

Vorlesungsskript

Num. Lin. Algebra

Inhaltsverzeichnis

1. Einleitung	2
2. Das Gauß-Verfahren I	4
2.1. Gaußsche Eliminationsverfahren und LR-Zerlegung	4
2.2. Pivot-Strategien	8
2.3. Cholesky-Verfahren für symm. pos. definite A	11
3. Fehleranalyse	12
3.1. Zahlendarstellung und Rundungsfehler	13
3.2. Kondition eines Problems	13
3.3. Stabilität von Algorithmen	16
4. Das Gauß-Verfahren II	20
4.1. Lösung von linearen Gleichungssystemen: Kondition	20
4.2. Stabilität der Gaußelimination	22
4.3. Nachiteration	23
5. Die QR-Zerlegung	27
5.1. Householder-Transformationen	27
5.2. Berechnung der QR -Zerlegung	29

Definitionen

1. Einleitung

Wichtige Aufgabenklassen der linearen Algebra sind **lineare Gleichungssysteme**.

Gegeben: $A \in \mathbb{R}^{m \times n}, b \in \mathbb{R}^m$

Gesucht: Ein/alle $x \in \mathbb{R}^n$ mit $Ax = b$

Herkunft:

- “direkt” aus der Anwendung, z.B. Beschreibung von Netzwerken, Tragwerk
- “indirekt” als Diskretisierung von stationären Prozessen, z.B. Belastung einer Membran
- “mittelbar” durch die Linearisierung nichtlinearer Modelle, z.B. Newton-Verfahren, Approximation von Lösungen gewöhnlicher DGL, notwendige Optimalitätsbedingungen

Klassifizierung:

- $m = n$: A quadratisch

Generische Situation: A regulär

$\implies \exists!$ Lösung

- $m < n$: “Unterbestimmtes System”

Generische Situation:

$$\begin{aligned} \operatorname{rg}(A) &= m \text{ (Vollrang)} \\ A &\hat{=} [A_1 A_2] \quad A_1 \in \mathbb{R}^{m \times m} \end{aligned}$$

Lösungsmenge:

$$\mathcal{L} = \{x \in \mathbb{R}^n \mid Ax = b\} = \{x = x^+ + h, h \in \ker(A)\}$$

= $(n - m)$ -dimensionale lineare Mannigfaltigkeit

Gesucht ist dann z.B. norm-minimale Lösung (Kap. 5)

- $m > n$: "Überbestimmtes System"

$$\text{lösbar} \iff b \in \text{im}(A) = \{y \in \mathbb{R}^m \mid \exists x : Ax = y\}$$

Generisch nicht lösbar!

Sinnvoll: Bestimme $\bar{x} \in \mathbb{R}^m$, so dass

$$\|A\bar{x} - b\| = \min_{x \in \mathbb{R}^m} \|Ax - b\|$$

$\|\cdot\| \hat{=}$ geeignete Norm, $\bar{x} \hat{=}$ Bestapproximierender für diese Norm.

Mögliche Ansätze:

- $\|\cdot\|_\infty : \|Ax - b\|_\infty = \max_{1 \leq i \leq m} |(Ax - b)_i|$

Ein nichtglattes Optimierungsproblem auch als lineares Optimierungsproblem formulierbar, schwierig zu lösen für m bzw. n groß.

- $\|\cdot\|_1 : \|Ax - b\|_1 = \sum_{i=1}^m |Ax - b|$

Wie bei $\|\cdot\|_\infty$ stückweise lineares Optimierungsproblem.

Aber stabil gegen Ausreißer.

- $\|\cdot\|_2 : \|Ax - b\|_2^2 = \sum_{i=1}^m ((Ax - b)_i)^2$

$\hat{=}$ lineares Quadratmittelpunktproblem, kleinste Quadrateproblem (Kap. 5)

Verfahren zur Lösung von LGS:

Direkte Verfahren:

- Transformation der Daten (A, b) in endlich viele in ein leichter zu lösendes LGS $\tilde{A}x = \tilde{b} \hat{=}$ CG-Verfahren
- Transformationen lassen sich oftmals als Faktorisierung von A interpretieren

$$A = L \cdot R \quad \text{bzw.} \quad A = Q \cdot R$$

- Dafür i.d.R. Zugriff auf Elemente von $A \implies$ limitiert die Größe der Matrix!

Kap. 2-5

Indirekte Verfahren:

- Ausgehend von einem Startvektor x^0 Iteration zur Berechnung von x^k mit $Ax^k \approx b$
Hierbei wird oftmals nur das Matrix-Vektor-Produkt Av benötigt! (Kap. 6)
- Eigenwertprobleme

Stabilitätsanalyse von Bauwerken. Verfahren dazu: numerische Optimierung

2. Das Gauß-Verfahren I

Jetzt: $A \in \mathbb{R}^{m \times n}$, $b \in \mathbb{R}^n$, $x : Ax = b$?

Satz 2.1: Existenz und Eindeutigkeit einer Lösung

Sei $A \in \mathbb{R}^{m \times n}$ eine Matrix mit $\det(A) \neq 0$ und $b \in \mathbb{R}^n$. Dann existiert genau ein $x \in \mathbb{R}^n$ mit

$$Ax = b$$

Beweis: lineare Algebra

□

\Rightarrow Anwendung von Algorithmen zur Berechnung von x sinnvoll! Wie?

2.1. Gaußsche Eliminationsverfahren und LR-Zerlegung

$\hat{=}$ direktes Verfahren für quadratisches System

Erste Idee: Systeme spezieller Struktur, z.B.

$$Rx = c, \quad R = \begin{pmatrix} r_{11} & \cdots & r_{1n} \\ 0 & \ddots & \cdots \\ 0 & 0 & r_{nn} \end{pmatrix} \in \mathbb{R}^{n \times n}, \quad c = \begin{pmatrix} c_1 \\ \vdots \\ c_n \end{pmatrix} \in \mathbb{R}^n$$

$$Rx = c$$

$$r_{nn}x_n = c_n \Rightarrow x_n = \frac{c_n}{r_{nn}}, \quad r_{nn} \neq 0$$

$$r_{n-1n-1}x_{n-1} + r_{n-1n}x_n = c_{n-1}$$

$$x_{n-1} = \frac{c_{n-1} - r_{n-1n}x_n}{r_{n-1n-1}}, \quad r_{n-1n-1} \neq 0$$

Algorithmus 2.2: Rückwärtssubstitution

$$\begin{aligned}
x_n &= \frac{c_n}{r_{nn}} \quad \text{falls } r_{nn} \neq 0 \\
&\vdots \\
x_i &= \frac{c_i - \sum_{j=i+1}^n r_{ij} x_j}{r_{ii}} \quad \text{falls } r_{ii} \neq 0 \\
&\vdots \\
x_1 &= \frac{c_1 - \sum_{j=2}^n r_{1j} x_j}{r_{11}} \quad \text{falls } r_{11} \neq 0
\end{aligned}$$

Algo. 2.2 anwendbar, wenn $\det(R) \neq 0$ (vgl. Theo. 2.1)

Wichtiger Aspekt dieser Vorlesung: Aufwandsabschätzung

Aufwand: i -te Zeile je $n - i$ Additionen und Multiplikationen und 1 Division
insgesamt:

$$\sum_{i=1}^n (i-1) = \frac{n(n-1)}{2} = \mathcal{O}(n^2)$$

Addition und Multiplikationen und n Divisionen.

Landau-Symbol: $\mathcal{O}(\cdot)$

$$f(n) = \mathcal{O}(g(n)) \iff \exists c > 0 : |f(n)| \leq C|g(n)|$$

Für ein lineares Gleichungssystem der Form

$$Lx = z, \quad L = \begin{pmatrix} l_{11} & & 0 \\ \vdots & \ddots & \\ l_{n1} & \dots & l_{nn} \end{pmatrix} \in \mathbb{R}^{n \times n} \quad z \in \begin{pmatrix} z_1 \\ \vdots \\ z_n \end{pmatrix} \in \mathbb{R}^n$$

gibt es einen analogen Algorithmus:

$$\begin{aligned}
x_1 &= \frac{z_1}{l_{11}} \quad l_{11} \neq 0 \\
&\vdots \\
x_n &= \frac{z_n - \sum_{i=1}^{n-1} l_{ni} x_i}{l_{nn}} \quad l_{nn} \neq 0
\end{aligned}$$

\implies Vorwärtssubstitution mit gleichem Aufwand $\mathcal{O}(n^2)$

Lösungsidee für ein allgemeines Gleichungssystem:

Faktorisiere $A = L \cdot R$ und berechne die Lösung x von $Ax = b$ durch

$$Ax = L \underbrace{Rx}_{=:z} = b$$

$$Lz = b \implies z = L^{-1}b \quad \text{Vorwärtssubstitution}$$

$$Rx = z \implies x = R^{-1}z \quad \text{Rückwärtssubstitution}$$

Mit Aufwand: $\mathcal{O}(n^2)$

Frage: Wie berechnet man Zerlegung $A = L \cdot R$

Man generiert eine Folge von Matrizen:

$$A = A^{(1)} \longrightarrow A^{(2)} \longrightarrow \dots \longrightarrow A^n = R$$

von Matrizen der Gestalt

$$A^{(k)} = \begin{pmatrix} a_{11}^{(1)} & \dots & \dots & \dots & \dots & a_{1n}^{(1)} \\ 0 & a_{22}^{(2)} & \dots & \dots & \dots & a_{2n}^{(2)} \\ & 0 & \ddots & & & \vdots \\ & & 0 & a_{kk}^{(k)} & \dots & a_{kn}^{(k)} \\ & & & \vdots & & \vdots \\ & & & a_{nk}^{(k)} & \dots & a_{nn}^{(k)} \end{pmatrix}$$

Wie?

Sei $x = (x_1, \dots, x_n)^T \in \mathbb{R}^n$, $x_k \neq 0 \hat{=} k$ -Spalte

Definiere: $l_{jk} = \frac{x_j}{x_k}$

$$l_k = \left(\underbrace{0, \dots, 0}_{k \text{ mal}}, l_{k+1k}, \dots, l_{nk} \right)^T$$

$e_k = k$ -ter Einheitsvektor

$$L_k = I_n - l_k e_k^T \in \mathbb{R}^{n \times n}$$

Dann gilt

$$L_k x = \begin{pmatrix} 1 & & & & 0 \\ & \ddots & & & \\ & & 1 & & \\ & & -l_{k+1k} & \ddots & \\ & & \vdots & & \\ & & -l_{nk} & & 1 \end{pmatrix} \begin{pmatrix} x_1 \\ \vdots \\ x_n \end{pmatrix} = \begin{pmatrix} x_1 \\ \vdots \\ x_k \\ 0 \\ \vdots \\ 0 \end{pmatrix}$$

Jeder Eliminationsschritt $A^{(k)} \longrightarrow A^{(k+1)}$ lässt sich damit als Multiplikation mit einer Matrix $L_k \in \mathbb{R}^{n \times n}$ von links

$$A^{k+1} = L_k A^{(k)} = \begin{pmatrix} I_k & 0 \\ 0 & * & I_{n-k} \end{pmatrix} \begin{pmatrix} A_{11}^{(k)} & A_{12}^{(k)} \\ A_{21}^{(k)} & A_{22}^{(k)} \end{pmatrix} = \begin{pmatrix} A_{11}^{(k)} & A_{12}^{(k)} \\ 0 & A_{22}^{(k+1)} \end{pmatrix} \quad * \in \mathbb{R}^{n-k,1}$$

Eine Matrix der Gestalt L_k heißt Frobeniusmatrix \rightarrow weitere Eigenschaften siehe Übung.

Der Eliminationsschritt ist genau dann durchführbar wenn $a_{kk}^{(k)} \neq 0$ gilt. Angenommen, dies gilt, dann erhält man

$$L_n \cdots L_2 L_1 A = R$$

$$A = \underbrace{L_1^{-1} \cdots L_{n-2}^{-1} L_{n-1}^{-1}}_{=: L} R$$

Induktiv beweist man

$$L = L_1^{-1} \cdots L_{n-1}^{-1} = \begin{pmatrix} 1 & & & & 0 \\ l_{21} & \ddots & & & \\ \vdots & l_{32} & \ddots & & \\ \vdots & \vdots & \vdots & \ddots & \\ \vdots & \vdots & & \ddots & \\ l_{n1} & l_{n2} & \cdots & l_{nn-1} & 1 \end{pmatrix}$$

Durch diese Struktur kann der Speicherplatz für A zum Speichern von L und R genutzt werden!

Algorithmus 2.3: LR -Zerlegung

Gegeben: $A \in \mathbb{R}^{n \times n}$

```

for  $i = 1, \dots, n$ 
  for  $j = i, \dots, n$ 
    for  $k = 1, \dots, i - 1$ 
       $a_{ij} = a_{ij} - a_{ik} * a_{kj}$ 
    end
  end
  for  $j = i + 1, \dots, n$ 
    for  $k = 1, \dots, i - 1$ 
       $a_{ji} = a_{ji} - a_{jk} * a_{ki}$ 
    end
     $a_{ji} = \frac{a_{ji}}{a_{ii}}$ 
  end
end
end

```

Aufwand für die Dreieckszerlegung $A = L \cdot R$

#Operationen =

$$\begin{aligned}\sum_{i=1}^{n-1} \left((n-i)^2 + (n-i) \right) &= \frac{1}{3}n^3 - \frac{1}{2}n^2 + \frac{1}{6}n + \frac{1}{2}n^2 - \frac{1}{2}n \\ &= \frac{1}{3}n^3 + \mathcal{O}(n^2) = \mathcal{O}(n^3)\end{aligned}$$

\Rightarrow kubischer Aufwand! Nur akzeptabel für moderates n !

Algorithmus 2.4: Gaußsche Eliminationsverfahren

Gegeben: $A \in \mathbb{R}^{n \times n}$, $b \in \mathbb{R}^n$

1) Berechne $A = L \cdot R \quad \mathcal{O}(n^3)$

2) Berechne z aus $Lz = b \quad \mathcal{O}(n^2)$

3) Berechne x aus $Rx = z \quad \mathcal{O}(n^2)$

\Rightarrow Gesamtaufwand (Operationen): $\mathcal{O}(n^3)$, (Speicherplatz): $n^2 + n$

Vorteil der Faktorisierung:

Zerlegung (teuer) kann für mehrere rechte Seiten nachgenutzt werden.

2.2. Pivot-Strategien

Beispiel 2.5: Algo 2.4 kann selbst für einfache Schritte scheitern:

$$Ax = b, \quad x = \begin{pmatrix} w \\ z \end{pmatrix}, \quad A = \begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix}, \quad \det(A) = -1, \quad b = \begin{pmatrix} c \\ e \end{pmatrix}$$

$$\Rightarrow a_{11} = 0 \quad \nmid$$

Bei der völlig äquivalenten Formulierung

$$\tilde{A}\tilde{x} = \tilde{b}, \quad \tilde{x} = \begin{pmatrix} w \\ z \end{pmatrix}, \quad \tilde{A} = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}, \quad \det(\tilde{A}) = 1, \quad \tilde{b} = \begin{pmatrix} e \\ c \end{pmatrix}$$

funktioniert Algo 2.4 mit

$$\begin{aligned}\tilde{A} &= I_2 = L \cdot R \\ L &= I_2 \quad R = I_2\end{aligned}$$

\Rightarrow Zeilenvertauschung in A und der rechten Seite, **nicht** in x bzw. \tilde{x} !

Die LR -Zerlegung versagt nicht nur bei verschwindenden Diagonalelementen, sondern auch wenn diese betragsmäßig klein im Vergleich zu den restlichen Elementen sind.

\leadsto Praktikum, Fehlertheorie (Kap. III)

Algorithmus 2.6: LR -Zerlegung mit Spaltenpivotisierung

1. $k = 1, A^{(1)} = A$

2. Spaltenpivotisierung

Bestimme $p \in \{k, \dots, n\}$ so, dass

$$|a_{pk}^{(k)}| \geq |a_{jk}^{(k)}| \text{ für } j = k, \dots, n$$

3. Vertausche die Zeilen p und k durch

$$A^{(k)} \rightarrow \tilde{A}^{(k)} \text{ mit } \tilde{a}_{ij}^{(k)} = \begin{cases} a_{kj}^{(k)} & \text{falls } i = p \\ a_{pj}^{(k)} & \text{falls } i = k \\ a_{ij}^{(k)} & \text{sonst} \end{cases}$$

4. Führen der Eliminationsschritte

$$\tilde{A}^{(k)} \rightarrow A^{(k+1)} \text{ setze } k = k + 1$$

5. Falls $k = n$ STOP

Sonst gehe zu 2.

Alternative Pivotisierungsstrategien:

- Zeilenpivotisierung und Spaltentausch
- vollständige Pivotisierung, d.h. Suche des betragsmäßig größten Elements in der Restmatrix

Aufwand:

- Sowohl Spalten- als auch Zeilenpivotisierung: Im schlimmsten Fall $\mathcal{O}(n^2)$ zusätzliche Operationen
- vollständige Pivotisierung: Im schlimmsten Fall $\mathcal{O}(n^3)$ zusätzliche Operationen

Formale Beschreibung von Algo 2.6? Dazu: Permutationsmatrizen $P_\pi \in \mathbb{R}^{n \times n}$

Jede Permutation $\pi : \{1, \dots, n\} \rightarrow \{1, \dots, n\}$ der Zahlen $1, \dots, n$ bestimmt eine Matrix

$$P_\pi = (e_{\pi(1)} \ \dots \ e_{\pi(n)})$$

Eine Zeilenvertauschung in A kann dann durch das Produkt $P_\pi A$ beschrieben werden, Spaltenvertauschung durch AP_π . Des Weiteren gilt $P_\pi^{-1} = P_\pi^T$, $\det(P_\pi) = \{-1, 1\}$.

Man kann beweisen, dass die LR -Zerlegung mit Spaltenpivotisierung theoretisch nur versagen kann, wenn $\det(A) = 0$

Satz 2.7: Durchführbarkeit der LR -Zerlegung

Für jede invertierbare Matrix A existiert eine Permutationsmatrix P derart, dass für PA die LR -Zerlegung mit Spaltenpivotisierung durchgeführt werden kann. D.h., man erhält $PA = LR$. Dabei kann man P so wählen, dass alle Elemente von L betragsmäßig kleiner gleich 1 sind, also $|L| \leq 1$

Beweis: Da A invertierbar ist, gilt $\det(A) \neq 0$. Damit existiert eine Permutationsmatrix P_{π_1} , so dass das erste Diagonalelement $\tilde{a}_{11}^{(1)}$ der Matrix

$$\tilde{A}^{(1)} = P_{\pi_1} A^{(1)}$$

von Null verschieden ist und das betragsmäßig größte Element in der ersten Spalte ist:

$$0 \neq |\tilde{a}_{11}^{(1)}| \geq |\tilde{a}_{i1}^{(1)}| \quad \text{für } i = 1, \dots, n$$

Nach dem ersten Eliminationsschritt erhalten wir

$$A^{(2)} = L_1 \tilde{A}^{(1)} = L_1 P_{\pi_1} A = \begin{pmatrix} \tilde{a}_{11}^{(1)} & * \\ 0 & \tilde{A}_2^{(2)} \end{pmatrix}$$

Wegen (*) gilt für L_1 :

$$|l_{i1}| = \left| \frac{\tilde{a}_{i1}^{(1)}}{\tilde{a}_{11}^{(1)}} \right| \leq 1 \quad i = 2, \dots, n$$

$$\Rightarrow |L_1| \leq 1, \quad \det(L_1) = 1$$

$$\det(A^{(2)}) = \underbrace{\det(L_1)}_{=1} \underbrace{\det(P_{\pi_1})}_{\in \{-1,1\}} \underbrace{\det(A)}_{\neq 0} \neq 0$$

$$\det(\tilde{A}^{(2)}) = \frac{\overbrace{\det(A^{(2)})}^{\neq 0}}{\tilde{a}_{11}^{(1)}} \neq 0$$

Induktiv erhält man

$$R = A^{(n)} = L_{n-1} R_{\pi_{n-1}} L_{n-2} P_{\pi_{n-2}} \cdots L_1 P_{\pi_1} A$$

mit $|L_k| \leq 1$ und P_{π_k} entweder die Identität oder zwei Zeilen $j_1, j_2 \geq k$ vertauschen. Deswegen gilt für die Frobeniusmatrix

$$L_k = \begin{pmatrix} 1 & & & 0 \\ & \ddots & & \\ & & 1 & \\ & & -l_{k+1k} & \\ & & \vdots & \ddots \\ & & -l_{nk} & & 1 \end{pmatrix}, \text{ dass}$$

$$\tilde{L}_k = P_{\pi_j} L_k P_{\pi_j}^{-1} = \begin{pmatrix} 1 & & & 0 \\ & \ddots & & \\ & & 1 & \\ & & -l_{\pi_j(k+1)k} & \\ & & \vdots & \ddots \\ & & -l_{\pi_j(n)k} & & 1 \end{pmatrix} \quad \text{für } j > k$$

Durch geschicktes Einfügen von $I = P_{\pi_{k+1}}^{-1} P_{\pi_{k+1}}$

$$R = A^{(n)} = L_{k-1} \left(P_{\pi_{n-1}} L_{n-2} P_{\pi_{n-1}}^{-1} \right) \left(P_{\pi_{n-1}} P_{\pi_{n-2}} L_{k-3} P_{\pi_{n-2}}^{-1} P_{\pi_{n-1}}^{-1} \right) \\ P_{\pi_{n-1}} P_{\pi_{n-2}} \cdot \dots \cdot \left(\dots L_1 P_{\pi_1} \dots P_{\pi_{n-1}}^{-1} \left(P_{\pi_{n-1}} \dots P_{\pi_1} A \right) \right)$$

$$\Rightarrow PA = \underbrace{\tilde{L}_1^{-1} \dots \tilde{L}_{n-1}^{-1}}_{=:L} R \text{ mit}$$

$$L = \begin{pmatrix} 1 & & & 0 \\ l_{\tilde{\pi}_1(l)1} & \ddots & & \\ & \ddots & \ddots & \\ \vdots & & & \\ l_{\tilde{\pi}_1(n)1} & \dots & l_{\tilde{\pi}_{n-1}(n)(n-1)} & 1 \end{pmatrix}$$

und $|L| \leq 1$

Bemerkungen:

- Gilt $PA = LR$, dann berechnet man

$$Ax = b$$

$$PAx = Pb$$

$$LRx = Pb$$

$$x = R^{-1} L^{-1} Pb$$

- Theoretisch sind die Formulierungen

$$Ax = b \quad DAx = Db$$

für eine invertierbare Diagonalmatrix D äquivalent. Bei der praktischen Lösung auf dem Rechner haben solche Skalierungen aber u.U. einen **dramatischen** Einfluß, vgl. Kap. III.

- Auf dem Rechner: Verbesserung der unexakten Lösung durch sogenannte Nachiteration möglich, vgl. Kap. IV.

2.3. Cholesky-Verfahren für symm. pos. definite A

Gesucht: A spd eine $L \in \mathbb{R}^{n \times n}$ ($\det(L) > 0$) s.d. $A = LL^T$

siehe Übungen

3. Fehleranalyse

Situation

ideal: Eingabe $x \longrightarrow$ Algorithmus/Problemstellung $f \longrightarrow$ Ausgabe $y = f(x)$

real: $\tilde{x} = x + \varepsilon \longrightarrow \tilde{f} \longrightarrow \tilde{y} = \tilde{f}(\tilde{x})$

Frage: $y \longleftrightarrow \tilde{y}$?

Ursachen für den Gesamtfehler $\tilde{y} - y$

- **Modellfehler**

- Idealisierungsfehler, z.B. in der Modellbildung
- Datenfehler

Modellfehler lassen in der Regel nicht vermeiden!

Frage: Wie wirken sich solche Fehler **unabhängig** vom gewählten Algorithmus aus?

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

Kondition eines Problems

- **numerische Fehler**

- Diskretisierungsfehler, kontinuierliches Problem versus diskretisierte Formulierung
- Abbruchfehler, eigentlich unendliche Algorithmen werden nach endlichen Schritten abgebrochen
- Approximationsfehler

$$\begin{aligned} \sin(x) &= \sum_{n=0}^{\infty} (-1)^n \frac{x^{2n+1}}{(2n+1)!} \\ \Rightarrow \widetilde{\sin}(x) &= \sum_{n=0}^k (-1)^n \frac{x^{2n+1}}{(2n+1)!} \end{aligned}$$

- Rechengenauigkeit, reelle Zahlen versus Gleitkommazahlen

Rundungsfehler und Approximationsfehler \Rightarrow **Stabilität** eines Algorithmus

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

Vernachlässigung von Fehlerbetrachtungen kann dramatische Auswirkungen haben:

- 1991: Untergang der Bohrsinsel Sleipner, Fehler in den Kräften von 47%
- 1. Golfkrieg: Eine Patriotrakete verpasst angreifende Rakete. Im Steuerprogramm der Patriotrakete durch Multiplikation mit 0.1. Nach 100 Betriebsstunden: Differenz der berechneten Zeit und tatsächlich vergangener Zeit von 0.34 Sekunden
- Absturz der ersten Ariane 5 Rakete (1996), Umwandlung einer 64 bit Gleitkommazahl in 16 bit integer Zahl in Software der Arian 4
- London Millenium Bridge (2000), flasche Abschätzung der Fußgängerkräfte

3.1. Zahlendarstellung und Rundungsfehler

→ Einführung in das wissenschaftliche Rechnen

3.2. Kondition eines Problems

Erwartungshaltung: kleiner Fehler in der Aufgabenstellung ($x \rightarrow \tilde{x}$ verursacht einen kleinen Fehler in der Lösung \tilde{y})

Beispiel 3.1: Störung eines LGS

Gegeben ist das lineare Gleichungssystem

$$\underbrace{\begin{pmatrix} 1.2969 & 0.8648 \\ 0.2161 & 0.1441 \end{pmatrix}}_{=:A} \begin{pmatrix} x_1 \\ x_2 \end{pmatrix} = \underbrace{\begin{pmatrix} 0.8642 \\ 0.1440 \end{pmatrix}}_{=:b}$$

mit $\det(A) \neq 0$ und der eindeutig bestimmten Lösung $x = (2 \ 2)^T$.

Jetzt: Störung der rechten Seite

$$b \rightsquigarrow \tilde{b} = \begin{pmatrix} 0.86419999 \\ 0.14400001 \end{pmatrix}$$

liefert die Lösung $\tilde{x} = (0.9911 \ -0.4870)^T$. Ursache?

Dazu: Formalisierung Eigenschaften der Problemstellung

Wichtig: Notation: x - Eingabe, f - Problemstellung, y - Ausgabe

Definition 3.2: Numerisches Problem

Ein numerisches Problem ist ein Paar (f, x) wobei $f : D \subset \mathbb{R}^n \rightarrow \mathbb{R}^m$ eine Abbildung, $x \in D$ die Eingabe und $y = f(x)$ die Ausgabe ist.

Beispiel 3.3:

- Auswertung von $\sin(x)$: $x = 1.7, y = f(x) = \sin(x) = \sin(1.7)$
- Bestimmung von Nullstelle von $g(t) = at^2 + bt + c$

Eingabe: $x = (a, b, c), y = f(x)$ definiert durch $g(f(a, b, c)) \stackrel{!}{=} 0$

Zur Lösung eines numerischen Problems können verschiedene Algorithmen genutzt werden

(Algorithmus: endliche Folge von Elementaroperationen, deterministisch bestimmt)

Hier: Die Kondition ist **unabhängig** vom gewählten Algorithmus!

Definition 3.4: wohl gestelltes Problem, schlecht gestelltes Problem

Das numerische Problem (f, x) heißt wohlgestellt, falls es eine konstante $L_{\text{abs}} \in \mathbb{R}^+$ gibt, so dass

$$\|f(\tilde{x}) - f(x)\| \leq L_{\text{abs}} \|x - \tilde{x}\|$$

für alle $\tilde{x} \rightarrow x$. Existiert keine solche Konstante L_{abs} , dann heißt (f, x) schlecht gestellt. Zur weiteren Analyse setzt man im wohldefinierten Fall

$$\kappa_{\text{abs}} := \inf\{L_{\text{abs}} \mid L_{\text{abs}} \geq 0 \text{ und } (*) \text{ gilt}\}$$

Gilt $x \neq 0 \neq f(x)$, definiert man analog κ_{rel} als die kleinste Konstante mit

$$\frac{\|f(\tilde{x}) - f(x)\|}{\|f(x)\|} \leq \kappa_{\text{rel}} \frac{\|\tilde{x} - x\|}{\|x\|}$$

für alle \tilde{x} nahe x .

Bemerkungen:

- Die **absolute Kondition** κ_{abs} beschreibt die Verstärkung des absoluten Fehlers, die **relative Kondition** κ_{rel} die Verstärkung des relativen Fehlers
- Bei nichtlinearen Problemen hängen κ_{abs} und κ_{rel} meist stark von der Umgebung ab \Rightarrow linearisierte Fehlertheorie!
- κ_{abs} und κ_{rel} hängen stark von den verwendeten Normen ab! $\|\cdot\|_2, \|\cdot\|_\infty, \|\cdot\|_p, \|\cdot\|_1$

Definition 3.5: absolute und relative Kondition

Die Konstante κ_{abs} gibt die absolute Kondition eines numerischen Problems (f, x) und κ_{rel} die relative Kondition.

Das numerische Problem (f, x) ist **schlecht konditioniert**, wenn κ_{abs} bzw. κ_{rel} “groß” sind und gut konditioniert, wenn κ_{abs} bzw. κ_{rel} “klein” sind.

Wie berechnet man $\kappa_{\text{abs}}/\kappa_{\text{rel}}$? Dafür: Mittelwertsatz der Differentialrechnung

Es sei $f : [a, b] \rightarrow \mathbb{R}$ stetig auf $[a, b]$ und diffbar auf (a, b) . Dann existiert $\bar{x} \in (a, b)$, so dass

$$f'(\bar{x}) = \frac{f(b) - f(a)}{b - a}$$

Anwendung in der Fehlertheorie: Für differenzierbare $f : \mathbb{R}^n \rightarrow \mathbb{R}$ existiert wegen der Taylorentwicklung für x und Δx ein $\bar{x} \in x + t\Delta x, t \in (0, 1)$ mit

$$\Delta y := \tilde{y} - y = f(x + \Delta x) - f(x) = \nabla f(\bar{x}) \Delta x$$

$\Rightarrow \|\nabla f(\bar{x})\|$ ist ein Fehlermaß $\sim \rightarrow x$.

Deswegen verwendet man den Term $\|\nabla f(x)\|$ als Maß für die Fehlerverstärkung des absoluten Eingabefehlers $\|\Delta x\| = \|\tilde{x} - x\|$.

Der relative Fehler ist meist von größerer Bedeutung. Für $n = 1$ und $x \cdot y \neq 0$

$$\frac{\Delta y}{y} \approx \nabla f(x) \frac{\Delta x}{y} = \underbrace{\left(\nabla f(x) \frac{x}{f(x)} \right)}_{=\kappa_{\text{rel}}} \frac{\Delta x}{x}$$

Verallgemeinerung auf $n > 1$: $\kappa_{\text{rel}} = \left| \nabla f(x)^T x \cdot \frac{1}{f(x)} \right|$ o **Beispiel 3.6:** Kondition der Addition

Problem: $f : \mathbb{R}^2 \rightarrow \mathbb{R}, f(x_1, x_2) = x_1 + x_2$ mit der l_1 -Norm

$$\nabla f(x) = (1, 1)^T \Rightarrow$$

$$\kappa_{\text{abs}} = \|\nabla f(x)\|_1 = \|(1 \ 1)^T\|_1 = 2$$

$$\kappa_{\text{rel}} = \left\| \nabla f(x)^T x \frac{1}{f(x)} \right\|_1 = \frac{|x_1| + |x_2|}{|x_1 + x_2|}$$

Für die Addition zweier Zahlen mit gleichen Vorzeichen ergibt sich $\kappa_{\text{rel}} = 2 \Rightarrow$ gut konditioniert!

Gleikommazahlen: Hämmerlin, Hoffmann: Numerische Mathematik, Springer (1994)

Beispiel 3.6: $f : \mathbb{R}^2 \rightarrow \mathbb{R}, f(x) = x_1 + x_2$

$$\kappa_{\text{rel}} = \left\| \nabla f(x)^T x \cdot \frac{1}{f(x)} \right\|_1 = \frac{|x_1| + |x_2|}{|x_1 + x_2|} = \star$$

x_1, x_2 **gleiche** Vorzeichen, z.B. $x_1, x_2 > 0$

$$\star = \frac{x_1 + x_2}{x_1 + x_2} = 1$$

\Rightarrow sehr gut konditioniert!

Beispiel 3.7: Subtraktion zweier Zahlen

Problem: $f : \mathbb{R}^2 \rightarrow \mathbb{R}, f(x_1, x_2) = x_1 - x_2$

Wähle die l_1 -Norm

$$f'(x) = \begin{pmatrix} 1 \\ -1 \end{pmatrix} \Rightarrow \kappa_{\text{abs}} = \|Df(x)\|_1 = \|(1 \ -1)^T\|_1 = 2$$

$$\kappa_{\text{rel}} = \left\| \nabla f(x)^T x \frac{1}{f(x)} \right\|_1 = \frac{|x_1| + |x_2|}{|x_1 - x_2|}$$

Subtraktion zwei fast gleicher Zahlen ist schlecht konditioniert da

$$|x_1 - x_2| \ll |x_1| + |x_2|$$

Für die Rechengenauigkeit $\text{eps} = 10^{-7}$ (einfache Genauigkeit)

$$x_1 = 1.23467 * \quad \text{Störung in der 7. Stelle}$$

$$x_2 = 1.23456 * \quad \text{Störung in der 7. Stelle}$$

$$x_1 - x_2 = 0.00011 * = 0.11 \cdot 10^{-3} \text{ Störung in der 3. Stelle}$$

\Rightarrow Man verliert 4 Stellen an Genauigkeit

$$\Rightarrow \kappa_{\text{rel}} \approx 10^4$$

Problemstellung, Kondition \leftrightarrow Algorithmus

Wichtiges Beispiel: Sekantenverfahren zur Lösung nichtlinearer Gleichungen

theoretisch: sehr schöne Konvergenzeigenschaften

praktisch: Erhebliche Probleme durch schlechte Kondition der Subtraktion

3.3. Stabilität von Algorithmen

Jetzt: Wie wirken sich Eingabefehler und Fehler während der Rechnung auf das Endergebnis aus?

Vorwärtsanalyse

Definition 3.8: Vorwärtsstabilität (komponentenweise)

Die Implementierung \tilde{f} heißt **vorwärtsstabil** wenn für alle x aus dem Definitionsbereich von f mit $f(x) \neq 0$ ein moderates, von x unabhängiges $C_V > 0$, so dass

$$\left| \frac{\tilde{f}(x) - f(x)}{f(x)} \right| \leq C_V \cdot \kappa_{\text{rel}} \cdot \text{eps}$$

mit eps als Rechengenauigkeit gilt.

Hier betrachtet man die Fehlerfortpflanzung, d.h. die Auswirkung bereits gemachter Fehler.

Dazu: x_1, x_2 sind die exakten Daten, $\Delta x_1, \Delta x_2$ sind die bisher gemachten Fehler mit $\left| \frac{\Delta x_1}{x_1} \right|, \left| \frac{\Delta x_2}{x_2} \right| \ll 1$

Was passiert bei exakter Durchführung einer arithmetischen Operation?

Lemma 3.9: Gegeben seien $x_1, x_2, \Delta x_1, \Delta x_2 \in \mathbb{R}$. Dann gelten mit $\circ \in \{+, -, \cdot, \div\}$ für den forgepflanzten Fehler

$$\Delta(x_1 \circ x_2) = (x_1 + \Delta x_1) \circ (x_2 + \Delta x_2) - x_1 \circ x_2$$

die Abschätzung:

$$\frac{\Delta(x_1 \pm x_2)}{x_1 \pm x_2} = \frac{x_1}{x_1 \pm x_2} \cdot \frac{\Delta x_1}{x_1} \pm \frac{x_2}{x_1 \pm x_2} \cdot \frac{\Delta x_2}{x_2}$$

$$\frac{\Delta(x_1 \cdot x_2)}{x_1 \cdot x_2} \approx \frac{\Delta x_1}{x_1} + \frac{\Delta x_2}{x_2}$$

$$\frac{\Delta\left(\frac{x_1}{x_2}\right)}{\frac{x_1}{x_2}} \approx \frac{\Delta x_1}{x_1} - \frac{\Delta x_2}{x_2}$$

Dabei bedeutet \approx ein Vernachlässigen von Termen höherer Ordnung, z.B. x_1^2, x_2^2

Beweis: Nachrechnen

□

Fazit: \pm können u.U. zu einer erheblichen Fehlerverstärkung führen!

\cdot, \div : Im wesentlichen unkritische Fehlerforpflanzung

Die Fehlerverstärkung tritt besonders dann auf, wenn $|x_1| \approx |x_2|, x_1 \pm x_2$ nahe Null.

Dieser Effekt heißt **Auslöschung**

Rückwärtsanalyse

$$\tilde{f}(x) \stackrel{?}{=} f(\tilde{x}) = f(x + \Delta x)$$

Erwartungshaltung: Δx nicht zu groß

Definition 3.10: Rückwärtsstabilität (komponentenweise)

Die Implementierung \tilde{f} heißt **rückwärtsstabil**, wenn für alle $x \neq 0$ aus dem Definitionsbereich von f und Δx mit $\tilde{f}(x) = f(x + \Delta x)$ die Abschätzung

$$\left| \frac{\Delta x}{x} \right| \leq C_R \cdot \text{eps}$$

für eps als Rechengenauigkeit und ein moderates von x unabhängiges $C_R > 0$ gilt.

D.h. kann $\tilde{f}(x)$ als exaktes Ergebnis einer gestörten Eingabe $\tilde{x} = x + \Delta x$ interpretieren?

Bemerkungen:

- Δx muss nicht existieren, z.B. außerhalb des Definitionsbereichs
- f nicht injektiv \Rightarrow u.U. existieren mehrere Kandidaten, dann wählt man \tilde{x} so, dass $\|x - \tilde{x}\|$ minimal ist

$$f(x) \longleftrightarrow \tilde{f}(\tilde{x}) \quad ?$$

Es gilt:

$$\left| \frac{\tilde{f}(x) - f(x)}{f(x)} \right| = \left| \frac{f(x + \Delta x) - f(x)}{f(x)} \right|$$

$$\approx \kappa_{\text{rel}} \left| \frac{x + \Delta x - x}{x} \right| \leq \kappa_{\text{rel}} \cdot C_R \cdot \text{eps}$$

Also: Für ein wohl gestelltes Problem ist eine rückwärtsstabile Implementierung auch immer vorwärtsstabil mit $C_V = C_R$

Fazit für den Gesamtfehler:

$$\begin{aligned} \|f(x) - \tilde{f}(\tilde{x})\| &= \|f(x) - f(\tilde{x}) + f(\tilde{x}) - \tilde{f}(\tilde{x})\| \\ &\leq \underbrace{\|f(x) - f(\tilde{x})\|}_{\text{Kondition}} + \underbrace{\|f(\tilde{x}) - \tilde{f}(\tilde{x})\|}_{\text{Stabilität}} \end{aligned}$$

Ein gut konditioniertes Problem und ein stabiler Algorithmus sichern gute numerische Ergebnisse!

Beispiel 3.11: Auslöschung

Betrachtet wird

$$f(x) = x^3 \left(\frac{x}{x^2 - 1} - \frac{1}{x} \right)$$

Funktionsauswertung?

Matlab, $x = 2$

$$y_1 = \frac{x}{x^2 - 1} = \frac{2}{3}, \quad y_2 = \frac{1}{x} = \frac{1}{2}, \quad y_3 = x^3 = 8, \quad y_4 = y_3 \cdot (y_1 - y_2) = \frac{4}{3}$$

$$x = 1.2 \cdot 10^7$$

$$y_1 = 8.33333333333339 \cdot 10^{-8}, \quad y_2 = 8.33333333333334 \cdot 10^{-8}, \quad y_3 = x^3 = 1.728^{21}$$

$$\Rightarrow y_4 = y_1 - y_2 = 5.691 \cdot 10^{-22}, \quad y_5 = y_3 \cdot y_4 = 0.983$$

Wir können f umschreiben zu

$$f(x) = x^3 \left(\frac{x}{x^2 - 1} - \frac{1}{x} \right) = \frac{1}{1 - x^{-2}} =: g(x) > 1$$

Stabilität beider Formulierung?

$$f'(x) = \dots = -\frac{2x}{(x^2 - 1)^2}$$

$$\kappa_{\text{rel}} = \frac{2}{x^2 - 1} \leq 1 \quad \text{für } x \geq 4$$

\Rightarrow Eingabefehler werden gedämpft!

$$\left| \frac{\tilde{f}(x) - f(x)}{f(x)} \right| \approx 0.02 = C_V \cdot \kappa_{\text{rel}} \cdot \text{eps}$$

$$\Rightarrow C_V > 10^{13}$$

\Rightarrow diese Implementierung ist nicht vorwärtsstabil!

Für $x = 1.2 \cdot 10^7$:

$$x^{-2} \approx 6.94 \cdot 10^{-15}$$

$$1 - x^{-2} \approx 1.0$$

$$\frac{1}{1 - x^{-2}} \approx 1.0$$

Stabilität? Exakte Rechnung:

$$\begin{aligned} g(x) &= \frac{1}{1 - x^{-2}} \\ &= \frac{1}{1 - (12 \cdot 10^7)^{-2}} \\ &= \frac{1}{1 - (1.44 \cdot 10^{14})^{-1}} \\ &= \frac{1}{0.9999999999999995} \\ &\approx 1.000000000000000711 \end{aligned}$$

$$\Rightarrow \left| \frac{\tilde{g}(x) - g(x)}{g(x)} \right| \approx 7.11 \cdot 10^{-15}$$

$$\kappa_{\text{rel}}(g) = \left| g'(x) \cdot \frac{x}{g(x)} \right| = \left| -\frac{2x}{(1 - x^{-2})^2} \cdot \frac{x}{\frac{1}{1 - x^{-2}}} \right| = \left| \frac{2}{x^{-2} - 1} \right| \leq 1$$

$\Rightarrow C_V \approx 10$, vorwärtsstabil

4. Das Gauß-Verfahren II

Wieder die Frage: Wann ist ein lineares Gleichungssystem lösbar? Jetzt mit Fehlertheorie!

Beispiel 4.1: Interpolationspolynome

siehe PDF-Datei

4.1. Lösung von linearen Gleichungssystemen: Kondition

Zwei Möglichkeiten:

1. Störung Δb in b . Löse $A(x + \Delta x) = b + \Delta b$. Größe von Δx ?
2. Störung ΔA in A . Existiert eine Lösung $x + \Delta x = \tilde{x}$ von $(A + \Delta A)(x + \Delta x) = b$. Größe von Δx ?

zu 1.:

Absolute Kondition:

$$\begin{aligned}\|\Delta x\| &= \|A^{-1} \Delta b\| \leq \|A^{-1}\| \cdot \|\Delta b\| \\ \Rightarrow \kappa_{\text{abs}} &= \|A^{-1}\|\end{aligned}$$

Relative Kondition:

$$\begin{aligned}\frac{\|\Delta x\|}{\|x\|} &\leq \text{cond}(A) \frac{\|\Delta b\|}{\|b\|} \\ \kappa_{\text{rel}} &= \text{cond}(A) := \|A\| \cdot \|A^{-1}\|\end{aligned}$$

zu Bsp. 3.1: Hier gilt für $\|\cdot\| = \|\cdot\|_2$

$$\begin{aligned}\kappa_{\text{abs}} &\approx 1.5803 \cdot 10^8 \\ \kappa_{\text{rel}} &\approx 1.5803 \cdot 10^8\end{aligned}$$

Desweiteren gilt: (siehe ÜA)

$$\text{cond}(A) = \frac{\max_{\|x\|=1} \|Ax\|}{\min_{\|x\|=1} \|Ax\|} \in [1, \infty[$$

- Damit erhält man $A \neq 0$ ist genau dann singulär, wenn $\text{cond}(A) = \infty$
- Außerdem gilt $\text{cond}(\alpha A) = \text{cond}(A)$ für $0 \neq \alpha \in \mathbb{R}$

\Rightarrow die Kondition einer Matrix ist skalierungsinvariant!

zu 2.: Warum existiert $(A + \Delta A)^{-1}$ zu der gestörten Matrix $A + \Delta A$?

$$A + \Delta A = A(I + A^{-1}\Delta A), \quad \det(A) \neq 0$$

Lemma 4.2: Neumannsche Reihe

Sei $C \in \mathbb{R}^{n \times n}$ mit $\|C\| < 1$ mit einer submultiplikativen Norm. Dann ist $I - C$ invertierbar und man schreibt

$$(I - C)^{-1} = \sum_{k=0}^{\infty} C^k$$

Weiterhin gilt

$$\|(I - C)^{-1}\| \leq \frac{1}{1 - \|C\|}$$

Beweis: ÜA

□

nach Carl Gottfried Neumann (1832 - 1925), deutsche Mathematiker

Satz 4.3: Störungssatz

Es sei $\text{cond}(A) \cdot \frac{\|\Delta A\|}{\|A\|} < 1$ und $x + \Delta x$ die Lösung von $(A + \Delta A)(x + \Delta x) = b + \Delta b$. Dann gilt (*)

$$\frac{\|\Delta x\|}{\|x\|} \leq \text{cond}(A) \left(1 - \text{cond}(A) \frac{\|\Delta A\|}{\|A\|} \right) \left(\frac{\|\Delta A\|}{\|A\|} + \frac{\|\Delta b\|}{\|b\|} \right)$$

für $x \neq 0$ und $b \neq 0$.

Beweis: ÜA

□

Wenn $\text{cond}(A) \frac{\|\Delta A\|}{\|A\|} \ll 1$ und (*) gilt

$$\frac{\|\Delta x\|}{\|x\|} \leq \text{cond}(A) \cdot \text{eps}$$

wenn der relative Fehler in A und b auf dem Niveau der Maschinengenauigkeit sind. Der Störungssatz liefert dann, dass man wegen der Kondition des numerischen Problems $(A, b) \mapsto x = A^{-1}b$ einen unvermeidbaren Fehler der Größenordnung $\text{cond}(A) \cdot \text{eps}$ erwarten kann.

Beispiel 4.4: Fortsetzung von Bsp. 3.1

$$A = \begin{pmatrix} 1.2969 & 0.8648 \\ 0.2161 & 0.1441 \end{pmatrix}, \quad \text{cond}(A) = 2.469 \cdot 10^8 \quad \text{für } \|\cdot\|_2, \quad b = \begin{pmatrix} 0.8642 \\ 0.1440 \end{pmatrix}$$

$$\Rightarrow x = \begin{pmatrix} 2 \\ -2 \end{pmatrix}$$

Lösung mit Matlab $x = A \setminus b$

$$x = \begin{pmatrix} 2.0000000001802645 \\ -2.0000000007450581 \end{pmatrix}$$

$$\Rightarrow \frac{\|\Delta x\|_2}{\|x\|_2} = 2.715 \cdot 10^{-9} \leq 2.469 \cdot 10^8 \cdot 10^{-16}$$

Dann: $\|\Delta b\|_2 = 1.614 \cdot 10^{-8}$

$$\begin{aligned} \frac{\|\Delta x\|_2}{\|x\|_2} &= 1.64679 \leq \text{cond}(A) \cdot 1 \cdot \frac{\|\Delta b\|_2}{\|b\|_2} \\ &= 2.469 \cdot 10^8 \cdot 1 \cdot 7.6799 \cdot 10^7 \\ &\approx 1.9177 \end{aligned}$$

Beispiel 4.5:

$$A = \begin{pmatrix} 2 & 0.999 \\ 4 & 2.003 \end{pmatrix}, \quad b = \begin{pmatrix} 1.001 \\ 1.997 \end{pmatrix} \Rightarrow x = \begin{pmatrix} 1 \\ -1 \end{pmatrix}$$

$$\Rightarrow A^{-1} = \begin{pmatrix} 200.2\bar{9} & -99.9\bar{9} \\ -399.\bar{9} & 199.9\bar{9} \end{pmatrix}$$

$$\Rightarrow \|A\|_\infty = 6.003$$

$$\Rightarrow \|A^{-1}\|_\infty = 5.99\bar{9}$$

$$\text{cond}_\infty(A) = 3607.7\bar{9}$$

Jetzt:

$$\tilde{A} = \begin{pmatrix} 2 & 1 \\ 4 & 2.003 \end{pmatrix}, \quad \tilde{b} = \begin{pmatrix} 1.001 \\ 2 \end{pmatrix}$$

$$A = \begin{pmatrix} 0 & 0.001 \\ 0 & 0 \end{pmatrix}, \quad \Delta b = \begin{pmatrix} 0 \\ 0.003 \end{pmatrix}$$

$$\Rightarrow \|\Delta A\|_\infty = 0.001, \|\Delta b\|_\infty = 0.003$$

$$\Rightarrow \frac{\|\Delta x\|_\infty}{\|x\|_\infty} \leq 14.4072$$

$$\tilde{x} = \begin{pmatrix} 0.8\bar{3} \\ -0.\bar{6} \end{pmatrix} \Rightarrow \frac{\|\Delta x\|}{\|x\|} = 1.49$$

gap

4.2. Stabilität der Gaußelimination

Stabilität der LR - bzw. Cholesky-Zerlegung.

Rückwärtsanalyse: Interpretation des Ergebnisses als Ergebnis exakter Rechnung mit gestörten Eingangsdaten.

Abschätzung von ΔA : Störungssatz 4.3 und Kondition von A .

1. Fall:

A spd Matrix

\Rightarrow Cholesky-Verfahren $\mathcal{O}(n^2)$

Für die berechnete Lösung \tilde{x} als exakte Lösung des gestörten Systems:

$$(A + \Delta A) \underbrace{\tilde{x}}_{=x+\Delta x} = b$$

mit $\frac{\|\Delta A\|_\infty}{\|A\|_\infty} \leq C_n \text{ eps}$ gilt, dass c_n "klein" ist und nur von der Dimension von A abhängt, d.h. A ist im Rahmen der Maschinengenauigkeit.

Des Weiteren erhält man

$$\frac{\|\Delta x\|_\infty}{\|x\|_\infty} = \frac{\|\tilde{x} - x\|_\infty}{\|x\|_\infty} \leq \frac{\text{cond}_\infty(A) \left(\frac{\|\Delta A\|_\infty}{\|A\|_\infty} \right)}{1 - \text{cond}_\infty(A) \left(\frac{\|\Delta A\|_\infty}{\|A\|_\infty} \right)} \lesssim \frac{\text{cond}_\infty(A) c_n \text{ eps}}{1 - \text{cond}_\infty(A) c_n \text{ eps}} \approx \text{cond}_\infty(A) c_n \text{ eps}$$

falls $\text{cond}_\infty(A) \frac{\|\Delta A\|_\infty}{\|A\|_\infty} \ll 1$. D.h. der Fehler in \tilde{x} bleibt bzgl. der Größenordnung im Rahmen der unvermeidbaren Fehler.

\Rightarrow Cholesky-Verfahren ist stabil

Eine genauere Analyse findet man bei Deuffhard/Hohmann und auch Stoer/Bulirsch.

2. Fall:

A nicht spd

$\Rightarrow LR$ -Zerlegung mit Spaltenpivotisierung

Dann gilt auch

$$(A + \Delta A) \tilde{x} = b \quad \text{mit} \quad \frac{\|\Delta A\|_\infty}{\|A\|_\infty} \leq \tilde{c}_n \text{ eps}$$

Allerdings ist \tilde{c}_n sehr viel größer als c_n . Trotzdem kann die LR -Zerlegung mit Spaltenpivotisierung noch als stabil angesehen werden.

4.3. Nachiteration

Gauß-Elimination liefert \tilde{x} als Approximation der Lösung x von $Ax = b$. Wie gut ist \tilde{x} ? Für die exakte Lösung x verschwindet das Residuum $r(x) = Ax - b = 0$.

Ist $r(\tilde{x}) = A\tilde{x} - b$ "klein" ein sinnvolles Kriterium?

Problem: Die Norm

$\|r(\tilde{x})\|$ kann durch Zeilenskalierung beliebig verändert werden

$$Ax = b \iff DAx = Db$$

Deswegen: Beurteilung anhand der Rückwärtsanalyse, basiert auf einem Resultat von Rigal und Gaches (1967)

Satz 4.6: Gegeben sei das numerische Problem $Ax = b$, d.h. $x = A^{-1}b$. Dann ist der normweise relative Rückwärtsfehler einer Näherungslösung \tilde{x} gegeben durch:

$$\|\Delta x\| = \frac{\|A\tilde{x} - b\|}{\|A\| \cdot \|\tilde{x}\| + \|b\|}$$

Praktischer Nutzen? Nachiteration > D.h.:

Ausgehend von einer fehlerhaften LR -Zerlegung, d.h.

$$\tilde{L}\tilde{R} \neq PA$$

und daher nur einer Näherungslösung $\tilde{x} =: x_0$ mit exaktem Defekt

$$d_0 := r(x_0) = b - Ax_0,$$

kann man die Zerlegung $\tilde{L}\tilde{R} \approx PA$ zur Lösung der sogenannten Defektgleichung

$$Aw = d_0 \implies \tilde{L}\tilde{R}w = d_0.$$

\implies Korrektur w_1 für x_0 , d.h.

$$x_1 = x_0 + w_1$$

liefert dann

$$Ax_1 = Ax_0 + Aw_1 = Ax_0 - b + b + d_0 = b$$

$\implies x_1$ wäre exakte Lösung, aber: w_1 ist auch fehlerbehaftet

Frage: Ist x_1 eine bessere Lösung als x_0 ?

Dazu: Fehleranalyse basierend auf dem Störungssatz, Satz 4.3.

$$\frac{\|x_0 - x\|}{\|x\|} \leq \frac{\text{cond}(A)}{1 - \text{cond}(A) \frac{\|PA - \tilde{L}\tilde{R}\|}{\|A\|}} \cdot \underbrace{\frac{\|PA - \tilde{L}\tilde{R}\|}{\|A\|}}_{\approx \text{eps}}$$

\implies Stellenverlust entspricht der Kondition von A . Die zusätzlich auftretenden Rechenfehler werden vernachlässigt. Ersetzt man den exakten Defekt d_0 durch den Ausdruck

$$\hat{d}_0 = b - \hat{A}x_0$$

mit einer genaueren Darstellung von A ,

$$\frac{\|A - \hat{A}\|}{\|A\|} \leq \tilde{\varepsilon} \ll \varepsilon$$

erhält man

$$\begin{aligned} x_1 &= x_0 + w_1 = x_0 + (\tilde{L}\tilde{R})^{-1}(b - \hat{A}x_0) = x_0 + (\tilde{L}\tilde{R})^{-1}(Ax - Ax_0 + (A - \hat{A})x_0) \\ &\Rightarrow x_1 - x = x_0 - x + (\tilde{L}\tilde{R})^{-1}A(x - x_0) + (\tilde{L}\tilde{R})^{-1}(A - \hat{A})x_0 = (\star) \end{aligned}$$

Es gilt:

$$\begin{aligned} \tilde{L}\tilde{R} &= A - A + \tilde{L}\tilde{R} \\ &= A(I - A^{-1}(A - \tilde{L}\tilde{R})) \end{aligned}$$

Dann liefert die Neumannsche Reihe (Lemma 4.2)

$$\begin{aligned} \|(\tilde{L}\tilde{R})^{-1}\| &\leq \|A^{-1}\| \cdot \|I - A^{-1}(A - \tilde{L}\tilde{R})^{-1}\| \\ &\leq \frac{\|A^{-1}\|}{1 - \|A^{-1}(A - \tilde{L}\tilde{R})\|} \\ &\leq \frac{\|A^{-1}\|}{1 - \|A^{-1}\| \|A - \tilde{L}\tilde{R}\|} \\ &= \frac{\|A^{-1}\|}{1 - \text{cond}(A) \frac{\|A - \tilde{L}\tilde{R}\|}{\|A\|}} \end{aligned}$$

Damit erhält man mit

$$(\star) = (\tilde{L}\tilde{R})^{-1}(\tilde{L}\tilde{R} - A)(x_0 - x) + (\tilde{L}\tilde{R})^{-1}(A - \hat{A})x_0$$

dass

$$\frac{\|x_1 - x\|}{\|x\|} \approx \text{cond}(A) \left(\underbrace{\frac{\|A - \tilde{L}\tilde{R}\|}{\|A\|}}_{\varepsilon} \underbrace{\frac{\|x_0 - x\|}{\|x\|}}_{\text{cond}(A) \cdot \varepsilon} + \underbrace{\frac{\|A - \hat{A}\|}{\|A\|}}_{\leq \tilde{\varepsilon} \ll \varepsilon} \frac{\|x_0\|}{\|x\|} \right)$$

wenn $\text{cond}(A) \frac{\|A - \tilde{L}\tilde{R}\|}{\|A\|} \ll 1$

$$(\text{cond}(A))^2 \cdot \left(\varepsilon^2 + \tilde{\varepsilon} \cdot \frac{\|x_0\|}{\|x\|} \right)$$

x_1 ist nicht die exakte Lösung, deswegen wendet man die Nachiteration iterativ an. In der Praxis wird der Faktor in x nach wenigen Korrekturschritten (2-3) auf die Größenordnung der Defektgleichung reduziert, d.h. oft hat man

$$\frac{\|x_3 - x\|}{\|x\|} \sim \tilde{\varepsilon}$$

Beispiel 4.7: Betrachtet wird das Gleichungssystem

$$\begin{pmatrix} 1.05 & 1.02 \\ 1.04 & 1.02 \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \end{pmatrix} = \begin{pmatrix} 12 \\ 12 \end{pmatrix}$$

Exakte Lösung

$$x = (-100, 103.9216)$$

LR -Zerlegung mit 3-stelliger Genauigkeit:

$$\begin{aligned} \tilde{L} &= \begin{pmatrix} 1 & 0 \\ 0.99 & 1 \end{pmatrix} & \tilde{R} &= \begin{pmatrix} 1.05 & 1.02 \\ 0 & 0.01 \end{pmatrix} \\ \tilde{L}\tilde{R} - A &= \begin{pmatrix} 0 & 0 \\ -5 \cdot 10^{-4} & -2 \cdot 10^{-4} \end{pmatrix} \checkmark \\ &\Rightarrow \|\tilde{L}\tilde{R} - A\| \leq \text{eps} \end{aligned}$$

Näherungslösung $x_0 = (-97.2, 101)$ zugehöriges Residuum.

$$d_0 = b - Ax_0 = \begin{cases} (0, 0)^T & \text{3-stellige Rechnung} \\ (-0.17, -0.14)^T & \text{6-stellige Rechnung} \end{cases}$$

Bei 3-stelliger Rechnung hat die Korrekturgleichung

$$\begin{pmatrix} 1 & 0 \\ 0.99 & 1 \end{pmatrix} \begin{pmatrix} 1.05 & 1.02 \\ 0 & 0.01 \end{pmatrix} \begin{pmatrix} w_0^1 \\ w_0^2 \end{pmatrix} = \begin{pmatrix} -0.17 \\ -0.14 \end{pmatrix}$$

die Lösung $w_1 = (-2.90, 2.83) \Rightarrow x_1 = x_0 + w_1 = \begin{pmatrix} -99.9 \\ 104 \end{pmatrix}$ ist eine deutlich bessere Lösung!

Vorsicht: Selbst, wenn die LR -Zerlegung mit Spaltenpivotisierung stabil ist, d.h. $\|PA - \tilde{L}\tilde{R}\|$ klein, kann die numerische Lösung \tilde{x} sehr ungenau sein, wenn $\text{cond}(A)$ sehr groß ist. Dann können kleine Störungen zu A zu großen Störungen in x führen.

Abhilfe: Vorkonditionierung

$$Ax = b \xrightarrow{\tilde{P}} \tilde{P}Ax = \tilde{b}$$

Vorkonditionierung, so dass $\text{cond}(\tilde{P}A) \ll \text{cond}(A)$

Beispiel 4.8: Hilbertmatrix

Die Hilbertmatrix ist definiert durch $A = \left(\frac{1}{i+j-1}\right) \in \mathbb{R}^{n \times n}$. Man kann zeigen: $\text{cond}(A)$ wächst exponentiell in n .

5. Die QR -Zerlegung

Jetzt: $A \in \mathbb{R}^{m \times n}$, $m \geq n$, $\text{rang}(A) = n$

$A \in \mathbb{C}^{m \times n}$ genauso möglich, dann symmetrisch \leadsto hermitesch

Ziel: Faktorisierung $A = QR$ mit R = rechte obere Dreiecksmatrix, $R \in \mathbb{R}^{m \times n}$, $Q \in \mathbb{R}^{m \times m}$ unitär, d.h. $Q^T Q = I_n$.

5.1. Householder-Transformationen

Alston Householder (1904-1993, 1958)

Eigenschaften unitärer Matrizen: Sei $Q \in \mathbb{R}^{m \times m}$ unitär, dann gilt

$$1. \quad \|Qx\|_2^2 = (Qx)^T Qx = x^T Q^T Qx = x^T x = \|x\|_2^2$$

$\Rightarrow Q$ normerhaltend

$$\|Qx\|_2 = \|x\|_2 \quad \forall x \in \mathbb{R}^m$$

$$2. \quad \|Q\|_2 = \max_{\|x\|_2=1} \|Qx\|_2 = 1$$

$$\|Q^{-1}\|_2 = \|Q^T\|_2 = \|Q\|_2 = 1$$

$$\text{cond}_2(Q) \leq \|Q\|_2 \|Q^{-1}\|_2 = 1$$

$$3. \quad P, Q \in \mathbb{R}^{m \times m} \text{ unitär} \Rightarrow P \cdot Q \text{ unitär}$$

Definition 5.1: Householder-Transformation

Sei $v \in \mathbb{R}^m \setminus \{0\}$. Die Matrix $P_v = I_m - \frac{2}{\|v\|_2^2} vv^T \in \mathbb{R}^{m \times m}$ heißt **Householder-Transformation**.

Lemma 5.2: Sei $v \in \mathbb{R}^m \setminus \{0\}$. Dann ist P_v eine symmetrische unitäre mit $P_v v = -v$ und für alle $w \in \mathbb{R}^m$ mit $w \perp v$, d.h. $w^T v = 0$, gilt $P_v w = w$.

Beweis:

Symmetrie: klar

$$P_v^T P_v \stackrel{?}{=} I_m$$

$$\begin{aligned}
P_v^T P_v &= \left(I_m - \frac{2}{\|v\|_2^2} v v^T \right)^T \left(I_m - \frac{2}{\|v\|_2^2} v v^T \right) \\
&= I_m - \frac{4}{\|v\|_2^2} v v^T + \frac{4}{\|v\|_2^2} \frac{v v^T v v^T}{\|v\|_2^2} \\
&= I_m
\end{aligned}$$

Für v erhält man:

$$P_v v = \left(I_m - \frac{2}{\|v\|_2^2} v v^T \right) v = v - \frac{2}{\|v\|_2^2} \overbrace{v v^T v}^{\|v\|_2^2} = -v$$

sowie für $w \in \mathbb{R}^m$ mit $w \perp v$

$$P_v w = \left(I_m - \frac{2}{\|v\|_2^2} v v^T \right) w = w - \frac{2}{\|v\|_2^2} \underbrace{v v^T w}_0 = w$$

□

Damit kann P_v als Spiegelung interpretiert werden.

Nun Einsatz von Householder-Transformationen um A auf die Dreiecksgestalt zu bringen.
Dazu:

Lemma 5.3: Gegeben sei $x \in \mathbb{R}^n \setminus \{0\}$. Für $v = x + \tau e_1$ mit

$$\tau = \begin{cases} \pm \frac{x_1}{\|x\|_2} \|x\|_2 & \text{falls } x_1 \neq 0 \\ -\|x\|_2 & \text{falls } x_1 = 0 \end{cases}$$

gelte $v \neq 0$. Dann ist

$$P_v x = \left(I_m - \frac{2}{\|v\|_2^2} v v^T \right) x = -\tau e_1$$

Beweis:

$$\begin{aligned}
\|x + \tau e_1\|_2^2 &= \|x\|_2^2 + 2\tau e_1^T x + \tau^2 \\
&= \underbrace{2(x + \tau e_1)^T x}_v
\end{aligned}$$

Dann gilt: $2v^T x = 2(x + \tau e_1)^T = \|x + \tau e_1\|^2 = \|v\|_2^2$

Mit der Definition von v folgt weiterhin

$$\frac{2}{\|v\|_2^2} v(v^T x) = v = x + \tau e_1 \implies \text{Behauptung}$$

□

Bemerkung: Damit im Fall $x_1 \neq 0$ ($x_1 \in \mathbb{C}$) bei der Berechnung von v keine Auslöschung auftritt, kann man das Vorzeichen von τ entsprechend wählen.

5.2. Berechnung der QR-Zerlegung

Satz 5.4: Existenz einer QR-Zerlegung

Sei $A \in \mathbb{R}^{m \times n}$ mit $m \geq n$ mit $\text{rang}(A) = n$. Dann existiert eine unitäre Matrix $Q \in \mathbb{R}^{m \times m}$ und eine obere Dreiecksmatrix $R \in \mathbb{R}^{m \times n}$ mit

$$A = Q \cdot R$$

so dass $v_{ii} \neq 0, i = 1, \dots, n$.

Beweis:

Idee: Nutze Householder-Transformationen, um die Spalten von R zu erhalten. D.h. $Q_n \cdot \dots \cdot A = R$ mit $Q_i \hat{=}$ Householder-Transformationen, dann $Q := Q_1^T Q_2^T \dots Q_n^T = Q_1 \dots Q_n$.

1. Schritt: $A_1 = A, 0 \neq x = a_1 \hat{=}$ 1 Spalte von A . $Q_1 \in \mathbb{R}^{m \times n}$ Householder-Transformation mit v gemäß Lemma 5.3.

$$\implies Q_1 x = Q a_1 = v_{11} e_1$$

mit $|v_{11}| = \|a_1\| \neq 0$ da 1 Vollrang besitzt.

Also

$$Q_1 A = \begin{pmatrix} v_{11} & v_{12} & \dots & v_{1m} \\ 0 & & & \\ \vdots & & A_2 & \\ 0 & & & \end{pmatrix}, \quad A_2 \in \mathbb{R}^{(m-1) \times (n-1)}$$

2. Schritt: $x = a_2 =$ 1 Spalte von $A_2, x \in \mathbb{R}^{m-1}$ und $\tilde{Q}_2 \in \mathbb{R}^{(m-1) \times (n-1)}$ als Householder-Transformation gemäß Lemma 5.3. Dann folgt mit $Q_2 := \begin{pmatrix} 1 & 0 \\ 0 & \tilde{Q}_2 \end{pmatrix}$ unitär, dass

$$Q_2 Q_1 A = \begin{pmatrix} v_{11} & v_{12} & \dots & \dots & v_{1m} \\ 0 & v_{22} & v_{23} & \dots & v_{2m} \\ \vdots & 0 & & & \\ \vdots & \vdots & & A_3 & \\ 0 & 0 & & & \end{pmatrix}$$

mit $v_{22} \neq 0$, da A Vollrang hat.

Nach n Schritten erhält man die gewünschte Zerlegung

□

Bemerkungen:

- “Naive” Householder-Transformation, d.h. Aufstellen der Matrix und Anwenden, erfordert $\mathcal{O}(m^2n)$ Multiplikationen.
- Besser $P_1 A = A - \frac{2}{\|v\|_2^2} v v^T A = A - \frac{2}{\|v\|_2^2} v w^T$
 $\implies \mathcal{O}(mn)$ Multiplikationen