3. LINEARE GLEICHUNGSSYSTEME

Gesucht ist die Lösung $x \in \mathbb{R}^n$ eines linearen Gleichungssystems $Ax = b$ mit quadratischer und invertierbarer Matrix $A \in \mathbb{R}^{n,n}$ und $b \in \mathbb{R}^n$:

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

3.1 Gauß-Elimination

Es sei $A = (a_{ij})_{i,j=1}^n$ gegeben. Unter der Annahme, dass A invertierbar ist, gibt es in der ersten Spalte von A mindestens ein von Null verschiedenes Element. Nach eventuellem Vertauschen von Zeilen können wir daher $a_{11} \neq 0$ annehmen. Dann kann man die Variable x_1 eliminieren, indem zunächst

$$l_{i1} = \frac{a_{i1}}{a_{11}}, \qquad i = 2, \ldots, n$$

berechnet und dann die i-te Zeile durch

$$(\text{Zeile } i) - l_{i1}(\text{Zeile } 1)$$

ersetzt. Man erhält

$$\begin{aligned} a_{11}^{(1)}x_1 + a_{12}^{(1)}x_2 + \ldots + a_{1n}^{(1)}x_n &= b_1^{(1)} \\ a_{22}^{(1)}x_2 + \ldots + a_{2n}^{(1)}x_n &= b_2^{(1)} \\ &\vdots \\ a_{n2}^{(1)}x_2 + \ldots + a_{nn}^{(1)}x_n &= b_n^{(1)} \end{aligned}$$

mit

$$\begin{aligned} a_{1j}^{(1)} &= a_{1j} & b_1^{(1)} &= b_1 \\ a_{ij}^{(1)} &= a_{ij} - l_{i1}a_{1j} & b_i^{(1)} &= b_i - l_{i1}b_1, \qquad i = 2, \ldots, n \end{aligned}$$

Dieses Vorgehen wiederholt man mit der Untermatrix der Dimension $(n-1) \times (n-1)$ von $A^{(1)} = (a_{ij}^{(1)})$. Die Untermatrix von $A^{(1)}$ ist ebenfalls invertierbar, da das Ausgangssystem $Ax = b$ eine eindeutige Lösung hat. Wir vertauschen gegebenenfalls die Zeilen so, dass das *Pivotelement* $a_{22}^{(1)} \neq 0$ ist und setzen

$$l_{i2} = \frac{a_{i2}^{(1)}}{a_{22}^{(1)}}, \qquad i = 3, \ldots, n.$$

Nach $n-1$ derartigen Schritten erhalten wir eine Folge

$$(A,b) \longrightarrow (A^{(1)},b^{(1)}) \longrightarrow (A^{(2)},b^{(2)}) \longrightarrow \ldots \longrightarrow (A^{(n-1)},b^{(n-1)}) =: (R,c).$$

Nach Konstruktion ist R eine obere Dreiecksmatrix R:

$$\begin{aligned} r_{11}x_1 + r_{12}x_2 + \ldots + r_{1n}x_n &= c_1 \\ r_{22}x_2 + \ldots + r_{2n}x_n &= c_2 \\ \ddots \qquad\qquad \vdots & \\ r_{nn}x_n &= c_n \end{aligned}$$

mit $r_{ii} \neq 0$. Lineare Gleichungssysteme mit Dreiecksmatrizen sind leicht zu lösen:

$$x_n = \frac{c_n}{r_{nn}}, \qquad x_i = \frac{1}{r_{ii}}\big(c_i - \sum_{j=i+1}^{n} r_{ij}x_j\big), \qquad i = n-1,\ldots,1.$$

Satz 3.1. *(LR-Zerlegung)*
Sei $A \in \mathbb{R}^{n,n}$ invertierbar. Dann liefert die Gauß-Elimination die LR-Zerlegung

$$PA = LR,$$

wobei P eine Permutationsmatrix und

$$L = \begin{bmatrix} 1 & & \\ l_{21} & 1 & \\ \vdots & \ddots & \ddots \\ l_{n1} & \cdots & l_{n,n-1}1 \end{bmatrix}, \qquad R = \begin{bmatrix} r_{11} & r_{12} & \cdots & r_{1n} \\ & r_{22} & \cdots & r_{2n} \\ & & \ddots & \vdots \\ & & & r_{nn} \end{bmatrix}$$

mit $r_{ii} \neq 0$ für $i = 1,\ldots,n$.

Beweis. Wir nehmen an, dass die notwendingen Zeilenvertauschungen für die Gauß-Elimination bereits zu Beginn durchgeführt wurden, d. h. wir ersetzen A durch PA und wenden darauf die Gauß-Elimination an. Die einzelnen Eliminationsschritte kann man durch Multiplikation mit den unteren Dreiecksmatrizen

$$L_i = I_n - V_i, \qquad V_i = \begin{bmatrix} 0 & \cdots & 0 & l_i & 0 & \cdots & 0 \end{bmatrix}, \qquad l_i = \begin{bmatrix} 0_i \\ l_{i+1,i} \\ \vdots \\ l_{n,i} \end{bmatrix}.$$

wie folgt schreiben:

$$A^{(1)} = L_1PA, \qquad A^{(k)} = L_kA^{(k-1)}, \qquad k = 2,\ldots,n-1.$$

Daraus ergibt sich

$$R = A^{(n-1)} = L_{n-1}A^{(n-2)} = \ldots = L_{n-1}L_{n-2}\cdots L_1PA$$

oder

$$PA = L_1^{-1}L_2^{-1}\cdots L_{n-1}^{-1}R.$$

Es bleibt noch $L = L_1^{-1}L_2^{-1}\cdots L_{n-1}^{-1}$ zu zeigen. Die spezielle Form von V_i impliziert $V_iV_k = 0$ für $i \le k$. Des Weiteren gilt

$$(I + V_i)L_i = (I + V_i)(I - V_i) = I + V_i - V_i + V_i^2 = I,$$

also ist $L_i^{-1} = I + V_i$ und

$$\begin{aligned} L_1^{-1}L_2^{-1}\cdots L_{n-1}^{-1} &= (I + V_1)\cdots(I + V_{n-1}) \\ &= I + V_1 + V_2 + \ldots + V_{n-1} = L, \end{aligned}$$

denn alle gemischten Produkte von $V_i's$ verschwinden. □

Bemerkung. Ist σ die Anzahl der Zeilenvertauschungen, dann gilt $\det P = (-1)^\sigma$ und man erhält

$$\det A = (-1)^\sigma \det L \det R = (-1)^\sigma \det R = (-1)^\sigma \prod_{i=1}^{n} r_{ii}.$$

Lösung des Gleichungssystems

Sobald man die LR-Zerlegung von A kennt, löst man das lineare Gleichungssystem $Ax = b$ durch Lösen zweier Dreieckssysteme:

$$PAx = L\underbrace{Rx}_{y} = Pb$$

ist äquivalent zu

$$Ly = Pb, \qquad Rx = y.$$

Das Dreieckssystem mit der unteren Dreiecksmatrix L löst man durch Vorwärtselimination, das mit der oberen Dreiecksmatrix R wie oben beschrieben durch Rückwärtselimination.

Rechenaufwand

Die Transformation von A auf $A^{(1)}$ kostet $n-1$ Divisionen sowie $(n-1)^2$ Multiplikationen und Additionen, die von $A^{(k-1)}$ auf $A^{(k)}$ kostet $n-k$ Divisionen sowie $(n-k)^2$ Multiplikationen und Additionen. Der Aufwand für die Divisionen ist gegenüber Multiplikationen und Additionen zu vernachlässigen. Insgesamt kann die LR-Zerlegung mit

$$\sum_{j=1}^{n-1} j^2 \approx \int_0^n x^2 dx = \frac{n^3}{3}$$

Operationen (Additionen und Multiplikationen) berechnet werden. Die Berechnung des Vektors c aus $Lc = b$ kostet $(n-1) + \ldots + 1 \approx n^2/2$ Operationen und die gleiche Zahl von Operationen benötigt man zur Lösung von $Rx = c$. Die Hauptarbeit steckt also in der Berechnung der LR-Zerlegung.

3.2 Wahl der Pivotelemente – Gleitpunktarithmetik

Während der Gauß-Elimination muss man gegebenenfalls Zeilenvertauschungen vornehmen. Hierfür hat man in der Regel mehrere Möglichkeiten. Im ersten Schritt kann man jede Zeile nehmen, für die $a_{i1} \neq 0$ gilt. In der Praxis verwendet man häufig die **Spaltenpivotsuche**: Als Pivotelement wählt man im $(k+1)$-ten Schritt das betragsmäßig größtmögliche Element $a_{j,k+1}^{(k)}$:

$$|a_{j,k+1}^{(k)}| = \max_{k+1 \leq i \leq n} |a_{i,k+1}^{(k)}|.$$

Auf Zeilenvertauschungen kann für spezielle Matrizen verzichtet werden. Hierzu gehören strikt diagonaldominante Matrizen.

Definition 3.2. *Eine Matrix $A \in \mathbb{R}^{n,n}$ heißt* **diagonaldominant***, falls für A*

$$|a_{ii}| \geq \sum_{j \neq i} |a_{ij}| \qquad \text{für alle } i = 1, \ldots, n,$$

gilt, wobei die Ungleichung für mindestens ein i eine echte Ungleichung ist. A heißt **strikt diagonaldominant***, wenn für alle i eine echte Ungleichung erfüllt ist.*

A heißt **(strikt) spaltendiagonaldominant***, falls A^T (strikt) diagonaldominant ist.*

Satz 3.3. *Ist A strikt spaltendiagonaldominant, dann wählt die Spaltenpivotsuche in jedem Eliminationsschritt des Gauß-Algorithmus das Diagonalelement $a_{ii}^{(i-1)}$ als Pivotelement. Insbesondere existiert also eine LR-Zerlegung von A und A ist nicht singulär.*

Beweis. Betrachten wir zunächst die Auswahl des ersten Pivotelements. Da A strikt spaltendiagonaldominant ist, ist $|a_{i1}| < |a_{11}|$. Im ersten Eliminationsschritt wird also das $(1,1)$-Element als Pivotelement gewählt.

Der Beweis läuft nun induktiv durch, wenn wir zeigen können, dass nach einem Schritt wieder eine strikt spaltendiagonaldominante Matrix entsteht. Dazu schreiben wir den Eliminationsschritt in der Blockform

$$\left[\begin{array}{c|c} a_{11} & v^T \\ \hline u & B \end{array}\right] = \left[\begin{array}{c|c} 1 & 0 \\ \hline u/a_{11} & I_{n-1} \end{array}\right] \left[\begin{array}{c|c} a_{11} & v^T \\ \hline 0 & B^{(1)}, \end{array}\right], \qquad B^{(1)} = B - \frac{1}{a_{11}} uv^T$$

Im nächsten Schritt wird mit $B^{(1)}$ genauso verfahren. Es ist

$$b_{ij}^{(1)} = a_{ij} - \frac{a_{i1}a_{1j}}{a_{11}}, \qquad 2 \leq i, j \leq n.$$

$B^{(1)}$ ist also genau dann strikt spaltendiagonaldominant, wenn

$$\left| a_{ii} - \frac{a_{i1}a_{1i}}{a_{11}} \right| > \sum_{\substack{j=2 \\ j \neq i}}^{n} \left| a_{ji} - \frac{a_{j1}a_{1i}}{a_{11}} \right|, \qquad i = 2, \ldots, n. \tag{3.1}$$

Wegen der strikten Spaltendiagonaldominanz von A gilt für $i \geq 2$ wegen

$$\sum_{\substack{j=2 \\ j \neq i}}^{n} |a_{ji}| = \sum_{\substack{j=1 \\ j \neq i}}^{n} |a_{ji}| - |a_{1i}| < |a_{ii}| - |a_{1i}|$$

tatsächlich

$$\begin{aligned}\sum_{\substack{j=2\\ j\neq i}}^{n}\left|a_{ji}-\frac{a_{j1}a_{1i}}{a_{11}}\right| &\leq \sum_{\substack{j=2\\ j\neq i}}^{n}|a_{ji}|+\left|\frac{a_{1i}}{a_{11}}\right|\sum_{\substack{j=2\\ j\neq i}}^{n}|a_{j1}|\\ &< |a_{ii}|-|a_{1i}|+|a_{1i}|\frac{|a_{11}|-|a_{i1}|}{|a_{11}|}\\ &= |a_{ii}|-\frac{|a_{i1}a_{1i}|}{|a_{11}|}.\end{aligned}$$

(3.1) folgt damit aus der umgekehrten Dreiecksungleichung. □

Die Wahl der sogenannten **Pivotelemente** kann wesentlichen Einfluss auf die Genauigkeit der berechneten Lösung haben. Um dieses Phänomen zu verstehen, erläutern wir zunächst die Gleitpunktdarstellung reeller Zahlen im Rechner.

Im Rechner wird eine reelle Zahl $x \neq 0$ in der Form

$$x = \pm a \cdot 10^b$$

dargestellt. Hierbei ist a die **Mantisse**, $0.1 \leq a < 1$ und die ganze Zahl b der **Exponent**. Die Darstellung ist eindeutig. Im Folgenden nehmen wir an, dass für die Mantisse eine feste Anzahl ℓ von Ziffern zur Verfügung, machen aber keine Einschränkung für den Exponent. Ist $\overline{a}$ die auf ℓ Ziffern gerundete Mantisse, dann können wir statt x nur mit

$$fl(x) = \pm\overline{a} \cdot 10^b$$

rechnen. Für $\ell = 8$ wird $\pi = 3.141592653\ldots$ durch

$$fl(\pi) = 0.31415927 \cdot 10^1$$

dargestellt. Die **Maschinengenauigkeit** `eps` ist die kleinste positive Zahl für die

$$fl(1 + \texttt{eps}) > 1.$$

Im Dezimalsystem mit ℓ-stelliger Genauigkeit ist $\texttt{eps} = 5 \cdot 10^{-\ell}$, im Binärsystem (Basis 2) ist $\texttt{eps} = 2^{-\ell}$, denn

$$\begin{aligned}fl(0.\underbrace{10\ldots0}_{\ell}49\ldots\cdot 10^1) &= 1,\\ fl(0.\underbrace{10\ldots0}_{\ell}50\ldots\cdot 10^1) &= 0.\underbrace{10\ldots1}_{\ell}\cdot 10^1 > 1.\end{aligned}$$

Satz 3.4. *Für jedes $x \neq 0$ ist*

$$|fl(x) - x| \leq \texttt{eps}|x|,$$

der relative Fehler ist also durch die Maschinengenauigkeit beschränkt.

Beweis. Es sei $x = a \cdot 10^b$ und $fl(x) = \overline{a} \cdot 10^b$. Bei ℓ signifikanten Stellen ist $|a - \overline{a}| \leq 5 \cdot 10^{-l-1}$, also

$$\frac{|fl(x) - x|}{|x|} = \frac{|\overline{a} - a| \cdot 10^b}{|a| \cdot 10^b} \leq \frac{5 \cdot 10^{-l-1}}{10^{-1}} = \texttt{eps},$$

da $a \geq 10^{-1}$. □

Wir können daher

$$fl(x) = x(1+\epsilon) \qquad \text{mit} \qquad |\epsilon| \leq \texttt{eps}$$

schreiben.

Beispiel. (Forsythe). Wir betrachten das System

$$\begin{aligned} 10^{-4}x_1 + x_2 &= 1 \\ x_1 + x_2 &= 2 \end{aligned}$$

mit der exakten Lösung

$$x_1 = \frac{1}{0.9999} = 1.00010001\ldots, \qquad x_2 = \frac{0.9998}{0.9999} = 0.99989998\ldots.$$

Bei dreistelliger Gleitpunktrechnung sieht das System wie folgt aus:

$$\begin{aligned} 0.100 \cdot 10^{-3}x_1 + 0.100 \cdot 10^{1}x_2 &= 0.100 \cdot 10^{1} \\ 0.100 \cdot 10^{1}x_1 + 0.100 \cdot 10^{1}x_2 &= 0.200 \cdot 10^{1} \end{aligned}$$

(a) Die Wahl des Pivotelements $a_{11} = 0.100 \cdot 10^{-3}$ ergibt $l_{21} = a_{21}/a_{11} = 0.100 \cdot 10^{5}$ und daraus

$$\begin{aligned} a_{22}^{(1)} &= 0.100 \cdot 10^{1} - 0.100 \cdot 10^{5} = -0.100 \cdot 10^{5}, \\ b_{2}^{(1)} &= 0.200 \cdot 10^{1} - 0.100 \cdot 10^{5} = -0.100 \cdot 10^{5}. \end{aligned}$$

Folglich ist $x_2 = 1$ die richtige Lösung, aber für x_1 ergibt sich mit

$$x_1 = (b_1 - a_{12}x_2)/a_{11} = 0,$$

ein völlig falscher Wert.

(b) Wählt man als Pivotelement $a_{21} = 1$, dann ist $l_{21} = 0.100 \cdot 10^{-3}$ und

$$\begin{aligned} a_{22}^{(1)} &= 0.100 \cdot 10^{1} - 0.100 \cdot 10^{-4} = 0.100 \cdot 10^{1}, \\ b_{2}^{(1)} &= 0.100 \cdot 10^{1} - 0.200 \cdot 10^{1} \cdot 0.100 \cdot 10^{-3} = 0.100 \cdot 10^{1}. \end{aligned}$$

Auch hier ergibt sich $x_2 = 1$, aber dieses Mal ist

$$x_1 = (b_1 - a_{12}x_2)/a_{11} = 0.200 \cdot 10^{1} - 1 = 1$$

ebenfalls korrekt.

◇

Wir analysieren jetzt die bei allen numerischen Algorithmen auftretenden elementaren Operationen (Addition, Subtraktion, Multiplikation und Division).

Die Kondition eines Problems

Ein Problem sei durch die Auswertung einer Abbildung $F : \mathbb{R}^n \to \mathbb{R}$ beschrieben. Wie wirken sich Störungen von $x = (x_1, \ldots, x_n)$ auf das Resultat $F(x)$ aus?

Definition 3.5. *Die* **Kondition** κ *von* F *ist die kleinste Zahl, für die gilt*

$$\frac{|\widehat{x}_i - x_i|}{|x_i|} \leq \texttt{eps} \Longrightarrow \frac{|F(\widehat{x}) - F(x)|}{|F(x)|} \leq \kappa \cdot \texttt{eps}.$$

Das Problem heißt **gut konditioniert**, *falls* κ *nicht zu groß ist und anderenfalls* **schlecht konditioniert**.

Bemerkung. Für $F : \mathbb{R} \to \mathbb{R}$ stetig differenzierbar gilt für $|\widehat{x} - x|/|x| < \texttt{eps}$

$$\frac{|F(\widehat{x}) - F(x)|}{|F(x)|} = \frac{|F(\widehat{x}) - F(x)|}{|\widehat{x} - x|} \frac{|\widehat{x} - x|}{|x|} \frac{|x|}{|F(x)|} \lesssim \frac{|F'(x)x|}{|F(x)|}\texttt{eps}, \quad \widehat{x} \to x.$$

Die Konditionszahl κ kann in diesem Fall also leicht durch Differenzieren berechnet werden:

$$\kappa = \frac{|F'(x)x|}{|F(x)|}.$$

In der Literatur findet man neben der hier definierten relativen Konditionszahl auch die absolute Konditionszahl. Diese ist durch

$$|F(\widehat{x}) - F(x)| \leq \kappa_{\text{abs}} |\widehat{x} - x|$$

definiert. Durch Grenzübergang $\widehat{x} \to x$ sieht man, dass $\kappa_{\text{abs}} = |F'(x)|$.
Beispiel. (Multiplikation zweier reeller Zahlen)
Es sei $F(x_1, x_2) = x_1 x_2$. Für die gestörten Werte

$$\widehat{x}_1 = x_1(1 + \epsilon_1), \qquad \widehat{x}_2 = x_2(1 + \epsilon_2), \qquad |\epsilon_i| \leq \texttt{eps}$$

ist

$$\frac{\widehat{x}_1 \widehat{x}_2 - x_1 x_2}{x_1 x_2} = (1 + \epsilon_1)(1 + \epsilon_2) - 1 = \epsilon_1 + \epsilon_2 + \epsilon_1 \epsilon_2.$$

Da eps klein ist, kann man das Produkt $\epsilon_1 \epsilon_2$ vernachlässigen und erhält

$$\left| \frac{\widehat{x}_1 \widehat{x}_2 - x_1 x_2}{x_1 x_2} \right| \leq 2\texttt{eps}.$$

Die Konditionszahl ist hier $\kappa = 2$, das Problem also gut konditioniert. $\diamond$
Beispiel. (Subtraktion zweier reeller Zahlen)
Für $F(x_1, x_2) = x_1 - x_2$ erhält man durch eine analoge Rechnung

$$\left| \frac{(\widehat{x}_1 - \widehat{x}_2) - (x_1 - x_2)}{x_1 - x_2} \right| = \left| \frac{x_1 \epsilon_1 - x_2 \epsilon_2}{x_1 - x_2} \right| \leq \frac{|x_1| + |x_2|}{|x_1 - x_2|}\texttt{eps} = \kappa\texttt{eps}.$$

Für $\text{sign} x_1 = -\text{sign} x_2$ (Addition) ist $\kappa = 1$ und das Problem gut konditioniert. Im Gegensatz dazu wird die Konditionszahl für $x_1 \approx x_2$ sehr groß. Die Subtraktion zweier etwa gleich großer Zahlen ist sehr schlecht konditioniert. Man spricht hier auch von Auslöschung. $\diamond$

Ist im Gauß-Algorithmus wie in unserem Beispiel l_{21} sehr groß bzw. a_{11} sehr klein, so ergibt sich wegen

$$\left.\begin{array}{rcccl} a_{22}^{(1)} & = & a_{22} - l_{21}a_{12} & \approx & -l_{21}a_{12} \\ b_2^{(1)} & = & b_2 - l_{21}b_1 & \approx & -l_{21}b_1 \end{array}\right\} \qquad \Longrightarrow x_2 = \frac{b_2^{(1)}}{a_{22}^{(1)}} \approx \frac{b_1}{a_{12}}$$

Dieser Wert ist in der Regel noch korrekt. Jedoch ist die Berechnung von

$$x_1 = (b_1 - a_{12}x_2)/a_{11},$$

für $a_{12}x_2 \approx b_1$ schlecht konditioniert, weil Auslöschung von signifikanten Stellen auftreten kann. Man muss daher große Werte l_{ij} vermeiden.

Die Spaltenpivotsuche garantiert $|l_{ij}| \leq 1$ für alle i, j.

3.3 Die Kondition einer Matrix

Erinnerung: $\|\cdot\|\colon \mathbb{C}^n \to \mathbb{R}_0^+$ ist eine **Vektornorm**, wenn für alle $x, y \in \mathbb{C}^n$, $\alpha \in \mathbb{C}$

(a) $\|x\| \geq 0$ und $(\|x\| = 0 \Longleftrightarrow x = 0)$

(b) $\|\alpha x\| = |\alpha| \cdot \|x\|$

(c) $\|x + y\| \leq \|x\| + \|y\|$

Definition 3.6. $\|\cdot\| : \mathbb{C}^{n,m} \to \mathbb{R}_0^+$ *ist eine* **Matrixnorm**, *wenn* (a)–(c) *für alle* $x, y \in \mathbb{C}^{n,m}$, $\alpha \in \mathbb{C}$ *und außerdem für alle* $A \in \mathbb{C}^{n,k}$, $B \in \mathbb{C}^{k,m}$ *gilt:*

(d) $\|AB\| \leq \|A\| \cdot \|B\|$ *(Submultiplikativität)*

Beispiel. Für $A \in \mathbb{C}^{m,n}$, ist $\|A\|_F^2 = \sum_{i=1}^m \sum_{j=1}^n |a_{ij}|^2$ die Frobenius Matrixnorm. $\diamond$

Lemma 3.7. *Sei* $\|\cdot\|$ *eine Vektornorm. Dann ist für* $A \in \mathbb{C}^{m,n}$

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

eine Matrixnorm, die von $\|\cdot\|$ **induzierte Matrixnorm**.

Bemerkung. Beachten Sie, dass mit $\|x\|$ die Vektornorm in $\mathbb{C}^n$ und mit $\|Ax\|$ die Vektornorm in $\mathbb{C}^m$ gemeint ist. Die induzierte Matrixnorm $\|A\|$ ist also die kleinste nichtnegative Zahl, für die gilt

$$\|Ax\| \leq \|A\| \cdot \|x\| \qquad \forall x \in \mathbb{C}^n.$$

Beweis. (a) und (b) sind klar. Für alle $x \in \mathbb{C}^n$ gilt

$$\begin{aligned} \|(A + B)x\| &\leq \|Ax\| + \|Bx\| \\ &\leq \|A\|\,\|x\| + \|B\|\,\|x\| \\ &= (\|A\| + \|B\|)\|x\|, \end{aligned}$$

also erfüllt $\|\cdot\|$ auch (c). Schließlich folgt aus

$$\|ABx\| \leq \|A\| \cdot \|Bx\| \leq \|A\|\|B\|\|x\| \qquad \forall x$$

die letzte Bedingung (d). $\square$

Im Folgenden betrachten wir stets induzierte Matrixnormen. Für diese Normen gilt offensichtlich $\|I\| = 1$. Falls nicht anders erwähnt, bezeichnet $\|\cdot\| = \|\cdot\|_2$ immer die Euklid'sche Norm und die dadurch induzierte Matrixnorm.

Satz 3.8. *Sei $A \in \mathbb{C}^{m,n}$. Dann gilt*

(a) $\|A\|_1 = \max\limits_{j=1}^{n} \sum\limits_{i=1}^{m} |a_{ij}|$ *(maximale Spaltensummennorm)*

(b) $\|A\|_2 = \sqrt{\text{größter EW von } A^H A}$ *(Spektralnorm)*

(c) $\|A\|_\infty = \max\limits_{i=1}^{m} \sum\limits_{j=1}^{n} |a_{ij}|$ *(maximale Zeilensummennorm)*

Beweis. (a) Für $x \in \mathbb{C}^n$ beliebig folgt aus

$$\begin{aligned}\|Ax\|_1 &= \sum_{i=1}^{m} \left| \sum_{j=1}^{n} a_{ij} x_j \right| \le \sum_{i=1}^{m} \sum_{j=1}^{n} |a_{ij}| \cdot |x_j| \\ &= \sum_{j=1}^{n} \left(\sum_{i=1}^{m} |a_{ij}| \right) \cdot |x_j| \le \max_j \left(\sum_{i=1}^{m} |a_{ij}| \right) \|x\|_1\end{aligned}$$

zunächst

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

Um Gleichheit zu zeigen, suchen wir ein y mit

$$\|Ay\|_1 = \max_{j=1}^{n} \sum_{i=1}^{m} |a_{ij}| \cdot \|y\|_1.$$

Hierzu wählen wir j_0 so, dass

$$\max_j \sum_{i=1}^{m} |a_{ij}| = \sum_{i=1}^{m} |a_{i,j_0}|$$

und $y = e_{j_0}$. Dann gilt

$$\|Ay\|_1 = \|Ae_{j_0}\|_1 = \sum_{i=1}^{m} |a_{i,j_0}| = \max_j \sum_{i=1}^{m} |a_{ij}| = \|A\|_1 \|y\|_1.$$

(b) Die Matrix $A^H A$ ist Hermitesch und positiv semidefinit:

$$x^H A^H A x = \|Ax\|_2^2 \ge 0.$$

Daher exitiert eine unitäre Matrix U (d. h. $U^H U = U U^H = I$) mit

$$U^H A^H A U = \operatorname{diag}(\lambda_1, \ldots, \lambda_n), \qquad \lambda_i \ge 0 \text{ EW von } A^H A.$$

Mit $\lambda_{\max}$ bezeichnen wir den größten Eigenwert von $A^H A$. Dann ist mit $x = Uy$

$$\begin{aligned}\|Ax\|_2^2 &= x^H A^H A x = y^H U^H A^H A U y \\ &= \sum_{i=1}^{n} \lambda_i |y_i|^2 \le \lambda_{\max} \sum_{i=1}^{n} |y_i|^2 = \lambda_{\max} y^H y \\ &= \lambda_{\max} x^H U^H U x = \lambda_{\max} \|x\|_2^2,\end{aligned}$$

also $\|A\|_2 \le \sqrt{\lambda_{\max}}$. Gleichheit gilt, wenn man für x einen Eigenvektor von $A^H A$ zum größten Eigenwert $\lambda_{\max}$ einsetzt.

(c) beweist man wie (a). □

Um die Kondition eines linearen Gleichungssystems $Ax = b$ zu untersuchen, betrachten wir ein zweites, gestörtes System $\widehat{A}\widehat{x} = \widehat{b}$ mit

$$\begin{aligned}\widehat{a}_{ij} &= a_{ij}(1+\epsilon_{ij}), & |\epsilon_{ij}| &\le \epsilon_A, \\ \widehat{b}_i &= b_i(1+\epsilon_i), & |\epsilon_i| &\le \epsilon_b.\end{aligned} \tag{3.2}$$

Es gilt dann in den Normen $\|\cdot\|_1$ und $\|\cdot\|_\infty$

$$\|\widehat{A} - A\| \le \epsilon_A \|A\|, \qquad \|\widehat{b} - b\| \le \epsilon_b \|b\|. \tag{3.3}$$

Satz 3.9. *Es sei A invertierbar und es gelte $Ax = b$, $\widehat{A}\widehat{x} = \widehat{b}$. Ist (3.3) erfüllt, dann gilt*

$$\frac{\|\widehat{x} - x\|}{\|x\|} \le \frac{\kappa(A)}{1 - \epsilon_A \kappa(A)} (\epsilon_A + \epsilon_b),$$

falls $\epsilon_A \kappa(A) < 1$, wobei $\kappa(A) = \|A\| \cdot \|A^{-1}\|$ die **Konditionszahl von** *A bezeichnet.*

Beweis. Aus $\widehat{b} - b = \widehat{A}\widehat{x} - Ax = (\widehat{A} - A)\widehat{x} + A(\widehat{x} - x)$ folgt

$$\widehat{x} - x = A^{-1}\big((\widehat{b} - b) - (\widehat{A} - A)\widehat{x}\big).$$

Des Weiteren folgt aus

$$\|b\| = \|Ax\| \le \|A\|\|x\| \qquad \text{und} \qquad \|\widehat{x}\| \le \|x\| + \|\widehat{x} - x\|$$

mit Hilfe von (3.3)

$$\begin{aligned}\|\widehat{x} - x\| &\le \|A^{-1}\| \cdot \big[\epsilon_b \|b\| + \epsilon_A \|A\| \|\widehat{x}\|\big] \\ &\le \kappa(A)\Big(\epsilon_b \|x\| + \epsilon_A (\|x\| + \|\widehat{x} - x\|)\Big).\end{aligned}$$

Auflösen nach $\|\widehat{x} - x\|$ liefert die Behauptung. □

Bemerkung. Die Abschätzung in Satz 3.9 ist optimal in dem Sinn, dass es zu jedem System $Ax = b$ ein gestörtes System $\widehat{A}\widehat{x} = \widehat{b}$ gibt, so dass (3.3) erfüllt ist und in der Abschätzung Gleichheit gilt. Trotzdem ist die Schranke für Rundungsfehlerabschätzungen oft zu pessimistisch.

Lemma 3.10. *(Eigenschaften der Konditionszahl)*
Es sei eine invertierbare Matrix A gegeben. Dann gilt

(a) $\kappa(A) \geq 1$,

(b) $\kappa(\alpha A) = \kappa(A)$ *für alle* $\alpha \in \mathbb{C}$, $\alpha \neq 0$,

(c) $\kappa(A) = \max_{\|y\|=1} \|Ay\| / \min_{\|z\|=1} \|Az\|$.

Beweis. (a) Da wir nur von Vektornormen induzierte Matrixnormen betrachten, folgt die Abschätzung aus $1 = \|I\| = \|AA^{-1}\| \leq \|A\|\|A^{-1}\|$.

(b) ist klar und (c) folgt aus

$$\|A^{-1}\| = \max_{x \neq 0} \frac{\|A^{-1}x\|}{\|x\|} = \max_{y \neq 0} \frac{\|y\|}{\|Ay\|} = \left(\min_{y \neq 0} \frac{\|Ay\|}{\|y\|} \right)^{-1}$$

mit Hilfe der Substitution $y = A^{-1}x$. □

Die Darstellung (c) erlaubt es, $\kappa(A)$ auch für rechteckige Matrizen zu definieren.

Beispiel. (Matrizen mit großer Kondition) Die Hilbert-Matrix H_n und die Vandermonde-Matrix V_n definiert durch

$$H_n = \left(\frac{1}{i+j-1} \right)_{i,j=1}^{n}, \qquad V_n = \left(c_j^{i-1} \right)_{i,j=1}^{n}, \quad c_j = \frac{j}{n}$$

haben die in Tabelle 3.1 dargestellten großen Kondionszahlen. Gleichungssysteme mit einer Vandermonde-Matrix sind uns schon bei den Ordnungsbedingungen für Quadraturformeln begegnet. ⋄

n	2	4	6	8	10	12
$\kappa(H_n)$	27	$2.8 \cdot 10^4$	$2.9 \cdot 10^7$	$3.4 \cdot 10^{10}$	$3.5 \cdot 10^{13}$	$3.8 \cdot 10^{16}$
$\kappa(V_n)$	8	$5.6 \cdot 10^2$	$3.7 \cdot 10^4$	$2.4 \cdot 10^6$	$1.6 \cdot 10^8$	$1.0 \cdot 10^{10}$

Tabelle 3.1: Konditionszahlen von Hilbert– und Vandermonde-Matrizen

Beispiel. (Matrizen mit kleiner Kondition)

(a) Unitäre Matrizen haben Spektralnorm eins, denn unitäre Matrizen erhalten die Norm:

$$\|Ux\|^2 = x^H U^H U x = x^H x = \|x\|^2$$

und $U^{-1} = U^H$ ist ebenfalls unitär.

(b) Spline-Interpolation für äquidistante Knoten ($h_i = h$). Die dort auftretende Matrix hat die Form

$$A = h^{-1}\text{tridiag}\,(1, 4, 1).$$

Da $\kappa(A)$ unabhängig von der Skalierung ist, setzen wir ohne Einschränkung $h = 1$. Es ist $\|A\|_\infty = 6$. Um die Norm von A^{-1} abschätzen zu können schreiben wir

$$A = 4(I + N), \qquad N = \text{tridiag}\,(1/4, 0, 1/4).$$

Mit der Formel für die geometrische Reihe (Übung) folgt

$$A^{-1} = \frac{1}{4}(I+N)^{-1} = \frac{1}{4}\sum_{j=0}^{\infty}(-1)^j N^j.$$

Aus $\|N\|_\infty \le \frac{1}{2}$ und der Dreiecksungleichung ergibt sich daraus

$$\|A^{-1}\|_\infty \le \frac{1}{4}\sum_{j=0}^{\infty}\|N\|^j \le \frac{1}{4}\sum_{j=0}^{\infty}2^{-j} = \frac{1}{2}.$$

Somit ist $\kappa_\infty(A) \le 3$ unabhängig von der Dimension von A (Spline-Interpolation mit beliebig vielen Knoten ist gut konditioniert).

◇

Für eine beliebige Matrix $B = (b_{ij})$ bezeichnen wir mit $|B| = (|b_{ij}|)$ den elementweisen Betrag von B. Sind B und C zwei Matrizen gleicher Dimension, so schreiben wir $B \le C$, falls $b_{ij} \le c_{ij}$ für alle i, j gilt.

Satz 3.11. *Es sei A invertierbar und es gelte $Ax = b$, $\widehat{A}\widehat{x} = \widehat{b}$. Ist (3.2) erfüllt, dann gilt*

$$\frac{\|\widehat{x}-x\|_\infty}{\|x\|_\infty} \le \frac{\kappa_c(A)}{1-\epsilon_A\kappa_c(A)}\,(\epsilon_A+\epsilon_b),$$

falls $\epsilon_A\kappa_c(A) < 1$, wobei $\kappa_c(A) = \| \, |A^{-1}|\cdot|A| \, \|_\infty$.

Beweis. Genau wie im Beweis von Satz 3.9 gilt komponentenweise

$$\begin{aligned}
|\widehat{x}-x| &\le |A^{-1}|\cdot\big[\epsilon_b|b| + \epsilon_A|A|\cdot|\widehat{x}|\big] \\
&\le |A^{-1}|\cdot\big[\epsilon_b|A|\cdot|x| + \epsilon_A|A|\,\big(|x| + |\widehat{x}-x|\big)\big] \\
&= |A^{-1}|\cdot|A|\Big(\epsilon_b|x| + \epsilon_A(|x| + |\widehat{x}-x|)\Big).
\end{aligned}$$

Die Behauptung folgt nun, indem man nach $|\widehat{x}-x|$ auflöst und die Maximumsnorm dieser Ungleichung nimmt. □

Bemerkung. Satz 3.11 ist eine Verschärfung von Satz 3.9, denn es gilt stets

$$\kappa_c(A) \le \kappa_\infty(A)$$

(Übung). Für invertierbare Diagonalmatrizen D gilt $|DA| = |D|\cdot|A|$ und $|(DA)^{-1}| = |A^{-1}D^{-1}| = |A^{-1}|\cdot|D^{-1}|$. Daher ist $\kappa_c(A)$ invariant unter Linksmultiplikation mit Diagonalmatrizen, d. h.

$$\kappa_c(DA) = \kappa_c(A) \qquad \text{für alle} \qquad D = \operatorname{diag}(d_1,\dots,d_n) \quad \text{mit} \quad d_i \neq 0.$$

Für jede solche Diagonalmatrix D ist damit $\kappa_c(D) = 1$.

3.4 Die Stabilität des Gauß-Algorithmus

Ein Algorithmus zur Lösung eines Problems P zu gegebenen Daten x besteht aus einer Folge elementarer Operationen $f_1, \ldots, f_n$ (Additionen, Subtraktionen, Multiplikationen, Divisionen, Wurzelziehen, ...). Er kann durch eine Abbildung F beschrieben werden:

$$F(x) = f_n(f_{n-1}(\ldots f_2(f_1(x))\ldots)).$$

Im Allgemeinen gibt es viele verschiedene Algorithmen, um $F(x)$ zu berechnen. Diese können sich sehr unterschiedlich verhalten.

Definition 3.12. *Ein Algorithmus zur Berechnung von $y = F(x)$ heißt* **numerisch stabil im Sinne der Vorwärtsanalysis**, *falls für das berechnete Resultat $\widehat{y}$*

$$\frac{|\widehat{y} - y|}{|y|} \leq C\,\kappa(F)\,\texttt{eps}$$

mit einer nicht zu großen Konstante C (zum Beispiel $C = O(n)$) gilt.

Beispiel. Zur Berechnung von

$$F(x) = \frac{e^x - 1}{x} = \sum_{j=1}^{\infty} \frac{x^j}{(j+1)!}$$

betrachten wir die beiden folgenden Algorithmen:

Algorithmus 1	Algorithmus 2
if $x = 0$	$y = e^x$
$F = 1$	if $y = 1$
else	$F = 1$
$F = (e^x - 1)/x$	else
end	$F = (y-1)/\log y$
	end

Algorithmus 1 ist für $|x| \ll 1$ schlecht konditioniert, da Auslöschung bei der Subtraktion auftritt. Algorithmus 2 scheint nicht sinnvoll zu sein, da neben der Exponentialfunktion auch noch der Logarithmus berechnet werden muss.

In Tabelle 3.2 sind die Resultate der beiden Algorithmen für einige Werte x dargestellt. Man erkennt, dass Algorithmus 2 in allen signifikanten Stellen korrekte Werte liefert, mit Ausnahme für $x = 10^{-15}$, wo die letzte Stelle eine 1 sein sollte. Im Gegensatz dazu werden die von Algorithmus 1 berechneten Werte immer ungenauer, je kleiner x wird. $\diamond$

Wichtiger als die Vorwärtsstabilität ist die Rückwärtsstabilität eines Algorithmus, wie sie von Wilkinson eingeführt wurde.

Definition 3.13. *Ein Algorithmus zur Berechnung von $F(x)$ heißt* **numerisch stabil im Sinne der Rückwärtsanalysis**, *falls das numerische Resultat $\widehat{y}$ als exaktes Resultat der Berechnung von F mit leicht gestörten Daten $\widehat{x}$ interpretiert werden kann, d. h. $\widehat{y} = F(\widehat{x})$, wobei*

$$\frac{|\widehat{x} - x|}{|x|} \leq C\texttt{eps}$$

mit einer nicht zu großen Konstante C und der Maschinengenauigkeit `eps`.

x	Algorithmus 1	Algorithmus 2
10^{-5}	1.000005000006965	1.000005000016667
10^{-6}	1.000000499962184	1.000000500000167
10^{-7}	1.000000049433680	1.000000050000002
10^{-8}	$9.999999939225290 \cdot 10^{-1}$	1.000000005000000
10^{-9}	1.000000082740371	1.000000000500000
10^{-10}	1.000000082740371	1.000000000050000
10^{-11}	1.000000082740371	1.000000000005000
10^{-12}	1.000088900582341	1.000000000000500
10^{-13}	$9.992007221626409 \cdot 10^{-1}$	1.000000000000050
10^{-14}	$9.992007221626409 \cdot 10^{-1}$	1.000000000000005
10^{-15}	1.110223024625157	1.000000000000000
10^{-16}	0	1

Tabelle 3.2: Berechnete Werte für $(e^x - 1)/x$ mit Algorithmus 1 bzw. 2

Um das obige Beispiel zu analysieren nehmen wir an, dass exp und log mit einem relativen Fehler in der Größenordnung `eps` berechnet werden. Dann berechnet Algorithmus 2 zunächst $\widehat{y} = e^x(1+\delta)$ mit $|\delta| \le \texttt{eps}$. Ist $\widehat{y} = 1$, dann ist $e^x(1+\delta) = 1$, also

$$x = -\log(1+\delta) = \sum_{j=1}^{\infty} (-1)^j \frac{\delta^j}{j}, \qquad |\delta| \le \texttt{eps}.$$

Daraus folgt, dass der richtig gerundete Wert von $F(x) = 1 + x/2 + x^2/6 + \ldots$ eins ist und F bis auf Maschinengenauigkeit korrekt berechnet wurde. Für $\widehat{y} \neq 1$ tritt bei jeder elementaren Operation ein relativer Fehler auf:

$$\widehat{F} = \frac{(\widehat{y}-1)(1+\epsilon_1)}{\log \widehat{y}(1+\epsilon_2)}(1+\epsilon_3), \qquad |\epsilon_j| \le \texttt{eps}, \quad j = 1,2,3. \tag{3.4}$$

Setzen wir $v = \widehat{y} - 1$, so gilt

$$\begin{aligned} g(\widehat{y}) := \frac{\widehat{y}-1}{\log \widehat{y}} &= \frac{v}{\log(1+v)} \\ &= \frac{v}{v - v^2/2 + v^3/3 - \ldots} = \frac{1}{1 - v/2 + v^2/3 - \ldots} \\ &= 1 + \frac{v}{2} + O(v^2) \end{aligned}$$

Damit ist für kleine x, also $y \approx 1$

$$g(\widehat{y}) - g(y) \approx \frac{\widehat{y}-y}{2} \approx \frac{e^x \delta}{2} \approx \frac{\delta}{2} \approx \frac{g(y)\delta}{2}.$$

Mit (3.4) folgt daraus, dass Algorithmus 2 rückwärts stabil mit relativem Fehler höchstens 3.5`eps` ist.

Wir untersuchen jetzt die Rundungsfehler bei der LR-Zerlegung einer Matrix A. Ohne Einschränkung nehmen wir an, dass Zeilenvertauschungen bereits zu Beginn durchgeführt wurden.

Satz 3.14. *Es sei $A \in \mathbb{R}^{n,n}$ eine Matrix von Gleitpunktzahlen. Falls bei der LR-Zerlegung von A keine 0-Pivots auftreten, gilt für die in Gleitpunktrechnung durch Gauß-Elimination erhaltenen Matrizen $\widehat{L}$, $\widehat{R}$*

$$|A - \widehat{L}\widehat{R}| \le (n+3)\texttt{eps}|\widehat{L}| \cdot |\widehat{R}| + O(\texttt{eps}^2).$$

Beweis. Der Beweis erfolgt mit Induktion nach n. Für $n = 1$ ist die Behauptung klar. Nehmen wir also an, dass die Ungleichung für alle $(n-1)\times(n-1)$ Gleitpunktmatrizen gilt. Wir zerlegen A gemäß

$$A = \begin{bmatrix} \alpha & w^T \\ v & B \end{bmatrix}, \qquad v, w \in \mathbb{R}^{n-1}, \quad B \in \mathbb{R}^{n-1,n-1}.$$

Die Gauß-Elimination berechnet

$$z = v/\alpha, \qquad A_1 = B - zw^T$$

bzw. in Gleitpunktrechnung $\widehat{z}$ und $\widehat{A}_1$. Aus

$$\widehat{z}_i = \frac{v_i}{\alpha}(1+\epsilon_i) = z_i(1+\epsilon_i), \qquad |\epsilon_i| \le \texttt{eps},$$

folgt $|\widehat{z} - z| \le |z|\texttt{eps}$. Für $\widehat{A}_1$ gilt

$$\big(\widehat{A}_1\big)_{ij} = \big(b_{ij} - \widehat{z}_i w_j(1+\epsilon_{ij})\big)(1+\epsilon'_{ij}), \qquad |\epsilon_{ij}|, |\epsilon'_{ij}| \le \texttt{eps}.$$

Wegen

$$|(1+\epsilon_i)(1+\epsilon_{ij})(1+\epsilon'_{ij}) - 1| \le 3\texttt{eps} + O(\texttt{eps}^2)$$

ergibt sich hieraus

$$|(\widehat{A}_1)_{ij} - (A_1)_{ij}| \le \texttt{eps}\big(|b_{ij}| + 3|z_i| \cdot |w_j|\big) + O(\texttt{eps}^2)$$

oder in kompakter Schreibweise

$$\begin{aligned} |\widehat{A}_1 - A_1| &\le \texttt{eps}\big(|B| + 3|z| \cdot |w|^T\big) + O(\texttt{eps}^2) \\ &\le \texttt{eps}\big(|A_1| + 4|z| \cdot |w|^T\big) + O(\texttt{eps}^2), \end{aligned}$$

da $B = A_1 + zw^T$. Der Gauß-Algorithmus berechnet im nächsten Schritt die LR-Zerlegung von $\widehat{A}_1$. Nach Induktionsvoraussetzung gilt hierfür

$$|\widehat{A}_1 - \widehat{L}_1\widehat{R}_1| \le (n+2)\texttt{eps}|\widehat{L}_1| \cdot |\widehat{R}_1| + O(\texttt{eps}^2).$$

Aus diesen beiden Abschätzungen folgt dann auch

$$|A_1| = |\widehat{L}_1\widehat{R}_1| + O(\texttt{eps}) \le |\widehat{L}_1| \cdot |\widehat{R}_1| + O(\texttt{eps}).$$

Wir haben damit die folgenden Zerlegungen:

$$\begin{aligned} \widehat{L}\widehat{R} &= \begin{bmatrix} 1 & 0 \\ \widehat{z} & \widehat{L}_1 \end{bmatrix} \begin{bmatrix} \alpha & w^T \\ 0 & \widehat{R}_1 \end{bmatrix} = \begin{bmatrix} \alpha & w^T \\ \alpha\widehat{z} & \widehat{z}w^T + \widehat{L}_1\widehat{R}_1 \end{bmatrix}, \\ A = LR &= \begin{bmatrix} 1 & 0 \\ z & L_1 \end{bmatrix} \begin{bmatrix} \alpha & w^T \\ 0 & R_1 \end{bmatrix} = \begin{bmatrix} \alpha & w^T \\ \alpha z & zw^T + L_1R_1 \end{bmatrix}. \end{aligned}$$

Subtraktion liefert

$$A - \widehat{L}\widehat{R} = \begin{bmatrix} 0 & 0 \\ \alpha(z-\widehat{z}) & (z-\widehat{z})w^T + (A_1 - \widehat{A}_1) + (\widehat{A}_1 - \widehat{L}_1\widehat{R}_1) \end{bmatrix}.$$

Mit obigen Abschätzungen ist wegen $5 \le n+3$

$$\begin{aligned} |A - \widehat{L}\widehat{R}| &\le \texttt{eps} \begin{bmatrix} 0 & 0 \\ |\alpha|\,|z| & |z|\cdot|w|^T + |A_1| + 4|z|\,|w|^T + (n+2)|\widehat{L}_1|\,|\widehat{R}_1| \end{bmatrix} + O(\texttt{eps}^2) \\ &\le (n+3)\texttt{eps} \begin{bmatrix} |\alpha| & |w|^T \\ |\alpha|\,|\widehat{z}| & |\widehat{z}|\,|w|^T + |\widehat{L}_1|\,|\widehat{R}_1| \end{bmatrix} + O(\texttt{eps}^2) \\ &= (n+3)\texttt{eps}|\widehat{L}|\,|\widehat{R}| + O(\texttt{eps}^2). \end{aligned}$$

Damit ist die Behauptung für alle n gezeigt. □

Die Gauß-Elimination ist also stabil im Sinne der Rückwärtsanalysis, falls $|\widehat{L}|$ und $|\widehat{R}|$ nicht zu groß sind. Bei der Spaltenpivotsuche gilt $|l_{ij}| \le 1$ für alle i, j und daraus folgt

$$\max_{ij}|r_{ij}| \le 2^{n-1} \max_{ij}|a_{ij}|$$

(Übung). Die Abschätzung ist meist zu pessimistisch, aber sie ist scharf. Ebenfalls in den Übungen lernen wir ein Beispiel kennen, bei dem der Faktor 2^{n-1} tatsächlich auftritt. Bei zufällig gewählten Matrizen beobachtet man nur einen Verstärkungsfaktor der Größenordnung n.

Satz 3.15. *Seien L, R untere bzw. obere Dreiecksmatrizen und b, c Vektoren von Gleitpunktzahlen. Die in Gleitpunktrechnung erhaltenen Ergebnisse $\widehat{x}$, $\widehat{y}$ für die Gleichungssysteme*

$$Ly = b, \qquad Rx = c$$

sind die exakten Lösungen von

$$\widehat{L}\widehat{y} = b, \qquad \widehat{R}\widehat{x} = c$$

mit $|L - \widehat{L}| \le n|L|\texttt{eps}$ und $|R - \widehat{R}| \le n|R|\texttt{eps}$.

Beweis. Übung. □

Satz 3.16. *Seien $\widehat{L}$ und $\widehat{R}$ wie in Satz 3.14 erhalten. Das in Gleitpunktrechnung erhaltene Ergebnis $\widetilde{x}$ von*

$$\widehat{L}\widehat{y} = b, \qquad \widehat{R}\widehat{x} = \widehat{y}$$

erfüllt

$$\widetilde{A}\widetilde{x} = b \quad \textit{mit} \quad |A - \widetilde{A}| \le 3(n+1)\texttt{eps}|\widehat{L}|\cdot|\widehat{R}| + O(\texttt{eps}^2).$$

Beweis. Ohne Rundungsfehler wäre $A = LR$, $Ly = b$, $Rx = y$ bzw. $Ax = b$. Statt L und R haben wir jetzt die gestörten Dreiecksmatrizen $\widehat{L}$ und $\widehat{R}$. Nach Satz 3.15 ist $\widetilde{x}$ die Lösung von

$$\widetilde{L}\widehat{y} = b, \qquad \widetilde{R}\widetilde{x} = \widehat{y}$$

mit $|\widetilde{L}-\widehat{L}| \le n|\widehat{L}|\mathtt{eps}$ und $|\widetilde{R}-\widehat{R}| \le n|\widehat{R}|\mathtt{eps}$. Setzen wir $\widetilde{A} = \widetilde{L}\widetilde{R}$, so folgt $\widetilde{A}\widetilde{x} = b$. Für den Fehler gilt

$$\begin{aligned}
|\widetilde{A}-A| &= |\widetilde{L}\widetilde{R} - \widetilde{L}\widehat{R} + \widetilde{L}\widehat{R} - \widehat{L}\widehat{R} + \widehat{L}\widehat{R} - A| \\
&\le \underbrace{|\widetilde{L}|}_{|\widehat{L}|+O(\mathtt{eps})} \cdot|\widetilde{R}-\widehat{R}| + |\widetilde{L}-\widehat{L}|\cdot|\widehat{R}| + |\widehat{L}\widehat{R}-A| \\
&\le n\mathtt{eps}|\widehat{L}||\widehat{R}| + n\mathtt{eps}|\widehat{L}|\cdot|\widehat{R}| + (n+3)\mathtt{eps}|\widehat{L}|\cdot|\widehat{R}| + O(\mathtt{eps}^2) \\
&= 3(n+1)\mathtt{eps}|\widehat{L}|\cdot|\widehat{R}| + O(\mathtt{eps}^2),
\end{aligned}$$

wobei die letzte Ungleichung aus Satz 3.14 folgt. □

3.5 Cholesky-Zerlegung

Wir betrachten zunächst die folgende Blockversion der LR-Zerlegung. Dazu sei $A \in \mathbb{R}^{n,n}$ in der Form

$$A = \begin{bmatrix} A_{11} & A_{12} \\ A_{21} & A_{22} \end{bmatrix} \qquad \text{mit nicht singulärem } A_{11} \in \mathbb{R}^{m,m}$$

und $A_{22} \in \mathbb{R}^{n-m,n-m}$ partitioniert. Demzufolge ist $A_{12} \in \mathbb{R}^{m,n-m}$ und $A_{21} \in \mathbb{R}^{n-m,m}$. Bei der Block-LR-Zerlegung eliminiert man analog zur LR-Zerlegung durch Multiplikation der oberen Blockzeile mit $-A_{21}A_{11}^{-1}$ und Addition zur unteren Blockzeile die Matrix A_{21}. Dies entspricht wie dort einer Multiplikation von links mit

$$\begin{bmatrix} I & 0 \\ -A_{21}A_{11}^{-1} & I \end{bmatrix}.$$

Wenden wir diese direkt auf ein lineares Gleichungssystem an, so ist

$$A\begin{bmatrix} x \\ y \end{bmatrix} = \begin{bmatrix} b \\ c \end{bmatrix}$$

äquivalent zu

$$\begin{bmatrix} A_{11} & A_{12} \\ 0 & A_{22} - A_{21}A_{11}^{-1}A_{12} \end{bmatrix}\begin{bmatrix} x \\ y \end{bmatrix} = \begin{bmatrix} b \\ c - A_{21}A_{11}^{-1}b \end{bmatrix}.$$

Definition 3.17. *Die Matrix*

$$S = A_{22} - A_{21}A_{11}^{-1}A_{12} \in \mathbb{R}^{n-m,n-m}$$

heißt **Schur-Komplement** *von A bezüglich A_{11}.*

Die Lösung des linearen Gleichungssystems kann mit Hilfe des Schur-Komplements durch

$$\begin{aligned}
y &= S^{-1}\big(c - A_{21}A_{11}^{-1}b\big) \\
x &= A_{11}^{-1}\big(b - A_{12}y\big)
\end{aligned} \tag{3.5}$$

berechnet werden. Ferner hat A wegen

$$\begin{bmatrix} I & 0 \\ -A_{21}A_{11}^{-1} & I \end{bmatrix}^{-1} = \begin{bmatrix} I & 0 \\ A_{21}A_{11}^{-1} & I \end{bmatrix}$$

die Blockzerlegung

$$A = \begin{bmatrix} I & 0 \\ A_{21}A_{11}^{-1} & I \end{bmatrix} \begin{bmatrix} A_{11} & A_{12} \\ 0 & S \end{bmatrix}.$$

Wir betrachten im Folgenden symmetrische und positiv definite Matrizen $A \in \mathbb{R}^{n,n}$. Zur Erinnerung: Eine Matrix $A \in \mathbb{R}^{n,n}$ heißt symmetrisch, $A = A^T$, genau dann, wenn $a_{ij} = a_{ji}$ für alle $i, j = 1, \ldots, n$ gilt. Eine reelle und symmetrische Matrix A heißt positiv definit genau dann, wenn $x^T Ax > 0$ für alle $x \in \mathbb{R}^n$, $x \neq 0$ gilt.

Lemma 3.18. *$A \in \mathbb{R}^{n,n}$ sei symmetrisch und positiv definit. Dann ist das Schur-Komplement S von A bzgl. A_{11} für jedes $m < n$ wohldefiniert und $S = S^T$ ist ebenfalls positiv definit.*

Beweis. Aus $A = A^T$ positiv definit folgt sofort $A_{11} = A_{11}^T$ und $A_{22} = A_{22}^T$ positiv definit, insbesondere nicht singulär. Damit ist S wohldefiniert. Ferner ist S symmetrisch:

$$S^T = A_{22}^T - A_{12}^T A_{11}^{-T} A_{21}^T = A_{22} - A_{21}A_{11}^{-1}A_{12} = S.$$

Um zu zeigen, dass S auch positiv definit ist, wählen wir einen beliebigen Vektor $y \in \mathbb{R}^{n-m}$, $y \neq 0$ aus. Für diesen gilt

$$y^T Sy = \begin{bmatrix} x \\ y \end{bmatrix}^T \begin{bmatrix} 0 \\ Sy \end{bmatrix} \qquad \forall x \in \mathbb{R}^m.$$

Hier wählen wir jetzt x so, dass

$$A \begin{bmatrix} x \\ y \end{bmatrix} = \begin{bmatrix} 0 \\ Sy \end{bmatrix}$$

gilt. Nach (3.5) mit $b = 0$ und $c = Sy$ ist $x = -A_{11}^{-1}A_{12}y$. Da A positiv definit ist, folgt daraus $y^T Sy > 0$ und damit die Behauptung. □

Definition 3.19. *Eine Zerlegung von $A = A^T \in \mathbb{R}^{n,n}$ der Form $A = LL^T$ mit unterer Dreiecksmatrix L mit positiven Diagonalelementen heißt* **Cholesky-Zerlegung** *von A, L heißt* **Cholesky-Faktor**.

Satz 3.20. *A ist genau dann symmetrisch und positiv definit, wenn eine Cholesky-Zerlegung von A existiert.*

Beweis. Es sei $A = LL^T$ eine Cholesky-Zerlegung. Dann ist

$$A^T = (L^T)^T L^T = LL^T = A,$$

A ist also symmetrisch. Für beliebiges $x \in \mathbb{R}^n$, $x \neq 0$ gilt

$$x^T Ax = x^T LL^T x = (L^T x)^T (L^T x) = \|L^T x\|_2^2 > 0,$$

also ist A auch positiv definit.

Umgekehrt sei $A = A^T$ positiv definit. Wir beweisen die Existenz der Cholesky-Zerlegung mit Induktion nach n. Für $n = 1$ ist $a_{11} > 0$, also ist $L = (\sqrt{a_{11}})$ der Cholesky-Faktor.

Wir nehmen an, dass die Behauptung für $n - 1$ bereits bewiesen ist und betrachten die Partitionierung

$$A = \left[\begin{array}{c|c} a_{11} & A_{12} \\ \hline A_{21} & A_{22} \end{array}\right], \qquad A_{21} = A_{12}^T.$$

Man sieht sofort, dass $l_{11} = \sqrt{a_{11}} > 0$. Das Schur-Komplement $S \in \mathbb{R}^{n-1,n-1}$ von A bzgl. a_{11} ist nach Lemma 3.18 symmetrisch und positiv definit. Nach Induktionsannahme gibt es daher eine untere Dreiecksmatrix L_1 mit positiven Diagonalelementen, so dass $S = L_1 L_1^T$. Wir zeigen jetzt, dass

$$L = \left[\begin{array}{c|c} l_{11} & 0 \\ \hline \frac{1}{l_{11}} A_{21} & L_1 \end{array}\right]$$

ein Cholesky-Faktor von A ist, d. h. $A = LL^T$ gilt. Man rechnet leicht

$$LL^T = \left[\begin{array}{c|c} l_{11}^2 & A_{21}^T \\ \hline A_{21} & B \end{array}\right], \qquad B = \frac{1}{l_{11}^2} A_{21} A_{21}^T + L_1 L_1^T$$

nach. Für B gilt wegen $S = A_{22} - A_{21} A_{11}^{-1} A_{12} = L_1 L_1^T$

$$B = \frac{1}{a_{11}} A_{21} A_{12} + A_{22} - A_{21} \frac{1}{a_{11}} A_{12} = A_{22}.$$

Somit ist $LL^T = A$. Da L_1 positive Diagonalelemente hat, gilt dasselbe auch für L. □

Zur Berechnung der Diagonalelemente vergleicht man in

$$LL^T = \begin{bmatrix} l_{11} & & & \\ l_{21} & l_{22} & & \\ \vdots & & \ddots & \\ l_{n1} & \cdots & \cdots & l_{nn} \end{bmatrix} \begin{bmatrix} l_{11} & l_{21} & \cdots & l_{n1} \\ & l_{22} & & \vdots \\ & & \ddots & \vdots \\ & & & l_{nn} \end{bmatrix} = \begin{bmatrix} a_{11} & a_{12} & \cdots & a_{1n} \\ a_{21} & a_{22} & \cdots & a_{2n} \\ \vdots & \vdots & & \vdots \\ a_{n1} & a_{n2} & \cdots & a_{nn} \end{bmatrix}$$

die Einträge spaltenweise von links nach rechts und in den Spalten von oben nach unten:

$$\begin{aligned} &(j,j): && a_{jj} = l_{j1}^2 + l_{j2}^2 + \ldots + l_{jj}^2 && \longrightarrow l_{jj} \\ &(i,j), i > j: && a_{ij} = l_{i1} l_{j1} + l_{i2} l_{j2} + \ldots + l_{i,j} l_{j,j} && \longrightarrow l_{ij}. \end{aligned}$$

Es ergibt sich Algorithmus 3.1. Man kann durch Überschreiben von A mit dem Cholesky-Faktor L ohne zusätzlichen Speicherbedarf auskommen.

Algorithmus 3.1 Cholesky-Verfahren

for $j = 1, \ldots, n$ **do**

$l_{jj} = \left(a_{jj} - \sum_{k=1}^{j-1} l_{jk}^2 \right)^{1/2}$ — $j-1$ Mult. u. Add., 1 Wurzel

for $i = j+1, \ldots, n$ **do**

$l_{ij} = \left(a_{ij} - \sum_{k=1}^{j-1} l_{ik} l_{jk} \right) / l_{jj}$ — $j-1$ Mult. u. Add., 1 Division

end for

end for

Bemerkung. Der Aufwand für das Cholesky-Verfahren ist

$$\sum_{j=1}^{n} \big[j + (n-j)j \big] \approx \int_0^n \big[x + x(n-x) \big] dx = \frac{1}{6} n^3 + O(n^2)$$

Additionen und Multiplikationen. Zum Vergleich: die LR-Zerlegung einer Matrix benötigt mit $\frac{1}{3} n^3 + O(n^2)$ etwa doppelt soviele Operationen.

Stabilität der Cholesky-Zerlegung

In Gleitpunktarithmetik berechnet man wegen Rundungsfehlern nur einen gestörten Cholesky-Faktor $\widehat{L}$ statt L. Analog zu Satz 3.14 kann man zeigen, dass

$$|A - \widehat{L}\widehat{L}^T| \leq (n+3)\texttt{eps}|\widehat{L}| \cdot |\widehat{L}^T| + O(\texttt{eps}^2),$$

gilt. Die Stabilität hängt also davon ab, ob $|\widehat{L}|$ groß werden kann. Mit Hilfe der Cauchy-Schwarz'schen Ungleichung ($|x^T y| \leq \|x\| \cdot \|y\|$) zeigt man die Abschätzung

$$\begin{aligned}\left(|L| \cdot |L^T|\right)_{ij} &= \sum_{k=1}^{n} |l_{ik}| \cdot |l_{jk}| \\ &\leq \left(\sum_{k=1}^{n} l_{ik}^2\right)^{1/2} \left(\sum_{k=1}^{n} l_{jk}^2\right)^{1/2} \\ &= \sqrt{a_{ii}}\sqrt{a_{jj}}.\end{aligned}$$

Lemma 3.21. *Ist $A \in \mathbb{R}^{n,n}$ symmetrisch und positiv definit, dann gilt*

$$a_{ii}a_{jj} > a_{ij}^2 \qquad \forall i, j = 1, \ldots, n, \quad i \neq j$$

Beweis. Wir betrachten die Hauptuntermatrix $\widetilde{A}$ von A, die aus den Zeilen und Spalten i und j gebildet wird:

$$\widetilde{A} = \begin{bmatrix} a_{ii} & a_{ij} \\ a_{ji} & a_{jj} \end{bmatrix}.$$

Da $\widetilde{A}$ ebenfalls symmetrisch und positiv definit ist, gilt

$$a_{ij} = a_{ji}, \qquad a_{ii} > 0, \qquad a_{jj} > 0$$

und $\det \widetilde{A} = a_{ii}a_{jj} - a_{ij}^2 > 0$. □

Damit folgt wegen $\widehat{L}\widehat{L}^T = A + O(\texttt{eps})$, dass

$$|\widehat{l}_{ij}| \leq \sqrt{a}(1 + O(\texttt{eps})), \qquad a = \max_i a_{ii} = \max_{ij} |a_{ij}|.$$

Die Cholesky-Zerlegung ist damit rückwärts stabil.

Zur Lösung eines linearen Gleichungssystems $Ax = b$ mit der Cholesky-Zerlegung $A = LL^T$ löst man

$$Ly = b, \qquad L^T x = y$$

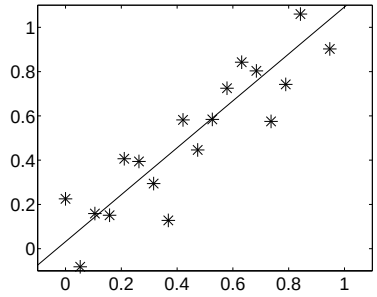
jeweils mit $n^2/2$ Operationen.

3.6 Lineare Ausgleichsrechnung

Wir betrachten jetzt das Problem, zu gegebenen Daten $y_1, \ldots, y_m$, die zu Zeiten $t_1, \ldots, t_m$ gemessen wurden, und zu gegebenen n Ansatzfunktionen $\varphi_i(t)$, $i = 1, \ldots, n$ (zum Beispiel Polynome $\varphi_i(t) = t^{i-1}$) Koeffizienten $x_i \in \mathbb{R}$ so zu bestimmen, dass

$$y_j \approx f(t_j), \qquad j = 1, \ldots, m, \qquad f(t) = \sum_{i=1}^{n} x_i \varphi_i(t).$$

Häufig hat man sehr viele Daten, möchte aber mit sehr wenigen Ansatzfunktionen auskommen, also $n \ll m$.



Bei der linearen Ausgleichsrechnung erfolgt die Approximation im Sinne der kleinsten Fehlerquadrate (Gauß, 1801). Für das *Residuum*

$$r = (r_j)_{j=1}^m, \qquad r_j = y_j - f(t_j)$$

werden die Koeffizienten x_i so gesucht, dass in der Eukliknorm $\|\cdot\| = \|\cdot\|_2$

$$\|r\|^2 = \sum_{j=1}^m r_j^2 = \min!.$$

Definieren wir

$$A = \begin{bmatrix} \varphi_1(t_1) & \cdots & \varphi_n(t_1) \\ \vdots & & \vdots \\ \varphi_1(t_m) & \cdots & \varphi_n(t_m) \end{bmatrix} \in \mathbb{R}^{m,n}, \qquad b = \begin{bmatrix} y_1 \\ \vdots \\ y_m \end{bmatrix} \in \mathbb{R}^m,$$

so ist $x \in \mathbb{R}^n$ gesucht, für das

$$\|r\| = \min!, \qquad r = b - Ax.$$

Satz 3.22. *Es sei $A \in \mathbb{R}^{m,n}$, $b \in \mathbb{R}^m$, $m \geq n$. Dann gilt*

$$\|b - Ax\| = \min_{v \in \mathbb{R}^n} \|b - Av\|$$

genau dann, wenn x die **Gauß'schen Normalengleichungen**

$$A^T A x = A^T b$$

löst.

Beweis. Es sei $\phi(v) := \|b - Av\|^2$ und x sei die Minimalstelle von ϕ. $y \in \mathbb{R}^n$ und $\epsilon \neq 0$ seien beliebig gewählt. Dann gilt

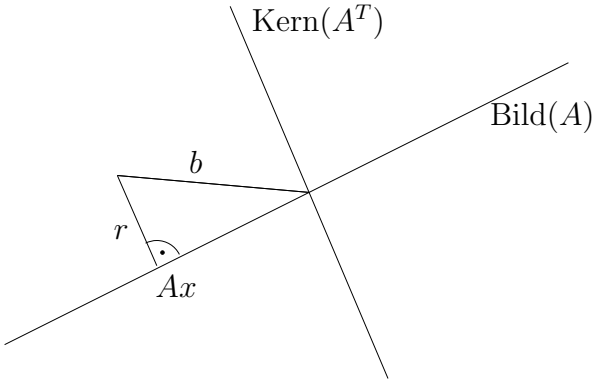
$$\begin{aligned} \phi(x + \epsilon y) &= (b - A(x + \epsilon y))^T (b - A(x + \epsilon y)) \\ &= \phi(x) - 2\epsilon (Ay)^T (b - Ax) + \epsilon^2 \|Ay\|^2. \end{aligned}$$

Betrachten wir jetzt den Grenzübergang $\epsilon \to 0$ mit geeignetem Vorzeichen von ϵ, so folgt,

$$\begin{aligned} \phi(x + \epsilon y) \geq \phi(x) \quad \forall y \in \mathbb{R}^n &\iff (Ay)^T (b - Ax) = 0 \quad \forall y \in \mathbb{R}^n \\ &\iff A^T (b - Ax) = 0. \end{aligned}$$

Die Umkehrung gilt offensichtlich ebenfalls. □

Geometrisch kann die Aussage so interpretiert werden, dass $\|b - Ax\| = \min!$ genau dann gilt, wenn der *Residuenvektor* $r = b - Ax$ senkrecht auf Bild(A) steht:



Bemerkung. Die Matrix A^TA ist symmetrisch und positiv semidefinit, denn

$$x^TA^TAx = \|Ax\|_2^2 \geq 0. \tag{3.6}$$

Lemma 3.23. *Es sei $A \in \mathbb{R}^{m,n}$ mit $m \geq n$. Dann sind folgende Aussagen äquivalent:*

(a) *A^TA ist positiv definit;*

(b) Kern$(A) = \{0\}$*;*

(c) Rang $(A) = n$.

Beweis. $x^TA^TAx = 0$ ist nach (3.6) äquivalent zu $Ax = 0$ bzw. $x \in$ Kern(A). Also ist A^TA positiv definit genau dann, wenn Kern$(A) = 0$ und damit (a) $\Longleftrightarrow$ (b).

(b) $\Longleftrightarrow$ (c) folgt aus Rang $(A) =$ dim Bild$(A) = n -$ dim Kern$(A) = n$. □

Lemma 3.24. *Es existiert eine Lösung der Gauß'schen Normalengleichung. Diese ist genau dann eindeutig, wenn* Rang $(A) = n$.

Beweis. Existenz: Wir zeigen Bild$(A^TA) =$ Bild(A^T). Wegen Bild$(A^T) =$ Kern$(A)^\perp$ ist

$$\text{Bild}(A^TA) = \text{Bild}(A^T) \Longleftrightarrow \text{Kern}(A^TA) = \text{Kern}(A).$$

Offensichtlich ist Kern$(A) \subset$ Kern(A^TA). Umgekehrt sei $x \in$ Kern(A^TA), also $A^TAx = 0$. Dann folgt auch $x^TA^TAx = \|Ax\|^2 = 0$ und daraus $x \in$ Kern(A).

Das Gleichungssystem $A^TAx = A^Tb$ hat eine eindeutige Lösung genau dann, wenn A^TA nicht singulär ist oder nach Lemma 3.23 Rang $(A) = n$ gilt. □

Ist Rang $(A) = n$, dann kann man die eindeutige Lösung des linearen Ausgleichsproblems mit Hilfe der Cholesky-Zerlegung berechnen. Dazu berechnet man A^TA mit $\frac{1}{2}mn^2$ und A^Tb mit mn Operationen. Schließlich kostet die Cholesky-Zerlegung $\frac{1}{6}n^3$ Operationen und das Lösen der beiden Dreieckssysteme n^2 Operationen. Das folgende Beispiel zeigt, dass die Lösung des linearen Ausgleichsproblems über die Normalengleichungen instabil ist:

Beispiel 3.1. Es sei

$$A = \begin{bmatrix} 1 & 1 \\ \epsilon & 0 \\ 0 & \epsilon \end{bmatrix}, \qquad b = \begin{bmatrix} 1 \\ 0 \\ 0 \end{bmatrix}, \qquad \epsilon^2 < \texttt{eps}.$$

Dann ist

$$A^T A = \begin{bmatrix} 1+\epsilon^2 & 1 \\ 1 & 1+\epsilon^2 \end{bmatrix}, \qquad A^T b = \begin{bmatrix} 1 \\ 1 \end{bmatrix}$$

Die Gauß'schen Normalengleichungen $A^T Ax = A^T b$ haben die exakte Lösung

$$x = \frac{1}{2+\epsilon^2} \begin{bmatrix} 1 \\ 1 \end{bmatrix} \approx \frac{1}{2} \begin{bmatrix} 1 \\ 1 \end{bmatrix},$$

aber in Gleitpunktrechnung ist $A^T A$ singulär, das Cholesky-Verfahren also nicht durchführbar. ◇

Das im Beispiel geschilderte Phänomen lässt sich wieder mit Hilfe der Kondition von A erklären. Hierzu verallgemeinern wir die Definition der Konditionszahl einer Matrix auf rechteckige Matrizen:

Definition 3.25. *Für $A \in \mathbb{R}^{m,n}$, $m \geq n$ mit* $\operatorname{Rang}(A) = n$ *heißt*

$$\kappa(A) = \frac{\max\limits_{\|v\|=1} \|Av\|}{\min\limits_{\|v\|=1} \|Av\|}$$

Konditionszahl *von A bezüglich $\|\cdot\|$.*

Bemerkung. Für quadratische Matrizen $A \in \mathbb{R}^{n,n}$ stimmt nach Lemma 3.10(c) die Definition mit der früheren überein.

Das Stabilitätsproblem in obigem Beispiel resultiert daraus, dass

$$\kappa(A^T A) = \kappa(A)^2 \gg \kappa(A).$$

Ein einfacher Beweis hierfür ist mit Hilfe der Singulärwertzerlegung (Abschnitt 3.8) möglich. Durch den Übergang zu den Normalengleichungen kann sich die Kondition des Problems wesentlich verschlechtern.

Wir stellen im nächsten Abschnitt eine Alternative zu den Normalengleichungen vor, die gegenüber der Cholesky-Zerlegung folgende Vorteile hat:

- Die Stabilität ist durch $\kappa(A)$ und nicht durch $\kappa(A)^2$ bestimmt
- Das Verfahren ist auch dann noch anwendbar, wenn $\operatorname{Rang}(A) < n$ gilt.

3.7 QR-Zerlegung

Gegeben sei eine Matrix $A \in \mathbb{R}^{m,n}$ mit $m \geq n$. Zunächst nehmen wir $\operatorname{Rang} A = n$ an.

Definition 3.26. *Zu $A \in \mathbb{R}^{m,n}$ mit $m \geq n$ nennt man eine Faktorisierung der Form*

$$A = QR, \qquad R = \begin{bmatrix} \widetilde{R} \\ 0 \end{bmatrix}, \qquad \widetilde{R} \in \mathbb{R}^{n,n} \tag{3.7}$$

mit einer orthogonalen Matrix $Q \in \mathbb{R}^{m,m}$ und einer oberen Dreiecksmatrix $\widetilde{R}$ **QR-Zerlegung** *von A.*

Wegen Rang $A = \text{Rang}\, R = \text{Rang}\, \widetilde{R} = n$ ist $\widetilde{R}$ nicht singulär.

Ziel dieses Abschnitts ist die effiziente und stabile Berechnung der QR-Zerlegung.

Die QR-Zerlegung hat in der Numerik viele wichtige Anwendungen. Man kann sie für $m = n$ zum Beispiel zur Lösung linearer Gleichungssysteme verwenden,

$$Ax = b \Longleftrightarrow QRx = b \Longleftrightarrow Rx = Q^T b.$$

Wir werden später sehen, dass dieses Verfahren zwar doppelt so teuer ist wie die Gauß-Elimination, aber bessere Stabilitätseigenschaften hat.

Für die Lösung linearer Ausgleichsprobleme nutzt man aus, dass die Multiplikation mit orthogonalen Matrizen die Euklid-Norm eines Vektors nicht ändert. Daher ist

$$\begin{aligned} \|b - Ax\|^2 &= \|Q^T(b - Ax)\|^2 \\ &= \|Q^T b - Rx\|^2 \qquad\qquad Q^T b =: \begin{bmatrix} c \\ d \end{bmatrix}, \quad c \in \mathbb{R}^n,\ d \in \mathbb{R}^{m-n} \\ &= \|\widetilde{R}x - c\|^2 + \|d\|^2 \end{aligned}$$

minimal genau dann, wenn $\widetilde{R}x = c$ und das Minimum ist $\|b - Ax\| = \|d\|$.

Die QR-Zerlegung spielt auch eine fundamentale Rolle beim QR-Algorithmus zur Lösung von Eigenwertproblemen. Hierzu verweisen wir auf weiterführende Vorlesungen.

Als wesentliches Hilfsmittel zur Berechnung der QR-Zerlegung dienen Householder-Transformationen:

Definition 3.27. *Für $v \in \mathbb{R}^m$, $v \neq 0$ heißt*

$$Q = I_m - \beta v v^T \in \mathbb{R}^{m,m}, \qquad \beta = \frac{2}{v^T v}$$

Householder-Matrix.

Satz 3.28. *Es sei $Q = I - \beta v v^T$, $\beta = 2/\|v\|^2$ eine Householder-Matrix. Dann gilt*

(a) $Q = Q^T$

(b) *Q ist orthogonal*

(c) $Qv = -v$

(d) *$Qw = w$ für alle $w \in \text{span}\{v\}^\perp$, also alle w in der Hyperebene senkrecht zu v.*

Beweis. Ohne Einschränkung nehmen wir $\|v\| = 1$, also $\beta = 2$ an.

(a) ist offensichtlich. (b) und (c) folgen durch einfaches Nachrechnen wegen $v^T v = 1$:

$$Q^T Q = Q^2 = (I - 2vv^T)(I - 2vv^T) = I - 4vv^T + 4vv^T vv^T = I$$

und

$$Qv = v - 2vv^T v = -v.$$

Um (d) zu zeigen, wählen wir ein beliebiges $w \in \text{span}\{v\}^\perp$ aus. Aus $v^T w = 0$ folgt mit

$$Qw = w - 2vv^T w = w$$

die Behauptung. □

Geometrisch ist die Householder-Transformation nach (c) und (d) eine Spiegelung an der Hyperebene $\text{span}\{v\}^\perp$.

Lemma 3.29. *Zu jedem $x \neq 0$, $x \notin \text{span}\{e_1\}$ gibt es eine Householder-Matrix $Q = I - \beta vv^T$ mit*

$$Qx = \alpha e_1.$$

v und α können wie folgt gewählt werden

$$v = x - \alpha e_1 \qquad \alpha = \pm\|x\|. \tag{3.8}$$

Beweis. Sei $x \notin \text{span}\{e_1\}$. Aus der Orthogonalität von Q folgt $\|Qx\| = \|x\| = |\alpha|$. Somit ist $\alpha = \pm\|x\|$. Mit

$$Qx = x - \beta v(v^T x) = \alpha e_1$$

folgt zunächst $\beta \neq 0$ und

$$v = \frac{x - \alpha e_1}{\beta v^T x}.$$

Da die Normierung von v keine Rolle spielt, können wir v wie in (3.8) wählen. □

Bemerkung. Bei der Bestimmung des Householder-Vektors sollte man einige wichtige Details beachten. Eines davon ist die Wahl des Vorzeichens von α, denn bei der Berechnung von v_1 kann ggf. Auslöschung eintreten. Hier wählt man entweder das Vorzeichen von α so, dass $\text{sign}(\alpha) = -\text{sign}(x_1)$ oder man verwendet für $\alpha = \|x\|$ und $x_1 > 0$ die von Parlett (1971) vorgeschlagene Formel

$$v_1 = x_1 - \|x\| = \frac{x_1^2 - \|x\|^2}{x_1 + \|x\|} = \frac{-(x_2^2 + \ldots + x_n^2)}{x_1 + \|x\|}.$$

In der Praxis ist es günstig, $v_1 = 1$ zu wählen. Eine stabile Implementierung mit dieser Wahl ist in Algorithmus 3.2 angegeben (aus Golub & van Loan (1996, §5.1.3)). Der Aufwand beträgt etwa n Multiplikationen, Additionen und Divisionen.

Die Multiplikation einer Householder-Matrix mit einem beliebigen Vektor erfolgt durch

$$Qx = (I - \beta vv^T)x = x - \beta(v^T x)v.$$

Sie kostet ein Skalarprodukt und ein SAXPY. Darunter versteht man eine Operation der Form $y := \alpha x + y$ mit Vektoren x, y und einem Skalar α. Es ist nicht nötig (und im Allgemeinen auch nicht sinnvoll), zunächst Q explizit zu berechnen.

Satz 3.30. *(QR-Zerlegung)*
Es sei $A \in \mathbb{R}^{m,n}$ mit $m \geq n$ und $\text{Rang}\,(A) = n$. Dann existiert eine QR-Zerlegung von A, d. h. eine orthogonale Matrix $Q \in \mathbb{R}^{m,m}$ und eine obere Dreiecksmatrix $R \in \mathbb{R}^{m,n}$ mit (3.7). Dabei sind alle Diagonalelemente r_{jj}, $j = 1, \ldots, n$ von $\widetilde{R}$ von Null verschieden.

Beweis. Zur Berechnung der QR-Zerlegung geht man spaltenweise vor. In jedem Schritt multiplizieren wir von links mit einer Householder-Matrix, um die obere Dreiecksform von R zu erhalten. Im ersten Schritt bestimmen wir eine Householder–Matrix $Q_1 = I - \beta v_1 v_1^T$ nach Lemma 3.29 so, dass die erste Spalte von A auf ein Vielfaches des ersten Einheitsvektors abgebildet wird, d. h. wir berechnen v_1 nach (3.8) mit $x = Ae_1$ und $\alpha = r_{11}$. Damit erhalten wir

$$Q_1 A = \left[\begin{array}{c|c} r_{11} & \star \\ \hline 0 & A_2 \end{array}\right]$$

Algorithmus 3.2 Berechnung des Householder-Vektors

```
function [v, β] =house(x)
n = length(x)
σ = x(2 : n)^T x(2 : n)
v = [1; x(2 : n)]
if σ = 0 then
  β = 0
else
  μ = sqrt(x(1)^2 + σ)
  if x(1) ≤ 0 then
    v(1) = x(1) − μ
  else
    v(1) = −σ/(x(1) + μ)
  end if
  β = 2v(1)^2/(σ + v(1)^2)
  v = v/v(1)
end if
```

Nach k Schritten haben wir Householdermatrizen $Q_1, \ldots, Q_k$ konstruiert mit

$$Q_k Q_{k-1} \cdots Q_1 A = \left[\begin{array}{ccc|c} r_{11} & \cdots & r_{1k} & \\ & \ddots & \vdots & R'_k \\ 0 & & r_{kk} & \\ \hline 0 & \cdots & 0 & A_{k+1} \end{array}\right]$$

wobei $R'_k \in \mathbb{R}^{k,n-k}$ und $A_{k+1} \in \mathbb{R}^{m-k,n-k}$. Im $(k+1)$-ten Schritt wählen wir jetzt wieder nach Lemma 3.29 eine Householdermatrix $\widetilde{Q}_{k+1} \in \mathbb{R}^{m-k.m-k}$ so, dass sie $A_{k+1}e_1$ auf $r_{k+1,k+1}e_1$ abbildet. Setzen wir dann

$$Q_{k+1} = \left[\begin{array}{c|c} I_k & 0 \\ \hline 0 & \widetilde{Q}_{k+1} \end{array}\right],$$

so ergibt sich

$$Q_{k+1} Q_k \cdots Q_1 A = \left[\begin{array}{ccc|cc} r_{11} & \cdots & r_{1,k} & & \\ & \ddots & \vdots & R'_k & \\ & & r_{k,k} & & \\ \hline 0 & \cdots & 0 & r_{k+1,k+1} & \star \\ \hline 0 & \cdots & 0 & 0 & A_{k+2} \end{array}\right]$$

Hierbei ist wichtig, dass die Multiplikation mit Q_{k+1} die ersten k Zeilen unverändert läßt.

Nach n Schritten ergibt sich insgesamt

$$\underbrace{Q_n Q_{n-1} \cdots Q_1}_{=:Q^T} A = R = \begin{bmatrix} r_{11} & \cdots & r_{1n} \\ & \ddots & \vdots \\ 0 & & r_{nn} \\ 0 & \cdots & 0 \end{bmatrix}$$

Q^T ist als Produkt orthogonaler Matrizen selbst orthogonal, also ist $A = QR$ die QR-Zerlegung. Wegen $\text{Rang}\,(A) = n$ ist $r_{jj} \neq 0$ für $j = 1, \ldots, n$. Im Fall $m = n$ ist $Q_n = I$, da die letzte Transformation nicht mehr nötig ist. □

Die Berechnung der QR-Zerlegung fassen wir in Algorithmus 3.3 zusammen.

Algorithmus 3.3 QR-Zerlegung mit Householder-Transformationen

```
for k = 1, ..., n do
  [v, β] = house(A(k : m, k))
  A(k : m, k : n) = A(k : m, k : n) − βv(v^T A(k : m, k : n))
  if k < m then
    A(k + 1 : m, k) = v(2 : m − k + 1)
  end if
end for
```

Nach Verlassen von Algorithmus 3.3 enthält A im oberen Dreiecksteil die Elemente von $\widetilde{R}$ und unterhalb der Diagonalen die Householder-Vektoren. Möchte man auf die Householder-Matrizen nach dem Algorithmus zugreifen, so sollte man zusätzlich noch die berechneten β's abspeichern.

Der Aufwand für den k-ten Schritt der QR-Zerlegung beträgt etwa

$$\underbrace{(m-k+1)}_{\texttt{house}} + \underbrace{(n-k+1)}_{k:n} 2 \underbrace{(m-k+1)}_{k:m} \approx 2(m-k+1)(n-k+1)$$

Additionen und Multiplikationen, der Gesamtaufwand

$$\begin{aligned} 2\sum_{k=1}^{n}(m-k+1)(n-k+1) &= 2\sum_{k=1}^{n} k(m-n+k) \approx 2\int_0^n x(m-n+x)dx \\ &= \frac{2}{3}n^3 + (m-n)n^2 + O(mn) = mn^2 - \frac{1}{3}n^3 + O(mn). \end{aligned}$$

Im Spezialfall $m = n$ ist der Aufwand mit $\frac{2}{3}n^3$ etwa doppelt so hoch wie der zur Berechnung der LR-Zerlegung. Für $m \gg n$ ist der Aufwand im Wesentlichen mn^2.

Man kann zeigen, dass Algorithmus 3.3 stabil im Sinne der Rückwärtsanalysis ist und dasselbe auch für die Lösung des linearen Ausgleichsproblems mit Hilfe der QR-Zerlegung gilt. Aus diesem Grund ist bei schlecht konditionierten Matrizen die QR-Zerlegung auch für $m = n$ gegenüber der LR-Zerlegung zu bevorzugen. Details dazu findet man in Higham (1996).

Beispiel. Wir betrachten jetzt noch einmal Beispiel 3.1. Dort hatten wir gezeigt, dass die Normalengleichungen zu

$$A = \begin{bmatrix} 1 & 1 \\ \epsilon & 0 \\ 0 & \epsilon \end{bmatrix}, \qquad b = \begin{bmatrix} 1 \\ 0 \\ 0 \end{bmatrix}, \qquad \epsilon^2 < \texttt{eps}.$$

in Gleitpunktarithmetik nicht mehr mit dem Cholesky-Verfahren gelöst werden können. Die QR-Zerlegung berechnet bis auf Terme der Größenordnung ϵ^2

$$r_{11} \approx -1, \qquad v_1 \approx \begin{bmatrix} 2 \\ \epsilon \\ 0 \end{bmatrix}, \qquad r_{22} \approx \epsilon\sqrt{2}, \qquad v_2 \approx \epsilon \begin{bmatrix} 0 \\ -1-\sqrt{2} \\ 1 \end{bmatrix}.$$

Damit ist

$$R \approx \begin{bmatrix} -1 & -1 \\ 0 & \epsilon\sqrt{2} \\ 0 & 0 \end{bmatrix}, \qquad Q^T b = Q_2 Q_1 b \approx \begin{bmatrix} -1 \\ \frac{\epsilon}{\sqrt{2}} \\ -\frac{\epsilon}{\sqrt{2}} \end{bmatrix}.$$

Die Lösung des Minimierungsproblems $\|b - Ax\| = \min$ mit der QR-Zerlegung liefert in Gleitpunktarithmetik die exakte Lösung. Dies ist dadurch zu verstehen, dass die Stabilität von der Kondition von $\widetilde{R}$ abhängt. Es gilt

$$\kappa(\widetilde{R}) = \kappa(R) = \kappa(QR) = \kappa(A).$$

Die Stabilität wird also durch $\kappa(A)$ und nicht wie bei den Normalengleichungen durch $\kappa^2(A)$ bestimmt. $\diamond$

Hat $A \in \mathbb{R}^{m,n}$, $m \geq n$ nicht vollen Spaltenrang, etwa $\text{Rang}\,(A) = p < n$, dann wäre in exakter Arithmetik ein Diagonalelement Null, $r_{jj} = 0$ für ein $j < n$. Der Algorithmus bricht dann vorzeitig ab. Man kann jedoch durch Spaltenpivotsuche eine Permutationsmatrix P so finden, dass

$$AP = QR = Q \begin{bmatrix} R_{11} & R_{12} \\ 0 & 0 \end{bmatrix}, \qquad R_{11} \in \mathbb{R}^{p,p} \text{ nicht singulär}, \; R_{12} \in \mathbb{R}^{p,n-p}$$

gilt.

Wir verzichten hier auf die Details und betrachten nur das Ausgleichsproblem im Fall, dass A nicht vollen Rang hat. Dazu sei nun allgemein

$$\mathcal{L} = \{x \in \mathbb{R}^n : \|b - Ax\|_2 = \min\}$$

die Lösungsmenge des linearen Ausgleichsproblems. Ist $x \in \mathcal{L}$ beliebig, dann ist

$$\mathcal{L} = \{y \mid y = x + z, \; z \in \text{Kern}(A)\}. \tag{3.9}$$

Die Lösung $x_{LS} \in \mathcal{L}$ mit $x_{LS} \in \text{Kern}(A)^\perp = \text{Bild}(A^T)$ heißt **kleinste Quadrate-Lösung**, denn ist $x \in \mathcal{L}$ beliebig, so ist wegen (3.9)

$$x = x_{LS} \oplus z, \qquad z \in \text{Kern}(A).$$

Der Satz von Pythagoras liefert

$$\|x\|_2^2 = \|x_{LS}\|_2^2 + \|z\|_2^2.$$

Damit ist

$$\|x_{LS}\|_2 = \min_{x \in \mathcal{L}} \|x\|_2$$

die Lösung mit minimaler Norm (*least squares solution*).

3.8 Singulärwertzerlegung und Pseudoinverse

Beim linearen Ausgleichsproblem spielt die Matrix A^TA bzw. im Komplexen A^HA offensichtlich eine wichtige Rolle.

Satz 3.31. *(Singulärwertzerlegung)*
Zu jeder Matrix $A \in \mathbb{C}^{m,n}$ gibt es unitäre Matrizen $U \in \mathbb{C}^{m,m}$ und $V \in \mathbb{C}^{n,n}$, so dass

$$U^HAV = \Sigma = \operatorname{diag}(\sigma_1, \ldots, \sigma_k) \in \mathbb{R}^{m,n}, \qquad k = \min\{m, n\}$$

mit $\sigma_1 \geq \sigma_2 \geq \cdots \geq \sigma_k \geq 0$.

Beweis. Wir beweisen die Behauptung mit Induktion nach m und n. Ohne Einschränkung nehmen wir $m \geq n$ an, denn sonst betrachten wir A^T statt A.

Seien $x \in \mathbb{C}^n$, $y \in \mathbb{C}^m$ mit $\|x\| = \|y\| = 1$ so gewählt, dass

$$Ax = \sigma y, \qquad \sigma = \|A\|.$$

Dies ist nach Definition von $\|A\| = \max_{\|x\|=1}\|Ax\|$ immer möglich. Wir ergänzen x zu einer Orthonormalbasis von $\mathbb{C}^n$ und y zu einer Orthonormalbasis von $\mathbb{C}^m$:

$$V = \begin{bmatrix} x & V_2 \end{bmatrix} \in \mathbb{C}^{n,n}, \qquad U = \begin{bmatrix} y & U_2 \end{bmatrix} \in \mathbb{C}^{m,m}.$$

Im Fall $n = 1$ ist keine Ergänzung von x nötig. Für beliebiges m gilt in diesem Fall

$$U^HAV = U^HAx = \sigma U^Hy = \sigma U^HUe_1 = \sigma e_1 = \Sigma.$$

Wir können also annehmen, dass die Behauptung für $(m-1) \times (n-1)$ Matrizen gilt und zeigen jetzt die Existenz für $m \times n$ Matrizen.

Nach Konstruktion sind U und V unitär und es gilt

$$U^HAV = \left[\begin{array}{c|c} \sigma & w^H \\ \hline 0 & B \end{array}\right] =: A_1.$$

Ferner folgt aus

$$\|A_1 \begin{bmatrix} \sigma \\ w \end{bmatrix}\|^2 = \|\begin{bmatrix} \sigma^2 + w^Hw \\ Bw \end{bmatrix}\|^2 \geq (\sigma^2 + w^Hw)^2$$

$\|A_1\|^2 \geq \sigma^2 + \|w\|^2$. Wegen $\sigma = \|A\| = \|A_1\|$ muss also $w = 0$ gelten. Die Behauptung schließt man nun durch Anwendung der Induktionsannahme auf B. □

Mit Hilfe der Singulärwertzerlegung haben A und A^H die Darstellung

$$A = \sum_{j=1}^{p} \sigma_j u_j v_j^H, \qquad A^H = \sum_{j=1}^{p} \sigma_j v_j u_j^H. \tag{3.10}$$

Ist A reell, dann können U, V ebenfalls reell gewählt werden.

Definition 3.32. *Die nicht negativen Zahlen $\sigma_1, \ldots, \sigma_k$ aus Satz 3.31 heißen* **Singulärwerte** *von A, die Spalten von $V = \begin{bmatrix} v_1 & \cdots & v_n \end{bmatrix}$ heißen* **rechte Singulärvektoren** *und die Spalten von $U = \begin{bmatrix} u_1 & \cdots & u_m \end{bmatrix}$ heißen* **linke Singulärvektoren**.

Satz 3.33. *Es sei $U^HAV = \Sigma$ die Singulärwertzerlegung von $A \in \mathbb{C}^{m,n}$ mit $m \geq n$. Dann gilt*

(a) $\{\sigma_1^2, \ldots, \sigma_n^2\} = \lambda(A^H A)$ *und die rechten Singulärvektoren* v_i *bilden eine Orthonormalbasis von Eigenvektoren von* $A^H A$.

(b) *Die Eigenwerte von* AA^H *sind* $\sigma_1^2, \ldots, \sigma_n^2$ *und* $m-n$ *Nullen. Die linken Singulärvektoren* u_i *sind orthonormierte Eigenvektoren von* AA^H.

(c) $\|A\| = \sigma_1$.

(d) *Ist* $m = n$ *und* A *nicht singulär, dann gilt* $\|A^{-1}\| = \sigma_n^{-1}$.

(e) *Ist* $\sigma_1 \geq \cdots \geq \sigma_p > \sigma_{p+1} = \cdots = \sigma_n = 0$, *dann ist* $\mathrm{Rang}\,(A) = p$.

(f) $Av_i = \sigma_i u_i$, $i = 1, \ldots, p$ *und* $Av_i = 0$, $i = p+1, \ldots, n$.

(g) $A^H u_i = \sigma_i v_i$, $i = 1, \ldots, p$ *und* $A^H u_i = 0$, $i = p+1, \ldots, m$.

Beweis. (a), (b) Es sei

$$\Sigma = \begin{bmatrix} \widetilde{\Sigma} \\ 0 \end{bmatrix}, \qquad \widetilde{\Sigma} = \mathrm{diag}(\sigma_1, \ldots, \sigma_n).$$

Dann gilt

$$A^H A = (U\Sigma V^H)^H (U\Sigma V^H) = V\Sigma^T \Sigma V^H = V\widetilde{\Sigma}^2 V^H$$

und

$$AA^H = (U\Sigma V^H)(U\Sigma V^H)^H = U\Sigma\Sigma^T U^H = U \begin{bmatrix} \widetilde{\Sigma}^2 & \\ & 0 \end{bmatrix} U^H.$$

(c) Folgt aus $\|A\| = \|\Sigma\| = \sigma_1$.

(d) Da $A^{-1} = V\Sigma^{-1}U^H$ die Singulärwertzerlegung von A^{-1} ist, folgt die Behauptung aus $\|\Sigma^{-1}\| = \sigma_n^{-1}$.

(e) Da U und V unitär sind, haben A und Σ gleichen Rang. Der Rang der Diagonalmatrix Σ ist offensichtlich p.

(f), (g) folgen aus der Darstellung (3.10). □

Definition 3.34. *Sei* $A = U\Sigma V^H$ *die Singulärwertzerlegung von* $A \in \mathbb{C}^{m,n}$ *mit* $\mathrm{Rang}\,(A) = p > 0$. *Dann heißt*

$$A^+ = V\Sigma^+ U^H, \qquad \Sigma^+ := \mathrm{diag}(\sigma_1^{-1}, \ldots, \sigma_p^{-1}, 0, \ldots, 0) \in \mathbb{C}^{n,m}$$

Pseudoinverse *oder* **Moore-Penrose-Pseudoinverse** *von* A.

Aus der zu (3.10) analogen Darstellung

$$A^+ = \sum_{j=1}^{p} \sigma_j^{-1} v_j u_j^H.$$

für die Pseudoinverse folgt unmittelbar

$$\mathrm{Kern}(A^+) = \mathrm{Kern}(A^H) = \mathrm{Bild}(A)^\perp, \qquad \mathrm{Bild}(A^+) = \mathrm{Bild}(A^H) = \mathrm{Kern}(A)^\perp.$$

Satz 3.35. *Sei $A = U\Sigma V^H$ die Singulärwertzerlegung von $A \in \mathbb{C}^{m,n}$ mit* Rang $A = p$ *und sei $b \in \mathbb{C}^m$. Dann ist*

$$x_{LS} = A^+ b = \sum_{i=1}^{p} \frac{u_i^H b}{\sigma_i} v_i$$

*die Lösung von $\|b - Ax\| =$ min! mit minimaler Norm (*least squares solution*) und es gilt*

$$\|b - Ax_{LS}\|^2 = \sum_{i=p+1}^{m} (u_i^H b)^2.$$

Beweis. Es gilt

$$\|b - Ax\|^2 = \|U^H b - U^H AVV^H x\|^2 = \|U^H b - \Sigma y\|^2,$$

wobei $y = V^H x$. Zerlegen wir $U = [U_p\ \widetilde{U}]$ und $y = \begin{bmatrix} y_p \\ \widetilde{y} \end{bmatrix}$, so gilt

$$\begin{aligned} \|b - Ax\|^2 &= \| \begin{bmatrix} U_p^H b \\ \widetilde{U}^H b \end{bmatrix} - \begin{bmatrix} \Sigma_p y_p \\ 0 \end{bmatrix} \|^2 \\ &= \|U_p^H b - \Sigma_p y_p\|^2 + \|\widetilde{U}^H b\|^2. \end{aligned}$$

Dieser Ausdruck wird minimal genau dann, wenn $y_p = \Sigma_p^{-1} U_p^H b$ und für beliebige $\widetilde{y}$. Wegen

$$x = Vy = V \begin{bmatrix} y_p \\ \widetilde{y} \end{bmatrix}$$

ist $\|x\|^2 = \|y\|^2 = \|y_p\|^2 + \|\widetilde{y}\|^2$ minimal für $\widetilde{y} = 0$, also ist

$$x = V \begin{bmatrix} y_p \\ \widetilde{y} \end{bmatrix} = V \begin{bmatrix} \Sigma_p^{-1} & 0 \\ 0 & 0 \end{bmatrix} U^H b = V\Sigma^+ U^H b$$

die Lösung mit minimaler Norm. □

Für Anwendungen wichtig ist die folgende Optimalitätseigenschaft:

Satz 3.36. *Es sei $A = U\Sigma V^H$ die Singulärwertzerlegung von A und*

$$A_k = U \begin{bmatrix} \Sigma_k & 0 \\ 0 & 0 \end{bmatrix} V^H = \sum_{j=1}^{k} \sigma_j u_j v_j^H, \qquad \Sigma_k = \mathrm{diag}(\sigma_1, \ldots, \sigma_k).$$

Dann gilt

$$\|A - B\| \geq \|A - A_k\| = \sigma_{k+1}$$

für alle $B \in \mathbb{C}^{m,n}$ mit Rang $B = k$.

Beweis. Wegen (3.10) gilt

$$\|A - A_k\| = \| \sum_{j=k+1}^{n} \sigma_j u_j v_j^H \| = \|U \mathrm{diag}(0, \ldots, 0, \sigma_{k+1}, \ldots, \sigma_n) V^H\| = \sigma_{k+1}.$$

Sei $B \in \mathbb{C}^{m,n}$ mit Rang $B = k$ beliebig gewählt. Dann ist $\dim \mathrm{Kern} B = n - k$ und es gilt

$$\mathrm{Kern} B \cap \mathrm{span}\{v_1, \ldots, v_{k+1}\} \neq \{0\}.$$

Wir wählen uns y aus dieser Menge mit $\|y\| = 1$. Dann ist wegen $By = 0$

$$\begin{aligned}\|A - B\|^2 \geq \|(A - B)y\|^2 &= \|Ay\|^2 \\ &= \|U\Sigma V^H y\|^2 \\ &= \|\Sigma V^H y\|^2 \\ &\geq \sigma_{k+1}^2.\end{aligned}$$

Die letzte Ungleichung gilt, da $(V^H y)_j = 0$ für $j \geq k + 2$ wegen $y \in \text{span}\{v_1, \ldots, v_{k+1}\}$. □