stellt bis zu 7 gültige Ziffern im Bereich $\mathcal{F} = [\min \mathcal{F}, \max \mathcal{F}]$ zur Verfügung; der Fehler für jede in $\mathcal{F}$ liegende reelle Zahl x durch ihre Approximation trunc(x) bzw. durch ihre Rundung round(x) ist absolut

$$\begin{aligned} |x - \mathrm{trunc}(x)| &\leq 10^{-3} \\ |x - \mathrm{round}(x)| &\leq \frac{1}{2} \cdot 10^{-3}, \end{aligned}$$

also z.B. $|\frac{2}{3} - 0.666|$ bzw. $|\frac{2}{3} - 0.667|$. Dagegen kann der relative Fehler

$$\left| \frac{x - \mathrm{trunc}(x)}{x} \right| \quad \text{bzw.} \quad \left| \frac{x - \mathrm{trunc}(x)}{\mathrm{trunc}(x)} \right|$$

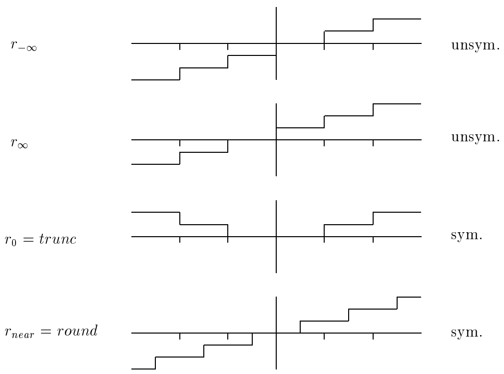
beliebig groß werden, es kann sogar passieren, daß keine Ziffer der Näherung stimmt! Für viele naturwissenschaftlichen Rechnungen reichen etwa obige 7 Stellen (7 *richtige* führende Stellen) aus, aber der abgedeckte Zahlenbereich erstreckt sich über viele Größenordnungen wie μV,...,kV, also 10^{-6}V,...,10^3V, obwohl viele analoge Voltmeter nur $\pm 1\%$ Ablesegenauigkeit haben.

Definition 1.3 *Eine Abbildung* $\circ : \mathbb{R} \to \mathcal{F}_n^\nu$ *mit den Eigenschaften*

$$\begin{aligned} \textit{Projektion :} \quad & \circ|_{\mathcal{F}_n^\nu} = id \\ \textit{Monotonie :} \quad & x \leq y \Leftarrow \circ x \leq \circ y \end{aligned}$$

heißt Rundung .

IEEE stellt 4 Rundungen zur Verfügung:



Sei $\tilde{x} \in \mathcal{F}_n^\nu$ eine Näherung von x, z.B. $\tilde{x} = \circ x$ mit $\circ \in \{r_{-\infty}, r_\infty, r_0, r_{near}\}$, so bezeichnet man

$$\Delta\tilde{x} := |\tilde{x} - x|$$

als *absoluten Fehler* des Näherungswertes $\tilde{x}$,

$$\bar{\varepsilon}_{\tilde{x}} := \left| \frac{\tilde{x} - x}{x} \right|$$

für $x \neq 0$ als *relativen Fehler* von $\tilde{x}$. Da x häufig nicht bekannt ist, man also nur weiß, daß $\tilde{x}$ maximal $\Delta\tilde{x}$ von x entfernt ist, benutzt man

$$\varepsilon_{\tilde{x}} := \left| \frac{\tilde{x} - x}{\tilde{x}} \right|$$

für $\tilde{x} \neq 0$ als relativen Fehler. Weiß man (z.B. bei $\tilde{x} = \circ x$ gerundet), daß $\tilde{x}$ genau zwischen zwei Rasterpunkten von $\mathcal{F}_3^4$ liegt, so kann man grob abschätzen:

$$\begin{aligned} \Delta\tilde{x} &\leq 10^{-3} \\ \bar{\varepsilon}_{\tilde{x}} &\leq \frac{10^{-3}}{x} \quad \forall x \neq 0 \\ \varepsilon_{\tilde{x}} &\leq \frac{10^{-3}}{\tilde{x}} \quad \forall \tilde{x} \neq 0 \end{aligned}$$

Ist die Art der Rundung nicht bekannt, ergibt sich also:

$$\begin{array}{lll} \tilde{x} = 0 & \Rightarrow x \in (-0.001, 0.001) \Rightarrow & \left|\frac{\Delta\tilde{x}}{x}\right| \leq \infty \\ \tilde{x} = 0.001 & \Rightarrow x \in (0.000, 0.002) \Rightarrow & \left|\frac{\Delta\tilde{x}}{x}\right| \leq \infty \\ \tilde{x} = 0.002 & \Rightarrow x \in (0.001, 0.003) \Rightarrow & \left|\frac{\Delta\tilde{x}}{x}\right| \leq 1 \\ \tilde{x} = 0.003 & \Rightarrow x \in (0.002, 0.004) \Rightarrow & \left|\frac{\Delta\tilde{x}}{x}\right| \leq 0.5 \\ \vdots & & \\ \tilde{x} = 9999.998 & \Rightarrow x \in (9999.997, 9999.999) \Rightarrow & \left|\frac{\Delta\tilde{x}}{x}\right| \leq 10^{-7} \\ \tilde{x} = 9999.999 & \Rightarrow ??? & \end{array}$$

Kennt man die Art der Rundung, so läßt sich genauer feststellen

1. Rundung zur Null:

$$\tilde{x} = 0 \quad \Rightarrow x \in (-0.001, 0.001) \Rightarrow \quad \left|\frac{\Delta\tilde{x}}{x}\right| \leq \infty$$

$$\begin{array}{lll} \tilde{x} = 0.001 & \Rightarrow x \in [0.001, 0.002) \Rightarrow & \left|\frac{\Delta\tilde{x}}{x}\right| \leq 1 \\ \tilde{x} = 0.002 & \Rightarrow x \in [0.002, 0.003) \Rightarrow & \left|\frac{\Delta\tilde{x}}{x}\right| \leq 0.5 \\ \vdots & & \end{array}$$

Ein analoges Ergebnis gilt für $r_{+\infty}$ und $r_{-\infty}$.

2. r_{near} : $\quad \Delta\tilde{x} = |x - \tilde{x}| \leq \frac{1}{2} \cdot 10^{-3}$

$$\begin{array}{lll} \tilde{x} = 0 & \Rightarrow x \in [-0.0005, 0.0005) \Rightarrow & \left|\frac{\Delta\tilde{x}}{x}\right| \leq \infty \\ \tilde{x} = 0.001 & \Rightarrow x \in [0.0005, 0.0015) \Rightarrow & \left|\frac{\Delta\tilde{x}}{x}\right| \leq 1 \\ \tilde{x} = 0.002 & \Rightarrow x \in [0.0015, 0.0025) \Rightarrow & \left|\frac{\Delta\tilde{x}}{x}\right| \leq 0.33 \\ \tilde{x} = 0.003 & \Rightarrow x \in [0.0025, 0.0035) \Rightarrow & \left|\frac{\Delta\tilde{x}}{x}\right| \leq 0.20 \\ \vdots & & \\ \tilde{x} = 9999.998 & \Rightarrow x \in [9999.9975, 9999.9985) \Rightarrow & \left|\frac{\Delta\tilde{x}}{x}\right| \leq \frac{1}{2} \cdot 10^{-7} \end{array}$$

Für naturwissenschaftliche/technische Anwendungen sind also etwa nur zwei führende Ziffern sinnvoll. Im Beispiel heißt das

$$(-1)^s \cdot \underbrace{(z_1 10^{-1} + z_2 10^{-2})}_{\text{norm. Mantisse}} \cdot \underbrace{10^e}_{\text{Basis}}$$

mit $s \in \{0,1\}, z_{1/2} \in \{0, \ldots, 9\}, z_1 \neq 0$ und einem Exponenten $e \in \{-6, -5, \ldots, 5\}$. Allgemein ergibt sich

$$(-1)^s \cdot \left(\sum_{i=1}^{N} z_i b^{-i} \right) \cdot b^e, \quad z_1 \neq 0 \qquad (\circ)$$

oder

$$(-1)^s \cdot \left(\sum_{i=0}^{N} z_i b^{-i} \right) \cdot b^e, \quad z_0 \neq 0$$

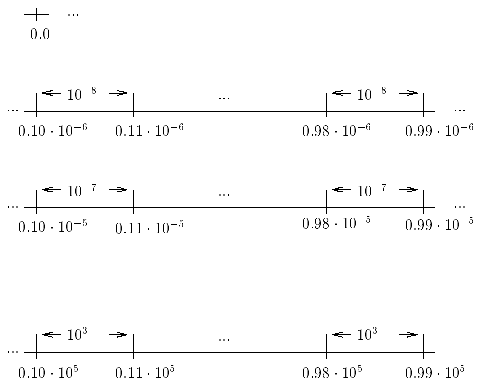
Zur Untersuchung des Zahlenbereichs der durch $(\circ)$ gegebenen Zahlen sei

$$\tilde{\mathcal{G}} := \left\{ (-1)^s \cdot \left(\sum_{i=1}^{N} z_i b^{-i} \right) \cdot b^e | s \in \{0,1\}, z_i \in \{0, 1, \ldots, b-1\}, z_1 \neq 0, e \in \{e_{min}, \ldots, e_{max}\} \right\}$$

Man beachte, daß $\tilde{\mathcal{G}}$ <u>nicht</u> die 0 enthält! Man setzt daher $\mathcal{G} := \tilde{\mathcal{G}} \cup \{0\}$, $\max \mathcal{G} := \max \tilde{\mathcal{G}}$ und $\min \mathcal{G} := \min \tilde{\mathcal{G}}$. Für $\tilde{\mathcal{G}}$ erhält man z.B. mit $b = 10, N = 2$ und $e \in \{-6, \ldots, 5\}$:

$$\begin{array}{rcl} \{-0.99 \cdot 10^5, -0.98 \cdot 10^5, & \ldots & , -0.11 \cdot 10^5, -0.10 \cdot 10^5, \\ -0.99 \cdot 10^4, -0.98 \cdot 10^4, & \ldots & , -0.11 \cdot 10^4, -0.10 \cdot 10^4, \\ & \ldots & \\ -0.99 \cdot 10^{-6}, -0.98 \cdot 10^{-6}, & \ldots & , -0.11 \cdot 10^{-6}, -0.10 \cdot 10^{-6}, \\ 0.10 \cdot 10^{-6}, 0.11 \cdot 10^{-6}, & \ldots & , 0.98 \cdot 10^{-6}, 0.99 \cdot 10^{-6}, \\ 0.10 \cdot 10^{-5}, 0.11 \cdot 10^{-5}, & \ldots & , 0.98 \cdot 10^{-5}, 0.99 \cdot 10^{-5}, \\ & \ldots & \\ 0.10 \cdot 10^4, 0.11 \cdot 10^4, & \ldots & 0.98 \cdot 10^4, 0.99 \cdot 10^4, \\ 0.10 \cdot 10^5, 0.11 \cdot 10^5, & \ldots & 0.98 \cdot 10^5, 0.99 \cdot 10^5\} \end{array}$$

Für die Rundungsfehler ergibt sich im Beispiel mit r_0 folgendes Bild:



Eine Rechnung liefert für die relativen Fehler:

$$\tilde{x} = 0.0 \quad \Rightarrow x \in (-0.10 \cdot 10^{-6}, 0.10 \cdot 10^{-6}) \Rightarrow \quad \left|\frac{\tilde{x}-x}{x}\right| \leq \infty$$

$$\tilde{x} = 0.10 \cdot 10^{-6} \quad \Rightarrow x \in [0.10 \cdot 10^{-6}, 0.11 \cdot 10^{-6}) \Rightarrow \quad \left|\frac{\tilde{x}-x}{x}\right| \leq \frac{10^{-8}}{10^{-7}} = 0.1$$

$$\tilde{x} = 0.11 \cdot 10^{-6} \quad \Rightarrow x \in [0.11 \cdot 10^{-6}, 0.12 \cdot 10^{-6}) \Rightarrow \quad \left|\frac{\tilde{x}-x}{x}\right| \leq \frac{10^{-8}}{0.11 \cdot 10^{-6}} \leq 0.\overline{09}$$

$$\vdots$$

$$\tilde{x} = 0.99 \cdot 10^{-6} \quad \Rightarrow x \in [0.99 \cdot 10^{-6}, 1.00 \cdot 10^{-6}) \Rightarrow \quad \left|\frac{\tilde{x}-x}{x}\right| \leq \frac{10^{-8}}{0.99 \cdot 10^{-6}} \leq 0.\overline{01}\ldots$$

Also mit $p = N$:

$$\boxed{\Delta\tilde{x} \leq b^{exponent(\tilde{x})-p}}$$

Ist nun $\tilde{x} \in S(b, p, emin, emax) \setminus \{0.0\}$ und dabei ohne Einschränkung $\tilde{x} > 0$, so gilt bei der Rundung r_0 für $x \in [\tilde{x}, \tilde{x} + b^{-p+exponent(\tilde{x})})$

$$\begin{aligned}
\left|\frac{x-\tilde{x}}{x}\right| &\leq \left|\frac{\tilde{x}}{x} - 1\right| \\
&\leq \left|\frac{\tilde{x}}{\tilde{x} + b^{-p+exponent(\tilde{x})}} - 1\right| \\
&= \left|\frac{mantisse(\tilde{x})}{mantisse(\tilde{x}) + b^{-p}} - 1\right| \\
&= \left|\frac{b^{-p}}{mantisse(\tilde{x}) + b^{-p}}\right| \\
&= \left(\frac{mantisse(\tilde{x})}{b^{-p}} + 1\right)^{-1} \\
&\approx \frac{b^{-p}}{mantisse(\tilde{x})} \\
&\in \left(b^{-p}, b^{1-p}\right],
\end{aligned}$$

wobei benutzt wurde, daß für große α gilt: $\frac{1}{1+\alpha} \approx \frac{1}{\alpha}$. Bei der Rundungsart r_{near} erhält man folgende Abschätzungen:

$$\tilde{x} = 0.0 \qquad\qquad \Rightarrow \quad ?$$

$$\tilde{x} = 0.10 \cdot 10^{-6} \qquad\qquad \Rightarrow \quad ?$$

$$\tilde{x} = 0.11 \cdot 10^{-6} \quad \Rightarrow x \in [0.105 \cdot 10^{-6}, 0.115 \cdot 10^{-6}) \Rightarrow \quad \left|\frac{\tilde{x}-x}{x}\right| \leq \frac{\frac{1}{2} \cdot 10^{-8}}{0.105 \cdot 10^{-6}} \leq 0.0476\ldots$$

$$\vdots$$

$$\tilde{x} = 0.99 \cdot 10^{-6} \quad \Rightarrow x \in [0.985 \cdot 10^{-6}, 0.995 \cdot 10^{-6}) \Rightarrow \quad \left|\frac{\tilde{x}-x}{x}\right| \leq \frac{\frac{1}{2} \cdot 10^{-8}}{0.985 \cdot 10^{-6}} \leq 0.005076\ldots$$

$$\vdots$$

$$\tilde{x} = 0.99 \cdot 10^{5} \qquad\qquad \Rightarrow \quad ?$$

Für r_{near} gilt also:

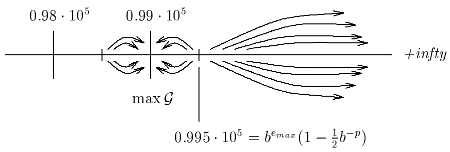
$$\boxed{\Delta\tilde{x} \leq \frac{1}{2} \cdot 10^{exponent(\tilde{x})-p}}$$

Damit die Formel von r_{near} auch für $\tilde{x} = 0.99 \cdot 10^5$ gilt, muß

$$x \in [0.985 \cdot 10^5, 0.995 \cdot 10^5)$$

auf $\tilde{x}$ gerundet werden; alle $x \geq 0.995 \cdot 10^5$ werden zu $+infty$.

$$\mathcal{G}(b, p, e_{min}, e_{max}) = \mathcal{G}(10, 2, -6, 5)$$



Nahe 0 arbeiten *alte* Arithmetiken mit

$$\tilde{x} = 0.10 \cdot 10^{-6} \quad \Rightarrow x \in [0.10 \cdot 10^{-6}, 0.15 \cdot 10^{-6}),$$

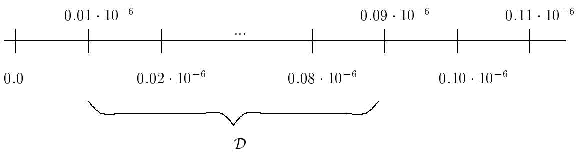
d.h. die Formel gilt auch hier, könnte jedoch noch verschärft werden, und

$$\tilde{x} = 0.0 \quad \Rightarrow x \in (-0.10 \cdot 10^{-6}, +0.10 \cdot 10^{-6}) \Rightarrow \quad \left\{ \begin{array}{rcl} \Delta\tilde{x} & = & 2 \cdot 0.10 \cdot 10^{-6} \\ \left|\frac{\tilde{x}-x}{x}\right| & < & \infty \end{array} \right.$$

Es ergibt sich folgendes **Problem**: Die Äquivalenz

$$\boxed{x - y > 0 \quad \Leftrightarrow \quad x > y}$$

gilt im allgemeinen hier nicht mehr. So wird z.B. die Differenz $0.11 \cdot 10^{-6} - 0.10 \cdot 10^{-6}$ nach 0 gerundet, weil $10^{-8} < 0.10 \cdot 10^{-6} = minReal$ ist, aber natürlich ist $0.11 \cdot 10^{-6} > 0.10 \cdot 10^{-6}$. Da Festpunktsysteme die Addition exakt ausführen, ergänzen wir $\mathcal{G}(b, p, e_{min}, e_{max})$ um Null herum um die Rasterpunkte des Festkommasystems mit $\Delta = b^{e_{min}-p}$, d.h. im Beispiel $\Delta = b^{-8}$.



Damit ergeben sich in $\mathcal{D}$ folgende Fehlerformeln für $\tilde{x} \in \mathcal{D} \setminus \{0.0\}$: Bei der Rundungsart r_0 ist $\Delta\tilde{x} = b^{minexp}$ und somit mit $\tilde{mant}(\tilde{x}) = \frac{\tilde{x}}{b^{minexp}}$

$$\begin{aligned}
\left|\frac{\tilde{x}-x}{\tilde{x}}\right| &\leq \frac{b^{minexp-p}}{\tilde{mant}(\tilde{x}) \cdot b^{minexp}} = \frac{b^{-p}}{\tilde{mant}(\tilde{x})} \\
\left|\frac{\tilde{x}-x}{x}\right| &\leq \left(\frac{\tilde{mant}(\tilde{x})}{b^{-p}} + 1\right)^{-1}.
\end{aligned}$$

Für r_{near} ist $\Delta\tilde{x} = \frac{1}{2}b^{minexp-p}$ und daher

$$\left|\frac{\tilde{x}-x}{x}\right| \leq \frac{1}{2} \frac{b^{-p}}{\tilde{mant}(\tilde{x}) - \frac{1}{2}b^{-p}}$$

Aufgabe 1.5 *Berechne für $0.01 \cdot 10^{-6}$ die Abschätzungen für r_0 und r_{near}.*

Hochgenaue Standardfunktionen können i.a. nur auf drei benachbarte Gitterpunkte genau berechnet werden:

$$\frac{1}{3} \cdot 3 \in [0.99...9, 1.00...01]$$

Für ein (unbekanntes) Funktionsergebnis x und die gelieferte Näherung $\tilde{x}$ gilt dann also z.B. bei der Rundung r_0:

$$\begin{aligned}
\tilde{x} = 0.15 \cdot 10^{-6} \quad &\Rightarrow \quad x \in [0.15 \cdot 10^{-6}, 0.17 \cdot 10^{-6}] \\
&\Rightarrow \quad \left|\frac{x-\tilde{x}}{x}\right| \leq \left|\frac{mant(\tilde{x})}{mant(\tilde{x}) + 2 \cdot 10^{-2}} - 1\right| = \left(\frac{mant(\tilde{x})}{2 \cdot 10^{-2}} + 1\right)^{-1} \\
&\in \quad [0.0198..., 0.166...]
\end{aligned}$$

für Argumente im normalisierten Bereich.

Aufgabe 1.6 *Wie sieht das im unnormierten Bereich aus?*