den reellen Zahlen entsprechen. In diesem Kapitel steht $\mathbb{F}$ für ein beliebiges Fließkommasystem $\mathbb{F}(\beta, r, s)$. Die genauen Parameter werden nur dann spezifiziert wenn dies wichtig ist.

Die Problematik der Fließkommaoperationen ist, dass in der Regel selbst für zwei Fließkommazahlen $x, y \in \mathbb{F}$ das Ergebnis der exakten Verknüpfung $x * y$ nicht in notwendigerweise in $\mathbb{F}$ liegt. So wird sich in der Regel bei der Multiplikation die Zahl der Mantissenstellen verdoppeln. Auch das Ergebnis einer Addition zweier Zahlen mit verschiedenen Exponenten erfordert in der Regel mehr Mantissenstellen.

Ein naheliegende Definition für die Fließkommaarithmetik ist daher

$$x \circledast y = \mathrm{rd}(x * y) \qquad * \in \{+, -, \cdot, /\}. \tag{3.2}$$

Ein Fließkommaarithmetik mit dieser Eigenschaft nennt man *exakt gerundet*.

Eine triviale Realisierung exakter Rundung ist jedoch nicht effizient. Betrachten wir die Addition von $x = 10^{10}$ und $y = 10^{-10}$ in $\mathbb{F}(10, 4, 2)$, Bevor die Mantissen addiert werden können sind beide Summanden auf den gleichen Exponenten zu bringen, also $x = 0.1 \cdot 10^{11}$ und $y = 10^{-10} = 10^{-10-11} \cdot 10^{11} = 10^{-21} \cdot 10^{11}$. Dementsprechend bräuchte man 21 Nachkommstellen um in einem Zwischenschritt $x*y$ exakt zu berechnen. Die anschließende Rundung schneidet viele der zusätzlichen Stellen wieder ab, das ist sehr ineffizient. Die Rundung gleich nach Egalisierung der Exponenten durch zuführen ist allerdings gefährlich und kann zu großen relativen Fehlern führen. Es zeigt sich, dass folgende Vorgehensweise zum Ziel führt:

1) Bringe beide Argumente auf den gleichen Exponenten, also die betragsmäßig kleinere Zahl auf den Exponenenten der betragsmäßig Größeren. Diese ist dann nicht mehr normiert.

2) Runde die geschobene Zahl auf $r + 2$ Mantissenstellen.

3) Addiere beide Zahlen mit $r + 2$ Mantissenstellen. Diese zusätzlichen Stellen nennt man *guard digits*.

4) Runde das Ergebnis auf r Stellen.

Mit zwei zusätzlichen Stellen erreicht man so ein exakte Rundung. Im IEEE 754 Standard sind sowohl die Grundoperationen als auch die Berechnung der Quadratwurzel exakt gerundet.

Die Fließkommaarithmetik unterscheidet sich von der Arithmetik in den reellen Zahlen, hier eine Liste relevanter Eigenschaften:

a) Es gibt Zahlen $y \in \mathbb{F}$ so dass $x \oplus y = x$ gilt.

b) Assoziativ und Distributivgesetz gelten nicht (unbedingt). Es kommt also auf die Reihenfolge der Operationen an. Z.B. gilt mit dem y aus a): $(x \oplus y) \ominus x = 0$ und $y \oplus (x \ominus x) = y$. Hier haben wir ausgenutzt, dass $x \ominus x = 0$.

c) Es gilt jedoch das Kommutativgesetz!

d) Auch folgende einfache Regeln gelten

$$\begin{aligned}(-x) \odot y &= -(x \odot y)\\ 1 \odot x &= x\\ x \odot y &= 0 \quad \text{nur dann wenn } x = 0 \text{ oder } y = 0\\ x \odot z &\leq y \odot z \quad \text{wenn } x \leq y \text{ und } z > 0\end{aligned}$$

3.2 Fehleranalyse

Wir kommen nun zu den Auswirkungen von Rundungsfehlern in Rechnungen die aus einer Folge von Grundoperation zusammengesetzt sind. Abstrakt betrachten wir die Berechnung als Auswertung einer vektorwertigen Funktion

$$F \ : \ \mathbb{R}^m \to \mathbb{R}^n.$$

Die Berechnung von F im Computer wird durch eine entsprechende Funktion

$$F' \ : \ \mathbb{F}^m \to \mathbb{F}^n$$

realisiert. F' wird durch endlich viele Elementaroperationen

$$F'(x') = \varphi_l(\dots \varphi_2(\varphi_1(x')) \dots)$$

realisiert. Jedes φ_i führt eine Grundoperation aus $\{\oplus, \ominus, \odot, \oslash\}$ durch und steuert einen unbekannten Teilfehler bei. Die numerische Realisierung F' von F ist in der Regel nicht eindeutig. Zum einen gibt es mathematisch äquivalente Möglichkeiten einen Ausdruck zu berechen, z.B $x(y+z) = xy + xz$ und selbst unterschiedlich Reihenfolgen der Zwischenschritte können zu unterschiedlichen Ergebnissen in $\mathbb{F}$ führen.

Im größeren praktischen Kontext sind auch die Eingaben der exakten Funktion F mit Unsicherheiten behaftet. Dann ist es interessant zu untersuchen wie sich Variationen in der Eingabe x gegenüber den akkumulierten Rundungsfehlern verhalten. Dies erreicht man mit der nun folgenden Analyse. Dazu betrachten wir die Aufspaltung

$$F(x) - F'(\mathrm{rd}(x)) = \underbrace{F(x) - F(\mathrm{rd}(x))}_{\text{Konditionsanalyse}} + \underbrace{F(\mathrm{rd}(x)) - F'(\mathrm{rd}(x))}_{\text{Rundungsfehleranalyse}}. \tag{3.3}$$

Die erste Differenz analysiert wie die exakte Abbildung F auf Änderungen in der Eingabe (hier Rundungsfehler) reagiert. Diese Analyse ist unabhängig von der Problematik der Fließkommaarithmetik. Die zweite Differenz analysiert wie sich die beiden Abbildungen F, F' bei der selben Eingabe Unterscheiden, also die eigentliche Rundungsfehleranalyse.

Schlussendlich wird man Fehlernormen analysieren. Dann folgt mittels Dreiecksungleichung

$$\|F(x) - F'(\mathrm{rd}(x))\| \leq \|F(x) - F(\mathrm{rd}(x))\| + \|F(\mathrm{rd}(x)) - F'(\mathrm{rd}(x))\|.$$

3.3 Differentielle Konditionsanalyse

Die folgende Analyse braucht den Satz von Tayler aus der Analysis. Da dieser Satz in der Vorlesung sehr häufig zitiert wird wollen wir ihn in zwei Varianten anführen.

Satz 3.1 (Satz von Taylor für Funktionen in einer Variable)**.** Sei $f : (a, b) \to \mathbb{R}$ $(r + 1)$ mal stetig differenzierbar und seien $x, z \in (a, b)$ gegeben. Dann gibt es ein $\xi \in (a, b)$ zwischen x und z (und abhängig von z), so dass gilt:

$$f(z) = \sum_{k=0}^{r} \frac{1}{k!} \frac{d^k f(x)}{dx^k} (z - x)^k + \frac{1}{(r + 1)!} \frac{d^{r+1} f(\xi)}{dx^{r+1}} (z - x)^{r+1}. \tag{3.4}$$

Dabei heisst $t_n(x, z) = \sum_{k=0}^{r} \frac{1}{k!} \frac{d^k f(x)}{dx^k} (z-x)^k$ *Taylorpolynom* in z um den *Entwicklungspunkt* x und $R_r(x, z) = \frac{1}{(r+1)!} \frac{d^{r+1} f(\xi)}{dx^{r+1}} (z - x)^{r+1}$ *Lagranges Restglied*.

Beweis. Siehe [Rannacher, 2018a, Satz 5.8] □

Bemerkung 3.2. Oft setzt man $z = x + h$ und somit $z - x = h$ mit der Idee, dass $|h|$ "klein" ist. Damit lauten die Taylorformel und das Restglied alternativ

$$f(x + h) = \sum_{k=0}^{r} \frac{1}{k!} \frac{d^k f(x)}{dx^k} h^k + \frac{1}{(r + 1)!} \frac{d^{r+1} f(\xi)}{dx^{r+1}} h^{r+1}. \tag{3.5}$$

Der Satz von Taylor kann auf $r + 1$ mal stetig partiell differenzierbare Funktionen $f : D \to \mathbb{R}$, mit D einer offenen Teilmenge des $\mathbb{R}^m$, erweitert werden. Um komplexe Notation zu vermeiden zitieren wir hier nur eine spezialisierte Variante des Satzes für $r = 1$. Mehr werden wir im folgenden nicht benötigen.

Satz 3.3 (Spezialfall des Satz von Taylor für Funktionen in mehreren Variablen)**.** Sei $D \subset \mathbb{R}^m$ eine offene Menge und $f : D \to \mathbb{R}$ zweimal stetig differenzierbar. Weiter seien $x, \eta \in \mathbb{R}^m$ derart, dass $x + s\eta \in D$ für alle $s \in [0, 1]$. Dann gibt es zu so einem η ein $\theta \in (0, 1)$ so dass

$$\begin{aligned} f(x + \eta) = f(x) &+ \sum_{i=1}^{m} \frac{\partial f(x)}{\partial x_i} \eta_i \\ &+ \frac{1}{2} \sum_{i=1}^{m} \frac{\partial^2 f(x + \theta\eta)}{\partial x_i^2} \eta_i^2 + \sum_{i=1}^{m} \sum_{i<j\leq m} \frac{\partial^2 f(x + \theta\eta)}{\partial x_i \partial x_j} \eta_i \eta_j. \end{aligned} \tag{3.6}$$

Dabei ist in der zweiten Zeile das Restglied $R_1(x, \eta) = \frac{1}{2} \sum_{i=1}^{m} \frac{\partial^2 f(x+\theta\eta)}{\partial x_i^2} \eta_i^2 + \sum_{i=1}^{m} \sum_{i<j\leq m} \frac{\partial^2 f(x+\theta\eta)}{\partial x_i \partial x_j} \eta_i \eta_j$ in Lagranger Form angegeben.

Beweis. Folgerung aus [Rannacher, 2018b, Satz 3.6] □

Bemerkung 3.4. Oft benötigen wir Abschätzungen des Restglieds. Mit $h := \max_{1\leq i\leq n} |\eta_i|$ gilt dann

$$|R_1(x, \eta)| \leq \left(\frac{1}{2} \sum_{i=1}^{m} \left| \frac{\partial^2 f(x + \theta\eta)}{\partial x_i^2} \right| + \sum_{i=1}^{m} \sum_{i<j\leq m} \left| \frac{\partial^2 f(x + \theta\eta)}{\partial x_i \partial x_j} \right| \right) h^2.$$

Für $h \to 0$ konvergiert der Wert in der Klammer und der Betrag des Restgliedes "geht wie h^2" gegen Null. Man schreibt hierfür auch "$|R_1(x, \eta)| = O(h^2)$".

Zur allgemeinen Quantifizierung der Konvergenzgeschwindigkeit dienen die

Definition 3.5 (Landau Symbole)**.** a) Man schreibt

$$g(t) = O(h(t)) \quad (t \to 0)$$

falls es ein $t_0 > 0$ und $c_0 \geq 0$ gibt so dass für alle $0 < t \leq t_0$ die Abschätzung

$$|g(t)| \leq c_0 |h(t)|$$

gilt. Man sagt "$g(t)$ geht mindestens wie $h(t)$ gegen Null".

b) Weiter schreibt man

$$g(t) = o(h(t)) \quad (t \to 0)$$

wenn es ein $t_0 > 0$ und eine Funktion $c(t)$ mit $\lim_{t\to 0} c(t) = 0$ gibt, so dass für alle $0 < t \leq t_0$ die Abschätzung

$$|g(t)| \leq c(t)|h(t)|$$

gilt. Geht $h(t)$ ebenfalls gegen Null dann drückt dies aus: "$g(t)$ geht schneller als $h(t)$ gegen Null".

c) Schließlich schreiben wir

$$g(t) \doteq h(t)$$

und sagen „$g(t)$ ist *in erster Näherung gleich* $h(t)$", wenn gilt

$$g(t) - h(t) = o(t).$$

□

Aus dem Satz von Taylor folgt mittels der Beobachtung $g(t) = O(t^2) \Rightarrow g(t) = o(t)$

$$f(x + \eta) \doteq f(x) + \sum_{i=1}^{m} \frac{\partial f(x)}{\partial x_i} \eta_i$$

Nach diesem Ausflug in die Analysis wenden wir uns nun wieder der Konditionsanalyse zu. Zur Analyse der Auswirkung von Änderungen in der Eingabe der Funktion F betrachten wir diese als zweimal stetig differenzierbare Abbildung. Nach dem Satz von Taylor 3.3 gilt damit für jedem Komponente F_i:

$$F_i(x + \Delta x) = F_i(x) + \sum_{j=1}^{m} \frac{\partial F_i}{\partial x_j}(x)\Delta x_j + R_{1,i}(x, \Delta x), \quad i = 1, \ldots n, \qquad (3.7)$$

mit einem Restglied $R_{1,i}(x, \Delta x) = O(h^2)$ und $h = \max_{1\leq i\leq n} |\Delta x_i|$. Wir formen um und gehen zur $\doteq$-Notation über:

$$F_i(x + \Delta x) - F_i(x) \doteq \sum_{j=1}^{m} \frac{\partial F_i}{\partial x_j}(x)\Delta x_j, \qquad i = 1, \ldots n.$$

Für die relative Änderung erhalten wir dann

$$\frac{F_i(x + \Delta x) - F_i(x)}{F_i(x)} \doteq \sum_{j=1}^{m} \frac{\partial F_i}{\partial x_j}(x) \frac{\Delta x_j}{F_i(x)} = \sum_{j=1}^{m} \underbrace{\left(\frac{\partial F_i}{\partial x_j}(x) \frac{x_j}{F_i(x)}\right)}_{\text{Verstärkungsfaktor } k_{ij}(x)} \frac{\Delta x_j}{x_j}, \quad i = 1, \ldots n.$$

Der Verstärkungsfaktor k_{ij} beschreibt wie stark sich die relative Änderung in der j-ten Komponente der Eingabe auf die relative Änderung in der i-ten Komponente des Ergebnisses auswirkt.

Definition 3.6. Wir nennen die Auswertung $y = F(x)$ "schlecht konditioniert" im Punkt x falls $|k_{ij}(x)| \gg 1$, andernfalls heißt die Auswertung "gut konditioniert". Bei $|k_{ij}(x)| < 1$ spricht man von Fehlerdämpfung, bei $|k_{ij}(x)| > 1$ von Fehlerverstärkung. □

In der Aufspaltung (3.3) entspricht $\Delta x = \mathrm{rd}(x) - x$, also gerade dem absoluten Rundungsfehler in der Eingabe. Damit gilt $\frac{\Delta x_j}{x_j} = \frac{1}{2}\beta^{1-r}$.

Beispiel 3.7 (Konditionierung der Grundoperationen)**.** a) Wir untersuchen die Konditionierung der Addition, also $F(x_1, x_2) = x_1 + x_2$. Offensichtlich gilt $\frac{\partial F}{\partial x_1} = 1$, $\frac{\partial F}{\partial x_2} = 1$ und damit nach obiger Formel

$$\frac{F(x_1 + \Delta x_1, x_2 + \Delta x_2) - F(x_1, x_2)}{F(x_1, x_2)} = \left(1 \cdot \frac{x_1}{x_1 + x_2}\right)\frac{\Delta x_1}{x_1} + \left(1 \cdot \frac{x_2}{x_1 + x_2}\right)\frac{\Delta x_2}{x_2}.$$

Für $x_1 \to -x_2$ werden beide Verstärkungsfaktoren sehr groß. In diesem Fall ist die Addition schlecht konditioniert. Dies gilt analog für die Subtraktion falls $x_1 \to x_2$.

b) Für die Multiplikation $F(x_1, x_2) = x_1 \cdot x_2$ gilt $\frac{\partial F}{\partial x_1} = x_2$, $\frac{\partial F}{\partial x_2} = x_1$ und damit

$$\frac{F(x_1 + \Delta x_1, x_2 + \Delta x_2) - F(x_1, x_2)}{F(x_1, x_2)} = \left(x_2 \cdot \frac{x_1}{x_1 \cdot x_2}\right)\frac{\Delta x_1}{x_1} + \left(x_1 \cdot \frac{x_2}{x_1 \cdot x_2}\right)\frac{\Delta x_2}{x_2}.$$

Die Verstärkungsfaktoren sind somit beide 1 unabhängig von x_1, x_2. Die Multiplikation ist eine gut konditionierte Operation!

c) Schließlich betrachten wir noch die Funktion $F(x_1, x_2) = x_1^2 - x_2^2$. Offensichtlich gilt $\frac{\partial F}{\partial x_1} = 2x_1$, $\frac{\partial F}{\partial x_2} = -2x_2$ und damit nach obiger Formel

$$\frac{F(x_1 + \Delta x_1, x_2 + \Delta x_2) - F(x_1, x_2)}{F(x_1, x_2)} = \left(2x_1 \cdot \frac{x_1}{x_1^2 - x_2^2}\right)\frac{\Delta x_1}{x_1} + \left(2x_2 \cdot \frac{x_2}{x_2^2 - x_1^2}\right)\frac{\Delta x_2}{x_2}.$$

3.4 Rundungsfehleranalyse

In der eigentlichen Rundungsfehleranalyse betrachten wir entsprechend der Aufspaltung (3.3) die Differenz $F(x) - F'(x)$ für Maschinenzahlen $x \in \mathbb{F}$. Entspricht F genau einer Rechenoperation $\circledast$ so gilt wegen der exakten Rundung

$$\frac{(x * y) - (x \circledast y)}{x * y} = \epsilon \qquad |\epsilon| \leq \mathrm{eps} = \frac{1}{2}\beta^{1-r}$$

Der genaue Rundungsfehler ϵ hängt von den Argumenten x und y ab und ist damit für jede Operation verschieden. Umstellen der letzten Beziehung liefert

$$x \circledast y = (x * y)(1 + \epsilon) \qquad \text{für ein } |\epsilon(x, y)| \leq \mathrm{eps}$$

In der Rundungsfehleranalyse lassen sich ebenfalls Verstärkungsfaktoren ähnlich wie in der Konditionsanalyse herleiten. Dies illustrieren wir mit Beispielen.

Beispiel 3.8. Wir untersuchen die Abbildung $F(x_1, x_2) = x_1^2 - x_2^2$ und zwei verschieden Realisierungen

$$F_a(x_1, x_2) = (x_1 \odot x_1) \ominus (x_2 \odot x_2),\ F_b(x_1, x_2) = (x_1 \ominus x_2) \odot (x_1 \oplus x_2).$$

a) In diesem Fall erhalten wir für die ersten beiden Operationen

$$\begin{aligned} u = x_1 \odot x_1 = x_1^2(1+\epsilon_1) \qquad & |\epsilon_1| \leq \text{eps}, \\ v = x_2 \odot x_2 = x_2^2(1+\epsilon_2) \qquad & |\epsilon_2| \leq \text{eps} \end{aligned}$$

und dann

$$\begin{aligned} F_a(x_1,x_2) &= u \ominus v \\ &= (x_1^2(1+\epsilon_1) - x_2^2(1+\epsilon_2))(1+\epsilon_3) \\ &= (x_1^2 + \epsilon_1 x_1^2 - x_2^2 - \epsilon_2 x_2^2)(1+\epsilon_3) \\ &= x_1^2 - x_2^2 + (\epsilon_1+\epsilon_3)x_1^2 - (\epsilon_2+\epsilon_3)x_2^2 + \epsilon_1\epsilon_3 x_1^2 - \epsilon_2\epsilon_3 x_2^2. \end{aligned}$$

Damit erhalten wir in erster Näherung für den relativen Rundungsfehler

$$\frac{F_a(x_1,x_2) - F(x_1,x_2)}{F(x_1,x_2)} \doteq \frac{x_1^2}{x_1^2 - x_2^2}(\epsilon_1+\epsilon_3) + \frac{x_2^2}{x_2^2 - x_1^2}(\epsilon_2+\epsilon_3).$$

Ein Vergleich mit Beispiel 3.7 ergibt bis auf Konstanten die gleichen Verstärkungsfaktoren.

b) Für die zweite Variante ergibt sich

$$\begin{aligned} u = x_1 \ominus x_2 = (x_1 - x_2)(1+\epsilon_1) \qquad & |\epsilon_1| \leq \text{eps}, \\ v = x_1 \oplus x_2 = (x_1 + x_2)(1+\epsilon_2) \qquad & |\epsilon_2| \leq \text{eps} \end{aligned}$$

und dann

$$\begin{aligned} F_b(x_1,x_2) &= u \odot v \\ &= (u \cdot v)(1+\epsilon_3) \\ &= ((x_1 - x_2)(1+\epsilon_1) \cdot (x_1 + x_2)(1+\epsilon_2))(1+\epsilon_3) \\ &= (x_1 - x_2)(x_1 + x_2)(1+\epsilon_1)(1+\epsilon_2)(1+\epsilon_3) \\ &= (x_1^2 - x_2^2)(1 + \epsilon_1 + \epsilon_2 + \epsilon_3 + \ldots \epsilon_1\epsilon_2\epsilon_3). \end{aligned}$$

Damit erhalten wir in erster Näherung für den relativen Rundungsfehler

$$\frac{F_b(x_1,x_2) - F(x_1,x_2)}{F(x_1,x_2)} \doteq \epsilon_1 + \epsilon_2 + \epsilon_3.$$

Hier ist der Verstärkungsfaktor 1!. □

In der Gesamtschau sind Konditionsanalyse und Rundungsfehleranalyse gemeinsam zu betrachten. Dazu die folgende

Definition 3.9. Wir nennen einen numerischen Algorithmus *numerisch stabil*, wenn die Verstärkungsfaktoren aus der Rundungsfehleranalyse die aus der Konditionsanalyse nicht übersteigen.

Nach dieser Definition sind beide Realisierungen aus Beispiel 3.8 numerisch stabil.

3.5 Die quadratische Gleichung

Als ein weiteres, wichtiges Beispiel betrachten wir die Lösung der quadratischen Gleichung mit der p,q-Formel. Dieses Beispiel zeigt auch, dass die Minimierung von Rundungsfehlern für jedes Problem speziell betrachtet werden muss.

Die Gleichung

$$x^2 - px + q = 0$$

hat für $p^2/4 > q \neq 0$ die beiden reellen und verschiedenen Lösungen

$$x_{1/2} = f_\pm(p,q) = \frac{p}{2} \pm \sqrt{\frac{p^2}{4} - q}.$$

Die *Konditionsanalyse* der (beiden!) Abbildungen $f_\pm(p,q)$ liefert (Berechnung der partiellen Ableitungen):

$$\frac{f_\pm(p+\Delta p, q+\Delta q) - f(p,q)}{f(p,q)} \doteq \left(1 \pm \frac{p}{2\sqrt{\frac{p^2}{4}-q}}\right) \frac{p}{p \pm 2\sqrt{\frac{p^2}{4}-q}} \frac{\Delta p}{p} \mp \frac{q}{\sqrt{\frac{p^2}{4}-q}\left(p \pm 2\sqrt{\frac{p^2}{4}-q}\right)} \frac{\Delta q}{q}.$$

Daraus ersehen wir die folgenden Fälle

1) Für $\frac{p^2}{4} \to q$ sind beide Lösungen schlecht konditioniert, da beide Verstärkungsfaktoren unweigerlich groß werden.

2) Für $q \to 0$ und $|p|$ weg von der Null kommt es auf das Vorzeichen von p an. Für $p < 0$ ist die negative Lösung $f_-(p,q) = \frac{p}{2} - \sqrt{\frac{p^2}{4} - q}$ in jedem Fall gut konditioniert und für $p > 0$ ist die positive Lösung $f_+(p,q) = \frac{p}{2} + \sqrt{\frac{p^2}{4} - q}$ in jedem Fall gut konditioniert.

Für die *Rundungsfehler* ist die Situation ähnlich. Für den ersten Fall oben kann man Probleme nicht vermeiden, es tritt unweigerlich Auslösung ein. Im zweiten Fall ist es vermeidet man Auslöschung in dem man die jeweils andere Lösung über den "Satz von Vieta"

$$x_1 + x_2 = p, \qquad x_1 \cdot x_2 = q,$$

ermittelt. D.h. man berechnet zunächst

$$w = \sqrt{(p \odot p) \oslash 4 \ominus q}.$$

Im Fall $p < 0$ rechnet man dann

$$x_2 = p \oslash 2 \ominus w, \qquad x_1 = q \oslash x_2$$

und im Fall $p > 0$ entsprechend

$$x_1 = p \oslash 2 \oplus w, \qquad x_2 = q \oslash x_1.$$

3.6 Auslöschung

Aus Beispiel 3.7 a) und Beispiel 3.8 a) entnehmen wir, dass die Addition (bzw. Subtraktion) die gefährlichen Operationen sind (da $x - y = x + (-y)$ sind Addition und Subtraktion keine verschiedenen Operationen). Das Grundübel ist die sogenannte Auslöschung, die immer dann auftritt wenn annähernd gleiche Zahlen von einander subtrahiert werden, bzw. wenn betragsmäßig annähernd gleiche Zahlen mit verschiedenen Vorzeichen addiert werden. Dabei treten Probleme erst dann auf, wenn die beiden Operanden selbst schon fehlerbehaftet sind (was aber in der Regel bei längeren Rechnungen der Fall ist). Wir illustrieren das mit einem Beispiel.

Beispiel 3.10 (Zur Auslöschung)**.** a) Für zwei beliebige *Maschinenzahlen* $x_1, x_2 \in \mathbb{F}$ gilt (unter der Bedingung $x_1 - x_2 \neq 0$) wegen der exakten Rundung der Fließkommaarithmetik (3.2):

$$\left| \frac{(x_1 \ominus x_2) - (x_1 - x_2)}{x_1 - x_2} \right| \leq \text{eps}.$$

b) Auslöschung tritt erst auf, wenn die beiden Argumente *selbst schon mit Fehlern behaftet sind*. Woher diese Fehler kommen ist dabei unerheblich. So kann in Beispiel 3.8 a) der Verstärkungsfaktor sehr groß werden wenn $|x_1 - x_2|$ klein ist. Sind $x_1 = m_1\beta^e, x_2 = (m_1 - \beta^{-r})\beta^e$ Maschinenzahlen, ist ein Verstärkungsfaktor der Größenordnung

$$\frac{x_1^2}{x_1^2 - x_2^2} = \frac{x_1^2}{(x_1 - x_2)(x_1 + x_2)} \approx \frac{m_1^2\beta^{2e}}{\beta^{-r}\beta^e 2m_1\beta^e} \approx \beta^{r-1}$$

möglich!

c) Betrachte die Berechnung $F'(\text{rd}(x_1), \text{rd}(x_2)) = \text{rd}(x_1) \ominus \text{rd}(x_2)$ in $\mathbb{F}(10, 4, 1)$ an einem konkreten Beispiel. Als Eingabe seien $x_1 = 0.11258762 \cdot 10^2$ und $x_2 = 0.11244891 \cdot 10^2$ gegeben. Offensichtlich gilt $x_1, x_2 \neq \mathbb{F}(10, 4, 1)$ und wir betrachten den relativen Fehler *inklusive Rundung* der Eingaben. Als exaktes Ergebnis erhalten wir

$$x_1 - x_2 = 0.13871 \cdot 10^{-1}.$$

Die (kaufmännische) Rundung der Eingaben liefert

$$\text{rd}(x_1) = 0.1126 \cdot 10^2, \quad \text{rd}(x_2) = 0.1124 \cdot 10^2$$

und die anschließende Fließkommaoperation liefert

$$\text{rd}(x_1) \ominus \text{rd}(x_2) = 0.2 \cdot 10^{-1}$$

wobei im übrigen *kein* weiterer Rundungsfehler eingeführt wird! Trotzdem ist keine Stelle im Gesamtergebnis korrekt und es ergibt sich als relativer Gesamtfehler

$$\frac{0.2 \cdot 10^{-1} - 0.13871 \cdot 10^{-1}}{0.13871 \cdot 10^{-1}} \approx 0.44 \approx 883\frac{10^{-3}}{2} = 883\,\text{eps}.$$

□

Als Regel kann man sich merken: *Setze die potentiell gefährlichen Operationen $\oplus$, $\ominus$ möglichst früh ein.*