Kapitel 2

Buch Dahmen-Reusken

RWTH Aachen University

2022

Normierte Räume

Definition

Sei $\boldsymbol{V}$ ein $\mathbb{R}$-Vektorraum. Eine Abbildung $\|\cdot\| : \boldsymbol{V} \to \mathbb{R}$ heißt Norm auf $\boldsymbol{V}$, falls

- $\|\boldsymbol{v}\| \geq \mathbf{0}\ \forall \boldsymbol{v} \in \boldsymbol{V}$ und $\|\boldsymbol{v}\| = \mathbf{0}$ nur wenn $\boldsymbol{v} = \mathbf{0}$.
- Für alle $\boldsymbol{a} \in \mathbb{R}$, $\boldsymbol{v} \in \boldsymbol{V}$ gilt $\|\boldsymbol{a}\,\boldsymbol{v}\| = |\boldsymbol{a}|\,\|\boldsymbol{v}\|$
- Für alle $\boldsymbol{v}, \boldsymbol{w} \in \boldsymbol{V}$ gilt die Dreiecksungleichung
$$\|\boldsymbol{v} + \boldsymbol{w}\| \leq \|\boldsymbol{v}\| + \|\boldsymbol{w}\|$$

Wenn eine Norm auf $\boldsymbol{V}$ definiert ist, nennt man $\boldsymbol{V}$ oft einen linearen normierten Raum.

Vektornormen

Beispiel

Sei $V = \mathbb{R}^n$, $x = (x_1, \dots, x_n) \in \mathbb{R}^n$.
Für jedes p mit $1 \leq p \leq \infty$ wird eine Norm definiert durch

$$\|x\|_p = \left(\sum_{i=1}^{n} |x_i|^p \right)^{\frac{1}{p}}$$

Speziell:

- 1-Norm: $\|x\|_1 = \sum_{i=1}^{n} |x_i|$

Vektornormen

Beispiel

Sei $V = \mathbb{R}^n$, $x = (x_1, \ldots, x_n) \in \mathbb{R}^n$.
Für jedes p mit $1 \leq p \leq \infty$ wird eine Norm definiert durch

$$\|x\|_p = \left(\sum_{i=1}^{n} |x_i|^p \right)^{\frac{1}{p}}$$

Speziell:

- 1-Norm: $\|x\|_1 = \sum_{i=1}^{n} |x_i|$
- ∞-Norm: $\|x\|_\infty = \max_{i=1,\ldots,n} |x_i|$

Vektornormen

Beispiel

Sei $V = \mathbb{R}^n$, $x = (x_1, \ldots, x_n) \in \mathbb{R}^n$.
Für jedes p mit $1 \leq p \leq \infty$ wird eine Norm definiert durch

$$\|x\|_p = \left(\sum_{i=1}^{n} |x_i|^p \right)^{\frac{1}{p}}$$

Speziell:

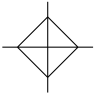
- 1-Norm: $\|x\|_1 = \sum_{i=1}^{n} |x_i|$
- ∞-Norm: $\|x\|_\infty = \max_{i=1,\ldots,n} |x_i|$
- 2-Norm: $\|x\|_2 = \left(\sum_{i=1}^{n} x_i^2 \right)^{\frac{1}{2}}$ (Euklidische Norm)

  Bemerkung: 2-Norm wird durch ein Skalarprodukt induziert.

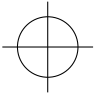
Vektornormen

Einheitskreise in $\mathbb{R}^2$: $\{x \in \mathbb{R}^2 : \|x\| = 1\}$

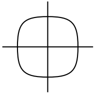
$$\|x\|_1 = \sum_{i=1}^{2} |x_i|$$



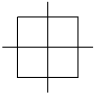
$$\|x\|_2 = \left(\sum_{i=1}^{2} |x_i|^2 \right)^{1/2} = \sqrt{x^T x}$$



$$\|x\|_p = \left(\sum_{i=1}^{2} |x_i|^p \right)^{1/p} \qquad (1 \leq p < \infty)$$



$$\|x\|_\infty = \max_{1 \leq i \leq 2} |x_i|$$



Weitere Beispiele

“Endlich-dimensionaler Vektorraum” beinhaltet nicht nur $\mathbb{R}^n$:

Beispiel

Die Menge

$$\Pi_m := \left\{ \sum_{i=0}^{m} a_i \, t^i \,\middle|\, a_i \in \mathbb{R} \right\}$$

ist ein $\mathbb{R}$-Vektorraum der Dimension $m + 1$.

Die Monome $M_i(t) := t^i,\ i = 0, \ldots, m$, dienen als Basis.

Weitere Beispiele

"Endlich-dimensionaler Vektorraum" beinhaltet nicht nur $\mathbb{R}^n$:

Beispiel

Die Menge

$$\Pi_m := \left\{ \sum_{i=0}^{m} a_i\, t^i \,\middle|\, a_i \in \mathbb{R} \right\}$$

ist ein $\mathbb{R}$-Vektorraum der Dimension $m+1$.

Die Monome $M_i(t) := t^i,\ i = 0, \ldots, m$, dienen als Basis.

Unendlich-dimensionaler Vektorraum: $V = C^0(I),\ I = [a,b] \subset \mathbb{R}$, mit Norm

$$\|f\|_p = \left(\int_a^b |f(x)|^p dx \right)^{\frac{1}{p}}$$

Satz 2.10

Auf einem endlich-dimensionalen Vektorraum V sind alle Normen äquivalent, d.h. zu je zwei Normen $\|\cdot\|_*$ und $\|\cdot\|_{**}$ existieren beschränkte, positive Konstanten c und C, so dass

$$c\|v\|_* \leq \|v\|_{**} \leq C\|v\|_*, \quad \text{für alle } v \in V$$

Satz 2.10

Auf einem endlich-dimensionalen Vektorraum V sind alle Normen äquivalent, d.h. zu je zwei Normen $\|\cdot\|_*$ und $\|\cdot\|_{**}$ existieren beschränkte, positive Konstanten c und C, so dass

$$c\|v\|_* \leq \|v\|_{**} \leq C\|v\|_*, \quad \text{für alle } v \in V$$

Beispiel

Sei $V = \mathbb{R}^n$, dann gilt

$$\|x\|_\infty \leq \|x\|_2 \leq \|x\|_1$$

und

$$\|x\|_2 \leq \sqrt{n}\|x\|_\infty$$

$$\|x\|_1 \leq \sqrt{n}\|x\|_2$$

$$\|x\|_1 \leq n\|x\|_\infty$$

Lineare Abbildungen und Operatornormen

X und Y: lineare normierte Räume (über $\mathbb{R}$) mit Normen $\|\cdot\|_X$, $\|\cdot\|_Y$.
Eine Abbildung $\mathcal{L} : X \to Y$ heißt linear, falls für $x, z \in X$ und $\alpha, \beta \in \mathbb{R}$ gilt

$$\mathcal{L}(\alpha x + \beta z) = \alpha \mathcal{L}(x) + \beta \mathcal{L}(z)$$

Beispiel

Eine Matrix $B \in \mathbb{R}^{m \times n}$

$$B = (b_{i,j})_{i,j=1}^{m,n} = \begin{pmatrix} b_{1,1} & \cdots & b_{1,n} \\ \vdots & & \vdots \\ b_{m,1} & \cdots & b_{m,n} \end{pmatrix}, \quad b_{i,j} \in \mathbb{R},$$

entspricht der Abbildung $B : \mathbb{R}^n \to \mathbb{R}^m$, $Bx = \sum_{j=1}^{n} x_j b_j$, mit b_j die j-te Spalte der Matrix B.

Lineare Abbildungen und Operatornormen

Beispiel

$X = \mathbb{R}^{m+1}$, $Y = \Pi_m$ mit Basis $\{\phi_0, \ldots, \phi_m\}$.

$$\mathcal{L}(a) := \sum_{i=0}^{m} a_i \phi_i, \quad a \in \mathbb{R}^{m+1}$$

Lineare Abbildungen und Operatornormen

Beispiel

$X = \mathbb{R}^{m+1}$, $Y = \Pi_m$ mit Basis $\{\phi_0, \ldots, \phi_m\}$.

$$\mathcal{L}(a) := \sum_{i=0}^{m} a_i \phi_i, \quad a \in \mathbb{R}^{m+1}$$

Operatornorm

Sei $\mathcal{L} : X \to Y$ eine lineare Abbildung.

$$\|\mathcal{L}\|_{X \to Y} := \sup_{x \neq 0} \frac{\|\mathcal{L}(x)\|_Y}{\|x\|_X}$$

Wichtige Eigenschaft:

$$\|\mathcal{L}(x)\|_Y \leq \|\mathcal{L}\|_{X \to Y} \|x\|_X, \quad \text{für all} \quad x \in X$$

Matrixnormen

Lineare Abbildung in Form einer Matrix, dann Matrixnorm:=Operatornorm.

Definition

Sei $A \in \mathbb{R}^{n\times n}$ und $\|\cdot\|$ eine Norm auf $\mathbb{R}^n$, dann ist durch

$$\|A\| = \sup_{x\neq 0} \frac{\|Ax\|}{\|x\|} = \sup_{\|x\|=1} \|Ax\|$$

eine dazugehörige Matrixnorm definiert.

Beachte

Definition gilt entsprechend auch für $A \in \mathbb{R}^{m\times n}$.

Die induzierte Matrixnorm $\|A\|$ ist die kleinste Zahl c, so dass gilt

$$\|Ax\| \leq c\|x\| \quad \text{für alle } x \in \mathbb{R}^n.$$

Matrixnormen

Lineare Abbildung in Form einer Matrix, dann Matrixnorm:=Operatornorm.

Definition

Sei $A \in \mathbb{R}^{n \times n}$ und $\|\cdot\|$ eine Norm auf $\mathbb{R}^n$, dann ist durch

$$\|A\| = \sup_{x \neq 0} \frac{\|Ax\|}{\|x\|} = \sup_{\|x\|=1} \|Ax\|$$

eine dazugehörige Matrixnorm definiert.

Es gilt:

- $\|A\| \geq 0$, und $\|A\| = 0$ nur wenn $A = 0$.
- Für alle $\alpha \in \mathbb{R}$: $\|\alpha\, A\| = |\alpha|\, \|A\|$
- Dreiecksungleichung:

$$\|A + B\| \leq \|A\| + \|B\|$$

Matrixnormen

Lineare Abbildung in Form einer Matrix, dann Matrixnorm:=Operatornorm.

Definition

Sei $A \in \mathbb{R}^{n \times n}$ und $\|\cdot\|$ eine Norm auf $\mathbb{R}^n$, dann ist durch

$$\|A\| = \sup_{x \neq 0} \frac{\|Ax\|}{\|x\|} = \sup_{\|x\|=1} \|Ax\|$$

eine dazugehörige Matrixnorm definiert.

und:

- $\|Ax\| \leq \|A\| \, \|x\|$
- $\|AB\| \leq \|A\| \, \|B\|$
- $\|I\| = 1$

Matrixnormen

Aus der Definition ergeben sich folgende Formeln für $A \in \mathbb{R}^{m \times n}$:

- 1-Norm: (max. Spaltensumme)

$$\|A\|_1 = \max_{j=1,\dots,n} \sum_{i=1}^{m} |a_{ij}|$$

Matrixnormen

Aus der Definition ergeben sich folgende Formeln für $A \in \mathbb{R}^{m \times n}$:

- 1-Norm: (max. Spaltensumme)

$$\|A\|_1 = \max_{j=1,\dots,n} \sum_{i=1}^{m} |a_{ij}|$$

- ∞-Norm: (max. Zeilensumme)

$$\|A\|_\infty = \max_{i=1,\dots,m} \sum_{j=1}^{n} |a_{ij}|$$

Matrixnormen

Aus der Definition ergeben sich folgende Formeln für $A \in \mathbb{R}^{m \times n}$:

- 1-Norm: (max. Spaltensumme)

$$\|A\|_1 = \max_{j=1,\ldots,n} \sum_{i=1}^{m} |a_{ij}|$$

- ∞-Norm: (max. Zeilensumme)

$$\|A\|_\infty = \max_{i=1,\ldots,m} \sum_{j=1}^{n} |a_{ij}|$$

- 2-Norm: (Spektralnorm)

$$\|A\|_2 = \sqrt{\lambda_{\max}(A^T A)}$$

wobei $\lambda_{\max}(A^T A)$ der größte Eigenwert von $A^T A$ ist.

Matrixnormen

Beispiel

Für $\boldsymbol{A} = \begin{pmatrix} \mathbf{2} & \mathbf{-3} \\ \mathbf{1} & \mathbf{1} \end{pmatrix}$ ergibt sich:

Matrixnormen

Beispiel

Für $\boldsymbol{A} = \begin{pmatrix} \mathbf{2} & \mathbf{-3} \\ \mathbf{1} & \mathbf{1} \end{pmatrix}$ ergibt sich:

$$\|\boldsymbol{A}\|_1 = 4, \qquad \|\boldsymbol{A}\|_\infty = \mathbf{5}, \qquad \|\boldsymbol{A}\|_2 = \sqrt{\tfrac{1}{2}\left(15 + 5\sqrt{5}\right)},$$

Matrixnormen

Beispiel

Für $\boldsymbol{A} = \begin{pmatrix} 2 & -3 \\ 1 & 1 \end{pmatrix}$ ergibt sich:

$$\|A\|_1 = 4, \qquad \|A\|_\infty = 5, \qquad \|A\|_2 = \sqrt{\frac{1}{2}\left(15 + 5\sqrt{5}\right)},$$

denn die Eigenwerte von $A^T A = \begin{pmatrix} 5 & -5 \\ -5 & 10 \end{pmatrix}$ kann man über

$$\det(A^T A - \lambda I) = 0 \iff (5-\lambda)(10-\lambda) - 25 = 0$$

bestimmen und damit

$$\lambda_1 = \frac{1}{2}\left(15 - 5\sqrt{5}\right), \qquad \lambda_2 = \frac{1}{2}\left(15 + 5\sqrt{5}\right).$$

Landau-Symbol

Landau-Symbol $\mathcal{O}$

Betrachte zwei Funktionen $g, h : \mathbb{R} \to \mathbb{R}$. Wir verwenden die Notation

$$g(x) = \mathcal{O}\left(h(x)\right) \qquad (x \to x_0)$$

wenn es Konstanten $C > 0$ und $\delta > 0$ gibt, so dass gilt:

$$|g(x)| \leq C|h(x)|, \qquad \forall x \text{ mit } |x - x_0| < \delta$$

- Anschauliche Bedeutung
  g wächst nicht wesentlich schneller als h (in einer Umgebung von x_0)
- Definition gilt entsprechend auch für $g, h : \mathbb{R}^n \to \mathbb{R}^m$.

Taylor-Entwicklung: Skalare Funktionen

Taylor-Entwicklung (von f um x_0)

Für hinreichend oft differenzierbares $f : \mathbb{R} \to \mathbb{R}$ gilt

$$\begin{aligned} f(x) \;&=\; f(x_0) + f'(x_0)\,(x - x_0) + \frac{f^{(2)}(x_0)}{2}(x - x_0)^2 + \dots \\ &\quad + \frac{f^{(k-1)}(x_0)}{(k-1)!}(x - x_0)^{k-1} + \frac{f^{(k)}(\xi)}{k!}(x - x_0)^k, \end{aligned}$$

wobei ξ eine Zahl zwischen x und x_0 ist.

$f^{(n)}(x_0)$ ist die n-te Ableitung von f an der Stelle x_0.

Taylor-Entwicklung: Skalare Funktionen

Taylor-Polynom vom Grad $k-1$ in x_0

$$\begin{aligned} p_{k-1}(x) &= f(x_0) + f'(x_0)\,(x-x_0) + \frac{f^{(2)}(x_0)}{2}(x-x_0)^2 \\ &\quad + \ldots + \frac{f^{(k-1)}(x_0)}{(k-1)!}(x-x_0)^{k-1}. \end{aligned}$$

Taylor-Entwicklung: Skalare Funktionen

Taylor-Polynom vom Grad $k-1$ in x_0

$$p_{k-1}(x) = f(x_0) + f'(x_0)(x-x_0) + \frac{f^{(2)}(x_0)}{2}(x-x_0)^2 + \ldots + \frac{f^{(k-1)}(x_0)}{(k-1)!}(x-x_0)^{k-1}.$$

- Für $k=1$ erhält man den Mittelwertsatz
$$\frac{f(x)-f(x_0)}{x-x_0} = f'(\xi),$$
wobei ξ eine Zahl zwischen x und x_0 ist.
- Oft verwendete Darstellung
$$f(x) = p_{k-1}(x) + \mathcal{O}(|x-x_0|^k) \qquad (x \to x_0)$$

Taylor-Entwicklung: Skalare Funktionen

Taylor-Polynom vom Grad $k-1$ in x_0

$$p_{k-1}(x) = f(x_0) + f'(x_0)(x-x_0) + \frac{f^{(2)}(x_0)}{2}(x-x_0)^2 + \ldots + \frac{f^{(k-1)}(x_0)}{(k-1)!}(x-x_0)^{k-1}.$$

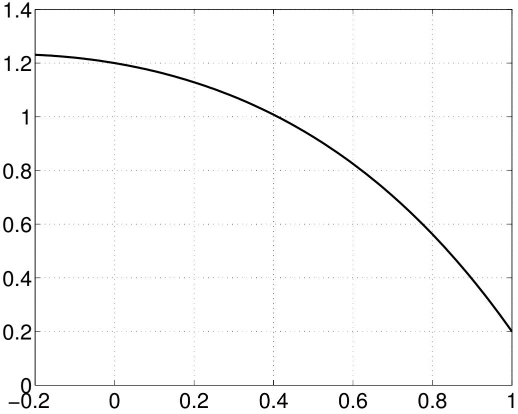
- Für $k = 1$ erhält man den Mittelwertsatz
  $$\frac{f(x) - f(x_0)}{x - x_0} = f'(\xi),$$
  wobei ξ eine Zahl zwischen x und x_0 ist.
- Oft verwendete Darstellung
  $$f(x) = p_{k-1}(x) + \mathcal{O}(|x-x_0|^k) \qquad (x \to x_0)$$

Siehe auch Matlab-Demo 2.23.

Taylor-Entwicklung: Skalare Funktionen

Taylor-Reihenentwicklung 0. , 1. und 2. Ordnung um $x_0 = 0$ von

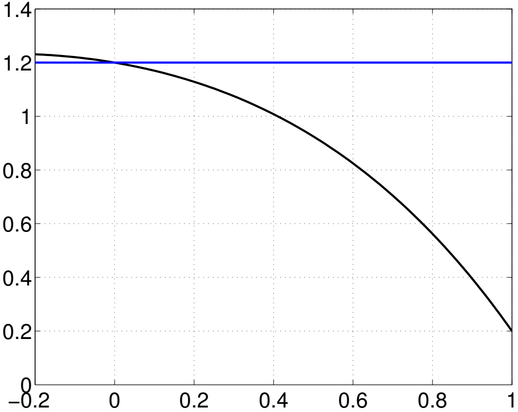
$$f(x) = -0.1x^4 - 0.15x^3 - 0.5x^2 - 0.25x + 1.2$$



Taylor-Entwicklung: Skalare Funktionen

Taylor-Reihenentwicklung 0. , 1. und 2. Ordnung um $\boldsymbol{x_0 = 0}$ von

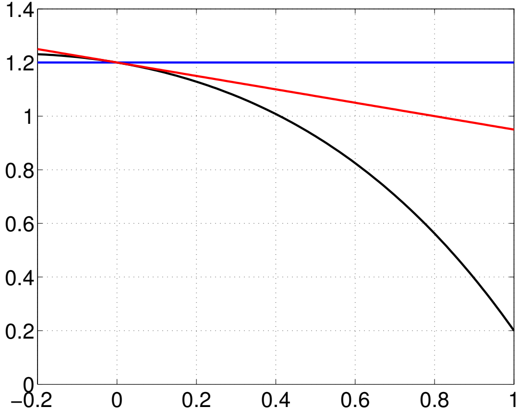
$$\boldsymbol{f(x) = -0.1x^4 - 0.15x^3 - 0.5x^2 - 0.25x + 1.2}$$



Taylor-Entwicklung: Skalare Funktionen

Taylor-Reihenentwicklung 0. , 1. und 2. Ordnung um $x_0 = 0$ von

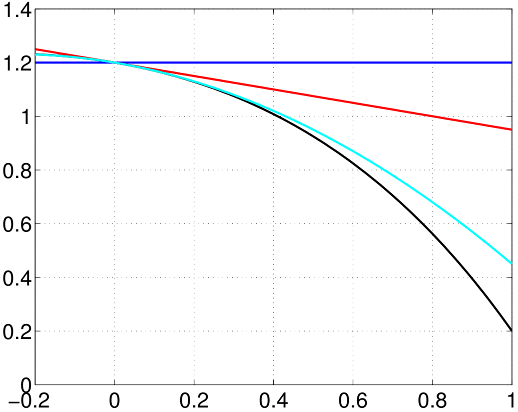
$$f(x) = -0.1x^4 - 0.15x^3 - 0.5x^2 - 0.25x + 1.2$$



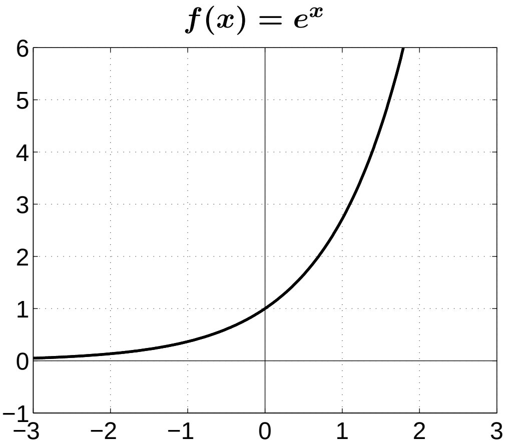
Taylor-Entwicklung: Skalare Funktionen

Taylor-Reihenentwicklung 0. , 1. und 2. Ordnung um $x_0 = 0$ von

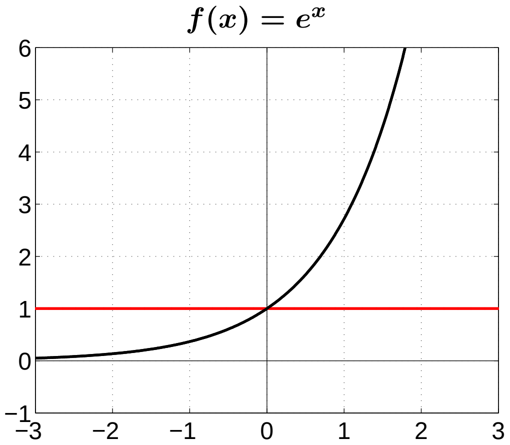
$$f(x) = -0.1x^4 - 0.15x^3 - 0.5x^2 - 0.25x + 1.2$$



Taylor-Entwicklung: Skalare Funktionen

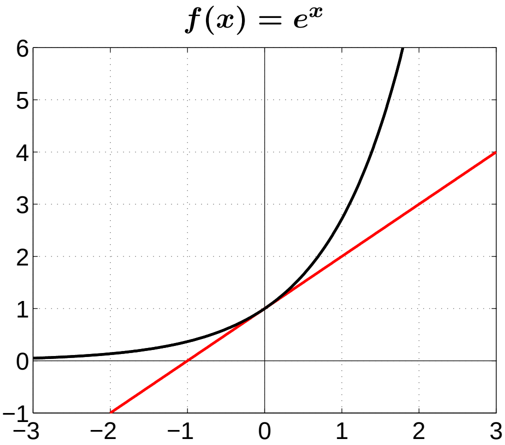


Taylor-Entwicklung: Skalare Funktionen



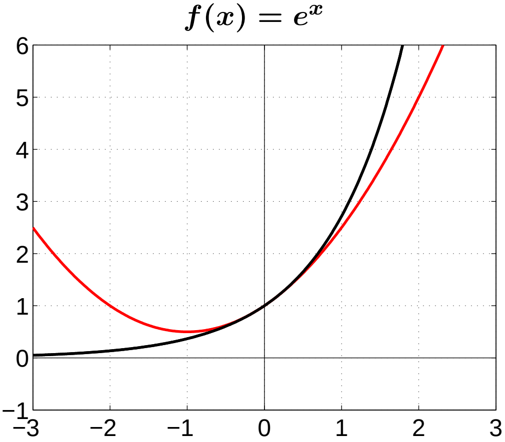
Taylor-Pol. 0. Grades in 0: $p_0(x) = 1$

Taylor-Entwicklung: Skalare Funktionen



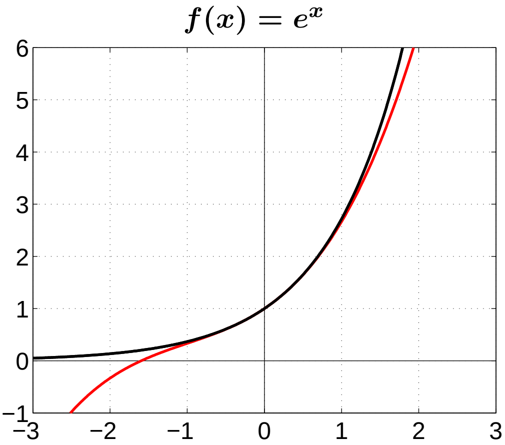
Taylor-Pol. 1. Grades in 0: $p_1(x) = 1 + x$

Taylor-Entwicklung: Skalare Funktionen



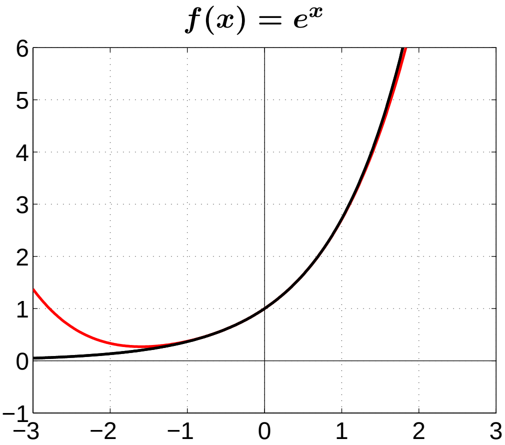
Taylor-Pol. 2. Grades in 0: $p_2(x) = 1 + x + \frac{x^2}{2!}$

Taylor-Entwicklung: Skalare Funktionen



Taylor-Pol. 3. Grades in 0: $p_3(x) = 1 + x + \frac{x^2}{2!} + \frac{x^3}{3!}$

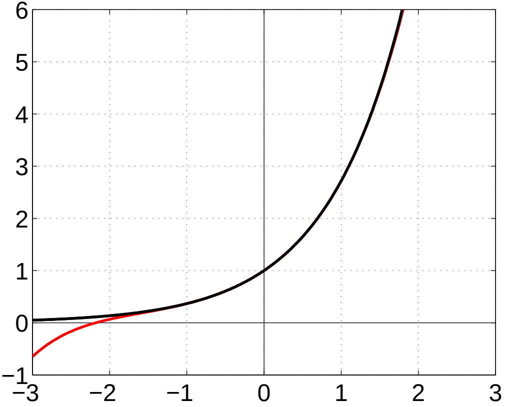
Taylor-Entwicklung: Skalare Funktionen



Taylor-Pol. 4. Grades in 0: $p_4(x) = 1 + x + \frac{x^2}{2!} + \frac{x^3}{3!} + \frac{x^4}{4!}$

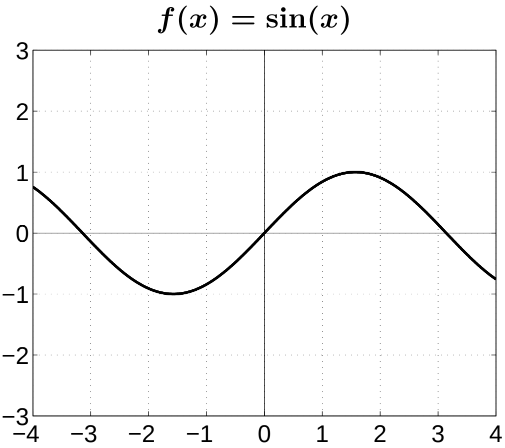
Taylor-Entwicklung: Skalare Funktionen

$$f(x) = e^x$$

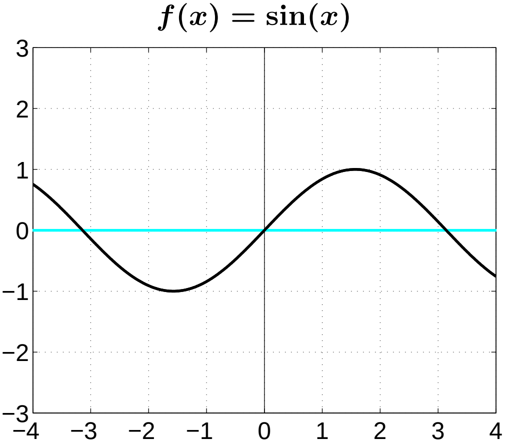


Taylor-Pol. 5. Grades in 0: $p_5(x) = 1 + x + \frac{x^2}{2!} + \frac{x^3}{3!} + \frac{x^4}{4!} + \frac{x^5}{5!}$

Taylor-Entwicklung: Skalare Funktionen

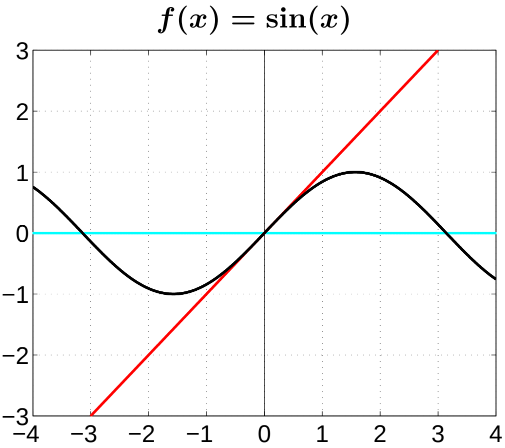


Taylor-Entwicklung: Skalare Funktionen



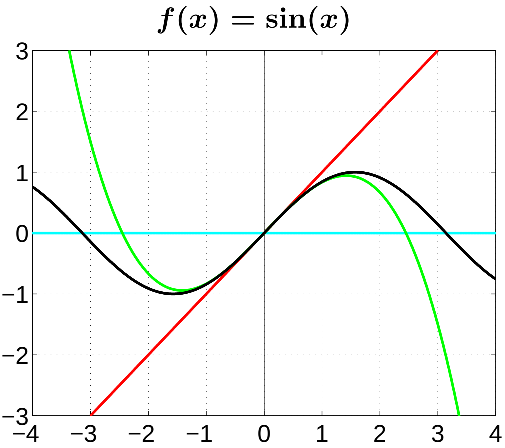
Taylor-Polynom 0. Grades in 0: $p_0(x) = 0$

Taylor-Entwicklung: Skalare Funktionen



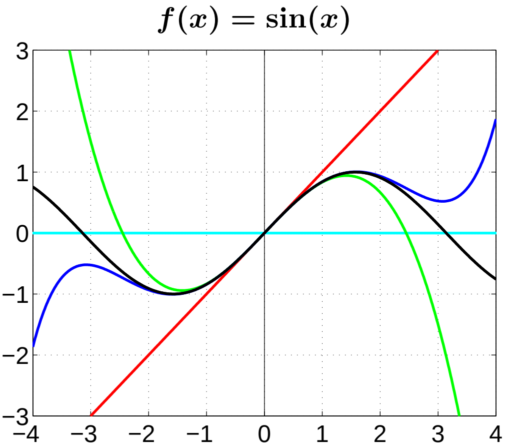
Taylor-Polynom 1. Grades in 0: $p_1(x) = x$

Taylor-Entwicklung: Skalare Funktionen



Taylor-Polynom 3. Grades in 0: $p_3(x) = x - \frac{x^3}{3!}$

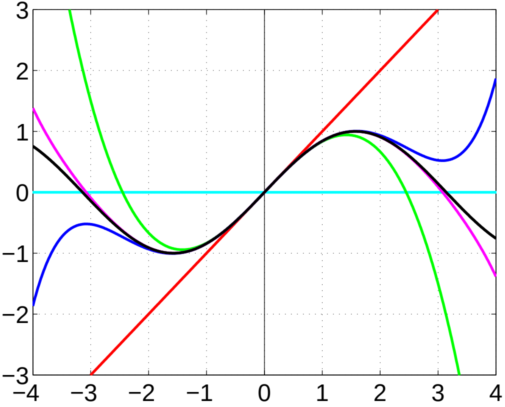
Taylor-Entwicklung: Skalare Funktionen



Taylor-Polynom 5. Grades in 0: $p_5(x) = x - \frac{x^3}{3!} + \frac{x^5}{5!}$

Taylor-Entwicklung: Skalare Funktionen

$$f(x) = \sin(x)$$



Taylor-Polynom 7. Grades in 0: $p_7(x) = x - \frac{x^3}{3!} + \frac{x^5}{5!} - \frac{x^7}{7!}$

Taylor-Entwicklung: Vektorwertige Funktionen

Taylor-Entwicklung (von f um x)

Für hinreichend oft differenzierbares $f : \mathbb{R}^n \to \mathbb{R}$ gilt

$$\begin{aligned} f(\tilde{x}) &= f(x) + \sum_{j=1}^{n} \frac{\partial f(x)}{\partial x_j} (\tilde{x}_j - x_j) \\ &+ \sum_{i,j=1}^{n} \frac{1}{2} \frac{\partial^2 f(x)}{\partial x_i \partial x_j} (\tilde{x}_i - x_i)(\tilde{x}_j - x_j) + \mathcal{O}(\|\tilde{x} - x\|_2^3). \end{aligned}$$

Taylor-Entwicklung: Vektorwertige Funktionen

Taylor-Entwicklung (von f um x)

Für hinreichend oft differenzierbares $f : \mathbb{R}^n \to \mathbb{R}$ gilt

$$\begin{aligned} f(\tilde{x}) &= f(x) + \sum_{j=1}^{n} \frac{\partial f(x)}{\partial x_j} (\tilde{x}_j - x_j) \\ &+ \sum_{i,j=1}^{n} \frac{1}{2} \frac{\partial^2 f(x)}{\partial x_i \partial x_j} (\tilde{x}_i - x_i)(\tilde{x}_j - x_j) + \mathcal{O}(\|\tilde{x} - x\|_2^3). \end{aligned}$$

- Gradient: $\nabla f(x) = \left(\frac{\partial f(x)}{\partial x_1}, \ldots, \frac{\partial f(x)}{\partial x_n} \right)^T$
- Hesse-Matrix: $f''(x) = \left(\frac{\partial^2 f(x)}{\partial x_i \partial x_j} \right)_{i,j=1}^{n}$

Taylor-Entwicklung: Vektorwertige Funktionen

Kompakte Schreibweise

$$\begin{aligned} f(\tilde{x}) &= f(x) + (\nabla f(x))^T (\tilde{x} - x) \\ &\quad + \frac{1}{2}(\tilde{x} - x)^T f''(x)(\tilde{x} - x) + \mathcal{O}(\|\tilde{x} - x\|_2^3). \end{aligned}$$

oder

$$f(\tilde{x}) = f(x) + (\nabla f(x))^T (\tilde{x} - x) + \mathcal{O}(\|\tilde{x} - x\|_2^2).$$

Falls $\|\tilde{x} - x\| \ll 1$:

$$f(\tilde{x}) \doteq f(x) + (\nabla f(x))^T (\tilde{x} - x)$$

$\doteq$: Terme höherer Ordnung werden vernachlässigt.

Fehlerquellen

Fehler im Resultat auf Grund von

- Datenfehlern (oder Eingabefehlern)

  $\Rightarrow$ Kondition eines Problems

  – können häufig nicht vermieden werden

Fehlerquellen

Fehler im Resultat auf Grund von

- Datenfehlern (oder Eingabefehlern)

  $\Rightarrow$ Kondition eines Problems

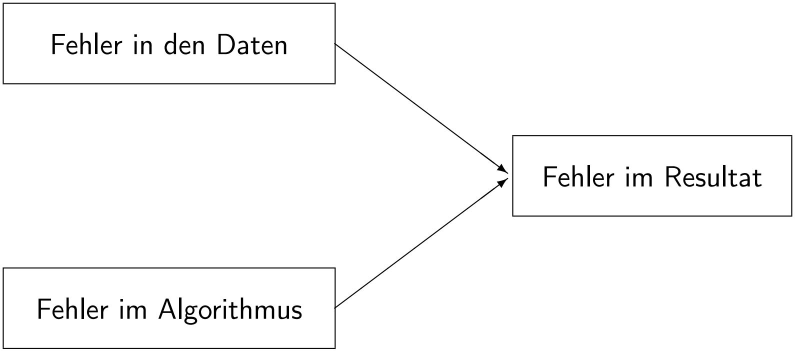
  – können häufig nicht vermieden werden

- Fehler(akkumulation) im Algorithmus (z.B. Rundungsfehler)

  $\Rightarrow$ Stabilität eines Algorithmus

  – kann man beeinflussen durch Anpassung des Verfahrens

Fehleranalyse: Kondition, Rundungsfehler, Stabilität



Kondition

Ziel: Analyse der Fehlerverstärkung bei Datenfehlern

Konzept der Kondition eines Problems

Kondition

Ziel: Analyse der Fehlerverstärkung bei Datenfehlern

Konzept der Kondition eines Problems

Beachte: Kondition

- ist unabhängig von einem speziellen Lösungweg (Algorithmus)
- gibt an, welche Genauigkeit man bestenfalls (bei exakter Rechnung) bei gestörten Eingangsdaten erwarten kann.

Kondition

Ziel: Analyse der Fehlerverstärkung bei Datenfehlern

Konzept der Kondition eines Problems

Beachte: Kondition

- ist unabhängig von einem speziellen Lösungweg (Algorithmus)
- gibt an, welche Genauigkeit man bestenfalls (bei exakter Rechnung) bei gestörten Eingangsdaten erwarten kann.

Wir fassen den "mathematischen Prozess" oder das "Problem" als Aufgabe auf, eine gegebene Funktion

$$f : X \to Y$$

an einer Stelle $x \in X$ auszuwerten.

Elementare Beispiele

- Die Berechnung der Multiplikation von x_1 und x_2:

$$f(x_1, x_2) = x_1\,x_2$$

und $X = \mathbb{R}^2$, $Y = \mathbb{R}$.

Elementare Beispiele

- Die Berechnung der Multiplikation von $\boldsymbol{x_1}$ und $\boldsymbol{x_2}$:

$$\boldsymbol{f(x_1, x_2) = x_1\, x_2}$$

  und $\boldsymbol{X} = \mathbb{R}^{\mathbf{2}}$, $\boldsymbol{Y} = \mathbb{R}$.

- Die Berechnung der Summe von $\boldsymbol{x_1}$ und $\boldsymbol{x_2}$:

$$\boldsymbol{f(x_1, x_2) = x_1 + x_2}$$

  und $\boldsymbol{X} = \mathbb{R}^{\mathbf{2}}$, $\boldsymbol{Y} = \mathbb{R}$.

Elementare Beispiele

- Man bestimme die kleinere Nullstelle der Gleichung

$$y^2 - 2\,x_1\,y + x_2 = 0,$$

mit $x_1^2 > x_2$. Die Lösung y^* ist

$$y^* = f(x_1, x_2) = x_1 - \sqrt{x_1^2 - x_2}.$$

mit

$$X = \left\{(x_1, x_2) \in \mathbb{R}^2 | x_1^2 > x_2\right\},$$

$$Y = \mathbb{R}.$$

Elementare Beispiele

- Bestimmung des Schnittpunktes zweier Geraden:

$$\begin{aligned} G_1 &= \{(y_1, y_2) \in \mathbb{R}^2 \mid a_{1,1}\, y_1 + a_{1,2}\, y_2 = x_1\} \\ G_2 &= \{(y_1, y_2) \in \mathbb{R}^2 \mid a_{2,1}\, y_1 + a_{2,2}\, y_2 = x_2\} \end{aligned}$$

wobei $(x_1, x_2)^T \in \mathbb{R}^2$ und $a_{i,j}$ gegeben seien.

Mit

$$A = \begin{pmatrix} a_{1,1} & a_{1,2} \\ a_{2,1} & a_{2,2} \end{pmatrix}$$

ergibt sich

$$Ay = x.$$

Elementare Beispiele

- Bestimmung des Schnittpunktes zweier Geraden: (Fortsetzung)

  Es gilt also

$$A \cdot y = \begin{pmatrix} a_{1,1} & a_{1,2} \\ a_{2,1} & a_{2,2} \end{pmatrix} \cdot \begin{pmatrix} y_1 \\ y_2 \end{pmatrix} = \begin{pmatrix} x_1 \\ x_2 \end{pmatrix} = x$$

  Annahme: $\det A \neq 0$
  Dann ist y durch

$$y = A^{-1} x$$

  gegeben. Also Auswertung der Funktion

$$f(x) = A^{-1} x,$$

  d.h. $X = Y = \mathbb{R}^2$.

Begriff der Kondition

Ungestörtes Problem

$$\underbrace{\boldsymbol{x}}_{\text{Eingabedaten}} \in \boldsymbol{X} \xrightarrow[\boldsymbol{f}]{\boxed{\text{Problem, Prozess}}} \underbrace{\boldsymbol{y} = \boldsymbol{f}(\boldsymbol{x})}_{\text{Ausgabedaten}} \in \boldsymbol{Y}$$

Begriff der Kondition

Ungestörtes Problem

$$\underbrace{x}_{\text{Eingabedaten}} \in X \xrightarrow[f]{\boxed{\text{Problem, Prozess}}} \underbrace{y = f(x)}_{\text{Ausgabedaten}} \in Y$$

Gestörtes Problem

$$\tilde{x} = x + \Delta x \xrightarrow[f]{\boxed{\text{Problem, Prozess}}} \tilde{y} = f(\tilde{x})$$

mit Eingabefehler $\Delta x = \tilde{x} - x$

Ausgabefehler $\Delta y = \tilde{y} - y = f(\tilde{x}) - f(x)$

Begriff der Kondition

Ungestörtes Problem

$$\underbrace{x}_{\text{Eingabedaten}} \in X \xrightarrow[f]{\boxed{\text{Problem, Prozess}}} \underbrace{y = f(x)}_{\text{Ausgabedaten}} \in Y$$

Gestörtes Problem

$$\tilde{x} = x + \Delta x \xrightarrow[f]{\boxed{\text{Problem, Prozess}}} \tilde{y} = f(\tilde{x})$$

mit Eingabefehler $\Delta x = \tilde{x} - x$

Ausgabefehler $\Delta y = \tilde{y} - y = f(\tilde{x}) - f(x)$

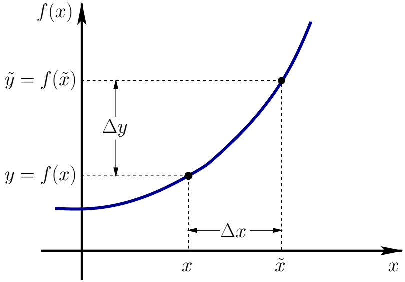
Ziel: Verhältnis Ausgabefehler Δy zu Eingabefehler Δx.

Begriff der Kondition

Arten von Fehlern

- absoluter Eingabefehler: $\|\boldsymbol{\Delta x}\|_{\boldsymbol{X}}$
- absoluter Ausgabefehler: $\|\boldsymbol{\Delta y}\|_{\boldsymbol{Y}}$
- relativer Eingabefehler: $\boldsymbol{\delta_x} = \dfrac{\|\boldsymbol{\Delta x}\|_{\boldsymbol{X}}}{\|\boldsymbol{x}\|_{\boldsymbol{X}}}$
- relativer Ausgabefehler: $\boldsymbol{\delta_y} = \dfrac{\|\boldsymbol{\Delta y}\|_{\boldsymbol{Y}}}{\|\boldsymbol{y}\|_{\boldsymbol{Y}}}$

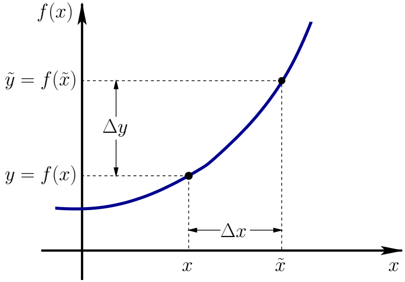
Begriff der Kondition: skalare Funktion



Allgemein:

- absoluter Eingabefehler: $\|\boldsymbol{\Delta x}\|_{\boldsymbol{X}}$
- absoluter Ausgabefehler: $\|\boldsymbol{\Delta y}\|_{\boldsymbol{Y}}$

Begriff der Kondition: skalare Funktion



Allgemein:

- relativer Eingabefehler: $\boldsymbol{\delta_x} = \dfrac{\|\boldsymbol{\Delta x}\|_{\boldsymbol{X}}}{\|\boldsymbol{x}\|_{\boldsymbol{X}}}$
- relativer Ausgabefehler: $\boldsymbol{\delta_y} = \dfrac{\|\boldsymbol{\Delta y}\|_{\boldsymbol{Y}}}{\|\boldsymbol{y}\|_{\boldsymbol{Y}}}$

Begriff der Kondition

Definition

Mit der relativen Kondition eines (durch f beschriebenen) Problems bezeichnet man das Verhältnis

$$\frac{\delta_y}{\delta_x}$$

des relativen Ausgabefehlers zum relativen Eingabefehler, d.h. die Sensitivität des Problems unter Störungen der Eingabedaten.

Begriff der Kondition

Definition

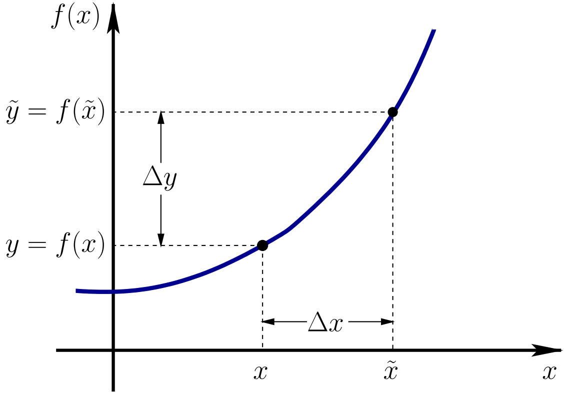
Mit der relativen Kondition eines (durch f beschriebenen) Problems bezeichnet man das Verhältnis

$$\frac{\delta_y}{\delta_x}$$

des relativen Ausgabefehlers zum relativen Eingabefehler, d.h. die Sensitivität des Problems unter Störungen der Eingabedaten.

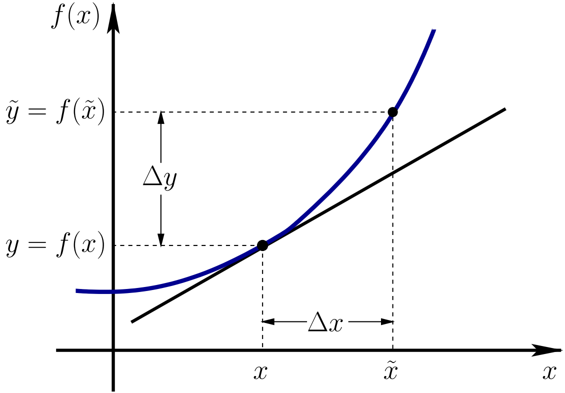
- Absolute Kondition: Verhältnis $\frac{\|\Delta y\|_Y}{\|\Delta x\|_X}$
- Mit Kondition wird meistens die relative Kondition gemeint.
- Ein Problem ist umso besser konditioniert, je kleinere Schranken für δ_y/δ_x (mit $\delta_x \to 0$) existieren.

Taylorentwicklung 1. Ordnung



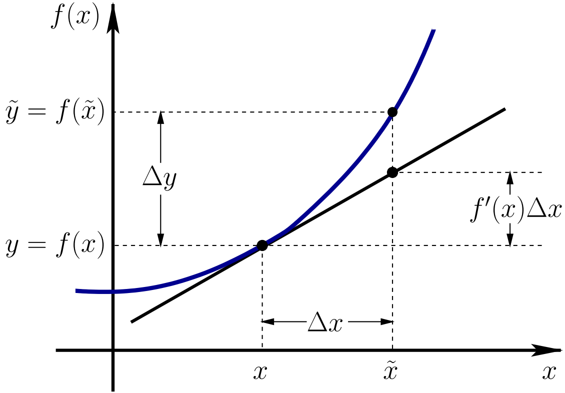
Relative / absolute Kondition: Verhältnis $\frac{\boldsymbol{\delta_y}}{\boldsymbol{\delta_x}}$ bzw. $\frac{\|\boldsymbol{\Delta y}\|_{\boldsymbol{Y}}}{\|\boldsymbol{\Delta x}\|_{\boldsymbol{X}}}$.

Taylorentwicklung 1. Ordnung



Relative / absolute Kondition: Verhältnis $\dfrac{\boldsymbol{\delta_y}}{\boldsymbol{\delta_x}}$ bzw. $\dfrac{\|\boldsymbol{\Delta y}\|_{\boldsymbol{Y}}}{\|\boldsymbol{\Delta x}\|_{\boldsymbol{X}}}$.

Taylorentwicklung 1. Ordnung



Relative / absolute Kondition: Verhältnis $\dfrac{\boldsymbol{\delta_y}}{\boldsymbol{\delta_x}}$ bzw. $\dfrac{\|\boldsymbol{\Delta y}\|_{\boldsymbol{Y}}}{\|\boldsymbol{\Delta x}\|_{\boldsymbol{X}}}$.

Kondition: $f : \mathbb{R} \to \mathbb{R}$

Taylor-Entwicklung 1. Ordnung von f um festes x

$$f(\tilde{x}) \doteq f(x) + f'(x)\,(\tilde{x} - x)$$

Daraus erhält man die Kondition für

- $f : \mathbb{R} \to \mathbb{R}$ (Eingabe: Skalar, Ausgabe: Skalar)

Kondition: $f : \mathbb{R} \to \mathbb{R}$

Taylor-Entwicklung 1. Ordnung von f um festes x

$$f(\tilde{x}) \doteq f(x) + f'(x)\,(\tilde{x} - x)$$

Daraus erhält man die Kondition für

- $f : \mathbb{R} \to \mathbb{R}$ (Eingabe: Skalar, Ausgabe: Skalar)

$$\left|\frac{f(\tilde{x}) - f(x)}{f(x)}\right| \;\doteq\; \kappa_{\mathrm{rel}}(x)\left|\frac{\tilde{x} - x}{x}\right|$$

mit

$$\kappa_{\mathrm{rel}}(x) := \left|f'(x)\frac{x}{f(x)}\right|$$

Kondition: $f : \mathbb{R}^n \to \mathbb{R}$

Für $f : \mathbb{R}^n \to \mathbb{R}$ lautet die Taylor-Reihenentwicklung 1. Ordnung

$$f(\tilde{x}) \doteq f(x) + (\nabla f(x))^T \cdot (\tilde{x} - x)$$

mit

$$x = \begin{pmatrix} x_1 \\ \vdots \\ x_n \end{pmatrix} \in \mathbb{R}^n.$$

Kondition: $f:\mathbb{R}^n \to \mathbb{R}$

Für $f:\mathbb{R}^n \to \mathbb{R}$ lautet die Taylor-Reihenentwicklung 1. Ordnung

$$f(\tilde{x}) \doteq f(x) + (\nabla f(x))^T \cdot (\tilde{x} - x)$$

mit

$$x = \begin{pmatrix} x_1 \\ \vdots \\ x_n \end{pmatrix} \in \mathbb{R}^n.$$

Hieraus folgt

$$\frac{f(\tilde{x}) - f(x)}{f(x)} \doteq \sum_{j=1}^{n} \left(\frac{\partial f(x)}{\partial x_j} \cdot \frac{x_j}{f(x)} \right) \cdot \frac{\tilde{x}_j - x_j}{x_j}$$

Kondition: $f : \mathbb{R}^n \to \mathbb{R}$

Mit den Verstärkungsfaktoren

$$\phi_j(x) = \frac{\partial f(x)}{\partial x_j} \cdot \frac{x_j}{f(x)}$$

erhält man

$$\underbrace{\frac{f(\tilde{x}) - f(x)}{f(x)}}_{\substack{\text{relativer Fehler} \\ \text{der Ausgabe}}} \doteq \sum_{j=1}^{n} \underbrace{\phi_j(x)}_{\substack{\text{Fehler-} \\ \text{verstärkung}}} \cdot \underbrace{\frac{\tilde{x}_j - x_j}{x_j}}_{\substack{\text{relativer Fehler} \\ \text{der Eingabe in } x_j}}$$

Kondition $f : \mathbb{R}^n \to \mathbb{R}$

Damit erhält man die Kondition für

- $f : \mathbb{R}^n \to \mathbb{R}$ (Eingabe: Vektor, Ausgabe: Skalar)

Kondition $f : \mathbb{R}^n \to \mathbb{R}$

Damit erhält man die Kondition für

- $f : \mathbb{R}^n \to \mathbb{R}$ (Eingabe: Vektor, Ausgabe: Skalar)

$$\left| \frac{f(\tilde{x}) - f(x)}{f(x)} \right| \dot{\leq} \kappa_{\text{rel}}(x) \sum_{j=1}^{n} \left| \frac{\tilde{x}_j - x_j}{x_j} \right|$$

mit

$$\kappa_{\text{rel}}(x) = \kappa_{\text{rel}}^{\infty}(x) := \max_j |\phi_j(x)|$$

und den Verstärkungsfaktoren

$$\phi_j(x) = \frac{\partial f(x)}{\partial x_j} \cdot \frac{x_j}{f(x)}$$

wobei $\dot{\leq}$ entsprechend $\doteq$ zu verstehen ist.

Beispiel 2.26.

Gegeben sei

$$f : \mathbb{R} \to \mathbb{R}, \quad f(x) = e^{3x^2}.$$

Relative Konditionszahl:

$$\kappa_{\mathrm{rel}}(x) = \left| f'(x) \frac{x}{f(x)} \right| = 6\,x^2.$$

$\rightsquigarrow$ für $|x|$ klein/groß ist f gut/schlecht konditioniert.

Beispiel 2.26.

Gegeben sei

$$f : \mathbb{R} \to \mathbb{R}, \quad f(x) = e^{3x^2}.$$

Relative Konditionszahl:

$$\kappa_{\mathrm{rel}}(x) = \left| f'(x) \frac{x}{f(x)} \right| = 6\,x^2.$$

$\rightsquigarrow$ für $|x|$ klein/groß ist f gut/schlecht konditioniert.

Beispiel

- $x = 0.1,\ \tilde{x} = 0.10001$: $\kappa_{\mathrm{rel}}(0.1) = 6 \cdot 10^{-2}$

$$\left| \frac{\tilde{x}-x}{x} \right| = 10^{-4} \to \left| \frac{f(x)-f(\tilde{x})}{f(x)} \right| = 6.03 \cdot 10^{-6}$$

Beispiel 2.26.

Gegeben sei

$$f : \mathbb{R} \to \mathbb{R}, \quad f(x) = e^{3x^2}.$$

Relative Konditionszahl:

$$\kappa_{\mathrm{rel}}(x) = \left| f'(x) \frac{x}{f(x)} \right| = 6\,x^2.$$

$\rightsquigarrow$ für $|x|$ klein/groß ist f gut/schlecht konditioniert.

Beispiel

- $x = 0.1,\ \tilde{x} = 0.10001$: $\kappa_{\mathrm{rel}}(0.1) = 6 \cdot 10^{-2}$

$$\left| \frac{\tilde{x} - x}{x} \right| = 10^{-4} \to \left| \frac{f(x) - f(\tilde{x})}{f(x)} \right| = 6.03 \cdot 10^{-6}$$

- $x = 4,\ \tilde{x} = 4.0004$: $\kappa_{\mathrm{rel}}(4) = 96$

$$\left| \frac{\tilde{x} - x}{x} \right| = 10^{-4} \to \left| \frac{f(x) - f(\tilde{x})}{f(x)} \right| = 9.65 \cdot 10^{-3}$$

Elementare Rechenoperationen

Kondition bei

- Multiplikation: $\boldsymbol{x = (x_1, x_2)^T}$, $\boldsymbol{f(x) = x_1\, x_2}$

Elementare Rechenoperationen

Kondition bei

- Multiplikation: $x = (x_1, x_2)^T$, $f(x) = x_1 \, x_2$

  $\kappa_{\mathrm{rel}}(x) = 1$ (von x unabhängig!)

  Multiplikation für alle Eingangsdaten gut konditioniert.
  Ein ähnliches Resultat gilt für die Division.

Elementare Rechenoperationen

Kondition bei

- Multiplikation: $\boldsymbol{x = (x_1, x_2)^T}$, $\boldsymbol{f(x) = x_1\, x_2}$

  $$\boldsymbol{\kappa}_{\mathrm{rel}}\boldsymbol{(x) = 1} \text{ (von } \boldsymbol{x} \text{ unabhängig!)}$$

  Multiplikation für alle Eingangsdaten gut konditioniert.
  Ein ähnliches Resultat gilt für die Division.

- Addition: $\boldsymbol{x = (x_1, x_2)^T}$, $\boldsymbol{f(x) = x_1 + x_2}$

Elementare Rechenoperationen

Kondition bei

- Multiplikation: $x = (x_1, x_2)^T$, $f(x) = x_1\, x_2$

  $$\kappa_{\rm rel}(x) = 1 \text{ (von } x \text{ unabhängig!)}$$

  Multiplikation für alle Eingangsdaten gut konditioniert.
  Ein ähnliches Resultat gilt für die Division.

- Addition: $x = (x_1, x_2)^T$, $f(x) = x_1 + x_2$

  $$\kappa_{\rm rel}(x) = \max\left\{ \left|\frac{x_1}{x_1+x_2}\right|, \left|\frac{x_2}{x_1+x_2}\right| \right\}$$

  Bei zwei Zahlen mit gleichem Vorzeichen: $\kappa_{\rm rel} \leq 1$.

  ABER: $\kappa_{\rm rel}(x) \gg 1$ wenn $x_1 \approx -x_2$.

Beispiel 2.29 (Nullstelle)

Bestimmung der kleineren Nullstelle y^* von $y^2 - 2x_1 y + x_2 = 0$:

$$x = (x_1, x_2)^T, \qquad y^* = f(x) = x_1 - \sqrt{x_1^2 - x_2}$$

- Partielle Ableitungen

$$\frac{\partial f(x)}{\partial x_1} = \frac{\sqrt{x_1^2 - x_2} - x_1}{\sqrt{x_1^2 - x_2}} = \frac{-y^*}{\sqrt{x_1^2 - x_2}}$$

$$\frac{\partial f(x)}{\partial x_2} = \frac{1}{2\sqrt{x_1^2 - x_2}}$$

- Verstärkungsfaktoren

$$\phi_j(x) = \frac{\partial f(x)}{\partial x_j} \cdot \frac{x_j}{f(x)}$$

Beispiel 2.29 (Nullstelle)

- Verstärkungsfaktoren

$$
\begin{aligned}
\phi_1(x) &= \frac{-y^*}{\sqrt{x_1^2 - x_2}} \cdot \frac{x_1}{y^*} = \frac{-x_1}{\sqrt{x_1^2 - x_2}} \\
\phi_2(x) &= \frac{1}{2\sqrt{x_1^2 - x_2}} \cdot \frac{x_2}{y^*} = \frac{1}{2} - \frac{1}{2}\phi_1(x)
\end{aligned}
$$

- Kondition: $\kappa_{\mathrm{rel}}(x) = \max_j |\phi_j(x)|$

Beispiel 2.29 (Nullstelle)

- Verstärkungsfaktoren

$$\phi_1(x) = \frac{-y^*}{\sqrt{x_1^2 - x_2}} \cdot \frac{x_1}{y^*} = \frac{-x_1}{\sqrt{x_1^2 - x_2}}$$

$$\phi_2(x) = \frac{1}{2\sqrt{x_1^2 - x_2}} \cdot \frac{x_2}{y^*} = \frac{1}{2} - \frac{1}{2}\phi_1(x)$$

- Kondition: $\kappa_{\mathrm{rel}}(x) = \max_j |\phi_j(x)|$

Kondition hängt stark von der Stelle (x_1, x_2) ab:

Beispiel 2.29 (Nullstelle)

- Verstärkungsfaktoren

$$\begin{aligned}\phi_1(x) &= \frac{-y^*}{\sqrt{x_1^2 - x_2}} \cdot \frac{x_1}{y^*} = \frac{-x_1}{\sqrt{x_1^2 - x_2}}\\ \phi_2(x) &= \frac{1}{2\sqrt{x_1^2 - x_2}} \cdot \frac{x_2}{y^*} = \frac{1}{2} - \frac{1}{2}\phi_1(x)\end{aligned}$$

- Kondition: $\kappa_{\mathrm{rel}}(x) = \max_j |\phi_j(x)|$

Kondition hängt stark von der Stelle (x_1, x_2) ab:

- Wenn $x_2 < 0$: $|\phi_1(x)| \leq 1$ und $\kappa_{\mathrm{rel}}(x) \leq 1$
- Wenn $x_2 \approx x_1^2$: $|\phi_1(x)| \gg 1$ und $\kappa_{\mathrm{rel}}(x) \gg 1$

Kondition: $f : \mathbb{R}^n \to \mathbb{R}^n$, f linear

Sei $f = B \in \mathbb{R}^{n \times n}$, mit $\det B \neq 0$, also

$$y = f(x) = Bx$$

bzw. für gestörte Daten

$$\tilde{y} = f(\tilde{x}) = B\tilde{x}$$

und damit

$$f(\tilde{x}) - f(x) = B\tilde{x} - Bx = B(\tilde{x} - x)$$

$$x = B^{-1}y$$

Kondition: $f : \mathbb{R}^n \to \mathbb{R}^n$, f linear

Wegen $\|x\| = \|B^{-1}y\| \leq \|B^{-1}\|\|y\|$ gilt

$$\frac{\|f(\tilde{x}) - f(x)\|}{\|f(x)\|} = \frac{\|B(\tilde{x} - x)\|}{\|y\|} \leq \underbrace{\|B\| \cdot \|B^{-1}\|}_{\kappa(B)} \frac{\|\tilde{x} - x\|}{\|x\|}$$

wobei

$$\kappa(B) \equiv \|B\| \cdot \|B^{-1}\|$$

die Konditionszahl der Matrix B ist.

Kondition: $f : \mathbb{R}^n \to \mathbb{R}^n$, f linear

Wegen $\|x\| = \|B^{-1}y\| \leq \|B^{-1}\|\|y\|$ gilt

$$\frac{\|f(\tilde{x}) - f(x)\|}{\|f(x)\|} = \frac{\|B(\tilde{x} - x)\|}{\|y\|} \leq \underbrace{\|B\| \cdot \|B^{-1}\|}_{\kappa(B)} \frac{\|\tilde{x} - x\|}{\|x\|}$$

wobei

$$\kappa(B) \equiv \|B\| \cdot \|B^{-1}\|$$

die Konditionszahl der Matrix B ist.

Beachte:
$\kappa(B) = \kappa(B^{-1})$ hängt nur von der Matrix B (und der Norm $\|\cdot\|$) ab.

Beispiel 2.34.

Die Bestimmung des Schnittpunkts der Geraden

$$\begin{aligned} 3\,u_1 + 1.001\,u_2 &= 1.999 \\ 6\,u_1 + 1.997\,u_2 &= 4.003. \end{aligned}$$

(fast parallel!) ergibt das Problem $u = A^{-1}b$ mit

$$A = \begin{pmatrix} 3 & 1.001 \\ 6 & 1.997 \end{pmatrix}, \qquad b = \begin{pmatrix} 1.999 \\ 4.003 \end{pmatrix}.$$

Die Lösung ist

$$u = \begin{pmatrix} 1 \\ \text{-}1 \end{pmatrix}.$$

Beispiel 2.34.

Die Bestimmung des Schnittpunkts der Geraden

$$\begin{aligned} 3\,u_1 + 1.001\,u_2 &= 1.999 \\ 6\,u_1 + 1.997\,u_2 &= 4.003. \end{aligned}$$

(fast parallel!) ergibt das Problem $u = A^{-1}b$ mit

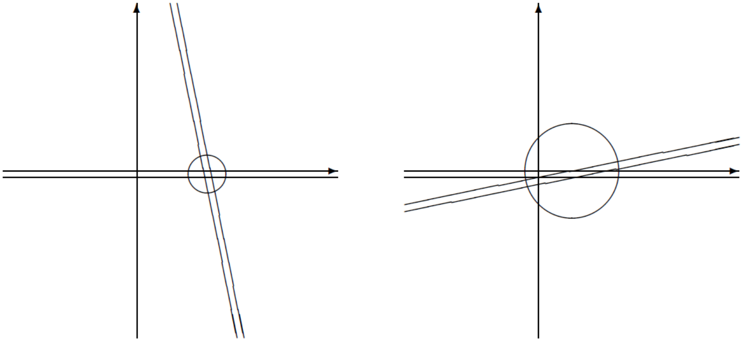
$$A = \begin{pmatrix} 3 & 1.001 \\ 6 & 1.997 \end{pmatrix}, \qquad b = \begin{pmatrix} 1.999 \\ 4.003 \end{pmatrix}.$$

Die Lösung ist

$$u = \begin{pmatrix} 1 \\ \text{-}1 \end{pmatrix}.$$

Also: $f : \mathbb{R}^2 \to \mathbb{R}^2$, $f(x) = A^{-1}x$.

Beispiel 2.34. Kondition bei Bestimmung eines Schnittpunktes



Beispiel 2.34.

Effekt einer Störung in $\boldsymbol{b}$:

$$\tilde{\boldsymbol{b}} = \begin{pmatrix} \mathbf{2.002} \\ \mathbf{4} \end{pmatrix}, \qquad \tilde{\boldsymbol{u}} = \boldsymbol{A}^{-1}\tilde{\boldsymbol{b}}.$$

Man erhält

$$\boldsymbol{A}^{-1} = \frac{-1}{0.015}\begin{pmatrix} 1.997 & -1.001 \\ -6 & 3 \end{pmatrix}, \qquad \tilde{\boldsymbol{u}} = \begin{pmatrix} 0.4004 \\ 0.8 \end{pmatrix}$$

Beispiel 2.34.

Effekt einer Störung in $\boldsymbol{b}$:

$$\tilde{\boldsymbol{b}} = \begin{pmatrix} \mathbf{2.002} \\ \mathbf{4} \end{pmatrix}, \qquad \tilde{\boldsymbol{u}} = \boldsymbol{A}^{-1}\tilde{\boldsymbol{b}}.$$

Man erhält

$$\boldsymbol{A}^{-1} = \frac{-1}{0.015}\begin{pmatrix} 1.997 & -1.001 \\ -6 & 3 \end{pmatrix}, \qquad \tilde{\boldsymbol{u}} = \begin{pmatrix} 0.4004 \\ 0.8 \end{pmatrix}$$

Wir betrachten die Maximumnorm:

$$\|\boldsymbol{x}\| = \|\boldsymbol{x}\|_\infty = \max_i |x_i|.$$

Beispiel 2.34.

Es gilt

- Störung der Daten

$$\frac{\|\tilde{b}-b\|_\infty}{\|b\|_\infty} = \frac{3\cdot 10^{-3}}{4.003} \approx 7.5\cdot 10^{-4}$$

- Änderung des Resultats

$$\frac{\|\tilde{u}-u\|_\infty}{\|u\|_\infty} = \frac{1.8}{1} \approx 1.8$$

Schlechte Kondition wird quantifiziert durch

$$\|A\|_\infty \, \|A^{-1}\|_\infty = 4798.2.$$

Kondition einer Basis

Sei V ein linearer normierter Raum mit Basis $\Phi = \{\phi_1, \ldots, \phi_n\}$. Die Koordinaten-Abbildung ist durch

$$\mathcal{L} : \mathbb{R}^n \to V, \qquad \mathcal{L}(a) = \sum_{i=1}^{n} a_i \phi_i,$$

gegeben.

Kondition

Es gilt

$$\min\left\{ C/|c| \;\middle|\; |c|\,\|a\| \le \Big\| \sum_{j=1}^{n} a_j \phi_j \Big\|_V \le C\|a\| \;\; \forall\, a \in \mathbb{R}^n \right\}$$
$$= \kappa(\mathcal{L}) = \|\mathcal{L}\|_{\mathbb{R}^n \to V} \|\mathcal{L}^{-1}\|_{V \to \mathbb{R}^n}$$

Kondition einer Basis

Gram-Matrix

Annahmen: $\|\cdot\|_V$ entspricht $(\cdot,\cdot)_V$, und $\|\cdot\| := \|\cdot\|_2$ auf $\mathbb{R}^n$.

$$\text{Gram-Matrix:} \quad G_{i,j} := (\phi_i, \phi_j)_V, \quad 1 \leq i, j \leq n,$$

definiert. Es gilt $\kappa(\mathcal{L}) = \sqrt{\kappa_2(G)}$.

Kondition einer Basis

Gram-Matrix

Annahmen: $\|\cdot\|_V$ entspricht $(\cdot,\cdot)_V$, und $\|\cdot\| := \|\cdot\|_2$ auf $\mathbb{R}^n$.

$$\text{Gram-Matrix:} \quad G_{i,j} := (\phi_i, \phi_j)_V, \quad 1 \le i, j \le n,$$

definiert. Es gilt $\kappa(\mathcal{L}) = \sqrt{\kappa_2(G)}$.

Beispiel

Sei $V = \Pi_m$, mit Dimension $n := m + 1$, Skalarprodukt $(f, g)_V := \int_0^1 f(t)g(t)\,dt$ und die Basis $\phi_i(t) = t^{i-1}$, $i = 1, \dots n$.

$$G_{i,j} = \int_0^1 \phi_i(t)\phi_j(t)\,dt = \int_0^1 t^{i+j-2}\,dt = \frac{1}{i+j-1}$$

Diese Matrix G wird Hilbert-Matrix genannt. Sie hat eine (sehr) große Konditionszahl, z.B. $3.75 \cdot 10^{16}$ für $n = 12$.

Zahlendarstellungen: Beispiel 2.40.

Wir betrachten als Beispiel die Zahl $\mathbf{123.75}$:

- Dezimalsystem (Basis 10)

$$\begin{aligned}
&\mathbf{123.75}\\
&= 1 \cdot 10^{2} + 2 \cdot 10^{1} + 3 \cdot 10^{0} + 7 \cdot 10^{-1} + 5 \cdot 10^{-2}\\
&= 10^{3}\,(1 \cdot 10^{-1} + 2 \cdot 10^{-2} + 3 \cdot 10^{-3} + 7 \cdot 10^{-4} + 5 \cdot 10^{-5})
\end{aligned}$$

Zahlendarstellungen: Beispiel 2.40.

Wir betrachten als Beispiel die Zahl $\mathbf{123.75}$:

- Dezimalsystem (Basis 10)

$$\begin{aligned}
&\mathbf{123.75}\\
&= 1 \cdot 10^{2} + 2 \cdot 10^{1} + 3 \cdot 10^{0} + 7 \cdot 10^{-1} + 5 \cdot 10^{-2}\\
&= 10^{3}\,(1 \cdot 10^{-1} + 2 \cdot 10^{-2} + 3 \cdot 10^{-3} + 7 \cdot 10^{-4} + 5 \cdot 10^{-5})
\end{aligned}$$

- Binärsystem (Basis 2)

$$\begin{aligned}
&\mathbf{123.75}\\
&= 1 \cdot 2^{6} + 1 \cdot 2^{5} + 1 \cdot 2^{4} + 1 \cdot 2^{3} + 0 \cdot 2^{2} + 1 \cdot 2^{1} + 1 \cdot 2^{0}\\
&\quad + 1 \cdot 2^{-1} + 1 \cdot 2^{-2}\\
&= 2^{7}\,(1 \cdot 2^{-1} + 1 \cdot 2^{-2} + 1 \cdot 2^{-3} + 1 \cdot 2^{-4} + 0 \cdot 2^{-5} + 1 \cdot 2^{-6}\\
&\quad + 1 \cdot 2^{-7} + 1 \cdot 2^{-8} + 1 \cdot 2^{-9})
\end{aligned}$$

Zahlendarstellung

Seien $b \in \mathbb{N}$, $b > 1$, fest gewählt. Jedes $x \in \mathbb{R}$, $x \neq 0$, lässt sich in der Form

$$x = \pm \left(\sum_{j=1}^{\infty} d_j \, b^{-j} \right) \cdot b^e$$

darstellen, mit $d_j \in \{0, 1, \ldots, b-1\}$, $d_1 \neq 0$, und e eine ganze Zahl.

Zahlendarstellung

Seien $b \in \mathbb{N}$, $b > 1$, fest gewählt. Jedes $x \in \mathbb{R}$, $x \neq 0$, lässt sich in der Form

$$x = \pm \left(\sum_{j=1}^{\infty} d_j \, b^{-j} \right) \cdot b^e$$

darstellen, mit $d_j \in \{0, 1, \ldots, b-1\}$, $d_1 \neq 0$, und e eine ganze Zahl.

- Dezimalsystem (Basis $b = 10$)

$$123.75 \Rightarrow 0.12375 \cdot 10^3$$

Zahlendarstellung

Seien $b \in \mathbb{N}$, $b > 1$, fest gewählt. Jedes $x \in \mathbb{R}$, $x \neq 0$, lässt sich in der Form

$$x = \pm \left(\sum_{j=1}^{\infty} d_j \, b^{-j} \right) \cdot b^e$$

darstellen, mit $d_j \in \{0, 1, \ldots, b-1\}$, $d_1 \neq 0$, und e eine ganze Zahl.

- Dezimalsystem (Basis $b = 10$)

$$123.75 \Rightarrow 0.12375 \cdot 10^3$$

- Binärsystem (Basis $b = 2$)

$$123.75 \Rightarrow 0.111101111 \cdot 2^{111}$$

Zahlendarstellung

Seien $b \in \mathbb{N}$, $b > 1$, fest gewählt. Jedes $x \in \mathbb{R}$, $x \neq 0$, lässt sich in der Form

$$x = \pm \left(\sum_{j=1}^{\infty} d_j \, b^{-j} \right) \cdot b^e$$

darstellen, mit $d_j \in \{0, 1, \ldots, b-1\}$, $d_1 \neq 0$, und e eine ganze Zahl.

- Dezimalsystem (Basis $b = 10$)
  $$123.75 \Rightarrow 0.12375 \cdot 10^3$$
- Binärsystem (Basis $b = 2$)
  $$123.75 \Rightarrow 0.111101111 \cdot 2^{111}$$
- Dezimalsystem (Basis $b = 10$)
  $$\frac{1}{3} \Rightarrow 0.33333..... \cdot 10^0$$

Normalisierte Gleitpunktdarstellung

Floating Point Representation:

$$\begin{aligned} x &= \pm\, 0.d_1 d_2 \ldots d_m \cdot b^e \\ &= \pm \left(\sum_{j=1}^{m} d_j\, b^{-j} \right) \cdot b^e \end{aligned}$$

wobei

- Basis $b \in \mathbb{N} \backslash \{1\}$

Normalisierte Gleitpunktdarstellung

Floating Point Representation:

$$\begin{aligned} x &= \pm\, 0.d_1 d_2 \ldots d_m \cdot b^e \\ &= \pm \left(\sum_{j=1}^{m} d_j\, b^{-j} \right) \cdot b^e \end{aligned}$$

wobei

- Basis $b \in \mathbb{N} \backslash \{1\}$
- Exponent $e \in \mathbb{Z}$ mit $r \leq e \leq R$

Normalisierte Gleitpunktdarstellung

Floating Point Representation:

$$\begin{aligned} x &= \pm\, 0.d_1 d_2 \dots d_m \cdot b^e \\ &= \pm \left(\sum_{j=1}^{m} d_j\, b^{-j} \right) \cdot b^e \end{aligned}$$

wobei

- Basis $b \in \mathbb{N} \backslash \{1\}$
- Exponent $e \in \mathbb{Z}$ mit $r \leq e \leq R$
- Mantisse $f = \pm\, 0.d_1 d_2 \dots d_m,\ d_j \in \{0, 1, \dots, b-1\}$

Normalisierte Gleitpunktdarstellung

Floating Point Representation:

$$\begin{aligned} x &= \pm\, 0.d_1 d_2 \dots d_m \cdot b^e \\ &= \pm \left(\sum_{j=1}^{m} d_j\, b^{-j} \right) \cdot b^e \end{aligned}$$

wobei

- Basis $b \in \mathbb{N} \backslash \{1\}$
- Exponent $e \in \mathbb{Z}$ mit $r \leq e \leq R$
- Mantisse $f = \pm\, 0.d_1 d_2 \dots d_m,\ d_j \in \{0, 1, \dots, b-1\}$
- Mantissenlänge m

Normalisierte Gleitpunktdarstellung

Floating Point Representation:

$$\begin{aligned} x &= \pm\, 0.d_1 d_2 \dots d_m \cdot b^e \\ &= \pm \left(\sum_{j=1}^{m} d_j\, b^{-j} \right) \cdot b^e \end{aligned}$$

wobei

- Basis $b \in \mathbb{N} \backslash \{1\}$
- Exponent $e \in \mathbb{Z}$ mit $r \leq e \leq R$
- Mantisse $f = \pm\, 0.d_1 d_2 \dots d_m,\ d_j \in \{0, 1, \dots, b-1\}$
- Mantissenlänge m
- Normalisierung: $d_1 \neq 0$ für $x \neq 0$

Maschinenzahlen

Nur endliche Anzahl von Zahlen darstellbar:

$$\boldsymbol{x = \pm \left(\sum_{j=1}^{m} d_j \, b^{-j} \right) \cdot b^e, \qquad r \leq e \leq R}$$

$\Rightarrow$ Maschinenzahlen $\mathbb{M}(\boldsymbol{b, m, r, R})$.

Betragsmäßig kleinste bzw. größte Zahl in $\mathbb{M}(\boldsymbol{b, m, r, R})$: $\boldsymbol{x}_{\mathrm{MIN}}$, $\boldsymbol{x}_{\mathrm{MAX}}$.

Maschinenzahlen

Nur endliche Anzahl von Zahlen darstellbar:

$$x = \pm \left(\sum_{j=1}^{m} d_j \, b^{-j} \right) \cdot b^e, \qquad r \leq e \leq R$$

$\Rightarrow$ Maschinenzahlen $\mathbb{M}(b, m, r, R)$.
Betragsmäßig kleinste bzw. größte Zahl in $\mathbb{M}(b, m, r, R)$: x_{MIN}, x_{MAX}.

Reduktionsabbildung $\mathrm{fl} : \mathbb{D} \to \mathbb{M}(b, m, r, R)$

Für $x \in \mathbb{D} := [-x_{\mathrm{MAX}}, -x_{\mathrm{MIN}}] \cup [x_{\mathrm{MIN}}, x_{\mathrm{MAX}}]$

$$\mathrm{fl}(x) := \pm \begin{cases} \left(\sum_{j=1}^{m} d_j \, b^{-j} \right) \cdot b^e & \text{falls } d_{m+1} < \frac{b}{2}, \\ \left(\sum_{j=1}^{m} d_j \, b^{-j} + b^{-m} \right) \cdot b^e & \text{falls } d_{m+1} \geq \frac{b}{2}, \end{cases}$$

d.h. die letzte Stelle der Mantisse wird um eins erhöht bzw. beibehalten, falls die Ziffer in der nächsten Stelle $\geq \frac{b}{2}$ bzw. $< \frac{b}{2}$ ist.

x_{MIN} und x_{MAX}

- Betragsmäßig kleinste ($\neq 0$) Zahl:

$$d_1 = 1,\ d_2 = \cdots = d_m = 0;\ e = r : \quad x_{\mathrm{MIN}} = b^{r-1}$$

- Betragsmäßig größte Zahl:

$$d_1 = \cdots = d_m = b - 1;\ e = R : \quad x_{\mathrm{MAX}} = (1 - b^{-m}) \cdot b^R$$

x_{MIN} und x_{MAX}

- Betragsmäßig kleinste ($\neq 0$) Zahl:

$$d_1 = 1,\ d_2 = \cdots = d_m = 0;\ e = r : \quad x_{\text{MIN}} = b^{r-1}$$

- Betragsmäßig größte Zahl:

$$d_1 = \cdots = d_m = b-1;\ e = R : \quad x_{\text{MAX}} = (1 - b^{-m}) \cdot b^R$$

Partition $\mathbb{R} = \mathbb{D} \cup \mathbb{D}_{\min} \cup \mathbb{D}_{\max}$, mit
$\mathbb{D} := [-x_{\text{MAX}}, -x_{\text{MIN}}] \cup [x_{\text{MIN}}, x_{\text{MAX}}]$, $\mathbb{D}_{\min} = (-x_{\text{MIN}}, x_{\text{MIN}})$,
$\mathbb{D}_{\max} := (-\infty, -x_{\text{MAX}}) \cup (x_{\text{MAX}}, \infty)$.

$$x \in \mathbb{D}_{\max} : \quad fl(x) := \operatorname{sign}(x)\infty$$

$$x \in \mathbb{D}_{\min} : \quad fl(x) := \begin{cases} 0 & \text{falls } |x| < \frac{1}{2} x_{\text{MIN}} \\ \operatorname{sign}(x) x_{\text{MIN}} & \text{falls } |x| \geq \frac{1}{2} x_{\text{MIN}} \end{cases}$$

Maschinengenauigkeit – Beispiel

Gleitpunktdarstellung: $b = 10$, $m = 6$

x	$\mathrm{fl}(x)$	$\left\lvert\frac{\mathrm{fl}(x)-x}{x}\right\rvert$
$\frac{1}{3} = 0.33333333\ldots$	$0.333333 * 10^0$	$1.0 * 10^{-6}$
$\sqrt{2} = 1.41421356\ldots$	$0.141421 * 10^1$	$2.5 * 10^{-6}$
$e^{-10} = 0.000045399927\ldots$	$0.453999 * 10^{-4}$	$6.6 * 10^{-7}$
$e^{10} = 22026.46579\ldots$	$0.220265 * 10^5$	$1.6 * 10^{-6}$
$\frac{1}{10} = 0.1$	$0.100000 * 10^0$	0.0

Gleitpunktdarstellung: $b = 2$, $m = 10$

x	$\mathrm{fl}(x)$	$\left\lvert\frac{\mathrm{fl}(x)-x}{x}\right\rvert$
$\frac{1}{3}$	$0.1010101011 * 2^{-1}$	$4.9 * 10^{-4}$
$\sqrt{2}$	$0.1011010100 * 2^{1}$	$1.1 * 10^{-4}$
e^{-10}	$0.1011111010 * 2^{-111}$	$3.3 * 10^{-4}$
e^{10}	$0.1010110000 * 2^{1111}$	$4.8 * 10^{-4}$
$\frac{1}{10}$	$0.1100110011 * 2^{-11}$	$2.4 * 10^{-4}$

Dahmen & Reusken

Maschinengenauigkeit

- Für den relativen Rundungsfehler erhält man
$$\left|\frac{\mathrm{fl}(x) - x}{x}\right| \leq \frac{\frac{b^{-m}}{2} b^e}{b^{-1} b^e} = \frac{b^{1-m}}{2}.$$

Maschinengenauigkeit

- Für den relativen Rundungsfehler erhält man

$$\left|\frac{\mathrm{fl}(x) - x}{x}\right| \leq \frac{\frac{b^{-m}}{2} b^e}{b^{-1} b^e} = \frac{b^{1-m}}{2}.$$

- Die (relative) Maschinengenauigkeit

$$\mathrm{eps} := \frac{b^{1-m}}{2}$$

charakterisiert das Auflösungsvermögen des Rechners, d.h.

$$\mathrm{eps} = \inf\{\delta > 0 \mid \mathrm{fl}(1 + \delta) > 1\}$$

Maschinengenauigkeit

- Für den relativen Rundungsfehler erhält man

$$\left|\frac{\mathrm{fl}(x) - x}{x}\right| \leq \frac{\frac{b^{-m}}{2} b^e}{b^{-1} b^e} = \frac{b^{1-m}}{2}.$$

- Die (relative) Maschinengenauigkeit

$$\mathrm{eps} := \frac{b^{1-m}}{2}$$

charakterisiert das Auflösungsvermögen des Rechners, d.h.

$$\mathrm{eps} = \inf\{\delta > 0 \mid \mathrm{fl}(1+\delta) > 1\}$$

- Der Rundungsfehler ε erfüllt $|\varepsilon| \leq \mathrm{eps}$ und es gilt

$$\mathrm{fl}(x) = x\,(1+\varepsilon).$$

Maschinengenauigkeit – Beispiel

Gleitpunktdarstellung: $\boldsymbol{b = 10}$, $\boldsymbol{m = 6} \rightarrow \mathrm{eps} = \boldsymbol{\frac{1}{2} \times 10^{-5}}$

x	$\mathrm{fl}(x)$	$\left\lvert\frac{\mathrm{fl}(x)-x}{x}\right\rvert$
$\frac{1}{3} = 0.33333333\ldots$	$0.333333 * 10^{0}$	$1.0 * 10^{-6}$
$\sqrt{2} = 1.41421356\ldots$	$0.141421 * 10^{1}$	$2.5 * 10^{-6}$
$e^{-10} = 0.000045399927\ldots$	$0.453999 * 10^{-4}$	$6.6 * 10^{-7}$
$e^{10} = 22026.46579\ldots$	$0.220265 * 10^{5}$	$1.6 * 10^{-6}$
$\frac{1}{10} = 0.1$	$0.100000 * 10^{0}$	0.0

Gleitpunktdarstellung: $\boldsymbol{b = 2}$, $\boldsymbol{m = 10} \rightarrow \mathrm{eps} = \boldsymbol{9.8 \times 10^{-4}}$

x	$\mathrm{fl}(x)$	$\left\lvert\frac{\mathrm{fl}(x)-x}{x}\right\rvert$
$\frac{1}{3}$	$0.1010101011 * 2^{-1}$	$4.9 * 10^{-4}$
$\sqrt{2}$	$0.1011010100 * 2^{1}$	$1.1 * 10^{-4}$
e^{-10}	$0.1011111010 * 2^{-111}$	$3.3 * 10^{-4}$
e^{10}	$0.1010110000 * 2^{1111}$	$4.8 * 10^{-4}$
$\frac{1}{10}$	$0.1100110011 * 2^{-11}$	$2.4 * 10^{-4}$

Dahmen & Reusken

IEEE Standard

- Double-precision floating-point

  64-bit Wort: 52 bits für f, 11 bits für e, 1 bit für das Vorzeichen

- Der Exponent e ist eine ganze Zahl im Intervall

$$-1022 \leq e \leq 1023$$

- In MATLAB:

	Binary	Decimal
`eps`	`2^(-52)`	`2.2204e-16`
`realmin`	`2^(-1022)`	`2.2251e-308`
`realmax`	`(2-eps)*2^1023`	`1.7977e+308`

Gleitpunktarithmetik

Exakte elementare arithmetische Operation von Maschinenzahlen $\not\Rightarrow$ Maschinenzahl

Beispiel

$\boldsymbol{b = 10,\ m = 3}$:

$$\mathbf{0.346 \cdot 10^2 + 0.785 \cdot 10^2 = 0.1131 \cdot 10^3 \neq 0.113 \cdot 10^3}$$

Ähnliches passiert bei Multiplikation und Division.

Exakte Arithmetik $\rightsquigarrow$ Gleitpunktarithmetik (Pseudoarithmetik),

z.B.: $\qquad + \rightsquigarrow \oplus.$

Gleitpunktarithmetik

Forderung

Für $\nabla \in \{+, -, \cdot, \div\}$ gelte

$$x \circledcirc y = \mathrm{fl}(x \nabla y) \quad \text{für } x, y \in \mathbb{M}(b, m, r, R).$$

Da $\mathrm{fl}(x) = x\,(1+\varepsilon)$, folgt somit, dass für $\nabla \in \{+, -, \cdot, \div\}$

$$x \circledcirc y = (x \nabla y)(1+\varepsilon) \quad \text{für } x, y \in \mathbb{M}(b, m, r, R)$$

und ein ε mit $|\varepsilon| \le \mathrm{eps}$ gilt.

Gleitpunktarithmetik

Forderung

Für $\nabla \in \{+, -, \cdot, \div\}$ gelte

$$x \circledtriangledown y = \mathrm{fl}(x \nabla y) \quad \text{für } x, y \in \mathbb{M}(b, m, r, R).$$

Da $\mathrm{fl}(x) = x\,(1 + \varepsilon)$, folgt somit, dass für $\nabla \in \{+, -, \cdot, \div\}$

$$x \circledtriangledown y = (x \nabla y)(1 + \varepsilon) \quad \text{für } x, y \in \mathbb{M}(b, m, r, R)$$

und ein ε mit $|\varepsilon| \leq \mathrm{eps}$ gilt.

Vorsicht bei Gleitpunktarithmetik:

- Grundlegende Regeln der Algebra, die bei exakter Arithmetik gelten, sind nicht mehr gültig.
- Reihenfolge der Verküpfung spielt eine Rolle (Assoziativität der Addition geht verloren).

Assoziativgesetz

Beispiel 2.45

Zahlensystem mit $b = 10,\ m = 3$. Maschinenzahlen

$$\begin{array}{rcccl} x & = & 6590 & = & 0.659 \cdot 10^4 \\ y & = & 1 & = & 0.100 \cdot 10^1 \\ z & = & 4 & = & 0.400 \cdot 10^1 \end{array}$$

Exakte Rechnung:

$$(x+y)+z = (y+z)+x = 6595.$$

Assoziativgesetz

Beispiel 2.45

Zahlensystem mit $b = 10,\ m = 3$. Maschinenzahlen

$$\begin{array}{rcccl} x & = & 6590 & = & 0.659 \cdot 10^4 \\ y & = & 1 & = & 0.100 \cdot 10^1 \\ z & = & 4 & = & 0.400 \cdot 10^1 \end{array}$$

Exakte Rechnung:

$$(x + y) + z = (y + z) + x = 6595.$$

Pseudoarithmetik:

$$x \oplus y = 0.659 \cdot 10^4 \quad \text{und} \quad (x \oplus y) \oplus z = 0.659 \cdot 10^4,$$

aber

$$y \oplus z = 0.500 \cdot 10^1 \quad \text{und} \quad (y \oplus z) \oplus x = 0.660 \cdot 10^4.$$

Distributivgesetz

Beispiel 2.46

Für $b = 10,\ m = 3, x = 0.156 \cdot 10^2$ und $y = 0.157 \cdot 10^2$

$$\begin{aligned} (x - y) \cdot (x - y) &= 0.01 \\ (x \ominus y) \odot (x \ominus y) &= 0.100 \cdot 10^{-1} \end{aligned}$$

aber

$$(x \odot x) \ominus (x \odot y) \ominus (y \odot x) \oplus (y \odot y) = -0.100 \cdot 10^1.$$

Auslöschung

Beispiel 2.47

Betrachte

$$x = 0.73563, \quad y = 0.73441, \quad x - y = 0.00122.$$

Bei 3-stelliger Rechnung ($b = 10,\ m = 3,\ \mathrm{eps} = \frac{1}{2} \times 10^{-2}$):

$$\begin{aligned} \tilde{x} &= \mathrm{fl}(x) = 0.736, \qquad |\delta_x| &= 0.50 \cdot 10^{-3} \\ \tilde{y} &= \mathrm{fl}(y) = 0.734, \qquad |\delta_y| &= 0.56 \cdot 10^{-3} \end{aligned}$$

Die relative Störung im Resultat:

$$\left| \frac{(\tilde{x} - \tilde{y}) - (x - y)}{x - y} \right| = \left| \frac{0.002 - 0.00122}{0.00122} \right| = 0.64$$

also sehr groß im Vergleich zu δ_x, δ_y.

Zusammenfassung Gleitpunktarithmetik

$$\boxed{\left|\frac{(x \circledtriangledown y) - (x \nabla y)}{(x \nabla y)}\right| \leq \mathrm{eps}, \quad x, y \in \mathbb{M}, \quad \nabla \in \{+, -, \cdot, \div\}}$$

Die relativen Rundungsfehler bei den elementaren Gleitpunktoperationen sind $\leq \mathrm{eps}$, wenn die Eingangsdaten x, y Maschinenzahlen sind.

Sei $f(x, y) = x \nabla y$, $x, y \in \mathbb{R}$, $\nabla \in \{+, -, \cdot, \div\}$ und κ_{rel} die relative Konditionszahl von f. Es gilt

$$\boxed{\begin{array}{llll} \nabla \in \{\cdot, \div\} & : & \kappa_{\mathrm{rel}} \leq 1 & \text{für alle } x, y, \\ \nabla \in \{+, -\} & : & \kappa_{\mathrm{rel}} \gg 1 & \text{wenn } |x \nabla y| \ll \max\{|x|, |y|\} \end{array}}$$

Sehr große Fehlerverstärkung bei $+, -$ möglich (Auslöschung).

Beispiele

In Matlab:

- $\mathtt{u} = 0.3/0.1$

Beispiele

In Matlab:

- $\mathtt{u} = 0.3/0.1$
  - Das Ergebnis ist nicht gleich 3, da Zähler etwas kleiner als 0.3 und Nenner etwas größer als 0.1.

Beispiele

In Matlab:

- $\mathtt{u} = 0.3/0.1$
  - Das Ergebnis ist nicht gleich 3, da Zähler etwas kleiner als 0.3 und Nenner etwas größer als 0.1.
- $\mathtt{a} = 2^{100};\ \mathtt{b} = \mathtt{a} + 2^{47};\ \mathtt{b} - \mathtt{a} = 0$

Beispiele

In Matlab:

- $\texttt{u} = 0.3/0.1$
  - Das Ergebnis ist nicht gleich 3, da Zähler etwas kleiner als 0.3 und Nenner etwas größer als 0.1.
- $\texttt{a} = 2^{100};\ \texttt{b} = \texttt{a} + 2^{47};\ \texttt{b} - \texttt{a} = 0$
  - Die relative Differenz zwischen a und b is kleiner als eps

Beispiele

In Matlab:

- $\texttt{u} = 0.3/0.1$
  - Das Ergebnis ist nicht gleich 3, da Zähler etwas kleiner als 0.3 und Nenner etwas größer als 0.1.
- $\texttt{a} = 2^{100}\texttt{; b} = \texttt{a} + 2^{47}\texttt{; b} - \texttt{a} = 0$
  - Die relative Differenz zwischen a und b is kleiner als eps
  - Es gibt keine Maschinenzahl zwischen $\mathbf{2^{100}}$ und $\mathbf{2^{100} + 2^{48}}$

Beispiele

In Matlab:

- $\mathtt{u} = 0.3/0.1$
  - Das Ergebnis ist nicht gleich 3, da Zähler etwas kleiner als 0.3 und Nenner etwas größer als 0.1.
- $\mathtt{a} = 2^{100};\ \mathtt{b} = \mathtt{a} + 2^{47};\ \mathtt{b} - \mathtt{a} = 0$
  - Die relative Differenz zwischen a und b is kleiner als eps
  - Es gibt keine Maschinenzahl zwischen $\mathbf{2^{100}}$ und $\mathbf{2^{100} + 2^{48}}$
- eps/3+eps/3+1-1= 2.220446049250313e-16
  eps/3+1+eps/3-1= 0

Beispiele

In Matlab:

- $\mathtt{u = 0.3/0.1}$
  - Das Ergebnis ist nicht gleich 3, da Zähler etwas kleiner als 0.3 und Nenner etwas größer als 0.1.
- $\mathtt{a = 2^{100};\ b = a + 2^{47};\ b - a = 0}$
  - Die relative Differenz zwischen a und b is kleiner als eps
  - Es gibt keine Maschinenzahl zwischen $\mathbf{2^{100}}$ und $\mathbf{2^{100} + 2^{48}}$
- eps/3+eps/3+1-1= 2.220446049250313e-16
  eps/3+1+eps/3-1= 0
  - Assoziativgesetz gilt nicht

Beispiele

- Auswerten der Funktion $f(x) = 1 - x\left(\frac{x+1}{x} - 1\right)$

  Exakt: $f(x) = 1 - x\,\frac{x+1-x}{x} = 0$ für alle $x > 0$

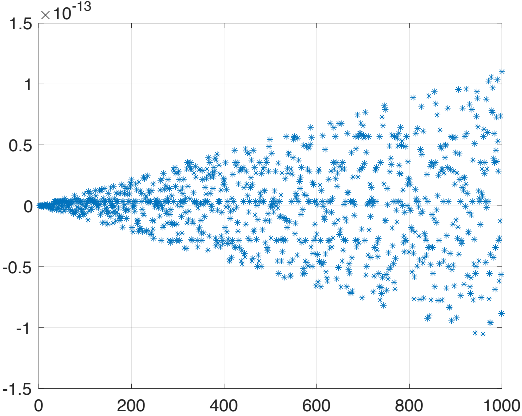
  Auswertung in Matlab:

Beispiele

- Auswerten der Funktion $f(x) = 1 - x\left(\frac{x+1}{x} - 1\right)$

  Exakt: $f(x) = 1 - x\,\frac{x+1-x}{x} = 0$ für alle $x > 0$

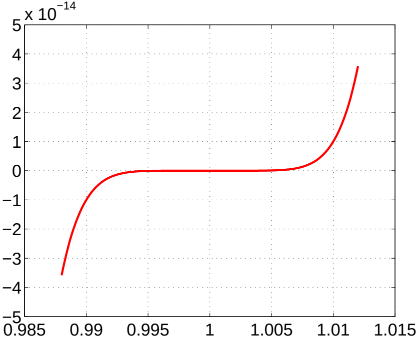
  Auswertung in Matlab:



Beispiel: Polynom 7. Grades

Matlab Plot

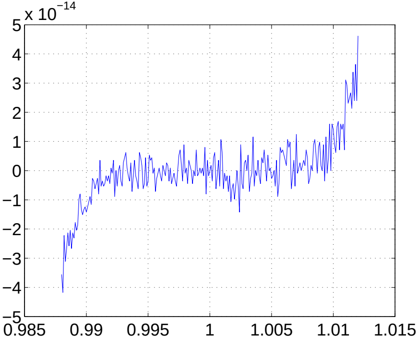
```
x = 0.988 : 0.0001 : 1.012;
y = (x - 1).^7;
plot(x,y)
```



Beispiel: Polynom 7. Grades

Matlab Plot

```
x = 0.988 : 0.0001 : 1.012;
y = x.^7 - 7 * x.^6 + 21 * x.^5 - 35 * x.^4
      +35 * x.^3 - 21 * x.^2 + 7 * x - 1;
plot(x,y)
```



Stabilität

Definition

Ein Algorithmus heißt gutartig oder stabil, wenn die durch ihn im Laufe der Rechnung erzeugten Fehler in der Größenordnung des durch die Kondition des Problems bedingten unvermeidbaren Fehlers bleiben.

Stabilität

Definition

Ein Algorithmus heißt gutartig oder stabil, wenn die durch ihn im Laufe der Rechnung erzeugten Fehler in der Größenordnung des durch die Kondition des Problems bedingten unvermeidbaren Fehlers bleiben.

- Kondition ist Eigenschaft des Problems
- Stabilität ist Eigenschaft des Verfahrens/Algorithmus

$\Rightarrow$ Wenn ein Problem schlecht konditioniert ist, kann man nicht erwarten, dass eine numerische Methode (ein stabiler Algorithmus) gute Ergebnisse liefert.

Stabilität

Definition

Ein Algorithmus heißt gutartig oder stabil, wenn die durch ihn im Laufe der Rechnung erzeugten Fehler in der Größenordnung des durch die Kondition des Problems bedingten unvermeidbaren Fehlers bleiben.

- Kondition ist Eigenschaft des Problems
- Stabilität ist Eigenschaft des Verfahrens/Algorithmus

$\Rightarrow$ Wenn ein Problem schlecht konditioniert ist, kann man nicht erwarten, dass eine numerische Methode (ein stabiler Algorithmus) gute Ergebnisse liefert.

Ziel: Numerische Methode soll Fehlerverstärkung nicht signifikant weiter vergrößern

Beispiel 2.50

Bestimmung der Lösung u^* von

$$y^2 - 2a_1 y + a_2 = 0$$

für $a_1 = 6.000227$, $a_2 = 0.01$.

Algorithmus I

$$u^* = f(a_1, a_2) = a_1 - \sqrt{a_1^2 - a_2}$$

$$\begin{array}{rrcl} & y_1 & = & a_1 \cdot a_1 \\ \longrightarrow & y_2 & = & y_1 - a_2 \\ \longrightarrow & y_3 & = & \sqrt{y_2} \\ \longrightarrow & u^* & = & a_1 - y_3 \end{array}$$

Beispiel 2.50

Algorithmus I

$$u^* = f(a_1, a_2) = a_1 - \sqrt{a_1^2 - a_2}.$$

In Gleitpunktarithmetik mit $b = 10$, $m = 5$ $\left(\text{eps} = \frac{1}{2} \cdot 10^{-4}\right)$:

$$\tilde{u}^* = 0.90000 \cdot 10^{-3}$$

Exakte Lösung:

$$u^* = 0.83336 \cdot 10^{-3}$$

Beispiel 2.50

Algorithmus I

$$u^* = f(a_1, a_2) = a_1 - \sqrt{a_1^2 - a_2}.$$

In Gleitpunktarithmetik mit $b = 10$, $m = 5$ $\left(\text{eps} = \frac{1}{2} \cdot 10^{-4}\right)$:

$$\tilde{u}^* = 0.90000 \cdot 10^{-3}$$

Exakte Lösung:

$$u^* = 0.83336 \cdot 10^{-3}$$

- Problem ist für diese Eingangsdaten $a_1,\ a_2$ gut konditioniert.
- Durch Algorithmus erzeugte Fehler sind sehr viel größer als der unvermeidbare Fehler.

$\Rightarrow$ Algorithmus I ist nicht stabil

Ursache: Auslöschung

Beispiel 2.50

Bestimmung der Lösung u^* von

$$y^2 - 2a_1 y + a_2 = 0$$

für $a_1 = 6.000227$, $a_2 = 0.01$.

Algorithmus II (Alternative)

$$u^* = \frac{a_2}{a_1 + \sqrt{a_1^2 - a_2}}$$

$$\begin{array}{rrcl} & y_1 & = & a_1 \cdot a_1 \\ \longrightarrow & y_2 & = & y_1 - a_2 \\ \longrightarrow & y_3 & = & \sqrt{y_2} \\ \longrightarrow & y_4 & = & a_1 + y_3 \\ \longrightarrow & u^* & = & \frac{a_2}{y_4} \end{array}$$

Beispiel 2.50

Algorithmus II

$$\boldsymbol{u^* = \frac{a_2}{a_1 + \sqrt{a_1^2 - a_2}}}$$

In Gleitpunktarithmetik mit $\boldsymbol{b = 10}$, $\boldsymbol{m = 5}$ $\left(\mathrm{eps} = \boldsymbol{\frac{1}{2} \cdot 10^{-4}}\right)$:

$$\boldsymbol{\tilde{u}^* = 0.83333 \cdot 10^{-3}}$$

Exakte Lösung:

$$\boldsymbol{u^* = 0.83336 \cdot 10^{-3}}$$

- Gesamtfehler bleibt im Rahmen der Maschinengenauigkeit.
- Auslöschung tritt nicht auf.

$\Rightarrow$ Algorithmus II ist stabil

Rückwärtsstabilität

Wunsch: Auswertung von $\boldsymbol{f} : \boldsymbol{X} \to \boldsymbol{Y}$

Wirklichkeit: berechnetes Ergebnis $\tilde{\boldsymbol{f}} : \boldsymbol{X} \to \boldsymbol{Y}$

wobei $\boldsymbol{f} \neq \tilde{\boldsymbol{f}}$ aufgrund von

- Rundungsfehlern (Maschinengenauigkeit),
- Gleitpunktarithmetik.

Rückwärtsstabilität

Wunsch: Auswertung von $f : X \to Y$

Wirklichkeit: berechnetes Ergebnis $\tilde{f} : X \to Y$

wobei $f \neq \tilde{f}$ aufgrund von

- Rundungsfehlern (Maschinengenauigkeit),
- Gleitpunktarithmetik.

Das Ziel

$$\frac{\|\tilde{f}(x) - f(x)\|}{\|f(x)\|} = \mathcal{O}(\mathrm{eps})$$

ist zu ehrgeizig.

Grund: Wenn Problem f schlecht konditioniert ist, werden Datenstörungen um Kondition $\kappa \gg 1$ des Problems verstärkt.

Rückwärtsstabilität

Ein Verfahren zur Berechnung von $f(x)$ liefert als Ergebnis $\tilde{f}(x)$.

Definition

Das Verfahren heißt rückwärts stabil, wenn

$$\tilde{f}(x) = f(\tilde{x})$$

für ein $\tilde{x}$ mit $\frac{\|x-\tilde{x}\|}{\|x\|} = \mathcal{O}(\mathrm{eps})$.

$\Rightarrow$ Ein rückwärts stabiler Algorithmus gibt die exakte Lösung des Problems mit nahezu richtigen Eingabedaten (d.h. $x \to \tilde{x} = x(1+\epsilon)$, $|\epsilon| \leq \mathrm{eps}$).

Rückwärtsstabilität

Satz

Wird ein rückwärts stabiler Algorithmus zur Lösung des Problems f mit Kondition $\kappa(x)$ angewendet, so gilt

$$\frac{\|\tilde{f}(x) - f(x)\|}{\|f(x)\|} = \mathcal{O}(\kappa(x)\,\mathrm{eps}).$$

Beweis:

$$\frac{\|\tilde{f}(x) - f(x)\|}{\|f(x)\|} = \frac{\|f(\tilde{x}) - f(x)\|}{\|f(x)\|} \lesssim \kappa(x)\,\underbrace{\frac{\|\tilde{x} - x\|}{\|x\|}}_{\mathcal{O}(\mathrm{eps})}.$$

Rückwärtsstabilität

Was haben wir gemacht?

Fehler im Algorithmus wurden

Rückwärtsstabilität

Was haben wir gemacht?

Fehler im Algorithmus wurden

zurückgespiegelt auf Fehler in den Daten.

Rückwärtsstabilität

Was haben wir gemacht?

Fehler im Algorithmus wurden

zurückgespiegelt auf Fehler in den Daten.

$$\Downarrow$$

Konsequenz:
Fehler in $\tilde{f}(x) = f(\tilde{x})$ ist unvermeidbar, wegen Kondition von f.

Rückwärtsanalyse

$$
\begin{tikzcd}
\boxed{\text{exakte Daten } x} \arrow[r, "{f\,[\text{exakt}]}"] \arrow[dd] \arrow[ddr, swap, "{\tilde f = f\,[\text{numerisch}]}"] & \boxed{f(x)} \arrow[d, dotted] \\
 & \begin{array}{c}\text{Fehler } f(x) - f(\tilde x)\text{:}\\ \text{Kondition des Problems}\end{array} \\
\boxed{\text{gestörte Daten } \tilde x} \arrow[r, "{f\,[\text{exakt}]}"] & \boxed{f(\tilde x) = \tilde f(x)} \arrow[u, dotted]
\end{tikzcd}
$$

Beispiel 2.54: Summation ist rückwärts stabil

Geg.: Maschinenzahlen $x_1,\ x_2,\ x_3$, Maschinengenauigkeit eps.
Ges.: Summe $S = (x_1 + x_2) + x_3$.

Man erhält

$$\widetilde{S} = ((x_1 + x_2)(1 + \varepsilon_2) + x_3)(1 + \varepsilon_3)$$

mit $|\varepsilon_i| \leq \mathrm{eps},\ i = 2, 3$.

Beispiel 2.54: Summation ist rückwärts stabil

Geg.: Maschinenzahlen $x_1,\ x_2,\ x_3$, Maschinengenauigkeit eps.
Ges.: Summe $S = (x_1 + x_2) + x_3$.

Man erhält

$$\widetilde{S} = ((x_1 + x_2)(1 + \varepsilon_2) + x_3)(1 + \varepsilon_3)$$

mit $|\varepsilon_i| \le \mathrm{eps},\ i = 2, 3$.

Daraus folgt

$$\begin{aligned} \widetilde{S} &= x_1 (1 + \varepsilon_2)(1 + \varepsilon_3) + x_2 (1 + \varepsilon_2)(1 + \varepsilon_3) + x_3 (1 + \varepsilon_3) \\ &\doteq x_1 (1 + \varepsilon_2 + \varepsilon_3) + x_2 (1 + \varepsilon_2 + \varepsilon_3) + x_3 (1 + \varepsilon_3) \\ &= x_1 (1 + \delta_1) + x_2 (1 + \delta_2) + x_3 (1 + \delta_3) \end{aligned}$$

wobei

$$|\delta_1| = |\delta_2| = |\varepsilon_2 + \varepsilon_3| \le 2\,\mathrm{eps}, \quad |\delta_3| = |\varepsilon_3| \le \mathrm{eps}$$

Beispiel 2.54: Summation ist rückwärts stabil

Es gilt

$$\begin{aligned}\widetilde{S} &= x_1\,(1+\delta_1) + x_2\,(1+\delta_2) + x_3\,(1+\delta_3)\\ &=: \tilde{x}_1 + \tilde{x}_2 + \tilde{x}_3,\end{aligned}$$

wobei

$$|\delta_1| = |\delta_2| = |\varepsilon_2 + \varepsilon_3| \leq 2\,\mathrm{eps}, \quad |\delta_3| = |\varepsilon_3| \leq \mathrm{eps}$$

$\Rightarrow$ Fehlerbehaftetes Resultat $\widetilde{S}$ als exaktes Ergebnis zu gestörten Eingabedaten $\tilde{x}_i = x_i(1+\delta_i)$.

Beispiel 2.54: Summation ist rückwärts stabil

Es gilt

$$\begin{aligned}\widetilde{S} &= x_1\,(1+\delta_1) + x_2\,(1+\delta_2) + x_3\,(1+\delta_3)\\ &=: \tilde{x}_1 + \tilde{x}_2 + \tilde{x}_3,\end{aligned}$$

wobei

$$|\delta_1| = |\delta_2| = |\varepsilon_2 + \varepsilon_3| \le 2\,\mathrm{eps}, \quad |\delta_3| = |\varepsilon_3| \le \mathrm{eps}$$

$\Rightarrow$ Fehlerbehaftetes Resultat $\widetilde{S}$ als exaktes Ergebnis zu gestörten Eingabedaten $\tilde{x}_i = x_i(1+\delta_i)$.

Der durch Rechnung bedingte Fehler ist höchstens

$$\begin{aligned}\left|\frac{f(\tilde{x})-f(x)}{f(x)}\right| &\le \kappa_{\mathrm{rel}}(x)\sum_{j=1}^{3}\left|\frac{\tilde{x}_j - x_j}{x_j}\right|\\ &\le \kappa_{\mathrm{rel}}(x)\sum_{j=1}^{3}|\delta_j| \le \kappa_{\mathrm{rel}}(x)\,5\,\mathrm{eps}.\end{aligned}$$

Beispiel 2.54

Der für die Summation $f(x) = f(x_1, x_2, x_3) = x_1 + x_2 + x_3$ unvermeidbare Fehler ist

$$\left|\frac{f(\tilde{x}) - f(x)}{f(x)}\right| \leq \kappa_{\mathrm{rel}}(x) \sum_{j=1}^{3} \left|\frac{\tilde{x}_j - x_j}{x_j}\right| \leq \kappa_{\mathrm{rel}}(x) 3\,\mathrm{eps},$$

wenn Daten höchstens mit Maschinengenauigkeit gestört werden ($\tilde{x}_i = x_i(1+\varepsilon),\ |\varepsilon| \leq \mathrm{eps}$).

Beispiel 2.54

Der für die Summation $f(x) = f(x_1, x_2, x_3) = x_1 + x_2 + x_3$ unvermeidbare Fehler ist

$$\left|\frac{f(\tilde{x}) - f(x)}{f(x)}\right| \leq \kappa_{\mathrm{rel}}(x) \sum_{j=1}^{3} \left|\frac{\tilde{x}_j - x_j}{x_j}\right| \leq \kappa_{\mathrm{rel}}(x) 3 \,\mathrm{eps},$$

wenn Daten höchstens mit Maschinengenauigkeit gestört werden $(\tilde{x}_i = x_i(1+\varepsilon),\ |\varepsilon| \leq \mathrm{eps})$.

$\Rightarrow$ Berechnung von S ist ein stabiler Algorithmus