

# Einführung in die Numerische Rechenmethode

## 1. Redunzianzrechnung

Bei unterschiedlichen "Redunaufgaben" treten Fehler auf, und zwar

- Datenfehler aufgrund ungenauer Eingabedaten
- Darstellungsfehler von Zahlen
- Fehler durch ungenaue Reduzierungen, z.B. wird man bei der Aufgabe

$$1:3 = 0,33333\dots$$

eigentlich nie fertig, d.h. man gibt irgendwann loschoppf auf und macht einen Fehler.

### 1.1 Zahlendarstellungen

Aus der Analysis ist bekannt, dass man jede Zahl  $x \in \mathbb{R}, x \neq 0$  bei einer gegebenen Basis  $b \in \mathbb{N}, b \geq 2$  in der Form

$$x = \sum_{k=-e+1}^{\infty} a_{k+e} b^{-k} = \sum_{k=-e+1}^{\infty} a_k b^{-k} b^e, \quad (1.1)$$

$$a_1, a_2, \dots \in \{0, 1, \dots, b-1\}, e \in \mathbb{Z}, b \in \{+, -\}$$

darstellen kann, wobei  $a_1 \neq 0$  ist  
(fordert man, dass eine unendliche Teilmenge  $N_1 \subset \mathbb{N}$  gibt mit  $a_k \neq b-1$  für  $k \in N_1$ , dann ist die Darstellung (1.1) eindeutig).

### (1.1) heißt Gleitpunkt-Darstellung

Als Basis  $b$  wird oft  $b=10$  (das wird normalerweise in der Schule benutzt) oder  $b=2$ . Man spricht dann vom Decimal- bzw. Dualsystem.

### 1.2 Allgemeine Gleitpunkt-Zahlensysteme

Da man auf Rechnern nicht beliebig viele Stellen zur Darstellung von Zahlen in der Form (1.1) zur Verfügung hat, z.B. für die Zahlen

$$\frac{1}{3} = \left( \sum_{k=1}^{\infty} 3 \cdot 10^{-k} \right) 10^0$$

im Dezimalsystem, oder

$$\frac{1}{3} = \left( \sum_{k=1}^{\infty} c_k 2^{-k} \right) 2^0 \quad \text{mit } c_{k-1}=0, c_k=0, k=1, 2, \dots$$

arbeitet man mit Gleitpunktzahlensystemen wie folgt.

#### Definition 1.1

Zu gegebener Basis  $b \geq 2$  und Bruchintervall  $t \in \mathbb{N}$  sowie für Exponentenintervalle  $e_{\min} < e_{\max}$  ist die Menge  $F = F(b, t, e_{\min}, e_{\max}) \subset \mathbb{R}$  durch

$$F = \left\{ G \left( \sum_{k=1}^t a_k b^{-k} \right) b^e \mid a_1, \dots, a_t \in \{0, 1, \dots, b-1\}, a_1 \neq 0, e \in \mathbb{Z}, \right. \quad (1.2)$$

$$\left. e_{\min} \leq e \leq e_{\max}, G \in \{+, -\} \right\} \cup \{0\}$$

ehält und wird System von Normalisierten Gleitpunktzahlen genannt.

Läßt man noch die Kombination  $e=e_{\min}, a_1=0$  zu dann erhält man  $\hat{F} \supset F$  das System von denormalisierten Gleitpunktzahlen.

Statt der Angabe von Exponentengröße  $e_{\min}, e_{\max} \in \mathbb{Z}$  wird bei einem Gleitpunktzahlensystem auch mit  $\epsilon$  die Stellenzahl des Exponenten e angegeben, so dass man statt

$$F = F(2, 24, -127, 127) \quad (1.3)$$

und

$$F = F(2, 24, 7)$$

Schreiben kann, da man mit einer 7-stelligen Dualzahl alle Exponenten von  $0_{\text{bin}} \pm 127$  darstellen kann.

Statt F wird aus M als Symbol für Gleitpunktzahlensysteme verwendet (M für binäre Zahlen), also z.B.

$$M = M(2, 24, 7). \quad (1.4)$$

Die Darstellung (1.3) mit der Angabe der minimalen und maximalen Exponenten ist allerdings oft präziser, da in vielen praktischen Gleitpunktzahlensystemen tatsächlich  $|e_{\min}| \neq e_{\max}$

ist, was bei der Darstellung (1.4) nicht zu erkennen ist.

(4)

### 1.3 Struktur und Eigenschaften des normalisierten Gleitpunkt-Zahlsystems F

Es ist offensichtlich, dass die Elemente von F symmetrisch um den Nullpunkt liegen, weshalb hier nur die positiven Elemente betrachtet werden sollen.

Konkret betrachten wir  $F = F(b, t, e_{\min}, e_{\max})$  und finden mit

$$x_{\min} = (1 \cdot b^{-1} + 0 \cdot b^{-2} + \dots + 0 \cdot b^{-t}) b^{e_{\min}} = b^{-1+e_{\min}} \quad (1.5)$$

die kleinste positive normierte Gleitpunkt-Zahl.  
Andererseits ergibt sich mit

$$\begin{aligned} x_{\max} &= ((b-1)b^{-1} + (b-1)b^{-2} + \dots + (b-1)b^{-t}) b^{e_{\max}} \\ &= (1 - b^{-1} + b^{-1} - b^{-2} + \dots - b^{-t}) b^{e_{\max}} \\ &= (1 - b^{-t}) b^{e_{\max}} \end{aligned} \quad (1.6)$$

die größte positive normierte Gleitpunkt-Zahl.  
Für die Mantissen von Zahlen aus F ergibt sich aus (1.5) und (1.6)

$$b^{-1} \leq a \leq 1 - b^{-t} \quad (1.7)$$

In  $\hat{F}$ , also der Menge der denormalisierten Gleitpunkt-Zahlen sind kleinere Zahlen als  $x_{\min}$  darstellbar, und zwar mit

$$\hat{x}_{\min} = (0 \cdot b^{-1} + 0 \cdot b^{-2} + \dots + 1 \cdot b^{-t}) b^{e_{\min}} = b^{-t+e_{\min}} \quad (1.8)$$

die kleinste positive denormalisierte Gleitpunkt-Zahl.

Mit der Festlegung einer Mantisse-Länge  $t$  ist die ~~Zahl~~ der möglichen Mantissen festgelegt, so dass in jedem Intervall  $[b^{e-1}, b^e]$  gleichviel Gleitpunktzahlen liegen, die außerdem aquidistant verteilt sind, und zwar mit dem Abstand

$$\Delta = b^{-t} \cdot b^e = b^{e-t}$$

Der kleinste Abstand einer beliebigenellen Zahl  $x \in [b^{e-1}, b^e]$  zum nächstgelegenen Element  $z$  aus  $F$  ist damit durch  $\frac{1}{2}\Delta$  begrenzt, d.h.

$$|z - x| \leq \frac{1}{2} b^{e-t}. \quad (1.9)$$

Die Gleichheit wird erreicht, wenn  $x$  genau zwischen zwei benachbarten Zahlen aus  $F$  liegt. Wegen  $b^{e-1} \leq x$ , folgt aus (1.9)

$$\frac{|z - x|}{|x|} \leq \frac{\frac{1}{2} b^{e-t}}{b^{e-1}} = \frac{1}{2} b^{-t+1} =: \text{eps} \quad (1.10)$$

mit  $\text{eps} = \frac{1}{2} b^{-t+1}$  der (maximale relative) Abstand der Zahlen  $\{x \in \mathbb{R} \mid x_{\min} \leq |x| \leq x_{\max}\}$  zum jeweils nächstgelegenen Element aus  $F$ .

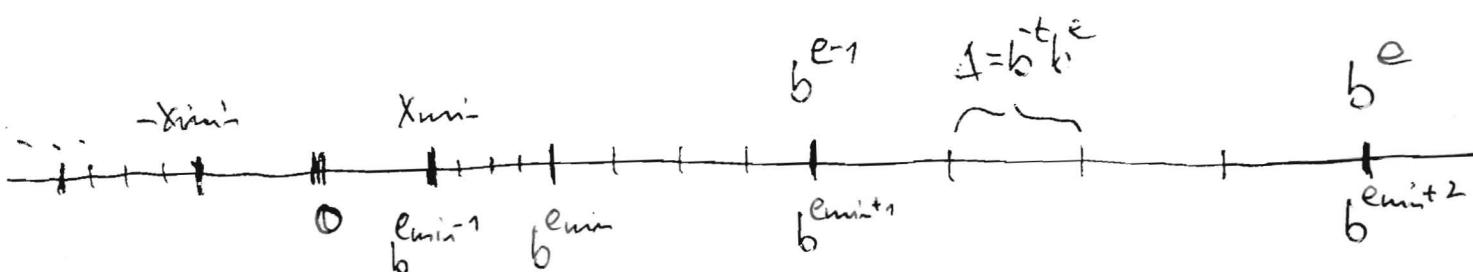


Bild 1 Verteilung normalisierter Gleitpunktzahlen  
(z.B. für  $b=2, t=1$ , folgt  $\Delta=2^{-1}$ )

Mit der Vemutris von  $\text{eps} = \frac{1}{2} 5^{t+1}$  lässt sich nun über die Bedingung

$$0,5 \cdot 10^{-4} \leq \text{eps} \leq 5 \cdot 10^{-4} \quad (1.11)$$

eine Zahl  $n \in \mathbb{N}$  bestimmen, und man spricht dann beim Gleitpunkt-Zahlensystem  $F$  von einer  $n$ -stellige Dezimalstellenarithmetik.

Als Beispiele von in der Praxis benutzten Gleitpunkt-Zahlensystemen sind hier IEEE-Standardsysteme

$$\begin{array}{ll} \hat{F}(2,2^4,-125,128) & (\text{einfach}), "real*4") \\ \hat{F}(2,53,-1021,1024) & (\text{doppelt}, "real*8") \end{array}$$

sowie die IBM-Systeme

$$\begin{array}{ll} F(16,6,-64,63) & (\text{einfach}) \\ F(16,14,-64,63) & (\text{doppelt}) \end{array}$$

genannt.

## 1.4 Reduktion mit Gleitpunkt-Zahlen

Einfache Reduktionen zeigen, dass Gleitpunkt-Zahlensysteme hinsichtlich der Addition/Subtraktion bzw. Multiplikation/Division nicht abgeschlossen sind, d.h. die Addition oder Multiplikation von Zahlen  $x, y \in F$  ergibt i. Allg. keine Zahl aus  $F$ .

Beispiel:  $F(10,4,-63,64)$ ,  $x = 0,1502 \cdot 10^2$ ,  $y = 0,1 \cdot 10^1$

$$x+y = 15,02 + 0,0001 = 15,02001 \approx 0,1502001 \cdot 10^2,$$

wir die Stellenzahl  $t=4$  nicht aus, um  $x+y$  in  $F$  zu erhalten.

Um in einem Gleitpunkt-Zahlsystem reden zu können braucht man letztendlich eine Abbildung aus  $\mathbb{R}$  in  $F$

### Definition 1.2

Zu einem gegebenen Gleitpunkt-Zahlsystem

$F = F(b, t, l_{\min}, l_{\max})$  mit gerader Basis  $b$  ist die Funktion  $rd: \{x \in \mathbb{R} \mid l_{\min} \leq |x| \leq l_{\max}\} \rightarrow \mathbb{R}$  durch

$$rd(x) = \begin{cases} G\left(\sum_{k=1}^t a_k b^{-k}\right) b^e & \text{falls } a_{k+1} \leq \frac{b}{2} - 1 \\ G\left(\sum_{k=1}^t a_k b^{-k} + b^{-t}\right) b^e & \text{falls } a_{k+1} \geq \frac{b}{2} \end{cases}$$

für  $x = G\left(\sum_{k=1}^{\infty} a_k b^{-k}\right) b^e$  erlaubt.  $rd(x)$  heißt auf  $t$  Stellen gerundeter Wert von  $x$

Dann kann nun folgende Eigenschaften für das Runden zeigen:

### Theorem 1.3

Zu einem gegebenen Gleitpunkt-Zahlsystem  $F = F(b, t, l_{\min}, l_{\max})$  gilt für jede reelle Zahl  $x$  mit  $|x| \in [l_{\min}, l_{\max}]$  die Eigenschaft  $rd(x) \in F$  und die Minimalabweichung

$$|rd(x) - x| = \min_{z \in F} |z - x|$$

### Beweis

Es gilt offensichtlich

$$\sum_{k=1}^t a_k b^{-k} \leq \sum_{k=1}^{\infty} a_k b^{-k} \leq \sum_{k=1}^t a_k b^{-k} + \underbrace{\sum_{k=t+1}^{\infty} (b-1)b^{-k}}_{= b^{-t} \text{ da Teleskop-Summe}}$$

Nach Multiplikation mit  $b^e$  erhalten wir

$$\underbrace{\left(\sum_{k=1}^t a_k b^{-k}\right) b^e}_{\geq b^{-1}} \leq \left(\sum_{k=1}^{\infty} a_k b^{-k}\right) b^e = |x| \leq \underbrace{\left(\sum_{k=1}^t a_k b^{-k} + b^{-t}\right) b^e}_{\leq 1}$$

d.h. die Schranken von  $|x|$  liegen im Intervall  $[b^{e-1}, b^e]$  und damit sind die beiden für  $\text{rd}(x)$  in Frage kommenden Werte

$$\mathfrak{G}\left(\sum_{k=1}^t a_k b^{-k}\right) b^e \quad \text{und} \quad \mathfrak{G}\left(\sum_{k=1}^t a_k b^{-1} + b^{-t}\right) b^e$$

die Nachbar von  $x$  aus dem Gleitpunkt-Zahlensystem  $F_1$ , also ist  $\text{rd}(x) \in F$ .

Es wird nun die Abschätzung

$$|\text{rd}(x) - x| \leq \frac{1}{2} b^{-t+e} \quad (1.12)$$

gezeigt.

Für  $a_{t+1} \leq \frac{b}{2} - 1$  (Abrunden) erhalten wir

$$\begin{aligned} |\text{rd}(x) - x| &= \left(\sum_{k=t+1}^{\infty} a_k b^{-k}\right) b^e = (a_{t+1} b^{-(t+1)} + \sum_{k=t+2}^{\infty} a_k b^{-k}) b^e \\ &\leq \left[\left(\frac{b}{2} - 1\right) b^{-(t+1)} + \sum_{k=t+2}^{\infty} (b-1) b^{-k}\right] b^e \\ &= b^{-(t+1)} \underbrace{\left[\left(\frac{b}{2} - 1\right) b^{-(t+1)} + b^{-(t+1)}\right]}_{\text{da Telephon-Folge}} b^e \\ &= \left[\left(\frac{b}{2} - 1\right) b^{-(t+1)} + b^{-(t+1)}\right] b^e = \frac{1}{2} b^{-t+e} \end{aligned}$$

Beim Abrunden, d.h.  $a_{t+1} \geq \frac{b}{2}$ , ergibt sich

$$\begin{aligned} |\text{rd}(x) - x| &= \left(b^{-t} - \sum_{k=t+1}^{\infty} a_k b^{-k}\right) b^e = \left(b^{-t} - \underbrace{a_{t+1} b^{-(t+1)}}_{\geq \frac{1}{2} b^{-t}} - \underbrace{\sum_{k=t+2}^{\infty} a_k b^{-k}}_{\geq 0}\right) b^e \\ &\leq \frac{1}{2} b^{-t+e} \end{aligned}$$

Da wir früher gelernt haben, dass  $\frac{1}{2}b^{-t+e}$  (9)  
 die Hälfte des Abstandes zweier Nachbars in  $F$   
 darstellt, folgt aus (1.12)

$$|rd(x) - x| = \min_{z \in F} |z - x| \quad (1.13)$$

Als Folge aus (1.12) erhält man wegen  $|x| \geq b^{e_t}$

$$\frac{|rd(x) - x|}{|x|} \leq \frac{1}{2}b^{-t+1} = \text{eps} \quad (\text{Maschinenepsilon}) \quad (1.14)$$

als Abschätzung für den relativen Rundungsfehler.

### Definition 1.4

$\text{eps} = \frac{1}{2}b^{-t+1}$  Als Schranke für den  
 relativen Rundungsfehler heißt Maschinenrau-  
 keit oder Rundeff mit  $u$  ( $u$  werden auch  
 die Bezeichnungen  $mdeps$  oder  $\epsilon^*$  verwendet, es gilt dann  
 $\text{eps} = \inf \{ \delta > 0 \mid rd(1+\delta) > 1 \}$ ).

Neben der Möglichkeit des Rundens mit  $u$  gibt  
 es auch als Alternative das Kürzen (englisch:  
 truncation):

### Definition 1.5

Zu  $F = F(b, t, c_{\min}, c_{\max})$  ist die Funktion  
 $tc: \{x \in \mathbb{R} \mid x_{\min} \leq |x| \leq x_{\max}\} \rightarrow F$  durch

$tc(x) = C \left( \sum_{k=1}^t a_k b^{-k} \right) b^e$  für  $x = C \left( \sum_{k=1}^{\infty} a_k b^{-k} \right) b^e$   
 erklärt.

### Bemerkung 1.6

Kürzen ist i. Allg. ungenauer als Runden und es  
 gilt  $\frac{|tc(x) - x|}{|x|} \leq 2 \text{eps}$  für  $x \in \mathbb{R}$  mit  $x_{\min} \leq |x| \leq x_{\max}$ .