Reinhard Mennicken
Ekkehard Wagenführer

Numerische Mathematik 1

Mathematik
Grundkurs

rororo vieweg

Dr. rer. nat. Reinhard Mennicken ist Professor
im Fachbereich Mathematik der Universität Regensburg

Dr. rer. nat. Ekkehard Wagenführer ist Akademischer Rat
im Fachbereich Mathematik der Universität Regensburg

(Kurzbiographien der Autoren stehen auf Seite 170)

Redaktion: Verlag Vieweg, Wiesbaden

Veröffentlicht im Rowohlt Taschenbuch Verlag GmbH,
Reinbek bei Hamburg, Mai 1977

Umschlagentwurf Werner Rebhuhn
Satz Vieweg, Braunschweig
Druck Clausen & Bosse, Leck/Schleswig
Printed in Germany
1180-ISBN 3 499 27028 5

Inhaltsverzeichnis

Vorwort

Dieses Buch, aufgegliedert in 3 Bände, enthält eine Einführung in die Numerische Mathematik. Es wendet sich an Studenten der Mathematik, der Informatik, der Wirtschaftswissenschaften, der Natur- und Ingenieurwissenschaften. Voraussetzung zur Lektüre sind Kenntnisse der Analysis und Linearen Algebra etwa im Umfang eines zweisemestrigen Studiums. Entstanden ist das Buch auf der Grundlage einer Vorlesungsreihe, die die Verfasser an der Universität Regensburg seit dem Wintersemester 1971/72 regelmäßig halten.

Der vorliegende 1. Band beschäftigt sich, wie aus dem Inhaltsverzeichnis ersichtlich ist, mit Verfahren zur Lösung linearer Gleichungssysteme sowie damit in Zusammenhang stehender Problemstellungen. Der in Kürze erscheinende 2. Band befaßt sich mit

Eigenwertberechnung,
Iterationsverfahren,
Interpolation.

Der 3. Band schließlich wird die Gebiete

Approximation,
Numerische Quadratur,
Numerische Integration von Differentialgleichungen

behandeln. Die Folge der einzelnen Abschnitte entspricht dem Aufbau der Vorlesungen; die etwas willkürlich erscheinende Aufteilung der Themen auf 3 Bände ist vornehmlich drucktechnisch bedingt.

Besonderes Gewicht legen die Autoren auf die Vermittlung der theoretischen Grundlagen. Insbesondere werden funktionalanalytische Begriffe und Zusammenhänge bereitgestellt und zur Begründung numerischer Verfahren herangezogen. Hauptziel ist es, den Leser zu befähigen, sich weitergehende Verfahren, auch in der Originalliteratur, eigenständig erarbeiten zu können; dies erscheint den Autoren bei der schnell-fortschreitenden Entwicklung der Numerischen Mathematik in hohem Maße bedeutsam.

Trotz dieser Betonung der theoretischen Begründung wird der Gesichtspunkt der praktischen Anwendbarkeit der Verfahren keineswegs vernachlässigt. Bei den einzelnen Algorithmen wird jeweils ausführlich der Rechenaufwand, der Speicherplatzbedarf und die numerische Stabilität diskutiert und eine möglichst weitgehende Fehleranalyse durchgeführt. Darüberhinaus sind die Rechenvorschriften stets so formuliert, daß danach Computer-Programme mühelos zu fertigen sind.

Danken möchten die Autoren allen, die an dieser Monographie mitgewirkt haben. Hervorzuheben sind insbesondere unsere Mitarbeiter Dr. *B. Sagraloff* und Dipl.-Math. *J. Wiesmüller*, die die verschiedenen Fassungen des Manuskripts kritisch durchgesehen haben. Schließlich gilt unser Dank dem Vieweg Verlag für seine Geduld und sein verständnisvolles Eingehen auf unsere zahlreichen Wünsche.

Regensburg, im März 1977

R. Mennicken
E. Wagenführer

1. Rundungsfehler bei Gleitkommarechnung

1.1. Zahlsysteme und Zahldarstellungen im Digitalrechner

Bekanntlich läßt sich jede reelle Zahl a als unendlicher Dezimalbruch darstellen; ist $a \neq 0$, kann man eine geeignete Zehnerpotenz als Faktor abspalten, so daß die Dezimalbruchentwicklung mit einer von Null verschiedenen Ziffer hinter dem Komma beginnt, z.B.

$$\tfrac{4}{3} = 1{,}333\ldots = 10^1 \cdot 0{,}1333\ldots$$

$$= 10^1 \sum_{\nu=1}^{\infty} a_\nu 10^{-\nu} \quad \text{mit} \quad a_1 = 1,\ a_\nu = 3 \quad (\nu \geq 2).$$

Eine Ziffernfolge, die in lauter Neunen endet, ist dabei nicht zugelassen, beispielsweise ist 0,4999... durch 0,5000... zu ersetzen. Daß man als „Basis" des Zahlsystems statt der 10 eine beliebige natürliche Zahl $g \geq 2$ wählen kann, zeigt

(1.1.1) **Satz.** *Es sei* $g \geq 2$ *natürliche Zahl,* $a \neq 0$ *sei reell. Dann existieren eindeutig*

$$(*) \qquad \begin{cases} \sigma \in \{+1, -1\}, & k \in \mathbb{Z}, \\ a_\nu \in \{0, 1, \ldots, g-1\} & (\nu = 1, 2, 3, \ldots) \ \textit{mit} \\ a_1 \neq 0, & \\ \forall n \in \mathbb{N} \quad \exists m \geq n & a_m \neq g-1, \end{cases}$$

so daß gilt

$$(1.1.2) \qquad a = \sigma g^k \sum_{\nu=1}^{\infty} a_\nu g^{-\nu}.$$

Beweis. Zunächst nehmen wir an, wir hätten eine Folge von a_ν mit den Eigenschaften (*). Dann ist die Reihe in (1.1.2) wegen $a_\nu \geq 0$ und

$$a_\nu g^{-\nu} \leq (g-1)\, g^{-\nu}$$

nach dem Majorantenkritierium konvergent; zusätzlich zeigt man

$$(1.1.3) \qquad \forall n \in \mathbb{N} \ \sum_{\nu=n}^{\infty} a_\nu g^{-\nu} < g^{-n+1}:$$

Wählt man nämlich m nach der letzten Zeile von (*), so wird $a_m \leq g-2$; da für alle sonstigen ν sicher $a_\nu \leq g-1$, schätzen wir folgendermaßen ab:

$$\sum_{\nu=n}^{\infty} a_\nu g^{-\nu} \leq \sum_{\substack{\nu=n \\ \nu \neq m}}^{\infty} a_\nu g^{-\nu} + a_m g^{-m} \leq (g-1) \sum_{\nu=n}^{\infty} g^{-\nu} - g^{-m}$$

$$= (g-1)\, g^{-n} \frac{1}{1-g^{-1}} - g^{-m} = g^{-n+1} - g^{-m} < g^{-n+1}.$$

Insbesondere wird

$$(1.1.4) \qquad g^{-1} \leq \sum_{\nu=1}^{\infty} a_\nu g^{-\nu} < 1.$$

Wenn also a die Darstellung (1.1.2) besitzt, ist notwendig

$$(1.1.5) \qquad \sigma = \operatorname{sign} a,$$

und mit dem angegebenen k gilt die Ungleichung

$$(1.1.6) \qquad g^{k-1} \leq |a| < g^k.$$

Definiert man nun

$$(1.1.7) \qquad \tilde{a} = g^{-k} |a|,$$

so ist nach (1.1.2)

$$\tilde{a} - a_1 g^{-1} = \sum_{\nu=2}^{\infty} a_\nu g^{-\nu},$$

und wegen (1.1.3), auf n = 2 angewandt,

$$0 \leq g(\tilde{a} - a_1 g^{-1}) < 1,$$

also

$$(1.1.8) \qquad a_1 \leq g \cdot \tilde{a} < a_1 + 1.$$

Für $n \geq 2$ hat man

$$\tilde{a} - \sum_{\nu=1}^{n-1} a_\nu g^{-\nu} - a_n g^{-n} = \sum_{\nu=n+1}^{\infty} a_\nu g^{-\nu}$$

und wegen (1.1.3) für n + 1:

$$0 \leq g^n \left(\tilde{a} - \sum_{\nu=1}^{n-1} a_\nu g^{-\nu} \right) - a_n < 1,$$

also

$$(1.1.9) \qquad a_n \le g^n \left(\tilde{a} - \sum_{\nu=1}^{n-1} a_\nu g^{-\nu} \right) < a_n + 1 .$$

Es sei nun ein beliebiges reelles $a \neq 0$ vorgegeben, dann wählt man σ gemäß (1.1.5) sowie

$$k = \min \{ \kappa \in \mathbb{Z} : |a| < g^\kappa \} .$$

Wegen $g^\kappa \to \infty$ $(\kappa \to \infty)$ ist die hier notierte Menge nicht leer, wegen $g^\kappa \to 0$ $(\kappa \to -\infty)$ auch nach unten beschränkt und besitzt deshalb ein eindeutig bestimmtes Minimum. k erfüllt die notwendige Bedingung (1.1.6), die wegen der strengen Monotonie von $(g^\kappa)_{\kappa \in \mathbb{Z}}$ das k auch eindeutig festlegt. Wir suchen nun für

$$\tilde{a} := g^{-k} |a|$$

eine Darstellung

$$\tilde{a} = \sum_{\nu=1}^{\infty} a_\nu g^{-\nu}$$

mit den Eigenschaften (*). Dazu konstruieren wir die a_ν $(\nu = 1, 2, \ldots)$ über die folgenden Rekursionen, die wegen (1.1.8) und (1.1.9) erfüllt sein müssen:

$$(1.1.10) \quad \begin{cases} a_1 = \max \{ \kappa \in \mathbb{Z} : \kappa \le g\tilde{a} \} , \\ a_n = \max \left\{ \kappa \in \mathbb{Z} : \kappa \le g^n \left(\tilde{a} - \sum_{\nu=1}^{n-1} a_\nu g^{-\nu} \right) \right\} \quad (n = 2, 3, \ldots), \end{cases}$$

und zeigen für alle $n \ge 1$ die Ungleichungen

$$(A_n) \quad \begin{cases} 0 \le a_n \le g - 1 , \\ 0 \le g^n \left(\tilde{a} - \sum_{\nu=1}^{n} a_\nu g^{-\nu} \right) < 1 . \end{cases}$$

Wir führen einen Induktionsbeweis: Aus (1.1.6) folgt

$$1 \le g \cdot \tilde{a} < g$$

und nach Definition von a_1

$$1 \le a_1 < g, \quad \text{also} \quad 1 \le a_1 \le g - 1 ,$$

außerdem

$$a_1 \le g \cdot \tilde{a} < a_1 + 1, \quad \text{also} \quad 0 \le g \cdot \tilde{a} - a_1 < 1 ,$$

womit (A_1) sowie $a_1 \neq 0$ gezeigt ist.

Für $n \geq 2$ folgt aus der zweiten Zeile von (A_{n-1}) nach Multiplikation mit g:

$$0 \leq g^n \left(\tilde{a} - \sum_{\nu=1}^{n-1} a_\nu g^{-\nu} \right) < g$$

und daher nach Definition von a_n

$$0 \leq a_n \leq g - 1$$

und

$$a_n \leq g^n \left(\tilde{a} - \sum_{\nu=1}^{n-1} a_\nu g^{-\nu} \right) < a_n + 1,$$

was nach Subtraktion von a_n auch den zweiten Teil von (A_n) liefert.

Nachdem (A_n) für jedes natürliche n gezeigt ist, folgern wir aus der zweiten Zeile:

$$0 \leq \tilde{a} - \sum_{\nu=1}^{n} a_\nu g^{-\nu} < g^{-n},$$

und durch Umkehrung der Vorzeichen und Addition von $\tilde{a}$:

$$\tilde{a} - g^{-n} < \sum_{\nu=1}^{n} a_\nu g^{-\nu} \leq \tilde{a}.$$

Da der Ausdruck links gegen $\tilde{a}$ konvergiert, wird

$$\lim_{n \to \infty} \sum_{\nu=1}^{n} a_\nu g^{-\nu} = \sum_{\nu=1}^{\infty} a_\nu g^{-\nu} = \tilde{a},$$

so daß wir eine Darstellung (1.1.2) von a gewonnen haben.

Wir nehmen nun an, es wäre die letzte Eigenschaft von (*) verletzt, also ab einem gewissen n alle a_ν gleich $g - 1$. Dann wäre

$$\tilde{a} = \sum_{\nu=1}^{n} a_\nu g^{-\nu} + \sum_{\nu=n+1}^{\infty} (g-1) g^{-\nu},$$

also

$$g^n \left(\tilde{a} - \sum_{\nu=1}^{n} a_\nu g^{-\nu} \right) = g^n (g-1) \sum_{\nu=n+1}^{\infty} g^{-\nu} = 1,$$

im Widerspruch zu (A_n). – Damit sind alle Eigenschaften (*) der durch (1.1.10) definierten Folge nachgewiesen. Die Eindeutigkeit von σ, k, a_ν $(\nu = 1, 2, 3, \ldots)$ liegt daran, daß alle Konstruktionen aus notwendigen Bedingungen herrühren.

Wir bemerken, daß man für spezielle, z.B. ganze oder rationale Zahlen a die Vorschrift (1.1.10) durch ein praktikableres Verfahren ersetzt, vergleiche dazu Übungsaufgabe 1.1!

Wir kommen nun zu den Zahldarstellungen im Digitalrechner. Bei den Rechenmaschinen unterscheidet man zwischen Digital- und Analogrechnern: der *Analogrechner* übersetzt Zahlen in „kontinuierliche" physikalische Größen, so beispielsweise der Rechenstab in Längen; dabei stellen sich natürlich Probleme der Meßgenauigkeit bei der Ein- und Ausgabe ein. Im Speicher eines *Digitalrechners* – die großen programmierbaren Maschinen sind heute durchweg Digitalrechner – werden Zahlen durch ein physikalisches System mit endlich vielen (diskreten) Zuständen beschrieben; darstellbar sind an Stelle von (1.1.2) nur Zahlen der Form

$$(1.1.11) \qquad a = \sigma\, g^k \sum_{\nu=1}^{t} a_\nu g^{-\nu},$$

wobei $t > 0$ eine feste natürliche Zahl ist.

(1.1.12) **Bezeichnung.** In (1.1.11) heißen

σ das Vorzeichen,

k der Exponent,

$\sum_{\nu=1}^{t} a_\nu g^{-\nu}$ die Mantisse,

t die Mantissenlänge und

a_ν $(\nu = 1, \ldots, t)$ die Ziffern.

Bezüglich des Exponenten gibt es zwei Möglichkeiten:

(i) Festkomma-Arithmetik

Man beschränkt sich auf Zahlen, die sich mit einem festen, vorgegebenen k darstellen lassen, wobei auch $a_1 = 0$ in (1.1.11) zugelassen ist: dann braucht man für k keinen Platz im Speicher. Ist etwa $k = 0$ vorgegeben, werden durch (1.1.11) nur Zahlen mit $0 \leqq |a| < 1$ beschrieben. Mit $k = t$ stehen in (1.1.11) gerade die ganzen Zahlen z mit

$$(1.1.13) \qquad |z| \leqq g^t - 1.$$

Ganze Zahlen schreibt man auch als

$$(1.1.14) \qquad z = \sigma \sum_{\nu=0}^{t} b_\nu g^\nu,$$

man hat dazu in (1.1.11) $a_{t-\nu} = b_\nu$ zu setzen. Die Festkomma-Arithmetik wird

dort angewendet, wo man von der Problemstellung her nicht erwartet, daß der vorgegebene Zahlenbereich überschritten wird, so bei der kaufmännischen Anwendung. Die für mathematische und naturwissenschaftliche Probleme geeigneten Programmiersprachen ALGOL und FORTRAN sehen die Festkomma-Arithmetik nur für ganze Zahlen (INTEGER), also $k = t$ vor. INTEGER – Größen werden vor allem zur Indizierung und zum Abzählen verwendet.

(ii) Gleitkomma-Arithmetik

Für die eigentlichen numerischen Rechnungen arbeitet man vorwiegend mit der Gleitkomma-Arithmetik. Benutzt werden Zahlen der Form

$$(1.1.11) \qquad a = \sigma g^k \sum_{\nu=1}^{t} a_\nu g^{-\nu},$$

für die bei fest vorgegebener Mantissenlänge $t > 0$ und festen ganzzahligen Schranken $K_- < K_+$ gilt:

$$(1.1.15) \qquad \begin{cases} a_\nu \in \{0, 1, \ldots, g-1\} \ (\nu = 1, \ldots, t), \quad \sigma = \pm 1, \\ a_1 \neq 0, \quad \text{falls} \quad a \neq 0, \\ K_- \leqq k \leqq K_+ . \end{cases}$$

Alle so darstellbaren Zahlen ungleich Null liegen in dem Bereich

$$g^{K_- - 1} \leqq |a| < g^{K_+}.$$

Ist $|a| < g^{K_- - 1}$, wird es durch die Null ersetzt, Zahlen mit größerem Betrag als g^{K_+} können nicht verarbeitet werden: in beiden Fällen spricht man vom *Exponentenüberlauf*. Numerische Verfahren sollte man so einrichten, daß keine Bereichsüberschreitung eintritt: das ist wegen des großen Zahlbereichs bei der Gleitkomma-Arithmetik im allgemeinen zu erreichen.

Es stellt sich die Frage nach einer geeigneten Basis g des benutzten Zahlsystems. In den großen Digitalrechnern ist die kleinste Einheit des Kernspeichers das *Bit* (binary digit), ein physikalisches System mit zwei möglichen Zuständen, die man als + und −, wahr und falsch oder als die Dualziffern 0 und 1 interpretiert.

Zur Darstellung von ganzen Zahlen (1.1.13) benutzt man meistens das *Dualsystem*, also $g = 2$. Bei Gleitkommazahlen hat das Dualsystem den Nachteil, daß man betraglich große Schranken K_- und K_+ für den Exponenten wählen muß, um einen befriedigenden Zahlbereich zu erhalten. Man verwendet gern als g eine Zweierpotenz, z. B. $g = 8$ *(Oktalsystem)* oder $g = 16$ *(Hexadezimalsystem)* und schreibt die Ziffern a_ν als Dualzahlen. Ist $g = 2^m$, werden m Dualziffern, also m Bits für jedes a_ν benötigt.

(1.1.16) **Beispiel.** Den Autoren stand eine Rechenanlage vom Typ SIEMENS 4004 zur Verfügung; im Kernspeicher dieser Anlage sind jeweils 8 Bits zu einem *Byte*, der aus technischen Gründen kleinsten adressierbaren Speichereinheit, zusammen-

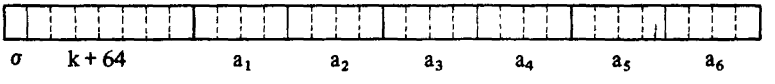
gefaßt. Als Basis des Zahlsystems wird $g = 16 = 2^4$ verwendet. Für die einfach langen Gleitkommazahlen (REAL * 4) stehen 4 Bytes zur Verfügung, davon 1 Byte für Vorzeichen (1 Bit) und Exponent (7 Bits). Man wählt

$$K_- = -64, \quad K_+ = 63$$

und speichert auf den 7 vorgesehenen Bits die Zahl $k + 64$, für die

$$0 \leqq k + 64 \leqq 127 = 2^7 - 1$$

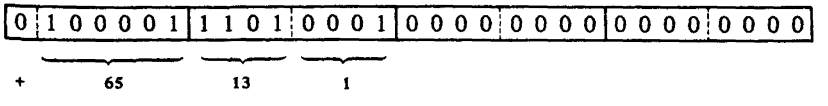
gilt. Die restlichen 3 Bytes werden mit insgesamt $t = 6$ Hexadezimalziffern belegt. Schematisch ist der REAL * 4-Speicherplatz wie folgt unterteilt:



In dem für das Vorzeichen reservierten Bit wird die Dualziffer 0 als +, 1 als − interpretiert. Beispielsweise wird die Zahl

$$a = 13{,}0625 = 16^1 (13 \cdot 16^{-1} + 1 \cdot 16^{-2})$$

wie folgt dargestellt:



Die doppeltgenauen Gleitkommazahlen in dieser Anlage belegen 8 Bytes (REAL * 8), davon wieder 1 Byte für Vorzeichen und Exponent, so daß man $t = 14$ Hexadezimalziffern für die Mantisse erhält.

Der Benutzer einer solchen Rechenanlage braucht im allgemeinen das maschineninterne Zahlsystem nicht zu beherrschen: die Umwandlung vom maschineninternen System ins Dezimalsystem und umgekehrt wird vom *Übersetzer*, also der Maschine selbst durchgeführt. Dabei muß im allgemeinen gerundet werden.

1.2. Runden auf t-stellige Mantisse

Nicht jede reelle Zahl a läßt sich in der Form (1.1.11) schreiben, man muß sie gegebenenfalls durch eine geeignete andere Zahl $\tilde{a}$ ersetzen, *runden*. $\tilde{a}$ ist dann eine Näherung von a und als solche mit einem Fehler behaftet.

(1.2.1) **Definition.** Es seien a und $\tilde{a}$ beliebige reelle Zahlen. Dann heißt

$\tilde{a} - a$ der *(absolute) Fehler* von $\tilde{a}$ als Näherung von a;

ist $a \neq 0$, heißt

$\frac{\tilde{a} - a}{a}$ der *relative Fehler* von $\tilde{a}$.

Wir wollen nun den Prozeß des Aufrundens beschreiben und Fehlerabschätzungen angeben. Hierbei beschränken wir uns auf die Gleitkomma-Arithmetik; entsprechende Überlegungen für die Festkomma-Arithmetik findet man bei Wilkinson [31]. Ferner wollen wir auf einen möglichen Exponentenüberlauf nicht eingehen, setzen also für die auftretenden Exponenten stets $K_- \leqq k \leqq K_+$ gemäß (1.1.15) voraus.

(1.2.2) **Definition.** Es sei $g \geqq 2$ gerade ganze Zahl; $t \in \mathbb{N}, > 0$, $a \in \mathbb{R}, \neq 0$ mit der in Satz (1.1.1) angegebenen Darstellung

$$a = \sigma g^k \sum_{\nu=1}^{\infty} a_\nu g^{-\nu}.$$

Dann ist folgende Zahl der *auf* t *Stellen gerundete Wert von* a:

$$\mathrm{rd}(a) = \begin{cases} \sigma g^k \sum_{\nu=1}^{t} a_\nu g^{-\nu}, & \text{falls } a_{t+1} < \frac{g}{2}, \\ \sigma g^k \left[\sum_{\nu=1}^{t} a_\nu g^{-\nu} + g^{-t} \right], & \text{falls } a_{t+1} \geqq \frac{g}{2}. \end{cases}$$

(1.2.3) **Satz.** *Unter der Voraussetzung von* (1.2.2) *gilt:*

(i) $\mathrm{rd}(a)$ *besitzt eine Darstellung* (1.1.11), *also*

$$\mathrm{rd}(a) = \sigma g^{k'} \sum_{\nu=1}^{t} a'_\nu g^{-\nu};$$

(ii) $|\mathrm{rd}(a) - a| \leqq \frac{1}{2} g^{k-t}$,

(iii) $\left| \frac{\mathrm{rd}(a) - a}{a} \right| \leqq \frac{1}{2} g^{-t+1}$,

(iv) $\left| \frac{\mathrm{rd}(a) - a}{\mathrm{rd}(a)} \right| \leqq \frac{1}{2} g^{-t+1}$.

Daß sich k und alle Ziffern a_ν ändern können, sieht man am folgenden Beispiel mit $g = 10$, $t = 3$:

Für $a = 0{,}9996$ ist $\mathrm{rd}(a) = 1{,}000 = 10^1 \cdot 0{,}100$.

Beweis des Satzes: Aussage (i) ist nur für $a_{t+1} \geqq \frac{g}{2}$ zu zeigen. Man unterscheidet zwei Fälle:

I. Wenn es ein $\nu \in \{1, \ldots, t\}$ mit $a_\nu < g - 1$ gibt, setzt man $k' = k$ und mit $l = \max\{\nu \in \{1, \ldots, t\}: a_\nu < g - 1\}$:

$$a'_\nu = a_\nu \ (\nu = 1, \ldots, l-1), \quad a'_l = a_l + 1, \quad a'_\nu = 0 \quad \text{für } \nu = l+1, \ldots, t.$$

II. Sind alle $a_\nu = g - 1$ $(\nu = 1, \ldots, t)$, so ist $rd(a) = \sigma g^k$, d.h. in der gewünschten Darstellung hat man $k' = k + 1$, $a_1' = 1$, $a_\nu' = 0$ $(\nu = 2, \ldots, t)$.

Zu (ii) berechnen wir im Fall $a_{t+1} < \frac{g}{2}$

$$-\sigma(rd(a) - a) = g^k \sum_{\nu=t+1}^{\infty} a_\nu g^{-\nu} = g^{k-t-1} a_{t+1} + g^k \sum_{\nu=t+2}^{\infty} a_\nu g^{-\nu},$$

wobei wegen $a_{t+1}, \frac{g}{2} \in \mathbb{N}$

$$a_{t+1} g^{k-t-1} \leq (\tfrac{g}{2} - 1)\, g^{k-t-1}$$

sowie nach (1.1.3)

$$g^k \sum_{\nu=t+2}^{\infty} a_\nu g^{-\nu} \leq g^{k-t-1}$$

erfüllt ist. Addition beider Ungleichungen liefert die Behauptung.

Im Fall $a_{t+1} \geq \frac{g}{2}$ erhalten wir

$$\sigma(rd(a) - a) = g^{k-t} - g^k a_{t+1} g^{-t-1} - g^k \sum_{\nu=t+2}^{\infty} a_\nu g^{-\nu}$$

$$= g^{k-t-1}(g - a_{t+1}) - g^k \sum_{\nu=t+2}^{\infty} a_\nu g^{-\nu}.$$

Wegen $\frac{g}{2} \leq a_{t+1} \leq g - 1$ wird

$$g^{k-t-1} \leq g^{k-t-1}(g - a_{t+1}) \leq \tfrac{1}{2}\, g^{k-t}$$

und nach (1.1.3)

$$-g^{k-t-1} < -g^k \sum_{\nu=t+2}^{\infty} a_\nu g^{-\nu} \leq 0\,.$$

Addition beider Zeilen liefert

$$0 < \sigma(rd(a) - a) \leq \tfrac{1}{2}\, g^{k-t}$$

und damit

$$|rd(a) - a| = \sigma(rd(a) - a) \leq \tfrac{1}{2}\, g^{k-t}.$$

In (iii) benutzen wir $a_1 \geq 1$, also $|a| \geq g^{k-1}$, so daß aus (ii)

$$\left| \frac{rd(a) - a}{a} \right| \leq \frac{1}{2} \frac{g^{k-t}}{g^{k-1}} = \frac{1}{2} g^{-t+1}$$

folgt. Zur Aussage (iv) beachtet man die Rundungsvorschrift (1.2.2), nach der in jedem Fall

$$|\mathrm{rd}(a)| \geq a_1 g^{k-1} \geq g^{k-1}$$

erfüllt ist.

Der Vollständigkeit halber definiert man

$$\mathrm{rd}(0) = 0$$

und erhält aus (1.2.3), (iii) und (iv)

(1.2.4) **Folgerung.** Es gelten die Darstellungen

$$\mathrm{rd}(a) = a(1+\epsilon) = \frac{a}{1+\eta}$$

mit gewissen $\epsilon, \eta \in \mathbb{R}$ und

$$|\epsilon|,\ |\eta| \leq \tfrac{1}{2}\, g^{-t+1}.$$

Beweis. Für $a = 0$ setzt man $\epsilon = \eta = 0$; für $a \neq 0$ wird

$$\epsilon = \frac{\mathrm{rd}(a) - a}{a}, \qquad \eta = \frac{a - \mathrm{rd}(a)}{\mathrm{rd}(a)}.$$

(1.2.5) **Definition.** Man bezeichnet

$$\tau := \tfrac{1}{2}\, g^{-t+1}$$

als *(relative) Rechengenauigkeit* der t-stelligen Gleitkomma-Arithmetik.

Wir kommen auf das Beispiel (1.1.16) zurück: will man Dezimalzahlen in die REAL*4-Speicherplätze der angegebenen Maschine bringen, werden sie vom Übersetzer in das Zahlsystem zur Basis $g = 16$ umgewandelt und auf $t = 6$ Stellen gerundet. Somit besitzen nach (1.2.3, iii) die im Kernspeicher befindlichen Größen einen relativen Fehler vom Betrag

$$\leq \tau = \tfrac{1}{2}\, 16^{-5} \approx \tfrac{1}{2}\, 10^{-6}.$$

Es ist also nicht sinnvoll, in der Dezimaldarstellung mehr als 7 Stellen für die Mantisse ein- und ausgeben zu lassen; da bei Ein- und Ausgabe gerundet wird, ändert sich die 7-te Dezimalstelle oft schon, wenn man eine Dezimalzahl in den Kernspeicher bringt und anschließend wieder ausdrucken läßt (vgl. Aufgabe 1.2). Bei doppeltgenauer Rechnung (REAL*8-Speicherplätzen) ist $t = 14$, folglich

$$\tau = \tfrac{1}{2}\, 16^{-13} < \tfrac{1}{2}\, 10^{-15};$$

die doppeltgenaue Arithmetik der Maschine entspricht also einer etwa 16-stelligen Dezimalarithmetik.

Da uns das Dezimalsystem am geläufigsten ist, mißt man die Genauigkeit einer beliebigen Näherung $\tilde{a}$ für die reelle Zahl a an der Zahl der im wesentlichen übereinstimmenden Dezimalstellen.

(1.2.6) **Definition.** Es sei

$$a = \sigma \cdot 10^k b \quad \text{mit} \quad 0{,}1 \leq b < 1;$$
$$\tilde{a} = \sigma \cdot 10^k \tilde{b} \quad \text{mit} \quad \tilde{b} \in \mathbb{R} \text{ beliebig.}$$

Dann sagt man: $\tilde{a}$ als Näherung von a hat s *signifikante Stellen,* oder: a ist durch $\tilde{a}$ bis auf s Dezimalen bestimmt, wenn gilt

$$s = \max \{t \in \mathbb{Z}: |b - \tilde{b}| \leq \tfrac{1}{2}\, 10^{-t+1}\}.$$

Im *Beispiel*

$$a = 10^1 \cdot 0{,}341732\ldots\ ,$$
$$\tilde{a} = 10^1 \cdot 0{,}34215\ ,$$

mit $b, \tilde{b}$ nach (1.2.6) hat man

$$\tfrac{1}{2}\, 10^{-4} < |b - \tilde{b}| < \tfrac{1}{2}\, 10^{-3},$$

also besitzt $\tilde{a}$ drei signifikante Stellen. Diese Tatsache markiert man auch gern durch Unterstreichen der signifikanten Stellen:

$$\tilde{a} = 0{,}\underline{342}15 \cdot 10^1.$$

In allen folgenden Betrachtungen wollen wir zur besseren Anschaulichkeit dezimale Gleitkommarechnung, also $g = 10$, zugrunde legen. Bei einem anderen g ändert sich nichts am Prinzip aller Überlegungen.

1.3. Gleitkomma-Operationen

Es bezeichne $\triangle$ eine der Rechenoperationen $+, -, \cdot, :$. Sind x und y Gleitkommazahlen mit t-stelliger Mantisse, so ist im allgemeinen $x \triangle y$ nicht mit t-stelliger Mantisse darstellbar. Ein *Beispiel* mit $t = 3$: Für

$$x = 0{,}123 \cdot 10^0, \quad y = 0{,}456 \cdot 10^{-3}$$

ist

$$x + y = 0{,}123456 \cdot 10^0$$

Um den Bereich der t-stelligen Zahlen nicht zu verlassen, muß man runden. Man bezeichnet das Resultat der beiden Schritte

1. möglichst genaue Berechnung von $x \triangle y$,
2. Runden des Ergebnisses auf t Stellen,

als

$$gl\,(x \triangle y).$$

Diese *Arithmetik* ist bei den einzelnen Rechnerfabrikaten verschieden organisiert. Wir wollen annehmen, daß für t-stellige Gleitkommazahlen x und y stets gilt:

(1.3.1) $gl(x \triangle y) = rd(x \triangle y)$,

also nach (1.2.4)

(1.3.2) $$gl(x \triangle y) = (x \triangle y)\cdot(1+\epsilon) = \frac{x \triangle y}{1+\eta} \quad \text{mit} \quad |\epsilon|, |\eta| \leq \tau .$$

Wie man die Arithmetik organisieren kann, um (1.3.1) zu erreichen, soll am Beispiel der Addition erläutert werden. Hat man x und y mit t-stelliger Mantisse,

$$x = 10^k \cdot a, \quad y = 10^l \cdot b \quad \text{mit} \quad 0{,}1 \leq |a|, |b| < 1 \quad \text{und} \quad l \leq k,$$

(die Addition wird kommutativ), so unterscheidet man zwei Fälle:

1. Ist $k - l > t$, setzt man direkt

$$gl(x + y) = x.$$

Man überzeugt sich leicht, daß auch $rd(x + y)$ den Wert x hat.

2. Im anderen Fall, $0 \leq k - l \leq t$, benutzt man *doppeltgenaue Zwischenspeicherung:* x und y werden als 2t-stellige Gleitkommazahlen zum gleichen Exponenten dargestellt und dann addiert, wobei für das Ergebnis (2t + 1) Stellen reserviert sind. Ist das Ergebnis nicht Null, wird es *normiert* (der Exponent eventuell geändert, so daß die Mantisse zwischen 0,1 und 1 liegt) und schließlich auf t Stellen gerundet. Dazu notieren wir

(1.3.3) *Zahlenbeispiele mit* t = 3:

(i) $x_1 = 0{,}433 \cdot 10^2$, $x_2 = 0{,}745 \cdot 10^0$;

$$\begin{array}{r} 0{,}433\,000 \cdot 10^2 \\ +\,0{,}007\,450 \cdot 10^2 \\ \hline 0{,}440\,450 \cdot 10^2 \end{array} \quad \rightarrow \underline{gl(x_1 + x_2) = 0{,}440 \cdot 10^2}$$

(ii) $x_1 = 0{,}215 \cdot 10^{-4}$, $x_2 = 0{,}998 \cdot 10^{-4}$;

$$\begin{array}{r} 0{,}215\,000 \cdot 10^{-4} \\ +\,0{,}998\,000 \cdot 10^{-4} \\ \hline 1{,}213\,000 \cdot 10^{-4} \end{array} \quad \rightarrow 0{,}121\,300 \cdot 10^{-3}$$
$$\rightarrow \underline{gl(x_1 + x_2) = 0{,}121 \cdot 10^{-3}}$$

(iii) $x_1 = 0{,}100 \cdot 10^1$, $x_2 = -0{,}998 \cdot 10^0$;

$$\begin{array}{r} 0{,}100\,000 \cdot 10^1 \\ -\,0{,}099\,800 \cdot 10^1 \\ \hline 0{,}000\,200 \cdot 10^1 \end{array} \quad \rightarrow \underline{gl(x_1 + x_2) = 0{,}200 \cdot 10^{-2}}$$

Im Beispiel (iii) spricht man von *Auslöschung* der ersten Dezimalstellen. Obwohl hier sogar $gl(x_1 + x_2) = x_1 + x_2$ herauskommt, ist die Addition entgegengesetzt fast gleicher (d.h. die Subtraktion fast gleicher) Zahlen eine gefährliche Fehlerquelle. Hierzu überlegen wir allgemein:

In vielen Fällen ist eine Summe $a_1 + a_2$ $(a_i \in \mathbb{R})$ zu bestimmen, wobei statt der a_i nur gewisse Gleitkommazahlen

$$x_i = a_i (1 + \rho_i) \qquad (i = 1, 2),$$

d.h. mit relativen Fehlern ρ_i, zur Verfügung stehen. Man berechnet dann $gl(x_1 + x_2)$ als Näherung von $a_1 + a_2$. Über die Mantissenlänge der Gleitkomma-Arithmetik setzen wir $t \geq 2$ voraus, also

$$\tau = \tfrac{1}{2} \cdot 10^{-t+1} \leq 0{,}05 .$$

Wegen (1.3.2) wird

$$gl(x_1 + x_2) = (x_1 + x_2)(1 + \epsilon) = [a_1(1 + \rho_1) + a_2(1 + \rho_2)](1 + \epsilon) = (a_1 + a_2) + F$$

mit dem absoluten Fehler

$$F = a_1(\epsilon + \rho_1(1 + \epsilon)) + a_2(\epsilon + \rho_2(1 + \epsilon)),$$

für den wegen $|\epsilon| \leq \tau \leq 0{,}05$ die Abschätzung

(1.3.4) $$|F| \leq |a_1|(\tau + 1{,}05\,|\rho_1|) + |a_2|(\tau + 1{,}05\,|\rho_2|)$$

erfüllt ist. Ist $a_1 + a_2 \neq 0$, kann man auch den relativen Fehler angeben; man hat dann

$$gl(x_1 + x_2) = (a_1 + a_2)(1 + \rho),$$

mit

$$\rho = \frac{F}{a_1 + a_2},$$

folglich

(1.3.5) $$|\rho| \leq \frac{|a_1|}{|a_1 + a_2|}(\tau + 1{,}05\,|\rho_1|) + \frac{|a_2|}{|a_1 + a_2|}(\tau + 1{,}05\,|\rho_2|) .$$

Wir hatten die x_i als beliebige Näherungen für die a_i angesehen: im günstigsten Fall gilt $x_i = rd(a_i)$, also $|\rho_i| \leq \tau$ $(i = 1, 2)$.

An (1.3.5) lesen wir folgendes ab:

(1.3.6) Im Fall $|a_1 + a_2| < \max(|a_1|, |a_2|)$, insbesondere für $a_2 \approx -a_1$, kann $|\rho|$ sehr viel größer als ρ_1 und ρ_2 werden. Aus diesem Grund heißt die Subtraktion fast gleicher Zahlen *numerisch instabil.*

Beispielsweise liefert

$$a_1 = 0{,}9995 \cdot 10^0 , \qquad a_2 = -0{,}9984 \cdot 10^0$$

bei 3-stelliger Gleitkommarechnung

$$rd(a_1) = x_1\,, \qquad rd(a_2) = x_2$$

mit den Größen von (1.3.3, iii). Als Näherung von $a_1 + a_2$ erhält man

$$gl(rd(a_1) + rd(a_2)) = gl(x_1 + x_2) = 0{,}200 \cdot 10^{-2}\,,$$

während sich als exakter Wert

$$a_1 + a_2 = 0{,}110 \cdot 10^{-2}$$

ergibt. Es liegt also ein relativer Fehler von etwa 82 % vor.

(1.3.7) Bei gleichen Vorzeichen von a_1 und a_2 wird $|a_1 + a_2| = |a_1| + |a_2|$, folglich

$$|\rho| \leq \tau + 1{,}05 \cdot \max(|\rho_1|, |\rho_2|);$$

die vorhandenen relativen Fehler werden also nicht wesentlich vergrößert. Sind speziell die $x_i = rd(a_i)$, so erhält man

$$|\rho| \leq 2{,}05\,\tau\,.$$

(1.3.8) Ist $|a_2|$ wesentlich kleiner als $|a_1|$, so wirkt sich auch ein großer relativer Fehler ρ_2 von x_2 nur wenig im Ergebnis aus. In solchen Fällen spricht man von *Fehlerdämpfung*. Wir wollen diesen Effekt demonstrieren durch die konkreten Annahmen

$$|a_2| \leq 0{,}01\,|a_1|, \quad |\rho_1| \leq \tau, \quad |\rho_2| \leq 10\,\tau\,.$$

Dann wird

$$|a_1 + a_2| \geq 0{,}99\,|a_1|, \quad \frac{|a_1|}{|a_1 + a_2|} \leq 1{,}01, \quad \frac{|a_2|}{|a_1 + a_2|} \leq 0{,}0101$$

und daher

$$|\rho| \leq 1{,}01 \cdot 2{,}05\,\tau + 0{,}0101 \cdot 11{,}5\,\tau \leq 2{,}2\,\tau\,.$$

Der überwiegende Fehleranteil stammt von ρ_1 und der Addition.

Die Subtraktion fast gleicher Zahlen muß immer dann vermieden werden, wenn man einen kleinen relativen Fehler im Ergebnis braucht, z.B. wenn anschließend dividiert werden soll. Nicht ganz zu Recht wird, wie man häufig hört, die „Division durch kleine Zahlen" als Grund für numerische Instabilität angegeben: in solchen Fällen sind durchweg Zähler oder Nenner aus Subtraktion fast gleicher Zahlen entstanden und besitzen daher einen großen relativen Fehler; die Division selbst vergrößert vorhandene relative Fehler nur unwesentlich (vgl. Übungsaufgabe 1.3).

Wie man numerische Instabilität vermeidet, soll an zwei Beispielen gezeigt werden:

(1.3.9) **Beispiel.** In der quadratischen Gleichung

$$ax^2 + bx + c = 0 \quad \text{sei} \quad |4ac| < b^2\,.$$

Dann besitzt die Gleichung die beiden reellen Lösungen

(i) $x_1 = \frac{1}{2a}\,[-b - \mathrm{sign}(b)\sqrt{b^2-4ac}]$,

(ii) $x_2 = \frac{1}{2a}\,[-b + \mathrm{sign}(b)\sqrt{b^2-4ac}]$.

Wenn $|4ac|$ viel kleiner als b^2 ist, hat man

$$\mathrm{sign}(b)\sqrt{b^2-4ac} \approx b\,,$$

und in der Berechnung von x_2 nach (ii) tritt genau der Fall (1.3.6) ein, während bei x_1 der Fall (1.3.7) vorliegt. Wegen

$$x_1 x_2 = \frac{c}{a}$$

berechnet man x_2 also besser über die numerisch stabile Formel

(iii) $$x_2 = \frac{2c}{-b - \mathrm{sign}(b)\sqrt{b^2-4ac}}\,.$$

(1.3.10) **Beispiel.** In einigen Rechenmaschinen sind die Exponentialfunktion, nicht aber die Hyperbelfunktionen fest eingebaut. Läßt man sich $\mathrm{Sinh}(x)$ über

(i) $\mathrm{Sinh}(x) = \frac{1}{2}\,(e^x - e^{-x})$

berechnen, so tritt für $|x| \leqq 0{,}1$ Auslöschung einer oder mehrerer Dezimalstellen ein: e^x und e^{-x} liegen beide nahe bei 1. Daher wird für $|x| \leqq 0{,}1$ der $\mathrm{Sinh}(x)$ durch einen Abschnitt der zugehörigen Potenzreihe ersetzt; bei etwa 7-stelliger Dezimalrechnung empfiehlt sich

(ii) $$S_5(x) := x + \frac{x^3}{6} + \frac{x^5}{120}\,.$$

Nach dem Satz von Taylor gilt mit einem $\vartheta \in]0,\,1[$:

$$|\mathrm{Sinh}(x) - S_5(x)| = \frac{x^6}{720}\,|\mathrm{Sinh}(\vartheta x)| \leqq \frac{x^6}{720}\,|\mathrm{Sinh}(x)|\,,$$

und daher für $0 < |x| \leqq 0{,}1$:

$$\frac{|S_5(x) - \mathrm{Sinh}(x)|}{|\mathrm{Sinh}(x)|} \leqq \frac{x^6}{720} \leqq \frac{1}{7{,}2}\cdot 10^{-8} \quad (< \tfrac{1}{2}\cdot 10^{-6});$$

der durch Abbruch der Potenzreihe entstandene relative Fehler liegt also unterhalb der Rechengenauigkeit. Es bleibt zu bemerken, daß der Ausdruck (ii) bei geeigneter Reihenfolge der Rechenoperationen in den Fall (1.3.8) der Fehlerdämpfung einzuordnen ist, so daß ein geringer Rundungsfehler resultiert. Dazu vgl. Aufgabe 1.4!

1.4. Zusammengesetzte Gleitkomma-Operationen

Rundungsfehler bei zusammengesetzten arithmetischen Ausdrücken behandelt man, indem man (1.3.2) nacheinander auf die einzelnen Operationen und jeweiligen Zwischenergebnisse anwendet. Daß es dabei wesentlich auf die Reihenfolge der Operationen ankommt – die Gleitkommarechnung erfüllt nicht das Assoziativ- und Distributivgesetz –, zeigt das folgende Beispiel mit $t = 3$: Für

$$x_1 = 0{,}481 \cdot 10^{-5}, \quad x_2 = 0{,}572 \cdot 10^{-5}, \quad x_3 = -0{,}963 \cdot 10^{-5}$$

ist

$$gl(x_1 + x_2) = 0{,}105 \cdot 10^{-4}, \quad \text{also} \quad gl(gl(x_1 + x_2) + x_3) = 0{,}870 \cdot 10^{-6};$$
$$gl(x_2 + x_3) = -0{,}391 \cdot 10^{-5}, \quad \text{also} \quad gl(x_1 + gl(x_2 + x_3)) = 0{,}900 \cdot 10^{-6}.$$

Wenn wir gelegentlich trotzdem etwa

$$gl(x_1 + x_2 + x_3)$$

schreiben, meinen wir die Ausführung der Operationen in einer naheliegenden Reihenfolge, hier

$$gl(gl(x_1 + x_3) + x_3).$$

Als Beispiel für den Einfluß der Rundungsfehler bei zusammengesetzten Gleitkomma-Operationen wollen wir an dieser Stelle die Polynomberechnung mit dem Horner-Schema behandeln; weitere Rundungsfehlerbetrachtungen werden wir im folgenden Text an die jeweils besprochenen numerischen Verfahren anschließen.

Gegeben sei ein Polynom n-ten Grades

$$P(x) = P_n(x) = a_0 x^n + a_1 x^{n-1} + \ldots + a_n \quad (a_i \in \mathbb{R},\ a_0 \neq 0) \tag{1.4.1}$$

sowie $\alpha \in \mathbb{R}$. Dann hat das Polynom $P_n(x) - P_n(\alpha)$ in α eine Nullstelle, also existiert ein Polynom $(n-1)$-ten Grades $P_{n-1}(x)$ mit

$$P_n(x) - P_n(\alpha) = (x - \alpha) P_{n-1}(x).$$

(1.4.2) **Satz** *(Horner-Algorithmus). Definiert man* $b_\nu (\nu = 0, \ldots, n)$ *durch*

$$P_{n-1}(x) =: b_0 x^{n-1} + b_1 x^{n-2} + \ldots + b_{n-1}, \qquad P_n(\alpha) =: b_n,$$

so gelten die Rekursionen

$$b_0 = a_0, \quad b_\nu = a_\nu + \alpha b_{\nu-1} \quad (\nu = 1, 2, \ldots, n).$$

Beweis. Aus der Definition von $P_{n-1}(x)$ erhält man

$$\begin{aligned} P_n(x) &= x \cdot P_{n-1}(x) + P_n(\alpha) - \alpha P_{n-1}(x) \\ &= x \cdot \sum_{\nu=0}^{n-1} b_\nu x^{n-1-\nu} + b_n - \sum_{\nu=0}^{n-1} (\alpha b_\nu) x^{n-1-\nu} \\ &= \sum_{\nu=0}^{n} b_\nu x^{n-\nu} - \sum_{\nu=1}^{n} (\alpha b_{\nu-1}) x^{n-\nu}, \end{aligned}$$

also

$$\sum_{\nu=0}^{n} a_\nu x^{n-\nu} = b_0 x^n + \sum_{\nu=1}^{n} (b_\nu - \alpha b_{\nu-1}) x^{n-\nu}.$$

Ein Koeffizientenvergleich liefert die Behauptung des Satzes.

Auf das Polynom $P_{n-1}(x)$ kann man wieder den Horner-Algorithmus anwenden usf.: man erhält sukzessiv Polynome vom Grad $n-\nu$ mit

$$(1.4.3) \qquad P_{n-\nu}(x) - P_{n-\nu}(\alpha) = (x-\alpha)\, P_{n-\nu-1}(x) \quad (\nu = 0, 1, \ldots, n-1),$$

wobei

$$P_0(x) = P_0(\alpha) = a_0$$

ein konstantes Polynom wird. In diesem Fall spricht man vom *verketteten Horner-Algorithmus.*

(1.4.4) **Satz.** *Definiert man gemäß* (1.4.3)

$$c_{n-\nu} = P_{n-\nu}(\alpha) \qquad (\nu = 0, 1, \ldots, n),$$

so wird

$$P_n(x) = \sum_{\nu=0}^{n} c_\nu (x-\alpha)^{n-\nu}$$

und

$$P_n^{(\nu)}(\alpha) = \nu!\, c_{n-\nu} \qquad (\nu = 0, 1, \ldots, n).$$

Der *Beweis* dieses Satzes wird als Übungsaufgabe (Nr. 1.6) empfohlen.

Man benutzt den Horner-Algorithmus

1. zur Berechnung von $P_n(\alpha)$,
2. zur Division von $P_n(x)$ durch $(x-\alpha)$, falls $P_n(\alpha) = 0$ ist,

und mit wiederholter Anwendung

3. zur Berechnung der höheren Ableitungen,
4. zur Entwicklung nach Potenzen von $(x-\alpha)$.

Bei Benutzung von (1.4.1) als Rechenvorschrift für $P_n(\alpha)$ käme man auf $\frac{1}{2}n(n+1)$ Multiplikationen gegenüber n in Satz (1.4.2): durch das Horner-Schema gewinnt man Rechenzeit und Genauigkeit. Oft liegen Polynome nicht in der Gestalt (1.4.1) vor; dann bieten sich im allgemeinen andere Algorithmen zur Berechnung von $P_n(\alpha)$ an. Die Division durch $(x-\alpha)$ im Fall $P_n(\alpha) = 0$ nennt man *Deflation* des Polynoms: man wendet sie häufig an, wenn man mehrere Nullstellen von $P_n(x)$ bestimmen will und eine gefunden hat. Das dividierte Polynom ist von niedrigerem Grad und besitzt – eventuell außer α – alle sonstigen Nullstellen von P_n.

Im folgenden setzen wir die Koeffizienten von P_n sowie α als t-stellige Gleitkommazahlen voraus. t sei so groß, daß für $\tau = \frac{1}{2}\, 10^{-t+1}$ gilt

$$(1.4.5) \qquad 2n\tau \leqq 0{,}09.$$

Führt man den Horner-Algorithmus mit t-stelliger Gleitkommarechnung durch, erhält man statt der b_ν in Satz (1.4.2) gewisse Näherungen $\tilde{b}_\nu$ mit

$$\tilde{b}_0 = a_0,$$
$$\tilde{b}_\nu = gl(a_\nu + gl(\alpha \tilde{b}_{\nu-1})) \qquad (\nu = 1, 2, \ldots, n).$$

Nach (1.3.2) gibt es reelle ϵ_ν, δ_ν mit $|\epsilon_\nu|, |\delta_\nu| \leqq \tau$, so daß

$$gl(\alpha \tilde{b}_{\nu-1}) = \alpha \tilde{b}_{\nu-1}(1 + \epsilon_\nu),$$
$$gl(a_\nu + gl(\alpha \tilde{b}_{\nu-1})) = (a_\nu + gl(\alpha \tilde{b}_{\nu-1}))(1 + \delta_\nu),$$

zusammen also

$$\tilde{b}_\nu = (a_\nu + \alpha \tilde{b}_{\nu-1}(1 + \epsilon_\nu))(1 + \delta_\nu) \qquad (\nu = 1, 2, \ldots, n) \tag{1.4.6}$$

erfüllt ist. Wir zeigen nun für $k = 0, 1, \ldots, n$:

$$\begin{cases} \tilde{b}_k = a_0\alpha^k(1 + E_0^{(k)}) + a_1\alpha^{k-1}(1 + E_1^{(k)}) + \ldots + a_k(1 + E_k^{(k)}), \\ 1 + E_\nu^{(k)} = (1 + \epsilon_{\nu+1}) \cdot \ldots \cdot (1 + \epsilon_k)(1 + \delta_\nu) \cdot \ldots \cdot (1 + \delta_k) \quad (\nu = 0, \ldots, k), \end{cases} \tag{1.4.7}$$

wobei $\delta_0 = 0$ definiert ist.

Den *Beweis* führen wir mit Induktion über k: für $k = 0$ ist $1 + E_0^{(0)} = 1 + \delta_0 = 1$ definiert, andererseits gilt $\tilde{b}_0 = a_0$. Für den Induktionsschluß von k auf $k + 1$ benutzen wir (1.4.6) und setzen anschließend die für k bereits angenommene Darstellung (1.4.7) ein:

$$\begin{aligned} \tilde{b}_{k+1} &= \alpha \tilde{b}_k(1 + \epsilon_{k+1})(1 + \delta_{k+1}) + a_{k+1}(1 + \delta_{k+1}) \\ &= \alpha \sum_{\nu=0}^{k} a_\nu \alpha^{k-\nu}(1 + E_\nu^{(k)})(1 + \epsilon_{k+1})(1 + \delta_{k+1}) + a_{k+1}(1 + \delta_{k+1}) \\ &= \sum_{\nu=0}^{k+1} a_\nu \alpha^{k+1-\nu}(1 + E_\nu^{(k+1)}) \end{aligned}$$

mit

$$1 + E_\nu^{(k+1)} = (1 + E_\nu^{(k)})(1 + \epsilon_{k+1})(1 + \delta_{k+1}) \qquad (\nu = 0, \ldots, k),$$
$$1 + E_{k+1}^{(k+1)} = 1 + \delta_{k+1}.$$

Nach Induktionsannahme gewinnt man (1.4.7) für $k + 1$ statt k. Vereinfacht man vorstehenden Induktionsbeweis auf den Fall $\epsilon_\nu = \delta_\nu = 0$ $(\nu = 0, \ldots, n)$, erhält man für die theoretischen Werte:

$$b_k = \sum_{\nu=0}^{k} a_\nu \alpha^{k-\nu} \qquad (k = 0, \ldots, n),$$

so daß wir aus (1.4.7) folgern

$$(1.4.8) \qquad \tilde{b}_k = b_k + \sum_{\nu=0}^{k} a_\nu \alpha^{k-\nu} E_\nu^{(k)} \qquad (k = 0, 1, \ldots, n).$$

Insbesondere interessiert uns der Fehler bei der Berechnung von $b_n = P(\alpha)$. Den entsprechenden, numerisch gewonnenen Wert bezeichnen wir als $\tilde{b}_n =: gl(P(\alpha))$, außerdem sei $E_\nu := E_\nu^{(n)}$ $(\nu = 0, \ldots, n)$ gesetzt. Mit diesen Umbenennungen schließen wir aus (1.4.7) und (1.4.8) die

(1.4.9) **Folgerung.** Es gilt mit gewissen reellen Größen E_ν $(\nu = 0, \ldots, n)$:

$$gl(P(\alpha)) = P(\alpha) + \sum_{\nu=0}^{n} a_{n-\nu} E_{n-\nu} \alpha^\nu \quad ,$$

wobei

$$(1-\tau)^{2\nu+1} \le 1 + E_{n-\nu} \le (1+\tau)^{2\nu+1} \quad . \qquad (\nu = 1, \ldots, n-1),$$

$$(1-\tau)^{2n} \le 1 + E_0 \le (1+\tau)^{2n}.$$

Zur Abschätzung der E_ν, allgemeiner der $E_\nu^{(k)}$, notieren wir den

(1.4.10) **Hilfssatz.** *Es seien* $\beta, \tau \in \mathbb{R}$ *mit* $\beta \ge 1$, $0 < \tau < 1$ *und* $(\beta - 1)\tau < 1$, *ferner* $E \in \mathbb{R}$ *mit*

$$(1-\tau)^\beta \le 1 + E \le (1+\tau)^\beta.$$

Dann gilt

$$|E| \le (1+\tau)^\beta - 1 \le \frac{\beta}{1-(\beta-1)\tau}\,\tau,$$

und im Fall $(\beta - 1)\tau \le 0{,}09$

$$|E| \le 1{,}1\,\beta\,\tau.$$

Beweis.

(i) Wir zeigen zunächst

$$(*) \qquad (1+\tau)^\beta + (1-\tau)^\beta - 2 \ge 0.$$

Hierzu benutzen wir die Taylorentwicklung der Funktion

$$f(\xi) := (1+\xi)^\beta + (1-\xi)^\beta - 2$$

um den Nullpunkt. Wegen $f(0) = f'(0) = 0$ erhalten wir

$$f(\tau) = \frac{\beta(\beta-1)}{2}\,\tau^2\,[(1+\eta)^{\beta-2} + (1-\eta)^{\beta-2}]$$

mit einem positiven $\eta < \tau$, womit (*) bewiesen ist. Es folgt unmittelbar

$$|E| \le (1+\tau)^\beta - 1.$$

(ii) Zur weiteren Abschätzung sei für $0 \le \xi \le 1$

$$g(\xi) := (1+\xi)^\beta - 1$$

gesetzt. Nach dem Mittelwertsatz gibt es ein ϑ mit $0 < \vartheta < \tau$ und

$$g(\tau) = \tau g'(\vartheta) = \tau\beta(1+\vartheta)^{\beta-1} \le \tau\beta(1+\tau)^{\beta-1} \ ;$$

es folgt

$$(1+\tau)^\beta \left(1 - \frac{\beta\tau}{1+\tau}\right) = (1+\tau)^\beta \ \frac{1-(\beta-1)\tau}{1+\tau} \le 1 \ .$$

Wegen $1-(\beta-1)\tau > 0$ darf man durch den Bruch im zweiten Ausdruck dividieren und erhält

$$(1+\tau)^\beta \le \frac{1+\tau}{1-(\beta-1)\tau} = \frac{1-(\beta-1)\tau+\beta\tau}{1-(\beta-1)\tau} = 1 + \frac{\beta\tau}{1-(\beta-1)\tau} \ .$$

Schließlich hat man im Fall $(\beta-1)\tau \le 0{,}09$:

$$\frac{1}{1-(\beta-1)\tau} \le \frac{1}{0{,}91} < 1{,}1 \ ,$$

was den Beweis vollendet.

Wegen (1.4.5) dürfen wir Hilfssatz (1.4.10) auf alle $E_{n-\nu}$ in (1.4.9) mit $\beta = 2\nu + 1$ anwenden (was für E_0 eine leichte Vergröberung der Fehlerschranke bedeutet). Wir erhalten

$$|E_{n-\nu}| \le 1{,}1 \cdot (2\nu+1)\tau \qquad (\nu = 0, 1, \ldots, n)$$

und daher

$$\left| \sum_{\nu=0}^{n} a_{n-\nu} E_{n-\nu} \alpha^\nu \right| \le 1{,}1 \ \tau \sum_{\nu=0}^{n} (2\nu+1) \, |a_{n-\nu}| \ |\alpha|^\nu \ .$$

Mit der Bezeichnung

$$\hat{P}(x) := \sum_{\nu=0}^{n} |a_{n-\nu}| \, x^\nu$$

wird

$$\hat{P}'(x) = \sum_{\nu=1}^{n} \nu \, |a_{n-\nu}| \, x^{\nu-1}$$

und weiter

$$\sum_{\nu=0}^{n} (2\nu+1) \, |a_{n-\nu}| \ |\alpha|^\nu = 2\,|\alpha|\,\hat{P}'(|\alpha|) + \hat{P}(|\alpha|) \ .$$

Insgesamt liefert (1.4.9) die Rundungsfehlerabschätzung

$$(1.4.11) \qquad |gl(P(\alpha)) - P(\alpha)| \leqq 1{,}1\,\tau\,\{2\,|\alpha|\,\hat{P}'(|\alpha|) + \hat{P}(|\alpha|)\}\,.$$

Wir notieren Konsequenzen für die *Nullstellenbestimmung von Polynomen:*

1. Es sei $\alpha \in \mathbb{R}$ eine einfache Nullstelle von P: dann werden wir die Näherung $\tilde{\alpha}$ von α als numerische Nullstelle von P ansehen, wenn

$$gl(P(\tilde{\alpha})) = \eta$$

mit einem betraglich kleinen η erfüllt ist. Nach (1.4.9) ist $\tilde{\alpha}$ Nullstelle eines gestörten Polynoms $P(x) + q(x)$ mit

$$q(x) = \sum_{\nu=0}^{n} a_{n-\nu} E_{n-\nu} x^{\nu} - \eta\,.$$

Wir betrachten für reelle x, ϵ die Funktion

$$F(x, \epsilon) = P(x) + \epsilon q(x)\,.$$

Da P in α eine einfache Nullstelle hat, gilt

$$F(\alpha, 0) = 0, \quad \frac{\partial F}{\partial x}(x, \epsilon)\,|_{(\alpha,0)} = P'(\alpha) \neq 0\,.$$

Nach dem Satz über implizite Funktionen gibt es Umgebungen U_0 um 0, U_α um α sowie ein beliebig oft differenzierbares φ: $U_0 \to U_\alpha$, so daß für alle $\epsilon \in U_0$ gilt $F(\varphi(\epsilon), \epsilon) = 0$. Wir nehmen an, der Wert $\epsilon = 1$ liege in U_0, was wegen (1.4.11) bei hinreichender Rechengenauigkeit und genügend kleinem $|\eta|$ erfüllt ist, und es sei $\tilde{\alpha}$ gerade die Nullstelle $\varphi(1)$ von $F(x, 1) = P(x) + q(x)$. Die letztere Bedingung läßt sich gegebenenfalls durch Betrachtung sämtlicher Nullstellen von P nachweisen.

Nach dem Satz von Taylor haben wir für $\epsilon \in U_0$

$$\varphi(\epsilon) = \varphi(0) + \epsilon\varphi'(0) + \epsilon^2 r(\epsilon)$$

mit einem Restglied $\epsilon^2 r(\epsilon)$. Wegen

$$\varphi(0) = \alpha, \; \varphi'(0) = -\frac{q(\alpha)}{P'(\alpha)}$$

erhalten wir in erster Näherung – durch Weglassen des Restglieds:

$$\tilde{\alpha} - \alpha = \varphi(1) - \varphi(0) \approx -\frac{q(\alpha)}{P'(\alpha)}\,.$$

Die Genauigkeit von $\tilde{\alpha}$ als Näherung von α hängt nicht nur von der Größe von $|q(\alpha)|$ ab, sondern auch von $P'(\alpha)^{-1}$. Der letztere Wert mißt die *Störanfälligkeit* der Nullstelle α. – Eine mehrfache Nullstelle α von P der Vielfachheit $v > 1$ spaltet bei Störung im allgemeinen in v verschiedene Nullstellen $\alpha_j(\epsilon)$ $(j = 1, \ldots, v)$ von $P(x) + \epsilon q(x)$ auf, wobei sich die $\alpha_j(\epsilon) - \alpha$ wie gewisse gebrochene Potenzen

von ϵ verhalten. Diese Effekte sind bei Wilkinson [32] und Kato [16] eingehender beschrieben; zum Beweis benötigt man eine im Rahmen der Funktionentheorie zu behandelnde Theorie der algebraischen Funktionen.

2. Nach (1.4.11) wächst in Abhängigkeit von α die Rundungsfehlerschranke mit $|\alpha|^n$; die entsprechenden Schranken für $|\tilde{b}_k - b_k|$ wachsen also mit $|\alpha|^k$ $(k=1,\ldots,n)$. Daher sind erwartungsgemäß bei großem α die berechneten Koeffizienten von $P_{n-1}(x)$ ungenau. Will man alle Nullstellen von $P_n(x)$ durch einzelne Berechnung und sukzessive Deflation des Polynoms bestimmen, sollte man mit der betragsmäßig kleinsten Nullstelle beginnen: diese Reihenfolge hat sich auch in der Praxis als optimal erwiesen.

Weitergehende Ausführungen zu diesem Problemkreis finden sich bei Wilkinson [31].

Übungsaufgaben zum 1. Kapitel

Aufgabe 1.1. Es sei $a = \frac{p}{q}$ mit $0 < p < q \in \mathbb{N}$; $g \in \mathbb{N}$, ≥ 2. Man bestimmt k durch

$$-k = \min\{\kappa \in \mathbb{N}: g^{\kappa+1} p \geq q\}$$

und sukzessiv für $n = 1, 2, 3, \ldots$ $a_n \in \mathbb{N}$, $r_n \in \{0, 1, \ldots, q-1\}$ durch die folgenden ganzzahligen Divisionen mit Rest:

$$g^{-k+1} p = a_1 q + r_1,$$
$$g \cdot r_{n-1} = a_n q + r_n \qquad (n = 2, 3, \ldots).$$

(i) Man zeige, daß k der Bedingung (1.1.6) genügt und die a_n die Rekursionen (1.1.10) erfüllen, so daß gemäß Satz (1.1.1) gilt:

$$a = g^k \sum_{\nu=1}^{\infty} a_\nu g^{-\nu}.$$

(ii) Man zeige, daß die Folge der a_n schließlich periodisch wird, d.h.

$$\exists\, n_0 \in \mathbb{N}\ \exists\, l \in \mathbb{N}, > 0 \quad \forall\, n \geq n_0 \quad a_{n+l} = a_n,$$

und gebe eine Abschätzung für die kleinste derartige Zahl l.

Aufgabe 1.2

(i) Man stelle $a = \frac{1}{160}$ als unendlichen Hexadezimalbruch, also nach Satz (1.1.1) mit $g = 16$ dar.

(ii) Welchen Wert hat die aus a durch Runden auf 6 Hexadezimalstellen gewonnene Zahl $\tilde{a}$? – Man stelle die Belegung eines REAL*4-Speicherplatzes durch $\tilde{a}$ wie in Beispiel (1.1.16) dar.

(iii) $\tilde{a}$ wird in eine Dezimal-Gleitkommazahl $\tilde{\tilde{a}}$ mit 7-stelliger Mantisse zurückverwandelt. Man vergleiche $\tilde{\tilde{a}}$ mit a.

Anmerkung: Die Hexadezimalziffern 10, 11, ..., 15 bezeichnet man üblicherweise mit den Großbuchstaben A, B, ..., F.

Ergebnisse: (i) $a = 16^{-1} \cdot (0.1\overline{9})_{\text{Hex.}}$, (ii) $\tilde{a} = 16^{-1} \cdot (0.19999A)_{\text{Hex.}} = a + \frac{2}{5} \cdot 16^{-7}$, (iii) $\tilde{\tilde{a}} = 10^{-2} \cdot 0{,}6250001$.

Aufgabe 1.3. Es seien $a_1, a_2 \in \mathbb{R}$, $a_2 \neq 0$. Man zeige, daß bei t-stelliger dezimaler Gleitkomma-Arithmetik ($t \geq 2$) mit τ gemäß (1.2.5) gilt:

$$gl\left(\frac{rd(a_1)}{rd(a_2)}\right) = \frac{a_1}{a_2}(1+\eta), \qquad |\eta| \leq \frac{3}{1-2\tau}\tau .$$

Aufgabe 1.4

(i) Man organisiere die Berechnung von

$$S_5(x) := x + \frac{x^3}{6} + \frac{x^5}{120}$$

so, daß man nur 4 Multiplikationen bzw. Divisionen benötigt und daß der letzte Schritt in der Addition von x besteht.

(ii) Man zeige, daß man bei t-stelliger dezimaler Gleitkommarechnung ($t \geq 2$) nach der in (i) angegebenen Vorschrift für Gleitkommazahlen x mit $|x| \leq 0{,}1$ die Rundungsfehlerabschätzung

$$|gl(S_5(x)) - S_5(x)| \leq 1{,}05\,|x|\,\tau$$

erhält.

Aufgabe 1.5

(i) Man ersetze die Funktionsvorschrift

$$f(x) = \frac{\cos(x) - 1}{x} \qquad (x \neq 0), \qquad f(0) = 0$$

durch eine für $0 < |x| \leq 0{,}1$ numerisch stabile Formel.

Hinweis: $\cos(2\xi) = ?$

(ii) Man gebe die Taylor-Entwicklung von f um $x = 0$ an und führe eine Restgliedabschätzung für das Taylorpolynom 3. Grades im Bereich $|x| \leq 0{,}1$ durch.

(iii) Zur Demonstration berechne man mit einem Taschenrechner $f(x)$ an mehreren Stellen $0 < |x| \leq 0{,}1$ (α) nach der angegebenen Formel, (β) nach der numerisch stabilen Formel, (γ) unter Benutzung des Taylorpolynoms 3. Grades.

Aufgabe 1.6. Man beweise den Satz (1.4.4) zum verketteten Horner-Algorithmus.

2. Lineare Gleichungssysteme, Eliminations- und Zerlegungsmethoden

Die in diesem Kapitel zu besprechenden Eliminationsverfahren sind die gebräuchlichsten Methoden zur Lösung linearer Gleichungssysteme, sie führen mit endlich vielen Rechenoperationen zum Ziel. Lediglich bei sehr umfangreichen Gleichungssystemen von spezieller Gestalt, wie sie bei der numerischen Behandlung von Randwertaufgaben bei Differentialgleichungen auftreten, wird man die im 6. Kapitel angegebenen Iterationsverfahren vorziehen.

2.1. Vorbemerkungen

Wir betrachten lineare Gleichungssysteme der Form

$$Ax = b, \tag{2.1.1}$$

in denen

$$A = \begin{pmatrix} a_{1,1} & \cdots & a_{1,k} \\ \vdots & & \vdots \\ a_{n,1} & \cdots & a_{n,k} \end{pmatrix}, \quad x = \begin{pmatrix} x_1 \\ \vdots \\ x_k \end{pmatrix}, \quad b = \begin{pmatrix} b_1 \\ \vdots \\ b_n \end{pmatrix}$$

reelle oder komplexe Matrizen bzw. Vektoren sind, für die wir häufig auch die kurze Schreibweise

$$A = (a_{i,j})_{(n,k)}, \quad x = (x_i)_{i=1}^{k}, \quad b = (b_i)_{i=1}^{n}$$

verwenden.

Ein allgemeineres Problem als (2.1.1) ist die Matrizengleichung

$$AX = B \tag{2.1.2}$$

mit der (n,k)-Matrix A, der (n,l)-Matrix B, in der eine (k,l)-Matrix X als Lösung gesucht ist. Bezeichnet

x^j $(j = 1, \ldots, l)$ die Spalten von X,
b^j $(j = 1, \ldots, l)$ die Spalten von B,

so ist (2.1.2) den l Gleichungssystemen

$$Ax^j = b^j \qquad (j = 1, \ldots, l)$$

äquivalent; daher lassen sich Überlegungen zur Lösung von (2.1.1) direkt auf (2.1.2) übertragen und umgekehrt.

Die Invertierung einer (n,n)-Matrix A läßt sich als Spezialfall von (2.1.2) behandeln; die Lösung von

$$AX = I, \qquad I := \text{Einheitsmatrix},$$

falls sie existiert, wird $X = A^{-1}$. Übrigens tritt die Berechnung der Matrix A^{-1} nicht so häufig auf, wie es auf den ersten Blick scheinen mag; meistens werden Vektoren $x = A^{-1}b$ gesucht, die man direkt durch Lösung von (2.1.1) bestimmt.

Zunächst betrachten wir den Fall der *Koeffizientenmatrix in Dreiecksgestalt*, also ein Gleichungssystem der Form

$$(2.1.3) \qquad Rx = b,$$

in dem $R = (r_{i,j})_{(n,n)}$ obere Dreiecksmatrix ist, d.h. $r_{i,j} = 0$ für $i > j$. Ausgeschrieben, lautet das System

$$\begin{aligned} r_{1,1}x_1 + r_{1,2}x_2 + \ldots + r_{1,n}x_n &= b_1 \\ r_{2,2}x_2 + \ldots + r_{2,n}x_n &= b_2 \\ &\vdots \quad\quad \vdots \\ r_{n,n}x_n &= b_n \,. \end{aligned}$$

Unter der Annahme, daß R regulär, also $r_{i,i} \neq 0$ $(i = 1, \ldots, n)$ ist, können wir die Komponenten von x rekursiv berechnen:

$$(2.1.4) \qquad \begin{cases} x_n = \dfrac{b_n}{r_{n,n}} \\ x_\nu = \dfrac{1}{r_{\nu,\nu}} (b_\nu - r_{\nu,\nu+1}x_{\nu+1} - \ldots - r_{\nu,n}x_n) \quad (\nu = n-1, \ldots, 1) \end{cases}$$

Wir bestimmen nun den Rechenaufwand zur Lösung von (2.1.3). Dabei wollen wir grundsätzlich nur die Multiplikationen und Divisionen sowie eventuell kompliziertere Rechenausdrücke wie etwa Wurzelziehen zählen, während wir Additionen und Subtraktionen wegen ihrer erheblich kürzeren Rechenzeiten vernachlässigen. Wegen

$$1 + 2 + \ldots + n = \tfrac{1}{2}\, n(n+1)$$

lesen wir an (2.1.4) unmittelbar ab:

(2.1.5) **Bemerkung.** Der Rechenaufwand zur Lösung eines Gleichungssystems des Typs (2.1.3) beträgt

$$\tfrac{1}{2}\, n(n+1) \approx \tfrac{1}{2}\, n^2 \text{ Operationen.}$$

Dementsprechend benötigt die Lösung eines Systems

$$RX = S$$

mit einer (n, l)-Matrix S ungefähr $\frac{1}{2}\, n^2 l$ Operationen.

(2.1.6) **Bezeichnung.** Es seien $e_1, \ldots, e_n$ die Einheitsspalten des $\mathbb{C}^n$, also die (n,1)-Matrizen

$$e_i = \begin{pmatrix} 0 \\ \vdots \\ 0 \\ 1 \\ 0 \\ \vdots \\ 0 \end{pmatrix} \leftarrow \text{i-te Zeile} \qquad (i = 1, \ldots, n).$$

Zur Invertierung der oberen Dreiecksmatrix R muß man die n Gleichungssysteme

$$(2.1.7) \qquad Rx^i = e_i \qquad (i = 1, \ldots, n)$$

lösen. Offenbar wird in der Lösung von (2.1.7) stets

$$x^i_{i+1} = x^i_{i+2} = \ldots = x^i_n = 0,$$

so daß jedes der Systeme (2.1.7) nur $\frac{1}{2}\,i(i+1)$ Operationen benötigt. Zur Invertierung einer Dreiecksmatrix braucht man also

$$\frac{1}{2}\left(\sum_{i=1}^{n} i^2 + \sum_{i=1}^{n} i\right) = \frac{1}{12}\,n(n+1)(2n+4) \approx \frac{1}{6}\,n^3$$

Operationen.

Analoge Überlegungen wie zu (2.1.3) gelten auch für Gleichungssysteme

$$Lx = b$$

mit einer unteren Dreiecksmatrix L.

Oft treten verallgemeinerte Dreiecksmatrizen auf:

(2.1.8) **Definition.** Eine (n,k)-Matrix $R = (r_{i,j})_{(n,k)}$ heißt obere Dreiecksmatrix (auch, wenn $n \neq k$) stets dann, wenn $r_{i,j} = 0$ für alle $i > j$ gilt.

Eine obere Dreiecksmatrix in diesem Sinn hat also die Form

$$\begin{pmatrix} r_{1,1} & \cdots & r_{1,k} \\ & \ddots & \vdots \\ 0 & & r_{n,n} \cdots r_{n,k} \end{pmatrix} \text{ für } k \geq n, \qquad \begin{pmatrix} r_{1,1} & \cdots & r_{1,k} \\ & \ddots & \vdots \\ 0 & & r_{k,k} \\ & 0 & \end{pmatrix} \text{ für } k < n.$$

Die Lösungsgesamtheit eines Systems

$$(2.1.9) \qquad RX = S,$$

in dem S eine (n,l)-Matrix, R eine obere (n,k)-Dreiecksmatrix bedeute, läßt sich im Fall $n \leq k$, $r_{i,i} \neq 0$ mit einigen Zusatzüberlegungen zu (2.1.3) angeben. Im

Fall $k < n$ ist bekanntlich (2.1.9) nur dann lösbar, wenn die letzten $n - k$ Zeilen von S verschwinden; wenn man S durch numerische Rechnung ermitteln muß, ist letzteres auf Grund von Rundungsfehlern im allgemeinen nicht entscheidbar. Auf überbestimmte Systeme, deren theoretische Lösbarkeit vorher bekannt ist, wendet man im allgemeinen Approximationsmethoden, die im Band 3 behandelt werden, (z.B. Gauß-Ausgleichung) an.

In den folgenden Abschnitten wollen wir allgemeine Gleichungssysteme (2.1.2) auf den Fall (2.1.9) zurückführen. Dazu zerlegen wir im wesentlichen die Koeffizientenmatrix A in

$$A = FR,$$

worin R eine obere Dreiecksmatrix im Sinne von (2.1.8) und F entweder eine invertierbare untere (n,n)-Dreiecksmatrix oder eine unitäre (n,n)-Matrix bedeutet.

Wir notieren die

(2.1.10) **Bemerkung.** Ist $A = FR$ mit den angegebenen Eigenschaften, so erhält man die Lösungsgesamtheit von $AX = B$ durch

(i) Ermittlung der (n,l)-Matrix S mit

$$FS = B,$$

(ii) Lösung des Gleichungssystems

$$RX = S.$$

Beweis. Wegen der Invertierbarkeit von F existiert genau eine (n,l)-Matrix S mit $FS = B$, weiter gilt für jede (k,l)-Matrix X:

$$RX = S \iff F(RX) = FS \iff AX = B.$$

Zur *Durchführung* bemerken wir:

(2.1.11) Die Matrix S kann eventuell gleichzeitig mit R bestimmt werden, indem statt A die erweiterte Matrix des Gleichungssystems

$$(A,B) = F(R,S)$$

zerlegt wird. Zur Berechnung von X ist dann nur *ein* System (2.1.9) zu lösen, und man braucht F nicht explizit zu kennen.

(2.1.12) Die Berechnung von F und Lösung beider Gleichungssysteme (i) und (ii) in (2.1.10) wird explizit durchgeführt, wenn ein Algorithmus zur gleichzeitigen Zerlegung von (A,B) nicht vorgesehen ist, wie bei der Cholesky-Zerlegung im Abschnitt 2.4, oder wenn Gleichungssysteme zu lösen sind, in denen bei gleichem A im Verlauf der Rechnung verschiedene rechte Seiten auftreten. Beispiele hierfür sind die inverse Potenzmethode (Abschnitt 5.1) und das vereinfachte Newton-Verfahren zur Lösung nichtlinearer Gleichungssysteme (Abschnitt 6.6).

Wir bemerken, daß die Zerlegungsmethoden weitere Anwendungen bei der Eigenwertberechnung (5. Kapitel) finden.

2.2. Gaußsches Eliminationsverfahren

Das Gaußsche Eliminationsverfahren wird meistens schon in der linearen Algebra behandelt; auch in der Praxis hat es sich als ein sehr zugkräftiges Verfahren zur Lösung linearer Gleichungssysteme erwiesen. Wir wollen es so formulieren, daß man danach ein Computerprogramm schreiben kann und daß die Rundungsfehler bei der Durchführung möglichst kleingehalten werden. – Die Eliminationsschritte werden durch gewisse untere Dreiecksmatrizen beschrieben, deren Eigenschaften wir zunächst zusammenstellen.

(2.2.1) **Bezeichnung.** Eine untere (n,n)-Dreiecksmatrix $L = (l_{i,j})_{(n,n)}$ heißt *normiert* genau dann, wenn gilt

$$l_{i,i} = 1 \qquad (i = 1, \ldots, n).$$

Bekanntlich bilden die normierten unteren Dreiecksmatrizen bezüglich Matrizenmultiplikation eine Gruppe; wichtige Vertreter dieser Gruppe sind die folgenden elementaren Dreiecksmatrizen:

(2.2.2) **Definition.** Es sei $\nu \in \{1, \ldots, n-1\}$ fest, d eine Spalte im $\mathbb{C}^n$ der Form

$$d = \sum_{i=\nu+1}^{n} \delta_i e_i ,$$

wobei die e_i nach (2.1.6) erklärt sind. Es bezeichne e_ν^t die zu e_ν transponierte $(1,n)$-Matrix, also die ν-te Einheitszeile und hiermit

$$L_\nu(d) := I - de_\nu^t = \begin{pmatrix} 1 & & & & 0 \\ & \ddots & & & \\ & & 1 & & \\ 0 & & -\delta_{\nu+1} & \ddots & \\ & & \vdots & & 0 \\ & & -\delta_n & & 1 \end{pmatrix}$$

ν-te Spalte

(2.2.3) **Hilfssatz.** Die $L_\nu(d)$ genügen den folgenden Rechenregeln:

(i) $\det L_\nu(d) = 1$,

(ii) $L_\nu(d)\, L_\mu(\hat{d}) = I - \hat{d}e_\mu^t - de_\nu^t \qquad (\nu \leqq \mu)$

(iii) $L_{\nu_1}(d_1)\, L_{\nu_2}(d_2) \ldots L_{\nu_k}(d_k) = I - d_1 e_{\nu_1}^t - d_2 e_{\nu_2}^t - \ldots - d_k e_{\nu_k}^t \quad (\nu_1 \leqq \nu_2 \leqq \ldots \leqq \nu_k)$;

(iv) $(L_\nu(d))^{-1} = L_\nu(-d)$;

(v) ist C (n,m)-Matrix mit den Zeilen c_μ^t, so gilt für die Zeilen $\hat{c}_\mu^t$ von $\hat{C} := L_\nu(d)C$:

$$\hat{c}_\mu^t = \begin{cases} c_\mu^t & (\mu = 1, \ldots, \nu), \\ c_\mu^t - \delta_\mu c_\nu^t & (\mu = \nu+1, \ldots, n). \end{cases}$$

Beweis. (i) liest man an der Dreiecksgestalt von $L_\nu(d)$ ab, zu (ii) berechnet man

$$L_\nu(d)\, L_\mu(\hat{d}) = (I - de_\nu^t)(I - \hat{d}e_\mu^t) = I - de_\nu^t - \hat{d}e_\mu^t + (de_\nu^t)(\hat{d}e_\mu^t).$$

Zur Bestimmung des letzten Ausdrucks benutzt man die Assoziativität der Matrizenmultiplikation, wonach gilt

$$(de_\nu^t)(\hat{d}e_\mu^t) = d(e_\nu^t\hat{d})e_\mu^t = (e_\nu^t\hat{d})de_\mu^t,$$

letzteres, da $e_\nu^t\hat{d}$ eine Zahl ist. Schließlich hat man

$$e_\nu^t\hat{d} = e_\nu^t \sum_{i=\mu+1}^{n} \hat{\delta}_i e_i = \sum_{i=\mu+1}^{n} \hat{\delta}_i (e_\nu^t e_i) = 0$$

wegen $\mu + 1 > \nu$. Gleichung (iii) zeigt man mit einem Induktionsbeweis, in den die Überlegungen zu (ii) eingehen. Weiter berechnet man an Hand von (ii)

$$L_\nu(d)\, L_\nu(-d) = I,$$

womit (iv) bewiesen ist. Schließlich wird in (v)

$$\hat{c}_\mu^t = e_\mu^t \hat{C} = e_\mu^t (I - de_\nu^t) C = (e_\mu^t - e_\mu^t de_\nu^t) C = c_\mu^t - (e_\mu^t d)\, c_\nu^t$$

mit

$$e_\mu^t d = \begin{cases} 0 & (\mu = 0, \ldots, \nu), \\ \delta_\mu & (\mu = \nu+1, \ldots, n). \end{cases}$$

Die Regel (v) des Hilfssatzes besagt, daß die Linksmultiplikation mit $L_\nu(d)$ gerade Zeilenoperationen in C bewirkt, wie sie als Gaußsche Eliminationsschritte von der linearen Algebra her bekannt sind. Als ersten Schritt beispielsweise zur Lösung des Gleichungssystems

$$a_{1,1}x_1 + a_{1,2}x_2 + a_{1,3}x_3 = b_1$$
$$a_{2,1}x_1 + a_{2,2}x_2 + a_{2,3}x_3 = b_2$$
$$a_{3,1}x_1 + a_{3,2}x_2 + a_{3,3}x_3 = b_3$$

subtrahiert man

$$\frac{a_{2,1}}{a_{1,1}} \times \text{1. Zeile von der 2. Zeile},$$

$$\frac{a_{3,1}}{a_{1,1}} \times \text{1. Zeile von der 3. Zeile},$$

falls $a_{1,1}$ von Null verschieden ist. Im Fall $a_{1,1} = 0$ muß man vorher Spalten- oder Zeilenvertauschungen vornehmen, die wir im folgenden durch Matrizenmultiplikationen beschreiben.

Es bezeichne S_n die Gruppe der Permutationen, also der bijektiven Abbildungen von $\{1, \ldots, n\}$ auf sich.

(2.2.4) **Definition.** Eine (n,n)-Matrix P heißt *Permutationsmatrix*, wenn es eine Permutation $\pi \in S_n$ gibt, bezüglich der $e^t_{\pi(1)}, \ldots, e^t_{\pi(n)}$ die Zeilen von P sind, d.h.

$$P = \sum_{i=1}^{n} e_i e^t_{\pi(i)} .$$

Als Eigenschaften der Permutationsmatrizen notieren wir

(2.2.5) **Hilfssatz.**

(i) P *ist Permutationsmatrix genau dann, wenn eine Permutation* $\sigma \in S_n$ *existiert, bezüglich der* $e_{\sigma(1)}, \ldots, e_{\sigma(n)}$ *die Spalten von* P *sind; dabei gilt* $\sigma = \pi^{-1}$.

(ii) *Mit* P *ist auch* P^t *Permutationsmatrix.*

(iii) *Permutationsmatrizen sind orthogonal, d.h.* $P^t = P^{-1}$.

(iv) *Ist* A *beliebige* (n,m)*-Matrix mit den Zeilen* $a^t_1, \ldots, a^t_n$, B(k,n)*-Matrix mit den Spalten* $b_1, \ldots, b_n$, *so hat*

PA *die Zeilen* $a^t_{\pi(1)}, \ldots, a^t_{\pi(n)}$,

BP *die Spalten* $b_{\sigma(1)}, \ldots, b_{\sigma(n)}$.

(v) *Die Permutationsmatrizen bilden eine Gruppe.*

(vi) *Die für* $i, j \in \{1, \ldots, n\}$ *durch*

$$P_{i,j} := I - (e_i - e_j)(e_i - e_j)^t$$

definierten Matrizen sind Permutationsmatrizen, die zur Permutation $\pi = \mathrm{id}$ *im Fall* $i = j$, $\pi =$ *Transposition von* i *und* j *im Fall* $i \neq j$ *gehören; ferner gilt*

(vii) $P^{-1}_{i,j} = P_{i,j}$,

(viii) *jede Permutationsmatrix läßt sich als Produkt gewisser* $P_{i,j}$ *darstellen.*

Beweis. In der Darstellung

$$P = \sum_{i=1}^{n} e_i e^t_{\pi(i)}$$

verwenden wir als Summationsindex $j = \pi(i)$, wonach

$$P = \sum_{j=1}^{n} e_{\pi^{-1}(j)} e^t_j$$

offenbar die Spalten $e_{\sigma(1)}, \ldots, e_{\sigma(n)}$ mit $\sigma = \pi^{-1}$ besitzt und daher auch

$$P^t = \sum_{j=1}^{n} e_j e^t_{\sigma(j)}$$

Permutationsmatrix ist. Damit haben wir (ii) und eine Teilaussage von (i) gezeigt; die Umkehrung und damit die vollständige Behauptung (i) erhalten wir durch Betrachtung von P^t statt P. Offenbar bilden die Zeilen einer Permutationsmatrix ein Orthonormalsystem bezüglich des üblichen Skalarprodukts im $\mathbb{C}^n$, woraus (iii) folgt. In (iv) erhalten wir die i-te Zeile von PA als

$$e_i^t(PA) = (e_i^t P)A = e_{\pi(i)}^t A = a_{\pi(i)}^t$$

und die j-te Spalte von B als

$$(BP)e_j = B(Pe_j) = Be_{\sigma(j)} = b_{\sigma(j)}.$$

Die Einheitsmatrix ist Permutationsmatrix, die zu $\pi = \mathrm{id} \in S_n$ gehört; nach (ii) und (iii) ist mit jeder Permutationsmatrix auch ihre Inverse eine solche; es bleibt zu (v) nur die Abgeschlossenheit bezüglich der Multiplikation zu zeigen. Dazu sei P gemäß (2.2.4), ferner

$$Q = \sum_{i=1}^{n} e_i e_{\tau(i)}^t \qquad (\tau \in S_n)$$

vorgegeben. Nach (iv) besitzt PQ die Zeilen

$$e_{\pi(i)}^t Q = \sum_{j=1}^{n} e_{\pi(i)}^t e_j e_{\tau(j)}^t = e_{\tau \circ \pi(i)}^t \qquad (i = 1, \ldots, n),$$

und hiermit ist

$$(*) \quad PQ = \sum_{i=1}^{n} e_i e_{\tau \circ \pi(i)}^t$$

Permutationsmatrix, die zur Zeilenpermutation $\tau \circ \pi$ gehört.

Für die unter (vi) definierten Matrizen rechnet man unmittelbar

$$e_\kappa^t P_{i,j} = e_{\pi(\kappa)}^t \qquad (\kappa = 1, \ldots, n)$$

mit den angegebenen Permutationen π nach. Die zu (vii) äquivalente Aussage

$$P_{i,j}^2 = I$$

erhält man durch Ausmultiplizieren oder aus (*) unter Benutzung von $\pi^2 = \mathrm{id}$. Ist P beliebige Permutationsmatrix zur Zeilenpermutation σ, so läßt sich σ als endliches Produkt gewisser Transpositionen schreiben. Das Produkt der entsprechenden $P_{i,j}$, in der umgekehrten Reihenfolge notiert, liefert P, wie man aus (*) induktiv folgert.

Über das Gaußsche Eliminationsverfahren notieren wir den

(2.2.6) **Satz.** *Es sei* A (n,k)*-Matrix,* B (n,l)*-Matrix. Wir behaupten:*

(i) *Es existieren eine* (n,n)*-Permutationsmatrix* P, *eine* (k,k)*-Permutationsmatrix* Q,

eine normierte untere (n,n)*-Dreiecksmatrix* L, *eine obere* (n,k)*-Dreiecksmatrix* R *und eine* (n,l)*-Matrix* S *mit*

$$PAQ = LR, \quad PB = LS.$$

(ii) *Für* (k,l)*-Matrizen* X *und* Y *mit*

$$X = QY$$

hat man

$$AX = B \iff RY = S.$$

Wir bemerken, daß PAQ aus A nach (2.2.5, iv) durch Permutationen von Zeilen und Spalten entsteht. Eine Zerlegung ohne Zeilen- und Spaltenpermutationen, also

$$A = LR$$

ist nicht allgemein möglich. Wir werden darauf noch eingehen.

Den *Beweis* des Satzes führen wir über ein *Konstruktionsverfahren:* Man definiert

$$C^{(1)} := (A, B) = \left(\begin{array}{ccc|ccc} c^{(1)}_{1,1} & \cdots & c^{(1)}_{1,k} & c^{(1)}_{1,k+1} & \cdots & c^{(1)}_{1,k+l} \\ \vdots & & & & & \vdots \\ c^{(1)}_{n,1} & \cdots & c^{(1)}_{n,k} & c^{(1)}_{n,k+1} & \cdots & c^{(1)}_{n,k+l} \end{array}\right)$$

und betrachtet zwei Fälle:

(I) $c^{(1)}_{i,j} = 0$ für alle $i = 1, \ldots, n;\ j = 1, \ldots, k$, also $A = 0$.
Dann ist mit

$$P = L = I, \quad Q = I, \quad R = A, \quad S = B$$

die gesuchte Zerlegung bereits angegeben.

(II) Im Fall, daß ein $c^{(1)}_{i,j} \neq 0$ $(1 \leqq i \leqq n,\ 1 \leqq j \leqq k)$ existiert, wählt man ein derartiges $c^{(1)}_{i_1,j_1} \neq 0$. Man bezeichnet dann die Zeile Nr. i_1 als Pivotzeile, die Spalte Nr. j_1 als Pivotspalte und $c^{(1)}_{i_1,j_1}$ als Pivotelement. Die Wahl des Pivotelements muß in einem Programm nach einer festen Vorschrift erfolgen; es sind folgende Auswahlvorschriften gebräuchlich:

(2.2.7) *Diagonale Pivotwahl:* Man wählt

$$c^{(1)}_{1,1} \qquad \text{(beim } \nu\text{-ten Eliminationsschritt } c^{(\nu)}_{\nu,\nu})$$

als Pivotelement. Dies ist nicht immer möglich, da $c^{(1)}_{1,1}$ Null sein kann.

(2.2.8) *Halbmaximale Pivotwahl:* Man wählt die 1. Spalte (beim ν-ten Schritt die ν-te Spalte) als Pivotspalte und bestimmt ein i_1 mit

$$|c^{(1)}_{i_1,1}| = \max_{i=1}^{n} |c^{(1)}_{i,1}|, \qquad \text{(später } |c^{(\nu)}_{i_\nu,\nu}| = \max_{i=\nu}^{n} |c^{(\nu)}_{i,\nu}|\text{)}.$$

Diese Pivotwahl ist möglich, wenn die 1. Spalte nicht aus lauter Nullen besteht.

(2.2.9) *Maximale Pivotwahl:* Man wählt ein Indexpaar $(i_1, j_1) \in \{1, \ldots, n\} \times \{1, \ldots, k\}$ mit

$$|c^{(1)}_{i_1,j_1}| = \max \{|c^{(1)}_{i,j}| : i = 1, \ldots, n;\ j = 1, \ldots, k\},$$

(beim ν-ten Schritt entsprechend

$$|c^{(\nu)}_{i_\nu,j_\nu}| = \max \{|c^{(\nu)}_{i,j}| : i = \nu, \ldots, n;\ j = \nu, \ldots, k\}).$$

Diese *maximale* oder *totale* Pivotwahl ist immer möglich.

Anschließend definiert man nach (2.2.5, vi) die (n,n)- bzw. (k,k)-Permutationsmatrizen

$$P^{(1)} = P_{1,i_1}, \qquad Q^{(1)} = P_{1,j_1}$$

sowie

$$\hat{Q}^{(1)} = \left(\begin{array}{c|c} Q^{(1)} & 0 \\ \hline 0 & I_l \end{array}\right)_{(k+l,\,k+l)}$$

und setzt hiermit

$$\hat{C}^{(1)} = P^{(1)} C^{(1)} \hat{Q}^{(1)} = \left(\begin{array}{ccc|ccc} \hat{c}^{(1)}_{1,1} & \cdots & \hat{c}^{(1)}_{1,k} & \hat{c}^{(1)}_{1,k+1} & \cdots & \hat{c}^{(1)}_{1,k+l} \\ \vdots & & \vdots & \vdots & & \vdots \\ \hat{c}^{(1)}_{n,1} & \cdots & \hat{c}^{(1)}_{n,k} & \hat{c}^{(1)}_{n,k+1} & \cdots & \hat{c}^{(1)}_{n,k+l} \end{array}\right),$$

so daß nunmehr $\hat{c}^{(1)}_{1,1} = c^{(1)}_{i_1,j_1} \neq 0$ gilt. Durch Ausmultiplikation der Teilmatrizen erhält man

$$\hat{C}^{(1)} = (P^{(1)} A Q^{(1)},\ P^{(1)} B).$$

Weiter wird

$$d_1 := \frac{1}{\hat{c}^{(1)}_{1,1}} \cdot \sum_{i=2}^{n} \hat{c}^{(1)}_{i,1}\, e_i$$

und hiermit

$$C^{(2)} := L_1(d_1)\, \hat{C}^{(1)}$$

definiert, so daß nach (2.2.3, v) $C^{(2)}$ die Form

$$C^{(2)} = \begin{pmatrix} c^{(2)}_{1,1} & c^{(2)}_{1,2} & \cdots & c^{(2)}_{1,k+l} \\ 0 & c^{(2)}_{2,2} & \cdots & c^{(2)}_{2,k+l} \\ \vdots & \vdots & & \vdots \\ 0 & c^{(2)}_{n,2} & \cdots & c^{(2)}_{n,k+l} \end{pmatrix}$$

erhält, wobei $c^{(2)}_{1,j} = \hat{c}^{(1)}_{1,j}$ für $j = 1, \ldots, k + l$ gilt.

Beim 2. Konstruktionsschritt kann als I. Fall $c^{(2)}_{i,j} = 0$ für alle $2 \leqq i \leqq n$, $2 \leqq j \leqq k$ eintreten; dann ist das Verfahren abzubrechen. Sonst wählt man ein Pivotelement $c^{(2)}_{i_2,j_2} \neq 0$ $(2 \leqq i_2 \leqq n,\ 2 \leqq j_2 \leqq k)$ nach der schon beim ersten Schritt angewendeten Vorschrift (2.2.7), (2.2.8) oder (2.2.9) und setzt

$$P^{(2)} = P_{2,i_2}, \quad Q^{(2)} = P_{2,j_2}, \quad \hat{Q}^{(2)} = \left(\begin{array}{c|c} Q^{(2)} & 0 \\ \hline 0 & I_l \end{array}\right)_{(k+l,\,k+l)},$$

$$\hat{C}^{(2)} = P^{(2)} C^{(2)} \hat{Q}^{(2)} = \begin{pmatrix} \hat{c}^{(2)}_{1,1} & \hat{c}^{(2)}_{1,2} & \cdots & \hat{c}^{(2)}_{1,k+l} \\ 0 & \hat{c}^{(2)}_{2,2} & \cdots & \hat{c}^{(2)}_{2,k+l} \\ \vdots & \vdots & & \vdots \\ 0 & \hat{c}^{(2)}_{n,2} & \cdots & \hat{c}^{(2)}_{n,k+l} \end{pmatrix}.$$

Anschließend konstruiert man $C^{(3)} = L_2(d_2)\hat{C}^{(2)}$ mit $d_2 = \dfrac{1}{\hat{c}^{(2)}_{2,2}} \sum_{i=3}^{n} \hat{c}^{(2)}_{i,2} e_i$, so daß $C^{(3)}$ die Gestalt

$$C^{(3)} = \begin{pmatrix} c^{(3)}_{1,1} & c^{(3)}_{1,2} & c^{(3)}_{1,3} & \cdots & c^{(3)}_{1,k+l} \\ 0 & c^{(3)}_{2,2} & c^{(3)}_{2,3} & \cdots & c^{(3)}_{2,k+l} \\ 0 & 0 & c^{(3)}_{3,3} & \cdots & c^{(3)}_{3,k+l} \\ \vdots & \vdots & \vdots & & \vdots \\ 0 & 0 & c^{(3)}_{n,3} & & c^{(3)}_{n,k+l} \end{pmatrix}$$

besitzt. Nach der gleichen Methode fährt man fort, bis das Verfahren nach $(m-1)$ Konstruktionsschritten – mit $m \leqq \min(n, k+1)$ – bei einer Matrix $C^{(m)}$ der Form

$$C^{(m)} = \left(\begin{array}{cccc|c} c_{1,1}^{(m)} & c_{1,2}^{(m)} & \cdots & c_{1,k}^{(m)} & \cdots \; c_{1,k+l}^{(m)} \\ 0 & c_{2,2}^{(m)} & \cdots & c_{2,k}^{(m)} & \cdots \; c_{2,k+l}^{(m)} \\ & \ddots & & \vdots & \vdots \\ 0 & & & & \end{array}\right)$$

endet. Der erste Teil von $C^{(m)}$, den wir mit R bezeichnen wollen, besitzt die gewünschte obere Dreiecksgestalt; die Matrix aus den letzten l Spalten von $C^{(m)}$ bezeichnen wir mit S. Nach Konstruktion gilt

(2.2.10)

$$C^{(m)} = L_{m-1}(d_{m-1})P^{(m-1)}L_{m-2}(d_{m-2})P^{(m-2)}\ldots L_1(d_1)P^{(1)}C^{(1)}\hat{Q}^{(1)}\ldots\hat{Q}^{(m-1)},$$

wobei offenbar $\hat{Q} := \hat{Q}^{(1)}\ldots\hat{Q}^{(m-1)}$ mit einer (k,k)-Permutationsmatrix Q die Darstellung

$$\hat{Q} = \left(\begin{array}{c|c} Q & 0 \\ \hline 0 & I_l \end{array}\right)$$

besitzt. Wir zeigen nun

(2.2.11) $\quad P_{i,j}L_\nu(d_\nu) = L_\nu(P_{i,j}d_\nu)P_{i,j} \qquad (i,j>\nu).$

Dazu berechnen wir

$$P_{i,j}L_\nu(d_\nu) = P_{i,j} - P_{i,j}d_\nu e_\nu^t = P_{i,j} - (P_{i,j}d_\nu)(e_\nu^t P_{i,j}) = L_\nu(P_{i,j}d_\nu)P_{i,j};$$

dabei beachten wir, daß für $i,j>\nu$ $e_\nu^t P_{i,j} = e_\nu^t$ gilt. Da außerdem $P_{i,j}e_\mu$ für alle $\mu \geqq \nu+1$ eine der Einheitsspalten $e_{\nu+1},\ldots,e_n$ liefert, besitzt

$$P_{i,j}d_\nu = \sum_{\mu=\nu+1}^{n} \delta_\mu(P_{i,j}e_\mu)$$

eine Darstellung wie d_ν, so daß $L_\nu(P_{i,j}d_\nu)$ elementare untere Dreiecksmatrix wird.

Aus (2.2.11) folgt für $\nu = 1, 2, \ldots, m-2$

$$P^{(m-1)}\ldots P^{(\nu+1)}L_\nu(d_\nu) = L_\nu(P^{(m-1)}\ldots P^{(\nu+1)}d_\nu)P^{(m-1)}\ldots P^{(\nu+1)}.$$

Dementsprechend gewinnen wir mit

$$\hat{d}_\nu := P^{(m-1)}\ldots P^{(\nu+1)}d_\nu \qquad (\nu = 1,\ldots,m-2),$$
$$\hat{d}_{m-1} := d_{m-1}, \qquad P := P^{(m-1)}\ldots P^{(1)}$$

und den elementaren unteren Dreiecksmatrizen $L_\nu(\hat{d}_\nu)$ aus (2.2.10) die Beziehung

$$(2.2.12) \qquad C^{(m)} = L_{m-1}(\hat{d}_{m-1}) \cdot \ldots \cdot L_1(\hat{d}_1)\, P C^{(1)} \hat{Q}\,.$$

Man setzt

$$L = [L_{m-1}(\hat{d}_{m-1}) \ldots L_1(\hat{d}_1)]^{-1},$$

also nach (2.2.3), (iv) und (iii)

$$(2.2.13) \qquad L = I + \hat{d}_1 e_1^t + \ldots + \hat{d}_{m-1} e_{m-1}^t$$

und erhält hiermit aus (2.2.16)

$$L(R, S) = P(A, B) \begin{pmatrix} Q & 0 \\ 0 & I_l \end{pmatrix} = (PAQ, PB),$$

so wie im Satz behauptet. Die Aussage (ii) des Satzes folgt unmittelbar durch Einsetzen.

Beim Programmieren der Gauß-Elimination wäre es einigermaßen umständlich, die Zeilen- und Spaltenvertauschungen durch Umspeichern zu realisieren; statt dessen bringt man die entsprechenden Permutationen in INTEGER-Zahlenfelder und benutzt sie zur Indizierung der Matrixkoeffizienten. Um die Gauß-Elimination in dieser Form zu beschreiben, bemerken wir, daß (2.2.12) statt für m, entsprechend abgeändert, auch für alle ν mit $2 \leqq \nu \leqq m$ gilt; mit P_ν, $\hat{Q}_\nu$, definiert durch

$$P_1 = I_n, \qquad P_\nu = P^{(\nu-1)} \cdot \ldots \cdot P^{(1)} \qquad (\nu \geqq 2),$$
$$\hat{Q}_1 = I_{k+l}, \qquad \hat{Q}_\nu = \hat{Q}^{(\nu-1)} \cdot \ldots \cdot \hat{Q}^{(1)} \qquad (\nu \geqq 2)$$

und gewissen unteren Dreiecksmatrizen $L^{(\nu)}$ sind die $C^{(\nu)}$ durch

$$(2.2.14) \qquad C^{(\nu)} = L^{(\nu)}(P_\nu C^{(1)} \hat{Q}_\nu) \qquad (\nu = 1, \ldots, m)$$

darstellbar. Bezeichnet $\pi_\nu \in S_n$, $\sigma_\nu \in S_{k+l}$ die durch

$$e_\kappa^t P_\nu = e_{\pi_\nu(\kappa)}^t \qquad (\kappa = 1, \ldots, n), \qquad \hat{Q}_\nu e_\kappa = e_{\sigma_\nu(\kappa)} \qquad (\kappa = 1, \ldots, k+l)$$

festgelegten Permutationen, so gilt nach (2.2.5, iv) die Darstellung

$$P_\nu C^{(1)} \hat{Q}_\nu = (c^{(1)}_{\pi_\nu(i), \sigma_\nu(j)})_{(n, k+l)}.$$

Diese Tatsache legt es nahe, die Matrizen $C^{(\nu)}$ durch Koeffizienten $\gamma^{(\nu)}_{i,j}$ in der Form

$$C^{(\nu)} := (\gamma^{(\nu)}_{\pi_\nu(i), \sigma_\nu(j)})_{(n, k+l)}$$

zu beschreiben. An Speicherplatz benötigt man im wesentlichen ein INTEGER-Zahlenfeld der Länge n für die $\pi_\nu(i)$ $(i = 1, \ldots, n)$, eins der Länge $(k + l)$ für die $\sigma_\nu(j)$ und ein REAL-Feld der Dimension $n \times (k + l)$ für die $\gamma^{(\nu)}_{i,j}$, – der Index ν wird zur Indizierung der Felder im Programm nicht gebraucht.

Durch Übertragung auf die angegebene Schreibweise gewinnen wir aus dem Beweis von Satz (2.2.6), getrennt nach den Möglichkeiten der Pivotwahl, die folgenden Algorithmen der Gauß-Elimination:

(2.2.15) *Algorithmus bei maximaler Pivotwahl:*

(i) Man setzt $\pi_1(i) = i,\ \sigma_1(j) = j,\ \gamma_{i,j}^{(1)} = c_{i,j}^{(1)}$ $(i = 1, \ldots, n;\ j = 1, \ldots, k + l)$.

Für $\nu = 1, \ldots, \min\{n-1, k\}$ lautet der ν-te Eliminationsschritt, solange das Verfahren nicht vorher abgebrochen ist:

(ii) $\begin{cases} \text{Wähle } (i_\nu, j_\nu) \in \{\nu, \ldots, n\} \times \{\nu, \ldots, k\} \text{ mit} \\ |\gamma^{(\nu)}_{\pi_\nu(i_\nu), \sigma_\nu(j_\nu)}| = \max\{|\gamma^{(\nu)}_{\pi_\nu(i), \sigma_\nu(j)}| : i = \nu, \ldots, n;\ j = \nu, \ldots, k\}. \end{cases}$

Ist dieses maximale Element Null, bricht das Verfahren hier ab, sonst setzt man

(iii) $\pi_{\nu+1}(\nu) = \pi_\nu(i_\nu),\ \pi_{\nu+1}(i_\nu) = \pi_\nu(\nu),\ \pi_{\nu+1}(i) = \pi_\nu(i)$ sonst,

(iv) $\sigma_{\nu+1}(\nu) = \sigma_\nu(j_\nu),\ \sigma_{\nu+1}(j_\nu) = \sigma_\nu(\nu),\ \sigma_{\nu+1}(j) = \sigma_\nu(j)$ sonst,

und bestimmt $C^{(\nu+1)}$ durch die folgenden Beziehungen, in denen kurz $\pi := \pi_{\nu+1}$, $\sigma := \sigma_{\nu+1}$ gesetzt ist:

(v) $\gamma^{(\nu+1)}_{\pi(i),\sigma(j)} := \gamma^{(\nu)}_{\pi(i),\sigma(j)}$ $\qquad (i = 1, \ldots, \nu;\ j = 1, \ldots, k + l)$;

(vi) $\gamma^{(\nu+1)}_{\pi(i),\sigma(\nu)} := 0$ $\qquad (i = \nu + 1, \ldots, n)$,

und weiter mit

(vii) $d_{\pi(i),\sigma(\nu)} := \dfrac{\gamma^{(\nu)}_{\pi(i),\sigma(\nu)}}{\gamma^{(\nu)}_{\pi(\nu),\sigma(\nu)}}$ $\qquad (i = \nu + 1, \ldots, n)$:

(viii) $\gamma^{(\nu+1)}_{\pi(i),\sigma(j)} := \gamma^{(\nu)}_{\pi(i),\sigma(j)} - d_{\pi(i),\sigma(\nu)}\,\gamma^{(\nu)}_{\pi(\nu),\sigma(j)}$ $(i = \nu + 1, \ldots, n;\ j = \nu + 1, \ldots, k + l)$.

Da die $\gamma^{(\nu+1)}_{i,j}$ die gleichen Speicherplätze wie die $\gamma^{(\nu)}_{i,j}$ belegen, erscheint (v) nicht als Anweisung in einem Programm. Auf das Abspeichern der Nullen in (vi) kann man verzichten und statt dessen die $d_{\pi(i),\sigma(\nu)}$ nach (vii) einspeichern: vergleiche dazu (2.2.20)!

Für die einfacheren Fälle notieren wir gegenüber (2.2.15) folgende Änderungen:

(2.2.16) *Algorithmus bei halbmaximaler Pivotwahl:*

Man setzt $\sigma_\nu = \mathrm{id}$ für alle ν, dann erübrigt sich ein Speichern der $\sigma_\nu(j)$, und die Vorschrift (iv) in (2.2.15) entfällt; (ii) ist durch

(ii′) $\begin{cases} \text{Wähle } i_\nu \in \{\nu, \ldots, n\} \text{ mit} \\ |\gamma^{(\nu)}_{\pi_\nu(i_\nu),\nu}| = \max\limits_{i=\nu}^{n} |\gamma^{(\nu)}_{\pi_\nu(i),\nu}| \end{cases}$

zu ersetzen.

(2.2.17) *Algorithmus bei diagonaler Pivotwahl:*
Es ist stets

$$\pi_\nu = \mathrm{id}, \quad \sigma_\nu = \mathrm{id}$$

zu setzen, die Vorschriften (ii), (iii) und (iv) in (2.2.15) entfallen.

In dem wichtigen Spezialfall

$k = n$, A invertierbar, $l = 1$, also $B = b \in \mathbb{C}^n$

wollen wir die Lösung $x = (x_i)_{i=1}^n$ des Gleichungssystems

$$Ax = b$$

angeben. Nach der Behauptung (ii) von Satz (2.2.6) haben wir das Gleichungssystem

$$Ry = s$$

zu lösen, wobei R wegen der Invertierbarkeit von A eine invertierbare obere Dreiecksmatrix ist, die wir zusammen mit s nach $m = n$ Eliminationsschritten erreichen. Nach unserem Algorithmus haben wir mit

$$\pi := \pi_n, \quad \sigma := \sigma_n$$

die Darstellung

$$R = (\gamma^{(n)}_{\pi(i),\sigma(j)})_{(n,n)}, \quad s = (\gamma^{(n)}_{\pi(i),n+1})_{i=1}^n,$$

letzteres wegen $\sigma(n+1) = n+1$. $y = (y_i)_{i=1}^n$ hat wegen $x = Qy$ die Komponenten

$$y_i = e_i^t y = e_i^t Q^t x = (Qe_i)^t x = x_{\sigma(i)} \qquad (i = 1, \ldots, n).$$

Läßt man den oberen Index n in den Koeffizienten von R und s weg, so ergibt sich aus (2.1.4) durch einfaches Einsetzen:

$$(2.2.18) \qquad \begin{cases} x_{\sigma(n)} = \dfrac{\gamma_{\pi(n),n+1}}{\gamma_{\pi(n),\sigma(n)}} \\ x_{\sigma(\nu)} = \dfrac{1}{\gamma_{\pi(\nu),\sigma(\nu)}} \left(\gamma_{\pi(\nu),n+1} - \displaystyle\sum_{j=\nu+1}^{n} \gamma_{\pi(\nu),\sigma(j)} x_{\sigma(j)} \right) \quad (\nu = n-1, \ldots, 1) \end{cases}$$

In der bisherigen Formulierung der Gauß-Elimination wird die permutierte erweiterte Koeffizientenmatrix (PAQ, PB) des vorgegebenen Gleichungssystems zerlegt; dabei sind die wegen der Pivotwahl auftretenden Permutationsmatrizen P und Q von B unabhängig. Wenn wir die Matrix B weglassen, was dem Fall $l = 0$ entspricht, bekommen wir einen Algorithmus zur Zerlegung

$$PAQ = LR.$$

Hiermit läßt sich die Determinante einer (n,n)-Matrix A berechnen. Man hat nämlich

$$\det A = (\det P)^{-1} (\det Q)^{-1} \det L \det R,$$

wobei mit den zu P und Q gehörenden Zeilen- bzw. Spaltenpermutationen π und σ offenbar

$$\det P = \operatorname{sign}\pi, \quad \det Q = \operatorname{sign}\sigma, \quad \det L = 1$$

gilt. Das Vorzeichen von $\pi = \pi_m$ im Fall der halbmaximalen und maximalen Pivotwahl bestimmt man wegen (2.2.15, iii) rekursiv durch

$$\operatorname{sign}\pi_{\nu+1} = \begin{cases} \operatorname{sign}\pi_\nu, & \text{falls } i_\nu = \nu, \\ -\operatorname{sign}\pi_\nu, & \text{falls } i_\nu \neq \nu; \end{cases}$$

ebenso wird bei maximaler Pivotwahl $\operatorname{sign}\sigma$ ermittelt. Mit den Diagonalelementen $r_{i,i} = \gamma^{(m)}_{\pi_m(i),\,\sigma_m(i)}$ von R ergibt sich schließlich als

(2.2.19) **Bemerkung**.

$$\det A = \operatorname{sign}\pi \cdot \operatorname{sign}\sigma \cdot \prod_{i=1}^{n} r_{i,i}.$$

Wir kehren zum allgemeinen Fall der (n,k)-Matrix A zurück. Um nach der Zerlegung der nicht erweiterten Koeffizientenmatrix in

$$PAQ = LR$$

die Matrixgleichung $AX = B$ zu lösen – vergleiche (2.1.12)! –, müssen wir S als die Lösung von

$$LS = PB$$

berechnen und benötigen dazu die explizite Darstellung von L.

(2.2.20) **Bemerkung.** $L = (l_{i,j})_{(n,n)}$ besitzt die Koeffizienten

$$l_{i,j} = \begin{cases} 1 & (i = j) \\ d_{\pi_m(i),\,\sigma_m(j)} & (i > j) \\ 0 & \text{sonst}, \end{cases}$$

wobei die Größen

$$d_{\pi_{\nu+1}(i),\,\sigma_{\nu+1}(\nu)} \qquad (1 \leqq \nu \leqq m-1;\ \nu+1 \leqq i \leqq n)$$

durch (2.2.15, vii) erklärt sind.

Der *Beweis* wird dem Leser als Übungsaufgabe 2.3 überlassen. Insbesondere erfordert die Bestimmung von L keine zusätzlichen Rechenschritte.

Den Rechenaufwand der Gauß-Elimination wollen wir nur für den Fall $k \geqq n$, $m = n$ bestimmen. Die Ausführung von (2.2.15, vii) beim ν-ten Schritt $(1 \leqq \nu \leqq n-1)$ erfordert $(n-\nu)$ Divisionen, hinzu kommen $(n-\nu)(k+l-\nu)$

Multiplikationen nach (2.2.15, viii). Als Gesamtzahl der Multiplikationen und Divisionen erhält man

$$\sum_{\nu=1}^{n-1}(n-\nu)(k+l+1-\nu)=\sum_{\mu=1}^{n-1}\mu(k+l+1-n+\mu)=$$

$$=(k+l+1-n)\frac{n(n-1)}{2}+\frac{n(n-1)(2n-1)}{6}=\frac{n(n-1)}{2}\left(k+l+1-\frac{n+1}{3}\right).$$

Besonders wichtig ist der Fall $k = n$, $l = 0$; wir notieren die

(2.2.21) **Bemerkung.** Ist A (n,n)-Matrix, so benötigt die Zerlegung

$$PAQ = LR$$

gerade

$\frac{1}{3}n(n-1)(n+1)$ Operationen.

Zur Lösung von $Ax = b$ ist s mit $Ls = Pb$ zu berechnen. Bei expliziter Auflösung des Gleichungssystems $Ls = Pb$ benötigt man, wie man analog zu (2.1.5) zeigt, $\frac{1}{2}n(n-1)$ Operationen. Um die gleiche Zahl von Operationen wächst der Rechenaufwand gegenüber (2.2.21), wenn man s durch Zerlegung der erweiterten Matrix $(PAQ, Pb) = L(R, s)$ bestimmt. Schärfer zeigt man, daß alle Rechenschritte und Zwischenergebnisse übereinstimmen, gleichgültig ob man s über das entsprechende Gleichungssystem oder zusammen mit der Zerlegung von PAQ ermittelt. Für welche der Möglichkeiten man sich entscheidet, ist also nur eine Frage der Programmorganisation. Wir erhalten zusammenfassend, wenn wir den Rechenaufwand zur Lösung von $Ry = s$ nach (2.1.5) berücksichtigen, den

(2.2.22) **Satz.** *Die Lösung des Gleichungssystems* $Ax = b$ *mit der invertierbaren* (n,n)*-Matrix* A, *dem* (n,1)*-Vektor* b *unter Benutzung der Gauß-Elimination erfordert*

$\frac{1}{3}n^3+n^2-\frac{n}{3}$ *Multiplikationen bzw. Divisionen.*

Zusätzliche Rechenzeit benötigen die Vergleichsoperationen (2.2.15, ii) bzw. (2.2.16, ii'), deren Anzahl bei maximaler Pivotwahl etwa der Zahl der Multiplikationen entspricht, bei halbmaximaler Pivotwahl erheblich geringer ist. Im Folgenden geben wir Bedingungen an, unter denen halbmaximale und diagonale Pivotwahl möglich ist.

(2.2.23) **Satz.** *Im Fall einer* (n,k)*-Matrix* A *mit* $k \leq n$ *und* $\operatorname{rg} A = k$, *speziell bei invertierbarer* (n,n)*-Matrix* A, *ist die Gauß-Elimination mit halbmaximaler Pivotwahl durchführbar; das Verfahren bricht erst nach* $m - 1 = \min\{n-1, k\}$ *Schritten ab.*

Beweis. Da die Matrix B keinen Einfluß auf die Pivotwahl hat, betrachten wir die im Beweis zu (2.2.6) auftretenden $C^{(\nu)}$ für den Fall $l = 0$. Wegen $\operatorname{rg} A = k$ besitzt A keine Nullspalte, daher ist beim 1. Eliminationsschritt das Pivotelement aus der

1. Spalte wählbar. – Nach $(\nu-1)$ Eliminationsschritten $(2 \leqq \nu \leqq k)$ mit halbmaximaler Pivotwahl erhält man

$$C^{(\nu)} = \begin{pmatrix} c^{(\nu)}_{1,1} & \cdots & \cdots & c^{(\nu)}_{1,k} \\ 0 & \ddots & & \vdots \\ \vdots & & c^{(\nu)}_{\nu-1,\nu-1} & \vdots \\ \vdots & & 0 \quad c^{(\nu)}_{\nu,\nu} & \vdots \\ \vdots & & \vdots & \vdots \\ 0 & \cdots\cdots\ 0 & c^{(\nu)}_{n,\nu} \cdots\cdots & c^{(\nu)}_{n,k} \end{pmatrix}$$

Unter der Annahme

$$c^{(\nu)}_{\nu,\nu} = \ldots = c^{(\nu)}_{n,\nu} = 0 \,,$$

wären die ersten ν Spalten und damit alle k Spalten von $C^{(\nu)}$ linear abhängig; dagegen folgt aus (2.2.14), nämlich

$$C^{(\nu)} = L^{(\nu)} P_\nu A \,,$$

die Beziehung

$$\operatorname{rg} C^{(\nu)} = \operatorname{rg} A = k$$

und hieraus die lineare Unabhängigkeit der Spalten von $C^{(\nu)}$. Im Fall $\nu \leqq m-1$ kann also der ν-te Eliminationsschritt mit einem Pivotelement aus der ν-ten Spalte durchgeführt werden.

Da bei der halbmaximalen Pivotwahl keine Spaltenvertauschungen auftreten, bemerken wir zur Gestalt von $C^{(\nu)}$ die Beziehungen

$$c^{(\nu)}_{i,j} = \hat{c}^{(i)}_{i,j} \qquad (1 \leqq i \leqq \nu-1;\ i \leqq j \leqq k).$$

Bei diagonaler Pivotwahl fallen auch die Zeilenvertauschungen weg, nach $(\nu-1)$ Eliminationsschritten $(\nu \geqq 2)$ erhält man die Gestalt

$$C^{(\nu)} = \begin{pmatrix} c^{(1)}_{1,1} & \cdots & \cdots & c^{(1)}_{1,k} \\ 0 & c^{(2)}_{2,2} & \cdots & c^{(2)}_{2,k} \\ \vdots & \ddots & & \vdots \\ \vdots & & c^{(\nu)}_{\nu,\nu} & \vdots \\ \vdots & & \vdots & \vdots \\ 0 & \cdots\ 0 & c^{(\nu)}_{n,\nu} \cdots & c^{(\nu)}_{n,k} \end{pmatrix}; \tag{2.2.24}$$

Unter der Annahme, daß diagonale Pivotwahl bis zum letzten Eliminationsschritt möglich ist, werden P und Q zu I_n bzw. I_k, was die Zerlegbarkeit von A in

$$A = LR$$

nach sich zieht. Diese Zerlegung ist nicht immer möglich, als Gegenbeispiel betrachten wir eine beliebige invertierbare (n,n)-Matrix $A = (a_{i,j})_{(n,n)}$ mit $a_{1,1} = 0$. Aus

$$A = LR, \quad L = (l_{i,j})_{(n,n)}, \quad R = (r_{i,j})_{(n,n)}$$

würde die Invertierbarkeit von L und R und damit

$$a_{1,1} = l_{1,1} r_{1,1} \neq 0$$

folgen. – Prinzipiell ist die diagonale Pivotwahl nicht zu empfehlen, auch wenn sie durchführbar sein sollte; es können nämlich sehr große Koeffizienten von L auftreten und numerische Instabilität verursachen: man vergleiche dazu die Übungsaufgabe 2.1 und die Ausführungen des Abschnitts 3.6: dort wird auch auf die Frage der Rundungsfehler bei maximaler und halbmaximaler Pivotwahl eingegangen.

Eine wichtige Klasse von Matrizen, die eine LR-Zerlegung besitzen und sich eventuell zur Gauß-Elimination bei diagonaler Pivotwahl eignen, bilden die positiv definiten Matrizen. Wir wollen das Skalarprodukt im $\mathbb{C}^n$ wie üblich durch

$$(x, y) := y^* x = \sum_{i=1}^{n} x_i \overline{y}_i \qquad (x = (x_i)_{i=1}^n,\ y = (y_i)_{i=1}^n \in \mathbb{C}^n)$$

erklären. Eine komplexe (n,n)-Matrix A heißt bekanntlich positiv definit, wenn A hermitesch, also $A = A^*$ ist und zusätzlich

$$\forall\, x \in \mathbb{C}^n,\ \neq 0 \qquad (Ax, x) > 0$$

gilt. Wir zeigen dazu den

(2.2.25) **Satz.** *Eine positiv definite* (n,n)*-Matrix* A *besitzt eine* LR*-Zerlegung, und die Diagonalelemente von* R *sind positiv.*

Beweis. Wir haben

$$c_{1,1}^{(1)} = a_{1,1} = (Ae_1, e_1) > 0$$

nach Voraussetzung, daher ist der erste Eliminationsschritt ohne Vertauschungen möglich. Nach $(\nu - 1)$ Eliminationsschritten, ohne Vertauschungen durchgeführt, erhalten wir $C^{(\nu)}$ in der Gestalt (2.2.24) mit $k = n$, und nach (2.2.14) wegen $P_\nu = Q_\nu = I$

$$C^{(\nu)} = L^{(\nu)} A$$

mit einer normierten unteren Dreiecksmatrix $L^{(\nu)} := (l_{i,j}^{(\nu)})_{(n,n)}$. Für

$$x := L^{(\nu)*} e_\nu \neq 0$$

gilt nach Voraussetzung

$$0 < (Ax,x) = (Ax, L^{(\nu)*} e_\nu) = (L^{(\nu)} A L^{(\nu)*} e_\nu, e_\nu) = (C^{(\nu)} L^{(\nu)*} e_\nu, e_\nu) =$$

$$= (e_\nu^* C^{(\nu)})(L^{(\nu)*} e_\nu) = \sum_{i=\nu}^{n} c_{\nu,i}^{(\nu)} \overline{l}_{\nu,i}^{(\nu)} = c_{\nu,\nu}^{(\nu)} ,$$

letzteres wegen der Form von $L^{(\nu)}$. Ist $\nu \leqq n-1$, läßt sich auch der ν-te Eliminationsschritt mit $c_{\nu,\nu}^{(\nu)}$ als Pivotelement durchführen. Schließlich wird

$$R = C^{(n)} = \begin{pmatrix} c_{1,1}^{(1)} & \cdots & c_{1,n}^{(1)} \\ & \ddots & \vdots \\ 0 & & \\ & & c_{n,n}^{(n)} \end{pmatrix}$$

mit $c_{\nu,\nu}^{(\nu)} > 0 \quad (\nu = 1, \ldots, n)$.

2.3. Kompakter Gauß-Algorithmus

Wir betrachten wieder das Gleichungssystem

$$Ax = B$$

mit einer (n,k)-Matrix A und einer (n,l)-Matrix B. Dabei setzen wir voraus, daß $k \geq n$ ist und im Sinne von Satz (2.2.6) $(n-1)$ Gauß-Eliminationsschritte bei diagonaler Pivotwahl ausführbar sind. Dann gilt für

$$C := (A,B)$$

die Dreieckszerlegung

$$C := LC^{(n)}$$

oder ausgeschrieben, mit $m := k + l$:

$$\begin{pmatrix} c_{1,1} & \cdots & c_{1,m} \\ \vdots & & \vdots \\ c_{n,1} & \cdots & c_{n,m} \end{pmatrix} = \begin{pmatrix} 1 & & 0 \\ d_{2,1} & \ddots & \\ \vdots & & \\ d_{n,1} & \cdots & d_{n,n-1} \; 1 \end{pmatrix} \begin{pmatrix} c_{1,1}^{(1)} & \cdots & c_{1,m}^{(1)} \\ & c_{2,2}^{(2)} & \vdots \\ 0 & & c_{n,n}^{(n)} \cdots c_{n,m}^{(n)} \end{pmatrix} ,$$

also

$$(2.3.1)\qquad \begin{cases} c_{i,j} = \sum_{\kappa=1}^{i-1} d_{i,\kappa}\, c_{\kappa,j}^{(\kappa)} + c_{i,j}^{(i)} & \text{für } i \le j, \\ c_{i,j} = \sum_{\kappa=1}^{j} d_{i,\kappa}\, c_{\kappa,j}^{(\kappa)} & \text{für } i > j. \end{cases}$$

Auf Grund dieser Beziehung lassen sich die Koeffizienten von L und $C^{(n)}$ in geeigneter Reihenfolge rekursiv berechnen. Wir gehen nach Crout vor, indem wir für $\nu = 1, \ldots, n$ die ν-te Zeile von $C^{(n)}$ und anschließend die ν-te Spalte von L bestimmen.

Wir erhalten als

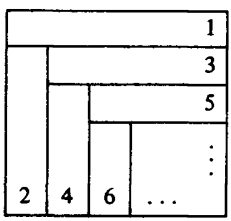
(2.3.2) *Algorithmus bei diagonaler Pivotwahl.*

$$\text{(i)}\quad \begin{cases} c_{1,j}^{(1)} := c_{1,j} & (j = 1, \ldots, m), \\ d_{i,1} := \dfrac{c_{i,1}}{c_{1,1}^{(1)}} & (i = 2, \ldots, n) \end{cases}$$

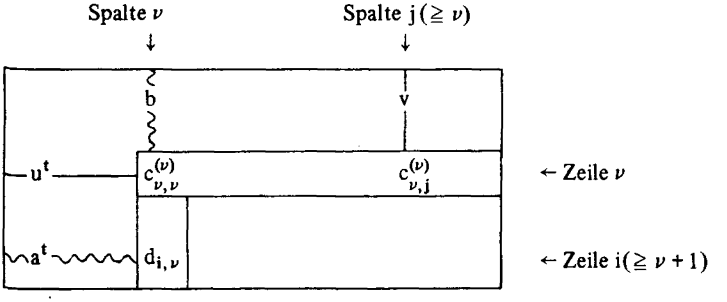
und für $\nu = 2, 3, \ldots, n$:

$$\text{(ii)}\quad \begin{cases} c_{\nu,j}^{(\nu)} := c_{\nu,j} - \sum_{\kappa=1}^{\nu-1} d_{\nu,\kappa}\, c_{\kappa,j}^{(\kappa)} & (j = \nu, \ldots, m), \\ d_{i,\nu} := \dfrac{1}{c_{\nu,\nu}^{(\nu)}} \left(c_{i,\nu} - \sum_{\kappa=1}^{\nu-1} d_{i,\kappa}\, c_{\kappa,\nu}^{(\kappa)} \right) & (i = \nu+1, \ldots, n), \quad \text{falls } \nu \le n-1. \end{cases}$$

Zur Veranschaulichung des Algorithmus und zum Rechnen von Hand verwenden wir ein rechteckiges (n,m)-Schema, in dem auf und oberhalb der „Diagonalen" die Koeffizienten von $C^{(n)}$, unterhalb der Diagonalen die $d_{i,\nu}$ eingetragen werden. Das Schema wird in der nachstehenden Reihenfolge ausgefüllt:



Sobald die ersten $(\nu-1)$ Zeilen und Spalten des Schemas eingetragen sind, berechnen wir im $(2\nu-1)$-ten bzw. (2ν)-ten Schritt die ν-te Zeile und ν-te Spalte gemäß folgender Skizze:



Mit den – von ν, i, j abhängigen – $(\nu-1)$-Vektoren

$$u = (d_{\nu,\kappa})_{\kappa=1}^{\nu-1},\quad v = (c^{(\kappa)}_{\kappa,j})_{\kappa=1}^{\nu-1},\quad a = (d_{i,\kappa})_{\kappa=1}^{\nu-1},\quad b = (c^{(\kappa)}_{\kappa,\nu})_{\kappa=1}^{\nu-1}$$

ergibt sich nämlich

$$c^{(\nu)}_{\nu,j} = c_{\nu,j} - u^t v, \qquad d_{i,\nu} = \frac{1}{c^{(\nu)}_{\nu,\nu}}\,(c_{i,\nu} - a^t b).$$

Wenn man in dem hier betrachteten Fall der diagonalen Pivotwahl – d.h. $\pi_\nu = \mathrm{id}$, $\sigma_\nu = \mathrm{id}$, $\gamma^{(\nu)}_{i,j} = c^{(\nu)}_{i,j}$ – die Formeln (2.2.15, viii) bezüglich ν aufsummiert, so erhält man für $\mu \leqq \min\{i,j\}$

$$c^{(\mu)}_{i,j} = c_{i,j} - \sum_{\kappa=1}^{\mu-1} d_{i,\kappa}\, c^{(\kappa)}_{\kappa,j}\,.$$

Wendet man also die in (2.3.2, ii) auftretenden Rechenoperationen in der natürlichen Reihenfolge an, so sind die Zwischensummen die gerade erwähnten $c^{(\mu)}_{\nu,j}$ bzw. $c^{(\mu)}_{i,\nu}$ $(\mu \leqq \nu)$. Daher stimmen sämtliche Rechenschritte im kompakten Gauß-Algorithmus mit denen der ausführlichen Form der Gauß-Elimination überein. Da die Zwischenergebnisse nicht explizit erscheinen, liefert der kompakte Gauß-Algorithmus für das Rechnen von Hand ein knappes und übersichtliches Rechenschema.

(2.3.3) **Zahlenbeispiel.** Wir betrachten das Gleichungssystem $Ax = b$,

$$A = \begin{pmatrix} 1{,}1 & 3{,}1 & 1{,}8 & 2{,}3 \\ 3{,}2 & -4{,}1 & 2{,}5 & 8{,}3 \\ 4{,}7 & 0{,}21 & 6{,}7 & 1{,}9 \\ 0{,}5 & 7{,}3 & 1{,}3 & 7{,}1 \end{pmatrix}, \qquad b = \begin{pmatrix} 1{,}2 \\ 3{,}4 \\ 5{,}6 \\ 7{,}3 \end{pmatrix}.$$

Aus der erweiterten Koeffizientenmatrix $C = (A, b)$ berechnen wir mit 3-stelliger dezimaler Gleitkommarechnung das Schema:

	R				s
	1,10	3,10	1,80	2,30	1,20
	2,91	−13,1	−2,74	1,61	−0,0900
L	4,27	0,992	1,73	−9,52	0,569
	0,455	−0,450	−0,433	2,67	6,96

und hieraus durch Lösung von $Rx = s$:

$$x_4 = 2{,}61; \quad x_3 = 14{,}6; \quad x_2 = -2{,}73; \quad x_1 = -20{,}5.$$

Zum Vergleich notieren wir die auf 3 Dezimalstellen gerundeten Komponenten der exakten Lösung:

$$x_4 = 2{,}62; \quad x_3 = 14{,}7; \quad x_2 = -2{,}75; \quad x_1 = -20{,}8.$$

Zum Programmieren bietet der kompakte Gauß-Algorithmus nur dann Vorteile, wenn die Maschine Skalarprodukte von Vektoren mit erhöhter Genauigkeit rechnet (akkumulierende Gleitkommarechnung, vgl. Wilkinson [31], S. 28 ff.).

Wegen der diagonalen Pivotwahl kann das Verfahren numerisch instabil werden oder ganz versagen. Beim Rechnen von Hand sieht man während der Durchführung, ob ein Diagonalelement Null wird oder zu große Koeffizienten von L entstehen, und kann sich gegebenenfalls durch Zeilenvertauschungen korrigieren. Im Beispiel (2.3.3) ist der größte Koeffizient von L kleiner als 5, daher gelangen wir ohne Zeilenvertauschungen zu einer numerisch akzeptablen Lösung.

Maximale Pivotwahl ist im kompakten Gauß-Algorithmus nicht möglich; wir wollen eine Variante mit halbmaximaler Pivotwahl angeben.

Dazu setzen wir, schwächer als bisher, voraus, daß $(n-1)$ Gauß-Eliminationsschritte bei halbmaximaler Pivotwahl möglich sind, so daß gilt
wahl möglich sind, so daß gilt

$$PC = LC^{(n)},$$

wobei die Permutationsmatrix P durch $\pi \in S_n$ mit $e_i^t P = e_{\pi(i)}^t$ $(i = 1, \ldots, n)$ beschrieben sei. Wir setzen $m = k + l$ und – vgl. (2.2.16), (2.2.20) –

$$C = (\gamma_{i,j})_{(n,m)}, \quad L = (d_{\pi(i),j})_{(n,n)}, \quad C^{(n)} = (\gamma^{(i)}_{\pi(i),j})_{(n,m)}.$$

Durch Ausmultiplikation erhalten wir

$$(2.3.4) \quad \begin{cases} \text{(i)} & \gamma_{\pi(i),j} = \sum_{\kappa=1}^{i-1} d_{\pi(i),\kappa}\, \gamma^{(\kappa)}_{\pi(\kappa),j} + \gamma^{(i)}_{\pi(i),j} & \text{für } i \leqq j, \\ \text{(ii)} & \gamma_{\pi(i),j} = \sum_{\kappa=1}^{j} d_{\pi(i),\kappa}\, \gamma^{(\kappa)}_{\pi(\kappa),j} & \text{für } i > j. \end{cases}$$

Die Permutation π ist durch die im Algorithmus (2.2.16) auftretenden π_ν ($\nu = 2, \ldots, n$) als $\pi = \pi_n$ gegeben; auf Grund der Rekursionen (2.2.15, iii), in denen wegen (2.2.16, ii') stets $i_\nu \geqq \nu$ gilt, haben wir

(*) $\quad \pi_\nu(i) = \pi_n(i) = \pi(i) \qquad (i = 1, \ldots, \nu - 1)$

und folglich

(**) $\quad \{\pi_\nu(\nu), \ldots, \pi_\nu(n)\} = \{\pi(\nu), \ldots, \pi(n)\}$.

Die zur Pivotwahl nach (2.2.16, ii') benötigten $\gamma^{(\nu)}_{\pi(i),\nu}$ ($i = \nu, \nu+1, \ldots, n$) genügen wegen (2.2.15, vii) und (**) den Gleichungen

$$\gamma^{(\nu)}_{\pi(i),\nu} = \gamma^{(\nu)}_{\pi(\nu),\nu}\, d_{\pi(i),\nu} \qquad (i = \nu + 1, \ldots, n)$$

und besitzen daher nach (2.3.4, ii) mit $j = \nu$ bzw. (2.3.4, i) mit $i = j = \nu$ die Darstellungen

$$(2.3.5) \qquad \gamma^{(\nu)}_{\pi(i),\nu} = \gamma_{\pi(i),\nu} - \sum_{\kappa=1}^{\nu-1} d_{\pi(i),\kappa}\, \gamma^{(\kappa)}_{\pi(\kappa),\nu} \qquad (i = \nu, \ldots, n)$$

Nach diesen Vorbemerkungen notieren wir den

(2.3.6) *Algorithmus bei halbmaximaler Pivotwahl.*

1. Schritt: Man wählt $i_1 \in \{1, \ldots, n\}$ mit $|\gamma_{i_1,1}| = \max_{i=1}^{n} |\gamma_{i,1}|$, setzt

$$\pi_2(1) = i_1, \quad \pi_2(i_1) = 1, \quad \pi_2(i) = i \qquad \text{sonst}$$

und weiter

(i) $\quad \gamma^{(1)}_{\pi_2(1),j} = \gamma_{\pi_2(1),j} \qquad (j = 1, \ldots, m)$,

(ii) $\quad d_{\pi_2(i),1} = \dfrac{\gamma_{\pi_2(i),1}}{\gamma_{\pi_2(1),1}} \qquad (i = 2, \ldots, n)$.

ν-ter Schritt ($2 \leqq \nu \leqq n$): Unter der Annahme, daß π_ν und die Koeffizienten $\gamma^{(i)}_{\pi_\nu(i),j}$ für $1 \leqq i \leqq \nu - 1$, $i \leqq j \leqq m$ sowie $d_{\pi_\nu(i),j}$ für $1 \leqq j \leqq \nu - 1$, $j + 1 \leqq i \leqq n$ vorliegen, berechnen wir

(iii) $\quad \gamma^{(\nu)}_{\pi_\nu(i),\nu} = \gamma_{\pi_\nu(i),\nu} - \sum_{\kappa=1}^{\nu-1} d_{\pi_\nu(i),\kappa}\, \gamma_{\pi_\nu(\kappa),\nu} \qquad (i = \nu, \ldots, n)$.

Im Fall $\nu \leqq n - 1$ wählen wir i_ν nach (2.2.16, ii') und bestimmen $\pi_{\nu+1}$ nach (2.2.15, iii); für $\nu = n$ wird $\pi_{n+1} = \pi_n$. Da mit (iii) bereits das Pivotelement $\gamma^{(\nu)}_{\pi_{\nu+1}(\nu),\nu}$ berechnet ist, setzen wir weiter:

(iv) $\quad \gamma^{(\nu)}_{\pi_{\nu+1}(\nu),j} = \gamma_{\pi_{\nu+1}(\nu),j} - \sum_{\kappa=1}^{\nu-1} d_{\pi_{\nu+1}(\nu),\kappa}\, \gamma_{\pi_{\nu+1}(\kappa),j} \qquad (j = \nu + 1, \ldots, m)$

und im Fall $\nu \leqq n-1$:

$$(v) \quad d_{\pi_{\nu+1}(i),\nu} = \frac{1}{\gamma^{(\nu)}_{\pi_{\nu+1}(\nu),\nu}} \gamma^{(\nu)}_{\pi_{\nu+1}(i),\nu} \qquad (i = \nu+1, \ldots, n).$$

Wir wollen nachweisen, daß die so konstruierten Größen den Beziehungen (2.3.4) und (2.3.5) genügen: auf Grund von (*) entsprechen (i) und (iv) den Gleichungen (2.3.4, i) mit $i = 1$ bzw. $i = \nu$; außerdem darf man in der in (iii) auftretenden Summe $\pi_\nu(\kappa)$ durch $\pi(\kappa)$ ersetzen. Weiter folgern wir aus (**), daß (ii) und (iii) den Beziehungen (2.3.4, ii) mit $j = 1$ bzw. (2.3.5) entsprechen, wobei nur die Reihenfolge der Gleichungen geändert ist. Durch Einsetzen von (iii) ergibt sich schließlich die Äquivalenz von (v) zu (2.3.4, ii) mit $j = \nu$. – Bei vorgegebenem π sind die Koeffizienten von L und $C^{(n)}$ durch (2.3.4) eindeutig bestimmt, wie man durch Induktion zeigt (vgl. (2.3.2)!); daher werden mit (2.3.6) die gleichen Zerlegungsmatrizen L und $C^{(n)}$ wie im Algorithmus (2.2.16) berechnet. Wie man sich überlegt, stimmen dabei auch die jeweiligen Zwischenergebnisse überein.

Bei Verwendung von (2.3.6) zum Rechnen von Hand kommt man nicht mehr mit einem (n,m)-Schema aus: sobald nämlich die ersten $(\nu-1)$ Zeilen von $C^{(n)}$ und $(\nu-1)$ Spalten von L vorliegen, sind zuerst die Hilfsgrößen in (2.3.6, iii) zu berechnen, und nach der Pivotwahl müssen in C und in den schon bekannten Spalten von L die Zeilen permutiert werden, erst dann werden die ν-te Zeile von $C^{(n)}$ und ν-te Spalte von L ergänzt. Wir überlassen es dem Leser, sich dieses Vorgehen an Hand der Übungsaufgabe 2.6 zu verdeutlichen. – Zum Programmieren ist der Algorithmus (2.3.6) besonders dann zu empfehlen, wenn die Maschine die auftretenden Produktsummen mit erhöhter Genauigkeit rechnet.

2.4. Cholesky-Zerlegung

Die Cholesky-Zerlegung liefert ein dem kompakten Gauß-Algorithmus verwandtes Verfahren bei positiv definiten Matrizen, das auf die Symmetrie dieser Matrizen besser eingeht. Wir notieren zunächst

(2.4.1) **Hilfssatz.** *Es sei* A *eine invertierbare* (n,n)*-Matrix mit einer* LR*-Zerlegung. Dann besitzt* A *eine Darstellung*

$$A = LDM,$$

in der L *eine normierte untere Dreiecksmatrix,* M *eine normierte obere Dreiecksmatrix und* D *eine Diagonalmatrix bezeichne. Diese Zerlegung ist eindeutig.*

Beweis. In der Darstellung $A = LR$ ist $R = (r_{i,j})_{(n,n)}$ invertierbare obere Dreiecksmatrix, die wir wegen $r_{i,i} \neq 0$ als

$$R = DM, \quad D := \operatorname{diag}(r_{1,1}, \ldots, r_{n,n})$$

schreiben können, womit die Existenz der angegebenen Zerlegung gezeigt ist.

Zum Nachweis der Eindeutigkeit nehmen wir zwei Darstellungen von A mit den angegebenen Eigenschaften, also

$$A = LDM = \tilde{L}\tilde{D}\tilde{M}$$

an. Dann wird $\hat{D}$, definiert durch

$$\hat{D} := \tilde{L}^{-1} LD = \tilde{D}\tilde{M}M^{-1}$$

gleichzeitig obere und untere Dreiecksmatrix, also Diagonalmatrix. Da $\tilde{L}^{-1}L$ und $\tilde{M}M^{-1}$ normierte Dreiecksmatrizen sind, liefert ein Vergleich der Diagonalelemente $\hat{D} = \tilde{D} = D$ und daher auch $\tilde{L}^{-1}L = \tilde{M}M^{-1} = I$, d.h. $L = \tilde{L}$ und $M = \tilde{M}$.

Die theoretische Begründung der Cholesky-Zerlegung ist der

(2.4.2) **Satz.** *Zu einer positiv definiten Matrix* A *gibt es eine obere Dreiecksmatrix* R *mit*

$$A = R^*R\,.$$

Beweis. Nach Satz (2.2.25) erfüllt A die Voraussetzung von Hilfssatz (2.4.1), und in der Zerlegung

$$A = LDM$$

sind die Diagonalelemente von D nach Konstruktion positiv. Außerdem ergibt sich

$$A = A^* = M^*D^*L^* = M^*DL^*\,,$$

mithin wegen der Eindeutigkeit der Zerlegung $L = M^*$. Bezeichnet $D = \operatorname{diag}(d_1, \ldots, d_n)$, so setzt man

$$\sqrt{D} = \operatorname{diag}(\sqrt{d_1}, \ldots, \sqrt{d_n}), \qquad R = \sqrt{D}\,M$$

und hat hiermit

$$A = (M^*\sqrt{D})(\sqrt{D}\,M) = (\sqrt{D}\,M)^*(\sqrt{D}\,M) = R^*R\,,$$

was zu zeigen war.

Zur Herleitung des entsprechenden Algorithmus setzen wir $A = (a_{i,j})_{(n,n)}$, $R = (r_{i,j})_{(n,n)}$ und erhalten aus $A = R^*R$ für die $a_{i,j}$ mit $j \geq i$ die Beziehungen

$$(2.4.3) \qquad a_{i,j} = \sum_{\kappa=1}^{n} \overline{r}_{\kappa,i} r_{\kappa,j} = \sum_{\kappa=1}^{i} \overline{r}_{\kappa,i} r_{\kappa,j} \qquad (1 \leq i \leq n;\ i \leq j \leq n)\,.$$

Da nach Satz (2.4.2) die $r_{i,i} = \sqrt{d_i} > 0$ gewählt werden können, ergibt sich hieraus der

(2.4.4) *Algorithmus der Cholesky-Zerlegung.*

Wegen $a_{1,1} = |r_{1,1}|^2$ mit $r_{1,1} > 0$ und $a_{1,j} = \overline{r}_{1,1} r_{1,j}$ $(j = 2, \ldots, n)$ setzen wir

$$(i) \quad \begin{cases} r_{1,1} = \sqrt{a_{1,1}}\ (>0), \\ r_{1,j} = \dfrac{a_{1,j}}{r_{1,1}} \qquad (j = 2, \ldots, n)\,. \end{cases}$$

Für $2 \leqq i \leqq n$ setzen wir nach Berechnung der ersten $(i-1)$ Zeilen von R

$$\text{(ii)} \quad r_{i,i} = \left(a_{i,i} - \sum_{\kappa=1}^{i-1} |r_{\kappa,i}|^2 \right)^{\frac{1}{2}}$$

und im Fall $i \leqq n-1$ anschließend

$$\text{(iii)} \quad r_{i,j} = \frac{1}{r_{i,i}} \left(a_{i,j} - \sum_{\kappa=1}^{i-1} \bar{r}_{\kappa,i} r_{\kappa,j} \right) \qquad (j = i+1, \ldots, n).$$

Im Verlauf des Algorithmus sind n Quadratwurzeln zu berechnen, hinzu kommen $[(n-i+1)\,i-1]$ Multiplikationen bzw. Divisionen zur Bestimmung der i-ten Zeile von R. Durch Aufsummieren erhalten wir:

(2.4.5) **Bemerkung.** Die Cholesky-Zerlegung einer positiv definiten (n,n)-Matrix A benötigt

n Wurzelberechnungen und

$\frac{1}{6}\, n(n^2+3n-4) \approx \frac{1}{6}\, n^3$ Multiplikationen bzw. Divisionen.

Bei großem n fallen die Wurzeln gegenüber den anderen Rechenoperationen nicht sehr ins Gewicht; wie ein Vergleich mit (2.2.21) zeigt, benötigt die Cholesky-Zerlegung gegenüber der Gauß-Elimination etwa den halben Rechenaufwand. Vorteilhaft für die numerische Stabilität ist die Beschränktheit der Koeffizienten von R, nämlich

$$|r_{\kappa,i}|^2 \leqq a_{i,i} \qquad (\kappa \leqq i);$$

hierzu und zum Vergleich mit der Gauß-Elimination verweisen wir auf die Aufgabe 2.7.

Für eine beliebige invertierbare (n,n)-Matrix A ist A^*A hermitesch und wegen

$$(A^*Ax, x) = (Ax, Ax) > 0 \quad \text{für } x \neq 0$$

auch positiv definit. Das Gleichungssystem $AX = B$ ist dem System

$$\text{(2.4.6)} \qquad (A^*A)\,X = A^*B$$

äquivalent. Aus zwei Gründen empfiehlt es sich jedoch nicht, auf (2.4.6) überzugehen; erstens erfordert die Berechnung von A^*A zusätzlich $\frac{1}{2}\, n^2 (n+1)$ Multiplikationen, zweitens kann auch bei invertierbarem A der numerisch ermittelte Wert von A^*A eine nichtinvertierbare Matrix werden. Dazu betrachten wir das

(2.4.7) **Beispiel.** A sei die folgende, aus t-stelligen Gleitkommazahlen bestehende Matrix

$$A = \begin{pmatrix} 1+\epsilon & 2-\eta \\ 1-\epsilon & 2+\eta \end{pmatrix}$$

mit $\epsilon, \eta \geqq 2\tau = 10^{-t+1}$ und $\epsilon^2, \eta^2 < \frac{1}{2}\tau$. Dann gilt exakt bzw. bei t-stelliger Gleitkommarechnung

$$A^*A = \begin{pmatrix} 2+2\epsilon^2 & 4-2\epsilon\eta \\ 4-2\epsilon\eta & 8+2\eta^2 \end{pmatrix}, \qquad gl(A^*A) = \begin{pmatrix} 2 & 4 \\ 4 & 8 \end{pmatrix},$$

also ist $gl(A^*A)$ nicht invertierbar. Ein konkretes Zahlenbeispiel mit $t = 3$ bringt Aufgabe 2.8.

2.5. Jordan-Elimination

Bei der Gauß-Elimination bringen wir die erweiterte Koeffizientenmatrix $C = (A, B)$ des vorgegebenen Gleichungssystems schrittweise auf die Gestalt $C^{(m)} = (R, S)$ mit einer oberen Dreiecksmatrix R, indem wir bei jedem Eliminationsschritt gewisse Vielfache der Pivotzeile von allen *nachfolgenden* Zeilen subtrahieren. Eine naheliegende Modifikation dieses Verfahrens besteht darin, Vielfache der Pivotzeile von *allen übrigen* Zeilen zu subtrahieren, so daß A im wesentlichen in eine Diagonalmatrix übergeht.

(2.5.1) **Definition.** Vorgegeben sei eine (n,m)-Matrix $C = (c_{i,j})_{(n,m)} \neq 0$. Ein *Jordanscher Eliminationsschritt* mit dem Pivotelement $c_{r,s} \neq 0$ besteht aus

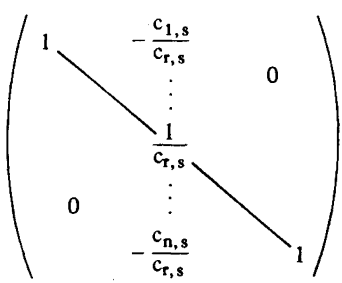
(i) Division der r-ten Zeile von C durch $c_{r,s}$, anschließend

(ii) Subtraktion geeigneter Vielfacher der r-ten Zeile von allen übrigen Zeilen,

so daß die s-te Spalte von C in die Einheitsspalte e_r übergeht.

(2.5.2) **Bemerkung.** Ein Jordanscher Eliminationsschritt bedeutet den Übergang von C auf $\tilde{C} = TC$ mit der invertierbaren (n,n)-Matrix

(i) $T =$



r-te Spalte

(ii) $\tilde{C} = (\tilde{c}_{i,j})_{(n,m)}$ besitzt die Koeffizienten

$$\tilde{c}_{r,j} = \frac{1}{c_{r,s}} c_{r,j} \qquad (j = 1, \ldots, m),$$

$$\tilde{c}_{i,j} = c_{i,j} - c_{i,s}\tilde{c}_{r,j} = c_{i,j} - \frac{c_{i,s}}{c_{r,s}} c_{r,j} \qquad (i \neq r;\ j = 1, \ldots, m).$$

Zur Lösung der Matrixgleichung $AX = B$ mit der (n,k)-Matrix A und der (n,l)-Matrix B setzt man $C^{(1)} = (A^{(1)}, B^{(1)}) := (A,B)$ und konstruiert sukzessive $C^{(\nu)} = (A^{(\nu)}, B^{(\nu)})$ $(\nu = 2, 3, \ldots)$ durch Anwendung von Jordanschen Eliminationsschritten, bei denen das Pivotelement stets aus der Teilmatrix $A^{(\nu)}$ und nicht aus einer früheren Pivotzeile oder -spalte genommen wird: in diesem Rahmen sind diagonale, halbmaximale und maximale Pivotwahl gebräuchlich.

Wir bemerken dazu, daß wir keine Zeilen- und Spaltenvertauschungen vornehmen: bei vollkommen analoger Formulierung zur Gauß-Elimination bestände der ν-te Jordansche Eliminationsschritt mit dem Pivotelement (r,s) in einer Zeilen- und einer Spaltenvertauschung, die das Pivotelement in die Diagonale bringt, und anschließender Transformation mit einem $\tilde{T}^{(\nu)}$ der Form (2.5.2, i) mit $r = s = \nu$, d.h.

$$C^{(\nu)} \to P_{r,\nu} C^{(\nu)} P_{s,\nu} \to \tilde{T}^{(\nu)} P_{r,\nu} C^{(\nu)} P_{s,\nu},$$

wobei die $P_{i,j}$ nach (2.2.5, vi) zu definieren sind. Um die Zeilen- und Spaltenvertauschungen weglassen zu können, müssen wir $T^{(\nu)}$ so wählen, daß

$$\tilde{T}^{(\nu)} P_{r,\nu} C^{(\nu)} P_{s,\nu} = P_{r,\nu} (T^{(\nu)} C^{(\nu)}) P_{s,\nu}$$

gilt. Dies erreichen wir offenbar mit der Matrix $T^{(\nu)} := P_{r,\nu} \tilde{T}^{(\nu)} P_{r,\nu}$ die die Gestalt (2.5.2, i) besitzt.

Das Verfahren bricht ab, sobald kein erlaubtes, von Null verschiedenes Pivotelement mehr existiert. Offenbar wird beim ν-ten Eliminationsschritt eine Einheitsspalte in $A^{(\nu+1)}$ erzeugt, ohne daß vorher gewonnene Einheitsspalten zerstört werden. Hieraus folgern wir den

(2.5.3) **Satz**.

(i) *Ist* A (n,k)*-Matrix mit* $k \geq n$ *und* $\operatorname{rg} A = n$, *so sind* n *Jordansche Eliminationsschritte mit maximaler Pivotwahl, im Fall* $k = n$ *auch mit halbmaximaler Pivotwahl möglich.*

(ii) *Unter der Voraussetzung wie in* (i) *bezeichne* $\pi(\nu)$ *und* $\sigma(\nu)$ *die Indizes der Pivotzeile bzw. -spalte beim* ν*-ten Eliminationsschritt, wobei* $\pi \in S_n$, $\sigma \in S_k$ *gewisse Permutationen sind. Dann besitzt* $A^{(n+1)}$ n *Einheitsspalten, nämlich*

$$A^{(n+1)} e_{\sigma(\nu)} = e_{\pi(\nu)} \qquad (\nu = 1, \ldots, n),$$

im Fall $k = n$ *ist also* $A^{(n+1)}$ *Permutationsmatrix.*

Der *Beweis* wird als Übungsaufgabe 2.9 empfohlen. – Da für alle $\nu = 1, \ldots, n$ stets

$$(A^{(\nu+1)}, B^{(\nu+1)}) = T^{(\nu)} (A^{(\nu)}, B^{(\nu)}) = (T^{(\nu)} A^{(\nu)}, T^{(\nu)} B^{(\nu)})$$

mit einer invertierbaren Matrix $T^{(\nu)}$ der Gestalt (2.5.2, i) gilt, ist für eine (k,l)-Matrix X offenbar

$$AX = B \iff A^{(n+1)} X = B^{(n+1)}.$$

Hieraus folgern wir unmittelbar

(2.5.4) **Bemerkung.** Mit $A^{(n+1)}$ gemäß (2.5.3, ii), $B^{(n+1)} =: (b_{i,j}^{(n+1)})_{(n,l)}$ erhalten wir eine Lösung $X = (x_{i,j})_{(k,l)}$ von $AX = B$ durch

$$x_{\sigma(i),j} := b_{\pi(i),j}^{(n+1)} \qquad (i = 1, \ldots, n;\ j = 1, \ldots, l);$$

$$x_{\kappa,j} := 0 \qquad (\kappa \in \{1, \ldots, k\} \setminus \{\sigma(1), \ldots, \sigma(n)\};\ j = 1, \ldots, l).$$

Die Permutationen π und σ wollen wir wie bei der Gauß-Elimination über gewisse $\pi_\nu \in S_n$, $\sigma_\nu \in S_k$ $(\nu = 1, \ldots, n+1)$ rekursiv konstruieren. Da wir die Jordanschen Eliminationsschritte ohne Zeilen- und Spaltenvertauschungen definiert haben, bezeichnen wir – anders als bei der Gauß-Elimination –

$$C^{(\nu)} =: (\gamma_{i,j}^{(\nu)})_{(n,k+l)} \qquad (\nu = 1, \ldots, n+1)$$

und notieren hiermit den

(2.5.5) *Jordan-Algorithmus bei maximaler Pivotwahl.*

Wir setzen $C^{(1)} := (A,B)$, $\pi_1 = \mathrm{id} \in S_n$, $\sigma_1 = \mathrm{id} \in S_k$. Anschließend verfahren wir für $\nu = 1, \ldots, n$ wie folgt:

(i) $\pi_{\nu+1}$ und $\sigma_{\nu+1}$ werden nach (2.2.15), (ii), (iii) und (iv) bestimmt. Ist das Pivotelement Null, müssen wir abbrechen, sonst berechnen wir mit $r := \pi_{\nu+1}(\nu)$, $s := \sigma_{\nu+1}(\nu)$

(ii)
$$\begin{cases} \gamma_{r,j}^{(\nu+1)} = \dfrac{1}{\gamma_{r,s}^{(\nu)}}\,\gamma_{r,j}^{(\nu)} & (j = 1, \ldots, k+l), \\ \gamma_{i,j}^{(\nu+1)} = \gamma_{i,j}^{(\nu)} - \gamma_{i,s}^{(\nu)}\,\gamma_{r,j}^{(\nu+1)} & (i \neq r;\ j = 1, \ldots, k+l). \end{cases}$$

Nach n Schritten (i) und (ii) – falls durchführbar – setzen wir $\pi := \pi_{n+1}$, $\sigma := \sigma_{n+1}$ und erhalten eine Lösung von $AX = B$ nach (2.5.4) ohne weitere Rechnung.

Da sich beim ν-ten Eliminationsschritt frühere Pivotspalten nicht ändern und die s-te Spalte in e_r übergeht, verursacht die Vorschrift (2.5.5, ii) keinen Rechenaufwand für $j = \sigma_{\nu+1}(1), \ldots, \sigma_{\nu+1}(\nu)$, so daß der ν-te Eliminationsschritt $n(k + l - \nu)$ Multiplikationen bzw. Divisionen erfordert. Summation über $\nu = 1, \ldots, n$ liefert im Fall $k = n$, $l = 1$:

(2.5.6) **Bemerkung.** Die Lösung eines Gleichungssystems $Ax = b$ mit invertierbarer (n,n)-Matrix A benötigt bei Verwendung der Jordan-Elimination

$\frac{1}{2}(n^3 + n^2)$ Multiplikationen und Divisionen.

Für einzelne Gleichungssysteme bevorzugt man daher die Gauß-Elimination, die nach (2.2.22) mit etwa $\frac{1}{3}n^3$ Rechenoperationen auskommt. Eine wichtige Anwendung der Jordan-Elimination ist jedoch die Matrix-Invertierung.

Zur Invertierung der (n,n)-Matrix A mit $\mathrm{rg}\,A = n$ ist das beschriebene Eliminationsverfahren auf $C = (A,B)$ mit $B = I$ anzuwenden. Beim ν-ten Eliminationsschritt mit dem Pivotelement (r,s) entsteht in $A^{(\nu+1)}$ die Einheitsspalte e_r,

während die Spalte e_r von $B^{(\nu)}$ in die r-te Spalte der Transformationsmatrix $T^{(\nu)}$ – vgl. (2.5.2, i) – übergeht. Das legt den Gedanken nahe, auf das Speichern der Einheitsspalten zu verzichten und beim ν-ten Schritt die Pivotspalte in die r-te Spalte von $T^{(\nu)}$ (= r-te Spalte von $B^{(\nu+1)}$) zu überführen. Mit den so zu konstruierenden Matrizen

$$F^{(\nu)} = (f_{i,j}^{(\nu)})_{(n,n)} \qquad (\nu = 1, \ldots, n+1)$$

erhalten wir

(2.5.7) *Algorithmus zur Matrix-Invertierung nach Jordan.*
Wir setzen $F^{(1)} = A$, $\pi_1 = \sigma_1 = \mathrm{id} \in S_n$ und führen für $\nu = 1, \ldots, n$ folgende Eliminationsschritte aus:

(i) Wir wenden (2.2.15, ii) auf $F^{(\nu)}$ an und bestimmen $\pi_{\nu+1}$, $\sigma_{\nu+1}$ nach (2.2.15, iii) und (iv).

(ii) Mit den abkürzenden Bezeichnungen

$$r = \pi_{\nu+1}(\nu), \quad s = \sigma_{\nu+1}(\nu), \quad F^{(\nu)} := (f_{i,j})_{(n,n)}, \quad F^{(\nu+1)} = (\tilde{f}_{i,j})_{(n,n)}$$

berechnen wir

$$\begin{cases} \tilde{f}_{r,s} = \dfrac{1}{f_{r,s}}\,, \\ \tilde{f}_{i,s} = -\dfrac{f_{i,s}}{f_{r,s}} & (i = 1, \ldots, n;\ i \neq r), \\ \tilde{f}_{r,j} = \dfrac{f_{r,j}}{f_{r,s}} & (j = 1, \ldots, n;\ j \neq s), \\ \tilde{f}_{i,j} = f_{i,j} - \tilde{f}_{r,j} f_{i,s} \ (= f_{i,j} + f_{r,j} \tilde{f}_{i,s}) & (i \neq r;\ j \neq s). \end{cases}$$

Nach den n Eliminationsschritten ergibt sich $A^{-1} =: (t_{i,j})_{(n,n)}$ aus den Beziehungen

(iii) $t_{\sigma(i),\pi(j)} = f_{\pi(i),\sigma(j)}^{(n+1)} \qquad (i, j = 1, \ldots, n)$,

wobei $\pi := \pi_{n+1}$, $\sigma := \sigma_{n+1}$ gesetzt sei.

Zur *Begründung* von (i) beachten wir, daß die in (2.2.15, ii) abgesuchten Spalten von $F^{(\nu)}$ keine früheren Pivotspalten sind und daher zu $A^{(\nu)}$ gehören. – Da beim ν-ten Eliminationsschritt die Spalte $\pi_{\nu+1}(\nu) = \pi(\nu)$ von $B^{(\nu+1)}$ in die Spalte $\sigma_{\nu+1}(\nu) = \sigma(\nu)$ von $F^{(\nu+1)}$ übertragen wird, besteht $F^{(n+1)}$ aus Spalten von $B^{(n+1)}$ in der Anordnung

$$F^{(n+1)} e_{\sigma(\nu)} = B^{(n+1)} e_{\pi(\nu)} \qquad (\nu = 1, \ldots, n).$$

A^{-1} als die Lösungsmatrix des Systems $AX = I$ ergibt sich nach (2.5.4) durch

$$t_{\sigma(i),\pi(j)} = b_{\pi(i),\pi(j)}^{(n+1)} = f_{\pi(i),\sigma(j)}^{(n+1)} \qquad (i, j = 1, \ldots, n),$$

womit auch (iii) nachgewiesen ist.

Die Algorithmen (2.5.5) und (2.5.7) lassen sich unschwer auf halbmaximale Pivotwahl übertragen.

Die Jordan-Elimination zur Invertierung einer Matrix wird häufig als *Austauschverfahren* beschrieben, wie wir im Folgenden erläutern:

Es seien A, B invertierbare (n,n)-Matrizen; hierzu betrachten wir für $x, y \in \mathbb{C}^n$ das Gleichungssystem

$$Ax + By = 0 .$$

Ist dann T eine invertierbare (n,n)-Matrix, so besitzt das durch (TA, TB) beschriebene Gleichungssystem die gleiche Lösungsgesamtheit wie das ursprüngliche. Wir betrachten speziell den Fall B = I, also das System

$$Ax + y = 0 , \tag{2.5.8}$$

das in einer nach y aufgelösten Form vorliegt. Durch Anwendung eines Jordanschen Eliminationsschritts auf die Matrix (A, I) mit dem Pivotelement $a_{r,s} \neq 0$ geht (2.5.8) in ein äquivalentes System

$$(TA)x + Ty = 0 \tag{*}$$

über. Wir schreiben

$$x = (x_j)_1^n = \sum_{j=1}^{n} x_j e_j , \qquad y = (y_i)_1^n = \sum_{i=1}^{n} y_i e_i$$

und setzen diese Darstellungen in (*) ein; dabei berücksichtigen wir die in (2.5.1) bzw. (2.5.2, i) erwähnten Beziehungen

$$(TA)e_s = e_r , \qquad Te_i = e_i \qquad (i \neq r) ,$$

womit wir (*) in die Form

$$\left(\sum_{\substack{j=1 \\ j \neq s}}^{n} x_j (TA) e_j + y_r (Te_r) \right) + \left(\sum_{\substack{i=1 \\ i \neq r}}^{n} y_i e_i + x_s e_r = 0 \right)$$

bringen. Wenn wir anschließend

$$x^{(2)} := \sum_{\substack{j=1 \\ j \neq s}}^{n} x_j e_j + y_r e_s , \qquad y^{(2)} := \sum_{\substack{i=1 \\ i \neq r}}^{n} y_i e_i + x_s e_r$$

setzen, so wird (*) zu

$$F^{(2)} x^{(2)} + y^{(2)} = 0 , \tag{**}$$

wobei $F^{(2)}$ wie in (2.5.7) definiert ist. In (**) ist nach den Komponenten von $y^{(2)}$, d.h. nach den Variablen y_i $(i \neq r)$ sowie x_s aufgelöst. Ein Jordanscher Elimina-

tionsschritt mit dem Pivotelement $(r, s) = (\pi(1), \sigma(1))$, auf (2.5.8) angewendet, bewirkt also den „Austausch" der Variablen x_s gegen die Variable y_r.

Auf (**) wird ein Jordanschritt mit den Pivotindizes $(\pi(2), \sigma(2))$ angewendet, wodurch die Variable $x_{\sigma(2)}$ gegen $y_{\pi(2)}$ ausgetauscht wird, usf. Nach n Schritten gelangen wir zu einem System

$$(***) \qquad F^{(n+1)} x^{(n+1)} + y^{(n+1)} = 0 ,$$

in dem die Komponenten von $x^{(n+1)}$ die – permutierten – y_i und die Komponenten von $y^{(n+1)}$ die – ebenfalls permutierten – x_i sind. Durch Zeilen- und Spaltenvertauschungen läßt sich (***) schließlich zu einem Gleichungssystem

$$\tilde{A}y + x = 0$$

umschreiben, in dem wegen der Äquivalenz zu (2.5.8) offenbar $\tilde{A} = A^{-1}$ ist.

Den Variablentausch benutzen wir beim Rechnen von Hand für eine knappe und übersichtliche Darstellung der Eliminationsschritte. Hierzu setzen wir vor die Zeilen der $F^{(\nu)}$ die Komponenten von $y^{(\nu)}$, d.h. die Variablen, nach denen aufgelöst ist, sowie an den Kopf der Spalten von $F^{(\nu)}$ die Komponenten von $x^{(\nu)}$. Für $\nu = 1$ steht also $y_1, \ldots, y_n$ vor den Zeilen, $x_1, \ldots, x_n$ über den Spalten von A; ferner tauschen wir bei jedem Eliminationsschritt die Variable vor der Pivotzeile gegen die über der Pivotspalte stehende Variable aus. So behalten wir während der ganzen Rechnung den Überblick über die bereits benutzten Pivotzeilen und -spalten und können schließlich A^{-1} ohne weiteres aus $F^{(n+1)}$ herauslesen.

(2.5.9) **Zahlenbeispiel.** Wir berechnen die Inverse von

$$A = \begin{pmatrix} 2{,}00 & 1{,}01 & 2{,}52 \\ 0{,}400 & 0{,}203 & -1{,}80 \\ 0{,}600 & -1{,}05 & 0{,}800 \end{pmatrix}$$

mit Jordan-Elimination bei maximaler Pivotwahl unter Benutzung von 3-stelliger Gleitkomma-Arithmetik:

	x_1	x_2	x_3
y_1	2,00	1,01	**2,52**
y_2	0,400	0,203	−1,80
y_3	0,600	−1,05	0,800

→

	x_1	x_2	y_1
x_3	0,794	0,401	0,397
y_2	**1,83**	0,925	0,709
y_3	−0,0350	−1,37	−0,317

	y_2	x_2	y_1
x_3	−0,434	0,00	0,0900
x_1	0,546	0,505	0,387
y_3	0,0191	**−1,35**	−0,303

→

	y_2	y_3	y_1
x_3	−0,434	0,00	0,0900
x_1	0,553	0,374	0,274
x_2	−0,0141	−0,741	0,224

Indem wir die Zeilen und Spalten des letzten Schemas so vertauschen, daß die Variablen in ihrer natürlichen Reihenfolge erscheinen, erhalten wir

$$A^{-1} = \begin{pmatrix} 0{,}274 & 0{,}553 & 0{,}374 \\ 0{,}224 & -0{,}0141 & -0{,}741 \\ 0{,}0900 & -0{,}434 & 0{,}00 \end{pmatrix} .$$

Den benötigten Rechenaufwand lesen wir am Algorithmus (2.5.7) unmittelbar ab:

(2.5.10) **Bemerkung.** Die Invertierung einer (n,n)-Matrix nach Jordan erfordert

n^3 Multiplikationen bzw. Divisionen.

Bei Anwendung der Gauß-Elimination zerfällt die Invertierung von A in drei Programmteile: 1. Zerlegung PAQ = LR, 2. Bestimmung von S mit LS = P, 3. Lösung von RY = S, dann ist $A^{-1} = QY$. Nach (2.2.21), (2.1.7) und (2.1.5) treten hierbei etwa n^3 Multiplikationen und Divisionen auf. Da es also im Rechenaufwand keinen wesentlichen Unterschied gibt, liegt der Vorteil des Jordan-Verfahrens gegenüber der Gauß-Elimination in der einfacheren Organisation.

Von besonderer Bedeutung sind die Jordanschen Eliminationsschritte beim Simplex-Verfahren in der Linearen Optimierung (4. Kapitel).

2.6. QR-Zerlegung nach Householder

In diesem Abschnitt kehren wir zu den eigentlichen Zerlegungsmethoden zurück, wobei wir die bisher auftretenden unteren Dreiecksmatrizen durch unitäre Matrizen ersetzen wollen. Im folgenden bezeichne für $x = (x_i)_{i=1}^n$, $y = (y_i)_{i=1}^n \in \mathbb{C}^n$ stets

$$(x,y) := y^*x = \sum_{i=1}^{n} x_i \bar{y}_i$$

und $|x|$ die daraus gebildete euklidische Norm

$$|x| := (x,x)^{\frac{1}{2}} = \left(\sum_{i=1}^{n} |x_i|^2 \right)^{\frac{1}{2}} .$$

Eine komplexe (n,n)-Matrix Q heißt bekanntlich *unitär* (im reellen Fall auch *orthogonal*), wenn $Q^*Q = QQ^* = I$, also $Q^{-1} = Q^*$ gilt oder, was das gleiche bedeutet, wenn die Spalten von Q eine Orthonormalbasis des $\mathbb{C}^n$ bilden. Wir erinnern an den aus der linearen Algebra bekannten

(2.6.1) **Satz.** *Zu einer invertierbaren* (n,n)*-Matrix* A *existiert eine unitäre Matrix* Q *und eine obere* (n,n)*-Dreiecksmatrix* R *mit*

$$A = QR .$$

Zum *Beweis* konstruieren wir aus den Spalten $a_1, \ldots, a_n$ von A orthogonale Spalten $q_1, \ldots, q_n \in \mathbb{C}^n$ mit dem

(2.6.2) *Orthogonalisierungsverfahren nach E. Schmidt.* Wir setzen

$$\text{(i)} \quad \tilde{q}_1 = a_1, \quad q_1 = \frac{1}{|\tilde{q}_1|}\,\tilde{q}_1$$

und definieren für $k = 2, \ldots, n$ rekursiv:

$$\text{(ii)} \begin{cases} \tilde{q}_k = a_k - \sum_{j=1}^{k-1} (a_k, q_j) q_j , \\ q_k = \frac{1}{|\tilde{q}_k|}\,\tilde{q}_k . \end{cases}$$

Wegen der linearen Unabhängigkeit der a_k sind die $\tilde{q}_k$ nicht Null, offenbar wird

$$a_k = |\tilde{q}_k|\, q_k + \sum_{j=1}^{k-1} (a_k, q_j)\, q_j \qquad (k = 1, \ldots, n),$$

also mit entsprechend gewählten $r_{j,k} \in \mathbb{C}$:

$$\text{(2.6.3)} \qquad a_k = \sum_{j=1}^{k} r_{j,k}\, q_j \qquad (k = 1, \ldots, n).$$

Definiert man Q als Matrix aus den Spalten $q_1, \ldots, q_n$, also

$$Q = \sum_{k=1}^{n} q_k\, e_k^t$$

und zusätzlich $R := (r_{j,k})_{(n,n)}$ mit $r_{j,k} = 0$ für $j > k$, so liefert (2.6.3) gerade die behauptete Zerlegung $A = QR$.

Beim Schmidtschen Orthogonalisierungsverfahren wird $\tilde{q}_k$ nach (2.6.2, ii) als Lot von a_k auf $\operatorname{span}(q_1, \ldots, q_{k-1})$, den von $q_1, \ldots, q_{k-1}$ aufgespannten Unterraum (= $\operatorname{span}(a_1, \ldots, a_{k-1})$) konstruiert. Ist der Abstand von a_k zu diesem Unterraum klein gegenüber $|a_k|$, tritt bei der Berechnung von $\tilde{q}_k$ in allen Komponenten starke Auslöschung von Dezimalstellen ein (vgl. (1.3.3, iii), (1.3.6)); wenn der numerische Wert von $\tilde{q}_k$ nicht sogar Null wird, braucht das berechnete q_k nicht annähernd orthogonal zu $q_1, \ldots, q_{k-1}$ zu werden, wie das Beispiel der Übungsaufgabe 2.10 zeigt.

In der QR-Zerlegung nach Householder wird Q als Produkt von elementaren unitären Matrizen konstruiert.

(2.6.4) **Definition.** Es sei $w = (w_i)_{i=1}^n \in \mathbb{C}^n$ mit $|w| = 1$. Dann bezeichnen wir als *Householder-Matrix*

$$H(w) := I - 2ww^* = I - 2 \begin{pmatrix} w_1 \\ \vdots \\ w_n \end{pmatrix} (\overline{w}_1, \ldots, \overline{w}_n) .$$

Hierzu notieren wir die

(2.6.5) **Bemerkungen.**

(i) Die Einheitsmatrix läßt sich nicht in der Form (2.6.4) darstellen, dennoch wollen wir I als Householder-Matrix $H(w)$ mit $w = 0$ bezeichnen, um spätere Fallunterscheidungen zu vermeiden.

(ii) Im reellen Fall bedeutet die Abbildung

$$x \mapsto H(w)x = x - 2(w,x)w$$

gerade die senkrechte Spiegelung von x an der Hyperebene

$$\{y \in \mathbb{R}^n : (y,w) = 0\} .$$

(iii) $H(w) = H(w)^* = H(w)^{-1}$,

d.h. Householder-Matrizen sind hermitesch, involutorisch und folglich unitär.

Zum *Beweis* von (iii) berechnen wir für $H(w) \neq I$

$$H(w)^* = I - 2w^{**}w^* = H(w) ,$$
$$H(w)H(w) = I - 2ww^* - 2ww^* + 4w(w^*w)w^* = I ,$$

letzteres auf Grund von $w^*w = 1$.

(2.6.6) **Hilfssatz.** *Es seien* $a, b \in \mathbb{C}^n$, $a \neq b$, $|a| = |b|$. *Dann existiert ein* $w \in \mathbb{C}^n$ *mit* $|w| = 1$ *und* $H(w)a = b$ *genau dann, wenn* $a^*b = b^*a$ *gilt.*

Beweis. Die Existenz der angegebenen Matrix $H(w)$ ist wegen

$$H(w)a = a - 2ww^*a = b$$

dazu äquivalent, daß es ein $w \in \mathbb{C}^n$ mit $|w| = 1$ und

$$(*) \quad a - b = 2ww^*a = 2(w^*a)w$$

gibt. Da (w^*a) eine Zahl ist, muß sich

$$(**) \quad w = \frac{\tau}{|a-b|}(a-b) \qquad (\tau \in \mathbb{C},\ |\tau| = 1)$$

schreiben lassen. Einsetzen in $(*)$ liefert sodann

$$(***) \quad a - b = |\tau|^2 \frac{2(a-b)^*a}{|a-b|^2}(a-b) ,$$

also

$$2(a-b)^*a = |a-b|^2 .$$

Dazu berechnet man auf Grund der Voraussetzung $a^*a = b^*b$

$$|a-b|^2 = (a-b)^*(a-b) = 2a^*a - b^*a - a^*b\,,$$
$$2(a-b)^*a = 2a^*a - 2b^*a\,.$$

Die Gleichheit beider Ausdrücke und damit Bedingung (***) ist genau dann gegeben, wenn $a^*b = b^*a$ gilt. Aus der Existenz eines $w \in \mathbb{C}^n$ mit $|w| = 1$ und $H(w)a = b$ folgt also $a^*b = b^*a$. Ist umgekehrt die letztere Bedingung erfüllt, wählt man w nach (**), z.B. mit $\tau = 1$ und erhält wegen (***) die Beziehung (*). Die Wahl von τ in (**) hat auf $H(w) = I - 2ww^*$ keinen Einfluß.

Als unmittelbare Folgerungen notieren wir

(2.6.7) **Bemerkungen**.

(i) Falls H(w) mit $H(w)a = b$ existiert, so ist es eindeutig bestimmt, w kann man (im Fall $a \neq b$) wählen als

$$w = \frac{1}{|a-b|}(a-b)\,.$$

(ii) Für $a, b \in \mathbb{R}^n$ ist $a^*b = b^*a$ stets erfüllt; im Fall $a \neq b$, $|a| = |b|$ existiert ein $w \in \mathbb{R}^n$ mit $H(w)a = b$.

(iii) Haben $a, b \in \mathbb{C}^n$ mit $a \neq b$, $|a| = |b|$, $a^*b = b^*a$ die Gestalt

$$a = \begin{pmatrix} 0 \\ \vdots \\ 0 \\ \alpha_{\nu+1} \\ \vdots \\ \alpha_n \end{pmatrix}, \qquad b = \begin{pmatrix} 0 \\ \vdots \\ 0 \\ \beta_{\nu+1} \\ \vdots \\ \beta_n \end{pmatrix}$$

mit $1 \leqq \nu < n$, so wird

$$w = \begin{pmatrix} 0 \\ \vdots \\ 0 \\ w_{\nu+1} \\ \vdots \\ w_n \end{pmatrix} = \begin{pmatrix} 0 \\ \tilde{w} \end{pmatrix}, \qquad H(w) = \left(\begin{array}{c|c} I_\nu & 0 \\ \hline 0 & I_{n-\nu} - 2\tilde{w}\tilde{w}^* \end{array}\right).$$

Nach diesen Vorüberlegungen beweisen wir die gegenüber (2.6.1) verallgemeinerte Zerlegungsaussage:

(2.6.8) **Satz** *(Householder). Zu einer beliebigen* (n,k)*-Matrix* A *existieren eine unitäre* (n,n)*-Matrix* Q *und eine obere* (n,k)*-Dreiecksmatrix* R *mit*

$$A = QR\,.$$

Mit dem *Beweis* geben wir gleichzeitig die Konstruktion von Q und R an.

1. Schritt: Spaltenweise notiert, sei

$$A = \left(a_1^{(1)}, \ldots, a_k^{(1)} \right) = \begin{pmatrix} a_{1,1} & \cdots & a_{1,k} \\ \vdots & & \vdots \\ a_{n,1} & \cdots & a_{n,k} \end{pmatrix} .$$

Im I. Fall, $a_{i,1} = 0$ für alle $i \in \{2, \ldots, n\}$, wählen wir $H_1 = I$; dann besitzt $H_1 A = A$ bereits die Gestalt

$$(2.6.9) \qquad H_1 A = \left(\begin{array}{c|ccc} r_{1,1} & r_{1,2} & \cdots & r_{1,k} \\ 0 & & & \\ \vdots & & a_2^{(2)} \ \cdots \ a_k^{(2)} & \\ 0 & & & \end{array} \right) ,$$

die wir mit dem 1. Konstruktionsschritt erreichen wollen.

Im II. Fall, daß ein $i \in \{2, \ldots, n\}$ mit $a_{i,1} \neq 0$ existiert, setzen wir zur Anwendung von Hilfssatz (2.6.6)

$$a := a_1^{(1)}, \quad b := \sigma \, |a_1^{(1)}| \, e_1 \quad \text{mit } \sigma \in \mathbb{C}, \ |\sigma| = 1 ,$$

womit $a \neq b$, $|a| = |b|$ erfüllt ist. Wegen

$$a^* b = \sigma \, |a_1^{(1)}| \, \bar{a}_{1,1}, \qquad b^* a = \bar{\sigma} \, |a_1^{(1)}| \, a_{1,1}$$

muß σ so gewählt werden, daß

$$\sigma \bar{a}_{1,1} = \bar{\sigma} a_{1,1}, \quad \text{d.h. } \bar{\sigma} a_{1,1} = \pm |a_{1,1}| \in \mathbb{R}$$

gilt. Wir entscheiden uns für σ mit

$$(2.6.10) \qquad \bar{\sigma} a_{1,1} = - |a_{1,1}| .$$

Nach Hilfssatz (2.6.6) haben wir mit

$$(2.6.11) \qquad \begin{cases} w_1 := \dfrac{1}{|a_1^{(1)} - \sigma | a_1^{(1)}| \, e_1|} (a_1^{(1)} - \sigma \, |a_1^{(1)}| \, e_1) , \\ H_1 := H(w_1) \end{cases}$$

unmittelbar

$$H_1 a_1^{(1)} = \sigma \, |a_1^{(1)}| \, e_1$$

und damit spaltenweise

$$H_1 A = (H_1 a_1^{(1)}, H_1 a_2^{(1)}, \ldots, H_1 a_k^{(1)}) ,$$

also in der gewünschten Form (2.6.9), wobei – wie auch im I. Fall –

$$r_{1,1} = \sigma \, |a_1^{(1)}| \ \text{ mit } |\sigma| = 1$$

gilt. Zur Berechnung des Nenners in (2.6.11) notieren wir

$$|a_1^{(1)} - \sigma\,|a_1^{(1)}|\,e_1|^2 = |a_{1,1} - \sigma\,|a_1^{(1)}|\,|^2 + \sum_{i=2}^{n} |a_{i,1}|^2\,,$$

und wegen (2.6.10)

$$\begin{aligned} |a_{1,1} - \sigma\,|a_1^{(1)}|\,|^2 &= |\bar{\sigma}\,a_{1,1} - |a_1^{(1)}|\,|^2 = |-|a_{1,1}| - |a_1^{(1)}|\,|^2 \\ &= |a_1^{(1)}|^2 + 2\,|a_1^{(1)}|\,|a_{1,1}| + |a_{1,1}|^2\,, \end{aligned}$$

so daß wir insgesamt

$$|a_1^{(1)} - \sigma\,|a_1^{(1)}|\,e_1|^2 = 2\,|a_1^{(1)}|\,(|a_{1,1}| + |a_1^{(1)}|) \tag{2.6.12}$$

erhalten. Durch die Wahl von σ in (2.6.10) werden bei der Berechnung des letztgenannten Wertes nur positive Zahlen summiert, was aus Gründen der numerischen Stabilität wünschenswert ist (vgl. (1.3.7)!).

2. Schritt: Zu der $(n-1, k-1)$-Teilmatrix in (2.6.9)

$$A_2 := (a_2^{(2)}, \ldots, a_k^{(2)})$$

wählen wir $H_2' := I_{n-1} - 2\tilde{w}_2\tilde{w}_2^*$ $(\tilde{w}_2 \in \mathbb{C}^{n-1})$ analog dem 1. Konstruktionsschritt und setzen dann

$$w_2 = \begin{pmatrix} 0 \\ \tilde{w}_2 \end{pmatrix} \in \mathbb{C}^n, \qquad H_2 = \left(\begin{array}{c|c} 1 & 0 \\ \hline 0 & H_2' \end{array}\right) = I_n - 2w_2w_2^*\,.$$

Es wird

$$H_2(H_1A) = \left(\begin{array}{c|ccc} r_{1,1} & r_{1,2} & \cdots & r_{1,k} \\ \hline 0 & & & \\ \vdots & & H_2'A_2 & \\ 0 & & & \end{array}\right) = \left(\begin{array}{cc|ccc} r_{1,1} & r_{1,2} & \cdots & \cdots & r_{1,k} \\ 0 & r_{2,2} & \cdots & \cdots & r_{2,k} \\ \hline \vdots & 0 & & & \\ \vdots & \vdots & & A_3 & \\ 0 & 0 & & & \end{array}\right)$$

und dabei $r_{2,2} = \sigma_2\,|a_2^{(2)}|$ mit $\sigma_2 \in \mathbb{C}$, $|\sigma_2| = 1$.

In derselben Weise fahren wir fort; nach $(m-1)$ Schritten, wobei $m = \min(n, k-1)$ bezeichne, haben wir mit einer oberen (n,k)-Dreiecksmatrix R

$$H_{m-1}H_{m-2}\ldots H_1A = R\,.$$

Hieraus folgt wegen $H_\nu^{-1} = H_\nu$ $(\nu = 1, \ldots, m-1)$ die Darstellung $A = (H_1\ldots H_{m-1})R$, wobei $Q := H_1\ldots H_{m-1}$ als Produkt von unitären Matrizen wieder unitär ist.

Aus dem vorstehenden Beweis lesen wir die folgenden Ergänzungen ab:

(2.6.13) **Bemerkung.** Für eine reelle Matrix A werden nach (2.6.7, ii) die H_ν und damit Q reell. Bei jedem Konstruktionsschritt erhält das in (2.6.10) definierte σ den Wert $+1$ oder -1.

(2.6.14) **Bemerkung.** Durch Anwendung der $(m-1)$ Householderschritte auf eine erweiterte Matrix (A, B), in der B eine (n, l)-Matrix ist, erhalten wir als Resultat eine $(n, k+l)$-Matrix

$$(R, S) = Q^*(A, B) = (Q^*A, Q^*B) .$$

Durch die spezielle Wahl $B = I$ gewinnen wir hiermit eine explizite Darstellung von Q^* und damit Q.

Ziel der anschließenden Überlegungen ist es, das im Beweis von Satz (2.6.8) entwickelte Konstruktionsverfahren mit möglichst geringem Rechenaufwand zu realisieren; dazu brauchen wir nur den 1. Schritt, also den Übergang von A auf $H_1 A$ zu beschreiben.

(2.6.15) *Algorithmus der Householder-Transformation.*
Ist $a_{i,1} = 0$ für alle $i = 2, \ldots, n$, wird $H_1 A = A$; sonst wählen wir nach (2.6.10)

$$\text{(i)} \quad \sigma = \begin{cases} 1, \text{ falls } a_{1,1} = 0, \\ -\dfrac{a_{1,1}}{|a_{1,1}|} \quad \text{sonst.} \end{cases}$$

In Anlehnung an (2.6.11) bestimmen wir $\mu := |a_1^{(1)}|$ über

$$\text{(ii)} \quad \mu^2 := \sum_{i=1}^{n} |a_{i,1}|^2, \qquad \mu := \sqrt[+]{\mu^2}$$

und hiermit nach (2.6.12)

$$\text{(iii)} \quad N := \mu^2 + \mu |a_{1,1}|$$

sowie

$$\text{(iv)} \begin{cases} u := a_1^{(1)} - \sigma \mu e_1 , \\ v := \dfrac{1}{N} u . \end{cases}$$

Mit dieser Bezeichnung ist offenbar $N = \frac{1}{2} |a_1^{(1)} - \sigma |a_1^{(1)}| e_1|^2$ und $H_1 = I - vu^*$. Man verzichtet darauf, H_1 explizit zu bestimmen, sondern berechnet

$$\text{(v)} \begin{cases} p^* := u^*A & (p \in \mathbb{C}^k) , \\ H_1 A := A - vp^* , \end{cases}$$

wobei die erste Spalte unmittelbar gleich $\sigma \mu e_1$ gesetzt wird. Zur Bestimmung von $H_1 B$ für eine (n, l)-Matrix B setzt man analog

$$\text{(vi)} \begin{cases} q^* := u^*B & (u \in \mathbb{C}^l) , \\ H_1 B := B - vq^* . \end{cases}$$

Die Vorschrift (v) – ebenso (vi) – wird mit

$$A - vp^* = A - vu^*A = (I - vu^*) A = H_1 A$$

begründet; sie vermeidet die explizite Matrizenmultiplikation von H_1 mit A, die aus n^2k Einzelmultiplikationen bestehen würde. Den Rechenaufwand des Householder-Verfahrens wollen wir für den reellen Fall ermitteln, hier wird $\sigma = \pm 1$. Beim 1. Konstruktionsschritt benötigt man außer Additionen zur Berechnung von

μ^2 – n Multiplikationen,
μ – 1 Wurzel,
N – 1 Multiplikation,
u – (nur 1 Addition),
v – n Divisionen,

und weiter nach (vi) bei beliebiger (n, l)-Matrix B für

q^* – $n \cdot l$ Multiplikationen,
$H_1 B$ – $n \cdot l$ Multiplikationen.

Da die erste Spalte von $H_1 A$ bereits vorliegt, sind in (v) für

p^* – $n(k-1)$ Multiplikationen,
$H_1 A$ – $n(k-1)$ Multiplikationen

auszuführen. Beim $(\nu+1)$-ten Konstruktionsschritt $(0 \leqq \nu \leqq m-2)$ haben wir n durch $n-\nu$, k durch $k-\nu$ zu ersetzen und erhalten somit $[2(k-\nu)(n-\nu)+1]$ Multiplikationen bzw. Divisionen und 1 Wurzel beim Übergang von $H_\nu \dots H_1 A$ auf $H_{\nu+1}(H_\nu \dots H_1 A)$ und weitere $2l(n-\nu)$ Multiplikationen für die entsprechende Transformation mit B. Speziell für $k = n$ wird $m = n-1$; durch Aufsummieren erhalten wir

(2.6.16) **Bemerkung.**

(i) Die QR-Zerlegung einer (n,n)-Matrix A nach Householder erfordert

$(n-1)$ Wurzelberechnungen und
$[\frac{2}{3}n^3 + n^2 + \frac{4}{3}n - 3] \approx \frac{2}{3}n^3$ Multiplikationen bzw. Divisionen.

(ii) Die Zerlegung einer (n, l)-Matrix B in $B = QS$ mit dem gleichen Q benötigt zusätzlich zu (i)

$l(n^2 + n - 2)$ Multiplikationen.

Wir diskutieren zwei Anwendungen der QR-Zerlegung, zunächst die

Reduktion eines Gleichungssystems auf obere Dreiecksgestalt.

Für ein Gleichungssystem $AX = B$ mit der (n,k)-Matrix A und der (n, l)-Matrix B erhalten wir nach (2.6.14) eine simultane Zerlegung $QR = A$, $QS = B$ und damit die Äquivalenz

$$AX = B \iff RX = S.$$

Auch wenn S mit $QS = B$ erst *nach* der Zerlegung von A bestimmt werden soll – vgl. (2.1.12) –, werden wir Q nicht explizit berechnen, sondern die beim ν-ten Schritt gemäß (2.6.15, iv) auftretenden $u^{(\nu)}$ und $v^{(\nu)}$ (bzw. $N^{(\nu)}$) $(\nu = 1, \dots, m-1)$

speichern und hiermit $S := H_{m-1} \dots H_1 B$ durch wiederholte Anwendung von (2.6.15, vi) ermitteln. Aus (2.6.16) und (2.1.5) folgern wir

(2.6.17) **Bemerkung.** Die Lösung des Gleichungssystems $Ax = b$ mit der invertierbaren (n,n)-Matrix A und dem (n, 1)-Vektor b unter Benutzung der Householder-Zerlegung erfordert $(n-1)$ Wurzelberechnungen und etwa

$(\frac{2}{3} n^3 + \frac{5}{2} n^2)$ Multiplikationen bzw. Divisionen.

Wie ein Vergleich mit (2.2.22) zeigt, benötigt die Householder-Reduktion etwa doppelt so viele Rechenoperationen wie die Gauß-Elimination, allerdings entfallen Pivotsuche und Zeilen- bzw. Spaltenvertauschungen. Bezüglich der Rundungsfehler erweisen sich beide Verfahren als etwa gleichwertig, was im nächsten Kapitel ausführlich erläutert wird. Da man bei der Gauß-Elimination meistens mit halbmaximaler Pivotwahl auskommt, gibt man diesem Verfahren wegen seines geringeren Rechenaufwandes den Vorzug.

(2.6.18) **Zahlenbeispiel.** Wir lösen das Gleichungssystem (2.3.3) mit Householder-Elimination unter Benutzung der im Beispiel (1.1.16) erwähnten REAL*4-Arithmetik; dazu lassen wir auch Q mit $Q(R,s) = (A,b)$ explizit berechnen. Wir erhalten folgende Werte:

$$Q = \begin{pmatrix} -0{,}1892328 & -0{,}3673748 & 0{,}07173681 & -0{,}9077885 \\ -0{,}5504981 & -0{,}4072523 & 0{,}7287234 & 0{,}007529378 \\ -0{,}8086442 & 0{,}1030207 & -0{,}5549393 & 0{,}1663833 \\ -0{,}08601534 & 0{,}8297991 & 0{,}3947841 & 0{,}3849415 \end{pmatrix},$$

$$(R,s) = \left(\begin{array}{cccc|c} -5{,}812916 & 0{,}8727131 & -7{,}245927 & -7{,}151311 & -7{,}254530 \\ 0 & 8{,}887770 & 1{,}412116 & 3{,}552075 & 5{,}690638 \\ 0 & 0 & -1{,}253943 & 7{,}961990 & 2{,}338003 \\ 0 & 0 & 0 & 1{,}023791 & 2{,}678076 \end{array}\right).$$

Die Auflösung des Gleichungssystems $Rx = s$ liefert:

$x_1 = -20{,}76250$, $x_2 = -2{,}747878$, $x_3 = 14{,}74490$, $x_4 = 2{,}615841$.

Wir erhalten mit Gauß-Elimination, maximaler Pivotwahl unter Benutzung der gleichen Arithmetik:

$x_1 = -20{,}76247$, $x_2 = -2{,}747877$, $x_3 = 14{,}74485$, $x_4 = 2{,}615837$.

Um die Genauigkeit vergleichen zu können, lösen wir das Gleichungssystem $Ax = b$ in doppeltgenauer Arithmetik, wobei wir nach wie vor die auf 6-stellige Hexadezimalzahlen (REAL*4) gerundeten Koeffizienten von (A,b) verwenden. Die Lösung, auf 7 Dezimalen gerundet, lautet

$$x_1 = -20{,}76267,\quad x_2 = -2{,}747911,\quad x_3 = 14{,}74500,\quad x_4 = 2{,}615857.$$

An Hand dieses Ergebnisses haben wir in den oben berechneten Werten die signifikanten Stellen unterstrichen: wir sehen, daß die Lösungen, nach Gauß und Householder berechnet, numerisch etwa gleichwertig sind.

Als 2. Anwendung der QR-Zerlegung notieren wir die

Orthogonalisierung von k *Vektoren des* $\mathbb{C}^n$.

Um k Vektoren $a_1, \ldots, a_k$ des $\mathbb{C}^n$ mit $k \leqq n$ zu orthogonalisieren oder ihre lineare Abhängigkeit festzustellen, bilden wir die (n,k)-Matrix A aus den Spalten a_i, also

$$A := (a_1, \ldots, a_k) = \sum_{i=1}^{k} a_i e_i^t$$

und zerlegen A = QR nach Householder. Es sei $R =: (r_{i,j})_{(n,k)}$; es bezeichne $q_1, \ldots, q_n$ die Spalten von Q. Da Q unitär ist, bilden die $q_1, \ldots, q_n$ ein Orthonormalsystem im $\mathbb{C}^n$. Weiter erhalten wir den

(2.6.19) **Satz.** *Für* $1 \leqq m \leqq k$ $(\leqq n)$ *sind folgende Aussagen äquivalent:*

(i) $r_{i,i} \neq 0 \qquad (i = 1, \ldots, m)$,

(ii) $a_1, \ldots, a_m$ *linear unabhängig,*

(iii) $\operatorname{span}(a_1, \ldots, a_m) = \operatorname{span}(q_1, \ldots, q_m)$.

Beweis. Aus A = QR schließen wir für die Spalten von A:

$$(2.6.20) \qquad a_j = \sum_{i=1}^{j} r_{i,j} q_i \qquad (j = 1, \ldots, m)$$

und daher stets

$$\operatorname{span}(a_1, \ldots, a_m) \subseteq \operatorname{span}(q_1, \ldots, q_m).$$

Da der rechts stehende Unterraum die Dimension m besitzt, gilt Gleichheit genau dann, wenn die $a_1, \ldots, a_m$ linear unabhängig sind; hiermit ist die Äquivalenz von (ii) und (iii) gezeigt. – Wir fassen (2.6.20) zu einer Matrixgleichung zusammen, nämlich

$$(a_1, \ldots, a_m)_{(n,m)} = (q_1, \ldots, q_m)_{(n,m)} \begin{pmatrix} r_{1,1} & \cdots & r_{1,m} \\ & \ddots & \vdots \\ 0 & & r_{m,m} \end{pmatrix}.$$

Da der Rang von $(q_1, \ldots, q_m)_{(n,m)}$ gleich m ist, sind die Spalten $a_1, \ldots, a_m$ genau dann linear unabhängig, wenn die Dreiecksmatrix $(r_{i,j})_{(m,m)}$ invertierbar ist; hieraus folgt die Äquivalenz von (i) und (ii).

Wir schließen unmittelbar die

(2.6.22) **Folgerung.** Im Fall $r_{i,i} \neq 0$ $(1 \leqq i \leqq k)$ sind die $a_1, \ldots, a_k$ linear unabhängig, und $q_1, \ldots, q_k$ sind orthonormierte Vektoren mit der Eigenschaft

$$\operatorname{span}(a_1, \ldots, a_m) = \operatorname{span}(q_1, \ldots, q_m)$$

für alle m mit $1 \leqq m \leqq k$.

Die QR-Zerlegung findet weitere Anwendungen beim linearen Ausgleichsproblem, das wir im Band 3 behandeln wollen, und in einigen Verfahren der Eigenwertberechnung.

Außer dem Schmidtschen Orthogonalisierungsverfahren und der Householder-Zerlegung sei noch die QR-Zerlegung nach Givens erwähnt, in der Q als Produkt von ebenen Drehungen konstruiert wird; näheres dazu bringt die Übungsaufgabe 2.11. Diese Methode ist nur auf reelle Matrizen anwendbar und erfordert gegenüber der Householder-Zerlegung etwa den doppelten Rechenaufwand.

Abschließend beweisen wir eine Eindeutigkeitsaussage zur QR-Zerlegung:

(2.6.23) **Satz.** *Ist* A *eine invertierbare* (n,n)*-Matrix und*

$$A = QR = \tilde{Q}\tilde{R}$$

mit unitären (n,n)*-Matrizen* $Q, \tilde{Q}$ *und oberen Dreiecksmatrizen* R *und* $\tilde{R}$, *so gilt mit geeigneten* $\sigma_i \in \mathbb{C}$, $|\sigma_i| = 1$

$$\tilde{Q} = Q \cdot \operatorname{diag}(\sigma_1, \ldots, \sigma_n),$$

d. h. die Spalten von Q *sind bis auf Vielfache vom Betrag* 1 *eindeutig bestimmt.*

Beweis. Aus der Invertierbarkeit von A folgt die Invertierbarkeit von R und $\tilde{R}$ und weiter

$$Q^*\tilde{Q} = R\tilde{R}^{-1}.$$

Die Matrix $B := R\tilde{R}^{-1}$ ist damit unitär und obere Dreiecksmatrix; es folgt, daß $B^* = B^{-1}$ gleichzeitig untere und obere Dreiecksmatrix, also Diagonalmatrix ist. Setzt man

$$B = \operatorname{diag}(\sigma_1, \ldots, \sigma_n),$$

so folgt aus $BB^* = \operatorname{diag}(\sigma_1\bar{\sigma}_1, \ldots, \sigma_n\bar{\sigma}_n) = I$ unmittelbar, daß die $|\sigma_i| = 1$ sind, und es gilt $\tilde{Q} = QB$ nach Definition von B.

Übungsaufgaben zum 2. Kapitel

Aufgabe 2.1. Man löse mit 3-stelliger Gleitkomma-Arithmetik (vgl. (1.3.1)–(1.3.3)!) das Gleichungssystem

$$\begin{pmatrix} 2 & 1{,}01 & 2{,}52 \\ 0{,}4 & 0{,}203 & -1{,}8 \\ 0{,}6 & -1{,}05 & 0{,}8 \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \\ x_3 \end{pmatrix} = \begin{pmatrix} 9{,}57 \\ -0{,}385 \\ -3{,}85 \end{pmatrix}$$

unter Benutzung der Gauß-Elimination
(i) bei diagonaler, (ii) bei halbmaximaler Pivotwahl.

Ergebnisse: (i) $x_1 = 3{,}53, \quad x_2 = 0{,}00, \quad x_3 = 1{,}00;$
(ii) $x_1 = 1{,}00, \quad x_2 = 5{,}01, \quad x_3 = 1{,}00;$

die exakte Lösung ist: $x_1 = 1{,}00, \quad x_2 = 5{,}00, \quad x_3 = 1{,}00.$

Aufgabe 2.2

(i) Das Gleichungssystem

$$\begin{pmatrix} 0{,}2 \cdot 10^{30} & 1{,}0002 & 1{,}401 \cdot 10^{-20} \\ 0{,}6 \cdot 10^{19} & 3{,}1006 \cdot 10^{-11} & 4{,}422 \cdot 10^{-31} \\ 0{,}4 \cdot 10^{-47} & 2{,}0005 \cdot 10^{-77} & -7{,}004 \cdot 10^{-97} \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \\ x_3 \end{pmatrix} = \begin{pmatrix} 9{,}57 \\ -0{,}41 \cdot 10^{-9} \\ -0{,}39 \cdot 10^{-76} \end{pmatrix}$$

soll mit Gauß-Elimination, halbmaximaler Pivotwahl gelöst werden. Man zeige, daß bei Verwendung der im Beispiel (1.1.16) angegebenen Maschinenarithmetik die Koeffizientenmatrix durch Exponentenüberlauf in eine nichtinvertierbare Matrix übergeht.

(ii) Man forme das Gleichungssystem so um, daß alle Koeffizienten der Matrix dem Betrag nach zwischen 1 und 10 liegen (Äquilibrierung), und löse anschließend das Gleichungssystem unter Benutzung eines Tischrechners.

Lösung: $x_1 = 0{,}35433 \cdot 10^{-26}, \; x_2 = -0{,}70223 \cdot 10^{3}, \; x_3 = 0{,}23424 \cdot 10^{20}.$

Aufgabe 2.3. Man zeige an Hand von (2.2.13), daß die Matrix L der Gauß-Zerlegung die in (2.2.20) angegebenen Koeffizienten besitzt.

Aufgabe 2.4. Es sei $A = (a_{i,j})_{(n,n)}$ invertierbare (n,n)-Matrix in oberer Hessenbergform, d.h. $a_{i,j} = 0$ für alle $i > j + 1$. Mit Gauß-Elimination bei halbmaximaler Pivotwahl werde $PA = LR$ zerlegt. Man zeige:

(i) L besitzt außer dem Diagonalelement in jeder Spalte höchstens einen von Null verschiedenen Koeffizienten,

(ii) Die Zerlegung benötigt $\frac{1}{2}(n^2 + n - 2)$ Multiplikationen bzw. Divisionen,

(iii) sind die $|a_{i,j}| \leqq 1$, so gelten für die Koeffizienten $c_{i,j}^{(\nu)}$ der $C^{(\nu)}$ aus Satz (2.2.6) zu $l = 0$ die Abschätzungen

$$|c_{i,j}^{(\nu)}| \leqslant \nu \qquad (\nu = 1, \ldots, n).$$

Hinweis: Man zeigt induktiv, daß die Matrizen $C^{(\nu)}$ $(\nu = 1, \ldots, n)$ obere Hessenbergform haben und daß gilt $|c_{i,j}^{(\nu)}| \leq \nu$ $(i \leq \nu)$, $|c_{i,j}^{(\nu)}| \leq 1$ $(i > \nu)$.

Aufgabe 2.5. Es sei $A = (a_{i,j})_{(n,n)}$ invertierbare Tridiagonalmatrix, d.h. $a_{i,j} = 0$ für alle i,j mit $|i - j| > 1$. Man zeige für die Gauß-Zerlegung bei halbmaximaler Pivotwahl $PA = LR$:

(i) in $R = (r_{i,j})_{(n,n)}$ gilt $r_{i,j} = 0$ für $j > i + 2$ $(i = 1, \ldots, n-2)$,

(ii) die Zerlegung benötigt höchstens $3(n-1)$ Operationen,

(iii) $\max \{|c_{i,j}^{(\nu)}| : i, j, \nu = 1, \ldots, n\} \leqq 2 \max_{i,j=1}^{n} |a_{i,j}|$.

Hinweis: Man nimmt o.E. $|a_{i,j}| \leq 1$ an und zeigt mit Induktion über ν: in den ersten $(\nu-1)$ Zeilen besitzt $C^{(\nu)}$ die in (i) behauptete Eigenschaft, in den Zeilen $\nu, \ldots, n$ hat $C^{(\nu)}$ Tridiagonalgestalt (mit $c^{(\nu)}_{\nu,\nu-1} = 0$), und es gilt $|c^{(\nu)}_{i,i}| \leq 2$ $(i = 1, \ldots, \nu)$, $|c^{(\nu)}_{i,j}| \leq 1$ sonst.

Aufgabe 2.6. Man löse – unter Benutzung eines Tischrechners – das im Zahlenbeispiel (2.3.3) angegebene Gleichungssystem mit dem kompakten Gauß-Algorithmus bei halbmaximaler Pivotwahl. Dabei sind die auftretenden Produktsummen mindestens 6-stellig zu berechnen.

Ergebnis: $x_1 = 20{,}7$; $x_2 = -2{,}74$; $x_3 = 14{,}7$; $x_4 = 2{,}61$.

Aufgabe 2.7. Es sei A positiv definite (n,n)-Matrix.

(i) Man zeige für die Koeffizienten von R in der Cholesky-Zerlegung $A = R^*R$ die Abschätzungen $|r_{\kappa,i}| \leq \sqrt{a_{i,i}}$ $(1 \leq i \leq n,\ i \leq \kappa \leq n)$.

(ii) Es sei $A = LU$, wobei $L = (l_{i,j})_{(n,n)}$ normierte untere Dreiecksmatrix, $U = (u_{i,j})_{(n,n)}$ obere Dreiecksmatrix bezeichne. Man zeige

$$|u_{i,j}| \leq \sqrt{a_{i,i}a_{j,j}} \qquad (1 \leq i \leq n,\ i \leq j \leq n),$$

$$|l_{i,j}| \leq \sqrt{\frac{a_{i,i}}{\lambda_1}} \qquad \text{mit } \lambda_1 = \min\{\lambda \in \mathbb{R}:\ \lambda \text{ Eigenwert von } A\}.$$

Hinweis zu (ii): Man stellt L und U mit Hilfe der Koeffizienten von R dar, zeigt $(Ax, x) \geq \lambda_1 (x, x)$ für alle $x \in \mathbb{C}^n$ und weiter $r^2_{\nu,\nu} = (Ax, x)$ mit $x = L^{*-1} e_\nu$ $(\nu = 1, \ldots, n)$.

Aufgabe 2.8. Für

$$A = \begin{pmatrix} 1{,}01 & 1{,}97 \\ 0{,}990 & 2{,}03 \end{pmatrix}$$

berechne man A*A und die LR-Zerlegung von A, beides mit 3-stelliger Gleitkommarechnung; – man überzeugt sich, daß das numerisch gewonnene R invertierbar ist, während $gl(A^*A)$ nicht invertierbar ist.

Aufgabe 2.9. Man beweise den Satz (2.5.3) über die Jordan-Elimination.

Aufgabe 2.10. Man berechne mit 3-stelliger Gleitkomma-Arithmetik die QR-Zerlegung von

$$A = \begin{pmatrix} 0{,}632 & 0{,}562 \\ 0{,}313 & 0{,}277 \end{pmatrix}$$

(i) mit dem Orthogonalisierungsverfahren von E. Schmidt,
(ii) mit dem Householder-Verfahren.

Aufgabe 2.11. Es sei A reelle (n,k)-Matrix. Beim 1. Schritt der QR-Zerlegung nach Givens wird eine orthogonale Matrix S_1 mit

$$S_1 A = \begin{pmatrix} r_{1,1} & r_{1,2} & \cdots & r_{1,k} \\ 0 & & A^{(2)} & \end{pmatrix}$$

als Produkt von ebenen Drehmatrizen bestimmt, nämlich

$$S_1 = S_{1,n} S_{1,n-1} \cdots S_{1,2}$$
$$S_{1,j} := S_{1,j}(\alpha_j) = I - (1 - \cos\alpha_j)(e_1 e_1^t + e_j e_j^t) + \sin\alpha_j (e_1 e_j^t - e_j e_1^t),$$

wobei die $\alpha_j \in \mathbb{R}$ geeignet zu wählen sind. Man zeige für $j = 2, \ldots, n$:

(i) $S_{1,j}(\alpha_j)$ $(\alpha_j \in \mathbb{R})$ ist orthogonale Matrix,

(ii) $S_{1,j}(\alpha_j) A =: (\tilde{a}_{\mu,\nu})_{(n,k)}$ unterscheidet sich von A höchstens in der 1-ten und j-ten Zeile.

(iii) Man bestimme $\gamma_j = \cos\alpha_j$, $\sigma_j = \sin\alpha_j$ (α_j selbst wird nicht berechnet), so daß $\tilde{a}_{j,1} = 0$ wird, und gebe die $\tilde{a}_{1,i}$ sowie $\tilde{a}_{j,i}$ an.

(iv) Man zeige, daß der Übergang von A auf $S_1 A$ $4k(n-1)$ Multiplikationen bzw. Divisionen und $(n-1)$ Wurzelberechnungen erfordert.

Hat man nach ν Schritten $(1 \leqq \nu \leqq \min(n-2,\ k-1))$ die Gestalt

$$S_\nu \ldots S_1 A = \begin{pmatrix} r_{1,1} & \cdots & \cdots & r_{1,k} \\ & \ddots & & \vdots \\ 0 & & r_{\nu,\nu} & \cdots \quad r_{\nu,k} \\ & & 0 & A^{(\nu+1)} \end{pmatrix}$$

erreicht, so transformiert man $A^{(\nu+1)}$ mit einer orthogonalen $(n-\nu, n-\nu)$-Matrix $\tilde{S}_{\nu+1}$, wie oben angegeben, und setzt hiermit

$$S_{\nu+1} := \begin{pmatrix} I_\nu & 0 \\ 0 & \tilde{S}_{\nu+1} \end{pmatrix}_{(n,n)}$$

Mit $m = \min(n-1, k)$ haben wir schließlich

$$R := S_{m-1} \ldots S_1 A = Q^* A$$

1 oberer Dreiecksgestalt.

r) Man zeige im Fall $k = n$, daß die QR-Zerlegung nach Givens $\frac{1}{2} n(n-1)$ Wurzeln und $\frac{4}{3}(n-1)\,n(n+1)$ Multiplikationen bzw. Divisionen erfordert.

Aufgabe 2.12. Es sei $A = (\alpha_{i,j})_{(n,n)}$ komplexe (n,n)-Matrix; $a_1, \ldots, a_n$ $(\in \mathbb{C}^n)$ seien die Spalten von A. – Mit Hilfe der QR-Zerlegung zeige man die Determinantenabschätzung nach Hadamard

$$|\det A| \leqq \prod_{j=1}^{n} (a_j, a_j)^{\frac{1}{2}} = \prod_{j=1}^{n} \left(\sum_{i=1}^{n} |\alpha_{i,j}|^2 \right)^{\frac{1}{2}}.$$

3. Fehlerbetrachtungen bei linearen Problemen

Der Herkunft nach, unterscheiden wir drei Arten von Fehlern:

(i) *Fehler in den Eingabedaten,* verursacht durch Meßungenauigkeiten oder durch Runden der Rechenmaschine beim Einlesen,

(ii) *Rundungsfehler,* verursacht durch das Runden der Maschine bei fast jeder Rechenoperation (vgl. (1.3.1)!),

(iii) *Abbruchfehler* oder *Verfahrensfehler:* sie treten immer dann auf, wenn eine als Grenzwert gegebene Größe (etwa der Wert einer unendlichen Reihe) durch einen endlichen Rechenausdruck (z.B. eine Partialsumme der Reihe) ersetzt wird.

Die Eliminationsverfahren gehören zu den *finiten* Verfahren, d.h. ihre Lösungen sind über endlich viele algebraische Operationen und nicht als Grenzwerte definiert, daher erscheinen hier keine Abbruchfehler.

Um die Fehler von Vektoren und Matrizen zu messen, führen wir in den folgenden Abschnitten Metriken und Normen ein. Im Interesse einer möglichst geschlossenen Darstellung wollen wir dazu auch Eigenschaften und Beispiele bringen, die zur Behandlung von Eliminationsaufgaben nicht benötigt werden, auf die wir aber in späteren Kapiteln zurückgreifen werden.

3.1. Metrische Räume

Vorgegeben sei eine Menge R und eine Abbildung $d: R \times R \to \mathbb{R}$.

(3.1.1) **Definition.** (R, d) heißt *metrischer Raum* mit der *Metrik* oder *Abstandsfunktion* d genau dann, wenn für alle $x, y, z \in R$ gilt:

(i) $d(x,y) \geq 0;\ d(x,y) = 0 \Leftrightarrow x = y,$

(ii) $d(x,y) = d(y,x),$

(iii) $d(x,z) \leq d(x,y) + d(y,z).$

Es sei nun (R, d) metrischer Raum, $x \in R$, $(x_n)_0^\infty$ eine Folge in R.

(3.1.2) **Definition.**

(i) Die Folge $(x_n)_0^\infty$ heißt *konvergent gegen* x, x heißt *Grenzwert* der Folge, abgekürzt geschrieben

$$x_n \to x \ (n \to \infty) \quad \text{oder} \quad \lim_{n \to \infty} x_n = x,$$

genau dann, wenn gilt

$$\forall \epsilon > 0 \ \exists N \in \mathbb{N} \ \forall n \geq N \ \ d(x_n, x) < \epsilon.$$

(ii) $(x_n)_0^\infty$ heißt *Cauchy-konvergent* (C-kgt.) genau dann, wenn gilt

$$\forall \epsilon > 0 \ \exists N \in \mathbb{N} \ \forall n, m \geq N \ \ d(x_n, x_m) < \epsilon$$

oder, kurz geschrieben:

$$d(x_n, x_m) \to 0 \qquad (n, m \to \infty) .$$

(3.1.3) **Bemerkung.** Für eine Folge $(x_n)_0^\infty$ im metrischen Raum bestätigt man durch Abschätzung unmittelbar die Aussagen:

(i) $\lim\limits_{n\to\infty} x_n = x, \ \lim\limits_{n\to\infty} x_n = y \ \Rightarrow \ x = y$;

(ii) $\exists x \in R \ \lim\limits_{n\to\infty} x_n = x \ \Rightarrow \ (x_n)_0^\infty$ Cauchy-konvergent.

Die Umkehrung von (3.1.3, ii) gilt nicht in jedem metrischen Raum.

(3.1.4) **Definition.** Ein metrischer Raum (R, d) heißt *vollständig* genau dann, wenn jede Cauchy-konvergente Folge in R einen Grenzwert in R besitzt.

Als *Beispiele für metrische Räume* erwähnen wir

(3.1.5) $R = \mathbb{Q}$ (Menge der rationalen Zahlen), $\mathbb{R}$ oder $\mathbb{C}$,
$d(x, y) := |x - y|$ (= Absolutbetrag von $x - y$),

und weiter mit der Bezeichnung $\mathbb{K} := \mathbb{R}$ oder $\mathbb{C}$:

(3.1.6) $R = \mathbb{K}^n$ mit den für $x = (\xi_i)_1^n$, $y = (\eta_i)_1^n$ wie folgt definierten Metriken:

$$d_\infty(x,y) = \max_{i=1}^{n} |\xi_i - \eta_i| ,$$

$$d_p(x,y) = \left(\sum_{i=1}^{n} |\xi_i - \eta_i|^p \right)^{\frac{1}{p}} \qquad \text{mit } 1 \leq p < \infty ,$$

$$d_p(x,y) = \sum_{i=1}^{n} |\xi_i - \eta_i|^p \qquad \text{mit } 0 < p \leq 1 .$$

Im Fall $p = 2$ erhalten wir $d_2(x,y)$ als den euklidischen Abstand zwischen x und y.

(3.1.7) Es sei $[a, b] \subseteq \mathbb{R}$ kompaktes Intervall; wir betrachten

$$R = C_0[a,b] := \{x : [a,b] \to \mathbb{K} \text{ stetig}\},$$
$$d(x,y) = \max_{t\in[a,b]} |x(t) - y(t)| .$$

(3.1.8) Für $R = C_1[a,b] := \{x : [a,b] \to \mathbb{K}$ stetig differenzierbar$\}$ geben wir verschiedene Metriken an, zunächst

$$d_1(x,y) = \max_{t\in[a,b]} |x(t) - y(t)| ,$$

also die Einschränkung der in (3.1.7) genannten Metrik auf die Teilmenge $C_1[a,b]$; außerdem sei

$$d_2(x,y) = \max_{t\in[a,b]} |x(t)-y(t)| + \max_{t\in[a,b]} |x'(t)-y'(t)|$$

und bei festem $t_0 \in [a,b]$:

$$d_3(x,y) = |x(t_0)-y(t_0)| + \max_{t\in[a,b]} |x'(t)-y'(t)| .$$

(3.1.9) Es sei M meßbare Menge im $\mathbb{R}^n$, $1 \leqq p < \infty$. Auf

$$R = \mathcal{L}_p(M) := \{x: M \to \mathbb{K} \text{ meßbar, } |x(t)|^p \text{ Lebesgue-integrabel}\}$$

ist durch

$$d(x,y) = \left(\int_M |x(t)-y(t)|^p \, dt\right)^{\frac{1}{p}}$$

zunächst eine *Pseudometrik* erklärt, d.h. statt (3.1.1, i) gilt wegen

$$d(x,y) = 0 \iff x(t) = y(t) \quad \text{fast überall in M}$$

die schwächere Bedingung

$$d(x,y) \geqq 0, \quad d(x,x) = 0 .$$

Um zu einem metrischen Raum zu gelangen, betrachten wir die durch

$$x \sim y \iff d(x,y) = 0 \quad (\iff x(t) = y(t) \text{ fast überall})$$

erklärte Äquivalenzrelation in R. Bezeichnet $\hat{x}$ die Äquivalenzklasse zu $x \in R$, so wird

$$\hat{R} = L_p(M) := \mathcal{L}_p(M)/\sim = \{\hat{x}: x \in R\}$$

mit

$$\hat{d}(\hat{x},\hat{y}) := d(x,y) \quad \text{für } x \in \hat{x},\ y \in \hat{y}$$

metrischer Raum, wie man anhand der Übungsaufgabe 3.1 nachrechnet. Im Fall $p = 2$ erhalten wir so den Hilbertschen Funktionenraum über M.

Bis auf die Metrik d_p mit $0 < p < 1$ in (3.1.6) sind die hier aufgeführten Metriken durch entsprechende Normen gemäß (3.2.3) erzeugt; zum Nachweis der Eigenschaften (3.1.1) rechnet man zweckmäßigerweise die Normeigenschaften nach.

Um zu zeigen, daß nicht allgemein im metrischen Raum R jede Cauchykonvergente Folge einen Grenzwert in R besitzt, notieren wir

(3.1.10) *Beispiele für nicht vollständige metrische Räume:*

(i) In $R = \mathbb{Q}$ mit $d(x,y) = |x-y|$ sei $(x_n)_0^\infty$ durch

$$x_0 = 1{,}4; \quad x_{n+1} = \frac{1}{2}\left(x_n + \frac{2}{x_n}\right) \quad (n = 0, 1, 2, \ldots)$$

definiert. Die Folge $(x_n)_0^\infty$ liegt offensichtlich in $\mathbb{Q}$, sie konvergiert in $\mathbb{R}$ gegen $\sqrt{2}$, ist also in $\mathbb{Q}$ Cauchy-konvergent, aber besitzt keinen Grenzwert in $\mathbb{Q}$.

(ii) Es sei $R = C_1[-1,1]$, $d_1(x,y) = \max_{t\in[-1,1]} |x(t)-y(t)|$.

Hier sind die Funktionen

$$x_n(t) = |t|^{1+\frac{1}{n}} \qquad (n = 1, 2, 3, \ldots)$$

sämtlich stetig differenzierbar und konvergieren in $[-1,1]$ gleichmäßig gegen $x(t) = |t|$ mit $x \notin C_1[-1,1]$. Daher ist die Folge $(x_n)_1^\infty$ in (R, d_1) Cauchy-konvergent, ohne in R einen Grenzwert zu besitzen. – Die Ausführung dieses Beispiels wird als Übungsaufgabe 3.2 empfohlen.

Eine Prüfung der übrigen Beispiele in (3.1.5)–(3.1.9) ergibt:

(3.1.11) $R = \mathbb{R}$ bzw. $\mathbb{C}$ ist vollständig, wie von der Analysis her bekannt ist.

(3.1.12) $R = \mathbb{K}^n$ mit jeder der angegebenen Metriken ist vollständig, da Konvergenz bezüglich der Metrik jeweils koordinatenweise Konvergenz bedeutet.

(3.1.13) Die Vollständigkeit von $(C_0[a,b], d)$ ist Inhalt des Satzes von Weierstraß über die gleichmäßige Konvergenz stetiger Funktionen, den wir hier beweisen wollen: dazu sei $(x_n)_0^\infty$ Cauchy-Folge in $(C_0[a,b], d)$, also

$$\max_{t\in[a,b]} |x_n(t) - x_m(t)| \to 0 \qquad (n, m \to \infty).$$

Dann ist für jedes feste $t \in [a,b]$ $(x_n(t))_0^\infty$ Cauchy-Folge in $\mathbb{K}$; wegen der Vollständigkeit von $\mathbb{K}$ existiert $x(t) \in \mathbb{K}$ mit

$$\lim_{n\to\infty} x_n(t) = x(t).$$

Auf diese Weise ist eine Funktion $x: [a,b] \to \mathbb{K}$ erklärt. Auf Grund der C-Konvergenz von $(x_n)_0^\infty$ haben wir:

$$\forall\, \epsilon > 0 \;\; \exists N \in \mathbb{N} \;\; \forall\, n, m \geqq N \;\; \forall\, t \in [a,b] \;\; |x_n(t) - x_m(t)| < \epsilon.$$

Betrachten wir in dieser Aussage bei festem $\epsilon > 0$, $n \geqq N$ den Grenzübergang $m \to \infty$, so folgt aus $\lim_{m\to\infty} x_m(t) = x(t)$:

$$\forall\, \epsilon > 0 \;\; \exists N \in \mathbb{N} \;\; \forall\, n \geqq N \;\; \forall\, t \in [a,b] \;\; |x_n(t) - x(t)| \leqq \epsilon,$$

d.h. $x_n(t) \to x(t)$ $(n \to \infty)$ gleichmäßig für $t \in [a,b]$. Zum Beweis der Stetigkeit von x geben wir $\epsilon > 0$ vor; dazu wählen wir uns ein $n \in \mathbb{N}$ mit der Eigenschaft, daß

$$\forall\, t \in [a,b] \;\; |x_n(t) - x(t)| < \frac{\epsilon}{3}$$

gilt. Wegen der Stetigkeit von x_n finden wir zu vorgegebenem $t_0 \in [a,b]$ ein $\delta > 0$, so daß für alle $t \in [a,b]$ mit $|t - t_0| < \delta$

$$|x_n(t) - x_n(t_0)| < \frac{\epsilon}{3}$$

erfüllt ist. Hieraus schließen wir für $|t - t_0| < \delta$:

$$|x(t) - x(t_0)| \leq |x(t) - x_n(t)| + |x_n(t) - x_n(t_0)| + |x_n(t_0) - x(t_0)| < \epsilon .$$

Damit liegt x in $C_0[a,b]$, und bezüglich der Metrik d haben wir $x_n \to x$ $(n \to \infty)$.

(3.1.14) $C_1[a,b]$ mit den in (3.1.8) angegebenen Metriken d_2 und d_3 ist vollständig, wie in Aufgabe 3.3 zu zeigen ist.

(3.1.15) Die Räume $L_p(M)$ sind vollständig metrische Räume nach dem Konvergenzsatz von Fischer – Riesz (vgl. z.B. Titchmarsh [27]).

Im Folgenden seien $(R_1, d_1), (R_2, d_2)$ beliebige metrische Räume und T eine Abbildung von R_1 in R_2 (man sagt statt Abbildung auch *Operator*).

(3.1.16) **Definition.** T heißt *beschränkt*, wenn ein $\gamma \geq 0$ existiert, so daß für alle $x, y \in R_1$

$$d_2(T(x), T(y)) \leq \gamma \cdot d_1(x, y) ;$$

für einen beschränkten Operator T nennen wir

$$|T| := \inf \{\gamma \geq 0 : \forall\, x, y \in R_1 \quad d_2(T(x), T(y)) \leq \gamma \cdot d_1(x,y)\}$$

den *Betrag von* T.

(3.1.17) **Bemerkung.** Ist T beschränkt, so gilt für alle $x, y \in R_1$

$$d_2(T(x), T(y)) \leq |T| \cdot d_1(x,y)$$

und damit

$$|T| = \min \{\gamma \geq 0 : \forall\, x, y \in R_1 \quad d_2(T(x), T(y)) \leq \gamma \cdot d_1(x,y)\} .$$

Zum *Beweis* seien $x, y \in R_1$ beliebig, aber fest vorgegeben; wir wählen $\gamma_n > |T|$ mit $\lim\limits_{n \to \infty} \gamma_n = |T|$. Dann haben wir für alle $n \in \mathbb{N}$

$$d_2(T(x), T(y)) \leq \gamma_n \cdot d_1(x,y) ,$$

daher dürfen wir die rechte Seite der Ungleichung durch ihren Grenzwert ersetzen.

Zum Begriff der Stetigkeit notieren wir die

(3.1.18) **Definition.** Es seien $(R_1, d_1), (R_2, d_2)$ metrische Räume, $T: R_1 \to R_2$ eine Abbildung.

(i) Für $z \in R_1$ heißt T *stetig in* z genau dann, wenn für alle Folgen $(x_n)_0^\infty$ in R_1 mit $\lim\limits_{n \to \infty} x_n = z$ gilt

$$\lim_{n \to \infty} T(x_n) = T(z) .$$

In diesem Fall schreiben wir kurz

$$T(x) \to T(z) \; (x \to z) \quad \text{oder} \quad \lim_{x \to z} T(x) = T(z) .$$

(ii) T heißt *stetig in* R_1 oder nur *stetig*, wenn T in jedem $z \in R_1$ stetig ist.

(3.1.19) **Bemerkung.** Jeder beschränkte Operator $T: R_1 \to R_2$ ist stetig.

Beweis. Bei vorgegebenem $z \in R_1$ haben wir für jede Folge $(x_n)_0^\infty$ in R_1 mit $\lim_{n\to\infty} x_n = z$

$$d_2(T(x_n), T(z)) \leqq |T| \cdot d_1(x_n, z); \quad d_1(x_n, z) \to 0 \quad (n \to \infty)$$

und daher auch $d_2(T(x_n), T(z)) \to 0$ $(n \to \infty)$.

Die Umkehrung von (3.1.19) ist im allgemeinen falsch; als Gegenbeispiel betrachtet man eine reelle, stetige Funktion, deren Differenzenquotienten nicht beschränkt sind.

(3.1.20) **Hilfssatz.** *Es seien* $(R_1, d_1), (R_2, d_2), (R_3, d_3)$ *metrische Räume,* $S: R_1 \to R_2$ *und* $T: R_2 \to R_3$ *beschränkte Operatoren. Dann gilt*

$$T \circ S: R_1 \to R_3 \quad \textit{ist beschränkt,}$$
$$|T \circ S| \leqq |T| \cdot |S|.$$

Beweis. Wir haben für $x, y \in R_1$

$$d_3(T(S(x)), T(S(y))) \leqq |T|\, d_2(S(x), S(y)) \leqq |T|\, |S|\, d_1(x,y);$$

daher ist $\gamma := |T|\, |S| \geqq 0$ eine Konstante mit

$$d_3(T \circ S(x), T \circ S(y)) \leqq \gamma \cdot d_1(x,y) \qquad (x, y \in R_1).$$

3.2. Normierte Vektorräume

Vorgegeben sei ein Vektorraum R über dem Körper $\mathbb{K} = \mathbb{R}$ oder $\mathbb{C}$ und eine Abbildung $\| \ \|: R \to \mathbb{R}$.

(3.2.1) **Definition.** $(R, \| \ \|)$ heißt normierter Vektorraum mit der Norm $\| \ \|$, wenn für alle $x, y \in R$, $\alpha \in \mathbb{K}$ gilt:

(i) $\|x\| \geqq 0, \quad \|x\| = 0 \iff x = 0;$

(ii) $\|\alpha x\| = |\alpha| \cdot \|x\|;$

(iii) $\|x + y\| \leqq \|x\| + \|y\|.$

Wir werden die Normen häufig mit einfachen Betragsstrichen schreiben: die Unterscheidung vom Betrag einer Zahl ergibt sich dann aus dem Zusammenhang.

Als Beispiele für normierte Vektorräume erwähnen wir

(3.2.2) $R = \mathbb{K}$, $|x|$ = Betrag von x;

(3.2.3) $R = \mathbb{K}^n$; für $x = (\xi_i)_1^n$ bezeichnen wir die Normen

$$\|x\|_p = \left(\sum_{i=1}^{n} |\xi_i|^p \right)^{\frac{1}{p}} \qquad (1 \leqq p < \infty) \text{ als } p\textit{-Norm},$$

speziell

$$\|x\|_2 = \left(\sum_{i=1}^{n} |\xi_i|^2 \right)^{\frac{1}{2}} \qquad \text{als } \textit{euklidische Norm},$$

$$\|x\|_\infty = \max_{i=1}^{n} |\xi_i| \qquad \text{als } \textit{Maximumsnorm}$$

und mit $w = (w_i)_1^n$, $w_i > 0$ $\qquad (i = 1, \ldots, n)$:

$$\|x\|_{p,w} = \left(\sum_{i=1}^{n} \left(\frac{|\xi_i|}{w_i} \right)^p \right)^{\frac{1}{p}} \qquad (1 \leqq p < \infty) \text{ als } \textit{gewichtete p-Norm},$$

$$\|x\|_w = \max_{i=1}^{n} \frac{|\xi_i|}{w_i} \quad , \qquad \text{als } \textit{gewichtete Maximumsnorm} \text{ oder einfach } \textit{w-Norm}.$$

(3.2.4) In $R = C_0[a,b]$ wird durch $\|x\| = \max_{t \in [a,b]} |x(t)|$ die *Maximums-* oder *Tschebyscheff-Norm* definiert.

Entsprechend (3.1.8) und (3.1.9) lassen sich in $C_1[a,b]$ und $L_p(M)$ Normen erklären. – Die Dreiecksungleichung für die p-Norm im $\mathbb{K}^n$ (und $L_p(M)$) heißt Minkowski-Ungleichung und wird für $p > 1$ mit der Hölderschen Ungleichung

$$\sum_{i=1}^{n} |\xi_i \eta_i| \leqq \left(\sum_{i=1}^{n} |\xi_i|^p \right)^{\frac{1}{p}} \left(\sum_{i=1}^{n} |\eta_i|^q \right)^{\frac{1}{q}} \qquad \left(\frac{1}{q} := 1 - \frac{1}{p} \right)$$

bewiesen. In numerischen Anwendungen treten außer der (gewichteten) Maximumsnorm die (gewichteten) p-Normen fast nur für $p = 1$ oder 2 auf; hierfür sind die Normeigenschaften leicht nachzurechnen.

(3.2.5) **Bemerkung.** Im normierten Vektorraum $(R, |\ |)$ ist durch $d: R \times R \to \mathbb{R}$ mit

$$d(x,y) := |x - y|$$

eine Metrik erklärt, d nennt man die *von der Norm erzeugte Metrik.*

Den *Beweis* erhalten wir durch einfaches Einsetzen. – Konvergenz im normierten Vektorraum bedeutet Konvergenz bezüglich der durch die Norm erzeugten Metrik. Man nennt einen normierten Vektorraum *vollständig* oder auch einen *Banach-Raum,* wenn der zugeordnete metrische Raum vollständig ist. Nach den Überlegungen (3.1.11)–(3.1.13) sind die in den Beispielen (3.2.2)–(3.2.4) genannten Räume sämtlich Banach-Räume.

Stetigkeit in normierten Vektorräumen bedeutet Stetigkeit in den zugehörigen metrischen Räumen; der folgende Hilfssatz zeigt, daß im normierten Vektorraum die Rechenoperationen und die Norm stetig sind.

(3.2.6) **Hilfssatz.** *Es sei* $(R, \|\ \|)$ *normierter Vektorraum über* $\mathbb{K}$, *es seien* $x, y \in R$, $(x_n)_0^\infty$, $(y_n)_0^\infty$ *Folgen in* R, $\alpha \in \mathbb{K}$, $(\alpha_n)_0^\infty$ *eine Folge in* $\mathbb{K}$. *Dann gilt*

(i) $\big|\, \|x\| - \|y\| \,\big| \leqq \|x - y\|$,

(ii) $\lim\limits_{n\to\infty} x_n = x \Rightarrow \lim\limits_{n\to\infty} \|x_n\| = \|x\|$ (*Konvergenz in* $\mathbb{R}$!),

(iii) $\lim\limits_{n\to\infty} \alpha_n = \alpha, \quad \lim\limits_{n\to\infty} x_n = x \Rightarrow \lim\limits_{n\to\infty} \alpha_n x_n = \alpha x$,

(iv) $\lim\limits_{n\to\infty} x_n = x, \quad \lim\limits_{n\to\infty} y_n = y \Rightarrow \lim\limits_{n\to\infty} (x_n + y_n) = x + y$.

Beweis. Auf Grund der Dreiecksungleichung erhalten wir

$$\|x\| = \|y + (x - y)\| \leqq \|y\| + \|x - y\| \Rightarrow \|x\| - \|y\| \leqq \|x - y\|$$

und ebenso durch Vertauschen von x und y

$$\|y\| - \|x\| = -(\|x\| - \|y\|) \leqq \|y - x\| = \|x - y\|$$

und damit die Aussage (i); es folgt (ii) unmittelbar wegen

$$\big|\, \|x_n\| - \|x\| \,\big| \leqq \|x_n - x\| \to 0 \quad (n \to \infty).$$

Zu (iii) schätzen wir folgendermaßen ab:

$$0 \leqq \|\alpha_n x_n - \alpha x\| = \|\alpha_n (x_n - x) + (\alpha_n - \alpha) x\| \leqq |\alpha_n|\, \|x_n - x\| + |\alpha_n - \alpha|\, \|x\|,$$

wobei $|\alpha_n - \alpha| \to 0$, $\|x_n - x\| \to 0$ $(n \to \infty)$, $|\alpha_n| \leqq M < \infty$ auf Grund der Konvergenz gilt, woraus $\|\alpha_n x_n - \alpha x\| \to 0$ $(n \to \infty)$ folgt. Zu (iv) notieren wir

$$0 \leqq \|(x_n + y_n) - (x + y)\| = \|(x_n - x) + (y_n - y)\|$$
$$\leqq \|x_n - x\| + \|y_n - y\| \to 0 \quad (n \to \infty).$$

Für Vektorräume R_1, R_2 über $\mathbb{K}$ bezeichnen wir

$$\mathrm{Hom}(R_1, R_2) := \{A : R_1 \to R_2 \text{ linear}\},$$
$$\mathrm{Hom}(R_1) := \mathrm{Hom}(R_1, R_1).$$

In Anlehnung an die Matrizenschreibweise setzen wir bei einer linearen Abbildung das Argument meistens ohne Klammern hinter den Namen der Abbildung, schreiben also Ax statt A(x). Führt man wie üblich durch

$$(A_1 + A_2)x := A_1 x + A_2 x, \quad (\alpha A)x := \alpha(Ax) \quad (\alpha \in \mathbb{K})$$

Addition und Multiplikation in $\mathrm{Hom}(R_1, R_2)$ ein, so wird $\mathrm{Hom}(R_1, R_2)$ bezüglich dieser Operationen ein Vektorraum über $\mathbb{K}$.

Es seien nun $(R_1, |\ |_1), (R_2, |\ |_2)$ normierte Vektorräume, $T \in \mathrm{Hom}(R_1, R_2)$. Dann gilt im Sinne der Definition (3.1.16), auf die von den Normen $|\ |_1$ und $|\ |_2$ erzeugten Metriken d_1 und d_2 angewendet, die

(3.2.7) **Bemerkung.**

(i) T ist beschränkt genau dann, wenn ein $\gamma \geqslant 0$ existiert, so daß für alle $x \in R_1$

$$|Tx|_2 \leqslant \gamma |x|_1 .$$

Für einen beschränkten Operator $T \in \mathrm{Hom}(R_1, R_2)$ haben wir

(ii) $|T| = \min \{\gamma \geqq 0 : \forall x \in R_1 \ |Tx|_2 \leqq \gamma |x|_1\}$,

das heißt

(iii) $$|T| = \begin{cases} 0, & \text{falls } R_1 = \{0\}, \\ \sup \left(\dfrac{|Tx|_2}{|x|_1} : x \in R_1, \neq 0 \right) & \text{sonst;} \end{cases}$$

insbesondere gilt für alle $x \in R_1$

(iv) $|Tx|_2 \leqq |T| \cdot |x|_1$.

Zum Beweis von (i) und (ii) überlegen wir uns für beliebiges $\gamma \geqq 0$ die Äquivalenzen

$$\begin{aligned} &\forall x, y \in R_1 \quad d_2(T(x), T(y)) \leqq \gamma \, d_1(x, y) \\ \Leftrightarrow \quad &\forall x, y \in R_1 \quad |Tx - Ty|_2 = |T(x-y)|_2 \leqq \gamma \cdot |x-y|_1 \\ \Leftrightarrow \quad &\forall z \in R_1 \quad |Tz|_2 \leqq \gamma \cdot |z|_1 . \end{aligned}$$

Zur Begründung der letzten Äquivalenz setzen wir einmal $x = z$, $y = 0$, und umgekehrt $z = x - y$. Die Aussage (iii) ergibt sich aus der Definition des Supremums als der kleinsten oberen Schranke.

Zur weiteren Charakterisierung von beschränkten linearen Operatoren zeigen wir den

(3.2.8) **Hilfssatz.** *Es seien* $(R_1, |\ |_1)$, $(R_2, |\ |_2)$ *normierte Vektorräume,* $T \in \mathrm{Hom}(R_1, R_2)$. *Dann sind die folgenden Aussagen äquivalent:*

(i) T *stetig (in* R_1),
(ii) T *stetig in* 0,
(iii) T *beschränkt*.

Beweis. Die Folgerung (i) ⇒ (ii) ist klar, (iii) ⇒ (i) ist mit (3.1.19) bewiesen; um (ii) ⇒ (iii) zu zeigen, nehmen wir an, es sei T nicht beschränkt, d.h.

$$\forall \gamma > 0 \quad \exists x \in R_1 \quad |Tx|_2 > \gamma |x|_1 .$$

Speziell für $\gamma = k \in \mathbb{N}$ erhält man $x_k \in R_1$ mit

$$|Tx_k|_2 > k \, |x_k|_1 \qquad (k = 1, 2, \ldots) .$$

Da wegen $Tx_k \neq 0$ auch $x_k \neq 0$ gilt, setzen wir

$$y_k = \frac{1}{k \, |x_k|_1} x_k \qquad (k = 1, 2, \ldots)$$

und haben hiermit

$$|y_k|_1 = \frac{1}{k} \to 0 \quad (k \to \infty), \quad |Ty_k|_2 = \frac{1}{k\,|x_k|_1}\,|Tx_k|_2 > 1\,,$$

im Widerspruch zur Stetigkeit in 0.

Als *Beispiel* für einen linearen beschränkten Operator im unendlichdimensionalen Fall betrachten wir $R = C_0[a,b]$ mit der Maximumsnorm (vgl. (3.2.4)) und definieren zu einer vorgegebenen Funktion

$$K: [a,b] \times [a,b] \to \mathbb{K} \text{ stetig},$$

die wir mit Argumenten als $K(s,t)$ schreiben, den Operator T durch

$$(Tx)(s) := \int_a^b K(s,t)\,x(t)\,dt \qquad (x \in C_0[a,b],\ s \in [a,b])\,.$$

An Hand der gleichmäßigen Stetigkeit von K weist man die Stetigkeit der in $[a,b]$ definierten Funktion Tx nach, insgesamt ist

$$T: C_0[a,b] \to C_0[a,b] \quad \text{linear}.$$

Für $x \in C_0[a,b]$ notieren wir

$$\|Tx\| = \max_{s\in[a,b]} |(Tx)(s)| \leqq \max_{s\in[a,b]} \int_a^b |K(s,t)|\,|x(t)|\,dt$$

$$\leqq \left(\max_{s\in[a,b]} \int_a^b |K(s,t)|\,dt \right) \cdot \|x\|\,,$$

also ist T beschränkter Operator mit $|T| \leqq \max_{s\in[a,b]} \int_a^b |K(s,t)|\,dt$.

(3.2.9) **Definition.** Für normierte Vektorräume $(R_1, |\ |_1)$, $(R_2, |\ |_2)$ bezeichne

$$L(R_1, R_2) := \{T: R_1 \to R_2 \text{ linear, beschränkt}\},$$
$$L(R_1) := L(R_1, R_1)\,.$$

Dazu zeigen wir die

(3.2.10) **Bemerkungen.**

(i) $L(R_1, R_2)$ ist Teilvektorraum von $\mathrm{Hom}(R_1, R_2)$ und mit $|\ |$, definiert nach (3.1.16), normierter Vektorraum. Wir bezeichnen $|\ |$ auch als *Operatornorm*.

(ii) Ist $(R_2, |\ |_2)$ vollständig, so ist auch $(L(R_1, R_2), |\ |)$ Banach-Raum.

Beweis. Offenbar liegt die Nullabbildung in $L(R_1, R_2)$; um weiter die Abgeschlossenheit von $L(R_1, R_2)$ bezüglich der Vektorraumoperationen zu zeigen, notieren wir für $T, S \in L(R_1, R_2)$, $\alpha, \beta \in \mathbb{K}$ und beliebiges $x \in R_1$ die Abschätzung

$$|(\alpha T + \beta S)x|_2 = |\alpha Tx + \beta Sx|_2 \leqq |\alpha|\,|Tx|_2 + |\beta|\,|Sx|_2 \leqq (|\alpha|\,|T| + |\beta|\,|S|)\,|x|_1 .$$

Hieraus folgt die Beschränktheit von $\alpha T + \beta S$ sowie

$$(*) \quad |\alpha T + \beta S| \leqq |\alpha|\,|T| + |\beta|\,|S| .$$

Die Normeigenschaft $|T| \geqq 0$ ist klar, außerdem gilt

$$|T| = 0 \Leftrightarrow \forall x \in R_1 \;\; |Tx|_2 = 0 \Leftrightarrow \forall x \in R_1 \;\; Tx = 0 \Leftrightarrow T = 0 .$$

Aus $(*)$ mit $S = 0$ folgt

$$|\alpha T| \leqq |\alpha|\,|T| \qquad (\alpha \in \mathbb{K})$$

und für $\alpha \neq 0$

$$|T| = \left| \frac{1}{\alpha}(\alpha T) \right| \leqq \frac{1}{|\alpha|}|\alpha T| \Rightarrow |\alpha|\,|T| \leqq |\alpha T| ,$$

also

$$|\alpha T| = |\alpha|\,|T| \qquad (\alpha \in \mathbb{K}) ,$$

denn für $\alpha = 0$ ist diese Beziehung wegen $|0| = 0$ erfüllt. Schließlich erhalten wir aus $(*)$ mit $\alpha = \beta = 1$:

$$|T + S| \leqq |T| + |S| .$$

Zum Nachweis von (ii) sei eine Cauchy-Folge $(A_n)_0^\infty$ in $(L(R_1, R_2), |\;|)$ vorgegeben, d.h. $|A_n - A_m| \to 0$ $(n, m \to \infty)$. Für festes $x \in R_1$ haben wir

$$|A_n x - A_m x|_2 = |(A_n - A_m)x|_2 \leqq |A_n - A_m|\,|x|_1 \to 0 \qquad (n, m \to \infty) ,$$

d.h. $(A_n x)_0^\infty$ ist Cauchy-Folge im Banach-Raum $(R_2, |\;|_2)$, folglich existiert ein $A(x) \in R_2$ mit $A_n x \to A(x)$ $(n \to \infty)$. Zunächst ist durch $x \mapsto A(x)$ eine Abbildung $A: R_1 \to R_2$ gegeben. Wir zeigen nacheinander:

$$A \text{ linear}, \quad A \text{ beschränkt}, \quad A_n \to A \;\; (n \to \infty) \text{ in } L(R_1, R_2), |\;|) .$$

Für $x, y \in R_1$, $\alpha, \beta \in \mathbb{K}$ haben wir

$$A_n(\alpha x + \beta y) = \alpha(A_n x) + \beta(A_n y) \qquad (n \in \mathbb{N}) .$$

Für $n \to \infty$ strebt der linke Ausdruck gegen $A(\alpha x + \beta y)$, der rechte nach (3.2.6), (iii) und (iv) gegen $\alpha A(x) + \beta A(y)$. Die Eindeutigkeit des Grenzwerts liefert $A(\alpha x + \beta y) = \alpha A(x) + \beta A(y)$ und damit die Linearität von A. – Zu vorgegebenem $\epsilon > 0$ gibt es ein $N \in \mathbb{N}$ mit

$$|A_n - A_m| < \epsilon \qquad (n, m \geqq N) ;$$

daher gilt für jedes $x \in R_1$

$$|(A_n - A_m)x|_2 \leqq |A_n - A_m|\,|x|_1 \leqq \epsilon\,|x|_1 \qquad (n, m \geqq N) .$$

Wir betrachten bei festen $x \in R_1$, $n \geqq N$ den Grenzwert des linksstehenden Ausdrucks für $m \to \infty$ und folgern mit (3.2.6, i):

$$|(A_n - A)x|_2 \leqq \epsilon |x|_1 \qquad (x \in R_1),$$

d.h. für $n \geqq N$ ist $A_n - A \in L(R_1, R_2)$ und damit

$$A = A_n - (A_n - A) \in L(R_1, R_2),$$

außerdem haben wir $|A_n - A| \leqq \epsilon$ $(n \geqq N)$ und daher $\lim\limits_{n \to \infty} A_n = A$.

In Hom(R), wobei R ein Vektorraum ist, wird die Multiplikation von T, $S \in \mathrm{Hom}(R)$ als Hintereinanderausführung der Abbildungen, nämlich

$$TS := T \circ S$$

erklärt. Im Fall des normierten Vektorraums $(R, |\ |)$ haben wir nach (3.1.20):

(3.2.11) $\quad T, S \in L(R) \Rightarrow TS \in L(R), \quad |TS| \leqq |T|\,|S|$;

damit ist die Multiplikation eine Operation in L(R). Als Stetigkeitsaussage gewinnen wir die

(3.2.12) **Bemerkung.** Es seien $(R, |\ |)$ normierter Vektorraum, $A, B \in L(R)$, $(A_k)_0^\infty$, $(B_k)_0^\infty$ Folgen in L(R). Dann gilt

$$\lim_{k \to \infty} A_k = A, \quad \lim_{k \to \infty} B_k = B \quad \Rightarrow \quad \lim_{k \to \infty} A_k B_k = AB\,.$$

Zum *Beweis* schätzen wir $|A_k B_k - AB|$ durch

$$|A_k B_k - AB| \leqq |(A_k - A)B_k| + |A(B_k - B)| \leqq |A_k - A|\,|B_k| + |A|\,|B_k - B|$$

ab, wobei $|A_k - A| \to 0$, $|B_k - B| \to 0$ $(k \to \infty)$ und gemäß (3.2.6, ii) mit einer positiven Konstanten M die Abschätzung $|B_k| \leqq M$ $(k \in \mathbb{N})$ gilt.

Für $C \in L(R)$ definieren wir

$$C^0 := I = \text{Identität in R}$$

und rekursiv

$$C^{n+1} := C(C^n) \qquad (n = 0, 1, 2, \ldots)\,.$$

Wegen (3.2.11) sind dann sämtliche $C^n \in L(R)$.

Nach diesen Vorbemerkungen beweisen wir den 1. *Stabilitätssatz von Banach* (auch Satz über die *Neumannsche Reihe* genannt):

(3.2.13) **Satz.**

(i) *Es sei* $(R, |\ |)$ *Banach-Raum,* $C \in L(R)$ *mit*

$$\sum_{n=0}^{\infty} |C^n| < \infty\,.$$

Dann ist $I-C$ *bijektiv,* $(I-C)^{-1} \in L(R)$, *und es gilt bezüglich der Norm in* $L(R)$

$$(I-C)^{-1} = \sum_{n=0}^{\infty} C^n := \lim_{k\to\infty}\left(\sum_{n=0}^{k} C^n\right).$$

(ii) *Speziell für* $C \in L(R)$ *mit* $|C| < 1$ *ist* $\sum_{n=0}^{\infty} |C^n| < \infty$ *erfüllt.*

Beweis. Bekanntlich ist $A = I - C$ bijektiv mit der Inversen B genau dann, wenn $BA = AB = I$ erfüllt ist. Wir setzen

$$B_k := \sum_{n=0}^{k} C^n \qquad (k = 0, 1, 2, \ldots);$$

dann sind die $B_k \in L(R)$, und für $k \geqslant l$ haben wir

$$|B_k - B_l| = \Big|\sum_{n=l+1}^{k} C^n\Big| \leqq \sum_{n=l+1}^{k} |C^n| \to 0 \qquad (k, l \to \infty);$$

daher ist die Folge $(B_k)_0^{\infty}$ in $L(R)$ Cauchy-konvergent. Wegen der Vollständigkeit von $L(R)$ existiert ein $B \in L(R)$ mit $B_k \to B$ $(k \to \infty)$. Man rechnet leicht die Beziehungen

$$(I-C)B_k = B_k(I-C) = I - C^{k+1} \qquad (k \in \mathbb{N})$$

nach, wobei wegen $|C^{k+1}| \to 0$ $(k \to \infty)$ die Folge $I - C^{k+1}$ gegen I konvergiert. Es folgt nach Hilfssatz (3.2.12), da konstante Folgen stets konvergent sind:

$$\lim_{k\to\infty}(I-C)B_k = (I-C)B, \qquad \lim_{k\to\infty} B_k(I-C) = B(I-C),$$

womit $(I-C)B = B(I-C) = I$ erfüllt ist. Zu (ii) folgern wir aus $|C| < 1$ und aus den durch Induktion zu zeigenden Ungleichungen $|C^n| \leqq |C|^n$ $(n \in \mathbb{N})$ die Reihenabschätzung

$$\sum_{n=0}^{\infty} |C^n| \leqq \sum_{n=0}^{\infty} |C|^n = \frac{1}{1-|C|} < \infty.$$

(3.2.14) **Definition.** Für einen normierten Vektorraum $(R, |\ |)$ bezeichne

$$J(R) := \{T \in L(R) : T \text{ bijektiv, } T^{-1} \in L(R)\}.$$

Wir bemerken dazu, daß die Eigenschaft $T^{-1} \in L(R)$ nach dem Satz vom abgeschlossenen Graphen stets erfüllt ist, wenn R Banach-Raum und $T \in L(R)$ ein bijektiver Operator ist (vgl. Kato [16], Kap. III).

Als Anwendung des Stabilitätssatzes gewinnen wir die

(3.2.15) **Folgerung.** Es sei R Banach-Raum und $A \in J(R)$, ferner $B \in L(R)$ mit

$$|A^{-1}(A-B)| < 1$$

oder stärker

$$|A^{-1}|\,|A-B| < 1\,.$$

Dann gilt

$$B \in J(R), \quad B^{-1} = \left[\sum_{n=0}^{\infty} (A^{-1}(A-B))^n\right] A^{-1}\,.$$

Beweis. Aus der im allgemeinen leichter zu prüfenden Bedingung $|A^{-1}|\,|A-B| < 1$ folgt $|A^{-1}(A-B)| \leqq |A^{-1}|\,|A-B| < 1$ unmittelbar. – Wir schreiben

$$B = A + (B-A) = A(I-C), \qquad C := A^{-1}(A-B)$$

und haben nach Voraussetzung $C \in L(R)$ mit $|C| < 1$. Nach Satz (3.2.13) existiert $(I-C)^{-1}$ mit $(I-C)^{-1} \in L(R)$; als Produkt invertierbarer Operatoren ist auch B invertierbar, und es gilt

$$B^{-1} = (I-C)^{-1}A^{-1} = \left(\sum_{n=0}^{\infty} C^n\right) A^{-1} \in L(R)\,.$$

3.3. Endlichdimensionale normierte Vektorräume

Für $\mathbb{K} = \mathbb{R}$ oder $\mathbb{C}$ erinnern wir an die im $\mathbb{K}^n$ nach (3.2.3) erklärte Norm

$$\|x\|_1 = \sum_{i=1}^{n} |\xi_i| \qquad \text{für} \quad x = (\xi_i)_1^n\,.$$

(3.3.1) **Hilfssatz.** *Es sei* $(R, |\ |)$ *normierter Vektorraum über* $\mathbb{K}$, $\dim R = n < \infty$. *Dann gilt*

$$(R, |\ |) \cong (\mathbb{K}^n, \|\ \|_1)$$

in dem Sinne, daß eine bijektive, lineare und stetige Abbildung $\varphi\colon R \to \mathbb{K}^n$ *mit stetiger Umkehrabbildung* φ^{-1} *existiert. Insbesondere ist* $(R, |\ |)$ *Banach-Raum.*

Beweis. Wir wählen eine Basis $(v_1, \ldots, v_n)$ von R und definieren $\varphi\colon R \to \mathbb{K}^n$ durch

$$\varphi(x) := (\lambda_i)_1^n, \quad \text{falls} \quad x = \sum_{i=1}^{n} \lambda_i v_i$$

Bekanntlich ist φ eine lineare, bijektive Abbildung. Die Beschränktheit und damit die Stetigkeit von φ beweisen wir indirekt, indem wir annehmen:

$$\forall\, k \in \mathbb{N} \quad \exists\, x_k \in R \quad \|\varphi(x_k)\|_1 > k\,|x_k|\,.$$

Dazu bezeichnen wir

$$y_k := \frac{1}{\|\varphi(x_k)\|_1} x_k \in R, \quad b_k := \varphi(y_k) \in \mathbb{K}^n \qquad (k \in \mathbb{N})$$

und erhalten hiermit $|y_k| < \frac{1}{k}$ $(k \geq 1)$, also $\lim\limits_{k\to\infty} y_k = 0$ sowie $\|b_k\|_1 = 1$ $(k \in \mathbb{N})$. Da die „Einheitskugel" des $\mathbb{K}^n$ kompakt ist, existiert eine Teilfolge $(b_{k_\nu})_{\nu=0}^{\infty}$ von $(b_k)_{k=0}^{\infty}$ und ein $b \in \mathbb{K}^n$ mit $\lim\limits_{\nu\to\infty} b_{k_\nu} = b$, wobei wegen $\|b_{k_\nu}\|_1 = 1$ auch $\|b\|_1 = 1$ gilt. Für die Komponenten von $b := (\beta_i)_{i=1}^n$, $b_k := (\beta_i^{(k)})_{i=1}^n$ folgt

$$\lim_{\nu\to\infty} \beta_i^{(k_\nu)} = \beta_i \qquad (i = 1, \ldots, n)$$

und nach (3.2.6, iii) und (iv):

$$\lim_{\nu\to\infty} y_{k_\nu} = \lim_{\nu\to\infty} \left(\sum_{i=1}^{n} \beta_i^{(k_\nu)} v_i \right) = \sum_{i=1}^{n} \beta_i v_i \neq 0 ,$$

letzteres wegen $\|b\|_1 = 1$. Das ist ein Widerspruch zu $\lim\limits_{k\to\infty} y_k = 0$.

Die Stetigkeit von φ^{-1} folgt aus der für $y = (\lambda_i)_1^n \in \mathbb{K}^n$ gültigen Abschätzung

$$|\varphi^{-1}(y)| = \left| \sum_{i=1}^{n} \lambda_i v_i \right| \leq \sum_{i=1}^{n} |\lambda_i| \, |v_i| \leq \beta \|y\|_1 ,$$

in der

$$\beta = \max_{i=1}^{n} |v_i|$$

gesetzt ist.

Zum Nachweis der Vollständigkeit von R sei eine Cauchy-Folge $(x_k)_0^\infty$ in R vorgegeben. Dann ist wegen

$$\|\varphi(x_k) - \varphi(x_m)\|_1 \leq \gamma |x_k - x_m| \qquad (k, m \in \mathbb{N})$$

auch $(\varphi(x_k))_0^\infty$ Cauchy-Folge im $\mathbb{K}^n$. Auf Grund der Vollständigkeit des $\mathbb{K}^n$ existiert ein $y \in \mathbb{K}^n$ mit $\lim\limits_{k\to\infty} \varphi(x_k) = y$; aus der Stetigkeit von φ^{-1} folgt für $x = \varphi^{-1}(y)$

$$\lim_{k\to\infty} x_k = \lim_{k\to\infty} \varphi^{-1}(\varphi(x_k)) = x .$$

(3.3.2) **Hilfssatz.** *Es seien* $(R_1, |\ |_1)$, $(R_2, |\ |_2)$ *normierte Vektorräume über* $\mathbb{K}$, $\dim R_1 = n < \infty$, $A \in \mathrm{Hom}(R_1, R_2)$. *Dann ist* A *stetig, also* $\mathrm{Hom}(R_1, R_2) = L(R_1, R_2)$.

Zum *Beweis* wählen wir eine Basis $(v_1, \ldots, v_n)$ von R_1; dann gilt nach Hilfssatz (3.3.1) mit einem $\gamma > 0$ für alle

$$x = \sum_{i=1}^{n} \lambda_i v_i \in R_1$$

die Abschätzung

$$\sum_{i=1}^{n} |\lambda_i| \leqq \gamma \cdot |x|_1 .$$

Aus der Darstellung

$$Ax = \sum_{i=1}^{n} \lambda_i (Av_i)$$

folgern wir unmittelbar die Ungleichungen

$$|Ax|_2 \leqq \max_{i=1}^{n} |Av_i|_2 \cdot \sum_{i=1}^{n} |\lambda_i| \leqq \beta \cdot |x|_1$$

mit der von x unabhängigen Konstanten

$$\beta := \gamma \cdot \max_{i=1}^{n} |Av_i|_2$$

und hieraus die Stetigkeit von A.

(3.3.3) **Folgerung.** Ist R endlichdimensionaler Vektorraum über $\mathbb{K}$, so sind je zwei Normen $|\ |_1$ und $|\ |_2$ in R zueinander äquivalent, d.h.

$$\exists \beta, \gamma \geqq 0 \quad \forall x \in R \quad |x|_1 \leqq \gamma |x|_2, \quad |x|_2 \leqq \beta |x|_1 .$$

Für Folgen $(x_k)_0^\infty$ in R gilt mit $x \in R$ stets

$$\lim_{k \to \infty} x_k = x \text{ bezüglich } |\ |_1 \iff \lim_{k \to \infty} x_k = x \text{ bezüglich } |\ |_2 .$$

Zum *Beweis* verwenden wir den Hilfssatz (3.3.2), wonach die Identität als lineare Abbildung von $(R, |\ |_1)$ in $(R, |\ |_2)$ stetig ist, folglich

$$\exists \beta \geqq 0 \quad \forall x \in R \quad |x|_2 \leqq \beta |x|_1 ;$$

entsprechendes gilt bei Vertauschung der Normen $|\ |_1$ und $|\ |_2$. Die Stetigkeit der Identität liefert auch, daß die Konvergenz einer Folge von der gewählten Norm unabhängig ist.

Insbesondere sind im $\mathbb{K}^n$ alle Normen zueinander äquivalent; zu je zwei der speziellen in (3.2.3) aufgeführten Normen lassen sich scharfe Abschätzungskonstanten β und γ angeben (vgl. Aufgabe 3.4!). Nach Hilfssatz (3.3.2) sind alle linearen Abbildungen von $\mathbb{K}^k$ in $\mathbb{K}^n$ $(k, n \in \mathbb{N})$ stetig. Bezeichnet

$$M(n \times k, \mathbb{K}) := \{A = (a_{i,j})_{(n,k)} : a_{i,j} \in \mathbb{K}\},$$

also den Vektorraum aller (n,k)-Matrizen mit Koeffizienten in $\mathbb{K}$, so gilt bekanntlich

$$M(n \times k, \mathbb{K}) \cong \mathrm{Hom}(\mathbb{K}^k, \mathbb{K}^n),$$

wobei die Isomorphie

$$A = (a_{i,j})_{(n,k)} \mapsto T \in \mathrm{Hom}(\mathbb{K}^k, \mathbb{K}^n)$$

durch

$$T(x) = \left(\sum_{j=1}^{k} a_{i,j}\, \xi_j \right)_{j=1}^{n} \qquad \text{für} \quad x = (\xi_j)_{j=1}^{k} \in \mathbb{K}^k$$

gegeben ist. Da sich $T(x)$ als das Matrizenprodukt von A mit der $(k,1)$-Spalte $(\xi_j)_1^k$ ausdrücken läßt, schreiben wir meistens Ax statt $T(x)$.

In den folgenden Überlegungen betrachten wir nur lineare Abbildungen von $\mathbb{K}^n$ in sich, also (n,n)-Matrizen, wobei im Urbild- und Bildraum jeweils dieselbe Norm gewählt sei.

(3.3.4) **Definition.** Ist $\|\ \|$ beliebige Norm im $\mathbb{K}^n (n \geqq 1)$, so bezeichnen wir für $A \in M(n \times n, \mathbb{K})$

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

als *Operatornorm von* A *zur Norm* $\|\ \|$ *im* $\mathbb{K}^n$.

Offensichtlich gilt

$$\|A\| = \sup \{ \|Ax\| : x \in \mathbb{K}^n, |x| = 1 \} = \max \{ \|Ax\| : x \in \mathbb{K}^n, |x| = 1 \},$$

letzteres, da die Menge $\{x \in \mathbb{K}^n : |x| = 1\}$ kompakt und die Abbildung $x \mapsto \|Ax\|$ stetig ist.

Da für einige Normen des $\mathbb{K}^n$ die zugehörigen Operatornormen von A schwer auszurechnen sind, führt man Normen im $M(n \times n, \mathbb{K})$ ein, die unmittelbar durch die Koeffizienten von A beschrieben werden.

(3.3.5) **Definition.** Eine Abbildung $\|\ \|: M(n \times n, \mathbb{K}) \to \mathbb{R}$ heißt *Matrixnorm* genau dann, wenn $\|\ \|$ Norm im Vektorraum $M(n \times n, \mathbb{K})$ ist und die zusätzliche Eigenschaft

$$\|AB\| \leqq \|A\| \cdot \|B\| \qquad (A, B \in M(n \times n, \mathbb{K}))$$

besitzt.

Mit einer Matrixnorm wird $M(n \times n, \mathbb{K})$ *normierte Algebra;* wegen $\dim M(n \times n, \mathbb{K}) = n^2 < \infty$ ist der entsprechende metrische Raum vollständig; daher ist $M(n \times n, \mathbb{K})$ sogar eine *Banach-Algebra.*

(3.3.6) **Definition.** Es sei $\|\ \|$ eine Norm im $\mathbb{K}^n$, $\|A\|$ die zugehörige Operatornorm zu $A \in M(n \times n, \mathbb{K})$, $\|\ \|$ Matrixnorm in $M(n \times n, \mathbb{K})$.

(i) Die Matrixnorm $\|\ \|$ heißt *passend zur Norm* $\|\ \|$ *im* $\mathbb{K}^n$ genau dann, wenn

$$\forall\, A \in M(n \times n, \mathbb{K}) \qquad \|A\| \leqq \|A\|,$$

das heißt

$$\forall A \in M(n \times n, \mathbb{K}) \qquad \forall x \in \mathbb{K}^n \qquad |Ax| \leq \|A\| \cdot |x| .$$

(ii) Die Matrixnorm $\| \ \|$ heißt *der Norm* $| \ |$ *im* $\mathbb{K}^n$ *zugeordnet*, wenn sie gleich der Operatornorm ist. Dies ist gleichbedeutend damit, daß $\| \ \|$ zur Norm $| \ |$ im $\mathbb{K}^n$ passende Matrixnorm ist und zusätzlich die Bedingung

$$\forall A \in M(n \times n, \mathbb{K}) \qquad \exists x \in \mathbb{K}^n \text{ mit } |x| = 1 \text{ und } |Ax| = \|A\|$$

erfüllt.

Dazu notieren wir die

(3.3.7) **Bemerkungen**.

(i) Die in (3.3.4) definierte Operatornorm ist Matrixnorm und der Norm $| \ |$ im $\mathbb{K}^n$ zugeordnet.

(ii) Zu einer beliebigen Matrixnorm $\| \ \|$ in $M(n \times n, \mathbb{K})$ exisitert eine Norm $| \ |$ im $\mathbb{K}^n$, so daß $\| \ \|$ eine dazu passende Matrixnorm ist.

Zum *Beweis* der Aussage (i) verweisen wir auf (3.2.10) und (3.2.11). Ist $\| \ \|$ beliebige Matrixnorm, so definieren wir für $x \in \mathbb{K}^n$

$$M(x) := x e_1^t = (x \mid 0 \mid \ldots \mid 0) \in M(n \times n, \mathbb{K})$$

und hiermit

$$|x| := \|M(x)\| .$$

Dann gewinnen wir für $x, y \in \mathbb{K}^n$, $\alpha \in \mathbb{K}$ die Beziehungen

$$|x| \geq 0; \quad |x| = 0 \iff \|M(x)\| = 0 \iff M(x) = 0 \iff x = 0;$$
$$|\alpha x| = \|M(\alpha x)\| = \|\alpha M(x)\| = |\alpha| \cdot \|M(x)\| = |\alpha| \cdot |x| ;$$
$$|x + y| = \|M(x + y)\| = \|M(x) + M(y)\| \leq \|M(x)\| + \|M(y)\| = |x| + |y| .$$

Damit sind die Normeigenschaften gezeigt; schließlich gilt für $A \in M(n \times n, \mathbb{K})$

$$|Ax| = \|M(Ax)\| = \|(Ax) e_1^t\| = \|AM(x)\| \leq \|A\| \cdot \|M(x)\| = \|A\| \cdot |x| .$$

(3.3.8) **Beispiele.** Als Matrixnormen, die zu entsprechenden Normen im $\mathbb{K}^n$ passen, notieren wir für $A = (a_{i,j})_{(n,n)}$:

$$\|A\|_1 = \max_{j=1}^{n} \left(\sum_{i=1}^{n} |a_{i,j}| \right) \qquad - \textit{Spaltensummennorm},$$

$$\|A\|_2 = \left(\sum_{i,j=1}^{n} |a_{i,j}|^2 \right)^{\frac{1}{2}} = (\text{spur}(A^*A))^{\frac{1}{2}} \quad - \textit{Frobenius-} \text{ oder } \textit{Schur-Norm},$$

$$\|A\|_\infty = \max_{i=1}^{n} \left(\sum_{j=1}^{n} |a_{i,j}| \right) \qquad - \textit{Zeilensummennorm},$$

$$\|A\|_G = n \cdot \max_{i,j=1}^{n} |a_{i,j}| \qquad - \textit{Gesamtnorm},$$

und mit $w = (w_i)_1^n$, $w_i > 0$:

$$\|A\|_w = \max_{i=1}^{n} \left(\frac{1}{w_i} \sum_{j=1}^{n} |a_{i,j}| w_j \right) \qquad - \textit{w-Norm}.$$

Wir erhalten folgende Eigenschaften:

(3.3.9) **Hilfssatz**.

(i) $\|A\|_1$ *ist die der Norm* $\|x\|_1$ *im* $\mathbb{K}^n$ *zugeordnete Matrixnorm;*

(ii) $\|A\|_2$ *ist zur euklidischen Norm im* $\mathbb{K}^n$ *passende, aber nicht zugeordnete Matrixnorm* $(n \geq 2)$;

(iii) $\|A\|_w$ *ist die der Norm* $\|x\|_w$ *im* $\mathbb{K}^n$, *speziell* $\|A\|_\infty$ *die der Norm* $\|x\|_\infty$ *im* $\mathbb{K}^n$ *zugeordnete Matrixnorm;*

(iv) $\|A\|_G$ *ist Matrixnorm, passend zu den Normen* $\|x\|_p$ $(1 \leq p < \infty)$ *und* $\|x\|_\infty$ *im* $\mathbb{K}^n$, *aber für* $n \geq 2$ *keiner dieser Normen zugeordnet.*

Beweis. Wir zeigen nur die Aussage (iii) und überlassen die restlichen Behauptungen dem Leser als Übungsaufgabe 3.6. Die ∞-Normen sind offensichtlich die mit $w_i = 1$ $(i = 1, \ldots, n)$ spezialisierten w-Normen. Wir zeigen zwei Teilaussagen:

(α) $\forall\, x \in \mathbb{K}^n,\ A \in M(n \times n, \mathbb{K}) \qquad \|Ax\|_w \leq \|A\|_w \|x\|_w$.

Für $A = (a_{i,j})_{(n,n)}$, $x = (\xi_i)_1^n$ gilt nämlich

$$\|Ax\|_w = \max_{i=1}^{n} \left(\frac{1}{w_i} \left| \sum_{j=1}^{n} a_{i,j} \xi_j \right| \right),$$

wobei für $i = 1, \ldots, n$

$$\left| \sum_{j=1}^{n} a_{i,j} \xi_j \right| \leq \sum_{j=1}^{n} |a_{i,j}|\, |\xi_j| = \sum_{j=1}^{n} |a_{i,j}| w_j \cdot \frac{|\xi_j|}{w_j} \leq \|x\|_w \sum_{j=1}^{n} |a_{i,j}| w_j$$

und daher, wie behauptet,

$$\max_{i=1}^{n} \left(\frac{1}{w_i} \sum_{j=1}^{n} |a_{i,j} \xi_j| \right) \leq \|x\|_w \max_{i=1}^{n} \left(\frac{1}{w_i} \sum_{j=1}^{n} |a_{i,j}| w_j \right)$$

erfüllt ist. – Als zweite Aussage beweisen wir

(β) $\forall\, A \in M(n \times n, \mathbb{K}) \quad \exists\, x \in \mathbb{K}^n, \neq 0 \qquad \|Ax\|_w \geq \|A\|_w \|x\|_w$:

Zu vorgegebenem $A \neq 0$ sei $k \in \{1, \ldots, n\}$ ein Zeilenindex mit

$$\frac{1}{w_k} \sum_{j=1}^{n} |a_{k,j}| w_j = \max_{i=1}^{n} \left(\frac{1}{w_i} \sum_{j=1}^{n} |a_{i,j}| w_j \right) = \|A\|_w \qquad (> 0).$$

Wir wählen $x = (\xi_j)_{j=1}^n$ mit

$$\xi_j = \begin{cases} 0, & \text{falls } a_{k,j} = 0 \\ \dfrac{\bar{a}_{k,j}}{|a_{k,j}|} w_j & \text{sonst.} \end{cases}$$

Da mindestens ein $a_{k,j} \neq 0$ existiert, ist $\|x\|_w = 1$. Für die k-te Komponente von $(\eta_i)_{i=1}^n := Ax$ gilt:

$$\eta_k = \sum_{j=1}^n a_{k,j}\,\xi_j = \sum_{\substack{j=1 \\ a_{k,j} \neq 0}}^n a_{k,j} \frac{\bar{a}_{k,j}}{|a_{k,j}|} w_j = \sum_{j=1}^n |a_{k,j}|\, w_j$$

und daher

$$\|Ax\|_w \geq \frac{\eta_k}{w_k} = \frac{1}{w_k} \sum_{j=1}^n |a_{k,j}|\, w_j = \|A\|_w \|x\|_w .$$

Hiernach ist $\|A\|_w$ für alle $A \in M(n \times n, \mathbb{K})$ gleich der Operatornorm von A bezüglich der w-Norm; gemäß (3.3.7, i) ist $\|A\|_w$ Matrixnorm.

Die Frobeniusnorm ist wegen $\|I\|_2 = \sqrt{n} > 1$ $(n \geq 2)$ nur passend zur euklidischen Norm. Ziel der folgenden Überlegungen ist es, die Operatornorm zur euklidischen Norm zu charakterisieren; dabei beschränken wir uns auf den Fall $\mathbb{K} = \mathbb{C}$, fassen also die (n, n)-Matrizen als Abbildungen von $\mathbb{C}^n$ in sich auf.

(3.3.10) **Definition.** Für $A \in M(n \times n, \mathbb{C})$ bezeichne

$$\rho(A) := \max \{|\lambda| : \lambda \in \mathbb{C} \text{ Eigenwert von } A\}$$

als *Spektralradius von* A.

Wir notieren einige naheliegenden Eigenschaften des Spektralradius:

(3.3.11) **Bemerkungen.** Es sei C invertierbare (n,n)-Matrix, $\|\ \|$ beliebige Matrixnorm in $M(n \times n, \mathbb{C})$. Dann gilt für $A \in M(n \times n, \mathbb{C})$

(i) $\rho(A^*) = \rho(A)$,

(ii) $\rho(C^{-1}AC) = \rho(A)$,

(iii) $\rho(A) \leq \|A\|$.

Beweis. Es gilt

$$\det(\lambda I - A) = \det((\lambda I - A)^*) = \det(\bar{\lambda} I - A^*),$$

daher ist λ Eigenwert von A genau dann, wenn $\bar{\lambda}$ Eigenwert von A^* ist, folglich

$$\rho(A^*) = \max \{|\bar{\lambda}| : \lambda \in \mathbb{C} \text{ Eigenwert von } A\} = \rho(A),$$

womit (i) gezeigt ist. Zu (ii) benutzen wir

$$\det(\lambda I - A) = \det(\lambda I - C^{-1}AC),$$

woraus folgt, daß A und $C^{-1}AC$ die gleichen Eigenwerte besitzen.

Zu einer beliebigen Matrixnorm existiert nach (3.3.7, ii) eine Norm $|\ |$ im $\mathbb{C}^n$, so daß für alle $x \in \mathbb{C}^n$ die Abschätzung

$$|Ax| \leqq \|A\| \cdot |x|$$

gilt. Wir wählen nun einen Eigenwert λ von A mit $|\lambda| = \rho(A)$, dazu einen Eigenvektor $x_0 \neq 0$ und erhalten hiermit

$$|Ax_0| = |\lambda x_0| = \rho(A)\,|x_0|\,,$$

daher

$$\rho(A) = \frac{|Ax_0|}{|x_0|} \leqq \|A\|\,.$$

Wir bezeichnen für $A \in M(n \times n, \mathbb{C})$

$\|A\|_S :=$ Operatornorm zu euklidischen Norm im $\mathbb{C}^n$

und notieren dazu den

(3.3.12) **Satz.**

(i) *Ist* A *normal, d.h.* $A^*A = AA^*$, *so gilt*

$$\|A\|_S = \rho(A)\,,$$

(ii) *für beliebiges* A *hat man*

$$\|A\|_S = (\rho(A^*A))^{\frac{1}{2}}\,.$$

Auf Grund dieser Eigenschaften heißt $\|A\|_S$ auch *Spektralnorm.* – Zum *Beweis* der Aussage (i) verweisen wir auf die Übungsaufgabe 3.8, zu Teil (ii) betrachten wir die positiv-semidefinite (hermitesche) Matrix A^*A, die nur reelle, nichtnegative Eigenwerte besitzt, so daß gilt

$$\rho(A^*A) = \max\{\mu \in \mathbb{R}:\ \mu \text{ Eigenwert von } A^*A\}\,.$$

Nach dem Minimum-Maximum-Prinzip, das in Abschnitt 5.7 (Band 2) bewiesen wird, oder auch nach Aufgabe 3.8 ist daher

$$\rho(A^*A) = \max\left\{\frac{(A^*Ax, x)}{(x, x)}:\ x \in \mathbb{C}^n,\ \neq 0\right\}\,,$$

andererseits

$$\max_{x \neq 0} \frac{(A^*Ax, x)}{(x, x)} = \max_{x \neq 0} \frac{(Ax, Ax)}{(x, x)} = \max_{x \neq 0} \left(\frac{\|Ax\|_2}{\|x\|_2}\right)^2 = \|A\|_S^2\,.$$

3.4. Störanfälligkeit linearer Gleichungssysteme

Zu einer vorgegebenen komplexen (n,n)-Matrix A und einer (n,1)-Spalte b soll ein (n,1)-Vektor x mit $Ax = b$ berechnet werden. Durch Fehler in den Eingabe-

daten – und durch Rundungsfehler während der Rechnung, siehe 3.5 bis 3.7 – löst man statt des gegebenen ein abgeändertes Gleichungssystem, nämlich

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

wobei $\tilde{A} = A + \Delta A$, $\tilde{b} = b + \Delta b$ mit im allgemeinen „kleinen" Matrizen ΔA und Δb gilt. Wir bezeichnen wie in (1.2.1) mit

$$\Delta x := \tilde{x} - x$$

den *Fehler von* $\tilde{x}$ *als Näherung von* x und interessieren uns für die Größe des Fehlers in Abhängigkeit von ΔA und Δb.

In den folgenden Überlegungen sei $\|\ \|$ eine Norm im $\mathbb{C}^n$, für $A \in M(n \times n, \mathbb{C})$ sei mit $|A|$ die nach (3.3.4) zugeordnete Matrixnorm bezeichnet. Zunächst erhalten wir den

(3.4.1) **Hilfssatz.** *Es seien* $A, \Delta A \in M(n \times n, \mathbb{C})$, A *invertierbar und* $|A^{-1}\Delta A| < 1$. *Dann ist* $A + \Delta A$ *invertierbar, und es gilt*

$$|(A + \Delta A)^{-1}| \leqq \frac{|A^{-1}|}{1 - |A^{-1}\Delta A|}.$$

Beweis. Offenbar erfüllt $B = A + \Delta A$ die Voraussetzung von (3.2.15) und ist daher invertierbar, außerdem gilt

$$|(A + \Delta A)^{-1}| = |\sum_{n=0}^{\infty} (-A^{-1}\Delta A)^n A^{-1}| \leqq |A^{-1}| \sum_{n=0}^{\infty} |A^{-1}\Delta A|^n = \frac{|A^{-1}|}{1 - |A^{-1}\Delta A|}.$$

(3.4.2) **Satz.** *Es seien* A *und* ΔA *wie in Hilfssatz* (3.4.1), *ferner* $b, \Delta b \in \mathbb{C}^n$ *mit* $b \neq 0$ *vorgegeben. Hierzu bezeichne* x *die Lösung von* $Ax = b$, Δx *sei durch*

$$(3.4.3) \qquad (A + \Delta A)(x + \Delta x) = b + \Delta b$$

definiert. Dann gilt die Abschätzung

$$(3.4.4) \qquad \frac{|\Delta x|}{|x|} \leqq \frac{|A^{-1}|\,|A|}{1 - |A^{-1}\Delta A|} \left(\frac{|\Delta A|}{|A|} + \frac{|\Delta b|}{|b|} \right)$$

und unter der stärkeren Voraussetzung $|A^{-1}|\,|\Delta A| < 1$ *weiter*

$$(3.4.5) \qquad \frac{|\Delta x|}{|x|} \leqq \frac{|A^{-1}|\,|A|}{1 - |A^{-1}|\,|\Delta A|} \left(\frac{|\Delta A|}{|A|} + \frac{|\Delta b|}{|b|} \right).$$

Beweis. Wegen der Invertierbarkeit von $A + \Delta A$ ist Δx in (3.4.3) eindeutig bestimmt; wir haben

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

und wegen $x = A^{-1} b$ weiter

$$\begin{aligned} \Delta x &= [(A + \Delta A)^{-1} - A^{-1}]\, b + (A + \Delta A)^{-1} \Delta b \\ &= (A + \Delta A)^{-1} [A - (A + \Delta A)] A^{-1} b + (A + \Delta A)^{-1} \Delta b \\ &= (A + \Delta A)^{-1} (-\Delta A) x + (A + \Delta A)^{-1} b \end{aligned}$$

und daher nach Hilfssatz (3.4.1)

$$|\Delta x| \leqq |\Delta A| \frac{|A^{-1}|}{1-|A^{-1}\Delta A|}|x| + \frac{|A^{-1}|}{1-|A^{-1}\Delta A|}|\Delta b| .$$

Da mit $b \neq 0$ auch $x \neq 0$ ist, können wir durch $|x|$ dividieren. Ferner benutzen wir die aus $Ax = b$, also $|b| \leqq |A| \cdot |x|$ folgende Abschätzung

$$\frac{1}{|x|} \leqq \frac{|A|}{|b|} ,$$

die unmittelbar (3.4.4) liefert.

Ist zusätzlich $|A^{-1}|\,|\Delta A| < 1$, erhalten wir (3.4.5) mittels der aus $|A^{-1}\Delta A| \leqq |A^{-1}|\,|\Delta A|$ folgenden Ungleichung

$$\frac{1}{1-|A^{-1}\Delta A|} \leqq \frac{1}{1-|A^{-1}|\,|\Delta A|} .$$

Wenn die Koeffizienten von ΔA so klein sind, daß $1 - |A^{-1}\Delta A| \approx 1$ beträgt, wird die Abhängigkeit des „relativen Fehlers" der Lösung von den Eingabefehlern im wesentlichen durch den Faktor $|A^{-1}| \cdot |A|$ beschrieben. Hiermit kommen wir zu der

(3.4.6) **Definition.** Für eine invertierbare (n,n)-Matrix A heißt

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

die *Konditionszahl* von A bezüglich der Norm $|\ |$.

Mit Hilfe der Konditionszahl können wir in (3.4.5)

$$1 - |A^{-1}|\,|\Delta A| = 1 - \kappa(A)\frac{|\Delta A|}{|A|}$$

umformen und erhalten an Stelle von (3.4.5) die Abschätzung

$$(3.4.7) \qquad \frac{|\Delta x|}{|x|} \leqq \frac{\kappa(A)}{1-\kappa(A)\frac{|\Delta A|}{|A|}}\left(\frac{|\Delta A|}{|A|} + \frac{|\Delta b|}{|b|}\right) \approx \kappa(A)\left(\frac{|\Delta A|}{|A|} + \frac{|\Delta b|}{|b|}\right).$$

Wir nennen ein Gleichungssystem *schlecht konditioniert*, wenn $\kappa(A)$ groß ist. $\kappa(A)$ hängt von der betrachteten Norm ab, wir können jedoch eine normunabhängige untere Schranke angeben:

(3.4.8) **Bemerkung.** Für eine invertierbare Matrix A gilt stets

$$\kappa(A) \geqq \frac{\rho(A)}{\min\{|\lambda| : \lambda \text{ Eigenwert von } A\}} \geqq 1 .$$

Beweis. Wir folgern aus (3.3.11, iii) unmittelbar

$$\kappa(A) = |A|\,|A^{-1}| \geqq \rho(A)\,\rho(A^{-1})$$

und beachten die Äquivalenz

$$\lambda \text{ Eigenwert von } A \iff \frac{1}{\lambda} \text{ Eigenwert von } A^{-1}.$$

Hieraus ergibt sich

$$\rho(A^{-1}) = (\min\{|\lambda| : \lambda \text{ Eigenwert von } A\})^{-1}.$$

Ein Gleichungssystem ist also sicher dann schlecht konditioniert, wenn die Eigenwerte von A betragsmäßig weit streuen.

Es entsteht das Problem, ein gegebenes Gleichungssystem $Ax = b$ vor seiner Lösung so umzuformen, daß $\kappa(A)$ möglichst klein wird. Allgemein wird empfohlen, das System durch Multiplikation von Zeilen und Spalten (letzteres bei Variablensubstitution) in ein *äquilibriertes* umzuformen, in dem die Koeffizienten von A betraglich etwa im Bereich der gleichen Zehnerpotenz liegen. Für diese totale Äquilibrierung gibt es bisher kein Computer-Programm (man löst dieses Problem durch genaues „Hinsehen", vgl. Aufgabe 2.2) und auch keine brauchbaren Aussagen über die Konditionszahlen. Dagegen wird eine *Zeilenäquilibrierung*, d.h. daß die Betragssummen aller Zeilen gleich sind, durch Multiplikation der Zeilen mit leicht anzugebenden Skalaren erreicht. Dem entspricht der Übergang von $Ax = b$ auf $(DA)x = Db$ mit einer invertierbaren Diagonalmatrix D. Der folgende Satz zeigt, daß eine zeilenäquilibrierte Matrix gegenüber den nicht äquilibrierten Matrizen die günstigste Konditionszahl bezüglich der Maximumsnorm $\|\ \|_\infty$ (vgl. (3.3.8)!) besitzt.

(3.4.8) **Satz.** *Es sei* $A = (a_{i,j})_{(n,n)}$ *invertierbare, zeilenäquilibrierte* (n,n)*-Matrix, es gelte*

$$\sum_{j=1}^{n} |a_{i,j}| = 1 \qquad (i = 1, \ldots, n).$$

D *sei invertierbare Diagonalmatrix. Dann gilt für die Konditionszahlen bezüglich der Maximumsnorm:*

$$\kappa_\infty(A) \leq \kappa_\infty(DA).$$

Beweis. Mit $D = \operatorname{diag}(d_1, \ldots, d_n)$ berechnen wir

$$\|DA\|_\infty = \max_{i=1}^{n}\left(\sum_{j=1}^{n} |d_i a_{i,j}|\right) = \max_{i=1}^{n}\left(|d_i| \cdot \sum_{j=1}^{n} |a_{i,j}|\right) = \max_{i=1}^{n} |d_i|.$$

Setzt man $A^{-1} =: (\tilde{a}_{i,j})_{(n,n)}$, so wird

$$\|(DA)^{-1}\|_\infty = \|A^{-1}D^{-1}\|_\infty = \max_{i=1}^{n}\left(\sum_{j=1}^{n} \left|\tilde{a}_{i,j} \cdot \frac{1}{d_j}\right|\right).$$

Nun gilt für jedes $i \in \{1, \ldots, n\}$

$$\sum_{j=1}^{n} |\tilde{a}_{i,j}| \frac{1}{|d_j|} \geqq \left(\sum_{j=1}^{n} |\tilde{a}_{i,j}| \right) \min_{j=1}^{n} \frac{1}{|d_j|},$$

daher auch

$$\max_{i=1}^{n} \left(\sum_{j=1}^{n} \left| \frac{\tilde{a}_{i,j}}{d_j} \right| \right) \geqq \min_{j=1}^{n} \left(\frac{1}{|d_j|} \right) \cdot \max_{i=1}^{n} \left(\sum_{j=1}^{n} |\tilde{a}_{i,j}| \right) = \frac{1}{\max\limits_{j=1}^{n} |d_j|} \cdot \|A^{-1}\|_\infty$$

Insgesamt wird wegen $\|A\|_\infty = 1$:

$$\kappa_\infty(DA) = \|DA\|_\infty \, \|(DA)^{-1}\|_\infty \geqq \|A^{-1}\| = \kappa_\infty(A) .$$

Umgekehrt ist nicht jede äquilibrierte Matrix gut konditioniert, dazu notieren wir das folgende

(3.4.9) **Beispiel.** Es seien $x_0, \ldots, x_n \in \mathbb{R}$, verschieden, dazu $f_0, \ldots, f_n \in \mathbb{C}$ beliebig. Gesucht ist dasjenige Polynom höchstens n-ten Grades

$$(*) \quad P(x) = a_0 + a_1 x + \ldots + a_n x^n$$

mit $P(x_i) = f_i$ $(i = 0, \ldots, n)$. – Diese *Interpolationsaufgabe* wollen wir im Band 2 ausführlich behandeln. Die Koeffizienten des gesuchten Polynoms erfüllen das Gleichungssystem

$$(**) \quad \sum_{j=0}^{n} x_i^j a_j = f_i \qquad (i = 0, \ldots, n)$$

mit der *Vandermonde*-Matrix

$$A = V_{n+1}(x_0, \ldots, x_n) = (x_i^j)_{0 \leqq i,j \leqq n}$$

als Koeffizientenmatrix. Die Determinante dieser Matrix ist bei verschiedenen $x_0, \ldots, x_n$ ungleich Null, folglich gibt es genau eine Lösung $(a_i)_0^n$ von (**).

Für die Werte

$$(***) \quad n = 5; \quad x_\nu = 1{,}1 + \nu \cdot 0{,}1 \qquad (\nu = 0, \ldots, 5),$$

die wir als Beispiel heranziehen wollen, ist A nahezu äquilibriert. Es ergibt sich jedoch

$$\kappa_\infty(A) \approx 0{,}19 \cdot 10^8 ;$$

das bedeutet, daß eine relative Störung der f_i um den Betrag $\epsilon\,(>0)$ eine relative Störung von etwa $2 \cdot 10^7 \cdot \epsilon$ in den a_j erwarten läßt.

Als Konsequenz ergibt sich:

Trotz ihrer theoretischen Bedeutung ist wegen ihrer allgemein schlechten Kondition die Vandermonde-Matrix für numerische Zwecke ungeeignet. Die Fehler in den Eingabedaten und, wie die folgenden Abschnitte zeigen, auch die Rundungsfehler während der Lösung des Gleichungssystems bewirken eine verhältnismäßig große Ungenauigkeit im Ergebnis.

Speziell bei den in (***) angegebenen Stützstellen hängen die Polynomkoeffizienten a_j sehr empfindlich von den f_i ab. Allgemein sucht man daher für das Interpolationspolynom eine andere, dem Problem besser angepaßte Darstellung als (*).

Wie aus dem 1. Kapitel bekannt ist, treten beim Einlesen eines Gleichungssystems in eine Rechenmaschine Rundungsfehler auf, die von der benutzten Arithmetik abhängen. Den Einfluß dieser Rundungsfehler auf das Ergebnis diskutieren wir im folgenden

(3.4.10) **Zahlenbeispiel.** Die Koeffizienten des Gleichungssystems im Beispiel (2.3.3) lassen sich nicht als endliche Hexadezimalzahlen schreiben. Wir lösen dieses Gleichungssystem auf der in (1.1.16) angegebenen Maschine mit doppeltgenauer Arithmetik, wobei wir die Koeffizientenmatrix einmal doppeltgenau (REAL*8), beim zweiten Mal einfachgenau (REAL*4), die rechte Seite stets doppeltgenau einlesen. Wir bezeichnen mit (A, b) die erweiterte Matrix aus den entsprechenden REAL*8-Größen, also auf 14 Hexadezimalziffern gerundet. Dann lösen wir beim ersten Mal $Ax = b$, beim zweiten Mal ein Gleichungssystem, in dem A auf 6 Hexadezimalziffern gerundet ist, also

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

mit

$$(*) \qquad \frac{\|\Delta A\|_\infty}{\|A\|_\infty} \leqq \tau = \tfrac{1}{2} \cdot 16^{-5} \approx \tfrac{1}{2} \cdot 10^{-6}.$$

Die Berechnung von A^{-1} liefert

$$(**) \qquad \|A^{-1}\|_\infty = 14{,}1; \qquad \kappa_\infty(A) = 2{,}6 \cdot 10^2.$$

Wir erhalten, auf 10 Dezimalstellen gerundet, folgende Lösungen

(i) $x_1 = -20{,}76272418;\ x_2 = -2{,}747919203;\ x_3 = 14{,}74503384;$
$x_4 = 2{,}615863138;$

und beim zweiten Mal $\tilde{x} = (\tilde{x}_i)_1^4$ mit

(ii) $\tilde{x}_1 = -20{,}76267022;\ \tilde{x}_2 = -2{,}747910721;\ \tilde{x}_3 = 14{,}74499749;$
$\tilde{x}_4 = 2{,}615856811.$

Die Koeffizienten von $\tilde{x}$ stimmen in 5 Dezimalstellen mit denen von x überein; wir berechnen für $\Delta x := \tilde{x} - x$:

$$(***) \qquad \frac{\|\Delta x\|_\infty}{\|x\|_\infty} = 0{,}26 \cdot 10^{-5} > 5\tau\,,$$

andererseits liefert die Fehlerabschätzung (3.4.7) nach (*), (**)

$$\frac{\|\Delta x\|_\infty}{\|x\|_\infty} \lesssim \kappa_\infty(A)\,\frac{\|\Delta A\|_\infty}{\|A\|_\infty} \leqq \tau\cdot\kappa_\infty(A) = 0{,}13\cdot 10^{-3}.$$

Hiermit ist der tatsächliche Fehler zwar stark überschätzt, jedoch bewirkt nach (***) die Störanfälligkeit von A ein Anwachsen des „relativen Fehlers" von x gegenüber den Eingabefehlern um mehr als den Faktor 5. Wegen der doppeltgenauen Arithmetik (ca. 16 Dezimalstellen) haben die Rundungsfehler bei der benutzten Gauß-Elimination auf die ersten 10 Dezimalstellen der Lösung keinen Einfluß; man vergleiche dazu den Abschnitt 3.6!

3.5. Rundungsfehler bei Gleichungssystemen in Dreiecksgestalt

Es soll die Lösung eines reellen linearen Gleichungssystems $Ax = b$ unter Benutzung von t-stelliger Gleitkommarechnung, wie sie im 1. Kapitel beschrieben wurde, ermittelt werden. Als Basis des Zahlsystems sei $g = 10$ gewählt, die Mantissenlänge sei $t \geqq 2$, und es bezeichne

$$\tau := \tfrac{1}{2}\cdot 10^{-t+1}.$$

Nach (1.3.2) gilt für je zwei t-stellige Gleitkommazahlen x, y und jede der Rechenoperationen $\triangle \in \{+, -, \cdot, :\}$:

$$gl(x \triangle y) = (x \triangle y)(1+\epsilon) = \frac{x \triangle y}{1+\eta}$$

mit gewissen, von x, y und $\triangle$ abhängigen reellen Größen ϵ und η, wobei

$$|\epsilon|, |\eta| \leqq \tau$$

gilt. Die mit Gleitkommarechnung ermittelte Lösung eines linearen Gleichungssystems wird in den folgenden Abschnitten als Lösung eines Systems mit gestörten Koeffizienten dargestellt, so daß anschließend die Überlegungen von 3.4 zum Tragen kommen. Da mit 3.4 auch die Rundungsfehler bei der Eingabe behandelt sind, beschränken wir uns auf t-stellige Gleitkommazahlen als Matrixkoeffizienten.

Wir betrachten zunächst Koeffizientenmatrizen in Dreiecksgestalt, wegen der einfacheren Notation speziell untere Dreiecksmatrizen und notieren hierzu die

(3.5.1) **Voraussetzung.** Vorgegeben sei ein Gleichungssystem $Lx = b$,

$$L = \begin{pmatrix} l_{1,1} & & 0 \\ \vdots & \ddots & \\ l_{n,1} & \cdots & l_{n,n} \end{pmatrix}, \qquad b = \begin{pmatrix} b_1 \\ \vdots \\ b_n \end{pmatrix}$$

mit reellen t-stelligen Gleitkommazahlen $l_{i,j}$, b_i und $l_{i,i} \neq 0$ $(i = 1, \ldots, n)$. Für t und damit für die Rechengenauigkeit in Beziehung zu n fordern wir

$$n\tau \leqq 0{,}09;$$

es bezeichne $x = (x_i)_1^n$ den exakten, $\tilde{x} = (\tilde{x}_i)_1^n$ den numerisch berechneten Lösungsvektor des Gleichungssystems.

Unter dieser Annahme gelten für x die Beziehungen

$$x_1 = \frac{b_1}{l_{1,1}}; \quad x_r = \frac{b_r - l_{r,1}x_1 - \ldots - l_{r,r-1}x_{r-1}}{l_{r,r}} \qquad (r = 2, \ldots, n).$$

Bei t-stelliger Gleitkommarechnung erhalten wir statt dessen

$$\tilde{x}_1 = gl\left(\frac{b_1}{l_{1,1}}\right)$$

$$\tilde{x}_r = gl\left(\frac{b_r - l_{r,1}\tilde{x}_1 - \ldots - l_{r,r-1}\tilde{x}_{r-1}}{l_{r,r}}\right) \qquad (r = 2, \ldots, n),$$

wobei wir den Zähler von $\tilde{x}_r$ in der natürlichen, unten genau gekennzeichneten Reihenfolge berechnen wollen. Jedes $\tilde{x}_r$ ist also die numerisch ermittelte Lösung einer Gleichung

$$\sum_{i=1}^{r} l_{r,i}\tilde{x}_i = b_r$$

bei schon bekannten $\tilde{x}_1, \ldots, \tilde{x}_{r-1}$. Wir beweisen hierzu den

(3.5.2) **Hilfssatz.** *Es sei* $1 \leqq r \leqq n$, $n\tau \leqq 0{,}09$; $b, l_1, \ldots, l_r$ *sowie* $\tilde{x}_1, \ldots, \tilde{x}_{r-1}$ *seien t-stellige Gleitkommazahlen, und* $l_r \neq 0$. *Dann gilt mit* $\tilde{x}_r$, *definiert durch*

$$\tilde{x}_r = gl\left(\frac{b - l_1\tilde{x}_1 - \ldots - l_{r-1}\tilde{x}_{r-1}}{l_r}\right)$$

eine Darstellung

$$\text{(i)} \quad b = \sum_{j=1}^{n} l_j\tilde{x}_j(1 + F_j),$$

wobei die $F_j \in \mathbb{R}$ *den Ungleichungen*

$$|F_j| \leqq \frac{1}{1 - n\tau} j\tau \leqq 1{,}1 \cdot j\tau \qquad (j = 1, \ldots, r)$$

genügen, im Fall $l_r = 1$ *schärfer die Ungleichung*

$$|F_r| \leqq \frac{1}{1 - n\tau}(r - 1)\tau \leqq 1{,}1(r - 1)\tau$$

erfüllt ist. Es folgt die Abschätzung

$$\text{(ii)} \quad \left|b - \sum_{j=1}^{r} l_j\tilde{x}_j\right| \leqq \frac{\tau}{1 - n\tau} \sum_{j=1}^{r} j\,|l_j|\,|\tilde{x}_j|,$$

und im Fall $l_r = 1$

(iii) $$\left| b - \sum_{j=1}^{r} l_j \tilde{x}_j \right| \leqq \frac{\tau}{1 - n\tau} \left(\sum_{j=1}^{r} j \, |l_j| \, |\tilde{x}_j| - |\tilde{x}_r| \right)$$

Zum *Beweis* verfolgen wir die Zwischenergebnisse bei der Berechnung von $\tilde{x}_r$. Dazu setzen wir

$$s_0 = b$$

und im Fall $r \geqq 2$:

$$s_\nu = gl\,(s_{\nu-1} - l_\nu \tilde{x}_\nu) \qquad (\nu = 1, \ldots, r-1)$$

und erhalten hiermit

$$\tilde{x}_r = gl\left(\frac{s_{r-1}}{l_r}\right).$$

Folglich gelten mit gewissen $\epsilon_\nu, \eta_\nu \in \mathbb{R}$, $|\epsilon_\nu|, |\eta_\nu| \leqq \tau$ die Gleichungen

$$s_\nu = [s_{\nu-1} - l_\nu \tilde{x}_\nu (1 + \epsilon_\nu)] \frac{1}{1 + \eta_\nu} \qquad (\nu = 1, \ldots, r-1),$$

$$\tilde{x}_r = \frac{s_{r-1}}{l_r} \cdot \frac{1}{1 + \eta_r},$$

wobei im Fall $l_r = 1$ offenbar $\eta_r = 0$ zu wählen ist. Durch Induktion über r zeigt man für $r \geqq 2$ die Darstellung

$$s_{r-1} = \prod_{\nu=1}^{r-1} (1 + \eta_\nu)^{-1} b - \sum_{j=1}^{r-1} l_j \tilde{x}_j (1 + \epsilon_j) \prod_{\nu=j}^{r-1} (1 + \eta_\nu)^{-1}$$

und erhält in jedem Fall

$$l_r \tilde{x}_r = \frac{s_{r-1}}{1 + \eta_r} = \prod_{\nu=1}^{r} (1 + \eta_\nu)^{-1} b - \sum_{j=1}^{r-1} l_j \tilde{x}_j (1 + \epsilon_j) \prod_{\nu=j}^{r} (1 + \eta_\nu)^{-1}.$$

Dabei ist die leere Summe bekanntlich Null, das leere Produkt Eins zu setzen. Multiplikation mit dem Produkt der $(1 + \eta_\nu)$ und Auflösen nach b liefert

$$b = l_r \tilde{x}_r \prod_{\nu=1}^{r} (1 + \eta_\nu) + \sum_{j=1}^{r-1} l_j \tilde{x}_j (1 + \epsilon_j) \prod_{\nu=1}^{j-1} (1 + \eta_\nu).$$

Demnach definieren wir

$$1 + F_r := \prod_{\nu=1}^{r} (1 + \eta_\nu), \quad 1 + F_j := (1 + \epsilon_j) \prod_{\nu=1}^{j-1} (1 + \eta_\nu) \qquad (j = 1, \ldots, r-1)$$

und erhalten unmittelbar

$$(1-\tau)^j \leqq 1 + F_j \leqq (1+\tau)^j \qquad (j = 1, \ldots, r),$$

im Fall $l_r = 1$ zusätzlich

$$(1-\tau)^{r-1} \leqq 1 + F_r \leqq (1+\tau)^{r-1}.$$

Hiermit ist die Darstellung (i) gezeigt, und Hilfssatz (1.4.10) liefert wegen

$$\frac{j}{1-(j-1)\tau} \leqq \frac{j}{1-n\tau}$$

auch die Abschätzungen der $|F_j|$. Aus (i) folgern wir

$$(3.5.3) \qquad b = \sum_{j=1}^{r} l_j \tilde{x}_j + \sum_{j=1}^{r} l_j \tilde{x}_j F_j$$

und damit

$$|b - \sum_{j=1}^{r} l_j \tilde{x}_j| \leqq \sum_{j=1}^{r} |l_j|\,|\tilde{x}_j|\,|F_j|.$$

Nach Abschätzung der $|F_j|$ erhalten wir unmittelbar (ii) bzw. (iii).

Die Ungleichungen (ii) und (iii) des Hilfssatzes werden wir erst im Abschnitt 3.6 brauchen; die Anwendung von (3.5.2, i) auf das in (3.5.1) vorgegebene Gleichungssystem liefert für die numerisch berechnete Lösung x die folgenden Beziehungen:

$$(3.5.4) \qquad \begin{cases} l_{1,1}(1+F_{1,1})\tilde{x}_1 & = b_1 \\ l_{r,1}(1+F_{r,1})\tilde{x}_1 + \ldots + l_{r,r}(1+F_{r,r})\tilde{x}_r & = b_r \qquad (r = 2, \ldots, n) \end{cases}$$

mit

$$|F_{r,j}| \leqq \frac{j}{1-n\tau}\tau \;\; (j = 1, \ldots, r); \quad |F_{r,r}| \leqq \frac{(r-1)}{1-n\tau}\tau, \quad \text{falls } l_{r,r} = 1.$$

Zur komponentenweise Abschätzung vereinbaren wir folgende

(3.5.5) **Bezeichnungen.** Für eine reelle oder komplexe (n,m)-Matrix $A = (a_{i,j})_{(n,m)}$ sei

$$\hat{A} := (|a_{i,j}|)_{(n,m)};$$

für reelle (n,m)-Matrizen $A = (a_{i,j})_{(n,m)}$ und $B = (b_{i,j})_{(n,m)}$ schreiben wir $A \leqq B$ oder auch $B \geqq A$ genau dann, wenn $a_{i,j} \leqq b_{i,j}$ für alle $i = 1, \ldots, n$; $j = 1, \ldots, m$ gilt; entsprechend bedeute $A < B$ oder auch $B > A$, daß $a_{i,j} < b_{i,j}$ für alle i und j erfüllt ist.

Die genannten Schreibweisen verwenden wir speziell auch für (n,1)-Matrizen, also Spalten des $\mathbb{C}^n$ bzw. $\mathbb{R}^n$.

Wir folgern aus (3.5.4) unmittelbar

(3.5.6) **Satz.** *Unter der Voraussetzung* (3.5.1) *ist der numerisch ermittelte Lösungsvektor* $\tilde{x} = (\tilde{x}_i)_1^n$ *die exakte Lösung eines gestörten Gleichungssystems*

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

in dem ΔL *die folgende Gestalt hat:*

$$\Delta L = \begin{pmatrix} l_{1,1}F_{1,1} & & 0 \\ \vdots & \ddots & \\ l_{n,1}F_{n,1} & \cdots & l_{n,n}F_{n,n} \end{pmatrix}$$

Mit der Bezeichnung

$$\delta = \begin{cases} 1, & \textit{falls } L \textit{ normierte untere Dreiecksmatrix ist,} \\ 0 & \textit{sonst}\,, \end{cases}$$

gilt die komponentenweise Abschätzung

$$\hat{\Delta L} \leqq \frac{\tau}{1 - n\tau} \begin{pmatrix} (1-\delta)\,|l_{1,1}| & & & 0 \\ |l_{2,1}| & (2-\delta)\,|l_{2,2}| & & \\ \vdots & 2\,|l_{3,2}| & \ddots & \\ \vdots & \vdots & & \\ |l_{n,1}| & 2\,|l_{n,2}| & \cdots & (n-\delta)\,|l_{n,n}| \end{pmatrix},$$

also mit $D := \operatorname{diag}(1, \ldots, n)$

$$\hat{\Delta L} \leqq \frac{\tau}{1 - n\tau}\,(\hat{L}D - \delta I) \leqq 1{,}1\,\tau\,(\hat{L}D - \delta I)\,.$$

Für eine obere Dreiecksmatrix erhalten wir als

(3.5.7) **Folgerung.** Ist $R = (r_{i,j})_{(n,n)}$ mit $r_{i,i} \neq 0$ obere Dreiecksmatrix aus t-stelligen Gleitkommazahlen, b wie in (3.5.1) und $n\tau \leqq 0{,}09$, so erfüllt die numerisch ermittelte Lösung $\tilde{x}$ von $Rx = b$ ein Gleichungssystem

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

in dem mit $E := \operatorname{diag}(n, n-1, \ldots, 1)$ die komponentenweise Abschätzung

$$\hat{\Delta R} \leqq \frac{\tau}{1 - n\tau}\,\hat{R}E \leqq 1{,}1\,\tau\,\hat{R}E$$

gilt.

Damit ist das Problem der Rundungsfehler bei Dreiecksmatrizen auf die im Abschnitt 3.4 behandelte Aufgabenstellung zurückgeführt. Durch Vergröberung gewinnen wir aus (3.5.6) die Abschätzung

(3.5.8) $$\hat{\Delta L} \leqq 1{,}1\,(n-\delta)\,\tau\,\hat{L} \leqq 1{,}1\,n\tau\,\hat{L}\,;$$

daher ist der Effekt der Rundungsfehler bei Berechnung von $\tilde{x}$ der gleiche, als hätte man die Koeffizienten von L mit relativen Fehlern vom Betrag $\leqq 1{,}1 \cdot n\tau$ versehen. Sind die Koeffizienten von L beispielsweise als Meßdaten mit relativen Fehlerschranken größer als $n\tau$ bekannt, spielen die Rundungsfehler demgegenüber keine Rolle mehr. In diesem Sinne ist die numerische Auflösung des Gleichungssystems $Lx = b$ ein gutartiges Verfahren. Wir notieren Abschätzungen für die wichtigsten Matrixnormen von ΔL:

(3.5.9) **Folgerung.**

(i) Bezeichnet $\| \ \|$ eine der Normen $\| \ \|_1$, $\| \ \|_2$ oder $\| \ \|_\infty$, so gilt

$$\|\Delta L\| \leqq 1{,}1\, \tau(n-\delta)\, \|L\|,$$

(ii) Ist g eine positive Konstante mit $|l_{i,j}| \leqq g$ $(i, j = 1, \ldots, n)$, so hat man

$$\|\Delta L\|_\infty \leqq \frac{1}{2}(n^2 + n - 2\delta) g \cdot \frac{\tau}{1 - n\tau},$$

$$\|\Delta L\|_1 \leqq \frac{1}{4}(n + 1 - \delta)^2 g \cdot \frac{\tau}{1 - n\tau},$$

$$\|\Delta L\|_2 \leqq \frac{1}{\sqrt{12}}(n+1)^2 g \frac{\tau}{1 - n\tau}.$$

Beweis. Die Aussage (i) schließen wir aus der Ungleichung (3.5.8), die sich auf die angegebenen Normen überträgt; zu (ii) benutzen wir mit

$$G = \begin{pmatrix} 1-\delta & & & \\ 1 & 2-\delta & & 0 \\ 1 & 2 & \ddots & \\ \vdots & \vdots & & \\ 1 & 2 & \ldots\ldots & n-\delta \end{pmatrix}_{(n,n)}$$

die komponentenweise Abschätzung

$$\hat{\Delta L} \leqq \frac{\tau}{1 - n\tau} g \cdot G$$

sowie

$$\|G\|_\infty = \sum_{\nu=1}^{n} \nu - \delta = \frac{1}{2}(n^2 + n - 2\delta),$$

$$\|G\|_1 = \max_{\kappa=1}^{n} \kappa(n - \kappa + 1 - \delta) \leqq \left(\frac{n+1-\delta}{2}\right)^2,$$

$$\|G\|_2^2 = \sum_{r=1}^{n} r^2(n - r + 1) - \delta n^2 \leqq (n+1)\frac{(n+1)\,n(2n+1)}{6} - \frac{n^2(n+1)^2}{4} \leqq \frac{1}{12}(n+1)^4.$$

Als Fehlerabschätzung für die numerisch berechnete Lösung $\tilde{x}$ von $Lx = b$ erhalten wir nach Satz (3.4.2) bzw. (3.4.7) und (3.5.9, i) bezüglich der Maximumsnorm:

(3.5.10) **Folgerung.** Im Fall $1{,}1\,\tau(n-\delta)\kappa_\infty(L) < 1$, $b \neq 0$ genügt der Fehler von $\tilde{x}$ als Näherung von x der Ungleichung

$$\frac{\|\tilde{x}-x\|_\infty}{\|x\|_\infty} \leqq \frac{\kappa_\infty(L)}{1-1{,}1\,\tau(n-\delta)\kappa_\infty(L)}\,1{,}1(n-\delta)\tau \approx n\tau\kappa_\infty(L)\,.$$

3.6. Rundungsfehler beim Gaußschen Eliminationsverfahren

Wir betrachten die Rundungsfehler, die bei der Zerlegung einer Matrix $PAQ = LR$ nach der Gauß-Elimination entstehen. Da für eine (n,l)-Matrix B bei der simultanen Zerlegung $PB = LS$ dieselben Rechenschritte auftreten wie bei der expliziten Auflösung des Gleichungssystems für S, wollen wir B zunächst weglassen, also $l = 0$ setzen. Wir notieren als

(3.6.1) **Voraussetzung.** Es sei $k \geqq n$, A eine (n,k)-Matrix aus t-stelligen reellen Gleitkommazahlen. Die Gauß-Elimination nach Satz (2.2.6) werde mit t-stelliger Gleitkommarechnung bei halbmaximaler bzw. maximaler Pivotwahl durchgeführt, das Verfahren ende nach $(n-1)$ Schritten; für die benutzte Rechengenauigkeit gelte

$$n\tau \leqq 0{,}09\,.$$

Die Fälle, in denen das Verfahren vorher abbricht oder $k < n$ ist, sind von geringer praktischer Bedeutung und mit einigen Zusatzüberlegungen zu erledigen.

Bei halbmaximaler oder maximaler Pivotwahl wird die Dreieckszerlegung einer permutierten Matrix PAQ berechnet, wobei P der Zeilenpermutation $\pi = \pi_n$, Q der Spaltenpermutation $\sigma = \sigma_n$ gemäß den Algorithmen (2.2.15) oder (2.2.16) entspricht. Wir beachten, daß wegen $\{\pi_{\nu+1}(\nu), \ldots, \pi_{\nu+1}(n)\} = \{\pi_n(\nu), \ldots, \pi_n(n)\}$, $\{\sigma_{\nu+1}(\nu), \ldots, \sigma_{\nu+1}(k)\} = \{\sigma_n(\nu), \ldots, \sigma_n(k)\}$ die Beziehungen (2.2.15, v–viii) gültig bleiben, wenn man $\pi_{\nu+1}$ und $\sigma_{\nu+1}$ durch π_n bzw. σ_n ersetzt. Das bedeutet, daß man von vornherein PAQ statt A betrachten, also $\pi = \mathrm{id}$, $\sigma = \mathrm{id}$ wählen kann. Daher nehmen wir ohne Einschränkung der Allgemeinheit an:

(3.6.2) **Voraussetzung.** Bei jedem Eliminationsschritt sei

(i) die diagonale Pivotwahl gleich der maximalen Pivotwahl, bzw.

(ii) die diagonale Pivotwahl gleich der halbmaximalen Pivotwahl.

Wir verwenden die

(3.6.3) **Bezeichnungen**

(i) Es sei $A^{(1)} = (a_{i,j}^{(1)})_{(n,k)} := A$ sowie für $\nu = 2, \ldots, n$

$$A^{(\nu)} = (a_{i,j}^{(\nu)})_{(n,k)}$$

die bei t-stelliger Gleitkommarechnung nach $(\nu-1)$ Eliminationsschritten entstandenen Matrizen; wir setzen

(ii) $$U := A^{(n)} = \begin{pmatrix} a_{1,1}^{(1)} & \cdots & \cdots & a_{1,k}^{(1)} \\ & a_{2,2}^{(2)} & & \vdots \\ 0 & & \ddots & \vdots \\ & & a_{n,n}^{(n)} & a_{n,k}^{(n)} \end{pmatrix}$$

und definieren $M = (m_{i,j})_{(n,n)}$ durch

(iii) $$m_{i,j} := \begin{cases} 0 \text{ für } i<j, \\ 1 \text{ für } i=j, \\ gl\left(\dfrac{a_{i,j}^{(j)}}{a_{j,j}^{(j)}}\right) \end{cases} \quad \text{für } 1 \leqq j \leqq n-1,\ j+1 \leqq i \leqq n.$$

Bei halbmaximaler (erst recht bei maximaler) Pivotwahl gilt dann die

(3.6.4) **Bemerkung.** M ist normierte untere Dreiecksmatrix mit

$$|m_{i,j}| \leqq 1 \qquad (1 \leqq i \leqq j \leqq n).$$

U und M ersetzen offenbar die in (2.2.6) erwähnten Matrizen R und L, – man vergleiche dazu (2.2.24) und (2.2.20). An Stelle der Gleichung LR = A gilt für die numerisch gewonnenen Matrizen nur MU = A + F mit einer gewissen (n,k)-Matrix F, für deren Komponenten wir im Folgenden zwei verschiedene Abschätzungen herleiten wollen. Wir benutzen die

(3.6.5) **Bezeichnung.** $g := \max\{|a_{i,j}^{(\nu)}|:\ 1 \leqq \nu \leqq n,\ 1 \leqq i \leqq n,\ 1 \leqq j \leqq k\}$

(3.6.6) **Satz.** *Unter den Annahmen* (3.6.1) *und* (3.6.2) *liefert die Gauß-Elimination bei t-stelliger Gleitkommarechnung Zerlegungsmatrizen* M *und* U *mit*

$$MU = A + F,$$

wobei $\hat{F}$ – *vgl.* (3.5.5) – *der Ungleichung*

(i) $$\hat{F} \leqq 2\tau g \begin{pmatrix} 0 & \cdots & \cdots & 0 \\ 1 & \cdots & \cdots & 1 \\ 1 & 2 & \cdots & 2 \\ \vdots & \vdots & \ddots & \vdots \\ 1 & 2 & \cdots & (n-1)\ \ldots\ (n-1) \end{pmatrix}_{(n,k)}$$

sowie mit $D := \operatorname{diag}(1, 2, \ldots, n)$ *der Abschätzung*

(ii) $$\hat{F} \leqq \frac{\tau}{1-n\tau}\,(\hat{M}D\hat{U} - \hat{U}).$$

genügt.

Beweis. Wir betrachten im Algorithmus (2.2.15), den wir mit $\pi = \mathrm{id}$, $\sigma = \mathrm{id}$ anwenden, zunächst die Rechenoperationen, die einen festen Koeffizienten $a_{i,j}^{(i)}$ $(i \leqq j)$ von U liefern. Dabei liegen die $a_{1,j}^{(1)} = a_{1,j}$ bereits vor; im Fall $i \geqq 2$ werden bei den Eliminationsschritten $\nu = 1, \ldots, i-1$ die Werte $a_{i,j}^{(\nu+1)}$ gemäß (2.2.15), (viii) als

$$a_{i,j}^{(\nu+1)} = gl(a_{i,j}^{(\nu)} - m_{i,\nu}\, a_{\nu,j}^{(\nu)}) \qquad (\nu = 1, \ldots, i-1) \tag{3.6.7}$$

berechnet, wobei die beim ν-ten Schritt benötigten Größen $a_{i,j}^{(\nu)}$, $a_{\nu,j}^{(\nu)}$ und $m_{i,\nu}$ jeweils vorher bestimmt worden sind.

Ein fester Koeffizient $m_{i,j}$ $(j < i)$ von M wird nach (2.2.15, vii) beim j-ten Eliminationsschritt berechnet. Dazu werden die Größen $a_{i,j}^{(j)}$ und $a_{j,j}^{(j)}$ benötigt, die im Fall $j \geqq 2$ in den Schritten $\nu = 1, \ldots, j-1$ ermittelt werden. Unter Berücksichtigung der Zwischenergebnisse für $a_{i,j}^{(j)}$ $(a_{j,j}^{(j)}$ siehe (3.6.7)!) erhalten wir $m_{i,j}$ über

$$\begin{cases} \text{(i)} & a_{i,j}^{(\nu+1)} = gl(a_{i,j}^{(\nu)} - m_{i,\nu}\, a_{\nu,j}^{(\nu)}) \qquad (\nu = 1, \ldots, j-1), \text{ falls } j \geqq 2, \\ \text{(ii)} & m_{i,j} = gl\left(\dfrac{a_{i,j}^{(j)}}{a_{j,j}^{(j)}}\right). \end{cases} \tag{3.6.8}$$

Zu (3.6.7) und (3.6.8, i) definieren wir absolute Rundungsfehler $\epsilon_{i,j}^{(\nu)}$ durch

$$a_{i,j}^{(\nu+1)} := a_{i,j}^{(\nu)} - m_{i,\nu}\, a_{\nu,j}^{(\nu)} + \epsilon_{i,j}^{(\nu)} \qquad (1 \leqq \nu \leqq \min(i,j) - 1) \tag{3.6.9}$$

und zu (3.6.8, ii) $\epsilon_{i,j}^{(j)}$ durch die Beziehung

$$0 = a_{i,j}^{(j)} - m_{i,j}\, a_{j,j}^{(j)} + \epsilon_{i,j}^{(j)}, \qquad \text{falls } j+1 \leqq i \leqq n. \tag{3.6.10}$$

Hierzu notieren wir die

(3.6.11) **Bemerkung.** Sämtliche in (3.6.9), (3.6.10) auftretenden $\epsilon_{i,j}^{(\nu)}$ genügen den Ungleichungen

$$|\epsilon_{i,j}^{(\nu)}| \leqq 2\tau g.$$

Zum *Beweis* betrachten wir zunächst ein nach (3.6.9) durch

$$(*) \qquad \epsilon_{i,j}^{(\nu)} := a_{i,j}^{(\nu+1)} - (a_{i,j}^{(\nu)} - m_{i,\nu}\, a_{\nu,j}^{(\nu)})$$

definiertes $\epsilon_{i,j}^{(\nu)}$ und beachten, daß mit gewissen, von i, j, ν abhängigen $\vartheta_1, \vartheta_2 \in \mathbb{R}$ mit $|\vartheta_1|, |\vartheta_2| \leqq \tau$ die Beziehung

$$a_{i,j}^{(\nu+1)} = gl[a_{i,j}^{(\nu)} - (m_{i,\nu}\, a_{\nu,j}^{(\nu)})] = [a_{i,j}^{(\nu)} - m_{i,\nu}\, a_{\nu,j}^{(\nu)}(1+\vartheta_1)]\,\frac{1}{1+\vartheta_2}$$

und folglich

$$(1+\vartheta_2)\, a_{i,j}^{(\nu+1)} + m_{i,\nu}\, a_{\nu,j}^{(\nu)}\, \vartheta_1 = a_{i,j}^{(\nu)} - m_{i,\nu}\, a_{\nu,j}^{(\nu)}$$

gilt. Durch Einsetzen in (*) gewinnen wir die Darstellung

$$\epsilon_{i,j}^{(\nu)} = -\vartheta_2\, a_{i,j}^{(\nu+1)} - m_{i,\nu}\, a_{\nu,j}^{(\nu)}\, \vartheta_1$$

und wegen $|m_{i,\nu}| \leqq 1$; $|a_{i,j}^{(\nu+1)}|$, $|a_{\nu,j}^{(\nu)}| \leqq g$ die behauptete Abschätzung $|\epsilon_{i,j}^{(\nu)}| \leqq 2\tau g$. Zu (3.6.10) benutzen wir

$$m_{i,j} = g l\left(\frac{a_{i,j}^{(j)}}{a_{j,j}^{(j)}}\right) = \frac{a_{i,j}^{(j)}}{a_{j,j}^{(j)}}\,(1+\eta_{i,j})$$

mit einem $|\eta_{i,j}| \leqq \tau$ und folgern die Gleichung

$$a_{i,j}^{(j)} - m_{i,j}\, a_{j,j}^{(j)} + a_{i,j}^{(j)}\,\eta_{i,j} = 0\,,$$

woraus wir die Darstellung $\epsilon_{i,j}^{(j)} = a_{i,j}^{(j)}\,\eta_{i,j}$ und die Abschätzung $|\epsilon_{i,j}^{(j)}| \leqq \tau g \leqq 2\tau g$ gewinnen.

Aufsummieren der Gleichungen (3.6.9) über ν und Addition von (3.6.10) im Fall $i > j$ liefert

$$a_{i,j}^{(i)} = a_{i,j}^{(1)} - \sum_{\nu=1}^{i-1} m_{i,\nu}\, a_{\nu,j}^{(\nu)} + \sum_{\nu=1}^{i-1} \epsilon_{i,j}^{(\nu)} \qquad (2 \leqq i \leqq n;\ i \leqq j \leqq k)$$

beziehungsweise

$$0 = a_{i,j}^{(1)} - \sum_{\nu=1}^{j} m_{i,\nu}\, a_{\nu,j}^{(\nu)} + \sum_{\nu=1}^{j} \epsilon_{i,j}^{(\nu)} \qquad (1 \leqq j \leqq n-1;\ j+1 \leqq i \leqq n)\,.$$

Mit den Größen

$$(3.6.12) \qquad f_{i,j} = \begin{cases} 0 & (i = 1;\ 1 \leqq j \leqq k)\,, \\ \displaystyle\sum_{\nu=1}^{i-1} \epsilon_{i,j}^{(\nu)} & (2 \leqq i \leqq n;\ i \leqq j \leqq k)\,, \\ \displaystyle\sum_{\nu=1}^{j} \epsilon_{i,j}^{(\nu)} & (1 \leqq j \leqq n-1;\ j+1 \leqq i \leqq n) \end{cases}$$

erhalten wir hieraus die Gleichungen

$$\sum_{\nu=1}^{i-1} m_{i,\nu}\, a_{\nu,j}^{(\nu)} + a_{i,j}^{(i)} = a_{i,j}^{(1)} + f_{i,j} \qquad (1 \leqq i \leqq n;\ i \leqq j \leqq k)\,,$$

$$\sum_{\nu=1}^{j} m_{i,\nu}\, a_{\nu,j}^{(\nu)} = a_{i,j}^{(1)} + f_{i,j} \qquad (1 \leqq j \leqq n-1;\ j+1 \leqq i \leqq n)\,,$$

und zusammengefaßt

$$MU = A + F, \quad F := (f_{i,j})_{(n,k)}.$$

Aus (3.6.11) und den Darstellungen (3.6.12) schließen wir

$$(3.6.13) \quad |f_{i,j}| \leqq \begin{cases} 2\tau g(i-1) & (1 \leqq i \leqq n;\ i \leqq j \leqq k), \\ 2\tau g \cdot j & (1 \leqq j \leqq n-1;\ j+1 \leqq i \leqq n). \end{cases}$$

Hiermit ist die Aussage (i) des Satzes (3.6.6) bewiesen. – Wir erkennen die Größen $a_{i,j}^{(\nu+1)}$ in (3.6.7) bzw. (3.6.8) als die Zwischenergebnisse bei der numerischen Auswertung der Vorschriften

$$(3.6.14) \quad \begin{cases} a_{i,j}^{(i)} = gl\left(a_{i,j}^{(1)} - \sum_{\kappa=1}^{i-1} m_{i,\kappa}\, a_{\kappa,j}^{(\kappa)}\right) & (i \leqq j), \\ m_{i,j} = gl\left(\dfrac{a_{i,j}^{(1)} - \sum_{\kappa=1}^{j-1} m_{i,\kappa}\, a_{\kappa,j}^{(\kappa)}}{a_{j,j}^{(j)}}\right) & (i \geqq j+1), \end{cases}$$

falls hierbei die Rechenoperationen in der gleichen Reihenfolge wie bei der Bestimmung von $\tilde{x}_r$ im Hilfssatz (3.5.2) ausgeführt werden. Mit (3.5.2), (iii) bzw. (ii) erhalten wir wegen

$$f_{i,j} = -a_{i,j}^{(1)} + \sum_{\kappa=1}^{\min(i,j)} m_{i,\kappa}\, a_{\kappa,j}^{(\kappa)}$$

unmittelbar

$$(3.6.15) \quad \begin{cases} \text{(i)} \quad |f_{i,j}| \leqq \dfrac{\tau}{1-n\tau}\left(\sum_{\nu=1}^{i} \nu\, |m_{i,\nu}|\ |a_{\nu,j}^{(\nu)}| - |a_{i,j}^{(i)}|\right) & (i \leqq j), \\ \text{(ii)} \quad |f_{i,j}| \leqq \dfrac{\tau}{1-n\tau}\left(\sum_{\nu=1}^{j} \nu\, |m_{i,\nu}|\ |a_{\nu,j}^{(\nu)}|\right) & (i \geqq j+1). \end{cases}$$

Auf Grund der Beziehungen

$$\hat{DU} = (\nu\, |a_{\nu,j}^{(\nu)}|)_{\substack{1 \leqq \nu \leqq n \\ 1 \leqq j \leqq k}}$$

und $a_{i,j}^{(i)} = 0$ für $i \geqq j+1$ erweisen sich die rechten Seiten von (3.6.15) als die Koeffizienten der Matrix

$$\frac{\tau}{1-n\tau}(\hat{M}\hat{DU} - \hat{U});$$

damit ist der Beweis des Satzes (3.6.6) abgeschlossen.

Über die wichtigsten Normen von F bringen wir den

(3.6.16) **Zusatz.** Im Fall $k = n$ gelten die Abschätzungen

$$\|F\|_\infty \leqq (n^2 + n - 2)g\tau,$$
$$\|F\|_1 \leqq (n^2 - n)g\tau,$$
$$\|F\|_2 \leqq 0{,}82\, n^2 g\tau.$$

Beweis. Mit der (n,n)-Matrix

$$G = \begin{pmatrix} 0 & \cdots & \cdots & 0 \\ 1 & 1 & \cdots & 1 \\ | & 2 & \cdots & 2 \\ | & | & \searrow & \vdots \\ 1 & 2 & \cdots & (n-1) \quad (n-1) \end{pmatrix}$$

gilt die komponentenweise Ungleichung $\hat{F} \leqq 2g\tau G$; wir berechnen

$$\|G\|_\infty = \sum_{\nu=1}^{n} \nu - 1 = \frac{1}{2}(n^2 + n - 2),$$

$$\|G\|_1 = \sum_{\nu=1}^{n-1} \nu = \frac{1}{2}(n^2 - n);$$

Zur Bestimmung der Frobeniusnorm beachten wir, daß in G ein fester Wert $\kappa \in \{1, \ldots, n-1\}$ gerade $(n - \kappa + 1)$-mal in einer Zeile und zusätzlich $(n - \kappa - 1)$-mal in einer Spalte vorkommt, so daß gilt

$$\|G\|_2^2 = \sum_{\kappa=1}^{n-1} (\kappa^2(n - \kappa + 1) + \kappa^2(n - \kappa - 1)) = 2 \sum_{\kappa=1}^{n-1} \kappa^2(n - \kappa) =$$
$$= 2n^2 \frac{(n-1)(2n-1)}{6} - \frac{n^2(n-1)^2}{2} = \frac{n^4 - n^2}{6} < \frac{n^4}{6},$$

folglich

$$\|G\|_2 \leqq \frac{n^2}{\sqrt{6}} \leqq 0{,}41\, n^2.$$

Die zur Abschätzung von $\hat{F}$ benötigte Matrix $\hat{M}D\hat{U} - \hat{U}$ läßt sich leicht nachträglich ausrechnen, wobei wegen der Summation positiver Zahlen nur geringe Rundungsfehler auftreten; auch die Größe g läßt sich im Verlauf der Eliminationsschritte ohne weiteres bestimmen, so daß uns der Satz (3.6.6) zwei brauchbare *a posteriori* Abschätzungen liefert. Um prinzipielle Aussagen über die numerische Qualität der Gauß-Elimination bei halbmaximaler bzw. maximaler Pivotwahl zu erhalten, suchen wir *a priori* Schranken für g. Da es uns hierbei nicht um exakte Fehlerabschätzungen geht, sollen in den folgenden Betrachtungen die Rundungs-

fehler unberücksichtigt bleiben. Wir gehen also von einer reellen (oder komplexen) Matrix

$$A = C^{(1)} = (c_{i,j}^{(1)})_{(n,k)}$$

aus und betrachten dazu die bei Rechnen ohne Rundungsfehler entstehenden

$$C^{(\nu)} = (c_{i,j}^{(\nu)})_{(n,k)} \qquad (\nu = 1, \ldots, n).$$

wobei wir die Voraussetzungen (3.6.1) und (3.6.2) sinngemäß für die exakte Rechnung übernehmen. Wir verwenden die

(3.6.17) **Bezeichnung.**

$$p_\nu := \max \{|c_{i,j}^{(\nu)}|: \nu \leqq i \leqq n,\ \nu \leqq j \leqq k\} \quad (\nu = 1, \ldots, n)$$

und notieren dazu die

(3.6.18) **Bemerkung.** Bei Rechnung ohne Rundungsfehler wird

$$g = \max_{\nu=1}^{n} p_\nu .$$

Beweis. Bei rundungsfehlerfreier Rechnung stimmen die $A^{(\nu)}$ mit den $C^{(\nu)}$ überein; die Ungleichung $\max_{\nu=1}^{n} p_\nu \leqq g$ ist nach (3.6.5) trivial; andererseits haben wir

$$|a_{i,j}^{(\nu)}| = |c_{i,j}^{(\nu)}| = |c_{i,j}^{(i)}| \leqq p_i \quad \text{für } i < \nu; \quad a_{i,j}^{(\nu)} = c_{i,j}^{(\nu)} = 0 \quad \text{für } j < \nu.$$

Eine erste Abschätzung der p_ν notieren wir in dem

(3.6.19) **Satz.** *Bei halbmaximaler und maximaler Pivotwahl gilt*

$$p_\nu \leqq 2^{\nu-1} p_1 \qquad (\nu = 1, \ldots, n)$$

und folglich bei rundungsfehlerfreier Rechnung

$$g \leqq 2^{n-1} p_1 .$$

Beweis. Für $1 \leqq \nu \leqq n-1$; $i, j \geqq \nu + 1$ haben wir nach (2.2.15), (vii) und (viii)

$$|c_{i,j}^{(\nu+1)}| = |c_{i,j}^{(\nu)} - d_{i,\nu} c_{\nu,j}^{(\nu)}|, \quad |d_{i,\nu}| = \frac{|c_{i,\nu}^{(\nu)}|}{|c_{\nu,\nu}^{(\nu)}|} \leqq 1 ,$$

also

$$|c_{i,j}^{(\nu+1)}| \leqq 2 \cdot p_\nu .$$

Es folgt $p_{\nu+1} \leqq 2 \cdot p_\nu$ und hieraus die Behauptung des Satzes.

Daß bei halbmaximaler Pivotwahl die in (3.6.19) angegebenen Schranken erreicht werden können, zeigt das

(3.6.20) **Beispiel.** Für die (n,n)-Matrix

$$A = \begin{pmatrix} 1 & 0 & \cdots & \cdots & 0 & 1 \\ -1 & 1 & 0 & & 0 & 1 \\ -1 & -1 & 1 & 0 & 0 & 1 \\ \vdots & \vdots & & \ddots & \vdots & \vdots \\ & & & 1 & 0 & 1 \\ & & & -1 & 1 & 1 \\ -1 & \cdots & \cdots & -1 & -1 & 1 \end{pmatrix}$$

rechnet man $p_\nu = 2^{\nu-1} = 2^{\nu-1} p_1$ $(\nu = 1, \ldots, n)$ leicht nach.

Zu einer schärferen Abschätzung der p_ν im Fall der maximalen Pivotwahl setzen wir zunächst $k = n$ voraus (vgl. Aufgabe 3.11!) und erhalten entsprechend der Voraussetzung (3.6.2, ii) die Darstellungen

$$p_\nu = |c^{(\nu)}_{\nu,\nu}| \qquad (\nu = 1, \ldots, n).$$

Wir zeigen die

(3.6.21) **Bemerkung.** Für die $(n-\nu+1,\ n-\nu+1)$-Teilmatrizen $\tilde{C}^{(\nu)}$ der $C^{(\nu)}$ gemäß der Aufteilung

$$C^{(\nu)} = \begin{pmatrix} c^{(1)}_{1,1} & \cdots & \cdots & \cdots & \cdots \\ 0 & \ddots & & & \\ \vdots & & c^{(\nu-1)}_{\nu-1,\nu-1} & \cdots & \\ \vdots & & 0 & & \\ \vdots & & \vdots & & \tilde{C}^{(\nu)} \\ 0 & \cdots & 0 & & \end{pmatrix},$$

wobei $\tilde{C}^{(1)} = C^{(1)}$ gesetzt ist, gilt

$$|\det \tilde{C}^{(\nu)}| = |c^{(\nu)}_{\nu,\nu} \cdot c^{(\nu+1)}_{\nu+1,\nu+1} \cdot \ldots \cdot c^{(n)}_{n,n}| = p_\nu \cdot p_{\nu+1} \cdot \ldots \cdot p_n \qquad (\nu = 1, \ldots, n).$$

Beweis. Man hat gemäß (2.2.14) für alle $\nu = 1, \ldots, n$

$$\det A = \det C^{(\nu)} = c^{(1)}_{1,1} \ldots c^{(\nu-1)}_{\nu-1,\nu-1} \cdot \det \tilde{C}^{(\nu)} = c^{(1)}_{1,1} \cdot \ldots \cdot c^{(n)}_{n,n}$$

und, weil das Verfahren laut Voraussetzung erst nach $(n-1)$ Schritten endet,

$$c^{(1)}_{1,1} \cdot \ldots \cdot c^{(\nu-1)}_{\nu-1,\nu-1} \neq 0.$$

(3.6.22) **Hilfssatz.**

$$p_\nu \cdot p_{\nu+1} \cdot \ldots \cdot p_n \leqq (n-\nu+1)^{\frac{n-\nu+1}{2}} p_\nu^{n-\nu+1} \qquad (\nu = 1, \ldots, n).$$

Beweis. Wir wenden die Determinantenabschätzung nach Hadamard (Aufgabe 2.12) auf $\tilde{C}^{(\nu)}$ an; da sich jeder Koeffizient von $\tilde{C}^{(\nu)}$ durch p_ν abschätzen läßt, wird die euklidische Norm jeder Spalte in $\tilde{C}^{(\nu)}$ höchstens $\sqrt{(n-\nu+1)}\,p_\nu$ und folglich

$$p_\nu \cdot \ldots \cdot p_n = |\det \tilde{C}^{(\nu)}| \leqq \prod_{j=\nu}^{n} (n-\nu+1)^{\frac{1}{2}} p_\nu = (n-\nu+1)^{\frac{n-\nu+1}{2}} p_\nu^{n-\nu+1}.$$

Wir erhalten nun nach Wilkinson [33] den

(3.6.23) **Satz.** *Falls auf die reelle (oder komplexe)* (n,n)*-Matrix* A *mit* $\operatorname{rg} A = n$ *Gauß-Elimination mit maximaler Pivotwahl angewendet wird, so gilt*

$$p_\nu \leqq \sqrt{\nu}\, f(\nu)\, p_1 \qquad (\nu = 2, 3, \ldots, n)$$

mit

$$f(\nu) := (2 \cdot 3^{\frac{1}{2}} \ldots \nu^{\frac{1}{\nu-1}})^{\frac{1}{2}} \qquad (\nu = 2, 3, \ldots, n).$$

Beweis. Wir zeigen den Satz zunächst für $\nu = n$, also

(3.6.24) $$p_n \leqq \sqrt{n}\, f(n)\, p_1 .$$

Wegen $p_\nu > 0$ können wir

$$q_\nu := \log p_\nu \qquad (\nu = 1, \ldots, n)$$

definieren und gewinnen aus (3.6.22) die Ungleichungen

$$(*) \quad \sum_{j=\nu}^{n} q_j \leqq \frac{n-\nu+1}{2} \log(n-\nu+1) + (n-\nu+1) \cdot q_\nu \qquad (\nu = 1, \ldots, n-1).$$

Wir subtrahieren q_ν auf beiden Seiten und dividieren durch $(n-\nu+1)\cdot(n-\nu)$, wodurch

$$\frac{1}{(n-\nu)(n-\nu+1)} \sum_{j=\nu+1}^{n} q_j \leqq \frac{1}{2(n-\nu)} \log(n-\nu+1) + \frac{1}{n-\nu+1} q_\nu$$

entsteht. Diese Ungleichungen für $\nu = 2, \ldots, n-1$ werden aufaddiert, hinzu nehmen wir die aus $(*)$ für $\nu = 1$ stammende Beziehung

$$\frac{1}{n-1} \sum_{j=1}^{n} q_j \leqq \frac{n}{2(n-1)} \log n + \frac{n}{n-1} q_1 = \frac{1}{2(n-1)} \log n + \frac{1}{2} \log n + \frac{n}{n-1} q_1 ,$$

und erhalten somit

$$(**) \quad \sum_{\nu=2}^{n-1} \frac{1}{(n-\nu)(n-\nu+1)} \sum_{j=\nu+1}^{n} q_j + \frac{1}{n-1} \sum_{j=1}^{n} q_j \leqq \frac{1}{2} \sum_{\nu=1}^{n-1} \frac{1}{n-\nu} \log(n-\nu+1) +$$

$$+ \log\sqrt{n} + \sum_{j=2}^{n-1} \frac{q_j}{n-j+1} + \frac{n}{n-1} q_1 .$$

Nach Vertauschen der Summationsreihenfolge wird aus der linken Seite

$$\sum_{j=3}^{n} q_j \left[\sum_{\nu=2}^{j-1} \frac{1}{(n-\nu)(n-\nu+1)} + \frac{1}{n-1} \right] + \frac{q_1}{n-1} + \frac{q_2}{n-1}$$

mit

$$\sum_{\nu=2}^{j-1} \frac{1}{(n-\nu)(n-\nu+1)} + \frac{1}{n-1} = \sum_{\nu=2}^{j-1} \left[\frac{1}{n-\nu} - \frac{1}{n-\nu+1} \right] + \frac{1}{n-1} = \frac{1}{n-j+1};$$

auf der linken Seite von (**) steht also

$$\sum_{j=2}^{n} \frac{1}{n-j+1} q_j + \frac{1}{n-1} q_1 .$$

Durch Subtraktion gleicher Ausdrücke auf beiden Seiten von (**) erhalten wir nun

$$q_n \leqq \frac{1}{2} \sum_{\nu=1}^{n-1} \frac{1}{n-\nu} \log(n-\nu+1) + \log\sqrt{n} + \left(\frac{n}{n-1} - \frac{1}{n-1} \right) q_1 = \log(\sqrt{n}\, f(n)) + q_1$$

und daraus (3.6.24). Zum Beweis der allgemeinen Aussage in (3.6.23) betrachten wir für $2 \leqq \nu \leqq n-1$ die Teilmatrizen

$$\tilde{A}_\nu := (c_{i,j}^{(1)})_{1 \leqq i,j \leqq \nu}$$

von A. Die ersten $(\nu - 1)$ Eliminationsschritte für A bewirken die entsprechenden Eliminationsschritte in $\tilde{A}_\nu$ und führen $\tilde{A}_\nu$ in eine obere Dreiecksmatrix über. Für die Pivotelemente beim μ-ten Schritt $(1 \leqq \mu \leqq \nu)$ gilt auf Grund der Annahme (3.6.2, i) offenbar

$$p_\mu = |c_{\mu,\mu}^{(\mu)}| = \max\{|c_{i,j}^{(\mu)}| : \mu \leqq i,j \leqq n\} = \max\{|c_{i,j}^{(\mu)}| : \mu \leqq i,j \leqq \nu\},$$

und daher für die Größen $\tilde{p}_\mu$ in $\tilde{A}_\nu$, die den p_ν in A entsprechen,

$$\tilde{p}_\mu = p_\mu \qquad (\mu = 1, \ldots, \nu).$$

Die Aussage (3.6.24), auf $\tilde{A}_\nu$ statt A angewendet, liefert also

$$p_\nu \leqq \sqrt{\nu}\, f(\nu) \cdot p_1 .$$

Die Funktion $\sqrt{n}\, f(n)$ wächst wesentlich langsamer als 2^{n-1}; wir geben einige Werte an:

n	10	20	50	100	200	1000
$\sqrt{n}\, f(n)$	19	67	530	3300	26000	7900000

Bei halbmaximaler Pivotwahl kann nach (3.6.20)

$$p_n = 2^{n-1} p_1$$

eintreten, das dazu angegebene Beispiel ist jedoch nicht typisch für praktisch auftretende Matrizen. Die numerische Erfahrung zeigt, daß im allgemeinen die p_ν nicht stärker anwachsen als bei maximaler Pivotwahl.

Für die maximale Pivotwahl kann man zeigen, daß $p_n = \sqrt{n}\, f(n) \cdot p_1$ für $n \geqq 4$ nicht erreicht werden kann; darüberhinaus ist keine Matrix mit $p_n > n \cdot p_1$ bekannt. Aus diesem Grund vermutet man an Stelle von (3.6.24) ein Verhalten wie

$$p_n \leqq n \cdot p_1 .$$

Bei Matrizen von spezieller Gestalt wie oberen Hessenbergmatrizen und insbesondere Tridiagonalmatrizen gilt eine derartige Abschätzung für p_n sogar im Fall halbmaximaler Pivotwahl. Hierzu notieren wir den in den Übungsaufgaben 2.4 und 2.5 zu beweisenden

(3.6.25) **Satz.** *Bei halbmaximaler Pivotwahl wird*

(i) *für eine obere Hessenbergmatrix* A

$$p_\nu \leqq \nu \cdot p_1 \qquad (\nu = 1, \ldots, n),$$

(ii) *für eine Tridiagonalmatrix* A

$$p_\nu \leqq 2 \cdot p_1 \qquad (\nu = 1, \ldots, n).$$

Wir kommen zur Fehlerabschätzung für die mit Gauß-Elimination numerisch ermittelte Lösung $\tilde{x}$ eines Gleichungssystems

$$Ax = b$$

unter der

(3.6.26) **Voraussetzung.** A sei invertierbare (n,n)-Matrix, $b \in \mathbb{R}^n$, $\neq 0$; die Koeffizienten von A und b seien t-stellige Gleitkommazahlen, und die Gauß-Zerlegung von A werde gemäß den Voraussetzungen (3.6.1) und (3.6.2) durchgeführt.

Die Lösung von $Ax = b$ zerfällt in 3 Schritte; bei exakter Rechnung werden nämlich L, R, s und x mit

(1) $A = L \cdot R$,
(2) $L s = b$,
(3) $R x = s$

bestimmt. Für die entsprechenden numerisch gewonnenen Matrizen M, U, v und $\tilde{x}$ erhalten wir nach Satz (3.6.6) sowie (3.5.6) und (3.5.7) die Beziehungen

(1′) $A + F = M U$,
(2′) $(M + \Delta M)\, v = b$,
(3′) $(U + \Delta U)\, \tilde{x} = v$.

Einsetzen von (3′) in (2′) liefert

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

und mit (1′) zusammen

$$(A + F + \Delta M \cdot U + M \cdot \Delta U + \Delta M \cdot \Delta U)\,\tilde{x} = b.$$

Mit den genannten Bezeichnungen notieren wir den

(3.6.27) **Satz.** *Unter der Voraussetzung* (3.6.26) *erfüllt die mit* t*-stelliger Gleitkommarechnung ermittelte Lösung* $\tilde{x}$ *von* $Ax = b$ *das Gleichungssystem*

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

in dem

$$\Delta A = F + \Delta M \cdot U + M \cdot \Delta U + \Delta M \cdot \Delta U$$

gesetzt ist. ΔA *läßt sich komponentenweise durch*

$$\hat{\Delta A} \leqq \frac{2{,}1\, n\tau}{1 - n\tau}\, \hat{M} \cdot \hat{U}$$

abschätzen.

Beweis. Es ist nur noch die Abschätzung von ΔA auszuführen. Zunächst haben wir

$$(*) \quad \hat{\Delta A} \leqq \hat{F} + \hat{\Delta M}\,\hat{U} + (\hat{M} + \hat{\Delta M})\,\hat{\Delta U}\ ;$$

ferner übernehmen wir aus (3.6.6, ii), (3.5.6) und (3.5.7) die Ungleichungen

$$\hat{F} \leqq \frac{\tau}{1 - n\tau}\,(\hat{M}D - I)\,\hat{U}, \qquad \hat{\Delta U} \leqq \frac{\tau}{1 - n\tau}\,\hat{U}\,E,$$

$$\hat{\Delta M} \leqq \frac{\tau}{1 - n\tau}\,(\hat{M}D - I) \leqq \frac{n-1}{1 - n\tau}\,\tau\,\hat{M}$$

mit $D := \operatorname{diag}(1, \ldots, n)$, $E := \operatorname{diag}(n, \ldots, 1)$. Auf Grund der Voraussetzung

$n\tau \leqq 0{,}09$ wird

$$\frac{n-1}{1 - n\tau}\,\tau \leqq 0{,}1\ ,$$

folglich

$$\hat{M} + \hat{\Delta M} \leqq 1{,}1\,\hat{M}.$$

Einsetzen der aufgeführten Ungleichungen in (*) liefert

$$\hat{\Delta A} \leqq \frac{\tau}{1 - n\tau}\,[2\,(\hat{M}D - I)\,\hat{U} + 1{,}1 \cdot \hat{M} \cdot \hat{U} \cdot E].$$

Der Koeffizient (i, j) in der rechts stehenden Matrix ist gerade

$$(**) \quad \frac{\tau}{1 - n\tau} \sum_{\nu=1}^{\min(i,j)} |m_{i,\nu}|\,(2\nu - 2\delta_{\nu,j} + 1{,}1(n + 1 - j))\,|a_{\nu,j}^{(\nu)}|\ ,$$

worin $\delta_{\nu,j}$ das Kroneckersymbol bedeutet. Hierzu zeigen wir die

(3.6.28) **Bemerkung.** Für alle ν, j mit $1 \leqq \nu \leqq j \leqq n$ gilt

$$2\nu - 2\delta_{\nu,j} + 1{,}1 \cdot (n+1-j) \leqq 2{,}1 \cdot n.$$

Wir haben nämlich für $\nu < j$

$$2\nu - 2\delta_{\nu,j} + 1{,}1(n+1-j) = \nu + 1{,}1(n+1-j) + \nu < \nu + 1{,}1(n+1-(j-\nu)) \leqq \nu + 1{,}1 \cdot n$$

und für $j = \nu$:

$$2\nu - 2\delta_{\nu,j} + 1{,}1(n+1-j) = 2(\nu-1) + 1{,}1(n-(\nu-1)) \leqq (\nu-1) + 1{,}1 \cdot n.$$

Mit Hilfe von (3.6.28) wird (**) durch

$$2{,}1\, n \frac{\tau}{1-n\tau} \sum_{\nu=1}^{\min(i,j)} |m_{i,\nu}|\, |a_{\nu,j}^{(\nu)}|,$$

also durch den Koeffizienten (i, j) der Matrix

$$\frac{2{,}1\, n\tau}{1-n\tau\tau} \hat{M}\hat{U}$$

abgeschätzt, womit der Satz (3.6.27) gezeigt ist.

Wegen $A \approx MU$ gilt

$$\hat{A} \approx \hat{M}\hat{U} \leqq \hat{M}\hat{U}.$$

Wenn die Koeffizienten von $\hat{M}\hat{U}$ nicht erheblich größer als die entsprechenden Koeffizienten von $\hat{A}$ werden, ist der Rundungsfehlereinfluß bei der Lösung des Gleichungssystems ähnlich, als hätte man die $a_{i,j}$ mit relativen Fehlern vom Betrag $\leqq 2n\tau$ gestört. Die Gauß-Elimination ist also im Fall $\hat{M}\hat{U} \approx \hat{A}$ numerisch stabil; durch Berechnung von $\hat{M}\hat{U}$ bietet sich die Möglichkeit, die numerische Stabilität des Verfahrens im Einzelfall nachzuprüfen. – Bei nicht zu großem n beeinflussen die Rundungsfehler erfahrungsgemäß das Ergebnis ähnlich wie das Runden der Eingangsdaten auf die benutzte Stellenzahl.

(3.6.29) **Zahlenbeispiel.** Das Gleichungssystem (2.3.3) liefert bei doppeltgenau eingelesenen Koeffizienten nach (3.4.10, i) die auf 7 Dezimalen gerundete Lösung

(i) $x_1 = -20{,}76272,\ x_2 = -2{,}747919,\ x_3 = 14{,}74503,\ x_4 = 2{,}615863;$

nach (2.6.18) bewirkt das Runden der Koeffizienten auf REAL*4-Genauigkeit, aber doppeltgenaue Rechnung das Ergebnis

(ii) $x_1 = -20{,}76267,\ x_2 = -2{,}747911,\ x_3 = 14{,}74500,\ x_4 = 2{,}615857;$

und bei durchgehender REAL*4-Arithmetik erhalten wir

(iii) $x_1 = -20{,}76247,\ x_2 = -2{,}747877,\ x_3 = 14{,}74485,\ x_4 = 2{,}615837.$

Ein Vergleich der Ergebnisse liefert eine Übereinstimmung in etwa den ersten 5 Dezimalen. Daß keine bessere Genauigkeit erzielt wird, liegt an der großen Konditionszahl von A.

Wir kommen zu einer *a priori* Abschätzung von ΔA im Satz (3.6.27).

Dazu benutzen wir

$$|m_{i,j}| \leqq 1, \quad |a_{i,j}^{(i)}| \leqq g \qquad (i,j = 1, \ldots, n),$$

so daß wir mit

$$V := \begin{pmatrix} 1 & & \\ \vdots & \ddots & 0 \\ \vdots & & \ddots \\ 1 & \ldots\ldots & 1 \end{pmatrix}$$

wegen

$$\hat{M} \leqq V, \quad \hat{U} \leqq g V^t \Rightarrow \hat{M}\hat{U} \leqq g V V^t$$

folgende Ungleichung erhalten:

(3.6.30) **Bemerkung.**

$$\hat{\Delta A} \leqq \frac{2{,}1\, n\tau}{1 - n\tau} g \begin{pmatrix} 1 & \text{———} & 1 \\ | & 2 & \text{—} & 2 \\ | & | & & \vdots \\ 1 & 2 & \ldots\ldots & n \end{pmatrix}$$

Hiermit lassen sich die wichtigsten Matrixnormen von ΔA abschätzen. Als Konsequenz notieren wir unter Benutzung von Satz (3.4.2) bzw. (3.4.7) eine Normabschätzung für $\tilde{x} - x$:

(3.6.31) **Folgerung.**

(i) $$\|\Delta A\|_\infty \leqq \frac{1{,}05}{1 - n\tau}\, \tau(n^3 + n^2)\, g \approx g n^3 \tau .$$

(ii) Im Fall $\|\Delta A\|_\infty \, \|A^{-1}\|_\infty < 1$ gilt

$$\frac{\|\tilde{x} - x\|_\infty}{\|x\|_\infty} \leqq \frac{\kappa_\infty(A)}{1 - \kappa_\infty(A)\frac{\|\Delta A\|_\infty}{\|A\|_\infty}} \cdot \frac{\|\Delta A\|_\infty}{\|A\|_\infty} \approx \kappa_\infty(A)\, \frac{g}{\|A\|_\infty}\, n^3 \cdot \tau .$$

Statt eines Wachstums mit n^3 hat man in der Praxis ein Verhalten der Rundungsfehler etwa proportional zu n beobachtet. – Eine weitere schwache Stelle der Abschätzungen in (3.6.31) ist die Größe g, die theoretisch stark anwachsen kann.

3.7. Rundungsfehler bei der Householder-Zerlegung

Wir wollen die Rundungsfehler untersuchen, die bei der Lösung eines Gleichungssystems Ax = b mit dem Householder-Verfahren entstehen. Dazu nehmen wir an:

(3.7.1) **Voraussetzung**.

(i) Es sei $A = (a_{i,j})_{(n,n)}$ invertierbare (n,n)-Matrix, $b = (b_i)_1^n \in \mathbb{R}^n$, $\neq 0$; die $a_{i,j}$ und b_i seien reelle t-stellige Gleitkommazahlen.

(ii) Für die Rechengenauigkeit $\tau = \frac{1}{2} \cdot 10^{-t+1}$ gelte $(2{,}5\,n + 6)\,\tau \leqq 0{,}09$.

(iii) In der benutzten Gleitkomma-Arithmetik sei für eine t-stellige Gleitkommazahl $a \geqq 0$ stets

$$gl(\sqrt{a}) = \sqrt{a}(1+\varphi) = \frac{\sqrt{a}}{1+\psi}; \quad |\varphi|, |\psi| \leqq \tau .$$

Es bezeichne im Folgenden

$$C_1 := (A, b) = (a_1, \ldots, a_{n+1}) = (a_{i,j})_{(n,n+1)}$$

und für $\nu = 2, \ldots, n$

$$C_\nu = (A_\nu, b^{(\nu)}) = \left(\begin{array}{ccc|c} r_{1,1} & \cdots & r_{1,n} & s_1 \\ 0 & & & \vdots \\ \vdots & \ddots & & \\ & r_{\nu-1,\nu-1} & \cdots & s_{\nu-1} \\ & 0 & & \\ \vdots & & A'_\nu & b^{(\nu)'} \\ 0 & \cdots\cdots\cdots & 0 & \end{array}\right)$$

die nach $(\nu - 1)$ Householderschritten bei t-stelliger Gleitkommarechnung aus C_1 *numerisch* entstandenen Matrizen, ferner

$$H'_\nu = I_{n-\nu+1} - 2w'_\nu w'^{*}_\nu \qquad (\nu = 1, \ldots, n-1)$$

die zu A'_ν (für $\nu = 1$ zu A) gemäß (2.6.11) gehörenden *theoretischen* Householder-Matrizen, dazu

$$H_1 := H'_1, \quad H_\nu := \left(\begin{array}{c|c} I_{\nu-1} & 0 \\ \hline 0 & H'_\nu \end{array}\right) \qquad (\nu = 2, \ldots, n-1).$$

Mit $\| \ \|$ sei stets die euklidische Norm im $\mathbb{R}^n$ bezeichnet. Wir führen die folgende abkürzende Schreibweise ein:

(3.7.2) **Bezeichnung**. Für reelles $\alpha > 0$ sei $J(\alpha)$ das Intervall

$$J(\alpha) := [(1-\tau)^\alpha, \ (1+\tau)^\alpha] .$$

Hiermit formulieren wir den

(3.7.3) **Hilfssatz.** *Es seien* x_i, y_i $(i = 1, \ldots, n)$ t-*stellige Gleitkommazahlen. Dann gilt eine Darstellung*

$$gl\left(\sum_{i=1}^{n} x_i y_i\right) = \sum_{i=1}^{n} x_i y_i (1 + E_i) = \sum_{i=1}^{n} x_i y_i \frac{1}{1 + F_i}$$

mit

$$1 + E_i,\ 1 + F_i \in J(n) \qquad (i = 1, \ldots, n).$$

Beweis. Bezeichnet s_ν die numerisch berechneten Zwischensummen, so hat man mit gewissen $\delta_i, \epsilon_i, \eta_i, \xi_i \in [-\tau, \tau]$

$$s_1 = gl(x_1 y_1) = x_1 y_1 (1 + \epsilon_1) = x_1 y_1 \frac{1}{1 + \eta_1}$$

und für $\nu = 2, \ldots, n$

$$s_\nu = gl(s_{\nu-1} + x_\nu y_\nu) = (s_{\nu-1} + x_\nu y_\nu (1 + \epsilon_\nu))(1 + \delta_{\nu-1}) = \left(s_{\nu-1} + \frac{x_\nu y_\nu}{1 + \eta_\nu}\right) \frac{1}{1 + \xi_{\nu-1}}.$$

Mit Induktion über n erhalten wir hieraus die behauptete Darstellung.

Über die Rundungsfehler beim 1. Householder-Schritt beweisen wir den

(3.7.4) **Satz.** *Unter der Voraussetzung* (3.7.1) *gilt für die numerisch gewonnene Matrix* $C_2 = (A_2, b^{(2)})$ *eine Darstellung*

$$A_2 = H_1 A + F_1, \qquad b^{(2)} = H_1 b + d^{(1)}$$

mit den Normabschätzungen

$$\|F_1\|_2 \leqq 1{,}1\,(5n + 15)\,\tau\,\|A\|_2,$$
$$\|d^{(1)}\| \leqq 1{,}1\,(5n + 15)\,\tau\,\|b\|.$$

Zum *Beweis* betrachten wir nur den Fall, daß nicht alle $a_{i,1}$ $(2 \leqq i \leqq n)$ verschwinden, und führen den Algorithmus (2.6.15) mit t-stelliger Gleitkommarechnung aus. Wir bezeichnen zu den in (2.6.15) angegebenen theoretischen Größen μ^2, μ, N usw. mit $\widetilde{\mu^2}$, $\tilde{\mu}$, $\tilde{N}$... jeweils die numerisch ermittelten Werte. Da der reelle Fall vorliegt, ist σ ein Vorzeichen, wird also ohne Fehler bestimmt. Zunächst zeigen wir

$$(3.7.5) \qquad \tilde{\mu} = \mu(1 + \epsilon) = \frac{\mu}{1 + \eta}; \qquad 1 + \epsilon,\ 1 + \eta \in J\left(\frac{n+2}{2}\right).$$

Dazu notieren wir für $\widetilde{\mu^2}$ eine Darstellung nach (3.7.3), nämlich

$$\widetilde{\mu^2} = \sum_{i=1}^{n} |a_{i,1}|^2 (1 + E_i) = \sum_{i=1}^{n} |a_{i,1}|^2 \frac{1}{1 + F_i}$$

und schließen hieraus

$$(*) \qquad \mu^2 \cdot \min_{i=1}^{n} (1 + E_i) \leqq \widetilde{\mu^2} \leqq \mu^2 \cdot \max_{i=1}^{n} (1 + E_i)$$

mit $\min(1+E_i),\ \max(1+E_i) \in J(n)$. Wegen $\mu > 0$ folgt

$$1+E := \frac{\widetilde{\mu}^2}{\mu^2} \in J(n).$$

Ersetzt man $(1+E_i)$ in (*) durch $(1+F_i)^{-1}$, erhält man

$$(**)\quad \mu^2 \leqq \widetilde{\mu}^2 \max_{i=1}^{n}(1+F_i), \qquad \mu^2 \geqq \widetilde{\mu}^2 \cdot \min_{i=1}^{n}(1+F_i),$$

daher wie oben

$$(1+F) := \frac{\mu^2}{\widetilde{\mu}^2} \in J(n),$$

insgesamt also

$$(3.7.5')\quad \widetilde{\mu}^2 = \mu^2(1+E) = \frac{\mu^2}{1+F}; \qquad 1+E,\ 1+F \in J(n).$$

Schließlich wird nach (3.7.1, iii)

$$\tilde{\mu} = \mathrm{gl}(\sqrt{\widetilde{\mu}^2}) = \mu \cdot \sqrt{1+E}\,(1+\varphi) = \frac{\mu}{\sqrt{1+F}} \frac{1}{1+\psi}$$

mit $1+\varphi,\ 1+\psi \in J(1)$. Wir setzen also

$$1+\epsilon := (1+\varphi)\sqrt{1+E} \in J\left(\frac{n+2}{2}\right), \quad 1+\eta := (1+\psi)\sqrt{1+F} \in J\left(\frac{n+2}{2}\right).$$

Unter Benutzung von (3.7.5), (3.7.5') berechnen wir weiter

$$\widetilde{N} = \mathrm{gl}(\widetilde{\mu}^2 + \tilde{\mu}\,|a_{1,1}|) = \left(\frac{\mu^2}{1+F} + \frac{\mu}{1+\eta}|a_{1,1}|\,\frac{1}{1+\gamma}\right)\frac{1}{1+\beta}$$

mit $|\gamma|,\ |\beta| \leqq \tau$; folglich gilt eine Darstellung

$$\widetilde{N} = \mu^2 \frac{1}{1+\vartheta_1} + \mu\,|a_{1,1}|\,\frac{1}{1+\vartheta_2}$$

mit $1+\vartheta_1 \in J(n+1)$, $1+\vartheta_2 \in J\left(\frac{n+6}{2}\right)$, also wegen $n \geqq 2$

$$1+\vartheta_i \in J(n+2) \qquad (i=1,2).$$

Wir schließen wie in (**):

$$(3.7.6)\quad \widetilde{N} = \frac{N}{1+\vartheta}, \qquad 1+\vartheta \in J(n+2).$$

Für die erste Komponente von $\tilde{u}$ ergibt sich nach (3.7.5) mit $|\delta_1| \leqq \tau$:

$$\tilde{u}_1 = -\sigma\,\mathrm{gl}(|a_{1,1}| + \tilde{\mu}) = -\sigma(|a_{1,1}| + \mu(1+\epsilon))(1+\delta_1)$$

$$= -\sigma(|a_{1,1}|(1+\epsilon_1) + \mu(1+\epsilon_2)), \qquad 1+\epsilon_i \in J\left(\frac{n+4}{2}\right)$$

und wie oben

$$(3.7.7) \qquad \tilde{u}_1 = u_1(1+\gamma_1); \qquad 1+\gamma_1 \in J\left(\frac{n+4}{2}\right).$$

Für die restlichen $\tilde{u}_i$ schreiben wir

$$(3.7.7') \qquad \tilde{u}_i = a_{i,1} = u_i(1+\gamma_i); \qquad \gamma_i = 0 \quad (i = 2, \ldots, n).$$

Wir erhalten nach (3.7.5) wegen

$$\tilde{v}_1 = -gl\left(\frac{\sigma}{\tilde{\mu}}\right) = -\frac{\sigma}{\mu}(1+\eta)(1+\zeta_1), \qquad |\zeta_1| \leqq \tau$$

unmittelbar

$$\tilde{v}_1 = v_1(1+\alpha_1), \qquad 1+\alpha_1 \in J\left(\frac{n+4}{2}\right)$$

und weiter mit $|\zeta_i| \leqq \tau$ für $i = 2, \ldots, n$:

$$\tilde{v}_i = gl\left(\frac{a_{i,1}}{\tilde{N}}\right) = \frac{a_{i,1}}{N}(1+\vartheta)(1+\zeta_i) = v_i(1+\alpha_i); \quad 1+\alpha_i \in J(n+3),$$

insgesamt also

$$(3.7.8) \qquad \tilde{v}_i = v_i(1+\alpha_i); \quad 1+\alpha_i \in J(n+3) \qquad (i = 1, \ldots, n).$$

Für die Komponenten von $\tilde{p} =: (\tilde{p}_j)_1^n$ und für das nach (2.6.15, vi) zu berechnende $\tilde{q}(\in \mathbb{R})$, das wir mit $\tilde{p}_{n+1}$ bezeichnen wollen, notieren wir die Beziehungen

$$\tilde{p}_j = gl\left(\sum_{i=1}^{n} \tilde{u}_i a_{i,j}\right) = \sum_{i=1}^{n} u_i a_{i,j}(1+\gamma_i)(1+\epsilon_{i,j}) \qquad (j = 1, \ldots, n+1)$$

mit $1+\epsilon_{i,j} \in J(n)$ nach (3.7.3), γ_i gemäß (3.7.7) bzw. (3.7.7'), also

$$(3.7.9) \qquad \tilde{p}_j = \sum_{i=1}^{n} u_i(a_{i,j} + a_{i,j}\zeta_{i,j}); \qquad 1+\zeta_{i,j} \in J\left(\frac{3n+4}{2}\right).$$

Die erste der Spalten $c_j^{(2)} = (c_{i,j}^{(2)})_{i=1}^n$ $(j = 1, \ldots, n+1)$ von C_2 wird $\sigma\tilde{\mu}e_1$ gesetzt; die übrigen Spalten besitzen die Komponenten

$$c_{i,j}^{(2)} = gl(a_{i,j} - \tilde{v}_i\tilde{p}_j) \qquad (1 \leqq i \leqq n; \quad 2 \leqq j \leqq n+1),$$

also

$$(3.7.10) \qquad c_{i,j}^{(2)} = a_{i,j} - \tilde{v}_i\tilde{p}_j + a_{i,j}\eta_{i,j} - \tilde{v}_i\tilde{p}_j\rho_{i,j}; \; |\eta_{i,j}| \leqq \tau, \; 1+\rho_{i,j} \in J(2).$$

Die in (3.7.8)–(3.7.10) auftretenden Größen fassen wir zu folgenden Matrizen bzw. Spalten zusammen:

$$\left.\begin{aligned} D &:= \operatorname{diag}(\alpha_1, \ldots, \alpha_n), \\ z_j &:= (a_{i,j}\,\zeta_{i,j})_{i=1}^{n}, \\ r_j &:= (a_{i,j}\,\eta_{i,j})_{i=1}^{n}, \\ t_j &:= (\tilde{v}_i\,\tilde{p}_j\,\rho_{i,j})_{i=1}^{n} \end{aligned}\right\} \quad (j = 2, \ldots, n+1).$$

Hiermit erhalten wir aus (3.7.8)–(3.7.10) die Darstellungen

$$\tilde{v} = (I + D)v,$$
$$\tilde{p}_j = u^*(a_j + z_j) \qquad (2 \leqq j \leqq n+1)$$

sowie

$$c_j^{(2)} = \begin{cases} \sigma\tilde{\mu}\,e_1 & (j = 1), \\ a_j - (I + D)vu^*(a_j + z_j) + r_j + t_j & (j = 2, \ldots, n+1). \end{cases} \tag{3.7.11}$$

Zum Vergleich mit H_1C_1 beachten wir, daß H_1C_1 die Spalten $a_j - vu^*a_j$, insbesondere $\sigma\mu e_1$ als 1. Spalte besitzt, und schließen aus (3.7.11) für die Matrizen $F_1, d^{(1)}$ in Satz (3.7.4) die

(3.7.12) **Bemerkung.** Die Spalten $g_1, \ldots, g_{n+1}$ von $G_1 := (F_1, d^{(1)})$ lassen sich wie folgt darstellen:

$$g_1 = \sigma(\tilde{\mu} - \mu)e_1,$$
$$g_j = -Dvu^*a_j - (I + D)vu^*z_j + r_j + t_j \qquad (j = 2, \ldots, n+1).$$

Um den Beweis von Satz (3.7.4) zu Ende zu führen, wollen wir die g_j bezüglich der euklidischen Norm abschätzen. Zunächst erhalten wir für die in (3.7.8) bis (3.7.10) auftretenden Größen aus Hilfssatz (1.4.10) die Ungleichungen

$$|\alpha_i| \leqq \alpha, \quad |\zeta_{i,j}| \leqq \beta, \quad |\rho_{i,j}| \leqq \gamma \qquad (1 \leqq i \leqq n;\ 2 \leqq j \leqq n+1)$$

mit den positiven Konstanten

$$\alpha := (1 + \tau)^{n+3} - 1; \quad \beta := (1 + \tau)^{\frac{3n+4}{2}} - 1; \quad \gamma := (1 + \tau)^2 - 1.$$

Wegen $|\eta_{i,j}| \leqq \tau$ erhalten wir für $j \geqq 2$ offenbar

$$\|r_j\| \leqq \tau\,\|a_j\|, \tag{3.7.13}$$

ferner

$$\|t_j\| \leqq \gamma\,|\tilde{p}_j| \cdot \|\tilde{v}\| = \gamma\,|u^*(a_j + z_j)|\ \|(I + D)v\|.$$

Unter Benutzung der Schwarzschen Ungleichung und der Abschätzung

$$\|z_j\| \leqq \beta\,\|a_j\|$$

wird hierin

$$|u^*(a_j + z_j)| \leqq \|u\| \cdot \|a_j + z_j\| \leqq (1+\beta)\,\|a_j\| \cdot \|u\|\,,$$

außerdem

$$\|(I + D)v\| \leqq (1+\alpha)\,\|v\|\,.$$

Aus $N = \frac{1}{2}\,\|u\|^2$ – vgl. (2.6.15, iv) – folgt $\|u\| \cdot \|v\| = 2$ und daher

$$\|t_j\| \leqq 2\gamma(1+\alpha)(1+\beta)\,\|a_j\|\,. \tag{3.7.14}$$

Ähnlich folgern wir

$$\|Dvu^*a_j\| \leqq 2\alpha\,\|a_j\| \tag{3.7.15}$$

und

$$\|(I + D)vu^*z_j\| \leqq (1+\alpha)\,\|v\| \cdot \|u\| \cdot \|z_j\| \leqq 2(1+\alpha)\beta\,\|a_j\|\,. \tag{3.7.16}$$

Die Zusammenfassung von (3.7.13)–(3.7.17) liefert

$$\|g_j\| \leqq \{2[(1+\alpha)(1+\beta)(1+\gamma) - 1] + \tau\}\,\|a_j\| \qquad (j = 2, \ldots, n+1)\,.$$

Hieraus gewinnen wir mit einer Abschätzung nach Hilfssatz (1.4.10) für $j = 2, \ldots, n+1$ die Ungleichungen

$$\|g_j\| \leqq \frac{1}{1-(2{,}5n+6)\tau}\,(5n+15)\tau\,\|a_j\| \leqq 1{,}1\,(5n+15)\tau\,\|a_j\|\,. \tag{3.7.17}$$

Für die erste Spalte von G_1 notieren wir nach (3.7.5) und Hilfssatz (1.4.10), wobei wir $\mu = \|a_1\|$ beachten,

$$\|g_1\| = |\tilde{\mu} - \mu| \leqq \frac{1}{1-0{,}5n\tau}\,(0{,}5n+1)\tau\,\|a_1\| \leqq 1{,}1\,(0{,}5n+1)\tau\,\|a_1\|\,;$$

insbesondere ist demgemäß (3.7.17) auch für $j = 1$ erfüllt. Da sich die Frobeniusnorm von G_1 als

$$\|G_1\|_2 = \left(\sum_{j=1}^{n} \|g_j\|^2\right)^{\frac{1}{2}}$$

schreiben läßt, erhalten wir

$$\|G_1\|_2 \leqq 1{,}1\,(5n+15)\tau\left(\sum_{j=1}^{n} \|a_j\|^2\right)^{\frac{1}{2}} = 1{,}1\,(5n+15)\tau\,\|A\|_2\,;$$

außerdem ist die im Satz (3.7.4) behauptete Abschätzung von $\|d^{(1)}\| = \|g_{n+1}\|$ durch (3.7.17) unmittelbar gegeben, womit der Beweis des Satzes (3.7.4) abgeschlossen ist.

Beim ν-ten Eliminationsschritt $(2 \leqq \nu \leqq n-1)$ wird nur die $(n-\nu+1, n-\nu)$-Teilmatrix $C'_\nu = (A'_\nu, b^{(\nu)'})$ von C_ν einer Householdertransformation unterworfen, es gilt also

$$(3.7.18) \qquad C_{\nu+1} = H_\nu C_\nu + G_\nu = \begin{pmatrix} r_{1,1} & \cdots & & s_1 \\ 0 & \ddots & & \vdots \\ \vdots & & r_{\nu-1,\nu-1} \cdots & s_{\nu-1} \\ \vdots & & 0 & H'_\nu C'_\nu + G'_\nu \\ 0 & \cdots & 0 & \end{pmatrix}$$

Wir unterteilen $G_\nu = (F_\nu, d^{(\nu)})$, $G'_\nu = (F'_\nu, d^{(\nu)'})$ und erhalten wegen $\|F_\nu\|_2 = \|F'_\nu\|_2$, $\|A'_\nu\|_2 \leqq \|A_\nu\|_2$ und entsprechenden Beziehungen mit $d^{(\nu)}$ und $b^{(\nu)}$ aus Satz (3.7.4) unmittelbar die

(3.7.19) **Folgerung.** Für die nach dem ν-ten Householderschritt $(1 \leqq \nu \leqq n-1)$ gewonnenen Matrizen $C_{\nu+1} = (A_{\nu+1}, b^{(\nu+1)})$ gelten Darstellungen

$$A_{\nu+1} = H_\nu A_\nu + F_\nu, \qquad b^{(\nu+1)} = H_\nu b^{(\nu)} + d^{(\nu)}$$

mit

$$\|F_\nu\|_2 \leqq 1{,}1\,(5(n-\nu+1)+15)\,\tau\,\|A_\nu\|_2\,,$$

$$\|d^{(\nu)}\| \leqq 1{,}1\,(5(n-\nu+1)+15)\tau\,\|b^{(\nu)}\|\,.$$

Zur weiteren Abschätzung der $\|A_\nu\|_2$ und $\|b^{(\nu)}\|$ stützen wir uns auf den

(3.7.20) **Hilfssatz.** *Für eine unitäre* (n,n)*-Matrix* U *und eine beliebige* (n,n)*-Matrix* F *ist stets*

$$\|UF\|_2 = \|F\|_2\,.$$

Beweis. Für $x \in \mathbb{C}^n$ gilt stets $\|Ux\| = \|x\|$, also

$$\|UF\|_2^2 = \sum_{j=1}^{n} \|UFe_j\|^2 = \sum_{j=1}^{n} \|Fe_j\|^2 = \|F\|_2^2\,.$$

(3.7.21) **Hilfssatz.** *Mit*

$$\rho := \prod_{j=2}^{n} (1 + 1{,}1\,(5j+15)\,\tau)$$

gelten die Ungleichungen

$$\|A_\nu\|_2 \leqq \rho\,\|A\|_2\,; \qquad \|b^{(\nu)}\| \leqq \rho\,\|b\| \qquad (\nu = 1, \ldots, n)\,.$$

Beweis. Zunächst ist wegen $\rho > 1$ die Behauptung für $\nu = 1$ richtig; für $\nu \geqq 2$ folgern wir aus (3.7.19) und (3.7.20)

$$\begin{aligned} \|A_\nu\|_2 &\leqq \|H_{\nu-1}A_{\nu-1}\|_2 + \|F_{\nu-1}\|_2 = \|A_{\nu-1}\|_2 + \|F_{\nu-1}\|_2 \\ &\leqq (1 + 1{,}1\,(5(n-\nu+2)+15)\,\tau)\,\|A_{\nu-1}\|_2 \end{aligned}$$

und induktiv

$$\|A_\nu\|_2 \leqq \prod_{j=n-\nu+2}^{n} (1 + 1{,}1\,(5j+15)\,\tau)\,\|A\|_2 \leqq \rho\,\|A\|_2 .$$

Analog wird $\|b^{(\nu)}\|$ abgeschätzt.

Nach $(n-1)$ Householder-Schritten gewinnen wir

$$C_n = (A_n, b^{(n)}) =: (R, s) ,$$

wobei R eine obere Dreiecksmatrix ist. Gemäß (3.7.18) haben wir

$$C_n = H_{n-1} \cdot \ldots \cdot H_1 C_1 + H_{n-1} \cdot \ldots \cdot H_2 G_1 + H_{n-1} \cdot \ldots \cdot H_3 G_2 + \ldots + G_{n-1} .$$

Diese Gleichung multiplizieren wir von links mit $H_1 \cdot \ldots \cdot H_{n-1} = (H_{n-1} \cdot \ldots \cdot H_1)^{-1}$ und setzen

$$U_j := H_1 H_2 \ldots H_j \qquad (j = 1, \ldots, n-1) .$$

Es ergibt sich

$$U_{n-1} C_n = C_1 + \sum_{j=1}^{n-1} U_j G_j$$

und nach Abtrennen der letzten Spalte

$$(3.7.22) \qquad \begin{cases} U_{n-1} R = A + \sum_{j=1}^{n-1} U_j F_j , \\ U_{n-1} s = b + \sum_{j=1}^{n-1} U_j d^{(j)} . \end{cases}$$

Wenn R invertierbar ist, betrachten wir als numerische Lösung des Gleichungssystems $Ax = b$ die mit Gleitkommarechnung ermittelte Lösung von $Rx = s$. Nach (3.5.7), (3.5.9) erfüllt diese Lösung $\tilde{x}$ ein gestörtes Gleichungssystem

$$(R + \Delta R)\,\tilde{x} = s$$

mit

$$\|\Delta R\|_2 \leqq 1{,}1\, n\tau\, \|R\|_2 .$$

Es folgt

$$U_{n-1}(R + \Delta R)\,\tilde{x} = U_{n-1} s,$$

durch Einsetzen der Gleichungen aus (3.7.22) erhalten wir weiter

$$\left(A + \sum_{j=1}^{n-1} U_j F_j + U_{n-1} \Delta R\right) \tilde{x} = b + \sum_{j=1}^{n-1} U_j d^{(j)} .$$

Aus dieser Darstellung folgern wir den

(3.7.23) **Satz.** *Falls unter der Voraussetzung* (3.7.1) *das Householder-Verfahren bei t-stelliger Gleitkommarechnung eine invertierbare obere Dreiecksmatrix* R *liefert, so erfüllt die numerisch gewonnene Lösung* $\tilde{x}$ *von* $Ax = b$ *ein Gleichungssystem*

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

mit

$$\|\Delta A\|_2 \leqq 2{,}75\, \rho\, (n+4)^2\, \|A\|_2\, \tau\,,$$
$$\|\Delta b\| \leqq 2{,}75\, \rho\, (n+4)^2\, \|b\|\, \tau\,.$$

Beweis. Es ist nach (3.7.20) zunächst

$$\|\Delta A\|_2 = \Big\| \sum_{j=1}^{n-1} U_j F_j + U_{n-1} \Delta R \Big\|_2 \leqq \sum_{j=1}^{n-1} \|F_j\|_2 + \|\Delta R\|_2$$

und nach (3.7.21), (3.7.19)

$$\|\Delta R\|_2 \leqq 1{,}1\, n\tau\, \|R\|_2 = 1{,}1\, n\tau\, \|A_n\|_2 \leqq 1{,}1\, \rho\, n\tau\, \|A\|_2\,,$$

sowie

$$\|F_j\|_2 \leqq 1{,}1 (5(n-j+1)+15)\tau\, \|A_j\|_2 \leqq 1{,}1\, \rho\, (5(n-j+1)+15)\tau\, \|A\|_2\,,$$

insgesamt also

$$\|\Delta A\|_2 \leqq 1{,}1\, \rho \left[\sum_{j=2}^{n} (5j+15) + n \right] \tau\, \|A\|_2\,.$$

Ebenso ergibt sich

$$\|\Delta b\| \leqq 1{,}1\, \rho \left[\sum_{j=2}^{n} (5j+15)\tau \right] \|b\|\,.$$

Eine Zwischenrechnung liefert

$$\sum_{j=2}^{n} (5j+15) + n = 2{,}5\,(n^2 + 7{,}4\,n - 6) \leqq 2{,}5\,(n+4)^2\,,$$

womit der Beweis vollendet ist.

Bei hinreichender Rechengenauigkeit kann der Faktor $1{,}1\,\rho$ durch einen Wert, der sehr nahe bei 1 liegt, ersetzt werden; daher ist der Effekt der Rundungsfehler etwa der gleiche, als hätte man die Koeffizienten von A und b mit relativen Fehlern vom Betrag $\leqq 2{,}5\, n^2 \tau$ gestört. Der Fehler bei der numerischen Durchführung der Gauß-Elimination entspricht nach Satz (3.6.27) im Fall $\hat{M}\hat{U} \approx \hat{A}$ relativen Fehlern der Eingangsdaten vom Betrag $\leqq 2n\tau$. Da die Bedingung $\hat{M}\hat{U} \approx \hat{A}$ in der Praxis meistens erfüllt ist, beobachtet man im allgemeinen bei der Gauß-

Elimination einen etwas kleineren Rundungsfehler als beim Householder-Verfahren; im Zahlenbeispiel (2.6.18) liefern beide Verfahren gleich gute Lösungen. Theoretisch haben wir nach (3.6.30) die Matrix $\hat{M}\hat{U}$ mit Hilfe der Größe g, die eventuell stark anwachsen kann, abgeschätzt; demgegenüber haben wir für das Householder-Verfahren eine günstigere a priori Fehlerschranke gefunden.

Schließlich schätzen wir noch $\tilde{x} - x$ bezüglich der euklidischen Norm ab.

(3.7.24) **Satz.** *Es gelte die Voraussetzung* (3.7.1), *ferner*

$$2{,}75\,\rho(n+4)^2\,\|A\|_2\cdot\|A^{-1}\|_S\cdot\tau < 1\,.$$

Dann ist die mit dem Householder-Verfahren bei t*-stelliger Gleitkomma-Arithmetik berechnete obere Dreiecksmatrix* R *invertierbar, und für den Fehler der numerisch gewonnenen Lösung* $\tilde{x}$ *von* $Ax = b$ *gilt die Abschätzung*

$$\frac{\|\tilde{x}-x\|}{\|x\|} \leqq \frac{\kappa(A)}{1-2{,}75\,\rho(n+4)^2\tau\,\|A\|_2\,\|A^{-1}\|_S}\,2{,}75\,\rho(n+4)^2(\sqrt{n}+1)\tau \approx 2{,}5\,n^{2{,}5}\tau\cdot\kappa(A)$$

wobei $\kappa(A) := \|A\|_S\,\|A^{-1}\|_S$ *bedeute.*

Beweis. Wir erinnern zunächst daran, daß $\|A\|_S$ die Operatornorm bezüglich der euklidischen Norm ist. Wie in Aufgabe 3.7 gezeigt werden soll, gelten für jede (n,n)-Matrix A die Ungleichungen

$$(3.7.25)\qquad \|A\|_S \leqq \|A\|_2 \leqq \sqrt{n}\,\|A\|_S\,.$$

Weiter haben wir, wie im Beweis zu (3.7.23) gezeigt,

$$\|A^{-1}\|_S \sum_{j=1}^{n-1}\|F_j\|_2 \leqq \|A^{-1}\|_S\,2{,}75\,\rho(n+4)^2\,\|A\|_2\cdot\tau < 1\,,$$

Wegen

$$\|\sum_{j=1}^{n-1} U_j F_j\|_S \leqq \sum_{j=1}^{n-1}\|U_j F_j\|_2 = \sum_{j=1}^{n-1}\|F_j\|_2\,;$$

ist daher auch

$$\|A^{-1}\|_S\,\|\sum_{j=1}^{n-1} U_j F_j\|_S < 1\,.$$

Hieraus folgt nach Hilfssatz (3.4.1) auf Grund der Darstellung (3.7.22) die Invertierbarkeit von $U_{n-1}R$ und somit auch von R.

Zur Fehlerabschätzung stützen wir uns auf den Satz (3.4.2). Hierzu beachten wir die unmittelbar aus (3.7.25) abzulesende Ungleichung

$$\frac{\|\Delta A\|_S}{\|A\|_S} \leqq \sqrt{n}\,\frac{\|\Delta A\|_2}{\|A\|_2}$$

sowie die Abschätzungen aus Satz (3.7.23).

Übungsaufgaben zum 3. Kapitel

Aufgabe 3.1. Es sei R eine Menge, $d: R \times R \to \mathbb{R}$ Pseudometrik in R, d.h.

(i) $d(x,y) \geqq 0, \quad d(x,x) = 0;$

(ii) $d(x,y) = d(y,x);$

(iii) $d(x,y) \leqq d(x,z) + d(z,y)$ für $x, y, z \in R.$

Es sei definiert

$$x \sim y : \iff d(x,y) = 0.$$

(α) Man zeige, daß $\sim$ eine Äquivalenzrelation in R ist.

Für jedes $x \in R$ bezeichne $\hat{x} := \{y \in R : x \sim y\}$; $\hat{R} = R/\sim := \{\hat{x} : x \in R\}$.

Wir definieren $\hat{d}: \hat{R} \times \hat{R} \to \mathbb{R}$ durch

$$\hat{d}(\hat{x}, \hat{y}) := d(x,y) \quad \text{für} \quad x \in \hat{x},\ y \in \hat{y}.$$

(β) Man zeige, daß $\hat{d}$ wohldefiniert und $(\hat{R}, \hat{d})$ metrischer Raum ist.

Aufgabe 3.2. Man zeige, daß die Funktionen $x_n(t) := |t|^{1+\frac{1}{n}}$ $(n = 1, 2, 3, \ldots)$ in $[-1, 1]$ stetig differenzierbar sind und gleichmäßig gegen $x(t) = |t|$ konvergieren.

Aufgabe 3.3. Es sei $[a,b] \subseteq \mathbb{R}$ kompaktes Intervall, $t_0 \in [a,b]$ fest.
Für $x \in C_1[a,b]$ bezeichne

$$\|x\|_2 := |x(t_0)| + \max_{a \leqq t \leqq b} |x'(t)|, \qquad \|x\|_3 := \max_{a \leqq t \leqq b} |x(t)| + \max_{a \leqq t \leqq b} |x'(t)|.$$

Man zeige:

(i) $\| \ \|_2$ und $\| \ \|_3$ sind Normen in $C_1[a,b]$;

(ii) die angegebenen Normen sind zueinander äquivalent; dazu gebe man positive Konstanten α, β an, so daß für alle $x \in C_1[a,b]$

$$\|x\|_2 \leqq \alpha \cdot \|x\|_3, \qquad \|x\|_3 \leqq \beta \|x\|_2$$

gilt.

(iii) Bezüglich beider Normen ist $C_1[a,b]$ vollständig.

Aufgabe 3.4. Im $\mathbb{K}^n$ $(n \geqq 1)$ seien für $1 \leqslant p \leqslant \infty$ die Normen $\|x\|_p$ nach (3.2.3) definiert. Man gebe (von p, p′ und n abhängige) Konstanten $\alpha, \beta > 0$ an mit den Eigenschaften

$$\begin{cases} \forall\, x \in \mathbb{K}^n \quad \|x\|_p \leqslant \alpha \|x\|_{p'}, \ \|x\|_{p'} \leqslant \beta \|x\|_p, \\ \exists\, x_1, x_2 \in \mathbb{K}^n, \neq 0 \quad \|x_1\|_p = \alpha \|x_1\|_{p'}, \ \|x_2\|_{p'} = \beta \|x_2\|_p. \end{cases}$$

Dabei sind folgende Fälle zu unterscheiden:

(i) $1 = p < p' < \infty$, (ii) $p = 1$, $p' = \infty$,

(iii) $1 < p < p' < \infty$, (iv) $1 < p < p' = \infty$.

Aufgabe 3.5. Es sei $1 \leqq p \leqq \infty$, dazu $1 \leqq q \leqq \infty$ mit $\frac{1}{p} + \frac{1}{q} = 1$. Zu fest vorgegebenem $a \in \mathbb{K}^n$ sei $\varphi \in \mathrm{Hom}(\mathbb{K}^n, \mathbb{K})$ durch

$$\varphi(x) := a^t x$$

definiert. Dann ist φ bezüglich der Normen $\| \ \|_p$ im $\mathbb{K}^n$, Absolutbetrag in $\mathbb{K}$ stetig; man zeige

$$|\varphi| = \|a\|_q .$$

Aufgabe 3.6. Man beweise die Aussagen (i), (ii) und (iv) von Hilfssatz (3.3.9).

Aufgabe 3.7. Man zeige für $A \in M(n \times n, \mathbb{C})$ die Ungleichungen

$$\|A\|_S \leqq \|A\|_2 \leqq \sqrt{n}\, \|A\|_S .$$

Hinweis: Man zeigt und verwendet $\|Ae_i\|_2 \leqq \|A\|_S$ $(i = 1, \ldots, n)$.

Aufgabe 3.8. Es sei $A \in M(n \times n, \mathbb{C})$ normal; man zeige

$$\|A\|_S = \rho(A) = \max \left\{ \left| \frac{(Ax, x)}{(x, x)} \right| : \ x \in \mathbb{C}^n, \ \neq 0 \right\} .$$

Hinweis: A ist normal genau dann, wenn eine Orthonormalbasis des $\mathbb{C}^n$ aus Eigenvektoren zu A existiert.

Aufgabe 3.9. In $M(n \times n, \mathbb{C})$ bezeichne $| \ |$ die Operatornorm zu einer Norm im $\mathbb{C}^n$. Es seien $A, X_0 \in M(n \times n, \mathbb{C})$, A invertierbar, X_0 Näherung von A^{-1} mit

$$AX_0 = I + F_0, \quad |F_0| < 1 .$$

Man zeige

(i) $$|A^{-1}| \leqq \frac{|X_0|}{1 - |F_0|} ,$$

(ii) die durch $X_{k+1} = X_k + X_k(I - AX_k)$ $(k = 0, 1, 2, \ldots)$ definierte Folge konvergiert gegen A^{-1}.

Hinweis: Man zeigt $I - AX_{k+1} = (I - AX_k)^2$.

Aufgabe 3.10. Es sei $| \ |$ eine beliebige Norm im $\mathbb{C}^n$, $A \in M(n \times n, \mathbb{C})$ invertierbar. Man gebe $b, \Delta b \in \mathbb{C}^n$, $\neq 0$ so an, daß für $x, \Delta x$ mit

$$Ax = b, \qquad A(x + \Delta x) = b + \Delta b$$

die Gleichung

$$\frac{|\Delta x|}{|x|} = \rho(A)\, \rho(A^{-1}) \, \frac{|\Delta b|}{|b|}$$

erfüllt ist.

Aufgabe 3.11. Man zeige $p_n \leqq \sqrt{n}\, f(n)$ wie in Satz (3.6.23) für den Fall, daß A eine (n,k)-Matrix mit $k > n$ und $\mathrm{rg}\, A = n$ ist.

Aufgabe 3.12.

(i) Man zeige, daß die Ungleichung

$$\hat{\Delta A} \leqq \frac{2{,}1\, n\tau}{1 - n\tau}\, \hat{M}\hat{U}$$

in Satz (3.6.27) auch für die Gauß-Elimination bei nicht halbmaximaler bzw. maximaler Pivotwahl gilt.

(ii) Für das Gleichungssystem der Aufgabe 2.1 im Fall der diagonalen Pivotwahl rechne man $\hat{M}\hat{U} \approx 6{,}22 \cdot 10^3$ nach und überzeuge sich, daß hiermit keine Abschätzung nach Satz (3.4.2) möglich ist.

(iii) Für die Lösung der Aufgabe 2.1 mit halbmaximaler Pivotwahl zeige man

$$\|\Delta A\|_\infty \leqq 0{,}032\, \|A\|_\infty ,$$
$$\|A^{-1}\|_\infty \, \|A\|_\infty < 1$$

und folgere die Fehlerabschätzung

$$\frac{\|\tilde{x} - x\|_\infty}{\|x\|_\infty} \leqq 0{,}270 .$$

Dazu schätze man $\|A^{-1}\|_\infty$ mit Hilfe der im Beispiel (2.5.9) ermittelten Näherung und Aufgabe 3.9 ab.

Beispiele aus Büchern und Briefen: Wie man zu Geld kommen kann

(Abb.: Fouqué in Münze)

«‹Ei›, schnarrte das Kerlchen . . .

. . . und lachte noch viel entsetzlich dummer, ‹schenkt mir doch erst ein Trinkgeld, denn ich hab ja Euer Rösselein aufgefangen; lägt Ihr doch ohne mich samt Euerm Rösselein in der Steinkluft da unten; hu!› – ‹Schneide nur keine Gesichter weiter›, sagte ich, ‹nimm dein Geld hin, wenn du auch lügst; denn siehe, der gute Bach dorten hat mich gerettet, nicht aber du, höchst ärmlicher Wicht!› – Und zugleich ließ ich ein Goldstück in seine wunderliche Mütze fallen, die er bettelnd vor mir abgezogen hatte.»

Ganz einfach also: Man muß nur ein Wicht sein und einen Ritter, der von Bertalda zu Undine unterwegs ist, vor dem Absturz bewahren.

Summa summarum: In Büchern und Briefen kann man auf wunderlichste Weise zu Geld kommen, doch nähm' es wunder, wenn einer mit Büchern und Briefen zu Geld käme, es seien denn Sparbücher und Pfandbriefe.

4. Lineare Optimierung

Optimierungsaufgaben treten als Organisationsprobleme in Wirtschaft und Technik, aber auch bei naturwissenschaftlichen und mathematischen Fragestellungen auf. Im Lauf der letzten Jahre hat sich die Optimierungstheorie zu einer eigenen mathematischen Disziplin entwickelt. Ihre heutige Bedeutung verdankt diese Theorie vor allem den Rechenmaschinen, die die praktische Lösung der gestellten Optimierungsaufgaben überhaupt erst ermöglichen. Wir wollen hier den einfachsten Fall, nämlich den der linearen Optimierung behandeln.

4.1. Vorbemerkungen

Wir betrachten die folgende

(4.1.1) **Optimierungsaufgabe.** Es sei $C \in M(n \times m, \mathbb{R})$ mit $m = q + n$, $q \geq 1$, $\operatorname{rg} C = n$; ferner seien $f \in \mathbb{R}^m$, $b \in \mathbb{R}^n$ und $\vartheta \in \mathbb{R}$. Dann ist ein $x \in \mathbb{R}^m$ gesucht, so daß der Wert der *Zielfunktion*

(i) $\quad z(x) := f^t x + \vartheta$

maximal wird unter den *Restriktionen*

(ii) $$\begin{cases} Cx = b\,, \\ x \geq 0\,. \end{cases}$$

Dabei ist die Ungleichung $x \geq 0$ komponentenweise gemeint, vgl. (3.5.5).

Ein wichtiger Spezialfall ist die folgende Aufgabe, in der $A \in M(n \times q, \mathbb{R})$, $d \in \mathbb{R}^q$, $b \in \mathbb{R}^n$ vorgegeben sind:

(4.1.2) Maximiere

(i) $\quad z(x) := d^t x \qquad (x \in \mathbb{R}^q)$

unter den Restriktionen

(ii) $$\begin{cases} Ax \leqq b\,, \\ x \geq 0\,. \end{cases}$$

Zur Reduktion auf die Form (4.1.1) definieren wir eine Spalte $x_s \in \mathbb{R}^n$, deren Komponenten wir als Schlupfvariable bezeichnen, durch

(*) $x_s := b - Ax$

und weiter

$$(4.1.3)\quad \begin{cases} x_e := \begin{pmatrix} x \\ x_s \end{pmatrix} \in \mathbb{R}^{q+n}, \\ C := (A, I_n)_{(n,\,q+n)}, \\ f^t := (d^t, \underbrace{0 \ldots 0}_{n}), \quad \vartheta = 0. \end{cases}$$

Wenn ein $x_e \in \mathbb{R}^{q+n}$ die Bedingungen $Cx_e = b$, $x_e \geqq 0$ erfüllt, so gilt offenbar $Ax \leqq b$, $x \geqq 0$; umgekehrt liefert ein $x \in \mathbb{R}^q$ mit der Eigenschaft (4.1.2), (ii) vermöge (*) ein $x_e \in \mathbb{R}^{q+n}$ mit (4.1.1), (ii); die Werte der Zielfunktionen sind für entsprechende x und x_e gleich.

Wir können auch solche Probleme auf die Form (4.1.1) bringen, in denen die Vorzeichenbedingung $x \geqq 0$ fehlt oder nur für einen Teil der Komponenten von x gefordert wird; dazu verweisen wir auf Aufgabe 4.5.

Optimierungsaufgaben der Form (4.1.2) lassen sich im Fall $q = 2$ graphisch lösen. Dazu betrachten wir das

(4.1.4) **Zahlenbeispiel.** Man maximiere

$$z(x) = 20\,x_1 + 60\,x_2$$

unter den Restriktionen

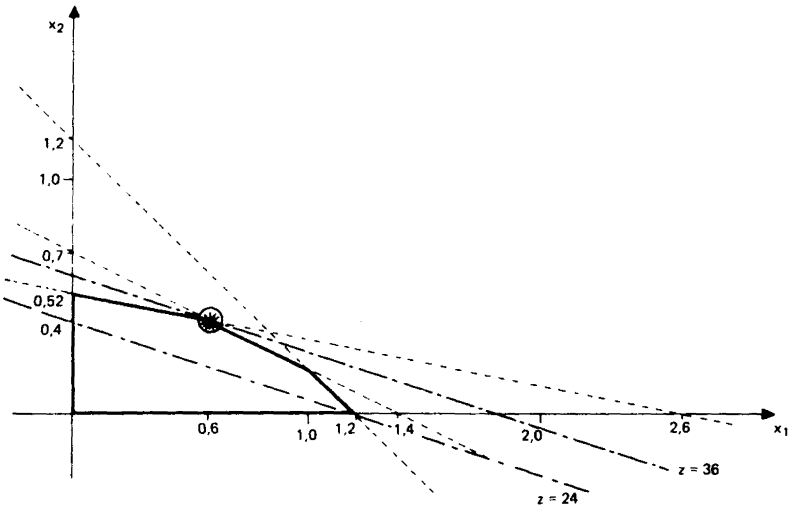
$$\begin{aligned} x_1 + \quad x_2 &\leqq 1{,}2\,, \\ 5\,x_1 + 10\,x_2 &\leqq 7{,}0\,, \\ 2\,x_1 + 10\,x_2 &\leqq 5{,}2\,, \\ x_1 \geqq 0\,, \quad x_2 &\geqq 0\,. \end{aligned}$$

Jede der aufgeführten Ungleichungen definiert eine Halbebene des $\mathbb{R}^2$, die auf einer Seite einer gewissen Geraden liegt; somit beschreiben die Restriktionen ein konvexes Polyeder, das wir in der folgenden Skizze fett eingerahmt haben. Durch

$$z = 20\,x_1 + 60\,x_2$$

ist bei variablem z eine Schar von parallelen Geraden gegeben, von denen wir zwei Vertreter ($z = 24$ und $z = 36$) eingetragen haben. Wir lösen die Optimierungsaufgabe, indem wir (wie es in diesem Beispiel möglich ist) eine solche Gerade aus-

wählen, die bei maximalem z einen Punkt mit dem Polyeder der Restriktionen gemeinsam hat. Offenbar wird $z = 36$ der maximale Wert der Zielfunktion, dazu $(x_1 ; x_2) = (0{,}6; 0{,}4)$ die eindeutige Lösung der Aufgabe.



Graphische Lösung von (4.1.4)

Zur Optimierungsaufgabe (4.1.1) vereinbaren wir folgende

(4.1.5) **Definition**.

(i) Es sei

$$\zeta := \{x \in \mathbb{R}^m : Cx = b,\ x \geqq 0\},$$

jedes $x \in \zeta$ heiße *zulässige Lösung* von (4.1.1).

(ii) $x^0 \in \mathbb{R}^m$ heiße *optimale Lösung* von (4.1.1) genau dann, wenn

$$x^0 \in \zeta, \quad z(x^0) = \max \{z(x) : x \in \zeta\}$$

gilt.

(iii) Im Fall, daß eine optimale Lösung existiert, bezeichne

$$z_{opt} := \max \{z(x) : x \in \zeta\}.$$

Es seien $c_1, \ldots, c_m$ die Spalten von C. Wegen $\operatorname{rg} C = n$ ist $\mathbb{R}^n = \operatorname{span}(c_1, \ldots, c_m)$. Folglich existieren $i_1, i_2, \ldots, i_n \in \{1, \ldots, m\}$, so daß $(c_{i_1}, \ldots, c_{i_n})$ eine Basis des $\mathbb{R}^n$ ist. Dazu existieren eindeutig $\xi_{i_1}, \ldots, \xi_{i_n} \in \mathbb{R}$ mit

$$b = \sum_{\nu=1}^{n} \xi_{i_\nu} c_{i_\nu}.$$

Ergänzt man durch $\xi_i = 0$ für $i \in \{1, \ldots, m\} \setminus \{i_1, \ldots, i_n\}$, so ist $x = (\xi_i)_{i=1}^m$ eine Lösung des Gleichungssystems $Cx = b$. In diesem Sinne notieren wir

(4.1.6) **Definition**.

(i) Es sei $x = (\xi_i)_1^m \in \mathbb{R}^m$ Lösung von $Cx = b$. Dann heißt x *Basislösung* von $Cx = b$, wenn Indizes $i_1, \ldots, i_n \in \{1, \ldots, m\}$ existieren, so daß die Spalten $c_{i_1}, \ldots, c_{i_n}$ linear unabhängig sind und für $i \notin \{i_1, \ldots, i_n\}$ $\xi_i = 0$ gilt.

(ii) Eine Basislösung x von $Cx = b$ heißt *ausgeartet*, wenn für mindestens ein $j \in \{i_1, \ldots, i_n\}$ $\xi_j = 0$ gilt, andernfalls heißt die Basislösung *nicht-ausgeartet*.

(iii) Zur Optimierungsaufgabe (4.1.1) bezeichne

$$\zeta_B := \{x \in \mathbb{R}^m : x \text{ Basislösung von } Cx = b,\ x \geq 0\} \subseteq \zeta$$

die Gesamtheit der *zulässigen Basislösungen*.

Hierzu zeigen wir die

(4.1.7) **Bemerkungen**.

(i) Zu je n linear unabhängigen Spalten von C existiert genau eine Basislösung.

(ii) Die Zahl der Basislösungen von $Cx = b$ ist höchstens $\binom{m}{n}$.

(iii) Es sei $x = (\xi_i)_1^m \in \mathbb{R}^m$ mit $Cx = b$. Dann ist x Basislösung von $Cx = b$ genau dann, wenn ein $k \leq n$ und Indizes $i_1, \ldots, i_k \in \{1, \ldots, m\}$ existieren, so daß $c_{i_1}, \ldots, c_{i_k}$ linear unabhängig sind und für $i \notin \{i_1, \ldots, i_k\}$ $\xi_i = 0$ gilt.

(iv) Ist L eine invertierbare (n, n)-Matrix und $C' = LC$, $b' = Lb$, so ist jede Basislösung von $Cx = b$ auch Basislösung von $C'x = b'$ und umgekehrt.

(v) Besitzt C die Gestalt

$$CQ = (A, I_n) \qquad (A \in M(n \times q, \mathbb{R}))\,,$$

mit einer (m, m)-Permutationsmatrix Q, so ist

$$x := Q \begin{pmatrix} 0 \\ b \end{pmatrix} \in \mathbb{R}^m$$

Basislösung von $Cx = b$. x ist nicht-ausgeartet genau dann, wenn sämtliche Komponenten von b von Null verschieden sind.

Beweis. Die Aussage (i) ist bereits gezeigt; es folgt, daß die Zahl der Basislösungen höchstens so groß ist wie die Anzahl aller Mengen, die aus n linear unabhängigen Spalten von C bestehen. – Wir kommen zu Aussage (iii): ist x Basislösung von $Cx = b$, so können wir $k = n$ wählen und (4.1.6), (i) anwenden. Umgekehrt setzen wir $k \leqq n$ und $i_1, \ldots, i_k$ mit $c_{i_1}, \ldots, c_{i_k}$ linear unabhängig, $\xi_i = 0$ für $i \notin \{i_1, \ldots, i_k\}$ voraus. Dann können wir wegen $\operatorname{rg} C = n$ die Spalten $c_{i_1}, \ldots, c_{i_k}$ durch Hinzunahme weiterer Spalten zu einer Basis $(c_{i_1}, \ldots, c_{i_n})$ des $\mathbb{R}^n$ ergänzen; wegen $\xi_i = 0$ für $i \notin \{i_1, \ldots, i_n\}$ ist x die zugehörige Basislösung von $Cx = b$. In (iv) besitzt LC die Spalten $Lc_1, \ldots, Lc_m$; wegen der Invertierbarkeit von L sind stets $c_{i_1}, \ldots, c_{i_n}$ linear unabhängig genau dann, wenn $Lc_{i_1}, \ldots, Lc_{i_n}$ linear unabhängig sind. In (v) rechnen wir $Cx = b$ sofort nach. Wir wollen nun zeigen, daß $x =: (\xi_i)_1^m$ Basislösung zu den Spalten $c_{\sigma(q+1)}, \ldots, c_{\sigma(q+n)}$ von C ist, wobei $\sigma \in S_m$ durch $Qe_i = e_{\sigma(i)}$ $(i = 1, \ldots, m)$ definiert ist. Wir haben zunächst

$$c_{\sigma(q+j)} = Ce_{\sigma(q+j)} = CQe_{q+j} = e_j \qquad (j = 1, \ldots, n),$$

also sind die $c_{\sigma(q+j)}$ $(j = 1, \ldots, n)$ linear unabhängig; andererseits folgt aus $Q^t x = \binom{0}{b}$ unmittelbar

$$0 = e_j^t Q^t x = e_{\sigma(j)}^t x = \xi_{\sigma(j)} = 0 \qquad (j = 1, \ldots, q),$$

d.h. $\xi_i = 0$ für $i \notin \{\sigma(q+1), \ldots, \sigma(q+n)\}$.

Im Fall, daß (4.1.1) aus der Optimierungsaufgabe (4.1.2) entstanden ist, entsprechen die zulässigen Basislösungen gerade den „Ecken" des durch die Restriktionen (4.1.2), (ii) beschriebenen Polyeders des $\mathbb{R}^q$; hierzu verweisen wir auf die Aufgabe 4.1.

(4.1.8) **Satz.** *Es sei* $x^0 \in \zeta$ *und*

$$z_0 := z(x^0) = f^t x^0 + \vartheta .$$

Dann gilt entweder

(1) $\sup \{z(x) : x \in \zeta\} = \infty$,

oder

(2) *es existiert* $x^1 \in \zeta_B$ *mit*

$$z_1 := f^t x^1 + \vartheta \geqq z_0 .$$

Beweis. Es sei $x^0 =: (x_i)_1^m$ mit $x_{j_1}, \ldots, x_{j_k} > 0$; $x_i = 0$ für $i \notin \{j_1, \ldots, j_k\}$, wobei $0 \leqq k \leqq m$ möglich ist. Ohne Einschränkung können wir

$$x_1, \ldots, x_k > 0; \qquad x_{k+1} = \ldots = x_m = 0$$

annehmen. Es seien $c_1, \ldots, c_m$ die Spalten von C; dann unterscheiden wir

(I) $k = 0$, oder die $c_1, \ldots, c_k$ sind linear unabhängig,

(II) die $c_1, \ldots, c_k$ sind linear abhängig.

Im ersten Fall ist $0 \leqq k \leqq n$ und nach (4.1.7, iii) x^0 selbst Basislösung von $Cx = b$; im zweiten Fall zeigen wir, daß entweder (1) eintritt oder ein $y = (y_i)_1^m \in \zeta$ existiert, das folgende Eigenschaften besitzt

$$\begin{cases} y_i = 0 & \text{für } i = k+1, \ldots, m, \\ y_{j_0} = 0 & \text{für mindestens ein } j_0 \in \{1, \ldots, k\}, \\ f^t y \geqq f^t x^0 . \end{cases}$$

Hat man ein derartiges y gefunden, so stellt man die vorangehenden Überlegungen bei verkleinertem k (!) bezüglich dieses y an; nach spätestens k Schritten ist die Behauptung des Satzes (4.1.8) bewiesen.

Im Fall (II) gibt es $\alpha_1, \ldots, \alpha_k \in \mathbb{R}$, nicht alle Null, mit

$$(*) \qquad \sum_{j=1}^{k} \alpha_j c_j = 0 .$$

Mit den Komponenten von f bezeichnen wir

$$\gamma := \sum_{j=1}^{k} \alpha_j f_j .$$

Da mit $(\alpha_j)_1^k$ auch $(-\alpha_j)_1^k$ der Gleichung (*) genügt, nehmen wir ohne Einschränkung folgendes an:

$$(**) \quad \begin{cases} \text{es sei stets } \gamma \leqq 0 ; \\ \text{im Fall } \gamma = 0 \text{ existiere ein } j \in \{1, \ldots, k\} \text{ mit } \alpha_j > 0. \end{cases}$$

Für $\lambda \in \mathbb{R}$ definieren wir $y(\lambda) = (y_i(\lambda))_1^m \in \mathbb{R}^m$ durch

$$y_i(\lambda) = x_i - \lambda \alpha_i \quad (i = 1, \ldots, k); \qquad y_i(\lambda) = 0 \quad (i = k+1, \ldots, m)$$

und erhalten

$$Cy(\lambda) = \sum_{j=1}^{m} y_j(\lambda) c_j = \sum_{j=1}^{k} x_j c_j - \lambda \sum_{j=1}^{k} \alpha_j c_j = b ,$$

$$z(y(\lambda)) = f^t y(\lambda) + \vartheta = \sum_{j=1}^{k} f_j y_j(\lambda) + \vartheta = z_0 - \lambda \gamma .$$

Es ist $y(\lambda) \in \zeta$ genau dann, wenn $y_i(\lambda) \geqq 0$ für alle $i = 1, \ldots, k$ gilt. Sind alle $\alpha_i \leqq 0$, so ist $y(\lambda) \in \zeta$ für jedes $\lambda > 0$. In diesem Fall haben wir nach (**) $\gamma < 0$, daher nimmt $z(y(\lambda))$ für $\lambda \to \infty$ beliebig große Werte an: es tritt der Fall (1) ein.

Wenn es Indizes $j \in \{1, \ldots, k\}$ mit $\alpha_j > 0$ gibt, definieren wir

$$\lambda_0 := \min \left\{ \frac{x_j}{\alpha_j} : j \in \{1, \ldots, k\}, \ \alpha_j > 0 \right\} \qquad (> 0)$$

und wählen hierzu $j_0 \in \{1, \ldots, k\}$ mit

$$\lambda_0 = \frac{x_{j_0}}{\alpha_{j_0}} .$$

Wir erhalten für $j = 1, \ldots, k$

$$y_j(\lambda_0) = x_j - \lambda_0 \alpha_j \quad \begin{cases} > 0 , & \text{falls } \alpha_j \leqq 0 , \\ \geqq 0 , & \text{falls } \alpha_j > 0 , \\ = 0 & \text{für } j = j_0 . \end{cases}$$

Außerdem gilt wegen $\gamma \leqq 0$

$$f^t y(\lambda_0) + \vartheta = z_0 - \lambda_0 \gamma \geqq z_0 = f^t x^0 + \vartheta .$$

Insgesamt hat $y := y(\lambda_0)$ die gewünschten Eigenschaften.
Als unmittelbare Folgerung gewinnen wir den

(4.1.9) **Satz.** *Unter der Voraussetzung*

$$\zeta \neq \emptyset , \qquad \sup \{z(x) : x \in \zeta\} < \infty$$

besitzt (4.1.1) *eine optimale Lösung. Es existiert sogar ein* $x^0 \in \zeta_B$ *mit*

$$z(x^0) = z_{opt} = \max \{z(x) : x \in \zeta\} ,$$

d. h. eine optimale Basislösung.

Beweis. Wir haben nach dem vorangehenden Satz

$$\sup \{z(x) : x \in \zeta\} \leqq \sup \{z(x) : x \in \zeta_B\} ;$$

wegen $\zeta_B \subseteq \zeta$ gilt sogar das Gleichheitszeichen. Nach (4.1.7), (ii) ist ζ_B eine endliche Menge, folglich existiert

$$\max \{z(x) : x \in \zeta_B\} = \max \{z(x) : x \in \zeta\} = z_{opt} .$$

4.2. Simplex-Verfahren

Prinzipiell könnte man die Optimierungsaufgabe (4.1.1) lösen, indem man alle Basislösungen von $Cx = b$ bestimmt. Da ein solches Vorgehen zu zeitraubend wäre, werden im Simplex-Verfahren nacheinander zulässige Basislösungen so konstruiert, daß die Werte der Zielfunktion bei jedem Schritt wachsen. Zur Beschreibung des Verfahrens benötigen wir noch Hilfsmittel zur Darstellung und Transformation von Optimierungsaufgaben.

Offenbar wird (4.1.1) durch folgende $(n + 1, m + 1)$-Matrix vollständig beschrieben:

$$(4.2.1) \qquad \Omega := \left(\begin{array}{c|c} C & b \\ \hline -f^t & \vartheta \end{array}\right) .$$

Um die Zuordnung der Mengen ζ und ζ_B zur betrachteten Optimierungsaufgabe hervorzuheben, bezeichnen wir diese von nun an genauer mit $\zeta(\Omega)$ bzw. $\zeta_B(\Omega)$.

Zur Transformation von Ω verwenden wir $(n+1, n+1)$-Matrizen der Form

$$\Lambda = \left(\begin{array}{c|c} L & 0 \\ \hline g^t & 1 \end{array}\right) \tag{4.2.2}$$

mit reeller, invertierbarer (n,n)-Matrix L sowie $g \in \mathbb{R}^n$.

Dazu zeigen wir den

(4.2.3) **Hilfssatz.**

(i) *Die Matrizen der Form* (4.2.2) *bilden bezüglich der Matrizenmultiplikation eine Gruppe.*

(ii) *Ist* Λ *von der Form* (4.2.2) *und hiermit*

$$\Omega' = \Lambda\Omega = \left(\begin{array}{c|c} C' & b' \\ \hline -f'^t & \vartheta' \end{array}\right) ,$$

so gilt

$$\begin{cases} \zeta(\Omega) = \zeta(\Omega'), \quad \zeta_B(\Omega) = \zeta_B(\Omega'); \\ x \in \zeta(\Omega) \Rightarrow f^t x + \vartheta = f'^t x + \vartheta' . \end{cases}$$

In diesem Fall nennen wir Ω und Ω' (oder auch die entsprechenden Optimierungsaufgaben) zueinander *äquivalent.*

Beweis. Zunächst ist die $(n+1, n+1)$-Einheitsmatrix von der Gestalt (4.2.2); es seien ferner

$$\Lambda_i = \left(\begin{array}{c|c} L_i & 0 \\ \hline g_i^t & 1 \end{array}\right) \qquad (i = 1, 2)$$

vorgegeben. Dann berechnet man

$$\Lambda_1 \Lambda_2 = \left(\begin{array}{c|c} L_1 L_2 & 0 \\ \hline g_1^t L_2 + g_2^t & 1 \end{array}\right) ,$$

womit die Form (4.2.2) gegeben ist. An Hand dieser Gleichung sieht man, daß

$$\Lambda_1^{-1} = \left(\begin{array}{c|c} L_1^{-1} & 0 \\ \hline -g_1^t L_1^{-1} & 1 \end{array}\right)$$

ist, was den Beweis von Teil (i) abschließt. Die Teilmatrizen von Ω' in (ii) berechnet man als

$$C' = LC, \quad b' = Lb, \quad -f'^t = -f^t + g^t C, \quad \vartheta' = g^t b + \vartheta .$$

Daher haben wir für $x \in \mathbb{R}^m$

$$Cx = b, \quad x \geqq 0 \iff C'x = b', \quad x \geqq 0 ;$$

und nach (4.1.7), (iv) erhalten wir x als Basislösung von $Cx = b$ genau dann, wenn x Basislösung von $C'x = b'$ ist. Für jedes $x \in \zeta(\Omega) = \zeta(\Omega')$ wird

$$f'^t x + \vartheta' = (f^t - g^t C)x + g^t b + \vartheta = f^t x + \vartheta - g^t(Cx - b) = f^t x + \vartheta .$$

Für das Folgende sei stets erfüllt die

(4.2.4) **Voraussetzung.** Zu der der Optimierungsaufgabe (4.1.1) zugeordneten Matrix Ω existiere eine Permutationsmatrix Π der Gestalt

$$\Pi = \left(\begin{array}{c|c} Q & 0 \\ \hline 0 & 1 \end{array}\right) , \quad Q \ (m,m)\text{-Permutationsmatrix} ,$$

mit

$$\Omega\Pi = \left(\begin{array}{c|c|c} A & I_n & b \\ \hline -d^t & 0 & \vartheta \end{array}\right) = \left(\begin{array}{ccc|ccc|c} a_{1,1} & \cdots & a_{1,q} & 1 & & 0 & b_1 \\ \vdots & & \vdots & & \ddots & & \vdots \\ a_{n,1} & \cdots & a_{n,q} & 0 & & 1 & b_n \\ \hline -d_1 & \cdots & -d_q & 0 & \cdots & 0 & \vartheta \end{array}\right) ;$$

dabei sei $b \geqq 0$.

Geht die Optimierungsaufgabe (4.1.1) aus einer Aufgabe der Form (4.1.2) hervor, so ist, falls $b \geqq 0$, die Voraussetzung (4.2.4) erfüllt, und zwar mit $\vartheta = 0$.

Zur Vereinfachung der Schreibweise bezeichnen wir mit

$a_1, \ldots, a_q$ die Spalten von A ,

$S := \{s \in \{1, \ldots, q\} : d_s > 0\}$.

Wir notieren nun den

(4.2.5) **Satz** *(über das Simplex-Verfahren).*

(i) $x := Q\begin{pmatrix} 0 \\ b \end{pmatrix} \in \mathbb{R}^{q+n}$ *ist zulässige Basislösung der Optimierungsaufgabe mit*

$z(x) = f^t x + \vartheta = \vartheta$.

(ii) *Im Fall* $S = \emptyset$, *d. h.* $d \leqq 0$ *ist* x *optimale Lösung,*

$z_{opt} = \vartheta$.

Im Fall $d < 0$ *ist* x *die einzige optimale Lösung.*

(iii) *Wenn ein* $s \in S$ *mit* $a_s \leqq 0$ *existiert, so ist* z *auf* $\zeta(\Omega)$ *nach oben unbeschränkt.*

(iv) *Ist* $S \neq \emptyset$ *und* $a_s \nleqq 0$ *für alle* $s \in S$, *so findet man eine Transformationsmatrix* $\Lambda \neq I$ *der Gestalt* (4.2.2), *mit der*

$$\Omega' := \Lambda\Omega$$

wieder die Voraussetzung (4.2.4) *erfüllt, wobei* $\vartheta' \geqq \vartheta$ *gilt.*

Beweis

(i) Nach (4.1.7), (v) ist x Basislösung von $Cx = b$; ferner ist natürlich mit b auch $x \geqq 0$. Weiter erhält man

$$f^t x + \vartheta = f^t Q \begin{pmatrix} 0 \\ b \end{pmatrix} + \vartheta = (d^t, 0) \begin{pmatrix} 0 \\ b \end{pmatrix} + \vartheta = \vartheta .$$

(ii) Mit d ist ebenfalls $f \leqq 0$; daher gilt für jedes $\tilde{x} \in \zeta(\Omega)$

$$f^t \tilde{x} + \vartheta = \sum_{i=1}^{m} f_i \tilde{x}_i + \vartheta \leqq \vartheta .$$

Ist $d < 0$, $\tilde{x} \in \zeta(\Omega)$ optimale Lösung, so ergibt sich wegen

$$f^t \tilde{x} + \vartheta = (f^t Q)(Q^t \tilde{x}) + \vartheta = (d^t, 0)(Q^t \tilde{x}) + \vartheta = \vartheta$$

mit $Q^t \tilde{x} =: \begin{pmatrix} y_1 \\ y_2 \end{pmatrix} \in \mathbb{R}^{q+n}$ notwendig $y_1 = 0$, mithin

$$b = C\tilde{x} = (CQ)(Q^t \tilde{x}) = (A, I_n) \begin{pmatrix} 0 \\ y_2 \end{pmatrix} = y_2 ,$$

d.h. $\tilde{x} = x$. – Zum Fall $a \not< 0$ vgl. Aufgabe 4.2!

(iii) Es sei $s \in S$ und $a_s \leqq 0$. Dann definieren wir für $\lambda > 0$

$$y(\lambda) := \begin{pmatrix} 0 \\ b \end{pmatrix} + \lambda \begin{pmatrix} e_s \\ -a_s \end{pmatrix} ,$$

wobei e_s der s-te Einheitsvektor des $\mathbb{R}^q$ sei, sowie

$$x(\lambda) := Qy(\lambda) .$$

Wegen $-a_s \geqq 0$ ist $x(\lambda) \geqq 0$ für alle $\lambda > 0$, und

$$Cx(\lambda) = CQy(\lambda) = (A, I) \begin{pmatrix} 0 \\ b \end{pmatrix} + \lambda (A, I) \begin{pmatrix} e_s \\ -a_s \end{pmatrix} = b + \lambda (Ae_s - a_s) = b ,$$

insgesamt also $x(\lambda) \in \zeta(\Omega)$. Ferner berechnen wir

$$f^t x(\lambda) + \vartheta = (d^t, 0)\, y(\lambda) + \vartheta = (d^t, 0) \begin{pmatrix} 0 \\ b \end{pmatrix} + \lambda (d^t, 0) \begin{pmatrix} e_s \\ -a_s \end{pmatrix} + \vartheta = \lambda \cdot d_s + \vartheta .$$

Wegen $d_s > 0$ nimmt $z(x(\lambda))$ $(\lambda > 0)$ beliebig große Werte an.

(iv) Zur Konstruktion von Λ wählt man $s \in S$ beliebig, dazu $r \in \{1, \ldots, n\}$ mit

$$(4.2.6) \qquad a_{r,s} > 0, \quad \frac{b_r}{a_{r,s}} = \min \left\{ \frac{b_i}{a_{i,s}} : i \in \{1, \ldots, n\}, \ a_{i,s} > 0 \right\} .$$

Häufig betrachtet man ein s mit $d_s = \max_{\sigma \in S} d_\sigma$, da man davon ausgeht, daß hierbei ϑ' größer wird als bei kleinem d_s (vgl. (4.2.8, iv)!).

Auf $\Omega\Pi$ wenden wir einen Jordanschen Eliminationsschritt mit dem Pivotelement $a_{r,s}$ an; dem entspricht nach (2.5.2) die Linksmultiplikation mit

$$(4.2.7) \qquad \Lambda := \left(\begin{array}{ccccc|c} 1 & & 0 & -\frac{a_{1,s}}{a_{r,s}} & & 0 \\ & \ddots & & \vdots & 0 & \vdots \\ 0 & & 1 & \vdots & & \\ \vdots & & 0 & \frac{1}{a_{r,s}} & & \vdots \\ & & \vdots & \vdots & & \\ & & & -\frac{a_{n,s}}{a_{r,s}} & 1 & 0 \\ \hline 0 \ldots 0 & & & \frac{d_s}{a_{r,s}} & 0 \ldots 0 & 1 \end{array}\right)$$

r-te Spalte

Λ besitzt offenbar die Gestalt (4.2.2): von dort übernehmen wir die Bezeichnung der Teilmatrizen und erhalten somit

$$\Lambda\Omega\Pi = \left(\begin{array}{ccccc|c} a'_{1,1} \ldots & a'_{1,s-1} & 0 & \ldots & a'_{1,q} & \\ \vdots & \vdots & \vdots & & \vdots & \\ a'_{r,1} \ldots & a'_{r,s-1} & 1 & \ldots & a'_{r,q} & L \qquad\qquad b' \\ \vdots & \vdots & \vdots & & \vdots & \\ a'_{n,1} \ldots & a'_{n,s-1} & 0 & \ldots & a'_{n,q} & \\ \hline -d'_1 \ldots & -d'_{s-1} & 0 & \ldots -d'_q & & 0 \ldots 0 \ \frac{d_s}{a_{r,s}} \ 0 \ldots 0 \quad \vartheta' \end{array}\right)$$

s-te Spalte (r + q)-te Spalte

Mit $P_{s,r+q}$, der $(m+1, m+1)$-Permutationsmatrix, deren Rechtsmultiplikation die Vertauschung der s-ten mit der $(r+q)$-ten Spalte bewirkt, sei $\Pi' := \Pi \cdot P_{s,q+r}$, außerdem $\Omega' := \Lambda\Omega$. Offenbar hat Π' mit einer (m,m)-Permutationsmatrix Q' die Form

$$\Pi' = \left(\begin{array}{c|c} Q' & 0 \\ \hline 0 & 1 \end{array}\right) ,$$

und es wird

$$\Omega'\Pi' = (\Lambda\Omega\Pi)P_{s,r+q} = \left(\begin{array}{c|c|c} A' & I_n & b' \\ \hline -d'^t & 0 & \vartheta' \end{array}\right) .$$

Setzen wir $A' = (a'_{i,j})_{(n,q)}$, $b' = (b'_i)_1^n$, $d' = (d'_j)_1^q$, so ergeben sich gemäß dem Jordanschen Eliminationsverfahren – vgl. (2.5.7) – die folgenden

(4.2.8) *Gleichungen für die Elemente von* Ω':

(i)
$$\left\{\begin{array}{ll} a'_{i,j} = a_{i,j} - \dfrac{a_{i,s}}{a_{r,s}} a_{r,j} & (i \neq r), \\ a'_{r,j} = \dfrac{1}{a_{r,s}} a_{r,j} & \end{array}\right\} \quad (j \neq s),$$

$$a'_{i,s} = -\frac{a_{i,s}}{a_{r,s}} \qquad (i \neq r),$$

$$a'_{r,s} = \frac{1}{a_{r,s}} ;$$

(ii)
$$\left\{\begin{array}{ll} b'_i = b_i - \dfrac{a_{i,s}}{a_{r,s}} b_r & (i \neq r), \\ b'_r = \dfrac{b_r}{a_{r,s}} ; & \end{array}\right.$$

(iii)
$$\left\{\begin{array}{ll} d'_j = d_j - \dfrac{d_s}{a_{r,s}} a_{r,j} & (j \neq s), \\ d'_s = -\dfrac{d_s}{a_{r,s}} ; & \end{array}\right.$$

(iv)
$$\vartheta' = \vartheta + \frac{d_s}{a_{r,s}} b_r .$$

Nun zu $b' \geqq 0$: wegen $b_r \geqq 0$, $a_{r,s} > 0$ ist natürlich $b'_r \geqq 0$; für $i \neq r$ hat man, falls $a_{i,s} \leqq 0$, ebenfalls elementarerweise

$$b'_i = b_i - \frac{a_{i,s}}{a_{r,s}} b_r \geqq b_i \geqq 0$$

sowie, falls $a_{i,s} > 0$, gemäß der Wahl von r

$$b'_i = b_i - \frac{a_{i,s}}{a_{r,s}} b_r = a_{i,s}\left(\frac{b_i}{a_{i,s}} - \frac{b_r}{a_{r,s}}\right) \geqq 0 .$$

Schließlich gilt nach (4.2.8), (iv) wegen $d_s > 0$ die Ungleichung

$$\vartheta' \geqq \vartheta .$$

Das Simplex-Verfahren besteht aus einer (eventuell) wiederholten Anwendung des Satzes (4.2.5). Vorgegeben sei eine Optimierungsaufgabe (4.1.1), deren beschreibende Matrix Ω die Bedingung (4.2.4), ohne Einschränkung mit $\Pi = I_{m+1}$, erfüllt, beispielsweise ein Problem der Form (4.1.2) mit $b \geqq 0$. Dann definieren wir

$$\Omega^{(1)} := \Omega = \left(\begin{array}{c|c|c} A & I_n & b \\ \hline -d^t & 0 & \vartheta \end{array}\right), \qquad \Pi^{(1)} := I_{m+1}.$$

Wir prüfen, ob $\Omega^{(1)}$ die Bedingung (ii) oder (iii) des Satzes (4.2.5) erfüllt: wenn ja, brechen wir ab; sonst konstruieren wir $\Omega^{(2)} = \Lambda^{(1)}\Omega^{(1)}$ nach dem Beweis der Aussage (iv). Genügt $\Omega^{(2)}$ nicht der Bedingung (ii) oder (iii) des Satzes, machen wir weiter mit $\Omega^{(3)} = \Lambda^{(2)}\Omega^{(2)}$ usf. So erhalten wir

$$(4.2.9) \qquad \Omega^{(\nu)} = \Lambda^{(\nu-1)}\Omega^{(\nu-1)} = \Lambda^{(\nu-1)} \ldots \Lambda^{(1)}\Omega^{(1)} \qquad (\nu = 2, 3, \ldots),$$

wobei die Transformationsmatrizen $\Lambda^{(\mu)}$ die Form (4.2.7), ihre Produkte die Gestalt (4.2.2) besitzen. Wir benutzen folgende

(4.2.10) **Bezeichnungen**

(i) $$\Omega^{(\nu)} =: \left(\begin{array}{c|c} C^{(\nu)} & b^{(\nu)} \\ \hline -f^{(\nu)t} & \vartheta_\nu \end{array}\right), \qquad \Lambda^{(\nu)} =: \left(\begin{array}{c|c} L^{(\nu)} & 0 \\ \hline g^{(\nu)t} & 1 \end{array}\right) \qquad (\nu = 1, 2, \ldots).$$

(ii) Wir definieren $T^{(1)} := I_n$, $h^{(1)} := 0$ $(\in \mathbb{R}^n)$ und für $\nu = 2, 3, \ldots$

$$\Lambda^{(\nu-1)} \ldots \Lambda^{(1)} =: \left(\begin{array}{c|c} T^{(\nu)} & 0 \\ \hline h^{(\nu)t} & 1 \end{array}\right).$$

(iii) Es seien die

$$\Pi^{(\nu)} = \left(\begin{array}{c|c} Q^{(\nu)} & 0 \\ \hline 0 & 1 \end{array}\right)_{(m+1,\, m+1)}$$

die zu den $\Omega^{(\nu)}$ gehörenden Permutationsmatrizen mit

$$\Omega^{(\nu)}\Pi^{(\nu)} =: \left(\begin{array}{c|c|c} A^{(\nu)} & I_n & b^{(\nu)} \\ \hline -d^{(\nu)t} & 0 & \vartheta_\nu \end{array}\right).$$

Sämtliche $\Omega^{(\nu)}$ $(\nu = 2, 3, \ldots)$ sind im Sinne des Hilfssatzes (4.2.3) zu $\Omega^{(1)}$ äquivalent. Nach (4.2.5), (i) gehört zu jedem $\Omega^{(\nu)}$ in kanonischer Weise

$$x^{(\nu)} := Q^{(\nu)}\begin{pmatrix} 0 \\ b^{(\nu)} \end{pmatrix} \in \zeta_B(\Omega^{(\nu)}) = \zeta_B(\Omega^{(1)}),$$

und es gilt $\vartheta_\nu = z(x^{(\nu)})$ $(\nu = 1, 2, \ldots)$. Wir wollen nun eine Bedingung angeben, unter der das Simplex-Verfahren nach endlich vielen Schritten zum Ziel führt.

(4.2.11) **Bemerkung.** Wenn jede Basislösung von $Cx = b$ nicht-ausgeartet ist, so endet das Simplex-Verfahren nach endlich vielen Schritten, d.h. es existiert ein $\nu \in \mathbb{N}$, so daß $\Omega^{(\nu)}$ die Bedingung (ii) oder (iii) des Satzes (4.2.5) erfüllt.

Beweis. Die Basislösungen $x^{(\nu)} = Q^{(\nu)} \begin{pmatrix} 0 \\ b^{(\nu)} \end{pmatrix}$ von $C^{(\nu)}x = b^{(\nu)}$ sind nach (4.1.7) genau dann nicht-ausgeartet, wenn sämtliche Komponenten von $b^{(\nu)}$ von Null verschieden, mithin positiv sind. Nun ist $x^{(\nu)}$ auch Basislösung von $Cx = b$ und infolgedessen nach Voraussetzung nicht-ausgeartet.

Weiter folgt aus $b^{(\nu)} > 0$ $(\nu = 1, 2, \ldots)$ nach (4.2.8), (iv)

$$\vartheta_{\nu+1} > \vartheta_\nu \qquad (\nu = 1, 2, \ldots)$$

und daher $\vartheta_\nu \neq \vartheta_\mu$ für $\nu \neq \mu$. Wegen $\vartheta_\nu = z(x^{(\nu)})$ impliziert dies $x^{(\nu)} \neq x^{(\mu)}$ für $\nu \neq \mu$. Das bedeutet, daß keine Basislösung mehrmals vorkommt; damit ist die Behauptung bewiesen.

Zum Programmieren des Simplex-Verfahrens wird man statt der vollen Matrizen $\Omega^{(\nu)}$ nur die Teilmatrizen $A^{(\nu)}$, $b^{(\nu)}$, $d^{(\nu)t}$, ϑ_ν speichern, die sich rekursiv nach (4.2.8) berechnen lassen. Außerdem benötigt man zur Darstellung der $Q^{(\nu)}$ (und $\Pi^{(\nu)}$) die Permutationen $\pi_\nu \in S_m$ mit

$$Q^{(\nu)} e_\kappa = e_{\pi_\nu(\kappa)} \qquad (\kappa = 1, \ldots, m) .$$

Wegen $\Pi^{(1)} = I_{m+1}$, $\Pi^{(\nu+1)} = \Pi^{(\nu)} P_{s_\nu, q+r_\nu}$ $(\nu = 1, 2, \ldots)$, wobei das Pivotelement in $A^{(\nu)}$ mit (r_ν, s_ν) indiziert sei, hat man $\pi_1 = \mathrm{id}$ und für $\nu = 1, 2, \ldots$

$$\pi_{\nu+1}(s_\nu) = \pi_\nu(q + r_\nu), \quad \pi_{\nu+1}(q + r_\nu) = \pi_\nu(s_\nu), \quad \pi_{\nu+1}(\kappa) = \pi_\nu(\kappa) \quad \text{sonst} .$$

Die Koeffizienten von $x^{(\nu)} = Q^{(\nu)} \begin{pmatrix} 0 \\ b^{(\nu)} \end{pmatrix} =: (x_j^{(\nu)})_1^m$ sind gegeben durch

$$(4.2.12) \qquad \begin{cases} x^{(\nu)}_{\pi_\nu(q+i)} = b_i^{(\nu)} & (i = 1, \ldots, n) , \\ x_j^{(\nu)} = 0 & (j \notin \{\pi_\nu(q+1), \ldots, \pi_\nu(q+n)\}) . \end{cases}$$

Auch zum Rechnen von Hand genügt ein verkürztes Schema, in dem die n Einheitsspalten weggelassen sind. Zur Beschreibung der Permutationen π_ν werden, ähnlich wie im Beispiel (2.5.9), die mit $\pi_\nu(q+1), \ldots, \pi_\nu(q+n)$ indizierten Variablen vor die Zeilen, die Variablen Nr. $\pi_\nu(1), \ldots, \pi_\nu(q)$ an den Kopf der Spalten von $A^{(\nu)}$ gesetzt. Bei einem Eliminationsschritt wird die vor der Pivotzeile stehende Variable gegen die Variable oberhalb der Pivotspalte ausgetauscht, wobei die Regeln (4.2.8) zur Anwendung kommen. Die zur Wahl der Pivotzeile benötigten Werte $\dfrac{b_i^{(\nu)}}{a_{i,s}^{(\nu)}}$ mit $a_{i,s}^{(\nu)} > 0$ trägt man in einer zusätzlichen Spalte des Schemas ein.

Wir betrachten folgendes

(4.2.13) **Zahlenbeispiel.** Zur Optimierungsaufgabe (4.1.4) gehört nach Einführung von Schlupfvariablen gemäß (4.1.3) folgende Ausgangsmatrix

$$\Omega^{(1)} = \left(\begin{array}{cc|ccc|c} 1 & 1 & 1 & 0 & 0 & 1{,}2 \\ 5 & 10 & 0 & 1 & 0 & 7{,}0 \\ 2 & 10 & 0 & 0 & 1 & 5{,}2 \\ \hline -20 & -60 & 0 & 0 & 0 & 0 \end{array}\right).$$

Wegen $b \geqq 0$ ist die Voraussetzung (4.2.4) erfüllt. Wenn wir die 2. Spalte als Pivotspalte wählen, lautet das Ausgangsschema

	x_1	x_2	b_i	$\frac{b_i}{a_{i,2}}$
x_3	1	1	1,2	1,2
x_4	5	10	7,0	0,7
x_5	2	[10]	5,2	0,52
	−20	−60	0	

Wir haben das nach (4.2.6) gewonnene Pivotelement eingerahmt und erhalten nach Vertauschung von x_2 mit x_5 und Anwendung von (4.2.8):

	x_1	x_5		
x_3	0,8	−0,1	0,68	0,85
x_4	[3,0]	−1,0	1,8	0,6
x_2	0,2	0,1	0,52	2,6
	−8,0	6,0	31,2	

Als Pivotspalte ist nur die erste Spalte möglich. Berechnung von $\frac{b_i^{(2)}}{a_{i,1}^{(2)}}$ für $a_{i,1}^{(2)} > 0$ liefert die 2. Zeile als Pivotzeile und als weiteres Schema

	x_4	x_5	
x_3	−0,2667	−0,1667	0,2
x_1	0,3333	−0,3333	0,6
x_2	−0,06667	0,1667	0,4
	2,667	3,333	36,0

Für $\Omega^{(3)}$ ist nunmehr die Bedingung (ii) von Satz (4.2.5) erfüllt. Aus unserem Schema erhalten wir

$$z_{opt} = 36{,}0$$

und nach (4.2.12) für die Komponenten der optimalen Lösung:

$$x_3 = 0{,}2; \quad x_1 = 0{,}6; \quad x_2 = 0{,}4; \quad x_4 = x_5 = 0 .$$

Die ursprünglich gegebene Aufgabe besitzt also die optimale Lösung

$$(x_1 ; x_2) = (0{,}6 ; 0{,}4) , \qquad z_{opt} = 36{,}0 .$$

Diese optimale Lösung genügt wegen $x_4 = x_5 = 0$ der 2. und 3. Restriktion mit dem Gleichheitszeichen, ferner ist $x_1 + x_2 = 1{,}0 = 1{,}2 - x_3 < 1{,}2$.

In einem Programm müssen die Vorschriften für die Pivotwahl konkretisiert werden, da eventuell mehrere Indizes s und r in Frage kommen. So präzisiert man (4.2.6) durch die Vorschrift

$$(4.2.14) \qquad r := \min \left\{ r' \in \{1, \ldots, n\} : a_{r',s} > 0, \ \frac{b_{r'}}{a_{r',s}} = \min \left\{ \frac{b_i}{a_{i,s}} : a_{i,s} > 0 \right\} \right\} .$$

Für den Fall, daß beim ν-ten Schritt des Simplex-Verfahrens in $b^{(\nu)}$ mindestens zwei Komponenten verschwinden (vgl. Aufgabe 4.2), kann mit einem gewissen $l > 0$ $\Omega^{(\nu+l)} = \Omega^{(\nu)}$ eintreten. Da die Pivotwahl nach einer festen Vorschrift erfolgt, wird dann auch $\Omega^{(\nu+l+1)} = \Omega^{(\nu+1)}$ usw.; das Verfahren gerät in einen Zyklus der Länge l, ohne daß jemals die Bedingung (ii) oder (iii) des Satzes (4.2.5) erreicht wird. Ein Beispiel hierzu liefert die Aufgabe 4.3.

Um Zyklen zu vermeiden, ist die Auswahlvorschrift von r in einer dem Problem angepaßteren Form zu präzisieren. Dazu führen wir den Begriff der lexikographischen Ordnung ein:
$x \in \mathbb{R}^N, \neq 0$ heißt *lexikographisch positiv*, in Zeichen $x \succ 0$, genau dann, wenn für

$$i_0 = \min \{ i \in \{1, \ldots, N\} : \ x_i \neq 0 \}$$

$x_i > 0$ zutrifft. Sind $x, y \in \mathbb{R}^N$, $x \neq y$, so heißt x *lexikographisch größer als* y, in Zeichen $x \succ y$, genau dann, wenn $x - y \succ 0$.

Wir stellen einige elementare Eigenschaften der lexikographischen Ordnung zusammen.

(4.2.15) **Hilfssatz.** *Für* $x, y, z \in \mathbb{R}^N$ *gilt*

(i) $\begin{cases} x \succ 0, \ y \succ 0 \quad \textit{oder} \quad y = 0 \Rightarrow x + y \succ 0 ; \\ x \succ 0, \ \alpha \in \mathbb{R}, > 0 \Rightarrow \alpha x \succ 0 ; \end{cases}$

(ii) $\begin{cases} x \succ y, \ y \succ z \Rightarrow x \succ z ; \\ x \succ y \Rightarrow x + z \succ y + z ; \end{cases}$

(iii) $x \succ y, \ \alpha \in \mathbb{R}, > 0 \Rightarrow \alpha x \succ \alpha y .$

Beweis. In (i) interessiert nur der Fall $y \succ 0$; es sei $i_1 = \min \{ i : x_i \neq 0 \}$, $i_2 = \min \{ i : y_i \neq 0 \}$, so daß $x_{i_1} > 0$, $y_{i_2} > 0$.

Wir können o.E. $i_1 \leqq i_2$ annehmen. Dann haben wir $x_i + y_i = 0$ für alle $i < i_1$ und $x_{i_1} + y_{i_1} \geqq x_{i_1} > 0$, daher $x + y \succ 0$. Ebenso gilt $\alpha x_i = 0$ für $i < i_1$, $\alpha x_{i_1} > 0$, d.h. $\alpha x \succ 0$.

Zu Teil (ii) benutzen wir (i), womit wir

$$x - y \succ 0, \quad y - z \succ 0 \Rightarrow x - z = (x - y) + (y - z) \succ 0$$

folgern. In (iii) haben wir nach Definition

$$(x + z) - (y + z) = x - y \succ 0$$

sowie nach (i)

$$\alpha x - \alpha y = \alpha(x - y) \succ 0 .$$

Sind $x, y \in \mathbb{R}^N$ mit $x \neq y$, so hat man offenbar $x \succ y$ genau dann, wenn für

$$i_0 = \min \{i \in \{1, 2, \ldots, N\} : x_i \neq y_i\}$$

$x_{i_0} > y_{i_0}$ gilt. Demgemäß ist $(\mathbb{R}^N, \succ)$ totalgeordnet; es ist nämlich stets genau eine der Beziehungen

$$x = y, \quad x \succ y, \quad y \succ x$$

erfüllt.

Ist $M \subseteq \mathbb{R}^N$, $a \in M$, so heißt a kleinstes Element von M bezüglich der lexikographischen Ordnung, kurz

$$a = \operatorname{lmin} M ,$$

wenn für jedes $b \in M$, $\neq a$ die Beziehung $b \succ a$ gilt. Da jede endliche Teilmenge einer totalgeordneten Menge genau ein kleinstes Element besitzt – Beweis durch Induktion über die Anzahl der Elemente –, hat man die

(4.2.16) **Bemerkung.** Jede endliche Teilmenge des $\mathbb{R}^N$ besitzt bezüglich der lexikographischen Ordnung genau ein kleinstes Element.

Für das Folgende nehmen wir an, daß ein Start des Simplex-Verfahrens mit $\Pi^{(1)} = I_{m+1}$ möglich ist, also

$$\Omega^{(1)} = \left(\begin{array}{c|c|c} A^{(1)} & I_n & b^{(1)} \\ \hline -d^{(1)t} & 0 & \vartheta_1 \end{array}\right)$$

gilt. Unter Berücksichtigung dieser Tatsache gewinnen wir aus (4.2.9), (4.2.10) durch blockweise Multiplikation die

(4.2.17) **Bemerkung.** Für $\nu = 1, 2, \ldots$ gelten die Beziehungen

(i) $C^{(\nu)} = T^{(\nu)} C^{(1)} = (T^{(\nu)} A^{(1)}, T^{(\nu)})$,

(ii) $b^{(\nu)} = T^{(\nu)} b^{(1)}$,

(iii) $-f^{(\nu)t} = -f^{(1)t} + h^{(\nu)t}C^{(1)} = (-d^{(1)t} + h^{(\nu)t}A^{(1)},\ h^{(\nu)t})$,

(iv) $\vartheta_\nu = \vartheta_1 + h^{(\nu)t}b^{(1)}$.

Offenbar sind die $h^{(\nu)t}$ und $T^{(\nu)}$ als Teilmatrizen in den $\Omega^{(\nu)}$ enthalten. Weiter folgern wir aus

$$\left(\begin{array}{c|c} T^{(\nu+1)} & 0 \\ \hline h^{(\nu+1)t} & 1 \end{array}\right) = \Lambda^{(\nu)} \left(\begin{array}{c|c} T^{(\nu)} & 0 \\ \hline h^{(\nu)t} & 1 \end{array}\right) = \left(\begin{array}{c|c} L^{(\nu)}T^{(\nu)} & 0 \\ \hline g^{(\nu)t}T^{(\nu)} + h^{(\nu)t} & 1 \end{array}\right)$$

sowie durch Ausmultiplikation von $\Omega^{(\nu+1)} = \Lambda^{(\nu)}\Omega^{(\nu)}$ die Rekursionen

$$(4.2.18)\quad \begin{cases} \text{(i)} & h^{(\nu+1)t} = g^{(\nu)t}T^{(\nu)} + h^{(\nu)t}, \\ \text{(ii)} & \vartheta_{\nu+1} = g^{(\nu)t}b^{(\nu)} + \vartheta_\nu, \qquad (\nu = 1, 2, 3, \ldots) \\ \text{(iii)} & (b^{(\nu+1)},\ T^{(\nu+1)}) = L^{(\nu)}(b^{(\nu)},\ T^{(\nu)}). \end{cases}$$

Wir kommen nun zur angekündigten Präzisierung der Auswahlvorschrift für r, und zwar allgemein beim ν-ten Schritt, natürlich unter der Voraussetzung (4.2.5, iv) für $\Omega^{(\nu)}\Pi^{(\nu)}$. Hierzu seien die Koeffizienten von $\Omega^{(\nu)}\Pi^{(\nu)}$ analog wie in (4.2.4) bezeichnet, jedoch zusätzlich mit einem oberen Index (ν) versehen.

Wie bereits erwähnt, wählen wir zunächst ein beliebiges $s = s_\nu \in \{1, \ldots, q\}$ mit $d_s^{(\nu)} > 0$. Damit ist die s-te Spalte in $\Omega^{(\nu)}\Pi^{(\nu)}$, also die $\pi_\nu(s)$-te Spalte in $\Omega^{(\nu)}$ als Pivotspalte bestimmt. Wir bezeichnen mit

$$y_i^{(\nu)t} \quad (i = 1, \ldots, n)$$

die Zeilen von $T^{(\nu)}$. Hiermit definieren wir für $i \in \{1, \ldots, n\}$ mit $a_{i,s}^{(\nu)} > 0$

$$w_i^{(\nu)} := \frac{1}{a_{i,s}^{(\nu)}}\,(b_i^{(\nu)},\ y_i^{(\nu)t})^t \in \mathbb{R}^{n+1}$$

und bestimmen schließlich $r = r_\nu$ durch

$$(4.2.19)\quad w_r^{(\nu)} = \operatorname{lmin}\{w_i^{(\nu)} : i \in \{1, \ldots, n\},\ a_{i,s}^{(\nu)} > 0\}.$$

Wir notieren als

(4.2.20) **Hilfssatz.** *r ist durch die Beziehung* (4.2.19) *eindeutig bestimmt und genügt der Auswahlvorschrift* (4.2.6).

Beweis. Gäbe es neben r ein weiteres $r' \in \{1, 2, \ldots, n\}$, welches (4.2.19) erfüllt, so wäre auf Grund der Eindeutigkeit des lexikographischen Minimums $w_r^{(\nu)} = w_{r'}^{(\nu)}$. Dies implizierte natürlich die lineare Abhängigkeit der r-ten und r'-ten Zeile von $T^{(\nu)}$, im Widerspruch zur Invertierbarkeit dieser Matrix.

Weiter erhält man gemäß der Definition des lexikographischen Minimums für alle $i = 1, \ldots, n$ mit $a_{i,s}^{(\nu)} > 0$

$$\frac{b_r^{(\nu)}}{a_{r,s}^{(\nu)}} = \frac{b_i^{(\nu)}}{a_{i,s}^{(\nu)}} \quad \text{oder} \quad \frac{b_r^{(\nu)}}{a_{r,s}^{(\nu)}} < \frac{b_i^{(\nu)}}{a_{i,s}^{(\nu)}},$$

womit auch die zweite Aussage klar ist.

Bei den weiteren Überlegungen dieses Abschnitts gehen wir davon aus, daß bei jedem Schritt des Simplex-Verfahrens die Vorschrift (4.2.19) zur Anwendung kommt. Als erstes notieren wir bezüglich der Zeilen der Matrix $(b^{(\nu)}, T^{(\nu)})$, nämlich

$$v_i^{(\nu)t} := (b_i^{(\nu)}, y_i^{(\nu)t}) \qquad (i = 1, \ldots, n)$$

die natürlich von Null verschieden sind, den

(4.2.21) **Hilfssatz.** *Sämtliche* $v_i^{(\nu)}$ *sind lexikographisch positiv.*

Beweis. Für $\nu = 1$ ist $(b^{(1)}, T^{(1)}) = (b, I)$; wegen $b \geqq 0$ sind alle Zeilen lexikographisch positiv. Weiter nehmen wir die Gültigkeit von (4.2.21) für ein beliebiges $\nu \geqq 1$ an. Nach dem Eliminationsschritt mit dem Pivotelement $a_{r,s}^{(\nu)}$ (>0) hat man nach (4.2.18), (iii) für $i \neq r$

$$v_i^{(\nu+1)} = v_i^{(\nu)} - \frac{a_{i,s}^{(\nu)}}{a_{r,s}^{(\nu)}} v_r^{(\nu)}$$

sowie

$$v_r^{(\nu+1)} = \frac{1}{a_{r,s}^{(\nu)}} v_r^{(\nu)} .$$

Nach (4.2.15), (i) ist $v_r^{(\nu+1)} \succ 0$. Ist $a_{i,s}^{(\nu)} < 0$ bzw. $= 0$, so wird

$$-\frac{a_{i,s}^{(\nu)}}{a_{r,s}^{(\nu)}} v_r^{(\nu)} \succ 0 \quad \text{bzw.} \succ = 0$$

und daher für dieses i abermals nach (4.2.15), (i) $v_i^{(\nu+1)} \succ 0$.
Schließlich ergibt sich, falls $a_{i,s}^{(\nu)} > 0$ ist, bei Beachtung von (4.2.15), (iii)

$$v_i^{(\nu+1)} = a_{i,s}^{(\nu)} (w_i^{(\nu)} - w_r^{(\nu)}) \qquad 0 .$$

Die Vektoren

$$u^{(\nu)t} := (\vartheta_\nu, h^{(\nu)t})$$

erfüllen nach (4.2.18), (ii), (iii) die Rekursionsformel

$$u^{(\nu+1)t} = u^{(\nu)t} + g^{(\nu)t}(b^{(\nu)}, T^{(\nu)})$$

$$= u^{(\nu)t} + \frac{d_s^{(\nu)}}{a_{r,s}^{(\nu)}} v_r^{(\nu)} .$$

Da $d_s^{(\nu)}$, $a_{r,s}^{(\nu)}$ positiv und $v_r^{(\nu)}$ lexikographisch positiv sind, erhalten wir als

(4.2.22) **Folgerung.** Für $\nu = 1, 2, \ldots$ gilt

$$u^{(\nu+1)} \succ u^{(\nu)} .$$

Danach gilt für $\nu \neq \mu$ $u^{(\nu)} \neq u^{(\mu)}$ und weiter auf Grund von (4.2.17), (iv) ebenfalls $h^{(\nu)} \neq h^{(\mu)}$. Wegen $C^{(\nu)}Q^{(\nu)} = (A^{(\nu)}, I_n)$ ist mit der (m,n)-Matrix

$$J = (e_{q+1}, \ldots, e_{q+n}) = \begin{pmatrix} 0 \\ \hline I_n \end{pmatrix}$$

nach (4.2.17), (i) die (n,n)-Matrix

$$C^{(1)}Q^{(\nu)}J = T^{(\nu)-1}C^{(\nu)}Q^{(\nu)}J = T^{(\nu)-1}$$

invertierbar. Beachtet man, daß $f^{(\nu)t}Q^{(\nu)} = (d^{(\nu)t}, 0)$, mithin $f^{(\nu)t}Q^{(\nu)}J$ die $(1,n)$-Nullzeile ist, so gewinnt man mit (4.2.17), (iii) für $h^{(\nu)t}$ das lineare Gleichungssystem

$$h^{(\nu)t}C^{(1)}Q^{(\nu)}J = f^{(1)t}Q^{(\nu)}J\,.$$

Die Annahme, daß beim Simplex-Verfahren für $\mu \neq \nu$ $Q^{(\mu)}J = Q^{(\nu)}J$ eintritt, führt wegen der Invertierbarkeit der Koeffizientenmatrix $C^{(1)}Q^{(\nu)}J$ zum Widerspruch $h^{(\nu)} = h^{(\mu)}$. Nun gibt es genau $\binom{m}{n} \cdot n!$ verschieden Matrizen der Form $Q^{(\nu)}J$; infolgedessen muß das Simplex-Verfahren nach spätestens $\binom{m}{n} \cdot n!$ Schritten abbrechen. – Man kann sogar zeigen (Übungsaufgabe 4.6), daß höchstens $\binom{m}{n}$ verschiedene $Q^{(\nu)}J$ auftreten können, so daß das Verfahren nach spätestens $\binom{m}{n}$ Schritten endet. –

Zusammenfassend notieren wir den

(4.2.23) **Satz.** *Erfüllt die zur Optimierungsaufgabe* (4.1.1) *gehörende Matrix die Voraussetzung* (4.2.4) *und wird das Simplex-Verfahren gemäß der Auswahlvorschrift* (4.2.19) *durchgeführt, so wird nach höchstens* $\binom{m}{n}$ *Schritten die Eigenschaft* (ii) *oder* (iii) *in* (4.2.5) *erreicht.*

Wir vermerken, daß die Vorschrift (4.2.19) gegenüber einfacheren Auswahlvorschriften für r gemäß (4.2.6) einen erheblichen Rechenzeit-Mehraufwand bedingt. Statt (4.2.19) verwenden die Computer-Programme daher die einfachere Auswahlvorschrift (4.2.14); dies ist gerechtfertigt, weil Zyklen der oben beschriebenen Art bisher nur in eigens dazu konstruierten Beispielen aufgetreten sind.

Modifikationen des Simplex-Verfahrens, etwa bei nicht vorzeichen-beschränkten Variablen findet man z.B. bei Collatz-Wetterling [4].

4.3. Zweiphasenmethode

In diesem Abschnitt beschäftigen wir uns mit Optimierungsaufgaben der Form (4.1.1), deren zugeordnete Matrix Ω jedoch nicht der Bedingung (4.2.4) genügt, sondern die Gestalt

$$(4.3.1) \qquad \Omega = \left(\begin{array}{c|c} C & b \\ \hline -f^t & \vartheta \end{array}\right) = \left(\begin{array}{c|c|c} A & I_{n_1} & b \\ \hline & 0 & \\ \hline -d^t & 0 & \vartheta \end{array}\right)$$

besitzt. Dabei sei $A \in M(n \times q, \mathbb{R})$, $0 \leq n_1 < n$, $m_1 = n_1 + q > n$, $\operatorname{rg} C = n$ und $b \in \mathbb{R}^n$, ≥ 0.

Zur Untersuchung einer derartigen Aufgabenstellung kommt man, wenn man von dem folgenden Optimierungsproblem ausgeht:

(4.3.2) Maximiere

$$z(\tilde{x}) = \tilde{d}^t \tilde{x} + \vartheta$$

unter den Restriktionen

$$\begin{cases} A_1 \tilde{x} \leq b_1 , \\ A_2 \tilde{x} = b_2 , \\ A_3 \tilde{x} \leq b_3 , \\ \tilde{x} \geq 0 \end{cases}$$

mit $A_i \in M(n_i \times \tilde{q}, \mathbb{R})$, $b_i \in \mathbb{R}^{n_i}$ $(i = 1, 2, 3)$, $\tilde{d} \in \mathbb{R}^{\tilde{q}}$ und $b_1, b_2 \geq 0$, $b_3 < 0$. Hierbei seien die $n_i \geq 0$, $n_2 + n_3 > 0$ – da sonst die Bedingung (4.2.4) erfüllt ist –, $\tilde{q} > n_2$ sowie $\operatorname{rg} A_2 = n_2$.

Definiert man nämlich $x_I \in \mathbb{R}^{n_1}$, $x_{III} \in \mathbb{R}^{n_3}$ durch

$$x_I := b_1 - A_1 \tilde{x}, \qquad x_{III} := b_3 - A_3 \tilde{x}$$

und setzt hiermit

$$x := \begin{pmatrix} \tilde{x} \\ x_{III} \\ x_I \end{pmatrix}$$

sowie

$$C := \begin{pmatrix} A_1 & 0 & I_{n_1} \\ A_2 & 0 & 0 \\ -A_3 & -I_{n_3} & 0 \end{pmatrix}, \quad b := \begin{pmatrix} b_1 \\ b_2 \\ -b_3 \end{pmatrix}, \quad f^t = (\tilde{d}^t, \underbrace{0, \ldots, 0}_{n_3 + n_1}),$$

$$n := n_1 + n_2 + n_3, \qquad q := \tilde{q} + n_3 ,$$

so erkennt man, daß die vorstehende Optimierungsaufgabe (4.3.2) auf ein Problem der Form (4.1.1) zurückgeführt ist, wobei Ω die Gestalt (4.3.1) besitzt. Hierzu hat man zu beachten, daß C wegen $\operatorname{rg} A_2 = n_2$ vom Rang n ist.

Die Behandlung der Optimierungsaufgabe (4.1.1) unter der Voraussetzung (4.3.1) zerfällt in zwei Phasen; in der 1. Phase wird geprüft, ob es überhaupt zulässige Lösungen gibt; ist $\zeta(\Omega) \neq \emptyset$, so wird Ω in eine Matrix Ω_I des Typs (4.2.4) transformiert. Die 2. Phase besteht sodann in der Anwendung des Simplex-Verfahrens auf Ω_I.

Zunächst betrachten wir zu Ω die „erweiterte" Matrix

$$(4.3.3) \qquad \tilde{\Omega} = \left(\begin{array}{c|c} \tilde{C} & b \\ \hline -\tilde{f}^t & \vartheta \end{array}\right) := \left(\begin{array}{c|c|c} A & I_n & b \\ \hline -d^t & 0 & \vartheta \end{array}\right) ;$$

diese ist eine $(n+1, m+1)$-Matrix vom Typ (4.2.4) mit $m = m_1 + (n - n_1) = q + n$.

Für $x \in \mathbb{R}^{m_1}$, $x_u \in \mathbb{R}^{n-n_1}$ – die Komponenten von x_u nennt man unechte Schlupfvariable – und

$$\tilde{x} = \begin{pmatrix} x \\ x_u \end{pmatrix}$$

stellt man unmittelbar die Äquivalenz

$$\tilde{C}\tilde{x} = b \iff Cx + \begin{pmatrix} 0 \\ x_u \end{pmatrix} = b$$

fest. Dies führt zu der

(4.3.4) **Bemerkung.** Es gilt $\zeta(\Omega) \neq \emptyset$ genau dann, wenn ein $\tilde{x} \in \zeta(\tilde{\Omega})$, $= \begin{pmatrix} x \\ x_u \end{pmatrix}$ mit $x_u = 0$ existiert.

Zur Prüfung der Existenz einer Lösung $\tilde{x} \in \zeta(\tilde{\Omega})$ mit $x_u = 0$ formulieren wir unter Benutzung des Vektors $\tilde{e}^t := (\underbrace{0, \ldots, 0}_{m_1}, \underbrace{1, \ldots, 1}_{n-n_1})$ die Optimierungsaufgabe

(4.3.5) Maximiere

$$w(\tilde{x}) = -\tilde{e}^t\tilde{x}$$

unter den Restriktionen

$$\begin{cases} \tilde{C}\tilde{x} = b, \\ \tilde{x} \geqq 0. \end{cases}$$

Dazu zeigen wir den

(4.3.6) **Hilfssatz.**

(i) *Das Problem* (4.3.5) *besitzt eine optimale Lösung* $\tilde{x}_{opt}$ *mit*

$$w_{opt} := w(\tilde{x}_{opt}) \leqq 0.$$

(ii) *Es gilt* $\zeta(\Omega) \neq \emptyset$, *d.h. es existiert ein* $\tilde{x} \in \zeta(\tilde{\Omega})$ *mit* $x_u = 0$ *genau dann, wenn* $w_{opt} = 0$ *ist.*

Beweis.

(i) Die (4.3.5) beschreibende Matrix $\tilde{\Omega}_w$ hat die Gestalt

$$(4.3.7) \qquad \tilde{\Omega}_w = \left(\begin{array}{c|c} \tilde{C} & b \\ \hline \tilde{e}^t & 0 \end{array}\right) =: \left(\begin{array}{c|c|c} A & I_n & b \\ \hline 0 & e^t & 0 \end{array}\right).$$

Diese Matrix geht durch Linksmultiplikation mit

$$(4.3.8) \qquad \Lambda_w^{(0)} := \left(\begin{array}{c|c} I_n & 0 \\ \hline -e^t & 1 \end{array}\right)$$

in

$$(4.3.9) \qquad \tilde{\Omega}_w^{(1)} = \left(\begin{array}{c|c|c} A & I_n & b \\ \hline -h^t & 0 & \eta \end{array}\right),$$

also eine Matrix des Typs (4.2.4) über. Nach Satz (4.2.5) ist $\zeta(\widetilde{\Omega}_w^{(1)}) \neq \emptyset$, folglich gemäß (4.2.3), (ii) auch $\zeta(\widetilde{\Omega}_w) \neq \emptyset$; außerdem haben wir $\tilde{x} \geqq 0$ für jedes $\tilde{x} \in \zeta(\widetilde{\Omega}_w)$ und daher $w(\tilde{x}) \leqq 0$.

(ii) Ist $w_{opt} = 0$, so existiert ein $\tilde{x} \in \zeta(\widetilde{\Omega}_w) = \zeta(\widetilde{\Omega})$ mit $w(\tilde{x}) = -\tilde{e}^t\tilde{x} = 0$ und folglich mit $x_u = 0$. Gibt es umgekehrt ein $\tilde{x} \in \zeta(\widetilde{\Omega}) = \zeta(\widetilde{\Omega}_w)$ mit $x_u = 0$, so ist natürlich $w_{opt} = w(\tilde{x}) = 0$.

Zur Lösung der Optimierungsaufgabe (4.3.5), d.h. zur Prüfung der Frage, ob $w_{opt} = 0$ ist, wenden wir das Simplex-Verfahren bezüglich der transformierten Matrix (4.3.9) an. Wir konstruieren also rekursiv

$$\widetilde{\Omega}_w^{(\nu+1)} = \Lambda_w^{(\nu)} \widetilde{\Omega}_w^{(\nu)} \qquad (\nu = 1, 2, 3, \ldots), \tag{4.3.10}$$

so daß mit $(m+1, m+1)$-Permutationsmatrizen $\widetilde{\Pi}^{(\nu)}$, insbesondere $\widetilde{\Pi}^{(1)} = I_{m+1}$, für $\nu = 1, 2, 3, \ldots$

$$\widetilde{\Omega}_w^{(\nu)} \widetilde{\Pi}^{(\nu)} = \left(\begin{array}{c|c|c} A^{(\nu)} & I_n & b^{(\nu)} \\ \hline -h^{(\nu)t} & 0 & \eta_\nu \end{array}\right), \quad b^{(\nu)} \geqq 0 \tag{4.3.11}$$

gilt. Gemäß (4.2.7) haben dabei die Transformationsmatrizen $\Lambda_w^{(\nu)}$ die Gestalt

$$\Lambda_w^{(\nu)} = \left(\begin{array}{c|c} L^{(\nu)} & 0 \\ \hline g_w^{(\nu)t} & 1 \end{array}\right), \qquad g_w^{(\nu)} = \frac{h_s^{(\nu)}}{a_{r,s}^{(\nu)}} e_r, \tag{4.3.12}$$

worin $r = r_\nu$, $s = s_\nu$ die Indizes der jeweiligen Pivotzeile bzw. Pivotspalte in $\widetilde{\Omega}_w^{(\nu)} \widetilde{\Pi}^{(\nu)}$ bedeuten.

Da die Aufgabe (4.3.5) eine optimale Lösung besitzt, ergibt das auf $\widetilde{\Omega}_w^{(1)}$ angewendete Simplex-Verfahren auf Grund das Satzes (4.2.23) nach endlich-vielen, etwa $l-1$ Schritten, eine Matrix $\widetilde{\Omega}_w^{(l)}$, die der in (4.2.5), (ii) angegebenen Bedingung genügt, d.h. es gilt im Sinne von (4.3.11)

$$h^{(l)} \leqq 0, \qquad \eta_l = w_{opt} \leqq 0.$$

Ist nun $w_{opt} < 0$, so besitzt die ursprüngliche, durch (4.3.1) beschriebene Optimierungsaufgabe keine zulässigen, mithin erst recht keine optimalen Lösungen. Eine weitere Diskussion erübrigt sich unter diesen Umständen.

Im weiteren beschäftigen wir uns daher mit dem Fall $w_{opt} = 0$. Hierzu sei zunächst angemerkt, daß zwar einerseits $w_{opt} = 0$, andererseits auf Grund von Rundungsfehlern η_l von Null verschieden sein kann; wir wollen daher eine leicht nachprüfbare, hinreichende Bedingung für $w_{opt} = 0$ notieren.

(4.3.13) **Bemerkung.** Es sei $\widetilde{Q}$ die (m, m)-Permutationsmatrix mit

$$\widetilde{\Pi}^{(l)} = \left(\begin{array}{c|c} \widetilde{Q} & 0 \\ \hline 0 & 1 \end{array}\right),$$

ferner $\tilde{\pi} \in S_m$ mit

$$\tilde{Q} e_\kappa = e_{\tilde{\pi}(\kappa)} \qquad (\kappa = 1, 2, \ldots, m) .$$

Ist dann die Bedingung

(4.3.14) $\quad \{m_1 + 1, \ldots, m\} \subseteq \{\tilde{\pi}(1), \ldots, \tilde{\pi}(q)\}$

erfüllt, so gilt $w_{opt} = 0$.

Zum *Beweis* ziehen wir den Satz (4.2.5), (ii) heran. Hiernach ist

$$(4.3.15) \qquad \tilde{x} = (x_i)_1^m := \tilde{Q} \begin{pmatrix} 0 \\ b^{(l)} \end{pmatrix}$$

eine optimale Lösung von (4.3.5), die nach (4.2.12) den Bedingungen

$$x_{\tilde{\pi}(j)} = 0 \qquad (j = 1, \ldots, q)$$

genügt. Auf Grund von (4.3.14) folgt $x_u = (x_i)_{m_1+1}^m = 0$ und mithin wie behauptet $w(\tilde{x}) = 0$.

Weiter beweisen wir den

(4.3.16) **Satz.** *Ist* $\zeta(\Omega) \neq \emptyset$, *d. h.* $w_{opt} = 0$, *so ist* Ω *mittels einer Matrix* Λ_I *der Form* (4.2.2) *auf eine Matrix*

$$\Omega_I = \Lambda_I \Omega$$

des Typs (4.2.4) *transformierbar.*

Beweis. Ausgehend von der Matrix $\tilde{\Omega}^{(1)} := \tilde{\Omega}$, die sich von $\tilde{\Omega}_w^{(1)}$ nur in der letzten Zeile unterscheidet, definieren wir rekursiv Matrizen

$$(4.3.17) \qquad \tilde{\Omega}^{(\nu+1)} = \Lambda^{(\nu)} \tilde{\Omega}^{(\nu)} \qquad (\nu = 1, 2, 3, \ldots) ,$$

so daß sich mit den Matrizen $\tilde{\Pi}^{(\nu)}$, $A^{(\nu)}$, $b^{(\nu)}$ aus (4.3.11)

$$(4.3.18) \qquad \tilde{\Omega}^{(\nu)} \tilde{\Pi}^{(\nu)} = \left(\begin{array}{c|c|c} A^{(\nu)} & I_n & b^{(\nu)} \\ \hline -d^{(\nu)t} & 0 & \vartheta_\nu \end{array}\right)$$

ergibt. Zu diesem Zweck wählen wir die Transformationsmatrizen $\Lambda^{(\nu)}$, indem wir unter Benutzung der Pivotindizes $r = r_\nu$, $s = s_\nu$ des Simplex-Verfahrens (4.3.10)

$$(4.3.19) \qquad g^{(\nu)} := \frac{d_s^{(\nu)}}{a_{r,s}^{(\nu)}} e_r$$

bestimmen und hiermit sowie mit den $L^{(\nu)}$ aus (4.3.12)

$$(4.3.20) \qquad \Lambda^{(\nu)} := \left(\begin{array}{c|c} L^{(\nu)} & 0 \\ \hline g^{(\nu)t} & 1 \end{array}\right)$$

setzen. Die Gleichung (4.3.18) ist dann erfüllt, weil auf Grund der Festsetzung (4.3.19) die letzte Komponente der s_ν-ten Spalte in $\Lambda^{(\nu)}\widetilde{\Omega}^{(\nu)}$ zu Null wird.

Nach Abschluß des Simplex-Verfahrens (4.3.10), d.h. nach $l-1$ Transformationsschritten brechen wir das Verfahren (4.3.17) ebenfalls ab. Gemäß Konstruktion haben wir

$$(4.3.21) \qquad \widetilde{\Omega}^{(l)} = \Lambda^{(l-1)} \dots \Lambda^{(1)} \widetilde{\Omega} .$$

Zur ursprünglichen Matrix Ω in (4.3.1) definieren wir entsprechend

$$(4.3.22) \qquad \Omega^{(l)} := \Lambda^{(l-1)} \cdot \dots \cdot \Lambda^{(1)} \Omega .$$

Die Tatsache, daß Ω aus $\widetilde{\Omega}$ durch Streichen der Spalten $m_1 + 1, \dots, m$ entsteht, überträgt sich offenbar auf die Matrizen $\Omega^{(l)}$ und $\widetilde{\Omega}^{(l)}$.

Wir zeigen nun zunächst den

(4.3.23) **Hilfssatz.** *Ist die Bedingung* (4.3.14) *erfüllt, so besitzt* $\Omega^{(l)}$ *die Eigenschaft* (4.2.4).

Beweis. Es sei $\kappa \in \{m_1 + 1, \dots, m\}$. Wir betrachten die κ-te Spalte in $\widetilde{\Omega}^{(l)}$, die sich mit dem Einheitsvektor $e_\kappa \in \mathbb{R}^{m+1}$ als $\widetilde{\Omega}^{(l)} e_\kappa$ schreiben läßt. Auf Grund von (4.3.14) existiert ein $j \in \{1, 2, \dots, q\}$ mit $\kappa = \bar{\pi}(j)$, folglich mit

$$\widetilde{\Omega}^{(l)} e_\kappa = \widetilde{\Omega}^{(l)} e_{\bar{\pi}(j)} = \widetilde{\Omega}^{(l)} \widetilde{\Pi}^{(l)} e_j .$$

Das Streichen der κ-ten Spalte in $\widetilde{\Omega}^{(l)}$ bedeutet also das Weglassen einer der ersten q Spalten in $\widetilde{\Omega}^{(l)} \widetilde{\Pi}^{(l)}$. Erhalten bleiben mithin sämtliche Einheitsspalten aus der Teilmatrix

$$(4.3.24) \qquad \left(\frac{I_n}{0}\right)$$

von $\widetilde{\Omega}^{(l)} \widetilde{\Pi}^{(l)}$. Da außerdem gemäß Konstruktion $b^{(l)} \geqq 0$ ist, besitzt $\Omega^{(l)}$ die Eigenschaft (4.2.4).

Mit $\Omega_I = \Omega^{(l)}$, $\Lambda_I = \Lambda^{(l-1)} \cdot \dots \cdot \Lambda^{(1)}$ ist in diesem Fall die Behauptung des Satzes (4.3.16) klar; ansonsten beweisen wir den

(4.3.25) **Hilfssatz.** *Ist die Bedingung* (4.3.14) *verletzt, so existieren Transformationsmatrizen* $\Lambda^{(l)}, \Lambda^{(l+1)}, \dots \Lambda^{(k-1)}$ *der Form* (4.2.2), *so daß die Matrix*

$$\Omega^{(k)} := \Lambda^{(k-1)} \cdot \dots \cdot \Lambda^{(l)} \Omega^{(l)}$$

der Bedingung (4.2.4) *genügt.*

Beweis. Wir setzen

$$(4.3.26) \qquad \Omega^{(l)} = \left(\begin{array}{c|c} C^{(l)} & b^{(l)} \\ \hline -f^{(l)t} & \vartheta_l \end{array}\right) .$$

Gemäß (4.3.20), (4.3.22) ist darin

$$C^{(l)} = L^{(l-1)} \cdot \ldots \cdot L^{(1)} C ,$$

woraus insbesondere $\operatorname{rg} C^{(l)} = \operatorname{rg} C = n$ folgt; die $b^{(l)}$, ϑ_l sind die bereits in (4.3.11) bzw. (4.3.18) definierten Größen.

Da die Eigenschaft (4.3.14) nicht gegeben ist, existieren gewisse Indizes $i \in \{1, 2, \ldots, n\}$ mit $\tilde{\pi}(q+i) \in \{m_1+1, \ldots, m\}$. Nach Überlegungen wie im Beweis des vorangehenden Hilfssatzes werden für diese i die Spalten e_i in der Teilmatrix (4.3.24) von $\widetilde{\Omega}^{(l)} \widetilde{\Pi}^{(l)}$ gestrichen, während die übrigen Einheitsspalten in die Matrix $\Omega^{(l)}$ übertragen werden. Unser Ziel ist es, die in $\Omega^{(l)}$ fehlenden Einheitsspalten durch geeignete Transformationen zu erzeugen. Dazu wählen wir ein $i_1 \in \{1, 2, \ldots, n\}$ mit $\tilde{\pi}(q+i_1) \in \{m_1+1, \ldots, m\}$. Auf Grund der Voraussetzung $w_{opt} = 0$ verschwindet die Komponente x_u der optimalen Lösung (4.3.15) des Problems (4.3.5); danach ist insbesondere $x_{\tilde{\pi}(q+i_1)} = 0$, mithin gemäß (4.2.12) auch $b_{i_1}^{(l)} = 0$. Wegen $\operatorname{rg} C^{(l)} = n$ verschwindet die i_1-te Zeile von $C^{(l)}$ nicht, es existiert also ein $j_1 \in \{1, \ldots, m_1\}$ mit $c_{i_1, j_1}^{(l)} \neq 0$. Bezüglich dieses Elements wenden wir auf die Matrix $\Omega^{(l)}$ einen Jordanschen Eliminationsschritt an und erhalten

$$\Omega^{(l+1)} = \Lambda^{(l)} \Omega^{(l)} = \left(\begin{array}{c|c} C^{(l+1)} & b^{(l+1)} \\ \hline -f^{(l+1)t} & \vartheta_{l+1} \end{array} \right) .$$

Hierbei besitzt $\Lambda^{(l)}$ die Gestalt (4.2.2). Ferner werden sämtliche Einheitsspalten $\neq e_{i_1}$ von $\Omega^{(l)}$ unverändert in die Matrix $\Omega^{(l+1)}$ übertragen; zusätzlich enthält $\Omega^{(l+1)}$ die Spalte $e_{i_1} \in \mathbb{R}^{n+1}$. Schließlich gilt – vgl. (4.2.8) – wegen $b_{i_1}^{(l)} = 0$

$$b^{(l+1)} = b^{(l)} \quad (\geqq 0) .$$

Ist $i_1 \in \{1, \ldots, n\}$ der einzige Index mit $\tilde{\pi}(q+i_1) \in \{m_1+1, \ldots, m\}$, so sind in $\Omega^{(l+1)}$ sämtliche Einheitsspalten $e_1, \ldots, e_n \in \mathbb{R}^{n+1}$ vorhanden, und $\Omega^{(l+1)}$ genügt infolgedessen der Bedingung (4.2.4). Liegt ein weiteres $i_2 \in \{1, \ldots, n\}$ mit $\tilde{\pi}(q+i_2) \in \{m_1+1, \ldots, m\}$ vor, so wiederholt man bezüglich $\Omega^{(l+1)}$ die vorstehend beschriebene Transformation usw. Nach spätestens n Schritten gewinnt man bei diesem Vorgehen eine Matrix $\Omega^{(k)}$, die der Bedingung (4.2.4) genügt, womit der Beweis des Hilfssatzes (4.3.25) erbracht ist.

Mit diesem Hilfssatz ist natürlich auch – man setze $\Omega_I = \Omega^{(k)}$, $\Lambda_I = \Lambda^{(k-1)} \cdot \ldots \cdot \Lambda^{(1)}$ – der Satz (4.3.16) vollständig bewiesen.

Angemerkt sei noch, daß bei dem Verfahren (4.3.17) nicht notwendig $\vartheta_{\nu+1} \geqq \vartheta_\nu$ in (4.3.18) eintritt. Das liegt darin begründet, daß wir die $\widetilde{\Omega}^{(\nu)}$ nicht durch direkte Anwendung des Simplex-Verfahrens auf $\widetilde{\Omega}$ gewinnen; es wird nämlich beim ν-ten Schritt die Pivotspalte, also s, nicht gemäß $d_s^{(\nu)} > 0$, sondern gemäß $h_s^{(\nu)} > 0$ bestimmt.

In der Praxis führt man übrigens die beiden Transformationsverfahren (4.3.10) und (4.3.17) wegen der Übereinstimmung in den ersten n Zeilen gleichzeitig durch; hierzu dient das folgende (n + 2, m + 1)-Matrix-Schema:

(4.3.27)
$$\left(\begin{array}{c|c|c} A^{(\nu)} & I_n & b^{(\nu)} \\ \hline -h^{(\nu)\,t} & 0 & \eta_\nu \\ \hline -d^{(\nu)\,t} & 0 & \vartheta_\nu \end{array}\right).$$

Auf die in der 1. Phase gewonnene Matrix Ω_I wenden wir – wie bereits erwähnt – das in Abschnitt 4.2 bereitgestellte Simplex-Verfahren an. Dieses Verfahren liefert eine Entscheidung darüber, ob die Zielfunktion auf $\zeta(\Omega) = \zeta(\Omega_I)$ nach oben beschränkt ist; wenn ja, wird eine optimale Lösung des Problems rechnerisch ermittelt.

(4.3.28) **Zahlenbeispiel.** Wir betrachten die gegenüber (4.1.4) abgeänderte Aufgabe: Maximiere

$$z(x) = 20\,x_1 + 60\,x_2$$

unter den Restriktionen

$$\begin{aligned} 5\,x_1 + 10\,x_2 &\leqq 7{,}0\,, \\ 2\,x_1 + 10\,x_2 &\leqq 5{,}2\,, \\ x_1 + \;\; x_2 &= 1{,}2\,, \\ x_1,\; x_2 &\geqq 0\,. \end{aligned}$$

Diese Aufgabenstellung führt, wie zu (4.3.2) allgemein gezeigt, auf die Matrix

$$\Omega = \left(\begin{array}{cccc|c} 5 & 10 & 1 & 0 & 7{,}0 \\ 2 & 10 & 0 & 1 & 5{,}2 \\ 1 & 1 & 0 & 0 & 1{,}2 \\ \hline -20 & -60 & 0 & 0 & 0 \end{array}\right).$$

Zur erweiterten Matrix $\widetilde{\Omega}$ gehört gemäß (4.3.7)

$$\widetilde{\Omega}_w = \left(\begin{array}{ccccc|c} 5 & 10 & 1 & 0 & 0 & 7{,}0 \\ 2 & 10 & 0 & 1 & 0 & 5{,}2 \\ 1 & 1 & 0 & 0 & 1 & 1{,}2 \\ \hline 0 & 0 & 0 & 0 & 1 & 0 \end{array}\right).$$

Wir berechnen $\widetilde{\Omega}_w^{(1)} = \Lambda_w^{(0)} \widetilde{\Omega}_w$, indem wir die vorletzte von der letzten Zeile in $\widetilde{\Omega}_w$ subtrahieren. In den folgenden Schemata fügen wir gemäß (4.3.27) an die $\widetilde{\Omega}_w^{(\nu)}$ die letzte Zeile von $\widetilde{\Omega}^{(\nu)}$ an, notieren die Matrizen aber ohne Durchführung der Spaltenpermutationen.

$$\widetilde{\Omega}_w^{(1)},\ \widetilde{\Omega}^{(1)} = \left(\begin{array}{ccccc|c} 5 & 10 & 1 & 0 & 0 & 7 \\ 2 & 10 & 0 & 1 & 0 & 5{,}2 \\ 1 & 1 & 0 & 0 & 1 & 1{,}2 \\ \hline -1 & -1 & 0 & 0 & 0 & -1{,}2 \\ \hline -20 & -60 & 0 & 0 & 0 & 0 \end{array}\right) \qquad \begin{array}{c} \frac{b_i}{a_{i,1}} \\ \downarrow \\ \begin{pmatrix} 1{,}4 \\ 2{,}6 \\ 1{,}2 \end{pmatrix} \end{array}$$

Nach Wahl der ersten Spalte als Pivotspalte (wegen $h_1^{(1)} > 0$) ermitteln wir die 3. Zeile als Pivotzeile und führen einen Jordanschen Eliminationsschritt aus. Es ergibt sich

$$\widetilde{\Omega}_w^{(2)},\ \widetilde{\Omega}^{(2)} = \left(\begin{array}{ccccc|c} 0 & 5 & 1 & 0 & -5 & 1 \\ 0 & 8 & 0 & 1 & -2 & 2{,}8 \\ 1 & 1 & 0 & 0 & 1 & 1{,}2 \\ \hline 0 & 0 & 0 & 0 & 1 & 0 \\ \hline 0 & -40 & 0 & 0 & 20 & 24 \end{array}\right) .$$

Hiermit erfüllt $\widetilde{\Omega}_w^{(2)}$ die Voraussetzung in (4.2.5), (ii); gleichzeitig ist die Bedingung (4.3.14) gegeben. Gemäß Hilfssatz (4.3.23) ist die Matrix $\Omega^{(2)}$, die aus $\widetilde{\Omega}^{(2)}$ durch Streichen der 5. Spalte entsteht, vom Typ (4.2.4). Wir haben also

$$\Omega_I = \Omega^{(2)} = \left(\begin{array}{cccc|c} 0 & 5 & 1 & 0 & 1 \\ 0 & 8 & 0 & 1 & 2{,}8 \\ 1 & 1 & 0 & 0 & 1{,}2 \\ \hline 0 & -40 & 0 & 0 & 24 \end{array}\right) \qquad \begin{array}{c} \frac{b_i^{(2)}}{c_{i,2}^{(2)}} \\ \downarrow \\ \begin{pmatrix} 0{,}2 \\ 0{,}35 \\ 1{,}2 \end{pmatrix} \end{array} .$$

Die 2. Phase besteht nun aus dem Simplex-Verfahren, angewendet auf $\Omega^{(2)}$; hierzu führen wir einen Transformationsschritt mit dem Pivotelement $c_{1,2}^{(2)}$ aus. Wir erhalten bereits eine Matrix mit der in (4.2.5), (ii) angegebenen Eigenschaft, nämlich

$$\Omega_{II} = \left(\begin{array}{cccc|c} 0 & 1 & 0{,}2 & 0 & 0{,}2 \\ 0 & 0 & -1{,}6 & 1 & 1{,}2 \\ 1 & 0 & -0{,}2 & 0 & 1{,}0 \\ \hline 0 & 0 & 8 & 0 & 32 \end{array}\right)$$

Die optimale Lösung lautet nunmehr

$$x_1 = 1{,}0; \quad x_2 = 0{,}2; \quad x_3 = 0; \quad x_4 = 1{,}2$$

sowie

$$z_{opt} = 32 .$$

Wir bemerken, daß auch für die Zweiphasenmethode verkürzte Schemata wie im Beispiel (4.2.13) angewendet werden können; die Durchführung des Beispiels (4.3.28) in der verkürzten Form überlassen wir dem Leser.

4.4. Die duale Optimierungsaufgabe

Zur Optimierungsaufgabe (4.1.2) formulieren wir für $u \in \mathbb{R}^n$ das folgende Problem:

(4.4.1) Minimiere

$$w(u) := b^t u$$

unter den Restriktionen

$$\begin{cases} A^t u \geqq d , \\ u \geqq 0 . \end{cases}$$

Dann heißt (4.4.1) die zu (4.1.2) *duale Aufgabe.* Wir können (4.4.1) auch in der uns geläufigeren Form schreiben, nämlich

(4.4.1′) Maximiere

$$-w(u) := -b^t u$$

unter den Restriktionen

$$\begin{cases} -A^t u \leqq -d , \\ u \geqq 0 . \end{cases}$$

Offenbar ist die zu (4.4.1′) – also (4.4.1) – duale Optimierungsaufgabe wieder das ursprüngliche Problem (4.1.2).

In leichter Abwandlung der bisherigen Schreibweise nennen wir

$$\zeta := \{x \in \mathbb{R}^q : Ax \leqq b,\ x \geqq 0\}$$

bzw.

$$\zeta_d := \{u \in \mathbb{R}^n : A^t u \geqq d,\ u \geqq 0\}$$

die Gesamtheit der zulässigen Lösungen von (4.1.2) bzw. (4.4.1). Hierzu zeigen wir den

(4.4.2) **Hilfssatz.** *Ist* $x \in \zeta$, $u \in \zeta_d$, *so gilt*

$$d^t x \leqq b^t u .$$

Beweis. Wir betrachten die Beziehung $d \leqq A^t u$ komponentenweise, multiplizieren die einzelnen Ungleichungen mit den entsprechenden Koeffizienten von x und erhalten wegen $x \geqq 0$ durch Summation

$$d^t x \leqq (A^t u)^t x = u^t A x .$$

Ebenso folgern wir aus $Ax \leqq b$ und $u \geqq 0$ weiter

$$u^t A x \leqq u^t b = b^t u .$$

Wir gewinnen unmittelbar die

(4.4.3) **Folgerung.** Ist $\zeta \neq \emptyset$ und $\zeta_d \neq \emptyset$, so besitzen beide Optimierungsaufgaben (4.1.2) und (4.4.1) optimale Lösungen. Für alle $x \in \zeta$, $u \in \zeta_d$ gelten die Ungleichungen

$$d^t x \leqq z_{opt} \leqq w_{opt} \leqq b^t u .$$

Beweis. Ist $u \in \zeta_d$ fest, so gilt nach (4.4.2) für alle $x \in \zeta$

$$z(x) \leqq b^t u .$$

Nach Satz (4.1.9) existiert eine optimale Lösung von (4.1.2), also ein $x_0 \in \zeta$ mit $z(x_0) = z_{opt}$. Wieder nach (4.4.2) haben wir für alle $u \in \zeta_d$

$$z_{opt} = d^t x_0 \leqq b^t u .$$

Hieraus folgt die Existenz einer optimalen Lösung u_0 von (4.4.1), und es gilt

$$z_{opt} \leqq b^t u_0 = w_{opt} .$$

Wir wollen nun unter anderem eine Umkehrung der Aussage (4.4.3) beweisen, nämlich aus der Existenz einer optimalen Lösung von (4.1.2) die Tatsache $\zeta_d \neq \emptyset$ folgern.

Offenbar wird (4.1.2) gemäß der Reduktion (4.1.3) durch die Matrix

$$\Omega^{(0)} := \left(\begin{array}{c|c|c} A & I_n & b \\ \hline -d^t & 0 & 0 \end{array}\right)$$

beschrieben. Hierzu zeigen wir den

(4.4.4) **Hilfssatz.** *Die Aufgabe* (4.1.2) *besitze eine optimale Lösung. Dann existiert eine* $(n+1, n+1)$*-Matrix der Form* (4.2.2), *also*

$$\Lambda = \left(\begin{array}{c|c} L & 0 \\ \hline g^t & 1 \end{array}\right) ,$$

so daß

$$(4.4.5) \qquad \Lambda \Omega^{(0)} =: \left(\begin{array}{c|c|c} LA & L & b_0 \\ \hline v_0^t & u_0^t & \vartheta_0 \end{array}\right)$$

die Bedingungen

$$b_0,\ u_0 \in \mathbb{R}^n,\ \geqq 0; \qquad v_0 \in \mathbb{R}^q,\ \geqq 0; \qquad \vartheta_0 = z_{opt}$$

erfüllt.

Beweis. Im Fall $b \geqq 0$ genügt $\Omega^{(0)}$ der Voraussetzung (4.2.4), also ist das Simplex-Verfahren direkt anwendbar. Nach Satz (4.2.23) erhalten wir nach endlich vielen – etwa l – Schritten

$$\Omega^{(l)} = (\Lambda^{(l-1)} \cdot \ldots \cdot \Lambda^{(0)})\, \Omega^{(0)} =: \Lambda\, \Omega^{(0)}$$

mit den in (4.2.5), (ii) angegebenen Eigenschaften und $b_0 \geqq 0$. Für diesen Fall ist der Hilfssatz also gezeigt.

Ist $b \geqq 0$ nicht erfüllt, so können wir ohne Einschränkung

$$b = \begin{pmatrix} b_1 \\ b_2 \end{pmatrix} \quad \text{mit} \quad b_1 \in \mathbb{R}^{n_1},\ \geqq 0; \quad b_2 \in \mathbb{R}^{n-n_1},\ < 0 \qquad (0 \leqq n_1 < n)$$

annehmen. Durch Multiplikation von $\Omega^{(0)}$ mit

$$\Lambda^{(0)} := \left(\begin{array}{c|c|c} I_{n_1} & 0 & 0 \\ \hline 0 & -I_{n-n_1} & 0 \\ \hline 0 & 0 & 1 \end{array}\right)$$

erhalten wir die Gestalt

$$\Lambda^{(0)}\Omega^{(0)} =: \Omega^{(1)} = \left(\begin{array}{c|c|c|c} \multirow{2}{*}{$\tilde{A}$} & I_{n_1} & 0 & \multirow{2}{*}{$\tilde{b}$} \\ \cline{2-3} & 0 & -I_{n-n_1} & \\ \hline -d^t & \multicolumn{2}{c|}{0} & 0 \end{array}\right)$$

mit $\tilde{b} \geqq 0$; eine leicht anzugebende Permutationsmatrix $\Pi^{(1)}$ liefert

$$\Omega^{(1)}\Pi^{(1)} = \left(\begin{array}{c|c|c|c} \multirow{2}{*}{$\tilde{A}$} & 0 & I_{n_1} & \multirow{2}{*}{$\tilde{b}$} \\ \cline{2-3} & -I_{n-n_1} & 0 & \\ \hline -d^t & \multicolumn{2}{c|}{0} & 0 \end{array}\right) .$$

Offenbar ist $\Omega^{(1)}\Pi^{(1)}$ von der Gestalt (4.3.1). Nun ist laut Voraussetzung $\zeta(\Omega^{(0)}) = \zeta(\Omega^{(1)}) \neq \emptyset$, mithin auch $\zeta(\Omega^{(1)}\Pi^{(1)}) \neq \emptyset$. Durch Anwendung von Satz (4.3.16) auf $\Omega^{(1)}\Pi^{(1)}$ gewinnen wir eine Matrix Λ_I der Form (4.2.2), mit der $\Lambda_I \Omega^{(1)}\Pi^{(1)}$, also auch

$$\Omega_I := \Lambda_I \Omega^{(1)} = \Lambda_I \Lambda^{(0)} \Omega^{(0)}$$

die Bedingung (4.2.4) erfüllt. Wegen Hilfssatz (4.2.3) und nach Voraussetzung

besitzt die durch Ω_I beschriebene Optimierungsaufgabe eine optimale Lösung. Daher liefert das Simplex-Verfahren gemäß Satz (4.2.23), auf Ω_I angewendet, eine Transformation auf eine Matrix

$$\Omega_{II} = \Lambda_{II}\Omega_I = (\Lambda_{II}\Lambda_I\Lambda^{(0)})\Omega^{(0)} =: \Lambda\Omega^{(0)} ,$$

die die Eigenschaft (4.2.5), (ii) besitzt.

Anschließend gewinnen wir den

(4.4.6) **Satz** *(Dualitätssatz). Das Problem* (4.1.2) *besitzt eine optimale Lösung genau dann, wenn zur Aufgabe* (4.4.1) *eine optimale Lösung existiert. In diesem Fall ist*

$$z_{opt} = w_{opt} ;$$

dabei ist u_0 *in* (4.4.5) *eine optimale Lösung von* (4.4.1).

Beweis. Von der ersten Aussage brauchen wir nur 1 Implikation zu zeigen, da umgekehrt auch die Aufgabe (4.1.2) zu (4.4.1) dual ist. Wir nehmen also an, daß (4.1.2) eine optimale Lösung besitzt. Dann berechnen wir die in (4.4.5) auftretenden Teilmatrizen und erhalten

$$(\alpha) \quad v_0^t = -d^t + g^tA ,$$

$$(\beta) \quad u_0^t = g^t ,$$

$$(\gamma) \quad z_{opt} = g^tb .$$

Es folgt aus (β) und (α)

$$A^tu_0 = A^tg = (g^tA)^t = v_0 + d ,$$

also

$$A^tu_0 - v_0 = d ;$$

wegen u_0, $v_0 \geqq 0$ ist u_0 hiermit zulässige Lösung von (4.4.1). Außerdem haben wir nach (β) und (γ)

$$w(u_0) = b^tu_0 = b^tg = z_{opt} ,$$

woraus nach (4.4.3) unmittelbar $w(u_0) = w_{opt}$ folgt.

Wir wollen nun Anwendungsmöglichkeiten der dualen Aufgabe (4.4.1) erläutern. Zunächst nehmen wir an, es sei uns eine zulässige Lösung x von (4.1.2) bekannt. Wenn dann ζ_d leer ist, so ist wegen Satz (4.4.6) die Zielfunktion z auf ζ nach oben unbeschränkt. Andernfalls liefert ein $u \in \zeta_d$ nach (4.4.3) eine Einschließung für z_{opt}: diese Möglichkeit wird dann angewendet, wenn es zu aufwendig erscheint, eine optimale Lösung von (4.1.2) zu bestimmen.

Manchmal ist die Aufgabe (4.4.1) leichter zu behandeln als die vorgegebene Aufgabe (4.1.2). Wenn wir mit den in 4.2 und 4.3 besprochenen Verfahren eine optimale Lösung von (4.4.1) konstruiert haben, können wir – da (4.1.2) das zu

(4.4.1) duale Problem ist – nach Satz (4.4.6) eine optimale Lösung von (4.4.1) direkt ablesen.

Die zu einem allgemeineren Optimierungsproblem als (4.1.2) duale Aufgabe wird in der Übungsaufgabe 4.7 angegeben.

Übungsaufgaben zum 4. Kapitel

Aufgabe 4.1. Die Optimierungsaufgabe (4.1.1) sei aus (4.1.2) durch die Reduktion (4.1.3) entstanden. Es sei $x \in \mathbb{R}^q$ mit $Ax \leqq b$, $x \geqq 0$ vorgegeben, dazu x_e gemäß (4.1.3). Man zeige:

x_e ist zulässige Basislösung von (4.1.1) genau dann, wenn x eine „Ecke" des durch (4.1.2), (ii) beschriebenen Polyeders ist, das heißt, wenn q linear unabhängige Zeilen der Matrix $\begin{pmatrix} A \\ I_q \end{pmatrix}$ existieren, so daß in den entsprechenden Zeilen der Ungleichungen $Ax \leqq b$, $x \geqq 0$ das Gleichheitszeichen gilt.

Aufgabe 4.2. Es erfülle Ω die Voraussetzung (4.2.4), es gelte $d \leqq 0$, aber nicht $d < 0$. In diesem Fall lassen sich eventuell mehrere optimale Lösungen von (4.1.1) angeben. Man zeige:

(i) Ist $s \in \{1, \ldots, q\}$ mit $d_s = 0$ und $a_s \leqq 0$, so ist für jedes $\lambda \geqq 0$

$$x(\lambda) := Qy(\lambda), \qquad y(\lambda) := \begin{pmatrix} 0 \\ b \end{pmatrix} + \lambda \begin{pmatrix} e_s \\ -a_s \end{pmatrix}$$

optimale Lösung von (4.1.1).

(ii) Ist $s \in \{1, \ldots, q\}$ mit $d_s = 0$ und $a_s \nleqq 0$, so sei r mit

$$\frac{b_r}{a_{r,s}} = \min \left\{ \frac{b_i}{a_{i,s}} : i \in \{1, \ldots, n\},\ a_{i,s} > 0 \right\}$$

gewählt; Ω' entstehe aus Ω durch einen Jordanschen Eliminationsschritt mit dem Pivotelement $a_{r,s}$. Dann erfüllt Ω' wieder die Bedingung (4.2.5), (ii) und liefert eine optimale Basislösung von (4.1.1), nämlich $x' := Q' \begin{pmatrix} 0 \\ b' \end{pmatrix}$. Dabei ist $x' \neq x := Q \begin{pmatrix} 0 \\ b \end{pmatrix}$ genau dann, wenn $b_r \neq 0$ gilt.

Aufgabe 4.3. Es erfülle $\Omega = \Omega^{(\nu)}$ die Voraussetzung (4.2.4); dabei gelte gemäß den in (4.2.12) verwendeten Bezeichnungen $b_{r_0}^{(\nu)} = 0$ für genau ein $r_0 \in \{1, \ldots, n\}$. Man zeige, daß bei der weiteren Durchführung des Simplex-Verfahrens mit einer beliebigen Pivotwahl gemäß (4.2.6) $(d_s > 0)$ stets

$$(*) \quad \{\pi_\nu(q+1), \ldots, \pi_\nu(q+n)\} \neq \{\pi_\mu(q+1), \ldots, \pi_\mu(q+n)\}$$

für alle $\mu > \nu$ gilt.

Anleitung: Aus $\vartheta_\mu > \vartheta_\nu$ folgt (*) unmittelbar; $\vartheta_\mu = \vartheta_\nu$ kann nur eintreten, wenn bei den Schritten $\nu, \nu+1, \ldots, \mu-1$ stets die r_0-te Zeile zur Pivotzeile wird. Man beachte die Vorzeichen von $d_s^{(\nu)}, d_s^{(\nu+1)}, \ldots, d_s^{(\mu)}$, wenn s die Pivotspalte beim ν-ten Schritt indiziert.

Aufgabe 4.4. (Yudin-Gol'shtein [35]). Es sei

$$\Omega^{(1)} := \left(\begin{array}{ccccccc|c} 4 & 1 & 0 & 1 & -2 & -3 & 0 & 0 \\ 1 & 0 & 1 & 4 & -3 & -2 & 0 & 0 \\ 1 & 0 & 0 & 1 & 1 & 1 & 1 & 1 \\ \hline 1 & 0 & 0 & -1 & 1 & -1 & 0 & 0 \end{array}\right)$$

Man rechne nach, daß das Simplex-Verfahren zyklisch wird, wenn in

$$\Omega^{(\nu)} = \left(\begin{array}{c|c} C^{(\nu)} & b^{(\nu)} \\ \hline -f^{(\nu)t} & \vartheta_\nu \end{array}\right), \quad C^{(\nu)} = (c_{i,j}^{(\nu)})_{(3;7)}, \quad f^{(\nu)} = (f_j^{(\nu)})_1^7, \quad b^{(\nu)} = (b_i^{(\nu)})_1^3$$

das Pivotelement $c_{r,s}^{(\nu)}$ nach der folgenden Vorschrift gewählt wird:

$$s := \min\{j \in \{1, \ldots, 7\}: -f_j^{(\nu)} < 0\},$$

$$r := \min\left\{r' \in \{1,2,3\}: \frac{b_{r'}^{(\nu)}}{c_{r',s}^{(\nu)}} = \min\left\{\frac{b_i^{(\nu)}}{c_{i,s}^{(\nu)}} : c_{i,s}^{(\nu)} > 0\right\}\right\}.$$

Aufgabe 4.5. (Nicht vorzeichenbeschränkte Variable)
Vorgegeben sei für $x = (x_i)_1^q \in \mathbb{R}^q$ folgende Optimierungsaufgabe:
Maximiere

$$z(x) := d^t x$$

unter den Restriktionen

$$\begin{cases} Ax \leqq b, \\ x_1, \ldots, x_t \geqq 0. \end{cases}$$

Hierbei sei $A \in M(n \times q, \mathbb{R})$; $d \in \mathbb{R}^q$; $b \in \mathbb{R}^n$, $\geqq 0$ sowie $0 \leqq t < q$. Zur Reduktion auf die Gestalt (4.1.2) führt man Variable x_j^+, x_j^- mit

$$x_j = x_j^+ - x_j^-; \quad x_j^+ \geqq 0; \quad x_j^- \geqq 0 \qquad (j = t+1, \ldots, q)$$

ein; entsprechend erweitert man A und d^t um jeweils $q-t$ Spalten. Man zeige, daß in jeder zulässigen Lösung, die das Simplex-Verfahren liefert, für alle $j \in \{t+1, \ldots, q\}$ stets $x_j^+ = 0$ oder $x_j^- = 0$ gilt.

Aufgabe 4.6. Ergänzend zum Beweis von Satz (4.2.23) zeige man, daß stets

$$Q^{(\nu)} J \neq Q^{(\mu)} JP \qquad (\mu \neq \nu)$$

gilt, falls P beliebige (n, n)-Permutationsmatrix ist. Hieraus folgere man, daß das Simplex-Verfahren nach höchstens $\binom{m}{n}$ Schritten abbricht.

Aufgabe 4.7. Es seien $q_i, n_i \in \mathbb{N}, \geqq 0$ $(i = 1, 2)$; $n := n_1 + n_2 > 0$; $q := q_1 + q_2 > 0$ sowie $b_i \in \mathbb{R}^{n_i}$, $d_j \in \mathbb{R}^{q_j}$, $A_{i,j} \in M(n_i \times q_j, \mathbb{R})$ $(i, j = 1, 2)$. Hiermit sei folgende Optimierungsaufgabe gegeben:

(I) Maximiere

$$z(x) := d_1^t x_1 + d_2^t x_2 \qquad x := \begin{pmatrix} x_1 \\ x_2 \end{pmatrix}, \quad x_j \in \mathbb{R}^{q_j}$$

unter den Restriktionen

$$\begin{cases} A_{1,1} x_1 + A_{1,2} x_2 \leqq b_1, \\ A_{2,1} x_1 + A_{2,2} x_2 = b_2, \\ x_1 \geqq 0. \end{cases}$$

Als zu (I) duale Aufgabe bezeichnet man das Problem

(II) Minimiere

$$w(u) := b_1^t u_1 + b_2^t u_2 \qquad u := \begin{pmatrix} u_1 \\ u_2 \end{pmatrix}, \quad u_i \in \mathbb{R}^{n_i}$$

unter den Restriktionen

$$\begin{cases} A_{1,1}^t u_1 + A_{2,1}^t u_2 \geqq d_1, \\ A_{1,2}^t u_1 + A_{2,2}^t u_2 = d_2 \\ u_1 \geqq 0. \end{cases}$$

Man bringe die Probleme (I) und (II) auf die Form (4.1.2) bzw. (4.4.1), indem man wie in Aufgabe 4.5 neue Variable einführt und die Gleichungen durch Paare von Ungleichungen ersetzt. Man überzeuge sich davon, daß die reduzierten Probleme im Sinne von Abschnitt 4.4 zueinander dual sind.

Aufgabe 4.8. Man löse – unter Benutzung eines Tischrechners – das Optimierungsproblem:

Maximiere

$$z(x) := 2x_1 + 5x_2 + 5x_3$$

unter den Restriktionen

$$\begin{aligned} 0{,}5\,x_1 + 3\,x_2 + 2\,x_3 &\leqq 8, \\ 2\,x_1 - 2\,x_2 - x_3 &\leqq 12, \\ x_1 - x_2 - 4\,x_3 &\leqq -2, \\ x_i \geqq 0 \quad (i = 1, 2, 3). & \end{aligned}$$

Zur Kontrolle lese man die optimale Lösung des dualen Problems an der Endmatrix ab.

Literatur

[1] *F. L. Bauer, J. Heinhold, K. Samelson, R. Sauer:* Moderne Rechenanlagen. Leitfäden der angewandten Mathematik, Bd. 5. Stuttgart, Teubner 1964.

[2] *I. S. Berensin, N. P. Shidkow:* Numerische Methoden 1, 2. Berlin, VEB Deutscher Verlag der Wissenschaften 1970, 1971.

[3] *L. Collatz:* Funktionalanalysis und numerische Mathematik. Die Grundlehren der mathematischen Wissenschaften. Berlin-Heidelberg-New York, Springer 1964.

[4] *L. Collatz, W. Wetterling:* Optimierungsaufgaben. 2. Aufl. Heidelberger Taschenbücher. Berlin-Heidelberg-New York, Springer 1971.

[5] *D. K. Faddejew, W. N. Faddejewa:* Numerische Methoden der linearen Algebra, München-Wien, Oldenbourg 1970.

[6] *G. E. Forsythe, C. B. Moler:* Computer-Verfahren für lineare algebraische Systeme. Verfahren der Datenverarbeitung. München-Wien, Oldenbourg 1971.

[7] *S. I. Gass:* Linear Programming. 3rd ed. New York, McGraw-Hill 1969.

[8] *G. Hämmerlin:* Numerische Mathematik I. Hochschultaschenbücher. Mannheim, Bibliographisches Inst. 1970.

[9] *R. W. Hamming:* Numerical Methods for Scientists and Engeneers. New York-Toronto-London. McGraw-Hill 1962.

[10] *P. Henrici:* Elemente der Numerischen Analysis I, II. Hochschultaschenbücher. Mannheim, Bibliographisches Inst. 1972.

[11] *F. B. Hildebrand:* Introduction to Numerical Analysis. New York-Toronto-London, McGraw-Hill 1956.

[12] *A. S. Householder:* Principles of Numerical Analysis. New York-Toronto-London, McGraw-Hill 1953.

[13] *A. S. Householder:* The Theory of Matrices in Numerical Analysis. New York, Blaisdell Publishing Company 1965.

[14] *E. Isaacson, H. B. Keller:* Analysis of Numerical Methods. New York, Wiley 1966.

[15] *L. V. Kantorovich, G. P. Akilov:* Functional Analysis in Normed Linear Spaces. Pergamon Press, 1964.

[16] *T. Kato:* Perturbation Theory for Linear Operators. Die Grundlehren der mathematischen Wissenschaften. New York, Springer 1966.

[17] *C. Lanczos:* Applied Analysis. Englewood Cliffs, Prentice Hall 1956.

[18] *W. E. Milne:* Numerical Calculus. Princeton University Press 1949.

[19] *B. Noble:* Applied Linear Algebra. Englewood Cliffs, Prentice Hall 1969.

[20] *A. Ralston:* A First Course in Numerical Analysis. New York, McGraw-Hill 1965.

[21] *A. Ralston, H. S. Wilf:* Mathematische Methoden für Digitalrechner I, II. München-Wien, Oldenbourg 1967, 1969.

[22] *H. R. Schwarz, H. Rutishauser, E. Stiefel:* Numerik symmetrischer Matrizen. Leitfäden der angewandten Mathematik, Bd. 11. Stuttgart, Teubner 1968.

[23] *E. Stiefel:* Einführung in die Numerische Mathematik. Leitfäden der angewandten Mathematik, Bd. 2. Stuttgart, Teubner 1963.

[24] *J. Stoer:* Einführung in die Numerische Mathematik I, 2. Aufl. Heidelberger Taschenbücher. Berlin-Heidelberg-New York, Springer 1976.

[25] *J. Stoer, R. Bulirsch:* Einführung in die Numerische Mathematik II. Heidelberger Taschenbücher. Berlin-Heidelberg-New York, Springer 1973.

[26] *F. Stummel, K. Hainer:* Praktische Mathematik. Teubner Studienbücher. Stuttgart, Teubner 1971.

[27] *E. C. Titchmarsh:* The Theory of Functions. 2nd ed. Oxford University Press 1939.

[28] *J. Varga:* Praktische Optimierung. München-Wien, Oldenbourg 1974.

[29] *H. Werner:* Praktische Mathematik I, Methoden der linearen Algebra. Hochschultext. Berlin-Heidelberg-New York, Springer 1970.

[30] *H. Werner, R. Schaback:* Praktische Mathematik II, Methoden der Analysis. Hochschultext. Berlin-Heidelberg-New York, Springer 1972.

[31] *J. H. Wilkinson:* Rundungsfehler. Heidelberger Taschenbücher. Berlin-Heidelberg-New York, Springer 1969.

[32] *J. H. Wilkinson:* The Algebraic Eigenvalue Problem. Oxford, Clarendon Press 1965.

[33] *J. H. Wilkinson:* Error analysis of direct methods of matrix inversion. J. Ass. comp. Mach. 8 (1961), 281–330.

[34] *J. H. Wilkinson, Ch. Reinsch:* Linear Algebra. Handbook for Automatic Computation, Vol. II. Grundlehren der mathematischen Wissenschaften. Berlin-Heidelberg-New York, Springer 1971.

[35] *D. B. Yudin, E. G. Gol'shtein:* Linear Programming. Israel Program for Scientific Translations, Jerusalem 1965.

In der angegebenen Literatur finden sich weitere Literaturhinweise.

Sachregister

Kurzbiographie der Autoren

Reinhard Mennicken wurde 1935 in Köln geboren. Studium der Mathematik und Physik 1957–1962 in Köln; dort 1963 Promotion und 1969 Habilitation. 1969–1971 Dozent an der Universität Konstanz. 1971–74 Abteilungsvorsteher und Professor an der Universität Regensburg. 1972–74 Lehrstuhlvertretungen in Erlangen und Braunschweig. 1974/75 o. Professor für Mathematik an der TU Braunschweig. Seit 1975 Professor an der Universität Regensburg und Leiter einer Abteilung für Angewandte Mathematik. Forschungsschwerpunkte sind Differentialgleichungen, Funktionalanalysis und Numerische Mathematik.

Ekkehard Wagenführer wurde 1944 in Apolda/Thüringen geboren. Studium der Mathematik und Physik 1963–69 in Köln; dort 1971 Promotion. Seit 1972 Akademischer Rat beim Fachbereich Mathematik der Universität Regensburg. Forschungsschwerpunkte sind Differentialgleichungen und Numerische Mathematik.