Übung 1: Maschinenzahlen

Für $b, t \in \mathbb{N}$, $e_{min}, e_{max} \in \mathbb{Z} \cup \{\pm\infty\}$ heißt

$$\mathbb{F}(b,t,e_{min},e_{max}) := \Big\{ \overset{\text{Vorzeichen}}{\sigma} \cdot (0,\overbrace{z_1 \dots z_t}^{\text{Mantisse}}) \cdot b^{\overset{\text{Exponent}}{e}} \;\Big|\; \sigma \in \{1,-1\} \dots$$

$$z_1 \neq 0,\ z_1,\dots,z_t \in \{0,\dots,b-1\},\ e \in \{e_{min},\dots,e_{max}\} \Big\} \cup \{0\}$$

die Menge der Fließkommazahlen zur Basis b und Mantissenlänge t.

Ist b gerade so ist der Rundungsoperator $rd : \mathbb{R} \to \mathbb{F}$ für

$$\mathbb{R} \ni x = \sigma \cdot (0,z_1 \dots z_t\, z_{t+1} \dots) \cdot b^e \quad \text{durch}$$

$$rd(x) = \begin{cases} \sigma \cdot (0,z_1 \dots z_t) \cdot b^e &, z_{t+1} \leq \frac{b}{2} - 1 \\ \sigma \cdot (0,z_1 \dots (z_t + 1)) \cdot b^e &, z_{t+1} \geq \frac{b}{2} \end{cases}$$

definiert.

Für die Beispiele betrachten wir im folgenden

$$\mathbb{F} = \mathbb{F}(10, 4, -\infty, \infty).$$

Damit ist z. B.

$$rd(\pi) = rd\,(+(0,31415\dots) \cdot 10^{1}) = 0,3142 \cdot 10^1$$

und

$$rd(\sqrt{2}) = rd\,(0,14142\dots \cdot 10^1) = 0,1414 \cdot 10^1$$

Es heißt

$$\text{macheps} := \varepsilon^* := \text{eps} := \frac{1}{2} b^{(-t+1)}$$

die Maschinengenauigkeit zu $\mathbb{F}(b, t, e_{min}, e_{max})$.

Satz: Es ist

$$\text{eps} = \inf\{\delta > 0 \mid rd(1+\delta) > 1\}$$

und für $x \in \mathbb{R}$ gilt (falls $e_{min} = -\infty$, $e_{max} = \infty$)

$$\left[\text{eps} \geq \left| \frac{rd(x) - x}{x} \right| \quad (=: |\varepsilon|) \right] \iff$$

$$\left[\exists \varepsilon \in \mathbb{R}, |\varepsilon| \leq \text{eps}, \text{ so dass } \quad \varepsilon = \frac{rd(x)-x}{x} \right] \iff$$

$$\left[\exists \varepsilon \in \mathbb{R}, |\varepsilon| \leq \text{eps}, \text{ mit } \quad rd(x) = x(1+\varepsilon) \right]$$

Beweis: VL $\boxtimes$

Die Maschinen- / Ersatz-Operationen $\tilde{+}, \tilde{-}, \tilde{\times}, \tilde{\div} : \mathbb{F} \times \mathbb{F} \to \mathbb{F}$ sind für $x, y \in \mathbb{F}$ durch

$$x \mathbin{\tilde{\circ}} y := rd(x \circ y) \qquad \circ \in \{+, -, \times, \div\}$$

$$\overset{\text{Satz}}{=} (x \circ y)(1+\varepsilon)$$

Rundungsfehler

gegeben. Weiter bezeichne

~~Es ergibt sich nun folgendes Problem:~~

$\operatorname{sqrt}(x) := \operatorname{rd}(\sqrt{x}) \overset{\text{Satz}}{=} \sqrt{x}\,(1+\underbrace{\varepsilon}_{\text{Rundungsfehler}})$,

das maschinelle Wurzelziehen.

Bei verketteten Ausdrücken können sich die Rundungsfehler hochschaukeln/akkumulieren.

Um diesen Effekt zu analysieren betrachten wir die

Fehleranalyse

am Beispiel der äquivalenten Ausdrücke

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

$$\boxed{b)\ \frac{x}{\sqrt{1+x}+1}}\ , \text{ für } |x| \approx 0 .$$

Dabei kann man zwei Methoden anwenden:

Methode 1: am Beispiel a)

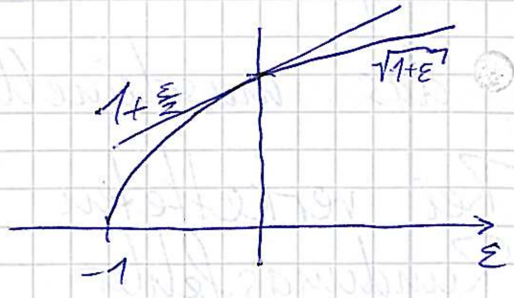
Ist $f: \mathbb{R} \to \mathbb{R}$ genügend oft differenzierbar, so heißt

$$f(x+\varepsilon) = f(x) + \varepsilon f'(x) + \frac{\varepsilon^2}{2} f''(x) + \ldots$$

die *Taylorentwicklung* von f. Ist nun $|\varepsilon|$ klein, so ist $|\varepsilon|^2$ sehr klein und daher nennt man

$$f(x+\varepsilon) \approx f(x) + \varepsilon f'(x)$$

Approximation erster Ordnung. Da $\frac{d}{dx}\sqrt{x} = \frac{1}{2\sqrt{x}}$ ist z. B.



$$\sqrt{1+\varepsilon} \approx \sqrt{1} + \varepsilon \frac{1}{2\sqrt{1}}$$

$$= 1 + \frac{\varepsilon}{2} \qquad (*).$$

Nun wird auf der Maschine der Ausdruck a) ausgewertet als

$$\mathrm{sqrt}(1 \tilde{+} x) \tilde{-} 1$$

$$= \mathrm{sqrt}\left((1+x)(1+\varepsilon_1)\right) \tilde{-} 1$$

$$= \left[\sqrt{(1+x)(1+\varepsilon_1)}\,(1+\varepsilon_2) - 1\right](1+\varepsilon_3)$$

(□) $f(\varepsilon_1, \varepsilon_2, \varepsilon_3)$ s. u.

$$= \left[\sqrt{1+x}\,\sqrt{1+\varepsilon_1}\,(1+\varepsilon_2) - 1\right](1+\varepsilon_3)$$

$$\overset{(*)}{\approx} \left[\sqrt{1+x}\left(1+\frac{\varepsilon_1}{2}\right)(1+\varepsilon_2) - 1\right](1+\varepsilon_3)$$

$$= \left[\sqrt{1+x}\left(1 + \frac{\varepsilon_1}{2} + \varepsilon_2 + \frac{\varepsilon_1\varepsilon_2}{2}\right) - 1\right](1+\varepsilon_3)$$

sehr klein

$$\approx \left[\sqrt{1+x}\left(1+\frac{\varepsilon_1}{2}+\varepsilon_2\right) - 1\right](1+\varepsilon_3)$$

$$= \sqrt{1+x}\left(1+\frac{\varepsilon_1}{2}+\varepsilon_2\right) - 1 + \sqrt{1+x}\left(\varepsilon_3 + \frac{\varepsilon_1\varepsilon_3}{2} + \varepsilon_2\varepsilon_3\right) - \varepsilon_3$$

sehr klein

$$\approx \sqrt{1+x} - 1 + \sqrt{1+x}\left(\frac{\varepsilon_1}{2} + \varepsilon_2 + \varepsilon_3\right) - \varepsilon_3$$

Damit ist der Fehler

$$\left| \mathrm{sqrt}(1 \tilde{+} x) \tilde{-} 1 - \left(\sqrt{1+x} - 1\right) \right|$$

$$\dot{\leq} \left|\sqrt{1+x}\right| \cdot \left| \frac{\varepsilon_1}{2} + \varepsilon_2 \right| + \left| \sqrt{1+x} - 1 \right| \cdot |\varepsilon_3|$$

$$\leq \sqrt{1+x} \cdot \frac{3}{2}\,\mathrm{eps} + \left|\sqrt{1+x} - 1\right| \cdot \mathrm{eps}$$

$$\underset{x\text{ klein}}{\overset{(*)}{\approx}} \left(1 + \frac{x}{2}\right) \frac{3}{2}\,\mathrm{eps} + \left| 1 + \frac{x}{2} - 1 \right| \cdot \mathrm{eps}$$

$$\leq \left(\frac{3}{2} + \frac{5}{4}|x| \right) \cdot \mathrm{eps}$$

<u>Methode 2:</u> am Beispiel (a)

Ist $f: \mathbb{R}^n \to \mathbb{R}$ genügend oft differenzierbar, so heißt

$$f(\varepsilon_1, \dots, \varepsilon_n) = f(0, \dots, 0) + Df\big|_0 \cdot \begin{bmatrix} \varepsilon_1 \\ \vdots \\ \varepsilon_n \end{bmatrix} + \frac{1}{2} \begin{bmatrix} \varepsilon_1 \\ \vdots \\ \varepsilon_n \end{bmatrix}^T D^2 f\big|_0 \begin{bmatrix} \varepsilon_1 \\ \vdots \\ \varepsilon_n \end{bmatrix} + \dots$$

die <u>mehrdimensionale Taylorentwicklung</u> von f. Analog heißt

$$f(\underbrace{\varepsilon_1, \dots, \varepsilon_n}_{=:\varepsilon}) \approx f(0, \dots, 0) + Df\big|_0 \begin{bmatrix} \varepsilon_1 \\ \vdots \\ \varepsilon_n \end{bmatrix}$$

die Approximation erster Ordnung.

~~Nun wird auf der Maschine der Ausdruck b) ausgewertet als,~~

Mit $f(\varepsilon_1, \varepsilon_2, \varepsilon_3)$ wie oben $^{(II)}$ ist nun $f(0,0,0)$

der exakte Ausdruck a) und damit ist der Fehler

$$|f(\varepsilon_1, \varepsilon_2, \varepsilon_3) - f(0,0,0)|$$

$$\approx \left| Df|_0 \begin{bmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \varepsilon_3 \end{bmatrix} \right| \qquad (\triangle)$$

$$\leq \|Df|_0\|_1 \cdot \left\| \begin{bmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \varepsilon_3 \end{bmatrix} \right\|_\infty \leq \|Df|_0\|_1 \cdot eps \, .$$

Mit

$$Df|_{(\varepsilon_1,\varepsilon_2,\varepsilon_3)} = \Big[\sqrt{1+x}\, \frac{1}{2\sqrt{1+\varepsilon_1}} (1+\varepsilon_2)(1+\varepsilon_3) \, , \, \sqrt{(1+x)(1+\varepsilon_1)}\,(1+\varepsilon_3),$$

$$\sqrt{(1+x)(1+\varepsilon_1)}\,(1+\varepsilon_2) - 1 \Big]$$

$$\Rightarrow Df|_{(0,0,0)} = \left[\frac{1}{2}\sqrt{1+x} \, , \, \sqrt{1+x} \, , \, \sqrt{1+x} - 1 \right]$$

$$\Rightarrow \|Df|_0\|_1 = \frac{3}{2}\sqrt{1+x} + |\sqrt{1+x} - 1|$$

Insgesamt also

$$|\operatorname{sqrt}(1 \tilde{+} x) \tilde{-} 1 - (\sqrt{1+x} - 1)|$$

$$\overset{(\square)}{=} |f(\varepsilon_1, \varepsilon_2, \varepsilon_3) - f(0,0,0)|$$

$$\overset{(\triangle)}{\leq} \|Df|_0\|_1 \cdot eps = \sqrt{1+x} \cdot \frac{3}{2} eps + |\sqrt{1+x} - 1| \cdot eps$$

$$\underset{x \text{ klein}}{\overset{(*)}{\approx}} \left(1 + \frac{x}{2}\right) \frac{3}{2} eps + \left|1 + \frac{x}{2} - 1\right| eps \leq \left(\frac{3}{2} + \frac{5}{4}|x| \right) \cdot eps$$

das gleiche Ergebnis wie bei Methode 1.

Methode 2: am Beispiel b)

Auf der Maschine ist Ausdruck b)

$$x \mathbin{\tilde{\div}} (\mathrm{sqrt}(1 \mathbin{\tilde{+}} x) \mathbin{\tilde{+}} 1)$$

$$= \frac{x\,(1+\varepsilon_1)}{\left[\sqrt{(1+x)(1+\varepsilon_2)}\,(1+\varepsilon_3)+1\right](1+\varepsilon_4)} =: f(\varepsilon_1, \varepsilon_2, \varepsilon_3, \varepsilon_4)$$

$$\Rightarrow Df|_0 = \left[\frac{x}{\sqrt{1+x}+1},\ -\frac{x}{2}\frac{\sqrt{1+x}}{(\sqrt{1+x}+1)^2},\ -x\frac{\sqrt{1+x}}{(\sqrt{1+x}+1)^2},\ -\frac{x}{\sqrt{1+x}+1}\right]$$

$$\Rightarrow \|Df|_0\|_1 = 2\frac{|x|}{\sqrt{1+x}+1} + \frac{3}{2}\frac{|x|\sqrt{1+x}}{(\sqrt{1+x}+1)^2}$$

$$\leq 2|x| + \frac{3}{2}|x|\underbrace{\sqrt{1+x}}_{\leq 2 \text{ für } |x|\leq 1}$$

$$\leq 5|x|$$

$$\Rightarrow \left| x \mathbin{\tilde{\div}} (\mathrm{sqrt}(1 \mathbin{\tilde{+}} x) \mathbin{\tilde{+}} 1) - \left(\frac{x}{\sqrt{1+x}+1}\right)\right|$$

$$\dot{\leq} \|Df|_0\|_1 \cdot eps \leq 5|x| \cdot eps$$