Übung 2 - Numerisches Programmieren

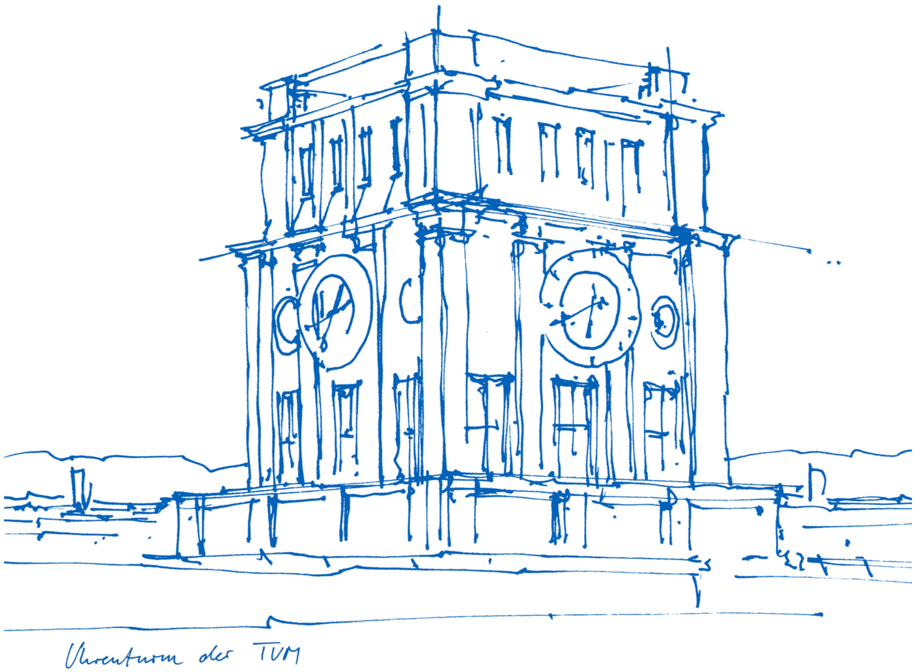
Michael Obersteiner

Technische Universität München

Fakultät für Informatik

Lehrstuhl für Wissenschaftliches Rechnen

Garching, 2. November 2021



Recap – IEEE Gleitkomma: Definition

- Darstellung einer Gleitkommazahl durch 3 Komponenten:
  - Mantisse
  - Exponent
  - Vorzeichen
- Beispiel: $-1{,}00101 * 2^{5}$
- Normalisierung: Immer **genau** eine 1 vor dem Komma!
  - $101{,}01 * 2^{5} \rightarrow 1{,}0101 * 2^{7}$
  - $0{,}10101 * 2^{5} \rightarrow 1{,}0101 * 2^{4}$
- Speicherlayout:

Vorzeichen	Exponent	Mantisse

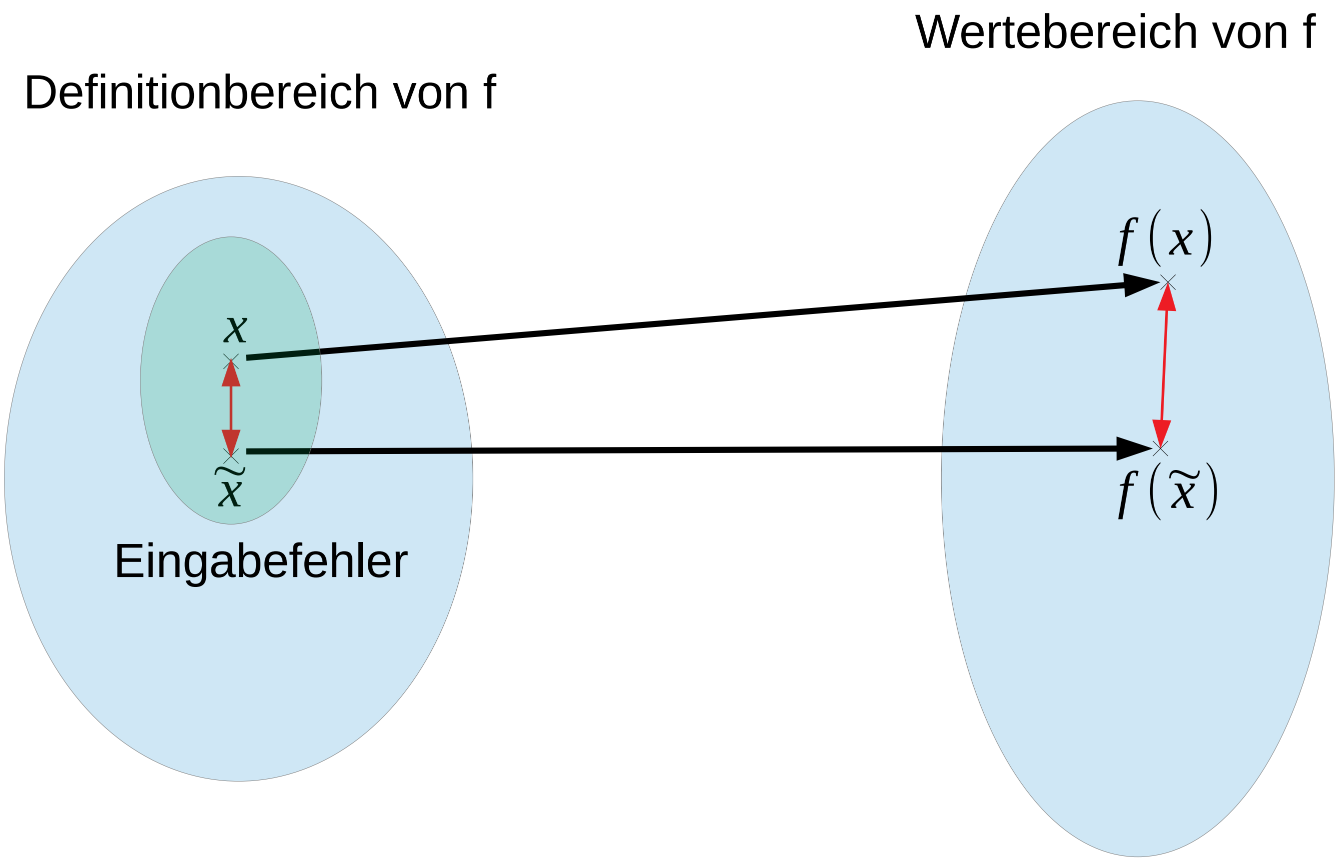
  - Vorzeichen (1 Bit): 1 Negativ 0 Positiv
  - Exponent (8 Bit): Exponent + Offset 127 (Vermeidet 2er Komplement)
  - Mantisse (23 Bit): Nachkommastellen

Recap – Gleitkomma: Maschinengenauigkeit

- Maß für die relative Genauigkeit der Gleitkommazahlen
- Definition: $\epsilon_{Ma} = max_x(rd(1+x)=1)$
- Faustregel: bei m Mantissenbits $2^{-(m+1)}$
- Aussage: $\epsilon_{rel}(x) = \left|\frac{rd(x)-x}{x}\right| < \epsilon_{Ma}$

- Andere Definitionen in der Literatur:
  - Kleinste Schrittweite $\rightarrow$ 2^{-m}

Übung 2 – Kondition



Übung 2 – Kondition und Stabilität

Kondition

- Abhängigkeit der Ausgabe von Eingabe (Verstärkungseffekt?)
- Ausgabefehler < Eingabefehler * c

$$\left|\frac{f(\tilde{x})-f(x)}{f(x)}\right| < \left|\frac{\tilde{x}-x}{x}\right| \cdot c(f,x)$$

- z.B. Approximation durch Ableitung:

$$c(f,x)=\left|\frac{x\cdot f'(x)}{f(x)}\right|$$

- Problemspezifisch und nicht von Implementierung abhängig!
- Vorkonditionierung kann helfen.

Stabilität

- Wie wirken sich interne Rundungsfehler auf Ausgabe aus?
- Relativer Fehler $\left|\frac{\tilde{f}(\tilde{x})-f(\tilde{x})}{f(\tilde{x})}\right|$
- Abschätzung durch Epsilontik: rd(x op y) = (x op y)(1+ε); $\varepsilon < \varepsilon_{Ma}$
- Bsp: rd(x + y) = (x + y)(1+ε)
- Implementierungsabhängig!
- Umformung der Operation kann helfen.

Übung 2 – Kondition

Bearbeitung Aufgabe 1

i) $f_1(x) = a \cdot x$

ii) $f_2(x) = (a - x)/b$

iii) $f_3(x) = 3e^x - 3$

Übung 2 – Kondition

Bearbeitung Aufgabe 2

$$\text{Schnittpunkte:} \quad g_1 : y = mx + 1 \qquad g_2 : y = x$$

i) Schnittpunkt in Abhängigkeit zu m (als f(m))

ii) Berechnung der Kondition von f(m) an Punkt $m = 1,005$

iii) Wie ist die tatsächliche Verstärkung ($m = 1,005$, $\tilde{m} = 1,01$)?

Übung 2 - Epsilontik

- Beispiel: $f(x) = a + b * x$

$$rd(a + b * x) = (a + b * x * (1 + \varepsilon_1)) * (1 + \varepsilon_2)$$

$$= a * (1 + \varepsilon_2) + b * x * (1 + \varepsilon_1) * (1 + \varepsilon_2)$$

$$= a * (1 + \varepsilon_2) + b * x * (1 + \varepsilon_1 + \varepsilon_2 + \varepsilon_1\varepsilon_2)$$

$$\approx a * (1 + \varepsilon_2) + b * x * (1 + \varepsilon_1 + \varepsilon_2)$$

$$= f(x) (1 + \varepsilon_2) + b * x * \varepsilon_1$$

Relativer Fehler:

$$|(f(x) - rd(f(x))) / f(x)| = |(f(x) * \varepsilon_2 + b * x * \varepsilon_1)/ f(x)|$$

$$\text{(Dreiecksungleichung)} \leq |(f(x) * \varepsilon_2)/f(x)| + |(b * x * \varepsilon_1)/ f(x)|$$

$$= |\varepsilon_2| + |(b * x * \varepsilon_1)/ f(x)| \leq \varepsilon_{Ma} + |(b * x * \varepsilon_{Ma})/ (a + b*x)|$$

Problem falls $x \approx -a/b$

Übung 2 – Stabilität

Bearbeitung Aufgabe 3

i) $f_1(x) = a \cdot x$

ii) $f_2(x) = (a - x)/b$

iii) $f_3(x) = 3e^x - 3$

Recap – Numerische Effekte: Auslöschung

- Subtraktion zweier ähnlicher Zahlen in Gleitkommaarithmetik
- **Problem:** Verlust an gültigen Stellen → ungenaue Ergebnisse

- Beispiel:
  - rd(1,00001…..) – rd(1,00000….) = 1,00001 – 1,00000 = **0,00001**
  - Aus 6 gültigen Stellen wird 1!
  - Würden wir nur 5 gültige Stellen verwenden so wäre das Ergebnis 0!

  → Weitere Rechnungen eventuell signifikant verfälscht (Teilen durch 0!)

- Manchmal vermeidbar durch Umformung (siehe Aufgabe 6):
- s → 0

$$s_{new}=\sqrt{2-\sqrt{4-s^2}}=\frac{\sqrt{2-\sqrt{4-s^2}}*\sqrt{2+\sqrt{4-s^2}}}{\sqrt{2+\sqrt{4-s^2}}}=\frac{|s|}{\sqrt{2+\sqrt{4-s^2}}}$$

Recap – Numerische Effekte: Absorption

- Addieren zweier Zahlen unterschiedlicher Größenordnung
- **Problem:** Minimale oder keine Änderung der größeren Zahl
  → Genauigkeitsverlust

- Beispiel:
  - 100000000 + 1 = rd(100000001) = 100000000

- Besonders problematisch bei for loops! (JavaScript kennt nur float!)
- Manchmal lösbar durch Änderung der Additionsreihenfolge:
  - -1000 + (1011 + 0.11) = -1000 + 1100 = 100 → 4 (ungenau!)
  - (-1000 + 1011) + 0.11 = 11 + 0.11 = 11.11 → 3.75 (exakt!)
- Typischerweise weniger problematisch als Auslöschung