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?

   Lösung

   Festkommazahlen haben eine feste Anzahl an Stellen vor dem Komma und hinter dem Komma. Das Format $\mathbb{F}(b, n_b, n_a)$ beschreibt alle Zahlen, die sich wie folgt darstellen lassen:

   $$\pm \underbrace{x_{n_b} \dots x_1}_{n_b \text{ Stellen}}, \underbrace{y_{n_a} \dots y_1}_{n_a \text{ Stellen}}$$

   Dabei sind $x_1, \dots, x_n, y_1, \dots, y_m$ Ziffern aus $\{0, \dots, b-1\}$.

   Lösung Ende

2. Nennen Sie die Bestandteile einer Gleitkommazahl und erklären Sie die Bedeutung des Gleitkommazahlenformats $\mathbb{G}(b, n_m)$. Wodurch kommt das „Gleiten“ zustande?

   Lösung

   Eine Gleitkommazahl besteht aus Vorzeichen, Basis, Mantisse und Exponent. Dabei beschreibt das Format $\mathbb{G}(b, n_m)$ alle Zahlen, die als

   $$\pm \underbrace{m_{n_m}, m_{n_m-1} \dots m_1}_{n_m \text{ Stellen}} \cdot b^e \quad e \in \mathbb{Z}$$

   geschrieben werden können (die Darstellung ist bei manchen Zahlen mehrdeutig). Dabei sind alle Stellen der Mantisse $m_1, \dots, m_{n_m}$ eine Ziffer aus $\{0, \dots, b-1\}$ und der Exponent $e \in \mathbb{Z}$ beliebig wählbar (Anmerkung: In der Praxis ist der Exponent jedoch beschränkt, da nur eine bestimmte Anzahl an Bits für diesen zur Verfügung stehen). Insbesondere sind alle GKZ-Formate in diesem Kurs (abzählbar) unendliche Mengen. Das Gleiten kommt dadurch zustande, dass die Mantisse mit der Basis in der Potenz zu jedem beliebigen Exponenten multipliziert werden kann.

   Lösung Ende

3. Welche der folgenden Zahlen ist im Format $\mathbb{G}(10, 3)$ exakt darstellbar? Begründen Sie Ihre Entscheidung.

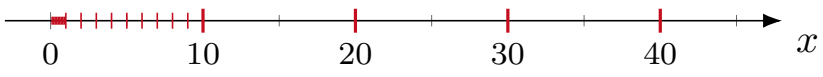
   a) 0,713 Darstellbar, da $0{,}713 = 7{,}13 \cdot 10^{-1}$.

   b) 12,79 Nicht darstellbar, da für die Mantisse $1{,}279$ vier Mantissenstellen benötigt werden.

   c) 211.000.000.000.000 Darstellbar, entspricht $2{,}11 \cdot 10^{14}$.

   d) 3,0001 Nicht darstellbar, da für die Mantisse $3{,}0001$ fünf Mantissenstellen benötigt werden.

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?

Lösung

Der relative Fehler entspricht dem absoluten Fehler, der durch den Betrag des eigentlichen Ergebnisses dividiert wird. Dadurch ist er unabhängig von der Größenordnung der Zahlen und somit für viele Zwecke aussagekräftiger.

Lösung Ende

6. Was ist die Maschinengenauigkeit ϵ? Welche Bedeutung, in Bezug auf Rundungsfehler sowie die Verteilung von Gleitkommazahlen, hat sie?

Lösung

Die Maschinengenauigkeit ist ein Maß für den Rundungsfehler, der bei Rechnungen mit Gleitkommazahlen auftritt. Diese lässt sich wie folgt berechnen:

$$\epsilon = b^{-n_m+1}.$$

Die Maschinengenauigkeit ist eine obere Schranke für den relativen Rundungsfehler:

$$\frac{|x - G(x)|}{|x|} \leq \epsilon.$$

Dies gilt für alle beliebigen $x \in \mathbb{R}^+$ und sinnvollen Rundungsfunktionen G (abrunden, aufrunden, kaufmännisch runden, ...). Zum Anderen ist sie die kleinste Zahl, die auf Eins addiert, nicht auf Eins abgerundet wird:

$$\epsilon = \arg\min_{\delta \in \mathbb{G}} G(1 + \delta) > 1, \quad \text{wobei } G \text{ abrundet.}$$

Dies erkennt man im obigen Zahlenstrahl: Die Maschinengenauigkeit ist der Abstand zweier benachbarter Gleitkommazahlen zwischen 1 und b (im Falle von $\mathbb{G}(10, 1)$ ist dann $\epsilon = 1$).

Lösung Ende

7. Diskutieren Sie die wesentlichen Vor- und Nachteile von Gleitkommazahlen gegenüber Festkommazahlen.

Lösung

Vorteile von Gleitkommazahlen:

- GKZ haben einen sehr großen Wertebereich. Es lassen sich sowohl sehr große als auch (im Betrag) sehr kleine Zahlen darstellen.
- Festkommazahlen haben hingegen einen beschränkten Wertebereich. Sie benötigen für Berechnungen, bei denen die Werte in stark unterschiedlichen Größenordnungen sind, viel Speicherplatz, um Über- und Unterläufe zu vermeiden.

Nachteile von Gleitkommazahlen:

- Die Zahlen sind nicht gleichmäßig über den Bereich verteilt. Je größer der Exponent wird, umso weiter weg sind zwei benachbarte Zahlen.
- Rechnungen mit Gleitkommazahlen sind fehleranfällig (dazu folgt unten mehr).

Lösung Ende

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.

   Lösung

   - Abrunden: $G(a) = 3{,}57$, $G(b) = 40{,}1$
   - Kaufmännisch: $G(a) = 3{,}58$, $G(b) = 40{,}1$

   Lösung Ende

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.

   Lösung

   - Abrunden: $\hat{x} = G(43{,}67) = 43{,}6$
   - Kaufmännisch: $\hat{x} = G(43{,}68) = 43{,}7$

   Lösung Ende

3. Berechnen Sie den absoluten Fehler E_a sowie den relativen Fehler E_r.

   Lösung

   Das eigentliche Ergebnis lautet $x = 43{,}702$.

   - Abrunden: $E_a = 0{,}102$, $E_r \approx 2{,}33 \cdot 10^{-3}$
   - Kaufmännisch: $E_a = 0{,}002$, $E_r \approx 4{,}58 \cdot 10^{-5}$

   Lösung Ende

4. Vergleichen Sie den relativen Fehler mit der Maschinengenauigkeit für das hier verwendete Gleitkommaformat.

   Lösung

   - $\epsilon = 10^{-2} = 0{,}01$

   Der relative Fehler ist in jedem Fall deutlich kleiner als die Maschinengenauigkeit.

   Lösung Ende

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.

   Lösung

- Abrunden: $G(a) = 11{,}1$, $G(b) = 11{,}1$
- Kaufmännisch: $G(a) = 11{,}2$, $G(b) = 11{,}1$

Lösung Ende

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.

Lösung

- Abrunden: $\hat{x} = 0$
- Kaufmännisch: $\hat{x} = 0{,}1$

Lösung Ende

3. Berechnen Sie den absoluten Fehler E_a sowie den relativen Fehler E_r.

Lösung

Das eigentliche Ergebnis lautet $x = 0{,}0292$.

- Abrunden: $E_a = 0{,}0292$, $E_r = 1{,}00$
- Kaufmännisch: $E_a = 0{,}0708$, $E_r \approx 2{,}42$

Lösung Ende

4. Vergleichen Sie den relativen Fehler erneut mit der Maschinengenauigkeit. Welches Phänomen können Sie an dieser Rechnung beobachten?

Lösung

Diesmal ist der relative Fehler, unabhängig davon welche Rundungsfunktion gewählt wurde, mehrere Größenordnungen größer als die Maschinengenauigkeit. Das beobachtete Phänomen nennt sich Auslöschung, bei der der relative Fehler beliebig groß werden kann. Dies geschieht dann, wenn zwei ähnlich große Zahlen mit gleichem Vorzeichen subtrahiert werden und das Ergebnis in einer anderen Größenordnung als die Zahlen ist.

Lösung Ende

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 Ja b) Subtraktion Nein c) Multiplikation Ja d) Division Ja

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

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

   Wichtige Anmerkung: Für die positiven Eigenschaften d), e) und f) gilt als Prämisse, dass a, b und c Zahlen aus dem Gleitkommaformat sind. Zwar existiert in der Theorie für Gleitkommazahlen eine totale Ordnung (bzgl. $\leq$) und es gilt die Transitivität bezüglich $=$, jedoch ist diese Eigenschaft für die Praxis wenig relevant, da man in der Regel mit gerundeten Ergebnissen arbeitet. Zum Beispiel: Falls zwei Zahlen x und y ähnlich groß sind (siehe nächste Aufgabe) und y ähnlich groß zu z ist, heißt es im Allgemeinen nicht, dass auch x ähnlich groß wie z ist.

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

Lösung

Da bei Berechnungen mit Gleitkommazahlen numerische Fehler entstehen, sollte man auf gar keinen Fall den == Operator verwenden, da dieser alle Bits auf Gleichheit überprüft. Stattdessen sollte geguckt werden, ob die Differenz der beiden Zahlen eine Toleranzgrenze nicht überschreitet:

$$|x - y| \leq k \cdot \epsilon$$

Dabei ist die Toleranzgrenze meist ein (kleines) Vielfaches der Maschinengenauigkeit. Anmerkung: Hier arbeiten wir mit einer absoluten Toleranz. Alternativ kann man zwei Zahlen mithilfe einer relativen Toleranz (oder mit einer Mischung aus beiden) vergleichen. Siehe dazu die Dokumentation von **np.isclose**.

Lösung Ende

$*$ 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.

Lösung

Wähle $a = 1$ und $b = 1 - \left(\frac{0{,}001}{k+1}\right)$. Zunächst ist das Ergebnis der Rechnung in den reellen Zahlen $x = a - b = \frac{0{,}001}{k+1}$. In Gleitkommazahlen erhalten wir folgendes Ergebnis:

$$G(a) = 1, \quad G(b) = 0{,}999, \quad \hat{x} = 0{,}001,$$

da $G\left(\frac{0{,}001}{k+1}\right)$ stets größer als Null ist und damit b immer auf $0{,}999$ abgerundet wird. Der relative Fehler ist dann gegeben durch

$$E_{\text{rel}} = \frac{\left|\frac{0{,}001}{k+1} - 0{,}001\right|}{\frac{0{,}001}{k+1}} = \frac{|0{,}001 - 0{,}001(k+1)|}{0{,}001} = |1 - (k+1)| = k.$$

Lösung Ende