Numerische Mathematik I

Vorlesung WS 2009/10 gehalten von

Gerlind Plonka-Hoch

Inhalt

- **Einführung**
- **Direkte Verfahren zur Lösung von LGS**
- **Iterative Verfahren zur Lösung von LGS**
- **Ausgleichsrechnung**
- **Matrizeneigenwertprobleme**
- **Nichtlineare Gleichungen**
- **Interpolation**
- **Numerische Integration**

Kapitel 1

Einführung

Was will Numerische Mathematik?

Viele Probleme der realen Welt lassen sich mathematisch formulieren und mit Hilfe des Computers lösen.

Forderung an die Lösungsverfahren

- Effizienz (Berechnung innerhalb einer problemabhängigen Zeitspanne)
- Genauigkeit (Lösung innerhalb eines gewissen Toleranzbereiches)

Aufgaben der Numerik

- Konstruktion von Verfahren zum Auffinden von Lösungen
- Analyse der Verfahren bzgl. Effizienz und Störungsanfälligkeit

Literatur

R. Schaback, H. Wendland *Numerische Mathematik*, Springer, **2005.**

M. Hanke-Bourgeois *Grundlagen der Numerischen Mathematik und des Wissenschaftlichen Rechnens*, B.G. Teubner, Stuttgart, **2002.**

G. Hämmerlin, K.-H. Hoffmann *Numerische Mathematik,* Springer, Berlin, **1989**

T. Sauer *Numerische Mathematik I*, Vorlesungsskript (Vorlesung gehalten im Wintersemester 1999/2000)

H.R. Schwarz *Numerische Mathematik,* B.G. Teubner Stuttgart, **1988**

J. Stoer *Numerische Mathematik I,* Springer, Berlin, **1999** (8. Auflage)

J. Stoer, R. Burlisch *Einführung in die numerische Mathematik II,* Springer, **1978**

MAPLE: W. Burkhardt *Erste Schritte mit Maple,* Springer, **1994**

Beispiel 1.1: (Berechnung der Ableitung)

Sei $f \in C^1(\mathbb{R})$ (einmal stetig differenzierbar) gegeben. Dann gilt

$$f'(x) \;=\; \lim_{h\to 0} \frac{f(x+h)-f(x)}{h}, \qquad x \in \mathbb{R}.$$

Daraus ergäbe sich zum Beispiel das Verfahren (MAPLE):

```
> # Berechnung der Ableitung für cos(1)
> printlevel:=0;   h:=1;   x:=1;
> for i from 0 by 1 to 40 do
  res := evalf((cos(x+h)-cos(x))/h);
  print(level=i, abstand=evalf(h), resultat=res);
  h:=h/2; end do;
```

Es ergibt sich dann

level=0	abstand=1	resultat=-0.9564491424
$\vdots$	$\vdots$	$\vdots$
level=10	abstand=0.00976525	resultat=-0.8417342464
$\vdots$	$\vdots$	$\vdots$
level=20	abstand=$0.9537 \cdot 10^{-6}$	resultat=-0.8417968121
$\vdots$	$\vdots$	$\vdots$
level=30	abstand=$0.9313 \cdot 10^{-9}$	resultat=-0.9663676416
$\vdots$	$\vdots$	$\vdots$
level=40	abstand=$0.909 \cdot 10^{-12}$	resultat=0

Das Ergebnis nähert sich zunächst dem korrekten Wert -0.841471 an und entfernt sich dann wieder davon. Das ist keine gute numerische Berechnungsvorschrift für Ableitungen! (Grund für die Abweichung: die Rechengenauigkeit des Rechners liefert nur bis zu einer bestimmten Nachkommastelle exakte Werte.)

Beispiel 1.2: (Lösung eines LGS)

Betrachte das LGS

$$\underline{\mathbf{A}} \cdot \underline{\mathbf{x}} \;=\; \underline{\mathbf{b}}, \qquad \underline{\mathbf{A}} \in \mathbb{R}^{n\times n},\ \underline{\mathbf{b}} \in \mathbb{R}^n,\ \underline{\mathbf{x}} \in \mathbb{R}^n$$

mit gegebenem $\underline{\mathbf{A}}, \underline{\mathbf{b}}$, gesucht ist $\underline{\mathbf{x}}$. Eine Berechnungsvorschrift aus der Linearen Algebra ist die ***Cramer'sche Regel***. Es sei

$$\underline{\mathbf{A}}_j \;=\; \begin{bmatrix} a_{11} & \cdots & a_{1(j-1)} & b_1 & a_{1(j+1)} & \cdots & a_{1n} \\ \vdots & & \vdots & \vdots & \vdots & & \vdots \\ a_{n1} & \cdots & a_{n(j-1)} & b_n & a_{n(j+1)} & \cdots & a_{nn} \end{bmatrix}$$

(das heißt, die j-te Spalte von $\underline{\mathbf{A}}$ wird durch $\underline{\mathbf{b}}$ ersetzt). Dann gilt: Ist das LGS eindeutig lösbar, so folgt

$$x_j \;=\; \frac{\det \underline{\mathbf{A}}_j}{\det \underline{\mathbf{A}}}.$$

Die Determinantenrechnung ist aber sehr aufwändig, nach dem Satz von Leibniz gilt für jedes $j \in \{1,\ldots,n\}$

$$\det \underline{\mathbf{A}} \;=\; (-1)^j \sum_{k=1}^{n} (-1)^k a_{jk} \cdot \det\left(\underline{\mathbf{A}}_{jk}\right),$$

wobei

$$\underline{\mathbf{A}}_{jk} = \begin{bmatrix} a_{11} & \cdots & a_{1(k-1)} & a_{1(k+1)} & \cdots & a_{1n} \\ \vdots & & \vdots & \vdots & & \vdots \\ a_{(j-1)1} & \cdots & a_{(j-1)(k-1)} & a_{(j-1)(k+1)} & \cdots & a_{(j-1)n} \\ a_{(j+1)1} & \cdots & a_{(j+1)(k-1)} & a_{(j+1)(k+1)} & \cdots & a_{(j+1)n} \\ \vdots & & \vdots & \vdots & & \vdots \\ a_{n1} & \cdots & a_{n(k-1)} & a_{n(k+1)} & \cdots & a_{nn} \end{bmatrix}.$$

Die Anzahl der Rechenoperationen zur Berechnung ***einer*** Determinante sei $f(n)$. Für $n = 2$ gilt dann

$$f(2) = 3,$$

denn im Falle einer $2 \times 2-$Determinante gilt

$$|\underline{\mathbf{A}}| = a_{11} \cdot a_{22} - a_{12} \cdot a_{21}.$$

Weiter ist allgemein

$$f(n) = n \cdot f(n-1) + (2n-1)$$

(n Multiplikationen, $n-1$ Additionen), das heißt also insgesamt für $n \geq 2$

$$\begin{aligned} f(n) &= n(n-1)f(n-2) + n(2n-3) + (2n-1) \\ &= \sum_{j=0}^{n-2} \frac{n!}{(n-j)!} (2n-2j-1). \end{aligned}$$

Für $n = 3$ ergibt sich damit

$$f(3) = 5 + \frac{6}{2} \cdot 3 = 14,$$

das heißt, der Aufwand beläuft sich auf

$$14 \cdot 4 = 56$$

Operationen allein zur Berechnung der 4 benötigten Determinanten zur Berechnung des gesamten Gleichungssystems. Für $n = 10$ ergibt sich

$$f(10) = 9862749.$$

Damit ergibt sich eine im Vergleich mit dem Gauß-Algorithmus sehr ungünstige Anzahl von Rechenoperationen, der Gauß-Algorithmus benötigt eine Anzahl von $C \cdot n^3$ (also für $n = 10 : \ C \cdot 1000$) Rechenoperationen. Wir werden Verfahren zur Lösung von LGS betrachten, die weniger als $2 \cdot n^3$ Operationen benötigen. Darüber hinaus ist die Berechnung von Determinanten sehr anfällig gegen kleine Störungen (instabil).

Beispiel 1.3: (Berechnung der Standardabweichung)

Es sei $X = (x_1, ..., x_n)^T \in \mathbb{R}^n$ ein Vektor von Messwerten. Dann heißt

$$E(X) = \frac{1}{n} \cdot \sum_{j=1}^{n} x_j$$

Erwartungswert und

$$\begin{aligned} V(X) &= \sqrt{\frac{1}{n-1}\sum_{j=1}^{n}(x_j - E(X))^2} = \sqrt{\frac{1}{n-1}\sum_{j=1}^{n}(x_j - \frac{1}{n}\sum_{k=1}^{n}x_k)^2} \\ &= \sqrt{\frac{1}{n-1}\left(\sum_{j=1}^{n}x_j^2 - \frac{2}{n}\sum_{j=1}^{n}\sum_{k=1}^{n}x_j x_k + \frac{1}{n}(\sum_{k=1}^{n}x_k)^2\right)} \\ &= \sqrt{\frac{1}{n-1}\left(\sum_{j=1}^{n}x_j^2 - \frac{1}{n}\left(\sum_{k=1}^{n}x_k\right)^2\right)} \end{aligned}$$

Standardabweichung.

Wir betrachten folgende Berechnungsvorschrift: Es sei

$$S_0 \,=\, Q_0 \,=\, 0,$$

und

$$S_j \,=\, S_{j-1} + x_j, \qquad Q_j \,=\, Q_{j-1} + x_j^2, \qquad\qquad (j = 1, ..., n).$$

Dann ergeben sich die oben definierten Ausdrücke durch

$$E(X) \,=\, \frac{S_n}{n}, \qquad V(X) \,=\, \sqrt{\frac{1}{n-1}\left(Q_n - \frac{S_n^2}{n}\right)}$$

Man kann dies mittels des folgenden MAPLE-Programmes realisieren (ungünstig!):

```
> restart;
> with(LinearAlgebra);
> v:=Vector([1.00000, 1.00001, 1.00002]);
> N:=Dimension(v);
> S:=0;   Q:=0;
> for i from 1 by 1 to N do
  S:=S+v[i];
  Q:=Q+v[i]^2; end do;
> E:=S/N;
> V:=sqrt((Q-S^2/N)/(N-1));
```

Dies liefert als Ausgabe

$$E \,:=\, 1.00001, \qquad V \,:=\, 0,$$

also ein falsches Ergebnis für die Standardabweichung V. Der Fehler bei der Standardabweichung entsteht, da MAPLE nur mit 10 Stellen hinter dem Komma genau rechnet (wenn man nichts anderes einstellt!).

Für ein stabileres Verfahren setzen wir

$$X_k \,:=\, (x_1, ..., x_k), \qquad M_k \,=\, E(X_k),$$

sowie

$$Q_k \,:=\, (k-1)\cdot V(X_k)^2 = \sum_{l=1}^{k}(x_l - M_k)^2$$

und

$$M_0 \,=\, Q_0 \,:=\, 0.$$

Wir Zeigen: Q_j, und M_j lassen sich rekursiv für $j = 1, ..., n$ berechnen,

$$M_j \,=\, M_{j-1} + \frac{x_j - M_{j-1}}{j}, \qquad Q_j \,=\, Q_{j-1} + (j-1) \cdot \frac{(x_j - M_{j-1})^2}{j}.$$

Damit ist dann

$$E(X) \,=\, M_n, \qquad V(X) \,=\, \sqrt{\frac{Q_n}{n-1}}.$$

Beweis:

Wir erhalten

$$M_j = \frac{1}{j}\sum_{l=1}^{j} x_l = \frac{1}{j}\sum_{l=1}^{j-1} x_l + \frac{x_j}{j} = \frac{(j-1)}{j} M_{j-1} + \frac{x_j}{j} = M_{j-1} + \frac{(x_j - M_{j-1})}{j}$$

und

$$\begin{aligned}
Q_j &= \sum_{l=1}^{j}(x_l - M_j)^2 = \sum_{l=1}^{j}\left(x_l - M_{j-1} - \frac{(x_j - M_{j-1})}{j}\right)^2 \\
&= \sum_{l=1}^{j-1}(x_l - M_{j-1})^2 - 2\sum_{l=1}^{j-1}(x_l - M_{j-1})\frac{(x_j - M_{j-1})}{j} + (j-1)\left(\frac{x_j - M_{j-1}}{j}\right)^2 \\
&\quad + \left(\frac{(j-1)(x_j - M_{j-1})}{j}\right)^2 \\
&= Q_{j-1} - 2\frac{(x_j - M_{j-1})}{j}((j-1)M_{j-1} - (j-1)M_{j-1}) + (x_j - M_{j-1})^2\frac{(j^2 - j)}{j^2} \\
&= Q_{j-1} + \frac{(j-1)}{j}(x_j - M_{j-1})^2.
\end{aligned}$$

□

Dann lässt sich das folgende (günstige) MAPLE-Programm erstellen:

```
> restart;
> with(LinearAlgebra);
> v:=vector([1.00000,1.00001,1.00002]);
> N:=Dimension(v);
> M:=0;   Q:=0;
> for i from 1 to N do
  Q:=Q+(i-1)*(v[i]-M)^2/i;   M:=M+(v[i]-M)/i;   end do;
> E:=M;
> V:=sqrt(Q/(N-1));
```

Dieses liefert als Ausgabe schließlich die richtigen Werte, nämlich

$$E \,=\, 1.00001, \qquad V \,=\, 0.00001.$$

Die Wahl des Berechnungsverfahrens ist also wesentlich für die Genauigkeit und die Effizienz.

Kapitel 2

Direkte Verfahren zur Lösung von LGS

Gegeben seien eine Matrix $\underline{\mathbf{A}} \in \mathbb{R}^{m \times n}$ und ein Vektor $\underline{\mathbf{b}} \in \mathbb{R}^m$, gesucht ist ein Vektor $\underline{\mathbf{x}} \in \mathbb{R}^n$, so dass

$$\underline{\mathbf{A}} \cdot \underline{\mathbf{x}} = \underline{\mathbf{b}}$$

gilt. Das Problem hat Lösungen, falls

$$rg(\underline{\mathbf{A}}) = rg(\underline{\mathbf{A}}, \underline{\mathbf{b}})$$

erfüllt ist und ist eindeutig lösbar, wenn $m = n$ und $\det(\underline{\mathbf{A}}) \neq 0$ gilt (dies ist gleichbedeutend mit $rg(\underline{\mathbf{A}}) = n$). Dann gilt

$$\underline{\mathbf{x}} = \underline{\mathbf{A}}^{-1} \cdot \underline{\mathbf{b}}.$$

Wir unterscheiden ***direkte*** und ***indirekte*** Verfahren zur Lösung von LGS.

2.1 Normen und Kondition

Um Aussagen über die numerische Stabilität eines LGS machen zu können, benötigen wir genauere Informationen über das LGS.

Beispiel:

Wir betrachten das LGS

$$\begin{array}{rcrcl} x_1 & + & 2x_2 & = & 3 \\ 2x_1 & + & cx_2 & = & 4 \end{array}$$

für die folgenden zwei Fälle:

a) Es sei $c = 3,999$. Dann erhält man als Lösung die Werte $x_1 = -3997$, $x_2 = 2000$.

b) Es sei $c = 4,001$. Dann liefert das LGS die Lösungen $x_1 = 4003$, $x_2 = -2000$.

Offensichtlich bewirken schon kleine Änderungen eines Koeffizienten große Änderungen der Lösung. In diesem Fall ist dies dadurch zu erklären, dass die Determinante der Koeffizientenmatrix für $c = 4$ gleich Null ist, das heißt, dass wir uns bei der Wahl des Koeffizienten nahe einer kritischen Stelle befunden haben. Die Lösung eines solchen LGS kann mittels MAPLE in der folgenden Weise berechnet werden:

```
> with(LinearAlgebra);
> sys1:={x1+2 * x2=3,   2 * x1 + c * x2=4};
> c:=3.999;
> solve(sys1, {x1,x2});
> c:=4.001;
> solve(sys1, {x1,x2});
```

Hierbei wurden alle entstehenden Ausgaben nicht berücksichtigt.

Definition 2.1: (Normen)

Eine Abbildung

$$\|\cdot\| : \mathbb{R}^n \longrightarrow \mathbb{R}$$

heißt ***(Vektor-)Norm***, wenn die folgenden drei Bedingungen erfüllt sind:

1) Es gilt

$$\|\underline{\mathbf{x}}\| \geq 0 \qquad \text{und} \qquad \|\underline{\mathbf{x}}\| = 0 \iff \underline{\mathbf{x}} = \underline{\mathbf{0}},$$

diese Bedingung ist die ***Positivität*** der Norm.

2) Es gilt

$$\|c \cdot \underline{\mathbf{x}}\| = |c| \cdot \|\underline{\mathbf{x}}\|, \qquad c \in \mathbb{R},$$

die ***positive Homogenität*** der Norm.

3) Es gilt die ***Dreiecksungleichung***, das heißt, für $\underline{\mathbf{x}}, \underline{\mathbf{y}} \in \mathbb{R}^n$ gilt

$$\|\underline{\mathbf{x}} + \underline{\mathbf{y}}\| \leq \|\underline{\mathbf{x}}\| + \|\underline{\mathbf{y}}\|.$$

Beispiele:

1) Die ***1-Norm*** ist in der folgenden Weise definiert:

$$\|\underline{\mathbf{x}}\|_1 := \sum_{j=1}^{n} |x_j|.$$

2) Die ***euklidische Norm*** berechnet man durch

$$\|\underline{\mathbf{x}}\|_2 := \sqrt{\sum_{j=1}^{n} x_j^2}.$$

3) Die ∞***-Norm*** **(Maximum-Norm)** ist definiert durch

$$\|\underline{\mathbf{x}}\|_\infty := \max_{j=1,\ldots,n} |x_j|.$$

4) Allgemein umfasst die p***-Norm*** für $1 \leq p < \infty$ die Fälle 1), 2), und als Grenzfall auch Fall 3). Sie ist definiert durch

$$\|\underline{\mathbf{x}}\|_p := \sqrt[p]{\sum_{j=1}^{n} |x_j|^p}.$$

Definition 2.2: (Matrixnorm)

Eine Abbildung

$$\|\cdot\| : \mathbb{R}^{n \times n} \longrightarrow \mathbb{R}$$

heißt ***Matrix-Norm,*** wenn für Matrizen $\underline{\mathbf{A}}, \underline{\mathbf{B}} \in \mathbb{R}^{n\times n}$ die folgenden drei Bedingungen gelten:

1) Es ist

$$\|\underline{\mathbf{A}}\| \geq 0 \qquad \text{und} \qquad \|\underline{\mathbf{A}}\| = 0 \iff \underline{\mathbf{A}} = \underline{\mathbf{0}} \quad \text{(Nullmatrix)}.$$

2) Es gilt

$$\|c \cdot \underline{\mathbf{A}}\| = |c| \cdot \|\underline{\mathbf{A}}\|, \qquad c \in \mathbb{R}.$$

3) Es gilt die Dreiecksungleichung

$$\|\underline{\mathbf{A}} + \underline{\mathbf{B}}\| \leq \|\underline{\mathbf{A}}\| + \|\underline{\mathbf{B}}\|.$$

Beispiele:

1) Man fasst die Matrix als Vektor des $\mathbb{R}^{n^2}$ auf und verwendet eine Vektornorm. Das liefert zum Beispiel die Norm

$$\|\underline{\mathbf{A}}\|_M = \max_{1\leq j,k\leq n} |a_{jk}|$$

als Verallgemeinerung der ∞-Norm. Die ***Frobenius-Norm,*** definiert durch

$$\|\underline{\mathbf{A}}\|_F = \sqrt{\sum_{j,k=1}^{n} |a_{jk}|^2},$$

ist die Verallgemeinerung der euklidischen Norm.

2) Zu einer beliebigen Vektornorm $\|\cdot\|_V$ lässt sich eine Matrixnorm oder ***Operator-Norm*** in der Form

$$\|\underline{\mathbf{A}}\|_M := \max_{\underline{\mathbf{x}}\neq\underline{\mathbf{0}}} \frac{\|\underline{\mathbf{A}}\underline{\mathbf{x}}\|_V}{\|\underline{\mathbf{x}}\|_V}$$

definieren.

Satz 2.3: (Matrixnormen)

Durch Einsetzen der 1-(Vektor)Norm in die Definition der Operatornorm ergibt sich die ***Spaltensummen-Norm*** für Matrizen

$$\|\underline{\mathbf{A}}\|_1 := \max_{k=1,\ldots,n} \sum_{j=1}^{n} |a_{jk}|.$$

Mit Hilfe der ∞-(Vektor)Norm ergibt sich entsprechend die ***Zeilensummen-Norm***

$$\|\underline{\mathbf{A}}\|_\infty := \max_{j=1,\ldots,n} \sum_{k=1}^{n} |a_{jk}|.$$

Durch Einsetzen der euklidischen (Vektor)Norm erhält man schließlich die ***Spektral-Norm***

$$\|\underline{\mathbf{A}}\|_2 := \sqrt{\rho(\underline{\mathbf{A}}^T\underline{\mathbf{A}})}.$$

Dabei ist $\rho(\underline{\mathbf{B}})$ der Spektralradius einer Matrix $\underline{\mathbf{B}}$, das heißt

$$\rho(\underline{\mathbf{B}}) = \max\{|\lambda_1|, |\lambda_2|, ..., |\lambda_n|\}$$

wobei λ_i $(i = 1, ..., n)$ die Eigenwerte von $\underline{\mathbf{B}}$ sind.

Beweis:

Wir zeigen für die $\|\cdot\|_1$-Norm, dass

$$\|\underline{\mathbf{A}}\|_1 := \max_{k=1,\ldots,n} \sum_{j=1}^{n} |a_{jk}| = \max_{\underline{\mathbf{x}} \neq \underline{\mathbf{0}}} \frac{\|\underline{\mathbf{A}} \cdot \underline{\mathbf{x}}\|_1}{\|\underline{\mathbf{x}}\|_1}$$

gilt. Dabei ist zu zeigen, dass

$$\|\underline{\mathbf{A}}\|_1 \geq \frac{\|\underline{\mathbf{A}} \cdot \underline{\mathbf{x}}\|_1}{\|\underline{\mathbf{x}}\|_1} \qquad \forall\ \underline{\mathbf{x}} \neq \underline{\mathbf{0}}$$

gilt und dass die Gleichheit auch angenommen werden kann. Es sei

$$\underline{\mathbf{A}} = [\,\underline{\mathbf{a}}_1 \ \cdots \ \underline{\mathbf{a}}_n\,]$$

das heißt, es sind die Vektoren $\underline{\mathbf{a}}_i$ die Spalten der Matrix $\underline{\mathbf{A}}$. Dann gilt für $\underline{\mathbf{x}} = (x_1, ..., x_n)^T$

$$\underline{\mathbf{A}} \cdot \underline{\mathbf{x}} = [\,\underline{\mathbf{a}}_1 \ \cdots \ \underline{\mathbf{a}}_n\,] \begin{bmatrix} x_1 \\ \vdots \\ x_n \end{bmatrix} = x_1\underline{\mathbf{a}}_1 + ... + x_n\underline{\mathbf{a}}_n.$$

Das liefert die Abschätzung

$$\begin{aligned} \frac{\|\underline{\mathbf{A}} \cdot \underline{\mathbf{x}}\|_1}{\|\underline{\mathbf{x}}\|_1} &= \frac{\|x_1\underline{\mathbf{a}}_1 + ... + x_n\underline{\mathbf{a}}_n\|_1}{\|\underline{\mathbf{x}}\|_1} \leq \frac{|x_1| \cdot \|\underline{\mathbf{a}}_1\|_1 + ... + |x_n| \cdot \|\underline{\mathbf{a}}_n\|_1}{\|\underline{\mathbf{x}}\|_1} \\ &\leq \max_{k=1,\ldots,n} \|\underline{\mathbf{a}}_k\|_1 \cdot \underbrace{\frac{|x_1| + ... + |x_n|}{\|\underline{\mathbf{x}}\|_1}}_{=1} = \max_{k=1,\ldots,n} \|\underline{\mathbf{a}}_k\|_1 = \max_{k=1,\ldots,n} \sum_{j=1}^{n} |a_{jk}|. \end{aligned}$$

Ist nun k_0 der Index, für den

$$\|\underline{\mathbf{a}}_{k_0}\|_1 = \max_{k=1,\ldots,n} \|\underline{\mathbf{a}}_k\|_1$$

gilt, so wähle man

$$\underline{\mathbf{x}} = \underline{\mathbf{e}}_{k_0} = (0, ..., 0, 1, 0, ..., 0)^T$$

(der k_0-te Eintrag ist eine 1, ansonsten nur Nullen) und erhält

$$\frac{\|\underline{\mathbf{A}} \cdot \underline{\mathbf{e}}_{k_0}\|_1}{\|\underline{\mathbf{e}}_{k_0}\|_1} = \frac{\|1 \cdot \underline{\mathbf{a}}_{k_0}\|_1}{1} = \max_{k=1,\ldots,n} \|\underline{\mathbf{a}}_k\|_1,$$

das heißt, in diesem (speziellen) Fall gilt sogar die Gleichheit.

Übung: Zeigen Sie, dass für die Operator-Norm alle drei Normaxiome erfüllt sind.

Definition 2.4: (Konsistenz und Verträglichkeit)

1) Eine Matrixnorm $\|\cdot\|_M$ heißt ***konsistent (submultiplikativ)***, falls

$$\|\underline{\mathbf{A}} \cdot \underline{\mathbf{B}}\|_M \leq \|\underline{\mathbf{A}}\|_M \cdot \|\underline{\mathbf{B}}\|_M \qquad \forall \quad \underline{\mathbf{A}}, \underline{\mathbf{B}} \in \mathbb{R}^{n \times n}$$

gilt.

2) Eine Matrixnorm $\|\cdot\|_M$ heißt ***mit einer Vektornorm*** $\|\cdot\|_V$ ***verträglich***, wenn

$$\|\underline{\mathbf{A}} \cdot \underline{\mathbf{x}}\|_V \leq \|\underline{\mathbf{A}}\|_M \cdot \|\underline{\mathbf{x}}\|_V$$

gilt.

Satz 2.5:

Die Operator-Norm

$$\|\underline{\mathbf{A}}\|_M := \max_{\underline{\mathbf{x}} \neq \underline{\mathbf{0}}} \frac{\|\underline{\mathbf{A}}\underline{\mathbf{x}}\|_V}{\|\underline{\mathbf{x}}\|_V}$$

ist konsistent und mit der zugehörigen Vektor-Norm verträglich.

Beweis:

Aus der Definition der Norm folgt die Abschätzung

$$\|\underline{\mathbf{A}}\|_M \geq \frac{\|\underline{\mathbf{A}} \cdot \underline{\mathbf{x}}\|_V}{\|\underline{\mathbf{x}}\|_V} \qquad \forall \quad \underline{\mathbf{x}} \neq \underline{\mathbf{0}},$$

das heißt also

$$\|\underline{\mathbf{A}} \cdot \underline{\mathbf{x}}\|_V \leq \|\underline{\mathbf{A}}\|_M \cdot \|\underline{\mathbf{x}}\|_V \qquad \forall \quad \underline{\mathbf{x}} \neq \underline{\mathbf{0}}.$$

Diese Ungleichung gilt auch für $\underline{\mathbf{x}} = \underline{\mathbf{0}}$. Damit ist die Verträglichkeit gezeigt. Betrachtet man jetzt für zwei Matrizen $\underline{\mathbf{A}}, \underline{\mathbf{B}}$ den Ausdruck

$$\|\underline{\mathbf{A}} \cdot \underline{\mathbf{B}}\|_M = \max_{\underline{\mathbf{x}} \neq \underline{\mathbf{0}}} \frac{\|\underline{\mathbf{A}} \cdot \underline{\mathbf{B}} \cdot \underline{\mathbf{x}}\|_V}{\|\underline{\mathbf{x}}\|_V},$$

so lässt sich dieser durch Ausnutzung der Verträglichkeit in der folgenden Weise abschätzen:

$$\begin{aligned} \|\underline{\mathbf{A}} \cdot \underline{\mathbf{B}}\|_M &= \max_{\underline{\mathbf{x}} \neq \underline{\mathbf{0}}} \frac{\|\underline{\mathbf{A}} \cdot \underline{\mathbf{B}} \cdot \underline{\mathbf{x}}\|_V}{\|\underline{\mathbf{x}}\|_V} \leq \max_{\underline{\mathbf{x}} \neq \underline{\mathbf{0}}} \frac{\|\underline{\mathbf{A}}\|_M \cdot \|\underline{\mathbf{B}}\underline{\mathbf{x}}\|_V}{\|\underline{\mathbf{x}}\|_V} \\ &= \|\underline{\mathbf{A}}\|_M \cdot \max_{\underline{\mathbf{x}} \neq \underline{\mathbf{0}}} \frac{\|\underline{\mathbf{B}}\underline{\mathbf{x}}\|_V}{\|\underline{\mathbf{x}}\|_V} = \|\underline{\mathbf{A}}\|_M \cdot \|\underline{\mathbf{B}}\|_M \end{aligned}$$

und man erhält die Konsistenz.

Definition 2.6: (Kondition einer Matrix)

Die ***Kondition einer invertierbaren Matrix*** $\underline{\mathbf{A}} \in \mathbb{R}^{n \times n}$ bezüglich der Matrix-Norm $\|\cdot\|_M$ ist definiert durch

$$\kappa(\underline{\mathbf{A}}) = \|\underline{\mathbf{A}}^{-1}\|_M \cdot \|\underline{\mathbf{A}}\|_M.$$

Beispiel:

Wir betrachten die Matrix

$$\underline{\mathbf{A}} = \begin{bmatrix} 1 & -3 \\ -5 & 2 \end{bmatrix}.$$

Für diese gilt

$$\|\underline{\mathbf{A}}\|_\infty = 7, \qquad \|\underline{\mathbf{A}}\|_1 = 6.$$

Wir berechnen nun $\|\underline{\mathbf{A}}\|_2$. Wegen

$$\|\underline{\mathbf{A}}\|_2 = \sqrt{\rho\left(\underline{\mathbf{A}}^T \underline{\mathbf{A}}\right)}$$

sind zunächst die Eigenwerte von $\underline{\mathbf{A}}^T \underline{\mathbf{A}}$ zu berechnen. Mit

$$\underline{\mathbf{A}}^T \underline{\mathbf{A}} = \begin{bmatrix} 1 & -5 \\ -3 & 2 \end{bmatrix} \begin{bmatrix} 1 & -3 \\ -5 & 2 \end{bmatrix} = \begin{bmatrix} 26 & -13 \\ -13 & 13 \end{bmatrix}$$

folgt für das charakteristische Polynom

$$\chi(\lambda) = \begin{vmatrix} \lambda - 26 & 13 \\ 13 & \lambda - 13 \end{vmatrix} = (\lambda - 26)(\lambda - 13) - 169 = \lambda^2 - 39\lambda + 169 \stackrel{!}{=} 0.$$

Dies liefert die Nullstellen

$$\lambda_{1,2} = \frac{39}{2} \pm \sqrt{\frac{9 \cdot 13^2}{4} - 13^2} = \frac{39}{2} \pm \frac{13}{2}\sqrt{5} = \frac{13}{2}(3 \pm \sqrt{5}).$$

Damit gilt für den Spektralradius

$$\rho\left(\underline{\mathbf{A}}^T \underline{\mathbf{A}}\right) = \frac{13}{2}(3 + \sqrt{5}),$$

und für die Spektralnorm erhalten wir

$$\|\underline{\mathbf{A}}\|_2 = \sqrt{\frac{13}{2}(3 + \sqrt{5})} \approx 5,83...$$

Schließlich erhält man für die Frobenius-Norm

$$\|\underline{\mathbf{A}}\|_F = \sqrt{1 + 9 + 25 + 4} = \sqrt{39} \approx 6,425... .$$

Betrachtet man die Kondition der Matrix $\underline{\mathbf{A}}$, so ist diese wegen $\det(\underline{\mathbf{A}}) = -13$ und damit

$$\underline{\mathbf{A}}^{-1} = \frac{1}{-13} \cdot \begin{bmatrix} 2 & 3 \\ 5 & 1 \end{bmatrix}$$

(bei 2×2-Matrizen ist die Bildung der Inversen einfach: mal teile durch die Determinante, vertausche die Elemente auf der Hauptdiagonalen und ändere das Vorzeichen der beiden übrigen Elemente!), das heißt

$$\|\underline{\mathbf{A}}^{-1}\|_1 = \frac{7}{13},$$

gegeben durch

$$\kappa_1(\underline{\mathbf{A}}) = \frac{7}{13} \cdot 6 = \frac{42}{13} \approx 3,2308...$$

2.2 Einfluss von Datenfehlern

Es sei

$$\underline{\mathbf{A}} \cdot \underline{\mathbf{x}} = \underline{\mathbf{b}}$$

das betrachtete LGS, wobei

$$\underline{\mathbf{A}} \in \mathbb{R}^{n \times n}, \qquad \underline{\mathbf{b}} \in \mathbb{R}^n, \qquad \underline{\mathbf{x}} \in \mathbb{R}^n$$

gelte und $\underline{\mathbf{A}}$ invertierbar sei. Es sei weiter $\|\cdot\|_V$ eine beliebige Vektornorm und $\|\cdot\|_M$ die zugehörige Operatornorm. Wir untersuchen den Einfluss von kleinen Änderungen in $\underline{\mathbf{A}}$ bzw. $\underline{\mathbf{b}}$ auf die Lösung $\underline{\mathbf{x}}$.

1) Es sei die rechte Seite $\underline{\mathbf{b}}$ gestört, das heißt, es wird das LGS

$$\underline{\mathbf{A}} \cdot \tilde{\underline{\mathbf{x}}} = \tilde{\underline{\mathbf{b}}} := \underline{\mathbf{b}} + \triangle\underline{\mathbf{b}}$$

betrachtet. Setzt man

$$\triangle\underline{\mathbf{x}} = \tilde{\underline{\mathbf{x}}} - \underline{\mathbf{x}} \quad \Longleftrightarrow \quad \tilde{\underline{\mathbf{x}}} = \underline{\mathbf{x}} + \triangle\underline{\mathbf{x}},$$

so lautet das LGS dann

$$\underline{\mathbf{A}}(\underline{\mathbf{x}} + \triangle\underline{\mathbf{x}}) = \underline{\mathbf{b}} + \triangle\underline{\mathbf{b}},$$

was gleichbedeutend ist mit

$$\underline{\mathbf{A}} \cdot \triangle\underline{\mathbf{x}} = \triangle\underline{\mathbf{b}}$$

(denn es gilt $\underline{\mathbf{A}} \cdot \underline{\mathbf{x}} = \underline{\mathbf{b}}$). Aufgrund der Verträglichkeit der Matrix-Norm mit der Vektornorm erhält man die Abschätzung

$$\begin{aligned} \|\triangle\underline{\mathbf{x}}\|_V &= \|\underline{\mathbf{A}}^{-1}\triangle\underline{\mathbf{b}}\|_V \\ &\leq \|\underline{\mathbf{A}}^{-1}\|_M \cdot \|\triangle\underline{\mathbf{b}}\|_V. \end{aligned}$$

Die relative Änderung der Lösung ist

$$\frac{\|\tilde{\underline{\mathbf{x}}} - \underline{\mathbf{x}}\|_V}{\|\underline{\mathbf{x}}\|_V} = \frac{\|\triangle\underline{\mathbf{x}}\|_V}{\|\underline{\mathbf{x}}\|_V}.$$

Wegen $\underline{\mathbf{A}} \cdot \underline{\mathbf{x}} = \underline{\mathbf{b}}$ folgt

$$\|\underline{\mathbf{b}}\|_V = \|\underline{\mathbf{A}} \cdot \underline{\mathbf{x}}\|_V \leq \|\underline{\mathbf{A}}\|_M \cdot \|\underline{\mathbf{x}}\|_V,$$

das heißt

$$\|\underline{\mathbf{x}}\|_V \geq \frac{\|\underline{\mathbf{b}}\|_V}{\|\underline{\mathbf{A}}\|_M}.$$

Also erhält man

$$\begin{aligned} \frac{\|\triangle\underline{\mathbf{x}}\|_V}{\|\underline{\mathbf{x}}\|_V} &\leq \frac{\|\underline{\mathbf{A}}^{-1}\|_M \cdot \|\triangle\underline{\mathbf{b}}\|_V}{\frac{\|\underline{\mathbf{b}}\|_V}{\|\underline{\mathbf{A}}\|_M}} \\ &= \underbrace{\|\underline{\mathbf{A}}^{-1}\|_M \cdot \|\underline{\mathbf{A}}\|_M}_{\kappa(\underline{\mathbf{A}})} \cdot \frac{\|\triangle\underline{\mathbf{b}}\|_V}{\|\underline{\mathbf{b}}\|_V}. \end{aligned}$$

Der Bruch beschreibt dabei die relative Änderung der rechten Seite.

Satz 2.7 (Einfluss der Abänderung der rechten Seite)

Bei Änderung der rechten Seite $\underline{\mathbf{b}}$ eines LGS $\underline{\mathbf{A}} \cdot \underline{\mathbf{x}} = \underline{\mathbf{b}}$ um $\triangle\underline{\mathbf{b}}$ erhalten wir die Abschätzung

$$\frac{\|\triangle\underline{\mathbf{x}}\|_V}{\|\underline{\mathbf{x}}\|_V} \leq \kappa(\underline{\mathbf{A}}) \cdot \frac{\|\triangle\underline{\mathbf{b}}\|_V}{\|\underline{\mathbf{b}}\|_V}$$

für die relative Änderung des Lösungsvektors $\underline{\mathbf{x}}$.

Wir untersuchen jetzt den Einfluss der Änderung der Matrix $\underline{\mathbf{A}}$ auf die Lösung $\underline{\mathbf{x}}$:

Lemma 2.8:

Ist $\underline{\mathbf{F}}$ eine $n \times n$-Matrix mit der Operatornorm

$$\|\underline{\mathbf{F}}\|_M < 1,$$

so existiert die Matrix

$$(\underline{\mathbf{I}} + \underline{\mathbf{F}})^{-1}$$

und es gilt

$$\|\,(\underline{\mathbf{I}} + \underline{\mathbf{F}})^{-1}\,\|_M \leq \frac{1}{1 - \|\underline{\mathbf{F}}\|_M}.$$

Beweis:

Wir wenden die Dreiecksungleichung nach unten an, diese lautet für $\underline{\mathbf{x}}$, $\underline{\mathbf{y}} \in \mathbb{R}^n$

$$\left| \|\underline{\mathbf{x}}\|_V - \|\underline{\mathbf{y}}\|_V \right| \leq \|\underline{\mathbf{x}} - \underline{\mathbf{y}}\|_V .$$

Man erhält so folgende Abschätzung (beachte: aus $\|\underline{\mathbf{F}}\|_M < 1$ folgt $\|\underline{\mathbf{F}} \cdot \underline{\mathbf{x}}\|_V \leq \|\underline{\mathbf{x}}\|_V$):

$$\begin{aligned} \|(\underline{\mathbf{I}} + \underline{\mathbf{F}}) \cdot \underline{\mathbf{x}}\|_V &= \|\underline{\mathbf{x}} + \underline{\mathbf{F}} \cdot \underline{\mathbf{x}}\|_V \geq \|\underline{\mathbf{x}}\|_V - \|\underline{\mathbf{F}} \cdot \underline{\mathbf{x}}\|_V \\ &\geq \|\underline{\mathbf{x}}\|_V - \|\underline{\mathbf{F}}\|_M \cdot \|\underline{\mathbf{x}}\|_V = \underbrace{(1 - \|\underline{\mathbf{F}}\|_M)}_{>0} \cdot \|\underline{\mathbf{x}}\|_V > 0 \qquad (\underline{\mathbf{x}} \neq \underline{\mathbf{0}}), \end{aligned}$$

das heißt, das LGS

$$(\underline{\mathbf{I}} + \underline{\mathbf{F}}) \cdot \underline{\mathbf{x}} = \mathbf{0}$$

besitzt nur die triviale (und damit die eindeutige) Lösung $\underline{\mathbf{x}} = \underline{\mathbf{0}}$. Das bedeutet aber, dass

$$(\underline{\mathbf{I}} + \underline{\mathbf{F}})^{-1}$$

existiert. Wir setzen jetzt zur Abkürzung $\underline{\mathbf{C}} := (\underline{\mathbf{I}} + \underline{\mathbf{F}})^{-1}$. Dann folgt (wie oben)

$$\begin{aligned} 1 &= \|\underline{\mathbf{I}}\|_M = \|(\underline{\mathbf{I}} + \underline{\mathbf{F}}) \cdot \underline{\mathbf{C}}\|_M = \|\underline{\mathbf{C}} + \underline{\mathbf{F}} \cdot \underline{\mathbf{C}}\|_M \\ &\underset{\triangle - Ungl.n.u.}{\geq} \|\underline{\mathbf{C}}\|_M - \|\underline{\mathbf{F}} \cdot \underline{\mathbf{C}}\|_M \geq \|\underline{\mathbf{C}}\|_M - \|\underline{\mathbf{F}}\|_M \cdot \|\underline{\mathbf{C}}\|_M = (1 - \|\underline{\mathbf{F}}\|_M) \cdot \|\underline{\mathbf{C}}\|_M . \end{aligned}$$

Damit ergibt sich durch Division auf beiden Seiten durch den (positiven!) Klammerausdruck

$$\|\underline{\mathbf{C}}\|_M \leq \frac{1}{1 - \|\underline{\mathbf{F}}\|_M}$$

und damit die Behauptung. □

Satz 2.9: (Einfluss der Änderung von $\underline{\mathbf{A}}$)

Es sei $\underline{\mathbf{A}}$ eine invertierbare $n \times n$-Matrix und

$$\underline{\mathbf{B}} = \underline{\mathbf{A}} \cdot (\underline{\mathbf{I}} + \underline{\mathbf{F}}) , \qquad \|\underline{\mathbf{F}}\|_M < 1,$$

die gestörte Matrix. Weiter seien $\underline{\mathbf{x}}$ und $\triangle\underline{\mathbf{x}}$ definiert durch

$$\underline{\mathbf{A}} \cdot \underline{\mathbf{x}} = \underline{\mathbf{b}}, \qquad \underline{\mathbf{B}} \cdot \tilde{\underline{\mathbf{x}}} = \underline{\mathbf{b}}, \qquad \tilde{\underline{\mathbf{x}}} = \underline{\mathbf{x}} + \triangle\underline{\mathbf{x}}.$$

Dann gilt die Abschätzung

$$\frac{\|\triangle\underline{\mathbf{x}}\|_V}{\|\underline{\mathbf{x}}\|_V} \leq \frac{\|\underline{\mathbf{F}}\|_M}{1 - \|\underline{\mathbf{F}}\|_M}.$$

Falls $\kappa(\underline{\mathbf{A}}) \cdot \dfrac{\|\underline{\mathbf{B}} - \underline{\mathbf{A}}\|_M}{\|\underline{\mathbf{A}}\|_M} < 1$ gilt, so ist

$$\frac{\|\triangle\underline{\mathbf{x}}\|_V}{\|\underline{\mathbf{x}}\|_V} \leq \frac{\kappa(\underline{\mathbf{A}}) \cdot \delta}{1 - \kappa(\underline{\mathbf{A}}) \cdot \delta}$$

mit $\delta = \dfrac{\|\underline{\mathbf{B}} - \underline{\mathbf{A}}\|_M}{\|\underline{\mathbf{A}}\|_M}$.

Beweis:

Wegen Lemma 2.8 existiert $\underline{\mathbf{B}}^{-1}$ und es gilt

$$\underline{\mathbf{B}}(\underline{\mathbf{x}} + \triangle\underline{\mathbf{x}}) = \underline{\mathbf{b}} \qquad \Longleftrightarrow \qquad \underline{\mathbf{B}}\triangle\underline{\mathbf{x}} = \underline{\mathbf{b}} - \underline{\mathbf{B}}\underline{\mathbf{x}},$$

das heißt, es ist

$$\begin{aligned} \triangle\underline{\mathbf{x}} &= \underline{\mathbf{B}}^{-1}(\underline{\mathbf{b}} - \underline{\mathbf{B}} \cdot \underline{\mathbf{x}}) = \underline{\mathbf{B}}^{-1}(\underline{\mathbf{A}} \cdot \underline{\mathbf{x}} - \underline{\mathbf{B}} \cdot \underline{\mathbf{x}}) \\ &= \underline{\mathbf{B}}^{-1}(\underline{\mathbf{A}} - \underline{\mathbf{B}}) \cdot \underline{\mathbf{x}} = \underline{\mathbf{B}}^{-1}(\underline{\mathbf{A}} - \underline{\mathbf{B}}) \cdot \underline{\mathbf{A}}^{-1}\underline{\mathbf{b}}. \end{aligned}$$

Da $\|\cdot\|_M$ die Operatornorm zur Vektornorm $\|\cdot\|_V$ ist, gilt die Abschätzung

$$\begin{aligned} \frac{\|\triangle\underline{\mathbf{x}}\|_V}{\|\underline{\mathbf{x}}\|_V} &= \frac{\|\underline{\mathbf{B}}^{-1}(\underline{\mathbf{A}} - \underline{\mathbf{B}}) \cdot \underline{\mathbf{A}}^{-1}\underline{\mathbf{b}}\|_V}{\|\underline{\mathbf{A}}^{-1}\underline{\mathbf{b}}\|_V} \\ &\leq \|\underline{\mathbf{B}}^{-1}(\underline{\mathbf{A}} - \underline{\mathbf{B}})\|_M \cdot \frac{\|\underline{\mathbf{A}}^{-1}\underline{\mathbf{b}}\|_V}{\|\underline{\mathbf{A}}^{-1}\underline{\mathbf{b}}\|_V} \\ &= \|\left[\underline{\mathbf{A}} \cdot (\underline{\mathbf{I}} + \underline{\mathbf{F}})\right]^{-1}(\underline{\mathbf{A}} - \underline{\mathbf{A}} - \underline{\mathbf{A}} \cdot \underline{\mathbf{F}})\|_M \\ &= \|(\underline{\mathbf{I}} + \underline{\mathbf{F}})^{-1}\underline{\mathbf{A}}^{-1}(-\underline{\mathbf{A}} \cdot \underline{\mathbf{F}})\|_M \\ &= |-1| \cdot \|(\underline{\mathbf{I}} + \underline{\mathbf{F}})^{-1}\underline{\mathbf{F}}\|_M \\ &\overset{Konsistenz}{\leq} \|(\underline{\mathbf{I}} + \underline{\mathbf{F}})^{-1}\|_M \cdot \|\underline{\mathbf{F}}\|_M \leq \frac{\|\underline{\mathbf{F}}\|_M}{1 - \|\underline{\mathbf{F}}\|_M}, \end{aligned}$$

wobei im letzten Schritt Lemma 2.8 benutzt wurde. Wegen

$$\underline{\mathbf{F}} = \underline{\mathbf{A}}^{-1}(\underline{\mathbf{B}} - \underline{\mathbf{A}})$$

folgt schließlich

$$\|\underline{\mathbf{F}}\|_M \leq \|\underline{\mathbf{A}}^{-1}\|_M \|\underline{\mathbf{B}} - \underline{\mathbf{A}}\|_M = \underbrace{\|\underline{\mathbf{A}}^{-1}\|_M \|\underline{\mathbf{A}}\|_M}_{\kappa(\underline{\mathbf{A}})} \cdot \underbrace{\frac{\|\underline{\mathbf{B}} - \underline{\mathbf{A}}\|_M}{\|\underline{\mathbf{A}}\|_M}}_{=:\delta}.$$

Für $\kappa(\underline{\mathbf{A}}) \cdot \delta < 1$ gilt dann

$$\frac{\|\underline{\mathbf{F}}\|_M}{1 - \|\underline{\mathbf{F}}\|_M} \leq \frac{\kappa(\underline{\mathbf{A}}) \cdot \delta}{1 - \kappa(\underline{\mathbf{A}}) \cdot \delta}$$

und damit die Behauptung. □

Beispiel:

Es sei das LGS

$$\underbrace{\begin{bmatrix} 1 & 2 \\ 2 & 3,999 \end{bmatrix}}_{:=\underline{\mathbf{A}}} \cdot \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} = \begin{bmatrix} 3 \\ 4 \end{bmatrix}$$

gegeben. Dann gilt

$$\|\underline{\mathbf{A}}\|_\infty = 5,999$$

und als Inverse erhält man

$$\underline{\mathbf{A}}^{-1} = -\frac{1}{0,001} \cdot \begin{bmatrix} 3,999 & -2 \\ -2 & 1 \end{bmatrix} = \begin{bmatrix} -3999 & 2000 \\ 2000 & -1000 \end{bmatrix},$$

was wiederum

$$\|\underline{\mathbf{A}}^{-1}\|_\infty = 5999$$

liefert. Daraus ergibt sich bezüglich der Zeilensummennorm $\|\cdot\|_\infty$ die Konditionszahl

$$\kappa(\underline{\mathbf{A}}) \;=\; 35988,001,$$

das heißt schließlich aber, dass die Bedingung

$$\kappa(\underline{\mathbf{A}}) \cdot \frac{\|\underline{\mathbf{B}} - \underline{\mathbf{A}}\|_\infty}{\|\underline{\mathbf{A}}\|_\infty} \;<\; 1$$

nur für sehr kleine Störungen $\underline{\mathbf{B}}$ von $\underline{\mathbf{A}}$ erfüllt ist. Setzt man zum Beispiel

$$\underline{\mathbf{B}} = \begin{bmatrix} 1 & 2 \\ 2 & 3,9991 \end{bmatrix} \qquad \Longrightarrow \qquad \underline{\mathbf{B}} - \underline{\mathbf{A}} = \begin{bmatrix} 0 & 0 \\ 0 & 0,0001 \end{bmatrix},$$

so erhält man

$$\|\underline{\mathbf{B}} - \underline{\mathbf{A}}\|_\infty \;=\; 0,0001.$$

Setzt man die gewonnenen Werte ein, so ergibt sich

$$\kappa(\underline{\mathbf{A}}) \cdot \delta = \kappa(\underline{\mathbf{A}}) \cdot \frac{\|\underline{\mathbf{B}} - \underline{\mathbf{A}}\|}{\|\underline{\mathbf{A}}\|} \;=\; \frac{3,5988001}{5,999} \;=\; 0,5999 \;<\; 1.$$

Man erhält dann aus

$$\frac{\|\triangle\underline{\mathbf{x}}\|_\infty}{\|\underline{\mathbf{x}}\|_\infty} \leq \frac{\kappa(\underline{\mathbf{A}}) \cdot \delta}{1 - \kappa(\underline{\mathbf{A}}) \cdot \delta} = \frac{0,5999}{1 - 0,5999} = 1,499375$$

eine sehr grobe Abschätzung. Diese ist völlig unakzeptabel, da die Fehlernorm $\|\Delta\underline{\mathbf{x}}\|_\infty$ größer sein kann als $\|\underline{\mathbf{x}}\|_\infty$. Vergleicht man diesen Wert jetzt mit

$$\underline{\mathbf{F}} \;=\; \underline{\mathbf{A}}^{-1}(\underline{\mathbf{B}} - \underline{\mathbf{A}}) \;=\; \begin{bmatrix} -3999 & 2000 \\ 2000 & -1000 \end{bmatrix} \begin{bmatrix} 0 & 0 \\ 0 & 0,0001 \end{bmatrix} = \begin{bmatrix} 0 & 0,2 \\ 0 & -0,1 \end{bmatrix}, \quad \|\underline{\mathbf{F}}\|_\infty \;=\; 0,2,$$

so ergibt sich

$$\frac{\|\triangle\underline{\mathbf{x}}\|_\infty}{\|\underline{\mathbf{x}}\|_\infty} \;=\; \frac{\|\underline{\mathbf{F}}\|_\infty}{1 - \|\underline{\mathbf{F}}\|_\infty} \;=\; \frac{0,2}{0,8} \;=\; 0,25.$$

Tatsächlich folgt im speziellen Beispiel

$$\underline{\mathbf{B}} \cdot \tilde{\underline{\mathbf{x}}} \;=\; \begin{bmatrix} 3 \\ 4 \end{bmatrix},$$

dass

$$\tilde{x}_1 \;=\; -4441,444... \qquad \tilde{x}_2 \;=\; 2222,222...$$

gilt, andererseits erhält man aus

$$\underline{\mathbf{A}} \cdot \underline{\mathbf{x}} \;=\; \begin{bmatrix} 3 \\ 4 \end{bmatrix},$$

die Lösungen

$$x_1 \;=\; -3997 \qquad x_2 \;=\; 2000.$$

Damit ist

$$\|\tilde{\underline{\mathbf{x}}} - \underline{\mathbf{x}}\|_\infty \;=\; 444,444..., \qquad \|\underline{\mathbf{x}}\|_\infty \;=\; 3997, \qquad \frac{\|\triangle\underline{\mathbf{x}}\|_\infty}{\|\underline{\mathbf{x}}\|_\infty} \;=\; 0,1111945.$$

2.3 Dreiecksmatrizen

Definition 2.10:

Eine Matrix

$$\underline{\mathbf{A}} = (a_{jk})_{j,k=1}^{n} \in \mathbb{R}^{n\times n}$$

heißt ***untere (obere) Dreiecksmatrix,*** wenn $a_{jk} = 0$ gilt für $j < k$ $(j > k)$.

Das LGS für eine untere Dreiecksmatrix $\underline{\mathbf{L}} = (l_{jk})_{j,k=1}^{n}$ lautet dann

$$\begin{array}{ccccccc} l_{11}x_1 & & & & & = & b_1 \\ l_{21}x_1 & + & l_{22}x_2 & & & = & b_2 \\ \vdots & & & \ddots & & & \vdots \\ l_{n1}x_1 & + & \cdots & \cdots & +l_{nn}x_n & = & b_n \end{array}$$

wobei $l_{jj} \neq 0$ gelte für alle $j = 1, ..., n$. Die Lösung dieses LGS ergibt sich durch die sogenannte ***Vorwärtselimination*** , es ergibt sich sukzessive

$$x_1 = \frac{b_1}{l_{11}}, \qquad x_2 = \frac{(b_2 - l_{21}x_1)}{l_{22}}, ...,$$

und schließlich

$$x_n = \frac{(b_n - l_{n1}x_1 - l_{n2}x_2 - \cdots - l_{n(n-1)}x_{n-1})}{l_{nn}},$$

das heißt, man erhält die Lösungsformel

$$x_j = \frac{\left(b_j - \sum_{k=1}^{j-1} l_{jk}x_k\right)}{l_{jj}}, \qquad j = 1, \ldots, n.$$

Analog erhält man für obere Dreiecksmatrizen $\underline{\mathbf{U}} = (u_{jk})_{j,k=1}^{n}$ das LGS

$$\begin{array}{ccccccccc} u_{11}x_1 & + & u_{12}x_2 & + & \cdots & + & u_{1n}x_n & = & b_1 \\ & & u_{22}x_2 & + & \cdots & + & u_{2n}x_n & = & b_2 \\ & & & & \ddots & & \vdots & & \vdots \\ & & & & & & u_{nn}x_n & = & b_n \end{array}$$

und daraus die Lösungsformel (***Rückwärtselimination***)

$$x_j = \frac{\left(b_j - \sum_{k=j+1}^{n} u_{jk}x_k\right)}{u_{jj}}, \qquad j = 1, \ldots, n.$$

Man kann die Vorwärtselimination mittels der folgenden MAPLE-Prozedur durchführen:

```
> #  Vorwärtselimination, überschreibt b
> with(LinearAlgebra);
> velim:= proc(L::Matrix,b::Vector)
  #  untere Dreiecksmatrix L, rechte Seite b
  local j, k, n;
  n:=Dimension(b);
  for j from 1 to n do
      for k from 1 to j − 1 do b[j] := b[j] − L[j, k] ∗ b[k] end do;
```

```
    b[j] := b[j]/L[j, j]
end do;
evalm(b);
end proc;
```

Die gesuchten Lösungen x_j sind nun auf $b[j]$, $j = 1, \ldots, n$, abgespeichert. Die hierbei benutzte Rechenvorschrift entspricht der oben angeführten Formel für die Vorwärtselimination. Der Aufruf der Prozedur erfolgt in der Form

```
> velim(L, b);
```

wobei zuvor natürlich die Matrix

```
> L:=Matrix([...])
```

und der Vektor

```
> b:=Vector([...]);
```

definiert sein müssen. Der Rechenaufwand der Vorwärtselimination umfasst im j-ten Schritt $j-1$ Multiplikationen, $j-1$ Additionen und eine Division, also insgesamt $2j-1$ arithmetische Operationen (***flops = floating point operations***). Der gesamte Vorgang (alle Schritte zusammen) erfordert dann

$$\sum_{j=1}^{n}(2j-1) \;=\; n(n+1) - n \;=\; n^2$$

flops. Analog lässt sich der Rechenaufwand der Rückwärtselimination bestimmen, auch hier ergibt sich ein Aufwand von n^2 flops.

2.4 Gauß-Elimination

Ziel der Gauß-Elimination ist es, die Lösung eines beliebigen LGS auf die Lösung von Dreieckssystemen zurückzuführen. Wir suchen also obere und untere Dreiecksmatrizen $\underline{\mathbf{U}}, \underline{\mathbf{L}}$ (die Bezeichnung leitet sich von den englischen Begriffen ***upper, lower*** ab) so, dass

$$\underline{\mathbf{A}} \;=\; \underline{\mathbf{L}} \cdot \underline{\mathbf{U}}$$

gilt und damit für das ursprüngliche LGS

$$\underline{\mathbf{A}} \cdot \underline{\mathbf{x}} = \underline{\mathbf{b}} \quad \Longleftrightarrow \quad \underline{\mathbf{L}} \cdot \underbrace{\underline{\mathbf{U}} \cdot \underline{\mathbf{x}}}_{:=\underline{\mathbf{y}}} = \underline{\mathbf{b}}$$

gilt. Somit ergibt sich die Lösung des Ausgangssystems durch Lösung der beiden Systeme

$$\underline{\mathbf{L}} \cdot \underline{\mathbf{y}} = \underline{\mathbf{b}} \qquad \underline{\mathbf{U}} \cdot \underline{\mathbf{x}} = \underline{\mathbf{y}}$$

in dieser Reihenfolge, also durch jeweils eine Vorwärts- und eine Rückwärtselimination.

Bemerkung:

Der Vorteil der $\underline{\mathbf{L}} \cdot \underline{\mathbf{U}}$-Zerlegung ist folgender: Muss das LGS $\underline{\mathbf{A}} \cdot \underline{\mathbf{x}} = \underline{\mathbf{b}}$ für verschiedene

rechte Seiten gelöst werden, so muss die $\underline{\mathbf{L}} \cdot \underline{\mathbf{U}}$-Zerlegung nur einmal durchgeführt werden, danach sind nur noch Dreieckssysteme zu lösen.

Beispiel:

Die Eliminationsschritte beim Gauß-Algorithmus lassen sich durch Multiplikation mit geeigneten Matrizen darstellen:

$$\left.\begin{array}{cccc|cccc} & & & & 1 & -2 & -1 & 0 \\ & & & & 2 & -6 & 1 & -5 \\ & & & & 3 & -6 & 1 & -7 \\ & & & & 1 & 2 & -3 & 4 \end{array}\right. = \underline{\mathbf{A}}$$

$$\underline{\mathbf{M}}_1 = \left.\begin{array}{cccc|cccc} 1 & 0 & 0 & 0 & 1 & -2 & -1 & 0 \\ -2 & 1 & 0 & 0 & 0 & -2 & 3 & -5 \\ -3 & 0 & 1 & 0 & 0 & 0 & 4 & -7 \\ -1 & 0 & 0 & 1 & 0 & 4 & -2 & 4 \end{array}\right. = \underline{\mathbf{M}}_1 \cdot \underline{\mathbf{A}}$$

$$\underline{\mathbf{M}}_2 = \left.\begin{array}{cccc|cccc} 1 & 0 & 0 & 0 & 1 & -2 & -1 & 0 \\ 0 & 1 & 0 & 0 & 0 & -2 & 3 & -5 \\ 0 & 0 & 1 & 0 & 0 & 0 & 4 & -7 \\ 0 & 2 & 0 & 1 & 0 & 0 & 4 & -6 \end{array}\right. = \underline{\mathbf{M}}_2 \cdot \underline{\mathbf{M}}_1 \cdot \underline{\mathbf{A}}$$

$$\underline{\mathbf{M}}_3 = \left.\begin{array}{cccc|cccc} 1 & 0 & 0 & 0 & 1 & -2 & -1 & 0 \\ 0 & 1 & 0 & 0 & 0 & -2 & 3 & -5 \\ 0 & 0 & 1 & 0 & 0 & 0 & 4 & -7 \\ 0 & 0 & -1 & 1 & 0 & 0 & 0 & 1 \end{array}\right. = \underline{\mathbf{M}}_3 \cdot \underline{\mathbf{M}}_2 \cdot \underline{\mathbf{M}}_1 \cdot \underline{\mathbf{A}}$$

Damit ist

$$\underline{\mathbf{U}} := \underline{\mathbf{M}}_3 \cdot \underline{\mathbf{M}}_2 \cdot \underline{\mathbf{M}}_1 \cdot \underline{\mathbf{A}}$$

eine obere Dreiecksmatrix. Dann ist aber auch

$$\underline{\mathbf{M}}_1^{-1} \cdot \underline{\mathbf{M}}_2^{-1} \cdot \underline{\mathbf{M}}_3^{-1} \cdot \underline{\mathbf{U}} = \underline{\mathbf{A}}$$

(alle Matrizen sind invertierbar) und es stellt sich die Frage, ob nicht die Matrix

$$\underline{\mathbf{M}}_1^{-1} \cdot \underline{\mathbf{M}}_2^{-1} \cdot \underline{\mathbf{M}}_3^{-1}$$

eine untere Dreiecksmatrix ist, denn dann könnte man die in der $\underline{\mathbf{L}} \cdot \underline{\mathbf{U}}$-Zerlegung gesuchten Matrizen auf diese Art und Weise leicht bestimmen.

Definition 2.11:

Eine ***Gauß-Transformation (Eliminationsmatrix)*** ist eine Matrix der Form

$$\underline{\mathbf{M}}_k = \underline{\mathbf{I}} - \underline{\mathbf{y}}_k \cdot \underline{\mathbf{e}}_k^T$$

mit

$$\underline{\mathbf{y}}_k = [\,0 \;\cdots\; 0\; y_{k+1} \;\cdots\; y_n\,]^T$$

und dem k-tem Einheitsvektor $\underline{\mathbf{e}}_k$, das heißt,

$$\underline{\mathbf{M}}_k = \begin{bmatrix} 1 & & & & & 0 \\ & \ddots & & & & \\ & & 1 & & & \\ & & & \ddots & & \\ & & -y_{k+1} & & & \\ & & \vdots & & \ddots & \\ 0 & & -y_n & & & 1 \end{bmatrix}.$$

Die von der Einheitsmatrix abweichende Spalte ist also jeweils die k-te.

Bemerkung:

1) Die Inverse der Gauß-Transformation $\underline{\mathbf{I}} - \underline{\mathbf{y}}_k \cdot \underline{\mathbf{e}}_k^T$ ist gegeben durch

$$\underline{\mathbf{I}} + \underline{\mathbf{y}}_k \cdot \underline{\mathbf{e}}_k^T,$$

denn es gilt

$$\left(\underline{\mathbf{I}} - \underline{\mathbf{y}}_k \cdot \underline{\mathbf{e}}_k^T\right)\left(\underline{\mathbf{I}} + \underline{\mathbf{y}}_k \cdot \underline{\mathbf{e}}_k^T\right) = \underline{\mathbf{I}} + \underline{\mathbf{y}}_k \cdot \underline{\mathbf{e}}_k^T - \underline{\mathbf{y}}_k \cdot \underline{\mathbf{e}}_k^T - \underline{\mathbf{y}}_k \underbrace{\underline{\mathbf{e}}_k^T \underline{\mathbf{y}}_k}_{=0} \underline{\mathbf{e}}_k^T = \underline{\mathbf{I}}.$$

2) Für eine Matrix mit den Zeilenvektoren $\underline{\mathbf{a}}_1, \ldots \underline{\mathbf{a}}_n$,

$$\underline{\mathbf{A}} = \begin{bmatrix} \underline{\mathbf{a}}_1^T \\ \vdots \\ \underline{\mathbf{a}}_n^T \end{bmatrix},$$

folgt

$$\underline{\mathbf{M}}_k \cdot \underline{\mathbf{A}} = \begin{bmatrix} \underline{\mathbf{a}}_1^T \\ \vdots \\ \underline{\mathbf{a}}_k^T \\ \underline{\mathbf{a}}_{k+1}^T - y_{k+1} \cdot \underline{\mathbf{a}}_k^T \\ \vdots \\ \underline{\mathbf{a}}_n^T - y_n \cdot \underline{\mathbf{a}}_k^T \end{bmatrix},$$

das heißt, von den Zeilen $\underline{\mathbf{a}}_j^T,\ j = k+1, ..., n$, wird das y_j-fache der k-ten Zeile $\underline{\mathbf{a}}_k^T$ abgezogen.

Definition 2.12:

Für eine Matrix $\underline{\mathbf{A}} \in \mathbb{R}^{n \times n}$ heißt

$$\underline{\mathbf{A}}_m = (a_{jk})_{j,k=1}^m \qquad (m = 1, \ldots, n)$$

m***-te Hauptminore***.

Satz 2.13:

Für eine Matrix $\underline{\mathbf{A}} \in \mathbb{R}^{n \times n}$ seien alle Hauptminoren invertierbar, das heißt, es sei

$$\det\left(\underline{\mathbf{A}}_m\right) \neq 0$$

für $m = 1, ..., n$. Dann gibt es eine untere Dreiecksmatrix $\underline{\mathbf{L}} = (l_{ij})_{i,j=1}^n$ und eine obere Dreiecksmatrix $\underline{\mathbf{U}} = (u_{jk})_{j,k=1}^n$ so, dass

$$\underline{\mathbf{A}} = \underline{\mathbf{L}} \cdot \underline{\mathbf{U}}$$

gilt und zusätzlich für die Diagonalemelente der unteren Dreiecksmatrix

$$l_{jj} = 1, \qquad (j = 1, ..., n)$$

gilt.

Beweis:

Idee: Finde die Gauß-Transformationen $\underline{\mathbf{M}}_j,\ j = 1, ..., n-1$, so dass

$$\underline{\mathbf{M}}_{n-1} \cdot \underline{\mathbf{M}}_{n-2} \cdots \underline{\mathbf{M}}_1 \cdot \underline{\mathbf{A}} = \underline{\mathbf{U}}$$

eine obere Dreiecksmatrix ist. Die Zerlegung lautet dann

$$\underline{\mathbf{A}} = \underbrace{\underline{\mathbf{M}}_1^{-1} \cdot \underline{\mathbf{M}}_2^{-1} \cdots \underline{\mathbf{M}}_{n-1}^{-1}}_{:=\underline{\mathbf{L}}} \cdot \underline{\mathbf{U}}.$$

Den Beweis führt man durch induktive Konstruktion:

1. Schritt:

Betrachtet man eine Matrix

$$\underline{\mathbf{A}} = \begin{bmatrix} a_{11} & \cdots & \cdots & a_{1n} \\ a_{21} & \ddots & & \vdots \\ \vdots & & \ddots & \vdots \\ a_{n1} & \cdots & \cdots & a_{nn} \end{bmatrix},$$

so erzeugt man in der ersten Spalte Nullen, indem man die erste Zeile mit $-\frac{a_{j1}}{a_{11}}$ multipliziert und zur j-ten Zeile addiert. (Hier ist $a_{11} = \det(\underline{\mathbf{A}}_1) \neq 0$.) Damit ergibt sich die erste Gauß-Transformation

$$\underline{\mathbf{M}}_1 = \begin{bmatrix} 1 & & & 0 \\ -\frac{a_{21}}{a_{11}} & \ddots & & \\ \vdots & & \ddots & \\ -\frac{a_{n1}}{a_{11}} & & & 1 \end{bmatrix},$$

und es gilt

$$\underline{\mathbf{M}}_1 \cdot \underline{\mathbf{A}} = \begin{bmatrix} a_{11} & a_{12} & \cdots & a_{1n} \\ 0 & b_{22} & \cdots & b_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ 0 & b_{n2} & \cdots & b_{nn} \end{bmatrix}$$

mit

$$b_{jk} = a_{jk} - \frac{a_{j1}}{a_{11}} \cdot a_{1k} \qquad (j, k = 2, ..., n).$$

Insbesondere ist

$$b_{22} = a_{22} - \frac{a_{21}}{a_{11}} \cdot a_{12} = \frac{1}{a_{11}} \cdot \underbrace{(a_{11} \cdot a_{22} - a_{21} \cdot a_{12})}_{\det(\underline{\mathbf{A}}_2) \neq 0} \neq 0.$$

Die Elimination lässt sich in entsprechender Weise fortsetzen. Im $(m-1)$-ten Schritt erhält man so

$$\underline{\mathbf{M}}_{m-1} \cdot \underline{\mathbf{M}}_{m-2} \cdots \underline{\mathbf{M}}_1 \cdot \underline{\mathbf{A}} = \left[\begin{array}{c|c} \underline{\mathbf{U}}_{m-1} & \underline{\mathbf{B}} \\ \hline \underline{\mathbf{0}} & \underline{\mathbf{C}} \end{array}\right]$$

mit

$$\underline{\mathbf{U}}_{m-1} \in \mathbb{R}^{(m-1)\times(m-1)}, \quad \underline{\mathbf{0}} \in \mathbb{R}^{(n-m+1)\times(m-1)},$$
$$\underline{\mathbf{B}} \in \mathbb{R}^{(m-1)\times(n-m+1)}, \quad \underline{\mathbf{C}} \in \mathbb{R}^{(n-m+1)\times(n-m+1)}.$$

Dabei ist $\underline{\mathbf{U}}_{m-1}$ von der Gestalt

$$\underline{\mathbf{U}}_{m-1} = \begin{bmatrix} u_{11} & \cdots & \cdots & u_{1,m-1} \\ 0 & \ddots & & \vdots \\ \vdots & \ddots & \ddots & \vdots \\ 0 & \cdots & 0 & u_{m-1,m-1} \end{bmatrix},$$

und es gilt $u_{jj} \neq 0$ $(j = 1, ..., m-1)$. Sei nun $\underline{\mathbf{b}}_1$ die erste Spalte von $\underline{\mathbf{B}}$. Da für die Hauptminore $\underline{\mathbf{A}}_m$ die Voraussetzung $\det(\underline{\mathbf{A}}_m) \neq 0$ erfüllt ist, folgt

$$\det(\underline{\mathbf{A}}_m) = \det \left[\begin{array}{c|c} \underline{\mathbf{U}}_{m-1} & \underline{\mathbf{b}}_1 \\ \hline \underline{\mathbf{0}}^T & c_{11} \end{array} \right] = \underbrace{\det(\underline{\mathbf{U}}_{m-1})}_{\neq 0} \cdot c_{11} \neq 0,$$

denn letztere Matrix ist aus $\underline{\mathbf{A}}_m$ nur durch Addieren von Vielfachen von Zeilen hervorgegangen. Also ist $c_{11} \neq 0$. Wählt man dann

$$\underline{\mathbf{M}}_m = \begin{bmatrix} 1 & & & & \\ & \ddots & & & \\ & & 1 & & \\ & & -\frac{c_{21}}{c_{11}} & \ddots & \\ & & \vdots & & \ddots \\ & & -\frac{c_{n1}}{c_{11}} & & 1 \end{bmatrix},$$

so liefert dies

$$\underline{\mathbf{M}}_m \cdot \underline{\mathbf{M}}_{m-1} \cdots \underline{\mathbf{M}}_1 \cdot \underline{\mathbf{A}} = \left[\begin{array}{c|c} \underline{\mathbf{U}}_m & \tilde{\underline{\mathbf{B}}} \\ \hline \underline{\mathbf{0}} & \tilde{\underline{\mathbf{C}}} \end{array} \right],$$

wobei

$$\underline{\mathbf{U}}_m \in \mathbb{R}^{m \times m}, \quad \tilde{\underline{\mathbf{B}}} \in \mathbb{R}^{m \times (n-m)}, \quad \underline{\mathbf{0}} \in \mathbb{R}^{(n-m) \times m}, \quad \tilde{\underline{\mathbf{C}}} \in \mathbb{R}^{(n-m) \times (n-m)}.$$

Es bleibt nur noch zu zeigen, dass $\underline{\mathbf{M}}_1^{-1} \ldots \underline{\mathbf{M}}_m^{-1}$ für $m = 1, \ldots, n-1$ eine untere Dreiecksmatrix ist. Der Beweis folgt jetzt mit folgendem Lemma 2.14. □

Lemma 2.14:

Für die Eliminationsmatrizen

$$\underline{\mathbf{M}}_j = \underline{\mathbf{I}} - \underline{\mathbf{y}}^{(j)} \cdot \underline{\mathbf{e}}_j^T, \quad j = 1, \ldots, m$$

mit $1 \leq m \leq n-1$ gilt

$$\underline{\mathbf{M}}_1^{-1} \cdot \underline{\mathbf{M}}_2^{-1} \cdots \underline{\mathbf{M}}_m^{-1} = \underline{\mathbf{I}} + \sum_{j=1}^{m} \underline{\mathbf{y}}^{(j)} \cdot \underline{\mathbf{e}}_j^T = \begin{bmatrix} 1 & & & & & & \\ y_2^{(1)} & 1 & & & & & \\ \vdots & y_3^{(2)} & 1 & & & & \\ \vdots & & \vdots & 1 & & & \\ \vdots & \vdots & & y_{m+1}^{(m)} & 1 & & \\ & & & \vdots & & \ddots & \\ y_n^{(1)} & y_n^{(2)} & \cdots & y_n^{(m)} & & & 1 \end{bmatrix}.$$

Beweis:

Wir beweisen das Lemma durch vollständige Induktion über m mit $1 \leq m \leq n-1$. Der Fall $m = 1$ wurde in der Bemerkung nach der Definition 2.11 der Gauß-Transformation bereits behandelt. Für $m < n-1$ folgt der Induktionsschritt :

$$\begin{aligned}
(\underline{\mathbf{M}}_1^{-1} \cdots \underline{\mathbf{M}}_m^{-1})\underline{\mathbf{M}}_{m+1}^{-1} &= \left(\underline{\mathbf{I}} + \sum_{j=1}^{m} \underline{\mathbf{y}}^{(j)} \underline{\mathbf{e}}_j^T\right)\left(\underline{\mathbf{I}} + \underline{\mathbf{y}}^{(m+1)} \underline{\mathbf{e}}_{m+1}^T\right) \\
&= \left(\underline{\mathbf{I}} + \underline{\mathbf{y}}^{(m+1)} \underline{\mathbf{e}}_{m+1}^T + \sum_{j=1}^{m} \underline{\mathbf{y}}^{(j)} \underline{\mathbf{e}}_j^T + \sum_{j=1}^{m} \underline{\mathbf{y}}^{(j)} \underbrace{\underline{\mathbf{e}}_j^T \underline{\mathbf{y}}^{(m+1)}}_{=0} \underline{\mathbf{e}}_{m+1}^T\right) \\
&= \underline{\mathbf{I}} + \sum_{j=1}^{m+1} \underline{\mathbf{y}}^{(j)} \underline{\mathbf{e}}_j^T .
\end{aligned}$$

Damit folgt die Behauptung. □

Bemerkung:

Nach Lemma 2.14 erhalten wir die Matrix $\underline{\mathbf{L}}$, indem wir die zu den jeweiligen Eliminationsmatrizen gehörigen Vektoren $\underline{\mathbf{y}}^{(j)}$ „aufsammeln“. Die Gauß-Elimination kann in MAPLE mittels der folgenden Prozedur realisiert werden:

```
> #   Gauß-Elimination mit Überschreiben (ohne Pivot)
> with(LinearAlgebra);
  gauss1:=proc(A::Matrix)
  local n, m, j, k;
  n:=RowDimension(A);
  for m from 1 to n do
      for j from m + 1 to n do
          A[j, m] := A[j, m]/A[m, m];
          for k from m + 1 to n do
              A[j, k] := A[j, k] − A[j, m] ∗ A[m, k]
          end do
      end do
  end do;
  evalm(A)
  end proc;
```

Der Rechenaufwand beträgt dabei $n - m$ Multiplikationen für die Berechnung der Eliminationsmatrix $\underline{\mathbf{M}}_m$ und $2(n-m)^2$ Operationen für die Multiplikation mit $\underline{\mathbf{M}}_m$, also insgesamt

$$(n-m) + 2(n-m)^2 \;=\; (n-m)(1+2n-2m)$$

Operationen pro Eliminationsschritt. Summiert man nun über alle m auf, so liefert dies

$$\begin{aligned}
&\sum_{m=1}^{n} (n-m)(1+2n-2m) = \sum_{m=1}^{n} (n + 2n^2 - 2mn - m - 2mn + 2m^2) \\
= \;& n^2 + 2n^3 - 4n \cdot \frac{n(n+1)}{2} - \frac{n(n+1)}{2} + \frac{2n(n+1)(2n+1)}{6} \\
= \;& \frac{2}{3}n^3 - \frac{n^2}{2} - \frac{n}{6}.
\end{aligned}$$

Das fertige Lösungsverfahren

Für die Berechnung der Lösung eines linearen Gleichungssystems $\underline{\mathbf{A}} \cdot \underline{\mathbf{x}} = \underline{\mathbf{b}}$ sind folgende Schritte durchzuführen:

1) Bestimmung der Zerlegung $\underline{\mathbf{A}} = \underline{\mathbf{L}} \cdot \underline{\mathbf{U}}$ mittels der Gauß-Transformation.

2) Bestimmung der Lösung $\underline{\mathbf{y}}$ des LGS $\underline{\mathbf{L}} \cdot \underline{\mathbf{y}} = \underline{\mathbf{b}}$ durch Vorwärtselimination.

3) Bestimmung der Lösung $\underline{\mathbf{x}}$ des LGS $\underline{\mathbf{U}} \cdot \underline{\mathbf{x}} = \underline{\mathbf{y}}$ durch Rückwärtselimination.

Bemerkung:

Man kann Vorwärts- und Rückwärtselimination direkt auf die in ***einer*** Matrix gespeicherte $\underline{\mathbf{L}} \cdot \underline{\mathbf{U}}$-Zerlegung anwenden, wenn man die Vorwärtselimination so modifiziert, dass sie alle Diagonalelemente als 1 annimmt (vgl. folgende Zusammenstellung).

Zusammenstellung von Prozeduren:

1) Die Vorwärtselimination wird in der folgenden Weise implementiert:

```
> #   Vorwärtselimination, überschreibt b
> #   Annahme, dass auf der Diagonalen Einsen stehen
> vorelim1:=proc(L::Matrix,b::Vector)
  local n, j, k;
  n:=Dimension(b);
  for j from 1 to n do
      for k from 1 to j − 1 do b[j] := b[j] − L[j, k] ∗ b[k] end do
  end do;
  evalm(b)
  end proc;
```

2) Die Prozedur der Rückwärtselimination lautet:

```
> #   Rückwärtselimination, überschreibt b
> ruecksubs:=proc(U::Matrix,b::Vector)
  local n, j, k;
  n:=Dimension(b);
  for j from n by (−1) to 1 do
    for k from n by (−1) to j + 1 do b[j] := b[j] − U[j, k] ∗ b[k] end do;
    b[j] := b[j]/U[j, j]
  end do;
  evalm(b)
  end proc;
```

3) Die komplette Lösungsprozedur für ein lineares Gleichungssystem (ohne Pivot) lautet dann wie folgt:

```
> loesAxb:= proc(A::Matrix, b::Vector)
  local gauss1, vorelim1, ruecksubs;
  gauss1:=proc(A) ... end proc;
  vorelim1:=proc(L, b) ... end proc;
  ruecksubs:=proc(U, b) ... end proc;
  gauss1(A);
  vorelim1(A, b);
```

```
ruecksubs(A, b);
evalm(b)
end proc;
```

2.5 Pivotsuche

Problem:

1) Die Bedingung, dass alle Hauptminoren einer Matrix $\underline{\mathbf{A}}$ invertierbar sind, ist aufwändig nachzuprüfen und oft gar nicht erfüllt. Wir brauchen aber einen Algorithmus, der für alle invertierbaren Matrizen $\underline{\mathbf{A}}$ anwendbar ist.

2) Bei der Gauß-Elimination werden Vielfache einer Zeile von $\underline{\mathbf{A}}$ zu ″weiter unten liegenden Zeilen″ addiert. Diese Vielfachen können betragsmäßig sehr groß werden, wenn das Diagonalelement sehr klein ist. Um dies zu verhindern, suchen wir eine Strategie, die die Division durch zu kleine Werte verhindert.

Definition 2.15:

Das Element a_{11} einer Matrix $\underline{\mathbf{A}} \in \mathbb{R}^{n\times n}$ heißt ***Pivot-Element.***

Definition 2.16:

Eine Matrix $\underline{\mathbf{P}} = (p_{ij})_{i,j=1}^{n} \in \mathbb{R}^{n\times n}$ heißt ***Permutationsmatrix,*** wenn sie in jeder Spalte und Zeile genau eine Eins und sonst nur Nullen enthält, das heißt, es gilt

$$p_{ij} \in \{0,1\}, \qquad \sum_{j=1}^{n} p_{ij} \;=\; 1, \qquad \sum_{i=1}^{n} p_{ij} \;=\; 1, \qquad (i,j = 1,...,n).$$

Wir bezeichnen die Menge der Permutationsmatrizen mit Π_n. Diese Menge bildet mit der Matrixmultiplikation eine Gruppe. Die ***Vertauschungsmatrix*** $\underline{\mathbf{P}}[j,k],\ j,k = 1,...,n$, ist definiert durch

$$\underline{\mathbf{P}}[j,k] \;=\; \begin{bmatrix} 1 & & & & & & & & \\ & \ddots & & & & & & & \\ & & 0 & \dots & 1 & & & & \\ & & & 1 & & & & & \\ & & \vdots & \ddots & \vdots & & & & \\ & & & & 1 & & & & \\ & & 1 & \dots & 0 & & & & \\ & & & & & 1 & & & \\ & & & & & & \ddots & & \\ & & & & & & & 1 & \end{bmatrix},$$

das heißt, die Diagonalelemente p_{jj} und p_{kk} werden gleich Null gesetzt und die Elemente p_{kj} und p_{jk} gleich Eins.

Bemerkung:

1) Vertauschungsmatrizen sind spezielle Permutationsmatrizen.

2) Die Matrix $\underline{\mathbf{P}}[j,k] \cdot \underline{\mathbf{A}}$ ist die Matrix, bei der die j-te und k-te Zeile von $\underline{\mathbf{A}}$ vertauscht sind. Umgekehrt ist die Matrix $\underline{\mathbf{A}} \cdot \underline{\mathbf{P}}[j,k]$ die Matrix, bei der j-te und k-te Spalte von $\underline{\mathbf{A}}$

vertauscht sind.

Satz 2.17:

Sei $\underline{\mathbf{A}}$ invertierbar. Dann gibt es eine Permutationsmatrix $\underline{\mathbf{P}} \in \Pi_n$, eine untere Dreiecksmatrix $\underline{\mathbf{L}} = (l_{j,k})_{j,k=1}^n \in \mathbb{R}^{n\times n}$ mit $l_{jj} = 1$, $j = 1, \ldots, n$, und eine invertierbare obere Dreiecksmatrix $\underline{\mathbf{U}} \in \mathbb{R}^{n\times n}$, so dass

$$\underline{\mathbf{P}} \cdot \underline{\mathbf{A}} = \underline{\mathbf{L}} \cdot \underline{\mathbf{U}}$$

gilt.

Beweis:

Später.

Folgerung 2.18:

Eine Matrix $\underline{\mathbf{A}} \in \mathbb{R}^{n\times n}$ ist genau dann invertierbar, wenn es eine Permutationsmatrix $\underline{\mathbf{P}} \in \Pi_n$ gibt (das heißt eine Zeilenumordnung von $\underline{\mathbf{A}}$), so dass alle Hauptminoren $(\underline{\mathbf{P}} \cdot \underline{\mathbf{A}})_m$ von $\underline{\mathbf{P}} \cdot \underline{\mathbf{A}}$, $m = 1, ..., n$, invertierbar sind.

Beweis:

Übung.

Bemerkung:

1) Man kann folgendes Lösungsverfahren anwenden: Wegen

$$\underline{\mathbf{A}} \cdot \underline{\mathbf{x}} = \underline{\mathbf{b}} \iff \underline{\mathbf{P}} \cdot \underline{\mathbf{A}} \cdot \underline{\mathbf{x}} = \underline{\mathbf{P}} \cdot \underline{\mathbf{b}} \iff \underline{\mathbf{L}} \cdot \underline{\mathbf{U}} \cdot \underline{\mathbf{x}} = \underline{\mathbf{P}} \cdot \underline{\mathbf{b}}$$

berechnet man zunächst eine Dreieckszerlegung $\underline{\mathbf{L}} \cdot \underline{\mathbf{U}}$ für $\underline{\mathbf{P}} \cdot \underline{\mathbf{A}}$ und löst dann

$$\underline{\mathbf{L}} \cdot \underline{\mathbf{y}} = \underline{\mathbf{P}} \cdot \underline{\mathbf{b}}$$

durch Vorwärtselimination und anschließend

$$\underline{\mathbf{U}} \cdot \underline{\mathbf{x}} = \underline{\mathbf{y}}$$

durch Rückwärtselimination.

2) Satz 2.17 ergibt als Verfahren die Gauß-Elimination mit Zeilenvertauschung, gesteuert durch ***Spalten-Pivotsuche***. Analog ist auch eine ***Zeilen-Pivotsuche*** möglich, das heißt, man betrachtet eine Zerlegung der Form

$$\underline{\mathbf{A}} \cdot \underline{\mathbf{Q}} = \underline{\mathbf{L}} \cdot \underline{\mathbf{U}}, \qquad \underline{\mathbf{Q}} \in \Pi_n,$$

oder eine ***Total-Pivotsuche*** mit einer Zerlegung

$$\underline{\mathbf{P}} \cdot \underline{\mathbf{A}} \cdot \underline{\mathbf{Q}} = \underline{\mathbf{L}} \cdot \underline{\mathbf{U}}, \qquad \underline{\mathbf{P}}, \underline{\mathbf{Q}} \in \Pi_n.$$

Im Falle der Zeilen-Pivotsuche löst man dann (in dieser Reihenfolge)

$$\underline{\mathbf{L}} \cdot \underline{\mathbf{y}} = \underline{\mathbf{b}}, \qquad \underline{\mathbf{U}} \cdot \underline{\mathbf{z}} = \underline{\mathbf{y}}, \qquad \underline{\mathbf{x}} = \underline{\mathbf{Q}} \cdot \underline{\mathbf{z}},$$

im Falle der Total-Pivotsuche (in dieser Reihenfolge)

$$\underline{\mathbf{L}} \cdot \underline{\mathbf{y}} = \underline{\mathbf{P}} \cdot \underline{\mathbf{b}}, \qquad \underline{\mathbf{U}} \cdot \underline{\mathbf{z}} = \underline{\mathbf{y}}, \qquad \underline{\mathbf{x}} = \underline{\mathbf{Q}} \cdot \underline{\mathbf{z}}.$$

3) Aufgrund der Invertierbarkeit von $\underline{\mathbf{A}}$ folgt aus $\underline{\mathbf{P}}\,\underline{\mathbf{A}} = \underline{\mathbf{L}}\,\underline{\mathbf{U}}$

$$0 \neq \det(\underline{\mathbf{A}}) = \det(\underline{\mathbf{P}}^{-1} \cdot \underline{\mathbf{L}} \cdot \underline{\mathbf{U}}) = \underbrace{\frac{1}{\det(\underline{\mathbf{P}})}}_{=1 \text{ oder } -1} \cdot \underbrace{\det(\underline{\mathbf{L}})}_{=1} \cdot \underbrace{\det(\underline{\mathbf{U}})}_{=\prod_{i=1}^{n} u_{ii}},$$

das heißt, es muss

$$u_{ii} \neq 0$$

gelten für $i = 1, ..., n$. Also ist $\underline{\mathbf{U}}$ regulär.

Beweis von Satz 2.17:

Der Beweis verläuft in Analogie zu dem von Satz 2.12. Wir konstruieren Permutationsmatrizen $\underline{\mathbf{P}}_j \in \Pi_n$ und Eliminationsmatrizen $\underline{\mathbf{M}}_j$ $(j = 1, ..., n)$, so dass

$$\begin{aligned} \underline{\mathbf{A}}^{(m)} &= \underline{\mathbf{M}}_m \cdot \underline{\mathbf{P}}_m \cdots \underline{\mathbf{M}}_1 \cdot \underline{\mathbf{P}}_1 \cdot \underline{\mathbf{A}} \\ &= \left[\begin{array}{cccc|c} u_{11} & \cdots & \cdots & u_{1m} & \\ 0 & \ddots & & \vdots & \underline{\mathbf{B}}_m \\ \vdots & \ddots & \ddots & \vdots & \\ 0 & \cdots & 0 & u_{mm} & \\ \hline & \multicolumn{3}{c}{\underline{\mathbf{0}}} & \tilde{\underline{\mathbf{A}}}_m \end{array}\right] \qquad (m = 0, ..., n) \end{aligned}$$

mit $\tilde{\underline{\mathbf{A}}}_m \in \mathbb{R}^{(n-m)\times(n-m)}$ gilt. Der Beweis erfolgt dann durch Induktion.

1. Schritt:

Wir betrachten die Matrix

$$\underline{\mathbf{A}} = \begin{bmatrix} a_{11} & \cdots & a_{1n} \\ \vdots & \ddots & \vdots \\ a_{n1} & \cdots & a_{nn} \end{bmatrix}.$$

Sei

$$|a_{j1}| \geq |a_{k1}| \qquad \text{für} \quad k = 1, ..., n,$$

das heißt, a_{j1} ist das betragsmäßig größte Element der ersten Spalte (Pivotsuche in der ersten Spalte). Dann wählen wir

$$\underline{\mathbf{P}}_1 = \underline{\mathbf{P}}[1, j] = \begin{bmatrix} 0 & & \cdots & & 1 & & & \\ & 1 & & & & & & \\ \vdots & & \ddots & & \vdots & & & \\ & & & 1 & & & & \\ 1 & & \cdots & & 0 & & & \\ & & & & & 1 & & \\ & & & & & & \ddots & \\ & & & & & & & 1 \end{bmatrix},$$

mit $p_{j1} = p_{1j} = 1$ und $p_{11} = p_{jj} = 0$. Damit folgt

$$\underline{\mathbf{P}}_1 \cdot \underline{\mathbf{A}} = \begin{bmatrix} a_{j1} & \cdots & \cdots & a_{jn} \\ a_{21} & \cdots & \cdots & a_{2n} \\ \vdots & & & \vdots \\ a_{j-1,1} & \cdots & \cdots & a_{j-1,n} \\ a_{11} & \cdots & \cdots & a_{1n} \\ a_{j+1,1} & \cdots & \cdots & a_{j+1,n} \\ \vdots & & & \vdots \\ a_{n1} & \cdots & \cdots & a_{nn} \end{bmatrix}.$$

Der erste Eliminationsschritt wird durch die Multiplikation mit der Eliminationsmatrix

$$\underline{\mathbf{M}}_1 = \begin{bmatrix} 1 & & & & \\ -\frac{a_{21}}{a_{j1}} & 1 & & & \\ -\frac{a_{31}}{a_{j1}} & 0 & \ddots & & \\ \vdots & \vdots & \ddots & \ddots & \\ -\frac{a_{n1}}{a_{j1}} & 0 & \cdots & 0 & 1 \end{bmatrix}$$

repräsentiert, dabei ist $a_{j1} \neq 0$, da sonst die Matrix $\underline{\mathbf{A}}$ singulär wäre (a_{j1} war das betragsmäßig größte Element der ersten Spalte, wäre dieses Null, so wären auch alle anderen Elemente der ersten Spalte gleich Null, das heißt die Determinante von $\underline{\mathbf{A}}$ wäre gleich Null). Alle Komponenten von $\underline{\mathbf{M}}_1$ sind vom Betrag außerdem ≤ 1 (denn in den Brüchen der ersten Spalte in $\underline{\mathbf{M}}_1$ steht im Zähler jeweils eine Zahl, die betragsmäßig höchstens genauso groß wie der jeweilige Nenner ist). Das Ergebnis des ersten Eliminationsschrittes lautet damit

$$\underline{\mathbf{M}}_1 \cdot \underline{\mathbf{P}}_1 \cdot \underline{\mathbf{A}} = \begin{bmatrix} a_{j1} & a_{j2} & \cdots & a_{jn} \\ 0 & \tilde{a}_{22}^{(1)} & \cdots & \tilde{a}_{2n}^{(1)} \\ \vdots & \vdots & \ddots & \vdots \\ 0 & \tilde{a}_{n2}^{(1)} & \cdots & \tilde{a}_{nn}^{(1)} \end{bmatrix}.$$

Weitere Eliminationsschritte verlaufen analog. Dabei ist zu beachten, dass das Pivot-Element nie Null wird, da sonst

$$\underline{\mathbf{A}}^{(m)} = \left[\begin{array}{cccc|cc} u_{11} & \cdots & \cdots & u_{1m} & \multicolumn{2}{c}{\multirow{3}{*}{$\underline{\mathbf{B}}$}} \\ 0 & \ddots & & \vdots & & \\ \vdots & \ddots & \ddots & \vdots & & \\ 0 & \cdots & 0 & u_{mm} & & \\ \hline \multicolumn{4}{c|}{\multirow{3}{*}{$\underline{\mathbf{0}}$}} & 0 & \\ & & & & \vdots & $\underline{\tilde{\mathbf{A}}} \in \mathbb{R}^{(n-m)\times(n-m-1)}$ \\ & & & & 0 & \end{array}\right]$$

gilt und somit die $(m+1)$-te Spalte linear abhängig von den ersten m Spalten ist. Das stellt aber einen Widerspruch zur Regularität von $\underline{\mathbf{A}}$ dar. Nach dem m-ten Gauß-Schritt erhält

man

$$\begin{aligned}\underline{\mathbf{A}}^{(m)} &= \underline{\mathbf{M}}_m \cdot \underline{\mathbf{P}}_m \cdot \underline{\mathbf{M}}_{m-1} \cdot \underline{\mathbf{P}}_{m-1} \cdots \underline{\mathbf{M}}_1 \cdot \underline{\mathbf{P}}_1 \cdot \underline{\mathbf{A}} \\ &= \left[\begin{array}{ccc|c} u_{11} & \cdots & u_{1m} & \\ & \ddots & \vdots & \underline{\mathbf{B}}_m \\ & & u_{mm} & \\ \hline & \underline{\mathbf{0}} & & \tilde{\underline{\mathbf{A}}}_m \end{array}\right]\end{aligned}$$

mit

$$\underline{\mathbf{B}}_m \in \mathbb{R}^{m\times n-m}, \qquad \tilde{\underline{\mathbf{A}}}_m \in \mathbb{R}^{n-m\times n-m}.$$

Nach $n-1$ Gauß-Schritten ergibt sich schließlich

$$\underline{\mathbf{M}}_{n-1} \cdot \underline{\mathbf{P}}_{n-1} \cdot \underline{\mathbf{M}}_{n-2} \cdot \underline{\mathbf{P}}_{n-2} \cdots \underline{\mathbf{M}}_1 \cdot \underline{\mathbf{P}}_1 \cdot \underline{\mathbf{A}} = \underline{\mathbf{U}}.$$

Es ist noch zu zeigen, dass

$$\begin{aligned}&\left(\underline{\mathbf{M}}_{n-1} \cdot \underline{\mathbf{P}}_{n-1} \cdot \underline{\mathbf{M}}_{n-2} \cdot \underline{\mathbf{P}}_{n-2} \cdots \underline{\mathbf{M}}_1 \cdot \underline{\mathbf{P}}_1\right)^{-1} \\ = \;& \underline{\mathbf{P}}_1^{-1} \cdot \underline{\mathbf{M}}_1^{-1} \cdots \underline{\mathbf{P}}_{n-2}^{-1} \cdot \underline{\mathbf{M}}_{n-2}^{-1} \cdot \underline{\mathbf{P}}_{n-1}^{-1} \cdot \underline{\mathbf{M}}_{n-1}^{-1} = \underline{\mathbf{P}}^{-1} \cdot \underline{\mathbf{L}} \qquad (*)\end{aligned}$$

gilt mit einer Permutationsmatrix $\underline{\mathbf{P}}$ und einer unteren Dreiecksmatrix $\underline{\mathbf{L}}$, wobei

$$l_{jj} = 1, \qquad j = 1, ..., n$$

ist. Dann würde folgen, dass

$$\underline{\mathbf{A}} = \underline{\mathbf{P}}^{-1} \cdot \underline{\mathbf{L}} \cdot \underline{\mathbf{U}} \quad \Longleftrightarrow \quad \underline{\mathbf{P}} \cdot \underline{\mathbf{A}} = \underline{\mathbf{L}} \cdot \underline{\mathbf{U}}$$

mit einer Permutationsmatrix $\underline{\mathbf{P}}$ gilt. Dabei ist $\underline{\mathbf{P}}^{-1} = \underline{\mathbf{P}}^T$.

Zum Beweis von $(*)$: Es war

$$\underline{\mathbf{P}}_j = \underline{\mathbf{P}}[j, k_j], \qquad k_j \geq j.$$

Beachte, dass für Vertauschungsmatrizen $\underline{\mathbf{P}}_j = \underline{\mathbf{P}}_j^{-1}$ gilt. Wir zeigen, dass sich für $m = n-1$, $n-2, ..., 1$ die Darstellung

$$\prod_{j=m}^{n-1} \underline{\mathbf{P}}_j \cdot \underline{\mathbf{M}}_j^{-1} = \underline{\mathbf{P}}_m \cdot \underline{\mathbf{M}}_m^{-1} \cdots \underline{\mathbf{P}}_{n-1} \cdot \underline{\mathbf{M}}_{n-1}^{-1} = \left[\begin{array}{c|c} \underline{\mathbf{I}}_{m-1} & \underline{\mathbf{0}} \\ \hline \underline{\mathbf{0}} & \underline{\mathbf{Q}}_m \cdot \underline{\mathbf{L}}_m \end{array}\right]$$

finden lässt mit einer Permutationsmatrix

$$\underline{\mathbf{Q}}_m \in \mathbb{R}^{(n-m+1)\times(n-m+1)}$$

und unterer Dreiecksmatrix $\underline{\mathbf{L}}_m$ mit $l_{jj} = 1$. Für $m = 1$ folgt daraus $(*)$. Den Beweis führen wir durch vollständige Induktion über m.

Induktionsanfang:
Für $m = n-1$ ist

$$\underline{\mathbf{P}}_{n-1} = \underline{\mathbf{P}}[n-1, n]$$

oder

$$\underline{\mathbf{P}}_{n-1} = \underline{\mathbf{P}}[n-1, n-1] = \underline{\mathbf{I}}.$$

Im zweiten Fall ist die Behauptung trivial, da

$$\underline{\mathbf{M}}_{n-1}^{-1} = \begin{bmatrix} 1 & & & \\ & \ddots & & \\ & & 1 & \\ & & y_{n,n-1} & 1 \end{bmatrix}.$$

Im ersten Fall ergibt sich

$$\underline{\mathbf{P}}_{n-1} \cdot \underline{\mathbf{M}}_{n-1}^{-1} = \left[\begin{array}{ccc|cc} 1 & & & & \\ & \ddots & & \multicolumn{2}{c}{\underline{\mathbf{0}}} \\ & & 1 & & \\ \hline \multicolumn{3}{c|}{\underline{\mathbf{0}}} & 0 & 1 \\ & & & 1 & 0 \end{array}\right] \left[\begin{array}{ccc|cc} 1 & & & & \\ & \ddots & & \multicolumn{2}{c}{\underline{\mathbf{0}}} \\ & & 1 & & \\ \hline \multicolumn{3}{c|}{\underline{\mathbf{0}}} & 1 & 0 \\ & & & y_{n,n-1} & 1 \end{array}\right]$$

$$= \left[\begin{array}{ccc|c} 1 & & & \\ & \ddots & & \underline{\mathbf{0}} \\ & & 1 & \\ \hline \multicolumn{3}{c|}{\underline{\mathbf{0}}} & \underline{\mathbf{Q}}_{n-1} \cdot \underline{\mathbf{L}}_{n-1} \end{array}\right],$$

wobei

$$\underline{\mathbf{Q}}_{n-1} = \begin{bmatrix} 0 & 1 \\ 1 & 0 \end{bmatrix}, \qquad \underline{\mathbf{L}}_{n-1} = \begin{bmatrix} 1 & 0 \\ y_{n,n-1} & 1 \end{bmatrix}.$$

Induktionsschritt:
Es sei für alle k mit $m \leq k \leq n-1$ gezeigt, dass

$$\underline{\mathbf{P}}_k \underline{\mathbf{M}}_k^{-1} \cdots \underline{\mathbf{P}}_{n-1} \underline{\mathbf{M}}_{n-1}^{-1} = \left[\begin{array}{c|c} \underline{\mathbf{I}}_{k-1} & \underline{\mathbf{0}} \\ \hline \underline{\mathbf{0}} & \underline{\mathbf{Q}}_k \cdot \underline{\mathbf{L}}_k \end{array}\right]$$

gilt. Dann folgt

$$\begin{aligned} & \underline{\mathbf{P}}_{m-1} \underline{\mathbf{M}}_{m-1}^{-1} \underline{\mathbf{P}}_m \underline{\mathbf{M}}_m^{-1} \cdots \underline{\mathbf{P}}_{n-1} \underline{\mathbf{M}}_{n-1}^{-1} \\ = \; & \underline{\mathbf{P}}_{m-1} \cdot \underline{\mathbf{M}}_{m-1}^{-1} \cdot \left[\begin{array}{c|c} \underline{\mathbf{I}}_{m-1} & \underline{\mathbf{0}} \\ \hline \underline{\mathbf{0}} & \underline{\mathbf{Q}}_m \cdot \underline{\mathbf{L}}_m \end{array}\right] \\ = \; & \underline{\mathbf{P}}_{m-1} \cdot \left[\begin{array}{c|ccc} \underline{\mathbf{I}}_{m-2} & \multicolumn{3}{c}{\underline{\mathbf{0}}} \\ \hline & 1 & & \\ \underline{\mathbf{0}} & y_{m,m-1} & \ddots & \\ & \vdots & & \ddots \\ & y_{n,m-1} & & 1 \end{array}\right] \cdot \left[\begin{array}{c|cc} \underline{\mathbf{I}}_{m-2} & \multicolumn{2}{c}{\underline{\mathbf{0}}} \\ \hline & 1 & \\ \underline{\mathbf{0}} & & \\ & \multicolumn{2}{c}{\underline{\mathbf{Q}}_m \cdot \underline{\mathbf{L}}_m} \end{array}\right] \\ = \; & \underline{\mathbf{P}}_{m-1} \cdot \left[\begin{array}{c|c} \underline{\mathbf{I}}_{m-2} & \underline{\mathbf{0}} \\ \hline \underline{\mathbf{0}} & \underline{\mathbf{C}}_{m-1} \end{array}\right]. \end{aligned}$$

Dabei ist

$$\begin{aligned}\underline{\mathbf{C}}_{m-1} &= \left[\begin{array}{c|c} 1 & \underline{\mathbf{0}}^T \\ \hline \underline{\mathbf{y}}_{m-1} & \underline{\mathbf{I}} \end{array}\right] \cdot \left[\begin{array}{c|c} 1 & \underline{\mathbf{0}}^T \\ \hline \underline{\mathbf{0}} & \underline{\mathbf{Q}}_m \cdot \underline{\mathbf{L}}_m \end{array}\right] = \left[\begin{array}{c|c} 1 & \underline{\mathbf{0}}^T \\ \hline \underline{\mathbf{y}}_{m-1} & \underline{\mathbf{Q}}_m \cdot \underline{\mathbf{L}}_m \end{array}\right] \\ &= \underbrace{\left[\begin{array}{c|c} 1 & \underline{\mathbf{0}}^T \\ \hline \underline{\mathbf{0}} & \underline{\mathbf{Q}}_m \end{array}\right]}_{\tilde{\underline{\mathbf{Q}}}_{m-1}} \cdot \underbrace{\left[\begin{array}{c|c} 1 & \underline{\mathbf{0}}^T \\ \hline \underline{\mathbf{y}}'_{m-1} & \underline{\mathbf{L}}_m \end{array}\right]}_{\underline{\mathbf{L}}_{m-1}}\end{aligned}$$

mit $\underline{\mathbf{y}}'_{m-1} = \underline{\mathbf{Q}}_m^{-1}\,\underline{\mathbf{y}}_{m-1}$. Also ist $\tilde{\underline{\mathbf{Q}}}_{m-1}$ eine Permutationsmatrix. Da außerdem

$$\underline{\mathbf{P}}_{m-1} = \underline{\mathbf{P}}[m-1,k]$$

mit $k \geq m-1$ gilt, folgt

$$\underline{\mathbf{P}}_{m-1} = \left[\begin{array}{c|c} \underline{\mathbf{I}}_{m-2} & \underline{\mathbf{0}} \\ \hline \underline{\mathbf{0}} & \underline{\mathbf{P}}[1,k'] \end{array}\right], \qquad k' = k-m+2,$$

und damit

$$\begin{aligned}&\prod_{j=m-1}^{n-1} \underline{\mathbf{P}}_j \cdot \underline{\mathbf{M}}_j^{-1} \\ &= \left[\begin{array}{c|c} \underline{\mathbf{I}}_{m-2} & \underline{\mathbf{0}} \\ \hline \underline{\mathbf{0}} & \underline{\mathbf{P}}[1,k'] \end{array}\right] \cdot \left[\begin{array}{c|c} \underline{\mathbf{I}}_{m-2} & \underline{\mathbf{0}} \\ \hline \underline{\mathbf{0}} & \tilde{\underline{\mathbf{Q}}}_{m-1} \cdot \underline{\mathbf{L}}_{m-1} \end{array}\right] \\ &= \left[\begin{array}{c|c} \underline{\mathbf{I}}_{m-2} & \underline{\mathbf{0}} \\ \hline \underline{\mathbf{0}} & \underline{\mathbf{P}}[1,k'] \cdot \tilde{\underline{\mathbf{Q}}}_{m-1} \cdot \underline{\mathbf{L}}_{m-1} \end{array}\right].\end{aligned}$$

Mit $\underline{\mathbf{Q}}_{m-1} = \underline{\mathbf{P}}[1,k'] \cdot \tilde{\underline{\mathbf{Q}}}_{m-1}$ folgt die Behauptung. □

Beispiel:

Betrachte die Matrix

$$\underline{\mathbf{A}} = \begin{bmatrix} 1 & 0 & 0 \\ 2 & 1 & 3 \\ 4 & 2 & 1 \end{bmatrix}.$$

Die erste Pivotsuche liefert

$$\underline{\mathbf{P}}[1,3] \cdot \underline{\mathbf{A}} = \begin{bmatrix} 0 & 0 & 1 \\ 0 & 1 & 0 \\ 1 & 0 & 0 \end{bmatrix} \cdot \begin{bmatrix} 1 & 0 & 0 \\ 2 & 1 & 3 \\ 4 & 2 & 1 \end{bmatrix} = \begin{bmatrix} 4 & 2 & 1 \\ 2 & 1 & 3 \\ 1 & 0 & 0 \end{bmatrix}.$$

Damit liefert der erste Eliminationsschritt

$$\underline{\mathbf{M}}_1 \cdot \underline{\mathbf{P}}[1,3] \cdot \underline{\mathbf{A}} = \begin{bmatrix} 1 & 0 & 0 \\ -\frac{1}{2} & 1 & 0 \\ -\frac{1}{4} & 0 & 1 \end{bmatrix} \cdot \begin{bmatrix} 4 & 2 & 1 \\ 2 & 1 & 3 \\ 1 & 0 & 0 \end{bmatrix} = \begin{bmatrix} 4 & 2 & 1 \\ 0 & 0 & \frac{5}{2} \\ 0 & -\frac{1}{2} & -\frac{1}{4} \end{bmatrix}.$$

Durch die zweite Pivotsuche erhält man

$$\underbrace{\underline{\mathbf{P}}[2,3]\cdot\underline{\mathbf{M}}_1\cdot\underline{\mathbf{P}}[1,3]}_{\underline{\tilde{\mathbf{L}}}^{-1}}\cdot\underline{\mathbf{A}} = \begin{bmatrix}1&0&0\\0&0&1\\0&1&0\end{bmatrix}\cdot\begin{bmatrix}4&2&1\\0&0&\frac{5}{2}\\0&-\frac{1}{2}&-\frac{1}{4}\end{bmatrix} = \underbrace{\begin{bmatrix}4&2&1\\0&-\frac{1}{2}&-\frac{1}{4}\\0&0&\frac{5}{2}\end{bmatrix}}_{=\underline{\mathbf{U}}}.$$

Das liefert schließlich

$$\begin{aligned}\tilde{\mathbf{L}} = \underline{\mathbf{P}}[1,3]^{-1}\cdot\underline{\mathbf{M}}_1^{-1}\cdot\underline{\mathbf{P}}[2,3]^{-1} &= \begin{bmatrix}0&0&1\\0&1&0\\1&0&0\end{bmatrix}\cdot\begin{bmatrix}1&0&0\\\frac{1}{2}&1&0\\\frac{1}{4}&0&1\end{bmatrix}\cdot\begin{bmatrix}1&0&0\\0&0&1\\0&1&0\end{bmatrix}\\ &= \underbrace{\begin{bmatrix}0&0&1\\0&1&0\\1&0&0\end{bmatrix}\cdot\begin{bmatrix}1&0&0\\0&0&1\\0&1&0\end{bmatrix}}_{=\begin{bmatrix}0&1&0\\0&0&1\\1&0&0\end{bmatrix}=\underline{\mathbf{P}}^T}\cdot\underbrace{\begin{bmatrix}1&0&0\\\frac{1}{4}&1&0\\\frac{1}{2}&0&1\end{bmatrix}}_{\underline{\mathbf{L}}}.\end{aligned}$$

Mit diesen Matrizen gilt dann

$$\underline{\mathbf{P}}\cdot\underline{\mathbf{A}} = \underline{\mathbf{L}}\cdot\underline{\mathbf{U}}.$$

Bemerkung:

Man kann zeigen, dass für eine untere Dreiecksmatrix $\underline{\mathbf{L}}$ und eine Vertauschungsmatrix $\underline{\mathbf{P}}$ eine untere Dreiecksmatrix $\underline{\mathbf{L}}'$ existiert, so dass

$$\underline{\mathbf{P}}\cdot\underline{\mathbf{L}}' = \underline{\mathbf{L}}\cdot\underline{\mathbf{P}}$$

gilt (Übung!).

MAPLE-Prozedur:

Die eigentliche Gauß-Elimination mit Spaltenpivotsuche wird durch folgende Prozedur realisiert:

```
> restart;
> with(LinearAlgebra);
> gaussSP:=proc(A)
  global p;
  local n, m, piv, pivj, x, j, k;
  n:=RowDimension(A::Matrix);
  p:=Vector(n, i-> i);
  x:=Vector(n);
  for m from 1 to n do
    piv:=abs(A[m, m]);        # Pivotelement (Initialisierung)
    pivj:=m;                  # Zeilenindex des Pivotelementes
```

```
    for j from m + 1 to n do
        if abs(A[j, m]) > piv then piv := abs(A[j, m]); pivj := j end if
    end do;
    k := p[m]; p[m] := p[pivj]; p[pivj] := k;
    for k from 1 to n do
        x[k] := A[m, k]; A[m, k] := A[pivj, k]; A[pivj, k] := x[k]
    end do;
    for j from m + 1 to n do
        A[j, m] := A[j, m]/A[m, m];
        for k from m + 1 to n do
            A[j, k] := A[j, k] − A[j, m] ∗ A[m, k]
        end do
    end do
end do
end proc;
```

Der Vektor p enthält die Permutationen der Zeilen von $\underline{\mathbf{A}}$, das heißt, hier werden die Vertauschungsschritte, die während der Pivotsuche vollzogen werden, gespeichert, so dass zum Schluss eine entsprechende Vertauschung im Lösungsvektor vorgenommen werden kann. Eine Lösung des Gleichungssystems mit (Spalten-) Pivotsuche ergibt sich dann in der folgenden Weise:

```
> loesAxb2:=proc(A::Matrix,b::Vector);
  local vorelim1, ruecksubs, gaussSP, j, bb, n;
  gaussSP:=proc(A::Matrix) ... end proc;
  vorelim1:=proc(L::Matrix,b::Vector) ... end proc;
  ruecksubs:=proc(U::Matrix,b::Vector) ... end proc;
  n:=Dimension(b);
  bb:=Vector(n);
  gaussSP(A);
  for j from 1 to n do
      bb[j] := b[p[j]];
  end do;
  vorelim1(A, bb);
  ruecksubs(A, bb);
  evalm(bb);
  end proc;
```

Dabei ist also die Lösung des LGS im Vektor *bb* abgespeichert.

2.6 Spezielle Lineare Gleichungssysteme

Cholesky-Zerlegung:

Wir betrachten symmetrische, positiv definite Matrizen.

Definition 2.19:

Eine Matrix $\underline{\mathbf{A}} \in \mathbb{R}^{n\times n}$ heißt ***symmetrisch***, falls $\underline{\mathbf{A}} = \underline{\mathbf{A}}^T$ gilt und ***(strikt) positiv***

definit, wenn für alle $\underline{\mathbf{x}} \in \mathbb{R}^n \setminus \{\underline{\mathbf{0}}\}$

$$\underline{\mathbf{x}}^T \cdot \underline{\mathbf{A}} \cdot \underline{\mathbf{x}} > 0.$$

Falls

$$\underline{\mathbf{x}}^T \cdot \underline{\mathbf{A}} \cdot \underline{\mathbf{x}} \geq 0$$

für alle $\underline{\mathbf{x}} \in \mathbb{R}^n \setminus \{\underline{\mathbf{0}}\}$ gilt, so heißt $\underline{\mathbf{A}}$ ***positiv semidefinit***.

Satz 2.20:

Ist $\underline{\mathbf{A}} \in \mathbb{R}^{n\times n}$ symmetrisch und positiv definit, dann gibt es eine untere Dreiecksmatrix $\underline{\mathbf{G}} \in \mathbb{R}^{n\times n}$, so dass

$$\underline{\mathbf{A}} = \underline{\mathbf{G}} \cdot \underline{\mathbf{G}}^T$$

gilt. Diese Zerlegung heißt ***Cholesky-Zerlegung***.

Beweis:

Wir zeigen zunächst, dass $\underline{\mathbf{A}}$ genau dann positiv definit ist, wenn alle Hauptminoren $\underline{\mathbf{A}}_m$, $m = 1, ..., n$, positiv definit sind.

$\Longleftarrow$:

Diese Beweisrichtung ist trivial, da für $m = n$ gilt, dass

$$\underline{\mathbf{A}}_m = \underline{\mathbf{A}}_n = \underline{\mathbf{A}}$$

ist und sich damit die positive Definitheit von $\underline{\mathbf{A}}$ als Spezialfall ergibt.

$\Longrightarrow$:

Angenommen es existiert eine Hauptminore $\underline{\mathbf{A}}_m$ und ein Vektor $\underline{\mathbf{y}} \in \mathbb{R}^m \setminus \{\underline{\mathbf{0}}\}$ so, dass

$$\underline{\mathbf{y}}^T \cdot \underline{\mathbf{A}}_m \cdot \underline{\mathbf{y}} \leq 0$$

gilt. Setze nun

$$\underline{\tilde{\mathbf{y}}}^T := [\underline{\mathbf{y}}^T \quad \underline{\mathbf{0}}^T] \in \mathbb{R}^n$$

mit $\underline{\mathbf{0}} \in \mathbb{R}^{n-m}$, dann gilt

$$[\underline{\mathbf{y}}^T \quad \underline{\mathbf{0}}^T] \cdot \underline{\mathbf{A}} \cdot \begin{bmatrix} \underline{\mathbf{y}} \\ \underline{\mathbf{0}} \end{bmatrix} = \underline{\mathbf{y}}^T \cdot \underline{\mathbf{A}}_m \cdot \underline{\mathbf{y}} \leq 0,$$

denn $\underline{\mathbf{A}}$ ist von der Gestalt

$$\underline{\mathbf{A}} = \left[\begin{array}{c|c} \underline{\mathbf{A}}_m & \cdots \\ \hline \cdots & \cdots \end{array}\right].$$

Das ist aber ein Widerspruch zur positiven Definitheit von $\underline{\mathbf{A}}$, das heißt die Annahme war falsch. Also sind alle Hauptminoren positiv definit. Insbesondere sind alle Hauptminoren $\underline{\mathbf{A}}_m$ invertierbar und nach Satz 2.13 existiert dann eine eindeutige $\underline{\mathbf{LU}}$-Zerlegung für $\underline{\mathbf{A}}$, das heißt, es existieren Matrizen $\underline{\mathbf{L}}, \underline{\mathbf{U}}$ mit $l_{jj} = 1, u_{jj} \neq 0$, $(j = 1, ..., n)$ und

$$\underline{\mathbf{A}} = \underline{\mathbf{L}} \cdot \underline{\mathbf{U}}.$$

Wählt man dann

$$\underline{\mathbf{D}} = \begin{bmatrix} u_{11} & & \\ & \ddots & \\ & & u_{nn} \end{bmatrix},$$

so kann $\underline{\mathbf{U}}$ in der Form

$$\underline{\mathbf{U}} = \underline{\mathbf{D}} \cdot \underline{\tilde{\mathbf{U}}}$$

mit

$$\underline{\tilde{\mathbf{U}}} = \begin{bmatrix} 1 & \frac{u_{12}}{u_{11}} & \cdots & \frac{u_{1n}}{u_{11}} \\ & \ddots & & \vdots \\ & & \ddots & \frac{u_{n-1,n}}{u_{n-1,n-1}} \\ & & & 1 \end{bmatrix}$$

geschrieben werden. Damit ist die eindeutige $\underline{\mathbf{L}} \cdot \underline{\mathbf{U}}$-Zerlegung gleichbedeutend mit

$$\underline{\mathbf{A}} = \underline{\mathbf{L}} \cdot \underline{\mathbf{D}} \cdot \underline{\tilde{\mathbf{U}}}.$$

Aufgrund der Symmetrie von $\underline{\mathbf{A}}$ folgt daraus

$$\underline{\mathbf{A}} = \underline{\mathbf{A}}^T \qquad \Longleftrightarrow \qquad \underline{\mathbf{L}} \cdot \underline{\mathbf{D}} \cdot \underline{\tilde{\mathbf{U}}} = \underline{\tilde{\mathbf{U}}}^T \cdot \underline{\mathbf{D}} \cdot \underline{\mathbf{L}}^T,$$

woraus sich wegen der Eindeutigkeit der LU-Zerlegung

$$\underline{\mathbf{L}} = \underline{\tilde{\mathbf{U}}}^T$$

und damit

$$\underline{\mathbf{A}} = \underline{\mathbf{L}} \cdot \underline{\mathbf{D}} \cdot \underline{\mathbf{L}}^T$$

ergibt. Außerdem gilt, wenn mit $\underline{\mathbf{e}}_j$ der j-te Einheitsvektor bezeichnet wird,

$$\begin{aligned} \left((\underline{\mathbf{L}}^T)^{-1} \cdot \underline{\mathbf{e}}_j\right)^T \cdot \underline{\mathbf{A}} \cdot \left((\underline{\mathbf{L}}^T)^{-1} \cdot \underline{\mathbf{e}}_j\right) &= \underline{\mathbf{e}}_j^T \cdot \underline{\mathbf{L}}^{-1} \cdot \underline{\mathbf{A}} \cdot (\underline{\mathbf{L}}^T)^{-1} \cdot \underline{\mathbf{e}}_j \\ &= \underline{\mathbf{e}}_j^T \cdot \underline{\mathbf{L}}^{-1} \cdot \underline{\mathbf{L}} \cdot \underline{\mathbf{D}} \cdot \underline{\mathbf{L}}^T \cdot (\underline{\mathbf{L}}^T)^{-1} \cdot \underline{\mathbf{e}}_j \\ &= \underline{\mathbf{e}}_j^T \cdot \underline{\mathbf{D}} \cdot \underline{\mathbf{e}}_j = u_{jj} > 0, \end{aligned}$$

da $\underline{\mathbf{A}}$ positiv definit ist. Wählt man dann (positive Wurzeln wählen!)

$$\underline{\mathbf{E}} = \begin{bmatrix} \sqrt{u_{11}} & & \\ & \ddots & \\ & & \sqrt{u_{nn}} \end{bmatrix} = \underline{\mathbf{E}}^T,$$

so liefert dies

$$\underline{\mathbf{E}} \cdot \underline{\mathbf{E}} = \underline{\mathbf{D}}.$$

Dann folgt

$$\underline{\mathbf{A}} = \underline{\mathbf{L}} \cdot \underline{\mathbf{D}} \cdot \underline{\mathbf{L}}^T = \underline{\mathbf{L}} \cdot \underline{\mathbf{E}} \cdot \underline{\mathbf{E}}^T \cdot \underline{\mathbf{L}}^T = (\underline{\mathbf{L}} \cdot \underline{\mathbf{E}}) (\underline{\mathbf{L}} \cdot \underline{\mathbf{E}})^T.$$

Mit

$$\underline{\mathbf{G}} = \underline{\mathbf{L}} \cdot \underline{\mathbf{E}},$$

stellt dies eine Zerlegung in der geforderten Form dar und der Beweis ist beendet. □

Zur Bestimmung der Cholesky-Zerlegung $\underline{\mathbf{A}} = \underline{\mathbf{G}} \cdot \underline{\mathbf{G}}^T$ geht man folgendermaßen vor:

Beispiel:

Wir betrachten die Matrix

$$\underline{\mathbf{A}} = \begin{bmatrix} 1 & 2 & 1 \\ 2 & 5 & 2 \\ 1 & 2 & 10 \end{bmatrix}.$$

Diese Matrix ist symmetrisch und positiv definit, da det $\underline{\mathbf{A}}_1 = a_{11} > 0$, det $\underline{\mathbf{A}}_2 = 1 > 0$ und det $\underline{\mathbf{A}}_3 = 9 > 0$. Wir nutzen den Ansatz

$$\underline{\mathbf{G}} = \begin{array}{ccc|ccc} & & & g_{11} & g_{21} & g_{31} \\ & & & 0 & g_{22} & g_{32} \\ & & & 0 & 0 & g_{33} \\ \hline g_{11} & 0 & 0 & 1 & 2 & 1 \\ g_{21} & g_{22} & 0 & 2 & 5 & 2 \\ g_{31} & g_{32} & g_{33} & 1 & 2 & 10 \end{array} = \underline{\mathbf{G}}^T$$

Für die erste Spalte ergeben sich die Zusammenhänge

$$g_{11}^2 = 1 \quad \Longrightarrow \quad g_{11} = 1, \qquad g_{21} \cdot g_{11} = 2 \quad \Longleftrightarrow \quad g_{21} = \frac{2}{g_{11}} = 2$$

und

$$g_{31} \cdot g_{11} = 1 \quad \Longleftrightarrow \quad g_{31} = \frac{1}{g_{11}} = 1.$$

Die zweite Spalte ergibt sich aus

$$g_{21}^2 + g_{22}^2 = 5 \quad \Longleftrightarrow \quad g_{22}^2 = 5 - g_{21}^2 = 5 - 4 = 1 \quad \Longrightarrow \quad g_{22} = 1$$

und

$$g_{31} \cdot g_{21} + g_{32} \cdot g_{22} = 2 \quad \Longleftrightarrow \quad g_{32} = \frac{2 - g_{31} \cdot g_{21}}{g_{22}} = \frac{2-2}{1} = 0.$$

Für die dritte Spalte ist nur noch g_{33} zu bestimmen, dies ergibt sich durch

$$g_{31}^2 + g_{32}^2 + g_{33}^2 = 10 \quad \Longleftrightarrow \quad g_{33}^2 = 10 - 1 = 9 \quad \Longrightarrow \quad g_{33} = 3.$$

Die gesuchte Matrix $\underline{\mathbf{G}}$ ist dann gegeben durch

$$\underline{\mathbf{G}} = \begin{bmatrix} 1 & 0 & 0 \\ 2 & 1 & 0 \\ 1 & 0 & 3 \end{bmatrix}.$$

Allgemein ergeben sich für die einzelnen Elemente der Matrix $\underline{\mathbf{G}}$ die Rechenvorschriften

$$a_{jk} = \sum_{r=1}^{j} g_{jr} \cdot g_{kr} \qquad\qquad 1 \le j \le k \le n,$$

woraus für $j = k$ speziell $a_{jj} = \sum_{r=1}^{j} g_{jr}^2$, also

$$g_{jj} = \sqrt{a_{jj} - \sum_{r=1}^{j-1} g_{jr}^2}$$

folgt. Die übrigen Elemente ergeben sich aus

$$g_{kj} = \frac{a_{kj} - \sum\limits_{r=1}^{j-1} g_{jr} \cdot g_{kr}}{g_{jj}} \qquad\qquad 1 \le j < k \le n.$$

Dabei ist zu beachten, dass aufgrund der Symmetrie von $\underline{\mathbf{A}}$ $a_{jk} = a_{kj}$ gilt. Die gesamte Zerlegung lässt sich durch folgende Prozedur in MAPLE realisieren (wir setzen voraus, dass die Eingabematrix die nötigen Voraussetzungen erfüllt, das heißt, dass sie positiv definit und symmetrisch ist) :

```
> restart;
> with(LinearAlgebra);
> cholesky:=proc(A::Matrix)
  local n, j, k, r, g, G;
  n:=RowDimension(A);
  G:=Matrix(n, n);
  for j from 1 to n do
      g := A[j, j];
      for r from 1 to j − 1 do g := g − G[j, r]^2 end do;
      G[j, j] :=sqrt(g);
      for k from j + 1 to n do
          g := A[k, j];
          for r from 1 to j − 1 do g := g − G[j, r] ∗ G[k, r] end do;
          G[k, j] := g/G[j, j]
      end do
  end do;
  evalm(G);
  end proc;
```