Wissenschaftliches Rechnen - Großübung 1.2

Themen: Gleitkommazahlen, Kondition

Ugo & Gabriel

8. November 2022

Aufgabe 1: Gleitkommazahlen

1. Wie ist der absolute Fehler durch eine fehlerbehaftete Funktion G definiert?

   Lösung

   $$E_a(x) = |G(x) - x|$$

   Lösung Ende

2. Wie ist der relative Fehler durch eine fehlerbehaftete Funktion G definiert?

   Lösung

   $$E_r(x) = \frac{|G(x) - x|}{|x|}$$

   Lösung Ende

3. Gegeben sei das dezimale Gleitkommazahlenformat $\mathbb{G}(10, 3)$ mit 3 Ziffern und das dezimale Festkommazahlenformat $\mathbb{F}(10, 2, 2)$ mit zwei Stellen vor und zwei Stellen nach dem Komma, sowie die Funktionen $G : \mathbb{R} \to \mathbb{G}(10, 3)$ und $F : \mathbb{R} \to \mathbb{F}(10, 2, 2)$, die jeweils auf die nächste darstellbare Zahl **abrunden**.

   a) Geben Sie den Abstand zwischen zwei Zahlen in den jeweiligen Formaten im Intervall $[0{,}1, 1[$ sowie $[10, 100[$ an

   b) Geben Sie die obere Grenze des absoluten Fehlers an, der sich durch G sowie F auf dem Intervallen $[0{,}1, 1[$ sowie $[10, 100[$ ergibt.

   c) Geben Sie die obere Grenze des relativen Fehlers an, der sich durch G sowie F auf dem Intervallen $[0{,}1, 1[$ sowie $[10, 100[$ ergibt.

   Lösung

a)

	$[0{,}1, 1[$	$[10, 100[$
$\mathbb{G}$	0,001	0,1
$\mathbb{F}$	0,01	0,01

b)

	$[0{,}1, 1[$	$[10, 100[$
$\mathbb{G}$	0,001	0,1
$\mathbb{F}$	0,01	0,01

c)

	$[0{,}1,1[$	$[10,100[$
$\mathbb{G}$	0,01	0,01
$\mathbb{F}$	0,1	0,001

Lösung Ende

4. Welche der folgenden Gesetze gelten für Festkommazahlen?
   a) Assoziativgesetz für die Addition: $a + (b + c) = (a + b) + c$ Ja
   b) Distributivgesetz: $a(b + c) = ab + ac$ Nein
   c) Transitivität bzgl. Kleiner: $a > b \wedge b > c \Rightarrow a > c$ Ja
   d) Transitivität bzgl. Gleich: $a = b \wedge b = c \Rightarrow a = c$ Ja
   e) Antisymmetrie $a \leq b \wedge b \leq a \Rightarrow a = b$ Ja

5. Geben Sie die zwei in der Vorlesung/Skript vorgestellten Definitionen der Maschinengenauigkeit an.

Lösung

$$\epsilon = \max_{x\in\mathbb{Q}^+} \frac{|x - G(x)|}{|x|} = \max_{x\in\mathbb{Q}^+} \frac{x - G(x)}{x},$$
$$\epsilon = \arg\min_{x\in\mathbb{Q}} G(1 + x) > 1,$$

Lösung Ende

6. Die zwei Definitionen sind äquivalent für den Fall $G(x) = \text{floor}(x)$, wobei floor auf die nächste Gleitkommazahl des gegebenen Gleitkommzahlenformates abrundet. Überprüfen Sie ob diese Definition für unterschiedliche Funktionen übereinstimmen, indem Sie die Werte der jeweiligen Definition berechnen:
   a) $G_\mathsf{f}(x) = \text{floor}(x)$, wobei floor auf die nächste Gleitkommazahl des gegebenen Gleitkommzahlenformates abrundet.
   b) $G_\mathsf{c}(x) = \text{ceil}(x)$, wobei ceil auf die nächste Gleitkommazahl des gegebenen Gleitkommzahlenformates aufrundet.
   c) $G_\mathsf{r}(x) = \text{round}(x)$, wobei round auf die nächste Gleitkommazahl des gegebenen Gleitkommzahlenformates kaufmännisch rundet.

Lösung

a) $\max_{x\in\mathbb{Q}^+} \frac{|x-G_\mathsf{f}(x)|}{|x|} = b^{1-n_m}$, $\left(\arg\min_{x\in\mathbb{Q}} G_\mathsf{f}(1 + x) > 1\right) = b^{1-n_m}$

b) $\max_{x\in\mathbb{Q}^+} \frac{|x-G_\mathsf{c}(x)|}{|x|} = b^{1-n_m}$, $\left(\arg\min_{x\in\mathbb{Q}} G_\mathsf{c}(1 + x) > 1\right) = 0$

c) $\max_{x\in\mathbb{Q}^+} \frac{|x-G_\mathsf{r}(x)|}{|x|} = \frac{1}{2}b^{1-n_m}$, $\left(\arg\min_{x\in\mathbb{Q}} G_\mathsf{r}(1 + x) > 1\right) = \frac{1}{2}b^{1-n_m}$

Lösung Ende

7. Geben Sie eine sinnvolle Obergrenze für den Fehler, der bei der Division zweier Gleitkommazahlen $x, y \in \mathbb{G}(b, n_m)$ entstehen kann, in Abhängigkeit der Mantissenstellen n_m und Basis b an (relativer Fehler von $\frac{G(x)}{G(y)}$).

Lösung

Bei der Abschätzung wie im Skript erhält man $|\frac{r_y g_x - g_y r_x}{g_x g_y + r_y g_x}| < b\epsilon$.

Lösung Ende

8. Gegeben sei das dezimale Gleitkommazahlenformat $\mathbb{G}(10, 3)$ mit 3 Ziffern sowie eine beliebige Zahl k. Geben Sie eine Subtraktion $x - y$ an, die einen größeren oder gleich großen relativen Fehler hat als/wie k.

Lösung

Wähle $x = 1$ und $y = 1 - (\frac{0{,}001}{k})$. Der Fehler ist dann gegeben durch

$$\frac{1 - 0{,}999}{1 - 1 - \frac{0{,}001}{k}} = \frac{0{,}001}{\frac{0{,}0001}{k}} = k.$$

Lösung Ende

Aufgabe 2: Kondition

Die Kondition[1] einer Matrix $\mathbf{A} \in \mathbb{R}^{n\times n}$ ist definiert als

$$\kappa(\mathbf{A}) = \frac{\max_{\|\mathbf{x}\|=1} \|\mathbf{A}\mathbf{x}\|}{\min_{\|\mathbf{x}\|=1} \|\mathbf{A}\mathbf{x}\|}$$

und charakterisiert den potentiellen numerischen Genauigkeitsverlust jener Matrix. Zunächst kann die Norm $\|\cdot\|$ beliebig gewählt werden. Wie (fast) überall sonst im Kurs wählen wir im Folgenden die euklidische/ℓ^2-Norm.

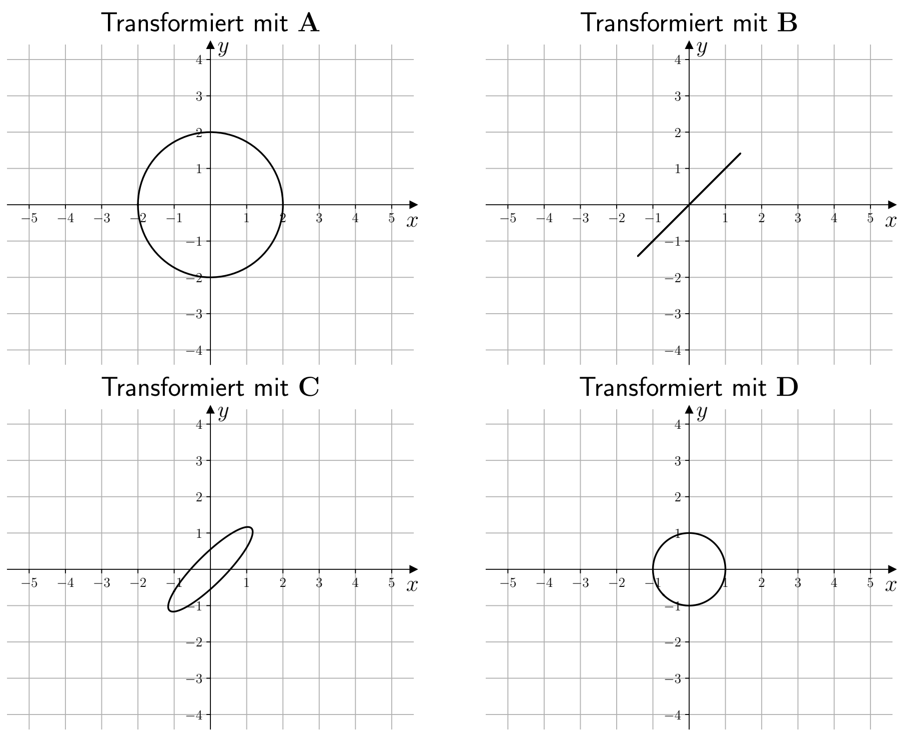
1. Wie sieht die Menge aus, die durch $\|\mathbf{x}\|_2 = 1$ beschrieben wird?

   Lösung

   Die Einheitskugel im $\mathbb{R}^n$.

   Lösung Ende

2. Gegeben seien vier lineare Transformationen $\mathbf{A}, \mathbf{B}, \mathbf{C}, \mathbf{D} \in \mathbb{R}^{2\times 2}$. Im Folgenden ist die Transformation des Einheitskreises unter diesen vier Transformationen zu sehen.



   Entscheiden Sie, ob die folgenden Aussagen gelten oder nicht.

   a) $\mathbf{A}$ ist orthogonal. Falsch
   b) $\mathbf{B}$ ist singulär. Korrekt
   c) $\mathbf{C}$ ist regulär. Korrekt
   d) $\mathbf{D}$ ist orthogonal. Korrekt
   e) $\mathbf{A}$ hat eine Kondition von 1. Korrekt

[1] Falls der Nenner zu Null wird, gilt per Konvention $\kappa(\mathbf{A}) = \infty$.

f) $\mathbf{A}$ hat eine größere Kondition als $\mathbf{D}$. Falsch

g) $\mathbf{C}$ hat eine größere Kondition als $\mathbf{B}$. Falsch

h) $\mathbf{C}$ hat eine größere Kondition als $\mathbf{D}$. Korrekt

3. Geben Sie, unter Zuhilfenahme der Erkenntnisse der vorherigen Aufgabe, eine geometrische Interpretation für die Kondition an.

Lösung

Die Kondition beschreibt die Verzerrung der Einheitskugel nach einer linearen Transformation bzw. den Quotienten aus der stärksten Verlängerung und der stärksten Verkürzung durch besagte Transformation.

Lösung Ende

4. Berechnen Sie die Kondition der folgenden Matrizen:

$$\mathbf{A} = \begin{bmatrix} 12 & 0 \\ 0 & 3 \end{bmatrix}, \qquad \mathbf{B} = \begin{bmatrix} 1/8 & 0 \\ 0 & 8 \end{bmatrix}$$

Lösung

a) $\kappa(\mathbf{A}) = 4$

b) $\kappa(\mathbf{B}) = 64$

Lösung Ende

5. Matrizen mit schlechter Kondition müssen nicht unbedingt einen hohen Genauigkeitsverlust aufweisen. Geben Sie eine Matrix $\mathbf{A} \in \mathbb{G}(10,3)^{3\times 3}$ mit einer endlichen Kondition von größer oder gleich 100 an, welche einen relativen Fehler von 0 für alle Berechnungen $\mathbf{Ax}$ mit $\mathbf{x} \in \mathbb{G}(10,3)^3$ aufweist.

Lösung

$$\mathbf{A} = \begin{bmatrix} 100 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix}$$

Lösung Ende