4 Lineare Gleichungssysteme I (Direkte Verfahren)

Seien A eine Matrix und b ein Vektor

$$A = (a_{jk})_{j,k=1}^{m,n} = \begin{bmatrix} a_{11} & \cdots & a_{1n} \\ \vdots & & \vdots \\ a_{m1} & \cdots & a_{mn} \end{bmatrix}, \qquad b = (b_j)_{j=1,\ldots,m} = \begin{bmatrix} b_1 \\ \vdots \\ b_m \end{bmatrix}.$$

Gesucht ist ein Vektor $x = (x_k)_{k=1,\ldots,n}$ mit der Eigenschaft

$$\begin{aligned} a_{11}x_1 + a_{12}x_2 + \cdots + a_{1n}x_n &= b_1 \\ &\vdots \\ a_{m1}x_1 + a_{m2}x_2 + \cdots + a_{mn}x_n &= b_m \end{aligned}$$

oder abgekürzt geschrieben: $Ax = b$.

Definition 4.1: *Das lineare Gleichungssystem $Ax = b$ heißt „unterbestimmt“ im Fall $m < n$, „quadratisch“ im Fall $m = n$ und „überbestimmt“ im Fall $m > n$.*

Das lineare Gleichungssystem ist genau dann lösbar, wenn $\mathrm{Rang}(A) = \mathrm{Rang}[A, b]$, mit der zusammengesetzten Matrix

$$[A, b] = \left[\begin{array}{ccc|c} a_{11} & \cdots & a_{1n} & b_1 \\ \vdots & & \vdots & \vdots \\ a_{m1} & \cdots & a_{mn} & b_m \end{array}\right].$$

Im „quadratischen“ Fall $m = n$ sind die folgenden Aussagen äquivalent:

(i) $Ax = b$ ist für jedes b eindeutig lösbar.

(ii) $\mathrm{Rang}(A) = n$.

(iii) $\det(A) \neq 0$.

(iv) Alle Eigenwerte von A sind ungleich Null.

Wir beschäftigen uns im Folgenden hauptsächlich mit der Lösung von quadratischen Gleichungssystemen. Die dazu verwendeten Verfahren lassen sich grob in zwei Klassen einteilen:

Definition 4.2: *Ein „direktes“ Verfahren zur Lösung des Gleichungssystems $Ax = b$ ist ein Algorithmus, der (bei Vernachlässigung von Rundungsfehlern) in endlich vielen Schritten die Lösung x liefert. Im Gegensatz dazu erzeugen die „iterativen“ Verfahren sukzessive eine Folge von Vektoren $(x^{(t)})_{t\in\mathbb{N}}$, die im Limes für $t \to \infty$ immer bessere Approximationen zur Lösung x sind.*

4.1 Störungstheorie

Wir beschäftigen uns zunächst mit dem Problem der „Konditionierung“ von quadratischen linearen Gleichungssystemen. Bei der Lösung eines Gleichungssystems $Ax = b$ treten zwei Fehlereinflüsse auf:

a) Fehler in der „theoretischen“ Lösung aufgrund von Eingangsfehlern in den Elementen von A und b,

b) Fehler in der „numerischen“ Lösung aufgrund des Rundungsfehlers im Verlaufe des Lösungsprozesses.

4.1.1 Vektor- und Matrizennormen

Zur Erfassung dieser Fehler benötigen wir ein Maß für die „Größe“ von Vektoren und Matrizen. Dazu dienen üblicherweise Normen auf dem n-dimensionalen Zahlenraum $\mathbb{K}^n$, $\mathbb{K} = \mathbb{R}$ oder $\mathbb{K} = \mathbb{C}$. (Im Hinblick auf spätere Anwendungen lassen wir im Folgenden auch komplexe Vektoren bzw. Matrizen zu.)

Definition 4.3: *Eine Abbildung* $\|\cdot\| : \mathbb{K}^n \to \mathbb{R}_+$ *heißt „Norm“, wenn sie folgende Eigenschaften besitzt:*

(N1) $\|x\| > 0\,, \quad x \in \mathbb{K}^n \setminus \{0\}$ *(Definitheit),*

(N2) $\|\alpha x\| = |\alpha|\,\|x\|\,, \quad x \in \mathbb{K}^n,\ \alpha \in \mathbb{K}$ *(positive Homogenität),*

(N3) $\|x + y\| \le \|x\| + \|y\|\,, \quad x, y \in \mathbb{K}^n$ *(Subadditivität).*

Beispiel 4.1: *Gebräuchliche Beispiele von Vektornormen sind:*

$$\begin{aligned} \|x\|_2 &:= \Big(\sum_{i=1}^{n} |x_i|^2\Big)^{1/2} && \text{„euklidische Norm“ } (l_2\text{-Norm)} \\ \|x\|_\infty &:= \max_{i=1,\ldots,n} |x_i| && \text{„Maximumnorm“ } (l_\infty\text{-Norm)} \\ \|x\|_1 &:= \sum_{i=1}^{n} |x_i| && \text{„}l_1\text{-Norm“} \end{aligned}$$

Mit Hilfe einer Norm $\|\cdot\|$ auf $\mathbb{K}^n$ lässt sich die Konvergenz einer Folge von Vektoren gegen einen Vektor erklären durch

$$x^{(t)} \to x \ \ (t \to \infty) \quad :\Longleftrightarrow \quad \|x^{(t)} - x\| \to 0 \ \ (t \to \infty)\,.$$

Die sog. „Dreiecksungleichung“ (N3) ergibt über die Beziehung $\|x\| = \|x - y + y\|$ die wichtige Ungleichung

$$\|x - y\| \ge \big|\|x\| - \|y\|\big|\,, \quad x, y \in \mathbb{K}^n\,, \tag{4.1.1}$$

welche u. a. die Stetigkeit von Normen als Funktionen von $\mathbb{K}^n$ in $\mathbb{R}$ impliziert.

Hilfssatz 4.1 (Normäquivalenz): *Auf dem endlich dimensionalen Vektorraum $\mathbb{K}^n$ sind alle Normen äquivalent, d. h.: Zu je zwei Normen $\|\cdot\|$, $\|\cdot\|'$ gibt es positive Konstanten m, M, mit denen gilt:*

$$m\,\|x\| \le \|x\|' \le M\,\|x\|\,, \quad x \in \mathbb{K}^n\,. \tag{4.1.2}$$

Beweis: Es genügt, die Behauptung für den Fall zu zeigen, dass eine der beiden Normen die Maximumnorm $\|\cdot\|_\infty$ ist. Sei $\|\cdot\|$ irgendeine zweite Norm. Bzgl. der kartesischen Einheitsvektoren $e_1, \ldots, e_n$ hat jeder Vektor $x \in \mathbb{K}^n$ die Darstellung $x = \sum_{i=1}^n x_i e_i$. Folglich gilt

$$\|x\| \le \gamma\,\|x\|_\infty\,, \quad \gamma := \sum_{i=1}^n \|e_i\|\,.$$

Die Norm $\|\cdot\|$ ist also auch stetig bzgl. der komponentenweisen Konvergenz von Vektoren. Die Punktmenge

$$S \equiv \{x \in \mathbb{K}^n,\ \|x\|_\infty = 1\} \subset \mathbb{K}^n$$

ist beschränkt und abgeschlossen (und damit kompakt). Die Norm $\|\cdot\|$ nimmt also als stetige Funktion auf S ihr Minimum und Maximum an. Es existieren also $x_0, x_1 \in S$, so dass

$$0 < \|x_0\| \le \|x\| \le \|x_1\| < \infty\,, \quad \forall x \in S\,.$$

Für beliebiges $y \in \mathbb{K}^n \setminus \{0\}$ ist $y/\|y\|_\infty \in S$ und folglich

$$\|x_0\| \le \|y\|/\|y\|_\infty \le \|x_1\|\,.$$

Mit $m \equiv \|x_0\|$ und $M \equiv \|x_1\|$ gilt daher

$$m\,\|y\|_\infty \le \|y\| \le M\,\|y\|_\infty\,, \quad \forall y \in \mathbb{K}^n,$$

was zu zeigen war. Q.E.D.

Die Beziehung (4.1.2) impliziert, dass die durch eine beliebige Norm induzierte Konvergenz von Vektoren stets äquivalent zur „komponentenweisen“ Konvergenz ist.

Wir betrachten nun den Vektorraum der $n \times n$-Matrizen $A \in \mathbb{K}^{n\times n}$. Offenbar kann dieser mit dem Vektorraum der n^2-Vektoren identifiziert werden. Somit übertragen sich alle Aussagen für Vektornormen auf Normen für Matrizen. Insbesondere sind alle Normen für $n \times n$-Matrizen äquivalent, und die Konvergenz von Matrizen ist die komponentenweise Konvergenz:

$$A^{(t)} \to A\ \ (t \to \infty) \quad \Longleftrightarrow \quad a_{jk}^{(t)} \to a_{jk}\ \ (t \to \infty)\,, \quad j,k = 1, \ldots, n\,.$$

Definition 4.4: *Eine Norm $\|\cdot\|$ auf $\mathbb{K}^{n\times n}$ heißt „verträglich“ mit einer Vektornorm $\|\cdot\|$ auf $\mathbb{K}^n$, wenn gilt:*

$$\|Ax\| \le \|A\|\,\|x\|\,, \quad x \in \mathbb{K}^n\,, \quad A \in \mathbb{K}^{n\times n}.$$

Sie heißt „Matrizennorm“, wenn sie submultiplikativ ist:

$$\|AB\| \le \|A\|\,\|B\|\,, \quad A, B \in \mathbb{K}^{n\times n},$$

Z. B. ist die Quadratsummennorm (sog. „Frobenius[1]-Norm“)

$$\|A\|_{\mathrm{Fr}} := \Big(\sum_{j,k=1}^{n} |a_{jk}|^2\Big)^{1/2}$$

eine mit der euklidischen Vektornorm verträgliche Matrizennorm. Für eine beliebige Vektornorm $\|\cdot\|$ auf $\mathbb{K}^n$ wird durch

$$\|A\| := \sup_{x\in\mathbb{K}^n\setminus\{0\}} \frac{\|Ax\|}{\|x\|} = \sup_{x\in\mathbb{K}^n,\|x\|=1} \|Ax\|$$

eine mit $\|\cdot\|$ verträgliche Matrizennorm erklärt (Übungsaufgabe). Diese heißt die von $\|\cdot\|$ erzeugte „natürliche“ Matrizennorm. Für natürliche Matrizennormen gilt

$$\|I\| = \sup_{x\in\mathbb{K}^n\setminus\{0\}} \frac{\|Ix\|}{\|x\|} = 1.$$

Hilfssatz 4.2: *Die natürlichen Matrizennormen zu $\|\cdot\|_\infty$ und $\|\cdot\|_1$ sind die „maximale Zeilensumme“ bzw. die „maximale Spaltensumme“:*

$$\|A\|_\infty := \max_{1\le j\le n} \sum_{k=1}^{n} |a_{jk}|, \qquad \|A\|_1 := \max_{1\le k\le n} \sum_{j=1}^{n} |a_{jk}|\,. \tag{4.1.3}$$

Beweis: Wir geben den Beweis nur für $\|\cdot\|_\infty$. Für $\|\cdot\|_1$ verläuft er analog.
(i) Die maximale Zeilensumme $\|\cdot\|_\infty$ ist eine Matrizennorm. Die Normeigenschaften (N1) - (N3) folgen mit Hilfe der entsprechenden Eigenschaften des Absolutbetrags, und für ein Matrizenprodukt AB gilt

$$\begin{aligned}\|AB\|_\infty &= \max_{1\le i\le n} \Big|\sum_{j=1}^{n}\Big(\sum_{k=1}^{n} a_{ik}b_{kj}\Big)\Big| \le \max_{1\le i\le n}\sum_{k=1}^{n}\Big(|a_{ik}|\sum_{j=1}^{n}|b_{kj}|\Big)\\ &\le \max_{1\le i\le n}\sum_{k=1}^{n}|a_{ik}|\ \max_{1\le k\le n}\sum_{j=1}^{n}|b_{kj}| = \|A\|_\infty\|B\|_\infty.\end{aligned}$$

(ii) Weiter ist die maximale Zeilensumme wegen

$$\|Ax\|_\infty = \max_{1\le j\le n} |\sum_{k=1}^{n} a_{jk}x_k| \le \max_{1\le j\le n}\sum_{k=1}^{n}|a_{jk}|\ \max_{1\le k\le n}|x_k| = \|A\|_\infty\|x\|_\infty$$

[1]Ferdinand Georg Frobenius (1849–1917): Deutscher Mathematiker; Professor in Zürich und Berlin; bedeutende Beiträge zur Theorie der Differentialgleichungen, zu Determinanten und Matrizen sowie zur Gruppentheorie.

verträglich mit $\|\cdot\|_\infty$, und es gilt

$$\sup_{\|x\|_\infty=1} \|Ax\|_\infty \le \|A\|_\infty.$$

(iii) Im Falle $\|A\|_\infty = 0$ ist $A = 0$, d. h. trivialerweise

$$\|A\|_\infty = \sup_{\|x\|_\infty=1} \|Ax\|_\infty.$$

(iv) Sei also $\|A\|_\infty > 0$ und $m \in \{1, \ldots, n\}$ ein Index mit der Eigenschaft

$$\|A\|_\infty = \max_{1\le j\le n} \sum_{k=1}^{n} |a_{jk}| = \sum_{k=1}^{n} |a_{mk}|.$$

Wir setzen für $k = 1, \ldots, n$:

$$z_k := \begin{cases} |a_{mk}|/a_{mk} & \text{für } a_{mk} \neq 0, \\ 0, & \text{sonst,} \end{cases}$$

d. h.: $z = (z_k)_{k=1}^n \in \mathbb{K}^n$, $\|z\|_\infty = 1$. Für $v := Az$ gilt dann

$$v_m = \sum_{k=1}^{n} a_{mk} z_k = \sum_{k=1}^{n} |a_{mk}| = \|A\|_\infty.$$

Folglich ist

$$\|A\|_\infty = v_m \le \|v\|_\infty = \|Az\|_\infty \le \sup_{\|y\|_\infty=1} \|Ay\|_\infty,$$

was zu zeigen war. Q.E.D.

4.1.2 Eigenwerte und Skalarprodukte

Die „Eigenwerte“ $\lambda \in \mathbb{K}$ einer Matrix $A \in \mathbb{K}^{n\times n}$ sind definiert als die (komplexen) Nullstellen ihres charakteristischen Polynoms $p(\lambda) = \det(A - \lambda I)$. Folglich existieren genau n (ihrer Vielfachheit als Nullstelle entsprechend oft gezählte) Eigenwerte λ, und zu jedem λ existiert mindestens ein „Eigenvektor“ $w \in \mathbb{K}^n \setminus \{0\} : Aw = \lambda w$. Sei nun $\|\cdot\|$ eine beliebige Vektornorm und $\|\cdot\|$ eine damit verträgliche Matrizennorm, (wobei die beiden Normen der Einfachheit halber wieder gleich bezeichnet werden). Mit einem normierten Eigenvektor zum Eigenwert λ gilt

$$|\lambda| = |\lambda|\,\|w\| = \|\lambda w\| = \|Aw\| \le \|A\|\,\|w\| = \|A\|, \tag{4.1.4}$$

d. h. alle Eigenwerte von A liegen in einer Kreisscheibe in $\mathbb{C}$ mit Mittelpunkt Null und Radius $\|A\|$. Speziell mit $\|A\|_\infty$ erhält man die Abschätzung

$$\max |\lambda| \le \|A\|_\infty = \max_{1\le j\le n} \sum_{k=1}^{n} |a_{jk}|. \tag{4.1.5}$$

Eine Matrix $A \in \mathbb{K}^{n\times n}$ heißt „hermitesch“, wenn gilt:

$$A = \bar{A}^T \quad \text{bzw.} \quad a_{jk} = \overline{a_{kj}},\ j,k = 1,\dots,n\,.$$

Reelle hermitesche Matrizen werden „symmetrisch“ genannt. Hermitesche Matrizen haben nur reelle Eigenwerte und besitzen dazu eine Orthonormalbasis von Eigenvektoren. Der Begriff der Symmetrie ist eng verknüpft mit dem des Skalarprodukts.

Definition 4.5: *Eine Abbildung* $(\cdot,\cdot) : \mathbb{K}^n \times \mathbb{K}^n \to \mathbb{K}$ *wird ein „Skalarprodukt“ genannt, wenn sie folgende Eigenschaften hat:*

(S1) $(x,y) = \overline{(y,x)}\,, \quad x,y \in \mathbb{K}^n$ *(Symmetrie),*

(S2) $(\alpha x + \beta y, z) = \alpha(x,z) + \beta(y,z)\,, \quad x,y,z \in \mathbb{K}^n,\ \alpha,\beta \in \mathbb{K}$ *(Linearität),*

(S3) $(x,x) > 0\,, \quad x \in \mathbb{K}^n \setminus \{0\}$ *(Definitheit).*

Ein Skalarprodukt auf $\mathbb{K}^n \times \mathbb{K}^n$ *erzeugt eine zugehörige Vektornorm durch*

$$\|x\| := (x,x)^{1/2}, \quad x \in \mathbb{K}^n.$$

Im Folgenden wird fast ausschließlich das sog. „euklidische“ Skalarprodukt verwendet:

$$(x,y)_2 = \sum_{j=1}^{n} x_j \overline{y_j}\,, \quad (x,x)_2 = \|x\|_2^2.$$

Mit Hilfe des euklidischen Skalarprodukts lässt sich die Eigenschaft einer Matrix, hermitesch zu sein, äquivalent ausdrücken durch:

$$A = \bar{A}^T \quad \Longleftrightarrow \quad (Ax,y)_2 = (x,Ay)_2, \quad x,y \in \mathbb{K}^n.$$

Die von der euklidischen Vektornorm erzeugte natürliche Matrizennorm heißt die „Spektralnorm“ und wird mit $\|\cdot\|_2$ bezeichnet. Diese Bezeichnung ist durch das folgende Resultat gerechtfertigt:

Hilfssatz 4.3: *Für die Spektralnorm hermitescher Matrizen* $A \in \mathbb{K}^{n\times n}$ *gilt*

$$\|A\|_2 = \max\{|\lambda|,\ \lambda \textit{ Eigenwert von } A\}\,. \tag{4.1.6}$$

Für allgemeine Matrizen $A \in \mathbb{K}^{n\times n}$ *gilt*

$$\|A\|_2 = \max\{|\lambda|^{1/2},\ \lambda \textit{ Eigenwert von } \bar{A}^T A\}. \tag{4.1.7}$$

Beweis: Bekanntlich besitzt eine hermitesche Matrix $A \in \mathbb{K}^{n\times n}$ nur reelle Eigenwerte und zwar genau n Stück (ihrer Vielfachheit entsprechend oft gezählt), $\lambda_1, \ldots, \lambda_n \in \mathbb{R}$. Ferner existiert ein zugehöriges „Orthonormalsystem“ von Eigenvektoren

$$\{w^1, \ldots, w^n\} \subset \mathbb{K}^n : Aw^i = \lambda_i w^i\,, \quad (w^i, w^j)_2 = \delta_{ij}, \quad i,j = 1, \ldots, n.$$

Jedes $x \in \mathbb{K}^n$ besitzt eine Darstellung der Form

$$x = \sum_{i=1}^{n} \alpha_i w^i\,, \quad \alpha_i = (x, w^i)_2,$$

und es gilt

$$\|x\|_2^2 = (x,x)_2 = \sum_{i,j=1}^{n} \alpha_i \bar{\alpha}_j\, (w^i, w^j)_2 = \sum_{i=1}^{n} |\alpha_i|^2,$$

$$\|Ax\|_2^2 = (Ax, Ax)_2 = \sum_{i,j=1}^{n} \lambda_i \alpha_i \overline{\lambda_j \alpha_j}\, (w^i, w^j)_2 = \sum_{i=1}^{n} \lambda_i^2\, |\alpha_i|^2.$$

Hiermit folgt

$$\|A\|_2^2 = \sup_{x\in\mathbb{K}^n\setminus\{0\}} \frac{\|Ax\|_2^2}{\|x\|_2^2} = \sup_{x\in\mathbb{K}^n\setminus\{0\}} \frac{\sum_{i=1}^n \lambda_i^2\,|\alpha_i|^2}{\sum_{i=1}^n |\alpha_i|^2} \le \max_{1\le i\le n} |\lambda_i|^2.$$

Wegen der Eigenwertschranke (4.1.4) ergibt sich damit die Behauptung. Q.E.D.

Eine Matrix $A \in \mathbb{K}^{n\times n}$ heißt „positiv definit“, wenn gilt:

$$(Ax, x)_2 \in \mathbb{R}\,, \quad (Ax, x)_2 > 0 \quad \forall x \in \mathbb{K}^n \setminus \{0\}.$$

Lemma 4.1: *Eine symmetrische Matrix $A \in \mathbb{R}^{n\times n}$ ist genau dann positiv definit, wenn alle ihre (reellen) Eigenwerte positiv sind. Alle ihre Hauptdiagonalelemente sind positiv, und ihr betragsmäßig größtes Element liegt auf der Hauptdiagonalen.*

Beweis: Seien $\lambda_1, \ldots, \lambda_n \in \mathbb{R}$ die (ihrer Vielfachheiten entsprechend oft gezählten) Eigenwerte der symmetrischen Matrix A und $\{w^1, \ldots, w^n\}$ eine zugehörige Orthonormalbasis von Eigenvektoren.

(i) Sei $\lambda \in \mathbb{R}$ Eigenwert und $v \in \mathbb{R}^n$, $\|v||_2 = 1$, ein zugehöriger Eigenvektor von A,

$$Av = \lambda v.$$

Aus der positiven Definitheit von A folgt $\lambda = \lambda \|v\|_2^2 = (Av, v)_2 > 0$. Sind umgekehrt alle Eigenwerte von A positiv, so folgt für beliebigen Vektor $v = \sum_{i=1}^n \alpha_i w^i \neq 0$:

$$(Av, v)_2 = \sum_{i,j=1}^{n} \lambda_i \alpha_i \alpha_j (w^i, w^j)_2 = \sum_{i=1}^{n} \lambda_i \alpha_i^2 > 0.$$

(ii) Mit dem k-ten kartesischen Einheitsvektor e^k liefert die positive Definitheit von A:

$$a_{kk} = \sum_{i,j=1}^{n} a_{ij}\delta_{ik}\delta_{jk} = (Ae^k, e^k)_2 > 0.$$

(iii) Sei $a_{ij} \neq 0$ ein betragsmäßig größtes Element von A, und sei $i \neq j$. Wir wählen $x = e^i - \mathrm{sign}(a_{ij})e^j \neq 0$ und erhalten wieder aus der positiven Definitheit von A den folgenden Widerspruch:

$$\begin{aligned} 0 < (Ax, x)_2 &= (Ae^i, e^i)_2 - 2\,\mathrm{sign}(a_{ij})(Ae^i, e^j)_2 + \mathrm{sign}(a_{ij})^2(Ae^j, e^j)_2 \\ &= a_{ii} - 2\,\mathrm{sign}(a_{ij})a_{ij} + a_{jj} = a_{ii} - 2|a_{ij}| + a_{jj} \leq 0. \end{aligned}$$

Dies vervollständigt den Beweis. Q.E.D.

Im Folgenden werden wir in Verbindung mit der Eigenschaft „positiv definit“ stets auch die Eigenschaft „hermitesch“ (bzw. „symmetrisch“ im Reellen) einer Matrix annehmen. Dies ist im Komplexen automatisch gegeben, im Rellen aber eine zusätzliche Bedingung. Wir werden später sehen, dass lineare Gleichungssysteme mit positiv definiten Koeffizientenmatrizen besonders günstige Lösbarkeitseigenschaften besitzen.

4.1.3 Fehleranalyse

Wir kommen nun zur Fehleranalyse für lineare Gleichungssysteme

$$Ax = b \tag{4.1.8}$$

mit regulärer Koeffizientenmatrix $A \in \mathbb{K}^{n\times n}$. Die Matrix A und der Vektor b seien mit Fehlern δA bzw. δb behaftet, so dass ein gestörtes System

$$\tilde{A}\tilde{x} = \tilde{b} \tag{4.1.9}$$

mit $\tilde{A} = A + \delta A$, $\tilde{b} = b + \delta b$ und $\tilde{x} = x + \delta x$ gelöst wird. Wir wollen den Fehler δx in Abhängigkeit von δA und δb abschätzen. Dazu sei im Folgenden $\|\cdot\|$ eine beliebige Vektornorm und entsprechend $\|\cdot\|$ die zugehörige natürliche Matrizennorm.

Hilfssatz 4.4: *Die Matrix $B \in \mathbb{K}^{n\times n}$ habe eine Norm $\|B\| < 1$. Dann ist die Matrix $I + B$ regulär, und es gilt*

$$\|(I + B)^{-1}\| \leq \frac{1}{1 - \|B\|}. \tag{4.1.10}$$

Beweis: Für alle $x \in \mathbb{K}^n$ gilt

$$\|(I + B)x\| \geq \|x\| - \|Bx\| \geq (1 - \|B\|)\|x\|.$$

Wegen $1 - \|B\| > 0$ ist also $I + B$ injektiv und folglich regulär. Mit der Abschätzung

$$\begin{aligned} 1 &= \|I\| = \|(I+B)(I+B)^{-1}\| = \|(I+B)^{-1} + B(I+B)^{-1}\| \\ &\geq \|(I+B)^{-1}\| - \|B\|\,\|(I+B)^{-1}\| = \|(I+B)^{-1}\|\,(1 - \|B\|) > 0 \end{aligned}$$

erhält man die behauptete Ungleichung. Q.E.D.

Nach diesen Vorbereitungen können wir den folgenden allgemeinen Störungssatz für lineare Gleichungssysteme beweisen:

Satz 4.1 (Störungssatz): *Die Matrix $A \in \mathbb{K}^{n\times n}$ sei regulär, und es sei*

$$\|\delta A\| < \frac{1}{\|A^{-1}\|}. \tag{4.1.11}$$

Dann ist die gestörte Matrix $\tilde{A} = A + \delta A$ ebenfalls regulär, und für den relativen Fehler der Lösung gilt:

$$\frac{\|\delta x\|}{\|x\|} \;\leq\; \frac{\mathrm{cond}(A)}{1 - \mathrm{cond}(A)\|\delta A\|/\|A\|} \left\{ \frac{\|\delta b\|}{\|b\|} + \frac{\|\delta A\|}{\|A\|} \right\}, \tag{4.1.12}$$

mit der sog. „Konditionszahl“ $\mathrm{cond}(A) := \|A\|\,\|A^{-1}\|$ von A.

Beweis: Aufgrund der Voraussetzung ist

$$\|A^{-1}\delta A\| \leq \|A^{-1}\|\,\|\delta A\| < 1,$$

so dass auch $A + \delta A = A[I + A^{-1}\delta A]$ nach Hilfssatz 4.4 regulär ist. Aus

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

folgt dann für $\delta x = \tilde{x} - x$

$$(A + \delta A)\delta x = \delta b - \delta A x,$$

und damit unter Verwendung von Hilfssatz 4.4

$$\begin{aligned} \|\delta x\| &\leq \|(A+\delta A)^{-1}\| \left\{ \|\delta b\| + \|\delta A\|\,\|x\| \right\} \\ &= \|\big(A(I + A^{-1}\delta A)\big)^{-1}\| \left\{ \|\delta b\| + \|\delta A\|\,\|x\| \right\} \\ &= \|(I + A^{-1}\delta A)^{-1}A^{-1}\| \left\{ \|\delta b\| + \|\delta A\|\,\|x\| \right\} \\ &\leq \|(I + A^{-1}\delta A)^{-1}\|\,\|A^{-1}\| \left\{ \|\delta b\| + \|\delta A\|\,\|x\| \right\} \\ &\leq \frac{\|A^{-1}\|}{1 - \|A^{-1}\delta A\|} \left\{ \|\delta b\| + \|\delta A\|\,\|x\| \right\} \\ &\leq \frac{\|A^{-1}\|\,\|A\|\,\|x\|}{1 - \|A^{-1}\|\,\|\delta A\|\,\|A\|\,\|A\|^{-1}} \left\{ \frac{\|\delta b\|}{\|A\|\,\|x\|} + \frac{\|\delta A\|}{\|A\|} \right\}. \end{aligned}$$

Wegen $\|b\| = \|Ax\| \leq \|A\|\,\|x\|$ folgt schließlich

$$\|\delta x\| \leq \frac{\mathrm{cond(A)}}{1 - \mathrm{cond(A)}\|\delta \mathrm{A}\|\|\mathrm{A}\|^{-1}} \left\{ \frac{\|\delta b\|}{\|b\|} + \frac{\|\delta A\|}{\|A\|} \right\} \|x\|,$$

was zu zeigen war. Q.E.D.

Die Konditionszahl $\mathrm{cond(A)}$ hängt offenbar von der bei ihrer Definition zugrundegelegten Vektornorm ab. Meistens verwendet man die Maximumnorm $\|\cdot\|_\infty$ oder die euklidische Norm $\|\cdot\|_2$. Im ersten Fall ist

$$\mathrm{cond}_\infty(\mathrm{A}) := \|\mathrm{A}\|_\infty \|\mathrm{A}^{-1}\|_\infty$$

mit der maximalen Zeilensumme $\|\cdot\|_\infty$. Speziell für „hermitesche“ Matrizen gilt nach Hilfssatz 4.3

$$\mathrm{cond}_2(\mathrm{A}) := \|\mathrm{A}\|_2 \|\mathrm{A}^{-1}\|_2 = \frac{|\lambda_{\max}|}{|\lambda_{\min}|} \tag{4.1.13}$$

mit den betragsmäßig größten bzw. kleinsten Eigenwerten $\lambda_{\max}$ und $\lambda_{\min}$ von A; die Größe $\mathrm{cond}_2(\mathrm{A})$ wird auch die „Spektralkonditionszahl“ von A genannt.

Ist $\mathrm{cond(A)}\|\delta \mathrm{A}\|\,\|\mathrm{A}\|^{-1} \ll 1$, so wird in Satz 4.1

$$\frac{\|\delta x\|}{\|x\|} \dot{\leq} \mathrm{cond(A)} \left\{ \frac{\|\delta \mathrm{b}\|}{\|\mathrm{b}\|} + \frac{\|\delta \mathrm{A}\|}{\|\mathrm{A}\|} \right\}, \tag{4.1.14}$$

d. h.: $\mathrm{cond(A)}$ ist in erster Näherung gerade der Verstärkungsfaktor, mit dem sich die relativen Fehler in A und b auf den in x auswirken. Diese Fehlerabschätzung erlaubt folgenden Schluss:

Regel 4.1: *Die Kondition von A sei $\mathrm{cond(A)} \sim 10^{\mathrm{s}}$. Sind dann die Elemente von A und b mit einem relativen Fehler der Art*

$$\frac{\|\delta A\|}{\|A\|} \sim 10^{-k}, \quad \frac{\|\delta b\|}{\|b\|} \sim 10^{-k}\ (k > s)$$

behaftet, so muss mit einem relativen Fehler im Ergebnis der Größenordnung

$$\frac{\|\delta x\|}{\|x\|} \sim 10^{s-k}$$

gerechnet werden, d. h.: Im Fall $\|\cdot\| = \|\cdot\|_\infty$ verliert man s Stellen Genauigkeit.

Beispiel 4.2: Wir betrachten die folgende Koeffizientenmatrix A:

$$A = \begin{bmatrix} 1.2969 & 0.8648 \\ 0.2161 & 0.1441 \end{bmatrix}, \quad A^{-1} = 10^8 \begin{bmatrix} 0.1441 & -0.8648 \\ -0.2161 & 1.2969 \end{bmatrix}$$

$$\|A\|_\infty = 2.1617\,, \quad \|A^{-1}\|_\infty = 1.513 \cdot 10^8 \Rightarrow \mathrm{cond(A)} \approx 3.3 \cdot 10^8.$$

Bei der Lösung des Gleichungssystems $Ax = b$ gehen also im ungünstigsten Fall 8 wesentliche Stellen an der Genauigkeit, mit der die Elemente a_{jk} und b_j gegeben sind, verloren. Dieses System ist sehr „schlecht konditioniert“.

Wir demonstrieren anhand der Spektralkondition, dass die Abschätzung in Satz 4.1 im wesentlichen scharf ist. Sei A eine positiv definite $n \times n$-Matrix mit kleinstem und größtem Eigenwert λ_1 bzw. λ_n sowie zugehörigen normierten Eigenvektoren w_1 bzw. w_n. Wir wählen

$$\delta A \equiv 0, \quad b \equiv w_n, \quad \delta b \equiv \varepsilon w_1 \, (\varepsilon \neq 0).$$

Dann haben die Gleichungen $Ax = b$ und $A\tilde{x} = b + \delta b$ die Lösungen

$$x = \frac{1}{\lambda_n}\, w_n\,, \quad \tilde{x} = \frac{1}{\lambda_n}\, w_n + \varepsilon\, \frac{1}{\lambda_1}\, w_1\,.$$

Folglich ist für $\delta x = \tilde{x} - x$

$$\frac{\|\delta x\|_2}{\|x\|_2} = \varepsilon\, \frac{\lambda_n}{\lambda_1}\, \frac{\|w_1\|_2}{\|w_n\|_2} = \mathrm{cond}_2(\mathrm{A})\, \frac{\|\delta \mathrm{b}\|_2}{\|\mathrm{b}\|_2}\,.$$

4.2 Gaußsches Eliminationsverfahren

Im Folgenden diskutieren wir „direkte“ Lösungsmethoden für (reelle) quadratische lineare Gleichungssysteme

$$Ax = b\,. \tag{4.2.15}$$

Besonders leicht lösbar sind gestaffelte Systeme, z. B. solche mit einer oberen Dreiecksmatrix $A = (a_{jk})$ als Koeffizientenmatrix

$$\begin{array}{rcrcrcl} a_{11}x_1 & + & a_{12}x_2 & + \ \ldots \ + & a_{1n}x_n & = & b_1 \\ & & a_{22}x_2 & + \ \ldots \ + & a_{2n}x_n & = & b_2 \\ & & & & \vdots & & \\ & & & & a_{nn}x_n & = & b_n \end{array}\,.$$

Im Falle $a_{jj} \neq 0$, $j = 1, \ldots, n$, erhält man die Lösung durch sog. „Rückwärtseinsetzen“:

$$x_n = \frac{b_n}{a_{nn}}\,, \quad j = n-1, \ldots, 1: \quad x_j = \frac{1}{a_{jj}}\,(b_j - \sum_{k=j+1}^{n} a_{jk}\, x_k)\,.$$

Dazu sind offensichtlich $\frac{1}{2}n^2 + O(n)$ *arithmetische Operationen* erforderlich (1 a.Op. := 1 Multiplikation (+ 1 Addition) oder 1 Division).

Das *klassische* direkte Verfahren zur Lösung allgemeiner (regulärer) Gleichungssysteme ist das Gaußsche Eliminationsverfahren (kurz „Gauß-Elimination“) . Dabei wird das gegebene System $Ax = b$ schrittweise in ein oberes Dreieckssystem $Rx = c$ umgeformt, welches dieselbe Lösung x besitzt und dann durch Rückwärtseinsetzen gelöst wird. Dazu stehen die folgenden elementaren Umformungen zur Verfügung:

- Vertauschung zweier Gleichungen,
- Addition des Vielfachen einer Gleichung zu einer anderen.

Die Vertauschung zweier Spalten von A ist ebenfalls zulässig, wenn die Unbekannten x_i entsprechend umnummeriert werden. In der praktischen Durchführung der Gauß-Elimination wendet man die elementaren Umformungen auf die zusammengesetzte Matrix $[A, b]$ an. Im Folgenden wird A als regulär angenommen. Zunächst setzt man $A^{(0)} \equiv A$, $b^{(0)} \equiv b$ und bestimmt $a_{r1}^{(0)} \neq 0$, $r \in \{1, \ldots, n\}$. (Solch ein Element existiert, da A sonst singulär wäre). Vertausche die 1-te und die r-te Zeile. Das Resultat sei $[\tilde{A}^{(0)}, \tilde{b}^{(0)}]$. Dann wird für $j = 2, \ldots, n$ das q_{j1}-fache der 1-ten Zeile von der j-ten Zeile abgezogen:

$$q_{j1} \equiv \tilde{a}_{j1}^{(0)} / \tilde{a}_{11}^{(0)} \; (= a_{r1}^{(0)} / a_{rr}^{(0)}), \quad a_{ji}^{(1)} := \tilde{a}_{ji}^{(0)} - q_{j1} \tilde{a}_{1i}^{(0)}, \quad b_j^{(1)} := \tilde{b}_j^{(0)} - q_{j1} \tilde{b}_1^{(0)}.$$

Das Resultat ist

$$[A^{(1)}, b^{(1)}] = \left[\begin{array}{cccc|c} \tilde{a}_{11}^{(0)} & \tilde{a}_{12}^{(0)} & \ldots & \tilde{a}_{1n}^{(0)} & \tilde{b}_1^{(0)} \\ 0 & a_{22}^{(1)} & \ldots & a_{2n}^{(1)} & b_2^{(1)} \\ \vdots & \vdots & & & \\ 0 & a_{n2}^{(1)} & \ldots & a_{nn}^{(1)} & b_n^{(1)} \end{array}\right].$$

Den Übergang $[A^{(0)}, b^{(0)}] \to [\tilde{A}^{(0)}, \tilde{b}^{(0)}] \to [A^{(1)}, b^{(1)}]$ kann man mit Hilfe von Matrizenmultiplikation beschreiben:

$$[\tilde{A}^{(0)}, \tilde{b}^{(0)}] = P_1 [A^{(0)}, b^{(0)}], \quad [A^{(1)}, b^{(1)}] \; = \; G_1 [\tilde{A}^{(0)}, \tilde{b}^{(0)}],$$

wobei P_1 eine „Permutationsmatrix“ und G_1 eine sog. „Frobenius-Matrix“ der folgenden Gestalt sind:

$$P_1 = \begin{array}{c} \begin{array}{ccccccccc} 1 & & & & \mathrm{r} & & & & \end{array} \\ \left[\begin{array}{ccccccccc} 0 & & \cdots & & 1 & & & & \\ & 1 & & & & & & & \\ \vdots & & \ddots & & \vdots & & & & \\ & & & 1 & & & & & \\ 1 & & \cdots & & 0 & & & & \\ & & & & & 1 & & & \\ & & & & & & \ddots & & \\ & & & & & & & 1 & \end{array}\right] \begin{array}{l} 1 \\ \\ \\ \\ \mathrm{r} \\ \\ \\ \\ \end{array} \end{array} \qquad G_1 = \begin{array}{c} \begin{array}{ccccc} & 1 & & & \\ & & & & \end{array} \\ \left[\begin{array}{ccccc} 1 & & & & \\ -q_{21} & 1 & & & \\ \vdots & & \ddots & & \\ -q_{n1} & & & 1 & \end{array}\right] \begin{array}{l} \\ 1 \\ \\ \\ \end{array} \end{array}$$

Beide Matrizen P_1 und G_1 sind regulär (Determinante $= \pm 1$), und es gilt

$$P_1^{-1} = P_1\,, \qquad G_1^{-1} = \begin{bmatrix} 1 & & & \\ q_{21} & 1 & & \\ \vdots & & \ddots & \\ q_{n1} & & & 1 \end{bmatrix}.$$

Die Gleichungssysteme $Ax = b$ und $A^{(1)}x = b^{(1)}$ haben offenbar dieselbe Lösung:

$$Ax = b \quad \Longleftrightarrow \quad A^{(1)}x = G_1P_1Ax = G_1P_1b = b^{(1)}\,.$$

Definition 4.6: *Das Element $a_{r1} = \tilde{a}_{11}^{(0)}$ heißt „Pivotelement“ und der ganze Teilschritte seiner Bestimmung „Pivotsuche“. Aus Gründen der numerischen Stabilität trifft man gewöhnlich die Wahl*

$$|a_{r1}| = \max_{1\le j\le n} |a_{j1}|\,. \tag{4.2.16}$$

Der ganze Prozeß inkl. Zeilenvertauschung wird dann „Spaltenpivotierung“ genannt. Sind die Elemente der Matrix A von sehr unterschiedlicher Größenordnung, so empfiehlt es sich, sog. „Totalpivotierung“ vorzunehmen. Diese besteht aus der Wahl

$$|a_{rs}| = \max_{1\le j,k\le n} |a_{jk}| \tag{4.2.17}$$

und anschließender Vertauschung der 1-ten mit der r-ten Zeile und der 1-ten mit der s-ten Spalte. Entsprechend der Spaltenvertauschung müssen dann die Unbekannten x_k umnumeriert werden. Bei großen Gleichungssystemen ist die totale Pivotierung meist zu aufwendig, so dass man sich mit der Spaltenpivotierung begnügt.

Die im 1. Schritt erzeugte Matrix $A^{(1)}$ ist wieder regulär. Dasselbe gilt auch für die um die erste Zeile und Spalte reduzierte Teilmatrix, so dass auf sie der Eliminationsprozess analog zu Schritt 1 angewendet werden kann. Durch Weiterführung dieses Eliminationsprozesses erhält man in $n-1$ Schritten eine Kette von Matrizen

$$[A,b] \to [A^{(1)},b^{(1)}] \to \ldots \to [A^{(n-1)},b^{(n-1)}] =: [R,c]\,,$$

wobei

$$[A^{(i)},b^{(i)}] = G_iP_i[A^{(i-1)},b^{(i-1)}]\,, \quad [A^{(0)},b^{(0)}] := [A,b]\,,$$

mit Permutationsmatrizen P_i und (regulären) Frobenius-Matrizen G_i der folgenden Form:

$$
P_i = \begin{array}{c} \\ \\ \\ \\ \\ \\ \\ \\ \\ \\ \\ \end{array}
\begin{array}{c}
\begin{array}{ccccccccccc} & & & & & & \mathrm{i} & & & \mathrm{r} & \end{array} \\
\left[\begin{array}{ccccccccccc}
1 & & & & & & & & & & \\
& \ddots & & & & & & & & & \\
& & 1 & & & & & & & & \\
& & & 0 & \cdots & & 1 & & & & \\
& & & & 1 & & & & & & \\
& & & \vdots & & \ddots & & \vdots & & & \\
& & & & & & 1 & & & & \\
& & & 1 & \cdots & & 0 & & & & \\
& & & & & & & 1 & & & \\
& & & & & & & & \ddots & & \\
& & & & & & & & & 1 &
\end{array}\right]
\end{array}
\begin{array}{c} \\ \\ \\ \\ \mathrm{i} \\ \\ \\ \\ \mathrm{r} \\ \\ \\ \end{array}
$$

$$
G_i = \begin{array}{c}
\begin{array}{ccccc} & & \mathrm{i} & & \end{array} \\
\left[\begin{array}{ccccc}
1 & & & & \\
& \ddots & & & \\
& & 1 & & \\
& & -q_{i+1,i} & 1 & \\
& & \vdots & & \ddots \\
& & -q_{ni} & & 1
\end{array}\right]
\end{array}
\begin{array}{c} \\ \\ \\ \mathrm{i} \\ \\ \\ \end{array}
$$

Das Endresultat

$$[R, c] = G_{n-1}P_{n-1}\ldots G_1P_1[A, b] \tag{4.2.18}$$

entspricht einem oberen Dreieckssystem $Rx = c$, welches dieselbe Lösung wie das Ausgangssystem $Ax = b$ besitzt.

Im i-ten Eliminationsschritt $[A^{(i-1)}, b^{(i-1)}] \to [A^{(i)}, b^{(i)}]$ werden in der i-ten Spalte die Elemente unterhalb der Diagonalen annuliert. Den frei werdenden Platz benutzt man zur Abspeicherung der wesentlichen Elemente $q_{i+1,i}, \ldots, q_{n,i}$ der Frobenius-Matrizen G_i^{-1} $(i = 1, \ldots, n-1)$. Da im i-ten Eliminationsschritt die vorausgehenden Zeilen 1 bis i nicht verändert werden, arbeitet man also mit Matrizen der folgenden Form:

$$
\left[\begin{array}{ccccccc|c}
r_{11} & r_{12} & \cdots & r_{1i} & r_{1,i+1} & \cdots & r_{1n} & c_1 \\
\lambda_{21} & r_{22} & \cdots & r_{2i} & r_{2,i+1} & \cdots & r_{2n} & c_2 \\
\lambda_{31} & \lambda_{32} & & r_{3i} & r_{3,i+1} & \cdots & r_{3n} & c_3 \\
\vdots & \vdots & \ddots & \vdots & \vdots & & \vdots & \vdots \\
\lambda_{i1} & \lambda_{i2} & & r_{ii} & r_{i,i+1} & \cdots & r_{in} & c_i \\
\hline
\lambda_{i+1,1} & \lambda_{i+2,2} & & \lambda_{i+1,i} & a_{i+1,i+1}^{(i)} & \cdots & a_{i+1,n}^{(i)} & b_{i+1}^{(i)} \\
\vdots & \vdots & & \vdots & \vdots & & \vdots & \vdots \\
\lambda_{n,1} & \lambda_{n,2} & \cdots & \lambda_{n,i} & a_{n,i+1}^{(i)} & \cdots & a_{n,n}^{(i)} & b_n^{(i)}
\end{array}\right]
$$

Dabei sind die Subdiagonalelemente $\lambda_{k+1,k}, \ldots, \lambda_{nk}$ der k-ten Spalte Permutationen der Elemente $q_{k+1,k}, \ldots, q_{nk}$ von G_k^{-1}, da die Zeilenvertauschungen (nur diese!) an der ge-

samten Matrix vorgenommen werden. Als Endresultat erhält man eine Matrix

$$\left[\begin{array}{cccc|c} r_{11} & & \cdots & r_{1n} & c_1 \\ l_{21} & r_{22} & & r_{2n} & c_2 \\ \vdots & \ddots & \ddots & \vdots & \vdots \\ l_{n1} & \cdots & l_{n,n-1} & r_{nn} & c_n \end{array}\right] .$$

Satz 4.2 (LR-Zerlegung): *Die Matrizen*

$$L = \begin{bmatrix} 1 & & & 0 \\ l_{21} & 1 & & \\ \vdots & \ddots & \ddots & \\ l_{n1} & \cdots & l_{n,n-1} & 1 \end{bmatrix}, \quad R = \begin{bmatrix} r_{11} & r_{12} & \cdots & r_{1n} \\ & r_{22} & \cdots & r_{2n} \\ & & \ddots & \vdots \\ 0 & & & r_{nn} \end{bmatrix}$$

bilden eine sog. „LR-Zerlegung“ der Matrix PA:

$$PA = LR, \qquad P = P_{n-1} \cdots P_1 . \tag{4.2.19}$$

Diese Zerlegung ist im Falle $P = I$ eindeutig bestimmt.

Beweis: Wir führen den Beweis für den Fall, dass keine Pivotierung erforderlich ist, d. h.: $P_i = I$. Dann ist $R = G_{n-1} \cdots G_1 A$ bzw.

$$G_1^{-1} \cdots G_{n-1}^{-1} R = A .$$

Wegen $L = G_1^{-1} \cdots G_{n-1}^{-1}$ folgt die Behauptung. Zum Nachweis der Eindeutigkeit, seien nun $A = L_1 R_1 = L_2 R_2$ zwei LR-Zerlegungen. Dann ist

$$L_2^{-1} L_1 = R_2 R_1^{-1} = I ,$$

da $L_2^{-1} L_1$ untere Dreiecksmatrix mit Einsen auf der Hauptdiagonalen und $R_2 R_1^{-1}$ obere Dreiecksmatrix ist. Folglich ist $L_1 = L_2$ und $R_1 = R_2$. Q.E.D.

Lemma 4.2: *Die zur Lösung eines $n \times n$ Gleichungssystems $Ax = b$ mit Hilfe der Gauß-Elimination erforderliche Anzahl von arithmetischen Operationen („a. Op.“) ist*

$$N_{\text{Gauß}}(n) = \tfrac{1}{3} n^3 + O(n^2).$$

Dasselbe gilt für die Bestimmung der Dreieckszerlegung $PA = LR$.

Beweis: Der k-te Eliminationsschritt

$$a_{ij}^{(k)} = a_{ij}^{(k-1)} - \frac{a_{ik}^{(k-1)}}{a_{kk}^{(k-1)}} a_{kj}^{(k-1)}, \quad b_i^{(k)} = b_i^{(k-1)} - \frac{a_{ik}^{(k-1)}}{a_{kk}^{(k-1)}} b_k^{(k-1)}, \quad i, j = k, \ldots, n,$$

erfordert $n-k$ Divisionen sowie $(n-k)+(n-k)^2$ Multiplikationen und Additionen; also zusammen

$$\sum_{k=1}^{n-1} k^2 + O(n^2) = \frac{1}{3}n^3 + O(n^2) \quad \text{a. Op.}$$

für die $n-1$ Schritte der Vorwärtselimination. Damit werden alle Elemente der Zerlegungsmatrizen L und R bestimmt. Q.E.D.

Beispiel 4.3: Mit $\boxed{\cdot}$ wird das Pivotelement markiert.

$$\begin{bmatrix} 3 & 1 & 6 \\ 2 & 1 & 3 \\ 1 & 1 & 1 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ x_3 \end{bmatrix} = \begin{bmatrix} 2 \\ 7 \\ 4 \end{bmatrix} \quad \rightarrow$$

Pivotierung

$\boxed{3}$	1	6	2
2	1	3	7
1	1	1	4

Elimination

3	1	6	2
2/3	1/3	−1	17/3
1/3	$\boxed{2/3}$	−1	10/3

$\rightarrow$

Pivotierung

3	1	6	2
1/3	2/3	−1	10/3
2/3	1/3	−1	17/3

Elimination

3	1	6	2
1/3	2/3	−1	10/3
2/3	1/2	−1/2	4

$\rightarrow$

$$\begin{aligned} x_3 &= -8 \\ x_2 &= \tfrac{3}{2}(\tfrac{10}{3} - x_3) = -7 \\ x_1 &= \tfrac{1}{3}(2 - x_2 - 6x_3) = 19\,. \end{aligned}$$

LR-Zerlegung:

$$P_1 = I\,, \quad P_2 = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 0 & 1 \\ 0 & 1 & 0 \end{bmatrix}\,,$$

$$PA = \begin{bmatrix} 3 & 1 & 6 \\ 1 & 1 & 1 \\ 2 & 1 & 3 \end{bmatrix} = LR = \begin{bmatrix} 1 & 0 & 0 \\ 1/3 & 1 & 0 \\ 2/3 & 1/2 & 1 \end{bmatrix} \begin{bmatrix} 3 & 1 & 6 \\ 0 & 2/3 & -1 \\ 0 & 0 & -1/2 \end{bmatrix}\,.$$

Beispiel 4.4: Zur Demonstration der Bedeutung der Pivotierung beim Gaußschen Eliminationsverfahren betrachten wir das folgende Gleichungssystem

$$\begin{bmatrix} 10^{-4} & 1 \\ 1 & 1 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} = \begin{bmatrix} 1 \\ 2 \end{bmatrix} \tag{4.2.20}$$

mit der exakten Lösung $x_1 = 1.00010001$, $x_2 = 0.99989999$. Bei 3-stelliger Gleitpunktrechnung mit korrekter Rundung erhält man:

a) ohne Pivotierung:

x_1	x_2	
$0.1 \cdot 10^{-3}$	$0.1 \cdot 10^{1}$	$0.1 \cdot 10^{1}$
0	$-0.1 \cdot 10^{5}$	$-0.1 \cdot 10^{5}$
$x_2 = 1\,,$	$x_1 = 0$	

b) mit Pivotierung:

x_1	x_2	
$0.1 \cdot 10^{1}$	$0.1 \cdot 10^{1}$	$0.2 \cdot 10^{1}$
0	$0.1 \cdot 10^{1}$	$0.1 \cdot 10^{1}$
$x_2 = 1\,,$	$x_1 = 1$	

Beispiel 4.5: Der positive Effekt der Spaltenpivotierung ist allerdings nur dann gesichert, wenn die (betragsmäßigen) Zeilensummen der Matrix in etwa gleich groß sind. Als Beispiel betrachte man das Gleichungssystem

$$\begin{bmatrix} 2 & 20000 \\ 1 & 1 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} = \begin{bmatrix} 20000 \\ 2 \end{bmatrix},$$

welches aus (4.2.20) durch Multiplikation der ersten Zeile mit dem Faktor 20000 hervorgeht. Da nun in der ersten Spalte das betragsgrößte Element in der Diagonalen steht, liefert der Gauß-Algorithmus mit und ohne Spaltenpivotierung dasselbe inakzeptable Resultat $(x_1, x_2)^T = (0, 1)^T$. Man führt daher vor der Rechnung eine sog. „Äquilibrierung“ durch, das heißt, man multipliziert mit einer Diagonalmatrix D

$$Ax = b \quad \rightarrow \quad DAx = Db\,, \qquad d_i = \Big(\sum\nolimits_{j=1}^{n} |a_{ij}|\Big)^{-1},$$

die alle Zeilensummen der Matrix auf 1 transformiert. Eine verbesserte Stabilisierung im Fall stark unterschiedlicher Größenordnung der Matrixeinträge ist die „totale“ Pivotierung. Vor der Durchführung wird hier eine Äquilibrierung (zeilenweise und spaltenweise) vorgenommen.

4.2.1 Konditionierung der Gauß-Elimination

Wir diskutieren nun noch die Konditionierung des Lösens eines linearen Gleichungssystems $Ax = b$ mit Hilfe der Gauß-Elimination. Die (reguläre) Matrix A besitze mit Spaltenpivotierung eine LR-Zerlegung der Form $PA = LR$. Dann gilt

$$R = L^{-1}PA, \quad R^{-1} = (PA)^{-1}L.$$

Wegen der Spaltenpivotierung sind die Elemente der Dreiecksmatrizen L und L^{-1} alle kleiner gleich eins, und es gilt somit

$$\mathrm{cond}_\infty(L) = \|L\|_\infty \|L^{-1}\|_\infty \leq n^2.$$

Folglich ist

$$\begin{aligned}\mathrm{cond}_\infty(R) &= \|R\|_\infty\|R^{-1}\|_\infty = \|L^{-1}PA\|_\infty\|(PA)^{-1}L\|_\infty \\ &\le \|L^{-1}\|_\infty\|PA\|_\infty\|(PA)^{-1}\|_\infty\|L\|_\infty \le n^2\,\mathrm{cond}_\infty(PA).\end{aligned}$$

Nach dem allgemeinen Störungssatz gilt dann für die Lösung des Systeme $LRx = Pb$:

$$\frac{\|\delta x\|_\infty}{\|x\|_\infty} \le \mathrm{cond}_\infty(L)\mathrm{cond}_\infty(R)\frac{\|\delta Pb\|_\infty}{\|Pb\|_\infty} \le n^4\,\mathrm{cond}_\infty(PA)\frac{\|\delta Pb\|_\infty}{\|Pb\|_\infty}.$$

Die Konditionierung des Ausgangssystems wird also durch die LR-Zerlegung im schlimmsten Fall mit n^4 verschlechtert. Dies ist aber eine extrem pessimistische Abschätzung und kann wesentlich verbessert werden.

Wir geben zum Abschluß noch ein Resultat von Wilkinson[2] an, das die Fortpflanzung des Rundungsfehlers im Verlaufe der Gauß-Elimination beschreibt.

Satz 4.3 (Rundungsfehlereinfluss): *Die Matrix $A \in \mathbb{R}^{n\times n}$ sei regulär, und das Gleichungssystem $Ax = b$ werde mit Gauß-Elimination mit Spaltenpivotierung gelöst. Dann ist die unter dem Einfluß von Rundungsfehlern tatsächlich berechnete Lösung $x + \delta x$ exakte Lösung eines gestörten Systems $(A + \delta A)(x + \delta x) = b$, wobei (*eps = *Maschinengenauigkeit)*

$$\frac{\|\delta A\|_\infty}{\|A\|_\infty} \le 1.01 \cdot 2^{n-1}(n^3 + 2n^2)\,\mathrm{eps}. \tag{4.2.21}$$

In Verbindung mit der Fehlerabschätzung von Satz 4.1 ergibt dieses Resultat die folgende Abschätzung für den Rundungsfehlereinfluss:

$$\frac{\|\delta x\|_\infty}{\|x\|_\infty} \le \frac{\mathrm{cond(A)}}{1 - \mathrm{cond(A)}\|\delta\mathrm{A}\|_\infty/\|\mathrm{A}\|_\infty}\left\{1.01 \cdot 2^{n-1}(n^3 + 2n^2)\,\mathrm{eps}\right\}. \tag{4.2.22}$$

Diese Abschätzung ist, wie die Praxis zeigt, viel zu pessimistisch, da sie am ungünstigsten Fall ausgerichtet ist und keine Rundungsfehlerauslöschungen berücksichtigt. Zur Erfassung der letzteren wäre eine statistische Theorie erforderlich. Außerdem gilt die obige Abschätzung allgemein für „vollbesetzte“ Matrizen. Für „dünnbesetzte“ Matrizen sind wesentlich günstigere Resultate zu erwarten. Insgesamt sieht man, dass der Gaußsche Eliminationsprozeß (in Abhängigkeit von der Dimension n) ein gutartiger numerischer Algorithmus ist, d. h.: Der Rundungsfehlereinfluss kann abgeschätzt werden allein in Abhängigkeit von der Kondition $\mathrm{cond(A)}$, die ja auch die Konditionierung der numerischen Aufgabe selber beschreibt.

[2]James Hardy Wilkinson (1919–1986): Englischer Numeriker und früher Informatiker; arbeitete seit 1946 am National Physical Laboratory in London mit Alan Turing an der Entwicklung der ersten digitalen Computer; fundamentale Beiträge zur numerischen linearen Algebra, insbes. zur Rundungsfehleranalyse; Mitbegründer der NAG Library (1970).

4.2.2 Nachiteration

Wir diskutieren nun noch einige Varianten und weitere Anwendungsmöglichkeiten der Gauß-Elimination. Die Gauß-Elimination überführt ein Gleichungssystem $Ax = b$ in ein oberes Dreieckssystem $Rx = c$, aus dem sich die Lösung x durch einfaches Rückwärtsauflösen berechnen lässt. Nach Satz 4.2 ist dieser Prozess gleichbedeutend mit der Erstellung einer Dreieckszerlegung $PA = LR$ und der anschließenden Lösung der beiden gestaffelten Systeme

$$Ly = Pb\,, \quad Rx = y\,. \tag{4.2.23}$$

Diese Variante des Gauß-Algorithmus ist insbesondere dann vorzuziehen, wenn dasselbe Gleichungssystem nacheinander für verschiedene rechte Seiten b gelöst werden soll. Aufgrund des unvermeidlichen Rundungsfehlers erhält man in der Praxis nur eine fehlerhafte LR-Zerlegung

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

und damit nur eine Näherungslösung x^0 mit dem (exakten) „Defekt“

$$\hat{d}^0 := Ax^0 - b \neq 0\,.$$

Unter Verwendung der bereits erstellten Dreieckszerlegung $\tilde{L}\tilde{R} \sim PA$ löst man nun (näherungsweise) die sog. „Defektgleichung“

$$Ak = \hat{d}^0, \qquad \tilde{L}\tilde{R}k^1 = \hat{d}^0, \tag{4.2.24}$$

und erhält daraus eine Korrektur k^1 für x^0 und setzen damit $x^1 := x^0 - k^1$. Hätte man die Defektgleichung exakt gelöst, d. h. $k^1 \equiv k$, so wäre

$$Ax^1 = Ax^0 - Ak = Ax^0 - \hat{d}^0 = Ax^0 - Ax^0 + b = b\,,$$

d. h.: $x^1 = x$ wäre die exakte Lösung des Systems $Ax = b$. I. Allg. wird x^1 auch bei fehlerhafter Lösung der Defektgleichung eine bessere Näherung zu x als x^0 sein. Dazu ist es jedoch erforderlich, den Defekt d mit *erhöhter* Genauigkeit zu berechnen. Dies wird durch die folgende Fehleranalyse belegt (der Einfachheit halber sei $P = I$):

Wir nehmen an, dass sich der relative Fehler bei der LR-Zerlegung der Matrix A durch eine kleine Zahl ε beschränken lässt. Nach dem allgemeinen Störungssatz 4.1 gilt dann die Abschätzung

$$\frac{\|x^0 - x\|}{\|x\|} \leq \frac{\mathrm{cond(A)}}{1 - \mathrm{cond(A)}\frac{\|\mathrm{A}-\tilde{\mathrm{L}}\tilde{\mathrm{R}}\|}{\|\mathrm{A}\|}} \underbrace{\frac{\|A - \tilde{L}\tilde{R}\|}{\|A\|}}_{\sim \varepsilon}\,.$$

Der Verlust von Stellen entspricht der Größe von $\mathrm{cond(A)}$. Zusätzlich auftretende Rundungsfehler werden vernachlässigt. Den exakten Defekt $\hat{d}^0$ ersetzen wir durch den Ausdruck $d^0 := \tilde{A}x^0 - b$, wobei $\tilde{A}$ eine genauere Approximation für A ist,

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

Nach Konstruktion gilt

$$\begin{aligned} x^1 = x^0 - k^1 &= x^0 - (\tilde{L}\tilde{R})^{-1}[\tilde{A}x^0 - b] \\ &= x^0 + (\tilde{L}\tilde{R})^{-1}[Ax^0 - Ax + (\tilde{A} - A)\,x^0], \end{aligned}$$

und daher

$$\begin{aligned} x^1 - x &= x^0 - x - (\tilde{L}\tilde{R})^{-1}A(x^0 - x) + (\tilde{L}\tilde{R})^{-1}(\tilde{A} - A)\,x^0 \\ &= (\tilde{L}\tilde{R})^{-1}[\tilde{L}\tilde{R} - A](x^0 - x) + (\tilde{L}\tilde{R})^{-1}(\tilde{A} - A)\,x^0. \end{aligned}$$

Wegen

$$\tilde{L}\tilde{R} = A - A + \tilde{L}\tilde{R} = A\big(I - A^{-1}(A - \tilde{L}\tilde{R})\big)$$

folgt mit Hilfssatz 4.4:

$$\begin{aligned} \|(\tilde{L}\tilde{R})^{-1}\| &\leq \|A^{-1}\|\,\|\,[I - A^{-1}(A - \tilde{L}\tilde{R})]^{-1}\| \\ &\leq \frac{\|A^{-1}\|}{1 - \|A^{-1}(A - \tilde{L}\tilde{R})\|} \leq \frac{\|A^{-1}\|}{1 - \|A^{-1}\|\,\|A - \tilde{L}\tilde{R}\|} = \frac{\|A^{-1}\|}{1 - \mathrm{cond}(A)\,\frac{\|A - \tilde{L}\tilde{R}\|}{\|A\|}}. \end{aligned}$$

Dies impliziert schließlich

$$\frac{\|x^1 - x\|}{\|x\|} \;\sim\; \mathrm{cond}(A)\,\Big[\underbrace{\frac{\|A - \tilde{L}\tilde{R}\|}{\|A\|}}_{\sim\,\varepsilon}\;\underbrace{\frac{\|x^0 - x\|}{\|x\|}}_{\sim\,\mathrm{cond(A)}\varepsilon} + \underbrace{\frac{\|\tilde{A} - A\|}{\|A\|}}_{\sim\,\tilde{\varepsilon}}\;\frac{\|x^0\|}{\|x\|}\Big]\,.$$

Diese Korrektur der Lösung kann natürlich iteriert werden, in dem die jeweils neuen Näherungen x^i wieder in die Defektgleichung eingesetzt werden. Diesen Prozess nennt man „Nachiteration“; in der Praxis wird der Fehler in x schon durch wenige Korrekturschritte (meist $2-3$) auf die Größenordnung der Genauigkeit der Defektauswertung gedrückt, d. h.: $\|x^3 - x\|/\|x\| \sim \tilde{\varepsilon}$.

Beispiel 4.6: Das Gleichungssystem

$$\begin{bmatrix} 1.05 & 1,02 \\ 1.04 & 1,02 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} = \begin{bmatrix} 1 \\ 2 \end{bmatrix}$$

hat die exakte Lösung $x = (-100,\, 103.921\ldots)^T$. Die Gauß-Elimination ergibt bei Verwendung 3-stelliger Gleitpunktarithmetik (mit korrekter Rundung) die genäherten Zerlegungsmatrizen

$$\tilde{L} = \begin{bmatrix} 1 & 0 \\ 0.990 & 1 \end{bmatrix}, \qquad \tilde{R} = \begin{bmatrix} 1.05 & 1.02 \\ 0 & 0.01 \end{bmatrix},$$

$$\tilde{L}\tilde{R} - A = \begin{bmatrix} 0 & 0 \\ 5 \cdot 10^{-4} & 2 \cdot 10^{-4} \end{bmatrix} \qquad \text{(im Rahmen der Maschinengenauigkeit korrekt)}\,.$$

Die damit bestimmte „Lösung“ $x^0 = (-97, 1.101)^T$ hat den Defekt

$$d^0 = Ax^0 - b = \begin{cases} (0,0)^T & \text{3-stellige Rechnung} \\ -(0,065, 0,035)^T & \text{6-stellige Rechnung}\,. \end{cases}$$

Die approximative Korrekturgleichung

$$\begin{bmatrix} 1 & 0 \\ 0.990 & 1 \end{bmatrix} \begin{bmatrix} 1.05 & 1.02 \\ 0 & 0.01 \end{bmatrix} \begin{bmatrix} k_1^1 \\ k_2^1 \end{bmatrix} = \begin{bmatrix} -0.065 \\ -0.035 \end{bmatrix}$$

hat (3-stellige Rechnung) die Lösung $k^1 = (2.9, -102.899)^T$. Die durch Nachkorrektur verbesserte Lösung ist also

$$x^1 = x^0 - k^1 = (-99.9, 104)^T,$$

welche deutlich genauer ist als die erste Näherung x^0.

4.2.3 Determinantenberechnung

Für quadratische Matrizen gilt der Determinantensatz

$$\det(AB) = \det(A)\,\det(B)\,. \tag{4.2.25}$$

Für die durch Gauß-Elimination aus der gegebenen Matrix A gewonnene Dreiecksmatrix

$$R = G_{n-1}P_{n-1}\cdots G_1P_1A$$

folgt somit unter Beachtung von

$$\det(P_i^{-1}) = \det(P_i) = -1\,, \quad \det(G_i^{-1}) = 1\,,$$

die Beziehung

$$\det(A) = \det(P_1^{-1}G_1^{-1}\cdots P_{n-1}^{-1}G_{n-1}^{-1}R) = \pm\det(R) = \pm\prod_{j=1}^{n} r_{jj}\,. \tag{4.2.26}$$

Das Vorzeichen in $\det(A)$ ist $+/-$, je nachdem, ob eine gerade oder ungerade Anzahl von Zeilenvertauschungen vorgenommen wurde. Lässt sich im Verlaufe des Eliminationsprozesses einmal in einer Spalte kein von Null verschiedenes Pivotelement finden, so ist die Matrix A singulär und folglich $\det(A) = 0$. (Man beachte, dass bei Rechnung in Gleitpunktarithmetik aufgrund des Rundungsfehlers durchaus auch im Falle $\det(A) = 0$ der tatsächlich berechnete Wert $\neq 0$ sein kann!)

4.2.4 Rangbestimmung

Ist die Elimination durchführbar, d. h. lässt sich immer ein Pivotelement $\neq 0$ finden, und ist schließlich auch das letzte Diagonalelement $a_{n,n}^{(n-1)} \neq 0$, so ist $\det(A) \neq 0$, d. h.

$$\mathrm{Rang}(A) = n$$

(dies natürlich nur bei Vernachlässigung der Rundungsfehler!). Gilt dagegen im i-ten Eliminationsschritt für alle Elemente in der i-ten Spalte

$$a_{ji}^{(i-1)} = 0\,, \quad j = i, \ldots, n\,,$$

so ist A singulär. In diesem Fall wird zur weiteren Rangberechnung Totalpivotierung vorgenommen:

$$|a_{rs}^{(i-1)}| = \max_{j,k=1,\ldots,n} |a_{jk}^{(i-1)}|\,.$$

(Zeilen- und Spaltenvertauschungen ändern $\mathrm{Rang}(A)$ nicht!) Gilt dann nach dem i-ten Eliminationsschritt

$$a_{jk}^{(i)} = 0\,, \quad j,k = i+1, \ldots, n\,,$$

so ist $\mathrm{Rang}(A) = i$. Dieser Prozess kann natürlich auch zur Rangbestimmung bei *nicht* quadratischen Matrizen verwendet werden.

4.2.5 Inversenberechnung (Gauß-Jordan-Algorithmus)

Grundsätzlich kann die Inverse A^{-1} einer regulären Matrix A wie folgt berechnet werden:

(i) Berechnung der LR-Zerlegung von PA;

(ii) Lösung der gestaffelten Systeme

$$Ly^{(i)} = Pe^{(i)},\ Rx^{(i)} = y^{(i)}, \quad i = 1, \ldots, n,$$

mit den kartesischen Basisvektoren $e^{(i)}$ des $\mathbb{R}^n$;

(iii) $A^{-1} = [x^{(1)}, \ldots, x^{(n)}]$.

Praktischer ist jedoch eine simultane Elimination (hier ohne Berechnung der Matrizen L und R), die direkt auf die Inverse führt (ohne Zeilenvertauschungen):

$$\begin{array}{c|ccc} & 1 & & 0 \\ A & & \ddots & \\ & 0 & & 1 \end{array} \quad \rightarrow \quad \begin{array}{c} \text{Vorwärtselimination} \\ \hline r_{11} \quad \cdots \quad r_{1n} \\ \qquad\quad \ddots \quad \vdots \\ \qquad\qquad\quad r_{nn} \end{array} \quad \begin{array}{c|ccc} & 1 & & 0 \\ \hline & & \ddots & \\ & * & & 1 \end{array}$$

$$\begin{array}{c|c|c} \multicolumn{3}{c}{\text{Rückwärtselimination}} \\ \hline r_{11} & 0 & \\ \ddots & & * \\ 0 & r_{nn} & \end{array} \quad \rightarrow \quad \begin{array}{c|c|c} \multicolumn{3}{c}{\text{Skalierung}} \\ \hline 1 & 0 & \\ \ddots & & A^{-1} \\ 0 & 1 & \end{array}$$

Beispiel 4.7: Es markiere $\boxed{\cdot}$ das Pivotelement.

$$A = \begin{bmatrix} 3 & 1 & 6 \\ 2 & 1 & 3 \\ 1 & 1 & 1 \end{bmatrix} \quad : \quad \begin{array}{c|c|c} \multicolumn{3}{c}{\text{Vorwärtselimination}} \\ \hline \boxed{3} & 1 & 6 & 1 & 0 & 0 \\ 2 & 1 & 3 & 0 & 1 & 0 \\ 1 & 1 & 1 & 0 & 0 & 1 \end{array} \quad \rightarrow$$

$$\rightarrow \quad \begin{array}{c|c|c} \multicolumn{3}{c}{\text{Zeilenvertauschung}} \\ \hline 3 & 1 & 6 & 1 & 0 & 0 \\ 0 & 1/3 & -1 & -2/3 & 1 & 0 \\ 0 & \boxed{2/3} & -1 & -1/3 & 0 & 1 \end{array} \quad \rightarrow \quad \begin{array}{c|c|c} \multicolumn{3}{c}{\text{Vorwärtselimination}} \\ \hline 3 & 1 & 6 & 1 & 0 & 0 \\ 0 & 2/3 & -1 & -1/3 & 0 & 1 \\ 0 & 1/3 & -1 & -2/3 & 1 & 0 \end{array} \quad \rightarrow$$

$$\rightarrow \quad \begin{array}{c|c|c} \multicolumn{3}{c}{\text{Rückwärtselimination}} \\ \hline 3 & 1 & 6 & 1 & 0 & 0 \\ 0 & 2/3 & -1 & -1/3 & 0 & 1 \\ 0 & 0 & -1/2 & -1/2 & 1 & -1/2 \end{array} \quad \rightarrow \quad \begin{array}{c|c|c} \multicolumn{3}{c}{\text{Rückwärtselimination}} \\ \hline 3 & 1 & 0 & -5 & 12 & -6 \\ 0 & 2/3 & 0 & 2/3 & -2 & 2 \\ 0 & 0 & -1/2 & -1/2 & 1 & -1/2 \end{array} \quad \rightarrow$$

$$\rightarrow \quad \begin{array}{c|c|c} \multicolumn{3}{c}{\text{Skalierung}} \\ \hline 3 & 0 & 0 & -6 & 15 & -9 \\ 0 & 2/3 & 0 & 2/3 & -2 & 2 \\ 0 & 0 & -1/2 & -1/2 & 1 & -1/2 \end{array} \quad \rightarrow \quad \begin{array}{c|c|c} & & & & & \\ \hline 1 & 0 & 0 & -2 & 5 & -3 \\ 0 & 1 & 0 & 1 & -3 & 3 \\ 0 & 0 & 1 & 1 & -2 & 1 \end{array}$$

$$\Rightarrow \quad A^{-1} = \begin{bmatrix} -2 & 5 & -3 \\ 1 & -3 & 3 \\ 1 & -2 & 1 \end{bmatrix} .$$

Eine alternative Methode zur Berechnung der Inversen einer Matrix ist das sog. „Austauschverfahren“ (auch „Gauß-Jordan[3]-Algorithmus“ genannt). Gegeben sei ein (nicht notwendig quadratisches) lineares Gleichungssystem

$$Ax = y \qquad \text{mit} \quad A \in \mathbb{R}^{m \times n}, \quad x \in \mathbb{R}^n, \quad y \in \mathbb{R}^m . \tag{4.2.27}$$

[3]Marie Ennemond Camille Jordan (1838–1922): Französischer Mathematiker; Professor in Paris; Beiträge zur Algebra, Gruppentheorie, Analysis und Topologie.

Eine Lösung wird berechnet durch sukzessiven Austausch der Komponenten von x gegen solche von y. Ist ein Matrixelement $a_{pq} \neq 0$, so wird die p-te Gleichung nach x_q aufgelöst:

$$x_q = -\frac{a_{p1}}{a_{pq}}x_1 - \ldots - \frac{a_{p,q-1}}{a_{pq}}x_{q-1} + \frac{1}{a_{pq}}y_p - \frac{a_{p,q+1}}{a_{pq}}x_{q+1} - \ldots - \frac{a_{pn}}{a_{pq}}x_n\,.$$

Durch Substitution von x_q in den anderen Gleichungen,

$$a_{j1}x_1 + \ldots + a_{j,q-1}x_{q-1} + a_{jq}\boxed{x_q} + a_{j,q+1}x_{q+1} + \ldots + a_{jn}x_n = y_j\,,$$

erhält man für $j = 1, \ldots, m$, $j \neq p$:

$$\left[a_{j1} - \frac{a_{jq}a_{p1}}{a_{pq}}\right] x_1 + \ldots + \left[a_{j,q-1} - \frac{a_{jq}a_{p,q-1}}{a_{pq}}\right] x_{q-1} + \frac{a_{jq}}{a_{pq}}y_p \quad +$$

$$+ \left[a_{j,q+1} - \frac{a_{jq}a_{p,q+1}}{a_{pq}}\right] x_{q+1} + \ldots + \left[a_{jn} - \frac{a_{jq}a_{pn}}{a_{pq}}\right] x_n \;=\; y_j\,.$$

Das Resultat ist ein zum Ausgangssystem äquivalentes System

$$\tilde{A}\begin{bmatrix} x_1 \\ \vdots \\ y_p \\ \vdots \\ x_n \end{bmatrix} = \begin{bmatrix} y_1 \\ \vdots \\ x_q \\ \vdots \\ y_m \end{bmatrix}\,, \tag{4.2.28}$$

wobei die Elemente der Matrix $\tilde{A}$ wie folgt bestimmt sind:

Pivotelement	:	$\tilde{a}_{pq} = 1/a_{pq}\,,$
Pivotzeile	:	$\tilde{a}_{pk} = a_{pk}/a_{pq}\,, \quad k = 1, \ldots, n\,, \quad k \neq q\,,$
Pivotspalte	:	$\tilde{a}_{jq} = a_{jq}/a_{pq}\,, \quad j = 1, \ldots, m\,, \quad j \neq p\,,$
sonstige	:	$\tilde{a}_{jk} = a_{jk} - a_{jq}a_{pk}/a_{pq}\,, \quad j = 1, \ldots, m,\ j \neq p,\ k = 1, \ldots, n\,,\ k \neq q.$

Gelingt es, durch Fortsetzung des Verfahrens alle Komponenten von x durch solche von y zu ersetzen, so hat man eine explizite Darstellung der Lösung von $y = A^{-1}x$. Im Fall $m = n$ ergibt sich so auch die Inverse A^{-1}, allerdings i. Allg. mit vertauschten Zeilen und Spalten. Bei der Festlegung des Pivotelementes empfiehlt es sich aus Stabilitätsgründen, unter allen in Frage kommenden a_{pq} jeweils eines von möglichst großem Betrag zu wählen.

Satz 4.4 (Gauß-Jordan-Algorithmus): *Es können genau* $r = \mathrm{Rang}(A)$ *Austauschschritte durchgeführt werden.*

Beweis: Das Verfahren breche nach r Austauschschritten ab. O.B.d.A. seien $x_1, \ldots, x_r$ gegen $y_1, \ldots, y_r$ ausgetauscht, so dass das resultierende System die folgende Gestalt hat:

$$\begin{array}{c} r\left\{\vphantom{\begin{array}{c}a\\b\end{array}}\right. \\ m-r\left\{\vphantom{\begin{array}{c}a\\b\end{array}}\right. \\ \end{array} \left[\begin{array}{c|c} * & * \\ \hline * & 0 \end{array}\right] \begin{array}{c} \underbrace{}_{r} \end{array} \begin{array}{c} \underbrace{}_{n-r} \end{array} \left[\begin{array}{c} y_1 \\ \vdots \\ y_r \\ x_{r+1} \\ \vdots \\ x_n \end{array}\right] = \left[\begin{array}{c} x_1 \\ \vdots \\ x_r \\ y_{r+1} \\ \vdots \\ y_m \end{array}\right] .$$

Wählt man nun $y_1 = \cdots = y_r = 0, \quad x_{r+1} = \lambda_1, \cdots, \quad x_n = \lambda_{n-r}$, so sind die $x_1, \cdots, x_r$ dadurch eindeutig bestimmt, und es folgt $y_{r+1} = \cdots = y_m = 0$. Für beliebige Werte $\lambda_1, \cdots, \lambda_{n-r}$ ist also

$$A \begin{bmatrix} x_1(\lambda_1, \cdots, \lambda_{n-r}) \\ \vdots \\ x_r(\lambda_1, \cdots, \lambda_{n-r}) \\ \lambda_1 \\ \vdots \\ \lambda_{n-r} \end{bmatrix} = \begin{bmatrix} 0 \\ \vdots \\ 0 \end{bmatrix} ,$$

d. h.: $\dim(\mathrm{Kern}(A)) \geq n - r$. Andererseits ist wegen der möglichen freien Wahl von $y_1, \cdots, y_r$ offenbar $\dim(\mathrm{Bild}(A)) \geq r$. Da $\dim(\mathrm{Bild}(A)) + \dim(\mathrm{Kern}(A)) = n$ ist, folgt $\mathrm{Rang}(A) = \dim(\mathrm{Bild}(A)) = r$. Q.E.D.

Für ein quadratisches Gleichungssystem mit regulärer Koeffizientenmatrix A ist das Gauß-Jordan-Verfahren zur Berechnung von A^{-1} also stets durchführbar.

Beispiel 4.8:

$$\begin{bmatrix} 1 & 2 & 1 \\ -3 & -5 & -1 \\ -7 & -12 & -2 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ x_3 \end{bmatrix} = \begin{bmatrix} y_1 \\ y_2 \\ y_3 \end{bmatrix}$$

Austauschschritte: Mit $\boxed{\cdot}$ wird das Pivotelement markiert.

x_1	x_2	x_3	
1	2	1	y_1
-3	-5	-1	y_2
-7	$\boxed{-12}$	-2	y_3

x_1	y_3	x_3	
$-1/6$	$-1/6$	$\boxed{2/3}$	y_1
$-1/12$	$5/12$	$-1/6$	y_2
$-7/12$	$-1/12$	$-1/6$	x_2

x_1	y_3	y_1	
$1/4$	$1/4$	$3/2$	x_3
$\boxed{-1/8}$	$3/8$	$-1/4$	y_2
$-5/8$	$-1/8$	$-1/4$	x_2

y_2	y_3	y_1	
-2	1	1	x_3
-8	3	-2	x_1
5	-2	1	x_2

Inverse: $\begin{bmatrix} -2 & -8 & 3 \\ 1 & 5 & -2 \\ 1 & -2 & 1 \end{bmatrix}$

Lemma 4.3: *Die zur Invertierung einer regulären $n \times n$-Matrix mit Hilfe der simultanen Elimination oder des Gauß-Jordan-Algorithmus erforderliche Anzahl von arithmetischen Operationen („a. Op.“) ist*

$$N_{\text{Gauß-Jordan}}(n) = n^3 + O(n^2).$$

Beweis: (i) Die $n-1$ Schritte der Vorwärtselimination an der Matrix A erfordert $\frac{1}{3}n^3 + O(n^2)$ a. Op.. Die simultane Bearbeitung der Spalten der Einheitsmatrix erfordert wegen der Dreicksgestalt von I zusätzliche $\frac{1}{6}n^3 + O(n^2)$ a. Op.. Die Rückwärtselimination zur Erstellung der Einheitsmatrix links erfordert schließlich nochmal

$$(n-1)n + (n-2)n + \ldots + n = \frac{n(n-1)}{2}n = \frac{1}{2}n^3 + O(n^2)$$

Multiplikationen und Additionen und nachfolgend n^2 Divisionen. Für die gesamte Berechnung der Inversen ergibt sich also:

$$\tfrac{1}{3}n^3 + \tfrac{1}{6}n^3 + \tfrac{1}{2}n^3 + O(n^2) = n^3 + O(n^2).$$

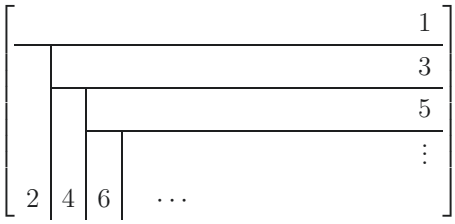
(ii) Beim Gauß-Jordan-Verfahren erfordert der k-te Austauschschritt $2n+1$ Divisionen in Pivotzeile und -spalte und $(n-1)^2$ Multiplikationen und Additionen für den Update der Restmatrix, also insgesamt $n^2 + O(n)$ a. Op. Zur Berechnung der Inversen sind n Austauschschritte durchzuführen, so dass sich ebenfalls ein Gesamtaufwand von $n^3 + O(n^2)$ a. Op. ergibt. Q.E.D.

4.2.6 Direkte LR-Zerlegung

Der Gauß-Algorithmus zur Berechnung der LR-Zerlegung $A = LR$ (falls sie existiert) kann auch in direkter Form geschrieben werden, bei der die Elemente l_{jk} von L und r_{jk} von R rekursiv berechnet werden. Die Gleichung $A = LR$ ergibt n^2 Bestimmungsgleichungen für die n^2 unbekannten Größen $r_{jk}\,,\ j \le k\,,\ l_{jk}\,,\ j > k\,(l_{jj} = 1)$:

$$a_{jk} = \sum_{i=1}^{\min(j,k)} l_{ji}\, r_{ik}. \tag{4.2.29}$$

Die Reihenfolge der Berechnung von l_{jk}, r_{jk} ist zunächst noch offen. Beim sog. „Algorithmus von Crout[4] “ wird die Matrix $A = LR$ wie folgt parkettiert:



Die einzelnen Schritte des Algorithmus sind dann $(l_{ii} \equiv 1)$:

$$
\begin{aligned}
&k = 1, \cdots, n: \qquad a_{1k} = \sum_{i=1}^{1} l_{1i} r_{ik} \;\Rightarrow\; r_{1k} := a_{1k}, \\
&j = 2, \cdots, n: \qquad a_{j1} = \sum_{i=1}^{1} l_{ji} r_{i1} \;\Rightarrow\; l_{j1} := r_{11}^{-1} a_{j1}, \\
&k = 2, \cdots, n: \qquad a_{2k} = \sum_{i=1}^{2} l_{2i} r_{ik} \;\Rightarrow\; r_{2k} := a_{2k} - l_{21} r_{1k}, \\
&\vdots
\end{aligned}
$$

und allgemein für $j = 1, \cdots, n$:

$$
\begin{aligned}
r_{jk} &:= a_{jk} - \sum_{i=1}^{j-1} l_{ji} r_{ik}\,, \quad k = j, j+1, \cdots, n\,, \\
l_{kj} &:= r_{jj}^{-1} \left(a_{kj} - \sum_{i=1}^{j-1} l_{ki} r_{ij} \right), \quad k = j+1, j+2, \cdots, n\,.
\end{aligned}
\tag{4.2.30}
$$

Die Gauß-Elimination und die direkte Dreieckszerlegung unterscheiden sich nur in der Reihenfolge der Operationen und sind algebraisch völlig äquivalent. Allerdings besteht bei der direkten Dreieckszerlegung wegen der längeren Summationsketten ähnlich wie beim Gram-Schmitt-Algorithmus zur Orthonormalisierung von Vektoren eine grösere Rundungsfehleranfälligkeit (Auslöschungsgefahr).

[4]Prescott D. Crout (1907–1984): US-Amerikanischer Mathematiker und Ingenieur; Professor am Massachusetts Institute of Technology (MIT); Beiträge zur numerischen Linearen Algebra („A short method for evaluating determinants and solving systems of linear equations with real or complex coefficients“, Trans. Amer. Inst. Elec. Eng. 60, 1235-1241, 1941) und zur numerischen Elektrodynamik.

4.3 Gleichungssysteme mit spezieller Struktur

4.3.1 Bandmatrizen

Die Anwendung der Gauß-Elimination zur Lösung „großer“ Gleichungssysteme ($n > 1000$) ist mit großen technischen Schwierigkeiten verbunden, wenn der Kernspeicher des Rechners nicht zur Speicherung der ganzen Koeffizientenmatrix ausreicht. In diesem Fall müssen externe Speicher verwendet werden, was wegen des Datentransfers die Rechenzeit in die Höhe treibt. Viele der in der Praxis auftretenden großen Matrizen besitzen jedoch eine besondere Struktur, welche es erlaubt, bei der Durchführung der Gauß-Elimination Speicherplatz zu sparen.

Definition 4.7: *Eine Matrix $A \in \mathbb{R}^{n,n}$ heißt „Bandmatrix“ vom Bandtyp (m_l, m_r) mit $0 \le m_l, m_r \le n-1$, wenn gilt:*

$$a_{jk} = 0 \quad \textit{für} \quad k < j - m_l \quad \textit{oder} \quad k > j + m_r \quad (j,k = 1,\ldots,n)\,.$$

Die Elemente von A sind also bis auf die Hauptdiagonale und höchstens $m_l + m_r$ Nebendiagonalen gleich Null. Die Größe $m = m_l + m_r + 1$ ist dann die sog. „Bandbreite“.

Beispiel 4.9: Wir geben einige einfache Beispiele von Bandmatrizen an:

Typ $(n-1, 0)$ untere Dreiecksmatrix

Typ $(0, n-1)$ obere Dreiecksmatrix

Typ $(1, 1)$ Tridiagonalmatrix

Beispiel einer (16×16)-Matrix vom Bandtyp $(4, 4)$:

$$A = \left[\begin{array}{cccc} B & -I & & \\ -I & B & -I & \\ & -I & B & -I \\ & & -I & B \end{array}\right] \Bigg\} 16 \qquad B = \left[\begin{array}{cccc} 4 & -1 & & \\ -1 & 4 & -1 & \\ & -1 & 4 & -1 \\ & & -1 & 4 \end{array}\right] \Bigg\} 4$$

$$I = \left[\begin{array}{cccc} 1 & & & \\ & 1 & & \\ & & 1 & \\ & & & 1 \end{array}\right] \Bigg\} 4$$

Satz 4.5 (Bandmatrix): *Ist $A \in \mathbb{R}^{n\times n}$ eine Bandmatrix vom Typ (m_l, m_r), für die das Gauß-Verfahren ohne Zeilenvertauschung durchführbar ist, dann sind auch alle reduzierten Matrizen Bandmatrizen desselben Typs, und die Faktoren L und R der Dreieckszerlegung von A sind Bandmatrizen vom Typ $(m_l, 0)$ bzw. $(0, m_r)$. Der Aufwand für die*

Berechnung der LR-Zerlegung einer Bandmatrix vom Typ (m_l, m_r) *ist*

$$N = \tfrac{1}{3} n m_l m_r + O(n(m_l + m_r)).$$

Beweis: Man erhält die Behauptung durch Nachrechnen (Übung). Q.E.D.

Zur Durchführung der Gauß-Elimination genügt es also bei Bandmatrizen, die Elemente im „Band“ zu speichern. Bei Größenordnungen $n \sim 10.000$ und $m \sim 100$ macht dies die Anwendung des Verfahrens erst möglich. Bei der obigen Modellmatrix ergibt sich ein reduzierter Speicherplatzbedarf von $16 \times 9 = 144$ (oder weniger) anstatt der $16 \times 16 = 256$ für die volle Matrix. (Die Ausnutzung der Symmetrie wird später noch diskutiert.)

Eine extreme Ersparnis ergibt sich natürlich bei den besonders einfach strukturierten Tridiagonalmatrizen

$$\begin{bmatrix} a_1 & b_1 & & \\ c_2 & \ddots & \ddots & \\ & \ddots & \ddots & b_{n-1} \\ & & c_n & a_n \end{bmatrix}.$$

Hier lassen sich die Elemente der LR-Zerlegung

$$L = \begin{bmatrix} 1 & & & \\ \gamma_2 & \ddots & & \\ & \ddots & 1 & \\ & & \gamma_n & 1 \end{bmatrix}, \qquad R = \begin{bmatrix} \alpha_1 & \beta_1 & & \\ & \ddots & \ddots & \\ & & \alpha_{n-1} & \beta_{n-1} \\ & & & \alpha_n \end{bmatrix}$$

durch einfache, rekursive Beziehungen bestimmen (Beweis durch Probe):

$$\begin{array}{rlllllllll} & & & & \alpha_1 & = & a_1 & , & \beta_1 & = & b_1\,, \\ i = 2, \ldots, n-1: & \gamma_i & = & c_i/\alpha_{i-1} & , & \alpha_i & = & a_i - \gamma_i \beta_{i-1} & , & \beta_i & = & b_i\,, \\ & \gamma_n & = & c_n/\alpha_{n-1} & , & \alpha_n & = & a_n - \gamma_n \beta_{n-1} & . & & & \end{array}$$

Hierzu sind offenbar nur $3n - 2$ Speicherplätze und $2n - 2$ a. Op. erforderlich. Dieser Spezialfall des Gauß-Verfahrens wird manchmal auch als „Thomas[5]-Algrothmus“ bezeichnet.

Häufig sind Bandmatrizen auch noch „dünn besetzt“, d. h.: Die meisten Elemente innerhalb des Bandes sind Null. Dieser Umstand kann bei der Gauß-Elimination jedoch

[5]Llewellyn Hilleth Thomas (1903–1992): Britischer Physiker und angewandter Mathematiker; arbeitete während des 2. Weltkriegs an einem ballistischen Forschungszentrum der US-Armee; ab 1945 am Computing Laboratory (von IBM gestiftet) der Columbia University und 1946–1968 dort Prof. füe Physik; danach bis 1978 an der North Carolina State-University; Beiträge zur angewandten Atomphysik; erfand u. a. den Magnetkernspeicher.

nicht zur Speicherersparnis ausgenutzt werden, da i. Allg. das ganze Band im Verlaufe des Verfahrens mit Elementen ungleich Null aufgefüllt wird.

Wesentlich für Satz 4.5 war, dass das Gauß-Verfahren ohne Zeilenvertauschungen durchgeführt werden kann, da andernfalls die Bandbreite anwächst. Wir betrachten im Folgenden zwei Klassen von Matrizen, bei denen dies der Fall ist.

4.3.2 Diagonaldominante Matrizen

Definition 4.8: *Eine Matrix* $A = (a_{ij})_{i,j=1}^n \in \mathbb{R}^{n\times n}$ *heißt „diagonaldominant“, wenn*

$$\sum_{k=1,k\neq j}^{n} |a_{jk}| \leq |a_{jj}|\,, \quad j = 1,\ldots,n\,. \tag{4.3.31}$$

Satz 4.6 (Existenz der LR-Zerlegung): *Die Matrix* $A \in \mathbb{R}^{n\times n}$ *sei regulär und diagonaldominant. Dann existiert eine LR-Zerlegung* $A = LR$*, die mit Gauß-Elimination ohne Pivotierung bestimmt werden kann.*

Beweis: Da A regulär und diagonaldominant ist, muss $a_{11} \neq 0$ sein. Folglich kann der erste Eliminationsschritt $A := A^{(0)} \to A^{(1)}$ *ohne* Pivotierung durchgeführt werden. Die Elemente $a_{jk}^{(1)}$ erhält man durch $a_{1k}^{(1)} = a_{1k},\ k = 1,\ldots,n$, und

$$j = 2,\ldots,n\,, \quad k = 1,\ldots,n\,: \qquad a_{jk}^{(1)} = a_{jk} - q_{j1}a_{1k}\,, \quad q_{j1} = \frac{a_{j1}}{a_{11}}\,.$$

Also gilt für $j = 2,\ldots,n$:

$$\begin{aligned}
\sum_{k=2,k\neq j}^{n} |a_{jk}^{(1)}| &\leq \sum_{k=2,k\neq j}^{n} |a_{jk}| + |q_{j1}| \sum_{k=2,k\neq j}^{n} |a_{1k}| \\
&\leq \underbrace{\sum_{k=1,k\neq j}^{n} |a_{jk}|}_{\leq |a_{jj}|} - |a_{j1}| + \underbrace{|q_{j1}|}_{=\left|\frac{a_{j1}}{a_{11}}\right|} \underbrace{\sum_{k=2}^{n} |a_{1k}|}_{\leq |a_{11}|} - |q_{j1}||a_{1j}| \\
&\leq |a_{jj}| - |q_{j1}a_{1j}| \leq |a_{jj} - q_{j1}a_{1j}| = |a_{jj}^{(1)}|.
\end{aligned}$$

Die Matrix $A^{(1)} = G_1A^{(0)}$ ist regulär und offenbar wieder diagonaldominant, und folglich ist $a_{22}^{(1)} \neq 0$. Diese Eigenschaft bleibt also bei Durchführung der Gauß-Elimination erhalten. Der ganze Prozess ist somit ohne Zeilenvertauschungen durchführbar. Q.E.D.

Bemerkung 4.1: Gilt in (4.3.31) für alle $j \in \{1,\ldots,n\}$ die strikte Ungleichung, so heißt die Matrix A „strikt diagonaldominant“. Der Beweis von Satz 4.6 zeigt, dass für eine solche die Gauß-Elimination stets ohne Pivotierung durchführbar ist, d. h.: Die Matrix

ist „regulär“. Die obige Modellmatrix ist zwar diagonaldominant, aber nicht strikt diagonaldominant. Dass sie trotzdem regulär ist, wird sich später aufgrund eines schärferen Kriteriums ergeben.

4.3.3 Positiv definite Matrizen

Wir erinnern daran, dass eine (symmetrische) Matrix $A \in \mathbb{R}^{n\times n}$ mit der Eigenschaft

$$(Ax, x)_2 > 0\,, \quad x \in \mathbb{R}^n \setminus \{0\}$$

„positiv definit“ genannt wird.

Satz 4.7 (Existenz der LR-Zerlegung): *Für positiv definite Matrizen $A \in R^{n\times n}$ ist die Gauß-Eliminationsverfahren ohne Zeilenvertauschung durchführbar, und die dabei auftretenden Pivotelemente $a_{ii}^{(i)}$ sind alle positiv.*

Beweis: Da A symmetrisch und positiv definit ist, ist notwendig $a_{11} > 0$, und die Beziehung

$$a_{jk}^{(1)} = a_{jk} - \frac{a_{j1}}{a_{11}} a_{1k} = a_{kj} - \frac{a_{k1}}{a_{11}} a_{1j} = a_{kj}^{(1)}$$

für $j, k = 2, \ldots, n$ zeigt, dass die im ersten Eliminationsschritt erzeugte $(n-1)\times(n-1)$-Matrix $\tilde{A}^{(1)} = (a_{jk}^{(1)})_{j,k=2,\ldots,n}$ ebenfalls symmetrisch ist. Wir wollen zeigen, dass sie auch positiv definit ist, so dass wieder $a_{22}^{(1)} > 0$. Der Eliminationsprozeß kann dann mit positivem Pivotelement fortgesetzt werden, und die Behauptung folgt durch Induktion.

Sei $\tilde{x} = (x_2, \ldots, x_n)^T \in \mathbb{R}^{n-1} \setminus \{0\}$ und $x = (x_1, \tilde{x})^T \in \mathbb{R}^n$ mit

$$x_1 = -\frac{1}{a_{11}} \sum_{k=2}^{n} a_{1k} x_k\,.$$

Dann ist

$$\begin{aligned} 0 < \sum_{j,k=1}^{n} a_{jk} x_j x_k &= \sum_{j,k=2}^{n} a_{jk} x_j x_k + 2x_1 \sum_{k=2}^{n} a_{1k} x_k + a_{11} x_1^2 \\ &\quad \underbrace{- \frac{1}{a_{11}} \sum_{j,k=2}^{n} a_{k1} a_{1j} x_k x_j + \frac{1}{a_{11}} \Big(\sum_{k=2}^{n} a_{1k} x_k \Big)^2}_{= 0 \ (a_{jk} = a_{kj})} \\ &= \sum_{j,k=2}^{n} \underbrace{\Big(a_{jk} - \frac{a_{k1} a_{1j}}{a_{11}} \Big)}_{= a_{jk}^{(1)}} x_j x_k + a_{11} \underbrace{\Big(x_1 + \frac{1}{a_{11}} \sum_{k=2}^{n} a_{1k} x_k \Big)^2}_{= 0} \end{aligned}$$

und somit $\tilde{x}^T \tilde{A}^{(1)} \tilde{x} > 0$. Q.E.D.

Für positiv definite Matrizen existiert also stets eine LR-Zerlegung $A = LR$ mit positiven Pivotelementen

$$r_{ii} = a_{ii}^{(i)} > 0\,, \quad i = 1, \ldots, n\,.$$

Wegen $A = A^T$ gilt aber auch

$$A = A^T = (LR)^T = (LD\tilde{R})^T = \tilde{R}^T D L^T$$

mit den Matrizen

$$\tilde{R} = \begin{bmatrix} 1 & r_{12}/r_{11} & \cdots & r_{1n}/r_{11} \\ & \ddots & \ddots & \vdots \\ & & 1 & r_{n-1,n}/r_{n-1,n-1} \\ 0 & & & 1 \end{bmatrix}, \qquad D = \begin{bmatrix} r_{11} & & 0 \\ & \ddots & \\ 0 & & r_{nn} \end{bmatrix}.$$

Mit der Eindeutigkeit der LR-Zerlegung folgt aus

$$A = LR = \tilde{R}^T D L^T$$

notwendig $L = \tilde{R}^T$ bzw. $R = DL^T$. Damit haben wir den folgenden Satz bewiesen.

Satz 4.8: *Positiv definite Matrizen gestatten eine sog. „Cholesky[6]-Zerlegung“.*

$$A = LDL^T = \tilde{L}\tilde{L}^T \tag{4.3.32}$$

mit der Matrix $\tilde{L} := LD^{1/2}$. Bei der Berechnung der Cholesky-Zerlegung genügt es, die Matrizen D und L zu bestimmen. Dies reduziert die benötigten Operationen auf

$$N_{\text{Cholesky}}(n) = \tfrac{1}{6}n^3 + O(n^2).$$

Der sog. „Algorithmus von Cholesky“ zur Berechnung der Zerlegungsmatrix

$$\tilde{L} = \begin{bmatrix} \tilde{l}_{11} & & 0 \\ \vdots & \ddots & \\ \tilde{l}_{n1} & \cdots & \tilde{l}_{nn} \end{bmatrix}$$

geht direkt von der Beziehung $A = \tilde{L}\tilde{L}^T$ aus, die man als ein System von $n(n+1)/2$ Gleichungen für die Größen $\tilde{l}_{jk}$, $k \le j$, auffassen kann. Ausmultiplizieren von

$$\begin{bmatrix} \tilde{l}_{11} & & 0 \\ \vdots & \ddots & \\ \tilde{l}_{n1} & \cdots & \tilde{l}_{nn} \end{bmatrix} \begin{bmatrix} \tilde{l}_{11} & \cdots & \tilde{l}_{n1} \\ & \ddots & \vdots \\ 0 & & \tilde{l}_{nn} \end{bmatrix} = \begin{bmatrix} a_{11} & \cdots & a_{1n} \\ \vdots & & \vdots \\ a_{n1} & \cdots & a_{nn} \end{bmatrix}$$

[6] Andr̈e Louis Cholesky (1975–1918): Französischer Mathematiker; Militärkarriere; Beiträge zur Numerischen Linearen Algebra.

ergibt in der ersten Spalte von $\tilde{L}$:

$$\tilde{l}_{11}^2 = a_{11}\,, \quad \tilde{l}_{21}\,\tilde{l}_{11} = a_{21}\,, \quad \ldots\,, \quad \tilde{l}_{n1}\,\tilde{l}_{11} = a_{n1}\,,$$

woraus sich

$$\tilde{l}_{11} = \sqrt{a_{11}}\,, \quad j = 2, \ldots, n\,: \quad \tilde{l}_{j1} = \frac{a_{j1}}{\tilde{l}_{11}} = \frac{a_{j1}}{\sqrt{a_{11}}}\,, \tag{4.3.33}$$

berechnet. Seien nun für ein $i \in \{2, \cdots, n\}$ die Elemente $\tilde{l}_{jk}$, $k = 1, \ldots, i-1$, $j = k, \ldots, n$, schon bekannt. Dann erhält man aus

$$\begin{aligned} \tilde{l}_{i1}^2 + \tilde{l}_{i2}^2 + \ldots + \tilde{l}_{ii}^2 &= a_{ii}\,, \quad \tilde{l}_{ii} > 0\,, \\ \tilde{l}_{j1}\tilde{l}_{i1} + \tilde{l}_{j2}\tilde{l}_{i2} + \ldots + \tilde{l}_{ji}\tilde{l}_{ii} &= a_{ji}, \end{aligned}$$

die nächsten Elemente $\tilde{l}_{ii}$ und $\tilde{l}_{ji}$, $j = i+1, \ldots, n$, gemäß

$$\begin{aligned} \tilde{l}_{ii} &= \sqrt{a_{ii} - \tilde{l}_{i1}^2 - \tilde{l}_{i2}^2 - \ldots - \tilde{l}_{i,i-1}^2}\,, \\ \tilde{l}_{ji} &= \tilde{l}_{ii}^{-1}\big\{a_{ji} - \tilde{l}_{j1}\tilde{l}_{i1} - \tilde{l}_{j2}\tilde{l}_{i2} - \ldots - \tilde{l}_{j,i-1}\tilde{l}_{i,i-1}\big\}, \quad j = i+1, \ldots, n, \end{aligned}$$

4.4 Nicht reguläre Systeme

Mit einer (nicht notwendig quadratischen) Matrix $A \in \mathbb{R}^{m\times n}$ und einem Vektor $b \in \mathbb{R}^m$ sei das Gleichungssystem

$$Ax = b \tag{4.4.34}$$

gegeben. Es wird hier auch $\mathrm{Rang}(A) < \mathrm{Rang}[A, b]$ zugelassen, d. h.: Das System muss nicht unbedingt im eigentlichen Sinne lösbar sein. In diesem Fall wird ein geeigneter erweiterter Lösungsbegriff eingeführt. Wir betrachten im Folgenden die auf Gauß zurückgehende sog. „Methode der kleinsten Fehlerquadrate". Dabei wird ein Vektor $\bar{x} \in \mathbb{R}^n$ gesucht, dessen Defekt $d \equiv A\bar{x} - b$ bzgl. der euklidischen Norm minimal ist. Dieser Lösungsbegriff fällt natürlich im Falle $\mathrm{Rang}(A) = \mathrm{Rang}[A, b]$ mit dem üblichen zusammen.

Satz 4.9 („Least-Squares"-Lösung): *Es existiert stets eine „Lösung" $\bar{x} \in \mathbb{R}^n$ von (4.4.34) mit kleinsten Fehlerquadraten („Least-Squares"-Lösung)*

$$\|A\bar{x} - b\|_2 = \min_{x\in\mathbb{R}^n} \|Ax - b\|_2. \tag{4.4.35}$$

Dies ist äquivalent dazu, dass $\bar{x}$ Lösung der sog. „Normalgleichung" ist:

$$A^T A\bar{x} = A^T b. \tag{4.4.36}$$

Im Falle $\mathrm{Rang}(A) = n$ ist $\bar{x}$ eindeutig bestimmt, andernfalls ist jede weitere Lösung von der Form $\bar{x} + y$ mit $y \in \mathrm{Kern}(A)$.

Beweis: (i) Sei $\bar{x}$ Lösung der Normalgleichung. Für ein beliebiges $x \in \mathbb{R}^n$ gilt dann

$$\begin{aligned}\|Ax-b\|_2^2 &= \|A\bar{x}-b+A(x-\bar{x})\|_2^2 \\ &= \|A\bar{x}-b\|_2^2 + 2\underbrace{(A\bar{x}-b}_{\in \mathrm{Kern}(A^T)}, \underbrace{A[x-\bar{x}])}_{\in \mathrm{Bild}(A)} + \|A(x-\bar{x})\|_2^2 \geq \|A\bar{x}-b\|_2^2,\end{aligned}$$

d. h.: $\bar{x}$ ist Minimallösung. Für eine Minimallösung $\bar{x}$ gilt umgekehrt notwendig

$$\begin{aligned}0 = \frac{\partial}{\partial x_i}\|Ax-b\|_{2|x=\bar{x}}^2 &= \frac{\partial}{\partial x_i}\Big(\sum_{j=1}^n \Big|\sum_{k=1}^n a_{jk}x_k - b_j\Big|^2\Big)_{|x=\bar{x}} \\ &= 2\sum_{j=1}^n a_{ji}\Big(\sum_{k=1}^n a_{jk}\bar{x}_k - b_j\Big) = 2(A^TA\bar{x}-A^Tb)_i,\end{aligned}$$

d. h.: $\bar{x}$ löst die Normalgleichung.

(ii) Wir untersuchen nun die Lösbarkeit der Normalgleichung. Das orthogonale Komplement von $\mathrm{Bild}(A)$ in $\mathbb{R}^m$ ist $\mathrm{Kern}(A^T)$. Also besitzt b eine eindeutige Zerlegung

$$b = s + r\,, \quad s \in \mathrm{Bild}(A)\,, \quad r \in \mathrm{Kern}(A^T).$$

Für ein $\bar{x} \in \mathbb{R}^n$ mit $A\bar{x} = s$ gilt dann

$$A^TA\bar{x} = A^Ts \;=\; A^Ts + A^Tr = A^Tb\,,$$

d. h.: $\bar{x}$ löst die Normalgleichung. Im Falle $\mathrm{Rang}(A) = n$ ist $\mathrm{Kern}(A) = \{0\}$ und $\mathrm{Bild}(A) = \mathbb{R}^n$. Aus $A^TAx = 0$ folgt also wegen $\mathrm{Kern}(A^T) \perp \mathrm{Bild}(A)$ notwendig $Ax = 0$ bzw. $x = 0$. Die Matrix $A^TA \in \mathbb{R}^{n\times n}$ ist regulär und folglich $\bar{x}$ eindeutig bestimmt. Im Falle $\mathrm{Rang}(A) < n$ gilt für jede weitere Lösung x_1 der Normalgleichung

$$b = Ax_1 + (b - Ax_1) \;\in\; \mathrm{Bild}(A) + \mathrm{Kern}(A^T)\,.$$

Wegen der Eindeutigkeit dieser orthogonalen Zerlegung ist notwendig $Ax_1 = A\bar{x}$ bzw. $\bar{x} - x_1 \in \mathrm{Kern}(A)$. Q.E.D.

4.4.1 Gaußsche Ausgleichsrechnung

Im Anschluss an Satz 4.9 betrachten wir als klassische Anwendung der Methode der kleinsten Fehlerquadrate, die sog. „Gaußsche Ausgleichsrechnung“ (kurz „Gauß-Ausgleich“). Die Aufgabenstellung ist dabei die folgende:

Zu gegebenen Funktionen $u_1, \ldots, u_n$ und Punkten $(x_j, y_j) \in \mathbb{R}^2$, $j = 1, \ldots, m$, $m > n$, ist eine Linearkombination

$$u(x) = \sum_{k=1}^n c_k u_k(x)$$

so zu bestimmen, dass die sog. „mittlere Abweichung“

$$\Delta_2 \equiv \Big(\sum_{j=1}^{m} |u(x_j) - y_j|^2 \Big)^{1/2}$$

möglichst klein wird. (Die sog. „Tschebyscheffsche Ausgleichsaufgabe“, bei der die „maximale Abweichung“

$$\Delta_\infty \equiv \max_{j=1,\ldots,m} |u(x_j) - y_j|$$

minimiert wird, ist i. Allg. wesentlich schwieriger zu behandeln.) Zur Lösung der Gauß-Ausgleichsaufgabe setzen wir

$$\begin{aligned} y &:= (y_1, \ldots, y_m)^T, \qquad c := (c_1, \ldots, c_n)^T, \\ a_k &:= (u_k(x_1), \ldots, u_k(x_m))^T, \quad k = 1, \ldots, n, \qquad A := [a_1, \ldots, a_n]. \end{aligned}$$

Zu minimieren ist also bzgl. $c \in \mathbb{R}^n$ das Funktional

$$F(c); = \|Ac - y\|_2.$$

Dies ist gleichbedeutend damit, für das (überbestimmte) Gleichungssystem $Ac = y$ eine „Lösung“ mit kleinsten Fehlerquadraten zu ermitteln. Im Falle $\mathrm{Rang}(A) = n$ ist die eindeutige „Lösung“ c dann bestimmt als Lösung der Normalgleichung

$$A^T A c = A^T y.$$

Ist speziell $u_k(x) = x^{k-1}$, so nennt man die „optimale“ Lösung

$$u(x) = \sum_{k=1}^{n} c_k x^{k-1}$$

die „Gauß-Ausgleichsparabel“ zu den Punkten (x_j, y_j), $j = 1, \ldots, m$. Wegen der Regularität der sog. „Vandermondschen[7] Determinante“

$$\det \begin{bmatrix} 1 & x_1 & \cdots & x_1^{n-1} \\ 1 & x_2 & \cdots & x_2^{n-1} \\ \vdots & \vdots & & \vdots \\ 1 & x_n & \cdots & x_n^{n-1} \end{bmatrix} = \prod_{j,k=1, j<k}^{n} (x_k - x_j) \neq 0$$

für paarweise verschiedene Stützstellen x_j ist dann stets $\mathrm{Rang}(A) = n$, d. h.: Die Ausgleichsparabel ist eindeutig bestimmt.

[7] Alexandre-Thophile Vandermonde (1735–1796): Französischer Mathematiker; begabter Musiker, kam spät zur Mathematik und publizierte hierzu nur vier Arbeiten (trotzdem Mitglied der Akademie der Wissenschaften in Paris); Beiträge zur Determinantentheorie und kombinatorischer Probleme (kurioserweise taucht die nach ihm benannte „Determinante“ in keiner dieser Arbeiten explizit auf).

Beispiel 4.10: Zu den Messdaten

x_i	-2	-1	0	1	2
y_i	0.5	0.5	2	3.5	3.5

soll mit Hilfe des Gauß-Ausgleichs eine lineare Funktion $y(x) = a + bx$ angepasst werden. Dies ist äquivalent zur Lösung des überbestimmten Gleichungssystems

$$\begin{bmatrix} 1 & -2 \\ 1 & -1 \\ 1 & 0 \\ 1 & 1 \\ 1 & 2 \end{bmatrix} \begin{bmatrix} a \\ b \end{bmatrix} = \begin{bmatrix} 0.5 \\ 0.5 \\ 2 \\ 3.5 \\ 3.5 \end{bmatrix}.$$

Die zugehörige Normalgleichung lautet:

$$\begin{bmatrix} 1 & 1 & 1 & 1 & 1 \\ -2 & -1 & 0 & 1 & 2 \end{bmatrix} \begin{bmatrix} 1 & -2 \\ 1 & -1 \\ 1 & 0 \\ 1 & 1 \\ 1 & 2 \end{bmatrix} \begin{bmatrix} a \\ b \end{bmatrix} = \begin{bmatrix} 1 & 1 & 1 & 1 & 1 \\ -2 & -1 & 0 & 1 & 2 \end{bmatrix} \begin{bmatrix} 0.5 \\ 0.5 \\ 2.0 \\ 3.5 \\ 3.5 \end{bmatrix}$$

$$\rightarrow \quad \begin{bmatrix} 5 & 0 \\ 0 & 10 \end{bmatrix} \begin{bmatrix} a \\ b \end{bmatrix} = \begin{bmatrix} 10 \\ 9 \end{bmatrix} \quad \rightarrow \quad \begin{bmatrix} a \\ b \end{bmatrix} = \begin{bmatrix} 2 \\ 0.9 \end{bmatrix}.$$

Es ergibt sich die Lösung $y(x) = 2 + 0.9x$ mit der mittleren Abweichung:

$$\Delta_2 = \Big(\sum_{i=1}^{5} |y(x_i) - y_i|^2 \Big)^{1/2} = \sqrt{0.9} < 1,$$

und der maximalen Abweichung:

$$\Delta_\infty = \max_{1 \le i \le 5} |y(x_i) - y_i| = 0.6.$$

Durch geometrische Anschauung erhält man in diesem Fall auch die Lösung des Tschebyscheff-Ausgleichproblems:

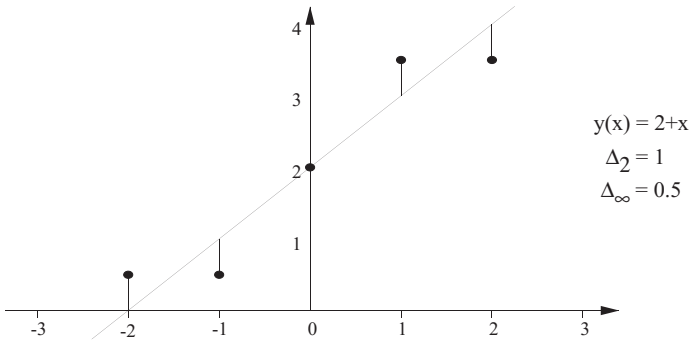


Abbildung 4.1: Lösung der Tschebyscheff-Ausgleichsaufgabe

Bemerkung 4.2: Wesentlich für die Anwendbarkeit des Gauß-Ausgleichs ist, dass für die zu bestimmenden Größen eine „lineare“ Beziehung gegeben ist, z. B. $y(x) = a + bx$. Ist die gegebene Beziehung (etwa aus physikalischen Gründen) nichtlinear, so kann man versuchen, aus ihr eine lineare Beziehung für unter Umständen andere Größen zu gewinnen, aus denen sich dann nachträglich die eigentlich gesuchten Größen bestimmen lassen; z. B.:

$$y(x) = \frac{a}{1+bx}\,.$$

$$\text{Umformung:}\quad \frac{1}{a} + \frac{b}{a}x = \frac{1}{y(x)}\,, \qquad \text{neue Größen:}\quad \tilde{a} = \frac{1}{a}\,,\quad \tilde{b} = \frac{b}{a}\,.$$

Zur Berechnung der Lösung mit kleinsten (Fehler-)Quadraten eines irregulären Systems $Ax = b$ muss die Normalgleichung $A^T Ax = A^T b$ gelöst werden. Dessen Matrix besitzt einige Besonderheiten, die in folgendem Lemma zusammengefasst sind.

Lemma 4.4: *Für eine Matrix $A \in \mathbb{K}^{m\times n}$ mit $m \geq n$ ist die Matrix $\bar{A}^T A \in \mathbb{K}^{n\times n}$ stets hermitesch (symmetrisch) und positiv semi-definit. Im Fall* $\mathrm{Rang}(A) = n$ *ist $\bar{A}^T A$ sogar positiv definit.*

Beweis: Nach den Regeln der Matrizenrechnung gilt:

$$(\bar{A}^T A)^T = A^T \bar{A} = \overline{\bar{A}^T A}, \qquad \bar{x}^T (\bar{A}^T A)x = \overline{(Ax)}^T Ax = \|Ax\|_2^2 \geq 0,$$

d. h.: $\bar{A}^T A$ ist hermitesch und positiv semi-definit. Im Fall $\mathrm{Rang}(A) = n$ ist die Matrix als Abbildung $A : \mathbb{K}^n \to \mathbb{K}^m$ $(n \leq m)$ injektiv, d. h.: $\|Ax\|_2 = 0$ impliziert $x = 0$. Die Matrix $\bar{A}^T A$ ist also positiv definit. Q.E.D.

Die Lösung des Normalgleichungssystems kann wegen der Symmetrie der Matrix A^TA prinzipiell mit dem Cholesky-Algorithmus erfolgen. I. Allg. ist sie aber sehr schlecht konditioniert; im Fall $m = n$ ist

$$\mathrm{cond(A^TA)} \sim \mathrm{cond(A)}^2 . \tag{4.4.37}$$

Beispiel 4.11: Bei 3-stelliger Rechnung erhält man

$$A = \begin{bmatrix} 1.07 & 1.10 \\ 1.07 & 1.11 \\ 1.07 & 1.15 \end{bmatrix} \quad \rightarrow \quad A^TA = \begin{bmatrix} 3.43 & 3.60 \\ 3.60 & 3.76 \end{bmatrix} .$$

Aber A^TA ist nicht positiv definit: $(-1,1)\cdot A^TA\cdot(-1,1)^T = -0.01$, d. h. Das Cholesky-Verfahren wird i. Allg. keine Lösung liefern!

Wir werden nun eine Methode betrachten, die es gestattet, die Cholesky-Zerlegun $A^TA = L^TL$ ohne explizites Ausmultipkizieren der Matrix A^TA zu berechnen. Für spätere Zwecke wird dabei der Fall komplexer Matrizen zugelassen.

Satz 4.10 (QR-Zerlegung): *Sei $A \in \mathbb{K}^{m\times n}$ eine rechteckige Matrix mit $m \geq n$ und* $\mathrm{Rang}(A) = n$ *. Dann existiert eine eindeutig bestimmte Matrix $Q \in \mathbb{K}^{m\times n}$ mit der Eigenschaft*

$$\bar{Q}^TQ = I \quad (\mathbb{K} = \mathbb{C}) , \qquad Q^TQ = I \quad (\mathbb{K} = \mathbb{R}), \tag{4.4.38}$$

und eine eindeutig bestimmte obere Dreiecksmatrix $R \in \mathbb{K}^{n\times n}$ mit reellen Diagonalelementen $r_{ii} > 0$, $i = 1,\ldots,n$, so dass

$$A = QR. \tag{4.4.39}$$

Wegen $\bar{Q}^TQ = I$ sind offenbar die Spalten von Q paarweise orthonormal; Q wird daher „orthogonale“ (genauer „orthonormale“) Matrix genannt (im Falle $m = n$ „unitäre“ Matrix).

Beweis: (i) Die Matrix Q wird durch sukzessive Orthonormalisierung der Spaltenvektoren a_k, $k = 1,\ldots,n$, von A erzeugt. Nach dem Gram-Schmidt-Verfahren setzt man

$$q_1 \equiv \|a_1\|_2^{-1} a_1$$

$$k = 2,\ldots,n : \qquad \tilde{q}_k \equiv a_k - \sum_{i=1}^{k-1} (a_k, q_i)_2 \, q_i , \quad q_k \equiv \|\tilde{q}_k\|_2^{-1} \, \tilde{q}_k.$$

Wegen $\mathrm{Rang}(A) = n$ sind die n Spaltenvektoren $\{a_1,\ldots,a_n\}$ linear unabhängig, und der Orthonormalisierungsprozess kann folglich nicht vorzeitig abbrechen.

(ii) Die Matrix $Q \equiv [q_1, \ldots, q_n]$ ist konstruktionsgemäß orthonormal. Ferner gilt für $k = 1, \ldots, n$:

$$a_k = \tilde{q}_k + \sum_{i=1}^{k-1} (a_k, q_i)_2 \, q_i = \|\tilde{q}_k\|_2 \, q_k + \sum_{i=1}^{k-1} (a_k, q_i)_2 \, q_i$$

bzw.

$$a_k = \sum_{i=1}^{k} r_{ik} \, q_k \,, \qquad r_{kk} \equiv \|\tilde{q}_k\|_2 \in \mathbb{R}_+ \,, \quad r_{ik} \equiv (a_k, q_i)_2 \,.$$

Setzt man noch $r_{ik} \equiv 0$ für $i > k$, so ist dies äquivalent zur Gleichung

$$A = QR$$

mit der oberen Dreiecksmatrix $R = (r_{ik}) \in \mathbb{K}^{n \times n}$.

(iii) Zum Beweis der Eindeutigkeit der QR-Zerlegung seien $A = Q_1 R_1$ und $A = Q_2 R_2$ zwei solche Zerlegungen. Da R_1 und R_2 regulär sind $(\det(R_i) > 0)$, gilt:

$$\begin{array}{rcccl} Q & := & \bar{Q}_2^T Q_1 & = & R_2 R_1^{-1} \quad \text{rechte obere Dreiecksmatrix}\,, \\ \bar{Q}^T & = & \bar{Q}_1^T Q_2 & = & R_1 R_2^{-1} \quad \text{rechte obere Dreiecksmatrix}\,. \end{array}$$

Wegen $\bar{Q}^T Q = R_1 R_2^{-1} R_2 R_1^{-1} = I$ ist Q *orthonormal* und *diagonal* mit $|\lambda_i| = 1$. Aus $QR_1 = R_2$ folgt $\lambda_i r_{ii}^1 = r_{ii}^2 > 0$ und damit $\lambda_i \in \mathbb{R}$ sowie $\lambda_i = 1$. Also ist $Q = I$, d. h.

$$R_1 = R_2 \,, \qquad Q_1 = A R_1^{-1} = A R_2^{-1} = Q_2$$

Dies vervollständigt den Beweis. Q.E.D.

Im Fall $\mathbb{K} = \mathbb{R}$ geht die Normalgleichung $A^T A x = A^T b$ bei Verwendung der QR-Zerlegung über in

$$A^T A x = R^T Q^T Q R x = R^T R x = R^T Q^T b,$$

bzw. wegen der Regularität von R^T,

$$Rx = Q^T b. \tag{4.4.40}$$

Dieses System ist nun durch Rückwärtseinsetzen lösbar. Wegen

$$A^T A = R^T R \tag{4.4.41}$$

ist mit R also die Cholesky-Zerlegung von $A^T A$ bestimmt, ohne $A^T A$ explizit berechnen zu müssen. Bei einer quadratischen Matrix $A \in \mathbb{R}^{n \times n}$ erfordert die Berechnung der QR-Zerlegung etwa den **doppelten** Aufwand wie zur Berechnung der LR-Zerlegung mit dem Gauß-Algorithmus: $N_{\mathrm{QR}}(n) = \frac{2}{3} n^3 + O(n^2)$.

4.4.2 Householder-Verfahren

Die in Satz 4.9 verwendete Gram-Schmidt-Orthogonalisierung zum Nachweis der Existenz der QR-Zerlegung ist ungeeignet zur praktischen Berechnung von Q und R, da aufgrund von Auslöschung die Orthonormalität der Spalten von Q rasch verloren geht. Die Gram-Schmidt-Orthogonalisierung ist kein numerisch gutartiger Algorithmus. Eine stabilere Methode zur Erstellung der Zerlegung $A = QR$ ist das sog. „Householder[8]-Verfahren“, welches wir nun beschreiben wollen. Zur Verwendung an späterer Stelle lassen wir dabei wieder komplexe Matrizen zu. Für einen Vektor $v \in \mathbb{K}^m$ nennt man

$$v\bar{v}^T := \begin{bmatrix} v_1 \\ \vdots \\ v_m \end{bmatrix} [\bar{v}_1, \ldots, \bar{v}_m] = \begin{bmatrix} |v_1|^2 & v_1\bar{v}_2 & \cdots & v_1\bar{v}_m \\ \vdots & & & \\ v_m\bar{v}_1 & v_m\bar{v}_2 & \cdots & |v_m|^2 \end{bmatrix} \in \mathbb{K}^{m\times m}$$

sein „dyadisches Produkt“ (im Gegensatz zum „skalaren Produkt“ $\bar{v}^T v = \|v\|_2^2$).

Definition 4.9: *Für einen Vektor $v \in \mathbb{K}^n$ mit $\|v\|_2 = 1$, heißt die Matrix*

$$S = I - 2v\bar{v}^T \in \mathbb{K}^{m\times m}$$

„Householder-Transformation“. Offenbar ist $S = \bar{S}^T = S^{-1}$ d. h.: S (und auch $\bar{S}^T$) ist hermitesch und unitär. Ferner ist das Produkt von zwei Householder-Transformationen wieder unitär.

Zur geometrischen Interpretation der Householder-Transformation S beschränken wir uns auf den $\mathbb{R}^2$ und betrachten dort für irgendeinen normierten Vektor v, $\|v\|_2 = 1$, die Basis $\{v, v^\perp\}$, wobei $v^T v^\perp = 0$. Für einen Vektor $u = \alpha v + \beta v^\perp \in \mathbb{R}^2$ ist dann

$$\begin{aligned} Su &= (I - 2vv^T)(\alpha v + \beta v^\perp) \\ &= \alpha v + \beta v^\perp - 2\alpha (v \underbrace{v^T) v}_{=1} - 2\beta (v \underbrace{v^T) v^\perp}_{=0} = -\alpha v + \beta v^\perp . \end{aligned}$$

Die Anwendung von $S = I - 2vv^T$ auf einen Vektor u bewirkt also in der Ebene $\mathrm{Span}\{v, u\}$ eine Spiegelung von u an der orthogonalen Achse $\mathrm{Span}\{v^\perp\}$.

Ausgehend von einer Matrix $A \in \mathbb{K}^{m\times n}$ erzeugt nun das Householder-Verfahren in n Schritten eine Folge von Matrizen

$$A := A^{(0)} \to \cdots \to A^{(i-1)} \to \cdots \to A^{(n)} := \tilde{R},$$

[8] Alston Scott Householder (1904–1993): US-Amerikanischer Mathematiker; Direktor des Oak Ridge Natioanl Laboratory (1948-1969), danach Professor an der University of Tennessee; Arbeiten zur mathematischen Biologie, aber am besten bekannt durch fundamentalen Beiträge zur Numerik, insbesondere zur numerischen Lineare Algebra.

wobei $A^{(i-1)}$ die Gestalt hat:

$$A^{(i-1)} = \left[\begin{array}{ccc|c|ccc} * & & & & & & * \\ & * & & & & & \\ & & \ddots & & & & \vdots \\ & & & * & & & \\ \hline & & & & * & \cdots & * \\ & & 0 & & & & \\ & & & & * & \cdots & * \end{array}\right] \begin{array}{l} \\ \\ \\ \\ i \\ \\ \\ \end{array}$$
$$\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad i$$

Im i-ten Schritt wird eine Householder-Transformation $S_i \in \mathbb{K}^{m\times m}$ so bestimmt, dass

$$S_i A^{(i-1)} = A^{(i)}.$$

Nach n Schritten ist dann

$$\tilde{R} = A^{(n)} = S_n S_{n-1} \cdots S_1 A =: \bar{\tilde{Q}}^T A,$$

wobei $\bar{\tilde{Q}} \in \mathbb{K}^{m\times m}$ als Produkt unitärer Matrizen selbst unitär ist, und $\tilde{R} \in \mathbb{K}^{m\times n}$ die folgende Gestalt hat:

$$\tilde{R} \;=\; \left[\begin{array}{ccc} r_{11} & \cdots & r_{1n} \\ & \ddots & \vdots \\ 0 & & r_{nn} \\ \hline 0 & \cdots & 0 \end{array}\right] \begin{array}{l} \Big\} \text{ n} \\ \Big\} \text{ m.} \end{array}$$

Dies ergibt die Darstellung

$$A = \bar{S}_1^T \cdots \bar{S}_n^T \tilde{R} = \tilde{Q}\tilde{R}.$$

Hieraus erhält man die gewünschte QR-Zerlegung von A einfach durch Streichen der letzten $m-n$ Spalten in $\tilde{Q}$ sowie der letzten $m-n$ Zeilen in $\tilde{R}$:

$$A \;=\; \underbrace{\left[\begin{array}{c|c} & \\ Q & * \\ & \end{array}\right]}_{\text{n}\qquad\text{m - n}} \cdot \left[\begin{array}{c} R \\ \hline 0 \end{array}\right] \begin{array}{l} \Big\} \text{ n} \\ \Big\} \text{ m} - \text{n} \end{array} \;=\; QR$$

Man beachte, dass hier die Diagonalelemente von R nicht notwendig positiv sein müssen, d. h.: Der Householder-Algorithmus liefert in der Regel nicht die durch Satz 4.10 gegebene „eindeutig bestimmte“ QR-Zerlegung mit Einträgen $r_{ii} > 0$.

Wir beschreiben nun den Transformationsprozess im Detail. Seien a_k die Spaltenvektoren der Matrix A.

1. Schritt: S_1 wird so gewählt, dass $S_1 a_1 \in \mathrm{Span}\{e_1\}$.

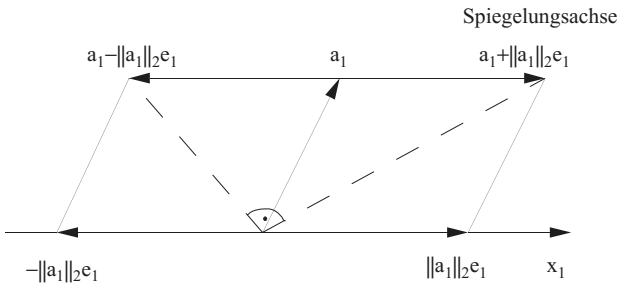


Abbildung 4.2: Schema der Householder-Transformation

Im Folgenden werden euklidische Norm und Skalarprodukt zur Abkürzung mit $\|\cdot\| = \|\cdot\|_2$ und $(\cdot,\cdot) = (\cdot,\cdot)_2$ bezeichnet. Der Vektor a_1 wird an der Achse $\mathrm{Span}\{a_1 + \|a_1\|e_1\}$ (oder $\mathrm{Span}\{a_1 - \|a_1\|e_1\}$) in die x_1-Achse gespiegelt. (Zur Vermeidung von Auslöschung wählt man gewöhnlich das Vorzeichen entsprechend $\mathrm{sgn}(a_{11})$.) Im Falle $a_{11} \geq 0$ ist also

$$v_1 = \frac{a_1 + \|a_1\|e_1}{\|a_1 + \|a_1\|e_1\|}, \qquad v_1^{\perp} = \frac{a_1 - \|a_1\|e_1}{\|a_1 - \|a_1\|e_1\|}.$$

Die Matrix $A^{(1)} = (I - 2v_1\bar{v}_1^T)A$ hat dann die Spaltenvektoren

$$a_1^{(1)} = -\|a_1\|e_1\,, \quad a_k^{(1)} = a_k - 2(a_k, v_1)v_1, \quad k = 2, \ldots, n\,.$$

Sei nun die transformierte Matrix $A^{(i-1)}$ schon berechnet.

i-ter Schritt: Für S_i machen wir den folgenden Ansatz:

$$S_i = \underbrace{\begin{bmatrix} I & 0 \\ \hline 0 & I - 2\tilde{v}_i\bar{\tilde{v}}_i^T \end{bmatrix}}_{\text{i - 1}} = I - 2v_i\bar{v}_i^T, \qquad v_i = \left.\begin{bmatrix} 0 \\ \vdots \\ 0 \\ \tilde{v}_i \end{bmatrix}\right\}\begin{matrix}\text{i - 1} \\ \text{m}\end{matrix}$$

Die Anwendung der (unitären) Matrix S_i von links auf $A^{(i-1)}$ lässt die ersten $i-1$ Zeilen- und Spalten von $A^{(i-1)}$ unverändert. Zur Konstruktion von v_i wenden wir die Überlegung vom 1. Schritt auf die Teilmatrix:

$$\tilde{A}^{(i-1)} = \begin{bmatrix} \tilde{a}_{ii}^{(i-1)} & \cdots & \tilde{a}_{in}^{(i-1)} \\ \vdots & & \vdots \\ \tilde{a}_{mi}^{(i-1)} & \cdots & \tilde{a}_{mn}^{(i-1)} \end{bmatrix} = \left[\tilde{a}_i^{(i-1)}, \ldots, \tilde{a}_n^{(i-1)}\right]$$

an. Es ist demnach

$$\tilde{v}_i = \frac{\tilde{a}_i^{(i-1)} - \|\tilde{a}_i^{(i-1)}\|\tilde{e}_i}{\| \quad \cdots \quad \|}, \quad \tilde{v}_i^{\perp} = \frac{\tilde{a}_i^{(i-1)} + \|\tilde{a}_i^{(i-1)}\|\tilde{e}_i}{\| \quad \cdots \quad \|},$$

und die Matrix $A^{(i)}$ hat die Spaltenvektoren

$$\begin{aligned} a_k^{(i)} &= a_k^{(i-1)}, \quad k = 1, \ldots, i-1, \\ a_i^{(i)} &= (a_{1i}^{(i-1)}, \ \ldots \ , a_{i-1,i}^{(i-1)}, \ \|\tilde{a}_i^{(i-1)}\|, 0, \ldots, 0)^T, \\ a_k^{(i)} &= a_k^{(i-1)} - 2(\tilde{a}_k^{(i-1)}, \tilde{v}_i)v_i, \quad k = i+1, \ldots, n. \end{aligned}$$

4.5 Die Singulärwertzerlegung

Die in den vorhergehenden Abschnitten vorgestellten Methoden zur Lösung linearer Gleichungssysteme oder Ausgleichsprobleme (Methode der kleinsten Fehlerquadrate) werden numerisch unzuverlässig, wenn die Matrizen sehr schlecht konditioniert sind. Es kann sein, dass eine eigentlich reguläre Matrix für die numerische Rechnung singulär erscheint. Die Bestimmung des Ranges einer Matrix ist mit der LR- oder QR-Zerlegung oft nicht mit genügender Sicherheit zu entscheiden. Die derzeit zuverlässigste Technik zur Behandlung nahezu rang-defizienter linearer Gleichungs- und Ausgleichsprobleme verwendet die sog. „Singulärwertzerlegung“ („singular value decomposition“, „SVD“) einer Matrix. Dabei handelt es sich um einer spezielle „orthogonale“ Zerlegung, welche die Matrix von beiden Seiten transformiert.

Es sei $A \in \mathbb{R}^{m\times n}$ gegeben. Weiter seien $Q \in \mathbb{R}^{m\times m}$ und $Z \in \mathbb{R}^{n\times n}$ orthogonal. Dann gilt

$$\|QAZ\|_2 = \|A\|_2, \tag{4.5.42}$$

so dass auch solche beidseitigen Transformationen die Kondition der Matrix A nicht verschlechtern. Für geeignete Matrizen Q und Z erhält man nun präzise Informationen über den Rang einer Matrix. Außerdem lässt sich das Ausgleichsproblem auch im Fall reduzierten Ranges befriedigend lösen. Allerdings ist die numerisch stabile Berechnung solcher Transformationen recht aufwendig, wie aus der Tabelle am Ende dieses Abschnittes hervorgeht.

Satz 4.11 (Singulärwertzerlegung): *Es sei* $A \in \mathbb{R}^{m\times n}$ *. Dann existieren orthogonale Matrizen* $V \in \mathbb{R}^{n\times n}$ *und* $U \in \mathbb{R}^{m\times m}$ *, so dass*

$$U^T AV = \Sigma = \mathrm{diag}(\sigma_1, \ldots, \sigma_p) \in \mathbb{R}^{m\times n}, \quad p = \min(m, n), \tag{4.5.43}$$

wobei $\sigma_1 \geq \sigma_2 \geq \cdots \geq \sigma_p \geq 0$ *.*

Je nachdem, ob $m \leq n$ oder $m \geq n$ ist, erhält Σ die Gestalt

$$\left(\begin{array}{ccc|c} \sigma_1 & & 0 & \\ & \ddots & & 0 \\ 0 & & \sigma_m & \end{array}\right) \qquad \text{oder} \qquad \left(\begin{array}{c} \sigma_1 \\ & \ddots \\ & & \sigma_n \\ \hline & 0 & \end{array}\right).$$

Man nennt die Werte σ_i die „singulären Werte“ der Matrix A. Aus (4.5.43) liest man unmittelbar ab, dass mit den Spaltenvektoren u_i, v_i von U, V gilt:

$$Av_i = \sigma_i u_i, \quad A^T u_i = \sigma_i v_i,$$

für $i = 1, \ldots, \min(m, n)$. Daraus ergibt sich

$$A^T A v_i = \sigma_i^2 v_i, \quad AA^T u_i = \sigma_i^2 u_i.$$

Die singulären Werte σ_i, $i = \ldots, \min(m, n)$ sind also gerade die Wurzeln der Eigenwerte von $A^T A$ bzw. AA^T.

Die Existenz einer Zerlegung der Form (4.5.43) lässt sich mit der letzten Überlegung unmittelbar darauf zurückführen, dass $A^T A$ sich durch orthogonale Matrizen auf Diagonalgestalt transformieren lässt,

$$Q^T (A^T A) Q = D.$$

Wir geben hier einen alternativen, mehr konstruktiven Beweis.

Beweis: Es sei $\sigma = \|A\|_2$. Wegen $\|A\|_2 = \max_{\|x\|_2=1} \|Ax\|_2$ existiert ein $x \in \mathbb{R}^n$ mit

$$Ax = \sigma y, \qquad \|x\|_2 = \|y\|_2 = 1.$$

Wir ergänzen x und y zu Orthonormalbasen des $\mathbb{R}^n$ und $\mathbb{R}^m$:

$$U = [y, \tilde{y}], \qquad V = [x, \tilde{x}].$$

Damit ergibt sich

$$A_1 \equiv U^T AV = \begin{pmatrix} \sigma & w^T \\ 0 & B \end{pmatrix}$$

mit einem Vektor $w \in \mathbb{R}^{n-1}$ und einer Matrix B. Da U und V orthogonal sind, folgt aus (4.5.42):

$$\|A_1\|_2 = \|A\|_2 = \sigma.$$

Andererseits gilt

$$\|A_1(\sigma, w)^T\|_2^2 = \|(\sigma^2 + \|w\|_2^2, Bw)^T\|_2^2 \geq (\sigma^2 + \|w\|_2^2)^2 = (\sigma^2 + \|w\|_2^2)\|(\sigma, w)^T\|_2^2$$

und somit $w \equiv 0$. Der Rest folgt mit vollständiger Induktion. Q.E.D.

Wir stellen nun einige einfache Folgerungen aus (4.5.43) zusammen. Die singulären Werte seien geordnet in der Form $\sigma_1 \geq \cdots \geq \sigma_r > \sigma_{r+1} = \cdots = \sigma_p = 0\,, p = \min(m, n)\,.$

- $\mathrm{Rang}(A) = r\,,$
- $\mathrm{Kern}(A) = \mathrm{Span}\{v_{r+1}, \ldots, v_n\}\,,$
- $\mathrm{Bild}(A) = \mathrm{Span}\{u_1, \ldots, u_r\}\,,$
- $A = U_r \Sigma_r V_r^T \equiv \sum_{i=1}^r \sigma_i u_i v_i^T$ (Singulärwertzerlegung von A),
- $\|A\|_2 = \sigma_1 = \sigma_{\max}\,,$
- $\|A\|_F = (\sigma_1^2 + \cdots + \sigma_r^2)^{1/2}\,.$

Wir betrachten nun das Problem der Bestimmung des „numerischen Rangs“ einer Matrix. Wir definieren

$$\mathrm{Rang}(A, \varepsilon) = \min_{\|A-B\|_2 \leq \varepsilon} \mathrm{Rang}(B)\,.$$

Man bezeichnet eine Matrix als „numerisch rang-defizient“, falls

$$\mathrm{Rang}(A, \varepsilon) < \min(m, n)\,, \qquad \varepsilon = \mathrm{eps}\|A\|_2\,.$$

Stammen die Einträge der Matrix z. B. aus Messreihen, so ist statt dessen ε an die Genauigkeit der Meßergebnisse zu knüpfen.

Satz 4.12 (Fehlerabschätzung): *Es seien $A\,, U\,, V\,, \Sigma$ wie in Satz* 4.11. *Falls $k < r = \mathrm{Rang(A)}\,,$ so gilt mit der abgeschnittenen Singulärwertzerlegung*

$$A_k = \sum_{i=1}^k \sigma_i u_i v_i^T$$

die Abschätzung

$$\min_{\mathrm{Rang}(B)=k} \|A - B\|_2 = \|A - A_k\|_2 = \sigma_{k+1}\,.$$

Als Konsequenz ergibt sich für $r_\varepsilon = \mathrm{Rang}(A, \varepsilon)$ die Beziehung

$$\sigma_1 \geq \cdots \geq \sigma_{r_\varepsilon} > \varepsilon \geq \sigma_{r_\varepsilon+1} \geq \cdots \geq \sigma_p\,, \quad p = \min(m, n)\,.$$

Beweis: Wegen

$$U^T A_k V = \mathrm{diag}(\sigma_1, \ldots, \sigma_k, 0, \ldots, 0)$$

folgt $\mathrm{Rang}(A_k) = k$. Weiter erhält man

$$U^T (A - A_k) V = \mathrm{diag}(0, \ldots, 0, \sigma_{k+1}, \ldots, \sigma_p)$$

und wegen der Orthogonalität von U und V somit

$$\|A - A_k\|_2 = \sigma_{k+1}.$$

Es bleibt zu zeigen, dass für jede andere Matrix B mit Rang k die Ungleichung

$$\|A - B\|_2 \geq \sigma_{k+1}$$

gilt. Dazu wählt man eine Orthonormalbasis $\{x_1, \ldots, x_{n-k}\}$ von $\mathrm{Kern}(B)$. Aus Dimensionsgründen gilt offensichtlich

$$\mathrm{Span}\{x_1, \ldots, x_{n-k}\} \cap \mathrm{Span}\{v_1, \ldots, v_{k+1}\} \neq \emptyset.$$

Sei z mit $\|z\|_2 = 1$ aus dieser Menge. Es gilt dann

$$Bz = 0\,, \qquad Az = \sum_{i=1}^{k+1} \sigma_i (v_i^T z) u_i$$

und somit

$$\|A - B\|_2^2 \geq \|(A - B)z\|_2^2 = \|Az\|_2^2 = \sum_{i=1}^{k+1} \sigma_i^2 (v_i^T z)^2 \geq \sigma_{k+1}^2\,.$$

Hier wurde ausgenutzt, dass $z = \sum_{i=1}^{k+1} (v_i^T z) v_i$ und deshalb

$$1 = \|z\|_2^2 = \sum_{i=1}^{k+1} (v_i^T z)^2\,.$$

Q.E.D.

Mit Hilfe der Singulärwertzerlegung lässt sich auch das Ausgleichsproblem elegant lösen. Es sei im Folgenden wieder $m \geq n$. Wir haben bereits gesehen, dass jede Minimallösung,

$$\|Ax - b\|_2 = \min$$

notwendig der Normalgleichung

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

genügt. Die Lösung ist jedoch nur im (numerisch nicht unbedingt eindeutig feststellbaren) Fall, dass $\mathrm{Rang}(A) = n$ maximal ist, eindeutig bestimmt. In diesem Fall ist $A^T A$ invertierbar und es gilt $x = (A^T A)^{-1} A^T b$.

Im Fall $\text{Rang}(A) < n$ besitzen die Normalgleichungen unendlich viele Lösungen. Eindeutigkeit erzielt man durch die Zusatzforderung, dass diejenige Lösung gesucht wird, die z. B, minimale euklidische Norm besitzt. Diese heißt die „Minimallösung“ des Ausgleichsproblems.

Satz 4.13 (Minimallösung): *Es sei* $A = U\Sigma V^T$ *die Singulärwertzerlegung der Matrix* $A \in \mathbb{R}^{m\times n}$ *und es sei* $r = \text{Rang}(A)$ *. Dann ist*

$$\bar{x} = \sum_{i=1}^{r} \frac{u_i^T b}{\sigma_i} v_i$$

die eindeutig bestimmte Lösung der Normalgleichung mit minimaler euklidischer Norm. Der Fehler genügt der Beziehung

$$\rho^2 = \|A\bar{x} - b\|_2^2 = \sum_{i=r+1}^{m} (u_i^T b)^2.$$

Beweis: Für jedes $x \in \mathbb{R}^n$ gilt die Identität

$$\|Ax - b\|_2^2 = \|AVV^T x - b\|_2^2 = \|U^T AVV^T x - U^T b\|_2^2 = \|\Sigma V^T x - U^T b\|_2^2.$$

Mit der Abkürzung $z = V^T x$ liefert dies

$$\|Ax - b\|_2^2 = \|\Sigma z - U^T b\|_2^2 = \sum_{i=1}^{r} (\sigma_i z_i - u_i^T b)^2 + \sum_{i=r+1}^{m} (u_i^T b)^2.$$

Ein Minimum erfüllt also notwendig $\sigma_i z_i = u_i^T b,\ i = 1, \ldots, r$. Unter allen z mit dieser Eigenschaft hat dasjenige mit $z_i = 0\,,\ i = r+1, \ldots, m\,,$ die minimale euklidische Norm. Die Identität für den Fehler ist offensichtlich. Q.E.D.

Die eindeutig bestimmte Minimallösung des Ausgleichsproblems lässt sich kompakt wie folgt darstellen: Es sei

$$\Sigma^+ = \text{diag}(\sigma_1^{-1}, \ldots, \sigma_r^{-1}, 0, \ldots, 0) \in \mathbb{R}^{n\times m}.$$

Wir nennen die Matrix

$$A^+ = V\Sigma^+ U^T \tag{4.5.44}$$

die „Pseudo-Inverse“ der Matrix A (oder auch die die „Penrose[9]-Inverse“ (1955)).

[9]Roger Penrose (1931–): Englischer Mathematiker; Professor am Birkbeck College in London (1964) und seit 1973 Professor an der Universität Oxford; fundamentale Beiträge in der Mathematik zur Theorie von Halbgruppen, zur Matrix-Analysis und zur Theorie von „Kachelungen“ sowie in der Theoretischen Physik zur Kosmologie, Relativitäts- und Quantentheorie.

Der letzte Satz besagt

$$\bar{x} = A^+ b\,, \qquad \rho = \|(I - AA^+)b\|_2. \tag{4.5.45}$$

Die Pseudo-Inverse ist die eindeutige Lösung von

$$\min_{X \in \mathbb{R}^{n\times m}} \|AX - I\|_F,$$

mit der Frobenius-Norm $\|\cdot\|_F$. Da die Identität in (4.5.45) für alle b gilt, folgt

$$\begin{aligned} \mathrm{Rang}(A) = n \quad &\Rightarrow \quad A^+ = (A^T A)^{-1} A^T \\ \mathrm{Rang}(A) = n = m \quad &\Rightarrow \quad A^+ = A^{-1}\,. \end{aligned}$$

In der numerischen Praxis ist bei der Definition der Pseudoinversen natürlich der (geeignet definierte) numerische Rang zu benutzen. Die numerisch stabile Berechnung der Singulärwertzerlegung ist recht aufwendig. Auf Einzelheiten kann hier nicht eingegangen werden; es sei auf das Buch von Golub/van Loan: „Matrix Computations“ verwiesen.

4.6 Übungsaufgaben

Übung 4.1: Man zeige, dass für jede Vektornorm $\|\cdot\|$ auf $\mathbb{K}^n$ durch

$$\|A\| := \sup\left\{\frac{\|Ax\|}{\|x\|},\, x \in \mathbb{K}^n, x \neq 0\right\} = \sup\left\{\|Ax\|,\, x \in \mathbb{K}^n, \|x\| = 1\right\}$$

eine mit ihr „verträgliche“ Matrizennorm erklärt ist. Diese wird als die von $\|\cdot\|$ erzeugte „natürliche“ Matrizennorm bezeichnet. Warum kann die Quadratsummennorm (sog. „Frobenius-Norm“)

$$\|A\|_{\mathrm{FR}} = \Big(\sum_{i,j=1}^{n} |a_{ij}|^2\Big)^{1/2}$$

keine natürliche Matrizennorm sein?

Übung 4.2: Man betrachte das lineare Gleichungssystem

$$\begin{pmatrix} 0,5 & 0,5 \\ 0,5 & -1 \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \end{pmatrix} = \begin{pmatrix} 1 \\ 1 \end{pmatrix}.$$

Wie groß sind die relativen Fehler $\|\delta x\|_1/\|x\|_1$ und $\|\delta x\|_\infty/\|x\|_\infty$, wenn der relative Fehler in den Matrixelementen höchstens $\pm 1\%$ und der in den Komponenten der rechten Seite höchstens $\pm 3\%$ beträgt? Man zeichne die Punktmenge im $\mathbb{R}^2$, in denen die Lösung $x+\delta x$ des gestörten Systems liegt. (Hinweis: Man berechne die Inverse der Koeffizientenmatrix und bestimme damit die l_1- und die l_∞-Kondition.)

Übung 4.3: a) Man löse durch Gauß-Elimination (ohne Pivotierung) das lineare Gleichungssystem $Ax = b$, wobei (Hinweis: Der Lösungsvektor ist ganzzahlig.)

$$A = \begin{bmatrix} -\frac{1}{2} & 9 & -2 & 1 \\ -\frac{3}{2} & 30 & -12 & 0 \\ 1 & -15 & 0 & -4 \\ 0 & -6 & 18 & 8 \end{bmatrix}, \qquad b = \begin{bmatrix} 3 \\ 3 \\ 2 \\ -4 \end{bmatrix}.$$

b) Man bestimme die LR-Zerlegung von A und berechne die Determinante $\det(A)$. c) Man bestimme die Inverse A^{-1} und die Konditionszahl $\mathrm{cond}_\infty(A) = \|A\|_\infty \|A^{-1}\|_\infty$.

Übung 4.4: Sei $A = (a_{ij})_{i,j=1}^n \in \mathbb{R}^{n\times n}$ eine symmetrische, positiv definite Matrix. Der Gauß-Algorithmus (ohne Pivotierung) erzeugt bei Anwendung auf A eine Folge von Matrizen $A = A^{(0)} \to \ldots \to A^{(k)} \to \ldots \to A^{(n-1)} = R$ mit einer oberen Dreiecksmatrix $R = (r_{ij})_{i,j=1}^n$ als Resultat. Man zeige, dass dieser Algorithmus wie folgt „stabil“ ist:

$$k = 1, \ldots, n-1: \qquad a_{ii}^{(k)} \le a_{ii}^{(k-1)}, \quad i = 1, \ldots, n, \quad \max_{1\le i,j\le n} |r_{ij}| \le \max_{1\le i,j\le n} |a_{ij}|.$$

(Hinweis: Man gehe von den Rekursionsformeln der Eliminationsprozesses aus.)

Übung 4.5 (Praktische Aufgabe): Das folgende MATLAB-Programm leistet die Berechnung der LR-Zerlegung $A = LR$ (sofern sie existiert) einer regulären Matrix:

```
function [L,R] = LR(A)
%-------------------------------------------------------------------
% Berechnet eine LR-Zerlegung (Gauss-Elimination).
% Eingabe: A nxn-regulre Matrix.
% Ausgabe: L nxn-untere Dreiecksmatrix, R nxn-obere Dreiecksmatrix.
%-------------------------------------------------------------------
[m,n] = size(A);
if (m ~= n), error('Matrix A ist nicht quadratisch'), end
for k=1:n-1
  A(k+1:n,k) = A(k+1:n,k)/A(k,k);
  A(k+1:n,k+1:n) = A(k+1:n,k+1:n)-A(k+1:n,k)*A(k,k+1:n);
end
L = eye(n,n) + tril(A,-1); R = triu(A); return;
```

(i) Man berechne die LR-Zerlegung der symmetrischen, positiv definiten Blockmatrix

$$A_n = \begin{bmatrix} B_m & -I_m & & \\ -I_m & B_m & \ddots & \\ & \ddots & \ddots & -I_m \\ & & -I_m & B_m \end{bmatrix} \in \mathbb{R}^{n\times n}, \qquad B_m = \begin{bmatrix} 4 & -1 & & \\ -1 & 4 & \ddots & \\ & \ddots & \ddots & -1 \\ & & -1 & 4 \end{bmatrix} \in \mathbb{R}^{m\times m}$$

$(n = m^2)$ mit der m-dimensionalen Einheitsmatrix I_m, für $m = 2^k$, $k = 1, \ldots, 7$. Die Leerstellen sind dabei mit Nullen besetzt gedacht. Mit Hilfe eines selbst erstellten Programms bestimme man mit Hilfe der gewonnenen LR-Zerlegung $A_n = L_n R_n$ noch die Cholesky-Zerlegung $A_n = L_n L_n^T$ und die Inverse A_n^{-1}. Die Genauigkeit überprüfe man jeweils durch Berechnung der Fehlernormen

$$\|A_n - L_n R_n\|_\infty, \quad \|A_n - L_n L_n^T\|_\infty, \quad \|A_n A_n^{-1} - I_n\|_\infty .$$

(ii) Was lässt sich über die l_∞-Konditionszahl $\mathrm{cond}_\infty(A_n) = \|A_n\|_\infty \|A_n^{-1}\|_\infty$ von A_n in Abhängigkeit von der Dimension n sagen?

Übung 4.6: Man zeige für allgemeine Matrizen $A \in \mathbb{K}^{n\times n}$ die Beziehung

$$\|A\|_2 := \sup\left\{\frac{\|Ax\|_2}{\|x\|_2},\, x \in \mathbb{K}^n, x \neq 0\right\} = \sup\left\{\sqrt{|\lambda|},\, \lambda \text{ Eigenwert von } \bar{A}^T A\right\}.$$

(Hinweis: Siehe den obigen Beweis für hermitesches A. Man beachte, dass für allgemeines A die Matrix $\bar{A}^T A$ stets hermitesch ist und somit eine Orthonormalbasis von Eigenvektoren besitzt.)

Übung 4.7: Unter einer „LR-Zelegung“ einer regulären Matrix $A \in \mathbb{R}^{n\times n}$ versteht man allgemein eine Produktzerlegung der Form $A = LR$ mit einer unteren Dreiecksmatrix L, mit Einsen auf der Hauptdiagonalen, und einer (regulären) oberen Dreiecksmatrix R.

(i) Man verifiziere, dass die regulären unteren Dreiecksmatrizen $L \in \mathbb{K}^{n\times n}$, mit Einsen auf der Hauptdiagonalen, und ebenso die allgemeinen regulären oberen Dreiecksmatrizen $R \in \mathbb{K}^{n\times n}$ bezüglich der üblichen Matrizenmultiplikation „Gruppen“ bilden. Sind diese Gruppen „abelsch“?

(ii) Man zeige damit, dass die mit dem Gauß-Verfahren erzeugte LR-Zerlegung einer regulären Matrix $A \in \mathbb{K}^{n\times n}$ (falls sie existiert) eindeutig bestimmt ist.

Übung 4.8: Gegeben sei das Gleichungssystem $Ax = b$ mit

$$A = \begin{bmatrix} 5 & -5 & 0 & 0 \\ -5 & 7 & -2 & 0 \\ 0 & -2 & 20 & -18 \\ 0 & 0 & -18 & 19 \end{bmatrix}, \qquad b = \begin{bmatrix} 5 \\ -7 \\ 20 \\ -17 \end{bmatrix}$$

und der Lösung $x = (2, 1, 2, 1)^T$.

a) Man bestimme eine Näherungslösung mit dem Cholesky-Algorithmus unter Verwendung 4-stelliger Arithmetik mit korrekter Rundung.

b) Man versuche, das Ergebnis durch einen Nachiterationsschritt unter Verwendung 8-stelliger Arithmetik für den Defekt zu verbessern.

Übung 4.9: Sei $A \in \mathbb{R}^{n\times n}$ eine reguläre Matrix, für die eine LR-Zerlegung existiert. In der Vorlesung wurde gezeigt, dass sich diese mit Gauß-Elimination (ohne Pivotierung) in $\frac{1}{3}n^3 + O(n^2)$ a. Op. berechnen lässt. Im Fall einer symmetrischen Matrix reduziert sich dieser Aufwand zu $\frac{1}{6}n^3 + O(n^2)$ a.Op. Dabei entspricht eine „a.Op.“ gerade einer Multiplikation (mit einer Addition) oder einer Division.

Frage: Wie sehen diese Aufwandszahlen für Band-Matrizen vom Typ (m_l, m_r) mit $m_l = m_r$ aus? Man konkretisiere dies anhand der in der nächsten (praktischen) Aufgabe betrachteten Modellmatrix für die Werte $m = 10^2$ bzw. $n = m^2 = 10^4$.

Übung 4.10 (Praktische Aufgabe): a) Man schreibe ein MATLAB-Programm zur Berechnung der Cholesky-Zerlegung von symmetrischen positiv definiten Matrizen mit Hilfe des Algorithmus von Cholesky und wende es an für die Modellmatrix

$$A_n = \begin{bmatrix} B_m & -I_m & & \\ -I_m & B_m & \ddots & \\ & \ddots & \ddots & -I_m \\ & & -I_m & B_m \end{bmatrix} \in \mathbb{R}^{n\times n}, \quad B_m = \begin{bmatrix} 4 & -1 & & \\ -1 & 4 & \ddots & \\ & \ddots & \ddots & -1 \\ & & -1 & 4 \end{bmatrix} \in \mathbb{R}^{m\times m}$$

$(n = m^2)$ mit der m-dimensionalen Einheitsmatrix I_m, für $m = 2, \ldots, 20$. Man überprüfe die Genauigkeit wieder durch die Probe $\|A_n - L_n^T L_n\|_\infty$. Welche Einsparungen an Speicherplatz und a. Op. ließen sich hier durch Ausnutzen der Matrixstruktur erzielen?

b) Man wende das Programm auf die Hilbert-Matrix

$$H_n = \begin{bmatrix} 1 & \frac{1}{2} & \frac{1}{3} & \cdots & \frac{1}{n} \\ \frac{1}{2} & \frac{1}{3} & \frac{1}{4} & \cdots & \frac{1}{n+1} \\ \frac{1}{3} & \frac{1}{4} & \frac{1}{5} & \cdots & \frac{1}{n+2} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ \frac{1}{n} & \frac{1}{n+1} & \frac{1}{n+2} & \cdots & \frac{1}{2n-1} \end{bmatrix},$$

an für $n = 2, \ldots, 20$ und plotte die Residuennorm $\|A_n - L_n^T L_n\|_\infty$. Welche Ergebnisse liefert hier das MATLAB-interne Programm zur Cholesky-Zerlegung?

Übung 4.11: Betrachtet werde das Gleichungssystem $Ax = b$ der Form

$$\begin{bmatrix} 1 & 3 & -4 \\ 3 & 9 & -2 \\ 4 & 12 & -6 \\ 2 & 6 & 2 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ x_3 \end{bmatrix} = \begin{bmatrix} 1 \\ 1 \\ 1 \\ 1 \end{bmatrix}.$$

a) Man untersuche, ob das System lösbar ist (mit Begründung).

b) Man bestimme eine Lösung nach der Methode der kleinsten Fehlerquadrate.

c) Ist diese Lösung eindeutig?

d) Ist die Matrix $A^T A$ positiv definit?

Übung 4.12: Wenn in dem Gleichungssystem von Aufgabe 9.1 einzelnen der Gleichungen bei der Lösung mehr Gewicht beibemessen werden soll, z.B. weil die zugehörigen Meßwerte zuverlässiger als die anderen sind, so kann dies dadurch berücksichtigt werden, daß statt $\|Ax - b\|_2$ eine gewichtete Quadratsumme $\|D(Ax - b)\|_2$ minimiert wird. Dabei ist $D = \text{diag}(d_{ii})$ eine Diagonalmatrix mit Elementen $d_{ii} > 0$. Wie lautet in diesem Fall das zugehörige Normalgleichungssystem?

Übung 4.13: Man berechne die QR-Zerlegung der Matrix

$$A = \begin{bmatrix} 0 & 0 & 0 & 2 \\ -2 & 1 & 0 & 0 \\ 0 & -2 & 1 & 0 \\ 0 & 0 & -2 & 1 \end{bmatrix}$$

mit Hilfe des Householder-Verfahrens.

Übung 4.14: Nach dem ersten Keplerschen Gesetz bewegt sich ein Komet im Sonnensystem auf einer ebenen Bahn von Ellipsen- oder Hyperbelform, wenn Störungen durch die Planeten vernachlässigt werden. Bezüglich eines in der Sonne zentrierten polaren (r, φ)-Koordinatensystems wird diese Bahn durch die sog. „Kegelschnittgleichung“

$$r = \frac{p}{1 - e\,\cos(\varphi)}$$

beschrieben mit der sog. „Exzentrizität“ e und einem Parameter p. Für $0 \le e < 1$ liegt eine Ellipse, für $e = 1$ eine Parabel und für $e > 1$ eine Hyperbel vor. Für einen neu entdeckten Kometen wurden die folgenden Beobachtungen gemacht:

Meßtag	15. Jan.	15. April	15. Juni	15. MAug.	15. Sept.	
r	10	5	2.5	1.3	1	(Einheiten)
$\cos(\varphi)$	$\sim 0,63$	$\sim 0,39$	$\sim 0,12$	$\sim -0,31$	$\sim -0,59$	

Man bestimme mit Hilfe der Gaußschen Ausgleichsrechnung den Typ der Kometenbahn. (Hinweis: Man schreibe die Kegelschnittgleichung zunächst in der Form $1/p - e/p\cos(\varphi) = 1/r$, die linear in $1/p$ und $1/e$ ist. Es genügt zweistellige Rechnung).

Übung 4.15 (Praktische Aufgabe): a) Man schreibe ein Programm zur Berechnung der QR-Zerlegung einer Matrix $A \in \mathbb{R}^{n \times n}$ mit dem Householder-Verfahren und teste es für die Matrix aus Aufgabe 9.3.

b) Die numerische Stabilität des Algorithmus untersuche man anhand der Hilbert-Matrix

$$H_n = \begin{bmatrix} 1 & \frac{1}{2} & \frac{1}{3} & \cdots & \frac{1}{n} \\ \frac{1}{2} & \frac{1}{3} & \frac{1}{4} & \cdots & \frac{1}{n+1} \\ \frac{1}{3} & \frac{1}{4} & \frac{1}{5} & \cdots & \frac{1}{n+2} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ \frac{1}{n} & \frac{1}{n+1} & \frac{1}{n+2} & \cdots & \frac{1}{2n-1} \end{bmatrix},$$

für $n = 2^k$, $k = 1, 2, \ldots, 8$. Man überprüfe die Genauigkeit der QR-Zerlegung anhand der Defektnorm $\|A - QR\|_\infty$.

c) Die QR-Zerlegung einer Matrix $A \in \mathbb{R}^{n \times n}$ liefert die Cholesky-Zerlegung der Matrix $A^T A$ gemäß $A^T A = R^T R$. Man überprüfe die Qualität dieser Zerlegung für die Hilbert-Matrix H_n anhand der Defektnorm $\|A^T A - R^T R\|_\infty$ und vergleiche dies mit der Cholesky-Zerlegung von $A^T A$.