4.2.5 Das Cholesky-Verfahren

Für allgemeine invertierbare Matrizen kann das Gauß-Verfahren ohne Pivotsuche zusammenbrechen und wir werden sehen, dass auch aus Gründen der numerischen Stabilität eine Pivotsuche ratsam ist. Für die wichtige Klasse der positiv definiten Matrizen ist jedoch das Gauß-Verfahren immer ohne Pivotsuche numerisch stabil durchführbar.

Definition 4.2.4 *Eine reelle Matrix $A \in \mathbb{R}^{n,n}$ heißt positiv definit, falls gilt*

$$A = A^T, \quad x^T Ax > 0 \quad \forall\, x \in \mathbb{R}^n \setminus \{0\}.$$

und positiv semi-definit, falls gilt

$$A = A^T, \quad x^T Ax \geq 0 \quad \forall\, x \in \mathbb{R}^n.$$

Allgemeiner heißt eine komplexe Matrix $A \in \mathbb{C}^{n,n}$ positiv definit, falls gilt

$$A = A^H, \quad x^H Ax > 0 \quad \forall\, x \in \mathbb{C}^n \setminus \{0\}.$$

und positiv semi-definit, falls gilt

$$A = A^H, \quad x^H Ax \geq 0 \quad \forall\, x \in \mathbb{C}^n.$$

Hierbei ist $A^H = (\bar{a}_{ji})_{1\leq i\leq n, 1\leq j\leq n}$, wobei $\bar{a}_{ji}$ komplexe Konjugation bezeichnet.

Positiv definite Matrizen treten sehr oft in Anwendungen auf, etwa bei der numerischen Lösung von elliptischen (z.B. Laplace-Gleichung) und parabolischen (z.B. Wärmeleitungsgleichung) partiellen Differentialgleichungen.

Satz 4.2.5 *Für eine positiv definite Matrix A existiert A^{-1} und ist wieder positiv definit. Zudem gilt: Alle Eigenwerte sind positiv, alle Hauptuntermatrizen $A_{kl} = (a_{ij})_{k\leq i,j\leq l}$, $1 \leq k \leq l \leq n$ sind wieder positiv definit und alle Hauptminoren $\det(A_{kl})$ sind positiv.*

Beweis: Elementare Übung aus der linearen Algebra. Siehe z.B. Stoer [St94]. □

Eine effiziente Variante des Gaußschen Verfahrens für Gleichungssysteme mit positiv definiter Matrix wurde von Cholesky angegeben. Das Cholesky-Verfahren beruht auf der folgenden Beobachtung

Satz 4.2.6 *Es sei $A \in \mathbb{R}^{n,n}$ positiv definit. Dann gibt es genau eine untere Dreieckmatrix L mit positiven Diagonaleinträgen $l_{ii} > 0$, so dass*

$$LL^T = A \quad \textit{(Cholesky-Zerlegung)}.$$

Ferner besitzt A eine eindeutige Dreieckszerlegung

$$\tilde{L}\tilde{R} = A,$$

wobei $\tilde{L} = LD^{-1}$, $\tilde{R} = DL^T$ mit $D = \operatorname{diag}(l_{11}, \ldots, l_{nn})$. Sie wird vom Gauß-Verfahren ohne Pivotsuche geliefert.

Der Beweis kann durch vollständige Induktion nach n erfolgen, wir wollen ihn aber nicht ausführen.

Die Cholesky-Zerlegung $LL^T = A$ berechnet man durch Auflösen der $\frac{n(n+1)}{2}$ Gleichungen (aus Symmetriegründen muss nur das untere Dreieck mit Diagonale betrachtet werden)

$$a_{ij} = \sum_{k=1}^{j} l_{ik} l_{jk}, \quad \text{für } j \le i, \ \ i = 1, \dots, n. \tag{4.13}$$

Man kann hieraus die Elemente von L spaltenweise in der Reihenfolge

$$l_{11}, \dots, l_{n1}, l_{22}, \dots, l_{n2}, \dots, l_{nn}$$

berechnen. Für die erste Spalte von L (setze $j = 1$) ergibt sich

$$\begin{aligned} a_{11} &= l_{11}^2, \text{also } \ l_{11} = \sqrt{a_{11}} \\ a_{i1} &= l_{i1} l_{11}, \text{also } \ l_{i1} = a_{i1}/l_{11}. \end{aligned}$$

Sukzessives Auflösen nach l_{ij}, $i = j, \dots, n$ liefert den folgenden Algorithmus.

Algorithmus 3 Cholesky-Verfahren zur Berechnung der Zerlegung $LL^T = A$

Für $j = 1, \dots, n$

$$l_{jj} = \sqrt{a_{jj} - \sum_{k=1}^{j-1} l_{jk}^2}$$

Für $i = j + 1, \dots, n$*:*

$$l_{ij} = \frac{a_{ij} - \sum_{k=1}^{j-1} l_{ik} l_{jk}}{l_{jj}}$$

Bemerkung: Das Cholesky-Verfahren hat einige schöne Eigenschaften:

- Da das Cholesky-Verfahren die Symmetrie ausnutzt, benötigt es neben n Quadratwurzeln nur noch ca. $n^3/6$ Operationen. Dies ist etwa die Hälfte der beim Gauß-Verfahren benötigten Operationen.
- Aus (4.13) folgt
  $$|l_{ij}| \le \sqrt{a_{ii}}, \quad j \le i, \ \ i = 1, \dots, n.$$
  Die Elemente der Matrix L können daher nicht zu groß werden. Dies ist ein wesentlicher Grund für die numerische Stabilität des Cholesky-Verfahrens.
- Das Cholesky-Verfahren ist die effizienteste allgemeine Testmethode auf positive Definitheit. Man muss hierbei Algorithmus 3 nur wie folgt erweitern:
  $$a = a_{jj} - \sum_{k=1}^{j-1} l_{jk}^2. \qquad \textit{Falls } a \le 0\textit{: STOP, } A \textit{ nicht positiv definit.}$$
  Sonst setze $l_{jj} = \sqrt{a}$.

4.3 Fehlerabschätzungen und Rundungsfehlereinfluß

Bei der Beschreibung der direkten Verfahren zur Lösung von Gleichungssystemen sind wir bisher davon ausgegangen, dass alle Ausgangsdaten exakt vorliegen und die Rechnung ohne Rundungsfehler durchgeführt wird. Dies ist jedoch unrealistisch, denn insbesondere bei großen Systemen können Rundungsfehler die Rechnung erheblich beeinflussen.

4.3.1 Fehlerabschätzungen für gestörte Gleichungssysteme

Wir studieren zunächst, wie stark sich die Lösung eines linearen Gleichungssystems bei Störung von Matrix und rechter Seite ändert. Vorgelegt sei ein lineares Gleichungssystem

$$Ax = b$$

und ein gestörtes System

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

mit ΔA und Δb "klein".

Frage: Wie klein ist $x - \tilde{x}$?

Diese Fragestellung ist von größter praktischer Bedeutung:

- Man kann abschätzen, wie sensitiv die Lösung bezüglich Störungen von Matrix und rechter Seite ist.
- Eine berechnete Näherungslösung (z.B. mit einer Implementierung des Gauß-Verfahrens) $\tilde{x}$ von $Ax = b$ ist exakte Lösung des Systems

  $$A\tilde{x} = b + \Delta b, \quad \text{mit dem Residuum } \Delta b = A\tilde{x} - b.$$

  Man kann nun aus dem leicht berechenbaren Residuum $\Delta b = A\tilde{x} - b$ Schranken an den unbekannten Fehler $\|x - \tilde{x}\|$ ableiten.

Es stellt sich heraus, dass die sogenannte Kondition einer Matrix diesen Störeinfluss beschreibt.

Zur Messung von $x - \tilde{x}$, Δb und ΔA benötigen wir einen "Längenbegriff" für Vektoren und Matrizen.

Definition 4.3.1 *Eine* Vektornorm *auf* $\mathbb{R}^n$ *ist eine Abbildung* $x \in \mathbb{R}^n \mapsto \|x\| \in [0, \infty[$ *mit folgenden Eigenschaften:*

a) $\|x\| = 0$ *nur für* $x = 0$.

b) $\|\alpha x\| = |\alpha|\,\|x\|$ *für alle* $\alpha \in \mathbb{R}$ *und alle* $x \in \mathbb{R}^n$.

c) $\|x + y\| \leq \|x\| + \|y\|$ *für alle* $x, y \in \mathbb{R}^n$ *(Dreiecksungleichung).*

Nun sollen auch *Matrix-Normen* eingeführt werden. Sei hierzu $\|\cdot\|$ eine beliebige Norm auf $\mathbb{R}^n$. Dann können wir auf $\mathbb{R}^{n,n}$ eine zugehörige Matrix-Norm definieren durch

$$\|A\| := \sup_{\|x\|=1} \|Ax\| = \sup_{x\neq 0} \frac{\|Ax\|}{\|x\|} \tag{4.14}$$

für $A \in \mathbb{R}^{n,n}$. Sie heißt die durch die *Vektornorm* $\|\cdot\|$ *indizierte Matrix-Norm.*

Sie hat wiederum die Eigenschaften

$\|A\| = 0$ nur für $A = 0$.

$\|\alpha A\| = |\alpha|\,\|A\|$ für alle $\alpha \in \mathbb{R}$ und alle $A \in \mathbb{R}^{n,n}$.

$\|A + B\| \leq \|A\| + \|B\|$ für alle $A, B \in \mathbb{R}^{n,n}$ (Dreiecksungleichung).

Zusätzlich sichert (4.14) die nützlichen Ungleichungen

$\|Ax\| \leq \|A\|\,\|x\|$ für alle $x \in \mathbb{R}^n$ und alle $A \in \mathbb{R}^{n,n}$ (Verträglichkeitsbedingung)

$\|AB\| \leq \|A\|\,\|B\|$ für alle $A, B \in \mathbb{R}^{n,n}$ (Submultiplikativität).

Beispiele:

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

$$\|x\|_1 = \sum_{i=1}^{n} |x_i| \ \text{ induziert } \ \|A\|_1 = \max_{j=1,\ldots,n} \sum_{i=1}^{n} |a_{ij}| \quad \text{(Spaltensummennorm)}$$

$$\|x\|_\infty = \max_{i=1,\ldots,n} |x_i| \ \text{ induziert } \ \|A\|_\infty = \max_{i=1,\ldots,n} \sum_{j=1}^{n} |a_{ij}| \quad \text{(Zeilensummennorm)}$$

Wir sind nun in der Lage, die bereits erwähnte Kondition einer Matrix einzuführen.

Definition 4.3.2 *Sei* $A \in \mathbb{R}^{n,n}$ *invertierbar und sei* $\|\cdot\|$ *eine induzierte Matrixnorm. Dann heißt die Zahl*

$$\mathit{cond}(A) = \|A\|\,\|A^{-1}\|$$

die Kondition *von* A *bezüglich der Matrixnorm.*

Man kann nun folgendes zeigen.

Satz 4.3.3 (Störeinfluss von Matrix und rechter Seite)
Sei $A \in \mathbb{R}^{n,n}$ invertierbar, $b, \Delta b \in \mathbb{R}^n$, $b \neq 0$ und $\Delta A \in \mathbb{R}^{n,n}$ mit $\|\Delta A\| < 1/\|A^{-1}\|$ mit einer beliebigen durch eine Norm $\|\cdot\|$ auf $\mathbb{R}^n$ induzierten Matrixnorm $\|\cdot\|$. Ist x die Lösung von

$$Ax = b$$

und $\tilde{x}$ die Lösung von

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

dann gilt

$$\frac{\|\tilde{x} - x\|}{\|x\|} \leq \frac{\mathit{cond}(A)}{1 - \mathit{cond}(A)\|\Delta A\|/\|A\|} \left(\frac{\|\Delta A\|}{\|A\|} + \frac{\|\Delta b\|}{\|b\|} \right).$$

Beweis: Wir betrachten der Einfachheit halber nur den Fall $\Delta A = 0$. Dann liefert Subtraktion der gestörten und ungestörten Gleichung

$$A(\tilde{x} - x) = \Delta b,$$

also

$$\|\tilde{x} - x\| = \|A^{-1}\Delta b\| \leq \|A^{-1}\| \|\Delta b\|.$$

Wegen $\|b\| = \|Ax\| \leq \|A\| \|x\|$ folgt $1/\|x\| \leq \|A\|/\|b\|$ und somit

$$\frac{\|\tilde{x} - x\|}{\|x\|} \leq \|A\| \|A^{-1}\| \frac{\|\Delta b\|}{\|b\|}.$$

□

Die Kondition bestimmt also die Sensitivität bezüglich Störungen von Matrix und rechter Seite.

4.3.2 Ergänzung: Rundungsfehlereinfluß beim Gauß-Verfahren

Auf einem Rechner wird eine Zahl $\neq 0$ nach IEEE-Standard dargestellt in der Form

$$\pm 1, \alpha_1\, \alpha_2\, \ldots\, \alpha_{t-1} \cdot 2^b, \quad \alpha_i \in \{0, 1\}, b \in \{b_-, \ldots, b_+\},$$

z.B. bei der heute üblichen doppelten Genauigkeit

$$t = 53 \text{ (ca. 15 Dezimalstellen)}, \; b_- = -1022, \; b_+ = 1023.$$

Alle elementaren Rechenoperationen sind nach IEEE-Standard so zu implementieren, dass das Ergebnis der Operation das gerundete exakte Ergebnis ist, ausser bei Exponenten-Über- oder Unterlauf. Bezeichnen wir mit $+_g$, $-_g$, usw. die Rechenoperationen auf einem Rechner, dann gilt also z.B.

$$x +_g y = \mathrm{rd}(x + y).$$

Hierbei rundet rd zur nächstgelegenen Gleitpunktzahl. Es gilt für den relativen Fehler

$$\frac{|x - \mathrm{rd}(x)|}{|x|} \le 2^{-t} =: \text{eps}, \quad \text{eps: Maschinengenauigkeit.}$$

Somit gilt bei jeder Rechenoperation $\circ_g \in \{+_g, -_g, *_g, /_g\}$

$$x \circ_g y = (x \circ y)(1 + \varepsilon), \quad |\varepsilon| \le \text{eps}.$$

Rundungsfehleranalyse für das Gauß-Verfahren

Durch eine elementare aber aufwendige Abschätzung der beim Gauß-Verfahren auftretenden Rundungsfehlerverstärkung erhält man folgendes Resultat.

Satz 4.3.4 *Sei $A \in \mathbb{R}^{n,n}$ invertierbar. Wendet man das Gauß-Verfahren auf einem Rechner mit Maschinengenauigkeit eps mit einer Pivot-Technik an, die $|l_{ij}| \le 1$ sicherstellt (z.B. Spaltenpivotsuche oder totale Pivotsuche), dann errechnet man $\bar{L}$, $\bar{R}$ mit*

$$\bar{L}\bar{R} = PAQ + F, \quad |f_{ij}| \le 2j\bar{a}\frac{eps}{1 - eps}.$$

Hierbei sind P, Q die aus der Pivotsuche resultierenden Permutationen und

$$\bar{a} = \max_k \bar{a}_k, \quad \bar{a}_k = \max_{i,j} |a_{ij}^{(k)}|. \tag{4.15}$$

Berechnet man mit Hilfe von $\bar{L}$, $\bar{R}$ durch Vorwärts- und Rückwärtssubstitution eine Näherungslösung $\bar{x}$ von $Ax = b$, dann existiert eine Matrix E mit

$$(A + E)\bar{x} = b, \quad |e_{ij}| \le \frac{2(n+1)eps}{1 - neps}(|\bar{L}||\bar{R}|)_{ij} \le \frac{2(n+1)eps}{1 - neps} n\bar{a}.$$

Hierbei bezeichnet $|\bar{L}| = (|\bar{l}_{ij}|)$, $|\bar{R}| = (|\bar{r}_{ij}|)$.

Beweis: Siehe Stoer [St94]. □

Bemerkung: Mit Satz 4.3.3 kann man nun auch den relativen Fehler der Näherungslösung $\bar{x}$ abschätzen. □

Einfluß der Pivot-Strategie

Die Größe von $\bar{a}$ in (4.15) hängt von der Pivotstrategie ab. Man kann folgendes zeigen:

- **Spaltenpivotsuche:** $\bar{a}_k \le 2^k \max_{i,j} |a_{ij}|$.

  Diese Schranke kann erreicht werden, ist aber in der Regel viel zu pessimistisch. In der Praxis tritt fast immer $\bar{a}_k \le 10 \max_{i,j} |a_{ij}|$ auf.

- **Spaltenpivotsuche bei Tridiagonalmatrizen:** $\bar{a}_k \le 2 \max_{i,j} |a_{ij}|$.
- **Vollständige Pivotsuche:** $\bar{a}_k \le f(k) \max_{i,j} |a_{ij}|, \quad f(k) = k^{1/2}(2^1 3^{1/2} \cdots k^{1/(k-1)})^{1/2}$.

  $f(n)$ wächst recht langsam. Es ist bislang kein Beispiel mit $\bar{a}_k \ge (k+1) \max_{i,j} |a_{ij}|$ entdeckt worden.