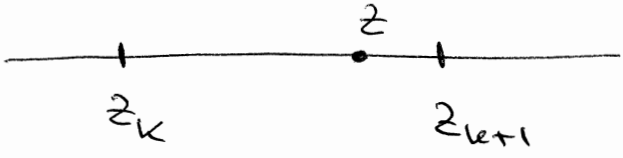
Betrachten wir Gleitpunktzahlen (normalisiert) der Form

$$z_{\mathbb{F}} = m \cdot b^e \quad \text{mit} \quad m > 0 \text{ und } m \geq 0{,}1$$

Dann gilt $\left| \frac{\square z - z}{z} \right| \leq eps = \begin{cases} b^{1-p} & \text{f. Absch.} \\ \frac{1}{2} b^{1-p} & \text{f. Runden} \end{cases}$

für alle $z \in \mathbb{R} \cap \mathbb{F}_N$ (d.h. für alle reelle Zahlen z, die im Bereich der normal. Maschinenzahlen liegen.)



Seien z_k und z_{k+1} Maschinenzahlen mit

$z_k = m_k b^e$ und $z_{k+1} = m_{k+1} b^e$, dann

gilt $|\square z - z| \leq |z_{k+1} - z_k| = b^{-p} \cdot b^e$

$$\Rightarrow \quad \left| \frac{\square z - z}{z} \right| \leq \frac{b^{-p} \cdot b^e}{m_k \cdot b^e} \leq \frac{b^{-p}}{0{,}1} = \frac{b^{-p}}{b^{-1}} =$$

$$= \underbrace{b^{1-p} \text{ für das Absch.}}$$

und $\left| \frac{\square - z}{z} \right| \leq \underbrace{\frac{1}{2} b^{1-p} \text{ für das Runden}}$

Der wesentliche Unterschied zwischen den Fest- und Gleitkommazahlen besteht darin, daß für die Festkommazahlen der absolute Abstand zwischen ihnen konstant ist, während für die Gleitkommazahlen der relative Abstand konstant ist (im normal. Bereich.)

Schätzung des relativen Gesamtrechenfehlers mit Hilfe der "(1+ε)"-Technik.

$\tilde{y} = \tilde{\ln}(1 \oplus x)$, x Maschinenzahl

$$1 \oplus x = (1+x)(1+\varepsilon_1), \quad |\varepsilon_1| \leq eps$$

$$\tilde{\ln}(1 \oplus x) = \ln(1 \oplus x)(1+\varepsilon_2), \quad |\varepsilon_2| \leq eps$$

Wir streben die Form an:

$$\tilde{\ln}(1 \oplus x) = \ln(1+x)(1+\delta);$$

wobei δ dann der Gesamtrechenfehler ist.

Rechnung:

$$\tilde{\ln}(1 \oplus x) = \underbrace{\ln\left((1+x)(1+\varepsilon_1)\right)}_{A}(1+\varepsilon_2).$$

Wir sehen uns zunächst den Term A an:

$$A = \ln\left((1+x)(1+\varepsilon_1)\right) = \ln\left((1+x) + \varepsilon_1(1+x)\right)$$

Wir haben also $\ln(a+b)$ mit b viel kleiner als $a \Rightarrow$ Linearisierung:

$$\ln(a+b) = \ln(a) + \left(\ln(a+b)\right)'\Big|_{b=0} \cdot b + O(b^2)$$

$$\approx \ln(a) + \frac{1}{a} \cdot b$$

d.h. $a = 1+x$, $b = \varepsilon_1(1+x) \Rightarrow$

$$\ln\left((1+x) + \varepsilon_1(1+x)\right) \approx \ln(1+x) + \varepsilon_1(1+x)\frac{1}{(1+x)}$$

$$= \ln(1+x) + \varepsilon_1 ;$$

Damit haben wir

$$\tilde{y} = \left(\ln(1+x) + \varepsilon_1\right)(1+\varepsilon_2) =$$

$$= \ln(1+x) + \varepsilon_2 \ln(1+x) + \varepsilon_1 + \underbrace{\varepsilon_1\varepsilon_2}_{\approx 0}$$

$$\doteq \ln(1+x) + \varepsilon_1 + \varepsilon_2 \ln(1+x) =$$

$$= \ln(1+x)\left(1 + \underbrace{\frac{\varepsilon_1}{\ln(1+x)}}_{\text{sehr groß für } x \approx 0} + \varepsilon_2\right)$$

Der Gesamtrechenfehler

$$\delta = \frac{\varepsilon_1}{\ln(1+x)} + \varepsilon_2 = \underbrace{\frac{1}{\ln(1+x)}}_{\text{großer Verstärkungsfaktor für den Additionsfehler.}} \cdot \varepsilon_1 + 1 \cdot \varepsilon_2 .$$

Dieser Fehler wird verglichen mit dem rel. Datenfehler

$$\tilde{\tilde{y}} = y(\tilde{x}) = y(x(1+\varepsilon)) \text{, dabei ist } \varepsilon \text{ der relative Datenfehler.}$$

$\Rightarrow$ Kondition $\tilde{\tilde{y}} = y(1+\tilde{\delta})$ mit

$$\tilde{\delta} = \underset{y \leftarrow x}{K} \cdot \varepsilon \approx 1 \cdot \varepsilon ,$$

vgl. Rechnungen auf meinen Seiten zur Vo. am 24.10.2005.

Damit sind die Verstärkungsfaktoren bei der RF viel größer als bei der Kondition und der Algorithmus instabil.

Auslöschung! Wegen $\ln(1+x) \approx x$ für kleine x bedeutet die Berechnung, daß wir die Operationen

$$(1 \oplus x) \ominus 1 \text{ ausführen!}$$