Technische Universität Berlin
Fachgebiet Computer Graphics

Prof. Marc Alexa, Ugo Finnendahl,
Dimitrios Bogiokas, Gabriel Dogadov

Wissenschaftliches Rechnen - Übung 1.2

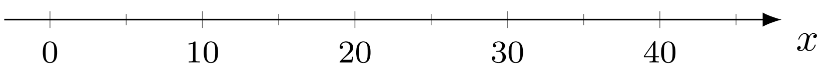
Gleitkommazahlen

06.11.2023 bis 10.11.2023

Aufgabe 1: Fest- und Gleitkommazahlen

Zum Darstellen von Dezimalzahlen in Computern werden hauptsächlich Gleitkommazahlen verwendet. Diese trifft man in modernen Programmiersprachen meist unter den Namen *float* und *double*. Ziel dieser Aufgabe ist es, die speziellen Eigenschaften von Gleitkommazahlen, im Kontrast zu Festkommazahlen, hervorzuheben.

1. Was ist eine Festkommazahl? Wie ist das Festkommazahlenformat $\mathbb{F}(b, n_b, n_a)$ aufgebaut?
2. Nennen Sie die Bestandteile einer Gleitkommazahl und erklären Sie die Bedeutung des Gleitkommazahlenformats $\mathbb{G}(b, n_m)$. Wodurch kommt das „Gleiten“ zustande?
3. Welche der folgenden Zahlen ist im Format $\mathbb{G}(10, 3)$ exakt darstellbar? Begründen Sie Ihre Entscheidung.
   a) 0,713
   b) 12,79
   c) 211.000.000.000.000
   d) 3,0001
4. Zeichnen Sie die exakt darstellbaren Gleitkommazahlen in $\mathbb{G}(10, 1)$ in den folgenden Zahlenstrahl so gut es geht hinein.



Aus den letzten Aufgaben wird ersichtlich, dass nicht alle reellen Zahlen als Gleitkommazahlen in $\mathbb{G}(b, n_m)$ exakt darstellbar sind. Hierzu sei $G : \mathbb{R} \to \mathbb{G}(b, n_m),\ x \mapsto G(x)$ eine Funktion, die eine reelle Zahl x auf eine in dem entsprechenden Format darstellbare Gleitkommazahl $G(x)$ sinnvoll rundet. Die bei der Rundung entstehenden Fehler sind wie folgt definiert:

$$\text{Absoluter Fehler: } E_a = |x - G(x)| \quad \text{und} \quad \text{Relativer Fehler: } E_r = \left|\frac{x - G(x)}{x}\right|.$$

5. Was ist der Unterschied zwischen dem absoluten und dem relativen Fehler und warum ist der relative Fehler im Allgemeinen aussagekräftiger als der absolute Fehler?
6. Was ist die Maschinengenauigkeit ϵ? Welche Bedeutung, in Bezug auf Rundungsfehler sowie die Verteilung von Gleitkommazahlen, hat sie?
7. Diskutieren Sie die wesentlichen Vor- und Nachteile von Gleitkommazahlen gegenüber Festkommazahlen.

Aufgabe 2: Rechnen in Gleitkommazahlen

Gegeben sind die Zahlen $a, b \in \mathbb{R}$ mit $a = 3{,}578$ und $b = 40{,}124$. Wir möchten nun die Rechnung $a + b$ im Gleitkommaformat $\mathbb{G}(10, 3)$ durchführen. Bei Berechnungen mit Gleitkommazahlen entstehen ebenso Fehler beim Runden der Ergebnisse, die analog definiert sind:

- Absoluter Fehler: $E_a = |x - \hat{x}|$ und

- Relativer Fehler: $E_r = \left|\frac{x-\hat{x}}{x}\right|$.

Dabei ist $\hat{x} = G(G(a) \bullet G(b))$ mit $\bullet \in \{+, -, \cdot, \div\}$ das Ergebnis der Rechnung in Gleitkommazahlen. Sie dürfen dabei auswählen, ob die Funktion G abrunden oder kaufmännisch runden soll.

1. Geben Sie die gerundeten Zahlen $G(a)$ und $G(b)$ an.
2. Berechnen Sie das Ergebnis der Rechnung in Gleitkommazahlen $\hat{x} = G(G(a) + G(b))$. Beachten Sie, dass die Addition von $G(a)$ und $G(b)$ der üblichen Addition in den reellen Zahlen entspricht.
3. Berechnen Sie den absoluten Fehler E_a sowie den relativen Fehler E_r.
4. Vergleichen Sie den relativen Fehler mit der Maschinengenauigkeit für das hier verwendete Gleitkommaformat.

Aufgabe 3: Subtraktion und Auslöschung

Nun möchten wir die Rechnung $a-b$ mit $a = 11{,}1556$ und $b = 11{,}1264$ im Format $\mathbb{G}(10, 3)$ durchführen und erneut den dadurch entstandenen relativen Fehler mit der Maschinengenauigkeit vergleichen. Erneut dürfen Sie entscheiden, ob die Funktion G abrunden oder kaufmännisch runden soll.

1. Geben Sie die gerundeten Zahlen $G(a)$ und $G(b)$ an.
2. Berechnen Sie das Ergebnis der Rechnung in Gleitkommazahlen $\hat{x} = G(G(a) - G(b))$. Beachten Sie, dass die Subtraktion von $G(a)$ und $G(b)$ der üblichen Subtraktion in den reellen Zahlen entspricht.
3. Berechnen Sie den absoluten Fehler E_a sowie den relativen Fehler E_r.
4. Vergleichen Sie den relativen Fehler erneut mit der Maschinengenauigkeit. Welches Phänomen können Sie an dieser Rechnung beobachten?

Aufgabe 4: Fehler und Schranken

1. Bei welchen Rechenoperationen mit *ausschließlich positiven* Gleitkommazahlen ist der relative Fehler durch ein konstantes Vielfaches der Maschinengenauigkeit beschränkt und bei welchen nicht?

   a) Addition b) Subtraktion c) Multiplikation d) Division

2. Welche der folgenden Gesetze gelten für Gleitkommazahlen? Dabei seien $a, b, c \in \mathbb{G}$ beliebig.

   a) Assoziativgesetz für die Addition: $a + (b + c) = (a + b) + c$
   b) Assoziativgesetz für die Multiplikation: $a(bc) = (ab)c$
   c) Distributivgesetz: $a(b + c) = ab + ac$
   d) Transitivität bzgl. Kleiner: $a > b \wedge b > c \Rightarrow a > c$
   e) Transitivität bzgl. Gleich: $a = b \wedge b = c \Rightarrow a = c$
   f) Antisymmetrie $a \leq b \wedge b \leq a \Rightarrow a = b$

3. Wie sollte man zwei Gleitkommazahlen $x, y \in \mathbb{G}$ im Programmcode auf Gleichheit überprüfen?

$*$ Gegeben sei das dezimale Gleitkommazahlenformat $\mathbb{G}(10, 3)$, wobei $G : \mathbb{R} \to \mathbb{G}(10, 3)$ jede Zahl auf die nächste Zahl in $\mathbb{G}(10, 3)$ abrundet. Zeigen Sie, dass der relative Fehler bei der Subtraktion beliebig groß werden kann, indem Sie für ein beliebiges $k \in \mathbb{R}^+$ zwei reelle Zahlen $a, b \in \mathbb{R}$, u.U. in Abhängigkeit von k, angeben, sodass der relative Fehler bei der Berechnung $a - b$ in GKZ größer oder gleich k beträgt.