

Skript Numerik I

von Prof. Dr. Luise Blank im WS14/15

Gesina Schwalbe

19. November 2014

Inhaltsverzeichnis

| | | |
|----------|---|-----------|
| 1 | Einführung | 3 |
| 1.1 | Anwendungen | 4 |
| 1.2 | „Gute“ Programme | 7 |
| 2 | Lineare Gleichungssysteme: Direkte Methoden | 9 |
| 2.1 | Gaußsches Eliminationsverfahren | 9 |
| 2.1.1 | Vorwärtselimination | 9 |
| 2.1.2 | Rückwärtselimination | 11 |
| 2.1.3 | Vorsicht | 12 |
| 2.1.4 | Weitere algorithmische Anmerkungen | 12 |
| 2.1.5 | Dreieckszerlegung | 13 |
| 2.1.6 | Vorwärtssubstitution | 13 |
| 2.1.7 | Gauß-Elimination zur Lösung von $Ax = b$ | 13 |
| 2.1.8 | Rechenaufwand gezählt in „flops“ | 13 |
| 2.1.9 | Definition: Landau-Symbole | 14 |
| 2.1.10 | Allgemeines zur Aufwandsbetrachtung | 15 |
| 2.1.11 | Formalisieren des Gauß-Algorithmus | 15 |
| 2.1.12 | Lemma (Eigenschaften der L_k -Matrizen) | 16 |
| 2.1.13 | Satz (LR- oder LU-Zerlegung) | 16 |
| 2.2 | Gaußsches Eliminationsverfahren mit Pivotisierung | 18 |
| 2.2.1 | Spaltenpivotisierung (=partielle/ halbmaximale Pivotisierung) . . | 18 |
| 2.2.2 | Bemerkung | 18 |
| 2.2.3 | Gauß-Elimination mit Spaltenpivotisierung | 20 |
| 2.2.4 | Satz: Dreieckszerlegung mit Permutationsmatrix | 20 |
| 2.2.5 | Lösen eines Gleichungssystems $Ax = b$ | 22 |
| 2.2.6 | Bemerkungen | 22 |
| 2.2.7 | Beispiel zur Pivotisierung | 22 |
| 3 | Fehleranalyse | 23 |
| 3.1 | Zahlendarstellung und Rundungsfehler | 23 |
| 3.1.1 | Definition: Gleitkommazahl | 24 |
| 3.1.2 | Bemerkung | 24 |
| 3.1.3 | Beispiel | 24 |
| 3.1.4 | Verteilung der Maschinenzahlen | 25 |
| 3.1.5 | Bezeichnungen | 26 |
| 3.1.6 | Rundungsfehler | 26 |

| | | |
|----------|---|-----------|
| 3.1.7 | Bemerkung | 27 |
| 3.1.8 | Auslöschung von signifikanten Stellen | 28 |
| 3.2 | Kondition eines Problems | 28 |
| 3.2.1 | Definition: Problem | 28 |
| 3.2.2 | Definition: absoluter und relativer Fehler | 29 |
| 3.2.3 | Wiederholung: Normen | 29 |
| 3.2.4 | Definition: Matrixnorm | 29 |
| 3.2.5 | Definition: Frobeniusnorm, p-Norm, Verträglichkeit | 30 |
| 3.2.6 | Bemerkungen | 30 |
| 3.2.7 | Definition: absolute und relative normweise Kondition | 31 |
| 3.2.8 | Lemma | 31 |
| 3.2.9 | Beispiel: Kondition der Addition | 32 |
| 3.2.10 | Beispiel: Lösen eines Gleichungssystems | 33 |
| 3.2.11 | Definition: Kondition einer Matrix | 35 |
| 3.2.12 | Lemma (Neumannsche Reihe) | 36 |
| 3.2.13 | Bemerkung | 37 |
| 3.2.14 | Beispiel: Kondition eines nichtlinearen Gleichungssystems | 37 |
| 3.2.15 | Beispiel | 38 |
| 3.2.16 | Definition: Komponentenweise Kondition | 39 |
| 3.2.17 | Lemma | 39 |
| 3.2.18 | Beispiel | 40 |
| 3.3 | Stabilität von Algorithmen | 41 |
| 3.3.1 | Bemerkung | 41 |
| 3.3.2 | Beispiel | 42 |
| 3.3.3 | Definition: Elementar ausführbar | 43 |
| 3.3.4 | Definition: Algorithmus, Implementation | 43 |
| 3.3.5 | Lemma (Fehlerfortpflanzung) | 43 |
| 3.3.6 | Korollar | 45 |
| 3.3.7 | Bemerkung | 45 |
| 3.3.8 | Bemerkung zur Sprechweise | 45 |
| 3.3.9 | Beispiel | 46 |
| 3.3.10 | Bemerkung | 47 |
| 3.3.11 | Satz | 47 |
| 3.3.12 | Satz (Sautter 1971) | 47 |
| 3.3.13 | Allgemeine Faustregeln für die LR-Zerlegung | 48 |
| 3.4 | Beurteilung von Näherungslösungen linearer GLS | 48 |
| 3.4.1 | Satz (Prager und Oettli, 1964) | 49 |
| 4 | Lineare Gleichungssysteme: Direkte Methoden (Fortsetzung) | 51 |
| 4.1 | Gaußsches Eliminationsverfahren mit Äquilibration und Nachiteration . | 51 |
| 4.1.1 | Äquilibration der Zeilen | 51 |
| 4.1.2 | Äquilibration der Spalten | 51 |
| 4.1.3 | Lemma | 52 |
| 4.1.4 | Nachiteration | 52 |

| | | |
|----------|--|-----------|
| 4.1.5 | Bemerkung (nach Skeel 1980) | 52 |
| 4.2 | Cholesky-Verfahren | 52 |
| 4.2.1 | Satz (Eigenschaften von symm., pos. def. Matrizen) | 53 |
| 4.2.2 | Folgerung | 54 |
| 4.2.3 | Cholesky-Zerlegung | 55 |
| 4.2.4 | Rechenaufwand in flops | 55 |
| 4.2.5 | Bemerkung | 55 |
| 4.3 | Lineare Ausgleichsprobleme | 55 |
| 4.3.1 | Beispiel | 55 |
| 4.3.2 | Lineares Ausgleichsproblem | 57 |
| 4.3.3 | Projektionssatz | 58 |
| 4.3.4 | Satz | 59 |
| 4.3.5 | Lösung der Normalgleichung | 60 |
| 4.3.6 | Satz | 60 |
| 4.3.7 | Bemerkung | 61 |
| 4.4 | Orthogonalisierungsverfahren | 62 |
| 4.4.1 | Givens-QR-Algorithmus | 65 |
| 4.4.2 | Bemerkungen | 65 |
| 4.4.3 | Aufand des Givens-QR-Algorithmus | 66 |
| 4.4.4 | Bemerkung | 67 |
| 4.4.5 | Speicherung | 69 |
| 4.4.6 | Aufwand für den Householder-QR-Algorithmus | 69 |
| 5 | Numerische Lösung nichtlinearer Gleichungssysteme | 71 |
| 5.1 | Einführung | 71 |
| 5.1.1 | Beispiele | 71 |
| 5.1.2 | Das Bisektionsverfahren | 72 |
| 5.1.3 | Korollar | 73 |
| 5.1.4 | Bemerkungen | 73 |
| 5.2 | Fixpunktiteration | 73 |
| 5.2.1 | Beispiel | 73 |
| 5.2.2 | Definition: Kontraktion | 74 |
| 5.2.3 | Lemma | 74 |
| 5.2.4 | Banachscher Fixpunktsatz | 75 |
| 5.2.5 | Bemerkung | 77 |
| 5.2.6 | Folgerungen | 77 |
| 5.3 | Konvergenzordnung und Fehlerabschätzungen | 77 |
| 5.3.1 | Definition: lineare Konvergenz | 77 |
| 5.3.2 | Bemerkung | 77 |
| 5.3.3 | Bemerkung | 78 |
| 5.3.4 | Definition: lokale und globale Konvergenz | 78 |
| 5.3.5 | Lemma | 78 |
| 5.3.6 | Folgerung: a posteriori Fehlerabschätzung | 79 |
| 5.3.7 | Bemerkung | 79 |

Inhaltsverzeichnis

| | | |
|------------------|---|-----------|
| 5.3.8 | Folgerung | 80 |
| 5.4 | Newton-Verfahren für skalare Gleichung | 80 |
| 5.4.1 | Iterationsschritt des Newton(-Kantorowitsch)-Verfahrens | 80 |
| 5.4.2 | Satz | 81 |
| 5.4.3 | Bemerkung | 81 |
| Literatur | | 85 |

Vorwort

Skriptfehler

An alle, die gedenken dieses Skript zur Numerikvorlesung im WS2014/15 zu nutzen:
Es wird keinerlei Anspruch auf Richtigkeit, Vollständigkeit und auch sicher nicht Schönheit (ich bin LaTeX-Anfänger) dieses Dokuments erhoben.

Ihr würdet mir aber unglaublich weiterhelfen, wenn ihr jede Anmerkung – das kann alles, von groben inhaltlichen Fehlern über Rechtschreibkorrekturen bis hin zu Wünschen/Anregungen/Tipps zur Typografie, sein – an mich weiterleitet!

Jegliche Anmerkungen bitte gleich und jederzeit an:

| |
|--|
| gesina.schwalbe@stud.uni-regensburg.de |
|--|

oder auch an gesina.schwalbe@gmail.com

Copyright

Was das Rechtliche angeht bitte beachten:
Urheber dieses Skriptes ist Prof. Dr. Luise Blank.
Dies ist nur eine genehmigte Vorlesungsmitschrift und unterliegt dem deutschen Urheberrecht, jegliche nicht rein private Verwendung muss demnach vorher mit Frau Blank abgesprochen werden.

Bilder oder „IMAGE MISSING“

Leider habe ich mich bisher noch nicht in die (ordentliche) Grafikerstellung in LaTeX-Dokumenten eingearbeitet, weshalb das Skript erstmal nicht mit Grafiken dienen kann – die Fehlstellen sind mit „IMAGE MISSING“ markiert.

Wenn ihr gerne Veranschaulichungen haben möchtet, könnt ihr jederzeit **die entsprechenden Bilder an mich schicken**, bitte mit Kapitelangabe. (Oder mich explizit

Inhaltsverzeichnis

darum bitten, dass ich mich übergangsweise um das Einscannen, Bearbeiten etc. kümmer(e). Dann werden sie an die entsprechenden Stellen eingebunden.

Erscheinungsdatum

Ich werde mich bemühen, das Skript jeweils am Vorlesungstag zumindest in unverbesserter Form online zu stellen, so dass v.a. diejenigen, die die Vorlesung nicht besuchen können, einen Überblick über den Stoff bekommen.

Innerhalb einer Woche sollte das Skript aktuell sein.

Anfangsteil

Nachdem ich nicht seit Semesterbeginn mittex, fehlt noch ein Großteil des ersten Kapitels. Ich hoffe, es bis Mitte des Semesters nachholen zu können, aber keine Garantie.

Ziel ist, dass es vor der Vorlesungszeit integriert ist.

06.10.2014

1 Einführung

Wozu?

Oft sind Probleme mit der gleichen Struktur zu lösen, z.B.

- $ax^2 + bx + c = 0 \Rightarrow x_{\pm} = -\frac{b}{2a} \pm \frac{1}{2a} \sqrt{b^2 - 4ac}$
- Bestimmung des größten gemeinsamen Teilers zweier Zahlen
 \leadsto euklidischer Algorithmus

Hierfür ist ein allgemeiner Algorithmus erwünscht.

Ein weiterer Fall ist, dass ein Problem zwar analytisch gelöst werden kann, es aber zu lange dauert bis das Ergebnis bestimmt ist, z.B. tauchen bei der numerischen Simulation von Strömungen bis zu 1 Million Unbekannte auf. Damit sind Systeme mit $\approx 10^6$ Gleichungen zu lösen, wofür effiziente Algorithmen notwendig sind. Näherungslösungen sind bei solchen Problemen häufig ausreichend.

Es gibt auch Probleme, die nicht analytisch gelöst werden können, z.B. bei Differentialgleichungen ist vielleicht Existenz- und Eindeutigkeit gewährleistet, aber keine konstruktive Methode zur Berechnung bekannt. Hierfür ist dann eine Näherungslösung gefragt.

Geschichte

Algorithmen gibt es schon lange bevor es Rechner gab, z.B. den euklidischen Algorithmus seit um 300 v.Chr.

Die Fragestellungen und der Blickwinkel verschieben sich jedoch in Abhängigkeit von den zu lösenden Problemen und der existierenden Computer-Hardware.

(Strömungsprobleme modelliert mit partiellen Differentialgleichungen, es stehen Parallelrechner, Vektorrechner, größere Speicher zur Verfügung, etc.)

1.1 Anwendungen

- Herd, Waschmaschine, Heizungsanlage
- Handy (über welchen Satelliten wird übertragen), Digitalkamera, MP3-Player
- Navigationssysteme
- Erstellung des Zugfahrplans
- Robotersteuerung (bis hin zu Roboterfußball)
- Fahrzeugindustrie
 - Fahrzeugbau (Crash-Simulation, Strömungsmodellierung)
 - Fahrzeugsteuerung
- Finanzmarkt z.B. Risikoanalyse im Wertpapierhandel
- Klimaaanalyse z.B. Vorhersage von Erdbeben, Hurrikans, Überflutungen
- Medizinische Versorgung z.B. Bildverarbeitung, Prognose für Epidemieentwicklungen, Beschreibung des Blutkreislaufs
- Raffinerie-Industrie
- Kontrolle und Optimierung chemischer und biologischer Prozesse, z.B. Regelung von Wärmezufuhr
- u.s.w.

Fragestellungen

Rechengeschwindigkeit, Rechenaufwand (Anzahl der Rechenoperationen, Rechenzeit auf welchem Rechner...), Komplexität des Algorithmus

Beispiel 1

Berechnung der Lösung eines Gleichungssystems

$\mathbf{Ax} = \mathbf{b}$ mit einer $n \times n$ -Matrix \mathbf{A}

1. Cramersche Regel

$$x_j = \frac{\det(A)_j}{\det A} \text{ (ersetze die } j\text{-te Spalte von } A \text{ durch } b)$$

$$\text{und } \det A = \sum_{\pi} \text{sign}(\pi) a_{1m_1} \cdot \dots \cdot a_{nm_n}$$

Benötigt etwa **$n!$ Multiplikationen und Additionen**. Bei einer 20×20 -Matrix \mathbf{A} (was heutzutage klein ist) wären dies $\approx 2,5 \cdot 10^{18}$ Operationen. Falls jede arithmetische Operation 10^{-6} Sekunden (also

eine Mikrosekunde) benötigt, ist eine Rechenzeit von mehr als **eine Millionen Jahre** nötig!

2. Gaußsches Eliminationsverfahren

Benötigt etwa n^3 Operationen, also ungefähr 8000 Operationen und weniger als **0,005 Sekunden** (Golub, Ortega: Scientific Computing).

3. Verhalten bei Störungen, Stabilität des Verfahrens (Eingabefehler, Rundungsfehler, Diskretisierungsfehler)

Beispiel

$\frac{1}{10^{-8}} = 10^8$ Störung des Nenners $\frac{1}{2 \cdot 10^{-8}} = 5 \cdot 10^7$

\leadsto kleine Störung im Nenner kann zu großen Störungen im Ergebnis führen.

Beispiel

$$x^2 + 314x - 2 = 0$$

Falls diese Gleichung mit der **p,q-Formel** (Mitternachtsformel) gelöst wird und immer auf 5 signifikante Stellen gerundet wird, ergibt sich ein

$$\text{relativer Fehler} = \frac{|\text{Fehler}|}{|\text{Lösung}|}$$

von $\approx 57\%$.

Eine geschickte Formatierung liefert ein Ergebnis mit einem relativen Fehler von $\approx 1,5 \cdot 10^{-5}$, d.h. die ersten **4 Stellen sind exakt**. Dabei werden für $ax^2 + bx + c = 0$ folgende Ausdrücke verwendet:

$$\begin{aligned} x_1 &= \frac{1}{2a} (-b - \text{sign}(b) \sqrt{b^2 - 4ac}) \\ x_2 &= \frac{2c}{-b - \text{sign}(b) \sqrt{b^2 - 4ac}} \end{aligned}$$

Lösung 0.0063693, p,q-Formel 0.01, letzte Formel 0.0063692

Genauigkeit des Verfahrens, Fehleranalyse, Konvergenzgeschwindigkeit, Konvergenzordnung

Beispiel

Numerische Approximation von Ableitungen

Für $f \in C^3(I)$ gilt die Taylor-Entwicklung:

$$f(x \pm h) = f(x) \pm h f'(x) + \frac{h^2}{2} f''(x) + R(x) \text{ mit } |R(x)| \leq ch^3$$

Vorwärtsgenommener Differenzenquotient

$$(D_h^+ f)(x) := \frac{f(x+h) - f(x)}{h} \approx f'(x),$$

$$\left| f'(x) - \frac{f(x+h) - f(x)}{h} \right| \leq ch,$$

konvergiert also mit linearer Abhängigkeit der Schrittweite h

Zentraler Differenzenquotient

$$(D_h^0 f)(x) := \frac{f(x+h) - f(x-h)}{2 \cdot h} \approx f'(x),$$

$$\left| f'(x) - \frac{f(x+h) - f(x-h)}{2 \cdot h} \right| \leq ch^2,$$

konvergiert mit quadratischer **Ordnung** bei **gleichem Aufwand**!

Some disasters attributable to bad numerical computing

(Last modified August 26, 1998 by Douglas N. Arnold, arnold@ima.umn.edu)

Have you been paying attention in your numerical analysis or scientific computation courses?

If not, it could be a costly mistake.

Here are some real life examples of what can happen when numerical algorithms are not correctly applied.

The Patriot Missile failure in Dharan, Saudi Arabia, on February 25, 1991 which resulted in 28 deaths, is ultimately attributable to poor handling of **rounding errors**.

The explosion of the Ariane 5 rocket just after lift-off on its maiden voyage off French Guiana, on June 4, 1996, was ultimately the consequence of a **simple overflow**.

The sinking of the Sleipner A offshore platform in Gandsfjorden near Stavanger, Norway, on August 23, 1991, resulted in a loss of nearly one billion dollars. It was found to be the result of **inaccurate finite element analysis**.

Weitere praxisrelevante Fragestellungen

- Nutze black box solver oder entwickle Lösungsmethode, welche auf das spezielle Problem angepaßt ist.
- Wie teuer ist die Implementierung?
(= wieviel Arbeitszeit)
- Ist der implementierte Algorithmus vielseitig einsetzbar?
(welche Problemklassen deckt er ab, welche Rechnerstruktur ist vorausgesetzt)

1.2 „Gute“ Programme

- zuverlässig (fehlerfrei)
- robust (z.B. behandeln Ausnahmesituationen und filtern ungeeignete Daten heraus)
- portierbar auf andere Rechenanlagen
- wartungsfreundlich (leicht zu ändern oder zu erweitern)
- gut dokumentiert
- ausgiebig getestet
soll in den Übungen trainiert werden

Eine Faustregel der numerischen Mathematik

Zu jedem noch so eleganten numerischen Verfahren gibt es ein Gegenbeispiel, für welches die Methode völlig versagt.

(Teubner Taschenbuch)

Welche Probleme werden hier behandelt?

1. Lineare Gleichungssysteme $Ax = b$

- „kleine“ bis „mittelgroße“ Matrizen
→ **direkte Methoden:** nach endlich vielen Schritten ist die **exakte Lösung** bis auf Rundungsfehler berechnet (z.B. Gauß-Elimination)
- strukturierte Matrizen

Symmetrie oder sogar Bandstruktur:

$$\begin{pmatrix} * & * & * & & 0 \\ * & * & * & * & \\ & \ddots & \ddots & \ddots & \ddots \\ & & * & * & * & * \\ 0 & & & * & * & * \end{pmatrix}$$

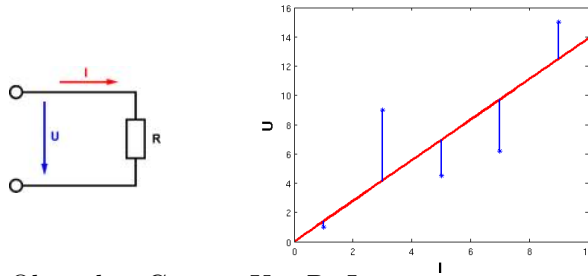
1 Einführung

- große Matrizen (mit zusätzlichen Eigenschaften)
→ **iterative Methoden**: kenne Startwert x_0 , berechne neue Approximation x_i unter Ausnutzung der vorherigen bis die **Näherungslösung** x_i „gut genug ist“.

2. Lineare Ausgleichsprobleme

Beispiel:

Wir messen den Zusammenhang zwischen Spannung U und Stromstärke I



Ohmsches Gesetz: $U = R \cdot I$

Gesucht ist der Widerstand R .

(U_i, I_i) seien die Messdaten mit möglichen Messfehlern.

Finde nun R , sodass $f(R) = \min_r \sum_i (U_i - r I_i)^2$

3. Lösung nichtlinearer Gleichungen, z.B.

- Berechnung von Nullstellen $g(x) = 0$
- Berechnung von Fixpunkten $f(x) = x$

4. Eigenwertwertberechnung $Ax = \lambda x$, $\lambda \in \mathbb{C}$

5. Interpolation

Setze Meßdaten zu einer kontinuierlichen Funktion fort,
aber wie „glatt“?

z.B. stückweise konstant, stückweise linear, oder falls sie eine Schwingung repräsentieren, berechne die zugehörige Fourierreihe

6. Berechnung von Integralen (Quadraturformeln)

Approximation von $\int_a^b f(x) dx$

Bei allem spielt die **Fehleranalyse** eine große Rolle und ihre Grundbegriffe werden in einem extra Abschnitt behandelt.

2 Lineare Gleichungssysteme: Direkte Methoden

Sei $A \in \mathbb{R}^{n \times n}$, $b \in \mathbb{R}^n$. Gesucht ist $x \in \mathbb{R}^n$ mit

$$A \cdot x = b$$

Weitere Voraussetzungen sind die Existenz und Eindeutigkeit einer Lösung. Bemerkung:

- Ein verlässlicher Lösungsalgorithmus überprüft dies und behandelt alle Fälle.
- Die Cramersche Regel ist ineffizient (s. Einführung).
- Das Inverse für $x = A^{-1} \cdot b$ aufzustellen ist ebenso ineffizient, denn es ist keine Lösung für alle $b \in \mathbb{R}^n$ verlangt und der Algorithmus wird evtl. instabil aufgrund vieler Operationen.

\Rightarrow Invertieren von Matrizen vermeiden!!

\Rightarrow Lösen des Linearen Gleichungssystems!!

2.1 Gaußsches Eliminationsverfahren

Das Verfahren wurde 1809 von Friedrich Gauß, 1759 von Joseph Louis Lagrange beschrieben und war seit dem 1. Jhd. v. Chr. in China bekannt.

2.1.1 Vorwärtselimination

Das Gaußverfahren gilt der Lösung eines linearen Gleichungssystems der Form

$$Ax = b$$

2 Lineare Gleichungssysteme: Direkte Methoden

mit $A = (a_{ij})_{i,j \leq n} \in K^{n \times n}$ Matrix und $b = (b_i)_{i \leq n} \in K^n$ Vektor.
Der zugehörige Algorithmus sieht folgendermaßen aus:

$$\begin{array}{cccccc} a_{11}x_1 & + & a_{12}x_2 & + & \cdots & + & a_{1n}x_n & = & b_1 \\ a_{21}x_1 & + & a_{22}x_2 & + & \cdots & + & a_{2n}x_n & = & b_2 \\ \vdots & & \vdots & & & & \vdots & & \vdots \\ a_{n1}x_1 & + & a_{n2}x_2 & + & \cdots & + & a_{nn}x_n & = & b_n \end{array}$$

\Downarrow

$$(i\text{-te Zeile}) - (1. \text{ Zeile}) \cdot \frac{a_{i1}}{a_{11}} \Rightarrow a_{i1} = 0$$

\Downarrow

$$\begin{array}{cccccc} a_{11}x_1 & + & a_{12}x_2 & + & \cdots & + & a_{1n}x_n & = & b_1 \\ & & + & a_{22}^{(1)}x_2 & + & \cdots & + & a_{2n}^{(1)}x_n & = & b_2^{(1)} \\ & & & \vdots & & & \vdots & & \vdots \\ & & & & & & a_{nn}^{(1)}x_n & = & b_n^{(1)} \end{array}$$

\Downarrow
 \vdots

mit

$$\begin{aligned} a_{ij}^{(1)} &= a_{ij} - a_{1j} \cdot \frac{a_{i1}}{a_{11}} && \text{für } i, j = 2, \dots, n \\ b_i^{(1)} &= b_i - b_1 \cdot \frac{a_{i1}}{a_{11}} && \text{für } i = 2, \dots, n \end{aligned}$$

08.10.2014

In jedem Schritt werden die Einträge der k -ten Spalte analog unterhalb der Diagonalen (also $k = 1, \dots, n-1$) eliminiert:

$$(i\text{-te Zeile}) - (k\text{-te Zeile}) \cdot \frac{a_{ik}}{a_{kk}} \quad \text{für } i = k+1, \dots, n$$

Die Reihe

$$A \rightarrow A^{(1)} \rightarrow A^{(2)} \rightarrow \dots \rightarrow A^{(n-1)}$$

wird bis zum n -ten Schritt fortgeführt, d.h. bis eine obere Dreiecksgestalt eintritt:

$$\underbrace{\begin{pmatrix} a_{11} & \cdots & \cdots & a_{1n} \\ & a_{22}^{(1)} & \cdots & a_{2n}^{(1)} \\ & & \ddots & \vdots \\ 0 & & & a_{nn}^{(n-1)} \end{pmatrix}}_{:=R} \cdot \begin{pmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{pmatrix} = \underbrace{\begin{pmatrix} b_1 \\ b_2^{(1)} \\ \vdots \\ b_n^{(n-1)} \end{pmatrix}}_{:=z}$$

$$Rx = z \tag{2.1.1}$$

wobei für $i = k + 1, \dots, n$ die Einträge wie folgt aussehen:

$$l_{ik} := \frac{a_{ik}^{(k-1)}}{a_{kk}^{(k-1)}} \quad (2.1.2)$$

$$a_{ij}^{(k)} = a_{ij}^{(k-1)} - a_{kj}^{(k-1)} \cdot l_{ik} \quad \text{für } j = k + 1, \dots, n \quad (2.1.3)$$

$$b_i^{(k)} = b_i^{(k-1)} - b_k^{(k-1)} \cdot l_{ik} \quad (2.1.4)$$

Dieser Prozess wird **Vorwärtselimination** genannt.

Der zugehörige Algorithmus ist:

```

for  $k = 1, \dots, n - 1$ 
|   for  $i = k + 1, \dots, n$ 
|   |    $l_{ik} = a_{ik} / a_{kk}$ 
|   |   for  $j = k + 1, \dots, n$ 
|   |   |    $a_{ij} = a_{ij} - l_{ik} a_{kj}$ 
|   |   end
|   |    $b_i = b_i - l_{ik} b_k$ 
|   end
end

```

2.1.2 Rückwärtselimination

Für die Lösung des Gleichungssystems ist dann noch die **Rückwärtssubstitution** nötig:

$$x_n = \frac{b_n^{(n-1)}}{a_{nn}^{(n-1)}} \quad (2.1.5)$$

$$x_{n-1} = \frac{b_{n-1}^{(n-2)} - a_{n-1,n}^{(n-1)} \cdot x_n}{a_{(n-1)(n-1)}^{(n-2)}} \quad (2.1.6)$$

$$x_k = \frac{b_k^{(k-1)} - \sum_{j=k+1}^n a_{kj}^{(k-1)} x_j}{a_{kk}^{(k-1)}} \quad (2.1.7)$$

Als Algorithmus:



Abbildung 2.1: Veranschaulichung der LR-Zerlegung

```

for  $k = n, n-1, \dots, 1$ 
|    $x_k = b_k$ 
|   for  $j = k+1, \dots, n$ 
|   |    $x_k = x_k - a_{kj}x_j$ 
|   end
|    $x_k = x_k / a_{kk}$ 
end
    
```

2.1.3 Vorsicht

Algorithmen 2.1.1 und 2.1.2 sind nur ausführbar, falls für die sog. **Pivotelemente** $a_{kk}^{(k-1)}$ gilt:

$$a_{kk}^{(k-1)} \neq 0 \quad \text{für } k = 1, \dots, n$$

Dies ist auch für invertierbare Matrizen nicht immer gewährleistet.

2.1.4 Weitere algorithmische Anmerkungen

Matrix A und Vektor b sollten möglichst **nie** überschrieben werden! (Stattdessen kann eine Kopie überschrieben werden.)

Das Aufstellen von A und b ist bei manchen Anwendungen das teuerste, sie gehen sonst verloren. In 2.1.1 wird das obere Dreieck von A überschrieben. Dies ist möglich, da in (2.1.3) nur die Zeilen $k+1, \dots, n$ mithilfe der k -ten bearbeitet werden. Am Ende steht R im oberen Dreieck von A und z in b .

Die l_{ik} werden spaltenweise berechnet und können daher anstelle der entsprechenden Nullen (in der Kopie) von A gespeichert werden, d.h.:

$$\tilde{L} := (l_{ik}) \tag{2.1.8}$$

und R werden sukzessive in A geschrieben. Der Vektor z und anschließend der Lösungsvektor x kann in (eine Kopie von) b geschrieben werden. Wird eine neue rechte Seite b betrachtet, muss 2.1.1 nicht komplett neu ausgeführt werden, da sich \tilde{L} nicht ändert. Es reicht 2.1.4 zu wiederholen.

2.1.5 Dreieckszerlegung

Die Dreieckszerlegung einer Matrix A entspricht dem Verfahren aus 2.1.1, nur ohne die Zeile (2.1.4).

2.1.6 Vorwärtssubstitution

Die Vorwärtssubstitution entspricht der in 2.1.4 bzw. dem Verfahren aus 2.1.1 ohne die Bestimmung von l_{ik} und R , also nur Schritt (2.1.4).

2.1.7 Gauß-Elimination zur Lösung von $Ax = b$

1 Dreieckszerlegung

2 Vorwärtssubstitution $b_i^{(k)} = b_i^{(k-1)} - b_k^{(k-1)} \cdot l_{ik}$

3 Rückwärtssubstitution $x_k = \frac{b_k^{(k-1)} - \sum_{j=k+1}^n a_{kj}^{(k-1)} x_j}{a_{kk}^{(k-1)}}$

2.1.8 Rechenaufwand gezählt in „flops“

„flops“ = floating point operations

1. Dreieckszerlegung

für $j = k + 1, \dots, n$ 1 Addition, 1 Multiplikation für a_{ij}

für $i = k + 1, \dots, n$ 1 Division zusätzl. für l_{ik}

Dies ist je für $k = 1, \dots, n - 1$, also ist die Zahl an Additionen und Multiplikationen

$$\begin{aligned} \sum_{k=1}^{n-1} (n-k)^2 &= \sum_{k=1}^{n-1} k^2 \\ &= \frac{(n-1)n(2n-1)}{6} \\ &= \frac{2n^3 - 3n^2 + n}{6}. \end{aligned}$$

Für große n sind das etwa $\frac{n^3}{3}$ Additionen und Multiplikationen und

$$\sum_{k=1}^{n-1} (n-k) = \frac{n^2 - n}{2} \approx \frac{n^2}{n}$$

2 Lineare Gleichungssysteme: Direkte Methoden

Divisionen.

Damit ergibt sich eine Gesamtanzahl an flops von

$$2 \cdot \frac{2n^3 - 3n^2 + n}{6} + \frac{n^2 - n}{2} = \frac{2}{3}n^3 - \frac{1}{2}n^2 - \frac{1}{6}n \approx \frac{2}{3}n^3$$

für große n .

2. Vorwärts- bzw. Rückwärtssubstitution

Hier ergeben sich je

$$\sum_{k=1}^{n-1} (n-k) = \frac{n^2 - n}{2} \approx \frac{n^2}{2}$$

Multiplikationen und Additionen sowie n Divisionen für die Rückwärtssubstitution und damit insgesamt

$$n^2 + n$$

flops.

Zusammenfassung

Die Dreieckszerlegung benötigt $\mathcal{O}(n^3)$ flops und die Vorwärts- bzw. Rückwärtssubstitution $\mathcal{O}(n^2)$ flops.

2.1.9 Definition: Landau-Symbole

Seien $f, g : D \rightarrow \mathbb{R}$, $D \subset \mathbb{R}$, $-\infty \leq a \leq \infty$ und $(a_n)_{n \in \mathbb{N}}, (b_n)_{n \in \mathbb{N}}$ Folgen in \mathbb{R} .

a) $f(x) = \mathcal{O}(g(x))$ für $x \rightarrow a$, falls

$$\exists U(a), c \in \mathbb{R} : \forall x \in U(a) : |f(x)| \leq c \cdot |g(x)|$$

(bzw. falls $\lim_{x \rightarrow a} \frac{|f(x)|}{|g(x)|} \leq c$)

b) $f(x) = o(g(x))$ für $x \rightarrow a$, falls $\lim_{x \rightarrow a} \frac{|f(x)|}{|g(x)|} = 0$

c) $a_n = \mathcal{O}(b_n)$ für $n \rightarrow \infty$, falls

$$\forall \varepsilon > 0 : \exists N \in \mathbb{N} : \forall n \geq N : |a_n| \leq \varepsilon |b_n|$$

2.1.10 Allgemeines zur Aufwandsbetrachtung

Die Anzahl der Rechenoperationen ist nicht immer ausschlaggebend für den Aufwand, da z.B.

Parallelrechner: In manchen Algorithmen sind Rechenschritte parallel ausführbar. Damit entspricht die Zeit nicht der Anzahl an Operationen und es wird zusätzlich „Kommunikationszeit“ benötigt.

Sortieralgorithmen: Die Indexverwaltung benötigt Zeit, aber keine/kaum Rechenoperationen

If-When-Abfragen: entsprechend

Rechenoperationen liefern jedoch oft eine gute Schätzung.

13.10.2014

2.1.11 Formalisieren des Gauß-Algorithmus

a) **Rückwärtssubstitution:** entspricht $Rx = b$

b) **Vorwärtssubstitution:**

$$b_i = b_i^{(k-1)} - l_{ik} b_k^{(k-1)} \quad i = k+1, \dots, n$$

$$b^{(k)} = b^{(k-1)} - l_k b_k^{(k-1)} \quad \text{mit } l_k := \begin{pmatrix} 0 \\ \vdots \\ 0 \\ l_{k+1,k} \\ \vdots \\ l_{n,k} \end{pmatrix}$$

Sei $l_k \in \mathbb{R}^n$ der k-te Einheitsvektor und

$$L_k := I - l_k e_k^T = \begin{pmatrix} 1 & & & & \\ 0 & \ddots & & & 0 \\ & & 1 & & \\ \vdots & -l_{k+1,k} & & 1 & \\ & \vdots & & & \ddots \\ & -l_{n,k} & & & & 1 \end{pmatrix} \quad (2.1.9)$$

dann gilt $b^{(k)} = L_k b^{(k-1)}$, also

$$z = L_{n-1} \cdot L_{n-2} \cdot \dots \cdot L_1 b$$

Sei

$$L := L_1^{-1} \cdot \dots \cdot L_{n-1}^{-1} \quad (2.1.10)$$

2 Lineare Gleichungssysteme: Direkte Methoden

Hiermit folgt dann, dass die Vorwärtssubstitution

$$Lz = b \quad (2.1.11)$$

entspricht.

c) **Dreieckszerlegung:** Wie für die Vorwärtssubstitution ergibt sich

$$A^{(k)} = L_k A^{(k-1)}$$

und somit $R = L_{n-1} A^{(n-2)} = L_{n-1} \dots L_1 A$ bzw.

$$A = L \cdot R \quad (2.1.12)$$

2.1.12 Lemma (Eigenschaften der L_k -Matrizen)

1. L_k ist eine Frobeniusmatrix, d.h. sie unterscheidet sich höchstens in einer Spalte von der Einheitsmatrix I .
2. $L_k^{-1} = I + l_k e_k^T$
3. Es gilt:

$$\begin{aligned} L &= L_1^{-1} \cdot \dots \cdot L_{n-1}^{-1} \\ &= I + \sum_{i=1}^{n-1} l_i e_i^T \\ &= \begin{pmatrix} 1 & & 0 \\ & \ddots & \\ l_{ij} & & 1 \end{pmatrix} \\ &= I + \tilde{L} \end{aligned} \quad (2.1.13)$$

Hiermit folgt:

2.1.13 Satz (LR- oder LU-Zerlegung)

Das obige Verfahren ((2.1.2) und (2.1.3)) erzeugt unter der Voraussetzung von nicht-nullwertigen Pivotelementen eine Faktorisierung

$$A = L \cdot R$$

wobei R eine obere Dreiecksmatrix und L eine untere, normierte Dreiecksmatrix ist, d.h. für $i = 1, \dots, n$ gilt $l_{ii} = 1$.

Weiterhin existiert zu jeder regulären Matrix höchstens eine solche Zerlegung.



Abbildung 2.2: Verfahren von Crout



Abbildung 2.3: Eckweises Verfahren zur LR-Zerlegung

Beweis. Zur Eindeutigkeit:

$$A = LR \Leftrightarrow a_{ij} = \sum_{k=1}^{\min j,k} l_{ik} r_{kj}$$

Da $l_{ii} = 1$ gilt, folgt

$$i \leq j: r_{ij} = a_{ij} - \sum_{k=1}^{i-1} l_{ik} r_{kj} \text{ und}$$

$$i > j: l_{ij} = \frac{1}{r_{jj}} \left(a_{ij} - \sum_{k=1}^{j-1} l_{ik} r_{kj} \right),$$

da mit A auch R regulär (=invertierbar) ist, und somit $\det(A) = \prod_{j=1}^n r_{jj} \neq 0$, also $r_{jj} \neq 0$ gilt.

Diese können rekursiv, also eindeutig, berechnet werden, wenn auch die Reihenfolge der Berechnung nicht eindeutig ist.

Mögliche Verfahren sind z.B. das **Verfahren von Crout**

oder eckweise

oder zeilenweise

oder wie in 2.1.1.

□



Abbildung 2.4: Zeilenweises Verfahren zur LR-Zerlegung

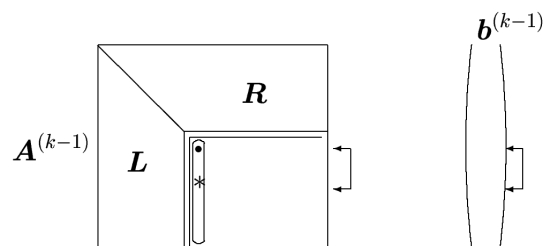
2.2 Gaußsches Eliminationsverfahren mit Pivotisierung

Beispiel Die Matrix $A = \begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix}$ ist invertierbar, aber die Gauß-Elimination versagt. Permutiere also die erste mit der zweiten Zeile und der Algorithmus wird anwendbar.

Allgemein Vermeide die Division durch betragsmäßig kleine Zahlen!

2.2.1 Spaltenpivotisierung (=partielle/ halbmaximale Pivotisierung)

Im k -ten Eliminationsschritt ist



1. Bestimme das Pivotelement $a_{pk}^{(k-1)}$ als betragsmäßig größtes der „Rest-Spalte“, d.h.

$$|a_{pk}^{(k-1)}| \geq |a_{jk}^{(k-1)}| \quad \text{für } j = k, \dots, n$$

2. Vertausche in $A^{(k-1)}$ die k -te mit der p -ten Zeile
3. Führe einen Gauß-Eliminationsschritt aus.

2.2.2 Bemerkung

- a) Hiermit gilt $|l_{jk}| \ll 1$.

- b) Anstelle von Spaltenpivotisierung kann eine **Zeilenpivotisierung** durchgeführt werden. Welche günstiger (in cpu-time) ist, hängt von der Rechnerarchitektur und der damit zusammenhängenden Umsetzung des Gauß-Algorithmus ab.
(Beispielsweise greifen Vektorrechner entweder auf die gesamte Spalte oder auf die gesamte Zeile einer Matrix zu und bevorzugen dementsprechend Operationen spalten- bzw. zeilenweise.)
- c) Der Aufwand enthält (bis auf $|\cdot|$) keine Rechenoperationen (flops), aber $\mathcal{O}(n^2)$ Vergleiche und Vertauschungen.
- d) Eine **vollständige Pivotsuche** sucht das betragsmäßig größte Element der gesamten Restmatrix und benötigt $\mathcal{O}(n^3)$ Vergleiche (sie wird so gut wie nie angewendet).

Damit die LR-Zerlegung unabhängig von der rechten Seite erstellt werden kann, müssen die Permutationen gespeichert werden. Hierfür verwendet man einen sog. **Permutationsvektor** Π , wobei

$$\Pi^{(k-1)}(r) = s$$

bedeutet, dass nach dem $(k-1)$ -ten Eliminationsschritt in der r -ten Zeile von $A^{(k-1)}$ die s -te bearbeitete Zeile von A steht, also

$$\begin{aligned}\Pi^{(k)}(k) &= \Pi^{(k-1)}(p) \\ \Pi^{(k)}(p) &= \Pi^{(k-1)}(k) \quad \text{und entsprechend} \\ \Pi^{(k)}(i) &= \Pi^{(k)}(i) \quad \text{für } i \neq k, p\end{aligned}$$

Für die **Permutationsmatrix**

$$P_{\Pi} = (e_{\Pi(1)}, \dots, e_{\Pi(n)}) \quad e_j := \begin{pmatrix} 0 \\ \vdots \\ 0 \\ 1 \\ 0 \\ \vdots \\ 0 \end{pmatrix} \leftarrow j\text{-te Stelle}$$

$$\text{mit } PA = LR$$

gilt

$$P^{-1} = P^T$$

und

$$\det P_{\Pi} = \text{sign}(\Pi) = \begin{cases} +1 & \text{falls } \Pi \text{ von gerader} \\ -1 & \text{falls } \Pi \text{ von ungerader} \end{cases} \quad \text{Anzahl an Transpositionen erzeugt wird}$$

2.2.3 Gauß-Elimination mit Spaltenpivotisierung

Der zugehörige Algorithmus zur Spaltenpivotisierung ist:

```

 $\pi(1:n) = [1:n]$ 
for  $k = 1, \dots, n-1$ 
|   bestimme Pivotzeile  $p$ , so dass
|        $|a_{pk}| = \max\{|a_{jk}|, j = k, \dots, n\}$ 
|    $\pi(k) \rightleftharpoons^1 \pi(p)$ 
|    $A(k, 1:n) \rightleftharpoons A(p, 1:n)$ 
|   if  $a_{kk} \neq 0$ 
|   |    $zeile = [k+1:n]$ 
|   |    $A(zeile, k) = A(zeile, k)/a_{kk}$ 
|   |    $A(zeile, zeile) = A(zeile, zeile) - A(zeile, k)A(k, zeile)$ 
|   else
|   |   „ $A$  ist singulär“
|   end
end

```

\rightleftharpoons bedeutet „vertausche mit“

2.2.4 Satz: Dreieckszerlegung mit Permutationsmatrix

Für jede invertierbare Matrix A existiert eine Permutationsmatrix P , so dass eine Dreieckszerlegung

$$PA = LR$$

existiert. P kann so gewählt werden, dass alle Elemente von L betragsmäßig kleiner oder gleich 1 sind, d.h.

$$|l_{ij}| \leq 1 \quad \forall i, j$$

Beweis. Da $\det A \neq 0$ ist, existiert eine Transposition τ_1 , s.d.

$$a_{11}^{(1)} = a_{\tau_1, 1} \neq 0$$

und

$$|a_{\tau_1, 1}| \geq |a_{i1}| \quad \forall i = 1, \dots, n.$$

2.2 Gaußsches Eliminationsverfahren mit Pivotisierung

Wir erhalten damit

$$L_1 P_{\tau_1} \cdot A = A^{(1)} = \begin{pmatrix} a_{11}^{(1)} & \cdots & \\ 0 & & \\ \vdots & & \\ 0 & & B^{(1)} \end{pmatrix}$$

und alle Elemente von L_1 sind betragsmäßig kleiner oder gleich 1 sowie $\det L_1 = 1$.
Daraus folgt

$$\begin{aligned} \det B^{(1)} &= \underbrace{\frac{1}{a_{11}^{(1)}}}_{\neq 0} \cdot \det A^{(1)} \\ &= \frac{1}{a_{\tau_1,1}^{(1)}} \cdot \det(L_1) \cdot \det(A) \\ &\neq 0. \end{aligned}$$

Also ist $B^{(1)}$ invertierbar.

Induktiv erhalten wir dann

$$R = A^{(n-1)} = L_{n-1} P_{\tau_{n-1}} \cdot \dots \cdot L_1 P_{\tau_1} \cdot A$$

Da τ_i nur zwei Zahlen $\geq i$ vertauscht, ist

15.10.2014

$$\Pi_i := \tau_{n-1} \circ \dots \circ \tau_i \quad \text{für } i = 1, \dots, (n-1)$$

eine Permutation der Zahlen $\{i, \dots, n\}$, d.h. insbesondere gilt:

$$\begin{aligned} \Pi_i(j) &= j && \text{für } j = 1, \dots, (i-1) \\ \Pi_i(j) &\in \{i, \dots, n\} && \text{für } j = i, \dots, n. \end{aligned}$$

$$P_{\Pi_{i+1}} = (e_1, \dots, e_i, e_{\Pi_{i+1}(i+1)}, \dots, e_{\Pi_{i+1}(n)}) = \begin{pmatrix} I_i & 0 \\ 0 & P_\sigma \end{pmatrix}$$

2 Lineare Gleichungssysteme: Direkte Methoden

Damit folgt:

$$\begin{aligned}
 P_{\Pi(i+1)} \cdot L_i \cdot P_{\Pi(i+1)}^{-1} &= P_{\Pi(i+1)} \cdot \left(\begin{array}{ccc|c} I_i & & & 0 \\ & -l_{i+1,i} & & \\ 0 & \vdots & & I_{n-i} \\ & -l_{n,i} & & \end{array} \right) \cdot \begin{pmatrix} I_i & 0 \\ 0 & P_{\sigma}^{-1} \end{pmatrix} \\
 &= \begin{pmatrix} I_i & 0 \\ 0 & P_{\sigma} \end{pmatrix} \cdot \dots \cdot \left(\begin{array}{ccc|c} I_i & & & 0 \\ & -l_{i+1,i} & & \\ 0 & \vdots & & P_{\sigma}^{-1} \\ & -l_{n,i} & & \end{array} \right) \\
 &= \left(\begin{array}{ccc|c} I_i & & & 0 \\ & -l_{\Pi(i+1)(i+1),i} & & \\ 0 & \vdots & & I_{n-i} \\ & -l_{\Pi(i+1)(n),i} & & \end{array} \right) \\
 &= I - (P_{\Pi(i+1)} l_i) e_i^T \\
 &=: \widehat{L}_i
 \end{aligned}$$

und

$$\begin{aligned}
 R &= L_{n-1} \\
 &\quad \cdot (P_{\tau_{n-1}} L_{n-2} P_{\tau_{n-1}}^{-1}) \\
 &\quad \cdot (P_{\tau_{n-1}} P_{\tau_{n-2}} L_{n-2} P_{\tau_{n-2}}^{-1} P_{\tau_{n-1}}^{-1}) \\
 &\quad \vdots \\
 &\quad \cdot (P_{\tau_{n-1}} \dots P_{\tau_1} L_1 P_{\tau_1} \dots P_{\tau_{n-1}}) \cdot A \\
 &= L_{n-1} \widehat{L}_{n-2} \dots \widehat{L}_1 P_{\Pi_1} \cdot A
 \end{aligned}$$

Nach Lemma 2.1.12 gilt daher, es existiert eine Permutation Π_1 mit

$$P_{\Pi_1} \cdot A = LR,$$

wobei R obere Dreiecksgestalt hat und

$$L = \begin{pmatrix} 1 & & & 0 \\ l_{\Pi_2(2),1} & \ddots & & \\ \vdots & \ddots & 1 & \\ l_{\Pi_n(n),1} & \dots & l_{\Pi_n(n),n-1} & 1 \end{pmatrix} \quad \text{mit } |l_{ij}| \leq 1$$

gilt. □

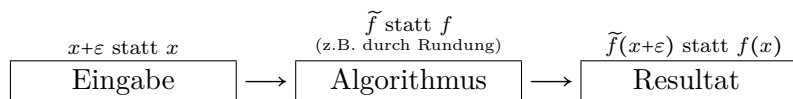
2.2.5 Lösen eines Gleichungssystems $Ax = b$

2.2.6 Bemerkungen

2.2.7 Beispiel zur Pivotisierung

15.10.2014

3 Fehleranalyse



Bei der Fehleranalyse liegt das Hauptaugenmerk auf

Eingabefehler

z.B. Rundungsfehler, Fehler in Messdaten, Fehler im Modell (falsche Parameter)

Fehler im Algorithmus

z.B. Rundungsfehler durch Rechenoperationen, Approximationen

(z.B. Ableitung durch Differenzenquotient oder die Berechnung von Sinus durch abgebrochene Reihenentwicklung)

1. *Frage* Wie wirken sich Eingabefehler auf das Resultat unabhängig vom gewählten Algorithmus aus?
2. *Frage* Wie wirken sich (Rundungs-)Fehler des Algorithmus aus?
Und wie verstärkt der Algorithmus Eingabefehler?

3.1 Zahlendarstellung und Rundungsfehler

Auf (Digital-)Rechnern können nur endlich viele Zahlen realisiert werden.

Die wichtigsten Typen sind:

- **ganze Zahlen** (integer):

$$z = \pm \sum_{i=0}^m z_i \beta_i \quad \text{mit} \quad \begin{array}{l} \beta = \text{Basis des Zahlensystems (oft } \beta = 2) \\ z_i \in \{0, \dots, \beta - 1\} \end{array}$$

- **Gleitpunktzahlen** (floating point)

3.1.1 Definition: Gleitkommazahl

Eine Zahl $x \in \mathbb{Q}$ mit einer Darstellung

$$x = \sigma \cdot (a_1.a_2 \dots a_t)_\beta \cdot \beta^e = \sigma \beta^e \cdot \sum_{\nu=1}^t a_\nu \beta^{-\nu+1}$$

| | |
|--|--|
| $\beta \in \mathbb{N}$ | Basis des Zahlensystems |
| $\sigma \in \{\pm 1\}$ | Vorzeichen |
| $m = (a_1.a_2 \dots a_t)_\beta$ $= \sum_{\nu=1}^t a_\nu \beta^{-\nu+1}$ | Mantisse |
| $a_i \in \{0, \dots, \beta - 1\}$ | Ziffern der Mantisse |
| $t \in \mathbb{N}$ | Mantissenlänge |
| $e \in \mathbb{Z}$ | mit $e_{\min} \leq e \leq e_{\max}$ Exponent |

heißt **Gleitkommazahl** mit t Stellen und Exponent e zur Basis β .

Ist $a_1 \neq 0$, so heißt x **normalisierte Gleitkommazahl**

3.1.2 Bemerkung

- a) 0 ist keine normalisierte Gleitkommazahl, da $a_1 = 0$ ist.
- b) $a_1 \neq 0$ stellt sicher, dass die Gleitkommadarstellung eindeutig ist.
- c) In der Praxis werden auch nicht-normalisierte Darstellungen verwendet.
- d) Heutige Rechner verwenden meist $\beta = 2$, aber auch $\beta = 8, \beta = 16$.

3.1.3 Beispiel

bit-Darstellung nach IEEE-Standard 754 von floating point numbers

Sei die Basis $\beta = 2$.

| | Speicherplatz | t | e_{\min} | e_{\max} |
|-------------------------------|-----------------|-----|------------|------------|
| einfache Genauigkeit (float) | 32bits = 4Bytes | 24 | -126 | 127 |
| doppelte Genauigkeit (double) | 64bits = 8Bytes | 52 | -1022 | 1023 |

Darstellung im Rechner (Bitmuster) für float:

$$\boxed{s \mid b_0 \dots b_7 \mid a_2 \dots a_{24}}$$

(Da $a_1 \neq 0$, also $a_1 = 1$ gilt, wird a_1 nicht gespeichert)

Interpretation ($s, b, a_i \in \{0, 1\} \forall i$)



Abbildung 3.1: Ungleichmäßige Verteilung der Maschinenzahlen im Dezimalsystem

- s Vorzeichenbit: $\sigma = (-1)^s \Rightarrow \begin{matrix} \sigma(0) = 1 \\ \sigma(1) = -1 \end{matrix}$
- $b = \sum_{i=0}^7 b_i \cdot 2^i \in \{1, \dots, 254\}$ speichert den Exponenten mit
 $e = b - \underbrace{127}_{\text{Basiswert}}$ (kein Vorzeichen nötig)
 Beachte: $b_0 = \dots = b_7 = 1$ sowie $b_0 = \dots = b_7 = 0$ sind bis auf Ausnahmen keine gültigen Exponenten
- $m = (a_1.a_2 \dots a_{24}) = 1 + \sum_{\nu=2}^{24} a_\nu 2^{1-\nu}$ stellt die Mantisse dar, $a_1 = 1$ wird nicht abgespeichert.
- Besondere Zahlen per Konvention:

$$x = 0: s \text{ bel.}, b = 0, m = 1 \quad \boxed{s \mid 0 \dots 0 \mid 0 \dots 0}$$

$$x = \pm\infty: s \text{ bel.}, b = 255, m = 1 \quad \boxed{s \mid 1 \dots 1 \mid 0 \dots 0}$$

$$x = \text{NaN} \quad s \text{ bel.}, b = 255, m \neq 1$$

$$x = (-1)^s \quad s \text{ bel.}, b = 0, m \neq 1 \text{ und } x \text{ hat die Form } x = (0 + \sum_{\nu=2}^{24} a_\nu \cdot 2^{1-\nu}) \cdot 2^{126} \text{ („denormalized“ number)}$$

20.10.2014

Betragsmäßig **größte Zahl**:

$$\boxed{0 \mid 01 \dots 1 \mid 1 \dots 1}$$

$$x_{\max} = (2 - 2^{-23}) \cdot 2^{127} \approx 3,4 \cdot 10^{38}$$

Betragsmäßig **kleinste Zahl**:

$$\boxed{0 \mid 0 \dots 0 \mid 0 \dots 01}$$

$$x_{\min} = (2 - 2^{-23}) \cdot 2^{-126} = 2^{-149} \approx 1,4 \cdot 10^{-45}$$

3.1.4 Verteilung der Maschinenzahlen

ungleichmäßig im Dezimalsystem, z. B.

$$x = \pm a_1.a_2a_3 \cdot 2^e \quad -2 \leq e \leq 1 \quad a_i \in \{0, 1\}$$

ist im Dualsystem gleichmäßig verteilt.

3.1.5 Bezeichnungen

overflow es ergibt sich eine Zahl, die betragsmäßig größer ist als die größte maschinendarstellbare Zahl

underflow entsprechend, betragsmäßig kleiner als die kleinste positive Zahl

Bsp.: overflow beim integer $b = e + 127$

$$\begin{array}{rcl} b & = & 254 \quad 11111110 \\ & + & 3 \quad 00000011 \\ b + 3 = 257 \bmod 2^8 & = & 1 \quad \overline{100000001} \end{array}$$

3.1.6 Rundungsfehler

Habe $x \in \mathbb{R}$ die normalisierte Darstellung

$$\begin{aligned} x &= \sigma \cdot \beta^e \left(\sum_{\nu=1}^t a_{\nu} \beta^{1-\nu} + \sum_{\nu=t+1}^{\infty} a_{\nu} \beta^{1-\nu} \right) \\ &= \sigma \cdot \beta^e \left(\sum_{\nu=1}^t a_{\nu} \beta^{1-\nu} + \beta^{1-t} \sum_{l=1}^{\infty} a_{t+l} \beta^{-l} \right) \end{aligned}$$

mit $e_{\min} \leq e \leq e_{\max}$, dann wird mit $fl(x)$ die gerundete Zahl bezeichnet, wobei $fl(x)$ eindeutig gegeben ist durch die Schranke an den **absoluten Rundungsfehler**

$$|fl(x) - x| \leq \begin{cases} \frac{1}{2} \beta^{e+1+t} & \text{bei symmetrischem Runden} \\ \beta^{e+1+t} & \text{bei Abschneiden} \end{cases}.$$

Für die **relative Rechengenauigkeit** folgt somit

$$\frac{|fl(x) - x|}{|x|} \leq \begin{cases} \frac{1}{2} \beta^{1-t} & \text{bei symmetrischem Runden} \\ \beta^{1-t} & \text{bei Abschneiden} \end{cases}.$$

Die **Maschinengenauigkeit** des Rechners ist daher durch

$$eps = \beta^{1-t} \quad (\text{für float} \approx 10^{-7}, \text{ für double} \approx 10^{-16})$$

gegeben.

Die Mantissenlänge bestimmt also die Maschinengenauigkeit. Bei einfacher Genauigkeit ist $fl(x)$ bis auf ungefähr 7 signifikante Stellen genau.

Im Folgenden betrachten wir symmetrisches Runden und definieren daher

$$\tau := \frac{1}{2} eps$$

Weiterhin gilt:



Abbildung 3.2: Eingabemenge einer Maschinenzahl

- a) Die kleinste Zahl am Rechner, welche größer als 1 ist, ist

$$1 + \epsilon$$

- b) Eine Maschinenzahl x repräsentiert eine Eingabemenge

$$E(x) = \{\tilde{x} \in \mathbb{R} : |\tilde{x} - x| \leq \tau|x|\}$$

3.1.7 Bemerkung

Gesetze der arithmetischen Operationen gelten i.A. nicht, z.B.

- x Maschinenzahl $\Rightarrow fl(x + \nu) = x$ für $|\nu| < \tau|x|$
- Assoziativ- und Distributivgesetze gelten nicht, z.B. für $\beta = 10$, $t = 3$, $a = 0,1$, $b = 105$, $c = -104$ gilt:

$$\begin{aligned} fl(a + fl(b + c)) &= 1,1 \\ fl(fl(a + b) + c) &= fl(fl(105, 1) + (-104)) \\ &= fl(105 - 104) \\ &= 1 \quad \nabla \end{aligned}$$

\Rightarrow Für einen Algorithmus ist die Reihenfolge der Operationen wesentlich! Mathematisch äquivalente Formulierungen können zu verschiedenen Ergebnissen führen.

3.1.8 Auslöschung von signifikanten Stellen

Sei $x = 9,995 \cdot 10^{-1}$, $y = 9,984 \cdot 10^{-1}$. Runde auf drei signifikante Stellen und berechne $x - y$:

$$\begin{aligned}\tilde{f}(x, y) &:= fl(fl(x) - fl(y)) = fl(1,00 \cdot 10^0 - 9,98 \cdot 10^{-1}) \\ &= fl(0,02 \cdot 10^{-1}) \\ &= fl(2,00 \cdot 10^{-3}) \\ f(x, y) &:= x - y \\ &:= 0,0011 = 1,1 \cdot 10^{-3}\end{aligned}$$

Daraus ergibt sich der relative Fehler

$$\frac{|\tilde{f}(x, y) - f(x, y)|}{|f(x, y)|} = \frac{|2 \cdot 10^{-3} - 1,1 \cdot 10^{-3}|}{|1,1 \cdot 10^{-3}|} = 82\%$$

Der Grund hierfür ist, dass das Problem der Subtraktion zweier annähernd gleich großer Zahlen schlecht konditioniert ist.

Zwei Regeln:

- 1) Umgehbare Subtraktion annähernd gleich großer Zahlen vermeiden!
- 2) Unumgängliche Subtraktion möglichst an den Anfang des Algorithmus stellen! (siehe später)

3.2 Kondition eines Problems

Es wird das Verhältnis

$$\frac{\text{Ausgabefehler}}{\text{Eingabefehler}}$$

untersucht.

3.2.1 Definition: Problem

Sei $f : U \subseteq \mathbb{R}^n \rightarrow \mathbb{R}^m$ mit U offen und sei $x \in U$. Dann bezeichne (f, x) das Problem, zu einem gegebenen x die Lösung $f(x)$ zu finden.

3.2.2 Definition: absoluter und relativer Fehler

Sei $x \in \mathbb{R}^n$ und $\tilde{x} \in \mathbb{R}^n$ eine Näherung an x . Weiterhin sei $\|\cdot\|$ eine Norm auf \mathbb{R}^n .

a) $\|\tilde{x} - x\|$ heißt **absoluter Fehler**

b) $\frac{\|\tilde{x} - x\|}{\|x\|}$ heißt **relativer Fehler**

Da der relative Fehler skalierungsinvariant ist, d.h. unabhängig von der Wahl von x ist, ist dieser i.d.R. von größerem Interesse. Beide Fehler hängen von der Wahl der Norm ab! Häufig werden Fehler auch komponentenweise gemessen:

$$\begin{aligned} \text{Für } i = 1, \dots, n : \quad & |\tilde{x}_i - x_i| \leq \delta && (\text{absolut}) \\ & |\tilde{x}_i - x_i| \leq \delta |x_i| && (\text{relativ}) \end{aligned}$$

3.2.3 Wiederholung: Normen

$$\text{Euklidische Norm (} l_2\text{-Norm):} \quad \|x\|_2 := \sqrt{\sum_{i=1}^n |x_i|^2}$$

IMAGE MISSING

$$\text{Maximumsnorm (} l_\infty\text{-Norm):} \quad \|x\|_\infty := \max\{|x_i| : i = 1, \dots, n\}$$

IMAGE MISSING

$$\text{Summennorm (} l_1\text{-Norm):} \quad \|x\|_1 := \sum_{i=1}^n |x_i|$$

IMAGE MISSING

$$\text{Hölder-Norm (} l_p\text{-Norm):} \quad \|x\|_p := \left(\sum_{i=1}^n |x_i|^p \right)^{\frac{1}{p}}$$

3.2.4 Definition: Matrixnorm

Auf dem \mathbb{R}^n sei die Norm $\|\cdot\|_a$ und auf dem \mathbb{R}^m die Norm $\|\cdot\|_b$ gegeben. Dann ist die zugehörige **Matrixnorm** gegeben durch:

$$\begin{aligned} \|A\|_{a,b} &:= \sup_{x \neq 0} \frac{\|Ax\|_b}{\|x\|_a} \\ &= \sup_{\|x\|_a=1} \|Ax\|_b \end{aligned} \tag{3.2.1}$$

Also ist $\|A\|_{a,b}$ die kleinste Zahl $c > 0$ mit

$$\|Ax\|_b \leq c \|x\|_a \quad \forall x \in \mathbb{R}^n$$

3.2.5 Definition: Frobeniusnorm, p-Norm, Verträglichkeit

Sei $A \in \mathbb{R}^{m \times n}$.

a) **Frobeniusnorm** (Schurnorm): $\|A\|_F := \sqrt{\sum_{i=1}^m \sum_{j=1}^n |a_{ij}|^2}$

b) **p-Norm**: $\|A\|_p := \|A\|_{p,p}$

c) Eine Matrixnorm heißt **verträglich** mit den Vektornormen $\|\cdot\|_a, \|\cdot\|_b$, falls gilt ¹:

$$\|Ax\|_b \leq \|A\| \cdot \|x\|_a \quad \forall x \in \mathbb{R}^n$$

3.2.6 Bemerkungen

a) Die Normen $\|\cdot\|_F$ und $\|\cdot\|_p$ sind **submultiplikativ**, d.h.

$$\|A \cdot B\| \leq \|A\| \cdot \|B\|$$

b) Die Norm $\|\cdot\|_{1,1}$ wird auch **Spaltensummennorm** genannt:

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

Sie ist das Maximum der Spaltensummen².

c) Die Norm $\|\cdot\|_{\infty,\infty}$ wird auch **Zeilensummennorm** genannt³:

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

d) Die Frobeniusnorm $\|\cdot\|_F$ ist verträglich mit der euklidischen Norm $\|\cdot\|_2$

e) Die Wurzeln aus den Eigenwerten von $A^T A$ heißen **Singulärwerte** σ_i von A. Mit ihnen kann die $\|\cdot\|_{2,2}$ Norm dargestellt werden⁴:

$$\begin{aligned} \|A\|_2 &= \max\{\sqrt{\mu} : A^T A \cdot x = \mu x \text{ für ein } x \neq 0\} \\ &= \sigma_{\max} \end{aligned}$$

22.10.2014

¹ Beachte: $\|A\|_{a,b}$ ist die kleinste Norm im Gegensatz zu $\|A\|$, welche hier beliebig ist.

²Beweis: siehe Übungsblatt 3

³Beweis: siehe Übungsblatt 3

⁴Beweis: siehe Übungsblatt 3

3.2a) Normweise Konditionsanalyse

3.2.7 Definition: absolute und relative normweise Kondition

Sei (f, x) ein Problem mit $f : U \subset \mathbb{R}^n \rightarrow \mathbb{R}^m$ und $\|\cdot\|_a$ auf \mathbb{R}^n und $\|\cdot\|_b$ auf \mathbb{R}^m eine Norm.

- a) Die **absolute normweise Kondition** eines Problems (f, x) ist die kleinste Zahl $\kappa_{abs} > 0$ mit

$$\begin{aligned} \|f(\tilde{x}) - f(x)\|_b &\leq \kappa_{abs}(f, x) \|\tilde{x} - x\|_a + o(\|\tilde{x} - x\|_a) \\ \left(f(\tilde{x}) - f(x) = \underbrace{f'(x)(\tilde{x} - x)}_{\text{Taylorentwicklung}} \pm o(\|\tilde{x} - x\|) \right) &\quad \text{für } \tilde{x} \rightarrow x \end{aligned} \quad (3.2.2)$$

- b) Die **relative normweise Kondition** eines Problems (f, x) mit $x \neq 0, f(x) \neq 0$ ist die kleinste Zahl $\kappa_{rel} > 0$ mit

$$\frac{\|f(\tilde{x}) - f(x)\|_b}{\|f(x)\|_b} \leq \kappa_{rel}(f, x) \frac{\|\tilde{x} - x\|_a}{\|x\|_a} + o\left(\frac{\|\tilde{x} - x\|_a}{\|x\|_a}\right) \quad \text{für } \tilde{x} \rightarrow x \quad (3.2.3)$$

- c) Sprechweise:

- falls κ „klein“ ist, ist das Problem „gut konditioniert“
- falls κ „groß“ ist, ist das Problem „schlecht konditioniert“

3.2.8 Lemma

Falls f differenzierbar ist, gilt

$$\kappa_{abs}(f, x) = \|Df(x)\|_{a,b} \quad (3.2.4)$$

und für $f(x) \neq 0$

$$\kappa_{rel}(f, x) = \frac{\|x\|_a}{\|f(x)\|_b} \cdot \|Df(x)\|_{a,b} \quad (3.2.5)$$

wobei $Df(x)$ die Jakobi-Matrix bezeichnet.

3.2.9 Beispiel: Kondition der Addition

$f(x_1, x_2) := x_1 + x_2, f: \mathbb{R}^2 \rightarrow \mathbb{R}$.
Wähle l_1 -Norm auf \mathbb{R}^2 (und \mathbb{R})

$$\begin{aligned} Df(x_1, x_2) &= (\nabla f^T) = \left(\frac{\partial}{\partial x_1} f, \frac{\partial}{\partial x_2} f \right) \\ &= (1, 1) \end{aligned} \quad (\text{Matrix!})$$

damit

$$\begin{aligned} \kappa_{abs}(f, x) &= \|Df(x)\|_{1,1} && (\text{Matrix-Norm!!}) \\ &= \|Df(x)\|_1 \\ &= 1 \\ \kappa_{rel}(f, x) &= \frac{\|x\|_1}{\|f(x)\|_1} \cdot \|Df(x)\|_1 \\ &= \frac{|x_1| + |x_2|}{|x_1 + x_2|} \end{aligned}$$

Daraus folgt: Die Addition zweier Zahlen mit gleichem Vorzeichen ergibt

$$\kappa_{rel} = 1$$

Die Subtraktion zweier annähernd gleich großer Zahlen ergibt eine sehr schlechte relative Konditionierung:

$$\kappa_{rel} \gg 1$$

Zum Beispiel in 3.1.8: Es ist

$$\begin{aligned} x &= \begin{pmatrix} 9,995 \\ -9,984 \end{pmatrix} \cdot 10^{-1} \\ \tilde{x} = fl(x) &= \begin{pmatrix} 1 \\ -9,98 \cdot 10^{-1} \end{pmatrix} \end{aligned}$$

also

$$\begin{aligned} \frac{|f(\tilde{x}) - f(x)|}{|f(x)|} &= \frac{0,9}{1,1} = 0,8\overline{1} \\ &\leq \kappa_{rel}(f, x) \cdot \frac{\|\tilde{x} - x\|_1}{\|x\|_1} \\ &= \kappa_{rel}(f, x) \cdot 4,6 \cdot 10^{-4} \end{aligned}$$

3.2.10 Beispiel: Lösen eines Gleichungssystems

Sei $A \in \mathbb{R}^{n \times n}$ invertierbar und $b \in \mathbb{R}^n$. Es soll

$$Ax = b$$

gelöst werden. Die möglichen Lösungen in A und in b lassen sich folgendermaßen ermitteln:

- a) Betrachte die Störungen in b :
Sei hierzu

$$f : b \mapsto x = A^{-1}b$$

Berechne dann $\kappa(f, b)$ und löse

$$\begin{aligned} A(x + \Delta x) &= b + \Delta b \\ f(b + \Delta b) - f(b) &= \Delta x \\ &= A^{-1} \cdot \Delta b && \text{da } x = A^{-1}b \\ \Rightarrow \|\Delta x\|_b &= \|A^{-1} \Delta b\|_b \\ &\leq \|A^{-1}\|_{a,b} \cdot \|\Delta b\|_b && \forall b, \Delta b \end{aligned}$$

wobei $\|\cdot\|$ auf $\mathbb{R}^{n \times n}$ die dem \mathbb{R}^n zugeordnete Matrix-Norm sei.

Die Abschätzung ist **scharf**, d.h. es gibt ein $\Delta b \in \mathbb{R}^n$, so dass „=“ gilt, nach Definition 3.2.4.

Also gilt⁵:

$$\kappa_{abs}(f, b) = \|A^{-1}\|_{a,b} \quad (3.2.6)$$

unabhängig von b . ($x \mapsto Ax \quad \kappa_{abs}$)

Ebenso folgt die scharfe Abschätzung

$$\begin{aligned} \frac{\|f(b + \Delta b) - f(b)\|}{\|f(b)\|} &= \frac{\|\Delta x\|}{\|x\|} \\ &= \frac{\|A^{-1} \Delta b\|}{\|x\|} \\ &\leq \frac{\|A^{-1}\| \cdot \|b\|}{\|x\|} \cdot \frac{\|\Delta b\|}{\|b\|} \end{aligned}$$

Damit

$$\kappa_{rel}(f, b) = \|A^{-1}\| \cdot \frac{\|b\|}{\|A^{-1} \cdot b\|} \quad (3.2.7)$$

⁵vgl. auch Lemma 3.2.8: $\kappa_{abs}(f, b) = \|Df(b)\|_{a,b} = \|A^{-1}\|_{a,b}$

3 Fehleranalyse

Da $\|b\| \leq \|A\| \cdot \|x\| = \|A\| \cdot \|A^{-1}b\|$ folgt:

$$\kappa_{rel}(f, b) \leq \|A\| \cdot \|A^{-1}\| \quad (3.2.8)$$

für alle (möglichen rechten Seiten) b .

3.2.8 ist scharf in dem Sinne, dass es ein $\widehat{b} \in \mathbb{R}^n$ gibt mit

$$\|\widehat{b}\| = \|A\| \cdot \|\widehat{x}\|$$

und somit

$$\kappa_{rel}(f, \widehat{b}) = \|A\| \cdot \|A^{-1}\|$$

b) Betrachte die Störungen in A :

Löse also

$$(A + \Delta A)(x + \Delta x) = b$$

Sei hierzu

$$\begin{aligned} f : A &\mapsto x = A^{-1}b \\ \mathbb{R}^{n \times n} &\rightarrow \mathbb{R}^n \end{aligned}$$

und berechne $\kappa(f, A)$ mittels Ableitung $Df(A) : \mathbb{R}^{n \times n} \rightarrow \mathbb{R}^n$:

$$\begin{aligned} C \mapsto Df(A)C &= \frac{d}{dt} \left((A + tC)^{-1} \cdot b \right) \Big|_{t=0} \\ &= \frac{d}{dt} \left((A + tC)^{-1} \right) \Big|_{t=0} \cdot b \end{aligned}$$

Weiterhin gilt

$$\frac{d}{dt} \left((A + tC)^{-1} \right) \Big|_{t=0} = -A^{-1}CA^{-1}, \quad (3.2.9)$$

da

$$\begin{aligned} 0 &= \frac{d}{dt} I \\ &= \frac{d}{dt} \left((A + tC)(A + tC)^{-1} \right) \\ &= C(A + tC)^{-1} + (A + tC) \cdot \frac{d}{dt} (A + tC)^{-1} \\ &\Leftrightarrow \frac{d}{dt} (A + tC)^{-1} = -(A + tC)^{-1} \cdot C(A + tC)^{-1}, \end{aligned}$$

falls $(A + tC)$ invertierbar ist. Für ein genügend kleines t ist das gewährleistet, da A invertierbar ist (s. Lemma 3.2.12).

$$\Rightarrow Df(A)C = -A^{-1}CA^{-1}b$$

Somit folgt

$$\begin{aligned}
\kappa_{abs}(f, A) &= \|Df(A)\| \\
&= \sup_{\substack{C \neq 0 \\ C \in \mathbb{R}^{n \times n}}} \frac{\|A^{-1}CA^{-1}b\|}{\|C\|} \\
&\leq \sup_{\substack{C \neq 0 \\ C \in \mathbb{R}^{n \times n}}} \frac{\|A^{-1}\| \cdot \|C\| \cdot \|A^{-1}b\|}{\|C\|} \\
&= \|A^{-1}\| \cdot \|x\| \\
&\leq \|A^{-1}\|^2 \cdot \|b\| \\
\kappa_{rel}(f, A) &= \frac{\|A\|}{\|f(A)\|} \cdot \|Df(A)\| \\
&\leq \|A\| \cdot \|A^{-1}\|
\end{aligned} \tag{3.2.10}$$

c) betrachte Störungen in A und b :

$$(A + \Delta A)(x + \Delta x) = (b + \Delta b)$$

Für κ müsste $\|(A, b)\|$ festgelegt werden. Dies wird jedoch nicht betrachtet. Es gilt aber folgende Abschätzung für invertierbare Matrizen $A \in \mathbb{R}^{n \times n}$ und Störungen $\Delta A \in \mathbb{R}^{n \times n}$ mit $\|A^{-1}\| \cdot \|\Delta A\| < 1$:

$$\frac{\|\Delta x\|}{\|x\|} \leq \|A\| \cdot \|A^{-1}\| \cdot (1 - \|A^{-1}\| \cdot \|\Delta A\|) \cdot \underbrace{\left(\frac{\|\Delta b\|}{\|b\|} + \frac{\|\Delta A\|}{\|A\|} \right)}_{\neq \frac{\|(\Delta A, \Delta b)\|}{\|(A, b)\|}} \tag{3.2.11}$$

Beweis. s. Übungsblatt

□

3.2.11 Definition: Kondition einer Matrix

Sei $\|\cdot\|$ eine Norm auf $\mathbb{R}^{n \times n}$ und $A \in \mathbb{R}^{n \times n}$ eine reguläre Matrix. Die Größe

$$\kappa_{\|\cdot\|}(A) = \text{cond}_{\|\cdot\|}(A) := \|A\| \cdot \|A^{-1}\|$$

heißt **Kondition der Matrix** bzgl. der Norm $\|\cdot\|$.

Ist $\|\cdot\|$ von einer Vektor-Norm $\|\cdot\|_p$ induziert, bezeichnet $\text{cond}_p(A)$ die $\text{cond}_{\|\cdot\|_p}(A)$. Wir schreiben $\text{cond}(A)$ für $\text{cond}_2(A)$.

$\text{cond}_{\|\cdot\|}(A)$ schätzt die relative Kondition eines linearen GLS $Ax = b$ für alle möglichen Störungen in b oder in A ab und diese Abschätzung ist scharf.

3 Fehleranalyse

Es stellt sich nun die Frage:

Wann existiert die Inverse der gestörten invertierbaren Matrix A ?

Hierzu werden wir die Relationen benötigen:

$$A + \Delta A = A(I + A^{-1}\Delta A)$$

und mit $C \in \mathbb{R}^{n \times n}$, $\|C\| < 1$

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

27.10.2014

3.2.12 Lemma (Neumannsche Reihe)

Sei $C \in \mathbb{R}^{n \times n}$ mit $\|C\| < 1$ und mit einer submultiplikativen Norm $\|\cdot\|$, so ist $(I - C)$ invertierbar und es gilt:

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

Weiterhin gilt:

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

Beweis. Es gilt zu zeigen, dass $\sum_{k=0}^{\infty} C^k$ existiert:

Sei $q := \|C\| < 1$, dann gilt:

$$\begin{aligned} \left\| \sum_{k=0}^m C^k \right\| &\leq \sum_{k=0}^m \|C^k\| && \text{Dreiecksungleichung} \\ &\leq \sum_{k=0}^m \|C\|^k && \text{da } \|\cdot\| \text{ submultiplikativ} \\ &= \sum_{k=0}^m q^k \\ &= \frac{1 - q^{m+1}}{1 - q} \\ &\leq \frac{1}{1 - \|C\|} && \forall m \in \mathbb{N}, \text{ da } q < 1 \text{ (geometr. Reihe)} \end{aligned}$$

Daraus folgt bereits, dass $\sum_{k=1}^{\infty} C^k$ existiert (nach Majorantenkriterium).
Weiter gilt dann:

$$\begin{aligned}(I - C) \sum_{k=1}^{\infty} C^k &= \lim_{m \rightarrow \infty} (I - C) \sum_{k=1}^m C^k \\ &= \lim_{m \rightarrow \infty} (C^0 - C^{m+1}) \\ &= I\end{aligned}$$

□

3.2.13 Bemerkung

a) Für symmetrische, positiv definite Matrix $A \in \mathbb{R}^{n \times n}$ gilt⁶:

$$\kappa_2(A) = \frac{\lambda_{\max}}{\lambda_{\min}} \quad (3.2.13)$$

b) Eine andere Darstellung von $\kappa(A)$ ist

$$\kappa(A) := \frac{\max_{\|x\|=1} \|Ax\|}{\min_{\|x\|=1} \|Ax\|} \in [0, \infty] \quad (3.2.14)$$

Diese ist auch für nicht invertierbare und rechteckige Matrizen wohldefiniert.
Dann gilt offensichtlich:

c) $\kappa(A) \geq 1$

d) $\kappa(\alpha A) = \kappa(A)$ für $0 \neq \alpha \in \mathbb{R}$ (skalierungsinvariant)

e) $A \neq 0$ und $A \in \mathbb{R}^{n \times n}$ ist genau dann singulär, wenn $\kappa(A) = \infty$.

Wegen der Skalierungsinvarianz ist die Kondition zur Überprüfung der Regularität von A besser geeignet als die Determinante.

3.2.14 Beispiel: Kondition eines nichtlinearen Gleichungssystems

Sei $f: \mathbb{R}^n \rightarrow \mathbb{R}^n$ stetig differenzierbar und $y \in \mathbb{R}^n$ gegeben.

Löse

$$f(x) = y$$

Gesucht:

$$\kappa(f^{-1}, y)$$

⁶Beweis: siehe Übungsblatt 3



Abbildung 3.3: Gute und schlechte Kondition

mit f^{-1} Ausgabe und y Eingabe.

Sei $Df(x)$ invertierbar, dann existiert aufgrund des Satzes für implizite Funktionen die inverse Funktion f^{-1} lokal in einer Umgebung von y mit $f^{-1}(y) = x$, sowie

$$D(f^{-1})(y) = (Df(x))^{-1}$$

Hiermit folgt:

$$\begin{aligned}\kappa_{abs}(f^{-1}, y) &= \|(Df(x))^{-1}\| \\ \kappa_{rel}(f^{-1}, y) &= \frac{\|f(x)\|}{\|x\|} \cdot \|(Df(x))^{-1}\|\end{aligned}\tag{3.2.15}$$

Für skalare Funktionen $f: \mathbb{R} \rightarrow \mathbb{R}$ folgt somit:

$$\kappa_{rel}(f^{-1}, y) = \frac{|f(x)|}{|x|} \cdot \frac{1}{|f'(x)|}$$

Falls $|f'(x)| \rightarrow 0$ ist es eine schlechte absolute Kondition.

Für $|f'(x)| \gg 0$ ist es eine gute absolute Kondition.

Damit bedeutet eine kleine Störung in y eine große Störung in x .

3.2b) Komponentenweise Konditionsanalyse

3.2.15 Beispiel

Falls A Diagonalgestalt hat, sind die Gleichungen unabhängig voneinander (entkoppelt). Die erwartete relative Kondition wäre dann – wie bei skalaren Gleichungen – stets gleich 1. Ebenso sind Störungen nur in der Diagonale zu erwarten. Jedoch:

$$\begin{aligned}A &= \begin{pmatrix} 1 & 0 \\ 0 & \varepsilon \end{pmatrix} \\ \Rightarrow A^{-1} &= \begin{pmatrix} 1 & 0 \\ 0 & \varepsilon^{-1} \end{pmatrix} \\ \Rightarrow \kappa_{\infty} = \kappa_2 &= \frac{1}{\varepsilon} \quad \text{für } 0 < \varepsilon \leq 1\end{aligned}$$

3.2.16 Definition: Komponentenweise Kondition

Sei (f, x) ein Problem mit $f(x) \neq 0$ und $x = (x_i)_{i=1, \dots, n}$ mit $x_i \neq 0$ für alle $i = 1, \dots, n$. Die **komponentenweise Kondition** von (f, x) ist die kleinste Zahl $\kappa_{rel} \geq 0$, so dass:

$$\frac{\|f(\tilde{x}) - f(x)\|_\infty}{\|f(x)\|_\infty} \leq \kappa_{rel} \cdot \max_i \frac{|\tilde{x}_i - x_i|}{|x_i|} + o\left(\max_i \frac{|\tilde{x}_i - x_i|}{|x_i|}\right) \quad \text{für } \tilde{x} \rightarrow x$$

Vorsicht:

$$\frac{\|\tilde{x} - x\|_\infty}{\|x\|_\infty} \neq \max_i \frac{|\tilde{x}_i - x_i|}{|x_i|}$$

3.2.17 Lemma

Sei f differenzierbar und fasse $|\cdot|$ komponentenweise auf, d.h. $|x| = \begin{pmatrix} |x_1| \\ \vdots \\ |x_n| \end{pmatrix}$. Dann gilt:

$$\kappa_{rel} = \frac{\| |Df(x)| \cdot |x| \|_\infty}{\|f(x)\|_\infty} \quad (3.2.16)$$

Beweis. Vergleiche seien ebenfalls komponentenweise zu verstehen.

Nach dem Satz von Taylor gilt:

$$\begin{aligned} f_i(\tilde{x}) - f_i(x) &= \left(\frac{\partial f_i}{\partial x_1}(x), \dots, \frac{\partial f_i}{\partial x_n}(x) \right) \cdot \begin{pmatrix} \tilde{x}_1 - x_1 \\ \vdots \\ \tilde{x}_n - x_n \end{pmatrix} + o(\|\tilde{x} - x\|) \\ \Rightarrow |f_i(\tilde{x}) - f_i(x)| &\leq |Df(x)| \cdot \begin{pmatrix} |x_1| \cdot \frac{\tilde{x}_1 - x_1}{|x_1|} \\ \vdots \\ |x_n| \cdot \frac{\tilde{x}_n - x_n}{|x_n|} \end{pmatrix} + o\left(\max_i \frac{|\tilde{x}_i - x_i|}{|x_i|}\right) \quad \text{da } x_i \text{ fest und } \tilde{x}_i \rightarrow x_i \\ &\leq |Df(x)| \cdot |x| \cdot \max_i \frac{|\tilde{x}_i - x_i|}{|x_i|} + o\left(\max_i \frac{|\tilde{x}_i - x_i|}{|x_i|}\right) \\ \Rightarrow \frac{\|f(\tilde{x}) - f(x)\|_\infty}{\|f(x)\|_\infty} &\leq \frac{\| |Df(x)| \cdot |x| \|_\infty}{\|f(x)\|_\infty} \cdot \max_i \frac{|\tilde{x}_i - x_i|}{|x_i|} + o\left(\max_i \frac{|\tilde{x}_i - x_i|}{|x_i|}\right) \end{aligned}$$

Wähle $\tilde{x}_i = x_j + h \cdot \text{sign} \frac{\partial f_i}{\partial x_j}(x)$ mit $h > 0$, dann gilt:

$$|Df_i(x)(\tilde{x} - x)| = Df_i(x)(\tilde{x} - x)$$

und in obiger Rechnung gilt Gleichheit.

Also folgt, dass

$$\frac{\| |Df(x)| \cdot |x| \|_\infty}{\|f(x)\|_\infty} = \kappa_{rel}$$

□

3.2.18 Beispiel

a) Komponentenweise Kondition der Multiplikation

$$\begin{aligned}
f: \mathbb{R}^2 &\rightarrow \mathbb{R}, f(x, y) := x \cdot y \\
\Rightarrow Df(x, y) &= (y, x) \\
\Rightarrow \kappa_{rel}(x, y) &= \frac{\left\| (|y|, |x|) \cdot \begin{pmatrix} |x| \\ |y| \end{pmatrix} \right\|_{\infty}}{|x \cdot y|} \\
&= \frac{2 \cdot |x| \cdot |y|}{|x \cdot y|} \\
&= 2
\end{aligned}$$

b) Komponentenweise Kondition eines linearen Gleichungssystems:
Löse $Ax = b$ mit möglichen Störungen in b , also zu

$$\begin{aligned}
f: b &\mapsto A^{-1}b \\
\kappa_{rel} &= \frac{\| |A^{-1}| \cdot |b| \|_{\infty}}{\| A^{-1}b \|_{\infty}}
\end{aligned}$$

Falls A eine Diagonalmatrix ist, folgt:

$$\kappa_{rel} = 1$$

c) Komponentenweise Kondition des Skalarproduktes:

$$\begin{aligned}
\langle x, y \rangle &:= \sum_{i=1}^n x_i y_i = x^T y \\
f: \mathbb{R}^2 &\rightarrow \mathbb{R}, f(x, y) = \langle x, y \rangle \\
\Rightarrow Df(x, y) &= (y^T, x^T) \\
\kappa_{rel} &= \frac{\left\| |(y^T, x^T)| \cdot \left| \begin{pmatrix} x \\ y \end{pmatrix} \right| \right\|_{\infty}}{\|\langle x, y \rangle\|_{\infty}} \\
&= \frac{2 \cdot |y^T| \cdot |x|}{|\langle x, y \rangle|} \\
&= 2 \cdot \frac{\langle |x|, |y| \rangle}{|\langle x, y \rangle|} \\
&= 2 \cdot \frac{\cos(|x|, |y|)}{\cos(x, y)}
\end{aligned}$$

$$\text{da } \cos(x, y) = \frac{\langle y, x \rangle}{\|x\|_2 \cdot \|y\|_2}.$$



Abbildung 3.4: Schlechte Kondition des Skalarprodukts bei nahezu senkrechten Vektoren



Abbildung 3.5: Stabilität eines Algorithmus

Falls x und y nahezu senkrecht aufeinander stehen, kann das Skalarprodukt sehr schlecht konditioniert sein.

Zum Beispiel für $x = \tilde{x} = \begin{pmatrix} 1 \\ 1 \end{pmatrix}$ und $y = \begin{pmatrix} 1 + 10^{-10} \\ -1 \end{pmatrix}$, $\tilde{y} = \begin{pmatrix} 1 \\ -1 \end{pmatrix}$.

3.3 Stabilität von Algorithmen

Bislang: Kondition eines gegebenen Problems (f, x) .

Nun stellt sich die Frage: *Was passiert durch das Implementieren am Rechner?*

Ein „stabiler“ Algorithmus sollte ein gut konditioniertes Problem nicht „kaputt machen“.

3.3a) Vorwärtsanalyse

Die Fehlerfortpflanzung durch die einzelnen Rechenschritte, aus denen die Implementierung aufgebaut ist, wird abgeschätzt.

3.3.1 Bemerkung

Für die Rechenoperationen $+$, $-$, \cdot , $/$, kurz ∇ , gilt:

$$\begin{aligned} fl(a \nabla b) &= (a \nabla b) \cdot (1 + \varepsilon) \\ &= (a \nabla b) \cdot \frac{1}{1 + \mu} \end{aligned} \tag{3.3.1}$$

mit $|\varepsilon|, |\mu| \leq eps$.

29.10.2014

3.3.2 Beispiel

Sei $f(x_1, x_2, x_3) := \frac{x_1 x_2}{x_3}$ mit Maschinenzahlen x_i und $x_3 \neq 0$ und sei der Algorithmus durch

$$f(x_1, x_2, x_3) = (f^{(2)} \circ f^{(1)})(x_1, x_2, x_3)$$

gegeben mit

$$\begin{aligned} f^{(1)}(x_1, x_2, x_3) &= (x_1 \cdot x_2, x_3) & \text{und} \\ f^{(2)}(y, z) &= \frac{y}{z} \end{aligned}$$

Die Implementierung \tilde{f} von f beinhaltet Rundungsfehler.

Sei $x = (x_1, x_2, x_3)$. Daraus folgt:

$$\begin{aligned} \tilde{f}^{(1)}(x) &= (fl(x_1 \cdot x_2), x_3) \\ &= (x_1 x_2 (1 + \varepsilon_1), x_3) \end{aligned}$$

mit $|\varepsilon_1| \leq eps$:

$$\begin{aligned} \tilde{f}(x) &= \tilde{f}^{(2)}(\tilde{f}^{(1)}(x)) \\ &= fl(f^{(2)}(x_1 x_2 (1 + \varepsilon_1), x_3)) \\ &= \frac{x_1 x_2 (1 + \varepsilon_1)}{x_3} \cdot (1 + \varepsilon_2) \\ &= f(x) \cdot (1 + \varepsilon_1)(1 + \varepsilon_2) \end{aligned}$$

mit $|\varepsilon_2| \leq eps$:

$$\begin{aligned} \frac{|\tilde{f}(x) - f(x)|}{|f(x)|} &= |\varepsilon_1 + \varepsilon_2 + \varepsilon_1 \cdot \varepsilon_2| \\ &\leq 2eps + eps^2 \end{aligned}$$

Dies ist eine „worst case“ Analyse, da immer der maximale Fehler angenommen wird, und gibt i.d.R. eine starke Überschätzung des Fehlers an. Für qualitative Aussagen sind sie jedoch unnützlich.

In Computersystemen stehen mehr Operationen wie ∇ zur Verfügung, die mit einer relativen Genauigkeit eps realisiert werden können.

Daher:

3.3.3 Definition: Elementar ausführbar

Eine Abbildung $\phi : U \subseteq \mathbb{R}^n \rightarrow \mathbb{R}^m$ heißt **elementar ausführbar**, falls es eine elementare Operation $\tilde{\phi} : \mathbb{F}^n \rightarrow \mathbb{F}^m$ gibt, wobei \mathbb{F} die Menge der Maschinenzahlen bezeichne mit

$$|\tilde{\phi}_i(x) - \phi_i(x)| \leq \text{eps} \cdot |\phi_i(x)| \quad \forall x \in \mathbb{F}^n \text{ und } i = 1, \dots, m. \quad (3.3.2)$$

$\tilde{\phi}$ heißt dann **Realisierung** von ϕ .

Bemerkung: aus (3.3.2) folgt für $1 \leq p \leq \infty$:

$$\|\tilde{\phi}(x) - \phi(x)\|_p \leq \text{eps} \cdot \|\phi(x)\|_n \quad \forall x \in \mathbb{F}^n \quad (3.3.3)$$

3.3.4 Definition: Algorithmus, Implementation

Sei $f : E \subseteq \mathbb{R}^n \rightarrow \mathbb{R}^m$ gegeben.

Ein Tupel $(f^{(1)}, \dots, f^{(l)})$ mit $l \in \mathbb{N}$ von elementar ausführbaren Abbildungen

$$f^{(i)} : U_1 \subseteq \mathbb{R}^{k_i} \rightarrow U_{i+1} \subseteq \mathbb{R}^{k_{i+1}}$$

mit $k_1 = n$ und $k_{l+1} = m$ heißt **Algorithmus** von f , falls

$$f = f^{(l)} \circ \dots \circ f^{(1)}$$

Das Tupel $(\tilde{f}^{(1)}, \dots, \tilde{f}^{(l)})$ mit Abbildungen $\tilde{f}^{(i)}$, welche Realisierungen der $f^{(i)}$ sind, heißt **Implementation** von $(f^{(1)}, \dots, f^{(l)})$. Die Komposition

$$\tilde{f} = \tilde{f}^{(l)} \circ \dots \circ \tilde{f}^{(1)}$$

heißt Implementation von f .

Im Allgemeinen gibt es verschiedene Implementierungen einer Abbildung f .

3.3.5 Lemma (Fehlerfortpflanzung)

Sei $x \in \mathbb{R}^n$ und $\tilde{x} \in \mathbb{F}^n$ mit $|\tilde{x}_i - x_i| \leq \text{eps}|x_i|$ für alle $i = 1, \dots, n$. Sei $(f^{(1)}, \dots, f^{(l)})$ ein Algorithmus für f und $(\tilde{f}^{(1)}, \dots, \tilde{f}^{(l)})$ eine zugehörige Implementation.

Mit den Abkürzungen

$$\begin{aligned} x^{(j+1)} &:= f^{(j)} \circ \dots \circ f^{(1)}(x) \\ x^{(1)} &:= x \end{aligned}$$

3 Fehleranalyse

und entsprechend mit $\tilde{x}^{(j+1)}$ gilt, falls $x^{(j+1)} \neq 0$ für alle $j = 0, \dots, (l-1)$ und $\|\cdot\|$ eine beliebige p-Norm ist:

$$\frac{\|\tilde{x}^{(j+1)} - x^{(j+1)}\|}{\|x^{(j+1)}\|} \leq eps \cdot \mathcal{K} + o(eps) \quad (3.3.4)$$

$$\mathcal{K}^{(j)} = (1 + \kappa^{(j)} + \kappa^{(j)} \cdot \kappa^{(j-1)} + \dots + \kappa^{(j)} \cdot \dots \cdot \kappa^{(1)})$$

wobei $\kappa^{(j)} := \kappa_{rel}(f^{(j)}, x^{(j)})$ die Kondition der elementar ausführbaren Operationen $f^{(j)}$ ist.

Beweis.

$$\begin{aligned} \frac{\|\tilde{x}^{(j+1)} - x^{(j+1)}\|}{\|x^{(j+1)}\|} &= \frac{\|\tilde{f}^{(j)}(\tilde{x}^{(j)}) - f^{(j)}(x^{(j)})\|}{\|f^{(j)}(x^{(j)})\|} \\ &\leq \frac{\|\tilde{f}(\tilde{x}) - f(\tilde{x})\|}{\|f(\tilde{x})\|} \cdot \frac{\|f(\tilde{x})\|}{\|f(x)\|} + \frac{\|f(\tilde{x}) - f(x)\|}{\|f(x)\|} \quad (\text{Index } j \text{ vernachlässigt}) \\ &\leq eps \left(1 + \frac{\|f(\tilde{x}) - f(x)\|}{\|f(x)\|}\right) + \frac{\|f(\tilde{x}) - f(x)\|}{\|f(x)\|} \\ &\stackrel{\text{nach 3.3.3}}{=} eps + (eps + 1) \cdot \left(\kappa^{(j)} \cdot \frac{\|\tilde{x}^{(j)} - x^{(j)}\|}{\|x^{(j)}\|}\right) + o\left(\frac{\|\tilde{x}^{(j)} - x^{(j)}\|}{\|x^{(j)}\|}\right) \end{aligned}$$

Nach Voraussetzung gilt Gleichung (3.3.4) mit $\mathcal{K}^{(0)} = 1$ für $j = 0$.

Für $j = 1$ folgt nach Voraussetzung mit Gleichung (3.3.3)

$$\begin{aligned} \frac{\|\tilde{x}^{(2)} - x^{(2)}\|}{\|x^{(2)}\|} &\leq eps + (eps + 1) \cdot \left(\kappa^{(1)} eps + o(eps)\right) \\ &= eps(1 + \kappa^{(1)}) + o(eps) \\ &= eps\mathcal{K}^{(1)} + o(eps) \end{aligned}$$

Womit der Induktionsanfang gezeigt ist.

Für den Induktionsschritt von $j-1$ zu j :

$$\begin{aligned} \frac{\|\tilde{x}^{(j+1)} - x^{(j+1)}\|}{\|x^{(j+1)}\|} &\leq eps + (1 + eps)\kappa^{(j)} \left[eps\mathcal{K}^{(j-1)} + o(eps)\right] \\ &\quad + (1 + eps) \cdot o\left(eps \cdot \mathcal{K}^{(j-1)} + o(eps)\right) \\ &= eps \left(1 + \kappa^{(j)} \cdot \mathcal{K}^{(j-1)}\right) + o(eps) \end{aligned}$$

Mit $\mathcal{K}^{(j)} = 1 + \kappa^{(j)} \cdot \mathcal{K}^{(j-1)}$ folgt die Behauptung. □

Hiermit folgt:

3.3.6 Korollar

Unter der Voraussetzung von Lemma 3.3.5 gilt:

$$\frac{\|\tilde{f}(\tilde{x}) - f(x)\|}{\|f(x)\|} \leq eps \cdot \left(1 + \kappa^{(l)} + \kappa^{(l)} \cdot \kappa^{(l-1)} + \dots + \kappa^{(l)} \cdot \dots \cdot \kappa^{(1)}\right) + o(eps) \quad (3.3.5)$$

3.3.7 Bemerkung

Mit Korollar 3.3.6 ist offensichtlich, dass schlecht konditionierte Probleme zu elementar ausführbaren Abbildungen so früh wie möglich ausgeführt werden sollten.

Nach Beispiel 3.2.9 ist die Subtraktion zweier annähernd gleicher Zahlen schlecht konditioniert. Deshalb sollte man unvermeidbare Subtraktionen möglichst früh durchführen. Allerdings hängt $\kappa^{(j)}$ nicht nur von $f^{(j)}$, sondern auch vom Zwischenergebnis $x^{(j)}$ ab, welches a priori unbekannt ist.

3.3.8 Bemerkung zur Sprechweise

Der Quotient

$$\frac{\overbrace{\frac{\|\tilde{f}(\tilde{x}) - f(x)\|}{\|f(x)\|}}^{\text{Gesamtfehler}}}{\underbrace{\frac{\|\tilde{f}(\tilde{x})\|}{\|f(x)\|}}_{\text{Fehler durch Problem}} \cdot \underbrace{\frac{\|\tilde{x} - x\|}{\|x\|}}_{\text{Eingabefehler}}} \quad (3.3.6)$$

gibt die **Güte des Algorithmus** an. Als Stabilitätsindikator kann also

$$\sigma(f, \tilde{f}, x) := \frac{\mathcal{K}}{\kappa_{rel}(f, x)} \quad (3.3.7)$$

verwendet werden und es gilt

$$\frac{\|\tilde{f}(\tilde{x}) - f(x)\|}{\|f(x)\|} < \underbrace{\sigma(f, \tilde{f}, x)}_{\text{Beitrag des Algorithmus}} \cdot \underbrace{\kappa_{rel}(f, x)}_{\text{Beitrag des Problems}} \cdot \underbrace{eps}_{\text{Rundungsfehler}} + o(eps)$$

Falls $\sigma(f, \tilde{f}, x) < 1$, dämpft der Algorithmus die Fehlerfortpflanzung der Eingabe- und Rundungsfehler und heißt **stabil**.

Für $\sigma(f, \tilde{f}, x) \gg 1$ heißt der Algorithmus **instabil**.



Abbildung 3.6: Diagramm zur Rückwärtsanalyse

3.3.9 Beispiel

Nach Gleichung (3.3.3) gilt für die Elementaroperationen $\mathcal{K} \leq 1$. Da für die Subtraktion zweier annähernd gleich großer Zahlen $\kappa_{rel} \gg 1$ gilt, ist der Stabilitätsfaktor zweier annähernd gleich großer Zahlen sehr klein und der Algorithmus also stabil, Falls es sich jedoch bei einer zusammengesetzten Abbildung $f = h \circ g$ bei der zweiten Abbildung h um eine Subtraktion handelt, gilt

$$\mathcal{K} = (1 + \kappa(sub) + \kappa(sub) \cdot \kappa(g))$$

und die Stabilität ist gefährdet. Genauere Abschätzungen und damit genauere Indikatoren können durch komponentenweise Betrachtungen erhalten werden.

3.3b) Rückwärtsanalyse

Die Fragestellung ist nun:

Kann $\tilde{f}(\hat{x})$ als exaktes Ergebnis von einer gestörten Eingabe \hat{x} unter der exakten Abbildung f aufgefasst werden?

Das würde heißen

$$\exists \hat{x} \in \mathbb{R}^n : f(\hat{x}) = \tilde{f}(\tilde{x}).$$

Dann schätze den Fehler

$$\|\hat{x} - x\|$$

bzw. für nicht injektive f

$$\min_{\hat{x} \in \mathbb{R}^n} \{ \|\hat{x} - x\| \mid f(\hat{x}) = \tilde{f}(\tilde{x}) \}$$

ab.

Ein Anwendungsbeispiel:

Die Eingangsdaten seien Messdaten \tilde{x} mit 1 % relativer Genauigkeit. Liefert die Rückwärtsanalyse, dass $\tilde{f}(\tilde{x})$ als exaktes Ergebnis $f(\hat{x})$ mit Eingangsdaten \hat{x} , die höchstens um 0,5 % schwanken, aufgefasst werden kann, so ist das Verfahren „geeignet“.

Die Rückwärtsanalyse ist

- in der Regel leichter durchführbar als die Vorwärtsanalyse und
- ebenfalls nur eine qualitative Schätzung der Genauigkeit der numerisch berechneten Werte.

3.3.10 Bemerkung

(siehe auch Folien)

Vorwärtsfehler \leq Kondition des Problems \cdot Rückwärtsfehler.

$$\|\tilde{f}(\tilde{x}) - f(x)\| \leq \kappa(f, x) \|\hat{x} - x\|$$

Beispiel: Rückwärtsanalyse der Gauß-Elimination (geht auf Wilkinson zurück)

3.3.11 Satz

$A \in \mathbb{R}^{n \times n}$ besitze eine LR-Zerlegung. Dann berechnet die Gauß-Elimination Matrizen \hat{L} und \hat{R} , so dass

$$\hat{L} \hat{R} = \hat{A}$$

und

$$\begin{aligned} |\hat{A} - A| &\leq \frac{eps}{1 - n \, eps} \left(|\hat{L}| \begin{pmatrix} 1 & & & 0 \\ & 2 & & \\ & & \ddots & \\ & & & n \end{pmatrix} |\hat{R}| - |\hat{R}| \right) \\ &\leq \frac{n \, eps}{1 - n \, eps} |\hat{L}| |\hat{R}| = n \, eps |\hat{L}| |\hat{R}| + \mathcal{O}(n^2 eps^2) \end{aligned}$$

falls $n \, eps \leq 1/2$.

Beweis. siehe Stoer[6].

□

3.3.12 Satz (Sautter 1971)

$A \in \mathbb{R}^{n \times n}$ besitze eine LR-Zerlegung. Dann berechnet das Gaußsche Eliminationsverfahren für das Gleichungssystem $Ax = b$ eine Lösung \bar{x} mit

$$\bar{A} \bar{x} = b$$

mit

$$|\bar{A} - A| \leq 2 n \, eps |\hat{L}| |\hat{R}| + \mathcal{O}(n^2 eps^2).$$

Beweis. siehe Deuffhard/Hohmann [1]. □

Weitere Abschätzungen existieren für Gauß-Elimination mit Pivotisierung und für spezielle Klassen von Matrizen.

3.3.13 Allgemeine Faustregeln für die LR-Zerlegung

- Falls die Matrix $n|\hat{L}||\hat{R}|$ die selbe Größenordnung wie $|A|$ besitzt, ist der Algorithmus „gutartig“;
- Für tridiagonale Matrizen ist der Algorithmus mit Spaltenpivotisierung stabil.
- Falls A oder A^T **strikt diagonal dominant** ist, d.h.

$$|a_{ii}| > \sum_{j=1, j \neq i}^n |a_{ij}| \text{ für alle } i = 1, \dots, n,$$

ist Spaltenpivotisierung überflüssig. Der Algorithmus ist stabil.

- Für symmetrische, positiv definite Matrizen sollte keine Pivotisierung durchgeführt werden, um die Symmetrie zu erhalten. Der Algorithmus ist stabil.

Vorsicht

Selbst wenn die LR-Zerlegung stabil ist, in dem Sinne dass $|\bar{A} - A|$ klein ist für $\bar{A}\bar{x} = b$, kann die numerische Lösung \bar{x} sehr ungenau sein, da der Vorwärtsfehler $|\bar{x} - x|$ auch von der Kondition abhängt.

Ein Beispiel hierzu ist die Hilbertmatrix

$$H = \left(\frac{1}{i+j-1} \right)_{i,j=1,\dots,n},$$

für die $\text{cond}(H)$ exponentiell mit der Dimension n wächst.

3.4 Beurteilung von Näherungslösungen linearer GLS

Zu $Ax = b$ liege eine Näherungslösung \tilde{x} vor.

3.4a) Im Sinne der Vorwärtsanalyse

Im Sinne der Vorwärtsanalyse und der Fehlerentwicklung durch das Problem gilt:

$$\frac{\|\tilde{x} - x\|}{\|x\|} \leq \text{cond}(A) \cdot \frac{\|\Delta b\|}{\|b\|}$$

nach Beispiel 3.2.10, mit dem Residuum

$$\begin{aligned} r(\tilde{x}) &:= A\tilde{x} - b \\ &= \tilde{b} - b \\ &= \Delta b \end{aligned} \tag{3.4.1}$$

Wie der absolute Fehler ist das Residuum skalierungsabhängig. Daher ist $\|r(\tilde{x})\|$ „klein“ ungeeignet, um Genauigkeitsaussagen zu treffen.

Um den Fehler in x abzuschätzen, ist die Betrachtung von

$$\frac{\|r(\tilde{x})\|}{\|b\|} \tag{3.4.2}$$

geeigneter.

Für große $\text{cond}(A)$ ist dieser Quotient jedoch weiterhin ungeeignet.

3.4b) Im Sinne der Rückwärtsanalyse

3.4.1 Satz (Prager und Oettli, 1964)

Sei \tilde{x} eine Näherungslösung für $Ax = b$. Falls

$$|r(\tilde{x})| \leq \varepsilon(|A||\tilde{x}| + |b|). \tag{3.4.3}$$

dann existiert eine Matrix \tilde{A} und ein Vektor \tilde{b} , so dass

$$\tilde{A}\tilde{x} = \tilde{b}$$

und

$$|\tilde{A} - A| \leq \varepsilon|A| \quad \text{und} \quad |\tilde{b} - b| \leq \varepsilon|b|. \tag{3.4.4}$$

Aufgrund von (3.4.3) wird der komponentenweise relative Rückwärtsfehler durch

$$\max_i \frac{|A\tilde{x} - b|_i}{(|A||\tilde{x}| + |b|)_i}$$

abgeschätzt.

Für den normweisen relativen Rückwärtsfehler gilt entsprechend (Rigal und Gaches 1967)

$$\frac{\|A\tilde{x} - b\|}{\|A\|\|\tilde{x}\| + \|b\|}.$$

4 Lineare Gleichungssysteme: Direkte Methoden (Fortsetzung)

4.1 Gaußsches Eliminationsverfahren mit Äquilibration und Nachiteration

Mit Skalierung $D_z A$ (**Zeilenskalierung**) oder $D_s A$ (**Spaltenskalierung**) mittels Diagonalmatrizen D_z, D_s lässt sich eine Pivotstrategie beliebig abändern. Jetzt ist die Frage: *Was ist eine „gute“ Skalierung?*

Skalierung ändert die Länge der Basisvektoren des Bild- bzw. des Urbildvektorraumes. Durch Normierung der Länge auf 1 wird die Pivotstrategie unabhängig von der gewählten Einheit.

Sei $A \in \mathbb{R}^{n \times m}$ und $\|\cdot\|$ eine Vektornorm.

4.1.1 Äquilibration der Zeilen

Alle Zeilen von $D_z A$ haben die gleiche Norm, z.B. $\|\cdot\| = 1$, wofür

$$D_z = \begin{pmatrix} \sigma_1 & & 0 \\ & \ddots & \\ 0 & & \sigma_n \end{pmatrix} \quad \text{mit } \sigma_i := \frac{1}{\|(a_{i1}, \dots, a_{im})\|} \quad (4.1.1)$$

gesetzt wird.

4.1.2 Äquilibration der Spalten

Alle Spalten von AD_s haben die gleiche Norm, z.B. $\|\cdot\| = 1$, wofür

$$D_s = \begin{pmatrix} \tau_1 & & 0 \\ & \ddots & \\ 0 & & \tau_m \end{pmatrix} \quad \text{mit } \tau_j := \left\| \begin{pmatrix} a_{1j} \\ \vdots \\ a_{nj} \end{pmatrix} \right\|^{-1} \quad (4.1.2)$$

gesetzt wird.

Äquilibration von Zeilen **und** Spalten führt zu einem nichtlinearen Gleichungssystem und ist i.d.R. aufwendig.

4.1.3 Lemma

Sei A zeilenäquilibriert bzgl. der l_1 -Norm, dann gilt:

$$\text{cond}_\infty(A) \leq \text{cond}_\infty(DA) \quad (4.1.3)$$

für alle regulären Diagonalmatrizen D .

Beweis. siehe Übungsaufgabe □

Wie in Kapitel 3 gesehen, kann die Näherungslösung \tilde{x} trotz Pivotisierung und Äquilibration noch sehr ungenau sein.

4.1.4 Nachiteration

Die Näherung \tilde{x} kann durch Nachiteration verbessert werden.
Falls \tilde{x} exakt ist, gilt:

$$r(\tilde{x}) := b - A\tilde{x} = 0 \quad (4.1.4)$$

ansonsten ist $A(x - \tilde{x}) = r(\tilde{x})$. Also löse die Korrekturgleichung

$$A\Delta x = r(\tilde{x}) \quad (4.1.5)$$

und setze

$$x^{(1)} := \tilde{x} + \Delta x$$

Wiederhole dies sooft, bis $x^{(i)}$ „genau genug“ ist. Die Lösung \tilde{x} wird durch Nachiteration meist mit sehr gutem Erfolg verbessert [genaueres in 7]

(4.1.5) wird mit der bereits vorhandenen LR-Zerlegung nur mit der neuen rechten Seite $r(\tilde{x})$ gelöst, d.h. eine vorwärts und eine Rückwärtssubstitution mit $\mathcal{O}(n^2)$ flops.

4.1.5 Bemerkung (nach Skeel 1980)

Die Gauß-Elimination mit Spaltenpivotsuche und einer Nachiteration ist komponentenweise stabil.

4.2 Cholesky-Verfahren

Im Folgenden sei A eine symmetrische, positiv definite Matrix in $\mathbb{R}^{n \times n}$, d.h. $A = A^T$ und $\langle x, Ax \rangle = x^T Ax > 0$ für alle $x \neq 0$.

(kurs: **spd Matrix**)

4.2.1 Satz (Eigenschaften von symm., pos. def. Matrizen)

Für jede spd Matrix $A \in \mathbb{R}^{n \times n}$ gilt:

- i) A ist invertierbar
- ii) $a_{ii} > 0$ für $i = 1, \dots, n$
- iii) $\max_{ij} |a_{ij}| = \max_i a_{ii}$
- iv) Bei der Gauß-Elimination ohne Pivotsuche ist jede Restmatrix wieder eine spd Matrix.

Beweis.

- i) folgt aus (??)
- ii) Sei e_i der i -te Einheitsvektor, so folgt $a_{ii} = e_i^T A e_i > 0$.
- iii) siehe Übungsaufgabe
- iv) Es gilt:

$$\begin{aligned}
 A^{(1)} &:= A = \begin{pmatrix} a_{11} & z^T \\ z & B^{(1)} \end{pmatrix} \\
 A^{(2)} &:= L_1 A^{(1)} = \begin{pmatrix} 1 & 0 & \dots & 0 \\ -\frac{z}{a_{11}} & & I & \end{pmatrix} = \begin{pmatrix} a_{11} & z^T \\ 0 & B^{(2)} \\ \vdots & \\ 0 & \end{pmatrix} \\
 \Rightarrow L_1 A^{(1)} L_1^T &= \begin{pmatrix} a_{11} & z^T \\ 0 & B^{(2)} \\ \vdots & \\ 0 & \end{pmatrix} \cdot \begin{pmatrix} 1 & -\frac{z}{a_{11}} \\ 0 & I \\ \vdots & \\ 0 & \end{pmatrix} \\
 &= \begin{pmatrix} a_{11} & 0 & \dots & 0 \\ 0 & & B^{(2)} & \\ \vdots & & & \\ 0 & & & \end{pmatrix}
 \end{aligned}$$

Weiterhin gilt:

$$x \neq 0 \Leftrightarrow L_1 x \neq 0$$

da L_1 invertierbar. Also gilt insgesamt:

$$\begin{aligned}
 \tilde{x}^T B^{(2)} \tilde{x} &= x^T L_1 A^{(1)} L_1^T x && \text{für } x := \begin{pmatrix} 0 \\ \tilde{x} \end{pmatrix} \\
 &= (L_1^T x)^T A (L_1^T x) > 0 && \forall \tilde{x} \neq 0
 \end{aligned}$$

4 Lineare Gleichungssysteme: Direkte Methoden (Fortsetzung)

und damit ist auch $B^{(2)}$ spd.

Induktiv folgt hiermit iv).

□ Insbesondere ergibt sich:

$$(L_{n-1} \cdots L_1) A^{(1)} (L_1^T \cdots L_{n-1}^T) = \begin{pmatrix} d_1 & & 0 \\ & \ddots & \\ 0 & & d_n \end{pmatrix},$$

wobei d_i das i -te Diagonalelement von $A^{(i)}$ ist und somit $d_i > 0$ für $i = 1, \dots, n$ gilt.

Sei $L := (L_1^{-1} \cdots L_{n-1}^{-1})$ wie in (2.1.8), so ergibt sich:

□

4.2.2 Folgerung

Für jede spd Matrix A existiert eine eindeutige Zerlegung der Form

$$A = LDL^T$$

wobei L eine reelle unipotente (d.h. $l_{ii} = 1$) (, normierte) untere Dreiecksmatrix und D eine positive Diagonalmatrix ist. Diese Zerlegung heißt **rationale Cholesky-Zerlegung**. Die Zerlegung

$$A = \bar{L} \bar{L}^T \tag{4.2.1}$$

mit der reellen unteren Dreiecksmatrix

$$\bar{L} = L \begin{pmatrix} \sqrt{d_1} & & 0 \\ & \ddots & \\ 0 & & \sqrt{d_n} \end{pmatrix} = LD^{\frac{1}{2}}$$

heißt **Cholesky-Zerlegung**. .

Wegen (4.2.1) gilt:

$$a_{kk} = \bar{l}_{k1}^2 + \cdots + \bar{l}_{kk}^2 \tag{4.2.2}$$

$$a_{ik} = \bar{l}_{i1} \bar{l}_{k1} + \cdots + \bar{l}_{ik} \bar{l}_{kk} \tag{4.2.3}$$

$$IMAGE MISSING \tag{4.2.4}$$

Demnach funktioniert spaltenweises und zeilenweises Berechnen.

Es ergibt sich folgender Algorithmus:

4.2.3 Cholesky-Zerlegung

Der Algorithmus der Cholesky-Zerlegung ist wie folgt:

```

for  $k = 1, \dots, n$ 
|    $l_{kk} = (a_{kk} - \sum_{j=1}^{k-1} l_{kj}^2)^{\frac{1}{2}}$ 
|   for  $i = k + 1, \dots, n$ 
|   |    $l_{ik} = (a_{ik} - \sum_{j=1}^{k-1} l_{ij} l_{kj}) / l_{kk}$ 
|   end
end

```

4.2.4 Rechenaufwand in flops

Es sind je

$\frac{1}{6}(n^2 - n)$ Additionen sowie Multiplikationen und

$\frac{1}{6}(3n^2 - 3n)$ Divisionen

also ca. $\frac{2}{3}n^2$ flops für große n notwendig.

Im Vergleich zur LR-Zerlegung halbiert sich in etwa der Aufwand.

4.2.5 Bemerkung

- a) Wegen (4.2.2) gilt $|\bar{l}_{kj}| \leq \sqrt{a_{kk}}$, d.h. die Matrizeneinträge können nicht zu groß werden.
- b) Für spd Matrizen ist der Cholesky-Algorithmus stabil nach (??)
- c) Da A symmetrisch ist, muss nur die untere Dreiecksmatrix gespeichert werden. In Algorithmen kann \bar{L} in eine Kopie dieser Dreiecksmatrix geschrieben werden.
- d) Fast singuläre Matrizen können durch die Diagonale erkannt werden.

10.11.2014

4.3 Lineare Ausgleichsprobleme

4.3.1 Beispiel

(s. Einführung)

Seien m Messungen (I_i, U_i) für die Stromstärke I und die Spannung U gegeben.

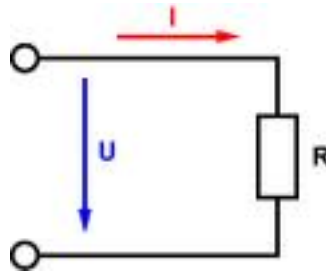


Abbildung 4.1: Schaltplan einer einfachen U-I-Messung

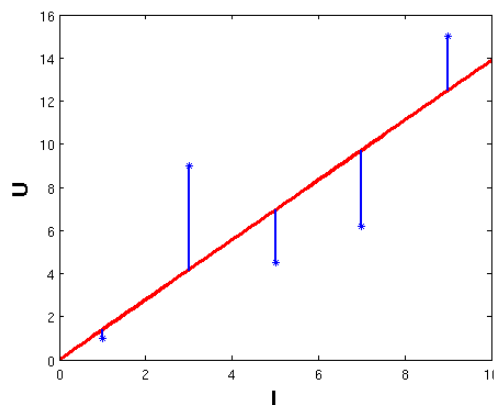


Abbildung 4.2: Linearausgleich einer U-I-Messung mit Ursprungsgerade als Modellfunktion

Das Ohmsche Gesetz liefert hierfür:

$$U = R \cdot I$$

Gesucht ist der zugehörige Widerstand R .

Wird jetzt davon ausgegangen dass die I_i exakt sind, wird das R gesucht, für das RI_i im Mittel den minimalen Abstand zu U_i hat. Genauer gesagt berechne

$$\min_{r \in \mathbb{R}} \sum_{i=1}^m (U_i - rI_i)^2$$

Vorsicht: Es wird **nicht** die Gerade (bzw. der lineare Untervektorraum) mit minimalem euklidischem Abstand zu (I_i, U_i) gesucht!

Dieses Problem ist nichtlinear und aufwendig zu lösen.

4.3.2 Lineares Ausgleichsproblem

Gegeben seien Messdaten (t_i, b_i) mit $t_i, b_i \in \mathbb{R}$ für $i = 1, \dots, m$ und die Abhängigkeit $b(t)$ werde beschrieben durch eine Modellfunktion, welche linear von den unbekannten Parametern x_1, \dots, x_n des Modells abhängt, d.h.

$$b(t) = a_1(t)x_1 + \dots + a_n(t)x_n$$

Für exakte Messdaten b_i würde

$$b(t_i) = b_i \quad \forall i \in \{1, \dots, m\}$$

gelten.

Im Allgemeinen werden jedoch $m \geq n$ Messwerte b_i bestimmt, und hiermit die n Parameter x_i so gewählt, dass die kleinsten **Fehlerquadrate auftreten**:

$$\min_{x_1, \dots, x_n} \sum_{i=1}^m (b_i - b(t_i))^2 \quad (4.3.1)$$

(Nach Gauß kann (4.3.1) auch aus der Maximum-Likelihood-Methode für einen stochastischen Ansatz hergeleitet werden.)

Definiere:

$$\begin{aligned} b &= (b_i)_{i=1, \dots, m} \in \mathbb{R}^m \\ x &= (x_j)_{j=1, \dots, n} \in \mathbb{R}^n \\ A &= (a_j(t_i))_{\substack{i=1, \dots, m \\ j=1, \dots, n}} \in \mathbb{R}^{m \times n} \end{aligned}$$

Damit ist (4.3.1) äquivalent zum **linearen Ausgleichsproblem**:

Zu gegebenem $b \in \mathbb{R}^m$ und $A \in \mathbb{R}^{m \times n}$ mit $m \geq n$ ist das $\bar{x} \in \mathbb{R}^n$ gesucht mit

$$\|b - A\bar{x}\|_2 = \min_{x \in \mathbb{R}^n} \|b - Ax\|_2 \quad (4.3.2)$$

Das entspricht der „Lösung“ eines überbestimmten, i.A. nicht erfüllbaren GLS $Ax = b$. Aufgrund der l_2 -Norm ist \bar{x} gegeben durch die orthogonale Projektion von b auf den Bildraum $R(A)$, wie gleich gezeigt wird.

IMAGE MISSING

4.3.3 Projektionssatz

Sei V ein reeller Vektorraum mit einem Skalarprodukt $\langle \cdot, \cdot \rangle$ und der induzierten Norm $\|v\| := \sqrt{\langle v, v \rangle}$. Sei $U \subset V$ ein endlich dimensionaler Untervektorraum und sei

$$U^\perp := \{v \in V \mid \langle v, u \rangle = 0 \quad \forall u \in U\}$$

Dann gilt:

- 1) Zu jedem $v \in V$ existiert genau ein $\bar{u} \in U$, so dass $v - \bar{u} \in U^\perp$, d.h.

$$\langle v - \bar{u}, u \rangle = 0 \quad \forall u \in U$$

Dies definiert die **orthogonale Projektion**

$$P : V \rightarrow U, \quad v \mapsto \bar{u} = Pv$$

- 2) Zu jedem $v \in V$ bestimmt $P \cdot v$ die eindeutige Lösung

$$\|v - Pv\| = \min_{u \in U} \|v - u\|$$

Also gilt mit einem eindeutigen $\bar{u} = Pv$, dass

$$\|v - \bar{u}\| = \min_{u \in U} \|v - u\| \iff \langle v - \bar{u}, u \rangle = 0 \quad \forall u \in U \quad (4.3.3)$$

Beweis. 1) Sei $\{u_1, \dots, u_n\}$ eine Orthonormalbasis von U und $\bar{u} \in U$.
Daraus folgt:

$$\exists! (\alpha_i)_{i=1, \dots, n} \in \mathbb{R} : \bar{u} = \sum_{i=1}^n \alpha_i u_i$$

Damit gilt:

$$\begin{aligned} 0 &= \langle v - \bar{u}, u \rangle && \forall u \in U \\ \iff 0 &= \langle v - \sum_{i=1}^n \alpha_i u_i, u_j \rangle && \forall j = 1, \dots, n \\ \iff \langle v, u_j \rangle &= \sum_{i=1}^n \alpha_i \langle u_i, u_j \rangle = \alpha_j \end{aligned}$$

Setze also

$$\begin{aligned} P \cdot v &= \bar{u} \\ &= \sum_{i=1}^n \langle v, u_i \rangle u_i \in U \end{aligned} \quad (4.3.4)$$

dann ist \bar{u} die eindeutig bestimmte Lösung für $v - \bar{u} \in U^\perp$

IMAGE MISSING

Sei $u \in U$. Dann gilt:

$$\begin{aligned}
 \|v - u\|^2 &= \|v - \bar{u} + \bar{u} - u\|^2 \\
 &= \|v - \bar{u}\|^2 + \underbrace{\langle v - \bar{u}, \overbrace{\bar{u} - u}^{\in U} \rangle}_{=0} + \|u - \bar{u}\|^2 \\
 &= \|v - \bar{u}\|^2 + \|u - \bar{u}\|^2
 \end{aligned} \tag{4.3.5}$$

(Dies ist anschaulich der Satz des Pythagoras.)

□

4.3.4 Satz

Der Vektor $\bar{x} \in \mathbb{R}^n$ ist genau dann Lösung des linearen Ausgleichsproblems

$$\min_{x \in \mathbb{R}^n} \|b - Ax\|_2,$$

falls er die Normalengleichung

$$A^T A \bar{x} = A^T b \tag{4.3.6}$$

erfüllt.

Insbesondere ist \bar{x} eindeutig, falls $A \in \mathbb{R}^{m \times n}$ maximalen Rang $n \leq m$ hat.

Beweis. Bezeichne $V = \mathbb{R}^m$, $U = R(A) = \{Ax \mid x \in \mathbb{R}^n\}$, $b \in \mathbb{R}^m$.

Nach (4.3.3) gilt:

$$\begin{aligned}
 \|b - A\bar{x}\|_2 &= \min_{x \in \mathbb{R}^n} \|b - Ax\|_2 \\
 \Leftrightarrow \quad \langle b - A\bar{x}, Ax \rangle &= 0 \quad \forall x \in \mathbb{R}^n \\
 \Leftrightarrow \quad \langle A^T(b - A\bar{x}), x \rangle &= 0 \quad \forall x \in \mathbb{R}^n \\
 \Leftrightarrow \quad A^T(b - A\bar{x}) &= 0 \\
 \Leftrightarrow \quad A^T A \bar{x} &= A^T b
 \end{aligned}$$

Nach dem Projektionssatz 4.3.3 existiert mindestens ein eindeutiges $\bar{y} = Pb$. Für dieses \bar{y} ist $\bar{x} \in \mathbb{R}^n$ mit $\bar{y} = A\bar{x}$ eindeutig bestimmt, falls A injektiv ist, d.h. falls $\text{rang}(A) = n$. □

Ähnlich zum Skalarprodukt ist die relative Kondition von (P, b) schlecht, falls b fast senkrecht zu U steht. Die relative Kondition des linearen Ausgleichsproblems hängt zusätzlich von $\text{cond}(A)$ ab.

4.3.5 Lösung der Normalgleichung

Falls $\text{rang}(A) = n$, ist $A^T A$ spd und das Cholesky-Verfahren ist anwendbar. Dafür ist

1. $A^T A$ zu berechnen:

Aufwand ca. $\frac{1}{2}n^2m$ Multiplikationen

Kondition häufig schlecht, da $\frac{1}{2}n^2$ Skalarprodukte berechnet werden

2. die Cholesky-Zerlegung von $A^T A$ durchzuführen:

Aufwand ca. $\frac{1}{6}n^3$ Multiplikationen

Kondition Für $A \in \mathbb{R}^{m \times n}$ mit $m \geq n$ und $\text{rang}(A) = n$ gilt:

$$\text{cond}_2(A^T A) = \text{cond}_2(A)^2 \quad (4.3.7)$$

(siehe Übungsaufgabe 19)

Also überwiegt für $m \gg n$ der Aufwand $A^T A$ zu berechnen. Die auftretenden Konditionen entsprechen i.d.R. nicht dem des Ausgangsproblems.

Damit ist die **Cholesky-Zerlegung für Normalgleichungen ungeeignet**.

4.3.6 Satz

Sei $A \in \mathbb{R}^{m \times n}$ mit $m \geq n$ und $\text{rang}(A) = n$, sei $b \in \mathbb{R}^m$ und besitze A eine Zerlegung

$$A = Q \begin{pmatrix} R \\ 0 \end{pmatrix}$$

mit einer orthogonalen Matrix $Q \in \mathbb{R}^{m \times m}$ und einer oberen Dreiecksmatrix $R \in \mathbb{R}^{n \times n}$.

Dann ist R invertierbar.

Bezeichne

$$\begin{pmatrix} \bar{b}_1 \\ \bar{b}_2 \end{pmatrix} := Q^T \cdot b \quad (4.3.8)$$

dann ist

$$\bar{x} = R^{-1} \bar{b}_1 \quad (4.3.9)$$

die Lösung des linearen Ausgleichsproblems und

$$\begin{aligned} \|\bar{b}_2\| &= \|b - A\bar{x}\| \\ &= \min_{x \in \mathbb{R}^n} \|b - Ax\| \end{aligned}$$

Zur Erinnerung:

$$\begin{aligned} Q \text{ orthogonal} &: \Leftrightarrow QQ^T = I \\ &\Leftrightarrow Q^{-1} = Q^T \end{aligned}$$

Weiterhin ist Q längenerhaltend, d.h. $\|Qv\|_2 = \|v\|_2$ und somit folgt

$$\begin{aligned} \|Q\|_2 &= \|Q^{-1}\|_2 = 1 \quad \text{und} \\ \text{cond}_2(Q) &= 1 \end{aligned} \tag{4.3.10}$$

Beweis. R ist invertierbar, da

$$\begin{aligned} \text{rang}(R) &= \text{rang}(Q^{-1} \cdot A) \\ &= \text{rang}(A) \\ &= n \end{aligned}$$

Außerdem gilt:

$$\begin{aligned} \|b - Ax\|_2^2 &= \left\| Q \left(Q^T b - \begin{pmatrix} R \\ 0 \end{pmatrix} x \right) \right\|_2^2 \\ &= \left\| Q^T b - \begin{pmatrix} Rx \\ 0 \end{pmatrix} \right\|_2^2 \\ &\stackrel{Q \text{ längenerhaltend}}{=} \left\| \bar{b}_1 - Rx \right\|_2^2 + \left\| \bar{b}_2 \right\|_2^2 \end{aligned}$$

wird minimal für $R\bar{x} = \bar{b}_1$

□

Da Q längenerhaltend ist, folgt mit 3.2.13 b) ($\text{cond}_A := \frac{\max\|Ax\|}{\min\|Ax\|}$) sofort:

$$\text{cond}_2(A) = \text{cond}_2(R)$$

Die auftretende Kondition entspricht also der des Ausgleichsproblems.

4.3.7 Bemerkung

Sei $A \in \mathbb{R}^{n \times n}$ invertierbar und habe eine QR-Zerlegung, d.h. es existiert eine orthogonale Matrix Q und eine obere Dreiecksmatrix R , so dass:

$$A = Q \cdot R$$

Dann kann das Gleichungssystem $Ax = b$ wie folgt gelöst werden:

1. Setze $z = Q^T b$, was Kondition 1 hat.
2. Löse durch Rückwärtssubstitution $Rx = z$.

12.11.2014

4.4 Orthogonalisierungsverfahren

Konstruiere eine QR-Zerlegung

$$A = Q \cdot \begin{pmatrix} R \\ 0 \end{pmatrix} \quad (4.4.1)$$

durch einen Eliminationsprozess:

$$A \rightarrow Q^{(1)} A \rightarrow Q^{(2)} Q^{(1)} A \rightarrow Q^{(p)} \dots Q^{(1)} A = \begin{pmatrix} R \\ 0 \end{pmatrix}. \quad (4.4.2)$$

mit orthogonalen Matrizen $Q^{(i)}$. Dann gilt

$$Q = Q^{(1)T} \dots Q^{(p)T} \quad (4.4.3)$$

Dies ist im Gegensatz zur LR -Zerlegung aufgrund von $\text{cond}(Q^{(i)}) = 1$ immer stabil.

Für $Q \in \mathbb{R}^{2 \times 2}$ gibt es zwei mögliche Anschauungen, nämlich:

- a) Drehung IMAGE MISSING
- b) Spiegelung IMAGE MISSING

4.4a) Givens-Rotation

Es wird eine Drehung auf den 1. Einheitsvektor durchgeführt:

IMAGE MISSING

$$a = \begin{pmatrix} a_1 \\ a_2 \end{pmatrix} \rightarrow \begin{pmatrix} \alpha \\ 0 \end{pmatrix} = \alpha e_1$$

d.h. Elimination von a_2 mit

$$\|\alpha e_1\|_2 = \|a\|_2$$

Also gilt für α

$$\alpha = \pm \|a\|_2$$

Drehungen werden beschrieben durch

$$Q = \begin{pmatrix} \cos(\theta) & \sin(\theta) \\ -\sin(\theta) & \cos(\theta) \end{pmatrix} =: \begin{pmatrix} c & s \\ -s & c \end{pmatrix} \quad \theta \in [0, 2\pi)$$

und es muss gelten

$$Qa = \begin{pmatrix} \alpha \\ 0 \end{pmatrix}$$

Hiermit folgt für $\|a\| = 0$

$$\begin{aligned} c &= 1, s = 0 \\ c &= \frac{a_1}{\alpha}, s = \frac{a_2}{\alpha} \quad \text{mit} \\ \alpha &= \pm \sqrt{a_1^2 + a_2^2} \end{aligned} \quad (4.4.4)$$

für $\|a\| \neq 0$.

Im Folgenden wird dies kurz mit

$$[c, s] = \text{givens}(a_1, a_2)$$

bezeichnet.

Als **Givens-Rotation** wird eine Matrix der Form

$$\Omega_{k,l} = \begin{pmatrix} 1 & & & & & & \\ & \ddots & & & & & \\ & & 1 & & & & \\ & & & \mathbf{c} & & \mathbf{s} & \\ & & & & 1 & & \\ & & & & & \ddots & \\ & & & & & & 1 \\ & & & -\mathbf{s} & & \mathbf{c} & \\ & & & & & & \ddots \\ & & & & & & & 1 \end{pmatrix} \begin{array}{l} \leftarrow k\text{-te Zeile} \\ \\ \\ \leftarrow l\text{-te Zeile} \end{array} \quad (4.4.5)$$

mit $c^2 + s^2 = 1$ und $k < l$ bezeichnet.

Es folgt:

$$\begin{aligned} \Omega_{kl}A &= \tilde{A} \quad \text{mit} \\ \tilde{a}_{ij} &= a_{ij} \quad \text{für } i \neq k, l \\ \tilde{a}_{kj} &= ca_{kj} + sa_{lj} \\ \tilde{a}_{lj} &= -sa_{kj} + ca_{lj} \end{aligned}$$

4 Lineare Gleichungssysteme: Direkte Methoden (Fortsetzung)

Demnach werden nur die k -te und l -te Zeile werden verändert.
Falls nun $[c, s] = \text{givens}(x_k, x_l)$ gilt

$$\Omega_{k,l} \cdot x = \begin{pmatrix} x_1 \\ \vdots \\ x_{k-1} \\ \alpha \\ x_{k+1} \\ \vdots \\ x_{l-1} \\ 0 \\ x_{l+1} \\ \vdots \\ x_n \end{pmatrix} \quad \text{mit } \alpha = \pm \left\| \begin{pmatrix} x_k \\ x_l \end{pmatrix} \right\|_2$$

d.h. eine Givens-Rotation erzeugt eine Null.
Da nun

$$\mathbb{R}^{m \times n} \ni \begin{pmatrix} * & * & * & \dots \\ 0 & * & & \\ 0 & 0 & \ddots & \\ \vdots & & & \\ 0 & 0 & 0 & \dots & * \end{pmatrix} = \begin{pmatrix} R \\ 0 \end{pmatrix}$$

gilt, sind

$$p = \sum_{j=1}^n (m - j)$$

Givens-Rotationen nötig, um eine QR-Zerlegung nach (4.4.1) zu erzeugen. Und eine Rotation, welche a_{ij} auf 0 setzt, ist durch zugehörige (c_{ij}, s_{ij}) gegeben.

Für eine 3x4-Matrix sieht das Verfahren folgendermaßen aus:

$$\begin{aligned} A = \begin{pmatrix} * & & & \\ * & & * & \\ * \swarrow & & & \\ * \swarrow & & & \end{pmatrix} &\xrightarrow{\Omega_{3,4}} \begin{pmatrix} * & & & \\ * \swarrow & & & \\ * \swarrow & & & \\ 0 & & & \end{pmatrix} \xrightarrow{\Omega_{1,2}} \begin{pmatrix} * \swarrow & & & \\ * \swarrow & & & \\ 0 & & & \\ 0 & & & \end{pmatrix} \\ \xrightarrow{\Omega_{2,3}} \begin{pmatrix} * & & & \\ 0 & & * & \\ 0 & & & \\ 0 & & & \end{pmatrix} &\xrightarrow{\Omega_{3,4}} \begin{pmatrix} * & * & & \\ 0 & * \swarrow & & * \\ 0 & * \swarrow & & \\ 0 & 0 & & \end{pmatrix} \xrightarrow{\Omega_{2,3}} \begin{pmatrix} * & * & * & \\ 0 & * & * & \\ 0 & 0 & * \swarrow & \\ 0 & 0 & * \swarrow & \end{pmatrix} \\ \xrightarrow{\Omega_{3,4}} \begin{pmatrix} * & * & * & \\ 0 & * & * & \\ 0 & 0 & * & \\ 0 & 0 & 0 & \end{pmatrix} &= \begin{pmatrix} R \\ 0 \end{pmatrix} \end{aligned}$$

Es ergibt sich:

4.4.1 Givens-QR-Algorithmus

```

for  $j = 1, \dots, n$ 
|   for  $i = m, m-1, \dots, j+1$ 
|   |   % setze  $a_{ij}$  auf 0
|   |    $[c, s] = \text{givens}(a_{i-1,j}, \dots, a_{ij})$ 
|   |   speichere  $c$  und  $s$  für  $a_{ij}$ 
|   |    $A(i-1:j, j:h) = \begin{pmatrix} c & s \\ -s & c \end{pmatrix} * A(i-1:j, j:n)$ 
|   end
end

```

4.4.2 Bemerkungen

- a) $A(i-1:i, 1:j-1) = 0$ und ist daher nicht zu berechnen oder zu speichern. Der Speicherplatz kann für die Speicherung der Givensrotationen benutzt werden.
- b) R steht anschließend in A .
- c) Die Bestimmung der Länge $|\alpha|$ wird so ausgeführt, dass overflow oder underflow vermieden wird. Weiterhin wird das Vorzeichen von c oder s festgelegt, so dass aufgrund von $c^2 + s^2 = 1$ nur ein Wert ρ gespeichert werden muss. Hiermit wird auch das Vorzeichen von α festgelegt:

$a_2 = 0$: Setze $c = 1, s = 0, \alpha = a_1$, merke $\rho = 1$.

$|a_2| > |a_1|$: Setze $\tau = \frac{a_1}{a_2}, s = \frac{1}{\sqrt{1+\tau^2}}, c = s \cdot \tau, \alpha = a_2 \sqrt{1+\tau^2}$, merke $\rho = \frac{c}{2}$.

$|a_1| \geq |a_2|$: Setze $\tau = \frac{a_2}{a_1}, c = \frac{1}{\sqrt{1+\tau^2}}, s = c \cdot \tau, \alpha = a_1 \sqrt{1+\tau^2}$, merke $\rho = \frac{2}{s}$.

- d) Aufgrund von c) muss nur ρ gespeichert werden:

$\rho = 1$: Setze $c = 1, s = 0$.

$|\rho| < 1$: Setze $c = 2\rho, s = \sqrt{1-c^2}$.

$|\rho| > 1$: Setze $s = \frac{2}{\rho}, c = \sqrt{1-s^2}$.

Hiermit können alle notwendigen Givens-Rotationen als untere Dreiecksmatrix zusammen mit R in A gespeichert werden.

4.4.3 Aufwand des Givens-QR-Algorithmus

a) $m \approx n$

→ ca. $\frac{4}{3}n^3$ Multiplikationen und $\frac{1}{2}n^2$ Quadratwurzeln nötig

Die Givens-QR-Zerlegung ist somit ungefähr viermal so aufwändig wie die Gauß-Elimination, dafür jedoch stabil.

b) $m \gg n$

→ ca. $2n^2m$ Multiplikationen und mn Quadratwurzeln nötig

Das Verfahren ist daher zwei- bis viermal so aufwändig wie das Cholesky-Verfahren für die Normalgleichungen, aber stabil.

c) Bei Hessenberg-Matrizen, d.h. Matrizen mit der Gestalt

$$A = \begin{pmatrix} * & \dots & * \\ * & * & \\ & \ddots & \ddots \\ 0 & & * \end{pmatrix}, \quad (4.4.6)$$

also $a_{ik} = 0 \forall i < k + 1$, sind nur $(n - 1)$ Givens-Rotationen auszuführen.

Diese Matrizen tauchen z.B. bei Eigenwertberechnungen auf und sind dort ein wichtiger Bestandteil der Verfahren.

4.4b) Householder-Reflexion

Es sei H eine Hyperebene im \mathbb{R}^m und zusätzlich ein Vektor $a \in \mathbb{R}^m$ gegeben.

IMAGE MISSING

Gesucht ist nun die Reflexion Q , so dass

$$Qa = \alpha e_1 = \begin{pmatrix} \alpha \\ 0 \\ \vdots \\ 0 \end{pmatrix} \quad \text{mit } \alpha = \pm \|a\|$$

Mit

$$v = a - \alpha e_1 \quad (4.4.7)$$

gegeben, welches senkrecht zu H steht.

Damit Stellenauslöschungen in v , d.h. in v_1 , vermieden werden, wähle ein entsprechendes Vorzeichen für α , also

$$\alpha = -\text{sign}(a_1) \|a\|_2 \quad (4.4.8)$$

Die zugehörige Reflexion Q ist gegeben durch

IMAGE MISSING

wobei

$$w = \left\langle \frac{v}{\|v\|}, x \right\rangle \cdot \frac{v}{\|v\|}$$

$$Qx = x - 2w = x - 2 \frac{v^T x}{v^T v} v = \left(I - 2 \frac{vv^T}{v^T v} \right) x \quad (4.4.9)$$

$$Q = I - 2 \frac{vv^T}{v^T v} \quad vv^T \in \mathbb{R}^{n \times n}, \quad v^T v \in \mathbb{R} \quad (4.4.10)$$

und heißt **Householder Reflexion** (wurde 1958 von Householder eingeführt).

Für die spezielle Wahl (4.4.7) mit (4.4.8) vom Vektor v folgt

$$\begin{aligned} vv^T &= \|v\|^2 = \|a\|^2 - 2\alpha \langle a, e_1 \rangle + \alpha^2 \\ &= -2\alpha(a_1 - \alpha) \\ &= -2\alpha v \end{aligned} \quad (4.4.11)$$

4.4.4 Bemerkung

- a) Q ist symmetrisch
- b) Q ist orthogonal
- c) Q ist involutorisch, d.h. $Q^2 = I$ (bzw. gilt $Q^{-1} = Q^T = Q$)

Die Householder Reflexion setzt nicht nur eine Null, sondern im Vektor gleich alle gewünschten Nullen

$$\begin{array}{ll} \textbf{Rotation:} & \begin{pmatrix} * \\ * \\ * \end{pmatrix} \rightarrow \begin{pmatrix} * \\ * \\ 0 \end{pmatrix} \rightarrow \begin{pmatrix} * \\ 0 \\ 0 \end{pmatrix} \\ \textbf{Reflexion:} & \begin{pmatrix} * \\ * \\ * \end{pmatrix} \rightarrow \begin{pmatrix} * \\ 0 \\ 0 \end{pmatrix} \end{array}$$

4 Lineare Gleichungssysteme: Direkte Methoden (Fortsetzung)

Um die erste Spalte in A auf die gewünschte Gestalt zu bringen, bestimme $Q^{(1)}$ wie oben, indem die erste Spalte als a gewählt wird:

$$A \rightarrow A^{(1)} = Q^{(1)} A = \begin{pmatrix} \alpha^{(1)} & & \\ 0 & * & \\ \vdots & & \\ 0 & & \end{pmatrix}$$

In der k -ten Spalte sollen nun die $(k-1)$ -ten Zeilen und die $(k-1)$ -ten Spalten bleiben und die Restmatrix verändert werden.

$$A^{(k-1)} = \begin{pmatrix} * & & & & \\ & \ddots & & & \\ & & * & & \\ & & 0 & | & \text{---} & \text{---} & \text{---} \\ & 0 & \vdots & | & T^{(k-1)} & & \\ & & \vdots & | & & & \\ & & 0 & | & & & \end{pmatrix} \begin{matrix} \leftarrow (k-1)\text{-te Zeile} \\ \\ \\ \uparrow (k-1)\text{-te Spalte} \end{matrix}$$

Setze also

$$Q^{(k)} = \begin{pmatrix} I_{k-1} & 0 \\ 0 & \overline{Q}^{(k)} \end{pmatrix} \quad (4.4.12)$$

wobei $\overline{Q}^{(k)}$ durch die erste Spalte von $T^{(k)}$, d.h.

$$a = (a_{i,k}^{k-1})_{i=k,\dots,m} \subset \mathbb{R}^{m+1-k} \quad (4.4.13)$$

bestimmt wird. Dann gilt

$$Q^{(k)} A^{(k-1)} = \begin{pmatrix} * & & & & \\ & \ddots & & & \\ & & * & & \\ & & 0 & | & * & & \\ & 0 & \vdots & | & 0 & * & \\ & & \vdots & | & \vdots & & \ddots \\ & & 0 & | & 0 & & & * \end{pmatrix}$$

Nach insgesamt

$$p = \min(m-1, n) \quad (4.4.14)$$

Abbildung 4.3: Zerlegung einer $m \times n$ -Matrix für $m < n$

Schritten erhalten wir für $A \in \mathbb{R}^{m \times n}$

$$Q^T A = Q^{(p)} \cdot \dots \cdot Q^{(1)} A = \begin{pmatrix} R \\ 0 \end{pmatrix}, \quad (4.4.15)$$

wobei Bemerkung 4.4.4 auch für $Q^T = Q^{(p)} \cdot \dots \cdot Q^{(1)}$ und somit auch für

$$Q = Q^{(1)} \cdot \dots \cdot Q^{(p)} \quad (4.4.16)$$

gilt.

17.11.2014

4.4.5 Speicherung

Gespeichert werden müssen die obere Dreiecksmatrix R und die **Householdervektoren** $v^{(i)} \in \mathbb{R}^{m+1-i}$. Die Diagonalelemente von R sind $r_{ii} = \alpha^{(i)}$.

Folgende Speicheraufteilung ist möglich:

$$A \rightarrow \left(\begin{array}{c|c|c|c|c} & & & & R \\ \hline v^{(1)} & & & & \\ \hline & v^{(2)} & & & \\ \hline & & \ddots & & \\ \hline & & & v^{(p)} & \\ \hline \end{array} \right) \quad \text{und} \quad \begin{pmatrix} \alpha^{(1)} \\ \vdots \\ \alpha^{(p)} \end{pmatrix}$$

Wohlgemerkt kann so auch $A \in \mathbb{R}^{m \times n}$ mit $m < n$ bearbeitet werden, dann wird A zu:

Falls zusätzlich $v^{(i)}$ so normiert ist, dass $v_1^{(i)} = 1$ ist, braucht diese Komponente nicht gespeichert werden und R kann komplett in A gespeichert werden.

4.4.6 Aufwand für den Householder-QR-Algorithmus

- Falls $m \approx n$ sind ungefähr $\frac{2}{3}n^3$ Multiplikationen notwendig und ist somit ungefähr doppelt so teuer wie die LR-Zerlegung, ist aber stabil.
- Falls $m \gg n$ sind ungefähr $2mn^2$ Multiplikationen notwendig. Der Aufwand ist daher ungefähr so hoch wie beim Cholesky-Verfahren für Normalgleichungen, aber stabil.

5 Numerische Lösung nichtlinearer Gleichungssysteme

Beispiel:

linear: $Ax = b$

nichtlinear: $f(x) = \sin(x) + x^3 - 4 = 0$

5.1 Einführung

5.1.1 Beispiele

- 1) $f(x) = x^2 - c = 0 \Leftrightarrow x = \pm\sqrt{c}$:
Berechnung der Wurzel
- 2) Sei p ein Polynom:
Nullstellenbestimmung
- 3) Löse das nichtlineare Randwertproblem

$$-\Delta u = f(u)$$

in $\Omega = (0, 1)^2$ mit $u = 0$ auf $\partial\Omega$.

Mit dem Differenzenverfahren¹ ergibt sich

$$A\vec{u} = h^2 \vec{f}(\vec{u})$$

ein System nichtlinearer Gleichungen.

Nullstellenbestimmung

Gegeben $D \subseteq \mathbb{R}^n, f : D \rightarrow \mathbb{R}^m$ stetig

Gesucht $x^* \in D$ mit $f(x^*) = 0$

¹s. Übungsaufgabe 2)



Abbildung 5.1: Beispiel zur Nullstellenexistenz

Fixpunktgleichung

Gegeben $D \subseteq \mathbb{R}^n, g: D \rightarrow \mathbb{R}^n$ stetig

Gesucht $x^* \in D$ mit $g(x^*) = x^*$

Falls $m=n$ ist dies äquivalent zur Nullstellenbestimmung.

5.1.2 Das Bisektionsverfahren

Sei $f: [a, b] \rightarrow \mathbb{R}$ stetig und $f(a) \cdot f(b) < 0$.

Dann folgt aus dem Zwischenwertsatz die Existenz mindestens einer Nullstelle $x^* \in (a, b)$.

Generiere eine Folge von Intervallen $[a^{(i)}, b^{(i)}] \in [a^{(i-1)}, b^{(i-1)}]$, die eine Nullstelle enthalten und mit $b^{(i)} - a^{(i)} \rightarrow 0$. Definiere

$$x^{(i+1)} = \frac{1}{2}(b^{(i)} + a^{(i)}) \quad (5.1.1)$$

und

$$[a^{(i+1)}, b^{(i+1)}] := \begin{cases} [x^{(i+1)}, b^{(i)}] & \text{für } f(a^{(i)}) \cdot f(x^{(i+1)}) > 0 \\ [a^{(i)}, x^{(i+1)}] & \text{für } f(a^{(i)}) \cdot f(x^{(i+1)}) < 0 \end{cases} \quad (5.1.2)$$

Für jedes $i \geq 1$ gilt somit

$$b^{(i)} - a^{(i)} = \frac{1}{2^i}(b - a)$$

und es existiert eine Nullstelle x^* in $[a^{(i)}, b^{(i)}] \in [a^{(i-1)}, b^{(i-1)}]$ für alle i .
Damit folgt

$$\begin{aligned} |x^{(i-1)} - x^*| &\leq \frac{1}{2}(b^{(i)} - a^{(i)}) \\ &= 2^{-(i+1)}(b - a) \rightarrow 0 \end{aligned}$$

Also $\lim_{i \rightarrow \infty} x^{(i)} = x^*$.

5.1.3 Korollar

Das oben angegebene Bisektionsverfahren konvergiert, falls $f : [a, b] \rightarrow \mathbb{R}$ stetig ist und $f(a) \cdot f(b) < 0$ gilt.

5.1.4 Bemerkungen

a) $x^{(i)}$ wird als Intervallmitte, also unabhängig von $f(x^{(i)})$ gewählt. die Konvergenzgeschwindigkeit hängt von der Länge des Intervalls $[a, b]$ ab und der Lage von x^* bezüglich der Intervallhalbierung ab.

Die Konvergenz kann demnach sehr langsam sein.

b) Ein Vorteil ist, dass keine Differenzierbarkeitsvoraussetzungen nötig sind.

c) Das Verfahren ist nicht für $f : D \rightarrow \mathbb{R}^n$ anwendbar.

5.2 Fixpunktiteration

Gesucht sei ein Fixpunkt $x^* \in D \subseteq \mathbb{R}^n$ der stetigen Funktion $g : D \rightarrow \mathbb{R}^n$, d.h.

$$x^* = g(x^*) \quad (5.2.1)$$

Idee: Nutze (5.2.1) zur Iteration, d.h. wähle $x^{(0)} \in D$, setze

$$x^{(k+1)} = g(x^{(k)}) \quad \text{für } k \in 0, 1, \dots \quad (5.2.2)$$

Es bedarf noch der Voraussetzung, dass $x^{(k)} \in D \quad \forall k$

Falls $x^{(k)}$ konvergiert, ist der Grenzwert x^* ein Fixpunkt, denn für stetiges g gilt:

$$\begin{aligned} x^* &= \lim_{k \rightarrow \infty} x^{(k+1)} = \lim_{k \rightarrow \infty} g(x^{(k)}) \\ &\stackrel{g \text{ stetig}}{=} g\left(\lim_{k \rightarrow \infty} x^{(k)}\right) = g(x^*) \end{aligned} \quad (5.2.3)$$

5.2.1 Beispiel

Löse $x - e^{-x} - 1 = 0$.

a) $x = 1 + e^{-x} =: g_1(x)$
 \rightarrow Konvergenz

b) $e^{-x} = x - 1 \Leftrightarrow x = -\ln(x - 1) =: g_2(x)$
 $\rightarrow g(x^{(2)})$ nicht definiert!



Abbildung 5.2: Konvergenz der Fixpunktiteration für $x = 1 + e^{-x}$



Abbildung 5.3: Versagen der Fixpunktiteration für $x = -\ln(x - 1)$

5.2.2 Definition: Kontraktion

Sei $D \subseteq \mathbb{R}^n$ abgeschlossen und $\|\cdot\|$ eine Norm auf dem \mathbb{R}^n . Eine Abbildung $g : D \rightarrow \mathbb{R}^n$ heißt **Kontraktion** bezüglich $\|\cdot\|$, falls es ein $\kappa \in [0, 1)$ gibt mit

$$\|g(u) - g(v)\| \leq \kappa \|u - v\| \quad \forall u, v \in D$$

Die kleinste solche Zahl κ heißt Kontraktionszahl von g .

Offensichtlich ist jede auf D kontrahierende Abbildung stetig.

5.2.3 Lemma

Sei $D = \overline{\Omega}$ mit $\Omega \subseteq \mathbb{R}^n$ offen und konvex und $\|\cdot\|$ eine Norm auf dem \mathbb{R}^n .

Falls $g : D \rightarrow \mathbb{R}^n$ eine stetig differenzierbare Funktion ist und bezüglich der zugeordneten Matrixnorm $\sup_{x \in \Omega} \|Dg(x)\| < 1$ gelte, so ist g kontrahierend bezüglich $\|\cdot\|$.

Beweis. Mit $u, v \in D$ gilt $u + t(v - u) \in D$, da D konvex ist.

Somit ist $h : [0, 1] \rightarrow \mathbb{R}^n$ mit $h(t) := g(u + t(v - u))$ wohldefiniert und stetig differenzierbar.



Abbildung 5.4: Grafische Veranschaulichung einer Kontraktion

Mit dem Hauptsatz der Differenzial- und Integralrechnung folgt:

$$\begin{aligned}
 \|g(u) - g(v)\| &= \|h(1) - h(0)\| \\
 &= \left\| \int_0^1 h'(t) dt \right\| \\
 &= \left\| \int_0^1 Dg(u + t(v - u)) \cdot (v - u) dt \right\| \\
 &\leq \int_0^1 \|Dg(u + t(v - u))\| dt \cdot \|v - u\| \\
 &\leq \underbrace{\sup_{x \in \Omega} \|Dg(x)\|}_{=: \kappa} \cdot \|v - u\|
 \end{aligned} \tag{5.2.4}$$

□

5.2.4 Banachscher Fixpunktsatz

Sei $D \subset \mathbb{R}^n$ abgeschlossen und die Abbildung $g : D \rightarrow \mathbb{R}^n$ eine Kontraktion. Dann gilt:

- 1) Es existiert genau ein Fixpunkt x^* von g .
- 2) Für jeden Startwert $x^{(0)} \in D$ konvergiert die Folge der Fixpunktiterierten

$$x^{(k+1)} := g(x^{(k)}) \xrightarrow{k \rightarrow \infty} x^* \tag{5.2.5}$$

- 3) Es gelte die a posteriori Fehlerabschätzung

$$\|x^{(k)} - x^*\| \leq \frac{\kappa}{1 - \kappa} \|x^{(k)} - x^{(k-1)}\| \tag{5.2.6}$$

und die a priori Fehlerabschätzung

$$\|x^{(k)} - x^*\| \leq \frac{\kappa^k}{1 - \kappa} \|x^{(1)} - x^{(0)}\| \tag{5.2.7}$$

19.11.2014

Beweis. zu 2) Sei $x_0 \in D$ beliebig. (5.2.5) ist wohldefiniert, da $g(D) \subset D$. $(x^{(k)})_{k \in \mathbb{N}}$

5 Numerische Lösung nichtlinearer Gleichungssysteme

bilden eine Cauchyfolge, da

$$\begin{aligned}
 \|x^{(k+1)} - x^{(k)}\| &= \|g(x^{(k)}) - g(x^{(k-1)})\| \\
 &\leq \kappa \|x^{(k)} - x^{(k-1)}\| \\
 &\leq \kappa^k \|x^{(1)} - x^{(0)}\| \\
 \implies \|x^{(k+l)} - x^{(k)}\| &\leq \sum_{i=0}^l \|x^{(k+i+1)} - x^{(k+i)}\| \\
 &\leq \sum_{i=0}^l \kappa^{k+1} \|x^{(1)} - x^{(0)}\| \\
 &\leq \frac{\kappa^k}{1 - \kappa} \|x^{(1)} - x^{(0)}\| \quad \forall l \in \mathbb{N}
 \end{aligned}$$

Daraus folgt, dass $\lim_{k \rightarrow \infty} x^{(k)} = x^*$ existiert und $x^* \in D$, da D abgeschlossen ist und somit vollständig.

zu 1) Da g stetig ist, ist $g(x^*) = x^*$ (siehe hierzu (5.2.3)).
 x^* ist eindeutiger Fixpunkt, da für einen weiteren Fixpunkt y^* gilt

$$0 \leq \|x^* - y^*\| = \|g(x^*) - g(y^*)\| \leq \kappa \|x^* - y^*\|$$

Da $\kappa < 1$, muss $\|x^* - y^*\| = 0$ sein und damit $x^* = y^*$.

zu 3) Betrachte

$$\begin{aligned}
 \|x^* - x^{(k)}\| &= \lim_{l \rightarrow \infty} \|x^{(k+l)} - x^{(k)}\| \\
 &\leq \frac{\kappa^k}{1 - \kappa} \|x^{(1)} - x^{(0)}\|
 \end{aligned}$$

bzw.

$$\begin{aligned}
 \lim_{l \rightarrow \infty} \|x^{(k+l)} - x^{(k)}\| &\leq \lim_{l \rightarrow \infty} \sum_{i=0}^{l-1} \|x^{(k+i+1)} - x^{(k+i)}\| \\
 &\leq \lim_{l \rightarrow \infty} \sum_{i=0}^{l-1} \kappa^{k+i+1} \|x^{(1)} - x^{(0)}\| \\
 &\leq \frac{\kappa^k}{1 - \kappa} \|x^{(1)} - x^{(0)}\|
 \end{aligned}$$

□

5.2.5 Bemerkung

- 1) Als Voraussetzung wäre bereits ausreichend:
 D ist vollständiger metrischer Raum mit Metrik d .
 Dann ersetze die Norm durch die Metrik d .
- 2) Im Allgemeinen ist der Nachweis $g(D) \subset D$ schwierig.

5.2.6 Folgerungen

Sei $x^* \in \mathbb{R}^n$, so dass $g(x^*) = x^*$ und sei g in einer Umgebung von $\overline{B_\varepsilon(x^*)} = \{x \in \mathbb{R}^n \mid \|x - x^*\| \leq \varepsilon\}$ stetig differenzierbar und es gelte $\|g'(x)\| < 1$ für $x \in \overline{B_\varepsilon(x^*)}$, so ist Satz 5.2.4 mit $D = \overline{B_\varepsilon(x^*)}$ anwendbar.

Beweis. Nutze (5.2.4) und Lemma 5.2.3. □

5.3 Konvergenzordnung und Fehlerabschätzungen

5.3.1 Definition: lineare Konvergenz

Eine Folge $(x^{(k)})_{k \in \mathbb{N}}$ mit $x^{(k)} \in \mathbb{R}^n$ **konvergiert** mit (mindestens) der **Ordnung** $p \geq 1$ gegen x^* , falls

$$\lim_{k \rightarrow \infty} x^{(k)} = x^*$$

und falls es ein $C > 0$ sowie $N \in \mathbb{N}$ gibt, so dass

$$\|x^{(k+1)} - x^*\| \leq C \|x^{(k)} - x^*\|^p \quad \forall k \geq N$$

Im Fall $p = 1$ ist zusätzlich $C < 1$ und man spricht von **linearer Konvergenz**.

Für $p = 2$ heißt es **quadratische Konvergenz**.

Gilt

$$\lim_{k \rightarrow \infty} \frac{\|x^{(k+1)} - x^*\|}{\|x^{(k)} - x^*\|} = 0,$$

so konvergiert die Folge **superlinear**.

5.3.2 Bemerkung

Die Fixpunktiteration konvergiert unter der Voraussetzung in 5.2.4 mindestens linear.

5.3.3 Bemerkung

- a) lineare Konvergenz hängt von der gewählten Norm ab.
- b) Hat die Folge bzgl. einer Vektornorm auf dem \mathbb{R}^n die Konvergenzordnung $p > 1$, hat sie diese bzgl. jeder Norm.

5.3.4 Definition: lokale und globale Konvergenz

- a) Ein iteratives Verfahren zur Bestimmung eines Wertes x^* hat die Konvergenzordnung p , falls es eine Umgebung U um x^* gibt, so dass für alle Startwerte aus $U \setminus \{x^*\}$ die erzeugte Folge mit Ordnung p konvergiert.
- b) Das Verfahren heißt **lokal konvergent**, falls es für alle Startwerte in einer Umgebung von x^* konvergiert.
- c) Das Verfahren heißt **global konvergent**, falls es im gesamten Definitionsbereich des zugehörigen Problems konvergiert.

5.3.5 Lemma

Sei $(x^{(k)})_{k \in \mathbb{N}}$ eine konvergente Folge in \mathbb{R} mit Grenzwert x^* .

- a) Falls

$$\lim_{k \rightarrow \infty} \frac{\|x^{(k+1)} - x^*\|}{\|x^{(k)} - x^*\|} = A \in (-1, 1), \quad A \neq 0 \quad (5.3.1)$$

hat die Folge genau die Konvergenzordnung 1. Weiter gilt mit $A_k = \frac{x^{(k)} - x^{(k-1)}}{x^{(k-1)} - x^{(k-2)}}$

$$\begin{aligned} \lim_{k \rightarrow \infty} \frac{A_k}{1 - A_k} \cdot \frac{x^{(k)} - x^{(k-1)}}{x^* - x^{(k)}} &= 1 \\ \lim_{k \rightarrow \infty} A_k &= A \end{aligned} \quad (5.3.2)$$

- b) Falls die Folge Konvergenzordnung $p > 1$ hat, gilt

$$\lim_{k \rightarrow \infty} \frac{x^{(k)} - x^{(k-1)}}{x^* - x^{(k)}} = 1 \quad (5.3.3)$$

zu Def. 5.3.1: Im Fall $p = 1$ ist zusätzlich $c < 1$ verlangt.

Beweis. (skizzenhaft, siehe Übungsaufgaben)

Sei $e^{(k)} := x^* - x^{(k)}$.

Nutze $x^{(k+1)} - x^{(k)} = e^{(k)} - e^{(k+1)}$:

a) Zeige

$$\lim_{k \rightarrow \infty} \frac{x^{(k)} - x^{(k-1)}}{e^{(k)}} = \frac{1-A}{A},$$

sowie

$$\lim_{k \rightarrow \infty} A_k = A,$$

so folgt die Behauptung.

b) Folgt aus $\lim_{k \rightarrow \infty} \frac{e^{(k+1)}}{e^{(k)}} = 0$.

□

5.3.6 Folgerung: a posteriori Fehlerabschätzung

a) Für $p = 1$ gilt

$$x^* - x^{(k)} \approx \frac{a_k}{1 - A_k} (x^{(k)} - x^{(k-1)}) \quad (5.3.4)$$

für große k und A_k in etwa konstant.

$|x^{(k)} - x^{(k-1)}|$ ist im Allgemeinen **keine** sinnvolle Schätzung des Fehlers $|x^* - x^{(k)}|$!

b) Für $p > 1$ gilt:

$$x^* - x^{(k)} \approx x^{(k+1)} - x^{(k)} \quad (5.3.5)$$

für große k .

5.3.7 Bemerkung

Für Folgen im \mathbb{R}^n gibt es für $p = 1$ kein Analogon zu (5.3.4). Falls $p > 1$, lässt sich (5.3.3) für die Normen der Differenzen zeigen, d.h.

$$\|x^* - x^{(k)}\| \approx \|x^{(k+1)} - x^{(k)}\| \quad (5.3.6)$$

Beweis. Nutze $\lim_{k \rightarrow \infty} \frac{\|e^{(k+1)}\|}{\|e^{(k)}\|} = 0$ und

$$\|e^{(k)}\| - \|e^{(k+1)}\| \leq \|x^{(k+1)} - x^{(k)}\| \leq \|e^{(k)}\| + \|e^{(k+1)}\|.$$

□



Abbildung 5.5: Veranschaulichung des Newton-Verfahrens an einem Funktionsgraphen

5.3.8 Folgerung

Falls $p > 1$ ist, kann p folgendermaßen approximiert werden:

$$p \approx \frac{\log(\|x^{(k+2)} - x^{(k+1)}\|)}{\log(\|x^{(k+1)} - x^{(k)}\|)}$$

Beweis. siehe Übungsaufgabe. □

5.4 Newton-Verfahren für skalare Gleichung

Sei $f : [a, b] \rightarrow \mathbb{R}$ differenzierbar. Dann gilt

$$f(x^*) = f(x) + f'(x)(x^* - x) + o(\|x - x^*\|),$$

d.h. f kann lokal gut durch die Tangente approximiert werden.
Betrachte die Nullstellengleichung $f(x^*) = 0$

und bestimme iterativ die Nullstelle der Tangentengleichung

$$0 = f(x) + f'(x)(\bar{x} - x) \Leftrightarrow \bar{x} = x - \frac{f(x)}{f'(x)}$$

Notwendig ist hier die Bedingung $f'(x) \neq 0$.

5.4.1 Iterationsschritt des Newton(-Kantorowitsch)-Verfahrens

$$x^{(k+1)} = x^{(k)} - \frac{f(x^{(k)})}{f'(x^{(k)})} \tag{5.4.1}$$

wird auch Tangentenverfahren genannt und stammt von J. Raphson (1630). Newton hat eine ähnliche Technik früher angewendet.

5.4.2 Satz

Sei $f \in C^1(a, b)$ und $x^* \in (a, b)$ eine einfache Nullstelle von f , d.h. $f'(x^*) \neq 0$.

Dann gibt es ein $\varepsilon > 0$, s.d. für jedes $x^{(0)} \in \overline{B_\varepsilon(x^*)}$ das Newton-Verfahren (5.4.1) superlinear gegen x^* konvergiert.

Falls $f \in C^2(a, b)$ tritt mindestens quadratische Konvergenz ein, d.h. das Verfahren konvergiert lokal quadratisch.

Beweis. Gleichung (5.4.1) definiert eine Fixpunktiteration mit $g(x) = x - \frac{f(x)}{f'(x)}$.

Für $f \in C^2(a, b)$ gilt

$$g'(x) = 1 - \frac{f'f' - ff''}{(f')^2}(x) = \frac{f(x)f''(x)}{(f'(x))^2}.$$

Da $f(x^*) = 0$ und $f'(x^*) \neq 0$ gilt $g'(x^*) = 0$.

Weiterhin gibt es eine Umgebung U_0 von x^* , in der $f(x) \neq 0 \ \forall x \in U_0$, da f stetig ist.

In U_0 ist somit g' stetig. Da $g'(x^*) = 0$ ist, existiert ein $\varepsilon > 0$ mit

$$g'(x) \leq \kappa < 1 \quad \forall x \in \overline{B_\varepsilon(x^*)}.$$

Da $g(x^*) = x^*$ ist, ist die Folgerung 5.2.6 anwendbar, also ist g eine Kontraktion. und $g(\overline{B_\varepsilon(x^*)}) \subset \overline{B_\varepsilon(x^*)}$. Der Banachsche Fixpunktsatz liefert Konvergenz für alle $x^{(0)} \in \overline{B_\varepsilon(x^*)}$.

Die quadratische Konvergenz folgt aus

$$\begin{aligned} |x^{(k-1)} - x^*| &= \left| x^{(k)} - \frac{f(x^{(k)})}{f'(x^{(k)})} - x^* + \frac{f(x^*)}{f'(x^*)} \right| \\ &= \frac{|f(x^{(k)}) - f(x^*) + f'(x^{(k)})(x^{(k)} - x^*)|}{|f'(x^{(k)})|} \\ &\leq \sup_{x \in \overline{B_\varepsilon(x^*)}} \frac{1}{|f'(x^*)|} \cdot \sup_{x \in \overline{B_\varepsilon(x^*)}} |f''(x)| \cdot \frac{1}{2} |x^{(k)} - x^*|^2 \end{aligned}$$

aufgrund der Taylorentwicklung und da $x^{(k-1)} \in \overline{B_\varepsilon(x^*)} \ \forall k \in \mathbb{N}$ (da g Kontraktion).

Für $f \in C^1$ siehe Hämmerlin[3]. □

5.4.3 Bemerkung

- a) Mehrfache Nullstellen könne im Allgemeinen nicht mit (5.4.1) bestimmt werden.
- b) Die Ableitung f' muss analytisch (als Funktion) gegeben sein.
- c) Die Lage und Größe des Konvergenzinterfalls ist a priori unbekannt.
(Hierfür könnte z.B. das Bisektionsverfahren Anwendung finden.)

Index

- Äquilibrierung
 - Spalten-, 51
 - Zeilen-, 51
- Algorithmus, 43
- Ausgleichsproblem
 - linear, 55, 57
- Banachscher Fixpunktsatz, 75
- Basis, 24
- Bildraum, 57
- Bisektionsverfahren, 72
- Cholesky-Zerlegung, 54, 55, 60
- double, 24
- Dreieckszerlegung, 9, 13
- elementar ausführbar, 43
- entkoppelt, 38
- Fehler, 23, 29
 - absoluter, 29
 - absoluter Rundungsfehler, 26
 - Fortpflanzung, 43
 - relativer, 29
- Fixpunktiteration, 72, 73
- floating point, 23, 24
- floating point operations, 13
- flops, 13
- Frobeniusmatrix, 16
- Güte
 - Algorithmus, 45
- Gauß-Eliminator, 13
- Gaußsches Eliminationsverfahren, 9, 15
- Genauigkeit
 - Maschinengenauigkeit, 26
 - relative Rechengenauigkeit, 26
- Givens-QR-Algorithmus, 65
- Givens-Rotation, 63
- Gleitkommazahl, 24
- Householder Reflexion, 67
- Householdervektoren, 69
- Implementation, 43
- integer, 23
- Kondition
 - Addition, 32
 - gut/schlecht konditioniert, 31
 - komponentenweise, 39
 - Matrix, 35
 - normweise, absolut, 31
 - normweise. relativ, 31
- Kontraktion, 74
- Konvergen
 - linear, 77
- Konvergenz
 - global, 78
 - lokal, 78
 - Ordnung, 77
 - quadratisch, 77
 - superlinear, 77
- längenerhaltend, 61
- Landau-Symbole, 14
- LR-Zerlegung, 15, 16
- LU-Zerlegung, 16
- Mantisse, 24
- Maximum-Likelihood-Methode, 57

Index

- Nachiteration, 52
- Neumannsche Reihe, 36
- Newton-Verfahren, 80
- Norm, 29
 - Euklidische Norm, 29
 - Frobeniusnorm, 30
 - Hölder-Norm, 29
 - Matrixnorm, 29
 - Maximumsnorm, 29
 - Spaltensummennorm, 30
 - submultiplikative, 30
 - Summennorm, 29
 - verträglich, 30
- normalisierte Gleitkommazahl, 24
- normweise Kondition, 31
- Nullstellenbestimmung, 71

- orthogonale Projektion, 58
- Orthogonalisierung, 62

- p-Norm, 30
- Permutationsmatrix, 19
- Pivotelement, 12
- Pivotisierung, 18
 - halbmaximale, 18
 - partielle, 18
 - Spalten-, 18
 - vollständige, 19
 - Zeilen-, 19
- Problem, 28
- Projektionssatz, 58

- Rückwärtsanalyse, 46
- Rückwärtssubstitution, 11
- Realisierung, 43
- Rechenaufwand, 13, 15
- Rundungsfehler, 23

- scharf, 33
- Singulärwert, 30
- Skalierung
 - Spalten-, 51
 - Zeilen-, 51
- spd Matrix, 52
- Stabilität, 45

- strikt diagonal dominant, 48
- Tangentenverfahren, 80

- unipotent, 54

- Verfahren von Crout, 17
- Vorwärtselimination, 9, 21
- Vorwärtssubstitution, 9, 11, 13

- Zahlendarstellung, 23
- Zeilensummennorm, 30

Literatur

- [1] P. Deuffhard und A. Hohmann. Numerische Mathematik. I, Eine algorithmisch orientierte Einführung. 3. Aufl. de Gruyter, 2002.
- [2] von G. Golub und J.M. Ortega. Scientific Computing.
- [3] K.H. Hoffmann G. Haemmerlin. Numerische Mathematik. Springer Berlin.
- [4] W.H. Press u. a. Numerical Recipes in C++. Cambridge University Press.
- [5] R.W. Hoppe R.W. Freund. Stoer/Bulirsch: Numerische Mathematik 1. Springer.
- [6] J. Stoer und R. Bulirsch. Numerische Mathematik 2. Springer.
- [7] W.Dahmen und A. Reusken. Numerik fuer Ingenieure und Naturwissenschaftler. Springer.