A. KIEŁBASIŃSKI
H. SCHWETLICK

Numerische lineare ALGEBRA

VEB DEUTSCHER
VERLAG DER
WISSENSCHAFTEN

Mathematik für Naturwissenschaft und Technik

Herausgegeben von H. Heinrich und H. Schubert

Band 18

Numerische lineare Algebra

Eine computerorientierte Einführung

von A. Kiełbasiński und H. Schwetlick

Mit 19 Abbildungen

VEB Deutscher Verlag der Wissenschaften
Berlin 1988

ISBN 3-326-00194-0
ISSN 0543-100X

Verlagslektor: Erika Arndt
Verlagshersteller: Norma Braun
Umschlaggestaltung: Hartwig Hoeftmann

Lizenz-Nr. 206 · 435/127/88
Printed in the German Democratic Republic
Gesamtherstellung: VEB Druckhaus „Maxim Gorki", DDR - 7400 Altenburg
LSV 1024
Bestellnummer: 571 555 2
03980

Vorwort

Die numerische lineare Algebra nimmt zweifellos eine zentrale Stellung innerhalb der Numerischen Mathematik ein. Von der Theorie her ist sie eines der am besten aufgearbeiteten Gebiete. Es gibt nur wenige Anwendungen der Mathematik, wo sie nicht direkt oder mittelbar angesprochen wird. Ihre Geschichte läßt sich über mehr als 2000 Jahre zurückverfolgen, und seit Beginn des Computerzeitalters ist ihre Entwicklung stürmischer als je zuvor vorangetrieben worden.

Der Zwang, Aufgaben der linearen Algebra auf verfügbaren Computern schnell, zuverlässig und in großer Anzahl lösen zu müssen, hat frühzeitig zur Auswahl, Entwicklung und Analyse von dafür geeigneten Algorithmen und zu deren sachgemäßer Implementierung — meist in Form von Programmpaketen — geführt. Es zeigte sich, daß die bei Computerrechnung mit endlicher Stellenzahl notwendig auftretenden Rundungsfehler einen wesentlichen Einfluß auf die berechneten Resultate haben können. Die Untersuchung des Verhaltens von Algorithmen bei Realisierung in Computerarithmetik wurde daher zu einem zentralen Gegenstand der numerischen linearen Algebra. Entscheidende Fortschritte brachte um die sechziger Jahre die im wesentlichen von WILKINSON entwickelte und perfekt angewendete Methode der „backward error analysis", die eine theoretische Erfassung der Rundungsfehler und daher gesicherte Aussagen über das Stabilitätsverhalten von Algorithmen erlaubt. In seinem Buch „The Algebraic Eigenvalue Problem" — gemeinhin „Bibel" der numerischen linearen Algebra genannt — hat WILKINSON 1965 eine auch heute noch gültige Zusammenfassung der wesentlichen Ergebnisse gegeben und damit die spätere Entwicklung entscheidend beeinflußt.

In den darauffolgenden Jahren entstanden weiterführende Darstellungen zu Teilgebieten der linearen Algebra, etwa LAWSON/HANSONS „Solving Least Squares Problems" aus dem Jahre 1974 und PARLETTS „The Symmetric Eigenvalue Problem" aus dem Jahre 1980. Alle diese anspruchsvollen Monographien erschließen sich jedoch dem Lernenden nicht leicht; sie sprechen mehr den erfahrenen, bereits mit speziellen Kenntnissen der numerischen linearen Algebra vertrauten Leser an. Obwohl die grundlegenden Ideen in entsprechend vereinfachter Form bereits Eingang in neuere einführende Gesamtdarstellungen der Numerischen Mathematik wie etwa die von STOER oder MAESS gefunden haben, fehlten Lehrbücher, die einen breiten Interessentenkreis systematisch zu einem tieferen Verständnis dieser neuen Entwicklung hätten führen können. Dieser Mangel wurde deutlicher, als exzellente Implementierungen der wichtigsten Algorithmen weltweit verfügbar wurden; wir erwähnen das „Handbook for Automatic Computation. Vol. II: Linear Algebra" aus dem Jahre

1971, die 1974 bzw. 1977 erschienenen beiden EISPACK-Versionen und das 1979 publizierte LINPACK.

Das vorliegende Lehrbuch ist — nach GOLUB/VAN LOANS „Matrix Computations" aus dem Jahre 1983 und der gleichnamigen „Numerischen linearen Algebra" von BUNSE/BUNSE-GERSTNER aus dem Jahre 1985 — nach Wissen der Autoren das dritte, das die erwähnte Lücke zwischen den anspruchsvollen Monographien und den einführenden Darstellungen zu schließen versucht. Es wendet sich an Mathematiker, aber auch an mathematisch interessierte Naturwissenschaftler und Ingenieure, die sich intensiver mit numerischer linearer Algebra vertraut machen wollen und nicht nur an einer kurzen, übersichtsartigen Zusammenfassung interessiert sind.

Um dieser Zielstellung gerecht zu werden, haben wir uns auf die Behandlung der direkten Verfahren für lineare Gleichungssysteme und Quadratmittelprobleme und die des Eigenwertproblems symmetrischer Matrizen beschränkt. Diese Gebiete sind jedoch vergleichsweise gründlich unter Einbeziehung verschiedenster Gesichtspunkte bis hin zu Fragen der Implementierung erörtert worden. Es wurde eine weitgehend abgeschlossene Darstellung gewählt, die im wesentlichen nur die aus der mathematischen Grundausbildung der Ingenieure bekannten Vorkenntnisse voraussetzt. Um mögliche Unterschiede in diesen Kenntnissen auszugleichen, sind die später benötigten Begriffe und Fakten in einem einführenden ersten Kapitel nochmals übersichtsartig zusammengestellt worden. Im übrigen Text wurden fast alle Beweise im Detail ausgeführt — anfangs ausführlicher, zum Schluß dann knapper —, so daß das Buch auch zum Selbststudium geeignet sein dürfte. In dieser Hinsicht unterscheidet sich die vorliegende Darstellung von der GOLUB/VAN LOANS, wo in den Beweisen zumeist auf die Originalliteratur oder andere Monographien verwiesen wird; dafür ist allerdings dort ein wesentlich umfangreicheres Gebiet behandelt worden. Verzichtet wurde lediglich auf die Herleitung von gewissen Ergebnissen der Rundungsfehleranalyse, wenn der Aufwand dafür ein vertretbares Maß überschritten hätte. Viel Wert wurde auf präzise Begriffsbildungen, die Herausarbeitung der algorithmischen Grundideen und — stärker als etwa bei BUNSE/BUNSE-GERSTNER — auf die Untersuchung des Einflusses der Rundungsfehler und besonders auf die sachgemäße Interpretation der Ergebnisse der Rundungsfehleranalyse gelegt. Nach den Erfahrungen der Autoren macht das letztgenannte Problem nicht nur Anfängern Schwierigkeiten, so daß für den Komplex „Aufgaben, Computer, Algorithmen" ein separates Kapitel vorgesehen wurde, in dem die Grundbegriffe der Fehleranalyse und verwandte Begriffe in systematischer Weise entwickelt werden. Der den Grundlagen gewidmete erste Teil schließt mit einem Kapitel über elementare Transformationsmatrizen. Die weiteren drei Teile haben die bereits genannten Problemklassen zum Gegenstand. Sie beginnen jeweils mit der Darstellung der wesentlichen theoretischen Eigenschaften, wobei insbesondere auf Ergebnisse der Störungstheorie und auf Residualkriterien eingegangen wird. In den algorithmisch orientierten Abschnitten wird zunächst der „mathematische" Algorithmus motiviert und formuliert. Es folgen Bemerkungen zur Implementierung, die Formulierung des „Computeralgorithmus" in einer sich selbst erklärenden, einfachen Pseudo-Programmiersprache, Aufwandsbetrachtungen, Rundungsfehleranalyse und deren Interpretation sowie ergänzende und weiterführende Bemerkungen. Da die Algorithmen zum Schluß

immer komplexer werden und eine detaillierte computernahe Formulierung mit akzeptablem Aufwand kaum noch möglich ist, wird ab Kapitel 11 auf die gesonderte Formulierung des Computeralgorithmus verzichtet.

Jedem Kapitel sind Bemerkungen nachgestellt, in denen die historische Entwicklung unter Angabe grundlegender Arbeiten skizziert wird und Hinweise auf weiterführende Resultate zu finden sind. Im Text selbst wird bis auf wenige Ausnahmen keine Literatur zitiert. Jeder der den drei behandelten Aufgabenklassen gewidmeten Teile schließt mit einer kurzen Zusammenfassung, in der die vorgestellten Algorithmen verglichen und Empfehlungen zu ihrer Anwendung gegeben werden. Außerdem wird auch auf im Text nicht behandelte Aufgaben- und Algorithmenklassen eingegangen. Die ins Literaturverzeichnis aufgenommenen Arbeiten können selbstverständlich nur als eine Auswahl angesehen werden; eine Zusammenstellung der fast unübersehbar gewordenen Literatur zur linearen Algebra geben die beiden von FADDEEVA und Mitarbeitern erarbeiteten Bibliographien.

Es ist den Autoren ein Bedürfnis, an dieser Stelle all denen herzlich zu danken, die ihnen im Laufe der mehr als vier Jahre dauernden Erarbeitung des Manuskriptes zur Seite gestanden haben: Die Universität Warschau und die Martin-Luther-Universität Halle ermöglichten uns auch unter schwierigen Bedingungen die für ein solches Projekt erforderlichen wechselseitigen Besuche. Förderlich wirkten sich auch Einladungen des erstgenannten Autors zu Gastvorlesungen an die Technische Universität Dresden aus, wo der zweitgenannte bis 1979 tätig gewesen ist. Herr Prof. Dr. Dr. h. c. H. HEINRICH hat als Mitherausgeber der Reihe, in der dieses Buch erscheint, zahlreiche Anregungen für die Formulierung der Schlußfassung gegeben. Unsere Kollegen Prof. Dr. G. ZIELKE und Dr. O. KNOTH waren uns bei der Beschaffung von Literatur behilflich; Prof. ZIELKE hat auch Teile des Manuskriptes kritisch gelesen. Viele andere Kollegen — stellvertretend sei Dr. V. TILLER genannt — unterstützten uns bei der Bewältigung der technischen Arbeiten tatkräftig. Der Deutsche Verlag der Wissenschaften und insbesondere Frau Dipl.-Math. E. ARNDT haben die Entstehung dieses Buches mit Verständnis und Geduld gefördert. Zu herzlichem Dank sind die Autoren Frau E. SCHÄFER verpflichtet, die den größten Teil des Manuskriptes sorgfältig geschrieben hat.

Unser besonderer Dank gilt unseren Familien, die während einer längeren Zeit mancher Belastung ausgesetzt waren und die durch ihr Verständnis, aber auch durch ihre aktive Unterstützung zum Gelingen des Projektes beigetragen haben.

Die Autoren hoffen, daß sich die Arbeit gelohnt hat und der Leser mit Hilfe dieses Buches ein tieferes Verständnis der Probleme der numerischen linearen Algebra gewinnt. Vor allem möge er in die Lage versetzt werden, Möglichkeiten und Grenzen der verschiedenen Algorithmen richtig einzuschätzen, die verfügbare Software effektiv zu nutzen und diese Software gegebenenfalls anzupassen, zu modifizieren oder zu ergänzen, wenn spezielle, von Standardaufgaben abweichende Probleme dies erfordern.

Warschau/Halle, Frühjahr 1986

A. KIEŁBASIŃSKI
H. SCHWETLICK

Inhalt

Teil I. Grundlagen 11

1. Grundbegriffe der linearen Algebra 11
   1.1. Vektoren, Matrizen, Normen 11
   1.2. Eigenwerte, Singulärwerte 31

2. Aufgaben, Computer, Algorithmen 45
   2.1. Aufgabenklassen der linearen Algebra 45
   2.2. Computerarithmetik 63
   2.3. Numerische Algorithmen und Grundlagen der Fehleranalyse 73

3. Elementare Transformationsmatrizen 96
   3.1. Permutationsmatrizen 97
   3.2. Nichtorthogonale elementare Transformationsmatrizen 100
   3.3. Orthogonale elementare Transformationsmatrizen 106

Teil II. Reguläre lineare Gleichungssysteme 127

4. Grundlegende Fakten über reguläre Gleichungssysteme 127
   4.1. Störungstheorie und Skalierung 127
   4.2. Prinzip der direkten Lösungsverfahren 142
   4.3. Reguläre Gleichungssysteme einfacher Struklur 143

5. Der Gaußsche Algorithmus 153
   5.1. Die Grundform des Gaußschen Algorithmus 153
   5.2. Pivotisierung 161
   5.3. Rundungsfehleranalyse 168
   5.4. Genauigkeitsabschätzung und iterative Verbesserung 174

6. Modifikationen des Gaußschen Algorithmus 184
   6.1. Dreiecksfaktorisierungen symmetrischer Matrizen 184
   6.2. Direkte Dreiecksfaktorisierungen 201
   6.3. Aufdatierung von Dreiecksfaktorisierungen 207
   6.4. Dreiecksfaktorisierung schwach besetzter Matrizen 220
   6.5. Berechnung und Aufdatierung der inversen Matrix 229

7. Zusammenfassung zum Teil II 237

Teil III. Lineare Quadratmittelprobleme 240

8. Rechteckige Gleichungssysteme und Quadratmittelprobleme 240
   8.1. Klassische Lösungen, Quadratmittellösungen, Pseudoinverse 240

8.2. Störungstheorie 251
8.3. Klassifikation und statistische Interpretation von Quadratmittelproblemen . . 264

9. Normalgleichungsverfahren 270
9.1. Aufstellung und Lösung der Normalgleichungen 270
9.2. Iterative Verbesserung 275

10. Orthogonalisierungsverfahren 277
10.1. Orthogonalisierung nach Gram-Schmidt 277
10.2. Orthogonalisierung nach Householder und Givens 285
10.3. Aufdatierung von $\boldsymbol{QR}$-Faktorisierungen 301

11. Rangdefiziente Quadratmittelprobleme 315
11.1. Numerische Berechnung der Singulärwertzerlegung 316
11.2. Diskrete Regularisierung 325
11.3. Kontinuierliche Regularisierung 338

12. Zusammenfassung zum Teil III 350

Teil IV. Eigenwertprobleme 353

13. Das spezielle symmetrische Eigenwertproblem 353
13.1. Grundlegende Eigenschaften, Störungstheorie, Residualkriterien 354
13.2. Das Jacobi-Verfahren 365
13.3. Vektor- und Teilraumiteration 373
13.4. Inverse Iteration 389
13.5. Orthogonale Ähnlichkeitstransformation auf Tridiagonalform 397
13.6. Schneiden des Spektrums und Bisektionsverfahren 406
13.7. Der $\boldsymbol{QR}$-Algorithmus 415

14. Das allgemeine symmetrische Eigenwertproblem 434
14.1. Grundlegende Eigenschaften 434
14.2. Explizite Reduktion auf ein spezielles Eigenwertproblem 438
14.3. Vektor- und Teilraumiteration 441

15. Zusammenfassung zum Teil IV 448

16. Software für Aufgabenklassen der numerischen linearen Algebra 450

Literatur 457

Sachverzeichnis 468

I. Grundlagen

1. Grundbegriffe der linearen Algebra

Dieses Kapitel dient der Einführung der später benötigten grundlegenden Begriffe und Aussagen der linearen Algebra. Ein großer Teil des behandelten Stoffes — speziell im Abschnitt 1.1 — ist Gegenstand der mathematischen Grundausbildung an den Hochschulen und wird deshalb nur übersichtsartig zusammengestellt. Weiterführende, dem Naturwissenschaftler und Ingenieur in der Regel nicht vertraute Ergebnisse werden in Form von Aussagen und Sätzen dargeboten. Allerdings muß auch hier auf detaillierte Beweise verzichtet werden; für diese sei auf die Spezialliteratur verwiesen, siehe Bemerkung B 1.1.

1.1. Vektoren, Matrizen, Normen

A. *Vektoren und elementare Vektoroperationen*

Der Vektorbegriff ist ein zentraler Begriff der linearen Algebra. Unter einem Vektor — genauer: einem *m-dimensionalen Spaltenvektor* — verstehen wir ein m-Tupel

$$\boldsymbol{x} = \begin{pmatrix} x_1 \\ x_2 \\ \vdots \\ x_m \end{pmatrix}$$

von reellen Zahlen $x_1, x_2, \ldots, x_m$, den *Komponenten* des Vektors $\boldsymbol{x}$. Die Menge aller m-dimensionalen Vektoren bezeichnen wir mit $\mathbf{R}^m$; für $\mathbf{R}^1$ — die Menge der reellen Zahlen — schreiben wir einfach $\mathbf{R}$. Für die Bezeichnung von Vektoren verwenden wir i. allg. kleine lateinische Buchstaben $\boldsymbol{x}, \boldsymbol{y}, \ldots$ Die zugehörigen unten indizierten Größen $x_i, y_i, \ldots$ stellen die Komponenten des jeweiligen Vektors dar; gelegentlich benutzen wir auch die Schreibweise $(\boldsymbol{x})_i$. Für reelle Zahlen, die nicht Komponenten von Vektoren sind, werden i. allg. kleine griechische Buchstaben gewählt.

Zwei Vektoren $\boldsymbol{x}, \boldsymbol{y} \in \mathbf{R}^m$ sind *gleich*, wenn $x_i = y_i$ $(i = 1, \ldots, m)$ ist. Der Vektor, dessen sämtliche Komponenten 0 sind, heißt *Nullvektor* $\boldsymbol{o} \in \mathbf{R}^m$.

In der Menge $\mathbf{R}^m$ ist für zwei Vektoren $\boldsymbol{x}, \boldsymbol{y}$ gemäß $(\boldsymbol{x} + \boldsymbol{y})_i := x_i + y_i$ die *Summe* $\boldsymbol{x} + \boldsymbol{y}$ definiert. Analog ist das *Produkt* $\lambda\boldsymbol{x}$ bzw. $\boldsymbol{x}\lambda$ von $\boldsymbol{x} \in \mathbf{R}^m$ mit einer Zahl $\lambda \in \mathbf{R}$ durch $(\lambda\boldsymbol{x})_i := (\boldsymbol{x}\lambda)_i := \lambda \cdot x_i$ erklärt. Das Zeichen „:=" wird dabei zur Definition der auf der Seite des Doppelpunktes stehenden Größe verwendet. Es gilt $\boldsymbol{x} + \boldsymbol{y} \in \mathbf{R}^m$ sowie $\lambda\boldsymbol{x} \in \mathbf{R}^m$ für alle $\boldsymbol{x}, \boldsymbol{y} \in \mathbf{R}^m$ und $\lambda \in \mathbf{R}$, und aus später klar werdenden Gründen wird $\mathbf{R}^m$ daher *m-dimensionaler linearer Raum* genannt.

Die soeben eingeführten Operationen genügen den Rechenregeln

$$\boldsymbol{x} + \boldsymbol{y} = \boldsymbol{y} + \boldsymbol{x}, \qquad (\boldsymbol{x} + \boldsymbol{y}) + \boldsymbol{z} = \boldsymbol{x} + (\boldsymbol{y} + \boldsymbol{z}) =: \boldsymbol{x} + \boldsymbol{y} + \boldsymbol{z}, \tag{1}$$

$$(\lambda\mu)\,\boldsymbol{x} = \lambda(\mu\boldsymbol{x}), \tag{2}$$

$$(\lambda + \mu)\,\boldsymbol{x} = \lambda\boldsymbol{x} + \mu\boldsymbol{x}, \qquad \lambda(\boldsymbol{x} + \boldsymbol{y}) = \lambda\boldsymbol{x} + \lambda\boldsymbol{y} \tag{3}$$

für alle $\boldsymbol{x}, \boldsymbol{y}, \boldsymbol{z} \in \mathbf{R}^m$ und alle $\lambda, \mu \in \mathbf{R}$.

Das folgende Beispiel zeigt, daß die auf der rechten und linken Seite der Gleichheitszeichen in (2), (3) stehenden Ausdrücke zwar dieselben Größen definieren, ihre Berechnung in der durch die Klammersetzung festgelegten Reihenfolge jedoch i. allg. einen unterschiedlichen Aufwand erfordert:

1.1.1. *Beispiel.* Man betrachte $\boldsymbol{u} := (\lambda + \mu)\,\boldsymbol{x}$ und $\boldsymbol{v} := \lambda\boldsymbol{x} + \mu\boldsymbol{x}$ für gegebenes $x \in \mathbf{R}^m$, $\lambda, \mu \in \mathbf{R}$, sowie die zugehörigen Berechnungsvorschriften

V_1: $\alpha := \lambda + \mu$
for $i := 1(1)m$ **do** $u_i := \alpha * x_i$,

V_2: **for** $i := 1(1)m$ **do** $v_i := (\lambda * x_i) + (\mu * x_i)$.

V_1 erfordert eine Addition und m Multiplikationen, während V_2 m Additionen und $2m$ Multiplikationen erfordert, obwohl beide Vorschriften denselben Vektor $\boldsymbol{u} = \boldsymbol{v}$ definieren. □

In Berechnungsvorschriften bedeutet das Zeichen „:=", daß der links stehenden Variablen der Wert des rechtsstehenden Ausdrucks zugewiesen wird.

Das Beispiel zeigt, daß in der numerischen linearen Algebra im Gegensatz zur theoretischen linearen Algebra zwischen der *Definition* einer Größe durch einen Ausdruck und der geeignet vorzunehmenden *numerischen Auswertung* dieses Ausdruckes durch eine Berechnungsvorschrift unterschieden werden muß. Neben dem Aufwand ist dabei auch der Einfluß der bei Rechnung mit endlicher Stellenzahl notwendig auftretenden Rundungsfehler zu beachten, siehe Abschnitt 2.2 und 2.3 für eine ausführliche Diskussion dieses Problemkreises.

B. *Linearkombinationen und Teilräume*

Wir wählen jetzt n Vektoren $\boldsymbol{a}^1, \ldots, \boldsymbol{a}^n \in \mathbf{R}^m$ und Zahlen $\lambda_1, \ldots, \lambda_n \in \mathbf{R}$ aus; zur Unterscheidung von Vektoren werden stets obere Indizes verwendet, während reelle Zahlen unten indiziert werden. Der Vektor

$$\boldsymbol{y} = \lambda_1\boldsymbol{a}^1 + \lambda_2\boldsymbol{a}^2 + \cdots + \lambda_n\boldsymbol{a}^n$$

ist dann wieder ein Vektor aus $\mathbf{R}^m$ und heißt *Linearkombination* der Vektoren $\boldsymbol{a}^1, \ldots, \boldsymbol{a}^n$ mit den *Koeffizienten* $\lambda_1, \ldots, \lambda_n$. Die Menge aller Linearkombinationen von

$\boldsymbol{a}^1, \ldots, \boldsymbol{a}^n$ wird mit

$$\operatorname{span}\{\boldsymbol{a}^1, \ldots, \boldsymbol{a}^n\} := \{\boldsymbol{y} \in \mathbf{R}^m : \boldsymbol{y} = \lambda_1 \boldsymbol{a}^1 + \cdots + \lambda_n \boldsymbol{a}^n;\ \lambda_1, \ldots, \lambda_n \in \mathbf{R}\}$$

bezeichnet.

Für $\boldsymbol{y} = \sum_{j=1}^{n} \lambda_j \boldsymbol{a}^j$, $\boldsymbol{y}' = \sum_{j=1}^{n} \lambda_j' \boldsymbol{a}^j$ und $\alpha \in \mathbf{R}$ gilt $\boldsymbol{y} + \boldsymbol{y}' = \sum_{j=1}^{n} (\lambda_j + \lambda_j') \boldsymbol{a}^j$ sowie $\alpha \boldsymbol{y} = \sum_{j=1}^{n} (\alpha \lambda_j) \boldsymbol{a}^j$, d. h., auch $\boldsymbol{y} + \boldsymbol{y}'$ und $\alpha \boldsymbol{y}$ sind Linearkombinationen von $\boldsymbol{a}^1, \ldots, \boldsymbol{a}^n$. Daher hat $X := \operatorname{span}\{\boldsymbol{a}^1, \ldots, \boldsymbol{a}^n\} \subset \mathbf{R}^m$ folgende Eigenschaften:

$$\text{Aus } \boldsymbol{y}, \boldsymbol{y}' \in X \text{ und } \alpha \in \mathbf{R} \text{ folgt } \boldsymbol{y} + \boldsymbol{y}' \in X \text{ sowie } \alpha \boldsymbol{y} \in X.$$

Eine Teilmenge X von $\mathbf{R}^m$ mit diesen Eigenschaften heißt (linearer) *Teilraum* von $\mathbf{R}^m$. Speziell heißt $\operatorname{span}\{\boldsymbol{a}^1, \ldots, \boldsymbol{a}^n\}$ der *durch die Vektoren* $\boldsymbol{a}^1, \ldots, \boldsymbol{a}^n$ *aufgespannte Teilraum*. Im Dreidimensionalen ($m = 3$) spannt ein Vektor $\boldsymbol{a}^1 \neq \boldsymbol{o}$ eine Gerade durch den Ursprung auf, und zwei Vektoren $\boldsymbol{a}^1, \boldsymbol{a}^2$, die nicht Vielfache voneinander sind, spannen eine durch den Ursprung gehende Ebene auf.

Speziell sind $X = \{\boldsymbol{o}\}$ — die nur aus dem Nullvektor bestehende Menge — und $X = \mathbf{R}^m$ selbst Teilräume von $\mathbf{R}^m$, so daß insbesondere die Bezeichnung von $\mathbf{R}^m$ als linearer Raum berechtigt ist.

Die m-dimensionalen *Koordinatenvektoren*

$$\boldsymbol{e}^i := \begin{pmatrix} 0 \\ \vdots \\ 0 \\ 1 \\ 0 \\ \vdots \\ 0 \end{pmatrix} \leftarrow i\text{-te Komponente } (i = 1, \ldots, m)$$

spannen den Raum $\mathbf{R}^m$ auf, denn es gilt $\boldsymbol{x} = \sum_{j=1}^{m} x_j \boldsymbol{e}^j$ für jedes $\boldsymbol{x} \in \mathbf{R}^m$. Andererseits brauchen m beliebige Vektoren nicht notwendig den ganzen Raum aufzuspannen, wie das folgende Beispiel zeigt.

1.1.2. *Beispiel.* Für

$$\boldsymbol{a}^1 = \begin{pmatrix} 1 \\ 0 \\ 0 \end{pmatrix}, \qquad \boldsymbol{a}^2 = \begin{pmatrix} 0 \\ 1 \\ 0 \end{pmatrix}, \qquad \boldsymbol{a}^3 = \begin{pmatrix} 1 \\ 1 \\ 0 \end{pmatrix}$$

gilt $\boldsymbol{a}^3 = \boldsymbol{a}^1 + \boldsymbol{a}^2$. Jede Linearkombination $\boldsymbol{y} = \lambda_1 \boldsymbol{a}^1 + \lambda_2 \boldsymbol{a}^2 + \lambda_3 \boldsymbol{a}^3$ läßt sich daher als $\boldsymbol{y} = \lambda_1 \boldsymbol{a}^1 + \lambda_2 \boldsymbol{a}^2 + \lambda_3 (\boldsymbol{a}^1 + \boldsymbol{a}^2) = (\lambda_1 + \lambda_3) \boldsymbol{a}^1 + (\lambda_2 + \lambda_3) \boldsymbol{a}^2$ schreiben, woraus

$$\operatorname{span}\{\boldsymbol{a}^1, \boldsymbol{a}^2, \boldsymbol{a}^3\} = \operatorname{span}\{\boldsymbol{a}^1, \boldsymbol{a}^2\}$$

folgt.

Es ist daher sinnvoll, zu fragen, wieviel Vektoren höchstens benötigt werden,

um einen vorgegebenen Teilraum aufzuspannen. Zur Erörterung dieser Frage wird der fundamentale Begriff der linearen Unabhängigkeit eingeführt: Die Vektoren $\boldsymbol{a}^1, \ldots, \boldsymbol{a}^n \in \mathbf{R}^m$ heißen *linear unabhängig*, wenn aus $\lambda_1 \boldsymbol{a}^1 + \cdots + \lambda_n \boldsymbol{a}^n = \boldsymbol{o}$ folgt, daß $\lambda_1 = \cdots = \lambda_n = 0$ ist, d. h., wenn nur die Linearkombination mit sämtlich verschwindenden Koeffizienten den Nullvektor ergibt. Äquivalent dazu ist die mehr geometrische Charakterisierung, daß keiner der Vektoren $\boldsymbol{a}^1, \ldots, \boldsymbol{a}^n$ beim Aufspannen von span $\{\boldsymbol{a}^1, \ldots, \boldsymbol{a}^n\}$ weggelassen werden kann. Im Beispiel 1.1.2 gilt $\boldsymbol{a}^1 + \boldsymbol{a}^2 - \boldsymbol{a}^3 = \boldsymbol{o}$ mit nicht identisch verschwindenden Koeffizienten, und $\boldsymbol{a}^3$ kann beim Aufspannen von span $\{\boldsymbol{a}^1, \boldsymbol{a}^2, \boldsymbol{a}^3\}$ weggelassen werden.

Die Vektoren $\boldsymbol{a}^1, \ldots, \boldsymbol{a}^n$ heißen *linear abhängig*, wenn sie nicht linear unabhängig sind. Dann kann wie im Beispiel mindestens einer dieser Vektoren als Linearkombination der übrigen ausgedrückt und daher beim Aufspannen von span $\{\boldsymbol{a}^1, \ldots, \boldsymbol{a}^n\}$ weggelassen werden. Dieser Prozeß ist so lange fortsetzbar, bis $r \leqq n$ linear unabhängige Vektoren erhalten werden — diese seien ohne Einschränkung der Allgemeinheit $\boldsymbol{a}^1, \ldots, \boldsymbol{a}^r$ —, die span $\{\boldsymbol{a}^1, \ldots, \boldsymbol{a}^n\}$ = span $\{\boldsymbol{a}^1, \ldots, \boldsymbol{a}^r\}$ aufspannen.

Eine Menge von linear unabhängigen Vektoren $\boldsymbol{a}^1, \ldots, \boldsymbol{a}^r$, die einen Teilraum X aufspannen, wird *Basis* von X genannt. Zu jedem Teilraum $X \neq \{\boldsymbol{o}\}$ gibt es unendlich viele verschiedene Basen, aber jede von diesen besteht aus derselben Anzahl r linear unabhängiger Vektoren, die *Dimension* von X — in Zeichen: $r = \dim(X)$ — genannt wird.

Die beiden extremalen Teilräume $X = \{\boldsymbol{o}\}$ bzw. $X = \mathbf{R}^m = \text{span}\,\{\boldsymbol{e}^1, \ldots, \boldsymbol{e}^m\}$ von $\mathbf{R}^m$ haben die Dimension 0 bzw. m, und für jeden Teilraum $X \subset \mathbf{R}^m$ gilt $0 \leqq r = \dim(X) \leqq m$. Insbesondere sind also mehr als m Vektoren aus $\mathbf{R}^m$ stets linear abhängig.

C. *Matrizen und elementare Matrixoperationen*

Neben dem Vektorbegriff spielt der Begriff einer Matrix in der linearen Algebra eine entscheidende Rolle. Unter einer Matrix — genauer: einer (m,n)-*Matrix* bzw. *Matrix vom Format* (m, n) — verstehen wir ein Schema

$$\boldsymbol{A} = \begin{pmatrix} a_{11} & a_{12} & \ldots & a_{1n} \\ a_{21} & a_{22} & \ldots & a_{2n} \\ \vdots & \vdots & & \vdots \\ a_{m1} & a_{m2} & \ldots & a_{mn} \end{pmatrix} = (a_{ij})$$

von $m \cdot n$ reellen Zahlen a_{ij} $(i = 1, \ldots, m;\ j = 1, \ldots, n)$, den *Elementen* der Matrix $\boldsymbol{A}$. Matrizen werden stets mit großen lateinischen, gelegentlich auch griechischen Buchstaben und ihre Elemente mit den zugehörigen kleinen, unten doppelt indizierten Buchstaben bezeichnet. Für a_{ij} schreiben wir auch $(\boldsymbol{A})_{ij}$. Zur Unterscheidung von Matrizen und deren Elementen werden obere, in Klammern gesetzte Indizes verwendet wie $\boldsymbol{A}^{(0)} = (a_{ij}^{(0)})$, $\boldsymbol{A}^{(1)} = (a_{ij}^{(1)})$ usw., sofern auf die Elemente Bezug genommen wird. Andernfalls kann auch unten indiziert werden, etwa $\boldsymbol{A}_1, \boldsymbol{A}_2$ usw.

Der erste Index i heißt *Zeilenindex*, der zweite Index j *Spaltenindex* des Elementes a_{ij}. Die Menge aller (m,n)-Matrizen wird mit $\mathbf{R}^{m,n}$ bezeichnet. In natürlicher Weise

können $(m,1)$-Matrizen als m-dimensionale *Spaltenvektoren*, $(1,n)$-Matrizen als n-dimensionale *Zeilenvektoren* aufgefaßt werden.

Der Spaltenvektor

$$\begin{pmatrix} a_{1j} \\ \vdots \\ a_{mj} \end{pmatrix} \in \mathbf{R}^{m,1}$$

heißt *j-te Spalte von* $\boldsymbol{A}$, der Zeilenvektor

$$(a_{i1}, \ldots, a_{in}) \in \mathbf{R}^{1,n}$$

i-te Zeile von $\boldsymbol{A}$. Im Fall $m = n$ heißt $\boldsymbol{A}$ *quadratisch* und von der *Ordnung* n. Die Elemente $a_{11}, a_{22}, \ldots, a_{ll}$, $l := \min\{m, n\}$, einer Matrix $\boldsymbol{A}$ werden *Diagonalelemente* von $\boldsymbol{A}$, die zugehörigen Positionen im Rechteckschema *Diagonale* bzw. *Hauptdiagonale* genannt.

Eine Matrix $\boldsymbol{A}$ heißt *Diagonalmatrix*, wenn $a_{ij} = 0$ für $i \neq j$ ist, d. h., wenn alle Nichtdiagonalelemente verschwinden. Wir verzichten dann auch auf die doppelte Indizierung und schreiben einfach $\boldsymbol{A} = \operatorname{diag}(a_1, \ldots, a_l) = \operatorname{diag}(a_i)$, wobei das Format angegeben werden muß bzw. aus dem Zusammenhang ersichtlich ist. Zum Beispiel sind

$$\boldsymbol{A}_1 = \begin{pmatrix} 5 & 0 & 0 \\ 0 & 7 & 0 \\ 0 & 0 & 9 \end{pmatrix}, \quad \boldsymbol{A}_2 = \begin{pmatrix} 5 & 0 & 0 & 0 \\ 0 & 7 & 0 & 0 \\ 0 & 0 & 9 & 0 \end{pmatrix}, \quad \boldsymbol{A}_3 = \begin{pmatrix} 5 & 0 & 0 \\ 0 & 7 & 0 \\ 0 & 0 & 9 \\ 0 & 0 & 0 \end{pmatrix}$$

Diagonalmatrizen diag (5, 7, 9) der Formate (3, 3), (3, 4) und (4, 3).

Eine Matrix $\boldsymbol{A}$ wird *Bandmatrix* der *Bandbreite* $l := p + q + 1$ genannt, wenn $a_{ij} = 0$ für $j < i - p$ und $i + q < j$ mit Zahlen $p, q \geqq 0$ gilt. Dies bedeutet, daß neben der Hauptdiagonalen nur p untere und q obere Nebendiagonalen von 0 verschiedene Elemente besitzen dürfen. Zum Beispiel sind

$$\boldsymbol{A}_1 = \begin{pmatrix} \times & \times & 0 & 0 \\ 0 & \times & \times & 0 \\ 0 & 0 & \times & \times \\ 0 & 0 & 0 & \times \\ 0 & 0 & 0 & 0 \end{pmatrix} \quad \text{und} \quad \boldsymbol{A}_2 = \begin{pmatrix} \times & \times & 0 & 0 & 0 \\ \times & \times & \times & 0 & 0 \\ 0 & \times & \times & \times & 0 \\ 0 & 0 & \times & \times & \times \\ 0 & 0 & 0 & \times & \times \end{pmatrix}.$$

Bandmatrizen der Bandbreite 2 bzw. 3; $\boldsymbol{A}_1$ heißt (obere) *Bidiagonalmatrix*, $\boldsymbol{A}_2$ *Tridiagonalmatrix*. Hierbei kennzeichnet „$\times$“ das *Besetztheitsmuster* von $\boldsymbol{A}$, d. h. diejenigen Positionen, wo von 0 verschiedene Elemente auftreten dürfen.

Eine Matrix $\boldsymbol{A}$ heißt *obere* bzw. *untere Trapezmatrix*, wenn $a_{ij} = 0$ ist für $i > j$ bzw. $i < j$. Eine obere Trapezmatrix mit $m \geqq n$ wird *obere Dreiecksmatrix* genannt,

eine untere Trapezmatrix mit $m \leqq n$ heißt *untere Dreiecksmatrix*. Beispiele sind

$$\boldsymbol{A}_1 = \begin{pmatrix} \times & \times & \times \\ 0 & \times & \times \\ 0 & 0 & \times \\ 0 & 0 & 0 \end{pmatrix}, \qquad \boldsymbol{A}_2 = \begin{pmatrix} \times & \times & \times & \times \\ 0 & \times & \times & \times \\ 0 & 0 & \times & \times \end{pmatrix},$$

$$\boldsymbol{A}_3 = \begin{pmatrix} \times & \times & \times \\ 0 & \times & \times \\ 0 & 0 & \times \end{pmatrix}, \qquad \boldsymbol{A}_4 = \begin{pmatrix} \times & 0 & 0 \\ \times & \times & 0 \\ \times & \times & \times \\ \times & \times & \times \end{pmatrix},$$

wobei $\boldsymbol{A}_2$ eine obere Trapezmatrix ist, $\boldsymbol{A}_1$ und $\boldsymbol{A}_3$ obere Dreiecksmatrizen sind und $\boldsymbol{A}_4$ eine untere Trapezmatrix ist. Schließlich heißt $\boldsymbol{A}$ *obere Hessenbergmatrix*, wenn $a_{ij} = 0$ für $i > j + 1$ gilt; für $m = n = 4$ ergibt sich das Besetztheitsmuster

$$\boldsymbol{A} = \begin{pmatrix} \times & \times & \times & \times \\ \times & \times & \times & \times \\ 0 & \times & \times & \times \\ 0 & 0 & \times & \times \end{pmatrix}.$$

Jede Matrix $\boldsymbol{A} \in \mathbf{R}^{m,n}$ läßt sich als Spaltenvektor $\boldsymbol{a} \in \mathbf{R}^{m \cdot n}$ auffassen, indem die Elemente von $\boldsymbol{A}$ nacheinander spaltenweise als Komponenten des Vektors $\boldsymbol{a}$ angeordnet werden. Dem Element a_{ij} entspricht dann die Komponente a_l von $\boldsymbol{a}$, wobei durch die Beziehung

$$l = (j - 1)\, m + i \qquad (i = 1, \ldots, m;\ j = 1, \ldots, n)$$

eine eineindeutige Zuordnung zwischen (i, j) und l gegeben ist. Bei Verwendung der Programmiersprache FORTRAN wird übrigens eine Matrix in dieser Form als einfachindiziertes Feld abgespeichert, wobei die Zeilenzahl m für die Indexzuordnung bekannt sein muß, vgl. Kapitel 16.

Für $\boldsymbol{A}, \boldsymbol{B} \in \mathbf{R}^{m,n}$ und $\lambda \in \mathbf{R}$ kann $\boldsymbol{A} + \boldsymbol{B}$, $\lambda \boldsymbol{A} \in \mathbf{R}^{m,n}$ definiert werden, indem die zugehörigen „langen" Vektoren denselben Operationen unterworfen werden. Dies führt auf $(\boldsymbol{A} + \boldsymbol{B})_{ij} = (\boldsymbol{A})_{ij} + (\boldsymbol{B})_{ij}$, $(\lambda \boldsymbol{A})_{ij} = (\boldsymbol{A}\lambda)_{ij} = \lambda(\boldsymbol{A})_{ij}$ $(i = 1, \ldots, m,$ $j = 1, \ldots, n)$, d. h., *Summe* und *Produkt mit einer Zahl* sind elementweise zu bilden. Damit wird $\mathbf{R}^{m,n}$ ein linearer Raum der Dimension $m \cdot n$. Die *Gleichheit* $\boldsymbol{A} = \boldsymbol{B}$ bedeutet $(\boldsymbol{A})_{ij} = (\boldsymbol{B})_{ij}$ $(i = 1, \ldots, m;\ j = 1, \ldots, n)$, die *Nullmatrix* $\boldsymbol{O} \in \mathbf{R}^{m,n}$ besteht aus den $m \cdot n$ Elementen Null, und die Regeln (1), (2), (3) gelten sinngemäß.

Jeder Matrix $\boldsymbol{A} \in \mathbf{R}^{m,n}$ kann die *transponierte Matrix* $\boldsymbol{A}^{\mathsf{T}} \in \mathbf{R}^{n,m}$ durch die Festlegung $(\boldsymbol{A}^{\mathsf{T}})_{ji} := (\boldsymbol{A})_{ij}$ $(i = 1, \ldots, m;\ j = 1, \ldots, n)$ zugeordnet werden. Sie entsteht aus $\boldsymbol{A}$ durch Spiegelung an der Diagonalen, im Beispiel

$$\boldsymbol{A} = \begin{pmatrix} 1 & 2 \\ 3 & 4 \\ 5 & 6 \end{pmatrix}, \qquad \boldsymbol{A}^{\mathsf{T}} = \begin{pmatrix} 1 & 3 & 5 \\ 2 & 4 & 6 \end{pmatrix}.$$

Durch Transposition geht der Spaltenvektor $\boldsymbol{x} = (x_i) \in \mathbf{R}^m = \mathbf{R}^{m,1}$ in den m-dimensionalen Zeilenvektor $\boldsymbol{x}^T = (x_1, \ldots, x_m) \in \mathbf{R}^{1,m}$ über, so daß $\boldsymbol{x}$ platzsparend als $\boldsymbol{x} = (x_1, \ldots, x_m)^T \in \mathbf{R}^m$ geschrieben werden kann. Die Transposition genügt den Regeln

$$(\boldsymbol{A}^T)^T = \boldsymbol{A}, \qquad (\boldsymbol{A} + \boldsymbol{B})^T = \boldsymbol{A}^T + \boldsymbol{B}^T, \qquad (\lambda \boldsymbol{A})^T = \lambda(\boldsymbol{A}^T) \tag{4}$$

für alle $\boldsymbol{A}, \boldsymbol{B} \in \mathbf{R}^{m,n}$ und alle $\lambda \in \mathbf{R}$.

Eine quadratische Matrix $\boldsymbol{A}$ wird *symmetrisch* genannt, wenn

$$\boldsymbol{A} = \boldsymbol{A}^T$$

gilt; z. B. ist

$$\boldsymbol{A} = \begin{pmatrix} 1 & 2 & 3 \\ 2 & 4 & 5 \\ 3 & 5 & 6 \end{pmatrix}$$

eine symmetrische Matrix. Wegen (4) sind Summe und skalare Vielfache symmetrischer Matrizen wieder symmetrisch. Die Menge $\mathbf{S}^{n,n}$ aller reellen symmetrischen Matrizen der Ordnung n bildet daher einen Teilraum von $\mathbf{R}^{n,n}$. Jede symmetrische Matrix ist durch die Elemente a_{ij} $(i \leqq j)$ im oberen Dreieck eindeutig festgelegt, so daß $\mathbf{S}^{n,n}$ von der Dimension $n(n+1)/2$ ist.

In vielen Anwendungen ist es zweckmäßig, eine Matrix nicht durch ihre Elemente, sondern durch regelmäßig angeordnete Teilmatrizen niedrigerer Dimension — sog. *Blöcke* — zu charakterisieren. Man spricht dann von Hyper- oder *Blockmatrizen*; Beispiele sind

$$\boldsymbol{A} = \begin{pmatrix} \boldsymbol{A}_{11} & \boldsymbol{A}_{12} & \boldsymbol{A}_{13} \\ \boldsymbol{A}_{21} & \boldsymbol{A}_{22} & \boldsymbol{A}_{23} \end{pmatrix}, \qquad \boldsymbol{B} = \begin{pmatrix} \boldsymbol{B}_{11} & \boldsymbol{B}_{12} & \boldsymbol{B}_{13} \\ \boldsymbol{B}_{21} & \boldsymbol{B}_{22} & \boldsymbol{B}_{23} \end{pmatrix}.$$

Dabei wird vorausgesetzt, daß nebeneinander stehende Blöcke dieselbe Zeilenzahl, übereinander stehende dieselbe Spaltenzahl besitzen. Falls $\boldsymbol{A}$ und $\boldsymbol{B}$ dieselbe Blockstruktur haben, ergibt sich die Summe offensichtlich blockweise zu

$$\boldsymbol{A} + \boldsymbol{B} = \left(\begin{array}{c|c|c} \boldsymbol{A}_{11} + \boldsymbol{B}_{11} & \boldsymbol{A}_{12} + \boldsymbol{B}_{12} & \boldsymbol{A}_{13} + \boldsymbol{B}_{13} \\ \hline \boldsymbol{A}_{21} + \boldsymbol{B}_{21} & \boldsymbol{A}_{22} + \boldsymbol{B}_{22} & \boldsymbol{A}_{23} + \boldsymbol{B}_{23} \end{array}\right).$$

Durch Transposition von $\boldsymbol{A}$ entsteht

$$\boldsymbol{A}^T = \begin{pmatrix} \boldsymbol{A}_{11}^T & \boldsymbol{A}_{21}^T \\ \boldsymbol{A}_{12}^T & \boldsymbol{A}_{22}^T \\ \boldsymbol{A}_{13}^T & \boldsymbol{A}_{23}^T \end{pmatrix}.$$

D. *Multiplikative Matrixoperationen*

Wir betrachten wieder n Vektoren $\boldsymbol{a}^1, \ldots, \boldsymbol{a}^n \in \mathbf{R}^m$, die wir als Spalten der Matrix

$$\boldsymbol{A} = (a_{ij}) = (\boldsymbol{a}^1, \ldots, \boldsymbol{a}^n) \in \mathbf{R}^{m,n}$$

auffassen. Jede Linearkombination $\boldsymbol{y} = x_1\boldsymbol{a}^1 + \cdots + x_n\boldsymbol{a}^n$ mit den im Vektor

$\boldsymbol{x} \in \mathbf{R}^n$ zusammengefaßten Koeffizienten $x_1, \ldots, x_n$ hat dann die Komponenten

$$y_i = \sum_{j=1}^{n} x_j (\boldsymbol{a}^j)_i = \sum_{j=1}^{n} a_{ij} x_j \qquad (i = 1, \ldots, m). \tag{5}$$

Wir nennen den Vektor $\boldsymbol{y} \in \mathbf{R}^m$ mit den durch (5) definierten Komponenten das *Produkt der Matrix* $\boldsymbol{A} \in \mathbf{R}^{m,n}$ *mit dem Vektor* $\boldsymbol{x} \in \mathbf{R}^n$ und schreiben dafür

$$\boldsymbol{y} = \boldsymbol{A}\boldsymbol{x}. \tag{6}$$

Wenn $\boldsymbol{x}$ seinerseits aus einem Vektor $\boldsymbol{w} \in \mathbf{R}^l$ durch Multiplikation mit der Matrix $\boldsymbol{B} \in \mathbf{R}^{n,l}$ gemäß

$$\boldsymbol{x} = \boldsymbol{B}\boldsymbol{w} \tag{7}$$

entstanden ist, erhebt sich die Frage, wie die Komponenten von $\boldsymbol{y} = \boldsymbol{A}\boldsymbol{x} = \boldsymbol{A}(\boldsymbol{B}\boldsymbol{w})$ mit denen von $\boldsymbol{w}$ zusammenhängen. Indem (7) komponentenweise als $x_j = \sum_{k=1}^{l} b_{jk} w_k$ geschrieben und in (5) eingesetzt wird, ergibt sich

$$y_i = \sum_{j=1}^{n} a_{ij} x_j = \sum_{j=1}^{n} a_{ij} \left(\sum_{k=1}^{l} b_{jk} w_k \right) = \sum_{k=1}^{l} \left(\sum_{j=1}^{n} a_{ij} b_{jk} \right) w_k = \sum_{k=1}^{l} c_{ik} w_k$$

mit

$$c_{ik} := \sum_{j=1}^{n} a_{ij} b_{jk} \qquad (i = 1, \ldots, m;\ k = 1, \ldots, l). \tag{8}$$

Fassen wir die c_{ik} als Elemente einer Matrix $\boldsymbol{C} = (c_{ik}) \in \mathbf{R}^{m,l}$ auf, so gilt folglich $\boldsymbol{y} = \boldsymbol{A}(\boldsymbol{B}\boldsymbol{w}) = \boldsymbol{C}\boldsymbol{w}$. Dieses Ergebnis legt es nahe, für zwei Matrizen $\boldsymbol{A} \in \mathbf{R}^{m,n}$, $\boldsymbol{B} \in \mathbf{R}^{n,l}$ das *Produkt*

$$\boldsymbol{C} = \boldsymbol{A}\boldsymbol{B} \tag{9}$$

von $\boldsymbol{A}$ *mit* $\boldsymbol{B}$ als (m, l)-Matrix mit den durch (8) erklärten Elementen zu definieren. Für den Sonderfall eines Spaltenvektors $\boldsymbol{B}$, d. h. für $\boldsymbol{B} = \boldsymbol{x} \in \mathbf{R}^{n,1}$, ist diese Definition mit der des Produktes $\boldsymbol{A}\boldsymbol{x}$ identisch.

Mit etwas Schreibarbeit kann nachgerechnet werden, daß die Matrixmultiplikation den Regeln

$$\boldsymbol{A}(\lambda \boldsymbol{B}) = (\lambda \boldsymbol{A})\,\boldsymbol{B} = \lambda(\boldsymbol{A}\boldsymbol{B}), \tag{10}$$

$$(\boldsymbol{A} + \boldsymbol{B})\,\boldsymbol{C} = \boldsymbol{A}\boldsymbol{C} + \boldsymbol{B}\boldsymbol{C}, \qquad \boldsymbol{A}(\boldsymbol{B} + \boldsymbol{C}) = \boldsymbol{A}\boldsymbol{B} + \boldsymbol{A}\boldsymbol{C}, \tag{11}$$

$$(\boldsymbol{A}\boldsymbol{B})^{\mathsf{T}} = \boldsymbol{B}^{\mathsf{T}}\boldsymbol{A}^{\mathsf{T}} \tag{12}$$

für alle Matrizen $\boldsymbol{A}, \boldsymbol{B}, \boldsymbol{C}$ passenden Formats und alle $\lambda \in \mathbf{R}$ genügt.

Für zwei Matrizen, $\boldsymbol{A}, \boldsymbol{B}$ können beide Produkte $\boldsymbol{A}\boldsymbol{B}$ und $\boldsymbol{B}\boldsymbol{A}$ genau dann gebildet werden, wenn $\boldsymbol{A} \in \mathbf{R}^{m,n}$ und $\boldsymbol{B} \in \mathbf{R}^{n,m}$ ist, aber selbst im Fall $m = n$ gilt i. allg.

$$\boldsymbol{A}\boldsymbol{B} \neq \boldsymbol{B}\boldsymbol{A},$$

wie das Beispiel

$$\boldsymbol{A} = \begin{pmatrix} 1 & 2 \\ 2 & 3 \end{pmatrix}, \qquad \boldsymbol{B} = \begin{pmatrix} 2 & 1 \\ 1 & 4 \end{pmatrix}, \qquad \boldsymbol{AB} = \begin{pmatrix} 4 & 9 \\ 7 & 14 \end{pmatrix} \neq \begin{pmatrix} 4 & 7 \\ 9 & 14 \end{pmatrix} = \boldsymbol{BA}$$

zeigt. Falls $\boldsymbol{AB} = \boldsymbol{BA}$ gilt, heißen $\boldsymbol{A}$ und $\boldsymbol{B}$ *vertauschbar*.

Für jede Matrix $\boldsymbol{A} \in \mathbf{R}^{m,n}$ können die beiden Produkte $\boldsymbol{A}^{\mathsf{T}}\boldsymbol{A} \in \mathbf{R}^{n,n}$ sowie $\boldsymbol{A}\boldsymbol{A}^{\mathsf{T}} \in \mathbf{R}^{m,m}$ stets gebildet werden, und beide sind symmetrisch. Das Produkt von oberen bzw. unteren Dreiecksmatrizen ist wieder eine obere bzw. untere Dreiecksmatrix, insbesondere ist das Produkt von Diagonalmatrizen wieder diagonal. Wie das obige Beispiel zeigt, ist dagegen das Produkt symmetrischer Matrizen i. allg. nicht symmetrisch.

Für $\boldsymbol{A} \in \mathbf{R}^{m,n}$, $\boldsymbol{B} \in \mathbf{R}^{n,l}$ und $\boldsymbol{C} \in \mathbf{R}^{l,k}$ kann $(\boldsymbol{AB})\,\boldsymbol{C}$ gebildet werden und ist vom Format (m, k). Solche Dreierprodukte genügen dem Assoziativgesetz

$$(\boldsymbol{AB})\,\boldsymbol{C} = \boldsymbol{A}(\boldsymbol{BC}) =: \boldsymbol{ABC}, \tag{13}$$

allerdings hat gerade hier die Klammersetzung häufig einen entscheidenden Einfluß auf die Anzahl der zur Berechnung erforderlichen Rechenoperationen und Zwischenspeicher, siehe Ü 1.1.4.

Eine besondere Rolle spielen die multiplikativen Verknüpfungen zweier Vektoren $\boldsymbol{a} \in \mathbf{R}^m$, $\boldsymbol{b} \in \mathbf{R}^n$. Während die Produkte $\boldsymbol{ab}$ und $\boldsymbol{ba}$ im Fall $m, n > 1$ nicht gebildet werden können, ist $\boldsymbol{ab}^{\mathsf{T}}$ definiert und stellt die Matrix

$$\boldsymbol{ab}^{\mathsf{T}} = (a_i b_j) = \begin{pmatrix} a_1 b_1 & \dots & a_1 b_n \\ \vdots & & \vdots \\ a_m b_1 & \dots & a_m b_n \end{pmatrix} \in \mathbf{R}^{m,n} \tag{14}$$

dar, die *dyadisches Produkt* von $\boldsymbol{a}$ mit $\boldsymbol{b}$ genannt wird. Ebenso kann $\boldsymbol{ba}^{\mathsf{T}} \in \mathbf{R}^{n,m}$ gebildet werden und ist gleich $(\boldsymbol{ab}^{\mathsf{T}})^{\mathsf{T}}$.

Das Produkt $\boldsymbol{b}^{\mathsf{T}}\boldsymbol{a}$ ist für Vektoren $\boldsymbol{a}, \boldsymbol{b} \in \mathbf{R}^n$ gleicher Dimension stets erklärt und ergibt

$$\boldsymbol{b}^{\mathsf{T}}\boldsymbol{a} = b_1 a_1 + \dots + b_n a_n \in \mathbf{R}. \tag{15}$$

Es genügt dem Kommutativgesetz $\boldsymbol{a}^{\mathsf{T}}\boldsymbol{b} = \boldsymbol{b}^{\mathsf{T}}\boldsymbol{a}$ und heißt *Skalarprodukt* von $\boldsymbol{a}$ und $\boldsymbol{b}$.

Unter Verwendung des Skalarproduktes läßt sich die folgende einprägsame Regel für die Berechnung der Elemente c_{ik} des Produktes $\boldsymbol{C} = \boldsymbol{AB}$, $\boldsymbol{A} \in \mathbf{R}^{m,n}$, $\boldsymbol{B} \in \mathbf{R}^{n,l}$ angeben: Falls $\boldsymbol{a}^{i\mathsf{T}}$ die Zeilen von $\boldsymbol{A}$ und $\boldsymbol{b}^k$ die Spalten von $\boldsymbol{B}$ gemäß

$$\boldsymbol{A} = \begin{pmatrix} — \boldsymbol{a}^{1\mathsf{T}} — \\ \vdots \\ — \boldsymbol{a}^{i\mathsf{T}} — \\ \vdots \\ — \boldsymbol{a}^{m\mathsf{T}} — \end{pmatrix}, \qquad \boldsymbol{B} = \begin{pmatrix} | & & | & & | \\ \boldsymbol{b}^1 & \dots & \boldsymbol{b}^k & \dots & \boldsymbol{b}^l \\ | & & | & & | \end{pmatrix}$$

bezeichnen, gilt $c_{ik} = \boldsymbol{a}^{i\mathsf{T}}\boldsymbol{b}^k$ $(i = 1, \dots, m;\ j = 1, \dots, l)$, d. h., c_{ik} ist das Skalarprodukt von $\boldsymbol{a}^i$ mit $\boldsymbol{b}^k$, oder kurz: $c_{ik} = (i\text{-te Zeile } \boldsymbol{A})$ mal $(k\text{-te Spalte } \boldsymbol{B})$. In analoger Weise wird bei der Produktbildung von Blockmatrizen vorgegangen. Als Beispiel

betrachten wir die Matrizen

$$\boldsymbol{A} = \left(\begin{array}{c|c|c} \boldsymbol{A}_{11} & \boldsymbol{A}_{12} & \boldsymbol{A}_{13} \\ \hline \boldsymbol{A}_{21} & \boldsymbol{A}_{22} & \boldsymbol{A}_{23} \end{array}\right) \quad \text{und} \quad \boldsymbol{B} = \left(\begin{array}{c|c} \boldsymbol{B}_{11} & \boldsymbol{B}_{12} \\ \hline \boldsymbol{B}_{21} & \boldsymbol{B}_{22} \\ \hline \boldsymbol{B}_{31} & \boldsymbol{B}_{32} \end{array}\right).$$

Das Produkt hat dann die Blockstruktur

$$\boldsymbol{AB} = \left(\begin{array}{c|c} \boldsymbol{A}_{11}\boldsymbol{B}_{11} + \boldsymbol{A}_{12}\boldsymbol{B}_{21} + \boldsymbol{A}_{13}\boldsymbol{B}_{31} & \boldsymbol{A}_{11}\boldsymbol{B}_{12} + \boldsymbol{A}_{12}\boldsymbol{B}_{22} + \boldsymbol{A}_{13}\boldsymbol{B}_{32} \\ \hline \boldsymbol{A}_{21}\boldsymbol{B}_{11} + \boldsymbol{A}_{22}\boldsymbol{B}_{21} + \boldsymbol{A}_{23}\boldsymbol{B}_{31} & \boldsymbol{A}_{21}\boldsymbol{B}_{12} + \boldsymbol{A}_{22}\boldsymbol{B}_{22} + \boldsymbol{A}_{23}\boldsymbol{B}_{32} \end{array}\right).$$

Dabei muß die Blockeinteilung selbstverständlich so sein, daß die Produkte $\boldsymbol{A}_{ij}\boldsymbol{B}_{jk}$ gebildet werden können. Allgemein ist das Produkt zweier Blockmatrizen $\boldsymbol{A} = (\boldsymbol{A}_{ij})$, $\boldsymbol{B} = (\boldsymbol{B}_{jk})$ eine Blockmatrix $\boldsymbol{C} = \boldsymbol{AB} = (\boldsymbol{C}_{ik})$ mit den Blöcken

$$\boldsymbol{C}_{ik} = \sum_{j=1}^{m} \boldsymbol{A}_{ij}\boldsymbol{B}_{jk},$$

wobei m die Anzahl der Blöcke in den „Zeilen" von $\boldsymbol{A}$ wie in den „Spalten" von $\boldsymbol{B}$ bezeichnet. Speziell ergibt sich aus den obigen Regeln im Fall $\boldsymbol{B} = (\boldsymbol{b}^1, \ldots, \boldsymbol{b}^l) \in \mathbf{R}^{n,l}$ die Darstellung des Produktes $\boldsymbol{AB}$ in der Form

$$\boldsymbol{AB} = \begin{pmatrix} | & & | \\ \boldsymbol{Ab}^1 & \ldots & \boldsymbol{Ab}^l \\ | & & | \end{pmatrix}.$$

Wir bemerken abschließend, daß sich die Multiplikation eines Vektors $\boldsymbol{x}$ mit einer Zahl λ den Regeln der Matrixmultiplikation unterordnet, wenn $\boldsymbol{x}\lambda$ statt $\lambda\boldsymbol{x}$ geschrieben wird, da dann die Formate $(n, 1)$ von $\boldsymbol{x}$ und $(1, 1)$ von λ im Sinne der Multiplikation zueinander passen.

E. *Lineare Abbildungen*

Wir betrachten wieder bei gegebener Matrix $\boldsymbol{A} \in \mathbf{R}^{m,n}$ das durch (5), (6) erklärte Produkt

$$\boldsymbol{y} = \boldsymbol{Ax}, \qquad \boldsymbol{x} \in \mathbf{R}^n. \tag{16}$$

Diese Beziehung kann so gedeutet werden, daß mittels der Matrix $\boldsymbol{A}$ jedem $\boldsymbol{x} \in \mathbf{R}^n$ in eindeutiger Weise ein $\boldsymbol{y} = \boldsymbol{Ax} \in \mathbf{R}^m$ zugeordnet wird. Wir sagen auch, daß die *Matrix* $\boldsymbol{A}$ durch (16) eine *Abbildung* von $\mathbf{R}^n$ in $\mathbf{R}^m$ erklärt, und schreiben $\boldsymbol{A}: \mathbf{R}^n \to \mathbf{R}^m$. Der Vektor $\boldsymbol{y} = \boldsymbol{Ax}$ wird *Bild* von $\boldsymbol{x}$ genannt; umgekehrt heißt bei gegebenem $\boldsymbol{y} \in \mathbf{R}^m$ jedes $\boldsymbol{x} \in \mathbf{R}^n$ mit $\boldsymbol{Ax} = \boldsymbol{y}$ ein *Urbild* von $\boldsymbol{y}$. Es gibt Matrizen $\boldsymbol{A}$, so daß nicht jedes $\boldsymbol{y}$ ein Urbild besitzt, z. B. hat im Fall $\boldsymbol{A} = \begin{pmatrix} 1 & 0 & 1 \\ 0 & 1 & 1 \\ 0 & 0 & 0 \end{pmatrix}$ der Vektor $\boldsymbol{y} = \begin{pmatrix} 0 \\ 0 \\ 1 \end{pmatrix}$ kein Urbild. Andererseits kann es zu einem Bild mehrere Urbilder geben; für obige Matrix hat $\boldsymbol{y} = \begin{pmatrix} 1 \\ 0 \\ 0 \end{pmatrix}$ die Urbilder $\boldsymbol{x} = \begin{pmatrix} 1 \\ 0 \\ 0 \end{pmatrix} + \xi \begin{pmatrix} 1 \\ 1 \\ -1 \end{pmatrix}$, $\xi \in \mathbf{R}$ beliebig.

Die durch (16) festgelegte Abbildung $\boldsymbol{A}\colon \mathbf{R}^n \to \mathbf{R}^m$ hat die Eigenschaften

$$\boldsymbol{A}(\boldsymbol{x} + \boldsymbol{x}') = \boldsymbol{A}\boldsymbol{x} + \boldsymbol{A}\boldsymbol{x}', \qquad \boldsymbol{A}(\lambda\boldsymbol{x}) = \lambda(\boldsymbol{A}\boldsymbol{x}) \tag{17}$$

und wird deshalb auch *lineare Abbildung* genannt. Die Menge

$$\mathcal{R}(\boldsymbol{A}) := \{\boldsymbol{y} \in \mathbf{R}^m\colon \boldsymbol{y} = \boldsymbol{A}\boldsymbol{x}, \boldsymbol{x} \in \mathbf{R}^n\} \tag{18}$$

aller möglichen Bilder heißt *Wertebereich* von $\boldsymbol{A}$. Wenn $\boldsymbol{a}^1, \ldots, \boldsymbol{a}^n$ die Spalten von $\boldsymbol{A}$ bezeichnen, ist $\mathcal{R}(\boldsymbol{A}) = \operatorname{span}\{\boldsymbol{a}^1, \ldots, \boldsymbol{a}^n\}$, so daß $\mathcal{R}(\boldsymbol{A})$ auch *Spaltenraum* von $\boldsymbol{A}$ genannt wird und insbesondere einen Teilraum von $\mathbf{R}^m$ darstellt. Unter dem *Rang* von $\boldsymbol{A}$ — in Zeichen: $\operatorname{rang}(\boldsymbol{A})$ — verstehen wir die Dimension von $\mathcal{R}(\boldsymbol{A})$, d. h. die Anzahl der linear unabhängigen Spalten von $\boldsymbol{A}$. Es zeigt sich, daß die Anzahl der linear unabhängigen Spalten gleich der Anzahl der linear unabhängigen Zeilen von $\boldsymbol{A}$ ist, d. h., es gilt $\operatorname{rang}(\boldsymbol{A}) = \operatorname{rang}(\boldsymbol{A}^{\mathsf{T}})$. Hieraus folgt $r = \operatorname{rang}(\boldsymbol{A}) \leqq \min\{m, n\}$. Im Fall $r = \min\{m, n\}$ sagt man, daß $\boldsymbol{A}$ *Vollrang* hat, andernfalls wird $\boldsymbol{A}$ *rangdefizient* genannt.

Wir betrachten jetzt eine Vollrangmatrix $\boldsymbol{A} \in \mathbf{R}^{m,n}$. Im Fall $m \geqq n$ muß dann $r = n$ gelten, d. h., alle n Spalten von $\boldsymbol{A}$ müssen linear unabhängig sein, und $\boldsymbol{A}$ wird *spaltenregulär* genannt. Entsprechend müssen im Fall $m \leqq n$ alle m Zeilen linear unabhängig sein, und $\boldsymbol{A}$ heißt *zeilenregulär*. Genau in diesem Fall ist $\mathcal{R}(\boldsymbol{A}) = \mathbf{R}^m$, d. h., *jedes* $\boldsymbol{y} \in \mathbf{R}^m$ hat *mindestens ein* Urbild $\boldsymbol{x} \in \mathbf{R}^n$, so daß man auch von einer durch $\boldsymbol{A}$ vermittelten *Abbildung* von $\mathbf{R}^n$ *auf* $\mathbf{R}^m$ spricht.

Eine quadratische Matrix $\boldsymbol{A} \in \mathbf{R}^{n,n}$ von vollem Rang $r = n$ ist sowohl spalten- als auch zeilenregulär und wird schlechthin *regulär* genannt; andernfalls heißt sie *singulär*.

Von besonderer Bedeutung sind diejenigen Vektoren, die durch eine Matrix $\boldsymbol{A} \in \mathbf{R}^{m,n}$ in den Nullvektor abgebildet werden. Sie bilden den Nullraum

$$\mathcal{N}(\boldsymbol{A}) := \{\boldsymbol{x} \in \mathbf{R}^n\colon \boldsymbol{A}\boldsymbol{x} = \boldsymbol{o}\} \tag{19}$$

von $\boldsymbol{A}$, der wegen (17) in der Tat ein Teilraum von $\mathbf{R}^n$ ist. Es gilt

$$\dim\big(\mathcal{N}(\boldsymbol{A})\big) + \dim\big(\mathcal{R}(\boldsymbol{A})\big) = \dim\big(\mathcal{N}(\boldsymbol{A})\big) + r = n, \tag{20}$$

insbesondere besteht $\mathcal{N}(\boldsymbol{A})$ nur aus dem Nullvektor genau dann, wenn $\boldsymbol{A}$ spaltenregulär ist. Es gibt dann zu jedem $\boldsymbol{y} \in \mathcal{R}(\boldsymbol{A})$ genau ein Urbild $\boldsymbol{x} \in \mathbf{R}^n$. Ist $\boldsymbol{A}$ zudem quadratisch, also regulär, so existiert folglich zu jedem $\boldsymbol{y} \in \mathbf{R}^n$ genau ein Urbild $\boldsymbol{x} \in \mathbf{R}^n$. Die so definierte Abbildung, die jedem $\boldsymbol{y} \in \mathbf{R}^n$ eindeutig das $\boldsymbol{x} \in \mathbf{R}^n$ mit $\boldsymbol{y} = \boldsymbol{A}\boldsymbol{x}$ zuordnet, heißt die zu $\boldsymbol{A}\colon \mathbf{R}^n \to \mathbf{R}^n$ *inverse Abbildung*. Sie ist ebenfalls linear und kann mit einer durch $\boldsymbol{A} \in \mathbf{R}^{n,n}$ eindeutig festgelegten Matrix $\boldsymbol{A}^{-1} \in \mathbf{R}^{n,n}$ in der Form

$$\boldsymbol{x} = \boldsymbol{A}^{-1}\boldsymbol{y}, \qquad \boldsymbol{y} \in \mathbf{R}^n, \tag{21}$$

geschrieben werden. Die Matrix $\boldsymbol{A}^{-1}$ heißt die zu $\boldsymbol{A}$ *inverse Matrix* und ist durch die äquivalenten Bedingungen

$$\boldsymbol{A}\boldsymbol{A}^{-1} = \boldsymbol{I} \quad \text{und} \quad \boldsymbol{A}^{-1}\boldsymbol{A} = \boldsymbol{I} \tag{22}$$

eindeutig charakterisiert, wobei

$$\boldsymbol{I} := \boldsymbol{I}_n := \begin{pmatrix} 1 & & & \\ & 1 & & 0 \\ & & \ddots & \\ & 0 & & 1 \end{pmatrix} \in \mathbf{R}^{n,n}$$

die *Einheitsmatrix* der Ordnung n bezeichnet. Für diese gilt $\boldsymbol{I}\boldsymbol{x} = \boldsymbol{x}$ für alle $\boldsymbol{x} \in \mathbf{R}^n$, d. h., sie realisiert die *identische Abbildung* von $\mathbf{R}^n$ auf sich. Überdies ist

$$\boldsymbol{I}_m \boldsymbol{A} = \boldsymbol{A}\boldsymbol{I}_n = \boldsymbol{A} \tag{23}$$

für alle $\boldsymbol{A} \in \mathbf{R}^{m,n}$.

Die Matrixinversion genügt den Regeln

$$(\boldsymbol{A}^{-1})^{-1} = \boldsymbol{A}, \qquad (\lambda \boldsymbol{A})^{-1} = \frac{1}{\lambda}\boldsymbol{A}^{-1}, \qquad (\boldsymbol{A}^{\mathsf{T}})^{-1} = (\boldsymbol{A}^{-1})^{\mathsf{T}} =: \boldsymbol{A}^{-\mathsf{T}} \tag{24}$$

für alle regulären $\boldsymbol{A} \in \mathbf{R}^{n,n}$ und alle $\lambda \in \mathbf{R}$ mit $\lambda \neq 0$.

Das Produkt zweier quadratischer Matrizen ist genau dann regulär, wenn jeder Faktor regulär ist. Ist dies der Fall, so gilt

$$(\boldsymbol{A}\boldsymbol{B})^{-1} = \boldsymbol{B}^{-1}\boldsymbol{A}^{-1}. \tag{25}$$

F. *Quadratische Formen*

Es sei jetzt $\boldsymbol{A} \in \mathbf{R}^{n,n}$ eine quadratische Matrix. Dann heißt die Abbildung, die jedem $\boldsymbol{x} \in \mathbf{R}^n$ die reelle Zahl

$$\boldsymbol{x}^{\mathsf{T}}\boldsymbol{A}\boldsymbol{x} = \sum_{i,j=1}^{n} a_{ij}x_i x_j \tag{26}$$

zuordnet, die durch $\boldsymbol{A}$ erzeugte *quadratische Form*. Die Matrix $\boldsymbol{A}$ heißt *positiv semidefinit*, wenn $\boldsymbol{x}^{\mathsf{T}}\boldsymbol{A}\boldsymbol{x} \geqq 0$ für alle $\boldsymbol{x} \in \mathbf{R}^n$ gilt, d. h., wenn die quadratische Form nur nichtnegative Werte annimmt. Gilt sogar $\boldsymbol{x}^{\mathsf{T}}\boldsymbol{A}\boldsymbol{x} > 0$ für alle $\boldsymbol{x} \neq \boldsymbol{o}$, so heißt $\boldsymbol{A}$ *positiv definit*. Wenn die quadratische Form sowohl positive als auch negative Werte annehmen kann, heißt $\boldsymbol{A}$ *indefinit*. Häufig werden die eingeführten Definitheitsbegriffe nur für symmetrische Matrizen erklärt, vgl. dazu Ü 1.1.5.

G. *Determinanten*

Jeder quadratischen Matrix $\boldsymbol{A} \in \mathbf{R}^{n,n}$ kann die als *Determinante* von $\boldsymbol{A}$ — in Zeichen: $\det(\boldsymbol{A})$ — bezeichnete reelle Zahl durch die folgende rekursive Definition zugeordnet werden:

(i) Für $n = 1$ ist $\det(\boldsymbol{A}) = a_{11}$.

(ii) Für $n > 1$ ist

$$\det(\boldsymbol{A}) := \sum_{i=1}^{n} (-1)^{i+j} a_{ij} \det(\boldsymbol{A}^{(i,j)}),$$

wobei $j \in \{1, \ldots, n\}$ ein beliebiger Spaltenindex ist und $\boldsymbol{A}^{(i,j)}$ diejenige quadratische Untermatrix von $\boldsymbol{A}$ bezeichnet, die durch Streichen der i-ten Zeile und j-ten Spalte aus $\boldsymbol{A}$ entsteht.

Die Vorschrift (ii) wird auch *Entwicklungssatz* genannt; man spricht auch von der Entwicklung von $\det(\boldsymbol{A})$ nach den Elementen der j-ten Spalte. Eine analoge Formel gilt für die Entwicklung nach der i-ten Zeile. Man kann zeigen, daß $\det(\boldsymbol{A})$ nicht vom verwendeten Spaltenindex j (bzw. analog vom Zeilenindex i) abhängt, so daß $\det(\boldsymbol{A})$ durch die obige Definition in der Tat eindeutig festgelegt ist.

Für $n = 2$ ergibt sich speziell mit $j = 1$

$$\det \begin{pmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{pmatrix} = (-1)^2 a_{11} \det(a_{22}) + (-1)^3 a_{21} \det(a_{12}) = a_{11}a_{22} - a_{12}a_{21}.$$

Prinzipiell kann $\det(\boldsymbol{A})$ nach der angegebenen Definition für beliebiges n berechnet werden. Aus Aufwandsgründen scheidet dieser Weg für praktische Zwecke allerdings aus, siehe Ü 1.1.7. Aufwandsgünstigere Vorschriften zur Berechnung von $\det(\boldsymbol{A})$ werden in Kapitel 5 beschrieben.

Für die Anwendungen sind die folgenden Eigenschaften von $\det(\boldsymbol{A})$ von Bedeutung:

(D_1) Bei Vertauschung zweier Zeilen bzw. Spalten ändert sich das Vorzeichen der Determinante.

(D_2) Die Determinante ändert sich nicht, wenn ein Vielfaches einer Zeile bzw. Spalte zu einer anderen Zeile bzw. Spalte addiert wird.

(D_3) Bei Multiplikation einer Zeile bzw. Spalte mit einem Faktor $\lambda \in \mathbf{R}$ ändert sich die Determinante um denselben Faktor λ, insbesondere gilt $\det(\lambda \boldsymbol{A}) = \lambda^n \det(\boldsymbol{A})$ für $\boldsymbol{A} \in \mathbf{R}^{n,n}$.

(D_4) $\det(\boldsymbol{A}) = \det(\boldsymbol{A}^T)$.

(D_5) $\det(\boldsymbol{AB}) = \det(\boldsymbol{A}) \det(\boldsymbol{B})$.

(D_6) $\det(\boldsymbol{A}) \neq 0$ genau dann, wenn $\boldsymbol{A}$ regulär ist.

(D_7) $\det(\boldsymbol{R}) = \det(\boldsymbol{R}_{11}) \det(\boldsymbol{R}_{22}) \cdots \det(\boldsymbol{R}_{ss})$
für jede obere Blockdreiecksmatrix

$$\boldsymbol{R} = \begin{pmatrix} \boldsymbol{R}_{11} & \boldsymbol{R}_{12} & \ldots & \boldsymbol{R}_{1s} \\ & \boldsymbol{R}_{22} & \ldots & \boldsymbol{R}_{2s} \\ & 0 & \ddots & \vdots \\ & & & \boldsymbol{R}_{ss} \end{pmatrix}$$

mit quadratischen Diagonalblöcken $\boldsymbol{R}_{11}, \boldsymbol{R}_{22}, \ldots, \boldsymbol{R}_{ss}$; insbesondere ist die Determinante einer Dreiecksmatrix gleich dem Produkt ihrer Diagonalelemente.

H. *Orthogonalität von Vektoren und orthogonale Matrizen*

Zwei Vektoren $\boldsymbol{p}, \boldsymbol{q} \in \mathbf{R}^m$ heißen *orthogonal*, wenn

$$\boldsymbol{p}^\mathsf{T}\boldsymbol{q} = 0$$

gilt. Offensichtlich ist der Nullvektor orthogonal zu jedem Vektor, und für nichtverschwindende Vektoren stimmt der Orthogonalitätsbegriff im Fall $m = 2, 3$ mit dem aus der analytischen Geometrie bekannten überein. Allgemeiner heißt ein *System* $\{\boldsymbol{p}^1, \ldots, \boldsymbol{p}^n\}$ von Vektoren $\boldsymbol{p}^1, \ldots, \boldsymbol{p}^n \in \mathbf{R}^m$ *orthogonal*, wenn

$$\boldsymbol{p}^{i\mathsf{T}}\boldsymbol{p}^j = 0 \qquad (i \neq j;\ i, j = 1, \ldots, n)$$

ist. Gilt zusätzlich $\boldsymbol{p}^{i\mathsf{T}}\boldsymbol{p}^i = 1$ $(i = 1, \ldots, n)$, so heißt das System *orthonormal*. Nichtverschwindende orthogonale, also insbesondere orthonormale Vektoren sind stets linear unabhängig. Ein Beispiel eines orthonormalen Systems bilden die Einheitsvektoren $\{\boldsymbol{e}^1, \ldots, \boldsymbol{e}^m\}$. Eine Matrix $\boldsymbol{Q} \in \mathbf{R}^{m,n}$, $m \geqq n$, heißt *spaltenorthogonal* bzw. *spaltenorthonormal*, wenn die Spalten $\boldsymbol{q}^1, \ldots, \boldsymbol{q}^n$ von $\boldsymbol{Q}$ ein orthogonales bzw. orthonormales System bilden. Dabei ist $\boldsymbol{Q}$ spaltenorthonormal genau dann, wenn

$$\boldsymbol{Q}^\mathsf{T}\boldsymbol{Q} = \boldsymbol{I}_n \tag{27}$$

gilt. Entsprechend heißt $\boldsymbol{P} \in \mathbf{R}^{m,n}$, $m \leqq n$, *zeilenorthogonal* bzw. *zeilenorthonormal*, wenn $\boldsymbol{P}^\mathsf{T}$ spaltenorthogonal bzw. spaltenorthonormal ist. Zeilenorthonormale Matrizen $\boldsymbol{P}$ sind durch die Bedingung

$$\boldsymbol{P}\boldsymbol{P}^\mathsf{T} = \boldsymbol{I}_m \tag{28}$$

charakterisiert.

Eine quadratische spaltenorthonormale Matrix $\boldsymbol{Q} \in \mathbf{R}^{n,n}$ heißt schlechthin *orthogonal*. Sie ist dann auch zeilenorthonormal und umgekehrt, d. h., eine orthogonale Matrix ist durch die gleichwertigen Bedingungen

$$\boldsymbol{Q}^\mathsf{T}\boldsymbol{Q} = \boldsymbol{I} \quad \text{bzw.} \quad \boldsymbol{Q}\boldsymbol{Q}^\mathsf{T} = \boldsymbol{I} \tag{29}$$

charakterisiert. Jede orthogonale Matrix ist wegen (29) regulär, und es gilt

$$\boldsymbol{Q}^{-1} = \boldsymbol{Q}^\mathsf{T}. \tag{30}$$

Aus (29) folgt unter Beachtung der Determinanteneigenschaften (D_5), (D_4) und (D_7) sofort $1 = \det(\boldsymbol{I}) = \det(\boldsymbol{Q}) \det(\boldsymbol{Q}^\mathsf{T}) = [\det(\boldsymbol{Q})]^2$, also

$$|\det(\boldsymbol{Q})| = 1 \tag{31}$$

für jedes orthogonale $\boldsymbol{Q}$.

Das Produkt orthogonaler Matrizen ist wieder orthogonal. Wir bemerken abschließend, daß eine Matrix $\boldsymbol{Q} \in \mathbf{R}^{n,n}$ genau dann orthogonal ist, wenn für alle $\boldsymbol{u}, \boldsymbol{v} \in \mathbf{R}^n$

$$(\boldsymbol{Q}\boldsymbol{u})^\mathsf{T}\,\boldsymbol{Q}\boldsymbol{v} = \boldsymbol{u}^\mathsf{T}\boldsymbol{v} \tag{32}$$

gilt, d. h., wenn die durch $\boldsymbol{Q}$ vermittelte Abbildung $\boldsymbol{y} = \boldsymbol{Q}\boldsymbol{x}$ das Skalarprodukt zweier Vektoren invariant läßt.

I. *Vektor- und Matrixnormen*

Für jedes $\boldsymbol{x} \in \mathbf{R}^n$ ist $\boldsymbol{x}^\mathsf{T}\boldsymbol{x} = \sum_{i=1}^{n} (x_i)^2 \geqq 0$, so daß die Größe

$$\|\boldsymbol{x}\|_2 := \sqrt{\boldsymbol{x}^\mathsf{T}\boldsymbol{x}} = \sqrt{\sum_{i=1}^{n} |x_i|^2} \tag{33}$$

stets definiert ist. Durch (33) wird jedem $\boldsymbol{x} \in \mathbf{R}^n$ die nichtnegative Zahl $\|\boldsymbol{x}\|_2$ zugeordnet, die *Euklidische Norm* von $\boldsymbol{x}$. Sie gibt für $n = 2$ und $n = 3$ die aus der analytischen Geometrie bekannte Länge des zu $\boldsymbol{x}$ gehörenden Ortsvektors wieder, siehe Abb. 1.1.1.

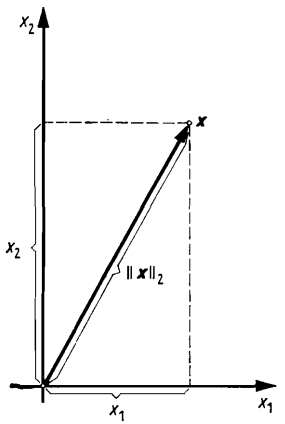


Abb. 1.1.1. Euklidische Norm für $n = 2$

Die Euklidische Norm hat die Eigenschaften

(N_1) $\|\boldsymbol{x}\|_2 \geqq 0$, $\|\boldsymbol{x}\|_2 = 0$ genau für $\boldsymbol{x} = \boldsymbol{o}$,

(N_2) $\|\lambda\boldsymbol{x}\|_2 = |\lambda| \, \|\boldsymbol{x}\|_2$,

(N_3) $\|\boldsymbol{x} + \boldsymbol{y}\|_2 \leqq \|\boldsymbol{x}\|_2 + \|\boldsymbol{y}\|_2$
für alle $\boldsymbol{x}, \boldsymbol{y} \in \mathbf{R}^n$ und alle $\lambda \in \mathbf{R}$. Die Eigenschaft (N_3) wird *Dreiecksungleichung* genannt; sie ist gleichwertig mit

(N_3') $|\|\boldsymbol{x}\|_2 - \|\boldsymbol{y}\|_2| \leqq \|\boldsymbol{x} - \boldsymbol{y}\|_2$ für alle $\boldsymbol{x}, \boldsymbol{y} \in \mathbf{R}^n$.

Ihre Gültigkeit folgt aus der sog. *Schwarzschen Ungleichung*

$$|\boldsymbol{x}^\mathsf{T}\boldsymbol{y}| \leqq \|\boldsymbol{x}\|_2 \, \|\boldsymbol{y}\|_2 \quad \text{für alle } \boldsymbol{x}, \boldsymbol{y} \in \mathbf{R}^n, \tag{34}$$

durch die Skalarprodukt und Euklidische Norm verknüpft sind. Dabei gilt in (34) das Gleichheitszeichen genau dann, wenn $\boldsymbol{x}$ und $\boldsymbol{y}$ linear abhängig sind. Man kann elementar nachrechnen, daß im Fall $n = 2$ und $n = 3$ der cos des durch zwei nichtverschwindende Vektoren $\boldsymbol{x}, \boldsymbol{y}$ gebildeten Winkels $\varphi = \sphericalangle(\boldsymbol{x}, \boldsymbol{y})$ durch

$$\cos\varphi = \frac{\boldsymbol{x}^\mathsf{T}\boldsymbol{y}}{\|\boldsymbol{x}\|_2 \, \|\boldsymbol{y}\|_2} \tag{35}$$

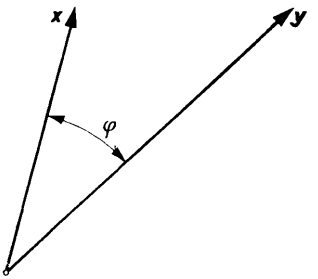


Abb. 1.1.2. Winkel zweier Vektoren

ausgedrückt werden kann, siehe Abb. 1.1.2. Da die rechte Seite von (35) für beliebiges n wegen (34) betragsmäßig höchstens gleich 1 ist, kann durch (35) auch für $n > 3$ zwischen $\boldsymbol{x}$ und $\boldsymbol{y}$ ein *Winkel* φ mit $0 \leqq \varphi \leqq \pi$ definiert werden.

Wir kehren jetzt wieder zur Euklidischen Norm zurück. Obwohl diese ein natürliches Maß für die „Größe" eines Vektors ist, können auch andere Vorschriften angegeben werden, die sich ebenfalls in dieser Weise interpretieren lassen und wie $\|\boldsymbol{x}\|_2$ die Eigenschaften (N_1), (N_2), (N_3) erfüllen. Beispiele dafür sind die

Summennorm

$$\|\boldsymbol{x}\|_1 := \sum_{i=1}^{n} |x_i| \tag{36}$$

und die *Maximumnorm*

$$\|\boldsymbol{x}\|_\infty := \max_{i=1,\ldots,n} |x_i|. \tag{37}$$

Der Unterschied zwischen den angegebenen Normen wird deutlich, wenn man sich die jeweiligen sog. *Einheitskugeln* $\mathcal{S} = \{\boldsymbol{x} \in \mathbf{R}^n : \|\boldsymbol{x}\| \leqq 1\}$ etwa für $n = 2$ veranschaulicht, siehe Abb. 1.1.3.

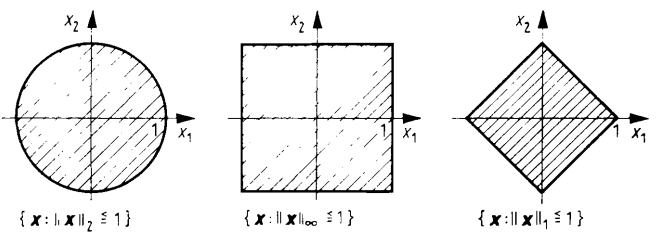


Abb. 1.1.3. Einheitskugeln für verschiedene Vektornormen im Fall $n = 2$

Jede der eingeführten Vektornormen läßt sich durch die übrigen abschätzen, siehe Ü 1.1.8.

Wir betrachten jetzt die durch eine Matrix $\boldsymbol{A} \in \mathbf{R}^{m,n}$ vermittelte Abbildung $\boldsymbol{y} = \boldsymbol{A}\boldsymbol{x}$ und fragen, um welchen Faktor sich $\|\boldsymbol{x}\|_p$ beim Übergang zu $\|\boldsymbol{y}\|_p = \|\boldsymbol{A}\boldsymbol{x}\|_p$ maximal verändert, d. h., wir suchen den maximalen Wert von $\|\boldsymbol{A}\boldsymbol{x}\|_p / \|\boldsymbol{x}\|_p$, wenn $\boldsymbol{x}$ alle nichtverschwindenden Vektoren aus $\mathbf{R}^n$ durchläuft. Dabei bezeichnet $\|\boldsymbol{x}\|_p$

bzw. $\|\boldsymbol{y}\|_p$ eine Norm des Typs $p \in \{1, 2, \infty\}$ im Urbildraum $\mathbf{R}^n$ bzw. im Bildraum $\mathbf{R}^m$. Der gesuchte Wert

$$\|\boldsymbol{A}\|_p := \max \left\{ \frac{\|\boldsymbol{A}\boldsymbol{x}\|_p}{\|\boldsymbol{x}\|_p} : \boldsymbol{x} \neq \boldsymbol{o} \right\} = \max \{\|\boldsymbol{A}\boldsymbol{x}\|_p : \|\boldsymbol{x}\|_p \leqq 1\} \tag{38}$$

existiert für jede Matrix $\boldsymbol{A} \in \mathbf{R}^{m,n}$. Er genügt den Normeigenschaften (N_1), (N_2), (N_3) und ist daher eine Norm auf $\mathbf{R}^{m,n}$, die die der Vektornorm $\|\boldsymbol{x}\|_p$ *zugeordnete Matrixnorm* genannt wird. Aus (38) folgt sofort die Gültigkeit der Abschätzung

$$\|\boldsymbol{A}\boldsymbol{x}\|_p \leqq \|\boldsymbol{A}\|_p \, \|\boldsymbol{x}\|_p \quad \text{für alle } \boldsymbol{x} \in \mathbf{R}^n. \tag{39}$$

Für mindestens ein $\boldsymbol{x} \neq \boldsymbol{o}$ steht in (39) das Gleichheitszeichen, so daß die Zahl $\|\boldsymbol{A}\|_p$ dort nicht durch eine kleinere ersetzt werden kann.

Für die Indizes $p = 1$ und $p = \infty$ lassen sich die den Vektornormen $\|\boldsymbol{x}\|_1$ und $\|\boldsymbol{x}\|_\infty$ zugeordneten Matrixnormen in einfacher Weise aus den Elementen von $\boldsymbol{A}$ gemäß

$$\|\boldsymbol{A}\|_1 = \max_{j=1,\ldots,n} \sum_{i=1}^{m} |a_{ij}| \tag{40}$$

bzw.

$$\|\boldsymbol{A}\|_\infty = \max_{i=1,\ldots,m} \sum_{j=1}^{n} |a_{ij}| \tag{41}$$

berechnen. Diese Formeln legen die Benennungen *Spaltensummennorm* für $\|\boldsymbol{A}\|_1$ und *Zeilensummennorm* für $\|\boldsymbol{A}\|_\infty$ nahe. Dagegen kann $\|\boldsymbol{A}\|_2$ nicht in solch einfacher Weise bestimmt werden. Wir werden im Abschnitt 1.2 zeigen, daß

$$\|\boldsymbol{A}\|_2 = \sqrt{\lambda_{\max}(\boldsymbol{A}^\mathsf{T}\boldsymbol{A})} \tag{42}$$

gilt, wobei $\lambda_{\max}(\boldsymbol{A}^\mathsf{T}\boldsymbol{A})$ den größten Eigenwert der symmetrischen und positiv semidefiniten Matrix $\boldsymbol{A}^\mathsf{T}\boldsymbol{A}$ bezeichnet, der stets nichtnegativ ist. Die Norm $\|\boldsymbol{A}\|_2$ heißt *Spektralnorm* von $\boldsymbol{A}$. Die in Analogie zur Euklidischen Vektornorm aus den Matrixelementen gebildete Norm

$$\|\boldsymbol{A}\|_F := \sqrt{\sum_{i=1}^{m} \sum_{j=1}^{n} |a_{ij}|^2} \tag{43}$$

ist natürlich auch eine Matrixnorm — d. h. eine Norm auf $\mathbf{R}^{m,n}$ — mit den Eigenschaften (N_1), (N_2), (N_3); sie wird *Frobeniusnorm* genannt. Für diese gilt

$$\|\boldsymbol{A}\|_F/\sqrt{r} \leqq \|\boldsymbol{A}\|_2 \leqq \|\boldsymbol{A}\|_F \tag{44}$$

mit $r = \operatorname{rang}(\boldsymbol{A}) \leqq \min\{m, n\}$. Insbesondere kann daher in der Ungleichung (39) für $p = 2$ die Zahl $\|\boldsymbol{A}\|_2$ durch die leicht berechenbare obere Schranke $\|\boldsymbol{A}\|_F$ ersetzt werden: Für alle $\boldsymbol{x} \in \mathbf{R}^n$ gilt

$$\|\boldsymbol{A}\boldsymbol{x}\|_2 \leqq \|\boldsymbol{A}\|_F \, \|\boldsymbol{x}\|_2. \tag{45}$$

Allerdings gibt es in (45) i. allg. kein $\boldsymbol{x} \neq \boldsymbol{o}$, so daß das Gleichheitszeichen steht.

Für den Fall einer Diagonalmatrix $\boldsymbol{D} = \operatorname{diag}(d_1, \ldots, d_l) \in \mathbf{R}^{m,n}$ mit $l = \min\{m, n\}$ gilt

$$\|\boldsymbol{D}\|_1 = \|\boldsymbol{D}\|_2 = \|\boldsymbol{D}\|_\infty = \max_{j=1,\ldots,l} |d_j|, \tag{46}$$

aber

$$\|\boldsymbol{D}\|_F = \sqrt{\sum_{j=1}^{l} (d_j)^2}. \tag{47}$$

Aus der in (32) festgestellten Invarianz des Skalarproduktes folgt eine wichtige Eigenschaft der mit dem Index 2 bzw. F gekennzeichneten Vektor- bzw. Matrixnormen, nämlich deren Invarianz unter orthogonalen Transformationen: Sind $\boldsymbol{Q} \in \mathbf{R}^{n,n}$, $\boldsymbol{P} \in \mathbf{R}^{m,m}$ orthogonale Matrizen, so gilt

$$\|\boldsymbol{Q}\boldsymbol{x}\|_2 = \|\boldsymbol{x}\|_2 \tag{48}$$

für alle $\boldsymbol{x} \in \mathbf{R}^n$ und

$$\|\boldsymbol{P}\boldsymbol{A}\boldsymbol{Q}\|_2 = \|\boldsymbol{A}\|_2, \qquad \|\boldsymbol{P}\boldsymbol{A}\boldsymbol{Q}\|_F = \|\boldsymbol{A}\|_F \tag{49}$$

für alle $\boldsymbol{A} \in \mathbf{R}^{m,n}$. Man beachte, daß die Einheitsmatrix auch orthogonal ist, so daß in (49) einer der Faktoren $\boldsymbol{P}$ oder $\boldsymbol{Q}$ auch fehlen darf. Insbesondere folgt aus (47), (48) und (49)

$$\|\boldsymbol{Q}\|_2 = 1, \qquad \|\boldsymbol{Q}\|_F = \sqrt{n} \tag{50}$$

für jedes orthogonale $\boldsymbol{Q}$.

Alle bisher betrachteten Matrixnormen $\|\boldsymbol{A}\|_p$, $p \in \{1, 2, \infty, F\}$, haben die Eigenschaft

$$|a_{ij}| \leqq \|\boldsymbol{A}\| \qquad (i = 1, \ldots, m;\ j = 1, \ldots, n). \tag{51}$$

Für die Indizes $1, \infty$ und F ist dies offensichtlich, für die Spektralnorm folgt (51) mittels (34) aus $|a_{ij}| = |\boldsymbol{e}^{i\mathsf{T}}(\boldsymbol{A}\boldsymbol{e}^j)| \leqq \|\boldsymbol{e}^i\|_2 \|\boldsymbol{A}\boldsymbol{e}^j\|_2 \leqq \|\boldsymbol{e}^i\|_2 \|\boldsymbol{A}\|_2 \|\boldsymbol{e}^j\|_2 = \|\boldsymbol{A}\|_2$. Überdies gilt für die erwähnten Normen stets

$$\|\boldsymbol{A}\boldsymbol{B}\| \leqq \|\boldsymbol{A}\| \|\boldsymbol{B}\|, \tag{52}$$

sofern das Produkt $\boldsymbol{A}\boldsymbol{B}$ gebildet werden kann. Im Fall $\|\boldsymbol{A}\|_p$, $p \in \{1, 2, \infty\}$, ist dies eine Folge der Eigenschaft (39), denn für jedes $\boldsymbol{x}$ gilt

$$\|(\boldsymbol{A}\boldsymbol{B})\,\boldsymbol{x}\|_p = \|\boldsymbol{A}(\boldsymbol{B}\boldsymbol{x})\|_p \leqq \|\boldsymbol{A}\|_p \|\boldsymbol{B}\boldsymbol{x}\|_p \leqq \|\boldsymbol{A}\|_p \|\boldsymbol{B}\|_p \|\boldsymbol{x}\|_p,$$

so daß $\|\boldsymbol{A}\|_p \|\boldsymbol{B}\|_p$ in der Tat eine obere Schranke für $\|\boldsymbol{A}\boldsymbol{B}\|_p$ ist. Für die Frobeniusnorm ergibt sich (52) unter Beachtung von (44) aus der Ungleichung

$$\|\boldsymbol{A}\boldsymbol{B}\|_F \leqq \|\boldsymbol{A}\|_2 \|\boldsymbol{B}\|_F. \tag{53}$$

Zum Nachweis von (53) setzen wir $\boldsymbol{C} := \boldsymbol{A}\boldsymbol{B} = (\boldsymbol{c}^1, \ldots, \boldsymbol{c}^l)$, $\boldsymbol{B} = (\boldsymbol{b}^1, \ldots, \boldsymbol{b}^l)$. Dann gilt $\boldsymbol{c}^j = \boldsymbol{A}\boldsymbol{b}^j$ $(j = 1, \ldots, l)$, wobei $\boldsymbol{c}^j$ bzw. $\boldsymbol{b}^j$ die Spalten von $\boldsymbol{C}$ bzw. $\boldsymbol{B}$ bezeichnen. Wegen (39) folgt hieraus $\|\boldsymbol{c}^j\|_2 \leqq \|\boldsymbol{A}\|_2 \|\boldsymbol{b}^j\|_2$, also

$$\|\boldsymbol{C}\|_F^2 = \sum_{j=1}^{l} (\|\boldsymbol{c}^j\|_2)^2 \leqq \|\boldsymbol{A}\|_2^2 \sum_{j=1}^{l} (\|\boldsymbol{b}^j\|_2)^2 = \|\boldsymbol{A}\|_2^2 \|\boldsymbol{B}\|_F^2$$

wie behauptet. Die Eigenschaft (52) wird *Multiplikativität* der Matrixnorm $\|\boldsymbol{A}\|$ genannt.

J. *Ordnungen und Absolutbeträge*

Bei der Analyse von Algorithmen der linearen Algebra treten häufig elementweise Abschätzungen von Vektoren und Matrizen auf. Um solche Abschätzungen einfach beschreiben zu können, führen wir im $\mathbf{R}^m$ eine *natürliche Halbordnung* genannte Ordnungsrelation „$\leqq$" durch die Festlegung

$$\boldsymbol{x} \leqq \boldsymbol{y}, \qquad \text{wenn } x_i \leqq y_i \qquad (i = 1, \ldots, m) \tag{54}$$

für $\boldsymbol{x}, \boldsymbol{y} \in \mathbf{R}^m$ ein. Analog wird für $\boldsymbol{A}, \boldsymbol{B} \in \mathbf{R}^{m,n}$

$$\boldsymbol{A} \leqq \boldsymbol{B}, \qquad \text{wenn } a_{ij} \leqq b_{ij} \qquad (i = 1, \ldots, m;\ j = 1, \ldots, n) \tag{55}$$

gesetzt. Im Gegensatz zur Situation bei reellen Zahlen gibt es Vektoren $\boldsymbol{x}, \boldsymbol{y}$, für die weder $\boldsymbol{x} \geqq \boldsymbol{y}$ noch $\boldsymbol{y} \geqq \boldsymbol{x}$ gilt, so daß die Benennung Halbordnung berechtigt ist. In entsprechender Weise wird für $\boldsymbol{x} \in \mathbf{R}^m$ ein *Betragsvektor* $|\boldsymbol{x}| \in \mathbf{R}^m$ als Vektor mit den Komponenten

$$(|\boldsymbol{x}|)_i = |x_i| \qquad (i = 1, \ldots, m) \tag{56}$$

und für $\boldsymbol{A} \in \mathbf{R}^{m,n}$ eine *Betragsmatrix* $|\boldsymbol{A}| \in \mathbf{R}^{m,n}$ als Matrix mit den Elementen

$$(|\boldsymbol{A}|)_{ij} = |a_{ij}| \qquad (i = 1, \ldots, m;\ j = 1, \ldots, n) \tag{57}$$

eingeführt. Ordnungsbeziehungen und Betragsbildung genügen dabei denselben Beziehungen wie im Bereich der reellen Zahlen, insbesondere dürfen Ungleichungen addiert und mit nichtnegativen Faktoren multipliziert werden, und die Beträge erfüllen die Beziehungen

$$|\boldsymbol{x}| = 0 \text{ genau für } \boldsymbol{x} = \boldsymbol{o}, \qquad |\lambda \boldsymbol{x}| = |\lambda|\,|\boldsymbol{x}| \quad \text{sowie} \quad |\boldsymbol{x} + \boldsymbol{y}| \leqq |\boldsymbol{x}| + |\boldsymbol{y}| \tag{58}$$

für alle $\boldsymbol{x}, \boldsymbol{y} \in \mathbf{R}^m$ und $\lambda \in \mathbf{R}$. Dieselben Beziehungen gelten auch, wenn $\boldsymbol{x}, \boldsymbol{y}$ durch Matrizen $\boldsymbol{A}, \boldsymbol{B} \in \mathbf{R}^{m,n}$ ersetzt werden. Falls das Produkt $\boldsymbol{AB}$ gebildet werden kann, ist überdies

$$|\boldsymbol{AB}| \leqq |\boldsymbol{A}|\,|\boldsymbol{B}|, \tag{59}$$

insbesondere gilt

$$|\boldsymbol{Ax}| \leqq |\boldsymbol{A}|\,|\boldsymbol{x}| \tag{60}$$

für alle $\boldsymbol{A} \in \mathbf{R}^{m,n}$ und $\boldsymbol{x} \in \mathbf{R}^n$.

Die eingeführten Ordnungsbeziehungen und Beträge sind mit den Normen in dem Sinne verträglich, daß

$$\| |\boldsymbol{x}| \| = \|\boldsymbol{x}\| \tag{61}$$

sowie

$$\text{aus } |\boldsymbol{x}| \leqq |\boldsymbol{y}| \text{ folgt } \|\boldsymbol{x}\| \leqq \|\boldsymbol{y}\| \tag{62}$$

gilt, wobei $\|\boldsymbol{x}\|$ eine der Vektornormen $\|\boldsymbol{x}\|_p$, $p \in \{1, 2, \infty\}$, ist.

Eine Norm mit der Eigenschaft (61) bzw. (62) heißt *absolut* bzw. *monoton*. Es ist sofort zu sehen, daß die analogen Eigenschaften

$$\| |\boldsymbol{A}| \| = \|\boldsymbol{A}\| \tag{63}$$

bzw.

$$\text{aus } |\boldsymbol{A}| \leqq |\boldsymbol{B}| \text{ folgt } \|\boldsymbol{A}\| \leqq \|\boldsymbol{B}\| \tag{64}$$

auch für die Matrixnormen $\|\boldsymbol{A}\|_1$, $\|\boldsymbol{A}\|_\infty$ und $\|\boldsymbol{A}\|_F$ gelten, während $\|\boldsymbol{A}\|_2$ weder absolut noch monoton ist, siehe Ü 1.1.11. Es gilt jedoch

$$\|\boldsymbol{A}\|_2 \leqq \| |\boldsymbol{A}| \|_2 \tag{65}$$

für jedes $\boldsymbol{A} \in \mathbf{R}^{m,n}$ sowie

$$\text{aus } |\boldsymbol{A}| \leqq |\boldsymbol{B}| \text{ folgt } \| |\boldsymbol{A}| \|_2 \leqq \| |\boldsymbol{B}| \|_2 . \tag{66}$$

Übungsaufgaben

Ü 1.1.1. Für $\boldsymbol{A} \in \mathbf{R}^{m,n}$, $\boldsymbol{x} \in \mathbf{R}^n$ kann $\boldsymbol{y} = \boldsymbol{A}\boldsymbol{x} \in \mathbf{R}^m$ wie folgt berechnet werden:

V_1: **for** $i := 1(1)m$ **do**
$\quad y_i := 0$
$\quad$ **for** $j := 1(1)n$ **do** $y_i := y_i + a_{ij} * x_j$,

oder

V_2: **for** $i := 1(1)m$ **do** $y_i := 0$
$\quad$ **for** $j := 1(1)n$ **do**
$\quad\quad$ **for** $i := 1(1)m$ **do** $y_i := y_i + a_{ij} * x_j$.

Man mache sich die Unterschiede zwischen beiden Varianten klar und überlege sich, daß V_2 z. B. dann vorzuziehen ist, wenn $\boldsymbol{A}$ spaltenweise auf einem Magnetband abgespeichert ist.

Ü 1.1.2. Es sei $\boldsymbol{S} \in \mathbf{R}^{m,n}$ eine reguläre Matrix und $X \subset \mathbf{R}^n$ ein Teilraum von $\mathbf{R}^n$. Man beweise, daß dann $\boldsymbol{S}X := \{\boldsymbol{y} = \boldsymbol{S}\boldsymbol{x} : \boldsymbol{x} \in X\}$ auch ein Teilraum von $\mathbf{R}^n$ ist und daß $\dim(\boldsymbol{S}X) = \dim(X)$ gilt. Außerdem ist $\{\boldsymbol{a}^1, \ldots, \boldsymbol{a}^r\}$ eine Basis von X genau dann, wenn $\{\boldsymbol{S}\boldsymbol{a}^1, \ldots, \boldsymbol{S}\boldsymbol{a}^r\}$ eine Basis von $\boldsymbol{S}X$ ist.

Ü 1.1.3. Es sei $\boldsymbol{A} := \boldsymbol{a}\boldsymbol{b}^\mathsf{T}$ das dyadische Produkt der Vektoren $\boldsymbol{a}, \boldsymbol{b} \in \mathbf{R}^n$, $\boldsymbol{a} \neq \boldsymbol{o}$, $\boldsymbol{b} \neq \boldsymbol{o}$. Man zeige

(i) $\mathcal{R}(\boldsymbol{A}) = \operatorname{span}\{\boldsymbol{a}\}$, $\operatorname{rang}(\boldsymbol{A}) = 1$ und $\mathcal{N}(\boldsymbol{A}) = \{\boldsymbol{x} \in \mathbf{R}^n : \boldsymbol{b}^\mathsf{T}\boldsymbol{x} = \boldsymbol{o}\}$, $\dim \mathcal{N}(\boldsymbol{A}) = n - 1$.

(ii) $\boldsymbol{A} = \boldsymbol{A}^\mathsf{T}$ genau dann, wenn $\boldsymbol{a}$ und $\boldsymbol{b}$ linear abhängig sind, d. h., wenn $\boldsymbol{a} = \lambda\boldsymbol{b}$ ist mit einer Zahl $\lambda \neq 0$.

(iii) $\|\boldsymbol{A}\|_2 = \|\boldsymbol{A}\|_F = \|\boldsymbol{a}\|_2 \|\boldsymbol{b}\|_2$, $\quad \|\boldsymbol{A}\|_\infty = \|\boldsymbol{a}\|_\infty \|\boldsymbol{b}\|_1$, $\quad \|\boldsymbol{A}\|_1 = \|\boldsymbol{a}\|_1 \|\boldsymbol{b}\|_\infty$.

Ü 1.1.4. Es sei $\boldsymbol{A} = \boldsymbol{a}\boldsymbol{b}^\mathsf{T}$ wie in Ü 1.1.3 erklärt. Dann kann $\boldsymbol{y} = \boldsymbol{A}\boldsymbol{x} = \boldsymbol{a}\boldsymbol{b}^\mathsf{T}\boldsymbol{x}$ nach den beiden Varianten $\boldsymbol{y} = (\boldsymbol{a}\boldsymbol{b}^\mathsf{T})\,\boldsymbol{x}$ bzw. $\boldsymbol{y} = \boldsymbol{a}(\boldsymbol{b}^\mathsf{T}\boldsymbol{x})$ berechnet werden. Man vergleiche beide Vorschriften hinsichtlich der Anzahl der erforderlichen Rechenoperationen und des — auch für Zwischenergebnisse — benötigten Speicherplatzes.

Ü 1.1.5. Man beweise: (i) Jede Matrix $\boldsymbol{A} \in \mathbf{R}^{n,n}$ kann in der Form $\boldsymbol{A} = \boldsymbol{B} + \boldsymbol{C}$ als Summe der symmetrischen Matrix $\boldsymbol{B} = (\boldsymbol{A} + \boldsymbol{A}^\mathsf{T})/2$ und der schiefsymmetrischen Matrix $\boldsymbol{C} = (\boldsymbol{A} - \boldsymbol{A}^\mathsf{T})/2$ dargestellt werden. Dabei heißt $\boldsymbol{C} \in \mathbf{R}^{n,n}$ *schiefsymmetrisch*, wenn $\boldsymbol{C} = -\boldsymbol{C}^\mathsf{T}$ gilt. Die Matrix $\boldsymbol{B}$ wird *symmetrischer Anteil* von $\boldsymbol{A}$ genannt.

(ii) Die quadratische Form einer schiefsymmetrischen Matrix $\boldsymbol{C}$ verschwindet identisch, d. h., es gilt $\boldsymbol{x}^\mathsf{T}\boldsymbol{C}\boldsymbol{x} = 0$ für alle $\boldsymbol{x} \in \mathbf{R}^n$. Hieraus folgt $\boldsymbol{x}^\mathsf{T}\boldsymbol{A}\boldsymbol{x} = \boldsymbol{x}^\mathsf{T}\boldsymbol{B}\boldsymbol{x}$ für alle $\boldsymbol{x} \in \mathbf{R}^n$, d. h., die quadratische Form einer beliebigen Matrix wird allein durch deren symmetrischen Anteil bestimmt.

Ü 1.1.6. Es sei $A = (a_{ij}) \in \mathbf{R}^{n,n}$ eine symmetrische und positiv definite Matrix. Man beweise:
(i) $a_{ii} > 0 \qquad (i = 1, \ldots, n)$,
(ii) $|a_{ij}| < (a_{ii} + a_{jj})/2 \qquad (i, j = 1, \ldots, n;\ i \neq j)$,
(iii) $|a_{ij}| < \sqrt{a_{ii} a_{jj}} \qquad (i, j = 1, \ldots, n;\ i \neq j)$.
Hinweis: Man betrachte $x^T A x$ für $x = \xi e^i + \eta e^j$, wobei $\xi = 1$, $\eta = 0$ ist für (i), $\xi = 1$, $\eta = \pm 1$ für (ii) und ξ, η beliebig für (iii).

Ü 1.1.7. Es ist zu zeigen, daß die Berechnung von $\det(A)$ für $A \in \mathbf{R}^{n,n}$ nach dem Entwicklungssatz $n!$ Multiplikationen und $(n-1)!$ Additionen erfordert.

Ü 1.1.8. Man beweise, daß für die Vektornormen $\|x\|_p$, $p \in \{1, 2, \infty\}$, die Ungleichungen

$$\|x\|_\infty \leqq \|x\|_1 \leqq n \|x\|_\infty, \qquad \|x\|_\infty \leqq \|x\|_2 \leqq \sqrt{n} \|x\|_\infty, \qquad \|x\|_2 \leqq \|x\|_1 \leqq \sqrt{n} \|x\|_2$$

für jedes $x \in \mathbf{R}^n$ gelten, und zeige, daß die hier auftretenden Konstanten nicht verbessert werden können.

Ü 1.1.9. Man zeige, daß für jedes $x \in \mathbf{R}^n$

$$\|x\|_2^2 \leqq \|x\|_1 \|x\|_\infty \leqq 0.5\left(\sqrt{n} + 1\right) \|x\|_2^2$$

gilt.

Ü 1.1.10. Man beweise: Wenn die Matrix $A \in \mathbf{R}^{m,n}$ die Blockstruktur

$$A = \left(\begin{array}{c|c} A_{11} & A_{12} \\ \hline A_{21} & A_{22} \end{array}\right)$$

besitzt, gilt

$$\|A_{ij}\|_p \leqq \|A\|_p \qquad (i, j = 1, 2)$$

für jede der Matrixnormen $p \in \{1, 2, \infty, F\}$.

Ü 1.1.11. Am Beispiel der Matrizen

$$A = \begin{pmatrix} 0.6 & 0.6 \\ 0.6 & 0.6 \end{pmatrix}, \qquad B = \frac{1}{\sqrt{2}} \begin{pmatrix} 1 & 1 \\ -1 & 1 \end{pmatrix}$$

ist zu zeigen, daß $\|A\|_2$ weder eine absolute noch eine monotone Norm ist.
Hinweis: Man beachte, daß B orthogonal ist, und stelle A in der Form $A = 0.6 e e^T$ mit $e = (1, 1)^T$ dar. Die Norm $\|A\|_2$ kann dann nach Ü 1.1.3 berechnet werden. Analog ist $\||B|\|_2$ bestimmbar.

Ü 1.1.12. Man beweise, daß für jede orthogonale obere Dreiecksmatrix A

$$|A| = I$$

gilt, d. h., A ist diagonal mit Diagonalelementen ± 1.

1.2. Eigenwerte, Singulärwerte

A. *Eigenwerte und charakteristisches Polynom*

Eine Zahl $\lambda \in \mathbf{R}$ wird *Eigenwert* der quadratischen Matrix $A \in \mathbf{R}^{n,n}$ genannt, wenn es einen Vektor $x \in \mathbf{R}^n$, $x \neq o$, gibt, für den

$$Ax = \lambda x \tag{1}$$

gilt. Jeder derartige Vektor $\boldsymbol{x}$ heißt ein zu λ gehörender *Eigenvektor*; wir nennen $\{\lambda, \boldsymbol{x}\}$ auch ein *Eigenpaar* der Matrix $\boldsymbol{A}$. Die Eigenvektoren sind also diejenigen Vektoren, die durch $\boldsymbol{A}$ in ein Vielfaches von sich selbst abgebildet werden, und der Proportionalitätsfaktor stellt gerade den Eigenwert dar. Die Bedingung $\boldsymbol{x} \neq \boldsymbol{o}$ ist dabei erforderlich, denn für $\boldsymbol{x} = \boldsymbol{o}$ ist (1) trivialerweise für beliebiges λ erfüllt. Mit $\boldsymbol{x}$ ist auch jedes nichtverschwindende Vielfache Eigenvektor von $\boldsymbol{A}$.

Die Gleichung (1) läßt sich äquivalent in der Form

$$(\boldsymbol{A} - \lambda \boldsymbol{I})\, \boldsymbol{x} = \boldsymbol{o} \tag{2}$$

schreiben, d. h. als homogenes Gleichungssystem $\boldsymbol{M}(\lambda)\, \boldsymbol{x} = \boldsymbol{o}$ mit der von λ abhängenden Koeffizientenmatrix $\boldsymbol{M}(\lambda) := \boldsymbol{A} - \lambda \boldsymbol{I}$. Die Eigenwerte von $\boldsymbol{A}$ sind diejenigen λ, für die (2) von $\boldsymbol{o}$ verschiedene Lösungen besitzt, für die also $\boldsymbol{A} - \lambda \boldsymbol{I}$ singulär ist. Unter Beachtung der Determinanteneigenschaft (D_6) aus Abschnitt 1.1.G ergibt sich hieraus

$$\det (\boldsymbol{A} - \lambda \boldsymbol{I}) = 0 \tag{3}$$

als notwendige und hinreichende Bedingung dafür, daß λ ein Eigenwert von $\boldsymbol{A}$ ist. Für $n = 1$ bedeutet dies

$$\det (a_{11} - \lambda) = (-\lambda) + a_{11} = 0,$$

für $n = 2$ geht (3) in

$$\det \begin{pmatrix} a_{11} - \lambda & a_{12} \\ a_{21} & a_{22} - \lambda \end{pmatrix} = (-\lambda)^2 - (a_{11} + a_{22})\, \lambda + a_{11} a_{22} - a_{12} a_{21} = 0$$

über, und allgemein erhält man

$$\det \begin{pmatrix} a_{11} - \lambda & a_{12} & \dots & a_{1n} \\ a_{21} & a_{22} - \lambda & \dots & a_{2n} \\ \vdots & \vdots & & \vdots \\ a_{n1} & a_{n2} & \dots & a_{nn} - \lambda \end{pmatrix}$$
$$= (-1)^n \lambda^n + c_{n-1} \lambda^{n-1} + \dots + c_1 \lambda + c_0 = 0$$

mit gewissen reellen Koeffizienten $c_0, \dots, c_{n-1}$, die Produktsummen der Elemente von $\boldsymbol{A}$ sind. Der Ausdruck $\det (\boldsymbol{A} - \lambda \boldsymbol{I}) =: p(\lambda)$ stellt ein Polynom vom Grad n mit dem Führungskoeffizienten $(-1)^n$ dar und wird *charakteristisches Polynom* der Matrix $\boldsymbol{A}$ genannt. Die Eigenwerte von $\boldsymbol{A}$ im Sinne der obigen Definition sind also genau die reellen Nullstellen des charakteristischen Polynoms.

Obwohl das charakteristische Polynom einer reellen Matrix nur reelle Koeffizienten hat, kann es auch komplexe Nullstellen besitzen, wie das Beispiel

$$\boldsymbol{A} = \begin{pmatrix} 2 & 1 \\ -2 & 4 \end{pmatrix}, \qquad \det (\boldsymbol{A} - \lambda \boldsymbol{I}) = \lambda^2 - 6\lambda + 10 = 0$$

mit den Nullstellen $\lambda_{1,2} = 3 \pm \sqrt{-1} = 3 \pm \mathrm{i}$ zeigt. Es ist daher angebracht, auch komplexe Eigenwerte $\lambda = \alpha + \mathrm{i}\beta$ zuzulassen. Für solche hat (2) die Gestalt

$$\boldsymbol{M}(\lambda)\, \boldsymbol{x} := (\boldsymbol{A} - \lambda \boldsymbol{I})\, \boldsymbol{x} = [(\boldsymbol{A} - \alpha \boldsymbol{I}) - \mathrm{i}\beta \boldsymbol{I}]\, \boldsymbol{x} = \boldsymbol{o},$$

d. h., die Koeffizientenmatrix $M(\lambda)$ wird ebenfalls komplex, und es müssen dann auch komplexe Eigenvektoren $x = u + iv$ als Lösungen von (2) betrachtet werden. Wir erinnern an dieser Stelle daran, daß jedes λ aus der *Menge* $\mathbf{C}$ *der komplexen Zahlen* die Gestalt $\lambda = \alpha + i\beta$ mit $\alpha, \beta \in \mathbf{R}$ hat. Dabei ist α der *Realteil*, β der *Imaginärteil* von λ, und i bezeichnet die imaginäre Einheit, für die $i^2 = -1$ gilt. Die *Addition* bzw. *Subtraktion* komplexer Zahlen geschieht durch Addition bzw. Subtraktion der Real- und Imaginärteile, während die *Multiplikation* durch

$$\lambda_1\lambda_2 = (\alpha_1 + i\beta_1)(\alpha_2 + i\beta_2) = (\alpha_1\alpha_2 - \beta_1\beta_2) + i(\alpha_1\beta_2 + \alpha_2\beta_1) \tag{4}$$

erklärt ist. Die zu $\lambda = \alpha + i\beta$ *konjugiert komplexe Zahl* ist $\bar{\lambda} = \alpha - i\beta$, und $|\lambda| = \sqrt{\lambda\bar{\lambda}} = \sqrt{\alpha^2 + \beta^2} \geqq 0$ ist der *absolute Betrag* von λ bzw. $\bar{\lambda}$. Im Fall $\lambda_2 \neq 0$ wird schließlich die komplexe *Division* gemäß

$$\frac{\lambda_1}{\lambda_2} = \frac{\lambda_1\bar{\lambda}_2}{\lambda_2\bar{\lambda}_2} = \frac{\alpha_1\alpha_2 + \beta_1\beta_2}{\alpha_2^2 + \beta_2^2} + i\,\frac{-\alpha_1\beta_2 + \alpha_2\beta_1}{\alpha_2^2 + \beta_2^2} \tag{5}$$

eingeführt. Analog zum Vorgehen im Abschnitt 1.1 wird dann der *m-dimensionale komplexe Vektorraum* $\mathbf{C}^m$ als Menge aller m-Tupel $x = u + iv$ mit $u, v \in \mathbf{R}^m$ definiert, in dem eine Addition und eine Multiplikation mit einem komplexen Skalar $\lambda \in \mathbf{C}$ komponentenweise erklärt ist. Entsprechend ist $\mathbf{C}^{m,n}$ die *Menge aller komplexen* (m,n)*-Matrizen* $A = U + iV$ mit $U, V \in \mathbf{R}^{m,n}$, wobei wieder alle Operationen im komplexen Sinne zu verstehen sind. Für $A \in \mathbf{C}^{m,n}$ ist $A^H \in \mathbf{C}^{n,m}$ diejenige Matrix, die durch Transposition und Übergang zum Konjugiertkomplexen aus A entsteht, d. h. $A^H := \bar{A}^T$. Das *Skalarprodukt* von $x, y \in \mathbf{C}^m$ wird als komplexe Zahl $y^Hx = \sum_j \bar{y}_j x_j$ eingeführt. Dann gilt $y^Hx = \overline{x^Hy}$, d. h., das Skalarprodukt ist nicht mehr kommutativ. Jedoch ist $x^Hx = \sum_j |x_j|^2$ stets reell und nichtnegativ, so daß $\|x\|_2 = \sqrt{x^Hx}$ für alle $x \in \mathbf{C}^m$ erklärt ist und die Normeigenschaften besitzt, wobei (N_2) als $\|\lambda x\| = |\lambda| \cdot \|x\|$ für alle $x \in \mathbf{C}^m$ und alle $\lambda \in \mathbf{C}$ mit dem oben eingeführten Absolutbetrag $|\lambda|$ zu verstehen ist. Indem die Absolutbeträge überall durch ihre komplexe Version und der obere Index „T" durch „H" ersetzt werden, übertragen sich die Begriffe und Aussagen des Abschnittes 1.1 sinngemäß auf den komplexen Fall.

Nach diesen Vorbereitungen können wir nun auch eine komplexe Matrix $A \in \mathbf{C}^{n,n}$ betrachten und in Erweiterung der eingangs gegebenen Definitionen komplexe Eigenwerte und Eigenvektoren zulassen. *Die Eigenwerte von* A *sind dann genau die reellen und komplexen Nullstellen des charakteristischen Polynoms.* Da die Nullstellen eines Polynoms stetige Funktionen der Koeffizienten, die Koeffizienten des charakteristischen Polynoms wieder stetige Funktionen der Matrixelemente sind, ergibt sich hieraus: *Die Eigenwerte einer Matrix sind stetige Funktionen der Matrixelemente. Für reelle Matrizen* $A \in \mathbf{R}^{n,n} \subset \mathbf{C}^{n,n}$ *treten komplexe Eigenwerte stets als Paare konjugiert komplexer Eigenwerte auf*, und die Eigenvektoren können entsprechend gewählt werden, siehe Ü 1.2.2.

Ein Eigenwert $\lambda_j \in \mathbf{C}$ von A kann u. U. mehrfache Nullstelle des charakteristischen Polynoms $p(\lambda)$ sein. Wir sagen, daß λ_j die *algebraische Vielfachheit* $\alpha_j \geqq 1$

besitzt, wenn λ_j eine α_j-fache Nullstelle von $p(\lambda)$ ist, d. h., wenn

$$p(\lambda_j) = p'(\lambda_j) = \cdots = p^{(\alpha_j - 1)}(\lambda_j) = 0, \qquad p^{(\alpha_j)}(\lambda_j) \neq 0$$

gilt. *Jede Matrix* $\boldsymbol{A} \in \mathbf{C}^{n,n}$ *hat* also *genau n entsprechend ihren algebraischen Vielfachheiten gezählte Eigenwerte.*

Die Anzahl γ_j der zu einem Eigenwert λ_j gehörenden linear unabhängigen Eigenvektoren wird *geometrische Vielfachheit* von λ_j genannt, d. h. $\gamma_j := \dim \mathcal{N}(\boldsymbol{A} - \lambda_j \boldsymbol{I})$. Dabei gilt stets

$$1 \leqq \gamma_j \leqq \alpha_j \leqq n, \tag{6}$$

und es kann Eigenwerte λ_j mit $\gamma_j < \alpha_j$ geben. Solche Eigenwerte sind notwendig algebraisch mehrfach. Ein Beispiel gibt die Matrix

$$\boldsymbol{A} = \begin{pmatrix} 3 & 1 \\ -1 & 1 \end{pmatrix} \quad \text{mit} \quad \det(\boldsymbol{A} - \lambda \boldsymbol{I}) = (\lambda - 2)^2.$$

Sie besitzt den algebraischen zweifachen Eigenwert $\lambda_1 = 2$, aber die zugehörigen Eigenvektoren als Lösungen von $(\boldsymbol{A} - 2\boldsymbol{I})\,\boldsymbol{x} = \boldsymbol{o}$ bilden nur den eindimensionalen, durch $\boldsymbol{x} = (1, -1)^\mathsf{T}$ aufgespannten Teilraum. Eine Matrix $\boldsymbol{A}$, die mindestens einen Eigenwert λ_j mit $\gamma_j < \alpha_j$ besitzt, heißt *defektiv*.

Eigenvektoren, die zu verschiedenen Eigenwerten gehören, sind stets linear unabhängig. Die Anzahl der linear unabhängigen Eigenvektoren einer Matrix ist gleich der Summe der geometrischen Vielfachheiten ihrer Eigenwerte.

Wir bemerken abschließend, daß $\boldsymbol{A}$ und $\boldsymbol{A}^\mathsf{T}$ dieselben Eigenwerte besitzen, denn wegen der Determinanteneigenschaft (D_4) haben $\boldsymbol{A}$ und $\boldsymbol{A}^\mathsf{T}$ dasselbe charakteristische Polynom. Die Eigenvektoren sind jedoch i. allg. verschieden.

B. *Ähnlichkeitstransformationen und Eigenwertzerlegung*

Wir wollen das Verhalten der Eigenwertgleichung (1) gegenüber Transformationen

$$\boldsymbol{x} \to \boldsymbol{y} = \boldsymbol{T}^{-1}\boldsymbol{x}, \qquad \boldsymbol{T} \in \mathbf{C}^{n,n} \text{ regulär}, \tag{7}$$

untersuchen, d. h., wir ersetzen $\boldsymbol{x}$ durch $\boldsymbol{Ty}$. Einsetzen von (7) in (1) führt auf das transformierte Problem

$$\boldsymbol{By} = \lambda \boldsymbol{y} \quad \text{mit} \quad \boldsymbol{B} := \boldsymbol{T}^{-1}\boldsymbol{AT}. \tag{8}$$

1.2.1. *Aussage.* Die Matrix $\boldsymbol{A} \in \mathbf{C}^{n,n}$ werde mittels der regulären Matrix $\boldsymbol{T} \in \mathbf{C}^{n,n}$ in $\boldsymbol{B} = \boldsymbol{T}^{-1}\boldsymbol{AT}$ transformiert. Dann haben $\boldsymbol{A}$ und $\boldsymbol{B}$ dieselben Eigenwerte, und sowohl ihre algebraischen als auch ihre geometrischen Vielfachheiten stimmen überein. Dabei ist $\boldsymbol{x}$ Eigenvektor von $\boldsymbol{A}$ zum Eigenwert λ genau dann, wenn $\boldsymbol{y} = \boldsymbol{T}^{-1}\boldsymbol{x}$ Eigenvektor von $\boldsymbol{B}$ zum selben Eigenwert ist.

Beweis. Aus $\boldsymbol{B} - \lambda \boldsymbol{I} = \boldsymbol{T}^{-1}\boldsymbol{AT} - \lambda \boldsymbol{T}^{-1}\boldsymbol{T} = \boldsymbol{T}^{-1}(\boldsymbol{A} - \lambda \boldsymbol{I})\,\boldsymbol{T}$ folgt wegen (D_5) sofort

$$\det(\boldsymbol{B} - \lambda \boldsymbol{I}) = \det(\boldsymbol{T}^{-1}) \det(\boldsymbol{A} - \lambda \boldsymbol{I}) \det(\boldsymbol{T}).$$

Analog ergibt sich aus $\boldsymbol{T}^{-1}\boldsymbol{T} = \boldsymbol{I}$ die Beziehung $\det(\boldsymbol{T}^{-1}) \det(\boldsymbol{T}) = \det(\boldsymbol{I}) = 1$, was auf $\det(\boldsymbol{B} - \lambda \boldsymbol{I}) = \det(\boldsymbol{A} - \lambda \boldsymbol{I})$, also die Gleichheit der charakteristischen Polynome von $\boldsymbol{B}$ und

$\boldsymbol{A}$ führt. Daher haben $\boldsymbol{B}$ und $\boldsymbol{A}$ dieselben Eigenwerte einschließlich ihrer algebraischen Vielfachheiten. Die Aussage über die Eigenvektoren folgt aus der Äquivalenz von (1) und (8) und zieht $\mathcal{N}(\boldsymbol{B} - \lambda\boldsymbol{I}) = \boldsymbol{T}^{-1}\mathcal{N}(\boldsymbol{A} - \lambda\boldsymbol{I})$ nach sich. Nach Ü 1.1.2 haben dann $\mathcal{N}(\boldsymbol{B} - \lambda\boldsymbol{I})$ und $\mathcal{N}(\boldsymbol{A} - \lambda\boldsymbol{I})$ dieselbe Dimension, d. h., auch die geometrischen Vielfachheiten sind dieselben. □

Die Transformation $\boldsymbol{A} \to \boldsymbol{B} = \boldsymbol{T}^{-1}\boldsymbol{A}\boldsymbol{T}$ mittels einer regulären Matrix $\boldsymbol{T}$ wird *Ähnlichkeitstransformation* genannt; $\boldsymbol{A}$ und $\boldsymbol{B}$ heißen *ähnlich*. Aussage 1.2.1 bedeutet: *Die Eigenwerte einer Matrix einschließlich ihrer algebraischen und geometrischen Vielfachheiten sind invariant unter Ähnlichkeitstransformationen.*

Für theoretische wie praktische Zwecke sind diejenigen Ähnlichkeitstransformationen von besonderem Interesse, die $\boldsymbol{A}$ in eine Matrix $\boldsymbol{B}$ überführen, deren Eigenwerte sofort abgelesen bzw. einfach berechnet werden können. Als Zielmatrizen mit dieser Eigenschaft bieten sich speziell Blockdreiecksmatrizen an.

1.2.2. *Aussage.* Die Eigenwerte der Blockdreiecksmatrix

$$\boldsymbol{R} = \begin{pmatrix} \boldsymbol{R}_{11} & \boldsymbol{R}_{12} & \dots & \boldsymbol{R}_{1s} \\ & \boldsymbol{R}_{22} & \dots & \boldsymbol{R}_{2s} \\ & & \ddots & \vdots \\ \boldsymbol{O} & & & \boldsymbol{R}_{ss} \end{pmatrix} \in \mathbf{C}^{n,n}$$

mit quadratischen Diagonalblöcken $\boldsymbol{R}_{jj}$ der Dimension n_j, $n_1 + \dots + n_s = n$, sind gerade die Eigenwerte aller dieser Diagonalblöcke.

Beweis. Mit der Determinanteneigenschaft (D_7) erhält man

$$\det(\boldsymbol{R} - \lambda\boldsymbol{I}) = \det(\boldsymbol{R}_{11} - \lambda\boldsymbol{I}) \det(R_{22} - \lambda\boldsymbol{I}) \cdots \det(\boldsymbol{R}_{ss} - \lambda\boldsymbol{I}),$$

denn $\boldsymbol{R} - \lambda\boldsymbol{I}$ ist eine Blockdreiecksmatrix derselben Struktur wie $\boldsymbol{R}$. □

Aus 1.2.2 ergibt sich als wichtiger Sonderfall: *Die Eigenwerte einer Dreiecksmatrix und speziell die einer Diagonalmatrix sind gleich den Diagonalelementen.*

Besonders einfach in bezug auf die Struktur ihrer Eigenwerte und Eigenvektoren sind reelle symmetrische Matrizen.

1.2.3. *Aussage.* Für jede reelle symmetrische Matrix $\boldsymbol{A}$ gilt:

(i) $\boldsymbol{A}$ besitzt nur reelle Eigenwerte, und die Eigenvektoren können reell gewählt werden.

(ii) Eigenvektoren zu verschiedenen Eigenwerten sind orthogonal.

(iii) Algebraische und geometrische Vielfachheiten der Eigenwerte stimmen überein, d. h., zu einem Eigenwert λ_j der algebraischen Vielfachheit α_j gehören auch α_j linear unabhängige Eigenvektoren.

Beweis. (i) Es sei $\lambda \in \mathbf{C}$ ein Eigenwert und $\boldsymbol{x} \in \mathbf{C}^n$ ein zugehöriger Eigenvektor, d. h., es gelte $\boldsymbol{y} := \boldsymbol{A}\boldsymbol{x} = \lambda\boldsymbol{x}$. Durch Linksmultiplikation mit $\boldsymbol{x}^{\mathsf{H}}$ folgt $\boldsymbol{x}^{\mathsf{H}}\boldsymbol{y} = \lambda\boldsymbol{x}^{\mathsf{H}}\boldsymbol{x}$, also $\lambda = \boldsymbol{x}^{\mathsf{H}}\boldsymbol{y}/\boldsymbol{x}^{\mathsf{H}}\boldsymbol{x}$. Nun ist $\boldsymbol{x}^{\mathsf{H}}\boldsymbol{x}$ reell und positiv, und für den Zähler gilt $\overline{\boldsymbol{x}^{\mathsf{H}}\boldsymbol{y}} = \boldsymbol{x}^{\mathsf{T}}\overline{\boldsymbol{y}} = \overline{\boldsymbol{y}}^{\mathsf{T}}\boldsymbol{x} = \boldsymbol{y}^{\mathsf{H}}\boldsymbol{x} = (\overline{\boldsymbol{A}\boldsymbol{x}})^{\mathsf{T}}\boldsymbol{x}$ $= \bar{\boldsymbol{x}}^{\mathsf{T}}\bar{\boldsymbol{A}}^{\mathsf{T}}\boldsymbol{x} = \boldsymbol{x}^{\mathsf{H}}\boldsymbol{A}\boldsymbol{x} = \boldsymbol{x}^{\mathsf{H}}\boldsymbol{y}$, da $\boldsymbol{A}$ reell und symmetrisch ist. Hieraus folgt $\bar{\lambda} = \overline{\boldsymbol{x}^{\mathsf{H}}\boldsymbol{y}/\boldsymbol{x}^{\mathsf{H}}\boldsymbol{x}}$ $= \boldsymbol{x}^{\mathsf{H}}\boldsymbol{y}/\boldsymbol{x}^{\mathsf{H}}\boldsymbol{x} = \lambda$, d. h., λ ist reell. Die Eigenvektoren als Lösungen des reellen Systems $(\boldsymbol{A} - \lambda\boldsymbol{I})\,\boldsymbol{x} = \boldsymbol{o}$ können dann stets reell gewählt werden.

(ii) Es seien $\lambda_i \neq \lambda_j$ zwei Eigenwerte von $\boldsymbol{A}$ mit zugehörigen reellen Eigenvektoren $\boldsymbol{x}^i$, $\boldsymbol{x}^j$. Aus $\boldsymbol{A}\boldsymbol{x}^i = \lambda_i\boldsymbol{x}^i$ folgt $(\boldsymbol{x}^j)^{\mathsf{T}}\boldsymbol{A}\boldsymbol{x}^i = \lambda_i(\boldsymbol{x}^j)^{\mathsf{T}}\boldsymbol{x}^i$, und ebenso erhält man $(\boldsymbol{x}^i)^{\mathsf{T}}\boldsymbol{A}\boldsymbol{x}^j = \lambda_j(\boldsymbol{x}^i)^{\mathsf{T}}\boldsymbol{x}^j$.

Subtraktion beider Gleichungen liefert unter Beachtung der Symmetrie von $\boldsymbol{A}$ die Beziehung $0 = (\lambda_i - \lambda_j)[\boldsymbol{x}^{i\mathsf{T}} \boldsymbol{x}^j]$. Wegen $\lambda_i \neq \lambda_j$ zieht dies $\boldsymbol{x}^{i\mathsf{T}} \boldsymbol{x}^j = 0$, also die Orthogonalität der Eigenvektoren nach sich.

Auf den Beweis von (iii) soll an dieser Stelle verzichtet werden; ein konstruktiver Beweis wird später in Abschnitt 13.2 mittels des Jacobi-Verfahrens gegeben werden. □

Aus (iii) folgt, daß im Fall eines Eigenwertes λ_j der algebraischen Vielfachheit α_j jeweils beliebige α_j linear unabhängige Vektoren aus dem α_j-dimensionalen Teilraum $\mathcal{N}(\boldsymbol{A} - \lambda_j \boldsymbol{I})$ als Eigenvektoren zu λ_j genommen werden können. Insbesondere können diese Vektoren als orthogonale Basis von $\mathcal{N}(\boldsymbol{A} - \lambda_j \boldsymbol{I})$ ausgewählt werden. Es gibt daher stets n orthogonale Eigenvektoren $\boldsymbol{u}^1, \ldots, \boldsymbol{u}^n$ zu den Eigenwerten $\lambda_1, \ldots, \lambda_n$, von denen gewisse entsprechend ihrer Vielfachheit mehrfach auftreten können. Ohne Einschränkung der Allgemeinheit können die Eigenvektoren als normiert bezüglich $\|\boldsymbol{x}\|_2$ vorausgesetzt werden. Die Gleichungen $\boldsymbol{A}\boldsymbol{u}^j = \lambda_j \boldsymbol{u}^j$ $(j = 1, \ldots, n)$ lassen sich dann mit der orthogonalen Matrix $\boldsymbol{U} := (\boldsymbol{u}^1, \ldots, \boldsymbol{u}^n)$ und der Diagonalmatrix $\boldsymbol{\Lambda} := \operatorname{diag}(\lambda_1, \ldots, \lambda_n)$ in Matrixform $\boldsymbol{AU} = \boldsymbol{U\Lambda}$ schreiben. Durch Linksmultiplikation mit $\boldsymbol{U}^{-1}$ folgt schließlich $\boldsymbol{U}^{-1}\boldsymbol{AU} = \boldsymbol{U}^{\mathsf{T}}\boldsymbol{AU} = \boldsymbol{\Lambda}$. Damit ist der folgende Satz bewiesen:

1.2.4. *Eigenwertzerlegung reeller symmetrischer Matrizen.* Zu jeder reellen symmetrischen Matrix $\boldsymbol{A} \in \mathbf{R}^{n,n}$ gibt es eine orthogonale Matrix $\boldsymbol{U} \in \mathbf{R}^{n,n}$ mit

$$\boldsymbol{U}^{\mathsf{T}}\boldsymbol{AU} = \boldsymbol{\Lambda} = \operatorname{diag}(\lambda_1, \ldots, \lambda_n) \in \mathbf{R}^{n,n}, \tag{9}$$

wobei $\lambda_1, \ldots, \lambda_n$ die Eigenwerte von $\boldsymbol{A}$ sind.

Satz 1.2.4 besagt: *Jede reelle symmetrische Matrix kann orthogonal ähnlich auf reelle Diagonalform transformiert werden.* In der nach $\boldsymbol{A}$ aufgelösten Form lautet (9)

$$\boldsymbol{A} = \boldsymbol{U\Lambda U}^{\mathsf{T}} = (\boldsymbol{u}^1, \ldots, \boldsymbol{u}^n) \begin{pmatrix} \lambda_1 \boldsymbol{u}^{1\mathsf{T}} \\ \vdots \\ \lambda_n \boldsymbol{u}^{n\mathsf{T}} \end{pmatrix} = \sum_{j=1}^{n} \lambda_j \boldsymbol{u}^j \boldsymbol{u}^{j\mathsf{T}}, \tag{10}$$

was die Benennung *Eigenwertzerlegung* motiviert.

Bei nichtsymmetrischen Matrizen ist die Situation weitaus komplizierter, da dann sowohl komplexe Eigenpaare als auch weniger als n linear unabhängige Eigenvektoren auftreten können. Wir schließen zunächst den letzten Fall aus, d. h., wir betrachten eine nichtdefektive Matrix $\boldsymbol{A}$, und $\boldsymbol{x}^1, \ldots, \boldsymbol{x}^n \in \mathbf{C}^n$ seien linear unabhängige Eigenvektoren zu den Eigenwerten $\lambda_1, \ldots, \lambda_n \in \mathbf{C}$. Mit der regulären Matrix $\boldsymbol{X} := (\boldsymbol{x}^1, \ldots, \boldsymbol{x}^n) \in \mathbf{C}^{n,n}$ und $\boldsymbol{\Lambda} := \operatorname{diag}(\lambda_1, \ldots, \lambda_n) \in \mathbf{C}^{n,n}$ folgt dann $\boldsymbol{AX} = \boldsymbol{X\Lambda}$, also $\boldsymbol{X}^{-1}\boldsymbol{AX} = \boldsymbol{\Lambda}$.

1.2.5. *Aussage.* Zu jeder nichtdefektiven Matrix $\boldsymbol{A} \in \mathbf{R}^{n,n}$ gibt es eine reguläre Matrix $\boldsymbol{X} \in \mathbf{C}^{n,n}$ mit

$$\boldsymbol{X}^{-1}\boldsymbol{AX} = \boldsymbol{\Lambda} = \operatorname{diag}(\lambda_1, \ldots, \lambda_n) \in \mathbf{C}^{n,n}. \tag{11}$$

Falls $\boldsymbol{A}$ nur reelle Eigenwerte hat, kann $\boldsymbol{X}$ reell gewählt werden.

Aussage 1.2.5 begründet, warum nichtdefektive Matrizen — d. h. Matrizen, bei denen die algebraische Vielfachheit der Eigenwerte mit der geometrischen überein-

stimmt — auch *diagonalähnlich* genannt werden. *Speziell sind Matrizen mit n verschiedenen Eigenwerten* wegen (6) *diagonalähnlich.*

Der Fall defektiver, also nichtdiagonalähnlicher Matrizen wird durch den folgenden Satz beschrieben:

1.2.6. *Satz.* Zu jeder Matrix $\boldsymbol{A} \in \mathbf{R}^{n,n}$ gibt es eine reguläre Matrix $\boldsymbol{T} \in \mathbf{C}^{n,n}$, so daß

$$\boldsymbol{T}^{-1}\boldsymbol{A}\boldsymbol{T} = \boldsymbol{J} = \begin{pmatrix} \boldsymbol{J}(\lambda_1, n_1) & & & \\ & \boldsymbol{J}(\lambda_2, n_2) & & \\ & & \ddots & \\ & & & \boldsymbol{J}(\lambda_s, n_s) \end{pmatrix} \tag{12}$$

mit Diagonalblöcken

$$\boldsymbol{J}(\lambda_j, n_j) = \begin{pmatrix} \lambda_j & 1 & & 0 \\ & \lambda_j & 1 & \\ & & \ddots & \ddots \\ & & & 1 \\ 0 & & & \lambda_j \end{pmatrix} \in \mathbf{C}^{n_j,n_j}$$

und $n_1 + \cdots + n_s = n$ gilt. Falls $\boldsymbol{A}$ nur reelle Eigenwerte hat, kann $\boldsymbol{T}$ reell gewählt werden.

Die Matrix $\boldsymbol{J}$ wird *Jordansche Normalform* von $\boldsymbol{A}$ genannt; die $\boldsymbol{J}(\lambda_j, n_j)$ heißen *Jordanblöcke*, für $n_j = 1$ ist $\boldsymbol{J}(\lambda_j, 1) = (\lambda_j)$. *Die Matrix $\boldsymbol{J}$ ist bis auf die Anordnung der Diagonalblöcke eindeutig durch $\boldsymbol{A}$ bestimmt.* Die Anzahl s der Diagonalblöcke ist gleich der Anzahl der linear unabhängigen Eigenvektoren von $\boldsymbol{A}$. Diese stehen in denjenigen Spalten von $\boldsymbol{T}$, die den ersten Spalten der Jordanblöcke entsprechen, vgl. Ü 1.2.3. Sie werden durch sog. Hauptvektoren zu einer Basis von $\mathbf{C}^n$ ergänzt. Eine diagonalähnliche Matrix hat genau n Jordanblöcke der Ordnung 1.

Ein Beispiel für die Jordansche Normalform einer Matrix ist

$$\boldsymbol{J} = \left(\begin{array}{cc|cc|ccc|cc} 4 & 1 & & & & & & & \\ & 4 & & & & & & & \\ \hline & & 4 & & & & & & \\ \hline & & & & 4 & 1 & & & \\ & & & & & 4 & 1 & & \\ & & & & & & 4 & & \\ \hline & & & & & & & 2 & 1 \\ & & & & & & & & 2 \end{array}\right).$$

Die Matrix $\boldsymbol{J}$ und damit auch $\boldsymbol{A}$ besitzt den algebraisch sechsfachen Eigenwert $\lambda = 4$ mit der geometrischen Vielfachheit 3 sowie den algebraisch doppelten Eigenwert $\lambda = 2$ mit der geometrischen Vielfachheit 1. Die drei linear unabhängigen Eigenvektoren zu $\lambda_1 = 4$ stehen in den Spalten 1, 3 und 4 von $\boldsymbol{T}$, der zu $\lambda_2 = 2$ gehörende in Spalte 7. Das Beispiel weist darauf hin, daß mehrere Jordanblöcke zum selben Eigenwert gehören können.

In Ü 1.2.4 wird gezeigt, daß die Einsen in der Nebendiagonale der Jordanschen Normalform durch beliebige Zahlen $\varepsilon > 0$ ersetzt werden können.

Die Ähnlichkeitstransformation aus 1.2.5 und 1.2.6 zielen wie die aus 1.2.4 auf eine Diagonal- bzw. Fast-Diagonalform hin, allerdings ist dann die Transformationsmatrix nicht mehr orthogonal. Ein alternatives, für die Algorithmenentwicklung günstigeres Vorgehen besteht darin, als Zielmatrix lediglich eine Dreiecksform anzustreben, dafür aber mit orthogonalen bzw. unitären Transformationen als deren komplexe Entsprechung zu arbeiten. Eine Matrix $\boldsymbol{U} \in \mathbf{C}^{n,n}$ heißt dabei *unitär*, wenn

$$\boldsymbol{U}^{\mathsf{H}}\boldsymbol{U} = \boldsymbol{I} \quad \text{bzw. äquivalent} \quad \boldsymbol{U}\boldsymbol{U}^{\mathsf{H}} = \boldsymbol{I}$$

gilt, und für solche Matrizen ist

$$\boldsymbol{U}^{\mathsf{H}} = \boldsymbol{U}^{-1},$$

vgl. (1.1.29) und (1.1.30).

1.2.7. *Satz von Schur.* Zu jeder Matrix $\boldsymbol{A} \in \mathbf{R}^{n,n}$ gibt es eine unitäre Matrix $\boldsymbol{U} \in \mathbf{C}^{n,n}$, so daß

$$\boldsymbol{U}^{\mathsf{H}}\boldsymbol{A}\boldsymbol{U} = \boldsymbol{R} \tag{13}$$

mit einer oberen Dreiecksmatrix $\boldsymbol{R} \in \mathbf{C}^{n,n}$ gilt. Falls $\boldsymbol{A}$ nur reelle Eigenwerte besitzt, kann $\boldsymbol{U}$ reell, d. h. orthogonal gewählt werden.

Jede Matrix läßt sich also *unitär ähnlich auf obere Dreiecksform transformieren, und die Diagonalelemente von* $\boldsymbol{R}$ *sind gerade die Eigenwerte von* $\boldsymbol{A}$.

Für die Anwendungen hat Satz 1.2.7 den Nachteil, daß im Fall einer Matrix mit komplexen Eigenwerten sowohl $\boldsymbol{R}$ als auch $\boldsymbol{U}$ komplex werden und gegebenenfalls in komplexer Arithmetik berechnet werden müssen. Andererseits treten für reelles $\boldsymbol{A}$ stets Paare von konjugiert komplexen Eigenwerten $\alpha \pm \mathrm{i}\beta$ auf, die selbst Eigenwerte von reellen (2, 2)-Matrizen sind, siehe Ü 1.2.2. Es liegt daher nahe, auf das explizite Auftreten der komplexen Eigenwerte von $\boldsymbol{A}$ in der Diagonalen von $\boldsymbol{R}$ zu verzichten und dort stattdessen reelle Zweierblöcke anzusiedeln, deren zwei Eigenwerte gerade ein Paar konjugiert komplexer oder zwei reelle Eigenwerte von $\boldsymbol{A}$ sind. Dann kann die gesamte Transformation sogar im Reellen durchgeführt werden, wie der folgende Satz als Modifikation von 1.2.7 zeigt.

1.2.8. *Satz.* Zu jeder Matrix $\boldsymbol{A} \in \mathbf{R}^{n,n}$ gibt es eine orthogonale Matrix $\boldsymbol{U} \in \mathbf{R}^{n,n}$ mit

$$\boldsymbol{U}^{\mathsf{T}}\boldsymbol{A}\boldsymbol{U} = \boldsymbol{R}, \tag{14}$$

wobei $\boldsymbol{R}$ eine obere Blockdreiecksmatrix

$$\boldsymbol{R} = \begin{pmatrix} \boldsymbol{R}_{11} & \boldsymbol{R}_{12} & \dots & \boldsymbol{R}_{1s} \\ & \boldsymbol{R}_{22} & \dots & \boldsymbol{R}_{2s} \\ & & \ddots & \vdots \\ & & & \boldsymbol{R}_{ss} \end{pmatrix} \in \mathbf{R}^{n,n}$$

mit quadratischen Diagonalblöcken $\boldsymbol{R}_{jj} \in \mathbf{R}^{n_j,n_j}$ der Dimension $n_j = 1$ oder $n_j = 2$ ist.

Nach 1.2.2 sind die Eigenwerte von $\boldsymbol{R}$ gerade die Eigenwerte der Diagonalblöcke $\boldsymbol{R}_{jj}$. Für Blöcke der Ordnung 1 sind dies die Diagonalelemente selbst, für (2,2)-Blöcke ergeben sie sich aus dem zugehörigen charakteristischen Polynom, d. h. aus Wurzeln einer quadratischen Gleichung. Die Sätze 1.2.7 und 1.2.8 bilden die theoretische Grundlage für den $\boldsymbol{QR}$-Algorithmus zur Eigenwertbestimmung nichtsymmetrischer Matrizen, auf den wir allerdings in diesem Buch nicht eingehen.

Wir geben abschließend zu diesem Komplex noch einige einfach berechenbare Schranken für die Eigenwerte einer Matrix an.

1.2.9. *Aussage.* Für jeden Eigenwert $\lambda_j \in \mathbf{C}$ einer Matrix $\boldsymbol{A} \in \mathbf{R}^{n,n}$ gilt

$$|\lambda_j| \leqq \|\boldsymbol{A}\|, \tag{15}$$

wobei $\|\boldsymbol{A}\|$ irgendeine multiplikative Matrixnorm, speziell also eine der Normen $\|\boldsymbol{A}\|_p$, $p \in \{1, 2, \infty, F\}$, sei.

Beweis. Es sei $\{\lambda, \boldsymbol{x}\}$ ein Eigenpaar von $\boldsymbol{A}$. Aus $\boldsymbol{A}\boldsymbol{x} = \lambda\boldsymbol{x}$ folgt dann auf Grund der vorausgesetzten Multiplikativität der Matrixnorm sofort $|\lambda|\,\|\boldsymbol{x}\| = \|\lambda\boldsymbol{x}\| = \|\boldsymbol{A}\boldsymbol{x}\| \leqq \|\boldsymbol{A}\|\,\|\boldsymbol{x}\|$, wegen $\boldsymbol{x} \neq \boldsymbol{o}$ also die Behauptung. □

Bessere Schranken liefert in der Regel der folgende Satz.

1.2.10. *Kreisesatz von Geršgorin.* Jeder Eigenwert $\lambda_j \in \mathbf{C}$ einer Matrix $\boldsymbol{A} \in \mathbf{R}^{n,n}$ liegt in mindestens einem der Kreise

$$\mathcal{K}_i := \left\{\lambda \in \mathbf{C}\colon |\lambda - a_{ii}| \leqq \sum_{\substack{j=1 \\ j \neq i}}^{n} |a_{ij}| =: r_i\right\} \qquad (i = 1, \ldots, n). \tag{16}$$

Wir bemerken zunächst, daß $\mathcal{K}_i$ einen Kreis der komplexen Ebene mit dem Mittelpunkt a_{ii} und dem Radius r_i darstellt, und beweisen jetzt die Aussage des Satzes. Es sei dazu $\{\lambda, \boldsymbol{x}\}$ ein Eigenpaar von $\boldsymbol{A}$. Dann wird i so gewählt, daß $|x_i| = \|\boldsymbol{x}\|_\infty$ gilt. Aus $(\boldsymbol{A}\boldsymbol{x})_i = \lambda x_i$ ergibt sich dann sofort $(\boldsymbol{A}\boldsymbol{x})_i - a_{ii}x_i = (\lambda - a_{ii})\,x_i$, also

$$|\lambda - a_{ii}|\,|x_i| = |(\boldsymbol{A}\boldsymbol{x})_i - a_{ii}x_i| \leqq \sum_{\substack{j=1 \\ j \neq i}}^{n} |a_{ij}x_j| = \sum_{\substack{j=1 \\ j \neq i}}^{n} |a_{ij}|\,|x_j| \leqq \left\{\sum_{\substack{j=1 \\ j \neq i}}^{n} |a_{ij}|\right\} |x_i|,$$

d. h. $\lambda \in \mathcal{K}_i$. □

1.2.11. *Verschärfung.* Hat die Vereinigung von r Kreisen $\mathcal{K}_i$ keine gemeinsamen Punkte mit den restlichen Kreisen, so liegen in ihr genau r Eigenwerte von $\boldsymbol{A}$. Insbesondere enthält jeder Kreis $\mathcal{K}_i$ genau einen Eigenwert, wenn alle Kreise disjunkt sind.

Dies ergibt sich aus 1.2.10, indem die Nichtdiagonalelemente stetig und monoton von 0 auf ihren vorgegebenen Wert transformiert werden und die stetige Abhängigkeit der Eigenwerte von den Matrixelementen berücksichtigt wird. □

C. *Kongruenztransformationen und quadratische Formen*

Wir betrachten jetzt das Verhalten der durch die symmetrische Matrix $\boldsymbol{A} \in \mathbf{S}^{n,n}$ erzeugten quadratischen Form $\boldsymbol{x}^{\mathrm{T}}\boldsymbol{A}\boldsymbol{x}$ gegenüber der Transformation (7). Einsetzen

von $\boldsymbol{x} = \boldsymbol{T}\boldsymbol{y}$ führt auf

$$\boldsymbol{x}^{\mathsf{T}}\boldsymbol{A}\boldsymbol{x} = \boldsymbol{y}^{\mathsf{T}}\boldsymbol{B}\boldsymbol{y} \quad \text{mit} \quad \boldsymbol{B} := \boldsymbol{T}^{\mathsf{T}}\boldsymbol{A}\boldsymbol{T}. \tag{17}$$

Die Transformation $\boldsymbol{A} \to \boldsymbol{B} = \boldsymbol{T}^{\mathsf{T}}\boldsymbol{A}\boldsymbol{T}$ mittels einer regulären Matrix $\boldsymbol{T}$ wird *Kongruenztransformation* genannt; $\boldsymbol{A}$ und $\boldsymbol{B}$ heißen *kongruent*. Dabei bleibt die Symmetrie der Matrix erhalten. Wir suchen wieder solche Transformationen, für die $\boldsymbol{B}$ und damit die zugehörige Form $\boldsymbol{y}^{\mathsf{T}}\boldsymbol{B}\boldsymbol{y}$ möglichst einfache Gestalt annehmen. Dies ist der Fall, wenn $\boldsymbol{B}$ diagonal ist. Hierfür bietet sich die Ähnlichkeitstransformation (9) an, die wegen der Orthogonalität der Transformationsmatrix $\boldsymbol{U}$ gleichzeitig eine Kongruenztransformation ist.

1.2.12. *Aussage.* Es sei $\boldsymbol{U}^{\mathsf{T}}\boldsymbol{A}\boldsymbol{U} = \boldsymbol{\Lambda} = \operatorname{diag}(\lambda_1, \ldots, \lambda_n)$ mit $\boldsymbol{U}^{\mathsf{T}}\boldsymbol{U} = \boldsymbol{I}$ die Eigenwertzerlegung der reellen symmetrischen Matrix $\boldsymbol{A} \in \mathbf{S}^{n,n}$. Dann gilt für jedes $\boldsymbol{x} \in \mathbf{R}^n$

$$\boldsymbol{x}^{\mathsf{T}}\boldsymbol{A}\boldsymbol{x} = \boldsymbol{y}^{\mathsf{T}}\boldsymbol{\Lambda}\boldsymbol{y} = \sum_{j=1}^{n} \lambda_j (y_j)^2 \tag{18}$$

mit $\boldsymbol{y} := \boldsymbol{U}^{\mathsf{T}}\boldsymbol{x}$.

Die hier gegebene Transformation der quadratischen Form auf eine Summe von Quadraten mit den Eigenwerten λ_j von $\boldsymbol{A}$ als Koeffizienten heißt *Hauptachsentransformation*. Aus 1.2.12 folgt insbesondere: *$\boldsymbol{A}$ ist positiv definit bzw. semidefinit genau dann, wenn alle Eigenwerte positiv bzw. nichtnegativ sind.*

Wir werden später im Abschnitt 6.1.A sehen, daß es auch einfachere, allerdings nichtorthogonale Kongruenztransformationen von $\boldsymbol{A}$ auf Diagonalform $\boldsymbol{\Lambda} = \operatorname{diag}(\delta_j)$ gibt. Die Zahlen δ_j sind dann nicht mehr die Eigenwerte von $\boldsymbol{A}$, aber ihre Vorzeichenverteilung ist dieselbe wie die der λ_j.

1.2.13. *Trägheitsgesetz für quadratische Formen.* Für die reelle symmetrische Matrix $\boldsymbol{A} \in \mathbf{S}^{n,n}$ und reguläres $\boldsymbol{T} \in \mathbf{R}^{n,n}$ gelte

$$\boldsymbol{T}^{\mathsf{T}}\boldsymbol{A}\boldsymbol{T} = \boldsymbol{\Lambda} = \operatorname{diag}(\delta_j).$$

Dann ist die Anzahl der positiven, negativen und verschwindenden Diagonalelemente von $\boldsymbol{\Lambda}$ unabhängig von der Transformationsmatrix $\boldsymbol{T}$ und identisch mit der Anzahl der positiven, negativen und verschwindenden Eigenwerte von $\boldsymbol{A}$.

Satz 1.2.13 läßt sich wie folgt lesen: *Die Vorzeichenverteilung der Eigenwerte reeller symmetrischer Matrizen ist invariant unter Kongruenztransformationen.*

D. *Äquivalenztransformationen und Singulärwertzerlegung*

Wir betrachten abschließend eine beliebige, nicht notwendig quadratische Matrix $\boldsymbol{A} \in \mathbf{R}^{m,n}$ und wollen untersuchen, wie sich die durch $\boldsymbol{A}$ erzeugte Abbildung

$$\boldsymbol{y} = \boldsymbol{A}\boldsymbol{x} \tag{19}$$

gegenüber Transformationen verhält. Da Urbild- und Bildraum hier i. allg. verschieden sind, müssen $\boldsymbol{x}$ und $\boldsymbol{y}$ unterschiedlichen Transformationen unterworfen werden,

d. h., wir betrachten Transformationen der Form

$$\boldsymbol{x} \to \boldsymbol{x}' = \boldsymbol{S}^{-1}\boldsymbol{x},\ \boldsymbol{S} \in \mathbf{R}^{n,n} \text{ regulär}, \quad \boldsymbol{y} \to \boldsymbol{y}' = \boldsymbol{T}^{-1}\boldsymbol{y},\ \boldsymbol{T} \in \mathbf{R}^{m,m} \text{ regulär}. \tag{20}$$

Einsetzen von (20) in (19) führt auf

$$\boldsymbol{y}' = \boldsymbol{B}\boldsymbol{x}' \quad \text{mit} \quad \boldsymbol{B} := \boldsymbol{T}^{-1}\boldsymbol{A}\boldsymbol{S} \in \mathbf{R}^{m,n}. \tag{21}$$

Die Transformation $\boldsymbol{A} \to \boldsymbol{B} = \boldsymbol{T}^{-1}\boldsymbol{A}\boldsymbol{S}$ mit regulären Matrizen $\boldsymbol{S}, \boldsymbol{T}$ wird *Äquivalenztransformation* genannt; $\boldsymbol{A}$ und $\boldsymbol{B}$ heißen *äquivalent*. Dabei gilt

$$\mathcal{N}(\boldsymbol{B}) = \boldsymbol{S}^{-1}\mathcal{N}(\boldsymbol{A}), \qquad \mathcal{R}(\boldsymbol{B}) = \boldsymbol{T}^{-1}\mathcal{R}(\boldsymbol{A}). \tag{22}$$

Hieraus folgt: *Der Rang einer Matrix ist invariant unter Äquivalenztransformationen;* man beachte Ü 1.1.2.

Wie bei den bisher betrachteten Transformationen ist es auch hier möglich, die Matrix $\boldsymbol{A}$ auf Diagonalform zu transformieren, und dies sogar mittels orthogonaler Matrizen.

1.2.14. *Singulärwertzerlegung.* Zu jeder Matrix $\boldsymbol{A} \in \mathbf{R}^{m,n}$ gibt es orthogonale Matrizen $\boldsymbol{U} \in \mathbf{R}^{m,m}$ und $\boldsymbol{V} \in \mathbf{R}^{n,n}$, so daß

$$\boldsymbol{U}^{\mathrm{T}}\boldsymbol{A}\boldsymbol{V} = \boldsymbol{\Sigma} = \operatorname{diag}(\sigma_1, \ldots, \sigma_l) \in \mathbf{R}^{m,n} \tag{23}$$

mit $l := \min\{m, n\}$ und

$$\sigma_1 \geqq \sigma_2 \geqq \cdots \geqq \sigma_r > 0, \qquad \sigma_{r+1} = \cdots = \sigma_l = 0$$

gilt. Die Zahlen $\sigma_1, \ldots, \sigma_l$ sind durch $\boldsymbol{A}$ eindeutig festgelegt und heißen *Singulärwerte* von $\boldsymbol{A}$.

Die Diagonalmatrix $\boldsymbol{\Sigma}$ ist also von der Gestalt

$$\boldsymbol{\Sigma} = \left[\begin{array}{ccc|c} \sigma_1 & & & \\ & \ddots & & 0 \\ & & \sigma_r & \\ \hline & 0 & & 0 \end{array}\right] \in \mathbf{R}^{m,n};$$

im Fall $r = l = n$ fehlen die rechten, im Fall $r = l = m$ die unteren Nullblöcke in $\boldsymbol{\Sigma}$.

Bevor wir 1.2.14 beweisen, sollen einige Folgerungen angegeben werden:

Unter Beachtung von (22) ergibt sich sofort

$$\mathcal{R}(\boldsymbol{A}) = \boldsymbol{U}\mathcal{R}(\boldsymbol{\Sigma}), \qquad \mathcal{N}(\boldsymbol{A}) = \boldsymbol{V}\mathcal{N}(\boldsymbol{\Sigma}), \tag{24}$$

insbesondere ist also $\operatorname{rang}(\boldsymbol{A}) = \operatorname{rang}(\boldsymbol{\Sigma}) = r$, d. h., *die Anzahl r der positiven Singulärwerte ist gleich dem Rang der Matrix.*

Auflösen von (23) nach $\boldsymbol{A}$ liefert

$$\boldsymbol{A} = \boldsymbol{U}\boldsymbol{\Sigma}\boldsymbol{V}^{\mathrm{T}} = \sum_{j=1}^{r} \sigma_j \boldsymbol{u}^j \boldsymbol{v}^{j\mathrm{T}}, \tag{25}$$

was die Benennung Singulärwertzerlegung motiviert. Die Spalten von $\boldsymbol{V}$ bzw. $\boldsymbol{U}$ werden gelegentlich *Singulärvektoren* genannt; für diese gilt

$$\boldsymbol{A}\boldsymbol{v}^j = \sigma_j \boldsymbol{u}^j \qquad (j = 1, \ldots, l).$$

Einsetzen von (25) liefert

$$(\boldsymbol{A}^\mathsf{T}\boldsymbol{A})\,\boldsymbol{V} = \boldsymbol{V}(\boldsymbol{\Sigma}^\mathsf{T}\boldsymbol{\Sigma}) = \boldsymbol{V}\,\mathrm{diag}\,(\sigma_1^2, \ldots, \sigma_r^2, 0, \ldots, 0), \tag{26}$$

und ebenso folgt

$$(\boldsymbol{A}\boldsymbol{A}^\mathsf{T})\,\boldsymbol{U} = \boldsymbol{U}(\boldsymbol{\Sigma}\boldsymbol{\Sigma}^\mathsf{T}) = \boldsymbol{U}\,\mathrm{diag}\,(\sigma_1^2, \ldots, \sigma_r^2, 0, \ldots, 0). \tag{27}$$

Die Zahlen $\sigma_1^2, \ldots, \sigma_r^2$ sind also *die positiven Eigenwerte sowohl von $\boldsymbol{A}^\mathsf{T}\boldsymbol{A}$ als auch von $\boldsymbol{A}\boldsymbol{A}^\mathsf{T}$, und die restlichen $n - r$ bzw. $m - r$ Eigenwerte von $\boldsymbol{A}^\mathsf{T}\boldsymbol{A}$ bzw. $\boldsymbol{A}\boldsymbol{A}^\mathsf{T}$ sind* 0. *Insbesondere gilt*

$$r = \mathrm{rang}\,(\boldsymbol{A}) = \mathrm{rang}\,(\boldsymbol{A}^\mathsf{T}\boldsymbol{A}) = \mathrm{rang}\,(\boldsymbol{A}\boldsymbol{A}^\mathsf{T}), \tag{28}$$

und $\boldsymbol{A}^\mathsf{T}\boldsymbol{A}$ wie $\boldsymbol{A}\boldsymbol{A}^\mathsf{T}$ sind positiv semidefinit. Die Spalten von $\boldsymbol{V}$ sind die Eigenvektoren von $\boldsymbol{A}^\mathsf{T}\boldsymbol{A}$, die Spalten von $\boldsymbol{U}$ die Eigenvektoren von $\boldsymbol{A}\boldsymbol{A}^\mathsf{T}$.

Wir kommen jetzt zum Beweis von 1.2.14. Die Matrix $\boldsymbol{A}^\mathsf{T}\boldsymbol{A}$ ist symmetrisch und wegen $\boldsymbol{x}^\mathsf{T}\boldsymbol{A}^\mathsf{T}\boldsymbol{A}\boldsymbol{x} = \|\boldsymbol{A}\boldsymbol{x}\|_2^2 \geqq 0$ positiv semidefinit. Nach 1.2.4 existiert dann ein orthogonales $\boldsymbol{V} = (\boldsymbol{v}^1, \ldots, \boldsymbol{v}^n) \in \mathbf{R}^{n,n}$ mit $\boldsymbol{V}^\mathsf{T}(\boldsymbol{A}^\mathsf{T}\boldsymbol{A})\,\boldsymbol{V} = \boldsymbol{\Lambda} = \mathrm{diag}\,(\lambda_i)$ und $\lambda_i \geqq 0$. Ohne Einschränkung der Allgemeinheit kann $\lambda_1 \geqq \cdots \geqq \lambda_r > 0$, $\lambda_{r+1} = \cdots = \lambda_n = 0$ mit einem gewissen $r \in \{0, \ldots, n\}$ angenommen werden. Aus $\boldsymbol{A}^\mathsf{T}\boldsymbol{A}\boldsymbol{v}^j = \lambda_j \boldsymbol{v}^j$ folgt $\lambda_j = \boldsymbol{v}^{j\mathsf{T}}\boldsymbol{A}^\mathsf{T}\boldsymbol{A}\boldsymbol{v}^j = \|\boldsymbol{A}\boldsymbol{v}^j\|_2^2$, also $\boldsymbol{A}\boldsymbol{v}^j \neq \boldsymbol{o}$ für $j = 1, \ldots, r$ und $\boldsymbol{A}\boldsymbol{v}^j = \boldsymbol{o}$ für $j = r + 1, \ldots, n$. Mit den Festlegungen $\sigma_j := \sqrt{\lambda_j}$ und $\boldsymbol{u}^j := \boldsymbol{A}\boldsymbol{v}^j/\sigma_j$ $(j = 1, \ldots, r)$ ergibt sich dann für $i, j = 1, \ldots, r$

$$\boldsymbol{u}^{i\mathsf{T}}\boldsymbol{u}^j = \frac{1}{\sigma_i\sigma_j}\,\boldsymbol{v}^{i\mathsf{T}}\boldsymbol{A}^\mathsf{T}\boldsymbol{A}\boldsymbol{v}^j = \frac{\lambda_j}{\sigma_i\sigma_j}\,\boldsymbol{v}^{i\mathsf{T}}\boldsymbol{v}^j = \begin{cases} 1 & \text{für} \quad i = j, \\ 0 & \text{für} \quad i \neq j. \end{cases}$$

Die Vektoren $\{\boldsymbol{u}^1, \ldots, \boldsymbol{u}^r\}$ bilden daher ein orthonormiertes System und können durch Hinzunahme weiterer Vektoren $\boldsymbol{u}^{r+1}, \ldots, \boldsymbol{u}^m$ zu einer orthonormalen Basis von $\mathbf{R}^m$ ergänzt werden. Dann ist $\boldsymbol{U} = (\boldsymbol{u}^1, \ldots, \boldsymbol{u}^m)$ orthogonal. Mit der Festlegung $\sigma_{r+1} = \cdots = \sigma_l = 0$ folgt wegen

$$(\boldsymbol{U}^\mathsf{T}\boldsymbol{A}\boldsymbol{V})_{ij} = \boldsymbol{u}^{i\mathsf{T}}\boldsymbol{A}\boldsymbol{v}^j = \begin{cases} \sigma_j \boldsymbol{u}^{i\mathsf{T}}\boldsymbol{u}^j & \text{für} \quad j = 1, \ldots, r, \\ 0 & \text{für} \quad j = r + 1, \ldots, n \end{cases}$$

die gesuchte Beziehung $\boldsymbol{U}^\mathsf{T}\boldsymbol{A}\boldsymbol{V} = \boldsymbol{\Sigma}$. □

Die in Abschnitt 1.1 eingeführten Matrixnormen $\|\boldsymbol{A}\|_2$ und $\|\boldsymbol{A}\|_F$ können in einfacher Weise durch die Singulärwerte ausgedrückt werden.

1.2.15. *Aussage.* Für jedes $\boldsymbol{A} \in \mathbf{R}^{m,n}$ gilt

$$\|\boldsymbol{A}\|_2 = \sigma_1, \qquad \|\boldsymbol{A}\|_F = \sqrt{(\sigma_1)^2 + \cdots + (\sigma_r)^2}. \tag{29}$$

Beweis. Wegen der Orthogonalinvarianz beider Normen ist $\|\boldsymbol{A}\| = \|\boldsymbol{\Sigma}\|$. Aus (1.1.46) und (1.1.47) ergibt sich dann die Behauptung. □

Der erste Teil von (29) ist wegen (26) überdies identisch mit (1.1.42), womit diese Formel nachträglich bewiesen ist; außerdem folgt (1.1.44) ebenfalls aus (29).

Abschließend geben wir einige einfache Beziehungen zwischen Eigen- und Singulärwerten quadratischer Matrizen an.

1.2.16. *Aussage.* Die quadratische Matrix $\boldsymbol{A} \in \mathbf{R}^{n,n}$ besitze die Singulärwerte $\sigma_1 \geqq \cdots \geqq \sigma_n$ und die gemäß $|\lambda_1| \geqq \cdots \geqq |\lambda_n|$ geordneten Eigenwerte $\lambda_1, \ldots, \lambda_n \in \mathbf{C}$. Dann gelten folgende Beziehungen:

$$\text{(i)} \quad \det(\boldsymbol{A}) = \prod_{j=1}^{n} \lambda_j, \qquad |\det(\boldsymbol{A})| = \prod_{j=1}^{n} \sigma_j. \tag{30}$$

$$\text{(ii)} \quad \sigma_n \leqq |\lambda_j| \leqq \sigma_1 \qquad (j = 1, \ldots, n). \tag{31}$$

$$\text{(iii)} \quad \sigma_j = |\lambda_j| \qquad (j = 1, \ldots, n), \qquad \text{falls } \boldsymbol{A} = \boldsymbol{A}^{\mathsf{T}} \text{ ist}. \tag{32}$$

Beweis. (i) Unter Beachtung von (1.1.31) folgt der erste Teil direkt aus (13), während sich der zweite analog aus (23) ergibt.

(ii) Die rechte Seite der Ungleichung ist der Sonderfall von 1.2.9 für $\|\boldsymbol{A}\|_2$. Die linke Seite ist im Fall $\sigma_n = 0$ trivial. Es sei also $\sigma_n > 0$, d. h., $\boldsymbol{A}$ sei regulär. Dann ist $\boldsymbol{A}^{-1}$ vorhanden und hat die Singulärwerte $1/\sigma_j$ sowie die Eigenwerte $1/\lambda_j$. Anwendung der eben bewiesenen Ungleichung auf $\boldsymbol{A}^{-1}$ liefert dann die linke Seite.

(iii) Aus (26) folgt für symmetrisches $\boldsymbol{A}$ unmittelbar $\boldsymbol{A}^2 = \boldsymbol{A}^{\mathsf{T}}\boldsymbol{A} = \boldsymbol{V}\boldsymbol{\Sigma}^2\boldsymbol{V}^{\mathsf{T}}$, d. h. $\lambda_j^2 = \sigma_j^2$ und damit $|\lambda_j| = \sigma_j$. Man beachte dabei, daß $\boldsymbol{A}^2$ die Eigenwerte λ_j^2 hat. □

Für nichtsymmetrisches $\boldsymbol{A}$ können die Singulärwerte beliebig stark von den Eigenwerten abweichen, vgl. Ü 1.2.8.

Übungsaufgaben

Ü 1.2.1. Man zeige, daß für $\boldsymbol{A} = \boldsymbol{U} + \mathrm{i}\boldsymbol{V}$ mit $\boldsymbol{U}, \boldsymbol{V} \in \mathbf{R}^{m,n}$, $\boldsymbol{a} = \boldsymbol{b} + \mathrm{i}\boldsymbol{c}$ mit $\boldsymbol{b}, \boldsymbol{c} \in \mathbf{R}^m$ und $\boldsymbol{x} = \boldsymbol{u} + \mathrm{i}\boldsymbol{v}$ mit $\boldsymbol{u}, \boldsymbol{v} \in \mathbf{R}^n$ die folgenden Beziehungen gleichwertig sind:

$$\text{(i)}\ \boldsymbol{A}\boldsymbol{x} = \boldsymbol{a}, \quad \text{(ii)}\ \bar{\boldsymbol{A}}\bar{\boldsymbol{x}} = \bar{\boldsymbol{a}}, \quad \text{(iii)} \left(\begin{array}{c|c} \boldsymbol{U} & -\boldsymbol{V} \\ \hline \boldsymbol{V} & \boldsymbol{U} \end{array}\right) \begin{pmatrix} \boldsymbol{u} \\ \boldsymbol{v} \end{pmatrix} = \begin{pmatrix} \boldsymbol{b} \\ \boldsymbol{c} \end{pmatrix}.$$

Überdies gilt

$$\boldsymbol{Q}_m^{\mathsf{H}} \left(\begin{array}{c|c} \boldsymbol{U} + \mathrm{i}\boldsymbol{V} & \boldsymbol{O} \\ \hline \boldsymbol{O} & \boldsymbol{U} - \mathrm{i}\boldsymbol{V} \end{array}\right) \boldsymbol{Q}_n = \left(\begin{array}{c|c} \boldsymbol{U} & -\boldsymbol{V} \\ \hline \boldsymbol{V} & \boldsymbol{U} \end{array}\right) \quad \text{mit} \quad \boldsymbol{Q}_k := \frac{1}{\sqrt{2}} \left(\begin{array}{c|c} \boldsymbol{I}_k & -\mathrm{i}\boldsymbol{I}_k \\ \hline \mathrm{i}\boldsymbol{I}_k & \boldsymbol{I}_k \end{array}\right),$$

und $\boldsymbol{Q}$ ist unitär.

Ü 1.2.2. Man beweise, daß für $\boldsymbol{A} \in \mathbf{R}^{n,n}$ die folgenden Aussagen äquivalent sind:

(i) $(\alpha + \mathrm{i}\beta, \boldsymbol{u} + \mathrm{i}\boldsymbol{v})$ ist Eigenpaar von $\boldsymbol{A}$.

(ii) $(\alpha - \mathrm{i}\beta, \boldsymbol{u} - \mathrm{i}\boldsymbol{v})$ ist Eigenpaar von $\boldsymbol{A}$.

(iii) Mit $\boldsymbol{\Gamma} := \begin{pmatrix} \alpha & \beta \\ -\beta & \alpha \end{pmatrix} \in \mathbf{R}^{2,2}$ und $\boldsymbol{X} := \left(\begin{array}{c|c} & \\ \boldsymbol{u}, & \boldsymbol{v} \\ & \end{array}\right) \in \mathbf{R}^{n,2}$, $\boldsymbol{X} \neq \boldsymbol{O}$, gilt $\boldsymbol{A}\boldsymbol{X} = \boldsymbol{X}\boldsymbol{\Gamma}$. Man beachte dabei, daß $\boldsymbol{\Gamma}$ gerade die Eigenwerte $\lambda_{1,2} = \alpha \pm \mathrm{i}\beta$ besitzt.

Ü 1.2.3. Man überlege sich, daß der Jordanblock $\boldsymbol{J}(\lambda, k) \in \mathbf{C}^{k,k}$ den Eigenwert λ der algebraischen Vielfachheit k und der geometrischen Vielfachheit 1 besitzt. Wie lautet ein zugehöriger Eigenvektor?

Ü 1.2.4. Es sei $\boldsymbol{J} = \boldsymbol{T}^{-1}\boldsymbol{A}\boldsymbol{T}$ die Jordansche Normalform von $\boldsymbol{A} \in \mathbf{R}^{n,n}$. Mit $\varepsilon > 0$ werde $\boldsymbol{D} := \operatorname{diag}(1, \varepsilon, \ldots, \varepsilon^{n-1})$, $\boldsymbol{T}_\varepsilon := \boldsymbol{T}\boldsymbol{D}_\varepsilon$ und $\boldsymbol{J}_\varepsilon := \boldsymbol{T}_\varepsilon^{-1}\boldsymbol{A}\boldsymbol{T}_\varepsilon$ gebildet. Man zeige, daß $\boldsymbol{J}_\varepsilon$ diejenige Matrix ist, die aus $\boldsymbol{J}$ entsteht, indem dort die Einsen in der oberen Nebendiagonalen durch ε ersetzt werden. Dabei gilt $\|\boldsymbol{T}_\varepsilon^{-1}\| \cdot \|\boldsymbol{T}_\varepsilon\| \to \infty$ für $\varepsilon \to 0$, d. h., die Transformationsmatrix $\boldsymbol{T}_\varepsilon$ wird für $\varepsilon \to 0$ beliebig schlecht konditioniert, siehe Abschnitt 4.1.

Ü 1.2.5. Man überlege sich, daß Satz 1.2.7 für den Spezialfall einer reellen symmetrischen Matrix in Satz 1.2.4 übergeht.
Hinweis: Man zeige $\boldsymbol{R}^{\mathsf{T}} = \boldsymbol{R}$ und beachte, daß eine symmetrische obere Dreiecksmatrix notwendig diagonal ist.

Ü 1.2.6. Es ist zu zeigen, daß genau dann $\|\boldsymbol{A}\|_F = \|\boldsymbol{A}\|_2$ gilt, wenn $\operatorname{rang}(\boldsymbol{A}) \leqq 1$ ist.

Ü 1.2.7. Betrachtet werde eine Matrix $\boldsymbol{A} \in \mathbf{R}^{m,n}$ vom Rang r mit der Singulärwertzerlegung (25). Man zeige, daß für festes $s \in \{1, \ldots, r\}$ die Matrix $\boldsymbol{A}_s := \sum_{j=1}^{s} \sigma_j \boldsymbol{u}^j \boldsymbol{v}^{j\mathsf{T}}$ diejenige Matrix aus der Menge $\{\boldsymbol{B} \in \mathbf{R}^{m,n}: \operatorname{rang}(\boldsymbol{B}) \leqq s\}$ ist, die $\|\boldsymbol{B} - \boldsymbol{A}\|_F$ minimiert. Dies gilt auch für die Spektralnorm, allerdings ist der Beweis komplizierter.

Ü 1.2.8. Man zeige, daß sich die Singulärwerte der Matrix

$$\boldsymbol{A}(\alpha) := \begin{pmatrix} 1 & \alpha \\ 0 & 1 \end{pmatrix}$$

für großes α wie $\sigma_1(\alpha) \sim \sqrt{2 + \alpha^2}$, $\sigma_2(\alpha) \sim 1/\sqrt{2 + \alpha^2}$ verhalten, während die Eigenwerte durch $\lambda_1(\alpha) = \lambda_2(\alpha) = 1$ für alle α gegeben sind.

Ü 1.2.9. Man zeige, daß für jede Matrix $\boldsymbol{A} \in \mathbf{R}^{m,n}$

$$(\|\boldsymbol{A}\|_2)^2 = (\|\boldsymbol{A}^{\mathsf{T}}\|_2)^2 = \|\boldsymbol{A}^{\mathsf{T}}\boldsymbol{A}\|_2 = \|\boldsymbol{A}\boldsymbol{A}^{\mathsf{T}}\|_2 \leqq \|\boldsymbol{A}\|_1 \|\boldsymbol{A}\|_\infty \tag{33}$$

gilt, und folgere hieraus für symmetrisches $\boldsymbol{A} = \boldsymbol{A}^{\mathsf{T}} \in \mathbf{R}^{n,n}$

$$\|\boldsymbol{A}\|_2 \leqq \||\boldsymbol{A}|\|_2 \leqq \|\boldsymbol{A}\|_1 = \|\boldsymbol{A}\|_\infty. \tag{34}$$

Hinweis: Man beachte $\sigma_1^2 = \lambda_{\max}(\boldsymbol{A}^{\mathsf{T}}\boldsymbol{A}) \leqq \|\boldsymbol{A}^{\mathsf{T}}\boldsymbol{A}\|_\infty \leqq \|\boldsymbol{A}^{\mathsf{T}}\|_\infty \|\boldsymbol{A}\|_\infty = \|\boldsymbol{A}\|_1 \|\boldsymbol{A}\|_\infty$.

Ü 1.2.10. Man beweise die Gültigkeit von

$$\|\boldsymbol{A}\|_2 = \max\{|\boldsymbol{y}^{\mathsf{T}}\boldsymbol{A}\boldsymbol{x}|: \|\boldsymbol{y}\|_2 \leqq 1, \|\boldsymbol{x}\|_2 \leqq 1\} \tag{35}$$

für jede Matrix $\boldsymbol{A} \in \mathbf{R}^{m,n}$ und von

$$\|\boldsymbol{A}\|_2 = \max\{|\boldsymbol{x}^{\mathsf{T}}\boldsymbol{A}\boldsymbol{x}|: \|\boldsymbol{x}\|_2 \leqq 1\} \tag{36}$$

für symmetrisches $\boldsymbol{A} = \boldsymbol{A}^{\mathsf{T}} \in \mathbf{R}^{n,n}$.

Bemerkungen zum Kapitel 1

B 1.1. Der in diesem Kapitel behandelte Stoff ist Gegenstand der meisten Lehrbücher der linearen Algebra und Matrixtheorie. Wir verweisen z. B. auf GANTMAHER [86] als theoretisch orientierte sowie FADDEEV/FADDEEVA [63] als numerisch orientierte detaillierte Darstellungen.

B 1.2. Der größte Teil der in diesem Buch dargestellten Gebiete läßt sich im Reellen behandeln, lediglich bei Eigenwertaufgaben nichtsymmetrischer Matrizen müssen komplexe Eigenwerte und damit komplexe Vektoren und Matrizen zugelassen werden. Aus diesem Grund wählen wir die Darstellung im Reellen als Basis und nehmen nur dann, wenn es der Gegenstand erfordert, eine Erweiterung ins Komplexe vor.

B 1.3. Die reelle Blockdreiecksversion 1.2.8 des klassischen Schurschen Satzes 1.2.7 ist Grundlage für den im Reellen ausführbaren doppelten $\boldsymbol{QR}$-Algorithmus zur Eigenwertbestimmung bei nichtsymmetrischen Matrizen. Sie ist mit einer Beweisskizze bei STEWART [73] nachzulesen. Das Trägheitsgesetz 1.2.13 wird meist nach SYLVESTER benannt.

B 1.4. Obwohl die Singulärwertzerlegung 1.2.14 für quadratische Matrizen bereits 1889 durch SYLVESTER und für allgemeine Matrizen 1939 durch ECKART/YOUNG gefunden wurde, hat sie erst in jüngster Zeit Eingang in die numerische lineare Algebra und die zugehörige Spezialliteratur gefunden, siehe etwa FORSYTHE/MOLER [67], STEWART [73], LAWSON/HANSON [74] und VOEVODIN [77].

2. Aufgaben, Computer, Algorithmen

Dieses Kapitel spielt für das Verständnis des vorliegenden Buches eine grundlegende Rolle. Es dient

- der Vorstellung der hier behandelten konkreten Aufgabenklassen
- der Diskussion des Begriffes „Numerisches Problem" und dessen wesentlichen Eigenschaften, speziell der Abhängigkeit der Resultate von den Eingangsdaten und deren Fehlern
- der Untersuchung der auf derzeitigen Computern realisierten Arithmetik
- der Einführung und Erläuterung des Begriffs „Numerischer Algorithmus" und damit zusammenhängender Probleme wie Auswirkung von Rundungsfehlern während der Computerrechnung.

In der klassischen Mathematikausbildung werden diese Probleme meist nur kurz bzw. überhaupt nicht behandelt. Der Leser sei deshalb aufgefordert, die folgenden Seiten besonders sorgfältig zu studieren, die Beispiele gründlich nachzuvollziehen und sich um das aktive Verständnis der z. T. nicht ganz einfachen Begriffe und Denkweisen zu bemühen. Auch in den späteren Kapiteln wird der hier angesprochene Problemkreis immer wieder an konkreten Aufgaben erläutert werden, so daß sich nach und nach der erwünschte Zustand der Vertrautheit und aktiven Beherrschung einstellen wird.

2.1. Aufgabenklassen der linearen Algebra

Die Aufgaben der linearen Algebra treten in der Regel nicht eigenständig, sondern einzeln bzw. mehrfach als Teilaufgaben bei der Lösung komplexer Probleme aus den verschiedensten Anwendungsgebieten auf.

A.1. *Lösung quadratischer linearer Gleichungssysteme*

Gegeben sind eine reguläre Koeffizientenmatrix $\boldsymbol{A} \in \mathbf{R}^{n,n}$ und eine rechte Seite $\boldsymbol{b} \in \mathbf{R}^n$. *Gesucht* ist die in diesem Fall eindeutige Lösung $\boldsymbol{x} \in \mathbf{R}^n$ des linearen Gleichungssystems

$$\boldsymbol{A}\boldsymbol{x} = \boldsymbol{b}. \tag{1}$$

Beispiele für das Auftreten solcher Gleichungssysteme sind

- Bilanzprobleme in Technik und Ökonomie
- Analyse linearer elektrischer und mechanischer Netzwerke (Systeme hoher Dimension mit vielen unregelmäßig verteilten Nullen in $\boldsymbol{A}$)

- Lösung nichtlinearer Gleichungssysteme und Minimierungsaufgaben mittels Newton-ähnlicher Verfahren (pro Iterationsschritt ein System $\boldsymbol{A}_k\boldsymbol{x}^k = \boldsymbol{b}^k$ zu lösen)
- Interpolation und Approximation von Kurven und Flächen mittels Spline- und anderer Funktionen (Systeme mit Bandmatrizen)
- Integration von Anfangswertaufgaben bei Systemen gewöhnlicher Differentialgleichungen mittels impliziter Verfahren (Systeme hoher Dimension mit vielen unregelmäßig verteilten Nullen, siehe Ü 2.1.1)
- Diskretisierung von Randwertaufgaben bei gewöhnlichen und partiellen Differentialgleichungen mittels Differenzenverfahren oder finiter Elemente (Systeme hoher Dimension mit vielen regelmäßig verteilten Nullen, Systeme mit Bandmatrizen, siehe Ü 2.1.2)
- Lösung von Anfangs-Randwert-Aufgaben bei partiellen Differentialgleichungen.

Bei den angegebenen Beispielen treten u. a. folgende *Sonderfälle* und *Modifikationen* der Aufgabenstellung auf:

- Lösung von Systemen mit Matrizen spezieller Struktur:
  $\boldsymbol{A}$ symmetrisch bzw. symmetrisch und positiv definit und/oder
  $\boldsymbol{A}$ von hoher Dimension mit vielen regelmäßig bzw. unregelmäßig verteilten Nullen, speziell $\boldsymbol{A}$ Bandmatrix (Diskretisierung selbstadjungierter Differentialgleichungen, Optimierungsalgorithmen)
- Lösung mehrerer Systeme $\boldsymbol{A}\boldsymbol{x}^k = \boldsymbol{b}^k$ $(k = 1, \ldots, q)$ mit derselben Koeffizientenmatrix (iterative Verbesserung, Lösung von Matrixgleichungen $\boldsymbol{A}\boldsymbol{X} = \boldsymbol{B}$, $\boldsymbol{X} = (\boldsymbol{x}^1, \ldots, \boldsymbol{x}^q)$, $\boldsymbol{B} = (\boldsymbol{b}^1, \ldots, \boldsymbol{b}^q) \in \mathbf{R}^{n,q}$)
- Lösung mehrerer Systeme $\boldsymbol{A}_k\boldsymbol{x}^k = \boldsymbol{b}^k$ $(k = 1, 2, \ldots)$ mit verschiedenen Koeffizientenmatrizen, wobei

  rang $(\boldsymbol{A}_{k+1} - \boldsymbol{A}_k) = 1$ oder 2 (Quasi-Newton-Verfahren)

  oder

  rang $(\boldsymbol{A}_{k+1} - \boldsymbol{A}_k) = n$, aber n groß und Besetztheitsmuster aller $\boldsymbol{A}_k$ identisch

  (Anfangswertaufgaben bei Differentialgleichungen).

Eine Matrix, die viele Nullelemente enthält, wird *schwach besetzte Matrix* (engl. „sparse matrix", russ. „разреженная матрица") genannt. Man spricht von schwacher Besetztheit, wenn die Anzahl der Nichtnullelemente kleiner als etwa 10% der Gesamtzahl aller Elemente ist; meist ist sie sogar wesentlich geringer. *Schwach besetzte lineare Gleichungssysteme* entstehen z. B. bei der Analyse von Netzwerken und bei der Diskretisierung von Differentialgleichungen und *erfordern* i. allg. *spezielle Lösungsverfahren*, deren ausführliche Behandlung über den Rahmen dieses Buches hinausgeht. Einige Basistechniken werden im Abschnitt 6.4 vorgestellt.

A 2. *Berechnung inverser Matrizen*

Gegeben ist eine reguläre Matrix $\boldsymbol{A} \in \mathbf{R}^{n,n}$. *Gesucht* ist die inverse Matrix

$$\boldsymbol{A}^{-1}.$$

Ein wichtiger *Sonderfall* ist die Inversion einer symmetrischen, positiv definiten Matrix.

Beispiele für Probleme, bei denen inverse Matrizen explizit berechnet werden müssen, kommen aus der Statistik.

2.1.1. *Warnung.* Ausdrücke der Form $\boldsymbol{X} = \boldsymbol{A}^{-1}\boldsymbol{B}$, $\boldsymbol{B} \in \mathbf{R}^{n,q}$, sollten i. allg. nicht durch Berechnung von $\boldsymbol{A}^{-1}$ und anschließende Multiplikation mit $\boldsymbol{B}$ berechnet werden, sondern billiger und genauer durch direkte Berechnung von $\boldsymbol{X}$ aus der Matrixgleichung $\boldsymbol{AX} = \boldsymbol{B}$.

Insbesondere sollte die Lösung $\boldsymbol{x}$ des regulären Systems $\boldsymbol{Ax} = \boldsymbol{b}$ i. allg. nicht durch numerische Auswertung der Formel $\boldsymbol{x} = \boldsymbol{A}^{-1}\boldsymbol{b}$ berechnet werden.

Eine *Modifikation* der Aufgabenstellung ist die Berechnung mehrerer inverser Matrizen $\boldsymbol{A}_k$, wobei $\boldsymbol{A}_{k+1}$ aus $\boldsymbol{A}_k$ durch

— eine Rang-1-Änderung, speziell Ändern einer Zeile oder Spalte

— Streichen einer Zeile und Spalte

— Hinzufügen einer Zeile und Spalte

entsteht, siehe Abschnitt 6.5.

A 3. *Berechnung von Determinanten*

Gegeben ist die quadratische, nicht notwendig reguläre Matrix $\boldsymbol{A} \in \mathbf{R}^{n,n}$. *Gesucht* ist die Zahl

$$\det(\boldsymbol{A}).$$

Beispiele für das Auftreten von Determinantenberechnungen sind

— Verfahren zur Lösung linearer bzw. nichtlinearer Eigenwertprobleme $\boldsymbol{M}(\lambda)\,\boldsymbol{x} = \boldsymbol{o}$, die auf der Bestimmung der Wurzeln der skalaren Gleichung $p(\lambda) := \det\big(\boldsymbol{M}(\lambda)\big) = 0$ beruhen

— Verfahren zur Berechnung von Rückkehrpunkten und Verzweigungspunkten bei parameterabhängigen nichtlinearen Gleichungssystemen.

Dabei heißt das *Eigenwertproblem* $\boldsymbol{M}(\lambda)\,\boldsymbol{x} = \boldsymbol{o}$ *linear*, wenn $\boldsymbol{M}(\lambda) = \boldsymbol{A}_0 + \lambda\boldsymbol{A}_1$ mit Matrizen $\boldsymbol{A}_0, \boldsymbol{A}_1 \in \mathbf{R}^{n,n}$ gilt; andernfalls heißt es *nichtlinear*. Zum Beispiel legt $\boldsymbol{M}(\lambda) := \boldsymbol{A}_0 + \lambda\boldsymbol{A}_1 + \lambda^2\boldsymbol{A}_2$ ein nichtlineares — nämlich quadratisches — Eigenwertproblem fest.

B 1. *Lösung linearer Quadratmittelprobleme*

Gegeben sind eine Koeffizientenmatrix $\boldsymbol{A} \in \mathbf{R}^{m,n}$ mit $m \geqq n$ und eine rechte Seite $\boldsymbol{b} \in \mathbf{R}^m$. Dann hat das überbestimmte lineare Gleichungssystem $\boldsymbol{Ax} = \boldsymbol{b}$ i. allg. keine Lösung. *Gesucht* ist deshalb eine Lösung $\boldsymbol{x} \in \mathbf{R}^n$ des linearen Quadratmittelproblems

$$\|\boldsymbol{Ax} - \boldsymbol{b}\|_2 \to \underset{\boldsymbol{x} \in \boldsymbol{R}^n}{\text{Minimum}}!, \tag{2}$$

d. h. ein Vektor $\boldsymbol{x}$, der das *Residuum* $\boldsymbol{r}(\boldsymbol{x}) := \boldsymbol{b} - \boldsymbol{Ax}$ des überbestimmten Systems in der Euklidischen Norm minimiert. Für (2) verwenden wir äquivalent die ab-

kürzende Schreibweise

$$\boldsymbol{A}\boldsymbol{x} \cong \boldsymbol{b}. \tag{3}$$

Beispiele für das Auftreten linearer Quadratmittelprobleme sind

- lineare Regressionsprobleme (siehe Ü 2.1.3)
- Verfahren vom Gauß-Newton-Typ zur Lösung nichtlinearer Regressionsprobleme.

Sonderfälle und *Modifikationen* der Aufgabenstellung sind

- Lösung von Quadratmittelproblemen mit schwach besetzten Matrizen hoher Dimension
- Lösung mehrerer Probleme $\boldsymbol{A}\boldsymbol{x}^k \cong \boldsymbol{b}^k$ $(k = 1, \ldots, q)$ mit derselben Koeffizientenmatrix (iterative Verbesserung von Quadratmittellösungen, Matrixprobleme $\|\boldsymbol{A}\boldsymbol{X} - \boldsymbol{B}\|_F \to \text{Min!}$, $\boldsymbol{B}, \boldsymbol{X} \in \mathbf{R}^{m,q}$, siehe Ü 2.1.4)
- Lösung mehrerer Probleme $\boldsymbol{A}_k\boldsymbol{x}^k \cong \boldsymbol{b}^k$ mit unterschiedlichen Matrizen $\boldsymbol{A}_k$, wobei $\boldsymbol{A}_{k+1}$ aus $\boldsymbol{A}_k$ entsteht durch
  Rang-1-Änderung, speziell Ändern einer Zeile oder Spalte,
  Hinzufügen oder Streichen einer Zeile oder Spalte
  (Lineare Regressionsprobleme mit Änderungen in Ansatzfunktionen und Meßwertsätzen, siehe Ü 2.1.3)
- Lösung regularisierter Quadratmittelprobleme im Fall schlecht konditionierter bzw. inkorrekt gestellter Aufgaben $\boldsymbol{A}\boldsymbol{x} \cong \boldsymbol{b}$. Dieses Problem hängt eng mit der numerischen Bestimmung des Ranges der Matrix $\boldsymbol{A}$ zusammen, siehe Kapitel 11
- Lösung linearer Quadratmittelprobleme mit linearen Gleichheitsnebenbedingungen:

  $\|\boldsymbol{A}\boldsymbol{x} - \boldsymbol{b}\|_2 \to$ Minimum bei $\boldsymbol{C}\boldsymbol{x} = \boldsymbol{d}$

  mit $\boldsymbol{C} \in \mathbf{R}^{p,n}$, $\boldsymbol{d} \in \mathbf{R}^p$ und $m \geqq n - p$.

B 2. *Berechnung von Pseudoinversen*

Gegeben ist eine Matrix $\boldsymbol{A} \in \mathbf{R}^{m,n}$ beliebigen Formats (m, n). *Gesucht* ist ihre *Pseudoinverse*

$$\boldsymbol{A}^+.$$

Dabei ist $\boldsymbol{A}^+ \in \mathbf{R}^{n,m}$ gerade diejenige, durch $\boldsymbol{A}$ eindeutig festgelegte Matrix, die jedem $\boldsymbol{b} \in \mathbf{R}^m$ gemäß $\boldsymbol{x}^+ := \boldsymbol{A}^+\boldsymbol{b}$ die bezüglich $\|\boldsymbol{x}\|_2$ normkleinste Lösung $\boldsymbol{x}^+$ von $\boldsymbol{A}\boldsymbol{x} \cong \boldsymbol{b}$ zuordnet, siehe Abschnitt 8.1.

Beispiele für Aufgaben, bei denen Pseudoinverse explizit benötigt werden, kommen aus der Statistik.

2.1.2. *Warnung.* Ausdrücke der Form $\boldsymbol{Y} = \boldsymbol{A}^+\boldsymbol{B}$, $\boldsymbol{B} \in \mathbf{R}^{m,q}$, sollten i. allg. nicht durch Berechnung von $\boldsymbol{A}^+$ und anschließende Multiplikation mit $\boldsymbol{B}$, sondern direkt durch Berechnung der Lösungen $\boldsymbol{y}^k$ von $\boldsymbol{A}\boldsymbol{y}^k \cong \boldsymbol{b}^k$ $(k = 1, \ldots, q)$ mit minimalem $\|\boldsymbol{y}^k\|_2$ ermittelt werden, wobei $\boldsymbol{Y} = (\boldsymbol{y}^1, \ldots, \boldsymbol{y}^q)$ und $\boldsymbol{B} = (\boldsymbol{b}^1, \ldots, \boldsymbol{b}^q)$.

Modifikationen der Aufgabe bestehen in Änderungen und Regularisierungen von $\boldsymbol{A}$ analog zu B 1.

C 1. *Das spezielle symmetrische Eigenwertproblem*

Gegeben ist eine symmetrische Matrix $\boldsymbol{A} \in \mathbf{S}^{n,n}$. *Gesucht* sind Eigenwerte und gegebenenfalls zugehörige orthonormierte Eigenvektoren des Eigenwertproblems

$$\boldsymbol{A}\boldsymbol{x} = \lambda \boldsymbol{x}. \tag{4}$$

Beispiele für die Herkunft solcher Probleme sind

- Diskretisierung spezieller Typen von Eigenwertaufgaben bei Differential- und Integralgleichungen
- allgemeine symmetrische Eigenwertprobleme, die auf spezielle reduziert worden sind, siehe Abschnitt 14.2.
- statistische Fragestellungen.

Sonderfälle des Problems stellen dar

- die Berechnung der p größten oder kleinsten Eigenwerte
- die Berechnung der in einem vorgegebenen Intervall $[a, b]$ liegenden Eigenwerte und gegebenenfalls Berechnung der zugehörigen Eigenvektoren
- Aufgaben mit Matrizen spezieller Struktur wie Bandmatrizen o. ä.

C 2. *Das allgemeine symmetrische Eigenwertproblem*

Gegeben sind die symmetrischen Matrizen $\boldsymbol{A}, \boldsymbol{B} \in \mathbf{S}^{n,n}$, wobei $\boldsymbol{B}$ positiv definit sei. *Gesucht* sind gewisse Eigenwerte und gegebenenfalls zugehörige Eigenvektoren des Matrizenpaares $\{\boldsymbol{A}, \boldsymbol{B}\}$ als Lösungen des sog. *allgemeinen symmetrischen Eigenwertproblems*

$$\boldsymbol{A}\boldsymbol{x} = \lambda \boldsymbol{B}\boldsymbol{x}. \tag{5}$$

Das wichtigste *Beispiel* für das Auftreten allgemeiner symmetrischer Eigenwertprobleme ist die Untersuchung des Eigenschwingungsverhaltens von mechanischen oder elektrischen Systemen, siehe Ü 2.1.6.

C 3. *Das spezielle nichtsymmetrische Eigenwertproblem*

Gegeben ist die nichtsymmetrische Matrix $\boldsymbol{A} \in \mathbf{R}^{n,n}$. *Gesucht* sind gewisse Eigenwerte und gegebenenfalls zugehörige Eigenvektoren von $\boldsymbol{A}$ als Lösungen von

$$\boldsymbol{A}\boldsymbol{x} = \lambda \boldsymbol{x}.$$

Beispiele für das Auftreten solcher Probleme sind

- Diskretisierung von nicht-selbstadjungierten Eigenwertproblemen bei Differentialgleichungen
- Bestimmung der allgemeinen Lösung von linearen Differentialgleichungssystemen $\dot{\boldsymbol{y}} = \boldsymbol{A}\boldsymbol{y}$, siehe Ü 2.1.5.

D. *Numerische Probleme*

Alle bisher beschriebenen Aufgabenklassen haben die folgende Struktur: Jede einzelne Aufgabe der Klasse ist durch eine endliche Zahl von *Eingangsdaten* mit gewissen Eigenschaften charakterisiert, denen durch die mathematische Aufgabenstellung in eindeutiger Weise eine endliche Zahl von *Ausgangsdaten* als Ergebnis bzw. Lösung zugeordnet wird.

2.1.3. *Beispiel. Eingangs- und Ausgangsdaten für ausgewählte Aufgabenklassen*

(i) Lösung beliebig regulärer Gleichungssysteme $\boldsymbol{Ax} = \boldsymbol{b}$ der Ordnung n.
Eingangsdaten: a_{ij}, b_i $(i, j = 1, \ldots, n)$, wobei $\boldsymbol{A}$ regulär ist
Ausgangsdaten: x_i $(i = 1, \ldots, n)$

(ii) Lösung regulärer Gleichungssysteme $\boldsymbol{Ax} = \boldsymbol{b}$ mit Tridiagonalmatrizen der Ordnung n.
Eingangsdaten: a_{ij}, b_i $(i, j = 1, \ldots, n, |i - j| \leqq 1)$, wobei $\boldsymbol{A}$ regulär ist
Ausgangsdaten: x_i $(i = 1, \ldots, n)$

(iii) Bestimmung der drei kleinsten Eigenwerte reeller symmetrischer Matrizen $\boldsymbol{A}$ der Ordnung n.
Eingangsdaten: a_{ij} $(i, j = 1, \ldots, n, i \leqq j)$
Ausgangsdaten: $\lambda_{n-2}, \lambda_{n-1}, \lambda_n$ □

Aufgaben des angeführten Typs werden *numerische Problemklassen* genannt. Eine solche numerische Problemklasse $\{\mathscr{E}, \boldsymbol{P}\}$ ist also eine Abbildung

$$\boldsymbol{P}: \mathscr{E} \subset \mathbf{R}^N \to \mathbf{R}^M, \tag{6}$$

die jedem N-dimensionalen Satz $\boldsymbol{e} = (e_1, \ldots, e_N)^T \in \mathscr{E}$ von Eingangsdaten aus dem zur Klasse gehörenden Definitionsbereich $\mathscr{E}$ in eindeutiger Weise einen M-dimensionalen Satz $\boldsymbol{a} = (a_1, \ldots, a_M)^T$ von Ausgangsdaten zugeordnet. Dabei kann $\boldsymbol{P}$ wie im Fall des regulären Gleichungssystems explizit angebbar sein — nämlich als $\boldsymbol{x} = \boldsymbol{A}^{-1}\boldsymbol{b}$ — bzw. wie im Fall der Eigenwertberechnung nur implizit definiert sein. Die Bestimmung von $\boldsymbol{a} = \boldsymbol{P}(\boldsymbol{e})$ für ein einzelnes, festes $\boldsymbol{e} \in \mathscr{E}$ heißt *numerisches Problem* $\{\boldsymbol{e}, \boldsymbol{P}\}$ aus der Klasse $\{\mathscr{E}, \boldsymbol{P}\}$. So ist die Lösung eines konkreten regulären Gleichungssystems $\boldsymbol{Ax} = \boldsymbol{b}$ für festes $\boldsymbol{A}, \boldsymbol{b}$ ein numerisches Problem aus der Klasse „Lösung regulärer Gleichungssysteme".

2.1.4. *Merksatz.* Numerische Probleme sind dadurch gekennzeichnet, daß die Eingangsdaten als fehlerbehaftet angesehen werden müssen.

Dabei sind drei Fehlerquellen zu unterscheiden, die in der Regel gleichzeitig auftreten:

— Die Eingangsdaten sind Ergebnisse von Messungen bzw. von vorausgegangenen Rechnungen und deshalb notwendig fehlerbehaftet

— Wenn $\boldsymbol{a} = \boldsymbol{P}(\boldsymbol{e})$ durch einen numerischen Algorithmus auf einem Computer berechnet werden soll, müssen die Eingangsdaten e_j als Computerzahlen im Speicher dargestellt werden. Selbst bei fehlerfreien Eingangsdaten wird dabei i. allg. ein — wenn auch kleiner — Darstellungsfehler hervorgerufen, vgl. Abschnitte 2.2 und 2.3

— Für viele Algorithmen läßt sich zeigen, daß die berechnete und durch Rundungsfehler verfälschte Computernäherung $\tilde{\boldsymbol{a}}$ für $\boldsymbol{a} = \boldsymbol{P}(\boldsymbol{e})$ die exakte Lösung eines benachbarten Problems mit den gestörten Eingangsdaten $\boldsymbol{e} + \delta\boldsymbol{e}$ ist, d. h., es gilt $\tilde{\boldsymbol{a}} = \boldsymbol{P}(\boldsymbol{e} + \delta\boldsymbol{e})$, wobei für die Störung $\delta\boldsymbol{e}$ eine kleine Schranke $\Delta\boldsymbol{e}$ mit $\|\delta\boldsymbol{e}\| \leqq \Delta\boldsymbol{e}$ angegeben werden kann. Das ist das Prinzip der rückwärtigen Fehleranalyse, vgl. Abschnitt 2.3. Die Berechnung und Abschätzung von $\delta\boldsymbol{e}$ ist oft wesentlich einfacher als die direkte Abschätzung von $\tilde{\boldsymbol{a}} - \boldsymbol{a}$ und hat sich speziell in der numerischen linearen Algebra als äußerst tragfähig erwiesen.

Unabhängig von der Art des Eingangsfehlers liegt dabei folgende Situation vor: Die Eingangsdaten $\boldsymbol{e}$ des konkreten Problems $\{\boldsymbol{e}, \boldsymbol{P}\}$ sind nur im Rahmen eines gewissen *Fehlerniveaus* bekannt, das hier der Einfachheit halber durch die kollektive Normschranke $\Delta\boldsymbol{e} > 0$ charakterisiert werden soll: Neben $\boldsymbol{e}$ sind alle Eingangsdaten $\tilde{\boldsymbol{e}}$ mit einem Fehler — genauer: *absolutem Fehler* —

$$\delta\boldsymbol{e} := \tilde{\boldsymbol{e}} - \boldsymbol{e}, \tag{7}$$

für den

$$\|\delta\boldsymbol{e}\| = \|\tilde{\boldsymbol{e}} - \boldsymbol{e}\| \leqq \Delta\boldsymbol{e} \tag{8}$$

gilt, als gleichberechtigt anzusehen. Selbstverständlich können auch individuelle Fehlerschranken des Typs $|\delta e_i| = |\tilde{e}_i - e_i| \leqq \Delta e_i$ verwendet werden.

Das fehlerbehaftete Problem $\{\boldsymbol{e}, \boldsymbol{P}\}$ mit dem Fehlerniveau $\Delta\boldsymbol{e}$ ist daher durch die Menge

$$\mathcal{E}(\boldsymbol{e}, \Delta\boldsymbol{e}) := \{\tilde{\boldsymbol{e}} \in \mathcal{E} \colon \|\tilde{\boldsymbol{e}} - \boldsymbol{e}\| \leqq \Delta\boldsymbol{e}\} \tag{9}$$

von gleichberechtigten Eingangsdaten charakterisiert. Als Lösung muß dann neben $\boldsymbol{a} = \boldsymbol{P}(\boldsymbol{e})$ jedes $\tilde{\boldsymbol{a}} = \boldsymbol{P}(\tilde{\boldsymbol{e}})$ mit $\tilde{\boldsymbol{e}} \in \mathcal{E}(\boldsymbol{e}, \Delta\boldsymbol{e})$ akzeptiert werden, d. h., alle Elemente der Menge

$$\mathcal{A}(\boldsymbol{e}, \Delta\boldsymbol{e}) := \boldsymbol{P}\big(\mathcal{E}(\boldsymbol{e}, \Delta\boldsymbol{e})\big) := \{\tilde{\boldsymbol{a}} = \boldsymbol{P}(\tilde{\boldsymbol{e}}) \colon \tilde{\boldsymbol{e}} \in \mathcal{E}(\boldsymbol{e}, \Delta\boldsymbol{e})\} \tag{10}$$

müssen als gleichberechtigte Lösungen angesehen werden. *Die Lösung des fehlerbehafteten Problems $\{\boldsymbol{e}, \boldsymbol{P}\}$ mit dem Fehlerniveau $\Delta\boldsymbol{e}$ besteht daher* eigentlich *in der Bestimmung der gesamten zulässigen Lösungsmenge $\mathcal{A}(\boldsymbol{e}, \Delta\boldsymbol{e})$.*

Damit das fehlerbehaftete numerische Problem vernünftig gestellt ist, muß $\mathcal{A}(\boldsymbol{e}, \Delta\boldsymbol{e})$ zumindest beschränkt sein, d. h., die Norm der Elemente von $\mathcal{A}$ darf eine obere Schranke nicht überschreiten. Als Beispiel betrachten wir eine fehlerbehaftete reguläre Matrix $\boldsymbol{M} \in \mathbf{R}^{n,n}$ mit dem Fehlerniveau $\Delta\boldsymbol{M}$. Bei zu großem $\Delta\boldsymbol{M}$ liegen in der Menge $\{\widetilde{\boldsymbol{M}} \colon \|\widetilde{\boldsymbol{M}} - \boldsymbol{M}\| \leqq \Delta\boldsymbol{M}\}$ auch singuläre Matrizen. Das Problem der Inversion von $\boldsymbol{M}$ ist dann zum Fehlerniveau $\Delta\boldsymbol{M}$ nicht vernünftig gestellt, denn in $\mathcal{E}(\boldsymbol{M}, \Delta\boldsymbol{M}) = \{\widetilde{\boldsymbol{M}} \colon \widetilde{\boldsymbol{M}}$ regulär und $\|\widetilde{\boldsymbol{M}} - \boldsymbol{M}\| \leqq \Delta\boldsymbol{M}\}$ gibt es Matrizen $\widetilde{\boldsymbol{M}}$ mit beliebig großem $\|\widetilde{\boldsymbol{M}}^{-1}\|$, d. h., $\mathcal{A}(\boldsymbol{M}, \Delta\boldsymbol{M})$ ist nicht beschränkt; siehe Abschnitt 4.1. In diesem Fall kann durch Verkleinerung von $\Delta\boldsymbol{M}$ — also genauere Bestimmung der Elemente von $\boldsymbol{M}$ — versucht werden, beschränkte Lösungsmengen zu erhalten. Im Abschnitt 4.1 werden wir sehen, daß dies für das diskutierte Beispiel der Inversion in der Tat möglich ist. Andernfalls muß die Problemstellung z. B. durch Regularisierung o. ä. modifiziert werden, vgl. Abschnitte 11.2 und 11.3. Ob dies sinnvoll und zulässig ist,

kann nicht allein vom Standpunkt der Mathematik, sondern nur vom Standpunkt des im Hintergrund stehenden realen Problems unter Einbeziehung des Naturwissenschaftlers oder Ingenieurs entschieden werden.

Es stellt sich heraus: *Die exakte Bestimmung der Lösungsmenge* $\mathcal{A}(\boldsymbol{e}, \Delta\boldsymbol{e})$ *ist für die Aufgabenklassen der linearen Algebra praktisch nicht möglich.* Man behilft sich daher mit der Konstruktion von Mengen, die $\mathcal{A}(\boldsymbol{e}, \Delta\boldsymbol{e})$ einschließen. Eine Möglichkeit dafür bildet die Intervallanalysis bzw. Intervallmathematik, siehe B 2.8. Wir gehen einen anderen, in der Regel weitaus billigeren Weg und bestimmen neben dem Repräsentanten $\boldsymbol{a} = \boldsymbol{P}(\boldsymbol{e})$ eine Fehlerschranke $\Delta\boldsymbol{a}$ so, daß

$$\|\delta\boldsymbol{a}\| = \|\boldsymbol{P}(\tilde{\boldsymbol{e}}) - \boldsymbol{P}(\boldsymbol{e})\| \leqq \Delta\boldsymbol{a} \quad \text{für alle } \tilde{\boldsymbol{e}} \in \mathcal{E}(\boldsymbol{e}, \Delta\boldsymbol{e}) \tag{11}$$

gilt. Dies ist gleichbedeutend mit

$$\mathcal{A}(\boldsymbol{e}, \Delta\boldsymbol{e}) \subset \mathcal{S}(\boldsymbol{a}, \Delta\boldsymbol{a}) := \{\tilde{\boldsymbol{a}} \in \mathbf{R}^M : \|\tilde{\boldsymbol{a}} - \boldsymbol{a}\| \leqq \Delta\boldsymbol{a}\}, \tag{12}$$

d. h., die Lösungsmenge wird mit einer Normkugel um $\boldsymbol{a}$ mit dem Radius $\Delta\boldsymbol{a}$ eingeschlossen, siehe Abb. 2.1.1 für eine Veranschaulichung unter Verwendung der Euklidischen Norm.

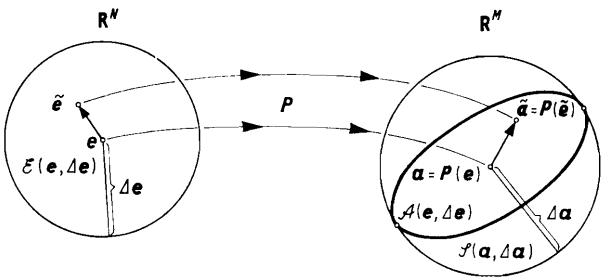


Abb. 2.1.1. Lösungsmenge und Fehlerschranken bei fehlerbehafteten Eingangsdaten

Der kleinste in (11) mögliche Wert für $\Delta\boldsymbol{a}$ wird mit $\Delta\boldsymbol{a}_{\text{opt}}$ bezeichnet und gibt das *optimale* (absolute) *Fehlerniveau* der Ausgangsdaten an:

$$\Delta\boldsymbol{a}_{\text{opt}} := \min\{\Delta\boldsymbol{a} > 0 : \|\boldsymbol{P}(\tilde{\boldsymbol{e}}) - \boldsymbol{P}(\boldsymbol{e})\| \leqq \Delta\boldsymbol{a} \text{ für alle } \tilde{\boldsymbol{e}} \in \mathcal{E}(\boldsymbol{e}, \Delta\boldsymbol{e})\}. \tag{13}$$

Eine äquivalente Festlegung ist

$$\Delta\boldsymbol{a}_{\text{opt}} = \max\{\|\boldsymbol{P}(\tilde{\boldsymbol{e}}) - \boldsymbol{P}(\boldsymbol{e})\| : \tilde{\boldsymbol{e}} \in \mathcal{E}(\boldsymbol{e}, \Delta\boldsymbol{e})\}. \tag{14}$$

Wir bemerken hier, daß „min" bzw. „max" eigentlich durch „inf" bzw. „sup" ersetzt werden müßten. Bei abgeschlossenem $\mathcal{E}(\boldsymbol{e}, \Delta\boldsymbol{e})$ und stetigem $\boldsymbol{P}$ — dies ist bei unseren Beispielen der Fall — sind jedoch die einfachen Begriffe „min" bzw. „max" zulässig.

Aus der Definition von $\Delta\boldsymbol{a}_{\text{opt}}$ folgt: *Bei einem Fehlerniveau* $\Delta\boldsymbol{e}$ *der Eingangsdaten ist der Fehler der Ausgangsdaten durch* $\Delta\boldsymbol{a}_{\text{opt}}$ *beschränkt, und es gibt zulässige Eingangsdaten* $\tilde{\boldsymbol{e}} \in \mathcal{E}(\boldsymbol{e}, \Delta\boldsymbol{a})$ *mit* $\|\boldsymbol{P}(\tilde{\boldsymbol{e}}) - \boldsymbol{P}(\boldsymbol{e})\| = \Delta\boldsymbol{a}_{\text{opt}}$, *d. h., das optimale Fehlerniveau*

wird angenommen. Man nennt $\Delta \boldsymbol{a}_{\text{opt}}$ aus diesem Grunde auch *unvermeidlichen Fehler* der Ausgangsdaten. *Eine höhere Genauigkeit ist bei der in den Eingangsdaten* $\mathcal{E}(\boldsymbol{e}, \Delta \boldsymbol{e})$ *liegenden Information nicht zu erwarten*, siehe Abb. 2.1.1, wo mit $\Delta \boldsymbol{a} = \Delta \boldsymbol{a}_{\text{opt}}$ gearbeitet wird.

Für die Berechnung von $\Delta \boldsymbol{a}$ muß die Änderung $\delta \boldsymbol{a} := \tilde{\boldsymbol{a}} - \boldsymbol{a} = \boldsymbol{P}(\tilde{\boldsymbol{e}}) - \boldsymbol{P}(\boldsymbol{e})$ für kleine Störungen $\tilde{\boldsymbol{e}} - \boldsymbol{e}$ untersucht und abgeschätzt werden. Dies ist Gegenstand der *Störungstheorie*, die exakte oder zumindest in der Größenordnung richtige Ausdrücke für $\boldsymbol{P}(\tilde{\boldsymbol{e}}) - \boldsymbol{P}(\boldsymbol{e})$ als Funktion von $\tilde{\boldsymbol{e}} - \boldsymbol{e}$ bzw. für deren Normen bereitstellt.

Eine natürliche Forderung für eine vernünftige Problemstellung $\{\boldsymbol{e}, \boldsymbol{P}\}$ ist die *Stetigkeit* von $\boldsymbol{P}$ im Punkt $\boldsymbol{e}$: Wenn $\tilde{\boldsymbol{e}}$ gegen $\boldsymbol{e}$ strebt, soll $\tilde{\boldsymbol{a}} = \boldsymbol{P}(\tilde{\boldsymbol{e}})$ gegen $\boldsymbol{a} = \boldsymbol{P}(\boldsymbol{e})$ streben; d. h., *kleine Änderungen der Eingangsdaten sollen kleine Änderungen der Ausgangsdaten bewirken.* Ein in diesem Sinne vernünftiges numerisches Problem wird als *korrekt gestellt* bezeichnet (engl. „properly posed", russ. „корректно поставлено"). Für numerische Zwecke ist die Eigenschaft „korrekt gestellt" zu schwach; die Stetigkeit im Punkt $\boldsymbol{e}$ muß relativ gutartig und quantitativ erfaßbar sein. Zur Motivierung betrachten wir das folgende Beispiel:

2.1.5. *Beispiel.* Stetigkeitsverhalten der Problemklasse

$$a = P(e) = 1/e, \qquad e \in \mathcal{E} := \{e \in \mathbf{R} : e > 0\}. \tag{15}$$

Für festes $e \in \mathcal{E}$ — also $e > 0$ — liegt auch jedes $\tilde{e}$ mit $|\tilde{e} - e| \leqq \Delta e$ in $\mathcal{E}$, sofern $\Delta e \leqq \Delta_0 e < e$ ist. In diesem Fall folgt

$$|P(\tilde{e}) - P(e)| = \left|\frac{1}{\tilde{e}} - \frac{1}{e}\right| = \frac{1}{e\tilde{e}}\,|\tilde{e} - e| \leqq \frac{1}{e(e - \Delta e)}\,|\tilde{e} - e|. \tag{16}$$

Die Änderung $\tilde{e} - e$ der Eingangsdaten kann sich also höchstens mit dem Verstärkungsfaktor $L(e, \Delta e) := 1/[e(e - \Delta e)]$ auf die Ausgangsdaten auswirken. Der Faktor $L(e, \Delta e)$ hängt von e und Δe ab, und für genügend kleines Δe unterscheidet er sich immer weniger von $L(e, 0) = 1/e^2$. □

Das obige eindimensionale Beispiel legt es nahe, die folgenden allgemeinen Begriffe einzuführen:

2.1.6. *Begriffe der Lipschitzstetigkeit.* Betrachtet wird die numerische Problemklasse $\{\mathcal{E}, \boldsymbol{P}\}$.

(i) Das *Problem* $\{\boldsymbol{e}, \boldsymbol{P}\}$ der Klasse $\{\mathcal{E}, \boldsymbol{P}\}$ heißt *lokal lipschitzstetig*, wenn es ein (i. allg. von $\boldsymbol{e}$ abhängiges) Fehlerniveau $\Delta_0 \boldsymbol{e} > 0$ gibt, so daß zu jedem $\Delta \boldsymbol{e} \leqq \Delta_0 \boldsymbol{e}$ eine *lokale Lipschitzkonstante* $L(\boldsymbol{e}, \Delta \boldsymbol{e}) > 0$ existiert mit

$$\|\boldsymbol{P}(\tilde{\boldsymbol{e}}) - \boldsymbol{P}(\boldsymbol{e})\| \leqq L(\boldsymbol{e}, \Delta \boldsymbol{e})\,\|\tilde{\boldsymbol{e}} - \boldsymbol{e}\| \tag{17}$$

für alle $\tilde{\boldsymbol{e}} \in \mathcal{E}$ mit $\|\tilde{\boldsymbol{e}} - \boldsymbol{e}\| \leqq \Delta \boldsymbol{e}$.

(ii) Die *Klasse* $\{\mathcal{E}, \boldsymbol{P}\}$ heißt *lokal lipschitzstetig*, wenn jedes Problem $\{\boldsymbol{e}, \boldsymbol{P}\}$, $\boldsymbol{e} \in \mathcal{E}$, der Klasse lokal lipschitzstetig ist.

(iii) Die *Klasse* $\{\mathcal{E}, \boldsymbol{P}\}$ heißt (schlechthin) *lipschitzstetig*, wenn eine (globale) *Lipschitzkonstante* $L > 0$ mit

$$\|\boldsymbol{P}(\tilde{\boldsymbol{e}}) - \boldsymbol{P}(\boldsymbol{e})\| \leqq L \, \|\tilde{\boldsymbol{e}} - \boldsymbol{e}\| \tag{18}$$

für alle $\tilde{\boldsymbol{e}}, \boldsymbol{e} \in \mathcal{E}$ existiert.

Dabei wird stets vorausgesetzt, daß die Lipschitzkonstanten minimal gewählt werden, so daß die Abschätzungen (17) bzw. (18) nicht verbessert werden können.

Wir bemerken noch, daß die Eigenschaft der Lipschitzstetigkeit unabhängig von den verwendeten Normen ist, während die Zahlenwerte der Lipschitzkonstanten natürlich von den Normen abhängen.

Sämtliche der oben eingeführten Stetigkeitsbegriffe treten bei den später behandelten Problemklassen tatsächlich auf. Als Beispiele führen wir an:

— Die Bestimmung der Normallösung eines Quadratmittelproblems $\boldsymbol{Ax} \cong \boldsymbol{b}$ ist ein lokal lipschitzstetiges Problem, falls $\boldsymbol{A}$ Vollrang hat. Andernfalls liegt nicht einmal Stetigkeit vor, so daß die Problemklasse, für beliebige Eingangsdaten $\{\boldsymbol{A}, \boldsymbol{b}\}$ die Normallösung zu bestimmen, nicht lokal lipschitzstetig ist.

— Die Lösung regulärer linearer Gleichungssysteme $\boldsymbol{Ax} = \boldsymbol{b}$ ist eine lokal lipschitzstetige Problemklasse.

— Die Eigenwertbestimmung symmetrischer Matrizen $\boldsymbol{A}$ ist eine schlechthin lipschitzstetige Problemklasse mit der Lipschitzkonstanten $L = 1$, sofern $\tilde{\boldsymbol{A}} - \boldsymbol{A}$ in der F-Norm gemessen wird.

2.1.7. *Bemerkung* (i) Die lokale Lipschitzstetigkeit ist eine starke Form der Stetigkeit, deren Vorliegen keineswegs selbstverständlich ist. Zur Illustration betrachten wir die Problemklasse

$$a = P(e) = e^{\alpha}, \qquad e \in \mathcal{E} := \{e \in \mathbf{R} : e \geqq 0\}, \qquad 0 < \alpha \leqq 1 \text{ fest}.$$

Für das durch $e = 0$ festgelegte Problem und beliebiges $\tilde{e} > 0$ gilt

$$|P(\tilde{e}) - P(e)| = |e^{\alpha}| = 1 \cdot |\tilde{e} - e|^{\alpha},$$

d. h., die Änderung $\tilde{e} - e$ wirkt sich nicht in erster, sondern in der Potenz α auf die Ausgangsdaten aus. Man spricht dann von Hölderstetigkeit. Etwa zur Definition der lokalen Hölderstetigkeit müßte (17) durch

$$\|P(\tilde{e}) - P(e)\| \leqq L(e, \Delta e) \, \|\tilde{e} - e\|^{\alpha(e)} \tag{19}$$

ersetzt werden, wobei $\alpha(e)$ den größtmöglichen Exponenten bezeichnet, für den (19) gilt; die übrigen Begriffe wären sinngemäß zu übertragen. Je kleiner α ist, um so schwächer ist diese Form der Stetigkeit: Im Beispiel ergibt sich für $e = 0$, $\tilde{e} - e = 10^{-6}$ etwa im Fall $\alpha = 1/2$ die Änderung $\delta a = P(\tilde{e}) - P(e) = 10^{-3}$, im Fall $\alpha = 1/12$ dagegen $\delta a = 10^{-1/2} = 0.3$! Man beachte, daß für $e > 0$ lokale Lipschitzstetigkeit vorliegt, also α in der Tat von e abhängen kann.

(ii) Es gibt auch in der linearen Algebra Probleme, die nur lokal hölderstetig mit $\alpha < 1$ sind. So können die Eigenwerte nichtsymmetrischer Matrizen $\boldsymbol{A}$ lokal hölder-

stetig mit $\alpha \geqq 1/n$ von $\boldsymbol{A}$ abhängen, und der Wert $\alpha = 1/n$ kann angenommen werden: Die Matrix

$$\boldsymbol{A}(\varepsilon) := \begin{bmatrix} 1 & 1 & & & & \\ & 1 & 1 & & & \\ \cdot & & \cdot & \cdot & & \\ \cdot & & & \cdot & \cdot & \\ \cdot & & & & \cdot & \cdot \\ \varepsilon & \cdot & \cdot & \cdot & \cdot & 1 & 1 \end{bmatrix} \in \mathbf{R}^{n,n}$$

hat das charakteristische Polynom $\det(\boldsymbol{A}(\varepsilon) - \lambda \boldsymbol{I}) = (1-\lambda)^n + (-1)^{n+1}\varepsilon$, so daß für die Eigenwerte $\lambda_j(\varepsilon)$ die Beziehung

$$|\lambda_j(\varepsilon) - \lambda_j(0)| = \sqrt[n]{\varepsilon} = \varepsilon^{1/n}$$

folgt; man beachte $\lambda_j(0) = 1$ $(j = 1, \ldots, n)$. Eine Störung $\varepsilon = 10^{-6}$ bewirkt also im Fall $n = 12$ eine Änderung der Eigenwerte um $|\delta\lambda_j| = 10^{-1/2} = 0.3$. Da wir das nichtsymmetrische Eigenwertproblem nicht behandeln, gehen wir auf die Hölderstetigkeit nicht weiter ein.

(iii) Falls $\boldsymbol{P}$ in der Umgebung von $\boldsymbol{e} \in \mathcal{E}$ stetige partielle Ableitungen erster Ordnung besitzt, ist das Problem $\{\boldsymbol{e}, \boldsymbol{P}\}$ lokal lipschitzstetig, vgl. Ü 2.1.7. □

Mittels der lokalen Lipschitzkonstanten läßt sich das optimale Fehlerniveau einfach abschätzen bzw. unter zusätzlichen Glattheitsvoraussetzungen sogar asymptotisch exakt angeben.

2.1.8. *Aussage.* Betrachtet wird das numerische Problem $\{\boldsymbol{e}, \boldsymbol{P}\}$.

(i) Wenn $\{\boldsymbol{e}, \boldsymbol{P}\}$ lokal lipschitzstetig ist, gilt für $\Delta\boldsymbol{e} \leqq \Delta_0\boldsymbol{e}$

$$\Delta\boldsymbol{a}_{\text{opt}} \leqq L(\boldsymbol{e}, \Delta\boldsymbol{e})\, \Delta\boldsymbol{e}. \tag{20}$$

(ii) Wenn P in der Umgebung von $\boldsymbol{e}$ genügend glatt im Sinne der Existenz stetiger partieller Ableitungen zweiter Ordnung ist, liegt lokale Lipschitzstetigkeit vor, die lokale Lipschitzkonstante $L(\boldsymbol{e}, \Delta\boldsymbol{e})$ hängt stetig von $\Delta\boldsymbol{e} \geqq 0$ ab, der Grenzwert

$$L(\boldsymbol{e}) := \lim_{\Delta\boldsymbol{e} \to +0} L(\boldsymbol{e}, \Delta\boldsymbol{e}) \tag{21}$$

existiert, und es gilt $L(\boldsymbol{e}, \Delta\boldsymbol{e}) = L(\boldsymbol{e}) + O(\Delta\boldsymbol{e})$ sowie

$$\Delta\boldsymbol{a}_{\text{opt}} = [L(\boldsymbol{e}) + O(\Delta\boldsymbol{e})]\, \Delta\boldsymbol{e} \tag{22}$$

für $\Delta\boldsymbol{e} \to 0$, $\Delta\boldsymbol{e} \leqq \Delta_0\boldsymbol{e}$ mit genügend kleinem $\Delta_0\boldsymbol{e} > 0$.

Für einen festen Exponenten $\beta > 0$ bezeichnet $O(\xi^\beta)$ dabei eine Funktion von $\xi \in \mathbf{R}$, die für $\xi \to 0$ mindestens so schnell wie ξ^β gegen 0 geht, d. h., es gilt

$$|O(\xi^\beta)| \leqq C\, |\xi|^\beta$$

für genügend kleines $|\xi|$ mit einer Konstanten $C > 0$.

Beweis. Aussage (i) folgt sofort aus (13) und (17). Teil (ii) ergibt sich aus der Taylorformel mit Restglied zweiter Ordnung für $\boldsymbol{P}$ an der Stelle $\boldsymbol{e}$, vgl. wieder Ü 2.1.7. □

Bei den später behandelten Problemen ist die Glattheitsforderung aus (ii) immer erfüllt. *Wir nehmen daher im folgenden stets an, daß* $\boldsymbol{P}$ *dieser Voraussetzung genügt und somit die Darstellung* (22) *für das optimale Fehlerniveau gültig ist.* Aus dieser folgt

$$\Delta \boldsymbol{a}_{\text{opt}} \approx L(\boldsymbol{e})\, \Delta \boldsymbol{e} \tag{23}$$

für genügend kleines $\Delta \boldsymbol{e}$, und (23) ist asymptotisch exakt im Sinne von

$$\frac{\Delta \boldsymbol{a}_{\text{opt}}}{\Delta \boldsymbol{e}} \to L(\boldsymbol{e}) \quad \text{für} \quad \Delta \boldsymbol{e} \to 0. \tag{24}$$

Die Zahl $L(\boldsymbol{e})$ wird *absolute Konditionszahl* des lokal lipschitzstetigen Problems $\{\boldsymbol{e}, \boldsymbol{P}\}$ genannt. *Sie gibt den maximal möglichen Verstärkungsfaktor an, mit dem sich eine kleine Störung der Eingangsdaten auf die Ausgangsdaten auswirken kann.*

In den meisten Fällen ist es zweckmäßig, mit relativen Fehlern und Fehlerschranken anstelle von absoluten zu arbeiten. Für eine skalare Größe $e_i \in \mathbf{R}$ heißt unter der Voraussetzung $e_i \neq 0$ der Quotient

$$\varepsilon_i := \varepsilon(e_i) := \frac{\tilde{e}_i - e_i}{e_i} = \frac{\delta e_i}{e_i} \tag{25}$$

relativer Fehler von e_i und wird üblicherweise in Prozent angegeben; gelegentlich wird auch mit $(\tilde{e}_i - e_i)/\tilde{e}_i$ gearbeitet, siehe B 2.1 und Ü 2.1.8. Durch ε_i kann die im Repräsentanten e_i enthaltene Information unabhängig von dessen Größenordnung und damit besser charakterisiert werden; z. B. läßt sich $-\log_{10} |\varepsilon_i|$ als Anzahl der gültigen Dezimalziffern in der Dezimaldarstellung von e_i deuten, vgl. Abschnitt 2.2. Bei Vektoren $\boldsymbol{e} = (e_i) \in \mathbf{R}^N$ können die individuellen relativen Fehler der Komponenten für jedes i gemäß (25) definiert werden. Häufig ist es jedoch einfacher und zweckmäßiger, mit kollektiven relativen Fehlern

$$\frac{\|\tilde{\boldsymbol{e}} - \boldsymbol{e}\|}{\|\boldsymbol{e}\|} \quad \text{bzw.} \quad \frac{\|\boldsymbol{P}(\tilde{\boldsymbol{e}}) - \boldsymbol{P}(\boldsymbol{e})\|}{\|\boldsymbol{P}(\boldsymbol{e})\|}$$

zu arbeiten. Dazu muß natürlich $\boldsymbol{e} \neq \boldsymbol{o}$ bzw. $\boldsymbol{P}(\boldsymbol{e}) \neq 0$ vorausgesetzt werden.

Wenn $\{\boldsymbol{e}, \boldsymbol{P}\}$ lokal lipschitzstetig ist, läßt sich (17) mit

$$K(\boldsymbol{e}, \Delta \boldsymbol{e}) := L(\boldsymbol{e}, \Delta \boldsymbol{e}) \frac{\|\boldsymbol{e}\|}{\|\boldsymbol{P}(\boldsymbol{e})\|}$$

äquivalent in der Form

$$\frac{\|\delta \boldsymbol{a}\|}{\|\boldsymbol{a}\|} = \frac{\|\boldsymbol{P}(\tilde{\boldsymbol{e}}) - \boldsymbol{P}(\boldsymbol{e})\|}{\|\boldsymbol{P}(\boldsymbol{e})\|} \leqq K(\boldsymbol{e}, \Delta \boldsymbol{e}) \frac{\|\tilde{\boldsymbol{e}} - \boldsymbol{e}\|}{\|\boldsymbol{e}\|}$$

schreiben, d. h., die relativen Fehler genügen einer zu (17) analogen Ungleichung mit der relativen lokalen Lipschitzkonstanten $K(\boldsymbol{e}, \Delta \boldsymbol{e})$ statt $L(\boldsymbol{e}, \Delta \boldsymbol{e})$. Mit $L(\boldsymbol{e}, \Delta \boldsymbol{e})$ ist auch $K(\boldsymbol{e}, \Delta \boldsymbol{e})$ optimal, d. h. nicht durch eine kleinere Konstante ersetzbar. Die

2.1.8(i) entsprechende Abschätzung des relativen optimalen Fehlerniveaus lautet dann

$$\frac{\Delta \boldsymbol{a}_{\text{opt}}}{\|\boldsymbol{a}\|} \leqq K(\boldsymbol{e}, \Delta \boldsymbol{e}) \frac{\Delta \boldsymbol{e}}{\|\boldsymbol{e}\|} \quad \text{für} \quad \frac{\Delta \boldsymbol{e}}{\|\boldsymbol{e}\|} \leqq \frac{\Delta_0 \boldsymbol{e}}{\|\boldsymbol{e}\|}, \tag{26}$$

und analog zu 2.1.8(ii) ist $K(\boldsymbol{e}, \Delta \boldsymbol{e})$ für genügend glattes $\boldsymbol{P}$ bezüglich $\Delta \boldsymbol{e}$ stetig, der Grenzwert

$$K(\boldsymbol{e}) := \lim_{\Delta \boldsymbol{e} \to +0} K(\boldsymbol{e}, \Delta \boldsymbol{e}) = L(\boldsymbol{e}) \, \|\boldsymbol{e}\| / \|\boldsymbol{P}(\boldsymbol{e})\| \tag{27}$$

existiert, und es gilt $K(\boldsymbol{e}, \Delta \boldsymbol{e}) = K(\boldsymbol{e}) + 0(\Delta \boldsymbol{e}/\|\boldsymbol{e}\|)$ sowie

$$\frac{\Delta \boldsymbol{a}_{\text{opt}}}{\|\boldsymbol{a}\|} = \left[K(\boldsymbol{e}) + 0\left(\frac{\Delta \boldsymbol{e}}{\|\boldsymbol{e}\|} \right) \right] \frac{\Delta \boldsymbol{e}}{\|\boldsymbol{e}\|}$$

für $\Delta \boldsymbol{e}/\|\boldsymbol{e}\| \to 0$, $\Delta \boldsymbol{e}/\|\boldsymbol{e}\| \leqq \Delta_0 \boldsymbol{e}/\|\boldsymbol{e}\|$, $\Delta_0 \boldsymbol{e}$ genügend klein. Dies bedeutet

$$\frac{\Delta \boldsymbol{a}_{\text{opt}}}{\|\boldsymbol{a}\|} \approx K(\boldsymbol{e}) \frac{\Delta \boldsymbol{e}}{\|\boldsymbol{e}\|} \tag{28}$$

für genügend kleines $\Delta \boldsymbol{e}/\|\boldsymbol{e}\|$, und (28) ist im Sinne von

$$\frac{\Delta \boldsymbol{a}_{\text{opt}}/\|\boldsymbol{a}\|}{\Delta \boldsymbol{e}/\|\boldsymbol{e}\|} \to K(\boldsymbol{e}) \quad \text{für} \quad \frac{\Delta \boldsymbol{e}}{\|\boldsymbol{e}\|} \to 0 \tag{29}$$

asymptotisch exakt. *Das optimale relative Fehlerniveau der Ausgangsdaten ist also das $K(\boldsymbol{e})$-fache des relativen Fehlerniveaus der Eingangsdaten,* sofern letzteres genügend klein ist. Der relative Fehler der Eingangsdaten kann sich höchstens mit dem Verstärkungsfaktor $K(\boldsymbol{e})$ als relativer Fehler der Ausgangsdaten auswirken. Die Zahl $K(\boldsymbol{e})$ wird deshalb *relative Konditionszahl* des Problems $\{\boldsymbol{e}, \boldsymbol{P}\}$ genannt.

Die folgenden Beispiele sollen die neu eingeführten Begriffe erläutern und zeigen, wie Lipschitzkonstanten und Konditionszahlen in konkreten Fällen berechnet werden können.

2.1.9. *Beispiel. Bestimmung der Euklidischen Norm* $r(\boldsymbol{x}) := \|\boldsymbol{x}\|_2$ für $\boldsymbol{x} \in \mathbf{R}^n$. Wegen der Normeigenschaft (N_3') aus 1.1.I gilt

$$|\delta r| = |\, \|\boldsymbol{x} + \delta \boldsymbol{x}\|_2 - \|\boldsymbol{x}\|_2| \leqq \|(\boldsymbol{x} + \delta \boldsymbol{x}) - \boldsymbol{x}\|_2 = \|\delta \boldsymbol{x}\|_2 \tag{30}$$

für alle $\boldsymbol{x}, \delta \boldsymbol{x} \in \mathbf{R}^n$, d. h., die Aufgabenklasse ist lipschitzstetig mit der Konstanten $L = 1$. Aus (30) folgt

$$\left| \frac{\delta r}{r} \right| = \frac{|\, \|\boldsymbol{x} + \delta \boldsymbol{x}\|_2 - \|\boldsymbol{x}\|_2|}{\|\boldsymbol{x}\|_2} \leqq \frac{\|\delta \boldsymbol{x}\|_2}{\|\boldsymbol{x}\|_2} \tag{31}$$

für alle $\boldsymbol{x}, \delta \boldsymbol{x} \in \mathbf{R}^n$ mit $\boldsymbol{x} \neq \boldsymbol{o}$, d. h., auch die relativen Fehler können sich nicht verstärken. Da in (30) und (31) für $\delta \boldsymbol{x} = \lambda \boldsymbol{x}$, $\boldsymbol{x} \neq \boldsymbol{o}$, $\lambda \neq 0$ beliebig, das Gleichheitszeichen steht, folgt $L(\boldsymbol{x}) = 1$ für alle $\boldsymbol{x}$ und $K(\boldsymbol{x}) = 1$ für alle $\boldsymbol{x} \neq \boldsymbol{o}$.

Um den Einfluß von δx nicht nur in einer Normabschätzung, sondern direkt erfassen zu können, betrachten wir

$$\begin{aligned}\delta r &= \|x + \delta x\|_2 - \|x\|_2 = (\|x\|_2^2 + 2x^T\delta x + \|\delta x\|_2^2)^{1/2} - (\|x\|_2^2)^{1/2} \\ &= [(\|x\|_2^2 + 2x^T\delta x + \|\delta x\|_2^2) - (\|x\|_2^2)]/[(\|x\|_2^2 + 2x^T\delta x + \|\delta x\|_2^2)^{1/2} \\ &\quad + (\|x\|_2^2)^{1/2}].\end{aligned}$$

Berücksichtigung von Gliedern erster Ordnung in δx führt auf

$$\delta r = \|x + \delta x\|_2 - \|x\|_2 = \frac{x^T\delta x}{\|x\|_2} + O(\|\delta x\|^2) \quad \text{für} \quad x \neq o. \tag{32}$$

Die Beziehung $v(\varepsilon) = O(\varepsilon^p)$ für $\varepsilon \in \mathbf{R}$ bedeutet dabei $\|v(\varepsilon)\| \leqq C\,|\varepsilon|^p$ für genügend kleines $|\varepsilon|$ mit einer Konstanten $C > 0$. Die Darstellung (32) besagt u. a., daß sich eine zu x orthogonale Störung δx in erster Ordnung nicht auf $\|x + \delta x\|_2$ auswirkt, während der Einfluß für $\delta x = \lambda x$ maximal wird. Dies ist anschaulich evident.

Zahlenbeispiel. Für $n = 2$ und $x = (0.9987, 0.05097)^T$ ist $r(x) = \|x\|_2 = 1.000000$. Das zu $\Delta x = 0.01$ gehörende optimale Fehlerniveau ist $\Delta r_{\text{opt}} = \Delta x = 0.01$.

(i) Für $\delta x = (0.01, 0)^T$ ergibt sich $r(x + \delta x) = 1.009987$, also

$$\delta r = 0.009987 \quad \text{und} \quad |\delta r|/\Delta r_{\text{opt}} = 99.87\%.$$

(ii) Für $\delta x = (0, 0.01)^T$ ergibt sich $r(x + \delta x) = 1.000559$, also

$$\delta r = 0.000559 \quad \text{und} \quad |\delta r|/\Delta r_{\text{opt}} = 5.59\%.$$

Das Beispiel zeigt, daß ein Fehler δr in der Größenordnung Δr_{opt} nicht für jede im Rahmen des Fehlerniveaus $\Delta x = 0.01$ zulässige Störung δx angenommen wird. Da der „wahre" Fehler δx der Eingangsdaten i. allg. nicht bekannt ist, muß jedoch mit einem Fehler von maximal $\Delta r_{\text{opt}} = 0.01$ gerechnet werden. □

Falls die Eingangsdaten e aus verschiedenen Gruppen — etwa $e = (A, b)$ bei linearen Gleichungssystemen bzw. $e = (x, y)$ beim Skalarprodukt — bestehen, ist es zweckmäßig, den Einfluß jeder Gruppe getrennt zu berücksichtigen und additiv zu überlagern. Man kommt dann zu sog. *partiellen Konditionszahlen* bezüglich jeder Gruppe. Wir demonstrieren diese Verallgemeinerung der bisher eingeführten Begriffe am Beispiel des Skalarproduktes x^Ty.

2.1.10. *Beispiel. Bestimmung des Skalarproduktes* $s = s(x, y) = x^Ty$ für $x, y \in \mathbf{R}^n$. Hier gilt

$$\delta s = (x + \delta x)^T (y + \delta y)^T - x^Ty = \delta x^Ty + x^T\delta y + \delta x^T\delta y,$$

also

$$\begin{aligned}|\delta s| &\leqq \|\delta x\|_2 \|y\|_2 + \|\delta y\|_2 \|x\|_2 + \|\delta x\|_2 \|\delta y\|_2 \\ &= (\|y\|_2 + 0.5\,\|\delta y\|_2)\,\|\delta x\|_2 + (\|x\|_2 + 0.5\,\|\delta x\|_2)\,\|\delta x\|_2.\end{aligned} \tag{33}$$

Für beliebige $\Delta x, \Delta y > 0$ und $\|\delta x\|_2 \leqq \Delta x$, $\|\delta y\|_2 \leqq \Delta y$ folgt

$$|\delta s| \leqq (\|y\|_2 + 0.5\Delta y)\,\|\delta x\|_2 + (\|x\|_2 + 0.5\Delta x)\,\|\delta y\|_2, \tag{34}$$

d. h., die Problemklasse ist lokal lipschitzstetig mit den partiellen lokalen Lipschitzkonstanten $L_x(\boldsymbol{x}, \boldsymbol{y}, \Delta\boldsymbol{x}, \Delta\boldsymbol{y}) := \|\boldsymbol{y}\|_2 + 0.5\Delta\boldsymbol{y}$, $L_y(\boldsymbol{x}, \boldsymbol{y}, \Delta\boldsymbol{x}, \Delta\boldsymbol{y}) := \|\boldsymbol{x}\|_2 + 0.5\Delta\boldsymbol{x}$.

Für $\boldsymbol{x}^T\boldsymbol{y} \neq 0$ ergibt sich aus (33)

$$\frac{|\delta s|}{|s|} \leqq \frac{\|\boldsymbol{x}\|_2 \, \|\boldsymbol{y}\|_2}{|\boldsymbol{x}^T\boldsymbol{y}|} \left\{ \frac{\|\delta\boldsymbol{x}\|_2}{\|\boldsymbol{x}\|_2} + \frac{\|\delta\boldsymbol{y}\|_2}{\|\boldsymbol{y}\|_2} + \frac{\|\delta\boldsymbol{x}\|_2}{\|\boldsymbol{x}\|_2} \cdot \frac{\|\delta\boldsymbol{y}\|_2}{\|\boldsymbol{y}\|_2} \right\}.$$

Ist $\|\delta\boldsymbol{x}\|_2/\|\boldsymbol{x}\|_2 \leqq \Delta\boldsymbol{x}/\|\boldsymbol{x}\|_2$ und $\|\delta\boldsymbol{y}\|_2/\|\boldsymbol{y}\|_2 \leqq \Delta\boldsymbol{y}/\|\boldsymbol{y}\|_2$, so folgt

$$\frac{|\delta s|}{|s|} \leqq \frac{\|\boldsymbol{x}\|_2 \, \|\boldsymbol{y}\|_2}{|\boldsymbol{x}^T\boldsymbol{y}|} \left\{ \left(1 + 0.5 \, \frac{\Delta\boldsymbol{y}}{\|\boldsymbol{y}\|_2}\right) \frac{\|\delta\boldsymbol{x}\|_2}{\|\boldsymbol{x}\|_2} + \left(1 + 0.5 \, \frac{\Delta\boldsymbol{x}}{\|\boldsymbol{x}\|_2}\right) \frac{\|\delta\boldsymbol{y}\|_2}{\|\boldsymbol{y}\|_2} \right\}. \tag{35}$$

Aus (34) bzw. (35) liest man für $\Delta\boldsymbol{x}, \Delta\boldsymbol{y} \to 0$ die absoluten bzw. relativen partiellen Konditionszahlen

$$L_x(\boldsymbol{x}, \boldsymbol{y}) := \|\boldsymbol{y}\|_2, \quad L_y(\boldsymbol{x}, \boldsymbol{y}) := \|\boldsymbol{x}\|_2 \quad \text{bzw.} \quad K_x(\boldsymbol{x}, \boldsymbol{y}) = K_y(\boldsymbol{x}, \boldsymbol{y}) = \frac{\|\boldsymbol{x}\|_2 \, \|\boldsymbol{y}\|_2}{|\boldsymbol{x}^T\boldsymbol{y}|} \tag{36}$$

ab. Aus (35) bzw. (36) folgt insbesondere, daß der relative Fehler des Skalarproduktes fast orthogonaler Vektoren beliebig groß werden kann, d. h., die Skalarproduktbestimmung fast orthogonaler Vektoren ist bezüglich des relativen Fehlers ein schlecht konditioniertes Problem.

Zahlenbeispiel. Für $n = 2$ und $\boldsymbol{x} = (0.9987, 0.05097)^T$, $\boldsymbol{y} = (1.023, -20.16)^T$ ist $s = \boldsymbol{x}^T\boldsymbol{y} = -0.005885$. Für die im Rahmen des absoluten Fehlerniveaus $\Delta\boldsymbol{x} = \Delta\boldsymbol{y} = 0.01$ zulässigen Störungen $\delta\boldsymbol{x} = (0, -0.01)^T$ und $\delta\boldsymbol{y} = (0.01, 0)^T$ ergibt sich $(\boldsymbol{x} + \delta\boldsymbol{x})^T (\boldsymbol{y} + \delta\boldsymbol{y}) = 0.205702$, also $\delta s = 0.2116$ und $\delta s/s = -35.95$. Die absoluten Konditionszahlen sind $L_x = \|\boldsymbol{y}\|_2 = 20.19$, $L_y = \|\boldsymbol{x}\|_2 = 1.000$, die relativen $K_x = K_y = \|\boldsymbol{x}\|_2 \, \|\boldsymbol{y}\|_2/|\boldsymbol{x}^T\boldsymbol{y}| = 3430$, d. h., die absoluten Fehler können sich etwa um den Faktor 20, die relativen Fehler jedoch um den Faktor 3430 verstärken. Die in Analogie zu (24) bzw. (29) berechneten Näherungswerte für Δs_{opt} bzw. $\Delta s_{\text{opt}}/|s|$ sind $\Delta s_{\text{opt}} \approx L_x\Delta\boldsymbol{x} + L_y\Delta\boldsymbol{y} = 0.1219$ bzw. $\Delta s_{\text{opt}}/|s| \approx K_x \cdot \Delta\boldsymbol{x}/\|\boldsymbol{x}\|_2 + K_y \cdot \Delta\boldsymbol{y}/\|\boldsymbol{y}\|_2 = 36.000$, d. h., mit der obigen Störung $\delta\boldsymbol{x}, \delta\boldsymbol{y}$ wird das unvermeidliche Fehlerniveau praktisch erreicht. □

Wir haben bisher korrekt gestellte Probleme $\{\boldsymbol{e}, \boldsymbol{P}\}$ aus der Klasse $\{\mathcal{E}, \boldsymbol{P}\}$ betrachtet, d. h. Probleme, die eine eindeutige, stetig von $\boldsymbol{e}$ abhängende Lösung $\boldsymbol{a} = \boldsymbol{P}(\boldsymbol{e})$ besitzen. Ist dies nicht der Fall, so heißt das Problem *inkorrekt* gestellt (engl. „ill-posed", „not properly posed", russ. „некорректно поставлено"). Es gibt dann evtl. keine Lösung, oder es existieren mehrere Lösungen, oder die Lösung hängt nicht stetig von den Eingangsdaten ab. Ein Beispiel für Mehrdeutigkeit ist die Eigenvektorbestimmung bei mehrfachen Eigenwerten symmetrischer Matrizen, siehe Ü 2.1.9. Hier kann durch eine Präzisierung der Problemstellung bzw. durch Hinzunahme von Bedingungen, welche die Eindeutigkeit garantieren, zu korrekten Aufgabenstellungen gelangt werden.

Im letzten Fall, d. h. bei Unstetigkeit von $\boldsymbol{P}$ in $\boldsymbol{e}$, muß die Problemstellung durch Veränderung von $\mathcal{E}$ bzw. $\boldsymbol{P}$ modifiziert und damit Stetigkeit erzwungen werden. Man spricht dann von einer *Regularisierung* des inkorrekt gestellten Problems.

2.1.11. *Beispiel.* Für $x \in \mathcal{E} := \mathbf{R}$ sei $x^+ := P(x)$ definiert als

$$P(x) = \begin{cases} 1/x & \text{für} \quad x \neq 0, \\ 0 & \text{für} \quad x = 0. \end{cases}$$

Für $x = 0$ und $\delta x \neq 0$ gilt $|P(x + \delta x) - P(x)| = 1/|\delta x|$, d. h., eine Störung $|\delta x| > 0$ von x wirkt sich wie $1/|\delta x|$ auf das Ergebnis aus. Für $x = 0$ liegt daher eine extrem inkorrekt gestellte Aufgabe vor. Mögliche Regularisierungen sind

$$P(x, \alpha) := \frac{x}{\alpha^2 + x^2} \quad \text{bzw.} \quad P(x, \alpha) := \begin{cases} 1/x & \text{für } |x| \geqq \alpha, \\ 0 & \text{für } |x| < \alpha, \end{cases}$$

wobei $\alpha > 0$ ein geeignet festzulegender Regularisierungsparameter ist. □

Wir werden in Kapitel 8 sehen, daß die Lösung des Quadratmittelproblems $\boldsymbol{A}\boldsymbol{x} \cong \boldsymbol{b}$ wie auch die Bestimmung der Pseudoinversen $\boldsymbol{A}^+$ im Sinne von 2.1.11 inkorrekt gestellte Aufgaben sind, sofern $\boldsymbol{A}$ rangdefizient ist. Zur Regularisierung werden beide angegebenen Möglichkeiten benutzt, siehe Abschnitte 11.2 und 11.3.

Übungsaufgaben

Ü 2.1.1. Zur Integration eines steifen Differentialgleichungssystems $\dot{\boldsymbol{y}} = \boldsymbol{f}(t, \boldsymbol{y})$, $\boldsymbol{y} = \boldsymbol{y}(t) = (y_1(t), \ldots, y_n(t))^T$, $t \in [0, T]$, kann das als Trapezregel bezeichnete implizite Verfahren

$$\boldsymbol{y}^{k+1} = \boldsymbol{y}^k + 0.5h_k[\boldsymbol{f}(t_{k+1}, \boldsymbol{y}^{k+1}) + \boldsymbol{f}(t_k, \boldsymbol{y}^k)], \qquad h_k > 0 \text{ Schrittweite},$$

verwendet werden. Man zeige, daß bei Anwendung auf die lineare Differentialgleichung

$$\dot{\boldsymbol{y}} = \boldsymbol{A}(t)\,\boldsymbol{y} + \boldsymbol{g}(t), \qquad \boldsymbol{A}(.)\colon [0, T] \to \mathbf{R}^{n,n}, \qquad \boldsymbol{g}(.)\colon [0, T] \to \mathbf{R}^n$$

in jedem Schritt das Gleichungssystem

$$[I - 0.5h_k\boldsymbol{A}(t_{k+1})]\,\boldsymbol{y}^{k+1} = \boldsymbol{y}^k + 0.5h_k[\boldsymbol{A}(t_k)\,\boldsymbol{y}^k + \boldsymbol{g}(t_{k+1}) + \boldsymbol{g}(t_k)]$$

zu lösen ist. Dabei ist $\boldsymbol{y}^k \in \mathbf{R}^n$ die Näherung für $\boldsymbol{y}(t_k)$.

Ü 2.1.2. Zur Lösung des Randwertproblems

$$-y'' + p(x)\,y = q(x), \qquad y(0) = y(1) = 0,$$

für die skalare Funktion $y = y(t)$ nach dem gewöhnlichen Differenzenverfahren wird das Erfülltsein der Differentialgleichung in den Gitterpunkten $x_i := i/(n + 1)$ $(i = 1, \ldots, n)$ gefordert und $y''(x_i)$ durch die Differenzenapproximation

$$y''(x_i) \approx \frac{1}{h^2}[y(x_{i-1}) - 2y(x_i) + y(x_{i+1})]$$

ersetzt, wobei $h := 1/(n + 1)$ ist. Man zeige, daß dadurch das lineare Gleichungssystem

$$-y_{i-1} + (2 + h^2p_i)\,y_i - y_{i-1} = h^2q_i \qquad (i = 1, \ldots, n),$$

$$y_0 = y_{n+1} = 0, \qquad p_i = p(x_i), \qquad q_i = q(x_i)$$

für die Näherungswerte $y_i \approx y(x_i)$ mit einer symmetrischen und tridiagonalen Koeffizientenmatrix entsteht.

Ü 2.1.3. Der Ansatz $y = \sum_{j=1}^{n} x_j\varphi_j(t)$ mit n Ansatzfunktionen $\varphi_j(t)$ und freien Parametern x_j soll so an $m \geqq n$ Meßwertpaare (t_i, y_i) $(i = 1, \ldots, m)$ angepaßt werden, daß

$$\sum_{i=1}^{m} \left| y_i - \sum_{j=1}^{n} x_j\varphi_j(t_i) \right|^2 \to \underset{x_1,\ldots,x_n}{\text{Minimum!}}$$

Man schreibe diese Aufgabe als lineares Quadratmittelproblem $\boldsymbol{A}\boldsymbol{x} \cong \boldsymbol{b}$. Wie ändert sich $\boldsymbol{A}$ bzw. $\boldsymbol{b}$ beim ändern, Streichen oder Hinzufügen einer Ansatzfunktion bzw. eines Meßwertpaares?

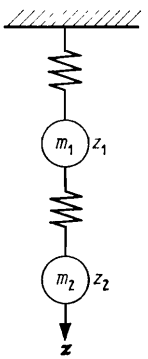
Ü 2.1.4. Es seien $\boldsymbol{A} \in \mathbf{R}^{m,n}$, $\boldsymbol{B} = (\boldsymbol{b}^1, \ldots, \boldsymbol{b}^q) \in \mathbf{R}^{m,q}$ sowie $\boldsymbol{X} = (\boldsymbol{x}^1, \ldots, \boldsymbol{x}^q) \in \mathbf{R}^{n,q}$ gegeben. Man zeige, daß $\boldsymbol{X}$ genau dann $\|\boldsymbol{A}\boldsymbol{X} - \boldsymbol{B}\|_F$ minimiert, wenn die Spalten $\boldsymbol{x}^k$ die Quadratmittelprobleme $\boldsymbol{A}\boldsymbol{x} \cong \boldsymbol{b}^k$ $(k = 1, \ldots, q)$ lösen.

Ü 2.1.5. Es sei $\boldsymbol{A} \in \mathbf{R}^{n,n}$ eine diagonalähnliche Matrix mit den Eigenwerten $\{\lambda_j\}$ und den linear unabhängigen Eigenvektoren $\{\boldsymbol{s}^j\}$. Man beweise, daß $\boldsymbol{y}(t) := \sum_{j=1}^{n} a_j \, \mathrm{e}^{\lambda_j t} \boldsymbol{s}^j$ für beliebige Koeffizienten $a_j \in \mathbf{C}$ die Differentialgleichung

$$\dot{\boldsymbol{y}} = \boldsymbol{A}\boldsymbol{y}$$

löst und daß $\boldsymbol{y}(t) := \mathrm{e}^{\boldsymbol{A}t}\boldsymbol{y}^0$ mit $\mathrm{e}^{\boldsymbol{A}t} := \boldsymbol{S}\,\mathrm{e}^{\boldsymbol{\Lambda}t}\boldsymbol{S}^{-1}$, $\mathrm{e}^{\boldsymbol{\Lambda}t} := \mathrm{diag}\,(\mathrm{e}^{\lambda_j t})$, $\boldsymbol{S} := (\boldsymbol{s}^1, \ldots, \boldsymbol{s}^n)$ die Lösung zum Anfangswert $\boldsymbol{y}(0) = \boldsymbol{y}^0$ ist.

Ü 2.1.6. Gegeben sei das folgende, unter dem Einfluß der Schwerkraft stehende schwingfähige System.



Die Bewegung der Punktmassen m_1, m_2 sei auf die vertikale z-Achse eingeschränkt, und k_i, l_i bezeichne Federkonstante und Länge der Feder i. Man überlege sich:

(i) Die Bewegungsgleichungen des Systems lauten

$$\boldsymbol{M}\ddot{\boldsymbol{z}} = -\boldsymbol{K}\boldsymbol{z} + \boldsymbol{p}$$

mit

$$\boldsymbol{z} = \begin{pmatrix} z_1 \\ z_2 \end{pmatrix}, \quad \boldsymbol{p} = \begin{pmatrix} m_1 g + k_1 l_1 - k_2 l_2 \\ m_2 g + k_2 l_2 \end{pmatrix}, \quad \boldsymbol{M} = \begin{pmatrix} m_1 & 0 \\ 0 & m_2 \end{pmatrix}, \quad \boldsymbol{K} = \begin{pmatrix} k_1 + k_2 & -k_2 \\ -k_2 & k_2 \end{pmatrix},$$

wobei $\boldsymbol{M}$ und $\boldsymbol{K}$ symmetrisch und positiv definit sind.

(ii) Die Gleichgewichtslage $\boldsymbol{z}^0$ ergibt sich aus dem linearen Gleichungssystem $\boldsymbol{K}\boldsymbol{z} = \boldsymbol{p}$.

(iii) Die Auslenkungen $\boldsymbol{u} := \boldsymbol{z} - \boldsymbol{z}^0$ aus der Gleichgewichtslage genügen der Differentialgleichung

$$\boldsymbol{M}\ddot{\boldsymbol{u}} = -\boldsymbol{K}\boldsymbol{u},$$

und die Eigenfrequenzen ω_j von periodischen Lösungen $\boldsymbol{u} = \mathrm{e}^{\mathrm{i}\omega t}\boldsymbol{x}$ ergeben sich aus dem allgemeinen symmetrischen Eigenwertproblem

$$\omega^2 \boldsymbol{M}\boldsymbol{x} = \boldsymbol{K}\boldsymbol{x}.$$

Ü 2.1.7. Es sei $\{e, P\}$ ein numerisches Problem, und $P: \mathbf{R}^N \to \mathbf{R}^M$ besitze in der Umgebung von e stetige partielle Ableitungen $\partial P_i/\partial e_j$. Man überlege sich:

(i) Das Problem $\{e, P\}$ ist lokal lipschitzstetig, und es gilt

$$\|P(\tilde{e}) - P(e)\| \leqq \sum_{j=1}^{N} L_j(e, \Delta e)\, |\tilde{e}_j - e_j|$$

mit

$$L_j(e, \Delta e) := \max \left\{ \left\| \frac{\partial P}{\partial e_j}(e_j) \right\| : \|\tilde{e} - e\| \leqq \Delta e \right\}.$$

Insbesondere sind die Zahlen $L_j(e) := \left\| \frac{\partial P}{\partial e_j}(e) \right\|$ die partiellen Konditionszahlen von $\{e, P\}$.

(ii) Besitzt P stetige partielle Ableitungen zweiter Ordnung, so ist

$$P(\tilde{e}) - P(e) = \sum_{j=1}^{N} \left[\frac{\partial P}{\partial e_j}(e) \right] (\tilde{e}_j - e_j) + O(\|\tilde{e} - e\|^2).$$

Dabei ist $\frac{\partial P}{\partial e_j}(e) = \left(\frac{\partial P_1}{\partial e_j}, \ldots, \frac{\partial P_M}{\partial e_j} \right)^{\mathsf{T}}$ der Vektor der partiellen Ableitungen von P nach e_j.

Ü 2.1.8. Die Zahl $e \neq 0$, $e \in \mathbf{R}$, sei im Rahmen des relativen Fehlerniveaus $\sigma > 0$ gegeben, d. h., e repräsentiere alle Zahlen der Menge $\mathscr{E}_r := \mathscr{E}_r(e, \sigma) := \{\tilde{e} \in \mathbf{R}: |(\tilde{e} - e)/e| \leqq \sigma\}$. Man zeige, daß alle zulässigen $\tilde{e} \in \mathscr{E}_r$ genau dann von 0 verschieden sind, wenn $\sigma < 1$ gilt.

Im Fall $\sigma < 1$ kann daher auch

$$\varepsilon' := \varepsilon'(e) := \frac{\tilde{e} - e}{\tilde{e}} \quad \text{statt} \quad \varepsilon = \varepsilon(e) = \frac{\tilde{e} - e}{e}$$

als relativer Fehler von e definiert werden. Man rechne nach, daß

$$\varepsilon' = \frac{\varepsilon}{1 + \varepsilon} \quad \text{sowie} \quad |\varepsilon'| \leqq \frac{|\varepsilon|}{1 - |\varepsilon|} \leqq \frac{\sigma}{1 - \sigma} =: \sigma'$$

für alle $\tilde{e} \in \mathscr{E}_r$ gilt. Für kleines relatives Fehlerniveau stimmen daher ε' und ε sowie die zugehörigen Schranken σ' und σ praktisch überein.

Ü 2.1.9 (Parlett [80a]). Gegeben ist die symmetrische Matrix

$$A(\varepsilon) := \begin{pmatrix} 1 + \varepsilon \cos(2/\varepsilon) & \varepsilon \sin(2/\varepsilon) \\ \varepsilon \sin(2/\varepsilon) & 1 - \varepsilon \cos(2/\varepsilon) \end{pmatrix}$$

für $\varepsilon \geqq 0$. Man überlege sich:

(i) $A(\varepsilon)$ hängt stetig von ε ab und besitzt die Eigenwerte $\lambda_1(\varepsilon) = 1 + \varepsilon$, $\lambda_2(\varepsilon) = 1 - \varepsilon$. Die zugehörigen orthonormalen Eigenvektoren sind im Fall $\varepsilon > 0$ bis auf das Vorzeichen eindeutig bestimmt und lauten

$$u^1(\varepsilon) = (\cos(1/\varepsilon), \sin(1/\varepsilon))^{\mathsf{T}}, \qquad u^2(\varepsilon) = (\sin(1/\varepsilon), -\cos(1/\varepsilon))^{\mathsf{T}}.$$

(ii) Im Fall $\varepsilon = 0$ hat $A(\varepsilon)$ den doppelten Eigenwert $\lambda_1(0) = \lambda_2(0) = 1$, aber die Eigenvektoren $u^1(\varepsilon), u^2(\varepsilon)$ haben für $\varepsilon \to +0$ keinen Grenzwert. Jedes $u \in \mathbf{R}^2$ mit $\|u\|_2 = 1$ ist ein normierter Eigenvektor von $A(0)$ und Häufungspunkt der Menge $\{u = u^1(\varepsilon), u^2(\varepsilon): \varepsilon > 0\}$.

2.2. Computerarithmetik

A. *Numerisches Rechnen*

Das derzeitige numerische Rechnen beruht fast ausschließlich auf der seit Jahrhunderten bekannten Entwicklung von reellen Zahlen nach Potenzen einer Basis und der Anordnung der Entwicklungskoeffizienten als Ziffern im zugehörigen Stellenwertsystem. Beispiele für endliche sowie unendliche Potenzentwicklungen zur Basis 10 mit den Koeffizienten bzw. Ziffern 0, 1, ..., 9 sind

$$\begin{aligned} 5\times 10^2 + 0\times 10^1 + 7\times 10^0 &= 507,\\ 6\times 10^{-3} + 6\times 10^{-4} + 6\times 10^{-5} + \cdots &= 0.00666\ldots,\\ 3\times 10^0 + 1\times 10^{-1} + 4\times 10^{-2} + 1\times 10^{-3} + \cdots &= 3.141\ldots \end{aligned}$$

Neben der übersichtlichen Schreibweise und den einfachen Vorschriften zur Ausführung der arithmetischen Operationen hat diese Zahldarstellung den Vorteil, daß die erste von 0 verschiedene Ziffer und ihre Stellung relativ zum Dezimalpunkt die wesentlichste Information über die Zahl enthält. Alle weiter rechts stehenden Ziffern liefern zusätzliche, jedoch weniger und weniger wesentliche Information.

Bei den meisten Rechnungen treten Zahlen unterschiedlicher Größenordnung auf, so daß sich die halblogarithmische Darstellung der Potenzentwicklung anbietet. Für die obigen Beispiele lautet diese

$$\begin{aligned} &0.507\times 10^3 && (= 507.),\\ &0.666\ldots\times 10^{-2} && (= 0.00666\ldots),\\ &0.3141\ldots\times 10^1 && (= 3.141\ldots). \end{aligned}$$

Allgemein kann jede reelle Zahl $x \neq 0$ in der Form

$$x = m_R \cdot 10^e = \pm 0.m_1m_2m_3\ldots\times 10^e, \qquad m_i \in \{0, 1, \ldots, 9\} \tag{1}$$

mit der Mantisse m_R und dem ganzzahligen Exponenten e geschrieben werden, wobei die Normalisierungsbedingung $m_1 \neq 0$ als erfüllt vorausgesetzt wird.

Die numerische Praxis erfordert die Beschränkung auf Potenzentwicklungen mit endlicher Stellenzahl. Da dann i. allg. nicht mehr alle Ziffern berücksichtigt werden können, entstehen Fehler.

2.2.1. *Bemerkung.* Numerisches Rechnen ist durch die Einschränkung auf endliche Stellenzahl gekennzeichnet. Dadurch werden i. allg. sowohl bei der Darstellung von Zahlen als auch bei der Ausführung arithmetischer Operationen Fehler erzeugt, die Darstellungs- bzw. Rundungsfehler genannt werden.

Mit wachsender Länge der Potenzentwicklung wächst die Genauigkeit, aber auch der Rechenaufwand, d. h. der zum Speichern der Zahlen benötigte Speicherplatz und die Rechenzeit für die Ausführung arithmetischer und anderer Operationen. Die Länge der benutzten Potenzentwicklung soll daher ein vernünftiger Kompromiß zwischen Genauigkeitsanforderungen und Rechenkosten sein. Handrechnung erlaubt die flexible Anpassung der Stellenzahl an den Rechengang. Computerrech-

nung ist Rechnung mit fester Stellenzahl, die für übliche Genauigkeitsforderungen meist höher als eigentlich nötig ist. Andererseits kann wegen der hohen Zahl der Rechenoperationen der summarische Einfluß der Rundungsfehler wesentlich größer werden als bei Handrechnung.

2.2.2. *Bemerkung.* Die Einschätzung der Genauigkeit von Zahlen, die Ergebnisse von Computerrechnungen sind, erfordert die Kenntnis der wesentlichen Eigenschaften der Computerarithmetik und eine Fehleranalyse des benutzten Verfahrens.

Die Arithmetiken der derzeitigen Computertypen unterscheiden sich in vielen Einzelheiten. Ihre wesentlichen Eigenschaften sind jedoch dieselben und können durch wenige Kenngrößen charakterisiert werden. Dies ist Gegenstand der nachfolgenden Ausführungen; als Beispiel gehen wir auf die Arithmetik der Rechner der ESER-Reihe und der Reihe IBM 370 und Nachfolger ein. Grundlegende Eigenschaften von numerischen Algorithmen und Elemente der Fehleranalyse werden im nachfolgenden Abschnitt 2.3 behandelt.

B. *Computerzahlen*

Ausgangspunkt für die Darstellung einer Computerzahl ist die halblogarithmische Darstellung (1), wobei jedoch nur eine feste Anzahl von Mantissenstellen und ein fester Exponentenbereich zugelassen werden. Aus Gründen der effektiven Realisierung der Hardware arbeiten die meisten Computer dabei nicht mit der Basis 10, sondern mit einer Zweierpotenz als Basis. Üblich sind die Basen 2, 8 oder 16; man spricht von binärer, oktaler oder hexadezimaler Arithmetik.

2.2.3. *Begriffe.* Es seien β, t, E_1 und E_2 positive ganze Zahlen, $\beta \geqq 2$. Dann besteht die Menge $\mathfrak{R} = \mathfrak{R}(\beta, t, E_1, E_2)$ der *Computerzahlen* (auch: *Gleitpunktzahlen*) zur *Basis* β mit der *Mantissenlänge* t und dem *Exponentenbereich* $[-E_1, E_2]$ aus allen Zahlen der Gestalt

$$x = m \cdot \beta^e.$$

Dabei ist e — der *Exponent* von x — eine beliebige ganze Zahl mit $-E_1 \leqq e \leqq E_2$, und m — die *Mantisse* von x — ist ein beliebiger t-stelliger Bruch zur Basis β der Form

$$m = \pm 0.m_1 m_2 \ldots m_t = \pm (m_1 \beta^{-1} + m_2 \beta^{-2} + \cdots + m_t \beta^{-t})$$

mit den Ziffern $m_i \in \{0, 1, \ldots, \beta - 1\}$ $(i = 1, \ldots, t)$, wobei folgendes gilt:

$$\left| \begin{array}{l} \text{entweder } m_1 \geqq 1 \\ \text{oder} \qquad m_1 = m_2 = \cdots = m_t = 0 \text{ und } e = -E_1. \end{array} \right. \tag{2}$$

Die Alternative (2) besagt, daß für jede Computerzahl $x \neq 0$ die erste Mantissenstelle m_1 von 0 verschieden ist; man spricht von *normalisierter Gleitpunktdarstellung*. Die Normalisierungsbedingung ist nur für die Computerzahl $x = 0$ verletzt; für diese gilt $m = 0$. Die Zuordnung des kleinsten Exponenten $e = -E_1$ zur Computerzahl $x = 0$ ist üblich und zweckmäßig, aber nicht zwingend.

Für den Exponentenbereich $[-E_1, E_2]$ gilt meist $E_1 = E_2 + 1$ oder $E_1 = E_2$, d. h., er ist fast symmetrisch.

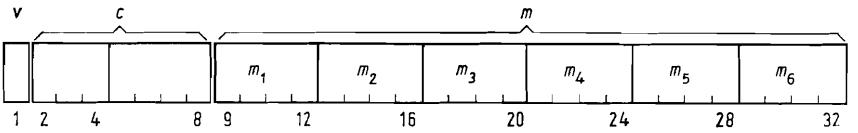
Zur Darstellung einer Zahl im Computer wird in der Regel ein *Computerwort* benutzt, das aus einer bestimmten Anzahl von Bits besteht; gelegentlich werden auch mehrere Computerworte verwendet. Dabei ist ein *Bit* die Einheit der binären Information und kann die Werte 0 oder 1 repräsentieren; jeweils 8 Bit werden als ein *Byte* bezeichnet.

2.2.4. *Beispiel.* Computer der ESER- und IBM-Reihen rechnen zur Basis $\beta = 16$ (hexadezimale Darstellung mit den Ziffern

$$m_i \in \{0, 1, \ldots, 9, 10, \ldots, 15\} = \{0, \ldots, 9, A, B, C, D, E, F\}).$$

Die Wortlänge ist 4 Byte, also 32 Bit.

(i) In einfacher Genauigkeit (REAL*4 bzw. REAL in FORTRAN) wird ein Wort zur Zahldarstellung verwendet mit der folgenden Aufteilung:



Dabei ist v das Vorzeichen von m ($v = 0$ für $m \geqq 0$, $v = 1$ für $m < 0$). Die Zahl $c := e + 64$ ist die Charakteristik von x und wird mit 7 bit binär codiert, so daß $0 \leqq c \leqq 2^7 - 1$, also $-64 \leqq e \leqq 63$ gilt. Die 6 hexadezimalen Mantissenziffern werden binär in den Positionen 9 bis 32 codiert. Die Menge der einfachgenauen Computerzahlen ist daher $\mathfrak{R}(16, 6, 64, 63)$.

(ii) In doppelter Genauigkeit (REAL*8 bzw. DOUBLE PRECISION in FORTRAN) wird ein Doppelwort zur Zahldarstellung verwendet. Das linke entspricht dem bei einfachgenauer Darstellung, das rechte nimmt in den Positionen 1 bis 32 die weiteren 8 hexadezimalen Mantissenziffern $m_7, \ldots, m_{14}$ in binärer Codierung auf. Die Menge der doppeltgenauen Computerzahlen ist daher $\mathfrak{R}(16, 14, 64, 63)$. □

Aus 2.2.3 folgt für jedes $x \in \mathbf{R}$, $x \neq 0$, die Beziehung $\beta^{-1} \leqq |m| \leqq 1 - \beta^{-t}$, also

$$\beta^{e-1} \leqq |x| \leqq \beta^e(1 - \beta^{-t}) < \beta^e. \tag{3}$$

Für zwei benachbarte, von 0 verschiedene Computerzahlen x, x' ergibt sich damit

$$\beta^{-t} < |x - x'|/|x| \leqq \beta^{1-t}.$$

Schließlich zeigt (3), daß für jedes $x \in \mathfrak{R}$

$$\text{entweder } x = 0 \quad \text{oder } \beta^{-E_1-1} \leqq |x| < \beta^{E_2}$$

gilt. Zusammenfassend erhalten wir die folgende Aussage:

2.2.5. *Eigenschaften der Computerzahlen.* Die Computerzahlen aus $\mathfrak{R}(\beta, t, E_1, E_2)$ bilden im Intervall [−MAX, +MAX], MAX $:= \beta^{E_2}$, ein endliches, symmetrisches und ungleichmäßiges Gitter. Das Intervall [−MIN, +MIN], MIN $:= \beta^{-E_1-1}$, ent-

hält nur die drei Computerzahlen $-\mathrm{MIN}, 0, +\mathrm{MIN}$. Die relative Feinheit des Gitters in den Intervallen $[-\mathrm{MAX}, -\mathrm{MIN}]$ und $[\mathrm{MIN}, \mathrm{MAX}]$ genügt den Ungleichungen

$$\beta^{-t} < \frac{|x - x'|}{|x|} \leqq \beta^{1-t},$$

wobei x, x' zwei benachbarte Computerzahlen aus diesen Intervallen sind.

2.2.6. *Beispiel.* Für ESER- und IBM-Computer gilt

$$\mathrm{MIN} = 16^{-65} = 5.4 \times 10^{-79}, \qquad \mathrm{MAX} = 16^{63} = 7.2 \times 10^{75}$$

in einfacher wie doppelter Genauigkeit. Die relative Feinheit ist durch die Ungleichungen

$$5.9 \times 10^{-8} = 16^{-6} < |x - x'|/|x| \leqq 16^{-5} = 9.5 \times 10^{-7}$$

in einfacher Genauigkeit bzw.

$$1.4 \times 10^{-17} = 16^{-14} < |x - x'|/|x| \leqq 16^{-13} = 2.2 \times 10^{-16}$$

in doppelter Genauigkeit charakterisiert. □

Für reale Computer ist β^{-t} stets eine sehr kleine Größe. *In Abschätzungen werden wir daher Summanden der Größenordnung β^{-t} gegenüber solchen der Größenordnung* 1 *vernachlässigen;* z. B. schreiben wir statt

$$|\lambda| \leqq 1.5 + 4 \cdot \beta^{-t}$$

einfach

$$|\lambda| \stackrel{\cdot}{\leqq} 1.5,$$

d. h., $\gamma \stackrel{\cdot}{\leqq} \delta$ bedeutet $\gamma \leqq \delta(1 + \varrho)$ mit $\varrho = O(\beta^{-t})$. Analog wird das Zeichen „$\stackrel{\cdot}{=}$“ eingeführt. Später werden wir in solchen Abschätzungen der Einfachheit halber auch den Punkt weglassen.

C. *Darstellung reeller Zahlen durch Computerzahlen*

Für die Computerrechnung müssen reelle Zahlen $x \in \mathbf{R}$ durch die endlich vielen Computerzahlen aus $\mathfrak{R}(\beta, t, E_1, E_2)$ approximiert werden.

Im Fall $|x| > \mathrm{MAX}$ ist dies nicht sinnvoll möglich, weil der Exponentenbereich nicht ausreicht, um wenigstens die Größenordnung von x richtig wiederzugeben. Man spricht von *Exponentenüberlauf*, und die Rechnung wird i. allg. abgebrochen.

Im Fall $|x| \leqq \mathrm{MAX}$ muß der Zahl $x \in \mathbf{R}$ jedoch in eindeutiger Weise eine Computerzahl $\mathrm{rd}\,(x) \in \mathfrak{R}$ zugeordnet werden. Für $|x| < \mathrm{MIN}$ ist dies $\mathrm{rd}\,(x) = 0$. Im Fall $0 < |x| < \mathrm{MIN}$ spricht man dann von *Exponentenunterlauf*.

Für $\mathrm{MIN} \leqq |x| \leqq \mathrm{MAX}$ sind zwei typische Zuordnungen üblich, nämlich

— *Rundung* (auch: symmetrische Rundung, engl. „rounding“, russ. „правильное округление“)

— *Abbrechen* (auch: unsymmetrische Rundung, engl. „chopping“).

Im Fall der Rundung ist rd (x) die x am nächsten gelegene Computerzahl. Wenn zwei solche Computerzahlen existieren, wird eine von beiden — z. B. die betragsgrößere — genommen.

Beim Abbrechen ist rd (x) die x nächstgelegene Computerzahl zwischen x und 0.

Man überzeugt sich, daß im Fall $\text{MIN} \leqq |x| \leqq \text{MAX}$ die Computerdarstellung rd (x) der reellen Zahl x wie folgt gefunden werden kann: Man bestimmt die halblogarithmische Darstellung $x = m_R \beta^e$ von x mit $\beta^{-1} \leqq |m_R| < 1$ und rundet die (normalisierte) Mantisse auf t Stellen bzw. bricht sie nach t Stellen ab. Dazu müssen die ersten $t+1$ bzw. t Stellen von m_R berechnet werden.

Wir nennen die Funktion

$$\text{rd}: [-\text{MAX}, \text{MAX}] \subset \mathbf{R} \to \mathfrak{R}$$

Rundungs- oder *Darstellungsfunktion*; die Größen

$$\text{rd}\,(x) - x \quad \text{bzw.} \quad \frac{\text{rd}\,(x) - x}{x} \text{ im Fall } x \neq 0$$

heißen *absoluter* bzw. *relativer Rundungs-* oder *Darstellungsfehler* von x.

Zur Untersuchung dieser Rundungsfehler betrachten wir die folgenden Fälle.

Fall 1: $x = 0$:
Dann ist rd $(x) = 0$, also rd $(x) - x = 0$.
Die Zahl $x = 0$ wird exakt dargestellt.

Fall 2: $0 < |x| < \text{MIN}$:
Dann ist rd $(x) = 0$, also

$$\text{rd}\,(x) - x = -x \quad \text{sowie} \quad \varepsilon(x) = \frac{\text{rd}\,(x) - x}{x} = -1.$$

Im Fall $0 < |x| < \text{MIN}$ ist der relative Rundungsfehler betragsmäßig gleich 100%. Die gesamte in $x \neq 0$ enthaltene Information geht bei Übergang zur Computerdarstellung rd $(x) = 0$ *verloren. Exponentenunterlauf kann daher* — muß aber nicht — *eine Quelle für wesentliche Genauigkeitsverluste sein in Abhängigkeit davon, wie die durch Unterlauf entstandenen Nullen in die nachfolgenden Rechnungen eingehen.*

Fall 3: $\text{MIN} \leqq |x| \leqq \text{MAX}$:
Aus der Definition von rd und der Abschätzung aus 2.2.5 folgt

$$|\varepsilon(x)| = \left|\frac{\text{rd}\,(x) - x}{x}\right| \leqq \nu := \begin{cases} 0.5\beta^{1-t} & \text{für symmetrische Rundung,} \\ \beta^{1-t} & \text{für unsymmetrische Rundung.} \end{cases} \tag{4}$$

Die Zahl ν heißt *relatives Rundungsfehlerniveau* der Arithmetik, gelegentlich auch *relative Maschinengenauigkeit*, und ist i. allg. sehr klein. Abschätzung (4) besagt: *Im Fall $\text{MIN} \leqq |x| \leqq \text{MAX}$ ist der relative Rundungsfehler klein. Die Rundung bewirkt nur einen geringen Verlust der in x enthaltenen Information.*

Die Abschätzung (4) läßt sich äquivalent auch als

$$\text{rd}\,(x) = x\big(1 + \varepsilon(x)\big) \quad \text{mit} \quad |\varepsilon(x)| \leqq \nu \tag{5}$$

schreiben und wie folgt deuten: *Die Computerdarstellung* rd (x) *ist im Fall* MIN $\leqq |x| \leqq$ MAX *die exakte Darstellung einer wenig gestörten Zahl* $x\big(1 + \varepsilon(x)\big)$.

Wenn $\varepsilon(x) = 0$ für $x = 0$ definiert wird, gilt (5) auch für $x = 0$.

Zusammenfassend erhalten wir:

2.2.7. *Eigenschaften der Rundungsfunktion.* Die Rundungsfunktion ordnet jeder reellen Zahl $x \in [-\text{MAX}, \text{MAX}]$ die Computerdarstellung rd $(x) \in \mathfrak{R}(\beta, t, E_1, E_2)$ zu. Dabei gilt

$$\text{rd}\,(x) = x\big(1 + \varepsilon(x)\big)$$

mit

$$\varepsilon(x) = 0 \quad \text{für} \quad x = 0,$$

$$\varepsilon(x) = -1 \quad \text{für} \quad 0 < |x| < \text{MIN},$$

$$|\varepsilon(x)| \leqq \nu \quad \text{für} \quad \text{MIN} \leqq |x| \leqq \text{MAX}$$

und dem relativen Rundungsfehlerniveau

$$\nu = \begin{cases} 0.5\beta^{1-t} & \text{für symmetrische Rundung,} \\ \beta^{1-t} & \text{für unsymmetrische Rundung.} \end{cases}$$

Im Fall $x \neq 0$ stellt $\varepsilon(x)$ den relativen Rundungsfehler von x dar.

2.2.8. *Bemerkung.* (i) Der Nachteil der unsymmetrischen Rundung besteht weniger in dem verdoppelten relativen Rundungsfehlerniveau, sondern mehr in der Tatsache, daß der relative Rundungsfehler $\varepsilon(x)$ stets nichtpositiv ist. Damit ist die Chance einer „stochastischen Auslöschung" relativer Rundungsfehler, z. B. bei Prozessen, in denen sich die relativen Fehler addieren, nicht mehr gegeben, vgl. Abschnitt 2.3.

(ii) Als einfaches Modell für die stochastische Untersuchung des relativen Rundungsfehlers $\varepsilon = \varepsilon(x)$ nehmen wir an, daß x eine auf $[a, b] \subset [\text{MIN}, \text{MAX}]$ gleichverteilte Zufallsgröße sei, wobei $(b - a)/a$ ein ganzzahliges Vielfaches der Basis β ist. Dann stellt ε auch eine Zufallsgröße dar, und im Fall symmetrischer Rundung gilt

$$E(\varepsilon) = 0 + O(\nu^2) \quad \text{sowie} \quad D^2(\varepsilon) = \frac{\nu^2}{3\beta} + O(\nu^3).$$

Für unsymmetrische Rundung ist dagegen

$$E(\varepsilon) = -\frac{\ln \beta}{2(\beta - 1)}\,\nu + O(\nu^2) \quad \text{sowie} \quad D^2(\varepsilon) = \left[\frac{1}{3\beta} - \frac{1}{4}\left(\frac{\ln \beta}{\beta - 1}\right)^2\right]\nu^2 + O(\nu^3).$$

Für Hinweise zur stochastischen Rundungsfehleranalyse siehe B 2.4. □

2.2.9. *Beispiel.* Bei Verwendung von FORTRAN und PL1 arbeiten ESER- und IBM-Computer mit unsymmetrischer Rundung.

In einfacher Genauigkeit ergibt sich mit $v = 9.5 \times 10^{-7}$

$$E(\varepsilon) = -8.8 \times 10^{-8}, \qquad \sqrt{D^2(\varepsilon)} = 1.1 \times 10^{-7}.$$

In doppelter Genauigkeit erhält man mit $v = 2.2 \times 10^{-16}$

$$E(\varepsilon) = -2.1 \times 10^{-17}, \qquad \sqrt{D^2(\varepsilon)} = 2.5 \times 10^{-17}.$$

Die statistischen Werte sind etwa eine Zehnerpotenz günstiger als die maximale Fehlerschranke v.

Hardwaremäßig ist bei den genannten Computern auch eine symmetrische Rundung realisiert, die jedoch nur über Assemblerprogramme genutzt werden kann. □

D. *Computerarithmetik*

Für zwei Computerzahlen $x, y \in \mathfrak{R}$ ist das exakte Ergebnis

$$z = x \diamond y, \qquad \diamond \in \{+, -, *, /\}$$

der arithmetischen Operation $\diamond$ in der Regel keine Computerzahl (bei der Division muß selbstverständlich $y \neq 0$ gefordert werden).

Als Beispiel betrachten wir $\mathfrak{R} = \mathfrak{R}(10, 3, 5, 5)$ und $x = 0.718 \times 10^2$, $y = 0.591 \times 10^{-4}$. Für diese Argumente ist

$$z_1 = x + y = 0.718000591 \times 10^2, \qquad z_3 = y * y = 0.349281 \times 10^{-8},$$

$$z_2 = x * y = 0.424338 \times 10^{-2}, \qquad z_4 = x/y = 0.121489\ldots \times 10^7.$$

Sämtliche Resultate gehören nicht zu $\mathfrak{R}$, denn in allen treten mehr als drei Mantissenstellen auf, und bei z_3 bzw. z_4 wird zusätzlich der Exponentenbereich nach unten bzw. nach oben überschritten.

Im folgenden bezeichne

$$\mathrm{fl}\,(x \diamond y) \in \mathfrak{R}$$

für $x, y \in \mathfrak{R}$ das Computerresultat der arithmetischen Operation $x \diamond y$ (nach der englischen Bezeichnung „floating point arithmetic" für Gleitpunktarithmetik).

Von einer guten Computerarithmetik sollte verlangt werden, daß dieses Computerresultat die Computerdarstellung des exakten Resultates ist, sofern kein Überlauf eintritt. Im obigen Beispiel müßte dann gelten:

$$\mathrm{fl}\,(x + y) = 0.718 \times 10^2, \qquad \mathrm{fl}\,(y * y) = 0 \quad \text{(Unterlauf)},$$

$$\mathrm{fl}\,(x * y) = 0.424 \times 19^{-2}, \qquad \mathrm{fl}\,(x/y) \text{ nicht erklärt} \quad \text{(Überlauf)}.$$

2.2.10. *Postulat. Die Computerarithmetik in* $\mathfrak{R} = \mathfrak{R}(\beta, t, E_1, E_2)$ *arbeite so, daß für* $x, y \in \mathfrak{R}$ *im Fall* $x \diamond y \in [-\mathrm{MAX}, \mathrm{MAX}]$

$$\mathrm{fl}\,(x \diamond y) = \mathrm{rd}\,(x \diamond y) \tag{6}$$

gilt.

2.2.11. *Aussage.* Unter der Voraussetzung 2.2.10 gilt für $x, y \in \mathfrak{R}$ mit $x \diamond y = 0$ oder $\mathrm{MIN} \leqq |x \diamond y| \leqq \mathrm{MAX}$

$$\mathrm{fl}\,(x \diamond y) = (x \diamond y)(1 + \varepsilon) \quad \text{mit} \quad |\varepsilon| \leqq \nu. \tag{7}$$

Dies folgt sofort aus 2.2.7 und besagt: *Der relative Rundungsfehler des Resultats ist im Fall* $\mathrm{fl}\,(x \diamond y) \neq 0$ *höchstens gleich* ν.

Falls $x' = x(1 + \varepsilon)$, $y' = y(1 + \varepsilon)$ geschrieben wird, läßt sich (7) als

$$\mathrm{fl}\,(x \pm y) = x' \pm y', \qquad \mathrm{fl}\,(x * y) = x * y' = x' * y, \qquad \mathrm{fl}\,(x/y) = x'/y$$

schreiben und wie folgt interpretieren: *Das Computerresultat einer arithmetischen Operation ist das exakte Resultat derselben Operation mit wenig gestörten Operanden.* Bei Bedarf können für $\diamond \in \{*, /\}$ die Störungen auf beide Operanden x, y verteilt werden, siehe Ü 2.2.3.

2.2.12. *Bemerkung.* Die Voraussetzung 2.2.10 ist für viele Computer — darunter für solche der ESER- und IBM-Reihen — erfüllt. Es gibt jedoch Computertypen, die eine schwächere Arithmetik besitzen, für die (6) nicht gilt. Dabei treten folgende Abweichungen auf:

(i) Die Darstellung (7) gilt mit der Abschätzung $|\varepsilon| \leqq K\nu$, wobei $K \leqq 3$ ist o. ä.

(ii) Für $\diamond \in \{+, -\}$ ist (7) durch

$$\mathrm{fl}\,(x \pm y) = x(1 + \varepsilon_x) \pm y(1 + \varepsilon_y), \qquad |\varepsilon_x|, |\varepsilon_y| \leqq \nu,$$

zu ersetzen. Hieraus folgt

$$|\mathrm{fl}\,(x \pm y) - (x \pm y)| \leqq \nu(|x| + |y|). \tag{8}$$

Der Fall (ii) tritt auf, wenn die Addition nur mit t Mantissenstellen ohne Schutzstelle ausgeführt wird. Aus (8) folgt für $x \pm y \neq 0$

$$\frac{|\mathrm{fl}\,(x \pm y) - (x \pm y)|}{|x \pm y|} \leqq \frac{|x| + |y|}{|x \pm y|}\,\nu,$$

d. h., im Fall $|x \pm y| \ll |x| + |y|$ kann der relative Rundungsfehler wesentlich größer als ν werden.

Es zeigt sich jedoch, daß diese Abweichungen von (6) *bzw.* (7) *die Genauigkeit der wichtigsten numerischen Prozesse der linearen Algebra nur unwesentlich beeinflussen. Wir nehmen daher im folgenden an, daß* 2.2.10 *erfüllt ist.*

2.2.13. *Bemerkung.* Für $x, y \in \mathfrak{R}$ gelte $x = m_x\beta^e$, $y = m_y\beta^e$, $\mathrm{sgn}\,(x) = \mathrm{sgn}\,(y)$ und $\mathrm{MIN} \leqq |x - y|$. Dann ist

$$\mathrm{fl}\,(x - y) = x - y, \tag{9}$$

d. h., *die Differenz zweier Computerzahlen desselben Vorzeichens und mit demselben Exponenten ist exakt.* Falls dabei mehrere führende Stellen von m_x und m_y übereinstimmen, spricht man von *Auslöschung.*

Dies folgt sofort aus dem Postulat 2.2.10, denn unter den angegebenen Voraussetzungen ist $x - y \in \mathfrak{R}$, also $\mathrm{fl}\,(x - y) = x - y$. Die Arithmetik aller derzeitiger Computer arbeitet jedoch so, daß 2.2.13 auch dann richtig bleibt, wenn 2.2.10 z. B. für Addition und Subtraktion nicht erfüllt ist.

Ein Beispiel für Auslöschung in $\mathfrak{R}(10, 8, 20, 20)$ ist

$$\mathrm{fl}\,(0.98765432 \times 10^1 - 0.98765321 \times 10^1) = 0.11100000 \times 10^{-4}.$$

Auslöschung kann auch bei um 1 verschiedenen Exponenten auftreten, wie durch

$$\mathrm{fl}\,(0.10000000 \times 10^1 - 0.99990000 \times 10^0) = 0.10000000 \times 10^{-3}$$

gezeigt wird. Allgemein ist das Auftreten von Auslöschung durch die Bedingung $|x \pm y| \ll |x| + |y|$ charakterisiert.

Bei Auslöschung gehen führende Mantissenstellen verloren. Dies kann — aber muß nicht — *zu einem wesentlichen Verlust der in x, y enthaltenen Information führen.*

Zur Erläuterung dieser Feststellung fassen wir x, y aus 2.2.13 als Computerdarstellungen der reellen Zahlen $\tilde{x}, \tilde{y} \in \mathbf{R}$ mit $\tilde{x} - \tilde{y} \neq 0$ auf. Dann ist $x = \mathrm{rd}\,(\tilde{x}) = \tilde{x}(1 + \varepsilon_x)$, $y = \mathrm{rd}\,(\tilde{y}) = \tilde{y}(1 + \varepsilon_y)$, wobei $\varepsilon_x, \varepsilon_y$ mit $|\varepsilon_x|, |\varepsilon_y| \leqq \nu$ die relativen Darstellungsfehler bezeichnen. Unter Beachtung von (9) folgt

$$\frac{|\mathrm{fl}\,(x - y) - (\tilde{x} - \tilde{y})|}{|\tilde{x} - \tilde{y}|} = \frac{|(x - y) - (\tilde{x} - \tilde{y})|}{|\tilde{x} - \tilde{y}|} = \frac{|\varepsilon_x \tilde{x} - \varepsilon_y \tilde{y}|}{|\tilde{x} - \tilde{y}|} \leqq \frac{|\tilde{x}| + |\tilde{y}|}{|\tilde{x} - \tilde{y}|}\,\nu.$$

Bei Auslöschung kann sich daher die Ungenauigkeit der Daten mit dem großen Faktor $(|\tilde{x}| + |\tilde{y}|)/|\tilde{x} - \tilde{y}|$ auf die Differenz übertragen, obwohl $\mathrm{fl}\,(x - y)$ fehlerfrei aus x, y berechnet wird!

Auslöschung kann auch bei der Auswertung komplizierterer Formeln auftreten und läßt sich in manchen Fällen durch geschickte Umformungen vermeiden. Beispiele sind

$$1 - \sqrt{1 - x} = \frac{\left(1 - \sqrt{1 - x}\right)\left(1 + \sqrt{1 - x}\right)}{1 + \sqrt{1 - x}} = \frac{x}{1 + \sqrt{1 - x}}, \quad |x| \text{ klein},$$

oder

$$1 - \cos x = 1 - \left(1 - 2\left(\sin\frac{x}{2}\right)^2\right) = 2\left(\sin\frac{x}{2}\right)^2, \quad |x| \text{ klein}.$$

Eine weitere fehlerfrei ausgeführte Operation ist die Multiplikation bzw. Division mit einer Potenz von β.

2.2.14. *Bemerkung.* Für $x, y \in \mathfrak{R}$ und eine Potenz β^k, k ganz, gilt

$$\mathrm{fl}\,(x * \beta^k) = x * \beta^k, \quad \mathrm{fl}\,(y/\beta^k) = y/\beta^k, \quad \text{falls MIN} \leqq \left\{\begin{matrix} |x\beta^k| \\ |y/\beta^k| \end{matrix}\right\} \leqq \text{MAX}.$$

In manchen numerischen Prozessen der linearen Algebra treten neben den arithmetischen Operationen auch Quadratwurzelberechnungen auf. Die entsprechende

Computerfunktion bezeichnen wir mit sqrt — gelegentlich auch mit fl $(\sqrt{x})$ — und fordern:

$$\text{Für } x \in \mathfrak{R},\ x \geqq 0, \text{ gelte sqrt } (x) := \text{rd } (\sqrt{x}). \tag{10}$$

2.2.15. *Bemerkung.* (i) Auf den meisten Computern ist über eine höhere Programmiersprache wie FORTRAN oder PL1 auch eine einfach- und doppeltgenaue *komplexe Arithmetik* nutzbar. Da wir ausschließlich mit reellen Eingangsdaten versehene Probleme und im Reellen arbeitende Algorithmen betrachten, gehen wir auf diese Arithmetik nicht ein.

(ii) Fast alle Computer verfügen außer über die beschriebene Gleitpunktarithmetik noch über eine *ganzzahlige*, rundungsfehlerfreie *Arithmetik* für Indexrechnungen o. ä., die auf numerische Prozesse keinen wesentlichen Einfluß hat.

Übungsaufgaben

Ü 2.2.1. Man überlege sich, daß die Menge $\mathfrak{R}(\beta, t, E_1, E_2)$ aus $2(E_1 + E_2 + 1)\,\beta^{t-1}(\beta - 1) + 1$ verschiedenen Zahlen besteht.

Ü 2.2.2. Es ist zu zeigen, daß das relative Rundungsfehlerniveau ν unter der Voraussetzung 2.2.10 durch

$$\nu = \min \{\nu' \in \mathfrak{R} : \text{fl } (1 + \nu') > 1\} \tag{11}$$

charakterisiert werden kann. Man schreibe ein FORTRAN-Programm, das ν auf der Grundlage von (11) mit einem relativen Fehler von höchstens 0.1 berechnet.

Ü 2.2.3. Man zeige, daß im Fall $|\varepsilon| \leqq \nu$

$$1 + \varepsilon = \frac{1}{1 + \varepsilon_1} = \frac{1 + \varepsilon_2}{1 + \varepsilon_3} = (1 + \varepsilon_4)^2 \tag{12}$$

mit $\varepsilon_1 \doteqdot -\varepsilon$, $\varepsilon_2 = -\varepsilon_3 \doteqdot \varepsilon/2$, $\varepsilon_4 \doteqdot \varepsilon/2$, also

$$|\varepsilon_1| \lessdot \nu, \qquad |\varepsilon_2|, |\varepsilon_3|, |\varepsilon_4| \lessdot \nu/2 \tag{13}$$

gilt. Für die Operationen $*$ und $/$ läßt sich dann (7) in der Gestalt

$$\text{fl } (x * y) = [x(1 + \varepsilon)] * y = x * [y(1 + \varepsilon)] = [x(1 + \varepsilon_4)] * [y(1 + \varepsilon_4)]$$

bzw.

$$\text{fl } (x/y) = [x(1 + \varepsilon)]/y = x/[y(1 + \varepsilon_1)] = [x(1 + \varepsilon_2)]/[y(1 + \varepsilon_3)]$$

mit den Schranken (13) schreiben, d. h., die Störungen können nach Wunsch auf jeden der einzelnen bzw. auf beide Operanden verteilt werden.

Ü 2.2.4. In $\mathfrak{R}(10, 4, 10, 10)$ seien die Zahlen $x = 0.2345 \times 10^{-2}$, $y = 0.6789 \times 10^{0}$, $z = 0.1234 \times 10^{2}$ gegeben. Man überprüfe, daß bei Rechnung mit unsymmetrischer Rundung

$$\text{fl } (\text{fl } (x + y) + z) = 0.1302 \times 10^{2} \neq \text{fl } (x + \text{fl } (y + z)) = 0.1301 \times 10^{2}$$

sowie

$$\text{fl } (x * \text{fl } (y + z)) = 0.3050 \times 10^{-1} \neq \text{fl } (\text{fl } (x * y) + \text{fl } (x * z)) = 0.3052 \times 10^{-1}$$

gilt. In Gleitpunktarithmetik gelten also weder das Assoziativgesetz noch das Distributivgesetz.

Ü 2.2.5. Bei der Berechnung von $\|\boldsymbol{v}\|_2 - |v_1|$ für $\boldsymbol{v} = (v_1, \ldots, v_n)^T \in \mathbf{R}^n$ tritt im Fall $|v_1| \gg |v_j|$ $(j = 2, \ldots, n)$ Auslöschung auf. Man überlege sich, daß dieser Nachteil bei Berechnung gemäß

$$\|\boldsymbol{v}\|_2 - |v_1| = \frac{\sum_{j=2}^{n} |v_j|^2}{\|\boldsymbol{v}\|_2 + |v_1|}$$

vermieden wird.

2.3. Numerische Algorithmen und Grundlagen der Fehleranalyse

Die Lösung des numerischen Problems $\{\boldsymbol{e}, \boldsymbol{P}\}$ aus der Klasse $\{\mathcal{E}, \boldsymbol{P}\}$ erfordert die Berechnung der zu $\boldsymbol{e}$ gehörenden Ausgangsdaten $\boldsymbol{a}^* = \boldsymbol{P}(\boldsymbol{e})$. Dazu wird ein *numerischer Algorithmus* benötigt, d. h. *eine Vorschrift* $\boldsymbol{V}_0 : \mathcal{E} \subset \mathbf{R}^N \to \mathbf{R}^M$, *die jedem Satz* $\boldsymbol{e} \in \mathcal{E}$ *von Eingangsdaten unter Ausführung endlich vieler arithmetischer Operationen* $\{+, -, *, /\}$ *in eindeutiger Weise einen Satz* $\boldsymbol{a}^* = \boldsymbol{V}_0(\boldsymbol{e})$ *von Ausgangsdaten zuordnet.* Ein numerischer Algorithmus ist also eine spezielle *Datentransformation* von $\boldsymbol{e} = (e_1, \ldots, e_N)^T$ in $\boldsymbol{a}^* = (a_1^*, \ldots, a_M^*)^T$, bei der nur endlich viele arithmetische Operationen erlaubt sind. Die Forderung nach der Endlichkeit der Zahl der Daten und Rechenoperationen ist dabei natürlich, denn auf einem realen Computer können in endlicher Zeit nur endlich viele Operationen mit endlich vielen Operanden ausgeführt werden. Es ist zweckmäßig, außer den arithmetischen Operationen auch die Berechnung von elementaren Funktionen — in der linearen Algebra speziell die von $\sqrt{x}$ — als Basisoperation zuzulassen.

Üblicherweise wird ein numerischer Algorithmus als Programm in einer höheren Programmiersprache wie FORTRAN, PL1 usw. angegeben. Da solche Programme i. allg. jedoch unübersichtlich und schwer lesbar sind, beschreiben wir Algorithmen in traditioneller mathematischer Schreibweise oder in einer sich selbst erklärenden Pseudo-Programmiersprache, wodurch der wesentliche mathematische Gehalt deutlicher sichtbar wird. Zwischen einer solchen kompakten Beschreibung und einem zuverlässigen Computerprogramm zur Lösung der entsprechenden Aufgaben liegt jedoch noch ein großer Schritt, vgl. Kapitel 16.

Numerische Algorithmen werden im Bereich der reellen Zahlen und exakten Rechenoperationen entwickelt. Bei der Realisierung auf einem Computer mit dem Zahlenbereich $\mathfrak{R}$ und der zugehörigen Computerarithmetik fl müssen jedoch alle Eingangsdaten und Zwischenergebnisse durch ihre Computerdarstellung und alle arithmetischen Operationen durch ihre Computerrealisierungen ersetzt werden. Bei iterativen Algorithmen, die theoretisch unendlich viele Iterationsschritte erfordern, um die gesuchte Lösung als Grenzwert der Iteriertenfolge zu liefern, muß außerdem ein geeignetes *Abbruchkriterium* vorhanden sein, das nach endlich vielen Schritten eine genügend genaue Näherung akzeptiert. Der konzeptionelle Algorithmus $\boldsymbol{V}_0$ geht damit in seine *Computerrealisierung* $\boldsymbol{V} : \mathfrak{E} \subset \mathfrak{R}^N \to \mathfrak{R}^M$ über. Mit $\mathfrak{R}^N$ ist hier die Menge aller N-dimensionalen Vektoren mit Computerzahlen aus $\mathfrak{R}$ als Komponenten bezeichnet worden, und $\mathfrak{E} \subset \mathfrak{R}^N$ ist der *Definitionsbereich* der Computerrealisierung $\boldsymbol{V}$.

Bei der Ausführung von $\boldsymbol{V}$ auf einem Computer entstehen notwendig *Rundungsfehler*, so daß selbst bei exakt darstellbaren Eingangsdaten $\boldsymbol{e} \in \mathfrak{E} \subset \mathfrak{R}^N$ und einem

endlichen konzeptionellen Algorithmus ein von $\boldsymbol{a}^*$ verschiedener Wert $\boldsymbol{a} = \boldsymbol{V}(\boldsymbol{e}) \in \mathfrak{R}^M$ berechnet wird. Die Erfassung der bei der Realisierung von $\boldsymbol{V}_0$ — d. h. beim Übergang von $\boldsymbol{V}_0$ zu $\boldsymbol{V}$ — erzeugten Rundungsfehler ist Gegenstand der *Fehleranalyse* des Algorithmus $\boldsymbol{V}_0$. *Das Verhalten gegenüber Rundungsfehlern ist ein wesentliches Charakteristikum eines numerischen Algorithmus.*

Neben dem Fehlerverhalten spielt der *Aufwand*, d. h. die *Anzahl der* auszuführenden *arithmetischen Operationen* und die *Anzahl der* für das Zwischenspeichern von Hilfsgrößen erforderlichen *Speicherplätze* S bei der Einschätzung von Algorithmen eine entscheidende Rolle. Wir werden später sehen, daß in der numerischen linearen Algebra Operationen des Typs $r := s + u * v$, also eine Multiplikation, der eine Addition folgt, am häufigsten vorkommen. Diese Kombination soll mit „opms" bezeichnet werden. Für eine einzelne Addition/Subtraktion, Multiplikation/Division bzw. Quadratwurzelberechnung sollen die Bezeichnungen „ops", „opm" bzw. „opr" verwendet werden.

Im folgenden werden wir einige einfache Basisalgorithmen der linearen Algebra genauer untersuchen und dabei wesentliche Begriffe und Prinzipien der Fehleranalyse herausarbeiten. *Dabei wird generell vorausgesetzt, daß bei der Realisierung weder Über- noch Unterlauf eintritt.*

A. *Die Summe*

Aufgabe: Für die Eingangsdaten $x_i \in \mathfrak{R}$ $(i = 1, \ldots, n)$ ist die Summe

$$z^* = \sum_{i=1}^{n} x_i$$

zu berechnen.

2.3.1. *Algorithmus zur Summation von n Zahlen.*
$z := 0$
for $i := 1(1)n$ **do** $z := z + x_i$
Aufwand: n ops (Wir geben nur solche Aufwandsgrößen an, die mindestens proportional zu n sind.)

Hier und im folgenden soll das Zeichen „:=" bedeuten, daß der rechtsstehende Ausdruck in der jeweils betrachteten Computerarithmetik fl *in der durch Klammern bzw. die Vorrangsregeln festgelegten Reihenfolge ausgewertet und das Computerresultat der linksstehenden Variablen als Wert zugewiesen wird.* Die Anweisung

$$r := s + u * v$$

ist daher im Sinne von

$$r = \mathrm{fl}\left(s + \mathrm{fl}\,(u * v)\right)$$

zu verstehen. Wir schreiben auch kurz

$$r = \mathrm{fl}\,(s + u * v),$$

da auf Grund der Vorrangregeln klar ist, daß das Produkt fl $(u * v)$ zuerst berechnet werden muß. Analog wird bei komplizierteren Ausdrücken verfahren.

Wenn die bei der Computerrealisierung von 2.3.1 nach dem i-ten Zyklus in z aufgelaufene Teilsumme mit z_i bezeichnet wird, läßt sich der reale Prozeß rekursiv durch

$$z_0 = 0,$$
$$z_i = \mathrm{fl}\,(z_{i-1} + x_i) \qquad (i = 1, \ldots, n) \tag{1}$$

beschreiben, und $z = z_n$ ist der berechnete Wert für die exakte Summe z^*.

Zur Erfassung der Rundungsfehler wenden wir 2.2.11 auf die Vorschrift (1) an und erhalten

$$z_i = (z_{i-1} + x_i)\,(1 + \varepsilon_i) \quad \text{mit} \quad |\varepsilon_i| \leqq \nu, \tag{2}$$

wobei die bei der i-ten Addition gemäß (2.2.7) auftretende Störung ε mit ε_i bezeichnet worden ist.

Aus (2) ergibt sich $z_1 = x_1(1 + \varepsilon_1)$, $z_2 = x_1(1 + \varepsilon_1)\,(1 + \varepsilon_2) + x_2(1 + \varepsilon_2)$, allgemein

$$z_j = \sum_{i=1}^{j} x_i \left\{ \prod_{k=i}^{j} (1 + \varepsilon_k) \right\} \qquad (j = 1, \ldots, n). \tag{3}$$

Zur Abschätzung der Produkte $\prod (1 + \varepsilon_k)$ stellen wir die folgende Aussage bereit:

2.3.2. *Aussage.* Es gelte $|\varepsilon_k| \leqq \nu$ $(k = 1, \ldots, m)$. Dann ist

$$\prod_{k=1}^{m} (1 + \varepsilon_k) = 1 + \varrho^{(m)} \quad \text{mit} \quad \varrho^{(m)} = \sum_{k=1}^{m} \varepsilon_k + O(\nu^2). \tag{4}$$

Im Fall $m\nu < 2$ gelten die Abschätzungen

$$|\varrho^{(m)}| \leqq \frac{m\nu}{1 - 0.5m\nu} \doteqdot m\nu. \tag{5}$$

Beweis. Aus

$$\prod_{k=1}^{m} (1 + \varepsilon_k) = 1 + \sum_{k=1}^{m} \varepsilon_k + \sum_{\substack{k,l=1 \\ k<l}}^{m} \varepsilon_k\varepsilon_l + \sum_{\substack{k,l,j=1 \\ k<l<j}}^{m} \varepsilon_k\varepsilon_l\varepsilon_j + \cdots + \varepsilon_1\varepsilon_2 \cdots \varepsilon_m \tag{6}$$

ergibt sich sofort (4). Zum Nachweis von (5) beachten wir, daß die erste Summe in (6) aus m Summanden, die zweite aus $m(m-1)/(1 \cdot 2)$, die dritte aus $m(m-1)\,(m-2)/(1 \cdot 2 \cdot 3)$ besteht usw. Im Fall $q := 0.5m\nu < 1$ ist daher

$$|\varrho^{(m)}| \leqq \frac{m\nu}{1} + \frac{m\nu(m-1)\,\nu}{1 \cdot 2} + \frac{m\nu(m-1)\,\nu(m-2)\,\nu}{1 \cdot 2 \cdot 3} + \cdots + \frac{m\nu(m-1)\,\nu \cdots 2\nu\,1\nu}{1 \cdot 2 \cdots (m-1)\,m}$$

$$\leqq m\nu(1 + q + q^2 + \cdots + q^{m-1}) \leqq m\nu(1 + q + q^2 + \cdots) = m\nu/(1 - q). \quad \square$$

Im Fall $m\nu \leqq 0.1$ ist $1/(1 - 0.5m\nu) \leqq 1.06$, also $|\varrho^{(m)}| \leqq 1.06m\nu$. *Im folgenden werden wir daher stets* $m\nu \leqq 0.1$ *voraussetzen und bei der Abschätzung von* $\varrho^{(m)}$ *die in erster Ordnung gültige einfachere Schranke* $m\nu$ *statt der exakten Schranke* $1.06m\nu$ *verwenden.* Für reale Dimensionen und Computer ist $m\nu$ in der Regel wesentlich kleiner, so daß dieses Vorgehen um so mehr berechtigt ist. Für $m \leqq 1000$ und $\nu = 9.5 \times 10^{-7}$

(einfachgenaue Arithmetik bei ESER- und IBM-Computern) ist z. B. $1/(1-0.5mv) \leqq 1.0005$!

Aussage 2.3.2 besagt: *Das Produkt aus m Faktoren* $(1+\varepsilon_k)$ *mit* $|\varepsilon_k| \leqq v$ *kann durch einen Faktor* $(1+\varrho^{(m)})$ *derselben Form ersetzt werden, wobei sich die Schranke von* $\varrho^{(m)}$ *auf das m-fache der Schranke der* ε_k *erhöht. In erster Näherung ist* $\varrho^{(m)}$ *gleich der Summe der Störungen* ε_k.

Für $j = n$ ergibt sich aus (3) unter Beachtung von 2.3.2 und wegen $\varepsilon_1 = 0$ — die erste Addition $z_1 = z_0 + x_1 = 0 + x_1$ wird fehlerfrei ausgeführt — die folgende Fehleraussage.

2.3.3. *Aussage.* Die zum Fehlerniveau v gemäß 2.3.1 berechnete Summe z läßt sich als

$$z = \sum_{i=1}^{n} x_i(1+\varepsilon_i^{(n)}) \tag{7}$$

mit relativen Störungen

$$\varepsilon_i^{(n)} = \prod_{k=i}^{n}(1+\varepsilon_k) - 1 = \sum_{k=i}^{n}\varepsilon_k + O(v^2), \qquad |\varepsilon_k| \leqq v, \tag{8}$$

darstellen. Die Störungen sind klein im Sinne von

$$|\varepsilon_i^{(n)}| \leqq \min(n-i+1, n-1)\, v \leqq (n-1)\, v, \tag{9}$$

und der erzeugte Rundungsfehler $z^* - z$ genügt der Ungleichung

$$|z^* - z| \leqq \left\{\sum_{i=1}^{n} \min(n-i+1, n-1)\,|x_i|\right\} v \leqq \left\{(n-1)\sum_{i=1}^{n}|x_i|\right\} v. \tag{10}$$

Die Abschätzungen (10) stellen sog. *a-priori-Schranken* dar, denn sie enthalten nur die Eingangsdaten x_i der Aufgabe. Eine i. allg. wesentlich günstigere *a-posteriori-Schranke*, die mit erst im Laufe oder nach Beendigung der Rechnung verfügbaren Größen arbeitet, kann wie folgt erhalten werden: Aus (2) folgt

$$z_i = z_{i-1} + x_i + (z_{i-1} + x_i)\,\varepsilon_i = z_{i-1} + x_i + z_i\varepsilon_i/(1+\varepsilon_i).$$

Für $\delta_i = \varepsilon_i/(1+\varepsilon_i)$ gilt $|\delta_i| \leqq v/(1-v) \doteqdot v$, so daß Summation auf

$$z = z_n - z_0 = \sum_{i=1}^{n}(z_i - z_{i-1}) = \sum_{i=1}^{n}(x_i + z_i\delta_i) = z^* + \sum_{i=1}^{n}\delta_i z_i,$$

also

$$|z^* - z| \leqq v\sum_{i=1}^{n}|z_i| \tag{11}$$

führt. Die Schranke (11) kann leicht in derselben Schleife wie z selbst berechnet werden.

2.3.4. *Bemerkungen.* (i) Die linke Schranke in (10) für den erzeugten Rundungsfehler wird minimal, wenn die x_i in betragsmäßig wachsender Folge summiert werden.

(ii) Der relative Rundungsfehler $|z^* - z|/|z^*|$ kann bei Summation von Zahlen unterschiedlichen Vorzeichens beliebig groß werden.

(iii) Der relative Rundungsfehler ist bei Summation von Zahlen einheitlichen Vorzeichens klein:

$$|z^* - z|/|z^*| \leqq (n-1)\nu. \quad \square$$

Dies ergibt sich sofort aus (10) und der Tatsache, daß der Quotient $\sum |x_i|/|\sum x_i|$ bei unterschiedlichem Vorzeichen der x_i beliebig groß werden kann, bei einheitlichem Vorzeichen jedoch den Wert 1 hat.

Aus 2.3.3 folgt ferner, daß die relativen Störungen $\varepsilon_i^{(n)}$ in erster Ordnung gleich der Summe $\varepsilon_i + \varepsilon_{i+1} + \cdots + \varepsilon_n$ der relativen Rundungsfehler der einzelnen Additionen sind. Bei symmetrischer Rundung werden die ε_k unterschiedliches Vorzeichen haben und können sich auslöschen. Unter geeigneten stochastischen Voraussetzungen wird dann statt der im ungünstigsten Fall zu erwartenden Relation $|\varepsilon_i^{(n)}| \doteqdot (n-i+1)\nu$ im Mittel ein Wert $|\varepsilon_i^{(n)}|$ in der Größenordnung von $\sqrt{(n-i+1)/(3\beta)}\,\nu$ zu erwarten sein, vgl. 2.2.8. Bei unsymmetrischer Rundung sind alle ε_k vom selben Vorzeichen, so daß der Auslöschungseffekt nicht auftreten kann.

In 2.3.4 wurde festgestellt, daß es Eingangsdaten $\{x_i\}$ gibt, die bei der Summation zu beliebig großen relativen Rundungsfehlern $|z^* - z|/|z^*|$ des Resultats führen, d. h., die berechnete Summe wird sehr ungenau. Es liegt daher die Frage nahe, ob diese Ungenauigkeit durch das verwendete Summationsverfahren oder durch andere Ursachen hervorgerufen wird. Zur Beantwortung dieser Frage erinnern wir daran, daß die Computerzahl $x \in \mathfrak{R}$ nicht nur sich selbst, sondern alle reellen Zahlen $\tilde{x}$ mit $x = \mathrm{rd}(\tilde{x})$ repräsentiert, vgl. Abschnitt 2.2. Die dadurch bedingte maximale relative Unsicherheit in x kann dann für $x \neq 0$ im günstigsten Fall den Wert ν/β, im ungünstigsten den Wert ν erreichen in Abhängigkeit von der Größenordnung der Mantisse, siehe 2.2.5. Da für reale Computer β nicht groß ist und wir auch die ungünstigste Situation erfassen müssen, sehen wir etwas vereinfachend die folgende Annahme als erfüllt an:

2.3.5. *Annahme.* Jedes Eingangsdatenelement $x \in \mathfrak{R}$, $x \neq 0$, enthält eine natürliche Unsicherheit auf dem relativen Fehlerniveau ν, die durch den Darstellungsfehler hervorgerufen wird. Durch $x \in \mathfrak{R}$, $x \neq 0$, werden alle reellen Zahlen $\tilde{x}$ mit

$$\tilde{x} = x(1+\vartheta), \qquad |\vartheta| \leqq \nu, \tag{12}$$

repräsentiert.

Im Fall der Summation repräsentiert also der Eingangsdatensatz $\{x_i\}_{i=1}^n$, $x_i \in \mathfrak{R}$, alle reellen Eingangsdaten

$$\tilde{x}_i = x_i(1+\vartheta_i) \quad \text{mit} \quad |\vartheta_i| \leqq \nu \qquad (i = 1, \ldots, n). \tag{13}$$

Die zu $\{\tilde{x}_i\}$ gehörende exakte Summe ist $\tilde{z} = \sum_{i=1}^{n} \tilde{x}_i$.

Nach diesen Vorbereitungen können wir die gestellte Frage beantworten, indem wir die Abschätzungen aus 2.3.3 in zweierlei Weise interpretieren.

Erste Interpretation: Unter Beachtung von (13) folgt

$$\tilde{z} = \sum_{i=1}^{n} x_i(1 + \vartheta_i) \quad \text{mit} \quad |\vartheta_i| \leqq \nu, \tag{14}$$

aus 2.3.3 liest man dagegen

$$z = \sum_{i=1}^{n} x_i(1 + \varepsilon_i^{(n)}) \quad \text{mit} \quad |\varepsilon_i^{(n)}| \leqq (n-1)\,\nu \tag{15}$$

ab. Ein Vergleich von (14) und (15) führt auf das folgende Resultat:

2.3.6. *Aussage.* Die mittels des Verfahrens 2.3.1 berechnete Summe z ist die exakte Summe der gestörten Eingangsdaten $\{x_i(1 + \varepsilon_i^{(n)})\}$, wobei $|\varepsilon_i^{(n)}| \leqq (n-1)\,\nu$ ist. Die durch die Rundungsfehler erzeugten Störungen $\varepsilon_i^{(n)}$ vergrößern die in den Eingangsdaten enthaltenen Unsicherheiten also höchstens um den Faktor $n - 1$.

Zweite Interpretation: Ein durch $\{x_i\}$ repräsentierter Eingangsdatensatz $\{\tilde{x}_i\}$ liefert die exakte Summe $\tilde{z}$, wobei wegen (14)

$$|\tilde{z} - z^*| = \left| \sum_{i=1}^{n} (\tilde{x}_i - x_i) \right| = \left| \sum_{i=1}^{n} x_i \vartheta_i \right| \leqq \nu \sum_{i=1}^{n} |x_i| \tag{16}$$

gilt. Diese Abschätzung ist in bezug auf (13) scharf, d. h., es gibt Störungen $|\vartheta_i| = \nu$, für die das Gleichheitszeichen steht. Die rechte Seite von (16) gibt also das unvermeidliche absolute Fehlerniveau $\Delta z_{\min}(x_i, \nu)$ der Summe an, falls jeder Summand im Rahmen des relativen Fehlerniveaus ν gegeben ist, und $\sum_{i=1}^{n} |x_i|$ ist die entsprechende Konditionszahl, vgl. Abschnitt 2.2.

Der Fehler von z kann nach 2.3.3 gemäß

$$|z^* - z| \leqq (n-1)\,\nu \sum_{i=1}^{n} |x_i| \tag{17}$$

abgeschätzt werden. Der Vergleich von (16) und (17) führt auf die folgende Aussage.

2.3.7. *Aussage.* Der im Verfahren 2.3.1 erzeugte Rundungsfehler von z ist höchstens das $(n-1)$-fache des unvermeidlichen Fehlerniveaus

$$\Delta z_{\text{opt}}(x_i, \nu) = \nu \sum_{i=1}^{n} |x_i|,$$

das durch die Unsicherheit der Eingangsdaten hervorgerufen wird.

Beide Interpretationen führen uns zu dem Ergebnis, daß die in den Eingangsdaten $\{x_i\}_{i=1}^{n}$ enthaltene Unsicherheit durch das Verfahren 2.3.1 nur quantitativ — und zwar höchstens um den Faktor $F = n - 1$ — vergrößert wird. Es ist jedoch wesentlich, den Unterschied zwischen beiden Interpretationen zu erkennen.

Erste Interpretation: *Das technische Mittel ist die Konstruktion von Störungen $\{\varepsilon_i^{(n)}\}$ der Eingangsdaten $\{x_i\}$ derart, daß das berechnete Ergebnis gerade das exakte Ergebnis*

zu den gestörten Eingangsdaten ist, und die Angabe von Schranken für diese Störungen. Dann werden die Schranken der Störungen mit dem Niveau des Darstellungsfehlers verglichen. *Die entsprechende Philosophie besteht darin, die während der Rechnung entstehenden Rundungsfehler als Störungen der Eingangsdaten zu interpretieren und mit deren natürlicher Unsicherheit zu vergleichen.* Dies ist die Idee der sog. *Rückwärtsanalyse* (engl. „backward error analysis").

Zweite Interpretation: *Das technische Mittel ist die Konstruktion von möglichst realistischen Abschätzungen sowohl für das unvermeidliche Fehlerniveau $\Delta z_{\mathrm{opt}}(x_i, \nu)$ des exakten Resultates z^* als auch für den durch das Verfahren erzeugten Rundungsfehler $|z^* - z|$.* Dann werden diese beiden Abschätzungen miteinander verglichen. *Die entsprechende Philosophie besteht darin, die durch Rechnung erzeugten Rundungsfehler in den Ausgangsdaten mit derem unvermeidlichen Fehlerniveau zu vergleichen, das durch die Unsicherheit der Eingangsdaten bedingt ist.* Das ist die Idee der sog. *Vorwärtsanalyse.*

Auf Grund von Aussage 2.3.6 (Ergebnis der Rückwärtsanalyse) werden wir das Summationsverfahren 2.3.1 als *numerisch gutartig mit der Fehlerkumulationskonstanten $F = n - 1$* bezeichnen, denn das berechnete Ergebnis ist das exakte Ergebnis zu gestörten Eingangsdaten, deren Störungsniveau höchstens das F-fache der Darstellungsunsicherheit erreicht. Auf Grund von Aussage 2.3.7 (Ergebnis der Vorwärtsanalyse) wird das Verfahren *numerisch stabil mit der Fehlerkumulationskonstanten $F = n - 1$* genannt, weil der durch das Verfahren erzeugte Rundungsfehler von z höchstens das F-fache des durch die Unsicherheit der Eingangsdaten bedingten unvermeidlichen Fehlers $\Delta z_{\mathrm{opt}}(x_i, \nu)$ betragen kann.

Aus den bisherigen Ausführungen wird klar: *Die numerische Gutartigkeit* (mit nicht zu großem F) *stellt die bestmögliche Qualität eines Algorithmus dar,* weil dann die in den Eingangsdaten enthaltene Information bestmöglich auf die Ausgangsdaten übertragen wird. Unter den gutartigen Algorithmen sind selbstverständlich diejenigen mit den kleinsten Fehlerkumulationskonstanten bezüglich ihrer Rundungsfehlerempfindlichkeit die günstigsten, allerdings muß ein kleineres F in der Regel durch einen höheren Aufwand erkauft werden.

Für genügend kleines ν und lokal lipschitzstetige Aufgabenklassen zieht die numerische Gutartigkeit die numerische Stabilität nach sich; die Umkehrung gilt i. allg. natürlich nicht. *Numerische Stabilität ist eine Mindestforderung an ein vernünftiges Verfahren. Liegt sie nicht vor, so können für gewisse Probleme beliebig große Genauigkeitsverluste in bezug auf das unvermeidliche Fehlerniveau auftreten.* Man beachte dabei, daß auch ein stabiles Verfahren beliebig große Rundungsfehler erzeugen kann, nämlich dann, wenn das unvermeidliche Fehlerniveau entsprechend hoch ist. Es wäre auch unangemessen zu verlangen, daß ein Verfahren die in den Daten bereits enthaltene Unsicherheit eliminiert.

Zur Illustration betrachten wir das folgende Beispiel.

2.3.8. *Beispiel.* In $\mathfrak{R}(10, 4, 9, 9)$ ist bei Rechnung mit unsymmetrischer Rundung (Abbrechen nach der vierten Mantissenstelle) die Summe $z^* = \sum_{i=1}^{6} x_i$ für die in der Tabelle angegebenen Summanden zu berechnen.

i	x_i	x_i'	$\varepsilon_i^{(6)} = (x_i' - x_i)/x_i$
1	0.1065×10^2	0.1060263×10^2	-4.4×10^{-3}
2	0.1298×10^0	0.12922×10^0	-4.5×10^{-3}
3	-0.5102×10^{-1}	-0.50840×10^{-1}	-3.5×10^{-3}
4	0.7982×10^{-1}	0.79600×10^{-1}	-2.8×10^{-3}
5	0.1399×10^0	0.13965×10^0	-1.8×10^{-3}
6	0.5979×10^{-1}	0.59740×10^{-1}	-0.8×10^{-3}

Man erhält $z = 10.96$, die exakte Summe ist $z^* = 11.00829$. Der erzeugte Rundungsfehler ist $\delta z = z^* - z = 0.04829$, und das durch den Darstellungsfehler $v = 10^{-3}$ bedingte unvermeidliche Fehlerniveau von z ist $\Delta z_{\text{opt}}(x_i, v) = v \sum_{i=1}^{6} |x_i| = 0.011$. Die tatsächliche Fehlerkumulation ist $|\delta z|/\Delta z_{\text{opt}} = 4.35$, also 87% des nach Aussage 2.3.7 maximal möglichen Wertes $F = n - 1 = 5$. Zum Vergleich sind in der zweiten Spalte gestörte Summanden $x_i' = x_i(1 + \varepsilon_i^{(6)})$ angegeben worden, deren Störungen $\varepsilon_i^{(6)}$ den Abschätzungen (9) genügen und deren exakte Summe $\sum_{i=1}^{6} x_i'$ gerade den Wert $z = 10.96$ ergibt. Wir bemerken schließlich, daß Summation der x_i nach wachsenden Beträgen auf den Wert 11.00, also das unsymmetrisch gerundete exakte Ergebnis führt. □

Für ein günstigeres, aber auch aufwendigeres Summationsverfahren mit $F = \log_2 n$ siehe Ü 2.3.1.

B. *Das Skalarprodukt*

Aufgabe: Für die Eingangsdaten $x_i, y_i \in \mathfrak{R}$ $(i = 1, \ldots, n)$ ist das Skalarprodukt

$$s^* = \sum_{i=1}^{n} x_i y_i$$

zu berechnen.

Bezeichnet $\mathfrak{R}^n$ wie eingangs die Menge aller Vektoren $\boldsymbol{x} = (x_1, \ldots, x_n)^T$ mit $x_i \in \mathfrak{R}$ $(i = 1, \ldots, n)$, so läßt sich die Aufgabe wie folgt formulieren:

Für $\boldsymbol{x}, \boldsymbol{y} \in \mathfrak{R}^n$ ist $s^* = \boldsymbol{x}^T\boldsymbol{y}$ zu berechnen.

2.3.9. *Algorithmus zur Berechnung des Skalarproduktes.*

$s := 0$

for $i := 1(1)n$ **do** $s := s + x_i * y_i$

Aufwand: n opms

Zur Rundungsfehleranalyse schreiben wir 2.3.9 in der äquivalenten Form

$$s_0 = 0,$$

$$s_i = \text{fl}\,(s_{i-1} + g_i), \qquad g_i = \text{fl}\,(x_i * y_i) \qquad (i = 1, \ldots, n).$$

Analog zum Vorgehen bei der Summation folgt unter Beachtung von 2.2.11

$$s_i = (s_{i-1} + g_i)(1 + \varepsilon_i), \qquad |\varepsilon_i| \leqq \nu, \tag{18}$$

$$g_i = x_i y_i(1 + \delta_i), \qquad |\delta_i| \leqq \nu \qquad (i = 1, \ldots, n). \tag{19}$$

Die Rekursion (18) ist bis auf die Bezeichnung identisch mit der Rekursion (2) bei der Summation. Aus 2.3.3 folgt daher

$$s = s_n = \sum_{i=1}^{n} g_i(1 + \varepsilon_i^{(n)}),$$

und $\varepsilon_i^{(n)}$ genügt den Abschätzungen (9). Einsetzen von (19) führt auf

$$s = \sum_{i=1}^{n} x_i y_i(1 + \delta_i)(1 + \varepsilon_i^{(n)}).$$

Nun gilt

$$(1 + \delta_i)(1 + \varepsilon_i^{(n)}) = 1 + \varphi_i^{(n)} \quad \text{mit} \quad \varphi_i^{(n)} = \delta_i + (1 + \delta_i)\,\varepsilon_i^{(n)}, \tag{20}$$

woraus wegen $1 + \nu \doteqdot 1$ und (9) sofort

$$|\varphi_i^{(n)}| \leqq \nu + (1 + \nu)\,|\varepsilon_i^{(n)}| \doteqdot \nu + (1 + \nu) \min(n - i + 1, n - 1)\,\nu$$
$$= \min(n - i + 2, n)\,\nu$$

folgt. Schließlich kann

$$(1 + \varphi_i^{(n)}) = (1 + \psi_i^{(n)})^2 \quad \text{mit} \quad |\psi_i^{(n)}| \doteqdot |\varphi_i^{(n)}|/2 \leqq 0.5 \min(n - i + 2, n)\,\nu$$

geschrieben werden, so daß jeder der Faktoren x_i und y_i mit einer Störung belastet werden kann.

2.3.10. *Rückwärtsanalyse der Skalarproduktberechnung.* Das gemäß Algorithmus 2.3.9 zum Rundungsfehlerniveau ν berechnete Skalarprodukt s ist das exakte Skalarprodukt

$$s = \sum_{i=1}^{n} x_i y_i(1 + \varphi_i^{(n)}) = \sum_{i=1}^{n} [x_i(1 + \psi_i^{(n)})]\,[y_i(1 + \psi_i^{(n)})] \tag{21}$$

mit gestörten Eingangsdaten, und die Störungen genügen den Abschätzungen

$$\left.\begin{matrix} |\varphi_i^{(n)}| \\ 2|\psi_i^{(n)}| \end{matrix}\right\} \doteqdot \min(n - i + 2, n)\,\nu \leqq n\nu. \tag{22}$$

Das Verfahren 2.3.9 ist also in bezug auf die Eingangsdaten $\boldsymbol{x}, \boldsymbol{y} \in \mathbf{R}^n$ numerisch gutartig mit $F = n/2$.

Wird $\boldsymbol{x}$ als fester Parameter und nur $\boldsymbol{y}$ als Eingangsdatenelement angesehen, so kann $s = \sum_{i=1}^{n} x_i[y_i(1 + \varphi_i^{(n)})]$ geschrieben werden, d. h., nur $\boldsymbol{y}$ braucht mit Störungen belastet zu werden, die dann allerdings unter Umständen doppelt so groß werden können. In dieser Interpretation — d. h. in bezug auf $\boldsymbol{y}$ — liegt also ebenfalls nume-

rische Gutartigkeit vor, aber mit $F = n$. In den späteren Anwendungen von 2.3.10 werden wir immer nur einen der Faktoren $\boldsymbol{x}, \boldsymbol{y}$ mit Fehlern belasten, weil dann die Abschätzungen einfacher werden.

Zur Vorwärtsanalyse bestimmen wir zunächst das durch den Darstellungsfehler hervorgerufene unvermeidliche Fehlerniveau Δs_{opt} und nehmen dazu an, daß durch x_i, y_i auch alle $\tilde{x}_i, \tilde{y}_i$ mit

$$\tilde{x}_i = x_i(1 + \xi_i), \qquad \tilde{y}_i = y_i(1 + \eta_i), \qquad |\xi_i|, |\eta_i| \leqq \nu \qquad (i = 1, \ldots, n)$$

repräsentiert werden. Mit $\tilde{s} = \tilde{\boldsymbol{y}}^{\mathsf{T}}\tilde{\boldsymbol{x}}$ ergibt sich dann

$$\begin{aligned} |\tilde{s} - s^*| &= \left| \sum_{i=1}^{n} \{x_i y_i (1 + \xi_i)(1 + \eta_i) - x_i y_i\} \right| \\ &\leqq \sum_{i=1}^{n} \{|\xi_i| + |\eta_i| + |\xi_i|\,|\eta_i|\}\,|x_i|\,|y_i| \\ &\leqq 2\nu \sum_{i=1}^{n} |x_i y_i| = 2\nu\,|\boldsymbol{x}|^{\mathsf{T}}|\boldsymbol{y}| \leqq 2\nu\,\|\boldsymbol{x}\|_2\,\|\boldsymbol{y}\|_2, \end{aligned}$$

vgl. Beispiel 2.1.10.

Aus (21) folgt weiter

$$|s^* - s| = \left| \sum_{i=1}^{n} x_i y_i \varphi_i^{(n)} \right| \leqq \sum_{i=1}^{n} |x_i y_i|\,|\varphi_i^{(n)}|,$$

mit (22) also das folgende Resultat.

2.3.11. *Vorwärtsanalyse der Skalarproduktberechnung.* Der durch Algorithmus 2.3.9 erzeugte Rundungsfehler $s^* - s$ genügt der Ungleichung

$$|s^* - s| \leqq \left\{ \sum_{i=1}^{n} \min(n - i + 2, n)\,|x_i y_i| \right\} \nu \leqq \left\{ n \sum_{i=1}^{n} |x_i y_i| \right\} \nu \leqq n\|\boldsymbol{x}\|_2\,\|\boldsymbol{y}\|_2\,\nu. \tag{23}$$

Das durch die Unsicherheit der Eingangsdaten hervorgerufene unvermeidliche Fehlerniveau ist

$$\Delta s_{\text{opt}}(x_i, y_i; \nu) \doteq \left\{ 2 \sum_{i=1}^{n} |x_i y_i| \right\} \nu \leqq 2\|\boldsymbol{x}\|_2\,\|\boldsymbol{y}\|_2\,\nu. \tag{24}$$

Das Verfahren 2.3.9 ist also numerisch stabil mit der Konstanten $F = n/2$.

2.3.12. *Bemerkung.* (i) Aus (20) und (8) folgt

$$\varphi_i^{(n)} = \delta_i + \varepsilon_i^{(n)} + O(\nu^2) = \delta_i + \varepsilon_i + \varepsilon_{i+1} + \cdots + \varepsilon_n + O(\nu^2), \tag{25}$$

wobei δ_i und ε_i durch (18), (19) erklärt sind. Bei symmetrischer Rundung ist daher zu erwarten, daß bei der Summation der $n - i + 2$ relativen Rundungsfehler δ_i, $\varepsilon_i, \ldots, \varepsilon_n$ stochastische Auslöschung eintritt.

(ii) Aus (23) folgt, daß bei einheitlichem Vorzeichen der Teilprodukte $x_i y_i$ der relative Fehler von s klein ist: $|s^* - s|/|s^*| \leqq n\nu$. Bei unterschiedlichem Vorzeichen

der Teilprodukte kann der relative Fehler von s jedoch beliebig groß werden, vgl. Beispiel 2.1.10.

(iii) Wird $\boldsymbol{x}$ als exakter Parameter betrachtet, so ergibt sich bezüglich der Darstellungsfehler von $\boldsymbol{y}$ die Schranke $\Delta s_{\mathrm{opt}}(y_i, \nu) = \nu \sum_{i=1}^{n} |x_i|\,|y_i|$, also der halbe Wert von (24). Bei festem $\boldsymbol{x}$ liegt dann bezüglich $\boldsymbol{y}$ Stabilität mit $F = n$ vor in Übereinstimmung zur Nachbemerkung von 2.3.10. □

Eine Möglichkeit zur Abschätzung des erzeugten Rundungsfehlers $|s^* - s|$ gibt die nachfolgende, analog zu (11) abgeleitete a-posteriori-Schranke: Aus (18), (19) folgt

$$\begin{aligned} s_i &= (s_{i-1} + g_i)(1 + \varepsilon_i) = s_{i-1} + g_i + (s_{i-1} + g_i)\,\varepsilon_i \\ &= s_{i-1} + g_i + s_i\varepsilon_i/(1 + \varepsilon_i) \end{aligned}$$

sowie

$$g_i = x_i y_i(1 + \delta_i) = x_i y_i + x_i y_i \delta_i = x_i y_i + g_i\delta_i/(1 + \delta_i),$$

also insgesamt

$$s_i = s_{i-1} + x_i y_i + g_i\delta_i/(1 + \delta_i) + s_i\varepsilon_i/(1 + \varepsilon_i).$$

Summation führt direkt auf

$$|s^* - s| = \left| \sum_{i=1}^{n} \{g_i\delta_i/(1 + \delta_i) + s_i\varepsilon_i/(1 + \varepsilon_i)\} \right| \leqq \nu \sum_{i=1}^{n} \{|g_i| + |s_i|\} =: d^*. \quad (26)$$

Der folgende Algorithmus ist eine Erweiterung von 2.3.9, in der gleichzeitig mit s die Schranke d durch Auswertung von (26) mit berechnet wird.

2.3.13. *Algorithmus zur Berechnung des Skalarproduktes und einer a-posteriori-Fehlerschranke.*

$s := d := 0$

for $i := 1(1)n$ **do**

$\quad g := x_i * y_i$

$\quad s := s + g$

$\quad d := d + |g| + |s|$

$d := \nu * d$

Aufwand: n opms + $2n$ ops, also $2n$ ops mehr als für Algorithmus 2.3.9.

Auch für schlechtkonditionierte Aufgaben, bei denen der Quotient $\sum |x_i y_i| / |\sum x_i y_i|$ bzw. $\|\boldsymbol{x}\|_2\,\|\boldsymbol{y}\|_2 / |\boldsymbol{x}^{\mathrm{T}}\boldsymbol{y}|$ sehr groß wird, vgl. 2.3.11 und 2.1.10, kann eine hohe relative Genauigkeit erreicht werden, indem die Teilprodukte $x_i y_i$ in höherer Genauigkeit berechnet und aufsummiert werden. Zusätzlich zur Standard-Gleitpunktarithmetik-fl mit dem Fehlerniveau ν muß dazu eine Arithmetik fl_1 mit dem kleineren Fehlerniveau $\nu_1 \ll \nu$ verfügbar sein. Zur Illustration betrachten wir das folgende Beispiel.

2.3.14. *Beispiel.* Berechnung von $s^* = \sum_{i=1}^{3} x_i y_i$ für die unten angegebenen Eingangsdaten x_i, $y_i \in \mathfrak{R} = \mathfrak{R}(10, 3, 9, 9)$. Die Standardarithmetik fl ist die in $\mathfrak{R}$ mit unsymmetrischer Rundung. Produktbildung $g_i = \mathrm{fl}_1(x_i * y_i)$ und Summation $s_i = \mathrm{fl}_1(s_{i-1} + g_i)$ erfolgen in der Arithmetik fl_1 von $\mathfrak{R}_1 = \mathfrak{R}(10, 6, 9, 9)$ mit unsymmetrischer Rundung, also mit doppelter Mantissenlänge. Die Fehlerniveaus sind $\nu = 10^{-2}$ bzw. $\nu_1 = 10^{-5}$. Zum Vergleich wurden zusätzlich die in der Standardarithmetik berechneten Werte mit angegeben.

i	x_i	y_i	$\mathrm{fl}(x_i * y_i)$	$\mathrm{fl}(s_{i-1} + g_i)$	$\mathrm{fl}_1(x_i * y_i)$	$\mathrm{fl}_1(s_{i-1} + g_i)$
1	0.341+00	0.107+02	0.364+01	0.364+01	0.364870+01	0.364870+01
2	0.567−01	0.164+00	0.929−02	0.364+01	0.929880−02	0.365799+01
3	−0.518+01	0.706+00	−0.365+01	−0.100−01	−0.365708+01	0.910000−03

Die Schreibweise $-0.518+01$ bedeutet hier und im folgenden stets $-0.518 \times 10^{+01}$ usw. Der exakte Wert ist $s^* = 0.9188 \times 10^{-3}$.

Bei Rechnung in fl ergibt sich $s = -0.100 \times 10^{-1}$, d. h., selbst das Vorzeichen ist falsch. Der relative Rundungsfehler $|s^* - s|/|s^*| = 11.9$ überschreitet das Niveau $\nu = 10^{-2}$ des Darstellungsfehlers der Daten um den Faktor 1190! Der durch Akkumulation der Summen in fl_1 berechnete und danach auf einfache Genauigkeit gerundete Wert $s^{ac} = 0.910 \times 10^{-3}$ hat dagegen einen Rundungsfehler $|s^* - s^{ac}|/|s^*| = 0.96 \times 10^{-2}$ in der Größenordnung von ν. Man beachte jedoch, daß diese hohe Genauigkeit von s^{ac} in bezug auf s^* nur dann einen Informationsgewinn bringt, wenn die Eingangsdaten x_i, y_i exakt bzw. sehr genau sind. Sind die Eingangsdaten dagegen nur im Rahmen des Darstellungsfehlers genau, so müssen alle $\tilde{x}_i$, $\tilde{y}_i$ mit $x_i = \mathrm{rd}(\tilde{x}_i)$ und $y_i = \mathrm{rd}(\tilde{y}_i)$ als gleichwertig angesehen werden. Über der Menge dieser zulässigen Eingangsdaten kann $\tilde{s} = \sum_{1=i}^{3} \tilde{x}_i \tilde{y}_i$ alle Werte zwischen $0.341 \times 10.7 + 0.0567 \times 0.164 - 5.19 \times 0.707 = -0.011331$ und $0.342 \times 10.8 + 0.0568 \times 0.165 - 5.18 \times 0.706 = 0.045892$ annehmen. Im Sinne des Darstellungsfehlers ist daher jedes $\tilde{z}$ zwischen $-0.011\ldots$ und $+0.045\ldots$ als zulässiges Resultat zu akzeptieren, vgl. Abschnitt 2.1. Speziell ist $s^* = 0.009188$ gegenüber $s = -0.01$ weder besser noch schlechter. Zum Beispiel ergibt sich $s = -0.01$ als exakter Wert, wenn x_3 durch $x_3' = -5.189$ und y_3 durch $y_3' = 3.6679988/5.189 = 0.70687\ldots$ ersetzt werden; die Computerdarstellungen der modifizierten Werte stimmen mit den ursprünglichen überein. □

Beispiel 2.3.14 zeigt, daß die Unsicherheit der Eingangsdaten durch keinen noch so genauen Algorithmus eliminiert werden kann. *Skalarproduktberechnung durch Akkumulation in höherer Genauigkeit ist daher nur dann sinnvoll, wenn die Eingangsdaten x, y als exakt angesehen werden können oder wenn n und damit die Fehlerkumulation $F = n/2$ von* 2.3.9 *sehr groß ist.* Im ersten Fall bringt Akkumulation mit höherer Genauigkeit tatsächlich eine wesentlich größere Genauigkeit, allerdings treten exakte Eingangsdaten sehr selten auf. Im zweiten Fall — n groß — kann die Fehlerkumulation durch Akkumulation in höherer Genauigkeit praktisch auf den Wert $F = 1$ gebracht werden. In diesem Fall läßt sich aber auch in der Standardarithmetik fl durch Verwendung der Kaskadensummation aus Ü 2.3.1 zur Summation der Produkte $\mathrm{fl}(x_i * y_i)$ eine Reduktion auf $F = (\log_2 n + 1)/2$ erreichen, also ein wesentlich kleinerer Wert als $F = n/2$ für die Standardsummation. Für $n = 2^{10} = 1024$ ist z. B. $(\log_2 n + 1)/2 = 5.5$, aber $n/2 = 512$!

C. *Darstellung von Vektoren und Matrizen in der Computerarithmetik*

Bei den Aufgabenklassen der linearen Algebra treten in der Regel Vektoren $\boldsymbol{b} = (b_i) \in \mathbf{R}^m$ und Matrizen $\boldsymbol{A} = (a_{ij}) \in \mathbf{R}^{m,n}$ als Eingangsdaten auf. Die zugehörigen Computerdarstellungen sollen mit rd $(\boldsymbol{b})$ bzw. rd $(\boldsymbol{A})$ bezeichnet werden; sie entstehen durch elementweise Rundung, d. h.

$$\text{rd}\,(\boldsymbol{b}) := \big(\text{rd}\,(b_i)\big) \in \mathfrak{R}^m, \qquad \text{rd}\,(\boldsymbol{A}) := \big(\text{rd}\,(a_{ij})\big) \in \mathfrak{R}^{m,n}. \tag{27}$$

Analog zu $\mathfrak{R}^m$ bezeichnet $\mathfrak{R}^{m,n}$ dabei die Menge aller (m,n)-Matrizen, deren Elemente Computerzahlen aus $\mathfrak{R}$ sind. Im Fall

$$b_i = 0 \text{ oder } \text{MIN} \leqq |b_i| \leqq \text{MAX} \quad \text{bzw.} \quad a_{ij} = 0 \text{ oder } \text{MIN} \leqq |a_{ij}| \leqq \text{MAX} \tag{28}$$

folgt aus 2.2.7

$$|\text{rd}\,(b_i) - b_i| \leqq \nu\,|b_i| \quad \text{bzw.} \quad |\text{rd}\,(a_{ij}) - a_{ij}| \leqq \nu\,|a_{ij}| \tag{29}$$
$$(i = 1, \ldots, m;\ j = 1, \ldots, n).$$

Unter Verwendung der in 1.1.J eingeführten Ordnungsbegriffe läßt sich (29) kurz als

$$|\text{rd}\,(\boldsymbol{b}) - \boldsymbol{b}| \leqq \nu\,|\boldsymbol{b}| \quad \text{bzw.} \quad |\text{rd}\,(\boldsymbol{A}) - \boldsymbol{A}| \leqq \nu\,|\boldsymbol{A}|$$

schreiben. Für die Vektornormen mit dem Index $p \in \{1, 2, \infty\}$ bzw. für die Matrixnormen mit dem Index $p \in \{1, 2, \infty, F\}$ folgt hieraus

$$\|\text{rd}\,(\boldsymbol{b}) - \boldsymbol{b}\| \leqq \nu\,\|\boldsymbol{b}\| \quad \text{bzw.} \quad \|\text{rd}\,(\boldsymbol{A}) - \boldsymbol{A}\| \leqq \nu\,\||\boldsymbol{A}|\|,$$

vgl. wieder Abschnitt 1.1.J. Auf der rechten Seite kann $\|\|\boldsymbol{A}\|\|$ für die Matrixnormen mit $p \in \{1, \infty, F\}$ durch $\|\boldsymbol{A}\|$ ersetzt werden, denn diese Normen sind wie die betrachteten Vektornormen absolut und monoton. Für $p = 2$ gilt wegen (1.1.44)

$$\||\boldsymbol{A}|\|_2 \leqq \||\boldsymbol{A}|\|_F = \|\boldsymbol{A}\|_F \leqq \sqrt{\min(m, n)}\,\|\boldsymbol{A}\|_2. \tag{30}$$

Zusammenfassend erhalten wir

2.3.15. *Aussage.* Für $\boldsymbol{b} \in \mathbf{R}^m$ bzw. $\boldsymbol{A} \in \mathbf{R}^{m,n}$ sei (28) erfüllt, d. h., bei Übergang zu den Computerdarstellungen rd $(\boldsymbol{b})$ bzw. rd $(\boldsymbol{A})$ trete weder Über- noch Unterlauf auf. Dann gilt

(i) Die Computerdarstellungen haben eine hohe individuelle Genauigkeit im Sinne von

$$|\text{rd}\,(\boldsymbol{b}) - \boldsymbol{b}| \leqq \nu\,|\boldsymbol{b}| \quad \text{bzw.} \quad |\text{rd}\,(\boldsymbol{A}) - \boldsymbol{A}| \leqq \nu\,|\boldsymbol{A}|. \tag{31}$$

(ii) Die hohe individuelle Genauigkeit überträgt sich auf die Normen im Sinne von

$$\|\text{rd}\,(\boldsymbol{b}) - \boldsymbol{b}\| \leqq \nu\,\|\boldsymbol{b}\| \tag{32}$$

bzw.

$$\|\text{rd}\,(\boldsymbol{A}) - \boldsymbol{A}\| \leqq \begin{cases} \nu\|\boldsymbol{A}\| & \text{für } p = 1, \infty, F, \\ \nu\,\||\boldsymbol{A}|\| \leqq \nu\sqrt{\min(m, n)}\,\|\boldsymbol{A}\| & \text{für } p = 2. \end{cases} \tag{33}$$

2.3.16. *Bemerkung.* (i) Eine im Sinne von

$$\|\tilde{\boldsymbol{b}} - \boldsymbol{b}\| \leqq v\,\|\boldsymbol{b}\| \quad \text{bzw.} \quad \|\tilde{\boldsymbol{A}} - \boldsymbol{A}\| \leqq v\,\||\boldsymbol{A}|\| \tag{34}$$

der Norm nach genaue Darstellung von $\boldsymbol{b}$ bzw. $\boldsymbol{A}$ durch $\tilde{\boldsymbol{b}}$ bzw. $\tilde{\boldsymbol{A}}$ ist gleichbedeutend damit, daß die betragsgroßen Elemente von $\boldsymbol{b}$ bzw. $\boldsymbol{A}$ durch die entsprechenden Elemente von $\tilde{\boldsymbol{b}}$ bzw. $\tilde{\boldsymbol{A}}$ mit hoher individueller Genauigkeit dargestellt werden.

(ii) Für viele Aufgaben der numerischen linearen Algebra kann die im Sinne von (31) hohe individuelle Genauigkeit aller Eingangsdaten nicht ausgenutzt werden. Um akzeptable Resultate zu garantieren, genügt eine der Norm nach hohe Genauigkeit der Eingangsdaten.

Zur Illustration betrachten wir die folgenden Beispiele, siehe auch Ü 2.3.2.

2.3.17. *Beispiel.* Es sei $\mathfrak{R} = \mathfrak{R}(10, 3, 9, 9)$, die Rundung sei unsymmetrisch.

(i) Der Vektor $\boldsymbol{b} = (6.6666, 0.0033333)^{\mathsf{T}} \in \mathbf{R}^2$ wird durch $\mathrm{rd}(\boldsymbol{b}) = (0.666 \times 10^{1}, 0.333 \times 10^{-2})^{\mathsf{T}} \in \mathfrak{R}^2$ mit dem kollektiven relativen Rundungsfehler $\|\mathrm{rd}(\boldsymbol{b}) - \boldsymbol{b}\|_2/\|\boldsymbol{b}\|_2 = 0.990 \times 10^{-3}$ dargestellt. Wird $\mathrm{rd}(\boldsymbol{b})$ durch $\tilde{\boldsymbol{b}} = (0.666 \times 10^{1}, 0.400 \times 10^{-2})^{\mathsf{T}} \in \mathfrak{R}^2$ ersetzt, so ergibt sich praktisch derselbe kollektive Darstellungsfehler $\|\tilde{\boldsymbol{b}} - \boldsymbol{b}\|/\|\boldsymbol{b}\| = 0.995 \times 10^{-3}$, obwohl $\tilde{b}_2$ in keiner Ziffer mit b_2 übereinstimmt und einen individuellen Darstellungsfehler $|\tilde{b}_2 - b_2|/|b_2| = 2 \times 10^{-1}$ von $20v$ aufweist.

(ii) Für $\boldsymbol{x} = (0.578, 0.713 \times 10^{-3})^{\mathsf{T}}$, $\boldsymbol{y} = (0.202, 0.513 \times 10^{1})^{\mathsf{T}} \in \mathfrak{R}^2$ soll $s = \mathrm{fl}(\boldsymbol{x}^{\mathsf{T}}\boldsymbol{y})$ berechnet werden. Es ergibt sich $s = \mathrm{fl}(\mathrm{fl}(x_1 * y_1) + \mathrm{fl}(x_2 * y_2)) = \mathrm{fl}(0.116 + 0.365 \times 10^{-2}) = 0.118$. Sind $\boldsymbol{x}, \boldsymbol{y}$ Computerdarstellungen der Vektoren $\tilde{\boldsymbol{x}}, \tilde{\boldsymbol{y}} \in \mathbf{R}^2$, so haben die Komponenten eine individuelle relative Unsicherheit von $v = 10^{-2}$. Diese bewirkt — unabhängig von den bei der Berechnung auftretenden Rundungsfehlern — im ersten Summanden von s eine absolute Unsicherheit von 10^{-3}, beim zweiten eine von 10^{-5}. Die Unsicherheit in s würde nicht wesentlich größer werden, wenn x_2 einen um den Faktor 100 größeren Fehler hätte. Nur bei spezieller Korrelation der Daten — wenn etwa y_2 in der Größenordnung 10^3 liegen würde — könnte die individuelle hohe Genauigkeit von x_2 ausgenutzt werden. □

Aus 2.3.16(i) folgt u. a., daß beim Übergang zu $\mathrm{rd}(\boldsymbol{b})$ bzw. $\mathrm{rd}(\boldsymbol{A})$ auftretender Unterlauf die Gültigkeit von (32) bzw. (33) nicht beeinträchtigt, sofern in $\boldsymbol{b}$ bzw. $\boldsymbol{A}$ mindestens ein Element in vernünftiger Größenordnung vorhanden ist.

D. *Einfache Vektor- und Matrixoperationen*

Wir betrachten hier nur einige einfache Beispiele, um die Technik und die Sprache der Rundungsfehleranalyse zu demonstrieren.

Aufgabe 1: Für $\boldsymbol{a}, \boldsymbol{b}, \boldsymbol{x} \in \mathfrak{R}^n$ ist

$$\boldsymbol{z}^* = (\boldsymbol{I} - \boldsymbol{a}\boldsymbol{b}^{\mathsf{T}})\,\boldsymbol{x}$$

zu berechnen.

Algorithmus: $\alpha := \boldsymbol{b}^{\mathsf{T}}\boldsymbol{x}$, $\boldsymbol{z} := \boldsymbol{x} - \boldsymbol{a} * \alpha$

Aufwand: $2n$ opms

Rundungsfehleranalyse: Nach 2.3.10 gilt

$$\alpha = \sum_{j=1}^{n} b_j x_j (1 + \varphi_j^{(n)}), \qquad |\varphi_j^{(n)}| \leqq nv,$$

und für die Komponenten z_i von $\boldsymbol{z}$ ergibt sich nach 2.2.11

$$z_i = \mathrm{fl}\,(x_i - a_i * \alpha) = [x_i - a_i * \alpha(1 + \delta_i)]\,(1 + \xi_i), \qquad |\delta_i|, |\xi_i| \leqq \nu, \quad (35)$$

also

$$z_i = x_i(1 + \xi_i) - a_i(1 + \delta_i)\,(1 + \xi_i)\left\{\sum_{j=1}^{n} b_j \left[\frac{1 + \varphi_j^{(n)}}{1 + \xi_j}\right] x_j(1 + \xi_j)\right\}.$$

Nach 2.3.2 ist $(1 + \delta_i)\,(1 + \xi_i) = 1 + \alpha_i$ mit $|\alpha_i| \leqq 2\nu$ und ebenso $(1 + \varphi_j^{(n)})/(1 + \xi_j) = 1 + \beta_j$ mit $|\beta_j| \leqq (n + 1)\,\nu$, vgl. Ü 2.2.3. Mit den durch $\delta x_i = \xi_i x_i$, $\delta a_i = \alpha_i a_i$ und $\delta b_i = \beta_i b_i$ festgelegten Störungen $\delta\boldsymbol{x}$, $\delta\boldsymbol{a}$ und $\delta\boldsymbol{b}$ folgt dann

$$\begin{aligned}\boldsymbol{z} &= (\boldsymbol{x} + \delta\boldsymbol{x}) - (\boldsymbol{a} + \delta\boldsymbol{a})\,(\boldsymbol{b} + \delta\boldsymbol{b})^{\mathsf{T}}\,(\boldsymbol{x} + \delta\boldsymbol{x}) \\ &= [\boldsymbol{I} - (\boldsymbol{a} + \delta\boldsymbol{a})\,(\boldsymbol{b} + \delta\boldsymbol{b})^{\mathsf{T}}]\,(\boldsymbol{x} + \delta\boldsymbol{x}),\end{aligned}$$

und die Störungen sind im Sinne von

$$|\delta\boldsymbol{a}| \leqq 2\nu\,|\boldsymbol{a}|, \qquad |\delta\boldsymbol{b}| \leqq (n + 1)\,\nu\,|\boldsymbol{b}| \quad \text{und} \quad |\delta\boldsymbol{x}| \leqq \nu\,|\boldsymbol{x}|$$

individuell klein. Der Algorithmus ist also numerisch gutartig mit $F = n + 1$, und zwar elementweise in bezug auf die Eingangsdaten $\boldsymbol{a}$, $\boldsymbol{b}$ und $\boldsymbol{x}$. Daß die Verteilung der Störungen in gewissen Grenzen willkürlich ist, zeigt die folgende gemischte Fehleranalyse, bei der $\boldsymbol{x}$ nicht, dafür aber $\boldsymbol{z}$ mit Störungen belastet wird. Aus (35) folgt auch

$$z_i/(1 + \xi_i) = z_i(1 + \eta_i) = x_i - a_i(1 + \delta_i)\left\{\sum_{j=1}^{n} b_j(1 + \varphi_j^{(n)})\,x_j\right\},$$

wobei η_i mit $|\eta_i| \leqq \nu$ über $1/(1 + \xi_i) = (1 + \eta_i)$ eingeführt worden ist. Mit den durch $\delta z_i = \eta_i z_i$, $\delta a_i = \delta_i a_i$ und $\delta b_i = \varphi_i^{(n)} b_i$ festgelegten Störungen $\delta\boldsymbol{z}$, $\delta\boldsymbol{a}$ und $\delta\boldsymbol{b}$, die sämtlich wegen

$$|\delta\boldsymbol{z}| \leqq \nu\,|\boldsymbol{z}|, \qquad |\delta\boldsymbol{a}| \leqq \nu\,|\boldsymbol{a}|, \qquad |\delta\boldsymbol{b}| \leqq n\nu\,|\boldsymbol{b}|$$

individuell klein sind, ergibt sich

$$\boldsymbol{z} + \delta\boldsymbol{z} = \boldsymbol{x} - (\boldsymbol{a} + \delta\boldsymbol{a})\,(\boldsymbol{b} + \delta\boldsymbol{b})^{\mathsf{T}}\,\boldsymbol{x} = [\boldsymbol{I} - (\boldsymbol{a} + \delta\boldsymbol{a})\,(\boldsymbol{b} + \delta\boldsymbol{b})^{\mathsf{T}}]\,\boldsymbol{x}.$$

Das berechnete Resultat $\boldsymbol{z}$ ist das um $\delta\boldsymbol{z}$ gestörte exakte Resultat zu den um $\delta\boldsymbol{a}$ und $\delta\boldsymbol{b}$ gestörten Eingangsdaten, während $\boldsymbol{x}$ nicht mit Störungen belastet wird. □

Aufgabe 2. Für $\boldsymbol{A} \in \Re^{m,n}$ und $\boldsymbol{x} \in \Re^n$ ist

$$\boldsymbol{y}^* = \boldsymbol{A}\boldsymbol{x}$$

zu berechnen.

Algorithmus: **for** $i := 1(1)m$ **do** $y_i := \sum_{j=1}^{n} a_{ij} * x_j$

Aufwand: mn opms

Rundungsfehleranalyse: Die i-te Komponente y_i ergibt sich als Skalarprodukt der i-ten Zeile von $\boldsymbol{A}$ mit $\boldsymbol{x}$, so daß wegen 2.3.10

$$y_i = \sum_{j=1}^{n} a_{ij}(1 + \varphi_{ij})\, x_j \quad \text{mit} \quad |\varphi_{ij}| \leqq n\nu \tag{36}$$

gilt. Mit der Störungsmatrix $\delta\boldsymbol{A}$, $(\delta\boldsymbol{A})_{ij} = \varphi_{ij} a_{ij}$ folgt daher:

Das berechnete Produkt $\boldsymbol{y}$ läßt sich in der Form

$$\boldsymbol{y} = (\boldsymbol{A} + \delta\boldsymbol{A})\,\boldsymbol{x} \quad \text{mit} \quad |\delta\boldsymbol{A}| \leqq n\nu\,|\boldsymbol{A}| \tag{37}$$

schreiben, d. h., die Matrix-Vektor-Multiplikation ist numerisch gutartig mit $F = n$ elementweise in bezug auf $\boldsymbol{A}$.

In (36) lassen sich auch die Komponenten von $\boldsymbol{x}$ mit den Störungen φ_{ij} belasten: Mit der Störung $\delta\boldsymbol{x}^{(i)}$, $(\delta\boldsymbol{x}^{(i)})_j = \varphi_{ij} x_j$, gilt

$$y_i = [\boldsymbol{A}(\boldsymbol{x} + \delta\boldsymbol{x}^{(i)})]_i \quad \text{mit} \quad |\delta\boldsymbol{x}^{(i)}| \leqq n\nu\,|\boldsymbol{x}|\,. \tag{38}$$

Da die Störungen $\delta\boldsymbol{x}^{(i)}$ von $\boldsymbol{x}$ für jede Komponente i. allg. verschieden sind, ist nur die Berechnung einer einzelnen Komponente y_i, nicht aber die Berechnung des gesamten Vektors $\boldsymbol{y} = (y_1, \ldots, y_m)^{\mathsf{T}}$ numerisch gutartig in bezug auf $\boldsymbol{x}$, d. h., es gibt i. allg. keine kleine Störung $\delta\boldsymbol{x}$ mit $\boldsymbol{y} = \boldsymbol{A}(\boldsymbol{x} + \delta\boldsymbol{x})$. Allerdings folgt aus (38) die numerische Stabilität in bezug auf $\boldsymbol{x}$, d. h. im Vergleich auf das durch den Darstellungsfehler von $\boldsymbol{x}$ hervorgerufene unvermeidliche Fehlerniveau von $\boldsymbol{y}^* = \boldsymbol{A}\boldsymbol{x}$: Wenn neben $\boldsymbol{x}$ alle $\tilde{\boldsymbol{x}} = \boldsymbol{x} + \delta\boldsymbol{x}$ mit $|\delta\boldsymbol{x}| \leqq \nu\,|\boldsymbol{x}|$ als zulässig angesehen werden, ergibt sich

$$|\tilde{\boldsymbol{y}} - \boldsymbol{y}^*| = |\boldsymbol{A}(\tilde{\boldsymbol{x}} - \boldsymbol{x})| \leqq |\boldsymbol{A}|\,|\delta\boldsymbol{x}| \leqq \nu\,|\boldsymbol{A}|\,|\boldsymbol{x}| = \boldsymbol{y}_{\text{opt}}(x_i, \nu) \tag{39}$$

als Vektor der unvermeidlichen Fehlerniveaus in jeder Komponente von $\boldsymbol{y}^*$. Andererseits folgt aus (38) sofort

$$|y_i^* - y_i| = |(\boldsymbol{A}\delta\boldsymbol{x}^{(i)})_i| \leqq (|\boldsymbol{A}|\,|\delta\boldsymbol{x}^{(i)}|)_i \leqq n\nu(|\boldsymbol{A}|\,|\boldsymbol{x}|)_i\,,$$

also

$$|\boldsymbol{y}^* - \boldsymbol{y}| \leqq n\nu\,|\boldsymbol{A}|\,|\boldsymbol{x}| \tag{40}$$

als Schranke für den erzeugten Rundungsfehler. Es liegt daher (komponentenweise) numerische Stabilität mit $F = n$ in bezug auf $\boldsymbol{x}$ vor. □

Aufgabe 3: Für $\boldsymbol{A} \in \mathfrak{R}^{m,n}$, $\boldsymbol{B} \in \mathfrak{R}^{n,l}$ ist

$$\boldsymbol{C}^* = \boldsymbol{A}\boldsymbol{B}$$

zu berechnen.

Algorithmus: **for** $i := 1(1)m$ **do**

$$\textbf{for } j := 1(1)l \textbf{ do } c_{ij} := \sum_{k=1}^{n} a_{ik} * b_{kj}$$

Aufwand: mnl opms

Rundungsfehleranalyse: Da die j-te Spalte $\boldsymbol{c}^{(j)}$ von $\boldsymbol{C}$ als Produkt von $\boldsymbol{A}$ mit der j-ten Spalte $\boldsymbol{b}^{(j)}$ von $\boldsymbol{B}$ berechnet wird, also $\boldsymbol{c}^{(j)} = \boldsymbol{A}\boldsymbol{b}^{(j)}$ gilt, können die Ergebnisse

von Aufgabe 2 direkt angewendet werden und liefern: Zu jedem j gibt es eine Störung $\delta A^{(j)} \in \mathbf{R}^{m,n}$, so daß die j-te Spalte der berechneten Ergebnismatrix als

$$c^{(j)} = (A + \delta A^{(j)})\, b^{(j)} \quad \text{mit} \quad |\delta A^{(j)}| \leqq \nu n\, |A| \tag{41}$$

geschrieben werden kann. Im allgemeinen existieren jedoch keine relativ kleinen Störungen δA und δB mit $C = (A + \delta A)\,(B + \delta B)$, so daß die Matrix-Matrix-Multiplikation nicht numerisch gutartig ist.

Zur Stabilitätsuntersuchung betrachten wir gestörte Faktoren $\tilde{A} = A + \delta A$, $\tilde{B} = B + \delta B$ mit individuell kleinen Störungen $|\delta A| \leqq \nu\, |A|$, $|\delta B| \leqq \nu\, |B|$. Dann folgt

$$\begin{aligned} |C - C^*| &= |(A + \delta A)\,(B + \delta B) - AB| \leqq |\delta A|\,|B| + |A|\,|\delta B| + |\delta A|\,|\delta B| \\ &\leqq (2\nu + \nu^2)\,|A|\,|B|, \end{aligned} \tag{42}$$

d. h., das elementweise Niveau des unvermeidlichen Fehlers ist $2\nu\,|A|\,|B|$. Aus (41) folgt

$$|c^{*(j)} - c^{(j)}| = |\delta A^{(j)} b^{(j)}| \leqq \nu n |A|\,|b^{(j)}|, \quad \text{also} \quad |C^* - C| \leqq \nu n |A|\,|B|. \tag{43}$$

Die Matrix-Matrix-Multiplikation ist daher numerisch stabil mit $F = n/2$ in bezug auf A und B. Wird nur die Darstellungsunsicherheit von B betrachtet und A als fixierter Parameter angesehen, so tritt in (42) der Faktor ν statt $2\nu + \nu^2$ auf. In diesem Modell ist $F = n$. □

E. *Vermeidung von Über- und Unterlauf*

Wir haben bisher vorausgesetzt, daß bei der Realisierung der Algorithmen in $\mathfrak{R}$ weder Über- noch Unterlauf eintritt. Bei den heute meist üblichen Werten der Unterlaufschranke MIN und der Überlaufsschranke MAX — vgl. Abschnitt 2.2 — deutet ein tatsächlich vorkommender Über- bzw. Unterlauf fast immer auf einen Programm- bzw. Anwendungsfehler hin. Die Hauptursache ist meist die fortlaufende Multiplikation großer bzw. kleiner Zahlen etwa bei der Determinantenberechnung oder bei der Berechnung der Exponentialfunktion sowie Division durch die Computernull, die meist gesondert angezeigt wird. Insbesondere sollte Über- bzw. Unterlauf bei der Berechnung von Zwischenergebnissen vermieden werden, wenn die Eingangs- und Ausgangsdaten betragsmäßig in [MIN, MAX] liegen. Durch geschickte Programmierung läßt sich das häufig erreichen; als Beispiel gehen wir auf die Berechnung der Euklidischen Norm eines Vektors ein.

Aufgabe: Für $x \in \mathfrak{R}^n$ ist $\|x\|_2 = \sqrt{\sum_{i=1}^{n} (x_i)^2}$ zu berechnen.

Die wörtliche Umsetzung dieser Formel führt in $\mathfrak{R}(10, 4, 9, 9)$ z. B. für $x = (0.3 \times 10^6, 0.1 \times 10^{-2})^T$ zu Überlauf beim Quadrieren von x_1, obwohl $\|x\|_2 = 0.300\ldots \times 10^6$ in $\mathfrak{R}$ gut darstellbar ist. Durch eine geeignete Skalierung läßt

sich der beschriebene Effekt vermeiden, etwa durch Auswertung von

$$\|\boldsymbol{x}\|_2 = |x_\mu| \sqrt{\sum_{i=1}^{n} (x_i/|x_\mu|)^2}\,, \qquad |x_\mu| = \max_i |x_i|$$

nach dem Algorithmus

xmax := 0
for i := 1(1)n **do if** $|x_i|$ > xmax **then** xmax := $|x_i|$
if xmax = 0
then normx := 0
else | s := 0
| **for** i := 1(1)n **do** s := s + (x_i/xmax) ↑ 2
| normx := xmax * sqrt (s)

Die Quotienten $x_i/|x_\mu|$ sind betragsmäßig durch 1 beschränkt, so daß kein Überlauf eintreten kann, und mindestens einer ist gleich 1, so daß evtl. eintretender Unterlauf die Genauigkeit des Resultats nicht beeinflußt. Unter der Annahme (2.2.10) und mit 2.3.10 läßt sich zeigen, daß der angegebene Algorithmus numerisch gutartig ist mit $F = (n + 6)/2$. Da alle Summanden nichtnegativ sind, ist der erzeugte Rundungsfehler klein: $|\,\|\boldsymbol{x}\|_2 - \text{normx}\,/\|\boldsymbol{x}\|_2 \leqq 0.5(n + 6)\,\nu$.

Durch raffiniertere Vorschriften kann der oben notwendige zweimalige Durchlauf der x_i — einmal zur Maximumsuche, einmal zur Aufsummierung der Quadrate — auf einen reduziert werden, siehe B 2.6 für Literaturhinweise.

Bei klassischer Auswertung gemäß normx := sqrt ($\boldsymbol{x}^T\boldsymbol{x}$) liegt numerische Gutartigkeit mit $F = (n + 2)/2$ vor.

F. *Zusammenfassung*

Nachdem die Grundbegriffe der Fehleranalyse an Hand mehrerer Beispiele eingeführt worden sind, sollen sie jetzt in allgemeiner Form fixiert werden.

Wir betrachten die numerische Problemklasse $\{\mathcal{E}, \boldsymbol{P}\}$, d. h. eine Abbildung $\boldsymbol{P}$: $\mathcal{E} \subset \mathbf{R}^N \to \mathbf{R}^M$, und einen (konzeptionellen) numerischen Algorithmus $\boldsymbol{V}_0$: $\mathcal{E} \subset \mathbf{R}^N \to \mathbf{R}^M$ zur Lösung von Aufgaben $\{\boldsymbol{e}, \boldsymbol{P}\}$ aus der Klasse $\{\mathcal{E}, \boldsymbol{P}\}$. Die Computerrealisierung von $\boldsymbol{V}_0$ in einer Gleitpunktarithmetik mit genügend kleinem ν sei $\boldsymbol{V}$: $\mathfrak{E} \subset \mathfrak{R}^N \to \mathfrak{R}^M$, wobei $\mathfrak{E}$ eine geeignete Einschränkung von $\mathcal{E}$ auf $\mathfrak{R}^N$ sei, d. h., es gelte $\mathfrak{E} \subset \mathcal{E}$. Durch $\boldsymbol{V}$ wird jedem Eingangsdatensatz $\boldsymbol{e} \in \mathfrak{E}$ die numerische Lösung $\boldsymbol{a} = \boldsymbol{V}(\boldsymbol{e}) \in \mathfrak{R}^M$ zugeordnet; die zu $\boldsymbol{e}$ gehörende exakte Lösung ist $\boldsymbol{a}^* = \boldsymbol{P}(\boldsymbol{e})$.

2.3.18. *Begriffe der Fehleranalyse.*

(i) Der numerische Algorithmus $\boldsymbol{V}$: $\mathfrak{E} \subset \mathfrak{R}^N \to \mathfrak{R}^M$ heißt in der Klasse $\{\mathfrak{E}, \boldsymbol{P}\}$ *numerisch gutartig*, wenn zu jedem $\boldsymbol{e} \in \mathfrak{E}$ eine Störung $\delta\boldsymbol{e} \in \mathbf{R}^N$ angegeben werden kann, so daß

$$\boldsymbol{a} = \boldsymbol{P}(\boldsymbol{e} + \delta\boldsymbol{e}) \quad \text{und} \quad \|\delta\boldsymbol{e}\| \leqq F_w \nu \|\boldsymbol{e}\| \tag{44}$$

mit einer von ν und $\boldsymbol{e}$ unabhängigen *Fehlerkumulationskonstanten* F_w gilt.

(ii) Für $\boldsymbol{e} \in \mathfrak{E}$ ist

$$\Delta \boldsymbol{a}_{\mathrm{opt}}(\boldsymbol{e}, \nu) = \max \{\|\boldsymbol{P}(\boldsymbol{e} + \delta\boldsymbol{e}) - \boldsymbol{P}(\boldsymbol{e})\| : \delta\boldsymbol{e} \in \mathbf{R}^N \text{ mit } \boldsymbol{e} + \delta\boldsymbol{e} \in \mathcal{E} \text{ und } \|\delta\boldsymbol{e}\| \leqq \nu \|\boldsymbol{e}\|\} \tag{45}$$

das durch die relative Darstellungsunsicherheit ν der Eingangsdaten bedingte *unvermeidliche Fehlerniveau* der Ausgangsdaten.

(iii) Der numerische Algorithmus $\boldsymbol{V}: \mathfrak{E} \subset \mathfrak{R}^N \to \mathfrak{R}^M$ heißt in der Klasse $\{\mathfrak{E}, \boldsymbol{P}\}$ *numerisch stabil*, wenn der erzeugte Rundungsfehler $\boldsymbol{a}^* - \boldsymbol{a} = \boldsymbol{P}(\boldsymbol{e}) - \boldsymbol{V}(\boldsymbol{e})$ des berechneten Resultates $\boldsymbol{a} = \boldsymbol{V}(\boldsymbol{e})$ für jedes $\boldsymbol{e} \in \mathfrak{E}$ der Ungleichung

$$\|\boldsymbol{a}^* - \boldsymbol{a}\| = \|\boldsymbol{P}(\boldsymbol{e}) - \boldsymbol{V}(\boldsymbol{e})\| \leqq F_s \Delta \boldsymbol{a}_{\mathrm{opt}}(\boldsymbol{e}, \nu) \tag{46}$$

mit einer von ν und $\boldsymbol{e}$ unabhängigen *Fehlerkumulationskonstanten* F_s genügt.

Dabei wird vorausgesetzt, daß F_w bzw. F_s minimal gewählt werden.

2.3.19. *Bemerkung.* (i) Numerische Gutartigkeit bedeutet: *Das berechnete Resultat ist das exakte Resultat eines benachbarten Problems mit gestörten Eingangsdaten, und das relative Störungsniveau ist höchstens das F_w-fache der relativen Darstellungsunsicherheit der Eingangsdaten. Ein numerisch gutartiges Verfahren* mit nicht zu großer Fehlerkumulationskonstanten *nutzt die in den Eingangsdaten enthaltene Information bestmöglich aus. Numerische Gutartigkeit ist die beste Eigenschaft, die ein numerisches Verfahren in bezug auf den Rundungsfehlereinfluß haben kann.*

(ii) Numerische Stabilität bedeutet: *Der erzeugte Rundungsfehler ist höchstens das F_s-fache des durch die Darstellungsunsicherheit der Eingangsdaten bedingten unvermeidlichen Fehlers der Ausgangsdaten. Ein numerisch nicht stabiler Algorithmus verursacht für gewisse* — i. allg. nicht für alle! — *Probleme* $\{\boldsymbol{e}, \boldsymbol{P}\}$ der Klasse $\{\mathfrak{E}, \boldsymbol{P}\}$ *beliebig hohe Genauigkeitsverluste. Numerische Stabilität ist daher eine Mindestforderung an ein vernünftiges numerisches Verfahren.*

(iii) Die Abschätzungen (44) bzw. (46) sind zur Fehlerabschätzung für ein konkretes einzelnes Problem i. allg. nicht geeignet, weil

— die Fehlerkumulationskonstanten minimal in bezug auf die gesamte Klasse $\{\mathfrak{E}, \boldsymbol{P}\}$, nicht aber bezüglich des vorliegenden Problems $\{\boldsymbol{e}, \boldsymbol{P}\}$ festgelegt sind. Für letzteres wird i. allg. eine geringere maximale Fehlerkumulation vorliegen als durch F_w bzw. F_s angegeben.

— selbst die kleinstmöglichen Fehlerkumulationskonstanten stets das ungünstigste Fehlerverhalten berücksichtigen und daher i. allg. zu pessimistisch sind.

— auf Grund der Kompliziertheit der Algorithmen der linearen Algebra und der dadurch hervorgerufenen technischen Schwierigkeiten bei der Fehleranalyse i. allg. nur vereinfachte und vergröberte obere Schranken für die „wahren" minimalen Fehlerkumulationskonstanten berechnet werden können.

Sofern die Fehleranalyse sachgerecht durchgeführt wird und Vergröberungen bei der Analyse verschiedener Verfahren auf vergleichbarem Niveau liegen, erlauben die Fehlerkumulationskonstanten jedoch die qualitative und quantitative Einschätzung und damit den Vergleich verschiedener Verfahren. Gegebenenfalls wird es dabei nötig sein, auch auf empirisch, durch umfangreiche Testrechnungen gewon-

nene Kenntnisse zurückzugreifen, um ein realistisches Bild zu bekommen. Zur Fehlerabschätzung für ein konkretes Problem sollten a-posteriori-Abschätzungen unter Verwendung der berechneten Lösung — z. B. mittels geeigneter Residualkriterien — durchgeführt werden.

(iv) Für lokal lipschitzstetige Aufgaben folgt aus 2.1.8

$$\Delta \boldsymbol{a}_{\mathrm{opt}}(\boldsymbol{e}, \nu) \doteq L(\boldsymbol{e})\, \nu\, \|\boldsymbol{e}\|,$$

und numerische Gutartigkeit zieht numerische Stabilität nach sich, wobei $F_s \doteq F_w$.

(v) Für gewisse Aufgabenklassen ist es sinnvoll und möglich, nur gewisse Teile der Eingangsdaten mit Störungen zu belasten. Ebenso kann numerische Stabilität auch nur in bezug auf gewisse Teile der Eingangsdaten definiert werden. Dabei ergeben sich i. allg. andere Fehlerkumulationskonstanten als bei Berücksichtigung aller Eingangsdaten.

(vi) Die Abschätzungen (44) bis (46) sind kollektive Normabschätzungen. Wenn statt der Normen Beträge verwendet werden, erhält man individuelle elementweise Abschätzungen und dementsprechend elementweise zu verstehende Gutartigkeit und Stabilität. □

Zum Abschluß geben wir ein Beispiel eines numerisch instabilen Verfahrens an.

2.3.20. *Beispiel.* Wir betrachten für $n = 2$ das lineare Gleichungssystem

$$a_{11}x_1 + a_{12}x_2 = b_1,$$
$$a_{21}x_1 + a_{22}x_2 = b_2.$$

In der Klasse der Aufgaben mit $a_{11} \neq 0$ und $a_{11}a_{22} - a_{12}a_{21} \neq 0$ können x_1 und x_2 wie folgt berechnet werden: Die erste Gleichung wird mit $-a_{21}/a_{11}$ multipliziert und zur zweiten addiert, womit sich

$$\left(a_{21} - \frac{a_{21}}{a_{11}}\, a_{11}\right) x_1 + \left(a_{22} - \frac{a_{21}}{a_{11}}\, a_{12}\right) x_2 = b_2 - \frac{a_{21}}{a_{11}}\, b_1$$

ergibt.

Der Koeffizient bei x_1 ist 0, so daß x_2 aus dieser Gleichung berechnet werden kann. Aus der ersten Originalgleichung ergibt sich dann x_1 gemäß

$$x_1 = (b_1 - a_{12}x_2)/a_{11}. \tag{47}$$

Man kann zeigen, daß dieses Verfahren (Gaußscher Algorithmus ohne Pivotisierung, siehe Kapitel 5) für die Berechnung von x_2 gutartig und damit stabil ist. Dagegen ist es numerisch instabil als Verfahren zur Berechnung von x_1.

Zahlenbeispiel: Rechnung in $\mathfrak{R}(10, 5, 9, 9)$ mit symmetrischer Rundung,

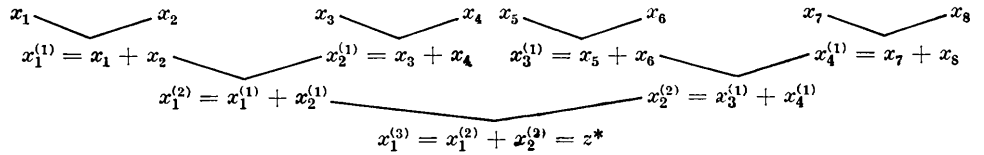
$$0.0021305x_1 + 6.7034x_2 = 19.041,$$
$$9.8702\quad x_1 - 1.8132x_2 = \ \ 4.7207.$$

Nach dem angegebenen Algorithmus erhält man $x_2 = 2.8401$, $x_1 = 1.4081$, die exakten Werte sind $x_2^* = 2.84018101\ldots$, $x_1^* = 1.00003203\ldots$ Der erzeugte Rundungsfehler $|x_1^* - x_1| = 0.408$ überschreitet das unvermeidliche Fehlerniveau $(\Delta x_1)_{\mathrm{opt}} = 4.9 \times 10^{-5}$ von x_1 um den Faktor 8000! Für die vorliegende Aufgabe ist dabei $(\Delta x_1)_{\mathrm{opt}}$ exakt berechnet worden. Werden die beiden Zeilen des Gleichungssystems vertauscht, so liefert derselbe Algorithmus $x_2 = 2.8402$,

$x_1 = 1.0000$, also die gerundeten Werte der exakten Lösung! In der ersten Version wird die Instabilität durch Formel (47) hervorgerufen, bei welcher der Zähler die Gestalt fl (19.041 − 19.038) = 0.003 hat, also zu Auslöschung führt, während der erzeugte Rundungsfehler in der nachfolgenden Division durch die kleine Zahl $a_{11} = 0.0021305$ sehr verstärkt wird, vgl. die Nachbemerkung zu 2.2.13. Selbst für den gerundeten Wert $x_2 = 2.8402$ der exakten Lösung würde sich aus (47) die erste Komponente zu $x_1 = \mathrm{fl}((19.041 - 19.038)/0.0021305) = 0.98498$ ergeben, also immer noch mit einem Fehler $|x_1^* - x_1| = 1.5 \times 10^{-2}$, der das unvermeidliche Fehlerniveau um den Faktor 300 überschreitet. □

Übungsaufgaben

Ü 2.3.1. Zur Berechnung der Summe $z^* = \sum_{i=1}^{n} x_i$ von $n = 2^p$ Zahlen $x_i \in \mathfrak{R}$ kann die sog. paarweise Summation — auch binäre oder Kaskadensummation genannt — verwendet werden, die für $p = 3$, $n = 2^3 = 8$ wie folgt abläuft:



Die allgemeine Vorschrift lautet:

for $i := 1(1)n$ **do** $x_i^{(0)} := x_i$
for $j := 1(1)p$ **do**
  imax $:= 2^{p-j}$
  for $i := 1(1)$imax **do** $x_i^{(j)} := x_{2i-1}^{(j-1)} + x_{2i}^{(j-1)}$
$z := x_1^{(p)}$

(i) Man zeige durch Induktion, daß für den berechneten Wert z

$$z = \sum_{i=1}^{n} x_i(1 + \mu_i^{(n)}) \quad \text{mit} \quad |\mu_i^{(n)}| \leqq pv = (\log_2 n)\, v$$

gilt. Das Verfahren ist also numerisch gutartig mit $F = \log_2 n$ im Vergleich zu $F = n - 1$ für die übliche Summationsvorschrift 2.3.1.

(ii) Man überlege sich, daß die Kaskadensummation unter Verwendung von jeweils einem INTEGER- und einem REAL-Feld der Länge p realisiert werden kann, und gebe ein entsprechendes FORTRAN-Programm an.

(iii) Wie ist das Verfahren zu modifizieren, damit es für beliebiges n angewendet werden kann? Man erweitere das FORTRAN-Programm aus (ii) entsprechend.

(iv) Die Kaskadensummation ist besonders günstig für Parallelrechnung, d. h. für Rechnung mit mehreren parallel arbeitenden Prozessoren.

Ü 2.3.2. Für $\boldsymbol{b} \in \mathbf{R}^m$ bzw. $\boldsymbol{A} \in \mathbf{R}^{m,n}$ ($\boldsymbol{b} \neq \boldsymbol{o}$, $\boldsymbol{A} \neq \boldsymbol{O}$) bezeichne b_μ bzw. $a_{\mu\nu}$ ein betragsgrößtes Element, d. h.

$$|b_\mu| = \max_i |b_i| \quad \text{bzw.} \quad |a_{\mu\nu}| = \max_{i,j} |a_{ij}|.$$

(i) Man beweise die Gültigkeit von

$$\|\boldsymbol{b}\|_p \leqq C_p\, |b_\mu| \quad \text{bzw.} \quad \|\,|\boldsymbol{A}|\,\|_p \leqq D_p\, |a_{\mu\nu}|$$

mit Konstanten

$$C_p = \begin{cases} m & \text{für } p = 1, \\ \sqrt{m} & \text{für } p = 2, \\ 1 & \text{für } p = \infty \end{cases} \quad \text{bzw.} \quad D_p = \begin{cases} m & \text{für } p = 1, \\ n & \text{für } p = \infty, \\ \sqrt{mn} & \text{für } p = 2, F. \end{cases}$$

(ii) Unter Verwendung von (i) zeige man, daß aus den Normabschätzungen

$$\frac{\|\tilde{\boldsymbol{b}} - \boldsymbol{b}\|_p}{\|\boldsymbol{b}\|_p} \leqq \nu \quad \text{bzw.} \quad \frac{\|\tilde{\boldsymbol{A}} - \boldsymbol{A}\|_p}{\||\boldsymbol{A}|\|_p} \leqq \nu \tag{48}$$

die individuellen Abschätzungen

$$\frac{|\tilde{b}_i - b_i|}{|b_\mu|} \leqq C_p \nu \quad \text{bzw.} \quad \frac{|\tilde{a}_{ij} - a_{ij}|}{|a_{\mu\nu}|} \leqq D_p \nu$$

folgen. Dies bedeutet, daß (48) nur für Elemente in der Größenordnung von $|b_\mu|$ bzw. $|a_{\mu\nu}|$ eine hohe individuelle relative Genauigkeit garantiert.

Ü 2.3.3. Man beweise die Behauptung aus Bemerkung 2.3.19(iv).

Ü 2.3.4. Bei der Lösung linearer Gleichungssysteme spielt das Residuum $\boldsymbol{r} = \boldsymbol{b} - \boldsymbol{A}\boldsymbol{x}$ einer Näherungslösung $\boldsymbol{x}$ eine Rolle.

(i) Man zeige, daß das gemäß

for $i := 1(1)n$ **do**

$y := 0$
for $j := 1(1)n$ **do** $y := y + a_{ij} * x_j$
$r_i := b_i - y$

in üblicher Weise berechnete Residuum $\boldsymbol{r} := \mathrm{fl}(\boldsymbol{b} - \boldsymbol{A}\boldsymbol{x})$ der Gleichung

$$r_i = (-y_i + b_i) + [\varepsilon_i/(1 + \varepsilon_i)]\, r_i, \qquad |\varepsilon_i| \leqq \nu, \tag{49}$$

genügt mit $y_i := \mathrm{fl}((\boldsymbol{A}\boldsymbol{x})_i)$. Mit (36), (37), (40) folgere man hieraus die a-priori-Abschätzung

$$|\delta \boldsymbol{r}| \leqq \nu(|\boldsymbol{r}| + n\,|\boldsymbol{A}|\,|\boldsymbol{x}|) \tag{50}$$

für den erzeugten Rundungsfehler $\delta \boldsymbol{r} = \boldsymbol{r} - (\boldsymbol{b} - \boldsymbol{A}\boldsymbol{x})$.

(ii) Man überlege sich, daß durch die erweiterte Vorschrift

for $i := 1(1)n$ **do**

$y := d := 0$
for $j := 1(1)n$ **do**
$[g := a_{ij} * x_j,\ y := y + g,\ d := d + |g| + |y|]$
$r_i := b_i - y,\ d_i := \nu * (d + |r_i|)$

neben $\boldsymbol{r}$ noch eine a-posteriori-Schranke $\boldsymbol{d} = (d_i)$ mit $|\delta \boldsymbol{r}| \leqq \boldsymbol{d}$ berechnet wird. Wie erhöht sich der Aufwand gegenüber (i)?

(iii) Man gebe zu beiden Versionen numerisch äquivalente Varianten an, bei denen die Elemente von $\boldsymbol{A}$ spaltenweise abgearbeitet werden.

Bemerkungen zum Kapitel 2

B 2.1. Wenn x den exakten Wert einer reellen Größe und y eine fehlerbehaftete Näherung für x bezeichnet, wird in der klassischen Fehlerrechnung $\delta := y - x$ als absoluter und $\varepsilon := (y - x)/x$ als relativer Fehler eingeführt. In der Regel — z. B. im Abschnitt 2.2 — gehen

wir ebenso vor. An anderen Stellen — etwa im Abschnitt 2.1 — vertauschen wir die Rolle von x und y, da dann die angestrebten Aussagen leichter zu formulieren sind. Man beachte, daß bei kleinem relativem Fehler beide Definitionen in erster Ordnung denselben Wert liefern, siehe Ü 2.1.8.

B 2.2. Die in den Abschnitten 2.1 und 2.3 eingeführten Begriffe zur Charakterisierung von numerischen Problemen und Algorithmen sind — unter demselben oder ähnlichem Namen und mit derselben oder ähnlicher Bedeutung — von einer Reihe von Autoren betrachtet worden, vgl. etwa WILKINSON [63, 65, 71], BAUER [66, 74], FADDEEV/FADDEEVA [68], BABUŠKA [69, 72], KAHAN [71] und viele andere. Unsere Darstellung folgt dem Konzept von KIEŁBASIŃSKI [78], allerdings in einer für die Probleme der linearen Algebra ausreichenden Vereinfachung.

B 2.3. Die hier benutzte Notation für Computerzahlen geht auf WILKINSON [63] zurück, wo auch die Grundlagen der Rundungsfehleranalyse an vielen Beispielen erläutert worden sind. WILKINSON selbst weist auf Arbeiten von VON NEUMANN/GOLDSTINE [47] und GIVENS [54] hin, in denen erstmals eine Rückwärtsanalyse vorkommt. In einer Vielzahl von Arbeiten ist seitdem der technische Apparat zur Fehleranalyse weiterentwickelt worden. Interessante Ergebnisse zur automatischen Rundungsfehleranalyse sind bei MILLER/WRATHALL [80] zu finden, wo der Computer als analytisches Hilfsmittel eingesetzt wird.

B 2.4. Eine detaillierte Beschreibung der auf realen Computern vorhandenen Gleitpunktarithmetiken gibt STERBENZ [74], vgl. auch BROWN [81]. Zur stochastischen Rundungsfehleranalyse siehe wieder STERBENZ, aber auch VOEVODIN [69b]. Eine Einführung in die Computerarithmetik ist bei VOEVODIN [77] und in den meisten neueren Lehrbüchern der Numerischen Mathematik zu finden.

B 2.5. Die Basis und Mantissenlänge der Zahldarstellung und die Art der Rundung lassen sich bei Bedarf automatisch mit den von MALCOLM [72] und GENTLEMAN/MAROVICH [74] angegebenen FORTRAN-Programmen ermitteln.

B 2.6. Als Beispiel für die computergerechte Programmierung unter Vermeidung von Genauigkeitsverlusten sowie Über- und Unterlauf kann das von BLUE [78] angegebene Programm zur Berechnung der Euklidischen Norm dienen. Dort werden die Komponenten des Vektors nur einmal durchlaufen. Ähnlich arbeiten die von LAWSON et al. [79] und DONGARRA et al. [79] angegebenen Programme.

B 2.7. Eine Alternative zur rundungsfehlerbehafteten Gleitpunktarithmetik stellen sog. exakte Arithmetiken dar, siehe GREGORY-KRISHNAMURTY [84]. Die Realisierung solcher Arithmetiken ist jedoch aufwendig, und sie sind i. allg. nur für spezielle Aufgaben mit fehlerfreien Eingangsdaten zu empfehlen.

B 2.8. Die bei Computerrechnung entstehenden Rundungsfehler lassen sich auch mit den Methoden der Intervallmathematik erfassen, siehe etwa MOORE [79] und NICKEL [80]. Solche Methoden erlauben für ein konkretes Problem und die mittels eines konkreten Algorithmus berechneten Resultate die Angabe von strengen Fehlerschranken, allerdings mit vergleichsweise hohem Aufwand. Sie erlauben jedoch nicht den Vergleich verschiedener Algorithmen für bestimmte Aufgabenklassen. Letzteres ist dagegen mit den Methoden der Rundungsfehleranalyse möglich, die wiederum für die Berechnung von Fehlerschranken für einzelne konkrete Aufgaben nicht besonders geeignet sind. Da wir hauptsächlich am qualitativen Vergleich von Algorithmen (über ihre Fehlerkumulationskonstanten) interessiert sind, gehen wir auf die Intervallmathematik nicht ein.

B 2.9. Durch Kombination von Intervallmethoden und Rechnung in variabler, mehrfach hoher Genauigkeit — beides läßt sich sowohl software- als auch neuerdings hardwaremäßig realisieren — können Ergebnisse mit beliebig vorgebbaren Fehlerschranken garantiert werden, vgl. das in jüngster Zeit von KULISCH et al. [85] u. a. entwickelte und von IBM vertriebene System ACRITH.

B 2.10. Für eine vorgegebene Problemklasse ist die Frage interessant, wieviel arithmetische Operationen mindestens zur Lösung eines beliebigen Problems der Klasse erforderlich sind und wie optimale Algorithmen beschaffen sind. Die Beantwortung dieser Fragen ist Gegenstand der sog. Kompliziertheitstheorie (engl. „complexity", russ. „теория сложности"). Da jedes Element der Eingangsdaten bei mindestens einer Operation beteiligt sein muß, ist eine untere Schranke für die Anzahl der Rechenoperationen von der Form $\sim K_1 N$, N Anzahl der skalaren Eingangsdaten. In diesem Sinne sind die im Abschnitt 2.3 angegebenen naiven Algorithmen für Summation und Skalarprodukt als optimal anzusehen. Eine nichttriviale Situation liegt bei der Matrixmultiplikation — Aufgabe 3 aus Abschnitt 2.3.D — vor. Für den Sonderfall $\boldsymbol{C} = \boldsymbol{AB}, \boldsymbol{A}, \boldsymbol{B} \in \mathbf{R}^{n,n}$, werden dort n^3 opms benötigt, während die untere Schranke von der Form $\sim K_1 n^2$ ist. Den ersten wesentlichen Beitrag zu diesem Problem hat STRASSEN [69] mit einem Verfahren geleistet, das $\sim K_2 n^{2.81}$ Operationen erfordert; inzwischen ist man etwa bei $\sim K_3 n^{2.5}$ angelangt. Man beachte jedoch, daß die Vorfaktoren K_i groß sind und daher die kleineren Exponenten erst für hohe Dimensionen n wirksam werden, wo andererseits die Voraussetzung der vollen Besetztheit von $\boldsymbol{A}$ und $\boldsymbol{B}$ meist nicht mehr realistisch ist. Unter gewissen Voraussetzungen hat MILLER [75] ferner gezeigt, daß ein numerisch stabiles Verfahren zur Matrixmultiplikation $\sim K_4 n^3$ Operationen erfordert. Aus den angegebenen Gründen haben die erwähnten „schnellen" Multiplikationsverfahren bislang keine praktische Bedeutung erlangt, und die naive Realisierung aus Abschnitt 2.3 kann bei Forderung der numerischen Stabilität als fast optimal angesehen werden.

3. Elementare Transformationsmatrizen

Die Aufgaben der linearen Algebra sind besonders einfach zu lösen, wenn die jeweiligen Koeffizientenmatrizen eine günstige Gestalt haben. Zum Beispiel ist ein Gleichungssystem mit einer Dreiecksmatrix viel leichter lösbar als ein System mit einer allgemeinen Matrix. Ganze Klassen von Algorithmen der linearen Algebra beruhen auf dieser Tatsache: Sie transformieren das zu lösende Problem durch eine endliche bzw. auch unendliche Folge von einfachen Transformationen in ein äquivalentes Problem mit einer Matrix der gewünschten günstigen Gestalt, z. B. in ein Problem mit einer Dreiecksmatrix. Dabei heißen zwei Probleme *äquivalent*, wenn sie dieselben Ausgangsdaten, also dieselben Lösungen besitzen. Die Transformation geschieht in der Regel durch Links- bzw. Rechtsmultiplikation mit geeigneten *elementaren Transformationsmatrizen*, wobei in der transformierten Matrix an den gewünschten Stellen nach und nach Nullen erzeugt werden.

Verfahren dieser Struktur werden häufig *direkte Verfahren* genannt; das vorliegende Buch ist vorwiegend solchen Verfahren gewidmet. Im Gegensatz dazu wird bei den sog. *iterativen Verfahren* die Datenmatrix nicht verändert, d. h., in jedem Iterationsschritt wird auf die originale Matrix der Eingangsdaten zurückgegriffen. Die Grenzen zwischen beiden Verfahrensklassen sind jedoch fließend.

Die erwähnten elementaren Transformationsmatrizen können als Werkzeuge der numerischen linearen Algebra angesehen werden, denn mit ihrer Hilfe lassen sich die unterschiedlichsten Algorithmen zur Lösung verschiedener Aufgabenklassen aufbauen.

Zur Illustration betrachten wir die Lösung des linearen Gleichungssystems

$$\begin{aligned} 3x_1 - 6x_2 + 2x_3 &= -3, \\ -2x_1 + 4x_2 + x_3 &= 16, \\ x_1 - x_2 + x_3 &= 4. \end{aligned} \tag{1}$$

Nach dem aus der Schule bekannten Gaußschen Eliminationsverfahren kann x_1 aus der zweiten und dritten Gleichung eliminiert werden, indem die erste mit 2/3 bzw. $-1/3$ multipliziert und zur zweiten bzw. dritten addiert wird:

$$\begin{aligned} 3x_1 - 6x_2 + 2x_3 &= -3, \\ 0 \cdot x_2 + \frac{7}{3} x_3 &= 14, \\ x_2 + \frac{1}{3} x_3 &= 5. \end{aligned} \tag{2}$$

Da in der neuen zweiten Gleichung auch der Koeffizient bei x_2 verschwindet, geht das System (2) durch Vertauschung der letzten beiden Zeilen in das System

$$\begin{aligned} 3x_1 - 6x_2 + 2x_3 &= -3, \\ x_2 + \frac{1}{3} x_3 &= 5, \\ \frac{7}{3} x_3 &= 14 \end{aligned} \tag{3}$$

mit einer Dreiecksmatrix als Koeffizientenmatrix über. Aus (3) lassen sich die Unbekannten in der Reihenfolge $x_3 = 6$, $x_2 = 3$ und $x_1 = 1$ leicht berechnen.

Der Übergang von (1) nach (2) kann in Matrixschreibweise als Linksmultiplikation der Koeffizientenmatrix und des Vektors der rechten Seiten mit der Transformationsmatrix

$$\boldsymbol{L}_1 = \begin{pmatrix} 1 & 0 & 0 \\ \frac{2}{3} & 1 & 0 \\ -\frac{1}{3} & 0 & 1 \end{pmatrix},$$

beschrieben werden, der von (2) nach (3) analog mittels

$$\boldsymbol{T}_{2,3} = \begin{pmatrix} 1 & 0 & 0 \\ 0 & 0 & 1 \\ 0 & 1 & 0 \end{pmatrix}.$$

Der Untersuchung dieser und weiterer Typen elementarer Transformationsmatrizen sind die folgenden Abschnitte gewidmet.

3.1. Permutationsmatrizen

Eine *Permutation* ist eine Abbildung (Zuordnung) $k\colon \{1,2,\ldots,n\} \to \{k(1),k(2),\ldots,k(n)\}$ mit der Eigenschaft, daß die Funktionswerte $k(i)$, $i = 1, \ldots, n$, jeden der Werte $1, \ldots, n$ genau einmal annehmen. Das n-Tupel $\{k(1), \ldots, k(n)\}$ heißt die durch k erzeugte *Permutation der Zahlen* $\{1, \ldots, n\}$; z. B. legt

$$\{k(1), k(2), k(3), k(4)\} = \{2, 4, 1, 3\}$$

für $n = 4$ eine Permutation k fest. Offensichtlich ist die Permutation k durch eine solche Tabelle $\{k(1), \ldots, k(n)\}$ eindeutig beschrieben.

Zu jeder Permutation k gibt es eine eindeutig festgelegte *inverse Permutation* l mit $l(k(i)) = i$ $(i = 1, \ldots, n)$, welche die durch k erzeugte Umordnung wieder rückgängig macht. Im Beispiel ist

$$\{l(1), l(2), l(3), l(4)\} = \{3, 1, 4, 2\}$$

die zu k inverse Permutation.

Durch eine Permutation k wird jedem Vektor $\boldsymbol{x} \in \mathbf{R}^n$ nach der Vorschrift

$$\boldsymbol{x} = (x_i) \to \boldsymbol{y} = (y_i) \quad \text{mit} \quad y_i := x_{k(i)} \qquad (i = 1, \ldots, n) \tag{4}$$

ein Vektor $\boldsymbol{y} \in \mathbf{R}^n$ zugeordnet, der die (durch k erzeugte) *Permutation des Vektors* $\boldsymbol{x}$ genannt wird. Der Übergang von $\boldsymbol{x}$ zu $\boldsymbol{y}$ gemäß (4) kann in eindeutiger Weise durch eine Matrix $\boldsymbol{P} \in \mathbf{R}^{n,n}$ in der Gestalt

$$\boldsymbol{y} = \boldsymbol{P}\boldsymbol{x} \tag{5}$$

geschrieben werden; im Beispiel ist

$$\boldsymbol{P} = \begin{pmatrix} 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 1 \\ 1 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 \end{pmatrix}.$$

3.1.1. *Definition.* Die Matrix $\boldsymbol{P} = (p_{ij}) \in \mathbf{R}^{n,n}$ mit

$$p_{ij} = \begin{cases} 1 & \text{für} \quad j = k(i), \\ 0 & \text{für} \quad j \neq k(i) \end{cases} \qquad (i, j = 1, \ldots, n)$$

heißt die durch k erzeugte Permutationsmatrix.

Offensichtlich ist eine Matrix $\boldsymbol{P}$ genau dann eine Permutationsmatrix, wenn in jeder Zeile und Spalte außer Nullen genau eine 1 vorkommt. Dabei gibt $k(i)$ den Spaltenindex der 1 in der i-ten Zeile an; entsprechend ist $l(j)$ der Zeilenindex der 1 in der j-ten Spalte.

Die Transponierte einer Permutationsmatrix und das Produkt von Permutationsmatrizen sind wieder Permutationsmatrizen. Wegen $\boldsymbol{P}^{\mathsf{T}}\boldsymbol{P} = \boldsymbol{P}\boldsymbol{P}^{\mathsf{T}} = \boldsymbol{I}$ gilt ferner

$$\boldsymbol{P}^{\mathsf{T}} = \boldsymbol{P}^{-1}, \tag{6}$$

d. h., Permutationsmatrizen sind orthogonal, und $\boldsymbol{P}^{\mathsf{T}}$ charakterisiert die inverse Permutation. Bei Anwendung von $\boldsymbol{P}$ auf eine Matrix $\boldsymbol{A} \in \mathbf{R}^{n,m}$ mit den Zeilen $\boldsymbol{a}^{i\mathsf{T}}$ bzw. auf eine Matrix $\boldsymbol{B} \in \mathbf{R}^{m,n}$ mit den Spalten $\boldsymbol{b}^j$ ergibt sich

$$\boldsymbol{P}\begin{pmatrix} - \, \boldsymbol{a}^{1\mathsf{T}} \, - \\ \vdots \\ - \, \boldsymbol{a}^{n\mathsf{T}} \, - \end{pmatrix} = \begin{pmatrix} - \, \boldsymbol{a}^{k(1)\mathsf{T}} \, - \\ \vdots \\ - \, \boldsymbol{a}^{k(n)\mathsf{T}} \, - \end{pmatrix} \quad \text{bzw.} \quad \begin{pmatrix} | & & | \\ \boldsymbol{b}^1 & \ldots & \boldsymbol{b}^n \\ | & & | \end{pmatrix}\boldsymbol{P} = \begin{pmatrix} | & & | \\ \boldsymbol{b}^{l(1)} & \ldots & \boldsymbol{b}^{l(n)} \\ | & & | \end{pmatrix}, \tag{7}$$

d. h., *Linksmultiplikation mit* $\boldsymbol{P}$ *bewirkt eine Permutation der Zeilen gemäß* k, *und Rechtsmultiplikation eine Permutation der Spalten gemäß der inversen Permutation* l.

Sollen die Spalten gemäß k permutiert werden, so ist von rechts mit $\boldsymbol{P}^{\mathsf{T}}$ zu multiplizieren.

Die einfachsten Permutationen bestehen in der Vertauschung von zwei Elementen, bei Vektoren also in der Vertauschung von zwei Komponenten $i \neq j$. Die entsprechenden Matrizen haben die Gestalt

$$\boldsymbol{T}_{ij} = \begin{array}{c} \begin{matrix} & i & & j & \end{matrix} \\ \begin{pmatrix} 1 & \vdots & & \vdots & \\ \dots & 0 & \dots & 1 & \dots \\ & \vdots & & \vdots & \\ \dots & 1 & \dots & 0 & \dots \\ & \vdots & & \vdots & 1 \end{pmatrix} \end{array} \begin{matrix} \\ \\ i\text{-te Zeile} \\ \\ j\text{-te Zeile} \\ \\ \end{matrix}$$

3.1.2. *Definition.* Eine Matrix $\boldsymbol{T}_{ij} \in \mathbf{R}^{n,n}$ $(i, j = 1, \dots, n;\ i \neq j)$, die sich von der Einheitsmatrix nur in den vier Elementen

$$(\boldsymbol{T}_{ij})_{ii} = (\boldsymbol{T}_{ij})_{jj} = 0, \qquad (\boldsymbol{T}_{ij})_{ij} = (\boldsymbol{T}_{ij})_{ji} = 1$$

unterscheidet, heißt *Vertauschungsmatrix*. Für $i = j$ wird $\boldsymbol{T}_{ii} = \boldsymbol{T}_{jj} = \boldsymbol{I}$ gesetzt (triviale Vertauschung).

Vertauschungsmatrizen sind symmetrisch, d. h. $\boldsymbol{T}_{ij} = \boldsymbol{T}_{ij}^{\mathsf{T}}$ und folglich $\boldsymbol{T}_{ij}^2 = \boldsymbol{I}$, und es gilt

$$\det(\boldsymbol{T}_{ij}) = \begin{cases} 1 & \text{für } i = j, \\ -1 & \text{für } i \neq j. \end{cases} \tag{8}$$

Wie im folgenden Beispiel gezeigt wird, läßt sich eine Permutation stets als Folge von Vertauschungen darstellen, wobei im i-ten Schritt das Element i mit einem geeigneten Element $s(i) \geq i$ vertauscht wird:

$$\begin{pmatrix}1\\2\\3\\4\end{pmatrix} \xrightarrow{\boldsymbol{P}} \begin{pmatrix}2\\4\\1\\3\end{pmatrix} \text{ äquivalent zu } \begin{pmatrix}1\\2\\3\\4\end{pmatrix} \xrightarrow{\boldsymbol{T}_{12}} \begin{pmatrix}2\\1\\3\\4\end{pmatrix} \xrightarrow{\boldsymbol{T}_{24}} \begin{pmatrix}2\\4\\3\\1\end{pmatrix} \xrightarrow{\boldsymbol{T}_{34}} \begin{pmatrix}2\\4\\1\\3\end{pmatrix},$$

also $s(1) = 2$, $s(2) = 4$, $s(3) = 4$ und $\boldsymbol{P} = \boldsymbol{T}_{34}\boldsymbol{T}_{24}\boldsymbol{T}_{12}$.

3.1.3. *Aussage.* Jede Permutationsmatrix läßt sich in eindeutiger Weise als Produkt von $n - 1$ Vertauschungsmatrizen gemäß

$$\boldsymbol{P} = \boldsymbol{T}_{n-1,s(n-1)}\boldsymbol{T}_{n-2,s(n-2)} \cdots \boldsymbol{T}_{2,s(2)}\boldsymbol{T}_{1,s(1)} \tag{9}$$

mit $i \leqq s(i) \leqq n$ $(i = 1, \dots, n-1)$ darstellen.

Die zu (9) inverse Permutation ist dann

$$\boldsymbol{P}^{-1} = \boldsymbol{P}^{\mathsf{T}} = \boldsymbol{T}_{1,s(1)}\boldsymbol{T}_{2,s(2)} \cdots \boldsymbol{T}_{n-2,s(n-2)}\boldsymbol{T}_{n-1,s(n-1)}. \tag{10}$$

In der numerischen linearen Algebra kommen Permutationen fast ausschließlich als Folge von Vertauschungen gemäß (9) vor. In vielen Fällen ist es sinnvoll, die Informationen über diese Permutationen in Form der Tabelle $\{s(1), s(2), \dots, s(n-1)\}$

in einem Integerfeld zu speichern. Zu gegebenem $\boldsymbol{x}$ kann dann $\boldsymbol{y} = \boldsymbol{P}\boldsymbol{x}$ sogar auf dem Platz von $\boldsymbol{x}$ — also in situ — berechnet werden; dasselbe gilt für die inverse Permutation.

In anderen Fällen ist eine Darstellung der Permutation als Tabelle $\{k(1), \ldots, k(n)\}$ günstiger. Eine solche Tabelle entsteht mittels der $s(i)$, wenn der nachfolgend beschriebene Algorithmus PER auf den Vektor $\boldsymbol{x} = (x_i) := (i)$ angewendet wird.

3.1.4. *Algorithmus.* Berechnung von $\boldsymbol{x} := \boldsymbol{P}\boldsymbol{x}$ (PER) bzw. $\boldsymbol{x} := \boldsymbol{P}^{-1}\boldsymbol{x}$ (PERINV)

PER: **for** $i := 1(1)n - 1$ **do**
  if $s(i) > i$ **then**
    $z := x_i$
    $x_i := x_{s(i)}$
    $x_{s(i)} := z$

PERINV: **for** $i := n - 1(-1)1$ **do**
  if $s(i) > i$ **then**
    $z := x_i$
    $x_i := x_{s(i)}$
    $x_{s(i)} := z$

Durch den Algorithmus PER wird $\boldsymbol{x}$ mit $\boldsymbol{P}\boldsymbol{x}$, durch PERINV wird $\boldsymbol{x}$ mit $\boldsymbol{P}^{-1}\boldsymbol{x}$ überspeichert. Beide Algorithmen sind Beispiele für sog. nichtnumerische Algorithmen, da in ihnen nur logische und Speicheroperationen, aber keine arithmetischen Operationen ausgeführt werden.

Wir weisen nochmals darauf hin, daß Permutationen bei der Programmierung *nur* über die Tabellen $\{k(i)\}$ bzw. $\{s(i)\}$ und *niemals* über die zugehörigen Matrizen realisiert werden; letztere dienen ausschließlich der mathematischen Beschreibung.

Übungsaufgaben

Ü 3.1.1. Man überlege sich Algorithmen, die aus der Tabelle $\{k(i)\}$ die Tabellen $\{l(i)\}$ der inversen Permutation sowie die Tabelle $\{s(i)\}$ für die Darstellung (9) berechnen.

Ü 3.1.2. Es ist zu zeigen, daß jede Permutation $\boldsymbol{P}$ auch eindeutig in der Form

$$\boldsymbol{P} = \boldsymbol{T}_{n,r(n)}\boldsymbol{T}_{n-1,r(n-1)} \cdots \boldsymbol{T}_{3,r(3)}\boldsymbol{T}_{2,r(2)} \tag{11}$$

mit $1 \leqq r(i) \leqq i$ $(i = 2, \ldots, n)$ geschrieben werden kann.

3.2. Nichtorthogonale elementare Transformationsmatrizen

Im folgenden betrachten wir Matrizen der Gestalt

$$\boldsymbol{M}_k = \begin{pmatrix} 1 & & & & \\ & \ddots & & & \\ & & 1 & & \\ & & m_{k+1} & \ddots & \\ & & \vdots & & \ddots \\ & & m_n & & 1 \end{pmatrix} \quad \text{bzw.} \quad \boldsymbol{L}_k = \begin{pmatrix} 1 & & & & \\ & \ddots & & & \\ & & 1 & & \\ & & l_{k+1} & \ddots & \\ & & \vdots & & \ddots \\ & & l_n & & 1 \end{pmatrix}, \tag{1}$$

wie sie uns in der Einführung zum dritten Kapitel bereits begegnet sind.

3.2.1. *Definition.*

(i) Eine Matrix $\boldsymbol{M}_k \in \mathbf{R}^{n,n}$ $(k = 1, \ldots, n)$ des Typs

$$\boldsymbol{M}_k = \boldsymbol{M}_k(\boldsymbol{m}) = \boldsymbol{I} + \boldsymbol{m}\boldsymbol{e}^{k\mathsf{T}}, \qquad \boldsymbol{m} = (m_i) \in \mathbf{R}^n,$$

mit

$$m_k = \boldsymbol{e}^{k\mathsf{T}}\boldsymbol{m} = 0$$

heißt *nichtorthogonale elementare Transformationsmatrix*, abgekürzt: *NT-Matrix*.

(ii) Eine NT-Matrix $\boldsymbol{L}_k \in \mathbf{R}^{n,n}$ $(k = 1, \ldots, n-1)$ des Typs

$$\boldsymbol{L}_k = \boldsymbol{L}_k(\boldsymbol{l}) = \boldsymbol{I} + \boldsymbol{l}\boldsymbol{e}^{k\mathsf{T}}, \qquad \boldsymbol{l} = (l_i) \in \mathbf{R}^n,$$

die gleichzeitig eine untere Dreiecksmatrix ist, d. h., für die

$$l_i = \boldsymbol{e}^{i\mathsf{T}}\boldsymbol{l} = 0 \qquad (i = 1, \ldots, k)$$

gilt, heißt *LNT-Matrix*.

Wegen

$$\begin{aligned}\boldsymbol{M}_k(\boldsymbol{m})\,\boldsymbol{M}_k(-\boldsymbol{m}) &= [\boldsymbol{I} + \boldsymbol{m}\boldsymbol{e}^{k\mathsf{T}}]\,[\boldsymbol{I} - \boldsymbol{m}\boldsymbol{e}^{k\mathsf{T}}] \\ &= \boldsymbol{I} + \boldsymbol{m}\boldsymbol{e}^{k\mathsf{T}} - \boldsymbol{m}\boldsymbol{e}^{k\mathsf{T}} - \boldsymbol{m}\boldsymbol{e}^{k\mathsf{T}}\boldsymbol{m}\boldsymbol{e}^{k\mathsf{T}} = \boldsymbol{I}\end{aligned}$$

ist jede NT-Matrix regulär, und die Inverse

$$\boldsymbol{M}_k(\boldsymbol{m})^{-1} = \boldsymbol{M}_k(-\boldsymbol{m}) = \boldsymbol{I} - \boldsymbol{m}\boldsymbol{e}^{k\mathsf{T}} \tag{2}$$

ist wieder eine NT-Matrix. Speziell gilt

$$\boldsymbol{L}_k(\boldsymbol{l})^{-1} = \boldsymbol{L}_k(-\boldsymbol{l}) = \boldsymbol{I} - \boldsymbol{l}\boldsymbol{e}^{k\mathsf{T}}, \tag{3}$$

d. h., auch die Inverse einer LNT-Matrix ist wieder eine LNT-Matrix. Die Formeln (2) und (3) besagen: *Die Inverse einer NT- bzw. LNT-Matrix entsteht durch Vorzeichenumkehr der nichttrivialen Elemente* m_i *bzw.* l_i. Da die LNT-Matrizen in der numerischen linearen Algebra eine besondere Rolle spielen, gehen wir im folgenden ausführlich auf diese Unterklasse der NT-Matrizen ein.

3.2.2. *Aussage.* Das Produkt

$$\boldsymbol{L} = \boldsymbol{L}_1\boldsymbol{L}_2 \cdots \boldsymbol{L}_{n-2}\boldsymbol{L}_{n-1}$$

der LNT-Matrizen

$$\boldsymbol{L}_k = \boldsymbol{L}_k(\boldsymbol{l}^k), \qquad \boldsymbol{l}^k = (0, \ldots, 0, l_{k+1,k}, \ldots, l_{nk})^{\mathsf{T}}$$

ist eine untere Dreiecksmatrix mit Einsdiagonale, und für die Nichtdiagonalelemente gilt

$$(\boldsymbol{L})_{ij} = l_{ij} \qquad (i > j,\ i, j = 1, \ldots, n).$$

Für $n = 4$ ist also

$$\begin{pmatrix} 1 & & & \\ l_{21} & 1 & & \\ l_{31} & & 1 & \\ l_{41} & & & 1 \end{pmatrix} \begin{pmatrix} 1 & & & \\ & 1 & & \\ & l_{32} & 1 & \\ & l_{42} & & 1 \end{pmatrix} \begin{pmatrix} 1 & & & \\ & 1 & & \\ & & 1 & \\ & & l_{43} & 1 \end{pmatrix} = \begin{pmatrix} 1 & & & \\ l_{21} & 1 & & \\ l_{31} & l_{32} & 1 & \\ l_{41} & l_{42} & l_{43} & 1 \end{pmatrix}.$$

Nachrechnen dieser Beziehung von rechts nach links zeigt, wie der Beweis zu 3.2.2. verläuft. Man beachte jedoch, daß das Produkt $\boldsymbol{L}_{n-1}\boldsymbol{L}_{n-2} \cdots \boldsymbol{L}_1\boldsymbol{L}_2$ zwar auch eine untere Dreiecksmatrix, aber i. allg. von $\boldsymbol{L}$ verschieden ist.

Wenn eine Matrix $\boldsymbol{B} \in \mathbf{R}^{n,r}$ mit den Zeilen $\boldsymbol{b}^{i\mathsf{T}}$ von links mit der LNT-Matrix $\boldsymbol{L}_k$ multipliziert wird, ergibt sich die transformierte Matrix $\bar{\boldsymbol{B}} = \boldsymbol{L}_k\boldsymbol{B}$ wie folgt:

$$\begin{pmatrix} 1 & & & & \\ & \ddots & & & \\ & & 1 & & \\ & & l_{k+1} & 1 & \\ & & \vdots & & \ddots \\ & & l_n & & 1 \end{pmatrix} \begin{pmatrix} \dots & \boldsymbol{b}^{1\mathsf{T}} & \dots \\ & \vdots & \\ \dots & \boldsymbol{b}^{k\mathsf{T}} & \dots \\ \dots & \boldsymbol{b}^{k+1\mathsf{T}} & \dots \\ & \vdots & \\ \dots & \boldsymbol{b}^{n\mathsf{T}} & \dots \end{pmatrix} = \begin{pmatrix} \dots\dots & \boldsymbol{b}^{1\mathsf{T}} & \dots\dots \\ & \vdots & \\ \dots\dots & \boldsymbol{b}^{k\mathsf{T}} & \dots\dots \\ \dots & \boldsymbol{b}^{k+1\mathsf{T}} + l_{k+1}\boldsymbol{b}^{k\mathsf{T}} & \dots \\ & \vdots & \\ \dots & \boldsymbol{b}^{n\mathsf{T}} + l_n\boldsymbol{b}^{k\mathsf{T}} & \dots \end{pmatrix}. \tag{4}$$

Man sieht: *Die Transformation $\bar{\boldsymbol{B}} = \boldsymbol{L}_k\boldsymbol{B}$ ändert die ersten k Zeilen von $\boldsymbol{B}$ nicht. Die Zeilen $i = k + 1, \dots, n$ von $\bar{\boldsymbol{B}}$ entstehen aus denen von $\boldsymbol{B}$ nach der Vorschrift*

$$\langle \text{neue } i\text{-te Zeile} \rangle := \langle \text{alte } i\text{-te Zeile} \rangle + l_i * \langle \text{alte } k\text{-te Zeile} \rangle$$

In elementweiser Notation erhalten wir

3.2.3. *Algorithmus zur LNT-Transformation.* Berechnung von $\bar{\boldsymbol{B}} = (\bar{b}_{ij}) := \boldsymbol{L}_k\boldsymbol{B}$ für $\boldsymbol{B} = (b_{ij}) \in \mathfrak{R}^{n,r}$ und die LNT-Matrix $\boldsymbol{L}_k = \boldsymbol{L}_k(\boldsymbol{l})$, $\boldsymbol{l} = (0, \dots, 0, l_{k+1}, \dots, l_n)^{\mathsf{T}} \in \mathfrak{R}^n$:

$$\bar{b}_{ij} := \begin{cases} b_{ij} & (i = 1, \dots, k), \\ b_{ij} + l_i * b_{kj} & (i = k + 1, \dots, n) \end{cases} \qquad (j = 1, \dots, r).$$

Aufwand: $(n - k)\,r$ opms

Algorithmus 3.2.3 kann auf dem Platz von $\boldsymbol{B}$ gemäß $\boldsymbol{B} := \boldsymbol{L}_k\boldsymbol{B}$ ausgeführt werden. Ob die neuen Elemente zeilen- oder spaltenweise berechnet werden sollten, hängt von den Umständen — z. B. der Art der Speicherung — ab.

Aus (4) ist ersichtlich, daß die Transformation $\bar{\boldsymbol{B}} = \boldsymbol{L}_k\boldsymbol{B}$ auch in der Form

$$\boldsymbol{L}_k^{(n)}\boldsymbol{B}^{(n)} = \left(\begin{array}{c|c} \boldsymbol{I}^{(k-1)} & \boldsymbol{O} \\ \hline \boldsymbol{O} & \boldsymbol{L}_1^{(n-k+1)} \end{array}\right) \left(\begin{array}{c} \boldsymbol{B}_1^{(k-1)} \\ \hline \boldsymbol{B}_2^{(n-k+1)} \end{array}\right) = \left(\begin{array}{c} \boldsymbol{B}_1^{(k-1)} \\ \hline \boldsymbol{L}_1^{(n-k+1)}\boldsymbol{B}_2^{(n-k+1)} \end{array}\right)$$

$$= \left(\begin{array}{c} \bar{\boldsymbol{B}}_1^{(k-1)} \\ \hline \bar{\boldsymbol{B}}_2^{(n-k+1)} \end{array}\right) = \bar{\boldsymbol{B}}^{(n)} \tag{6}$$

geschrieben werden kann, wobei $\boldsymbol{B}_1^{(k-1)}$ aus den ersten $k-1$, $\boldsymbol{B}_2^{(n-k+1)}$ aus den letzten $n-k+1$ Zeilen von $\boldsymbol{B} = \boldsymbol{B}^{(n)}$ besteht. Mit $\boldsymbol{I}^{(k-1)}$ ist die $(k-1)$-dimensionale Einheitsmatrix und mit $\boldsymbol{L}_1^{(n-k+1)}$ die LNT-Matrix $\boldsymbol{L}_1 = \boldsymbol{L}_1(\boldsymbol{l}^{(n-k+1)})$ der Dimension $n-k+1$ mit dem erzeugenden Vektor $\boldsymbol{l}^{(n-k+1)} = (0, l_{k+1}, \ldots, l_n)^T$ bezeichnet worden. *Es genügt daher, LNT-Matrizen für den Index $k = 1$ zu betrachten, da der Fall $k \neq 1$ gemäß* (6) *auf diese Standardsituation zurückgeführt werden kann.*

Die LNT-Matrix $\boldsymbol{L}_1$ wird benutzt, um in einem vorgegebenen Vektor $\boldsymbol{a} = (a_i) \in \mathbf{R}^n$ unterhalb der ersten Komponente Nullen zu erzeugen. Nach 3.2.3 ergibt sich für $\bar{\boldsymbol{a}} = \boldsymbol{L}_1\boldsymbol{a}$

$$\bar{a}_1 = a_1, \qquad \bar{a}_i = a_i + l_i * a_1 \qquad (i = 2, \ldots, n).$$

Die Bedingungen $\bar{a}_i = 0$ lassen sich daher erfüllen, wenn $a_1 \neq 0$ ist, und dann mit $l_i = -a_i/a_1$ für beliebige Werte von a_i.

3.2.4. *Aussage.* Für $\boldsymbol{a} = (a_i) \in \mathbf{R}^n$, $a_1 \neq 0$, transformiert die durch

$$l_1 = 0, \qquad l_i = a_i/a_1 \qquad (i = 2, \ldots, n) \tag{7}$$

festgelegte LNT-Matrix $\boldsymbol{L}_1 = \boldsymbol{L}_1(-\boldsymbol{l})$, $\boldsymbol{l} = (l_i) \in \mathbf{R}^n$, den Vektor $\boldsymbol{a}$ in

$$\bar{\boldsymbol{a}} = \boldsymbol{L}_1(-\boldsymbol{l})\,\boldsymbol{a} = a_1\boldsymbol{e}^1 = (a_1, 0, \ldots, 0)^T.$$

Die Standardanwendung der LNT-Matrizen besteht in der Transformation von $\boldsymbol{a}$ in $\bar{\boldsymbol{a}} = \boldsymbol{L}_1\boldsymbol{a}$ gemäß 3.2.4 und der analogen Transformation einer Matrix $\boldsymbol{B}$ in $\bar{\boldsymbol{B}} = \boldsymbol{L}_1\boldsymbol{B}$.

3.2.5. *Eliminationsschritt mittels LNT-Matrix.*

Aufgabe: Für $\boldsymbol{a} = (a_i) \in \mathfrak{R}^n$ mit $a_1 \neq 0$ und $\boldsymbol{B} = (b_{ij}) \in \mathfrak{R}^{n,r}$ ist $\boldsymbol{L}_1 = \boldsymbol{L}_1(-\boldsymbol{l})$, $\boldsymbol{l} = (0, l_2, \ldots, l_n)^T \in \mathfrak{R}^n$ so festzulegen, daß

$$\bar{\boldsymbol{a}} = \boldsymbol{L}_1(-\boldsymbol{l})\,\boldsymbol{a} = (a_1, 0, \ldots, 0)^T$$

gilt. Mit dem derart bestimmten $\boldsymbol{L}_1$ ist $\bar{\boldsymbol{B}} = \boldsymbol{L}_1\boldsymbol{B}$ zu berechnen.

Algorithmus:

S1 (Festlegung von $\boldsymbol{l}$): $l_i := a_i/a_1$ $(i = 2, \ldots, n)$

S2 (Erste Zeile): $\bar{a}_1 := a_1$, $\bar{b}_{1j} := b_{1j}$ $(j = 1, \ldots, r)$

S3 (Übrige Zeilen): $\bar{a}_i := 0$, $\bar{b}_{ij} := b_{ij} - l_i * b_{1j}$ $(i = 2, \ldots, n,\ j = 1, \ldots, r)$

Aufwand: $(n-1)\,r$ opms $+ (n-1)$ opm

In den Anwendungen ist $\boldsymbol{a}$ in der Regel die erste Spalte einer Matrix $\boldsymbol{M} \in \mathfrak{R}^{n,r+1}$, und $\boldsymbol{B}$ besteht aus den restlichen Spalten 2 bis $r+1$. Der Algorithmus kann in situ durch Überspeichern von

$$\left(\begin{array}{c|c} a_1 & \ldots\ \boldsymbol{b}^{1T}\ \ldots \\ \hline a_2 & \ldots\ \boldsymbol{b}^{2T}\ \ldots \\ \vdots & \vdots \\ a_n & \ldots\ \boldsymbol{b}^{nT}\ \ldots \end{array}\right) \quad \text{mit} \quad \left(\begin{array}{c|c} a_1 & \ldots\ \boldsymbol{b}^{1T}\ \ldots \\ \hline l_2 & \ldots\ \bar{\boldsymbol{b}}^{2T}\ \ldots \\ \vdots & \vdots \\ l_n & \ldots\ \bar{\boldsymbol{b}}^{nT}\ \ldots \end{array}\right)$$

ausgeführt werden. Die erzeugten Nullen brauchen selbstverständlich nicht explizit gespeichert zu werden, und S2 wird überflüssig.

Mit den im Abschnitt 2.3 bereitgestellten Hilfsmitteln kann eine Rundungsfehleranalyse des Algorithmus 3.2.5 durchgeführt werden, siehe Ü 3.2.3. Wir geben hier nur die Ergebnisse an.

3.2.6. *Rundungsfehleranalyse.* Für $\boldsymbol{a} \in \mathfrak{R}^n$ mit $a_1 \neq 0$ und $\boldsymbol{B} \in \mathfrak{R}^{n,r}$ werde $\boldsymbol{L}_1 = \boldsymbol{L}_1(-\boldsymbol{l})$, $\boldsymbol{l} = (0, l_2, \ldots, l_n)^\mathsf{T}$, $\bar{\boldsymbol{a}} \in \mathfrak{R}^n$ und $\bar{\boldsymbol{B}} \in \mathfrak{R}^{n,r}$ nach 3.2.5 berechnet. Dann ist

$$\bar{\boldsymbol{a}} = (a_1, 0, \ldots, 0)^\mathsf{T} = \boldsymbol{L}_1(\boldsymbol{a} + \delta\boldsymbol{a}) \tag{8}$$

und

$$\bar{\boldsymbol{B}} = \boldsymbol{L}_1(\boldsymbol{B} + \delta\boldsymbol{B}) \tag{9}$$

mit Störungen $\delta\boldsymbol{a}$ und $\delta\boldsymbol{B}$, für die

$$\delta a_1 = 0, \qquad \delta b_{1j} = 0 \tag{10}$$

sowie

$$|\delta a_i| \leqq \nu\,|a_i|, \tag{11}$$

$$|\delta b_{ij}| \leqq \nu(|b_{ij}| + 2\,|\bar{b}_{ij}|), \tag{12}$$

$$|\delta b_{ij}| \leqq \nu(|b_{ij}| + 2\,|l_i|\,|b_{1j}|) \tag{13}$$

$(i = 2, \ldots, n,\ j = 1, \ldots, r)$ gilt.

Wir fassen jetzt die Eingangsdaten zur Matrix $\boldsymbol{M} = (\boldsymbol{a} \mid \boldsymbol{B})$ zusammen, analog wird $\delta\boldsymbol{M} = (\delta\boldsymbol{a} \mid \delta\boldsymbol{B})$ gebildet. Aus (11), (13) ergibt sich dann

$$|\delta\boldsymbol{M}| \leqq \nu(|\boldsymbol{M}| + 2\,|(\boldsymbol{o} \mid \boldsymbol{l}\boldsymbol{b}^{1\mathsf{T}})|),$$

wobei $\boldsymbol{b}^{1\mathsf{T}}$ die erste Zeile von $\boldsymbol{B}$ bezeichnet. Unter Beachtung von Ü 1.1.3 folgt hieraus $\|\delta\boldsymbol{M}\| \leqq \nu(\|\boldsymbol{M}\| + 2\,\|\boldsymbol{l}\|\,\|\boldsymbol{B}\|)$, wegen $\|\boldsymbol{B}\| \leqq \|\boldsymbol{M}\|$ also

$$\|\delta\boldsymbol{M}\| \leqq (1 + 2\,\|\boldsymbol{l}\|)\,\|\boldsymbol{M}\|. \tag{14}$$

Kombination von (11) und (12) liefert dagegen

$$|\delta\boldsymbol{M}| \leqq \nu\left\{|\boldsymbol{M}| + 2\left(\begin{array}{c|c} \boldsymbol{O} & \boldsymbol{O} \\ \hline \boldsymbol{O} & |\boldsymbol{M}^{(2)}| \end{array}\right)\right\},$$

wobei $\boldsymbol{M}^{(2)}$ aus den Zeilen 2 bis n von $\bar{\boldsymbol{B}}$, d. h. aus den in 3.2.5 tatsächlich neu berechneten Elementen besteht. Hieraus ergibt sich

$$\|\delta\boldsymbol{M}\| \leqq \nu(\|\boldsymbol{M}\| + 2\,\|\boldsymbol{M}^{(2)}\|). \tag{15}$$

Als Normen können in (14), (15) die Matrixnormen mit $p = 1, \infty, F$ verwendet werden, in (14) ist für $\boldsymbol{l}$ dann die Vektornorm mit $p = 1, \infty, 2$ zu nehmen. Im Fall der Spektralnorm ist auf den rechten Seiten $\|\boldsymbol{M}\|_2$ durch $\||\boldsymbol{M}|\|_2$ zu ersetzen usw., vgl. Abschnitt 2.3.C.

Die aus 3.2.6 folgenden Abschätzungen (14) und (15) besagen: *Der Eliminationsschritt* 3.2.5 *stellt einen numerisch gutartigen Algorithmus dar, wenn die Elininations-*

koeffizienten l_i *im Sinne von* $\|\boldsymbol{l}\| \leqq K_0$ *beschränkt sind* — aus (14) folgt dann $F = 1 + 2K_0$ —, *oder wenn* $\|\boldsymbol{M}^{(2)}\| \leqq K_1 \|\boldsymbol{M}\|$ *gilt, d. h., wenn die neu berechneten Elemente nicht zu groß werden* — aus (15) liest man dann $F = 1 + 2K_1$ ab.

3.2.7. *Bemerkung.* (i) Wir werden später sehen, daß die Bedingung $\|\boldsymbol{M}^{(2)}\| \leqq K_1 \|\boldsymbol{M}\|$ für gewisse Aufgabenklassen automatisch erfüllt ist, z. B. bei Anwendung von 3.2.5 im Gaußschen Algorithmus für Systeme mit diagonaldominanten oder symmetrischen definiten Matrizen, vgl. Kapitel 5. Obwohl dann die l_i durchaus groß werden können, liegt numerische Gutartigkeit vor.

(ii) In anderen Aufgabenklassen kann sowohl $\|\boldsymbol{l}\|$ als auch $\|\boldsymbol{M}^{(2)}\|$ beliebig groß werden, so daß der Algorithmus instabil wird. In diesem Fall kann die Beschränktheit der l_i durch die im folgenden beschriebene *Stabilisierung* von 3.2.5 erzwungen werden:

S0: Festlegung des Index s, $1 \leqq s \leqq n$, einer betragsgrößten Komponente von $\boldsymbol{a}$, d. h.

$$|a_s| = \max\{|a_i| : i = 1, \ldots, n\}. \tag{16}$$

Ersatz von $\boldsymbol{a}$ durch $\hat{\boldsymbol{a}} = \boldsymbol{T}_{1s}\boldsymbol{a}$, d. h. Vertauschung der Komponenten a_1 und a_s von $\boldsymbol{a}$. Analog Ersatz von $\boldsymbol{B}$ durch $\hat{\boldsymbol{B}} = \boldsymbol{T}_{1s}\boldsymbol{B}$, d. h. Vertauschung der Zeilen 1 und s von $\boldsymbol{B}$.

S1: Festlegung von $\hat{\boldsymbol{l}} = (0, \hat{l}_2, \ldots, \hat{l}_n)^\mathsf{T}$ gemäß $\hat{l}_i := \hat{a}_i/\hat{a}_1$ $(i = 2, \ldots, n)$. Wegen (16) ist

$$|\hat{l}_i| \leqq 1 \qquad (i = 2, \ldots, n), \quad \text{also} \quad \|\hat{\boldsymbol{l}}\|_\infty \leqq 1. \tag{17}$$

S2, S3: Wie in 3.2.5 mit $\hat{l}_i$, $\hat{a}_i$ und $\hat{b}_{ij}$ statt l_i, a_i und b_{ij}.

Diese Stabilisierung ist gleichbedeutend mit dem Ersatz von $\boldsymbol{L}_1(-\boldsymbol{l})$ durch $\boldsymbol{L}_1(-\hat{\boldsymbol{l}})\,\boldsymbol{T}_{1s}$. □

3.2.8. *Definition.* Eine LNT-Matrix $\boldsymbol{L}_k(\hat{\boldsymbol{l}}) = \boldsymbol{I} + \hat{\boldsymbol{l}}\boldsymbol{e}^{k\mathsf{T}}$, $\hat{\boldsymbol{l}} = (0, \ldots, 0, \hat{l}_{k+1}, \ldots, \hat{l}_n)^\mathsf{T}$, mit

$$|\hat{l}_i| \leqq 1 \qquad (i = k+1, \ldots, n)$$

heißt *stabilisierte LNT-Matrix*, abgekürzt: *SLNT-Matrix*.

Im Fall (17) geht die Abschätzung (14) für die ∞-Norm in

$$\|\delta\boldsymbol{M}\|_\infty \leqq 3\nu \|\boldsymbol{M}\| \tag{18}$$

über. Aus dieser Ungleichung folgt: *Der wie oben stabilisierte Algorithmus* 3.2.5 *ist in der Klasse aller Aufgaben mit* $\boldsymbol{a} \in \mathfrak{R}^n$, $\boldsymbol{a} \neq \boldsymbol{o}$, $\boldsymbol{B} \in \mathfrak{R}^{n,r}$ *numerisch gutartig mit* $F = 3$ *in der Maximumnorm.*

Übungsaufgaben

Ü 3.2.1. Man beweise: Für $\boldsymbol{a} = (a_i) \in \mathbf{R}^n$ mit $a_k \neq 0$ transformiert die Matrix $\boldsymbol{M}_k = \boldsymbol{M}_k(-\boldsymbol{m})$, $\boldsymbol{m} = (m_i) \in \mathbf{R}^n$ mit $m_k = 0$, $m_i = a_i/a_k$ für $i \neq k$ den Vektor $\boldsymbol{a}$ in $\bar{\boldsymbol{a}} = \boldsymbol{M}_k\boldsymbol{a} = a_k\boldsymbol{e}^k = (0, \ldots, 0, a_k, 0, \ldots, 0)^\mathsf{T}$.

Ü 3.2.2. Für $1 \leqq k < i \leqq s \leqq n$ ist die Identität $\boldsymbol{T}_{is}\boldsymbol{L}_k(\boldsymbol{l}) = \boldsymbol{L}_k(\boldsymbol{T}_{is}\boldsymbol{l})\,\boldsymbol{T}_{is}$ zu beweisen.

Ü 3.2.3. Man leite die in 3.2.6 angegebenen Fehleraussagen ab.

Hinweis: Die Schlüsseloperation aus Schritt S3 von 3.2.5 läßt sich in der Form

$$\bar{b}_{ij} = [b_{ij} - l_i b_{1j}(1 + \varepsilon)]\,(1 + \vartheta) \quad \text{mit} \quad \varepsilon = \varepsilon_{ij}, \quad \vartheta = \vartheta_{ij}, \quad |\varepsilon|, |\vartheta| \leqq \nu \tag{19}$$

schreiben, wobei $l_i = \mathrm{fl}(a_i/a_1)$ der berechnete Koeffizient ist. Hieraus folgt

$$\bar{b}_{ij} = b_{ij} + \delta b_{ij} - l_i b_{1j} \tag{20}$$

mit

$$\delta b_{ij} = b_{ij}\vartheta - l_i b_{1j}(\varepsilon + \vartheta + \varepsilon\vartheta) = b_{ij}\vartheta - l_i b_{1j}\eta, \qquad |\eta| \leqq 2\nu. \tag{21}$$

Mit der Festlegung $\delta b_{1j} = 0$ ergibt sich (9) aus (20), und aus (21) kann die Abschätzung (13) gefolgert werden. Auflösen von (19) nach b_{ij} und Einsetzen in (21) führt unter Beachtung von $\bar{b}_{1j} = b_{1j}$ auf die alternative Darstellung

$$\delta b_{ij} = \delta\bar{b}_{ij} = \bar{b}_{ij}\vartheta/(1 + \vartheta) - l_i\bar{b}_{1j}\varepsilon. \tag{22}$$

Schließlich erhält man durch Auflösen von (20) nach $l_i b_{1j}$ und Einsetzen in (22)

$$\delta b_{ij} = \bar{b}_{ij}\eta/[(1 + \vartheta)\,(1 + \varepsilon)] - b_{ij}\varepsilon/(1 + \varepsilon) \tag{23}$$

und damit die Abschätzung (12). Zum Nachweis von (11) beachte man

$$0 = \bar{a}_i = -l_i a_1 + a_i + \delta a_i \qquad (i = 2, \ldots, n)$$

und $l_i = (a_i/a_1)\,(1 + \varrho_i)$ mit $|\varrho_i| \leqq \nu$.

3.3. Orthogonale elementare Transformationsmatrizen

Neben den im vorangegangenen Abschnitt behandelten nichtorthogonalen elementaren Transformationsmatrizen stellen orthogonale elementare Transformationsmatrizen (OT-Matrizen) ein weiteres universelles Hilfsmittel der linearen Algebra dar. Sie spielen insbesondere bei Algorithmen zur Lösung von Quadratmittelproblemen und symmetrischen Eigenwertproblemen eine fundamentale Rolle.

Zu den OT-Matrizen zählen speziell Spiegelungs- und Drehungsmatrizen, die beide Gegenstand dieses Abschnitts sind. Zu den OT-Matrizen gehören ferner Permutationsmatrizen und orthogonale Diagonalmatrizen. Erstere sind im Abschnitt 3.1 ausführlich diskutiert worden, letztere sind von der einfachen Gestalt $\boldsymbol{D} = \mathrm{diag}\,(d_i)$, $|d_i| = 1$, also $d_i = \pm 1$, und können lediglich Vorzeichenänderungen bewirken, so daß sie nicht weiter untersucht zu werden brauchen.

Zur Definition und Charakterisierung von allgemeinen orthogonalen Matrizen sei auf Abschnitt 1.1.H verwiesen.

A. *Spiegelungsmatrizen*

3.3.1. *Definition.* Eine Matrix $\boldsymbol{H} \in \mathbf{R}^{n,n}$ der Gestalt

$$\boldsymbol{H} = \boldsymbol{I} - 2\boldsymbol{u}\boldsymbol{u}^{\mathsf{T}}, \qquad \boldsymbol{u} \in \mathbf{R}^n, \qquad \|\boldsymbol{u}\|_2 = 1 \tag{1}$$

heißt *Spiegelungsmatrix.*

Nach Definition ist eine Spiegelungsmatrix symmetrisch, und wegen $\boldsymbol{u}^T\boldsymbol{u} = 1$ gilt $\boldsymbol{H}^2 = (\boldsymbol{I} - 2\boldsymbol{u}\boldsymbol{u}^T)(\boldsymbol{I} - 2\boldsymbol{u}\boldsymbol{u}^T) = \boldsymbol{I} - 2\boldsymbol{u}\boldsymbol{u}^T - 2\boldsymbol{u}\boldsymbol{u}^T + 4\boldsymbol{u}\boldsymbol{u}^T\boldsymbol{u}\boldsymbol{u}^T = \boldsymbol{I}$, also

$$\boldsymbol{H} = \boldsymbol{H}^T = \boldsymbol{H}^{-1}. \tag{2}$$

Für $\boldsymbol{x} \in \mathbf{R}^n$ läßt sich $\boldsymbol{y} = \boldsymbol{H}\boldsymbol{x} = \boldsymbol{x} - 2\boldsymbol{z}$ mit $\boldsymbol{z} = \boldsymbol{u}(\boldsymbol{u}^T\boldsymbol{x})$ schreiben. Da $\boldsymbol{z}$ gerade die Projektion von $\boldsymbol{x}$ in Richtung $\boldsymbol{u}$ ist, kann die Abbildung

$$\boldsymbol{x} \to \boldsymbol{y} = \boldsymbol{H}\boldsymbol{x}$$

geometrisch wie folgt gedeutet werden, vgl. Abb. 3.3.1: *Das Bild* $\boldsymbol{y}$ *entsteht aus* $\boldsymbol{x}$ *durch Spiegelung an der Ursprungsebene mit dem Normalenvektor* $\boldsymbol{u}$. Dies erklärt die Bezeichnung Spiegelungsmatrix. Da solche Matrizen erstmalig von HOUSEHOLDER für numerische Zwecke verwendet worden sind, werden sie auch als Householdersche Spiegelungsmatrizen oder kurz als *Householder-Spiegelungen* bezeichnet, vgl. B 3.2.

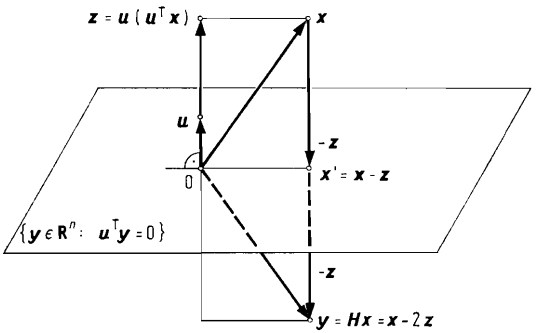


Abb. 3.3.1. Interpretation der Abbildung $\boldsymbol{x} \to \boldsymbol{y} = \boldsymbol{H}\boldsymbol{x}$ als Spiegelung

Aus (1) liest man ab, daß jedes Vielfache $\boldsymbol{v} = \mu\boldsymbol{u}$ von $\boldsymbol{u}$, $\mu \in \mathbf{R}$, $\mu \neq 0$, gemäß

$$\boldsymbol{H} = \boldsymbol{I} - \frac{\boldsymbol{v}\boldsymbol{v}^T}{\gamma} \quad \text{mit} \quad \gamma = \boldsymbol{v}^T\boldsymbol{v}/2 \tag{3}$$

dieselbe Spiegelungsmatrix wie (1) definiert. Für die numerische Praxis ist die Darstellung (3) wegen der größeren Freiheit in der Wahl des Vektors $\boldsymbol{v}$ jedoch günstiger als (1).

Für $\boldsymbol{b} \in \mathbf{R}^n$ gilt

$$\bar{\boldsymbol{b}} = \boldsymbol{H}\boldsymbol{b} = \left(\boldsymbol{I} - \frac{\boldsymbol{v}\boldsymbol{v}^T}{\gamma}\right)\boldsymbol{b} = \boldsymbol{b} - \boldsymbol{v}\beta \quad \text{mit} \quad \beta = \boldsymbol{v}^T\boldsymbol{b}/\gamma. \tag{4}$$

Wendet man (4) auf die Spalten $\boldsymbol{b}^j$ einer Matrix $\boldsymbol{B} \in \mathbf{R}^{n,r}$ an, so ergibt sich für die spaltenweise Berechnung von $\bar{\boldsymbol{B}} = \boldsymbol{H}\boldsymbol{B}$ die Vorschrift

$$\langle \text{neue } j\text{-te Spalte} \rangle := \langle \text{alte } j\text{-te Spalte} \rangle - \beta_j * \langle \text{Vektor } \boldsymbol{v} \rangle$$

mit dem für jede Spalte j zu berechnenden Faktor $\beta_j = (\boldsymbol{v}^T\boldsymbol{b}^j)/\gamma$. Die Matrix $\boldsymbol{H}$ braucht also nicht explizit gebildet zu werden.

3.3.2. *Algorithmus zur Householder-Transformation.* Für $\boldsymbol{B} = (\boldsymbol{b}^1, \ldots, \boldsymbol{b}^r) \in \mathfrak{R}^{n,r}$ und $\boldsymbol{v} \in \mathfrak{R}^n, \boldsymbol{v} \neq \boldsymbol{o}$, ist $\bar{\boldsymbol{B}} = (\bar{\boldsymbol{b}}^1, \ldots, \bar{\boldsymbol{b}}^r) = \boldsymbol{HB}$ zu berechnen, wobei $\boldsymbol{H} = \boldsymbol{I} - \boldsymbol{v}\boldsymbol{v}^\mathsf{T}/\gamma$ mit $\gamma = \boldsymbol{v}^\mathsf{T}\boldsymbol{v}/2$ durch $\boldsymbol{v}$ gegeben ist.

S0: $\gamma := \boldsymbol{v}^\mathsf{T}\boldsymbol{v}/2$

S1: **for** $j := 1(1)r$ **do** $\beta := \boldsymbol{v}^\mathsf{T}\boldsymbol{b}^j/\gamma$

$\bar{\boldsymbol{b}}^j := \boldsymbol{b}^j - \boldsymbol{v} * \beta$

Aufwand: $\sim n(2r + 1)$ opms

Die Skalarprodukte sollen dabei in üblicher Weise gemäß 2.3.9 berechnet werden. Algorithmus 3.3.2 kann in situ gemäß $\bar{\boldsymbol{B}} := \boldsymbol{HB}$ ausgeführt werden.

3.3.3. *Bemerkung.* (i) In den uns interessierenden Anwendungen fällt die Größe $\gamma = \boldsymbol{v}^\mathsf{T}\boldsymbol{v}/2$ bei der Festlegung von $\boldsymbol{v}$ ohnehin an, vgl. 3.3.4. Es ist daher zweckmäßig, die Spiegelung $\boldsymbol{H}$ durch die Daten $\{\boldsymbol{v}, \gamma\} = \{v_1, \ldots, v_n, \gamma\}$ zu repräsentieren. Bei der Bildung von $\bar{\boldsymbol{B}} = \boldsymbol{BH}$ gemäß 3.3.2 entfällt dann die Neuberechnung von γ im Schritt S0. Wir können dann außerdem $\boldsymbol{v}$ so skalieren, daß $v_1 = \gamma$ gilt, so daß kein zusätzlicher Speicherplatz für γ gebraucht wird, siehe 3.3.5(ii).

(ii) Algorithmus 3.3.2 kann bei Bedarf zeilenweise in situ realisiert werden, wenn Schritt S1 wie folgt abgeändert wird:

S1′: **for** $j := 1(1)r$ **do** $\beta_j := (\boldsymbol{v}^\mathsf{T}\boldsymbol{b}^j)/\gamma$

for $i := 1(1)n$ **do**

for $j := 1(1)r$ **do** $\bar{b}_{ij} := b_{ij} - v_i * \beta_j$

Dabei wird ein Hilfsfeld der Länge r zum Speichern der Koeffizienten $\{\beta_1, \ldots, \beta_r\}$ benötigt. □

Im folgenden untersuchen wir, ob mittels einer Householder-Spiegelung ähnlich wie mit einer LNT-Matrix in einem Vektor $\boldsymbol{a}$ unterhalb der ersten Komponente Nullen erzeugt werden können, d. h., ob

$$\boldsymbol{Ha} = \bar{\boldsymbol{a}} = \varrho\boldsymbol{e}^1 = (\varrho, 0, \ldots, 0)^\mathsf{T} \tag{5}$$

durch geeignete Wahl von $\boldsymbol{v}$ erzielt werden kann. Wird auf beiden Seiten von (5 die Euklidische Norm gebildet, so ergibt sich wegen deren Orthogonalinvarianz $\|\boldsymbol{a}\|_2 = \|\boldsymbol{Ha}\|_2 = \|\varrho\boldsymbol{e}^1\| = |\varrho|$, also

$$\varrho = \pm\|\boldsymbol{a}\|_2. \tag{6}$$

Analog zu (4) folgt weiter $\boldsymbol{Ha} = \boldsymbol{a} - \boldsymbol{v}(\boldsymbol{v}^\mathsf{T}\boldsymbol{a})/\gamma = \varrho\boldsymbol{e}^1$, also $\boldsymbol{v}(\boldsymbol{v}^\mathsf{T}\boldsymbol{a})/\gamma = \boldsymbol{a} - \varrho\boldsymbol{e}^1$. Falls $\boldsymbol{a}$ nicht schon von der gewünschten Form $\boldsymbol{a} = a_1\boldsymbol{e}^1$ ist, ergibt sich für jede der beiden Festlegungen (6) ein von $\boldsymbol{o}$ verschiedener Vektor $\boldsymbol{a} - \varrho\boldsymbol{e}^1$, so daß $\boldsymbol{v} = \lambda(\boldsymbol{a} - \varrho\boldsymbol{e}^1)$ für jedes $\lambda \neq 0$ die gestellte Aufgabe löst. Der zugehörige γ-Wert ist

$$\gamma = \lambda^2(\boldsymbol{a} - \varrho\boldsymbol{e}^1)^\mathsf{T}(\boldsymbol{a} - \varrho\boldsymbol{e}^1)/2 = \lambda^2(\|\boldsymbol{a}\|_2^2 - 2\varrho a_1 + \varrho^2)/2$$

$$= \lambda^2\varrho(\varrho - a_1) = -\lambda\varrho v_1.$$

3.3.4. *Aussage.* Für $\boldsymbol{a} \in \mathbf{R}^n$, $\boldsymbol{a} \neq a_1 \boldsymbol{e}^1$, und jedes $\lambda \neq 0$ transformiert die durch

$$\boldsymbol{v} = (v_i) = \lambda(\boldsymbol{a} - \varrho \boldsymbol{e}^1), \qquad \varrho = \pm\|\boldsymbol{a}\|_2, \tag{7}$$

festgelegte Householder-Spiegelung

$$\boldsymbol{H} = \boldsymbol{I} - \boldsymbol{v}^{\mathsf{T}}\boldsymbol{v}/\gamma, \qquad \gamma = \boldsymbol{v}^{\mathsf{T}}\boldsymbol{v}/2 = -\lambda\varrho v_1 \tag{8}$$

den Vektor $\boldsymbol{a}$ in

$$\bar{\boldsymbol{a}} = \boldsymbol{H}\boldsymbol{a} = \varrho \boldsymbol{e}^1 = (\varrho, 0, \ldots, 0)^{\mathsf{T}}.$$

Die beiden Möglichkeiten $\varrho = \pm\|\boldsymbol{a}\|_2$ aus 3.3.4 sind in Abb. 3.3.2 für $n = 2$ und $\lambda = 1$ mit den zugehörigen Vektoren $\boldsymbol{v}^+ = \boldsymbol{a} - \|\boldsymbol{a}\|_2 \boldsymbol{e}^1$, $\boldsymbol{v}^- = \boldsymbol{a} + \|\boldsymbol{a}\|_2 \boldsymbol{e}^1$ dargestellt worden. Man beachte, daß $\boldsymbol{v}^{+\mathsf{T}}\boldsymbol{v}^- = 0$ gilt, d. h., die Normalenvektoren der beiden möglichen spiegelnden Ebenen sind orthogonal.

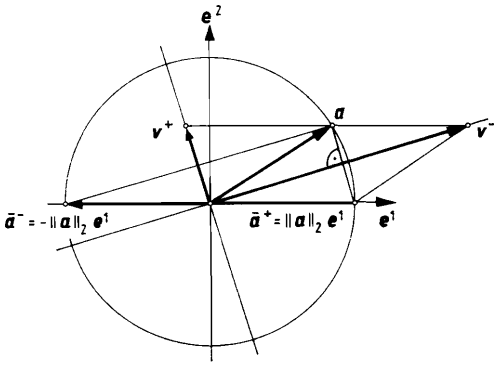


Abb. 3.3.2. Householder-Transformation 3.3.4 für $n = 2$ und $\lambda = 1$

3.3.5. *Bemerkung.* (i) Wird $\varrho = \|\boldsymbol{a}\|_2$ gesetzt, so ergibt sich aus (7) und (8)

$$v_1 = \lambda(a_1 - \|\boldsymbol{a}\|_2).$$

Die wörtliche numerische Auswertung von $a_1 - \|\boldsymbol{a}\|_2$ ist für $a_1 \leqq 0$ stabil. Für $a_1 > 0$ tritt jedoch im Fall $a_1 \approx \|\boldsymbol{a}\|_2$ bei der Differenzbildung Auslöschung auf, so daß ein instabiles Verfahren entsteht. Unter Beachtung der Identität

$$a_1 - \|\boldsymbol{a}\|_2 = -\left(\sum_{i=2}^{n} a_i^2\right) \Big/ (a_1 + \|\boldsymbol{a}\|_2)$$

kann dieser negative Effekt vermieden werden, vgl. Ü 2.2.5. Insgesamt liefert die Auswertung gemäß

$$v_1 = \begin{cases} \lambda(a_1 - \|\boldsymbol{a}\|_2) & \text{für} \quad a_1 \leqq 0, \\ -\lambda\left(\sum_{i=2}^{n} a_i^2\right) \Big/ (a_1 + \|\boldsymbol{a}\|_2) & \text{für} \quad a_1 > 0, \end{cases} \tag{9}$$

$$v_i = \lambda a_i \qquad (i = 2, \ldots, n)$$

ein stabiles Verfahren zur Berechnung von $\boldsymbol{v}$ und $\gamma = -\lambda \|\boldsymbol{a}\|_2 v_1$.

(ii) Eine einfachere Stabilisierung kann erreicht werden, wenn das Vorzeichen in (6) so gewählt wird, daß $\gamma = \lambda^2\varrho(\varrho - a_1) = \lambda^2(\|\boldsymbol{a}\|_2^2 - \varrho a_1)$ maximal wird. Dies ist bei der Festlegung

$$\varrho = \begin{cases} \|\boldsymbol{a}\|_2 & \text{für} \quad a_1 \leqq 0, \\ -\|\boldsymbol{a}\|_2 & \text{für} \quad a_1 > 0 \end{cases} \tag{10}$$

der Fall. Für den Skalierungsfaktor λ bietet sich wegen (8) der Wert $\lambda = -1/\varrho$ an, denn dann ist $\gamma = v_1$, vgl. 3.3.3(i). Damit ergibt sich

$$\boldsymbol{v} = (v_i) = \boldsymbol{e}^1 - (1/\varrho)\,\boldsymbol{a}, \tag{11}$$

also

$$v_1 = \gamma = 1 - a_1/\varrho = 1 + |a_1|/\|\boldsymbol{a}\|_2,$$
$$v_i = -a_i/\varrho \qquad (i = 2, \ldots, n).$$

Wegen $1 \leqq \gamma = \|\boldsymbol{v}\|_2^2/2 \leqq 2$ gilt dabei

$$\sqrt{2} \leqq \|\boldsymbol{v}\|_2 \leqq 2,$$

d. h., $\boldsymbol{v}$ ist gut skaliert, und bei der Berechnung von $\beta_j := \boldsymbol{v}^T\boldsymbol{b}^j/\gamma$ in 3.3.2 ist weder mit Unter- noch mit Überlauf zu rechnen. Im folgenden soll $\boldsymbol{v}$ stets gemäß (10), (11) festgelegt werden. Dies ist für beliebiges $\boldsymbol{a} \neq \boldsymbol{o}$, also auch im Fall $\boldsymbol{a} = a_1\boldsymbol{e}^1$, zulässig. □

Die typische Anwendung der Householder-Spiegelungen besteht wie die der LNT-Matrizen in der Transformation eines Vektors $\boldsymbol{a}$ in ein Vielfaches von $\boldsymbol{e}^1$ und der gleichzeitigen Transformation einer Matrix $\boldsymbol{B}$.

3.3.6. *Eliminationsschritt mittels Householder-Spiegelung.*

Aufgabe: Für $\boldsymbol{a} = (a_i) \in \Re^n$, $\boldsymbol{a} \neq \boldsymbol{o}$, und $\boldsymbol{B} = (\boldsymbol{b}^1, \ldots, \boldsymbol{b}^r) \in \Re^{n,r}$ sind $\boldsymbol{v} = (v_i) \in \Re^n$ und $\varrho \in \Re$ so festzulegen, daß $\bar{\boldsymbol{a}} = \boldsymbol{H}\boldsymbol{a} = \varrho\boldsymbol{e}^1 = (\varrho, 0, \ldots, 0)^T$ mit $\boldsymbol{H} = \boldsymbol{I} - \boldsymbol{v}\boldsymbol{v}^T/\gamma$ und $\gamma = \boldsymbol{v}^T\boldsymbol{v}/2 = v_1$ gilt. Für das so definierte $\boldsymbol{H}$ ist $\bar{\boldsymbol{B}} = (\bar{\boldsymbol{b}}^1, \ldots, \bar{\boldsymbol{b}}^r) = \boldsymbol{H}\boldsymbol{B}$ zu berechnen.

Algorithmus:

S0 (Festlegung von ϱ und $\boldsymbol{v}$): $\varrho := \|\boldsymbol{a}\|_2$
if $a_1 > 0$ **then** $\varrho := -\varrho$
$v_1 := 1 - a_1/\varrho$
for $i := 2(1)n$ **do** $v_i := -a_i/\varrho$

S1 (Festlegung von $\bar{\boldsymbol{a}}$): $\bar{a}_1 := \varrho$
for $i := 2(1)n$ **do** $\bar{a}_i := 0$

S2 (Berechnung von $\boldsymbol{B}$): **for** $j := 1(1)r$ **do**
$\beta := \boldsymbol{v}^T\boldsymbol{b}^j/v_1$
$\bar{\boldsymbol{b}}^j := \boldsymbol{b}^j - \boldsymbol{v} * \beta$

Aufwand: $\sim n[(2r + 1)$ opms $+ 1$ opm$] + 1$ opr, wenn $\|\boldsymbol{a}\|_2$ mit $\sim n$ opms $+ 1$ opr berechnet wird.

Der Algorithmus kann in situ durch Überspeichern von $\boldsymbol{a}$ mit $\boldsymbol{v}$ und $\boldsymbol{B}$ mit $\bar{\boldsymbol{B}}$ durchgeführt werden. Das neu berechnete Element $\bar{a}_1 = \varrho$ muß dann auf einem ge-

sonderten Platz aufgehoben werden, während die $n-1$ erzeugten Nullen $\bar{a}_2, \ldots, \bar{a}_n$ selbstverständlich nicht gespeichert werden. Bei dieser Realisierung geht

$$\left(\begin{array}{c|ccc} a_1 & \vdots & \vdots & \vdots \\ a_2 & \boldsymbol{b}^1 & \boldsymbol{b}^2 & \ldots\ \boldsymbol{b}^r \\ \vdots & \vdots & \vdots & \vdots \\ a_n & \vdots & \vdots & \vdots \end{array}\right) \text{ in } \left(\begin{array}{c|ccc} v_1 & \vdots & \vdots & \vdots \\ v_2 & \bar{\boldsymbol{b}}^1 & \bar{\boldsymbol{b}}^2 & \ldots\ \bar{\boldsymbol{b}}^r \\ \vdots & \vdots & \vdots & \vdots \\ v_n & \vdots & \vdots & \vdots \end{array}\right)$$

über. Falls $\boldsymbol{a}$ schlecht skaliert ist, sollte bei der Berechnung von $\|\boldsymbol{a}\|_2$ Über- oder Unterlauf vermieden werden, siehe 2.3.E.

Die Rundungsfehleranalyse führt auf die folgenden Ergebnisse:

3.3.7. *Rundungsfehleranalyse.* Algorithmus 3.3.6 werde für $\boldsymbol{a} \in \mathfrak{R}^n$, $\boldsymbol{a} \neq \boldsymbol{o}$, und $\boldsymbol{B} \in \mathfrak{R}^{n,r}$ durchgeführt, und $\bar{\boldsymbol{a}} = (\varrho, 0, \ldots, 0)^T \in \mathfrak{R}^n$ sowie $\bar{\boldsymbol{B}} \in \mathfrak{R}^{n,r}$ seien die berechneten Werte. Mit $\boldsymbol{H}^*$ werde die Spiegelungsmatrix bezeichnet, die bei exakter Ausführung von 3.3.6 durch $\boldsymbol{a}$ festgelegt wird. Dann gilt

$$\bar{\boldsymbol{a}} = (\varrho, 0, \ldots, 0)^T = \boldsymbol{H}^*(\boldsymbol{a} + \delta\boldsymbol{a}) \tag{12}$$

sowie

$$\bar{\boldsymbol{B}} = \boldsymbol{H}^*(\boldsymbol{B} + \delta\boldsymbol{B}) = \boldsymbol{H}^*\boldsymbol{B} + \delta\bar{\boldsymbol{B}} \tag{13}$$

mit Störungen $\delta\boldsymbol{a}$, $\delta\boldsymbol{B}$ und $\delta\bar{\boldsymbol{B}}$, die den Abschätzungen

$$|\delta\boldsymbol{a}| \leqq \nu[(n+2)/2]\,|\boldsymbol{a}| \tag{14}$$

und

$$\|\delta\boldsymbol{b}^j\|_2 \leqq \nu F\,\|\boldsymbol{b}^j\|_2, \qquad \|\delta\bar{\boldsymbol{b}}^j\|_2 \leqq \nu F\,\|\delta\bar{\boldsymbol{b}}^j\|_2 \tag{15}$$

mit

$$F := F(n) := K(n)n,$$

$$K(n) := 3.14(1 + 3/n), \quad \text{also} \quad F(n) \leqq 4 \ \text{ für } n \geqq 10 \tag{16}$$

genügen, wobei $\boldsymbol{b}^j$ bzw. $\bar{\boldsymbol{b}}^j$ die Spalten von $\boldsymbol{B}$ bzw. $\bar{\boldsymbol{B}}$ bezeichnen.

Aus den spaltenweisen Abschätzungen (15) ergibt sich sofort

$$\|\delta\boldsymbol{B}\|_F \leqq \nu F\,\|\boldsymbol{B}\|_F \quad \text{bzw.} \quad \|\delta\bar{\boldsymbol{B}}\|_F \leqq \nu F\,\|\bar{\boldsymbol{B}}\|_F \tag{17}$$

als kollektive Abschätzung in der Frobeniusnorm.

Die Ergebnisse von 3.3.7 besagen: *Die Ausführung eines Eliminationsschrittes mittels Householder-Spiegelungen gemäß* 3.3.6 *ist ein numerisch gutartiger Prozeß mit* $F = 4n$ *für* $n \geqq 10$ *in der Euklidischen bzw. Frobenius-Norm, und die Störungen sind spaltenweise klein.*

A. *Drehungsmatrizen*

Die zweite Klasse elementarer orthogonaler Matrizen besteht aus Matrizen des Typs

$$\boldsymbol{G}_{pq} = \begin{pmatrix} & & \boldsymbol{p} & & \boldsymbol{q} & \\ & 1 & \vdots & & \vdots & \\ & \ldots & c & \ldots & s & \ldots \\ & & \vdots & & \vdots & \\ & \ldots & -s & \ldots & c & \ldots \\ & & \vdots & & \vdots & 1 \end{pmatrix} \begin{matrix} \\ \\ p\text{-te Zeile} \\ \\ q\text{-te Zeile} \\ \\ \end{matrix} \quad \text{mit} \quad c^2 + s^2 = 1.$$

3.3.8. *Definition.* Eine Matrix $\boldsymbol{G}_{pq} \in \mathbf{R}^{n,n}$ $(1 \leqq p < q \leqq n)$, die sich von der Einheitsmatrix nur in den vier Elementen

$$(\boldsymbol{G}_{pq})_{pp} = (\boldsymbol{G}_{pq})_{qq} = c, \qquad (\boldsymbol{G}_{pq})_{pq} = -(\boldsymbol{G}_{pq})_{qp} = s$$

unterscheidet und für die

$$c^2 + s^2 = 1 \tag{18}$$

gilt, heißt *Drehungsmatrix*. Im Fall $p = q$ wird $\boldsymbol{G}_{pp} = \boldsymbol{G}_{qq} = \boldsymbol{I}$ gesetzt.

Aus (18) folgt in der Tat $\boldsymbol{G}\boldsymbol{G}^{\mathsf{T}} = \boldsymbol{I}$, also die Orthogonalität von $\boldsymbol{G}$. Die Transformation

$$\boldsymbol{x} \to \boldsymbol{y} = \boldsymbol{G}_{pq}\boldsymbol{x}$$

lautet komponentenweise

$$\begin{aligned} y_p &= c * x_p + s * x_q, \\ y_q &= -s * x_p + c * x_q, \\ y_i &= x_i \qquad (i \neq p, q;\ i = 1, \ldots, n). \end{aligned} \tag{19}$$

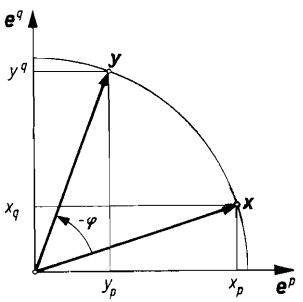


Abb. 3.3.3. Interpretation der Abbildung $\boldsymbol{x} \to \boldsymbol{y} = \boldsymbol{G}_{pq}\boldsymbol{x}$ als ebene Drehung

Sie verändert nur die Komponenten p und q von $\boldsymbol{x}$ und läßt sich als Drehung um den Winkel $-\varphi$ in der durch $\boldsymbol{e}^p$ und $\boldsymbol{e}^q$ aufgespannten Ebene deuten, wobei φ durch die Bedingungen

$$c = \cos\varphi, \qquad s = \sin\varphi$$

festgelegt ist, siehe Abb. 3.3.3. Man nennt $\boldsymbol{G}_{pq}$ deshalb auch *ebene Drehung* oder — nach GIVENS, der diese Matrizen für die numerische lineare Algebra wiederentdeckt hat, vgl. B 3.3 — *Givens-Drehung*.

Ist $\boldsymbol{B} \in \mathbf{R}^{n,r}$ eine Matrix mit den Zeilen $\boldsymbol{b}^{i\mathsf{T}}$, so ergeben sich die Zeilen $\bar{\boldsymbol{b}}_i^{\mathsf{T}}$ von $\bar{\boldsymbol{B}} = \boldsymbol{G}_{pq}\boldsymbol{B}$ analog zu (19) gemäß

$$\langle\text{neue } p\text{-te Zeile}\rangle := c * \langle\text{alte } p\text{-te Zeile}\rangle + s * \langle\text{alte } q\text{-te Zeile}\rangle,$$

$$\langle\text{neue } q\text{-te Zeile}\rangle := -s * \langle\text{alte } p\text{-te Zeile}\rangle + c * \langle\text{alte } q\text{-te Zeile}\rangle;$$

die übrigen Zeilen von $\bar{\boldsymbol{B}}$ werden unverändert von $\boldsymbol{B}$ übernommen.

Wenn diese Vorschrift auch in situ realisiert werden soll, muß eines der neuen Elemente zwischengespeichert werden wie im folgenden Algorithmus:

3.3.9. *Algorithmus zur Givens-Transformation.* Bei gegebener Givens-Drehung $\boldsymbol{G}_{pq} \in \mathfrak{R}^{n,n}$ mit $p < q$ und den Drehungsparametern $c, s \in \mathfrak{R}$ ist $\boldsymbol{B} = (b_{ij}) \in \mathfrak{R}^{n,r}$ mit $\overline{\boldsymbol{B}} = \boldsymbol{G}_{pq}\boldsymbol{B}$ zu überspeichern.

for $j := 1(1)r$ **do**
$\quad \mu := c * b_{pj} + s * b_{qj}$
$\quad b_{qj} := -s * b_{pj} + c * b_{qj}$
$\quad b_{pj} := \mu$

Aufwand: $4r$ opm + $2r$ ops

Wie die bisher eingeführten Transformationsmatrizen werden auch Givens-Drehungen verwendet, um durch geeignete Festlegung von c und s in einem Vektor $\boldsymbol{a} = (a_i)$ an vorgegebenen Stellen Nullen zu erzeugen. Da beim Übergang von $\boldsymbol{a}$ zu $\overline{\boldsymbol{a}} = \boldsymbol{G}_{pq}\boldsymbol{a}$ nur die Komponenten p und q verändert werden, kann nur eines dieser Elemente zu Null gemacht werden, etwa

$$\bar{a}_q = -s * a_p + c * a_q = 0. \tag{20}$$

Wegen der Orthogonalität von $\boldsymbol{G}_{pq}$ muß dann $|\bar{a}_p| = \sqrt{a_p^2 + a_q^2}$ gelten. Die Bedingung (20) läßt sich wie folgt erfüllen:

3.3.10. *Aussage.* Es seien p, q mit $1 \leqq p < q \leqq n$ und $\boldsymbol{a} = (a_i) \in \mathbf{R}^n$ mit $a_p^2 + a_q^2 > 0$ gegeben. Dann transformiert die durch

$$c = a_p/\varrho, \qquad s = a_q/\varrho \quad \text{mit} \quad \varrho = \pm\sqrt{a_p^2 + a_q^2} \tag{21}$$

festgelegte Givens-Drehung $\boldsymbol{G}_{pq}$ den Vektor $\boldsymbol{a}$ in $\overline{\boldsymbol{a}} = (\bar{a}_i) = \boldsymbol{G}_{pq}\boldsymbol{a}$ mit

$$\bar{a}_p = \varrho, \qquad \bar{a}_q = 0 \quad \text{und} \quad \bar{a}_i = a_i \qquad (i \neq p, q).$$

Dies läßt sich durch Einsetzen sofort überprüfen.

Für die numerische Realisierung sind die Formeln (21) wegen des möglichen Über- bzw. Unterlaufs bei der Berechnung von $a_p^2 + a_q^2$ nicht zu empfehlen, vgl. Abschnitt 2.3.E. Eine zuverlässige Realisierung von 3.3.10 gemäß

$$\varrho = \begin{cases} a_p\sqrt{1 + (a_q/a_p)^2} & \text{im Fall} \quad |a_p| \geqq |a_q|, \\ a_q\sqrt{1 + (a_p/a_q)^2} & \text{im Fall} \quad |a_p| < |a_q| \end{cases}$$

wird durch das nachfolgende Verfahren gegeben:

3.3.11. *Algorithmus.* Berechnung der Drehungsparameter c, s und der Größen ϱ, ξ aus $a_p, a_q \in \mathfrak{R}$.

if $|a_p| \geqq |a_q|$ **then**
$\quad$ **if** $a_p \neq 0$
$\quad$ **then** $\xi := a_q/a_p$, $\varrho :=$ sqrt $(1 + \xi^2)$, $c := 1/\varrho$
$\qquad s := \xi := \xi/\varrho$, $\varrho := a_p * \varrho$
$\quad$ **else** $[c := 1, s := \varrho := \xi := 0]$

else | $\xi := a_p/a_q$, $\varrho := \mathrm{sqrt}\,(1 + \xi^2)$, $s := 1/\varrho$
$c := \xi/\varrho$, $\varrho := a_q * \varrho$
if $c \neq 0$ **then** $\xi := 1/c$ **else** $\xi := 1$

Aufwand: 6 opm + 1 ops + 1 opr

Der Algorithmus arbeitet auch im Fall $a_p = a_q = 0$, wo $c = 1$, $s = 0$, also $\boldsymbol{G}_{pq} = \boldsymbol{I}$ gesetzt wird. Er stellt neben c, s, ϱ außerdem die Zahl ξ bereit, aus der c, s für spätere Anwendungen von $\boldsymbol{G}_{pq}$ einfach rekonstruiert werden kann. Es genügt daher, statt der beiden Zahlen c, s nur die eine Zahl ξ z. B. auf dem Platz von $\bar{a}_q = 0$ zu speichern, siehe B 3.3.

3.3.12. *Algorithmus.* Rekonstruktion von c, s aus $\xi \in \mathfrak{R}$.

if $|\xi| \leqq 1$ **then** | **if** $\xi \neq 1$
then $c := \mathrm{sqrt}\,(1 - \xi^2)$, $s := \xi$
else $c := 0$, $s := 1$
else $[c := 1/\xi, s := \mathrm{sqrt}\,(1 - c^2)]$

Aufwand: 2 opm + 1 ops + 1 opr

Man beachte dabei, daß für das vom Algorithmus 3.3.11 bereitgestellte ξ genau eine der Bedingungen $|\xi| < 1/\sqrt{2}$, $|\xi| > \sqrt{2}$ oder $\xi = 1$ erfüllt ist, so daß weder bei der Berechnung von $1 - \xi^2$ noch bei der von $1 - c^2$ Auslöschung eintreten kann.

Soll ein Vektor $\boldsymbol{a} = \boldsymbol{a}^1 \in \mathbf{R}^n$ analog zu 3.3.4 mittels Givens-Drehungen in den Vektor $\bar{\boldsymbol{a}} = (\varrho, 0, \ldots, 0)^\mathsf{T}$ transformiert werden, so kann dies schrittweise gemäß

$$\boldsymbol{a}^1 \to \boldsymbol{a}^2 = \boldsymbol{G}_{12}\boldsymbol{a}^1 \to \boldsymbol{a}^3 = \boldsymbol{G}_{13}\boldsymbol{a}^2 \to \cdots \to \bar{\boldsymbol{a}} = \boldsymbol{a}^n = \boldsymbol{G}_{1n}\boldsymbol{a}^{n-1} \tag{22}$$

geschehen, wobei $\boldsymbol{G}_{1i}$ in der i-ten Komponente von $\boldsymbol{a}^{i-1}$ eine Null erzeugt. Dabei bleiben die vorher bereits erzeugten Nullen offensichtlich erhalten.

3.3.13. *Aussage.* Zu jedem Vektor $\boldsymbol{a} \in \mathbf{R}^n$ gibt es eine Folge $\boldsymbol{G}_{12}, \boldsymbol{G}_{13}, \ldots, \boldsymbol{G}_{1n}$ von Givens-Drehungen, so daß

$$\boldsymbol{G} = \boldsymbol{G}_{1n}\boldsymbol{G}_{1,n-1} \cdots \boldsymbol{G}_{12} \tag{23}$$

den Vektor $\boldsymbol{a}$ in

$$\bar{\boldsymbol{a}} = \boldsymbol{G}\boldsymbol{a} = \varrho\boldsymbol{e}^1 = (\varrho, 0, \ldots, 0)^\mathsf{T} \quad \text{mit} \quad \varrho = \pm\|\boldsymbol{a}\|_2$$

transformiert.

Nachfolgend geben wir eine in-situ-Realisierung von 3.3.13 an. Wie in den vorangegangenen Abschnitten wird zusätzlich $\bar{\boldsymbol{B}} = \boldsymbol{G}\boldsymbol{B}$ für gegebenes $\boldsymbol{B} \in \mathfrak{R}^{n,r}$ berechnet, und zwar analog zu (22) gemäß

$$\boldsymbol{B}^{(1)} = \boldsymbol{B}, \qquad \boldsymbol{B}^{(i)} = \boldsymbol{G}_{1i}\boldsymbol{B}^{(i-1)} \qquad (i = 2, \ldots, n), \qquad \bar{\boldsymbol{B}} = \boldsymbol{B}^{(n)}. \tag{24}$$

3.3.14. *Eliminationsschritt mittels einer Folge von Givens-Drehungen.*

Aufgabe: Für $\boldsymbol{a} = (a_i) \in \mathfrak{R}^n$ und $\boldsymbol{B} = (b_{ij}) \in \mathfrak{R}^{n,r}$ sind die Drehungen $\boldsymbol{G}_{1i} \in \mathfrak{R}^{n,n}$ $(i = 2, \ldots, n)$ so festzulegen, daß $\bar{\boldsymbol{a}} = \boldsymbol{G}_{1n} \cdots \boldsymbol{G}_{12}\boldsymbol{a} = \varrho\boldsymbol{e}^1$ gilt. Der Vektor $\boldsymbol{a}$ ist mit $(\varrho, \xi_2, \ldots, \xi_n)^\mathsf{T}$ zu überspeichern, wobei ξ_i die der Drehung $\boldsymbol{G}_{1i}$ gemäß 3.3.11

zugeordnete Zahl bezeichne. Die Matrix $\boldsymbol{B}$ ist durch $\overline{\boldsymbol{B}} = \boldsymbol{G}_{1n} \cdots \boldsymbol{G}_{12}\boldsymbol{B}$ zu überspeichern.

Algorithmus:

for $i := 2(1)n$ **do**

S0 (Festlegung von $\boldsymbol{G}_{1i}$): Berechne
$c := c(a_1, a_i)$, $s := s(a_1, a_i)$, $\varrho := \varrho(a_1, a_i)$, $\xi = \xi(a_1, a_i)$ gemäß Algorithmus 3.3.11 mit $a_p = a_1$ und $a_q = a_i$

S1 (Transformation $\boldsymbol{a}^{i-1} \to \boldsymbol{a}^i := \boldsymbol{G}_{1i}\boldsymbol{a}^{i-1}$, Reservierung von $\xi = \xi_i$):
$a_1 := \varrho$, $a_i := \xi$

S2 (Transformation $\boldsymbol{B}^{(i-1)} \to \boldsymbol{B}^{(i)} := \boldsymbol{G}_{1i}\boldsymbol{B}^{(i-1)}$):
for $j := 1(1)r$ **do**
$\mu := c * b_{1j} + s * b_{ij}$
$b_{ij} := -s * b_{1j} + c * b_{ij}$
$b_{1j} := \mu$

Aufwand: Festlegung der Drehungsparameter und Berechnung von $\bar{\boldsymbol{a}}$:
$(n-1)$ (6 opm + 1 ops + 1 opr)
Berechnung von $\overline{\boldsymbol{B}}$: $(n-1)\,r$ (4 opm + 2 ops)

Durch 3.3.14 wird

$$\left(\begin{array}{c|ccc} a_1 & b_{11} & \dots & b_{1r} \\ a_2 & b_{21} & \dots & b_{2r} \\ \vdots & \vdots & & \vdots \\ a_n & b_{n1} & \dots & b_{nr} \end{array}\right) \quad \text{mit} \quad \left(\begin{array}{c|ccc} \varrho & \bar{b}_{11} & \dots & \bar{b}_{1r} \\ \xi_2 & \bar{b}_{21} & \dots & \bar{b}_{2r} \\ \vdots & \vdots & & \vdots \\ \xi_n & \bar{b}_{n1} & \dots & \bar{b}_{nr} \end{array}\right)$$

überspeichert. Man beachte dabei, daß die Produktmatrix $\boldsymbol{G}$ in 3.3.14 nicht explizit gebildet wird. Aus den Zahlen ξ_i $(i = 2, \dots, n)$ können die Drehungen $\boldsymbol{G}_{1i}$ bei Bedarf mittels Algorithmus 3.3.12 rekonstruiert werden.

3.3.15. *Rundungsfehleranalyse.* Für $\boldsymbol{a} \in \mathfrak{R}^n$ und $\boldsymbol{B} \in \mathfrak{R}^{n,r}$ werde Algorithmus 3.3.14 durchgeführt, und $\bar{\boldsymbol{a}} = (\varrho, 0, \dots, 0)^T \in \mathfrak{R}^n$ sowie $\overline{\boldsymbol{B}} \in \mathfrak{R}^{n,r}$ seien die berechneten Resultate. Mit $\boldsymbol{G}_{1i}^* \in \mathbf{R}^{n,n}$ werde die exakte Givens-Drehung bezeichnet, die den berechneten Vektor $\boldsymbol{a}^{i-1} \in \mathfrak{R}^n$ entsprechend 3.3.11 exakt in $\boldsymbol{G}_{1i}\boldsymbol{a}^{i-1}$ mit $(\boldsymbol{G}_{1i}\boldsymbol{a}^{i-1})_i = 0$ transformiert, und es sei $\boldsymbol{G}^* = \boldsymbol{G}_{1n}^* \cdots \boldsymbol{G}_{12}^*$ das exakte Produkt der Drehungen $\boldsymbol{G}_{1i}^*$. Dann gelten die Beziehungen

$$\bar{\boldsymbol{a}} = (\varrho, 0, \dots, 0)^T = \boldsymbol{G}^*(\boldsymbol{a} + \delta\boldsymbol{a}) \tag{25}$$

und

$$\overline{\boldsymbol{B}} = \boldsymbol{G}^*(\boldsymbol{B} + \delta\boldsymbol{B}) = \boldsymbol{G}^*\boldsymbol{B} + \delta\overline{\boldsymbol{B}}. \tag{26}$$

Die Störungen $\delta\boldsymbol{a}$, $\delta\boldsymbol{B}$ und $\delta\overline{\boldsymbol{B}}$ genügen den Abschätzungen

$$\|\delta\boldsymbol{a}\|_2 \leqq \nu(2n+2)\,\|\boldsymbol{a}\|_2, \tag{27}$$

$$\|\delta\boldsymbol{b}^j\|_2 \leqq \nu[6(n-1)]\,\|\boldsymbol{b}^j\|_2 \tag{28}$$

sowie

$$\|\delta\bar{\boldsymbol{b}}^j\|_2 \leqq \nu[6(n-1)]\,\|\bar{\boldsymbol{b}}^j\|_2, \tag{29}$$

wobei $\boldsymbol{b}^j$ die j-te Spalte von $\boldsymbol{B}$ bezeichnet usw.

Aus (28), (29) folgen die analogen Abschätzungen

$$\|\delta \boldsymbol{B}\|_F \lessapprox \nu[6(n-1)]\,\|\boldsymbol{B}\|_F \quad \text{bzw.} \quad \|\delta \bar{\boldsymbol{B}}\|_F \lessapprox \nu[6(n-1)]\,\|\bar{\boldsymbol{B}}\|_F \tag{30}$$

in der Frobeniusnorm.

Die Analyse 3.3.15 besagt: *Die Realisierung eines Eliminationsschrittes mittels einer Folge von Givens-Drehungen gemäß* 3.3.14 *ist ein numerisch gutartiger Algorithmus mit* $F = 2n + 2$ *bezüglich* $\boldsymbol{a}$ *und mit* $F = 6(n-1)$ *bezüglich der Spalten von* $\boldsymbol{B}$ *in der Euklidischen Norm.*

3.3.16. *Bemerkungen.* (i) Ein Vergleich der Aufwandskenngrößen von 3.3.6 und 3.3.14 zeigt, daß die Transformation $(\boldsymbol{a}, \boldsymbol{B}) \to (\bar{\boldsymbol{a}}, \bar{\boldsymbol{B}})$ mit $\bar{\boldsymbol{a}} = \varrho \boldsymbol{e}^1$ mittels $n-1$ Givens-Drehungen etwa doppelt soviel Multiplikationen und n-mal soviel Wurzelberechnungen erfordert wie die mittels einer einzelnen Householder-Spiegelung. Im Normalfall, d. h. für voll besetztes $\boldsymbol{a}$ und $\boldsymbol{B}$, ist daher die Elimination mittels einer Spiegelung vorzuziehen.

(ii) Wenn der steuernde Vektor $\boldsymbol{a}$ und die Matrix $\boldsymbol{B}$ schwach besetzt und ohne spezielle Struktur sind, läßt sich die Elimination mittels Givens-Drehungen wegen deren größerer Flexibilität trotz des höheren Operationszählers meist effektiver realisieren. Man beachte dabei, daß bei schwach besetzten Problemen der Organisationsaufwand für die Berücksichtigung der Besetztheitsstruktur einen erheblichen Anteil des Gesamtaufwandes ausmacht.

(iii) Wie im folgenden gezeigt wird, läßt sich der Aufwand für die Ausführung von Givens-Drehungen durch eine geschickte implizite Realisierung etwa halbieren, so daß sie von der Anzahl der Operationen her mit den Spiegelungen konkurrieren können. Bei Anwendung zur orthogonalen Dreiecksfaktorisierung sind die Schranken für die Rundungsfehler sogar deutlich günstiger als bei Verwendung von Spiegelungen, vgl. Abschnitt 10.2.

C. *Implizite Givens-Drehungen.*

Für $\boldsymbol{a} = (a_i) \in \mathbf{R}^n$, $\boldsymbol{B} = (b_{ij}) \in \mathbf{R}^{n,r}$ betrachten wir wieder die Transformation

$$\boldsymbol{a} \to \bar{\boldsymbol{a}} = \boldsymbol{G}_{pq}\boldsymbol{a}, \qquad \boldsymbol{B} \to \bar{\boldsymbol{B}} = \boldsymbol{G}_{pq}\boldsymbol{B}. \tag{31}$$

Dabei soll die Givens-Drehung $\boldsymbol{G}_{pq}$ so festgelegt sein, daß

$$\bar{a}_q = -s * a_p + c * a_q = 0 \tag{32}$$

gilt, vgl. 3.3.10. Die oben bereits angekündigte implizite Realisierung von (31), (32) beruht auf einer skalierten Darstellung der zu transformierenden Größen gemäß

$$\begin{aligned} \boldsymbol{a} &= \operatorname{diag}\left(\sqrt{\varkappa_i}\right) \boldsymbol{d}, \qquad \boldsymbol{B} = \operatorname{diag}\left(\sqrt{\varkappa_i}\right) \boldsymbol{E}, \\ \bar{\boldsymbol{a}} &= \operatorname{diag}\left(\sqrt{\bar{\varkappa}_i}\right) \bar{\boldsymbol{d}}, \qquad \bar{\boldsymbol{B}} = \operatorname{diag}\left(\sqrt{\bar{\varkappa}_i}\right) \bar{\boldsymbol{E}} \end{aligned} \tag{33}$$

mit Skalierungsfaktoren $\varkappa_i, \bar{\varkappa}_i > 0$ $(i = 1, \ldots, n)$, die wir in den Vektoren

$$\boldsymbol{\varkappa} = (\varkappa_i), \qquad \bar{\boldsymbol{\varkappa}} = (\bar{\varkappa}_i) \in \mathbf{R}^n$$

zusammenfassen. Wird $\boldsymbol{G}_{pq}$ ebenfalls in der skalierten Form

$$\boldsymbol{G}_{pq} = \operatorname{diag}\left(\sqrt{\bar{\varkappa}_i}\right) \boldsymbol{F}_{pq} \operatorname{diag}\left(1/\sqrt{\varkappa_i}\right) \tag{34}$$

mit der Matrix

$$\boldsymbol{F}_{pq} = \begin{array}{c} \begin{matrix} & & p & & q & \end{matrix} \\ \begin{pmatrix} 1 & \vdots & & \vdots & & \\ \cdots & \alpha & \cdots & \beta & \cdots & \\ & \vdots & & \vdots & & \\ \cdot & -\gamma & \cdots & \delta & \cdots & \\ & \vdots & & \vdots & 1 & \end{pmatrix} \end{array} \begin{matrix} \\ \\ p\text{-te Zeile} \\ \\ q\text{-te Zeile} \\ \\ \end{matrix}$$

geschrieben, so läßt sich die Transformation (31) äquivalent als

$$\boldsymbol{d} \to \bar{\boldsymbol{d}} = \boldsymbol{F}_{pq}\boldsymbol{d}, \qquad \boldsymbol{E} \to \bar{\boldsymbol{E}} = \boldsymbol{F}_{pq}\boldsymbol{E} \tag{36}$$

schreiben, und (32) geht in

$$\bar{d}_q = -\gamma * d_p + \delta * d_q = 0 \tag{36}$$

über. Mit den skalierten Größen lautet die Aufgabe wie folgt: *Für gegebene Eingangsdaten* $\{\boldsymbol{\varkappa}, \boldsymbol{d}, \boldsymbol{E}\}$ *sind die Ausgangsdaten* $\{\bar{\boldsymbol{\varkappa}}, \bar{\boldsymbol{d}}, \bar{\boldsymbol{E}}\}$ *und die Parameter* $\alpha, \beta, \gamma, \delta$ *von* $\boldsymbol{F}_{pq}$ *so zu bestimmen, daß die durch* (34) *festgelegte Matrix* $\boldsymbol{G}_{pq}$ *eine Drehung ist und* $\bar{d}_q = 0$ *gilt.*

Die Gleichung (34) bedeutet elementweise $\sqrt{\bar{\varkappa}_i/\varkappa_i} = 1$, also

$$\bar{\varkappa}_i = \varkappa_i \qquad (i \neq p, q;\ i = 1, \ldots, n) \tag{37}$$

sowie

$$c = \sqrt{\frac{\bar{\varkappa}_p}{\varkappa_p}}\,\alpha = \sqrt{\frac{\bar{\varkappa}_q}{\varkappa_q}}\,\delta, \quad s = \sqrt{\frac{\bar{\varkappa}_p}{\varkappa_q}}\,\beta = \sqrt{\frac{\bar{\varkappa}_q}{\varkappa_p}}\,\gamma. \tag{38}$$

Dabei muß

$$c^2 + s^2 = 1 \tag{39}$$

gelten, damit $\boldsymbol{G}_{pq}$ eine Drehung ist. Die noch frei wählbaren sechs Größen $\alpha, \beta, \gamma, \delta$ und $\bar{\varkappa}_p, \bar{\varkappa}_q$ brauchen also nur vier Bedingungsgleichungen (36), (38), (39) zu erfüllen. Wir versuchen daher, zwei der Parameter $\alpha, \beta, \gamma, \delta$ so vorzugeben, daß die Berechnung von $\bar{\boldsymbol{E}} = \boldsymbol{F}_{pq}\boldsymbol{E}$ möglichst einfach wird. Da c und s i. allg. von 0 verschieden sind, scheiden Nullen als mögliche Werte wegen (38) aus. Es bietet sich daher an, zwei der Parameter als Einsen zu wählen, und zwar $\alpha = \delta = 1$ im Fall $c \neq 0$ bzw. $\beta = \gamma = 1$ im Fall $s \neq 0$. Die nichttrivialen Elemente von $\boldsymbol{F}_{pq}$ werden dann zu

$$\begin{pmatrix} 1 & \beta \\ -\gamma & 1 \end{pmatrix} \quad \text{bzw.} \quad \begin{pmatrix} \alpha & 1 \\ -1 & \delta \end{pmatrix},$$

so daß die Berechnung von $\boldsymbol{F}_{pq}\boldsymbol{E}$ nur noch $2r$ statt der sonst nötigen $4r$ Multiplikationen erfordert.

Wir kommen jetzt zur Festlegung der Parameter und setzen zunächst $d_p^2 + d_q^2 > 0$ voraus; wegen 3.3.10 und (33) gilt dann

$$c^2 = \frac{a_p^2}{a_p^2 + a_q^2} = \frac{\varkappa_p d_p^2}{\varkappa_p d_p^2 + \varkappa_q d_q^2}, \quad s^2 = \frac{a_q^2}{a_p^2 + a_q^2} = \frac{\varkappa_q d_q^2}{\varkappa_p d_p^2 + \varkappa_q d_q^2}. \tag{40}$$

Fall I: $c \neq 0$, d. h. $d_p \neq 0$.

Wir setzen hier

$$\alpha = \delta = 1, \tag{41}$$

und aus (36) folgt sofort

$$\gamma = d_q/d_p. \tag{42}$$

Wegen (41) liefert die erste der Gleichungen (38) die Bedingung $\bar{\varkappa}_p/\varkappa_p = \bar{\varkappa}_q/\varkappa_q$. Mit dieser und (42) folgt aus der zweiten Gleichung (38)

$$\beta = (\varkappa_q/\varkappa_p)\,\gamma = (\bar{\varkappa}_q/\bar{\varkappa}_p)\,\gamma. \tag{43}$$

Aus der ersten Gleichung (38) liest man weiter $\bar{\varkappa}_i = \varkappa_i c^2$, $i = p, q$, wegen

$$\frac{1}{c^2} = \frac{c^2 + s^2}{c^2} = 1 + \frac{s^2}{c^2} = 1 + \frac{\varkappa_q d_q^2}{\varkappa_p d_p^2} = 1 + \beta\gamma$$

also

$$\bar{\varkappa}_p = \varkappa_p/\tau, \qquad \bar{\varkappa}_q = \varkappa_q/\tau \quad \text{mit} \quad \tau = 1 + \beta\gamma \tag{44}$$

ab. Das neue Element $\bar{d}_p$ ergibt sich schließlich zu

$$\bar{d}_p = d_p + \beta d_q = d_p + \beta\gamma d_p = d_p\tau. \tag{45}$$

Fall II: $s \neq 0$, d. h. $d_q \neq 0$.

Dann kann

$$\beta = \gamma = 1 \tag{46}$$

gesetzt werden, und analog zum Fall I ergeben sich

$$\delta = d_p/d_q, \qquad \alpha = (\varkappa_p/\varkappa_q)\,\delta = (\bar{\varkappa}_q/\bar{\varkappa}_p)\,\delta, \tag{47}$$

$$\bar{\varkappa}_p = \varkappa_q/\tau, \qquad \bar{\varkappa}_q = \varkappa_p/\tau \quad \text{mit} \quad \tau = 1 + \alpha\delta \tag{48}$$

sowie

$$\bar{d}_p = d_q\tau. \tag{49}$$

3.3.17. *Aussage.* Gegeben seien $\boldsymbol{d} = (d_i) \in \mathbf{R}^n$ mit $d_p^2 + d_q^2 > 0$, $p \neq q$, und die Skalierungsfaktoren $\varkappa_i > 0$ $(i = 1, \ldots, n)$. Dann können die Parameter $\alpha, \beta, \gamma, \delta$ von $\boldsymbol{F}_{pq}$ und die Skalierungsfaktoren $\bar{\varkappa}_i > 0$ $(i = 1, \ldots, n)$ nach den Formeln (37), (41) bis (49) so festgelegt werden, daß

entweder $\alpha = \delta = 1$ oder $\beta = \gamma = 1$

gilt und die durch (34) definierte Matrix $\boldsymbol{G}_{pq}$ eine Givens-Drehung ist, die

$$\boldsymbol{a} = \operatorname{diag}\left(\sqrt{\varkappa_i}\right)\boldsymbol{d} \quad \text{in} \quad \bar{\boldsymbol{a}} = \operatorname{diag}\left(\sqrt{\bar{\varkappa}_i}\right)\bar{\boldsymbol{d}} = \boldsymbol{G}_{pq}\boldsymbol{a}$$

transformiert, wobei $\bar{\boldsymbol{d}} = \boldsymbol{F}_{pq}\boldsymbol{d}$ mit

$$\bar{d}_p = \sigma, \qquad \bar{d}_q = 0 \quad \text{und} \quad \bar{d}_i = d_i \qquad (i \neq p, q)$$

gilt. Für die analoge Transformation von

$$\boldsymbol{B} = \operatorname{diag}\left(\sqrt{\varkappa_i}\right)\boldsymbol{E} \quad \text{in} \quad \bar{\boldsymbol{B}} = \operatorname{diag}\left(\sqrt{\bar{\varkappa}_i}\right)\bar{\boldsymbol{E}} = \boldsymbol{G}_{pq}\boldsymbol{E}$$

mit $\boldsymbol{E} = (e_{ij}) \in \mathbf{R}^{n,r}$ gilt $\bar{\boldsymbol{E}} = \boldsymbol{F}_{pq}\boldsymbol{E}$, also

$$\bar{e}_{ij} = e_{ij} \qquad (j = 1, \ldots, r;\ i \neq p, q;\ i = 1, \ldots, n)$$

und

$$\bar{e}_{pj} = e_{pj} + \beta e_{qj},\ \bar{e}_{qj} = -\gamma e_{pj} + e_{qj}\ (j = 1, \ldots, r) \quad \text{im Fall } \alpha = \delta = 1 \tag{50}$$

bzw.

$$\bar{e}_{pj} = \alpha e_{pj} + e_{qj},\ \bar{e}_{qj} = -e_{pj} + \delta e_{qj}\ (j = 1, \ldots, r) \quad \text{im Fall } \beta = \gamma = 1. \tag{51}$$

Wir nennen die Transformation $\bar{\boldsymbol{d}} = \boldsymbol{F}_{pq}\boldsymbol{d}$ einschließlich der zugehörigen Transformation von $\varkappa$ in $\bar{\varkappa}$ implizite Realisierung der durch $\boldsymbol{G}_{pq}$ gemäß 3.3.10 festgelegten Givens-Transformation und sprechen kurz von *impliziten Givens-Drehungen*. Gelegentlich werden diese Transformationen auch modifizierte Givens-Drehungen bzw. — da bei der Realisierung im Gegensatz zu 3.3.10 keine Quadratwurzeln zu ziehen sind — quadratwurzelfreie Givens-Drehungen genannt; siehe B 3.4 für einen historischen Kommentar.

Um die Parameter von $\boldsymbol{F}_{pq}$ und die Skalierungsfaktoren in vernünftigen Größenordnungen zu halten, gehen wir wie in 3.3.11 vor und verwenden die Formeln aus Fall I für $|a_p| \geqq |a_q|$, d. h. $\varkappa_p d_p^2 > \varkappa_q d_q^2$; die aus Fall II werden für $|a_p| < |a_q|$ verwendet. Den Ausnahmefall $d_p = d_q = 0$ ordnen wir wie dort dem Fall I zu und setzen einfach $\boldsymbol{F}_{pq} = \boldsymbol{G}_{pq} = \boldsymbol{I}$, $\bar{\varkappa} = \varkappa$. In Analogie zu 3.3.11 berechnen wir außerdem eine Größe ξ, aus der sowohl das Vorliegen von Fall I oder II wie auch die Werte der Parameter einfach rekonstruiert werden können.

3.3.18. *Algorithmus.* Berechnung der Parameter $\alpha, \beta, \gamma, \delta$ von $\boldsymbol{F}_{pq}$, der Skalierungsfaktoren $\bar{\varkappa}_p, \bar{\varkappa}_q$ und der Größen $\sigma = \bar{d}_p$ sowie ξ aus $d_p, d_q, \varkappa_p, \varkappa_q \in \mathfrak{R}$.

if $\varkappa_p d_p^2 \geqq \varkappa_q d_p^2$ **then**
- $\alpha := \delta := 1$
- **if** $d_p \neq 0$
- **then**
  - $\gamma := d_q/d_p$, $\xi := \varkappa_p/\varkappa_q$, $\beta := \gamma/\xi$
  - $\xi := \gamma/(1 + \xi)$
- **else** $[\gamma := \beta := \xi := 0]$
- $\tau := 1 + \gamma * \beta$, $\bar{\varkappa}_p := \varkappa_p/\tau$, $\bar{\varkappa}_q := \varkappa_q/\tau$, $\sigma := d_p * \tau$

else
- $\beta := \gamma := 1$
- $\delta := d_p/d_q$, $\xi := \varkappa_p/\varkappa_q$, $\alpha := \delta * \xi$
- **if** $\alpha \neq 0$ **then** $\xi := (1 + \xi)/\alpha$ **else** $\xi := 1$
- $\tau := 1 + \alpha * \delta$, $\bar{\varkappa}_p := \varkappa_q/\tau$, $\bar{\varkappa}_q := \varkappa_p/\tau$, $\sigma := d_q * \tau$

Aufwand: 12 opm + 2 ops (einschließlich der zur Fallunterscheidung nötigen Multiplikationen)

Man beachte dabei, daß die drei Fälle

$$\varkappa_p d_p^2 \geqq \varkappa_q d_q^2, \qquad 0 = \varkappa_p d_p^2 < \varkappa_q d_q^2 \quad \text{bzw.} \quad 0 < \varkappa_p d_p^2 < \varkappa_q d_q^2$$

durch

$$|\xi| \leqq 1/2, \qquad \xi = 1 \quad \text{bzw.} \quad |\xi| > 2$$

charakterisiert sind und daher zuverlässig aus ξ rekonstruiert werden können. Für den Fall $\varkappa_p d_p^2 \geqq \varkappa_q d_q^2$, also $|\xi| \leqq 1/2$, verwenden wir im folgenden auch die Bezeichnung „Fall 1"; mit „Fall 2" werden die beiden restlichen Möglichkeiten bezeichnet.

Die in 3.3.18 vorgenommene Fallunterscheidung garantiert

$$|\gamma| \leqq \sqrt{\varkappa_p/\varkappa_q},\ |\beta| \leqq \sqrt{\varkappa_q/\varkappa_p} \quad \text{im Fall 1}$$

bzw. (52)

$$|\alpha| \leqq \sqrt{\varkappa_p/\varkappa_q},\ |\delta| \leqq \sqrt{\varkappa_q/\varkappa_p} \quad \text{im Fall 2},$$

also

$$1 \leqq \tau = \alpha\delta + \beta\gamma \leqq 2. \tag{53}$$

Die Rekonstruktion der Kenngrößen von $\boldsymbol{F}_{pq}$ ist analog zu 3.3.12 nach den folgenden Algorithmen möglich, wobei einmal auf die Anfangswerte $\varkappa_i$, zum anderen auf die Endwerte $\bar{\varkappa}_i$ der Skalierungsparameter zurückgegriffen werden kann, vgl. (43), (44) bzw. (47), (48).

3.3.19. *Algorithmus.* Rekonstruktion von $\alpha, \beta, \gamma, \delta$ und $\bar{\varkappa}_p, \bar{\varkappa}_q$ aus $\xi, \varkappa_p, \varkappa_q \in \mathfrak{R}$.

if $|\xi| \leqq 1$ **then** | **if** $\xi \neq 1$
then Fall 1: | $\alpha := \delta := 1$
$\lambda := \varkappa_p/\varkappa_q,\ \gamma := \xi * (1 + \lambda),\ \beta := \gamma/\lambda$
$\tau := 1 + \gamma * \beta,\ \bar{\varkappa}_p := \varkappa_p/\tau,\ \bar{\varkappa}_q := \varkappa_q/\tau$
else Fall 2: | $\beta := \gamma := 1,\ \alpha := \delta := 0$
$\bar{\varkappa}_p := \varkappa_q,\ \bar{\varkappa}_q := \varkappa_p$
else Fall 2: | $\beta := \gamma := 1$
$\lambda := \varkappa_p/\varkappa_q,\ \alpha := (1 + \lambda)/\xi,\ \delta := \alpha/\lambda$
$\tau := 1 + \alpha * \delta,\ \bar{\varkappa}_p := \varkappa_q/\tau,\ \bar{\varkappa}_q := \varkappa_p/\tau$

Aufwand: 6 opm + 2 ops

3.3.20. *Algorithmus.* Rekonstruktion von $\alpha, \beta, \gamma, \delta$ und $\varkappa_p, \varkappa_q$ aus $\xi, \bar{\varkappa}_p, \bar{\varkappa}_q \in \mathfrak{R}$.

if $|\xi| \leqq 1$ **then** | **if** $\xi \neq 1$
then Fall 1: | $\alpha := \delta := 1$
$\lambda := \bar{\varkappa}_p/\bar{\varkappa}_q,\ \gamma := \xi * (1 + \lambda),\ \beta := \gamma/\lambda$
$\tau := 1 + \gamma * \beta,\ \varkappa_p := \bar{\varkappa}_p * \tau,\ \varkappa_q := \bar{\varkappa}_q * \tau$
else Fall 2: | $\beta := \gamma := 1,\ \alpha := \delta := 0$
$\varkappa_p := \bar{\varkappa}_q,\ \varkappa_q := \bar{\varkappa}_p$
else Fall 2: | $\beta := \gamma := 1$
$\lambda := \bar{\varkappa}_q/\bar{\varkappa}_p,\ \alpha := (1 + \lambda)/\xi,\ \delta := \alpha/\lambda$
$\tau := 1 + \alpha * \delta,\ \varkappa_p := \bar{\varkappa}_q * \tau,\ \varkappa_q := \bar{\varkappa}_p * \tau$

Aufwand: 6 opm + 2 ops

Die impliziten Givens-Drehungen zeigen ihre Effektivität erst dann, wenn sie in einer längeren Folge auftreten. Wir demonstrieren dies für die in 3.3.14 formulierte

Aufgabe und realisieren die dort ausgeführten Transformationen

$$\begin{aligned} \boldsymbol{a}^1 &= \boldsymbol{a},\ \boldsymbol{a}^{i-1} \to \boldsymbol{a}^i = \boldsymbol{G}_{1i}\boldsymbol{a}^{i-1}\ (i = 2, \ldots, n), & \bar{\boldsymbol{a}} &= \boldsymbol{a}^n, \\ \boldsymbol{B}^{(1)} &= \boldsymbol{B},\ \boldsymbol{B}^{(i-1)} \to \boldsymbol{B}^{(i)} = \boldsymbol{G}_{1i}\boldsymbol{B}^{(i-1)}\ (i = 2, \ldots, n), & \bar{\boldsymbol{B}} &= \boldsymbol{B}^{(n)} \end{aligned} \tag{54}$$

mittels der Matrizen $\boldsymbol{F}_{1i}$ und der Skalierungsfaktoren $\boldsymbol{\varkappa}^{(i)} = (\varkappa_l^{(i)}) \in \mathbf{R}^n$ in der impliziten Form

$$\begin{aligned} \boldsymbol{d}^1 &= \boldsymbol{d},\ \boldsymbol{d}^{i-1} \to \boldsymbol{d}^i = \boldsymbol{F}_{1i}\boldsymbol{d}^{i-1}\ (i = 2, \ldots, n), & \bar{\boldsymbol{d}} &= \boldsymbol{d}^n, \\ \boldsymbol{E}^{(1)} &= \boldsymbol{E},\ \boldsymbol{E}^{(i-1)} \to \boldsymbol{E}^{(i)} = \boldsymbol{F}_{1i}\boldsymbol{E}^{(i-1)}\ (i = 2, \ldots, n), & \bar{\boldsymbol{E}} &= \boldsymbol{E}^{(n)}. \end{aligned} \tag{55}$$

Dabei hängen die skalierten Größen aus (55) mit den unskalierten Größen aus (54) über

$$\boldsymbol{a}^i = \operatorname{diag}\left(\sqrt{\varkappa_l^{(i)}}\right)\boldsymbol{d}^i, \qquad \boldsymbol{B}^{(i)} = \operatorname{diag}\left(\sqrt{\varkappa_l^{(i)}}\right)\boldsymbol{E}^{(i)} \tag{56}$$

und

$$\boldsymbol{G}_{1i} = \operatorname{diag}\left(\sqrt{\varkappa_l^{(i)}}\right)\boldsymbol{F}_{1i}\operatorname{diag}\left(1/\sqrt{\varkappa_l^{(i-1)}}\right) \tag{57}$$

zusammen, wobei $\boldsymbol{\varkappa}^{(1)} = \boldsymbol{\varkappa}$ und $\boldsymbol{\varkappa}^{(n)} = \bar{\boldsymbol{\varkappa}}$ ist.

3.3.21. *Eliminationsschritt mittels einer Folge impliziter Givens-Drehungen*

Aufgabe: Gegeben seien $\boldsymbol{a} \in \mathbf{R}^n$, $\boldsymbol{B} \in \mathbf{R}^{n,r}$ in der skalierten Form

$$\boldsymbol{a} = \operatorname{diag}\left(\sqrt{\varkappa_l}\right)\boldsymbol{d}, \qquad \boldsymbol{B} = \operatorname{diag}\left(\sqrt{\varkappa_l}\right)\boldsymbol{E} \tag{58}$$

durch $\bar{\boldsymbol{d}} = (d_i) \in \mathfrak{R}^n$, $\boldsymbol{E} = (e_{ij}) \in \mathfrak{R}^{n,r}$ und $\boldsymbol{\varkappa} = (\varkappa_l) \in \mathfrak{R}^n$, $\varkappa_l > 0$. Die Transformation (54) ist mit Givens-Drehungen $\boldsymbol{G}_{1i}$ in der impliziten Version (55) so durchzuführen, daß

$$\bar{\boldsymbol{a}} = \boldsymbol{G}_{1n} \cdots \boldsymbol{G}_{12}\boldsymbol{a} = \varrho\boldsymbol{e}^1 = (\varrho, 0, \ldots, 0)^\mathsf{T}$$

gilt. Die Resultate sind mit $\bar{\boldsymbol{d}} = \sigma\boldsymbol{e}^1 = (\sigma, 0, \ldots, 0)^\mathsf{T}$ gemäß

$$\bar{\boldsymbol{a}} = \operatorname{diag}\left(\sqrt{\bar{\varkappa}_l}\right)\bar{\boldsymbol{d}}, \qquad \bar{\boldsymbol{B}} = \operatorname{diag}\left(\sqrt{\bar{\varkappa}_l}\right)\bar{\boldsymbol{E}} \tag{59}$$

skaliert durch $\bar{\boldsymbol{d}} \in \mathfrak{R}^n$, $\bar{\boldsymbol{E}} \in \mathfrak{R}^{n,r}$ und $\bar{\boldsymbol{\varkappa}} \in \mathfrak{R}^n$ darzustellen. Die Berechnung soll in situ durch Überspeichern von $\boldsymbol{d}$ mit $(\sigma, \xi_2, \ldots, \xi_n)^\mathsf{T}$, von $\boldsymbol{\varkappa}$ durch $\bar{\boldsymbol{\varkappa}}$ und von $\boldsymbol{E}$ durch $\bar{\boldsymbol{E}}$ erfolgen, wobei ξ_i die zu $\boldsymbol{F}_{1i}$ gemäß 3.3.18 bestimmte Kenngröße bezeichnet.

Algorithmus:

for $i := 2(1)n$ **do**

S0 (Festlegung der impliziten Givens-Drehung $\boldsymbol{F}_{1i}$): Berechne $\{\alpha, \beta, \gamma, \delta, \bar{\varkappa}_1, \bar{\varkappa}_i, \sigma, \xi\}$ aus $\{d_1, d_i, \varkappa_1, \varkappa_i\}$ gemäß Algorithmus 3.3.18 mit $p = 1$, $q = i$.

S1 ($\boldsymbol{d}^{i-1} \to \boldsymbol{d}^i = \boldsymbol{F}_{1i}\boldsymbol{d}^{i-1}$, $\boldsymbol{\varkappa}^{i-1} \to \boldsymbol{\varkappa}^i$, Reservierung von $\xi = \xi_i$):

$d_1 := \sigma$, $d_i := \xi$, $\varkappa_1 := \bar{\varkappa}_1$, $\varkappa_i := \bar{\varkappa}_i$

S2 ($\boldsymbol{E}^{(i-1)} \to \boldsymbol{E}^{(i)} = \boldsymbol{F}_{1i}\boldsymbol{E}^{(i-1)}$):
  for $j := 1(1)r$ **do**
    if $|\xi| < 1$
    then Fall 1: $\mu := e_{1j} + \beta * e_{ij}$; $e_{ij} := -\gamma * e_{1j} + e_{ij}$; $e_{1j} := \mu$
    else Fall 2: $\mu := \alpha * e_{1j} + e_{ij}$; $e_{ij} := -e_{1j} + \delta * e_{ij}$; $e_{1j} := \mu$

Aufwand: Festlegung von $\boldsymbol{F}_{1i}$ und Berechnung von $\bar{\boldsymbol{d}}, \bar{\boldsymbol{\varkappa}}$: $(n-1)$ (12 opm + 2 ops)
Berechnung von $\bar{\boldsymbol{E}}$: $(n-1)\, r$(2 opm + 2 ops)

Durch 3.3.21 wird

$$\left(\begin{array}{c|c} d_1 & e_{11} \dots e_{1r} \\ d_2 & e_{21} \dots e_{2r} \\ \vdots & \vdots \quad\quad \vdots \\ d_n & e_{n1} \dots e_{nr} \end{array}\right) \text{ mit } \left(\begin{array}{c|c} \sigma & \bar{e}_{11} \dots \bar{e}_{1r} \\ \xi_2 & \bar{e}_{21} \dots \bar{e}_{2r} \\ \vdots & \vdots \quad\quad \vdots \\ \xi_n & \bar{e}_{n1} \dots \bar{e}_{nr} \end{array}\right) \text{ und } \begin{pmatrix} \varkappa_1 \\ \varkappa_2 \\ \vdots \\ \varkappa_n \end{pmatrix} \text{ mit } \begin{pmatrix} \bar{\varkappa}_1 \\ \bar{\varkappa}_2 \\ \vdots \\ \bar{\varkappa}_n \end{pmatrix}$$

überspeichert.

Die Ergebnisse der Rundungsfehleranalyse von 3.3.21 sind weitgehend analog zu denen von 3.3.14.

3.3.22. *Rundungsfehleranalyse.* Algorithmus 3.3.21 werde mit den Eingangsdaten $\boldsymbol{\varkappa}, \boldsymbol{d} \in \mathfrak{R}^n$, $\boldsymbol{E} \in \mathfrak{R}^{n,r}$ durchgeführt, und $\bar{\boldsymbol{\varkappa}}, \bar{\boldsymbol{d}} = \sigma \boldsymbol{e}^1 = (\sigma, 0, \dots, 0)^T$ sowie $\bar{\boldsymbol{E}}$ seien die berechneten Resultate. Mit $\boldsymbol{G}_{1i}^*$ werde die exakte Givens-Drehung bezeichnet, die $(\boldsymbol{a}^{i-1})^*$ exakt in $\boldsymbol{G}_{1i}^*(\boldsymbol{a}^{i-1})^*$ mit $[\boldsymbol{G}_{1i}^*(\boldsymbol{a}^{i-1})^*]_i = 0$ transformiert, wobei $(\boldsymbol{a}^{i-1})^*$ das exakte Produkt diag $\left(\sqrt{\varkappa_l^{i-1}}\right) \boldsymbol{d}^{i-1}$ ist, und $\boldsymbol{G}^* = \boldsymbol{G}_{1n}^* \dots \boldsymbol{G}_{12}^*$ sei das exakte Produkt der $\boldsymbol{G}_{1i}^*$. Dann gelten für die gemäß (58) bzw. (59) definierten unskalierten Größen $\boldsymbol{a}, \boldsymbol{B}$ bzw. $\bar{\boldsymbol{a}}, \bar{\boldsymbol{B}}$ die Aussagen von 3.3.15.

Wir bemerken dazu nur, daß sich die geringfügigen Unterschiede zwischen 3.3.14 und 3.3.21 auf die Schlußabschätzungen (27) bis (29) nicht auswirken. Die Elimination mittels impliziter Givens-Drehungen ist also — in bezug auf die unskalierten Originaldaten — ein numerisch gutartiger Prozeß.

3.3.23. *Bemerkungen.* (i) Der entscheidende Vorteil von 3.3.21 gegenüber 3.3.14 liegt in der Halbierung der Anzahl von Multiplikationen, die für die Transformation von $\boldsymbol{B}$ bzw. $\boldsymbol{E}$ in $\bar{\boldsymbol{B}}$ bzw. $\bar{\boldsymbol{E}}$ erforderlich sind. Die Vermeidung von Quadratwurzelberechnungen ist eine weitere angenehme, aber hier weniger wesentliche Eigenschaft der impliziten Givens-Drehungen.

(ii) In den Anwendungen — siehe Abschnitt 10.2 — tritt 3.3.21 in der Regel als Teilschritt eines n-stufigen Verfahrens auf, so daß sich die Anfangswerte $\boldsymbol{\varkappa} = (\varkappa_l)$ der Skalierungsfaktoren als Endwerte der vorherigen Stufe ergeben. In der ersten Stufe wird $\varkappa_l = 1$ $(l = 1, \dots, n)$ gesetzt, oder die Faktoren ergeben sich in natürlicher Weise aus der praktischen Aufgabenstellung, etwa als Gewichte bei einem linearen Regressionsproblem. Für gewisse Aufgabentypen wie die Lösung linearer

Quadratmittelprobleme braucht $\bar{\boldsymbol{B}} = \operatorname{diag}\left(\sqrt{\bar{\varkappa}_l}\right)\bar{\boldsymbol{E}}$ nicht explizit bekannt zu sein; die Lösungen lassen sich allein aus $\bar{\boldsymbol{E}}$ berechnen. Bei Bedarf kann $\bar{\boldsymbol{B}}$ mit dem Aufwand von höchstens n opr + nr opm selbstverständlich gebildet werden.

(iii) Falls 3.3.21 mit $\varkappa_l = 1$ $(l = 1, \ldots, n)$ gestartet wird, folgt wegen (53)

$$\bar{\varkappa}_l \geqq 1/2^{n-1} \qquad (l = 1, \ldots, n). \tag{60}$$

Für großes n kann daher in ungünstigen Fällen bei der Berechnung der Skalierungsfaktoren Unterlauf eintreten.

(iv) Die erwähnte schnelle Verkleinerung der Skalierungsfaktoren kann durch eine zu 3.2.7(ii) analoge Pivotisierung vermieden werden. Dazu wird der Zeilenindex s, $1 \leqq s \leqq n$, mit

$$\varkappa_s d_s^2 = \max_{i=1,\ldots,n} \varkappa_i d_i^2 \tag{61}$$

bestimmt, und statt der Eingangsdaten $\{\boldsymbol{\varkappa}, \boldsymbol{d}, \boldsymbol{E}\}$ werden in 3.3.21 die gemäß

$$\hat{\boldsymbol{\varkappa}} = \boldsymbol{T}_{1s}\boldsymbol{\varkappa}, \qquad \hat{\boldsymbol{d}} = \boldsymbol{T}_{1s}\boldsymbol{d}, \qquad \hat{\boldsymbol{E}} = \boldsymbol{T}_{1s}\boldsymbol{E} \tag{62}$$

permutierten Eingangsdaten verwendet, d. h., die Zeilen 1 und s werden vertauscht. Dies ist gleichbedeutend mit dem Ersatz von

$$\boldsymbol{F} = \boldsymbol{F}_{1n} \cdots \boldsymbol{F}_{12} \quad \text{durch} \quad \boldsymbol{F}\boldsymbol{T}_{1s} = \boldsymbol{F}_{1n} \cdots \boldsymbol{F}_{12}\boldsymbol{T}_{1s}. \tag{63}$$

Im Originalproblem bedeutet (61)

$$a_s^2 = \max_{i=1,\ldots,n} a_i^2, \tag{64}$$

d. h., die maximale Komponente von $\boldsymbol{a}$ wird in die erste Zeile gebracht. Wegen

$$(a_1^{i-1})^2 = (\hat{a}_1)^2 + (\hat{a}_2)^2 + \cdots + (\hat{a}_{i-1})^2 \qquad (i = 2, \ldots, n) \tag{65}$$

gilt dann

$$\varkappa_1^{i-1}(d_1^{i-1})^2 = (a_1^{i-1})^2 \geqq (\hat{a}_i)^2 = \varkappa_i^{i-1}(d_i^{i-1})^2 \qquad (i = 2, \ldots, n),$$

d. h., für jedes $i = 2, \ldots, n$ liegt stets Fall 1 vor. Die Algorithmen 3.3.18 bis 3.3.21 reduzieren sich daher auf die für Fall 1 vorgesehenen Teile. Aus (65) folgt weiter die Darstellung

$$\tau = \tau_i = (a_1^i)^2/(a_1^{i-1})^2 = (\hat{a}_1^2 + \cdots + \hat{a}_i^2)/(\hat{a}_1^2 + \cdots + \hat{a}_{i-1}^2)$$

für das im i-ten Schritt berechnete τ, aus der sich wegen $\bar{\varkappa}_1 = \varkappa_1/(\tau_2 \cdots \tau_n)$, $\bar{\varkappa}_i = \varkappa_i/\tau_i$ — es liegt stets Fall 1 vor! — sofort

$$\bar{\varkappa}_1 \geqq \varkappa_1/n, \qquad \bar{\varkappa}_i \geqq \varkappa_i/2 \qquad (i = 2, \ldots, n) \tag{66}$$

ergibt. Auch für sehr großes n kann daher praktisch kein Unterlauf eintreten. Implizite Givens-Drehungen sollten daher möglichst in der beschriebenen pivotisierten Form angewendet werden. □

Wir betrachten abschließend das Problem, nach Ausführung von 3.3.21 für eine weitere Matrix $\boldsymbol{V} \in \mathbf{R}^{n,m}$ die Transformation

$$\bar{\boldsymbol{V}} = \boldsymbol{G}\boldsymbol{V} = \boldsymbol{G}_{1n} \cdots \boldsymbol{G}_{12}\boldsymbol{V} \tag{67}$$

durchzuführen, wobei $\boldsymbol{V}$ und $\overline{\boldsymbol{V}}$ in der skalierten Form

$$\boldsymbol{V} = \operatorname{diag}\left(\sqrt{\varkappa_l}\right)\boldsymbol{W}, \qquad \overline{\boldsymbol{V}} = \operatorname{diag}\left(\sqrt{\overline{\varkappa}_l}\right)\overline{\boldsymbol{W}} \tag{68}$$

durch $\varkappa$, $\boldsymbol{W}$ bzw. $\overline{\varkappa}$, $\overline{\boldsymbol{W}}$ gegeben sein sollen. Es gilt dann

$$\overline{\boldsymbol{W}} = \boldsymbol{F}\boldsymbol{W} = \boldsymbol{F}_{1n} \cdots \boldsymbol{F}_{12}\boldsymbol{W}, \tag{69}$$

und die Berechnung kann schrittweise gemäß

$$\boldsymbol{W}^{(1)} = \boldsymbol{W},\ \boldsymbol{W}^{(i-1)} \to \boldsymbol{W}^{(i)} = \boldsymbol{F}_{1i}\boldsymbol{W}^{(i-1)}\ (i = 2, \ldots, n),\ \overline{\boldsymbol{W}} = \boldsymbol{W}^{(n)} \tag{70}$$

erfolgen. Da bei der Rekonstruktion der Parameter von $\boldsymbol{F}_{1i}$ gemäß 3.3.19 die bezüglich des i-ten Schrittes „alten" Skalierungsfaktoren $\varkappa^{i-1}$ benötigt werden, werden wie in 3.3.21 neben (70) die Transformationen

$$\varkappa^1 = \varkappa,\ \varkappa^{i-1} \to \varkappa^i\ (i = 2, \ldots, n),\ \overline{\varkappa} = \varkappa^n \tag{71}$$

ausgeführt. Man beachte dabei, daß bei Anwendung von 3.3.19 auf $\boldsymbol{F}_{1i}$ die dort vorkommenden „alten" Skalierungsfaktoren $\varkappa$ die Bedeutung von $\varkappa^{i-1}$ haben; die „neuen" Faktoren $\overline{\varkappa}$ entsprechen den $\varkappa^i$.

3.3.24. *Algorithmus.* Für gegebenes $\varkappa \in \mathfrak{R}^n$, $\boldsymbol{W} = (w_{ij}) \in \mathfrak{R}^{n,m}$ ist $\varkappa \in \mathfrak{R}^n$, $\overline{\boldsymbol{W}} = (\overline{w}_{ij}) \in \mathfrak{R}^{n,m}$ unter Verwendung der Größen $\{\xi_2, \ldots, \xi_n\}$ aus Algorithmus 3.3.21 so zu berechnen, daß (69) gilt. Die Eingangsdaten $\{\varkappa, \boldsymbol{W}\}$ sind durch $\{\overline{\varkappa}, \overline{\boldsymbol{W}}\}$ zu überspeichern.

for $i := 2(1)n$ **do**
- S1: Berechne $\{\alpha, \beta, \gamma, \delta, \overline{\varkappa}_1, \overline{\varkappa}_i\}$ aus $\{\xi_i, \varkappa_1, \varkappa_i\}$ gemäß Algorithmus 3.3.19 mit $\xi = \xi_i$, $\varkappa_p = \varkappa_1$, $\varkappa_q = \varkappa_i$.
- S2 (Transformation $\varkappa^{i-1} \to \varkappa^i$):
  $\varkappa_1 := \overline{\varkappa}_1$, $\varkappa_i := \overline{\varkappa}_i$
- S3 (Transformation $\boldsymbol{W}^{(i-1)} \to \boldsymbol{W}^{(i)} = \boldsymbol{F}_{1i}\boldsymbol{W}^{(i-1)}$):
  for $j := 1(1)m$ **do**
  - **if** $|\xi| < 1$
  - **then** Fall 1: $\mu := w_{1j} + \beta * w_{ij}$; $w_{ij} := -\gamma * w_{1j} + w_{ij}$; $w_{1j} := \mu$
  - **else** Fall 2: $\mu := \alpha * w_{1j} + w_{ij}$; $w_{ij} := -w_{1j} + \delta * w_{ij}$; $w_{1j} := \mu$

Aufwand: $(n - 1)\,[(2m + 6)\,\text{opm} + (2m + 2)\,\text{ops}]$

3.3.25. *Bemerkungen.* (i) Im Gegensatz zu den klassischen Givens-Drehungen kann $\boldsymbol{F}_{1i}$ nicht direkt aus ξ_i reproduziert werden, da dazu entweder $\varkappa^{i-1}$ oder $\varkappa^i$ erforderlich ist, die intermediären Skalierungsfaktoren jedoch aus Platzgründen nicht aufgehoben werden können.

(ii) Bei praktischen Aufgaben werden die $\boldsymbol{F}_{1i}$ nur in der direkten Reihenfolge $\boldsymbol{F}_{12}, \boldsymbol{F}_{13}, \ldots, \boldsymbol{F}_{1n}$ bzw. in der inversen Reihenfolge $\boldsymbol{F}_{1n}, \boldsymbol{F}_{1,n-1}, \ldots, \boldsymbol{F}_{12}$ benötigt. Ein Beispiel für den ersten Fall ist 3.3.24, wo die $\varkappa^i$ in der analogen Reihenfolge gemäß 3.3.18 aufdatiert werden; statt S3 kann in 3.3.24 natürlich eine beliebige Operation

mit $\boldsymbol{F}_{1i}$ stehen. Falls die inverse Reihenfolge benötigt wird, ist 3.3.24 wie folgt zu modifizieren:

for $i := n(-1)2$ **do**

S1: Berechne $\{\alpha, \beta, \gamma, \delta, \varkappa_1, \varkappa_i\}$ aus $\{\xi_i, \bar{\varkappa}_1, \bar{\varkappa}_i\}$ gemäß Algorithmus 3.3.20 mit $\xi = \xi_i, \bar{\varkappa}_p = \bar{\varkappa}_1, \bar{\varkappa}_q = \bar{\varkappa}_i$

S2 (Transformation $\boldsymbol{\varkappa}^i \to \boldsymbol{\varkappa}^{i-1}$): $\bar{\varkappa}_1 := \varkappa_1, \bar{\varkappa}_i := \varkappa_i$

S3: Ausführung einer beliebigen Operation mit $\boldsymbol{F}_{1i}$

Die inverse Reihenfolge wird z. B. bei der Berechnung von $\boldsymbol{Z} = \boldsymbol{G}^{-1}\boldsymbol{V}$ durch Auswertung von

$$\boldsymbol{Z} = \operatorname{diag}\left(1/\sqrt{\varkappa_l}\right) \boldsymbol{F}_{12}^{\mathsf{T}}\boldsymbol{F}_{13}^{\mathsf{T}} \cdots \boldsymbol{F}_{1n}^{\mathsf{T}}\left[\operatorname{diag}\left(\sqrt{\bar{\varkappa}_l}\right) \operatorname{diag}\left(\sqrt{\varkappa_l}\right) \boldsymbol{W}\right] \tag{72}$$

benötigt; man beachte dabei die Gültigkeit von

$$\boldsymbol{G}^{-1} = \boldsymbol{G}^{\mathsf{T}} = \left[\operatorname{diag}\left(\sqrt{\bar{\varkappa}_l}\right) \boldsymbol{F} \operatorname{diag}\left(1/\sqrt{\varkappa_l}\right)\right]^{\mathsf{T}}. \tag{73}$$

Übungsaufgaben

Ü 3.3.1. Man zeige, daß sich im Fall $n = 2$ jede Householder-Spiegelung $\boldsymbol{H} = \boldsymbol{I} - 2\boldsymbol{u}\boldsymbol{u}^{\mathsf{T}}$, $\boldsymbol{u} = (\gamma, \sigma)^{\mathsf{T}}$, $\gamma^2 + \sigma^2 = 1$, in der Gestalt

$$\boldsymbol{H} = \begin{pmatrix} c & s \\ s & -c \end{pmatrix} \quad \text{mit} \quad c^2 + s^2 = 1$$

schreiben läßt. Wie hängen $\{c, s\}$ und $\{\gamma, \sigma\}$ zusammen? Geometrische Deutung! Die analog zu $\boldsymbol{G}_{pq}$ für $p \neq q$ und $n \geqq 2$ gebildeten Matrizen

$$\boldsymbol{H}_{pq} = \begin{pmatrix} 1 & \vdots & & \vdots & \\ \cdots & c & \cdots & s & \cdots \\ & \vdots & & \vdots & \\ \cdots & s & \cdots & -c & \cdots \\ & \vdots & & \vdots & 1 \end{pmatrix} \begin{matrix} \\ p\text{-te Zeile} \\ \\ q\text{-te Zeile} \\ \\ \end{matrix}$$

werden deshalb auch *ebene Spiegelungen* bzw. *Givens-Spiegelungen* genannt. Für sie gilt $\det(\boldsymbol{H}_{pq}) = -1$ im Gegensatz zu $\det(\boldsymbol{G}_{pq}) = 1$. Sie können sinngemäß wie Givens-Drehungen verwendet werden.

Ü 3.3.2. Gegeben seien $\boldsymbol{a}, \boldsymbol{b} \in \mathbf{R}^n$ mit $\boldsymbol{a} \neq \boldsymbol{b}$, $\|\boldsymbol{a}\|_2 = \|\boldsymbol{b}\|_2$. Man konstruiere eine Householder-Spiegelung $\boldsymbol{H}$, für die $\boldsymbol{H}\boldsymbol{a} = \boldsymbol{b}$ gilt.

Ü 3.3.3. Man zeige, daß die in Ü 3.3.2 mittels einer Householder-Spiegelung gelöste Aufgabe auch durch eine Folge $\boldsymbol{G} = \boldsymbol{G}_{p(n)q(n)} \cdots \boldsymbol{G}_{p(2)q(2)}$ von $n - 1$ Givens-Drehungen $\boldsymbol{G}_{p(i)q(i)}$ $(i = 2, \ldots, n)$ gelöst werden kann, wobei jede Drehung die Komponente $b_{q(i)}$ erzeugt.

Ü 3.3.4. Man überlege sich, daß $\boldsymbol{G}^{(i)} = \boldsymbol{G}_{1i} \cdots \boldsymbol{G}_{13}\boldsymbol{G}_{12}$ $(i = 2, \ldots, n)$ von der Gestalt

$$\boldsymbol{G}^{(i)} = \begin{pmatrix} & & & & \overset{i\text{-teSpalte}}{\downarrow} & & \\ \times & \times & \times & \cdots & \times & & \\ \times & \times & & & & & \\ \times & \times & \times & & & & \\ \vdots & \vdots & \vdots & \ddots & & & \\ \times & \times & \times & \cdots & \times & & \leftarrow i\text{-teZeile} \\ & & & & & 1 & \\ & & & & & & \ddots \\ & & & & & & 1 \end{pmatrix}$$

ist, und nütze diesen Fakt in einem Algorithmus zur Berechnung von $\boldsymbol{G} = \boldsymbol{G}^{(n)}$ aus. Die Drehungen $\boldsymbol{G}_{1i}$ seien dabei durch die Zahlen ξ_i repräsentiert. Der Aufwand ist $(n-1)$ $\times$ (2 opm + 1 ops + 1 opr) für die Rekonstruktion der c_i, s_i und $(n-1)\,n$ opm für die Produktbildung.

Ü 3.3.5. Der in 3.3.13, 3.3.14 bzw. 3.3.21 beschriebene Eliminationsschritt kann z. B. auch durch die Folge $G = \boldsymbol{G}_{12}\boldsymbol{G}_{23}\cdots\boldsymbol{G}_{n-2,n-1}\boldsymbol{G}_{n-1,n}$ von ebenen Drehungen realisiert werden. Welche Gestalt hat die explizit berechnete Matrix $\boldsymbol{G}$ in diesem Fall?

Ü 3.3.6. Man zeige, daß das neue Element b_{ij} im Schritt S 2 von 3.3.14 auch gemäß

$$b_{ij} := \begin{cases} b_{ij} - z * (\mu + b_{1j}) & \text{für} \quad c \geqq 0, \\ z * (\mu - b_{1j}) - b_{ij} & \text{für} \quad c < 0 \end{cases}$$

mit $z := s/(1 + |c|)$ berechnet werden kann. Der Aufwand zur Berechnung von $\bar{\boldsymbol{B}}$ ist dann $\sim (n-1)\,r \times 3$ opms gegenüber $(n-1)\,r \times$ (4 opm + 2 ops) in 3.3.14. Dieses Vorgehen ist daher zweckmäßig, wenn eine Multiplikation wesentlich aufwendiger als eine Addition ist.

Ü 3.3.7. Man lege $\boldsymbol{v} \in \mathbf{R}^n$ analog zu (10), (11) so fest, daß $\boldsymbol{a} = (a_i) \in \mathbf{R}^n$ in $\bar{\boldsymbol{a}} = (\bar{a}_i) = (\boldsymbol{I} - \boldsymbol{v}\boldsymbol{v}^{\mathrm{T}}/\gamma)\,\boldsymbol{a}$, $\gamma = \boldsymbol{v}^{\mathrm{T}}\boldsymbol{v}/2$, mit

$$\bar{\boldsymbol{a}}_i = \begin{cases} 0 & \text{für} \quad p \leqq i \leqq q, \\ \varrho & \text{für} \quad i = k, \\ a_i & \text{sonst} \end{cases}$$

übergeht, wobei entweder $1 \leqq k < p \leqq q \leqq n$ oder $1 \leqq p \leqq q < k \leqq n$.

Bemerkungen zum Kapitel 3

B 3.1. Seit Wilkinson [65] gehört die systematische Verwendung elementarer Transformationsmatrizen zum Standard in der Literatur über numerische lineare Algebra. In dem zitierten Buch ist auch eine Fehleranalyse der Grundoperationen mit solchen Matrizen zu finden, der hier im wesentlichen gefolgt wird.

B 3.2. Spiegelungsmatrizen wurden erstmalig von Householder [58] mit der Vorzeichenfestlegung aus 3.3.5(ii) für ϱ zur orthogonalen Elimination verwendet. Als Skalierungsfaktor wird in der Literatur meist $\lambda = 1$ gewählt, was auf die einfache Darstellung $\boldsymbol{v} = \boldsymbol{a} - \varrho\boldsymbol{e}^1$, $\gamma = -\varrho v_1$ führt und vom Aufwand und der Fehleranalyse her etwas günstiger ist als die Festlegung $\lambda = -1/\varrho$. Allerdings ist $\boldsymbol{v}$ dann unter Umständen nicht gut skaliert, so daß Über- oder Unterlauf entstehen kann. Dies wird mit der hier gebrauchten Festlegung $\lambda = -1/\varrho$, die von den LINPACK-Autoren Dongarra et al. [79] verwendet wird, vermieden. Die stabile Realisierung mit $\varrho > 0$ gemäß 3.3.5(i) geht auf Parlett zurück und kann z. B. in dessen Buch [80a] nachgelesen werden.

B 3.3. Bereits Jacobi [1846!] benutzte ebene Drehungen in dem nach ihm benannten Verfahren zur Eigenwertbestimmung symmetrischer Matrizen. Für dieselbe Aufgabe sind sie später von Givens [54] eingesetzt worden, der sie dann auch [58] für den beschriebenen Eliminationsschritt und allgemeiner für das Problem aus Ü 3.3.2 verwendet hat. Die effektive Speicherung der Drehungsparameter mittels einer reellen Größe ist von Stewart [76b] vorgeschlagen worden; Algorithmus 3.3.12 ist eine Modifikation des dort beschriebenen Vorgehens.

B 3.4. Die implizite Realisierung von Givens-Drehungen mittels geeigneter Skalierungen geht auf Gentleman [73] und Hammerling [74] zurück. Die effektive Speicherung der Drehungsparameter gemäß 3.3.18 scheint neu zu sein.

II. Reguläre lineare Gleichungssysteme

4. Grundlegende Fakten über reguläre Gleichungssysteme

Wir betrachten in diesem Kapitel reguläre lineare Gleichungssysteme, d. h. Systeme mit quadratischen und zudem regulären Koeffizientenmatrizen, und untersuchen,

- wie sich Störungen der Koeffizientenmatrix und der rechten Seite auf die Lösung auswirken und welchen Einfluß eine Skalierung des Systems hat,
- nach welchem Prinzip die Gleichungssysteme in Systeme einfacher Struktur mit orthogonalen oder Dreiecksmatrizen transformiert werden können,
- wie solche Systeme einfacher Struktur gelöst werden.

4.1. Störungstheorie und Skalierung

A. *Störungstheorie regulärer Gleichungssysteme*

Vorgelegt sei das lineare Gleichungssystem

$$\boldsymbol{A}\boldsymbol{x} = \boldsymbol{b} \tag{1}$$

mit der regulären Koeffizientenmatrix $\boldsymbol{A} \in \mathbf{R}^{n,n}$ und der rechten Seite $\boldsymbol{b} \in \mathbf{R}^n$. Wegen der vorausgesetzten Regularität von $\boldsymbol{A}$ besitzt (1) für jedes $\boldsymbol{b}$ die eindeutige Lösung

$$\boldsymbol{x} = \boldsymbol{A}^{-1}\boldsymbol{b}, \tag{2}$$

vgl. Abschnitt 1.1. Wir betrachten jetzt neben (1) das gestörte System

$$(\boldsymbol{A} + \delta\boldsymbol{A})\,(\boldsymbol{x} + \delta\boldsymbol{x}) = \boldsymbol{b} + \delta\boldsymbol{b} \tag{3}$$

mit Störungen $\delta\boldsymbol{A} \in \mathbf{R}^{n,n}$, $\delta\boldsymbol{b} \in \mathbf{R}^n$ und fragen, wann (3) eine eindeutige Lösung $\boldsymbol{x} + \delta\boldsymbol{x}$ besitzt und wie gegebenenfalls $\|\delta\boldsymbol{x}\|$ durch $\|\delta\boldsymbol{A}\|$ und $\|\delta\boldsymbol{b}\|$ abgeschätzt werden kann.

Für die eindeutige Lösbarkeit von (3) ist die Regularität von $\boldsymbol{A} + \delta\boldsymbol{A}$ notwendig und hinreichend.

Wegen

$$\boldsymbol{A} + \delta\boldsymbol{A} = \boldsymbol{A}(\boldsymbol{I} + \boldsymbol{A}^{-1}\delta\boldsymbol{A}) = \boldsymbol{A}(\boldsymbol{I} - \boldsymbol{P}), \qquad \boldsymbol{P} := -\boldsymbol{A}^{-1}\delta\boldsymbol{A} \tag{4}$$

ist $\boldsymbol{A} + \delta\boldsymbol{A}$ genau dann regulär, wenn $\boldsymbol{I} - \boldsymbol{P}$ regulär ist. Es genügt daher, die Regularität von $\boldsymbol{I} - \boldsymbol{P}$ zu untersuchen.

4.1.1. *Aussage.* Für die Matrix $\boldsymbol{P} \subset \mathbf{R}^{n,n}$ gelte

$$\|\boldsymbol{P}\| < 1. \tag{5}$$

Dann ist $\boldsymbol{I} - \boldsymbol{P}$ regulär, und die Inverse genügt der Abschätzung

$$\|(\boldsymbol{I} - \boldsymbol{P})^{-1}\| \leqq 1/(1 - \|\boldsymbol{P}\|). \tag{6}$$

Beweis. Für jedes $\boldsymbol{x} \in \mathbf{R}^n$ gilt

$$\|(\boldsymbol{I} - \boldsymbol{P})\,\boldsymbol{x}\| \geqq |\,\|\boldsymbol{x}\| - \|\boldsymbol{P}\boldsymbol{x}\|\,| \geqq \|\boldsymbol{x}\| - \|\boldsymbol{P}\|\,\|\boldsymbol{x}\| = (1 - \|\boldsymbol{P}\|)\,\|\boldsymbol{x}\|. \tag{7}$$

Wegen (5) ist $1 - \|\boldsymbol{P}\| > 0$, so daß aus (7) für $\boldsymbol{x} \neq \boldsymbol{o}$ sofort $(\boldsymbol{I} - \boldsymbol{P})\,\boldsymbol{x} \neq \boldsymbol{o}$, also die Regularität von $\boldsymbol{I} - \boldsymbol{P}$ folgt. Wenn $\boldsymbol{x} := (\boldsymbol{I} - \boldsymbol{P})^{-1}\,\boldsymbol{y}$ gesetzt wird, geht (7) in

$$\|\boldsymbol{y}\| \geqq (1 - \|\boldsymbol{P}\|)\,\|(\boldsymbol{I} - \boldsymbol{P})^{-1}\,\boldsymbol{y}\| \quad \text{für alle } \boldsymbol{y} \in \mathbf{R}^n$$

über. Nach der Definition der Matrixnorm folgt hieraus (6), vgl. (1.1.39). □

Wir erinnern daran, daß wir stets eine der Vektornormen mit dem Index $p \in \{1, 2, \infty\}$ und für Matrizen die zugeordneten Matrixnormen desselben Index verwenden. Abweichungen — etwa die Verwendung der Frobeniusnorm — werden explizit angegeben.

Anwendung von 4.1.1 auf (4) führt auf das folgende Resultat:

4.1.2. *Störungslemma.* Es sei $\boldsymbol{A} \in \mathbf{R}^{n,n}$ eine reguläre Matrix, und die Störungsmatrix $\delta\boldsymbol{A} \in \mathbf{R}^{n,n}$ genüge der Bedingung

$$\varkappa := \|\boldsymbol{A}^{-1}\|\,\|\delta\boldsymbol{A}\| < 1. \tag{8}$$

Dann ist auch die gestörte Matrix $\boldsymbol{A} + \delta\boldsymbol{A}$ regulär, und es gelten

$$\|(\boldsymbol{A} + \delta\boldsymbol{A})^{-1}\| \leqq \|\boldsymbol{A}^{-1}\|/(1 - \varkappa) \tag{9}$$

sowie

$$\|(\boldsymbol{A} + \delta\boldsymbol{A})^{-1} - \boldsymbol{A}^{-1}\| \leqq [\|\boldsymbol{A}^{-1}\|^2/(1 - \varkappa)]\,\|\delta\boldsymbol{A}\|. \tag{10}$$

Beweis. Wegen (8) ist $\|\boldsymbol{P}\| \leqq \|\boldsymbol{A}^{-1}\|\,\|\delta\boldsymbol{A}\| < 1$. Aus 4.1.1 folgt dann sofort die Regularität von $\boldsymbol{I} - \boldsymbol{P}$, also die von $\boldsymbol{A} + \delta\boldsymbol{A}$, und die Abschätzung $\|(\boldsymbol{A} + \delta\boldsymbol{A})^{-1}\| = \|(\boldsymbol{I} - \boldsymbol{P})^{-1}\,\boldsymbol{A}^{-1}\| \leqq \|(\boldsymbol{I} - \boldsymbol{P})^{-1}\|\,\|\boldsymbol{A}^{-1}\| \leqq \|\boldsymbol{A}^{-1}\|/(1 - \varkappa)$. Die Ungleichung (10) ergibt sich schließlich aus der Identität

$$(\boldsymbol{A} + \delta\boldsymbol{A})^{-1} - \boldsymbol{A}^{-1} = -(\boldsymbol{A} + \delta\boldsymbol{A})^{-1}\,\delta\boldsymbol{A}\boldsymbol{A}^{-1}. \quad \square$$

Lemma 4.1.2 besagt: *Eine im Sinne von* $\|\delta\boldsymbol{A}\| \leqq \Delta\boldsymbol{A} < 1/\|\boldsymbol{A}^{-1}\|$ *kleine Störung läßt die reguläre Matrix* $\boldsymbol{A}$ *regulär, und es gilt*

$$\|(\boldsymbol{A} + \delta\boldsymbol{A})^{-1} - \boldsymbol{A}^{-1}\| \leqq L(\boldsymbol{A}, \Delta\boldsymbol{A})\,\|\delta\boldsymbol{A}\|, \qquad L(\boldsymbol{A}, \Delta\boldsymbol{A}) := \|\boldsymbol{A}^{-1}\|^2/(1 - \|\boldsymbol{A}^{-1}\|\,\Delta\boldsymbol{A}).$$

Dies bedeutet: *Die Aufgabe* $\boldsymbol{A} \to \boldsymbol{A}^{-1}$ *ist auf der Menge der regulären Matrizen lokal*

lipschitzstetig; die absolute bzw. relative Konditionszahl ist

$$L(\boldsymbol{A}) = \|\boldsymbol{A}^{-1}\|^2 \quad bzw. \quad K(\boldsymbol{A}) = \|\boldsymbol{A}\|\,\|\boldsymbol{A}^{-1}\|,$$

vgl. Abschnitt 2.1.

Unter der Voraussetzung (8) besitzt das gestörte System (3) die eindeutige Lösung $\boldsymbol{x} + \delta\boldsymbol{x}$. Ausmultiplizieren von (3) liefert dann unter Beachtung von $\boldsymbol{A}\boldsymbol{x} = \boldsymbol{b}$ die Beziehung

$$\delta\boldsymbol{x} = (\boldsymbol{A} + \delta\boldsymbol{A})^{-1}\,(-\delta\boldsymbol{A}\,\boldsymbol{x} + \delta\boldsymbol{b}). \tag{11}$$

Anwendung von 4.1.2 führt auf das folgende Resultat:

4.1.3. *Satz.* Gegeben seien das lineare Gleichungssystem

$$\boldsymbol{A}\boldsymbol{x} = \boldsymbol{b}, \qquad \boldsymbol{A} \in \mathbf{R}^{n,n} \text{ regulär},$$

mit der Lösung $\boldsymbol{x} = \boldsymbol{A}^{-1}\boldsymbol{b}$ sowie das gestörte System

$$(\boldsymbol{A} + \delta\boldsymbol{A})\,(\boldsymbol{x} + \delta\boldsymbol{x}) = \boldsymbol{b} + \delta\boldsymbol{b} \quad \text{mit} \quad \varkappa := \|\boldsymbol{A}^{-1}\|\,\|\delta\boldsymbol{A}\| < 1.$$

(i) Dann ist $\boldsymbol{A} + \delta\boldsymbol{A}$ regulär, und für die eindeutige Lösung $\boldsymbol{x} + \delta\boldsymbol{x}$ des gestörten Systems gilt

$$\delta\boldsymbol{x} = \delta\boldsymbol{x}' + O\big(\|\delta\boldsymbol{A}\|\,(\|\delta\boldsymbol{A}\| + \|\delta\boldsymbol{b}\|)\big),$$

wobei

$$\delta\boldsymbol{x}' := \boldsymbol{A}^{-1}\{-\delta\boldsymbol{A}\,\boldsymbol{x} + \delta\boldsymbol{b}\} \tag{12}$$

den bezüglich $\delta\boldsymbol{A}, \delta\boldsymbol{b}$ linearen Teil von $\delta\boldsymbol{x}$ bezeichnet.

(ii) Die Störung $\delta\boldsymbol{x}$ genügt den Abschätzungen

$$\|\delta\boldsymbol{x}\| \leqq \frac{\|\boldsymbol{A}^{-1}\|}{1-\varkappa}\,\{\|\delta\boldsymbol{A}\|\,\|\boldsymbol{x}\| + \|\delta\boldsymbol{b}\|\} \tag{13}$$

und im Fall $\boldsymbol{b} \neq \boldsymbol{o}$

$$\frac{\|\delta\boldsymbol{x}\|}{\|\boldsymbol{x}\|} \leqq \frac{1}{1-\varkappa}\left\{\|\boldsymbol{A}\|\,\|\boldsymbol{A}^{-1}\|\,\frac{\|\delta\boldsymbol{A}\|}{\|\boldsymbol{A}\|} + \frac{\|\boldsymbol{A}^{-1}\|\,\|\boldsymbol{b}\|}{\|\boldsymbol{x}\|}\,\frac{\|\delta\boldsymbol{b}\|}{\|\boldsymbol{b}\|}\right\}. \tag{14}$$

Beweis. Aus (11) folgt $\delta\boldsymbol{x} = \boldsymbol{A}^{-1}(-\delta\boldsymbol{A}\,\boldsymbol{x} + \delta\boldsymbol{b}) + \boldsymbol{d}$ mit $\boldsymbol{d} := [(\boldsymbol{A} + \delta\boldsymbol{A})^{-1} - \boldsymbol{A}^{-1}] \times (-\delta\boldsymbol{A}\,\boldsymbol{x} + \delta\boldsymbol{b})$, unter Beachtung von (10) also (12). Durch direkte Abschätzung von (11) ergibt sich mit (9) die Ungleichung (13), aus der (14) nach Division durch $\|\boldsymbol{x}\|$ folgt. □

Die Abschätzung (14) besagt: *Die Lösung regulärer Gleichungssysteme ist eine lokal lipschitzstetige Aufgabenklasse, und die partiellen relativen Konditionszahlen sind*

$$K_{\boldsymbol{A}}(\boldsymbol{A}, \boldsymbol{b}) = \|\boldsymbol{A}\|\,\|\boldsymbol{A}^{-1}\| \tag{15}$$

sowie

$$K_{\boldsymbol{b}}(\boldsymbol{A}, \boldsymbol{b}) = \frac{\|\boldsymbol{A}^{-1}\|\,\|\boldsymbol{b}\|}{\|\boldsymbol{x}\|}. \tag{16}$$

Wegen $\|\boldsymbol{x}\| = \|\boldsymbol{A}^{-1}\boldsymbol{b}\| \leqq \|\boldsymbol{A}^{-1}\|\,\|\boldsymbol{b}\|$ und $\|\boldsymbol{b}\| = \|\boldsymbol{A}\boldsymbol{x}\| \leqq \|\boldsymbol{A}\|\,\|\boldsymbol{x}\|$ gilt dabei

$$1 \leqq K_{\boldsymbol{b}}(\boldsymbol{A}, \boldsymbol{b}) \leqq K_{\boldsymbol{A}}(\boldsymbol{A}, \boldsymbol{b}).$$

Die Zahl

$$\operatorname{cond}(\boldsymbol{A}) := \|\boldsymbol{A}\|\,\|\boldsymbol{A}^{-1}\| \geqq 1$$

heißt daher schlechthin *Konditionszahl* von $\boldsymbol{A}$; gegebenenfalls wird $\operatorname{cond}(\boldsymbol{A})$ mit dem Index der verwendeten Norm versehen. Ist $\varkappa = \|\boldsymbol{A}^{-1}\|\,\|\delta\boldsymbol{A}\| = \operatorname{cond}(\boldsymbol{A})\,\|\delta\boldsymbol{A}\|/\|\boldsymbol{A}\| \ll 1$ und wird $K_{\boldsymbol{b}}$ durch $\operatorname{cond}(\boldsymbol{A})$ abgeschätzt, so geht (14) näherungsweise in

$$\frac{\|\delta\boldsymbol{x}\|}{\|\boldsymbol{x}\|} \lessapprox \operatorname{cond}(\boldsymbol{A}) \left\{\frac{\|\delta\boldsymbol{A}\|}{\|\boldsymbol{A}\|} + \frac{\|\delta\boldsymbol{b}\|}{\|\boldsymbol{b}\|}\right\}$$

über, und dies bedeutet: *Relative Störungen in der Matrix und der rechten Seite können sich höchstens mit dem Verstärkungsfaktor* $\operatorname{cond}(\boldsymbol{A})$ *auf die Lösung auswirken.* Demgemäß heißt eine Matrix *schlecht konditioniert*, wenn $\operatorname{cond}(\boldsymbol{A}) \gg 1$ — z. B. $\operatorname{cond}(\boldsymbol{A}) \geqq 10^4$ — gilt; das Gleichungssystem (1) heißt dann auch schlecht konditioniert.

B. *Die Qualität der Störungsabschätzungen*

Wir befassen uns im folgenden mit der Güte der Abschätzungen (13) und (14) und zeigen insbesondere, daß sie sich für $\delta\boldsymbol{A} \to \boldsymbol{O}$ nicht verbessern lassen, d. h., daß durch (15), (16) in der Tat die relativen Konditionszahlen von $\boldsymbol{x} = \boldsymbol{A}^{-1}\boldsymbol{b}$ gegeben sind. Dazu betrachten wir alle Systeme (3) mit den durch

$$\|\delta\boldsymbol{A}\| \leqq \Delta\boldsymbol{A}, \qquad \|\delta\boldsymbol{b}\| \leqq \Delta\boldsymbol{b} \tag{17}$$

charakterisierten Eingangsdaten $\{\boldsymbol{A} + \delta\boldsymbol{A}, \boldsymbol{b} + \delta\boldsymbol{b}\}$ und den im Fall $\|\boldsymbol{A}^{-1}\|\,\Delta\boldsymbol{A} < 1$ eindeutigen Lösungen $\boldsymbol{x} + \delta\boldsymbol{x}$, $\delta\boldsymbol{x}$ nach (11). Um den verkomplizierenden Einfluß von $\delta\boldsymbol{A}$ im Term $(\boldsymbol{A} + \delta\boldsymbol{A})^{-1}$ auszuschalten, untersuchen wir zunächst die linearisierten Störungen $\delta\boldsymbol{x}' = \boldsymbol{A}^{-1}\{-\delta\boldsymbol{A}\,\boldsymbol{x} + \delta\boldsymbol{b}\}$ gemäß (12) und stellen für diese den folgenden Hilfssatz bereit.

4.1.4. *Lemma.* Für alle $\delta\boldsymbol{A}, \delta\boldsymbol{b}$ gemäß (17) genügt

$$\delta\boldsymbol{z} := -\delta\boldsymbol{A}\,\boldsymbol{x} + \delta\boldsymbol{b} \tag{18}$$

der Abschätzung

$$\|\delta\boldsymbol{z}\| \leqq \Delta\boldsymbol{z} := \Delta\boldsymbol{A}\,\|\boldsymbol{x}\| + \Delta\boldsymbol{b}, \tag{19}$$

und diese Abschätzung ist exakt: Wenn $\delta\boldsymbol{A}, \delta\boldsymbol{b}$ die Normkugeln (17) durchlaufen, durchläuft $\delta\boldsymbol{z}$ die Normkugel (19). Insbesondere läßt sich jeder beliebige Vektor $\delta\boldsymbol{z}$ mit $\|\delta\boldsymbol{z}\| = \Delta\boldsymbol{z}$ in der Form (18) mit $\|\delta\boldsymbol{A}\| = \Delta\boldsymbol{A}$, $\|\delta\boldsymbol{b}\| = \Delta\boldsymbol{b}$ darstellen.

Beweis. Aus (18) folgt sofort (19). Gegeben sei nun $\delta\boldsymbol{z}$ mit $\|\delta\boldsymbol{z}\| \leqq \Delta\boldsymbol{z}$, also $\delta\boldsymbol{z} = \lambda(\Delta\boldsymbol{z})\,\boldsymbol{z}$ mit $0 \leqq \lambda \leqq 1$ und $\|\boldsymbol{z}\| = 1$. Dann wird $\delta\boldsymbol{z}$ durch $\delta\boldsymbol{b} := \lambda(\Delta\boldsymbol{b})\,\boldsymbol{z}$ und

$$-\delta\boldsymbol{A} := \lambda(\Delta\boldsymbol{A})\,\boldsymbol{z}\boldsymbol{y}^{\mathsf{T}}, \quad \boldsymbol{y} := \begin{cases} (\operatorname{sgn}(x_1), \ldots, \operatorname{sgn}(x_n))^{\mathsf{T}} & \text{für } p = 1, \\ \boldsymbol{x}/\|\boldsymbol{x}\|_2 & \text{für } p = 2, \\ (\operatorname{sgn}(x_j))\,\boldsymbol{e}^j & \text{für } p = \infty \text{ mit } |x_j| = \|\boldsymbol{x}\|_\infty \end{cases}$$

im Fall $\boldsymbol{x} \neq \boldsymbol{o}$ bzw. $\delta\boldsymbol{A} := \boldsymbol{O}$ im Fall $\boldsymbol{x} = \boldsymbol{o}$ reproduziert; man beachte Abschnitt 1.1.I und Ü 1.1.3. Die Zusatzbemerkung ergibt sich für $\lambda = 1$. □

4.1.5. *Aussage.* Für alle δA, δb gemäß (17) genügen die linearisierten Störungen

$$\delta x' := A^{-1}\delta z = A^{-1}\{-\delta A\, x + \delta b\}$$

der Abschätzung

$$\|\delta x'\| \leqq \Delta x' := \|A^{-1}\|\, \Delta z = \|A^{-1}\|\, \{\Delta A\, \|x\| + \Delta b\}, \tag{20}$$

und diese Abschätzung ist scharf: Es gibt Störungen δA, δb mit $\|\delta A\| = \Delta A$, $\|\delta b\| = \Delta b$, für die in (20) das Gleichheitszeichen steht.

Beweis. Die Schranke (20) ist trivial. Zum Nachweis der Schärfe beachten wir, daß zu A^{-1} ein z mit $\|z\| = 1$ und $\|A^{-1}z\| = \|A^{-1}\|\, \|z\|$ existiert, vgl. Abschnitt 1.1.I. Mit $\delta z := (\Delta z)\, z$ und 4.1.4 folgt dann die Behauptung. □

4.1.6. *Folgerung.* Es gelte $\|A^{-1}\|\, \Delta A < 1$. Dann ist die für alle δA, δb gemäß (17) gültige und zu (13) äquivalente Abschätzung

$$\|\delta x\| \leqq \Delta x := \Delta x'/(1 - \|A^{-1}\|\, \Delta A) = \|A^{-1}\|\, \{\Delta A\, \|x\| + \Delta b\}/(1 - \|A^{-1}\|\, \Delta A) \tag{21}$$

scharf für $\Delta A \to 0$: Zu $\Delta A, \Delta b$ mit $\Delta x > 0$ gibt es Störungen δA, δb mit $\|\delta A\| = \Delta A$, $\|\delta b\| = \Delta b$, so daß

$$\|\delta x\|/\Delta x \to 1 \quad \text{für} \quad \Delta A \to 0 \tag{22}$$

gilt.

Dies besagt: *Die Abschätzungen* (13) *und* (14) *sind für* $\Delta A \to 0$ *scharf bezüglich der Gesamtheit der durch* (17) *charakterisierten Störungen.* Für eine einzelne feste Störung braucht (13) bzw. (14) natürlich nicht scharf zu sein, siehe Ü 4.1.5. Welcher Teil der Normkugel (20) durch die in bezug auf (17) zulässigen linearisierten Störungen $\delta x'$ ausgefüllt wird, läßt sich für die 2-Norm in einfacher Weise explizit angeben.

4.1.7. *Satz.* Wenn δA, δb die Normkugeln $\|\delta A\|_2 \leqq \Delta A$, $\|\delta b\|_2 \leqq \Delta b$ durchlaufen, durchlaufen die linearisierten Störungen $\delta x'$ das Ellipsoid $\mathfrak{E}' = \mathfrak{E}'(A, b, \Delta A, \Delta b)$ mit dem Mittelpunkt O und den Halbachsen

$$\pm\alpha_j v^j, \qquad \alpha_j := \Delta z/\sigma_j \qquad (j = 1, \ldots, n),$$

wobei σ_j und v^j die Singulärwerte bzw. Singulärvektoren von A nach 1.2.14 bezeichnen.

Beweis. Mit der Singulärwertzerlegung $A = U\Sigma V^T$ gilt $\delta x' = A^{-1}\delta z = V\Sigma^{-1}U^T\delta z$. Nach 4.1.4 durchläuft δz die Normkugel $\|\delta z\|_2 \leqq \Delta z$, und dasselbe gilt wegen der Orthogonalität von U für $\delta y := (\delta y_j) = U^T\delta z$. Die zulässigen Störungen lassen sich dann durch

$$\delta x' = \sum_{j=1}^{n} v^j(\delta y_j/\sigma_j) \quad \text{mit} \quad \sum_{j=1}^{n} (\delta y_j)^2 \leqq (\Delta z)^2 \tag{23}$$

darstellen. Die Größen $\xi_j := \delta y_j/\sigma_j$ sind die Komponenten von $\delta x'$ bezüglich der orthonormierten Basis $\{v^1, \ldots, v^n\}$, mit denen sich (23) äquivalent

$$\delta x' = \sum_{j=1}^{n} v^j\xi_j \quad \text{mit} \quad \sum_{j=1}^{n} \left(\frac{\xi_j}{\alpha_j}\right)^2 = \sum_{j=1}^{n} \left(\frac{\delta y_j}{\Delta z}\right)^2 \leqq 1$$

schreiben läßt. Dies ist gerade die Gleichung des Ellipsoids $\mathfrak{E}'$. □

Die längste Halbachse von $\mathscr{E}'$ hat die Länge $\alpha_n = \Delta z/\sigma_n = \Delta x'$, so daß die Normkugel $\mathscr{K}' = \mathscr{K}'(\boldsymbol{A}, \boldsymbol{b}, \Delta\boldsymbol{A}, \Delta\boldsymbol{b}) := \{\delta x' : \|\delta x'\| \leqq \Delta x'\}$ aus (20) in der Tat die umschriebene Kugel zu $\mathscr{E}'$ darstellt. Das Verhältnis der längsten zur kürzesten Halbachse von $\mathscr{E}'$ ist

$$\alpha_n/\alpha_1 = \sigma_1/\sigma_n = \operatorname{cond}_2(\boldsymbol{A}). \tag{24}$$

Je schlechter die Matrix konditioniert ist, um so mehr weicht also $\mathscr{E}'$ von der Kugel $\mathscr{K}'$ ab, siehe Abb. 4.1.1.

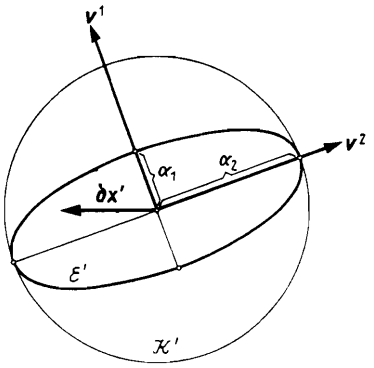


Abb. 4.1.1. Linearisierte Störungen für $n = 2$ im Fall der 2-Norm

Bei der Anwendung von (13) zur Abschätzung von $\|\delta x\|$ entsteht folgendes Problem: Häufig sind für die Störungen $\delta\boldsymbol{A} = (\delta a_{ij})$, $\delta\boldsymbol{b} = (\delta b_i)$ der Eingangsdaten individuelle Schranken mit

$$|\delta a_{ij}| \leqq \Delta a_{ij}, \qquad |\delta b_i| \leqq \Delta b_i \qquad (i, j = 1, \ldots, n) \tag{25}$$

bekannt, vgl. Abschnitt 2.3.C. Wenn diese in der Matrix $\Delta\boldsymbol{A} := (\Delta a_{ij})$ bzw. im Vektor $\Delta\boldsymbol{b} := (\Delta b_i)$ zusammengefaßt werden, läßt sich (25) kurz als

$$|\delta\boldsymbol{A}| \leqq \Delta\boldsymbol{A}, \qquad |\delta\boldsymbol{b}| \leqq \Delta\boldsymbol{b} \tag{26}$$

schreiben, siehe Abschnitt 1.1.J. Für die hier betrachteten Normen $p \in \{1, 2, \infty\}$ folgt aus (26)

$$\|\delta\boldsymbol{A}\| \leqq \|\Delta\boldsymbol{A}\| =: \Delta\boldsymbol{A}, \qquad \|\delta\boldsymbol{b}\| \leqq \|\Delta\boldsymbol{b}\| =: \Delta\boldsymbol{b}, \tag{27}$$

also (17) mit den in (27) definierten Werten von $\Delta\boldsymbol{A}$ und $\Delta\boldsymbol{b}$. Die $n^2 + n$ Schranken aus (25) werden dann durch die zwei Zahlen aus (27) repräsentiert. Obwohl die zugehörigen Schranken (13) in bezug auf (27) asymptotisch scharf sind, wird dies in bezug auf (26) i. allg. nicht zu erwarten sein. Es ist daher sinnvoll zu fragen, unter welchen Bedingungen die Schranke (13) das durch (26) charakterisierte individuelle Fehlerverhalten realistisch widerspiegelt. Da wir an komponentenweisen Abschätzungen für δx interessiert sind, bietet sich in (13) die Verwendung der ∞-Norm an.

Der Einfachheit halber betrachten wir wieder die linearisierten Störungen $\delta x'$ und gehen analog zum Fall der Normschranken vor.

4.1.8. *Lemma.* Für alle $\delta A, \delta b$ gemäß (26) genügt $\delta z = -\delta A\, x + \delta b$ der Abschätzung

$$|\delta z| \leqq \Delta z := \Delta A\, |x| + \Delta b, \tag{28}$$

und diese Abschätzung ist exakt, d. h., sie beschreibt genau die Menge der bezüglich (26) zulässigen Störungen δz. Insbesondere läßt sich jeder beliebige Vektor δz mit $|\delta z| = \Delta z$ in der Form (18) darstellen mit Störungen $\delta A, \delta b$, für die $|\delta A| = \Delta A$ und $|\delta b| = \Delta b$ gilt.

Beweis. Wegen

$$|\delta z| = |-\delta A\, x + \delta b| \leqq |\delta A\, x| + |\delta b| \leqq |\delta A|\, |x| + |\delta b| \leqq \Delta A\, |x| + \Delta b$$

gilt (28). Für den Beweis der Zusatzbemerkung setze man $\delta a_{ij} := \pm \Delta a_{ij}$, $\delta b_i := \pm \Delta b_i$ und verteile die Vorzeichen so, daß $|-\delta A\, x + \delta b| = |\delta A|\, |x| + |\delta b|$ gilt; dies ist offenbar stets möglich. □

Zur Untersuchung von $\delta x'$ benötigen wir das folgende Lemma.

4.1.9. *Lemma.* Für $B \in \mathbf{R}^{n,n}$ und δz mit $|\delta z| \leqq \Delta z$ werde $\delta x' := B\, \delta z$ gebildet. Dann gilt

(i) $$|\delta x'| \leqq |B|\, |\delta z| \leqq \Delta x' := |B|\, \Delta z, \tag{29}$$

und diese Abschätzung ist komponentenweise scharf: Zu jedem i gibt es ein δz mit $|\delta z| = \Delta z$, so daß in der i-ten Komponente von (29) das Gleichheitszeichen steht.

(ii) $$\|\delta x'\|_\infty \leqq \|B\|_\infty\, \|\Delta z\|_\infty, \tag{30}$$

und diese Abschätzung ist scharf, wenn $\Delta z_i = \zeta$ $(i = 1, \ldots, n)$ gilt: Es gibt dann ein δz mit $|\delta z| = \Delta z$, für welches in (30) das Gleichheitszeichen steht.

Beweis. Die Herleitung von (29) ist trivial. Zum Nachweis der Schärfe setzen wir $(\delta z)_j = \pm \Delta z_j$ und verteilen die Vorzeichen so, daß $|(B \delta z)_i| = (|B|\, |\delta z|)_i$ gilt. Wegen der Monotonie und Absolutheit der ∞-Norm folgt (30) direkt aus (29). Sei nun $\Delta z_i = \zeta$ für alle i. Nach Definition von $\|B\|_\infty$ gibt es ein z mit $|z_j| = 1$ $(j = 1, \cdots, n)$ und $\|Bz\|_\infty = \|B\|_\infty\, \|z\|_\infty$, vgl. (1.1.41). Für $\delta z := \zeta z$ ist dann (30) scharf, und es gilt $|\delta z| = \Delta z$. □

4.1.10. *Aussage.* Für alle $\delta A, \delta b$ gemäß (26) werde

$$\delta x' = A^{-1} \delta z = A^{-1}\{-\delta A\, x + \delta b\}$$

gebildet. Dann gilt

(i) $$|\delta x'| \leqq |A^{-1}|\, \Delta z \quad \text{mit} \quad \Delta z := \Delta A\, |x| + \Delta b, \tag{31}$$

und diese Abschätzung ist komponentenweise scharf: Zu jedem i gibt es Störungen $\delta A, \delta b$ mit $|\delta A| = \Delta A$, $|\delta b| = \Delta b$, so daß in der i-ten Komponente von (31) das Gleichheitszeichen steht.

(ii) $$\|\delta x'\|_\infty \leqq \|A^{-1}\|_\infty \|\Delta z\|_\infty = \|A^{-1}\|_\infty \|\Delta A\ |x| + \Delta b\|_\infty. \tag{32}$$

Im Fall $\Delta z_i = \zeta$ $(i = 1, \ldots, n)$ ist diese Abschätzung scharf: Es gibt Störungen $\delta A, \delta b$ mit $|\delta A| = \Delta A$, $|\delta b| = \Delta b$, so daß in (32) das Gleichheitszeichen steht.

(iii) $$\|\delta x'\|_\infty \leqq \|A^{-1}\|_\infty \{\|\Delta A\|_\infty \|x\|_\infty + \|\Delta b\|_\infty\}. \tag{33}$$

Im Fall

$$|x_i| = \xi, \qquad \Delta b_i = \beta, \qquad \sum_{j=1}^{n} \Delta a_{ij} = \sigma \qquad (i = 1, \ldots, n) \tag{34}$$

ist (33) scharf analog zu (ii).

Beweis. Nach 4.1.8 wird für δA, δb nach (26) der Bereich $|\delta z| \leqq \Delta z$ voll ausgeschöpft. Anwendung von 4.1.9 mit $B := A^{-1}$ liefert dann die Aussagen (i) und (ii). Aus (32) folgt sofort (33). Es bleibt zu zeigen, daß (33) unter der Voraussetzung (34) scharf ist. Aus (34) folgt nun $\Delta z_i = \sum_j \Delta a_{ij} |x_j| + \Delta b_i = \sigma\xi + \beta$, d. h., Δz erfüllt die Voraussetzungen von (ii), und (32) ist scharf. Wegen $\|\Delta z\|_\infty = \sigma\xi + \beta = \|\Delta A\|_\infty \|x\|_\infty + \|\Delta b\|_\infty$ stimmen die rechten Seiten von (32) und (33) überein, so daß auch (33) scharf ist. □

Im folgenden wird ein Vektor mit betragsgleichen Komponenten *äquilibriert* genannt. Analog heißt eine Matrix *zeilenäquilibriert*, falls die Zeilensummen der Beträge gleich sind.

Wenn die Störungen $\delta A, \delta b$ den Bedingungen (26) genügen, geht (13) für die ∞-Norm in

$$\|\delta x\|_\infty \leqq \frac{\|A^{-1}\|_\infty}{1 - \|A^{-1}\|_\infty \|\Delta A\|_\infty} \{\|\Delta A\|_\infty \|x\|_\infty + \|\Delta b\|_\infty\} \tag{35}$$

über, vgl. 4.1.6. Aus 4.1.10 lassen sich dann hinreichende Bedingungen ablesen, unter denen die Schranke (35) realistisch ist.

4.1.11. *Aussage.* Es sei $\|A^{-1}\|_\infty \|\Delta A\|_\infty < 1$, und die Komponenten x_i von x mögen sich betragsmäßig nicht zu sehr unterscheiden. Dann ist die Abschätzung (35) realistisch in bezug auf die Störungen (26), sofern ΔA zeilenäquilibriert und Δb äquilibriert ist.

4.1.12. *Bemerkung.* Falls die Elemente von A dasselbe relative Fehlerniveau haben d. h., falls

$$|\delta a_{ij}| \leqq \varepsilon\ |a_{ij}| =: \Delta a_{ij}$$

gilt, ist $\sum_j \Delta a_{ij} = \varepsilon \sum_j |a_{ij}|$ und ΔA folglich genau dann zeilenäquilibriert, wenn A selbst diese Eigenschaft hat. Analoge Aussagen gelten für b und Δb.

Wir bemerken abschließend, daß (35) und allgemein die Abschätzungen aus 4.1.3 für nicht ausreichend äquilibrierte Schranken ΔA und Δb das bezüglich (26) wahre Fehlerniveau wesentlich überschreiten können. Zur Illustration geben wir das folgende Beispiel an, siehe auch Ü 4.1.7.

4.1.13. *Beispiel.* Für

$$\boldsymbol{A} = \begin{pmatrix} 1 & 0.5 & 0 \\ 0.97 & 0.48 & 0 \\ 0.75 & 0 & -0.75 \end{pmatrix}, \quad \Delta\boldsymbol{A} = \boldsymbol{O}, \quad \boldsymbol{b} = \begin{pmatrix} 3.50 \\ 3.37 \\ -6.75 \end{pmatrix}, \quad \Delta\boldsymbol{b} = 10^{-3} \begin{pmatrix} 1 \\ 1 \\ 75 \end{pmatrix}$$

ist

$$\boldsymbol{A}^{-1} = \begin{pmatrix} -96 & 100 & 0 \\ 194 & -200 & 0 \\ -96 & 100 & -4/3 \end{pmatrix}, \quad \boldsymbol{x} = \begin{pmatrix} 1 \\ 5 \\ 10 \end{pmatrix}, \quad |\delta\boldsymbol{x}| \leqq |\boldsymbol{A}^{-1}|\, \Delta\boldsymbol{b} = 10^{-3} \begin{pmatrix} 196 \\ 394 \\ 296 \end{pmatrix}.$$

Die letzte Abschätzung ist nach 4.1.10 scharf. Die Abschätzung (35) liefert dagegen

$$\|\delta\boldsymbol{x}\|_\infty \leqq \|\boldsymbol{A}^{-1}\|_\infty \, \|\Delta\boldsymbol{b}\|_\infty = 10^{-3} \times 394 \times 75,$$

also eine um den Faktor 75 zu grobe Schranke.

Wenn die dritte Gleichung des Systems durch 75 dividiert wird, erhält man das durch

$$\hat{\boldsymbol{A}} = \begin{pmatrix} 1 & 0.5 & 0 \\ 0.97 & 0.48 & 0 \\ 0.01 & 0 & -0.01 \end{pmatrix}, \quad \Delta\hat{\boldsymbol{A}} = \boldsymbol{O}, \quad \hat{\boldsymbol{b}} = \begin{pmatrix} 3.50 \\ 3.37 \\ -0.09 \end{pmatrix}, \quad \Delta\hat{\boldsymbol{b}} = 10^{-3} \begin{pmatrix} 1 \\ 1 \\ 1 \end{pmatrix}$$

und

$$\hat{\boldsymbol{A}}^{-1} = \begin{pmatrix} -96 & 100 & 0 \\ 194 & -200 & 0 \\ -96 & 100 & -100 \end{pmatrix}$$

charakterisierte äquivalente System mit äquilibriertem $\Delta\boldsymbol{b}$. Die Schranke (35) liefert hier

$$\|\delta\boldsymbol{x}\|_\infty \leqq \|\hat{\boldsymbol{A}}^{-1}\|_\infty \, \|\Delta\hat{\boldsymbol{b}}\|_\infty = 10^{-3} \times 394 \times 1,$$

also eine scharfe Abschätzung des möglichen Fehlers. □

C. *Skalierung*

Das eben diskutierte Beispiel zeigt, daß sich die für eine realistische Fehlerabschätzung günstigen Eigenschaften von $\Delta\boldsymbol{A}$ bzw. $\Delta\boldsymbol{b}$ unter Umständen erzielen lassen, wenn Eingangsdaten und Fehlerschranken zeilenweise mit geeigneten positiven Faktoren multipliziert werden. Dieser Übergang vom Originalsystem in das äquivalente System

$$\bar{\boldsymbol{A}}\boldsymbol{x} = \bar{\boldsymbol{b}} \quad \text{mit} \quad \bar{\boldsymbol{A}} := \boldsymbol{D}\boldsymbol{A}, \qquad \bar{\boldsymbol{b}} := \boldsymbol{D}\boldsymbol{b} \tag{36}$$

und den Fehlerschranken

$$\Delta\bar{\boldsymbol{A}} := \boldsymbol{D}\,\Delta\boldsymbol{A}, \qquad \Delta\bar{\boldsymbol{b}} := \boldsymbol{D}\,\Delta\boldsymbol{b} \tag{37}$$

mittels der Diagonalmatrix

$$\boldsymbol{D} = \operatorname{diag}(d_i), \qquad d_i > 0,$$

wird *Skalierung* — genauer: *Zeilenskalierung* — genannt.

Bei der Skalierung werden die Eingangsdaten $\{\boldsymbol{A}, \boldsymbol{b}\}$ des Originalsystems zu $\{\bar{\boldsymbol{A}}, \bar{\boldsymbol{b}}\}$ verändert. Ein numerischer Algorithmus $\boldsymbol{V}$ zur Gleichungsauflösung wird daher bei

Anwendung auf das skalierte System (36) i. allg. eine etwas andere Lösung $\bar{\boldsymbol{x}} = \boldsymbol{V}(\bar{\boldsymbol{A}}, \bar{\boldsymbol{b}})$ liefern als bei Anwendung auf das Originalsystem, wo sich $\boldsymbol{x} = \boldsymbol{V}(\boldsymbol{A}, \boldsymbol{b})$ ergibt.

Die praktisch verwendeten direkten Lösungsverfahren $\boldsymbol{V}$ sind alle numerisch gutartig in bezug auf $\boldsymbol{A}$, siehe Kapitel 5 und 11. Es gibt dann eine Störung $\delta\boldsymbol{A}_R$, so daß die berechnete Lösung $\boldsymbol{x}$ dem gestörten System

$$(\boldsymbol{A} + \delta\boldsymbol{A}_R)\,\boldsymbol{x} = \boldsymbol{b} \quad \text{mit} \quad \|\delta\boldsymbol{A}_R\| \leqq \nu F \|\boldsymbol{A}\| \tag{38}$$

genügt, wobei die Kumulationskonstante F nur vom Verfahren und der Dimension n, nicht aber von $\boldsymbol{A}$ abhängt, vgl. Abschnitt 2.3. Der erzeugte Rundungsfehler $\delta\boldsymbol{x}_R := \boldsymbol{x} - \boldsymbol{x}^*$ gegenüber der exakten Lösung $\boldsymbol{x}^* = \boldsymbol{A}^{-1}\boldsymbol{b}$ kann dann nach (13) bzw. (35) zu

$$\|\delta\boldsymbol{x}_R\|_\infty \leqq \nu F \operatorname{cond}_\infty(\boldsymbol{A})\, \|\boldsymbol{x}^*\|_\infty / [1 - \nu F \operatorname{cond}_\infty(\boldsymbol{A})] \tag{39}$$

abgeschätzt werden. Für das skalierte System gilt (39) mit $\boldsymbol{DA}$ anstelle von $\boldsymbol{A}$, so daß die Frage entsteht, ob sich $\operatorname{cond}_\infty(\boldsymbol{DA})$ und damit die Schranke (39) durch geeignete Skalierung von $\boldsymbol{A}$ minimieren läßt. Der folgende Satz gibt eine positive Antwort.

4.1.14. *Satz.* Es sei

$$\boldsymbol{A} = \begin{pmatrix} -\boldsymbol{a}^{1\mathrm{T}}- \\ \vdots \\ -\boldsymbol{a}^{n\mathrm{T}}- \end{pmatrix} \in \mathbf{R}^{n,n}$$

eine reguläre Matrix mit den Zeilen $\boldsymbol{a}^{i\mathrm{T}}$, und mit $\boldsymbol{D} = \operatorname{diag}(d_i)$,

$$d_i := \frac{\max\limits_k \sum\limits_j |a_{kj}|}{\sum\limits_j |a_{ij}|} = \frac{\max\limits_k \|\boldsymbol{a}^k\|_1}{\|\boldsymbol{a}^i\|_1}, \tag{40}$$

werde $\bar{\boldsymbol{A}} := \boldsymbol{DA}$ gebildet. Dann gelten

$$\|\boldsymbol{A}\|_\infty = \|\bar{\boldsymbol{A}}\|_\infty, \qquad \|\boldsymbol{A}^{-1}\|_\infty / \max d_i \leqq \|\bar{\boldsymbol{A}}^{-1}\|_\infty \leqq \|\boldsymbol{A}^{-1}\|_\infty \tag{41}$$

und

$$\operatorname{cond}_\infty(\boldsymbol{A}) / \max d_i \leqq \operatorname{cond}_\infty(\bar{\boldsymbol{A}}) \leq \operatorname{cond}_\infty(\boldsymbol{A}). \tag{42}$$

Beweis. Die Gleichheit von $\|\boldsymbol{A}\|_\infty$ und $\|\bar{\boldsymbol{A}}\|_\infty$ folgt sofort aus (40), denn $\bar{\boldsymbol{A}}$ ist zeilenäquilibriert, und die Zeilensummen von $\bar{\boldsymbol{A}}$ sind gleich der maximalen Zeilensumme von $\boldsymbol{A}$. Wegen $\|\boldsymbol{D}\|_\infty = \max d_i$ und $\|\boldsymbol{D}^{-1}\|_\infty = 1/\min d_i = 1$ folgt weiter $\|\bar{\boldsymbol{A}}^{-1}\|_\infty = \|\boldsymbol{A}^{-1}\boldsymbol{D}^{-1}\|_\infty \leqq \|\boldsymbol{A}^{-1}\|_\infty \|\boldsymbol{D}^{-1}\|_\infty = \|\boldsymbol{A}^{-1}\|_\infty$ sowie $\|\boldsymbol{A}^{-1}\|_\infty = \|\bar{\boldsymbol{A}}^{-1}\boldsymbol{D}\|_\infty \leqq \|\bar{\boldsymbol{A}}^{-1}\|_\infty \|\boldsymbol{D}\|_\infty = \|\bar{\boldsymbol{A}}^{-1}\|_\infty \max d_i$, also (41). Aus der Definition von $\operatorname{cond}(\boldsymbol{A})$ ergibt sich (42). □

Die Abschätzung (42) besagt: *Unter allen durch Zeilenskalierung aus einer Matrix hervorgehenden Matrizen hat jede zeilenäquilibrierte die kleinste Kondition in der ∞-Norm.* Man beachte dabei, daß sich die Kondition von $\boldsymbol{DA}$ nicht ändert, wenn $\boldsymbol{D}$ durch $\lambda\boldsymbol{D}$, $\lambda > 0$, ersetzt wird, d. h., es kommt nur auf das Verhältnis der Skalierungsfaktoren d_i an. Die in (40) vorgenommene Normierung $\min d_i = 1$ wird nur für die (absoluten) Abschätzungen (41) benötigt.

In bezug auf die Schranke (39) *für den erzeugten Rundungsfehler* δx_R *ist nach* 4.1.14 *also eine Zeilenäquilibrierung des Systems optimal.*

In der Regel wird der Fehler δx_R durch weitere Anteile δx_D, δx_M zum Gesamtfehler $\delta x := \delta x_R + \delta x_D + \delta x_M$ ergänzt. Wenn die Eingangsdatenelemente nicht bereits Computerzahlen sind, tritt ein relativ durch ν beschränkter, also sehr kleiner Darstellungsfehler δA_D, δb_D auf. Nach 2.3.15 gilt für diesen

$$\|\delta A_D\|_\infty \leqq \nu \|A\|_\infty, \qquad \|\delta b_D\|_\infty \leqq \nu \|b\|_\infty.$$

Der zugehörige Fehler δx_D in der Lösung ist daher durch

$$\|\delta x_D\|_\infty \leqq 2\nu \operatorname{cond}_\infty(A) \|x^*\|_\infty / \big(1 - \nu \operatorname{cond}_\infty(A)\big) \tag{43}$$

beschränkt, also etwa um den Faktor $F/2$ kleiner als der erzeugte Rundungsfehler und wird deshalb i. allg. vernachlässigt werden können.

Wenn die Eingangsdaten das Resultat realer Messungen sind oder durch komplizierte, eventuell nur näherungsweise ausgeführte Berechnungen erhalten wurden, treten weitere Fehler δA_M, δb_M auf, die deutlich über dem Niveau des Darstellungsfehlers liegen. Man beachte, daß ν für reale Computer in der Größenordnung $10^{-6} \ldots 10^{-14}$ liegt, während Meßfehler meist eine relative Genauigkeit von $10^{-1} \ldots 10^{-3}$ aufweisen. Solche „großen" Fehler δA_M, δb_M wollen wir unabhängig von ihrer konkreten Quelle *Meßfehler* nennen, und der durch sie hervorgerufene Fehler δx_M in der Lösung dominiert i. allg. die Anteile δx_R und δx_D deutlich. Die durch Zeilenäquilibrierung mögliche Verkleinerung von δx_R wird daher bedeutungslos, dagegen ist die realistische Abschätzung des durch die Meßfehler hervorgerufenen Fehleranteils δx_M wesentlich. Diese kann jedoch durch die Zeilenäquilibrierung verhindert werden; vgl. Beispiel 4.1.13, wo sich mit der fast zeilenäquilibrierten Matrix A eine wesentlich schlechtere Abschätzung als für die nicht äquilibrierte Matrix $\hat{A}$ ergibt. *Aus* 4.1.9 *und* 4.1.10 *wird klar, daß bei dominanten Meßfehlern nicht die Matrix* A, *sondern der Vektor* $\Delta z = \Delta A|x| + \Delta b$ *bzw. die Schranken* ΔA *und* Δb *zu äquilibrieren sind.*

4.1.15. *Skalierungsregeln.*

(R_1) Wenn die Eingangsdaten keine Meßfehler aufweisen, skaliere man so, daß A zeilenäquilibriert wird:

$$d_i := \max_k \sum_j |a_{kj}| / \sum_j |a_{ij}|.$$

(R_2) Wenn die rechte Seite b einen Meßfehler δb_M mit $|\delta b_M| \leqq \Delta b$ besitzt und A im Rahmen des Darstellungsfehlers exakt ist oder einen gegenüber Δb vernachlässigbaren Fehler aufweist, skaliere man so, daß Δb äquilibriert wird:

$$d_i := \max_k \Delta b_k / \Delta b_i.$$

(R_3) Wenn die Matrix A einen Meßfehler δA_M mit $|\delta A_M| \leqq \Delta A$ besitzt und b im Rahmen des Darstellungsfehlers exakt ist oder einen gegenüber ΔA ver-

nachlässigbaren Fehler besitzt, skaliere man so, daß $\Delta \boldsymbol{A}$ zeilenäquilibriert wird:

$$d_i := \max_k \sum_j \Delta a_{kj} / \sum_j \Delta a_{ij}.$$

(R_4) Wenn die Matrix $\boldsymbol{A}$ und die rechte Seite $\boldsymbol{b}$ Meßfehler mit $|\delta \boldsymbol{A}_M| \leqq \Delta \boldsymbol{A}$ und $|\delta \boldsymbol{b}_M| \leqq \Delta \boldsymbol{b}$ aufweisen, die beide zum Fehlervektor

$$\Delta \boldsymbol{z} = \Delta \boldsymbol{A}\,|\boldsymbol{x}| + \Delta \boldsymbol{b} \tag{44}$$

beitragen, skaliere man so, daß $\Delta \boldsymbol{z}$ äquilibriert wird:

$$d_i := \max_k \Delta z_k / \Delta z_i.$$

4.1.16. *Bemerkung.* (i) Die Regel (R_1) ist lediglich in bezug auf die kollektiven Schranken (38), (39) optimal und berücksichtigt nicht die vom verwendeten Lösungsverfahren abhängige Struktur der individuellen äquivalenten Störungen $(\delta \boldsymbol{A}_R)_{ij}$. Sie braucht daher für ein konkretes Verfahren und eine konkrete Matrix nicht optimal zu sein; siehe auch Abschnitt 5.4. Da $\delta \boldsymbol{A}_R$ eine einzelne, feste Störung darstellt, wird die Abschätzung (39) auch dann meist zu pessimistisch sein, wenn (38) realistisch ist, d. h., wenn $\|\delta \boldsymbol{A}_R\| \approx vF\,\|\boldsymbol{A}\|$ sein sollte. Aus diesem Grund ist (39) i. allg. nicht zur realen Abschätzung des erzeugten Rundungsfehlers geeignet. Dieser kann nach anderen Methoden besser abgeschätzt werden, etwa über das Residuum $\boldsymbol{r} = \boldsymbol{b} - \boldsymbol{A}\boldsymbol{x}$ von $\boldsymbol{x}$ nach den im folgenden Teilabschnitt D angegebenen Methoden.

(ii) Durch den Fehlervektor $\Delta \boldsymbol{z}$ gemäß (44) wird erfaßt, welchen Anteil die Fehler $\Delta \boldsymbol{A}_M$ und $\Delta \boldsymbol{b}_M$ zum Gesamtfehler liefern können. Abgesehen vom Sonderfall $\Delta \boldsymbol{A} = \boldsymbol{O}$ — hier geht Regel (R_4) in (R_2) über — ist dazu i. allg. jedoch die Kenntnis von $|\boldsymbol{x}|$ erforderlich. Die für eine im Sinne von 4.1.10(ii) entsprechend (R_4) optimale Skalierung ist daher erst a posteriori — d. h. nach Berechnung von $\boldsymbol{x}$ — möglich und führt auf einen zweistufigen Prozeß:

Stufe 1: Skaliere $\boldsymbol{A}$ so, daß der bei Lösung von $\boldsymbol{A}\boldsymbol{x} = \boldsymbol{b}$ erzeugte Rundungsfehler akzeptabel ist (also Skalierung nach (R_1) oder für viele praktisch auftretende Systeme keine Skalierung). Berechne $\boldsymbol{x}$.

Stufe 2: Berechne $\Delta \boldsymbol{z}$ und $\boldsymbol{D}$ nach (R_4). Schätze $\|\boldsymbol{A}^{-1}\boldsymbol{D}^{-1}\|_\infty$ ab.

Wegen $\|\Delta \boldsymbol{z}\|_\infty = \|\boldsymbol{D}\,\Delta \boldsymbol{z}\|_\infty$ folgt dann aus (32)

$$\|\delta \boldsymbol{x}'_M\|_\infty \leqq \|\boldsymbol{A}^{-1}\boldsymbol{D}^{-1}\|_\infty \, \|\Delta \boldsymbol{z}\|_\infty \tag{45}$$

als scharfe Abschätzung für den linearisierten Fehler $\delta \boldsymbol{x}'_M$. Division der Schranke aus (45) durch $1 - \|\boldsymbol{A}^{-1}\boldsymbol{D}^{-1}\|_\infty \|\boldsymbol{D}\,\Delta \boldsymbol{A}\|_\infty$ ergibt eine Schranke für $\|\delta \boldsymbol{x}_M\|_\infty$, siehe Ü 4.1.6 für $p = \infty$.

(iii) Bei Verwendung des Gaußschen Algorithmus erfordert die erste Stufe $\sim n^3/3$ opms, und die zweite kann mit $O(n^2)$ opms — also billig — realisiert werden, siehe Abschnitt 5.4 für Details.

(iv) Wenn die a-posteriori-Skalierung gemäß (ii) nicht zweckmäßig erscheint, sollte a priori nach (R_2) oder (R_3) skaliert werden. Falls Schätzungen für die Größenordnung der $|x_i|$ bekannt sind, kann (R_4) näherungsweise a priori ausgeführt werden.

(v) Wenn die Störungen stochastischen Charakter haben, sollte der Fehlervektor $\Delta \boldsymbol{z}$ nicht gemäß (44), sondern nach der Vorschrift

$$(\Delta \boldsymbol{z})_i := \left\{ \sum_{j=1}^{n} (\Delta a_{ij} x_j)^2 + \Delta b_i^2 \right\}^{1/2} \tag{46}$$

festgelegt werden. □

Wir erwähnen abschließend, daß neben der Zeilenskalierung zusätzlich eine *Spaltenskalierung* gemäß

$$\boldsymbol{DAE}(\boldsymbol{E}^{-1}\boldsymbol{x}) = \boldsymbol{DAE}\hat{\boldsymbol{x}} = \boldsymbol{Db}, \qquad \hat{\boldsymbol{x}} = \boldsymbol{E}^{-1}\boldsymbol{x} \tag{47}$$

mit $\boldsymbol{E} = \operatorname{diag}(e_i)$, $e_i > 0$, durchgeführt werden kann. Dies entspricht einer Änderung des Maßstabes, in dem die Unbekannten x_j gemessen werden. Auf diese kompliziertere Skalierung soll hier jedoch nicht eingegangen werden, siehe B 4.3 für Literaturhinweise.

D. *Residualkriterien*

Wir kehren wieder zu den Schranken aus 4.1.3 zurück und bemerken, daß diese im Fall $\delta \boldsymbol{A} = \boldsymbol{O}$ durch untere Schranken ergänzt werden können: Wenn $\boldsymbol{x}$ bzw. $\boldsymbol{x} + \delta \boldsymbol{x}$ die Lösungen von

$$\boldsymbol{Ax} = \boldsymbol{b} \quad \text{bzw.} \quad \boldsymbol{A}(\boldsymbol{x} + \delta \boldsymbol{x}) = \boldsymbol{b} + \delta \boldsymbol{b}$$

bezeichnen, gilt

$$\|\delta \boldsymbol{b}\| / \|\boldsymbol{A}\| \leqq \|\delta \boldsymbol{x}\| \leqq \|\boldsymbol{A}^{-1}\| \, \|\delta \boldsymbol{b}\|; \tag{48}$$

man beachte $\|\delta \boldsymbol{b}\| = \|\boldsymbol{A}\delta \boldsymbol{x}\| \leqq \|\boldsymbol{A}\| \, \|\delta \boldsymbol{x}\|$. Wegen $\|\boldsymbol{b}\| / \|\boldsymbol{A}\| \leqq \|\boldsymbol{x}\| = \|\boldsymbol{A}^{-1}\boldsymbol{b}\| \leqq \|\boldsymbol{A}^{-1}\| \, \|\boldsymbol{b}\|$ folgt hieraus im Fall $\boldsymbol{b} \neq \boldsymbol{o}$

$$\frac{1}{\operatorname{cond}(\boldsymbol{A})} \frac{\|\delta \boldsymbol{b}\|}{\|\boldsymbol{b}\|} \leqq \frac{\|\delta \boldsymbol{x}\|}{\|\boldsymbol{x}\|} \leqq \operatorname{cond}(\boldsymbol{A}) \frac{\|\delta \boldsymbol{b}\|}{\|\boldsymbol{b}\|}. \tag{49}$$

Falls $\tilde{\boldsymbol{x}} := \boldsymbol{x} + \delta \boldsymbol{x}$ gesetzt wird, stellt $-\delta \boldsymbol{b} = \boldsymbol{b} - \boldsymbol{A}\tilde{\boldsymbol{x}} =: \boldsymbol{r}(\tilde{\boldsymbol{x}})$ gerade das *Residuum* von $\tilde{\boldsymbol{x}}$ bezüglich des Systems $\boldsymbol{Ax} = \boldsymbol{b}$ dar. In dieser Terminologie können (48), (49) wie folgt interpretiert werden.

4.1.17. *Aussage.* Es sei $\boldsymbol{x}$ die Lösung des regulären Gleichungssystems $\boldsymbol{Ax} = \boldsymbol{b}$. Dann gilt

$$\|\boldsymbol{b} - \boldsymbol{A}\tilde{\boldsymbol{x}}\| / \|\boldsymbol{A}\| \leqq \|\tilde{\boldsymbol{x}} - \boldsymbol{x}\| \leqq \|\boldsymbol{A}^{-1}\| \, \|\boldsymbol{b} - \boldsymbol{A}\tilde{\boldsymbol{x}}\| \tag{50}$$

für jedes $\tilde{\boldsymbol{x}}$, und im Fall $\boldsymbol{b} \neq \boldsymbol{o}$ ist

$$\frac{1}{\operatorname{cond}(\boldsymbol{A})} \frac{\|\boldsymbol{b} - \boldsymbol{A}\tilde{\boldsymbol{x}}\|}{\|\boldsymbol{b}\|} \leqq \frac{\|\tilde{\boldsymbol{x}} - \boldsymbol{x}\|}{\|\boldsymbol{x}\|} \leqq \operatorname{cond}(\boldsymbol{A}) \frac{\|\boldsymbol{b} - \boldsymbol{A}\boldsymbol{x}\|}{\|\boldsymbol{b}\|}. \tag{51}$$

Der relative Fehler einer Näherung $\tilde{\boldsymbol{x}}$ für $\boldsymbol{x}$ kann also um den Faktor $\operatorname{cond}(\boldsymbol{A})$ nach oben oder unten von der relativen Residuumsgröße abweichen. *Aus der Größe des Residuums, d. h. aus einer Einsetzprobe, kann daher i. allg. nicht auf die Größe*

des Fehlers geschlossen werden, sondern es ist eine Abschätzung von $\|A^{-1}\|$ bzw. cond (A) erforderlich. Im Gegensatz dazu folgt jedoch aus der Kleinheit des Residuums $r(\tilde{x})$, daß $\tilde{x}$ ein gegenüber $Ax = b$ nur wenig gestörtes System löst und umgekehrt:

4.1.18. *Satz.* Gegeben seien Zahlen $\Delta A, \Delta b \geqq 0$ und $A \in \mathbf{R}^{n,n}$, $b, \tilde{x} \in \mathbf{R}^n$. Dann existieren genau dann Störungen $\delta A, \delta b$, für die

$$(A + \delta A)\,\tilde{x} = b + \delta b \quad \text{mit} \quad \|\delta A\| \leqq \Delta A, \qquad \|\delta b\| \leqq \Delta b \tag{52}$$

gilt, wenn die Bedingung

$$\|r(\tilde{x})\| \leqq \|b - A\tilde{x}\| = \Delta b + \Delta A\,\|\tilde{x}\| \tag{53}$$

erfüllt ist.

Beweis. Das System (52) wird genau dann durch $\tilde{x}$ gelöst, wenn $r := b - A\tilde{x} = \delta A\,\tilde{x} - \delta b$ gilt. Mit den Bezeichnungen $\delta z := -r$ und $x := \tilde{x}$ kann Lemma 4.1.4 angewendet werden und besagt: Wenn δA und δb die Kugeln aus (52) durchlaufen, durchläuft r die Kugel $\|r\| \leqq \Delta b + \Delta A\,\|\tilde{x}\|$. Dies ist die zu beweisende Aussage. □

Mit $\Delta A := \varepsilon\,\|A\|$, $\Delta b := \varepsilon\,\|b\|$ läßt sich 4.1.18 in relativen statt absoluten Größen formulieren.

4.1.19. *Folgerung.* Für $\varepsilon \geqq 0$ und $A \in \mathbf{R}^{n,n}$, $b, \tilde{x} \in \mathbf{R}^n$ gibt es genau dann Störungen $\delta A, \delta b$ mit

$$(A + \delta A)\,\tilde{x} = b + \delta b \quad \text{und} \quad \|\delta A\| \leqq \varepsilon\,\|A\|, \qquad \|\delta b\| \leqq \varepsilon\,\|b\|, \tag{54}$$

wenn

$$\|r(\tilde{x})\| = \|b - A\tilde{x}\| \leqq \varepsilon\{\|b\| + \|A\|\,\|\tilde{x}\|\} \tag{55}$$

gilt.

Folgerung 4.1.19 besagt: *Die Lösungen $\tilde{x}$ der relativ wenig gestörten Gleichungen $(A + \delta A)\,\tilde{x} = b + \delta b$ sind durch die relative Kleinheit der entsprechenden Residuen $r(\tilde{x})$ gekennzeichnet.*

4.1.20. *Bemerkung.* (i) Die praktische Bedeutung der Residualkriterien 4.1.18 und 4.1.19 besteht darin, daß durch numerische Berechnung von $\|r(\tilde{x})\|$ überprüft werden kann, ob $\tilde{x}$ im Rahmen des Fehlerniveaus $\Delta A, \Delta b$ der Eingangsdaten als Lösung von $Ax = b$ akzeptiert werden kann. Dabei müssen die bei der Berechnung auftretenden Rundungsfehler mit berücksichtigt werden. Man vergleiche dazu Abschnitt 5.4.

(ii) Die zu vorgegebenem $r(\tilde{x})$ gemäß 4.1.18 bzw. 4.1.4 konstruierten Störungen sind i. allg. nicht eindeutig festgelegt. So kann z. B. für die 2-Norm in 4.1.4 auch $-\delta A := \lambda(\Delta A)\,H$ gesetzt werden, wobei H eine Spiegelungsmatrix mit $Hx = \|x\|_2 z$ ist. Eine solche Spiegelung läßt sich analog zu 3.3.4 konstruieren. Bei dieser Festlegung ist δA symmetrisch.

(iii) Zu 4.1.18 existiert das folgende elementweise Analogon: Gegeben seien $\Delta A \in \mathbf{R}^{n,n}$, $\Delta b \in \mathbf{R}^n$ mit $\Delta A \geqq O$, $\Delta b \geqq o$ und $A \in \mathbf{R}^{n,n}$, $b, \tilde{x} \in \mathbf{R}^n$. Dann gibt es

genau dann Störungen δA, δb mit

$$(A + \delta A)\,\tilde{x} = b + \delta b \quad \text{und} \quad |\delta A| \leqq \Delta A, \qquad |\delta b| \leqq \Delta b, \tag{56}$$

wenn

$$|r(\tilde{x})| = |b - A\tilde{x}| \leqq \Delta b + \Delta A\,|x| \tag{57}$$

gilt. Im Sinne von

$$|\delta A| \leqq \varepsilon\,|A|, \qquad |\delta b| \leqq \varepsilon\,|b| \tag{58}$$

elementweise kleine Störungen sind daher gleichwertig zu einem im Sinne von

$$|r(\tilde{x})| = |b - A\tilde{x}| \leqq \varepsilon\{|b| + |A|\,|\tilde{x}|\} \tag{59}$$

kleinen Residuum. □

Übungsaufgaben

Ü 4.1.1. Man zeige, daß unter der Voraussetzung $\|P\| < 1$

$$(I - P)^{-1} = \sum_{k=0}^{\infty} P^k = I + P + P^2 + \cdots + P^k + \cdots \tag{60}$$

gilt. Die unendliche Reihe (60) heißt *Neumannsche Reihe* von P.

Hinweis: Man beweise für die Teilsumme $S_l = \sum_{k=0}^{l} P^k$ die Identität

$$(I - P)^{-1} - S_l = (I - P)^{-1} P^{l+1}. \tag{61}$$

Ü 4.1.2. Man beweise unter Verwendung von (61) die Gültigkeit von

$$(A + \delta A)^{-1} = \sum_{k=0}^{l} (-A^{-1}\delta A)^k A^{-1} + O(\|\delta A\|^{l+1}) \tag{62}$$

und gebe eine strenge Schranke für das Restglied an. Was ergibt sich für $l = 1, 2, 3$?

Ü 4.1.3. Man überlege sich, daß 4.1.2 und 4.1.3 auch für $\varkappa := \|A^{-1}\delta A\|$ anstelle von $\varkappa = \|A^{-1}\|\,\|\delta A\|$ gültig sind. Die Bedingung $\|A^{-1}\delta A\| < 1$ ist schwächer als $\|A^{-1}\|\,\|\delta A\| < 1$, aber schwieriger zu überprüfen.

Ü 4.1.4. Es sei $A = U\Sigma V^T$ die Singulärwertzerlegung von A gemäß 1.2.14. Mit $c := U^T b$ kann dann die Lösung von $Ax = b$ in der Form $x = V\Sigma^{-1}c$ geschrieben werden. Man zeige, daß unter der Voraussetzung

$$(c_n)^2 \geqq 0.01(c_1^2 + \cdots + c_n^2)/n \tag{63}$$

in der 2-Norm die Abschätzung

$$K_b(A, b) \leqq 10\sqrt{n} \tag{64}$$

unabhängig von $\mathrm{cond}_2(A)$ gilt, vgl. (16)

Ü 4.1.5. Für das Gleichungssystem $\boldsymbol{A}\boldsymbol{x} = \boldsymbol{b}$ mit $\boldsymbol{A} = \begin{pmatrix} 1 & 1 \\ 1 & 0.99 \end{pmatrix}$, $\boldsymbol{b} = \begin{pmatrix} 1 \\ 1 \end{pmatrix}$ ist $\boldsymbol{A}^{-1} = \begin{pmatrix} -99 & 100 \\ 100 & -100 \end{pmatrix}$, $\boldsymbol{x} = \begin{pmatrix} 1 \\ 0 \end{pmatrix}$ und $\operatorname{cond}_\infty(\boldsymbol{A}) = 400$. Für die Störungen

(i) $\delta\boldsymbol{A} = 10^{-3}\begin{pmatrix} 1 & -1 \\ -1 & 1 \end{pmatrix}$, $\delta\boldsymbol{b} = 10^{-3}\begin{pmatrix} -1 \\ 1 \end{pmatrix}$, (ii) $\delta\boldsymbol{A} = 10^{-3}\begin{pmatrix} 1 & 1 \\ 1 & -1 \end{pmatrix}$, $\delta\boldsymbol{b} = 10^{-3}\begin{pmatrix} -1 \\ -1 \end{pmatrix}$

ist $\delta\boldsymbol{x}$ sowie $\|\delta\boldsymbol{x}\|_\infty$ zu berechnen und mit den Schranken (13) zu vergleichen. Man begründe die unterschiedlichen Resultate.

Ü 4.1.6. Für eine der Normen $p = 1$ oder $p = \infty$ gelte $\|\boldsymbol{A}^{-1}\|_p \, \|\delta\boldsymbol{A}\|_p < 1$. Mit den Bezeichnungen von 4.1.3 beweise man die Gültigkeit der Abschätzungen

$$|\delta\boldsymbol{x}| \leqq (\boldsymbol{I} - |\boldsymbol{A}^{-1}|\,|\delta\boldsymbol{A}|)^{-1}\,|\boldsymbol{A}^{-1}|\,\{|\delta\boldsymbol{A}|\,|\boldsymbol{x}| + |\delta\boldsymbol{b}|\} \tag{65}$$

und

$$\||\delta\boldsymbol{x}|\|_p \leqq \frac{\|\boldsymbol{A}^{-1}\|_p}{1 - \|\boldsymbol{A}^{-1}\|_p \, \|\delta\boldsymbol{A}\|_p} \, \||\delta\boldsymbol{A}|\,|\boldsymbol{x}| + |\delta\boldsymbol{b}|\|_p. \tag{66}$$

Hinweis: Man gehe von $\delta\boldsymbol{x} = (\boldsymbol{I} + \boldsymbol{A}^{-1}\delta\boldsymbol{A})^{-1}\boldsymbol{A}^{-1}\{-\delta\boldsymbol{A}\boldsymbol{x} + \delta\boldsymbol{b}\}$ aus und beachte $|(\boldsymbol{I} - \boldsymbol{P})^{-1}| \leqq (\boldsymbol{I} - |\boldsymbol{P}|)^{-1}$ sowie die Monotonie und Absolutheit der verwendeten Normen.

Ü 4.1.7. Man zeige am Beispiel

$$\boldsymbol{B} = \begin{pmatrix} 10 & 1 \\ 1 & 1 \end{pmatrix}, \qquad \Delta\boldsymbol{z} = \begin{pmatrix} 1 \\ 10 \end{pmatrix},$$

daß die Aussage (ii) aus Lemma 4.1.9 für nicht äquilibriertes $\Delta\boldsymbol{z}$ i. allg. falsch ist.

Ü 4.1.8. Man beweise die Äquivalenz aus Bemerkung 4.1.20 (iii) unter Verwendung von Lemma 4.1.8 analog zum Beweis von Satz 4.1.18.

Ü 4.1.9. Man zeige: Erfüllen die positiv definite Matrix $\boldsymbol{A} \in \mathbf{S}^{n,n}$ und die Störung $\delta\boldsymbol{A} \in \mathbf{S}^{n,n}$ die Bedingung (8), so ist $\boldsymbol{A} + \delta\boldsymbol{A}$ ebenfalls positiv definit.

Hinweis: Mit $\boldsymbol{A} + \delta\boldsymbol{A}$ ist nach 4.1.2 auch $\boldsymbol{B}(\lambda) := \boldsymbol{A} + \lambda\delta\boldsymbol{A}$ für alle $\lambda \in [0, 1]$ regulär, d. h., die Eigenwerte von $\boldsymbol{B}(\lambda)$ können für $\lambda \in [0, 1]$ das Vorzeichen nicht wechseln.

4.2. Prinzip der direkten Lösungsverfahren

Wir betrachten wieder das reguläre Gleichungssystem

$$\boldsymbol{A}\boldsymbol{x} = \boldsymbol{b}. \tag{1}$$

Das Prinzip der direkten Verfahren zur Lösung von (1) besteht darin, die Matrix $\boldsymbol{A}$ durch eine Folge regulärer elementarer Transformationsmatrizen auf obere Dreiecksform zu bringen. Wenn $\boldsymbol{T}$ das Produkt der Transformationsmatrizen bezeichnet, gilt also

$$\boldsymbol{T}\boldsymbol{A} = \boldsymbol{R}, \quad \boldsymbol{T} \text{ regulär}, \boldsymbol{R} \text{ obere Dreiecksmatrix}. \tag{2}$$

Im Kapitel 3 wurde bereits angedeutet, daß es aus Gründen der numerischen Stabilität in manchen Fällen nötig, in anderen Fällen zumindest günstig ist, die Matrix $\boldsymbol{A}$ dabei gewissen Zeilen- und gegebenenfalls auch Spaltenvertauschungen zu unterwerfen, d. h., statt (2) gilt

$$\boldsymbol{T}\bar{\boldsymbol{A}} = \boldsymbol{R} \quad \text{mit} \quad \bar{\boldsymbol{A}} = \boldsymbol{P}_z\boldsymbol{A}\boldsymbol{P}_s^{\mathrm{T}}, \qquad \boldsymbol{R} \text{ und } \boldsymbol{T} \text{ wie oben}. \tag{3}$$

Dabei sind $\boldsymbol{P}_z$, $\boldsymbol{P}_s$ zwei Permutationsmatrizen, von denen $\boldsymbol{P}_z$ auf die Zeilen und $\boldsymbol{P}_s$ auf die Spalten von $\boldsymbol{A}$ wirkt. Offensichtlich stellt (3) eine spezielle Äquivalenztransformation von $\boldsymbol{A}$ in $\boldsymbol{R}$ dar. Wenn (3) nach $\bar{\boldsymbol{A}}$ aufgelöst wird, ergibt sich

$$\bar{\boldsymbol{A}} = \boldsymbol{T}^{-1}\boldsymbol{R} = \boldsymbol{F}\boldsymbol{R} \quad \text{mit} \quad \boldsymbol{F} := \boldsymbol{T}^{-1}. \tag{4}$$

Praktisch wird $\boldsymbol{T}$ als Produkt von nichtorthogonalen elementaren Transformationsmatrizen, speziell als solches von LNT-Matrizen, oder als Produkt orthogonaler elementarer Transformationsmatrizen wie Householder-Spiegelungen oder Givens-Drehungen gewählt. Im ersten Fall spricht man vom *Gaußschen Algorithmus* und verwandten Verfahren, die Gegenstand der nachfolgenden Kapitel 5 und 6 sind. Als Produkt von LNT-Matrizen ist $\boldsymbol{T}$ ebenfalls eine untere Dreiecksmatrix, und dasselbe trifft für die Inverse $\boldsymbol{F} = \boldsymbol{T}^{-1}$ zu, die deshalb mit $\boldsymbol{L}$ bezeichnet wird. Es gilt dann

$$\bar{\boldsymbol{A}} = \boldsymbol{L}\boldsymbol{R}, \tag{5}$$

d. h., es liegt eine *Dreiecks-* oder *LR-Faktorisierung* vor.

Im zweiten Fall spricht man von *Orthogonalisierungsverfahren*, die in allgemeinerem Zusammenhang im Kapitel 10 behandelt werden. Als Produkt orthogonaler Matrizen ist $\boldsymbol{T}$ und damit auch $\boldsymbol{F}$ orthogonal, so daß $\boldsymbol{F}$ üblicherweise mit $\boldsymbol{Q}$ bezeichnet wird. Die Faktorisierung (4) ist dann von der Form

$$\bar{\boldsymbol{A}} = \boldsymbol{Q}\boldsymbol{R}, \tag{6}$$

sie heißt *orthogonale Dreiecks-* oder *QR-Faktorisierung*.

In der nach $\boldsymbol{A}$ aufgelösten Form lautet (3)

$$\boldsymbol{A} = \boldsymbol{P}_z^{\mathrm{T}}\boldsymbol{F}\boldsymbol{R}\boldsymbol{P}_s, \tag{7}$$

so daß das Ausgangssystem zu

$$\boldsymbol{P}_z^{\mathrm{T}}\boldsymbol{F}\boldsymbol{R}\boldsymbol{P}_s\boldsymbol{x} = \boldsymbol{b} \tag{8}$$

äquivalent ist. Mit den Bezeichnungen

$$\bar{\boldsymbol{x}} = \boldsymbol{P}_s\boldsymbol{x}, \qquad \bar{\boldsymbol{b}} = \boldsymbol{P}_z\boldsymbol{b} \tag{9}$$

läßt sich (8) in der Form $\boldsymbol{F}\boldsymbol{R}\bar{\boldsymbol{x}} = \bar{\boldsymbol{b}}$ bzw. als

$$\boldsymbol{F}\bar{\boldsymbol{c}} = \bar{\boldsymbol{b}} \quad \text{mit} \quad \boldsymbol{R}\bar{\boldsymbol{x}} = \bar{\boldsymbol{c}} \tag{10}$$

schreiben. Bei bekannter Faktorisierung (7) ist das Originalsystem (1) daher zu den beiden Gleichungssystemen (10) äquivalent, wobei $\bar{\boldsymbol{x}}$, $\bar{\boldsymbol{b}}$ über (9) mit $\boldsymbol{x}$, $\boldsymbol{b}$ verknüpft sind. Diese Äquivalenz ist der Ausgangspunkt für die später zu behandelnden direkten Lösungsverfahren und führt auf das nachfolgende algorithmische Schema.

4.2.1. *Basisalgorithmus zur Lösung von* $\boldsymbol{A}\boldsymbol{x} = \boldsymbol{b}$.

S1 (Faktorisierung von $\boldsymbol{A}$): Bestimme Permutationsmatrizen $\boldsymbol{P}_z$, $\boldsymbol{P}_s$, eine obere Dreiecksmatrix $\boldsymbol{R}$ und eine untere Dreiecksmatrix oder orthogonale Matrix $\boldsymbol{F}$ mit

$$\bar{\boldsymbol{A}} = \boldsymbol{P}_z\boldsymbol{A}\boldsymbol{P}_s^{\mathrm{T}} = \boldsymbol{F}\boldsymbol{R} \tag{11}$$

S2 (Berechnung der zur rechten Seite $\boldsymbol{b}$ gehörenden Lösung $\boldsymbol{x}$ von $\boldsymbol{Ax} = \boldsymbol{b}$):
S2.1: Bestimme $\bar{\boldsymbol{b}} := \boldsymbol{P}_z\boldsymbol{b}$
S2.2: Berechne $\bar{\boldsymbol{c}}$ als Lösung des linearen Gleichungssystems $\boldsymbol{F}\bar{\boldsymbol{c}} = \bar{\boldsymbol{b}}$
S2.3: Berechne $\bar{\boldsymbol{x}}$ als Lösung des linearen Gleichungssystems $\boldsymbol{R}\bar{\boldsymbol{x}} = \bar{\boldsymbol{c}}$
S2.4: Bestimme $\boldsymbol{x} := \boldsymbol{P}_s^T\bar{\boldsymbol{x}}$

4.2.2. *Bemerkung.* (i) Unter Verwendung der Zerlegung (11) geht das Gleichungssystem

$$\boldsymbol{A}^T\boldsymbol{y} = \boldsymbol{f}$$

in

$$\boldsymbol{P}_s^T\boldsymbol{R}^T\boldsymbol{F}^T\boldsymbol{P}_z\boldsymbol{y} = \boldsymbol{f}$$

über und kann analog zum Schritt S2 wie folgt gelöst werden:

S3 (Berechnung der zur rechten Seite $\boldsymbol{d}$ gehörenden Lösung $\boldsymbol{y}$ von $\boldsymbol{A}^T\boldsymbol{y} = \boldsymbol{d}$)
S3.1: Bestimme $\bar{\boldsymbol{f}} := \boldsymbol{P}_s\boldsymbol{f}$
S3.2: Berechne $\bar{\boldsymbol{g}}$ als Lösung von $\boldsymbol{R}^T\bar{\boldsymbol{g}} = \bar{\boldsymbol{f}}$
S3.3: Berechne $\bar{\boldsymbol{y}}$ als Lösung von $\boldsymbol{F}^T\bar{\boldsymbol{y}} = \bar{\boldsymbol{g}}$
S3.4: Bestimme $\boldsymbol{y} := \boldsymbol{P}_z^T\bar{\boldsymbol{y}}$

(ii) Offensichtlich hängt Schritt S1 nur von der Matrix $\boldsymbol{A}$ ab, während die Schritte S2 bzw. S3 nur auf die in S1 berechneten Faktoren und Permutationen und auf die jeweiligen rechten Seiten zurückgreifen. Wenn mehrere Gleichungssysteme mit derselben Koeffizientenmatrix $\boldsymbol{A}$ oder $\boldsymbol{A}^T$ zu lösen sind, braucht S1 nur einmal durchlaufen zu werden, und für jede rechte Seite genügt ein Durchlauf von S2 bzw. S3. Es ist daher zweckmäßig und heute üblich, den Faktorisierungsschritt S1 und den „Solver"-Schritt S2 bzw. S3 in getrennten Programmen zu realisieren. Man berücksichtige dabei, daß S1 für voll besetztes $\boldsymbol{A}$ i. allg. $O(n^3)$ opms, die Solver S2 bzw. S3 dagegen nur $O(n^2)$ opms erfordern.

(iii) Wenn die Faktorisierung S1 auf $\boldsymbol{A}^T$ statt $\boldsymbol{A}$ angewendet wird, ergibt sich die Zerlegung $\boldsymbol{P}_z\boldsymbol{A}^T\boldsymbol{P}_s^T = \boldsymbol{FR}$, also

$$\hat{\boldsymbol{P}}_z\boldsymbol{A}\hat{\boldsymbol{P}}_s^T = \hat{\boldsymbol{L}}\hat{\boldsymbol{F}} \tag{12}$$

mit $\hat{\boldsymbol{P}}_z := \boldsymbol{P}_s$, $\hat{\boldsymbol{P}}_s := \boldsymbol{P}_z$, der unteren Dreiecksmatrix $\hat{\boldsymbol{L}} := \boldsymbol{R}^T$ und der oberen Dreiecksmatrix oder orthogonalen Matrix $\hat{\boldsymbol{F}} := \boldsymbol{F}^T$. Im Fall $\hat{\boldsymbol{F}} = \hat{\boldsymbol{R}}$ liegt wieder eine $\boldsymbol{LR}$-Faktorisierung, im Fall $\hat{\boldsymbol{F}} = \hat{\boldsymbol{Q}}$ dagegen eine $\boldsymbol{LQ}$-Faktorisierung vor. Wir bemerken an dieser Stelle, daß (12) auch direkt durch zeilenweise Transformation von $\boldsymbol{A}$ auf untere Dreiecksform durch Anwendung transponierter elementarer Transformationsmatrizen von rechts auf $\boldsymbol{A}$ berechnet werden kann. Dies ist z. B. bei linearen Quadratmittelproblemen mit Gleichheitsnebenbedingungen und allgemein in der Optimierung zweckmäßig, wo unterbestimmte lineare Gleichungssysteme zu lösen sind. □

Auf die Realisierung der Permutationen sind wir bereits im Abschnitt 3.1 eingegangen. Die Lösung der in den Solvern S2, S3 auftretenden Gleichungssysteme mit orthogonalen oder Dreiecksmatrizen ist Gegenstand des folgenden Abschnitts.

Übungsaufgabe

Ü 4.2.1. Man beweise: Ist die reguläre Matrix $\boldsymbol{A}$ gemäß (11) faktorisiert, so sind $\boldsymbol{F}$ und $\boldsymbol{R}$ notwendig regulär, d. h., die Systeme in S2.2, S2.3 bzw. S3.2, S3.3 sind eindeutig lösbar.

4.3. Reguläre Gleichungssysteme einfacher Struktur

Unter regulären Gleichungssystemen einfacher Struktur verstehen wir Systeme mit regulären Dreiecksmatrizen oder mit orthogonalen Matrizen. Im folgenden befassen wir uns mit der Auflösung solcher Systeme und der Fehleranalyse der zugehörigen Algorithmen.

A. *Gleichungssysteme mit regulären Dreiecksmatrizen*

Reguläre Systeme mit Dreiecksmatrizen können in zwei zueinander symmetrischen Formen auftreten, nämlich als Systeme mit unteren Dreiecksmatrizen

$$\begin{aligned} l_{11}x_1 &= b_1, \\ l_{21}x_1 + l_{22}x_2 &= b_2, \qquad l_{ii} \neq 0 \quad (i = 1, \ldots, n), \\ &\vdots \\ l_{n1}x_1 + l_{n2}x_2 + \cdots + l_{nn}x_n &= b_n, \end{aligned} \tag{1}$$

oder als Systeme mit oberen Dreiecksmatrizen

$$\begin{aligned} r_{11}x_1 + r_{12}x_2 + \cdots + r_{1n}x_n &= c_1, \\ r_{22}x_2 + \cdots + r_{2n}x_n &= c_2, \qquad r_{ii} \neq 0 \quad (i = 1, \ldots, n). \\ &\vdots \\ r_{nn}x_n &= c_n, \end{aligned} \tag{2}$$

Die Eingangsdaten sind $\{l_{ij}, b_i\}$ bzw. $\{r_{ij}, c_i\}$, als Ausgangsdaten treten die Komponenten $\{x_j\}$ des Lösungsvektors auf.

Durch die Umindizierung $i \leftrightarrow n - i + 1$, $j \leftrightarrow n - j + 1$ geht (2) in (1) über und umgekehrt, so daß wir uns im folgenden auf untere Dreieckssysteme (1) beschränken, die wir kurz als

$$\boldsymbol{Lx} = \boldsymbol{b} \tag{3}$$

schreiben. Als häufig auftretende Sonderfälle berücksichtigen wir dabei

- untere *Einsdreiecksmatrizen*: $l_{ii} = 1 \quad (i = 1, \ldots, n)$,
- untere *Bidiagonalmatrizen*: $l_{ij} = 0 \quad (i > j + 1)$,
- Diagonalmatrizen: $l_{ij} = 0 \quad (i > j)$.

Die zugehörigen Besetztheitsmuster sind für $n = 4$

$$\begin{pmatrix} 1 & & & \\ \times & 1 & & \\ \times & \times & 1 & \\ \times & \times & \times & 1 \end{pmatrix}, \quad \begin{pmatrix} \times & & & \\ \times & \times & & \\ & \times & \times & \\ & & \times & \times \end{pmatrix}, \quad \begin{pmatrix} \times & & & \\ & \times & & \\ & & \times & \\ & & & \times \end{pmatrix}.$$

Auf Grund der speziellen Gestalt des Systems (1) lassen sich die Unbekannten nacheinander in der Reihenfolge $x_1, x_2, \ldots, x_n$ gemäß

$$x_1 := b_1/l_{11},$$
$$x_2 := (b_2 - l_{21} * x_1)/l_{22},$$
$$\vdots$$

berechnen, was auf den folgenden Algorithmus führt:

4.3.1. *Lösung eines Gleichungssystems mit regulärer unterer Dreiecksmatrix.*
Aufgabe: Für $\boldsymbol{L} = (l_{ij}) \in \mathfrak{R}^{n,n}$ mit $l_{ij} = 0$ $(i < j)$ und $l_{ii} \neq 0$ sowie $\boldsymbol{b} = (b_i) \in \mathfrak{R}^n$ ist $\boldsymbol{x} = (x_i) \in \mathfrak{R}^n$ als Lösung von $\boldsymbol{Lx} = \boldsymbol{b}$ zu berechnen.
Algorithmus:

$$\textbf{for } i := 1(1)n \textbf{ do } x_i := \left(b_i - \sum_{j=1}^{i-1} l_{ij} * x_j\right) / l_{ii} \tag{4}$$

Aufwand: $\sim n^2/2$ opms

Der Klammerausdruck in (4) soll in der Form

$$s := b_i$$
$$\textbf{for } j := 1(1)i - 1 \textbf{ do } s := s - l_{ij} * x_j \tag{5}$$

ausgewertet werden. Eine Summe bzw. Laufanweisung, bei welcher die obere Grenze kleiner als die untere ist, wird dabei hier und im folgenden als leer angesehen.

Offensichtlich kann 4.3.1 in situ realisiert werden, d. h., der Vektor $\boldsymbol{b}$ kann mit der Lösung $\boldsymbol{x}$ überspeichert werden.

Das Zeichen „$\sim$“ bedeutet bei Aufwandsangaben Gleichheit bis auf niedrigere Potenzen von n, d. h., $\boldsymbol{A} \sim K_p \boldsymbol{n}^p$ bedeutet $\boldsymbol{A} = K_p \boldsymbol{n}^p + K_{p-1} \boldsymbol{n}^{p-1} + \cdots + K_0$ mit gewissen Konstanten $K_0, \ldots, K_{p-1}$.

4.3.2. *Rundungsfehleranalyse.* Algorithmus 4.3.1 werde für die dort spezifizierten Eingangsdaten $\{\boldsymbol{L}, \boldsymbol{b}\}$ durchgeführt. Dann genügt die berechnete Lösung $\boldsymbol{x}$ der Gleichung

$$(\boldsymbol{L} + \delta\boldsymbol{L})\,\boldsymbol{x} = \boldsymbol{b} \tag{6}$$

mit einer unteren Dreiecksmatrix $\delta\boldsymbol{L}$, die komponentenweise klein ist im Sinne von

$$|(\delta\boldsymbol{L})_{ij}| \lessdot\!\!\cdot\; \nu j\, |l_{ij}| \qquad (i, j = 1, \ldots, n;\ i \geqq j). \tag{7}$$

Beweis. Der Zähler z von (4) läßt sich bei Berechnung nach (5) in der Form $z = \mathrm{fl}\,(z_0 - z_1 - \cdots - z_{i-1})$ mit $z_0 := b_i$ und $z_j := \mathrm{fl}\,(l_{ij} * x_j)$ $(j = 1, \ldots, i-1)$ schreiben. Aus 2.3.3 folgt dann unter Beachtung von $z_j = l_{ij}x_j(1+\delta_j)$ sofort $z = b_i(1+\varepsilon_1)\cdots(1+\varepsilon_{i-1}) - l_{i1}x_1(1+\varepsilon_1)\cdots(1+\varepsilon_{i-1})(1+\delta_1) - \cdots - l_{i,i-1}x_{i-1}(1+\varepsilon_{i-1})(1+\delta_{i-1})$ mit $|\varepsilon_j|, |\delta_j| \leqq \nu$. Berücksichtigung der Division durch l_{ii} führt auf

$$x_i = \frac{z}{l_{ii}(1+\mu_i)}, \qquad |\mu_i| \leqq \nu,$$

man vergleiche Ü 2.2.3. Einsetzen der Darstellung von z und Division von Zähler und Nenner durch $(1+\varepsilon_1)\cdots(1+\varepsilon_{i-1})$ ergibt schließlich

$$x_i = \frac{b_i - l_{i1}x_1(1+\delta_1) - \cdots - l_{i,i-1}x_{i-1}\dfrac{(1+\delta_{i-1})}{(1+\varepsilon_1)\cdots(1+\varepsilon_{i-2})}}{l_{ii}\dfrac{(1+\mu_i)}{(1+\varepsilon_1)\cdots(1+\varepsilon_{i-1})}}$$

$$= \left(b_i - \sum_{j=1}^{i-1} l_{ij}(1+\beta_{ij})\,x_j\right) \Big/ (l_{ii}(1+\beta_{ii}))$$

mit $|\beta_{ij}| \leqq \nu j$. Wenn $(\delta \boldsymbol{L})_{ij} := l_{ij}\beta_{ij}$ gesetzt wird, ist dies gerade (6) und (7). □

Man beachte, daß in (6) nur die Koeffizientenmatrix $\boldsymbol{L}$ mit den durch die Rundungsfehler hervorgerufenen Störungen belastet zu werden braucht.

Aus (7) folgt die kollektive Normabschätzung

$$\|\delta \boldsymbol{L}\|_\infty \leqq \nu n \,\|\boldsymbol{L}\|_\infty. \tag{8}$$

Die Ergebnisse von 4.3.2 besagen: *Algorithmus* 4.3.1 *ist numerisch gutartig in bezug auf* $\boldsymbol{L}$, *und zwar elementweise mit* $F = n$.

Der Fehler von $\boldsymbol{x}$ gegenüber der exakten Lösung $\boldsymbol{x}^* = \boldsymbol{L}^{-1}\boldsymbol{b}$ kann entsprechend 4.1.14 in der Form

$$\frac{\|\boldsymbol{x} - \boldsymbol{x}^*\|}{\|\boldsymbol{x}^*\|} \leqq \frac{\mathrm{cond}\,(\boldsymbol{L})}{1 - (\|\delta \boldsymbol{L}\|/\|\boldsymbol{L}\|)\,\mathrm{cond}\,(\boldsymbol{L})} \frac{\|\delta \boldsymbol{L}\|}{\|\boldsymbol{L}\|}$$

abgeschätzt werden. In der ∞-Norm gilt nach (8) $\|\delta \boldsymbol{L}\|_\infty / \|\boldsymbol{L}\|_\infty \leqq \nu n$, so daß im Fall $\nu n\,\mathrm{cond}_\infty\,(\boldsymbol{L}) \ll 1$ mit einem relativen Fehler in der Größenordnung von

$$\nu n\,\mathrm{cond}_\infty\,(\boldsymbol{L}), \qquad \mathrm{cond}_\infty\,(\boldsymbol{L}) := \|\boldsymbol{L}\|_\infty \,\|\boldsymbol{L}^{-1}\|_\infty \tag{9}$$

gerechnet werden muß. Praktische Erfahrungen zeigen jedoch, daß sich der tatsächliche relative Fehler fast immer in der Größenordnung

$$\nu \Phi(n)$$

bewegt, wobei $\Phi(n)$ eine schwach mit n wachsende Funktion ist. Der gemäß (9) theoretisch mögliche Verstärkungsfaktor $\mathrm{cond}_\infty\,(\boldsymbol{L})$ tritt also fast nie auf, was auf das günstige individuelle, durch (8) nicht erfaßte Fehlerverhalten zurückzuführen ist.

Die bisherigen Überlegungen zeigen, daß Algorithmus 4.3.1 hinsichtlich des Aufwandes (proportional zur Anzahl der Eingangsdaten), des Speicherbedarfes (in-situ-Realisierung möglich) und des Fehlerverhaltens (extrem gutartig) als fast optimal anzusehen ist.

4.3.3. *Bemerkung.* (i) Wenn $\boldsymbol{L}$ speziell eine Einsdreiecks-, Bidiagonal- oder Diagonalmatrix darstellt, vereinfacht sich in 4.3.1 die Laufanweisung (4) zu

$$x_i := b_i - \sum_{j=1}^{i-1} l_{ij} * x_j$$

bzw.

$$x_i := (b_i - l_{i,i-1} * x_{i-1})/l_{ii}$$

bzw.

$$x_i := b_i/l_{ii}.$$

Im Fall der Bidiagonal- bzw. Diagonalmatrix ist $\delta\boldsymbol{L}$ in 4.3.2 ebenfalls bidiagonal bzw. diagonal, und der Faktor j in (7) ist durch 2 bzw. 1 zu ersetzen.

(ii) Die Elemente von $\boldsymbol{L}$ werden in 4.3.1 bei Realisierung gemäß (4), (5) zeilenweise durchlaufen. Spaltenweise Abarbeitung ist ebenfalls möglich: Eine in-situ-Realisierung ist durch

for $j := 1(1)n$ **do**
  $b_j := b_j/l_{jj}$
  for $i := j + 1(1)n$ **do** $b_i := b_i - l_{ij} * b_j$ (10)

gegeben, wobei $\boldsymbol{b}$ mit $\boldsymbol{x}$ überspeichert wird.

(iii) Für das System (2) mit der oberen Dreiecksmatrix $\boldsymbol{R}$ lautet die zu (4), (5) analoge, zeilenweise arbeitende Vorschrift

for $i := n(-1)1$ **do**
  $s := c_i$
  for $j := i + 1(1)n$ **do** $s := s - r_{ij} * x_j$
  $x_i := s/r_{ii}$

Die Unbekannten werden also in der inversen Reihenfolge $x_n, x_{n-1}, \ldots, x_1$ berechnet. Eine spaltenweise in-situ-Version entsprechend (10) ist

for $j := n(-1)1$ **do**
  $c_j := c_j/r_{jj}$ (11)
  for $i := 1(1)j - 1$ **do** $c_i := c_i - r_{ij} * c_j$

Die übrigen Aussagen können sinngemäß übernommen werden, insbesondere gilt für das berechnete $\boldsymbol{x}$

$$(\boldsymbol{R} + \delta\boldsymbol{R})\,\boldsymbol{x} = \boldsymbol{c} \quad \text{mit} \quad |(\delta\boldsymbol{R})_{ij}| \leqq \nu(n - j + 1)\,|r_{ij}|.$$

B. *Gleichungssysteme mit orthogonalen Matrizen*

Wir betrachten jetzt ein Gleichungssystem

$$\boldsymbol{Q}\boldsymbol{x} = \boldsymbol{b} \tag{12}$$

mit einer orthogonalen Matrix $\boldsymbol{Q} \in \mathbf{R}^{n,n}$. Wegen der Orthogonalität von $\boldsymbol{Q}$ gilt

$$\boldsymbol{Q}^{\mathrm{T}}\boldsymbol{Q} = \boldsymbol{I} = \boldsymbol{Q}\boldsymbol{Q}^{\mathrm{T}}, \tag{13}$$

folglich

$$\boldsymbol{Q}^{-1} = \boldsymbol{Q}^{\mathrm{T}}, \tag{14}$$

vgl. Abschnitt 1.1.H, so daß die Lösung $\boldsymbol{x}^* = \boldsymbol{Q}^{-1}\boldsymbol{b}$ von (12) die Gestalt

$$\boldsymbol{x}^* = \boldsymbol{Q}^{\mathsf{T}}\boldsymbol{b}$$

hat. Damit liegt bereits ein Verfahren zur Lösung von (12) vor.

4.3.4. *Lösung eines Gleichungssystems mit einer orthogonalen Matrix.*
Aufgabe: Für $\boldsymbol{Q} = (q_{ij}) \in \mathfrak{R}^{n,n}$ und $\boldsymbol{b} = (b_i) \in \mathfrak{R}^n$ ist $\boldsymbol{x} = (x_i) \in \mathfrak{R}^n$ als Lösung von $\boldsymbol{Qx} = \boldsymbol{b}$ zu bestimmen.

Algorithmus: $\boldsymbol{x} := \boldsymbol{Q}^{\mathsf{T}}\boldsymbol{b}$, d. h. $x_i := \sum_{j=1}^{n} q_{ji} * b_j \quad (i = 1, \ldots, n)$ (15)

Aufwand: $\sim n^2$ opms

Bei der Berechnung von x treten einmal bei der Ausführung der Matrix-Vektor Multiplikation (15) Rundungsfehler auf, vgl. Aufgabe 2 aus Abschnitt 2.3.D. Zum anderen wird die tätsächlich benutzte Matrix $\boldsymbol{Q}$, deren Elemente Computerzahlen sind, i. allg. nicht orthogonal sein. Bereits die Computerdarstellung $\boldsymbol{Q} = \mathrm{rd}\,(\boldsymbol{Q}^*)$ einer exakt orthogonalen Matrix $\boldsymbol{Q}^* \in \mathbf{R}^{n,n}$ ist i. allg. nicht mehr orthogonal. In der numerischen linearen Algebra entsteht $\boldsymbol{Q}$ in der Regel als Ergebnis vorangegangener Rechnungen und erfüllt deshalb (13) und (14) erst recht nicht, so daß (15) auch bei fehlerfreier Realisierung nicht die exakte Lösung $\boldsymbol{x}^* = \boldsymbol{Q}^{-1}\boldsymbol{b}$ liefern würde. Es ist daher nötig, die Abweichung der Matrix $\boldsymbol{Q}$ von der Orthogonalität geeignet zu berücksichtigen, etwa durch das Residuum der Gleichungen (13).

4.3.5. *Rundungsfehleranalyse.* Algorithmus 4.3.4 werde für die dort spezifizierten Eingangsdaten $\{\boldsymbol{Q}, \boldsymbol{b}\}$ durchgeführt, wobei $\boldsymbol{Q}$ die abgeschwächte Orthogonalitätsbedingung

$$\boldsymbol{Q}^{\mathsf{T}}\boldsymbol{Q} = \boldsymbol{I} + \boldsymbol{G} \quad \text{mit} \quad \|\boldsymbol{G}\|_2 = \nu M_1 < 1 \tag{16}$$

oder

$$\boldsymbol{Q}\boldsymbol{Q}^{\mathsf{T}} = \boldsymbol{I} + \boldsymbol{H} \quad \text{mit} \quad \|\boldsymbol{H}\|_2 = \nu M_2 < 1 \tag{17}$$

erfüllen möge. Dann ist $\boldsymbol{Q}$ regulär, und das System $\boldsymbol{Qx} = \boldsymbol{b}$ besitzt die eindeutige Lösung $\boldsymbol{x}^* = \boldsymbol{Q}^{-1}\boldsymbol{b}$. Die nach 4.3.4 berechnete Lösung $\boldsymbol{x}$ genügt der Gleichung

$$\boldsymbol{Qx} = \boldsymbol{b} + \delta\boldsymbol{b}, \tag{18}$$

wobei

$$\frac{\|\delta\boldsymbol{b}\|_2}{\|\boldsymbol{b}\|_2} \leqq \nu \begin{cases} M_1 \sqrt{\dfrac{1 + \nu M_1}{1 - \nu M_1}} + (1 + \nu M_1)\, n^{3/2} & \text{im Fall (16)}, \\ M_2 + (1 + \nu M_2)\, n^{3/2} & \text{im Fall (17)}. \end{cases} \tag{19}$$

Der relative Fehler von $\boldsymbol{x}$ kann gemäß

$$\frac{\|\delta\boldsymbol{x}\|_2}{\|\boldsymbol{x}^*\|_2} \leqq \nu \begin{cases} M_1 + (1 + \nu M_1)\, n^{3/2} & \text{im Fall (16)}, \\ M_2 \sqrt{\dfrac{1 + \nu M_2}{1 - \nu M_2}} + (1 + \nu M_2)\, n^{3/2} & \text{im Fall (17)} \end{cases} \tag{20}$$

mit $\delta\boldsymbol{x} := \boldsymbol{x} - \boldsymbol{x}^*$ abgeschätzt werden.

Beweis. Die Matrix $\boldsymbol{Q}$ erfülle die Bedingung (16). Dann gilt

$$\|\boldsymbol{Q}^{\mathsf{T}}\boldsymbol{Q}\|_2 = \|\boldsymbol{I} + \boldsymbol{G}\|_2 \leqq \|\boldsymbol{I}\|_2 + \|\boldsymbol{G}\|_2 \leqq 1 + \nu M_1.$$

Nach Ü 1.2.9 ist $\|\boldsymbol{Q}^{\mathsf{T}}\boldsymbol{Q}\|_2 = \|\boldsymbol{Q}\|_2^2 = \|\boldsymbol{Q}^{\mathsf{T}}\|_2^2$, so daß

$$\|\boldsymbol{Q}\|_2 \leqq \sqrt{1 + \nu M_1} \tag{21}$$

folgt. Wegen $\|\boldsymbol{G}\|_2 < 1$ und 4.1.1 ist $\boldsymbol{I} + \boldsymbol{G}$, also $\boldsymbol{Q}^{\mathsf{T}}\boldsymbol{Q}$ regulär mit $\|(\boldsymbol{Q}^{\mathsf{T}}\boldsymbol{Q})^{-1}\| \leqq 1/(1 - \|\boldsymbol{G}\|_2) \leqq 1/(1 - \nu M_1)$. Dann ist auch $\boldsymbol{Q}$ regulär, siehe (1.2.28), und wie oben folgt $\|(\boldsymbol{Q}^{\mathsf{T}}\boldsymbol{Q})^{-1}\|_2 = \|\boldsymbol{Q}^{-1}\|_2^2$, mithin

$$\|\boldsymbol{Q}^{-1}\|_2 \leqq 1/\sqrt{1 - \nu M_1}. \tag{22}$$

Wir betrachten nun die durch (15) erzeugten Rundungsfehler. Nach (2.3.37) genügt der berechnete Vektor $\boldsymbol{x} := \boldsymbol{Q}^{\mathsf{T}}\boldsymbol{b}$ der Gleichung

$$\boldsymbol{x} = (\boldsymbol{Q}^{\mathsf{T}} + \delta\boldsymbol{Q}^{\mathsf{T}})\,\boldsymbol{b} \quad \text{mit} \quad |\delta\boldsymbol{Q}| \leqq \nu n\,|\boldsymbol{Q}|. \tag{23}$$

Analog zum Beweis von 2.3.15 folgt aus der letzten Abschätzung

$$\|\delta\boldsymbol{Q}\|_2 \leqq \nu n\sqrt{n}\,\|\boldsymbol{Q}\|_2 = \nu n^{3/2}\,\|\boldsymbol{Q}\|_2. \tag{24}$$

Multiplikation von (23) mit $\boldsymbol{Q}$ liefert

$$\boldsymbol{Q}\boldsymbol{x} = (\boldsymbol{Q}\boldsymbol{Q}^{\mathsf{T}} + \boldsymbol{Q}\delta\boldsymbol{Q}^{\mathsf{T}})\,\boldsymbol{b}.$$

Das erste $\boldsymbol{Q}$ auf der rechten Seite kann wegen (16) durch $\boldsymbol{Q} = \boldsymbol{Q}^{-\mathsf{T}} + \boldsymbol{Q}^{-\mathsf{T}}\boldsymbol{G}$ ersetzt werden, womit sich

$$\boldsymbol{Q}\boldsymbol{x} = (\boldsymbol{I} + \boldsymbol{Q}^{-\mathsf{T}}\boldsymbol{G}\boldsymbol{Q}^{\mathsf{T}} + \boldsymbol{Q}\delta\boldsymbol{Q}^{\mathsf{T}})\,\boldsymbol{b} = \boldsymbol{b} + \delta\boldsymbol{b}$$

mit $\delta\boldsymbol{b} := (\boldsymbol{Q}^{-\mathsf{T}}\boldsymbol{G}\boldsymbol{Q}^{\mathsf{T}} + \boldsymbol{Q}\delta\boldsymbol{Q}^{\mathsf{T}})\,\boldsymbol{b}$, wegen (21), (22) und (24) also

$$\begin{aligned}\|\delta\boldsymbol{b}\|_2 &\leqq (\|\boldsymbol{Q}^{-1}\|_2\,\|\boldsymbol{G}\|_2\,\|\boldsymbol{Q}\|_2 + \|\boldsymbol{Q}\|_2\,\|\delta\boldsymbol{Q}\|_2)\,\|\boldsymbol{b}\|_2 \\ &\leqq \nu\left\{M_1\sqrt{1 + \nu M_1}/\sqrt{1 - \nu M_1} + (1 + \nu M_1)\,n^{3/2}\right\}\|\boldsymbol{b}\|_2\end{aligned}$$

ergibt. Zur Abschätzung des relativen Fehlers von $\boldsymbol{x}$ setzen wir $\boldsymbol{b} = \boldsymbol{Q}\boldsymbol{x}^*$ in (23) ein und erhalten unter Berücksichtigung von (16)

$$\boldsymbol{x} = (\boldsymbol{Q}^{\mathsf{T}}\boldsymbol{Q} + \delta\boldsymbol{Q}^{\mathsf{T}}\boldsymbol{Q})\,\boldsymbol{x}^* = (\boldsymbol{I} + \boldsymbol{G} + \delta\boldsymbol{Q}^{\mathsf{T}}\boldsymbol{Q})\,\boldsymbol{x}^*,$$

folglich

$$\begin{aligned}\|\boldsymbol{x} - \boldsymbol{x}^*\|_2 &\leqq (\|\boldsymbol{G}\|_2 + \|\delta\boldsymbol{Q}\|_2\,\|\boldsymbol{Q}\|_2)\,\|\boldsymbol{x}^*\|_2 \\ &\leqq \nu\{M_1 + (1 + \nu M_1)\,n^{3/2}\}\,\|\boldsymbol{x}^*\|_2.\end{aligned}$$

Dabei sind (19) und (20) im Fall (16) bewiesen. Für den Fall (17) kann analog vorgegangen werden. □

Satz 4.3.5 läßt sich wie folgt interpretieren: *Wenn M_i nicht zu groß ist* — etwa in der Größenordnung n oder $n^{3/2}$ —, *also $\nu M_i \ll 1$ gilt und $\boldsymbol{Q}$ nur wenig von einer orthogonalen Matrix abweicht, ist Algorithmus* 4.3.5 *numerisch gutartig in bezug auf $\boldsymbol{b}$ mit $F = M_i + n^{3/2}$, und der relative Fehler von $\boldsymbol{x}$ ist durch νF beschränkt, also klein. Weicht $\boldsymbol{Q}$ dagegen stark von der Orthogonalität ab,* d. h., ist M_i sehr groß, *so können die Schranken* (19) *und* (20) *nicht mehr als akzeptabel angesehen werden. Der relative Fehler von $\boldsymbol{x}$ kann sehr groß werden, und Algorithmus* 4.3.4 *ist dann zur Lösung von $\boldsymbol{Q}\boldsymbol{x} = \boldsymbol{b}$ nicht mehr geeignet.*

4.3.6. *Bemerkung.* (i) Unter der schärferen Voraussetzung

$$\nu M_i \leqq 1/2 \tag{25}$$

können die Abschätzungen (19) und (20) durch die gröberen, aber einfacheren Schranken

$$\frac{\|\delta \boldsymbol{b}\|_2}{\|\boldsymbol{b}\|_2}, \frac{\|\delta \boldsymbol{x}\|_2}{\|\boldsymbol{x}\|_2} \leqq \nu \{\sqrt{3}\, M_i + 1.5 n^{3/2}\} \tag{26}$$

ersetzt werden, wobei $i = 1$ bzw. $i = 2$ für (16) bzw. (17) ist.

(ii) Wie weit eine Abweichung von der Orthogonalität bei der Matrix $\boldsymbol{Q}$ toleriert werden kann, hängt auch von der Genauigkeit der rechten Seite $\boldsymbol{b}$ ab. Wenn diese einen Meßfehler $\delta \boldsymbol{b}_M$ mit dem Fehlerniveau

$$\|\delta \boldsymbol{b}_M\|_2 / \|\boldsymbol{b}\|_2 = \nu L \leqq 1/2$$

aufweist, kann eine Abweichung νM_i derselben Größenordnung toleriert werden, da sie die Genauigkeit von $\boldsymbol{x}$ nicht wesentlich verschlechtert.

(iii) Wenn die durch νM_i charakterisierte Qualität der Orthogonalität von $\boldsymbol{Q}$ nicht a priori bekannt ist, müßte etwa $\|\boldsymbol{G}\|_2$ durch $\|\boldsymbol{Q}^T\boldsymbol{Q} - \boldsymbol{I}\|_F$ abgeschätzt werden, um den Vergleich von M_i mit L in (ii) durchführen zu können. Dies erfordert $O(n^3)$ Operationen. Unter der Voraussetzung (25), d. h. bei möglicherweise starker, aber nicht katastrophaler Abweichung von der Orthogonalität, gilt jedoch $\mathrm{cond}_2(\boldsymbol{Q}) \leqq \sqrt{3}$, so daß nach 4.1.17 die a-posteriori-Abschätzung

$$\frac{\|\boldsymbol{x}^* - \boldsymbol{x}\|_2}{\|\boldsymbol{x}^*\|_2} \leqq \sqrt{3}\, \frac{\|\boldsymbol{b} - \boldsymbol{Q}\boldsymbol{x}\|_2}{\|\boldsymbol{b}\|_2} \tag{27}$$

für das gemäß 4.3.4 berechnete $\boldsymbol{x}$ gilt. Aus dem nachträglich mit $O(n^2)$ Operationen berechneten Residuum $\|\boldsymbol{b} - \boldsymbol{Q}\boldsymbol{x}\|_2$ läßt sich also auf die Güte von $\boldsymbol{x}$ schließen. □

4.3.7. *Bemerkung.* In vielen Anwendungen — vgl. 10.2 — ist $\boldsymbol{Q}$ nicht explizit bekannt, sondern z. B. implizit als Produkt einer Folge elementarer orthogonaler Transformationsmatrizen definiert, so daß Algorithmus 4.3.4 nicht direkt angewendet werden kann. Die Abschätzungen aus 4.3.5 und 4.3.6 bleiben jedoch gültig, wenn $\boldsymbol{x} = \boldsymbol{Q}^T\boldsymbol{b}$ nach irgendeiner Vorschrift berechnet wird, für die

$$\boldsymbol{x} = \boldsymbol{Q}^T\boldsymbol{b} + \delta_1 \boldsymbol{x}, \qquad \|\delta_1 \boldsymbol{x}\|_2 \leqq \nu K \|\boldsymbol{Q}\|_2 \|\boldsymbol{b}\|_2 \tag{28}$$

mit einer (akzeptablen) Kumulationskonstanten K gilt. In den Abschätzungen ist dann nur $n^{3/2}$ durch K zu ersetzen. □

Übungsaufgaben

Ü 4.3.1. Man gebe eine zeilen- bzw. spaltenweise Realisierung von (15) zur Berechnung von $\boldsymbol{x} = \boldsymbol{Q}^T\boldsymbol{b}$ an.

Ü 4.3.2. Für die gemäß 4.3.4 berechnete Lösung $\boldsymbol{x}$ von $\boldsymbol{Q}\boldsymbol{x} = \boldsymbol{b}$ werde das Residuum $\boldsymbol{r} := \boldsymbol{b} - \boldsymbol{Q}\boldsymbol{x}$ ebenfalls in der durch ν charakterisierten Arithmetik berechnet. Man zeige, daß die

berechneten Werte unter der Voraussetzung (25) der Abschätzung

$$\|\boldsymbol{r}\|_2/\|\boldsymbol{b}\|_2 \leqq \nu\{\sqrt{3}\, M_i + 3n^{3/2}\}$$

genügen.

Ü 4.3.3. Wenn die Matrix $\boldsymbol{Q} = (\boldsymbol{q}^1, \ldots, \boldsymbol{q}^n)$ mit den Spalten $\boldsymbol{q}^j$ stärker von der Orthogonalität abweicht, kann nach BJÖRCK [67] das folgende Verfahren zur Lösung von $\boldsymbol{Qx} = \boldsymbol{b}$ verwendet werden:

$\boldsymbol{b}^1 := \boldsymbol{b}$
for $i := 1(1)n$ **do**
$\quad x_i := \boldsymbol{q}^{i\mathrm{T}}\boldsymbol{b}^i$ (29)
$\quad \boldsymbol{b}^{i+1} := \boldsymbol{b}^i - x_i * \boldsymbol{q}^i$

Als Teil eines speziellen Verfahrens — nämlich des modifizierten Gram-Schmidt-Verfahrens, siehe Abschnitt 10.1 — ergibt diese Vorschrift eine wesentlich höhere Genauigkeit als das Verfahren 4.3.4, d. h. als die Vorschrift

for $i := 1(1)n$ **do** $x_i := \boldsymbol{q}^{i\mathrm{T}}\boldsymbol{b}$. (30)

Man beweise, daß (29) und (30) bei exakter Realisierung und exakt orthogonalem $\boldsymbol{Q}$ identisch sind und charakterisiere den Rechenaufwand und Speicherbedarf von (29) im Vergleich zu (30) bei in-situ-Realisierung, d. h., wenn $\boldsymbol{b}$ mit den $\boldsymbol{b}^i$ überspeichert wird. Man überlege sich ferner, daß $\boldsymbol{b}^{n+1}$ für beliebiges $\boldsymbol{Q}$ bei exakter Realisierung von (29) das Residuum $\boldsymbol{b} - \boldsymbol{Qx}$ darstellt.

Bemerkungen zum Kapitel 4

B 4.1. Der Begriff der Kondition hat eine lange Tradition in der Mathematik. Im Zusammenhang mit der Matrixinversion und der Lösung linearer Gleichungssysteme ist die Zahl cond $(\boldsymbol{A})$ wohl erstmals von VON NEUMANN/GOLDSTINE [47] und TURING [48] verwendet worden. Die Ergebnisse der Störungstheorie sind in allen Lehrbüchern zu finden, allerdings meist mit der oberen Schranke cond $(\boldsymbol{A})$ für $K_{\boldsymbol{b}}(\boldsymbol{A}, \boldsymbol{b})$ in (4.1.14). Die so vergröberte Abschätzung ist dann nicht mehr für jedes $\{\boldsymbol{A}, \boldsymbol{b}\}$ scharf für $\|\delta\boldsymbol{A}\| \to 0$.

B 4.2. Die Schärfe der (linearisierten) Normschranken (4.1.13), (4.1.14) hat VAN DER SLUIS [70] untersucht, wo auch gewisse spaltenweise Konditionszahlen betrachtet werden. Individuelle Abschätzungen sind bei BAUER [66] zu finden, siehe auch MILLER [75]. Ähnliche Probleme behandelt SKEEL [81]. Der Frage, welche Norm zur Abschätzung gewählt werden sollte, wenden sich KAHAN [66] und FADDEEV/FADDEEVA [70] zu. Die LINPACK-Autoren DONGARRA et al. [79] empfehlen eine Zeilen- und Spaltenskalierung, welche die Elemente von $\Delta\boldsymbol{A}$ in etwa dieselbe Größenordnung bringt. Im Fall $\Delta\boldsymbol{b} = \boldsymbol{o}$ ist dies mit (R_4) aus 4.1.12 identisch, denn dann ist $\Delta\boldsymbol{z} = \Delta\boldsymbol{A}\,|\boldsymbol{x}|$ für jedes $\boldsymbol{x}$ äquilibriert und (30) realistisch. Die Festlegung von $\boldsymbol{D}$ mit dem Ziel der Äquilibrierung von $\Delta\boldsymbol{z}$ scheint neu zu sein.

B 4.3. Das Problem der Konditionsminimierung durch geeignete Skalierung ist in einer Reihe von Arbeiten behandelt worden; wir zitieren FORSYTH/STRAUS [55], BAUER [63, 66, 69] — dort ist Satz 4.1.14 zu finden —, HEINRICH [63], BUSINGER [68], VAN DER SLUIS [69] und FENNER/LOIZOU [77]. In diesen Arbeiten wird auch auf andere als ∞-Normen und simultane Zeilen- und Spaltenskalierung nach (4.1.17) eingegangen. Zum Beispiel wird cond $(\bar{\boldsymbol{A}})$ minimal über $\bar{\boldsymbol{A}} = \boldsymbol{DAE}$, wenn $\bar{\boldsymbol{A}}$ und $\bar{\boldsymbol{A}}^{-1}$ zeilenäquilibriert sind, allerdings ist dann die Bestimmung der optimalen Skalierungsfaktoren in $\boldsymbol{D}$ und $\boldsymbol{E}$ nicht mehr so einfach wie im Fall der Zeilenskalierung (BAUER [66]).

B 4.4. Die Abschnitte 4.2 und 4.3 folgen in ihren Grundideen den Darstellungen von WILKINSON [61, 63, 65].

5. Der Gaußsche Algorithmus

Der Gaußsche Algorithmus — lange vor GAUSS bekannt, siehe B 5.1 — war und ist *das* Standardverfahren zur Lösung linearer Gleichungssysteme. Seine Einfachheit erlaubt die Vermittlung und Anwendung bereits in der Schule, aber auch reale Gleichungssysteme mit Hunderten und Tausenden von Unbekannten werden heute nach Varianten dieses klassischen Verfahrens gelöst.

5.1. Die Grundform des Gaußschen Algorithmus

Zu lösen ist das lineare Gleichungssystem

$$\begin{aligned} a_{11}x_1 + \cdots + a_{1n}x_n &= b_1, \\ \vdots \qquad\qquad \vdots \qquad\quad &\vdots \\ a_{n1}x_1 + \cdots + a_{nn}x_n &= b_n, \end{aligned}$$

kurz

$$\boldsymbol{A}\boldsymbol{x} = \boldsymbol{b}, \tag{1}$$

mit der regulären Koeffizientenmatrix $\boldsymbol{A} = (a_{ij}) \in \mathbf{R}^{n,n}$ und der rechten Seite $\boldsymbol{b} = (b_i) \in \mathbf{R}^n$.

A. *Der Gaußsche Algorithmus als Verfahren der schrittweisen Elimination*

Das Prinzip des Gaußschen Algorithmus ist die schrittweise Elimination der Unbekannten durch geeignete Kombinationen der Gleichungen und damit die Überführung des Systems (1) in ein äquivalentes System

$$\boldsymbol{R}\boldsymbol{x} = \boldsymbol{c} \tag{2}$$

von oberer Dreiecksform.

Im ersten Schritt wird x_1 aus den Gleichungen $i = 2, \ldots, n$ eliminiert, indem ein Vielfaches der ersten Gleichung zur i-ten addiert wird, so daß dort der Koeffizient bei x_1 verschwindet. Dies ist gleichbedeutend mit dem Auflösen der ersten Gleichung nach x_1 und Einsetzen in die übrigen Gleichungen. Wenn die erste Spalte von $\boldsymbol{A}$ mit $\boldsymbol{a}$ bezeichnet wird und die restlichen zur Matrix $\boldsymbol{B}$ zusammengefaßt werden, liegt gerade der in 3.2.5 behandelte Eliminationsschritt mittels einer LNT-Matrix vor.

Für die Durchführbarkeit dieses Schrittes muß das *Pivotelement* a_{11} von 0 verschieden sein. Da auch für eine reguläre Matrix $a_{11} = 0$ gelten kann, wird ein nichtverschwindender Koeffizient $a_{s(1),1}$, $1 \leqq s(1) \leqq n$, in der ersten Spalte von $\boldsymbol{A}$ gesucht und durch Vertauschen der Zeilen 1 und $s(1)$ des Systems in die Position $(1, 1)$ gebracht, d. h., von dem Originalsystem $\boldsymbol{A}^{(1)}\boldsymbol{x} = \boldsymbol{b}^1$, $\boldsymbol{A}^{(1)} := \boldsymbol{A}$, $\boldsymbol{b}^1 := \boldsymbol{b}$, wird zu dem zeilenvertauschten System $\hat{\boldsymbol{A}}^{(1)}\boldsymbol{x} = \hat{\boldsymbol{b}}^1$ übergegangen, wobei jetzt

$$\hat{a}_{11}^{(1)} = a_{s(1),1}^{(1)} \neq 0$$

gilt. Dieser Übergang ist stets möglich, denn andernfalls hätte $\boldsymbol{A}$ nur Nullen in der ersten Spalte und wäre damit singulär im Widerspruch zur Voraussetzung.

Die gesuchte Transformation ist dann nach 3.2.5 durch die LNT-Matrix

$$\hat{\boldsymbol{L}}_1 := \boldsymbol{L}_1(-\hat{\boldsymbol{l}}^1), \qquad \hat{\boldsymbol{l}}^1 := (0, \hat{l}_{21}, \ldots, \hat{l}_{n1})^\mathrm{T}$$

mit

$$\hat{l}_{i1} := \hat{a}_{i1}^{(1)}/\hat{a}_{11}^{(1)} \qquad (i = 2, \ldots, n) \tag{3}$$

eindeutig festgelegt. Das transformierte System $\boldsymbol{A}^{(2)}\boldsymbol{x} = \boldsymbol{b}^2$ hat die Koeffizienten

$$\boldsymbol{A}^{(2)} = \hat{\boldsymbol{L}}_1\hat{\boldsymbol{A}}^{(1)} = \left(\begin{array}{c|ccc} a_{11}^{(2)} & a_{12}^{(2)} & \ldots & a_{1n}^{(2)} \\ \hline 0 & a_{22}^{(2)} & \ldots & a_{2n}^{(2)} \\ \vdots & \vdots & & \vdots \\ 0 & a_{n2}^{(2)} & \ldots & a_{nn}^{(2)} \end{array}\right) =: \left(\begin{array}{c|ccc} r_{11} & r_{12} & \ldots & r_{1n} \\ \hline & & & \\ \boldsymbol{o} & & \boldsymbol{M}^{(2)} & \\ & & & \end{array}\right),$$

$$\boldsymbol{b}^2 = \hat{\boldsymbol{L}}_1\hat{\boldsymbol{b}}^1 = \left(\begin{array}{c} b_1^{(2)} \\ \hline b_2^{(2)} \\ \vdots \\ b_n^{(2)} \end{array}\right) =: \left(\begin{array}{c} c_1 \\ \hline \\ \boldsymbol{d}^2 \\ \\ \end{array}\right),$$

wobei nur die Elemente von $\boldsymbol{M}^{(2)}$ und $\boldsymbol{d}^2$ neu berechnet zu werden brauchen:

$$a_{ij}^{(2)} := \hat{a}_{ij}^{(1)} - \hat{l}_{i1} * \hat{a}_{1j}^{(1)}, \qquad b_i^2 := \hat{b}_i^1 - \hat{l}_{i1} * \hat{b}_1^1 \qquad (i, j = 2, \ldots, n). \tag{4}$$

Wenn noch $\boldsymbol{M}^{(1)} := \boldsymbol{A}$, $\boldsymbol{d}^1 := \boldsymbol{b}$ gesetzt wird, ist also das Ausgangsproblem

$$\boldsymbol{M}^{(1)}\boldsymbol{x} = \boldsymbol{d}^1 \tag{5}$$

äquivalent zum System $\boldsymbol{A}^{(2)}\boldsymbol{x} = \boldsymbol{b}^2$, das in die erste Zeile

$$r_{11}x_1 + r_{12}x_2 + \cdots + r_{1n}x_n = c_1 \tag{6}$$

und das Restsystem

$$\boldsymbol{M}^{(2)}\boldsymbol{x}^2 = \boldsymbol{d}^2 \tag{7}$$

der Ordnung $n - 1$ für $\boldsymbol{x}^2 := (x_2, \ldots, x_n)^\mathrm{T}$ zerfällt. Da sich beim Übergang von $\boldsymbol{A}^{(1)}$ zu $\hat{\boldsymbol{A}}^{(1)}$ höchstens das Vorzeichen der Determinante ändert und $\hat{\boldsymbol{A}}^{(1)}$ dieselbe Determinante wie $\boldsymbol{A}^{(2)}$ besitzt — siehe Abschnitt 1.1.G —, gilt

$$\det(\boldsymbol{A}^{(2)}) = r_{11} \det(\boldsymbol{M}^{(2)}) = \pm\det(\boldsymbol{A}^{(1)}), \tag{8}$$

so daß $\boldsymbol{M}^{(2)}$ wie $\boldsymbol{A}^{(1)}$ regulär sein muß. Bei bekannter Lösung $\boldsymbol{x}^2$ von (7) kann die noch fehlende Komponente x_1 sofort aus (6) berechnet werden. Damit ist die Lösung des Gleichungssystems (5) der Ordnung n auf die Lösung des Gleichungssystems (7) der Ordnung $n - 1$ zurückgeführt worden.

Im zweiten Schritt wird derselbe Eliminationsprozeß mit $n - 1$ statt n auf das reduzierte System (7) angewendet usw. Auf diese Weise entsteht eine Folge von

Gleichungssystemen

$$\boldsymbol{A}^{(k+1)}\boldsymbol{x} = \boldsymbol{b}^{k+1} \qquad (k = 1, \ldots, n-1) \tag{9}$$

mit

$$\boldsymbol{A}^{(k)} = \left(\begin{array}{ccc|ccc} a_{11}^{(k)} & \cdots & a_{1,k-1}^{(k)} & a_{1k}^{(k)} & \cdots & a_{1n}^{(k)} \\ & \ddots & \vdots & \vdots & & \vdots \\ & & a_{k-1,k-1}^{(k)} & a_{k-1,k}^{(k)} & \cdots & a_{k-1,n}^{(k)} \\ \hline & & & a_{kk}^{(k)} & \cdots & a_{kn}^{(k)} \\ & \boldsymbol{O} & & \vdots & & \vdots \\ & & & a_{nk}^{(k)} & \cdots & a_{nn}^{(k)} \end{array}\right) =: \left(\begin{array}{c|c} \boldsymbol{O} & \boldsymbol{R}^{(k)} \\ \hline & \boldsymbol{M}^{(k)} \end{array}\right),$$

$$\boldsymbol{b}^{k} = \left(\begin{array}{c} b_1^k \\ \vdots \\ b_{k-1}^k \\ \hline b_k^k \\ \vdots \\ b_n^k \end{array}\right) =: \left(\begin{array}{c} \boldsymbol{c}^k \\ \hline \boldsymbol{d}^k \end{array}\right), \tag{10}$$

d. h., $\boldsymbol{A}^{(k)}$ ist bereits bis zur Spalte $k-1$ von oberer Dreiecksform. Im k-ten Eliminationsschritt werden dann die gewünschten Nullen in der k-ten Spalte von $\boldsymbol{A}^{(k+1)}$ erzeugt, indem der beschriebene „erste Schritt" auf das Restsystem $\boldsymbol{M}^{(k)}\boldsymbol{x}^k = \boldsymbol{d}^k$, $\boldsymbol{x}^k := (x_k, \ldots, x_n)^T$ der Ordnung $n-k+1$ angewendet wird. Sinngemäße Übertragung der Formeln (3) und (4) liefert die folgende Darstellung des Gesamtprozesses:

5.1.1. *Gaußscher Algorithmus.*
Initialisierung: $\boldsymbol{A}^{(1)} := \boldsymbol{A}$, $\boldsymbol{b}^1 := \boldsymbol{b}$
for $k := 1(1)n-1$ **do**
k-ter Eliminationsschritt:
Pivotsuche: Finde Index $s(k)$, $k \leqq s(k) \leqq n$, mit $a_{s(k),k}^{(k)} \neq 0$
Vertauschung: Festlegung von $\hat{\boldsymbol{A}}^{(k)}$, $\hat{\boldsymbol{b}}^k$ aus $\boldsymbol{A}^{(k)}$, $\boldsymbol{b}^k$ durch Vertauschen der Zeilen k und $s(k)$
Berechnung der Eliminationskoeffizienten:

$$\hat{l}_{ik} := \hat{a}_{ik}^{(k)} / \hat{a}_{kk}^{(k)} \qquad (i = k+1, \ldots, n)$$

Transformation der Matrix:

$$a_{ij}^{(k+1)} := \begin{cases} 0 & \text{für} \quad i = k+1, \ldots, n,\ j = k, \\ \hat{a}_{ij}^{(k)} - \hat{l}_{ik} * \hat{a}_{kj}^{(k)} & \text{für} \quad i, j = k+1, \ldots, n, \\ \hat{a}_{ij}^{(k)} & \text{sonst} \end{cases}$$

Transformation der rechten Seite:

$$b_i^{k+1} := \begin{cases} \hat{b}_i^k - \hat{l}_{ik} * \hat{b}_k^k & \text{für } i = k+1, \ldots, n, \\ \hat{b}_i^k & \text{sonst} \end{cases}$$

Aufwand: Berechnung von $\boldsymbol{R} = \boldsymbol{A}^{(n)}$: $\sim n^3/3$ opms;
Berechnung von $\boldsymbol{c} = \boldsymbol{b}^n$: $\sim n^2/2$ opms.

Zur Erläuterung soll das folgende einfache Beispiel dienen:

5.1.2. *Beispiel.* Gegeben ist das Gleichungssystem $\boldsymbol{Ax} = \boldsymbol{b}$ mit

$$\boldsymbol{A} = \boldsymbol{A}^{(1)} = \begin{pmatrix} 0 & 2 & -1 & -2 \\ 2 & -2 & 4 & -1 \\ 1 & 1 & 1 & 1 \\ -2 & 1 & -2 & 1 \end{pmatrix}, \quad \boldsymbol{b} = \boldsymbol{b}^1 = \begin{pmatrix} -7 \\ 6 \\ 10 \\ -2 \end{pmatrix}.$$

Erster Schritt ($k = 1$): $s(1) = 2$, d. h. Vertauschen der Zeilen 1 und 2

$$\hat{\boldsymbol{A}}^{(1)} = \begin{pmatrix} 2 & -2 & 4 & -1 \\ 0 & 2 & -1 & -2 \\ 1 & 1 & 1 & 1 \\ -2 & 1 & -2 & 1 \end{pmatrix}, \quad \hat{\boldsymbol{b}}^1 = \begin{pmatrix} 6 \\ -7 \\ 10 \\ -2 \end{pmatrix}.$$

Mit $\hat{l}_{21} = 0/2 = 0$, $\hat{l}_{31} = 1/2 = 0.5$, $\hat{l}_{41} = -2/2 = -1$ folgt

$$\boldsymbol{A}^{(2)} = \begin{pmatrix} 2 & -2 & 4 & -1 \\ \boxed{0} & 2 & -1 & -2 \\ 0 & 2 & -1 & 1.5 \\ 0 & -1 & 2 & 0 \end{pmatrix}, \quad \boldsymbol{b}^2 = \begin{pmatrix} 6 \\ -7 \\ 7 \\ 4 \end{pmatrix}.$$

Zweiter Schritt ($k = 2$): $s(2) = 2$, d. h. $\hat{\boldsymbol{A}}^{(2)} := \boldsymbol{A}^{(2)}$, $\hat{\boldsymbol{b}}^2 := \boldsymbol{b}^2$.
Mit $\hat{l}_{32} = 2/2 = 1$, $\hat{l}_{42} = -1/2 = -0.5$ folgt

$$\boldsymbol{A}^{(3)} = \begin{pmatrix} 2 & -2 & 4 & -1 \\ \boxed{0} & 2 & -1 & -2 \\ 0 & \boxed{0} & 0 & 3.5 \\ 0 & 0 & 1.5 & -1 \end{pmatrix}, \quad \boldsymbol{b}^3 = \begin{pmatrix} 6 \\ -7 \\ 14 \\ 0.5 \end{pmatrix}.$$

Dritter Schritt ($k = 3$): $s(3) = 4$, d. h. Vertauschen der Zeilen 3 und 4

$$\hat{\boldsymbol{A}}^{(3)} = \begin{pmatrix} 2 & -2 & 4 & -1 \\ 0 & 2 & -1 & -2 \\ 0 & 0 & 1.5 & -1 \\ 0 & 0 & 0 & 3.5 \end{pmatrix}, \quad \hat{\boldsymbol{b}}^3 = \begin{pmatrix} 6 \\ -7 \\ 0.5 \\ 14 \end{pmatrix}.$$

Mit $\hat{l}_{43} = 0/1.5 = 0$ ergibt sich $\boldsymbol{A}^{(4)} = \hat{\boldsymbol{A}}^{(3)}$, $\boldsymbol{b}^{(4)} = \hat{\boldsymbol{b}}^{(3)}$, also

$$\boldsymbol{R} = \boldsymbol{A}^{(4)} = \begin{pmatrix} 2 & -2 & 4 & -1 \\ \boxed{0} & 2 & -1 & -2 \\ 0 & \boxed{0} & 1.5 & -1 \\ 0 & 0 & 0 & 3.5 \end{pmatrix}, \quad \boldsymbol{c} = \boldsymbol{b}^4 = \begin{pmatrix} 6 \\ -7 \\ 0.5 \\ 14 \end{pmatrix}.$$

Der Lösungsvektor $\boldsymbol{x} = (1, 2, 3, 4)^T$ ergibt sich gemäß (4.3.11) aus dem Gleichungssystem $\boldsymbol{Rx} = \boldsymbol{c}$ mit der oberen Dreiecksmatrix $\boldsymbol{R}$. □

Das Beispiel macht deutlich, daß der Gaußsche Algorithmus in situ realisiert werden kann: $\boldsymbol{A}$ kann mit $\hat{\boldsymbol{A}}^{(k)}$ und $\boldsymbol{A}^{(k+1)}$, $\boldsymbol{b}$ mit $\hat{\boldsymbol{b}}^k$ und $\boldsymbol{b}^{k+1}$ überschrieben werden, und die Eliminationskoeffizienten $\hat{l}_{ik}$ können auf dem Platz der im k-ten Schritt erzeugten Nullen, d. h. in den Positionen (i, k) von $\boldsymbol{A}$ gespeichert werden

5.1.3. *Aussage.* Der Gaußsche Algorithmus 5.1.1 ist in exakter Arithmetik für jede reguläre Matrix $\boldsymbol{A} \in \mathbf{R}^{n,n}$ und jede rechte Seite $\boldsymbol{b} \in \mathbf{R}^n$ durchführbar und liefert nach $n-1$ Schritten das System $\boldsymbol{A}^{(n)}\boldsymbol{x} = \boldsymbol{b}^n$, wobei

$$\boldsymbol{A}^{(n)} = \begin{pmatrix} \hat{a}_{11}^{(1)} & \cdots & \hat{a}_{1n}^{(1)} \\ & \hat{a}_{22}^{(2)} & \cdots & \hat{a}_{2n}^{(2)} \\ & & \ddots & \\ & & \hat{a}_{n-1,n-1}^{(n-1)} & \hat{a}_{n-1,n}^{(n-1)} \\ & \boldsymbol{O} & & a_{nn}^{(n)} \end{pmatrix} =: \boldsymbol{R} = (r_{ij}), \qquad \boldsymbol{b}^n = \begin{pmatrix} \hat{b}_1^1 \\ \hat{b}_2^2 \\ \vdots \\ \hat{b}_{n-1}^{n-1} \\ b_n^n \end{pmatrix} =: \boldsymbol{c} = (c_i)$$

und

$$r_{kk} = a_{s(k),n}^{(k)} \neq 0 \qquad (k = 1, \ldots, n), \qquad s(n) := n, \tag{11}$$

ist.

Dies folgt unmittelbar aus den Überlegungen bei der Herleitung des Verfahrens.

B. *Matrixformulierung des Gaußschen Algorithmus*

Unter Verwendung der in den Abschnitten 3.1 und 3.2 eingeführten elementaren Transformationsmatrizen läßt sich der Gaußsche Algorithmus auch in Matrixnotation formulieren. Der im k-ten Eliminationsschritt vorzunehmende Übergang von $\boldsymbol{A}^{(k)}$ zu $\hat{\boldsymbol{A}}^{(k)}$ erfolgt gemäß

$$\hat{\boldsymbol{A}}^{(k)} := \boldsymbol{T}_{k,s(k)}\boldsymbol{A}^{(k)} \tag{12}$$

und der von $\hat{\boldsymbol{A}}^{(k)}$ zu $\boldsymbol{A}^{(k+1)}$ nach

$$\boldsymbol{A}^{(k+1)} = \left(\begin{array}{c|c} \boldsymbol{R}^{(k)} & \\ \hline \boldsymbol{O} & \hat{\boldsymbol{L}}_1^{(n-k+1)}\hat{\boldsymbol{M}}^{(k)} \end{array}\right) = \left(\begin{array}{c|c} \boldsymbol{I}^{(k-1)} & \boldsymbol{O} \\ \hline \boldsymbol{O} & \hat{\boldsymbol{L}}_1^{(n-k+1)} \end{array}\right) \hat{\boldsymbol{A}}^{(k)} =: \hat{\boldsymbol{L}}_k\hat{\boldsymbol{A}}^{(k)}. \tag{13}$$

Dabei ist $\hat{\boldsymbol{M}}^{(k)}$ diejenige Teilmatrix von $\hat{\boldsymbol{A}}^{(k)}$, der $\boldsymbol{M}^{(k)}$ in $\boldsymbol{A}^{(k)}$ entspricht, d. h., die durch Zeilenvertauschung aus dieser hervorgeht, und es wurde

$$\hat{\boldsymbol{L}}_1^{(n-k+1)} := \boldsymbol{L}_1^{(n-k+1)}(-\hat{\boldsymbol{l}}^{1,n-k+1}), \quad \hat{\boldsymbol{l}}^{1,n-k+1} := (0, \hat{l}_{k+1,k}, \ldots, \hat{l}_{nk})^{\mathsf{T}} \in \mathbf{R}^{n-k+1}$$

sowie

$$\hat{\boldsymbol{L}}_k := \boldsymbol{L}_k(-\hat{\boldsymbol{l}}^k), \qquad \hat{\boldsymbol{l}}^k := (0, \ldots, 0, \hat{l}_{k+1,k}, \ldots, \hat{l}_{n,k})^{\mathsf{T}} \in \mathbf{R}^n \tag{14}$$

gesetzt, vgl. (3.2.6).

Für den k-ten Eliminationsschritt ergibt sich damit

$$\boldsymbol{A}^{(k+1)} := \hat{\boldsymbol{L}}_k\boldsymbol{T}_{k,s(k)}\boldsymbol{A}^{(k)}, \qquad \boldsymbol{b}^k := \hat{\boldsymbol{L}}_k\boldsymbol{T}_{k,s(k)}\boldsymbol{b}^k \qquad (k = 1, \ldots, n-1), \tag{15}$$

und durch (14), (15) ist die gewünschte Matrixdarstellung des Verfahrens gegeben.

C. *Interpretation als Dreiecksfaktorisierung*

Hintereinanderschaltung der Transformationen (15) führt auf

$$\boldsymbol{A}^{(n)} = \hat{\boldsymbol{L}}_{n-1}\boldsymbol{T}_{n-1,s(n-1)}\boldsymbol{A}^{(n-1)} = \hat{\boldsymbol{L}}_{n-1}\boldsymbol{T}_{n-1,s(n-1)}\hat{\boldsymbol{L}}_{n-2}\boldsymbol{T}_{n-2,s(n-2)}\boldsymbol{A}^{(n-2)} = \cdots,$$

wegen $\boldsymbol{R} = \boldsymbol{A}^{(n)}$ und $\boldsymbol{A} = \boldsymbol{A}^{(1)}$ also auf

$$\boldsymbol{R} = \hat{\boldsymbol{L}}_{n-1}\boldsymbol{T}_{n-1,s(n-1)} \cdots \hat{\boldsymbol{L}}_2\boldsymbol{T}_{2,s(2)}\hat{\boldsymbol{L}}_1\boldsymbol{T}_{1,s(1)}\boldsymbol{A}. \tag{16}$$

Nach Ü 3.2.2 gilt wegen $1 < 2 \leqq s(2) \leqq n$

$$\boldsymbol{T}_{2,s(2)}\boldsymbol{L}_1(-\hat{\boldsymbol{l}}^1)\,\boldsymbol{T}_{1,s(1)} = \boldsymbol{L}_1(-\boldsymbol{T}_{2,s(2)}\hat{\boldsymbol{l}}^1)\,\boldsymbol{T}_{2,s(2)}\boldsymbol{T}_{1,s(1)},$$

wegen $1 < 2 < 3 \leqq s(3) \leqq n$ analog

$$\begin{aligned}
&[\boldsymbol{T}_{3,s(3)}\boldsymbol{L}_2(-\hat{\boldsymbol{l}}^2)]\,[\boldsymbol{T}_{2,s(2)}\boldsymbol{L}_1(-\hat{\boldsymbol{l}}^1)\,\boldsymbol{T}_{1,s(1)}]\\
&= [\boldsymbol{L}_2(-\boldsymbol{T}_{3,s(3)}\hat{\boldsymbol{l}}^2)\,\boldsymbol{T}_{3,s(3)}]\,[\boldsymbol{L}_1(-\boldsymbol{T}_{2,s(2)}\hat{\boldsymbol{l}}^1)\,\boldsymbol{T}_{2,s(2)}\boldsymbol{T}_{1,s(1)}]\\
&= \boldsymbol{L}_2(-\boldsymbol{T}_{3,s(3)}\hat{\boldsymbol{l}}^2)\,\boldsymbol{L}_1(-\boldsymbol{T}_{3,s(3)}\boldsymbol{T}_{2,s(2)}\hat{\boldsymbol{l}}^1)\,\boldsymbol{T}_{3,s(3)}\boldsymbol{T}_{2,s(2)}\boldsymbol{T}_{1,s(1)}
\end{aligned}$$

usw., d. h., die $\boldsymbol{T}_{k,s(k)}$ können nach rechts gebracht werden, wenn sie gleichzeitig auf die Argumentvektoren $\hat{\boldsymbol{l}}_j$ der $\boldsymbol{L}_j$ mit $j < k$ angewendet werden. Damit geht (16) in

$$\boldsymbol{R} = \boldsymbol{L}_{n-1}\boldsymbol{L}_{n-2} \cdots \boldsymbol{L}_2\boldsymbol{L}_1\boldsymbol{P}_0\boldsymbol{A} \tag{17}$$

über, wobei

$$\boldsymbol{L}_k := \boldsymbol{L}_k(-\boldsymbol{l}^k), \quad \boldsymbol{l}^k := (0, \ldots, 0, l_{k+1,k}, \ldots, l_{n,k})^{\mathsf{T}} := \boldsymbol{P}_k\hat{\boldsymbol{l}}^k \quad (k = 1, \ldots, n) \tag{18}$$

und

$$\boldsymbol{P}_k := \begin{cases} \boldsymbol{T}_{n-1,s(n-1)} \cdots \boldsymbol{T}_{k+1,s(k+1)} & \text{für } k = 0, 1, \ldots, n-2, \\ \boldsymbol{I} & \text{für } k = n-1 \end{cases} \tag{19}$$

ist. Nach (3.2.3) ist jedes $\boldsymbol{L}_k$ regulär mit der Inversen $\boldsymbol{L}_k^{-1} = \boldsymbol{L}_k(\boldsymbol{l}^k)$, so daß (17) mit $\boldsymbol{L}_1^{-1} \cdots \boldsymbol{L}_{n-1}^{-1}$ durchmultipliziert werden kann und auf

$$\boldsymbol{PA} = \boldsymbol{L}_1(\boldsymbol{l}^1) \cdots \boldsymbol{L}_{n-1}(\boldsymbol{l}^{n-1})\,\boldsymbol{R} = \boldsymbol{LR}, \qquad \boldsymbol{P} := \boldsymbol{P}_0 \tag{20}$$

führt. Das Produkt $\boldsymbol{L} = \boldsymbol{L}_1(\boldsymbol{l}^1) \cdots \boldsymbol{L}_{n-1}(\boldsymbol{l}^{n-1})$ ist nach 3.2.2 eine untere Einsdreiecksmatrix mit den Komponenten l_{ik} von $\boldsymbol{l}^k$ als Elementen der k-ten Spalte. Damit kann (20) wie folgt interpretiert werden:

5.1.4. *Satz.* Für jede reguläre Matrix $\boldsymbol{A} \in \mathbf{R}^{n,n}$ liefert der Gaußsche Algorithmus in exakter Arithmetik eine $\boldsymbol{LR}$-Faktorisierung von $\boldsymbol{A}$ gemäß

$$\boldsymbol{PA} = \boldsymbol{LR} = \begin{pmatrix} 1 & & & \boldsymbol{O} \\ l_{21} & 1 & & \\ l_{31} & l_{32} & 1 & \\ \vdots & \vdots & & \ddots \\ l_{n1} & l_{n2} & \ldots & 1 \end{pmatrix} \begin{pmatrix} r_{11} & r_{12} & \ldots & r_{1n} \\ & r_{22} & \ldots & r_{2n} \\ & & \ddots & \vdots \\ \boldsymbol{O} & & \ddots & \vdots \\ & & & r_{nn} \end{pmatrix} \tag{21}$$

mit $r_{kk} \neq 0$ $(k = 1, \ldots, n)$. Dabei ist $\boldsymbol{R} = \boldsymbol{A}^{(n)}$ die durch 5.1.1 erzeugte obere Dreiecksmatrix, $\boldsymbol{L}$ die gemäß (18) bis (20) festgelegte Matrix der Eliminationskoeffizienten und $\boldsymbol{P} = \boldsymbol{T}_{n-1,s(n-1)} \cdots \boldsymbol{T}_{1,s(1)}$ die durch die Vertauschungen bei der Pivotsuche charakterisierte Permutation der Zeilen von $\boldsymbol{A}$.

Die Faktorisierung (21) ist durch $\boldsymbol{P}$ und $\boldsymbol{A}$ eindeutig festgelegt.

Beweis. Es bleibt die Eindeutigkeit der Faktorisierung (21) zu zeigen: Es gelte also $\boldsymbol{PA} = \boldsymbol{LR} = \tilde{\boldsymbol{L}}\tilde{\boldsymbol{R}}$ mit einer weiteren unteren Einsdreiecksmatrix $\tilde{\boldsymbol{L}}$ und einer (notwendig regulären) oberen Dreiecksmatrix $\tilde{\boldsymbol{R}}$. Dann folgt $\tilde{\boldsymbol{L}}^{-1}\boldsymbol{L} = \tilde{\boldsymbol{R}}\boldsymbol{R}^{-1} =: \boldsymbol{D}$. Als Produkt unterer Dreiecksmatrizen ist $\tilde{\boldsymbol{L}}^{-1}\boldsymbol{L}$ eine untere Dreiecksmatrix, und ebenso ist $\tilde{\boldsymbol{R}}\boldsymbol{R}^{-1}$ als Produkt oberer Dreiecksmatrizen von oberer Dreiecksform. Dann muß $\boldsymbol{D}$ diagonal sein und wegen $\boldsymbol{L} = \tilde{\boldsymbol{L}}\boldsymbol{D}$ Diagonalelemente $d_i = 1$ besitzen. Also ist $\boldsymbol{D} = \boldsymbol{I}$ und folglich $\boldsymbol{L} = \tilde{\boldsymbol{L}}$, $\boldsymbol{R} = \tilde{\boldsymbol{R}}$. □

Die Permutationsmatrix $\boldsymbol{P}$ ist i. allg. natürlich nicht eindeutig festgelegt, denn in 5.1.1 gibt es i. allg. mehrere als Pivot geeignete Elemente $a^{(k)}_{s(k),k}$, $k \leqq s(k) \leqq n$.

5.1.5. *Bemerkung.* Die Vorschrift (18) besagt: Der Vektor $\boldsymbol{l}^k$ entsteht aus dem im k-ten Eliminationsschritt benutzten Vektor $\hat{\boldsymbol{l}}^k$ der Eliminationskoeffizienten, indem sämtliche nachfolgenden Vertauschungen $\boldsymbol{T}_{k+1,s(k+1)}, \ldots, \boldsymbol{T}_{n-1,s(n-1)}$ auf letzteren angewendet werden. Bei in-situ-Realisierung von 5.1.1 und Abspeichern der $\hat{l}_{ik}$ auf dem Platz von a_{ik} bedeutet dies, daß die vollen Zeilen von $\boldsymbol{A}$ einschließlich der bereits berechneten Eliminationskoeffizienten vertauscht werden.

5.1.6. *Beispiel.* Für die in 5.1.2 betrachtete Matrix ergibt sich mit $\boldsymbol{P} = \boldsymbol{T}_{3,s(3)}\boldsymbol{T}_{2,s(2)}\boldsymbol{T}_{1,s(1)} = \boldsymbol{T}_{34}\boldsymbol{T}_{12}$ die Faktorisierung

$$\boldsymbol{PA} = \begin{pmatrix} 2 & -2 & 4 & -1 \\ 0 & 2 & -1 & -2 \\ -2 & 1 & -2 & 1 \\ 1 & 1 & 1 & 1 \end{pmatrix} = \begin{pmatrix} 1 & & & \\ 0 & 1 & & \\ -1 & -0.5 & 1 & \\ 0.5 & 1 & 0 & 1 \end{pmatrix} \begin{pmatrix} 2 & -2 & 4 & -1 \\ & 2 & -1 & -2 \\ & & 1.5 & -1 \\ & & & 3.5 \end{pmatrix} = \boldsymbol{LR}. \; \square$$

Bei bekannter Dreiecksfaktorisierung (21) läßt sich die Lösung $\boldsymbol{x}$ des Systems (1) nach dem im Abschnitt 4.2 dargestellten Schema berechnen.

5.1.7. *Lösung von $\boldsymbol{Ax} = \boldsymbol{b}$ bei bekannter $\boldsymbol{LR}$-Faktorisierung.* Es sei $\boldsymbol{PA} = \boldsymbol{LR}$ die $\boldsymbol{LR}$-Faktorisierung der regulären Matrix $\boldsymbol{A} \in \mathbf{R}^{n,n}$ gemäß 5.1.4. Dann ergibt sich die Lösung $\boldsymbol{x}$ des Gleichungssystems $\boldsymbol{Ax} = \boldsymbol{b}$ für jedes $\boldsymbol{b} \in \mathbf{R}^n$ wie folgt:

S2.1: Bestimme $\bar{\boldsymbol{b}} = \boldsymbol{Pb}$

S2.2: Berechne $\boldsymbol{c}$ als Lösung von $\boldsymbol{Lc} = \bar{\boldsymbol{b}}$

S2.3: Berechne $\boldsymbol{x}$ als Lösung von $\boldsymbol{Rx} = \boldsymbol{c}$

Aufwand: $\sim n^2$ opms

5.1.8. *Bemerkung.* (i) Die Schritte S2.1 und S2.2 heißen *Vorwärtselimination*. Der dabei berechnete Vektor $\boldsymbol{c}$ ist identisch mit dem Vektor $\boldsymbol{c} = \boldsymbol{b}^n$, der sich nach dem Gaußschen Algorithmus 5.1.1 aus der rechten Seite $\boldsymbol{b}$ ergibt, siehe Ü 5.1.2. Der Schritt S2.3 heißt *Rücksubstitution*.

(ii) Zur Lösung von $\boldsymbol{A}^{\mathrm{T}}\boldsymbol{y} = \boldsymbol{d}$ unter Verwendung von (21) siehe Ü 5.1.3.

(iii) Wegen 5.1.4 und 5.1.7 ist es zweckmäßig, bei der Lösung von (1) in zwei

Schritten vorzugehen und diese auch in zwei getrennten Programmen zu realisieren, vgl. Bemerkung 4.2.2(ii):

Schritt 1: *Faktorisierung von* $\boldsymbol{A}$, d. h. Durchführung des Gaußschen Algorithmus nur für die Matrix $\boldsymbol{A}$ mit dem Ergebnis $\boldsymbol{P}, \boldsymbol{L}, \boldsymbol{R}$, wobei $\boldsymbol{P}$ durch die Zahlen $\{s(1), \ldots, s(n-1)\}$ repräsentiert wird.

Schritt 2: *Lösung des Systems* (1) bei vorliegender Faktorisierung von $\boldsymbol{A}$ nach 5.1.7.

Schritt 1 erfordert $\sim n^3/3$ opms, Schritt 2 dagegen nur $\sim n^2$ opms, so daß das beschriebene Vorgehen besonders günstig ist, wenn mehrere Gleichungssysteme mit derselben Koeffizientenmatrix zu lösen sind. Der aufwendige Faktorisierungsschritt braucht dann nur einmal durchgeführt zu werden.

(iv) Wenn $\boldsymbol{D} = \operatorname{diag}(r_{kk})$ gesetzt wird, kann (21) in der Form

$$\boldsymbol{PA} = \boldsymbol{LR} = \boldsymbol{LD}(\boldsymbol{D}^{-1}\boldsymbol{R}) = \boldsymbol{LDR}_1 = \boldsymbol{L}_1\boldsymbol{R}_1 \tag{22}$$

geschrieben werden, wobei $\boldsymbol{R}_1 := \boldsymbol{D}^{-1}\boldsymbol{R}$ eine obere Einsdreiecksmatrix und $\boldsymbol{L}_1 := \boldsymbol{LD}$ eine untere Dreiecksmatrix mit den Diagonalelementen r_{ii} bezeichnet. Die Faktorisierung $\boldsymbol{PA} = \boldsymbol{L}_1\boldsymbol{R}_1$ kann auch direkt berechnet werden, indem beim Gaußschen Algorithmus die Pivotzeile zu Beginn des k-ten Schrittes durch das Pivotelement dividiert wird. Dies ist dann auch für $k = n$ erforderlich. □

Aus der Faktorisierung (21) läßt sich in einfacher Weise die Determinante von $\boldsymbol{A}$ berechnen.

5.1.9. *Aussage.* Es sei $\boldsymbol{A} \in \mathbf{R}^{n,n}$ regulär mit der $\boldsymbol{LR}$-Faktorisierung (21) gemäß 5.1.4. Dann gilt

$$\det(\boldsymbol{A}) = \det(\boldsymbol{P}) \prod_{k=1}^{n} r_{kk} = (-1)^{\mu} \prod_{k=1}^{n} r_{kk}, \tag{23}$$

wobei μ die Anzahl der nichttrivialen Vertauschungen in $\boldsymbol{P} = \boldsymbol{T}_{n-1,s(n-1)} \cdots \boldsymbol{T}_{1,s(1)}$, d. h. die Anzahl der $k \in \{1, \ldots, n-1\}$ mit $k < s(k)$ bezeichnet.

Beweis. Aus $\boldsymbol{A} = \boldsymbol{P}^{\mathrm{T}}\boldsymbol{LR}$ folgt nach den Determinantenregeln aus 1.1.G sofort $\det(\boldsymbol{A}) = \det(\boldsymbol{P}) \det(\boldsymbol{L}) \det(\boldsymbol{R}) = \det(\boldsymbol{P}) \cdot 1 \cdot r_{11} \cdots r_{nn}$. Nun ist

$$\det(\boldsymbol{T}_{k,s(k)}) = \begin{cases} 1 & \text{für } k = s(k), \\ -1 & \text{für } k < s(k) \end{cases}$$

nach (3.1.8), womit sich sofort (23) ergibt. □

5.1.10. *Bemerkung.* Der Gaußsche Algorithmus 5.1.1 gestattet also für jede Matrix $\boldsymbol{A} \in \mathbf{R}^{n,n}$ die effektive Berechnung von $\det(\boldsymbol{A})$: Für reguläres $\boldsymbol{A}$ ist er bis zum $(n-1)$-ten Schritt durchführbar und liefert $\det(\boldsymbol{A})$ gemäß (23) als Produkt der Pivots mit eventueller Vorzeichenumkehr. Für singuläres $\boldsymbol{A}$ bricht der Gaußsche Algorithmus ab, weil in einem der Schritte $k = 1, \ldots, n-1$ kein von 0 verschiedenes Pivot gefunden werden kann, bzw. es ist $r_{nn} = 0$. In diesem Fall ist $\det(\boldsymbol{A}) = 0$ zu setzen. Bei der numerischen Realisierung der Formel (23) kann leicht Über- bzw. Unterlauf eintreten, so daß hier besondere Vorsichtsmaßnahmen erforderlich sind, siehe 5.2.5. □

Übungsaufgaben

Ü 5.1.1. Man zeige: Wenn $\boldsymbol{A}$ symmetrisch ist und $s(1) = 1$ gewählt werden kann, ist auch $\boldsymbol{M}^{(1)}$ symmetrisch.

Ü 5.1.2. Man weise nach, daß der in 5.1.7 berechnete Vektor $\boldsymbol{c} = \boldsymbol{L}^{-1}\boldsymbol{P}\boldsymbol{b}$ identisch mit $\boldsymbol{b}^n$ aus 5.1.1 ist.

Ü 5.1.3. Man überlege sich, daß das Gleichungssystem $\boldsymbol{A}^T\boldsymbol{y} = \boldsymbol{d}$ unter Verwendung der $\boldsymbol{LR}$-Faktorisierung (21) gemäß

S3.2: Berechne $\boldsymbol{g}$ als Lösung von $\boldsymbol{R}^T\boldsymbol{g} = \boldsymbol{d}$

S3.3: Berechne $\overline{\boldsymbol{y}}$ als Lösung von $\boldsymbol{L}^T\overline{\boldsymbol{y}} = \boldsymbol{g}$

S3.4: Bestimme $\boldsymbol{y} = \boldsymbol{P}^T\overline{\boldsymbol{y}}$

gelöst werden kann; vgl. auch Bemerkung 4.2.2 (i).

5.2. Pivotisierung

Die Bedingung $\hat{a}_{kk}^{(k)} \neq 0$ reicht i. allg. nicht aus, um die numerische Stabilität des Gaußschen Algorithmus zu garantieren, siehe Beispiel 2.3.20. Bei der Rundungsfehleranalyse im Abschnitt 5.3 werden wir sehen, daß numerische Stabilität vorliegt, wenn die Matrizen $\boldsymbol{M}^{(k)}$ gegenüber $\boldsymbol{M}^{(1)} = \boldsymbol{A}$ nicht beliebig groß werden. Für gewisse Matrizenklassen ist diese Bedingung auch ohne Pivotisierung — d. h. bei der Wahl des Diagonalelementes $a_{kk}^{(k)}$ als Pivot — erfüllt. Außer in diesen Spezialfällen muß jedoch durch geeignete Stabilisierungsmaßnahmen ein zu starkes Anwachsen der $\boldsymbol{M}^{(k)}$ verhindert werden. Dies wird durch verschiedene *Pivotisierungsstrategien* erreicht.

A. *Spaltenpivotisierung*

Wenn im k-ten Schritt das Pivot als ein betragsgrößtes Element der ersten Spalte von $\boldsymbol{M}^{(k)}$, also in der k-ten Spalte von $\boldsymbol{A}^{(k)}$ gesucht wird, spricht man von *Spaltenpivotisierung*. Es gilt dann

$$|\hat{a}_{kk}^{(k)}| = |a_{s(k),k}^{(k)}| = \max\{|a_{ik}^{(k)}| : i = k, \ldots, n\} \tag{1}$$

und somit

$$|\hat{l}_{ik}| = |\hat{a}_{ik}^{(k)}/\hat{a}_{kk}^{(k)}| \leqq 1 \qquad (i = k+1, \ldots, n), \tag{2}$$

d. h., die Transformationsmatrix $\hat{\boldsymbol{L}}_k = \boldsymbol{L}_k(-\hat{\boldsymbol{l}}^k)$ wird eine SLNT-Matrix im Sinne von 3.2.8. Aus (2) folgt

$$|a_{ij}^{(k+1)}| = |\hat{a}_{ij}^{(k)} - \hat{l}_{ik} * \hat{a}_{kj}^{(k)}| \leqq |\hat{a}_{ij}^{(k)}| + |\hat{a}_{kj}^{(k)}| \qquad (i, j = k+1, \ldots, n), \tag{3}$$

mit $\boldsymbol{M}^{(k)} = (\boldsymbol{m}^{k1}, \ldots, \boldsymbol{m}^{kn})$ also

$$\|\boldsymbol{m}^{k+1,j}\|_\infty \leqq 2\,\|\boldsymbol{m}^{k,j}\|_\infty \quad (j = k+1, \ldots, n) \quad \text{und} \quad \|\boldsymbol{M}^{(k+1)}\|_\infty \leqq 2\,\|\boldsymbol{M}^{(k)}\|_\infty.$$

Durch Induktion über k ergeben sich daraus die folgenden Abschätzungen:

5.2.1. *Aussage.* Im Fall der Spaltenpivotisierung (1) gilt für $k = 2, \ldots, n$

$$\|\boldsymbol{m}^{k,j}\|_\infty \leqq \gamma_k \|\boldsymbol{a}^j\|_\infty, \qquad \|\boldsymbol{M}^{(k)}\|_\infty \leqq \gamma_k \|\boldsymbol{A}\|_\infty \tag{4}$$

mit

$$\gamma_k := 2^{k-1}. \tag{5}$$

Dabei bezeichnet $\boldsymbol{a}^j$ die j-te Spalte von $\boldsymbol{A}$.

5.2.2. *Bemerkung.* (i) Es gibt Matrizen $\boldsymbol{A}$, für die 5.2.1 scharf ist, d. h., für die sich der ungünstige Verstärkungsfaktor 2^{k-1} tatsächlich einstellt, siehe Ü 5.2.2.

(ii) Umfangreiche Beobachtungen haben gezeigt, daß bei praktisch vorkommenden Gleichungssystemen die Abschätzung (4) mit

$$\gamma_k \leqq \gamma \approx 10 \tag{6}$$

erfüllt ist, d. h., das exponentielle Anwachsen tritt praktisch nicht auf. □

Auf Grund des im Sinne von (6) günstigen realen Verhaltens kann die Spaltenpivotisierung trotz der pessimistischen strengen Schranke (5) als i. allg. ausreichende Stabilisierung des Gaußschen Algorithmus empfohlen werden. Sie wird in fast allen derzeit bekannten Implementierungen verwendet. Wir geben deshalb eine computernahe Beschreibung der für die Lösung von $\boldsymbol{Ax} = \boldsymbol{b}$ nach dem Gaußschen Algorithmus mit Spaltenpivotisierung wesentlichen Teilschritte an und gehen dabei grundsätzlich spaltenweise vor, da zweidimensionale Felder z. B. in FORTRAN spaltenweise abgespeichert werden, vgl. Kapitel 16.

5.2.3. *Dreiecksfaktorisierung einer Matrix $\boldsymbol{A}$ mit Spaltenpivotisierung.*

Aufgabe: Für $\boldsymbol{A} = (a_{ij}) \in \mathfrak{R}^{n,n}$ ist nach dem Gaußschen Algorithmus mit Spaltenpivotisierung eine Dreiecksfaktorisierung $\boldsymbol{PA} = \boldsymbol{LR}$ gemäß 5.1.4 zu berechnen. Die Matrix $\boldsymbol{A}$ ist mit den signifikanten Elementen von $\boldsymbol{L}$ und $\boldsymbol{R}$ zu überspeichern; die Permutationsmatrix $\boldsymbol{P}$ ist durch die Zahlen $\{s(1), \ldots, s(n-1)\}$ zu charakterisieren. Falls in einem Eliminationsschritt k kein $r_{kk} \neq 0$ gefunden werden kann oder wenn $r_{nn} = 0$ ist, soll abgebrochen und $ie = -1$ gesetzt werden, andernfalls wird $ie = 0$ gesetzt, und $\boldsymbol{R}$ ist dann regulär.

Algorithmus:

$ie := 0$
for $k := 1(1)n - 1$ **do**
  Pivotsuche: $z := 0$
  for $i := k(1)n$ **do if** $|a_{ik}| > z$ **then** $[z := |a_{ik}|, s := i]$
  if $z = 0$ **then** $[ie := -1,$ **stop**$]$
  $s(k) := s$
  Vertauschung: **if** $k < s$ **then**
    for $j := 1(1)n$ **do** $[z := a_{kj}, a_{kj} := a_{sj}, a_{sj} := z]$
  Berechnung der l_{ik}: **for** $i := k + 1(1)n$ **do** $a_{ik} := a_{ik}/a_{kk}$
  Spaltenweise Berechnung von $\boldsymbol{M}^{(k+1)}$:
  for $j := k + 1(1)n$ **do**
    for $i := k + 1(1)n$ **do** $a_{ij} := a_{ij} - a_{ik} * a_{kj}$

Test von r_{nn}: **if** $a_{nn} = 0$ **then** [$ie := -1$, **stop**]
Aufwand: $\sim n^3/3$ opms, $\sim n$ S (für $s(1), \ldots, s(n-1)$)

Bei der Realisierung als SUBROUTINE ist die stop-Anweisung durch RETURN zu ersetzen.

5.2.4. *Lösung des Gleichungssystems* $\boldsymbol{Ax} = \boldsymbol{P}^T\boldsymbol{LR}\boldsymbol{x} = \boldsymbol{b}$.
Aufgabe: Für gegebene Dreiecksfaktorisierung $\boldsymbol{PA} = \boldsymbol{LR}$ gemäß 5.2.3 mit regulärem $\boldsymbol{R}$ und gegebene rechte Seite $\boldsymbol{b} = (b_i) \in \mathfrak{R}^n$ ist die Lösung $\boldsymbol{x} \in \mathfrak{R}^n$ des Gleichungssystems $\boldsymbol{Ax} = \boldsymbol{b}$ gemäß 5.1.7 zu berechnen und auf dem Platz von $\boldsymbol{b}$ zu speichern.

Algorithmus:
S2.1 ($\boldsymbol{b} := \boldsymbol{Pb}$): **for** $k := 1(1)n-1$ **do if** $k < s(k)$
then [$z := b_k$, $b_k := b_{s(k)}$, $b_{s(k)} := z$]
S2.2 ($\boldsymbol{b} := \boldsymbol{L}^{-1}\boldsymbol{b}$): **for** $k := 1(1)n-1$ **do**
for $i := k+1(1)n$ **do** $b_i := b_i - a_{ik} * b_k$
S2.3 ($\boldsymbol{b} := \boldsymbol{R}^{-1}\boldsymbol{b}$): **for** $k := n(-1)1$ **do**
$b_k := b_k/a_{kk}$
for $i := 1(1)k-1$ **do** $b_i := b_i - a_{ik} * b_k$
Aufwand: $\sim n^2$ opms

Vor Lösung von $\boldsymbol{Ax} = \boldsymbol{b}$ mittels 5.2.4 muß abgefragt werden, ob 5.2.3 mit $ie = 0$ ausgeführt worden ist.

Der Vollständigkeit halber geben wir noch ein Programm zur Berechnung der Determinante aus der Dreiecksfaktorisierung an, wobei wegen der Über- bzw. Unterlaufgefahr eine nichtstandardisierte Darstellung von $\det(\boldsymbol{A})$ verwendet wird. Der Algorithmus arbeitet auch, wenn 5.2.3 mit $ie = -1$ abgebrochen worden ist, vgl. Bemerkung 5.1.10.

5.2.5. *Berechnung von* $\det(\boldsymbol{A}) = \det(\boldsymbol{P}^T\boldsymbol{LR})$.
Aufgabe: Aus den Ausgangsdaten von 5.2.3 ist $\det(\boldsymbol{A})$ im Fall $ie = 0$ gemäß 5.1.9 in der Form $\det(\boldsymbol{A}) = d \times 16^e$ mit $1/16 \leqq |d| < 1$ und ganzzahligem Exponenten e zu berechnen; im Fall $ie = -1$, d. h. für numerisch singuläres $\boldsymbol{A}$, ist $d = 0$, $e := 0$ zu setzen.

Algorithmus:
$e := 0$
if $ie = 0$
then $d := 1$
for $k := 1(1)n$ **do**
$d := d * a_{kk}$, **if** $k < n$ **then** [**if** $k < s(k)$ **then** $d := -d$]
M1: **if** $|d| < 1/16$ **then** [$d := d * 16$, $e := e - 1$, **goto** M1]
M2: **if** $|d| \geqq 1$ **then** $d := d/16$, $e := e + 1$, **goto** M2]
else $d := 0$
Aufwand: $\sim n$ opm + Aufwand für Normalisierung von d

5.2.6. *Bemerkung.* (i) Statt der expliziten Dreiecksfaktorisierung (5.1.20) kann auch die implizite Darstellung (5.1.16) für den Lösungsprozeß verwendet werden. Die Elemente $\hat{l}_{ik}$ von $\hat{\boldsymbol{l}}^k$ werden dann im unteren Dreieck von $\boldsymbol{A}$ gespeichert und nicht mit vertauscht, vgl. 5.1.5. In der Vertauschungslaufanweisung von 5.2.3 muß dabei „$j := 1(1)n$" durch „$j := k(1)n$" ersetzt werden. Der Vektor $\boldsymbol{c} = \boldsymbol{b}^n$ wird dann gemäß (5.1.15) berechnet, d. h., in 5.2.4 sind S2.1 und S2.2 durch

S2.1 + 2: **for** $k := 1(1)n - 1$ **do**
$z := b_k,\ b_k := b_{s(k)},\ b_{s(k)} := z$
for $i := k + 1(1)n$ **do** $b_i := b_i - a_{ik} * b_k$

zu ersetzen. Dadurch wird der Vertauschungsaufwand in 5.2.3 etwa halbiert.

(ii) Durch Einführung eines Permutationsvektors $\{is(1), \ldots, is(n)\}$ kann die explizite physische Vertauschung in 5.2.3 und 5.2.4 umgangen werden, indem statt auf die i-te Zeile auf die Zeile $is(i)$ zurückgegriffen wird. Die Anfangsbelegung ist $is(i) := i$ $(i = 1, \ldots, n)$, und im k-ten Schritt werden $is(k)$ und $is(s)$ vertauscht, wenn $s = s(k)$ der Zeilenindex des Pivotelements ist. Dies ist zweckmäßig, wenn $\boldsymbol{A}$ extern gespeichert oder wie im Fall schwach besetzter Matrizen eine physische Umordnung aufwendig ist, vgl. Abschnitt 6.4. □

B. *Zeilenpivotisierung*

Bisher wurde das Pivot in der ersten Spalte von $\boldsymbol{M}^{(k)}$, d. h. in der k-ten Spalte von $\boldsymbol{A}^{(k)}$ gesucht. Analog ist natürlich auch eine Pivotsuche in der ersten Zeile von $\boldsymbol{M}^{(k)}$, d. h. in der k-ten Zeile von $\boldsymbol{A}^{(k)}$ möglich. Wenn dabei ein betragsgrößtes Element gemäß

$$|a^{(k)}_{k,\hat{s}(k)}| = \max\{|a^{(k)}_{kj}| : j = k, \ldots, n\} \tag{7}$$

als Pivot gewählt und durch Vertauschen der Spalten k und $\hat{s}(k)$ in die Position (k, k) gebracht wird, spricht man von *Zeilenpivotisierung*. In der Matrixformulierung ist dann (5.1.12) durch

$$\hat{\boldsymbol{A}}^{(k)} := \boldsymbol{A}^{(k)}\boldsymbol{T}_{k,\hat{s}(k)} \tag{8}$$

zu ersetzen. Analog zum Abschnitt 5.1.C ergibt sich damit eine Dreiecksfaktorisierung

$$\boldsymbol{A}\hat{\boldsymbol{P}}^{\mathsf{T}} = \hat{\boldsymbol{L}}\hat{\boldsymbol{R}} \tag{9}$$

mit

$$\hat{\boldsymbol{L}} = \boldsymbol{L}_1(\hat{\boldsymbol{l}}^1)\cdots\boldsymbol{L}_{n-1}(\hat{\boldsymbol{l}}^{n-1}), \qquad \hat{\boldsymbol{P}}^{\mathsf{T}} = \boldsymbol{T}_{1,\hat{s}(1)}\cdots\boldsymbol{T}_{n-1,\hat{s}(n-1)}. \tag{10}$$

Die Matrizen $\hat{\boldsymbol{L}}_k = \boldsymbol{L}_k(\hat{\boldsymbol{l}}^k)$ sind i. allg. keine SLNT-Matrizen. Aus (7) folgt jedoch

$$|\hat{a}^{(k)}_{kj}/\hat{a}^{(k)}_{kk}| \leqq 1 \qquad (j = k + 1, \ldots, n) \tag{11}$$

und damit

$$|a^{(k+1)}_{ij}| = \left|\hat{a}^{(k)}_{ij} - \frac{\hat{a}^{(k)}_{ik} * \hat{a}^{(k)}_{kj}}{\hat{a}^{(k)}_{kk}}\right| \leqq |\hat{a}^{(k)}_{ij}| + |\hat{a}^{(k)}_{ik}| \qquad (i, j = k + 1, \ldots, n) \tag{12}$$

als Analogon zur Abschätzung (3). Die Abschätzungen aus Teilabschnitt A gelten deshalb auch für die Zeilenpivotisierung, wenn sie zeilenweise und in der 1-Norm formuliert werden.

C. *Vollständige Pivotisierung*

Der Spielraum für die Wahl des Pivotelementes wird maximal ausgenutzt, wenn das Pivot als ein betragsgrößtes Element der gesamten Restmatrix $\boldsymbol{M}^{(k)}$ gewählt wird, d. h., wenn $s(k)$, $\hat{s}(k)$ gemäß

$$|a_{s(k),\hat{s}(k)}^{(k)}| = \max\{|a_{ij}^{(k)}| : i, j = k, \ldots, n\} \tag{13}$$

festgelegt und das Pivotelement $a_{s(k),\hat{s}(k)}^{(k)}$ durch Vertauschen der Zeilen k, $s(k)$ und der Spalten k, $\hat{s}(k)$ in die Position (k, k) gebracht wird. Es gilt dann

$$\hat{\boldsymbol{A}}^{(k)} = \boldsymbol{T}_{k,s(k)}\boldsymbol{A}^{(k)}\boldsymbol{T}_{k,\hat{s}(k)}, \tag{14}$$

was auf eine Dreiecksfaktorisierung des Typs

$$\boldsymbol{PA}\hat{\boldsymbol{P}}^{\mathrm{T}} = \bar{\boldsymbol{L}}\bar{\boldsymbol{R}} \tag{15}$$

führt. Wegen (13) gelten dabei sowohl die für die Spaltenpivotisierung wie auch die für die Zeilenpivotisierung gültigen Abschätzungen.

5.2.7. *Bemerkung.* (i) Bei vollständiger Pivotisierung gilt

$$\mu_k := \max\{|a_{ij}^{(k)}| : i, j = k, \ldots, n\} \leqq \varrho_k \mu_1 \qquad (k = 2, \ldots, n)$$

mit der Wachstumsschranke

$$\varrho_k := k^{1/2}[2 \cdot 3^{1/2} \cdots k^{1/(k-1)}]^{1/2} \leqq 1.8\, k^{0.25 \ln k}. \tag{16}$$

Im Gegensatz zu 5.2.1 ist diese Schranke jedoch nicht scharf, und bis jetzt sind keine reellen Matrizen bekannt geworden, für die

$$\mu_k \leqq k\mu_1$$

nicht gilt.

(ii) Trotz des günstigeren Stabilitätsverhaltens wird die vollständige Pivotisierung für reguläre Matrizen praktisch nur in Ausnahmefällen verwendet, da der Aufwand im Vergleich zur Spalten- oder Zeilenpivotisierung sehr hoch ist ($\sim n^3/3$ Lese- und Vergleichsoperationen gegenüber $\sim n^2/2$) und die einfacheren Strategien i. allg. zufriedenstellende Ergebnisse liefern. □

D. *Dreiecksfaktorisierung diagonaldominanter und symmetrischer definiter Matrizen*

Eine Matrix $\boldsymbol{A} = (a_{ij}) \in \mathbf{R}^{n,n}$ heißt *diagonaldominant* — genauer: *zeilendiagonaldominant* —, wenn

$$|a_{ii}| > \sum_{j=1, j\neq i}^{n} |a_{ij}| \qquad (i = 1, \ldots, n) \tag{17}$$

gilt. Nach 1.2.10 ist eine diagonaldominante Matrix regulär, denn wegen (17) kann $\lambda = 0$ in keinem Geršgorin-Kreis von $\boldsymbol{A}$ liegen.

5.2.8. *Aussage.* Für eine diagonaldominante Matrix $\boldsymbol{A}$ ist die Dreiecksfaktorisierung nach dem Gaußschen Algorithmus ohne Pivotisierung durchführbar, wobei folgende Aussagen gelten:

(i) Alle Matrizen $\boldsymbol{M}^{(k)}$ sind diagonaldominant,

(ii) $$\|\boldsymbol{M}^{(k)}\|_\infty \leqq \|\boldsymbol{M}^{(k-1)}\|_\infty \leqq \|\boldsymbol{A}\|_\infty \qquad (k = 2, \ldots, n). \tag{18}$$

Beweis. Es genügt, den Übergang von $\boldsymbol{A} = \boldsymbol{M}^{(1)}$ zu $\boldsymbol{M}^{(2)}$ zu betrachten. Wegen (17) ist $a_{11} \neq 0$ und kann als Pivot gewählt werden, womit sich

$$a_{ij}^{(2)} = a_{ij} - a_{i1} * a_{1j}/a_{11} \qquad (i, j = 2, \ldots, n) \tag{19}$$

ergibt. Nun ist (17) äquivalent zu

$$\tau_i := \sum_{j=1, j\neq i}^{n} |a_{ij}|/|a_{ii}| < 1 \qquad (i = 1, \ldots, n).$$

Aufsummieren von (19) liefert

$$\sum_{j=2}^{n} |a_{ij}^{(2)}| \leqq \sum_{j=2}^{n} |a_{ij}| + |a_{i1}| \sum_{j=2}^{n} |a_{1j}|/|a_{11}| = \sum_{j=2}^{n} |a_{ij}| + |a_{i1}|\,\tau_1 \leqq \sum_{j=1}^{n} |a_{ij}|, \tag{20}$$

also $\|\boldsymbol{M}^{(2)}\|_\infty \leqq \|\boldsymbol{M}^{(1)}\|_\infty$. Zum Nachweis der Diagonaldominanz von $\boldsymbol{M}^{(2)}$ folgern wir aus (19) mit $p := |a_{i1}|/|a_{ii}|$, $q := |a_{1i}|/|a_{11}|$, $p, q < 1$, zunächst $|a_{ii}^{(2)}| \geqq |a_{ii}| - |a_{i1}|\,|a_{1i}|/|a_{11}| = |a_{ii}|\,(1 - pq) > 0$ und damit unter Beachtung von (20)

$$\frac{\sum_{j=2, j\neq i}^{n} |a_{ij}^{(2)}|}{|a_{ii}^{(2)}|} \leqq \frac{\sum_{j=2, j\neq i}^{n} |a_{ij}| + |a_{i1}|\,(\tau_1 - q)}{|a_{ii}|\,(1 - pq)} = \frac{\tau_i - p + p(\tau_1 - q)}{1 - pq}$$

$$= \tau_i - p\,\frac{q(1 - \tau_i) + (1 - \tau_1)}{1 - pq} \leqq \tau_i < 1,$$

d. h., auch $\boldsymbol{M}^{(2)}$ ist diagonaldominant, und zwar mindestens so stark wie $\boldsymbol{M}^{(1)}$. □

Aussage 5.2.8 *besagt, daß für (zeilen-)diagonaldominante Matrizen ohne Pivotisierung gearbeitet werden kann.* Analoge Aussagen gelten auch für spaltendiagonaldominante Matrizen, d. h. für Matrizen $\boldsymbol{A}$, für die $\boldsymbol{A}^\mathsf{T}$ zeilendiagonaldominant ist.

Eine weitere Matrizenklasse, für deren Dreiecksfaktorisierungen keine Pivotisierung erforderlich ist, stellt die Klasse der symmetrischen definiten Matrizen dar. Wir beschränken uns auf symmetrische und positiv definite Matrizen, denn für negativ definites $\boldsymbol{A}$ ist $(-\boldsymbol{A})$ positiv definit.

5.2.9. *Aussage.* Für eine symmetrische und positiv definite Matrix $\boldsymbol{A}$ ist die Dreiecksfaktorisierung nach dem Gaußschen Algorithmus ohne Pivotisierung durchführbar, wobei folgende Eigenschaften gelten:

(i) Alle Matrizen $\boldsymbol{M}^{(k)}$ sind symmetrisch und positiv definit.

(ii) $$\|\boldsymbol{M}^{(k)}\|_F < \|\boldsymbol{M}^{(k-1)}\|_F \leqq \|\boldsymbol{A}\|_F \qquad (k = 2, \ldots, n). \tag{21}$$

Beweis. Nach Ü 1.1.6 ist $a_{11} > 0$ und kann als Pivot gewählt werden. Wir betrachten wieder den ersten Schritt und schreiben $\boldsymbol{A}$ bzw. $\boldsymbol{A}^{(2)}$ in der Gestalt

$$\boldsymbol{A} = \boldsymbol{A}^{(1)} = \left(\begin{array}{c|c} a_{11} & \boldsymbol{a}^\mathsf{T} \\ \hline \boldsymbol{a} & \boldsymbol{W} \end{array}\right), \qquad \boldsymbol{A}^{(2)} = \left(\begin{array}{c|c} a_{11} & \boldsymbol{a}^\mathsf{T} \\ \hline \boldsymbol{o} & \boldsymbol{M} \end{array}\right) \quad \text{mit} \quad \boldsymbol{M} = \boldsymbol{M}^{(2)}.$$

Nach 3.2.1 ist $\boldsymbol{L}_1 = \boldsymbol{L}_1(-\boldsymbol{l}) = \boldsymbol{I} - \boldsymbol{l}(\boldsymbol{e}^1)^\mathsf{T}$ mit $\boldsymbol{l} = \frac{1}{a_{11}} \begin{pmatrix} 0 \\ \boldsymbol{a} \end{pmatrix}$ und folglich

$$\boldsymbol{M} = \boldsymbol{W} - \frac{\boldsymbol{a}\boldsymbol{a}^\mathsf{T}}{a_{11}} = \boldsymbol{M}^\mathsf{T}. \tag{22}$$

Zum Nachweis der positiven Definitheit setzen wir $\boldsymbol{x} = \begin{pmatrix} \eta \\ \boldsymbol{y} \end{pmatrix}$, $\eta \in \mathbf{R}$, $\boldsymbol{y} \in \mathbf{R}^{n-1}$, und beachten, daß

$$\boldsymbol{x}^\mathsf{T}\boldsymbol{A}\boldsymbol{x} = a_{11}\eta^2 + 2(\boldsymbol{a}^\mathsf{T}\boldsymbol{y})\,\eta + \boldsymbol{y}^\mathsf{T}\boldsymbol{W}\boldsymbol{y} = a_{11}\left(\eta + \frac{\boldsymbol{a}^\mathsf{T}\boldsymbol{y}}{a_{11}}\right)^2 + \boldsymbol{y}^\mathsf{T}\boldsymbol{M}\boldsymbol{y} > 0 \tag{23}$$

für alle $\boldsymbol{x} \neq \boldsymbol{o}$ ist. Dann muß $\boldsymbol{y}^\mathsf{T}\boldsymbol{M}\boldsymbol{y} > 0$ für $\boldsymbol{y} \neq \boldsymbol{o}$ sein, d. h., $\boldsymbol{M}$ ist positiv definit. Wir betrachten jetzt eine Spiegelung $\boldsymbol{H} \in \mathbf{R}^{n-1,n-1}$ gemäß 3.3.4 mit $\boldsymbol{H}\boldsymbol{a} = \pm\|\boldsymbol{a}\|_2\,\boldsymbol{e}^1$. Für $\overline{\boldsymbol{M}} := \boldsymbol{H}\boldsymbol{M}\boldsymbol{H}^\mathsf{T}$ und $\overline{\boldsymbol{W}} := \boldsymbol{H}\boldsymbol{W}\boldsymbol{H}^\mathsf{T}$ folgt dann aus (22) $\overline{\boldsymbol{W}} = \overline{\boldsymbol{M}} + (\|\boldsymbol{a}\|_2^2/a_{11})\,\boldsymbol{e}^1\boldsymbol{e}^{1\mathsf{T}}$, also $\|\overline{\boldsymbol{W}}\|_F^2 = \|\overline{\boldsymbol{M}}_F^2\| + 2\overline{m}_{11}\|\boldsymbol{a}\|_2^2 a_{11} + \|\boldsymbol{a}\|_2^4/a_{11}^2 \geqq \|\overline{\boldsymbol{M}}\|_F^2$, denn $\overline{\boldsymbol{M}}$ ist wie $\boldsymbol{M}$ positiv definit, so daß auch $\overline{m}_{11} > 0$ gilt. Wegen der Orthogonalinvarianz der Frobeniusnorm erhält man schließlich

$$\|\boldsymbol{A}\|_F^2 > \|\boldsymbol{W}\|_F^2 = \|\overline{\boldsymbol{W}}\|_F^2 \geqq \|\overline{\boldsymbol{M}}\|_F^2 = \|\boldsymbol{M}\|_F^2. \quad \square$$

Aus dem Beweis ergibt sich sofort die folgende Aussage:

5.2.10. *Aussage.* Für eine symmetrische und positiv definite Matrix $\boldsymbol{A}$ liefert der Gaußsche Algorithmus ohne Pivotisierung eine Dreiecksfaktorisierung $\boldsymbol{A} = \boldsymbol{L}\boldsymbol{R}$ mit $\boldsymbol{L}^\mathsf{T} = \boldsymbol{D}^{-1}\boldsymbol{R}$, d. h., es gilt

$$\boldsymbol{A} = \boldsymbol{L}\boldsymbol{D}\boldsymbol{L}^\mathsf{T} \tag{24}$$

mit $\boldsymbol{D} = \operatorname{diag}(d_i)$, $d_i = r_{ii} > 0$ $(i = 1, \ldots, n)$, und einer unteren Einsdreiecksmatrix $\boldsymbol{L}$.

Auf die Berechnung von Dreiecksfaktorisierungen symmetrischer Matrizen und die dabei mögliche Halbierung von Speicherplatz und Rechenaufwand gehen wir im Abschnitt 6.1 nochmals im Detail ein.

Übungsaufgaben

Ü 5.2.1. Man beweise im Fall der Spaltenpivotisierung die Gültigkeit von $\|\boldsymbol{M}^{(k)}\|_\infty \leqq k\,\|\boldsymbol{A}\|_\infty$ im Fall einer oberen Hessenbergmatrix $\boldsymbol{A}$, d. h. einer Matrix mit $a_{ij} = 0$ für $i > j + 1$, und von $\|\boldsymbol{M}^{(k)}\|_\infty \leqq 2\,\|\boldsymbol{A}\|_\infty$ im Fall einer Tridiagonalmatrix $\boldsymbol{A}$.

Ü 5.2.2. Man überlege sich, daß für die Matrix

$$\boldsymbol{A} = \left(\begin{array}{cccc|c} 1 & & & & a \\ -1 & 1 & & & a \\ -1 & -1 & 1 & & a \\ \vdots & \vdots & & \ddots & \vdots \\ & & & 1 & a \\ \hline -1 & -1 & \ldots & -1 & a \end{array}\right)$$

im Fall der Spaltenpivotisierung

$$\|\boldsymbol{A}\|_\infty = |a| + n - 1, \qquad \|\boldsymbol{M}^{(k)}\|_\infty = 2^{k-1}\,|a| + n - k$$

gilt, so daß der Quotient $\|\boldsymbol{M}^{(k)}\|_\infty / \|\boldsymbol{A}\|_\infty$ der Schranke $\gamma_k = 2^{k-1}$ aus 5.2.1 für $|a| \to \infty$ beliebig nahe kommt.

Ü 5.2.3. Es sei $\boldsymbol{A}$ regulär. Man zeige: Wenn $\boldsymbol{PA}^\mathsf{T} = \boldsymbol{LR}$ die nach dem Gaußschen Algorithmus mit Spaltenpivotisierung berechnete Dreiecksfaktorisierung von $\boldsymbol{A}^\mathsf{T}$ und $\boldsymbol{A}\hat{\boldsymbol{P}}^\mathsf{T} = \hat{\boldsymbol{L}}\hat{\boldsymbol{R}}$ die mittels Zeilenpivotisierung berechnete Faktorisierung bezeichnen, gilt

$$\hat{\boldsymbol{P}} = \boldsymbol{P}, \qquad \hat{\boldsymbol{L}} = \boldsymbol{R}^\mathsf{T}\boldsymbol{D}^{-1}, \qquad \hat{\boldsymbol{R}} = \boldsymbol{D}\boldsymbol{L}^\mathsf{T} \quad \text{mit} \quad \boldsymbol{D} = \operatorname{diag}(r_{ii}),$$

sofern bei der Zeilenpivotisierung dieselbe Strategie wie bei der Spaltenpivotisierung verwendet wird.

Ü 5.2.4. Man gebe zu 5.2.3, 5.2.4 und 5.2.5 analoge Algorithmen an, die mit impliziten Vertauschungen entsprechend Bemerkung 5.2.6(ii) arbeiten.

Ü 5.2.5. Es ist zu zeigen, daß der Gaußsche Algorithmus genau dann ohne Pivotisierung mit regulärem $\boldsymbol{R}$ durchgeführt werden kann, wenn die Hauptabschnittsdeterminanten $\det(\boldsymbol{A}_k)$ von $\boldsymbol{A}$ mit

$$\boldsymbol{A}_k := \begin{pmatrix} a_{11} \dots a_{1k} \\ \vdots \quad\; \vdots \\ a_{k1} \dots a_{kk} \end{pmatrix}$$

der Bedingung

$$\det(\boldsymbol{A}_k) \neq 0 \qquad (k = 1, \dots, n) \tag{25}$$

genügen, und daß bei Erfülltsein von (25)

$$r_{kk} = \det(\boldsymbol{A}_k)/\det(\boldsymbol{A}_{k-1}) \qquad (k = 1, \dots, n) \tag{26}$$

mit $\det(\boldsymbol{A}_0) := 1$ gilt.

Ü 5.2.6. Man spricht von *Diagonalpivotisierung*, wenn das Pivotelement als betragsmaximales Diagonalelement von $\boldsymbol{M}^{(k)}$ gesucht wird, d. h., wenn $a^{(k)}_{s(k),s(k)}$ mit

$$|a^{(k)}_{s(k),s(k)}| = \max\{|a^{(k)}_{ii}| : k \leqq i \leqq n\} \tag{27}$$

als Pivot gewählt wird. Man zeige, daß für diagonaldominante und symmetrische definite Matrizen die Diagonalpivotisierung identisch mit der vollständigen Pivotisierung ist.

Ü 5.2.7. Man überlege sich: Die Matrix $\boldsymbol{A} \in \mathbf{S}^{n,n}$ ist positiv definit genau dann, wenn sie sich in der Form $\boldsymbol{A} = \boldsymbol{L}\boldsymbol{D}\boldsymbol{L}^\mathsf{T}$ mit einer unteren Einsdreiecksmatrix $\boldsymbol{L}$ und $\boldsymbol{D} = \operatorname{diag}(d_i)$, $d_i > 0$, faktorisieren läßt.

5.3. Rundungsfehleranalyse

Die bei der Pivotisierung vorgenommenen Zeilen- bzw. Spaltenvertauschungen entsprechen einer Umnumerierung der Gleichungen bzw. Unbekannten. Für die folgenden Überlegungen kann daher o. B. d. A. vorausgesetzt werden, daß diese Vertauschungen bereits vor Ausführung des Gaußschen Algorithmus vorgenommen worden sind und während der Elimination stets $s(k) = k = \hat{s}(k)$ gewählt wird. Damit werden die Vertauschungen und das Zeichen „$\hat{}$“ für die vertauschten Größen überflüssig.

A. *Fehleranalyse der Dreiecksfaktorisierung*

Mit der obigen Vereinbarung lautet der k-te Eliminationsschritt

$$\boldsymbol{A}^{(k+1)} = \left(\begin{array}{c|c} \boldsymbol{I}^{(k-1)} & \boldsymbol{O} \\ \hline \boldsymbol{O} & \boldsymbol{L}_1^{(n-k\cdot 1)} \end{array}\right)\left(\begin{array}{c|c} & \boldsymbol{R}^{(k)} \\ \hline \boldsymbol{O} & \boldsymbol{M}^{(k)} \end{array}\right) = \left(\begin{array}{c|c} & \boldsymbol{R}^{(k)} \\ \hline \boldsymbol{O} & \boldsymbol{L}_1^{(n-k+1)}\boldsymbol{M}^{(k)} \end{array}\right), \tag{1}$$

vgl. (5.1.13). Es genügt daher, die Rundungsfehler bei der Transformation

$$\boldsymbol{M}^{(k)} \to \overline{\boldsymbol{M}}^{(k)} := \left(\begin{array}{c|c} a_{kk}^{(k)} & a_{k,k+1}^{(k)} \dots a_{kn}^{(k)} \\ \hline \boldsymbol{o} & \boldsymbol{M}^{(k+1)} \end{array}\right) = \boldsymbol{L}_1^{(n-k\ 1)}\boldsymbol{M}^{(k)}$$

zu untersuchen, wobei $\boldsymbol{L}_1^{(n-k\cdot 1)} := \boldsymbol{L}_1(-\boldsymbol{l}^{1,n-k+1})$ ist. Dies ist gerade der in 3.2.5 betrachtete Eliminationsschritt. Der berechnete Vektor $\boldsymbol{l}^{1,n-k+1} = (0, l_{k+1,k}, \dots, l_{nk})^T$ und die berechnete Matrix $\overline{\boldsymbol{M}}^{(k)}$ bzw. $\boldsymbol{M}^{(k+1)}$ genügen dann wegen 3.2.6 der Gleichung

$$\overline{\boldsymbol{M}}^{(k)} = \boldsymbol{L}_1^{(n-k\cdot 1)}(\boldsymbol{M}^{(k)} + \delta\boldsymbol{M}^{(k)}), \tag{2}$$

und für die Störungsmatrix $\delta\boldsymbol{M}^{(k)}$ gilt

$$|\delta\boldsymbol{M}^{(k)}| \leqq \nu\left\{|\boldsymbol{M}^{(k)}| + 2\left(\begin{array}{c|c} 0 & \boldsymbol{o}^T \\ \hline \boldsymbol{o} & |\boldsymbol{M}^{(k+1)}| \end{array}\right)\right\}, \tag{3}$$

also

$$\|\delta\boldsymbol{M}^{(k)}\| \leqq \nu(\|\boldsymbol{M}^{(k)}\| + 2\,\|\boldsymbol{M}^{(k+1)}\|), \tag{4}$$

vgl. Formel (3.2.15) und deren Herleitung.

Einsetzen von (2) in (1) führt auf

$$\boldsymbol{A}^{(k+1)} = \boldsymbol{L}_k(\boldsymbol{A}^{(k)} + \delta\boldsymbol{A}^{(k)}) \tag{5}$$

mit

$$\boldsymbol{L}_k = \boldsymbol{L}_k(-\boldsymbol{l}^k), \qquad \boldsymbol{l}^k = (0, \dots, 0, l_{k+1,k}, \dots, l_{nk})^T$$

und

$$\delta\boldsymbol{A}^{(k)} = \left(\begin{array}{c|c} \boldsymbol{O} & \boldsymbol{O} \\ \hline \boldsymbol{O} & \delta\boldsymbol{M}^{(k)} \end{array}\right), \qquad \|\delta\boldsymbol{A}^{(k)}\| = \|\delta\boldsymbol{M}^{(k)}\|. \tag{6}$$

Für den berechneten Faktor $\boldsymbol{R}$ der Dreiecksfaktorisierung gilt dann

$$\begin{aligned} \boldsymbol{R} = \boldsymbol{A}^{(n)} &= \boldsymbol{L}_{n-1}(\boldsymbol{A}^{(n-1)} + \delta\boldsymbol{A}^{(n-1)}) \\ &= \boldsymbol{L}_{n-1}[\boldsymbol{L}_{n-2}(\boldsymbol{A}^{(n-2)} + \delta\boldsymbol{A}^{(n-2)}) + \delta\boldsymbol{A}^{(n-1)}] \\ &= \boldsymbol{L}_{n-1}\boldsymbol{L}_{n-2}[\boldsymbol{A}^{(n-2)} + \delta\boldsymbol{A}^{(n-2)} + \delta\boldsymbol{A}^{(n-1)}], \end{aligned}$$

denn wegen der speziellen Blockgestalt (6) von $\delta\boldsymbol{A}^{(k)}$ ist

$$\boldsymbol{L}_{n-2}\delta\boldsymbol{A}^{(n-1)} = [\boldsymbol{I} - \boldsymbol{l}^{n-2}(\boldsymbol{e}^{n-2})^T]\,\delta\boldsymbol{A}^{(n-1)} = \delta\boldsymbol{A}^{(n-1)}.$$

Sukzessive Fortsetzung dieser Überlegungen führt auf

$$\boldsymbol{R} = \boldsymbol{L}_{n-1}\boldsymbol{L}_{n-2} \cdots \boldsymbol{L}_2\boldsymbol{L}_1[\boldsymbol{A} + \delta\boldsymbol{A}^{(1)} + \cdots + \delta\boldsymbol{A}^{(n-1)}].$$

Da sich $\boldsymbol{L} = \boldsymbol{L}_1^{-1} \cdots \boldsymbol{L}_{n-1}^{-1} = (l_{ik})$ ohne Rundungsfehler aus den Koeffizienten l_{ik} ergibt, vgl. 5.1.C, folgt unter Beachtung von (4), (6) die Darstellung

$$\boldsymbol{LR} = \boldsymbol{A} + \delta\boldsymbol{A}, \qquad \delta\boldsymbol{A} := \sum_{k=1}^{n-1} \delta\boldsymbol{A}^{(k)}$$

mit

$$\|\delta\boldsymbol{A}\| \leqq v \left(\|\boldsymbol{A}\| + 3 \sum_{k=2}^{n} \|\boldsymbol{M}^{(k)}\| \right). \tag{7}$$

5.3.1. *Rundungsfehleranalyse der Dreiecksfaktorisierung nach Gauß.* Für $\boldsymbol{A} \in \mathfrak{R}^{n,n}$ sei der Gaußsche Algorithmus durchführbar, und $\boldsymbol{L}, \boldsymbol{R} \in \mathfrak{R}^{n,n}$ seien die berechneten Dreiecksfaktoren. Dann gilt

$$\boldsymbol{LR} = \boldsymbol{A} + \delta\boldsymbol{A} \tag{8}$$

mit einer Störung $\delta\boldsymbol{A}$, die der Abschätzung

$$\|\delta\boldsymbol{A}\| \leqq vF[\boldsymbol{A}]\, \|\boldsymbol{A}\| \tag{9}$$

genügt, wobei

$$F[\boldsymbol{A}] := 1 + 3 \sum_{k=2}^{n} \|\boldsymbol{M}^{(k)}\| / \|\boldsymbol{A}\| \tag{10}$$

ist.

Dies besagt: *In jeder Matrizenklasse* $\mathfrak{A} \subset \mathfrak{R}^{n,n}$, *in welcher der Gaußsche Algorithmus durchführbar ist und die Zahlen* $F[\boldsymbol{A}]$ *im Sinne von*

$$F[\boldsymbol{A}] \leqq F := F[\mathfrak{A}] := \max \{F[\boldsymbol{A}]: \boldsymbol{A} \in \mathfrak{A}\} < \infty \tag{11}$$

gleichmäßig beschränkt sind, ist die Gaußsche Dreiecksfaktorisierung numerisch gutartig mit F *gemäß* (11). Gutartigkeit liegt also vor, wenn die $\boldsymbol{M}^{(k)}$ nicht beliebig größer als $\boldsymbol{A}$ werden können.

Mit den Wachstumsaussagen für $\boldsymbol{M}^{(k)}$ aus Abschnitt 5.2 ergibt sich dann das folgende Resultat:

5.3.2. *Satz.* Die Gaußsche Dreiecksfaktorisierung ist durchführbar und numerisch gutartig

(i) ohne Pivotisierung für diagonaldominante bzw. symmetrische definite Matrizen $\boldsymbol{A}$ mit $F \leqq 3n$ bezüglich der ∞- bzw. F-Norm,

(ii) mit Spalten-, Zeilen- oder vollständiger Pivotisierung für alle regulären Matrizen $\boldsymbol{A}$ mit $F \leqq 3 \cdot 2^n$ bezüglich der Normen mit $p = 1, \infty, F$,

sofern die Kondition von $\boldsymbol{A}$ im Sinne von

$$vF \operatorname{cond}(\boldsymbol{A}) < 1 \tag{12}$$

mit den angegebenen Werten von F nicht zu groß ist.

Beweis. Die Aussagen aus Abschnitt 5.2 sind nicht direkt auf $\boldsymbol{A}$ anwendbar, da sie nur in exakter Arithmetik gültig sind. Für die betrachteten Matrizenklassen ist jedoch der erste Eliminationsschritt stets durchführbar. Aus der Herleitung von 5.3.1 ist zu sehen, daß die

berechnete Matrix $\boldsymbol{M}^{(2)}$ und die berechneten Koeffizienten l_{i1} das exakte Resultat eines Eliminationsschrittes für eine gestörte Matrix $\tilde{\boldsymbol{A}}^{(1)} = \boldsymbol{A} + \delta\boldsymbol{A}^{(1)}$ sind, wobei $\delta\boldsymbol{A}^{(1)}$ der Ungleichung (9) genügt. Auf Grund von (12) und des Störungslemmas 4.1.2 ist $\tilde{\boldsymbol{A}}^{(1)}$ für alle solche Störungen regulär, im definiten Fall also auch definit, siehe Ü 4.1.9, so daß die Aussagen aus 5.2 auf $\tilde{\boldsymbol{A}}^{(1)}$ statt $\boldsymbol{A}$ angewendet werden können und die jeweiligen Schranken für $\boldsymbol{M}^{(2)}$ liefern. Analog ergeben sich $\boldsymbol{M}^{(3)}$ und die l_{i2} als exaktes Resultat für eine gestörte Matrix $\tilde{\boldsymbol{A}}^{(2)}$ usw. Durch Induktion folgt dann die Behauptung.

Für den diagonaldominanten Fall müßte noch gezeigt werden, daß auch die berechneten Matrizen $\boldsymbol{M}^{(k)}$ diagonaldominant sind. Dies ist bei im Sinne von

$$\tau_i = \sum_{\substack{j=1 \\ j \neq i}}^{n} |a_{ij}|/|a_{ii}| < \left(\frac{1-\nu}{1+\nu}\right)^i \doteq 1 - 2i\nu \quad (i = 1, \ldots, n)$$

genügend starker Diagonaldominanz der Fall, wir verzichten jedoch auf einen detaillierten Beweis. Praktisch ist diese Bedingung kaum schärfer als die eigentliche Dominanzbedingung $\tau_i < 1$ $(i = 1, \ldots, n)$, vgl. den Beweis zu 5.2.8. □

5.3.3. *Bemerkung.* (i) Die Bedingung (12) garantiert, daß sämtliche Matrizen aus der Umgebung $\mathcal{M} := \{\boldsymbol{A} + \delta\boldsymbol{A} : \|\delta\boldsymbol{A}\| \leqq \nu F \|\boldsymbol{A}\|\}$ von $\boldsymbol{A}$ regulär sind, und sichert daher die Durchführbarkeit der Dreiecksfaktorisierung mit regulärem $\boldsymbol{R}$ in der durch ν charakterisierten Gleitpunktarithmetik. Auf die Bedingung (12) kann verzichtet werden, wenn bei der Dreiecksfaktorisierung im k-ten Schritt $a_{kk}^{(k)} := \nu \|\boldsymbol{A}\|$ gesetzt wird, falls mit der jeweiligen Pivotstrategie kein von 0 verschiedenes Pivot gefunden werden kann. Durch diese zusätzliche Störung wird die Gutartigkeit der Faktorisierung nicht beeinträchtigt. Allerdings können dann in der Menge $\mathcal{M}$ auch singuläre Matrizen liegen, insbesondere kann $\boldsymbol{A}$ selbst singulär sein. Die Lösung von $\boldsymbol{Ax} = \boldsymbol{b}$ ist dann eine numerisch inkorrekt gestellte Aufgabe, und die mit der obigen Modifikation erzwungene Dreiecksfaktorisierung ist i. allg. nicht zur Lösung geeignet; das berechnete $\boldsymbol{x}$ ist i. allg. von der Größenordnung $1/\nu$. In gewissen Ausnahmefällen — etwa bei der sog. inversen Iteration zur Eigenwertbestimmung, siehe Abschnitt 13.4 — ist dieser Fall jedoch sogar erwünscht.

(ii) In 5.3.2 kann der Faktor 3 bei sorgfältigerer Abschätzung durch 2 ersetzt werden.

(iii) Für praktisch auftretende Gleichungssysteme gilt 5.3.2 (ii) mit der empirischen Schranke $F \leqq \gamma n$, $\gamma \approx 10$, vgl. 5.2.2.

(iv) Im Fall der vollständigen Pivotisierung ist

$$F \leqq 1.8 n^2 n^{0.25 \ln n},$$

vgl. 5.2.7. Für praktisch auftretende Systeme liegt F in der Größenordnung von 1.

(v) Bei Bedarf kann μ_k bzw. $\|\boldsymbol{M}^{(k)}\|$ im Laufe der Elimination mit berechnet werden. Ein zu schnelles Wachstum dieser Größen signalisiert einen Verlust an numerischer Gutartigkeit. □

B. *Fehleranalyse der Gleichungsauflösung*

Wir betrachten jetzt den gesamten Lösungsprozeß, bei dem sich an die Dreiecksfaktorisierung $\boldsymbol{A} = \boldsymbol{LR}$ die Vorwärtselimination $\boldsymbol{c} = \boldsymbol{L}^{-1}\boldsymbol{b}$ und die Rücksubstitution $\boldsymbol{x} = \boldsymbol{R}^{-1}\boldsymbol{c}$ gemäß 5.1.7 anschließen. Nach 4.3.2 genügen die berechneten Größen $\boldsymbol{c}$

bzw. $\boldsymbol{x}$ den gestörten Gleichungen

$$(\boldsymbol{L} + \delta\boldsymbol{L})\,\boldsymbol{c} = \boldsymbol{b} \quad \text{bzw.} \quad (\boldsymbol{R} + \delta\boldsymbol{R})\,\boldsymbol{x} = \boldsymbol{c} \tag{13}$$

mit den elementweisen Schranken

$$|\delta l_{ik}| \leqq \nu k\,|l_{ik}| \quad \text{bzw.} \quad |\delta r_{kj}| \leqq \nu(n - j + 1)\,|r_{kj}|\,. \tag{14}$$

Aus (13) folgt unter Berücksichtigung von (8)

$$(\boldsymbol{L} + \delta\boldsymbol{L})\,(\boldsymbol{R} + \delta\boldsymbol{R})\,\boldsymbol{x} = (\boldsymbol{A} + \delta_1\boldsymbol{A})\,\boldsymbol{x} = \boldsymbol{b} \tag{15}$$

mit

$$\delta_1\boldsymbol{A} := \delta\boldsymbol{A} + \delta\boldsymbol{A}', \qquad \delta\boldsymbol{A}' := \boldsymbol{L}\delta\boldsymbol{R} + \delta\boldsymbol{L}(\boldsymbol{R} + \delta\boldsymbol{R}), \tag{16}$$

d. h., die äquivalenten Störungen $\delta_1\boldsymbol{A}$ für den Gesamtprozeß entstehen durch Überlagerung der Störungen $\delta\boldsymbol{A}$ für die Faktorisierung und der Störungen $\delta\boldsymbol{A}'$ für die Lösung der beiden Dreieckssysteme.

Wegen (14) ist

$$|\delta a'_{ij}| = \left|\sum_{k=1}^{\min(i,j)} [l_{ik}\delta r_{kj} + \delta l_{ik}(r_{kj} + \delta r_{kj})]\right|$$

$$\leqq \nu \sum_{k=1}^{\min(i,j)} |l_{ik}r_{kj}|\,(n - j + 1 + k) \leqq \nu(n + 1) \sum_{k=1}^{\min(i,j)} |l_{ik}r_{kj}|\,.$$

Analog zu (3.2.19) gilt mit $|\varepsilon|, |\vartheta| \leqq \nu$

$$l_{ik}r_{kj} = l_{ik}a_{kj}^{(k)} = a_{ij}^{(k)}/(1 + \varepsilon) - a_{ij}^{(k+1)}/[(1 + \varepsilon)\,(1 + \vartheta)],$$

also $|l_{ik}r_{kj}| \leqq |a_{ij}^{(k)}| + |a_{ij}^{(k-1)}|$ und daher

$$|\delta a'_{ij}| \leqq \nu(n + 1)\left(|a_{ij}^{(1)}| + 2\sum_{k=2}^{\min(i,j)-1} |a_{ij}^{(k)}|\right).$$

Die letzte Ungleichung kann äquivalent in der Form

$$|\delta\boldsymbol{A}'| \leqq \nu(n + 1)\left[|\boldsymbol{A}| + 2\sum_{k=2}^{n}\left(\begin{array}{c|c} \boldsymbol{O} & \boldsymbol{O} \\ \hline \boldsymbol{O} & |\boldsymbol{M}^{(k)}| \end{array}\right)\right] \tag{17}$$

geschrieben werden und führt auf

$$\|\delta\boldsymbol{A}'\| \leqq \nu(n + 1)\left(\|\boldsymbol{A}\| + 2\sum_{k=2}^{n} \|\boldsymbol{M}^{(k)}\|\right) = \nu F'[\boldsymbol{A}]\,\|\boldsymbol{A}\| \tag{18}$$

mit

$$F'[\boldsymbol{A}] := (n + 1)\left(1 + 2\sum_{k=2}^{n} \|\boldsymbol{M}^{(k)}\|/\|\boldsymbol{A}\|\right). \tag{19}$$

Ein Vergleich mit (9), (10) zeigt, daß $F' \leqq (n + 1)\,F$ gilt. Unter Beachtung von (16) ergibt sich daher das folgende Resultat.

5.3.4. *Rundungsfehleranalyse des Gaußschen Algorithmus.* Für $\boldsymbol{A} \in \mathfrak{R}^{n,n}$ und $\boldsymbol{b} \in \mathfrak{R}^n$ sei der Gaußsche Algorithmus durchführbar, und $\boldsymbol{x} \in \mathfrak{R}^n$ sei die nach diesem Verfahren berechnete Lösung des Gleichungssystems $\boldsymbol{Ax} = \boldsymbol{b}$. Dann gilt

$$(\boldsymbol{A} + \delta_1\boldsymbol{A})\,\boldsymbol{x} = \boldsymbol{b} \tag{20}$$

mit einer Störung $\delta_1\boldsymbol{A}$, für die

$$\|\delta_1\boldsymbol{A}\| \leqq \nu F_1[\boldsymbol{A}]\,\|\boldsymbol{A}\| \quad \text{mit} \quad F_1[\boldsymbol{A}] \leqq (n+2)\,F[\boldsymbol{A}] \tag{21}$$

und $F[\boldsymbol{A}]$ gemäß (10) gilt.

Aus der Störung $\delta_1\boldsymbol{A}$ kann auf den Fehler $\delta\boldsymbol{x} = \boldsymbol{x} - \boldsymbol{x}^*$ geschlossen werden.

5.3.5. *Folgerung.* Der erzeugte Rundungsfehler $\delta\boldsymbol{x} = \boldsymbol{x} - \boldsymbol{x}^*$, $\boldsymbol{x}^* = \boldsymbol{A}^{-1}\boldsymbol{b}$, genügt der Abschätzung

$$\|\delta\boldsymbol{x}\| \leqq \varkappa\,\|\boldsymbol{x}\| \quad \text{mit} \quad \varkappa := \nu F_1[\boldsymbol{A}]\,\mathrm{cond}\,(\boldsymbol{A}), \tag{22}$$

und im Fall $\varkappa < 1$ gilt

$$\|\delta\boldsymbol{x}\| \leqq \varkappa/(1-\varkappa)\,\|\boldsymbol{x}^*\|. \tag{23}$$

Beweis. Aus (20) folgt $\delta\boldsymbol{x} = -\boldsymbol{A}^{-1}\delta_1\boldsymbol{A}\boldsymbol{x}$ und damit (22), während sich (23) aus 4.1.3 ergibt. □

5.3.4 besagt: *In jeder Matrizenklasse* $\mathfrak{A} \in \mathfrak{R}^{n,n}$, *für die* (11) *gilt, ist die Berechnung von* $\boldsymbol{x}$ *nach dem Gaußschen Algorithmus numerisch gutartig mit der Fehlerkumulationskonstanten* $F_1 \leqq (n+2)\,F$. *Durch* 5.3.5 *wird die numerische Stabilität mit demselben* F_1 *ausgedrückt.*

5.3.6. *Bemerkung.* (i) Die bei der Lösung der Dreieckssysteme auftretende spezielle Fehlerkorrelation bewirkt, daß der durch den Anteil $\delta\boldsymbol{A}'$ von $\delta_1\boldsymbol{A}$ erzeugte Fehler $-\boldsymbol{A}^{-1}\delta\boldsymbol{A}'\boldsymbol{x}$ im allgemeinen wesentlich kleiner als die Schranke $\|\boldsymbol{A}^{-1}\|\,\|\delta\boldsymbol{A}'\|\,\|\boldsymbol{x}\|$ ist, und zwar um so mehr, je größer cond $(\boldsymbol{A})$ ist. Praktisch kann daher F_1 in 5.3.5 durch $\tilde{F}_1$,

$$F_1 \approx \begin{cases} F, & \text{falls cond } (\boldsymbol{A}) \text{ groß}, \\ nF, & \text{falls cond } (\boldsymbol{A}) \text{ klein}, \end{cases}$$

ersetzt werden. Aus den angegebenen Gründen führt auch eine höhere Genauigkeit bei der Lösung der Dreieckssysteme i. allg. nicht zu einer höheren Genauigkeit von $\boldsymbol{x}$.

(ii) Die elementweisen Schranken (3) und (17) zeigen, daß $\delta\boldsymbol{A}$ und $\delta\boldsymbol{A}'$ zeilen- bzw. spaltenweise relativ klein sind, sofern $\boldsymbol{M}^{(k)}$ zeilen- bzw. spaltenweise in vergleichbarer Größenordnung mit $\boldsymbol{A}$ liegt. Letzteres ist bei Zeilen- bzw. Spaltenpivotisierung der Fall. Man kann zeigen, daß die Zeilenpivotisierung in etwa dieselbe Wirkung wie eine im Sinne von 4.1.C optimale Zeilenskalierung hat. Die Spaltenpivotisierung wirkt dagegen etwa wie eine in bezug auf das Residuum $\boldsymbol{b} - \boldsymbol{Ax}$ optimale Spaltenskalierung. □

Übungsaufgaben

Ü 5.3.1. Man ergänze die Beweisskizze zu 5.3.2.

Ü 5.3.2. Man beweise: Wenn $d = \det(\boldsymbol{A})$ gemäß 5.1.9 berechnet wird, gilt für den berechneten Wert d

$$d = [\det(\boldsymbol{A} + \delta\boldsymbol{A})](1 + \varepsilon)$$

mit $\delta\boldsymbol{A}$ gemäß 5.3.1 und $|\varepsilon| \leqq \nu(n - 1)$, sofern kein Unter- oder Überlauf eintritt.

5.4. Genauigkeitsabschätzung und iterative Verbesserung

Bei gegebenen Eingangsdaten $\{\boldsymbol{A}, \boldsymbol{b}\}$ liefert der Gaußsche Algorithmus nach erfolgreicher Ausführung die berechnete Lösung $\boldsymbol{x} \in \Re^n$ und die berechneten Dreiecksfaktoren $\boldsymbol{L}, \boldsymbol{R} \in \Re^{n,n}$. Auf Grund der auftretenden Rundungsfehler ist $\boldsymbol{x}$ i. allg. nur ein Näherungswert für die exakte Lösung $\boldsymbol{x}^* = \boldsymbol{A}^{-1}\boldsymbol{b}$. Daher entstehen die Fragen,

- wie der erzeugte Rundungsfehler $\delta\boldsymbol{x} = \boldsymbol{x} - \boldsymbol{x}^*$ abgeschätzt werden kann,
- wie zu $\boldsymbol{x}$ eine Korrektur $\boldsymbol{h}$ berechnet werden kann, so daß $\boldsymbol{x}' = \boldsymbol{x} + \boldsymbol{h}$ eine bessere Näherung für $\boldsymbol{x}^*$ als $\boldsymbol{x}$ ist.

A. *Fehlerschranken und Korrekturen*

Im vorangegangenen Abschnitt wurde gezeigt, daß die berechnete Lösung $\boldsymbol{x}$ das gestörte System

$$(\boldsymbol{A} + \delta_1\boldsymbol{A})\,\boldsymbol{x} = \boldsymbol{b} \quad \text{mit} \quad \|\delta_1\boldsymbol{A}\| \leqq \nu F_1 \|\boldsymbol{A}\| \tag{1}$$

löst und daß

$$\|\delta\boldsymbol{x}\| \leqq \varkappa\|\boldsymbol{x}\| \quad \text{mit} \quad \varkappa := \nu F_1 \operatorname{cond}(\boldsymbol{A}) \tag{2}$$

gilt. Die Schranken in (1) und (2) sind für die praktische Fehlerabschätzung nicht besonders geeignet, da sie mit den a-priori-Werten für F_1 gemäß 5.3.2 und 5.3.4 i. allg. zu pessimistisch sind. Es bietet sich daher an, unter Verwendung des Residuums

$$\boldsymbol{r}^* = \boldsymbol{r}^*(\boldsymbol{x}) = \boldsymbol{b} - \boldsymbol{A}\boldsymbol{x} \tag{3}$$

eine a-posteriori-Abschätzung durchzuführen und dabei auf Teilabschnitt 4.1.D zurückzugreifen. Nach 4.1.18 existiert zu $\boldsymbol{x}$ eine Störung $\delta\boldsymbol{A}$ mit

$$(\boldsymbol{A} + \delta\boldsymbol{A})\,\boldsymbol{x} = \boldsymbol{b} \quad \text{und} \quad \|\delta\boldsymbol{A}\| \leqq \|\boldsymbol{b} - \boldsymbol{A}\boldsymbol{x}\|/\|\boldsymbol{x}\|, \tag{4}$$

und nach 4.1.17 gilt die Fehlerabschätzung

$$\|\delta\boldsymbol{x}\| \leqq \|\boldsymbol{A}^{-1}\|\,\|\boldsymbol{b} - \boldsymbol{A}\boldsymbol{x}\|. \tag{5}$$

Bei der Auswertung von (5) wird eine Schranke bzw. zumindest ein Schätzwert für $\|\boldsymbol{A}^{-1}\|$ benötigt; dieses Problem wird im Teilabschnitt C behandelt. Außerdem muß das Residuum $\boldsymbol{r}^* = \boldsymbol{b} - \boldsymbol{A}\boldsymbol{x}$ numerisch berechnet werden, was in der Form

$\boldsymbol{r} = \mathrm{fl}(\boldsymbol{b} - \boldsymbol{A}\boldsymbol{x})$ wie in Ü 2.3.4 geschehen möge. Dabei wird der Rundungsfehler $\delta\boldsymbol{r} := \boldsymbol{r} - \boldsymbol{r}^*$ erzeugt. Wenn $\boldsymbol{d}$ eine komponentenweise Schranke für $\delta\boldsymbol{r}$ ist, d. h., wenn

$$|\delta\boldsymbol{r}| = |\boldsymbol{r} - \boldsymbol{r}^*| \leqq \boldsymbol{d} \tag{6}$$

gilt, folgt $|\boldsymbol{r}^*| \leqq |\boldsymbol{r}| + |\boldsymbol{r}^* - \boldsymbol{r}| \leqq |\boldsymbol{r}| + \boldsymbol{d}$, also $\|\boldsymbol{r}^*\| \leqq \| |\boldsymbol{r}| + \boldsymbol{d}\|$. Damit ergibt sich das nachstehende Resultat.

5.4.1. *A-posteriori-Abschätzung unter Verwendung von* $\boldsymbol{r}$. Für reguläres $\boldsymbol{A} \in \mathfrak{R}^{n,n}$ und $\boldsymbol{b}, \boldsymbol{x} \in \mathfrak{R}^n$ werde $\boldsymbol{r}, \boldsymbol{d} \in \mathfrak{R}^n$ mit

$$\boldsymbol{r} := \mathrm{fl}(\boldsymbol{b} - \boldsymbol{A}\boldsymbol{x}), \qquad |\boldsymbol{r} - (\boldsymbol{b} - \boldsymbol{A}\boldsymbol{x})| \leqq \boldsymbol{d}$$

berechnet. Dann gilt

$$(\boldsymbol{A} + \delta\boldsymbol{A})\,\boldsymbol{x} = \boldsymbol{b} \quad \text{mit} \quad \|\delta\boldsymbol{A}\| \leqq \Delta\boldsymbol{A} := \| |\boldsymbol{r}| + \boldsymbol{d}\| / \|\boldsymbol{x}\| \tag{7}$$

sowie

$$\|\delta\boldsymbol{x}\| \leqq \|\boldsymbol{A}^{-1}\| \, \| |\boldsymbol{r}| + \boldsymbol{d}\| . \tag{8}$$

Aufwand: $\sim n^2$ opms (für $\boldsymbol{r}$) + $\sim 2n^2$ ops (für $\boldsymbol{d}$ bei Berechnung nach Ü 2.3.4).

5.4.2. *Bemerkung.* (i) Der bei der Addition $|\boldsymbol{r}| + \boldsymbol{d}$ und der Normbildung in (7), (8) auftretende Rundungsfehler kann vernachlässigt werden, da die beteiligten Vektoren nichtnegative Komponenten besitzen.

(ii) Praktisch wird häufig der Fehlerterm $\boldsymbol{d}$ vernachlässigt. Das kann zu falschen Schranken führen, wenn $|\delta\boldsymbol{r}|$ und $\boldsymbol{d}$ in der Größenordnung von $|\boldsymbol{r}|$ liegen. Man beachte dabei, daß für $\delta\boldsymbol{r}$ die a-priori-Abschätzung

$$|\delta\boldsymbol{r}| \leqq \nu(|\boldsymbol{r}| + n\,|\boldsymbol{A}|\,|\boldsymbol{x}|) \tag{9}$$

gilt, siehe Ü 2.3.4.

(iii) Falls das Residuum durch Akkumulation der Skalarprodukte in höherer Genauigkeit ν_1 berechnet wird, gilt

$$|\delta\boldsymbol{r}| \leqq \nu\,|\boldsymbol{r}| + \nu_1 n\,|\boldsymbol{A}|\,|\boldsymbol{x}| . \tag{10}$$

In diesem Fall ist fast immer $\boldsymbol{d} \approx \nu\,|\boldsymbol{r}|$, so daß die Fehlerschranke $\boldsymbol{d}$ weggelassen werden kann. □

Um zu einer verbesserten Näherung $\boldsymbol{x}'$ zu kommen, gehen wir von dem Ansatz $\boldsymbol{x}^* = \boldsymbol{x} + \boldsymbol{h}^*$, d. h.

$$\boldsymbol{h}^* = \boldsymbol{x}^* - \boldsymbol{x} = \boldsymbol{A}^{-1}\boldsymbol{b} - \boldsymbol{x} = \boldsymbol{A}^{-1}(\boldsymbol{b} - \boldsymbol{A}\boldsymbol{x}) = \boldsymbol{A}^{-1}\boldsymbol{r}^*$$

aus. Die zu $\boldsymbol{x}^*$ führende Korrektur $\boldsymbol{h}^*$ genügt also dem Gleichungssystem

$$\boldsymbol{A}\boldsymbol{h}^* = \boldsymbol{r}^* \tag{11}$$

mit derselben Koeffizientenmatrix $\boldsymbol{A}$ wie das Ausgangssystem, aber dem Residuum $\boldsymbol{r}^*$ als rechter Seite. Praktisch wird (11) unter Verwendung der bereits vorliegenden Dreiecksfaktorisierung von $\boldsymbol{A}$ mit $O(n^2)$ opms gelöst, wobei statt $\boldsymbol{r}^*$ das berechnete

Residuum $\boldsymbol{r}$ genommen wird. Die als Lösung von $\boldsymbol{Ah} = \boldsymbol{r}$ tatsächlich berechnete Korrektur $\boldsymbol{h}$ weist daher den Fehler

$$\boldsymbol{h} - \boldsymbol{h}^* = \boldsymbol{h} - \boldsymbol{A}^{-1}\boldsymbol{r}^* = (\boldsymbol{h} - \boldsymbol{A}^{-1}\boldsymbol{r}) + \boldsymbol{A}^{-1}\delta\boldsymbol{r} \tag{12}$$

auf. Der erste Term rechts entspricht dem bei der Lösung von $\boldsymbol{Ah} = \boldsymbol{r}$ entstandenem Fehler und kann analog zu (1), (2) durch

$$\|\boldsymbol{h} - \boldsymbol{A}^{-1}\boldsymbol{r}\| \leqq \varkappa \|\boldsymbol{h}\| \tag{13}$$

abgeschätzt werden. Der zweite Term berücksichtigt den bei der Berechnung von $\boldsymbol{r}$ hervorgerufenen Fehler und läßt sich mit $\boldsymbol{d}$ aus (6) in der Form

$$\|\boldsymbol{A}^{-1}\delta\boldsymbol{r}\| \leqq \|\boldsymbol{A}^{-1}\| \, \|\delta\boldsymbol{r}\| \leqq \|\boldsymbol{A}^{-1}\| \, \|\boldsymbol{d}\| \tag{14}$$

abschätzen.

Die berechnete Korrektur $\boldsymbol{h}$ kann nun in zweierlei Hinsicht verwendet werden: zur Berechnung der — hoffentlich! — besseren Näherung

$$\boldsymbol{x}' = \boldsymbol{x} + \boldsymbol{h} \tag{15}$$

für $\boldsymbol{x}^*$ mit der aus (12), (13), (14) folgenden Fehlerschranke

$$\|\boldsymbol{x}' - \boldsymbol{x}^*\| = \|\boldsymbol{h} - \boldsymbol{h}^*\| \leqq \varkappa \|\boldsymbol{h}\| + \|\boldsymbol{A}^{-1}\| \, \|\boldsymbol{d}\| \tag{16}$$

oder durch Umstellen von (12) gemäß $\boldsymbol{x} - \boldsymbol{x}^* = -\boldsymbol{h} + (\boldsymbol{h} - \boldsymbol{A}^{-1}\boldsymbol{r}) + \boldsymbol{A}^{-1}\delta\boldsymbol{r}$ zur Berechnung der — hoffentlich! — besseren Fehlerschranke

$$\|\delta\boldsymbol{x}\| = \|\boldsymbol{x} - \boldsymbol{x}^*\| \leqq (1 + \varkappa) \|\boldsymbol{h}\| + \|\boldsymbol{A}^{-1}\| \, \|\boldsymbol{d}\|. \tag{17}$$

Im Fall

$$\varkappa \ll 1 \quad \text{und} \quad \boldsymbol{d} \ll |\boldsymbol{r}| \tag{18}$$

liefert (8) die Schranke

$$\|\delta\boldsymbol{x}\| \lessapprox \|\boldsymbol{A}^{-1}\| \, \|\boldsymbol{r}\|,$$

während (17) die i. allg. bessere Schranke $\|\delta\boldsymbol{x}\| \lessapprox \|\boldsymbol{h}\| \approx \|\boldsymbol{A}^{-1}\boldsymbol{r}\|$ ergibt. In ähnlicher Weise folgt, daß $\boldsymbol{x}'$ dann eine bessere Näherung für $\boldsymbol{x}^*$ als $\boldsymbol{x}$ ist.

B. *Iterative Verbesserung*

Es liegt nahe, die eben beschriebene Korrektur mit $\boldsymbol{x}'$ anstelle von $\boldsymbol{x}$ zu wiederholen und damit zu einer iterativen Verbesserung zu gelangen.

5.4.3. *Iterative Verbesserung der Lösung von* $\boldsymbol{Ax} = \boldsymbol{b}$

Aufgabe: Zu gegebenem $\boldsymbol{A} \in \Re^{n,n}$, $\boldsymbol{b} \in \Re^n$ und gegebener Dreiecksfaktorisierung $\boldsymbol{P}_Z\boldsymbol{A}\boldsymbol{P}_S^{\mathrm{T}} = \boldsymbol{L}\boldsymbol{R}$ von $\boldsymbol{A}$ mit $\boldsymbol{L}, \boldsymbol{R} \in \Re^{n,n}$ und Permutationen $\boldsymbol{P}_S, \boldsymbol{P}_Z$ soll die Lösung des Gleichungssystems $\boldsymbol{Ax} = \boldsymbol{b}$ berechnet und iterativ verbessert werden.

Algorithmus:

S0: $\boldsymbol{x}^0 := \boldsymbol{o}$

for $k := 0(1)k_{\max}$ **do**

S1: Berechne $\boldsymbol{r}^k := \mathrm{fl}\,(\boldsymbol{b} - \boldsymbol{A}\boldsymbol{x}^k)$

S2: Berechne $\boldsymbol{h} = \boldsymbol{h}^k$ als Lösung von $\boldsymbol{A}\boldsymbol{h} = \boldsymbol{r}^k$ unter Verwendung der Dreiecksfaktorisierung von $\boldsymbol{A}$ aus den Gleichungen

$$\boldsymbol{L}\boldsymbol{g} = \boldsymbol{P}_Z\boldsymbol{r}^k, \qquad \boldsymbol{R}\bar{\boldsymbol{h}} = \boldsymbol{g}, \qquad \boldsymbol{h} = \boldsymbol{P}_S^{\mathrm{T}}\bar{\boldsymbol{h}}$$

S3: Setze $\boldsymbol{x}^{k+1} := \boldsymbol{x}^k + \boldsymbol{h}^k$

Aufwand: $\sim 2n^2$ opms pro Schritt $k \geqq 1$, $\sim n^2$ opms für $k = 0$

Schritt S1 wird dabei wie in Ü 2.3.4 realisiert. Im Fall der Spaltenpivotisierung erfordert S2 die Ausführung des „Solvers" 5.2.4 mit $\boldsymbol{r}^k$ statt $\boldsymbol{b}$; die Dreiecksfaktorisierung muß dazu vorher nach 5.2.3 berechnet worden sein.

Wir bemerken noch, daß im ersten Schritt ($k = 0$) $\boldsymbol{r}^0 = \boldsymbol{b}$ gilt, d. h., $\boldsymbol{x}^1$ stimmt mit der nach dem Gaußschen Algorithmus berechneten ersten Näherung für $\boldsymbol{x}^*$ überein.

5.4.4. *Fehleranalyse.* Es seien $\boldsymbol{x}^k$, $\boldsymbol{r}^k$, $\boldsymbol{h}^k$ ($k = 1, \ldots, k_{\max}$) die nach Algorithmus 5.4.3 berechneten Größen, und es gelte

$$|\boldsymbol{r}^k - (\boldsymbol{b} - \boldsymbol{A}\boldsymbol{x}^k)| \leqq \nu\,|\boldsymbol{r}^k| + \nu\varphi\,|\boldsymbol{A}|\,|\boldsymbol{x}^k| \tag{19}$$

sowie

$$(\boldsymbol{A} + \delta\hat{\boldsymbol{A}}_k)\,\boldsymbol{h}^k = \boldsymbol{r}^k \quad \text{mit} \quad \|\delta\hat{\boldsymbol{A}}_k\| \leqq \nu F_1\,\|\boldsymbol{A}\|. \tag{20}$$

Wenn die Konstanten φ, F_1 der Bedingung

$$\bar{\varkappa} := \nu[F_1 + \varphi + 1]\mathrm{c}\,\mathrm{ond}\,(\boldsymbol{A}) \leqq 0.2 \tag{21}$$

genügen, dann gibt es Störungen $\delta\boldsymbol{A}_k$, $\vartheta\boldsymbol{A}_k$ und $\boldsymbol{\xi}^k$, so daß

$$[\boldsymbol{A} + (\delta\boldsymbol{A}_k + \vartheta\boldsymbol{A}_k)]\,(\boldsymbol{x}^k + \boldsymbol{\xi}^k) = \boldsymbol{b} \tag{22}$$

gilt mit

$$|\delta\boldsymbol{A}_k| \leqq \nu\varphi\,|\boldsymbol{A}|, \qquad \|\vartheta\boldsymbol{A}_k\| \leqq \nu\vartheta_k\,\|\boldsymbol{A}\|, \qquad |\boldsymbol{\xi}^k| \leqq \nu\,|\boldsymbol{x}^k|, \tag{23}$$

wobei

$$[\bar{\varkappa}/(1 - 2\bar{\varkappa})]\,[2\varphi + (1 + 2\hat{\varkappa}^{k-1}/\nu)/\mathrm{cond}\,(\boldsymbol{A})]/(1 - \hat{\varkappa}^k), \quad \hat{\varkappa} = \bar{\varkappa}/(1 - \bar{\varkappa}). \tag{24}$$

Der Fehler von $\boldsymbol{x}^k$ genügt der Abschätzung

$$\|\boldsymbol{x}^k - \boldsymbol{x}^*\| \leqq \nu[1 + \hat{\varkappa}^k/\nu + \varphi\,\mathrm{cond}\,(\boldsymbol{A})/(1 - 2\bar{\varkappa})]\,\|\boldsymbol{x}^*\|. \tag{25}$$

Der Beweis ist sehr technisch und soll deshalb übergangen werden. Als Normen können die mit $p = 1, \infty$ genommen werden; für $p = 2, F$ sind etwas andere Konstanten erforderlich, auf deren Angabe wir verzichten.

5.4.5. *Bemerkung.* (i) Bedingung (19) ist bei einfachgenauer Berechnung von $\boldsymbol{r}^k$ mit $\varphi := n$, bei Akkumulation der Skalarprodukte mit höherer Genauigkeit $\nu_1 \ll \nu$ mit $\varphi := n\nu_1/\nu$ erfüllt, siehe 5.4.2. Bedingung (20) ist mit dem von der Matrix $\boldsymbol{A}$ und der Pivotstrategie abhängigem $F_1 = F_1[\boldsymbol{A}]$ aus Abschnitt 5.3 erfüllt.

(ii) Für genügend großes k folgt aus (24) wegen $\bar{x} \leqq 0.2$

$$\vartheta_k \lessapprox [2\varphi + 1/\mathrm{cond}\,(\boldsymbol{A})]/3,$$

mit (23) also für die Gesamtstörung

$$\|\delta \boldsymbol{A}_k + \vartheta \boldsymbol{A}_k\| \lessapprox \nu\big[(5/3)\,\varphi + 1/\big(3\,\mathrm{cond}\,(\boldsymbol{A})\big)\big]\,\|\boldsymbol{A}\|. \tag{26}$$

Es liegt also numerische Gutartigkeit vor, wobei die Kumulationskonstante nur von φ — der Genauigkeit der Residuenberechnung — abhängt, nicht aber von der durch F_1 charakterisierten Genauigkeit des Gaußschen Algorithmus als Basisverfahren. In analoger Weise folgt aus (25) für genügend großes k

$$\|\boldsymbol{x}^k - \boldsymbol{x}^*\| \lessapprox \nu[1 + (5/3)\,\varphi\,\mathrm{cond}\,(\boldsymbol{A})]\,\|\boldsymbol{x}^*\|, \tag{27}$$

d. h., der relative Fehler ist unabhängig von F_1 beschränkt. Bei Berechnung der Residuen in höherer Genauigkeit ist $\varphi\,\mathrm{cond}\,(\boldsymbol{A})$ klein, vgl. wieder 5.2.4, so daß (27) dann praktisch die volle relative (einfache) Genauigkeit von $\boldsymbol{x}^k$ bedeutet.

(iii) Als Abbruchkriterium kann z. B. die Bedingung

$$\|\boldsymbol{h}^k\| > \|\boldsymbol{h}^{k-1}\|/2$$

verwendet werden, d. h., es wird abgebrochen, wenn sich die Korrekturen nicht mehr schnell genug verkleinern. □

Die iterative Verbesserung kann als eine Ergänzung des Gaußschen Algorithmus angesehen werden, die bei genügend kleinem ν, nicht zu schlechter Dreiecksfaktorisierung und nicht zu schlecht konditioniertem $\boldsymbol{A}$ in wenigen Korrekturschritten einen Gesamtprozeß liefert, der in einem besseren Sinne als der Gaußsche Algorithmus numerisch gutartig ist.

C. *Abschätzung der Norm der Inversen*

Für die Abschätzung des erzeugten Rundungsfehlers gemäß 5.4.1, aber auch für die Abschätzung des durch Meßfehler von $\boldsymbol{A}$ und $\boldsymbol{b}$ hervorgerufenen Fehlers nach Abschnitt 4.1.B bzw. C wird $\|\boldsymbol{A}^{-1}\|$ benötigt. Da es i. allg. nur auf die Größenordnung ankommt, kann ein Schätzwert mit etwa 50% Fehler durchaus als akzeptabel angesehen werden. Dabei bietet es sich an, auf die Dreiecksfaktorisierung

$$\boldsymbol{P}_Z \boldsymbol{A} \boldsymbol{P}_S^{\mathsf{T}} = \boldsymbol{L}\boldsymbol{R} \tag{28}$$

von $\boldsymbol{A}$ zurückzugreifen.

Im Abschnitt 6.5 werden wir sehen, daß die explizite Berechnung von $\boldsymbol{A}^{-1}$ aus den Dreiecksfaktoren $\sim 2/3n^3$ opms, also das Doppelte des Aufwandes für die Dreiecksfaktorisierung selbst kostet und daher aus Aufwandsgründen ausscheidet. Wir versuchen daher, mit geringerem Aufwand einen akzeptablen Schätzwert zu bestimmen.

Als Hilfsaufgabe betrachten wir dazu die Berechnung eines Schätzwertes für $\|\boldsymbol{C}^{-1}\|_1$, wenn $\boldsymbol{C}$ eine reguläre Dreiecksmatrix ist. Nach Definition der Matrixnorm gilt

$$\|\boldsymbol{C}^{-1}\| = \max\{\|\boldsymbol{C}^{-1}\boldsymbol{u}\|/\|\boldsymbol{u}\| : \boldsymbol{u} \neq \boldsymbol{o}\} = \max\{\|\boldsymbol{v}\|/\|\boldsymbol{u}\| : \boldsymbol{u} \neq \boldsymbol{o}, \boldsymbol{C}\boldsymbol{v} = \boldsymbol{u}\}. \tag{29}$$

Bei gegebenem $\boldsymbol{u}$ ergibt sich dabei $\boldsymbol{v} = \boldsymbol{C}^{-1}\boldsymbol{u}$ als Lösung des Dreieckssystems $\boldsymbol{C}\boldsymbol{v} = \boldsymbol{u}$ im Fall einer unteren Dreiecksmatrix gemäß

$$
\begin{aligned}
v_1 &= \frac{u_1}{c_{11}},\\
v_2 &= -\frac{c_{21}}{c_{22}}\,v_1 + \frac{u_2}{c_{22}},\\
v_3 &= -\frac{c_{31}}{c_{33}}\,v_1 - \frac{c_{32}}{c_{33}}\,v_2 + \frac{u_3}{c_{33}},\\
&\vdots\\
v_i &= -\frac{c_{i1}}{c_{ii}}\,v_1 - \frac{c_{i2}}{c_{ii}}\,v_2 - \frac{c_{i3}}{c_{ii}}\,v_3 - \cdots - \frac{c_{i,i-1}}{c_{ii}}\,v_{i-1} + \frac{u_i}{c_{ii}},\\
&\vdots\\
v_n &= -\frac{c_{n1}}{c_{nn}}\,v_1 - \frac{c_{n2}}{c_{nn}}\,v_2 - \frac{c_{n3}}{c_{nn}}\,v_3 - \cdots - \frac{c_{n,n-1}}{c_{nn}}\,v_{n-1} + \frac{u_n}{c_{nn}}.
\end{aligned}
\tag{30}
$$

Wir wählen jetzt $\boldsymbol{u} = (u_k)$ als Vektor mit den Komponenten $u_k = \pm 1$ und verteilen die Vorzeichen so, daß $\|\boldsymbol{v}\|_1 = \sum_{i=1}^{n} |v_i|$ möglichst groß wird. Im ersten Schritt wird $u_1 = 1$ gesetzt. Damit sind v_1 und die in der ersten Spalte nach dem Gleichheitszeichen stehenden Summanden $-(c_{i1}/c_{ii})\,v_1$ von v_i $(i = 2, \ldots, n)$ festgelegt. Im zweiten Schritt wird für jede der beiden Möglichkeiten $u_2 = u_2^{+,-} = \pm 1$ das zugehörige $v_2 = v_2^{+,-} = -(c_{21}/c_{22})\,v_1 + u_2^{+,-}/c_{22}$ berechnet und das Vorzeichen in Abhängigkeit davon festgelegt, welche der beiden Summen

$$
|v_2^{+,-}| + \sum_{i=3}^{n} \left| -\frac{c_{i1}}{c_{ii}}\,v_1 - \frac{c_{i2}}{c_{ii}}\,v_2^{+,-} \right|
$$

die größere ist. Diese Summen werden als Näherung für $|v_2| + \sum_{i=3}^{n} |v_i|$ angesehen, wobei die $v_3, \ldots, v_n$ durch die aus den ersten beiden Spalten gebildeten Teilsummen ersetzt worden sind. Man beachte, daß v_1 bereits festgelegt ist und auf den Größenvergleich keinen Einfluß mehr hat. Vor dem k-ten Schritt sind in dieser Weise $u_1, \ldots, u_{k-1}$ sowie $v_1, \ldots, v_{k-1}$ fixiert. Das Vorzeichen von $u_k = \pm 1$ wird dann so gewählt, daß die aus den ersten k Spalten gebildeten Teilsummen der v_i $(i \geqq k)$ betragsmäßig summiert einen möglichst großen Wert ergeben. Dieses Vorgehen führt auf den folgenden Algorithmus:

5.4.6. *Heuristische Maximierung von* $\|\boldsymbol{C}^{-1}\boldsymbol{u}\|_1/\|\boldsymbol{u}\|_1$.

Aufgabe: Für die reguläre Dreiecksmatrix $\boldsymbol{C} = (c_{ij}) \in \mathfrak{R}^{n,n}$, $c_{ii} \neq 0$ $(i = 1, \ldots, n)$ werden $\boldsymbol{u} = (u_i) \in \mathfrak{R}^n$ mit $u_i = \pm 1$ und $\boldsymbol{v} = (v_i) = \boldsymbol{C}^{-1}\boldsymbol{u}$ so bestimmt, daß $\|\boldsymbol{v}\|_1 = \sum_{i=1}^{n} |v_i|$ möglichst groß wird.

Algorithmus (falls C untere Dreiecksmatrix; für obere Dreiecksmatrix analog):
$u_1 := 1$, $v_1 := 1/c_{11}$, **for** $i := 2(1)n$ **do** $v_i := -c_{i1} * v_1/c_{ii}$
for $k := 2(1)n$ **do**
  $z := 1/c_{kk}$, $u_k := v_k - z$, $v_k := v_k + z$
  SM $:= |u_k|$, SP $:= |v_k|$
  for $i := k + 1(1)n$ **do**
    $z := c_{ik}/c_{ii}$, $u_i := v_i - z * u_k$, $v_i := v_i - z * v_k$
    SM $:=$ SM $+ |u_i|$, SP $:=$ SP $+ |v_i|$
  if SM $>$ SP **then** [$u_k := -1$, **for** $i := k(1)n$ **do** $v_i := u_i$]
    else $u_k := 1$
Aufwand: $\sim 2n^2$ ops $+ \sim 3/2n^2$ opm

Für die in 5.4.6 berechneten Vektoren $\boldsymbol{u}, \boldsymbol{v}$ gilt nach (29)

$$\frac{\|\boldsymbol{v}\|_1}{\|\boldsymbol{u}\|_1} \leqq \|\boldsymbol{C}^{-1}\|_1, \tag{31}$$

und nach Konstruktion kann erwartet werden, daß der Quotient $\|\boldsymbol{v}\|_1/\|\boldsymbol{u}\|_1$ in der Größenordnung von $\|\boldsymbol{C}^{-1}\|_1$ liegt.

Wir kehren jetzt wieder zum Ausgangsproblem zurück, eine Näherung ζ für $\zeta^* = \|\boldsymbol{A}^{-1}\|$ zu berechnen. Aus (28) und der Invarianz der Matrixnormen mit $p = 1, 2, \infty, F$ gegenüber Zeilen- und Spaltenvertauschungen folgt

$$\|\boldsymbol{A}^{-1}\| = \|(\boldsymbol{LR})^{-1}\|.$$

Für die ∞-Norm gilt dann

$$\zeta^*_\infty = \|\boldsymbol{A}^{-1}\|_\infty = \|\boldsymbol{A}^{-\mathsf{T}}\|_1 = \|(\boldsymbol{LR})^{-\mathsf{T}}\|_1 = \max\{\|\boldsymbol{z}\|_1/\|\boldsymbol{x}\|_1 : \boldsymbol{x} \neq \boldsymbol{o},\ (\boldsymbol{LR})^{\mathsf{T}}\boldsymbol{z} = \boldsymbol{x}\}. \tag{32}$$

Wenn $\boldsymbol{z}$ bei gegebenem $\boldsymbol{x}$ in der Form

$$\boldsymbol{R}^{\mathsf{T}}\boldsymbol{y} = \boldsymbol{x}, \qquad \boldsymbol{L}^{\mathsf{T}}\boldsymbol{z} = \boldsymbol{y} \tag{33}$$

aus zwei Dreieckssystemen berechnet wird, ergibt sich

$$\frac{\|\boldsymbol{z}\|_1}{\|\boldsymbol{x}\|_1} = \frac{\|\boldsymbol{z}\|_1}{\|\boldsymbol{y}\|_1} \cdot \frac{\|\boldsymbol{y}\|_1}{\|\boldsymbol{x}\|_1} = \frac{\|(\boldsymbol{L}^{\mathsf{T}})^{-1}\,\boldsymbol{y}\|_1}{\|\boldsymbol{y}\|_1} \cdot \frac{\|(\boldsymbol{R}^{\mathsf{T}})^{-1}\,\boldsymbol{x}\|_1}{\|\boldsymbol{x}\|_1} \leqq \zeta^*_\infty. \tag{34}$$

Da der Maximierungsalgorithmus nur auf einen der beiden Quotienten $\|\boldsymbol{z}\|_1/\|\boldsymbol{y}\|_1$ oder $\|\boldsymbol{y}\|_1/\|\boldsymbol{x}\|_1$ angewendet werden kann, muß überlegt werden, welcher der beiden Dreiecksfaktoren $\boldsymbol{L}^{\mathsf{T}}$ bzw. $\boldsymbol{R}^{\mathsf{T}}$ stärker zu $(\boldsymbol{LR})^{-\mathsf{T}}$ beiträgt, d. h., in welchem sich die Kondition von $\boldsymbol{A}$ vorrangig niederschlägt. Dies hängt selbstverständlich von der Art der Dreiecksfaktorisierung (28) ab.

Wir betrachten zunächst den Fall der Spalten- bzw. vollständigen Pivotisierung oder den einer spaltendiagonaldominanten Matrix $\boldsymbol{A}$. Hier gilt

$$|l_{ij}| \leqq 1 \qquad (i > j), \tag{35}$$

und es folgt

$$\|\boldsymbol{L}^{\mathsf{T}}\|_1 \leqq n, \qquad \|(\boldsymbol{L}^{\mathsf{T}})^{-1}\|_1 = \|\boldsymbol{L}^{-1}\|_\infty \leqq 2^{n-1}$$

unabhängig von der Matrix $\boldsymbol{A}$, siehe Ü 5.4.1. Eine beliebig schlechte Kondition der Matrix $\boldsymbol{A}$ kann sich also nicht im Faktor $\boldsymbol{L}$ niederschlagen. Praktisch ist $\|\boldsymbol{L}^{-1}\|_\infty$ meist viel kleiner als der maximal mögliche Wert 2^{n-1}, so daß der erste Quotient in (34) wegen

$$\frac{1}{\|\boldsymbol{L}^{\mathsf{T}}\|_1} \leqq \frac{\|\boldsymbol{L}^{-\mathsf{T}}\boldsymbol{y}\|_1}{\|\boldsymbol{y}\|_1} \leqq \|\boldsymbol{L}^{-\mathsf{T}}\|_1 \tag{36}$$

nur wenig variiert. In diesem Fall sollte also der zweite Quotient $\|\boldsymbol{y}\|_1/\|\boldsymbol{x}\|_1$ maximiert werden.

Im Fall der Zeilen- bzw. vollständigen Pivotisierung oder dem einer zeilendiagonaldominanten Matrix $\boldsymbol{A}$ gilt

$$|r_{ij}| \leqq |r_{ii}| \qquad (i < j), \tag{37}$$

vgl. (5.2.7). Wenn hier

$$\boldsymbol{LR} = (\boldsymbol{LD})\,(\boldsymbol{D}^{-1}\boldsymbol{R}) =: \boldsymbol{L}_1\boldsymbol{R}_1 \quad \text{mit} \quad \boldsymbol{D} := \operatorname{diag}(r_{ii}) \tag{38}$$

gesetzt wird, ist $\boldsymbol{R}_1$ wegen (37) eine obere Einsdreiecksmatrix mit

$$|(\boldsymbol{R}_1)_{ij}| \leqq 1 \qquad (i < j), \tag{39}$$

d. h., $\boldsymbol{R}_1^{\mathsf{T}}$ hat qualitativ dieselben Eigenschaften wie $\boldsymbol{L}$ im Fall der Spaltenpivotisierung, und eine mögliche schlechte Kondition schlägt sich in $\boldsymbol{L}_1$ nieder. In der zu (33), (34) analogen Darstellung mit $\boldsymbol{L}_1$ statt $\boldsymbol{L}$ und $\boldsymbol{R}_1$ statt $\boldsymbol{R}$ sollte daher der erste Quotient $\|\boldsymbol{z}\|_1/\|\boldsymbol{y}\|_1 = \|\boldsymbol{L}_1^{-\mathsf{T}}\boldsymbol{y}\|_1/\|\boldsymbol{y}\|_1$ maximiert werden.

Im Fall einer symmetrischen und positiv definiten Matrix $\boldsymbol{A}$ gilt $r_{ii} > 0$ und

$$\boldsymbol{LR} = (\boldsymbol{LD}^{1/2})\,(\boldsymbol{D}^{-1/2}\boldsymbol{R}) =: \boldsymbol{L}_2\boldsymbol{R}_2 \quad \text{mit} \quad \boldsymbol{D}^{1/2} := \operatorname{diag}\left(\sqrt{r_{ii}}\right),\ \boldsymbol{R}_2 = \boldsymbol{L}_2^{\mathsf{T}}, \tag{40}$$

vgl. 5.2.10. Hier kann wie im Fall der Spaltenpivotisierung vorgegangen werden, wobei $\boldsymbol{L}$ durch $\boldsymbol{L}_2$ und $\boldsymbol{R}$ durch $\boldsymbol{R}_2 = \boldsymbol{L}_2^{\mathsf{T}}$ zu ersetzen ist; siehe Ü 5.4.2. Zusammenfassend erhalten wir die folgenden Vorschriften:

5.4.7. *Schätzung von* $\|(\boldsymbol{LR})^{-1}\|_\infty$.

Aufgabe: Die Matrix $\boldsymbol{A}$ sei gemäß (28) durch eine untere Einsdreiecksmatrix $\boldsymbol{L}$ und eine reguläre obere Dreiecksmatrix $\boldsymbol{R}$ gegeben. Aus den Dreiecksfaktoren $\boldsymbol{L}, \boldsymbol{R}$ ist ein Schätzwert ζ_∞ für $\zeta_\infty^* := \|(\boldsymbol{LR})^{-1}\|_\infty$ zu berechnen.

Algorithmus: Bestimme $\boldsymbol{x}, \boldsymbol{z}$ mit $(\boldsymbol{LR})^{\mathsf{T}}\boldsymbol{z} = \boldsymbol{x}$ wie folgt:

Fall 1 (Spalten- oder vollständige Pivotisierung; $\boldsymbol{A}$ spaltendiagonaldominant):

S1.1: Führe 5.4.6 mit $\boldsymbol{C} := \boldsymbol{R}^{\mathsf{T}}$ aus, setze $\boldsymbol{x} := \boldsymbol{u}$, $\boldsymbol{y} := \boldsymbol{v}$

S2.1: Berechne $\boldsymbol{z}$ aus $\boldsymbol{L}^{\mathsf{T}}\boldsymbol{z} = \boldsymbol{y}$

Fall 2 (Zeilen- oder vollständige Pivotisierung; $\boldsymbol{A}$ zeilendiagonaldominant):

S1.2: Setze $\boldsymbol{D} := \operatorname{diag}(r_{ii})$, $\boldsymbol{L}_1 := \boldsymbol{LD}$, $\boldsymbol{R}_1 := \boldsymbol{D}^{-1}\boldsymbol{R}$,
führe 5.4.6 mit $\boldsymbol{C} := \boldsymbol{L}_1^{\mathsf{T}}$ aus, setze $\boldsymbol{y} := \boldsymbol{u}$, $\boldsymbol{z} := \boldsymbol{v}$

S2.2: Berechne $\boldsymbol{x} := \boldsymbol{R}_1^{\mathsf{T}}\boldsymbol{y}$

Fall 3 ($\boldsymbol{A}$ symmetrisch und positiv definit):

S1.3: Setze $\boldsymbol{D}^{1/2} := \operatorname{diag}\left(\sqrt{r_{ii}}\right)$, $\boldsymbol{L}_2 := \boldsymbol{L}\boldsymbol{D}^{1/2}$, $\boldsymbol{R}_2 := \boldsymbol{D}^{-1/2}\boldsymbol{R} = \boldsymbol{L}_2^{\mathrm{T}}$, führe 5.4.6 mit $\boldsymbol{C} := \boldsymbol{L}_2$ aus, setze $\boldsymbol{x} := \boldsymbol{u}$, $\boldsymbol{y} := \boldsymbol{v}$

S2.3: Berechne $\boldsymbol{z}$ aus $\boldsymbol{L}_2^{\mathrm{T}}\boldsymbol{z} = \boldsymbol{y}$

S3: Setze $\zeta_\infty := \|\boldsymbol{z}\|_1/\|\boldsymbol{x}\|_1$.

Aufwand: $\sim 5/2n^2$ ops + $\sim 2n^2$ opm im Fall 1, in den Fällen 2 und 3 zusätzlich $\sim n^2/2$ opm zur Bildung von $\boldsymbol{L}_1$ bzw. $\boldsymbol{L}_2$, falls die Dreiecksfaktorisierung nicht so modifiziert worden ist, daß sie diese Faktoren direkt liefert.

5.4.8. *Bemerkung.* (i) Wenn $\boldsymbol{z}$ wie in 5.4.7 bestimmt und daraus $\boldsymbol{w} = (\boldsymbol{L}\boldsymbol{R})^{-1}\boldsymbol{z}$ gemäß

$$\boldsymbol{L}\boldsymbol{t} = \boldsymbol{z}, \qquad \boldsymbol{R}\boldsymbol{w} = \boldsymbol{t} \tag{41}$$

berechnet wird, gilt

$$\zeta'_\infty := \|\boldsymbol{w}\|_\infty/\|\boldsymbol{z}\|_\infty \leqq \|(\boldsymbol{L}\boldsymbol{R})^{-1}\|_\infty, \tag{42}$$

und i. allg. wird auch ζ'_∞ in der Größenordnung von ζ^*_∞ liegen. Die Größe

$$\hat{\zeta}_\infty := \max\{\zeta_\infty, \zeta'_\infty\} = \max\{\|\boldsymbol{z}\|_1/\|\boldsymbol{x}\|_1, \|\boldsymbol{w}\|_\infty/\|\boldsymbol{z}\|_\infty\} \leqq \zeta^*_\infty$$

ist daher i. allg. eine bessere Schätzung für ζ^*_∞ als jede der beiden einzelnen Werte ζ_∞ bzw. ζ'_∞. Der Mehraufwand zur Lösung der beiden Dreieckssysteme (41) ist $\sim n^2$ opms, also klein. Umfangreiche Experimente zeigen, daß $\hat{\zeta}_\infty$ fast immer einen relativen Fehler von weniger als 50% aufweist, siehe B 5.6.

(ii) Obwohl $\boldsymbol{z}$ und $\boldsymbol{x}$ mit dem Ziel der Maximierung von $\|\boldsymbol{z}\|_1/\|\boldsymbol{x}\|_1$ als Näherung für ζ^*_∞ konstruiert worden sind, stellen die Zahlen

$$\zeta_1 := \|\boldsymbol{z}\|_\infty/\|\boldsymbol{x}\|_\infty \quad \text{und} \quad \zeta'_1 := \|\boldsymbol{w}\|_1/\|\boldsymbol{z}\|_1$$

meist gute Schätzwerte für $\|(\boldsymbol{L}\boldsymbol{R})^{-1}\|_1$ dar. Ebenso sind

$$\zeta_2 := \|\boldsymbol{z}\|_2/\|\boldsymbol{x}\|_2 \quad \text{und} \quad \zeta'_2 := \|\boldsymbol{w}\|_2/\|\boldsymbol{z}\|_2$$

meist gute Näherungen für $\|(\boldsymbol{L}\boldsymbol{R})^{-1}\|_2$.

(iii) Algorithmus 5.4.6 kann leicht so modifiziert werden, daß $\boldsymbol{C}$ in der faktorisierten Form $\boldsymbol{C} = \boldsymbol{D}'\boldsymbol{C}'\boldsymbol{E}'$ mit regulären Diagonalmatrizen $\boldsymbol{D}' = \operatorname{diag}(d'_i)$, $\boldsymbol{E}' = \operatorname{diag}(e'_i)$ akzeptiert wird; man braucht bloß c_{ij} überall durch $d'_i c'_{ij} e'_j$ zu ersetzen. Der Mehraufwand ist $\sim n^2/2$ opm, im Fall $\boldsymbol{E}' = \boldsymbol{I}$ sogar nur n opm.

(iv) Wenn eine mittels Zeilenpivotisierung berechnete Dreiecksfaktorisierung $\boldsymbol{A}\boldsymbol{P}_S^{\mathrm{T}} = \boldsymbol{L}\boldsymbol{R}$ von $\boldsymbol{A}$ bekannt ist, kann die in (iii) beschriebene Modifikation zur a-posteriori-Abschätzung von $\|(\hat{\boldsymbol{D}}\boldsymbol{A})^{-1}\|_\infty$ verwendet werden, wobei $\hat{\boldsymbol{D}} = \operatorname{diag}(\hat{d}_i)$ eine reguläre Skalierungsmatrix bezeichnet. Diese Aufgabe tritt bei der in bezug auf die Fehlerabschätzung optimalen Skalierung von $\boldsymbol{A}$ auf, siehe 4.1.15 und 4.1.16. In 5.4.7, Fall 2, ist dazu $\boldsymbol{L}$ bzw. $\boldsymbol{L}_1$ durch $\hat{\boldsymbol{D}}\boldsymbol{L}$ bzw. $\hat{\boldsymbol{D}}\boldsymbol{L}_1$ zu ersetzen, d. h., man erhält

S1: Führe 5.4.6 in der gemäß (iii) modifizierten Form mit $\boldsymbol{C} := \boldsymbol{D}\boldsymbol{L}^{\mathrm{T}}\hat{\boldsymbol{D}} = \boldsymbol{L}_1^{\mathrm{T}}\hat{\boldsymbol{D}}$ durch, setze $\boldsymbol{y} := \boldsymbol{u}$, $\boldsymbol{z} := \boldsymbol{v}$

S2: Berechne $\boldsymbol{x} := \boldsymbol{R}^{\mathsf{T}}\boldsymbol{D}^{-1}\boldsymbol{y} = \boldsymbol{R}_1^{\mathsf{T}}\boldsymbol{y}$

S3: Setze $\zeta_{\infty,\hat{\boldsymbol{b}}} := \|\boldsymbol{z}\|_1/\|\boldsymbol{x}\|_1$

Dann ist $\zeta_{\infty,\hat{\boldsymbol{b}}}$ ein Schätzwert für $\zeta^*_{\infty,\hat{\boldsymbol{b}}} := \|(\hat{\boldsymbol{D}}\boldsymbol{A})^{-1}\|_\infty$.
Der Aufwand beträgt $\sim 5/2n^2$ (ops + opm).

(v) Die tatsächlich berechneten Dreiecksfaktoren genügen statt (28) der Beziehung

$$\boldsymbol{P}_Z(\boldsymbol{A} + \delta\boldsymbol{A})\,\boldsymbol{P}_S^{\mathsf{T}} = \boldsymbol{L}\boldsymbol{R},$$

siehe 5.3.1, so daß ζ_∞ in Wirklichkeit eine Schätzung für $\|(\boldsymbol{A} + \delta\boldsymbol{A})^{-1}\|_\infty$ ist. Für genügend kleines $\delta\boldsymbol{A}$ — dies ist in der Regel der Fall — gilt dann nach 4.1.2

$$\|(\boldsymbol{A} + \delta\boldsymbol{A})^{-1}\|_\infty = \|\boldsymbol{A}^{-1}\|_\infty + O(\|\delta\boldsymbol{A}\|_\infty),$$

so daß die Verfälschung durch $\delta\boldsymbol{A}$ i. allg. vernachlässigt werden kann. □

Übungsaufgaben

Ü 5.4.1. Es sei $\boldsymbol{L} = (l_{ij}) = \boldsymbol{I} - \boldsymbol{G}$ eine untere Einsdreiecksmatrix. Man beweise die Gültigkeit von

(i) $\boldsymbol{G}^n = \boldsymbol{O}$ und $\boldsymbol{L}^{-1} = \boldsymbol{I} + \sum_{k=1}^{n-1} \boldsymbol{G}^k$.

(ii) $\|\boldsymbol{L}\|_p \leqq \|\boldsymbol{L}^*\|_p = n$ und $\|\boldsymbol{L}^{-1}\|_p \leqq \|(\boldsymbol{L}^*)^{-1}\|_p = 2^{n-1}$ für alle L mit $|l_{ij}| \leqq 1$, wobei $\boldsymbol{L}^* = (l_{ij}^*)$ die untere Einsdreiecksmatrix mit $l_{ij}^* := -1$ $(i > j)$ und $p = 1, \infty$ ist. Was ergibt sich im Fall $|l_{ij}| \leqq q$ $(i > j)$?

Hinweis: In (i) beachte man Ü 4.1.1, in (ii) zeige man zunächst $|\boldsymbol{L}^{-1}| \leqq (\boldsymbol{L}^*)^{-1}$.

Ü 5.4.2. Es sei $\boldsymbol{A}$ symmetrisch und positiv definit mit der symmetrischen Dreiecksfaktorisierung $\boldsymbol{A} = \boldsymbol{L}_2\boldsymbol{L}_2^{\mathsf{T}}$, $\boldsymbol{L}_2$ wie in 5.4.7. Man zeige, daß

$$\|\boldsymbol{A}\|_2 = (\|\boldsymbol{L}_2\|_2)^2 = (\|\boldsymbol{L}_2^{\mathsf{T}}\|_2)^2 \quad \text{sowie} \quad \|\boldsymbol{A}^{-1}\|_2 = (\|\boldsymbol{L}_2^{-1}\|_2)^2 = (\|\boldsymbol{L}_2^{-\mathsf{T}}\|_2)^2$$

gilt.

Hinweis: Man beachte Ü 1.2.9.

Bemerkungen zum Kapitel 5

B 5.1. Das heute nach Gauss benannte Eliminationsverfahren war bereits vor etwa 2000 Jahren im alten China bekannt. Die Beschreibung mittels elementarer Transformationsmatrizen geht auf Turing [48] zurück und ist seit Wilkinson [61, 63, 65] Standard in der numerischen linearen Algebra. Eine leicht verständliche Beschreibung haben Forsythe/Moler [67] gegeben. Ausführliche Darstellungen einschließlich historischer Kommentare finden sich z. B. bei Faddeev/Faddeeva [63], Householder [64] und Stewart [73].

B 5.2. Trotz seiner einfachen Struktur stellt der Gaußsche Algorithmus einen schwer analysierbaren numerischen Prozeß dar, und die Frage nach einer in bezug auf den erzeugten Rundungsfehler optimalen Pivotstrategie ist nach wie vor ungeklärt. Die erste adäquate Rundungsfehleranalyse wurde für den symmetrischen und definiten Fall durch von Neumann/Goldstine [47] gegeben, allerdings nur für die damals übliche Festpunktarithmetik. Grundlegende Ergebnisse für den allgemeinen Fall — auch für Gleitpunktarithmetik — gehen auf Wilkinson [61, 63, 65] zurück.

B 5.3. Von besonderer Bedeutung sind die von WILKINSON erhaltenen Aussagen über die numerische Gutartigkeit des Gaußschen Algorithmus in Abhängigkeit von den betrachteten Matrizenklassen und den verwendeten Pivotisierungsstrategien. Die konkreten Normschranken für die äquivalenten Störungen sind dabei mehr als eine Folgerung anzusehen, entscheidend ist die Existenz von solchen Störungsschranken.

B 5.4. Detaillierte Untersuchungen verschiedener Pivotisierungsstrategien werden bis in die jüngste Zeit durchgeführt. Aus der Vielzahl der Arbeiten seien VAN DER SLUIS [70] und SKEEL [79, 80, 81] erwähnt.

B 5.5. Die Rundungsfehleranalyse zur iterativen Verbesserung wurde erstmals von WILKINSON durchgeführt; weitere Arbeiten zu dieser Thematik sind u. a. MOLER [67], JANKOWSKI/WOŹNIAKOWSKI [77] und SKEEL [80].

B 5.6. Es gibt eine Vielzahl von Arbeiten zur exakten Abschätzung von $\|\boldsymbol{C}^{-1}\|$ im Fall einer regulären Dreiecksmatrix $\boldsymbol{C}$, wir zitieren etwa LEMEIRE [75]. Die Schranken sind allerdings meist zu pessimistisch bzw. zu aufwendig zu berechnen. Die im Abschnitt 5.4 angegebenen heuristischen Methoden, die auf CLINE et al. [79] zurückgehen, brachten hier eine Wende, indem sie mit vertretbarem Aufwand die Berechnung akzeptabler unterer Schranken zulassen. Numerische Tests mit zufällig erzeugten Testmatrizen haben STEWART [80] und O'LEARY [80] durchgeführt. In der Originalarbeit von CLINE et al. wird ζ' gemäß 5.4.8(i) als Schätzer verwendet, während O'LEARY ζ und $\hat{\zeta}$ empfiehlt. Auf die Möglichkeit der a-posteriori-Abschätzung von $\|(\hat{\boldsymbol{D}}\boldsymbol{A})^{-1}\|_\infty$ unter Verwendung einer zeilenpivotisierten $\boldsymbol{LR}$-Faktorisierung von $\boldsymbol{A}$ gemäß 5.4.8(iv) scheint bisher nicht hingewiesen worden zu sein. Für weitere Resultate sei auf GRIMES/LEWIS [81] und CLINE et al. [82] verwiesen. In gewissen Ausnahmefällen können die Konditionsschätzer auch unbrauchbare Werte liefern, siehe CLINE/REW [83].

6. Modifikationen des Gaußschen Algorithmus

Gegenstand dieses Kapitels sind Modifikationen des Gaußschen Algorithmus, die sich von der $(n-1)$-stufigen Grundform aus 5.1 bzw. 5.2 einmal durch die Ausnutzung spezieller Eigenschaften der Matrix $\boldsymbol{A}$ wie Symmetrie oder schwache Besetztheit, zum anderen durch die Art der Berechnung der Elemente der Dreiecksfaktoren $\boldsymbol{L}$ und $\boldsymbol{R}$ unterscheiden. Außerdem gehen wir auf die Berechnung inverser Matrizen ein.

6.1. Dreiecksfaktorisierungen symmetrischer Matrizen

Es sei $\boldsymbol{A} \in \mathbf{S}^{n,n}$ eine symmetrische Matrix, die durch die Elemente im unteren Dreieck gegeben sei; zur Speicherung symmetrischer Matrizen siehe Bemerkung 6.1.14 am Schluß dieses Abschnitts.

Wir suchen im folgenden Modifikationen des Gaußschen Algorithmus, bei denen sich die Symmetrie von $\boldsymbol{A}$ in einer Halbierung des Aufwands gegenüber dem nichtsymmetrischen Fall niederschlägt, d. h., die mit $\sim n^3/6$ opms auskommen und auf nur einem Dreieck von $\boldsymbol{A}$ mit $\sim n^2/2$ S realisiert werden können. Die Möglichkeiten zur Ausnutzung der Symmetrie hängen dabei wesentlich davon ab, ob $\boldsymbol{A}$ definit oder indefinit ist. Im definiten Fall genügt es, sich auf positiv definites $\boldsymbol{A}$ zu beschränken, denn für negativ definites $\boldsymbol{A}'$ ist $\boldsymbol{A} := -\boldsymbol{A}'$ positiv definit.

A. *Positiv definite symmetrische Matrizen*

In Abschnitt 5.2 wurde gezeigt, daß der Gaußsche Algorithmus für eine positiv definite Matrix $\boldsymbol{A} \in \mathbf{S}^{n,n}$ ohne Pivotisierung durchführbar ist. Dabei gilt

$$\boldsymbol{L}^T = \boldsymbol{D}^{-1}\boldsymbol{R} \quad \text{mit} \quad \boldsymbol{D} = \operatorname{diag}(d_k), \qquad d_k := r_{kk} > 0 \qquad (k = 1, \ldots, n), \tag{1}$$

d. h., die Faktorisierung $\boldsymbol{A} = \boldsymbol{L}\boldsymbol{R}$ kann in der symmetrischen Form

$$\boldsymbol{A} = \boldsymbol{L}\boldsymbol{D}\boldsymbol{L}^T \tag{2}$$

mit der Einsdreiecksmatrix $\boldsymbol{L}$ und der positiven Diagonalmatrix $\boldsymbol{D}$ geschrieben werden.

Die im k-ten Schritt

$$\boldsymbol{A}^{(k+1)} := \boldsymbol{L}_k \boldsymbol{A}^{(k)} \qquad (k = 1, \ldots, n-1), \qquad \boldsymbol{A}^{(1)} := \boldsymbol{A} \tag{3}$$

gemäß der Partitionierung

$$\boldsymbol{A}^{(k)} = \begin{pmatrix} a_{11}^{(k)} & \cdots & \cdots & a_{1n}^{(k)} \\ & \ddots & & \\ & & a_{kk}^{(k)} & \cdots & a_{kn}^{(k)} \\ \boldsymbol{O} & & \vdots & & \vdots \\ & & a_{nk}^{(k)} & \cdots & a_{nn}^{(k)} \end{pmatrix} =: \begin{pmatrix} & \boldsymbol{R}^{(k)} & \\ \hline \boldsymbol{O} & \boldsymbol{M}^{(k)} & \end{pmatrix} \begin{matrix} k \\ \\ n \end{matrix} \tag{4}$$

vorkommenden quadratischen Restmatrizen $\boldsymbol{M}^{(k)}$ und $\boldsymbol{M}^{(k+1)}$ sind außerdem wie $\boldsymbol{M}^{(1)} = \boldsymbol{A}$ symmetrisch und positiv definit, siehe 5.2.9. Von den Elementen

$$a_{ij}^{(k+1)} = a_{ji}^{(k+1)} := a_{ij}^{(k)} - l_{ik} * a_{kj}^{(k)} \qquad (i, j = k+1, \ldots, n) \tag{5}$$

der Matrix $\boldsymbol{M}^{(k+1)}$ brauchen daher nur die im unteren Dreieck berechnet zu werden; die übrigen entstehen aus diesen durch Spiegelung an der Hauptdiagonalen. Der Gesamtprozeß kann dann auf dem Platz des unteren Dreiecks von $\boldsymbol{A}$ ausgeführt werden, wobei die erzeugten Nullen mit den Eliminationskoeffizienten

$$l_{ik} := a_{ik}^{(k)} / a_{kk}^{(k)} \qquad (i = k+1, \ldots, n) \tag{6}$$

überspeichert werden. Bei spaltenweiser Berechnung der $a_{ij}^{(k+1)}$ müssen dabei die in (5) benötigten Elemente $a_{kj}^{(k)}$ $(j = k+1, \ldots, n)$ — dies sind gerade die Elemente $r_{k,k+1}, \ldots, r_{kn}$ der k-ten Zeile von $\boldsymbol{R}$ — in einem Feld $\boldsymbol{r}$ der Länge n zwischengespeichert werden. Der sonst für $\boldsymbol{R}$ benötigte Speicherplatz von $\sim n^2/2$ S reduziert sich damit auf $\sim n$ S für $\boldsymbol{r}$. Eine zu 5.2.3 analoge Computerrealisierung des beschriebenen Verfahrens lautet wie folgt; mit $\mathfrak{S}^{n,n}$ bezeichnen wir dabei die Menge der symmetrischen Matrizen aus $\mathfrak{R}^{n,n}$:

6.1.1. ***LDL*^{T}-*Faktorisierung einer symmetrischen und positiv definiten Matrix***

Aufgabe: Für die positiv definit Matrix $\boldsymbol{A} \in \mathfrak{S}^{n,n}$ ist die Dreiecksfaktorisierung $\boldsymbol{A} = \boldsymbol{L}\boldsymbol{D}\boldsymbol{L}^T$ zu berechnen. Die Matrix $\boldsymbol{A}$ ist durch die Elemente des unteren Dreiecks gegeben, und die signifikanten Elemente von $\boldsymbol{L}, \boldsymbol{D} \in \mathbf{R}^{n,n}$ sind auf dem Platz des unteren Dreiecks von $\boldsymbol{A}$ zu speichern. Wenn $d_k = r_{kk} \leqq 0$ für ein $k \in \{1, \ldots, n\}$

gilt, soll abgebrochen und $ie = -1$ gesetzt werden; andernfalls wird $ie = 0$ gesetzt, und es gilt $d_k > 0$ $(k = 1, \ldots, n)$.

Algorithmus:

$ie := 0$
for $k := 1(1)n - 1$ **do**
  if $a_{kk} \leqq 0$ **then** [$ie := -1$, **stop**]
  for $i := k + 1(1)n$ **do** [$r_i := a_{ik}$, $a_{ik} := r_i/a_{kk}$]
  for $j := k + 1(1)n$ **do**
    for $i := j(1)n$ **do** $a_{ij} := a_{ij} - a_{ik} * r_j$
if $a_{nn} \leqq 0$ **then** [$ie := -1$, **stop**]

Aufwand: $\sim n^3/6$ opms, $\sim n$ S (für $r_2, \ldots, r_n$)

Für $n = 4$ ergibt sich bei erfolgreicher Faktorisierung ($ie = 0$) das folgende Belegungsmuster:

$$\begin{pmatrix} & \\ & \end{pmatrix} \begin{bmatrix} a_{11} & & & \\ a_{21} & a_{22} & & \\ a_{31} & a_{32} & a_{33} & \\ a_{41} & a_{42} & a_{43} & a_{44} \end{bmatrix} \rightarrow \begin{pmatrix} r_{12} \\ r_{23} \\ r_{34} \end{pmatrix} \begin{bmatrix} d_1 & & & \\ l_{21} & d_2 & & \\ l_{31} & l_{32} & d_3 & \\ l_{41} & l_{42} & l_{43} & d_4 \end{bmatrix}$$

6.1.2. *Bemerkung.* (i) In exakter Arithmetik ist 6.1.1 für eine symmetrische Matrix $\boldsymbol{A}$ genau dann mit $d_k > 0$ — d. h. $ie = 0$ — durchführbar, wenn $\boldsymbol{A}$ positiv definit ist, vgl. 5.2.10 und Ü 5.2.7. Das Verfahren 6.1.1 kann daher zur konstruktiven Überprüfung der positiven Definitheit einer symmetrischen Matrix $\boldsymbol{A}$ verwendet werden und liefert bei positiver Antwort gleichzeitig die dann existierende Faktorisierung $\boldsymbol{A} = \boldsymbol{LDL}^T$ mit $d_k > 0$. Analoge Aussagen gelten für Computerrechnung, sofern $\boldsymbol{A}$ nicht zu schlecht konditioniert ist, siehe 6.1.3 unten.

(ii) Bei zeilenweiser Berechnung der Elemente des unteren Dreiecks von $\boldsymbol{M}^{(k-1)}$ kann auf das Hilfsfeld $\boldsymbol{r}$ verzichtet werden. Die letzten drei Zeilen der Laufanweisung für k sind dann wie folgt abzuändern:

for $i := k + 1(1)n$ **do**
  $\varrho := a_{ik}$, $a_{ik} := \varrho/a_{kk}$
  for $j := k + 1(1)i$ **do** $a_{ij} := a_{ij} - \varrho * a_{jk}$

Man beachte dabei, daß (5) in der Form

$$a_{ij}^{(k+1)} = a_{ij}^{(k)} - l_{ik} * a_{jk}^{(k)} = a_{ij}^{(k)} - a_{ik}^{(k)} * l_{jk} \qquad (k + 1 \leqq j \leqq i)$$

geschrieben werden kann. Statt $\sim n$ S für $\boldsymbol{r}$ wird hier nur noch 1 S für ϱ zum Zwischenspeichern von $a_{ik}^{(k)} = a_{ki}^{(k)} = r_{ki}$ benötigt.

(iii) Die Lösung des Gleichungssystems $\boldsymbol{Ax} = \boldsymbol{LDL}^T\boldsymbol{x} = \boldsymbol{b}$ kann für jedes $\boldsymbol{b} \in \mathbf{R}^n$ analog zu 5.2.4 unter Verwendung der $\boldsymbol{LDL}^T$-Faktorisierung in drei Schritten durch Lösung von

S2.1: $\boldsymbol{Lc} = \boldsymbol{b}$, S2.2: $\boldsymbol{Dd} = \boldsymbol{c}$, S2.3: $\boldsymbol{L}^T\boldsymbol{x} = \boldsymbol{d}$

erfolgen. Die Determinante ergibt sich analog zu 5.2.5 gemäß

$$\det(\boldsymbol{A}) = \prod_{k=1}^{n} d_k. \quad \square$$

Zur Rundungsfehleranalyse von 6.1.1 können die Ergebnisse von 5.3 herangezogen werden, vgl. 5.3.2. Allerdings ist die dort angewendete Beweistechnik nicht auf den Fall symmetrischer Matrizen ausgerichtet, und die äquivalenten Störungen $\delta\boldsymbol{A}$ bzw. $\delta_1\boldsymbol{A}$ sind i. allg. nicht symmetrisch, was der Aufgabenklasse nicht angepaßt ist. Mit feineren und aufwendigeren Mitteln lassen sich dagegen auch symmetrische Störungen konstruieren und bessere Schranken für die Kumulationskonstanten herleiten. Aus Platzgründen geben wir nur die Ergebnisse an.

6.1.3. *Rundungsfehleranalyse.* Die $\boldsymbol{LDL}^T$-Faktorisierung 6.1.1 ist für die positiv definite Matrix $\boldsymbol{A} \in \mathbb{G}^{n,n}$ mit $d_k > 0$ $(k = 1, \ldots, n)$ durchführbar, sofern

$$vF\,\|\boldsymbol{A}^{-1}\|_2\,\|\boldsymbol{A}\|_F \leqq 0.5 \quad \text{mit} \quad F := n + 2\ln(n/2) \sim n \tag{7}$$

gilt. Zu den berechneten Faktoren $\boldsymbol{L}$, $\boldsymbol{D}$ existiert eine Störung $\delta\boldsymbol{A} \in \mathbf{S}^{n,n}$ mit

$$\boldsymbol{A} + \delta\boldsymbol{A} = \boldsymbol{LDL}^T \quad \text{und} \quad \|\delta\boldsymbol{A}\|_F \leqq vF\,\|\boldsymbol{A}\|_F. \tag{8}$$

Die zu $\boldsymbol{b} \in \mathfrak{R}^n$ gehörende und gemäß 6.1.2(iii) berechnete Lösung $\boldsymbol{x} \in \mathfrak{R}^n$ von $\boldsymbol{Ax} = \boldsymbol{b}$ ist die exakte Lösung des durch die Matrix $\delta_1\boldsymbol{A} \in \mathbf{S}^{n,n}$ gestörten Systems

$$(\boldsymbol{A} + \delta_1\boldsymbol{A})\,\boldsymbol{x} = \boldsymbol{b}, \quad \text{wobei} \quad \|\delta_1\boldsymbol{A}\|_p \leqq vF_1\,\|\boldsymbol{A}\|_p \tag{9}$$

gilt mit

$$F_1 := \begin{cases} 2n^{3/2} + F \sim 2n^{3/2} & \text{für} \quad p = F, \\ n^{3/2} + n + n^{1/2}F \sim 2n^{3/2} & \text{für} \quad p = 2. \end{cases} \tag{10}$$

Die $\boldsymbol{LDL}^T$-Faktorisierung einer symmetrischen definiten Matrix und die Lösung von $\boldsymbol{Ax} = \boldsymbol{b}$ mittels der Faktoren $\boldsymbol{L}$ und $\boldsymbol{D}$ sind also numerisch gutartige Prozesse.

Eine weitere symmetrische Faktorisierung kann erhalten werden, wenn

$$\hat{\boldsymbol{L}} := \boldsymbol{LD}^{1/2} \quad \text{mit} \quad \boldsymbol{D}^{1/2} := \operatorname{diag}\left(\sqrt{d_k}\right) \tag{11}$$

gesetzt wird. Die Matrix $\hat{\boldsymbol{L}}$ hat dann die Gestalt

$$\hat{\boldsymbol{L}} = \begin{pmatrix} \hat{l}_{11} & & & \\ \hat{l}_{21} & \hat{l}_{22} & & \\ \vdots & \vdots & \ddots & \\ \hat{l}_{n1} & \hat{l}_{n2} & \ldots & \hat{l}_{nn} \end{pmatrix} \quad \text{mit} \quad \hat{l}_{ik} = l_{ik}\sqrt{d_k} \quad (i = k, \ldots, n), \tag{12}$$

und (2) geht in $\boldsymbol{A} = \boldsymbol{LD}^{1/2}\boldsymbol{D}^{1/2}\boldsymbol{L}^T = (\boldsymbol{LD}^{1/2})\,(\boldsymbol{LD}^{1/2})^T$, also

$$\boldsymbol{A} = \hat{\boldsymbol{L}}\hat{\boldsymbol{L}}^T \tag{13}$$

über. Die symmetrische Faktorisierung (13) heißt *Cholesky-* oder $\boldsymbol{LL}^T$*-Faktorisierung*

der symmetrischen und positiv definiten Matrix $\boldsymbol{A}$. Wegen (12) und (6) gilt

$$\hat{l}_{kk} = \sqrt{d_k} = \sqrt{a_{kk}^{(k)}}, \qquad \hat{l}_{ik} = a_{ik}^{(k)}/\hat{l}_{kk} \qquad (i = k+1, \ldots, n), \tag{14}$$

und (5) geht in die symmetrische Formel

$$a_{ij}^{(k+1)} = a_{ji}^{(k+1)} := a_{ij}^{(k)} - \hat{l}_{ik} * \hat{l}_{jk} \qquad (i, j = k+1, \ldots, n) \tag{15}$$

über. Mit (14), (15) ergibt sich dann das nachfolgende, nach CHOLESKY benannte *Verfahren* zur Berechnung von $\hat{\boldsymbol{L}}$:

6.1.4. $\boldsymbol{LL^T}$*-Faktorisierung einer symmetrischen positiv definiten Matrix* $\boldsymbol{A}$ *ohne Pivotisierung.*

Aufgabe: Für positiv definites $\boldsymbol{A} \in \mathbb{S}^{n,n}$ ist die Cholesky-Faktorisierung $\boldsymbol{A} = \hat{\boldsymbol{L}}\hat{\boldsymbol{L}}^T$ zu berechnen. Die Matrix $\boldsymbol{A}$ ist durch die Elemente des unteren Dreiecks gegeben, und die signifikanten Elemente von $\hat{\boldsymbol{L}} \in \mathbf{R}^{n,n}$ sind auf dem Platz des unteren Dreiecks von $\boldsymbol{A}$ zu speichern. Wenn $a_{kk}^{(k)} \leqq 0$ für ein $k \in \{1, \ldots, n\}$ gilt, soll abgebrochen und $ie = -1$ gesetzt werden, andernfalls wird $ie = 0$ gesetzt, und es gilt $\hat{l}_{kk} > 0$ $(k = 1, \ldots, n)$.

Algorithmus:

$ie := 0$
for $k := 1(1)n$ **do**
  if $a_{kk} \leqq 0$ **then** [$ie := -1$, **stop**]
  $a_{kk} := \mathrm{sqrt}\,(a_{kk})$
  for $i := k + 1(1)n$ **do** $a_{ik} := a_{ik}/a_{kk}$
  for $j := k + 1(1)n$ **do**
    for $i := j(1)n$ **do** $a_{ij} := a_{ij} - a_{ik} * a_{jk}$

Aufwand: $\sim n^3/6$ opms $+ n$ opr

Bei erfolgreicher Faktorisierung ($ie = 0$) ergibt sich für $n = 4$ das folgende Belegungsmuster für $\boldsymbol{A}$:

$$\begin{pmatrix} a_{11} & & & \\ a_{21} & a_{22} & & \\ a_{31} & a_{32} & a_{33} & \\ a_{41} & a_{42} & a_{43} & a_{44} \end{pmatrix} \rightarrow \begin{pmatrix} \hat{l}_{11} & & & \\ \hat{l}_{21} & \hat{l}_{22} & & \\ \hat{l}_{31} & \hat{l}_{32} & \hat{l}_{33} & \\ \hat{l}_{41} & \hat{l}_{42} & \hat{l}_{43} & \hat{l}_{44} \end{pmatrix}.$$

6.1.5. *Bemerkung.* (i) Die Aussagen von Bemerkung 6.1.2 gelten sinngemäß für die Cholesky-Faktorisierung 6.1.4. Insbesondere kann die Lösung des Gleichungssystems $\boldsymbol{Ax} = \hat{\boldsymbol{L}}\hat{\boldsymbol{L}}^T\boldsymbol{x} = \boldsymbol{b}$ in zwei Schritten durch Lösung von

S2.1: $\hat{\boldsymbol{L}}\boldsymbol{c} = \boldsymbol{b}$, S2.2: $\hat{\boldsymbol{L}}^T\boldsymbol{x} = \boldsymbol{c}$

erfolgen. Die Determinante ergibt sich zu

$$\det(\boldsymbol{A}) = \left(\prod_{k=1}^{n} \hat{l}_{kk}\right)^2. \tag{16}$$

(ii) Die Rundungsfehleranalyse 6.1.3 bleibt für die Cholesky-Faktorisierung 6.1.4 und die Lösung von $\boldsymbol{A}\boldsymbol{x} = \boldsymbol{b}$ nach (i) gültig, wenn F durch $F := n + 1 + 0.5 \ln n$ ersetzt wird. Die asymptotischen Beziehungen $F \sim n$ und $F_1 \sim 2n^{3/2}$ ändern sich dabei nicht.

Auf die Durchführbarkeitsbedingung (7) kann in beiden symmetrischen Faktorisierungen verzichtet werden, wenn $a_{kk} := \nu \|\boldsymbol{A}\|$ im Fall $a_{kk} \leqq 0$ gesetzt wird. In der betrachteten Matrizenklasse wird dadurch die Gutartigkeit der Verfahren nicht beeinträchtigt, allerdings sind die derart berechneten Faktorisierungen i. allg. nicht mehr zur Lösung von $\boldsymbol{A}\boldsymbol{x} = \boldsymbol{b}$ geeignet, vgl. 5.3.3(i).

(iii) Beide Faktorisierungen 6.1.1 und 6.1.4 können mit Diagonalpivotisierung gemäß Ü 5.2.6 realisiert werden. Zu Beginn des k-ten Schrittes wird dazu ein Index $s = s(k)$ mit $k \leqq s \leqq n$ derart bestimmt, daß

$$|a_{ss}^{(k)}| = \max \{|a_{ii}^{(k)}| : k \leqq i \leqq n\} \tag{17}$$

gilt. Danach werden die Spalten k und s sowie die Zeilen k und s von $\boldsymbol{A}^{(k)}$ vertauscht. Wenn die bereits berechneten l_{ij} bzw. $\hat{l}_{ij}$ $(j < k)$ analog vertauscht werden, ergeben sich die Faktorisierungen

$$\bar{\boldsymbol{A}} = \boldsymbol{P}\boldsymbol{A}\boldsymbol{P}^{\mathsf{T}} = \boldsymbol{L}\boldsymbol{D}\boldsymbol{L}^{\mathsf{T}} = \hat{\boldsymbol{L}}\hat{\boldsymbol{L}}^{\mathsf{T}} \tag{18}$$

mit der Permutationsmatrix $\boldsymbol{P} := \boldsymbol{T}_{n-1,s(n-1)} \cdots \boldsymbol{T}_{2,s(2)}\boldsymbol{T}_{1,s(1)}$. Dabei gilt

$$d_k \geqq l_{jk}^2 d_k + l_{j,k+1}^2 d_{k+1} + \cdots + l_{j,j-1}^2 d_{j-1} + d_j$$

bzw.

$$\hat{l}_{kk}^2 \geqq \hat{l}_{jk}^2 + \hat{l}_{j,k+1}^2 + \cdots + \hat{l}_{jj}^2 \tag{19}$$

für $1 \leqq k < j \leqq n$, insbesondere ist

$$d_1 = (\hat{l}_{11})^2 \geqq d_2 = (\hat{l}_{22})^2 \geqq \cdots \geqq d_n = (\hat{l}_{nn})^2 > 0, \tag{20}$$

siehe Ü 6.1.1. Der Quotient

$$\gamma_2 := d_1/d_n = (\hat{l}_{11}/\hat{l}_{nn})^2 \leqq \operatorname{cond}_2(\boldsymbol{A}) = [\operatorname{cond}_2(\hat{\boldsymbol{L}})]^2 \tag{21}$$

ist eine untere Schranke für $\operatorname{cond}_2(\boldsymbol{A})$, welche die Größenordnung fast immer ausreichend genau wiedergibt und einfacher zu berechnen ist als durch Schätzung von $\|\hat{\boldsymbol{L}}^{-1}\|_2$ nach 5.4.7, 5.4.8, siehe Ü 6.1.2. □

Welche der Faktorisierungen $\boldsymbol{A} = \boldsymbol{L}\boldsymbol{D}\boldsymbol{L}^{\mathsf{T}}$ oder $\boldsymbol{A} = \hat{\boldsymbol{L}}\hat{\boldsymbol{L}}^{\mathsf{T}}$ man verwenden sollte, hängt von den Umständen und auch etwas vom Geschmack ab. Die Cholesky-Faktorisierung läßt sich etwas einfacher programmieren und bietet mehr Freiheit in der Reihenfolge der Berechnung der $a_{ij}^{(k)}$; das Auftreten der n Wurzelberechnungen kann kaum als Nachteil angesehen werden. Demgegenüber ist die $\boldsymbol{L}\boldsymbol{D}\boldsymbol{L}^{\mathsf{T}}$-Faktorisierung mit der verwandten Faktorisierung desselben Typs für indefinites symmetrisches $\boldsymbol{A}$ kompatibel. Für manche Anwendungen ist es außerdem zweckmäßig, die Diagonalelemente $d_k = (\hat{l}_{kk})^2$ explizit verfügbar zu haben.

B. *Indefinite symmetrische Matrizen*

Wir wollen im folgenden untersuchen, ob die symmetrischen Faktorisierungen (2) bzw. (13) auch für reguläres indefinites $\boldsymbol{A} \in \mathbf{S}^{n,n}$ existieren und stabil berechnet werden können. Für die Cholesky-Faktorisierung (13) läßt sich diese Frage sofort verneinen, da diese Darstellung für reguläres $\hat{\boldsymbol{L}}$ die positive Definitheit von $\boldsymbol{A}$ impliziert. Dagegen ist die $\boldsymbol{LDL}^{\mathrm{T}}$-Faktorisierung potentiell geeignet, auch indefinite Matrizen darzustellen, was durch das Auftreten positiver und negativer Diagonalelemente in $\boldsymbol{D}$ gekennzeichnet ist, vgl. Ü 6.1.4. Ein Beispiel ist die Matrix

$$\boldsymbol{A} = \begin{pmatrix} -1 & -1 \\ -1 & 1 \end{pmatrix} = \begin{pmatrix} 1 & 0 \\ 1 & 1 \end{pmatrix} \begin{pmatrix} -1 & 0 \\ 0 & 2 \end{pmatrix} \begin{pmatrix} 1 & 1 \\ 0 & 1 \end{pmatrix} = \boldsymbol{LDL}^{\mathrm{T}} \tag{22}$$

mit den Eigenwerten $\lambda_{1,2} = \pm\sqrt{2}$. Die Zerlegung (22) kann nach 6.1.1 berechnet werden, wenn dort der Test „$a_{kk} \leqq 0$" durch „$a_{kk} = 0$" ersetzt wird. Andererseits zeigt die Matrix

$$\boldsymbol{A} = \begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix} \tag{23}$$

mit den Eigenwerten $\lambda_{1,2} = \pm 1$, daß nicht jede reguläre indefinite symmetrische Matrix eine Zerlegung $\boldsymbol{A} = \boldsymbol{LDL}^{\mathrm{T}}$ mit regulärem $\boldsymbol{D}$ besitzt. Algorithmus 6.1.1 bricht hier bereits im ersten Schritt wegen $d_1 = 0$ ab. Dies läßt sich auch nicht durch Diagonalpivotisierung nach 6.1.5, d. h. durch Zulassung symmetrieerhaltender Permutationen gemäß

$$\tilde{\boldsymbol{A}} = \boldsymbol{PAP}^{\mathrm{T}} = \boldsymbol{LDL}^{\mathrm{T}} \tag{24}$$

beheben, denn die Matrix (23) ist unter solchen simultanen Zeilen- und Spaltenvertauschungen invariant.

Das Versagen eines Algorithmus für ein spezielles Problem — hier die Matrix (23) — bedeutet meist, daß auch im Fall der Durchführbarkeit für benachbarte Probleme mit instabilem Verhalten zu rechnen ist. Diese Erfahrung bestätigt sich auch hier:

6.1.6. *Beispiel.* Es sei $0 < \varepsilon \ll 1$ und

$$\boldsymbol{A} := \begin{pmatrix} \varepsilon & 1 \\ 1 & \varepsilon \end{pmatrix}, \quad \boldsymbol{b} := \begin{pmatrix} 1 \\ 0 \end{pmatrix}.$$

Die Algorithmen 6.1.1 und 6.1.2(iii) mögen mit dem oben beschriebenen modifiziertem Test so durchgeführt werden, daß $\mathrm{fl}(\alpha + \varepsilon^2\alpha) = \alpha$ für $\alpha \neq 0$ gesetzt, aber alle übrigen Operationen exakt durchgeführt werden. Dies entspricht qualitativ einer Arithmetik mit $\varepsilon^2 < \nu/2$. Bei dieser Vereinbarung ergibt sich

$$\boldsymbol{D} = \begin{pmatrix} \varepsilon & 0 \\ 0 & -1/\varepsilon \end{pmatrix}, \quad \boldsymbol{L} = \begin{pmatrix} 1 & 0 \\ 1/\varepsilon & 1 \end{pmatrix}, \quad \boldsymbol{x} = \begin{pmatrix} 0 \\ 1 \end{pmatrix};$$

die exakten Größen sind

$$\boldsymbol{D}^* = \begin{pmatrix} \varepsilon & 0 \\ 0 & \varepsilon - 1/\varepsilon \end{pmatrix}, \quad \boldsymbol{L}^* = \begin{pmatrix} 1 & 0 \\ 1/\varepsilon & 1 \end{pmatrix}, \quad \boldsymbol{x}^* = \frac{1}{1-\varepsilon^2} \begin{pmatrix} -\varepsilon \\ 1 \end{pmatrix}.$$

Für den allein durch die Rundung $\mathrm{fl}(\varepsilon - 1/\varepsilon) = -1/\varepsilon$ bei der Berechnung von d_2 hervorgerufenen Fehler $\delta\boldsymbol{x} = \boldsymbol{x} - \boldsymbol{x}^*$ gilt dann $\|\delta\boldsymbol{x}\|_\infty = \varepsilon + O(\varepsilon^3)$. Der Darstellungsfehler $\delta\boldsymbol{A}_D$ von $\boldsymbol{A}$ zur Maschinengenauigkeit ν kann dagegen höchstens zu einem Fehler

$$\|\delta\boldsymbol{x}\|_\infty \leqq \|\boldsymbol{A}^{-1}\|_\infty \, \|\delta\boldsymbol{A}_D\|_\infty \, \|\boldsymbol{x}^*\|_\infty \leqq \nu \doteqdot \Delta\boldsymbol{x}_{\mathrm{opt}}(A, \nu)$$

führen. Im Fall $\varepsilon^2 = \nu/3$ überschreitet daher der erzeugte Rundungsfehler das optimale Fehlerniveau um den Faktor $\varepsilon/\nu = 1/\sqrt{3\nu}$. Da der Quotient nicht unabhängig von ν beschränkt werden kann, liegt Instabilität vor. Die Lösung von $\boldsymbol{A}\boldsymbol{x} = \boldsymbol{b}$ mittels spaltenpivotisierter $\boldsymbol{LR}$-Faktorisierung würde auf die berechneten Faktoren

$$\boldsymbol{L} = \begin{pmatrix} 1 & 0 \\ \varepsilon & 1 \end{pmatrix}, \boldsymbol{R} = \begin{pmatrix} 1 & \varepsilon \\ 0 & 1 \end{pmatrix} \text{ von } \bar{\boldsymbol{A}} = \boldsymbol{P}\boldsymbol{A} = \begin{pmatrix} 1 & \varepsilon \\ \varepsilon & 1 \end{pmatrix}, \quad \text{also auf} \quad \boldsymbol{x} = \begin{pmatrix} -\varepsilon \\ 1 \end{pmatrix}$$

führen. Der erzeugte Rundungsfehler ist hier $\|\delta\boldsymbol{x}\|_\infty = \varepsilon^2 + O(\varepsilon^3) \doteqdot \nu/3$ und liegt wie erwartet in der Größenordnung von $\Delta\boldsymbol{x}_{\mathrm{opt}}(\boldsymbol{A}, \nu)$. □

Das beschriebene instabile Verhalten ist darauf zurückzuführen, daß die Diagonalelemente der Matrizen $\boldsymbol{M}^{(k)}$ im Unterschied zum positiv definiten Fall im Laufe der Faktorisierung beliebig klein im Vergleich zu den Nichtdiagonalelementen werden können. Dies führt zu einem unbeschränkten Wachstum der $\boldsymbol{M}^{(k)}$, im Beispiel ist $\boldsymbol{M}^{(2)} = (1/\varepsilon)$. Die zur Sicherung der Stabilität daher notwendige Zulassung von Nichtdiagonalpivots zerstört dagegen i. allg. die Symmetrie und verhindert die gewünschte Aufwandshalbierung.

Als Ausweg aus dieser Situation bietet sich an, einem Eliminationsschritt mit dem Nichtdiagonalpivot $a_{s\hat{s}}^{(k)}$ $(s \neq \hat{s})$ sofort einen weiteren mit dem spiegelbildlich gelegenen Pivot $a_{\hat{s}s}^{(k+1)}$ folgen zu lassen. Nach einem solchen Doppelschritt ist die Restmatrix $\boldsymbol{M}^{(k+2)}$ wieder symmetrisch. Dieses Vorgehen ist im wesentlichen identisch mit der simultanen Elimination des Variablenpaares $\{x_s, x_{\hat{s}}\}$, siehe Ü 6.1.6.

Zur Beschreibung einer solchen *Blockelimination* betrachten wir einen ersten Schritt, bei dem der Block (x_1, x_p) der ersten p Variablen eliminiert werden soll, und setzen dazu

$$\boldsymbol{A} = \left(\begin{array}{c|c} \boldsymbol{D} & \boldsymbol{C}^{\mathsf{T}} \\ \hline \boldsymbol{C} & \boldsymbol{B} \end{array}\right), \quad \boldsymbol{L}_1 = \left(\begin{array}{c|c} \boldsymbol{I}_p & \boldsymbol{O} \\ \hline -\boldsymbol{G} & \boldsymbol{I}_{n-p} \end{array}\right); \tag{25}$$

dabei ist $\boldsymbol{D} \in \mathbf{S}^{p,p}$, $\boldsymbol{C}, \boldsymbol{G} \in \mathbf{R}^{n-p,p}$, $\boldsymbol{B} \in \mathbf{S}^{n-p,n-p}$ mit $p \in \{1, 2\}$. Die Matrix $\boldsymbol{G}$ der Eliminationskoeffizienten soll so bestimmt werden, daß

$$\boldsymbol{L}_1\boldsymbol{A} = \left(\begin{array}{c|c} \boldsymbol{D} & \boldsymbol{C}^{\mathsf{T}} \\ \hline \boldsymbol{C} - \boldsymbol{G}\boldsymbol{D} & \boldsymbol{B} - \boldsymbol{G}\boldsymbol{C}^{\mathsf{T}} \end{array}\right) = \left(\begin{array}{c|c} \boldsymbol{D} & \boldsymbol{C}^{\mathsf{T}} \\ \hline \boldsymbol{O} & \boldsymbol{M} \end{array}\right) \tag{26}$$

gilt. Dies ist für beliebiges C genau dann möglich, wenn der Pivotblock D regulär ist. Falls letzteres zutrifft, folgt

$$G = CD^{-1}, \qquad M = B - GC^{\mathsf{T}} = B - CD^{-1}C^{\mathsf{T}},$$

insbesondere ist also M wie A symmetrisch. Um die volle Symmetrie bereits im ersten Schritt sichtbar zu machen, führen wir noch eine Rechtsmultiplikation mit L_1^{T} aus und erhalten

$$\bar{A} := L_1 A L_1^{\mathsf{T}} = \left(\begin{array}{c|c} D & C^{\mathsf{T}} \\ \hline O & M \end{array}\right)\left(\begin{array}{c|c} I_p & -G^{\mathsf{T}} \\ \hline O & I_{n-p} \end{array}\right) = \left(\begin{array}{c|c} D & C^{\mathsf{T}} - DG^{\mathsf{T}} \\ \hline O & M \end{array}\right) = \left(\begin{array}{c|c} D & O \\ \hline O & M \end{array}\right), \tag{27}$$

denn wegen der Symmetrie von D ist $C^{\mathsf{T}} - DG^{\mathsf{T}} = O$. Selbstverständlich braucht die Rechtsmultiplikation mit L_1^{T} nicht explizit durchgeführt zu werden, da sie lediglich die Symmetrie in $\bar{A}$ erzeugt und die signifikanten, bereits in $L_1 A$ vorkommenden Blöcke D und M nicht mehr verändert.

Die Regularität von D muß bei diesem Vorgehen gegebenenfalls durch geeignete symmetrieerhaltende Zeilen- und Spaltenvertauschungen, d. h. durch Übergang

$$A \to TAT^{\mathsf{T}}$$

mit einer Permutationsmatrix T erzwungen werden: Falls ein $a_{ss} \neq 0$ existiert, kann dieses mit der Festlegung $T := T_{1s}$ in die Position (1, 1) gebracht und ein einfacher Eliminationsschritt mit $p = 1$ ausgeführt werden. Existiert kein nichtverschwindendes Diagonalelement, so wird a_{s1} $(2 \leqq s \leqq n)$ in der ersten Spalte von A mit $\delta := a_{11}a_{ss} - a_{s1}^2 \neq 0$ gesucht und mittels $T := T_{2s}$ in die Position (2, 1) gebracht. Ein solches Element existiert für reguläres A stets, denn andernfalls wäre $a_{s1} = 0$ für alle $s = 1, \ldots, n$. Dann ist $\det(D) = \delta \neq 0$, und es wird ein Blockeliminationsschritt mit $p = 2$ ausgeführt.

In beiden Fällen $p = 1$ bzw. $p = 2$ ist die Restmatrix M wie A selbst regulär und symmetrisch, so daß derselbe Eliminationsprozeß auf M statt A angewendet und A folglich sukzessive auf obere Blockdiagonalform mit Diagonalblöcken der Dimension 1 oder 2 transformiert werden kann.

Beginnend mit $A^{(1)} := A$, sei die Blockdiagonalform in den Zeilen und Spalten 1 bis $k - 1$ bereits erzeugt worden, d. h., $A^{(k)}$ sei von der Gestalt

$$A^{(k)} = \left(\begin{array}{c|ccc} D^{(1)} & & & O \\ & \square & & \\ & & \square & \\ O & & & M^{(k)} \end{array}\right) = \left(\begin{array}{c|ccc} D^{(1)} & & & O \\ & \square & & \\ & & \square & \\ O & & & \begin{array}{c|c} D^{(k)} & C^{(k)\mathsf{T}} \\ \hline C^{(k)} & B^{(k)} \end{array} \end{array}\right). \tag{28}$$

Im Schritt k werden dann $p = p(k) \in \{1, 2\}$ und eine Vertauschungsmatrix $T_k := T_{t(k),s(k)}$ mit $k \leqq t(k) \leqq s(k) \leqq n$ so festgelegt, daß der (p, p)-Pivotblock $D^{(k)}$ von

$$A^{(k)} := T_k A^{(k)} T_k \tag{29}$$

regulär ist; man beachte $\boldsymbol{T}_k = \boldsymbol{T}_k^{\mathsf{T}}$. Zur Vereinfachung der Schreibweise bezeichnen wir die durch die Vertauschungen (29) entstandene Matrix wieder mit $\boldsymbol{A}^{(k)}$ und nicht mit $\hat{\boldsymbol{A}}^{(k)}$ wie in 5.1, 5.2. Der nachfolgende, wie oben symmetrisierte Eliminationsschritt

$$\boldsymbol{A}^{(k+p)} := \boldsymbol{L}_k \boldsymbol{A}^{(k)} \boldsymbol{L}_k^{\mathsf{T}} \tag{30}$$

mit der Blockeliminationsmatrix

$$\boldsymbol{L}_k := \begin{pmatrix} \boldsymbol{I}_{k-1} & & \\ & \boldsymbol{I}_p & \\ & -\boldsymbol{G}_k & \boldsymbol{I}_{n-p-k+1} \end{pmatrix}, \quad \boldsymbol{G}_k = \begin{pmatrix} l_{k+1,k} \\ \vdots \\ l_{nk} \end{pmatrix} \quad \text{für} \quad p = 1,$$

$$\boldsymbol{G}_k = \begin{pmatrix} l_{k+2,k} & l_{k+2,k+1} \\ \vdots & \vdots \\ l_{nk} & l_{n,k+1} \end{pmatrix} \quad \text{für} \quad p = 2, \tag{31}$$

erzeugt bei der Festlegung

$$\boldsymbol{G}_k := \boldsymbol{C}^{(k)} [\boldsymbol{D}^{(k)}]^{-1} \tag{32}$$

die gewünschten Nullen auf dem Platz von $\boldsymbol{C}^{(k)}$ und transformiert $\boldsymbol{B}^{(k)}$ in die neue Restmatrix

$$\boldsymbol{M}^{(k+p)} := \boldsymbol{B}^{(k)} - \boldsymbol{G}_k \boldsymbol{C}^{(k)\mathsf{T}}. \tag{33}$$

Unter Beachtung von

$$[\boldsymbol{D}^{(k)}]^{-1} = 1/\delta_k, \qquad \delta_k := a_{kk}^{(k)} \tag{34}$$

im Fall $p = 1$ bzw.

$$[\boldsymbol{D}^{(k)}]^{-1} = \frac{1}{\delta_k} \begin{pmatrix} a_{k+1,k+1}^{(k)} & -a_{k+1,k}^{(k)} \\ -a_{k+1,k}^{(k)} & a_{kk}^{(k)} \end{pmatrix}, \quad \delta_k := a_{kk}^{(k)} a_{k+1,k+1}^{(k)} - (a_{k+1,k}^{(k)})^2 \tag{35}$$

im Fall $p = 2$ lassen sich die Eliminationskoeffizienten l_{ij} ($j = k$ bzw. $j = k, k+1$) mittels (32) explizit durch die Elemente von $\boldsymbol{A}^{(k)}$ darstellen.

Zusammenfassend erhalten wir den folgenden Basisalgorithmus:

6.1.7. *Symmetrische Blockfaktorisierung der symmetrischen Matrix* $\boldsymbol{A}$.

Initialisierung: $\boldsymbol{A}^{(1)} := \boldsymbol{A}$, $k := 1$

k-ter Schritt:

S1: Pivotwahl: Bestimme $p = p(k) \in \{1, 2\}$ und Vertauschungsmatrix $\boldsymbol{T}_k := \boldsymbol{T}_{t(k),s(k)}$, $k \leqq t(k) \leqq s(k) \leqq n$ derart, daß $\boldsymbol{A}^{(k)} := \boldsymbol{T}_k \boldsymbol{A}^{(k)} \boldsymbol{T}_k$ der Bedingung $\delta_k \neq 0$ mit δ_k gemäß (34), (35) genügt.

S2: Bestimmung der Eliminationskoeffizienten und Transformation der Matrix $\boldsymbol{A}^{(k)}$:

S2.1: Falls $p = 1$, setze

$$a_{ij}^{(k+1)} := a_{ji}^{(k+1)} := \begin{cases} 0 & \text{für } i = k+1, \ldots, n,\ j = k, \\ a_{ij}^{(k)} - l_{ik} * a_{kj}^{(k)} & \text{für } i, j = k+1, \ldots, n, \\ a_{ij}^{(k)} & \text{sonst} \end{cases}$$

mit

$$l_{ik} := a_{ik}^{(k)}/\delta_k, \qquad i = k+1, \ldots, n$$

S2.2: Falls $p = 2$ und $k < n-1$, setze

$$a_{ij}^{(k+2)} := a_{ji}^{(k+2)}$$

$$:= \begin{cases} 0 & \text{für } i = k+2, \ldots, n;\ j = k, k+1, \\ a_{ij}^{(k)} - l_{ik} * a_{kj}^{(k)} - l_{i,k+1} * a_{k+1,j}^{(k)} & \text{für } i, j = k+2, \ldots, n, \\ a_{ij}^{(k)} & \text{sonst} \end{cases}$$

mit

$$l_{ik} := (a_{ik}^{(k)} * a_{k+1,k+1}^{(k)} - a_{i,k+1}^{(k)} * a_{k+1,k}^{(k)})/\delta_k,$$

$$l_{i,k+1} := (-a_{ik}^{(k)} * a_{k+1,k}^{(k)} + a_{i,k+1}^{(k)} * a_{kk}^{(k)})/\delta_k \quad \text{für } i = k+2, \ldots, n.$$

S3: Indexerhöhung: Setze $k := k + p$. Falls $k < n$, gehe nach S1.

Aufwand: $\sim n^3/6$ opms (bei Ausnutzung der Symmetrie von $\boldsymbol{M}^{(k+p)}$)

Bei Auswahl eines (2, 2)-Pivotblockes $\boldsymbol{D}^{(k)}$ wird die Stufe $k+1$ übersprungen und sofort zu $k+2$ übergegangen. Um zu einer einheitlichen und einfachen Bezeichnung zu gelangen, setzen wir in diesem Fall $\boldsymbol{T}_{k+1} := \boldsymbol{L}_{k+1} := \boldsymbol{I}$, $\boldsymbol{A}^{(k+1)} := \boldsymbol{A}^{(k)}$. Mit dieser Vereinbarung läßt sich der Gesamtprozeß in der Form

$$\boldsymbol{D} := \boldsymbol{A}^{(n)} = \boldsymbol{L}_{n-1}\boldsymbol{T}_{n-1}\boldsymbol{L}_{n-2}\boldsymbol{T}_{n-2} \cdots \boldsymbol{L}_1\boldsymbol{T}_1\boldsymbol{A}\boldsymbol{T}_1\boldsymbol{L}_1^{\mathsf{T}} \cdots \boldsymbol{T}_{n-2}\boldsymbol{L}_{n-2}^{\mathsf{T}}\boldsymbol{T}_{n-1}\boldsymbol{L}_{n-1}^{\mathsf{T}} \tag{36}$$

darstellen, und $\boldsymbol{D}$ ist nach Konstruktion von Blockdiagonalform

$$\boldsymbol{D} = \begin{pmatrix} \boxed{\boldsymbol{D}^{(1)}} & & & \\ & \boxed{\boldsymbol{D}^{(1+p_1)}} & & \\ & & \ddots & \\ & & & \boxed{\boldsymbol{D}^{(l)}} \end{pmatrix}. \tag{37}$$

Dabei ist $l = n-1$ oder $l = n$ in Abhängigkeit davon, ob der letzte Pivotblock die Dimension 2 oder 1 hat.

Wenn die nach dem k-ten Schritt vorkommenden Vertauschungen $\boldsymbol{T}_{k+1}, \ldots, \boldsymbol{T}_{n-1}$ gemäß

$$\begin{pmatrix} \boldsymbol{O} \\ \boldsymbol{G}_k \end{pmatrix} := \boldsymbol{T}_{n-1} \cdots \boldsymbol{T}_{k+1} \begin{pmatrix} \boldsymbol{O} \\ \boldsymbol{G}_k \end{pmatrix} \quad (k = 1, \ldots, n-2) \tag{38}$$

auf die Eliminationskoeffizienten $\boldsymbol{G}_k$ angewendet werden, geht (36) mit den so ver-

tauschten Größen in

$$\boldsymbol{D} = \boldsymbol{A}^{(n)} = \boldsymbol{L}_{n-1}\boldsymbol{L}_{n-2} \cdots \boldsymbol{L}_1 \boldsymbol{T}_{n-1}\boldsymbol{T}_{n-2} \cdots \boldsymbol{T}_1 \boldsymbol{A} \boldsymbol{T}_1 \cdots \boldsymbol{T}_{n-2}\boldsymbol{T}_{n-1}\boldsymbol{L}_1^{\mathsf{T}} \cdots \boldsymbol{L}_{n-2}^{\mathsf{T}}\boldsymbol{L}_{n-1}^{\mathsf{T}},$$

folglich in

$$\boldsymbol{L}\boldsymbol{D}\boldsymbol{L}^{\mathsf{T}} := \boldsymbol{L}_1^{-1} \cdots \boldsymbol{L}_{n-1}^{-1} \boldsymbol{D} \boldsymbol{L}_{n-1}^{-\mathsf{T}} \cdots \boldsymbol{L}_1^{-\mathsf{T}} = \boldsymbol{T}_{n-1} \cdots \boldsymbol{T}_1 \boldsymbol{A} \boldsymbol{T}_1 \cdots \boldsymbol{T}_{n-1} =: \boldsymbol{P}\boldsymbol{A}\boldsymbol{P}^{\mathsf{T}} \tag{39}$$

über, vgl. 5.1.C. Wie dort entsteht dabei $\boldsymbol{L}_k^{-1}$ aus $\boldsymbol{L}_k$, indem $-\boldsymbol{G}_k$ durch $\boldsymbol{G}_k$ ersetzt wird, vgl. (31).

Als Ergebnis erhalten wir das folgende Resultat.

6.1.8. *Satz.* Algorithmus 6.1.7 ist in exakter Arithmetik für jede reguläre Matrix $\boldsymbol{A} \in \mathbf{S}^{n,n}$ durchführbar und liefert eine Faktorisierung

$$\boldsymbol{P}\boldsymbol{A}\boldsymbol{P}^{\mathsf{T}} = \boldsymbol{L}\boldsymbol{D}\boldsymbol{L}^{\mathsf{T}} \tag{40}$$

mit

$$\boldsymbol{L} = \begin{pmatrix} \boldsymbol{I}_{p(1)} & & & 0 \\ \boldsymbol{G}_1 & \boldsymbol{I}_{p(2)} & & \\ & \boldsymbol{G}_2 & \ddots & \\ & & & \boldsymbol{I}_{p(l)} \end{pmatrix} = \begin{pmatrix} \boldsymbol{I}_{p(1)} & & & 0 \\ & \boldsymbol{I}_{p(2)} & & \\ & l_{ij} & \ddots & \\ & & & \boldsymbol{I}_{p(l)} \end{pmatrix},$$

$$\boldsymbol{D} = \begin{pmatrix} \boldsymbol{D}^{(1)} & & & \\ & \boldsymbol{D}^{(1+p(1))} & & \\ & & \ddots & \\ & & & \boldsymbol{D}^{(l)} \end{pmatrix}. \tag{41}$$

Dabei ist $\boldsymbol{L}$ die Matrix der gemäß (38) vertauschten Eliminationskoeffizienten, $\boldsymbol{D}$ die Blockdiagonalmatrix der Pivotblöcke und $\boldsymbol{P} := \boldsymbol{T}_{n-1} \cdots \boldsymbol{T}_1$ die durch die Vertauschungen definierte Permutationsmatrix.

6.1.9. *Bemerkung.* (i) Algorithmus 6.1.7 kann unter Ausnutzung der Symmetrie der $\boldsymbol{M}^{(k)}$ zeilenweise auf dem Platz des unteren Dreiecks von $\boldsymbol{A}$ ausgeführt werden, wobei die l_{ij} auf dem Platz der zu 0 gemachten $a_{ij}^{(k)}$ gespeichert werden. Die Vertauschungen (38) bedeuten dann, daß die bereits berechneten l_{ij} wie die Zeilen von $\boldsymbol{A}^{(k)}$ mit vertauscht werden. Für $n = 5$ ergibt sich etwa folgendes Belegungsschema ($p_1 = 2$, $p_3 = 1$, $p_4 = 2$):

$$\begin{pmatrix} a_{11} & & & & \\ a_{21} & a_{22} & & & \\ a_{31} & a_{32} & a_{33} & & \\ a_{41} & a_{42} & a_{43} & a_{44} & \\ a_{51} & a_{52} & a_{53} & a_{54} & a_{55} \end{pmatrix} \rightarrow \begin{pmatrix} d_{11} & & & & \\ d_{21} & d_{22} & & & \\ l_{31} & l_{32} & d_{33} & & \\ l_{41} & l_{42} & l_{43} & d_{44} & \\ l_{51} & l_{52} & l_{53} & d_{54} & d_{55} \end{pmatrix}.$$

Man beachte, daß die (2,2)-Pivots $\boldsymbol{D}^{(1)}$, $\boldsymbol{D}^{(4)}$ symmetrisch sind und nur das untere Dreieck gespeichert zu werden braucht. Da die zugehörigen Diagonalblöcke von $\boldsymbol{L}$ die Gestalt $\boldsymbol{I}_2$ haben, treten keine Probleme bei der Speicherung der Außerdiagonalelemente d_{21} und d_{54} auf. Die zur Rekonstruktion von $\boldsymbol{P}$ erforderliche Information wird in einem Feld der Länge n gespeichert; für Details sei auf die Literatur verwiesen, siehe B 6.1.

(ii) Bei erfolgreicher Faktorisierung mit regulärem $\boldsymbol{D}$ kann das Gleichungssystem $\boldsymbol{Ax} = \boldsymbol{P}^{\mathsf{T}}\boldsymbol{LDL}^{\mathsf{T}}\boldsymbol{Px} = \boldsymbol{b}$ in Analogie zu 4.2.1 für jedes $\boldsymbol{b} \in \mathbf{R}^n$ unter Verwendung der Faktorisierung (40) wie folgt gelöst werden:

S2.1: Berechne $\boldsymbol{c}$ aus $\boldsymbol{Lc} = \bar{\boldsymbol{b}} := \boldsymbol{Pb}$

S2.2: Berechne $\boldsymbol{d}$ aus $\boldsymbol{Dd} = \boldsymbol{c}$

S2.3: Berechne $\bar{\boldsymbol{x}}$ aus $\boldsymbol{L}^{\mathsf{T}}\bar{\boldsymbol{x}} = \boldsymbol{d}$, setze $\boldsymbol{x} := \boldsymbol{P}^{\mathsf{T}}\bar{\boldsymbol{x}}$.

Das System $\boldsymbol{Dd} = \boldsymbol{c}$ zerfällt dabei in Teilsysteme der Dimension 1 oder 2 entsprechend der Blockstruktur von $\boldsymbol{D}$. Der Aufwand ist $\sim n^2$ opms.

(iii) Die zur Darstellung (40) benötigten Vertauschungen der l_{ij} gemäß (38) können auch weggelassen werden, d. h., 6.1.7 kann in der dort angegebenen Version auf dem Platz des unteren Dreiecks von $\boldsymbol{A}$ ausgeführt werden. Dann muß direkt mit der Darstellung (36) gearbeitet werden, die sich äquivalent in der Form

$$\boldsymbol{A} = \tilde{\boldsymbol{L}}\boldsymbol{D}\tilde{\boldsymbol{L}}^{\mathsf{T}} \quad \text{mit} \quad \tilde{\boldsymbol{L}} := (\boldsymbol{L}_{n-1}\boldsymbol{T}_{n-1} \cdots \boldsymbol{L}_1\boldsymbol{T}_1)^{-1} = \boldsymbol{T}_1\boldsymbol{L}_1^{-1}\boldsymbol{T}_2\boldsymbol{L}_2^{-1} \cdots \boldsymbol{T}_{n-1}\boldsymbol{L}_{n-1}^{-1} \tag{42}$$

schreiben läßt. Die Matrix $\tilde{\boldsymbol{L}}$ ist selbst i. allg. keine untere Dreiecksmatrix, allerdings lassen sich die signifikanten Elemente l_{ij} der einzelnen Faktoren $\boldsymbol{L}_k$ auf dem Platz des unteren Blockdreiecks von $\boldsymbol{A}$ speichern. Die Lösung von $\boldsymbol{Ax} = \tilde{\boldsymbol{L}}\boldsymbol{D}\tilde{\boldsymbol{L}}^{\mathsf{T}}\boldsymbol{x} = \boldsymbol{b}$ erfolgt dann gemäß

S2.1: Berechne $\tilde{\boldsymbol{c}}$ aus $\tilde{\boldsymbol{L}}\tilde{\boldsymbol{c}} = \boldsymbol{b}$, d. h., berechne $\tilde{\boldsymbol{c}} = \tilde{\boldsymbol{L}}^{-1}\boldsymbol{b} = \boldsymbol{L}_{n-1}\boldsymbol{T}_{n-1} \cdots \boldsymbol{L}_1\boldsymbol{T}_1\boldsymbol{b}$

S2.2: Berechne $\tilde{\boldsymbol{d}}$ aus $\boldsymbol{D}\tilde{\boldsymbol{d}} = \tilde{\boldsymbol{c}}$

S2.3: Berechne $\boldsymbol{x}$ aus $\tilde{\boldsymbol{L}}^{\mathsf{T}}\boldsymbol{x} = \tilde{\boldsymbol{d}}$, d. h., berechne $\boldsymbol{x} = \tilde{\boldsymbol{L}}^{-\mathsf{T}}\tilde{\boldsymbol{d}} = \boldsymbol{T}_1\boldsymbol{L}_1^{\mathsf{T}} \cdots \boldsymbol{T}_{n-1}\boldsymbol{L}_{n-1}^{\mathsf{T}}\tilde{\boldsymbol{d}}$

vgl. 5.2.6(i). Die Anzahl der Rechenoperationen ist identisch mit der bei der Realisierung mit $\boldsymbol{L}$ gemäß (ii).

(iv) Die Determinante von $\boldsymbol{A}$ ergibt sich zu

$$\det(\boldsymbol{A}) = \det(\boldsymbol{D}^{(1)}) \det(\boldsymbol{D}^{(1+p_1)}) \cdots \det(\boldsymbol{D}^{(l)}) = \prod_{k=1}^{n} \delta_k; \tag{43}$$

formal muß dabei $\delta_{k+1} := 1$ gesetzt werden, sofern $\boldsymbol{D}^{(k)}$ ein (2,2)-Pivot ist.

(v) Für singuläres $\boldsymbol{A}$ versagt 6.1.7, weil im Schritt S1 kein regulärer Pivotblock gefunden werden kann, oder der letzte Block $\boldsymbol{D}^{(l)}$ ist singulär. Dies kann wie in 5.2.3 abgefragt und durch einen Fehlerindikator angezeigt werden. Bei der Determinantenberechnung ist dann wie in 5.2.5 mit $\det(\boldsymbol{A}) := 0$ abzubrechen. □

Selbstverständlich ist das Verfahren 6.1.7 ohne Zusatzforderungen an die Pivotwahl für Computerrechnung ungeeignet, vgl. 6.1.6. Wie beim Gaußschen Algorithmus muß auch hier durch geeignete Wahl von $p(k)$ und $\boldsymbol{T}_k$ ein zu starkes Anwachsen der

$\boldsymbol{M}^{(k)}$ verhindert werden. In Analogie zur vollständigen Pivotisierung beim Gaußschen Algorithmus könnte versucht werden, den Betrag von δ_k über alle zugelassenen Möglichkeiten zu maximieren. Dazu muß

$$v_k := \max \{|a_{ii}^{(k)}|, |a_{ii}^{(k)}a_{jj}^{(k)} - (a_{ij}^{(k)})^2| : i, j = k, \ldots, n;\ i > j\}$$

bestimmt werden. Falls $v_k = |a_{ss}^{(k)}|$ ist, wird $p = 1$ und $\boldsymbol{T}^{(k)} = \boldsymbol{T}_{ks}$ gesetzt. Im Fall $v_k = |a_{ss}^{(k)}a_{\hat{s}\hat{s}}^{(k)} - (a_{s\hat{s}}^{(k)})^2|$ wird $p = 2$ gesetzt und $a_{s\hat{s}}^{(k)}$ in die Position $(k+1, k)$ gebracht. Dazu sind zwei simultane Zeilen- und Spaltenvertauschungen erforderlich. Zur Berechnung aller v_k werden jedoch mindestens $\sim n^3/6$ opm benötigt. Das ist etwa derselbe Aufwand wie für die Berechnung von $\boldsymbol{L}$ und $\boldsymbol{D}$ selbst. Dieses Vorgehen scheidet daher aus Aufwandsgründen aus. Wie beim Gaußschen Algorithmus lassen sich jedoch auch hier aufwandsgünstigere teilweise Pivotisierungsstrategien angeben, die zu akzeptablem Stabilitätsverhalten führen. Bei der nachfolgend beschriebenen Strategie wird pro Schritt mindestens eine, höchstens jedoch zwei Spalten von $\boldsymbol{M}^{(k)}$ abgesucht, so daß sie etwa der Spaltenpivotisierung beim Gaußschen Algorithmus entspricht.

6.1.10. *Spaltenpivotisierung nach Bunch-Kaufman-Parlett:* Es sei α mit $0 < \alpha < 1$ fest vorgegeben, und $\boldsymbol{A}^{(k)}$ bezeichne die zu Beginn des k-ten Schrittes von 6.1.7 vorliegende Matrix der Gestalt (28).

S1.1: Bestimme Index $s = s(k)$, $k \leqq s \leqq n$, mit

$$\lambda_k := |a_{sk}^{(k)}| = \max \{|a_{ik}^{(k)}| : i = k+1, \ldots, n\}$$

S1.2: **if** $|a_{kk}^{(k)}| \geqq \alpha\lambda_k$

S1.3: **then** $[p(k) := 1,\ \boldsymbol{T}_k := \boldsymbol{I}]$

else

S1.4: Bestimme $\sigma_k := \max \{|a_{is}^{(k)}| : i = k, \ldots, n;\ i \neq s\}$

S1.5: **if** $|a_{kk}^{(k)}|\,\sigma_k \geqq \alpha(\lambda_k)^2$

S1.6: **then** $[p(k) := 1,\ \boldsymbol{T}_k := \boldsymbol{I}]$

else

S1.7: **if** $|a_{ss}^{(k)}| \geqq \alpha \cdot \sigma_k$

S1.8: **then** $[p(k) := 1,\ \boldsymbol{T}_k := \boldsymbol{T}_{ks}]$

S1.9: **else** $[p(k) := 2,\ \boldsymbol{T}_k := \boldsymbol{T}_{k+1,s}]$

Der Faktorisierungsalgorithmus 6.1.7 mit der Pivotstrategie 6.1.10 wird *Bunch-Kaufman-Parlett-Faktorisierung* — kurz: *BKP-Faktorisierung* — der symmetrischen Matrix $\boldsymbol{A}$ genannt, wobei für α der unten bestimmte optimale Wert $\alpha = \alpha_0$ gewählt wird. Das Wachstumsverhalten der $\boldsymbol{M}^{(k)}$ für diesen Algorithmus wird durch die folgende Aussage charakterisiert.

6.1.11. *Aussage.* Algorithmus 6.1.7 ist für jede reguläre Matrix $\boldsymbol{A} \in \mathbf{S}^{n,n}$ mit der Pivotstrategie 6.1.10 für jedes $\alpha \in (0, 1)$ in exakter Arithmetik durchführbar. Für jeden Index $k \in \{1, \ldots, n-1\}$, für den der Schritt $k \to k + p_k$ mit $p_k \in \{1, 2\}$

durchgeführt wird, gilt dabei

$$\mu_{k+1} \leqq \mu_k(1 + 1/\alpha) \quad \text{im Fall} \quad p_k = 1 \tag{44}$$

$$\mu_{k+2} \leqq \mu_k\big(1 + 2/(1 - \alpha)\big) \quad \text{im Fall} \quad p_k = 2, \tag{45}$$

wobei

$$\mu_k := \max\{|a_{ij}^{(k)}| : i, j = k, \ldots, n\} \qquad (k = 1, \ldots, n) \tag{46}$$

gesetzt wurde. Die für den Übergang $k \to k + 2$ gültige Schranke

$$\mu_{k+2} \leqq \max\{(1 + 1/\alpha)^2, \quad 1 + 2/(1 - \alpha)\}\, \mu_k =: [\varrho(\alpha)]^2 \mu_k \tag{47}$$

wird für $\alpha = \alpha_0 := \big(1 + \sqrt{17}\big)/9 = 0.6404$ minimal und hat den Wert $\varrho_0 := \varrho(\alpha_0) = \big(1 + \sqrt{17}\big)/2 \leqq 2.57$. Insbesondere gilt bei Verwendung von $\alpha = \alpha_0$

$$\mu_k \leqq (2.57)^{k-1}\, \mu_1 \qquad (k = 1, \ldots, n). \tag{48}$$

Beweis. Wir bemerken zunächst, daß

$$\lambda_k = |a_{sk}^{(k)}| = |a_{ks}^{(k)}| \leqq \max\{|a_{is}^{(k)}| : k \leqq i \leqq n;\ i \neq s\} = \sigma_k \leqq \mu_k \tag{49}$$

gilt. Der Einfachheit halber betrachten wir im folgenden nur den ersten Schritt und lassen den oberen Index $k = 1$ weg.

Fall 1: $p = 1$. Hier gilt nach S2.1 aus 6.1.7

$$a_{ij}^{(2)} = a_{ij} - a_{i1}a_{j1}/a_{11} \qquad (i, j = 2, \ldots, n).$$

Wenn S1.3 ausgeführt wird, folgt mit S1.2 und (49)

$$|a_{ij}^{(2)}| \leqq \mu_1 + |a_{i1}|\,|a_{j1}|/|a_{11}| \leqq \mu_1 + \lambda_1^2/(\alpha\lambda_1) \leqq \mu_1(1 + 1/\alpha).$$

Wenn S1.6 ausgeführt wird, ergibt sich mit S1.5 und (49)

$$|a_{ij}^{(2)}| \leqq \mu_1 + \lambda_1^2/(\alpha\lambda_1^2/\sigma_1) = \mu_1 + \sigma_1/\alpha \leqq \mu_1(1 + 1/\alpha).$$

Wird schließlich S1.8 ausgeführt, so folgt aus S1.7 — man beachte die Vertauschungen —

$$|a_{ij}^{(2)}| \leqq \mu_1 + \sigma_1^2/(\alpha\sigma_1) = \mu_1 + \sigma_1/\alpha \leqq \mu_1(1 + 1/\alpha),$$

womit (44) bewiesen ist.

Fall 2: $p = 2$. Hier ist nach S2.2 aus 6.1.7

$$a_{ij}^{(3)} = a_{ij} - [(a_{i1}a_{22} - a_{i2}a_{21})\, a_{j1} + (-a_{i1}a_{21} + a_{i2}a_{11})\, a_{j2}]/\delta_1 \qquad (i, j = 3, \ldots, n).$$

Wegen S1.9, S1.5 und S1.7 gilt

$$|a_{21}| = \lambda_1, \qquad |a_{11}| < \alpha\lambda_1^2/\sigma_1, \qquad |a_{22}| < \alpha\sigma_1, \tag{50}$$

folglich

$$|\delta_1| = -\delta_1 = a_{21}^2 - a_{11}a_{22} \geqq \lambda_1^2 - (\alpha\lambda_1^2/\sigma_1)\,(\alpha\sigma_1) = \lambda_1^2(1 - \alpha^2) > 0, \tag{51}$$

insbesondere ist also $\delta_1 = \det(\boldsymbol{D}^{(1)})$ negativ. Unter Beachtung von (50), (51) und $|a_{i1}| \leqq \lambda_1$, $|a_{i2}| \leqq \sigma_1$ $(i = 3, \ldots, n)$ erhalten wir

$$\begin{aligned} |a_{ij}^{(3)}| &\leqq \mu_1 + [(\lambda_1\alpha\sigma_1 + \sigma_1\lambda_1)\,\lambda_1 + (\lambda_1^2 + \sigma_1(\alpha\lambda_1^2/\sigma_1))\,\sigma_1]/[\lambda_1^2(1 - \alpha^2)] \\ &= \mu_1 + 2\sigma_1/(1 - \alpha) \leqq \mu_1\big(1 + 2/(1 - \alpha)\big), \end{aligned}$$

also (45). Die Abschätzung (47) ergibt sich aus (44), (45) unter Beachtung der Tatsache, daß zwei aufeinanderfolgende Schritte mit einem (1,1)-Pivot ein durch

$$\mu_{k+2} \leqq (1 + 1/\alpha)\, \mu_{k+1} \leqq (1 + 1/\alpha)^2\, \mu_k$$

beschränktes Wachstum hervorrufen. Das Minimum von $\varrho(\alpha)$ wird im Fall

$$(1 + 1/\alpha)^2 = 1 + 2/(1 - \alpha)$$

erreicht, was auf die quadratische Gleichung $4\alpha^2 - \alpha - 1 = 0$ mit der Wurzel $\alpha_0 \in (0, 1)$ führt. □

6.1.12. *Rundungsfehleranalyse.* Die reguläre Matrix $\boldsymbol{A} \in \mathbb{C}^{n,n}$ genüge der Bedingung

$$\varkappa = \nu F \operatorname{cond}_\infty(\boldsymbol{A}) < 1 \quad \text{mit} \quad F := 5.5n^2\mu_{\max}/\|\boldsymbol{A}\|_\infty \leqq 2.2n^2(2.57)^n, \tag{52}$$

wobei $\mu_{\max} := \max\{\mu_k : 1 \leqq k \leqq n\}$ ist mit μ_k gemäß (46). Dann ist die BKP-Faktorisierung durchführbar. Zu den berechneten Faktoren $\boldsymbol{L}, \boldsymbol{D} \in \mathbb{R}^{n,n}$ existiert eine symmetrische Störung $\delta\boldsymbol{A}$ mit

$$\boldsymbol{P}(\boldsymbol{A} + \delta\boldsymbol{A})\,\boldsymbol{P}^T = \boldsymbol{L}\boldsymbol{D}\boldsymbol{L}^T \quad \text{und} \quad \|\delta\boldsymbol{A}\|_\infty \leqq \nu F\, \|\boldsymbol{A}\|_\infty. \tag{53}$$

Die zu $\boldsymbol{b} \in \mathbb{R}^n$ gehörende und gemäß 6.1.9(ii) bzw. (iii) berechnete Lösung $\boldsymbol{x}$ von $\boldsymbol{A}\boldsymbol{x} = \boldsymbol{b}$ ist die exakte Lösung des durch eine Matrix $\delta_1\boldsymbol{A}$ gestörten Systems

$$(\boldsymbol{A} + \delta_1\boldsymbol{A})\,\boldsymbol{x} = \boldsymbol{b}, \quad \text{wobei} \quad \|\delta_1\boldsymbol{A}\|_\infty \leqq \nu F_1\, \|\boldsymbol{A}\|_\infty \text{ mit } F_1 \leqq (n + 1)\, F \text{ ist.} \tag{54}$$

6.1.13. *Bemerkung.* (i) Das nach 6.1.11 eventuell mögliche exponentielle Wachstum der μ_k tritt praktisch fast nie auf; es gilt fast immer

$$\mu_k \leqq \gamma\mu_1, \quad \text{also} \quad F \leqq 5.5\gamma n^2 \tag{55}$$

mit γ in der Größenordnung von 1. *Die BKP-Faktorisierung und die Lösung von* $\boldsymbol{A}\boldsymbol{x} = \boldsymbol{b}$ *mittels der Faktoren* $\boldsymbol{L}$ *und* $\boldsymbol{D}$ *sind also numerisch gutartige Prozesse* mit den Kumulationskonstanten F nach (52), (55) und F_1 nach (54).

(ii) Bei der Pivotstrategie 6.1.10 kann ein (2,2)-Pivotblock nur für indefinites $\boldsymbol{A}$ auftreten, denn jedem solchen Block entspricht ein Eigenwertpaar aus einem positiven und einem negativen Eigenwert, siehe Ü 6.1.5. Für positiv definites $\boldsymbol{A}$ ist daher $\boldsymbol{D}$ eine positive Diagonalmatrix, d. h., die BKP-Faktorisierung ist dann identisch mit der $\boldsymbol{L}\boldsymbol{D}\boldsymbol{L}^T$-Faktorisierung 6.1.1, wobei allerdings eine spezielle Form der Diagonalpivotisierung verwendet wird. Man beachte, daß in 6.1.10 auch im positiv definiten Fall $\boldsymbol{T}_k = \boldsymbol{T}_{ks}$ mit $k < s \leqq n$ gewählt werden kann und folglich $a_{kk}^{(k)}$ mit $a_{ss}^{(k)}$ vertauscht wird. □

6.1.14. *Bemerkung.* (i) Sämtliche in diesem Abschnitt beschriebenen symmetrischen Faktorisierungen können bzw. werden auf dem Platz des unteren Dreiecks von $\boldsymbol{A}$ ausgeführt. Selbstverständlich kann auch mit $\boldsymbol{R} := \boldsymbol{L}^T$ auf dem Platz des oberen Dreiecks gearbeitet werden. Dies ist bei Implementierung in FORTRAN i. allg. günstiger, weil die Algorithmen auf eine zeilenweise Berechnung von $\boldsymbol{L}$, also eine spaltenweise Berechnung von $\boldsymbol{R}$ orientiert sind und Felder in FORTRAN spaltenweise gespeichert werden.

(ii) Das jeweils nicht benötigte Dreieck von $\boldsymbol{A}$ kann zum Speichern der Originalmatrix $\boldsymbol{A}$ benutzt werden, wenn diese später etwa für die iterative Verbesserung von $\boldsymbol{x}$ analog zu 5.4 benötigt wird. Die Diagonalelemente müssen dann gesondert aufgehoben werden. Falls $\boldsymbol{A}$ nicht benötigt wird, kann das obere bzw. untere Dreieck zeilen- oder spaltenweise in einem eindimensionalen Feld der Länge $n(n+1)/2$ kompakt gespeichert werden. Wegen der komplizierteren Indexrechnung werden die Programme dabei unübersichtlicher, die Ausführungszeiten unterscheiden sich bei geeigneter Implementierung kaum von den mit zweidimensionalen Feldern arbeitenden Implementierungen. □

Übungsaufgaben

Ü 6.1.1. Man zeige, daß für das Cholesky-Verfahren 6.1.4

$$a_{jj}^{(j)} = a_{jj}^{(k)} - \hat{l}_{jk}^2 - \hat{l}_{j,k+1}^2 - \cdots - \hat{l}_{j,j-1}^2 \qquad (j > k)$$

gilt, und folgere hieraus (19) und (20).

Ü 6.1.2. Es sei $\hat{\boldsymbol{L}} = (\hat{l}_{ij})$ eine untere Dreiecksmatrix. Man zeige

$$|\hat{l}_{ii}| \leqq \|\hat{\boldsymbol{L}}\|_p \qquad (i = 1, \ldots, n)$$

und für reguläres $\hat{\boldsymbol{L}}$

$$1/|\hat{l}_{nn}| \leqq \|\hat{\boldsymbol{L}}^{-1}\|_p$$

für jede der Normen $p \in \{1, 2, \infty\}$ und leite hieraus (21) ab. Gilt (21) auch für $p \in \{1, \infty\}$?

Ü 6.1.3. Man gebe zu 6.1.1 bzw. 6.1.4 analoge Algorithmen zur $\boldsymbol{LDL}^T$- bzw. $\hat{\boldsymbol{L}}\hat{\boldsymbol{L}}^T$-Faktorisierung mit Diagonalpivotisierung entsprechend 6.1.5(iii) an.

Ü 6.1.4. Die symmetrische Matrix $\boldsymbol{A}$ besitze die Faktorisierung $\boldsymbol{A} = \boldsymbol{LDL}^T$ mit einer unteren Einsdreiecksmatrix $\boldsymbol{L}$ und einer Diagonalmatrix $\boldsymbol{D} = \operatorname{diag}(d_i)$. Man zeige, daß die Anzahl der positiven, verschwindenden bzw. negativen Eigenwerte von $\boldsymbol{A}$ gleich der Anzahl der positiven, verschwindenden bzw. negativen Diagonalelemente von $\boldsymbol{D}$ ist, und folgere hieraus, daß $\boldsymbol{A}$ genau dann positiv definit ist, wenn $\boldsymbol{D}$ nur positive Diagonalelemente besitzt.

Hinweis: Man beachte 1.2.C und die Tatsache, daß $\boldsymbol{A}$ und $\boldsymbol{D}$ kongruent sind.

Ü 6.1.5. Es sei $\boldsymbol{A}$ symmetrisch und regulär mit der BKP-Faktorisierung $\boldsymbol{PAP}^T = \boldsymbol{LDL}^T$ gemäß 6.1.7/6.1.10. Man zeige:

(i) Wenn in $\boldsymbol{D}$ ein (2,2)-Block $\boldsymbol{D}^{(k)}$ auftritt, besitzt dieser einen Eigenwert $\lambda_k > 0$ und einen Eigenwert $\lambda_{k+1} < 0$.

Hinweis: Man beachte $\det(\boldsymbol{D}^{(k)}) = \delta_k < 0$, vgl. (51).

(ii) Die Anzahl der positiven Eigenwerte von $\boldsymbol{A}$ ist gleich der Anzahl der positiven (1,1)-Blöcke von $\boldsymbol{D}$ plus der Anzahl der (2,2)-Blöcke; die Anzahl der negativen Eigenwerte von $\boldsymbol{A}$ ist gleich der Anzahl der negativen (1,1)-Blöcke plus der Anzahl der (2,2)-Blöcke von $\boldsymbol{D}$.

Ü 6.1.6. Es sei $\boldsymbol{A}$ wie in (25) mit $p = 2$, $a_{21} \neq 0$ und $\delta = \det(\boldsymbol{D}) = a_{11}a_{22} - a_{21}^2 \neq 0$. Man führe von $\boldsymbol{A}$ ausgehend zwei einfache Eliminationsschritte mit den Pivots (2, 1) und (1, 2) gemäß

$$\boldsymbol{A}^{(2)} := \boldsymbol{L}_2(-\boldsymbol{l}^2)\,\boldsymbol{A} = [\boldsymbol{I} - \boldsymbol{l}^2\boldsymbol{e}^{2T}]\,\boldsymbol{A},$$

$$\boldsymbol{A}^{(3)} := \boldsymbol{L}_1(-\boldsymbol{l}^1)\,\boldsymbol{A}^{(2)} = [\boldsymbol{I} - \boldsymbol{l}^1\boldsymbol{e}^{1T}]\,\boldsymbol{A}^{(2)}$$

mit

$$\boldsymbol{l}^2 := (l_{12}, 0, l_{32}, \ldots, l_{n2})^\mathsf{T}, \qquad l_{i2} := a_{i2}/a_{21} \qquad (i = 1, \ldots, n;\ i \neq 2),$$
$$\boldsymbol{l}^1 := (0, 0, l_{31}, \ldots, l_{n1})^\mathsf{T}, \qquad l_{i1} := a_{i1}^{(2)}/a_{12}^{(2)} \qquad (i = 3, \ldots, n)$$

durch und zeige, daß $\boldsymbol{A}^{(3)} = \bar{\boldsymbol{L}}_1 \boldsymbol{A}$ mit

$$\boldsymbol{A}^{(3)} = \left(\begin{array}{c|c} \bar{\boldsymbol{D}} & \bar{\boldsymbol{C}} \\ \hline \boldsymbol{O} & \boldsymbol{M} \end{array}\right), \quad \bar{\boldsymbol{L}}_1 = \left(\begin{array}{c|c} \boldsymbol{F} & \boldsymbol{O} \\ \hline -\boldsymbol{G} & \boldsymbol{I}_{n-2} \end{array}\right), \quad \boldsymbol{F} = \begin{pmatrix} 1 & l_{12} \\ 0 & 1 \end{pmatrix}, \quad \bar{\boldsymbol{D}} = \begin{pmatrix} 0 & a_{12}^{(2)} \\ a_{21} & a_{22} \end{pmatrix}$$

sowie

$$\bar{\boldsymbol{D}} = \boldsymbol{F}\boldsymbol{D}, \qquad \bar{\boldsymbol{C}}^\mathsf{T} = \boldsymbol{F}\boldsymbol{C}^\mathsf{T}$$

gilt. Dabei sind $\boldsymbol{M}$ und $\boldsymbol{G}$ die aus dem Blockeliminationsschritt (26) herrührenden Matrizen. Die Restmatrix $\boldsymbol{M}$ und die Eliminationskoeffizienten l_{ij} $(i = 3, \ldots, n;\ j = 1, 2)$ sind also dieselben wie bei der Blockelimination von $\{x_1, x_2\}$, während der Diagonalblock $\boldsymbol{D}$ auf die Gestalt $\bar{\boldsymbol{D}}$ transformiert wird.

Hinweis: Man beachte $\bar{\boldsymbol{L}}_1 = \boldsymbol{L}_1(-\boldsymbol{l}^1)\,\boldsymbol{L}_2(-\boldsymbol{l}^2) = \boldsymbol{I} - \boldsymbol{l}^1\boldsymbol{e}^{1\mathsf{T}} - (\boldsymbol{l}^2 - l_{12}\boldsymbol{l}^1)\,\boldsymbol{e}^{2\mathsf{T}}$.

6.2. Direkte Dreiecksfaktorisierungen

A. *Nichtsymmetrische Matrizen*

Im Abschnitt 5.1 wurde gezeigt, daß der Gaußsche Algorithmus mit Pivotsuche in den Spalten zu jedem regulären $\boldsymbol{A} = (a_{ij}) \in \mathbf{R}^{n,n}$ eine Faktorisierung

$$\boldsymbol{P}\boldsymbol{A} = \boldsymbol{L}\boldsymbol{R} \tag{1}$$

mit Dreiecksmatrizen

$$\boldsymbol{L} = \begin{pmatrix} 1 & & & 0 \\ & \ddots & & \\ l_{i1} & \ldots & l_{i,i-1} & \ddots \\ \vdots & & & \ddots \\ l_{n1} & \ldots & \ldots & l_{n,n-1}\ 1 \end{pmatrix}, \qquad \boldsymbol{R} = \begin{pmatrix} r_{11} & \ldots & r_{1j} & \ldots & r_{1n} \\ & \ddots & \vdots & & \vdots \\ & & r_{jj} & \ldots & r_{jn} \\ & & & \ddots & \vdots \\ & & & & r_{nn} \end{pmatrix},$$

$$r_{jj} \neq 0 \qquad (j = 1, \ldots, n) \tag{2}$$

und einer Permutationsmatrix $\boldsymbol{P}$ liefert, siehe 5.1.4. Im folgenden werden wir die Elemente $\{l_{ij}, r_{ij}\}$ von $\boldsymbol{L}$ und $\boldsymbol{R}$ direkt aus der Gleichung (1) bestimmen. Zur Vereinfachung betrachten wir zunächst den Fall $\boldsymbol{P} = \boldsymbol{I}$, d. h., wir setzen voraus, daß die Zerlegung

$$\boldsymbol{A} = \boldsymbol{L}\boldsymbol{R} \tag{3}$$

mit $\boldsymbol{L}$, $\boldsymbol{R}$ gemäß (2) existiert. Bedingungen dafür sind in Ü 5.2.5 angegeben worden, siehe auch 5.2.D. Unter Beachtung der Dreiecksgestalt von $\boldsymbol{L}$, $\boldsymbol{R}$ liest sich (3) elementweise als

$$a_{ij} = \sum_{p=1}^{n} l_{ip} r_{pj} = \sum_{p=1}^{\min\{i,j\}} l_{ip} r_{pj} \qquad (i, j = 1, \ldots, n),$$

also $$a_{ij} = \sum_{p=1}^{i} l_{ip} r_{pj} = \sum_{p=1}^{i-1} l_{ip} r_{pj} + r_{ij} \qquad (i \leqq j)$$

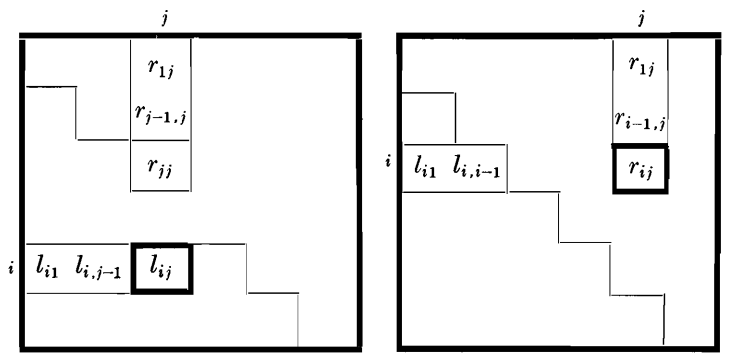
bzw. $$a_{ij} = \sum_{p=1}^{j} l_{ip} r_{pj} = \sum_{p=1}^{j-1} l_{ip} r_{pj} + l_{ij} r_{jj} \qquad (i > j).$$

Auflösen nach r_{ij} bzw. l_{ij} liefert

$$r_{ij} = a_{ij} - \sum_{p=1}^{i-1} l_{ip} r_{pj} \qquad (i \leqq j) \tag{4}$$

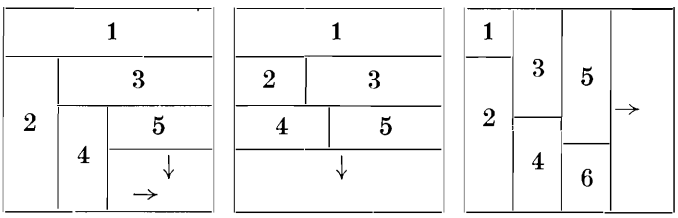
bzw. $$l_{ij} = (a_{ij} - \sum_{p=1}^{j-1} l_{ip} r_{pj}) / r_{jj} \qquad (i > j). \tag{5}$$

Die Gleichungen (4), (5) enthalten die Produkte der gesuchten Elemente $\{l_{ij}, r_{ij}\}$, allerdings in einer ganz speziellen Form: Auf den rechten Seiten kommen nur die links von der Position (i, j) stehenden Elemente der i-ten Zeile von $\boldsymbol{L}$ und die oberhalb dieser Position stehenden Elemente der j-ten Spalte von $\boldsymbol{R}$ vor. Wenn die nichttrivialen Elemente von $\boldsymbol{L}$ und $\boldsymbol{R}$ gemäß $(\boldsymbol{L} \setminus \boldsymbol{R})$ in einer (n, n)-Matrix zusammengefaßt werden, ergibt sich folgende Situation:



Fall $i > j$ Fall $i \leqq j$

Die l_{ij} und r_{ij} können daher sukzessive aus (4), (5) in jeder Reihenfolge berechnet werden, bei der links und oberhalb der Position (i, j) bereits alle Elemente vorhanden sind. Unter den in diesem Sinne zulässigen Reihenfolgen haben die folgenden praktische Bedeutung erlangt:



(R_1) (R_2) (R_3)

In (R_1) werden die Zeilen von $\boldsymbol{R}$ und die Spalten von $\boldsymbol{L}$ abwechselnd berechnet, in (R_2) bzw. (R_3) wird ($\boldsymbol{L} \setminus \boldsymbol{R}$) zeilen- bzw. spaltenweise aufgebaut. Als Beispiel geben wir für die Reihenfolge (R_1) die detaillierten Formeln an.

6.2.1. *Direkte $\boldsymbol{LR}$-Faktorisierung ohne Pivotisierung.* Die reguläre Matrix $\boldsymbol{A} \in \mathbf{R}^{n,n}$ besitze die $\boldsymbol{LR}$-Faktorisierung (2), (3). Dann ist das Verfahren

for $k := 1(1)n$ **do**

$$\left|\begin{array}{l} \textbf{for } j := k(1)n \textbf{ do } r_{kj} := a_{kj} - \sum_{p=1}^{k-1} l_{kp} * r_{pj} \\ \textbf{for } i := k+1(1)n \textbf{ do } l_{ik} := \left(a_{ik} - \sum_{p=1}^{k-1} l_{ip} * r_{pk}\right) \Big/ r_{kk} \end{array}\right.$$

in exakter Arithmetik durchführbar und liefert die nichttrivialen Elemente von $\boldsymbol{L}$ und $\boldsymbol{R}$.

In der deutschsprachigen Literatur wird das Verfahren 6.2.1 häufig nach Gauss-Banachiewicz benannt, siehe B 6.2. Es entsteht aus der stufenweisen Grundform des Gaußschen Algorithmus, indem alle zum Element l_{ij} bzw. r_{ij} beitragenden Operationen zusammengefaßt und in der Form (4) bzw. (5) geschrieben werden. Man spricht deshalb auch vom *verketteten Gaußschen Algorithmus*. Insbesondere werden in 6.2.1 dieselben Rechenoperationen wie in 5.1 ausgeführt, wenn die Skalarprodukte in üblicher Weise in Standardarithmetik — also nach 2.3.9 — berechnet werden.

Wegen der fehlenden Pivotisierung ist die direkte Faktorisierung nicht für beliebiges reguläres $\boldsymbol{A}$ durchführbar und numerisch gutartig, vgl. 5.2 und 5.3. Im folgenden soll gezeigt werden, daß die zur Stabilisierung führende Spaltenpivotisierung in einfacher Weise realisiert werden kann. Dazu beachten wir, daß

$$a_{kk}^{(k)} := r_{kk} = a_{kk} - \sum_{p=1}^{k-1} l_{kp} r_{pk} \tag{6}$$

gerade das Pivotelement des k-ten Hauptschrittes des Gaußschen Algorithmus darstellt. Wenn statt $a_{kk}^{(k)}$ das Element $a_{sk}^{(k)}$ mit $k < s \leqq n$ genommen würde, müßte (6) durch

$$\hat{a}_{sk} := a_{sk} - \sum_{p=1}^{k-1} l_{sp} r_{pk} \qquad (k+1 \leqq s \leqq n) \tag{7}$$

ersetzt werden. Ein Vergleich mit 6.2.1 zeigt, daß (7) den Zähler von l_{sk} darstellt. also ohnehin berechnet werden muß. Die Spaltenpivotisierung läuft dann darauf hinaus, unter allen Elementen (6), (7) ein betragsgrößtes zu bestimmen und die zugehörige Zeile mit der k-ten zu vertauschen. Eine zu 5.2.3 analoge Computerrealisierung auf dem Platz von $\boldsymbol{A}$ lautet dann wie folgt.

6.2.2. *Direkte $\boldsymbol{LR}$-Faktorisierung mit Spaltenpivotisierung*

Aufgabe: Für $\boldsymbol{A} = (a_{ij}) \in \mathbf{R}^{n,n}$ ist durch direkte Faktorisierung mit Spaltenpivotisierung eine Darstellung $\boldsymbol{PA} = \boldsymbol{LR}$ zu berechnen. Die Matrix $\boldsymbol{A}$ ist mit den

signifikanten Elementen von $\boldsymbol{L}, \boldsymbol{R} \in \mathbf{R}^{n,n}$ zu überspeichern, die Permutationsmatrix $\boldsymbol{P} = \boldsymbol{T}_{n-1,s(n-1)} \cdots \boldsymbol{T}_{1,s(1)}$ ist durch die Zahlen $\{s(1), \ldots, s(n-1)\}$ zu charakterisieren. Falls für ein $k \in \{1, \ldots, n\}$ kein $r_{kk} \neq 0$ gefunden werden kann, ist abzubrechen und $ie = -1$ zu setzen, andernfalls wird $ie = 0$ gesetzt, und $\boldsymbol{R}$ ist regulär.

Algorithmus:

$ie := 0$
for $k := 1(1)n$ **do**

Pivotsuche: $z := 0$
for $i := k(1)n$ **do** $a_{ik} := a_{ik} - \sum_{p=1}^{k-1} a_{ip} * a_{pk}$
if $|a_{ik}| > z$ **then** $[z := |a_{ik}|, s := i]$
if $z = 0$ **then** $[ie := -1,$ **stop**$]$
$s(k) := s$
Vertauschung:
if $k < s$ **then for** $j := 1(1)n$ **do** $[z := a_{kj}, a_{kj} := a_{sj}, a_{sj} := z]$
Berechnung der r_{kj}:
for $j := k + 1(1)n$ **do** $a_{kj} := a_{kj} - \sum_{p=1}^{k-1} a_{kp} * a_{pj}$
Berechnung der l_{ik}:
for $i := k + 1(1)n$ **do** $a_{ik} := a_{ik}/a_{kk}$

Aufwand: $\sim n^3/3$ opms, $\sim n$ S (für $s(1), \ldots, s(n-1)$)

6.2.3. *Bemerkung.* (i) Bei Auswertung der Skalarprodukte gemäß 2.3.9 liefert 6.2.2 dieselben Ausgangsdaten wie 5.2.3, so daß die Ergebnisse der Rundungsfehleranalyse aus 5.3 wörtlich für die nach 6.2.2 berechnete Faktorisierung gelten. Der Vorteil von 6.1.2 liegt in der etwas kompakteren Form und der Tatsache, daß die Rechenoperationen im wesentlichen in Form von Skalarprodukten anfallen. In 6.2.2 können daher zur Skalarproduktberechnung spezielle Routinen — gegebenenfalls als Assemblerprogramme und mit Akkumulation in höherer Genauigkeit — verwendet werden, was in der stufenweisen Grundform 5.2.3 nicht möglich ist.

(ii) Spaltenpivotisierung ist auch bei Verwendung der Reihenfolge (R_3) möglich. In analoger Weise kann für die Reihenfolgen (R_1) und (R_2) eine Zeilenpivotisierung durch Vertauschung der Spalten realisiert werden. Dagegen ist die vollständige Pivotisierung mit dem Prinzip der direkten Faktorisierung nicht verträglich, da dazu alle $a_{ij}^{(k)}$ $(i, j = k, \ldots, n)$ bekannt sein müssen. □

B. *Symmetrische definite Matrizen*

Für eine symmetrische und positiv definite Matrix $\boldsymbol{A}$ ist der Gaußsche Algorithmus ohne Pivotisierung durchführbar, und es gilt

$$\boldsymbol{L}^{\mathsf{T}} = \boldsymbol{D}^{-1}\boldsymbol{R} \quad \text{mit} \quad \boldsymbol{D} = \operatorname{diag}(d_k), \qquad d_k := r_{kk} > 0 \quad (k = 1, \ldots, n), \tag{8}$$

also

$$\boldsymbol{A} = \boldsymbol{L}\boldsymbol{D}\boldsymbol{L}^{\mathsf{T}}, \tag{9}$$

vgl. 6.1.A. Wegen (8) kann in diesem Fall die Laufanweisung zur Berechnung der l_{ik} in 6.2.1 durch

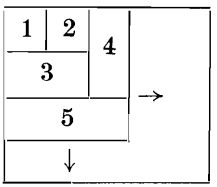
for $i := k + 1(1)n$ **do** $l_{ik} := r_{ki}/r_{kk}$

ersetzt werden, so daß sich der Rechenaufwand auf $\sim n^3/6$ opms reduziert. Allerdings werden wie im nichtsymmetrischen Fall n^2 S zum Speichern der l_{ij} und r_{ij} benötigt, bei in-situ-Realisierung also alle Plätze von $\boldsymbol{A}$. Wünschenswert ist dagegen ein Verfahren, bei dem sich wie in 6.1.1 die Symmetrie auch im Speicherplatzbedarf niederschlägt und das mit $\sim n^2/2$ S für $\boldsymbol{L}$ und $\boldsymbol{D}$ auskommt. Der sich anbietende Gedanke, auf das explizite Auftreten der r_{kj} überhaupt zu verzichten und sie überall durch $l_{jk}d_k$ zu ersetzen, führt bei Anwendung auf 6.2.1 zu der Vorschrift

for $k := 1(1)n$ **do**

$$\left|\begin{array}{l} d_k := a_{kk} - \sum_{p=1}^{k-1} l_{kp} * l_{kp} * d_p \\ \textbf{for } i := k + 1(1)n \textbf{ do } l_{ik} := \left(a_{ik} - \sum_{p=1}^{k-1} l_{ip} * l_{kp} * d_p\right) \Big| d_k \end{array}\right.$$

Wegen der hier auftretenden Dreifachprodukte werden zur Realisierung $\sim(n^3/3$ opm $+ n^3/6$ ops), also fast soviel Rechenoperationen wie im nichtsymmetrischen Fall benötigt, so daß dieses Vorgehen ausscheidet und doch mit den r_{ij} gearbeitet werden muß. Zur Senkung des Speicherbedarfs bietet sich dabei eine solche Reihenfolge an, bei der im k-ten Schritt nur auf die Elemente der k-ten Spalte von $\boldsymbol{R}$ zurückgegriffen wird. Dies ist für die durch das Schema



(R_4)

charakterisierte Reihenfolge (R_4) der Fall, bei der die l_{ij} zeilenweise berechnet werden. Unter Verwendung der Bezeichnung

$$r_j := r_{jk} = l_{kj}d_j \qquad (j = 1, \ldots, k-1)$$

ergibt sich mit dieser Reihenfolge der folgende Algorithmus:

6.2.4. *Direkte* $\boldsymbol{LDL^T}$*-Faktorisierung.* Es sei $\boldsymbol{A} \in \mathbf{S}^{n,n}$ positiv definit. Dann ist das Verfahren

for $k := 1(1)n$ **do**

$$\left|\begin{array}{l} \textbf{for } j := 1(1)k-1 \textbf{ do } \left[\begin{array}{l} r_j := a_{kj} - \sum_{p=1}^{j-1} l_{jp} * r_p \\ l_{kj} := r_j/d_j \end{array}\right] \\ d_k := a_{kk} - \sum_{p=1}^{k-1} l_{kp} * r_p \end{array}\right.$$

in exakter Arithmetik mit $d_k > 0$ $(k = 1, \ldots, n)$ durchführbar und liefert die Faktorisierung $\boldsymbol{A} = \boldsymbol{LDL}^T$.

Aufwand: $\sim n^3/6$ opms, $\sim n$ S (für $r_1, \ldots, r_{n-1}$)

6.2.5. *Bemerkung.* (i) Algorithmus 6.2.4 kann analog zu 6.1.1 auf dem Platz des unteren Dreiecks von $\boldsymbol{A}$ ausgeführt werden. Wenn auf die Möglichkeit der Berechnung von d_k über ein Skalarprodukt verzichtet wird, können die $r_j = r_{jk}$ sogar auf dem Platz von l_{kj} — bei in-situ-Realisierung also auf dem von a_{kj} — gespeichert werden. Dazu ist 6.2.4 wie folgt zu modifizieren, vgl. auch 6.1.2(ii):

for $k := 1(1)n$ **do**

$$\left|\begin{array}{l} \textbf{for } j := 1(1)k-1 \textbf{ do } l_{kj} := a_{kj} - \sum_{p=1}^{j-1} l_{jp} * l_{kp} \\ d_k := a_{kk} \\ \textbf{for } j := 1(1)k-1 \textbf{ do } [\varrho := l_{kj},\ l_{kj} := \varrho/d_j,\ d_k := d_k - l_{kj} * \varrho]. \end{array}\right.$$

Ein zu 6.2.4 analoger Algorithmus kann für die Reihenfolge (R_3), d. h. für spaltenweise Berechnung von $\boldsymbol{L}$ und $\boldsymbol{D}$ angegeben werden, während die Diagonalpivotisierung entsprechend 6.1.5(iii) nicht mit dem Prinzip der direkten Berechnung von $\boldsymbol{L}$ und $\boldsymbol{D}$ verträglich ist.

(ii) Bei Auswertung der Skalarprodukte nach 2.3.9 liefert 6.2.4 dieselben Ausgangsdaten wie 6.1.1, so daß die Ergebnisse der Rundungsfehleranalyse 6.1.3 wörtlich übernommen werden können. □

Für die Cholesky-Faktorisierung

$$\boldsymbol{A} = \hat{\boldsymbol{L}}\hat{\boldsymbol{L}}^T \tag{10}$$

mit

$$\hat{\boldsymbol{L}} := \boldsymbol{LD}^{1/2}, \quad \text{also} \quad \hat{l}_{ij} = l_{ij}\sqrt{d_j} \qquad (i \geqq j), \tag{11}$$

geht 6.2.4 in das folgende, nach CHOLESKY benannte *Verfahren* zur Berechnung von $\hat{\boldsymbol{L}}$ über. Die bei der $\boldsymbol{LDL}^T$-Faktorisierung beobachteten Speicherkonflikte treten dabei wie in der $(n-1)$-stufigen Grundform 6.1.4 nicht mehr auf.

6.2.6. *Direkte $\hat{\boldsymbol{L}}\hat{\boldsymbol{L}}^T$-Faktorisierung.* Es sei $\boldsymbol{A} \in \mathbf{S}^{n,n}$ positiv definit. Dann ist das Verfahren

for $k := 1(1)n$ **do**

$$\left|\begin{array}{l} \textbf{for } j := 1(1)k-1 \textbf{ do } \hat{l}_{kj} := \left(a_{kj} - \sum_{p=1}^{j-1} \hat{l}_{jp} * \hat{l}_{kp}\right) \Big/ \hat{l}_{jj} \\ \hat{l}_{kk} := \operatorname{sqrt}\left(a_{kk} - \sum_{p=1}^{k-1} \hat{l}_{kp}^2\right) \end{array}\right.$$

in exakter Arithmetik mit $\hat{l}_{kk} > 0$ $(k = 1, \ldots, n)$ durchführbar und liefert die signifikanten Elemente von $\hat{\boldsymbol{L}}$ in der Faktorisierung $\boldsymbol{A} = \hat{\boldsymbol{L}}\hat{\boldsymbol{L}}^T$.

Aufwand: $\sim n^3/6$ opms $+$ n opr

Die Formel für $\hat{l}_{kk}$ aus 6.2.6 läßt sich auch in der nach a_{kk} aufgelösten Form

$$\hat{l}_{k1}^2 + \hat{l}_{k2}^2 + \cdots + \hat{l}_{kk}^2 = a_{kk}$$

schreiben. Hieraus folgt

$$|\hat{l}_{kj}| \leqq \sqrt{a_{kk}} \quad (k \geqq j) \quad \text{sowie} \quad \|\hat{\boldsymbol{L}}\|_F = \left(\sum_{k=1}^{n} a_{kk}\right)^{1/2}, \tag{12}$$

d. h., der Dreiecksfaktor $\hat{\boldsymbol{L}}$ ist auch ohne Pivotisierung a priori beschränkt, vgl. 5.2.9 und Ü 6.1.1.

6.2.7. *Bemerkung.* (i) Das Cholesky-Verfahren 6.2.6 kann analog zu 6.1.4 auf dem Platz des unteren Dreiecks von $\boldsymbol{A}$ ausgeführt werden. Spaltenweise Berechnung der $\hat{l}_{ik}$ ist nach der Vorschrift

for $k := 1(1)n$ **do**

$$\left|\begin{array}{l} a_{kk} := \operatorname{sqrt}\left(a_{kk} - \sum_{p=1}^{k-1} (a_{kp})^2\right) \\ \textbf{for } i := k+1(1)n \textbf{ do } a_{ik} := \left(a_{ik} - \sum_{p=1}^{k-1} a_{ip} * a_{kp}\right) \Big/ a_{kk} \end{array}\right.$$

möglich, während die Diagonalpivotisierung entsprechend 6.1.5(iii) nicht mit dem Prinzip der direkten Berechnung von $\hat{\boldsymbol{L}}$ verträglich ist.

(ii) Bei üblicher Auswertung der Skalarprodukte gemäß 2.3.9 sind die Ausgangsdaten von 6.2.6 mit denen von 6.1.4 identisch, so daß die Rundungsfehleranalyse aus 6.1.5(ii) auch hier gültig ist.

(iii) Direkte Methoden zur Berechnung der BKP-Faktorisierung indefiniter symmetrischer Matrizen sind nicht bekannt und wegen der komplizierten Pivotstrategie wohl auch nicht möglich. □

Übungsaufgaben

Ü 6.2.1. Man gebe die 6.2.1. entsprechenden Verfahren zur direkten $\boldsymbol{LR}$-Faktorisierung für die Reihenfolgen (R_2) und (R_3) an und mache sich die Unterschiede im Zugriff auf die Elemente von $\boldsymbol{A}$, $\boldsymbol{L}$ und $\boldsymbol{R}$ klar. Wie sind die Algorithmen zu modifizieren, wenn für (R_2) mit Zeilenpivotisierung und für (R_3) mit Spaltenpivotisierung gearbeitet werden soll?

Ü 6.2.2. Man modifiziere 6.2.4 so, daß die Elemente von $\boldsymbol{L}$ und $\boldsymbol{D}$ spaltenweise nach der Reihenfolge (R_3) berechnet werden.

6.3. Aufdatierung von Dreiecksfaktorisierungen

Bei gewissen praktischen Aufgaben — etwa in der linearen und nichtlinearen Optimierung, bei der Minimierung nichtlinearer Funktionen oder bei der Lösung nichtlinearer Gleichungssysteme — ist nicht nur ein einzelnes Gleichungssystem, sondern

eine Folge

$$\boldsymbol{A}_k \boldsymbol{x} = \boldsymbol{b}^k \qquad (k = 1, \ldots, K) \tag{1}$$

von K Systemen zu lösen, wobei $\boldsymbol{A}_{k+1}$ aus $\boldsymbol{A}_k$ durch eine Rang-1-Modifikation

$$\boldsymbol{A}_{k+1} = \boldsymbol{A}_k + \boldsymbol{u}^k \boldsymbol{v}^{k\mathsf{T}} \tag{2}$$

entsteht. Es wäre dann wünschenswert, die Dreiecksfaktorisierung von $\boldsymbol{A}_k \in \mathbf{R}^{n,n}$ nicht für jedes k mit dem Aufwand von

$$\sim K_1 n^3 \text{ opms} \tag{3}$$

($K_1 = 1/3$ für nichtsymmetrisches $\boldsymbol{A}_k$, $K_1 = 1/6$ für symmetrisches $\boldsymbol{A}_k$) neu berechnen zu müssen, sondern die Faktoren von $\boldsymbol{A}_{k+1}$ aus denen von $\boldsymbol{A}_k$ mit geringerem Aufwand zu bestimmen. Im folgenden werden wir zeigen, daß dies in der Tat mit

$$\sim K_2 n^2 \text{ opms} \tag{4}$$

möglich ist.

Zur Vereinfachung der Schreibweise bezeichnen wir $\boldsymbol{A}_k$ mit $\boldsymbol{A}$ und die durch die Modifikation (2) entstehende Matrix mit

$$\bar{\boldsymbol{A}} = \boldsymbol{A} + \boldsymbol{u}\boldsymbol{v}^{\mathsf{T}}; \tag{5}$$

die Faktoren werden analog gekennzeichnet. Die Matrizen $\boldsymbol{A}$ und $\bar{\boldsymbol{A}}$ seien beide regulär.

A. *Nichtsymmetrische Matrizen*

Wir betrachten zunächst die folgende

Hilfsaufgabe: Zu gegebenem

$$\boldsymbol{r} = \begin{pmatrix} r_1 \\ r_2 \end{pmatrix}, \qquad \boldsymbol{L} = \begin{pmatrix} 1 & 0 \\ l_{21} & 1 \end{pmatrix} \quad \text{mit} \quad |l_{21}| \leqq \lambda$$

sind $\boldsymbol{F} = \begin{pmatrix} \alpha & \beta \\ \gamma & \delta \end{pmatrix}$ regulär und $\boldsymbol{T} = \boldsymbol{T}_{11} = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}$ oder $\boldsymbol{T} = \boldsymbol{T}_{12} = \begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix}$ so festzulegen, daß

$$\boldsymbol{L}\boldsymbol{r} = \boldsymbol{T}(\boldsymbol{T}\boldsymbol{L}\boldsymbol{F}^{-1})(\boldsymbol{F}\boldsymbol{r}) =: \boldsymbol{T}\bar{\boldsymbol{L}}\bar{\boldsymbol{r}} \tag{6}$$

mit

$$\bar{\boldsymbol{L}} := \boldsymbol{T}\boldsymbol{L}\boldsymbol{F}^{-1} = \begin{pmatrix} 1 & 0 \\ \bar{l}_{21} & 1 \end{pmatrix} \quad \text{und} \quad \bar{\boldsymbol{r}} := \boldsymbol{F}\boldsymbol{r} = \begin{pmatrix} \bar{r}_1 \\ 0 \end{pmatrix} \tag{7}$$

gilt. □

Für die weiteren Überlegungen führen wir die Größen

$$\boldsymbol{L}\boldsymbol{r} = \begin{pmatrix} 1 & 0 \\ l_{21} & 1 \end{pmatrix} \begin{pmatrix} r_1 \\ r_2 \end{pmatrix} = \begin{pmatrix} r_1 \\ l_{21} r_1 + r_2 \end{pmatrix} =: \begin{pmatrix} q_1 \\ q_2 \end{pmatrix} =: q$$

ein und unterscheiden zwei Fälle.

Fall 1: $|q_1| \geqq |q_2|$.
Hier wird $\boldsymbol{T} := \boldsymbol{T}_{11}$ gesetzt. Aus (7) folgt dann $\bar{\boldsymbol{L}} = \boldsymbol{L}\boldsymbol{F}^{-1}$, also $\boldsymbol{F} = \bar{\boldsymbol{L}}^{-1}\boldsymbol{L}$. Als Produkt unterer Einsdreiecksmatrizen muß $\boldsymbol{F}$ daher auch von Einsdreiecksform sein:

$$\boldsymbol{F} = \begin{pmatrix} 1 & 0 \\ \gamma & 1 \end{pmatrix}, \qquad \boldsymbol{F}^{-1} = \begin{pmatrix} 1 & 0 \\ -\gamma & 1 \end{pmatrix}. \tag{8}$$

Der noch freie Parameter γ ergibt sich aus der zweiten Bedingung (7), d. h. aus

$$\bar{\boldsymbol{r}} = \begin{pmatrix} \bar{r}_1 \\ 0 \end{pmatrix} = \boldsymbol{F}\boldsymbol{r} = \begin{pmatrix} r_1 \\ \gamma r_1 + r_2 \end{pmatrix}$$

im Fall $r_1 \neq 0$ zu

$$\gamma := -r_2/r_1$$

und kann wegen $|r_1| = |q_1| \geqq |q_2| = |l_{21}r_1 + r_2| \geqq |r_2| - \lambda\,|r_1|$ gemäß $|\gamma| \leqq 1 + \lambda$ abgeschätzt werden. Daraus folgt

$$\bar{l}_{21} = l_{21} - \gamma = q_2/q_1$$

mit $|\bar{l}_{21}| \leqq 1$. Im Fall $r_1 = r_2 = 0$ setzen wir $\gamma = 0$, also $\boldsymbol{F} = \boldsymbol{I}$, was auf $\bar{l}_{21} = l_{21}$ mit $|\bar{l}_{21}| \leqq \lambda$ führt.

Fall 2: $|q_1| < |q_2|$.
Wir setzen jetzt $\boldsymbol{T} := \boldsymbol{T}_{12}$. Aus (7) folgt dann

$$\boldsymbol{F} = \bar{\boldsymbol{L}}^{-1}\boldsymbol{T}_{12}\boldsymbol{L} = \begin{pmatrix} 1 & 0 \\ -\bar{l}_{21} & 1 \end{pmatrix}\begin{pmatrix} l_{21} & 1 \\ 1 & 0 \end{pmatrix} = \left(\begin{array}{c|c} l_{21} & 1 \\ \hline -\bar{l}_{21}l_{21} + 1 & -\bar{l}_{21} \end{array}\right).$$

Der noch freie Wert $\bar{l}_{21}$ ergibt sich wieder aus der zweiten Forderung (7), d. h. aus

$$\bar{\boldsymbol{r}} = \begin{pmatrix} \bar{r}_1 \\ 0 \end{pmatrix} = \boldsymbol{F}\boldsymbol{r} = \begin{pmatrix} l_{21}r_1 + r_2 \\ -\bar{l}_{21}(l_{21}r_1 + r_2) + r_1 \end{pmatrix} = \begin{pmatrix} q_2 \\ -\bar{l}_{21}q_2 + q_1 \end{pmatrix}$$

zu

$$l_{21} := q_1/q_2$$

mit $|\bar{l}_{21}| \leqq 1$. Damit erhalten wir

$$\boldsymbol{F} = \begin{pmatrix} \alpha & 1 \\ \gamma & \delta \end{pmatrix}, \qquad \boldsymbol{F}^{-1} = \begin{pmatrix} -\delta & 1 \\ \gamma & -\alpha \end{pmatrix} \tag{9}$$

mit

$$\alpha := l_{21}, \qquad \delta := -\bar{l}_{21} = -q_1/q_2, \qquad \gamma := -\bar{l}_{21}l_{21} + 1 = \alpha\delta + 1 = r_2/q_2,$$

wobei $|\alpha| \leqq \lambda$, $|\delta| \leqq 1$ und $|\gamma| \leqq 1 + \lambda$ ist.

Wenn $\boldsymbol{F}$ von links auf einen Spaltenvektor bzw. $\boldsymbol{F}^{-1}$ von rechts auf einen Zeilenvektor gemäß

$$\bar{\boldsymbol{w}} := \boldsymbol{F}\boldsymbol{w} \quad \text{bzw.} \quad \bar{\boldsymbol{f}}^{\mathsf{T}} := \boldsymbol{f}^{\mathsf{T}}\boldsymbol{F}^{-1} \quad \text{mit} \quad \boldsymbol{w}, \boldsymbol{f} \in \mathbf{R}^2$$

angewendet wird, ergibt sich im Fall 1 mit (8)

$$\begin{aligned} \bar{w}_1 &:= w_1, \\ \bar{w}_2 &:= w_2 + \gamma w_1 \end{aligned} \quad \text{bzw.} \quad \begin{aligned} \bar{f}_1 &:= f_1 - \gamma f_2, \\ \bar{f}_2 &:= f_2, \end{aligned}$$

d. h., die Berechnung erfordert jeweils 1 opms. Im Fall 2 würde sich mit (9) dagegen

$$\begin{aligned} \bar{w}_1 &:= w_2 + \alpha w_1, \\ \bar{w}_2 &:= \delta w_2 + \gamma w_1 \end{aligned} \quad \text{bzw.} \quad \begin{aligned} \bar{f}_1 &:= -\delta f_1 + \gamma f_2, \\ \bar{f}_2 &:= f_1 - \alpha f_2 \end{aligned}$$

ergeben, d. h., der Aufwand wäre jeweils 3 opm + 2 ops. Unter Beachtung von $\gamma = \alpha\delta + 1$ läßt sich jedoch eine Multiplikation einsparen, indem nach der in exakter Arithmetik äquivalenten Vorschrift

$$\begin{aligned} \bar{w}_1 &:= w_2 + \alpha w_1, \\ \bar{w}_2 &:= w_1 + \delta \bar{w}_1 \end{aligned} \quad \text{bzw.} \quad \begin{aligned} \bar{f}_2 &:= f_1 - \alpha f_2, \\ \bar{f}_1 &:= f_2 - \delta \bar{f}_2 \end{aligned} \tag{10}$$

vorgegangen wird. Dies entspricht der Darstellung von $\boldsymbol{F}$ bzw. $\boldsymbol{F}^{-1}$ gemäß

$$\boldsymbol{F} = \begin{pmatrix} 1 & 0 \\ \delta & 1 \end{pmatrix} \begin{pmatrix} \alpha & 1 \\ 1 & 0 \end{pmatrix} \quad \text{bzw.} \quad \boldsymbol{F}^{-1} = \begin{pmatrix} 0 & 1 \\ 1 & -\alpha \end{pmatrix} \begin{pmatrix} 1 & 0 \\ -\delta & 1 \end{pmatrix}, \tag{11}$$

bei der nur α und δ explizit vorkommen. Der Aufwand ist jetzt jeweils 2 opms.

In den nachfolgenden Anwendungen zur Aufdatierung von Dreiecksfaktorisierungen stellen die Matrizen $\boldsymbol{L}$, $\bar{\boldsymbol{L}}$ der Hilfsaufgabe gewisse (2,2)-Diagonalblöcke von (n,n)-Einsdreiecksmatrizen $\boldsymbol{L}$, $\bar{\boldsymbol{L}}$ in den Positionen $\{k, k+1\}$ dar, und die Vektoren $\boldsymbol{r}$, $\bar{\boldsymbol{r}}$ der Hilfsaufgabe sind die zugehörigen Zweierblöcke $(r_k, r_{k+1})^T$ bzw. $(\bar{r}_k, \bar{r}_{k+1})^T$ von n-Vektoren $\boldsymbol{r}$, $\bar{\boldsymbol{r}}$. Die Matrix $\boldsymbol{F}$ tritt demgemäß in der Gestalt

$$\boldsymbol{F}_{k,k+1} = \left(\begin{array}{c|c|c} \boldsymbol{I}_{k-1} & \boldsymbol{O} & \boldsymbol{O} \\ \hline \boldsymbol{O} & \boldsymbol{F} & \boldsymbol{O} \\ \hline \boldsymbol{O} & \boldsymbol{O} & \boldsymbol{I}_{n-k-1} \end{array}\right) = \left(\begin{array}{c|cc|c} 1 & & & \\ & \ddots & & \\ & & 1 & \\ \hline & & & \\ & \alpha & \beta & \\ & \gamma & \delta & \\ \hline & & & 1 \\ & & & & \ddots \\ & & & & & 1 \end{array}\right) \tag{12}$$

auf, und $\boldsymbol{T}_{12}$ ist durch $\boldsymbol{T}_{k,k+1}$ zu ersetzen. In dieser Erweiterung erhalten wir als Zusammenfassung der obigen Überlegungen das folgende Resultat.

6.3.1. *Dreiecksformerhaltender Elementareliminationsschritt.* Gegeben seien eine untere Eindreiecksmatrix $\boldsymbol{L} = (l_{ij}) \in \mathbf{R}^{n,n}$, ein Vektor $\boldsymbol{r} = (r_i) \in \mathbf{R}^n$ und ein Index $k \in \{1, \ldots, n-1\}$. Die Matrix $\boldsymbol{F}_{k,k+1}$ des Typs (12) und die Vertauschungsmatrix $\boldsymbol{T}_{ks}$ mit $s = s(k) \in \{k, k+1\}$ seien durch die Zahlen $\{k, l_{k+1,k}, r_k, r_{k+1}\}$

wie folgt festgelegt:

S1: $q_k := r_k$, $q_{k+1} := l_{k+1,k} * r_k + r_{k+1}$

S2: **if** $|q_k| \geqq |q_{k+1}|$

then Fall 1:

$s(k) := k$

$\alpha := \delta := 1$, $\beta := 0$, **if** $q_k = 0$ **then** $\gamma := 0$ **else** $\gamma := -r_{k+1}/r_k$

else Fall 2:

$s(k) := k + 1$

$\alpha := l_{k+1,k}$, $\beta := 1$, $\gamma := r_{k+1}/q_{k+1}$, $\delta := -q_k/q_{k+1}$

Dann gilt

$$\boldsymbol{L}\boldsymbol{r} = \boldsymbol{T}_{ks}\bar{\boldsymbol{L}}\bar{\boldsymbol{r}} \quad \text{mit} \quad \bar{\boldsymbol{L}} := \boldsymbol{T}_{ks}\boldsymbol{L}\boldsymbol{F}_{k,k+1}^{-1}, \qquad \bar{\boldsymbol{r}} := \boldsymbol{F}_{k,k+1}\boldsymbol{r}.$$

Die Matrix $\bar{\boldsymbol{L}}$ ist eine untere Einsdreiecksmatrix und unterscheidet sich von $\boldsymbol{L}$ bis auf die im Fall 2 vorzunehmenden Vertauschungen der Zeilen $\{k, k+1\}$ nur in den Spalten $\{k, k+1\}$, und der Vektor $\bar{\boldsymbol{r}}$ unterscheidet sich von $\boldsymbol{r}$ nur in den Komponenten $\{k, k+1\}$. Dabei ist

$$\bar{l}_{k+1,k} = \begin{cases} l_{k+1,k} & \text{im Fall 1, } q_k = 0, \\ q_{k+1}/q_k & \text{im Fall 1, } q_k \neq 0, \\ q_k/q_{k+1} & \text{im Fall 2;} \end{cases} \qquad \bar{r}_k = \begin{cases} q_k & \text{im Fall 1,} \\ q_{k+1} & \text{im Fall 2;} \end{cases} \quad \bar{r}_{k+1} = 0.$$

Der Vollständigkeit halber sind in 6.3.1 auch die explizit nicht benötigten Parameter α, β, δ im Fall 1 bzw. β, γ im Fall 2 mit angegeben worden.

Unter der Voraussetzung

$$|l_{k+1,k}| \leqq \lambda \tag{13}$$

folgt aus 6.3.1

$$|\bar{l}_{k+1,k}| \leqq \max\{\lambda, 1\}, \tag{14}$$

und die signifikanten Blöcke $\boldsymbol{F}$ bzw. $\boldsymbol{F}^{-1}$ von $\boldsymbol{F}_{k,k+1}$ bzw. $\boldsymbol{F}_{k,k+1}^{-1}$ genügen den Abschätzungen

$$|\boldsymbol{F}| \leqq \left(\begin{array}{c|c} \max\{\lambda, 1\} & 1 \\ \hline 1 + \lambda & 1 \end{array}\right), \qquad |\boldsymbol{F}^{-1}| \leqq \left(\begin{array}{c|c} 1 & 1 \\ \hline 1 + \lambda & \max\{\lambda, 1\} \end{array}\right), \tag{15}$$

insbesondere ist

$$\|\boldsymbol{F}\|_\infty \leqq 2 + \lambda, \qquad \|\boldsymbol{F}^{-1}\|_\infty \leqq 1 + \lambda + \max\{\lambda, 1\}. \tag{16}$$

Wenn die Einsdreiecksmatrix $\boldsymbol{L}$ den nach dem Gaußschen Algorithmus mit Spaltenpivotisierung berechneten Dreiecksfaktor darstellt, sind die Elemente betragsmäßig durch 1 beschränkt, speziell gilt

$$|l_{k+1,k}| \leqq 1 \qquad (k = 1, \ldots, n-1) \tag{17}$$

für die in 6.3.1 verwendeten Subdiagonalelemente. Wegen (14) gilt dann (17) auch für die modifizierte Matrix $\bar{\boldsymbol{L}}$, d. h., die Gültigkeit von (17) ist invariant unter den Elementareliminationsschritten 6.3.1. Die übrigen Elemente von $\{\boldsymbol{L}, \boldsymbol{r}\}$ können beim Übergang zu $\{\bar{\boldsymbol{L}}, \bar{\boldsymbol{r}}\}$ durchaus anwachsen, wegen (15), (16) jedoch nicht beliebig größer werden: Unter der Voraussetzung (17) kann das Betragsmaximum nach der Elimination höchstens um den Faktor 3 größer als vorher sein.

Algorithmus 6.3.1 wird jetzt $(n-1)$-mal angewendet, um sukzessive die Komponenten r_i $(i = n, \ldots, 2)$ von r zu 0 zu machen und dabei die Dreiecksform eines linken L-Faktors zu erhalten.

6.3.2. *Dreiecksformerhaltende Transformation von* $\boldsymbol{r}$ *in* $\varrho_1\boldsymbol{e}^1$. Gegeben seien die untere Einsdreiecksmatrix $\boldsymbol{L} \in \mathbf{R}^{n,n}$ und der Vektor $\boldsymbol{r} \in \mathbf{R}^n$. Die Matrizen $\boldsymbol{L}_1$ und der Vektor $\boldsymbol{r}^1$ seien rekursiv definiert durch die Vorschrift

S1: Setze $\boldsymbol{L}^{(n-1)} := \boldsymbol{L}$, $\boldsymbol{r}^{(n-1)} := \boldsymbol{r}$

S2: **for** $k := n - 1(-1)1$ **do**

> Lege $\{\boldsymbol{F}_{k,k+1}, \boldsymbol{T}_{k,s(k)}\}$ gemäß 6.3.1 durch
> $\{k, (\boldsymbol{L}^{(k)})_{k+1,k}, r_k^{(k)}, r_{k+1}^{(k)}\}$ fest.
> Bestimme $\boldsymbol{L}^{(k-1)} := \boldsymbol{T}_{k,s(k)}\boldsymbol{L}^{(k)}\boldsymbol{F}_{k,k+1}^{-1}$, $\boldsymbol{r}^{(k-1)} := \boldsymbol{F}_{k,k+1}\boldsymbol{r}^{(k)}$

S3: Setze $\boldsymbol{L}_1 := \boldsymbol{L}^{(0)}$, $\boldsymbol{r}^1 := \boldsymbol{r}^{(0)}$.

Dann gilt

$$\boldsymbol{L}\boldsymbol{r} = \boldsymbol{P}_1^{\mathsf{T}}(\boldsymbol{P}_1\boldsymbol{L}\boldsymbol{F}_1^{-1})\,(\boldsymbol{F}_1\boldsymbol{r}) = \boldsymbol{P}_1^{\mathsf{T}}\boldsymbol{L}_1\boldsymbol{r}^1$$

mit der unteren Einsdreiecksmatrix $\boldsymbol{L}_1 := \boldsymbol{P}_1\boldsymbol{L}\boldsymbol{F}_1^{-1}$ und dem Vektor $\boldsymbol{r}^1 := \boldsymbol{F}_1\boldsymbol{r} = \varrho_1\boldsymbol{e}^1$, wobei

$$\boldsymbol{F}_1 := \boldsymbol{F}_{12}\boldsymbol{F}_{23}\cdots\boldsymbol{F}_{n-1,n}, \qquad \boldsymbol{P}_1 := \boldsymbol{T}_{1,s(1)}\boldsymbol{T}_{2,s(2)}\cdots\boldsymbol{T}_{n-1,s(n-1)}$$

ist.

Im Fall $n = 4$ erfolgt die Transformation 6.3.2 nach folgendem Muster:

$$\left(\begin{array}{cccc|c} 1 & & & & \times \\ \times & 1 & & & \times \\ \times & \times & 1 & & \times \\ \times & \times & \times & 1 & \times \end{array}\right) \rightarrow \left(\begin{array}{cccc|c} 1 & & & & \times \\ \times & 1 & & & \times \\ \otimes & \otimes & 1 & & \boxed{\times} \\ \otimes & \otimes & \boxed{\times} & 1 & 0 \end{array}\right)$$

$$[\boldsymbol{L} \mid \boldsymbol{r}]$$

$$\rightarrow \left(\begin{array}{cccc|c} 1 & & & & \times \\ \otimes & 1 & & & \boxed{\times} \\ \otimes & \boxed{\times} & 1 & & 0 \\ \times & \boxed{\times} & \boxed{\times} & 1 & 0 \end{array}\right) \rightarrow \left(\begin{array}{cccc|c} 1 & & & & \boxed{\times} \\ \boxed{\times} & 1 & & & 0 \\ \boxed{\times} & \boxed{\times} & 1 & & 0 \\ \boxed{\times} & \boxed{\times} & \times & 1 & 0 \end{array}\right)$$

$$[\boldsymbol{L}_1 \mid \boldsymbol{r}^1]$$

Die eingerahmten Felder geben dabei diejenigen Elemente an, die im jeweiligen

Schritt neu berechnet werden; gegebenenfalls zu vertauschende Elemente sind durch Kreise gekennzeichnet.

Es sei jetzt $\boldsymbol{PA} = \boldsymbol{LR}$ eine Dreiecksfaktorisierung von $\boldsymbol{A}$. Dann kann (5) in der Form

$$\bar{\boldsymbol{A}} = \boldsymbol{P}^{\mathsf{T}}\boldsymbol{LR} + \boldsymbol{uv}^{\mathsf{T}} = \boldsymbol{P}^{\mathsf{T}}\boldsymbol{L}(\boldsymbol{R} + \boldsymbol{rv}^{\mathsf{T}}) \quad \text{mit} \quad \boldsymbol{r} := \boldsymbol{L}^{-1}\boldsymbol{Pu} \tag{18}$$

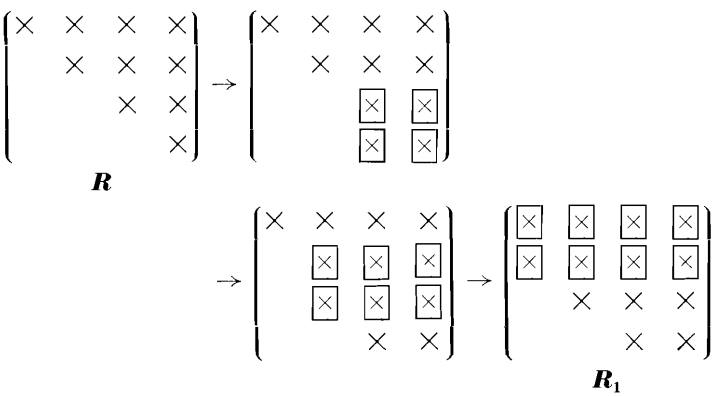
geschrieben werden. Anwendung der Transformation 6.3.2 führt auf

$$\bar{\boldsymbol{A}} = \boldsymbol{P}^{\mathsf{T}}\boldsymbol{P}_1^{\mathsf{T}}(\boldsymbol{P}_1\boldsymbol{L}\boldsymbol{F}_1^{-1})\,(\boldsymbol{F}_1\boldsymbol{R} + \boldsymbol{F}_1\boldsymbol{r}\boldsymbol{v}^{\mathsf{T}}) =: \boldsymbol{P}_2^{\mathsf{T}}\boldsymbol{L}_1(\boldsymbol{R}_1 + \boldsymbol{r}^1\boldsymbol{v}^{\mathsf{T}}) \tag{19}$$

mit $\boldsymbol{P}_2 := \boldsymbol{P}_1\boldsymbol{P}$ und $\boldsymbol{R}_1 := \boldsymbol{F}_1\boldsymbol{R}$. Dabei wird $\boldsymbol{R}_1$ analog zu $\boldsymbol{r}^1$ rekursiv gemäß

$$\boldsymbol{R}^{(n-1)} := \boldsymbol{R}, \text{ for } k := n - 1(1)1 \text{ do } \boldsymbol{R}^{(k-1)} := \boldsymbol{F}_{k,k+1}\boldsymbol{R}^{(k)}, \quad \boldsymbol{R}_1 := \boldsymbol{R}^{(0)} \tag{20}$$

berechnet. Für $n = 4$ verläuft die Bildung von $\boldsymbol{R}_1$ nach dem Muster



Die Subdiagonalelemente von $\boldsymbol{R}$ werden also sukzessive aufgefüllt. Wir erinnern daran, daß eine Matrix der Form $\boldsymbol{R}_1$, d. h. eine Matrix $\boldsymbol{R}_1$ mit

$$(\boldsymbol{R}_1)_{ij} = 0 \qquad (i > j + 1) \tag{21}$$

obere Hessenbergmatrix genannt wird; in analoger Weise ist eine *untere Hessenbergmatrix* definiert, vgl. 1.1.C.

Mit $\boldsymbol{R}_1$ ist in (19) dann auch $\boldsymbol{R}_2 := \boldsymbol{R}_1 + \boldsymbol{r}^1\boldsymbol{v}^{\mathsf{T}} = \boldsymbol{R}_1 + \varrho_1\boldsymbol{e}^1$ von oberer Hessenbergform, denn lediglich zur ersten Zeile von $\boldsymbol{R}_1$ wird $\varrho_1\boldsymbol{v}^{\mathsf{T}}$ addiert. Mit $\boldsymbol{L}_2 := \boldsymbol{L}_1$ läßt (19) sich somit als

$$\bar{\boldsymbol{A}} = \boldsymbol{P}_2^{\mathsf{T}}\boldsymbol{L}_2\boldsymbol{R}_2 \tag{22}$$

schreiben. Die noch störenden Subdiagonalelemente $(\boldsymbol{R}_2)_{k+1,k}$ $(k = 1, \ldots, n - 1)$ von $\boldsymbol{R}_2$ können dann wieder durch $n - 1$ Elementareliminationsschritte sukzessive zu 0 gemacht werden, wobei die k-te Spalte von $\boldsymbol{R}_2$ im k-ten Schritt die Rolle von $\boldsymbol{r}$ übernimmt.

6.3.3. *Dreiecksformerhaltende Transformation von $\boldsymbol{R}_2$ in $\overline{\boldsymbol{R}}$.* Gegeben seien die untere Einsdreiecksmatrix $\boldsymbol{L}_2 \in \mathbf{R}^{n,n}$ und die obere Hessenbergmatrix $\boldsymbol{R}_2 \in \mathbf{R}^{n,n}$. Die Matrizen $\overline{\boldsymbol{L}}, \overline{\boldsymbol{R}}$ seien rekursiv definiert durch die Vorschrift

S1: Setze $\boldsymbol{L}^{(1)} := \boldsymbol{L}_2$, $\boldsymbol{R}^{(1)} := \boldsymbol{R}_2$

S2: **for** $k := 1(1)n - 1$ **do**

| Lege $\{\tilde{\boldsymbol{F}}_{k,k+1}, \tilde{\boldsymbol{T}}_{k,s(k)}\}$ gemäß 6.3.1 durch
| $\{k, (\boldsymbol{L}^{(k)})_{k+1,k}, (\boldsymbol{R}^{(k)})_{kk}, (\boldsymbol{R}^{(k)})_{k+1,k}\}$ fest.
| Bestimme $\boldsymbol{L}^{(k+1)} := \tilde{\boldsymbol{T}}_{k,s(k)} \boldsymbol{L}^{(k)} \tilde{\boldsymbol{F}}^{-1}_{k,k+1}$, $\boldsymbol{R}^{(k+1)} := \tilde{\boldsymbol{F}}_{k,k+1} \boldsymbol{R}^{(k)}$

S3: Setze $\overline{\boldsymbol{L}} := \boldsymbol{L}^{(n)}$, $\overline{\boldsymbol{R}} := \boldsymbol{R}^{(n)}$

Dann gilt

$$\boldsymbol{L}_2\boldsymbol{R}_2 = \boldsymbol{P}_3^{\mathsf{T}}(\boldsymbol{P}_3\boldsymbol{L}_2\boldsymbol{F}_3^{-1})\,(\boldsymbol{F}_3\boldsymbol{R}_2) = \boldsymbol{P}_3^{\mathsf{T}}\overline{\boldsymbol{L}}\overline{\boldsymbol{R}} \tag{23}$$

mit der unteren Einsdreiecksmatrix $\overline{\boldsymbol{L}} := \boldsymbol{P}_3\boldsymbol{L}_2\boldsymbol{F}_3^{-1}$, der oberen Dreiecksmatrix $\overline{\boldsymbol{R}} := \boldsymbol{F}_3\boldsymbol{R}_2$ und

$$\boldsymbol{F}_3 := \tilde{\boldsymbol{F}}_{n\,1,n} \cdots \tilde{\boldsymbol{F}}_{23}\tilde{\boldsymbol{F}}_{12}, \qquad \boldsymbol{P}_3 := \tilde{\boldsymbol{T}}_{n-1,s(n-1)} \cdots \tilde{\boldsymbol{T}}_{2,s(2)}\tilde{\boldsymbol{T}}_{1,s(1)}. \tag{24}$$

Einsetzen von (23) in (22) führt auf die gesuchte Faktorisierung von $\bar{\boldsymbol{A}}$. Zusammenfassend erhalten wir den folgenden Algorithmus:

6.3.4. *Aufdatierung einer Dreiecksfaktorisierung bei Rang-1-Modifikation*

Aufgabe: Bestimme eine Dreiecksfaktorisierung $\overline{\boldsymbol{P}}\bar{\boldsymbol{A}} = \overline{\boldsymbol{L}}\overline{\boldsymbol{R}}$ von

$$\bar{\boldsymbol{A}} = \boldsymbol{A} + \boldsymbol{u}\boldsymbol{v}^{\mathsf{T}}$$

bei gegebener Dreiecksfaktorisierung $\boldsymbol{P}\boldsymbol{A} = \boldsymbol{L}\boldsymbol{R}$ von $\boldsymbol{A}$ mit unteren Einsdreiecksmatrizen $\boldsymbol{L}, \overline{\boldsymbol{L}}$, oberen Dreiecksmatrizen $\boldsymbol{R}, \overline{\boldsymbol{R}}$ und Permutationsmatrizen $\boldsymbol{P}, \overline{\boldsymbol{P}}$.

Algorithmus:

S1: Stelle $\boldsymbol{u}$ in der Form

$$\boldsymbol{u} = \boldsymbol{P}^{\mathsf{T}}\boldsymbol{L}\boldsymbol{r},\ \boldsymbol{r} \text{ Lösung des Dreieckssystems } \boldsymbol{L}\boldsymbol{r} = \boldsymbol{P}\boldsymbol{u}, \tag{25}$$

dar. Schreibe $\bar{\boldsymbol{A}}$ als

$$\bar{\boldsymbol{A}} = \boldsymbol{P}^{\mathsf{T}}\boldsymbol{L}(\boldsymbol{R} + \boldsymbol{r}\boldsymbol{v}^{\mathsf{T}}) \tag{26}$$

S2: Führe $n - 1$ Elementareliminationsschritte gemäß 6.3.2 zur Transformation von $\boldsymbol{r}$ in $\varrho_1\boldsymbol{e}^1$ aus und transformiere $\boldsymbol{R}$ gemäß (20) auf obere Hessenbergform $\boldsymbol{R}_1 := \boldsymbol{F}_1\boldsymbol{R}$. Schreibe (26) als

$$\bar{\boldsymbol{A}} = (\boldsymbol{P}^{\mathsf{T}}\boldsymbol{P}_1^{\mathsf{T}})\,(\boldsymbol{P}_1\boldsymbol{L}\boldsymbol{F}_1^{-1})\,(\boldsymbol{F}_1\boldsymbol{R} + \boldsymbol{F}_1\boldsymbol{r}\boldsymbol{v}^{\mathsf{T}}) = \boldsymbol{P}_2^{\mathsf{T}}\boldsymbol{L}_1(\boldsymbol{R}_1 + \varrho_1\boldsymbol{e}^1\boldsymbol{v}^{\mathsf{T}}) = \boldsymbol{P}_2^{\mathsf{T}}\boldsymbol{L}_2\boldsymbol{R}_2 \tag{27}$$

mit der Permutationsmatrix $\boldsymbol{P}_2 := \boldsymbol{P}_1\boldsymbol{P}$, der unteren Einsdreiecksmatrix $\boldsymbol{L}_2 := \boldsymbol{L}_1 := \boldsymbol{P}_1\boldsymbol{L}\boldsymbol{F}_1^{-1}$ und der oberen Hessenbergmatrix $\boldsymbol{R}_2 = \boldsymbol{R}_1 + \varrho_1\boldsymbol{e}^1\boldsymbol{v}^{\mathsf{T}} = \boldsymbol{F}_1\boldsymbol{R} + \boldsymbol{F}_1\boldsymbol{r}\boldsymbol{v}^{\mathsf{T}}$.

S3: Führe $n - 1$ Elementareliminationsschritte gemäß 6.3.3 zur Transformation

von $\boldsymbol{R}_2$ auf obere Dreiecksform $\bar{\boldsymbol{R}}$ aus. Schreibe (27) als

$$\bar{\boldsymbol{A}} = (\boldsymbol{P}_2^{\mathsf{T}}\boldsymbol{P}_3^{\mathsf{T}})\,(\boldsymbol{P}_3\boldsymbol{L}_2\boldsymbol{F}_3^{-1})\,(\boldsymbol{F}_3\boldsymbol{R}_2) = \bar{\boldsymbol{P}}^{\mathsf{T}}\bar{\boldsymbol{L}}\bar{\boldsymbol{R}} \tag{28}$$

mit der Permutationsmatrix $\bar{\boldsymbol{P}} := \boldsymbol{P}_3\boldsymbol{P}_2$, der unteren Einsdreiecksmatrix $\bar{\boldsymbol{L}} := \boldsymbol{P}_3\boldsymbol{L}_2\boldsymbol{F}_3^{-1}$ und der oberen Dreiecksmatrix $\bar{\boldsymbol{R}} := \boldsymbol{F}_3\boldsymbol{R}_2$.

Aufwand: $\sim K_2 n^2$ opms mit $5/2 \leqq K_2 \leqq 9/2$

6.3.5. *Bemerkung.* (i) Algorithmus 6.3.4 läßt sich auf dem Platz von $\boldsymbol{L}$, $\boldsymbol{R}$ und $\boldsymbol{u}$ durchführen, indem $\boldsymbol{u}$ mit $\boldsymbol{r}$ und danach mit den Subdiagonalelementen von $\boldsymbol{R}_2$ überspeichert wird. Die Permutationsmatrix $\boldsymbol{P}$ kann z. B. durch die Tabelle $\{k(1), \ldots, k(n)\}$ als Integervektor entsprechend 3.1 repräsentiert werden. Der Multiplikation $\boldsymbol{T}_{i,s(i)}\boldsymbol{P}$ entspricht dann die Vertauschung der Elemente i und $s(i)$ der Tabelle, vgl. 5.2.6(ii). Der Aufwand ist mindestens $\sim 5/2n^2$ opms (wenn nur Fall 1 eintritt) und höchstens $\sim 9/2n^2$ opms (wenn stets Fall 2 eintritt).

(ii) Wenn bei der Faktorisierung neben Zeilenvertauschungen auch solche der Spalten gemäß

$$\boldsymbol{P}_Z\boldsymbol{A}\boldsymbol{P}_S^{\mathsf{T}} = \boldsymbol{L}\boldsymbol{R} \tag{29}$$

zugelassen werden, läßt sich für den Sonderfall $\boldsymbol{v} = \boldsymbol{e}^k$, d. h. für Spaltenmodifikationen

$$\bar{\boldsymbol{A}} = \boldsymbol{A} + \boldsymbol{u}\boldsymbol{e}^{k\mathsf{T}} \tag{30}$$

eine $\boldsymbol{LR}$-Faktorisierung

$$\bar{\boldsymbol{P}}_Z\bar{\boldsymbol{A}}\bar{\boldsymbol{P}}_S^{\mathsf{T}} = \bar{\boldsymbol{L}}\bar{\boldsymbol{R}} \tag{31}$$

von $\bar{\boldsymbol{A}}$ wie folgt berechnen:

S1: Schreibe (30) bei gegebener Faktorisierung (29) in der Form

$$\bar{\boldsymbol{A}} = \boldsymbol{P}_Z^{\mathsf{T}}\boldsymbol{L}\boldsymbol{R}\boldsymbol{P}_S + \boldsymbol{u}\boldsymbol{e}^{k\mathsf{T}} = \boldsymbol{P}_Z^{\mathsf{T}}\boldsymbol{L}[\boldsymbol{R} + \boldsymbol{r}\boldsymbol{e}^{l\mathsf{T}}]\,\boldsymbol{P}_S \tag{32}$$

mit

$$\boldsymbol{r} := \boldsymbol{L}^{-1}\boldsymbol{P}_Z\boldsymbol{u} \quad \text{und} \quad \boldsymbol{e}^l := \boldsymbol{P}_S\boldsymbol{e}^k.$$

S2: Es bezeichne $\boldsymbol{P}_1$ diejenige Permutationsmatrix, die

$$\boldsymbol{A} = (\boldsymbol{a}^1, \ldots, \boldsymbol{a}^l, \ldots, \boldsymbol{a}^n) \text{ in } \boldsymbol{A}\boldsymbol{P}_1^{\mathsf{T}} = (\boldsymbol{a}^1, \ldots, \boldsymbol{a}^{l-1}, \boldsymbol{a}^{l+1}, \ldots, \boldsymbol{a}^n, \boldsymbol{a}^l)$$

transformiert. Schreibe (32) als

$$\bar{\boldsymbol{A}} = \boldsymbol{P}_Z^{\mathsf{T}}\boldsymbol{L}[\boldsymbol{R}\boldsymbol{P}_1^{\mathsf{T}} + \boldsymbol{r}\boldsymbol{e}^{l\mathsf{T}}\,\boldsymbol{P}_1^{\mathsf{T}}]\,\boldsymbol{P}_1\boldsymbol{P}_S =: \boldsymbol{P}_Z^{\mathsf{T}}\boldsymbol{L}_2\boldsymbol{R}_2\boldsymbol{P}_2 \tag{33}$$

mit $\boldsymbol{L}_2 := \boldsymbol{L}$, der Permutationsmatrix $\boldsymbol{P}_2 := \boldsymbol{P}_1\boldsymbol{P}_S$ und der oberen Hessenbergmatrix

$\boldsymbol{R}_2 := \boldsymbol{R}\boldsymbol{P}_1^{\mathsf{T}} + \boldsymbol{r}\boldsymbol{e}^{l\mathsf{T}}\boldsymbol{P}_1^{\mathsf{T}} = \boldsymbol{R}\boldsymbol{P}_1^{\mathsf{T}} + \boldsymbol{r}\boldsymbol{e}^{n\mathsf{T}}$ des Typs

$$\boldsymbol{R}_2 = \begin{array}{c} \begin{matrix} & & & l-1 & l & & \\ \end{matrix} \\ \begin{pmatrix} \times & \cdots & \times & \times & \cdots & \times \\ & \ddots & & & & \\ & & \times & \times & \cdots & \times \\ & & & \times & \cdots & \times \\ & & & \times & \ddots & \\ & & & & \ddots & \times \\ & & & & & \times \end{pmatrix} \begin{matrix} \\ \\ l-1 \\ l \\ l+1 \\ \\ \end{matrix} \end{array}$$

S3: Führe $n - l$ Elementareliminationsschritte analog zu 6.3.3 zur Transformation der Subdiagonalelemente $(\boldsymbol{R}_2)_{k+1,k}$ $(k = l, \ldots, n-1)$ in 0 gemäß

$$\boldsymbol{L}_2\boldsymbol{R}_2 = \boldsymbol{P}_3^{\mathsf{T}}(\boldsymbol{P}_3\boldsymbol{L}_2\boldsymbol{F}_3^{-1})\,(\boldsymbol{F}_3\boldsymbol{R}_2) = \boldsymbol{P}_3^{\mathsf{T}}\overline{\boldsymbol{L}}\,\overline{\boldsymbol{R}}$$

durch, wobei $\overline{\boldsymbol{L}} := \boldsymbol{P}_3\boldsymbol{L}_2\boldsymbol{F}_3^{-1}$ eine untere Einsdreiecksmatrix, $\overline{\boldsymbol{R}} := \boldsymbol{F}_3\boldsymbol{R}_2$ eine obere Dreiecksmatrix und

$$\boldsymbol{F}_3 := \tilde{\boldsymbol{F}}_{n-1,n} \cdots \tilde{\boldsymbol{F}}_{l,l+1}, \qquad \boldsymbol{P}_3 := \tilde{\boldsymbol{T}}_{n-1,s(n-1)} \cdots \tilde{\boldsymbol{T}}_{l,s(l)}$$

ist. Schreibe (33) in der Form

$$\bar{\boldsymbol{A}} = \boldsymbol{P}_Z^{\mathsf{T}}\boldsymbol{L}_2\boldsymbol{R}_2\boldsymbol{P}_2 = \boldsymbol{P}_Z^{\mathsf{T}}\boldsymbol{P}_3^{\mathsf{T}}\overline{\boldsymbol{L}}\,\overline{\boldsymbol{R}}\boldsymbol{P}_2 =: \overline{\boldsymbol{P}}_Z^{\mathsf{T}}\overline{\boldsymbol{L}}\,\overline{\boldsymbol{R}}\,\overline{\boldsymbol{P}}_S \tag{34}$$

mit den Permutationsmatrizen $\overline{\boldsymbol{P}}_Z := \boldsymbol{P}_3\boldsymbol{P}_Z$ und $\overline{\boldsymbol{P}}_S := \boldsymbol{P}_2 = \boldsymbol{P}_1\boldsymbol{P}_S$.

Die in S2 von 6.3.4 erforderlichen $n - 1$ Elementareliminationsschritte werden hier durch die Spaltenvertauschungen gemäß P_1 ersetzt, und in S3 reduziert sich die Anzahl der Elementareliminationsschritte von $n - 1$ auf $n - l$.

(iii) In 6.3.2 und S2 von 6.3.4 genügt es, nur $n - 2$ Schritte zur Erzeugung der Nullen in den Komponenten (r_i) $(i = n, \ldots, 3)$ durchzuführen, da bereits damit die Hessenbergform von $\boldsymbol{R}_2$ erreicht wird.

(iv) Man kann zeigen, daß Algorithmus 6.3.4 im folgenden Sinne numerisch gutartig ist: Wenn $\boldsymbol{P}(\boldsymbol{A} + \delta\boldsymbol{A}) = \boldsymbol{L}\boldsymbol{R}$ gilt, existiert eine Störung $\delta\bar{\boldsymbol{A}}$, die den Beziehungen

$$\overline{\boldsymbol{P}}(\bar{\boldsymbol{A}} + \delta\bar{\boldsymbol{A}}) = \overline{\boldsymbol{L}}\,\overline{\boldsymbol{R}} \quad \text{und} \quad \|\delta\bar{\boldsymbol{A}}\| \leqq \|\delta\boldsymbol{A}\| + v\,K\,\mathrm{cond}\,(\overline{\boldsymbol{L}})\,(\|\boldsymbol{A}\| + \|\bar{\boldsymbol{A}}\|)$$

genügt mit einer nur von n abhängigen Konstanten $K = K(n)$. □

B. *Symmetrische definite Matrizen*

Wir setzen $\boldsymbol{A}$ als symmetrisch und positiv definit voraus. Damit die modifizierte Matrix $\bar{\boldsymbol{A}}$ dann auch symmetrisch ist, muß die Rang-1-Korrektur $\boldsymbol{u}\boldsymbol{v}^{\mathsf{T}}$ ebenfalls symmetrisch sein, d. h., es muß $\boldsymbol{u} = \alpha\boldsymbol{v}$ und folglich

$$\bar{\boldsymbol{A}} = \boldsymbol{A} + \alpha\boldsymbol{v}\boldsymbol{v}^{\mathsf{T}} \tag{35}$$

mit einem $\alpha \in \mathbf{R}$ gelten. Ohne Einschränkung der Allgemeinheit kann dabei $\boldsymbol{v} \neq \boldsymbol{o}$

und $\alpha = 1$ oder $\alpha = -1$ vorausgesetzt werden, denn die Korrektur läßt sich stets in der Gestalt $\alpha \boldsymbol{v}\boldsymbol{v}^\mathsf{T} = \operatorname{sgn}(\alpha)\,(\sqrt{|\alpha|}\,\boldsymbol{v})\,(\sqrt{|\alpha|}\,\boldsymbol{v})^\mathsf{T} = \pm\tilde{\boldsymbol{v}}\tilde{\boldsymbol{v}}^\mathsf{T}$ schreiben.

Wir fragen als nächstes, unter welchen Voraussetzungen $\boldsymbol{A}$ positiv definit ist. Unter Verwendung der Cholesky-Faktorisierung

$$\boldsymbol{A} = \boldsymbol{L}\boldsymbol{L}^\mathsf{T}$$

ergibt sich aus (35)

$$\bar{\boldsymbol{A}} = \boldsymbol{L}\boldsymbol{L}^\mathsf{T} + \alpha\boldsymbol{v}\boldsymbol{v}^\mathsf{T} = \boldsymbol{L}(\boldsymbol{I} + \alpha\boldsymbol{r}\boldsymbol{r}^\mathsf{T})\,\boldsymbol{L}^\mathsf{T}, \qquad \boldsymbol{r} \text{ Lösung von } \boldsymbol{L}\boldsymbol{r} = \boldsymbol{v}. \tag{36}$$

Die Matrizen $\bar{\boldsymbol{A}}$ und $\boldsymbol{M} := \boldsymbol{I} + \alpha\boldsymbol{r}\boldsymbol{r}^\mathsf{T}$ sind also kongruent und besitzen folglich dieselben Definitheitseigenschaften, vgl. 1.2.C. Im Fall $\alpha = 1$ gilt

$$\boldsymbol{x}^\mathsf{T}\boldsymbol{M}\boldsymbol{x} = \boldsymbol{x}^\mathsf{T}\boldsymbol{x} + (\boldsymbol{x}^\mathsf{T}\boldsymbol{r})^2 \geqq \boldsymbol{x}^\mathsf{T}\boldsymbol{x}$$

für alle $\boldsymbol{x} \in \mathbf{R}^n$, d. h., durch die Addition der positiv definiten Rang-1-Korrektur $\boldsymbol{v}\boldsymbol{v}^\mathsf{T}$ verschlechtert sich die positive Definitheit unabhängig von $\boldsymbol{v}$ nicht. Im Fall $\alpha = -1$ gilt wegen $|\boldsymbol{x}^\mathsf{T}\boldsymbol{r}| \leqq \|\boldsymbol{x}\|_2\,\|\boldsymbol{r}\|_2$ dagegen

$$\boldsymbol{x}^\mathsf{T}\boldsymbol{M}\boldsymbol{x} = \boldsymbol{x}^\mathsf{T}\boldsymbol{x} - (\boldsymbol{x}^\mathsf{T}\boldsymbol{r})^2 \geqq \|\boldsymbol{x}\|_2^2\,(1 - \|\boldsymbol{r}\|_2^2),$$

und für $\boldsymbol{x} = \lambda\boldsymbol{r}$ steht das Gleichheitszeichen. In diesem Fall verschlechtert sich also die positive Definitheit. Sie geht verloren, wenn die Bedingung

$$\|\boldsymbol{r}\|_2^2 = \|\boldsymbol{L}^{-1}\boldsymbol{v}\|_2^2 = \boldsymbol{v}^\mathsf{T}(\boldsymbol{L}\boldsymbol{L}^\mathsf{T})^{-1}\,\boldsymbol{v} = \boldsymbol{v}^\mathsf{T}\boldsymbol{A}^{-1}\boldsymbol{v} < 1 \tag{37}$$

nicht mehr erfüllt ist. Bei den Aufdatierungsalgorithmen wird deshalb auch zwischen den Fällen $\alpha = 1$ und $\alpha = -1$ unterschieden. Wegen der Symmetrie der Faktorisierung können dabei die im nichtsymmetrischen Fall verwendeten Elementareliminationsmatrizen $\boldsymbol{F}_{k,k+1}$ durch Givens-Drehungen $\boldsymbol{G}_{ij}$ ersetzt werden, womit sich wegen $\|\boldsymbol{G}_{ij}\|_2 = 1$ günstigere Stabilitätseigenschaften ergeben.

6.3.6. *Aufdatierung einer Cholesky-Faktorisierung bei Addition einer Rang-1-Korrektur.*

Aufgabe: Bestimme die Cholesky-Faktorisierung $\bar{\boldsymbol{A}} = \bar{\boldsymbol{L}}\bar{\boldsymbol{L}}^\mathsf{T}$ von

$$\bar{\boldsymbol{A}} = \boldsymbol{A} + \boldsymbol{v}\boldsymbol{v}^\mathsf{T} \tag{38}$$

bei gegebener Cholesky-Faktorisierung $\boldsymbol{A} = \boldsymbol{L}\boldsymbol{L}^\mathsf{T}$ der positiv definiten Matrix $\boldsymbol{A} \in \mathbf{S}^{n,n}$.

Algorithmus:

Setze $\boldsymbol{L}_1 := (\boldsymbol{L} \mid \boldsymbol{v})$ und schreibe (38) in der Form

$$\bar{\boldsymbol{A}} = \boldsymbol{L}\boldsymbol{L}^\mathsf{T} + \boldsymbol{v}\boldsymbol{v}^\mathsf{T} = \boldsymbol{L}_1\boldsymbol{L}_1^\mathsf{T}. \tag{39}$$

Bestimme n Givens-Drehungen $\boldsymbol{G}_{j,n+1}$ $(j = 1, \ldots, n)$ so, daß die Elemente $(\boldsymbol{L}_1^\mathsf{T})_{n+1,j} = v_j$ in der $(n+1)$-ten Zeile von $\boldsymbol{L}_1^\mathsf{T}$ sukzessive zu 0 gemacht werden. Dann gilt

$$\boldsymbol{G}\boldsymbol{L}_1^\mathsf{T} := \boldsymbol{G}_{n,n+1} \cdots \boldsymbol{G}_{1,n+1}\boldsymbol{L}_1^\mathsf{T} =: \boldsymbol{L}_2^\mathsf{T} =: \begin{pmatrix} \bar{\boldsymbol{L}}^\mathsf{T} \\ \boldsymbol{o}^\mathsf{T} \end{pmatrix} \tag{40}$$

mit orthogonalem $\boldsymbol{G} := \boldsymbol{G}_{n,n+1} \cdots \boldsymbol{G}_{1,n+1}$ und der oberen Dreiecksmatrix $\bar{\boldsymbol{L}}^T$ bzw. $\boldsymbol{L}_2^T$. Schreibe (39) in der Form

$$\bar{\boldsymbol{A}} = \boldsymbol{L}_1\boldsymbol{L}_1^T = (\boldsymbol{L}_1\boldsymbol{G}^T)(\boldsymbol{G}\boldsymbol{L}_1^T) = \boldsymbol{L}_2\boldsymbol{L}_2^T = (\bar{\boldsymbol{L}} \mid \boldsymbol{o})\begin{pmatrix}\bar{\boldsymbol{L}}^T\\ \hline \boldsymbol{o}^T\end{pmatrix} = \bar{\boldsymbol{L}}\bar{\boldsymbol{L}}^T. \tag{41}$$

Aufwand: $\sim n^2(2\ \text{opm} + 1\ \text{ops}) + \sim n\ \text{opr}$

6.3.7. *Bemerkung.* (i) In-situ-Realisierung von 6.3.6 auf dem Platz von $\boldsymbol{L}$ und $\boldsymbol{v}$ ist möglich. Die Transformation (40) verläuft dabei für $n = 3$ nach dem folgenden Muster:

$$\underset{\boldsymbol{L}_1^T}{\left(\begin{array}{ccc}\times & \times & \times\\ & \times & \times\\ & & \times\\ \hline \times & \times & \times\end{array}\right)} \xrightarrow{\boldsymbol{G}_{14}} \left(\begin{array}{ccc}\boxtimes & \boxtimes & \boxtimes\\ & \times & \times\\ & & \times\\ \hline 0 & \boxtimes & \boxtimes\end{array}\right) \xrightarrow{\boldsymbol{G}_{24}} \left(\begin{array}{ccc}\times & \times & \times\\ & \boxtimes & \boxtimes\\ & & \times\\ \hline 0 & 0 & \boxtimes\end{array}\right) \xrightarrow{\boldsymbol{G}_{34}} \underset{\boldsymbol{L}_2^T}{\left(\begin{array}{ccc}\times & \times & \times\\ & \times & \times\\ & & \boxtimes\\ \hline 0 & 0 & 0\end{array}\right)}$$

(ii) Die Drehungsparameter $\{c, s\}$ von $\boldsymbol{G}_{j,n+1}$ können stets so gewählt werden, daß $(\bar{\boldsymbol{L}})_{jj} > 0$ $(j = 1, \ldots, n)$ gilt. Dies ist bei der Festlegung $c := l_{jj}/\varrho$, $s := v_j/\varrho$ mit $\varrho := \sqrt{l_{jj}^2 + v_j^2}$ der Fall. Bei Festlegung nach 3.3.1 kann $(\bar{\boldsymbol{L}})_{jj} < 0$ für gewisse j gelten. In diesem Fall braucht nur das Vorzeichen der j-ten Zeile von $\bar{\boldsymbol{L}}^T$ geändert zu werden, denn es gilt $\bar{\boldsymbol{L}}\bar{\boldsymbol{L}}^T = (\bar{\boldsymbol{L}}\boldsymbol{D})(\boldsymbol{D}\bar{\boldsymbol{L}}^T)$ für jedes $\boldsymbol{D} = \operatorname{diag}(d_i)$ mit $d_i = \pm 1$. □

Bei Subtraktion von $\boldsymbol{v}\boldsymbol{v}^T$ ist ein etwas aufwendigerer Algorithmus erforderlich.

6.3.8. *Aufdatierung einer Cholesky-Faktorisierung bei Subtraktion einer Rang-1-Korrektur.*

Aufgabe: Bestimme die Cholesky-Faktorisierung $\bar{\boldsymbol{A}} = \bar{\boldsymbol{L}}\bar{\boldsymbol{L}}^T$ der als positiv definit vorausgesetzten Matrix

$$\bar{\boldsymbol{A}} = \boldsymbol{A} - \boldsymbol{v}\boldsymbol{v}^T \tag{42}$$

bei gegebener Cholesky-Faktorisierung $\boldsymbol{A} = \boldsymbol{L}\boldsymbol{L}^T$ der positiv definiten Matrix $\boldsymbol{A} \in \mathbf{S}^{n,n}$.

Algorithmus:

S1: Berechne $\boldsymbol{r}$ als Lösung des Dreieckssystems

$$\boldsymbol{L}\boldsymbol{r} = \boldsymbol{v}. \tag{43}$$

Falls $\|\boldsymbol{r}\|_2 \geqq 1$, stop ($\bar{\boldsymbol{A}}$ nicht positiv definit).
Setze

$$\boldsymbol{r}^1 := \begin{pmatrix}\boldsymbol{r}\\ \hline \varrho\end{pmatrix} \quad \text{mit} \quad \varrho := \sqrt{1 - \|\boldsymbol{r}\|_2^2}; \tag{44}$$

man beachte

$$\|\boldsymbol{r}^1\|_2^2 = \|\boldsymbol{r}\|_2^2 + \varrho^2 = 1.$$

S2: Bestimme n Givens-Drehungen $\boldsymbol{G}_{i,n+1}$ $(i = n, \ldots, 1)$ so, daß die Komponenten r_i^1 von $\boldsymbol{r}^1$ sukzessive zu 0 gemacht werden. Dann gilt $\boldsymbol{G}\boldsymbol{r}^1 := \boldsymbol{G}_{1,n+1} \cdots \boldsymbol{G}_{n,n+1}\boldsymbol{r}^1 = \begin{pmatrix} \boldsymbol{o} \\ \sigma \end{pmatrix} = \sigma \boldsymbol{e}^{n+1}$ mit $\sigma = \pm\|\boldsymbol{r}^1\| = \pm 1$. Setze $\boldsymbol{L}_1 := (\boldsymbol{L} \mid \boldsymbol{o})$ und bilde

$$\boldsymbol{G}\boldsymbol{L}_1^\mathsf{T} = \boldsymbol{G}_{1,n+1} \cdots \boldsymbol{G}_{n,n+1}\boldsymbol{L}_1^\mathsf{T} =: \boldsymbol{L}_2^\mathsf{T} =: \begin{pmatrix} \overline{\boldsymbol{L}}^\mathsf{T} \\ \overline{\boldsymbol{v}}^\mathsf{T} \end{pmatrix}.$$

Dabei ist

$$\boldsymbol{v} = \boldsymbol{L}\boldsymbol{r} = \boldsymbol{L}_1\boldsymbol{r}^1 = (\boldsymbol{L}_1\boldsymbol{G}^\mathsf{T})(\boldsymbol{G}\boldsymbol{r}^1) = \boldsymbol{L}_2\sigma\boldsymbol{e}^{n+1} = (\overline{\boldsymbol{L}} \mid \overline{\boldsymbol{v}}) \begin{pmatrix} \boldsymbol{o} \\ \sigma \end{pmatrix} = \sigma\overline{\boldsymbol{v}},$$

also $\overline{\boldsymbol{v}} = \sigma\boldsymbol{v}$ und folglich

$$\boldsymbol{L}\boldsymbol{L}^\mathsf{T} = \boldsymbol{L}_1\boldsymbol{L}_1^\mathsf{T} = (\boldsymbol{L}_1\boldsymbol{G}^\mathsf{T})(\boldsymbol{G}\boldsymbol{L}_1^\mathsf{T}) = \boldsymbol{L}_2\boldsymbol{L}_2^\mathsf{T} = \overline{\boldsymbol{L}}\overline{\boldsymbol{L}}^\mathsf{T} + \overline{\boldsymbol{v}}\overline{\boldsymbol{v}}^\mathsf{T} = \overline{\boldsymbol{L}}\overline{\boldsymbol{L}}^\mathsf{T} + \sigma^2\boldsymbol{v}\boldsymbol{v}^\mathsf{T},$$

d. h., es gilt

$$\overline{\boldsymbol{L}}\overline{\boldsymbol{L}}^\mathsf{T} = \boldsymbol{L}\boldsymbol{L}^\mathsf{T} - \boldsymbol{v}\boldsymbol{v}^\mathsf{T} = \bar{\boldsymbol{A}}. \tag{45}$$

Aufwand: $\sim n^2(5/2 \text{ opm} + 3/2 \text{ ops}) + n \text{ opr}$

6.3.9. *Bemerkung.* (i) In-situ-Realisierung von 6.3.8 auf dem Platz von $\boldsymbol{L}$ und $\boldsymbol{v}$ ist möglich. Für $n = 3$ erfolgt dabei die Transformation $(\boldsymbol{r}^1 \mid \boldsymbol{L}_1^\mathsf{T}) \to (\sigma\boldsymbol{e}^{n+1} \mid \boldsymbol{L}_2^\mathsf{T})$ nach folgendem Muster:

$$\left(\begin{array}{c|ccc} \times & \times & \times & \times \\ \times & & \times & \times \\ \times & & & \times \\ \hline \times & & & \end{array}\right) \xrightarrow{\boldsymbol{G}_{34}} \left(\begin{array}{c|ccc} \times & \times & \times & \times \\ \times & & \times & \times \\ 0 & & & \boxtimes \\ \hline \boxtimes & & & \boxtimes \end{array}\right)$$

$$\xrightarrow{\boldsymbol{G}_{24}} \left(\begin{array}{c|ccc} \times & \times & \times & \times \\ 0 & & \boxtimes & \boxtimes \\ 0 & & & \times \\ \hline \boxtimes & & \boxtimes & \boxtimes \end{array}\right) \xrightarrow{\boldsymbol{G}_{14}} \left(\begin{array}{c|ccc} 0 & \boxtimes & \boxtimes & \boxtimes \\ 0 & & \times & \times \\ 0 & & & \times \\ \hline \boxtimes & \boxtimes & \boxtimes & \boxtimes \end{array}\right)$$

(ii) Die Drehungen $\boldsymbol{G}_{i,n+1}$ können so gewählt werden, daß $(\overline{\boldsymbol{L}})_{ii} > 0$ gilt, bzw. es läßt sich $(\overline{\boldsymbol{L}}_{ii}) > 0$ $(i = 1, \ldots, n)$ durch Vorzeichenwechsel in den Zeilen von $\overline{\boldsymbol{L}}^\mathsf{T}$ erzielen, vgl. 6.3.7(ii). □

Das Stabilitätsverhalten der Aufdatierungsalgorithmen hängt davon ab, ob $\alpha = 1$ oder $\alpha = -1$ ist.

6.3.10. *Bemerkung.* Es gelte $\boldsymbol{A} + \delta\boldsymbol{A} = \boldsymbol{L}\boldsymbol{L}^\mathsf{T}$ mit einer symmetrischen und bezüglich $\boldsymbol{A}$ kleinen Störung $\delta\boldsymbol{A}$. Dann gibt es im Fall $\alpha = 1$ eine symmetrische Störung $\delta\bar{\boldsymbol{A}}$, so daß der gemäß 6.3.6 berechnete Faktor $\overline{\boldsymbol{L}}$ der Beziehung $\bar{\boldsymbol{A}} + \delta\bar{\boldsymbol{A}} = \overline{\boldsymbol{L}}\overline{\boldsymbol{L}}^\mathsf{T}$ genügt, und $\delta\bar{\boldsymbol{A}}$ ist klein in bezug auf $\bar{\boldsymbol{A}}$. Im Fall $\alpha = 1$ ist der Aufdatierungsalgorithmus 6.3.6 also stets numerisch gutartig. Im Fall $\alpha = -1$ trifft dies für 6.3.8 auch zu, sofern $\|\bar{\boldsymbol{A}}\|$ nicht wesentlich kleiner als $\|\boldsymbol{A}\|$ ist. □

Etwa in der nichtlinearen Optimierung treten symmetrische und nach Konstruktion positiv definite Matrizen $\boldsymbol{A}, \bar{\boldsymbol{A}}$ auf, die sich durch eine Rang-2-Modifikation unterscheiden. Zur Aufdatierung der zugehörigen Cholesky-Faktorisierungen siehe Ü 6.3.2 unten. Die BKP-Faktorisierung finiter symmetrischer Matrizen läßt sich ebenfalls aufdatieren, allerdings ist der Algorithmus relativ kompliziert, siehe B 6.3.

Übungsaufgaben

Ü 6.3.1. Es sei $\boldsymbol{M} = \boldsymbol{I} + \alpha \boldsymbol{r}\boldsymbol{r}^{\mathsf{T}}$ mit $\boldsymbol{r} \in \mathbf{R}^n$, $\boldsymbol{r} \neq \boldsymbol{o}$, $\alpha \in \mathbf{R}$, $\alpha \neq 0$. Man zeige, daß $\boldsymbol{M}$ den einfachen Eigenwert $\lambda_1 := 1 + \alpha \|\boldsymbol{r}\|_2^2$ mit dem zugehörigen Eigenvektor $\boldsymbol{r}$ sowie den $(n-1)$-fachen Eigenwert $\lambda_2 = \cdots = \lambda_n = 1$ besitzt, wobei der zugehörige Eigenraum aus allen zu $\boldsymbol{r}$ orthogonalen Vektoren besteht.

Ü 6.3.2. Es sei $\boldsymbol{M} \in \mathbf{S}^{n,n}$ eine Matrix mit rang $(\boldsymbol{M}) = 2$. Man zeige:

(i) $\boldsymbol{M}$ läßt sich in der Form

$$\boldsymbol{M} = \alpha_1 \boldsymbol{v}^1 \boldsymbol{v}^{1\mathsf{T}} + \alpha_2 \boldsymbol{v}^2 \boldsymbol{v}^{2\mathsf{T}} \quad \text{mit } \alpha_1, \alpha_2 \in \{-1, +1\} \quad \text{und} \quad \boldsymbol{v}^{1\mathsf{T}} \boldsymbol{v}^2 = 0 \tag{46}$$

darstellen.

(ii) Für die Matrix

$$\boldsymbol{M} := \boldsymbol{u}\boldsymbol{v}^{\mathsf{T}} + \boldsymbol{v}\boldsymbol{u}^{\mathsf{T}}, \qquad \boldsymbol{u}, \boldsymbol{v} \in \mathbf{R}^n \text{ linear unabhängig}, \tag{47}$$

lautet die Zerlegung (46)

$$\boldsymbol{M} = \boldsymbol{v}^1 \boldsymbol{v}^{1\mathsf{T}} - \boldsymbol{v}^2 \boldsymbol{v}^{2\mathsf{T}} \tag{48}$$

mit

$$\boldsymbol{v}^1 := \varrho \left(\frac{\boldsymbol{u}}{\|\boldsymbol{u}\|_2} + \frac{\boldsymbol{v}}{\|\boldsymbol{v}\|_2} \right), \qquad \boldsymbol{v}^2 := \varrho \left(\frac{\boldsymbol{u}}{\|\boldsymbol{u}\|_2} - \frac{\boldsymbol{v}}{\|\boldsymbol{v}\|_2} \right), \qquad \varrho := \left(\frac{\|\boldsymbol{u}\| \, \|\boldsymbol{v}\|}{2} \right)^{1/2}. \tag{49}$$

(iii) Sind $\boldsymbol{A}$ und $\bar{\boldsymbol{A}} := \boldsymbol{A} + \boldsymbol{M}$ beide symmetrisch und positiv definit, so kann die Cholesky-Faktorisierung von $\bar{\boldsymbol{A}}$ aus der von $\boldsymbol{A}$ durch zweimalige Anwendung der Rang-1-Aufdatierungsalgorithmen auf die Modifikationen

$$\boldsymbol{A} \to \boldsymbol{A}_1 := \boldsymbol{A} + \alpha_1 \boldsymbol{v}^1 \boldsymbol{v}^{1\mathsf{T}} \to \bar{\boldsymbol{A}} := \boldsymbol{A}_1 + \alpha_2 \boldsymbol{v}^2 \boldsymbol{v}^{2\mathsf{T}} \tag{50}$$

berechnet werden. Im Fall unterschiedlichen Vorzeichens der α_i muß dabei $\alpha_1 = +1$ gewählt werden, damit die intermediäre Matrix $\boldsymbol{A}_1$ positiv definit ist und der zugehörige Cholesky-Faktor $\boldsymbol{L}_1$ existiert.

6.4. Dreiecksfaktorisierung schwach besetzter Matrizen

Eine Reihe von praktischen Problemen führt auf lineare Gleichungssysteme hoher Dimension mit schwach besetzten Koeffizientenmatrizen, vgl. 2.1.A1. Solche Matrizen sind durch die Eigenschaft

$$\eta(\boldsymbol{A}) \ll n^2$$

charakterisiert, wobei $\eta(\boldsymbol{A})$ die Anzahl der *Nichtnullelemente* — im folgenden mit NNE abgekürzt — von $\boldsymbol{A}$ bezeichnet. Für realistische Anwendungen liegt n in der Größenordnung von 10^2 bis 10^4 und größer, so daß es notwendig wird, die schwache Besetztheit sowohl im Hinblick auf den benötigten Speicherplatz — bei voll be-

setzten Matrizen $\sim K_1 n^2$ S — als auch bezüglich des Rechenaufwandes — bei voll besetzten Matrizen $\sim K_2 n^3$ opms — auszunutzen. In vielen Fällen ist die mittlere Anzahl der NNE pro Zeile unabhängig von n konstant gleich m, so daß $\eta(\boldsymbol{A}) = mn$ mit n wächst, aber der Besetztheitsgrad $\eta(\boldsymbol{A})/n^2 = m/n$ mit wachsendem n immer kleiner wird.

Die Vorgehensweise bei der Dreiecksfaktorisierung schwach besetzter Matrizen hängt wesentlich vom Besetztheitsmuster von $\boldsymbol{A}$, d. h. von der Anordnung der NNE ab.

A. *Bandmatrizen*

Eine Matrix $\boldsymbol{A} = (a_{ij}) \in \mathbf{R}^{n,n}$ mit

$$a_{ij} = 0 \quad \text{für} \quad j < i - p \text{ und } j > i + q \tag{1}$$

heißt *Bandmatrix* des Typs $\{p, q\}$; die Zahlen p bzw. q heißen *untere* bzw. *obere Bandbreite*, und $l := p + q + 1$ wird *Bandbreite* schlechthin genannt, vgl. 1.1.C. Die durch die Diagonale sowie die benachbarten p unteren und q oberen Nebendiagonalen gebildeten Plätze werden das *Band* von $\boldsymbol{A}$ genannt. Ein Beispiel mit $n = 8$, $p = 2$, $q = 3$ stellt die Matrix

$$\boldsymbol{A} = \begin{array}{c} \\ p \\ \\ \\ \\ \\ \\ \\ \end{array} \overbrace{}^{q} \begin{pmatrix} 11 & 12 & 13 & 14 & & & & \\ 21 & 22 & 23 & 24 & 25 & & & \\ 31 & 32 & 33 & 34 & 35 & 36 & & \\ & 42 & 43 & 44 & 45 & 46 & 47 & \\ & & 53 & 54 & 55 & 56 & 57 & 58 \\ & & & 64 & 65 & 66 & 67 & 68 \\ & & & & 75 & 76 & 77 & 78 \\ & & & & & 86 & 87 & 88 \end{pmatrix} \tag{2}$$

dar. Die signifikanten Elemente von $\boldsymbol{A}$ können zeilenweise in einem rechteckigen Feld der Dimension (n, l) oder spaltenweise in einem Feld der Dimension (l, n) gespeichert werden; für die Matrix (1) führt dies auf

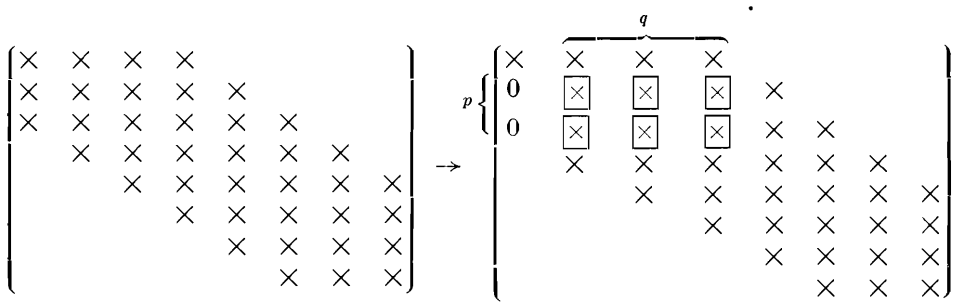
$$\begin{pmatrix} & & 11 & 12 & 13 & 14 \\ & 21 & 22 & 23 & 24 & 25 \\ 31 & 32 & 33 & 34 & 35 & 36 \\ 42 & 43 & 44 & 45 & 46 & 47 \\ 53 & 54 & 55 & 56 & 57 & 58 \\ 64 & 65 & 66 & 67 & 68 & \\ 75 & 76 & 77 & 78 & & \\ 86 & 87 & 88 & & & \end{pmatrix} \quad \text{bzw.} \quad \begin{pmatrix} & & & 14 & 25 & 36 & 47 & 58 \\ & & 13 & 24 & 35 & 46 & 57 & 68 \\ & 12 & 23 & 34 & 45 & 56 & 67 & 78 \\ 11 & 22 & 33 & 44 & 55 & 66 & 77 & 88 \\ 21 & 32 & 43 & 54 & 65 & 76 & 87 & \\ 31 & 42 & 53 & 64 & 75 & 86 & & \end{pmatrix}. \tag{3}$$

Die $p(p + 1)/2 + q(q + 1)/2$ Plätze in den Ecken werden zum Speichern des Bandes nicht benötigt; für genügend großes n spielen sie in bezug auf den Gesamtspeicherplatz keine Rolle.

Wir betrachten im folgenden den ersten Schritt des Gaußschen Algorithmus, in der Terminologie von 5.1 also den Übergang von $A = A^{(1)}$ zu $A^{(2)}$, und unterscheiden dabei zwei Fälle:

Fall 1: *Keine Pivotisierung*, d. h., $a_{11} \neq 0$ wird als Pivot gewählt.

Für die Matrix (2) ergibt sich dabei das Muster

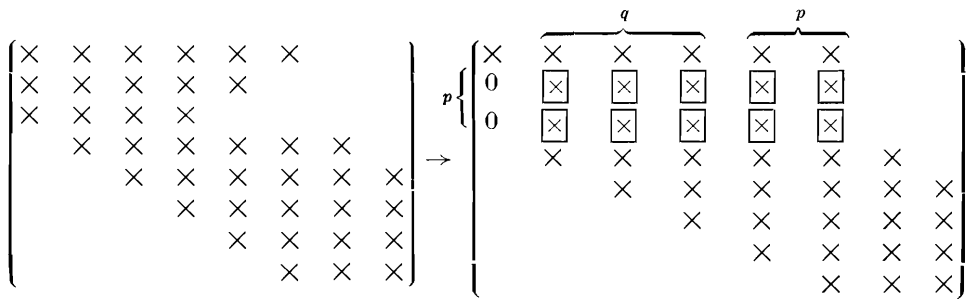


$A = A^{(1)} = \hat{A}^{(1)}$ $\qquad$ $A^{(2)}$

Beim Übergang zu $A^{(2)}$ werden nur die eingerahmten $p * q$ Elemente neu berechnet, insbesondere bleiben die Nullen außerhalb des Bandes erhalten. Das Verfahren kann daher auf dem Platz des Bandes von A ausgeführt werden. Die zu 0 gemachten Elemente werden mit den Eliminationskoeffizienten l_{ik} überspeichert.

Fall 2: *Spaltenpivotisierung*, d. h. a_{s1} mit $|a_{s1}| = \max\{|a_{i1}| : 1 \leqq i \leqq 1 + p\}$ wird als Pivot gewählt, und die Zeilen 1 und s werden vertauscht.

Für die Matrix (2) ergibt sich hier im ungünstigsten Fall $s = p + 1$ das Muster



$\hat{A}^{(1)}$ $\qquad$ $A^{(2)}$

Beim Übergang zu $A^{(2)}$ werden die ersten $1 + p$ Zeilen bis zur Spalte $1 + q + p$ aufgefüllt. Im nächsten Schritt kann derselbe Effekt wieder auftreten, wobei allerdings der entstehende Dreiecksfaktor R die obere Bandbreite $q + p$ nicht überschreitet.

Zusammenfassend erhalten wir das folgende Resultat:

6.4.1. *Aussage.* Es sei $\boldsymbol{A} \in \mathbf{R}^{n,n}$ eine reguläre Bandmatrix des Typs $\{p, q\}$. Dann gilt:

(i) Wenn die Dreiecksfaktorisierung nach dem Gaußschen Algorithmus ohne Pivotisierung durchführbar ist (etwa für diagonaldominantes oder symmetrisches und positiv definites $\boldsymbol{A}$, vgl. 5.2.D), sind die Dreiecksfaktoren $\boldsymbol{L}$ und $\boldsymbol{R}$ Bandmatrizen des Typs $\{p, 0\}$ und $\{0, q\}$. Die Faktorisierung kann auf dem Platz des Bandes von $\boldsymbol{A}$ ausgeführt werden; der Aufwand ist $\sim n(p$ opm $+ p * q$ opms). Im symmetrischen und positiv definiten Fall ist die Faktorisierung $\boldsymbol{A} = \boldsymbol{L}\boldsymbol{D}\boldsymbol{L}^{\mathsf{T}}$ auf dem unteren Teil des Bandes von $\boldsymbol{A}$ durchführbar, und der Aufwand reduziert sich auf $\sim n(p$ opm $+ p(p+1)/2$ opms). Bei der Cholesky-Faktorisierung $\boldsymbol{A} = \boldsymbol{L}\boldsymbol{L}^{\mathsf{T}}$ kommen n opr hinzu.

(ii) Wenn die Dreiecksfaktorisierung nach dem Gaußschen Algorithmus mit Spaltenpivotisierung durchgeführt werden muß, können maximal p weitere obere Nebendiagonalen aufgefüllt werden, d. h., $\boldsymbol{R}$ ist eine Bandmatrix des Typs $\{0, q + p\}$. Die Eliminationskoeffizienten $\hat{l}_{ik}$ können auf dem Platz der zu 0 gemachten Elemente im unteren Teil des Bandes gespeichert werden. Das Verfahren ist daher auf dem Platz einer $\{p, q + p\}$-Bandmatrix ausführbar, und der Aufwand ist maximal $\sim n(p$ opm $+ p(q + p)$ opms).

6.4.2. *Bemerkung.* (i) Bei Rechnung ohne Pivotisierung kann die Dreiecksfaktorisierung sowohl nach den $(n - 1)$-stufigen Grundformen des Gaußschen Algorithmus aus 5.1 oder 6.1 als auch nach den direkten Verfahren aus 6.2 berechnet werden. Bei Rechnung mit Spaltenpivotisierung treten wie beim Vorgehen ohne Pivotisierung für jedes $k \in \{1, \ldots, n-1\}$ höchstens p nichttriviale Eliminationskoeffizienten $\hat{l}_{ik}$ $(i = k + 1, \ldots, \min(k + p, n))$ auf, die auf dem Platz des unteren Bandes gespeichert werden können. Vertauschung der $\hat{l}_{ik}$ gemäß (5.1.18) analog zu den Zeilen der $\boldsymbol{A}^{(j)}$ $(j > k)$ führt auf die Dreiecksmatrix $\boldsymbol{L} = (l_{ik})$ der Faktorisierung

$$\boldsymbol{P}\boldsymbol{A} = \boldsymbol{L}\boldsymbol{R}. \tag{4}$$

Dabei ist $\boldsymbol{L}$ jedoch i. allg. keine $\{p, 0\}$-Bandmatrix, obwohl in jeder Spalte höchstens p nichtverschwindende l_{ik} außerhalb der Diagonalen vorkommen. Die Darstellung (4) mit dem explizit gebildeten Dreiecksfaktor $\boldsymbol{L}$ ist daher für Bandmatrizen im Fall der Spaltenpivotisierung nicht zweckmäßig. Die angepaßte Darstellung — vgl. auch 5.2.6(i) — ist die implizite Dreiecksfaktorisierung

$$\boldsymbol{A} = \hat{\boldsymbol{L}}\boldsymbol{R} \tag{5}$$

gemäß (5.1.16) mit

$$\hat{\boldsymbol{L}} := \boldsymbol{T}_{1,s(1)}\boldsymbol{L}_1(\hat{\boldsymbol{l}}^1) \cdots \boldsymbol{T}_{n-1,s(n-1)}\boldsymbol{L}_{n-1}(\hat{\boldsymbol{l}}^{n-1}) \tag{6}$$

und

$$\hat{\boldsymbol{l}}^k := (0, \ldots, 0, \hat{l}_{k+1,k}, \ldots, \hat{l}_{\min(k+p,n),k}, 0, \ldots, 0)^{\mathsf{T}}.$$

Die Matrix $\hat{\boldsymbol{L}}$ wird durch die Zahlen $\{\hat{l}_{ik}, s(k) : k = 1, \ldots, n-1, i = k + 1, \ldots, \min(k + p, n)\}$ repräsentiert. Für Details der Implementierung sei auf die Spezialliteratur verwiesen, siehe B 6.4.

(ii) Das Gleichungssystem $\boldsymbol{Ax} = \boldsymbol{b}$ kann unter Verwendung der ohne Pivotisierung berechneten Dreiecksfaktorisierung $\boldsymbol{A} = \boldsymbol{LR}$ durch Lösung von $\boldsymbol{Lc} = \boldsymbol{b}$ und $\boldsymbol{Rx} = \boldsymbol{c}$ mit $\sim n(1\,\text{opm} + (p+q)\,\text{opms})$ gelöst werden. Im Fall einer symmetrischen und definiten Matrix werden $\sim n(1\,\text{opm} + 2p\,\text{opms})$ bei Verwendung der $\boldsymbol{LDL}^{\mathrm{T}}$-Faktorisierung und $\sim n(2\,\text{opm} + 2p\,\text{opms})$ bei Verwendung der Cholesky-Faktorisierung benötigt. Wenn die Faktorisierung (5) mit Spaltenpivotisierung berechnet wurde, ist $\boldsymbol{c}$ als Lösung von $\hat{\boldsymbol{L}}\boldsymbol{c} = \boldsymbol{b}$ gemäß

$$\boldsymbol{c} = \boldsymbol{L}_{n-1}(-\hat{\boldsymbol{l}}^{n-1})\,\boldsymbol{T}_{n-1,s(n-1)} \cdots \boldsymbol{L}_1(-\hat{\boldsymbol{l}}^{1})\,\boldsymbol{T}_{1,s(1)}\boldsymbol{b}$$

zu bestimmen, man beachte (6) und 5.2.6(i). Der Aufwand ist maximal $\sim n(1\,\text{opm} + (2p+q)\,\text{opms})$.

(iii) Die beschriebenen Bandalgorithmen erzeugen dieselben Ausgangsdaten wie die entsprechenden Algorithmen für vollbesetztes $\boldsymbol{A}$; sie sind also insbesondere numerisch gutartig, wobei die Fehlerkumulationskonstanten $F = F(n, p, q)$ kleinere Werte als im vollbesetzten Fall haben. Als Algorithmen zur Lösung von $\boldsymbol{Ax} = \boldsymbol{b}$ sind sie fast optimal in dem Sinn, daß der Quotient $\varrho(n) := $ (Anzahl der Rechenoperationen)/(Anzahl der Eingangsdatenelemente) unabhängig von n durch eine Konstante beschränkt ist, vgl. B 2.10. Bei voll besetzter Matrix gilt für den Gaußschen Algorithmus dagegen $\varrho(n) \sim Kn$.

(iv) Bei Diagonal- oder vollständiger Pivotisierung wird die Bandstruktur i. allg. zerstört. Dasselbe trifft für die bei der BKP-Faktorisierung indefiniter symmetrischer Matrizen verwendete Pivotisierungsstrategie zu, so daß aus der Symmetrie einer Bandmatrix bei fehlender Definitheit i. allg. kein Vorteil gezogen werden kann.

(v) Im Band auftretende Nullen werden bei den diskutierten Bandtechniken nicht berücksichtigt. Falls jedoch eine sog. *Blockbandstruktur* mit Blöcken $\boldsymbol{A}_{ij}$ von Bandgestalt vorliegt, falls $\boldsymbol{A}$ also etwa vom Typ

$$\boldsymbol{A} = \begin{pmatrix} \boldsymbol{A}_{11} & \boldsymbol{A}_{12} & & & \\ \boldsymbol{A}_{21} & \boldsymbol{A}_{22} & \boldsymbol{A}_{23} & & \\ & \boldsymbol{A}_{32} & \boldsymbol{A}_{33} & \boldsymbol{A}_{34} & \\ & \ddots & \ddots & \ddots & \end{pmatrix}$$

ist, kann analog zum Vorgehen in 6.1.B zur *Blockelimination* übergegangen werden. Die Inversen der Pivotblöcke sollen dabei nicht explizit gebildet, sondern durch die zugehörigen Faktorisierungen repräsentiert werden. □

Den einfachsten Sonderfall einer Bandmatrix stellt eine *Tridiagonalmatrix*, d. h. eine $\{1, 1\}$-Bandmatrix

$$\boldsymbol{A} = \begin{pmatrix} d_1 & e_1 & & & & \\ c_2 & d_2 & e_2 & & & \\ & c_3 & d_3 & e_3 & & \\ & & \ddots & \ddots & \ddots & \\ & & & c_{n-1} & d_{n-1} & e_{n-1} \\ & & & & c_n & d_n \end{pmatrix} \tag{7}$$

dar, die zweckmäßigerweise durch die drei Vektoren $\boldsymbol{c} = (c_i)$, $\boldsymbol{d} = (d_i)$, $\boldsymbol{e} = (e_i) \in \mathbf{R}^n$ dargestellt wird. Der erste Schritt des Gaußschen Algorithmus führt je nach Wahl von d_1 bzw. c_2 als Pivot auf

$$\boldsymbol{A}^{(2)} = \begin{pmatrix} d_1 & e_1 & & \\ 0 & d_2 - l_{21}e_1 & e_2 & \\ & c_3 & d_3 & e_3 \\ & \ddots & \ddots & \ddots \end{pmatrix} \quad \text{bzw.} \quad \boldsymbol{A}^{(2)} = \begin{pmatrix} c_2 & d_2 & e_2 & \\ 0 & e_1 - \hat{l}_{21}d_2 & -\hat{l}_{21}e_2 & \\ & c_3 & d_3 & e_3 \\ & \ddots & \ddots & \ddots \end{pmatrix};$$

der Eliminationskoeffizient ist

$$l_{21} := c_2/d_1 \quad \text{bzw.} \quad \hat{l}_{21} := d_1/c_2.$$

Die entstehende $(n-1)$-dimensionale Restmatrix $\boldsymbol{M}^{(2)}$ ist wieder tridiagonal, so daß analog zum ersten Schritt fortgefahren werden kann. Bei Rechnung ohne Pivotisierung ergibt sich das folgende Verfahren.

6.4.3. *Dreiecksfaktorisierung einer Tridiagonalmatrix $\boldsymbol{A}$ ohne Pivotisierung und Lösung von $\boldsymbol{Ax} = \boldsymbol{b}$.*

Aufgabe: Für die durch (7) gegebene reguläre Tridiagonalmatrix $\boldsymbol{A} \in \mathbf{R}^{n,n}$ ist die Faktorisierung $\boldsymbol{A} = \boldsymbol{LR}$ nach dem als durchführbar vorausgesetzte Gaußschen Algorithmus zu berechnen; die Matrix $\boldsymbol{A}$ ist mit den signifikanten Elementen von $\boldsymbol{L}$ und $\boldsymbol{R}$ zu überspeichern. Unter Verwendung von $\boldsymbol{L}, \boldsymbol{R}$ ist $\boldsymbol{x}$ als Lösung von $\boldsymbol{Ax} = \boldsymbol{b}$ zu berechnen und auf dem Platz von $\boldsymbol{b}$ zu speichern.

Algorithmus:

S1 (Dreiecksfaktorisierung $\boldsymbol{A} = \boldsymbol{LR}$):
for $k := 2(1)n$ **do** $[c_k := c_k/d_{k-1},\ d_k := d_k - c_k * e_{k-1}]$

S2 (Lösung von $\boldsymbol{LRx} = \boldsymbol{b}$):

S2.1 ($\boldsymbol{b} := \boldsymbol{L}^{-1}\boldsymbol{b}$): **for** $k := 2(1)n$ **do** $b_k := b_k - c_k * b_{k-1}$

S2.2 ($\boldsymbol{b} := \boldsymbol{R}^{-1}\boldsymbol{b}$): $b_n := b_n/d_n$
for $k := n - 1(-1)1$ **do** $b_k := (b_k - e_k * b_{k+1})/d_k$

Aufwand: $\sim n$(1 opm + 1 opms) für $\boldsymbol{L}, \boldsymbol{R}$, $\sim n$(1 opm + 2 opms) für $\boldsymbol{x}$

6.4.4. *Bemerkung.* (i) Wenn $\boldsymbol{A}$ symmetrisch und positiv definit ist, gilt $e_{k-1} = c_k$, und die Laufanweisung in S1 ist durch

$$\textbf{for } k := 2(1)n \textbf{ do } [z := \mp c_k,\ c_k := c_k/d_{k-1},\ d_k := d_k - c_k * z]$$

zu ersetzen. Damit ergibt sich die Faktorisierung $\boldsymbol{A} = \boldsymbol{LDL}^{\mathrm{T}}$.

(ii) Bei Rechnung mit Spaltenpivotisierung werden zusätzlich $\sim n$ S für die zweite Nebendiagonale von $\boldsymbol{R}$ benötigt, außerdem sind die Indizes $\{s(1), \ldots, s(n-1)\}$ zu speichern. Wenn nur ein einzelnes System $\boldsymbol{Ax} = \boldsymbol{b}$ zu lösen ist und S1 und S2.1 wie in 5.1.1 gemeinsam in einer Laufanweisung realisiert werden, brauchen die $\hat{l}_{ik}$ und $s(k)$ nicht aufgehoben zu werden. In diesem Fall kann die Lösung auf dem Platz der rechten Seite $\boldsymbol{b}$ und der drei Diagonalen von $\boldsymbol{A}$ berechnet werden, indem die neu entstehende zweite obere Nebendiagonale auf dem Platz von $\boldsymbol{c}$ gespeichert wird. □

B. *Schwach besetzte Matrizen beliebiger Struktur*

Algorithmen zur Dreiecksfaktorisierung schwach besetzter Matrizen beliebiger Struktur gehen davon aus, daß die NNE beliebig über die Matrix verteilt sein können; evtl. vorhandene regelmäßige Strukturen werden ignoriert.

In manchen Fällen ist es möglich, die Matrix $\boldsymbol{A}$ durch Zeilen- und Spaltenvertauschungen

$$\bar{\boldsymbol{A}} = \boldsymbol{P}_Z \boldsymbol{A} \boldsymbol{P}_S^{\mathsf{T}} \tag{8}$$

für nichtsymmetrisches $\boldsymbol{A}$ bzw.

$$\bar{\boldsymbol{A}} = \boldsymbol{P}\boldsymbol{A}\boldsymbol{P}^{\mathsf{T}} \tag{9}$$

für symmetrisches und definites $\boldsymbol{A}$ so umzuordnen, daß $\bar{\boldsymbol{A}}$ von Bandgestalt mit minimaler Bandbreite ist. Die zugehörigen Algorithmen zur Bestimmung der Permutationsmatrizen $\boldsymbol{P}_Z$, $\boldsymbol{P}_S$ bzw. $\boldsymbol{P}$ sind graphentheoretischer Natur und können im Rahmen einer Einführung nicht diskutiert werden, siehe B 6.5 für Hinweise.

Für symmetrisches definites $\boldsymbol{A}$ ist durch Übergang von Band- zu Einhüllenden-Techniken eine weitere Senkung des Speicherbedarfs und Rechenaufwandes möglich. Die *Einhüllende* der Matrix $\boldsymbol{A}$ besteht dabei aus der Diagonalen und allen Positionen (i, j), für die ein l mit $a_{il} \neq 0$ $(1 \leqq l \leqq j < i)$ oder $a_{lj} \neq 0$ $(1 \leqq l \leqq i < j)$ existiert. Im nachfolgenden Beispiel sind die zur Einhüllenden gehörenden Positionen umrandet worden.

$$\boldsymbol{A} = \begin{pmatrix} \times & \times & & & & & \\ \times & \times & \times & & \times & & \times \\ & \times & \times & \times & & & \\ & & \times & \times & \times & & \\ & \times & & \times & \times & \times & \\ & & & & \times & \times & \times \\ & & \times & & & \times & \times \end{pmatrix}.$$

Aus der zeilenweisen Version 6.2.6 des Cholesky-Verfahrens liest man sofort das folgende Resultat ab:

6.4.5. *Aussage:* Die Matrix $\boldsymbol{A} \in \mathbf{S}^{n,n}$ sei positiv definit. Dann ist die Einhüllende von $\boldsymbol{A}$ gleich der Einhüllenden von $(\boldsymbol{L} + \boldsymbol{L}^{\mathsf{T}})$, wobei $\boldsymbol{A} = \boldsymbol{L}\boldsymbol{L}^{\mathsf{T}}$ die Cholesky-Faktorisierung von $\boldsymbol{A}$ bezeichnet.

Dies besagt: *Der Cholesky-Faktor $\boldsymbol{L}$ kann auf dem unteren Teil der Einhüllenden von $\boldsymbol{A}$ berechnet werden.* Wenn die Anzahl der zur Einhüllenden von $\boldsymbol{A}$ gehörigen Plätze *Profil von A* — in Zeichen: $p(\boldsymbol{A})$ — genannt wird, ist $\big(p(\boldsymbol{A}) + n\big)/2$ gerade die Anzahl der für die Faktorisierung benötigten Speicherplätze. Es ist daher sinnvoll, durch geeignete simultane Zeilen- und Spaltenvertauschungen gemäß (9) das Profil von $\boldsymbol{A}$ zu minimieren, siehe B 6.5 für Hinweise.

Wenn Band- oder Einhüllenden-Techniken nicht angewendet werden können oder nicht zum Erfolg führen, muß zu universellen Verfahren für schwach besetzte

Matrizen übergegangen werden. Grundlage ist wieder der Gaußsche Algorithmus, von dem bereits $k-1$ Schritte durchgeführt sein mögen. Die zu Beginn des k-ten Schrittes vorliegende Restmatrix ist dann

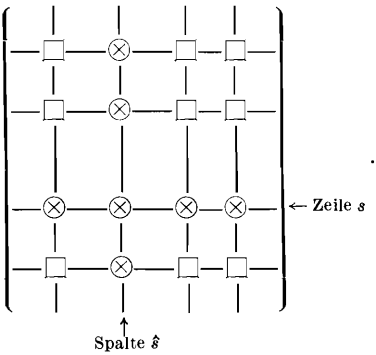
$$\boldsymbol{M}^{(k)} = \begin{bmatrix} a_{kk}^{(k)} & \dots & a_{k\hat{s}}^{(k)} & \dots & a_{kn}^{(k)} \\ \vdots & & \vdots & & \vdots \\ a_{sk}^{(k)} & \dots & a_{s\hat{s}}^{(k)} & \dots & a_{sn}^{(k)} \\ \vdots & & \vdots & & \vdots \\ a_{nk}^{(k)} & \dots & a_{n\hat{s}}^{(k)} & \dots & a_{nn}^{(k)} \end{bmatrix}.$$

Wenn das Element $a_{s\hat{s}}^{(k)} \neq 0$ als Pivot gewählt wird, gehen die $a_{ij}^{(k)}$ in

$$a_{ij}^{(k+1)} := a_{ij}^{(k)} - \frac{a_{i\hat{s}}^{(k)} * a_{sj}^{(k)}}{a_{s\hat{s}}^{(k)}} \quad (i, j = k, \dots, n;\ i \neq s,\ j \neq \hat{s}) \tag{10}$$

über, sofern auf die physische Vertauschung der Zeilen k, s und Spalten $k, \hat{s}$ verzichtet wird. Ein Nullelement $a_{ij}^{(k)}$ kann also nur dann in ein Nichtnullelement $a_{ij}^{(k+1)}$ übergehen, wenn sowohl das zugehörige Pivotspaltenelement $a_{i\hat{s}}^{(k)}$ als auch das Pivotzeilenelement $a_{sj}^{(k)}$ von 0 verschieden sind. Für schwach besetztes $\boldsymbol{A}$ wird das nur selten der Fall sein, so daß die meisten Nullen aus $\boldsymbol{M}^{(k)}$ nach $\boldsymbol{M}^{(k+1)}$ übernommen werden. Zur Illustration möge das folgende Beispiel dienen, bei dem die Positionen, in denen NNE entstehen können, eingerahmt worden sind:

$\boldsymbol{M}^{(k)} =$



Die Größe

$$\eta(\boldsymbol{M}^{(k+1)}) - \eta(\boldsymbol{M}^{(k)}) \tag{11}$$

heißt *lokale Auffüllung* im k-ten Schritt (engl. „local fill in"). Wenn $\sigma_i^{(k)}$ bzw. $\hat{\sigma}_j^{(k)}$ die Anzahl der NNE in der Zeile $(a_{ik}^{(k)}, \dots, a_{in}^{(k)})$ bzw. Spalte $(a_{kj}^{(k)}, \dots, a_{nj}^{(k)})^T$ von $\boldsymbol{M}^{(k)}$ bezeichnet, kann die lokale Auffüllung wie folgt abgeschätzt werden:

6.4.6. *Aussage.* Bei der Wahl von $a_{s\hat{s}}^{(k)} \neq 0$ $(k \leqq s, \hat{s} \leqq n)$ als Pivot gilt

$$\eta(\boldsymbol{M}^{(k+1)}) - \eta(\boldsymbol{M}^{(k)}) \leqq (\sigma_s^{(k)} - 1) * (\hat{\sigma}_{\hat{s}}^{(k)} - 1) =: \mu_{s\hat{s}}^{(k)}.$$

Die Zahlen $\mu_{s\hat{s}}^{(k)}$ heißen *Markowitz-Kosten* der Elemente $a_{s\hat{s}}^{(k)} \neq 0$. Man beachte, daß durch (11) die gesamte im k-ten Schritt auftretende Auffüllung erfaßt wird,

denn bei der Berechnung der $\hat{l}_{ik}$ und der Übernahme der Pivotzeile als k-te Zeile von $\boldsymbol{R}$ entstehen keine NNE.

Die bisherigen Überlegungen zeigen, daß ein universelles Verfahren zur Dreiecksfaktorisierung schwach besetzter Matrizen zwei Forderungen erfüllen muß:

— Auswahl einer *Pivotreihenfolge* $\{s(k), \hat{s}(k)\}$ $(k = 1, \ldots, n-1)$ bzw. äquivalent Festlegung von Permutationsmatrizen $\boldsymbol{P}_Z$ und $\boldsymbol{P}_S$ derart, daß die Dreiecksfaktorisierung von $\boldsymbol{A}$ stabil durchführbar ist und die durch

$$\eta(\boldsymbol{L} + \boldsymbol{R}) - \eta(\boldsymbol{A})$$

gegebene Auffüllung während des Eliminationsprozesses möglichst klein ist.

— Auswahl einer *Kompaktspeichertechnik*, bei welcher nur die NNE von $\boldsymbol{A}$ und die im Laufe der Elimination durch Auffüllung entstehenden NNE einschließlich zugehöriger Indexinformationen zum Auffinden bzw. Einfügen von Elementen gespeichert werden.

Die in der ersten Forderung genannte stabile Durchführbarkeit spielt je nach Eigenschaften der Matrix $\boldsymbol{A}$ eine unterschiedliche Rolle: Für symmetrisches und positiv definites $\boldsymbol{A}$ werden nur symmetrische Zeilen- und Spaltenvertauschungen zugelassen, d. h., die Pivots müssen auf der Diagonalen von $\boldsymbol{M}^{(k)}$ gewählt werden, und es gilt $s(k) = \hat{s}(k)$ bzw. $\boldsymbol{P}_Z = \boldsymbol{P}_S = \boldsymbol{P}$, vgl. (9). Aus 5.2 und 6.1 ist bekannt, daß die Cholesky-Faktorisierung für jede solche Pivotreihenfolge ein numerisch gutartiger Prozeß ist unabhängig von den konkreten Zahlenwerten der NNE. Die Pivotreihenfolge kann also allein auf Grund des Besetztheitsmusters von $\boldsymbol{A}$ festgelegt werden. Demgemäß bestehen Implementierungen für symmetrische und definite schwach besetzte Systeme in der Regel aus drei Teilprogrammen:

P1: Festlegung der Pivotreihenfolge auf Grund des Besetztheitsmusters von $\boldsymbol{A}$

P2: Numerische Berechnung des Cholesky-Faktors unter Verwendung der in P1 ermittelten Pivotreihenfolge

P3: Lösung von $\boldsymbol{Ax} = \boldsymbol{b}$ unter Verwendung des in P2 berechneten Cholesky-Faktors von $\boldsymbol{A}$.

Der Zeitaufwand für die in P1 auszuführenden nichtnumerischen Operationen ist i. allg. wesentlich größer als der für P2. Da in der Praxis meist mehrere, unter Umständen sogar sehr viele Systeme derselben Besetztheitsstruktur, aber mit unterschiedlichen Zahlenwerten zu lösen sind, braucht der aufwendige erste Schritt P1 nur einmal ausgeführt zu werden.

Im nichtsymmetrischen Fall hängt das Stabilitätsverhalten im Unterschied zum symmetrischen und definiten Fall auch von den Zahlenwerten der a_{ij} ab, so daß die Pivotreihenfolge nicht ausschließlich auf Grund des Besetztheitsmusters festgelegt werden kann. Wenn das Pivot in bezug auf die restlichen NNE der Pivotspalte relativ klein ist, werden die l_{ik} und i. allg. auch die $a_{ij}^{(k+1)}$ sehr groß, und die Gutartigkeit bzw. Stabilität geht verloren, vgl. 5.2 und 5.3. Die Freiheit in der Wahl des Pivots $a_{s\hat{s}}^{(k)}$ muß also auf Kosten einer möglicherweise größeren Auffüllung eingeschränkt werden, etwa durch die Bedingung

$$|a_{s\hat{s}}^{(k)}| \geqq \alpha \max \{|a_{i\hat{s}}^{(k)}| : k \leqq i \leqq n\} \quad \text{mit} \quad 0 < \alpha \leqq 1. \tag{12}$$

Aus (12) folgt

$$|\hat{l}_{ik}| \leqq 1/\alpha,$$

d. h., das Elementwachstum ist gemäß

$$\max\{|a_{ij}^{(k+1)}|: k+1 \leqq i, j \leqq n\} \leqq (1 + 1/\alpha) \max\{|a_{ij}^{(k)}|: k \leqq i, j \leqq n\} \quad (13)$$

beschränkt. Für $\alpha = 1$ liegt gerade der Fall der Spaltenpivotisierung in der Spalte $\hat{s}$ vor, allerdings kann der Spaltenindex $\hat{s} \in \{k, \ldots, n\}$ noch frei gewählt werden. Durch Vorgabe von α kann der Nutzer einen Kompromiß zwischen den beiden gegensätzlichen Forderungen „numerische Gutartigkeit" (α nahe bei 1) und „geringe Auffüllung" (α nahe bei 0) zu erzielen versuchen.

Bei Verwendung einer einmal bestimmten Pivotreihenfolge für eine weitere Matrix desselben Besetztheitsmusters ist die Bedingung (12) nicht notwendig erfüllt. Es macht sich dann erforderlich, das Wachstum der $a_{ij}^{(k)}$ im Laufe der Elimination zu kontrollieren und bei zu starkem Anwachsen gegebenenfalls eine neue Pivotreihenfolge festzulegen. Eine weitere Möglichkeit besteht darin, den Genauigkeitsverlust bei der Faktorisierung durch einige Schritte der iterativen Verbesserung entsprechend 5.4 auszugleichen.

Wir bemerken abschließend, daß für schwach besetzte Gleichungssysteme, die bei der Diskretisierung von partiellen Differentialgleichungen etwa nach der Methode der finiten Elemente bzw. nach dem Differenzenverfahren entstehen, eine Reihe von speziellen Lösungsverfahren bekannt ist. Diese Verfahren nutzen die spezielle Gestalt der Koeffizientenmatrix zielgerichtet aus und sind daher für gewisse Problemklassen günstiger als universelle Lösungsverfahren, siehe B 6.6.

Übungsaufgaben

Ü 6.4.1. Man schreibe ein zu 6.4.3 analoges Programm für den Fall einer symmetrischen und positiv definiten Koeffizientenmatrix unter Verwendung der Faktorisierung $\boldsymbol{A} = \boldsymbol{LDL}^T$ bzw. $\boldsymbol{A} = \boldsymbol{LL}^T$.

Ü 6.4.2. Man gebe ein zu 6.4.3 analoges Programm zur Faktorisierung einer nichtsymmetrischen Tridiagonalmatrix mit Spaltenpivotisierung und zur Lösung von $\boldsymbol{Ax} = \boldsymbol{b}$ an.

6.5. Berechnung und Aufdatierung der inversen Matrix

In den Anwendungen wird die Inverse $\boldsymbol{A}^{-1}$ einer regulären Matrix $\boldsymbol{A} \in \mathbf{R}^{n,n}$ in den seltensten Fällen explizit benötigt, vgl. die Ausführungen am Ende dieses Abschnitts. Wenn die Berechnung von $\boldsymbol{A}^{-1}$ doch erforderlich ist, gehen wir wie bisher von der Dreiecksfaktorisierung

$$\boldsymbol{PA} = \boldsymbol{LR} \quad (1)$$

von $\boldsymbol{A}$ aus, die nach einer der Varianten des Gaußschen Algorithmus berechnet sein möge. Aus (1) folgt

$$\boldsymbol{A}^{-1} = \boldsymbol{R}^{-1}\boldsymbol{L}^{-1}\boldsymbol{P}. \quad (2)$$

Zur Auswertung der Darstellung (2) bestimmen wir in einem ersten Schritt die Inverse $\boldsymbol{S} := \boldsymbol{R}^{-1}$ aus der Matrixgleichung

$$\boldsymbol{SR} = \boldsymbol{I} \tag{3}$$

analog zum Vorgehen beim verketteten Gaußschen Algorithmus in 6.2. Da $\boldsymbol{S}$ wie $\boldsymbol{R}$ eine obere Dreiecksmatrix ist, kann (3) elementweise in der Form

$$\sum_{p=i}^{j} s_{ip} r_{pj} = \begin{cases} 1 & \text{für } i = j, \\ 0 & \text{für } i \neq j \end{cases} \quad (i, j = 1, \ldots, n),$$

geschrieben werden. Für $i > j$ ist dies trivial erfüllt, da die linksstehende Summe leer ist. Für $i = j$ ergibt sich $s_{jj} r_{jj} = 1$, also

$$s_{jj} := 1/r_{jj} \qquad (j = 1, \ldots, n). \tag{4}$$

Für $i < j$ folgt $\sum_{p=i}^{j-1} s_{ip} r_{pj} + s_{ij} r_{jj} = 0$, d. h. unter Beachtung von (4)

$$s_{ij} := -\left(\sum_{p=i}^{j-1} s_{ip} r_{pj} \right) s_{jj}. \tag{5}$$

Zur Bestimmung von s_{ij} werden die links von $\{i, j\}$ stehenden Elemente von $\boldsymbol{S}$ und die unterhalb von $\{i, j\}$ stehenden Elemente von $\boldsymbol{R}$ benötigt. Eine Berechnung auf dem Platz von $\boldsymbol{R}$ ist daher möglich, wenn $\boldsymbol{S}$ spaltenweise von links nach rechts und in den Spalten von oben nach unten bestimmt wird.

Im zweiten Schritt wird $\boldsymbol{T} := \boldsymbol{SL}^{-1}$ gebildet. Aus der äquivalenten Gleichung

$$\boldsymbol{TL} = \boldsymbol{S} \tag{6}$$

ergibt sich unter Beachtung der Dreiecksform von $\boldsymbol{L}$

$$\sum_{p=j}^{n} t_{ip} l_{pj} = s_{ij},$$

also wegen $l_{jj} = 1$

$$t_{ij} := s_{ij} - \sum_{p=j+1}^{n} t_{ip} l_{pj} \qquad (i, j = 1, \ldots, n); \tag{7}$$

man beachte $s_{ij} = 0$ für $i > j$. Auf der rechten Seite treten die rechts von der Position (i, j) stehenden Elemente von $\boldsymbol{T}$ und die Eliminationskoeffizienten der j-ten Spalte auf. Eine spaltenweise in-situ-Realisierung ist daher möglich, wenn die l_{pj} $(p = j + 1, \ldots, n)$ in einem Hilfsfeld der Länge n ausgelagert werden, damit die entsprechenden t_{pj} in die Positionen (p, j) gebracht werden können.

Im dritten Schritt entsteht $\boldsymbol{A}^{-1} = \boldsymbol{TP}$ durch Spaltenvertauschung aus $\boldsymbol{T}$.

Zusammenfassend erhalten wir den folgenden Inversionsalgorithmus:

6.5.1. *Berechnung von $\boldsymbol{A}^{-1}$ unter Verwendung der $\boldsymbol{LR}$-Faktorisierung von $\boldsymbol{A}$.*

Aufgabe: Für gegebene und gemäß 5.2.3 auf dem Platz von $\boldsymbol{A}$ gespeicherte Dreiecksfaktorisierung $\boldsymbol{PA} = \boldsymbol{LR}$ mit $\boldsymbol{L}, \boldsymbol{R} \in \mathfrak{R}^{n,n}$, $\boldsymbol{R}$ regulär, ist die Inverse $\boldsymbol{A}^{-1}$ zu berechnen und auf dem Platz von $\boldsymbol{A}$ zu speichern.

Algorithmus:

S1 (Berechnung von $\boldsymbol{S} = \boldsymbol{R}^{-1}$):

for $j := 1(1)n$ **do**

$a_{jj} := 1/a_{jj}$

for $i := 1(1)j-1$ **do** $a_{ij} := -\left(\sum_{p=i}^{j-1} a_{ip} * a_{pj}\right) * a_{jj}$

S2 (Berechnung von $\boldsymbol{T} = \boldsymbol{R}^{-1}\boldsymbol{L}^{-1}$):

for $j := n-1(-1)1$ **do**

for $i := j+1(1)n$ **do** $[l_i := a_{ij},\ a_{ij} := 0]$

for $i := 1(1)n$ **do** $a_{ij} := a_{ij} - \sum_{p=j+1}^{n} a_{ip} * l_p$

S3 (Berechnung von $\boldsymbol{A}^{-1} = \boldsymbol{TP}$):

for $j := n-1(-1)1$ **do**

$s = s(j)$

if $j < s$ **then**

for $i := 1(1)n$ **do** $[z := a_{ij},\ a_{ij} := a_{is},\ a_{is} := z]$

Aufwand: S1: $\sim n^3/6$ opms, S2: $\sim n^3/2$ opms, S1 + S2: $\sim 2n^3/3$ opms, $\sim n$ S (für $l_2, \ldots, l_n$)

6.5.2. *Bemerkung.* (i) Wenn die $\sim n^3/3$ opms für die $\boldsymbol{LR}$-Faktorisierung mit gezählt werden, erfordert die Matrixinversion nach 6.5.1 insgesamt $\sim n^3$ opms. Eine n- bzw. $(n-1)$-stufige Version von S1 bzw. S2 ist möglich, siehe Ü 6.5.1.

(ii) Für symmetrisches $\boldsymbol{A}$ kann von den billigeren symmetrischen Faktorisierungen aus 6.1 ausgegangen werden. Im positiv definiten Fall würde sich etwa mit der Cholesky-Faktorisierung $\boldsymbol{A} = \boldsymbol{LL}^{\mathsf{T}}$ das folgende Vorgehen empfehlen:

S1: Berechne $\boldsymbol{S} := (\boldsymbol{L}^{\mathsf{T}})^{-1}$ mit $\sim n^3/6$ opms

S2: Bilde $\boldsymbol{A}^{-1} := \boldsymbol{SS}^{\mathsf{T}}$ mit $\sim n^3/6$ opms

Der Aufwand reduziert sich dabei auf die Hälfte. Für indefinites $\boldsymbol{A}$ kann unter Verwendung der BKP-Faktorisierung analog vorgegangen werden. Man beachte, daß $\boldsymbol{A}^{-1}$ symmetrisch ist, also nur ein Dreieck berechnet zu werden braucht.

(iii) Die Rundungsfehleranalyse von 6.5.1 unter Einbeziehung der $\boldsymbol{LR}$-Faktorisierung zeigt, daß das Verfahren für nicht zu schlecht konditioniertes $\boldsymbol{A} \in \mathfrak{R}^{n,n}$ durchführbar ist. Für die berechnete Inverse $\boldsymbol{X} \approx \boldsymbol{A}^{-1}$ gilt dabei: Zu jedem $i \in \{1, \ldots, n\}$ gibt es eine Störung $\delta\boldsymbol{A}^{(i)}$, so daß die i-te Zeile $\boldsymbol{e}^{i\mathsf{T}}\boldsymbol{X}$ von $\boldsymbol{X}$ der Beziehung

$$\boldsymbol{e}^{i\mathsf{T}}\boldsymbol{X}(\boldsymbol{A} + \delta\boldsymbol{A}^{(i)}) = \boldsymbol{e}^{i\mathsf{T}} \quad \text{mit} \quad \|\delta\boldsymbol{A}^{(i)}\|_2 \leqq \nu F\, \|\boldsymbol{A}\|_2 \tag{8}$$

genügt, wobei $F = F(n)$ eine schwach mit n wachsende, von der Matrizenklasse und dem Faktorisierungsverfahren abhängende Kumulationskonstante ist. *Das Verfahren* 6.5.1 *ist also numerisch gutartig in bezug auf jede Zeile von* $\boldsymbol{X}$. Dagegen gibt es i. allg. keine kleine Störung $\delta\boldsymbol{A}$, so daß $\boldsymbol{X} = (\boldsymbol{A} + \delta\boldsymbol{A})^{-1}$ gilt, d. h., Gutartigkeit

bezüglich $\boldsymbol{X}$ liegt nicht vor. Aus (8) folgt

$$\|\boldsymbol{XA} - \boldsymbol{I}\|_F \leqq vF \|\boldsymbol{A}\|_2 \|\boldsymbol{X}\|_F \leqq v\sqrt{n}\, F \|\boldsymbol{A}\|_2 \|\boldsymbol{X}\|_2 \doteqdot v\hat{F} \operatorname{cond}_2(\boldsymbol{A}) \tag{9}$$

mit $\hat{F} \approx \sqrt{n}\, F$, d. h., *das Residuum* $\boldsymbol{XA} - \boldsymbol{I}$ *ist klein, und* $\boldsymbol{X}$ *ist eine gute Linksinverse*. Dagegen muß $\boldsymbol{X}$ keine gute Rechtsinverse sein, obwohl dies häufig wegen spezieller Fehlerkorrelationen der Fall ist: Es kann nur

$$\|\boldsymbol{AX} - \boldsymbol{I}\|_F = \|\boldsymbol{A}(\boldsymbol{XA} - \boldsymbol{I})\boldsymbol{A}^{-1}\|_F \leqq \|\boldsymbol{XA} - \boldsymbol{I}\|_F \operatorname{cond}_2(\boldsymbol{A}) \doteqdot v\hat{F}[\operatorname{cond}_2(\boldsymbol{A})]^2 \tag{10}$$

gezeigt werden. Wegen $\boldsymbol{X} - \boldsymbol{A}^{-1} = (\boldsymbol{XA} - \boldsymbol{I})\boldsymbol{A}^{-1}$ folgt schließlich aus (9)

$$\frac{\|\boldsymbol{X} - \boldsymbol{A}^{-1}\|_F}{\|\boldsymbol{A}^{-1}\|_2} \leqq vF \|\boldsymbol{A}\|_2 \|\boldsymbol{X}\|_F \doteqdot v\hat{F} \operatorname{cond}_2(\boldsymbol{A}), \tag{11}$$

also die *numerische Stabilität* von 6.5.1. □

Als alternative Möglichkeit soll ein nach GAUSS und JORDAN benanntes Vorgehen beschrieben werden; hinsichtlich der Namensgebung siehe B 6.6. Dabei wird die Matrix $\boldsymbol{A}$ nach der Vorschrift

$$\boldsymbol{A}^{(1)} := \boldsymbol{A}, \qquad \boldsymbol{A}^{(k+1)} := \boldsymbol{D}_k \boldsymbol{M}_k \boldsymbol{A}^{(k)} \qquad (k = 1, \ldots, n) \tag{12}$$

sukzessive mittels NT-Matrizen $\boldsymbol{M}_k$ und Diagonalmatrizen $\boldsymbol{D}_k$ der Gestalt

$$\boldsymbol{M}_k = \begin{pmatrix} 1 & & & -m_{1k} & & \\ & \ddots & & \vdots & & \\ & & 1 & -m_{k-1,k} & & \\ & & & 1 & & \\ & & & -m_{k+1,k} & 1 & \\ & & & \vdots & & \ddots \\ & & & -m_{nk} & & 1 \end{pmatrix} \quad \text{bzw.} \quad \boldsymbol{D}_k = \begin{pmatrix} 1 & & & & \\ & \ddots & & & \\ & & 1 & & \\ & & & d_k & \\ & & & & 1 \\ & & & & & \ddots \\ & & & & & & 1 \end{pmatrix} \tag{13}$$

äquivalent in die Einheitsmatrix $\boldsymbol{I}$ transformiert. Die Matrix $\boldsymbol{A}^{(k)}$ stimmt in den ersten $k-1$ Spalten bereits mit $\boldsymbol{I}$ überein, d. h., es ist

$$\boldsymbol{A}^{(k)} = \left(\begin{array}{ccc|ccc} 1 & & & a_{1k}^{(k)} & \cdots & a_{1n}^{(k)} \\ & \ddots & & \vdots & & \vdots \\ & & 1 & a_{k-1,k-1}^{(k)} & \cdots & a_{k-1,n}^{(k)} \\ \hline & & & a_{kk}^{(k)} & \cdots & a_{kn}^{(k)} \\ & & & \vdots & & \vdots \\ & & & a_{nk}^{(k)} & \cdots & a_{nn}^{(k)} \end{array}\right) = \left(\begin{array}{c|c} \boldsymbol{I}_{k-1} & \boldsymbol{N}^{(k)} \\ \hline \boldsymbol{O} & \boldsymbol{M}^{(k)} \end{array}\right). \tag{14}$$

Die Koeffizienten m_{ik} werden so bestimmt, daß in der k-ten Spalte von $\boldsymbol{M}_k\boldsymbol{A}^{(k)}$ außerhalb des Diagonalelementes Nullen entstehen. Dies führt auf

$$m_{ik} := a_{ik}^{(k)}/a_{kk}^{(k)} \quad (i = 1, \ldots, n,\ i \neq k). \tag{15}$$

Dabei muß das Pivotelement $a_{kk}^{(k)}$ natürlich von 0 verschieden sein, was der Einfachheit halber im folgenden vorausgesetzt werden soll. Mit der Festlegung

$$d_k := 1/a_{kk}^{(k)} \tag{16}$$

wird schließlich das Diagonalelement $a_{kk}^{(k)}$ von $\boldsymbol{M}_k\boldsymbol{A}^{(k)}$ zu 1 gemacht. Nach n Schritten ergibt sich dann $\boldsymbol{A}^{(n+1)} = \boldsymbol{D}_n\boldsymbol{M}_n \cdots \boldsymbol{D}_1\boldsymbol{M}_1\boldsymbol{A}$. Da $\boldsymbol{A}^{(n+1)}$ nach Konstruktion gleich $\boldsymbol{I}$ ist, folgt

$$\boldsymbol{A}^{-1} = \boldsymbol{D}_n\boldsymbol{M}_n \cdots \boldsymbol{D}_1\boldsymbol{M}_1 . \tag{17}$$

Die Darstellung (17) wird *Produktform der Inversen* genannt. Die signifikanten Koeffizienten m_{ik} bzw. d_k der Faktoren können auf dem Platz von a_{ik} bzw. a_{kk} gespeichert werden. Eine eventuell vorhandene Bandgestalt oder andersartige schwache Besetztheit schlägt sich in der Matrix der $\{m_{ik}, d_k\}$ nieder, während das explizit gebildete $\boldsymbol{A}^{-1}$ i. allg. voll besetzt ist.

Falls ein Ausdruck der Form $\boldsymbol{X} = \boldsymbol{A}^{-1}\boldsymbol{B}$ zu berechnen ist, kann dies durch sukzessive Multiplikation mit den Faktoren geschehen. Speziell ergibt sich für $\boldsymbol{B} = \boldsymbol{b} \in \mathbf{R}^n$ gemäß

$$\boldsymbol{x} = \boldsymbol{A}^{-1}\boldsymbol{b} = \boldsymbol{D}_n\boldsymbol{M}_n \cdots \boldsymbol{D}_1\boldsymbol{M}_1\boldsymbol{b} \tag{18}$$

das sog. *Gauß-Jordan-Verfahren* zur Lösung des linearen Gleichungssystems $\boldsymbol{A}\boldsymbol{x} = \boldsymbol{b}$. Die rechte Seite $\boldsymbol{b}$ wird dabei wie eine zusätzliche Spalte von $\boldsymbol{A}$ transformiert. Der Gesamtaufwand ist $\sim n^3/2$ opms, also größer als beim Gaußschen Algorithmus.

Für den Sonderfall $\boldsymbol{B} = \boldsymbol{I}$ wird $\boldsymbol{X} = \boldsymbol{A}^{-1}$, und die zugehörige Transformation ist dann

$$\boldsymbol{X}^{(1)} := \boldsymbol{I}, \qquad \boldsymbol{X}^{(k+1)} := \boldsymbol{D}_k\boldsymbol{M}_k\boldsymbol{X}^{(k)} \qquad (k = 1, \ldots, n). \tag{19}$$

Die $\boldsymbol{X}^{(k)}$ werden nach derselben Vorschrift wie die $\boldsymbol{A}^{(k)}$ transformiert, lediglich die Startmatrizen sind verschieden, vgl. (12). Die Matrix $\boldsymbol{X}^{(k)}$ unterscheidet sich nur in den ersten $k-1$ Spalten von $\boldsymbol{I}$. Im k-ten Schritt geht die k-te Spalte $\boldsymbol{e}^k$ von $\boldsymbol{X}^{(k)}$ in

$$\boldsymbol{D}_k\boldsymbol{M}_k\boldsymbol{e}^k = (-m_{1k}, \ldots, -m_{k-1,k}, d_k, -m_{k+1,k}, \ldots, -m_{nk})^{\mathsf{T}}$$

über, und die Spalten 1 bis $k-1$ von $\boldsymbol{X}^{(k)}$ werden genauso transformiert wie die Spalten $k+1$ bis n von $\boldsymbol{A}^{(k)}$. Die signifikanten Elemente von $\boldsymbol{X}^{(k)}$ und $\boldsymbol{A}^{(k)}$ können daher gemäß

$$\boldsymbol{Z}^{(k)} := \left(\begin{array}{ccc|ccc} x_{11}^{(k)} & \ldots & x_{1,k-1}^{(k)} & a_{1k}^{(k)} & \ldots & a_{1n}^{(k)} \\ \vdots & & \vdots & \vdots & & \vdots \\ x_{n1}^{(k)} & \ldots & x_{n,k-1}^{(k)} & a_{nk}^{(k)} & \ldots & a_{nn}^{(k)} \end{array}\right)$$

in einer (n, n)-Matrix $\boldsymbol{Z}^{(k)}$ zusammengefaßt und simultan transformiert werden. Bei in-situ-Realisierung ergibt sich dann das folgende Verfahren der *Gauß-Jordan-Inversion*:

6.5.3. *Berechnung von A^{-1} nach dem Gauß-Jordan-Verfahren ohne Pivotisierung.* Die reguläre Matrix $A \in \mathbf{R}^{n,n}$ besitze die Dreiecksfaktorisierung $A = LR$. Dann ist das Verfahren

for $k := 1(1)n$ **do**

  Berechnung der k-ten Spalte: $a_{kk} := 1/a_{kk}$
  for $i := 1(1)n$ **do** [**if** $i \neq k$ **then** $a_{ik} := -a_{ik} * a_{kk}$]
  Berechnung der übrigen Spalten:
  for $j := 1(1)n$ **do**

    if $j \neq k$ **then**

      for $i := 1(1)n$ **do** [**if** $i \neq k$ **then** $a_{ij} := a_{ij} + a_{ik} * a_{kj}$]
      $a_{kj} := a_{kj} * a_{kk}$

in exakter Arithmetik durchführbar und überspeichert A mit A^{-1}.

Aufwand: $\sim n^3$ opms

6.5.4. *Bemerkung.* (i) Das Verfahren 6.5.3 erfordert denselben Aufwand wie 6.5.1. Sein Vorteil ist das kompakte Programm, allerdings ist das Rundungsfehlerverhalten etwas schlechter, vgl. (ii). Die nichtpivotisierte Version 6.5.3 ist i. allg. nur für diagonaldominante oder symmetrische und definite Matrizen geeignet; im letzten Fall braucht nur mit einem Dreieck gearbeitet zu werden. Für andere Matrizen muß eine der Pivotisierungsstrategien aus 5.2 angewendet werden. Man beachte, daß die Restmatrizen $M^{(k)}$ des Gauß-Jordan-Verfahrens und des Gaußschen Algorithmus bei gleicher Pivotisierung identisch sind. Im Fall der Spaltenpivotisierung gilt $|m_{ik}| \leqq 1$ für $i = k + 1, \ldots, n$, aber i. allg. nicht für $i = 1, \ldots, k - 1$. Bei Spaltenpivotisierung ist eine als Stiefelsches Rotationsverfahren bezeichnete Implementierung zweckmäßig, bei der die Pivotzeile stets in die n-te Zeile gebracht wird und die übrigen zyklisch nach oben verschoben werden. Das Merken und Rückgängigmachen der Vertauschungen wird dabei überflüssig.

(ii) Die Rundungsfehleranalyse der *Gauß-Jordan-Inversion* zeigt, daß das Verfahren für nicht zu schlecht konditioniertes A durchführbar und *numerisch stabil* im Sinne von

$$\frac{\|X - A^{-1}\|}{\|A^{-1}\|} \leqq vF \operatorname{cond}(A) \tag{20}$$

ist, wobei $F = F(n)$ von der Matrizenklasse, der Pivotisierungsstrategie und der Norm abhängt. *Dagegen liegt numerische Gutartigkeit i. allg. nicht vor*, und beide Residuen $\|AX - I\|$, $\|XA - I\|$ können von der Größenordnung $vF[\operatorname{cond}(A)]^2$ sein, d. h., die berechnete Inverse X braucht weder eine gute Rechts- noch Linksinverse zu sein. Ebenso ist die *Berechnung der Lösung x von $Ax = b$ nach dem Gauß-Jordan-Verfahren* (18) *ein numerisch stabiler, aber i. allg. kein numerisch gutartiger Prozeß* und sollte daher durch den billigeren und gutartigen Gaußschen Algorithmus ersetzt werden. □

Wir diskutieren als nächstes das Verhalten von A^{-1} bei Rang-1-Änderungen von A.

6.5.5. *Aufdatierung der inversen Matrix.* Es sei $\boldsymbol{A} \in \mathbf{R}^{n,n}$ eine reguläre Matrix. Dann ist

$$\bar{\boldsymbol{A}} := \boldsymbol{A} + \boldsymbol{u}\boldsymbol{v}^{\mathsf{T}}, \qquad \boldsymbol{u}, \boldsymbol{v} \in \mathbf{R}^n, \tag{21}$$

genau dann regulär, wenn

$$\boldsymbol{v}^{\mathsf{T}}\boldsymbol{A}^{-1}\boldsymbol{u} \neq -1 \tag{22}$$

gilt. Ist die Regularitätsbedingung (22) erfüllt, so gilt

$$\bar{\boldsymbol{A}}^{-1} = (\boldsymbol{A} + \boldsymbol{u}\boldsymbol{v}^{\mathsf{T}})^{-1} = \boldsymbol{A}^{-1} - \frac{\boldsymbol{A}^{-1}\boldsymbol{u}\boldsymbol{v}^{\mathsf{T}}\boldsymbol{A}^{-1}}{1 + \boldsymbol{v}^{\mathsf{T}}\boldsymbol{A}^{-1}\boldsymbol{u}}. \tag{23}$$

Beweis. Falls (22) gilt, folgt (23) durch direktes Nachrechnen, indem die rechte Seite mit $\boldsymbol{A} + \boldsymbol{u}\boldsymbol{v}^{\mathsf{T}}$ multipliziert wird. Im Fall $\boldsymbol{v}^{\mathsf{T}}\boldsymbol{A}^{-1}\boldsymbol{u} = -1$ ist $\boldsymbol{w} := \boldsymbol{A}^{-1}\boldsymbol{u} \neq \boldsymbol{o}$ und $\bar{\boldsymbol{A}}\boldsymbol{w} = (1 + \boldsymbol{v}^{\mathsf{T}}\boldsymbol{A}^{-1}\boldsymbol{u})\,\boldsymbol{u} = \boldsymbol{o}$, also $\bar{\boldsymbol{A}}$ singulär. □

Die Aufdatierungsformel (23) wird nach WOODBURY und SHERMAN/MORRISON benannt, siehe B 6.7. Sie zeigt, daß eine Rang-1-Änderung von $\boldsymbol{A}$ eine Rang-1-Änderung von $\boldsymbol{A}^{-1}$ bewirkt, und $\bar{\boldsymbol{A}}^{-1}$ kann mit $\sim 3n^2$ opms aus $\boldsymbol{A}^{-1}$ berechnet werden, vgl. Ü 6.5.2.

Wir bemerken abschließend, daß inverse Matrizen in den Anwendungen meist in Ausdrücken der Form

$$\boldsymbol{X} = \boldsymbol{A}^{-1}\boldsymbol{B} \quad \text{bzw.} \quad \boldsymbol{Y} = \boldsymbol{C}\boldsymbol{A}^{-1} \tag{24}$$

mit $\boldsymbol{B}, \boldsymbol{C}^{\mathsf{T}} \in \mathbf{R}^{n,r}$ auftreten. *Die naive Auswertung durch explizite Berechnung von* $\boldsymbol{A}^{-1}$ *und Multiplikation mit* $\boldsymbol{B}$ *bzw.* $\boldsymbol{C}$ *mit* $\sim n^2(n + r)$ opms *sollte grundsätzlich vermieden werden.* Stattdessen wird zu den äquivalenten Matrixgleichungen

$$\boldsymbol{A}\boldsymbol{X} = \boldsymbol{B} \quad \text{bzw.} \quad \boldsymbol{A}^{\mathsf{T}}\boldsymbol{Y}^{\mathsf{T}} = \boldsymbol{C}^{\mathsf{T}} \tag{25}$$

übergegangen und $\boldsymbol{X} = (\boldsymbol{x}^1, \ldots, \boldsymbol{x}^r)$ spaltenweise, $\boldsymbol{Y} = (\boldsymbol{y}^1, \ldots, \boldsymbol{y}^r)^{\mathsf{T}}$ zeilenweise aus den Gleichungssystemen

$$\boldsymbol{A}\boldsymbol{x}^j = \boldsymbol{b}^j \quad \text{bzw.} \quad \boldsymbol{A}^{\mathsf{T}}\boldsymbol{y}^j = \boldsymbol{c}^j \qquad (j = 1, \ldots, r) \tag{26}$$

berechnet, wobei $\boldsymbol{B} = (\boldsymbol{b}^1, \ldots, \boldsymbol{b}^r)$ und $\boldsymbol{C}^{\mathsf{T}} = (\boldsymbol{c}^1, \ldots, \boldsymbol{c}^r)$ ist. Dies erfordert eine Dreiecksfaktorisierung und die Behandlung r rechter Seiten, also $\sim n^2(n/3 + r)$ opms, und ist vom Rundungsfehlereinfluß günstiger, vgl. Ü 6.5.4.

Übungsaufgaben

Ü 6.5.1. Man überlege sich, daß S1, S2 aus 6.5.1 in den folgenden n- bzw. $(n - 1)$-stufigen Versionen realisiert werden können:

S1′: **for** $k := 1(1)n$ **do**

```
a_kk := 1/a_kk
for i := 1(1)k − 1 do a_ik := −a_ik * a_kk
for j := k + 1(1)n do
    r := a_kj, a_kj := 0
    for i := 1(1)k do a_ij := a_ij + a_ik * r
```

S2′: **for** $k := n - 1(-1)1$ **do**

| **for** $i := k + 1(1)n$ **do** $[l_i := a_{ik}, a_{ik} := 0]$
| **for** $j := k + 1(1)n$ **do**
| **for** $i := 1(1)n$ **do** $a_{ik} := a_{ik} - a_{ij} * l_j$

Hinweis: S1′ entsteht aus S1, indem der Beitrag des im k-ten Hauptschritt berechneten s_{ik} zu jeder der Summen $\sum_{p=i}^{j-1} s_{ip} r_{pj}$ $(k < j)$ aus (5) sofort auf dem Platz von s_{ij} akkumuliert wird. Überdies ist S1′ gerade das Gauß-Jordan-Inversionsverfahren für den Sonderfall $\boldsymbol{A} = \boldsymbol{R}$. S2′ entsteht durch Auswertung von $\boldsymbol{T} = \boldsymbol{S}\boldsymbol{L}^{-1} = \boldsymbol{S}\boldsymbol{L}_{n-1}(-\boldsymbol{l}^{n-1}) \cdots \boldsymbol{L}_1(-\boldsymbol{l}^1)$.

Ü 6.5.2. Es sei $\bar{\boldsymbol{A}} = \boldsymbol{A} + \boldsymbol{u}\boldsymbol{v}^\top$ mit regulärem $\boldsymbol{A}$, $\bar{\boldsymbol{A}}$. Man überzeuge sich, daß der nachfolgende Algorithmus eine Umsetzung der Formel (23) mit $\sim 3n^2$ opms darstellt, bei der $\boldsymbol{B} := \boldsymbol{A}^{-1}$ mit $\bar{\boldsymbol{B}} := \bar{\boldsymbol{A}}^{-1}$ überspeichert wird:

$$\boldsymbol{w} := \boldsymbol{B}\boldsymbol{u}, \qquad \alpha := 1 + \boldsymbol{v}^\top \boldsymbol{w}$$
$$\boldsymbol{u} := \boldsymbol{w}/\alpha, \qquad \boldsymbol{w} := \boldsymbol{B}^\top \boldsymbol{v}$$
$$\boldsymbol{B} := \boldsymbol{B} - \boldsymbol{u}\boldsymbol{w}^\top.$$

Ü 6.5.3. Man spezifiziere die Änderungsformel (23) für die Sonderfälle der Änderung einer Zeile, einer Spalte und eines Elementes von $\boldsymbol{A}$.

Ü 6.5.4. Man zeige: (i) Wenn $\boldsymbol{X}$ eine im Sinne von

$$\|\boldsymbol{X}\boldsymbol{A} - \boldsymbol{I}\| \leqq \nu F \operatorname{cond}(\boldsymbol{A})$$

gute Linksinverse ist, genügt $\boldsymbol{x} := \boldsymbol{X}\boldsymbol{b}$ der Abschätzung

$$\|\boldsymbol{x} - \boldsymbol{A}^{-1}\boldsymbol{b}\| \leqq \nu F \operatorname{cond}(\boldsymbol{A}) \|\boldsymbol{A}^{-1}\boldsymbol{b}\|,$$

d. h., $\boldsymbol{X}$ liefert auch eine gute Lösung von $\boldsymbol{A}\boldsymbol{x} = \boldsymbol{b}$.
(ii) Wenn $\boldsymbol{X}$ eine im Sinne von

$$\|\boldsymbol{A}\boldsymbol{X} - \boldsymbol{I}\| \leqq \nu F \operatorname{cond}(\boldsymbol{A})$$

gute Rechtsinverse ist, genügt $\boldsymbol{x} := \boldsymbol{X}\boldsymbol{b}$ dagegen der Abschätzung

$$\|\boldsymbol{x} - \boldsymbol{A}^{-1}\boldsymbol{b}\| \leqq \nu F [\operatorname{cond}(\boldsymbol{A})]^2 \|\boldsymbol{A}^{-1}\boldsymbol{b}\|,$$

und der Exponent 2 kann i. allg. nicht durch 1 ersetzt werden.

Bemerkungen zum Kapitel 6

B 6.1. Die Faktorisierung $\boldsymbol{A} = \boldsymbol{L}\boldsymbol{L}^\top$ einer symmetrisch positiv definiten Matrix und das zugehörige Verfahren zur direkten Berechnung von $\boldsymbol{L}$ wurden von dem französischen Offizier CHOLESKY eingeführt und zur Lösung von Normalgleichungen linearer Quadratmittelprobleme verwendet; zu den letzteren Begriffen siehe Kapitel 9. Nach dem Tod CHOLESKYS ist das Verfahren durch dessen Freund BENOIT 1924 veröffentlicht worden. Algorithmen zur symmetrischen Faktorisierung indefiniter symmetrischer Matrizen sind dagegen erst in den letzten 15 Jahren entwickelt worden, siehe BUNCH/PARLETT [71] für eine detaillierte Darstellung der Grundideen und historische Kommentare, vgl. auch BUNCH [71] und BUNCH/KAUFMAN [77]. Die hier verwendete Pivotisierungsstrategie geht auf BUNCH/KAUFMAN/PARLETT [76] zurück und wird auch von den LINPACK-Autoren DONGARRA et al. [79] verwendet. Instruktive Vergleiche mit alternativen Methoden — insbesondere mit der von AASEN [71] vorgeschlagenen Kongruenztransformation auf Tridiagonalform — haben BARWELL/GEORGE [76] durchgeführt.

B 6.2. Direkte Dreiecksfaktorisierungen nichtsymmetrischer Matrizen sind u. a. von BANACHIEWICZ, CROUT und DOOLITTLE angegeben worden. Sie unterscheiden sich in der Zuordnung

der Einsdiagonalen zu $\boldsymbol{L}$ oder $\boldsymbol{R}$ und in der Berechnungsreihenfolge. Für Details und Quellenangaben muß auf die Spezialliteratur — etwa HOUSEHOLDER [64] und FADDEEV/FADDEEVA [63] — verwiesen werden.

B 6.3. Die hier beschriebenen Algorithmen zur Aufdatierung von Cholesky-Faktorisierungen gehen im Fall $\alpha = 1$ auf GOLUB [65] und im Fall $\alpha = -1$ auf SAUNDERS [72] zurück und sind auch in LINPACK aufgenommen worden. Alternative und z. T. etwas billigere Varianten haben FLETCHER/POWELL [74], BERSENEV [79], DAX [83] u. a. angegeben. SORENSEN [77] behandelt die Aufdatierung von BKP-Faktorisierungen.

Zur Aufdatierung von Dreiecksfaktorisierungen nichtsymmetrischer Matrizen wurden bisher nur Algorithmen publiziert, die entweder ohne Pivotisierung arbeiten — also nur für spezielle Matrizenklassen stabil sind — oder bei denen die Dreiecksform eines Faktors allmählich zerstört wird, siehe etwa BENNETT [65] oder BARTELS/STOER/ZENGER [71]. Eine Wende brachte der Elementareliminationsschritt 6.3.1, der erstmals von BURMEISTER [76] in einem unveröffentlichten Manuskript für den Fall der Spaltenmodifikation entsprechend 6.3.5(ii) vorgeschlagen worden ist. Durch die Autoren wurde dieser Algorithmus auf den Fall beliebiger Rang-1-Modifikationen erweitert und auf sein Fehlerverhalten untersucht. Unabhängig und etwa gleichzeitig ist diese Idee von FLETCHER/MATTHEWS [83] gefunden worden. Eine Übersicht über Modifikationstechniken haben GILL/MURRAY [77] gegeben, vgl. auch 10.3 und die dort zitierte Literatur über die Aufdatierung von $\boldsymbol{QR}$-Faktorisierungen.

B 6.4. Algorithmen zur Faktorisierung von Bandmatrizen sind in WILKINSON/REINSCH [71] und DONGARRA et al. [79] implementiert worden, und zwar auch für den nicht ganz einfachen Fall nichtsymmetrischer Matrizen und Rechnung mit Spaltenpivotisierung.

B 6.5. Über die Lösung schwach besetzter Gleichungssysteme existiert eine umfangreiche Literatur. Wir verweisen lediglich auf die Sammelbände von BUNCH/ROSE [76], DUFF/STEWART [79] und DUFF [81], die Monographien von GEORGE/LIU [81] und PISSANETZKY [84] sowie den Übersichtsartikel von IKRAMOV [82], wo zahlreiche weiterführende Literaturhinweise zu finden sind. Zum erstenmal scheint die schwache Besetztheit einer Matrix bei Eliminationsverfahren von MARKOWITZ [57] in der linearen Optimierung ausgenutzt worden zu sein.

B 6.6. Der Inversionsalgorithmus 6.5.1 ist eine verkettete Variante des LINPACK-Verfahrens, das in Ü 6.5.1 angegeben worden ist. Das Gauß-Jordan-Verfahren war vermutlich weder GAUSS noch dem Geodäten JORDAN bekannt, trotzdem hat sich der Name in der Literatur eingebürgert. Die Fehleranalyse geht auf PETERS/WILKINSON [75] zurück. Die Implementierung der spaltenpivotisierten Version nach dem Rotationsprinzip gemäß 6.5.4(ii) kann bei STIEFEL [61] nachgelesen werden. Das Gauß-Jordansche Inversionsverfahren ist identisch mit dem sog. Austauschverfahren der linearen Optimierung.

B 6.7. Die Änderungsformel (23) wurde von SHERMAN/MORRISON 1953 angegeben; bereits vorher hatte WOODBURY 1950 eine analoge Formel für Rang-m-Änderungen $\boldsymbol{A} + \boldsymbol{U}\boldsymbol{V}^T$, $\boldsymbol{U}, \boldsymbol{V} \in \mathbf{R}^{n,m}$, aufgestellt; siehe ZIELKE [70] für eine detaillierte Diskussion solcher Änderungsformeln und historische Kommentare.

7. Zusammenfassung zum Teil II

Unter den direkten, d. h. in exakter Arithmetik in endlich vielen Schritten zum Ziel führenden Verfahren zur Lösung regulärer Gleichungssysteme nimmt der Gaußsche Algorithmus mit seinen verschiedenen Modifikationen eine herausragende Rolle ein. Zweckmäßig und heute üblich ist die Interpretation als Dreiecksfaktorisierung und die Trennung des Faktorisierungsprozesses von der Behandlung einer rechten Seite,

was insbesondere bei der Lösung mehrerer Gleichungssysteme mit derselben Koeffizientenmatrix vorteilhaft ist.

Für die wesentlichen Klassen der diagonaldominanten und symmetrischen positiv definiten Matrizen ist die Dreiecksfaktorisierung ohne Pivotisierung durchführbar und numerisch gutartig. Für nicht zu diesen Klassen gehörende Matrizen muß die Gutartigkeit durch eine geeignete Pivotisierungsstrategie erzwungen werden, was allerdings zu einer Erhöhung des Organisationsaufwandes führt. Es besteht heute Übereinstimmung in der Ansicht, daß die vergleichsweise billige Spaltenpivotisierung in den meisten Fällen ausreichend ist, obwohl die das schlechteste Verhalten erfassenden theoretischen Abschätzungen recht pessimistisch sind, siehe 5.2. und 5.3.

Wenn die Pivotisierung nicht ausreicht, um eine akzeptable Dreiecksfaktorisierung zu liefern, kann i. allg. durch wenige Schritte der iterativen Verbesserung gemäß 5.4 eine akzeptable Lösung des Gleichungssystems erhalten werden. Falls die iterative Verbesserung nicht konvergiert, muß zu einer Arithmetik mit höherer Genauigkeit übergegangen werden. Eine weitere Möglichkeit stellt in diesem Fall der Ersatz der Dreiecksfaktorisierung durch orthogonale Faktorisierungen dar. Mit verdoppeltem Aufwand läßt sich die $\boldsymbol{QR}$-Faktorisierung berechnen, siehe 10.2 und speziell 10.2.11(iii). Ein Höchstmaß an Information gibt die Singulärwertzerlegung, die allerdings den vierfachen Aufwand der Dreiecksfaktorisierung erfordert, siehe 11.1.

Ein wesentlicher Vorteil des Gaußschen Algorithmus ist seine Anpassungsfähigkeit. Dies betrifft einmal die Möglichkeit, spezielle Eigenschaften der Matrix zur Aufwandsreduktion auszunutzen. Für symmetrische Matrizen kann der Aufwand durch Verwendung der Cholesky-Faktorisierung im definiten Fall oder der neueren BKP-Faktorisierung im indefiniten Fall halbiert werden, siehe 6.1. Wenn die Matrix Bandstruktur hat, überträgt sich diese auf die Dreiecksfaktoren. Falls ohne Pivotisierung gearbeitet werden kann, bleibt die Bandbreite erhalten; bei Spaltenpivotisierung hat der $\boldsymbol{R}$-Faktor i. allg. eine größere Bandbreite als der entsprechende Teil der Originalmatrix, siehe 6.4. Bei nicht zu großer Bandbreite sind daher auch Systeme sehr hoher Dimension effektiv lösbar. In Verbindung mit Kompaktspeichertechniken kann der Gaußsche Algorithmus auch zur Lösung großer Gleichungssysteme mit schwach besetzten Koeffizientenmatrizen, die keine Bandstruktur haben, verwendet werden, siehe wieder 6.4.

Zum anderen kann der Gaußsche Algorithmus speziellen Eigenschaften des Computers, der Programmiersprache und des Betriebssystems angepaßt werden. Bei spaltenweiser Speicherung von zweidimensionalen Feldern läßt sich unnötiger Seitenwechsel durch spaltenweise Orientierung der inneren Schleifen vermeiden; dasselbe Vorgehen ist auch für gewisse Vektorrechner zweckmäßig, vgl. auch Kapitel 16. Wenn Skalarprodukte billig mit höherer Genauigkeit akkumuliert werden können, empfehlen sich die direkten Faktorisierungen aus 6.2. Für Parallelcomputer stellt allerdings die $\boldsymbol{QR}$-Faktorisierung nach GIVENS — siehe 10.2. — eine Konkurrenz für die Dreiecksfaktorisierung dar, da sie keine Pivotisierung erfordert und besser parallelisierbar ist. Dies trifft insbesondere für die sog. „systolic arrays" zu; zu diesen siehe Kapitel 16.

Schließlich lassen sich Dreiecksfaktorisierungen bei Rang-1-Änderungen der Matrix billig und numerisch stabil aufdatieren, was für gewisse Aufgabenklassen — etwa die

Lösung nichtlinearer Gleichungssysteme und linearer wie nichtlinearer Optimierungsaufgaben — von Bedeutung ist. Zusammenfassend kann festgestellt werden, daß der Gaußsche Algorithmus für nicht zu schlecht konditionierte Systeme mittlerer Dimension — n in der Größenordnung von einigen Hundert — und für schwach besetzte Systeme hoher Dimension mit Bandstruktur oder unregelmäßiger Besetztheitsstruktur das bevorzugte Verfahren darstellt und kaum durch Konkurrenten gefährdet ist.

Etwas anders ist die Situation bei schwach besetzten Systemen hoher Dimension und regelmäßiger Besetztheitsstruktur, wie sie bei der Diskretisierung partieller Differentialgleichungen nach der Methode der finiten Elemente oder dem Differenzenverfahren entstehen. Neben speziellen Versionen des Cholesky-Verfahrens — siehe etwa GEORGE/LIU [81] — treten hier iterative Verfahren als ernsthafte Konkurrenten auf. Auf solche Verfahren kann im Rahmen dieser Einführung nicht eingegangen werden. Wir erwähnen lediglich die klassischen Darstellungen VARGA [62], YOUNG [71] und HAGEMAN/YOUNG [81] und von den neueren Entwicklungen die Methode der alternierenden Dreiecksfaktorisierung — siehe SAMARSKIĬ/NIKOLAEV [78] —, die Mehrgitterverfahren — siehe etwa HACKBUSCH [85] — und iterative Verfahren mit Vorkonditionierungen. Ein Beispiel für die letztgenannte Klasse sind die ICCG-Verfahren, bei denen eine sog. unvollständige Cholesky-Faktorisierung („Incomplete Cholesky") als Vorkonditionierung für das Verfahren der konjugierten Gradienten („Conjugate Gradients") verwendet wird. Zu den Verfahren der konjugierten Gradienten sei auf KUZNECOV [83] verwiesen.

Auf sehr spezielle Gleichungssysteme konnten wir nicht eingehen, etwa auf solche mit Töplitz-Matrizen — siehe VOEVODIN/TYRTYŠNIKOV [83] — und Ljapunov-Gleichungen — siehe IKRAMOV [84]. Diese und andere Gleichungstypen werden auch von GOLUB/VAN LOAN [83] behandelt.

III. Lineare Quadratmittelprobleme

Nachdem wir im vorangegangenen Teil II Gleichungssysteme mit quadratischen und regulären Koeffizientenmatrizen betrachtet haben, lassen wir jetzt beliebige rechteckige Koeffizientenmatrizen $\boldsymbol{A} \in \mathbf{R}^{m,n}$ zu. Die zugehörigen Gleichungssysteme werden dann auch *rechteckig* genannt: Im Fall $m \geqq n$ spricht man von *überbestimmten*, im Fall $m < n$ von *unterbestimmten* Systemen. Teil III ist der Quadratmittellösung überbestimmter linearer Gleichungssysteme gewidmet. In den einführenden Abschnitten gehen wir dabei auch auf unterbestimmte Systeme und klassische Lösungen ein.

8. Rechteckige Gleichungssysteme und Quadratmittelprobleme

8.1. Klassische Lösungen, Quadratmittellösungen, Pseudoinverse

A. *Lösungen und Quadratmittellösungen rechteckiger Gleichungssysteme*

Gegeben ist das rechteckige Gleichungssystem

$$\boldsymbol{A}\boldsymbol{x} = \boldsymbol{b} \tag{1}$$

mit $\boldsymbol{A} \in \mathbf{R}^{m,n}$, $\boldsymbol{b} \in \mathbf{R}^m$. Gesucht sind alle Lösungen $\boldsymbol{x} \in \mathbf{R}^n$ von (1), deren Gesamtheit die *Lösungsmenge* $\mathcal{L}_0 = \mathcal{L}_0(\boldsymbol{A}, \boldsymbol{b}) := \{\boldsymbol{x} \in \mathbf{R}^n : \boldsymbol{A}\boldsymbol{x} = \boldsymbol{b}\}$ bildet. Zur weiteren Untersuchung von (1) ziehen wir die Singulärwertzerlegung

$$\boldsymbol{A} = \boldsymbol{U}\boldsymbol{\Sigma}\boldsymbol{V}^{\mathsf{T}} \tag{2}$$

von $\boldsymbol{A}$ mit $\boldsymbol{\Sigma} = \operatorname{diag}(\sigma_1, \ldots, \sigma_r, 0, \ldots, 0)$, $\sigma_1 \geqq \cdots \geqq \sigma_r > 0$, $r = \operatorname{rang}(\boldsymbol{A})$ und orthogonalen Matrizen $\boldsymbol{U} \in \mathbf{R}^{m,m}$, $\boldsymbol{V} \in \mathbf{R}^{n,n}$ heran, vgl. 1.2.14. Einsetzen in (1) führt unter Beachtung von $\boldsymbol{U}^{-1} = \boldsymbol{U}^{\mathsf{T}}$ auf das äquivalente System

$$\boldsymbol{\Sigma}\boldsymbol{\xi} = \boldsymbol{\beta} \tag{3}$$

für die transformierten Größen

$$\boldsymbol{\xi} = (\xi_j) := \boldsymbol{V}^{\mathsf{T}}\boldsymbol{x}, \qquad \boldsymbol{\beta} = (\beta_i) := \boldsymbol{U}^{\mathsf{T}}\boldsymbol{b}, \tag{4}$$

und (3) läßt sich komponentenweise als

$$\begin{aligned} \sigma_1\xi_1 &= \beta_1, \\ &\vdots \\ \sigma_r\xi_r &= \beta_r, \\ 0 &= \beta_{r+1}, \\ &\vdots \\ 0 &= \beta_m \end{aligned} \tag{5}$$

schreiben. Wegen $\boldsymbol{x} = \boldsymbol{V}\boldsymbol{\xi}$ sind die ξ_j gerade die Komponenten von $\boldsymbol{x}$ in der orthonormierten Basis des Urbildraumes $\mathbf{R}^n$, die durch die Spalten von $\boldsymbol{V}$ gebildet wird. Analog stellen die β_i die Komponenten von $\boldsymbol{b} = \boldsymbol{U}\boldsymbol{\beta}$ in der durch die Spalten von $\boldsymbol{U}$ gebildeten orthonormierten Basis des Bildraumes $\mathbf{R}^m$ dar. In den genannten Basen nimmt das Ausgangssystem die Form (5) an, d. h., die Gleichungen sind entkoppelt, und das System zerfällt in die r Gleichungen

$$\sigma_i\xi_i = \beta_i \qquad (i = 1, \ldots, r) \tag{6}$$

für $\boldsymbol{\xi}$ und die $m - r$ Bedingungen

$$0 = \beta_i \qquad (i = r + 1, \ldots, m) \tag{7}$$

an die transformierte rechte Seite $\boldsymbol{\beta}$. Das System (1) ist also genau dann lösbar, wenn (7) erfüllt ist. Wegen

$$\mathcal{R}(\boldsymbol{\Sigma}) = \{\boldsymbol{\beta} = (\beta_i) \in \mathbf{R}^m : \beta_{r+1} = \cdots = \beta_m = 0\}$$

ist (7) zu $\boldsymbol{\beta} \in \mathcal{R}(\boldsymbol{\Sigma})$, also zu $\boldsymbol{b} = \boldsymbol{U}\boldsymbol{\beta} \in \boldsymbol{U}\mathcal{R}(\boldsymbol{\Sigma}) = \mathcal{R}(\boldsymbol{A})$ äquivalent, vgl. (1.2.24). Die *Lösbarkeitsbedingung* (7) läßt sich daher auch als

$$\boldsymbol{b} \in \mathcal{R}(\boldsymbol{A}) \tag{8}$$

schreiben und ist uns in dieser Gestalt bereits im Abschnitt 1.1.E begegnet, denn die Auflösung von (1) bedeutet, den Vektor $\boldsymbol{b}$ als Linearkombination der Spalten von $\boldsymbol{A}$ mit den Koeffizienten x_j darzustellen. Wenn (8) erfüllt ist, wird (1) *konsistent* genannt, andernfalls heißt das System *inkonsistent*. Im Fall $m = r$, d. h. für eine zeilenreguläre Matrix, ist die Lösbarkeitsbedingung (8) wegen $\mathcal{R}(A) = \mathbf{R}^m$ für jede rechte Seite erfüllt.

Es sei jetzt (1) konsistent. Dann sind die ersten r Komponenten von $\boldsymbol{\xi}$ durch (6) eindeutig festgelegt, während die restlichen in (5) überhaupt nicht auftreten und deshalb beliebig gewählt werden können. Die Lösungen $\boldsymbol{x}$ des Originalsystems (1) ergeben sich wegen (4) gemäß $\boldsymbol{x} = \boldsymbol{V}\boldsymbol{\xi}$ aus den Lösungen $\boldsymbol{\xi}$ des transformierten Systems (5), sie sind also durch

$$\boldsymbol{x} = \boldsymbol{x}^N + \hat{\boldsymbol{x}} \tag{9}$$

mit

$$\boldsymbol{x}^N = \boldsymbol{V}\boldsymbol{\xi}^N, \quad \boldsymbol{\xi}^N = (\xi_1, \ldots, \xi_r, 0, \ldots, 0)^\mathsf{T}, \quad \xi_j = \beta_j/\sigma_j \quad (j = 1, \ldots, r) \tag{10}$$

und

$$\hat{\boldsymbol{x}} = \boldsymbol{V}\hat{\boldsymbol{\xi}}, \quad \hat{\boldsymbol{\xi}} = (0, \ldots, 0, \xi_{r+1}, \ldots, \xi_n)^\mathsf{T}, \quad \xi_j \text{ beliebig} \quad (j = r + 1, \ldots, n) \tag{11}$$

festgelegt. Dabei ist $\boldsymbol{x}^N$ eine spezielle Lösung des *inhomogenen Systems* $\boldsymbol{Ax} = \boldsymbol{b}$ (nämlich diejenige kleinster Euklidischer Norm, wie wir später sehen werden). Wegen

$$\mathcal{N}(\boldsymbol{\Sigma}) = \{\boldsymbol{\xi} = (\xi_j) : \xi_1 = \cdots = \xi_r = 0\}$$

läßt sich $\hat{\boldsymbol{\xi}}$ aus (11) durch $\hat{\boldsymbol{\xi}} \in \mathcal{N}(\boldsymbol{\Sigma})$, $\hat{\boldsymbol{x}}$ also durch $\hat{\boldsymbol{x}} = \boldsymbol{V}\hat{\boldsymbol{\xi}} \in \boldsymbol{V}\mathcal{N}(\boldsymbol{\Sigma}) = \mathcal{N}(\boldsymbol{A})$ charakterisieren. Daher stellt $\hat{\boldsymbol{x}}$ in (9), (11) gerade irgendein Element aus dem $(n - r)$-dimensionalen Nullraum $\mathcal{N}(\boldsymbol{A})$, d. h. irgendeine Lösung des *homogenen Systems* $\boldsymbol{Ax} = \boldsymbol{o}$ dar. Die Lösungsmenge $\mathcal{L}_0$ kann folglich in der Form

$$\mathcal{L}_0(\boldsymbol{A}, \boldsymbol{b}) = \{\boldsymbol{x} + \boldsymbol{x}^N + \hat{\boldsymbol{x}} : \hat{\boldsymbol{x}} \in \mathcal{N}(\boldsymbol{A})\} =: \boldsymbol{x}^N + \mathcal{N}(\boldsymbol{A})$$

geschrieben werden.

Wenn $\mathcal{X}$ ein Teilraum von $\mathbf{R}^n$ der Dimension s und $\boldsymbol{y} \in \mathbf{R}^n$ ist, wird die Menge

$$\mathcal{L} = \{\boldsymbol{x} \in \mathbf{R}^n : \boldsymbol{x} = \boldsymbol{y} + \boldsymbol{z}, \boldsymbol{z} \in \mathcal{X}\} = \boldsymbol{y} + \mathcal{X}$$

allgemein *lineare Mannigfaltigkeit* der Dimension $s = \dim(\mathcal{X})$ genannt. Mit $\boldsymbol{x}' \in \mathcal{L}$ gilt $\boldsymbol{x}'' \in \mathcal{L}$ genau dann, wenn $\boldsymbol{x}' - \boldsymbol{x}'' \in \mathcal{X}$ ist. Insbesondere kann $\mathcal{L}$ auch in der Form $\mathcal{L} = \boldsymbol{y}' + \mathcal{X}$ mit beliebigem $\boldsymbol{y}' \in \mathcal{L}$ geschrieben werden, d. h., für die Darstellung von $\mathcal{L}$ kann ein beliebiger Bezugspunkt $\boldsymbol{y}' \in \mathcal{L}$ gewählt werden.

Mit dieser Terminologie können wir sagen: *Das konsistente System* (1) *hat die* $(n - r)$*-dimensionale lineare Mannigfaltigkeit* $\mathcal{L}_0 = \boldsymbol{x}^* + \mathcal{N}(\boldsymbol{A})$ *zur Lösungsmenge, wobei* $\boldsymbol{x}^*$ *irgendeine Lösung von* (1) *darstellt.*

In den Anwendungen sind die Eingangsdaten $\{\boldsymbol{A}, \boldsymbol{b}\}$ i. allg. fehlerbehaftet. Wenn (1) etwa ein streng überbestimmtes, konsistentes System ist, tritt der Teil (7) im transformierten System (3) wegen $r \leqq n < m$ stets auf. Eine beliebig kleine Störung $\delta\beta_i \neq 0$ in einer der Komponenten $\beta_i = 0$, $i \in \{r + 1, \ldots, m\}$, von $\boldsymbol{\beta}$ wandelt dann das lösbare Originalsystem in ein gestörtes System mit der rechten Seite $\boldsymbol{\beta} + \delta\boldsymbol{\beta}$ bzw. $\boldsymbol{b} + \delta\boldsymbol{b} = \boldsymbol{U}(\boldsymbol{\beta} + \delta\boldsymbol{\beta})$ um, das nicht mehr lösbar ist. Für solche Aufgabenklassen ist daher die Forderung nach der exakten Erfüllung von (1) unnötig streng bzw. sogar unsachgemäß und sollte durch die Forderung ersetzt werden, die p-Norm des Residuums

$$\boldsymbol{r} = \boldsymbol{r}(\boldsymbol{x}) = \boldsymbol{b} - \boldsymbol{Ax}$$

zu minimieren. Obwohl auch die Fälle $p = 1$ und $p = \infty$ von Interesse sind, hat der Fall $p = 2$, d. h. die Verwendung der Euklidischen Norm, die größte Bedeutung für die Anwendungen erlangt. Da wir im Teil III ausschließlich die Euklidische Vektornorm und meist die Spektralnorm einer Matrix verwenden, lassen wir den Index 2 zur Kennzeichnung dieser Normen der Einfachheit halber weg. Das abgeschwächte Ersatzproblem schreibt sich dann als

$$\|\boldsymbol{b} - \boldsymbol{Ax}\| \to \underset{\boldsymbol{x} \in \mathbf{R}^n}{\text{Minimum}}! \quad \text{bzw. kurz:} \quad \boldsymbol{Ax} \cong \boldsymbol{b},$$

und ist äquivalent zu

$$(\boldsymbol{b} - \boldsymbol{Ax})^{\mathsf{T}}(\boldsymbol{b} - \boldsymbol{Ax}) = \sum_{i=1}^{m} [(\boldsymbol{b} - \boldsymbol{Ax})_i]^2 \to \underset{\boldsymbol{x} \in \mathbf{R}^n}{\text{Minimum}}! . \tag{12}$$

Es wird *lineares Quadratmittelproblem* (engl. „linear least squares problem", russ. „линейная проблема наименьших кбадратоб") genannt; man spricht auch von der *Methode der kleinsten (Fehler-) Quadrate* bzw. vom *linearen Ausgleichsproblem*.

Ein Vektor $\boldsymbol{x} \in \mathbf{R}^n$, für den

$$\|\boldsymbol{b} - \boldsymbol{A}\boldsymbol{x}\| \leqq \|\boldsymbol{b} - \boldsymbol{A}\boldsymbol{x}'\| \quad \text{für alle } \boldsymbol{x}' \in \mathbf{R}^n$$

gilt, heißt *Lösung des Quadratmittelproblems* $\boldsymbol{A}\boldsymbol{x} \cong \boldsymbol{b}$ bzw. *Quadratmittellösung des linearen Gleichungssystems* $\boldsymbol{A}\boldsymbol{x} = \boldsymbol{b}$; die Gesamtheit der Lösungen bildet die Lösungsmenge $\mathscr{L} = \mathscr{L}(\boldsymbol{A}, \boldsymbol{b})$. Für konsistente Systeme gilt offensichtlich $\mathscr{L}_0(\boldsymbol{A}, \boldsymbol{b}) = \mathscr{L}(\boldsymbol{A}, \boldsymbol{b})$, so daß eine Verallgemeinerung des klassischen Lösungsbegriffes vorliegt.

Wir betrachten jetzt einen beliebigen Vektor $\boldsymbol{x} \in \mathbf{R}^n$. Wenn $\boldsymbol{A}$ wie eingangs durch die Singulärwertzerlegung ausgedrückt wird, ergibt sich mit den Transformationen (4) die Darstellung

$$\boldsymbol{r} = \boldsymbol{b} - \boldsymbol{A}\boldsymbol{x} = \boldsymbol{U}\boldsymbol{U}^{\mathsf{T}}\boldsymbol{b} - \boldsymbol{U}\boldsymbol{\Sigma}\boldsymbol{V}^{\mathsf{T}}\boldsymbol{x} = \boldsymbol{U}(\boldsymbol{\beta} - \boldsymbol{\Sigma}\boldsymbol{\xi}). \tag{13}$$

Unter Beachtung der Orthogonalinvarianz der Euklidischen Norm folgt daraus

$$\|\boldsymbol{r}\|^2 = \|\boldsymbol{\beta} - \boldsymbol{\Sigma}\boldsymbol{\xi}\|^2 = \sum_{i=1}^{r} (\beta_i - \sigma_i\xi_i)^2 + \sum_{i=r+1}^{m} \beta_i^2. \tag{14}$$

Der erste Summand ist stets nichtnegativ und verschwindet genau dann, wenn

$$\sigma_i\xi_i = \beta_i, \quad \text{also} \quad \xi_i = \beta_i/\sigma_i \qquad (i = 1, \ldots, r)$$

gilt. Die restlichen Komponenten haben auf das Residuum keinen Einfluß und können beliebig gewählt werden. Der Vergleich mit (9) bis (11) zeigt, daß die Quadratmittellösungen identisch sind mit den Lösungen des Gleichungssystems $\boldsymbol{A}\boldsymbol{x} = \bar{\boldsymbol{b}}$, zu dem die rechte Seite

$$\bar{\boldsymbol{b}} := \boldsymbol{U}\bar{\boldsymbol{\beta}}, \qquad \bar{\boldsymbol{\beta}} := (\beta_1, \ldots, \beta_r, 0, \ldots, 0)^{\mathsf{T}} \tag{15}$$

gehört.

8.1.1. *Satz.* Es sei $\boldsymbol{A} \in \mathbf{R}^{m,n}$ gegeben. Dann gilt

(i) Das Quadratmittelproblem $\boldsymbol{A}\boldsymbol{x} \cong \boldsymbol{b}$ ist für jede rechte Seite $\boldsymbol{b} \in \mathbf{R}^m$ lösbar.

(ii) $\boldsymbol{x}$ ist Lösung von $\boldsymbol{A}\boldsymbol{x} \cong \boldsymbol{b}$ genau dann, wenn $\boldsymbol{x}$ das Gleichungssystem

$$\boldsymbol{A}\boldsymbol{x} = \bar{\boldsymbol{b}} \tag{16}$$

mit $\bar{\boldsymbol{b}}$ gemäß (15) löst.

(iii) Jede Lösung $\boldsymbol{x}$ läßt sich in der Form

$$\boldsymbol{x} = \boldsymbol{x}^N + \hat{\boldsymbol{x}} \quad \text{mit } \boldsymbol{x}^N, \hat{\boldsymbol{x}} \text{ gemäß (10), (11)} \tag{17}$$

schreiben.

Satz 8.1.1 besagt: *Die lineare Mannigfaltigkeit*

$$\mathscr{L}(\boldsymbol{A}, \boldsymbol{b}) = \boldsymbol{x}^N + \mathscr{N}(\boldsymbol{A}) \tag{18}$$

der Dimension $\dim \mathcal{N}(\boldsymbol{A}) = n - r$, $r = \operatorname{rang}(\boldsymbol{A})$, *stellt die Lösungsmenge von* $\boldsymbol{Ax} \cong \boldsymbol{b}$ *dar.* Die Lösung ist eindeutig, wenn $r = n$ gilt, d. h., wenn $\boldsymbol{A}$ spaltenregulär ist.

Wir bemerken ferner, daß das optimale Residuum $\boldsymbol{r} = \boldsymbol{b} - \boldsymbol{Ax}$ für alle Lösungen $\boldsymbol{x}$ den Wert

$$\boldsymbol{r} = \boldsymbol{b} - \bar{\boldsymbol{b}} = \boldsymbol{U}(\beta - \bar{\beta}) = \boldsymbol{U}\varrho, \qquad \varrho := (0, \ldots, 0, \beta_{r+1}, \ldots, \beta_m)^T$$

besitzt, also durch $\boldsymbol{A}$ und $\boldsymbol{b}$ eindeutig festgelegt ist, vgl. (13).

Nachdem die Singulärwertzerlegung von $\boldsymbol{A}$ über die Residuumsdarstellung (14) in einfacher Weise zu Aussagen über Lösbarkeit und Lösungsmenge des Quadratmittelproblems $\boldsymbol{Ax} \cong \boldsymbol{b}$ geführt hat, soll im folgenden eine von der Singulärwertzerlegung unabhängige, geometrisch motivierte Charakterisierung der Quadratmittellösungen hergeleitet werden. Wir betrachten dazu den Fall $m = 3$, $n = \operatorname{rang}(\boldsymbol{A}) = r = 2$, $\boldsymbol{A} = (\boldsymbol{a}^1, \boldsymbol{a}^2) \in \mathbf{R}^{3,2}$, $\boldsymbol{b} \in \mathbf{R}^3$, $\boldsymbol{b} \notin \mathcal{R}(A)$. Für $\boldsymbol{x} = (x_1, x_2)^T \in \mathbf{R}^2$ durchläuft $\boldsymbol{y} = \boldsymbol{Ax} = \boldsymbol{a}^1 x_1 + \boldsymbol{a}^2 x_2$ den Wertebereich $\mathcal{R}(\boldsymbol{A}) = \operatorname{span}\{\boldsymbol{a}^1, \boldsymbol{a}^2\}$. Die Minimierung von $\|\boldsymbol{b} - \boldsymbol{Ax}\|$ bedeutet also, den Vektor $\boldsymbol{b}$ im Sinne der Euklidischen Norm optimal durch eine Linearkombination der Spalten von $\boldsymbol{A}$ zu approximieren, siehe Abb. 8.1.1.

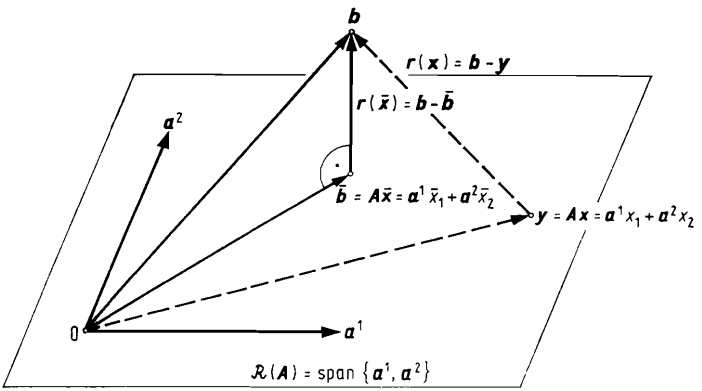


Abb. 8.1.1. Quadratmittellösung $\bar{\boldsymbol{x}}$ und minimales Residuum $\boldsymbol{b} - \bar{\boldsymbol{b}}$

B. *Orthogonale Projektionen, Projektoren und Komplemente*

Aus Abb. 8.1.1 wird anschaulich klar, daß $\boldsymbol{r} = \boldsymbol{b} - \bar{\boldsymbol{b}}$, $\bar{\boldsymbol{b}} = \boldsymbol{A}\bar{\boldsymbol{x}} \in \mathcal{R}(\boldsymbol{A})$, genau dann kleinste Euklidische Länge besitzt, wenn $\boldsymbol{r}$ orthogonal zu allen Vektoren $\boldsymbol{y} \in \mathcal{R}(\boldsymbol{A})$ und $\bar{\boldsymbol{b}}$ in diesem Sinne die orthogonale Projektion von $\boldsymbol{b}$ auf $\mathcal{R}(\boldsymbol{A})$ ist. Im folgenden soll der für das Verständnis der Quadratmittelprobleme wesentliche Begriff des orthogonalen Projektors präzise definiert und charakterisiert werden. Dabei lassen wir statt $\mathcal{R}(\boldsymbol{A})$ einen beliebigen Teilraum $\mathcal{Y}$ zu.

8.1.2. *Aussage.* Es sei $\mathcal{Y}$ ein Teilraum von $\mathbf{R}^m$. Dann gilt:

(i) Zu jedem $\boldsymbol{b} \in \mathbf{R}^m$ gibt es genau ein $\bar{\boldsymbol{b}} \in \mathcal{Y}$ mit

$$\boldsymbol{y}^T(\boldsymbol{b} - \bar{\boldsymbol{b}}) = 0 \quad \text{für alle } \boldsymbol{y} \in \mathcal{Y}, \quad \text{kurz: } \boldsymbol{b} - \bar{\boldsymbol{b}} \perp \mathcal{Y}. \tag{19}$$

Der Vektor $\bar{\boldsymbol{b}}$ heißt *orthogonale Projektion* von $\boldsymbol{b}$ auf $\mathcal{Y}$.

(ii) $\bar{\boldsymbol{b}}$ ist orthogonale Projektion von $\boldsymbol{b}$ auf $\mathcal{Y}$ genau dann, wenn $\bar{\boldsymbol{b}}$ die Aufgabe

$$\|\boldsymbol{b} - \boldsymbol{y}\| \to \underset{\boldsymbol{y} \in \mathcal{Y}}{\text{Minimum}}! \tag{20}$$

löst.

(iii) Es gibt genau eine Matrix $\boldsymbol{P} \in \mathbf{R}^{m,m}$, so daß

$$\bar{\boldsymbol{b}} = \boldsymbol{P}\boldsymbol{b} \tag{21}$$

für alle $\boldsymbol{b} \in \mathbf{R}^m$ gilt. $\boldsymbol{P}$ heißt *orthogonaler Projektor* bzw. kurz *Projektor* auf $\mathcal{Y}$. Wenn $\dim \mathcal{Y} = r$ und $\{\boldsymbol{g}^1, \ldots, \boldsymbol{g}^r\}$ eine Basis von $\mathcal{Y}$ ist, kann $\boldsymbol{P}$ in der Form

$$\boldsymbol{P} = \boldsymbol{G}(\boldsymbol{G}^\mathsf{T}\boldsymbol{G})^{-1}\boldsymbol{G}^\mathsf{T} \tag{22}$$

mit $\boldsymbol{G} := (\boldsymbol{g}^1, \ldots, \boldsymbol{g}^r)$ geschrieben werden.

Beweis. Es sei $\{\boldsymbol{g}^1, \ldots, \boldsymbol{g}^r\}$ irgendeine Basis von $\mathcal{Y}$ und $\boldsymbol{G} := (\boldsymbol{g}^1, \ldots, \boldsymbol{g}^r)$. Dann ist $\operatorname{rang}(\boldsymbol{G}) = \operatorname{rang}(\boldsymbol{G}^\mathsf{T}\boldsymbol{G}) = r$, vgl. (1.2.28), so daß $\boldsymbol{G}^\mathsf{T}\boldsymbol{G}$ regulär und $\boldsymbol{P}$ aus (22) wohldefiniert ist. Wegen $\bar{\boldsymbol{b}} = \boldsymbol{P}\boldsymbol{b} = \boldsymbol{G}\{(\boldsymbol{G}^\mathsf{T}\boldsymbol{G})^{-1}\boldsymbol{G}^\mathsf{T}\boldsymbol{b}\}$ ist $\bar{\boldsymbol{b}} \in \mathcal{R}(\boldsymbol{G}) = \mathcal{Y}$. Da jedes $\boldsymbol{y} \in \mathcal{Y}$ in der Form $\boldsymbol{y} = \boldsymbol{G}\boldsymbol{w}$ mit $\boldsymbol{w} \in \mathbf{R}^r$ dargestellt werden kann, ergibt sich

$$\boldsymbol{y}^\mathsf{T}(\boldsymbol{b} - \bar{\boldsymbol{b}}) = (\boldsymbol{G}\boldsymbol{w})^\mathsf{T}(\boldsymbol{b} - \boldsymbol{P}\boldsymbol{b}) = \boldsymbol{w}^\mathsf{T}(\boldsymbol{G}^\mathsf{T}\boldsymbol{b} - \boldsymbol{G}^\mathsf{T}\boldsymbol{P}\boldsymbol{b}) = \boldsymbol{o},$$

also (19). Es bleibt zu zeigen, daß $\bar{\boldsymbol{b}} = \boldsymbol{P}\boldsymbol{b}$ durch (19) eindeutig festgelegt ist. Wenn (19) auch für ein $\bar{\boldsymbol{c}} \in \mathcal{Y}$ erfüllt ist, folgt $\boldsymbol{y}^\mathsf{T}(\boldsymbol{b} - \bar{\boldsymbol{b}}) = \boldsymbol{y}^\mathsf{T}(\boldsymbol{b} - \bar{\boldsymbol{c}}) = 0$, also $\boldsymbol{y}^\mathsf{T}(\bar{\boldsymbol{b}} - \bar{\boldsymbol{c}}) = 0$ für alle $\boldsymbol{y} \in \mathcal{Y}$. Wegen $\bar{\boldsymbol{b}}, \bar{\boldsymbol{c}} \in \mathcal{Y}$ ist auch $\bar{\boldsymbol{b}} - \bar{\boldsymbol{c}} \in \mathcal{Y}$, so daß speziell $\boldsymbol{y} = \bar{\boldsymbol{b}} - \bar{\boldsymbol{c}}$ gewählt werden kann, was auf $(\bar{\boldsymbol{b}} - \bar{\boldsymbol{c}})^\mathsf{T}(\bar{\boldsymbol{b}} - \bar{\boldsymbol{c}}) = 0$, also $\bar{\boldsymbol{b}} = \bar{\boldsymbol{c}}$ führt. Daher ist $\bar{\boldsymbol{b}}$ eindeutig durch (19) festgelegt, so daß auch $\boldsymbol{P}$ eindeutig durch $\mathcal{Y}$ bestimmt ist und nicht von der speziell gewählten Basis $\{\boldsymbol{g}^1, \ldots, \boldsymbol{g}^r\}$ abhängt.

Es bleibt die Äquivalenz von (i) und (ii) zu zeigen. Nun gilt für beliebiges $\boldsymbol{y} \in \mathcal{Y}$

$$\begin{aligned}\|\boldsymbol{b} - \boldsymbol{y}\|^2 &= \|(\boldsymbol{b} - \bar{\boldsymbol{b}}) + (\bar{\boldsymbol{b}} - \boldsymbol{y})\|^2 = \|\boldsymbol{b} - \bar{\boldsymbol{b}}\|^2 + 2(\boldsymbol{b} - \bar{\boldsymbol{b}})^\mathsf{T}(\bar{\boldsymbol{b}} - \boldsymbol{y}) + \|\bar{\boldsymbol{b}} - \boldsymbol{y}\|^2 \\ &= \|\boldsymbol{b} - \bar{\boldsymbol{b}}\|^2 + \|\bar{\boldsymbol{b}} - \boldsymbol{y}\|^2,\end{aligned}$$

denn wegen $\bar{\boldsymbol{b}} - \boldsymbol{y} \in \mathcal{Y}$ verschwindet das Skalarprodukt, siehe (19). Offensichtlich wird dann $\|\boldsymbol{b} - \boldsymbol{y}\|$ über $\boldsymbol{y} \in \mathcal{Y}$ genau dann minimal, wenn $\boldsymbol{y} = \bar{\boldsymbol{b}}$ ist. □

Zur Illustration stellen wir die vorkommenden Größen für den Fall $m = 2$, $\mathcal{Y} = \operatorname{span}\{\boldsymbol{g}^1\}$, $\dim \mathcal{Y} = r = 1$, in Abb. 8.1.2 dar. Der Projektor $\boldsymbol{Q} = \boldsymbol{I} - \boldsymbol{P}$ und der Teilraum $\mathcal{Y}^\perp$ werden weiter unten eingeführt und können zunächst unbeachtet bleiben.

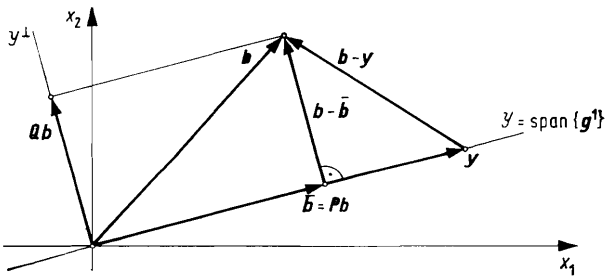


Abb. 8.1.2. Orthogonale Projektion und zugehörige Größen

Orthogonale Projektoren lassen sich in einfacher Weise charakterisieren.

8.1.3. *Aussage.* $\boldsymbol{P} \in \mathbf{R}^{m,m}$ ist genau dann ein orthogonaler Projektor, wenn

$$\boldsymbol{P} = \boldsymbol{P}^{\mathsf{T}} \quad \text{und} \quad \boldsymbol{P}^2 = \boldsymbol{P} \tag{23}$$

gilt. Wenn (23) erfüllt ist, stellt $\boldsymbol{P}$ den Projektor auf $\mathcal{Y} = \mathcal{R}(\boldsymbol{P})$ dar.

Beweis. Es sei $\boldsymbol{P}$ der Projektor auf $\mathcal{Y}$. Dann kann $\boldsymbol{P}$ nach (22) dargestellt werden, und (23) ist erfüllt, wie man durch Einsetzen bestätigt. Genügt $\boldsymbol{P}$ umgekehrt den Forderungen (23), gilt für jedes $\boldsymbol{y} \in \mathcal{Y} := \mathcal{R}(\boldsymbol{P})$, d. h. für jedes $\boldsymbol{y} = \boldsymbol{P}\boldsymbol{v}$, $\boldsymbol{v} \in \mathbf{R}^m$,

$$\boldsymbol{y}^{\mathsf{T}}(\boldsymbol{b} - \boldsymbol{P}\boldsymbol{b}) = (\boldsymbol{P}\boldsymbol{v})^{\mathsf{T}}(\boldsymbol{b} - \boldsymbol{P}\boldsymbol{b}) = \boldsymbol{v}^{\mathsf{T}}(\boldsymbol{P} - \boldsymbol{P}^2)\,\boldsymbol{b} = 0,$$

d. h., $\bar{\boldsymbol{b}} = \boldsymbol{P}\boldsymbol{b}$ ist in der Tat die Projektion von $\boldsymbol{b}$ auf $\mathcal{Y}$. □

Wenn $\mathcal{Y}$ ein Teilraum von $\mathbf{R}^m$ ist, wird das *orthogonale Komplement* $\mathcal{Y}^\perp$ von $\mathcal{Y}$ als Menge aller zu $\mathcal{Y}$ orthogonalen Vektoren gemäß

$$\mathcal{Y}^\perp := \{\boldsymbol{z} \in \mathbf{R}^m : \boldsymbol{z} \perp \mathcal{Y}\} = \{\boldsymbol{z} \in \mathbf{R}^m : \boldsymbol{z}^{\mathsf{T}}\boldsymbol{y} = 0 \text{ für alle } \boldsymbol{y} \in \mathcal{Y}\} \tag{24}$$

eingeführt. Es ist sofort zu sehen, daß $\mathcal{X} := \mathcal{Y}^\perp$ wieder ein Teilraum von $\mathbf{R}^m$ ist und durch die Bedingungen

$$\mathbf{R}^m = \mathcal{Y} + \mathcal{X} \quad \text{und} \quad \mathcal{Y} \perp \mathcal{X}, \quad \text{d. h. } \boldsymbol{y}^{\mathsf{T}}\boldsymbol{z} = 0 \text{ für alle } \boldsymbol{y} \in \mathcal{Y},\ \boldsymbol{z} \in \mathcal{X}, \tag{25}$$

eindeutig festgelegt wird, insbesondere gilt $(\mathcal{Y}^\perp)^\perp = \mathcal{Y}$. Die folgende Aussage zeigt, daß zwischen den Projektoren auf $\mathcal{Y}$ und $\mathcal{X}$ ein einfacher Zusammenhang besteht.

8.1.4. *Aussage.* Es sei $\mathcal{Y}$ ein Teilraum von $\mathbf{R}^m$, und $\boldsymbol{P}$ bezeichne den orthogonalen Projektor auf $\mathcal{Y}$. Dann ist

$$\boldsymbol{Q} := \boldsymbol{I} - \boldsymbol{P} \tag{26}$$

der orthogonale Projektor auf $\mathcal{X} := \mathcal{Y}^\perp$.

Beweis. $\boldsymbol{Q}$ genügt (23) und ist also ein Projektor. Für $\boldsymbol{z} \in \mathcal{X}$ gilt weiter $\boldsymbol{z}^{\mathsf{T}}(\boldsymbol{b} - \boldsymbol{Q}\boldsymbol{b}) = \boldsymbol{z}^{\mathsf{T}}(\boldsymbol{b} - (\boldsymbol{I} - \boldsymbol{P})\,\boldsymbol{b}) = \boldsymbol{z}^{\mathsf{T}}\boldsymbol{P}\boldsymbol{b} = 0$ für alle $\boldsymbol{b}$ wegen $\boldsymbol{P}\boldsymbol{b} \in \mathcal{Y}$, d. h., $\boldsymbol{Q}$ projiziert auf $\mathcal{X}$. □

Aus 8.1.4 folgt, daß jedes $\boldsymbol{x} \in \mathbf{R}^m$ gemäß

$$\boldsymbol{x} = \boldsymbol{y} + \boldsymbol{z} = \boldsymbol{P}\boldsymbol{x} + \boldsymbol{Q}\boldsymbol{x}, \qquad \boldsymbol{y} = \boldsymbol{P}\boldsymbol{x} \in \mathcal{Y}, \qquad \boldsymbol{z} = \boldsymbol{Q}\boldsymbol{x} \in \mathcal{X}, \tag{27}$$

in eindeutiger Weise als Summe zweier Vektoren aus $\mathcal{Y}$ und $\mathcal{X}$ dargestellt werden kann, wobei $\|\boldsymbol{P}\boldsymbol{x} + \boldsymbol{Q}\boldsymbol{x}\|^2 = \|\boldsymbol{P}\boldsymbol{x}\|^2 + 2(\boldsymbol{P}\boldsymbol{x})^{\mathsf{T}}\boldsymbol{Q}\boldsymbol{x} + \|\boldsymbol{Q}\boldsymbol{x}\|^2 = \|\boldsymbol{P}\boldsymbol{x}\|^2 + 2\boldsymbol{x}^{\mathsf{T}}\boldsymbol{P}\boldsymbol{Q}\boldsymbol{x} + \|\boldsymbol{Q}\boldsymbol{x}\|^2$, wegen $\boldsymbol{P}\boldsymbol{Q} = \boldsymbol{P}(\boldsymbol{I} - \boldsymbol{P}) = \boldsymbol{P} - \boldsymbol{P}^2 = \boldsymbol{O}$ also

$$\|\boldsymbol{x}\|^2 = \|\boldsymbol{P}\boldsymbol{x}\|^2 + \|\boldsymbol{Q}\boldsymbol{x}\|^2 \tag{28}$$

gilt. Dies ist gerade der Satz des PYTHAGORAS in dem durch $\boldsymbol{x}$, $\boldsymbol{P}\boldsymbol{x}$ und $\boldsymbol{Q}\boldsymbol{x}$ aufgespannten rechtwinkligen Dreieck, siehe Abb. 8.1.2, wo $\boldsymbol{b}$ statt $\boldsymbol{x}$ vorkommt.

C. *Weitere Charakterisierung von Quadratmittellösungen*

Wir kehren jetzt wieder zum Quadratmittelproblem $\boldsymbol{A}\boldsymbol{x} \cong \boldsymbol{b}$ zurück. Dieses läßt sich auch in der Form (20) mit $\mathcal{Y} = \mathcal{R}(\boldsymbol{A})$ schreiben, denn jedes $\boldsymbol{y} \in \mathcal{R}(\boldsymbol{A})$ ist von der Gestalt $\boldsymbol{y} = \boldsymbol{A}\boldsymbol{x}$, $\boldsymbol{x} \in \mathbf{R}^n$. Nach 8.1.2(ii) ist dann $\boldsymbol{x}$ Lösung genau dann, wenn $\boldsymbol{y} = \boldsymbol{A}\boldsymbol{x}$

$= \bar{\boldsymbol{b}} = \boldsymbol{P}_A \boldsymbol{b}$ gilt, wobei $\boldsymbol{P}_A$ den Projektor auf $\mathcal{R}(\boldsymbol{A})$ bezeichnet. Man rechnet nach, daß $\bar{\boldsymbol{b}}$ aus (15) in der Tat die (mittels der Singulärwertzerlegung bestimmte) Projektion von $\boldsymbol{b}$ auf $\mathcal{R}(\boldsymbol{A})$ ist, siehe Ü 8.1.5. Nach 8.1.2(i) kann $\bar{\boldsymbol{b}}$ auch durch die Bedingung $\boldsymbol{b} - \bar{\boldsymbol{b}} \perp \mathcal{R}(\boldsymbol{A})$, d. h. durch $\boldsymbol{r} \perp \mathcal{R}(\boldsymbol{A})$ mit $\boldsymbol{r} = \boldsymbol{b} - \boldsymbol{A}\boldsymbol{x}$, charakterisiert werden. Dies ist zu $\boldsymbol{r} \in \mathcal{R}(\boldsymbol{A})^\perp$ äquivalent. Das orthogonale Komplement zu $\mathcal{R}(\boldsymbol{A})$ läßt sich aber durch $\boldsymbol{A}^\mathsf{T}$ beschreiben.

8.1.5. *Aussage.* Für jede Matrix $\boldsymbol{A} \in \mathbf{R}^{n,n}$ gilt

$$\mathcal{R}(\boldsymbol{A})^\perp = \mathcal{N}(\boldsymbol{A}^\mathsf{T}).$$

Beweis. Es gilt $\boldsymbol{z} \in \mathcal{R}(\boldsymbol{A})^\perp$ genau dann, wenn $\boldsymbol{y}^\mathsf{T}\boldsymbol{z} = 0$ ist für alle $\boldsymbol{y} \in \mathcal{R}(\boldsymbol{A})$, d. h. für alle $\boldsymbol{y} = \boldsymbol{A}\boldsymbol{x}$, $\boldsymbol{x} \in \mathbf{R}^n$. Dies bedeutet $(\boldsymbol{A}\boldsymbol{x})^\mathsf{T}\boldsymbol{z} = \boldsymbol{x}^\mathsf{T}\boldsymbol{A}^\mathsf{T}\boldsymbol{z} = 0$ für alle $\boldsymbol{x} \in \mathbf{R}^n$, also $\boldsymbol{A}^\mathsf{T}\boldsymbol{z} = \boldsymbol{o}$ oder äquivalent $\boldsymbol{z} \in \mathcal{N}(\boldsymbol{A}^\mathsf{T})$. □

Die Bedingung $\boldsymbol{r} \in \mathcal{R}(\boldsymbol{A})^\perp$ ist also zu $\boldsymbol{r} \in \mathcal{N}(\boldsymbol{A}^\mathsf{T})$, d. h. zu $\boldsymbol{A}^\mathsf{T}\boldsymbol{r} = \boldsymbol{A}^\mathsf{T}(\boldsymbol{b} - \boldsymbol{A}\boldsymbol{x}) = \boldsymbol{o}$ äquivalent.

Zusammenfassend erhalten wir die folgenden Charakterisierungen der Quadratmittellösungen.

8.1.6. *Satz.* Die folgenden Aussagen sind äquivalent:

(i) $\boldsymbol{x}$ ist Lösung des Quadratmittelproblems $\boldsymbol{A}\boldsymbol{x} \cong \boldsymbol{b}$.

(ii) $\boldsymbol{x}$ ist Lösung des Gleichungssystems

$$\boldsymbol{A}\boldsymbol{x} = \boldsymbol{P}_A \boldsymbol{b}. \tag{29}$$

(iii) $\boldsymbol{x}$ genügt der Bedingung

$$\boldsymbol{r} = \boldsymbol{b} - \boldsymbol{A}\boldsymbol{x} \perp \mathcal{R}(\boldsymbol{A}). \tag{30}$$

(iv) $\boldsymbol{x}$ ist Lösung des Gleichungssystems

$$\boldsymbol{A}^\mathsf{T}\boldsymbol{A}\boldsymbol{x} = \boldsymbol{A}^\mathsf{T}\boldsymbol{b}. \tag{31}$$

Die Gleichungen (31) werden die zu $\boldsymbol{A}\boldsymbol{x} \cong \boldsymbol{b}$ gehörenden *Normalgleichungen* genannt: sie sind wie das Gleichungssystem (29) auf Grund der speziellen Gestalt der rechten Seite stets konsistent.

D. *Normallösung und Pseudoinverse*

In 8.1.1 wurde festgestellt, daß die $(n - r)$-dimensionale lineare Mannigfaltigkeit

$$\mathcal{L} = \mathcal{L}(\boldsymbol{A}, \boldsymbol{b}) = \boldsymbol{x}^N + \mathcal{N}(\boldsymbol{A})$$

die Lösungsmenge von $\boldsymbol{A}\boldsymbol{x} \cong \boldsymbol{b}$ darstellt. Im Fall $r < n$ existieren also unendlich viele Lösungen, und es ist sinnvoll, unter diesen eine spezielle Lösung auszuzeichnen. Wenn die Norm von $\boldsymbol{x}$ als Kriterium gewählt wird, ergibt sich aus der Darstellung (17) wegen der Orthogonalität von $\boldsymbol{V}$

$$\|\boldsymbol{x}\|^2 = \|\boldsymbol{\xi}\|^2 = \|\boldsymbol{\xi}^N\|^2 + \|\hat{\boldsymbol{\xi}}\|^2 = \sum_{j=1}^{r} (\beta_j/\sigma_j)^2 + \sum_{j=r+1}^{n} \xi_j^2 \tag{32}$$

für jedes $\boldsymbol{x} = \boldsymbol{V}\boldsymbol{\xi} \in \mathcal{L}$. Hieraus liest man ab, daß $\|\boldsymbol{x}\|$ genau dann minimal wird, wenn

die frei wählbaren Parameter $\xi_{r+1}, \ldots, \xi_n$ gleich 0 gesetzt werden, d. h., wenn $\hat{\boldsymbol{x}} = \boldsymbol{o}$ und folglich $\boldsymbol{x} = \boldsymbol{x}^N$ gilt. Dies ist offensichtlich dann und nur dann der Fall, wenn $\boldsymbol{x}$ orthogonal zu $\mathcal{N}(A)$ ist, siehe Abb. 8.1.3.

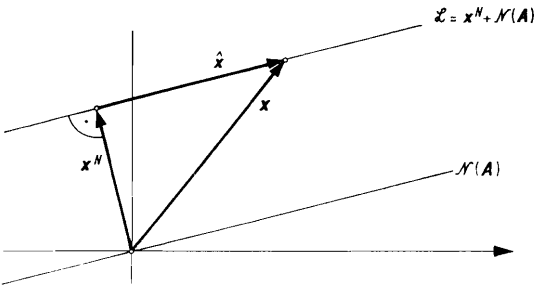


Abb. 8.1.3. Lösungsmenge und Normallösung für $n = 2$, $r = 1$

8.1.7. *Aussage.* Es sei $\mathcal{L}$ die Lösungsmenge des Quadratmittelproblems $\boldsymbol{Ax} \cong \boldsymbol{b}$. Dann gilt

(i) Unter allen Lösungen gibt es genau eine mit kleinster Euklidischer Norm, nämlich $\boldsymbol{x}^N$, d. h., die Aufgabe

$$\|\boldsymbol{x}\| \to \underset{\boldsymbol{x} \in \mathcal{L}}{\text{Minimum}}! \tag{33}$$

ist eindeutig durch $\boldsymbol{x}^N$ lösbar.

(ii) Für $\boldsymbol{x} \in \mathcal{L}$ gilt $\boldsymbol{x} = \boldsymbol{x}^N$ genau dann, wenn

$$\boldsymbol{x} \perp \mathcal{N}(\boldsymbol{A}). \tag{34}$$

Wegen (i) heißt $\boldsymbol{x}^N$ *Minimum-Norm-Lösung*, wegen (ii) auch *Normallösung* von $\boldsymbol{Ax} \cong \boldsymbol{b}$. Ihrer Kürze wegen verwenden wir die letztgenannte Bezeichnung. Für spaltenreguläre Probleme ist $\boldsymbol{x}^N$ die einzige Lösung von $\boldsymbol{Ax} \cong \boldsymbol{b}$.

Wir wollen jetzt untersuchen, wie $\boldsymbol{x}^N$ von $\boldsymbol{b}$ abhängt. Dazu benutzen wir wieder die Singulärwertzerlegung (2) und die Darstellung $\boldsymbol{x}^N = \boldsymbol{V\xi}^N$ mit $\boldsymbol{\xi}^N$ aus (10), vgl. 8.1.1. Wenn

$$\boldsymbol{\Sigma}^+ := \left(\begin{array}{ccc|c} 1/\sigma_1 & & & 0 \\ & \ddots & & \\ & & 1/\sigma_r & \\ \hline & 0 & & 0 \end{array}\right) \in \mathbf{R}^{n,m} \quad \text{und} \quad \boldsymbol{A}^+ := \boldsymbol{V\Sigma}^+\boldsymbol{U}^\top \in \mathbf{R}^{n,m} \tag{35}$$

gesetzt wird, ergibt sich unter Beachtung von $\boldsymbol{\beta} = \boldsymbol{U}^\top\boldsymbol{b}$

$$\boldsymbol{x}^N = \boldsymbol{V\Sigma}^+\boldsymbol{U}^\top\boldsymbol{b} = \boldsymbol{A}^+\boldsymbol{b}, \tag{36}$$

d. h., *die Abbildung* $\boldsymbol{b} \to \boldsymbol{x}^N$ *ist linear und wird durch die Matrix* $\boldsymbol{A}^+$ *beschrieben.* Der folgende Satz zeigt, daß $\boldsymbol{A}^+$ auch in völlig anderer Weise eingeführt werden kann.

8.1.8. *Satz.*

(i) Zu jeder Matrix $\boldsymbol{A} \in \mathbf{R}^{m,n}$ gibt es genau eine Matrix $\boldsymbol{A}^+ \in \mathbf{R}^{n,m}$, so daß

$$\boldsymbol{x}^N = \boldsymbol{A}^+\boldsymbol{b} \quad \text{für alle } \boldsymbol{b} \in \mathbf{R}^m$$

die Normallösung von $\boldsymbol{Ax} \cong \boldsymbol{b}$ darstellt. Wenn (2) die Singulärwertzerlegung von $\boldsymbol{A}$ bezeichnet, ist $\boldsymbol{A}^+$ durch (35) gegeben.

(ii) $\boldsymbol{A}^+$ ist durch die Bedingungen

$$(\mathrm{P}_1)\quad \boldsymbol{AA}^+\boldsymbol{A} = \boldsymbol{A}, \qquad (\mathrm{P}_2)\quad \boldsymbol{A}^+\boldsymbol{AA}^+ = \boldsymbol{A}^+,$$

$$(\mathrm{P}_3)\quad (\boldsymbol{AA}^+)^\mathsf{T} = \boldsymbol{AA}^+, \qquad (\mathrm{P}_4)\quad (\boldsymbol{A}^+\boldsymbol{A})^\mathsf{T} = \boldsymbol{A}^+\boldsymbol{A}$$

eindeutig festgelegt.

Beweis. Die Eindeutigkeit von $\boldsymbol{A}^+$ ergibt sich aus der Eindeutigkeit der Normallösung. Es bleibt zu zeigen, daß $\boldsymbol{A}^+$ äquivalent durch $(\mathrm{P}_1), \ldots, (\mathrm{P}_4)$ charakterisiert werden kann. Daß $\boldsymbol{A}^+$ aus (35) diese Bedingungen erfüllt, läßt sich sofort durch Einsetzen überprüfen. Es sei umgekehrt jetzt $\boldsymbol{B} \in \mathbf{R}^{n,m}$ eine Matrix, die den Bedingungen aus (ii) mit $\boldsymbol{B}$ anstelle von $\boldsymbol{A}^+$ genügt. Dann folgt

$$\boldsymbol{B} \underset{2}{=} \boldsymbol{BAB} \underset{1}{=} \boldsymbol{BAA}^+\boldsymbol{AB} \underset{4}{=} \boldsymbol{A}^\mathsf{T}\boldsymbol{B}^\mathsf{T}\boldsymbol{A}^+\boldsymbol{A} = \boldsymbol{A}^\mathsf{T}\boldsymbol{B}^\mathsf{T}\boldsymbol{A}^\mathsf{T}(\boldsymbol{A}^+)^\mathsf{T}\boldsymbol{B} \underset{1}{=} \boldsymbol{A}^\mathsf{T}(\boldsymbol{A}^+)^\mathsf{T}\boldsymbol{B} = \boldsymbol{A}^+\boldsymbol{AB}$$

$$= \boldsymbol{A}^+\boldsymbol{B}^\mathsf{T}\boldsymbol{A}^\mathsf{T} \underset{3}{=} \boldsymbol{A}^+\boldsymbol{AA}^+\boldsymbol{B}^\mathsf{T}\boldsymbol{A}^\mathsf{T} \underset{2}{=} \boldsymbol{A}^+(\boldsymbol{A}^+)^\mathsf{T}\boldsymbol{A}^\mathsf{T}\boldsymbol{B}^\mathsf{T}\boldsymbol{A}^\mathsf{T} \underset{3}{=} \boldsymbol{A}^+(\boldsymbol{A}^+)^\mathsf{T}\boldsymbol{A}^\mathsf{T} \underset{1}{=} \boldsymbol{A}^+\boldsymbol{AA}^+ \underset{2}{=} \boldsymbol{A}^+.$$

Die Zahl i unter dem Gleichheitszeichen gibt dabei an, welche Bedingung (P_i) für $\boldsymbol{A}^+$ bzw. $\boldsymbol{B}$ benutzt worden ist. □

Die Matrix $\boldsymbol{A}$ heißt *Moore-Penrose-Inverse* oder kurz *Pseudoinverse* von $\boldsymbol{A}$; man spricht auch von der *verallgemeinerten Inversen*. Die Bedingungen $(\mathrm{P}_1), \ldots, (\mathrm{P}_4)$ werden *Penrose-Bedingungen* genannt. Zur Geschichte der verallgemeinerten Inversen siehe B 8.3.

Durch Einsetzen überzeugt man sich von der Gültigkeit der Rechenregeln

$$\boldsymbol{O}^+ = \boldsymbol{O}, \qquad (\boldsymbol{A}^+)^\mathsf{T} = (\boldsymbol{A}^\mathsf{T})^+ \quad \text{und} \quad (\lambda\boldsymbol{A})^+ = (1/\lambda)\,\boldsymbol{A}^+ \quad \text{für } \lambda \in \mathbf{R},\ \lambda \neq 0. \tag{37}$$

Weitere nützliche Eigenschaften der Pseudoinversen, deren Richtigkeit ebenfalls durch Einsetzen in die Penrose-Bedingungen bestätigt werden kann, sind in der folgenden Aussage zusammengefaßt.

8.1.9. *Aussage.* Es sei $\boldsymbol{A} \in \mathbf{R}^{m,n}$ eine Matrix vom Rang r. Dann gilt

(i)
$$\boldsymbol{A}^+ = \begin{cases} (\boldsymbol{A}^\mathsf{T}\boldsymbol{A})^{-1}\boldsymbol{A}^\mathsf{T}, & \text{falls } n = r \leqq m, \\ \boldsymbol{A}^\mathsf{T}(\boldsymbol{AA}^\mathsf{T})^{-1}, & \text{falls } m = r \leqq n, \\ \boldsymbol{A}^{-1}, & \text{falls } m = r = n. \end{cases}$$

(ii) $\boldsymbol{A}^+ = \boldsymbol{G}^+\boldsymbol{F}^+$, falls $\boldsymbol{A} = \boldsymbol{FG}$ mit $\boldsymbol{F} \in \mathbf{R}^{m,r}$, $\boldsymbol{G} \in \mathbf{R}^{r,n}$ (und folglich rang $(\boldsymbol{F})$ = rang $(\boldsymbol{G})$ = r).

(iii) $\boldsymbol{A}^+ = \boldsymbol{VB}^+\boldsymbol{U}^\mathsf{T}$, falls $\boldsymbol{A} = \boldsymbol{UBV}^\mathsf{T}$ mit $\boldsymbol{B} \in \mathbf{R}^{m,n}$ und orthogonalen Matrizen $\boldsymbol{U} \in \mathbf{R}^{m,m}$, $\boldsymbol{V} \in \mathbf{R}^{n,n}$.

(iv) $\boldsymbol{A}^+ = (\underbrace{\boldsymbol{B}^\top}_{k} \mid \underbrace{\boldsymbol{O}}_{m-k})\}^n$, falls $\boldsymbol{A} = \underbrace{\left(\begin{matrix}\boldsymbol{B}\\ \hline \boldsymbol{O}\end{matrix}\right)}_{n}\begin{matrix}\}k\\ \}m-k\end{matrix}$.

(v) $[\operatorname{diag}(\sigma_1, \ldots, \sigma_l)]^+ = \operatorname{diag}(\sigma_1^+, \ldots, \sigma_l^+)$, wobei

$$\sigma^+ := \begin{cases} 1/\sigma & \text{für } \sigma \neq 0, \\ 0 & \text{für } \sigma = 0, \end{cases} \qquad \sigma \in \mathbf{R}.$$

Schließlich lassen sich die mit $\boldsymbol{A}$ verbundenen Projektoren explizit durch $\boldsymbol{A}$ und $\boldsymbol{A}^+$ ausdrücken.

8.1.10. *Aussage.* Es sei $\boldsymbol{A} \in \mathbf{R}^{m,n}$ gegeben. Dann ist

$$\boldsymbol{P}_{\boldsymbol{A}} = \boldsymbol{A}\boldsymbol{A}^- \quad \text{bzw.} \quad \boldsymbol{P}_{\boldsymbol{A}^\top} = \boldsymbol{A}^+\boldsymbol{A}$$

der Projektor auf $\mathcal{R}(\boldsymbol{A})$ bzw. $\mathcal{R}(\boldsymbol{A}^\top)$, und

$$\boldsymbol{Q}_{\boldsymbol{A}} = \boldsymbol{I} - \boldsymbol{A}\boldsymbol{A}^+ \quad \text{bzw.} \quad \boldsymbol{Q}_{\boldsymbol{A}^\top} = \boldsymbol{I} - \boldsymbol{A}^+\boldsymbol{A}$$

stellt den Projektor auf $\mathcal{N}(\boldsymbol{A}^\top)$ bzw. $\mathcal{N}(\boldsymbol{A})$ dar.

Beweis. Wir setzen $\boldsymbol{P} := \boldsymbol{P}_{\boldsymbol{A}} = \boldsymbol{A}\boldsymbol{A}^+$. Wegen (P_3) und (P_1) gilt (23), d. h., $\boldsymbol{P}$ ist ein Projektor. Es bleibt $\mathcal{R}(\boldsymbol{P}) = \mathcal{R}(\boldsymbol{A})$ zu zeigen. Für jedes $\boldsymbol{y}$ gilt nun $\boldsymbol{P}\boldsymbol{y} = \boldsymbol{A}\boldsymbol{A}^+\boldsymbol{y}$, also $\mathcal{R}(\boldsymbol{P}) \subset \mathcal{R}(\boldsymbol{A})$. Andererseits ist wegen (P_1) $\boldsymbol{A}\boldsymbol{x} = \boldsymbol{A}\boldsymbol{A}^+\boldsymbol{A}\boldsymbol{x} = \boldsymbol{P}\boldsymbol{A}\boldsymbol{x}$, also $\mathcal{R}(\boldsymbol{A}) \subset \mathcal{R}(\boldsymbol{P})$ und damit $\mathcal{R}(\boldsymbol{P}) = \mathcal{R}(\boldsymbol{A})$. Die übrigen Aussagen ergeben sich aus 8.1.4 und der Tatsache, daß $\mathcal{R}(\boldsymbol{A})^\perp = \mathcal{N}(\boldsymbol{A}^\top)$ sowie $\mathcal{R}(\boldsymbol{A}^\top)^\perp = \mathcal{N}(\boldsymbol{A})$ gilt; letztere Eigenschaft folgt bei Anwendung von 8.1.5 auf $\boldsymbol{A}^\top$. □

Übungsaufgaben

Ü 8.1.1. Man zeige: Wenn $\{\boldsymbol{q}^1, \ldots, \boldsymbol{q}^r\}$ eine orthonormierte Basis von $\mathcal{Y}$ ist, hat der orthonormale Projektor $\boldsymbol{P}$ auf $\mathcal{Y}$ die Gestalt

$$\boldsymbol{P} = \sum_{i=1}^{r} \boldsymbol{q}^i \boldsymbol{q}^{i\top}. \tag{38}$$

Der Projektor $\boldsymbol{Q} = \boldsymbol{I} - \boldsymbol{P}$ kann in der Produktform

$$\boldsymbol{Q} = \boldsymbol{I} - \boldsymbol{P} = [\boldsymbol{I} - \boldsymbol{q}^1\boldsymbol{q}^{1\top}][\boldsymbol{I} - \boldsymbol{q}^2\boldsymbol{q}^{2\top}] \cdots [\boldsymbol{I} - \boldsymbol{q}^r\boldsymbol{q}^{r\top}] \tag{39}$$

geschrieben werden.

Ü 8.1.2. Man beweise, daß $\boldsymbol{P} \in \mathbf{R}^{m,m}$ genau dann ein orthogonaler Projektor ist, wenn

$$\boldsymbol{P} = \boldsymbol{U}\boldsymbol{\Pi}\boldsymbol{U}^\top \tag{40}$$

mit orthogonalem $\boldsymbol{U} \in \mathbf{R}^{m,m}$ und einer Diagonalmatrix $\boldsymbol{\Pi} = \operatorname{diag}(\Pi_i)$, $\Pi_i \in \{0, 1\}$, gilt, d. h., Projektoren sind symmetrische Matrizen mit den Eigenwerten 0 oder 1. Man gebe eine orthonormierte Basis von $\mathcal{R}(\boldsymbol{P})$ an.

Ü 8.1.3. Es sei $\mathcal{Y} \subset \mathbf{R}^m$ ein Teilraum und $\mathcal{Z} = \mathcal{Y}^\perp$. Man beweise, daß $\boldsymbol{P} \in \mathbf{R}^{m,m}$ genau dann der orthonormale Projektor auf $\mathcal{Y}$ ist, wenn

$$\boldsymbol{P}\boldsymbol{y} = \boldsymbol{y} \text{ für alle } \boldsymbol{y} \in \mathcal{Y} \quad \text{und} \quad \boldsymbol{P}\boldsymbol{z} = \boldsymbol{o} \text{ für alle } \boldsymbol{z} \in \mathcal{Z} \tag{41}$$

gilt.

Hinweis: Es gilt $\boldsymbol{P} = \boldsymbol{P}^\top$ genau dann, wenn $\boldsymbol{u}^\top(\boldsymbol{P}\boldsymbol{v}) = (\boldsymbol{P}\boldsymbol{u})^\top\boldsymbol{v}$ ist für alle $\boldsymbol{u}, \boldsymbol{v} \in \mathbf{R}^m$.

Ü 8.1.4. Man zeige: Für jeden orthogonalen Projektor $\boldsymbol{P}$ gilt

$$\|\boldsymbol{P}\| = \begin{cases} 1, & \text{falls } \boldsymbol{P} \neq \boldsymbol{O}, \\ 0, & \text{falls } \boldsymbol{P} = \boldsymbol{O}, \end{cases} \tag{42}$$

sowie

$$\|\boldsymbol{P}\|_F = \sqrt{r}, \qquad r = \operatorname{rang}(\boldsymbol{P}) = \dim \mathcal{R}(\boldsymbol{P}). \tag{43}$$

Ü 8.1.5. Man drücke die Projektoren $\boldsymbol{P}_A$, $\boldsymbol{Q}_A$, $\boldsymbol{P}_{A^T}$, $\boldsymbol{Q}_{A^T}$ aus 8.1.10 unter Verwendung der Singulärwertzerlegung durch $\boldsymbol{U}$, $\boldsymbol{V}$ und $\boldsymbol{\Sigma}$ aus.

Ü 8.1.6. Man beweise

$$\|\boldsymbol{A}^+\| = \|\boldsymbol{\Sigma}^+\| = 1/\sigma_r. \tag{44}$$

Ü 8.1.7. Man zeige: Wenn $\boldsymbol{v} \in \mathbf{R}^m$ als $(m, 1)$-Matrix aufgefaßt wird, gilt im Fall $\boldsymbol{v} \neq \boldsymbol{o}$

$$\boldsymbol{v}^+ = \boldsymbol{v}^T/\|\boldsymbol{v}\|^2 \quad \text{sowie} \quad \boldsymbol{P}_{\boldsymbol{v}} = \boldsymbol{v}\boldsymbol{v}^T/\|\boldsymbol{v}\|^2, \tag{45}$$

wobei $\boldsymbol{P}_{\boldsymbol{v}}$ den Projektor auf $\mathcal{R}(\boldsymbol{v}) = \operatorname{span}\{\boldsymbol{v}\}$ bezeichnet.

Ü 8.1.8. Wie lautet $(\boldsymbol{a}\boldsymbol{b}^T)^+$ für $\boldsymbol{a}, \boldsymbol{b} \in \mathbf{R}^m$, $\boldsymbol{a} \neq \boldsymbol{o}$, $\boldsymbol{b} \neq \boldsymbol{o}$?

Ü 8.1.9. Man zeige, daß die Vektoren $\boldsymbol{P}_A\boldsymbol{b}$, $\boldsymbol{A}^-\boldsymbol{b}$ und $\boldsymbol{A}^T\boldsymbol{b}$ entweder alle verschwinden oder alle von 0 verschieden sind.

Ü 8.1.10. Man beweise: Unter allen Lösungen $\boldsymbol{x} \in \mathcal{L}$ von $\boldsymbol{A}\boldsymbol{x} \cong \boldsymbol{b}$ hat

$$\bar{\boldsymbol{x}} := \boldsymbol{x}^0 + \boldsymbol{A}^+(\boldsymbol{b} - \boldsymbol{A}\boldsymbol{x}^0) = (\boldsymbol{I} - \boldsymbol{A}^+\boldsymbol{A})\,\boldsymbol{x}^0 + \boldsymbol{A}^+\boldsymbol{b}$$

den kleinsten Euklidischen Abstand von $\boldsymbol{x}^0$. Skizze analog zu Abb. 8.1.3!
Hinweis: Man führe $\boldsymbol{y} := \boldsymbol{x} - \boldsymbol{x}^0$ als neue Variable ein.

8.2. Störungstheorie

Wir betrachten das lineare Quadratmittelproblem

$$\boldsymbol{A}\boldsymbol{x} \cong \boldsymbol{b}, \tag{1}$$

dessen eindeutige Normallösung durch

$$\boldsymbol{x} = \boldsymbol{x}^N = \boldsymbol{A}^+\boldsymbol{b} \tag{2}$$

gegeben ist. Im Fall einer spaltenregulären Matrix stellt (2) die einzige Lösung von (1) dar, andernfalls diejenige kleinster Euklidischer Norm.

Uns interessiert, wie sich Störungen $\delta\boldsymbol{A}$, $\delta\boldsymbol{b}$ der Eingangsdaten auf die Normallösung auswirken. Das gestörte Problem

$$(\boldsymbol{A} + \delta\boldsymbol{A})(\boldsymbol{x} + \delta\boldsymbol{x}) \cong \boldsymbol{b} + \delta\boldsymbol{b} \tag{3}$$

besitzt die Normallösung

$$\boldsymbol{x} + \delta\boldsymbol{x} = (\boldsymbol{A} + \delta\boldsymbol{A})^+(\boldsymbol{b} + \delta\boldsymbol{b}), \tag{4}$$

und die Änderung δx gegenüber der Normallösung x des ungestörten Problems ist

$$\delta x = [(A + \delta A)^+ - A^+]\, b + (A + \delta A)^+ \delta b. \tag{5}$$

Wir untersuchen daher zunächst, wie sich A^+ gegenüber Störungen von A verhält.

A. *Störungstheorie der Pseudoinversen*

Um zu zeigen, wie sich Störungen von A in unterschiedlicher Weise auf A^+ auswirken können, betrachten wir zwei einfache Beispiele.

8.2.1. *Beispiel.* Die Matrix $A \neq O$ sei durch ihre Singulärwertzerlegung

$$A = U \Sigma V^T$$

mit $\Sigma = \operatorname{diag}(\sigma_1, \ldots, \sigma_r, 0, \ldots, 0)$, $\sigma_1 \geqq \cdots \geqq \sigma_r > 0$, $r = \operatorname{rang}(A)$ und orthogonalem U, V gegeben. Die Störung δA werde in der Form

$$\delta A = U \delta \Sigma V^T$$

geschrieben, so daß die gestörte Matrix $A + \delta A$ die Gestalt

$$A + \delta A = U(\Sigma + \delta \Sigma)\, V^T$$

besitzt.

Für die durch

$$(\delta \Sigma)_{ij} = \begin{cases} \varepsilon & \text{für } i = j = r, \\ 0 & \text{sonst} \end{cases} \tag{6}$$

definierte spezielle Störung $\delta \Sigma$ ist

$$\Sigma = \left(\begin{array}{ccc|c} \sigma_1 & & & \\ & \ddots & & \\ & & \sigma_r & 0 \\ \hline & 0 & & 0 \end{array}\right), \quad \Sigma + \delta \Sigma = \left(\begin{array}{ccc|c} \sigma_1 & & & \\ & \ddots & & \\ & & \sigma_r + \varepsilon & 0 \\ \hline & 0 & & 0 \end{array}\right).$$

Im Fall

$$\varkappa := \|A^+\|\, \|\delta A\| = |\varepsilon| / \sigma_r < 1$$

— man beachte $\|\delta A\| = \|\delta \Sigma\| = |\varepsilon|$ und $\|A^+\| = 1/\sigma_r$ — folgt

$$\sigma_r + \varepsilon \geqq \sigma_r - |\varepsilon| = \sigma_r(1 - \varkappa) > 0, \tag{7}$$

d. h., $\sigma_r + \varepsilon$ bleibt positiv, und es gilt

$$\operatorname{rang}(A + \delta A) = \operatorname{rang}(\Sigma + \delta \Sigma) = \operatorname{rang}(\Sigma) = \operatorname{rang}(A).$$

Die Pseudoinversen sind dann durch

$$A^+ = V \Sigma^+ U^T \quad \text{bzw.} \quad (A + \delta A)^+ = V(\Sigma + \delta \Sigma)^+ U^T$$

mit

$$\Sigma^+ = \left(\begin{array}{ccc|c} \frac{1}{\sigma_1} & & & \\ & \ddots & & \\ & & \frac{1}{\sigma_{r-1}} & \\ & & & \frac{1}{\sigma_r} \\ \hline & 0 & & 0 \end{array}\right) \quad \text{bzw.} \quad (\Sigma + \delta\Sigma)^+ = \left(\begin{array}{ccc|c} \frac{1}{\sigma_1} & & & \\ & \ddots & & \\ & & \frac{1}{\sigma_{r-1}} & \\ & & & \frac{1}{\sigma_r + \varepsilon} \\ \hline & 0 & & 0 \end{array}\right)$$

gegeben, vgl. (8.1.35). Wegen (7) gilt

$$\|(\Sigma + \delta\Sigma)^+\| = \max\left\{\frac{1}{\sigma_1}, \frac{1}{\sigma_{r-1}}, \frac{1}{\sigma_r + \varepsilon}\right\} \leqq \frac{1}{\sigma_r - |\varepsilon|} = \frac{1}{\sigma_r(1 - \varkappa)},$$

also

$$\|(A + \delta A)^+\| = \|(\Sigma + \delta\Sigma)^+\| \leqq \|A^+\|/(1 - \varkappa). \tag{8}$$

Analog folgt

$$\|(\Sigma + \delta\Sigma)^+ - \Sigma^+\| = \left|\frac{1}{\sigma_r + \varepsilon} - \frac{1}{\sigma_r}\right| = \frac{|\varepsilon|}{\sigma_r(\sigma_r + \varepsilon)} \leqq \frac{|\varepsilon|}{\sigma_r^2(1 - \varkappa)},$$

d. h.

$$\|(A + \delta A)^+ - A^+\| \leqq [\|A^+\|^2/(1 - \varkappa)]\, \|\delta A\|. \quad \square \tag{9}$$

Die Abschätzungen (8), (9) entsprechen wörtlich den Abschätzungen (4.1.9), (4.1.10) für die Inversen gestörter regulärer Matrizen. Wie bei der Matrixinversion *liegt* hier also *bei der Störung* (6), *die den Rang von* A *nicht erhöht, lokale Lipschitzstetigkeit vor; insbesondere bleiben die Pseudoinversen beschränkt.*

Das zweite Beispiel zeigt, daß auch bösartigere Störungen δA existieren.

8.2.2. *Beispiel.* Es sei A wie in Beispiel 8.2.1, und zusätzlich gelte

$$r = \operatorname{rang}(A) < l := \min(m, n),$$

d. h., A sei rangdefizient. Die Störung $\delta\Sigma$ sei durch

$$(\delta\Sigma)_{ij} = \begin{cases} \varepsilon & \text{für} \quad i = j = r + 1, \\ 0 & \text{sonst} \end{cases} \tag{10}$$

festgelegt. Im Fall $|\varepsilon| = \|\delta\Sigma\| = \|\delta A\| \neq 0$ ergibt sich dann

$$\Sigma + \delta\Sigma = \left(\begin{array}{ccc|c} \sigma_1 & & & \\ & \ddots & & \\ & & \sigma_r & \\ & & & \varepsilon \\ \hline & 0 & & 0 \end{array}\right), \qquad (\Sigma + \delta\Sigma)^+ = \left(\begin{array}{ccc|c} \frac{1}{\sigma_1} & & & \\ & \ddots & & \\ & & \frac{1}{\sigma_r} & \\ & & & \frac{1}{\varepsilon} \\ \hline & 0 & & 0 \end{array}\right),$$

also

$$\operatorname{rang}(\boldsymbol{A}+\delta\boldsymbol{A}) = \operatorname{rang}(\boldsymbol{\Sigma}+\delta\boldsymbol{\Sigma}) = r+1 > r = \operatorname{rang}(\boldsymbol{\Sigma}) = \operatorname{rang}(\boldsymbol{A}),$$

und es folgt

$$\|(\boldsymbol{\Sigma}+\delta\boldsymbol{\Sigma})^+\| = \max\{1/\sigma_r, 1/|\varepsilon|\} \geqq 1/|\varepsilon|, \quad \|(\boldsymbol{\Sigma}+\delta\boldsymbol{\Sigma})^+ - \boldsymbol{\Sigma}^+\| = 1/|\varepsilon|,$$

also

$$\|(\boldsymbol{A}+\delta\boldsymbol{A})^+\| \geqq 1/\|\delta\boldsymbol{A}\|, \qquad \|(\boldsymbol{A}+\delta\boldsymbol{A})^+ - \boldsymbol{A}^+\| \geqq 1/\|\delta\boldsymbol{A}\|. \quad \square$$

Gegenüber der rangerhöhenden Störung (10) *ist die Pseudoinverse also unstetig und unbeschränkt, und zwar um so stärker, je kleiner die Störung ist. Es liegt also ein inkorrekt gestelltes Problem mit extrem schlechtem Unstetigkeitsverhalten vor*, vgl. Abschnitt 2.1.D und das Modellbeispiel 2.1.11.

Im folgenden werden wir zeigen, daß es genau die in den beiden Beispielen angegebenen Klassen von gutartigen und bösartigen Störungen gibt. Die gutartigen Störungen sind durch die Bedingung $\operatorname{rang}(\boldsymbol{A}+\delta\boldsymbol{A}) \leqq \operatorname{rang}(\boldsymbol{A})$ charakterisiert und führen zu lokal lipschitzstetiger Abhängigkeit. Die bösartigen Störungen sind diejenigen mit $\operatorname{rang}(\boldsymbol{A}+\delta\boldsymbol{A}) > \operatorname{rang}(\boldsymbol{A})$. Gegenüber solchen Störungen liegt Unstetigkeit vor, und die Pseudoinversen sind nicht beschränkt.

Als Hilfsmittel benötigen wir Aussagen über das Störungsverhalten der Singulärwerte.

8.2.3. *Satz.* Es gelte $\boldsymbol{A}, \delta\boldsymbol{A} \in \mathbf{R}^{m,n}$, und σ_i bzw. $\sigma_i + \delta\sigma_i$ seien die gemäß

$$\sigma_1 \geqq \sigma_2 \geqq \cdots \geqq \sigma_l$$

bzw.

$$\sigma_1 + \delta\sigma_1 \geqq \sigma_2 + \delta\sigma_2 \geqq \cdots \geqq \sigma_l + \delta\sigma_l$$

geordneten Singulärwerte von $\boldsymbol{A}$ bzw. $\boldsymbol{A}+\delta\boldsymbol{A}$, wobei $l := \min(m, n)$ ist. Dann gelten die Abschätzungen

$$|\delta\sigma|_i \leqq \|\delta\boldsymbol{A}\| \tag{11}$$

sowie

$$\left\{\sum_{i=1}^{l}(\delta\sigma_i)^2\right\}^{1/2} \leqq \|\delta\boldsymbol{A}\|_F. \tag{12}$$

Beweis. Wir führen die Matrizen

$$\boldsymbol{B} := \left(\begin{array}{c|c} \boldsymbol{O} & \boldsymbol{A} \\ \hline \boldsymbol{A}^T & \boldsymbol{O} \end{array}\right) \quad \text{und} \quad \delta\boldsymbol{B} := \left(\begin{array}{c|c} \boldsymbol{O} & \delta\boldsymbol{A} \\ \hline \delta\boldsymbol{A}^T & \boldsymbol{O} \end{array}\right)$$

ein. Unter Verwendung der Singulärwertzerlegung von $\boldsymbol{A}$ folgt

$$\boldsymbol{B} = \left(\begin{array}{c|c} \boldsymbol{O} & \boldsymbol{U}\boldsymbol{\Sigma}\boldsymbol{V}^T \\ \hline \boldsymbol{V}\boldsymbol{\Sigma}^T\boldsymbol{U}^T & \boldsymbol{O} \end{array}\right) = \boldsymbol{S}\boldsymbol{C}\boldsymbol{S}^T \quad \text{mit} \quad \boldsymbol{C} := \left(\begin{array}{c|c} \boldsymbol{O} & \boldsymbol{\Sigma} \\ \hline \boldsymbol{\Sigma}^T & \boldsymbol{O} \end{array}\right), \quad \boldsymbol{S} := \left(\begin{array}{c|c} \boldsymbol{U} & \boldsymbol{O} \\ \hline \boldsymbol{O} & \boldsymbol{V} \end{array}\right).$$

Wegen der Orthogonalität von $\boldsymbol{S}$ sind $\boldsymbol{C}$ und $\boldsymbol{B}$ ähnlich, besitzen also dieselben Eigenwerte.

Die Eigenwertgleichung $\boldsymbol{C}\boldsymbol{z} = \lambda \boldsymbol{z}$ läßt sich mit

$$\boldsymbol{z} = \begin{pmatrix} \boldsymbol{y} \\ \boldsymbol{x} \end{pmatrix} \quad \text{als} \quad \begin{pmatrix} \boldsymbol{\Sigma}\boldsymbol{x} \\ \boldsymbol{\Sigma}^{\mathrm{T}}\boldsymbol{y} \end{pmatrix} = \lambda \begin{pmatrix} \boldsymbol{y} \\ \boldsymbol{x} \end{pmatrix}$$

schreiben, d. h., sie zerfällt in die $|m - n|$ Gleichungen

$$0 = \lambda y_i \quad (i = l + 1, \ldots, m) \quad \text{oder} \quad 0 = \lambda x_i \quad (i = l + 1, \ldots, n)$$

und die l zweidimensionalen Eigenwertprobleme

$$\begin{pmatrix} 0 & \sigma_i \\ \sigma_i & 0 \end{pmatrix} \begin{pmatrix} y_i \\ x_i \end{pmatrix} \lambda = \begin{pmatrix} y_i \\ x_i \end{pmatrix} \quad (i = 1, \ldots, l),$$

von denen jedes die Eigenwerte $\pm\sigma_i$ besitzt. $\boldsymbol{B}$ hat also die $2l$ Eigenwerte $\pm\sigma_i$, ergänzt durch $|m - n|$ Nullen, und analoge Aussagen gelten mit $\pm(\sigma_i + \delta\sigma_i)$ für $\boldsymbol{B} + \boldsymbol{\delta B}$. Nun gilt $\|\boldsymbol{B}\| = \|\boldsymbol{A}\|$, $\|\boldsymbol{B}\|_{\boldsymbol{F}} = \sqrt{2}\,\|\boldsymbol{A}\|_{\boldsymbol{F}}$, und analoge Beziehungen gelten für $\boldsymbol{\delta B}$ und $\boldsymbol{\delta A}$. Damit ergeben sich (11) und (12) direkt aus 13.1.B. □

Satz 8.2.3 besagt: *Die Singulärwerte sind lipschitzstetige Funktionen der Matrix, und die Lipschitzkonstante hat den Wert* 1. Es liegt also ein besonders gutartiges Stetigkeitsverhalten vor.

Wir können jetzt die angekündigte Aussage über rangerhöhende Störungen formulieren.

8.2.4. *Satz.* Es sei $\boldsymbol{A} \in \mathbf{R}^{m,n}$ eine rangdefiziente Matrix und $\boldsymbol{\delta A} \neq \boldsymbol{O}$ eine Störung mit

$$\operatorname{rang}(\boldsymbol{A} + \boldsymbol{\delta A}) > \operatorname{rang}(\boldsymbol{A}). \tag{13}$$

Dann gilt

$$\|(\boldsymbol{A} + \boldsymbol{\delta A})^{+}\| \geqq 1/\|\boldsymbol{\delta A}\|. \tag{14}$$

Beweis. Es sei $r = \operatorname{rang}(\boldsymbol{A})$, und σ_i bzw. $\tilde{\sigma}_i = \sigma_i + \delta\sigma_i$ seien die wie in 8.2.3 geordneten Singulärwerte von $\boldsymbol{A}$ bzw. $\boldsymbol{A} + \boldsymbol{\delta A}$. Wegen (13) ist $\operatorname{rang}(\boldsymbol{A} + \boldsymbol{\delta A}) \geqq r + 1$, also $\tilde{\sigma}_{r+1} = \sigma_{r+1} + \delta\sigma_{r+1} = \delta\sigma_{r+1} > 0$. Nach (11) ist andererseits $|\delta\sigma_{r+1}| \leqq \|\boldsymbol{\delta A}\|$, womit sich sofort

$$\|(\boldsymbol{A} + \boldsymbol{\delta A})^{+}\| = \max\{1/\tilde{\sigma}_i : \tilde{\sigma}_i > 0\} \geqq 1/\tilde{\sigma}_{r+1} \geqq 1/\|\boldsymbol{\delta A}\|$$

ergibt. □

Aus (14) folgt

$$\|(\boldsymbol{A} + \boldsymbol{\delta A})^{+} - \boldsymbol{A}^{+}\| \geqq \|(\boldsymbol{A} + \boldsymbol{\delta A})^{+}\| - \|\boldsymbol{A}^{+}\| \geqq (1/\|\boldsymbol{\delta A}\|) - \|\boldsymbol{A}^{+}\|, \tag{15}$$

d. h., $(\boldsymbol{A} + \boldsymbol{\delta A})^{+}$ weicht von $\boldsymbol{A}^{+}$ beliebig viel ab, wenn $\boldsymbol{\delta A}$ genügend klein ist.

Wir zeigen jetzt, daß Störungen, die den Rang nicht erhöhen, stets gutartig sind.

8.2.5. *Satz.* Es sei $\boldsymbol{A} \in \mathbf{R}^{m,n}$ gegeben, und die Störung $\boldsymbol{\delta A} \in \mathbf{R}^{m,n}$ genüge den Bedingungen

$$\operatorname{rang}(\boldsymbol{A} + \boldsymbol{\delta A}) \leqq \operatorname{rang}(\boldsymbol{A}) \tag{16}$$

und

$$\varkappa := \|\boldsymbol{A}^{+}\|\,\|\boldsymbol{\delta A}\| < 1. \tag{17}$$

Dann ist

$$\operatorname{rang}(\boldsymbol{A}+\boldsymbol{\delta A}) = \operatorname{rang}(\boldsymbol{A}), \tag{18}$$

und es gilt

$$\|(\boldsymbol{A}+\boldsymbol{\delta A})^{+}\| \leqq \|\boldsymbol{A}^{+}\|/(1-\varkappa). \tag{19}$$

Die Änderung $(\boldsymbol{A}+\boldsymbol{\delta A})^{+}-\boldsymbol{A}^{+}$ ist in erster Ordnung durch

$$(\boldsymbol{A}+\boldsymbol{\delta A})^{+}-\boldsymbol{A}^{+} = -\boldsymbol{A}^{+}\boldsymbol{\delta A}\boldsymbol{A}^{+}+\boldsymbol{A}^{+}\boldsymbol{A}^{+\mathsf{T}}\boldsymbol{\delta A}^{\mathsf{T}}\boldsymbol{Q}_{\boldsymbol{A}}+\boldsymbol{Q}_{\boldsymbol{A}^{\mathsf{T}}}\boldsymbol{\delta A}^{\mathsf{T}}\boldsymbol{A}^{+\mathsf{T}}\boldsymbol{A}^{+} + \boldsymbol{O}(\|\boldsymbol{\delta A}\|^{2}) \tag{20}$$

mit

$$\boldsymbol{Q}_{\boldsymbol{A}} := \boldsymbol{I}-\boldsymbol{A}\boldsymbol{A}^{+} \quad \text{und} \quad \boldsymbol{Q}_{\boldsymbol{A}^{\mathsf{T}}} := \boldsymbol{I}-\boldsymbol{A}^{+}\boldsymbol{A} \tag{21}$$

gegeben und genügt der Abschätzung

$$\|(\boldsymbol{A}+\boldsymbol{\delta A})^{+}-\boldsymbol{A}^{+}\| \leqq \mu[\|\boldsymbol{A}^{+}\|^{2}/(1-\varkappa)]\,\|\boldsymbol{\delta A}\|, \tag{22}$$

wobei

$$\mu := \begin{cases} (1+\sqrt{5})/2 = 1.62, & \text{falls} \quad \operatorname{rang}(\boldsymbol{A}) < \min(m,n), \\ \sqrt{2} = 1.41, & \text{falls} \quad \operatorname{rang}(\boldsymbol{A}) = \min(m,n) < \max(m,n), \\ 1, & \text{falls} \quad \operatorname{rang}(\boldsymbol{A}) = m = n. \end{cases} \tag{23}$$

Beweis. Es sei $r = \operatorname{rang}(\boldsymbol{A})$, und σ_i bzw. $\tilde{\sigma}_i = \sigma_i + \delta\sigma_i$ seien die geordneten Singulärwerte von $\boldsymbol{A}$ bzw. $\boldsymbol{A}+\boldsymbol{\delta A}$. Wegen (11) und (17) gilt

$$\tilde{\sigma}_i = \sigma_i + \delta\sigma_i \geqq \sigma_i - |\delta\sigma_i| \geqq \sigma_r - \|\boldsymbol{\delta A}\| = (1-\varkappa)/\|\boldsymbol{A}^{+}\| > 0 \qquad (i = 1, \ldots, r),$$

also $\operatorname{rang}(\boldsymbol{A}+\boldsymbol{\delta A}) \geqq r$. Mit (16) folgt dann $\tilde{\sigma}_i = 0$ $(i = r+1, \ldots, \min(m,n))$, mithin (18) und (19). Als nächstes untersuchen wir die Änderung

$$\boldsymbol{G} := \tilde{\boldsymbol{A}}^{+}-\boldsymbol{A}^{+}, \qquad \tilde{\boldsymbol{A}}^{+} := (\boldsymbol{A}+\boldsymbol{\delta A})^{+}.$$

Mit den Projektoren

$$\boldsymbol{P} := \boldsymbol{P}_{\boldsymbol{A}} = \boldsymbol{A}\boldsymbol{A}^{+}, \qquad \boldsymbol{S} := \boldsymbol{P}_{\boldsymbol{A}^{\mathsf{T}}} = \boldsymbol{A}^{+}\boldsymbol{A}, \qquad \tilde{\boldsymbol{P}} := \tilde{\boldsymbol{A}}\tilde{\boldsymbol{A}}^{+}, \qquad \tilde{\boldsymbol{S}} := \tilde{\boldsymbol{A}}^{+}\tilde{\boldsymbol{A}}$$

ergibt sich

$$\begin{aligned} \boldsymbol{G} &= [\tilde{\boldsymbol{S}}+(\boldsymbol{I}-\tilde{\boldsymbol{S}})]\,(\tilde{\boldsymbol{A}}^{+}-\boldsymbol{A}^{+})\,[\boldsymbol{P}+(\boldsymbol{I}-\boldsymbol{P})] \\ &= \tilde{\boldsymbol{S}}\tilde{\boldsymbol{A}}^{+}\boldsymbol{P}+\tilde{\boldsymbol{S}}\tilde{\boldsymbol{A}}^{+}(\boldsymbol{I}-\boldsymbol{P})-\tilde{\boldsymbol{S}}\boldsymbol{A}^{+}\boldsymbol{P}-\tilde{\boldsymbol{S}}\boldsymbol{A}^{+}(\boldsymbol{I}-\boldsymbol{P}) \\ &\quad +(\boldsymbol{I}-\tilde{\boldsymbol{S}})\,\tilde{\boldsymbol{A}}^{+}\boldsymbol{P}+(\boldsymbol{I}-\tilde{\boldsymbol{S}})\,\tilde{\boldsymbol{A}}^{+}(\boldsymbol{I}-\boldsymbol{P})-(\boldsymbol{I}-\tilde{\boldsymbol{S}})\,\boldsymbol{A}^{+}\boldsymbol{P}-(\boldsymbol{I}-\tilde{\boldsymbol{S}})\,\boldsymbol{A}^{+}(\boldsymbol{I}-\boldsymbol{P}). \end{aligned}$$

Wegen $\tilde{\boldsymbol{S}}\tilde{\boldsymbol{A}}^{+} = \tilde{\boldsymbol{A}}^{+}$, $\boldsymbol{A}^{+}\boldsymbol{P} = \boldsymbol{A}^{+}$ vereinfacht sich diese Darstellung zu

$$\boldsymbol{G} = [\tilde{\boldsymbol{A}}^{+}\boldsymbol{P}-\tilde{\boldsymbol{S}}\boldsymbol{A}^{+}]+[\tilde{\boldsymbol{A}}^{+}(\boldsymbol{I}-\boldsymbol{P})]+[-(\boldsymbol{I}-\tilde{\boldsymbol{S}})\,\boldsymbol{A}^{+}] =: \boldsymbol{G}_1+\boldsymbol{G}_2+\boldsymbol{G}_3. \tag{24}$$

Unter Berücksichtigung von $\boldsymbol{A}^{\mathsf{T}}(\boldsymbol{I}-\boldsymbol{P}) = \boldsymbol{O}$, $(\boldsymbol{I}-\tilde{\boldsymbol{S}})\,\tilde{\boldsymbol{A}}^{\mathsf{T}} = \boldsymbol{O}$ folgt schließlich

$$\begin{aligned} \boldsymbol{G}_1 &= \tilde{\boldsymbol{A}}^{+}\boldsymbol{A}\boldsymbol{A}^{+}-\tilde{\boldsymbol{A}}^{+}\tilde{\boldsymbol{A}}\boldsymbol{A}^{+} = -\tilde{\boldsymbol{A}}^{+}\boldsymbol{\delta A}\boldsymbol{A}^{+}, \\ \boldsymbol{G}_2 &= \tilde{\boldsymbol{A}}^{+}(\boldsymbol{I}-\boldsymbol{P}) = \tilde{\boldsymbol{A}}^{+}\tilde{\boldsymbol{P}}(\boldsymbol{I}-\boldsymbol{P}) = \tilde{\boldsymbol{A}}^{+}\tilde{\boldsymbol{A}}^{+\mathsf{T}}\tilde{\boldsymbol{A}}^{\mathsf{T}}(\boldsymbol{I}-\boldsymbol{P}) \\ &= \tilde{\boldsymbol{A}}^{+}\tilde{\boldsymbol{A}}^{+\mathsf{T}}(\tilde{\boldsymbol{A}}^{\mathsf{T}}-\boldsymbol{A}^{\mathsf{T}})\,(\boldsymbol{I}-\boldsymbol{P}) = \tilde{\boldsymbol{A}}^{+}\tilde{\boldsymbol{A}}^{+\mathsf{T}}\boldsymbol{\delta A}^{\mathsf{T}}(\boldsymbol{I}-\boldsymbol{P}), \\ \boldsymbol{G}_3 &= -(\boldsymbol{I}-\tilde{\boldsymbol{S}})\,\boldsymbol{A}^{+} = -(\boldsymbol{I}-\tilde{\boldsymbol{S}})\,\boldsymbol{A}^{+}\boldsymbol{A}\boldsymbol{A}^{+} = -(\boldsymbol{I}-\tilde{\boldsymbol{S}})\,\boldsymbol{A}^{\mathsf{T}}\boldsymbol{A}^{+\mathsf{T}}\boldsymbol{A}^{+} \\ &= -(\boldsymbol{I}-\tilde{\boldsymbol{S}})\,(\boldsymbol{A}^{\mathsf{T}}-\tilde{\boldsymbol{A}}^{\mathsf{T}})\,\boldsymbol{A}^{+\mathsf{T}}\boldsymbol{A}^{+} = (\boldsymbol{I}-\tilde{\boldsymbol{S}})\,\boldsymbol{\delta A}^{\mathsf{T}}\boldsymbol{A}^{+\mathsf{T}}\boldsymbol{A}^{+}. \end{aligned} \tag{25}$$

Zur Abschätzung der $\boldsymbol{G}_i$ ziehen wir (19) heran und erhalten sofort

$$\|\boldsymbol{G}_1\| \leqq \|\tilde{\boldsymbol{A}}^+\| \, \|\boldsymbol{A}^+\| \, \|\delta\boldsymbol{A}\| \leqq [\|\boldsymbol{A}^+\|^2/(1-\varkappa)] \, \|\delta\boldsymbol{A}\| =: \alpha$$

und

$$\|\boldsymbol{G}_3\| \leqq \|\boldsymbol{A}^+\|^2 \, \|\delta\boldsymbol{A}\| \leqq \alpha. \tag{26}$$

Um den Faktor $\|\tilde{\boldsymbol{A}}^+\|^2$ bei der Abschätzung von $\boldsymbol{G}_2$ zu vermeiden, verwenden wir Ü 8.2.1. Mit dieser folgt

$$\|\boldsymbol{G}_2\| = \|\tilde{\boldsymbol{A}}^+\tilde{\boldsymbol{P}}(\boldsymbol{I}-\boldsymbol{P})\| \leqq \|\tilde{\boldsymbol{A}}^+\| \, \|\tilde{\boldsymbol{P}}(\boldsymbol{I}-\boldsymbol{P})\| = \|\tilde{\boldsymbol{A}}^+\| \, \|\boldsymbol{P}(\boldsymbol{I}-\tilde{\boldsymbol{P}})\| \leqq \alpha, \tag{27}$$

denn wegen $\tilde{\boldsymbol{A}}^{\mathsf{T}}(\boldsymbol{I}-\tilde{\boldsymbol{P}}) = \boldsymbol{O}$ gilt

$$\begin{aligned}\boldsymbol{P}(\boldsymbol{I}-\tilde{\boldsymbol{P}}) &= \boldsymbol{A}\boldsymbol{A}^+(\boldsymbol{I}-\tilde{\boldsymbol{P}}) = \boldsymbol{A}^{+\mathsf{T}}\boldsymbol{A}^{\mathsf{T}}(\boldsymbol{I}-\tilde{\boldsymbol{P}}) = \boldsymbol{A}^{+\mathsf{T}}(\boldsymbol{A}^{\mathsf{T}}-\tilde{\boldsymbol{A}}^{\mathsf{T}})\,(\boldsymbol{I}-\tilde{\boldsymbol{P}}) \\ &= -\boldsymbol{A}^{+\mathsf{T}}\delta\boldsymbol{A}^{\mathsf{T}}(\boldsymbol{I}-\tilde{\boldsymbol{P}}).\end{aligned}$$

Zusammenfassend ergibt sich $\|\boldsymbol{G}\| \leqq 3\alpha$, also (22) mit $\mu = 3$. Die verbesserten Werte (23) für μ erhält man durch elementare geometrische Überlegungen unter Ausnutzung von Orthogonalitätsbeziehungen, siehe Ü 8.2.2. Aus (22) folgt speziell

$$(\boldsymbol{A}+\delta\boldsymbol{A})^+ = \boldsymbol{A}^+ + \boldsymbol{O}(\|\delta\boldsymbol{A}\|),$$

so daß $\tilde{\boldsymbol{A}}^+$ in (24), (25) durch $\boldsymbol{A}^+$ ersetzt werden kann, um zur linearisierten Änderungsformel (20) zu gelangen. □

8.2.6. *Bemerkung.* (i) Wie die Beispiele 8.2.1 und 8.2.2 zeigen, kann auf keine der Voraussetzungen (16) und (17) verzichtet werden, um die Beschränktheit von $(\boldsymbol{A}+\delta\boldsymbol{A})^+$ zu sichern. Außerdem zeigen sie, daß die Abschätzungen (14) und (19) scharf sind, d. h., zu jedem $\Delta\boldsymbol{A}$ mit $\|\boldsymbol{A}^+\| \, \Delta\boldsymbol{A} < 1$ gibt es Störungen $\delta\boldsymbol{A}$ mit $\|\delta\boldsymbol{A}\| = \Delta\boldsymbol{A}$, so daß das Gleichheitszeichen steht.

(ii) Das Rangerhöhungsverbot (16) ist für jede beliebige Störung $\delta\boldsymbol{A}$ erfüllt, wenn $\boldsymbol{A}$ Vollrang hat, also spalten- oder zeilenregulär ist. *Für Vollrangmatrizen ist die Bestimmung der Pseudoinversen daher ein korrekt gestelltes Problem,* was übrigens bereits aus den in diesem Fall gültigen expliziten Darstellungen 8.1.9(i) für $\boldsymbol{A}^-$ folgt. *Für rangdefizientes $\boldsymbol{A}$ ist die Bestimmung von $\boldsymbol{A}^+$ ein inkorrekt gestelltes Problem, sofern die Störungen nicht gemäß* (16) *eingeschränkt werden.*

(iii) Für reguläres $\boldsymbol{A}$ geht (22) in die Abschätzung (4.1.10) des Störungslemmas über, während sich (20) wegen $\boldsymbol{Q}_{\boldsymbol{A}} = \boldsymbol{Q}_{\boldsymbol{A}^{\mathsf{T}}} = 0$ auf die bekannte Formel

$$(\boldsymbol{A}+\delta\boldsymbol{A})^{-1} - \boldsymbol{A}^{-1} = -\boldsymbol{A}^{-1}\delta\boldsymbol{A}\boldsymbol{A}^{-1} + O(\|\delta\boldsymbol{A}\|^2)$$

reduziert, siehe Ü 4.1.2.

B. *Störungstheorie der Normallösung*

Wir kehren jetzt zu den Quadratmittelproblemen (1), (3) zurück und untersuchen die Empfindlichkeit von $\boldsymbol{x}$ gegenüber nicht rangerhöhenden Störungen $\delta\boldsymbol{A}$.

8.2.7. *Satz.* Gegeben seien die Quadratmittelprobleme

$$\boldsymbol{A}\boldsymbol{x} \cong \boldsymbol{b} \quad \text{und} \quad (\boldsymbol{A}+\delta\boldsymbol{A})\,(\boldsymbol{x}+\delta\boldsymbol{x}) \cong \boldsymbol{b}+\delta\boldsymbol{b}$$

mit den Normallösungen

$$\boldsymbol{x} = \boldsymbol{A}^+\boldsymbol{b} \quad \text{bzw.} \quad \boldsymbol{x}+\delta\boldsymbol{x} = (\boldsymbol{A}+\delta\boldsymbol{A})^+\,(\boldsymbol{b}+\delta\boldsymbol{b}).$$

Die Störung δA genüge den Bedingungen

$$\operatorname{rang}(A + \delta A) \leqq \operatorname{rang}(A) \quad \text{sowie} \quad \varkappa := \|A^+\| \, \|\delta A\| < 1.$$

(i) Dann gilt

$$\delta x = \delta x' + O\big(\|\delta A\| \, (\|\delta A\| + \|\delta b\|)\big),$$

wobei

$$\delta x' := A^+\{-\delta A\, x + \delta b\} + A^+ A^{+\mathsf{T}} \delta A^{\mathsf{T}} r + (I - A^+ A)\, \delta A^{\mathsf{T}} A^{+\mathsf{T}} x \tag{28}$$

den bezüglich $\delta A, \delta b$ linearen Teil von δx darstellt und

$$r = b - Ax = (I - AA^+)\, b$$

das Residuum von x bezeichnet.

(ii) Die Störung δx genügt den Abschätzungen

$$\|\delta x\| \leqq \frac{\|A^+\|}{1 - \varkappa} \{\|\delta A\| \, [\omega \, \|x\| + \|A^+\| \, \|r\|] + \|\delta b\|\} \tag{29}$$

und (im Fall $x \neq o$)

$$\frac{\|\delta x\|}{\|x\|} \leqq \frac{1}{1 - \varkappa} \left\{ \left[\omega \operatorname{cond}(A) + \frac{\|r\|}{\|A\| \, \|x\|} \big(\operatorname{cond}(A)\big)^2 \right] \frac{\|\delta A\|}{\|A\|} + \frac{\|A^+\| \, \|b\|}{\|x\|} \cdot \frac{\|\delta b\|}{\|b\|} \right\} \tag{30}$$

mit

$$\operatorname{cond}(A) := \|A\| \, \|A^+\| \tag{31}$$

und

$$\omega := \begin{cases} \sqrt{2}, & \text{falls} \quad \operatorname{rang}(A) < n, \\ 1, & \text{falls} \quad \operatorname{rang}(A) = n. \end{cases}$$

Beweis. Mit den Bezeichnungen aus dem Beweis zu 8.2.5 und (5) gilt

$$\delta x = Gb + \tilde{A}^+ \delta b = G_1 b + G_2 b + G_3 b + \tilde{A}^+ \delta b. \tag{32}$$

Unter Beachtung von (19), (25) und (8.1.42) folgt

$$\|G_1 b\| = \|\tilde{A}^+ \delta A A^+ b\| = \|\tilde{A}^+ \delta A x\| \leqq [\|A^+\|/(1 - \varkappa)] \, \|\delta A\| \, \|x\| =: \beta,$$

$$\|G_3 b\| = \|(I - \tilde{S})\, \delta A^{\mathsf{T}} A^{+\mathsf{T}} A^+ b\| \leqq \|A^+\| \, \|\delta A\| \, \|x\| \leqq \beta,$$

wegen der Orthogonalität von $G_1 b$ und $G_3 b$ also

$$\|G_1 b + G_3 b\|^2 = \|G_1 b\|^2 + \|G_3 b\|^2 \leqq 2\beta^2.$$

Wegen $(I - P)\, b = r = (I - P)\, r$ gilt

$$G_2 b = \tilde{A}^+ (I - P)\, b = \tilde{A}^+ (I - P)\, r = G_2 r, \tag{33}$$

und mit (27) folgt

$$\|\boldsymbol{G}_2\boldsymbol{b}\| \leqq \alpha\,\|\boldsymbol{r}\|, \qquad \|\tilde{\boldsymbol{A}}^+\delta\boldsymbol{b}\| \leqq [\|\boldsymbol{A}^+\|/(1-\varkappa)]\,\|\delta\boldsymbol{b}\| =: \gamma.$$

Damit erhalten wir

$$\|\delta\boldsymbol{x}\| \leqq \|\boldsymbol{G}_1\boldsymbol{b} + \boldsymbol{G}_3\boldsymbol{b}\| + \|\boldsymbol{G}_2\boldsymbol{b}\| + \|\tilde{\boldsymbol{A}}^+\delta\boldsymbol{b}\| \leqq \sqrt{2}\,\beta + \alpha\,\|\boldsymbol{r}\| + \gamma.$$

Dies ist gerade (29), und Division durch $\|\boldsymbol{x}\|$ liefert (30). Im Fall $\operatorname{rang}(\boldsymbol{A}) = n$ ist $\boldsymbol{G}_3 = \boldsymbol{O}$, so daß $\sqrt{2}$ durch 1 ersetzt werden kann. Die Linearisierungsformel (28) ergibt sich mit (33) aus (32), wenn $\tilde{\boldsymbol{A}}$ überall durch $\boldsymbol{A}$ ersetzt wird. □

Satz 8.2.7 besagt: *Die Bestimmung der Normallösung eines Quadratmittelproblems ist eine lokal lipschitzstetige Aufgabenklasse, sofern keine rangerhöhenden Störungen von $\boldsymbol{A}$ zugelassen werden.*

Die partiellen relativen Konditionszahlen sind

$$K_{\boldsymbol{A}}(\boldsymbol{A}, \boldsymbol{b}) = \omega\,\operatorname{cond}(\boldsymbol{A}) + \frac{\|\boldsymbol{r}\|}{\|\boldsymbol{A}\|\,\|\boldsymbol{x}\|}\,[\operatorname{cond}(\boldsymbol{A})]^2, \quad K_{\boldsymbol{b}}(\boldsymbol{A}, \boldsymbol{b}) = \frac{\|\boldsymbol{A}^+\|\,\|\boldsymbol{b}\|}{\|\boldsymbol{x}\|}, \tag{34}$$

und es gilt

$$1 \leqq K_{\boldsymbol{b}}(\boldsymbol{A}, \boldsymbol{b}) \leqq K_{\boldsymbol{A}}(\boldsymbol{A}, \boldsymbol{b}). \tag{35}$$

Die durch (31) eingeführte Zahl $\operatorname{cond}(\boldsymbol{A})$ ist für reguläres $\boldsymbol{A}$ mit der in 4.1 eingeführten Konditionszahl identisch und wird deshalb auch *(Spektral-) Konditionszahl der rechteckigen Matrix $\boldsymbol{A}$* genannt. Für sie gilt

$$\operatorname{cond}(\boldsymbol{A}) = \sigma_1/\sigma_r, \quad \operatorname{cond}(\boldsymbol{A}) \geqq 1 \text{ für } \boldsymbol{A} \neq \boldsymbol{O}, \quad \operatorname{cond}(\boldsymbol{A}) = 0 \text{ für } \boldsymbol{A} = \boldsymbol{O}. \tag{36}$$

8.2.8. *Bemerkung.* (i) Für konsistente Gleichungssysteme — zu diesen gehören diejenigen mit zeilenregulärer Matrix $\boldsymbol{A}$, vgl. 8.1.A — ist $\boldsymbol{r} = \boldsymbol{o}$, so daß (30) in

$$\frac{\|\delta\boldsymbol{x}\|}{\|\boldsymbol{x}\|} \leqq \frac{1}{1-\varkappa}\left\{\omega\,\operatorname{cond}(\boldsymbol{A})\,\frac{\|\delta\boldsymbol{A}\|}{\|\boldsymbol{A}\|} + \frac{\|\boldsymbol{A}^+\|\,\|\boldsymbol{b}\|}{\|\boldsymbol{x}\|}\,\frac{\|\delta\boldsymbol{b}\|}{\|\boldsymbol{b}\|}\right\} \tag{37}$$

übergeht. Bis auf den Faktor $\omega \leqq \sqrt{2}$ entspricht (37) der für reguläre Systeme gültigen Abschätzung (4.1.14). *Die Normallösung konsistenter rechteckiger Quadratmittelprobleme hat also dasselbe qualitative Störungsverhalten gegenüber kleinen, den Rang nicht erhöhenden Störungen wie die Lösung regulärer Gleichungssysteme.* Insbesondere ist $K_{\boldsymbol{A}}$ — bis auf den für $\operatorname{rang}(\boldsymbol{A}) < n$ auftretenden Faktor $\omega = \sqrt{2}$ — mit $\operatorname{cond}(\boldsymbol{A})$ identisch und hängt nur von der Matrix $\boldsymbol{A}$ ab. Für reguläres $\boldsymbol{A}$ geht (37) bzw. (30) in (4.1.14) über.

(ii) Für inkonsistente Gleichungssysteme ist $\boldsymbol{r} \neq \boldsymbol{o}$, und in $K_{\boldsymbol{A}}$ tritt zusätzlich zu $\omega\,\operatorname{cond}(\boldsymbol{A})$ der zu $[\operatorname{cond}(\boldsymbol{A})]^2$ proportionale Term

$$\frac{\|\boldsymbol{r}\|}{\|\boldsymbol{A}\|\,\|\boldsymbol{x}\|}\,[\operatorname{cond}(\boldsymbol{A})]^2 \tag{38}$$

auf.

Für genügend kleines $\|\boldsymbol{r}\|$ dominiert der erste Summand in (34), d. h., es gilt

$$K_{\boldsymbol{A}} \approx \omega\,\operatorname{cond}(\boldsymbol{A}),$$

und das Störungsverhalten ist ähnlich wie im konsistenten Fall. Für großes $\|\boldsymbol{r}\|$ dominiert (38), d. h., $K_{\boldsymbol{A}}$ wird proportional zu [cond $(\boldsymbol{A})$]^2. *Im Gegensatz zur Situation bei regulären Gleichungssystemen hängt also* $K_{\boldsymbol{A}}$ *bei Quadratmittelproblemen nicht nur von* $\boldsymbol{A}$, *sondern über das Residuum* $\boldsymbol{r}$ *auch von der rechten Seite* $\boldsymbol{b}$ *ab und kann proportional zu* [cond $(\boldsymbol{A})$]^2 *sein.*

(iii) Die Bedingung rang $(\boldsymbol{A} + \delta\boldsymbol{A}) \leqq$ rang $(\boldsymbol{A})$ ist für alle $\boldsymbol{A}$ erfüllt, wenn $\boldsymbol{A}$ spalten- oder zeilenregulär ist. Im ersten Fall ist $\boldsymbol{x}$ die eindeutige Quadratmittellösung von (1), im zweiten Fall ist (1) konsistent und $\boldsymbol{r} = \boldsymbol{o}$. □

Wir zeigen abschließend an einem Beispiel, wie sich rangerhöhende Störungen auf die Normallösung auswirken können.

8.2.9. *Beispiel.* Es seien $\boldsymbol{A}$, $\delta\boldsymbol{A}$ wie in 8.2.2, und $\boldsymbol{b}$, $\delta\boldsymbol{b}$ seien durch

$$\boldsymbol{\beta} = \boldsymbol{U}^{\mathrm{T}}\boldsymbol{b}, \qquad \delta\boldsymbol{\beta} = \boldsymbol{U}^{\mathrm{T}}\delta\boldsymbol{b}$$

gegeben. Dann gilt für $\varepsilon \neq 0$

$$\delta\boldsymbol{x} = \boldsymbol{V}\delta\boldsymbol{\xi} \quad \text{mit} \quad \delta\boldsymbol{\xi} = (\delta\xi_j), \qquad \delta\xi_j = \begin{cases} \delta\beta_j/\sigma_j & \text{für} \quad j = 1, \ldots, r, \\ (\beta_{r+1} + \delta\beta_{r+1})/\varepsilon & \text{für} \quad j = r+1, \\ 0 & \text{sonst}, \end{cases}$$

also

$$\|\delta\boldsymbol{x}\|^2 = \|\delta\boldsymbol{\xi}\|^2 = \|\boldsymbol{\Sigma}^{+}\delta\boldsymbol{\beta}\|^2 + [(\beta_{r+1} + \delta\beta_{r+1})/\varepsilon]^2.$$

Im Fall $\beta_{r+1} + \delta\beta_{r+1} = 0$ wirkt sich ε überhaupt nicht auf $\delta\boldsymbol{x}$ aus, obwohl die Pseudoinverse wie $1/|\varepsilon|$ wächst. Die zu $1/|\varepsilon|$ proportionale bösartige Wirkung tritt nur im Fall $\beta_{r+1} + \delta\beta_{r+1} \neq 0$ auf. □

Der Einfluß einer rangerhöhenden Störung $\delta\boldsymbol{A}$ *auf* $\delta\boldsymbol{x}$ *hängt also auch von* $\boldsymbol{b} + \delta\boldsymbol{b}$ *ab.* Die Störung wirkt nicht bösartig, wenn im diagonalisierten Problem $(\boldsymbol{\Sigma} + \delta\boldsymbol{\Sigma}) \times (\boldsymbol{\xi} + \delta\boldsymbol{\xi}) \cong \boldsymbol{\beta} + \delta\boldsymbol{\beta}$ die Komponenten $(\boldsymbol{\beta} + \delta\boldsymbol{\beta})_i$ für $i = r+1, \ldots, l; l = \min(m, n)$, verschwinden. Wir werden später im Kapitel 11 solche rangerhöhenden, aber trotzdem gutartigen Störungen künstlich zur Regularisierung rangdefizienter Quadratmittelprobleme einführen.

C. *Die Qualität der Störungsabschätzungen*

Die Qualität der Abschätzung (29) bzw. (30) kann unter Verwendung der linearisierten Störung $\delta\boldsymbol{x}'$ aus (28) analog zu 4.1.B untersucht werden. Wie beim Beweis von 4.1.5 wird gezeigt, daß

$$\boldsymbol{A}^{+}\{-\delta\boldsymbol{A}\boldsymbol{x} + \delta\boldsymbol{b}\} \quad \text{durch} \quad \|\boldsymbol{A}^{+}\|\,\{\|\delta\boldsymbol{A}\|\,\|\boldsymbol{x}\| + \|\delta\boldsymbol{b}\|\} \tag{39}$$

scharf abgeschätzt wird. Ebenso wird

$$\boldsymbol{A}^{+}\boldsymbol{A}^{+\mathrm{T}}\delta\boldsymbol{A}^{\mathrm{T}}\boldsymbol{r} \quad \text{durch} \quad \|\boldsymbol{A}^{+}\|^2\,\|\delta\boldsymbol{A}\|\,\|\boldsymbol{r}\| \tag{40}$$

scharf abgeschätzt. Der im Fall rang $(\boldsymbol{A}) = n$ wegfallende und zu $\boldsymbol{A}^{+}\delta\boldsymbol{A}\boldsymbol{x}$ orthogonale dritte Term

$$(\boldsymbol{I} - \boldsymbol{A}^{+}\boldsymbol{A})\,\delta\boldsymbol{A}^{\mathrm{T}}\boldsymbol{A}^{+\mathrm{T}}\boldsymbol{x} \tag{41}$$

von $\delta x'$ erhöht den Faktor 1 bei $\|\delta A\|\,\|x\|$ in (29) auf $\sqrt{2}$, spielt also praktisch keine Rolle. Sofern einer der beiden Summanden von K_A deutlich dominiert, sind die Abschätzungen von 8.2.7 also asymptotisch fast scharf. Eine genauere Untersuchung der zu den Termen (39), (40) gehörenden Fehlerellipsoide analog zu 4.1.7 zeigt, daß diese Abschätzungen den maximal möglichen Fehler stets asymptotisch qualitativ richtig wiedergeben.

D. *Skalierung*

Bei regulären Gleichungssystemen haben wir eine Skalierung mit einer Diagonalmatrix D vorgenommen, um die Kondition von DA zu reduzieren oder die Fehlerabschätzung bezüglich elementweiser Schranken für δA, δb realistisch zu gestalten, vgl. 4.1.C. Es bleibt zu überlegen, ob für Quadratmittelprobleme ein analoges Vorgehen, d. h., der Übergang von

$$Ax \simeq b \quad \text{zu} \quad DAx \simeq Db \quad \text{mit} \quad D = \operatorname{diag}(d_i), \quad d_i > 0 \quad (i = 1, \ldots, m),$$

sinnvoll ist. Wir unterscheiden wieder zwei Fälle.

Fall 1: $Ax = b$ *konsistent*, $r = o$. Dann sind die Lösungen des skalierten und unskalierten Quadratmittelproblems identisch und stimmen mit den Lösungen der jeweiligen Gleichungssysteme überein. Die Formulierung als Quadratmittelproblem ist also überflüssig. In diesem Fall können die Überlegungen aus 4.1.C sinngemäß übernommen werden, denn der Term (40) verschwindet in der linearisierten Störung $\delta x'$, und (41) trägt nur unwesentlich zu $\|\delta x'\|$ bei. Fall 1 tritt stets ein, wenn A zeilenregulär ist.

Fall 2: $Ax = b$ *inkonsistent*, $r \neq o$. Dies ist in der Regel — d. h., für fast alle b — der Fall, wenn $m > n$ gilt, $Ax = b$ also streng überbestimmt ist, vgl. 8.1.A. Beim Übergang vom unskalierten zum skalierten Quadratmittelproblem ändert sich i. allg. die Lösungsmenge. Deshalb muß eine Skalierung nach den obigen numerischen Gesichtspunkten verworfen werden, sondern es ist so zu skalieren, daß die sich dabei einstellende Lösung im Hinblick auf die im Hintergrund stehende reale Aufgabe „richtig" bzw. zumindest „vernünftig" ist. Da die auf überbestimmte Gleichungen führenden Probleme in der Regel statistischer Natur sind, wird die natürliche Skalierung bzw. Wichtung des zugehörigen Quadratmittelproblems durch die statistischen Eigenschaften des Hintergrundproblems festgelegt und muß als gegeben angesehen werden, siehe Abschnitt 8.3.

D. *Residualabschätzungen*

Da 4.1.4 auch für rechteckige Matrizen gilt, kann 4.1.18 wörtlich übernommen werden.

8.2.10. *Aussage.* Es seien $\Delta A, \Delta b \geqq 0$ und $A \in \mathbf{R}^{m,n}$, $b \in \mathbf{R}^m$, $\tilde{x} \in \mathbf{R}^n$ gegeben. Dann existieren genau dann Störungen $\delta A, \delta b$ mit

$$(A + \delta A)\,\tilde{x} = b + \delta b \quad \text{und} \quad \|\delta A\| \leqq \Delta A, \qquad \|\delta b\|_i \leqq \Delta b, \tag{42}$$

wenn

$$\|\boldsymbol{b} - \boldsymbol{A}\tilde{\boldsymbol{x}}\| \leqq \Delta\boldsymbol{b} + \Delta\boldsymbol{A}\,\|\tilde{\boldsymbol{x}}\| \tag{43}$$

ist.

Falls $\tilde{\boldsymbol{r}} = \boldsymbol{b} - \boldsymbol{A}\tilde{\boldsymbol{x}}$ im Sinne von (43) klein ist, stellt $\tilde{\boldsymbol{x}}$ also eine Lösung — i. allg. jedoch nicht die Normallösung — des benachbarten Systems (42) dar. Für konsistente Systeme, bei denen $\boldsymbol{r} = \boldsymbol{b} - \boldsymbol{A}\boldsymbol{x}$ verschwindet, ist dies ein brauchbares Resultat. Für inkonsistente Systeme, bei denen das minimale Residuum von $\boldsymbol{o}$ verschieden ist, hat es jedoch keine praktische Bedeutung. Für solche Systeme bietet sich an, statt des inkonsistenten Systems (1) die stets konsistenten Normalgleichungen

$$\boldsymbol{g} := \boldsymbol{A}^{\mathsf{T}}\boldsymbol{r} = \boldsymbol{A}^{\mathsf{T}}(\boldsymbol{b} - \boldsymbol{A}\boldsymbol{x}) = \boldsymbol{A}^{\mathsf{T}}\boldsymbol{b} - \boldsymbol{A}^{\mathsf{T}}\boldsymbol{A}\boldsymbol{x} =: \boldsymbol{c} - \boldsymbol{M}\boldsymbol{x}$$

als Ausgangspunkt zu nehmen. Gilt

$$\|\tilde{\boldsymbol{g}}\| = \|\boldsymbol{A}^{\mathsf{T}}\tilde{\boldsymbol{r}}\| \leqq \varepsilon\,\|\boldsymbol{A}\|\,\|\tilde{\boldsymbol{r}}\|, \tag{44}$$

so folgt durch Anwendung von 8.2.10 auf $\boldsymbol{M}\boldsymbol{x} = \boldsymbol{c}$ mit $\Delta\boldsymbol{c} := 0$, $\Delta\boldsymbol{M} := \varepsilon\,\|\boldsymbol{A}\|\,\|\tilde{\boldsymbol{r}}\|/\|\tilde{\boldsymbol{x}}\|$ die Existenz einer Störung $\delta\boldsymbol{M}$ mit

$$(\boldsymbol{A}^{\mathsf{T}}\boldsymbol{A} + \delta\boldsymbol{M})\,\tilde{\boldsymbol{x}} = \boldsymbol{A}^{\mathsf{T}}\boldsymbol{b} \quad \text{und} \quad \|\delta\boldsymbol{M}\| \leqq \Delta\boldsymbol{M}, \tag{45}$$

wobei $\delta\boldsymbol{M}$ nach 4.1.20(ii) sogar symmetrisch gewählt werden kann. Der Vektor $\tilde{\boldsymbol{x}}$ löst also die gestörten Normalgleichungen (45), aber diese sind — leider! — nicht die Normalgleichungen eines gestörten Quadratmittelproblems, so daß bezüglich der eigentlichen Aufgabenstellung $\boldsymbol{A}\boldsymbol{x} \simeq \boldsymbol{b}$ keine Gutartigkeit gefolgert werden kann.

Wir gehen daher einen anderen Weg und wenden 8.2.10 auf das System

$$\boldsymbol{A}^{\mathsf{T}}\boldsymbol{r} = \boldsymbol{g} = \boldsymbol{o}$$

an. Mit $\Delta\boldsymbol{g} := 0$, $\Delta\boldsymbol{A}^{\mathsf{T}} := \varepsilon\,\|\boldsymbol{A}\|$ folgt dann die Existenz einer Störung $\delta_1\boldsymbol{A}$ mit

$$(\boldsymbol{A} + \delta_1\boldsymbol{A})^{\mathsf{T}}\,\tilde{\boldsymbol{r}} = (\boldsymbol{A} + \delta_1\boldsymbol{A})^{\mathsf{T}}\,(\boldsymbol{b} - \boldsymbol{A}\tilde{\boldsymbol{x}}) = \boldsymbol{O} \quad \text{und} \quad \|\delta_1\boldsymbol{A}\| \leqq \varepsilon\,\|\boldsymbol{A}\|. \tag{46}$$

Das folgende Lemma zeigt, daß den gestörten Normalgleichungen (46) die Normalgleichungen eines gestörten Quadratmittelproblems zugeordnet werden können, wobei noch etwas allgemeinere Störungen zugelassen werden dürfen.

8.2.11. *Lemma.* Für $\boldsymbol{A} \in \mathbf{R}^{m,n}$, $\boldsymbol{b} \in \mathbf{R}^m$, $\tilde{\boldsymbol{x}} \in \mathbf{R}^n$ und die Störungen $\delta_1\boldsymbol{A}$, $\delta_2\boldsymbol{A}$ gelte

$$(\boldsymbol{A} + \delta_1\boldsymbol{A})^{\mathsf{T}}\,[\boldsymbol{b} - (\boldsymbol{A} + \delta_2\boldsymbol{A})\,\tilde{\boldsymbol{x}}] = \boldsymbol{O}.$$

Dann gibt es eine Störung $\delta\boldsymbol{A}$ mit

$$\|\delta\boldsymbol{A}\| \leqq \{\|\delta_1\boldsymbol{A}\|^2 + \|\delta_2\boldsymbol{A}\|^2\}^{1/2}, \tag{47}$$

so daß

$$(\boldsymbol{A} + \delta\boldsymbol{A})^{\mathsf{T}}\,[\boldsymbol{b} - (\boldsymbol{A} + \delta\boldsymbol{A})\,\tilde{\boldsymbol{x}}] = \boldsymbol{O}$$

gilt, d. h., $\tilde{\boldsymbol{x}}$ löst das Quadratmittelproblem $(\boldsymbol{A} + \delta\boldsymbol{A})\,\tilde{\boldsymbol{x}} \simeq \boldsymbol{b}$.

Beweis. Im Fall $\boldsymbol{s} := \boldsymbol{b} - (\boldsymbol{A} + \delta_2\boldsymbol{A})\,\tilde{\boldsymbol{x}} = \boldsymbol{o}$ setze man $\delta\boldsymbol{A} := \delta_2\boldsymbol{A}$, andernfalls

$$\delta\boldsymbol{A} := \boldsymbol{P}\delta_1\boldsymbol{A} + (\boldsymbol{I} - \boldsymbol{P})\,\delta_2\boldsymbol{A} \quad \text{mit} \quad \boldsymbol{P} := \boldsymbol{s}\boldsymbol{s}^{\mathsf{T}}/\|\boldsymbol{s}\|^2.$$

Nach Ü 8.1.7 ist $\boldsymbol{P}$ der Projektor auf span $\{\boldsymbol{s}\}$. Mit $\boldsymbol{D} := \delta_1\boldsymbol{A} - \delta_2\boldsymbol{A}$ folgt

$$\boldsymbol{t} := \boldsymbol{b} - (\boldsymbol{A} + \delta\boldsymbol{A})\,\tilde{\boldsymbol{x}} = \boldsymbol{b} - (\boldsymbol{A} + \delta_2\boldsymbol{A})\,\tilde{\boldsymbol{x}} - \boldsymbol{P}\boldsymbol{D}\tilde{\boldsymbol{x}} = \lambda\boldsymbol{s},$$

wobei $\lambda := 1 - (\boldsymbol{D}\tilde{\boldsymbol{x}})^{\mathsf{T}}\,\boldsymbol{s}/\|\boldsymbol{s}\|^2$ ist. Mithin gilt

$$(\boldsymbol{A} + \delta\boldsymbol{A})^{\mathsf{T}}\boldsymbol{t} = [\boldsymbol{A} + \delta_1\boldsymbol{A} - (\boldsymbol{I} - \boldsymbol{P})\,\boldsymbol{D}]^{\mathsf{T}}\,\boldsymbol{t}$$
$$= \lambda[(\boldsymbol{A} + \delta_1\boldsymbol{A})^{\mathsf{T}}\,\boldsymbol{s} - \boldsymbol{D}^{\mathsf{T}}(\boldsymbol{I} - \boldsymbol{P})\,\boldsymbol{s}] = \boldsymbol{o}. \quad \square$$

Offensichtlich kann in (47) auch die Frobeniusnorm verwendet werden.

Bei Anwendung von 8.2.11 auf (46) ergibt sich mit $\delta_2\boldsymbol{A} = \boldsymbol{O}$ das nachfolgende gewünschte Resultat:

8.2.12. *Aussage.* Das zu $\tilde{\boldsymbol{x}}$ gehörende Residuum

$$\tilde{\boldsymbol{g}} = \boldsymbol{A}^{\mathsf{T}}\tilde{\boldsymbol{r}} = \boldsymbol{A}^{\mathsf{T}}(\boldsymbol{b} - \boldsymbol{A}\tilde{\boldsymbol{x}})$$

bezüglich der Normalgleichungen von $\boldsymbol{A}\boldsymbol{x} \cong \boldsymbol{b}$ genüge der Abschätzung

$$\|\tilde{\boldsymbol{g}}\| \leqq \varepsilon\,\|\boldsymbol{A}\|\,\|\tilde{\boldsymbol{r}}\|. \tag{48}$$

Dann existiert eine Störung $\delta\boldsymbol{A}$, so daß $\boldsymbol{x}$ das Quadratmittelproblem

$$(\boldsymbol{A} + \delta\boldsymbol{A})\,\tilde{\boldsymbol{x}} \cong \boldsymbol{b} \quad \text{mit} \quad \|\delta\boldsymbol{A}\| \leqq \varepsilon\,\|\boldsymbol{A}\| \tag{49}$$

löst.

Wir bemerken noch, daß (48) hinreichend, aber keineswegs notwendig für (49) ist. Insofern ist 8.2.12 schwächer als 8.2.10, wo (42) und (43) äquivalent sind.

Übungsaufgaben

Ü 8.2.1. Für $\boldsymbol{A}, \tilde{\boldsymbol{A}} \in \mathbf{R}^{m,n}$ gelte rang $(\boldsymbol{A})$ = rang $(\tilde{\boldsymbol{A}})$. Man zeige

$$\|\tilde{\boldsymbol{P}}(\boldsymbol{I} - \boldsymbol{P})\| = \|\boldsymbol{P}(\boldsymbol{I} - \tilde{\boldsymbol{P}})\| \quad \text{für} \quad \boldsymbol{P} := \boldsymbol{A}\boldsymbol{A}^{+}, \quad \tilde{\boldsymbol{P}} := \tilde{\boldsymbol{A}}\tilde{\boldsymbol{A}}^{+}.$$

Hinweis: Unter Verwendung der Singulärwertzerlegungen stelle man $\boldsymbol{P}$, $\tilde{\boldsymbol{P}}$ gemäß

$$\boldsymbol{P} = \boldsymbol{U}\boldsymbol{\Pi}\boldsymbol{U}^{\mathsf{T}}, \qquad \tilde{\boldsymbol{P}} = \tilde{\boldsymbol{U}}\boldsymbol{\Pi}\tilde{\boldsymbol{U}}^{\mathsf{T}}$$

mit $\boldsymbol{\Pi} = \left(\begin{array}{c|c} \boldsymbol{I}_r & \boldsymbol{O} \\ \hline \boldsymbol{O} & \boldsymbol{O} \end{array}\right)$ dar und partitioniere $\boldsymbol{W} := \tilde{\boldsymbol{U}}^{\mathsf{T}}\boldsymbol{U}$ gemäß $\boldsymbol{W} = \left(\begin{array}{c|c} \boldsymbol{W}_{11} & \boldsymbol{W}_{12} \\ \hline \boldsymbol{W}_{21} & \boldsymbol{W}_{22} \end{array}\right)$. Man beweise $\|\tilde{\boldsymbol{P}}(\boldsymbol{I} - \boldsymbol{P})\| = \|\boldsymbol{W}_{12}\|$, $\|\boldsymbol{P}(\boldsymbol{I} - \tilde{\boldsymbol{P}})\| = \|\boldsymbol{W}_{21}\|$ und folgere $\|\boldsymbol{W}_{12}\| = \|\boldsymbol{W}_{21}\|$ aus der Orthogonalität von $\boldsymbol{W}$.

Ü 8.2.2. Für $\boldsymbol{y} \in \mathbf{R}^m$ und $\boldsymbol{G}_i$ gemäß (25) mit $\|\boldsymbol{G}_i\| \leqq \alpha$ werde

$$\boldsymbol{z} = \boldsymbol{G}_1\boldsymbol{y} + \boldsymbol{G}_2\boldsymbol{y} + \boldsymbol{G}_3\boldsymbol{y} =: \boldsymbol{z}_1 + \boldsymbol{z}_2 + \boldsymbol{z}_3$$

gebildet. Man zeige:

(i) $\boldsymbol{z}_1 \perp \boldsymbol{z}_3$, $\boldsymbol{z}_2 \perp \boldsymbol{z}_3$, folglich $\|\boldsymbol{z}\|^2 = \|\boldsymbol{z}_1 + \boldsymbol{z}_2\|^2 + \|\boldsymbol{z}_3\|^2$.

(ii) Mit $\boldsymbol{u} := \boldsymbol{P}\boldsymbol{y}$, $\boldsymbol{v} := (\boldsymbol{I} - \boldsymbol{P})\,\boldsymbol{y}$, $\boldsymbol{P} := \boldsymbol{A}\boldsymbol{A}^+$, gilt $\|\boldsymbol{u}\|^2 + \|\boldsymbol{v}\|^2 = \|\boldsymbol{y}\|^2$ und

$$\boldsymbol{G}_1\boldsymbol{y} = \boldsymbol{G}_1\boldsymbol{u}, \qquad \boldsymbol{G}_2\boldsymbol{y} = \boldsymbol{G}_2\boldsymbol{v}, \qquad \boldsymbol{G}_3\boldsymbol{y} = \boldsymbol{G}_3\boldsymbol{u},$$

wegen $\|\boldsymbol{z}_1 + \boldsymbol{z}_2\| \leqq \|\boldsymbol{G}_1\boldsymbol{y}\| + \|\boldsymbol{G}_2\boldsymbol{y}\| = \|\boldsymbol{G}_1\boldsymbol{u}\| + \|\boldsymbol{G}_2\boldsymbol{v}\| \leqq \alpha(\|\boldsymbol{u}\| + \|\boldsymbol{v}\|)$ also

$$\|\boldsymbol{z}\|^2 \leqq \alpha^2\{(\|\boldsymbol{u}\| + \|\boldsymbol{v}\|)^2 + \|\boldsymbol{u}\|^2\} =: \alpha^2\,\|\boldsymbol{y}\|^2\,N.$$

Mit $\|\boldsymbol{u}\| = \|\boldsymbol{y}\|\cos\varphi$, $\|\boldsymbol{v}\| = \|\boldsymbol{y}\|\sin\varphi$ folgt

$$N = (\cos\varphi + \sin\varphi)^2 + \cos^2\varphi = (3 + 2\sin 2\varphi + \cos 2\varphi)/2 \leqq \left[\left(1 + \sqrt{5}\right)/2\right]^2.$$

(iii) Im Fall $r = n$ ist $\boldsymbol{z}_3 = \boldsymbol{o}$, also $\|\boldsymbol{z}\|^2 = \|\boldsymbol{z}_1 + \boldsymbol{z}_2\|^2 \leqq 2\alpha^2\,\|\boldsymbol{y}\|^2$.
Im Fall $r = m$ ist $\boldsymbol{z}_2 = \boldsymbol{o}$, also $\|\boldsymbol{z}\|^2 = \|\boldsymbol{z}_1\|^2 + \|\boldsymbol{z}_3\|^2 \leqq 2\alpha^2\,\|\boldsymbol{y}\|^2$.
Im Fall $r = m = n$ ist $\|\boldsymbol{z}\| = \|\boldsymbol{z}_1\| \leqq \alpha\,\|\boldsymbol{y}\|$.

Ü 8.2.3. Man beweise, daß $\operatorname{cond}(\boldsymbol{A}) = 1$ genau dann gilt, wenn $\boldsymbol{A}$ von der Gestalt

$$\boldsymbol{A} = \sigma\overline{\boldsymbol{U}}\,\overline{\boldsymbol{V}}^{\mathsf{T}} = \sigma\sum_{i=1}^{r}\bar{\boldsymbol{u}}^i\bar{\boldsymbol{v}}^{i\mathsf{T}}$$

mit $\sigma > 0$ und spaltenorthonormalen Matrizen $\overline{\boldsymbol{U}} \in \mathbf{R}^{m,r}$, $\overline{\boldsymbol{V}} \in \mathbf{R}^{n,r}$, $r = \operatorname{rang}(\boldsymbol{A})$ ist.

8.3. Klassifikation und statistische Interpretation von Quadratmittelproblemen

Wir betrachten das Quadratmittelproblem

$$\boldsymbol{A}\boldsymbol{x} \simeq \boldsymbol{b} \quad \text{mit} \quad \boldsymbol{A} \in \mathbf{R}^{m,n}, \qquad \boldsymbol{b} \in \mathbf{R}^m \tag{1}$$

und setzen $m \geqq n$ voraus, d. h., (1) sei das zum überbestimmten und i. allg. inkonsistenten Gleichungssystem

$$\boldsymbol{A}\boldsymbol{x} = \boldsymbol{b} \tag{2}$$

gehörende Ersatzproblem. Vom numerischen Standpunkt lassen sich die Probleme des Typs (1) in zwei Klassen einteilen.

1. *Spaltenreguläre Probleme:* $\operatorname{Rang}(\boldsymbol{A}) = r = n \leqq m$
   Solche Probleme sind stets eindeutig lösbar, und die Lösung $\boldsymbol{x}$ hängt lokal lipschitzstetig von den Eingangsdaten ab. Sie sind also *korrekt gestellt* und können i. allg. befriedigend gelöst werden.
2. *Rangdefiziente Probleme:* $\operatorname{Rang}(\boldsymbol{A}) = r < n \leqq m$
   Solche Probleme haben unendlich viele Lösungen, und die Normallösung als Lösung kleinster Euklidischer Norm hängt unstetig von den Eingangsdaten ab. Sie sind daher *inkorrekt gestellt* und können ohne zusätzliche Information — z. B. über den Rang von $\boldsymbol{A}$, um bösartige rangerhöhende Störungen in $\boldsymbol{A}$ auszuschließen — nicht befriedigend gelöst werden.

Die Quelle für überbestimmte inkonsistente Gleichungssysteme (2) ist meist das lineare statistische Modell

$$\boldsymbol{b} = \boldsymbol{A}\boldsymbol{x}^* + \boldsymbol{\varepsilon}. \tag{3}$$

Dabei stellt ε einen m-dimensionalen zufälligen Vektor mit dem Erwartungswert

$$E(\varepsilon) = 0 \tag{4}$$

und der Kovarianzmatrix

$$E(\varepsilon\varepsilon^T) = \sigma^2 I \tag{5}$$

dar, und für A gelte rang $(A) = n < m$. Der Vektor x^* ist ein nicht zufälliger, unbekannter Parametervektor, d r aus dem Zufallsvektor b geschätzt werden soll. Wenn die eindeutige Lösung

$$x = A^+ b = (A^T A)^{-1} A^T b \tag{6}$$

von (1) als Schätzer für x^* verwendet wird, gilt

$$E(x) = A^+ E(b) = A^+ A x^* = x^*, \tag{7}$$

d. h., x ist eine erwartungstreue Schätzung, und die Kovarianzmatrix ergibt sich zu

$$E\big((x - x^*)(x - x^*)^T\big) = \sigma^2 A^+ A^{+T} = \sigma^2 (A^T A)^{-1}. \tag{8}$$

Bei bekanntem σ^2 sind unter Verwendung der Kovarianzmatrix $\sigma^2(A^T A)^{-1}$ Aussagen über die Qualität der Schätzung x möglich. Ist σ^2 nicht bekannt, so kann

$$\|b - Ax\|^2/(m - n) \tag{9}$$

als Schätzung für σ^2 genommen werden. Falls ε einer m-dimensionalen Normalverteilung mit den Parametern (4), (5) genügt, ist x gerade die Maximum-Likelihood-Schätzung. Wir bemerken noch, daß auch nicht erwartungstreue Schätzungen — etwa solche nach dem Prinzip der „ridge regression" — praktische Bedeutung haben, vgl. Kapitel 11.3 und B 11.3.

Für eine detaillierte Beschreibung des statistischen Hintergrundes muß auf die einschlägige statistische Fachliteratur verwiesen werden, siehe B 8.5. Aus den skizzierten Problemen wird jedoch klar, daß neben der Lösung x in der Regel noch weitere numerische Informationen als zusätzliche Ausgangsdaten gewünscht werden, etwa

(I_1) die ausgeglichenen Werte Ax von b und das minimale Residuum $r = b - Ax$, zumindest aber dessen Norm $\|r\|^2$,

(I_2) alle oder einige Elemente von $C := (A^T A)^{-1}$ sowie Ausdrücke der Form BCB^T bzw. $F^T C^{-1} F$ für gegebene Matrizen B bzw. F,

(I_3) der Rang r von A, im Fall $r < n$ die Indizes von r Spalten von A, die $\mathcal{R}(A)$ aufspannen, die Koeffizienten von $n - r$ Abhängigkeitsbeziehungen zwischen den Spalten von A, ggf. eine (orthogonale) Basis von $\mathcal{N}(A)$,

(I_4) die Indizes von möglichst wenig Spalten von A, so daß ein auf diese Spalten reduziertes Modell (3) ein Residuum r mit $\|r\| \leqq \delta$ liefert, wobei δ eine vorgegebene bzw. im Laufe der Rechnung festgelegte Schranke darstellt,

(I_5) ein Schätzwert für $\|A^+\|$ oder eine andere Größe, die Aussagen über die Empfindlichkeit der Lösung gegenüber Störungen zuläßt.

Die Aufgaben (I_3) und (I_4) stehen in Zusammenhang mit der *Modellwahl*. Ziel ist dabei, mit möglichst wenigen Spalten von $\boldsymbol{A}$ einen akzeptablen Ausgleich zu erhalten. Man beachte dabei, daß sich beim Streichen von Spalten aus einer spaltenregulären Matrix die Kondition i. allg. verkleinert, siehe Ü 8.3.3. Eine minimale Spaltenzahl ist daher nicht nur vom statistischen, sondern auch vom numerischen Standpunkt her erwünscht.

Eine Möglichkeit zur Beschaffung der in (I_4) geforderten Information besteht darin, alle Quadratmittelprobleme zu lösen, bei denen $\boldsymbol{A}$ auf k der n Spalten reduziert ist ($k = 1, \ldots, n$), und die zugehörigen Residuen zu vergleichen. Das erfordert die Lösung einer Folge von $2^n - 1$ Quadratmittelproblemen, von denen jedes durch Hinzufügen, Streichen oder Ändern einer Spalte aus dem vorhergegangenen entstanden ist. Man spricht auch von *schrittweiser Regression*, siehe B 8.9. Solche Probleme treten auch bei der Lösung von Quadratmittelproblemen mit linearen Ungleichungsnebenbedingungen auf; zu diesen siehe B 8.8. Da andererseits jeder Zeile von (3) ein Meßwert zugeordnet werden kann, tritt dasselbe Problem bezüglich der Zeilen von (1) beim Hinzufügen, Streichen oder Ändern von Meßwertsätzen auf. Es besteht daher ein Interesse an Algorithmen, die eine Folge von Quadratmittelproblemen, die durch Hinzufügen, Streichen oder Ändern von Spalten bzw. Zeilen auseinander hervorgehen, effektiv zu lösen gestatten, vgl. Abschnitt 2.1.B1 und Ü 2.1.3.

Wir betrachten abschließend den Fall, daß $\boldsymbol{\varepsilon}$ eine von $\sigma^2\boldsymbol{I}$ verschiedene Kovarianzmatrix

$$E(\boldsymbol{\varepsilon}\boldsymbol{\varepsilon}^\top) = \boldsymbol{W} \tag{10}$$

besitzt, wobei $\boldsymbol{W}$ notwendig symmetrisch und positiv semidefinit ist. Wir setzen $\boldsymbol{W}$ zusätzlich als positiv definit voraus. Dann ist die Maximum-Likelihood-Schätzung für $\boldsymbol{x}^*$ die Lösung des sog. *gewichteten* oder *verallgemeinerten Quadratmittelproblems*

$$\boldsymbol{\varepsilon}^\top \boldsymbol{W}^{-1}\boldsymbol{\varepsilon} = \boldsymbol{r}^\top \boldsymbol{W}^{-1}\boldsymbol{r} = (\boldsymbol{b} - \boldsymbol{A}\boldsymbol{x})^\top \boldsymbol{W}^{-1}(\boldsymbol{b} - \boldsymbol{A}\boldsymbol{x}) \to \underset{\boldsymbol{x}\in\mathbb{R}^n}{\text{Minimum!}} \tag{11}$$

Die Wichtung der Residuen mit $\boldsymbol{W}^{-1}$ kann daher als natürliche Wichtung angesehen werden.

Wir wollen (11) auf ein ungewichtetes Quadratmittelproblem zurückführen und verwenden dazu die nach dem Cholesky-Verfahren mit $\sim m^3/6$ opms berechenbare symmetrische Dreiecksfaktorisierung

$$\boldsymbol{W} = \boldsymbol{L}\boldsymbol{L}^\top \tag{12}$$

von $\boldsymbol{W}$, wobei $\boldsymbol{L}$ eine untere Dreiecksmatrix mit positiven Diagonalelementen ist, vgl. 5.2.D und 6.1.A. Mit den Transformationen

$$\bar{\boldsymbol{b}} := \boldsymbol{L}^{-1}\boldsymbol{b}, \qquad \bar{\boldsymbol{A}} := \boldsymbol{L}^{-1}\boldsymbol{A} \tag{13}$$

und

$$\bar{\boldsymbol{r}} := \boldsymbol{L}^{-1}\boldsymbol{r} = \boldsymbol{L}^{-1}(\boldsymbol{b} - \boldsymbol{A}\boldsymbol{x}) = \bar{\boldsymbol{b}} - \bar{\boldsymbol{A}}\boldsymbol{x} \tag{14}$$

geht (11) in das ungewichtete Problem

$$\boldsymbol{r}^\top \boldsymbol{W}^{-1}\boldsymbol{r} = \bar{\boldsymbol{r}}^\top \boldsymbol{I}\bar{\boldsymbol{r}} = (\bar{\boldsymbol{b}} - \bar{\boldsymbol{A}}\boldsymbol{x})^\top (\bar{\boldsymbol{b}} - \bar{\boldsymbol{A}}\boldsymbol{x}) = \|\bar{\boldsymbol{b}} - \bar{\boldsymbol{A}}\boldsymbol{x}\|^2 \to \underset{\boldsymbol{x}\in\mathbb{R}^n}{\text{Minimum!}} \tag{15}$$

über. Die transformierten gewichteten Größen werden dabei direkt aus den Dreieckssystemen

$$\boldsymbol{L}\bar{\boldsymbol{b}} = \boldsymbol{b}, \qquad \boldsymbol{L}\bar{\boldsymbol{A}} = \boldsymbol{A}$$

mit $\sim m^2 n/2$ opms berechnet. (Die Bezeichnung $\bar{\boldsymbol{b}}$ hat hier nichts mit irgendwelchen Projektionen zu tun, sondern wurde in Anlehnung an 4.1.C gewählt.)

Für die meisten praktischen Ausgleichsaufgaben ist $\boldsymbol{W}$ diagonal, d. h.

$$\boldsymbol{W} = \operatorname{diag}(w_i), \qquad w_i > 0 \quad (i = 1, \ldots, m). \tag{16}$$

Das bedeutet im Fall der Normalverteilung, daß die Komponenten ε_i des Fehlervektors unabhängig sind und die i. allg. unterschiedlichen Varianzen $E(\varepsilon_i^2) = w_i$ haben. Dies entspricht unabhängigen Messungen mit unterschiedlicher Genauigkeit. Dann ist

$$\boldsymbol{L} = \operatorname{diag}\left(\sqrt{w_i}\right) \qquad (i = 1, \ldots, m)$$

ebenfalls diagonal, und mit der Skalierungsmatrix

$$\boldsymbol{D} = \operatorname{diag}(d_i) := \boldsymbol{L}^{-1}, \qquad d_i := 1/\sqrt{w_i} \qquad (i = 1, \ldots, m) \tag{17}$$

geht (15) in

$$\|\bar{\boldsymbol{b}} - \bar{\boldsymbol{A}}\boldsymbol{x}\|^2 = \|\boldsymbol{D}(\boldsymbol{b} - \boldsymbol{A}\boldsymbol{x})\|^2 = \|\boldsymbol{D}\boldsymbol{b} - \boldsymbol{D}\boldsymbol{A}\boldsymbol{x}\|^2 \to \underset{\boldsymbol{x}\in\mathbf{R}^n}{\text{Minimum!}} \tag{18}$$

über. Die i-te Zeile des Originalproblems ist also mit $1/\sqrt{w_i}$ zu multiplizieren.

Wenn die Kovarianzmatrix $\boldsymbol{W}$ schlecht konditioniert ist, wird das gewichtete Problem (11) *steif* genannt. Mit den Eigenwerten ω_i von $\boldsymbol{W}$ läßt sich dies durch

$$\operatorname{cond}(\boldsymbol{W}) = \max \omega_i / \min \omega_i \gg 1 \tag{19}$$

charakterisieren, d. h., $\boldsymbol{W}$ hat Eigenwerte stark unterschiedlicher Größenordnung. Für diagonales $\boldsymbol{W}$ ist $w_i = \omega_i$, so daß (19) in

$$\max w_i \gg \min w_i \tag{20}$$

übergeht. Dieser Fall tritt auf, wenn gewisse Komponenten von $\boldsymbol{b}$ wesentlich genauer gemessen werden können als die übrigen. Man beachte, daß im Fall $d_l = 1/\sqrt{w_l} \to \infty$ für ein festes $l \in \{1, \ldots, m\}$ die Lösung von (18) in die Lösung des Quadratmittelproblems

$$\sum_{\substack{i=1 \\ i\neq l}}^{m} [(\boldsymbol{b} - \boldsymbol{A}\boldsymbol{x})_i]^2 / w_i \to \text{Minimum!} \quad \text{bei} \quad \boldsymbol{x} \in \mathbf{R}^n, \qquad (\boldsymbol{b} - \boldsymbol{A}\boldsymbol{x})_l = 0 \tag{21}$$

mit Gleichungsnebenbedingungen übergeht.

Für steife Probleme mit nichtdiagonalem $\boldsymbol{W}$ kann die explizite Ausführung der Standardtransformation (13) nicht empfohlen werden, da sie i. allg. zu einem hohen Genauigkeitsverlust führt. Wir gehen auf solche Probleme nicht weiter ein, siehe B 8.10 und die dort angegebene Literatur.

Übungsaufgaben

Ü 8.3.1. Man zeige:
(i) Es sei

$$\boldsymbol{W} = \boldsymbol{U}\boldsymbol{\Omega}\boldsymbol{U}^{\mathsf{T}}, \qquad \boldsymbol{\Omega} = \operatorname{diag}(\omega_i) > 0, \qquad \boldsymbol{U} \text{ orthogonal},$$

die Eigenwertzerlegung der symmetrischen und positiv definiten Matrix $\boldsymbol{W}$. Dann gilt

$$\boldsymbol{W} = \boldsymbol{W}^{1/2}\boldsymbol{W}^{1/2}, \tag{22}$$

wobei

$$\boldsymbol{W}^{1/2} := \boldsymbol{U}\boldsymbol{\Omega}^{1/2}\boldsymbol{U}^{\mathsf{T}}, \qquad \boldsymbol{\Omega}^{1/2} := \operatorname{diag}\left(\sqrt{\omega_i}\right) \tag{23}$$

die eindeutig festgelegte positiv definite Quadratwurzel aus $\boldsymbol{W}$ bezeichnet.

(ii) Mit $\hat{\boldsymbol{D}} := \boldsymbol{\Omega}^{-1/2} = \operatorname{diag}\left(1/\sqrt{\omega_i}\right)$ und

$$\hat{\boldsymbol{b}} := \boldsymbol{U}^{\mathsf{T}}\boldsymbol{b}, \qquad \hat{\boldsymbol{A}} := \boldsymbol{U}^{\mathsf{T}}\boldsymbol{A}, \qquad \hat{\boldsymbol{r}} := \boldsymbol{U}^{\mathsf{T}}\boldsymbol{r} \tag{24}$$

geht (11) in das diagonal gewichtete Quadratmittelproblem

$$\boldsymbol{r}^{\mathsf{T}}\boldsymbol{W}^{-1}\boldsymbol{r} = \|\boldsymbol{W}^{-1/2}\boldsymbol{r}\|^2 = \|\hat{\boldsymbol{D}}(\hat{\boldsymbol{b}} - \hat{\boldsymbol{A}}\boldsymbol{x})\|^2 \to \text{Minimum!} \tag{25}$$

über.

Ü 8.3.2. Man überlege sich, daß (15) unter Verwendung von (12) in der Form

$$\|\boldsymbol{z}\|^2 \to \text{Minimum!} \quad \text{bei } \boldsymbol{x} \in \mathbf{R}^n, \boldsymbol{z} \in \mathbf{R}^m \text{ mit } \boldsymbol{b} = \boldsymbol{A}\boldsymbol{x} + \boldsymbol{L}\boldsymbol{z} \tag{26}$$

als Quadratmittelproblem der Dimension $n + m$ mit Gleichungsnebenbedingungen geschrieben werden kann.

Ü 8.3.3. Wenn $\boldsymbol{B} \in \mathbf{R}^{m,n-1}$ aus $\boldsymbol{A} \in \mathbf{R}^{m,n}$, $m \geqq n \geqq 2$, durch Streichen einer Spalte entsteht, gilt für die geordneten Singulärwerte β_i von $\boldsymbol{B}$ und α_i von $\boldsymbol{A}$ die Einschließung

$$\alpha_1 \geqq \beta_1 \geqq \alpha_2 \geqq \beta_2 \geqq \cdots \geqq \beta_{n-1} \geqq \alpha_n \geqq 0,$$

siehe Lawson/Hanson [74, Theorem 5.12]. Man folgere hieraus im Fall rang $(\boldsymbol{A}) = n$

$$\operatorname{cond}(\boldsymbol{B}) \leqq \operatorname{cond}(\boldsymbol{A}).$$

Bemerkungen zum Kapitel 8

B 8.1. Die Lösung des Minimierungsproblems $\|\boldsymbol{b} - \boldsymbol{A}\boldsymbol{x}\|_p \to \text{Minimum!}$ heißt L_p-*Lösung* des Gleichungssystems $\boldsymbol{A}\boldsymbol{x} = \boldsymbol{b}$; für $p = 1$ spricht man auch von einer *Lösung kleinster absoluter Abweichungen*, für $p = \infty$ von einer *Čebyšev-Lösung*. Die mit der ∞-Norm beabsichtigte Äquilibrierung des Residuums $\boldsymbol{r}$ kann meist auch billiger in der 2-Norm durch geeignete (unter Umständen adaptive) Skalierung erreicht werden. Man beachte dabei, daß die Berechnung der L_∞-Lösung zur Lösung eines speziellen linearen Optimierungsproblems äquivalent und daher i. allg. aufwendig ist. Die L_1-Lösungen haben gegenüber den Quadratmittellösungen den Vorteil, nicht so empfindlich auf einzelne große Komponenten von $\boldsymbol{r}$ zu reagieren, also „robuster" gegenüber „Ausreißern" zu sein. Dem steht die weitaus größere Kompliziertheit in der theoretischen Durchdringung und praktischen Berechnung gegenüber.

B 8.2. Das für identisch normalverteilte Fehler optimale Prinzip der kleinsten Fehlerquadrate wurde 1809 von Gauss bei geodätischen und astronomischen Berechnungen verwendet und trägt deshalb auch seinen Namen. Unabhängig ist es bereits 1806 von Lagrange eingeführt

worden, was damals zu einem erbitterten Prioritätsstreit geführt hat, siehe etwa LINNIK [62] für einen historischen Abriß.

B 8.3. Die als Pseudoinverse bezeichnete Matrix wurde unabhängig von MOORE 1920 und PENROSE 1955 über unterschiedliche Definitionen eingeführt, die sich später als gleichwertig herausstellten. Zur Geschichte der Pseudoinversen siehe ZIELKE [78], wo auch viele nützliche Literaturhinweise zu finden sind.

B 8.4. Die Darstellung der Störungsaussagen für A^+ und x^N folgt der bei LAWSON/HANSON [74]. Solche Aussagen sind für beliebiges $r = \text{Rang}(A)$ erstmals von WEDIN [73] bewiesen worden, obwohl für gewisse Sonderfälle auch vorher schon analoge Ergebnisse bekannt waren. Eine umfangreiche Zusammenfassung gibt STEWART [77]; dort ist auch Lemma 8.2.11 in einer im wesentlichen äquivalenten Form enthalten.

B 8.5. Über den statistischen Hintergrund linearer Quadratmittelprobleme kann man sich bei LINNIK [62], DRAPER/SMITH [66], K. M. S. HUMAK [77, 83] u. a. informieren.

B 8.6. Wenn im linearen Modell auch die Matrix A durch zufällige Fehler verfälscht ist, sollte neben x auch A in die Schätzung einbezogen und ein ausgeglichener Wert B für A aus dem erweiterten Problem

$$\text{(TLS)} \qquad \|b - Bx\|^2 + \|A - B\|_F^2 \to \text{Minimum!} \quad \text{bei } x \in \mathbf{R}^n, B \in \mathbf{R}^{m,n}$$

bestimmt werden. Solche Probleme werden *totale Quadratmittelprobleme* genannt. Wegen des quadratischen Terms Bx lassen sich diese nicht direkt auf gewöhnliche Quadratmittelprobleme zurückführen und erfordern spezielle Lösungsverfahren, siehe GOLUB/VAN LOAN [80]. Probleme des Typs (TLS) entstehen z. B. bei der Anpassung der skalaren linearen Modellgleichung

$$b = a_1 x_1 + \cdots + a_n x_n$$

an m Meßwertsätze $\{b_i, (a_1)_i, \ldots, (a_n)_i\}$, wenn sowohl b_i als auch die $(a_j)_i$ fehlerbehaftet sind. Man spricht dann auch von *linearen Modellen mit Fehlern in den Variablen.*

B 8.7. Gauß-Newton-Verfahren zur Lösung nichtlinearer Quadratmittelprobleme

$$\|r(x)\| \to \text{Minimum!} \quad \text{bei } x \in \mathbf{R}^n$$

mit einer nichtlinearen Vektorfunktion $r\colon \mathbf{R}^n \to \mathbf{R}^m$ erfordern in jedem Schritt die Lösung eines linearen Quadratmittelproblems, siehe BARD [73], SCHWETLICK [79], BJÖRCK [81] und GILL/MURRAY/WRIGHT [82]. Nichtlineare Modelle mit Fehlern in den Variablen führen auf lineare Quadratmittelprobleme spezieller Struktur, die angepaßte Lösungsverfahren benötigen, siehe SCHWETLICK/TILLER [85].

B 8.8. In gewissen praktischen Aufgaben treten zusätzliche Gleichungs- oder Ungleichungsnebenbedingungen auf, die zu Quadratmittelproblemen des Typs

$$\text{(ELS)} \qquad \|b - Ax\| \to \text{Minimum} \quad \text{bei } x \in \mathbf{R}^n, Cx = d$$

bzw.

$$\text{(ILS)} \qquad \|b - Ax\| \to \text{Minimum} \quad \text{bei } x \in \mathbf{R}^n, Cx = d, Ex \geqq f$$

führen. Während (ELS) durch geeignete Transformationen auf ein Problem ohne Nebenbedingungen zurückgeführt werden kann, siehe LAWSON/HANSON [74, Chapt. 20–22], erfordert (ILS) den Einsatz von Optimierungstechniken und ist wesentlich aufwendiger zu lösen, siehe wieder LAWSON/HANSON [74, Chapt. 23] und STOER [71], SCHITTKOWSKI/STOER [79], HASKELL/HANSON [81], LÖTSTEDT [84] für geeignete Verfahren. Wir gehen auf beide Aufgabenklassen nicht weiter ein.

B 8.9. Algorithmen zur schrittweisen Regression sind auf der Basis der Normalgleichungen von EFROYMSON [60], auf der Basis der Orthogonalisierungsverfahren von ELDÉN [72] und LAWSON/HANSON [74] beschrieben worden, siehe auch OSBORNE [76].

B 8.10. Steife diagonalgewichtete Quadratmittelprobleme werden bei LAWSON/HANSON [74, Chapt. 17] und BJÖRCK [78] diskutiert. Normalgleichungsverfahren sind für solche Probleme nicht geeignet, während $\boldsymbol{QR}$-Faktorisierungen von $\boldsymbol{DA}$ mit geeigneten Zusatzmaßnahmen zu brauchbaren Resultaten führen, siehe 10.2.5.

Für nichtdiagonales, schlechtkonditioniertes $\boldsymbol{W}$ empfiehlt BJÖRCK [81] zwei Vorgehensweisen:

- Ersatz der Cholesky-Faktorisierung (8.3.12) durch (8.3.22) unter Verwendung der Eigenwertzerlegung von $\boldsymbol{W}$ und Rückführung auf das diagonalgewichtete Problem (8.3.25) gemäß Ü 8.3.1,
- Formulierung als nichtgewichtetes Problem mit Gleichungsnebenbedingungen gemäß Ü 8.3.2 und Lösung mittels angepaßter Orthogonalisierungstechniken nach PAIGE [79].

9. Normalgleichungsverfahren

Normalgleichungsverfahren beruhen auf der Charakterisierung der Lösungen des Quadratmittelproblems $\boldsymbol{Ax} \cong \boldsymbol{b}$ als Lösungen der zugehörigen Normalgleichungen

$$\boldsymbol{A}^T\boldsymbol{Ax} = \boldsymbol{A}^T\boldsymbol{b}.$$

9.1. Aufgabenstellung und Lösung der Normalgleichungen

Wir bemerken zunächst, daß die Normalgleichungen im Fall $\boldsymbol{A} \in \mathbf{R}^{m,n}$ ein quadratisches, stets konsistentes Gleichungssystem der Ordnung n darstellen. Die Koeffizientenmatrix ist für jedes $\boldsymbol{A}$ symmetrisch und positiv semidefinit; sie ist positiv definit genau dann, wenn $\boldsymbol{A}$ spaltenregulär ist, vgl. 1.2.14 und den nachfolgenden Text. Wir setzen $\boldsymbol{A}$ daher im folgenden als spaltenregulär voraus. Es bietet sich dann an, die eindeutige Lösung der Normalgleichungen durch Cholesky-Faktorisierung von $\boldsymbol{A}^T\boldsymbol{A}$ zu berechnen.

9.1.1. *Lösung von $\boldsymbol{Ax} \cong \boldsymbol{b}$ über die Normalgleichungen mittels Cholesky-Faktorisierung.*

Aufgabe. Für spaltenreguläres $\boldsymbol{A} \in \mathfrak{R}^{m,n}$ ist die Koeffizientenmatrix $\boldsymbol{M} = \boldsymbol{A}^T\boldsymbol{A}$ der Normalgleichungen zu berechnen und nach dem Cholesky-Verfahren gemäß $\boldsymbol{M} = \boldsymbol{LL}^T$ mit einer unteren Dreiecksmatrix $\boldsymbol{L} \in \mathfrak{R}^{n,n}$ zu faktorisieren. Zu gegebenem $\boldsymbol{b} \in \mathfrak{R}^m$ ist die rechte Seite $\boldsymbol{c} = \boldsymbol{A}^T\boldsymbol{b}$ der Normalgleichungen

$$\boldsymbol{Mx} = \boldsymbol{c} \tag{1}$$

von $\boldsymbol{Ax} \cong \boldsymbol{b}$ zu berechnen und die zugehörige Lösung $\boldsymbol{x} \in \mathfrak{R}^n$ unter Verwendung von $\boldsymbol{L}$ zu bestimmen.

Algorithmus:

S 1 (Berechnung und Cholesky-Faktorisierung von $\boldsymbol{M}$):

S 1.1: Berechne $\boldsymbol{M} := \boldsymbol{A}^T\boldsymbol{A}$

S 1.2: Bestimme die Cholesky-Faktorisierung $\boldsymbol{M} = \boldsymbol{LL}^T$ von $\boldsymbol{M}$ mit einer unteren Dreiecksmatrix $\boldsymbol{L}$ mit positiven Diagonalelementen.

S2 (Berechnung von $\boldsymbol{c} = \boldsymbol{A}^T\boldsymbol{b}$ und Lösung von $\boldsymbol{L}\boldsymbol{L}^T\boldsymbol{x} = \boldsymbol{c}$):
S2.1: Berechne $\boldsymbol{c} := \boldsymbol{A}^T\boldsymbol{b}$
S2.2: Löse die Dreieckssysteme $\boldsymbol{L}\boldsymbol{d} = \boldsymbol{c}$, $\boldsymbol{L}^T\boldsymbol{x} = \boldsymbol{d}$
Aufwand: S1: $\sim (m + n/3)\, n^2/2$ opms, S2: $\sim (m + n)\, n$ opms

9.1.2. *Bemerkung.* (i) Für praktische Probleme ist m meist deutlich größer als n, so daß der Gesamtaufwand durch den Term $\sim mn^2/2$, d. h. durch die Berechnung von $\boldsymbol{M}$ bestimmt wird. Es ist zweckmäßig, ein Dreieck von $\boldsymbol{M}$ unter Verwendung von zusätzlichen $\sim n^2/2$ S gesondert zu speichern; die Originalmatrix $\boldsymbol{A}$ steht dann weiter zur Verfügung, etwa zur iterativen Verbesserung. Falls ein diagonalgewichtetes Problem $\boldsymbol{D}\boldsymbol{A}\boldsymbol{x} \cong \boldsymbol{D}\boldsymbol{b}$ zu lösen ist, erfordert die Berechnung von $\boldsymbol{D}\boldsymbol{A}$ und $\boldsymbol{D}\boldsymbol{b}$ weitere $\sim mn$ opms, spielt also keine Rolle.

(ii) Die Normalgleichungen (1) können auch nach einem anderen Verfahren gelöst werden, etwa durch schrittweise Diagonalisierung nach dem Gauß-Jordan-Verfahren, vgl. 6.4. Dies empfehlen manche Autoren für die schrittweise Regression, siehe B 8.9.

(iii) In exakter Arithmetik gilt $\boldsymbol{M} = \boldsymbol{A}^T\boldsymbol{A}$, also

$$\|\boldsymbol{A}\|_2^2 = \|\boldsymbol{M}\|_2 \leqq \sum_{i=1}^{n} (\boldsymbol{M})_{ii} = \|\boldsymbol{A}\|_F^2, \tag{2}$$

$$\|\boldsymbol{A}^+\|_2^2 = \|\boldsymbol{M}^{-1}\|_2 \leqq \sum_{i=1}^{n} (\boldsymbol{M}^{-1})_{ii} = \|\boldsymbol{A}^+\|_F^2. \quad \square$$

9.1.3. *Rundungsfehleranalyse.* Für spaltenreguläres $\boldsymbol{A} \in \mathfrak{R}^{m,n}$ und beliebiges $\boldsymbol{b} \in \mathfrak{R}^m$ ist Algorithmus 9.1.1 durchführbar, sofern $\boldsymbol{A}$ im Sinne von

$$\hat{\varkappa} := \nu N_1[\mathrm{cond}\,(\boldsymbol{A})]^2 \leqq 0.5 \quad \text{mit} \quad N_1 := mn + F_1 \sim mn \tag{3}$$

nicht zu schlecht konditioniert ist, wobei

$$F_1 := n^{3/2} + n + n^{1/2}F \sim 2n^{3/2}, \qquad F := n + 1 + 0.5 \ln n \sim n$$

wie in 6.1.5(ii) gesetzt ist. Die berechneten Größen $\boldsymbol{M}, \boldsymbol{L}$ und $\boldsymbol{c}, \boldsymbol{x}$ genügen den Beziehungen

$$\boldsymbol{M} = \boldsymbol{A}^T\boldsymbol{A} + \boldsymbol{\delta}_0\boldsymbol{M}, \qquad \|\boldsymbol{\delta}_0\boldsymbol{M}\| \leqq \nu mn\, \|\boldsymbol{A}\|^2, \tag{4}$$

$$\boldsymbol{c} = (\boldsymbol{A} + \boldsymbol{\delta}_0\boldsymbol{A})^T\boldsymbol{b}, \qquad \|\boldsymbol{\delta}_0\boldsymbol{A}\| \leqq \nu m \sqrt{n}\, \|\boldsymbol{A}\|, \tag{5}$$

$$\boldsymbol{M} + \boldsymbol{\delta}_1\boldsymbol{M} = \boldsymbol{L}\boldsymbol{L}^T, \qquad \|\boldsymbol{\delta}_1\boldsymbol{M}\| \leqq \nu nF\, \|\boldsymbol{M}\| \sim \nu n^{3/2}\, \|\boldsymbol{M}\|, \tag{6}$$

$$(\boldsymbol{M} + \boldsymbol{\delta}_2\boldsymbol{M})\,\boldsymbol{x} = \boldsymbol{c}, \qquad \|\boldsymbol{\delta}_2\boldsymbol{M}\| \leqq \nu F_1\, \|\boldsymbol{M}\| \sim \nu 2n^{3/2}\, \|\boldsymbol{M}\|. \tag{7}$$

Beweis. Aus (2.3.43) folgt $|\boldsymbol{M} - \boldsymbol{A}^T\boldsymbol{A}| \leqq \nu m\, |\boldsymbol{A}^T|\, |\boldsymbol{A}|$, mit (2.3.30) also $\|\boldsymbol{\delta}_0\boldsymbol{M}\| \leqq \|\boldsymbol{\delta}_0\boldsymbol{M}\|_F \leqq \nu m\, \|\boldsymbol{A}\|_F^2 \leqq \nu mn\, \|\boldsymbol{A}\|^2$. Analog ergibt sich (5) aus (2.3.37). Wegen (3), (4) gilt nun

$$\|(\boldsymbol{A}^T\boldsymbol{A})^{-1}\|\, \|\boldsymbol{\delta}_0\boldsymbol{M}\| \leqq \nu mn[\mathrm{cond}\,(\boldsymbol{A})]^2 =: \varkappa \leqq \hat{\varkappa} \leqq 0.5,$$

so daß $\boldsymbol{M}$ nach Ü 4.1.9 positiv definit ist, und aus 4.1.2 folgt $\|\boldsymbol{M}^{-1}\| \leqq \|(\boldsymbol{A}^T\boldsymbol{A})^{-1}\|/(1 - \varkappa)$ Mit $\|\boldsymbol{M}\| \leqq (1 + \nu mn)\, \|\boldsymbol{A}\|^2 \doteq \|\boldsymbol{A}\|^2$ zieht dies $\mathrm{cond}\,(\boldsymbol{M}) \leqq [\mathrm{cond}\,(\boldsymbol{A})]^2/(1 - \varkappa)$ nach sich Da-

mit erhalten wir

$$vF\,\|\boldsymbol{M}^{-1}\|_2\,\|\boldsymbol{M}\|_F \leqq v\sqrt{n}\,F\,\mathrm{cond}\,(\boldsymbol{M}) \leqq vF_1\,\mathrm{cond}\,(\boldsymbol{M}) \leqq vF_1[\mathrm{cond}\,(\boldsymbol{A})]^2/(1-\varkappa)$$
$$= (\hat{\varkappa}-\varkappa)/(1-\varkappa) \leqq (0.5-\varkappa)/(1-\varkappa) \leqq 0.5.$$

Die Bedingung (6.1.7) ist daher für $\boldsymbol{M}$ statt $\boldsymbol{A}$ erfüllt, und (6), (7) ergeben sich dann sofort aus 6.1.3 und 6.1.5(ii). □

9.1.4. *Bemerkung.* (i) Die Bedingung (3) garantiert, daß die berechnete Matrix $\boldsymbol{M}$ positiv definit und die Cholesky-Faktorisierung durchführbar ist. Eine Mindestforderung an $\boldsymbol{A}$ ist dagegen, daß die Lösung $\boldsymbol{x}$ stetig von Störungen in der Größenordnung des Darstellungsfehlers abhängt. Für spaltenreguläres $\boldsymbol{A}$ führt dies nach 8.2.7 auf $\|\boldsymbol{A}^+\|\,\|\delta\boldsymbol{A}_D\| < 1$. Da der Darstellungsfehler $\delta\boldsymbol{A}_D$ nach 2.3.15 durch $\|\delta\boldsymbol{A}_D\| \leqq v\sqrt{n}\,\|\boldsymbol{A}\|$ abgeschätzt werden kann, ergibt sich

$$v\sqrt{n}\,\mathrm{cond}\,(\boldsymbol{A}) < 1 \tag{8}$$

als Bedingung dafür, daß das Problem $\boldsymbol{A}\boldsymbol{x} \cong \boldsymbol{b}$ bezüglich des Fehlerniveaus v korrekt gestellt ist. Das Auftreten des Faktors $[\mathrm{cond}\,(\boldsymbol{A})]^2$ in (3) schränkt also die Unterklasse der vernünftig gestellten Probleme, für die 9.1.1 durchführbar ist, wesentlich ein.

(ii) Das folgende Beispiel zeigt, daß auf eine Bedingung des Typs (3) nicht verzichtet werden kann. Für

$$\boldsymbol{A} = \begin{pmatrix} 1 & 1 & 1 \\ \varepsilon & 0 & 0 \\ 0 & \varepsilon & 0 \\ 0 & 0 & \varepsilon \end{pmatrix} \quad \text{ist} \quad \boldsymbol{M}^* = \boldsymbol{A}^\top\boldsymbol{A} = \begin{pmatrix} 1+\varepsilon^2 & 1 & 1 \\ 1 & 1+\varepsilon^2 & 1 \\ 1 & 1 & 1+\varepsilon^2 \end{pmatrix},$$

und es gilt $[\mathrm{cond}\,(\boldsymbol{A})]^2 \sim 3/\varepsilon^2$, denn $\boldsymbol{A}^\top\boldsymbol{A}$ hat die Eigenwerte $\{3+\varepsilon^2, \varepsilon^2, \varepsilon^2\}$. Für $\varepsilon \in \mathfrak{R}$ mit $10v^2 \leqq \varepsilon^2 < v/2$ ist (3) verletzt und $\boldsymbol{M} = \mathrm{fl}\,(\boldsymbol{A}^\top\boldsymbol{A})$ wegen $\mathrm{fl}\,(1+\varepsilon^2) = 1$ singulär, aber $\boldsymbol{A}$ erfüllt (8). □

9.1.5. *Folgerung.* Unter den Voraussetzungen von 9.1.3 genügt der erzeugte Rundungsfehler $\delta\boldsymbol{x} = \boldsymbol{x} - \boldsymbol{x}^*$, $\boldsymbol{x}^* = \boldsymbol{A}^+\boldsymbol{b} = (\boldsymbol{A}^\top\boldsymbol{A})^{-1}\,\boldsymbol{A}^\top\boldsymbol{b}$, der Abschätzung

$$\|\delta\boldsymbol{x}\| \leqq \frac{v[\mathrm{cond}\,(\boldsymbol{A})]^2}{1-\hat{\varkappa}}\left\{N_1\,\|\boldsymbol{x}^*\| + N_2\,\frac{\|\boldsymbol{b}\|}{\|\boldsymbol{A}\|}\right\}, \quad N_2 := m\sqrt{n}. \tag{9}$$

Beweis. Es gilt nach 9.1.3

$$(\boldsymbol{A}^\top\boldsymbol{A} + \delta\boldsymbol{M})\,\boldsymbol{x} = \boldsymbol{A}^\top\boldsymbol{b} + \delta\boldsymbol{c} \quad \text{mit} \quad \delta\boldsymbol{M} := \delta_0\boldsymbol{M} + \delta_2\boldsymbol{M}, \qquad \delta\boldsymbol{c} := \delta_0\boldsymbol{A}^\top\boldsymbol{b}, \tag{10}$$

wobei $\|\delta\boldsymbol{M}\| \leqq vN_1\,\|\boldsymbol{A}\|^2$ und $\|\delta\boldsymbol{c}\| \leqq vm\sqrt{n}\,\|\boldsymbol{A}\|\,\|\boldsymbol{b}\|$ ist. Dies führt wegen $\|(\boldsymbol{A}^\top\boldsymbol{A})^{-1}\|\,\|\delta\boldsymbol{M}\| \leqq vN_1[\mathrm{cond}\,(\boldsymbol{A})]^2 = \hat{\varkappa} < 1$ und 4.1.3 sofort auf (9). □

Die Abschätzung (9) ist in dem Sinne realistisch, daß eine Schranke der Form

$$\|\delta\boldsymbol{x}\| \leqq \frac{vN[\mathrm{cond}\,(\boldsymbol{A})]^2}{1-\hat{\varkappa}}\left\{\|\boldsymbol{x}^*\| + \frac{\|\boldsymbol{b}\|}{\|\boldsymbol{A}\|}\right\} =: \frac{\Delta\boldsymbol{x}'_{\mathrm{gen}}(\boldsymbol{A}, \boldsymbol{b}, v)}{1-\hat{\varkappa}} \tag{11}$$

mit $1 \leqq N \leqq N_1$ sicher angenommen wird.

Wir wollen im folgenden untersuchen, ob 9.1.1 ein numerisch stabiles Verfahren ist. Die zum Fehlerniveau ν gehörenden Darstellungsfehler sind nach 2.1.13 durch $\|\delta A\| \leqq \nu \sqrt{n}\, \|A\|$, $\|\delta b\| \leqq \nu \|b\|$ beschränkt. Für die zugehörigen linearisierten Störungen $\delta x'$ gilt nach 8.2.7

$$\|\delta x'\| \leqq \nu \operatorname{cond}(A) \left\{ \sqrt{n} \left[\|x^*\| + \operatorname{cond}(A) \frac{\|r\|}{\|A\|} \right] + \frac{\|b\|}{\|A\|} \right\} =: \Delta x'_{\min}(A, b, \nu), \tag{12}$$

und diese Schranke ist realistisch. Mit der Schranke $\Delta x'_{\text{gen}}$ aus (11) für den linearisierten erzeugten Rundungsfehler $\delta x'_{\text{gen}}$ folgt dann

$$\frac{\Delta x'_{\text{gen}}(A, b, \nu)}{\Delta x'_{\min}(A, b, \nu)} \geqq \frac{N}{\sqrt{n}} \operatorname{cond}(A) \Big/ \left\{ 1 + \frac{\|r\| \operatorname{cond}(A)}{\|A\| \, \|A^+ b\| + \|b\|} \right\}.$$

Die untere Schranke kann so groß wie $\sim 1/\sqrt{\nu}$ werden, denn in der durch (3) charakterisierten Unterklasse spaltenregulärer Quadratmittelprobleme gibt es solche, für die $\operatorname{cond}(A) \sim 1/\sqrt{\nu}$ und trotzdem

$$\|r\| \operatorname{cond}(A) \leqq N_3(\|A\| \, \|A^+ b\| + \|b\|)$$

gilt, nämlich die mit genügend kleinem $\|r\|$ oder die mit $\|A^+ b\|$ in der Größenordnung von $\|A^+\| \, \|b\|$. Daher kann $\Delta x'_{\text{gen}}/\Delta x'_{\min}$ nicht durch eine von A und ν unabhängige Zahl F nach oben beschränkt werden. Da die linearisierten Schranken für $\delta A \to O$ die korrekten Schranken beliebig genau approximieren, liegt keine Stabilität vor.

9.1.6. *Aussage.* Über der Klasse der spaltenregulären Quadratmittelprobleme, die der Bedingung (3) genügen, ist Algorithmus 9.1.1 nicht numerisch stabil.

9.1.7. *Bemerkung.* (i) Aus den angegebenen Abschätzungen folgt andererseits, daß 9.1.1 über der Unterklasse der spaltenregulären Quadratmittelprobleme, für die neben (3) noch

$$\|r\| \geqq \|b\|/c_1 \quad \text{und} \quad \|x^*\| = \|A^+ b\| \leqq c_2 \|b\|/\|A\| \tag{13}$$

mit Konstanten $c_1, c_2 > 0$ gilt, der Quotient $\Delta x'_{\text{gen}}/\Delta x'_{\min}$ durch

$$F \leqq c_1(1 + c_2) N_1 \tag{14}$$

beschränkt ist. Dies bedeutet: *Für stark inkonsistente Probleme mit kleinen Lösungen ist* 9.1.1 *numerisch stabil;* es kann sogar numerische Gutartigkeit gezeigt werden. Aus (13) folgt

$$K_A \geqq \operatorname{cond}(A) + [\operatorname{cond}(A)]^2/(c_1 c_2), \tag{15}$$

d. h., die Unterklasse besteht aus denjenigen Quadratmittelproblemen, für die K_A durch $[\operatorname{cond}(A)]^2$ bestimmt wird, vgl. 8.2.8(ii).

(ii) Wenn z und x wie in 5.4.7, Fall 3, bestimmt werden (dort ist L_2 durch L zu ersetzen), gilt

$$\zeta_2 := \|z\|_2/\|x\|_2 \leqq \|(L^T L)^{-1}\| \doteqdot \|M^{-1}\| \doteqdot \|A^+\|^2. \tag{16}$$

Andererseits ist

$$\eta_2 := \sum_{i=1}^{n} (M)_{ii} \doteqdot \|L\|_F^2 \doteqdot \|A\|_F^2 \leqq n\,\|A\|^2, \tag{17}$$

so daß $\zeta_2\eta_2/n$ eine untere Schranke für $[\mathrm{cond}\,(A)]^2$ darstellt, welche die Größenordnung richtig wiedergibt. Falls die für die Gültigkeit von (3) notwendige und leicht überprüfbare Bedingung

$$\hat{\varkappa} := vm\zeta_2\eta_2 < 1$$

nicht erfüllt ist, muß damit gerechnet werden, daß der nach 9.1.1 berechnete Cholesky-Faktor L wie auch die Lösung x selbst unbrauchbar sein können.

(iii) Die Nachteile von 9.1.1 lassen sich beheben, wenn die Schritte S1 und S2 bei einfachgenauen Eingangsdaten $\{A, b\}$ mit höherer Genauigkeit $v_1 \leqq v^2$ realisiert werden und x danach wieder auf einfache Genauigkeit v gerundet wird. In diesem Fall kann $[\mathrm{cond}\,(A)]^2$ in (3) durch $\mathrm{cond}\,(A)$ ersetzt werden, und das berechnete x ist Lösung von $(A + \delta A)\,x \cong b + \delta b$ mit $\|\delta A\| \leqq vF\,\|A\|$, $\|\delta b\| \leqq vF\,\|b\|$, $F \sim mn$, d. h., es liegt numerische Gutartigkeit vor. Dazu sind nur $\sim n^2/2$ doppeltgenaue Speicherplätze erforderlich, allerdings erhöht sich der arithmetische Aufwand beträchtlich.

(iv) Für steife diagonalgewichtete Probleme $DAx \cong Db$ sind Normalgleichungsverfahren nicht geeignet. Als Modellfall betrachten wir die Gewichtsmatrix $D = \left(\begin{array}{c|c} wI_{m1} & \\ \hline & I_{m2} \end{array}\right)$ und das entsprechend partitionierte $A = \left(\frac{A_1}{A_2}\right)$ mit $w \gg 1$ und $1 \leqq m_1 < n$. Die Matrix der Normalgleichungen ist dann

$$(DA)^T DA = w^2 A_1^T A_1 + A_2^T A_2;$$

sie unterscheidet sich für großes w nicht von der singulären Matrix $w^2 A_1^T A_1$. □

Übungsaufgaben

Ü 9.1.1. Es sei $A \in \mathbf{R}^{m,n}$ spaltenregulär, und die Cholesky-Faktorisierung $A^T A = LL^T$ sei bekannt. Mit C werde $(A^T A)^{-1} = L^{-T}L^{-1}$ bezeichnet, vgl. 8.3. Man zeige:
(i) Für $F \in \mathbf{R}^{n,l}$ kann $X := F^T C^{-1} F = F^T (A^T A)\,F$ gemäß

$$X = F^T LL^T F = \bar{F}^T \bar{F} \quad \text{mit} \quad \bar{F} := L^T F \tag{18}$$

in $\sim (n + l)\,ln/2$ opms berechnet werden, während die direkte Berechnung gemäß

$$X = F^T A^T AF = \hat{F}^T \hat{F} \quad \text{mit} \quad \hat{F} := AF$$

$\sim (2n + l)\,lm/2$ opms erfordert, also mindestens das (m/n)-fache des Aufwands von (18).

(ii) Für $\boldsymbol{B} \in \mathbf{R}^{l,n}$ kann $\boldsymbol{Y} := \boldsymbol{BCB}^T$ ohne explizite Bildung von $\boldsymbol{C} = (\boldsymbol{A}^T\boldsymbol{A})^{-1}$ gemäß

$$\boldsymbol{Y} = \boldsymbol{BL}^{-T}\boldsymbol{L}^{-1}\boldsymbol{B} = \bar{\boldsymbol{B}}\bar{\boldsymbol{B}}^T \quad \text{mit} \quad \bar{\boldsymbol{B}} := \boldsymbol{BL}^{-T} \tag{19}$$

berechnet werden, wobei $\bar{\boldsymbol{B}}$ aus dem Dreieckssystem

$$\bar{\boldsymbol{B}}\boldsymbol{L}^T = \boldsymbol{B} \quad \text{bzw.} \quad \boldsymbol{L}\bar{\boldsymbol{B}}^T = \boldsymbol{B}^T$$

bestimmt wird. Dies erfordert insgesamt $\sim (n + l)\, ln/2$ opms. Für $\boldsymbol{B} := \boldsymbol{A}$ ist

$$\boldsymbol{Y} = \boldsymbol{A}(\boldsymbol{A}^T\boldsymbol{A})^{-1}\boldsymbol{A} = (\boldsymbol{AL}^{-T})\,(\boldsymbol{AL}^{-T})^T$$

der Projektor $\boldsymbol{P}_{\boldsymbol{A}}$ auf $\mathcal{R}(\boldsymbol{A})$.

Ü 9.1.2. Man beweise: Wenn

$$\bar{\boldsymbol{A}} := \boldsymbol{A} + \boldsymbol{uv}^T, \qquad \boldsymbol{u} \neq \boldsymbol{o}, \qquad \boldsymbol{v} \neq \boldsymbol{o}, \tag{20}$$

durch Rang-1-Änderung aus $\boldsymbol{A}$ entsteht, gilt

$$\bar{\boldsymbol{A}}^T\bar{\boldsymbol{A}} = \boldsymbol{A}^T\boldsymbol{A} - \boldsymbol{yy}^T + \boldsymbol{zz}^T \quad \text{mit} \quad \boldsymbol{y} := \boldsymbol{A}^T\boldsymbol{u}/\|\boldsymbol{u}\|, \qquad \boldsymbol{z} := \boldsymbol{y} + \|\boldsymbol{u}\|\,\boldsymbol{v}, \tag{21}$$

d. h., $\bar{\boldsymbol{A}}^T\bar{\boldsymbol{A}}$ entsteht aus $\boldsymbol{A}^T\boldsymbol{A}$ durch eine Rang-2-Änderung, die in der symmetrischen Form (21) als Differenz zweier positiv semidefiniter Rang-1-Terme geschrieben werden kann.

9.2. Iterative Verbesserung

Im vorangegangenen Abschnitt wurde gezeigt, daß das Normalgleichungsverfahren 9.1.1 zwei wesentliche Nachteile aufweist, nämlich:

- Die Klasse der durch 9.1.1 lösbaren spaltenregulären Quadratmittelprobleme wird durch die Bedingung (3) wesentlich eingeschränkt.
- Das Verfahren ist numerisch instabil.

Wir zeigen im folgenden, daß der zweite Nachteil durch iterative Verbesserung analog zu 5.4.3 behoben werden kann; der erste Nachteil bleibt dabei leider bestehen.

9.2.1. *Normalgleichungsverfahren mit iterativer Verbesserung*

Aufgabe: Für gegebenes spaltenreguläres $\boldsymbol{A} \in \mathfrak{R}^{m,n}$ sei $\boldsymbol{L} \in \mathfrak{R}^{n,n}$ der nach 9.1.1 berechnete Cholesky-Faktor von $\boldsymbol{M} := \boldsymbol{A}^T\boldsymbol{A}$. Unter Verwendung von $\boldsymbol{L}$ ist für gegebenes $\boldsymbol{b} \in \mathfrak{R}^m$ die Lösung des Quadratmittelproblems $\boldsymbol{Ax} \cong \boldsymbol{b}$ über die Normalgleichungen zu berechnen und iterativ zu verbessern.

Algorithmus:

S0: $\boldsymbol{x}^0 = \boldsymbol{o}$

for $k := 0(1)k_{\max}$ **do**

S1: Berechne $\boldsymbol{r}^k := \mathrm{fl}\,(\boldsymbol{b} - \boldsymbol{Ax}^k)$ und $\boldsymbol{g}^k := \mathrm{fl}\,(\boldsymbol{A}^T\boldsymbol{r}^k)$

S2: Berechne $\boldsymbol{h} = \boldsymbol{h}^k$ als Lösung von $\boldsymbol{Mh} = \boldsymbol{g}^k$ unter Verwendung der Cholesky-Faktorisierung $\boldsymbol{M} = \boldsymbol{LL}^T$ von $\boldsymbol{M}$ aus den Dreieckssystemen $\boldsymbol{Lf} = \boldsymbol{g}^k$, $\boldsymbol{L}^T\boldsymbol{h} = \boldsymbol{f}$

S3: Setze $\boldsymbol{x}^{k+1} := \boldsymbol{x}^k + \boldsymbol{h}^k$

Aufwand: $\sim (2m + n)\, n$ opms pro Schritt $k \geqq 1$, $\sim (m + n)\, n$ opms für $k = 0$.

Für $k = 0$ ist $\boldsymbol{x}^1 = \boldsymbol{h}^0$ gerade die in 9.1.1 berechnete Lösung $\boldsymbol{x}$.

9.2.2. *Fehleranalyse.* Die spaltenreguläre Matrix $\boldsymbol{A} \in \mathfrak{R}^{m,n}$ genüge der Bedingung

$$\bar{\varkappa} := \nu \overline{N}_1 [\mathrm{cond}\,(\boldsymbol{A})]^2 \leqq 0.23\,, \qquad \overline{N}_1 := (m+1)(n+1) + 2n^2. \tag{1}$$

Dann ist 9.2.1 für jedes $\boldsymbol{b} \in \mathfrak{R}^m$ durchführbar, und die berechneten $\boldsymbol{x}^k$ $(k = 1, \ldots, k_{\max})$ sind Lösungen der gestörten Quadratmittelprobleme

$$(\boldsymbol{A} + \delta \boldsymbol{A}_k)\, \boldsymbol{x}^k \cong \boldsymbol{b}\,, \tag{2}$$

wobei

$$\|\delta \boldsymbol{A}_k\| \leqq 1.3\nu\left(\sqrt{m^2 n + n^3} + \bar{\varkappa}^{k-1}\, \overline{N}_1\, \mathrm{cond}\,(\boldsymbol{A})\right) \|\boldsymbol{A}\| \tag{3}$$

ist. Mit $\boldsymbol{x}^* = \boldsymbol{A}^+ \boldsymbol{b}$, $\boldsymbol{r}^* = \boldsymbol{b} - \boldsymbol{A}\boldsymbol{x}^*$ gelten die Fehlerabschätzungen

$$\|\boldsymbol{x}^k - \boldsymbol{x}^*\| \leqq \bar{\varkappa}^k \|\boldsymbol{x}^*\| + 1.3\nu\, \mathrm{cond}\,(\boldsymbol{A})\, \{n^{3/2} \|\boldsymbol{x}^*\| + mn^{1/2}\, \mathrm{cond}\,(\boldsymbol{A})\, \|\boldsymbol{r}^*\|/\|\boldsymbol{A}\|\}\,. \tag{4}$$

Der Beweis ist sehr technisch und soll übergangen werden; er benutzt Lemma 8.2.11.

9.2.3. *Bemerkung.* (i) Für genügend großes k folgt aus (2), (3) die *numerische Gutartigkeit*, aus (4) die *numerische Stabilität des Verfahrens* 9.2.1. Da dies für $k = 1$, d. h. für die nicht verbesserte Lösung gemäß 9.1.1, nicht gefolgert werden kann, vgl. 9.1.6, spielt die iterative Verbesserung für das Normalgleichungsverfahren 9.1.1 eine andere Rolle als für die Lösung regulärer Gleichungssysteme nach dem Gaußschen Algorithmus in 5.4. Für das Verfahren 9.1.1 ist sie erforderlich, um überhaupt Gutartigkeit zu garantieren, während sie beim Gaußschen Algorithmus die für das Grundverfahren ohnehin vorliegende Gutartigkeit lediglich verbessert.

(ii) Als Abbruchkriterium kann wie in 5.4.5 die Bedingung

$$\|\boldsymbol{h}^k\| > \|\boldsymbol{h}^{k-1}\|/2$$

gewählt werden. Falls die Korrekturen nicht schnell genug kleiner werden, ist entweder die Grenzgenauigkeit erreicht, oder (1) ist nicht erfüllt, so daß die iterative Verbesserung ohnehin fragwürdig ist. Der letzte Fall kann durch Schätzung von $\bar{\varkappa} \sim \hat{\varkappa} \approx \tilde{\varkappa}$ analog zu 9.1.7(ii) diagnostiziert werden. Die Konvergenzgeschwindigkeit hängt entscheidend von $\bar{\varkappa}$ ab.

(iii) Wenn $\boldsymbol{M}$ und $\boldsymbol{L}$ mit erhöhter Genauigkeit $\nu_1 \leqq \nu^2$ berechnet und $\boldsymbol{L}$ danach wieder auf einfache Genauigkeit ν gerundet wird, kann $\bar{\varkappa}$ in 9.2.2 durch

$$\bar{\varkappa}_1 := \nu \overline{N}_1\, \mathrm{cond}\,(\boldsymbol{A})$$

ersetzt werden.

(iv) Wenn die Residuen $\boldsymbol{r}^k$ und $\boldsymbol{g}^k$ mit höherer Genauigkeit $\nu_1 \leqq \nu^2$ berechnet und danach wieder auf einfache Genauigkeit ν gerundet werden, liefert 9.2.2 für genügend großes k die volle relative Genauigkeit im Sinne von

$$\|\boldsymbol{x}^k - \boldsymbol{x}^*\| \leqq \nu N_3\, \{\|\boldsymbol{x}^*\| + \mathrm{cond}\,(\boldsymbol{A})\, \|\boldsymbol{r}^*\|/\|\boldsymbol{A}\|\}\,, \qquad N_3 \approx 1.$$

Ob diese Genauigkeit nötig und sinnvoll ist, hängt von der Genauigkeit der Eingangsdaten ab. □

Bemerkungen zum Kapitel 9

B 9.1. Die Lösung von Quadratmittelproblemen über die Normalgleichungen geht bereits auf GAUSS zurück. Die Normalgleichungen werden deshalb auch *Gaußsche Normalgleichungen* und deren Koeffizientenmatrix $\boldsymbol{A}^T\boldsymbol{A}$ *Gaußsche Transformierte* von $\boldsymbol{A}$ genannt. In der statistischen Literatur wird auch fast ausschließlich dieser klassische Weg beschrieben, obwohl er vom numerischen Standpunkt her deutliche Nachteile gegenüber den moderneren Orthogonalisierungsverfahren aufweist.

B 9.2. Die Fehleranalyse der iterativen Verbesserung gemäß 9.2.2 und speziell der Nachweis der numerischen Gutartigkeit für genügend großes k geht auf KIEŁBASIŃSKI zurück und scheint neu zu sein. Ein verwandtes Verfahren hat BJÖRCK [78] untersucht. In der letztgenannten Arbeit sind auch instruktive numerische Beispiele zu finden.

10. Orthogonalisierungsverfahren

Orthogonalisierungsverfahren zur Lösung von Quadratmittelproblemen $\boldsymbol{A}\boldsymbol{x} \cong \boldsymbol{b}$ benutzen die Darstellung von $\boldsymbol{A} \in \mathbf{R}^{m,n}$ gemäß $\boldsymbol{A} = \boldsymbol{Q}\boldsymbol{R}$ als Produkt einer spaltenorthonormalen Matrix $\boldsymbol{Q} \in \mathbf{R}^{m,l}$ und einer oberen Dreiecks- bzw. Trapezmatrix $\boldsymbol{R} \in \mathbf{R}^{l,n}$; man spricht von einer *$\boldsymbol{QR}$-Faktorisierung* von $\boldsymbol{A}$. Die verschiedenen Varianten unterscheiden sich in der Wahl der Formate — üblich sind $l = n$ oder $l = m$ — und in den Konstruktionsvorschriften für $\boldsymbol{Q}$ und $\boldsymbol{R}$. Unter Verwendung einer solchen $\boldsymbol{QR}$-Faktorisierung lassen sich die Quadratmittellösungen wie auch andere interessierende Größen, etwa $\boldsymbol{P}_{\boldsymbol{A}}$ oder $\boldsymbol{A}^+$, in einfacher Weise bestimmen.

10.1. Orthogonalisierung nach Gram-Schmidt

Wir setzen $\boldsymbol{A} \in \mathbf{R}^{m,n}$ als spaltenregulär voraus und suchen eine Faktorisierung

$$\boldsymbol{A} = \boldsymbol{Q}\boldsymbol{R} \tag{1}$$

mit einer spaltenorthonormalen Matrix $\boldsymbol{Q} \in \mathbf{R}^{m,n}$ und einer oberen Dreiecksmatrix $\boldsymbol{R} \in \mathbf{R}^{n,n}$, d. h., es gelte

$$\boldsymbol{Q}^T\boldsymbol{Q} = \boldsymbol{I}_n \quad \text{und} \quad \boldsymbol{R} = (r_{ij}) \quad \text{mit} \quad r_{ij} = 0 \qquad (i > j;\ i, j = 1, \ldots, n). \tag{2}$$

Wegen der Spaltenregularität von $\boldsymbol{A}$ muß $\boldsymbol{R}$ regulär sein, also $r_{jj} \neq 0$ $(j = 1, \ldots, n)$ gelten. Wenn die Spalten von $\boldsymbol{A}$ bzw. $\boldsymbol{Q}$ gemäß

$$\boldsymbol{A} = (\boldsymbol{a}^1, \ldots, \boldsymbol{a}^n), \qquad \boldsymbol{Q} = (\boldsymbol{q}^1, \ldots, \boldsymbol{q}^n)$$

mit $\boldsymbol{a}^j$ bzw. $\boldsymbol{q}^j$ bezeichnet werden, liest sich (1) spaltenweise als

$$\boldsymbol{a}^j = \sum_{i=1}^{j} \boldsymbol{q}^i r_{ij} = \boldsymbol{q}^1 r_{1j} + \boldsymbol{q}^2 r_{2j} + \cdots + \boldsymbol{q}^j r_{jj} \qquad (j = 1, \ldots, n), \tag{3}$$

d. h., die Elemente der j-ten Spalte von $\boldsymbol{R}$ sind gerade die Koeffizienten von $\boldsymbol{a}^j$ in der orthonormalen Basis $\{\boldsymbol{q}^1, \ldots, \boldsymbol{q}^j\}$. Umgekehrt gilt wegen der Regularität von $\boldsymbol{R}$ auch $\boldsymbol{Q} = \boldsymbol{A}\tilde{\boldsymbol{R}}$ mit der oberen Dreiecksmatrix $\tilde{\boldsymbol{R}} = \boldsymbol{R}^{-1}$, also die zu (3) analoge Darstellung von $\boldsymbol{q}^j$ als Linearkombination von $\boldsymbol{a}^1, \ldots, \boldsymbol{a}^j$. Hieraus folgt

$$\operatorname{span}\{\boldsymbol{a}^1, \ldots, \boldsymbol{a}^j\} = \operatorname{span}\{\boldsymbol{q}^1, \ldots, \boldsymbol{q}^j\} \qquad (j = 1, \ldots, n); \tag{4}$$

insbesondere ist also $\{\boldsymbol{q}^1, \ldots, \boldsymbol{q}^n\}$ eine orthonormale Basis von $\mathcal{R}(\boldsymbol{A})$. Andererseits folgt aus (4) die Gültigkeit der Darstellung (3). Dies bedeutet: Die Faktorisierung von $\boldsymbol{A}$ gemäß (1), (2) ist gleichbedeutend mit der Konstruktion einer orthonormalen Basis $\{\boldsymbol{q}^1, \ldots, \boldsymbol{q}^n\}$, für die (4) gilt. Daß eine solche Basis existiert, wird weiter unten konstruktiv bewiesen.

Unter Verwendung von (1), (2) läßt sich $\boldsymbol{A}^+$ nach 8.1.9 in der Form

$$\boldsymbol{A}^+ = \boldsymbol{R}^+\boldsymbol{Q}^+ = \boldsymbol{R}^{-1}\boldsymbol{Q}^\mathsf{T} \tag{5}$$

darstellen. Die eindeutige Lösung von $\boldsymbol{A}\boldsymbol{x} \cong \boldsymbol{b}$ ist dann durch

$$\boldsymbol{x} = \boldsymbol{A}^+\boldsymbol{b} = \boldsymbol{R}^{-1}\boldsymbol{Q}^\mathsf{T}\boldsymbol{b} \tag{6}$$

gegeben, und das zugehörige minimale Residuum ist

$$\boldsymbol{r} = \boldsymbol{b} - \boldsymbol{A}\boldsymbol{x} = \boldsymbol{b} - \boldsymbol{Q}\boldsymbol{Q}^\mathsf{T}\boldsymbol{b}. \tag{7}$$

Das Problem $\boldsymbol{A}\boldsymbol{x} \cong \boldsymbol{b}$ kann daher wie folgt gelöst werden.

10.1.1. *Basisalgorithmus zur Lösung von* $\boldsymbol{A}\boldsymbol{x} \cong \boldsymbol{b}$ *mittels Faktorisierung* $\boldsymbol{A} = \boldsymbol{Q}\boldsymbol{R}$

S1 ($\boldsymbol{Q}\boldsymbol{R}$-Faktorisierung von $\boldsymbol{A}$): Bestimme $\boldsymbol{Q} \in \mathbf{R}^{m,n}$ mit $\boldsymbol{Q}^\mathsf{T}\boldsymbol{Q} = \boldsymbol{I}_n$ und eine obere Dreiecksmatrix $\boldsymbol{R} \in \mathbf{R}^{n,n}$, so daß $\boldsymbol{A} = \boldsymbol{Q}\boldsymbol{R}$ gilt.

S2 (Berechnung der zur rechten Seite $\boldsymbol{b}$ gehörenden Lösung $\boldsymbol{x} = \boldsymbol{R}^{-1}\boldsymbol{Q}^\mathsf{T}\boldsymbol{b}$ und des Residuums $\boldsymbol{r} = \boldsymbol{b} - \boldsymbol{Q}\boldsymbol{Q}^\mathsf{T}\boldsymbol{b}$):

S2.1: Berechne $\boldsymbol{c} = \boldsymbol{Q}^\mathsf{T}\boldsymbol{b}$ und $\boldsymbol{r} = \boldsymbol{b} - \boldsymbol{Q}\boldsymbol{c}$

S2.2: Bestimme $\boldsymbol{x}$ als Lösung des Dreieckssystems $\boldsymbol{R}\boldsymbol{x} = \boldsymbol{c}$.

Zur Berechnung der $\boldsymbol{Q}\boldsymbol{R}$-Faktorisierung gehen wir von (3) aus und nehmen an, daß bereits ein System orthonormaler Vektoren $\{\boldsymbol{q}^1, \ldots, \boldsymbol{q}^{k-1}\}$, für das (4) für $j = 1, \ldots, k-1$ gilt, bestimmt worden ist. Wir suchen dann $\boldsymbol{q}^k$ in der Form

$$\boldsymbol{q}^k = \boldsymbol{p}^k / r_{kk} \quad \text{mit} \quad \boldsymbol{p}^k = r_{kk}\boldsymbol{q}^k = \boldsymbol{a}^k - \sum_{i=1}^{k-1} \boldsymbol{q}^i r_{ik}.$$

Die Forderung, daß $\boldsymbol{q}^k$ zu $\{\boldsymbol{q}^1, \ldots, \boldsymbol{q}^{k-1}\}$ orthogonal sei, führt wegen der Orthonormalität von $\{\boldsymbol{q}^1, \ldots, \boldsymbol{q}^{k-1}\}$ auf

$$0 = r_{kk}\boldsymbol{q}^{j\mathsf{T}}\boldsymbol{q}^k = \boldsymbol{q}^{j\mathsf{T}}\boldsymbol{a}^k - \sum_{i=1}^{k-1} \boldsymbol{q}^{j\mathsf{T}}\boldsymbol{q}^i r_{ik} = \boldsymbol{q}^{j\mathsf{T}}\boldsymbol{a}^k - r_{jk} \quad (j = 1, \ldots, k-1),$$

so daß bis auf r_{kk} alle Koeffizienten r_{jk} festgelegt sind. Der noch fehlende Wert ergibt sich aus der Normierungsbedingung $\|\boldsymbol{q}^k\| = \|\boldsymbol{p}^k\|/|r_{kk}| = 1$ zu

$$r_{kk} = \pm\|\boldsymbol{p}^k\|.$$

Wegen der linearen Unabhängigkeit der $\boldsymbol{a}^j$ muß dabei $\boldsymbol{p}^k \neq \boldsymbol{o}$ gelten, denn andernfalls wäre wegen (4) $\boldsymbol{a}^k \in \operatorname{span}\{\boldsymbol{a}^1, \ldots, \boldsymbol{a}^{k-1}\}$.

Zusammenfassend erhalten wir den folgenden, Gram-Schmidt-Orthogonalisierung genannten Algorithmus.

10.1.2. *Gram-Schmidt-Orthogonalisierung.* Für jede spaltenreguläre Matrix $\boldsymbol{A} = (\boldsymbol{a}^1, \ldots, \boldsymbol{a}^n) \in \mathbf{R}^{m,n}$ werden durch die Vorschrift

$$\boldsymbol{p}^k := \boldsymbol{a}^k - \sum_{i=1}^{k-1} r_{ik}\boldsymbol{q}^i, \qquad r_{ik} := \boldsymbol{q}^{i\mathsf{T}}\boldsymbol{a}^k \qquad (k = 1, \ldots, n), \tag{8}$$

$$r_{kk} := \|\boldsymbol{p}^k\|, \qquad \boldsymbol{q}^k := \boldsymbol{p}^k/r_{kk}$$

in exakter Arithmetik eine spaltenorthonormale Matrix $\boldsymbol{Q} = (\boldsymbol{q}^1, \ldots, \boldsymbol{q}^n) \in \mathbf{R}^{m,n}$ und eine obere Dreiecksmatrix $\boldsymbol{R} = (r_{ij}) \in \mathbf{R}^{n,n}$ mit $r_{kk} > 0$ $(k = 1, \ldots, n)$ erzeugt, so daß $\boldsymbol{A} = \boldsymbol{QR}$ gilt.

Man sagt auch, daß $\{\boldsymbol{q}^1, \ldots, \boldsymbol{q}^n\}$ durch Gram-Schmidt-Orthogonalisierung der Spalten $\{\boldsymbol{a}^1, \ldots, \boldsymbol{a}^n\}$ von $\boldsymbol{A}$ entstanden ist.

10.1.3. *Bemerkung.* (i) Für spaltenreguläres $\boldsymbol{A}$ ist die $\boldsymbol{QR}$-Faktorisierung (1), (2) bis auf das Vorzeichen der Spalten von $\boldsymbol{Q}$ und entsprechender Zeilen von $\boldsymbol{R}$ eindeutig festgelegt: Aus

$$\boldsymbol{A} = \boldsymbol{QR} = \bar{\boldsymbol{Q}}\bar{\boldsymbol{R}}$$

mit $\boldsymbol{Q}^\mathsf{T}\boldsymbol{Q} = \bar{\boldsymbol{Q}}^\mathsf{T}\bar{\boldsymbol{Q}} = \boldsymbol{I}_n$ und oberen Dreiecksmatrizen $\boldsymbol{R}$, $\bar{\boldsymbol{R}}$ folgt

$$\bar{\boldsymbol{Q}} = \boldsymbol{QD}, \qquad \bar{\boldsymbol{R}} = \boldsymbol{DR} \quad \text{mit} \quad \boldsymbol{D} = \operatorname{diag}(d_i), \qquad d_i = \pm 1,$$

siehe Ü 10.1.3.

(ii) Aus (1), (2) ergibt sich

$$\boldsymbol{A}^\mathsf{T}\boldsymbol{A} = \boldsymbol{R}^\mathsf{T}\boldsymbol{Q}^\mathsf{T}\boldsymbol{QR} = \boldsymbol{R}^\mathsf{T}\boldsymbol{R},$$

d. h., in exakter Arithmetik ist $\boldsymbol{R}^\mathsf{T}$ bis auf das Vorzeichen der Spalten mit dem eindeutig festgelegten Cholesky-Faktor $\boldsymbol{L}$ von $\boldsymbol{A}^\mathsf{T}\boldsymbol{A}$ identisch. Für das durch 10.1.2 definierte $\boldsymbol{R}$ gilt wegen $r_{kk} > 0$ daher $\boldsymbol{R}^\mathsf{T} = \boldsymbol{L}$. □

Das folgende Beispiel zeigt, daß der elegante und kompakte Gram-Schmidt-Prozeß in einer katastrophalen Weise numerisch instabil ist.

10.1.4. *Beispiel.* Es sei $\varepsilon > 0$ und

$$\boldsymbol{A} = \begin{pmatrix} 1 & 1 & 1 \\ \varepsilon & 0 & 0 \\ 0 & \varepsilon & 0 \\ 0 & 0 & \varepsilon \end{pmatrix} \tag{9}$$

die Matrix aus 9.1.4(ii). Algorithmus 10.1.2 werde so ausgeführt, daß $\mathrm{fl}(1+\varepsilon^2)=1$ gesetzt, aber alle übrigen Operationen exakt realisiert werden; dies entspricht qualitativ dem Fall $\varepsilon^2 < \nu/2$. Dann ergibt sich

$$\boldsymbol{Q} = \begin{bmatrix} 1 & 0 & 0 \\ \varepsilon & -\frac{1}{\sqrt{2}} & -\frac{1}{\sqrt{2}} \\ 0 & \frac{1}{\sqrt{2}} & 0 \\ 0 & 0 & \frac{1}{\sqrt{2}} \end{bmatrix}, \qquad \boldsymbol{R} = \begin{pmatrix} 1 & 1 & 1 \\ 0 & \varepsilon\sqrt{2} & 0 \\ 0 & 0 & \varepsilon\sqrt{2} \end{pmatrix}.$$

Wieder in exakter Arithmetik folgt

$$\boldsymbol{A} = \boldsymbol{QR}, \qquad \boldsymbol{Q}^{\mathrm{T}}\boldsymbol{Q} = \begin{bmatrix} 1+\varepsilon^2 & -\frac{\varepsilon}{\sqrt{2}} & -\frac{\varepsilon}{\sqrt{2}} \\ -\frac{\varepsilon}{\sqrt{2}} & 1 & \frac{1}{2} \\ -\frac{\varepsilon}{\sqrt{2}} & \frac{1}{2} & 1 \end{bmatrix},$$

d. h., $\boldsymbol{Q}$ und $\boldsymbol{R}$ liefern eine exakte Faktorisierung, aber $\boldsymbol{Q}$ *ist extrem nichtorthogonal*. Die exakten Faktoren sind

$$\boldsymbol{Q}^* = \begin{bmatrix} 1 & \frac{\varepsilon}{\sqrt{2}} & \frac{\varepsilon}{\sqrt{6}} \\ \varepsilon & -\frac{1}{\sqrt{2}} & -\frac{1}{\sqrt{6}} \\ 0 & \frac{1}{\sqrt{2}} & -\frac{1}{\sqrt{6}} \\ 0 & 0 & \frac{2}{\sqrt{6}} \end{bmatrix} (1+O(\varepsilon^2)), \quad \boldsymbol{R}^* = \begin{bmatrix} 1 & 1 & 1 \\ 0 & \varepsilon\sqrt{2} & \frac{\varepsilon}{\sqrt{2}} \\ 0 & 0 & \frac{\varepsilon\sqrt{6}}{2} \end{bmatrix} (1+O(\varepsilon^2)).$$

Die Schreibweise $\boldsymbol{F} = \boldsymbol{G}(1+O(\varepsilon^2))$ bedeutet dabei $f_{ij} = g_{ij}(1+O(\varepsilon^2))$ für alle i, j. Für die rechten Seiten $\boldsymbol{b}^1 := \boldsymbol{e}^1$, $\boldsymbol{b}^2 := \boldsymbol{e}^3$ ergeben sich mit den berechneten Faktoren $\boldsymbol{Q}, \boldsymbol{R}$ die numerischen Lösungen $\boldsymbol{x}^1, \boldsymbol{x}^2$, mit den exakten Faktoren $\boldsymbol{Q}^*, \boldsymbol{R}^*$ die exakten Lösungen $\boldsymbol{x}^{*,1}, \boldsymbol{x}^{*,2}$ zu

$$\boldsymbol{x}^1 = \begin{pmatrix} 1 \\ 0 \\ 0 \end{pmatrix}, \quad \boldsymbol{x}^{*,1} = \begin{pmatrix} 1/3 \\ 1/3 \\ 1/3 \end{pmatrix} (1+O(\varepsilon^2)),$$

$$\boldsymbol{x}^2 = \frac{1}{\varepsilon}\begin{pmatrix} -1/2 \\ 1/2 \\ 0 \end{pmatrix}, \quad \boldsymbol{x}^{*,2} = \frac{1}{\varepsilon}\begin{pmatrix} -1/3 \\ 2/3 \\ -1/3 \end{pmatrix} (1+O(\varepsilon^2)).$$

Die allein durch die Rundung $\mathrm{fl}(1+\varepsilon^2)=1$ generierten relativen Fehler sind $\|\delta\boldsymbol{x}^1\|/\|\boldsymbol{x}^{*,1}\| = 1.4$, $\|\delta\boldsymbol{x}^2\|/\|\boldsymbol{x}^{*,2}\| = 0.5$. Ein Darstellungsfehler von $\boldsymbol{A}$ zum Fehlerniveau ν kann im Fall $\varepsilon^2 = \nu/2$ dagegen nach 8.2.7 nur eine Störung $\delta\boldsymbol{x}^i$ hervorrufen, für die $\|\delta\boldsymbol{x}^i\|/\|\boldsymbol{x}^{*,i}\|$ von der Größenordnung $\varepsilon \approx \nu^{1/2}$ ist. □

Die klassische Form 10.1.2 *der Gram-Schmidt-Orthogonalisierung ist also für Computerrechnung nicht geeignet.*

Um zu einer numerisch brauchbaren Version zu kommen, schreiben wir (8) als

$$\boldsymbol{p}^k = \boldsymbol{a}^k - \sum_{i=1}^{k-1} \boldsymbol{q}^i \boldsymbol{q}^{i\mathsf{T}} \boldsymbol{a}^k = (\boldsymbol{I} - \boldsymbol{P}_{k-1})\, \boldsymbol{a}^k \quad \text{mit} \quad \boldsymbol{P}_{k-1} := \sum_{i=1}^{k-1} \boldsymbol{q}^i \boldsymbol{q}^{i\mathsf{T}}. \tag{10}$$

Nach Ü 8.1.1 läßt sich der Projektor $\boldsymbol{I} - \boldsymbol{P}_{k-1}$ in der Produktform

$$\boldsymbol{I} - \boldsymbol{P}_{k-1} = (\boldsymbol{I} - \boldsymbol{q}^{k-1}\boldsymbol{q}^{k-1\mathsf{T}}) \cdots (\boldsymbol{I} - \boldsymbol{q}^2\boldsymbol{q}^{2\mathsf{T}})\,(\boldsymbol{I} - \boldsymbol{q}^1\boldsymbol{q}^{1\mathsf{T}}) \tag{11}$$

darstellen. Einsetzen von (11) in (10) und rekursive Auswertung führt dann auf

$$\begin{aligned} &\boldsymbol{p}^{k,1} := \boldsymbol{a}^k, \\ &r_{ik} := \boldsymbol{q}^{i\mathsf{T}} \boldsymbol{p}^{k,i}, \qquad \boldsymbol{p}^{k,i+1} := \boldsymbol{p}^{k,i} - r_{ik}\boldsymbol{q}^i \qquad (i = 1, \ldots, k-1), \\ &\boldsymbol{p}^k := \boldsymbol{p}^{k,k} \end{aligned} \tag{12}$$

als in exakter Arithmetik äquivalente Berechnungsvorschrift für $\boldsymbol{p}^k$. Dabei gilt

$$\boldsymbol{q}^{i\mathsf{T}} \boldsymbol{p}^{k,i} = \boldsymbol{q}^{i\mathsf{T}} (\boldsymbol{I} - \boldsymbol{P}_{i-1})\, \boldsymbol{a}^k = \boldsymbol{q}^{i\mathsf{T}} \boldsymbol{a}^k \qquad (i = 1, \ldots, k-1),$$

d. h., die in (12) auftretenden Koeffizienten r_{ik} sind mit denen aus (8) identisch, und $\boldsymbol{p}^{k,i}$ ist gerade die $(i-1)$-te Teilsumme $\boldsymbol{a}^k - r_{1k}\boldsymbol{q}^1 - \cdots - r_{i-1,k}\boldsymbol{q}^{i-1}$ bei der Berechnung von $\boldsymbol{p}^k$ gemäß (8). Im Unterschied zu (8) wird jedoch in (12) der auf $\boldsymbol{q}^i$ projizierte Anteil $r_{ik}\boldsymbol{q}^i$ sofort von $\boldsymbol{a}^k$ abgezogen, bevor die nachfolgende Projektion gebildet wird. Wegen

$$\|\boldsymbol{a}^k\|^2 = \|\boldsymbol{p}^{k,i}\|^2 + r_{1k}^2 + \cdots + r_{i-1,k}^2 \tag{13}$$

werden die $\boldsymbol{p}^{k,i}$ dabei immer kürzer. In Computerarithmetik werden die r_{ik} gemäß (12) genauer berechnet werden als nach (8), denn die Schranke für den erzeugten Rundungsfehler ist im ersten Fall proportional zu $\|\boldsymbol{p}^{k,i}\|$, im zweiten zu $\|\boldsymbol{a}^k\|$. Es ist daher zu erwarten, daß der nach (12) berechnete Vektor $\boldsymbol{p}^k$ in geringerem Maße durch Rundungsfehler beeinflußt wird.

Derselbe Trick wird dann auch bei der Berechnung von $\boldsymbol{c} = (c_i) = \boldsymbol{Q}^{\mathsf{T}}\boldsymbol{b}$ angewendet und führt auf

$$\boldsymbol{b}^1 := \boldsymbol{b}, \qquad c_i := \boldsymbol{q}^{i\mathsf{T}} \boldsymbol{b}^i, \qquad \boldsymbol{b}^{i+1} := \boldsymbol{b}^i - c_i \boldsymbol{q}^i \qquad (i = 1, \ldots, k-1). \tag{14}$$

In exakter Arithmetik gilt

$$\boldsymbol{q}^{i\mathsf{T}} \boldsymbol{b}^i = \boldsymbol{q}^{i\mathsf{T}} (\boldsymbol{I} - \boldsymbol{P}_{i-1})\, \boldsymbol{b} = \boldsymbol{q}^{i\mathsf{T}} \boldsymbol{b},$$

und

$$\boldsymbol{b}^{n+1} = (\boldsymbol{I} - \boldsymbol{P}_n)\, \boldsymbol{b} = (\boldsymbol{I} - \boldsymbol{Q}\boldsymbol{Q}^{\mathsf{T}})\, \boldsymbol{b} = \boldsymbol{b} - \boldsymbol{Q}\boldsymbol{c} = \boldsymbol{b} - \boldsymbol{A}\boldsymbol{x} = \boldsymbol{r}$$

ist das minimale Residuum, vgl. Ü 4.3.3.

Die Realisierung von 10.1.1 gemäß (12), (14) wird *modifiziertes Gram-Schmidt-Verfahren* — kurz: *MGS-Verfahren* — genannt. Wir geben eine Realisierung an, bei der $\boldsymbol{p}^k$ und die $\boldsymbol{p}^{k,i}$ auf dem Platz von $\boldsymbol{q}^k$, die $\boldsymbol{b}^i$ auf dem von $\boldsymbol{r}$ und $\boldsymbol{c}$ auf dem von $\boldsymbol{x}$ gespeichert werden.

10.1.5. *Orthogonalisierung von A und Lösung von $Ax \cong b$ nach dem modifizierten Gram-Schmidt-Verfahren*

Aufgabe: Für gegebenes spaltenreguläres $A = (a^1, \ldots, a^n) \in \mathbf{R}^{m,n}$ ist eine QR-Faktorisierung von A mit $Q = (q^1, \ldots, q^n) \in \mathbf{R}^{m,n}$ und einer oberen Dreiecksmatrix $R = (r_{ij}) \in \mathbf{R}^{n,n}$ nach den modifizierten Gram-Schmidt-Formeln (12) zu berechnen. Mit dieser ist $x \in \mathbf{R}^n$ und $r \in \mathbf{R}^m$ gemäß (6), (7) bei modifizierter Auswertung von $Q^T b$ nach (14) zu bestimmen.

Algorithmus:

S1 (QR-Faktorisierung von A mittels modifizierter Gram-Schmidt-Orthogonalisierung):
for $k := 1(1)n$ **do**
S1.0: $q^k := a^k$
S1.1: **for** $i := 1(1)k - 1$ **do** $[r_{ik} := q^{iT} q^k,\ q^k := q^k - r_{ik} * q^i]$
S1.2: $r_{kk} := \|q^k\|,\ q^k := q^k / r_{kk}$

S2 (Modifizierte Berechnung von $c = Q^T b$ und $r = b - QQ^T b$, Lösung von $Rx = c$):
S2.1: $r := b$
for $i := 1(1)n$ **do** $[x_i := q^{iT} r,\ r := r - x_i * q^i]$
S2.2: **for** $k := n(-1)1$ **do**
$x_k := x_k / r_{kk}$
for $i := 1(1)k - 1$ **do** $x_i := x_i - r_{ik} * x_k$

Aufwand: S1: $\sim mn^2$ opms, S2: $\sim \left(2m + \frac{n}{2}\right) n$ opms

10.1.6. *Bemerkung.* (i) In-situ-Realisierung durch Überspeichern von A mit Q bzw. von b mit r ist möglich. Dann sind noch $\sim n^2/2$ S für R und n S für x erforderlich.

(ii) Das nichtmodifizierte Gram-Schmidt-Verfahren ergibt sich aus 10.1.5, wenn die Anweisungen $r_{ik} := q^{iT} q^k$ bzw. $x_i := q^{iT} r$ durch $r_{ik} := q^{iT} a^k$ bzw. $x_i := q^{iT} b$ ersetzt werden. Die Programme unterscheiden sich daher fast nicht, insbesondere ist der Aufwand für MGS identisch mit dem für das klassische Verfahren. Die kleinen Unterschiede in den Computerprogrammen haben jedoch große Auswirkungen, denn im Gegensatz zum instabilen Gram-Schmidt-Prozeß ist der modifizierte numerisch gutartig. Dies ist nicht das einzige Beispiel dafür, daß Verfahren, die in exakter Arithmetik äquivalent sind, in Gegenwart von Rundungsfehlern ein völlig unterschiedliches Stabilitätsverhalten aufweisen können.

(iii) In 10.1.5 wird R spaltenweise aufgebaut. Eine zeilenweise Berechnung ist gemäß

for $k := 1(1)n$ **do** $p^{k,1} := a^k$
for $k := 1(1)n$ **do**
$r_{kk} := \|p^{k,k}\|,\ q^k := p^{k,k} / r_{kk}$
for $j := k + 1(1)n$ **do** $[r_{kj} := q^{kT} p^{j,k},\ p^{j,k+1} := p^{j,k} - r_{kj} * q^k]$

möglich, wobei alle $p^{j,k}$ selbstverständlich auf dem Platz von q^j bzw. a^j gespeichert werden. Dabei kann eine „Pivotisierung" durch Spaltenvertauschung wie folgt

durchgeführt werden: Zu Beginn des k-ten Schrittes werden die Spalten k und $\hat{s}(k)$ vertauscht, wobei $\hat{s} = \hat{s}(k)$ mit $k \leqq \hat{s} \leqq n$ durch

$$\|\boldsymbol{p}^{\hat{s},k}\| = \max\{\|\boldsymbol{p}^{j,k}\| : k \leqq j \leqq n\}$$

festgelegt ist. Dadurch wird r_{kk} maximal unter allen Kandidaten. Die dabei benötigten Normen können wegen (13) gemäß

$$\|\boldsymbol{p}^{j,k+1}\|^2 = \|\boldsymbol{p}^{j,k}\|^2 - (r_{kj})^2$$

rekursiv aus $\|\boldsymbol{p}^{j,1}\| = \|\boldsymbol{a}^j\|$ bestimmt werden, wobei auf Auslöschung zu achten ist, siehe B 10.4. Der zur Normierung verwendete Wert sollte daher nochmals direkt berechnet werden. In dieser Modifikation liefert 10.1.5 eine $\boldsymbol{QR}$-Faktorisierung

$$\bar{\boldsymbol{A}} = \boldsymbol{A}\boldsymbol{P}_S^T = \boldsymbol{Q}\boldsymbol{R} \quad \text{mit} \quad \boldsymbol{P}_S^T := \boldsymbol{T}_{1,\hat{s}(1)}\boldsymbol{T}_{2,\hat{s}(2)} \cdots \boldsymbol{T}_{n-1,\hat{s}(n-1)}, \tag{15}$$

für die

$$r_{kk}^2 \geqq r_{kj}^2 + r_{k+1,j}^2 + \cdots + r_{jj}^2 \quad (j = k+1, \ldots, n;\ k = 1, \ldots, n-1) \tag{16}$$

gilt, insbesondere ist

$$r_{11} \geqq r_{22} \geqq \cdots \geqq r_{nn}.$$

Wegen des Auftretens der Permutationsmatrix $\boldsymbol{P}_S$ in (15) müssen die Komponenten x_i der Lösung $\boldsymbol{x}$ im derart modifizierten Algorithmus 10.1.5 am Schluß noch gemäß $\boldsymbol{x} := \boldsymbol{P}_S^T\boldsymbol{x}$ permutiert werden, um die Spaltenvertauschungen wieder rückgängig zu machen.

(iv) Der in (iii) beschriebene pivotisierte MGS-Prozeß kann auch für rangdefizientes $\boldsymbol{A}$ mit $\operatorname{rang}(\boldsymbol{A}) = r \leqq n \leqq m$ angewendet werden. In exakter Arithmetik bricht das Verfahren nach r Schritten wegen $\|\boldsymbol{p}^{j,r+1}\| = 0$ $(j = r+1, \ldots, n)$ ab, und $\{\boldsymbol{q}^1, \ldots, \boldsymbol{q}^r\}$ ist eine orthogonale Basis von $\mathcal{R}(\boldsymbol{A})$. Bei Computerrealisierung ist ein geeigneter Abbruchtest zu verwenden, siehe Kapitel 11. □

10.1.7. *Rundungsfehleranalyse.* Algorithmus 10.1.5 ist für spaltenreguläres $\boldsymbol{A} \in \mathbf{R}^{m,n}$ und beliebiges $\boldsymbol{b} \in \mathbf{R}^m$ mit $r_{kk} \neq 0$ $(k = 1, \ldots, n)$ durchführbar, sofern

$$8.6\nu F \operatorname{cond}(\boldsymbol{A}) < 1 \quad \text{mit} \quad F := 1.4mn^{3/2} \tag{17}$$

gilt. Die berechneten Faktoren $\boldsymbol{Q}, \boldsymbol{R}$ genügen den Beziehungen

$$\boldsymbol{A} + \delta_0\boldsymbol{A} = \boldsymbol{Q}\boldsymbol{R} \quad \text{mit} \quad \|\delta_0\boldsymbol{A}\| \leqq \nu(F/m)\,\|\boldsymbol{A}\| \tag{18}$$

und

$$\boldsymbol{Q}^T\boldsymbol{Q} - \boldsymbol{I} = \boldsymbol{G} \quad \text{mit} \quad \|\boldsymbol{G}\| \leqq 3\nu F \operatorname{cond}(\boldsymbol{A}). \tag{19}$$

Die berechnete Lösung $\boldsymbol{x}$ ist exakte Lösung des Quadratmittelproblems

$$(\boldsymbol{A} + \delta_1\boldsymbol{A})\,\boldsymbol{x} \cong \boldsymbol{b} + \delta_1\boldsymbol{b}, \tag{20}$$

und für das berechnete Residuum $\boldsymbol{r}$ gilt

$$\boldsymbol{r} = (\boldsymbol{b} + \delta_2\boldsymbol{b}) - (\boldsymbol{A} + \delta_2\boldsymbol{A})\,\boldsymbol{x}. \tag{21}$$

Die Matrizen $\boldsymbol{A} + \delta_i \boldsymbol{A}$ sind spaltenregulär, und es gilt

$$\|\delta_i \boldsymbol{A}\| \leqq \nu F \|\boldsymbol{A}\|, \qquad \|\delta_i \boldsymbol{b}\| \leqq \nu\left(F/\sqrt{n}\right) \|\boldsymbol{b}\| \qquad (i = 1, 2). \tag{22}$$

Die Störungsmatrix $\delta_0 \boldsymbol{A}$ ist überdies spaltenweise klein im Sinne von

$$\|\delta_0 \boldsymbol{a}^j\| \leqq \nu(1.4j) \|\boldsymbol{a}^j\| \qquad (j = 1, \ldots, n). \tag{23}$$

Aus Platzgründen verzichten wir auf den nicht ganz einfachen Beweis dieser Aussagen, siehe B 10.1.

Die Analyse 10.1.7 zeigt: *Die Berechnung der Quadratmittellösung $\boldsymbol{x}$ und des zugehörigen Residuums $\boldsymbol{r}$ nach dem modifizierten Gram-Schmidt-Verfahren ist ein numerisch gutartiger Prozeß.*

10.1.8. *Bemerkung.* (i) Die Klasse der spaltenregulären Quadratmittelprobleme, für die 10.1.5 durchführbar ist, wird durch die Bedingung (17) charakterisiert und besteht daher praktisch aus allen spaltenregulären Problemen, die zum Fehlerniveau ν vernünftig gestellt sind, vgl. 9.1.4(i). Insbesondere tritt hier der bei den Normalgleichungsverfahren vorkommende Faktor $[\mathrm{cond}\,(\boldsymbol{A})]^2$ nicht auf. Bei festem Rundungsfehlerniveau ist daher das MGS-Verfahren den Normalgleichungsverfahren in bezug auf die Anwendungsbreite grundsätzlich überlegen. Dies muß allerdings mit einem fast verdoppelten Aufwand bezahlt werden.

(ii) Wenn $\boldsymbol{Q}$ nach dem klassischen Gram-Schmidt-Prozeß berechnet wird, kann

$$\|\boldsymbol{Q}^{\mathrm{T}}\boldsymbol{Q} - \boldsymbol{I}\| \leqq \min\{n^{3/2}, \nu F[\mathrm{cond}\,(\boldsymbol{A})]^{n-1}\} \tag{24}$$

gezeigt werden, und diese Schranke ist realistisch. Demgegenüber ist die Schranke

$$\|\boldsymbol{Q}^{\mathrm{T}}\boldsymbol{Q} - \boldsymbol{I}\| \leqq 3\nu F\,\mathrm{cond}\,(\boldsymbol{A}) < 1 \tag{25}$$

für das nach MGS berechnete $\boldsymbol{Q}$ deutlich besser, hängt aber immer noch von cond $(\boldsymbol{A})$ ab. Das MGS-Verfahren ist daher in bezug auf den Faktor $\boldsymbol{Q}$ nicht numerisch gutartig. Aus (25) kann bei klassischer Berechnung von $\boldsymbol{c} = \boldsymbol{Q}^{\mathrm{T}}\boldsymbol{b}$ die Gutartigkeit bezüglich $\boldsymbol{x}$ auch nicht gefolgert werden, vgl. 4.3.5. Für das Vorliegen der Gutartigkeit bezüglich $\boldsymbol{x}$ und $\boldsymbol{r}$ erweist sich die modifizierte Berechnung dieser Größen gemäß 10.1.5 wesentlich; bei dieser wird die durch (25) nicht erfaßte günstige Korrelation der Fehler in $\boldsymbol{Q}$ und $\boldsymbol{R}$ ausgenutzt und ergibt eine von cond $(\boldsymbol{A})$ unabhängige Kumulationskonstante, siehe B 10.1.

(iii) Ein ausreichend spaltenorthonormales $\boldsymbol{Q}$ kann mit $\sim 2m \cdot n^2$ opms, also doppeltem Aufwand, durch das *Gram-Schmidt-Verfahren mit Re-Orthogonalisierung* gemäß

for $k := 1(1)n$ **do**
  $\boldsymbol{q}^k := \boldsymbol{a}^k$
  for $i := 1(1)k-1$ **do** $[r_{ik} := \boldsymbol{q}^{i\mathrm{T}}\boldsymbol{q}^k,\ \boldsymbol{q}^k := \boldsymbol{q}^k - r_{ik} * \boldsymbol{q}^i]$
  for $i := 1(1)k-1$ **do** $[s_{ik} := \boldsymbol{q}^{i\mathrm{T}}\boldsymbol{q}^k,\ \boldsymbol{q}^k := \boldsymbol{q}^k - s_{ik} * \boldsymbol{q}^i,\ r_{ik} := r_{ik} + s_{ik}]$
  $r_{kk} := \|\boldsymbol{q}^k\|,\ \boldsymbol{q}^k := \boldsymbol{q}^k / r_{kk}$

berechnet werden. Für die so berechneten Matrizen $\boldsymbol{Q}, \boldsymbol{R}$ gilt (18), und (19) kann zu

$$\|\boldsymbol{Q}^{\mathrm{T}}\boldsymbol{Q} - \boldsymbol{I}\| \leqq \nu F_1 \tag{26}$$

mit F_1 in der Größenordnung von F verschärft werden. Die Abweichung von der Orthonormalität hängt also nicht mehr von cond $(\boldsymbol{A})$ ab und ist stets klein. Die Berechnung von $\boldsymbol{Q}$ nach dem Gram-Schmidt-Verfahren mit Re-Orthogonalisierung ist daher ein numerisch gutartiger Prozeß. Die Aussagen (20) bis (22) gelten dann unabhängig davon, ob $\boldsymbol{c} = \boldsymbol{Q}^T\boldsymbol{b}$ modifiziert oder klassisch gemäß $c_i = \boldsymbol{q}^{iT}\boldsymbol{b}$ berechnet wird, siehe wieder 4.3.5. □

Übungsaufgaben

Ü 10.1.1. Man zeige, daß das zu beliebigem $\boldsymbol{x} \in \mathbf{R}^n$ gehörende Residuum $\boldsymbol{r} = \boldsymbol{b} - \boldsymbol{A}\boldsymbol{x}$ unter Verwendung der $\boldsymbol{QR}$-Faktorisierung (1), (2) von $\boldsymbol{A}$ in der Form

$$\boldsymbol{r} = \boldsymbol{u} + \boldsymbol{v} \tag{27}$$

mit

$$\boldsymbol{u} = \boldsymbol{Q}\boldsymbol{Q}^T\boldsymbol{r} = \boldsymbol{Q}(\boldsymbol{Q}^T\boldsymbol{b} - \boldsymbol{R}\boldsymbol{x}) \perp \boldsymbol{v} := (\boldsymbol{I} - \boldsymbol{Q}\boldsymbol{Q}^T)\,\boldsymbol{r} = (\boldsymbol{I} - \boldsymbol{Q}\boldsymbol{Q}^T)\,\boldsymbol{b} \tag{28}$$

dargestellt werden kann und daß

$$\|\boldsymbol{r}\|^2 = \|\boldsymbol{u}\|^2 + \|\boldsymbol{v}\|^2 = \|\boldsymbol{Q}^T\boldsymbol{b} - \boldsymbol{R}\boldsymbol{x}\|^2 + \|(\boldsymbol{I} - \boldsymbol{Q}\boldsymbol{Q}^T)\,\boldsymbol{b}\|^2 \tag{29}$$

gilt. Aus (29) kann der Basisalgorithmus 10.1.1 direkt abgelesen werden.

Ü 10.1.2. Es sei $\boldsymbol{A} = \boldsymbol{Q}\boldsymbol{R}$ eine $\boldsymbol{QR}$-Faktorisierung der spaltenregulären Matrix $\boldsymbol{A} \in \mathbf{R}^{m,n}$. Man überlege sich, daß die Normallösung $\boldsymbol{y}$ des zeilenregulären Systems

$$\boldsymbol{A}^T\boldsymbol{y} = \boldsymbol{f}, \qquad \boldsymbol{f} \in \mathbf{R}^n, \tag{30}$$

wie folgt berechnet werden kann, vgl. auch 4.2.2 (i).

S3.1: Bestimme $\boldsymbol{g}$ als Lösung des Dreieckssystems $\boldsymbol{R}^T\boldsymbol{g} = \boldsymbol{f}$

S3.2: Berechne $\boldsymbol{y} := \boldsymbol{Q}\boldsymbol{g}$.

Ü 10.1.3. Man beweise 10.1.3 (i).

Hinweis: Man betrachte $\bar{\boldsymbol{Q}}^T\boldsymbol{Q} = \bar{\boldsymbol{R}}\boldsymbol{R}^{-1} = \boldsymbol{S}$, zeige $\boldsymbol{S}^T = \boldsymbol{S}^{-1}$ und benutze Ü 1.1.12.

Ü 10.1.4. Man rechne Beispiel 10.1.4 für das modifizierte Gram-Schmidt-Verfahren 10.1.5 durch und überprüfe, daß sich dann bis auf einen Faktor $1 + O(\varepsilon^2)$ die exakten Resultate ergeben und daß die Schranke (25) angenommen wird.

10.2. Orthogonalisierung nach Householder und Givens

Wir suchen für die als spaltenregulär vorausgesetzte Matrix $\boldsymbol{A} \in \mathbf{R}^{m,n}$ wieder eine $\boldsymbol{QR}$-Faktorisierung

$$\boldsymbol{A} = \hat{\boldsymbol{Q}}\hat{\boldsymbol{R}}, \tag{1}$$

im Unterschied zum Ansatz in 10.1 jedoch mit orthogonalem $\hat{\boldsymbol{Q}} \in \mathbf{R}^{m,m}$ und einer oberen Dreiecksmatrix $\hat{\boldsymbol{R}} \in \mathbf{R}^{m,n}$, d. h., es gelte

$$\hat{\boldsymbol{Q}}^T\hat{\boldsymbol{Q}} = \boldsymbol{I}_m \quad \text{und} \quad \hat{\boldsymbol{R}} = (r_{ij}) \quad \text{mit} \quad r_{ij} = 0 \quad \text{für} \quad i > j$$
$$(i = 1, \ldots, m;\ j = 1, \ldots, n). \tag{2}$$

Wenn $\hat{\boldsymbol{Q}}$ und $\hat{\boldsymbol{R}}$ gemäß

$$\hat{\boldsymbol{Q}} = (\boldsymbol{Q} \mid \tilde{\boldsymbol{Q}}), \qquad \hat{\boldsymbol{R}} = \begin{pmatrix} \boldsymbol{R} \\ \hline \boldsymbol{O} \end{pmatrix} = \begin{pmatrix} \boxslash \\ \hline \boldsymbol{O} \end{pmatrix} \tag{3}$$

mit $\boldsymbol{Q} \in \mathbf{R}^{m,n}$, $\tilde{\boldsymbol{Q}} \in \mathbf{R}^{m,m-n}$ und $\boldsymbol{R} \in \mathbf{R}^{n,n}$ partitioniert werden, geht (1) in

$$\boldsymbol{A} = \boldsymbol{Q}\boldsymbol{R}$$

über. Dies bedeutet: Die Blöcke $\boldsymbol{Q}$ und $\boldsymbol{R}$ sind bis auf das Vorzeichen der Spalten von $\boldsymbol{Q}$ und der Zeilen von $\boldsymbol{R}$ eindeutig festgelegt und identisch mit der in diesem Sinne eindeutigen $\boldsymbol{QR}$-Faktorisierung von $\boldsymbol{A}$ nach (10.1.1), (10.1.2), vgl. 10.1.3. Die Spalten der Matrix $\tilde{\boldsymbol{Q}}$ bilden eine orthonormale Basis des orthogonalen Komplements $\mathcal{N}(\boldsymbol{A}^\mathsf{T})$ von $\mathcal{R}(\boldsymbol{A}) = \mathcal{R}(\boldsymbol{Q})$ und unterliegen keinen weiteren Beschränkungen.

Aus (1) bis (3) folgt nach 8.1.9

$$\boldsymbol{A}^+ = \hat{\boldsymbol{R}}^+\hat{\boldsymbol{Q}}^\mathsf{T} = (\boldsymbol{R}^{-1} \mid \boldsymbol{O})\,\hat{\boldsymbol{Q}}^\mathsf{T}, \tag{4}$$

so daß die eindeutige Lösung des Quadratmittelproblems $\boldsymbol{A}\boldsymbol{x} \cong \boldsymbol{b}$ durch

$$\boldsymbol{x} = \boldsymbol{A}^+\boldsymbol{b} = (\boldsymbol{R}^{-1} \mid \boldsymbol{O})\,\hat{\boldsymbol{Q}}^\mathsf{T}\boldsymbol{b} \tag{5}$$

gegeben ist. Das zugehörige minimale Residuum besitzt die Darstellung

$$\boldsymbol{r} = \boldsymbol{b} - \boldsymbol{A}\boldsymbol{x} = \boldsymbol{b} - \hat{\boldsymbol{Q}}\hat{\boldsymbol{R}}\hat{\boldsymbol{R}}^+\hat{\boldsymbol{Q}}^\mathsf{T}\boldsymbol{b} = \boldsymbol{b} - \hat{\boldsymbol{Q}}\left(\begin{array}{c|c} \boldsymbol{I}_n & \boldsymbol{O} \\ \hline \boldsymbol{O} & \boldsymbol{O} \end{array}\right)\hat{\boldsymbol{Q}}^\mathsf{T}\boldsymbol{b} = \hat{\boldsymbol{Q}}\left(\begin{array}{c|c} \boldsymbol{O} & \boldsymbol{O} \\ \hline \boldsymbol{O} & \boldsymbol{I}_{m-n} \end{array}\right)\hat{\boldsymbol{Q}}^\mathsf{T}\boldsymbol{b}. \tag{6}$$

Das Problem $\boldsymbol{A}\boldsymbol{x} \cong \boldsymbol{b}$ kann daher nach dem folgenden Basisalgorithmus gelöst werden.

10.2.1. *Basisalgorithmus zur Lösung von* $\boldsymbol{A}\boldsymbol{x} \cong \boldsymbol{b}$ *mittels Faktorisierung* $\boldsymbol{A} = \hat{\boldsymbol{Q}}\hat{\boldsymbol{R}}$.

S 1 ($\boldsymbol{QR}$-Faktorisierung von $\boldsymbol{A}$): Bestimme eine orthogonale Matrix $\hat{\boldsymbol{Q}} \in \mathbf{R}^{m,m}$ und eine obere Dreiecksmatrix $\hat{\boldsymbol{R}} = \begin{pmatrix} \boldsymbol{R} \\ \hline \boldsymbol{O} \end{pmatrix} \in \mathbf{R}^{m,n}$, so daß $\boldsymbol{A} = \hat{\boldsymbol{Q}}\hat{\boldsymbol{R}}$ gilt.

S 2 (Berechnung der zur rechten Seite $\boldsymbol{b}$ gehörenden Lösung $\boldsymbol{x}$ und des Residuums $\boldsymbol{r} = \boldsymbol{b} - \boldsymbol{A}\boldsymbol{x}$):

S 2.1: Berechne

$$\hat{\boldsymbol{c}} = \begin{pmatrix} \boldsymbol{c} \\ \hline \tilde{\boldsymbol{c}} \end{pmatrix} = \hat{\boldsymbol{Q}}^\mathsf{T}\boldsymbol{b} \quad \text{und} \quad \boldsymbol{r} = \hat{\boldsymbol{Q}}\begin{pmatrix} \boldsymbol{o} \\ \hline \tilde{\boldsymbol{c}} \end{pmatrix}$$

S 2.2: Bestimme $\boldsymbol{x}$ als Lösung des Dreieckssystems $\boldsymbol{R}\boldsymbol{x} = \boldsymbol{c}$.

Wir bemerken, daß die Matrix $\hat{\boldsymbol{Q}}$ für die Berechnung von $\boldsymbol{x}$ und $\boldsymbol{r}$ nicht explizit bekannt zu sein braucht. Es genügt, wenn $\hat{\boldsymbol{Q}}^\mathsf{T}\boldsymbol{b}$ bzw. $\hat{\boldsymbol{Q}}\boldsymbol{f}$ für gegebenes $\boldsymbol{b}$ bzw. $\boldsymbol{f}$ mit ausreichender Genauigkeit berechnet werden kann. Wegen der Orthogonalität von $\hat{\boldsymbol{Q}}$ gilt ferner $\|\boldsymbol{r}\| = \|\tilde{\boldsymbol{c}}\|$, so daß auf die Berechnung von $\boldsymbol{r}$ verzichtet werden kann, wenn nur nach $\|\boldsymbol{r}\|$ gefragt ist.

Um zu einer Konstruktionsvorschrift für $\hat{Q}$ und $\hat{R}$ zu kommen, schreiben wir (1) in der äquivalenten Form

$$\hat{Q}^{\mathsf{T}}A = \hat{R};$$

man beachte, daß $\hat{Q}$ orthogonal ist. Das Auffinden einer Faktorisierung (1), (2) ist also gleichwertig zum Auffinden einer orthogonalen Matrix $\hat{Q}^{\mathsf{T}}$, die A durch Linksmultiplikation auf obere Dreiecksform $\hat{R}$ bringt. Es bietet sich dann an, die Matrix $\hat{Q}^{\mathsf{T}}$ als Produkt der im Abschnitt 3.3 eingeführten orthogonalen elementaren Transformationsmatrizen zu konstruieren, also die Matrix A durch eine Folge solcher Transformationen sukzessive auf Dreiecksform zu bringen.

Im folgenden soll gezeigt werden, wie A mittels Householder-Spiegelungen spaltenweise auf Dreiecksform gebracht werden kann. Wenn man vom Typ der Transformationsmatrix absieht, wird dabei wie beim Gaußschen Algorithmus vorgegangen.

Im ersten Schritt wählen wir die Householder-Spiegelung $H_1 = \overline{H}_1$ so, daß

$$A =: A^{(1)} =: M^{(1)} =: (a^1 \mid B^{(1)})$$

in

$$A^{(2)} = H_1 A^{(1)} = \overline{H}_1 M^{(1)} = (\bar{a}^1 \mid \overline{B}^{(1)}) = \left(\begin{array}{c|c} \varrho_1 & \\ \hline o & \overline{B}^{(1)} \end{array}\right) = \left(\begin{array}{c|c} r_{11} & r_{12} \dots r_{1n} \\ \hline o & M^{(2)} \end{array}\right)$$

transformiert wird. Dies ist nach 3.3.6 stets möglich. Im zweiten Schritt wird die analoge Transformation mit $\overline{H}_2$ auf $M^{(2)}$ angewendet usw. Die nach $k-1$ Schritten erhaltene Matrix $A^{(k)}$ ist dann in den ersten $k-1$ Spalten bereits von oberer Dreiecksform

$$A^{(k)} = \left(\begin{array}{ccc|ccc} a_{11}^{(k)} & \dots & a_{1,k-1}^{(k)} & a_{1k}^{(k)} & \dots & a_{1n}^{(k)} \\ & \ddots & \vdots & \vdots & & \vdots \\ & & a_{k-1,k-1}^{(k)} & a_{k-1,k}^{(k)} & \dots & a_{k-1,n}^{(k)} \\ \hline & & & a_{kk}^{(k)} & \dots & a_{kn}^{(k)} \\ & O & & \vdots & & \vdots \\ & & & a_{mk}^{(k)} & \dots & a_{mn}^{(k)} \end{array}\right) =: \left(\begin{array}{c|c} \diagdown & R^{(k)} \\ \hline O & M^{(k)} \end{array}\right).$$

Im k-ten Schritt wird

$$M^{(k)} = \left(\begin{array}{c|ccc} a_{kk}^{(k)} & a_{k,k+1}^{(k)} & \dots & a_{kn}^{(k)} \\ \vdots & \vdots & & \vdots \\ a_{mk}^{(k)} & a_{m,k+1}^{(k)} & \dots & a_{mn}^{(k)} \end{array}\right) =: (a^k \mid B^{(k)})$$

gemäß 3.3.6 durch die Householder-Spiegelung $\overline{H}_k \in \mathbf{R}^{m-k+1,m-k+1}$ in

$$\overline{H}_k M^{(k)} = (\bar{a}^k \mid \overline{B}^{(k)}) = \left(\begin{array}{c|c} \varrho_k & \\ \hline o & \overline{B}^{(k)} \end{array}\right) =: \left(\begin{array}{c|c} r_{kk} & r_{k,k+1} \dots r_{kn} \\ \hline o & M^{(k+1)} \end{array}\right) \tag{7}$$

transformiert. Dies ist bei der Festlegung

$$\bar{\boldsymbol{H}}_k = \boldsymbol{I} - \frac{\bar{\boldsymbol{v}}^k \bar{\boldsymbol{v}}^{k\mathsf{T}}}{\gamma_k} \quad \text{mit} \quad \bar{\boldsymbol{v}}^k = \begin{pmatrix} v_{kk} \\ v_{k+1,k} \\ \vdots \\ v_{mk} \end{pmatrix} = \begin{pmatrix} 1 - a_{kk}^{(k)}/\varrho_k \\ -a_{k+1,k}^{(k)}/\varrho_k \\ \vdots \\ -a_{mk}^{(k)}/\varrho_k \end{pmatrix} = \boldsymbol{e}^1 - \frac{1}{\varrho_k}\boldsymbol{a}^k, \tag{8}$$

$$\varrho_k := \begin{cases} \|\boldsymbol{a}^k\| & \text{für } a_{kk}^{(k)} \leqq 0, \\ -\|\boldsymbol{a}^k\| & \text{für } a_{kk}^{(k)} > 0 \end{cases} \quad \text{und} \quad \gamma_k := v_{kk} = \|\bar{\boldsymbol{v}}^k\|^2/2 \tag{9}$$

der Fall. Mit der erweiterten Householder-Spiegelung

$$\boldsymbol{H}_k := \left(\begin{array}{c|c} \boldsymbol{I}_{k-1} & \boldsymbol{O} \\ \hline \boldsymbol{O} & \bar{\boldsymbol{H}}_k \end{array}\right) = \boldsymbol{I} - \frac{\boldsymbol{v}^k \boldsymbol{v}^{k\mathsf{T}}}{\gamma_k}, \quad \boldsymbol{v}^k := \begin{pmatrix} \boldsymbol{o} \\ \bar{\boldsymbol{v}}^k \end{pmatrix}_k^{k-1} = \begin{pmatrix} 0 \\ \vdots \\ 0 \\ v_{kk} \\ \vdots \\ v_{mk} \end{pmatrix} \in \mathbf{R}^m \tag{10}$$

läßt sich der k-te Schritt durch

$$\boldsymbol{A}^{(k+1)} := \boldsymbol{H}_k \boldsymbol{A}^{(k)} = \left(\begin{array}{c|c} & \boldsymbol{R}^{(k)} \\ \hline \boldsymbol{O} & \bar{\boldsymbol{H}}_k \boldsymbol{M}^{(k)} \end{array}\right) = \left(\begin{array}{c|c} & \boldsymbol{R}^{(k)} \\ \hline \boldsymbol{O} & \begin{array}{c|c} \varrho_k & \\ \hline \boldsymbol{o} & \bar{\boldsymbol{B}}^{(k)} \end{array} \end{array}\right) = \left(\begin{array}{c|c} & \boldsymbol{R}^{(k+1)} \\ \hline \boldsymbol{O} & \boldsymbol{M}^{(k+1)} \end{array}\right)$$

beschreiben, und $\boldsymbol{A}^{(k+1)}$ ist bis zur k-ten Spalte von oberer Dreiecksform.

Die nach n Schritten erhaltene Matrix

$$\hat{\boldsymbol{R}} := \boldsymbol{A}^{(n+1)} = \boldsymbol{H}_n \boldsymbol{A}^{(n)} = \boldsymbol{H}_n \boldsymbol{H}_{n-1} \cdots \boldsymbol{H}_1 \boldsymbol{A}^{(1)} = \hat{\boldsymbol{Q}}^{\mathsf{T}} \boldsymbol{A}$$

mit

$$\hat{\boldsymbol{Q}}^{\mathsf{T}} := \boldsymbol{H}_n \boldsymbol{H}_{n-1} \cdots \boldsymbol{H}_1, \qquad \hat{\boldsymbol{Q}} = \boldsymbol{H}_1 \boldsymbol{H}_2 \cdots \boldsymbol{H}_n$$

ist dann nach Konstruktion von oberer Dreiecksform, so daß eine $\boldsymbol{QR}$-Faktorisierung des Typs (1), (2) vorliegt. Im Fall $m = n$ ist $\boldsymbol{A}^{(n)}$ bereits eine Dreiecksmatrix, so daß nur $n - 1$ Schritte erforderlich sind. Der Gesamtprozeß wird *Householder-Orthogonalisierung* der Matrix $\boldsymbol{A}$ genannt.

10.2.2. *Householder-Orthogonalisierung.* Es sei $\boldsymbol{A} \in \mathbf{R}^{m,n}$ eine spaltenreguläre Matrix. Dann sind die Transformationen

$$\boldsymbol{A}^{(k+1)} := \boldsymbol{H}_k \boldsymbol{A}^{(k)} \quad (k = 1, \ldots, l), \qquad \boldsymbol{A}^{(1)} := \boldsymbol{A}, \quad l := \min\{m - 1, n\},$$

mit Householder-Spiegelungen $\boldsymbol{H}_k$ gemäß (8) bis (10) in exakter Arithmetik durchführbar und ergeben eine $\boldsymbol{QR}$-Faktorisierung $\boldsymbol{A} = \hat{\boldsymbol{Q}}\hat{\boldsymbol{R}}$ mit der orthogonalen Matrix

$$\hat{\boldsymbol{Q}} := \boldsymbol{H}_1 \boldsymbol{H}_2 \cdots \boldsymbol{H}_{l-1} \boldsymbol{H}_l$$

und der oberen Dreiecksmatrix $\hat{\boldsymbol{R}} = \begin{pmatrix} \boldsymbol{R} \\ \boldsymbol{O} \end{pmatrix} = \boldsymbol{A}^{(l-1)}$, wobei $\varrho_k = r_{kk} \neq 0$ $(k = 1, \ldots, n)$ ist.

Beweis. Es ist noch die Durchführbarkeit zu zeigen, die wegen $\gamma_k = (|a_{kk}^{(k)}| + \|\boldsymbol{a}^k\|)/\|\boldsymbol{a}^k\|$ zur Bedingung $\boldsymbol{a}^k \neq \boldsymbol{o}$ $(k = 1, \ldots, n)$ gleichwertig ist. Es sei also erstmals $\boldsymbol{a}^j = \boldsymbol{o}$ für ein gewisses $j \in \{1, \ldots, n\}$. Dann sind die ersten j Spalten von $\boldsymbol{A}^{(j)}$ linear abhängig, und wegen $\boldsymbol{H}_{j-1} \cdots \boldsymbol{H}_1 \boldsymbol{A} = \boldsymbol{A}^{(j)}$ trifft dies auch für die ersten j Spalten von $\boldsymbol{A}$ zu. Dies widerspricht der vorausgesetzten Spaltenregularität von $\boldsymbol{A}$. □

10.2.3. *Bemerkung.* (i) Bei Berechnung von $\overline{\boldsymbol{B}}^{(k)}$ gemäß 3.3.6 erfordert die Householder-Orthogonalisierung $\sim(m - n/3)\, n^2$ opms. Für $m \gg n$ ist sie daher etwa so teuer wie die MGS-Orthogonalisierung. Im Fall $m \approx n$ ist der Aufwand mit dem für das Normalgleichungsverfahren aus 9.1 vergleichbar.

(ii) Mit zusätzlichen n S zum Speichern der Diagonalelemente $\varrho_k = r_{kk}$ von $\boldsymbol{R}$ kann die Householder-Orthogonalisierung in situ realisiert werden, indem die signifikanten Elemente v_{ik} $(i \geqq k)$ der $\boldsymbol{v}^k$ auf dem Platz der zu 0 gemachten a_{ik} und die r_{kj} $(j > k)$ auf dem Platz der nicht mehr benötigten a_{kj} gespeichert werden. Für $m = 4$, $n = 3$ sind dann Anfangs- und Endbelegung durch

$$\begin{pmatrix} * & * & * \\ \hline a_{11} & a_{12} & a_{13} \\ a_{21} & a_{22} & a_{23} \\ a_{31} & a_{32} & a_{33} \\ a_{41} & a_{42} & a_{43} \end{pmatrix} \rightarrow \begin{pmatrix} \varrho_1 & \varrho_2 & \varrho_3 \\ \hline v_{11} & r_{12} & r_{13} \\ v_{21} & v_{22} & r_{23} \\ v_{31} & v_{32} & v_{33} \\ v_{41} & v_{42} & v_{43} \end{pmatrix}$$

gegeben. Im Gegensatz zur MGS-Orthogonalisierung ist also für $\boldsymbol{R}$ kein zusätzlicher Speicherplatz erforderlich. □

Um die tatsächlich auszuführenden Operationen deutlich zu machen, geben wir eine in-situ-Realisierung des Householder-Verfahrens entsprechend 10.2.3(ii) in detaillierter Form an.

10.2.4. *Orthogonale Faktorisierung der Matrix $\boldsymbol{A}$ nach Householder.*
Aufgabe: Für spaltenreguläres $\boldsymbol{A} \in \mathbf{R}^{m,n}$ ist durch Householder-Orthogonalisierung gemäß 10.2.2 eine Faktorisierung $\boldsymbol{A} = \hat{\boldsymbol{Q}}\hat{\boldsymbol{R}}$ mit einer oberen Dreiecksmatrix $\hat{\boldsymbol{R}} = \begin{pmatrix} \boldsymbol{R} \\ \boldsymbol{O} \end{pmatrix}$, $\boldsymbol{R} = (r_{ij}) \in \mathbf{R}^{n,n}$, und einer in der Produktform

$$\hat{\boldsymbol{Q}} = \boldsymbol{H}_1 \boldsymbol{H}_2 \cdots \boldsymbol{H}_l, \qquad l := \min\{m - 1, n\}, \tag{11}$$

gegebenen Matrix $\hat{\boldsymbol{Q}}$ zu bestimmen. Die Faktoren $\boldsymbol{H}_k$ sind in der Form

$$\boldsymbol{H}_k = \boldsymbol{I} - \boldsymbol{v}^k \boldsymbol{v}^{k\mathsf{T}}/\gamma_k, \qquad \gamma_k = \boldsymbol{v}^{k\mathsf{T}} \boldsymbol{v}^k/2 \qquad (k = 1, \ldots, l), \tag{12}$$

durch $\boldsymbol{v}^k = (0, \ldots, 0, v_{kk}, \ldots, v_{mk})^{\mathsf{T}} \in \mathbf{R}^m$ festgelegt, wobei $\boldsymbol{v}^k$ gemäß $\gamma_k = v_{kk}$ skaliert ist. Die Matrix $\boldsymbol{A}$ ist durch die signifikanten Elemente v_{ik} $(i \geqq k)$ von $\boldsymbol{v}^k$ bzw. r_{kj} $(k < j)$ von $\boldsymbol{R}$ zu überschreiben; die Diagonalelemente $\varrho_k = r_{kk}$ von $\boldsymbol{R}$ sind gesondert zu speichern.

Algorithmus:

$l := \min\{m-1, n\}$

for $k := 1(1)l$ **do**

S1: $\varrho_k := \|(a_{kk}, \ldots, a_{mk})^T\|$, **if** $a_{kk} > 0$ **then** $\varrho_k := -\varrho_k$
$a_{kk} := 1 - a_{kk}/\varrho_k$
for $i := k+1(1)m$ **do** $a_{ik} := -a_{ik}/\varrho_k$

S2: **for** $j := k+1(1)n$ **do**
$\beta := 0$
for $i := k(1)m$ **do** $\beta := \beta + a_{ik} * a_{ij}$
$\beta := \beta/a_{kk}$
for $i := k(1)m$ **do** $a_{ij} := a_{ij} - \beta * a_{ik}$

if $l < n$ **then** $\varrho_n := a_{nn}$

Aufwand: $\sim(m - n/3)\, n^2$ opms $+ n$ opr, n S (für $\varrho_1, \ldots, \varrho_n$).

Die Householder-Spiegelung $\boldsymbol{H}_k$ kann entsprechend 3.3.13 durch eine Folge

$$\boldsymbol{G}_k := \boldsymbol{G}_{km}\boldsymbol{G}_{k,m-1} \cdots \boldsymbol{G}_{k,k+2}\boldsymbol{G}_{k,k+1} \tag{13}$$

von $m-k$ Givens-Drehungen $\boldsymbol{G}_{ki}$ ersetzt werden, die nacheinander die Elemente $a_{ik}^{(k)}$ $(i = k+1, \ldots, m)$ zu 0 machen. Dabei geht $\boldsymbol{A}^{(k+1)} = \boldsymbol{H}_k\boldsymbol{A}^{(k)}$ in

$$\boldsymbol{A}^{(k+1)} = \boldsymbol{G}_k\boldsymbol{A}^{(k)} = \boldsymbol{G}_{km} \cdots \boldsymbol{G}_{k,k+1}\boldsymbol{A}^{(k)} \qquad (k = 1, \ldots, l) \tag{14}$$

über, und $\hat{\boldsymbol{Q}}$ ist durch

$$\hat{\boldsymbol{Q}}^T = \boldsymbol{G}_l \cdots \boldsymbol{G}_1, \quad \text{also} \quad \boldsymbol{Q} = \boldsymbol{G}_1^T \cdots \boldsymbol{G}_l^T$$

bzw. ausführlich durch

$$\hat{\boldsymbol{Q}} = \boldsymbol{G}_{12}^T \cdots \boldsymbol{G}_{1m}^T\boldsymbol{G}_{23}^T \cdots \boldsymbol{G}_{2m}^T \cdots \boldsymbol{G}_{lm}^T \tag{15}$$

gegeben. Dieser Prozeß wird *Givens-Orthogonalisierung* von $\boldsymbol{A}$ genannt.

10.2.5. *Bemerkung.* (i) Bei der expliziten Realisierung von (14), d. h. bei der Realisierung der zu (7) analogen Transformation

$$\boldsymbol{M}^{(k)} = (\boldsymbol{a}^k \mid \boldsymbol{B}^{(k)}) \to \bar{\boldsymbol{G}}_k\boldsymbol{M}^{(k)} = (\bar{\boldsymbol{a}}^k \mid \bar{\boldsymbol{B}}^{(k)}) = \left(\begin{array}{c|c} \varrho_k & \\ \hline \boldsymbol{o} & \bar{\boldsymbol{B}}^{(k)} \end{array}\right) \tag{16}$$

mit

$$\boldsymbol{G}_k = \left(\begin{array}{c|c} \boldsymbol{I}_{k-1} & \boldsymbol{O} \\ \hline \boldsymbol{O} & \bar{\boldsymbol{G}}_k \end{array}\right)$$

nach 3.3.14, erfordert die Givens-Orthogonalisierung $\sim(m - n/3)\, n^2$ (2 opm + 1 ops) und $\sim(m - n/2)\, n$ opr im Vergleich zu $\sim(m - n/3)\, n^2$ opms und n opr für die Householder-Orthogonalisierung, ist also etwa doppelt so teuer. Für n in der Größenordnung von 10 liegt der Aufwand für die Berechnung der $\sim(m - n/2)\, n$ Wurzeln in der Größenordnung von $(m - n/3)\, n^2$ opms, kann also nicht vernachlässigt werden. In-situ-Realisierung ist möglich, indem jeder Drehung $\boldsymbol{G}_{ki}$ gemäß 3.3.11 die Zahl ξ_{ik} zugeordnet und letztere auf dem Platz des zu 0 gemachten a_{ik} $(i > k)$

gespeichert wird; die a_{kj} $(j \geqq k)$ werden mit r_{kj} überschrieben. Das zu 10.2.3(ii) analoge Schema ist dann

$$\begin{pmatrix} a_{11} & a_{12} & a_{13} \\ a_{21} & a_{22} & a_{23} \\ a_{31} & a_{32} & a_{33} \\ a_{41} & a_{42} & a_{43} \end{pmatrix} \to \begin{pmatrix} r_{11} & r_{12} & r_{13} \\ \xi_{21} & r_{22} & r_{23} \\ \xi_{31} & \xi_{32} & r_{33} \\ \xi_{41} & \xi_{42} & \xi_{43} \end{pmatrix}.$$

(ii) Wenn die Matrix $\boldsymbol{A}$ in der skalierten Form

$$\boldsymbol{A} = \operatorname{diag}\left(\sqrt{\varkappa_i}\right)\boldsymbol{D}, \quad \text{d. h.} \quad a_{ij} = \sqrt{\varkappa_i}\, d_{ij} \qquad (i = 1, \ldots, m;\ j = 1, \ldots, n) \tag{17}$$

durch die Zahlen $\{d_{ij}, \varkappa_i\}$ dargestellt wird und die Transformationen (14) bzw. (16) in der impliziten Form gemäß 3.3.21 realisiert werden, spricht man von *impliziter Givens-Orthogonalisierung*. Die Transformationen $\boldsymbol{G}_{ki}$ sind dann in der skalierten Form durch $\boldsymbol{F}_{ki}$ und zugehörige intermediäre Skalierungsvektoren $\boldsymbol{\varkappa}^{(k,i)}$ und $\bar{\boldsymbol{\varkappa}}^{(k,i)}$ gegeben. Die Matrix $\boldsymbol{F}_{ki}$ kann gemäß 3.3.18 durch die dort festgelegte Zahl ξ_{ik} repräsentiert werden. Die Skalierungsvektoren $\boldsymbol{\varkappa}^{(k,i)}$, $\bar{\boldsymbol{\varkappa}}^{(k,i)}$ sind nicht zugänglich, können jedoch in direkter Reihenfolge aus den Anfangswerten $\boldsymbol{\varkappa} = (\varkappa_i)$ gemäß 3.3.19, in inverser Reihenfolge aus den Endwerten $\bar{\boldsymbol{\varkappa}} = (\bar{\varkappa}_i)$ gemäß 3.3.20 unter Verwendung der ξ_{ik} rekonstruiert werden. Die Matrix $\hat{\boldsymbol{Q}}$ aus (15) ist in diesem Sinne implizit durch die Zahlen $\{\xi_{ik}, \varkappa_i\}$ bzw. $\{\xi_{ik}, \bar{\varkappa}_i\}$ gegeben. Die Matrix $\hat{\boldsymbol{R}} = \boldsymbol{A}^{(l+1)}$ fällt in der skalierten Form

$$\hat{\boldsymbol{R}} = \begin{pmatrix} \boldsymbol{R} \\ \hline \boldsymbol{O} \end{pmatrix} = \operatorname{diag}\left(\sqrt{\bar{\varkappa}_i}\right)\hat{\boldsymbol{Z}}, \quad \hat{\boldsymbol{Z}} = \left(\begin{array}{c} \boldsymbol{Z} \\ \hline \boldsymbol{O} \end{array}\right), \quad \text{d. h.} \quad r_{ij} = \sqrt{\bar{\varkappa}_i}\, z_{ij}$$

$$(i \leqq j;\ i, j = 1, \ldots, n)$$

an und wird durch die Zahlen $\{z_{ij}, \bar{\varkappa}_i\}$ dargestellt.

Die implizite Givens-Orthogonalisierung ist also ein Prozeß, der den Eingangsdaten $\{d_{ij}, \varkappa_i\}$ die Ausgangsdaten $\{z_{ij}\ (i \leqq j), \xi_{ij}\ (i > j), \bar{\varkappa}_i\}$ zuordnet.

Der Aufwand ist $\sim(m - n/3)\, n^2$ (1 opm + 1 ops), entspricht also dem der Householder-Orthogonalisierung; Quadratwurzeln treten nicht auf. In-situ-Realisierung analog zu 10.2.3(ii) ist möglich. Das zugehörige Belegungsschema ist

$$\left(\begin{array}{c|ccc} \varkappa_1 & d_{11} & d_{12} & d_{13} \\ \varkappa_2 & d_{21} & d_{22} & d_{23} \\ \varkappa_3 & d_{31} & d_{32} & d_{33} \\ \varkappa_4 & d_{41} & d_{42} & d_{43} \end{array}\right) \to \left(\begin{array}{c|ccc} \bar{\varkappa}_1 & z_{11} & z_{12} & z_{13} \\ \bar{\varkappa}_2 & \xi_{21} & z_{22} & z_{23} \\ \bar{\varkappa}_3 & \xi_{31} & \xi_{32} & z_{33} \\ \bar{\varkappa}_4 & \xi_{41} & \xi_{42} & \xi_{43} \end{array}\right).$$

(iii) Zur Faktorisierung von $\boldsymbol{A}$ werden üblicherweise $\varkappa_i := 1$ und $\boldsymbol{D} := \boldsymbol{A}$ als Eingangsdaten für die implizite Givens-Orthogonalisierung gewählt. Ist jedoch die Matrix eines diagonalgewichteten Quadratmittelproblems $\hat{\boldsymbol{D}}\hat{\boldsymbol{A}}\boldsymbol{x} \cong \hat{\boldsymbol{D}}\hat{\boldsymbol{b}}$ mit $\hat{\boldsymbol{D}} = \operatorname{diag}(\hat{d}_i)$, $\hat{d}_i = 1/\sqrt{w_i}$, zu faktorisieren, vgl. 8.3, so sollte $\boldsymbol{D} := \hat{\boldsymbol{A}}$ und

$$\varkappa_i := w_{\max}/w_i \quad (i = 1, \ldots, m) \quad \text{mit} \quad w_{\max} = \max\{w_i : i = 1, \ldots, m\}$$

gesetzt werden, denn damit ist $\boldsymbol{A} = \operatorname{diag}\left(\sqrt{\varkappa_i}\right)\boldsymbol{D} = \sqrt{w_{\max}}\,\hat{\boldsymbol{D}}\hat{\boldsymbol{A}}$ bis auf den skalaren Faktor $\sqrt{w_{\max}}$ mit $\hat{\boldsymbol{D}}\hat{\boldsymbol{A}}$ identisch.

(iv) Falls die Skalierungsfaktoren im Laufe der impliziten Givens-Orthogonalisierung zu klein werden und zu Unterlauf führen könnten, vgl. 3.3.23, ist eine *Reskalierung* erforderlich. Dabei wird das zu kleine aktuelle $\varkappa_i$ durch $\varkappa_i * \omega^2$ mit einem Faktor $\omega \gg 1$ ersetzt, und die i-te Zeile von $\boldsymbol{D}^{(k)}$ wird durch ω dividiert. Wenn ω eine ganzzahlige Potenz der Basis ist, entstehen dabei keine Rundungsfehler.

(v) Die Reskalierung wird überflüssig, wenn eine „Spaltenpivotisierung" — d. h. Pivotsuche in den Spalten — nach der folgenden Vorschrift durchgeführt wird, vgl. 3.3.23(iv): Zu Beginn des k-ten Schrittes werden die Zeilen k und s von $\{\boldsymbol{\varkappa}^k, \boldsymbol{D}^{(k)}\}$ vertauscht, wobei $s = s(k)$ mit $k \leqq s \leqq m$ durch

$$\varkappa_s^{(k)}[d_{sk}^{(k)}]^2 = \max\{\varkappa_i^{(k)}[d_{ik}^{(k)}]^2 : k \leqq i \leqq m\} \tag{18}$$

festgelegt ist und $\boldsymbol{\varkappa}^k = (\varkappa_i^{(k)})$ die Skalierungsfaktoren von $\boldsymbol{A}^{(k)} = \operatorname{diag}\left(\sqrt{\varkappa_i^{(k)}}\right)\boldsymbol{D}^{(k)}$ bezeichnet. Damit ergibt sich eine orthogonale Faktorisierung

$$\boldsymbol{A} = \hat{\boldsymbol{Q}}\hat{\boldsymbol{R}} \quad \text{mit} \quad \hat{\boldsymbol{Q}} = \boldsymbol{T}_{1,s(1)}\boldsymbol{G}_1^{\mathsf{T}} \cdots \boldsymbol{T}_{l,s(l)}\boldsymbol{G}_l^{\mathsf{T}} \tag{19}$$

in impliziter, skalierter Form. Da wegen (18) immer Fall 1 eintritt, vereinfacht sich auch die Berechnung und Rekonstruktion der Drehungsparameter.

(vi) Die in (v) beschriebenen Zeilenvertauschungen sind auch bei der Householder- oder expliziten Givens-Orthogonalisierung möglich. Dabei ist (18) durch

$$|a_{sk}^{(k)}| = \max\{|a_{ik}^{(k)}| : k \leqq i \leqq m\} \tag{20}$$

zu ersetzen. Mit dieser zusätzlichen Pivotisierung sind die genannten Verfahren auch für steife diagonalgewichtete Probleme geeignet. Wenn auf die Pivotisierung verzichtet wird, sollten die Zeilen zumindest nach abnehmenden Gewichten $\hat{d}_i = 1/\sqrt{w_i}$ geordnet werden.

(vii) Außer den genannten Zeilenvertauschungen sind auch Spaltenvertauschungen analog zu 10.1.6(iii) möglich. Wenn die Spalten der zu Beginn des k-ten Schrittes vorliegenden Restmatrix $\boldsymbol{M}^{(k)}$ gemäß

$$\boldsymbol{M}^{(k)} = [\boldsymbol{m}^{k,k}, \boldsymbol{m}^{k,k+1}, \ldots, \boldsymbol{m}^{k,n}]$$

bezeichnet werden, wird $\hat{s} = \hat{s}(k)$ mit $k \leqq \hat{s} \leqq n$ und

$$\|\boldsymbol{m}^{k,\hat{s}}\| = \max\{\|\boldsymbol{m}^{k,j}\| : k \leqq j \leqq n\} \tag{21}$$

gesucht und eine Vertauschung der Spalten k und $\hat{s}(k)$ von $\boldsymbol{A}^{(k)}$ vorgenommen. Als Ergebnis entsteht eine $\boldsymbol{QR}$-Faktorisierung

$$\bar{\boldsymbol{A}} = \boldsymbol{A}\boldsymbol{P}_S^{\mathsf{T}} = \hat{\boldsymbol{Q}}\hat{\boldsymbol{R}} \quad \text{mit} \quad \boldsymbol{P}_S^{\mathsf{T}} = \boldsymbol{T}_{1,\hat{s}(1)}\boldsymbol{T}_{2,\hat{s}(2)} \cdots \boldsymbol{T}_{n-1,\hat{s}(n-1)} \tag{22}$$

und

$$r_{kk}^2 \geqq r_{kj}^2 + r_{k+1,j}^2 + \cdots + r_{jj}^2 \quad (j = k+1, \ldots, n;\ k = 1, \ldots, n-1); \tag{23}$$

insbesondere gilt

$$|r_{11}| \geqq r_{22}| \geqq \cdots \geqq |r_{nn}|.$$

Die benötigten Normen $\|\boldsymbol{m}^{k,j}\|$ können analog zu 10.1.6 aus $\|\boldsymbol{m}^{1,j}\| = \|\boldsymbol{a}^j\|$ rekursiv berechnet werden. Wie dort bricht das Verfahren in dieser pivotisierten Version im Fall rang $(\boldsymbol{A}) = r < n$ nach r Schritten wegen $\|\boldsymbol{m}^{r+1,j}\| = 0$ $(j = r+1, \ldots, n)$ ab, sofern exakt gerechnet wird. Es ist daher mit geeigneten Abbruchkriterien auch für rangdefizientes $\boldsymbol{A}$ geeignet, siehe Kapitel 11.2. □

Im Unterschied zur MGS-Orthogonalisierung liefert die Householder-Orthogonalisierung den orthogonalen Faktor $\hat{\boldsymbol{Q}}$ nicht explizit, sondern in der Produktdarstellung

$$\hat{\boldsymbol{Q}}^\mathsf{T} = \boldsymbol{H}_l \cdots \boldsymbol{H}_2 \boldsymbol{H}_1 \quad \text{bzw.} \quad \hat{\boldsymbol{Q}} = \boldsymbol{H}_1 \boldsymbol{H}_2 \cdots \boldsymbol{H}_l,$$

wobei jeder Faktor $\boldsymbol{H}_k$ gemäß (12) durch den Vektor $\boldsymbol{v}^k$ repräsentiert ist. Das ist durchaus kein Nachteil, denn zur Berechnung von $\boldsymbol{x}$ und $\boldsymbol{r}$ nach dem Basisalgorithmus 10.2.1 wie auch für weitere Anwendungen genügt es, wenn $\hat{\boldsymbol{Q}}^\mathsf{T}\boldsymbol{b}$ und $\hat{\boldsymbol{Q}}\boldsymbol{f}$ für gegebenes $\boldsymbol{b}$ und $\boldsymbol{f}$ einfach berechnet werden kann. Dies ist mittels der Rekursion

$$\begin{aligned} &\boldsymbol{b}^1 := \boldsymbol{b}, \quad \boldsymbol{b}^{k+1} := \boldsymbol{H}_k \boldsymbol{b}^k \quad (k = 1, \ldots, l), \quad \hat{\boldsymbol{Q}}^\mathsf{T}\boldsymbol{b} := \boldsymbol{b}^{l+1}, \\ &\boldsymbol{f}^{l+1} := \boldsymbol{f}, \quad \boldsymbol{f}^k := \boldsymbol{H}_k \boldsymbol{f}^{k+1} \quad (k = l, \ldots, 1), \quad \hat{\boldsymbol{Q}}\boldsymbol{f} := \boldsymbol{f}^1 \end{aligned} \tag{24}$$

möglich. Die zugehörigen Computerrealisierungen lauten wie folgt:

10.2.6. *Berechnung von $\hat{\boldsymbol{Q}}^\mathsf{T}\boldsymbol{b}$ und $\hat{\boldsymbol{Q}}\boldsymbol{f}$ mittels der Produktdarstellung von $\hat{\boldsymbol{Q}}$.*

Aufgabe: Die Matrix $\hat{\boldsymbol{Q}}$ sei durch die Ausgangsdaten $\boldsymbol{v}^k \in \mathfrak{R}^m$ $(k = 1, \ldots, l)$ von 10.2.4 gemäß (11), (12) definiert, die signifikanten Elemente v_{ik} $(i \geqq k)$ seien auf dem Platz von $\boldsymbol{A}$ gespeichert. Für $\boldsymbol{b}, \boldsymbol{f} \in \mathfrak{R}^m$ ist $\hat{\boldsymbol{c}} = \hat{\boldsymbol{Q}}^\mathsf{T}\boldsymbol{b}$ bzw. $\hat{\boldsymbol{g}} = \hat{\boldsymbol{Q}}\boldsymbol{f}$ zu berechnen und auf dem Platz von $\boldsymbol{b}$ bzw. $\boldsymbol{f}$ zu speichern.

Algorithmus:

$l := \min\{m-1, n\}$
A1 $(\boldsymbol{b} := \hat{\boldsymbol{Q}}^\mathsf{T}\boldsymbol{b})$:
for $k := 1(1)l$ **do**
  $\beta := 0$
  for $i := k(1)m$ **do** $\beta := \beta + a_{ik} * b_i$
  $\beta := \beta / a_{kk}$
  for $i := k(1)m$ **do** $b_i := b_i - a_{ik} * \beta$
A2 $(\boldsymbol{f} := \hat{\boldsymbol{Q}}\boldsymbol{f})$:
for $k := l(-1)1$ **do**
  $\beta := 0$
  for $i := k(1)m$ **do** $\beta := \beta + a_{ik} * f_i$
  $\beta := \beta / a_{kk}$
  for $i := k(1)m$ **do** $f_i := f_i - a_{ik} * \beta$

Aufwand: jeweils $\sim(2m - n)\, n$ opms.

Man beachte, daß die explizite Bildung von $\hat{\boldsymbol{Q}}$ auch aus Aufwandsgründen nicht zweckmäßig ist: Außer den zur Berechnung erforderlichen Rechenoperationen — zu diesen siehe Ü 10.2.2 — wären dafür m^2 S zum Speichern von $\hat{\boldsymbol{Q}}$ und jeweils $\sim m^2$ opms für die Berechnung von $\hat{\boldsymbol{Q}}^T\boldsymbol{b}$ bzw. $\hat{\boldsymbol{Q}}\boldsymbol{f}$ erforderlich. *Die implizite Produktdarstellung mit Speicherung der erzeugenden Daten v_{ik} auf dem Platz von $\boldsymbol{A}$ ist deshalb als natürliche Darstellung von $\hat{\boldsymbol{Q}}$ anzusehen. Dasselbe trifft sinngemäß für die Givens-Faktorisierung zu.*

10.2.7. *Rundungsfehleranalyse.* Die Householder-Orthogonalisierung 10.2.4 ist für spaltenreguläres $\boldsymbol{A} \in \mathfrak{R}^{m,n}$ mit $\varrho_k = r_{kk} \neq 0$ $(k = 1, \ldots, n)$ durchführbar, sofern

$$\varkappa := vF \operatorname{cond}(\boldsymbol{A}) < 1 \quad \text{mit} \quad F := Kmn^{3/2} \tag{25}$$

gilt. Dabei ist $K := K(m) = 3.14(1 + 3/m)$ wie in 3.3.7, insbesondere gilt

$$K(m) \leqq 4 \quad \text{für} \quad m \geqq 10. \tag{26}$$

Durch 10.2.4 ist eine exakt orthogonale Matrix $\hat{\boldsymbol{Q}} \in \mathbf{R}^{m,m}$ definiert, so daß mit dem berechneten Dreiecksfaktor $\hat{\boldsymbol{R}} \in \mathfrak{R}^{m,n}$

$$\boldsymbol{A} + \delta\boldsymbol{A} = \hat{\boldsymbol{Q}}\hat{\boldsymbol{R}} \quad \text{mit} \quad \|\delta\boldsymbol{A}\| \leqq vF\,\|\boldsymbol{A}\| \tag{27}$$

gilt. Die Ausgangsdaten $\boldsymbol{v}^k \in \mathfrak{R}^m$ $(k = 1, \ldots, l)$ von 10.2.4 liefern eine numerische Darstellung von $\hat{\boldsymbol{Q}}$ derart, daß die durch 10.2.6 für jedes $\boldsymbol{b}, \boldsymbol{f} \in \mathfrak{R}^m$ berechenbaren Vektoren $\hat{\boldsymbol{c}} = \mathrm{fl}(\hat{\boldsymbol{Q}}^T\boldsymbol{b})$ und $\hat{\boldsymbol{g}} = \mathrm{fl}(\hat{\boldsymbol{Q}}\boldsymbol{f})$ den Beziehungen

$$\begin{aligned} &\mathrm{fl}(\hat{\boldsymbol{Q}}^T\boldsymbol{b}) = \hat{\boldsymbol{Q}}^T(\boldsymbol{b} + \delta\boldsymbol{b}) \quad \text{mit} \quad \|\delta\boldsymbol{b}\| \leqq v\left(F/\sqrt{n}\right)\|\boldsymbol{b}\|, \\ &\mathrm{fl}(\hat{\boldsymbol{Q}}\boldsymbol{f}) = \hat{\boldsymbol{Q}}(\boldsymbol{f} + \delta\boldsymbol{f}) \quad \text{mit} \quad \|\delta\boldsymbol{f}\| \leqq v\left(F/\sqrt{n}\right)\|\boldsymbol{f}\| \end{aligned} \tag{28}$$

genügen.

Dabei ist die Störungsmatrix $\delta\boldsymbol{A}$ spaltenweise klein im Sinne von

$$\|\delta\boldsymbol{a}^j\| \leqq vKmj\,\|\boldsymbol{a}^j\| \qquad (j = 1, \ldots, n). \tag{29}$$

Beweis. Wir setzen zusätzlich $\mu := v\,3.14(m + 3)^2 \leqq 1$ voraus, was praktisch keine Einschränkung bedeutet. Der Beweis ist auch ohne diese Voraussetzung möglich, verläuft dann aber wesentlich komplizierter.

Wir nehmen an, daß bereits $k - 1$ Schritte des Householder-Verfahrens durchgeführt worden sind und daß

$$\boldsymbol{A}^{(k)} = \hat{\boldsymbol{Q}}_{k-1}(\boldsymbol{A} + \delta\bar{\boldsymbol{A}}^{(k)}) \tag{30}$$

mit exakt orthogonalem $\hat{\boldsymbol{Q}}_{k-1}$ und einer Störungsmatrix $\delta\bar{\boldsymbol{A}}^{(k)}$ gilt, die der Abschätzung

$$\|\delta\bar{\boldsymbol{A}}^{(k)}\|_F \leqq vKm(k-1)\,\|\boldsymbol{A}\|_F \tag{31}$$

genügt. Dann folgt

$$\|\boldsymbol{A}^{(k)}\|_F \leqq \|\boldsymbol{A}\|_F + \|\delta\bar{\boldsymbol{A}}^{(k)}\|_F \leqq [1 + vKm(k-1)]\,\|\boldsymbol{A}\|_F. \tag{32}$$

Aus (31) ergibt sich ferner

$$\|\boldsymbol{A}^+\|\,\|\delta\bar{\boldsymbol{A}}^{(k)}\| \leqq \|\boldsymbol{A}^+\|\,\|\delta\bar{\boldsymbol{A}}^{(k)}\|_F \leqq vKm(k-1)\sqrt{n}\operatorname{cond}(\boldsymbol{A}) < \varkappa < 1.$$

Die Matrix $\boldsymbol{A} + \delta\bar{\boldsymbol{A}}^{(k)}$ ist dann nach 8.2.5 spaltenregulär, und dasselbe trifft wegen (30) auch für

$$\boldsymbol{A}^{(k)} = \begin{pmatrix} \boldsymbol{R}^{(k)} \\ \hline \boldsymbol{O} \mid \boldsymbol{M}^{(k)} \end{pmatrix} \quad \text{mit} \quad \boldsymbol{M}^{(k)} = (\boldsymbol{a}^k \mid \boldsymbol{B}^{(k)})$$

zu. Insbesondere ist $\boldsymbol{a}^k \neq \boldsymbol{o}$, so daß der k-te Schritt mit $\varrho_k = r_{kk} \neq 0$ ausgeführt werden kann. Es sei jetzt $\boldsymbol{H}_k^*$ diejenige Spiegelungsmatrix, die der Spalte $\boldsymbol{a}^k$ nach den Formeln (8) bis (10) in exakter Arithmetik zugeordnet wird. Nach 3.3.7 genügt dann die berechnete Matrix $\boldsymbol{A}^{(k+1)}$ der Beziehung

$$\boldsymbol{A}^{(k+1)} = \boldsymbol{H}_k^*(\boldsymbol{A}^{(k)} + \delta\boldsymbol{A}^{(k)}) \tag{33}$$

mit

$$\|\delta\boldsymbol{A}^{(k)}\|_F \leqq \nu\, 3.14(m - k + 4)\,\|\boldsymbol{A}^{(k)}\|_F \leqq \nu K m\lambda\,\|\boldsymbol{A}\|_F;$$

die letzte Ungleichung folgt aus (32) mit

$$\lambda := \frac{m - k + 4}{m + 3}\,[1 + \nu K m(k - 1)] = 1 - \frac{k - 1}{m + 3}\left(1 - \mu\,\frac{m - k + 4}{m + 3}\right) \leqq 1.$$

Es gilt also

$$\|\delta\boldsymbol{A}^{(k)}\|_F \leqq \nu K m\,\|\boldsymbol{A}\|_F. \tag{34}$$

Einsetzen von (30) in (33) liefert

$$\boldsymbol{A}^{(k+1)} = \boldsymbol{H}_k^*[\hat{\boldsymbol{Q}}_{k-1}(\boldsymbol{A} + \delta\bar{\boldsymbol{A}}^{(k)}) + \delta\boldsymbol{A}^{(k)}] = \hat{\boldsymbol{Q}}_k(\boldsymbol{A} + \delta\bar{\boldsymbol{A}}^{(k+1)})$$

mit

$$\hat{\boldsymbol{Q}}_k = \boldsymbol{H}_k^*\hat{\boldsymbol{Q}}_{k-1}, \qquad \delta\bar{\boldsymbol{A}}^{(k+1)} = \delta\bar{\boldsymbol{A}}^{(k)} + \hat{\boldsymbol{Q}}_{k-1}^{\mathsf{T}}\delta\boldsymbol{A}^{(k)},$$

d. h., (30) gilt auch für den Index $k + 1$. Aus (31) und (34) folgt außerdem

$$\|\delta\bar{\boldsymbol{A}}^{(k+1)}\|_F \leqq \|\delta\bar{\boldsymbol{A}}^{(k)}\|_F + \|\delta\boldsymbol{A}^{(k)}\|_F \leqq \nu K m k\,\|\boldsymbol{A}\|_F,$$

also (31) für $k + 1$ statt k. Damit sind (30), (31) für jedes $k = 1, \ldots, l + 1$ gültig, insbesondere ist

$$\hat{\boldsymbol{R}} = \boldsymbol{A}^{(l+1)} = \hat{\boldsymbol{Q}}_l(\boldsymbol{A} + \delta\bar{\boldsymbol{A}}^{(l+1)}) =: \hat{\boldsymbol{Q}}^{\mathsf{T}}(\boldsymbol{A} + \delta\boldsymbol{A})$$

mit

$$\|\delta\boldsymbol{A}\| \leqq \|\delta\boldsymbol{A}\|_F = \|\delta\bar{\boldsymbol{A}}^{(l+1)}\|_F \leqq \nu K m l\,\|\boldsymbol{A}\|_F \leqq \nu F\,\|\boldsymbol{A}\|.$$

Der Beweis von (28) verläuft analog. □

Der Leser lasse sich nicht dadurch irritieren, daß die in 10.2.7 erwähnte und im Beweis definierte exakt orthogonale Matrix $\hat{\boldsymbol{Q}} = \boldsymbol{H}_1^* \cdots \boldsymbol{H}_l^*$ nicht zugänglich ist. Wesentlich ist:

— *Es existiert eine exakt orthogonale Matrix $\hat{\boldsymbol{Q}}$, die mit dem berechneten Dreiecksfaktor $\hat{\boldsymbol{R}}$ die exakte $\boldsymbol{QR}$-Faktorisierung der wenig gestörten Matrix $\boldsymbol{A} + \delta\boldsymbol{A}$ liefert.*

— *Die* nicht zugängliche *exakt orthogonale Matrix $\hat{\boldsymbol{Q}}$ wird numerisch durch die berechneten Vektoren $\boldsymbol{v}^k$ ($k = 1, \ldots, l$)* im folgenden Sinne *gut repräsentiert:* Die mittels der $\boldsymbol{v}^k$ nach 10.2.6 berechneten Vektoren $\mathrm{fl}(\hat{\boldsymbol{Q}}^{\mathsf{T}}\boldsymbol{b})$ bzw. $\mathrm{fl}(\hat{\boldsymbol{Q}}\boldsymbol{f})$ sind das exakte Produkt der Matrizen $\hat{\boldsymbol{Q}}^{\mathsf{T}}$ bzw. $\hat{\boldsymbol{Q}}$ mit den wenig gestörten Argumenten $\boldsymbol{b} + \delta\boldsymbol{b}$ bzw. $\boldsymbol{f} + \delta\boldsymbol{f}$.

Die Householder-Orthogonalisierung von $\boldsymbol{A}$ und die Berechnung von $\hat{\boldsymbol{Q}}^{\mathsf{T}}\boldsymbol{b}$ bzw. $\hat{\boldsymbol{Q}}\boldsymbol{f}$ mittels der $\boldsymbol{v}^k$ sind also numerisch gutartige Prozesse.

10.2.8. *Bemerkung.* (i) Für die in 10.2.5 beschriebene explizite oder implizite Givens-Orthogonalisierung und die zu 10.2.6 analoge Berechnung von $\hat{\boldsymbol{Q}}^T\boldsymbol{b}$ bzw. $\hat{\boldsymbol{Q}}\boldsymbol{f}$ bleibt 10.2.7 gültig, wobei F durch

$$F := 6(m+n)\,n^{1/2} \tag{35}$$

zu ersetzen ist. Zum Nachweis von (35) ist eine sorgfältige Abschätzungstechnik nötig; eine Übertragung des Beweises von 10.2.7 unter Verwendung von 3.3.15 bzw. 3.3.22 würde auf $F = 6\,mn^{3/2}$ führen; siehe B 10.3 für Literaturhinweise. *Die explizite oder implizite Givens-Orthogonalisierung ist also ein numerisch gutartiger Prozeß mit der wesentlich günstigeren Fehlerkumulationskonstanten* (35) *als bei der Householder-Orthogonalisierung.* Trotzdem wird die Householder-Faktorisierung derzeit der nach GIVENS meist vorgezogen, da sie für vollbesetztes $\boldsymbol{A}$ einerseits billiger als die explizite Givens-Faktorisierung, andererseits einfacher zu verstehen und zu realisieren ist als die implizite Givens-Faktorisierung. *Unter Beachtung aller Gesichtspunkte wie Aufwand, Rundungsfehlerverhalten und Flexibilität in der Anwendung besitzt die implizite Givens-Orthogonalisierung jedoch deutliche Vorteile gegenüber den übrigen Orthogonalisierungsverfahren.* Zugehörige hochwertige Software liegt allerdings in allgemein zugänglicher Form noch nicht vor.

(ii) Bei Bedarf kann $\hat{\boldsymbol{Q}}$ spaltenweise gemäß $\tilde{\boldsymbol{q}}^j := \mathrm{fl}(\hat{\boldsymbol{Q}}\boldsymbol{e}^j)$ $(j = 1, \ldots, m)$ nach 10.2.6 berechnet werden. Wegen der speziellen Form von $\boldsymbol{f} = \boldsymbol{e}^j$ vereinfacht sich dabei die Berechnung etwas, siehe Ü 10.2.2 für Hinweise zur Realisierung und Aufwandsangaben. Für die derart berechnete Matrix $\tilde{\boldsymbol{Q}} = (\tilde{\boldsymbol{q}}^1, \ldots, \tilde{\boldsymbol{q}}^m)$ gilt

$$\|\tilde{\boldsymbol{Q}}^T\tilde{\boldsymbol{Q}} - \boldsymbol{I}\|_F \leqq 2\nu K(m)\,m^{3/2}n, \tag{36}$$

d. h., sie ist unabhängig von cond $(\boldsymbol{A})$ ausreichend orthogonal. □

Wir kehren jetzt wieder zur Berechnung der Quadratmittellösung $\boldsymbol{x}$ und des Residuums $\boldsymbol{r}$ nach dem Basisalgorithmus 10.2.1 zurück. Eine Computerrealisierung von S1 ist bereits durch 10.2.4 gegeben. Im folgenden wird eine analoge Realisierung von S2 unter Verwendung der Auswertungsvorschriften 10.2.6 für Ausdrücke der Form $\hat{\boldsymbol{Q}}^T\boldsymbol{b}$ bzw. $\hat{\boldsymbol{Q}}\boldsymbol{f}$ ohne explizite Berechnung von $\hat{\boldsymbol{Q}}$ angegeben.

10.2.9. *Lösung des Quadratmittelproblems* $\hat{\boldsymbol{Q}}\hat{\boldsymbol{R}}\boldsymbol{x} \cong \boldsymbol{b}$.

Aufgabe: Unter Verwendung der gemäß 10.2.4 berechneten und auf dem Platz von $\boldsymbol{A}$ gespeicherten Householder-Faktorisierung $\boldsymbol{A} = \hat{\boldsymbol{Q}}\hat{\boldsymbol{R}}$ von $\boldsymbol{A} \in \mathfrak{R}^{m,n}$ mit $r_{kk} = \varrho_k \neq 0$ $(k = 1, \ldots, n)$ ist die zur rechten Seite $\boldsymbol{b} \in \mathfrak{R}^m$ gehörende Lösung $\boldsymbol{x} \in \mathfrak{R}^n$ von $\boldsymbol{A}\boldsymbol{x} \cong \boldsymbol{b}$ und das zugehörige Residuum $\boldsymbol{r} = \boldsymbol{b} - \boldsymbol{A}\boldsymbol{x} \in \mathfrak{R}^m$ sowie dessen Norm $n\boldsymbol{r} = \|\boldsymbol{r}\| \in \mathfrak{R}$ zu berechnen. Dabei ist $\boldsymbol{b}$ mit $\boldsymbol{r}$ zu überspeichern.

Algorithmus:

S2.1.1: Berechne $\boldsymbol{b} := \mathrm{fl}(\hat{\boldsymbol{Q}}^T\boldsymbol{b})$ nach Algorithmus A 1 aus 10.2.6

S2.1.2: Partitioniere $\boldsymbol{b}$ gemäß

$$\boldsymbol{b} = \begin{pmatrix} \boldsymbol{c} \\ \hline \tilde{\boldsymbol{c}} \end{pmatrix} \begin{matrix} 1 \\ \vdots \\ n \\ n+1 \\ \vdots \\ m \end{matrix}, \quad \text{setze} \quad \boldsymbol{x} := \boldsymbol{c}, \quad \boldsymbol{b} := \begin{pmatrix} \boldsymbol{o} \\ \hline \tilde{\boldsymbol{c}} \end{pmatrix}$$

Berechne $n\boldsymbol{r} := \|\tilde{\boldsymbol{c}}\|$

S2.1.3: Berechne $\boldsymbol{b} := \mathrm{fl}(\hat{\boldsymbol{Q}}\boldsymbol{b})$ nach Algorithmus A2 aus 10.2.6.
S2.2: Berechne $\boldsymbol{x} := \mathrm{fl}(\boldsymbol{R}^{-1}\boldsymbol{x})$ analog zu S2.2 von 10.1.5
Aufwand: $\sim (2mn - n^2/2)$ opms für $\boldsymbol{x}$ und $n\boldsymbol{r}$, $\sim (2mn - n^2)$ opms für $\boldsymbol{r}$.

10.2.10. *Rundungsfehleranalyse.* Die spaltenreguläre Matrix $\boldsymbol{A} \in \mathfrak{R}^{m,n}$ genüge der Bedingung

$$\varkappa = \nu F \operatorname{cond}(\boldsymbol{A}) < 1 \quad \text{mit} \quad F := Kmn^{3/2}, \tag{37}$$

$K := K(m)$ wie in 10.2.7. Dann kann die Householder-Faktorisierung $\boldsymbol{A} = \hat{\boldsymbol{Q}}\hat{\boldsymbol{R}}$ gemäß 10.2.4 mit $r_{kk} \neq 0$ $(k = 1, \ldots, n)$ berechnet werden, und Algorithmus 10.2.9 ist mit den so bestimmten Faktoren für jedes $\boldsymbol{b} \in \mathfrak{R}^m$ durchführbar. Die berechnete Lösung $\boldsymbol{x}$ ist exakte Lösung des spaltenregulären Quadratmittelproblems

$$(\boldsymbol{A} + \delta_1\boldsymbol{A})\,\boldsymbol{x} \cong \boldsymbol{b} + \delta\boldsymbol{b} \quad \text{mit} \quad \|\delta_1\boldsymbol{A}\| \leqq \nu F_1 \|\boldsymbol{A}\|, \qquad F_1 := F + n^{3/2} \sim F, \tag{38}$$

und für das berechnete Residuum $\boldsymbol{r}$ gilt

$$\boldsymbol{r} + \delta\boldsymbol{r} = (\boldsymbol{b} + \delta\boldsymbol{b}) - (\boldsymbol{A} + \delta_1\boldsymbol{A})\,\boldsymbol{x} \tag{39}$$

mit

$$\|\delta\boldsymbol{b}\| \leqq \nu\left(F/\sqrt{n}\right)\|\boldsymbol{b}\|, \qquad \|\delta\boldsymbol{r}\| \leqq \nu\left(F/\sqrt{n}\right)\|\boldsymbol{r}\|. \tag{40}$$

Beweis. Aus 10.2.7 folgt die Durchführbarkeit der Householder-Orthogonalisierung und die Darstellung $\boldsymbol{A} + \delta\boldsymbol{A} = \hat{\boldsymbol{Q}}\hat{\boldsymbol{R}}$ sowie

$$\mathrm{fl}(\hat{\boldsymbol{Q}}^{\mathsf{T}}\boldsymbol{b}) = \hat{\boldsymbol{c}} = \begin{pmatrix}\boldsymbol{c}\\ \tilde{\boldsymbol{c}}\end{pmatrix} = \hat{\boldsymbol{Q}}^{\mathsf{T}}(\boldsymbol{b} + \delta\boldsymbol{b}), \quad \boldsymbol{r} = \mathrm{fl}\left(\hat{\boldsymbol{Q}}\begin{pmatrix}\boldsymbol{o}\\ \tilde{\boldsymbol{c}}\end{pmatrix}\right) = \hat{\boldsymbol{Q}}\left\{\begin{pmatrix}\boldsymbol{o}\\ \tilde{\boldsymbol{c}}\end{pmatrix} + \delta\boldsymbol{c}\right\}$$

mit geeigneten Störungen $\delta\boldsymbol{A}$, $\delta\boldsymbol{b}$, $\delta\boldsymbol{c}$. Nach 4.3.2 löst das berechnete $\boldsymbol{x}$ das gestörte Dreiecksystem $(\boldsymbol{R} + \delta\boldsymbol{R})\,\boldsymbol{x} = \boldsymbol{c}$, wobei

$$\|\delta\boldsymbol{R}\| \leqq \|\delta\boldsymbol{R}\|_F \leqq \nu n \|\boldsymbol{R}\|_F \leqq \nu n^{3/2}\|\boldsymbol{R}\|.$$

Wenn

$$\delta_1\boldsymbol{A} := \delta\boldsymbol{A} + \hat{\boldsymbol{Q}}\begin{pmatrix}\delta\boldsymbol{R}\\ \boldsymbol{O}\end{pmatrix}, \quad \delta\boldsymbol{r} := -\hat{\boldsymbol{Q}}\delta\boldsymbol{c}$$

gesetzt wird, ergibt sich daher

$$\boldsymbol{A} + \delta_1\boldsymbol{A} = \hat{\boldsymbol{Q}}\left\{\hat{\boldsymbol{R}} + \begin{pmatrix}\delta\boldsymbol{R}\\ \boldsymbol{O}\end{pmatrix}\right\} = \hat{\boldsymbol{Q}}\begin{pmatrix}\boldsymbol{R} + \delta\boldsymbol{R}\\ \boldsymbol{O}\end{pmatrix}, \quad \boldsymbol{r} + \delta\boldsymbol{r} = \hat{\boldsymbol{Q}}\begin{pmatrix}\boldsymbol{o}\\ \tilde{\boldsymbol{c}}\end{pmatrix},$$

also

$$(\boldsymbol{A} + \delta_1\boldsymbol{A})^{\mathsf{T}}(\boldsymbol{r} + \delta\boldsymbol{r}) = \begin{pmatrix}\boldsymbol{R} + \delta\boldsymbol{R}\\ \boldsymbol{O}\end{pmatrix}^{\mathsf{T}}\begin{pmatrix}\boldsymbol{o}\\ \tilde{\boldsymbol{c}}\end{pmatrix} = \boldsymbol{o}. \tag{41}$$

Überdies ist $\boldsymbol{A} + \delta_1\boldsymbol{A}$ wie $\boldsymbol{R} + \delta\boldsymbol{R}$ spaltenregulär, und es gilt

$$\boldsymbol{r} + \delta\boldsymbol{r} = \hat{\boldsymbol{Q}}\hat{\boldsymbol{c}} - \hat{\boldsymbol{Q}}\begin{pmatrix}\boldsymbol{c}\\ \boldsymbol{o}\end{pmatrix} = \boldsymbol{b} + \delta\boldsymbol{b} - \hat{\boldsymbol{Q}}\begin{pmatrix}\boldsymbol{R} + \delta\boldsymbol{R}\\ \boldsymbol{O}\end{pmatrix}\boldsymbol{x} = (\boldsymbol{b} + \delta\boldsymbol{b}) - (\boldsymbol{A} + \delta_1\boldsymbol{A})\,\boldsymbol{x},$$

also (39), und (41) stellt gerade die Normalgleichungen des gestörten Problems (38) dar. Die Abschätzungen (40) ergeben sich wegen $\|\delta\boldsymbol{r}\| = \|\delta\boldsymbol{c}\|$ direkt aus (28). Die Schranke für $\delta_1\boldsymbol{A}$ erhält man aus (27) unter Beachtung von $\|\boldsymbol{A}\| \doteqdot \|\boldsymbol{R}\|$ gemäß

$$\|\delta_1\boldsymbol{A}\| \leqq \|\delta\boldsymbol{A}\| + \|\delta\boldsymbol{R}\| \leqq \nu F\|\boldsymbol{A}\| + \nu n^{3/2}\|\boldsymbol{R}\| \doteqdot \nu(F + n^{3/2})\|\boldsymbol{A}\|. \quad \square$$

10.2.11. *Bemerkung.* (i) Die Rundungsfehleranalyse 10.2.10 gilt mit günstigeren Kumulationskonstanten

$$F := 6(m + n)\, n^{1/2}, \qquad F_1 := F + n^{3/2} = (6m + 7n)\, n^{1/2} \tag{42}$$

auch für den Fall, daß die Faktorisierung $\boldsymbol{A} = \hat{\boldsymbol{Q}}\hat{\boldsymbol{R}}$ mittels expliziter oder impliziter Givens-Drehungen berechnet und $\hat{\boldsymbol{Q}}^T\boldsymbol{b}$ bzw. $\hat{\boldsymbol{Q}}\boldsymbol{f}$ in zu 10.2.6 analoger Weise bestimmt wird, vgl. 10.2.8(i). *Die Berechnung von* $\boldsymbol{x}$ *und* $\boldsymbol{r}$ *gemäß* 10.2.9 *ist daher sowohl bei Verwendung der Householder- als auch der expliziten oder impliziten Givens-Faktorisierung ein numerisch gutartiger Prozeß.* Wegen der deutlich kleineren Kumulationskonstanten ist die Givens-Faktorisierung dabei favorisiert.

(ii) Die für die Durchführbarkeit der Householder- bzw. Givens-Faktorisierung mit $r_{kk} \neq 0$ hinreichende Bedingung (25), (37) ist mit der analogen Bedingung (10.1.17) für das MGS-Verfahren vergleichbar und schränkt die Klasse der zum Fehlerniveau v vernünftig gestellten spaltenregulären Quadratmittelprobleme nicht wesentlich ein. Auch im Fall $r_{kk} = 0$ kann die Durchführbarkeit mit nichtverschwindenden Diagonalelementen für jedes $\boldsymbol{A} \neq \boldsymbol{O}$ erzwungen werden, indem $r_{kk} := v\,\|\boldsymbol{A}\|$ gesetzt wird, vgl. 5.3.3(i) für die analoge Situation bei der $\boldsymbol{LR}$-Faktorisierung. Allerdings ist die so modifizierte $\boldsymbol{QR}$-Faktorisierung wie dort nur für spezielle Zwecke verwendbar und insbesondere nicht direkt zur Lösung rangdefizienter Quadratmittelprobleme geeignet.

(iii) Im Fall $m = n$ liefern alle bisher behandelten Orthogonalisierungsverfahren für ausreichend reguläres $\boldsymbol{A}$ eine Faktorisierung $\boldsymbol{A} = \boldsymbol{QR}$ mit orthogonalem $\boldsymbol{Q}$ und einer regulären Dreiecksmatrix $\boldsymbol{R}$. Für die berechneten bzw. in Produktform implizit definierten Faktoren gilt dabei

$$\boldsymbol{A} + \delta\boldsymbol{A} = \boldsymbol{QR} \quad \text{mit} \quad \|\delta\boldsymbol{A}\| \leqq vF\,\|\boldsymbol{A}\|, \tag{43}$$

und die gemäß 10.1.5 bzw. 10.2.9 berechnete Lösung $\boldsymbol{x}$ löst das gestörte reguläre System

$$(\boldsymbol{A} + \delta_1\boldsymbol{A})\,\boldsymbol{x} = \boldsymbol{b} \quad \text{mit} \quad \|\delta_1\boldsymbol{A}\| \leqq vF_1\,\|\boldsymbol{A}\|. \tag{44}$$

Die Kumulationskonstanten F, F_1 und die Anzahl der zur Berechnung der $\boldsymbol{QR}$-Faktorisierung erforderlichen Rechenoperationen sind in der nachfolgenden Tabelle zusammengestellt. Für die Householder-Faktorisierung wurde die für $m \geqq 10$ gültige Schranke $K(m) \leqq 4$ verwendet. Zum Vergleich sind außerdem die entsprechenden Werte für die Gauß-Faktorisierung $\boldsymbol{P}_Z\boldsymbol{A} = \boldsymbol{LR}$ mit Spaltenpivotisierung aus 5.3 mit aufgenommen worden.

Verfahren	F	F_1	Aufwand für Faktorisierung
MGS	$1.4n^{3/2}$	$1.4n^{3/2}$	$\sim n^3$ opms
Householder	$4n^{5/2}$	$4n^{5/2}$	$\sim 2n^3/3$ opms
Givens explizit	$12n^{3/2}$	$13n^{3/2}$	$\sim 2n^3/3$(2 opm + 1 ops) + $n^2/2$ opr
Givens implizit	$12n^{3/2}$	$13n^{3/2}$	$\sim 2n^3/3$ opms
Gauß	$C(n)\,n^{1/2}$	$C(n)\,n^{3/2}$	$\sim n^3/3$ opms

Bei der Gauß-Faktorisierung gilt $C(n) \leqq 2 \cdot 2^n$ für die volle Klasse der genügend regulären Systeme. Die praktisch vorkommenden Gleichungssysteme gehören zur Unterklasse, für die $C(n) \approx n$ gesetzt werden kann, vgl. 5.3. Der Faktor $n^{1/2}$ wird durch den Übergang zur Spektralnorm hervorgerufen. □

Wir wenden uns abschließend der iterativen Verbesserung von Quadratmittellösungen zu und beachten dazu, daß die exakte Lösung $\boldsymbol{x}^* = \boldsymbol{A}^+\boldsymbol{b}$ und das zugehörige exakte Residuum $\boldsymbol{r}^* = \boldsymbol{b} - \boldsymbol{A}\boldsymbol{x}^*$ die eindeutige Lösung des regulären Gleichungssystems

$$\left(\begin{array}{c|c} \boldsymbol{I} & \boldsymbol{A} \\ \hline \boldsymbol{A}^\mathsf{T} & \boldsymbol{O} \end{array}\right)\left(\begin{array}{c} \boldsymbol{r} \\ \hline \boldsymbol{x} \end{array}\right) = \left(\begin{array}{c} \boldsymbol{b} \\ \hline \boldsymbol{o} \end{array}\right) \tag{45}$$

sind, siehe Ü 10.2.4. Wenn $\{\boldsymbol{r}, \boldsymbol{x}\}$ Näherungen für $\{\boldsymbol{r}^*, \boldsymbol{x}^*\}$ sind, stellt

$$\left(\begin{array}{c} \boldsymbol{e} \\ \hline \boldsymbol{f} \end{array}\right) := \left(\begin{array}{c} \boldsymbol{b} - \boldsymbol{A}\boldsymbol{x} - \boldsymbol{r} \\ \hline -\boldsymbol{A}^\mathsf{T}\boldsymbol{r} \end{array}\right) \tag{46}$$

das zugehörige Residuum bezüglich (45) dar. Analog zu 5.4 kann dann der Ansatz

$$\left(\begin{array}{c} \boldsymbol{r}^* \\ \hline \boldsymbol{x}^* \end{array}\right) = \left(\begin{array}{c} \boldsymbol{r} \\ \hline \boldsymbol{x} \end{array}\right) + \left(\begin{array}{c} \boldsymbol{s} \\ \hline \boldsymbol{h} \end{array}\right) \tag{47}$$

gemacht werden. Die Korrektur $\left(\begin{array}{c} \boldsymbol{s} \\ \hline \boldsymbol{h} \end{array}\right)$ genügt in exakter Arithmetik dem System

$$\left(\begin{array}{c|c} \boldsymbol{I} & \boldsymbol{A} \\ \hline \boldsymbol{A}^\mathsf{T} & \boldsymbol{O} \end{array}\right)\left(\begin{array}{c} \boldsymbol{s} \\ \hline \boldsymbol{h} \end{array}\right) = \left(\begin{array}{c} \boldsymbol{e} \\ \hline \boldsymbol{f} \end{array}\right), \tag{48}$$

das unter Verwendung der Faktorisierung $\boldsymbol{A} = \hat{\boldsymbol{Q}}\hat{\boldsymbol{R}}$, $\hat{\boldsymbol{R}} = \left(\begin{array}{c} \boldsymbol{R} \\ \hline \boldsymbol{O} \end{array}\right)$, wie folgt gelöst werden kann:

S1: Berechne $\hat{\boldsymbol{Q}}^\mathsf{T}\boldsymbol{e} =: \left(\begin{array}{c} \boldsymbol{u} \\ \hline \boldsymbol{v} \end{array}\right)\begin{array}{l} 1 \\ \vdots \\ n \\ n+1 \\ \vdots \\ m \end{array}$

S2: Bestimme $\boldsymbol{w}$ aus $\boldsymbol{R}^\mathsf{T}\boldsymbol{w} = \boldsymbol{f}$

S3: Bestimme $\boldsymbol{h}$ aus $\boldsymbol{R}\boldsymbol{h} = \boldsymbol{u} - \boldsymbol{w}$ (49)

S4: Berechne $\boldsymbol{s} := \hat{\boldsymbol{Q}}\left(\begin{array}{c} \boldsymbol{w} \\ \hline \boldsymbol{v} \end{array}\right)$.

Dann stellt

$$\left(\begin{array}{c} \bar{\boldsymbol{r}} \\ \hline \bar{\boldsymbol{x}} \end{array}\right) := \left(\begin{array}{c} \boldsymbol{r} \\ \hline \boldsymbol{x} \end{array}\right) + \left(\begin{array}{c} \boldsymbol{s} \\ \hline \boldsymbol{h} \end{array}\right) \tag{50}$$

bei Computerrechnung eine i. allg. verbesserte Näherung für $\left(\begin{array}{c} \boldsymbol{r}^* \\ \hline \boldsymbol{x}^* \end{array}\right)$ dar.

10.2.12. *Bemerkung.* (i) Der Übergang von $\{\boldsymbol{x}, \boldsymbol{r}\}$ zu $\{\bar{\boldsymbol{x}}, \bar{\boldsymbol{r}}\}$ gemäß (50) kann als einzelner Schritt eines zu 5.4.3 analogen Verfahrens zur iterativen Verbesserung

von Quadratmittellösungen angesehen werden. Man beachte dabei, daß $\boldsymbol{r}$ und $\bar{\boldsymbol{r}}$ hier Näherungen für $\boldsymbol{r}^* = \boldsymbol{b} - \boldsymbol{A}\boldsymbol{x}^*$ bezeichnen und nicht die zu $\boldsymbol{x}$ und $\bar{\boldsymbol{x}}$ gehörenden Residuen.

(ii) Wenn eine Faktorisierung $\boldsymbol{A} = \boldsymbol{Q}\boldsymbol{R}$ des Typs (10.1.1), (10.1.2) bekannt ist, läßt sich eine zu (49) analoge Vorschrift zur Lösung von (48) angeben, siehe Ü 10.2.4.

(iii) Wegen der Gutartigkeit der Orthogonalisierungsverfahren aus 10.1 und 10.2 und der günstigen Werte der Fehlerkumulationskonstanten ist eine iterative Verbesserung i. allg. zwecklos, sofern die Residuen $\{\boldsymbol{e}, \boldsymbol{f}\}$ nur in einfacher Genauigkeit ν berechnet werden. Wenn die Eingangsdaten $\{\boldsymbol{A}, \boldsymbol{b}\}$ jedoch als exakt betrachtet werden können und die Residuen in höherer Genauigkeit ν_1, $\nu_1 \leqq \nu^2$, berechnet werden, kann durch die iterative Verbesserung die volle erreichbare Genauigkeit im Sinne von

$$\begin{pmatrix} \|\delta \boldsymbol{r}\| \\ \|\delta \boldsymbol{x}\| \end{pmatrix} \leqq [\nu + \nu_1 F' \operatorname{cond}(\boldsymbol{A})] \begin{pmatrix} \|\boldsymbol{r}^*\| + \|\boldsymbol{x}^*\|/\|\boldsymbol{A}^+\| \\ \|\boldsymbol{x}^*\| + \|\boldsymbol{A}^+\|\,\|\boldsymbol{r}^*\| \end{pmatrix}$$

erhalten werden. Da diese Situation in den Anwendungen selten vorkommt, gehen wir auf die iterative Verbesserung nicht im Detail ein. Man beachte auch, daß bei den Normalgleichungsverfahren durch die iterative Verbesserung die Gutartigkeit überhaupt erst erreicht wird, also dort eine andere Situation vorliegt als bei den a-priorigutartigen Orthogonalisierungsverfahren. □

Übungsaufgaben

Ü 10.2.1. Man schreibe das Residuum $\boldsymbol{r} = \boldsymbol{b} - \boldsymbol{A}\boldsymbol{x}$ für beliebiges $\boldsymbol{x} \in \mathbf{R}^n$ unter Verwendung der Faktorisierung $\boldsymbol{A} = \hat{\boldsymbol{Q}}\hat{\boldsymbol{R}}$ in der Form

$$\boldsymbol{r} = \hat{\boldsymbol{Q}}(\hat{\boldsymbol{c}} - \hat{\boldsymbol{R}}\boldsymbol{x}), \qquad \hat{\boldsymbol{c}} := \hat{\boldsymbol{Q}}^{\mathsf{T}}\boldsymbol{b} = \begin{pmatrix} \boldsymbol{c} \\ \tilde{\boldsymbol{c}} \end{pmatrix}, \qquad \hat{\boldsymbol{R}} = \begin{pmatrix} \boldsymbol{R} \\ \boldsymbol{O} \end{pmatrix} \tag{51}$$

und folgere hieraus

$$\|\boldsymbol{r}\|^2 = \|\hat{\boldsymbol{c}} - \hat{\boldsymbol{R}}\boldsymbol{x}\|^2 = \|\boldsymbol{c} - \boldsymbol{R}\boldsymbol{x}\|^2 + \|\tilde{\boldsymbol{c}}\|^2. \tag{52}$$

Aus (52) läßt sich der Basisalgorithmus 10.2.1 direkt ablesen.

Ü 10.2.2. Es sei $\hat{\boldsymbol{Q}} = \boldsymbol{H}_1 \cdots \boldsymbol{H}_l$ mit $l = \min\{m - 1, n\}$ gemäß 10.2.4 in Produktform durch die erzeugenden Vektoren $\boldsymbol{v}^k$ $(k = 1, \ldots, l)$ gegeben.

(i) Man zeige die Gültigkeit von

$$\hat{\boldsymbol{Q}}\boldsymbol{e}^j = \boldsymbol{H}_1 \cdots \boldsymbol{H}_l \boldsymbol{e}^j = \boldsymbol{H}_1 \cdots \boldsymbol{H}_{\tilde{l}(j)} \boldsymbol{e}^j \qquad (j = 1, \ldots, m) \tag{53}$$

mit

$$\tilde{l}(j) := \begin{cases} j & \text{für } j = 1, \ldots, l-1, \\ l & \text{für } j \geqq l. \end{cases}$$

(ii) Man nutze (53) zur spaltenweisen Berechnung von $\hat{\boldsymbol{Q}} = (\boldsymbol{q}^1, \ldots, \boldsymbol{q}^m)$ gemäß $\boldsymbol{q}^j = \hat{\boldsymbol{Q}}\boldsymbol{e}^j$ in zu 10.2.6 analoger Weise aus und überlege sich, daß dazu $\sim 2(m^2 - mn + n^2/3)\,n$ opms erforderlich sind. Wenn nur die ersten n Spalten $\{\boldsymbol{q}^1, \ldots, \boldsymbol{q}^n\}$ benötigt werden, reduziert sich der Aufwand auf $\sim (m - n/3)\,n^2$ opms.

Ü 10.2.3. Man überlege sich, daß die Normallösung des zeilenregulären Systems

$$A^T y = f \tag{54}$$

unter Verwendung der Faktorisierung $A = \hat{Q}\hat{R}$ mit in Produktform gegebenem $\hat{Q}$ wie folgt berechnet werden kann, vgl. auch 4.2.2(i) und Ü 10.1.2:

S3.1: Bestimme g als Lösung von $R^T g = f$, setze $\hat{g} := \begin{pmatrix} g \\ o \end{pmatrix}$.

S3.2: Berechne $y := \mathrm{fl}(\hat{Q}\hat{g})$ nach 10.2.6.

Ü 10.2.4. Für $A \in \mathbf{R}^{m,n}$ mit $m \geqq n$ sei $M(A) := \left(\begin{array}{c|c} I & A \\ \hline A^T & O \end{array}\right)$.

(i) Man zeige, daß $M(A)$ genau dann regulär ist, wenn rang $(A) = n$ gilt und daß im Fall der Spaltenregularität von A

$$M(A)^{-1} = \left(\begin{array}{c|c} I - AA^+ & A^{+T} \\ \hline A^+ & -A^+A^{+T} \end{array}\right)$$

gilt.

(ii) Man überprüfe, daß die Lösung von (48) durch die Vorschriften (49) gegeben ist.

(iii) Man zeige, daß das System (48) unter Verwendung einer Faktorisierung $A = QR$ des Typs (10.1.1), (10.1.2) wie folgt gelöst werden kann:

S1: Berechne $u := Q^T e$

S2: Bestimme w aus $R^T w = f$

S3: Bestimme h aus $Rh = u - w$

S4: Berechne $s := e - Q(u - w)$.

10.3. Aufdatierung von QR-Faktorisierungen

Bei praktischen Aufgaben ist häufig nicht nur ein einzelnes Quadratmittelproblem zu lösen, sondern eine Folge

$$A_k x \cong b^k \qquad (k = 1, \ldots, K) \tag{1}$$

von K Problemen, die im folgenden Sinne benachbart sind: A_{k+1} entsteht aus A_k durch

— eine Rang-1-Modifikation

— Hinzufügen oder Streichen einer Spalte

— Hinzufügen oder Streichen einer Zeile,

vgl. 2.1.B1 und 9.3. Es ist dann günstig, die aufwendige QR-Faktorisierung nur für die Startmatrix A_1 zu berechnen und die Faktoren von A_{k+1} aus denen von A_k durch Aufdatierung mit geringem Aufwand zu bestimmen.

Zur Vermeidung von mehrfachen Indizes bezeichnen wir im folgenden die Matrix A_k mit A und die modifizierte Matrix A_{k+1} mit $\bar{A}$. Es gelte $A \in \mathbf{R}^{m,n}$ mit $m \geqq n$, und A wie $\bar{A}$ seien spaltenregulär.

Je nach Art der Faktorisierung unterscheiden wir zwei Fälle:

Fall 1: Gegeben ist die Faktorisierung

$$A = QR$$

mit $Q \in \mathbf{R}^{m,n}$, $Q^T Q = I_n$ und der oberen Dreiecksmatrix $R \in \mathbf{R}^{n,n}$.

Fall 2: Gegeben ist die Faktorisierung

$$\boldsymbol{A} = \hat{\boldsymbol{Q}}\hat{\boldsymbol{R}}$$

mit orthogonalem $\hat{\boldsymbol{Q}} \in \mathbf{R}^{m,m}$ und der oberen Dreiecksmatrix $\hat{\boldsymbol{R}} = \begin{pmatrix} \boldsymbol{R} \\ \hline \boldsymbol{O} \end{pmatrix} \in \mathbf{R}^{m,n}$.

Fall 1 entspricht der in 10.1, Fall 2 der in 10.2 behandelten Situation.

Gesucht ist die analoge Faktorisierung

$$\bar{\boldsymbol{A}} = \bar{\boldsymbol{Q}}\bar{\boldsymbol{R}} \quad \text{bzw.} \quad \bar{\boldsymbol{A}} = \hat{\bar{\boldsymbol{Q}}}\hat{\bar{\boldsymbol{R}}}$$

der modifizierten Matrix $\bar{\boldsymbol{A}}$. Für die Berechnung der Faktoren $\bar{\boldsymbol{Q}}, \bar{\boldsymbol{R}}$ bzw. $\hat{\bar{\boldsymbol{Q}}}, \hat{\bar{\boldsymbol{R}}}$ nach den in 10.1 und 10.2 beschriebenen Methoden sind

$$\sim (K_1 mn^2 + K_2 n^3) \text{ opms} \tag{2}$$

erforderlich. Wir werden unten sehen, daß die Aufdatierung der Faktoren $\boldsymbol{Q}, \boldsymbol{R}$ bzw. $\hat{\boldsymbol{Q}}, \hat{\boldsymbol{R}}$ von $\boldsymbol{A}$ dagegen mit einem Aufwand von

$$\sim (K_4 mn + K_5 n^2) \text{ opms} \tag{3}$$

möglich, also billiger als die Neuberechnung ist. Falls bei Faktorisierungen des Typs $\boldsymbol{A} = \hat{\boldsymbol{Q}}\hat{\boldsymbol{R}}$ die Matrix $\hat{\boldsymbol{Q}}$ explizit berechnet wird, kommt in (2) bzw. (3) ein Term $\sim K_3 m^2 n$ bzw. $\sim K_6 m^2$ hinzu. Allerdings wird $\hat{\boldsymbol{Q}}$ in den meisten Anwendungen überhaupt nicht explizit benötigt, siehe 10.3.6 und 10.3.7 unten.

Die im folgenden für den Fall 1 angegebenen Aufdatierungsalgorithmen beruhen sämtlich auf der Zerlegung eines Vektors $\boldsymbol{d} \in \mathbf{R}^m$ in der Form

$$\boldsymbol{d} = \boldsymbol{P}\boldsymbol{d} + (\boldsymbol{I} - \boldsymbol{P})\,\boldsymbol{d} = \boldsymbol{Q}\boldsymbol{r} + \boldsymbol{p} \tag{4}$$

mit $\boldsymbol{P} := \boldsymbol{Q}\boldsymbol{Q}^T$ als Projektor auf $\mathcal{R}(\boldsymbol{Q}) = \mathcal{R}(\boldsymbol{A})$ und

$$\boldsymbol{r} := \boldsymbol{Q}^T\boldsymbol{d}, \qquad \boldsymbol{p} := \boldsymbol{d} - \boldsymbol{Q}\boldsymbol{r}. \tag{5}$$

Dabei gilt

$$\boldsymbol{Q}\boldsymbol{r} \in \mathcal{R}(\boldsymbol{Q}), \qquad \boldsymbol{p} \perp \mathcal{R}(\boldsymbol{Q})$$

und

$$\|\boldsymbol{d}\|^2 = \|\boldsymbol{Q}\boldsymbol{r}\|^2 + \|\boldsymbol{p}\|^2 = \|\boldsymbol{r}\|^2 + \|\boldsymbol{p}\|^2,$$

vgl. 8.1.B. Die analoge Darstellung für den Fall 2 lautet

$$\boldsymbol{d} = \hat{\boldsymbol{Q}}\hat{\boldsymbol{r}} \quad \text{mit} \quad \hat{\boldsymbol{r}} := \hat{\boldsymbol{Q}}^T\boldsymbol{d}, \qquad \|\hat{\boldsymbol{r}}\| = \|\boldsymbol{d}\|; \tag{6}$$

wegen $\mathcal{R}(\hat{\boldsymbol{Q}}) = \mathbf{R}^m$ tritt der komplementäre Vektor $\boldsymbol{p}$ nicht auf.

10.3.1. *Aufdatierung bei Rang-1-Modifikation*

Aufgabe: Bestimme Faktorisierung von

$$\bar{\boldsymbol{A}} = \boldsymbol{A} + \boldsymbol{u}\boldsymbol{v}^T, \qquad \boldsymbol{u} \in \mathbf{R}^m, \quad \boldsymbol{v} \in \mathbf{R}^n. \tag{7}$$

Algorithmus 1: Gegeben ist die Faktorisierung $\boldsymbol{A} = \boldsymbol{Q}\boldsymbol{R}$.

S1: Zerlege $\boldsymbol{u}$ gemäß

$$\boldsymbol{u} = \boldsymbol{Q}\boldsymbol{r} + \boldsymbol{p} \quad \text{mit} \quad \boldsymbol{r} := \boldsymbol{Q}^{\mathsf{T}}\boldsymbol{u}, \qquad \boldsymbol{p} := \boldsymbol{u} - \boldsymbol{Q}\boldsymbol{r} \perp \mathcal{R}(\boldsymbol{Q}). \tag{8}$$

Berechne $\varrho := \|\boldsymbol{p}\|$. Falls $\varrho > 0$, setze $\boldsymbol{q} := \boldsymbol{p}/\varrho$. Andernfalls bestimme beliebigen Vektor $\boldsymbol{q} \perp \mathcal{R}(\boldsymbol{Q})$ mit $\|\boldsymbol{q}\| = 1$. Schreibe (7) mit $\boldsymbol{p} = \varrho\boldsymbol{q}$ in der Form

$$\bar{\boldsymbol{A}} = \boldsymbol{Q}\boldsymbol{R} + (\boldsymbol{Q}\boldsymbol{r} + \varrho\boldsymbol{q})\,\boldsymbol{v}^{\mathsf{T}} = (\boldsymbol{Q} \mid \boldsymbol{q}) \left(\begin{pmatrix} \boldsymbol{R} \\ \boldsymbol{o}^{\mathsf{T}} \end{pmatrix} + \begin{pmatrix} \boldsymbol{r} \\ \varrho \end{pmatrix} \boldsymbol{v}^{\mathsf{T}} \right) =: \boldsymbol{Q}_1(\boldsymbol{R}_0 + \boldsymbol{r}^0\boldsymbol{v}^{\mathsf{T}}) \tag{9}$$

mit der spaltenorthonormalen Matrix $\boldsymbol{Q}_1 := (\boldsymbol{Q} \mid \boldsymbol{q})$ und $\boldsymbol{R}_0 := \begin{pmatrix} \boldsymbol{R} \\ \boldsymbol{o}^{\mathsf{T}} \end{pmatrix}$, $\boldsymbol{r}^0 := \begin{pmatrix} \boldsymbol{r} \\ \varrho \end{pmatrix}$.

S2: Bestimme n Givens-Drehungen $\boldsymbol{G}_{n,n+1}, \ldots, \boldsymbol{G}_{12}$ so, daß die Komponenten $(\boldsymbol{r}^0)_i$ $(i = n+1, \ldots, 2)$ von $\boldsymbol{r}^0$ sukzessive zu 0 gemacht werden, vgl. 3.3.14 und Ü 3.3.5. Dann gilt

$$\boldsymbol{G}\boldsymbol{r}^0 := \boldsymbol{G}_{12} \cdots \boldsymbol{G}_{n,n+1}\boldsymbol{r}^0 = \varrho_1\boldsymbol{e}^1, \qquad \varrho_1 = \pm\|\boldsymbol{r}^0\|,$$

und

$$\boldsymbol{G}\boldsymbol{R}_0 = \boldsymbol{G}_{12} \cdots \boldsymbol{G}_{n,n+1}\boldsymbol{R}_0 =: \boldsymbol{R}_1$$

ist von oberer Hessenbergform. Damit geht (9) in

$$\bar{\boldsymbol{A}} = (\boldsymbol{Q}_1\boldsymbol{G}^{\mathsf{T}})\,(\boldsymbol{G}\boldsymbol{R}_0 + \boldsymbol{G}\boldsymbol{r}^0\boldsymbol{v}^{\mathsf{T}}) = (\boldsymbol{Q}_1\boldsymbol{G}^{\mathsf{T}})\,(\boldsymbol{R}_1 + \varrho_1\boldsymbol{e}^1\boldsymbol{v}^{\mathsf{T}}) =: \boldsymbol{Q}_2\boldsymbol{R}_2 \tag{10}$$

mit spaltenorthonormalem $\boldsymbol{Q}_2 := \boldsymbol{Q}_1\boldsymbol{G}^{\mathsf{T}}$ und der oberen Hessenbergmatrix $\boldsymbol{R}_2 := \boldsymbol{R}_1 + \varrho_1\boldsymbol{e}^1\boldsymbol{v}^{\mathsf{T}}$ über.

S3: Bestimme n Givens-Drehungen $\tilde{\boldsymbol{G}}_{12}, \ldots, \tilde{\boldsymbol{G}}_{n,n+1}$ so, daß die Subdiagonalelemente $(\boldsymbol{R}_2)_{j+1,j}$ $(j = 1, \ldots, n)$ von $\boldsymbol{R}_2$ sukzessive zu 0 gemacht werden. Dann ist

$$\tilde{\boldsymbol{G}}\boldsymbol{R}_2 := \tilde{\boldsymbol{G}}_{n,n+1} \cdots \tilde{\boldsymbol{G}}_{12}\boldsymbol{R}_2 =: \boldsymbol{R}_3 =: \begin{pmatrix} \bar{\boldsymbol{R}} \\ \boldsymbol{o}^{\mathsf{T}} \end{pmatrix}$$

von oberer Dreiecksform, und $\boldsymbol{Q}_3 := \boldsymbol{Q}_2\tilde{\boldsymbol{G}}^{\mathsf{T}} =: (\bar{\boldsymbol{Q}} \mid \bar{\boldsymbol{q}})$ ist spaltenorthonormal. Schreibe (10) in der Form

$$\bar{\boldsymbol{A}} = \boldsymbol{Q}_2\boldsymbol{R}_2 = (\boldsymbol{Q}_2\tilde{\boldsymbol{G}}^{\mathsf{T}})\,(\tilde{\boldsymbol{G}}\boldsymbol{R}_2) = \boldsymbol{Q}_3\boldsymbol{R}_3 = (\bar{\boldsymbol{Q}} \mid \bar{\boldsymbol{q}}) \begin{pmatrix} \bar{\boldsymbol{R}} \\ \boldsymbol{o}^{\mathsf{T}} \end{pmatrix} = \bar{\boldsymbol{Q}}\bar{\boldsymbol{R}}.$$

Man beachte dabei, daß im Ausnahmefall $\varrho = 0$ in S2 und S3 jeweils $\boldsymbol{G}_{n,n+1} = \tilde{\boldsymbol{G}}_{n,n+1} = \boldsymbol{I}$ gilt. Der Vektor $\boldsymbol{q}$ trägt daher zu $\bar{\boldsymbol{Q}}$ und $\bar{\boldsymbol{R}}$ nichts bei und braucht deshalb überhaupt nicht explizit bestimmt zu werden.

Algorithmus 2: Gegeben ist die Faktorisierung $\boldsymbol{A} = \hat{\boldsymbol{Q}}\hat{\boldsymbol{R}}$

S1: Stelle $\boldsymbol{u}$ in der Form

$$\boldsymbol{u} = \hat{\boldsymbol{Q}}\hat{\boldsymbol{r}}, \qquad \hat{\boldsymbol{r}} := \hat{\boldsymbol{Q}}^{\mathsf{T}}\boldsymbol{u} \tag{11}$$

dar und schreibe (7) als

$$\bar{A} = \hat{Q}\hat{R} + \hat{Q}\hat{r}v^{\mathsf{T}} = \hat{Q}(\hat{R} + \hat{r}v^{\mathsf{T}}). \tag{12}$$

S2: Bestimme $m-1$ Givens-Drehungen $G_{m-1,m}, \ldots, G_{12}$ so, daß

$$G\hat{r} := G_{12} \cdots G_{m-1,m}\hat{r} = \varrho_1 e^1, \qquad \varrho_1 = \pm\|\hat{r}\| = \pm\|u\|,$$

gilt. Dann ist

$$G\hat{R} = G_{12} \cdots G_{m-1,m}\begin{pmatrix} R \\ \hline O \end{pmatrix} = G_{12} \cdots G_{\hat{l},\hat{l}+1}\begin{pmatrix} R \\ \hline O \end{pmatrix} =: \hat{R}_1$$

von oberer Hessenbergform, wobei $\hat{l} := \min\{m-1, n\}$. Damit geht (12) in

$$\bar{A} = (\hat{Q}G^{\mathsf{T}})(G\hat{R} + G\hat{r}v^{\mathsf{T}}) = (\hat{Q}G^{\mathsf{T}})(\hat{R}_1 + \varrho_1 e^1 v^{\mathsf{T}}) =: \hat{Q}_2\hat{R}_2 \tag{13}$$

mit orthogonalem $\hat{Q}_2 := \hat{Q}G^{\mathsf{T}}$ und der oberen Hessenbergmatrix $\hat{R}_2 := \hat{R}_1 + \varrho_1 e^1 v^{\mathsf{T}}$ über.

S3: Bestimme $\hat{l}$ Givens-Drehungen $\tilde{G}_{12}, \cdots, \tilde{G}_{\hat{l},\hat{l}+1}$ so, daß $\hat{R}_2$ sukzessive auf obere Dreiecksform

$$\tilde{G}\hat{R}_2 := \tilde{G}_{\hat{l},\hat{l}+1} \cdots \tilde{G}_{12}\hat{R}_2 =: \hat{\bar{R}} =: \begin{pmatrix} \bar{R} \\ \hline O \end{pmatrix}$$

transformiert wird. Schreibe (13) mit der orthogonalen Matrix $\hat{\bar{Q}} := \hat{Q}_2\tilde{G}^{\mathsf{T}}$ als

$$\bar{A} = (\hat{Q}_2\tilde{G}^{\mathsf{T}})(\tilde{G}\hat{R}_2) = \hat{\bar{Q}}\hat{\bar{R}}. \quad \square$$

Im Fall $m = 5$, $n = 3$ verläuft die Transformation aus S2 von Algorithmus 2 nach folgendem Muster:

$$\hat{R} = \begin{pmatrix} \times & \times & \times \\ 0 & \times & \times \\ 0 & 0 & \times \\ 0 & 0 & 0 \\ 0 & 0 & 0 \end{pmatrix} \overset{G_{45}}{\to} \begin{pmatrix} \times & \times & \times \\ 0 & \times & \times \\ 0 & 0 & \times \\ 0 & 0 & 0 \\ 0 & 0 & 0 \end{pmatrix} \overset{G_{34}}{\to} \begin{pmatrix} \times & \times & \times \\ 0 & \times & \times \\ 0 & 0 & \boxed{\times} \\ 0 & 0 & \boxed{\times} \\ 0 & 0 & 0 \end{pmatrix} \overset{G_{23}}{\to} \begin{pmatrix} \times & \times & \times \\ 0 & \boxed{\times} & \boxed{\times} \\ 0 & \boxed{\times} & \boxed{\times} \\ 0 & 0 & \times \\ 0 & 0 & 0 \end{pmatrix}$$

$$\overset{G_{12}}{\to} \begin{pmatrix} \boxed{\times} & \boxed{\times} & \boxed{\times} \\ \boxed{\times} & \boxed{\times} & \boxed{\times} \\ 0 & \times & \times \\ 0 & 0 & \times \\ 0 & 0 & 0 \end{pmatrix} = \hat{R}_1.$$

Es entsteht also tatsächlich eine obere Hessenbergmatrix. Die eingerahmten Felder geben dabei diejenigen Elemente an, die sich im jeweiligen Teilschritt verändern.

10.3.2. *Aufdatierung bei Hinzufügen einer Spalte*
Aufgabe: Bestimme Faktorisierung von

$$\bar{A} = (A \mid a), \qquad a \in \mathbf{R}^m. \tag{14}$$

Algorithmus 1: Gegeben ist die Faktorisierung $A = QR$. Zerlege a gemäß

$$a = Qr + p \quad \text{mit} \quad r := Q^T a, \qquad p = a - Qr. \tag{15}$$

Setze $\varrho := \|p\|$, $q := p/\varrho$ und schreibe (14) in der Form

$$\bar{A} = (QR \mid Qr + \varrho q) = (Q \mid q) \left(\begin{array}{c|c} R & r \\ \hline o^T & \varrho \end{array}\right) =: \bar{Q}\bar{R}$$

mit der spaltenorthonormalen Matrix $\bar{Q} := (Q \mid q)$ und der oberen Dreiecksmatrix $\bar{R} = \left(\begin{array}{c|c} R & r \\ \hline o^T & \varrho \end{array}\right)$. Wegen rang $(\bar{A}) = n + 1$ ist dabei $a \notin \mathcal{R}(A)$, also $p \neq o$.

Algorithmus 2: Gegeben ist die Faktorisierung $A = \hat{Q}\hat{R}$.

S1: Stelle a in der Form

$$a = \hat{Q}\hat{r} \quad \text{mit} \quad \hat{r} := \hat{Q}^T a =: \left(\begin{array}{c} r \\ \hline \tilde{r} \end{array}\right) \begin{array}{l} 1 \\ \vdots \\ n \\ n+1 \\ \vdots \\ m \end{array} \tag{16}$$

dar. Schreibe (14) als

$$\bar{A} = (\hat{Q}\hat{R} \mid \hat{Q}\hat{r}) = \hat{Q}(\hat{R} \mid \hat{r}) =: \hat{Q}\hat{R}_1, \qquad \hat{R}_1 := (\hat{R} \mid \hat{r}). \tag{17}$$

S2: Bestimme orthogonales $\tilde{H} \in \mathbf{R}^{m-n,m-n}$ als Householder Spiegelung oder als Produkt von $m - n - 1$ Givens-Drehungen so, daß

$$\tilde{H}\tilde{r} = \varrho e^1, \qquad \varrho = \pm\|\tilde{r}\|,$$

gilt. Setze

$$H := \left(\begin{array}{c|c} I_n & O \\ \hline O & \tilde{H} \end{array}\right)$$

und schreibe (17) in der Form

$$\bar{A} = \hat{Q}\hat{R}_1 = (\hat{Q}H^T)(H\hat{R}_1) =: \hat{\hat{Q}}\hat{\hat{R}}$$

mit orthogonalem $\hat{\hat{Q}} := \hat{Q}H^T$ und der oberen Dreiecksmatrix

$$\hat{\hat{R}} := H\hat{R}^1 = \left(\begin{array}{c|c} I_n & O \\ \hline O & H \end{array}\right)\left(\begin{array}{c|c} R & r \\ \hline O & \tilde{r} \end{array}\right) = \left(\begin{array}{c|c} R & r \\ \hline O & \varrho e^1 \end{array}\right). \quad \square$$

Offensichtlich stellen die Algorithmen 10.3.2 gerade den $(n + 1)$-ten Schritt des Gram-Schmidt- bzw. Householder- oder Givens-Verfahrens dar.

10.3.3. *Aufdatierung bei Streichen einer Spalte*

Aufgabe: Bestimme Faktorisierung von $\bar{A}$, wobei

$$\bar{A} = (\bar{a}^1, \ldots, \bar{a}^{n-1}) := (a^1, \ldots, a^{l-1}, a^{l+1}, \ldots, a^n), \qquad n > 1, \tag{18}$$

durch Streichen der l-ten Spalte $\boldsymbol{a}^l$ aus $\boldsymbol{A} = (\boldsymbol{a}^1, \ldots, \boldsymbol{a}^l, \ldots, \boldsymbol{a}^n)$ entsteht ($1 \leqq l \leqq n$).

Algorithmus 1: Gegeben ist die Faktorisierung $\boldsymbol{A} = \boldsymbol{QR}$. Bilde

$$\boldsymbol{R}_0 := \begin{pmatrix} r_{11} & \cdots & r_{1,l-1} & r_{1,l+1} & \cdots & r_{1,n} \\ & \ddots & \vdots & & & \\ & & r_{l-1,l-1} & r_{l-1,l+1} & \cdots & r_{l-1,n} \\ & & & r_{l,l+1} & \cdots & r_{l,n} \\ & \boldsymbol{O} & & r_{l+1,l+1} & \cdots & r_{l+1,n} \\ & & & & \ddots & \vdots \\ & & & & & r_{n,n} \end{pmatrix} \in \mathbf{R}^{m,n-1} \tag{19}$$

aus $\boldsymbol{R} \in \mathbf{R}^{n,n}$ durch Streichen der l-ten Spalte. Bestimme $n - l$ Givens-Drehungen $\boldsymbol{G}_{l,l+1}, \ldots, \boldsymbol{G}_{n-1,n}$ so, daß die Subdiagonalelemente $(\boldsymbol{R}_0)_{j+1,j}$ ($j = l, \ldots, n-1$) von $\boldsymbol{R}_0$ sukzessive zu 0 gemacht werden. Dann ist

$$\boldsymbol{GR}_0 := \boldsymbol{G}_{n-1,n} \cdots \boldsymbol{G}_{l,l+1}\boldsymbol{R}_0 =: \boldsymbol{R}_1 =: \begin{pmatrix} \bar{\boldsymbol{R}} \\ \hline \boldsymbol{o}^\mathsf{T} \end{pmatrix}$$

mit einer oberen Dreiecksmatrix $\bar{\boldsymbol{R}} \in \mathbf{R}^{n-1,n-1}$. Bilde $\boldsymbol{Q}_1 := \boldsymbol{QG}^\mathsf{T} =: (\bar{\boldsymbol{Q}} \mid \bar{\boldsymbol{q}})$ und schreibe (18) in der Form

$$\bar{\boldsymbol{A}} = \boldsymbol{QR}_0 = (\boldsymbol{QG}^\mathsf{T})(\boldsymbol{GR}_0) = \boldsymbol{Q}_1\boldsymbol{R}_1 = (\bar{\boldsymbol{Q}} \mid \bar{\boldsymbol{q}}) \begin{pmatrix} \bar{\boldsymbol{R}} \\ \hline \boldsymbol{o}^\mathsf{T} \end{pmatrix} = \bar{\boldsymbol{Q}}\bar{\boldsymbol{R}}.$$

Algorithmus 2: Gegeben ist die Faktorisierung $\boldsymbol{A} = \hat{\boldsymbol{Q}}\hat{\boldsymbol{R}}$. Bilde $\hat{\boldsymbol{R}}_0 = \begin{pmatrix} \boldsymbol{R}_0 \\ \hline \boldsymbol{O} \end{pmatrix} \in \mathbf{R}^{m,n-1}$ analog zu Algorithmus 1 aus $\hat{\boldsymbol{R}}$ durch Streichen der l-ten Spalte. Bestimme $n - l$ Givens-Drehungen $\boldsymbol{G}_{l,l+1}, \ldots, \boldsymbol{G}_{n-1,n}$ so, daß

$$\boldsymbol{G}\hat{\boldsymbol{R}}_0 := \boldsymbol{G}_{n-1,n} \cdots \boldsymbol{G}_{l,l+1}\hat{\boldsymbol{R}}_0 =: \overset{\wedge}{\boldsymbol{R}} =: \begin{pmatrix} \bar{\boldsymbol{R}} \\ \hline \boldsymbol{O} \end{pmatrix} \in \mathbf{R}^{m,n-1}$$

von oberer Dreiecksform ist. Bilde $\overset{\wedge}{\boldsymbol{Q}} := \hat{\boldsymbol{Q}}\boldsymbol{G}^\mathsf{T}$ und schreibe (18) in der Form

$$\bar{\boldsymbol{A}} = \hat{\boldsymbol{Q}}\hat{\boldsymbol{R}}_0 = (\hat{\boldsymbol{Q}}\boldsymbol{G}^\mathsf{T})(\boldsymbol{G}\hat{\boldsymbol{R}}_0) = \overset{\wedge}{\boldsymbol{Q}}\overset{\wedge}{\boldsymbol{R}}.$$

10.3.4. *Aufdatierung bei Hinzufügen einer Zeile*

Aufgabe: Bestimme Faktorisierung von

$$\bar{\boldsymbol{A}} = \begin{pmatrix} \boldsymbol{a}^\mathsf{T} \\ \hline \boldsymbol{A} \end{pmatrix}, \quad \boldsymbol{a} \in \mathbf{R}^n. \tag{20}$$

Algorithmus 1: Gegeben ist die Faktorisierung $\boldsymbol{A} = \boldsymbol{QR}$. Schreibe (20) in der Form

$$\bar{\boldsymbol{A}} = \begin{pmatrix} \boldsymbol{a}^\mathsf{T} \\ \hline \boldsymbol{QR} \end{pmatrix} = \left(\begin{array}{c|c} 1 & \boldsymbol{o}^\mathsf{T} \\ \hline \boldsymbol{o} & \boldsymbol{Q} \end{array}\right) \begin{pmatrix} \boldsymbol{a}^\mathsf{T} \\ \hline \boldsymbol{R} \end{pmatrix} =: \boldsymbol{Q}_1\boldsymbol{R}_1 \tag{21}$$

mit der spaltenorthonormalen Matrix $\boldsymbol{Q}_1$ und der oberen Hessenbergmatrix $\boldsymbol{R}_1$.

Bestimme n Givens-Drehungen $G_{12}, \ldots, G_{n,n+1}$ so, daß

$$G R_1 := G_{n,n+1} \cdots G_{12} R_1 =: \begin{pmatrix} \bar{R} \\ \hline o^T \end{pmatrix}$$

von oberer Dreiecksform ist. Schreibe (21) mit der spaltenorthonormalen Matrix

$$Q_1 G^T =: (\bar{Q} \mid \bar{q})$$

in der Form

$$\bar{A} = (Q_1 G^T)(G R_1) = (\bar{Q} \mid \bar{q}) \begin{pmatrix} \bar{R} \\ \hline o^T \end{pmatrix} = \bar{Q}\bar{R}.$$

Algorithmus 2: Gegeben ist die Faktorisierung $A = \hat{Q}\hat{R}$. Schreibe (20) in der Form

$$A = \begin{pmatrix} a^T \\ \hline \hat{Q}\hat{R} \end{pmatrix} = \left(\begin{array}{c|c} 1 & o^T \\ \hline o & \hat{Q} \end{array}\right) \begin{pmatrix} a^T \\ \hline \hat{R} \end{pmatrix} =: \hat{Q}_1 \hat{R}_1 \tag{22}$$

mit orthogonalem $\hat{Q}_1$ und der oberen Hessenbergmatrix $\hat{R}_1$. Transformiere $\hat{R}_1$ mittels n Givens-Drehungen $G_{12}, \ldots, G_{n,n+1}$ auf obere Dreiecksform

$$G\hat{R}_1 := G_{n,n+1} \cdots G_{12} \hat{R}_1 =: \hat{\bar{R}} =: \begin{pmatrix} \bar{R} \\ \hline O \end{pmatrix}.$$

Schreibe (22) mit der orthogonalen Matrix $\hat{\bar{Q}} := \hat{Q}_1 G^T$ in der Form

$$\bar{A} = (\hat{Q}_1 G^T)(G\hat{R}_1) = \hat{\bar{Q}}\hat{\bar{R}}. \quad \square$$

10.3.5. *Aufdatierung bei Streichen einer Zeile*

Aufgabe: Bestimme Faktorisierung der Matrix $\bar{A}$, die aus A durch Streichen der l-ten Zeile ($1 \leq l \leq m$) entsteht, d. h.

$$\bar{A} = \begin{bmatrix} -a^{1T} & - \\ \vdots & \\ -a^{l-1T} & - \\ -a^{l+1T} & - \\ \vdots & \\ -a^{mT} & - \end{bmatrix}, \quad \text{wobei} \quad A = \begin{bmatrix} -a^{1T} & - \\ \vdots & \\ -a^{lT} & - \\ - & \vdots \ - \\ -a^{mT} & - \end{bmatrix}, \quad m > n. \tag{23}$$

Algorithmus 1: Gegeben ist die Faktorisierung $A = QR$

S0: Bestimme Permutationsmatrix P_Z so, daß

$$\tilde{A} := P_Z A = \begin{bmatrix} -a^T & - \\ -a^{1T} & - \\ - & \vdots \ - \\ -a^{l-1T} & - \\ -a^{l-1T} & - \\ - & \vdots \ - \\ -a^{mT} & - \end{bmatrix} = \begin{pmatrix} a^T \\ \hline \bar{A} \end{pmatrix} \quad \text{mit} \quad a := a^l \tag{24}$$

gilt. Dann ist

$$\tilde{\boldsymbol{A}} = \boldsymbol{P}_Z\boldsymbol{A} = \boldsymbol{P}_Z\boldsymbol{Q}\boldsymbol{R} = \tilde{\boldsymbol{Q}}\boldsymbol{R} \quad \text{mit} \quad \tilde{\boldsymbol{Q}} := \boldsymbol{P}_Z\boldsymbol{Q}. \tag{25}$$

S1: Betrachte die Rang-1-Modifikation

$$\boldsymbol{A}_1 := \begin{pmatrix} \boldsymbol{o}^\mathsf{T} \\ \hline \boldsymbol{A} \end{pmatrix} = \tilde{\boldsymbol{A}} - \boldsymbol{e}^1\boldsymbol{a}^\mathsf{T}. \tag{26}$$

Zerlege $\boldsymbol{e}^1$ gemäß

$$\boldsymbol{e}^1 = \tilde{\boldsymbol{Q}}\boldsymbol{r} + \boldsymbol{p}, \qquad \boldsymbol{r} := \tilde{\boldsymbol{Q}}^\mathsf{T}\boldsymbol{e}^1, \qquad \boldsymbol{p} := \boldsymbol{e}^1 - \tilde{\boldsymbol{Q}}\boldsymbol{r}. \tag{27}$$

Dabei ist $\boldsymbol{r}^\mathsf{T}$ die erste Zeile von $\tilde{\boldsymbol{Q}}$, braucht also nicht gesondert berechnet zu werden. Setze

$$\varrho := \|\boldsymbol{p}\| = \sqrt{1 - \|\tilde{\boldsymbol{Q}}^\mathsf{T}\boldsymbol{e}^1\|^2} = \sqrt{1 - \|\boldsymbol{r}\|^2}. \tag{28}$$

Falls $\varrho > 0$, setze $\boldsymbol{q} := \boldsymbol{p}/\varrho$. Andernfalls bestimme $\tilde{\boldsymbol{p}} \neq \boldsymbol{o}$, $\tilde{\boldsymbol{p}} \perp \mathcal{R}(\tilde{\boldsymbol{Q}})$ wie folgt: Wähle Index $s \in \{2, \ldots, m\}$ mit $\|\tilde{\boldsymbol{Q}}^\mathsf{T}\boldsymbol{e}^s\| < 1$. Ein solcher Index existiert, denn wegen $\|\tilde{\boldsymbol{Q}}^\mathsf{T}\boldsymbol{e}^i\| \leqq 1$ $(i = 1, \ldots, m)$ wäre sonst $\|\tilde{\boldsymbol{Q}}^\mathsf{T}\boldsymbol{e}^i\| = 1$ $(i = 2, \ldots, m)$. Da nach (28) im Fall $\varrho = 0$ auch $\|\tilde{\boldsymbol{Q}}^\mathsf{T}\boldsymbol{e}^1\| = 1$ gilt, folgte $n = \|\tilde{\boldsymbol{Q}}\|_F^2 = \sum_{i=1}^{m} \|\tilde{\boldsymbol{Q}}^\mathsf{T}\boldsymbol{e}^i\|^2 = m$ im Widerspruch zu $m > n$. Setze

$$\tilde{\boldsymbol{r}} := \tilde{\boldsymbol{Q}}^\mathsf{T}\boldsymbol{e}^s, \qquad \tilde{\boldsymbol{p}} := \boldsymbol{e}^s - \tilde{\boldsymbol{Q}}\tilde{\boldsymbol{r}}, \qquad \tilde{\varrho} := \|\tilde{\boldsymbol{p}}\|, \qquad \boldsymbol{q} := \tilde{\boldsymbol{p}}/\tilde{\varrho}, \tag{29}$$

wobei $\tilde{\varrho}^2 = 1 - \|\tilde{\boldsymbol{Q}}^\mathsf{T}\boldsymbol{e}^s\|^2 > 0$ ist nach Wahl von s. In beiden Fällen ist

$$\boldsymbol{p} = \boldsymbol{e}^1 - \tilde{\boldsymbol{Q}}\boldsymbol{r} = \varrho\boldsymbol{q}, \qquad \|\boldsymbol{q}\| = 1, \qquad \boldsymbol{q} \perp \mathcal{R}(\boldsymbol{Q}).$$

Schreibe (26) in der Form

$$\begin{aligned} \boldsymbol{A}_1 = \tilde{\boldsymbol{A}} - \boldsymbol{e}^1\boldsymbol{a}^\mathsf{T} &= \tilde{\boldsymbol{Q}}\boldsymbol{R} - (\tilde{\boldsymbol{Q}}\boldsymbol{r} + \varrho\boldsymbol{q})\,\boldsymbol{a}^\mathsf{T} = (\tilde{\boldsymbol{Q}} \mid \boldsymbol{q}) \left(\begin{pmatrix} \boldsymbol{R} \\ \hline \boldsymbol{o}^\mathsf{T} \end{pmatrix} - \begin{pmatrix} \boldsymbol{r} \\ \hline \varrho \end{pmatrix} \boldsymbol{a}^\mathsf{T} \right) \\ &=: \boldsymbol{Q}_1(\boldsymbol{R}_0 - \boldsymbol{r}^0\boldsymbol{a}^\mathsf{T}) \end{aligned} \tag{30}$$

mit der spaltenorthonormalen Matrix $\boldsymbol{Q}_1 := (\tilde{\boldsymbol{Q}} \mid \boldsymbol{q})$ und

$$\boldsymbol{R}_0 := \begin{pmatrix} \boldsymbol{R} \\ \hline \boldsymbol{o}^\mathsf{T} \end{pmatrix}, \quad \boldsymbol{r}^0 := \begin{pmatrix} \boldsymbol{r} \\ \hline \varrho \end{pmatrix}.$$

Dabei gilt

$$\boldsymbol{Q}_1^\mathsf{T}\boldsymbol{e}^1 = \begin{pmatrix} \tilde{\boldsymbol{Q}}^\mathsf{T} \\ \hline \boldsymbol{q}^\mathsf{T} \end{pmatrix} \boldsymbol{e}^1 = \begin{pmatrix} \tilde{\boldsymbol{Q}}^\mathsf{T}\boldsymbol{e}^1 \\ \hline \boldsymbol{q}^\mathsf{T}\boldsymbol{e}^1 \end{pmatrix} = \begin{pmatrix} \boldsymbol{r} \\ \hline \varrho \end{pmatrix} = \boldsymbol{r}^0, \tag{31}$$

denn im Fall $\varrho > 0$ ist nach (28), (29)

$$q_1 := \boldsymbol{e}^{1\mathsf{T}}\boldsymbol{q} = \boldsymbol{e}^{1\mathsf{T}}\boldsymbol{p}/\varrho = (1 - \|\tilde{\boldsymbol{Q}}^\mathsf{T}\boldsymbol{e}^1\|^2)/\varrho = \varrho,$$

und im Fall $\varrho = 0$ gilt wegen der Spaltenorthonormalität von $\boldsymbol{Q}_1$

$$1 \geqq \|\tilde{\boldsymbol{Q}}_1^\mathsf{T}\boldsymbol{e}^1\|^2 = \|\tilde{\boldsymbol{Q}}^\mathsf{T}\boldsymbol{e}^1\|^2 + q_1^2 = 1 + q_1^2,$$

also $q_1 = 0 = \varrho$.

S2: Wähle G wie in 10.3.1 als Folge von Givens-Drehungen so, daß

$$G\boldsymbol{r}^0 := G_{12} \cdots G_{n,n+1}\boldsymbol{r}^0 = \varrho_1 \boldsymbol{e}^1, \qquad \varrho_1 = \pm\|\boldsymbol{r}^0\| = \pm 1 \tag{32}$$

gilt. Dann ist

$$GR_0 = G_{12} \cdots G_{n,n+1}R_0 =: R_1$$

von oberer Hessenbergform. Schreibe (30) in der Form

$$A_1 = (Q_1 G^\mathsf{T})\,(GR_0 - G\boldsymbol{r}^0\boldsymbol{a}^\mathsf{T}) = (Q_1 G^\mathsf{T})\,(R_1 - \varrho_1 \boldsymbol{e}^1\boldsymbol{a}^\mathsf{T}) =: Q_2 R_2 \tag{33}$$

mit der oberen Hessenbergmatrix

$$R_2 := R_1 - \varrho_1 \boldsymbol{e}^1\boldsymbol{a}^\mathsf{T} =: \begin{pmatrix} \bar{\boldsymbol{r}}^\mathsf{T} \\ \hline \bar{R} \end{pmatrix}$$

und der spaltenorthonormalen Matrix

$$Q_2 := Q_1 G^\mathsf{T} = \left(\begin{array}{c|c} \pm 1 & \boldsymbol{o}^\mathsf{T} \\ \hline \bar{\boldsymbol{q}} & \bar{Q} \end{array}\right) = \left(\begin{array}{c|c} \pm 1 & \boldsymbol{o}^\mathsf{T} \\ \hline \boldsymbol{o} & \bar{Q} \end{array}\right).$$

Man beachte dabei, daß aus (31), (32) die Beziehung $Q_2^\mathsf{T}\boldsymbol{e}^1 = GQ_1^\mathsf{T}\boldsymbol{e}^1 = G\boldsymbol{r}^0 = \varrho_1\boldsymbol{e}^1$ folgt und daß $\|Q_2\boldsymbol{e}^1\|^2 = 1 + \|\bar{\boldsymbol{q}}\|^2 = 1$, also $\bar{\boldsymbol{q}} = \boldsymbol{o}$ gilt, denn Q_2 hat wie Q_1 normierte Spalten. Damit geht (33) in

$$\begin{pmatrix} \boldsymbol{o}^\mathsf{T} \\ \hline \bar{A} \end{pmatrix} = A_1 = Q_2 R_2 = \left(\begin{array}{c|c} \pm 1 & \boldsymbol{o}^\mathsf{T} \\ \hline \boldsymbol{o} & \bar{Q} \end{array}\right) \begin{pmatrix} \bar{\boldsymbol{r}} \\ \hline \bar{R} \end{pmatrix} = \begin{pmatrix} \pm\bar{\boldsymbol{r}} \\ \hline \bar{Q}\bar{R} \end{pmatrix},$$

also $\bar{\boldsymbol{r}} = \boldsymbol{o}$ und

$$\bar{A} = \bar{Q}\bar{R}$$

über.

Algorithmus 2: Gegeben ist die Faktorisierung $A = \hat{Q}\hat{R}$.

S0: Bestimme P_Z wie im Schritt S0 von Algorithmus 1. Dann gilt

$$\tilde{A} = P_Z A = P_Z\hat{Q}\hat{R} = \tilde{Q}\hat{R} \quad \text{mit} \quad \tilde{Q} := P_Z\hat{Q}.$$

S1: Schreibe (26) in der Form

$$A_1 = \tilde{A} - \boldsymbol{e}^1\boldsymbol{a}^\mathsf{T} = \tilde{Q}(\hat{R} - \hat{\boldsymbol{r}}\boldsymbol{a}^\mathsf{T}) \quad \text{mit} \quad \hat{\boldsymbol{r}} := \tilde{Q}^\mathsf{T}\boldsymbol{e}^1 \tag{34}$$

S2: Bestimme $m - 1$ Givens-Drehungen so, daß

$$G\hat{\boldsymbol{r}} = G_{12} \cdots G_{m-1,m}\hat{\boldsymbol{r}} = \varrho_1\boldsymbol{e}^1, \qquad \varrho_1 = \pm\|\hat{\boldsymbol{r}}\| = \pm\|\boldsymbol{e}^1\| = \pm 1$$

gilt. Dann ist

$$G\hat{R} := G_{12} \cdots G_{m-1,m}\hat{R} = G_{12} \cdots G_{\hat{l},\hat{l}+1}\hat{R} =: \hat{R}_1, \qquad \hat{l} := \min\{m - 1, n\},$$

von oberer Hessenbergform. Schreibe (34) als

$$A_1 = (\tilde{Q}G^\mathsf{T})\,(G\hat{R} - G\hat{\boldsymbol{r}}\boldsymbol{a}^\mathsf{T}) = (\tilde{Q}G^\mathsf{T})\,(\hat{R}_1 - \varrho_1\boldsymbol{e}^1\boldsymbol{a}^\mathsf{T}) =: \hat{Q}_2\hat{R}_2 \tag{35}$$

mit der oberen Hessenbergmatrix

$$\hat{\boldsymbol{R}}_2 := \hat{\boldsymbol{R}}_1 - \varrho_1 \boldsymbol{e}^1 \boldsymbol{a}^T = \begin{pmatrix} \bar{\boldsymbol{r}}^T \\ \hline \hat{\hat{\boldsymbol{R}}} \end{pmatrix}$$

und der orthogonalen Matrix

$$\hat{\boldsymbol{Q}}_2 := \tilde{\boldsymbol{Q}} \boldsymbol{G}^T = \left(\begin{array}{c|c} \pm 1 & \boldsymbol{o}^T \\ \hline \bar{\boldsymbol{q}} & \hat{\hat{\boldsymbol{Q}}} \end{array}\right) = \left(\begin{array}{c|c} \pm 1 & \boldsymbol{o}^T \\ \hline \boldsymbol{o} & \hat{\hat{\boldsymbol{Q}}} \end{array}\right);$$

man beachte $\hat{\boldsymbol{Q}}_2^T \boldsymbol{e}^1 = \varrho_1 \boldsymbol{e}^1$ und die Spaltennormiertheit von $\hat{\boldsymbol{Q}}_2$. Damit geht (35) in

$$\begin{pmatrix} \boldsymbol{o}^T \\ \hline \bar{\boldsymbol{A}} \end{pmatrix} = \boldsymbol{A}_1 = \left(\begin{array}{c|c} \pm 1 & \boldsymbol{o}^T \\ \hline \boldsymbol{o} & \hat{\hat{\boldsymbol{Q}}} \end{array}\right) \begin{pmatrix} \bar{\boldsymbol{r}}^T \\ \hline \hat{\hat{\boldsymbol{R}}} \end{pmatrix} = \begin{pmatrix} \pm \bar{\boldsymbol{r}}^T \\ \hline \hat{\hat{\boldsymbol{Q}}} \hat{\hat{\boldsymbol{R}}} \end{pmatrix},$$

also $\bar{\boldsymbol{r}} = \boldsymbol{o}$ und

$$\bar{\boldsymbol{A}} = \hat{\hat{\boldsymbol{Q}}} \hat{\hat{\boldsymbol{R}}} \tag{36}$$

über. □

Die Algorithmen aus 10.3.5 sind auf den Fall $\boldsymbol{u} = -\boldsymbol{e}^1$ zugeschnittene Varianten von 10.3.1, wobei im Algorithmus 1 der für $\boldsymbol{p} = \boldsymbol{o}$ benötigte Vektor $\boldsymbol{q} \perp \mathcal{R}(\boldsymbol{Q})$ jedoch explizit berechnet wird.

10.3.6. *Bemerkung.* Die Anwendung der beschriebenen Aufdatierungsalgorithmen hängt von der Art der Faktorisierung und der Aufgabenstellung ab:

(i) Für Faktorisierungen $\boldsymbol{A}_k = \boldsymbol{Q}_k \boldsymbol{R}_k$ nach Fall 1 sollten beide Faktoren $\boldsymbol{Q}_k$ und $\boldsymbol{R}_k$ explizit mitgeführt werden. Aus ihnen lassen sich alle weiteren benötigten Größen wie Quadratmittellösungen $\boldsymbol{x}^k$, Residuen $\boldsymbol{b}^k - \boldsymbol{A}\boldsymbol{x}^k$ u. a. einfach berechnen. Der Aufwand für die Aufdatierung ist vom Typ (3), also klein.

(ii) Die explizite Mitführung beider Faktoren $\boldsymbol{Q}_k, \boldsymbol{R}_k$ bzw. $\hat{\boldsymbol{Q}}_k, \hat{\boldsymbol{R}}_k$ empfiehlt sich ebenfalls, wenn eine Folge regulärer linearer Gleichungssysteme $\boldsymbol{A}_k \boldsymbol{x}^k = \boldsymbol{b}^k$ mit

$$\boldsymbol{A}_{k+1} = \boldsymbol{A}_k + \boldsymbol{u}^k (\boldsymbol{v}^k)^T \in \mathbf{R}^{n,n} \tag{37}$$

zu lösen ist und die rechten Seiten nicht a priori bekannt sind. Man beachte dabei, daß die Faktorisierungen gemäß Fall 1 und Fall 2 für $m = n$ bis auf das Vorzeichen der Spalten von $\boldsymbol{Q}$ und Zeilen von $\boldsymbol{R}$ identisch sind. Probleme dieses Typs treten in der Optimierung und bei der Lösung nichtlinearer Gleichungssysteme auf.

(iii) Für Faktorisierungen $\boldsymbol{A}_k = \hat{\boldsymbol{Q}}_k \hat{\boldsymbol{R}}_k$ gemäß Fall 2 ist die explizite Mitführung von $\hat{\boldsymbol{Q}}_k$ aus Aufwandsgründen nicht zu empfehlen, sofern $m \gg n$ ist. Ein Kompromiß für nicht zu großes K wäre die Berechnung von $\hat{\boldsymbol{Q}}_1$ in Produktform nach dem Householder- oder Givens-Verfahren und das Abspeichern der zugehörigen Daten v_{ik} bzw. ξ_{ik} auf dem Platz von $\boldsymbol{A}_1$ nach 10.2.4. Die bei der Aufdatierung benötigten Givens-Drehungen $\boldsymbol{G}_{ij}$ werden durch die zugehörigen ξ_{ij} repräsentiert; für diese muß dann zusätzlicher Speicherplatz bereitgestellt werden. Der Mehraufwand an Speicherplatz

zum Abspeichern der Produktform von $\hat{Q}_k$ und an Rechenoperationen zur Berechnung von $\hat{Q}_k^T b$ bzw. $\hat{Q}_k f$ in zu 10.2.6 analoger Form ist bei dieser Version proportional zu K.

(iv) Wenn die vorkommenden rechten Seiten b^j, die Vektoren u^j der Rang-1-Modifikationen und die hinzuzufügenden Spalten a^j a priori bekannt sind, können die Dreiecksfaktoren $\hat{R}_k = \hat{Q}_k^T A_k$, die transformierten Größen $\hat{Q}_k^T(b^j, u^j, a^j)$ und damit die Quadratmittellösungen x^k und die Residuumsnormen $nr_k := \|b^k - A_k x^k\|$ ohne explizite Kenntnis des Orthogonalfaktors $\hat{Q}_k$ berechnet werden. Dasselbe trifft für das Streichen von Spalten und Hinzufügen von Zeilen zu. Das Streichen der l-ten Zeile nach dem in 10.3.5 angegebenen Algorithmus 2 erfordert die Kenntnis von $\tilde{Q}^T e^l = \hat{Q}^T e^l$, und dieser Vektor kann ebenfalls wie b^j, u^j oder a^j aufdatiert werden. Es gibt jedoch andere Algorithmen, die Zeilenstreichungen ohne Rückgriff auf $\hat{Q}^T e^l$ ermöglichen. Für Details und Literaturhinweise sei auf das nachfolgende Beispiel 10.3.7 und B 10.6 verwiesen.

(v) Die Realisierung der Givens-Drehungen kann sowohl in expliziter wie auch in impliziter Form erfolgen. Im letzten Fall sind die Skalierungsfaktoren mitzuführen, siehe B 10.3.

10.3.7. *Beispiel.* Gegeben sind die Quadratmittelprobleme

$$A_k x^k \cong b \qquad (k = 1, 2, 3) \tag{38}$$

mit

$$A_1 := (a^1, a^2, a^3), \qquad A_2 := (a^1, a^2, a^3, a^4), \qquad A_3 := A_2 + uv^T. \tag{39}$$

Sie werden durch die Eingangsdaten

$$B := (a^1, a^2, a^3, a^4, u, b) = (A_1 \mid a^4 \mid u \mid b) \in \mathbf{R}^{m,6} \quad \text{und} \quad v \in \mathbf{R}^4 \tag{40}$$

dargestellt. Wenn $\hat{Q} \in \mathbf{R}^{m,m}$ irgendeine orthogonale Matrix bezeichnet, sind die durch die transformierten Eingangsdaten

$$\hat{Q}^T B = (\hat{Q}^T A_1 \mid \hat{Q}^T a^4 \mid \hat{Q}^T u \mid \hat{Q}^T b)$$

festgelegten Probleme zu den durch B charakterisierten äquivalent: Sie haben dieselben Lösungen x^k und dieselben Residuumsnormen $nr_k = \|\hat{Q}^T(b - A_k x^k)\| = \|b - A_k x^k\|$. Die Probleme (38), (39) können daher wie folgt gelöst werden:

S1 ($k = 1$): Bestimme $\hat{Q}_1^T$ als Produkt von Householder-Spiegelungen oder Givens-Drehungen gemäß 10.2.4 so, daß

$$B_1 := \hat{Q}_1^T B =: (\hat{R}_1 \mid \hat{r}^{4,1} \mid \hat{r}^1 \mid \hat{c}^1) = \hat{Q}_1^T(A_1 \mid a^4 \mid u \mid b)$$

mit einer oberen Dreiecksmatrix $\hat{R}_1 = \hat{Q}_1^T A_1$ gilt.
Berechne x^1 und nr_1 aus $\hat{R}_1$ und $\hat{c}^1 = \hat{Q}_1^T b$ gemäß 10.2.9.

S2 ($k = 2$): Bestimme H entsprechend 10.3.2 (Algorithmus 2, S2) mit

$$\hat{r} := \hat{r}^{4,1} = \hat{Q}_1^T a^4.$$

Transformiere B_1 in

$$B_2 := HB_1 =: (\hat{R}_2 \mid \hat{r}^2 \mid \hat{c}^2) = H\hat{Q}_1^T B = \hat{Q}_2^T\big((A_1 \mid a^4) \mid u \mid b\big) = \hat{Q}_2^T(A_2 \mid u \mid b)$$

mit $\hat{\boldsymbol{Q}}_2 := \hat{\boldsymbol{Q}}_1 \boldsymbol{H}^\mathsf{T}$ und der oberen Dreiecksmatrix $\hat{\boldsymbol{R}}_2 := (\boldsymbol{H}\hat{\boldsymbol{R}}_1 \mid \boldsymbol{H}\hat{\boldsymbol{r}}^{4,1}) = \hat{\boldsymbol{Q}}_2^\mathsf{T}\boldsymbol{A}_2$.
Berechne $\boldsymbol{x}^2$ und $n\boldsymbol{r}_2$ aus $\hat{\boldsymbol{R}}_2$ und $\hat{\boldsymbol{c}}^2 = \hat{\boldsymbol{Q}}_2^\mathsf{T}\boldsymbol{b}$.

S3 ($k = 3$): Bestimme $\boldsymbol{G}$ entsprechend 10.3.1 (Algorithmus 2, S2) mit

$$\hat{\boldsymbol{r}} := \hat{\boldsymbol{r}}^2 = \hat{\boldsymbol{Q}}_2^\mathsf{T}\boldsymbol{u}.$$

Transformiere $\boldsymbol{B}_2$ in

$$\boldsymbol{B}_{21} := \boldsymbol{G}\boldsymbol{B}_2 =: (\hat{\boldsymbol{R}}_{21} \mid \varrho_1\boldsymbol{e}^1 \mid \hat{\boldsymbol{c}}^{2,1}) = \boldsymbol{G}\hat{\boldsymbol{Q}}_2^\mathsf{T}\boldsymbol{B} = \hat{\boldsymbol{Q}}_{21}^\mathsf{T}(\boldsymbol{A}_2 \mid \boldsymbol{u} \mid \boldsymbol{b})$$

mit der oberen Hessenbergmatrix $\hat{\boldsymbol{R}}_{21} := \boldsymbol{G}\hat{\boldsymbol{R}}_2$ und $\hat{\boldsymbol{Q}}_{21} := \hat{\boldsymbol{Q}}_2\boldsymbol{G}^\mathsf{T}$.
Führe die Rang-1-Modifikation $\boldsymbol{A}_3 = \boldsymbol{A}_2 + \boldsymbol{u}\boldsymbol{v}^\mathsf{T}$ mit den durch $\hat{\boldsymbol{Q}}_{21}^\mathsf{T}$ transformierten Daten $\boldsymbol{B}_{21}$ gemäß

$$\tilde{\boldsymbol{B}}_{22} := (\hat{\boldsymbol{R}}_{21} + \varrho_1\boldsymbol{e}^1\boldsymbol{v}^\mathsf{T} \mid \varrho_1\boldsymbol{e}^1 \mid \hat{\boldsymbol{c}}^{2,1}) = \hat{\boldsymbol{Q}}_{21}^\mathsf{T}(\boldsymbol{A}_2 + \boldsymbol{u}\boldsymbol{v}^\mathsf{T} \mid \boldsymbol{u} \mid \boldsymbol{b}) = \hat{\boldsymbol{Q}}_{21}^\mathsf{T}(\boldsymbol{A}_3 \mid \boldsymbol{u} \mid \boldsymbol{b})$$

durch, wobei $\hat{\boldsymbol{R}}_{22} := \hat{\boldsymbol{R}}_{21} + \varrho_1\boldsymbol{e}^1\boldsymbol{v}^\mathsf{T}$ von oberer Hessenbergform ist, und streiche die nicht mehr benötigte fünfte Spalte von $\tilde{\boldsymbol{B}}_{22}$, d. h., setze

$$\boldsymbol{B}_{22} := (\hat{\boldsymbol{R}}_{22} \mid \hat{\boldsymbol{c}}^{2,1}) = \hat{\boldsymbol{Q}}_{21}^\mathsf{T}(\boldsymbol{A}_3 \mid \boldsymbol{b}).$$

Bestimme $\tilde{\boldsymbol{G}}$ entsprechend 10.3.1 (Algorithmus 2, S3) mit $\hat{\boldsymbol{R}}_2 := \hat{\boldsymbol{R}}_{22}$.
Transformiere $\boldsymbol{B}_{22}$ in

$$\boldsymbol{B}_3 := \tilde{\boldsymbol{G}}\boldsymbol{B}_{22} = (\hat{\boldsymbol{R}}_3 \mid \hat{\boldsymbol{c}}^3) = \tilde{\boldsymbol{G}}\hat{\boldsymbol{Q}}_{21}^\mathsf{T}\overline{\boldsymbol{B}} = \hat{\boldsymbol{Q}}_3^\mathsf{T}(\boldsymbol{A}_3 \mid \boldsymbol{b})$$

mit $\hat{\boldsymbol{Q}}_3 := \hat{\boldsymbol{Q}}_{21}\tilde{\boldsymbol{G}}^\mathsf{T} = \hat{\boldsymbol{Q}}_2\boldsymbol{G}^\mathsf{T}\tilde{\boldsymbol{G}}^\mathsf{T}$ und der oberen Dreiecksmatrix $\hat{\boldsymbol{R}}_3 = \hat{\boldsymbol{Q}}_3^\mathsf{T}\boldsymbol{A}_3$, wobei $\overline{\boldsymbol{B}}$ aus $\boldsymbol{B}$ durch Streichen der fünften Spalte entsteht.
Berechne $\boldsymbol{x}^3$ und $n\boldsymbol{r}_3$ aus $\hat{\boldsymbol{R}}_3$ und $\hat{\boldsymbol{c}}^3 = \hat{\boldsymbol{Q}}_3^\mathsf{T}\boldsymbol{b}$. □

Die Grundidee des Vorgehens in 10.3.7 ist die folgende: Statt die Orthogonalfaktoren $\hat{\boldsymbol{Q}}_k$ nach der Vorschrift

$$\hat{\boldsymbol{Q}}_{k+1} := \hat{\boldsymbol{Q}}_k\boldsymbol{G}_k^\mathsf{T} \tag{41}$$

mit den im k-ten Schritt vorkommenden orthogonalen Transformationen $\boldsymbol{G}_k^\mathsf{T}$ explizit aufzudatieren und damit die transformierten Eingangsdaten

$$\boldsymbol{B}_{k+1} := \hat{\boldsymbol{Q}}_{k+1}^\mathsf{T}\boldsymbol{B} \tag{42}$$

zu berechnen, werden diese gemäß $\boldsymbol{B}_{k+1} = (\hat{\boldsymbol{Q}}_k\boldsymbol{G}_k^\mathsf{T})^\mathsf{T}\boldsymbol{B} = \boldsymbol{G}_k\hat{\boldsymbol{Q}}_k^\mathsf{T}\boldsymbol{B}$, d. h.

$$\boldsymbol{B}_{k+1} := \boldsymbol{G}_k\boldsymbol{B}_k \tag{43}$$

aufdatiert. Die Matrix $\boldsymbol{G}_k$ ist dabei das Produkt sämtlicher auf $\hat{\boldsymbol{R}}_k$ angewandter Drehungen bzw. Spiegelungen. In dieser Weise lassen sich auch Spaltenstreichungen, Zeilenhinzufügungen und — falls $\boldsymbol{e}^{l_j}$ mit in $\boldsymbol{B}$ aufgenommen wird — das Streichen von Zeilen l_j realisieren. In allen Fällen wird weder $\hat{\boldsymbol{Q}}_k^\mathsf{T}$ noch $\boldsymbol{G}_k$ explizit benötigt; die Transformationen werden sukzessive auf das aktuelle $\boldsymbol{B}_k$ angewendet und brauchen auch nicht gespeichert zu werden.

10.3.8. *Bemerkung.* (i) Für die Realisierung der Aufdatierungsalgorithmen ist wesentlich, daß die in exakter Arithmetik gültigen Orthogonalitätsbeziehungen auch bei

Computerrechnung in ausreichendem Maße gewährleistet sind. Die Startfaktoren Q_1 bzw. $\hat{Q}_1$ müssen daher ausreichend orthonormal sein. Im Fall 1 sollte Q_1 daher entweder nach dem MGS-Verfahren mit Re-Orthogonalisierung oder durch Householder- bzw. Givens-Orthogonalisierung und spaltenweise Berechnung von $Q = (q^1, \ldots, q^n)$ bestimmt werden; im Fall 2 empfiehlt sich die Householder- oder Givens-Orthogonalisierung.

In den zum Fall 1 gehörenden Aufdatierungsalgorithmen müssen die dort zu berechnenden orthogonalen Zerlegungen (4), (5) ein p liefern, das genügend orthogonal zu $\mathcal{R}(Q)$ ist. Wenn p klein im Verhältnis zu d ist, tritt bei der Berechnung von $p = d - Qr$ Auslöschung auf, die die Orthogonalität wesentlich beeinträchtigen kann. In diesem Fall sollte re-orthogonalisiert werden, etwa nach der Vorschrift

$r := Q^T d,\ p := d - Qr$

f $\|p\| \leqq 0.1\ \|d\|$ then $[s := Q^T p,\ p := p - Qs,\ r := r + s]$.

Dabei ist es günstig, wenn die Realisierung in modifizierter Form analog zu 10.1.5 erfolgt. Der Faktor 0.1 hat dabei nur exemplarischen Charakter. Er bewirkt, daß die aufwandsverdoppelnde Re-Orthogonalisierung nur dann ausgeführt wird, wenn die angenäherte Orthogonalität um mehr als den Faktor 10 gestört werden könnte.

In 10.3.5 (Algorithmus 1, S1) sollte die Bedingung $\|\tilde{Q}^T e^i\| < 1$ als erfüllt angesehen werden, wenn die berechnete Norm der i-ten Zeile von $\tilde{Q}$ der Bedingung

$$\|\tilde{Q}^T e^i\| \leqq 1 - mv$$

genügt, andernfalls wird $\|\tilde{Q}^T e^i\|$ als 1 angesehen. In Entsprechung zur obigen Regel sollte p bzw. $\tilde{p}$ re-orthogonalisiert werden, wenn

$$\sqrt{0.99} \leqq \|\tilde{Q}^T e^i\| \leqq 1 - mv \quad \text{für} \quad i = 1 \text{ bzw. } i = s$$

gilt.

(ii) Man kann zeigen, daß die Aufdatierungsalgorithmen für das Hinzufügen von Spalten oder Zeilen in folgendem Sinne numerisch gutartig sind: Wenn $A + \delta A = QR$ mit kleinem δA gilt, ist auch $\bar{A} + \delta\bar{A} = \bar{Q}\bar{R}$ mit kleinem $\delta\bar{A}$. Wenn δA spaltenweise klein in bezug auf A ist, trifft dies auch für $\delta\bar{A}$ zu. Gutartigkeit liegt auch für Rang-1-Modifikationen und das Streichen von Spalten oder Zeilen vor, sofern die modifizierte Matrix $\bar{A}$ in der Norm nicht signifikant kleiner als A ist. Wenn $\|\bar{A}\|$ wesentlich kleiner als $\|A\|$ ist, wird $\|\delta\bar{A}\|$ im allgemeinen nicht klein in bezug auf $\|\bar{A}\|$ sein, so daß die berechneten Faktoren $\bar{Q}, \bar{R}$ eine weit von A entfernte Matrix reproduzieren. In diesem Fall empfiehlt sich eine Neuberechnung der Faktorisierung aus $\bar{A}$. Aus Platzgründen soll auf die präzise Formulierung und den Beweis dieser Aussagen verzichtet werden. □

Übungsaufgaben

Ü 10.3.1. Man überlege sich: Wenn $\bar{A}$ aus A durch Modifikation der l-ten Spalte gemäß

$$\bar{A} = A + ue^{lT} \tag{44}$$

entsteht und $\boldsymbol{P}_S$ diejenige Permutationsmatrix bezeichnet, die $\boldsymbol{A} = (\boldsymbol{a}^1, \ldots, \boldsymbol{a}^l, \ldots, \boldsymbol{a}^n)$ in $\hat{\boldsymbol{A}} = \boldsymbol{A}\boldsymbol{P}_S^\mathsf{T} = (\boldsymbol{a}^1, \ldots, \boldsymbol{a}^{l-1}, \boldsymbol{a}^{l+1}, \ldots, \boldsymbol{a}^n, \boldsymbol{a}^l)$ transformiert, kann eine $\boldsymbol{QR}$-Faktorisierung von $\boldsymbol{A}\boldsymbol{P}_S^\mathsf{T}$ wie folgt berechnet werden:

S1: Wie 10.3.1 (Algorithmus 1, S1)

S2: Schreibe (9) in der Form

$$\bar{\boldsymbol{A}}\boldsymbol{P}_S^\mathsf{T} = \boldsymbol{Q}_1(\boldsymbol{R}_0 + \boldsymbol{r}^0\boldsymbol{e}^{l\mathsf{T}})\,\boldsymbol{P}_S^\mathsf{T} = \boldsymbol{Q}_1(\boldsymbol{R}_0\boldsymbol{P}_S^\mathsf{T} + \boldsymbol{r}^0\boldsymbol{e}^{n\mathsf{T}}) =: \boldsymbol{Q}_2\boldsymbol{R}_2 \tag{45}$$

mit $\boldsymbol{Q}_2 := \boldsymbol{Q}_1$ und der oberen Hessenbergmatrix $\boldsymbol{R}_2 := \boldsymbol{R}_0\boldsymbol{P}_S^\mathsf{T} + \boldsymbol{r}^0\boldsymbol{e}^{n\mathsf{T}}$, die eine zu (19) analoge Gestalt hat.

S3: Bestimme $n - l + 1$ Givens-Drehungen $\tilde{\boldsymbol{G}}_{l,l+1}, \ldots, \tilde{\boldsymbol{G}}_{n,n+1}$ so, daß die Subdiagonalelemente $(\boldsymbol{R}_2)_{j+1,j}$ $(j = l, \ldots, n)$ sukzessive zu 0 gemacht werden. Dann ist

$$\tilde{\boldsymbol{G}}\boldsymbol{R}_2 := \tilde{\boldsymbol{G}}_{n,n+1} \cdots \tilde{\boldsymbol{G}}_{l,l+1}\boldsymbol{R}_2 =: \boldsymbol{R}_3 =: \begin{pmatrix} \bar{\boldsymbol{R}} \\ \hline \boldsymbol{o}^\mathsf{T} \end{pmatrix}$$

von oberer Dreiecksform, und $\boldsymbol{Q}_3 := \boldsymbol{Q}_2\tilde{\boldsymbol{G}}^\mathsf{T} =: (\bar{\boldsymbol{Q}} \mid \bar{\boldsymbol{q}})$ ist spaltenorthonormal. Schreibe (45) in der Form

$$\bar{\boldsymbol{A}}\boldsymbol{P}_S^\mathsf{T} = \boldsymbol{Q}_2\boldsymbol{R}_2 = (\boldsymbol{Q}_2\tilde{\boldsymbol{G}}^\mathsf{T})\,(\tilde{\boldsymbol{G}}\boldsymbol{R}_2) = \boldsymbol{Q}_3\boldsymbol{R}_3 = (\bar{\boldsymbol{Q}} \mid \bar{\boldsymbol{q}}) \begin{pmatrix} \bar{\boldsymbol{R}} \\ \hline \boldsymbol{o}^\mathsf{T} \end{pmatrix} = \bar{\boldsymbol{Q}}\bar{\boldsymbol{R}}.$$

In analoger Weise kann im Fall 2 vorgegangen werden. Die in 10.3.1(S2) erforderlichen Givens-Drehungen werden dabei auf Kosten der durch $\boldsymbol{P}_S$ bewirkten Spaltenvertauschungen eingespart.

Ü 10.3.2. Man überlege sich, daß im Schritt S2 von 10.3.1 und 10.3.5 die letzte Drehung $\boldsymbol{G}_{12}$ eingespart werden kann. Wie lauten dann die Formeln für $\boldsymbol{R}_2$ bzw. $\hat{\boldsymbol{R}}_2$?

Ü 10.3.3. Gegeben sei das Quadratmittelproblem $\boldsymbol{A}\boldsymbol{x} \cong \boldsymbol{b}$. Wenn $\bar{\boldsymbol{A}}\bar{\boldsymbol{x}} = \bar{\boldsymbol{b}}$ durch Streichen der l-ten Zeile $(\boldsymbol{a}^{l\mathsf{T}}, b_l)$ von $(\boldsymbol{A} \mid \boldsymbol{b})$ entsteht und $\bar{\bar{\boldsymbol{A}}}\bar{\bar{\boldsymbol{x}}} = \bar{\bar{\boldsymbol{b}}}$ durch Hinzufügen des i-fachen der l-ten Zeile gemäß

$$(\bar{\bar{\boldsymbol{A}}} \mid \bar{\bar{\boldsymbol{b}}}) = \left(\begin{array}{c|c} \mathrm{i}\boldsymbol{a}^{l\mathsf{T}} & \mathrm{i}b_l \\ \hline \boldsymbol{A} & \boldsymbol{b} \end{array}\right), \quad \mathrm{i} \text{ imaginäre Einheit mit } \mathrm{i}^2 = -1,$$

gebildet wird, gilt für jedes $\boldsymbol{x} \in \mathbf{R}^n$

$$\|\bar{\boldsymbol{b}} - \bar{\boldsymbol{A}}\boldsymbol{x}\|^2 = \|\bar{\bar{\boldsymbol{b}}} - \bar{\bar{\boldsymbol{A}}}\boldsymbol{x}\|^2.$$

In bezug auf die Residuumnorm hat das Hinzufügen des i-fachen einer Zeile also denselben Effekt wie das Streichen der Zeile. Man kann zeigen, daß die Aufdatierung der $\boldsymbol{QR}$-Faktorisierung von $\boldsymbol{A}$ bei Übergang zu $\bar{\bar{\boldsymbol{A}}}$ gemäß 10.3.4 in reeller Arithmetik ausgeführt werden kann.

Ü 10.3.4. Man ermittle die Koeffizienten K_4 und K_5 in (3) für die angegebenen Aufdatierungsalgorithmen unter der Voraussetzung, daß im Fall 1 stets mit Re-Orthogonalisierung gearbeitet wird und im Fall 2 die Matrix $\hat{\boldsymbol{Q}}_k$ nicht berechnet wird.

Bemerkungen zum Kapitel 10

B 10.1. Das Gram-Schmidt-Verfahren gehört zum klassischen Bestand der Analysis und wird seit Jahrzehnten als konstruktives Hilfsmittel zur Beschaffung orthogonaler Basen verwendet. Hinweise auf die unter Umständen extreme Instabilität sind bei Rice [66] zu finden, wo auch die bei gleichem Aufwand wesentlich günstigere modifizierte Version beschrieben ist. Der Nachweis der numerischen Stabilität des MGS-Verfahrens (und nicht der numerischen Gutartigkeit, wie in Lawson/Hanson [74, S. 130/132] irrtümlich berichtet wird) geht auf

Björck [67a] zurück und nutzt die spezielle Fehlerkorrelation in $\boldsymbol{Q}$ und $\boldsymbol{R}$ aus. Unter Benutzung des Stewart-Lemmas 8.2.11 ist die numerische Gutartigkeit von Kiełbasiński [unveröffentlichtes Manuskript] gezeigt worden.

B 10.2. Die Verwendung von Householder-Spiegelungen zur Erzeugung von Nullen geht auf Householder [58] zurück. Ihre Anwendung auf rechteckige Matrizen und zur Lösung von Quadratmittelproblemen wurde von Golub [65] und Businger/Golub [65] beschrieben. Die Fehleranalyse der $\boldsymbol{QR}$-Faktorisierung mittels Householder-Transformationen wurde von Wilkinson [65] durchgeführt.

B 10.3. Im Fall $m = n$ ist die orthogonale Dreiecksfaktorisierung mittels Givens-Drehungen von Givens [58] beschrieben worden. Wegen des doppelt so hohen Aufwands hat sie — abgesehen von Sonderfällen wie z. B. schwach besetztem $\boldsymbol{A}$ oder Modifikationstechniken gemäß Abschnitt 10.3 — zunächst nur wenig praktische Bedeutung erlangt. Das änderte sich mit dem Bekanntwerden der impliziten Realisierungsmöglichkeit, siehe B 3.4. Systematische Darstellungen der Nutzung von Givens-Drehungen sind bei Gentleman [73, 75], Wilkinson [77] und Voevodin [77] zu finden. Dort können auch die günstigen Kumulationskonstanten aus 10.2.8 nachgelesen werden, siehe auch Rath [82].

B 10.4. Zur Aufdatierung der Spaltennormen bei der Realisierung der Orthogonalisierungsverfahren mit Spaltenvertauschungen gemäß 10.1.6(iii) bzw. 10.2.5(vii) sei auf Björck [67] bzw. auf die LINPACK-Dokumentation von Dongarra et al. [79] verwiesen.

B 10.5. Grundlegende Ergebnisse zur iterativen Verbesserung von Quadratmittellösungen gehen auf Björck [67b, 68, 78] und Björck/Golub [67] zurück. Dabei ist wesentlich, daß sowohl $\boldsymbol{x}$ als auch $\boldsymbol{r}$ verbessert werden. Die sich auf den ersten Blick anbietende, zu 5.4 analoge Verbesserung allein von $\boldsymbol{x}$ nach der Vorschrift

$$\text{Bestimme } \boldsymbol{h} \text{ als Lösung von } \boldsymbol{Ah} \cong \boldsymbol{b} - \boldsymbol{Ax}, \text{ setze } \bar{\boldsymbol{x}} := \boldsymbol{x} + \boldsymbol{h}$$

ist nur im Fall kleiner Residuen brauchbar, siehe Golub [65] und Golub/Wilkinson [66].

B 10.6. Über Aufdatierungsalgorithmen von $\boldsymbol{QR}$-Faktorisierungen existiert eine umfangreiche Literatur. Wir zitieren die Arbeiten von Lawson/Hanson [74], Gill/Golub/Murray/Saunders [74], Gill/Murray/Saunders [75], Daniel/Gragg/Kaufman/Stewart [76] und Paige [80], wo neben den hier angegebenen Verfahren auch zahlreiche andere gefunden werden können. Für statistische Anwendungen sei auf Golub [69], Eldén [72], Golub/Styan [73] und Lawson/Hanson [74] verwiesen. Speziell wird erstmals bei Golub [69] das Streichen einer Zeile durch das in bezug auf die Quadratmittellösung äquivalente Hinzufügen des i-fachen der zu streichenden Zeile realisiert (i imaginäre Einheit). Rundungsfehleranalysen für Aufdatierungsalgorithmen sind z. B. bei Daniel et al. [76] und Paige [80] zu finden. Allerdings ist das Problem des Rundungsfehlereinflusses bei Aufdatierungsalgorithmen noch nicht völlig geklärt, und es gibt bis jetzt auch keine umfassende Darstellung des Gesamtkomplexes.

11. Rangdefiziente Quadratmittelprobleme

Im Fall einer rangdefizienten Matrix $\boldsymbol{A}$ hat das Quadratmittelproblem $\boldsymbol{Ax} \cong \boldsymbol{b}$ unendlich viele Lösungen, deren Gesamtheit eine lineare Mannigfaltigkeit bildet. Die Lösungsmenge kann durch die Normallösung $\boldsymbol{x} = \boldsymbol{A}^+\boldsymbol{b}$ — die Lösung kleinster Euklidischer Norm — und eine orthogonale Basis des Nullraumes von $\boldsymbol{A}$ charakterisiert werden, vgl. 8.1.

Im Abschnitt 8.2 haben wir gesehen, daß die Normallösung eine unstetige und unbeschränkte Funktion der Eingangsdaten $\{\boldsymbol{A}, \boldsymbol{b}\}$ ist, d. h., die Berechnung der

Normallösung ist ein inkorrekt gestelltes Problem. Ursache für dieses irreguläre Verhalten sind rangerhöhende Störungen der Matrix $\boldsymbol{A}$. Zur Überwindung der Inkorrektheit bieten sich daher zwei Möglichkeiten an:

— Ausschließung der bösartigen Störungen bei Beibehaltung der Zielfunktion $\|\boldsymbol{Ax} - \boldsymbol{b}\|$, d. h. Einschränkung des Definitionsgebietes der Aufgabe

— Änderung der Zielfunktion bei Beibehaltung des Definitionsgebietes derart, daß die für die ursprüngliche Aufgabe bösartigen Störungen für die modifizierte Aufgabe gutartig sind.

In beiden Fällen wird dabei zu einem modifizierten, korrekt gestelltem Problem übergegangen. Man spricht von *Regularisierung*, und die zugehörigen Algorithmen werden *Regularisierungsverfahren* genannt.

In diesem Kapitel werden wir beide genannten Möglichkeiten zur Regularisierung ausnutzen; die erste führt auf die *diskreten*, die zweite auf die *kontinuierlichen Regularisierungsverfahren*. In beiden Verfahrensklassen hängt der Regularisierungsprozeß und damit die regularisierte Lösung von einem Parameter ab. Die Wahl dieses Parameters kann nicht allein nach numerischen Gesichtspunkten erfolgen, sondern erfordert zusätzliche Informationen über erwünschte Eigenschaften der Lösung. Weil sich solche Eigenschaften nicht immer mathematisch präzise formulieren lassen und der Einfluß des Parameters auf die Lösung a priori nicht eingeschätzt werden kann, ist die Ausführung der Regularisierungsverfahren im Dialog mit dem sachkundigen Anwender oder Bearbeiter einer völlig automatisierten Ausführung i. allg. vorzuziehen.

Da die numerisch berechnete Singulärwertzerlegung — im folgenden mit SVD (engl. „Singular Value Decomposition") abgekürzt — die zuverlässigste Grundlage für die Lösung rangdefizienter Quadratmittelprobleme darstellt und für gewisse diskrete Regularisierungsverfahren explizit benötigt wird, gehen wir zunächst auf die Berechnung dieser Zerlegung ein.

11.1. Numerische Berechnung der Singulärwertzerlegung

Für eine gegebene Matrix $\boldsymbol{A} \in \mathbf{R}^{m,n}$, $m \geqq n$, erfolgt die Berechnung der Singulärwertzerlegung

$$\boldsymbol{A} = \boldsymbol{U}\boldsymbol{\Sigma}\boldsymbol{V}^{\mathrm{T}}, \tag{1}$$

$\boldsymbol{U} \in \mathbf{R}^{m,m}$ und $\boldsymbol{V} \in \mathbf{R}^{n,n}$ orthogonal, $\boldsymbol{\Sigma} = \operatorname{diag}(\sigma_1, \ldots, \sigma_n) \in \mathbf{R}^{m,n}$, $\sigma_1 \geqq \cdots \geqq \sigma_n \geqq 0$, in zwei Stufen:

In der ersten Stufe wird $\boldsymbol{A}$ durch endlich viele orthogonale Äquivalenztransformationen auf obere Bidiagonalform gebracht:

11.1.1. *Bidiagonalisierung einer Matrix mittels Householder-Spiegelungen.* Die Matrix $\boldsymbol{A} \in \mathbf{R}^{m,n}$ werde nach der Vorschrift

$$\boldsymbol{A}^{(1)} := \boldsymbol{A}, \qquad \boldsymbol{A}^{(k+1)} := \boldsymbol{H}_k \boldsymbol{A}^{(k)} \hat{\boldsymbol{H}}_k \qquad (k = 1, \ldots, n) \tag{2}$$

mit Householder-Spiegelungen $\boldsymbol{H}_k \in \mathbf{R}^{m,m}$, $\hat{\boldsymbol{H}}_k \in \mathbf{R}^{n,n}$ transformiert. Die Spiege-

lung $\boldsymbol{H}_k$ sei so festgelegt, daß die Elemente in der k-ten Spalte von $\bar{\boldsymbol{A}}^{(k)} := \boldsymbol{H}_k \boldsymbol{A}^{(k)}$ unterhalb des Diagonalelementes zu 0 gemacht werden, und $\hat{\boldsymbol{H}}_k$ werde so gewählt, daß die Elemente in der k-ten Zeile von $\boldsymbol{A}^{(k+1)} := \bar{\boldsymbol{A}}^{(k)} \hat{\boldsymbol{H}}_k$ rechts vom Nebendiagonalelement annulliert werden. Dann gilt in exakter Arithmetik

$$\boldsymbol{U}_1^{\mathsf{T}} \boldsymbol{A} \boldsymbol{V}_1 = \boldsymbol{A}^{(n+1)} = \begin{pmatrix} \boldsymbol{B} \\ \hline \boldsymbol{O} \end{pmatrix} \tag{3}$$

mit einer Bidiagonalmatrix

$$\boldsymbol{B} = \begin{pmatrix} \alpha_1 & \gamma_2 & & & \\ & \alpha_2 & \gamma_3 & & \\ & & \ddots & \ddots & \\ & & & \ddots & \gamma_n \\ & & & & \alpha_n \end{pmatrix} \in \mathbf{R}^{n,n} \tag{4}$$

und den orthogonalen Transformationsmatrizen

$$\begin{aligned} &\boldsymbol{U}_1^{\mathsf{T}} := \boldsymbol{H}_n \cdots \boldsymbol{H}_2 \boldsymbol{H}_1 \in \mathbf{R}^{m,m}, \qquad \boldsymbol{H}_n = \boldsymbol{I}_m \quad \text{im Fall } m = n, \\ &\boldsymbol{V}_1 := \hat{\boldsymbol{H}}_1 \hat{\boldsymbol{H}}_2 \cdots \hat{\boldsymbol{H}}_n \in \mathbf{R}^{n,n}, \qquad \hat{\boldsymbol{H}}_{n-1} = \hat{\boldsymbol{H}}_n = \boldsymbol{I}_n. \end{aligned} \tag{5}$$

Aufwand:
$\sim 2(m - n/3)\, n^2$ opms $+ \sim 2n$ opr und $\sim 2n$ S für $\boldsymbol{B}$ ohne explizite Berechnung von $\boldsymbol{U}_1$, $\boldsymbol{V}_1$
$\sim 2n^3/3$ opms und $\sim n^2$ S für explizite Berechnung von $\boldsymbol{V}_1$
$\sim 2(m^2 - mn + n^2/3)\, n$ opms und m^2 S für explizite Berechnung von $\boldsymbol{U}_1$

Für $m = 5$, $n = 4$ verlaufen die ersten Schritte der Transformation nach dem Muster

$$\begin{pmatrix} \times & \times & \times & \times \\ \times & \times & \times & \times \\ \times & \times & \times & \times \\ \times & \times & \times & \times \\ \times & \times & \times & \times \end{pmatrix} \to \begin{pmatrix} \times & \times & \times & \times \\ \boxed{0} & \times & \times & \times \\ \boxed{0} & \times & \times & \times \\ \boxed{0} & \times & \times & \times \\ \boxed{0} & \times & \times & \times \end{pmatrix} \to \begin{pmatrix} \times & \times & \boxed{0} & \boxed{0} \\ 0 & \times & \times & \times \\ 0 & \times & \times & \times \\ 0 & \times & \times & \times \\ 0 & \times & \times & \times \end{pmatrix} \to \begin{pmatrix} \times & \times & 0 & 0 \\ 0 & \times & \times & \times \\ 0 & \boxed{0} & \times & \times \\ 0 & \boxed{0} & \times & \times \\ 0 & \boxed{0} & \times & \times \end{pmatrix}$$

$\boldsymbol{A}^{(1)}$ $\qquad$ $\boldsymbol{H}_1 \boldsymbol{A}^{(1)} = \bar{\boldsymbol{A}}^{(1)}$ $\qquad$ $\bar{\boldsymbol{A}}^{(1)} \hat{\boldsymbol{H}}_1 = \boldsymbol{A}^{(2)}$ $\qquad$ $\boldsymbol{H}_2 \boldsymbol{A}^{(2)} = \bar{\boldsymbol{A}}^{(2)}$

$$\to \begin{pmatrix} \times & \times & 0 & 0 \\ 0 & \times & \times & \boxed{0} \\ 0 & 0 & \times & \times \\ 0 & 0 & \times & \times \\ 0 & 0 & \times & \times \end{pmatrix} \to \cdots.$$

$\bar{\boldsymbol{A}}^{(2)} \hat{\boldsymbol{H}}_2 = \boldsymbol{A}^{(3)}$

Die im jeweiligen Teilschritt neu entstehenden Nullen sind dabei eingerahmt worden.

11.1.2. *Bemerkung.* (i) Wenn $a_{ik}^{(k)} = 0$ $(i = k, \ldots, m)$ gilt, ist die Spiegelung $\boldsymbol{H}_k$ nicht eindeutig bestimmt. In diesem Fall soll $\boldsymbol{H}_k := \boldsymbol{I}_m$ gesetzt werden, und im Fall der Unbestimmtheit von $\hat{\boldsymbol{H}}_k$ soll analog $\hat{\boldsymbol{H}}_k := \boldsymbol{I}_n$ gesetzt werden.

(ii) Der Bidiagonalisierungsprozeß kann analog zur $\boldsymbol{QR}$-Faktorisierung in situ unter Verwendung von zusätzlichen $\sim 2n$ S realisiert werden, vgl. 10.2.3. Die Matrizen $\boldsymbol{U}_1, \boldsymbol{V}_1$ sind dann in der Produktform (5) durch die zu $\boldsymbol{H}_k$ bzw. $\hat{\boldsymbol{H}}_k$ gehörenden Vektoren $\boldsymbol{v}^k$ bzw. $\hat{\boldsymbol{v}}^k$ gegeben. Die Berechnung von $\boldsymbol{U}_1\boldsymbol{f}$, $\boldsymbol{U}_1^{\mathsf{T}}\boldsymbol{f}$ bzw. $\boldsymbol{V}_1\boldsymbol{g}$, $\boldsymbol{V}_1^{\mathsf{T}}\boldsymbol{g}$ für $\boldsymbol{f} \in \mathbf{R}^m$, $\boldsymbol{g} \in \mathbf{R}^n$ ist analog zu 10.2.6 mit $\sim (2m - n)\, n$ opms bzw. $\sim n^2$ opms möglich.

(iii) Die Aufwandsangaben für die explizite Berechnung von $\boldsymbol{U}_1$ bzw. $\boldsymbol{V}_1$ beziehen sich auf die Auswertung gemäß $\boldsymbol{U}_1 := \boldsymbol{H}_1\big(\cdots(\boldsymbol{H}_{n-1}(\boldsymbol{H}_n))\cdots\big)$ bzw. $\boldsymbol{V}_1 := \hat{\boldsymbol{H}}_1\big(\cdots(\hat{\boldsymbol{H}}_{n-2} \times (\hat{\boldsymbol{H}}_{n-1}))\cdots\big)$, vgl. Ü 10.2.2. Im Fall $m \gg n$ sollte die Berechnung von $\boldsymbol{U}_1$ wegen des Terms $2m^2n$ tunlichst vermieden werden, während die $\sim 2n^3/3$ opms für $\boldsymbol{V}_1$ kaum ins Gewicht fallen. Die Teilmatrix der ersten n Spalten von $\boldsymbol{U}_1$ kann nach der Vorschrift $\boldsymbol{H}_1\left(\cdots\left(\boldsymbol{H}_{n-1}\left(\boldsymbol{H}_n\begin{pmatrix}\boldsymbol{I}_n\\ \boldsymbol{O}\end{pmatrix}\right)\right)\cdots\right)$ mit $\sim (m - n/3)\, n^2$ opms ermittelt werden.

(iv) Für genügend großes m/n wird der Bidiagonalisierungsaufwand reduziert, wenn $\boldsymbol{A}$ zunächst auf obere Dreiecksform gebracht und die Dreiecksmatrix danach bidiagonalisiert wird, siehe Ü 11.1.1.

(v) Die Spiegelungen $\boldsymbol{H}_k$ bzw. $\hat{\boldsymbol{H}}_k$ können durch jeweils $m - k$ bzw. $n - k - 1$ explizite oder implizite Givens-Drehungen ersetzt werden. □

In der zweiten Stufe des Prozesses zur Berechnung der SVD wird die Bidiagonalmatrix $\boldsymbol{B}$ orthogonal äquivalent auf Diagonalform $\tilde{\boldsymbol{\Sigma}} = \boldsymbol{U}_2^{\mathsf{T}}\boldsymbol{B}\boldsymbol{V}_2$ transformiert, d. h., es wird die SVD von $\boldsymbol{B}$ bestimmt. Dazu bietet sich an, den Beweis zu 1.2.14 mit $\boldsymbol{B}$ statt $\boldsymbol{A}$ konstruktiv nachzuvollziehen, also die Zahlen $\lambda_i := \sigma_i^2$ als Eigenwerte der symmetrischen Tridiagonalmatrix

$$\boldsymbol{T} := \boldsymbol{B}^{\mathsf{T}}\boldsymbol{B} = \begin{pmatrix} \alpha_1^2 & \alpha_1\gamma_2 & & & & & \\ \alpha_1\gamma_2 & \alpha_2^2 + \gamma_2^2 & \alpha_2\gamma_3 & & & & \\ & \ddots & \ddots & \ddots & & & \\ & & \ddots & \ddots & \ddots & & \\ & & & \ddots & \ddots & \ddots & \\ & & & & \alpha_{n-1}^2 + \gamma_{n-1}^2 & \alpha_{n-1}\gamma_n \\ & & & & \alpha_{n-1}\gamma_n & \alpha_n^2 + \gamma_n^2 \end{pmatrix} \tag{6}$$

zu bestimmen und $\boldsymbol{U}_2, \boldsymbol{V}_2$ mittels der zugehörigen Eigenvektoren festzulegen. Als einen für diese Eigenwertaufgabe besonders geeigneten Algorithmus werden wir in 13.7 den sogenannten $\boldsymbol{QR}$-Algorithmus mit Verschiebungen kennenlernen. Bei Berechnung der λ_i nach diesem oder einem anderen Verfahren ist mindestens mit einem Fehler

$$|\delta\lambda_i| \leqq \nu\, \|\boldsymbol{T}\|_2 = \nu\, \|\boldsymbol{B}\|_2^2 = \nu\, \|\boldsymbol{A}\|_2^2 = \nu\sigma_1^2 \tag{7}$$

in der Größenordnung des optimalen Fehlerniveaus von λ_i bezüglich $\boldsymbol{T}$ zu rechnen,

siehe 13.1.2. Für $\sigma_i := \sqrt{\lambda_i}$ ergibt sich damit mindestens ein Fehler

$$|\delta\sigma_i| = \left|\sqrt{\lambda_i + \delta\lambda_i} - \sqrt{\lambda_i}\right| = \frac{|\delta\lambda_i|}{\sqrt{\lambda_i + \delta\lambda_i} + \sqrt{\lambda_i}} \approx \frac{|\delta\lambda_i|}{2\sigma_i} \leqq \nu \frac{\sigma_1^2}{2\sigma_i}. \tag{8}$$

Andererseits ist das optimale Fehlerniveau von σ_i bezüglich $\boldsymbol{B}$ nach 8.2.3 durch

$$|\delta\sigma_i| \leqq \nu \, \|\boldsymbol{B}\|_2 = \nu \, \|\boldsymbol{A}\|_2 = \nu\sigma_1 \tag{9}$$

charakterisiert, also um den Faktor $0{,}5\sigma_1/\sigma_i$ kleiner als beim Umweg über $\boldsymbol{T}$. Die kleineren Singulärwerte werden daher nicht mit ausreichender Genauigkeit berechnet; bezüglich dieser Werte liegt keine numerische Stabilität vor. Ein ähnlicher negativer Effekt ist uns bereits beim Übergang vom Quadratmittelproblem $\boldsymbol{Ax} \cong \boldsymbol{b}$ zu den theoretisch äquivalenten Normalgleichungen $\boldsymbol{A}^\mathsf{T}\boldsymbol{Ax} = \boldsymbol{A}^\mathsf{T}\boldsymbol{b}$ begegnet, vgl. 9.1.

Im folgenden soll versucht werden, den ungünstigen Einfluß des expliziten Übergangs zu $\boldsymbol{B}^\mathsf{T}\boldsymbol{B}$ zu vermeiden und $\boldsymbol{B}$ direkt durch eine theoretisch unendliche Folge

$$\boldsymbol{B}^{(1)} := \boldsymbol{B}, \qquad \boldsymbol{B}^{(k+1)} := \boldsymbol{G}_k \boldsymbol{B}^{(k)} \hat{\boldsymbol{G}}_k^\mathsf{T} \qquad (k = 1, 2, \ldots) \tag{10}$$

von orthogonalen Äquivalenztransformationen auf Diagonalform $\tilde{\boldsymbol{B}}$ zu transformieren. Dabei werden $\boldsymbol{G}_k$, $\hat{\boldsymbol{G}}_k$ als Produkte

$$\boldsymbol{G}_k := \boldsymbol{G}_{n-1,n} \cdots \boldsymbol{G}_{23}\boldsymbol{G}_{12}, \qquad \hat{\boldsymbol{G}}_k^\mathsf{T} := \hat{\boldsymbol{G}}_{12}^\mathsf{T}\hat{\boldsymbol{G}}_{23}^\mathsf{T} \cdots \hat{\boldsymbol{G}}_{n-1,n}^\mathsf{T} \tag{11}$$

von Givens-Drehungen so festgelegt, daß die Matrizen $\boldsymbol{B}^{(k)}$ wie $\boldsymbol{B}$ von Bidiagonalform

$$\boldsymbol{B}^{(k)} = \begin{pmatrix} \alpha_1^{(k)} & \gamma_2^{(k)} & & & \\ & \alpha_2^{(k)} & \ddots & & \\ & & \ddots & \ddots & \\ & & & \ddots & \gamma_n^{(k)} \\ & & & & \alpha_n^{(k)} \end{pmatrix} \tag{12}$$

sind und $\boldsymbol{T}^{(k)} := \boldsymbol{B}^{(k)\mathsf{T}}\boldsymbol{B}^{(k)}$ in exakter Arithmetik gerade die Matrizen wären, die der erwähnte $\boldsymbol{QR}$-Algorithmus aus $\boldsymbol{T} = \boldsymbol{T}^{(1)}$ erzeugt. Man beachte, daß die $\boldsymbol{T}^{(k)}$ dieselbe Struktur wie $\boldsymbol{T}$ haben. Der Einfachheit halber lassen wir bei den Drehungen und den Elementen von $\boldsymbol{B}^{(k)}$ bzw. $\boldsymbol{T}^{(k)}$ den oberen Iterationsindex k wie in (11) weg. Wir setzen ferner voraus, daß alle α_i und γ_i von 0 verschieden sind. Dies ist keine Einschränkung, denn andernfalls zerfällt $\boldsymbol{B}$ —gegebenenfalls nach einer Zusatztransformation — in Teilmatrizen niedrigerer Dimension, siehe Ü 11.1.3.

Die genannten Forderungen — Bidiagonalität von $\boldsymbol{B}^{(k)}$ und Äquivalenz mit dem $\boldsymbol{QR}$-Algorithmus — lassen sich wie folgt erfüllen: Zunächst wird wie in 13.7 der Verschiebungsparameter $\varrho = \varrho_k$ als derjenige Eigenwert der (2,2)-Teilmatrix

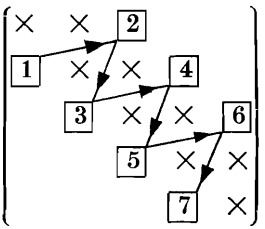
$$\begin{pmatrix} \alpha_{n-1}^2 + \gamma_{n-1}^2 & \alpha_{n-1}\gamma_n \\ \alpha_{n-1}\gamma_n & \alpha_n^2 + \gamma_n^2 \end{pmatrix}$$

in der rechten unteren Ecke von $\boldsymbol{T}^{(k)}$ gewählt, der dem Diagonalelement $\alpha_n^2 + \gamma_n^2$ am nächsten liegt. Danach wird die erste Drehung $\hat{\boldsymbol{G}}_{12}$ so bestimmt, daß das zweite

Element $\alpha_1\gamma_2$ der ersten Spalte $(\alpha_1^2 - \varrho, \alpha_1\gamma_2, 0, \ldots, 0)^\mathsf{T}$ von $\boldsymbol{T}^{(k)} - \varrho\boldsymbol{I} = \boldsymbol{B}^\mathsf{T}\boldsymbol{B} - \varrho\boldsymbol{I}$ bei der Transformation in $\hat{\boldsymbol{G}}_{12}(\boldsymbol{B}^\mathsf{T}\boldsymbol{B} - \varrho\boldsymbol{I})$ zu 0 gemacht wird. Mit dem so festgelegten $\hat{\boldsymbol{G}}_{12}$ wird $\boldsymbol{B}^{(k)}\hat{\boldsymbol{G}}_{12}^\mathsf{T}$ gebildet, wobei allerdings die Bidiagonalform in der Position $\{2, 1\}$ zerstört wird. Alle weiteren Drehungen werden dann in der durch

$$\boldsymbol{B}^{(k+1)} := \boldsymbol{G}_{n-1,n}\{\cdots \boldsymbol{G}_{23}\{[\boldsymbol{G}_{12}(\boldsymbol{B}^{(k)}\hat{\boldsymbol{G}}_{12}^\mathsf{T})]\,\hat{\boldsymbol{G}}_{23}^\mathsf{T}\}\cdots \hat{\boldsymbol{G}}_{n-1,n}^\mathsf{T}\} \tag{13}$$

festgelegten Reihenfolge lediglich dazu benutzt, die Bidiagonalform sukzessive wieder herzustellen: Durch Multiplikation von links mit $\boldsymbol{G}_{12}$ wird in $\{2, 1\}$ wieder eine Null erzeugt, allerdings entsteht dabei in $\{1, 3\}$ ein neues Nichtnullelement. Dieses wird durch Rechtsmultiplikation mit $\hat{\boldsymbol{G}}_{23}^\mathsf{T}$ annulliert, wobei in $\{3, 2\}$ ein von 0 verschiedenes Element entsteht usw. Die Reihenfolge, in der Nichtnullelemente entstehen und im nächsten Teilschritt wieder annulliert werden, ist für $n = 5$ durch das folgende Schema gegeben:



Wegen der schnellen, praktisch meist kubischen oder besseren Konvergenz des $\boldsymbol{QR}$-Algorithmus konvergieren die Produkte $\alpha_l^{(k)}\gamma_{l+1}^{(k)}$ — das wären die Nebendiagonalelemente von $\boldsymbol{T}^{(k)}$ — sehr schnell gegen 0. Daher muß für nicht zu großes k eines der Elemente $\alpha_l^{(k)}$ bzw. $\gamma_{l+1}^{(k)}$ im Sinne von

$$|\alpha_l^{(k)}| \leqq \nu\,\|\boldsymbol{B}\| \quad \text{bzw.} \quad |\gamma_{l+1}^{(k)}| \leqq \nu\,\|\boldsymbol{B}\| \tag{14}$$

genügend klein sein und kann im Rahmen der Computergenauigkeit als 0 angesehen werden. Dies entspricht einer Störung $\delta\boldsymbol{B}$ der Matrix $\boldsymbol{B}$ mit $\|\delta\boldsymbol{B}\| \leqq \nu\,\|\boldsymbol{B}\|$, beeinträchtigt also die Genauigkeit praktisch nicht. Beim Nullsetzen von $\gamma_{l+1}^{(k)}$ zerfällt die Matrix $\boldsymbol{B}^{(k)}$ direkt, beim Nullsetzen von $\alpha_l^{(k)}$ nach einer Zusatztransformation in Blöcke niedrigerer Dimension, siehe wieder Ü 11.1.3. Das Verfahren wird dann mit jedem der Blöcke einzeln und damit billiger fortgesetzt. Meist tritt (14) für $l = n - 1$ ein, siehe 13.7. Falls $|\gamma_n^{(k)}|$ klein ist und 0 gesetzt wird, zerfällt $\boldsymbol{B}^{(k)}$ in den $(1,1)$-Block $\alpha_n^{(k)}$, der als bereits feststehendes Diagonalelement angesehen wird, und die durch Streichen der n-ten Zeile und Spalte entstehende $(n - 1)$-dimensionale Restmatrix; man spricht von *Deflation*. Wenn $\alpha_{n-1}^{(k)}$ gleich 0 gesetzt werden kann, ist eine analoge Deflation nach Ausführung der in Ü 11.1.3 beschriebenen Zusatztransformationen sogar zweimal möglich, wobei einmal eine 0 abgespalten wird.

Alle diese Maßnahmen haben zur Folge, daß sich der theoretisch unendliche Diagonalisierungsprozeß (10) praktisch wie ein finiter Prozeß verhält: Nach k Schritten kann

$$\tilde{\boldsymbol{B}} := \boldsymbol{B}^{(k+1)} = \operatorname{diag}(\alpha_i^{(k+1)}) = \boldsymbol{G}_k \cdots \boldsymbol{G}_1\boldsymbol{B}\hat{\boldsymbol{G}}_1^\mathsf{T} \cdots \hat{\boldsymbol{G}}_k^\mathsf{T} \tag{15}$$

im Rahmen der Computergenauigkeit als Diagonalmatrix angesehen werden. Der

Abbruchindex k ist dabei meist nicht viel größer als n, d. h., pro Diagonalelement ist im Mittel nur eine Schleife (10) erforderlich. In den seltensten Fällen überschreitet k den Wert $2n$.

Da die Diagonalelemente von $\tilde{\boldsymbol{B}}$ i. allg. weder nichtnegativ noch der Größe nach geordnet sind, setzen wir abschließend

$$\tilde{\boldsymbol{\Sigma}} = \operatorname{diag}(\sigma_1, \ldots, \sigma_n) := \boldsymbol{P}(\boldsymbol{D}\tilde{\boldsymbol{B}})\,\boldsymbol{P}^{\mathsf{T}} \tag{16}$$

mit

$$\boldsymbol{D} = \operatorname{diag}(d_i), \qquad d_i := \begin{cases} 1 & \text{für } \alpha_i^{(k\cdot 1)} \geqq 0, \\ -1 & \text{für } \alpha_i^{(k+1)} < 0 \end{cases}$$

und einer Permutationsmatrix $\boldsymbol{P}$, die $\sigma_1 \geqq \cdots \geqq \sigma_n$ sichert.

Die gesamte Transformation von $\boldsymbol{B}$ in $\tilde{\boldsymbol{\Sigma}}$ läßt sich dann durch

$$\tilde{\boldsymbol{\Sigma}} = \boldsymbol{P}\boldsymbol{D}\boldsymbol{G}_k \cdots \boldsymbol{G}_1 \boldsymbol{B} \hat{\boldsymbol{G}}_1^{\mathsf{T}} \cdots \hat{\boldsymbol{G}}_k^{\mathsf{T}} \boldsymbol{P}^{\mathsf{T}} = \boldsymbol{U}_2^{\mathsf{T}} \boldsymbol{B} \boldsymbol{V}_2 \tag{17}$$

mit

$$\boldsymbol{U}_2^{\mathsf{T}} := \boldsymbol{P}\boldsymbol{D}\boldsymbol{G}_k \cdots \boldsymbol{G}_1, \qquad \boldsymbol{V}_2^{\mathsf{T}} := \boldsymbol{P}\hat{\boldsymbol{G}}_k \cdots \hat{\boldsymbol{G}}_1 \tag{18}$$

beschreiben. Aus (3) und (17) ergibt sich schließlich

$$\boldsymbol{A} = \boldsymbol{U}_1 \begin{pmatrix} \boldsymbol{B} \\ \hline \boldsymbol{O} \end{pmatrix} \boldsymbol{V}_1^{\mathsf{T}} = \boldsymbol{U}_1 \begin{pmatrix} \boldsymbol{U}_2 \tilde{\boldsymbol{\Sigma}} \boldsymbol{V}_2^{\mathsf{T}} \\ \hline \boldsymbol{O} \end{pmatrix} \boldsymbol{V}_1^{\mathsf{T}} = \boldsymbol{U}_1 \left(\begin{array}{c|c} \boldsymbol{U}_2 & \boldsymbol{O} \\ \hline \boldsymbol{O} & \boldsymbol{I} \end{array}\right) \begin{pmatrix} \tilde{\boldsymbol{\Sigma}} \\ \hline \boldsymbol{O} \end{pmatrix} \boldsymbol{V}_2^{\mathsf{T}} \boldsymbol{V}_1^{\mathsf{T}} \tag{19}$$

als die gesuchte SVD von $\boldsymbol{A}$. Zusammenfassend erhalten wir das folgende Resultat.

11.1.3. *Berechnung der SVD.* Für $\boldsymbol{A} \in \mathbf{R}^{m,n}$ werde der nachfolgende zweistufige Algorithmus ausgeführt:

S1: Transformiere $\boldsymbol{A}$ nach dem Verfahren 11.1.1 gemäß

$$\boldsymbol{U}_1^{\mathsf{T}} \boldsymbol{A} \boldsymbol{V}_1 = \begin{pmatrix} \boldsymbol{B} \\ \hline \boldsymbol{O} \end{pmatrix}$$

auf Bidiagonalform.

S2: Transformiere $\boldsymbol{B}$ nach den Formeln (10) bis (18) gemäß

$$\boldsymbol{U}_2^{\mathsf{T}} \boldsymbol{B} \boldsymbol{V}_2 = \tilde{\boldsymbol{\Sigma}}$$

auf Diagonalform $\tilde{\boldsymbol{\Sigma}} = \operatorname{diag}(\sigma_j)$, $\sigma_1 \geqq \cdots \geqq \sigma_n \geqq 0$.

Dann ist die Singulärwertzerlegung $\boldsymbol{A} = \boldsymbol{U}\boldsymbol{\Sigma}\boldsymbol{V}^{\mathsf{T}}$ durch die Faktoren

$$\boldsymbol{U} := \boldsymbol{U}_1 \left(\begin{array}{c|c} \boldsymbol{U}_2 & \boldsymbol{O} \\ \hline \boldsymbol{O} & \boldsymbol{I}_{m-n} \end{array}\right), \quad \boldsymbol{\Sigma} := \begin{pmatrix} \tilde{\boldsymbol{\Sigma}} \\ \hline \boldsymbol{O} \end{pmatrix}, \quad \boldsymbol{V} := \boldsymbol{V}_1 \boldsymbol{V}_2 \tag{20}$$

gegeben.

Aufwand: $\sim [2(m - n/3)\, n^2$ opms $+ 2n^2$ opr] und $\sim 2n$ S für $\boldsymbol{\Sigma}$ ohne Berechnung von $\boldsymbol{U}$ und $\boldsymbol{V}$; zur Darstellung von $\boldsymbol{U}, \boldsymbol{V}$ siehe Bemerkung 11.1.6.

11.1.4. *Bemerkung.* (i) Die Berechnung von $\boldsymbol{\Sigma}$ ist mit zusätzlichen $\sim 2n$ S in situ möglich, wobei $\boldsymbol{U}_1, \boldsymbol{V}_1$ ohne Mehraufwand in der Produktform (5) anfallen, siehe

11.1.2(ii). Wir setzen generell voraus, daß in der Stufe S 2 höchstens $2n$ $\boldsymbol{QR}$-Schritte erforderlich sind, und zwar zwei pro Singulärwert. Wegen der Dimensionsreduzierung mittels Deflation sind dann insgesamt $\sim 2[2(n-1)+2(n-2)+\cdots+2] \sim 2n^2$ Givens-Drehungen auszuführen; de facto sind es meist weniger. Die Stufe S 2 erfordert daher einen Aufwand von höchstens $\sim 2n^2$ opr $+ O(n^2)$ übrigen Operationen, die gegenüber den $\sim 2(m - n/3)\, n^2$ opms der Bidiagonalisierungsstufe S1 vernachlässigt werden können, d. h., die Berechnung der Singulärwerte aus der Bidiagonalform kostet praktisch nichts.

(ii) Die Kleinheitsbedingungen (14) werden meist durch raffiniertere Bedingungen wie

$$|\alpha_l^{(k)}| \leqq \nu(|\gamma_l^{(k)}| + |\gamma_{l+1}^{(k)}|) \quad \text{bzw.} \quad |\gamma_{l+1}^{(k)}| \leqq \nu(|\alpha_l^{(k)}| + |\alpha_{l+1}^{(k)}|) \tag{21}$$

ersetzt, die bei günstig strukturiertem $\boldsymbol{B}$ die Berechnung der kleinen Singulärwerte mit höherer Genauigkeit erlauben als bei Verwendung von (14). □

Die Art der Darstellung der Matrizen $\boldsymbol{U}, \boldsymbol{V}$ hängt von der beabsichtigten Nutzung der SVD ab. Wir erinnern daran, daß die Normallösung des Quadratmittelproblems $\boldsymbol{Ax} \cong \boldsymbol{b}$ mittels der SVD zumindest in der Theorie wie folgt berechnet werden kann, vgl. 8.1.D:

11.1.5. *Basisalgorithmus zur Lösung von* $\boldsymbol{Ax} \cong \boldsymbol{b}$ *mittels SVD*

S1 (Faktorisierung der Matrix $\boldsymbol{A}$ und Rangbestimmung)

S1.1: Berechne Singulärwertzerlegung $\boldsymbol{A} = \boldsymbol{U\Sigma V}^T$

S1.2: Bestimme $r = \operatorname{rang}(\boldsymbol{A})$ aus der Bedingung

$$\sigma_1 \geqq \cdots \geqq \sigma_r > 0 = \sigma_{r+1} = \cdots = \sigma_n \tag{22}$$

S2 (Berechnung der zur rechten Seite $\boldsymbol{b}$ gehörenden Normallösung $\boldsymbol{x} = \boldsymbol{A}^+\boldsymbol{b} = \boldsymbol{V\Sigma}^+\boldsymbol{U}^T\boldsymbol{b}$, des Residuums $\boldsymbol{r} = \boldsymbol{b} - \boldsymbol{Ax} = \boldsymbol{U}(\boldsymbol{I} - \boldsymbol{\Sigma\Sigma}^+)\,\boldsymbol{U}^T\boldsymbol{b}$ sowie dessen Norm $n\boldsymbol{r} = \|\boldsymbol{r}\|$):

S2.1: Berechne $\boldsymbol{\beta} := \boldsymbol{U}^T\boldsymbol{b}$

S2.2: Setze $\boldsymbol{\varrho} := (\boldsymbol{I} - \boldsymbol{\Sigma\Sigma}^+)\,\boldsymbol{\beta} = (0, \ldots, 0, \beta_{r+1}, \ldots, \beta_m)^T$,
berechne $n\boldsymbol{r} := \|\boldsymbol{\varrho}\|$ und $\boldsymbol{r} := \boldsymbol{U\varrho}$.

S2.3: Setze $\boldsymbol{\xi} := \boldsymbol{\Sigma}^+\boldsymbol{\beta} = (\beta_1/\sigma_1, \ldots, \beta_r/\sigma_r, 0, \ldots, 0)^T$,
berechne $\boldsymbol{x} := \boldsymbol{V\xi}$

11.1.6. *Bemerkung.* (i) Zur Berechnung der Normallösung $\boldsymbol{x}$ werden die ersten r Spalten von $\boldsymbol{V}$ benötigt, und die Spalten $r+1$ bis n bilden eine orthonormale Basis des Nullraumes $\mathcal{N}(\boldsymbol{A})$, der gemäß $\mathcal{L} = \mathcal{L}(\boldsymbol{A}, \boldsymbol{b}) = \boldsymbol{x} + \mathcal{N}(\boldsymbol{A})$ in die Lösungsmenge $\mathcal{L}$ von $\boldsymbol{Ax} \cong \boldsymbol{b}$ eingeht. Es ist daher zweckmäßig, die Matrix $\boldsymbol{V}$ explizit zu berechnen: Zunächst wird $\boldsymbol{V}_1$ wie in 11.1.2 gebildet und danach durch Rechtsmultiplikation mit den Drehungen $\hat{\boldsymbol{G}}_{j,j+1}^T$ aufdatiert. Dies erfordert insgesamt $\sim n^3(2/3$ opms $+ 4$ opm $+ 2$ ops), vgl. 11.1.4, und die Berechnung von $\boldsymbol{V\xi}$ kostet dann $\sim n^2$ opms. Die alternative implizite Produktdarstellung

$$\boldsymbol{V} = \hat{\boldsymbol{H}}_1 \cdots \hat{\boldsymbol{H}}_{n-2}\boldsymbol{V}_2 \tag{23}$$

mit explizit berechnetem V_2 erfordert $\sim n^3$(4 opm + 2 ops), ist also etwas billiger. Sie ist jedoch unhandlicher, und die Berechnung von $V\xi = \hat{H}_1(\cdots \hat{H}_{n-2}(V_2\xi)\cdots)$ entsprechend 10.2.6 kostet $\sim 2n^2$ opms.

(ii) Die zu (i) analoge explizite Berechnung von U benötigt $\sim[2(m^2 - mn + n^2/3)n$ opms + mn^2(4 opm + 2 ops)] und m^2 S, scheidet also für $m \gg n$ aus Aufwandsgründen aus. Da zur Festlegung von x jedoch nur die Teilmatrix $U_{1,r} = (u^1, \ldots, u^r)$ der ersten r Spalten von U gebraucht wird, genügt es, $U_{1,n}$ gemäß

$$U_{1,n} := H_1\left(\cdots\left(H_{n-1}\left(H_n\begin{pmatrix}U_2\\ O\end{pmatrix}\right)\right)\cdots\right) \tag{24}$$

mit $\sim[(2m - n)\, n^2$ opms + n^3(4 opm + 2 ops)] und mn S explizit zu berechnen. Das Residuum kann dann gemäß $r := b - U_{1,r}(U_{1,r}^{\mathsf{T}} b)$ bestimmt werden, vgl. (10.1.7).

(iii) Billiger als die explizite Berechnung von $U_{1,n}$ ist die Verwendung der impliziten Produktform

$$U = H_1 H_2 \cdots H_n \left(\begin{array}{c|c} U_2 & O \\ \hline O & I_{m-n}\end{array}\right) \tag{25}$$

von U mit explizit berechnetem U_2; der Aufwand ist $\sim n^3$(4 opm + 2 ops) + n^2 S. Analog zu 10.2.6 kann dann $U^{\mathsf{T}}b$ bzw. Uy mit $\sim 2mn$ opms berechnet werden.

(iv) Wenn die rechte Seite b a priori bekannt ist und U später nicht mehr benötigt wird, kann $\beta = U^{\mathsf{T}}b$ parallel zu $\Sigma = U^{\mathsf{T}}AV$ bestimmt werden, indem alle von links auf A angewendeten Transformationen sofort auch auf b angewendet werden. Damit wird die explizite Berechnung von U, $U_{1,n}$ oder U_2 vermieden, und der Aufwand zur Bildung von $U^{\mathsf{T}}b$ reduziert sich auf $\sim[(2m - n)\, n$ opms + n^2(4 opm + 2 ops)]. Das Residuum muß dann gemäß $r := b - Ax$ aus den Originaldaten berechnet werden.

(v) Für die verschiedenen Darstellungen von U sind in der nachfolgenden Tabelle die Aufwandszahlen in opms für die Faktorisierung $A = U\Sigma V^{\mathsf{T}}$ und die Berechnung von x und r für eine rechte Seite b zusammengestellt worden. Dabei ist der durch die Givens-Drehungen verursachte Term (4 opm + 2 ops) generell durch 3 opms ersetzt worden, vgl. Ü 3.3.6. Zum Vergleich wurden die entsprechenden Angaben für die QR-Faktorisierung nach Householder gemäß 10.2.4 und 10.2.9 mit angeführt.

Variante	Aufwand für $\{U, \Sigma, V\}$ bzw. $\{\hat{Q}, \hat{R}\}$	x	r
V, U explizit nach (i), (ii)	$4m^2n + 3mn^2 + 11n^3/3$	$m^2 + n^2$	m^2
V, $U_{1,n}$ explizit nach (i), (ii)	$4mn^2 + 5n^2$	$mn + n^2$	mn
V explizit, U Produktform nach (iii)	$2mn^2 + 6n^3$	$2mn + n^2$	$2mn$
V explizit, $U^{\mathsf{T}}b$ direkt nach (iv)	$2mn^2 + 3n^3$	$2mn + 3n^2$	mn
QR-Faktorisierung $A = \hat{Q}\hat{R}$	$mn^2 - n^3/3$	$2mn - n^2/2$	$2mn - n^2$

Der Faktorisierungsaufwand für die zu empfehlenden Varianten (iii), (iv) ist im Fall $m \gg n$ also etwa doppelt so groß wie für die Householder-Faktorisierung, vgl. jedoch Ü 11.1.1. □

Wir geben abschließend ohne Beweis die Ergebnisse der Rundungsfehleranalyse an.

11.1.7. *Rundungsfehleranalyse.* Für $A \in \mathbf{R}^{m,n}$ werde Algorithmus 11.1.3 ausgeführt, und es sei k die Anzahl der in S2 benötigten QR-Schritte; i. allg. ist $k \approx n$. Dann gibt es exakt orthogonale Matrizen $U \in \mathbf{R}^{m,m}$, $V \in \mathbf{R}^{n,n}$, so daß die berechnete Diagonalmatrix $\Sigma \in \mathbf{R}^{m,n}$ der Beziehung

$$A + \delta A = U\Sigma V^T \quad \text{mit} \quad \|\delta A\|_F \leqq \nu F \|A\|_F, \qquad F \sim 8mn + 12kn, \tag{26}$$

genügt. Die exakten Orthogonalfaktoren U, V werden durch die gemäß 11.1.6 explizit berechneten Matrizen $\tilde{U}, \tilde{V}$ im folgenden Sinne gut dargestellt: Für alle $\xi \in \mathbf{R}^n$, $b \in \mathbf{R}^m$ gilt für die mittels $\tilde{U}, \tilde{V}$ berechneten Vektoren $x := \mathrm{fl}(\tilde{V}\xi)$, $\beta := \mathrm{fl}(\tilde{U}^T b)$

$$\beta = U^T(b + \delta b) \text{ mit } \|\delta b\| \leqq \nu m(4n + 6k)\|b\|$$

und

$$x = V(\xi + \delta\xi) \quad \text{mit} \quad \|\delta\xi\| \leqq \nu n(4n + 6k)\|\xi\|. \tag{27}$$

Dasselbe trifft zu, wenn β, x mittels der impliziten Produktdarstellungen (23), (25) berechnet oder wenn U, V durch U^T, V^T ersetzt wird.

Dies besagt: *Die numerische Berechnung der Diagonalmatrix Σ und die Darstellungen der Faktoren U, V in der in* 11.1.6 *beschriebenen Form sind numerisch gutartige Prozesse mit akzeptablen Kumulationskonstanten.*

Übungsaufgaben

Ü 11.1.1. Man zeige: Im Fall $m > 5n/3$ ist die Bidiagonalisierung von $A \in \mathbf{R}^{m,n}$ nach der folgenden Modifikation von 11.1.1

S1.1: Transformiere A gemäß Algorithmus 10.2.4 nach der Vorschrift

$$Q^T A = \left(\frac{R}{O}\right) \tag{28}$$

mittels Householder-Spiegelungen $Q^T := H_n \cdots H_1$ auf obere Dreiecksform.

S1.2: Transformiere $R \in \mathbf{R}^{n,n}$ gemäß 11.1.1 nach der Vorschrift

$$\bar{U}_1^T R V_1 = B \tag{29}$$

mittels Householder-Spiegelungen $\bar{U}_1^T := \bar{H}_{n-1} \cdots \bar{H}_2$, $V_1 := \hat{H}_1 \cdots \hat{H}_{n-2}$ auf Bidiagonalform B

billiger als die direkte Anwendung von 11.1.1 auf A. Dabei gilt

$$U_1^T A V_1 = \left(\frac{B}{O}\right) \quad \text{mit} \quad U_1 := Q \left(\begin{array}{c|c} \bar{U}_1 & O \\ \hline O & I_{m-n} \end{array}\right), \tag{30}$$

und der Aufwand ist $\sim [(m + n)\, n^2$ opms $+ 3n$ opr].

Ü 11.1.2. Man zeige: Wenn $\boldsymbol{B}$ die gemäß (3), (4) zu $\boldsymbol{A}$ gehörende Bidiagonalmatrix bezeichnet, gilt

$$\|\boldsymbol{A}\|_2^2 \leqq [\max\{|\alpha_i| + |\gamma_i| : i = 1, \ldots, n\}]\,[\max\{|\alpha_i| + |\gamma_{i+1}| : i = 1, \ldots, n\}],$$

wobei $\gamma_1 = \gamma_n = 0$ zu setzen ist.
Hinweis: Man beachte $\|\boldsymbol{A}\|_2 = \|\boldsymbol{B}\|_2$ und Ü 1.2.9.

Ü 11.1.3. Es sei $\boldsymbol{B}$ eine Bidiagonalmatrix gemäß (4). Man überlege sich:
(i) Ist $\gamma_{l+1} = 0$ für ein $l \in \{1, \ldots, n-1\}$, so zerfällt $\boldsymbol{B}$ gemäß

$$\boldsymbol{B} = \left(\begin{array}{ccc|c} \alpha_1 & \gamma_2 & & \\ & \ddots & \ddots & \\ & & \alpha_{l-1} & \gamma_l \\ & & & \alpha_l \\ \hline & & & \end{array}\right.\left.\begin{array}{c} \\ \\ \\ 0 \\ \hline \end{array}\right.$$

$$\left.\begin{array}{ccc} & & \\ & & \\ & & \\ & & \\ \hline \alpha_{l+1} & \gamma_{l+2} & \\ & \ddots & \ddots \\ & & \gamma_n \\ & & \alpha_n \end{array}\right)$$

in zwei Bidiagonalblöcke der Dimension l bzw. $n - l$.
(ii) Ist $\alpha_l = 0$ für ein $l \in \{1, \ldots, n-1\}$, so existieren Drehungen $\boldsymbol{G}_{l,l+1}, \ldots, \boldsymbol{G}_{ln}$, so daß

$$\boldsymbol{B} = \left(\begin{array}{ccc|c} \alpha_1 & \gamma_2 & & \\ & \ddots & \ddots & \\ & & \alpha_{l-1} & \gamma_l \\ & & 0 & \\ \hline & & & \gamma_{l+1} \end{array}\right.\left.\begin{array}{c} \\ \\ \\ \\ \hline \end{array}\right.$$

$$\left.\begin{array}{ccc} & & \\ & & \\ & & \\ & & \\ \hline \alpha_{l+1} & \gamma_{l+2} & \\ & \ddots & \ddots \\ & & \gamma_n \\ & & \alpha_n \end{array}\right) \quad \text{in} \quad \overline{\boldsymbol{B}} = \left(\begin{array}{ccc|c} \alpha_1 & \gamma_2 & & \\ & \ddots & \ddots & \\ & & \alpha_{l+1} & \gamma_l \\ & & 0 & \\ \hline & & & 0 \end{array}\right.\left.\begin{array}{c} \\ \\ \\ \\ \hline \end{array}\right.$$

$$\left.\begin{array}{ccc} & & \\ & & \\ & & \\ & & \\ \hline \overline{\alpha}_{l+1} & \overline{\gamma}_{l+2} & \\ & \ddots & \ddots \\ & & \overline{\gamma}_n \\ & & \overline{\alpha}_n \end{array}\right) \tag{31}$$

übergeht mit $\overline{\boldsymbol{B}} = \boldsymbol{G}_{ln} \cdots \boldsymbol{G}_{l,l+1}\boldsymbol{B}$.
(iii) Ist $\alpha_n = 0$, so existieren Drehungen $\hat{\boldsymbol{G}}_{n-1,n}, \ldots, \hat{\boldsymbol{G}}_{1n}$, so daß

$$\boldsymbol{B} = \left(\begin{array}{ccc|c} \alpha_1 & \gamma_2 & & \\ & \ddots & \ddots & \\ & & \gamma_{n-1} & \\ & & \alpha_{n-1} & \gamma_n \\ \hline & & & 0 \end{array}\right.\left.\begin{array}{c} \\ \\ \\ \\ \hline \end{array}\right) \quad \text{in} \quad \overline{\boldsymbol{B}} = \left(\begin{array}{ccc|c} \overline{\alpha}_1 & \overline{\gamma}_2 & & \\ & \ddots & \ddots & \\ & & \overline{\gamma}_{n-1} & \\ & & \overline{\alpha}_{n-1} & \\ \hline & & & \end{array}\right.\left.\begin{array}{c} \\ \\ \\ \\ \hline \end{array}\right) \tag{32}$$

übergeht mit $\overline{\boldsymbol{B}} = \boldsymbol{B}\hat{\boldsymbol{G}}_{n-1,n}^{\mathsf{T}} \cdots \hat{\boldsymbol{G}}_{1n}^{\mathsf{T}}$.

11.2. Diskrete Regularisierung

A. *Diskrete Regularisierung mittels SVD*

Wir betrachten das Quadratmittelproblem $\boldsymbol{A}\boldsymbol{x} \cong \boldsymbol{b}$ mit $\boldsymbol{A} \in \mathfrak{R}^{m,n}$, $\boldsymbol{b} \in \mathfrak{R}^m$. Die Computermatrix $\boldsymbol{A}$ repräsentiere die rangdefiziente exakte Matrix $\boldsymbol{A}^* \in \mathbf{R}^{m,n}$, die den unbekannten Rang $r = \operatorname{rang}(\boldsymbol{A}^*) < n$ und die unbekannten Singulärwerte

$$\sigma_1^* \geqq \sigma_2^* \geqq \cdots \geqq \sigma_r^* > 0 = \sigma_{r+1}^* = \cdots = \sigma_n^* \tag{1}$$

besitzen möge. Dabei wird zugelassen, daß sich $\boldsymbol{A}$ nicht nur durch den Darstellungsfehler $\delta\boldsymbol{A}_D$, sondern auch durch einen Meßfehler $\delta\boldsymbol{A}_M$ von $\boldsymbol{A}^*$ unterscheidet: Es gelte

$$\boldsymbol{A} = \boldsymbol{A}^* + \delta\boldsymbol{A}_D + \delta\boldsymbol{A}_M \quad \text{mit} \quad \|\delta\boldsymbol{A}_D\|_F \leqq \nu \|\boldsymbol{A}^*\|_F =: \Delta\boldsymbol{A}_D,$$
$$\|\delta\boldsymbol{A}_M\|_F \leqq \Delta\boldsymbol{A}_M. \tag{2}$$

Bei der Bestimmung der Normallösung nach dem Basisalgorithmus 11.1.5 wird zunächst die SVD von $\boldsymbol{A}$ gemäß 11.1.3 berechnet. Dann gilt für die berechneten Faktoren $\{\boldsymbol{U}, \boldsymbol{\Sigma}, \boldsymbol{V}\}$ nach 11.1.7

$$\boldsymbol{A} + \delta\boldsymbol{A}_R = \boldsymbol{U}\boldsymbol{\Sigma}\boldsymbol{V}^\mathsf{T} \quad \text{mit} \quad \|\delta\boldsymbol{A}_R\|_F \leqq \nu F \|\boldsymbol{A}\|_F =: \Delta\boldsymbol{A}_R, \quad F \sim 8mn + 12kn. \tag{3}$$

Unter Beachtung von (2) ergibt sich daraus

$$\boldsymbol{A}^* + \delta\boldsymbol{A} = \boldsymbol{U}\boldsymbol{\Sigma}\boldsymbol{V}^\mathsf{T} \tag{4}$$

mit

$$\delta\boldsymbol{A} := \delta\boldsymbol{A}_D + \delta\boldsymbol{A}_M + \delta\boldsymbol{A}_R, \qquad \varepsilon := \|\delta\boldsymbol{A}\|_F \leqq \Delta\boldsymbol{A}_D + \Delta\boldsymbol{A}_M + \Delta\boldsymbol{A}_R =: \Delta\boldsymbol{A}. \tag{5}$$

Die berechnete SVD ist also die exakte Zerlegung einer um $\delta\boldsymbol{A}$ gestörten Matrix, und das Störungsniveau ε liegt in der Größenordnung des durch Darstellungs- und Meßfehler hervorgerufenen Niveaus. In diesem Sinne ist die berechnete SVD akzeptabel. Insbesondere folgt aus (4), (5) nach 8.2.3

$$|\sigma_i - \sigma_i^*| \leqq \varepsilon \qquad (i = 1, \ldots, n), \tag{6}$$

d. h., die exakten Singulärwerte σ_i^* werden ausreichend gut approximiert. Trotzdem ist die berechnete SVD nicht direkt zur Bestimmung von $(\boldsymbol{A}^*)^+$ und $\boldsymbol{x}^* = (\boldsymbol{A}^*)^+ \boldsymbol{b}$ geeignet: Auf Grund der Rundungs- und Meßfehler werden fast immer alle berechneten Singulärwerte

$$\sigma_1 \geqq \sigma_2 \geqq \cdots \geqq \sigma_r \geqq \sigma_{r+1} \geqq \cdots \geqq \sigma_n \tag{7}$$

sogar positiv sein, so daß S 1.2 von 11.1.5 den Rang n und damit einen zu hohen Wert liefert. Für $i > r$ folgt nun aus (1) sofort $\delta\sigma_i = \sigma_i - \sigma_i^* = \sigma_i$, mit (6) also

$$\varepsilon \geqq \sigma_{r+1} \geqq \cdots \geqq \sigma_n. \tag{8}$$

Die berechneten σ_i $(i > r)$ sind demnach kleine positive Zahlen, die wegen des zu groß festgelegten Ranges mit $1/\sigma_i$ in die Pseudoinverse $(\boldsymbol{A}^* + \delta\boldsymbol{A})^+ = \boldsymbol{V}\boldsymbol{\Sigma}^+\boldsymbol{U}^\mathsf{T}$ eingehen, also eine bösartige, rangerhöhende Störung von $\boldsymbol{\Sigma}^*$ darstellen. Die Bösartigkeit wirkt sich auch auf die Normallösung $\boldsymbol{x} = (\boldsymbol{A}^* + \delta\boldsymbol{A})^+ \boldsymbol{b} = \boldsymbol{V}\boldsymbol{\Sigma}^+\boldsymbol{U}^\mathsf{T}\boldsymbol{b}$ aus, sofern in $\boldsymbol{\beta} = \boldsymbol{U}^\mathsf{T}\boldsymbol{b}$ nicht alle Komponenten $\beta_{r+1}, \ldots, \beta_n$ verschwinden, was wegen der Rundungsfehler unwahrscheinlich ist; siehe 8.2.B.

Der beschriebene unerwünschte Effekt kann vermieden werden, indem $\boldsymbol{\Sigma}$ durch

Nullsetzen der „kleinen“ berechneten σ_i regularisiert wird:

$$\text{Statt } \boldsymbol{\Sigma} = \left(\begin{array}{c|c} \begin{matrix}\sigma_1 & & \\ & \ddots & \\ & & \sigma_p\end{matrix} & \\ \hline & \begin{matrix}\sigma_{p+1} & & \\ & \ddots & \\ & & \sigma_n\end{matrix} \end{array}\right) \quad \text{wird} \quad \boldsymbol{\Sigma}_\alpha := \left(\begin{array}{c|c} \begin{matrix}\sigma_1 & & \\ & \ddots & \\ & & \sigma_p\end{matrix} & \\ \hline & \boldsymbol{O} \end{array}\right) \tag{9}$$

verwendet, wobei der sog. *Pseudorang* $p = p(\alpha)$ in Abhängigkeit vom *Regularisierungsparameter* $\alpha \geqq 0$ durch die Bedingung

$$\sigma_1 \geqq \cdots \geqq \sigma_p > \alpha \geqq \sigma_{p+1} \geqq \cdots \geqq \sigma_n \tag{10}$$

festgelegt wird. Die derart regularisierende Singulärwertzerlegung $\boldsymbol{U}\boldsymbol{\Sigma}_\alpha\boldsymbol{V}^\mathsf{T}$ ist die exakte SVD der regularisierten Matrix

$$\boldsymbol{A}_\alpha := \boldsymbol{U}\boldsymbol{\Sigma}_\alpha\boldsymbol{V}^\mathsf{T} = \boldsymbol{A}^* + \delta\boldsymbol{A}_\alpha, \tag{11}$$

durch die eine regularisierte Pseudoinverse $\boldsymbol{A}_\alpha^+ := \boldsymbol{V}\boldsymbol{\Sigma}_\alpha^+\boldsymbol{U}^\mathsf{T}$ und eine regularisierte Normallösung $\boldsymbol{x}_\alpha := \boldsymbol{A}_\alpha^+\boldsymbol{b}$ definiert werden. Mit dieser Modifikation geht 11.1.5 in das folgende regularisierte Verfahren über:

11.2.1. *Lösung von* $\boldsymbol{Ax} \cong \boldsymbol{b}$ *mittels diskreter Regularisierung der SVD.*

S1 (Festlegung der regularisierten SVD):

S1.1: Bestimme Singulärwertzerlegung $\boldsymbol{A} = \boldsymbol{U}\boldsymbol{\Sigma}\boldsymbol{V}^\mathsf{T}$ nach 11.1.3.

S1.2: Wähle Regularisierungsniveau $\alpha \geqq 0$, bestimme Pseudorang $p = p(\alpha)$ gemäß (10), definiere $\boldsymbol{A}_\alpha$ nach (9), (11).

S2 (Berechnung der zu $\boldsymbol{b}$ gehörenden regularisierten Normallösung $\boldsymbol{x}_\alpha = \boldsymbol{A}_\alpha^+\boldsymbol{b} = \boldsymbol{V}\boldsymbol{\Sigma}_\alpha^+\boldsymbol{U}^\mathsf{T}\boldsymbol{b}$, des Residuums $\boldsymbol{r}_\alpha = \boldsymbol{b} - \boldsymbol{A}\boldsymbol{x}_\alpha = \boldsymbol{U}(\boldsymbol{I} - \boldsymbol{\Sigma}\boldsymbol{\Sigma}_\alpha^+)\,\boldsymbol{U}^\mathsf{T}\boldsymbol{b}$ und dessen Norm $n\boldsymbol{r}_\alpha = \|\boldsymbol{r}_\alpha\|$):

S2.1: Berechne $\boldsymbol{\beta} := \boldsymbol{U}^\mathsf{T}\boldsymbol{b}$

S2.2: Setze $\boldsymbol{\varrho}_\alpha := (\boldsymbol{I} - \boldsymbol{\Sigma}\boldsymbol{\Sigma}_\alpha^+)\,\boldsymbol{\beta} = (0, \ldots, 0, \beta_{p+1}, \ldots, \beta_m)^\mathsf{T}$, berechne $n\boldsymbol{r}_\alpha := \|\boldsymbol{\varrho}_\alpha\|$ und $\boldsymbol{r}_\alpha := \boldsymbol{U}\boldsymbol{\varrho}_\alpha$

S2.3: Setze $\boldsymbol{\xi}_\alpha := \boldsymbol{\Sigma}_\alpha^+\boldsymbol{\beta} = (\beta_1/\sigma_1, \ldots, \beta_p/\sigma_p, 0, \ldots, 0)^\mathsf{T}$, berechne $\boldsymbol{x}_\alpha := \boldsymbol{V}\boldsymbol{\xi}_\alpha$

Zur Festlegung des Regularisierungsparameters α beachten wir, daß die regularisierte Matrix $\boldsymbol{A}_\alpha$ durch die Störung

$$\delta\boldsymbol{A}_\alpha := \boldsymbol{A}_\alpha - \boldsymbol{A}^* = \delta\boldsymbol{A} - \boldsymbol{U}(\boldsymbol{\Sigma} - \boldsymbol{\Sigma}_\alpha)\,\boldsymbol{V}^\mathsf{T}, \quad \|\delta\boldsymbol{A}_\alpha\| \leqq \varepsilon + \sigma_{p(\alpha)+1} \leqq \varepsilon + \alpha \tag{12}$$

aus $\boldsymbol{A}^*$ entsteht, siehe (10), (11) und (4), (5). Nun sind die regularisierten Größen $\boldsymbol{A}_\alpha^+, \boldsymbol{x}_\alpha$ nach 8.2.5 bzw. 8.2.7 lokal lipschitzstetige Funktionen von $\delta\boldsymbol{A}_\alpha$, wenn die Bedingungen

$$p(\alpha) = \operatorname{rang}(\boldsymbol{\Sigma}_\alpha) = \operatorname{rang}(\boldsymbol{A}_\alpha) = \operatorname{rang}(\boldsymbol{A}^* + \delta\boldsymbol{A}_\alpha) \leqq \operatorname{rang}(\boldsymbol{A}^*) = r \tag{13}$$

und

$$\|(\boldsymbol{A}^*)^+\| \, \|\delta \boldsymbol{A}_\alpha\| < 1 \tag{14}$$

erfüllt sind; in (13) steht dann sogar das Gleichheitszeichen.

Die Ungleichung (13) ist wegen (8) im Fall

$$\varepsilon \leqq \alpha \tag{15}$$

sicher gültig. Durch die Festlegung (15) werden also rangerhöhende und damit bösartige Störungen $\delta \boldsymbol{A}_\alpha$ ausgeschlossen. Aus (12) folgt andererseits $\|(\boldsymbol{A}^*)^+\| \, \|\delta \boldsymbol{A}_\alpha\| \leqq (\varepsilon + \alpha)/\sigma_r^*$ so daß (14) im Fall

$$\alpha < \sigma_r^* - \varepsilon \tag{16}$$

erfüllt ist. Die beiden gegensätzlich wirkenden Bedingungen (15) und (16) lassen sich genau dann gleichzeitig erfüllen, wenn $\varepsilon \leqq \alpha < \sigma_r^* - \varepsilon$, also

$$2\varepsilon < \sigma_r^* = 1/\|(\boldsymbol{A}^*)^+\| \tag{17}$$

gilt; der kleinste zulässige α-Wert ist dann

$$\alpha := \alpha_{\text{opt}} := \varepsilon. \tag{18}$$

Für die mit diesem α berechneten regularisierten Größen gelten folgende Aussagen:

11.2.2. *Fehleranalyse.* Die Matrix $\boldsymbol{A}^* \in \mathbf{R}^{m,n}$ mit rang $(\boldsymbol{A}^*) = r$ und den Singulärwerten (1) werde durch die Computermatrix $\boldsymbol{A} \in \mathfrak{R}^{m,n}$ im Sinne von (4), (5) dargestellt. Für $\boldsymbol{A}$ und $\boldsymbol{b} \in \mathfrak{R}^m$ werde Algorithmus 11.2.1 mit $\alpha := \varepsilon$ durchgeführt. Dann genügen die in S1 berechneten Faktoren der regularisierten SVD der Beziehung

$$\boldsymbol{A}^* + \delta \boldsymbol{A}_\varepsilon = \boldsymbol{A}_\varepsilon = \boldsymbol{U}\boldsymbol{\Sigma}_\varepsilon \boldsymbol{V}^T \quad \text{mit} \quad \|\delta \boldsymbol{A}_\varepsilon\| \leqq 2\varepsilon, \tag{19}$$

und es gilt

$$\text{rang}\,(\boldsymbol{A}_\varepsilon) = \text{rang}\,(\boldsymbol{A}^* + \delta \boldsymbol{A}_\varepsilon) \leqq \text{rang}\,(\boldsymbol{A}^*). \tag{20}$$

Zu den in S2 berechneten regularisierten Größen $\boldsymbol{x}_\varepsilon$, $\boldsymbol{r}_\varepsilon$ existieren Störungen derart, daß

$$\hat{\boldsymbol{x}}_\varepsilon := \boldsymbol{x}_\varepsilon + \delta \boldsymbol{x}_\varepsilon \quad \text{mit} \quad \|\delta \boldsymbol{x}_\varepsilon\| \leqq \nu n(4n + 6k) \, \|\boldsymbol{x}_\varepsilon\| \tag{21}$$

die Normallösung des gestörten Quadratmittelproblems

$$(\boldsymbol{A}^* + \delta \boldsymbol{A}_\varepsilon)\, \hat{\boldsymbol{x}}_\varepsilon \cong \boldsymbol{b} + \delta \boldsymbol{b}_\varepsilon \quad \text{mit} \quad \|\delta \boldsymbol{b}_\varepsilon\| \leqq \nu m(4n + 6k) \, \|\boldsymbol{b}\| \tag{22}$$

ist, und es gilt

$$\hat{\boldsymbol{r}}_\varepsilon := \boldsymbol{r}_\varepsilon + \delta \boldsymbol{r}_\varepsilon = (\boldsymbol{b} + \delta \boldsymbol{b}_\varepsilon) - (\boldsymbol{A}^* + \delta \boldsymbol{A}_\varepsilon)\, \hat{\boldsymbol{x}}_\varepsilon$$

$$\text{mit} \quad \|\delta \boldsymbol{r}_\varepsilon\| \leqq \nu m(4n + 6k) \, \|\boldsymbol{r}_\varepsilon\|. \tag{23}$$

Wenn zusätzlich die Bedingung (17) erfüllt ist, gilt

$$\text{rang}\,(\boldsymbol{A}^* + \delta \boldsymbol{A}_\varepsilon) = \text{rang}\,(\boldsymbol{A}^*), \tag{24}$$

und $\boldsymbol{A}_\varepsilon^-$, $\hat{\boldsymbol{x}}_\varepsilon$ und $\hat{\boldsymbol{r}}_\varepsilon$ sind lokal lipschitzstetige Funktionen der Störungen $\delta \boldsymbol{A}_\varepsilon$, $\delta \boldsymbol{b}_\varepsilon$.

Beweis. Die zentralen Aussagen über A_ε^+ und die Rangbeziehungen wurden bereits hergeleitet; die restlichen ergeben sich aus den Formeln für x_ε und r_ε unter Beachtung von (11.1.27). □

11.2.2 besagt: *Die regularisierte, durch* $\{U, \Sigma_\varepsilon, V\}$ *gegebene SVD ist die exakte SVD der regularisierten Matrix* A_ε, *die sich von* A^* *durch eine Störung* δA_ε *in der Größenordnung des Fehlerniveaus* ε *unterscheidet. Sie stellt daher eine akzeptable SVD von* A^* *dar. Die regularisierte Normallösung* x_ε *und das zugehörige Residuum* r_ε *sind die leicht gestörte Normallösung und das leicht gestörte Residuum eines benachbarten Quadratmittelproblems. Die Störung* δA_ε *ist dabei nach Konstruktion nicht rangerhöhend, und bei Gültigkeit von* (17) *liegt sogar lokal lipschitzstetige Abhängigkeit vor.*

11.2.3. *Bemerkung.* (i) Wenn $b \in \Re^m$ der durch Darstellungs- und Meßfehler verfälschte Repräsentant der unbekannten rechten Seite $b^* \in \mathbf{R}^m$ ist, braucht nur

$$b = b^* + \delta b, \quad \delta b := \delta b_D + \delta b_M, \quad \|\delta b_D\| \leqq \nu \|b^*\|, \quad \|\delta b_M\| \leqq \Delta b_M$$

gesetzt zu werden.

(ii) Bei Erfülltsein von (17), d. h. im Fall $\varkappa := 2\varepsilon \|(A^*)^+\| < 1$, folgt aus (19) und 8.2.5

$$\frac{\|(A^*)^+ - A_\varepsilon^+\|}{\|(A^*)^+\|} \leqq \varepsilon \frac{(1+\sqrt{5})\,\|(A^*)^+\|}{1-\varkappa}, \tag{25}$$

und für $x^* := (A^*)^+ b$ sowie die exakt aus $\{U, \Sigma_\varepsilon, V\}$ berechnete regularisierte Lösung $A_\varepsilon^+ b$ gilt nach 8.2.7

$$\frac{\|x^* - A_\varepsilon^+ b\|}{\|x^*\|} \leqq \varepsilon \frac{2\,\|(A^*)^+\|}{1-\varkappa}\left[\sqrt{2} + \frac{\|(A^*)^+\|\,\|r^*\|}{\|x^*\|}\right] \tag{26}$$

mit $r^* = b - A^* x^*$. Unter Verwendung von (21), (22) kann $x^* - x_\varepsilon$ in analoger Weise abgeschätzt werden. Zur Schranke aus (26) kommen dabei noch Terme proportional zu ν hinzu.

(iii) Es scheint, als ob das Verfahren 11.2.1 mit $\alpha = \varepsilon$ rangdefiziente Probleme stets befriedigend löst. Das ist leider nicht immer der Fall: Zum einen ist das Fehlerniveau ε meist nicht genau bekannt. Bestenfalls kennt man Schranken die eine

$$\varepsilon_{\min} \leqq \varepsilon \leqq \varepsilon_{\max},$$

Unsicherheit der Rangbestimmung in den Grenzen

$$p_{\min} := p(\varepsilon_{\max}) \leqq p(\varepsilon) \leqq p(\varepsilon_{\min}) =: p_{\max}$$

nach sich ziehen. Zum anderen ist das gutartige, lokal lipschitzstetige Verhalten der regularisierten Größen an die Bedingung (17) gebunden, d. h., es liegt nur vor, wenn das Fehlerniveau deutlich kleiner als σ_r^* ist. Da r nicht bekannt ist, läßt sich diese Bedingung praktisch nicht überprüfen. Wenn (17) nicht erfüllt ist, brauchen sich (13) und (14) für kein $\alpha > 0$ gleichzeitig erfüllen zu lassen, d. h., *jeder* Regularisierungsversuch liefert entweder eine rangerhöhende oder eine zu große Störung. Eine solche Konfliktsituation kann nur durch Verwendung einer genügend hohen Genauig-

keit vermieden werden, sofern keine Meßfehler auftreten. Im Fall $\Delta \boldsymbol{A}_M = 0$ ist nämlich $\varepsilon \leqq \Delta \boldsymbol{A}_D + \Delta \boldsymbol{A}_R \leqq \nu(1 + F)\, \|\boldsymbol{A}^*\|_F$, so daß (17) für genügend kleines ν erfüllt ist. Bei Abwesenheit von Meßfehlern sollte dabei $\varepsilon_{\min} \approx \nu\, \|\boldsymbol{A}\| = \nu\sigma_1$ als untere Schranke für den ε-Bereich gewählt werden. Bei signifikanten Meßfehlern braucht (17) auch bei noch so hoher Computergenauigkeit nicht erfüllt zu sein. □

Die obigen Ausführungen legen es nahe, den Regularisierungsparameter $\alpha \geqq 0$ als variabel anzusehen und alle zugehörigen regularisierten Lösungen $\boldsymbol{x}_\alpha$ als mögliche Kandidaten für die gesuchte Lösung zu betrachten. Da $p(\alpha)$ nur die Werte $\{0, 1, \ldots, n\}$ annehmen kann, gibt es i. allg. $n + 1$ verschiedene $\boldsymbol{x}_\alpha$, deshalb auch der Name *diskrete Regularisierung*. Um den Einfluß von α auf die regularisierten Größen $\boldsymbol{x}_\alpha$, $\boldsymbol{r}_\alpha$ zu untersuchen, gehen wir von (4) aus und nehmen der Einfachheit halber an, daß $\{\boldsymbol{x}_\alpha, \boldsymbol{r}_\alpha\}$ fehlerfrei aus $\{\boldsymbol{b}, \boldsymbol{U}, \boldsymbol{\Sigma}, \boldsymbol{V}\}$ berechnet werden. Die in Gleitpunktarithmetik berechneten Größen unterscheiden sich von diesen nur unwesentlich. Mit den Bezeichnungen von 11.2.1 gilt dann

$$\|\boldsymbol{x}_\alpha\|^2 = \|\boldsymbol{\xi}_\alpha\|^2 = \sum_{j=1}^{p(\alpha)} (\beta_j/\sigma_j)^2 \leqq \sum_{j=1}^{\operatorname{rang}(\hat{\boldsymbol{A}})} (\beta_j/\sigma_j)^2 = \|\boldsymbol{x}^N\|^2 \tag{27}$$

bzw.

$$\|\boldsymbol{r}_\alpha\|^2 = \|\boldsymbol{\varrho}_\alpha\|^2 = \sum_{i=p(\alpha)+1}^{m} (\beta_i)^2 \geqq \sum_{i=\operatorname{rang}(\hat{\boldsymbol{A}})+1}^{m} (\beta_i)^2 = \|\boldsymbol{r}^N\|^2 \tag{28}$$

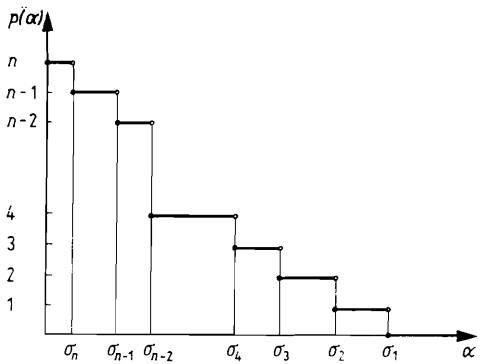


Abb. 11.2.1. $p(\alpha)$ als Funktion von α

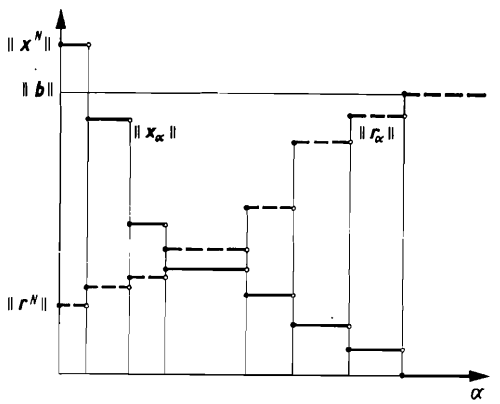


Abb. 11.2.2. $\|\boldsymbol{x}_\alpha\|$ und $\|\boldsymbol{r}_\alpha\|$ als Funktion von α

mit $\hat{\boldsymbol{A}} := \boldsymbol{A}^* + \delta\boldsymbol{A}$, $\boldsymbol{x}^N := \boldsymbol{x}_0 := \boldsymbol{A}^+\boldsymbol{b}$, $\boldsymbol{r}^N := \boldsymbol{r}_0 := \boldsymbol{b} - \boldsymbol{A}\boldsymbol{x}^N$. Die Normen $\|\boldsymbol{x}_\alpha\|$ und $\|\boldsymbol{r}_\alpha\|$ sind also stückweise konstante Funktionen, wobei $\|\boldsymbol{x}_\alpha\|$ monoton fällt, $\|\boldsymbol{r}_\alpha\|$ dagegen monoton wächst, siehe Abb. 11.2.1 und 11.2.2.

Um ein realistischeres Bild zu erhalten, müßte man sich die α-, $\|\boldsymbol{x}_\alpha\|$- und $\|\boldsymbol{r}_\alpha\|$-Achse logarithmisch geteilt denken. Im Fall $\Delta\boldsymbol{A}_M = 0$, $\nu \approx 10^{-6}$ ist eine typische Verteilung der berechneten Singulärwerte durch die Tabelle

i	1	2	...	r	$r+1$	$r+2$	...	n
σ_i	15.783	13.021	...	0.00352	0.75×10^{-4}	0.29×10^{-4}	...	0.83×10^{-5}

gegeben, wobei $\sigma_{r+1}, \ldots, \sigma_n$ in der Größenordnung von $\nu\,\|\boldsymbol{A}\| \approx \nu\sigma_1 = 1.5 \times 10^{-5}$ liegen und σ_r deutlich größer ist, also $p = r$ leicht erkannt werden kann. Mit fallendem α, d. h. mit wachsendem p, fällt $\|\boldsymbol{r}_\alpha\|$, während $\|\boldsymbol{x}_\alpha\|$ i. allg. nicht zu stark wächst. Wenn α in die Größenordnung von ε kommt, wächst $\|\boldsymbol{x}_\alpha\|$ i. allg. sehr stark an, meist um mehrere Zehnerpotenzen, während $\|\boldsymbol{r}_\alpha\|$ sich entsprechend der Zahlenwerte der β_i weiter — aber i. allg. nicht drastisch — verkleinert, siehe (27), (28). An Hand des Diagramms 11.2.2 kann der Bearbeiter dann denjenigen Pseudorang $p(\alpha)$ wählen, dessen zugehöriges $\boldsymbol{x}_\alpha$ und $\boldsymbol{r}_\alpha$ seinen Vorstellungen über die erwartete Lösung am besten entspricht. Geeignete Auswahlkriterien wären etwa

$$\|\boldsymbol{x}_\alpha\| \to \text{Minimum!} \quad \text{bei} \quad \|\boldsymbol{r}_\alpha\| \leqq \tau \tag{29}$$

oder

$$\|\boldsymbol{r}_\alpha\| \to \text{Minimum!} \quad \text{bei} \quad \|\boldsymbol{x}_\alpha\| \leqq \omega \tag{30}$$

mit vorgegebenen, aus der realen Aufgabenstellung abzuleitenden Schranken $\tau, \omega > 0$. In (29) muß dabei $\tau \geqq \|\boldsymbol{r}^N\|$ gefordert werden, damit überhaupt zulässige Lösungen existieren, vgl. (28). Zur Lösung von (30) wird p — von 0 beginnend — solange erhöht, daß die Nebenbedingung gerade noch erfüllt ist. Für (29) wird p solange erhöht, bis die Nebenbedingung erstmals erfüllt ist. Selbstverständlich erfolgt die Steuerung wie beschrieben direkt mit p und nicht mit α.

B. *Diskrete Regularisierung mittels anderer Faktorisierungen*

Um die relativ aufwendige Berechnung der SVD zu vermeiden, können auch andere Faktorisierungen zur diskreten Regularisierung herangezogen werden. Als Beispiel betrachten wir im folgenden die $\boldsymbol{QR}$-Faktorisierung mittels Householder-Spiegelungen, vgl. 10.2. Wegen der erwarteten Rangdefizienz kann dabei nicht spaltenweise wie im Vollrangfall vorgegangen werden, sondern es sind Spaltenvertauschungen gemäß 10.2.5(vii) erforderlich. Wenn das Verfahren mit $\boldsymbol{A} \in \mathfrak{R}^{m,n}$ gestartet wird, erhalten wir nach dem k-ten Schritt

$$\boldsymbol{A}^{(k+1)} = \begin{pmatrix} \boldsymbol{R}^{(k+1)} & \\ \boldsymbol{O} & \boldsymbol{M}^{(k+1)} \end{pmatrix} \begin{matrix} \\ k \end{matrix} \begin{matrix} \\ k+1 \end{matrix} = \boldsymbol{H}_k \cdots \boldsymbol{H}_1 \boldsymbol{A} \boldsymbol{T}_{1,\hat{s}(k)} \cdots \boldsymbol{T}_{k,\hat{s}(k)}$$

mit der Restmatrix

$$\boldsymbol{M}^{(k+1)} = (\boldsymbol{m}^{k+1,k+1}, \boldsymbol{m}^{k+1,k+2}, \ldots, \boldsymbol{m}^{k+1,n}).$$

Im Fall $k = r = \operatorname{rang}(\boldsymbol{A}^*)$ müßte bei fehlerfreiem $\boldsymbol{A} = \boldsymbol{A}^*$ und in exakter Arithmetik $\boldsymbol{M}^{(k+1)} = \boldsymbol{O}$ gelten. Bei Anwesenheit von Rundungs- und Meßfehlern wird jedoch auch in diesem Fall fast immer $\boldsymbol{M}^{(k+1)} \neq \boldsymbol{O}$ sein, so daß der nächste Schritt ausgeführt werden kann. Allerdings ist zu erwarten, daß $\boldsymbol{M}^{(k+1)}$ klein ist. Dann wird auch das nächste Diagonalelement klein sein, denn wegen der Spaltenvertauschungsstrategie gilt

$$|r_{k+1,k+1}| = \max\{\|\boldsymbol{m}^{k+1,j}\| : j = k+1, \ldots, n\}, \tag{31}$$

vgl. (10.2.21). Sollte doch $\boldsymbol{M}^{(k+1)} = \boldsymbol{O}$ sein, so wird einfach $\boldsymbol{H}_j := \boldsymbol{I}_m$, $\boldsymbol{T}_{j,\hat{s}(j)} := \boldsymbol{I}_n$ $(j = k+1, \ldots, n)$ gesetzt. Mit dieser Festlegung erhalten wir auch im rangdefizienten Fall stets eine Faktorisierung von $\boldsymbol{A}$, und nach 10.2.5 und 10.2.7 gilt

$$(\boldsymbol{A} + \delta\boldsymbol{A}_R)\,\boldsymbol{P}^{\mathsf{T}} = \boldsymbol{Q}\boldsymbol{R} \quad \text{mit} \quad \|\delta\boldsymbol{A}_R\| \leqq \nu F\,\|\boldsymbol{A}\| =: \Delta\boldsymbol{A}_R, \qquad F \leqq 4mn^{3/2}. \tag{32}$$

Dabei beschreibt $\boldsymbol{P}^{\mathsf{T}} := \boldsymbol{T}_{1,\hat{s}(1)} \cdots \boldsymbol{T}_{n-1,\hat{s}(n-1)}$ die Spaltenvertauschungen, $\boldsymbol{Q} := \boldsymbol{H}_1 \cdots \boldsymbol{H}_n$ ist eine exakt orthogonale Matrix und $\boldsymbol{R}$ die berechnete obere Dreiecksmatrix mit

$$|r_{11}| \geqq |r_{22}| \geqq \cdots \geqq |r_{nn}|.$$

Der in 10.2.7 geforderte Vollrang und die Bedingung (10.2.25) werden dort nur gebraucht, um die Durchführbarkeit mit $r_{kk} \neq 0$ zu sichern, was bei der obigen Modifikation überflüssig wird. Wenn zusätzlich wieder Darstellungs- und Meßfehler analog zu (2) zugelassen werden, ergibt sich schließlich

$$(\boldsymbol{A}^* + \delta\boldsymbol{A})\,\boldsymbol{P}^{\mathsf{T}} = \boldsymbol{Q}\boldsymbol{R} \quad \text{mit} \quad \varepsilon := \|\delta\boldsymbol{A}\| \leqq \Delta\boldsymbol{A}_D + \Delta\boldsymbol{A}_M + \Delta\boldsymbol{A}_R =: \Delta\boldsymbol{A}. \tag{33}$$

Da fast immer $r_{kk} \neq 0$ $(k = 1, \ldots, n)$, also $\operatorname{rang}(\boldsymbol{R}) = n$ sein wird, ist $\delta\boldsymbol{A}$ i. allg. eine rangerhöhende bösartige Störung. Zum Zweck der Regularisierung wird dann analog zu (10) der Pseudorang $p = p(\alpha)$ durch

$$|r_{11}| \geqq \cdots \geqq |r_{pp}| > \alpha \geqq |r_{p+1,p+1}| \geqq \cdots \geqq |r_{nn}| \tag{34}$$

festgelegt. Wegen (31) ist dies gleichwertig mit der Definition von p als kleinstem Index aus $\{0, 1, \ldots, n\}$, für den

$$\|\boldsymbol{m}^{k+1,j}\| \leqq \alpha \qquad (j = p+1, \ldots, n) \tag{35}$$

gilt, d. h., eine im Sinne von (35) spaltenweise kleine Restmatrix $\boldsymbol{M}^{(k+1)}$ wird als vernachlässigbar angesehen. Die Regularisierung besteht dann im Ersatz von

$$\boldsymbol{R} = \left(\begin{array}{c|c} \boldsymbol{R}_{11} & \boldsymbol{R}_{12} \\ \hline \boldsymbol{O} & \boldsymbol{R}_{22} \\ \hline \multicolumn{2}{c}{\boldsymbol{O}} \end{array}\right)\begin{matrix}1\\p\\p+1\\n\\n+1\\m\end{matrix} \quad \text{durch} \quad \boldsymbol{R}_\alpha := \left(\begin{array}{c|c} \boldsymbol{R}_{11} & \boldsymbol{R}_{12} \\ \hline \boldsymbol{O} & \boldsymbol{O} \\ \hline \multicolumn{2}{c}{\boldsymbol{O}} \end{array}\right) = \left(\begin{array}{c} \boldsymbol{R}_1 \\ \hline \boldsymbol{O} \\ \hline \end{array}\right)\begin{matrix}1\\p\\p+1\\m\end{matrix} \tag{36}$$

mit

$$\boldsymbol{R}_1 := (\boldsymbol{R}_{11} \mid \boldsymbol{R}_{12}) := \left(\begin{array}{ccc|ccc} r_{11} & \cdots & r_{1p} & r_{1,p+1} & \cdots & r_{1n} \\ & \ddots & \vdots & \vdots & & \vdots \\ 0 & & r_{pp} & r_{p,p+1} & \cdots & r_{pn} \end{array}\right). \tag{37}$$

Die derart regularisierte $\boldsymbol{QR}$-Faktorisierung ist die exakte Faktorisierung der regularisierten Matrix

$$\boldsymbol{A}_\alpha := \boldsymbol{Q}\boldsymbol{R}_\alpha\boldsymbol{P} = \boldsymbol{A}^* + \delta\boldsymbol{A}_\alpha, \tag{38}$$

wobei wegen (33)

$$\delta\boldsymbol{A}_\alpha := \boldsymbol{A}_\alpha - \boldsymbol{A}^* = \delta\boldsymbol{A} - \boldsymbol{Q}(\boldsymbol{R} - \boldsymbol{R}_\alpha)\,\boldsymbol{P}, \qquad \|\delta\boldsymbol{A}_\alpha\| \leqq \varepsilon + \sqrt{n-p}\,\alpha \tag{39}$$

gilt. Man beachte, daß aus (35)

$$\|\boldsymbol{R} - \boldsymbol{R}_\alpha\|_F = \|\boldsymbol{R}_{22}\|_F = \|\boldsymbol{M}^{(p+1)}\|_F \leqq \sqrt{n-p}\,\alpha$$

folgt.

Aus der Faktorisierung (38) ergibt sich nach 8.1.9

$$\boldsymbol{A}_\alpha^+ = \boldsymbol{P}^\mathsf{T}\boldsymbol{R}_\alpha^+\boldsymbol{Q}^\mathsf{T} = \boldsymbol{P}^\mathsf{T}(\boldsymbol{R}_1^+ \mid \boldsymbol{O})\,\boldsymbol{Q}^\mathsf{T}. \tag{40}$$

Zur Darstellung von $\boldsymbol{R}_1^+$ transformieren wir $\boldsymbol{R}_1$ gemäß

$$\boldsymbol{R}_1\bar{\boldsymbol{Q}}^\mathsf{T} = (\boldsymbol{R}_{11} \mid \boldsymbol{R}_{12})\,\bar{\boldsymbol{H}}_p \cdots \bar{\boldsymbol{H}}_2\bar{\boldsymbol{H}}_1 = (\bar{\boldsymbol{R}}_{11} \mid \boldsymbol{O}) =: \bar{\boldsymbol{R}}_1 \tag{41}$$

mittels p Householder-Spiegelungen $\bar{\boldsymbol{H}}_k = \boldsymbol{I} - \bar{\boldsymbol{v}}^k(\bar{\boldsymbol{v}}^k)^\mathsf{T}/\bar{\gamma}_k$ in $\bar{\boldsymbol{R}}_1$, wobei $\bar{\boldsymbol{R}}_{11}$ eine obere Dreiecksmatrix sein soll. Dies ist möglich, wenn $\bar{\boldsymbol{v}}^k$ in der Form

$$\bar{\boldsymbol{v}}^k = (0, \ldots, 0, \bar{v}_{kk}, 0, \ldots, 0, \bar{v}_{p+1,k}, \ldots, \bar{v}_{nk})^\mathsf{T} \qquad (k = p, \ldots, 1)$$

angesetzt wird und damit Nullen in den Positionen $(k, p+1)$ bis (k, n) von $\boldsymbol{R}_1$ erzeugt werden, vgl. Ü 3.3.7. Aus (41) folgt dann wieder nach 8.1.9

$$\boldsymbol{R}_1^+ = (\bar{\boldsymbol{R}}_1\bar{\boldsymbol{Q}})^+ = \bar{\boldsymbol{Q}}^\mathsf{T}\bar{\boldsymbol{R}}_1^+ = \bar{\boldsymbol{Q}}^\mathsf{T}\left(\frac{\bar{\boldsymbol{R}}_{11}^{-1}}{\boldsymbol{O}}\right), \tag{42}$$

mit (40) also

$$\boldsymbol{A}_\alpha^+ = \boldsymbol{P}^\mathsf{T}\bar{\boldsymbol{Q}}^\mathsf{T}\left(\begin{array}{c|c} \bar{\boldsymbol{R}}_{11}^{-1} & \boldsymbol{O} \\ \hline \boldsymbol{O} & \boldsymbol{O} \end{array}\right)\boldsymbol{Q}^\mathsf{T}. \tag{43}$$

Die Produktdarstellung (43) ist die Grundlage für das folgende Verfahren:

11.2.4. *Lösung von $\boldsymbol{A}\boldsymbol{x} \simeq \boldsymbol{b}$ mittels diskreter Regularisierung der $\boldsymbol{QR}$-Faktorisierung.*

S1 (Festlegung der regularisierten $\boldsymbol{QR}$-Faktorisierung):

S1.1: Bestimme Faktorisierung $\boldsymbol{A}\boldsymbol{P}^\mathsf{T} = \boldsymbol{Q}\boldsymbol{R}$ gemäß 10.2.5(vii) mit der oben beschriebenen Modifikation im Fall $\boldsymbol{M}^{(k+1)} = \boldsymbol{O}$

S1.2: Wähle Regularisierungsniveau $\alpha \geqq 0$, bestimme Pseudorang $p = p(\alpha)$ gemäß (34) und lege $\boldsymbol{R}_\alpha$ nach (36) fest.

S1.3: Transformiere $(\boldsymbol{R}_{11} \mid \boldsymbol{R}_{12})$ gemäß (41) in $(\bar{\boldsymbol{R}}_{11} \mid \boldsymbol{O}) = \boldsymbol{R}_1 \bar{\boldsymbol{Q}}^\top$.

S2 (Berechnung der regularisierten Normallösung $\boldsymbol{x}_\alpha = \boldsymbol{A}_\alpha^+ \boldsymbol{b}$, des Residuums $\boldsymbol{r}_\alpha = \boldsymbol{b} - \boldsymbol{A}\boldsymbol{x}_\alpha$ und dessen Norm $n\boldsymbol{r}_\alpha = \|\boldsymbol{r}_\alpha\|$):

S2.1: Berechne

$$\boldsymbol{c} = \begin{pmatrix} \boldsymbol{c}^1 \\ \boldsymbol{c}^2 \end{pmatrix} := \boldsymbol{Q}^\top \boldsymbol{b}, \qquad \boldsymbol{c}^1 \in \mathbf{R}^p, \quad \boldsymbol{c}^2 \in \mathbf{R}^{m-p}$$

S2.2: Bestimme $\boldsymbol{z}^1$ aus dem Dreieckssystem $\bar{\boldsymbol{R}}_{11}\boldsymbol{z}^1 = \boldsymbol{c}^1$, setze $\boldsymbol{z} := \begin{pmatrix} \boldsymbol{z}^1 \\ \boldsymbol{o} \end{pmatrix}$ mit $\boldsymbol{o} \in \mathbf{R}^{n-p}$

S2.3: Berechne $\boldsymbol{y} := \begin{pmatrix} \boldsymbol{y}^1 \\ \boldsymbol{y}^2 \end{pmatrix} := \bar{\boldsymbol{Q}}^\top \boldsymbol{z} = \bar{\boldsymbol{H}}_p \cdots \bar{\boldsymbol{H}}_1 \boldsymbol{z}$, setze $\boldsymbol{x}_\alpha := \boldsymbol{P}^\top \boldsymbol{y}$

S2.4: Berechne $\boldsymbol{d}^2 := \boldsymbol{c}^2 - \boldsymbol{R}_{22}\boldsymbol{y}^2$, setze $n\boldsymbol{r}_\alpha := \|\boldsymbol{d}^2\|$ und $\boldsymbol{d} := \begin{pmatrix} \boldsymbol{o} \\ \boldsymbol{d}^2 \end{pmatrix} \in \mathbf{R}^m$, berechne $\boldsymbol{r}_\alpha := \boldsymbol{Q}\boldsymbol{d}$

Der Ausdruck für $\boldsymbol{r}_\alpha$ ergibt sich dabei aus $\boldsymbol{r}_\alpha = \boldsymbol{b} - \boldsymbol{A}\boldsymbol{x}_\alpha = \boldsymbol{Q}(\boldsymbol{c} - \boldsymbol{R}\boldsymbol{y})$ unter Beachtung von

$$\boldsymbol{R}\boldsymbol{y} = \begin{pmatrix} \boldsymbol{R}_{11}\boldsymbol{y}^1 + \boldsymbol{R}_{12}\boldsymbol{y}^2 \\ \hline \boldsymbol{R}_{22}\boldsymbol{y}^2 \end{pmatrix} = \begin{pmatrix} \boldsymbol{c}^1 \\ \hline \boldsymbol{R}_{22}\boldsymbol{y}^2 \end{pmatrix}.$$

Wir fragen jetzt wieder, wann die für das gutartige Verhalten von $\delta\boldsymbol{A}_\alpha$ zu fordernden Bedingungen

$$p(\alpha) = \operatorname{rang}(\boldsymbol{R}_\alpha) = \operatorname{rang}(\boldsymbol{A}_\alpha) = \operatorname{rang}(\boldsymbol{A}^* + \delta\boldsymbol{A}_\alpha) \leqq \operatorname{rang}(\boldsymbol{A}^*) = r \tag{44}$$

und

$$\|(\boldsymbol{A}^*)^+\| \, \|\delta\boldsymbol{A}_\alpha\| < 1 \tag{45}$$

erfüllt sind. Zur Überprüfung von (44) betrachten wir die ersten p Spalten der Gleichung (33). Wenn die entsprechenden Teilmatrizen mit dem unteren Index 0 gekennzeichnet werden, ergibt sich

$$(\boldsymbol{A}^*\boldsymbol{P}^\top)_0 = (\boldsymbol{Q}\boldsymbol{R})_0 - (\delta\boldsymbol{A} \cdot \boldsymbol{P}^\top)_0 = \boldsymbol{Q}\boldsymbol{R}_0 - (\delta\boldsymbol{A} \cdot \boldsymbol{P}^\top)_0.$$

Da $\boldsymbol{R}_0$ nach Konstruktion Vollrang p hat, folgt im Fall

$$\|(\boldsymbol{Q}\boldsymbol{R}_0)^+\| \, \|(\delta\boldsymbol{A} \cdot \boldsymbol{P}^\top)_0\| < 1 \tag{46}$$

aus 8.2.5 $p = \operatorname{rang}(\boldsymbol{Q}\boldsymbol{R}_0) = \operatorname{rang}\big((\boldsymbol{A}^*\boldsymbol{P}^\top)_0\big)$, wegen $\operatorname{rang}\big((\boldsymbol{A}^*\boldsymbol{P}^\top)_0\big) \leqq \operatorname{rang}(\boldsymbol{A}^*\boldsymbol{P}^\top) = \operatorname{rang}(\boldsymbol{A}^*)$ also (44). Nun gelten

$$\|(\boldsymbol{Q}\boldsymbol{R}_0)^+\| = \|\boldsymbol{R}_0^+\| = \left\| \begin{pmatrix} \boldsymbol{R}_{11} \\ \hline \boldsymbol{O} \end{pmatrix}^+ \right\| = \|(\boldsymbol{R}_{11}^{-1} \mid \boldsymbol{O})\| = \|\boldsymbol{R}_{11}^{-1}\|$$

und $\|(\delta\boldsymbol{A} \cdot \boldsymbol{P}^\top)_0\| \leqq \|\delta\boldsymbol{A} \cdot \boldsymbol{P}^\top\| = \|\delta\boldsymbol{A}\| = \varepsilon$, so daß (46) und damit (44) sicher erfüllt sind, wenn $\boldsymbol{R}_{11} = \boldsymbol{R}_{11}(\alpha)$ der Bedingung

$$\|\boldsymbol{R}_{11}(\alpha)^{-1}\| \, \varepsilon < 1 \tag{47}$$

genügt. Zur Abschätzung von $\|\boldsymbol{R}_{11}^{-1}\|$ beachten wir den folgenden Satz: Für jede reguläre und im Sinne von $\|\boldsymbol{R}_{11}\boldsymbol{e}^j\|^2 = \sum_{i=1}^{j} r_{ij}^2 = 1$ $(j = 1, \ldots, p)$ spaltennormierte obere Dreiecksmatrix $\boldsymbol{R}_{11} \in \mathbf{R}^{p,p}$, deren Elemente der Ungleichung (10.2.23) genügen und die daher durch $\boldsymbol{QR}$-Faktorisierung mit Spaltenvertauschungen aus einer spaltennormierten Matrix $\boldsymbol{A}$ entstanden sein kann, gilt

$$|r_{pp}| \geqq \sigma_p(\boldsymbol{R}_{11}) = 1/\|\boldsymbol{R}_{11}^{-1}\| \geqq |r_{pp}|/2^{p-1}, \tag{48}$$

siehe B 11.2. Die Abschätzung nach unten ist dabei leider fast scharf, siehe Ü 11.2.2. Da (48) auch für nicht spaltennormiertes $\boldsymbol{R}_{11}$ gilt, folgt mit (34)

$$\|\boldsymbol{R}_{11}^{-1}\|\,\varepsilon \leqq 2^{p-1}\varepsilon/|r_{pp}| < 2^{p-1}\varepsilon/\alpha,$$

und die für die Gültigkeit von (44) hinreichende Bedingung (47) ist erfüllt, wenn

$$2^{p-1}\varepsilon \leqq \alpha \tag{49}$$

gilt. Während bei Verwendung der SVD bereits durch die Festlegung $\varepsilon \leqq \alpha$ eine Rangerhöhung ausgeschlossen wird, tritt hier zusätzlich der Faktor 2^{p-1} auf, der bei hohem Pseudorang $p = p(\alpha)$ ein gegenüber ε wesentlich vergrößertes Regularisierungsniveau bedingt. Daß ein solches ungünstiges Verhalten tatsächlich vorliegen kann, zeigt das folgende Beispiel.

11.2.5. *Beispiel.* Es sei $\boldsymbol{R}_{11} := \boldsymbol{R}(n, c, s)$ die Matrix aus Ü 11.2.2 für $n := 15$ und $c := 1 - s^2$, $s := 0.01^{1/14}$. Bei dieser Festlegung ergibt sich

$$1 = |r_{11}| \geqq \cdots \geqq |r_{15,15}| = 0.01.$$

Für jedes $\alpha < 0.01$ — also für relativ große α — ist $p(\alpha) = 15$, d. h., $\boldsymbol{R}_{11}$ wird als regulär angesehen. Nach Ü 11.2.2 gilt jedoch

$$\sigma_{15}(\boldsymbol{R}_{11}) \leqq |r_{nn}|/K(n, c) = 0.01/815 = 1.2 \times 10^{-5},$$

so daß $\boldsymbol{R}_{11}$ bereits im Rahmen des Fehlerniveaus $\varepsilon \approx 10^{-5}$ singulär ist. □

Wir wenden uns jetzt der zweiten Bedingung (45) zu. Sie ist wegen (39) sicher erfüllt, falls $(\varepsilon + \sqrt{n-p}\,\alpha)/\sigma_r^* < 1$, also die (16) entsprechende Ungleichung

$$\alpha < (\sigma_r^* - \varepsilon)/\sqrt{n-p} \tag{50}$$

gilt. Die beiden für das gutartige Verhalten von $\boldsymbol{A}$ hinreichenden Bedingungen (49), (50) sind daher genau dann gleichzeitig erfüllbar, wenn

$$\left(1 + \sqrt{n-r}\;2^{r-1}\right)\varepsilon \leqq \sigma_r^* = 1/\|(\boldsymbol{A}^*)^-\| \tag{51}$$

gilt.

11.2.6. *Bemerkung.* (i) Die Bedingung (51) ist für großes r wesentlich einschränkender als die bei Verwendung der SVD auftretende analoge Bedingung (17). Wenn sie nicht erfüllt ist, kann es spaltennormierte Matrizen $\boldsymbol{A}$ geben, so daß die hinreichenden Bedingungen (44) und (45) durch kein α befriedigt werden können, d. h., jeder Regularisierungsversuch würde entweder eine rangerhöhende oder eine

zu große Störung einführen. Es liegt in diesem Fall eine unlösbare Konfliktsituation in der Wahl von α vor. *Die numerische Rangbestimmung und Regularisierung auf der Grundlage der* $\boldsymbol{QR}$*-Faktorisierung mit Spaltenvertauschungen ist daher mit einer wesentlich höheren Unsicherheit verbunden als die mittels SVD.*

(ii) Selbstverständlich braucht die beschriebene Konfliktsituation nicht immer aufzutreten. Für viele praktisch vorkommende Matrizen werden die Bedingungen (44), (45) für $\alpha \approx \varepsilon$ erfüllt sein. In diesem Fall gelten für die berechneten Größen $\boldsymbol{A}_\alpha$ und $\boldsymbol{x}_\alpha$ zu 11.2.2 analoge Aussagen, jedoch ist $\boldsymbol{r}_\alpha$ i. allg. nicht das leicht gestörte Residuum eines leicht gestörten Nachbarproblems. Überdies gilt dann

$$\frac{1}{\sqrt{n-p+1}} \frac{|r_{11}|}{|r_{pp}|} \leqq \frac{\max\{|\bar{r}_{jj}| : j = 1, \ldots, p\}}{\min\{|\bar{r}_{jj}| : j = 1, \ldots, p\}} \leqq \operatorname{cond}(\boldsymbol{R}_1) = \operatorname{cond}(\boldsymbol{A}_\varepsilon) \approx \operatorname{cond}(\boldsymbol{A}^*), \tag{52}$$

man vgl. 6.1.5(iii) und beachte die Bildung von $\bar{\boldsymbol{R}}_1$. Die unteren Schranken können als Konditionsschätzer verwendet werden.

(iii) Wenn alle zu $p \in \{0, \ldots, n\}$ gehörenden regularisierten Lösungen gesucht sind, sollten diese in der Reihenfolge $p := n(-1)0$ berechnet werden. Zur Realisierung von (41) sind dann Givens-Drehungen günstiger, weil beim Übergang von p zu $p-1$ nur die ersten $p-1$ Elemente der p-ten Spalte von $\boldsymbol{R}_1 = \boldsymbol{R}_1^{(p)}$ durch Rechtsmultiplikation mit $\boldsymbol{G}_{p-1,p}^\mathsf{T}, \ldots, \boldsymbol{G}_{1p}^\mathsf{T}$ annulliert zu werden brauchen.

(iv) Eine zu 11.2.4 analoge Regularisierung ist auf der Grundlage der MGS-Faktorisierung mit Spaltenvertauschungen gemäß 10.1.6(iii) möglich.

(v) Da die (n, n)-Dreiecksmatrix $\tilde{\boldsymbol{R}} = \left(\begin{array}{c|c} \boldsymbol{R}_{11} & \boldsymbol{R}_{12} \\ \hline \boldsymbol{O} & \boldsymbol{R}_{22} \end{array}\right)$ im Faktor $\boldsymbol{R} = \left(\begin{array}{c} \tilde{\boldsymbol{R}} \\ \hline \boldsymbol{O} \end{array}\right)$ von $\boldsymbol{AP}^\mathsf{T} = \boldsymbol{QR}$ in exakter Arithmetik bis auf das Vorzeichen der Spalten mit dem Faktor $\boldsymbol{L}$ der mit Diagonalpivotisierung gemäß 6.1.5(iii) berechneten Cholesky-Faktorisierung $\boldsymbol{P}(\boldsymbol{A}^\mathsf{T}\boldsymbol{A})\boldsymbol{P}^\mathsf{T} = \boldsymbol{L}\boldsymbol{L}^\mathsf{T}$ übereinstimmt, kann 11.2.4 äquivalent als diskrete Regularisierung der Normalgleichungen formuliert werden, siehe Ü 11.2.3. Die Ungleichung (10.2.23) geht dann in (6.1.19) über. Bei Computerrechnung treten allerdings die durch die explizite Bildung von $\boldsymbol{M} = \boldsymbol{A}^\mathsf{T}\boldsymbol{A}$ hervorgerufenen Probleme auf, siehe Kapitel 9.

(vi) Wenn $r = \operatorname{rang}(\boldsymbol{A}^*)$ bekannt ist, braucht für die Gutartigkeit von $\delta\boldsymbol{A}_\alpha$ nur noch (45) bzw. (50) gefordert zu werden. Für $\alpha := \varepsilon$ ist letztere Forderung kaum einschränkender als (17), so daß die Bedenken gegen eine Verwendung der $\boldsymbol{QR}$-Faktorisierung in diesem Fall entfallen. □

Übungsaufgaben

Ü 11.2.1. Die Matrix $\boldsymbol{A} \in \mathbf{R}^{m,n}$ besitze die exakte Singulärwertzerlegung $\boldsymbol{A} = \boldsymbol{U}\boldsymbol{\Sigma}\boldsymbol{V}^\mathsf{T}$. Die Zahl $p = p(\alpha, \boldsymbol{A})$ sei gemäß (10) festgelegt, und $\boldsymbol{A}_\alpha = \boldsymbol{A}_\alpha(\alpha, \boldsymbol{A})$ sei die durch (9), (11) definierte regularisierte Matrix. Man zeige, daß

$$p(\alpha, \boldsymbol{A}) = \min\{\operatorname{rang}(\boldsymbol{A} + \delta\boldsymbol{A}) : \|\delta\boldsymbol{A}\| \leqq \alpha\}$$

gilt und $\boldsymbol{B} = \boldsymbol{A}_\alpha$ die Lösung von

$$\min \{\|\boldsymbol{A} - \boldsymbol{B}\|_F : \operatorname{rang}(\boldsymbol{B}) \leqq p(\alpha, \boldsymbol{A})\}$$

ist.

Ü 11.2.2. Für $c, s > 0$ werde die Matrix

$$\boldsymbol{R} = \boldsymbol{R}(n, c, s) := \begin{pmatrix} 1 & -c & -c & -c & \dots & -c \\ & s & -cs & -cs & \dots & -cs \\ & & s^2 & -cs^2 & \dots & -cs^2 \\ & & & \ddots & \ddots & \\ & & & & \ddots & \\ & & & & & s^{n-1} \end{pmatrix},$$

$$r_{ij} := \begin{cases} 0 & \text{für } i > j, \\ s^{i-1} & \text{für } i = j, \\ -cs^{i-1} & \text{für } i < j \end{cases}$$

betrachtet. Man zeige:

(i) Es gilt $\boldsymbol{u} = \boldsymbol{R}\boldsymbol{v}$ für $\boldsymbol{u} := s^{n-1}\boldsymbol{e}^n$ und

$$\boldsymbol{v} := (c(1+c)^{n-2}, c(1+c)^{n-3}, \dots, c(1+c), c, 1)^T.$$

folglich

$$\sigma_n(\boldsymbol{R}) = 1/\|\boldsymbol{R}^{-1}\| \leqq \|\boldsymbol{u}\|/\|\boldsymbol{v}\| =: |r_{nn}|/K(n, c)$$

mit

$$[K(n, c)]^2 = 1 + \frac{c^2[(1+c)^{2n-2} - 1]}{(1+c)^2 - 1}.$$

(ii) Im Fall $c := s := 1$ sind alle Diagonalelemente gleich 1, aber es gilt

$$\sigma_n(\boldsymbol{R})/|r_{nn}| \leqq 1/K(n, 1) \leqq \sqrt{3}/2^{n-1}.$$

Die Elemente von $\boldsymbol{R}$ erfüllen jedoch die Bedingung (10.2.23) nicht, d. h., $\boldsymbol{R}$ kann nicht durch $\boldsymbol{QR}$-Faktorisierung mit Spaltenvertauschungen entstanden sein.

(iii) Bei der Festlegung $0 < c < 1$, $s := \sqrt{1 - c^2}$ ist (10.2.23) dagegen erfüllt, und $\boldsymbol{R}$ ist spaltennormiert. Außerdem gilt

$$\sigma_n(\boldsymbol{R})/|r_{nn}| \leqq 1/K(n, c) \xrightarrow[c \to 1-0]{} 1/K(n, 1) \leqq \sqrt{3}/2^{n-1}.$$

Ü 11.2.3. Es sei $\boldsymbol{PMP}^T = \boldsymbol{LL}^T$ die in exakter Arithmetik nach 6.1.5 (iii) berechnete Cholesky-Faktorisierung von $\boldsymbol{M} := \boldsymbol{A}^T\boldsymbol{A}$; im Fall $\max\{a_{jj}^{(k+1)} : k+1 \leqq j \leqq n\} = 0$ werde dabei $\boldsymbol{L}_{k-1} = \cdots \boldsymbol{L}_{n-1} = \boldsymbol{I}$ gesetzt (bei Computerrechnung ist der Test $\max\{a_{jj}^{(k+1)} : k+1 \leqq j \leqq n\} \leqq 0$ zu verwenden und zusätzlich $\boldsymbol{M}^{(k+1)} := \boldsymbol{O}$ zu setzen). Man zeige:

(i) Wenn $p = p(\alpha)$ aus

$$l_{11} \geqq \cdots \geqq l_{pp} > \alpha \geqq l_{p+1,p+1} \geqq \cdots \geqq l_{nn}$$

bestimmt und

$$\boldsymbol{L} = \left(\begin{array}{c|c} \boldsymbol{L}_{11} & \boldsymbol{O} \\ \hline \boldsymbol{L}_{21} & \boldsymbol{L}_{22} \end{array}\right) \quad \text{zu} \quad \boldsymbol{L}_\alpha := \left(\begin{array}{c|c} \boldsymbol{L}_{11} & \boldsymbol{O} \\ \hline \boldsymbol{L}_{21} & \boldsymbol{O} \end{array}\right) \quad \text{mit} \quad \boldsymbol{L}_{11} \in \mathbf{R}^{p,p}$$

regularisiert wird, gilt

$$(\boldsymbol{A}_\alpha \boldsymbol{P}^\mathsf{T})^\mathsf{T} (\boldsymbol{A}_\alpha \boldsymbol{P}^\mathsf{T}) = \boldsymbol{P}\boldsymbol{A}_\alpha^\mathsf{T}\boldsymbol{A}_\alpha \boldsymbol{P}^\mathsf{T} = \boldsymbol{L}_\alpha \boldsymbol{L}_\alpha^\mathsf{T},$$

wobei $\boldsymbol{A}_\alpha \boldsymbol{P}^\mathsf{T} = (\boldsymbol{A}_1(\alpha) \mid \boldsymbol{A}_2(\alpha))$ aus $\boldsymbol{A}\boldsymbol{P}^\mathsf{T} = (\boldsymbol{A}_1 \mid \boldsymbol{A}_2)$ nach der Vorschrift $\boldsymbol{A}_1(\alpha) := \boldsymbol{A}_1$, $\boldsymbol{A}_2(\alpha) := \boldsymbol{A}_1 \boldsymbol{L}_{11}^{-\mathsf{T}} \boldsymbol{L}_{21}^\mathsf{T}$ gebildet wird.

(ii) Die Normalgleichungen

$$(\boldsymbol{A}_\alpha \boldsymbol{P}^\mathsf{T})^\mathsf{T} (\boldsymbol{A}_\alpha \boldsymbol{P}^\mathsf{T})\, \boldsymbol{y} = (\boldsymbol{A}_\alpha \boldsymbol{P}^\mathsf{T})^\mathsf{T} \boldsymbol{b}, \qquad \boldsymbol{y} := \begin{pmatrix} \boldsymbol{y}^1 \\ \boldsymbol{y}^2 \end{pmatrix} := \boldsymbol{P}\boldsymbol{x}_\alpha$$

des regularisierten Quadratmittelproblems $\boldsymbol{A}_\alpha \boldsymbol{x}_\alpha \cong \boldsymbol{b}$ reduzieren sich auf

$$\boldsymbol{L}_{11}\boldsymbol{L}_{11}^\mathsf{T}\boldsymbol{y}^1 + \boldsymbol{L}_{11}\boldsymbol{L}_{21}^\mathsf{T}\boldsymbol{y}^2 = \boldsymbol{A}_1^\mathsf{T}\boldsymbol{b} =: \boldsymbol{c}^1$$

und haben die Lösungsmenge

$$\boldsymbol{L} = \boldsymbol{L}(\boldsymbol{A}_\alpha, \boldsymbol{b}) = \{\boldsymbol{x}_\alpha = \boldsymbol{P}^\mathsf{T}\boldsymbol{y} : \boldsymbol{y}^1 = \boldsymbol{L}_{11}^{-\mathsf{T}}\boldsymbol{L}_{11}^{-1}\boldsymbol{c}^1 - \boldsymbol{L}_{11}^{-\mathsf{T}}\boldsymbol{L}_{21}^\mathsf{T}\boldsymbol{y}^2,\ \boldsymbol{y}^2 \in \mathbf{R}^{n-p} \text{ beliebig}\}.$$

(iii) Die Normallösung ist durch dasjenige $\boldsymbol{y}^2$ gegeben, das

$$\|\boldsymbol{y}\|^2 = \|\boldsymbol{y}^1\|^2 + \|\boldsymbol{y}^2\|^2 = \left\| \begin{bmatrix} \boldsymbol{L}_{11}^{-\mathsf{T}}\boldsymbol{L}_{11}^{-1}\boldsymbol{c}^1 \\ \hline \boldsymbol{o} \end{bmatrix} - \begin{bmatrix} \boldsymbol{L}_{11}^{-\mathsf{T}}\boldsymbol{L}_{21}^\mathsf{T} \\ \hline \boldsymbol{I} \end{bmatrix} \boldsymbol{y}^2 \right\|^2 \tag{53}$$

minimiert und das sich als eindeutige Lösung der zu (53) gehörenden Normalgleichungen

$$\{(\boldsymbol{L}_{21}\boldsymbol{L}_{11}^{-1})\,(\boldsymbol{L}_{21}\boldsymbol{L}_{11}^{-1})^\mathsf{T} + \boldsymbol{I}\}\,\boldsymbol{y}^2 = \boldsymbol{L}_{21}\boldsymbol{L}_{11}^{-1}\boldsymbol{L}_{11}^{-\mathsf{T}}\boldsymbol{L}_{11}^{-1}\boldsymbol{c}^1$$

ergibt.

11.3. Kontinuierliche Regularisierung

Wie bisher sei $\boldsymbol{A}^* \in \mathbf{R}^{m,n}$ eine rangdefiziente Matrix mit $r = \operatorname{rang}(\boldsymbol{A}^*) < n$, die im Sinne von (11.2.2) durch $\boldsymbol{A} \in \mathfrak{R}^{m,n}$ repräsentiert wird. Die kontinuierliche Regularisierung als eine zweite Regularisierungstechnik besteht dann darin, die Originalaufgabe

$$\|\boldsymbol{b} - \boldsymbol{A}\boldsymbol{x}\|^2 \to \text{Minimum!},$$

$\boldsymbol{b} \in \mathfrak{R}^m$, durch die regularisierte Aufgabe

$$\|\boldsymbol{b} - \boldsymbol{A}\boldsymbol{x}\|^2 + \alpha^2 \|\boldsymbol{x}\|^2 = \left\| \begin{pmatrix} \boldsymbol{b} \\ \boldsymbol{o} \end{pmatrix} - \begin{pmatrix} \boldsymbol{A} \\ \alpha\boldsymbol{I} \end{pmatrix} \boldsymbol{x} \right\|^2 \to \text{Minimum!} \tag{1}$$

zu ersetzen, d. h., statt $\boldsymbol{A}\boldsymbol{x} \cong \boldsymbol{b}$ wird das Problem

$$\boldsymbol{A}_\alpha \boldsymbol{x} \cong \boldsymbol{b}_\alpha \quad \text{mit} \quad \boldsymbol{A}_\alpha := \begin{pmatrix} \boldsymbol{A} \\ \alpha\boldsymbol{I} \end{pmatrix} \in \mathbf{R}^{m+n,n}, \qquad \boldsymbol{b}_\alpha := \begin{pmatrix} \boldsymbol{b} \\ \boldsymbol{o} \end{pmatrix} \in \mathbf{R}^{m+n} \tag{2}$$

gelöst. Die zugehörigen Normalgleichungen $\boldsymbol{A}_\alpha^\mathsf{T}\boldsymbol{A}_\alpha \boldsymbol{x} = \boldsymbol{A}_\alpha^\mathsf{T}\boldsymbol{b}_\alpha$ haben die Form

$$\boldsymbol{M}_\alpha \boldsymbol{x} := (\boldsymbol{A}^\mathsf{T}\boldsymbol{A} + \alpha^2\boldsymbol{I})\,\boldsymbol{x} = \boldsymbol{A}^\mathsf{T}\boldsymbol{b}. \tag{3}$$

Aus der Gültigkeit von

$$\boldsymbol{x}^\mathsf{T}\boldsymbol{M}_\alpha \boldsymbol{x} = \|\boldsymbol{A}_\alpha \boldsymbol{x}\|^2 = \|\boldsymbol{A}\boldsymbol{x}\|^2 + \alpha^2 \|\boldsymbol{x}\|^2 \geqq \alpha^2 \|\boldsymbol{x}\|^2 \quad \text{für alle} \quad \boldsymbol{x} \in \mathbf{R}^n$$

folgt

$$\sqrt{\lambda_i(\boldsymbol{M}_\alpha)} = \sigma_i(\boldsymbol{A}_\alpha) \geqq \alpha \qquad (i = 1, \ldots, n). \tag{4}$$

Insbesondere ist $\boldsymbol{A}_\alpha$ im Fall $\alpha > 0$ spaltenregulär, so daß $\boldsymbol{A}_\alpha^+$ durch

$$\boldsymbol{A}_\alpha^+ = (\boldsymbol{A}_\alpha^\mathsf{T}\boldsymbol{A}_\alpha)^{-1}\,\boldsymbol{A}_\alpha^\mathsf{T} = \boldsymbol{M}_\alpha^{-1}(\boldsymbol{A}^\mathsf{T} \mid \alpha\boldsymbol{I}) =: (\boldsymbol{S}_\alpha \mid \boldsymbol{T}_\alpha) \tag{5}$$

gegeben ist und die äquivalenten Probleme (1) bis (3) die eindeutige Lösung

$$\boldsymbol{x}_\alpha = \boldsymbol{A}_\alpha^+\boldsymbol{b}_\alpha = \boldsymbol{S}_\alpha\boldsymbol{b} + \boldsymbol{T}_\alpha\boldsymbol{o} = (\boldsymbol{A}^\mathsf{T}\boldsymbol{A} + \alpha^2\boldsymbol{I})^{-1}\,\boldsymbol{A}^\mathsf{T}\boldsymbol{b} \tag{6}$$

besitzen.

A. *Kontinuierliche Regularisierung mittels SVD*

Die Berechnung und Analyse der regularisierten Größen $\boldsymbol{S}_\alpha$ und $\boldsymbol{x}_\alpha$ sowie des zugehörigen Residuums $\boldsymbol{r}_\alpha = \boldsymbol{b} - \boldsymbol{A}\boldsymbol{x}_\alpha$ ist am leichtesten unter Verwendung der SVD möglich. Der Einfachheit halber ignorieren wir dabei zunächst die bei Computerrechnung auftretenden Rundungsfehler und betrachten nur den Einfluß der Regularisierung. Es sei dazu

$$\boldsymbol{A} = \boldsymbol{U}\boldsymbol{\Sigma}\boldsymbol{V}^\mathsf{T}, \qquad \boldsymbol{\Sigma} = \operatorname{diag}(\sigma_i), \qquad \sigma_1 \geqq \cdots \geqq \sigma_p > 0 = \sigma_{p+1} = \cdots = \sigma_n \tag{7}$$

die exakte SVD einer beliebigen Matrix vom Rang p, die gemäß $\boldsymbol{A} = \boldsymbol{A}^* + \delta\boldsymbol{A}$ durch eine Störung $\delta\boldsymbol{A}$ aus $\boldsymbol{A}^*$ entsteht und nicht notwendig die oben eingeführte Computerdarstellung von $\boldsymbol{A}^*$ zu sein braucht. Später werden wir insbesondere $\delta\boldsymbol{A} = \boldsymbol{o}$ setzen, so daß (7) die exakte SVD von $\boldsymbol{A}^*$ ist, und $\boldsymbol{x}_\alpha$ wird dann die exakte, zu $\boldsymbol{A}^*$ gehörende regularisierte Lösung $\boldsymbol{x}_\alpha^*$.

Einsetzen von (7) in (5) liefert für die regularisierte Pseudoinverse

$$\boldsymbol{S}_\alpha = \boldsymbol{V}[(\boldsymbol{\Sigma}^\mathsf{T}\boldsymbol{\Sigma} + \alpha^2\boldsymbol{I})^{-1}\,\boldsymbol{\Sigma}^\mathsf{T}]\,\boldsymbol{U}^\mathsf{T} =: \boldsymbol{V}\boldsymbol{\Omega}_\alpha\boldsymbol{U}^\mathsf{T} \tag{8}$$

mit

$$\boldsymbol{\Omega}_\alpha = \operatorname{diag}\left(\frac{\sigma_1}{\sigma_1^2 + \alpha^2}, \ldots, \frac{\sigma_p}{\sigma_p^2 + \alpha^2}, 0, \ldots, 0\right) \in \mathbf{R}^{n,m}. \tag{9}$$

Unter Verwendung dieser Darstellung ergibt sich der folgende Algorithmus zur Berechnung von $\boldsymbol{x}_\alpha$, vgl. 11.2.1:

11.3.1. *Lösung von $\boldsymbol{A}\boldsymbol{x} \cong \boldsymbol{b}$ mittels kontinuierlicher Regularisierung der SVD.*

S1 (Festlegung der regularisierten Pseudoinversen $\boldsymbol{S}_\alpha$):

S1.1: Berechne Singulärwertzerlegung $\boldsymbol{A} = \boldsymbol{U}\boldsymbol{\Sigma}\boldsymbol{V}^\mathsf{T}$

S1.2: Wähle Regularisierungsniveau $\alpha > 0$, definiere $\boldsymbol{S}_\alpha$, $\boldsymbol{\Omega}_\alpha$ gemäß (8), (9).

S2 (Berechnung der zu $\boldsymbol{b}$ gehörenden regularisierten Lösung $\boldsymbol{x}_\alpha$, des Residuums $\boldsymbol{r}_\alpha = \boldsymbol{b} - \boldsymbol{A}\boldsymbol{x}_\alpha$ und der Normen $n\boldsymbol{x}_\alpha = \|\boldsymbol{x}_\alpha\|$ sowie $n\boldsymbol{r}_\alpha = \|\boldsymbol{r}_\alpha\|$):

S2.1: Berechne $\beta := \boldsymbol{U}^\mathsf{T}\boldsymbol{b}$

S2.2: Bestimme $\varrho_\alpha := (\boldsymbol{I} - \boldsymbol{\Sigma}\boldsymbol{\Omega}_\alpha)\,\beta = \left(\frac{\alpha^2\beta_1}{\sigma_1^2 + \alpha^2}, \ldots, \frac{\alpha^2\beta_p}{\sigma_p^2 + \alpha^2}, \beta_{p+1}, \ldots, \beta_m\right)^\mathsf{T}$,

berechne $n\boldsymbol{r}_\alpha := \|\varrho_\alpha\|$ und $\boldsymbol{r}_\alpha := \boldsymbol{U}\varrho_\alpha$

S 2.3: Bestimme $\boldsymbol{\xi}_\alpha := \boldsymbol{\Omega}_\alpha \boldsymbol{\beta} = \left(\frac{\sigma_1\beta_1}{\sigma_1^2+\alpha^2}, \ldots, \frac{\sigma_p\beta_p}{\sigma_p^2+\alpha^2}, 0, \ldots, 0\right)^T$,
berechne $n\boldsymbol{x}_\alpha := \|\boldsymbol{\xi}_\alpha\|$ und $\boldsymbol{x}_\alpha := \boldsymbol{V}\boldsymbol{\xi}_\alpha$

Aus (7) bis (9) folgt

$$\boldsymbol{A}^+ - \boldsymbol{S}_\alpha = \boldsymbol{V}(\boldsymbol{\Sigma}^+ - \boldsymbol{\Omega}_\alpha)\,\boldsymbol{U}^T = \boldsymbol{V}\,\mathrm{diag}\left(\frac{\alpha^2}{\sigma_1(\sigma_1^2+\alpha^2)}, \ldots, \frac{\alpha^2}{\sigma_p(\sigma_p^2+\alpha^2)}, 0, \ldots, 0\right)\boldsymbol{U}^T,$$

wegen $\|\boldsymbol{A}^+\| = 1/\sigma_p$ also

$$\|\boldsymbol{A}^+ - \boldsymbol{S}_\alpha\| = \|\boldsymbol{\Sigma}^+ - \boldsymbol{\Omega}_\alpha\| = \frac{\alpha^2}{\sigma_p(\sigma_p^2+\alpha^2)} \leqq \frac{\alpha^2}{\sigma_p^3} = \alpha^2\,\|\boldsymbol{A}^+\|^3. \tag{10}$$

Die Abschätzung von $\boldsymbol{x}_\alpha$ gegen die Normallösung $\boldsymbol{x}^N = \boldsymbol{A}^+\boldsymbol{b} = \boldsymbol{V}\boldsymbol{\Sigma}^+\boldsymbol{\beta}$ führt auf

$$\|\boldsymbol{x}^N - \boldsymbol{x}_\alpha\|^2 = \|(\boldsymbol{\Sigma}^+ - \boldsymbol{\Omega}_\alpha)\,\boldsymbol{\beta}\|^2 = \sum_{j=1}^{p}\left[\frac{\alpha^2\beta_j}{\sigma_j(\sigma_j^2+\alpha^2)}\right]^2 \leqq \left[\frac{\alpha^2}{\sigma_p^2}\right]^2 \sum_{j=1}^{p}\left(\frac{\beta_j}{\sigma_j}\right)^2,$$

d. h.

$$\|\boldsymbol{x}^N - \boldsymbol{x}_\alpha\| \leqq \alpha^2\,\|\boldsymbol{A}^+\|^2\,\|\boldsymbol{x}^N\|. \tag{11}$$

Für den Fehler von $\boldsymbol{r}_\alpha$ gegenüber $\boldsymbol{r}^N := \boldsymbol{b} - \boldsymbol{A}\boldsymbol{x}^N$ ergibt sich analog

$$\|\boldsymbol{r}^N - \boldsymbol{r}_\alpha\|^2 = \|\boldsymbol{\Sigma}(\boldsymbol{\Sigma}^+ - \boldsymbol{\Omega}_\alpha)\,\boldsymbol{\beta}\|^2 = \sum_{j=1}^{p}\left[\frac{\alpha^2\beta_j}{\sigma_j^2+\alpha^2}\right]^2 \leqq \left[\frac{\alpha^2}{\sigma_p}\right]^2 \sum_{j=1}^{p}\left(\frac{\beta_j}{\sigma_j}\right)^2,$$

mithin

$$\|\boldsymbol{r}^N - \boldsymbol{r}_\alpha\| \leqq \alpha^2\,\|\boldsymbol{A}^+\|\,\|\boldsymbol{x}^N\|. \tag{12}$$

Aus den Abschätzungen (10) bis (12) folgt für $\alpha \to 0$ die Gültigkeit von

$$\boldsymbol{S}_\alpha \to \boldsymbol{A}^+, \qquad \boldsymbol{x}_\alpha \to \boldsymbol{x}^N \quad \text{und} \quad \boldsymbol{r}_\alpha \to \boldsymbol{r}^N,$$

d. h., *die exakt berechneten regularisierten Größen approximieren Pseudoinverse, Normallösung und deren Residuum beliebig genau.* Insbesondere können $\boldsymbol{S}_\alpha$, $\boldsymbol{x}_\alpha$ und $\boldsymbol{r}_\alpha$ auch für $\alpha = 0$ durch die entsprechenden Grenzwerte definiert werden. Man beachte ferner, daß der durch die Störung $\begin{pmatrix}\boldsymbol{O}\\ \alpha\boldsymbol{I}\end{pmatrix}$ von $\begin{pmatrix}\boldsymbol{A}\\ \boldsymbol{O}\end{pmatrix}$ bedingte Block

$$\boldsymbol{T}_\alpha = \boldsymbol{M}_\alpha^{-1}\alpha\boldsymbol{I} = \boldsymbol{V}\,\mathrm{diag}\left(\frac{\alpha}{\sigma_1^2+\alpha^2}, \ldots, \frac{\alpha}{\sigma_p^2+\alpha^2}, \frac{1}{\alpha}, \ldots, \frac{1}{\alpha}\right)\boldsymbol{V}^T$$

von $\boldsymbol{A}_\alpha^-$ im Fall $p < n$ für $\alpha \to 0$ unbeschränkt wächst, sich also bösartig auswirkt. Diese Bösartigkeit überträgt sich jedoch nicht auf $\boldsymbol{x}_\alpha$, da der entsprechende Block von $\boldsymbol{b}_\alpha$ nach Konstruktion $\boldsymbol{o}$ ist, vgl. (6).

Für die Normen $\|\boldsymbol{x}_\alpha\| = \|\boldsymbol{\xi}_\alpha\|$ bzw. $\|\boldsymbol{r}_\alpha\| = \|\boldsymbol{\varrho}_\alpha\|$ gilt

$$\|\boldsymbol{x}_\alpha\|^2 = \sum_{j=1}^{p}\left[\frac{\sigma_j\beta_j}{\sigma_j^2+\alpha^2}\right]^2 \leqq \sum_{j=1}^{p}\left(\frac{\beta_j}{\sigma_j}\right)^2 = \|\boldsymbol{x}^N\|^2 \tag{13}$$

bzw.

$$\|\boldsymbol{r}_\alpha\|^2 = \sum_{i=1}^{p} \left[\frac{\alpha^2\beta_i}{\sigma_i^2 + \alpha^2}\right]^2 + \sum_{i=p+1}^{m} (\beta_i)^2 \geqq \sum_{i=p+1}^{m} (\beta_i)^2 = \|\boldsymbol{r}^N\|^2, \tag{14}$$

und ihr Verhalten als Funktion von α wird durch die folgende Aussage charakterisiert:

11.3.2. *Aussage.* Die Normen $\|\boldsymbol{x}_\alpha\|$ bzw. $\|\boldsymbol{r}_\alpha\|$ sind für $0 \leqq \alpha < \infty$ monoton fallende bzw. wachsende Funktionen von α mit

$$\lim_{\alpha \to +0} \|\boldsymbol{x}_\alpha\| = \|\boldsymbol{x}^N\| \geqq \|\boldsymbol{x}_\alpha\| \geqq 0 = \lim_{\alpha \to \infty} \|\boldsymbol{x}_\alpha\|$$

bzw.

$$\lim_{\alpha \to +0} \|\boldsymbol{r}_\alpha\| = \|\boldsymbol{r}^N\| \leqq \|\boldsymbol{r}_\alpha\| \leqq \|\boldsymbol{b}\| = \lim_{\alpha \to \infty} \|\boldsymbol{r}_\alpha\|.$$

Im Fall $\boldsymbol{x}^N = \boldsymbol{A}^+\boldsymbol{b} \neq \boldsymbol{o}$ liegt strenge Monotonie vor, andernfalls sind beide Normen konstant.

Beweis. Differentiation von (13) bzw. (14) liefert

$$\frac{\mathrm{d}}{\mathrm{d}\alpha} \|\boldsymbol{x}_\alpha\|^2 = -4\alpha \sum_{j=1}^{p} \frac{\sigma_j^2\beta_j^2}{(\sigma_j^2 + \alpha^2)^3} \leqq 0 \tag{15}$$

bzw.

$$\frac{\mathrm{d}}{\mathrm{d}\alpha} \|\boldsymbol{r}_\alpha\|^2 = 4\alpha^3 \sum_{j=1}^{p} \frac{\sigma_j^2\beta_j^2}{(\sigma_j^2 + \alpha^2)^3} \geqq 0, \tag{16}$$

und im Fall $\boldsymbol{x}^N \neq \boldsymbol{o}$, d. h. $(\beta_1, \ldots, \beta_p) \neq (0, \ldots, 0)$, gilt strenge Ungleichheit. □

In Abb. 11.3.1 ist ein mögliches Bild der Funktionen $\|\boldsymbol{x}_\alpha\|$, $\|\boldsymbol{r}_\alpha\|$ skizziert worden.

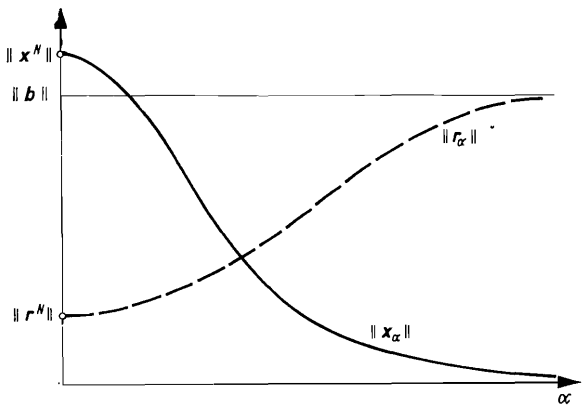


Abb. 11.3.1. $\|\boldsymbol{x}_\alpha\|$ und $\|\boldsymbol{r}_\alpha\|$ als Funktionen von α

Wie im Fall der diskreten Regularisierung bietet es sich an, alle $\{\boldsymbol{x}_\alpha : 0 \leqq \alpha\}$ als mögliche Kandidaten für eine regularisierte Lösung anzusehen und unter diesen einen auf Grund der konkreten Problemstellung besonders geeigneten auszuwählen.

Da die Kandidaten eine einparametrige stetige Schar bilden — deshalb auch der Name *kontinuierliche Regularisierung* —, wird man sich praktisch auf endlich viele Repräsentanten beschränken müssen, etwa auf die Berechnung von $\boldsymbol{x}_\alpha$ für $\alpha = 10^l\nu\,\|\boldsymbol{A}\| = 10^l\nu\sigma_1$, $l = 0, 1, \ldots$, so daß $10^l\nu \leqq 1$ ist. Eine andere Möglichkeit besteht wieder in der Festlegung von α analog zu (11.2.29) bzw. (11.2.30). Der folgende Satz zeigt, daß die im Sinne von (11.2.29) bzw. (11.2.30) optimalen $\boldsymbol{x}_\alpha$ die Nebenbedingungen mit Gleichheit erfüllen und gewisse zusätzliche Extremaleigenschaften besitzen.

11.3.3. *Satz.* Es sei $\boldsymbol{x}^N \neq \boldsymbol{o}$. Dann gilt:

(i) Zu jedem ω mit $0 < \omega \leqq \|\boldsymbol{x}^N\|$ gibt es genau ein $\alpha = \alpha_1(\omega) \geqq 0$ mit $\|\boldsymbol{x}_\alpha\| = \omega$, und $\boldsymbol{x}_\alpha$ ist eindeutige Lösung der Aufgabe

$$\|\boldsymbol{b} - \boldsymbol{A}\boldsymbol{x}\| \to \text{Minimum!} \quad \text{bei} \quad \|\boldsymbol{x}\| \leqq \omega. \tag{17}$$

(ii) Zu jedem τ mit $\|\boldsymbol{r}^N\| \leqq \tau < \|\boldsymbol{b}\|$ gibt es genau ein $\alpha = \alpha_2(\tau) \geqq 0$ mit $\|\boldsymbol{r}_\alpha\| = \tau$, und $\boldsymbol{x}_\alpha$ ist eindeutige Lösung von

$$\|\boldsymbol{x}\| \to \text{Minimum!} \quad \text{bei} \quad \|\boldsymbol{b} - \boldsymbol{A}\boldsymbol{x}\| \leqq \tau. \tag{18}$$

Beweis. Wegen 11.3.2 sind $\|\boldsymbol{x}_\alpha\|$ und $\|\boldsymbol{r}_\alpha\|$ streng monotone Funktionen mit den Wertebereichen $(0, \|\boldsymbol{x}^N\|]$ und $[\|\boldsymbol{r}^N\|, \|\boldsymbol{b}\|)$, so daß α in beiden Fällen eindeutig festgelegt ist.

Zu (i): Es sei $\|\boldsymbol{x}_\alpha\| = \omega$. Da $\boldsymbol{x}_\alpha$ die Aufgabe (1) löst, gilt $\|\boldsymbol{b} - \boldsymbol{A}\boldsymbol{x}\|^2 + \alpha^2\|\boldsymbol{x}\|^2 \geqq \|\boldsymbol{b} - \boldsymbol{A}\boldsymbol{x}_\alpha\|^2 + \alpha^2\|\boldsymbol{x}_\alpha\|^2$, also

$$\|\boldsymbol{b} - \boldsymbol{A}\boldsymbol{x}\|^2 - \|\boldsymbol{b} - \boldsymbol{A}\boldsymbol{x}_\alpha\|^2 \geqq \alpha^2[\|\boldsymbol{x}_\alpha\|^2 - \|\boldsymbol{x}\|^2]$$

für alle $\boldsymbol{x}$, d. h., $\boldsymbol{x}_\alpha$ löst (17). Im Fall $\alpha > 0$ ist $\boldsymbol{x}_\alpha$ die einzige Lösung von (1), so daß für $\boldsymbol{x} \neq \boldsymbol{x}_\alpha$ strenge Ungleichheit gilt und $\boldsymbol{x}_\alpha$ auch die einzige Lösung von (17) ist. Im Fall $\alpha = 0$ ist $\boldsymbol{x}_\alpha = \boldsymbol{x}^N$. Alle weiteren Lösungen müssen denselben Zielfunktionswert wie $\boldsymbol{x}^N$ haben, also Lösungen von $\boldsymbol{A}\boldsymbol{x} \cong \boldsymbol{b}$ sein. Wegen der Eindeutigkeit der Normallösung haben jedoch alle solche $\boldsymbol{x} \neq \boldsymbol{x}^N$ eine größere Norm als $\|\boldsymbol{x}^N\| = \omega$, sind also nicht zulässig.

Zu (ii): Es gelte $\|\boldsymbol{r}_\alpha\| = \tau$. Nach (i) gilt dann $\|\boldsymbol{b} - \boldsymbol{A}\boldsymbol{x}\| > \|\boldsymbol{b} - \boldsymbol{A}\boldsymbol{x}_\alpha\| = \|\boldsymbol{r}_\alpha\|$ für alle $\boldsymbol{x} \neq \boldsymbol{x}_\alpha$ mit $\|\boldsymbol{x}\| \leqq \|\boldsymbol{x}_\alpha\|$, d. h., es gibt kein zulässiges $\boldsymbol{x}$, das einen kleineren Zielfunktionswert als $\boldsymbol{x}_\alpha$ liefert. □

11.3.4. *Bemerkung.* (i) Zur Bestimmung von α kann ein skalares Nullstellenverfahren — etwa das Newton-Verfahren — auf die Gleichungen

$$f(\alpha) := \|\boldsymbol{x}_\alpha\| - \omega = 0 \quad \text{bzw.} \quad g(\alpha) = \|\boldsymbol{r}_\alpha\| - \tau = 0$$

oder i. allg. besser auf die äquivalenten Gleichungen

$$\hat{f}(\alpha) := \frac{1}{\|\boldsymbol{x}_\alpha\|} - \frac{1}{\omega} = 0 \quad \text{bzw.} \quad \hat{g}(\alpha) = \frac{1}{\|\boldsymbol{r}_\alpha\|} - \frac{1}{\tau} = 0$$

angewendet werden. Die Ableitungen lassen sich analog zu (15), (16) leicht berechnen.

(ii) Im Fall $\|\boldsymbol{x}^N\| < \omega$ ist $\boldsymbol{x}^N$ die normkleinste Lösung von (17). Im Fall $\tau < \|\boldsymbol{r}^N\|$ hat (18) keine Lösung, und für $\|\boldsymbol{b}\| \leqq \tau$ ist $\boldsymbol{x} = \boldsymbol{o}$ die einzige Lösung. □

Wir kehren jetzt wieder zur ursprünglichen Fragestellung zurück, die unbekannte, zu $\boldsymbol{A}^*$ gehörende Normallösung $\boldsymbol{x}^* = (\boldsymbol{A}^*)^+\boldsymbol{b}$ durch die zur gestörten Matrix

$\boldsymbol{A} = \boldsymbol{A}^* + \delta\boldsymbol{A}$ gehörende regularisierte Lösung $\boldsymbol{x}_\alpha = \boldsymbol{S}_\alpha \boldsymbol{b}$ zu approximieren. Um den Regularisierungsparameter α möglichst gut festlegen zu können, benötigen wir eine Abschätzung von $\|\boldsymbol{x}^* - \boldsymbol{x}_\alpha\|$ als Funktion von α. Zur Ableitung einer solchen Schranke fügen wir

$$\boldsymbol{x}_\alpha^* = \boldsymbol{S}_\alpha^* \boldsymbol{b} = [(\boldsymbol{A}^*)^\mathsf{T} \boldsymbol{A}^* + \alpha^2 \boldsymbol{I}]^{-1} (\boldsymbol{A}^*)^\mathsf{T} \boldsymbol{b}$$

als Hilfspunkt ein und schätzen gemäß

$$\|\boldsymbol{x}^* - \boldsymbol{x}_\alpha\| \leqq \|\boldsymbol{x}^* - \boldsymbol{x}_\alpha^*\| + \|\boldsymbol{x}_\alpha^* - \boldsymbol{x}_\alpha\|$$

ab. Eine Schranke für den ersten Term ergibt sich aus (11), wenn dort $\boldsymbol{A} = \boldsymbol{A}^*$ gesetzt wird. Zur Abschätzung von $\boldsymbol{x}_\alpha^* - \boldsymbol{x}_\alpha = (\boldsymbol{S}_\alpha^* - \boldsymbol{S}_\alpha)\,\boldsymbol{b}$ stellen wir $\boldsymbol{S}_\alpha^* - \boldsymbol{S}_\alpha$ für $\alpha > 0$ in der Form

$$\begin{aligned}
\boldsymbol{M}_\alpha(\boldsymbol{S}_\alpha^* - \boldsymbol{S}_\alpha) &= \boldsymbol{M}_\alpha \boldsymbol{S}_\alpha^* - \boldsymbol{A}^\mathsf{T} = (\boldsymbol{A}^\mathsf{T}\boldsymbol{A} + \alpha^2 \boldsymbol{I})\,\boldsymbol{S}_\alpha^* - \boldsymbol{A}^\mathsf{T} \\
&= -\boldsymbol{A}^\mathsf{T}(\boldsymbol{I} - \boldsymbol{A}\boldsymbol{S}_\alpha^*) + \alpha^2 \boldsymbol{S}_\alpha^* \\
&= -\boldsymbol{A}^\mathsf{T}(\boldsymbol{I} - \boldsymbol{A}\boldsymbol{S}_\alpha^*) + (\boldsymbol{A}^*)^\mathsf{T}(\boldsymbol{I} - \boldsymbol{A}^*\boldsymbol{S}_\alpha^*) \\
&= -\boldsymbol{A}^\mathsf{T}(\boldsymbol{A}^* - \boldsymbol{A})\,\boldsymbol{S}_\alpha^* + (\boldsymbol{A}^* - \boldsymbol{A})^\mathsf{T}(\boldsymbol{I} - \boldsymbol{A}^*\boldsymbol{S}_\alpha^*),
\end{aligned}$$

also

$$\boldsymbol{S}_\alpha^* - \boldsymbol{S}_\alpha = -\boldsymbol{S}_\alpha(\boldsymbol{A}^* - \boldsymbol{A})\,\boldsymbol{S}_\alpha^* + \boldsymbol{M}_\alpha^{-1}(\boldsymbol{A}^* - \boldsymbol{A})^\mathsf{T}(\boldsymbol{I} - \boldsymbol{A}^*\boldsymbol{S}_\alpha^*) \tag{19}$$

dar. Anwendung auf $\boldsymbol{b}$ liefert mit $\boldsymbol{r}_\alpha^* := \boldsymbol{b} - \boldsymbol{A}^*\boldsymbol{x}_\alpha^*$

$$\boldsymbol{x}_\alpha^* - \boldsymbol{x}_\alpha = -\boldsymbol{S}_\alpha(\boldsymbol{A}^* - \boldsymbol{A})\,\boldsymbol{x}_\alpha^* + \boldsymbol{M}_\alpha^{-1}(\boldsymbol{A}^* - \boldsymbol{A})^\mathsf{T}\boldsymbol{r}_\alpha^*. \tag{20}$$

Es gelte nun

$$\boldsymbol{A} = \boldsymbol{A}^* + \delta\boldsymbol{A}_D + \delta\boldsymbol{A}_M, \qquad \|\boldsymbol{A}^* - \boldsymbol{A}\| = \|\delta\boldsymbol{A}_D + \delta\boldsymbol{A}_M\| = \varepsilon, \tag{21}$$

vgl. (11.2.2). Dann ergibt sich aus (20)

$$\begin{aligned}
\|\boldsymbol{x}^* - \boldsymbol{x}_\alpha\| &\leqq \|\boldsymbol{x}^* - \boldsymbol{x}_\alpha^*\| + \|\boldsymbol{x}_\alpha^* - \boldsymbol{x}_\alpha\| \\
&\leqq \|\boldsymbol{x}^* - \boldsymbol{x}_\alpha^*\| + \varepsilon[\|\boldsymbol{S}_\alpha\|\,\|\boldsymbol{x}_\alpha^*\| + \|\boldsymbol{M}_\alpha^{-1}\|\,\|\boldsymbol{r}_\alpha^*\|],
\end{aligned}$$

wegen $\|\boldsymbol{r}_\alpha^*\| \leqq \|\boldsymbol{r}^*\| + \|\boldsymbol{r}_\alpha^* - \boldsymbol{r}^*\|$ und der auch für die aus $\boldsymbol{A}^*$ berechneten Größen gültigen Ungleichungen (11) bis (13) also

$$\|\boldsymbol{x}^* - \boldsymbol{x}_\alpha\| \leqq \{\alpha^2\,\|(\boldsymbol{A}^*)^+\|^2 + \varepsilon[\|\boldsymbol{S}_\alpha\| + \|\boldsymbol{M}_\alpha^{-1}\|\,(\|\boldsymbol{r}^*\|/\|\boldsymbol{x}^*\| + \alpha^2\|(\boldsymbol{A}^*)^+\|)]\}\,\|\boldsymbol{x}^*\|. \tag{22}$$

Aus (4) folgt nun $\|\boldsymbol{M}_\alpha^{-1}\| \leqq 1/\alpha^2$. Zur Abschätzung von $\boldsymbol{S}_\alpha$ beachten wir (8), (9) und erhalten

$$\|\boldsymbol{S}_\alpha\| = \|\boldsymbol{\Omega}_\alpha\| = \max\left\{\frac{\sigma}{\sigma^2 + \alpha^2} : \sigma = \sigma_i\right\} \leqq \frac{1}{2\alpha} =: \gamma(\alpha), \tag{23}$$

denn die Funktion $\varphi_\alpha(\sigma) := \sigma/(\sigma^2 + \alpha^2)$ hat für $0 \leqq \sigma$ genau ein Maximum an der Stelle $\sigma = \alpha$, und der maximale Funktionswert ist $\varphi_\alpha(\alpha) = 1/(2\alpha)$. Unter Verwendung

der Abkürzungen $\mu := \|(\boldsymbol{A}^*)^+\|$, $\varrho := \|\boldsymbol{r}^*\|/\|\boldsymbol{x}^*\|$ kann dann in (22) weiter gemäß

$$\|\boldsymbol{x}^* - \boldsymbol{x}_\alpha\| \leqq \left\{\alpha^2\mu^2 + \varepsilon\left[\frac{\varrho}{\alpha^2} + \frac{1}{2\alpha} + \mu\right]\right\}\|\boldsymbol{x}^*\| =: \eta_\varepsilon(\alpha)\,\|\boldsymbol{x}^*\| \tag{24}$$

abgeschätzt werden. Die Schrankenfunktion $\eta_\varepsilon(\alpha)$ hat das in Abb. 11.3.2 skizzierte Bild.

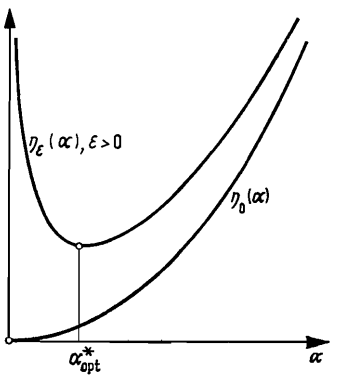


Abb. 11.3.2. $\eta_\varepsilon(\alpha)$ für $\varepsilon = 0$ und $\varepsilon > 0$

Wünschenswert ist daher die Festlegung von $\alpha = \alpha^*_{\text{opt}}$, so daß $\eta_\varepsilon(\alpha)$ minimal wird. Wir unterscheiden dazu zwei Fälle:

Fall 1: $\varrho \neq 0$, d. h. $\boldsymbol{r}^* = \boldsymbol{b} - \boldsymbol{A}^*\boldsymbol{x}^* \neq \boldsymbol{o}$ (*inkonsistenter Fall*)
Um die auf eine Gleichung vierten Grades führende exakte Minimierung zu umgehen, vernachlässigen wir den gegenüber ϱ/α^2 für $\alpha \to 0$ kleinen Term $1/(2\alpha)$ und bestimmen $\alpha = \alpha_{\text{opt}}$ als Minimumstelle von $\alpha^2\mu^2 + \varepsilon[\varrho/\alpha^2 + \mu]$, was auf

$$\alpha_{\text{opt}} = \left(\frac{\varepsilon\varrho}{\mu^2}\right)^{1/4} \quad \text{mit} \quad \eta_\varepsilon(\alpha_{\text{opt}}) = \varepsilon^{1/2}(2\mu\varrho^{1/2})\left(1 + \frac{\varepsilon^{1/4}}{4\varrho^{3/4}\mu^{1/2}}\right) \tag{25}$$

führt. Durch elementare Überlegungen überzeugt man sich davon, daß die Ordnung $O(\varepsilon^{1/2})$ auch bei der Wahl von $\alpha = \alpha^*_{\text{opt}}$ als exakter Minimumstelle von η_ε nicht verbessert werden kann.

Fall 2: $\varrho = 0$, d. h. $\boldsymbol{r}^* = \boldsymbol{b} - \boldsymbol{A}^*\boldsymbol{x}^* = \boldsymbol{o}$ (*konsistenter Fall*)
Hier ergibt sich α als Minimumstelle von $\alpha^2\mu^2 + \varepsilon[1/(2\alpha) + \mu]$ zu

$$\alpha_{\text{opt}} = \alpha^*_{\text{opt}} = \left(\frac{\varepsilon}{4\mu^2}\right)^{1/3} \quad \text{mit} \quad \eta_\varepsilon(\alpha_{\text{opt}}) = \varepsilon^{2/3}\left(\frac{3\mu^{2/3}}{4^{2/3}}\right)\left(1 + \frac{\varepsilon^{1/3}4^{2/3}\mu^{1/3}}{3}\right), \tag{26}$$

d. h., es kann eine Fehlerordnung $O(\varepsilon^{2/3})$ garantiert werden.

Bei feinerer Abschätzung von $\boldsymbol{S}_\alpha$ als durch $\gamma(\alpha)$ gemäß (23) läßt sich im Fall 2 sogar die Fehlerordnung $O(\varepsilon)$ erreichen, sofern das Fehlerniveau ε im Sinne von

$$2\varepsilon < \sigma_r^* = 1/\|(\boldsymbol{A}^*)^+\| = 1/\mu \tag{27}$$

genügend klein ist. Wir erinnern daran, daß (27) gerade die für eine korrekte α-Wahl bei diskreter Regularisierung mittels SVD geforderte Bedingung (11.2.17) ist. Aus

(27) folgt nämlich wegen $|\sigma_i^* - \sigma_i| \leqq \varepsilon$, also $\sigma_i^* - \varepsilon \leqq \sigma_i \leqq \sigma_i^* + \varepsilon$ $(i = 1, \ldots, n)$ die Ungleichung

$$0 \leqq \sigma_n \leqq \cdots \leqq \sigma_{r+1} \leqq \varepsilon < \sigma_r^* - \varepsilon \leqq \sigma_r \leqq \cdots \leqq \sigma_1 . \tag{28}$$

Die Singulärwerte von $\boldsymbol{A}$ liegen also entweder links von ε oder rechts von $\sigma_r^* - \varepsilon$, d. h., zwischen ε und $\sigma_r^* - \varepsilon$ tritt eine sog. Spektrumslücke auf. Die Schranke (23) kann dann gemäß

$$\|\boldsymbol{S}_\alpha\| \leqq \max \left\{ \frac{\sigma}{\sigma^2 + \alpha^2} : 0 \leqq \sigma \leqq \varepsilon \text{ und } \sigma_r^* - \varepsilon \leqq \sigma \right\} =: \gamma_\varepsilon(\alpha) \leqq \gamma(\alpha) = \frac{1}{2\alpha} \tag{29}$$

mit

$$\gamma_\varepsilon(\alpha) = \begin{cases} \dfrac{1}{2\alpha} & \text{für} \quad 0 < \alpha \leqq \varepsilon \text{ und } \sigma_r^* - \varepsilon \leqq \alpha , \\ \max \{\varphi_\alpha(\varepsilon), \varphi_\alpha(\sigma_r^* - \varepsilon)\} & \text{für} \quad \varepsilon < \alpha < \sigma_r^* - \varepsilon \end{cases} \tag{30}$$

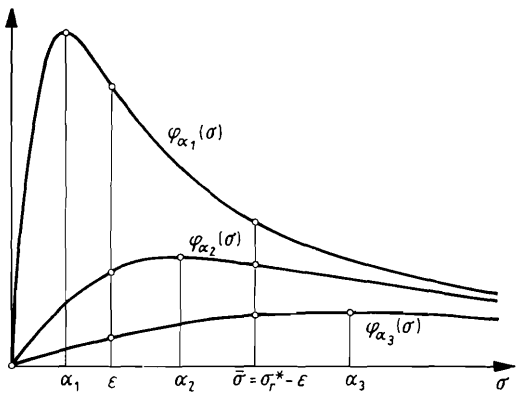


Abb. 11.3.3. $\varphi_{\alpha_i}(\varepsilon)$ für $\alpha_1 < \varepsilon < \alpha_2 < \sigma_r^* - \varepsilon < \alpha_3$ als Funktionen von σ

verfeinert werden, siehe Abb. 11.3.3. Man kann elementar überprüfen, daß $\alpha^2\mu^2 + \varepsilon[\gamma_\varepsilon(\alpha) + \mu]$ für dasjenige α minimal wird, das im Intervall $\varepsilon < \alpha < \sigma_r^* - \varepsilon =: \bar{\sigma}$ liegt und zu gleichen Randwerten $\varphi_\alpha(\varepsilon)$ und $\varphi_\alpha(\bar{\sigma})$ führt. Dies ergibt

$$\alpha_{\text{opt}} = (\varepsilon\bar{\sigma})^{1/2} \quad \text{und} \quad \gamma_\varepsilon(\alpha_{\text{opt}}) = 1/(\bar{\sigma} + \varepsilon) = 1/\sigma_r^* = \mu , \tag{31}$$

also

$$\begin{aligned} \|\boldsymbol{x}^* - \boldsymbol{x}_{\alpha_{\text{opt}}}\| &\leqq \{\alpha_{\text{opt}}^2\mu^2 + \varepsilon[\gamma_\varepsilon(\alpha_{\text{opt}}) + \mu]\} \, \|\boldsymbol{x}^*\| = \{\varepsilon(\bar{\sigma}\mu)\,\mu + \varepsilon[\mu + \mu]\} \, \|\boldsymbol{x}^*\| \\ &\leqq 3\varepsilon\mu \, \|\boldsymbol{x}^*\| , \end{aligned} \tag{32}$$

man beachte $\bar{\sigma}\mu = (\sigma_r^* - \varepsilon)/\sigma_r^* \leqq 1$.

Zusammenfassend erhalten wir das folgende Resultat.

11.3.5. *Aussage.* Unter der Voraussetzung $\|\boldsymbol{A}^* - \boldsymbol{A}\| \leqq \varepsilon$ kann durch geeignete Wahl von $\alpha = \alpha_{\text{opt}} > 0$ für den Fehler von $\boldsymbol{x}_\alpha := \boldsymbol{S}_\alpha \boldsymbol{b}$ gegenüber $\boldsymbol{x}^* := (\boldsymbol{A}^*)^+ \boldsymbol{b}$

die Schranke

$$\|\boldsymbol{x}^* - \boldsymbol{x}_\alpha\|/\|\boldsymbol{x}^*\| \leqq K\varepsilon^\theta(1 + N\varepsilon^{\theta/2}) \tag{33}$$

mit

$$\theta := \begin{cases} 1/2 & \text{für } \boldsymbol{r}^* \neq \boldsymbol{o}, \\ 2/3 & \text{für } \boldsymbol{r}^* = \boldsymbol{o}, \\ 1 & \text{für } \boldsymbol{r}^* = \boldsymbol{o} \text{ und } 2\varepsilon\,\|(\boldsymbol{A}^*)^+\| < 1 \end{cases} \tag{34}$$

garantiert werden, wobei $K > 0$ und $N \geqq 0$ die durch (25), (26) bzw. (32) gegebenen, nur von $\boldsymbol{A}^*$ und $\boldsymbol{b}$ abhängenden Zahlenwerte haben.

11.3.6. *Bemerkung.* (i) Im konsistenten Fall wird bei Gültigkeit von (27) mit $\alpha = \alpha_{\text{opt}}$ nach (32) die Genauigkeit

$$\|\boldsymbol{x}^* - \boldsymbol{x}_\alpha\|/\|\boldsymbol{x}^*\| \leqq 3\varepsilon\,\|(\boldsymbol{A}^*)^+\|$$

garantiert. Das ist praktisch die volle erreichbare Genauigkeit, denn nach 8.2.7 muß selbst bei Ranggleichheit von $\boldsymbol{A}^*$ und $\boldsymbol{A}$ mit einem Fehler

$$\|\boldsymbol{x}^* - \boldsymbol{x}\|/\|\boldsymbol{x}\| \leqq \varepsilon\sqrt{2}\,\|(\boldsymbol{A}^*)^+\|/(1 - \varepsilon\,\|(\boldsymbol{A}^*)^+\|) \leqq 2\sqrt{2}\,\varepsilon\,\|(\boldsymbol{A}^*)^+\|$$

gerechnet werden.

(ii) Im inkonsistenten Fall wird für beliebiges ε eine Genauigkeit $O(\varepsilon^{1/2})$ garantiert. Bei diskreter Regularisierung muß dagegen ε im Sinne von (27) klein sein, allerdings ist die Fehlerordnung dann $O(\varepsilon)$, vgl. 11.2.3(ii). Grob gesprochen läßt sich also für $\boldsymbol{r}^* \neq \boldsymbol{o}$ mittels kontinuierlicher Regularisierung höchstens etwa die Hälfte der gültigen Ziffern erreichen, die mittels diskreter Regularisierung erzielt werden können.

(iii) Bei Computerrechnung ist $\boldsymbol{A}$ in (21) durch $\boldsymbol{A} = \boldsymbol{U}\boldsymbol{\Sigma}\boldsymbol{V}^\mathsf{T} - \delta\boldsymbol{A}_R$ zu ersetzen, vgl. (11.2.3). Das Fehlerniveau ε kann sich dabei maximal um $\|\delta\boldsymbol{A}_R\| \leqq \nu F\,\|\boldsymbol{A}\|$ vergrößern, siehe (11.2.5). Außerdem sind die bei der Berechnung von $\boldsymbol{x}_\alpha$ bzw. $\boldsymbol{r}_\alpha$ aus $\{\boldsymbol{U}, \boldsymbol{\Sigma}, \boldsymbol{V}, \boldsymbol{b}\}$ auftretenden Rundungsfehler mit zu berücksichtigen. Wir verzichten auf eine detaillierte, zu 11.2.2 analoge Angabe der Resultate einer solchen Fehleranalyse und bemerken lediglich, daß sie sich qualitativ nicht wesentlich von denen aus 11.3.5 unterscheiden. Bei Bedarf kann auch ein Fehler $\boldsymbol{b} = \boldsymbol{b}^* + \delta\boldsymbol{b}$ der rechten Seiten mit berücksichtigt werden.

(iv) Falls α nicht nach dem bereits beschriebenen heuristischen Vorgehen durch Beobachtung von $\|\boldsymbol{x}_\alpha\|$ und $\|\boldsymbol{r}_\alpha\|$ für verschiedene α-Werte bzw. durch Vorgabe von $\omega = \|\boldsymbol{x}_\alpha\|$ oder $\tau = \|\boldsymbol{r}_\alpha\|$ bestimmt wird, können die angegebenen Formeln für α_{opt} zur Festlegung von α verwendet werden. Die in den Formeln auftretenden σ_i^* sind dann durch die berechneten σ_i zu ersetzen, wobei $r = p(\varepsilon)$ durch (11.2.10) festgelegt wird; unter der Bedingung (27) ist diese Festlegung korrekt. Für andere Prinzipien zur Festlegung von α sei auf B 11.3 verwiesen. □

Als Mittel zur Regularisierung rangdefizienter oder schlecht konditionierter Quadratmittelprobleme ist daher die diskrete Regularisierung i. allg. vorzuziehen. Es gibt jedoch Aufgaben, die aus anderen Gründen die Lösung kontinuierlicher Regu-

larisierungsaufgaben für einen oder mehrere α-Werte erfordern, vgl. B 11.3. In diesem Fall ist es vom Aufwand her zweckmäßig, mit der Bidiagonalform statt mit der SVD zu arbeiten.

B. *Kontinuierliche Regularisierung mittels der Bidiagonalform*

Es sei

$$\boldsymbol{A} = \boldsymbol{U}_1 \begin{pmatrix} \boldsymbol{B} \\ \boldsymbol{O} \end{pmatrix} \boldsymbol{V}_1^{\mathsf{T}}, \quad \boldsymbol{B} \in \mathbf{R}^{n,n} \text{ bidiagonal}, \tag{35}$$

die nach **11.1.1** berechnete Bidiagonalform von $\boldsymbol{A}$. Mit den Bezeichnungen $\begin{pmatrix} \tilde{\boldsymbol{\beta}} \\ \bar{\boldsymbol{\beta}} \end{pmatrix} := \boldsymbol{U}_1^{\mathsf{T}} \boldsymbol{b}$ und $\boldsymbol{\xi} := \boldsymbol{V}_1^{\mathsf{T}} \boldsymbol{x}$ läßt sich dann (1) in der Form

$$\begin{aligned} \|\boldsymbol{b} - \boldsymbol{A}\boldsymbol{x}\|^2 + \alpha^2 \|\boldsymbol{x}\|^2 &= \left\| \begin{pmatrix} \tilde{\boldsymbol{\beta}} \\ \bar{\boldsymbol{\beta}} \end{pmatrix} - \begin{pmatrix} \boldsymbol{B} \\ \boldsymbol{O} \end{pmatrix} \boldsymbol{\xi} \right\|^2 + \alpha^2 \|\boldsymbol{\xi}\|^2 \\ &= \|\tilde{\boldsymbol{\beta}} - \boldsymbol{B}\boldsymbol{\xi}\|^2 + \alpha^2 \|\boldsymbol{\xi}\|^2 + \|\bar{\boldsymbol{\beta}}\|^2 \end{aligned} \tag{36}$$

schreiben, so daß sich die Berechnung von $\boldsymbol{x}_\alpha = \boldsymbol{V}_1 \boldsymbol{\xi}_\alpha$ auf die von $\boldsymbol{\xi}_\alpha$ als Lösung von

$$\|\tilde{\boldsymbol{\beta}} - \boldsymbol{B}\boldsymbol{\xi}\|^2 + \alpha^2 \|\boldsymbol{\xi}\|^2 = \left\| \begin{pmatrix} \tilde{\boldsymbol{\beta}} \\ \boldsymbol{o} \end{pmatrix} - \begin{pmatrix} \boldsymbol{B} \\ \alpha \boldsymbol{I} \end{pmatrix} \boldsymbol{\xi} \right\|^2 \to \text{Minimum!} \tag{37}$$

reduziert. Zur Lösung von (37) beachten wir, daß der Block $\alpha \boldsymbol{I}$ von links nach rechts durch Multiplikation mit $\sim 2n$ Givens-Drehungen bei Erhalt der Bidiagonalform von $\boldsymbol{B}$ annulliert werden kann; für $n = 4$ ist der erste Hauptschritt durch das folgende Muster gegeben:

$$\begin{bmatrix} \times & \times & & \\ & \times & \times & \\ & & \times & \times \\ & & & \times \\ \hline \times & & & \\ & \times & & \\ & & \times & \\ & & & \times \end{bmatrix} \xrightarrow{G_{15}} \begin{bmatrix} \boxed{\times} & \boxed{\times} & & \\ & \times & \times & \\ & & \times & \times \\ & & & \times \\ \hline 0 & \boxed{\times} & & \\ & \times & & \\ & & \times & \\ & & & \times \end{bmatrix} \xrightarrow{G_{56}} \begin{bmatrix} \times & \times & & \\ & \times & \times & \\ & & \times & \times \\ & & & \times \\ \hline 0 & 0 & & \\ & \boxed{\times} & & \\ & & \times & \\ & & & \times \end{bmatrix}.$$

Wie üblich sind dabei die neu zu berechnenden Elemente eingerahmt worden. Mit dieser Technik ergibt sich das nachstehend beschriebene Verfahren.

11.3.7. *Lösung von $\boldsymbol{A}\boldsymbol{x} \cong \boldsymbol{b}$ mittels kontinuierlicher Regularisierung der Bidiagonalform.*

S1 (Festlegung der regularisierten Bidiagonalform):

S1.1: Berechne Bidiagonalform (35) von $\boldsymbol{A}$ nach **11.1.1**

S1.2: Wähle Regularisierungsniveau $\alpha > 0$

S2 (Berechnung der zu $\boldsymbol{b}$ gehörenden Größen $\boldsymbol{x}_\alpha$, $\boldsymbol{r}_\alpha = \boldsymbol{b} - \boldsymbol{A}\boldsymbol{x}_\alpha$ und deren Normen $n\boldsymbol{x}_\alpha = \|\boldsymbol{x}_\alpha\|$, $n\boldsymbol{r}_\alpha = \|\boldsymbol{r}_\alpha\|$):

S2.1: Berechne

$$\begin{pmatrix}\tilde{\boldsymbol{\beta}}\\ \bar{\boldsymbol{\beta}}\end{pmatrix} := \begin{pmatrix}\tilde{\boldsymbol{\beta}}\\ \bar{\boldsymbol{\beta}}\end{pmatrix} := \boldsymbol{U}_1^{\mathsf{T}}\boldsymbol{b}$$

S2.2: Bestimme $\bar{\boldsymbol{Q}}^{\mathsf{T}} := \boldsymbol{G}_n \cdots \boldsymbol{G}_2\boldsymbol{G}_1$ mit

$$\boldsymbol{G}_i := \boldsymbol{G}_{n+i,\,n+i+1}\boldsymbol{G}_{i,\,n+i} \quad (i = 1, \ldots, n-1), \qquad \boldsymbol{G}_n := \boldsymbol{G}_{n,\,2n}$$

als Produkt von $2n - 1$ Givens-Drehungen so, daß

$$\bar{\boldsymbol{Q}}^{\mathsf{T}}\begin{pmatrix}\boldsymbol{B}\\ \alpha\boldsymbol{I}\end{pmatrix} = \begin{pmatrix}\bar{\boldsymbol{B}}\\ \boldsymbol{O}\end{pmatrix}$$

mit einer Bidiagonalmatrix $\bar{\boldsymbol{B}}$ gilt.

S2.3: Berechne $\begin{pmatrix}\tilde{\gamma}\\ \hat{\gamma}\end{pmatrix} := \bar{\boldsymbol{Q}}^{\mathsf{T}}\begin{pmatrix}\tilde{\boldsymbol{\beta}}\\ \boldsymbol{o}\end{pmatrix}$, bestimme $\boldsymbol{\xi}_\alpha$ als Lösung des Bidiagonalsystems $\bar{\boldsymbol{B}}\boldsymbol{\xi}_\alpha = \tilde{\gamma}$, setze $n\boldsymbol{x}_\alpha := \|\boldsymbol{\xi}_\alpha\|$ und $\boldsymbol{x}_\alpha := \boldsymbol{V}_1\boldsymbol{\xi}_\alpha$

S2.4: Berechne $\begin{pmatrix}\tilde{\varrho}\\ \hat{\varrho}\end{pmatrix} := \bar{\boldsymbol{Q}}\begin{pmatrix}\boldsymbol{o}\\ \hat{\gamma}\end{pmatrix}$, setze $\boldsymbol{\varrho}_\alpha := \begin{pmatrix}\tilde{\varrho}\\ \bar{\boldsymbol{\beta}}\end{pmatrix}$, $n\boldsymbol{r}_\alpha := \|\boldsymbol{\varrho}_\alpha\|$, $\boldsymbol{r}_\alpha := \boldsymbol{U}_1\boldsymbol{\varrho}_\alpha$

11.3.8. *Bemerkungen.* (i) Die Matrizen $\boldsymbol{U}_1, \boldsymbol{V}_1$ können in Produktform auf dem Platz von $\boldsymbol{A}$ gespeichert werden, vgl. 11.1. Das Problem der effektiven Berechnung und Darstellung der in der SVD vorkommenden Faktoren $\boldsymbol{U}, \boldsymbol{V}$ wird daher umgangen. Für $m \gg n$ ist der Aufwand zur Berechnung von $\{\boldsymbol{U}_1, \boldsymbol{B}, \boldsymbol{V}_1\}$ etwa $\sim mn^2$, entspricht also dem einer $\boldsymbol{QR}$-Faktorisierung von $\boldsymbol{A}$. Die Berechnung von $\boldsymbol{x}_\alpha$ bzw. $\boldsymbol{r}_\alpha$ aus $\{\boldsymbol{U}_1, \boldsymbol{B}, \boldsymbol{V}_1, \boldsymbol{\beta} = \boldsymbol{U}_1^{\mathsf{T}}\boldsymbol{b}\}$ erfordert $\sim K_1n^2$ bzw. $\sim K_2mn$ Operationen, die Berechnung allein von $\|\boldsymbol{x}_\alpha\|$ und $\|\boldsymbol{r}_\alpha\|$ sogar nur $\sim K_3n$ Operationen.

(ii) Statt der Bidiagonalform könnte auch die $\boldsymbol{QR}$-Faktorisierung von $\boldsymbol{A}$ mit Spaltenvertauschungen verwendet werden, siehe Ü 11.3.3. Der Faktorisierungsaufwand ist für $m \gg n$ etwa derselbe wie in 11.3.7, allerdings erfordert die Berechnung von $\boldsymbol{x}_\alpha$ bzw. $\boldsymbol{r}_\alpha$ aus $\{\boldsymbol{Q}, \boldsymbol{R}, \boldsymbol{P}, \boldsymbol{c} = \boldsymbol{Q}^{\mathsf{T}}\boldsymbol{b}\}$ dann $\sim K_4n^3$ bzw. $\sim K_5mn^2$ Operationen, so daß diese Version teurer als 11.3.7 ist.

(iii) Bei Berechnung von $\boldsymbol{x}_\alpha$ aus den regularisierten Normalgleichungen (3) ist für jedes α eine erneute Cholesky-Faktorisierung von $\boldsymbol{M}_\alpha$ mit $\sim n^3/6$ opms erforderlich.

(iv) In verschiedenen Anwendungen ist es zweckmäßig, statt (1) das gewichtete regularisierte Problem

$$\|\boldsymbol{b} - \boldsymbol{A}\boldsymbol{x}\|^2 + \alpha^2\|\boldsymbol{D}\boldsymbol{x}\|^2 = \left\|\begin{pmatrix}\boldsymbol{b}\\ \boldsymbol{o}\end{pmatrix} - \begin{pmatrix}\boldsymbol{A}\\ \alpha\boldsymbol{D}\end{pmatrix}\boldsymbol{x}\right\|^2 \to \text{Minimum!} \tag{38}$$

zu verwenden, wobei $\boldsymbol{D} \in \mathbf{R}^{n,n}$ eine reguläre Gewichtsmatrix bezeichnet. Für den praktisch wichtigen Sonderfall $\boldsymbol{D} = \operatorname{diag}(d_i)$, $d_i > 0$, lassen sich 11.3.1 und 11.3.7 fast wörtlich übertragen; für allgemeinere Wichtung siehe Ü 11.3.4. □

Übungsaufgaben

Ü 11.3.1. Man zeige, daß für jedes $\boldsymbol{x} = \boldsymbol{x}_\alpha + \boldsymbol{h}$ die Identität

$$\|\boldsymbol{b} - \boldsymbol{A}\boldsymbol{x}\|^2 - \|\boldsymbol{b} - \boldsymbol{A}\boldsymbol{x}_\alpha\|^2 = \alpha^2[-2\boldsymbol{h}^\mathsf{T}\boldsymbol{x}_\alpha - \|\boldsymbol{h}\|^2] + \alpha^2\|\boldsymbol{h}\|^2 + \|\boldsymbol{A}\boldsymbol{h}\|^2 \tag{39}$$

gilt, und folgere hieraus, daß $\boldsymbol{x}_\alpha$ mit $\|\boldsymbol{x}_\alpha\| = \omega$ die einzige Lösung von (17) ist.
Hinweis: Es gilt $\|\boldsymbol{x}\| \leqq \|\boldsymbol{x}_\alpha\|$ genau dann, wenn $2\boldsymbol{h}^\mathsf{T}\boldsymbol{x}_\alpha + \|\boldsymbol{h}\|^2 \leqq 0$ ist.

Ü 11.3.2. Man zeige, daß sich aus (19) die Abschätzung

$$\|(\boldsymbol{A}^*)^+ - \boldsymbol{S}_\alpha\| \leqq \left\{\alpha^2\mu^2 + \varepsilon\left[\frac{\hat{\varrho}}{\alpha^2} + \frac{1}{2\alpha} + \mu\right]\right\}\|(\boldsymbol{A}^*)^+\|,$$

$$\hat{\varrho} := \frac{\|\boldsymbol{I} - (\boldsymbol{A}^*)\,(\boldsymbol{A}^*)^+\|}{\|(\boldsymbol{A}^*)^+\|} \leqq \frac{1}{\mu} \tag{40}$$

ergibt, und folgere hieraus die Existenz eines $\alpha = \alpha_{\text{opt}}$ mit

$$\|(\boldsymbol{A}^*)^+ - \boldsymbol{S}_\alpha\|/\|(\boldsymbol{A}^*)^+\| \leqq K\varepsilon^{1/2}(1 + N\varepsilon^{1/4}). \tag{41}$$

Man vergleiche die Schranke (41) mit der bei diskreter Regularisierung unter der Voraussetzung (27) erreichbaren Genauigkeit, siehe 11.2.3 (ii).

Ü 11.3.3. Man überlege sich, daß $\boldsymbol{x}_\alpha$, $\boldsymbol{r}_\alpha$ unter Verwendung der $\boldsymbol{QR}$-Faktorisierung mit Spaltenvertauschungen wie folgt berechnet werden können:

S 1: Berechne Faktorisierung $\boldsymbol{A}\boldsymbol{P}^\mathsf{T} = \boldsymbol{Q}\boldsymbol{R} = \boldsymbol{Q}\begin{pmatrix}\tilde{\boldsymbol{R}}\\ \hline \boldsymbol{O}\end{pmatrix}$ analog zu 11.2.B:

S 2.1: Berechne

$$\boldsymbol{c} := \begin{pmatrix}\tilde{\boldsymbol{c}}\\ \hline \bar{\boldsymbol{c}}\end{pmatrix} := \boldsymbol{Q}^\mathsf{T}\boldsymbol{b}.$$

Dann gilt $\|\boldsymbol{b} - \boldsymbol{A}\boldsymbol{x}\|^2 + \alpha^2\|\boldsymbol{x}\|^2 = \|\tilde{\boldsymbol{c}} - \tilde{\boldsymbol{R}}\boldsymbol{y}\|^2 + \alpha^2\|\boldsymbol{y}\|^2 + \|\bar{\boldsymbol{c}}\|^2$ mit $\boldsymbol{y} := \boldsymbol{P}\boldsymbol{x}$.

S 2.2: Bestimme $\bar{\boldsymbol{Q}}^\mathsf{T} := \boldsymbol{G}_1 \cdots \boldsymbol{G}_{n-1}\boldsymbol{G}_n$ mit

$$\boldsymbol{G}_i := \boldsymbol{G}_{n,n+i} \cdots \boldsymbol{G}_{i+1,n+i}\boldsymbol{G}_{i,n+i} \qquad (i = n, \ldots, 1)$$

als Produkt von $n(n+1)/2$ Givens-Drehungen so, daß

$$\bar{\boldsymbol{Q}}^\mathsf{T}\begin{pmatrix}\tilde{\boldsymbol{R}}\\ \hline \alpha\boldsymbol{I}\end{pmatrix} = \begin{pmatrix}\bar{\boldsymbol{R}}\\ \hline \boldsymbol{O}\end{pmatrix}$$

mit einer oberen Dreiecksmatrix $\bar{\boldsymbol{R}}$ gilt. Berechne $\begin{pmatrix}\tilde{\boldsymbol{f}}\\ \hline \hat{\boldsymbol{f}}\end{pmatrix} := \bar{\boldsymbol{Q}}^\mathsf{T}\begin{pmatrix}\tilde{\boldsymbol{c}}\\ \hline \boldsymbol{o}\end{pmatrix}$. Dann gilt

$$\|\tilde{\boldsymbol{c}} - \tilde{\boldsymbol{R}}\boldsymbol{y}\|^2 + \alpha^2\|\boldsymbol{y}\|^2 = \|\tilde{\boldsymbol{f}} - \bar{\boldsymbol{R}}\boldsymbol{y}\|^2 + \|\hat{\boldsymbol{f}}\|^2.$$

S 2.3: Berechne $\boldsymbol{y}$ als Lösung des Dreieckssystems $\bar{\boldsymbol{R}}\boldsymbol{y} = \tilde{\boldsymbol{f}}$, setze $\boldsymbol{x}_\alpha := \boldsymbol{P}^\mathsf{T}\boldsymbol{y}$.

S 2.4: Berechne

$$\begin{pmatrix}\tilde{\boldsymbol{d}}\\ \hline \hat{\boldsymbol{d}}\end{pmatrix} := \bar{\boldsymbol{Q}}\begin{pmatrix}\boldsymbol{o}\\ \hline \hat{\boldsymbol{f}}\end{pmatrix}, \quad \text{setze } \boldsymbol{d} := \begin{pmatrix}\tilde{\boldsymbol{d}}\\ \hline \bar{\boldsymbol{c}}\end{pmatrix}.$$

Dann gilt $\boldsymbol{d} = \boldsymbol{c} - \boldsymbol{R}\boldsymbol{y} = \boldsymbol{Q}^\mathsf{T}(\boldsymbol{b} - \boldsymbol{A}\boldsymbol{x}_\alpha)$. Bestimme $n\boldsymbol{r}_\alpha := \|\boldsymbol{d}\|$ und $\boldsymbol{r}_\alpha := \boldsymbol{Q}\boldsymbol{d}$.

Ü 11.3.4. Es sei $\boldsymbol{W} \in \mathbf{R}^{n,n}$ eine symmetrische und positiv definite Gewichtsmatrix. Man führe das gewichtete regularisierte Problem

$$\|\boldsymbol{b} - \boldsymbol{A}\boldsymbol{x}\|^2 + \alpha^2\boldsymbol{x}^\mathsf{T}\boldsymbol{W}^{-1}\boldsymbol{x} \to \text{Minimum!} \tag{42}$$

unter Verwendung der Cholesky-Faktorisierung $\boldsymbol{W} = \boldsymbol{L}\boldsymbol{L}^T$ von $\boldsymbol{W}$ auf das ungewichtete Problem

$$\|\boldsymbol{b} - \boldsymbol{A}\boldsymbol{L}\boldsymbol{y}\|^2 + \alpha^2 \|\boldsymbol{y}\|^2 \to \text{Minimum!} \tag{43}$$

mit $\boldsymbol{x} = \boldsymbol{L}\boldsymbol{y}$ zurück.

Bemerkungen zum Kapitel 11

B 11.1. Das Bidiagonalisierungsverfahren 11.1.1 geht auf GOLUB/KAHAN [65], das darauf aufbauende Verfahren 11.1.3 zur Berechnung der SVD auf GOLUB/REINSCH [71] zurück; vgl. auch GOLUB/KAHAN [68]. Zur Implementierung sei auf DONGARRA et al. [79] verwiesen. Die für $m \gg n$ nur halb so teure Version aus Ü 11.1.1 wird bei LAWSON/HANSON [74] erwähnt; ein FORTRAN-Programm hat CHAN [82] angegeben.

B 11.2. Die hier „diskrete Regularisierung der SVD" genannte Regularisierungstechnik wird bei GOLUB/REINSCH [70] beschrieben. Algorithmus 11.2.4 zur diskreten Regularisierung der $\boldsymbol{QR}$-Faktorisierung ist LAWSON/HANSON [74] entlehnt, wo auch detaillierte Hinweise zur Implementierung zu finden sind. Eine inkorrekte Formel für die Berechnung des Residuums — Formel (14.7) in LAWSON/HANSON — wurde dabei korrigiert. Die Abschätzung (11.2.48) geht auf FADDEEV et al. [68] zurück, ein Beweis kann bei LAWSON/HANSON nachgelesen werden. Die Beispiele aus Ü 11.2.2 stammen von KAHAN [66], siehe wieder LAWSON/HANSON.

B 11.3. Die Idee der kontinuierlichen Regularisierung geht u. a. auf LEVENBERG [44] und TYHONOV [63] zurück; ausführliche Darstellungen und weiterführende Literaturhinweise geben LAWSON/HANSON [74], TYHONOV/ARSENIN [79], TYHONOV et al. [83] und HOFMANN [86]. Bei der Lösung nichtlinearer Quadratmittelprobleme mittels sog. regularisierter Gauß-Newton-Verfahren — häufig auch nach MARQUARDT benannt — wird entweder α oder ω gemäß 11.3.3(i) vorgegeben. Im letzten Fall spricht man vom sog. „trust region approach". Die Steuerung von α durch $\|\boldsymbol{r}_\alpha\|$ über Defektprinzipien ist von MOROZOV [66, 73, 74/84] u. a. untersucht worden. Andere, auf dem Prinzip der „cross validation" beruhende Varianten sind bei GOLUB/HEATH/WAHBA [79] und FRIEDRICH/HOFMANN/TAUTENHAHN [79] zu finden. In der statistischen Literatur ist die kontinuierliche Regularisierung unter dem Namen „ridge regression" bekannt, und α heißt „ridge parameter". Die Identität (11.3.19), die daraus folgende Abschätzung und die Abschätzung (11.3.33) mit $\theta = 1/2$ bzw. $\theta = 2/3$ im Fall $\boldsymbol{r}^* = \boldsymbol{o}$ bzw. $\boldsymbol{r}^* \neq \boldsymbol{o}$ gehen auf VOEVODIN [69, 79] zurück, vgl. auch VOEVODIN/KUZNECOV [84, § 26]. Die unter der Voraussetzung (27) gültige feinere Abschätzung mit $\gamma_\varepsilon(\alpha)$ nach (30) und die daraus für $\boldsymbol{r}^* = \boldsymbol{o}$ folgende Abschätzung (33) mit $\theta = 1$ hat KIEŁBASIŃSKI [76, unveröffentlichtes Manuskript] angegeben. Auf die Nützlichkeit der Bidiagonalform zur Berechnung von $\boldsymbol{x}_\alpha$ ist von VOEVODIN [69] hingewiesen worden, allerdings wird dort das bidiagonale Problem durch Übergang zu den tridiagonalen Normalgleichungen gelöst.

12. Zusammenfassung zum Teil III

Das Vorgehen bei der Lösung linearer Quadratmittelprobleme $\boldsymbol{A}\boldsymbol{x} \cong \boldsymbol{b}$ hängt wesentlich davon ab, ob die Matrix $\boldsymbol{A} \in \mathbf{R}^{m,n}$ Vollrang n hat oder ob $\boldsymbol{A}$ möglicherweise rangdefizient ist, also $r = \text{rang}\,(\boldsymbol{A}) < n$ gilt.

Im gutartigen Vollrangfall mit nicht zu großem $\|\boldsymbol{A}^+\| = 1/\sigma_n(\boldsymbol{A})$ sind Orthogonalisierungsverfahren die heute bevorzugte Methode, wobei die Householder-Faktorisierung am weitesten verbreitet und am besten durch Software hoher Qualität

verfügbar gemacht ist. Vom Aufwand und dem Stabilitätsverhalten vergleichbar ist die modifizierte Gram-Schmidt-Orthogonalisierung. Sie sollte verwendet werden, wenn eine orthonormale Basis von $\mathcal{R}(\boldsymbol{A})$ explizit benötigt wird; andernfalls hat die Householder-Faktorisierung leichte Vorteile. Aufwandsmäßig liegt auch die $\boldsymbol{QR}$-Faktorisierung mittels impliziter Givens-Drehungen in derselben Größenordnung. Die Fehlerkumulationskonstanten sind jedoch deutlich kleiner als bei Verwendung von Householder-Spiegelungen. Implizite wie auch die doppelt so teuren expliziten Givens-Drehungen erlauben außerdem auf Grund ihrer flexibleren Einsatzmöglichkeiten die Ausnutzung spezieller Strukturen der Matrix $\boldsymbol{A}$, was etwa bei der Aufdatierung von $\boldsymbol{QR}$- und $\boldsymbol{LL}^T$-Faktorisierungen, der Diagonalisierung von Bidiagonalmatrizen und der $\boldsymbol{QR}$-Faktorisierung schwachbesetzter Matrizen zum Tragen kommt. Leider ist zur Zeit keine Software auf der Grundlage impliziter Givens-Drehungen allgemein verfügbar, ein Zustand, der sich möglicherweise bis zum Erscheinen dieses Buches geändert hat.

Die im Fall $m \gg n$ nur halb so aufwendigen Normalgleichungsverfahren haben gegenüber den Orthogonalisierungsverfahren den Nachteil, daß die Klasse der damit behandelbaren Probleme bei gleicher Computergenauigkeit wesentlich kleiner ist und daß speziell im Fall von fast konsistenten Systemen mit kleinem Residuum $\boldsymbol{r} = \boldsymbol{b} - \boldsymbol{Ax}$ ein größerer Genauigkeitsverlust eintritt. Der zweite Nachteil läßt sich durch iterative Verbesserung oder durch Aufstellung und Lösung der Normalgleichungen in höherer Genauigkeit beheben. Für nicht zu schlecht konditionierte Aufgaben mit nicht zu kleinen Residuen sind die Normalgleichungsverfahren daher eine Alternative zu den Orthogonalisierungsverfahren; wegen ihres geringen Aufwands sind sie besonders für Klein- und Kleinstrechner attraktiv.

Grundsätzlich anders ist die Situation bei rangdefizienten Problemen mit nicht zu großem $\|\boldsymbol{A}^+\| = 1/\sigma_r(\boldsymbol{A})$, die sich von Vollrangproblemen mit einer Singulärwertverteilung gemäß

$$\sigma_1 \geqq \cdots \geqq \sigma_r \gg \sigma_{r+1} \geqq \cdots \geqq \sigma_n > 0$$

vom numerischen Standpunkt nicht unterscheiden. Die Inkorrektheit dieser Aufgaben — ε-Störungen können sich wie $1/\varepsilon$ in der Normallösung und der Pseudoinversen auswirken — macht die Anwendung von Regularisierungstechniken erforderlich. Wesentlicher Bestandteil der diskreten Regularisierungsverfahren ist eine numerische Rangbestimmung, die zuverlässig nur mittels der numerisch berechneten Singulärwertzerlegung möglich ist, und zwar nur dann, wenn das Fehlerniveau ε bekannt ist und deutlich unter $\sigma_r(\boldsymbol{A})$ liegt. In ähnlicher Weise lassen sich $\boldsymbol{QR}$-Faktorisierungen mit Spaltenvertauschungen bzw. $\boldsymbol{LL}^T$-Faktorisierungen der Normalgleichungsmatrix mit Zeilen- und Spaltenvertauschungen zur Rangbestimmung und diskreten Regularisierung verwenden, allerdings erhöht sich dabei das Risiko einer falschen Rangfestlegung u. U. wesentlich. Eine alternative Regularisierungsmöglichkeit bilden die kontinuierlichen Regularisierungsverfahren. In beiden Verfahrensklassen ist vielfach ein interaktives Vorgehen angebracht, bei dem der Wert des Regularisierungsparameters α (bzw. direkt der Pseudorang $p = p(\alpha)$) in Abhängigkeit von Kenngrößen wie $\|\boldsymbol{x}_\alpha\|$ und $\|\boldsymbol{r}_\alpha\|$ der regularisierten Lösung vom Bearbeiter festgelegt wird.

Das in der Statistik häufig erforderliche sequentielle Ändern des Problems durch Ändern/Streichen/Hinzufügen von Spalten/Zeilen läßt sich mittels geeigneter Aufdatierungstechniken ökonomisch realisieren, und zwar sowohl bei den verschiedenen $\boldsymbol{QR}$-Faktorisierungen als auch bei der $\boldsymbol{LL}^T$-Faktorisierung der Matrix der Normalgleichungen.

Probleme mit extrem unterschiedlichen Zeilenskalierungen — sog. steife Probleme — erfordern besondere Techniken bzw. Zusatzmaßnahmen, da eine Umskalierung wegen des festgelegten statistischen Hintergrundes i. allg. nicht möglich ist. Spaltenskalierungen verändern das Residuum nicht und sind zulässig. Falls nur Darstellungsfehler vorliegen, sollten die Normen der Spalten in derselben Größenordnung liegen.

Auf Quadratmittelprobleme hoher Dimension mit schwach besetzten Koeffizientenmatrizen konnte im Rahmen dieser Einführung nicht eingegangen werden. Bewährt haben sich hier $\boldsymbol{QR}$-Faktorisierungen mittels Givens-Drehungen, aber auch iterative Verfahren für die Normalgleichungen. Wir verweisen auf die Übersichtsartikel von Björck [76], Duff/Reid [76], George/Heath/Plemmons [81], George/Heath/Ng [83] und Heath [83] und die darin zitierte Literatur.

Auf die Erzeugung von Testmatrizen mit vorgegebenen Eigenschaften wie Kondition und Rang geht Zielke [85] in einer umfangreichen Arbeit ein.

IV. Eigenwertprobleme

13. Das spezielle symmetrische Eigenwertproblem

Es sei $\mathbf{S}^{n,n} := \{\boldsymbol{A} \in \mathbf{R}^{n,n}: \boldsymbol{A} = \boldsymbol{A}^T\}$ die Menge der symmetrischen (n, n)-Matrizen. Das spezielle symmetrische Eigenwertproblem lautet dann:

Für eine gegebene Matrix $\boldsymbol{A} \in \mathbf{S}^{n,n}$ sind die Eigenwerte $\lambda_j = \lambda_j[\boldsymbol{A}]$, d. h. die Zahlen λ_j, für die

$$\boldsymbol{A}\boldsymbol{u}^j = \lambda_j \boldsymbol{u}^j \quad \text{mit} \quad \boldsymbol{u}^j \neq \boldsymbol{o}$$

gilt, und gegebenenfalls die zugehörigen Eigenvektoren $\boldsymbol{u}^j$ zu berechnen.

Je nach Herkunft der Aufgabe treten dabei spezifische Forderungen auf wie

- Berechnung des vollen Eigensystems, d. h. aller Eigenwerte und Eigenvektoren
- nur Berechnung aller Eigenwerte
- nur Berechnung gewisser Eigenwerte und gegebenenfalls zugehöriger Eigenvektoren, etwa Berechnung der l kleinsten oder größten Eigenwerte oder Berechnung der Eigenwerte aus einem vorgegebenen Intervall $[a, b]$,

vgl. auch 2.1.C1. Das spezielle Eigenwertproblem tritt in den Anwendungen so oft auf, daß es zweckmäßig ist, für jede der o. g. spezifischen Aufgaben — und manche andere — spezielle Algorithmen zu entwickeln. Derzeit genutzte Programmpakete enthalten deshalb eine Menge von Basisprozeduren, die sich in flexibler Weise zur Behandlung verschiedener spezifischer Aufgaben einsetzen lassen.

Eine wesentliche Erfahrung des numerischen Rechnens, die sich auch theoretisch begründen läßt, soll schon hier erwähnt werden: *Die Eigenwerte* $\{\lambda_j\}$ *von* $\boldsymbol{A} = (a_{ij}) \in \mathbf{S}^{n,n}$ *sollen i. allg. niemals als Nullstellen des charakteristischen Polynoms* $p(\lambda) = \det(\boldsymbol{A} - \lambda \boldsymbol{I}) = \sum_{k=0}^{n} c_k \lambda^k = 0$ *unter Verwendung der numerisch berechneten oder auch exakt bekannten Koeffizienten* $\{c_k\}$ *bestimmt werden. Der Berechnungsweg*

$$\{a_{ij}\} \to \{c_k\} \to \{\lambda_l\}$$

ist i. allg. instabil, häufig sogar in katastrophaler Weise. Der Weg über das charakteristische Polynom scheidet daher i. allg. aus, und zwar auch für $n = 2$, siehe Ü 13.1.1.

Wir bemerken abschließend, daß das spezielle symmetrische Eigenwertproblem wesentlich günstigere Eigenschaften im Hinblick auf Anzahl der benötigten arithmetischen Operationen, Speicherplatzanforderungen und erreichbare Genauigkeit aufweist als das spezielle Eigenwertproblem nichtsymmetrischer Matrizen.

13.1. Grundlegende Eigenschaften, Störungstheorie, Residualkriterien

A. *Allgemeine Eigenschaften*

Die Eigenschaften des durch die Matrix $\boldsymbol{A} \in \mathbf{R}^{n,n}$ bzw. $\boldsymbol{A} \in \mathbf{S}^{n,n}$ definierten speziellen Eigenwertproblems sind bereits im Abschnitt 1.2 zusammengestellt worden. Wir erinnern daran, daß sich der Fall einer symmetrischen Matrix $\boldsymbol{A}$ von dem einer nichtsymmetrischen im wesentlichen dadurch unterscheidet, daß

- alle Eigenwerte reell sind
- eine orthonormale Basis von Eigenvektoren existiert
- die Anzahl der positiven, negativen und verschwindenden Eigenwerte invariant unter Kongruenztransformationen ist.

In diesem Abschnitt seien die entsprechend ihrer Vielfachheit gezählten Eigenwerte $\{\lambda_j\}$ von $\boldsymbol{A}$ gemäß

$$\lambda_1 \leqq \lambda_2 \leqq \cdots \leqq \lambda_{n-1} \leqq \lambda_n \tag{1}$$

geordnet, und $\{\boldsymbol{u}^1, \ldots, \boldsymbol{u}^n\}$ sei ein zugehöriges orthonormales System von Eigenvektoren, d. h., es gelte

$$\boldsymbol{A}\boldsymbol{u}^j = \lambda_j \boldsymbol{u}^j \quad \text{mit} \quad \boldsymbol{u}^{i\mathsf{T}}\boldsymbol{u}^j = \begin{cases} 1 & \text{für} \quad i = j, \\ 0 & \text{für} \quad i \neq j \end{cases} \quad (i, j = 1, \ldots, n). \tag{2}$$

Wenn

$$\boldsymbol{U} := (\boldsymbol{u}^1, \ldots, \boldsymbol{u}^n) \quad \text{und} \quad \boldsymbol{\Lambda} := \operatorname{diag}(\lambda_1, \ldots, \lambda_n) \tag{3}$$

gesetzt wird, läßt sich (2) in der Gestalt

$$\boldsymbol{A}\boldsymbol{U} = \boldsymbol{U}\boldsymbol{\Lambda} \quad \text{mit} \quad \boldsymbol{U}^{\mathsf{T}}\boldsymbol{U} = \boldsymbol{I}$$

schreiben, d. h., $\boldsymbol{U}$ ist orthogonal, und es gilt

$$\boldsymbol{U}^{-1}\boldsymbol{A}\boldsymbol{U} = \boldsymbol{U}^{\mathsf{T}}\boldsymbol{A}\boldsymbol{U} = \boldsymbol{\Lambda}. \tag{4}$$

Die Matrix $\boldsymbol{A}$ wird also durch Ähnlichkeitstransformation mit der orthogonalen Matrix $\boldsymbol{U}$ auf Diagonalform $\boldsymbol{\Lambda}$ gebracht, und diese Transformation ist gleichzeitig eine Kongruenztransformation, vgl. 1.2.4 und 1.2.13. Wenn $\boldsymbol{A}$ mittels einer orthogonalen Matrix $\boldsymbol{Q} \in \mathbf{R}^{n,n}$ gemäß

$$\bar{\boldsymbol{A}} = \boldsymbol{Q}^{\mathsf{T}}\boldsymbol{A}\boldsymbol{Q} \tag{5}$$

ähnlich in $\bar{\boldsymbol{A}} \in \mathbf{S}^{n,n}$ transformiert wird, ändern sich die Eigenwerte nicht, d. h., $\bar{\boldsymbol{A}}$ hat auch die Eigenwerte $\{\lambda_j\}$, und die Spalten der orthogonalen Matrix

$$\bar{\boldsymbol{U}} = (\bar{\boldsymbol{u}}^1, \ldots, \bar{\boldsymbol{u}}^n) = \boldsymbol{Q}^T \boldsymbol{U}$$

bilden ein orthonormales System zugehöriger Eigenvektoren. Ist umgekehrt ein Eigenpaar $\{\lambda, \bar{\boldsymbol{u}}\}$ von $\bar{\boldsymbol{A}}$ bekannt, so erhält man durch die Rücktransformation

$$\boldsymbol{u} = \boldsymbol{Q}\bar{\boldsymbol{u}} \tag{6}$$

ein Eigenpaar $\{\lambda, \boldsymbol{u}\}$ von $\boldsymbol{A}$, siehe 1.2.1.

Wir werden speziell gewählte Ähnlichkeitstransformationen (5) benutzen, um das durch $\boldsymbol{A}$ definierte Ausgangsproblem in das Eigenwertproblem einer Matrix $\bar{\boldsymbol{A}}$ mit einfacherer Gestalt zu transformieren.

Eine weitere, praktisch wie theoretisch nützliche Transformation stellt die sog. Spektralverschiebung oder kurz Verschiebung (engl. „shift", russ. „сдвиг")

$$\bar{\boldsymbol{A}} = \boldsymbol{A} - \mu \boldsymbol{I} \tag{7}$$

um $\mu \in \mathbf{R}$ dar. Offensichtlich ist $\{\lambda, \boldsymbol{u}\}$ Eigenpaar von $\boldsymbol{A}$ genau dann, wenn $\{\bar{\lambda}, \boldsymbol{u}\}$ mit

$$\bar{\lambda} = \lambda - \mu \tag{8}$$

Eigenpaar von $\bar{\boldsymbol{A}}$ ist. Durch (7) werden also die Eigenvektoren nicht verändert, während die Eigenwerte um das durch den reellen Verschiebungsparameter μ charakterisierte Stück auf der reellen Achse verschoben werden.

B. *Störungstheorie*

Wir betrachten jetzt neben der Matrix $\boldsymbol{A} \in \mathbf{S}^{n,n}$ mit den Eigenpaaren $\{\lambda_j, \boldsymbol{u}^j\}$ gemäß (2) noch die gestörte Matrix

$$\boldsymbol{A} + \delta\boldsymbol{A} \quad \text{mit} \quad \delta\boldsymbol{A} \in \mathbf{S}^{n,n},$$

und es sei $\{\mu, \boldsymbol{u}\}$ ein beliebiges Eigenpaar von $\boldsymbol{A} + \delta\boldsymbol{A}$ mit $\|\boldsymbol{u}\| = 1$, d. h., es gelte

$$(\boldsymbol{A} + \delta\boldsymbol{A})\,\boldsymbol{u} = \mu \boldsymbol{u}, \qquad \|\boldsymbol{u}\| = 1. \tag{9}$$

Als Norm verwenden wir stets die Euklidische Vektornorm bzw. die Spektralnorm, sofern nichts Gegenteiliges explizit gesagt wird.

Wenn $\boldsymbol{u}$ gemäß

$$\boldsymbol{u} = \boldsymbol{U}\boldsymbol{z} = \sum_{j=1}^{n} z_j \boldsymbol{u}^j \quad \text{mit} \quad \|\boldsymbol{z}\| = 1 \tag{10}$$

als Linearkombination der Eigenvektoren von $\boldsymbol{A}$ ausgedrückt wird, geht (9) nach Multiplikation mit $\boldsymbol{U}^T$ von links in

$$\boldsymbol{U}^T(\boldsymbol{A} + \delta\boldsymbol{A})\,\boldsymbol{U}\boldsymbol{z} = \mu \boldsymbol{z} \quad \text{mit} \quad \|\boldsymbol{z}\| = 1$$

über, woraus wegen (4) und der Orthogonalität von $\boldsymbol{U}$

$$\boldsymbol{y} := (\mu \boldsymbol{I} - \boldsymbol{A})\,\boldsymbol{z} = \boldsymbol{U}^{\mathrm{T}} \delta \boldsymbol{A} \boldsymbol{U} \boldsymbol{z}, \tag{11}$$

also

$$\|\boldsymbol{y}\| \leqq \|\delta \boldsymbol{A}\|\,\|\boldsymbol{z}\| = \|\delta \boldsymbol{A}\| \tag{12}$$

folgt. Nun gilt

$$\|\boldsymbol{y}\|^2 = \sum_{j=1}^{n} (\mu - \lambda_j)^2 z_j^2 \geqq \left\{\min_j |\mu - \lambda_j|\right\}^2 \sum_{j=1}^{n} z_j^2 = \left\{\min_j |\mu - \lambda_j|\right\}^2,$$

so daß sich aus (12) sofort der folgende Satz ergibt:

13.1.1. *Bauer-Fike-Theorem.* Es gelte $\boldsymbol{A}, \delta \boldsymbol{A} \in \mathbf{S}^{n,n}$, es seien $\{\lambda_1, \ldots, \lambda_n\}$ die Eigenwerte von $\boldsymbol{A}$, und μ sei irgendein Eigenwert von $\boldsymbol{A} + \delta \boldsymbol{A}$. Dann gilt

$$\min_{j=1,\ldots,n} |\mu - \lambda_j| \leqq \|\delta \boldsymbol{A}\|.$$

Eine genauere Analyse zeigt, daß auch die folgende schärfere Aussage gültig ist:

13.1.2. *Satz.* Es gelte $\boldsymbol{A}, \delta \boldsymbol{A} \in \mathbf{S}^{n,n}$. Es seien $\{\lambda_j\}$ die gemäß

$$\lambda_1 \leqq \lambda_2 \leqq \cdots \leqq \lambda_n$$

geordneten Eigenwerte von $\boldsymbol{A}$, und $\{\mu_j\}$ seien die analog gemäß

$$\mu_1 \leqq \mu_2 \leqq \cdots \leqq \mu_n$$

geordneten Eigenwerte von $\boldsymbol{A} + \delta \boldsymbol{A}$. Dann gilt

$$|\mu_j - \lambda_j| \leqq \|\delta \boldsymbol{A}\| \qquad (j = 1, \ldots, n).$$

Unter Verwendung der Frobeniusnorm kann schließlich die folgende simultane Abschätzung bewiesen werden:

13.1.3. *Wielandt-Hoffman-Theorem.* Unter den Voraussetzungen von 13.1.2 gilt

$$\sqrt{\sum_{j=1}^{n} (\mu_j - \lambda_j)^2} \leqq \|\delta \boldsymbol{A}\|_F.$$

Die letzten beiden Sätze besagen: *Die Eigenwerte symmetrischer Matrizen sind lipschitzstetige Funktionen der Matrix, und die Lipschitzkonstante hat in den angegebenen Normen den Wert* 1. *Das symmetrische Eigenwertproblem ist in diesem Sinne also extrem gut konditioniert.*

Um zu einer besseren Einsicht in die individuelle Empfindlichkeit der einzelnen Eigenwerte und Eigenvektoren zu gelangen, betrachten wir wieder ein Eigenpaar $\{\mu, \boldsymbol{u}\}$ von $\boldsymbol{A} + \delta \boldsymbol{A}$, das der Beziehung (9) genügt. Aus (10) folgt dann

$$\boldsymbol{z} = (z_i) = \boldsymbol{U}^{\mathrm{T}} \boldsymbol{u} \quad \text{mit} \quad z_i = \boldsymbol{u}^{i\mathrm{T}} \boldsymbol{u} = \cos\left(\sphericalangle\,(\boldsymbol{u}^i, \boldsymbol{u})\right). \tag{13}$$

Mit $\boldsymbol{y} = (y_i)$ gemäß (11) ergibt sich ferner

$$y_i = (\mu - \lambda_i)\, z_i = \boldsymbol{u}^{i\mathrm{T}} \delta \boldsymbol{A} \boldsymbol{u} \qquad (i = 1, \ldots, n). \tag{14}$$

Es sei jetzt j ein Index mit $z_j \neq 0$, d. h., $\boldsymbol{u}$ sei nicht orthogonal zu $\boldsymbol{u}^j$. Aus (14) folgt dann

$$|\mu - \lambda_j| = |\boldsymbol{u}^{j\mathsf{T}} \delta \boldsymbol{A} \boldsymbol{u}|/|z_j| \leqq \|\delta \boldsymbol{A} \boldsymbol{u}\|/|z_j|.$$

Im Fall $\mu \neq \lambda_i$ läßt sich abschätzen, wie $\boldsymbol{u}$ und $\boldsymbol{u}^i$ von der Orthogonalität abweichen, denn (13) und (14) ziehen dann

$$\left|\cos\left(\sphericalangle\,(\boldsymbol{u}^i, \boldsymbol{u})\right)\right| = |z_i| = |\boldsymbol{u}^{i\mathsf{T}} \delta \boldsymbol{A} \boldsymbol{u}|/|\mu - \lambda_i| \leqq \|\delta \boldsymbol{A} \boldsymbol{u}\|/|\mu - \lambda_i|$$

nach sich. Ist für festes j sogar $\gamma_j = \min\{|\mu - \lambda_i| : i \neq j\} > 0$, so ergibt sich mit (10), (14) und (11)

$$\left[\sin\left(\sphericalangle\,(\boldsymbol{u}^j, \boldsymbol{u})\right)\right]^2 = 1 - z_j^2 = \sum_{i \neq j} z_i^2 = \sum_{i \neq j} \frac{|y_i|^2}{|\mu - \lambda_i|^2} \leqq \frac{\|\boldsymbol{y}\|^2}{\gamma_j^2} = \frac{\|\delta \boldsymbol{A} \boldsymbol{u}\|^2}{\gamma_j^2}$$

als Abschätzung für den Winkel zwischen $\boldsymbol{u}$ und $\boldsymbol{u}^j$.

Zusammenfassend erhalten wir das folgende Ergebnis:

13.1.4. *Aussage.* Es gelte $\boldsymbol{A}, \delta \boldsymbol{A} \in \mathbf{S}^{n,n}$, und es sei $\{\lambda_i, \boldsymbol{u}^i\}$ ein orthonormales Eigensystem von $\boldsymbol{A}$. Dann gilt für jedes Eigenpaar $\{\mu, \boldsymbol{u}\}$ von $\boldsymbol{A} + \delta \boldsymbol{A}$ mit $\|\boldsymbol{u}\| = 1$

(i) im Fall $\boldsymbol{u}^{j\mathsf{T}} \boldsymbol{u} \neq 0$

$$|\mu - \lambda_j| \leqq \frac{\|\delta \boldsymbol{A} \boldsymbol{u}\|}{|\boldsymbol{u}^{j\mathsf{T}} \boldsymbol{u}|} \leqq \frac{\|\delta \boldsymbol{A}\|}{|\boldsymbol{u}^{j\mathsf{T}} \boldsymbol{u}|}, \tag{15}$$

(ii) im Fall $\mu \neq \lambda_i$

$$\left|\cos\left(\sphericalangle\,(\boldsymbol{u}^i, \boldsymbol{u})\right)\right| \leqq \frac{\|\delta \boldsymbol{A} \boldsymbol{u}\|}{|\mu - \lambda_i|} \leqq \frac{\|\delta \boldsymbol{A}\|}{|\mu - \lambda_i|}, \tag{16}$$

(iii) im Fall $\gamma_j := \min\{|\mu - \lambda_i| : i \neq j\} > 0$

$$\sin\left(\sphericalangle\,(\boldsymbol{u}^j, \boldsymbol{u})\right) \leqq \frac{\|\delta \boldsymbol{A} \boldsymbol{u}\|}{\gamma_j} \leqq \frac{\|\delta \boldsymbol{A}\|}{\gamma_j}. \tag{17}$$

13.1.5. *Bemerkung.* (i) Mit dem i. allg. nicht normierten Vektor $\boldsymbol{v} := \boldsymbol{u}/\boldsymbol{u}^{j\mathsf{T}} \boldsymbol{u} = \boldsymbol{u}/z_j$ läßt sich (15) äquivalent als

$$|\mu - \lambda_j| \leqq \|\delta \boldsymbol{A} \boldsymbol{v}\| \leqq \|\delta \boldsymbol{A}\| \, \|\boldsymbol{v}\|$$

schreiben. Dabei gilt im Fall $z_j > 0$

$$\boldsymbol{v} = \boldsymbol{u}^j + \delta \boldsymbol{u}^j \quad \text{mit} \quad \boldsymbol{u}^j \perp \delta \boldsymbol{u}^j, \tag{18}$$

also

$$\|\boldsymbol{v}\|^2 = 1 + \|\delta \boldsymbol{u}^j\|^2, \qquad \|\delta \boldsymbol{u}^j\| = \|\boldsymbol{v}\| \sin\left(\sphericalangle\,(\boldsymbol{u}^j, \boldsymbol{u})\right), \tag{19}$$

vgl. Abb. 13.1.1; im Fall $z_j < 0$ ersetze man $\boldsymbol{u}^j$ durch $-\boldsymbol{u}^j$.

(ii) Es seien jetzt $\{\mu_j\}$ die gemäß 13.1.2 geordneten Eigenwerte von $\boldsymbol{A} + \delta \boldsymbol{A}$, und $\{\boldsymbol{v}^j\}$ sei ein zugehöriges orthonormales System von Eigenvektoren. Wenn $\{\mu, \boldsymbol{u}\}$

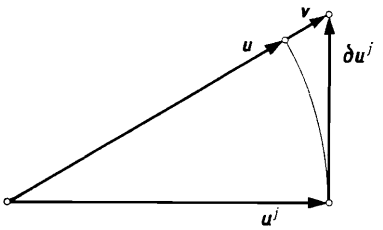


Abb. 13.1.1. Eigenvektoren $\boldsymbol{u}^j$, $\boldsymbol{u}$ und $\boldsymbol{v}$

$= \{\mu_j, \boldsymbol{v}^j\}$ gesetzt wird, besagt Aussage (ii): Falls μ_j genügend von λ_i getrennt und $\delta\boldsymbol{A}$ genügend klein ist, sind die zugehörigen Eigenvektoren $\boldsymbol{v}^j$ von $\boldsymbol{A} + \delta\boldsymbol{A}$ und $\boldsymbol{u}^i$ von $\boldsymbol{A}$ fast orthogonal. Da nach 13.1.2

$$|\mu - \lambda_i| = |\mu_j - \lambda_i| \geqq |\lambda_j - \lambda_i| - |\lambda_j - \mu_j| \geqq |\lambda_j - \lambda_i| - \|\delta\boldsymbol{A}\| \tag{20}$$

gilt, ist $\mu = \mu_j$ genügend von λ_i getrennt, wenn λ_j genügend von λ_i getrennt und $\delta\boldsymbol{A}$ genügend klein ist.

Dagegen besagt Aussage (iii): Falls μ_j genügend von *allen* λ_i $(i \neq j)$ getrennt und $\delta\boldsymbol{A}$ genügend klein ist, weichen die Richtungen von $\boldsymbol{v}^j$ und $\boldsymbol{u}^j$ nur wenig voneinander ab. Da aus (20)

$$\gamma_j \geqq \min\{|\lambda_j - \lambda_i| : i \neq j\} - \|\delta\boldsymbol{A}\| \tag{21}$$

folgt, ist letzteres sicher der Fall, wenn λ_j von allen übrigen Eigenwerten λ_i genügend getrennt und $\delta\boldsymbol{A}$ genügend klein ist, also eine genügend große *Spektrumslücke* um λ_j vorliegt.

(iii) Wenn gewisse Eigenwerte λ_i $(i \neq j)$ im Sinne von $|\lambda_i - \lambda_j| = O(\|\delta\boldsymbol{A}\|)$ pathologisch dicht zu λ_j benachbart sind, gilt $|\mu_j - \lambda_i| = O(\|\delta\boldsymbol{A}\|)$ und somit $\gamma_j = O(\|\delta\boldsymbol{A}\|)$, so daß (16) und (17) i. allg. keine brauchbaren Schranken liefern. Daß dies ohne Vorliegen einer genügend großen Spektrumslücke auch nicht erwartet werden kann, zeigt das Beispiel aus Ü 2.1.9. Ist ein solcher „*Haufen*" benachbarter Eigenwerte jedoch ausreichend von den restlichen Eigenwerten getrennt, so lassen sich analoge Schranken für den geeignet zu definierenden Winkel zwischen den Räumen, die durch die entsprechenden Eigenvektoren $\boldsymbol{u}^j$ bzw. $\boldsymbol{v}^j$ aufgespannt werden, angeben.

(iv) In gewissen wichtigen Fällen, auf die wir im folgenden eingehen werden, gilt $\|\delta\boldsymbol{A}\boldsymbol{u}\| \ll \|\delta\boldsymbol{A}\|\,\|\boldsymbol{u}\| = \|\delta\boldsymbol{A}\|$. Die Schranken mit $\|\delta\boldsymbol{A}\boldsymbol{u}\|$ sind dann wesentlich besser als diejenigen, die durch die Vergröberung $\|\delta\boldsymbol{A}\boldsymbol{u}\| \leqq \|\delta\boldsymbol{A}\|$ aus diesen entstehen. Wegen der Symmetrie von $\delta\boldsymbol{A}$ und der Gültigkeit von $|\boldsymbol{u}^{i\mathsf{T}}\delta\boldsymbol{A}\boldsymbol{u}| = |(\delta\boldsymbol{A}\boldsymbol{u}^i)^{\mathsf{T}}\boldsymbol{u}| \leqq \|\delta\boldsymbol{A}\boldsymbol{u}^i\|$ kann überdies der Term $\|\delta\boldsymbol{A}\boldsymbol{u}\|$ in (15) und (16) durch $\|\delta\boldsymbol{A}\boldsymbol{u}^j\|$ bzw. $\|\delta\boldsymbol{A}\boldsymbol{u}^i\|$ ersetzt werden. □

Zur Anwendung der Störungsabschätzungen betrachten wir eine Matrix $\boldsymbol{A}^* \in \mathbf{S}^{n,n}$, die durch ihre Computerdarstellung $\boldsymbol{A} \in \mathfrak{R}^{n,n}$, $\boldsymbol{A} = \boldsymbol{A}^{\mathsf{T}}$, repräsentiert wird. Dann gilt

$$\|\delta\boldsymbol{A}_D\| = \|\boldsymbol{A}^* - \boldsymbol{A}\| \leqq \nu\sqrt{n}\,\|\boldsymbol{A}^*\|,$$

siehe 2.3.15, so daß aus 13.1.2 die Abschätzung

$$|\lambda_j^* - \lambda_j| \leqq \nu \sqrt{n}\, \|\boldsymbol{A}^*\| = \nu \sqrt{n} \max \{|\lambda_i^*| : i = 1, \ldots, n\} \tag{22}$$

folgt. Durch (22) ist das optimale Fehlerniveau von λ_j gegeben.

Es seien jetzt $\{\mu_j\}$ die nach einem numerisch gutartigen Verfahren aus $\boldsymbol{A}$ berechneten Eigenwertnäherungen, d. h., die $\{\mu_j\}$ sind die exakten Eigenwerte einer gestörten Matrix $\boldsymbol{A} + \delta\boldsymbol{A}$, $\delta\boldsymbol{A} \in \mathbf{S}^{n,n}$, wobei

$$\|\delta\boldsymbol{A}\| \leqq \nu F \|\boldsymbol{A}\| \tag{23}$$

gilt. Dann liefert 13.1.2 für $\delta\lambda_j := \mu_j - \lambda_j$ die zu (22) analoge Schranke

$$|\delta\lambda_j| \leqq \|\delta\boldsymbol{A}\| \leqq \nu F \max \{|\lambda_i| : i = 1, \ldots, n\},$$

und für den relativen Fehler gilt

$$\frac{|\delta\lambda_j|}{|\lambda_j|} \leqq \nu F \frac{\max \{|\lambda_i| : i = 1, \ldots, n\}}{|\lambda_j|}. \tag{24}$$

Die Schranke (24) garantiert nur für die betragsgroßen Eigenwerte λ_j einen kleinen relativen Fehler, während die betragskleinen μ_j durchaus stark von den zugehörigen exakten λ_j abweichen können. *Die Forderung, auch die betragskleinen Eigenwerte von $\boldsymbol{A}$ mit hoher relativer Genauigkeit zu berechnen, ist deshalb i. allg. unrealistisch und nur schwer bzw. überhaupt nicht zu erfüllen.* Wenn jedoch $\boldsymbol{A}$ eine spezielle Struktur hat und die Störung $\delta\boldsymbol{A}$ dieser Struktur angepaßt ist, liefern die individuellen Abschätzungen aus 13.1.4 auch kleine Schranken für den relativen Fehler der betragskleinen Eigenwerte. Beispiele solcher Matrizen sind die sog. *gestuften* (engl. „graded") Matrizen, bei denen die Beträge der Elemente von links oben nach rechts unten abnehmen. Zur Illustration betrachten wir das folgende Zahlenbeispiel.

13.1.6. *Beispiel.* Gegeben sei die gestufte Matrix

$$\boldsymbol{A} = \begin{pmatrix} 10000 & 100 & 1 \\ 100 & 100 & 1 \\ 1 & 1 & 1 \end{pmatrix} \quad \text{mit} \quad \begin{matrix} \lambda_1 = 0.9899 \\ \lambda_2 = 99.0 \\ \|\boldsymbol{A}\| = \lambda_3 = 10001 \end{matrix} \quad \text{und} \quad \boldsymbol{u}^1 = \begin{pmatrix} 0.000001 \\ -0.010100 \\ 0.999950 \end{pmatrix}.$$

Für $\nu := 10^{-6}$, $F := 10$ und eine im Sinne von (23) kleine Störung ist $\|\delta\boldsymbol{A}\| \lessapprox 10^{-1}$, und (24) führt auf

$$|\delta\lambda_1/\lambda_1| \lessapprox 10^{-6} \times 10 \times 10^4 = 10^{-1}, \tag{25}$$

d. h., der kleinste Eigenwert λ_1 kann bis zu 10% gestört werden. Wenn jedoch die Störung $\delta\boldsymbol{A}$ in zu $\boldsymbol{A}$ analoger Weise gestuft ist, d. h., wenn

$$|\delta\boldsymbol{A}| \leqq \nu F |\boldsymbol{A}| = 10^{-5} |\boldsymbol{A}| \tag{26}$$

gilt, ist wieder $\|\delta\boldsymbol{A}\| \leqq 10^{-5} \|\boldsymbol{A}\| \approx 10^{-1}$, aber

$$|\delta\boldsymbol{A}\boldsymbol{u}^1| \leqq 10^{-5} |\boldsymbol{A}|\, |\boldsymbol{u}^1| \approx 10^{-5} \begin{pmatrix} 2 \\ 2 \\ 1 \end{pmatrix}, \qquad \|\delta\boldsymbol{A}\boldsymbol{u}^1\| \lessapprox 3 \times 10^{-5}.$$

Aus (17) ergibt sich mit $\gamma_1 \approx 98$ zunächst $s := \sin\left(\sphericalangle\, (\boldsymbol{u}^1, \boldsymbol{u})\right) \lessapprox 0.1/98 \approx 1 \times 10^{-3}$, wobei $\boldsymbol{u}$

den zu $\mu = \lambda_1 + \delta\lambda_1$ gehörenden normierten Eigenvektor von $\boldsymbol{A} + \delta\boldsymbol{A}$ bezeichnet. Dies führt auf

$$|\boldsymbol{u}^{1\mathrm{T}}\boldsymbol{u}| = \left|\cos\left(\sphericalangle\,(\boldsymbol{u}^1, \boldsymbol{u})\right)\right| = \sqrt{1 - s^2} \gtrsim 1 - 5 \times 10^{-7}$$

und weiter mit (15) unter Beachtung von 13.1.5(iv) auf

$$|\delta\lambda_1/\lambda_1| \leqq \|\delta\boldsymbol{A}\boldsymbol{u}^1\|/|\lambda_1 \boldsymbol{u}^{1\mathrm{T}}\boldsymbol{u}| \lesssim 3 \times 10^{-5}, \tag{27}$$

also eine wesentlich bessere Schranke als (25). Man beachte, daß hier

$$\|\delta\boldsymbol{A}\boldsymbol{u}^1\| \approx 3 \times 10^{-5} \ll 10^{-1} \approx \|\delta\boldsymbol{A}\|\,\|\boldsymbol{u}^1\| = \|\delta\boldsymbol{A}\|$$

gilt. Ebenso kann die oben benutzte Schranke für s verbessert werden: Aus (19) folgt $\|\delta\boldsymbol{u}^1\| = \sqrt{1 + \|\delta\boldsymbol{u}^1\|^2}\, s$, also

$$\|\delta\boldsymbol{u}^1\| = s/\sqrt{1 - s^2} \lesssim 1 \times 10^{-3}/(1 - 5 \times 10^{-7}) \approx 1 \times 10^{-3}$$

und mit (18) — vgl. auch Abb. 13.1.1 —

$$\begin{aligned}\|\delta\boldsymbol{A}\boldsymbol{u}\| = |\boldsymbol{u}^{1\mathrm{T}}\boldsymbol{u}|\,\|\delta\boldsymbol{A}\boldsymbol{v}\| &\leqq \|\delta\boldsymbol{A}\boldsymbol{v}\| \leqq \|\delta\boldsymbol{A}\boldsymbol{u}^1\| + \|\delta\boldsymbol{A}\|\,\|\delta\boldsymbol{u}^1\| \lesssim 3 \times 10^{-5} + 1 \times 10^{-4} \\ &= 1.3 \times 10^{-4}.\end{aligned}$$

Nach (17) ergibt sich daraus die verbesserte Schranke

$$s \leqq \|\delta\boldsymbol{A}\boldsymbol{u}\|/\gamma_1 \lesssim 1.3 \times 10^{-6}.$$

Durch wiederholte Anwendung dieses Prinzips erhält man schließlich $s \lesssim 3.1 \times 10^{-7}$ und $\|\delta\boldsymbol{u}^1\| \lesssim 1.3 \times 10^{-7}$, d. h., $\boldsymbol{u}$ und $\boldsymbol{u}^1$ weichen nur sehr wenig voneinander ab. □

Das Beispiel zeigt, daß für die Berechnung der betragskleinen Eigenwerte gestufter Matrizen solche numerisch gutartigen Algorithmen verwendet werden sollten, bei denen die den erzeugten Rundungsfehlern äquivalente Störung $\delta\boldsymbol{A}$ in zu $\boldsymbol{A}$ analoger Weise gestuft ist.

C. *Residualkriterien*

Wie bei linearen Gleichungssystemen und Quadratmittelproblemen können auch bei Eigenwertaufgaben geeignete Residuen zur Kontrolle und gegebenenfalls Verbesserung der Genauigkeit verwendet werden. Wir beginnen mit der folgenden Aussage:

13.1.7. *Aussage.* Es sei $\boldsymbol{A} \in \mathbf{S}^{n,n}$, $\mu \in \mathbf{R}$, $\boldsymbol{u} \in \mathbf{R}^n$ mit $\|\boldsymbol{u}\| = 1$. Dann sind die folgenden Aussagen äquivalent:

(i) Es existiert ein $\delta\boldsymbol{A} \in \mathbf{S}^{n,n}$ mit

$$(\boldsymbol{A} + \delta\boldsymbol{A})\,\boldsymbol{u} = \mu\boldsymbol{u} \quad \text{und} \quad \|\delta\boldsymbol{A}\| \leqq \varepsilon. \tag{28}$$

(ii) Es gilt

$$\|\boldsymbol{r}\| \leqq \varepsilon \quad \text{mit} \quad \boldsymbol{r} := \boldsymbol{A}\boldsymbol{u} - \mu\boldsymbol{u}. \tag{29}$$

Der Vektor $\boldsymbol{r} = (\boldsymbol{A} - \mu\boldsymbol{I})\,\boldsymbol{u}$ heißt *Residuum* des Paares $\{\mu, \boldsymbol{u}\}$ in bezug auf das Eigenwertproblem der Matrix $\boldsymbol{A}$.

Beweis. Aus (28) folgt $\boldsymbol{r} = -\delta\boldsymbol{A}\boldsymbol{u}$, also $\|\boldsymbol{r}\| \leqq \|\delta\boldsymbol{A}\| \leqq \varepsilon$. Gelte umgekehrt (29), d. h., es sei $\eta := \|\boldsymbol{r}\| \leqq \varepsilon$. Dann gibt es eine Matrix $\boldsymbol{H} \in \mathbf{S}^{n,n}$ mit $\|\boldsymbol{H}\| = 1$ und $\boldsymbol{r} = \eta\boldsymbol{H}\boldsymbol{u}$, etwa $\boldsymbol{H} := \boldsymbol{I}$ im Fall $\eta = 0$ oder im Fall $\eta > 0$ die Spiegelungsmatrix, die $\boldsymbol{u}$ in $\boldsymbol{r}/\eta$ abbildet, siehe Ü 3.3.2. Dies zieht $\boldsymbol{A}\boldsymbol{u} - \mu\boldsymbol{u} = \eta\boldsymbol{H}\boldsymbol{u}$, also (28) mit $\delta\boldsymbol{A} := -\eta\boldsymbol{H}$, $\|\delta\boldsymbol{A}\| = \eta \leqq \varepsilon$ nach sich. □

Aus 13.1.7 *folgt insbesondere, daß* $\{\mu, \boldsymbol{u}\}$ *genau dann ein Eigenpaar der im Sinne von*

$$\|\delta\boldsymbol{A}\| \leqq \nu F \, \|\boldsymbol{A}\|$$

zu $\boldsymbol{A}$ *benachbarten Matrix* $\boldsymbol{A} + \delta\boldsymbol{A}$ *ist, wenn*

$$\|\boldsymbol{r}\| = \|\boldsymbol{A}\boldsymbol{u} - \mu\boldsymbol{u}\| \leqq \nu F \, \|\boldsymbol{A}\|$$

gilt, d. h., wenn das relative Residuum $\|\boldsymbol{r}\|/\|\boldsymbol{A}\|$ *durch* νF *beschränkt ist.*

Da verschiedene Algorithmen nur Näherungen für Eigenwerte, andere nur Näherungen für Eigenvektoren liefern, entstehen sofort die Fragen:

— Wie soll μ bei gegebener Eigenvektornäherung $\boldsymbol{u}$ gewählt werden, damit $\{\mu, \boldsymbol{u}\}$ ein möglichst gutes Eigenpaar darstellt?
— Wie soll $\boldsymbol{u}$ bei gegebener Eigenwertnäherung μ gewählt werden, damit $\{\mu, \boldsymbol{u}\}$ ein möglichst gutes Eigenpaar ist?

Entsprechend 13.1.7 soll dabei die Güte von $\{\mu, \boldsymbol{u}\}$ durch die Norm des Residuums $\boldsymbol{r}$ charakterisiert werden.

Wir untersuchen im folgenden das erste Problem.

13.1.8. *Aussage.* Es seien $\boldsymbol{A} \in \mathbf{S}^{n,n}$ und $\boldsymbol{u} \in \mathbf{R}^n$ mit $\|\boldsymbol{u}\| = 1$ gegeben. Dann gilt:

(i) Das Quadratmittelproblem

$$\|\boldsymbol{A}\boldsymbol{u} - \mu\boldsymbol{u}\| \to \underset{\mu \in \mathbf{R}}{\text{Minimum!}} \tag{30}$$

hat die eindeutige Lösung

$$\mu = \boldsymbol{u}^{\mathsf{T}}\boldsymbol{A}\boldsymbol{u} =: \varrho. \tag{31}$$

Das zugehörige kürzeste Residuum $\boldsymbol{r} := \boldsymbol{A}\boldsymbol{u} - \varrho\boldsymbol{u}$ ist zu $\boldsymbol{u}$ orthogonal, d. h., es gilt

$$\boldsymbol{u}^{\mathsf{T}}\boldsymbol{r} = 0. \tag{32}$$

(ii) Das Paar $\{\varrho, \boldsymbol{u}\}$ ist Eigenpaar einer Matrix $\boldsymbol{A} + \delta\boldsymbol{A}$, wobei $\delta\boldsymbol{A} \in \mathbf{S}^{n,n}$ ist mit $\|\delta\boldsymbol{A}\| \leqq \|\boldsymbol{r}\|$.

(iii) Es gibt ein Eigenpaar $\{\lambda_j, \boldsymbol{u}^j\}$ von $\boldsymbol{A}$ mit

$$|\varrho - \lambda_j| = \min\{|\varrho - \lambda_i| : i = 1, \ldots, n\} \leqq \|\boldsymbol{r}\|. \tag{33}$$

Im Fall $\gamma_j := \min\{|\varrho - \lambda_i| : i \neq j\} > 0$ gilt

$$\sin\left(\sphericalangle(\boldsymbol{u}, \boldsymbol{u}^j)\right) \leqq \|\boldsymbol{r}\|/\gamma_j \tag{34}$$

und

$$|\varrho - \lambda_j| \leqq \|\boldsymbol{r}\|^2/\gamma_j. \tag{35}$$

Beweis. Die Normalgleichungen zu (30) lauten $\boldsymbol{u}^{\mathsf{T}}\boldsymbol{u}\mu = \boldsymbol{u}^{\mathsf{T}}\boldsymbol{A}\boldsymbol{u}$, woraus sofort (31) und (32) folgen, vgl. 8.1.6. Die Aussage (ii) ergibt sich aus 13.1.7 und impliziert (33) nach 13.1.1, während sich (34) aus 13.1.4 ergibt. Es bleibt der Nachweis von (35). Wir setzen dazu $\eta_i := \lambda_i - \varrho$ $(i = 1, \ldots, n)$ und numerieren die Eigenwerte $\{\lambda_j\}$ von $\boldsymbol{A}$ so, daß

$$|\eta_1| = |\lambda_1 - \varrho| \leqq |\eta_2| = |\lambda_2 - \varrho| \leqq \cdots \leqq |\eta_n| = |\lambda_n - \varrho| \tag{36}$$

gilt. Dann ist

$$\gamma_1 = |\eta_2|, \tag{37}$$

und (35) liest sich als

$$|\eta_1\eta_2| \leqq \|\boldsymbol{r}\|^2. \tag{38}$$

Mit der Eigenwertzerlegung (4) und $\boldsymbol{z} = \boldsymbol{U}^{\mathsf{T}}\boldsymbol{u}$ ergibt sich nun

$$\boldsymbol{r} = (\boldsymbol{A} - \varrho\boldsymbol{I})\,\boldsymbol{u} = \boldsymbol{U}(\boldsymbol{\Lambda} - \varrho\boldsymbol{I})\,\boldsymbol{U}^{\mathsf{T}}\boldsymbol{u} = \boldsymbol{U}(\boldsymbol{\Lambda} - \varrho\boldsymbol{I})\,\boldsymbol{z}, \tag{39}$$

wegen der Orthogonalität von $\boldsymbol{U}$ also

$$\|\boldsymbol{r}\|^2 = \|(\boldsymbol{\Lambda} - \varrho\boldsymbol{I})\,\boldsymbol{z}\|^2 = \sum_{i=1}^{n} \eta_i^2 z_i^2. \tag{40}$$

Die Orthogonalitätsrelation (32) geht in

$$0 = \boldsymbol{u}^{\mathsf{T}}\boldsymbol{r} = \boldsymbol{z}^{\mathsf{T}}(\boldsymbol{\Lambda} - \varrho\boldsymbol{I})\,\boldsymbol{z} = \sum_{i=1}^{n} \eta_i z_i^2 \tag{41}$$

über, und außerdem gilt

$$1 = \boldsymbol{u}^{\mathsf{T}}\boldsymbol{u} = \boldsymbol{z}^{\mathsf{T}}\boldsymbol{z} = \sum_{i=1}^{n} z_i^2. \tag{42}$$

Wenn (41) mit $\sigma(|\eta_2| - |\eta_1|)$, $\sigma := \operatorname{sgn}(\eta_1)$, und (42) mit $-|\eta_1\eta_2|$ multipliziert und zu (40) addiert wird, folgt

$$\begin{aligned} \|\boldsymbol{r}\|^2 - |\eta_1\eta_2| &= \sum_{i=1}^{n} \{\eta_i^2 + \eta_i\sigma(|\eta_2| - |\eta_1|) - |\eta_1\eta_2|\}\, z_i^2 \\ &= \sum_{i=1}^{n} (\eta_i - \eta_1)\,(\eta_i + \sigma\,|\eta_2|)\, z_i^2 \geqq 0, \end{aligned} \tag{43}$$

also die zu beweisende Ungleichung (38). Man beachte, daß die beiden Klammerausdrücke in (43) wegen (36) für jedes i dasselbe Vorzeichen haben, das Produkt also nicht negativ ist. □

Die dem Vektor $\boldsymbol{u}$ mit $\|\boldsymbol{u}\| = 1$ gemäß (31) zugeordnete optimale Eigenwert-Näherung $\varrho = \varrho(\boldsymbol{u}) = \boldsymbol{u}^{\mathsf{T}}\boldsymbol{A}\boldsymbol{u}$ wird *Rayleigh-Quotient* von $\boldsymbol{u}$ bezüglich $\boldsymbol{A}$ genannt; für jeden nicht notwendig normierten Vektor $\boldsymbol{v} \neq \boldsymbol{o}$ ist er durch

$$\varrho(\boldsymbol{v}) := \frac{\boldsymbol{v}^{\mathsf{T}}\boldsymbol{A}\boldsymbol{v}}{\boldsymbol{v}^{\mathsf{T}}\boldsymbol{v}} \tag{44}$$

definiert. Man beachte, daß $\varrho(\lambda\boldsymbol{v}) = \varrho(\boldsymbol{v})$ für alle $\lambda \neq 0$, $\boldsymbol{v} \neq \boldsymbol{o}$ gilt, also $\varrho(\boldsymbol{v})$ nur von der Richtung, nicht aber von Orientierung und Länge des Argumentvektors $\boldsymbol{v}$ abhängt.

Wir bemerken an dieser Stelle, daß die Schranken (34), (35) wie die Schranke (17) aus 13.1.4 meist nur theoretische Bedeutung haben, da realistische untere Schranken

für γ_j nur selten bekannt sind. Das folgende Zahlenbeispiel zeigt, daß sich solche Schranken beschaffen lassen, wenn λ_j genügend gut von den übrigen Eigenwerten λ_i getrennt ist und die λ_i genügend gut durch gewisse Rayleigh-Quotienten ϱ_i approximiert werden.

13.1.9. *Beispiel.* Betrachtet werde die Matrix

$$\boldsymbol{A} = \begin{pmatrix} 0.00 & 0.08 & 0.01 \\ 0.08 & 1.00 & 0.02 \\ 0.01 & 0.02 & 1.20 \end{pmatrix}.$$

Da die Diagonalelemente bis auf a_{11} dominant sind, bietet es sich an, die Einheitsvektoren $\boldsymbol{e}^1, \boldsymbol{e}^2, \boldsymbol{e}^3$ als Näherungen $\boldsymbol{u}$ für die gesuchten Eigenvektoren $\boldsymbol{u}^1, \boldsymbol{u}^2, \boldsymbol{u}^3$ zu wählen. Wenn ϱ_j den zu $\boldsymbol{u} = \boldsymbol{e}^j$ gehörenden Rayleigh-Quotienten und $\eta_j = \|\boldsymbol{r}^j\|$ die Norm des zugehörigen kürzesten Residuums bezeichnet, ergeben sich die Zahlenwerte

$$\begin{aligned} \boldsymbol{u} = \boldsymbol{e}^1\colon\ & \varrho_1 = 0.0, & \eta_1^2 &= 0.0065, & \eta_1 &= 0.0806, \\ \boldsymbol{u} = \boldsymbol{e}^2\colon\ & \varrho_2 = 1.0, & \eta_2^2 &= 0.0068, & \eta_2 &= 0.0825, \\ \boldsymbol{u} = \boldsymbol{e}^3\colon\ & \varrho_3 = 1.2, & \eta_3^2 &= 0.0005, & \eta_3 &= 0.0224. \end{aligned}$$

Aus (33) folgt dann

$$|\varrho_1 - \lambda_1| \leqq 0.0806, \qquad |\varrho_2 - \lambda_2| \leqq 0.0825, \qquad |\varrho_3 - \lambda_3| \leqq 0.0224 \tag{45}$$

bzw.

$$-0.0806 \leqq \lambda_1 \leqq 0.0806, \qquad 0.9175 \leqq \lambda_2 \leqq 1.0825, \qquad 1.1776 \leqq \lambda_3 \leqq 1.2224; \tag{46}$$

da die Einschließungsintervalle (46) disjunkt sind, muß in jedem genau ein Eigenwert λ_j liegen. Aus (46) ergeben sich die folgenden unteren Schranken für $\gamma_j = \min\{|\varrho_j - \lambda_i| : i \neq j\}$:

$$\gamma_1 \geqq 0.9175, \qquad \gamma_2 \geqq 0.1776, \qquad \gamma_3 \geqq 0.1175. \tag{47}$$

Die Abschätzung (35) liefert dann $|\varrho_j - \lambda_j| \leqq \eta_j^2/\gamma_j$, mit den angegebenen unteren Schranken (47) also

$$|\varrho_1 - \lambda_1| \leqq 0.0071, \qquad |\varrho_2 - \lambda_2| \leqq 0.038, \qquad |\varrho_3 - \lambda_3| \leqq 0.0043$$

und damit deutlich bessere Schranken als (45). □

Wir wenden uns abschließend der zweiten der oben gestellten Fragen zu, nämlich wie $\boldsymbol{u}$ mit $\|\boldsymbol{u}\| = 1$ zu wählen ist, damit $\{\mu, \boldsymbol{u}\}$ bei gegebener Eigenwertnäherung μ ein möglichst gutes Eigenpaar von $\boldsymbol{A}$ im Sinne eines minimalen Residuums wird.

13.1.10. *Aussage.* Die Matrix $\boldsymbol{A} \in \mathbf{S}^{n,n}$ besitze die Eigenwerte $\{\lambda_i\}$, und $\{\boldsymbol{u}^i\}$ sei eine orthonormale Basis zugehöriger Eigenvektoren. Es gelte $\mu \in \mathbf{R}$, und $j \in \{1, \ldots, n\}$ sei durch

$$\min\{|\lambda_i - \mu| : i = 1, \ldots, n\} = |\lambda_j - \mu| \tag{48}$$

definiert. Dann hat die Aufgabe

$$\|\boldsymbol{A}\boldsymbol{u} - \mu\boldsymbol{u}\| \to \text{Minimum! bei } \|\boldsymbol{u}\| = 1 \tag{49}$$

die Lösung $\boldsymbol{u} = \boldsymbol{u}^j$, und für das minimale Residuum gilt

$$\|\boldsymbol{A}\boldsymbol{u}^j - \mu\boldsymbol{u}^j\| = |\lambda_j - \mu|. \tag{50}$$

Beweis. Mit (3) und der Transformation $\boldsymbol{u} = \boldsymbol{U}\boldsymbol{z}$ läßt sich (49) äquivalent in der Form

$$\|\boldsymbol{A}\boldsymbol{z} - \mu\boldsymbol{z}\|^2 = \sum_{i=1}^{n} (\lambda_i - \mu)^2 z_i^2 \to \text{Minimum!} \quad \text{bei } \|\boldsymbol{z}\|^2 = \sum_{i=1}^{n} z_i^2 = 1 \tag{51}$$

schreiben, aus der die angegebene Lösung sofort abgelesen werden kann. □

Die Bestimmung eines optimalen $\boldsymbol{u} = \boldsymbol{u}^j$ läuft also darauf hinaus, einen normierten Eigenvektor zu demjenigen Eigenwert λ_j zu finden, der am dichtesten an μ liegt. Im Abschnitt 13.4 werden wir zeigen, daß diese Aufgabe näherungsweise sehr effektiv mittels der sog. inversen Iteration gelöst werden kann.

Übungsaufgaben

Ü 13.1.1. Es wird die Matrix $\boldsymbol{A} = \begin{pmatrix} \alpha & \beta \\ \beta & \gamma \end{pmatrix} \in \mathbf{S}^{2,2}$ und deren charakteristisches Polynom

$$\det(\boldsymbol{A} - \lambda\boldsymbol{I}) = \lambda^2 + c_1\lambda + c_0, \qquad c_1 := -(\alpha + \gamma), \qquad c_0 := \alpha\gamma - \beta^2$$

betrachtet.

(i) Man berechne $\{c_1, c_0\}$ in dezimaler Gleitpunktarithmetik mit acht Mantissenstellen für $\alpha := \gamma := 1$, $\beta := 7 \times 10^{-5}$ und bestimme $\{\lambda_1, \lambda_2\}$ als Nullstellen des so errechneten charakteristischen Polynoms. Man vergleiche die Ergebnisse mit den exakten Eigenwerten

$$\lambda_{\max}^* \doteq 1.00007, \qquad \lambda_{\min}^* \doteq 0.99993.$$

(ii) Man berechne λ_1, λ_2 für dieselben Daten und in derselben Arithmetik nach dem folgenden Algorithmus

V: $\xi := \alpha + \gamma, \qquad \delta := \alpha * \gamma - \beta^2, \qquad \omega := \sqrt{(\alpha - \gamma)^2 + 4 * \beta^2}$
if $\xi = 0$ **then** $[\lambda_{\max} := \omega/2;\ \lambda_{\min} := -\lambda_{\max}]$
else $[\lambda_{\max} := \operatorname{sgn}(\xi)\,(|\xi| + \omega)/2,\ \lambda_{\min} := \delta/\lambda_{\max}]$

Man überprüfe, daß (V) in exakter Arithmetik die Eigenwerte von $\boldsymbol{A}$ liefert, wobei $|\lambda_{\max}| \geqq |\lambda_{\min}|$ ist.

(iii) Man zeige, daß die nach (ii) berechneten Größen einen gemäß

$$|\lambda_{\max} - \lambda_{\max}^*| \leqq \nu(5\,|\lambda_{\max}^*|) = 5\nu\,\|\boldsymbol{A}\|_2,$$

$$|\lambda_{\min} - \lambda_{\min}^*| \leqq \nu(7\,|\lambda_{\min}^*| + |\lambda_{\max}^*|) \leqq 8\nu\,\|\boldsymbol{A}\|_2$$

beschränkten Fehler haben.

Ü 13.1.2. Es sei $\boldsymbol{A} = \boldsymbol{U} + \mathrm{i}\boldsymbol{V} \in \mathbf{C}^{n,n}$, $\boldsymbol{U}, \boldsymbol{V} \in \mathbf{R}^{n,n}$, eine *Hermitesche Matrix*, d. h., es gelte $\boldsymbol{A}^H = (\boldsymbol{A})^T = \boldsymbol{A}$, vgl. 1.2. Der Matrix $\boldsymbol{A} \in \mathbf{C}^{n,n}$ werde die Matrix

$$\boldsymbol{B} := \left(\begin{array}{c|c} \boldsymbol{U} & -\boldsymbol{V} \\ \hline \boldsymbol{V} & \boldsymbol{U} \end{array}\right) \in \mathbf{R}^{2n,2n} \tag{52}$$

zugeordnet. Man zeige:

(i) $\boldsymbol{U}$ ist symmetrisch, $\boldsymbol{V}$ ist schiefsymmetrisch, $\boldsymbol{B}$ ist symmetrisch.

(ii) Für $\boldsymbol{u}, \boldsymbol{v} \in \mathbf{R}^n$ werde $\boldsymbol{z} := \begin{pmatrix} \boldsymbol{u} \\ \boldsymbol{v} \end{pmatrix}$ und $\boldsymbol{z}' := \begin{pmatrix} -\boldsymbol{v} \\ \boldsymbol{u} \end{pmatrix}$ definiert, und es sei $\lambda \in \mathbf{R}$. Dann ist $\{\lambda, \boldsymbol{z}\}$ genau dann Eigenpaar von $\boldsymbol{B}$, wenn $(\lambda, \boldsymbol{z}')$ ein Eigenpaar ist.

(iii) Für $\lambda \in \mathbf{R}$ und $\boldsymbol{u}, \boldsymbol{v} \in \mathbf{R}^n$ ist $\{\lambda, \boldsymbol{u} + \mathrm{i}\boldsymbol{v}\}$ genau dann Eigenpaar von $\boldsymbol{A}$, wenn $\{\lambda, \boldsymbol{z}\}$ und $\{\lambda, \boldsymbol{z}'\}$ Eigenpaare von $\boldsymbol{B}$ sind.

Das Eigenwertproblem der komplexen Hermiteschen Matrix $\boldsymbol{A}$ kann also auf das Eigenwertproblem der reellen symmetrischen Matrix $\boldsymbol{B}$ doppelter Dimension zurückgeführt werden. Vom Aufwand her ist jedoch die direkte Lösung des Eigenwertproblems für $\boldsymbol{A}$ in komplexer Arithmetik i. allg. nur halb so teuer wie bei Übergang zu $\boldsymbol{B}$, da Speicherplatz und Operationszahlen wie $\sim K_1 n^2$ bzw. $\sim K_2 n^3$ von n abgängen.

Ü 13.1.3. Man zeige: Die Aufgabe (49) ist äquivalent zu

$$\|\delta \boldsymbol{A}\| \to \text{Minimum!}$$

bei

$$\delta \boldsymbol{A} \in \mathbf{S}^{n,n}, \qquad \boldsymbol{u} \in \mathbf{R}^n \quad \text{mit} \quad (\boldsymbol{A} + \delta \boldsymbol{A})\,\boldsymbol{u} = \mu \boldsymbol{u} \quad \text{und} \quad \|\boldsymbol{u}\| = 1, \tag{53}$$

und die minimalen Zielfunktionswerte stimmen überein.

13.2. Das Jacobi-Verfahren

Das Jacobi-Verfahren ist das älteste numerische Verfahren zur Lösung des symmetrischen Eigenwertproblems und zugleich das erste „moderne" im Sinne der Eignung für Computerrechnung. Obwohl es inzwischen effektivere Algorithmen gibt, findet dieses Verfahren wegen seiner Eleganz, Einfachheit und numerischen Gutartigkeit auch heute noch seine Anwendungsgebiete.

A. *Das Basisverfahren*

Die Grundidee des Verfahrens besteht darin, die Matrix $\boldsymbol{A} \in \mathbf{S}^{n,n}$ durch eine Folge orthogonaler Ähnlichkeitstransformationen

$$\boldsymbol{A}_1 := \boldsymbol{A}, \qquad \boldsymbol{A}_{k+1} := \boldsymbol{G}_k^{\mathsf{T}} \boldsymbol{A}_k \boldsymbol{G}_k \qquad (k = 1, 2, \ldots) \tag{1}$$

mittels geeignet gewählter Drehungsmatrizen $\boldsymbol{G}_k := \boldsymbol{G}_{pq}$ sukzessive auf Diagonalform zu bringen. Aus (1) folgt

$$\boldsymbol{A}_{k+1} = \boldsymbol{V}_{k+1}^{\mathsf{T}} \boldsymbol{A} \boldsymbol{V}_{k+1} \tag{2}$$

mit den orthogonalen Transformationsmatrizen

$$\boldsymbol{V}_1 := \boldsymbol{I}, \qquad \boldsymbol{V}_{k+1} := \boldsymbol{G}_1 \cdots \boldsymbol{G}_{k-1} \boldsymbol{G}_k = \boldsymbol{V}_k \boldsymbol{G}_k, \tag{3}$$

insbesondere sind alle $\boldsymbol{A}_k$ wie $\boldsymbol{A}$ symmetrisch und haben dieselben Eigenwerte $\{\lambda_i\}$.

Die im k-ten Schritt verwendete Givens-Drehung hängt von den Indizes p, q und den Drehungsparametern c, s mit $c^2 + s^2 = 1$ ab, siehe 3.3.B. Diese Größen sollen so festgelegt werden, daß $\boldsymbol{A}_{k+1} = (a_{ij}^{(k+1)})$ einer Diagonalmatrix möglichst nahe kommt. Als Maß für die Abweichung bietet sich in natürlicher Weise die Zahl

$$\omega_k := \omega(\boldsymbol{A}_k) := \sum_{i<j} [a_{ij}^{(k)}]^2 \tag{4}$$

an. Wenn

$$\boldsymbol{M}_k := \operatorname{diag}(a_{11}^{(k)}, \ldots, a_{nn}^{(k)}) \quad \text{und} \quad \boldsymbol{R}_k := \boldsymbol{A}_k - \boldsymbol{M}_k \tag{5}$$

gesetzt wird, ergibt sich aus (2) und (4)

$$\boldsymbol{M}_k = \boldsymbol{A}_k - \boldsymbol{R}_k = \boldsymbol{V}_k^{\mathsf{T}} (\boldsymbol{A} + \delta \boldsymbol{A}_k)\, \boldsymbol{V}_k \tag{6}$$

mit

$$\delta \boldsymbol{A}_k := -\boldsymbol{V}_k \boldsymbol{R}_k \boldsymbol{V}_k^{\mathsf{T}} \in \mathbf{S}^{n,n}, \qquad \|\delta \boldsymbol{A}_k\|_F = \|\boldsymbol{R}_k\|_F = \sqrt{2\omega_k}. \tag{7}$$

Die Diagonalelemente von $\boldsymbol{A}_k$ sind also die exakten Eigenwerte der gestörten Matrix $\boldsymbol{A} + \delta \boldsymbol{A}_k$, und die Spalten von $\boldsymbol{V}_k$ sind die zugehörigen Eigenvektoren. Nach 13.1.3 gilt dann

$$|a_{ii}^{(k)} - \lambda_{l(i)}| \leqq \sqrt{\sum_{i=1}^{n} (a_{ii}^{(k)} - \lambda_{l(i)})^2} \leqq \|\delta \boldsymbol{A}_k\|_F = \sqrt{2\omega_k}. \tag{8}$$

Für jedes k ist dabei $\{l(1), \ldots, l(n)\} = \{l(1, k), \ldots, l(n, k)\}$ eine Permutation derart, daß die $\{\lambda_{l(i)}\}$ in derselben Weise wie die $\{a_{ii}^{(k)}\}$ geordnet sind. Die $\{a_{ii}^{(k)}\}$ approximieren also die entsprechend geordneten Eigenwerte von $\boldsymbol{A}$ mit einem durch $\sqrt{2\omega_k}$ beschränkten absoluten Fehler.

Mit dem Kriterium (4) lautet die Aufgabe dann, ein $\boldsymbol{G}_k = \boldsymbol{G}_{pq}(c, s)$ so zu wählen, daß

$$\omega_k - \omega_{k+1} =: \Phi_k(p, q, c, s) \to \text{Maximum!} \tag{9}$$

bei

$$1 \leqq p < q \leqq n \quad \text{und} \quad c^2 + s^2 = 1$$

gilt. Zur Vereinfachung der Schreibweise lassen wir wie üblich den Index k weg und kennzeichnen die zum Index $k + 1$ gehörenden Größen durch einen Querstrich, also $\boldsymbol{A} = \boldsymbol{A}_k$, $\bar{\boldsymbol{A}} = \boldsymbol{A}_{k+1}$ usw. Bei der Berechnung von $\bar{\boldsymbol{A}} = \boldsymbol{G}_{pq}^{\mathsf{T}} \boldsymbol{A} \boldsymbol{G}_{pq}$ aus $\boldsymbol{A}$ ändern sich nur die Elemente der Zeilen und Spalten p und q von $\boldsymbol{A}$, und zwar nach den Vorschriften

$$\begin{aligned} \bar{a}_{pp} &:= c^2 a_{pp} - 2cs a_{pq} + s^2 a_{qq}, \\ \bar{a}_{qq} &:= s^2 a_{pp} + 2cs a_{pq} + c^2 a_{qq}, \\ \bar{a}_{pq} &:= \bar{a}_{qp} = cs(a_{pp} - a_{qq}) + (c^2 - s^2)\, a_{pq} \end{aligned} \tag{10}$$

für die Elemente in Kreuzungspunkten der Zeilen und Spalten p, q und nach

$$\begin{aligned} \bar{a}_{ip} &:= \bar{a}_{pi} := c a_{ip} - s a_{iq}, \\ \bar{a}_{iq} &:= \bar{a}_{qi} := s a_{ip} + c a_{iq} \end{aligned} \qquad (i \neq p, q) \tag{11}$$

für die übrigen Elemente. Wegen der Orthogonalität von $\boldsymbol{G}_{pq}$ folgt aus (11)

$$(\bar{a}_{ip})^2 + (\bar{a}_{iq})^2 = (a_{ip})^2 + (a_{iq})^2 \qquad (i \neq p, q).$$

Die Quadratsumme der Nichtdiagonalelemente wird also nur durch die Elemente in den Positionen $\{p, q\}$ und $\{q, p\}$ verändert, so daß die Änderung Φ durch

$$\Phi(p, q, c, s) = \omega_k - \omega_{k+1} = \omega(\boldsymbol{A}) - \omega(\bar{\boldsymbol{A}}) = (a_{pq})^2 - (\bar{a}_{pq})^2 \tag{12}$$

gegeben ist. Bei festem $\{p, q\}$ ist die Reduktion von ω_k maximal, wenn $\{c, s\}$ so gewählt wird, daß

$$\bar{a}_{pq} = 0 \tag{13}$$

gilt. Dies führt auf

$$\omega(\bar{\boldsymbol{A}}) = \omega(\boldsymbol{A}) - (a_{pq})^2. \tag{14}$$

Eine Drehung $\boldsymbol{G}_{pq}$, die (13) realisiert, heißt *Jacobi-Drehung* in der durch $\{p, q\}$ festgelegten Ebene. Die zugehörigen Parameter $\{c, s\}$ müssen dazu so gewählt werden, daß

$$cs(a_{pp} - a_{qq}) + (c^2 - s^2)\, a_{pq} = 0 \tag{15}$$

gilt. Wir setzen im folgenden $a_{pq} \neq 0$ voraus, denn andernfalls läßt sich $\omega(\boldsymbol{A})$ nach (12) überhaupt nicht reduzieren. Dann kann in (15) nicht $c = 0$ gelten, so daß

$$t := s/c \tag{16}$$

als neue Größe eingeführt werden kann. Damit geht (15) in die quadratische Gleichung

$$t^2 + 2\delta t - 1 = 0 \tag{17}$$

mit

$$\delta := \frac{a_{qq} - a_{pp}}{2a_{pq}} \tag{18}$$

über. Die betragskleinste Lösung von (17) kann nach der Vorschrift

$$t := \begin{cases} 1/\left(|\delta| + \sqrt{1 + \delta^2}\right) & \text{für} \quad \delta \geqq 0, \\ -1/\left(|\delta| + \sqrt{1 + \delta^2}\right) & \text{für} \quad \delta < 0 \end{cases} \tag{19}$$

numerisch stabil berechnet werden, siehe Ü 13.2.3. Aus (16) folgt dann

$$c := 1/\sqrt{1 + t^2}, \qquad s := tc, \tag{20}$$

wenn c positiv gewählt wird. Unter Verwendung von t, s und

$$\tau := s/(1 + c) \tag{21}$$

lassen sich die Transformationsformeln (10), (11) in der in exakter Arithmetik äquivalenten, für Computerrechnung jedoch wesentlich günstigeren Korrekturform

$$\begin{aligned} \bar{a}_{pp} &:= a_{pp} - ta_{pq}, \\ \bar{a}_{qq} &:= a_{qq} + ta_{pq}, \\ \bar{a}_{pq} &:= \bar{a}_{qp} := 0 \end{aligned} \tag{22}$$

bzw.

$$\begin{aligned} \bar{a}_{ip} &:= \bar{a}_{pi} := a_{ip} - s(a_{iq} + \tau a_{ip}), \\ \bar{a}_{iq} &:= \bar{a}_{qi} := a_{iq} + \tau(a_{ip} + \bar{a}_{ip}) \end{aligned} \qquad (i \neq p, q) \tag{23}$$

schreiben. Gegenüber der sonst üblichen Darstellung

$$\bar{a}_{iq} = \bar{a}_{qi} = a_{iq} + s(a_{ip} - \tau a_{iq})$$

führt die zweite Gleichung (23) auf die Einsparung einer Multiplikation, vgl. Ü 3.3.6.

Wenn die Wahl des Pivotelementes $a_{pq} \neq 0$ noch offengelassen wird, ergibt sich das folgende Basisverfahren.

13.2.1. *Jacobi-Basisverfahren*

S0 (Initialisierung): Setze $\boldsymbol{A}_1 := \boldsymbol{A}$, $k := 1$, berechne $\omega_1 := \omega(\boldsymbol{A})$, wähle $\varepsilon > 0$

S1 (Abbruchtest): **if** $2\omega_k \leqq \varepsilon^2$ **then stop**

S2 (Pivotwahl): Wähle $p = p(k)$, $q = q(k)$ mit $1 \leqq p < q \leqq n$ und $a_{pq}^{(k)} \neq 0$.

S3 (Bestimmung der Drehungsparameter): Bestimme $t = t_k$, $c = c_k$, $s = s_k$, $\tau = \tau_k$ nach (18) bis (21)

S4 (Berechnung von A_{k+1}): Berechne $\boldsymbol{A}_{k+1} := \bar{\boldsymbol{A}}$ gemäß (22), (23) aus $\boldsymbol{A} := \boldsymbol{A}_k$

S5 (Berechnung von ω_{k+1}): Setze $\omega_{k+1} := \omega_k - [a_{pq}^{(k)}]^2$

S6 (Vorbereitung des neuen Schrittes): Setze $k := k + 1$, **goto** S1

Aufwand: [$\sim n$(4 ops + 3 opm) + 2 opr] pro Schritt

13.2.2. *Bemerkung.* (i) Das Verfahren kann in situ — und zwar auf dem Platz nur eines Dreiecks einschließlich der Diagonalen — realisiert werden.

(ii) Wenn der Abbruchtest S1 nach k Schritten erfüllt ist, gilt wegen (8)

$$|\lambda_{l(i)} - a_{ii}^{(k+1)}| \leqq \varepsilon. \tag{24}$$

Im allgemeinen sollte $\varepsilon \geqq \nu \|\boldsymbol{A}\|_F$ gewählt werden, da eine höhere Genauigkeit ohnehin nicht erreichbar ist, vgl. 13.1.B. Mit feineren Abbruchtests läßt sich unter Umständen für die betragskleinen Eigenwerte eine höhere relative Genauigkeit erzielen.

(iii) Wenn ω_k in die Größenordnung von $k\nu\omega_1$ kommt, sollte ω_{k+1} nicht durch Aufdatierung von ω_k, sondern direkt aus $\boldsymbol{A}_{k+1}$ berechnet werden. Es genügt, dies nur einmal für ein geeignetes $k = k_0$ durchzuführen.

(iv) Im Konvergenzfall $\omega_k \to 0$ gilt $|a_{ij}^{(k)}| \to 0$ $(i \neq j)$. Nach (18) wird dann $|\delta| = |\delta_k|$ häufig groß sein. Wegen

$$|\tau| \leqq \frac{|s|}{1 + \sqrt{2}/2} < |s| \leqq |t| \leqq \frac{1}{1 + |\delta|} \tag{25}$$

sind dann τ_k, s_k und t_k betragsklein, so daß die Korrekturformeln (22), (23) wesentliche Genauigkeitsvorteile gegenüber den naiven Realisierungen (10), (11) bringen, vgl. Ü 13.2.4. Der bei der Berechnung der Diagonalelemente gemäß (22) erzeugte Rundungsfehler kann weiter verkleinert werden, wenn die auftretenden Korrekturen $\pm t a_{pq}$ nicht sofort zu a_{pp} bzw. a_{qq} geschlagen, sondern etwa über $N = n(n-1)/2$ aufeinanderfolgende Schritte in einem Hilfsfeld akkumuliert und erst dann insgesamt addiert werden.

(v) Wenn das vollständige Eigenwertproblem zu lösen ist, d. h., wenn neben den $\{\lambda_i\}$ auch die zugehörigen Eigenvektoren $\{\boldsymbol{u}^i\}$ berechnet werden sollen, müssen die

Transformationsmatrizen V_k gemäß (3) mit berechnet werden, vgl. (6). Dies ist auf dem Platz einer (n, n)-Matrix in zu (23) analoger Weise möglich und kostet zusätzliche $\sim n(3 \text{ ops} + 3 \text{ opm})$ pro Schritt, vgl. Ü 3.3.6. □

B. *Pivotisierungsstrategien, Konvergenz, Fehleranalyse*

Zur Realisierung des Basisverfahrens 13.2.1 muß der Teilschritt S 2 noch spezifiziert werden, d. h., es muß festgelegt werden, welche Pivotindizes $p = p(k)$, $q = q(k)$ mit $a_{pq}^{(k)} \neq 0$ im k-ten Schritt gewählt werden sollen. Aus (14) ist abzulesen, daß die bereits von JACOBI verwendete *Maximalpivotwahl*

$$\{p, q\} \text{ so, daß } |a_{pq}^{(k)}| = \max \{|a_{ij}^{(k)}| : 1 \leqq i < j \leqq n\} \tag{26}$$

die größte Reduktion von ω_k bringt, d. h., das durch (26) festgelegte Paar $\{p, q\}$ und die zugehörigen Jacobi-Parameter $\{c, s\}$ lösen (9). Es gilt dann

$$\omega_k = \omega(\boldsymbol{A}_k) \leqq N(a_{pq}^{(k)})^2, \qquad N := n(n-1)/2, \tag{27}$$

also

$$\omega_{k+1} = \omega_k - (a_{pq}^{(k)})^2 \leqq (1 - 1/N)\, \omega_k = \varkappa \omega_k, \qquad \varkappa := 1 - 1/N < 1. \tag{28}$$

Das durch (26) festgelegte Verfahren wird *klassisches Jacobi-Verfahren* genannt. Die maximale Reduktion von ω_k muß bei diesem Verfahren jedoch mit dem Aufwand von $N \sim n^2/2$ Lese- und Vergleichsoperationen pro Schritt zur Suche von $a_{pq}^{(k)}$ bezahlt werden. Es entsteht daher die Frage, ob sich dieser Pivotisierungsaufwand reduzieren läßt.

Als billigste Variante bietet sich an, das Pivot zyklisch alle Positionen $\{i, j\}$, $i < j$, durchlaufen zu lassen, etwa spaltenweise in der durch

$$\{1,2\},\ \{1,3\},\ \{2,3\},\ \ldots,\ \{j-2, j-1\},\ \{1, j\},\ \{2, j\},\ \ldots,\ \{j-1, j\},\ \{1, j+1\},\ \ldots,\ \{n-1, n\} \tag{29}$$

gegebenen Reihenfolge. Die zu den N Pivotpositionen (29) gehörenden Jacobi-Schritte werden ein *Zyklus* (engl. „sweep") genannt, und das zugehörige Verfahren heißt *zyklisches Jacobi-Verfahren*. Selbstverständlich wird dabei die k-te Drehung übergangen, wenn $a_{pq}^{(k)} = 0$ ist. Wir bemerken dabei ausdrücklich, daß nach einem Zyklus keineswegs alle Nichtdiagonalelemente annulliert sind, denn die erzeugten Nullen werden in den nachfolgenden Schritten i. allg. zerstört.

Als Nachteil des zyklischen Jacobi-Verfahrens muß angesehen werden, daß ein Schritt mit einem Pivot, für das $0 < (a_{pq}^{(k)})^2 \ll \omega_k$ gilt, trotz des Aufwands von $O(n)$ Operationen keine bemerkbare Reduktion von ω_k bringt. Durch Einführung eines sog. *Schwellwertes* (engl. „threshold") für die zulässige Pivotgröße kann dieser Effektivitätsverlust, der vor allem in den ersten Zyklen auftritt, vermieden werden. Bei einem solchen *Schwellwert-Jacobi-Verfahren* werden die Nichtdiagonalelemente wie beim zyklischen Verfahren etwa in der Reihenfolge (29) durchlaufen, aber die Drehung wird nur ausgeführt, wenn der Pivotbetrag nicht unter dem Schwellwert liegt. Falls das quadratische Mittel $\sqrt{\omega_k/N}$ der Nichtdiagonalelemente als Schwell-

wert gewählt wird, führt dies auf die Festlegung

$$\{p, q\} \text{ erstes Folgepaar von } \{p(k-1), q(k-1)\} \text{ im Zyklus (29)}$$
$$\text{mit } (a_{pq}^{(k)})^2 \geqq \omega_k/N. \tag{30}$$

Die Abschätzungen (27), (28) sind dann ebenfalls erfüllt. Für Hinweise auf andere Schwellwertstrategien siehe B 13.2.

Die folgende Aussage zeigt, daß (28) die Konvergenz von $\boldsymbol{A}_k$ gegen eine Diagonalmatrix $\boldsymbol{M}$ impliziert.

13.2.3. *Aussage.* Das Verfahren 13.2.1 werde in exakter Arithmetik mit $\varepsilon = 0$ und einer Pivotwahl durchgeführt, für die (28) mit einem $\varkappa < 1$ erfüllt ist, also etwa als klassisches oder Schwellwert-Jacobi-Verfahren. Dann gilt

$$\lim \boldsymbol{A}_k = \boldsymbol{M} = \operatorname{diag}(\lambda_{m(1)}, \ldots, \lambda_{m(n)}) \tag{31}$$

mit einer geeigneten Permutation $\{m(i)\}$, und die Konvergenzgeschwindigkeit ist durch die Ungleichungen

$$\left.\begin{aligned} &i \neq j,\ k \geqq 1\colon\ |a_{ij}^{(k)}| \leqq \|\boldsymbol{R}_k\|_F \\ &i = j,\ k \geqq k_0\colon\ |a_{ii}^{(k)} - \lambda_{m(i)}| \leqq \|\boldsymbol{M}_k - \boldsymbol{M}\|_F \end{aligned}\right\} \leqq \sqrt{2\omega_k} \leqq (\sqrt{\varkappa})^{k-1}\sqrt{2\omega_1} \tag{32}$$

charakterisiert, wobei $\boldsymbol{R}_k$ und $\boldsymbol{M}_k$ wie in (5) und k_0 ein genügend hoher Index ist.

Beweis. Für $i \neq j$ folgt (32) sofort aus (28). Für $i = j$ setzen wir $\mu_k := a_{ii}^{(k)}$. Aus (22), (25) und (32), Fall $i \neq j$, ergibt sich dann mit $\varrho := \sqrt{\varkappa}$

$$|\mu_{k+1} - \mu_k| \leqq |t_k|\, |a_{p(k)q(k)}^{(k)}| \leqq \varrho^{k-1}\sqrt{2\omega_1},$$

also für jedes $r = 0, 1, \ldots$

$$\begin{aligned} |\mu_{k+r+1} - \mu_k| &\leqq |\mu_{k+r+1} - \mu_{k+r}| + \cdots + |\mu_{k+1} - \mu_k| \leqq [\varrho^{k+r-1} + \cdots + \varrho^{k-1}]\sqrt{2\omega_1} \\ &\leqq \varrho^{k-1}[1 + \varrho + \varrho^2 + \cdots]\sqrt{2\omega_1} = [\varrho^{k-1}/(1-\varrho)]\sqrt{2\omega_1}. \end{aligned}$$

Die Folge $\{a_{ii}^{(k)}\} = \{\mu_k\}$ ist somit eine Cauchy-Folge und daher konvergent gegen ein $\mu =: \mu_i$, d. h., es gilt (31) mit $\boldsymbol{M} = \operatorname{diag}(\mu_i)$. Die μ_i müssen dann die Eigenwerte von $\boldsymbol{A}$ sein, so daß $\mu_i = \lambda_{m(i)}$ mit einer geeigneten Permutation $\{m(i)\}$ gilt. Für genügend großes k kann daher die in (8) vorkommende und vom Index k abhängende Permutation $\{l_k(i)\}$ identisch zu $\{m(i)\}$ gewählt werden, so daß (32) auch für $i = j$ gilt. Man beachte, daß aus (8) allein nicht die Konvergenz der Diagonalelemente folgt, da diese beim Übergang von k zu $k+1$ z. B. ihre Plätze wechseln können, vgl. Ü 13.2.2. □

13.2.4. *Bemerkung.* (i) Ungleichung (32) besagt, daß die Elemente von $\boldsymbol{A}_k$ mindestens linear mit dem Konvergenzfaktor $\sqrt{\varkappa} < 1$ konvergieren. Dies ist eine theoretisch nützliche, praktisch aber schwache Aussage, da $\varkappa$ nahe bei 1 liegt. Soll etwa aus (32) abgeschätzt werden, nach wieviel vollen Zyklen r bzw. Schritten $k = Nr$ das Abbruchkriterium $2\omega_{k+1} \leqq \varepsilon^2 := \varepsilon_{\text{rel}}^2 \|\boldsymbol{A}\|_F^2$ mit $\varepsilon_{\text{rel}} \geqq \nu$ erfüllt ist, so ergibt sich wegen $2\omega_{k+1} \leqq 2\omega_1\varkappa^k \leqq \|\boldsymbol{A}\|_F^2\,\varkappa^k$ die Bedingung $\varkappa^k \leqq \varepsilon_{\text{rel}}^2$, also $k = Nr \leqq (-2\ln\varepsilon_{\text{rel}})/(-\ln\varkappa)$ und wegen $-\ln\varkappa \geqq 1 - \varkappa = 1/N$ schließlich

$$r = k/N \geqq 2\ln(1/\varepsilon_{\text{rel}}). \tag{33}$$

Im Fall $\varepsilon_{\text{rel}} := 10^{-6}$ bzw. 10^{-12} liefert dies die Schranke $r_{\max} = 28$ bzw. 55 für die Anzahl der benötigten Zyklen. Umfangreiche numerische Experimente zeigen jedoch, daß auch für sehr kleines ε_{rel} und ν der Abbruchtest fast immer nach vier und praktisch nie später als nach zehn Zyklen erfüllt ist, also

$$r = k/N \leqq 4 \ldots 10 \tag{34}$$

unabhängig von ε gilt.

(ii) Das in (i) beschriebene reale Konvergenzverhalten ist ein typisches Kennzeichen sog. überlinearer Konvergenz, die in Form der *asymptotisch quadratischen Konvergenz* beim Jacobi-Verfahren in der Tat vorliegt: Falls $\eta_k := \max\{|a_{ij}^{(k)}| : i < j\}$ genügend klein, also k genügend groß ist, gilt

$$\eta_{k+N} \leqq C(\eta_k)^2 \tag{35}$$

mit einer Konstanten $C > 0$. Der Beweis dieser Aussage ist kompliziert und kann hier nicht gegeben werden, siehe B 13.2; die wesentlichen Ideen sind in Ü 13.2.5 zu finden. □

Bei Computerrechnung und Abbruch nach $k = Nr$ Schritten wird der durch (32) beschriebene Abbruchfehler durch die Rundungsfehler überlagert. Eine detaillierte Fehleranalyse zeigt, daß zu der berechneten Matrix $\boldsymbol{A}_{k+1}$ eine Störung $\delta\bar{\boldsymbol{A}}_{k+1} \in \mathbf{S}^{n,n}$ mit

$$\boldsymbol{A}_{k+1} = \boldsymbol{V}_{k+1}^{\mathsf{T}}(\boldsymbol{A} + \delta\bar{\boldsymbol{A}}_{k+1})\,\boldsymbol{V}_{k+1}, \quad \|\delta\bar{\boldsymbol{A}}_{k+1}\|_F \leqq \nu F\,\|\boldsymbol{A}\|_F, \quad F \leqq 18n^{3/2}r \tag{36}$$

existiert, wobei $\boldsymbol{V}_{k+1} := \boldsymbol{G}_1^*\boldsymbol{G}_2^* \cdots \boldsymbol{G}_k^*$ exakt orthogonal und $\boldsymbol{G}_k^*$ die der berechneten Matrix $\boldsymbol{A}_k$ zugeordnete exakte Jacobi-Drehung ist. Auf Grund der speziellen Struktur des Verfahrens und bei sorgfältiger Implementierung gemäß 13.2.1 und 13.2.2(iv) ist jedoch die äquivalente Störung wesentlich kleiner als durch (36) ausgedrückt wird. Praktisch gilt fast immer

$$F \leqq k/n \sim (n/2)\,r, \quad \text{und meistens ist sogar } F \leqq 10, \tag{37}$$

siehe B 13.2. Wenn das berechnete $\boldsymbol{A}_{k+1}$ analog zu (5), (6) dargestellt wird, ergibt sich aus (36)

$$\begin{aligned}\boldsymbol{M}_{k+1} &= \boldsymbol{A}_{k+1} - \boldsymbol{R}_{k+1} = \boldsymbol{V}_{k+1}^{\mathsf{T}}(\boldsymbol{A} + \delta\bar{\boldsymbol{A}}_{k+1} - \boldsymbol{V}_{k+1}\boldsymbol{R}_{k+1}\boldsymbol{V}_{k+1}^{\mathsf{T}})\,\boldsymbol{V}_{k+1} \\ &=: \boldsymbol{V}_{k+1}^{\mathsf{T}}(\boldsymbol{A} + \delta\hat{\boldsymbol{A}}_{k+1})\,\boldsymbol{V}_{k+1}\end{aligned} \tag{38}$$

mit

$$\|\delta\hat{\boldsymbol{A}}_{k+1}\|_F \leqq \|\delta\bar{\boldsymbol{A}}_{k+1}\|_F + \|\boldsymbol{R}_{k+1}\|_F \leqq \nu F\,\|\boldsymbol{A}\|_F + \sqrt{2\omega_{k+1}} \leqq (\nu F + \varepsilon_{\text{rel}})\,\|\boldsymbol{A}\|_F. \tag{39}$$

Dies führt auf das folgende Ergebnis, bei dem sowohl Abbruchfehler als auch Rundungsfehler erfaßt sind:

13.2.5. *Fehleranalyse.* Das Jacobi-Verfahren werde in Computerarithmetik mit $\varepsilon = \varepsilon_{\text{rel}}\,\|\boldsymbol{A}\|_F \geqq \nu\,\|\boldsymbol{A}\|_F$ und einer Pivotwahl, für die (28) gilt, durchgeführt. Dann bricht das Verfahren nach $k = Nr$ Schritten ab, wobei $r \leqq 4 \ldots 10$ ist, und die

berechneten Diagonalelemente $\mu_i = a_{ii}^{(k+1)}$ von $\boldsymbol{A}_{k+1}$ sind die exakten Eigenwerte einer Matrix

$$\boldsymbol{A} + \delta\boldsymbol{A} \quad \text{mit} \quad \delta\boldsymbol{A} \in \mathbf{S}^{n,n}, \qquad \|\delta\boldsymbol{A}\|_F \leqq (\nu F + \varepsilon_{\text{rel}}) \|\boldsymbol{A}\|_F \tag{40}$$

und F nach (36) bzw. (37).

13.2.6. *Bemerkung.* (i) Aus (40) folgt nach 13.1.3 die zu (8) analoge Abschätzung

$$|a_{ii}^{(k+1)} - \lambda_{l(i)}| \leqq \sqrt{\sum_{i=1}^{n} (a_{ii}^{(k+1)} - \lambda_{l(i)})^2} \leqq \|\delta\boldsymbol{A}\|_F \leqq (\nu F + \varepsilon_{\text{rel}}) \|\boldsymbol{A}\|_F \tag{41}$$

mit einer geeigneten Permutation $\{l(i)\}$, die der Ordnung der $\{a_{ii}^{(k+1)}\}$ entspricht.

(ii) *Für* $\varepsilon_{\text{rel}} := \nu$ *besagt* (40), *daß das Jacobi-Verfahren zur Berechnung der Eigenwerte von* $\boldsymbol{A}$ *ein numerisch gutartiger Prozeß ist, und* (41) *drückt die numerische Stabilität aus.*

(iii) Aus (38) folgt, daß die Spalten $\boldsymbol{v}^j$ von $\boldsymbol{V} := \boldsymbol{V}_{k+1}$ die zu den $\mu_j = a_{jj}^{(k+1)}$ gehörenden orthonormalen Eigenvektoren von $\boldsymbol{A} + \delta\boldsymbol{A}$ sind. Wenn $\tilde{\boldsymbol{V}} := \tilde{\boldsymbol{V}}_{k+1}$ die gemäß (3) berechnete Näherung für $\boldsymbol{V}$ bezeichnet, gilt

$$\|\tilde{\boldsymbol{V}} - \boldsymbol{V}\|_F \leqq \nu F_1 \quad \text{mit} \quad F_1 \leqq 6n^2, \tag{42}$$

d. h., $\boldsymbol{V}$ wird durch das berechnete $\tilde{\boldsymbol{V}}$ ausreichend genau approximiert. Insbesondere ist $\tilde{\boldsymbol{V}}$ wegen

$$\|\tilde{\boldsymbol{V}}^{\mathrm{T}}\tilde{\boldsymbol{V}} - \boldsymbol{I}\|_2 \leqq \|\tilde{\boldsymbol{V}} - \boldsymbol{V}\|_2 (1 + 2\|\tilde{\boldsymbol{V}} - \boldsymbol{V}\|_2) \leqq 2\nu F_1 \tag{43}$$

ausreichend orthogonal.

(iv) Falls der Eigenwert $\lambda_{l(i)} \approx a_{ii}^{(k)}$ von $\boldsymbol{A}$ einfach und genügend von den übrigen Eigenwerten getrennt ist, stellt die Spalte $\tilde{\boldsymbol{v}}^i$ von $\tilde{\boldsymbol{V}}$ nach 13.1.4 eine ausreichend gute Näherung für den exakten Eigenvektor $\boldsymbol{u}^i$ von $\boldsymbol{A}$ dar. Im Fall mehrfacher Eigenwerte oder Eigenwerthaufen, die von den übrigen Eigenwerten genügend getrennt sind, approximiert der von den entsprechenden Spalten $\tilde{\boldsymbol{v}}^j$ aufgespannte Raum den entsprechenden Eigenraum ausreichend genau.

(v) Bei Abbruch nach vier Zyklen — also $4N \sim 2n^2$ Schritten — kostet die Berechnung der Eigenwerte allein $\sim 6n^3$ opms, sofern ein Schritt mit $\sim 3n$ opms gezählt wird, was nach 13.2.1 berechtigt ist. Wenn auch die Eigenvektoren gesucht sind, muß $\boldsymbol{V}_k$ nach (3) aufdatiert werden. Bei Realisierung nach Ü 3.3.6 kostet dies ebenfalls $\sim 3n$ opms pro Schritt, also insgesamt nochmals $\sim 6n^3$ opms. Zur Berechnung des vollständigen Eigensystems werden daher $\sim 12n^3$ opms benötigt.

(vi) In Abschnitt 13.7 werden wir den $\boldsymbol{QR}$-Algorithmus kennenlernen, der die Eigenwerte mit einem Neuntel, das Eigensystem mit einem Drittel des hier benötigten Aufwandes bei vergleichbarer Qualität zu berechnen erlaubt. Aus diesem Grund ist die praktische Anwendung des Jacobi-Verfahrens auf Spezialfälle beschränkt, etwa die Berechnung der Eigenwerte mit geringer Genauigkeit — als Beispiel $\varepsilon_{\text{rel}} \approx 10^{-1}$ — oder die Berechnung der Eigenwerte mit voller Genauigkeit, wenn die Nichtdiagonalelemente klein sind — z. B. im Sinne von $\max\{|a_{ij}| : i \neq j\} \leqq \sqrt{\nu}\,\|\boldsymbol{A}\|_F$. In beiden Fällen besteht die Chance, mit ein bis zwei Zyklen zum Ziel

zu kommen. Auch für kleines n — etwa $n \leqq 10$ — kann sich das Jacobi-Verfahren gegenüber dem $\boldsymbol{QR}$-Algorithmus als konkurrenzfähig erweisen.

(vii) Wegen seiner Einfachheit und leichten Parallelisierbarkeit wird sich die Bedeutung des Jacobi-Verfahrens für nichtklassische Computerarchitekturen — etwa Parallelcomputer — möglicherweise erhöhen. □

Übungsaufgaben

Ü 13.2.1. Man zeige: Wenn der Drehwinkel φ der Jacobi-Drehung $\boldsymbol{G}_{pq}$ durch $c = \cos\varphi$, $s = \sin\varphi$ festgelegt wird, gilt $t = \tan\varphi$, $\delta = \cot(2\varphi)$, $\tau = \tan(\varphi/2)$.

Ü 13.2.2. Man überprüfe, daß die zur betragsgrößten Lösung $\hat{t} = -1/t$ von (17) gehörenden Parameter $\hat{c} := -s$, $\hat{s} := c$ eine weitere Jacobi-Drehung $\boldsymbol{G}_{pq}(\hat{c}, \hat{s})$ liefern. Dabei gilt für $|\delta| \to \infty$

$$\begin{pmatrix} c & s \\ -s & c \end{pmatrix} \to \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}, \quad \text{aber} \quad \begin{pmatrix} \hat{c} & \hat{s} \\ -\hat{s} & \hat{c} \end{pmatrix} \to \begin{pmatrix} 0 & 1 \\ -1 & 0 \end{pmatrix}.$$

Ü 13.2.3. Man zeige, daß die aus den Eingangsdaten $\{a_{pp}, a_{qq}, a_{pq}\}$ numerisch gemäß (18), (19), (20) berechneten Werte $\tilde{t}$, $\tilde{c}$, $\tilde{s}$ den Ungleichungen

$$|t - \tilde{t}| \leqq 6\nu\,|t|, \qquad |c - \tilde{c}| \leqq 9\nu\,|c|, \qquad |s - \tilde{s}| \leqq 9\nu\,|s| \tag{44}$$

genügen. Wie sollte die Berechnungsvorschrift (19) modifiziert werden, damit für großes $|\delta|$ kein Überlauf eintritt?

Ü 13.2.4. Es sei $\boldsymbol{a} := (a_{ip}, a_{iq})^{\mathsf{T}}$, $\bar{\boldsymbol{a}} := (\bar{a}_{ip}, \bar{a}_{iq})^{\mathsf{T}}$. Man zeige, daß im Fall $|s| \ll 1$ der bei der Berechnung von $\bar{\boldsymbol{a}}$ nach (11) erzeugte Rundungsfehler $\delta\bar{\boldsymbol{a}}$ durch

$$\|\delta\bar{\boldsymbol{a}}\| \leqq 12\nu\,\|\bar{\boldsymbol{a}}\|$$

beschränkt ist, während bei Verwendung von (23) eine Schranke in der Größenordnung

$$\|\delta\bar{\boldsymbol{a}}\| \leqq \nu\,\|\bar{\boldsymbol{a}}\|$$

zu erwarten ist. In ähnlicher Weise wirkt sich der Ersatz von (10) durch (22) aus.

Ü 13.2.5. Es sei $\min\{|\lambda_i - \lambda_j| : i \neq j\} =: 2\gamma > 0$, und k sei so groß, daß

$$\eta_k := \max\{|a_{ij}^{(k)}| : i < j\} \ll \gamma \leqq \min\{|a_{ii}^{(l)} - a_{jj}^{(l)}| : i \neq j\}$$

für alle $l \geqq k$ gilt. Man überlege sich:

(i) Wegen (25), (22), (23) ändert die Jacobi-Drehung $\boldsymbol{G}_{pq}$ im k-ten Schritt außer a_{pq} alle übrigen Elemente um höchstens η_k^2/γ, so daß $\eta_{k+1} \leqq \eta_k + (\eta_k)^2/\gamma \approx \eta_k$ gilt.

(ii) Wenn ab Schritt k ein voller Zyklus (29) von N Jacobi-Schritten durchlaufen wird, kann ein in diesem Zyklus annulliertes Element $\bar{a}_{pq}$ durch höchstens $n - 2$ nachfolgende Jacobi-Drehungen modifiziert werden, so daß

$$\eta_{k+N} \leqq n(\eta_k)^2/\gamma$$

gilt.

13.3. Vektor- und Teilraumiteration

Gegenstand dieses Abschnittes sind iterative Verfahren zur Approximation derjenigen Eigenvektoren der Matrix $\boldsymbol{A} \in \mathbf{S}^{n,n}$, die zu sog. dominanten, d. h. betragsmäßig größten und von den übrigen getrennten Eigenwerten gehören. Bei der

Realisierung dieser Verfahren fallen in natürlicher Weise auch Approximationen für die entsprechenden Eigenwerte an. Die Algorithmen selbst haben keinen allzu großen direkten Anwendungsbereich. Sie sind jedoch Grundlage für leistungsfähige Weiterentwicklungen wie die inverse Iteration und der $\boldsymbol{QR}$-Algorithmus, aber auch für die meisten modernen Verfahren zur Lösung des Eigenwertproblems schwach besetzter Matrizen hoher Dimension.

Im folgenden seien die Eigenwerte $\{\lambda_j\}$ von $\boldsymbol{A}$ nach absteigenden Beträgen gemäß

$$|\lambda_n| \leqq |\lambda_{n-1}| \leqq \cdots \leqq |\lambda_2| \leqq |\lambda_1| \tag{1}$$

geordnet, und $\{\boldsymbol{u}^j\}$ sei ein orthonormales System zugehöriger Eigenvektoren. Die Eigenwerte $\{\lambda_1, \ldots, \lambda_p\}$ heißen dann *dominant*, wenn

$$|\lambda_{p+1}| < |\lambda_p| \tag{2}$$

gilt. Den p dominanten Eigenwerten entspricht der durch die zugehörigen Eigenvektoren aufgespannte Teilraum

$$\mathcal{S}_p = \mathcal{S}_p(\boldsymbol{A}) := \operatorname{span}\{\boldsymbol{u}^1, \ldots, \boldsymbol{u}^p\}, \tag{3}$$

der *p-ter dominanter Teilraum* von $\boldsymbol{A}$ genannt wird. Offensichtlich ist $\{\boldsymbol{u}^1, \ldots, \boldsymbol{u}^p\}$ eine orthonormale Basis von $\mathcal{S}_p$, und es gilt $\dim \mathcal{S}_p = p$. Aus $\boldsymbol{A}\boldsymbol{u}^j = \lambda_j \boldsymbol{u}^j \in \mathcal{S}_p$ $(j = 1, \ldots, p)$ folgt außerdem

$$\boldsymbol{A}\mathcal{S}_p = \{\boldsymbol{y} = \boldsymbol{A}\boldsymbol{x} : \boldsymbol{x} \in \mathcal{S}_p\} \subset \mathcal{S}_p, \tag{4}$$

d. h., $\mathcal{S}_p$ ist invariant unter der durch $\boldsymbol{A}$ vermittelten linearen Abbildung. Ein Teilraum $\mathcal{S}_p$, für den (4) gilt, heißt *invarianter Teilraum* von $\boldsymbol{A}$. Wir werden später sehen, daß umgekehrt jedem invarianten Teilraum von $\boldsymbol{A}$ der Dimension p genau p Eigenwerte und Eigenvektoren von $\boldsymbol{A}$ zugeordnet werden können. Wegen der Trennung der dominanten Eigenwerte von den restlichen wird man daher erwarten können, aus einer genügend guten Approximation von $\mathcal{S}_p$ auch genügend gute Approximationen für die Eigenpaare $\{\lambda_j, \boldsymbol{u}^j\}$ $(j = 1, \ldots, p)$ beschaffen zu können.

Um zu den gesuchten Approximationen für $\mathcal{S}_p$ zu kommen, beachten wir, daß die Eigenwertzerlegung

$$\begin{aligned}\boldsymbol{A} = \boldsymbol{U}\boldsymbol{\Lambda}\boldsymbol{U}^{\mathsf{T}} = [&\lambda_1 \boldsymbol{u}^1 \boldsymbol{u}^{1\mathsf{T}} + \cdots + \lambda_p \boldsymbol{u}^p \boldsymbol{u}^{p\mathsf{T}}] \\ &+ [\lambda_{p+1} \boldsymbol{u}^{p+1} \boldsymbol{u}^{p+1\mathsf{T}} + \cdots + \lambda_n \boldsymbol{u}^n \boldsymbol{u}^{n\mathsf{T}}]\end{aligned} \tag{5}$$

von $\boldsymbol{A}$ die analoge Zerlegung

$$\begin{aligned}\boldsymbol{A}^k = \boldsymbol{U}\boldsymbol{\Lambda}^k\boldsymbol{U}^{\mathsf{T}} = [&\lambda_1^k \boldsymbol{u}^1 \boldsymbol{u}^{1\mathsf{T}} + \cdots + \lambda_p^k \boldsymbol{u}^p \boldsymbol{u}^{p\mathsf{T}}] \\ &+ [\lambda_{p+1}^k \boldsymbol{u}^{p+1} \boldsymbol{u}^{p+1\mathsf{T}} + \cdots + \lambda_n^k \boldsymbol{u}^n \boldsymbol{u}^{n\mathsf{T}}]\end{aligned} \tag{6}$$

für die k-te Potenz $\boldsymbol{A}^k$ von $\boldsymbol{A}$ nach sich zieht. Durch die Potenzbildung wird die Dominanz der ersten p Terme mit wachsendem k immer größer, so daß $\boldsymbol{A}^k$ für genügend großes k praktisch durch die Summanden in der ersten Klammer repräsentiert wird.

Wenn A^k auf einen beliebigen Vektor $v \in \mathbf{R}^n$ angewendet wird, ergibt sich daher

$$A^k v = \sum_{j=1}^{n} \lambda_j^k u^j [u^{j\top} v] \approx \left[\lambda_1^k u^1 [u^{1\top} v] + \cdots + \lambda_p^k u^p [u^{p\top} v]\right], \tag{7}$$

sofern die Koeffizienten $u^{j\top} v$ $(j = 1, \ldots, p)$ nicht sämtlich verschwinden. Für geeignet gewähltes v und genügend großes k ist also $A^k v$ „fast" eine Linearkombination der ersten p Eigenvektoren, d. h., $A^k v$ liegt „fast" in $\mathcal{S}_p$. Wenn (7) für p geeignet gewählte und linear unabhängige Vektoren $\{v\}$ gebildet wird, werden die zugehörigen Vektoren $\{A^k v\}$ einen Teilraum aufspannen, der $\mathcal{S}_p$ für genügend großes k ausreichend gut approximiert. Die Berechnung von $A^k v$ erfolgt selbstverständlich rekursiv gemäß $A^{k+1} v := A(A^k v)$. Da $\|A^k v\|$ im Fall $|\lambda_1| > 1$ unbeschränkt wächst, im Fall $|\lambda_1| < 1$ dagegen gegen 0 geht, muß bei der praktischen Berechnung eine Normierung vorgenommen werden, um unnötigen Über- bzw. Unterlauf zu vermeiden. Weil für die Teilraumapproximation nur die Richtungen wesentlich sind, hat das auf die Approximationsgüte keinen Einfluß.

Wir beginnen mit der genauen Untersuchung des einfachsten Falles $p = 1$.

A. *Vektoriteration*

Hier genügt es, einen Vektor v zu iterieren. Das zugehörige Verfahren lautet wie folgt:

13.3.1. *Basisverfahren der Vektoriteration*
S0 (Initialisierung): Wähle v^1 mit $\|v^1\| = 1$, setze $k := 1$
S1 (Iteration): Berechne $w^{k+1} := A v^k$
S2 (Normierung): Setze $v^{k+1} := w^{k+1}/\|w^{k+1}\|$
S3: Setze $k := k + 1$, **goto** S1
Aufwand pro Schritt: eine Auswertung von $Ax + \sim n(2 \text{ opm} + 1 \text{ ops}) + 1 \text{ opr}$

13.3.2. *Bemerkung.* (i) Im Unterschied zu anderen Eigenwertverfahren wird die Matrix A in 13.3.1 nicht verändert und braucht auch explizit nicht verfügbar zu sein. Es genügt, daß eine Prozedur vorhanden ist, die zu gegebenem x das Bild $y = Ax$ berechnet und im übrigen als „black box" behandelt werden kann. Dies ist besonders bei schwach besetzten Matrizen hoher Dimension, wie sie etwa bei der Diskretisierung von Eigenwertaufgaben bei Differentialgleichungen entstehen, vorteilhaft.

(ii) Statt der 2-Norm kann auch die ∞-Norm zur Normierung verwendet werden, etwa gemäß

S2′: Bestimme $s = s(k+1)$ mit $|(w^{k+1})_s| = \max\{|(w^{k+1})_i| : 1 \leqq i \leqq n\}$,
setze $\omega_{k+1} := (w^{k+1})_s$ und $v^{k+1} := w^{k+1}/\omega_{k+1}$.

Bei dieser Festlegung gilt $\|v^{k+1}\|_\infty = (v^{k+1})_s = 1$. Für die Konvergenzanalyse und die Eigenwertapproximation ist jedoch die 2-Norm günstiger.

(iii) Wegen der Gültigkeit der Darstellung

$$v^{k+1} = A^k v^1/\tau_{k+1}, \qquad \tau_{k+1} := \|w^2\| \cdot \|w^3\| \cdots \|w^{k+1}\| \tag{8}$$

heißt Algorithmus 13.3.1 auch *Potenzmethode* (engl. „power method", russ. „степенной метод"). In der deutschsprachigen Literatur ist auch der Name „von Mises-Iteration" gebräuchlich, siehe B 13.3. □

Das Konvergenzverhalten der Vektoriteration wird durch den folgenden Satz beschrieben:

13.3.3. *Satz.* Es sei λ_1 der dominierende Eigenwert von $\boldsymbol{A}$, und $\boldsymbol{u}^1$ bezeichne den zugehörigen normierten Eigenvektor. Der Startvektor $\boldsymbol{v}^1$ genüge der Bedingung

$$\sigma := \boldsymbol{u}^{1\mathrm{T}}\boldsymbol{v}^1 > 0. \tag{9}$$

Dann gilt für die gemäß 13.3.1 in exakter Arithmetik erzeugte Folge $\{\boldsymbol{v}^k\}$

$$0 \leqq \tan \varphi_{k+1} \leqq \varkappa \tan \varphi_k \leqq \varkappa^k \tan \varphi_1 = \varkappa^k \sqrt{1-\sigma^2}/\sigma \tag{10}$$

mit

$$\varkappa := |\lambda_2/\lambda_1| < 1 \tag{11}$$

und

$$\varphi_k := \sphericalangle\,(\boldsymbol{u}^1, \vartheta_k \boldsymbol{v}^k) \in [0, \pi/2), \qquad \vartheta_k := (\operatorname{sgn} \lambda_1)^{k+1} \in \{+1, -1\}. \tag{12}$$

Beweis. Wir führen die zueinander komplementären und bezüglich $\boldsymbol{A}$ invarianten Teilräume $\mathscr{S}_1 := \operatorname{span}\{\boldsymbol{u}^1\}$ und $\mathscr{Z}_2 := \operatorname{span}\{\boldsymbol{u}^2, \ldots, \boldsymbol{u}^n\} = \mathscr{S}_1^{\perp}$ ein, vgl. 8.1.B. Die Zerlegung von $\vartheta_1 \boldsymbol{v}^1$ in die Komponenten bezüglich $\mathscr{S}_1$ und $\mathscr{Z}_2$ kann dann in der Form

$$\boldsymbol{v}^1 = \vartheta_1 \boldsymbol{v}^1 = \boldsymbol{u}^1 \cos \varphi_1 + \boldsymbol{z}^1 \sin \varphi_1 \quad \text{mit} \quad \boldsymbol{z}^1 \in \mathscr{Z}_2, \qquad \|\boldsymbol{z}^1\| = 1, \tag{13}$$

geschrieben werden. Wegen (9) gilt dabei $\sigma = \cos \sphericalangle\,(\boldsymbol{u}^1, \boldsymbol{v}^1) = \cos \varphi_1 > 0$, also $0 \leqq \varphi_1 < \pi/2$ und $\tan \varphi_1 = \sqrt{1-\sigma^2}/\sigma \geqq 0$. Linksmultiplikation von (13) mit $\operatorname{sgn}(\lambda_1)\boldsymbol{A}$ führt auf

$$\operatorname{sgn}(\lambda_1)\boldsymbol{A}\vartheta_1\boldsymbol{v}^1 = \vartheta_2 \boldsymbol{A}\boldsymbol{v}^1 = \vartheta_2 \boldsymbol{w}^2 = \operatorname{sgn}(\lambda_1)\{\lambda_1 \boldsymbol{u}^1 \cos \varphi_1 + \boldsymbol{A}\boldsymbol{z}^1 \sin \varphi_1\}$$

mit $\boldsymbol{A}\boldsymbol{z}^1 \in \mathscr{Z}_2$, also $\|\boldsymbol{w}^2\| \geqq |\lambda_1| \cos \varphi_1 > 0$ und somit auf die Darstellung

$$\vartheta_2 \boldsymbol{v}^2 = \vartheta_2 \boldsymbol{w}^2/\|\boldsymbol{w}^2\| = \boldsymbol{u}^1 \cos \varphi_2 + \boldsymbol{z}^2 \sin \varphi_2 \quad \text{mit} \quad \boldsymbol{z}^2 \in \mathscr{Z}_2, \qquad \|\boldsymbol{z}^2\| = 1, \tag{14}$$

wobei

$$\cos \varphi_2 := \frac{|\lambda_1| \cos \varphi_1}{\|\boldsymbol{w}^2\|}, \qquad \sin \varphi_2 := \frac{\|\boldsymbol{A}\boldsymbol{z}^1\| \sin \varphi_1}{\|\boldsymbol{w}^2\|}, \qquad \boldsymbol{z}^2 := \frac{\operatorname{sgn}(\lambda_1)\boldsymbol{A}\boldsymbol{z}^1}{\|\boldsymbol{A}\boldsymbol{z}^1\|}$$

im Fall $\boldsymbol{A}\boldsymbol{z}^1 \neq \boldsymbol{o}$; für $\boldsymbol{A}\boldsymbol{z}^1 = \boldsymbol{o}$ setzen wir einfach $\boldsymbol{z}^2 := \boldsymbol{u}^2$. Aus (14) folgt

$$0 \leqq \varphi_2 < \pi/2 \quad \text{und} \quad 0 \leqq \tan \varphi_2 = [\|\boldsymbol{A}\boldsymbol{z}^1\|/|\lambda_1|] \tan \varphi_1.$$

Nach (13) gilt nun

$$\boldsymbol{z}^1 = \sum_{j=2}^{n} c_j \boldsymbol{u}^j \quad \text{mit} \quad \sum_{j=2}^{n} c_j^2 = 1,$$

woraus sich

$$\boldsymbol{A}\boldsymbol{z}^1 = \sum_{j=2}^{n} \lambda_j c_j \boldsymbol{u}^j, \quad \text{also} \quad \|\boldsymbol{A}\boldsymbol{z}^1\|^2 = \sum_{j=2}^{n} \lambda_j^2 c_j^2 \leqq \lambda_2^2 \sum_{j=2}^{n} c_j^2 = \lambda_2^2$$

ergibt. Daher ist $\|\boldsymbol{A}\boldsymbol{z}^1\|/|\lambda_1| \leqq |\lambda_2/\lambda_1|$, d. h., (10) gilt für $k = 1$. Die Gültigkeit für alle $k \geqq 1$ folgt in analoger Weise durch Induktion. □

13.3.4. *Bemerkung.* (i) Die Charakterisierung der „Entfernung" zweier Richtungen $\boldsymbol{u}, \boldsymbol{v}$ durch die Größe $\left|\tan\left(\sphericalangle(\boldsymbol{u}, \boldsymbol{v})\right)\right|$ scheint adäquat zu sein: Orthogonale Richtungen sind unendlich entfernt, während Vielfache eines Vektors die Entfernung 0, also dieselbe Richtung haben. Wir weisen darauf hin, daß der Vorzeichenfaktor ϑ_k im Verfahren 13.3.1 selbst überhaupt nicht vorkommt. Er wurde in 13.3.3 nur eingeführt, um $\vartheta_k \boldsymbol{v}^k$ wie $\boldsymbol{u}^1$ zu orientieren. Dies garantiert $0 \leqq \varphi_k \leqq \pi/2$ und erlaubt die einfache Formulierung von Konvergenzaussagen, siehe (ii) unten. Auf die Güte der Approximation der Eigenrichtung span $\{\boldsymbol{u}^1\}$ durch $\boldsymbol{v}^k$ hat ϑ_k keinen Einfluß, denn offensichtlich ist $\boldsymbol{v}^k$ eine genau so gute Approximation wie $-\boldsymbol{v}^k$. Ebenso ist die Forderung $\sigma > 0$ keine Einschränkung; im Fall $\sigma < 0$ braucht nur $\boldsymbol{u}^1$ durch $-\boldsymbol{u}^1$ ersetzt zu werden. Benötigt wird eigentlich nur $\sigma \neq 0$.

(ii) Wegen

$$\|\vartheta_k \boldsymbol{v}^k - \boldsymbol{u}^1\| = \sqrt{2(1 - \cos\varphi_k)} = 2 \sin(\varphi_k/2) \leqq \tan\varphi_k \tag{15}$$

folgt aus (10) die Konvergenz der geeignet orientierten Vektoren $\{\pm\boldsymbol{v}^k\}$ gegen $\boldsymbol{u}^1$.

(iii) Wenn $\vartheta_k \boldsymbol{v}^k$ analog zu (13) in der Form

$$\vartheta_k \boldsymbol{v}^k = \boldsymbol{u}^1 \cos\varphi_k + \boldsymbol{z}^k \sin\varphi_k \quad \text{mit} \quad \boldsymbol{z}^k \in \mathcal{Z}_2, \qquad \|\boldsymbol{z}^k\| = 1, \tag{16}$$

dargestellt wird, hat $\boldsymbol{z}^k$ für genügend großes k im Fall $|\lambda_2| > |\lambda_3|$ die Richtung von $\boldsymbol{u}^2$, vgl. auch (7). Es gilt dann $\tan\varphi_{k+1} \approx \varkappa \tan\varphi_k$, d. h., die Schranke (10) ist realistisch; schnellere Konvergenz ist nicht zu erwarten. □

Die zur Eigenvektornäherung $\boldsymbol{v}^k$ gehörende beste Eigenwertapproximation ist nach 13.1.8 der Rayleigh-Quotient

$$\varrho_k := \varrho(\boldsymbol{v}^k) = \boldsymbol{v}^{k\mathsf{T}} \boldsymbol{A} \boldsymbol{v}^k = \boldsymbol{v}^{k\mathsf{T}} \boldsymbol{w}^{k+1}. \tag{17}$$

Aus (16) folgt nun

$$\boldsymbol{A}(\vartheta_k \boldsymbol{v}^k) = \boldsymbol{u}^1 \lambda_1 \cos\varphi_k + \boldsymbol{A}\boldsymbol{z}^k \sin\varphi_k \quad \text{mit} \quad \boldsymbol{A}\boldsymbol{z}^k \in \mathcal{Z}_2 \perp \mathcal{S}_1, \tag{18}$$

also

$$\begin{aligned} \varrho_k = \varrho(\vartheta_k \boldsymbol{v}^k) &= (\boldsymbol{u}^1 \cos\varphi_k + \boldsymbol{z}^k \sin\varphi_k)^{\mathsf{T}} (\boldsymbol{u}^1 \lambda_1 \cos\varphi_k + \boldsymbol{A}\boldsymbol{z}^k \sin\varphi_k) \\ &= \lambda_1 \cos^2\varphi_k + \boldsymbol{z}^{k\mathsf{T}} \boldsymbol{A} \boldsymbol{z}^k \sin^2\varphi_k = \lambda_1 - [\lambda_1 - \boldsymbol{z}^{k\mathsf{T}} \boldsymbol{A} \boldsymbol{z}^k] \sin^2\varphi_k. \end{aligned}$$

Wegen

$$|\lambda_1 - \boldsymbol{z}^{k\mathsf{T}} \boldsymbol{A} \boldsymbol{z}^k| \leqq |\lambda_1| + \|\boldsymbol{z}^k\| \cdot \|\boldsymbol{A}\boldsymbol{z}^k\| \leqq |\lambda_1| + |\lambda_2| \leqq 2\,|\lambda_1| = 2\,\|\boldsymbol{A}\|$$

ergibt sich daraus die Abschätzung

$$|\varrho_k - \lambda_1| \leqq 2\,\|\boldsymbol{A}\| \sin^2\varphi_k \leqq 2\,\|\boldsymbol{A}\| \tan^2\varphi_k. \tag{19}$$

Die Rayleigh-Quotienten konvergieren also linear mit dem Konvergenzfaktor $\varkappa^2$ und damit schneller als die Eigenvektornäherungen, wo der Fehler durch $\tan\varphi_k$ beschränkt ist und $\varkappa$ als Konvergenzfaktor auftritt, siehe (15).

Aus der für genügend großes k gültigen Beziehung

$$\boldsymbol{w}^{k+1} = \boldsymbol{A}\boldsymbol{v}^k \approx \lambda_1 \boldsymbol{v}^k \tag{20}$$

lassen sich auch einfacher berechenbare Eigenwertnäherungen ableiten. Durch Normbildung folgt $\|\boldsymbol{w}^{k+1}\| \approx \|\lambda_1 \boldsymbol{v}^k\| = |\lambda_1|$, so daß

$$\gamma_k := \operatorname{sgn}(\lambda_1)\, \|\boldsymbol{w}^{k+1}\| \tag{21}$$

eine Näherung für λ_1 darstellt. Der Fehler ist durch

$$|\gamma_k - \lambda_1| \leqq 2\, \|\boldsymbol{A}\| \sin^2(\varphi_k/2) \leqq (\|\boldsymbol{A}\|/2) \tan^2 \varphi_k \tag{22}$$

beschränkt, siehe Ü 13.3.2, also vergleichbar mit dem von ϱ_k. Da $\|\boldsymbol{w}^{k+1}\|$ ohnehin berechnet werden muß, braucht in (21) nur noch das Vorzeichen $\operatorname{sgn}(\lambda_1)$ bestimmt zu werden. Zu diesem Zweck lesen wir (20) komponentenweise als $(\boldsymbol{w}^{k+1})_j \approx \lambda_1 (\boldsymbol{v}^k)_j$. Die Quotienten

$$\omega_{k,j} := (\boldsymbol{w}^{k+1})_j / (\boldsymbol{v}^k)_j\,, \qquad (\boldsymbol{v}^k)_j \neq 0, \tag{23}$$

stellen also ebenfalls Näherungen für λ_1 dar, und für genügend großes k ist

$$\operatorname{sgn}(\omega_{k,j}) = \operatorname{sgn}(\lambda_1). \tag{24}$$

Zu empfehlen ist die Festlegung von $j = s(k)$ als Index der betragsgrößten Komponente $(\boldsymbol{v}^k)_{s(k)}$ von $\boldsymbol{v}^k$. Zur Approximation von λ_1 selbst sind die Quotienten $\omega_{k,j}$ nicht so gut geeignet, denn die im Fall $|(\boldsymbol{u}^1)_j| > \tan \varphi_k$ gültige Fehlerschranke

$$|\omega_{k,j} - \lambda_1| \leqq 2\, \|\boldsymbol{A}\| \tan \varphi_k / [\,|(\boldsymbol{u}^1)_j| - \tan \varphi_k] \tag{25}$$

enthält $\tan \varphi_k$ statt $\tan^2 \varphi_k$ und liefert daher nur Konvergenz mit dem Faktor $\varkappa$, siehe Ü 13.3.3.

Bei der Computerrealisierung von 13.3.1 wird der Abbruchfehler durch die erzeugten Rundungsfehler überlagert, so daß 13.3.3 nur in der folgenden modifizierten Form gilt:

13.3.5. *Fehleranalyse.* Die Vektoriteration 13.3.1 werde in Computerarithmetik so durchgeführt, daß die berechneten Vektoren $\{\boldsymbol{w}^{k+1}\}$ der Beziehung

$$\boldsymbol{w}^{k+1} = (\boldsymbol{A} + \delta \boldsymbol{A}_k)\, \boldsymbol{v}^k \quad \text{mit} \quad \|\delta \boldsymbol{A}_k\| \leqq \nu F\, \|\boldsymbol{A}\| \tag{26}$$

genügen, und mit $\varkappa$ aus (11) und $\varepsilon := \nu(F + n/2)$ seien die Bedingungen

$$\varkappa + 10\varepsilon < 1 \quad \text{sowie} \quad |\boldsymbol{u}^{1\mathrm{T}} \boldsymbol{v}^1| > \bar{\varepsilon} := \varepsilon/(1 - \varkappa - \varepsilon) \tag{27}$$

erfüllt. Dann gilt für die den berechneten Vektoren $\{\boldsymbol{v}^k\}$ gemäß (12) zugeordneten Winkel $\{\varphi_k\}$ die Abschätzung

$$0 \leqq \tan \varphi_k \leqq \tau_k \quad \text{mit} \quad \tau_k \geqq \tau_{k+1} \quad \text{und} \quad \lim_{k \to \infty} \tau_k = \bar{\varepsilon}. \tag{28}$$

Wenn k_0 ein im Sinne von $\tan \varphi_{k_0} < 1$ genügend großer Index ist, gilt zudem für $k \geqq k_0$

$$\begin{aligned} &\tan \varphi_{k+1} \leqq \hat{\varkappa} \tan \varphi_k + \hat{\varepsilon} \leqq \hat{\varkappa}^{k-k_0+1} \tan \varphi_{k_0} + \hat{\varepsilon}/(1-\hat{\varkappa}) \\ &\text{mit} \quad \hat{\varkappa} := (\varkappa+\varepsilon)/\left(1-\sqrt{2}\,\varepsilon\right) \doteqdot \varkappa, \qquad \hat{\varepsilon} := \varepsilon/\left(1-\sqrt{2}\,\varepsilon\right) \doteqdot \varepsilon. \end{aligned} \tag{29}$$

Für alle $k \geqq 1$ ist außerdem

$$\|\vartheta_k \boldsymbol{v}^k - \boldsymbol{u}^1\| \leqq \tan \varphi_k + \nu n/2, \tag{30}$$

und die gemäß

$$\varrho_k := \boldsymbol{v}^{k\mathsf{T}} \boldsymbol{w}^{k+1} \quad \text{bzw.} \quad \gamma_k := (\operatorname{sgn} \lambda_1) \sqrt{\boldsymbol{w}^{k+1\mathsf{T}} \boldsymbol{w}^{k+1}}$$

berechneten Eigenwertnäherungen genügen den Abschätzungen

$$|\varrho_k - \lambda_1| \leqq \|\boldsymbol{A}\| \left[2 \sin^2 \varphi_k + \nu(F + 2n)\right] \tag{31}$$

bzw.

$$|\gamma_k - \lambda_1| \leqq \|\boldsymbol{A}\| \, \eta_k/\left(1 + \sqrt{1-\eta_k}\right) \quad \text{mit} \quad \eta_k := \sin^2 \varphi_k + 2\nu(F+n). \tag{32}$$

Der Beweis verläuft ähnlich wie der zu 13.3.3 und (15), (19), (22), wobei (26) und die Ergebnisse aus 2.3.B beachtet werden müssen.

13.3.6. *Bemerkung.* (i) Die erste der Bedingungen (27) fordert, daß $\varkappa = |\lambda_2/\lambda_1|$ nicht zu nahe bei 1 liegt. Die zweite soll ausschließen, daß der Startvektor $\boldsymbol{v}^1$ fast orthogonal zu $\boldsymbol{u}^1$ ist, denn dann kann die in 13.3.5 beschriebene Konvergenz nicht gefolgert werden. Praktisch tritt jedoch für jeden Startvektor $\boldsymbol{v}^1$ Konvergenz auf: Auch für exakt orthogonale $\boldsymbol{u}^1, \boldsymbol{v}^1$ erzeugen die Rundungsfehler fast immer nichtverschwindende Komponenten von $\boldsymbol{v}^k$ $(k > 1)$ in Richtung von $\boldsymbol{u}^1$, die sich in den nachfolgenden Schritten weiter verstärken.

(ii) Für voll besetztes $\boldsymbol{A}$ ist (26) bei „normaler" Berechnung von $\boldsymbol{w}^{k+1} = \mathrm{fl}(\boldsymbol{A}\boldsymbol{v}^k)$ mit $F = n^{3/2}$ erfüllt, vgl. 2.3.D und 2.3.15. In den meisten Anwendungen ist $\boldsymbol{A}$ jedoch schwach besetzt, so daß sich wesentlich kleinere Werte für F ergeben.

(iii) Aus (26) folgt mit der Abkürzung $\hat{\boldsymbol{v}}^k := \vartheta_k \boldsymbol{v}^k$

$$(\boldsymbol{A} + \delta\boldsymbol{A}_k)\,\hat{\boldsymbol{v}}^k = \gamma_k \hat{\boldsymbol{v}}^{k+1} = \gamma_k \hat{\boldsymbol{v}}^k + \gamma_k(\hat{\boldsymbol{v}}^{k+1} - \hat{\boldsymbol{v}}^k), \tag{33}$$

mithin

$$\begin{aligned} \|\boldsymbol{A}\hat{\boldsymbol{v}}^k - \gamma_k \hat{\boldsymbol{v}}^k\| = \|\boldsymbol{A}\boldsymbol{v}^k - \gamma_k \boldsymbol{v}^k\| &\leqq \|\delta\boldsymbol{A}_k\| + \|\hat{\boldsymbol{v}}^{k+1} - \hat{\boldsymbol{v}}^k\| \, |\gamma_k| \\ &\leqq \{\nu F + \|\hat{\boldsymbol{v}}^{k+1} - \hat{\boldsymbol{v}}^k\|\} \, \|\boldsymbol{A}\|. \end{aligned}$$

Nach 13.1.7 existiert dann eine Störung $\delta\hat{\boldsymbol{A}}_k \in \mathbf{S}^{n,n}$ mit

$$\|\delta\hat{\boldsymbol{A}}_k\| \leqq \varepsilon_k \|\boldsymbol{A}\|, \qquad \varepsilon_k := \nu F + \|\hat{\boldsymbol{v}}^{k+1} - \hat{\boldsymbol{v}}^k\|, \tag{34}$$

so daß $\{\gamma_k, \boldsymbol{v}^k\}$ Eigenpaar von $\boldsymbol{A} + \delta\hat{\boldsymbol{A}}_k$ ist. Dies liefert eine nützliche a-posteriori-Abschätzung, die z. B. als Abbruchkriterium verwendet werden kann. Mit (29),

(30) läßt sich ε_k für $k \geqq k_0$ weiter gemäß

$$\varepsilon_k \leqq \nu F + \|\hat{\boldsymbol{v}}^{k+1} - \boldsymbol{u}^1\| + \|\hat{\boldsymbol{v}}^k - \boldsymbol{u}^1\| \leqq \nu[F + (2F + n)/(1 - \hat{\varkappa})] + 2\hat{\varkappa}^{k-k_0} \tan \varphi_{k_0} \tag{35}$$

abschätzen. Sofern $(2F + n)/(1 - \hat{\varkappa})$ akzeptabel, also $\hat{\varkappa}$ nicht zu nahe bei 1 ist, liegt daher für genügend großes k *numerische Gutartigkeit* vor. Ähnliche Aussagen gelten für $\{\varrho_k, \boldsymbol{v}^k\}$, denn nach 13.1.8 ist dann $\|\boldsymbol{A}\boldsymbol{v}^k - \varrho_k \boldsymbol{v}^k\| \leqq \|\boldsymbol{A}\boldsymbol{v}^k - \gamma_k \boldsymbol{v}^k\|$.

(iv) Eine ausreichende Genauigkeit läßt sich mit vertretbarem Aufwand nur erreichen, wenn $\hat{\varkappa}$ genügend klein ist. Diese Forderung schränkt die direkte Anwendung von 13.3.1 auf ganz spezielle Aufgaben ein. Man beachte, daß durch eine Spektralverschiebung $\bar{\boldsymbol{A}} = \boldsymbol{A} - \mu \boldsymbol{I}$ höchstens die an den Enden des Spektrums gelegenen Eigenwerte zu dominanten gemacht werden können, vgl. auch Ü 13.3.1.

(v) Die richtig orientierten Näherungen $\{\hat{\boldsymbol{v}}^k\}$ ergeben sich nach der Vorschrift

$$\hat{\boldsymbol{w}}^{k+1} := \boldsymbol{A}\hat{\boldsymbol{v}}^k, \qquad \gamma_k := (\operatorname{sgn} \lambda_1)\, \|\hat{\boldsymbol{w}}^{k+1}\|, \qquad \hat{\boldsymbol{v}}^{k+1} := \hat{\boldsymbol{w}}^{k+1}/\gamma_k \quad (k = 1, 2, \ldots), \tag{36}$$

wobei sgn λ_1 etwa durch (24) mit $j = j(k)$ als Index der betragsgrößten Komponente von $\hat{\boldsymbol{v}}^k$ festgelegt wird. □

Wir beschließen diesen Teilabschnitt mit einigen Kommentaren über Modifikationen und Erweiterungen der Vektoriteration.

13.3.7. *Bemerkung.* (i) Wenn λ_1 ein doppelter dominanter Eigenwert ist, d. h., wenn $\lambda_1 = \lambda_2$ und $|\lambda_1| > |\lambda_3|$ gilt, erzeugt 13.3.1 eine Folge $\{\vartheta_k \boldsymbol{v}^k\}$, die mit dem Konvergenzfaktor $\varkappa = |\lambda_3/\lambda_1|$ gegen einen normierten Eigenvektor $\bar{\boldsymbol{u}}^1 \in \mathcal{S}_2 = \operatorname{span}\{\boldsymbol{u}^1, \boldsymbol{u}^2\}$ konvergiert. Wird 13.3.1 mit einem zweiten Startwert $\tilde{\boldsymbol{v}}^1 \neq \boldsymbol{v}^1$ gestartet, so konvergieren die zugehörigen Iterierten $\{\vartheta_k \tilde{\boldsymbol{v}}^k\}$ fast immer gegen einen weiteren normierten Eigenvektor $\tilde{\boldsymbol{u}}^2 \in \mathcal{S}_2$, der von $\bar{\boldsymbol{u}}^1$ linear unabhängig ist. Ein nachfolgender Orthogonalisierungsschritt

$$\boldsymbol{q}^2 := \tilde{\boldsymbol{u}}^2 - [\bar{\boldsymbol{u}}^{1\mathsf{T}} \tilde{\boldsymbol{u}}^2]\, \bar{\boldsymbol{u}}^1, \qquad \bar{\boldsymbol{u}}^2 := \boldsymbol{q}^2/\|\boldsymbol{q}^2\| \tag{37}$$

liefert dann einen Eigenvektor $\bar{\boldsymbol{u}}^2$ derart, daß $\{\bar{\boldsymbol{u}}^1, \bar{\boldsymbol{u}}^2\}$ eine orthonormale Basis von $\mathcal{S}_2$ bildet. Um Genauigkeitsverluste bei der numerischen Realisierung von (37) zu vermeiden, empfiehlt es sich, solche Orthogonalisierungsschritte schon in gewissen Abständen mit den Iterierten $\tilde{\boldsymbol{v}}^k$ durchzuführen, damit $\tilde{\boldsymbol{u}}^2$ hinreichend linear unabhängig von $\bar{\boldsymbol{u}}^1$ ist. In analoger Weise kann im Fall eines p-fachen dominanten Eigenwertes λ_1 vorgegangen werden.

(ii) Es gelte $|\lambda_1| > |\lambda_2| > |\lambda_3|$, und das Eigenpaar $\{\lambda_1, \boldsymbol{u}^1\}$ sei bereits bestimmt worden. Wir betrachten jetzt die Matrix

$$\bar{\boldsymbol{A}} := \boldsymbol{A} - \lambda_1 \boldsymbol{u}^1 \boldsymbol{u}^{1\mathsf{T}} = 0 \boldsymbol{u}^1 \boldsymbol{u}^{1\mathsf{T}} + \sum_{j=2}^{n} \lambda_j \boldsymbol{u}^j \boldsymbol{u}^{j\mathsf{T}}, \tag{38}$$

die die Eigenwerte $\{0, \lambda_2, \ldots, \lambda_n\}$ und die zugehörigen Eigenvektoren $\{\boldsymbol{u}^1, \boldsymbol{u}^2, \ldots, \boldsymbol{u}^n\}$ hat. Durch die Transformation von λ_1 in 0 ist also λ_2 zum dominanten Eigenwert

von $\bar{A}$ geworden, und das Paar $\{\lambda_2, u^2\}$ kann durch Vektoriteration mit $\bar{A}$ bestimmt werden. Dieses *Deflation* genannte Vorgehen ist gemäß $\bar{\bar{A}} := \bar{A} - \lambda_2 u^2 u^{2\mathsf{T}}$ usw. fortsetzbar, sofern die Eigenwerte betragsmäßig getrennt sind. Bei der numerischen Realisierung dieses Prozesses müssen die verwendeten Näherungen allerdings eine hohe Genauigkeit aufweisen.

(iii) Der durch die explizite Deflation (38) erzielte Effekt kann auch implizit erreicht werden: Wenn 13.3.1 in exakter Arithmetik mit einem Startvektor v^1, für den $\sigma = u^{1\mathsf{T}} v^1 = 0$ gilt, durchgeführt wird, ist $v^k \in \mathcal{X}_2 = \operatorname{span}\{u^2, \ldots, u^n\}$, vgl. (7). Überdies sind die v^k mit denjenigen Iterierten identisch, die sich aus v^1 mit $\bar{A}$ statt A ergeben, d. h., v^k approximiert u^2. Bei Computerrechnung muß die Bedingung $v^k \in \mathcal{X}_2$, d. h. $v^k \perp u^1$, durch in gewissen Abständen vorgenommene Orthogonalisierungsschritte

$$q^k := v^k - [u^{1\mathsf{T}} v^k]\, u^1, \qquad v^k := q^k / \|q^k\|, \tag{39}$$

erzwungen werden. Andernfalls würden die durch Rundungsfehler erzeugten Komponenten von v^k in Richtung u^1 schnell anwachsen und die Konvergenz gegen u^2 verhindern, vgl. 13.3.6(i). Dieser Prozeß läßt sich fortsetzen, um mittels Orthogonalisierung bezüglich $\operatorname{span}\{u^1, u^2\}$ eine Vektorfolge in $\mathcal{X}_3 = \operatorname{span}\{u^3, \ldots, u^n\}$ zu konstruieren usw. □

B. *Teilraumiteration*

Bei der Vektoriteration wurde eine Folge $\{v^k\}$ normierter Vektoren erzeugt derart, daß $\operatorname{span}\{v^k\}$ den dominanten Teilraum $\mathcal{S}_1 = \operatorname{span}\{u^1\}$ approximiert. Falls die zu p dominanten Eigenwerten $\{\lambda_1, \ldots, \lambda_p\}$ gehörenden Eigenvektoren $\{u^1, \ldots, u^p\}$ gesucht sind, könnten unter der Zusatzvoraussetzung $|\lambda_{p+1}| < |\lambda_p| < \cdots < |\lambda_2| < |\lambda_1|$ die eben beschriebenen Deflationstechniken angewendet werden. Es ist i. allg. jedoch wesentlich günstiger, direkt Approximationen $\mathcal{Y}_k$ für den p-ten dominanten Teilraum $\mathcal{S}_p$ zu berechnen. Dazu bietet es sich an, einen Startteilraum $\mathcal{Y}_1$ der Dimension p mit der Matrix A zu iterieren.

13.3.8. *Basisverfahren der Teilraumiteration.*
S0 (Initialisierung): Wähle Teilraum $\mathcal{Y}_1 \subset \mathbf{R}^n$ mit $\dim(\mathcal{Y}_1) = p$, setze $k := 1$.
S1 (Iteration): Definiere $\mathcal{Y}_{k+1} := A\mathcal{Y}_k = \{Ay : y \in \mathcal{Y}_k\}$
S2: Setze $k := k + 1$, **goto** S1.

Für die Konvergenzanalyse benötigen wir ein Maß für die Entfernung zweier Teilräume. Es sei dazu $\mathcal{S} \subset \mathbf{R}^n$ ein Teilraum und $y \in \mathbf{R}^n$, $y \neq o$. Dann definieren wir den *Winkel* $\sphericalangle(y, \mathcal{S})$ *zwischen* y *und* $\mathcal{S}$ durch

$$\sphericalangle(y, \mathcal{S}) := \min\{\sphericalangle(y, s) : s \in \mathcal{S}, s \neq o\}. \tag{40}$$

Das Minimum in (40) existiert stets, und es gilt $0 \leqq \sphericalangle(y, \mathcal{S}) \leqq \pi/2$ sowie

$$\sin\big(\sphericalangle(y, \mathcal{S})\big) = \sin\big(\sphericalangle(y, P_{\mathcal{S}} y)\big) = \min\{\|y - s\| : s \in \mathcal{S}\}/\|y\|, \tag{41}$$

wobei $P_{\mathcal{S}}$ den orthogonalen Projektor auf $\mathcal{S}$ bezeichnet, vgl. 8.1.2 und insbesondere Abb. 8.1.1. Damit (41) auch im Fall $P_{\mathcal{S}}\boldsymbol{y} = 0$ sinnvoll bleibt, setzen wir $\sphericalangle(\boldsymbol{y}, \boldsymbol{o}) := \pi/2$ für jedes $\boldsymbol{y}$.

Wenn $\mathcal{Y} \subset \mathbf{R}^n$ ein Teilraum ist, wird der *maximale Winkel* $\sphericalangle(\mathcal{Y}, \mathcal{S})$ — kurz: *Winkel* — zwischen $\mathcal{Y}$ und $\mathcal{S}$ durch

$$\sphericalangle(\mathcal{Y}, \mathcal{S}) := \max\{\sphericalangle(\boldsymbol{y}, \mathcal{S}) : \boldsymbol{y} \in \mathcal{Y}, \boldsymbol{y} \neq \boldsymbol{o}\} \tag{42}$$

definiert, der *minimale Winkel* $\sphericalangle_{\min}(\mathcal{Y}, \mathcal{S})$ dagegen durch

$$\sphericalangle_{\min}(\mathcal{Y}, \mathcal{S}) := \min\{\sphericalangle(\boldsymbol{y}, \mathcal{S}) : \boldsymbol{y} \in \mathcal{Y}, \boldsymbol{y} \neq \boldsymbol{o}\}. \tag{43}$$

Es ist sofort einzusehen, daß für beliebige Teilräume $\mathcal{Y}, \mathcal{S} \subset \mathbf{R}^n$

$$\sphericalangle_{\min}(\mathcal{Y}, \mathcal{S}) = \sphericalangle_{\min}(\mathcal{S}, \mathcal{Y}) \tag{44}$$

gilt, aber i. allg. ist

$$\sphericalangle(\mathcal{Y}, \mathcal{S}) \neq \sphericalangle(\mathcal{S}, \mathcal{Y});$$

man betrachte eine Gerade und eine Ebene im Raum $\mathbf{R}^3$. Unter der zusätzlichen Voraussetzung $\dim(\mathcal{Y}) = \dim(\mathcal{S})$ gilt jedoch

$$\sphericalangle(\mathcal{Y}, \mathcal{S}) = \sphericalangle(\mathcal{S}, \mathcal{Y}) = \sphericalangle(\mathcal{Y}^\perp, \mathcal{S}^\perp) \tag{45}$$

sowie

$$\sin\left(\sphericalangle(\mathcal{Y}, \mathcal{S})\right) = \cos\left(\sphericalangle_{\min}(\mathcal{Y}, \mathcal{S}^\perp)\right). \tag{46}$$

Zum Beweis von (45), (46) verweisen wir auf die Spezialliteratur, siehe B 13.4 für Hinweise.

Die Größe $\tan\left(\sphericalangle(\mathcal{Y}, \mathcal{S})\right)$ kann im Fall $\dim(\mathcal{Y}) = \dim(\mathcal{S})$ zur Charakterisierung der Entfernung zwischen $\mathcal{Y}$ und $\mathcal{S}$ verwendet werden. Dabei gilt $\tan\left(\sphericalangle(\mathcal{Y}, \mathcal{S})\right) = 0$ genau dann, wenn $\mathcal{Y} = \mathcal{S}$ ist, und die Entfernung ist unendlich, wenn einer der Teilräume eine Richtung enthält, die orthogonal zum anderen Teilraum ist.

Man beachte, daß der Winkel zwischen zwei Teilräumen stets im Intervall $[0, \pi/2]$ liegt, während der Winkel zwischen zwei Vektoren Werte im gesamten Intervall $[0, \pi]$ annehmen kann, denn die Vektoren können entgegengesetzt orientiert sein. Letzteres spielt bei der Definition (40) keine Rolle, denn mit $\boldsymbol{s}$ liegt auch $-\boldsymbol{s}$ in $\mathcal{S}$, so daß mit $\sphericalangle(\boldsymbol{y}, \boldsymbol{s})$ auch der Winkel $\sphericalangle(\boldsymbol{y}, -\boldsymbol{s}) = \pi - \sphericalangle(\boldsymbol{y}, \boldsymbol{s})$ in der Menge liegt, über der das Minimum gebildet wird. Insbesondere gilt für zwei Vektoren $\boldsymbol{u}, \boldsymbol{v}$

$$\sphericalangle(\operatorname{span}\{\boldsymbol{u}\}, \operatorname{span}\{\boldsymbol{v}\}) = \sphericalangle(\boldsymbol{u}, \boldsymbol{v})$$

nur im Fall $\boldsymbol{u}^T\boldsymbol{v} \geqq 0$, also $0 \leqq \sphericalangle(\boldsymbol{u}, \boldsymbol{v}) \leqq \pi/2$; im Fall $\boldsymbol{u}^T\boldsymbol{v} \leqq 0$ ist

$$\sphericalangle(\operatorname{span}\{\boldsymbol{u}\}, \operatorname{span}\{\boldsymbol{v}\}) = \pi - \sphericalangle(\boldsymbol{u}, \boldsymbol{v}) = \sphericalangle(-\boldsymbol{u}, \boldsymbol{v}) = \sphericalangle(\boldsymbol{u}, -\boldsymbol{v}).$$

Mit den eingeführten Größen läßt sich der nachfolgende Konvergenzsatz formulieren:

13.3.9. *Satz.* Es seien $\{\lambda_1, \ldots, \lambda_p\}$ dominante Eigenwerte von $\boldsymbol{A}$, und $\mathcal{S}_p$ sei der zugehörige p-te dominante Teilraum. Der Startteilraum $\mathcal{Y}_1 \subset \mathbf{R}^n$ genüge den

Bedingungen

$$\dim(\mathcal{Y}_1) = p \quad \text{und} \quad \sigma := \cos\left(\sphericalangle(\mathcal{Y}_1, \mathcal{S}_p)\right) > 0. \tag{47}$$

Dann gilt für die durch 13.3.8 definierte Folge $\{\mathcal{Y}_k\}$ von Teilräumen $\dim(\mathcal{Y}_k) = p$ und

$$\tan\varphi_{k+1} \leqq \varkappa \tan\varphi_k \leqq \varkappa^k \tan\varphi_1 = \varkappa^k \sqrt{1-\sigma^2}/\sigma \tag{48}$$

mit

$$\varkappa := |\lambda_{p+1}/\lambda_p| < 1 \quad \text{und} \quad \varphi_k := \sphericalangle(\mathcal{Y}_k, \mathcal{S}_p). \tag{49}$$

Der Beweis verläuft analog zu dem von 13.3.3, wobei $\mathcal{Y}_k$ in Komponenten bezüglich der komplementären Teilräume $\mathcal{S}_p = \operatorname{span}\{\boldsymbol{u}^1, \ldots, \boldsymbol{u}^p\}$ und $\mathcal{X}_{p+1} := \operatorname{span}\{\boldsymbol{u}^{p+1}, \ldots, \boldsymbol{u}^n\} = \mathcal{S}_p^{\perp}$ zu zerlegen ist. □

Zur Implementierung von 13.3.8 liegt es nahe, den Teilraum $\mathcal{Y}_k$ gemäß

$$\mathcal{Y}_k = \mathcal{R}(\boldsymbol{V}_k), \qquad \boldsymbol{V}_k := (\boldsymbol{v}^{k,1}, \ldots, \boldsymbol{v}^{k,p}) \in \mathbf{R}^{n,p} \tag{50}$$

durch die Spalten einer geeigneten Matrix $\boldsymbol{V}_k$ aufzuspannen. Dann spannen die Spalten von

$$\boldsymbol{W}_{k+1} := (\boldsymbol{w}^{k+1,1}, \ldots, \boldsymbol{w}^{k+1,p}) := \boldsymbol{A}\boldsymbol{V}_k = (\boldsymbol{A}\boldsymbol{v}^{k,1}, \ldots, \boldsymbol{A}\boldsymbol{v}^{k,p}) \tag{51}$$

den Teilraum $\mathcal{Y}_{k+1} = \boldsymbol{A}\mathcal{Y}_k$ auf. Falls diese Spalten zur Vermeidung von Über- bzw. Unterlauf gemäß

$$\boldsymbol{V}_{k+1} := \boldsymbol{W}_{k+1} \operatorname{diag}(1/\|\boldsymbol{w}^{k+1,1}\|, \ldots, 1/\|\boldsymbol{w}^{k+1,p}\|) \tag{52}$$

normiert werden, ändert sich der Wertebereich nicht, d. h., es gilt $\mathcal{Y}_{k+1} = \mathcal{R}(\boldsymbol{V}_{k+1}) = \mathcal{R}(\boldsymbol{W}_{k+1})$. Wenn die Startbasis normiert ist, sind die Vektoren $\{\boldsymbol{v}^{k,i}\}$ dann identisch mit denjenigen, die sich nach der Vektoriteration 13.3.1 aus $\boldsymbol{v}^{1,i}$ ergeben. Im Fall $|\lambda_1| > |\lambda_2|$ approximieren daher alle diese Vektoren für genügend großes k die Richtung $\boldsymbol{u}^1$ oder $-\boldsymbol{u}^1$ beliebig genau, so daß die Basis $\{\boldsymbol{v}^{k,1}, \ldots, \boldsymbol{v}^{k,p}\}$ mit wachsendem k immer schiefwinkliger wird und vom numerischen Standpunkt immer schlechter zur Repräsentation von $\mathcal{Y}_k$ geeignet ist. Aus diesem Grunde ist es zweckmäßig, mit einer orthonormalen Basis von $\mathcal{Y}_{k+1}$ zu arbeiten, also $\boldsymbol{V}_{k+1}$ als orthonormale Matrix festzulegen. Dabei empfiehlt es sich aus später klar werdenden Gründen, mit mehr als p Vektoren zu arbeiten, also Teilräume der Dimension $m \geqq p$ zu betrachten.

13.3.10. *Teilraumiteration mittels orthonormaler Basen.*

S0 (Initialisierung): Wähle m mit $p \leqq m \leqq n$ und spaltenorthonormale Startmatrix $\boldsymbol{V}_1 \in \mathbf{R}^{n,m}$, setze $k := 1$.

S1 (Iteration): Berechne $\boldsymbol{W}_{k+1} := \boldsymbol{A}\boldsymbol{V}_k$

S2 (Orthonormalisierung): Bestimme spaltenorthonormale Matrix $\boldsymbol{Q}_{k+1} \in \mathbf{R}^{n,m}$ und reguläre obere Dreiecksmatrix $\boldsymbol{R}_{k+1} \in \mathbf{R}^{m,m}$ mit

$$\boldsymbol{W}_{k+1} = \boldsymbol{Q}_{k+1}\boldsymbol{R}_{k+1}$$

S3: Setze $\boldsymbol{V}_{k+1} := \boldsymbol{Q}_{k+1}$
S4: Setze $k := k + 1$, **goto** S1
Aufwand pro Schritt: m Auswertungen von $\boldsymbol{Ax}$ (für S1), $\sim K_1 nm^2$ opms (für S2)

13.3.11. *Bemerkung.* (i) Wenn die Voraussetzungen von 13.3.9 mit m statt p erfüllt sind, gilt 13.3.9 auch für die in exakter Arithmetik ausgeführte Teilraumiteration 13.3.10. Insbesondere ist dann $\dim(\mathcal{Y}_{k+1}) = \operatorname{rang}(\boldsymbol{W}_{k+1}) = m$, so daß S2 in der Tat mit spaltenorthonormalem $\boldsymbol{Q}_{k+1}$ und regulärem $\boldsymbol{R}_{k+1}$ durchgeführt werden kann. Für eine ausreichend gute Approximation von $\mathcal{Y}_{k+1}$ durch die Spalten von $\boldsymbol{Q}_{k+1}$ brauchen dabei keine allzu hohen Orthogonalitätsforderungen gestellt zu werden, so daß sich das modifizierte Gram-Schmidt-Verfahren 10.1.5 zur Realisierung anbietet. Bei dieser Version ist in den Aufwandsangaben $K_1 = 1$ zu setzen. Bei höheren Orthogonalitätsforderungen muß i. allg. mit Re-Orthogonalisierung gearbeitet werden, wobei sich der Aufwand verdoppelt, vgl. 10.1.8(iii).

(ii) Wenn für festes p mit $1 \leqq p \leqq m$ die Partitionierungen $\boldsymbol{V}_k = (\boldsymbol{V}_k' \mid \boldsymbol{V}_k'')$ mit $\boldsymbol{V}_k' \in \mathbf{R}^{n,p}$, $\boldsymbol{V}_k'' \in \mathbf{R}^{n,m-p}$ und analog $\boldsymbol{Q}_k = (\boldsymbol{Q}_k' \mid \boldsymbol{Q}_k'')$, $\boldsymbol{W}_k = (\boldsymbol{W}_k' \mid \boldsymbol{W}_k'')$ sowie $\boldsymbol{R}_k = \left(\begin{array}{c|c} \boldsymbol{R}_k' & \boldsymbol{R}_k'' \\ \hline \boldsymbol{O} & \boldsymbol{R}_k''' \end{array}\right)$ vorgenommen werden, gilt $\boldsymbol{W}_{k+1}' = \boldsymbol{A}\boldsymbol{V}_k'$ und $\boldsymbol{W}_{k+1}' = \boldsymbol{Q}_{k+1}'\boldsymbol{R}_{k+1}'$, d. h., die Spalten von $\boldsymbol{V}_k'$ bilden eine orthonormale Basis desjenigen Teilraums $\mathcal{Y}_k'$ der Dimension p, der durch Teilraumiteration aus $\mathcal{Y}_1'$ entsteht. Algorithmus 13.3.10 liefert also gleichzeitig auch Approximationen für die invarianten Teilräume $\mathcal{S}_p$ mit $p \leqq m$, sofern die Eigenwerte $\{\lambda_1, \ldots, \lambda_p\}$ dominant sind.

(iii) Für die Computerrealisierung von 13.3.10 lassen sich zu 13.3.5 analoge Aussagen beweisen. Aus Aufwands- und Genauigkeitsgründen ist auch hier wesentlich, daß $\varkappa = |\lambda_{p+1}/\lambda_p|$ deutlich kleiner als 1 ist, vgl. 13.3.6. □

C. *Der Rayleigh-Ritz-Algorithmus*

Im folgenden wenden wir uns dem Problem zu, wie den durch $\boldsymbol{Q}_k$ repräsentierten Teilraumapproximationen $\mathcal{Y}_k$ möglichst gute Näherungen $\{\mu_j, \boldsymbol{v}^j\}$ für die gesuchten Eigenpaare von $\boldsymbol{A}$ zugeordnet werden können. Wenn in Analogie zu 13.1.8 die Quadratsummen der Residuumsnormen als Kriterium für die Güte der Approximationen verwendet werden, ergibt sich das folgende Resultat.

13.3.12. *Satz.* Es sei $\mathcal{Y} = \mathcal{R}(\boldsymbol{Q})$ ein m-dimensionaler Teilraum, der durch die Spalten der spaltenorthonormalen Matrix $\boldsymbol{Q} = (\boldsymbol{q}^1, \ldots, \boldsymbol{q}^m) \in \mathbf{R}^{n,m}$ $(1 \leqq m \leqq n)$ aufgespannt wird. Aus $\boldsymbol{Q}$ und $\boldsymbol{A} \in \mathbf{S}^{n,n}$ werde die Matrix

$$\boldsymbol{P} := (\varrho_{ij}) := \boldsymbol{Q}^{\mathsf{T}}\boldsymbol{A}\boldsymbol{Q} \in \mathbf{S}^{m,m} \tag{53}$$

gebildet, und

$$\boldsymbol{P} = \boldsymbol{X}\boldsymbol{M}\boldsymbol{X}^{\mathsf{T}} \tag{54}$$

sei die zugehörige Eigenwertzerlegung mit $\boldsymbol{M} = \operatorname{diag}(\mu_1, \ldots, \mu_m)$ und ortho-

gonalem $\boldsymbol{X} \in \mathbf{R}^{m,m}$. Dann löst $\{\boldsymbol{M}, \boldsymbol{V}\}$ mit

$$\boldsymbol{V} := (\boldsymbol{v}^1, \ldots, \boldsymbol{v}^m) := \boldsymbol{QX} \tag{55}$$

die Aufgabe

$$\|\boldsymbol{AV} - \boldsymbol{VM}\|_F^2 = \sum_{j=1}^{m} \|\boldsymbol{Av}^j - \mu_j \boldsymbol{v}^j\|^2 \to \text{Minimum!} \tag{56}$$

bei

$$\boldsymbol{M} = \operatorname{diag}(\mu_j) \in \mathbf{R}^{m,m}, \qquad \boldsymbol{V} \in \mathbf{R}^{n,m} \text{ spaltenorthonormal mit } \mathcal{R}(\boldsymbol{V}) = \mathcal{R}(\boldsymbol{Q}). \tag{57}$$

Beweis. Wegen $\mathcal{R}(\boldsymbol{V}) = \mathcal{R}(\boldsymbol{Q})$ ist $\boldsymbol{V} = \boldsymbol{QX}$ mit regulärem $\boldsymbol{X} \in \mathbf{R}^{m,m}$, und wegen der geforderten Spaltenorthonormalität muß $\boldsymbol{I} = \boldsymbol{V}^\mathsf{T}\boldsymbol{V} = \boldsymbol{X}^\mathsf{T}\boldsymbol{Q}^\mathsf{T}\boldsymbol{QX} = \boldsymbol{X}^\mathsf{T}\boldsymbol{X}$ gelten, also $\boldsymbol{X}$ orthogonal sein. Für jedes solche $\boldsymbol{V}$ gilt dann für die Residualmatrix $\boldsymbol{R} = \boldsymbol{AV} - \boldsymbol{VM}$

$$\|\boldsymbol{R}\|_F = \|\boldsymbol{RX}^\mathsf{T}\|_F = \|\boldsymbol{AQ} - \boldsymbol{Q}(\boldsymbol{XMX}^\mathsf{T})\|_F,$$

so daß sich die Aufgabe mit der Bezeichnung $\boldsymbol{B} := (\boldsymbol{b}^1, \ldots, \boldsymbol{b}^m) := \boldsymbol{AQ}$ auf

$$\|\boldsymbol{R}\|_F^2 = \|\boldsymbol{B} - \boldsymbol{QP}\|_F^2 = \sum_{j=1}^{m} \|\boldsymbol{b}^j - \boldsymbol{Q}\varrho^j\|^2 \to \text{Minimum!} \tag{58}$$

bei

$$\boldsymbol{P} := (\varrho^1, \ldots, \varrho^m) := \boldsymbol{XMX}^\mathsf{T}, \qquad \boldsymbol{M} \text{ diagonal}, \quad \boldsymbol{X} \text{ orthogonal},$$

reduziert. Wenn die Nebenbedingungen zunächst ignoriert werden, hat (58) die Lösungen $\varrho^j := (\boldsymbol{Q}^\mathsf{T}\boldsymbol{Q})^{-1}\boldsymbol{Q}^\mathsf{T}\boldsymbol{b}^j = \boldsymbol{Q}^\mathsf{T}\boldsymbol{b}^j$, also $\boldsymbol{P} = \boldsymbol{Q}^\mathsf{T}\boldsymbol{B} = \boldsymbol{Q}^\mathsf{T}\boldsymbol{AQ}$. Da die Matrix $\boldsymbol{P}$ symmetrisch ist, kann sie mittels ihrer Eigenwertzerlegung (54) in der geforderten Form dargestellt werden. □

Die in 13.3.12 beschriebene Zuordnung $\{\boldsymbol{A}, \boldsymbol{Q}\} \to \{\boldsymbol{M}, \boldsymbol{V}\}$ wird *Rayleigh-Ritz-Algorithmus* — kurz: *RR-Algorithmus* — genannt. Die Größen $\{\mu_j\}$ und $\{\boldsymbol{v}^j\}$ heißen *Ritzsche Eigenwerte* bzw. *Eigenvektoren* bezüglich des Teilraumes $\mathcal{Y} = \mathcal{R}(\boldsymbol{Q})$. Die Matrix $\boldsymbol{P} = (\varrho_{ij})$ mit den Elementen $\varrho_{ij} = (\boldsymbol{q}^i)^\mathsf{T}\boldsymbol{Aq}^j$ stellt eine Verallgemeinerung des Rayleigh-Quotienten dar und wird *Projektion von* $\boldsymbol{A}$ auf $\mathcal{Y} = \mathcal{R}(\boldsymbol{Q})$ genannt, vgl. Ü 13.3.5.

13.3.13. *Rayleigh-Ritz-Algorithmus.* Es sei $\boldsymbol{A} \in \mathbf{S}^{n,n}$, und $\boldsymbol{Q} \in \mathbf{R}^{n,m}$ sei spaltenorthonormal.

S1: Bilde $\boldsymbol{B} := \boldsymbol{AQ} \in \mathbf{R}^{n,m}$, berechne $\boldsymbol{P} := \boldsymbol{Q}^\mathsf{T}\boldsymbol{B} \in \mathbf{S}^{m,m}$

S2: Berechne Eigenwertzerlegung $\boldsymbol{P} = \boldsymbol{XMX}^\mathsf{T}$ mit diagonalem $\boldsymbol{M}$ und orthogonalem $\boldsymbol{X}$

S3: Berechne $\boldsymbol{V} := \boldsymbol{QX}$

Aufwand: m Auswertungen von $\boldsymbol{Ax}$ + $\sim nm^2/2$ opms (für S1),
$\sim K_2 m^3$ opms (für S2, K_2 abhängig vom verwendeten Verfahren),
$\sim nm^2$ opms (für S3)

13.3.14. *Bemerkung.* (i) Falls $\mathcal{Y}$ ein invarianter Teilraum von $\boldsymbol{A}$ ist, also $\mathcal{R}(\boldsymbol{AQ}) \subset \mathcal{R}(\boldsymbol{Q})$ gilt, muß $\boldsymbol{AQ} = \boldsymbol{QZ}$ mit einer Matrix $\boldsymbol{Z} \in \mathbf{R}^{m,m}$ sein. Dann folgt $\boldsymbol{P} = \boldsymbol{Q}^\mathsf{T}\boldsymbol{AQ}$

$= \boldsymbol{Q}^T\boldsymbol{Q}\boldsymbol{Z} = \boldsymbol{Z}$, mithin $\boldsymbol{A}\boldsymbol{Q} = \boldsymbol{Q}\boldsymbol{P}$ und durch Rechtsmultiplikation mit $\boldsymbol{X}$ schließlich

$$\boldsymbol{A}\boldsymbol{V} = \boldsymbol{V}\boldsymbol{M}, \quad \text{d. h.} \quad \boldsymbol{A}\boldsymbol{v}^j = \mu_j \boldsymbol{v}^j \qquad (j = 1, \ldots, m). \tag{59}$$

Im Fall eines invarianten Teilraumes verschwindet also die Residualmatrix $\boldsymbol{R} = \boldsymbol{A}\boldsymbol{V} - \boldsymbol{V}\boldsymbol{M}$, und die Ritzschen Eigenpaare $\{\mu_j, \boldsymbol{v}^j\}$ von $\boldsymbol{A}$ bezüglich $\mathcal{Y}$ sind identisch mit gewissen m Eigenpaaren $\{\lambda_{l(j)}, \boldsymbol{u}^{l(j)}\}$ von $\boldsymbol{A}$, wobei $\mathcal{Y} = \operatorname{span}\{u^{l(1)}, \ldots, u^{l(m)}\}$ ist.

(ii) Wenn die Spalten von $\boldsymbol{Q}$ bereits gute Näherungen für die gesuchten Eigenvektoren darstellen, weicht $\boldsymbol{P}$ nur wenig von einer Diagonalmatrix ab. Zur Berechnung der Eigenwertzerlegung in S2 bietet sich dann das Jacobi-Verfahren an. □

Falls $\boldsymbol{R} \neq \boldsymbol{O}$, also $\mathcal{Y}$ nicht invariant ist, lassen sich unter Verwendung von $\boldsymbol{R}$ die folgenden Aussagen über die Güte der Ritzschen Näherungen $\{\mu_j, \boldsymbol{v}^j\}$ machen:

13.3.15. *Satz.* Es sei $\boldsymbol{Q} \in \mathbf{R}^{n,m}$ spaltenorthonormal, und $\boldsymbol{M} = \operatorname{diag}(\mu_j)$ sowie $\boldsymbol{V} = (\boldsymbol{v}^1, \ldots, \boldsymbol{v}^m)$ seien die in exakter Arithmetik berechneten Ritzschen Eigenwerte bzw. Eigenvektoren mit der zugehörigen Residualmatrix $\boldsymbol{R} := (\boldsymbol{r}^1, \ldots, \boldsymbol{r}^m) := \boldsymbol{A}\boldsymbol{V} - \boldsymbol{V}\boldsymbol{M}$. Dann gilt:

(i) Es gibt Störungen $\delta\boldsymbol{A}_j \in \mathbf{S}^{n,n}$ mit

$$(\boldsymbol{A} + \delta\boldsymbol{A}_j)\,\boldsymbol{v}^j = \mu_j \boldsymbol{v}^j \quad \text{und} \quad \|\delta\boldsymbol{A}_j\| \leqq \|\boldsymbol{r}^j\| \qquad (j = 1, \ldots, m). \tag{60}$$

(ii) Es gibt Eigenwerte $\{\lambda_{l(1)}, \ldots, \lambda_{l(m)}\}$ von $\boldsymbol{A}$ mit

$$|\mu_j - \lambda_{l(j)}| \leqq \|\boldsymbol{R}\| \quad \text{und} \quad \sqrt{\sum_{j=1}^{n} (\mu_j - \lambda_{l(j)})^2} \leqq \sqrt{2}\,\|\boldsymbol{R}\|_F. \tag{61}$$

Beweis. Aussage (i) folgt direkt aus 13.1.7, zum Nachweis von (ii) sei auf die Spezialliteratur verwiesen, siehe B 13.4. □

13.3.16. *Bemerkung.* (i) Aus 13.3.15(i) folgt nach 13.1.2, daß im Intervall $[\mu_j - \|\boldsymbol{r}^j\|, \mu_j + \|\boldsymbol{r}^j\|]$ mindestens ein Eigenwert λ_l von $\boldsymbol{A}$ liegt. Wenn alle diese Intervalle disjunkt sind, ist dies eine schärfere Aussage als (ii). Falls μ_j den Eigenwert λ_l genügend gut approximiert und λ_l von den übrigen Eigenwerten ausreichend getrennt ist, kann auch der Winkel zwischen $\pm\boldsymbol{v}^j$ und dem entsprechenden Eigenvektor $\boldsymbol{u}^l$ abgeschätzt werden. Bei mehrfachen Eigenwerten und Eigenwerthaufen gelten analoge Abschätzungen für die zugehörigen Teilräume, auf die hier jedoch nicht eingegangen werden kann, vgl. 13.1.5.

(ii) Bei Computerrealisierung des RR-Algorithmus ist wesentlich, daß die numerisch berechnete Matrix $\boldsymbol{Q}$ die im Sinne von

$$\|\boldsymbol{Q}^T\boldsymbol{Q} - \boldsymbol{I}_m\| \leqq \nu F_0, \qquad F_0 \text{ akzeptabel}, \tag{62}$$

höchstmögliche Qualität der Orthogonalität hat, vgl. 10.1.8. Falls (62) erfüllt ist und die Eigenwertzerlegung in S2 von 13.3.13 nach einem numerisch gutartigen Verfahren vorgenommen wird, gelten die Aussagen von 13.3.15 auch für die nume-

risch berechneten Ritzschen Näherungen. Dabei ist in den Schranken ein Term der Größenordnung $vF_1\,\|\boldsymbol{A}\|$ mit akzeptablem F_1 zu addieren. □

Wir gehen abschließend auf mögliche Kombinationen der Teilraumiteration mit dem RR-Algorithmus ein.

13.3.17. *Bemerkung.* (i) Die einfachste Möglichkeit zur Kopplung von Teilraumiteration und RR-Algorithmus besteht darin, die nach Schritt S2 von 13.3.10 vorliegende spaltenorthonormale Matrix $\boldsymbol{Q}_{k+1} =: \boldsymbol{Q}$ als Ausgangspunkt für den RR-Algorithmus zu wählen. Dann sollte natürlich nicht $\boldsymbol{Q}_{k+1}$, sondern die denselben Teilraum $\mathcal{Y}_{k+1} = \mathcal{R}(\boldsymbol{Q}_{k+1}) = \mathcal{R}(\boldsymbol{V})$ repräsentierende Matrix $\boldsymbol{V} =: \boldsymbol{V}_{k+1}$ der Ritzschen Eigenvektoren als neue Basismatrix gewählt werden. Bei Konvergenz der Ritzschen Eigenvektoren gegen Eigenvektoren von $\boldsymbol{A}$ ist dann $\boldsymbol{P}$ fast diagonal, vgl. 13.3.14(ii). In dieser Version ist S3 von 13.3.10 also durch

S3′: Setze $\boldsymbol{Q} := \boldsymbol{Q}_{k+1}$, berechne Ritzsche Näherungen $\{\boldsymbol{M}, \boldsymbol{V}\}$ nach 13.3.13, setze $\boldsymbol{M}_{k+1} := \boldsymbol{M}$, $\boldsymbol{V}_{k+1} := \boldsymbol{V}$

zu ersetzen. Pro Schritt sind dann $2m$ Auswertungen von $\boldsymbol{Ax}$ erforderlich, nämlich m in S1 von 13.3.10 und m in S1 von 13.3.13.

(ii) Unter Verwendung der im folgenden Schritt ohnehin zu berechnenden Matrix $\boldsymbol{AV}_{k+1}$ läßt sich die Residualmatrix $\boldsymbol{R} = \boldsymbol{R}_{k+1} = \boldsymbol{AV}_{k+1} - \boldsymbol{V}_{k+1}\boldsymbol{M}_{k+1}$ bilden und damit gemäß 13.3.15 eine Einschätzung der Genauigkeit der Ritzschen Näherungen $\{\boldsymbol{M}_{k+1}, \boldsymbol{V}_{k+1}\}$ vornehmen. Dies kann als Grundlage für ein Abbruchkriterium zur Teilraumiteration genommen werden.

(iii) Für die in (i) beschriebene Variante der sogenannten *simultanen Iteration* gilt unter der Voraussetzung

$$0 \leqq \lambda_n \leqq \cdots \leqq \lambda_{m+1} < \lambda_m < \cdots < \lambda_2 < \lambda_1 \tag{63}$$

die Abschätzung

$$\sin\left(\sphericalangle\,(\boldsymbol{u}^j, \boldsymbol{v}^{k+1,j})\right) = O(|\lambda_{m+1}/\lambda_j|^k) \qquad (j = 1, \ldots, m), \tag{64}$$

sofern die Ritzschen Eigenwerte in Analogie zu (63), also gemäß $\mu_m \leqq \cdots \leqq \mu_2 \leqq \mu_1$ numeriert werden. Die Konvergenz ist also um so besser, je größer $m + 1 - j$ ist. Zur Berechnung der ersten p Eigenwerte ist es daher zweckmäßig, die Spaltenzahl m etwas größer als p zu wählen. Trotz des höheren Aufwandes pro Schritt ist der Gesamtaufwand wegen der schnelleren Konvergenz dann meist niedriger als im Fall $p = m$, sofern λ_{m+1} merklich kleiner als λ_{p+1} ist.

(iv) Die Aussagen von (iii) lassen sich auch auf den indefiniten Fall und nicht notwendig getrennte dominante Eigenwerte $\{\lambda_1, \ldots, \lambda_m\}$ übertragen. Die Formulierung und erst recht die Beweise sind allerdings wesentlich komplizierter, so daß hier darauf verzichtet wird. Bei mehrfachen Eigenwerten bzw. Eigenwerthaufen ist natürlich nicht mit individueller Konvergenz der Ritzschen Eigenvektoren zu rechnen, allerdings approximieren die zugehörigen Teilräume die entsprechenden invarianten Eigenvektorräume von $\boldsymbol{A}$, vgl. 13.1.B.

(v) Mit **nur** m Auswertungen von $\boldsymbol{Ax}$ pro Schritt kommt die folgende, in der

Prozedur ritzit von RUTISHAUSER [71] verwendete Modifikation aus, bei der S3 von 13.3.10 durch

S3.1′: Bilde $\overline{\boldsymbol{P}} := \boldsymbol{R}_{k+1}\boldsymbol{R}_{k+1}^{\mathsf{T}}$

S3.2′: Berechne Eigenwertzerlegung $\overline{\boldsymbol{P}} = \overline{\boldsymbol{X}}\overline{\boldsymbol{M}}^2\overline{\boldsymbol{X}}^{\mathsf{T}}$ von $\overline{\boldsymbol{P}}$ mit diagonalem $\overline{\boldsymbol{M}}^2 = \operatorname{diag}(\bar{\mu}_1^2, \ldots, \bar{\mu}_m^2)$ und orthogonalem $\overline{\boldsymbol{X}}$

S3.3′: Setze $\boldsymbol{V}_{k+1} := \boldsymbol{Q}_{k+1}\overline{\boldsymbol{X}}$

ersetzt wird. Zur Analyse dieser zu empfehlenden Version muß auf die Literatur verwiesen werden, siehe B 13.4.

(vi) Aus Aufwandsgründen ist es zweckmäßig, den RR-Algorithmus nur in gewissen Abständen mit der Teilraumiteration zu koppeln.

(vii) Für alternative Möglichkeiten zur Festlegung der approximierenden Teilräume siehe B 13.5. □

Übungsaufgaben

Ü 13.3.1. Es seien $\lambda_n \leqq \lambda_{n-1} \leqq \cdots \leqq \lambda_2 < \lambda_1$ die Eigenwerte von $\boldsymbol{A}$. Dann hat die durch die Verschiebung $\bar{\boldsymbol{A}} := \boldsymbol{A} - \mu\boldsymbol{I}$ entstehende Matrix $\bar{\boldsymbol{A}}$ die Eigenwerte $\bar{\lambda}_i = \lambda_i - \mu$. Man zeige, daß die Größe

$$\varkappa = \varkappa(\mu; \lambda_1, \ldots, \lambda_n) := \max\{|\bar{\lambda}_i| : i \neq 1\}/|\bar{\lambda}_1|$$

für $\mu_{\text{opt}} := (\lambda_2 + \lambda_n)/2$ minimal wird, und der minimale Wert ist

$$\varkappa_{\text{opt}} = \varkappa(\mu_{\text{opt}}; \lambda_1, \ldots, \lambda_n) = \frac{1-\tau}{1+\tau} \quad \text{mit} \quad 0 < \tau := \frac{\lambda_1 - \lambda_2}{\lambda_1 - \lambda_n} \leqq 1.$$

Wie kann λ_n unter der Voraussetzung $\lambda_n < \lambda_{n-1}$ in einen dominanten Eigenwert von $\bar{\boldsymbol{A}}$ transformiert werden?

Ü 13.3.2. Man zeige unter Verwendung von (18), daß

$$\gamma_k^2 = \|\boldsymbol{w}^{k+1}\|^2 = \lambda_1^2 - [\lambda_1^2 - \|\boldsymbol{A}\boldsymbol{z}^k\|^2] \sin^2 \varphi_k$$

gilt und folgere hieraus (22).

Ü 13.3.3. Man stelle $\omega_{k,j}$ unter Verwendung von (16) und (18) in der Form

$$\omega_{k,j} = \frac{(\vartheta_k \boldsymbol{w}^{k+1})_j}{(\vartheta_k \boldsymbol{v}^k)_j} = \lambda_1 + \frac{[(\boldsymbol{A}\boldsymbol{z}^k)_j - \lambda_1(\boldsymbol{z}^k)_j] \sin \varphi_k}{(\boldsymbol{u}^1)_j \cos \varphi_k + (\boldsymbol{z}^k)_j \sin \varphi_k}$$

dar und leite hieraus die Abschätzung (25) her.

Ü 13.3.4. Es seien $\boldsymbol{P} \in \mathbf{R}^{n,p}$, $\boldsymbol{Q} \in \mathbf{R}^{n,q}$ spaltenorthonormal. Man zeige, daß

$$\|\boldsymbol{P}^{\mathsf{T}}\boldsymbol{Q}\| \leqq 1$$

gilt und der minimale Winkel zwischen den Teilräumen $\mathcal{Y} := \mathcal{R}(\boldsymbol{P})$, $\mathcal{Z} := \mathcal{R}(\boldsymbol{Q})$ durch

$$\cos(\sphericalangle_{\min}(\mathcal{Y}, \mathcal{Z})) = \|\boldsymbol{P}^{\mathsf{T}}\boldsymbol{Q}\|$$

gegeben ist.

Hinweis: Man beachte Ü 1.2.10.

Ü 13.3.5. Es sei $\mathcal{Y} = \mathcal{R}(\boldsymbol{Q})$ der durch die Spalten der spaltenorthonormalen Matrix $\boldsymbol{Q} \in \mathbf{R}^{n,p}$ aufgespannte Teilraum. Dann kann die durch $\boldsymbol{A}$ vermittelte lineare Abbildung $\boldsymbol{x} \in \mathbf{R}^n \to \boldsymbol{Ax} \in \mathbf{R}^n$ wie folgt bezüglich $\mathcal{Y}$ approximiert werden:

$$\boldsymbol{x} \in \mathbf{R}^n \to \boldsymbol{y} := \boldsymbol{P}_{\mathcal{Y}}\boldsymbol{x} \in \mathcal{Y} \to \boldsymbol{Ay} \in \mathbf{R}^n \to \boldsymbol{z} := \boldsymbol{P}_{\mathcal{Y}}\boldsymbol{Ay} \in \mathcal{Y}.$$

Dabei bezeichnet $\boldsymbol{P}_{\mathcal{Y}} := \boldsymbol{QQ}^{\mathsf{T}}$ den Projektor auf $\mathcal{Y}$. Man zeige: Wenn

$$\boldsymbol{y} = \boldsymbol{P}_{\mathcal{Y}}\boldsymbol{x} = \boldsymbol{QQ}^{\mathsf{T}}\boldsymbol{x} = \boldsymbol{Q}\hat{\boldsymbol{y}}, \qquad \hat{\boldsymbol{y}} := \boldsymbol{Q}^{\mathsf{T}}\boldsymbol{x} \quad \text{und} \quad \boldsymbol{z} := \boldsymbol{Q}\hat{\boldsymbol{z}}$$

gesetzt wird, gilt

$$\hat{\boldsymbol{z}} = \boldsymbol{P}\hat{\boldsymbol{y}} \quad \text{mit} \quad \boldsymbol{P} := \boldsymbol{Q}^{\mathsf{T}}\boldsymbol{AQ}.$$

Die Matrix $\boldsymbol{P}$ heißt *Projektion von $\boldsymbol{A}$ auf $\mathcal{Y} = \mathcal{R}(\boldsymbol{Q})$*.

13.4. Inverse Iteration

Die Vektor- und Teilraumiteration aus 13.3 hat den Nachteil, daß sie nur Approximationen für Eigenvektoren bzw. Eigenräume liefert, die zu dominanten Eigenwerten gehören; außerdem ist die Konvergenz i. allg. langsam. Meistens sind jedoch andere Eigenvektoren gesucht, etwa der zu einem vorgegebenen Eigenwert λ_j gehörende Eigenvektor $\boldsymbol{u}^j$, wobei eine Näherung μ_j für λ_j bekannt ist, vgl. 13.1.10, oder die zu p betragskleinsten Eigenwerten gehörenden Eigenvektoren. Solche Eigenwerte lassen sich i. allg. auch durch eine Spektralverschiebung nicht zu dominanten machen. Eine Möglichkeit, auch die beschriebenen Aufgaben sehr effektiv mittels Vektoriteration zu lösen, beruht auf den folgenden beiden Fakten:

— Wenn $\boldsymbol{A} = \boldsymbol{U\Lambda U}^{\mathsf{T}} = \boldsymbol{U}\,\mathrm{diag}\,(\lambda_j)\,\boldsymbol{U}^{\mathsf{T}}$ die Eigenwertzerlegung von $\boldsymbol{A}$ bezeichnet und $\boldsymbol{A}$ regulär ist, gilt $\boldsymbol{A}^{-1} = \boldsymbol{U\Lambda}^{-1}\boldsymbol{U}^{\mathsf{T}} = \boldsymbol{U}\,\mathrm{diag}\,(1/\lambda_j)\,\boldsymbol{U}^{\mathsf{T}}$, d. h., die Inverse besitzt die reziproken Eigenwerte, während die Eigenvektoren unverändert bleiben. Die betragskleinsten Eigenwerte von $\boldsymbol{A}$ gehen also in die betragsgrößten von $\boldsymbol{A}^{-1}$ über.

— Durch eine Spektralverschiebung

$$\bar{\boldsymbol{A}} := \boldsymbol{A}_\mu := \boldsymbol{A} - \mu\boldsymbol{I} \tag{1}$$

kann *jeder* Eigenwert λ_j von $\boldsymbol{A}$ zum betragskleinsten Eigenwert $\bar{\lambda}_j = \lambda_j[\boldsymbol{A}_\mu] = \lambda_j - \mu$ von $\boldsymbol{A}_\mu$ gemacht werden, sofern $|\lambda_j - \mu|$ genügend klein ist, vgl. Abb. 13.4.1.

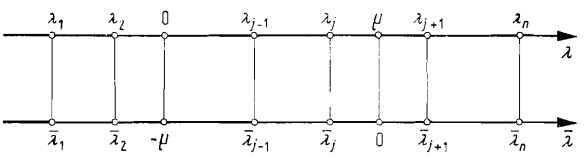


Abb. 13.4.1. Spektralverschiebung $\bar{\lambda}_j = \lambda_j - \mu$

Im folgenden seien die Eigenwerte λ_j von $\boldsymbol{A}$ gemäß

$$|\lambda_1 - \mu| \leqq |\lambda_2 - \mu| \leqq \cdots \leqq |\lambda_n - \mu| \tag{2}$$

numeriert, und es gelte

$$0 < |\lambda_1 - \mu|, \tag{3}$$

d. h., $\boldsymbol{A}_\mu$ sei regulär. Dann existiert $\hat{\boldsymbol{A}} := (\boldsymbol{A} - \mu\boldsymbol{I})^{-1}$, und es ist

$$\hat{\boldsymbol{A}} = (\boldsymbol{A} - \mu\boldsymbol{I})^{-1} = \boldsymbol{U}(\boldsymbol{\Lambda} - \mu\boldsymbol{I})^{-1}\,\boldsymbol{U}^{\mathsf{T}} = \boldsymbol{U}\,\mathrm{diag}\left(1/(\lambda_j - \mu)\right)\boldsymbol{U}^{\mathsf{T}}. \tag{4}$$

Die Matrix $\hat{\boldsymbol{A}}$ besitzt also die Eigenpaare $\{\hat{\lambda}_j, \boldsymbol{u}^j\}$, $\hat{\lambda}_j := 1/(\lambda_j - \mu)$. Insbesondere ist $\hat{\lambda}_1 = 1/(\lambda_1 - \mu)$ wegen (2) betragsgrößter Eigenwert von $\hat{\boldsymbol{A}}$. Der zugehörige Eigenvektor $\boldsymbol{u}^1$ kann dann im Fall der Dominanz von $\hat{\lambda}_1$, d. h. im Fall

$$|\lambda_1 - \mu| < |\lambda_2 - \mu|, \tag{5}$$

durch Vektoriteration mit der Matrix $\hat{\boldsymbol{A}}$ gefunden werden; man spricht von *inverser Iteration.*

13.4.1. *Basisverfahren der inversen Iteration.*
S0 (Initialisierung): Wähle μ mit $\mu \neq \lambda_j$ $(j = 1, \ldots, n)$ und $\boldsymbol{v}^1$ mit $\|\boldsymbol{v}^1\| = 1$, setze $k := 1$
S1 (Iteration): Berechne $\boldsymbol{w}^{k+1} := \hat{\boldsymbol{A}}\boldsymbol{v}^k = (\boldsymbol{A} - \mu\boldsymbol{I})^{-1}\,\boldsymbol{v}^k$
S2 (Normierung): Setze $\boldsymbol{v}^{k+1} := \boldsymbol{w}^{k+1}/\|\boldsymbol{w}^{k+1}\|$
S3: Setze $k := k + 1$, **goto** S1

13.4.2. *Bemerkung.* (i) Bei der Realisierung von S1 wird i. allg. nicht die explizit berechnete Inverse $\boldsymbol{A}_\mu^{-1}$ verwendet, sondern $\boldsymbol{w}^{k+1}$ wird als Lösung des Gleichungssystems

$$\boldsymbol{A}_\mu\boldsymbol{w}^{k+1} = (\boldsymbol{A} - \mu\boldsymbol{I})\,\boldsymbol{w}^{k+1} = \boldsymbol{v}^k \tag{6}$$

bestimmt. Dazu bietet sich eine Dreiecksfaktorisierung von $\boldsymbol{A}_\mu$ an, etwa die Spaltenpivotisierte Gauß-Faktorisierung $\boldsymbol{P}_\mu\boldsymbol{A}_\mu = \boldsymbol{L}_\mu\boldsymbol{R}_\mu$ aus 5.2.A oder die symmetrische BKP-Faktorisierung $\boldsymbol{P}_\mu\boldsymbol{A}_\mu\boldsymbol{P}_\mu^{\mathsf{T}} = \boldsymbol{L}_\mu\boldsymbol{D}_\mu\boldsymbol{L}_\mu^{\mathsf{T}}$ aus 6.1.B. Letztere nutzt zwar die Symmetrie von $\boldsymbol{A}_\mu$ voll aus, kann aber aus einer eventuellen Bandgestalt von $\boldsymbol{A}$ keine Vorteile ziehen. In beiden Fällen gilt für die berechnete Lösung $\boldsymbol{w}^{k+1}$ von (6)

$$(\boldsymbol{A}_\mu + \delta\boldsymbol{A}_k)\,\boldsymbol{w}^{k+1} = \boldsymbol{v}^k \quad \text{mit} \quad \|\delta\boldsymbol{A}_k\| \leqq \nu F\,\|\boldsymbol{A}_\mu\|, \tag{7}$$

wobei F die Kumulationskonstante des Lösungsverfahrens bezeichnet. Man beachte, daß $\boldsymbol{A}_\mu$ i. allg. nicht definit ist, so daß die einfache Cholesky-Faktorisierung nicht verwendet werden kann. Da die Koeffizientenmatrix für alle k dieselbe ist, braucht die Faktorisierung nur einmal berechnet zu werden.

(ii) In wichtigen Anwendungen der inversen Iteration ist $\boldsymbol{A}$ tridiagonal, siehe 13.6. In diesem Fall sollte die spaltenpivotisierte Gauß-Faktorisierung nach 6.4.4 verwendet werden, deren Berechnung dann wie die Lösung von (6) nur $O(n)$ opms kostet. Dasselbe gilt für Bandmatrizen.

(iii) Bei Realisierung in exakter Arithmetik wird das Konvergenzverhalten der Folge $\{\boldsymbol{v}^k\}$ durch 13.3.3 bei Anwendung auf $\boldsymbol{A}_\mu^{-1}$ statt $\boldsymbol{A}$ beschrieben: **Falls** (2), (3),

(5) und (13.3.9) gelten, konvergiert $\{\vartheta_k \boldsymbol{v}^k\}$ gegen $\boldsymbol{u}^1$, und der Konvergenzfaktor ist

$$\hat{\varkappa} := |\lambda_1 - \mu| / |\lambda_2 - \mu|. \tag{8}$$

Die Konvergenz ist also um so schneller, je besser die Eigenwertnäherung μ ist.

(iv) Bei Computerrealisierung wird das ideale Verhalten der Folge $\{\boldsymbol{v}^k\}$ durch die Rundungsfehler gestört. Wegen der erforderlichen Gleichungsauflösung ist die Fehleranalyse etwas aufwendiger als bei der einfachen Vektoriteration. Wir betrachten zunächst den Fall, daß $\boldsymbol{A}_\mu$ im Sinne von

$$vF \operatorname{cond}(\boldsymbol{A}_\mu) = vF\, |\lambda_n - \mu| / |\lambda_1 - \mu| =: \bar{\varkappa} < 1 \tag{9}$$

mit F aus (7) ausreichend regulär ist, d. h., daß μ keine zu gute Approximation für λ_1 darstellt. Dann folgt aus 4.1.2 die Regularität von $\boldsymbol{A}_\mu + \delta\boldsymbol{A}_k$, und man erhält

$$\begin{aligned} &\boldsymbol{u}^{k+1} = (\boldsymbol{A}_\mu + \delta\boldsymbol{A}_k)^{-1} \boldsymbol{v}^k = (\hat{\boldsymbol{A}} + \delta\hat{\boldsymbol{A}}_k)\, \boldsymbol{v}^k \\ &\text{mit} \quad \delta\hat{\boldsymbol{A}}_k := (\boldsymbol{A}_\mu + \delta\boldsymbol{A}_k)^{-1} - \boldsymbol{A}_\mu^{-1} \end{aligned} \tag{10}$$

und

$$\|\delta\hat{\boldsymbol{A}}_k\| \leqq [\|\hat{\boldsymbol{A}}\|^2/(1 - \bar{\varkappa})]\, \|\delta\boldsymbol{A}_k\| \leqq v\hat{F}\, \|\hat{\boldsymbol{A}}\|, \qquad \hat{F} := [\operatorname{cond}(\boldsymbol{A}_\mu)/(1 - \bar{\varkappa})]\, F. \tag{11}$$

Die Voraussetzungen von 13.3.5 sind also mit $\{\hat{\boldsymbol{A}}, \hat{F}\}$ statt $\{\boldsymbol{A}, F\}$ erfüllt, sofern

$$\hat{\varkappa} + 10v(\hat{F} + n/2) = |\lambda_1 - \mu| / |\lambda_2 - \mu| + 10v(\hat{F} + n/2) < 1 \tag{12}$$

gilt.

(v) Die in (iv) beschriebene Situation liegt in der Regel vor, wenn die inverse Iteration mit $\mu = 0$ — also ohne Verschiebungen — zur Bestimmung des betragskleinsten Eigenwertes verwendet wird. Falls die p betragskleinsten Eigenwerte und zugehörigen Eigenvektoren gesucht sind, sollte natürlich die Teilraumiteration mit $\hat{\boldsymbol{A}} := \boldsymbol{A}_0^{-1} := \boldsymbol{A}^{-1}$ statt $\boldsymbol{A}$ benutzt werden. Die Dominanzforderung (13.3.2) geht dann in

$$0 < |\lambda_1| \leqq \cdots \leqq |\lambda_p| < |\lambda_{p+1}| \tag{13}$$

über, und die Aussagen aus 13.3.B,C sind sinngemäß gültig. □

Falls $|\lambda_1 - \mu|$ sehr klein ist, kann nicht mehr wie in (iv) geschlossen werden, da (9) i. allg. nicht erfüllt ist. Wir gehen auf diesen wichtigen Spezialfall im folgenden genauer ein. Es sei dazu μ eine im Sinne von

$$|\mu - \lambda| \leqq vF_0\, \|\boldsymbol{A}\| =: \varepsilon_0 \tag{14}$$

sehr gute Näherung für irgendeinen Eigenwert λ von $\boldsymbol{A}$. Wir numerieren die Eigenwerte wieder gemäß (2), d. h., $\lambda = \lambda_1$ sei der μ am nahesten gelegene Eigenwert. Gesucht ist ein Vektor $\boldsymbol{v}$ mit $\|\boldsymbol{v}\| = 1$, so daß $\{\mu, \boldsymbol{v}\}$ eine akzeptable Eigenpaarapproximation darstellt. Wenn die Norm des Residuums $\boldsymbol{r} = (\boldsymbol{A} - \mu\boldsymbol{I})\, \boldsymbol{v}$ als Kriterium für die Güte von $\boldsymbol{v}$ verwendet wird, ist nach 13.1.10 jeder zu λ_1 gehörende Eigenvektor $\boldsymbol{u}^1$ optimal, so daß sich die inverse Iteration zur Berechnung von $\boldsymbol{v}$ anbietet. Dabei ist es zweckmäßig, die folgende spezielle Version zu verwenden.

13.4.3. *Berechnung einer akzeptablen Eigenvektorapproximation $\boldsymbol{v}$ bei gegebener Eigenwertnäherung μ mittels inverser Iteration.*

Aufgabe: Zur gegebenen Eigenwertapproximation μ mit (14) ist $\boldsymbol{v}$ mit $\|\boldsymbol{v}\| = 1$ so zu bestimmen, daß $\{\mu, \boldsymbol{v}\}$ ein akzeptables Eigenpaar von $\boldsymbol{A}$ ist.

Algorithmus:

S0 (Initialisierung):

S0.1: Berechne Faktorisierung $\boldsymbol{A}_\mu = \boldsymbol{P}_\mu^{\mathsf{T}} \boldsymbol{L}_\mu \boldsymbol{R}_\mu$ gemäß 5.2.3 mit regulärem $\boldsymbol{R}_\mu$

S0.2: Berechne $\boldsymbol{w}^1$ als Lösung von $\boldsymbol{R}_\mu \boldsymbol{w}^1 = \boldsymbol{e} := (1, \ldots, 1)^{\mathsf{T}}$

S0.3: Setze $\boldsymbol{v}^1 := \boldsymbol{w}^1 / \|\boldsymbol{w}^1\|$

S0.4: Setze $k := 1$ und $\varepsilon := \sqrt{n}\,(\varepsilon_0 + \varepsilon_1) = \nu\sqrt{n}\,(F_0 + F_1)\,\|\boldsymbol{A}\|$ mit F_0 aus (14) und $F_1 := (2 + \nu F_0)\,F \doteq 2F$, $\varepsilon_1 := \nu F_1 \|\boldsymbol{A}\|$, F aus (7)

S1 (Iteration): Bestimme $\boldsymbol{w}^{k+1}$ als Lösung von $\boldsymbol{P}_\mu^{\mathsf{T}} \boldsymbol{L}_\mu \boldsymbol{R}_\mu \boldsymbol{w}^{k+1} = \boldsymbol{v}^k$

S2 (Normierung): Setze $\omega_{k+1} := \|\boldsymbol{w}^{k+1}\|$ und $\boldsymbol{v}^{k+1} := \boldsymbol{w}^{k+1}/\omega_{k+1}$

S3 (Abbruchtest): **if** $\varepsilon * \omega_{k+1} < 1$ **then** [$k := k + 1$, **goto** S1]

Aufwand: $\sim n^3/3$ opms (für S0) + $\sim kn^2$ opms (für k Iterationsschritte S1, S2, falls $\boldsymbol{A}$ voll besetzt ist)

13.4.4. *Aussage.* Für die Computerrealisierung von 13.4.3 gilt:

(i) Falls das Verfahren nach k Schritten wegen $\varepsilon * \omega_{k+1} \geqq 1$ mit $\boldsymbol{v} := \boldsymbol{v}^{k+1}$ abbricht, gibt es eine Störung $\delta\boldsymbol{A} \in \mathbf{S}^{n,n}$, so daß $\{\mu, \boldsymbol{v}\}$ der Beziehung

$$(\boldsymbol{A} + \delta\boldsymbol{A})\,\boldsymbol{v} = \mu\boldsymbol{v} \quad \text{mit} \quad \|\delta\boldsymbol{A}\| \leqq \nu\left\{\sqrt{n}\,F_0 + \left(\sqrt{n} + 1\right)F_1\right\}\|\boldsymbol{A}\| \sim \varepsilon\,\|\boldsymbol{A}\| \tag{15}$$

genügt, d. h., $\{\mu, \boldsymbol{v}\}$ ist ein akzeptables Eigenpaar von $\boldsymbol{A}$.

(ii) Das Abbruchkriterium $\varepsilon * \omega_{k+1} \geqq 1$ ist fast immer nach wenigen Schritten $k = 1, 2, 3, \ldots$ erfüllt.

Beweis. Aus (14) folgt $|\mu| \leqq |\lambda_1| + \varepsilon_0 \leqq (1 + \nu F_0)\,\|\boldsymbol{A}\|$, so daß in (7) weiter gemäß

$$\|\delta\boldsymbol{A}_k\| \leqq \nu F\,\|\boldsymbol{A}_\mu\| \leqq \nu F[\|\boldsymbol{A}\| + |\mu|] \leqq \nu F(2 + \nu F_0)\,\|\boldsymbol{A}\| = \nu F_1\,\|\boldsymbol{A}\| = \varepsilon_1 \tag{16}$$

abgeschätzt werden kann.

Zu (i): Aus (7) und der Iterationsvorschrift folgt

$$\boldsymbol{r} = (\boldsymbol{A} - \mu\boldsymbol{I})\,\boldsymbol{v} = (\boldsymbol{A} - \mu\boldsymbol{I})\,\boldsymbol{v}^{k+1} = \boldsymbol{v}^k/\omega_{k+1} - \delta\boldsymbol{A}_k \boldsymbol{v}^{k+1}.$$

Wegen $\varepsilon * \omega_{k+1} \geqq 1$ und (16) zieht dies

$$\|\boldsymbol{r}\| \leqq 1/\omega_{k+1} + \varepsilon_1 \leqq \varepsilon + \varepsilon_1 = \nu\left\{\sqrt{n}\,F_0 + \left(\sqrt{n} + 1\right)F_1\right\}\|\boldsymbol{A}\|$$

und wegen 13.1.7 schließlich die Behauptung nach sich.

Zu (ii): Wir skizzieren den Beweis nur. Nach (7) und (16) gilt

$$(\tilde{\boldsymbol{A}} - \mu\boldsymbol{I})\,\boldsymbol{w}^2 = \boldsymbol{v}^1 \quad \text{mit} \quad \tilde{\boldsymbol{A}} := \boldsymbol{A} + \delta\boldsymbol{A}_1, \qquad \|\delta\boldsymbol{A}_1\| \leqq \varepsilon_1 \tag{17}$$

für das berechnete $\boldsymbol{w}^2$. Wir nehmen an, daß $\delta\boldsymbol{A}_1$ und damit $\tilde{\boldsymbol{A}}$ symmetrisch ist, siehe dazu 13.4.5(ii) unten. Es sei $\{\tilde{\lambda}_j, \tilde{\boldsymbol{u}}^j\}$ ein orthonormales Eigensystem von $\tilde{\boldsymbol{A}}$. Wegen (17) und 13.1.2 kann dieses System so indiziert werden, daß $|\lambda_1 - \tilde{\lambda}_1| \leqq \|\delta\boldsymbol{A}_1\| \leqq \varepsilon_1$ gilt, woraus unter Beachtung von (14)

$$|\tilde{\lambda}_1 - \mu| \leqq |\tilde{\lambda}_1 - \lambda_1| + |\lambda_1 - \mu| \leqq \varepsilon_1 + \varepsilon_0 = \varepsilon/\sqrt{n} \tag{18}$$

folgt. Mit $\tilde{\boldsymbol{U}} := (\tilde{\boldsymbol{u}}^1, \ldots, \tilde{\boldsymbol{u}}^n)$, $\boldsymbol{z} := \tilde{\boldsymbol{U}}^T \boldsymbol{v}^1$, $\boldsymbol{y} := \tilde{\boldsymbol{U}}^T \boldsymbol{w}^2$ läßt sich nun (17) in der Form $(\tilde{\lambda}_i - \mu)\, y_i = z_i$ $(i = 1, \ldots, n)$ schreiben. Im Fall $z_i \neq 0$ muß daher $\tilde{\lambda}_i - \mu \neq 0$ und $y_i = z_i/(\tilde{\lambda}_i - \mu)$ gelten. Daher ist

$$\omega_2^2 = \|\boldsymbol{w}^2\|^2 = \|\boldsymbol{y}\|^2 = \sum_{i=1}^{n} y_i^2 \geqq \sum_{i:\, z_i \neq 0} z_i^2/(\tilde{\lambda}_i - \mu)^2 \geqq n \left\{ \sum_{i \in I} z_i^2 \right\} \Big/ \varepsilon^2,$$

wobei I die Menge derjenigen Indizes bezeichnet, für die $z_i \neq 0$ und $|\tilde{\lambda}_i - \mu| \leqq \varepsilon/\sqrt{n}$ gilt. Wegen (18) gehört mindestens $i = 1$ zu dieser Indexmenge, sofern $z_1 = \tilde{\boldsymbol{u}}^{1T} \boldsymbol{v}^1 \neq 0$ ist, was fast immer erwartet werden kann. Wenn $\boldsymbol{v}^1$ zufällig gewählt worden wäre, ergäbe sich unter einfachen Voraussetzungen an die Verteilung der Erwartungswert $E(|z_1|) = 1/\sqrt{n}$. In den meisten Fällen wird daher

$$|z_1| \geqq 1/\sqrt{n} \tag{19}$$

gelten. Aus (19) folgt aber $\varepsilon * \omega_2 \geqq \sqrt{n}\, |z_1| \geqq 1$, und die Iteration bricht mit $\boldsymbol{v} = \boldsymbol{v}^2$ nach dem ersten Schritt ab. Der Startvektor $\boldsymbol{v}^1$ wird aber nicht zufällig gewählt, sondern $\boldsymbol{v}^1 = \boldsymbol{w}^1/\|\boldsymbol{w}^1\|$ ergibt sich aus $\boldsymbol{w}^1$, und $\boldsymbol{w}^1$ ist Lösung von

$$(\boldsymbol{A} - \mu \boldsymbol{I})\, \boldsymbol{w}^1 = \boldsymbol{v}^0 := \boldsymbol{P}_\mu^T \boldsymbol{L}_\mu \boldsymbol{e}.$$

Dies entspricht einem zusätzlichen Schritt mit dem speziellen Startvektor $\boldsymbol{v}^0$ und erhöht die Chance für einen Abbruch mit $\boldsymbol{v} = \boldsymbol{v}^2$ wesentlich. Falls der Abbruchtest in S3 für $k = 1$ noch nicht erfüllt sein sollte, zeigt eine genauere — allerdings recht komplizierte — Analyse, daß ω_{k+1} in den weiteren Schritten sehr schnell wächst, so daß im ungünstigsten Fall nach 2, 3 Schritten ein akzeptables Paar $\{\mu, \boldsymbol{v}\}$ vorliegt. □

13.4.5. *Bemerkung.* (i) Es ist $\operatorname{cond}(\boldsymbol{A}_\mu) = |\lambda_n - \mu|/|\lambda_1 - \mu| \geqq [|\lambda_n - \mu|/(\|\boldsymbol{A}\|\, F_0)]/\nu$ wegen (14), d. h., für akzeptables F_0 ist $\operatorname{cond}(\boldsymbol{A}_\mu)$ in der Größenordnung $1/\nu$. Die für die Durchführbarkeit der Gaußschen Dreiecksfaktorisierung mit regulärem $\boldsymbol{R}_\mu$ hinreichende Bedingung (9) ist daher i. allg. nicht erfüllt, vgl. 5.3.2. Praktisch wird $\boldsymbol{R}_\mu$ natürlich fast immer regulär sein, allerdings werden gewisse Pivots $(\boldsymbol{R}_\mu)_{jj}$ betragsmäßig sehr klein werden. Sollte ausnahmsweise sogar $(\boldsymbol{R}_\mu)_{jj} = 0$ gelten, so wird $(\boldsymbol{R}_\mu)_{jj} := \nu\, \|\boldsymbol{A}\|$ gesetzt, vgl. 5.3.3(i). Dabei bleibt (7) gültig.

(ii) Die im Beweis von 13.4.4 vorausgesetzte Symmetrie der Störung $\delta \boldsymbol{A}_k$ ist keine Einschränkung: Ist (17) mit einer nichtsymmetrischen Störung erfüllt, so folgt nach 4.1.18 für das Residuum die Abschätzung $\|\boldsymbol{r}^{k+1}\| = \|\boldsymbol{v}^k - \boldsymbol{A}_\mu \boldsymbol{w}^{k+1}\| \leqq \varepsilon_1$. Nach 4.1.20(ii) kann dann $\delta \boldsymbol{A}_k$ durch eine äquivalente symmetrische Störung $\delta \mathring{\boldsymbol{A}}_k$, für die (17) ebenfalls gilt, ersetzt werden.

(iii) Praktisch wird man ε in der Größenordnung von $\nu \sqrt{n}\, \|\boldsymbol{A}\|$ wählen und dabei $\|\boldsymbol{A}\|$ z. B. durch $\|\boldsymbol{A}\|_\infty$ nach oben abschätzen.

(iv) Wenn mehrere im Sinne von (14) akzeptable Eigenwertnäherungen $\mu, \mu', \mu'', \ldots$ vorliegen, kann 13.4.3 für jede von diesen durchgeführt werden und liefert Eigenvektornäherungen $\boldsymbol{v}, \boldsymbol{v}', \boldsymbol{v}'', \ldots$ Die Eigenpaarapproximationen $\{\mu, \boldsymbol{v}\}, \{\mu', \boldsymbol{v}'\}, \{\mu'', \boldsymbol{v}''\}, \ldots$ sind dann alle exakte Eigenpaare gestörter Matrizen $\boldsymbol{A} + \delta \boldsymbol{A}, \boldsymbol{A} + \delta \boldsymbol{A}', \ldots$ mit $\|\delta \boldsymbol{A}\|, \|\delta \boldsymbol{A}'\|, \ldots \leqq \varepsilon\, \|\boldsymbol{A}\| =: \hat{\varepsilon}$, also akzeptabel. Da die Störungen $\delta \boldsymbol{A}, \delta \boldsymbol{A}', \ldots$ jedoch i. allg. verschieden sind, brauchen die berechneten Eigenvektoren auch im Fall verschiedener Eigenwerte nicht exakt orthogonal zu sein; nach 13.1.B ist mit einer Abweichung von

$$|\boldsymbol{v}^T \boldsymbol{v}'| \lessapprox \|\delta \boldsymbol{A} - \delta \boldsymbol{A}'\|/|\mu - \mu'| \leqq 2\hat{\varepsilon}/|\mu - \mu'| \quad \text{usw.} \tag{20}$$

zu rechnen. Insbesondere bei dicht benachbarten und speziell bei mehrfachen Eigenwerten als deren Grenzfall ist daher eine Re-Orthogonalisierung nach dem MGS-Verfahren aus 10.1 gemäß

$$\boldsymbol{v}' := \boldsymbol{v}' - (\boldsymbol{v}^{\mathsf{T}}\boldsymbol{v}')\,\boldsymbol{v}, \qquad \boldsymbol{v}' := \boldsymbol{v}'/\|\boldsymbol{v}'\|,$$

$$\boldsymbol{v}'' := \boldsymbol{v}'' - (\boldsymbol{v}^{\mathsf{T}}\boldsymbol{v}'')\,\boldsymbol{v}, \qquad \boldsymbol{v}'' := \boldsymbol{v}'' - \left((\boldsymbol{v}')^{\mathsf{T}}\,\boldsymbol{v}''\right)\boldsymbol{v}', \qquad \boldsymbol{v}'' := \boldsymbol{v}''/\|\boldsymbol{v}''\|$$

.....

erforderlich. Die nach der Orthogonalisierung vorliegenden Paare $\{\mu, \boldsymbol{v}\}$, $\{\mu', \boldsymbol{v}'\}$, ... sind in fast demselben Sinne akzeptabel wie vorher. □

Normalerweise deutet ein Wert von cond $(\boldsymbol{A}_\mu)$ in der Größenordnung $1/\nu$ darauf hin, daß die berechnete Lösung $\boldsymbol{w}$ von $\boldsymbol{A}_\mu\boldsymbol{w} = \boldsymbol{v}$ einen sehr großen Fehler haben kann und daher praktisch wertlos ist, vgl. 4.1.A. Im Fall der inversen Iteration 13.4.3 trifft der erste Teil dieser Behauptung auch zu, allerdings zeigt der Beweis zu 13.4.4, daß sich dieser i. allg. große Fehler praktisch nicht auf die Güte der berechneten Richtung $\boldsymbol{w}/\|\boldsymbol{w}\|$ auswirkt, sofern ein numerisch gutartiges Verfahren verwendet wird und $\boldsymbol{w}$ im Sinne von $\|\boldsymbol{w}\| = O\big(1/(\nu\,\|\boldsymbol{A}\|)\big)$ genügend groß ist. Das berechnete $\boldsymbol{w}$ ist dann i. allg. eine sehr schlechte Näherung für die exakte Lösung $\boldsymbol{A}_\mu^{-1}\boldsymbol{v}$ des Gleichungssystems, aber die Richtung $\boldsymbol{w}/\|\boldsymbol{w}\|$ approximiert die von $\boldsymbol{u}^1$ gut, falls λ_1 ausreichend isoliert ist. Dies ist auch durch die folgenden, mehr geometrischen Überlegungen einzusehen: Für das berechnete $\boldsymbol{w}$ gilt wegen der Gutartigkeit $(\boldsymbol{A}_\mu + \delta\boldsymbol{A})\,\boldsymbol{w} = \boldsymbol{v}$ mit $\|\delta\boldsymbol{A}\| \leqq \nu F\,\|\boldsymbol{A}_\mu\| =: \varepsilon$ und akzeptablem F. Wir betrachten den Fehler $\delta\boldsymbol{w} := \boldsymbol{w} - \boldsymbol{w}^*$ von $\boldsymbol{w}$ bezüglich der exakten Lösung $\boldsymbol{w}^* = \boldsymbol{A}_\mu^{-1}\boldsymbol{v}$. Nach 4.1.7 liegt der linearisierte Fehler $\delta\boldsymbol{w}'$ in einem Ellipsoid $\mathscr{E}'$ um den Nullpunkt und mit den Halbachsen $\alpha_j\boldsymbol{v}^j = (\Delta\boldsymbol{z}/\sigma_j)\,\boldsymbol{v}^j$, wobei $\Delta\boldsymbol{z} := \varepsilon\,\|\boldsymbol{w}^*\|$ gilt und σ_j und $\boldsymbol{v}^j$ die Singulärwerte bzw. Singulärvektoren von $\boldsymbol{A}_\mu$ sind. Wegen der Symmetrie ist $\sigma_j = |\hat{\lambda}_j - \mu|$ und $\boldsymbol{v}^j = \operatorname{sgn}(\lambda_j - \mu)\,\boldsymbol{u}^j$. Unter der zusätzlichen Voraussetzung $|\lambda_1 - \mu| \ll |\lambda_2 - \mu|$ ist $\alpha_1 \gg \alpha_2 \geqq \cdots \geqq \alpha_n$, d. h., das Ellipsoid entartet dann in ein nadelförmiges Gebilde, dessen Längsachse die Richtung des zu λ_1 gehörenden Eigenvektors $\boldsymbol{u}^1$ hat. Die bezüglich $\delta\boldsymbol{A}$ linearisierte Lösung $\boldsymbol{w}' := \boldsymbol{w}^* + \delta\boldsymbol{w}'$ liegt dann in dem verschobenen Ellipsoid $\boldsymbol{w}^* + \mathscr{E}'$, siehe Abb. 13.4.2. Der Einfachheit halber lassen wir im folgenden das Zeichen „'" weg.

Man sieht, daß $\delta\boldsymbol{w}$ fast immer die Richtung von $\boldsymbol{u}^1$ haben wird. Bei gleichsinniger Orientierung von $\boldsymbol{w}^*$ und $\delta\boldsymbol{w}$ wirkt sich selbst ein sehr großer Fehler günstig auf die Approximation von $\boldsymbol{u}^1$ aus, bei ungleichsinniger verschlechtert er diese nur wenig,

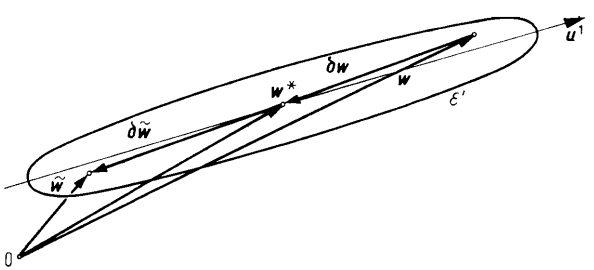


Abb. 13.4.2. Fehlerellipsoid bei inverser Iteration

sofern $\|\boldsymbol{w}\|$ genügend groß bleibt. Nur im praktisch unwahrscheinlichen Sonderfall $\boldsymbol{\delta w} \approx -\boldsymbol{w}^*$ — also für kleines $\|\boldsymbol{w}\|$ — weicht die durch $\boldsymbol{w}$ festgelegte Richtung wesentlich von $\boldsymbol{u}^1$ ab. Falls ein Lösungsverfahren verwendet wird, das nur numerisch stabil mit derselben Konstanten F ist, müßte das Ellipsoid durch die umschriebene Kugel ersetzt werden. Es könnten dann auch große Fehler orthogonal zu $\boldsymbol{u}^1$ auftreten, die zu einem völlig unbrauchbaren $\boldsymbol{w}$ führen.

Wenn λ_1 ein p-facher Eigenwert ist, entartet $\mathscr{E}'$ etwa im Fall $p = 2$ in ein scheibenförmiges Gebilde, das sich nur wenig von einem Kreis mit dem großen Radius α_1 in der durch $\{\boldsymbol{u}^1, \boldsymbol{u}^2\}$ aufgespannten Ebene — dem Eigenraum zu λ_1 — unterscheidet. Auch dann approximiert $\boldsymbol{w}$ fast immer einen in dieser Ebene liegenden Eigenvektor zu λ_1 gut. Dasselbe ist der Fall, wenn λ_1 zu einem aus p benachbarten Eigenwerten bestehenden Eigenwerthaufen gehört, der von den restlichen genügend getrennt ist.

Falls μ keine gute Eigenwertapproximation ist, liegt es nahe, während der inversen Iteration 13.4.1 die Näherung μ sukzessive mit zu verbessern. Da bei festem $\boldsymbol{v} = \boldsymbol{v}^k$ der Rayleigh-Quotient $\varrho(\boldsymbol{v}^k)$ nach 13.1.8 die im Sinne einer minimalen Residuumsnorm optimale Eigenwertnäherung ist, bietet sich dazu die Festlegung $\mu = \mu_k = \varrho(\boldsymbol{v}^k)$ an. Dies liefert die *Rayleigh-Quotienten-Iteration* — kurz: *RQ-Iteration* — genannte Modifikation der inversen Iteration.

13.4.6. *Rayleigh-Quotienten-Iteration.*

S0 (Initialisierung): Wähle $\boldsymbol{v}^1$ mit $\|\boldsymbol{v}^1\| = 1$ und $\varepsilon > 0$ in der Größenordnung $\nu \sqrt{n}\, \|\boldsymbol{A}\|$, setze $k := 1$

S1 (Iteration)

S1.1 (Berechnung der Verschiebung ϱ_k): Bestimme $\varrho_k := \varrho(\boldsymbol{v}^k) = \boldsymbol{v}^{k\mathsf{T}} \boldsymbol{A} \boldsymbol{v}^k$

S1.2 (Inverse Iteration mit $\mu = \varrho_k$): Berechne $\boldsymbol{w}^{k+1}$ aus $(\boldsymbol{A} - \varrho_k \boldsymbol{I})\, \boldsymbol{w}^{k+1} = \boldsymbol{v}^k$

S2 (Normierung): Berechne $\omega_{k+1} := \|\boldsymbol{w}^{k+1}\|$, setze $\boldsymbol{v}^{k+1} := \boldsymbol{w}^{k+1}/\omega_{k+1}$

S3 (Abbruchtest): **if** $\varepsilon * \omega_{k+1} < 1$ **then** [$k := k + 1$, **goto** S1]

Aufwand pro Schritt: $\sim n^2$ opms (für S1.1) + $\sim K_1 n^3$ opms (für S1.2), falls $\boldsymbol{A}$ voll besetzt ist.

Im Unterschied zu 13.4.1 bzw. 13.4.3 sind bei der RQ-Iteration die Koeffizientenmatrizen der in jedem Schritt zu lösenden Gleichungssysteme verschieden, so daß die Dreiecksfaktorisierungen jedesmal neu berechnet werden müssen. In den Anwendungen ist $\boldsymbol{A}$ jedoch meist von Bandgestalt und insbesondere tridiagonal, so daß dies nur $\sim Kn$ opms kostet, vgl. 6.4.A.

13.4.7. *Aussage.* Die RQ-Iteration 13.4.6 werde mit $\varepsilon = 0$ in exakter Arithmetik durchgeführt. Dann gilt:

(i) Die Normen der Residuen $\boldsymbol{r}^k := (\boldsymbol{A} - \varrho_k \boldsymbol{I})\, \boldsymbol{v}^k$ fallen monoton, d. h., es ist

$$\|\boldsymbol{r}^{k+1}\| \leqq \|\boldsymbol{r}^k\|, \tag{21}$$

und die Folge $\{\varrho_k, \vartheta_k \boldsymbol{v}^k\}$ mit $\vartheta_k \in \{+1, -1\}$ so, daß $(\vartheta_k \boldsymbol{v}^k)^{\mathsf{T}}\, \boldsymbol{u}^j \geqq 0$ gilt, konvergiert fast immer gegen ein Eigenpaar $\{\lambda_j, \boldsymbol{u}^j\}$ von $\boldsymbol{A}$.

(ii) Falls $\{\varrho_k\}$ gegen den Eigenwert λ_j konvergiert, ist die Konvergenz kubisch im Sinne der Gültigkeit von

$$\tan \varphi_{k+1} \leqq \eta_1 (\tan \varphi_k)^3 \quad \text{mit} \quad \eta_1 \approx 2\,\|\boldsymbol{A}\|/\gamma \tag{22}$$

für $k \geqq k_0$, k_0 hinreichend groß, wobei $\varphi_k := \sphericalangle\,(\boldsymbol{u}^j, \vartheta_k \boldsymbol{v}^k)$ und

$$\gamma := \min\{|\lambda_j - \lambda_i| : \lambda_i \neq \lambda_j\} \tag{23}$$

ist. Es gilt zudem

$$|\lambda_j - \varrho_k| \leqq 2\,\|\boldsymbol{A}\| \tan^2 \varphi_k \tag{24}$$

sowie

$$\|\boldsymbol{r}^k\| \leqq \eta_2 \tan \varphi_k \quad \text{mit} \quad \eta_2 \approx 2\,\|\boldsymbol{A}\|. \tag{25}$$

Beweis. Wir beginnen mit (ii): Es gelte also $\lim \varrho_k = \lambda_j$, und o. B. d. A. sei $\lambda_j = \lambda_1$. Falls φ_k wie oben definiert ist, liefert (13.3.19) gerade die Abschätzung (24). Zum Beweis von (22) beachten wir, daß $\boldsymbol{v}^{k+1}$ aus $\boldsymbol{v}^k$ durch einen Schritt der Vektoriteration mit $\hat{\boldsymbol{A}} := (\boldsymbol{A} - \varrho_k \boldsymbol{I})^{-1}$ entsteht. Nach 13.3.3 gilt daher

$$\tan \varphi_{k+1} \leqq \varkappa \tan \varphi_k \tag{26}$$

mit

$$\varkappa := \frac{\max\{|\hat{\lambda}_i| : i \neq 1\}}{|\hat{\lambda}_1|} = \frac{|\lambda_1 - \varrho_k|}{\min\{|\lambda_i - \varrho_k| : i \neq 1\}}. \tag{27}$$

Es sei zunächst λ_1 einfach, also $\gamma = \min\{|\lambda_1 - \lambda_i| : i \neq 1\} > 0$, und k_0 sei so groß, daß $|\lambda_1 - \varrho_k| < \gamma$ für $k \geqq k_0$ gilt. Dann folgt aus (26), (27) und $|\lambda_i - \varrho_k| \geqq |\lambda_1 - \lambda_i| - |\lambda_1 - \varrho_k|$ $\geqq \gamma - |\lambda_1 - \varrho_k| > 0$

$$\tan \varphi_{k+1} \leqq \frac{2\,\|\boldsymbol{A}\|\,(\tan \varphi_k)^3}{\gamma - 2\,\|\boldsymbol{A}\| \tan^2 \varphi_k},$$

also (22). Zum Nachweis von (25) stellen wir $\boldsymbol{r}^k$ in der Form

$$\vartheta_k \boldsymbol{r}^k = (\boldsymbol{A} - \varrho_k \boldsymbol{I})\,\hat{\boldsymbol{v}}^k = \boldsymbol{A}(\hat{\boldsymbol{v}}^k - \boldsymbol{u}^1) + \lambda_1(\boldsymbol{u}^1 - \hat{\boldsymbol{v}}^k) + (\lambda_1 - \varrho_k)\,\hat{\boldsymbol{v}}^k$$

mit $\hat{\boldsymbol{v}}^k := \vartheta_k \boldsymbol{v}^k$ dar. Mit (13.3.15) und (24) ergibt sich dann

$$\|\boldsymbol{r}^k\| \leqq 2\,\|\boldsymbol{A}\|\,\|\hat{\boldsymbol{v}}^k - \boldsymbol{u}^1\| + |\lambda_1 - \varrho_k| \leqq 2\,\|\boldsymbol{A}\|\,(\tan \varphi_k + \tan^2 \varphi_k),$$

also (25). Falls λ_1 mehrfach ist, kann analog mit γ gemäß (23) geschlossen werden, vgl. 13.3.7 (i).

Zu (i): Für die Residuumsnormen gilt wegen der Optimalität des Rayleigh-Quotienten

$$\begin{aligned}\|\boldsymbol{r}^{k+1}\| &= \|(\boldsymbol{A} - \varrho_{k+1}\boldsymbol{I})\,\boldsymbol{v}^{k+1}\| \leqq \|(\boldsymbol{A} - \varrho_k \boldsymbol{I})\,\boldsymbol{v}^{k+1}\| = |\boldsymbol{v}^{k\mathsf{T}}(\boldsymbol{A} - \varrho_k \boldsymbol{I})\,\boldsymbol{v}^{k+1}| \\ &= |\boldsymbol{v}^{k+1\mathsf{T}}(\boldsymbol{A} - \varrho_k \boldsymbol{I})\,\boldsymbol{v}^k| \leqq \|(\boldsymbol{A} - \varrho_k \boldsymbol{I})\,\boldsymbol{v}^k\| = \|\boldsymbol{r}^k\|.\end{aligned}$$

Das zweite Gleichheitszeichen folgt aus der Tatsache, daß $\boldsymbol{y}^k := (\boldsymbol{A} - \varrho_k \boldsymbol{I})\,\boldsymbol{v}^{k+1} = \boldsymbol{v}^k/\omega_{k+1}$ gilt, also $\boldsymbol{y}^k$ gleich seiner Projektion auf $\boldsymbol{v}^k$ ist. Zur Konvergenzuntersuchung beachten wir, daß neben $\|\boldsymbol{v}^k\| = 1$ noch $|\varrho_k| \leqq \|\boldsymbol{A}\|$ gilt, d. h., die Paare $\{\varrho_k, \boldsymbol{v}^k\}$ liegen in einer beschränkten und abgeschlossenen Menge. Es gibt dann nach dem Satz von WEIERSTRASS eine konvergente Teilfolge $\{\varrho_{k_i}, \boldsymbol{v}^{k_i}\}$ mit dem Grenzwert $\{\bar{\varrho}, \bar{\boldsymbol{v}}\}$, $|\bar{\varrho}| \leqq \|\boldsymbol{A}\|$, $\|\bar{\boldsymbol{v}}\| = 1$, und aus Stetigkeitsgründen gilt

$$\lim_{i \to \infty} \|(\boldsymbol{A} - \varrho_{k_i}\boldsymbol{I})\,\boldsymbol{v}^{k_i}\| = \lim_{i \to \infty} \|\boldsymbol{r}^{k_i}\| = \|(\boldsymbol{A} - \bar{\varrho}\boldsymbol{I})\,\bar{\boldsymbol{v}}\| =: \sigma. \tag{28}$$

Wegen der Monotonie der Residuen existiert nun stets der Grenzwert

$$\lim_{k\to\infty} \|\boldsymbol{r}^k\| = \sigma \geqq 0. \tag{29}$$

Falls $\sigma = 0$ ist, folgt $\|(\boldsymbol{A} - \bar{\varrho}\boldsymbol{I})\,\bar{\boldsymbol{v}}\| = 0$ aus (28) und (29), d. h., $\{\bar{\varrho}, \bar{\boldsymbol{v}}\} = \{\lambda_j, \boldsymbol{u}^j\}$ ist ein Eigenpaar von $\boldsymbol{A}$. Es gibt dann ein genügend großes $l := k_i$, so daß $\eta_1 \tan^2 \varphi_l \leqq q < 1$ gilt. Aus dem Beweis zu (ii) folgt dann

$$\tan \varphi_{k+1} \leqq [\eta_1 \tan^2 \varphi_k] \tan \varphi_k \leqq q \tan \varphi_k$$

für alle $k \geqq l$, d. h., die gesamte Folge $\{\varrho_k, \vartheta_k \boldsymbol{v}^k\}$ konvergiert gegen $\{\lambda_j, \boldsymbol{v}^j\}$. Der praktisch nicht auftretende Fall $\sigma > 0$ ist komplizierter zu analysieren, wir verweisen auf die Literatur, siehe B 13.6. □

13.4.8. *Bemerkung.* (i) Bei exakter Realisierung von 13.4.6 kann durch Zufall $\boldsymbol{A} - \varrho_k \boldsymbol{I}$ singulär sein. Dann ist ϱ_k ein Eigenwert von $\boldsymbol{A}$, und ein zugehöriger Eigenvektor läßt sich aus dem homogenen Gleichungssystem $(\boldsymbol{A} - \varrho_k \boldsymbol{I})\,\boldsymbol{v} = \boldsymbol{o}$ berechnen. Bei Computerrealisierung wird dies praktisch nicht auftreten. Wenn dabei doch $\boldsymbol{A} - \varrho_k \boldsymbol{I}$ numerisch singulär sein sollte, also kein von 0 verschiedenes Pivot bei der Dreiecksfaktorisierung gefunden werden kann, ist wie in 13.4.5 (i) zu verfahren.

(ii) Bei Computerrechnung wird das durch 13.4.2 beschriebene ideale Konvergenzverhalten durch die Rundungsfehler gestört, wegen 13.4.4 jedoch nicht verschlechtert, sondern eher verbessert. Es ist also auch bei Computerrealisierung praktisch immer mit schneller Konvergenz zu rechnen.

(iii) Im allgemeinen kann nichts darüber ausgesagt werden, gegen welchen Eigenwert λ_j die Rayleigh-Quotienten ϱ_k konvergieren. Insbesondere braucht die Konvergenz nicht gegen denjenigen Eigenwert zu erfolgen, der minimalen Abstand von ϱ_1 hat. □

Übungsaufgabe

Ü 13.4.1. Man zeige: Wenn bei der RQ-Iteration $\|\boldsymbol{r}^{k+1}\| = \|\boldsymbol{r}^k\|$ ist, gilt $\varrho_k = \varrho_{k+1}$, und $\boldsymbol{v}^k$ ist ein Eigenvektor von $(\boldsymbol{A} - \varrho_k \boldsymbol{I})^2$.

13.5. Orthogonale Ähnlichkeitstransformation auf Tridiagonalform

Die im vorherigen Abschnitt behandelten Varianten der inversen Iteration wurden direkt mit der Matrix $\boldsymbol{A} \in \mathbf{S}^{n,n}$ des zu lösenden Eigenwertproblems $\boldsymbol{A}\boldsymbol{x} = \lambda \boldsymbol{x}$ durchgeführt. Eine alternative Vorgehensweise besteht darin, das an anderen Stellen bereits praktizierte Prinzip der Transformation des Ausgangsproblems in ein äquivalentes Problem einfacherer Gestalt auch für die Lösung des Eigenwertproblems auszunützen. Als Transformation kommt dabei eine orthogonale Ähnlichkeitstransformation in Frage, bei der Symmetrie und Eigenwerte erhalten bleiben, vgl. 13.1.A.

Da eine Ähnlichkeitstransformation auf Diagonalform in endlich vielen Schritten i. allg. nicht möglich ist — die Diagonalelemente wären dann bereits die gesuchten

Eigenwerte —, bietet sich als Ziel eine Tridiagonalmatrix

$$\boldsymbol{T} = \begin{pmatrix} a_1 & b_2 & & \\ b_2 & a_2 & b_3 & \\ \ddots & \ddots & \ddots & \\ & b_{n-1} & a_{n-1} & b_n \\ & & b_n & a_n \end{pmatrix} =: \operatorname{trid}(a_1, \ldots, a_n, b_2, \ldots, b_n)$$

an. Man beachte, daß $\boldsymbol{A}$ durch $\sim n^2/2$ Daten $\{a_{ij}: i \leqq j\}$ repräsentiert wird, während $\boldsymbol{T}$ durch nur $\sim 2n$ Daten $\{a_1, \ldots, a_n, b_2, \ldots, b_n\}$ festgelegt ist. Beim Übergang von $\boldsymbol{A}$ zu $\boldsymbol{T}$ tritt daher eine wesentliche Datenreduktion ein. Durch die Transformation

$$\boldsymbol{A} \to \boldsymbol{T} = \boldsymbol{Q}^{\mathsf{T}}\boldsymbol{A}\boldsymbol{Q}, \qquad \boldsymbol{Q} \in \mathbf{R}^{n,n} \text{ orthogonal}, \tag{1}$$

wird also das Eigenwertproblem einer allgemeinen symmetrischen Matrix $\boldsymbol{A}$ auf das einer symmetrischen Tridiagonalmatrix $\boldsymbol{T}$ zurückgeführt. Das letztere läßt sich jedoch wesentlich effektiver lösen, wie wir in den Abschnitten 13.6 und 13.7 sehen werden.

A. *Transformation mittels Householder-Spiegelungen*

Im folgenden wird gezeigt, daß die Transformation (1) nach der Vorschrift

$$\boldsymbol{A}^{(1)} := \boldsymbol{A}, \quad \boldsymbol{A}^{(k+1)} := \boldsymbol{H}_k \boldsymbol{A}^{(k)} \boldsymbol{H}_k \quad (k = 1, \ldots, n-2), \quad \boldsymbol{T} := \boldsymbol{A}^{(n-1)} \tag{2}$$

unter Verwendung von Householder-Spiegelungen bzw. Einheitsmatrizen $\boldsymbol{H}_k$ in $n-2$ Schritten durchgeführt werden kann, wobei im k-ten Schritt die Nullen in der k-ten Spalte und Zeile erzeugt werden. Es sei also $\boldsymbol{A}^{(k)}$ bereits von der Form

$$\boldsymbol{A}^{(k)} = \left(\begin{array}{c|c} \boldsymbol{T}^{(k)} & \boldsymbol{a}^{k\mathsf{T}} \\ \hline \boldsymbol{a}^k & \boldsymbol{B}^{(k)} \end{array}\right) \tag{3}$$

mit

$$\boldsymbol{T}^{(k)} = \operatorname{trid}(a_1, \ldots, a_k, b_2, \ldots, b_k) \in \mathbf{S}^{k,k}, \qquad \boldsymbol{a}^k \in \mathbf{R}^{n-k}, \qquad \boldsymbol{B}^{(k)} \in \mathbf{S}^{n-k,n-k}.$$

Im Fall $\boldsymbol{a}^k = \boldsymbol{o}$ sind die geforderten Nullen bereits vorhanden, so daß

$$\boldsymbol{H}_k := \boldsymbol{I} \tag{4}$$

gewählt werden kann. Andernfalls setzen wir $\boldsymbol{H}_k$ als Householder-Spiegelung

$$\boldsymbol{H}_k := \left(\begin{array}{c|c} \boldsymbol{I}_k & \\ \hline & \overline{\boldsymbol{H}}_k \end{array}\right) \tag{5}$$

an und bestimmen

$$\overline{\boldsymbol{H}}_k := \boldsymbol{I} - \overline{\boldsymbol{v}}^k \overline{\boldsymbol{v}}^{k\mathsf{T}} / \overline{\gamma}_k \in \mathbf{S}^{n-k,n-k}, \qquad \overline{\gamma}_k := \overline{\boldsymbol{v}}^{k\mathsf{T}} \overline{\boldsymbol{v}}^k / 2 \tag{6}$$

so, daß

$$\boldsymbol{a}^k = \begin{pmatrix} a_{k+1,k}^{(k)} \\ a_{k+2,k}^{(k)} \\ \vdots \\ a_{nk}^{(k)} \end{pmatrix} \quad \text{in} \quad \bar{\boldsymbol{a}}^k := \bar{\boldsymbol{H}}_k \boldsymbol{a}^k = \varrho_k \boldsymbol{e}^1 = \begin{pmatrix} \varrho_k \\ 0 \\ \vdots \\ 0 \end{pmatrix}$$

übergeht. Dies ist bei der Festlegung

$$\bar{\boldsymbol{v}}^k = (\boldsymbol{v}_{k+1,k}, \ldots, \boldsymbol{v}_{nk})^{\mathsf{T}} := \boldsymbol{e}^1 - \boldsymbol{a}^k/\varrho_k \tag{7}$$

mit

$$\varrho_k := \begin{cases} \|\boldsymbol{a}^k\| & \text{für} \quad a_{k+1,k}^{(k)} \leqq 0 \\ -\|\boldsymbol{a}^k\| & \text{für} \quad a_{k+1,k}^{(k)} > 0, \end{cases} \tag{8}$$

der Fall; es gilt dann $\bar{\gamma}_k = v_{k+1,k}$, vgl. 3.3.A und Abschnitt 10.2. Für $\boldsymbol{A}^{(k+1)}$ ergibt sich damit

$$\boldsymbol{A}^{(k+1)} = \boldsymbol{H}_k \boldsymbol{A}^{(k)} \boldsymbol{H}_k = \left(\begin{array}{c|c} \boldsymbol{T}^{(k)} & \bar{\boldsymbol{a}}^{k\mathsf{T}} \\ \hline \bar{\boldsymbol{a}}^k & \bar{\boldsymbol{B}}^{(k)} \end{array}\right) = \left(\begin{array}{c|c} \boldsymbol{T}^{(k+1)} & \boldsymbol{a}^{k+1\mathsf{T}} \\ \hline \boldsymbol{a}^{k+1} & \boldsymbol{B}^{(k+1)} \end{array}\right)$$

mit

$$\bar{\boldsymbol{B}}^{(k)} := \bar{\boldsymbol{H}}_k \boldsymbol{B}^{(k)} \bar{\boldsymbol{H}}_k. \tag{9}$$

Im Unterschied zur Householder-Faktorisierung aus 10.2 tritt die Spiegelung $\bar{\boldsymbol{H}}_k$ in (9) auf beiden Seiten von $\boldsymbol{B}^{(k)}$ auf, und $\bar{\boldsymbol{B}}^{(k)}$ ist wie $\boldsymbol{B}^{(k)}$ symmetrisch. Es ist daher nicht zweckmäßig, die transformierte Matrix $\bar{\boldsymbol{B}}$ gemäß $\bar{\boldsymbol{B}} := (\bar{\boldsymbol{H}}\boldsymbol{B})\,\bar{\boldsymbol{H}}$ oder $\bar{\boldsymbol{B}} := \bar{\boldsymbol{H}}(\boldsymbol{B}\bar{\boldsymbol{H}})$ zu berechnen, denn $\bar{\boldsymbol{H}}\boldsymbol{B}$ und $\boldsymbol{B}\bar{\boldsymbol{H}}$ sind i. allg. nicht symmetrisch; zur Vereinfachung ignorieren wir den Index k. Die Symmetrie wird dagegen voll ausgenutzt, wenn (9) unter Beachtung von (6) in der Form

$$\begin{aligned} \bar{\boldsymbol{B}} = \bar{\boldsymbol{H}}\boldsymbol{B}\bar{\boldsymbol{H}} &= \boldsymbol{B} - \bar{\boldsymbol{v}}\bar{\boldsymbol{v}}^{\mathsf{T}}\boldsymbol{B}/\bar{\gamma} - \boldsymbol{B}\bar{\boldsymbol{v}}\bar{\boldsymbol{v}}^{\mathsf{T}}/\bar{\gamma} + \bar{\boldsymbol{v}}\bar{\boldsymbol{v}}^{\mathsf{T}}\boldsymbol{B}\bar{\boldsymbol{v}}\bar{\boldsymbol{v}}^{\mathsf{T}}/\bar{\gamma}^2 \\ &= \boldsymbol{B} - \bar{\boldsymbol{v}}\bar{\boldsymbol{w}}^{\mathsf{T}} - \bar{\boldsymbol{w}}\bar{\boldsymbol{v}}^{\mathsf{T}} + [\bar{\boldsymbol{v}}^{\mathsf{T}}\bar{\boldsymbol{w}}/\bar{\gamma}]\,\bar{\boldsymbol{v}}\bar{\boldsymbol{v}}^{\mathsf{T}} \end{aligned} \tag{10}$$

mit

$$\bar{\boldsymbol{w}} := \boldsymbol{B}\bar{\boldsymbol{v}}/\bar{\gamma} \tag{11}$$

geschrieben wird. Die gesonderte Berechnung des dritten dyadischen Produktes in (10) kann schließlich auch noch vermieden werden, indem dieses zur Hälfte den beiden anderen Produkten zugeschlagen wird, was auf

$$\bar{\boldsymbol{B}} := \boldsymbol{B} - (\bar{\boldsymbol{p}}\bar{\boldsymbol{v}}^{\mathsf{T}} + \bar{\boldsymbol{v}}\bar{\boldsymbol{p}}^{\mathsf{T}}) \tag{12}$$

mit

$$\bar{\boldsymbol{p}} := \bar{\boldsymbol{w}} - [\bar{\boldsymbol{v}}^{\mathsf{T}}\bar{\boldsymbol{w}}/(2\bar{\gamma})]\,\bar{\boldsymbol{v}} \tag{13}$$

führt.

Nach $n-2$ derart ausgeführten Schritten ergibt sich die gesuchte Tridiagonalmatrix

$$\boldsymbol{A}^{(n-1)} = \left(\begin{array}{c|c} \boldsymbol{T}^{(n-1)} & \begin{matrix}\\ \boldsymbol{a}^{n-1}\end{matrix} \\ \hline \begin{matrix}\\ \boldsymbol{a}^{n-1}\end{matrix} & \boldsymbol{B}^{(n-1)} \end{array}\right) =: \boldsymbol{T} = \operatorname{trid}(a_1, \ldots, a_n, b_2, \ldots, b_n)$$

mit $\boldsymbol{a}^{n-1} = \boldsymbol{a}^{n-1\mathsf{T}} =: b_n \in \mathbf{R}$ und $\boldsymbol{B}^{(n-1)} =: a_n \in \mathbf{R}$, und es gilt

$$\boldsymbol{T} = \boldsymbol{H}_{n-2} \cdots \boldsymbol{H}_2 \boldsymbol{H}_1 \boldsymbol{A} \boldsymbol{H}_1 \boldsymbol{H}_2 \cdots \boldsymbol{H}_{n-2},$$

also (1) mit der durch

$$\boldsymbol{Q} := \boldsymbol{H}_1 \boldsymbol{H}_2 \cdots \boldsymbol{H}_{n-2} \tag{14}$$

definierten orthogonalen Matrix $\boldsymbol{Q}$. Zusammenfassend erhalten wir den folgenden Algorithmus.

13.5.1. *Tridiagonalisierung mittels Householder-Spiegelungen.* Für gegebenes $\boldsymbol{A} \in \mathbf{S}^{n,n}$ werde die Transformation (2) mit $\boldsymbol{H}_k$ gemäß (4) oder (5) bis (8) in exakter Arithmetik durchgeführt. Dann gilt

$$\boldsymbol{T} = \boldsymbol{Q}^{\mathsf{T}} \boldsymbol{A} \boldsymbol{Q}$$

mit der Tridiagonalmatrix $\boldsymbol{T} := \boldsymbol{A}^{(n-1)}$ und der Produktdarstellung (14) für die orthogonale Transformationsmatrix $\boldsymbol{Q}$.

Aufwand: $\sim 2n^3/3$ opms $+ \sim n$ opr für $\boldsymbol{T}$, falls $\boldsymbol{A}$ voll besetzt ist und $\overline{\boldsymbol{B}}^{(k)}$ gemäß (11) bis (13) aus $\boldsymbol{B}^{(k)}$ berechnet wird.

13.5.2. *Bemerkung.* (i) Algorithmus 13.5.1 kann auf dem Platz des unteren Dreiecks von $\boldsymbol{A}$ ausgeführt werden. Falls $\boldsymbol{H}_k$ als Householder-Spiegelung gewählt wird, werden die Elemente a_{ik} mit den Komponenten $\{v_{ik}: i \geqq k+1\}$ des Vektors $\overline{\boldsymbol{v}}^k$ überspeichert. Es gilt dann $v_{k+1,k} = \overline{\gamma}_k \geqq 1$, vgl. 3.3.5. Falls $\boldsymbol{H}_k = \boldsymbol{I}$ gesetzt wurde, ordnen wir $\boldsymbol{H}_k$ den Vektor $\overline{\boldsymbol{v}}^k = \boldsymbol{o}$ zu. Da dies nur für $\boldsymbol{a}^k = \boldsymbol{o}$ eintreten kann, sind die entsprechenden Nullen bereits vorhanden. Insbesondere ist dann $v_{k+1,k} = 0$, so daß beide Fälle zuverlässig rekonstruiert werden können. Die Subdiagonalelemente $b_{k+1} = \varrho_k$ von $\boldsymbol{T}$ müssen dann in einem gesonderten Feld der Länge n aufgehoben werden. Der Aufwand im k-ten Teilschritt wird für voll besetztes $\boldsymbol{A}$ durch die Berechnung von $\overline{\boldsymbol{w}}$ gemäß (11) und die Bildung von $\overline{\boldsymbol{B}}$ gemäß (12) mit zusammen $\sim 2(n-k)^2$ opms bestimmt; man beachte, daß nur ein Dreieck von $\overline{\boldsymbol{B}}$ berechnet zu werden braucht.

(ii) Die Berechnung von $\boldsymbol{Q}^{\mathsf{T}}\boldsymbol{b}$ bzw. $\boldsymbol{Q}\boldsymbol{f}$ aus der durch die $\{v_{ik}: i \geqq k+1\}$ repräsentierten Produktform (14) ist analog zu 10.2.6 mit jeweils $\sim n^2$ opms möglich; dabei sind die Fallunterscheidungen aus (i) zu berücksichtigen. Falls $\boldsymbol{Q}$ explizit benötigt wird, kann die Berechnung mit zusätzlichen $\sim 2n^3/3$ opms analog zu Ü 10.2.2 erfolgen. □

Wir geben ohne Beweis die Ergebnisse der Rundungsfehleranalyse an.

13.5.3. *Rundungsfehleranalyse.* Algorithmus 13.5.1 werde für symmetrisches $A \in \mathfrak{R}^{n,n}$ durchgeführt, und $T \in \mathfrak{R}^{n,n}$ sei die berechnete symmetrische Tridiagonalmatrix. Dann gilt

$$A + \delta A = QTQ^T \tag{15}$$

mit exakt orthogonalem $Q \in \mathbf{R}^{n,n}$ und einer Störung $\delta A \in \mathbf{S}^{n,n}$, für die

$$\|\delta A\|_F \leqq \nu F_0 \|A\|_F \quad \text{mit} \quad F_0 := 4.2n^2 \tag{16}$$

und

$$\|\delta A\| \leqq \nu F_1 \|A\| \quad \text{mit} \quad F_1 := 0.4n^{5/2}(1 + 9.5/n^{1/2})$$

gilt.

13.5.4. *Bemerkung.* (i) Die Matrix Q in (15) ist das exakte Produkt (14) der Matrizen H_k, die den berechneten $A^{(k)}$ in exakter Arithmetik zugeordnet sind. Die exakte Matrix Q wird sowohl durch die mittels der berechneten $\{v_{ik}\}$ definierte Produktform (14) als auch durch die explizit analog zu Ü 10.2.2 berechnete Matrix $Q = \mathrm{fl}(H_1 H_2 \cdots H_{n-2})$ gut repräsentiert, vgl. 10.2 und 11.1 für analoge Überlegungen.

(ii) Die Matrix T und die Spiegelungen H_k und damit deren Produkt Q sind stetige Funktionen von A, sofern der Fall $H_k = I$ ausgeschlossen wird, also $\|a^k\| = |\varrho_k| = |b_{k+1}| \neq 0$ $(k = 1, \ldots, n-2)$ gilt. Trotzdem sind die berechneten Faktoren T und Q oft sehr empfindlich gegenüber Störungen von A und Änderungen der Computerarithmetik, siehe B 13.7 für Hinweise auf eindrucksvolle Beispiele in der Spezialliteratur. Die möglicherweise stark differierenden berechneten Matrizen T sind jedoch alle exakt orthogonal ähnlich zu $A + \delta A$ mit kleinem δA, d. h., es liegt numerische Gutartigkeit vor, und die Eigenwerte von T approximieren die von A gut.

(iii) Ist A eine im Sinne von 13.1.B absteigend gestufte Matrix, also etwa

$$A = \begin{pmatrix} 100 & 10 & 1 \\ 10 & 10 & 1 \\ 1 & 1 & 1 \end{pmatrix},$$

so sind T und δA bei Rechnung nach 13.5.1 meistens in derselben Weise gestuft, so daß auch die beitragskleinen Eigenwerte von A durch T ausreichend repräsentiert werden. Ist A jedoch in inverser Weise gestuft, also etwa

$$A = \begin{pmatrix} 1 & 1 & 1 \\ 1 & 10 & 10 \\ 1 & 10 & 100 \end{pmatrix},$$

so gehen beim Übergang zu T mittels 13.5.1 Informationen über die betragskleinen Eigenwerte verloren. In diesem Fall sollte mit

$$\bar{A} := PAP, \qquad P = P^T := \begin{pmatrix} 0 & 0 & 1 \\ 0 & 1 & 0 \\ 1 & 0 & 0 \end{pmatrix}, \quad \text{allgemein} \quad P = \sum_{j=1}^{n} e^{n-j+1} e^{jT}$$

gearbeitet werden, oder 13.5.1 sollte in inverser Reihenfolge realisiert werden, d. h., im ersten Schritt sollten die Nullen in der n-ten Spalte und Zeile erzeugt werden, siehe Ü 13.5.2. □

B. *Tridiagonalisierung mittels Givens-Drehungen und Anwendung auf Bandmatrizen*

In Analogie zum Vorgehen bei der $\boldsymbol{QR}$-Faktorisierung in 10.2 kann die Householder-Spiegelung $\boldsymbol{H}_k$ durch ein Produkt von $n-k-1$ Givens-Drehungen ersetzt werden: Die Vorschrift $\boldsymbol{A}^{(k+1)} := \boldsymbol{H}_k \boldsymbol{A}^{(k)} \boldsymbol{H}_k$ wird ersetzt durch

$$\boldsymbol{A}^{(k+1)} := \boldsymbol{G}_k^{\mathsf{T}} \boldsymbol{A}^{(k)} \boldsymbol{G}_k := [\boldsymbol{G}_{k+1,k+2} \cdots (\boldsymbol{G}_{n-1,n} \boldsymbol{A}^{(k)} \boldsymbol{G}_{n-1,n}^{\mathsf{T}}) \cdots \boldsymbol{G}_{k+1,k+2}^{\mathsf{T}}]; \tag{17}$$

selbstverständlich sind auch hier andere Drehungsindizes möglich, siehe Ü 13.5.3. Die Drehung $\boldsymbol{G}_{i-1,i}$ wird dabei so gewählt, daß $\boldsymbol{G}_{i-1,i}$ $(k+2 \leqq i \leqq n)$ bei Anwendung auf die aktuelle Transformierte in der Position $\{i, k\}$ die gewünschte Null erzeugt.

13.5.5. *Bemerkung.* (i) Bei voll besetztem $\boldsymbol{A}$ und expliziter Realisierung der Givens-Drehungen ist der Aufwand $\sim n^3$ opms $+ \sim n^2/2$ opr, also etwa um die Hälfte höher als in 13.5.1, wenn die Bildung von $\bar{\boldsymbol{A}} := \boldsymbol{G}_{i-1,i} \boldsymbol{A} \boldsymbol{G}_{i-1,i}^{\mathsf{T}}$ für das jeweils aktuelle $\boldsymbol{A}$ analog zu Ü 3.3.6 erfolgt. Man beachte, daß Drehungsindizes und Drehungsparameter $\{c, s\}$ beim Jacobi-Verfahren jedoch anders festgelegt sind als hier. Diese auf GIVENS zurückgehende, historisch älteste Version der Tridiagonalisierung — siehe B 13.7 — scheidet daher aus Aufwandsgründen aus. Bei impliziter Realisierung sind $\sim 2n^3/3$ opms wie bei 13.5.1 erforderlich, siehe wieder B 13.7 für weiterführende Hinweise zur Implementierung.

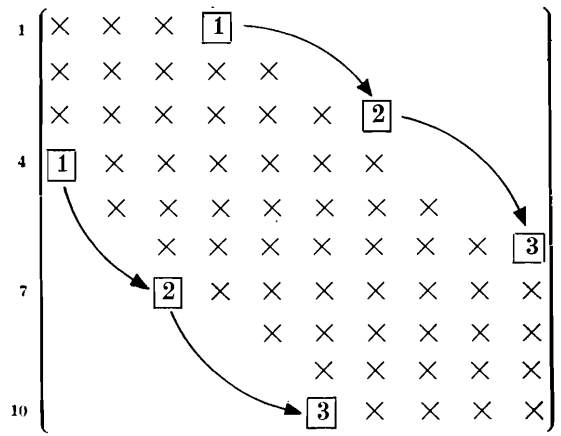
(ii) Falls $\boldsymbol{A}$ schwach besetzt ist, zeigen Givens-Drehungen i. allg. deutliche Vorteile gegenüber Householder-Spiegelungen, da nur die Nichtnullelemente (= NNE) annulliert zu werden brauchen, was mittels Givens-Drehungen effektiver möglich ist. Wir demonstrieren dies im folgenden für den wichtigen Sonderfall einer Bandmatrix. Man beachte, daß die Bandgestalt durch 13.5.1 sukzessive zerstört wird. □

Es sei also $\boldsymbol{A} \in \mathbf{S}^{n,n}$ eine Bandmatrix des Typs $\{s, s\}$ mit der Bandbreite $l = 2s+1$, vgl. 6.4.A, d. h., $\boldsymbol{A}$ habe die Gestalt

$$\boldsymbol{A} = \begin{pmatrix} \ddots & \cdots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \cdots & \ddots \end{pmatrix} \Big\}{\scriptstyle s}. \tag{18}$$

Wir erläutern das Vorgehen am Beispiel einer Matrix mit $n = 10$, $s = 3$, vgl. das Schema auf S. 403, und betrachten dazu nur den ersten Hauptschritt $k = 1$ bei insitu-Realisierung. Im ersten Schritt wird $\boldsymbol{G}_{34}$ so festgelegt, daß $\boldsymbol{A} := \boldsymbol{G}_{34} \boldsymbol{A} \boldsymbol{G}_{34}^{\mathsf{T}}$ in der Position $\{4, 1\}$ (und der jeweils symmetrischen Position, wir erwähnen diese nicht gesondert) eine 0 hat. Dabei wird in der Position $\{7, 3\}$ außerhalb des Bandes das mit $\boxed{2}$ bezeichnete NNE erzeugt. Um die Bandgestalt wieder herzustellen, wird dieses

Element durch eine zweite Drehung $\boldsymbol{G}_{67}$ nach der Vorschrift $\boldsymbol{A} := \boldsymbol{G}_{67}\boldsymbol{A}\boldsymbol{G}_{67}^{\mathsf{T}}$ annulliert, wobei allerdings in $\{10, 6\}$ ein zweites NNE entsteht, nämlich $\boxed{3}$. Dieses wird durch eine weitere Zusatzdrehung $\boldsymbol{G}_{9,10}$ annulliert, womit die Bandgestalt wieder vorliegt. Im nächsten Teilschritt wird in $\{3, 1\}$ mittels $\boldsymbol{G}_{23}$ eine 0 erzeugt, und die in $\{6, 2\}$ bzw. $\{9, 5\}$ entstehenden NNE außerhalb des Bandes werden mittels $\boldsymbol{G}_{56}$ bzw. $\boldsymbol{G}_{89}$ annulliert usw.



Im allgemeinen Fall lautet eine Grobbeschreibung des Verfahrens wie folgt:

13.5.6. *Tridiagonalisierung einer Bandmatrix mittels Givens-Drehungen.* Für die Bandmatrix $\boldsymbol{A} \in \mathbf{S}^{n,n}$ vom Typ $\{s, s\}$ werde der folgende Algorithmus in exakter Arithmetik ausgeführt:

for $k := 1(1)\, n - 2$ **do**
  k-ter Hauptschritt:
  for $i := \min\{k + s, n\}\ (-1)\ k + 2$ **do**
    S1 (Annullierung von a_{ik}):
      Lege $\boldsymbol{G}_{i-1,i}$ so fest, daß $(\boldsymbol{G}_{i-1,i}\boldsymbol{A})_{ik} = 0$
      Bilde $\boldsymbol{A} := \boldsymbol{G}_{i-1,i}\boldsymbol{A}\boldsymbol{G}_{i-1,i}^{\mathsf{T}}$
    S2 (Wiederherstellung der Bandgestalt):
      $p := i$
      for $p := p + s$ **while** $p \leqq n$ **do**
        S3 (Annullierung von $a_{p,p-s-1}$):
          Lege $\boldsymbol{G}_{p-1,p}$ so fest, daß $(\boldsymbol{G}_{p-1,p}\boldsymbol{A})_{p,p-s-1} = 0$
          Bilde $\boldsymbol{A} := \boldsymbol{G}_{p-1,p}\boldsymbol{A}\boldsymbol{G}_{p-1,p}^{\mathsf{T}}$

Dann ist $\boldsymbol{A}$ mit der Tridiagonalmatrix $\boldsymbol{T} = \mathrm{trid}\,(a_1, \ldots, a_n, b_2, \ldots, b_n)$ überspeichert, und $\boldsymbol{T}$ ist orthogonal ähnlich zu $\boldsymbol{A}$.

13.5.7. *Bemerkung.* (i) Bei expliziter Realisierung der Drehungen analog zu Ü 3.3.6 ist der Aufwand $\sim 3(s-1)\, n^2$ opms $+ \sim [(s-1)/(2s)]\, n^2$ opr, wenn nur $\boldsymbol{T}$ berechnet

wird. Im Fall $s \leqq n/5$ ist 13.5.6 also billiger als 13.5.1. Mittels impliziter Givens-Drehungen läßt sich der Aufwand etwa um ein Drittel reduzieren.

(ii) Da n bei einer Bandmatrix meist groß ist, werden nur wenige Eigenwerte zu berechnen sein. Diese können aus $\boldsymbol{T}$ nach dem Bisektionsverfahren 13.6.8 bestimmt werden. Zugehörige Eigenvektoren können dann mittels inverser Iteration nach 13.4.3 ermittelt werden. Da das Speichern der Transformationsmatrix $\boldsymbol{Q}$ aufwendig ist, sollte die inverse Iteration direkt mit der originalen Bandmatrix $\boldsymbol{A}$ unter Verwendung einer nach 6.4.A berechneten spaltenpivotisierten Gauß-Faktorisierung erfolgen, vgl. auch 13.6.11(vi). □

C. *Eindeutigkeit der Faktoren $\boldsymbol{T}$ und $\boldsymbol{Q}$*

Für gewisse Realisierungen des $\boldsymbol{QR}$-Algorithmus aus 13.7 ist die nachfolgende Eindeutigkeitsaussage von Bedeutung.

13.5.8. *Satz.* Für $\boldsymbol{A} \in \mathbf{S}^{n,n}$ gelte

$$\boldsymbol{A} = \boldsymbol{Q}\boldsymbol{T}\boldsymbol{Q}^{\mathsf{T}} = \bar{\boldsymbol{Q}}\bar{\boldsymbol{T}}\bar{\boldsymbol{Q}}^{\mathsf{T}} \tag{19}$$

mit Tridiagonalmatrizen $\boldsymbol{T} = \operatorname{trid}(a_i, b_i)$, $\bar{\boldsymbol{T}} = \operatorname{trid}(\bar{a}_i, \bar{b}_i)$ und orthogonalen Matrizen $\boldsymbol{Q} = (\boldsymbol{q}^1, \ldots, \boldsymbol{q}^n)$, $\bar{\boldsymbol{Q}} = (\bar{\boldsymbol{q}}^1, \ldots, \bar{\boldsymbol{q}}^n)$. Wenn die Bedingung

$$b_i \neq 0 \qquad (i = 2, \ldots, n) \tag{20}$$

erfüllt ist und

$$\boldsymbol{q}^1 = \boldsymbol{Q}\boldsymbol{e}^1 = \bar{\boldsymbol{Q}}\boldsymbol{e}^1 = \bar{\boldsymbol{q}}^1 \tag{21}$$

gilt, ist

$$\bar{\boldsymbol{Q}} = \boldsymbol{Q}\boldsymbol{D} \quad \text{und} \quad \bar{\boldsymbol{T}} = \boldsymbol{D}\boldsymbol{T}\boldsymbol{D} \tag{22}$$

mit

$$\boldsymbol{D} = \operatorname{diag}(d_i), \qquad d_1 = 1, \qquad |d_i| = 1 \qquad (i = 2, \ldots, n). \tag{23}$$

Beweis. Aus (19) folgt

$$\bar{\boldsymbol{T}}\boldsymbol{P} = \boldsymbol{P}\boldsymbol{T} \quad \text{mit orthogonalem } \boldsymbol{P} = (\boldsymbol{p}^1, \ldots, \boldsymbol{p}^n) := \bar{\boldsymbol{Q}}^{\mathsf{T}}\boldsymbol{Q}, \tag{24}$$

und wegen (21) ist $\boldsymbol{e}^1 = \bar{\boldsymbol{Q}}^{\mathsf{T}}\boldsymbol{Q}\boldsymbol{e}^1 = \boldsymbol{P}\boldsymbol{e}^1 = \boldsymbol{p}^1$. Die erste Spalte der Identität (24) lautet daher

$$\bar{\boldsymbol{T}}\boldsymbol{P}\boldsymbol{e}^1 = \bar{\boldsymbol{T}}\boldsymbol{p}^1 = \bar{\boldsymbol{T}}\boldsymbol{e}^1 = \bar{a}_1\boldsymbol{e}^1 + \bar{b}_2\boldsymbol{e}^2 = \boldsymbol{P}\boldsymbol{T}\boldsymbol{e}^1 = a_1\boldsymbol{p}^1 + b_2\boldsymbol{p}^2 = a_1\boldsymbol{e}^1 + b_2\boldsymbol{p}^2. \tag{25}$$

Linksmultiplikation mit $\boldsymbol{e}^{1\mathsf{T}}$ führt auf $\bar{a}_1 = a_1$, also $\bar{b}_2\boldsymbol{e}^2 = b_2\boldsymbol{p}^2$ und damit $|\bar{b}_2| = |b_2|$ sowie $\boldsymbol{p}^2 = d_2\boldsymbol{e}^2$ mit $d_2 := \bar{b}_2/b_2$, $|d_2| = 1$; man beachte (20). Für die zweite Spalte folgt in analoger Weise

$$\begin{aligned}\bar{\boldsymbol{T}}\boldsymbol{P}\boldsymbol{e}^2 = \bar{\boldsymbol{T}}\boldsymbol{p}^2 = d_2\bar{\boldsymbol{T}}\boldsymbol{e}^2 &= d_2(\bar{b}_2\boldsymbol{e}^1 + \bar{a}_2\boldsymbol{e}^2 + \bar{b}_3\boldsymbol{e}^3)\\ &= \boldsymbol{P}\boldsymbol{T}\boldsymbol{e}^2 = b_2\boldsymbol{p}^1 + a_2\boldsymbol{p}^2 + b_3\boldsymbol{p}^3 = b_2\boldsymbol{e}^1 + d_2a_2\boldsymbol{e}^2 + b_3\boldsymbol{p}^3.\end{aligned}$$

Dies liefert $b_3\boldsymbol{p}^3 = (d_2\bar{b}_2 - b_2)\,\boldsymbol{e}^1 + d_2(\bar{a}_2 - a_2)\,\boldsymbol{e}^2 + d_2\bar{b}_3\boldsymbol{e}^3$. Die erste Klammer verschwindet wegen der Festlegung von d_2. Multiplikation mit $\boldsymbol{e}^{2\mathsf{T}}$ liefert $\bar{a}_2 = a_2$, mithin $\boldsymbol{p}^3 = d_3\boldsymbol{e}^3$, $d_3 := d_2\bar{b}_3/b_3$, $|d_3| = 1$. Durch Induktion folgt schließlich $\boldsymbol{P} = \boldsymbol{D}$ mit $\boldsymbol{D}$ aus (23), womit (24) in (22) übergeht. □

13.5.9. *Bemerkung.* (i) Spalten- bzw. elementweise liest sich (22) als

$$\bar{\boldsymbol{q}}^j = d_j \boldsymbol{q}^j, \qquad \bar{a}_j = a_j \ (j = 1, \ldots, n), \qquad \bar{b}_j = d_{j-1} d_j b_j \ (j = 2, \ldots, n) \tag{26}$$

und bedeutet: *Die Faktoren $\boldsymbol{Q}$ und $\boldsymbol{T}$ der Faktorisierung $\boldsymbol{A} = \boldsymbol{Q}\boldsymbol{T}\boldsymbol{Q}^\mathsf{T}$ sind unter der Voraussetzung* (20) *durch die erste Spalte $\boldsymbol{q}^1$ von $\boldsymbol{Q}$ im wesentlichen festgelegt*, nämlich bis auf das Vorzeichen der Spalten von $\boldsymbol{Q}$ und der Nebendiagonalelemente $\{b_i\}$ von $\boldsymbol{T}$. Eine vergleichbare Aussage über die Eindeutigkeit der $\boldsymbol{QR}$-Faktorisierung einer spaltenregulären Matrix haben wir in 10.1.3 kennengelernt. Hier sind allerdings mehr Freiheitsgrade vorhanden, so daß $\boldsymbol{q}^1$ vorgegeben werden kann. Daß dies in der Tat bis auf gewisse Ausnahmesituationen möglich ist, zeigt der Lanczos-Algorithmus aus Ü 13.5.6.

(ii) Die Bedingung (21) kann durch $\boldsymbol{q}^n = \bar{\boldsymbol{q}}^n$ ersetzt werden, d. h., statt $\boldsymbol{q}^1$ kann $\boldsymbol{q}^n$ vorgegeben werden. Der Beweis ist dann in umgekehrter Richtung — beginnend mit Spalte n von (24) — zu führen. □

Übungsaufgaben

Ü 13.5.1. Man schreibe ein Programm zur Berechnung von $\boldsymbol{w} := \boldsymbol{B}\boldsymbol{v}$, wenn $\boldsymbol{B} \in \mathbf{S}^{n,n}$ durch die Elemente des unteren Dreiecks $\{b_{ij}: i \geqq j\}$ repräsentiert ist.

Ü 13.5.2. Man modifiziere 13.5.1 so, daß im ersten Schritt $\boldsymbol{A}^{(2)} := \boldsymbol{H}_1 \boldsymbol{A}^{(1)} \boldsymbol{H}_1$ Nullen in den Positionen $\{1, n\}, \{2, n\}, \ldots, \{n-2, n\}$ und $\{n, 1\}, \{n, 2\}, \ldots, \{n, n-2\}$ erzeugt werden usw. Welche Gestalt müssen dann $\boldsymbol{H}_1$ und $\bar{\boldsymbol{v}}^1$ aus $\bar{\boldsymbol{H}}_1 = \boldsymbol{I} - \bar{\boldsymbol{v}}^1 \bar{\boldsymbol{v}}^{1\mathsf{T}} / \bar{\gamma}_1$ haben?

Hinweis: Man beachte Ü 3.3.7.

Ü 13.5.3. Man überlege sich, daß die Transformation $\boldsymbol{A}^{(k)} \to \boldsymbol{A}^{(k+1)}$ auch in der Form

$$\boldsymbol{A}^{(k+1)} := \left\{ \boldsymbol{G}_{k+1,k+2} \cdots \left[\boldsymbol{G}_{k+1,n-1} \left(\boldsymbol{G}_{k+1,n} \boldsymbol{A}^{(k)} \boldsymbol{G}_{k+1,n}^\mathsf{T} \right) \boldsymbol{G}_{k+1,n-1}^\mathsf{T} \right] \cdots \boldsymbol{G}_{k+1,k+2}^\mathsf{T} \right\}$$

realisiert werden kann.

Ü 13.5.4. Es sei

$$\boldsymbol{C} = \begin{pmatrix} a_1 & c_1 & & & \\ b_2 & a_2 & c_2 & & \\ & \ddots & \ddots & \ddots & \\ & & \ddots & \ddots & c_{n-1} \\ & & & b_n & a_n \end{pmatrix} \quad \text{mit} \quad c_j * b_{j+1} > 0 \quad (j = 1, \ldots, n-1) \tag{27}$$

eine nicht notwendig symmetrische Tridiagonalmatrix. Man konstruiere eine Diagonalmatrix $\boldsymbol{D} = \operatorname{diag}(d_i)$, $d_i > 0$, so daß $\boldsymbol{T} := \boldsymbol{D}^{-1}\boldsymbol{C}\boldsymbol{D}$ symmetrisch ist. Eine Matrix des Typs (27) heißt *quasisymmetrisch*; bezüglich des Eigenwertproblems kann sie wie eine symmetrische Matrix behandelt werden.

Ü 13.5.5. Man zeige, daß die durch 13.5.1 erzeugte Matrix $\boldsymbol{Q}$ die erste Spalte $\boldsymbol{q}^1 = \boldsymbol{e}^1$ besitzt.

Ü 13.5.6. Man schreibe (1) als $\boldsymbol{A}\boldsymbol{Q} = \boldsymbol{Q}\boldsymbol{T}$, mit $\boldsymbol{Q} = (\boldsymbol{q}^1, \ldots, \boldsymbol{q}^n)$ also spaltenweise als $\boldsymbol{A}\boldsymbol{q}^k = b_k \boldsymbol{q}^{k-1} + a_k \boldsymbol{q}^k + b_{k+1} \boldsymbol{q}^{k+1}$ bzw.

$$b_{k+1} \boldsymbol{q}^{k+1} = \boldsymbol{A}\boldsymbol{q}^k - a_k \boldsymbol{q}^k - b_k \boldsymbol{q}^{k-1} =: \boldsymbol{p}^{k+1} \qquad (k = 1, \ldots, n) \tag{28}$$

mit $b_1 := b_{n+1} := 0$, $\boldsymbol{q}^0 := \boldsymbol{q}^{n+1} := \boldsymbol{o}$.

(i) Man überlege sich: Es seien b_k, $\boldsymbol{q}^{k-1}$ und $\boldsymbol{q}^k$ bekannt. Linksmultiplikation von (28) mit $\boldsymbol{q}^{k\mathsf{T}}$ führt auf $a_k = \boldsymbol{q}^{k\mathsf{T}}\boldsymbol{A}\boldsymbol{q}^k$. Im Fall $\boldsymbol{p}^{k+1} \neq \boldsymbol{o}$ ist dann die Richtung von $\boldsymbol{q}^{k+1}$ durch $\boldsymbol{p}^{k+1}$ festgelegt, andernfalls kann $\boldsymbol{q}^{k+1}$ orthogonal zu span $\{\boldsymbol{q}^1, \ldots, \boldsymbol{q}^k\}$ beliebig mit $\|\boldsymbol{q}^{k+1}\| = 1$ gewählt werden. Dies führt auf die folgende Grundform des sogenannten *Lanczos-Algorithmus*:

$b_1 := 0$, $\boldsymbol{q}^0 := \boldsymbol{o}$, wähle $\boldsymbol{q}^1$ mit $\|\boldsymbol{q}^1\| = 1$
for $k := 1(1)n$ **do**

$a_k := \boldsymbol{q}^{k\mathsf{T}}\boldsymbol{A}\boldsymbol{q}^k$, **if** $k = n$ **then** stop
$\boldsymbol{p}^{k+1} := \boldsymbol{A}\boldsymbol{q}^k - a_k\boldsymbol{q}^k - b_k\boldsymbol{q}^{k-1}$, $b_{k+1} := \pm\|\boldsymbol{p}^{k+1}\|$
if $b_{k+1} \neq 0$ **then** $\boldsymbol{q}^{k+1} := \boldsymbol{p}^{k+1}/b_{k+1}$
else [wähle $\boldsymbol{q}^{k+1} \perp$ span $\{\boldsymbol{q}^1, \ldots, \boldsymbol{q}^k\}$ mit $\|\boldsymbol{q}^{k+1}\| = 1$]

(ii) Man zeige durch Induktion über k, daß

$$\boldsymbol{q}^{i\mathsf{T}}\boldsymbol{q}^{k+1} = \begin{cases} 0 & \text{für } i = 1, \ldots, k, \\ 1 & \text{für } i = k+1 \end{cases} \qquad \boldsymbol{q}^{i\mathsf{T}}\boldsymbol{A}\boldsymbol{q}^{k+1} = \begin{cases} 0 & \text{für } i = 1, \ldots, k-1, \\ b_{k+1} & \text{für } i = k \end{cases}$$

gilt, d. h., in exakter Arithmetik erzeugt der Lanczos-Algorithmus die Faktorisierung $\boldsymbol{A} = \boldsymbol{Q}\boldsymbol{T}\boldsymbol{Q}^{\mathsf{T}}$. In der angegebenen Form ist das Verfahren jedoch instabil; die berechneten Vektoren $\{\boldsymbol{q}^k\}$ können stark von der Orthogonalität abweichen.

13.6. Schneiden des Spektrums und Bisektionsverfahren

In diesem Abschnitt behandeln wir Algorithmen, die besonders zur Berechnung einiger weniger Eigenwerte der symmetrischen Tridiagonalmatrix

$$\boldsymbol{T} = \operatorname{trid}(a_1, \ldots, a_n, b_2, \ldots, b_n)$$

geeignet sind, etwa zur Berechnung der k kleinsten oder größten Eigenwerte oder der in einem vorgegebenen Intervall liegenden Eigenwerte. Wir können dabei

$$b_i \neq 0 \qquad (i = 2, \ldots, n) \tag{1}$$

voraussetzen; die Matrix $\boldsymbol{T}$ heißt dann *nichtzerfallend*. Dies ist keine Einschränkung: Ist etwa $b_{k+1} = 0$ für ein gewisses $k \in \{1, \ldots, n-1\}$, so zerfällt $\boldsymbol{T}$ gemäß

$$\boldsymbol{T} = \left(\begin{array}{ccc|ccc} a_1 & b_2 & & & & \\ b_2 & \ddots & \ddots & & & \\ & \ddots & \ddots & b_k & & \\ & & b_k & a_k & & \\ \hline & & & a_{k+1} & b_{k+2} & \\ & & & b_{k+2} & \ddots & \ddots \\ & & & & \ddots & b_n \\ & & & & b_n & a_n \end{array}\right) =: \left(\begin{array}{c|c} \boldsymbol{T}_1 & \\ \hline & \boldsymbol{T}_2 \end{array}\right)$$

in die beiden Blöcke $\boldsymbol{T}_1$ und $\boldsymbol{T}_2$ der Dimensionen k bzw. $n-k$. Jedem Eigenpaar $\{\lambda_1, \boldsymbol{u}^1\}$ von $\boldsymbol{T}_1$ bzw. $\{\lambda_2, \boldsymbol{u}^2\}$ von $\boldsymbol{T}_2$ entspricht das Eigenpaar $\left\{\lambda_1, \begin{pmatrix}\boldsymbol{u}^1 \\ \boldsymbol{o}\end{pmatrix}\right\}$ bzw. $\left\{\lambda_2, \begin{pmatrix}\boldsymbol{o} \\ \boldsymbol{u}^2\end{pmatrix}\right\}$ von $\boldsymbol{T}$, d. h., die Eigenpaare von $\boldsymbol{T}$ ergeben sich unmittelbar aus denen von $\boldsymbol{T}_1$ und $\boldsymbol{T}_2$.

13.6.1. *Bemerkung.* Das durch verschwindende Nebendiagonalelemente bedingte Zerfallen des Eigenwertproblems von $\boldsymbol{T}$ in das mehrerer Probleme kleinerer Dimension mit nichtzerfallenden Blöcken $\boldsymbol{T}_i$ reduziert i. allg. den Gesamtaufwand. Es ist daher zweckmäßig, auch ein von 0 verschiedenes, aber ausreichend kleines b_{k+1} als 0 anzusehen. Als grobes Kriterium kann

$$|b_{k+1}| \leqq \nu \, \|\boldsymbol{T}\| =: \varepsilon \tag{2}$$

genommen werden. Raffinierter ist das Kriterium

$$|b_{k+1}| \leqq \nu\{|a_k| + |a_{k+1}|\} =: \varepsilon. \tag{3}$$

In beiden Fällen entspricht das Nullsetzen von b_{k+1} einer Störung von $\boldsymbol{A}$ (bzw. $\boldsymbol{T}$) mit $\|\delta \boldsymbol{A}\| \leqq \varepsilon$. □

Eine Folgerung von (1) wird durch die nachstehende Aussage angegeben.

13.6.2. *Aussage.* Die Eigenwerte $\lambda_j = \lambda_j[\boldsymbol{T}]$ einer nichtzerfallenden Tridiagonalmatrix $\boldsymbol{T} \in \mathbf{S}^{n,n}$ sind sämtlich einfach, d. h., in der natürlichen Ordnung gilt

$$\lambda_1 < \lambda_2 < \cdots < \lambda_{n-1} < \lambda_n. \tag{4}$$

Beweis. Für jedes λ sind die ersten $n-1$ Spalten von $\boldsymbol{T} - \lambda \boldsymbol{I}$ wegen (1) linear unabhängig, d. h., es gilt rang $(\boldsymbol{T} - \lambda \boldsymbol{I}) \geqq n - 1$. Im Fall $\lambda = \lambda_j$ ist $\boldsymbol{T} - \lambda_j \boldsymbol{I}$ singulär, so daß dann rang $(\boldsymbol{T} - \lambda_j \boldsymbol{I}) = n - 1$ folgt. Der zu λ_j gehörende Eigenraum $\mathcal{N}(\boldsymbol{T} - \lambda_j \boldsymbol{I})$ ist also eindimensional, d. h., es gibt nicht mehr als einen linear unabhängigen Eigenvektor zu λ_j. □

Aussage 13.6.2 schließt nicht aus, daß auch für deutlich von 0 verschiedene Nebendiagonalelemente gewisse Eigenwerte eng benachbart sein können. Ein viel zitiertes Beispiel dafür stellt die Wilkinson-Matrix

$$\boldsymbol{T} = \operatorname{trid}(10, 9, \ldots, 1, 0, 1, \ldots, 9, 10; 1, \ldots, 1) \in \mathbf{S}^{21,21} \tag{5}$$

mit den Eigenwerten

$$\lambda_{20} = 10.74619\,41829\,0339\ldots, \qquad \lambda_{21} = 10.74619\,41829\,0332\ldots$$

dar, siehe B 13.8.

A. *Die symmetrische Faktorisierung und das Schneiden des Spektrums*

Die in diesem Abschnitt beschriebenen Algorithmen beruhen auf der Möglichkeit, für gegebenes $\mu \in \mathbf{R}$ die Anzahl $s(\mu) \in \{0, \ldots, n\}$ derjenigen Eigenwerte λ_j von $\boldsymbol{T}$, die kleiner als μ sind, einfach berechnen zu können. Bei der Numerierung (4) ist $s = s(\mu)$ durch

$$\lambda_1 < \lambda_2 < \cdots < \lambda_s < \mu \leqq \lambda_{s+1} < \cdots < \lambda_n \tag{6}$$

festgelegt. Durch μ wird also das Spektrum in die beiden Bereiche $\{\lambda_1, \ldots, \lambda_s\}$ und $\{\lambda_{s+1}, \ldots, \lambda_n\}$ „zerschnitten", die links bzw. rechts von μ auf der λ-Achse liegen. Offensichtlich ist $s(\mu)$ gleich der Anzahl der negativen Eigenwerte $\lambda_j[\boldsymbol{T}_\mu] = \lambda_j - \mu$

von

$$\boldsymbol{T}_\mu := \boldsymbol{T} - \mu \boldsymbol{I}. \tag{7}$$

Um $s(\mu)$ einfach berechnen zu können, nutzen wir diesen Fakt aus und erinnern uns an die folgende Aussage, siehe Ü 6.1.4: Wenn die symmetrische Dreiecksfaktorisierung

$$\boldsymbol{T}_\mu = \boldsymbol{L}_\mu \boldsymbol{D}_\mu \boldsymbol{L}_\mu^{\mathsf{T}} \tag{8}$$

mit

$$\boldsymbol{D}_\mu = \begin{pmatrix} d_1 & & & & \\ & d_2 & & & \\ & & d_3 & & \\ & & & \ddots & \\ & & & & d_n \end{pmatrix}, \qquad \boldsymbol{L}_\mu = \begin{pmatrix} 1 & & & & \\ l_2 & 1 & & & \\ & l_3 & 1 & & \\ & & \ddots & \ddots & \\ & & & l_n & 1 \end{pmatrix} \tag{9}$$

existiert, sind $\boldsymbol{T}_\mu$ und $\boldsymbol{D}_\mu$ kongruent, und *die Anzahl der negativen Diagonalelemente von* $\boldsymbol{D}_\mu$ *ist gleich der Anzahl der negativen Eigenwerte von* $\boldsymbol{T}_\mu$, *also gleich* $s(\mu)$. Daß $\boldsymbol{L}_\mu$ bei Existenz der Zerlegung bidiagonal ist, folgt sofort aus der Tridiagonalität von $\boldsymbol{T}_\mu$, vgl. 6.4.1.

Elementweise Auswertung von (8) führt auf die Identitäten

$$a_1 - \mu = d_1, \qquad a_k - \mu = d_k + l_k * d_{k-1} * l_k \qquad (k = 2, \ldots, n), \tag{10}$$

$$b_k = l_k * d_{k-1} \qquad (k = 2, \ldots, n). \tag{11}$$

Wegen (1) und (11) ist

$$d_{k-1} \neq 0 \qquad (k = 2, \ldots, n), \tag{12}$$

so daß nach $l_k = b_k/d_{k-1}$ aufgelöst werden kann. Wenn auch (10) nach d_k aufgelöst und rechts $l_k * d_{k-1}$ durch b_k ersetzt wird, ergibt sich der folgende Algorithmus zur Berechnung von $\{d_k, l_k\}$ aus $\{a_k, b_k, \mu\}$:

13.6.3. ***LDL*$^{\mathsf{T}}$*-Faktorisierung von* $\boldsymbol{T}_\mu$. Für $\mu \in \mathbf{R}$ und nichtzerfallendes $\boldsymbol{T} = \mathrm{trid}(a_1, \ldots, a_n, b_2, \ldots, b_n)$ existiere die Zerlegung (8), (9). Dann ist der Algorithmus

$ie := 0, \quad d_1 := a_1 - \mu$
for $k := 2(1)n$ **do**
  if $d_{k-1} = 0$ **then** [$ie := -1$, **stop**]
  $l_k := b_k/d_{k-1}, \quad d_k := a_k - l_k * b_k - \mu$

in exakter Arithmetik mit $ie = 0$ durchführbar und liefert die signifikanten Elemente von $\boldsymbol{D}_\mu$ und $\boldsymbol{L}_\mu$.

Aufwand: $\sim 2n$ opms, $\sim 2n$ S (für $\{d_k, l_k\}$).

Für die Berechnung von $s(\mu)$ — wir sagen dazu auch „*Schneiden des Spektrums mit* μ“ (engl. „slicing the spectrum“) — sind nur die Vorzeichen der d_k erforderlich. Es bietet sich daher an, den Ausdruck für l_k direkt in die Formel für d_k einzusetzen,

was auf

$$d_k := a_k - (b_k)^2/d_{k-1} - \mu$$

führt. Wenn alle d_k auf dem Platz von d gespeichert werden, erhalten wir die folgende Modifikation von 13.6.3:

13.6.4. *Schneiden des Spektrums von T mit μ.* Für $\mu \in \mathbf{R}$ und nichtzerfallendes $T = \text{trid}\,(a_1, \ldots, a_n, b_2, \ldots, b_n)$ existiere die Zerlegung (8), (9). Dann ist der Algorithmus

S1 (Berechnung der Hilfsgrößen $c_k = (b_k)^2$):
for $k := 2(1)n$ **do** $c_k := (b_k)^2$

S2 (Berechnung von $s = s(\mu)$ aus $\{a_k, c_k, \mu\}$):
$ie := 0$, $d := a_1 - \mu$, **if** $d < 0$ **then** $s := 1$ **else** $s := 0$
for $k := 2(1)n$ **do**
  if $d = 0$ **then** [$ie := -1$, **stop**]
  $d := a_k - c_k/d - \mu$
  if $d < 0$ **then** $s := s + 1$

in exakter Arithmetik mit $ie = 0$ durchführbar und liefert die Zahl $s = s(\mu)$ als Anzahl der Eigenwerte von T, die kleiner als μ sind.

Aufwand: $\sim n$ opm + $\sim n$ S (für $\{c_k\}$), $\sim n(2$ ops + 1 opm) (für s)

13.6.5. *Bemerkung.* (i) Algorithmus 13.6.3 ist identisch mit der symmetrischen Version 6.4.4(i) des Faktorisierungsalgorithmus 6.4.3 für eine Tridiagonalmatrix. Die Voraussetzung der Durchführbarkeit, d. h. der Existenz der Zerlegung (8), ist äquivalent zu (12) und praktisch nicht einschneidend: Offensichtlich ist $T_\mu = L_\mu R_\mu$, $R_\mu := D_\mu L_\mu^T$ die Gauß-Faktorisierung von T_μ. Nach Ü 5.2.5 existiert diese genau dann mit (12), wenn

$$\tau_k := \det\left((T_\mu)_k\right) \neq 0 \qquad (k = 1, \ldots, n-1) \tag{13}$$

gilt, wobei

$$(T_\mu)_k := T_k - \mu I_k, \qquad T_k := \text{trid}\,(a_1, \ldots, a_k, b_2, \ldots, b_k) \tag{14}$$

ist. Nun ist τ_k der Wert des charakteristischen Polynoms der k-ten Hauptabschnittsmatrix $(T)_k$ von T und kann an höchstens k verschiedenen Stellen verschwinden, so daß (13) für insgesamt höchstens $\sum_{k=1}^{n-1} k \sim n^2/2$ Werte von μ nicht erfüllt ist. Wenn zufälligerweise doch $d_{k-1} = 0$ ist, sollte mit einem leicht modifizierten Wert von μ neu gestartet werden, oder d_{k-1} sollte durch eine kleine, von 0 verschiedene Zahl ersetzt werden, siehe 13.6.7(iv).

(ii) Zur Eigenwertberechnung muß 13.6.4 mehrfach für verschiedene μ-Werte ausgeführt werden. Es ist daher sinnvoll, den von μ unabhängigen Schritt S1 nur einmal zu Beginn der Rechnung auszuführen. Man beachte, daß in 13.6.4 nur die Quadrate der b_k vorkommen. Dies ist nicht verwunderlich, denn die Eigenwerte von T sind unabhängig vom Vorzeichen der b_k, siehe Ü 13.6.1.

(iii) Nach Ü 5.2.5 gilt

$$d_k = \tau_k/\tau_{k-1} \qquad (k = 1, \ldots, n), \qquad \tau_0 := 1, \tag{15}$$

so daß $s = s(\mu)$ auch als Anzahl der Vorzeichenwechsel in der Folge $\{\tau_0, \tau_1, \ldots, \tau_n\}$ berechnet werden kann. Die Zahlen $\{\tau_k\}$ lassen sich mittels einer dreigliedrigen Rekursionsformel einfach ermitteln, siehe Ü 13.6.2. □

13.6.6. *Rundungsfehleranalyse.* Für $\mu \in \mathfrak{R}$ und nichtzerfallendes $\boldsymbol{T} = \operatorname{trid}(a_1, \ldots, a_n, b_2, \ldots, b_n)$ sei Algorithmus 13.6.3 bzw. 13.6.4 mit $ie = 0$ durchführbar. Dann gilt für die durch die berechneten Ausgangsdaten $\{d_k, l_k\}$ bzw. $\{d_k\}$ festgelegten Matrizen $\boldsymbol{D}_\mu$ und $\boldsymbol{L}_\mu$

$$\boldsymbol{T}_\mu + \delta\boldsymbol{T}_\mu = \boldsymbol{T} + \delta\boldsymbol{T}_\mu - \mu\boldsymbol{I} = \hat{\boldsymbol{L}}_\mu \boldsymbol{D}_\mu \hat{\boldsymbol{L}}_\mu^{\mathsf{T}} \tag{16}$$

mit einer Dreiecksmatrix

$$\hat{\boldsymbol{L}}_\mu = \begin{pmatrix} 1 & & & \\ \hat{l}_2 & 1 & & \\ & \hat{l}_3 & 1 & \\ & & \ddots & \ddots \\ & & & \hat{l}_n \quad 1 \end{pmatrix} \in \mathbf{R}^{n,n}$$

und einer tridiagonalen Störung $\delta\boldsymbol{T}_\mu \in \mathbf{S}^{n,n}$, die den Abschätzungen

$$|\delta\boldsymbol{T}_\mu| \leqq \nu[2\,|\boldsymbol{T}| + |\mu|\,\boldsymbol{I}] \tag{17}$$

sowie

$$|\hat{l}_k - l_k| \leqq 2\nu\,|l_k| \qquad (k = 2, \ldots, n) \tag{18}$$

genügen.

13.6.7. *Bemerkung.* (i) Praktisch werden die angegebenen Algorithmen nur mit μ-Werten, für die

$$\begin{aligned} &|\mu| \leqq \|\boldsymbol{T}\|_\infty = \max\{|a_k| + |b_k| + |b_{k+1}| : k = 1, \ldots, n\}, \\ &b_1 := b_{n+1} := 0 \end{aligned} \tag{19}$$

gilt, durchgeführt. Aus (17) folgt dann

$$\|\delta\boldsymbol{T}_\mu\|_\infty \leqq \nu\{2\,\|\boldsymbol{T}\|_\infty + |\mu|\} \leqq 3\nu\,\|\boldsymbol{T}\|_\infty, \tag{20}$$

d. h., die berechnete Diagonalmatrix $\boldsymbol{D}_\mu$ ist exakt kongruent zu $\boldsymbol{T}_\mu + \delta\boldsymbol{T}_\mu$, und $\delta\boldsymbol{T}_\mu$ ist klein gegenüber $\boldsymbol{T}$. Die berechnete Zahl $s = s(\mu)$ ist also die Anzahl der Eigenwerte von $\boldsymbol{T} + \delta\boldsymbol{T}_\mu$, die kleiner als μ sind. *Die Berechnung von $s = s(\mu)$ nach* 13.6.4 *ist daher ein numerisch gutartiger Prozeß.*

(ii) Wenn $|\mu|$ in der Größenordnung von $\min\{|a_k|\}$ liegt, ist $\delta\boldsymbol{T}_\mu$ eine elementweise kleine Störung von $\boldsymbol{T}$. Dies zieht nach sich, daß unter Verwendung von $s(\mu)$ oft auch die betragskleinsten Eigenwerte einer gestuften Tridiagonalmatrix $\boldsymbol{T}$ ausreichend genau lokalisiert werden können.

(iii) Aus 13.6.6 folgt

$$\boldsymbol{T}_\mu + \delta\hat{\boldsymbol{T}}_\mu = \boldsymbol{L}_\mu \boldsymbol{D}_\mu \boldsymbol{L}_\mu^T \quad \text{mit} \quad \delta\hat{\boldsymbol{T}}_\mu := \delta\boldsymbol{T}_\mu + \delta\tilde{\boldsymbol{T}}_\mu, \tag{21}$$

wobei

$$|\delta\tilde{\boldsymbol{T}}_\mu| \leqq 2\nu \begin{pmatrix} 0 & |b_2| & & & \\ |b_2| & 2\,|l_2 b_2| & |b_3| & & \\ & \ddots & \ddots & \ddots & \\ & & & |b_n| & 2\,|l_n b_n| \end{pmatrix} . \tag{22}$$

ist, siehe Ü 13.6.3. Da die Produkte $|l_k b_k|$ groß werden können, ist $\delta\tilde{\boldsymbol{T}}_\mu$ nicht notwendig klein gegenüber $\boldsymbol{T}_\mu$, d. h., die berechneten Faktoren $\boldsymbol{D}_\mu$, $\boldsymbol{L}_\mu$ sind i. allg. nicht die exakten Faktoren einer leicht gestörten Matrix. *Der Faktorisierungsalgorithmus* 13.6.3 *ist also i. allg. nicht numerisch gutartig.*

(iv) Bei der Berechnung von $d_k := \mathrm{fl}\,(a_k - l_k * b_k - \mu)$ gilt

$$d_k = \{[a_k - l_k * b_k(1 + \varepsilon_k)]\,(1 + \delta_k) - \mu\}\,(1 + \xi_k) \tag{23}$$

mit $|\varepsilon_k|, |\delta_k|, |\xi_k| \leqq \nu$, also

$$|d_k - [a_k - l_k * b_k - \mu]| \leqq \nu[2\,|a_k| + 3\,|l_k * b_k| + |\mu|] =: \Delta d_k .$$

Eine Störung des berechneten d_k in der Größenordnung von Δd_k ändert also das Fehlerniveau von d_k nicht wesentlich. Dies kann ausgenutzt werden, um in 13.6.3 vorzeitigen erfolglosen Abbruch wegen $d_{k-1} = 0$ zu vermeiden, etwa durch Verwendung der folgenden Modifikation:

$ie := 0,\ d_1 := a_1 - \mu + \nu(|a_1| + |\mu|)$
if $d_1 = 0$ **then** [$ie := -1$, **stop**]
for $k := 2(1)n$ **do**
  $l_k := b_k/d_{k-1},\ p := l_k * b_k,\ d_k := a_k - p - \mu$
  if $d_k = 0$ **then** $d_k := \nu(|a_k| + |p| + |\mu|)$

Bis auf den Ausnahmefall $a_1 = \mu = 0$ ist damit unerwünschter Abbruch ausgeschlossen, und die Abschätzungen in 13.6.6 werden nur unwesentlich schlechter. Analog kann 13.6.4 modifiziert werden; die Korrekturgröße ist dann $d_k := \nu\{|a_k| + |c_k/d_{k-1}| + |\mu|\}$. Die durch die Modifikation hervorgerufene Überlaufgefahr ist gering, wenn $\|\boldsymbol{T}\|_\infty$ nicht zu groß ist. □

B. *Das Bisektionsverfahren*

Wir zeigen im folgenden, wie durch das Schneiden des Spektrums gemäß 13.6.4 jeder Eigenwert λ_j von $\boldsymbol{T}$ mit vorgegebenem Index j bezüglich der Numerierung (4) in exakter Arithmetik beliebig genau lokalisiert werden kann. Es sei dazu $[\alpha, \beta]$ ein Intervall mit

$$s(\alpha) < j \leqq s(\beta). \tag{24}$$

Nach Definition von $s(\mu)$ ist dies zu

$$\alpha \leqq \lambda_j < \beta \tag{25}$$

äquivalent, vgl. (6). Um die Lokalisierungsaussage (25) zu verbessern, berechnen wir den Mittelpunkt $\mu := (\alpha + \beta)/2$ des Intervalls $[\alpha, \beta]$ und bestimmen $s = s(\mu)$ nach 13.6.4. Dann können zwei Fälle auftreten:

Fall 1: $s(\mu) < j$. Dann sind weniger als j Eigenwerte kleiner als μ, d. h., es gilt

$$\mu \leqq \lambda_j < \beta.$$

Fall 2: $s(\mu) \geqq j$. Dann sind mindestens j Eigenwerte kleiner als μ, d. h., es gilt

$$\alpha \leqq \lambda_j < \mu.$$

Im ersten Fall wird $[\alpha, \beta]$ durch $[\mu, \beta]$ ersetzt, im zweiten durch $[\alpha, \mu]$, wobei die Intervallänge halbiert wird. Fortsetzung dieses Vorgehens führt auf das im folgenden angegebene Bisektionsverfahren:

13.6.8. *Das Bisektionsverfahren zur Lokalisierung von λ_j.* Gegeben seien die nichtzerfallende Tridiagonalmatrix $\boldsymbol{T} = \operatorname{trid}(a_1, \ldots, a_n, b_2, \ldots, b_n)$, ein Index $j \in \{1, \ldots, n\}$, Zahlen $\{\alpha, \beta\}$ mit (25) und eine Abbruchschranke $\varepsilon > 0$. Dann werden die Eingangsdaten $\{\alpha, \beta\}$ durch den in exakter Arithmetik ausgeführten Algorithmus

S 1 (Bisektionsschritt): $\delta := \beta - \alpha$
  if $\delta > \varepsilon$ **then**
    $\mu := \alpha + \delta/2$, berechne $s = s(\mu)$ nach 13.6.4
    if $s < j$ **then** $\alpha := \mu$ **else** $\beta := \mu$
    goto S 1

mit Zahlen $\{\alpha, \beta\}$ überspeichert, für die (25) mit $\beta - \alpha \leqq \varepsilon$ gilt.

Aufwand: $\sim kn$(2 ops + 1 opm), wobei k — die Anzahl der Bisektionsschritte — die kleinste ganze Zahl größer gleich $\log_2 \left((\beta - \alpha)/\varepsilon\right)$ ist.

13.6.9. *Bemerkung.* (i) Wenn $[\alpha_1, \beta_1]$ das Startintervall bezeichnet, gilt nach k Bisektionsschritten $\delta_{k+1} = \beta_{k+1} - \alpha_{k+1} = (\beta_1 - \alpha_1)/2^k$, d. h., die Abbruchbedingung $\delta_{k+1} \leqq \varepsilon$ ist zu

$$k \geqq \log_2 \left((\beta - \alpha)/\varepsilon\right) \tag{26}$$

äquivalent. Dies führt auf die oben angegebene Schrittzahl.

(ii) Als Starteinschließung (25) kann die aus dem Satz von GERŠGORIN folgende Einschließung mit

$$\begin{aligned} \alpha &:= G_{\min} := \min \{a_k - |b_k| - |b_{k+1}| : k = 1, \ldots, n\}, \\ \beta &:= G_{\max} := \max \{a_k + |b_k| + |b_{k+1}| : k = 1, \ldots, n\}, \end{aligned} \tag{27}$$

$b_1 := b_{n+1} := 0$, verwendet werden, die für jedes j gültig ist, siehe 1.2.10. Im Fall $j = n$ könnte $\lambda_n = G_{\max}$, also $s(G_{\max}) = n - 1$ sein. Dann ist theoretisch $G_{\max}$ durch eine wenig größere Zahl zu ersetzen. Praktisch tritt dieser Fall nicht auf. Für jedes $\mu \in [\alpha, \beta]$ gilt bei der Festlegung (27)

$$|\mu| \leqq \max \{|\alpha|, |\beta|\} = G_{\max} = \|\boldsymbol{T}\|_\infty, \tag{28}$$

d. h., (19) ist in allen Schritten erfüllt.

(iii) Mit Hilfe des Bisektionsverfahrens 13.6.8 können in offensichtlicher Weise z. B. die p kleinsten Eigenwerte $\{\lambda_1, \ldots, \lambda_p\}$, die p größten Eigenwerte $\{\lambda_{n-p+1}, \ldots, \lambda_n\}$, die im Intervall $[\alpha, \beta)$ liegenden Eigenwerte $\{\lambda_i, \ldots, \lambda_j\}$ o. ä. berechnet werden. Im letzten Fall ist $i = s(\alpha) + 1$, $j = s(\beta)$. Wenn wie in den angegebenen Fällen mehrere Eigenwerte $\{\lambda_{j_1}, \ldots, \lambda_{j_p}\}$ zu lokalisieren sind und sich die aktuellen Einschließungsintervalle $[\alpha^{(l)}, \beta^{(l)}]$ für λ_{j_l} überlappen, können mit einem berechneten $s = s(\mu)$ alle Intervalle aufdatiert werden, die μ als inneren Punkt enthalten. Dadurch kann in der Anfangsphase eine wesentliche Aufwandssenkung erzielt werden. In der Schlußphase werden die Einschließungsintervalle in der Regel disjunkt sein, so daß jedes für sich behandelt werden muß. □

13.6.10. *Rundungsfehleranalyse.* Es sei $[\alpha, \beta]$ das durch 13.6.8 berechnete Einschließungsintervall. Dann gibt es eine tridiagonale Störung $\delta T \in \mathbf{S}^{n,n}$, so daß

$$\alpha \leqq \lambda_j[T + \delta T] < \beta \quad \text{mit} \quad \beta - \alpha \leqq \varepsilon \tag{29}$$

gilt, wobei

$$|\delta T| \leqq \nu[2\,|T| + \max\{|\alpha|, |\beta|\}\, I] \tag{30}$$

ist.

Beweis. Nach 13.6.6 existieren für das Schlußintervall $[\alpha, \beta]$ Störungen δT_α, δT_β mit

$$\begin{aligned} &\alpha \leqq \lambda_j[T + \delta T_\alpha], \qquad |\delta T_\alpha| \leqq \nu\{2\,|T| + |\alpha|\, I\}, \\ &\lambda_j[T + \delta T_\beta] < \beta, \qquad |\delta T_\beta| \leqq \nu\{2\,|T| + |\beta|\, I\}. \end{aligned} \tag{31}$$

Im Fall $\alpha \leqq \lambda_j[T] < \beta$ gelten (29) und (30) mit $\delta T = O$. Im Fall $\lambda_j[T] < \alpha$ ist

$$\lambda_j[T + 0 \cdot \delta T_\alpha] < \alpha \leqq \lambda_j[T + 1 \cdot \delta T_\alpha].$$

Wegen der stetigen Abhängigkeit der Eigenwerte von der Matrix gibt es daher ein ξ, $0 < \xi \leqq 1$, so daß (29) für $\delta T := \xi \cdot \delta T_\alpha$ erfüllt ist, und wegen (31) gilt auch (30). Im Fall $\beta \leqq \lambda_j[T]$ kann analog geschlossen werden. □

13.6.11. *Bemerkung.* (i) Für die Anfangseinschließung (27) folgt aus (30) wegen (28)

$$\|\delta T\|_\infty \leqq 3\nu\, \|T\|_\infty, \tag{32}$$

d. h., im berechneten Intervall $[\alpha, \beta]$ liegt der j-te Eigenwert der zu T im Sinne von (32) benachbarten Matrix $T + \delta T$. Dies bedeutet die *numerische Gutartigkeit* des Bisektionsverfahrens.

(ii) Aus (29) folgt

$$|\lambda_j[T + \delta T] - \mu| \leqq (\beta - \alpha)/2 \leqq \varepsilon/2 \quad \text{mit} \quad \mu := (\alpha + \beta)/2.$$

Andererseits gilt nach 13.1.2 und (32)

$$|\lambda_j[T + \delta T] - \lambda_j[T]| \leqq \|\delta T\| \leqq \|\,|\delta T|\,\| \leqq \|\delta T\|_\infty \leqq 3\nu\, \|T\|_\infty,$$

für das vorletzte Ungleichheitszeichen siehe Ü 1.2.9, so daß sich insgesamt

$$|\lambda_j[\boldsymbol{T}] - \mu| \leqq \varepsilon/2 + 3\nu \, \|\boldsymbol{T}\|_\infty \tag{33}$$

als Abschätzung des Fehlers von μ gegen $\lambda_j[\boldsymbol{T}]$ ergibt.

(iii) Der Wahl $\varepsilon := \nu \, \|\boldsymbol{T}\|_\infty$ entspricht i. allg. die größtmögliche Genauigkeit des Bisektionsverfahrens; aus (33) folgt dann

$$|\lambda_j[\boldsymbol{T}] - \mu| \leqq 3.5\nu \, \|\boldsymbol{T}\|_\infty \leqq 3.5\nu \sqrt{3} \, \|\boldsymbol{T}\| \approx 6.1\nu \, \|\boldsymbol{T}\|, \tag{34}$$

man beachte Ü 13.6.4. Andererseits ergibt sich aus Ü 13.6.4 und 13.1.2 das optimale Fehlerniveau von $\lambda_j[\boldsymbol{T}]$ zu

$$|\lambda_j[\boldsymbol{T} + \delta\boldsymbol{T}_D] - \lambda_j[\boldsymbol{T}]| \leqq \|\delta\boldsymbol{T}_D\| \leqq \sqrt{3} \, \nu \, \|\boldsymbol{T}\| \approx 1.7\nu \, \|\boldsymbol{T}\|, \tag{35}$$

siehe auch (13.1.22) für den Fall einer allgemeinen Matrix, wo $\sqrt{n}$ statt $\sqrt{3}$ steht. Der Vergleich von (34) und (35) zeigt, daß das Bisektionsverfahren mit der obigen ε-Wahl hinsichtlich der Genauigkeit fast optimal ist. Wenn für die Starteinschließung $\beta - \alpha \approx \|\boldsymbol{T}\|_\infty$ gilt, sind jedoch der Größenordnung nach $\log_2(1/\nu)$ Schritte erforderlich. Selbst mit den in 13.6.9(iii) beschriebenen Methoden zur Effektivitätserhöhung werden dann zur Bestimmung aller Eigenwerte der Größenordnung nach $\sim \log_2(1/\nu) \times n^2$ opms benötigt. Dies ist deutlich mehr als bei Verwendung der Varianten des $\boldsymbol{QR}$-Algorithmus aus 13.7. In der Literatur wird empfohlen, das Bisektionsverfahren nur zu verwenden, wenn weniger als $n/4$ Eigenwerte gesucht sind. Andernfalls ist es billiger, alle Eigenwerte mittels des $\boldsymbol{QR}$-Algorithmus zu berechnen.

(iv) Falls die betragskleinen Eigenwerte einer gestuften Tridiagonalmatrix $\boldsymbol{T}$ zu bestimmen sind, sollte $\varepsilon \approx \nu \, \|\boldsymbol{T}_0\|_\infty$ gewählt werden, wobei $\boldsymbol{T}_0$ den betragskleinen Teil von $\boldsymbol{T}$ bezeichnet.

(v) Da das Bisektionsverfahren i. allg. eine sehr genaue Näherung μ für den gesuchten Eigenwert λ von $\boldsymbol{T}$ liefert, kann eine Näherung $\boldsymbol{v}$ für den zugehörigen Eigenvektor $\boldsymbol{u}$ durch wenige Schritte der inversen Iteration 13.4.3 berechnet werden. Man beachte jedoch, daß die für das Schneiden des Spektrums geeignete $\boldsymbol{LDL}^\mathsf{T}$-Faktorisierung aus 13.6.3 für die Gleichungsauflösung bei der inversen Iteration i. allg. ungeeignet ist, denn sie ist nicht numerisch gutartig, siehe 13.6.7(iii) und Text nach 13.4.5. Es sollte die in 13.4.3 vorgesehene pivotisierte Gauß-Faktorisierung verwendet werden, die nur $\sim n(1 \text{ opm} + 2 \text{ opms})$ kostet, siehe 6.4.1. Da $\boldsymbol{T}$ in der Regel nach dem Tridiagonalisierungsverfahren 13.5.1 gemäß $\boldsymbol{T} = \boldsymbol{Q}^\mathsf{T}\boldsymbol{A}\boldsymbol{Q}$ aus einer voll besetzten Matrix $\boldsymbol{A}$ entstanden ist, muß der Eigenvektor $\boldsymbol{v}$ von $\boldsymbol{T}$ noch in den Eigenvektor $\boldsymbol{Qv}$ von $\boldsymbol{A}$ rücktransformiert werden. Mit der Produktform von $\boldsymbol{Q}$ erfordert dies $\sim n^2$ opms pro Eigenvektor, siehe 13.5.2(ii).

(vi) Wenn einige Eigenwerte einer Bandmatrix $\boldsymbol{A}$ gesucht sind, braucht $\boldsymbol{A}$ nicht unbedingt erst nach 13.5.6 tridiagonalisiert zu werden, sondern das Schneiden des Spektrums kann direkt mit $\boldsymbol{A} - \mu\boldsymbol{I}$ unter Verwendung der etwa nach 6.1.1 berechneten $\boldsymbol{LDL}^\mathsf{T}$-Faktorisierung $\boldsymbol{A} - \mu\boldsymbol{I} = \boldsymbol{L}_\mu\boldsymbol{D}_\mu\boldsymbol{L}_\mu^\mathsf{T}$ durchgeführt werden. Der Test „$a_{kk} \leqq 0$“ ist dabei durch „$a_{kk} = 0$“ zu ersetzen, da $\boldsymbol{A} - \mu\boldsymbol{I}$ i. allg. nicht positiv definit ist. Im Gegensatz zum tridiagonalen Fall kann jedoch die berechnete Diagonalmatrix $\boldsymbol{D}_\mu$ zu einer Matrix $\boldsymbol{A} + \delta\boldsymbol{A}$ kongruent sein, die sich von $\boldsymbol{A}$ stark unter-

scheidet. Damit ist insbesondere im Fall großer $|l_{ij}|$ zu rechnen. Wenn $a_{kk} = 0$ ist oder große $|l_{ij}|$ auftreten, sollte die Faktorisierung deshalb mit einem leicht veränderten Wert von μ wiederholt werden. Meist gelingt es auf diese Weise, den gesuchten Eigenwert so genau zu approximieren, daß einige nachfolgende Schritte der inversen oder der RQ-Iteration ein akzeptables Eigenpaar liefern. Für jede der genannten Iterationen sollte jedoch eine numerisch gutartige Faktorisierung — etwa die spaltenpivotisierte Gauß-Faktorisierung — benutzt werden, vgl. 13.4. □

Übungsaufgaben

Ü 13.6.1. Man konstruiere eine Matrix $\boldsymbol{D} = \operatorname{diag}(d_i)$ mit $|d_i| = 1$, so daß

$$\operatorname{trid}(a_1, \ldots, a_n, |b_2|, \ldots, |b_n|) = \boldsymbol{D}[\operatorname{trid}(a_1, \ldots, a_n, b_2, \ldots, b_n)]\,\boldsymbol{D}$$

gilt. Wegen $\boldsymbol{D} = \boldsymbol{D}^{\mathsf{T}} = \boldsymbol{D}^{-1}$ sind die beiden Tridiagonalmatrizen also ähnlich und besitzen dieselben Eigenwerte. Letztere hängen daher nur von den Beträgen der Nebendiagonalelemente ab.

Ü 13.6.2. Man zeige durch Entwicklung von $\det((\boldsymbol{T}_\mu)_k)$ nach der letzten Spalte, daß die Hauptabschnittsdeterminanten τ_k von $\boldsymbol{T}_\mu$ der Rekursion

$$\tau_k = (a_k - \mu)\,\tau_{k-1} - (b_k)^2\,\tau_{k-2} \qquad (k = 2, \ldots, n) \tag{36}$$

mit $\tau_0 = 1$ genügen. Die Folge $\{\tau_0, \ldots, \tau_n\}$ und damit $s = s(\mu)$ als Zahl der Vorzeichenwechsel in dieser Folge läßt sich also mit $\sim n(2 \text{ ops} + 2 \text{ opm})$ bestimmen, falls $c_k = b_k^2$ bekannt ist, allerdings kann für großes n Über- bzw. Unterlauf eintreten. Wenn zu dessen Vermeidung zu den Quotienten $d_k := \tau_k/\tau_{k-1}$ übergegangen wird, geht (36) in die Rekursion aus 13.6.4 über.

Ü 13.6.3. Man beweise die Gültigkeit von (21) mit $\delta\tilde{\boldsymbol{T}}_\mu := -[\delta\boldsymbol{L}_\mu\boldsymbol{D}_\mu\boldsymbol{L}_\mu^{\mathsf{T}} + \boldsymbol{L}_\mu\boldsymbol{D}_\mu\delta\boldsymbol{L}_\mu^{\mathsf{T}} + \delta\boldsymbol{L}_\mu\boldsymbol{D}_\mu\delta\boldsymbol{L}_\mu^{\mathsf{T}}]$, $\delta\boldsymbol{L}_\mu := \hat{\boldsymbol{L}}_\mu - \boldsymbol{L}_\mu$ und schätze $|\delta\tilde{\boldsymbol{T}}_\mu|$ gemäß (22) ab.

Ü 13.6.4. Man zeige, daß für jede Tridiagonalmatrix $\boldsymbol{T} = \operatorname{trid}(a_j, b_j)$

$$\|\boldsymbol{T}\|_\infty \leqq \sqrt{3}\,\|\boldsymbol{T}\| \tag{37}$$

gilt, und folgere hieraus die Abschätzung

$$\|\delta\boldsymbol{T}_D\| \leqq \nu\sqrt{3}\,\|\boldsymbol{T}\| \tag{38}$$

für den Darstellungsfehler $\delta\boldsymbol{T}_D = \operatorname{rd}(\boldsymbol{T}) - \boldsymbol{T}$ von $\boldsymbol{T}$.

Hinweis: Wenn k die Zeile mit maximaler Betragssumme von $\boldsymbol{T}$ bezeichnet, gilt

$$\|\boldsymbol{T}\|_\infty \leqq \sqrt{3}\,\sqrt{|b_k|^2 + |a_k|^2 + |b_{k+1}|^2} = \sqrt{3}\,\sqrt{(\boldsymbol{T}^2)_{kk}}.$$

Zum Nachweis von (38) beachte man 2.3.15 und Ü 1.2.9.

13.7. Der QR-Algorithmus

Der vor etwa 25 Jahren entdeckte QR-Algorithmus — zur Geschichte siehe B 13.9 — ist der effektivste der derzeit bekannten Algorithmen zur Berechnung aller Eigenwerte und gegebenenfalls Eigenvektoren einer symmetrischen Tridiagonalmatrix. In Kombination mit den im Abschnitt 13.5 diskutierten Tridiagonalisierungsver-

fahren stellt er das Verfahren zur Lösung des vollständigen Eigenwertproblems voll besetzter symmetrischer Matrizen mittlerer Dimension dar.

Da ein großer Teil der Theorie auch für beliebige symmetrische Matrizen gilt, setzen wir zunächst außer der Symmetrie keine weiteren Eigenschaften der Matrix voraus.

A. *Motivierung, Grundform und Zusammenhang mit Teilraumiteration*

Ausgangspunkt der folgenden Überlegungen ist der Algorithmus 13.3.10 zur Teilraumiteration, der mit $p = m = n$ zur Berechnung aller Eigenvektoren von $\boldsymbol{A}$ verwendet werden kann. Um später mit der beim $\boldsymbol{QR}$-Algorithmus üblichen Bezeichnung nicht zu kollidieren, schreiben wir 13.3.10 in der Form

$$\text{Wähle } \overline{\boldsymbol{V}}_1, \text{ bilde } \overline{\boldsymbol{W}}_{k+1} := \boldsymbol{A}\overline{\boldsymbol{V}}_k, \text{ faktorisiere } \overline{\boldsymbol{W}}_{k+1} \text{ gemäß } \overline{\boldsymbol{W}}_{k+1} = \overline{\boldsymbol{V}}_{k+1}\overline{\boldsymbol{R}}_k \quad (1)$$

($k = 1, 2, \ldots$); in 13.3.10 steht $\overline{\boldsymbol{R}}_{k+1}$ statt $\overline{\boldsymbol{R}}_k$. Dabei sind $\{\overline{\boldsymbol{V}}_k\}$ und $\{\overline{\boldsymbol{R}}_k\}$ Folgen orthogonaler bzw. oberer Dreiecksmatrizen.

Wenn die aus $\boldsymbol{A}$ durch orthogonale Ähnlichkeitstransformation mit $\overline{\boldsymbol{V}}_k$ gemäß

$$\bar{\boldsymbol{A}}_k = \overline{\boldsymbol{V}}_k^{\mathrm{T}}\boldsymbol{A}\overline{\boldsymbol{V}}_k \quad (2)$$

entstehenden Matrizen $\bar{\boldsymbol{A}}_k$ eingeführt werden, ergibt sich wegen (1)

$$\bar{\boldsymbol{A}}_k = \overline{\boldsymbol{V}}_k^{\mathrm{T}}\overline{\boldsymbol{W}}_{k+1} = \overline{\boldsymbol{V}}_k^{\mathrm{T}}\overline{\boldsymbol{V}}_{k+1}\overline{\boldsymbol{R}}_k = \bar{\boldsymbol{Q}}_k\overline{\boldsymbol{R}}_k \quad (3)$$

mit

$$\bar{\boldsymbol{Q}}_k := \overline{\boldsymbol{V}}_k^{\mathrm{T}}\overline{\boldsymbol{V}}_{k+1}, \quad \text{also} \quad \overline{\boldsymbol{V}}_{k+1} = \overline{\boldsymbol{V}}_k\bar{\boldsymbol{Q}}_k = \overline{\boldsymbol{V}}_1\bar{\boldsymbol{Q}}_1\bar{\boldsymbol{Q}}_2 \cdots \bar{\boldsymbol{Q}}_k, \quad (4)$$

d. h., $\overline{\boldsymbol{R}}_k$ ist ein Dreiecksfaktor der $\boldsymbol{QR}$-Faktorisierung (3) von $\bar{\boldsymbol{A}}_k$, und $\overline{\boldsymbol{V}}_{k+1}$ entsteht gemäß (4) aus $\overline{\boldsymbol{V}}_k$ durch Aufdatierung mit dem zugehörigen Orthogonalfaktor $\bar{\boldsymbol{Q}}_k$. Mit (4), (3) folgt weiter

$$\bar{\boldsymbol{A}}_{k+1} = \overline{\boldsymbol{V}}_{k+1}^{\mathrm{T}}\boldsymbol{A}\overline{\boldsymbol{V}}_{k+1} = \bar{\boldsymbol{Q}}_k^{\mathrm{T}}\bar{\boldsymbol{A}}_k\bar{\boldsymbol{Q}}_k = \bar{\boldsymbol{Q}}_k^{\mathrm{T}}\bar{\boldsymbol{Q}}_k\overline{\boldsymbol{R}}_k\bar{\boldsymbol{Q}}_k = \overline{\boldsymbol{R}}_k\bar{\boldsymbol{Q}}_k, \quad (5)$$

d. h., $\bar{\boldsymbol{A}}_{k+1}$ ist das Produkt der Faktoren $\bar{\boldsymbol{Q}}_k$, $\overline{\boldsymbol{R}}_k$ der Faktorisierung (3) in umgekehrter Reihenfolge. Da die $\boldsymbol{QR}$-Faktorisierung für eine reguläre Matrix nach 10.1.3 bis auf das Vorzeichen der Spalten von $\boldsymbol{Q}$ und der zugehörigen Zeilen von $\boldsymbol{R}$ eindeutig ist, könnte man versuchen, die Formeln (3) bis (5) zur Erzeugung der Folgen $\{\bar{\boldsymbol{A}}_k\}$, $\{\bar{\boldsymbol{Q}}_k\}$, $\{\overline{\boldsymbol{R}}_k\}$, $\{\overline{\boldsymbol{V}}_k\}$ zu verwenden. Um die durch möglicherweise andere Vorzeichenfestlegung bei der $\boldsymbol{QR}$-Faktorisierung hervorgerufenen Unterschiede auszudrücken, bezeichnen wir die derart definierten Matrizen ohne Querstrich. Bei der sich anbietenden Festlegung $\boldsymbol{V}_1 := \boldsymbol{I}$, also $\boldsymbol{A}_1 = \boldsymbol{A}$, führt dies auf die folgende Vorschrift, die die Grundform des *$\boldsymbol{QR}$-Algorithmus* darstellt:

13.7.1. ***$\boldsymbol{QR}$-Algorithmus ohne Verschiebungen.***

S0 (Initialisierung): Setze $\boldsymbol{A}_1 := \boldsymbol{A}$, $\boldsymbol{V}_1 := \boldsymbol{I}$, $k := 1$

S1 (Faktorisierung von $\boldsymbol{A}_k$): Berechne $\boldsymbol{QR}$-Faktorisierung $\boldsymbol{A}_k = \boldsymbol{Q}_k\boldsymbol{R}_k$

S2 (Bildung von $\boldsymbol{A}_{k+1}$, $\boldsymbol{V}_{k+1}$): Bestimme $\boldsymbol{A}_{k+1} := \boldsymbol{R}_k\boldsymbol{Q}_k$, $\boldsymbol{V}_{k+1} := \boldsymbol{V}_k\boldsymbol{Q}_k$

S3: Setze $k := k + 1$, **goto** S1

Es gilt dann wie in (5)

$$A_{k+1} = V_{k+1}^{\mathsf{T}} A V_{k+1} = Q_k^{\mathsf{T}} A_k Q_k, \tag{6}$$

und die Transformationsmatrix

$$V_{k+1} = V_k Q_k = Q_1 Q_2 \cdots Q_k \tag{7}$$

als Produkt der Q-Faktoren ist orthogonal.

13.7.2. *Aussage.* Es sei $A \in \mathbf{S}^{n,n}$ regulär, d. h., $\lambda = 0$ sei kein Eigenwert von A. Die Folgen $\{\bar{V}_k\}$, $\{\bar{R}_k\}$ seien nach der Teilraumiteration 13.3.10 mit $\bar{V}_1 := I$ erzeugt, und $\{\bar{A}_k\}$, $\{\bar{Q}_k\}$ bezeichne die durch (2) bzw. (4) zugeordneten Matrizen. Die Folgen $\{A_k\}$, $\{Q_k\}$, $\{R_k\}$, $\{V_k\}$ seien durch den QR-Algorithmus 13.7.1 definiert. Dann gilt

$$\bar{V}_k = V_k D_{k-1} \tag{8}$$

und

$$\bar{A}_k = D_{k-1} A_k D_{k-1}, \qquad \bar{Q}_k = D_{k-1} Q_k D_k, \qquad \bar{R}_k = D_k R_k D_{k-1} \tag{9}$$

für $k = 1, 2, \ldots$ mit

$$D_0 := I \quad \text{und} \quad |D_k| = I, \quad \text{d. h.} \quad D_k = \operatorname{diag}(\pm 1). \tag{10}$$

Beweis. Angenommen, es gelten (8) bis (10) bis zum Index k. Dann folgt

$$\bar{V}_{k+1} = \bar{V}_k \bar{Q}_k = V_k D_{k-1} D_{k-1} Q_k D_k = V_k Q_k D_k = V_{k+1} D_k,$$

also (8) für den Index $k+1$. Weiter ist

$$\bar{A}_{k+1} = \bar{Q}_k^{\mathsf{T}} \bar{A}_k \bar{Q}_k = (D_k Q_k^{\mathsf{T}} D_{k-1})\,(D_{k-1} A_k D_{k-1})\,(D_{k-1} Q_k D_k) = D_k A_{k+1} D_k.$$

Mit dieser Beziehung ergibt sich

$$\bar{Q}_{k+1} \bar{R}_{k+1} = \bar{A}_{k+1} = D_k A_{k+1} D_k = (D_k Q_{k+1})\,(R_{k+1} D_k).$$

Da $\bar{A}_{k+1}$ zu A ähnlich und folglich regulär ist, gibt es nach 10.1.3 ein D_{k+1} mit $|D_{k+1}| = I$, so daß $\bar{Q}_{k+1} = (D_k Q_{k+1}) D_{k+1}$, $\bar{R}_{k+1} = D_{k+1}(R_{k+1} D_k)$ gilt, denn $D_k Q_{k+1}$ und $R_{k+1} D_k$ sind wie $\bar{Q}_{k+1}$ und $\bar{R}_{k+1}$ Faktoren von $\bar{A}_{k+1}$. Die Gültigkeit für $k = 1$ wird analog gezeigt bzw. folgt aus $\bar{V}_1 = V_1 = I$. □

13.7.3. *Bemerkung.* Im Beweis wurde nur ausgenutzt, daß die Folgen $\{\bar{V}_k\}$, $\{\bar{R}_k\}$, $\{\bar{A}_k\}$, $\{\bar{Q}_k\}$ den zu 13.7.1 analogen Gleichungen genügen. *Es wurde also eigentlich gezeigt, daß die durch* 13.7.1 *definierten Matrizenfolgen im wesentlichen* — d. h. im Sinne von (8) bis (10) — *durch A eindeutig festgelegt sind*, sofern A regulär ist. Wird zusätzlich z. B. $(R_k)_{jj} > 0$ $(j = 1, \ldots, n)$ gefordert, was sich durch gleichzeitiges Vertauschen der Vorzeichen der j-ten Spalte von Q_k und der j-ten Zeile von R_k stets erreichen ließe, ist $D_k = I$, und die Folgen sind sogar im strengen Sinne eindeutig festgelegt. □

Da die Spalten von V_k und $\bar{V}_k$ bis auf das Vorzeichen identisch sind, übertragen sich die Konvergenzaussagen für die Teilraumiteration auf den QR-Algorithmus.

13.7.4. *Satz.* Für die Matrix $A \in \mathbf{S}^{n,n}$ mit den Eigenwerten $\{\lambda_j\}$,

$$0 < |\lambda_n| \leqq |\lambda_{n-1}| \leqq \cdots \leqq |\lambda_2| \leqq |\lambda_1|, \tag{11}$$

und zugehörigen orthonormierten Eigenvektoren $\{u^j\}$ werde der QR-Algorithmus 13.7.1 in exakter Arithmetik ausgeführt. Es gelte

$$|\lambda_{p+1}| < |\lambda_p| \tag{12}$$

und

$$\sigma := \cos\left(\sphericalangle\,(\mathcal{Y}_1, \mathcal{S}_p)\right) > 0 \tag{13}$$

mit

$$\mathcal{Y}_1 := \operatorname{span}\{e^1, \ldots, e^p\}, \qquad \mathcal{S}_p := \operatorname{span}\{u^1, \ldots, u^p\}.$$

Die durch 13.7.1 erzeugten Matrizen seien gemäß

$$V_k = (v^{k,1}, \ldots, v^{k,n}) = (V_1^{(k)} \mid V_2^{(k)}),$$

$$A_k = \left(\begin{array}{c|c} A_{11}^{(k)} & A_{12}^{(k)} \\ \hline A_{21}^{(k)} & A_{22}^{(k)} \end{array}\right), \qquad Q_k = \left(\begin{array}{c|c} Q_{11}^{(k)} & Q_{12}^{(k)} \\ \hline Q_{21}^{(k)} & Q_{22}^{(k)} \end{array}\right), \qquad R_k = \left(\begin{array}{c|c} R_{11}^{(k)} & R_{12}^{(k)} \\ \hline O & R_{22}^{(k)} \end{array}\right)$$

partitioniert mit $V_1^{(k)} \in \mathbf{R}^{n,p}$, $A_{11}^{(k)} \in \mathbf{R}^{p,p}$ usw., und es sei

$$\mathcal{Y}_k := \mathcal{R}(V_1^{(k)}) = \operatorname{span}\{v^{k,1}, \ldots, v^{k,p}\}. \tag{14}$$

Dann gilt

$$\tan\varphi_{k+1} \leqq \varkappa \tan\varphi_k \leqq \varkappa^k \tan\varphi_1 = \varkappa^k \sqrt{1-\sigma^2}/\sigma =: \varepsilon_{k+1} \tag{15}$$

mit

$$\varkappa := |\lambda_{p+1}/\lambda_p| < 1 \quad \text{und} \quad \varphi_k := \sphericalangle\,(\mathcal{Y}_k, \mathcal{S}_p) = \sphericalangle\,(\mathcal{Y}_k^{\perp}, \mathcal{X}_{p+1}); \tag{16}$$

dabei ist

$$\mathcal{Y}_k^{\perp} = \mathcal{R}(V_2^{(k)}) = \operatorname{span}\{v^{k,p+1}, \ldots, v^{k,n}\}, \quad \mathcal{X}_{p+1} := \operatorname{span}\{u^{p+1}, \ldots, u^n\} = \mathcal{S}_p^{\perp}.$$

Überdies gelten die Abschätzungen

$$\|A_{12}^{(k)}\| = \|A_{21}^{(k)}\| \leqq (1+\varkappa)\,\varepsilon_k\,\|A\| \leqq 2\varepsilon_k\,\|A\|, \tag{17}$$

$$\|Q_{12}^{(k)}\|, \|Q_{21}^{(k)}\| \leqq (1+\varkappa)\,\varepsilon_k \leqq 2\varepsilon_k, \tag{18}$$

$$\|R_{12}^{(k)}\| \leqq (1+\varkappa)^2\,\varepsilon_k\,\|A\| \leqq 4\varepsilon_k\,\|A\|. \tag{19}$$

Beweis. Die Gültigkeit von (15), (16) folgt sofort aus 13.3.9 und 13.3.11 (ii) unter Beachtung von (13.3.45). Zum Nachweis von (18) gehen wir von $Q_k = V_k^{\mathsf{T}} V_{k+1}$ aus und erhalten

$$Q_{21}^{(k)} = V_2^{(k)\mathsf{T}} V_1^{(k+1)} = [U^{\mathsf{T}} V_2^{(k)}]^{\mathsf{T}}\,[U^{\mathsf{T}} V_1^{(k+1)}] \tag{20}$$

mit $U = (u^1, \ldots, u^p \mid u^{p+1}, \ldots, u^n) =: (U_1 \mid U_2)$. Wir setzen jetzt

$$U^{\mathsf{T}} V_1^{(k+1)} = \left(\begin{array}{c} U_1^{\mathsf{T}} V_1^{(k+1)} \\ \hline U_2^{\mathsf{T}} V_1^{(k+1)} \end{array}\right) =: \left(\begin{array}{c} F \\ \hline G \end{array}\right).$$

Wegen $\|U_1\| = \|V_1^{(k+1)}\| = 1$ ist $\|F\| \leqq 1$. Zur Abschätzung von $\|G\|$ beachten wir, daß

$$\|G\| = \cos\left(\sphericalangle_{\min}(\mathcal{Y}_{k+1}, \mathcal{X}_{p+1})\right) = \sin\varphi_{k+1} \tag{21}$$

gilt, vgl. Ü 13.3.4 und (13.3.46). Mit (15) folgt hieraus $\|G\| \leqq \varepsilon_{k+1}$. In analoger Weise erhält man für

$$U^{\mathsf{T}} V_2^{(k)} =: \begin{pmatrix} G' \\ F' \end{pmatrix}$$

die Abschätzungen $\|F'\| \leqq 1$ und $\|G'\| \leqq \varepsilon_k$. Damit kann (20) gemäß

$$\|Q_{21}^{(k)}\| = \|G'^{\mathsf{T}} F + F'^{\mathsf{T}} G\| \leqq \|G'\|\,\|F\| + \|F'\|\,\|G\| \leqq \varepsilon_k + \varepsilon_{k+1} \leqq (1+\varkappa)\,\varepsilon_k$$

abgeschätzt werden, und dieselbe Abschätzung gilt für $Q_{12}^{(k)}$. Zum Beweis von (17) beachten wir $A_k = Q_k R_k$, also $A_{21}^{(k)} = Q_{21}^{(k)} R_{11}^{(k)}$. Mit $\|R_{11}^{(k)}\| \leqq \|R_k\| = \|A_k\| = \|A\|$ und (18) ergibt sich (17). Schließlich folgt aus $R_k = A_{k+1} Q_k^{\mathsf{T}}$ die Darstellung $R_{12}^{(k)} = A_{11}^{(k+1)} Q_{21}^{(k)\mathsf{T}} + A_{12}^{(k+1)} Q_{22}^{(k)\mathsf{T}}$. Mit $\|A_{11}^{(k+1)}\| \leqq \|A_{k+1}\| = \|A\|$, (18), (17) und $\|Q_{22}^{(k)}\| \leqq 1$ ergibt sich daraus die noch ausstehende Ungleichung (19). □

13.7.5. *Bemerkung.* (i) Die Ungleichung (15) besagt, daß $\mathcal{Y}_k$ gegen $\mathcal{S}_p$ und $\mathcal{Y}_k^{\perp}$ gegen $\mathcal{X}_{p+1}$ linear mit dem Konvergenzfaktor $\varkappa < 1$ konvergieren, und die Konvergenz ist monoton im Sinne von $\tan\varphi_{k+1}/\tan\varphi_k \leqq \varkappa$. Die Abschätzungen (17) bis (19) zeigen, daß die Außerdiagonalblöcke von A_k, Q_k und R_k mit der Majorante ε_k gegen 0 gehen. Obwohl $\varepsilon_{k+1}/\varepsilon_k = \varkappa$ gilt, braucht $\|A_{12}^{(k+1)}\|/\|A_{12}^{(k)}\|$ usw. selbst nicht durch $\varkappa$ beschränkt zu sein, also keine monotone Konvergenz vorzuliegen.

(ii) Meist wird (12) für mehrere Indizes erfüllt sein. Es gelte etwa

$$\cdots \leqq |\lambda_{p'+1}| < |\lambda_{p'}| \leqq \cdots \leqq |\lambda_{p+1}| < |\lambda_p| \leqq \cdots,$$

und (13) sei ebenfalls für p und p' erfüllt. Mit der Partitionierung

$$A_k = \left(\begin{array}{c|c|c} A_{11}^{(k)} & A_{12}^{(k)} & A_{13}^{(k)} \\ \hline A_{21}^{(k)} & A_{22}^{(k)} & A_{23}^{(k)} \\ \hline A_{31}^{(k)} & A_{32}^{(k)} & A_{33}^{(k)} \end{array}\right) \begin{matrix} p \\ p+1 \\ \\ p' \\ p'+1 \end{matrix}$$

folgt dann aus 13.7.4

$$\|A_{21}^{(k)}\|, \|A_{31}^{(k)}\| \leqq \left\| \begin{pmatrix} A_{21}^{(k)} \\ A_{31}^{(k)} \end{pmatrix} \right\| \leqq 2\varepsilon_k \|A\|$$

und

$$\|A_{31}^{(k)}\|, \|A_{32}^{(k)}\| \leqq \|(A_{31}^{(k)} \mid A_{32}^{(k)})\| \leqq 2\varepsilon'_k \|A\|,$$

mithin

$$\|A_{31}^{(k)}\| \leqq 2 \min\{\varepsilon_k, \varepsilon'_k\} \|A\|, \tag{22}$$

wobei $\varepsilon_{k+1} = \varkappa^k \tan\varphi_1$, $\varepsilon'_{k+1} = (\varkappa')^k \tan\varphi'_1$ mit $\varkappa = |\lambda_{p+1}/\lambda_p|$, $\varkappa' = |\lambda_{p'+1}/\lambda_{p'}|$ ist.

(iii) Falls (12), (13) für $p = 1, 2, \ldots, n-1$ erfüllt sind, gilt (22) für jedes $p < p'$. Da speziell $|a_{p'+1,p}^{(k)}| \leqq \|A_{31}^{(k)}\|$ ist, folgt daraus mit $i := p' + 1$, $j := p$ die elementweise Abschätzung

$$|a_{ij}^{(k+1)}| = O\left((\varkappa_{ij})^k\right) \quad \text{mit} \quad \varkappa_{ij} := \min\{|\lambda_i/\lambda_{i-1}|, |\lambda_{j+1}/\lambda_j|\}. \tag{23}$$

Die Außerdiagonalelemente konvergieren also — i. allg. verschieden schnell — gegen 0, und die Diagonalelemente approximieren folglich die gesuchten Eigenwerte mit wachsendem k beliebig genau. □

In der bisher betrachteten Grundform 13.7.1 ist der $\boldsymbol{QR}$-Algorithmus wenig attraktiv: Für voll besetztes $\boldsymbol{A}$ kostet eine $\boldsymbol{QR}$-Faktorisierung mittels Householder-Spiegelungen als billigster Methode $\sim 2n^3/3$ opms, siehe 10.2.11, und für die Bildung von $\boldsymbol{R}_k\boldsymbol{Q}_k$ sind nochmals $\sim 2n^3/3$ opms nötig, so daß ein $\boldsymbol{QR}$-Schritt insgesamt $\sim 4n^3/3$ opms erfordert. Die Konvergenz ist nur linear mit dem Faktor $\varkappa = |\lambda_{p+1}/\lambda_p|$, also i. allg. langsam, so daß viele solcher teuren $\boldsymbol{QR}$-Schritte ausgeführt werden müssen.

Im folgenden diskutieren wir eine Reihe von Maßnahmen, die zur Überwindung der genannten Nachteile führen und die hohe Effektivität des entsprechend verfeinerten $\boldsymbol{QR}$-Algorithmus bewirken.

B. *Deflation und Verschiebungen*

Unter den Voraussetzungen von 13.7.4 werden die Außerdiagonalblöcke von

$$\boldsymbol{A}_k = \left(\begin{array}{c|c} \boldsymbol{A}_{11}^{(k)} & \boldsymbol{A}_{12}^{(k)} \\ \hline \boldsymbol{A}_{21}^{(k)} & \boldsymbol{A}_{22}^{(k)} \end{array}\right) \tag{24}$$

für hinreichend großes k im Sinne von

$$\|\boldsymbol{A}_{12}^{(k)}\| = \|\boldsymbol{A}_{21}^{(k)}\| \leqq \varepsilon := \nu F \, \|\boldsymbol{A}\| \tag{25}$$

mit akzeptablem F genügend klein sein und können im Rahmen der Computergenauigkeit vernachlässigt werden, d. h., $\boldsymbol{A}_k$ kann durch

$$\hat{\boldsymbol{A}}_k = \left(\begin{array}{c|c} \boldsymbol{A}_{11}^{(k)} & \boldsymbol{O} \\ \hline \boldsymbol{O} & \boldsymbol{A}_{22}^{(k)} \end{array}\right)$$

ersetzt werden. Dies entspricht einer Störung

$$\delta\hat{\boldsymbol{A}}_k := -\left(\begin{array}{c|c} \boldsymbol{O} & \boldsymbol{A}_{12}^{(k)} \\ \hline \boldsymbol{A}_{21}^{(k)} & \boldsymbol{O} \end{array}\right), \qquad \|\delta\hat{\boldsymbol{A}}_k\| \leqq \varepsilon,$$

von $\boldsymbol{A}_k$, wegen (6) also einer Störung

$$\delta\tilde{\boldsymbol{A}}_k := \boldsymbol{V}_k\delta\hat{\boldsymbol{A}}_k\boldsymbol{V}_k^{\mathsf{T}}, \qquad \|\delta\tilde{\boldsymbol{A}}_k\| = \|\delta\hat{\boldsymbol{A}}_k\| \leqq \varepsilon,$$

von $\boldsymbol{A}$. Das Eigenwertproblem der Dimension n zerfällt damit in die zwei kleineren Probleme mit den Diagonalblöcken $\boldsymbol{A}_{11}^{(k)}$ und $\boldsymbol{A}_{22}^{(k)}$ der Dimensionen p bzw. $n-p$; man spricht auch hier von *Deflation*. Das Verfahren kann dann mit jedem der Diagonalblöcke einzeln und damit insgesamt billiger fortgesetzt werden, wobei $\boldsymbol{V}_k$ entsprechend aufzudatieren ist. Wird etwa mit dem unteren Block weitergearbeitet und ist $\boldsymbol{A}_{22}^{(k)} = \boldsymbol{Q}_{22}^{(k)}\boldsymbol{R}_{22}^{(k)}$ dessen $\boldsymbol{QR}$-Faktorisierung, so ist der nächste Schritt mit

$$\boldsymbol{Q}_k := \left(\begin{array}{c|c} \boldsymbol{I} & \boldsymbol{O} \\ \hline \boldsymbol{O} & \boldsymbol{Q}_{22}^{(k)} \end{array}\right) \tag{26}$$

durch

$$\boldsymbol{A}_{k+1} = \left(\begin{array}{c|c} \boldsymbol{A}_{11}^{(k+1)} & \boldsymbol{O} \\ \hline \boldsymbol{O} & \boldsymbol{A}_{22}^{(k+1)} \end{array}\right) = \boldsymbol{Q}_k^{\mathsf{T}} \boldsymbol{A}_k \boldsymbol{Q}_k = \left(\begin{array}{c|c} \boldsymbol{A}_{11}^{(k)} & \boldsymbol{O} \\ \hline \boldsymbol{O} & \boldsymbol{R}_{22}^{(k)} \boldsymbol{Q}_{22}^{(k)} \end{array}\right) \tag{27}$$

und

$$\boldsymbol{V}_{k+1} = (\boldsymbol{V}_1^{(k+1)} \mid \boldsymbol{V}_1^{(k+1)}) = \boldsymbol{V}_k \boldsymbol{Q}_k = (\boldsymbol{V}_1^{(k)} \mid \boldsymbol{V}_2^{(k)} \boldsymbol{Q}_{22}^{(k)}) \tag{28}$$

definiert. Dabei wurde wieder $\boldsymbol{A}_k$ statt $\hat{\boldsymbol{A}}_k$ geschrieben.

Je schneller $\boldsymbol{A}_{21}^{(k)}$ gegen $\boldsymbol{O}$ geht, um so weniger QR-Schritte sind bis zur Erfüllung der Annullierungsbedingung (25) nötig und um so geringer wird der Gesamtaufwand. Da die Konvergenzgeschwindigkeit durch $\varkappa = |\lambda_{p+1}/\lambda_p|$ bestimmt wird, versuchen wir in Analogie zum Vorgehen bei der inversen Iteration in 13.4, durch eine geeignete Spektralverschiebung

$$\bar{\boldsymbol{A}} := \boldsymbol{A} - \mu \boldsymbol{I} \tag{29}$$

einen kleineren Konvergenzfaktor zu erreichen. Wie dort numerieren wir die Eigenwerte $\lambda_j - \mu$ von $\bar{\boldsymbol{A}}$ nach wachsenden Beträgen und setzen

$$0 < |\lambda_1 - \mu| \ll |\lambda_2 - \mu| \leqq \cdots \leqq |\lambda_n - \mu| \tag{30}$$

voraus; in der (11) entsprechenden Numerierung ist dann $\lambda_n[\bar{\boldsymbol{A}}] = \lambda_1 - \mu$, $\lambda_{n-1}[\bar{\boldsymbol{A}}] = \lambda_2 - \mu$ usw. Bezüglich $\bar{\boldsymbol{A}}$ ergibt sich damit für $p = n - 1$ der Konvergenzfaktor

$$\bar{\varkappa} = |\lambda_n[\bar{\boldsymbol{A}}]/\lambda_{n-1}[\bar{\boldsymbol{A}}]| = |(\lambda_1 - \mu)/(\lambda_2 - \mu)| \ll 1. \tag{31}$$

Die nach 13.7.1 aus $\bar{\boldsymbol{A}}$ erzeugte Folge $\{\bar{\boldsymbol{A}}_k\}$ der QR-Iterierten ist durch

$$\bar{\boldsymbol{A}}_1 := \bar{\boldsymbol{A}}, \text{ faktorisiere } \bar{\boldsymbol{A}}_k = \bar{\boldsymbol{Q}}_k \bar{\boldsymbol{R}}_k, \text{ bilde } \bar{\boldsymbol{A}}_{k+1} := \bar{\boldsymbol{R}}_k \bar{\boldsymbol{Q}}_k \quad (k = 1, 2, \ldots) \tag{32}$$

definiert. Wenn $\bar{\boldsymbol{A}}_k$ gemäß

$$\bar{\boldsymbol{A}}_k = \left(\begin{array}{c|c} \bar{\boldsymbol{A}}_{11}^{(k)} & \boldsymbol{a}^k \\ \hline \boldsymbol{a}^{k\mathsf{T}} & \bar{a}_{nn}^{(k)} \end{array}\right) \tag{33}$$

partitioniert wird, entspricht $\boldsymbol{a}^k \in \mathbf{R}^{n-1}$ dem Block $\boldsymbol{A}_{12}^{(k)}$ in 13.7.4. Unter der (13) entsprechenden Voraussetzung

$$\cos \bar{\varphi}_1 > 0, \qquad \bar{\varphi}_1 := \sphericalangle\,(\operatorname{span}\{\boldsymbol{e}^1, \ldots, \boldsymbol{e}^{n-1}\}, \mathcal{S}_{n-1}) = \sphericalangle\,(\boldsymbol{e}^n, \boldsymbol{u}^n) \tag{34}$$

gilt dann (17), also

$$\|\boldsymbol{a}^k\| \leqq (1 + \bar{\varkappa})\, \bar{\varepsilon}_k \, \|\bar{\boldsymbol{A}}\| \quad \text{mit} \quad \bar{\varepsilon}_{k+1} = \bar{\varkappa} \bar{\varepsilon}_k = \bar{\varkappa}^k \tan \bar{\varphi}_1. \tag{35}$$

Sofern $\tan \bar{\varphi}_1$ nicht zu groß ist, wird der Block $\boldsymbol{a}^k$ also schnell klein werden, und $\bar{a}_{nn}^{(k)}$ konvergiert gegen $\lambda_n[\bar{\boldsymbol{A}}] = \lambda_1 - \mu$.

Wenn analog zu (29)

$$\bar{\boldsymbol{A}}_k = \boldsymbol{A}_k - \mu \boldsymbol{I}$$

gesetzt wird, schreibt sich (32) als

$$\boldsymbol{A}_1 := \boldsymbol{A}, \text{ faktorisiere } \boldsymbol{A}_k - \mu\boldsymbol{I} = \bar{\boldsymbol{Q}}_k\bar{\boldsymbol{R}}_k, \text{ bilde } \boldsymbol{A}_{k+1} := \bar{\boldsymbol{R}}_k\bar{\boldsymbol{Q}}_k + \mu\boldsymbol{I}$$
$$(k = 1, 2, \ldots). \tag{36}$$

Wegen

$$\boldsymbol{A}_{k+1} = \bar{\boldsymbol{R}}_k\bar{\boldsymbol{Q}}_k + \mu\boldsymbol{I} = \bar{\boldsymbol{Q}}_k^{\mathsf{T}}(\boldsymbol{A}_k - \mu\boldsymbol{I})\,\bar{\boldsymbol{Q}}_k + \mu\boldsymbol{I} = \bar{\boldsymbol{Q}}_k^{\mathsf{T}}\boldsymbol{A}_k\bar{\boldsymbol{Q}}_k = \bar{\boldsymbol{V}}_{k+1}^{\mathsf{T}}\boldsymbol{A}\bar{\boldsymbol{V}}_{k+1} \tag{37}$$

sind die $\boldsymbol{A}_k$ orthogonal ähnlich zur Originalmatrix $\boldsymbol{A}$, die $\bar{\boldsymbol{A}}_k$ dagegen zu $\bar{\boldsymbol{A}}$. Bei Partitionierung von $\boldsymbol{A}_k$ in zu (33) analoger Weise ergibt sich

$$\boldsymbol{A}_k = \left(\begin{array}{c|c} \boldsymbol{A}_{11}^{(k)} & \boldsymbol{a}^k \\ \hline \boldsymbol{a}^{k\mathsf{T}} & a_{nn}^{(k)} \end{array}\right) = \left(\begin{array}{c|c} \bar{\boldsymbol{A}}_{11}^{(k)} + \mu\boldsymbol{I} & \boldsymbol{a}^k \\ \hline \boldsymbol{a}^{k\mathsf{T}} & \bar{a}_{nn}^{(k)} + \mu \end{array}\right), \tag{38}$$

d. h., der Außerdiagonalblock $\boldsymbol{a}^k$ ist derselbe wie in $\bar{\boldsymbol{A}}_k$ und geht gemäß (35) gegen $\boldsymbol{o}$. Sobald $\|\boldsymbol{a}^k\| \leqq \varepsilon$ mit ε aus (25) gilt, wird zu

$$\hat{\boldsymbol{A}}_k = \left(\begin{array}{c|c} \boldsymbol{A}_{11}^{(k)} & \boldsymbol{o} \\ \hline \boldsymbol{o}^{\mathsf{T}} & a_{nn}^{(k)} \end{array}\right) \tag{39}$$

übergegangen. Das Diagonalelement $a_{nn}^{(k)}$ ist dann eine akzeptable Approximation für λ_1, und das Verfahren wird mit $\boldsymbol{A}_{11}^{(k)} \in \mathbf{S}^{n-1,n-1}$ fortgesetzt. Die Dimension reduziert sich dabei um 1. Sollten noch andere Außerdiagonalblöcke von $\boldsymbol{A}_k$ im Sinne von (25) vernachlässigbar sein, so zerfällt $\boldsymbol{A}_{11}^{(k)}$ in weitere Diagonalblöcke niedrigerer Dimension, die dann jeder für sich weiter bearbeitet werden können.

Die bisherigen Überlegungen zeigen, daß der $\boldsymbol{QR}$-Algorithmus mit einer Verschiebung μ, die einen Eigenwert von $\boldsymbol{A}$ möglichst gut approximiert, durchgeführt werden sollte. Dabei bietet es sich an, $\mu = \mu_k$ in der Vorschrift (36) in jedem Schritt auf Grund der aktuellen Information neu festzulegen. Dies führt auf den folgenden Algorithmus:

13.7.6. ***QR-Algorithmus mit Verschiebungen.***
S0: Setze $\boldsymbol{A}_1 := \boldsymbol{A}$, $\boldsymbol{V}_1 := \boldsymbol{I}$, $k := 1$
S1: Wähle Verschiebung μ_k
S2: Berechne $\boldsymbol{QR}$-Faktorisierung $\boldsymbol{A}_k - \mu_k\boldsymbol{I} = \boldsymbol{Q}_k\boldsymbol{R}_k$
S3: Bilde $\boldsymbol{A}_{k+1} := \boldsymbol{R}_k\boldsymbol{Q}_k + \mu_k\boldsymbol{I}$, $\boldsymbol{V}_{k+1} := \boldsymbol{V}_k\boldsymbol{Q}_k$
S4: Setze $k := k + 1$, **goto** S1

Auch für die durch 13.7.6 definierten Matrizen gilt (37) mit $\mu = \mu_k$, d. h., die $\boldsymbol{A}_k$ sind orthogonal ähnlich zu $\boldsymbol{A}$.

13.7.7. *Bemerkung.* Unter der Voraussetzung

$$\mu_k \text{ ist kein Eigenwert von } \boldsymbol{A} \tag{40}$$

ist $\boldsymbol{A}_k - \mu_k\boldsymbol{I}$ regulär, so daß die Faktoren $\boldsymbol{Q}_k$, $\boldsymbol{R}_k$ und damit die durch 13.7.6 erzeugten Matrizenfolgen im Sinne von (8) bis (10) im wesentlichen eindeutig durch

A und $\{\mu_k\}$ festgelegt sind, vgl. 13.7.3. Für die Durchführbarkeit ist (40) jedoch nicht erforderlich, denn die QR-Faktorisierung existiert auch für singuläres $A_k - \mu_k I$. Bei spaltenweiser Erzeugung von R_k durch Linksmultiplikation mit elementaren orthogonalen Matrizen entsprechend 10.2 treten dann in einer Spalte j in den Positionen $\{j, j\}$ bis $\{n, j\}$ Nullen auf, so daß sofort zur nächsten Spalte übergegangen werden kann. Bei der für uns interessanten Anwendung auf Tridiagonalmatrizen führt ein singuläres $A_k - \mu_k I$ in exakter Arithmetik zum Abspalten des Eigenwertes μ_k im nächsten Schritt, siehe 13.7.12(ii). □

Wir diskutieren im folgenden zwei Möglichkeiten zur Festlegung von μ_k. Da $a_{nn}^{(k)}$ unter den Voraussetzungen (30), (34) gegen einen Eigenwert von A konvergiert, liegt es nahe, dieses letzte Diagonalelement als Verschiebung zu benutzen.

13.7.8. *Rayleigh-Quotienten-Verschiebungen.* Spezifiziere S1 in 13.7.6 gemäß
S1: Setze $\mu_k := a_{nn}^{(k)}$

Daß die auf die Rayleigh-Quotienten-Iteration 13.4.6 hinweisende Bezeichnung dieser Verschiebung berechtigt ist, zeigt die folgende Aussage:

13.7.9. *Aussage.* Der QR-Algorithmus 13.7.6 werde mit der Verschiebungsstrategie 13.7.8 realisiert, und es gelte (40). Dann sind die n-ten Spalten $v^{k,n} := V_k e^n$ der Matrizen V_k in exakter Arithmetik bis auf das Vorzeichen identisch mit den Iterierten v^k, die sich nach der RQ-Iteration 13.4.6 mit dem Startvektor $v^1 := e^n$ ergeben.

Beweis. Wegen (37) gilt

$$\mu_k = a_{nn}^{(k)} = e^{n\mathsf{T}} A_k e^n = e^{n\mathsf{T}} V_k^{\mathsf{T}} A V_k e^n = v^{k,n\mathsf{T}} A v^{k,n} = \varrho(v^{k,n}),$$

d. h., μ_k ist der zu $v^{k,n}$ gehörende Rayleigh-Quotient. Aus der Iterationsvorschrift und (37) folgt weiter

$$\begin{aligned}(A - \mu_k I)\, v^{k+1,n} &= V_k (A_k - \mu_k I)^{\mathsf{T}}\, V_k^{\mathsf{T}} v^{k+1,n} = V_k (Q_k R_k)^{\mathsf{T}}\, V_k^{\mathsf{T}} v^{k+1,n} \\ &= V_k R_k^{\mathsf{T}} Q_k^{\mathsf{T}} V_k^{\mathsf{T}} v^{k+1,n} = V_k R_k^{\mathsf{T}} V_{k-1}^{\mathsf{T}} v^{k+1,n} = V_k R_k^{\mathsf{T}} e^n \\ &= r_{nn}^{(k)} V_k e^n = r_{nn}^{(k)} v^{k,n}.\end{aligned}$$

Wegen (40) muß $r_{nn}^{(k)} \neq 0$ sein, so daß $w^{k+1,n} := v^{k+1,n}/r_{nn}^{(k)}$ der Gleichung $(A - \mu_k I)\, w^{k+1,n} = v^{k,n}$ genügt. Daraus folgt die Behauptung. □

Für das zu $\{\mu_k, v^{k,n}\}$ gehörende Residuum gilt

$$(A - \mu_k I)\, v^{k,n} = V_k (A_k - \mu_k I)\, V_k^{\mathsf{T}} v^{k,n} = V_k (A_k - \mu_k I)\, e^n,$$

also mit den Bezeichnungen aus (38)

$$r^{k,n} := (A - \mu_k I)\, v^{k,n} = V_k \begin{pmatrix} a^k \\ \hline 0 \end{pmatrix}, \qquad \|r^{k,n}\| = \|a^k\|. \tag{41}$$

Falls $\{\mu_k, v^{k,n}\}$ ein Eigenpaar von A ist, ergibt sich $\|r^{k,n}\| = \|a^k\| = 0$, so daß $\mu_k = a_{nn}^{(k)}$ abgespalten werden kann und das Verfahren abbricht. Andernfalls können die Konvergenzaussagen für die RQ-Iteration aus 13.4.7 übernommen werden:

13.7.10. *Satz.* Für den $\boldsymbol{QR}$-Algorithmus mit RQ-Verschiebungen fällt $\|\boldsymbol{a}^k\|$ monoton, und fast immer gilt $\|\boldsymbol{a}^k\| \to 0$, $a_{nn}^{(k)} \to \lambda_j$, $\vartheta_k \boldsymbol{v}^{k,n} \to \boldsymbol{u}^j$ mit geeignetem $\vartheta_k \in \{+1, -1\}$. Für genügend großes k ist dabei

$$\tan \varphi_{k+1} \leqq \eta_1 (\tan \varphi_k)^3 \tag{42}$$

mit $\eta_1 > 0$ und $\varphi_k := \sphericalangle(\boldsymbol{u}^j, \vartheta_k \boldsymbol{v}^{k,n})$, und es gelten die Abschätzungen

$$|\lambda_j - a_{nn}^{(k)}| \leqq 2\,\|\boldsymbol{A}\|\,(\tan \varphi_k)^2, \qquad \|\boldsymbol{a}^k\| \leqq \eta_2 \tan \varphi_k \quad \text{mit} \quad \eta_2 > 0. \tag{43}$$

13.7.8 konvergiert also fast immer und monoton, und die Konvergenz ist asymptotisch kubisch.

Die in $\boldsymbol{A}_k$ enthaltene Information wird durch die folgende, auf WILKINSON zurückgehende Verschiebungsstrategie noch besser ausgenutzt:

13.7.11. *Wilkinson-Verschiebungen.* Spezifiziere S1 in 13.7.6 wie folgt:

S1: Berechne die Eigenwerte $\{\mu', \mu''\}$ der in der rechten unteren Ecke von $\boldsymbol{A}_k$ stehenden (2, 2)-Matrix

$$\boldsymbol{P}_k := \begin{pmatrix} a_{n-1,n-1}^{(k)} & a_{n-1,n}^{(k)} \\ a_{n,n-1}^{(k)} & a_{nn}^{(k)} \end{pmatrix} \in \mathbf{S}^{2,2} \tag{44}$$

Ordne $\{\mu', \mu''\}$ gemäß $|\mu' - a_{nn}^{(k)}| < |\mu'' - a_{nn}^{(k)}|$ bzw. $|\mu'| \leqq |\mu''|$, wenn $|\mu' - a_{nn}^{(k)}| = |\mu'' - a_{nn}^{(k)}|$

Setze $\mu_k := \mu'$

Man beachte, daß $\boldsymbol{P}_k$ gerade die zu $\boldsymbol{Q} := (\boldsymbol{v}^{k,n-1}, \boldsymbol{v}^{k,n})$ gehörende Matrix aus dem Rayleigh-Ritz-Algorithmus ist und $\{\mu', \mu''\}$ die entsprechenden Ritzschen Eigenwerte bezüglich span $\{\boldsymbol{v}^{k,n-1}, \boldsymbol{v}^{k,n}\}$ sind, siehe 13.3.C, so daß 13.7.11 als natürliche Erweiterung von 13.7.8 auf den zweidimensionalen Fall angesehen werden kann. Zur numerischen Realisierung siehe Ü 13.7.2. Wir werden später sehen, daß 13.7.11 bei Anwendung auf eine Tridiagonalmatrix sogar globale Konvergenz garantiert.

Bei Verwendung der obigen Verschiebungsstrategien besteht die Tendenz, daß die Eigenwerte in betragsmäßig wachsender Reihenfolge abgespalten werden. Allerdings braucht das nicht immer der Fall zu sein.

C. *Der $\boldsymbol{QR}$-Algorithmus für Tridiagonalmatrizen*

Mit den oben angegebenen Verschiebungen erhöht sich die Konvergenzgeschwindigkeit des $\boldsymbol{QR}$-Algorithmus drastisch; in der Regel sind nur einige Schritte bis zur Abspaltung eines Eigenwertes nötig. Selbst wenn pro Eigenwert nur ein solcher Schritt erforderlich wäre, würde jedoch der Aufwand zur Berechnung aller Eigenwerte einer voll besetzten Matrix $\boldsymbol{A}$ bei Berücksichtigung der Dimensionsreduktion $\sum_{j=1}^{n} 4j^3/3 \sim n^4/3$ opms betragen, also um eine n-Potenz höher sein als beim Jacobi-Verfahren, vgl. 13.2. Die Kosten für einen $\boldsymbol{QR}$-Schritt können jedoch signifikant gesenkt werden, wenn $\boldsymbol{A}$ vor Ausführung des $\boldsymbol{QR}$-Algorithmus nach den in 13.5

beschriebenen Methoden orthogonal ähnlich auf symmetrische Tridiagonalform

$$\boldsymbol{T} = \operatorname{trid}(a_1, \ldots, a_n, b_2, \ldots, b_n) = \boldsymbol{Q}^{\mathsf{T}}\boldsymbol{A}\boldsymbol{Q} \tag{45}$$

transformiert wird. Nach 13.5.1 ist dies mittels Householder-Spiegelungen mit $\sim 2n^3/3$ opms möglich, kostet also nur die Hälfte eines einzelnen $\boldsymbol{QR}$-Schrittes für voll besetztes $\boldsymbol{A}$. Anschließend wird der $\boldsymbol{QR}$-Algorithmus 13.7.6 mit $\boldsymbol{A}_1 := \boldsymbol{T}$ durchgeführt.

Es zeigt sich, daß mit der Startmatrix $\boldsymbol{A}_1$ auch alle folgenden $\boldsymbol{QR}$-Iterierten $\boldsymbol{A}_k$ tridiagonal sind: Es sei $\boldsymbol{A}_k$ tridiagonal. Zur Berechnung der $\boldsymbol{QR}$-Faktorisierung

$$\boldsymbol{A}_k - \mu_k \boldsymbol{I} = \boldsymbol{Q}_k \boldsymbol{R}_k \tag{46}$$

bietet es sich an, das in der j-ten Spalte stehende Subdiagonalelement mittels einer Givens-Drehung $\boldsymbol{G}_{j,j+1} = \boldsymbol{G}^{(k)}_{j,j+1}$ zu annullieren, was auf

$$\boldsymbol{G}_{n-1,n}\Big(\cdots\Big(\boldsymbol{G}_{23}\big(\boldsymbol{G}_{12}(\boldsymbol{A}_k - \mu_k\boldsymbol{I})\big)\Big)\cdots\Big) = \boldsymbol{R}_k \tag{47}$$

führt, vgl. 10.2. Dann ist

$$\boldsymbol{Q}_k = \boldsymbol{G}_{12}^{\mathsf{T}}\boldsymbol{G}_{23}^{\mathsf{T}} \cdots \boldsymbol{G}_{n-1,n}^{\mathsf{T}}, \tag{48}$$

und wegen der Tridiagonalform von $\boldsymbol{A}_k - \mu_k\boldsymbol{I}$ hat $\boldsymbol{R}_k$ die Gestalt

$$\boldsymbol{R}_k = \begin{pmatrix} \times & \times & \times & & & \\ & \times & \times & \times & & \\ & & \ddots & \ddots & \ddots & \\ & & & \times & \times & \times \\ & & & & \times & \times \\ & & & & & \times \end{pmatrix},$$

ist also eine obere Dreiecksmatrix der Bandbreite 3 vom Typ $\{0, 2\}$; zur Bezeichnung siehe 6.4.A. Bei Bildung von

$$\boldsymbol{A}_{k+1} - \mu_k\boldsymbol{I} = \boldsymbol{R}_k\boldsymbol{Q}_k = \Big(\cdots\big((\boldsymbol{R}_k\boldsymbol{G}_{12}^{\mathsf{T}})\,\boldsymbol{G}_{23}^{\mathsf{T}}\big)\cdots\Big)\,\boldsymbol{G}_{n-1,n}^{\mathsf{T}} \tag{49}$$

werden nacheinander die Spalten j und $j+1$ kombiniert. Die Matrix $\boldsymbol{R}_k\boldsymbol{Q}_k$ ist daher eine obere Hessenberg-Matrix, wegen der Symmetrie also notwendig tridiagonal.

13.7.12. *Bemerkung.* (i) Wir haben gezeigt, daß die Tridiagonalform von $\boldsymbol{A}_k$ unter der gemäß (46) bis (49) realisierten $\boldsymbol{QR}$-Transformation

$$\boldsymbol{A}_k \to \boldsymbol{A}_{k+1} = \boldsymbol{Q}_k^{\mathsf{T}}\boldsymbol{A}_k\boldsymbol{Q}_k = \boldsymbol{R}_k\boldsymbol{Q}_k + \mu_k\boldsymbol{I}$$

invariant ist. Unter der zusätzlichen Voraussetzung (40) gilt dies auch für jede andere Realisierung des $\boldsymbol{QR}$-Schrittes, denn die transformierten Matrizen $\boldsymbol{A}_{k+1}$ können sich dann höchstens im Vorzeichen der Elemente unterscheiden, vgl. 13.7.3 und 13.7.7.

(ii) Wenn $\boldsymbol{A}_k$ nicht zerfällt, also

$$a_{i,i-1}^{(k)} = b_i^{(k)} \neq 0 \qquad (i = 2, \ldots, n) \tag{50}$$

gilt, und (40) nicht erfüllt, also μ_k ein Eigenwert von $\boldsymbol{A}$ ist, hat die letzte Zeile von $\boldsymbol{A}_{k+1}$ bei Rechnung in exakter Arithmetik die Gestalt $(0, \ldots, 0, a_{nn}^{(k+1)})$, siehe Ü 13.7.3. Es ist also $a_{n,n-1}^{(k-1)} = 0$, und $a_{nn}^{(k+1)} = \mu_k$ wird als Eigenwert erkannt und kann abgespalten werden. Die Voraussetzung (50) stellt dabei keine Einschränkung dar, denn andernfalls hätte bereits im vorhergehenden Schritt zu nichtzerfallenden Teilproblemen niedrigerer Dimension übergegangen werden können, vgl. 13.7.B.

(iii) Bei Verwendung der RQ-Verschiebungen besagt 13.7.10, daß $|a_{n.n-1}^{(k)}| = |b_n^{(k)}|$ in exakter Arithmetik fast immer monoton und asymptotisch kubisch gegen 0 geht. Bei Verwendung der Wilkinson-Verschiebungen kann sogar gezeigt werden, daß $b_n^{(k)}$ *immer* gegen 0 konvergiert, d. h., es liegt *globale Konvergenz* vor. Die Konvergenz ist dabei mindestens quadratisch im Sinne der Gültigkeit von

$$|b_n^{(k+1)}| \leqq C_1 \, |b_n^{(k)}|^2 \quad \text{mit} \quad C_1 > 0$$

und meistens sogar besser als kubisch. Für eine Formulierung des letztgenannten Sachverhaltes und die Beweise muß auf die Spezialliteratur verwiesen werden, siehe B 13.9. □

Für die numerische Realisierung ist wesentlich, daß die ersten j Spalten der in (47) auftretenden intermediären Matrix

$$\boldsymbol{G}_{j,j+1}\Big(\boldsymbol{G}_{j-1,j}\Big(\cdots\big(\boldsymbol{G}_{12}(\boldsymbol{A}_k - \mu_k \boldsymbol{I})\big)\Big)\Big) =: \boldsymbol{A}_k^{(j+1)}$$

bei den nachfolgenden Linksmultiplikationen mit $\boldsymbol{G}_{j+1,j+2}$ usw. weder geändert noch benutzt werden. Sie stellen daher bereits die entsprechenden Spalten von $\boldsymbol{R}_k$ dar, und die auf diese Spalten gemäß (49) von rechts angewendeten Drehungen $\boldsymbol{G}_{12}^{\mathsf{T}}, \ldots, \boldsymbol{G}_{j-1,j}^{\mathsf{T}}$ können bereits ausgeführt werden, ohne daß sie die noch benötigten Spalten $j+1, \ldots, n$ von $\boldsymbol{A}_k^{(j+1)}$ beeinflussen. Die Matrix $\boldsymbol{R}_k$ braucht als Ganzes also überhaupt nicht gebildet und gespeichert zu werden, sondern die Transformation kann direkt in der durch die Klammern gekennzeichneten Reihenfolge gemäß

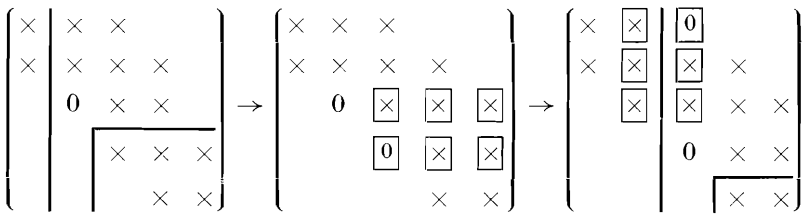
$$\boldsymbol{A}_{k+1} - \mu_k \boldsymbol{I} = \boldsymbol{R}_k \boldsymbol{Q}_k = \cdots \Big\{\boldsymbol{G}_{45}\Big[\big(\boldsymbol{G}_{34}\big\{\big[\boldsymbol{G}_{23}\boldsymbol{G}_{12}(\boldsymbol{A}_k - \mu_k \boldsymbol{I})\big]\, \boldsymbol{G}_{12}^{\mathsf{T}}\big\}\big)\, \boldsymbol{G}_{23}^{\mathsf{T}}\Big]\Big\}\, \boldsymbol{G}_{34}^{\mathsf{T}} \cdots \tag{51}$$

realisiert werden. Da hier die Verschiebungen explizit ausgeführt werden, spricht man auch vom ***QR***-Algorithmus mit explizit realisierten Verschiebungen oder kurz vom *expliziten* ***QR****-Algorithmus*. Um unnötige Indizierungen zu vermeiden, geben wir einen Schritt bei in-situ-Realisierung auf dem Platz von $\boldsymbol{A}$ an, wobei außerdem $\boldsymbol{V}_k$ auf dem Platz von $\boldsymbol{V}$ entsprechend (7) aufdatiert wird.

13.7.13. *Expliziter* ***QR****-Schritt.*

Setze $\boldsymbol{A} := \boldsymbol{A}_k - \mu_k \boldsymbol{I}$, $\boldsymbol{V} := \boldsymbol{V}_k$, $\boldsymbol{G}_{01} := \boldsymbol{I}$
for $j := 1(1)n - 1$ **do**
  Lege $\boldsymbol{G}_{j,j+1}$ so fest, daß $(\boldsymbol{G}_{j,j+1}\boldsymbol{A})_{j+1,j} = 0$ gilt
  Bilde $\boldsymbol{A} := \boldsymbol{G}_{j,j+1}\boldsymbol{A}$, $\boldsymbol{A} := \boldsymbol{A}\boldsymbol{G}_{j-1,j}^{\mathsf{T}}$, $\boldsymbol{V} := \boldsymbol{V}\boldsymbol{G}_{j,j+1}^{\mathsf{T}}$,
Setze $\boldsymbol{A} := \boldsymbol{A}\boldsymbol{G}_{n-1,n}^{\mathsf{T}}$, $\boldsymbol{A}_{k+1} := \boldsymbol{A} + \mu_k \boldsymbol{I}$, $\boldsymbol{V}_{k+1} := \boldsymbol{V}$

Der Teilschritt $j \to j+1$ wird für $n = 5$ und $j = 3$ durch das folgende Muster illustriert, wobei sich verändernde Elemente wie üblich eingerahmt worden sind.



$\boldsymbol{A}$ $\qquad$ $\boldsymbol{A} := \boldsymbol{G}_{34}\boldsymbol{A}$ $\qquad$ $\boldsymbol{A} := \boldsymbol{A}\boldsymbol{G}_{23}^{\mathsf{T}}$

Die durch die stark gezogenen Linien links bzw. rechts unten abgegrenzten Elemente sind bereits die von $\boldsymbol{R}_k\boldsymbol{Q}_k$ bzw. noch die von $\boldsymbol{A}_k - \mu_k\boldsymbol{I}$.

13.7.14. *Bemerkung.* (i) Unter Beachtung der Symmetrie läßt sich 13.7.13 bei expliziter Realisierung der Givens-Drehungen mit $\sim n(K_1 \text{ ops} + K_2 \text{ opm} + 1 \text{ opr})$ ausführen. Die genauen Werte der K_i hängen von den Details der Implementierung ab; für die bei WILKINSON/REINSCH [71] enthaltene Version ist $K_1 = 4$ und $K_2 = 13$. Die Berechnung von $\boldsymbol{V}_{k+1}$ erfordert dagegen bei Realisierung gemäß Ü 3.3.6 weitere $\sim 3n^2$ opms, also wesentlich mehr. Unter Verwendung einiger Hilfsspeicher kann die Transformation $\boldsymbol{A}_k \to \boldsymbol{A}_{k+1}$ auf den $\sim 2n$ Plätzen eines Dreiecks von $\boldsymbol{T}$ realisiert werden, und $\boldsymbol{V}_k$ kann offensichtlich mit $\boldsymbol{V}_{k+1}$ überspeichert werden.

(ii) Beim expliziten QR-Algorithmus ist es zweckmäßig, in Abänderung von 13.7.6 sofort

$$\boldsymbol{A}_{k+1} = \boldsymbol{R}_k\boldsymbol{Q}_k = \boldsymbol{Q}_k^{\mathsf{T}}(\boldsymbol{A}_k - \mu_k\boldsymbol{I})\,\boldsymbol{Q}_k$$

zu setzen, also die Addition von $\mu_k\boldsymbol{I}$ bei der Bildung von $\boldsymbol{A}_{k+1}$ wegzulassen. Für die derart definierte Folge gilt

$$\boldsymbol{A}_{k+1} = \boldsymbol{V}_{k+1}^{\mathsf{T}}[\boldsymbol{A} - (\mu_1 + \cdots + \mu_k)\,\boldsymbol{I}]\,\boldsymbol{V}_{k+1},$$

d. h., $\boldsymbol{A}_{k+1} + \sigma_{k+1}\boldsymbol{I}$ mit $\sigma_{k+1} := \sum_{j=1}^{k} \mu_j$ ist orthogonal ähnlich zu $\boldsymbol{A}$. Die Verschiebungen müssen dann gemäß $\sigma_1 := 0$, $\sigma_{k+1} := \sigma_k + \mu_k$ $(k = 1, 2, \ldots)$ gesondert akkumuliert werden. □

Eine zu 13.7.13 alternative Realisierung des QR-Schrittes beruht auf der Tatsache, daß $\boldsymbol{Q}_k$ und damit $\boldsymbol{A}_{k+1}$ und $\boldsymbol{V}_{k+1}$ im wesentlichen eindeutig durch $\boldsymbol{A}_k$ festgelegt sind, wenn die erste Spalte von $\boldsymbol{Q}_k$ vorgegeben wird und $\boldsymbol{A}_k$ nicht zerfällt, vgl. 13.5.8. Für die durch (48) definierte Matrix $\boldsymbol{Q}_k$ gilt nun

$$\boldsymbol{Q}_k\boldsymbol{e}^1 = \boldsymbol{G}_{12}^{\mathsf{T}}\boldsymbol{G}_{23}^{\mathsf{T}} \cdots \boldsymbol{G}_{n-1,n}^{\mathsf{T}}\boldsymbol{e}^1 = \boldsymbol{G}_{12}^{\mathsf{T}}\boldsymbol{e}^1, \tag{52}$$

d. h., die erste Spalte von $\boldsymbol{Q}_k$ ist gleich der ersten Spalte von $\boldsymbol{G}_{12}^{\mathsf{T}}$ und damit durch die erste Spalte von $\boldsymbol{A}_k - \mu_k\boldsymbol{I}$ festgelegt, denn $\boldsymbol{G}_{12}$ transformiert letztere in ein Vielfaches von $\boldsymbol{e}^1$. Jede Matrix

$$\bar{\boldsymbol{Q}}_k := \boldsymbol{G}_{12}^{\mathsf{T}}\bar{\boldsymbol{G}}_{23}^{\mathsf{T}}\bar{\boldsymbol{G}}_{34}^{\mathsf{T}} \cdots \bar{\boldsymbol{G}}_{n-1,n}^{\mathsf{T}} \tag{53}$$

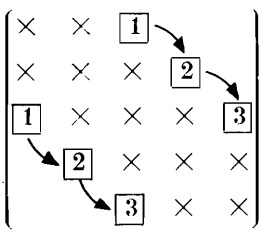
besitzt folglich für beliebige $\bar{G}_{23}, \ldots, \bar{G}_{n-1,n}$ dieselbe erste Spalte wie $\boldsymbol{Q}_k$. Wenn es gelingt, die Drehungen $\bar{G}_{23}, \ldots, \bar{G}_{n-1,n}$ so festzulegen, daß

$$\bar{\boldsymbol{A}}_{k+1} := \bar{\boldsymbol{Q}}_k^{\mathsf{T}} \boldsymbol{A}_k \bar{\boldsymbol{Q}}_k = \bar{\boldsymbol{G}}_{n-1,n} \cdots \bar{\boldsymbol{G}}_{23} \boldsymbol{G}_{12} \boldsymbol{A}_k \boldsymbol{G}_{12}^{\mathsf{T}} \bar{\boldsymbol{G}}_{23}^{\mathsf{T}} \cdots \bar{\boldsymbol{G}}_{n-1,n}^{\mathsf{T}} \tag{54}$$

tridiagonal ist, muß $\{\bar{\boldsymbol{A}}_{k+1}, \bar{\boldsymbol{Q}}_k\}$ im Sinne von 13.5.8 im wesentlichen mit $\{\boldsymbol{A}_{k+1}, \boldsymbol{Q}_k\}$ übereinstimmen. Nun ist $\boldsymbol{F}_2 := \boldsymbol{G}_{12} \boldsymbol{A}_k \boldsymbol{G}_{12}^{\mathsf{T}}$ von der Form

$$\boldsymbol{F}_2 = \begin{pmatrix} \times & \times & \boxed{\times} & & & \\ \times & \times & \times & & & \\ \boxed{\times} & \times & \times & \times & & \\ & & \times & \times & \times & \\ & & & \ddots & \ddots & \ddots \end{pmatrix},$$

weicht also nur in den Positionen (3, 1) und (1, 3) von der Tridiagonalform ab. Wenn $\bar{\boldsymbol{G}}_{23}$ so festgelegt wird, daß $(\bar{\boldsymbol{G}}_{23} \boldsymbol{F}_2)_{31} = 0$ gilt, hat $\boldsymbol{F}_3 := \bar{\boldsymbol{G}}_{23} \boldsymbol{F} \bar{\boldsymbol{G}}_{23}^{\mathsf{T}}$ in diesen Positionen Nullen, allerdings sind in $\{4, 2\}$ und $\{2, 4\}$ neue Nichtnullelemente erzeugt worden. Diese werden durch Übergang zu $\boldsymbol{F}_4 := \bar{\boldsymbol{G}}_{34} \boldsymbol{F}_3 \bar{\boldsymbol{G}}_{34}^{\mathsf{T}}$ in analoger Weise annulliert usw. Nach $n - 2$ Schritten ist die Tridiagonalform wieder hergestellt, vgl. 11.1 und 13.5.B für ein ähnliches Vorgehen. Die Reihenfolge, in der Nichtnullelemente außerhalb des tridiagonalen Bandes auftreten und wieder annulliert werden, wird für $n = 5$ durch das folgende Muster charakterisiert.



Da beim derart ausgeführten $\boldsymbol{QR}$-Schritt die Verschiebung μ_k nicht mehr explizit auftritt — sie wird lediglich bei der Festlegung der Drehungsparameter von $\boldsymbol{G}_{12}$ benötigt —, heißt diese Version auch $\boldsymbol{QR}$-Algorithmus mit impliziter Realisierung der Verschiebungen oder kurz *impliziter $\boldsymbol{QR}$-Algorithmus*. Eine 13.7.13 entsprechende Grobbeschreibung des Einzelschrittes lautet wie folgt:

13.7.15. *Impliziter $\boldsymbol{QR}$-Schritt.*
Setze $\boldsymbol{A} := \boldsymbol{A}_k$, $\boldsymbol{V} := \boldsymbol{V}_k$
Lege $\boldsymbol{G}_{12}$ so fest, daß

$$[\boldsymbol{G}_{12}(\boldsymbol{A} - \mu_k \boldsymbol{I})]_{21} = [\boldsymbol{G}_{12}(a_{11} - \mu_k, a_{21}, 0, \ldots, 0)^{\mathsf{T}}]_2 = 0$$

gilt. Bilde $\boldsymbol{A} := \boldsymbol{G}_{12} \boldsymbol{A} \boldsymbol{G}_{12}^{\mathsf{T}}$, $\boldsymbol{V} := \boldsymbol{V} \boldsymbol{G}_{12}^{\mathsf{T}}$
for $i := 3(1)n$ **do**
  Lege $\bar{\boldsymbol{G}}_{i-1,i}$ so fest, daß $(\bar{\boldsymbol{G}}_{i-1,i} \boldsymbol{A})_{i,i-2} = 0$ gilt
  Bilde $\boldsymbol{A} := \bar{\boldsymbol{G}}_{i-1,i} \boldsymbol{A} \bar{\boldsymbol{G}}_{i-1,i}^{\mathsf{T}}$, $\boldsymbol{V} := \boldsymbol{V} \bar{\boldsymbol{G}}_{i-1,i}^{\mathsf{T}}$
Setze $\boldsymbol{A}_{k+1} := \boldsymbol{A}$, $\boldsymbol{V}_{k+1} := \boldsymbol{V}$

13.7.16. *Bemerkung.* (i) Von der Implementierung her ist der implizite QR-Algorithmus etwas günstiger als der explizite. Der Aufwand ist nur unwesentlich geringer; für die bei WILKINSON/REINSCH [71] enthaltene Version werden $\sim n(6$ ops $+$ 11 opm $+$ 1 opr) zur Berechnung von A_{k+1} aus A_k benötigt. Die Bildung von V_{k+1} entspricht dem expliziten Verfahren und kostet $\sim 3n^2$ opms. In-situ-Realisierung ist bezüglich A_k und V_k möglich.

(ii) Der eigentliche Vorteil des impliziten QR-Algorithmus liegt darin, daß $A_k - \mu_k I$ nicht explizit gebildet und daher ein möglicher Informationsverlust bei der Berechnung von $a_{jj}^{(k)} - \mu_k$ bei betragsmäßig großem μ_k und kleinem $a_j^{(k)}$ vermieden wird. Selbst für ansteigend gestufte Matrizen liefert der implizite Algorithmus daher oft auch die betragskleinen Eigenwerte mit ausreichender Genauigkeit.

(iii) Besonders für den impliziten QR-Algorithmus ist wesentlich, daß nach jedem Schritt *alle* Subdiagonalelemente — nicht nur das letzte! — auf Kleinheit etwa im Sinne von

$$|b_i^{(k)}| \leqq \nu(|a_{i-1}^{(k)}| + |a_i^{(k)}|), \qquad i \in \{2, \ldots, n\}, \tag{55}$$

getestet und bei Erfülltsein von (55) gleich 0 gesetzt werden, vgl. 13.6.1. Wenn ein kleines $b_i^{(k)}$ mit $2 \leqq i \leqq n-1$ nicht zur Aufspaltung in kleinere Teilprobleme genutzt wird, erhöht sich nicht nur wegen der verschenkten Dimensionsreduktion der Aufwand für einen QR-Schritt, sondern das Konvergenzverhalten kann sich wesentlich verschlechtern: Mit $b_i^{(k)}$ wird meist auch $b_i^{(k+1)}$ klein sein. Dies bedeutet, daß die ersten Spalten $\{1, \ldots, i-1\}$ von $\bar{Q}_k$ — man beachte $A_k = \bar{Q}_k \bar{A}_{k+1} \bar{Q}_k^T$ — nur schwach mit den letzten Spalten $\{i, \ldots, n\}$ gekoppelt sind, was z. B. aus dem Beweis von 13.5.8 oder aus der Rekursionsformel des Lanczos-Algorithmus aus Ü 13.5.6 zu ersehen ist. Die in der ersten Spalte von $\bar{Q}_k$ enthaltene Information über die Verschiebung μ_k, die mittels der Drehungen $\bar{G}_{j-1,j}$ implizit auf die übrigen Spalten übertragen werden soll, geht dann bei der Durchführung des k-ten QR-Schrittes teilweise oder vollständig verloren; die Drehungen $\bar{G}_{i-1,i}, \ldots, \bar{G}_{n-1,n}$ werden deformiert, vgl. auch 13.5.4(ii) für ähnliche Überlegungen bei Verwendung von Householder-Spiegelungen. Die durch 13.5.8 in exakter Arithmetik garantierte Äquivalenz der explizit bzw. implizit berechneten Matrizen A_{k+1} bzw. $\bar{A}_{k+1}$ liegt bei Computerrechnung nicht mehr vor, was zum Verlust der schnellen Konvergenz führt.

(iv) Aus den in (iii) dargelegten Gründen ist es zweckmäßig, das in den EISPACK-Routinen verwendete, relativ einfache Kriterium (55) durch ein selektiveres zu ersetzen. Wir empfehlen die Verwendung von

$$|b_i| \leqq \nu(\beta_i + \gamma_i)(1 + \alpha_i/\beta_i), \qquad i \in \{2, \ldots, n\}, \tag{56}$$

mit

$$\alpha_i := |a_{i-1}| + |a_i|, \qquad \beta_i := |b_{i-1}| + |b_i| + |b_{i+1}|, \qquad \gamma_i := |a_{i-1} - a_i|;$$

der obere Index k wurde weggelassen. Da dieser Test aufwendig ist, sollte er nur ausgeführt werden, wenn beispielsweise $|b_i| \leqq \sqrt{\nu}\, \|T\|_\infty$ gilt. □

Nach einer nicht zu großen Zahl von Schritten wird (55) oder (56) für alle i erfüllt sein, d. h., A_{k+1} ist diagonal. Die Diagonalelemente approximieren dann die gesuchten Eigenwerte, und die Spalten von V_{k+1} sind zugehörige Eigenvektornäherungen zu T.

Mit $\boldsymbol{Q}$ aus (45) ergeben sich schließlich die Eigenvektornäherungen zu $\boldsymbol{A}$ als Spalten von $\boldsymbol{QV}_{k+1}$. Wenn die Eigenvektoren auch gesucht sind, sollte daher sofort $\boldsymbol{QV}_{k+1}$ durch Aufdatierung von $\boldsymbol{QV}_k$ mit $\boldsymbol{Q}_k$ ermittelt werden, d. h., 13.7.6 sollte mit $\boldsymbol{V}_1 := \boldsymbol{Q}$ statt $\boldsymbol{V}_1 := \boldsymbol{I}$ gestartet werden.

Wir geben im folgenden das Grobschema einer in-situ-Realisierung des $\boldsymbol{QR}$-Algorithmus für eine Tridiagonalmatrix $\boldsymbol{T}$ an, bei der die obigen Überlegungen berücksichtigt werden. Die Indizes p, q mit $1 \leqq p \leqq q \leqq n$ geben dabei die Position des gerade behandelten, untersten nichtzerfallenden Diagonalblockes von $\boldsymbol{T}$ an.

13.7.17. *$\boldsymbol{QR}$-Algorithmus für Tridiagonalmatrizen.* Gegeben seien die Tridiagonalmatrix $\boldsymbol{T} = \operatorname{trid}(a_1, \ldots, a_n, b_2, \ldots, b_n) \in \mathbf{S}^{n,n}$ und die orthogonale Matrix $\boldsymbol{V} \in \mathbf{R}^{n,n}$.

Algorithmus:

S0 (Initialisierung): Setze $k := 1$, $q := n$

S1 (Abbruchtest): **If** $q = 1$ **then stop**

S2 (Nullsetzen kleiner Subdiagonalelemente und Festlegung des untersten nichtzerfallenden Diagonalblockes $\boldsymbol{T}^{(p,q)} := \operatorname{trid}(a_p, \ldots, a_q, b_{p+1}, \ldots, b_q)$ der Dimension $m := q - p + 1$ von $\boldsymbol{T}$):

for $p := q(-1)2$ **do**

| **if** [$|b_p|$ klein im Sinne von (56)]
| **then** | $b_p := 0$
| | **if** $p = q$ **then** [$q := q - 1$, **goto** S1] **else goto** S3

$p := 1$

S3 (k-ter $\boldsymbol{QR}$-Schritt): Führe $\boldsymbol{QR}$-Schritt 13.7.13 bzw. 13.7.15 mit der Wilkinson-Verschiebung 13.7.11 für

$$\boldsymbol{A}_k := \boldsymbol{T}^{(p,q)} \in \mathbf{S}^{m,m}, \qquad \boldsymbol{V}_k := \boldsymbol{V}^{(p,q)} \in \mathbf{R}^{n,m}$$

auf dem Platz von $\boldsymbol{T}$ bzw. $\boldsymbol{V}$ durch, wobei $\boldsymbol{V}^{(p,q)}$ die aus den Spalten p bis q von $\boldsymbol{V}$ bestehende Teilmatrix bezeichnet

S4: Setze $k := k + 1$, **goto** S1

13.7.18. *Bemerkung.* (i) Umfangreiche numerische Experimente zeigen, daß die Eingangsmatrix $\boldsymbol{T}$ im Mittel nach $k \approx 1.7n$ $\boldsymbol{QR}$-Schritten diagonalisiert ist, d. h., bei Abbruch ist $\boldsymbol{T} = \boldsymbol{T}_{k+1} = \operatorname{diag}(a_j^{(k,1)}) =: \operatorname{diag}(\hat{\mu}_j) =: \boldsymbol{M}$. Für die zuerst berechneten Eigenwerte $\hat{\mu}_n, \hat{\mu}_{n-1}, \ldots$ werden dabei i. allg. etwas mehr — etwa 4 bis 6 — Schritte benötigt, die zuletzt berechneten Eigenwerte $\ldots, \hat{\mu}_3, \hat{\mu}_2, \hat{\mu}_1$ können meist schon nach einem, höchstens zwei Schritten abgespalten werden. Wenn pro Eigenwert zwei $\boldsymbol{QR}$-Schritte veranschlagt werden und ein Schritt der Dimension m ohne Aufdatierung von $\boldsymbol{V}$ mit $\sim m(10\,\text{opms} + 1\,\text{opr})$ gezählt wird, vgl. 13.7.14 und 13.7.16, kostet die Berechnung aller Eigenwerte von $\boldsymbol{T}$ bei dieser Modellvorstellung $\sim n^2(10\,\text{opms} + 1\,\text{opr})$, ist also extrem billig. Wegen der zusätzlich auftretenden Deflation außerhalb der letzten Zeile — in 13.7.17 durch $p > 1$ charakterisiert — ist der Aufwand praktisch sogar meist geringer. Die zugehörigen Eigenvektorapproximationen ergeben sich mit der Eingangsbelegung $\boldsymbol{V} := \boldsymbol{I}$ durch Aufdatierung von $\boldsymbol{V}$ mit maximal $\sim 3n^3$ opms.

(ii) Zur Berechnung aller Eigenwerte der voll besetzten Matrix A muß zunächst $T = Q^T AQ$ mit $\sim 2n^3/3$ opms nach 13.5.1 berechnet werden, so daß insgesamt $\sim n^2[(2n/3 + 10)$ opms $+ 1$ opr$] \sim 2n^3/3$ opms erforderlich sind. Wenn auch die Eigenvektoren gefordert werden, wird Q mit $\sim 2n^3/3$ opms explizit gebildet — siehe 13.5.2 — und in 13.7.17 als Eingangsbelegung $V := Q$ verwendet. Mit den zur Aufdatierung benötigten $\sim 3n^3$ opms kostet daher die Berechnung der aus den Eigenvektorapproximationen zu A gebildeten Ausgangsmatrix $V = QV_{k+1}$ maximal $\sim 11n^3/3$ opms, ist also wesentlich teurer als die Berechnung der Eigenwerte allein.

(iii) Sind nur wenige Eigenvektoren gesucht, so sollten die entsprechenden Eigenvektoren von T mittels inverser Iteration mit jeweils $\sim K_3 n$ opms berechnet werden, siehe 13.4.3. Unter Verwendung der Produktform von Q sind die Eigenvektoren v^j von T dann noch in Eigenvektoren Qv^j von A zu transformieren, was jeweils $\sim n^2$ opms kostet. Insbesondere bei mehrfachen und dicht benachbarten Eigenwerten ist dabei i. allg. eine Re-Orthogonalisierung erforderlich, vgl. 13.4.5 (iv).

(iv) Wenn A eine Bandmatrix ist, kann die Tridiagonalisierung nach 13.5.6 erfolgen. Eine alternative Möglichkeit besteht darin, den QR-Algorithmus direkt mit $A_1 := A$ durchzuführen, denn die Bandgestalt ist invariant unter der QR-Transformation. Bei impliziter Realisierung wird dabei wie in 13.5.B vorgegangen. Ist die Dimension n groß und sind nur wenige Eigenwerte und gegebenenfalls Eigenvektoren gesucht, so empfiehlt sich das Bisektionsverfahren bzw. die inverse Iteration direkt mit der Bandmatrix A, vgl. 13.6.11(vi) und 13.5.7(ii).

(v) Man kann zeigen, daß die nach 13.7.17 bei Abbruch nach k Schritten vorliegende, berechnete Diagonalmatrix $M = T_{k+1}$ der Gleichung

$$A + \delta A = VMV^T \quad \text{mit} \quad \delta A \in \mathbf{S}^{n,n}, \qquad \|\delta A\| \leqq \nu F \|A\|, \tag{57}$$

und akzeptablem F genügt, wobei V eine gewisse exakt orthogonale Matrix ist. *Die Berechnung der Eigenwerte von A nach dem QR-Algorithmus ist also ein numerisch gutartiger Prozeß.* Das exakt orthogonale V wird durch die mit $V = V_1 := Q$, Q aus 13.5.1, in 13.7.17 berechnete Matrix $\tilde{V} = V_{k+1}$ im Sinne von

$$\|V - \tilde{V}\| \leqq \nu F_1, \qquad F_1 \text{ akzeptabel}, \tag{58}$$

ausreichend genau approximiert. *Die berechneten Eigenvektornäherungen sind daher ausreichend orthogonal und akzeptabel im Sinne kleiner Residuen.*

(vi) Wenn T eine ansteigend gestufte Matrix ist, werden die Verschiebungen am Anfang groß sein. Die im betragskleinen oberen Teil enthaltene Information über die kleinen Eigenwerte wird daher — beim expliziten QR-Algorithmus in katastrophaler Weise, beim impliziten weniger stark — verlorengehen. Um dies zu vermeiden, kann der QR-Algorithmus in inverser Reihenfolge von rechts nach links durchgeführt und die Verschiebung durch die links oben stehende (2,2)-Matrix festgelegt, also die QR-Faktorisierung von $A_k - \mu_k I$ durch die analog gebildete QL-Faktorisierung $A_k - \mu_k I = Q_k L_k$ mit einer unteren Dreiecksmatrix L_k ersetzt werden. Diese Version wird *QL-Algorithmus* genannt und heute in den Programmpaketen vorwiegend verwendet, und zwar in impliziter Realisierung als *impliziter QL-Algo-*

rithmus. Wenn $\boldsymbol{A}$ bereits ansteigend gestuft ist, sollte die Tridiagonalisierung ebenfalls in inverser Reihenfolge von rechts nach links erfolgen, siehe 13.5.4(iii) und Ü 13.5.2. □

Übungsaufgaben

Ü 13.7.1. Man zeige, daß für die Grundform 13.7.1 des $\boldsymbol{QR}$-Algorithmus $\boldsymbol{R}_k = \boldsymbol{V}_{k+1}^{\mathsf{T}} \boldsymbol{A} \boldsymbol{V}_k$ gilt, und folgere hieraus die Gültigkeit von

$$\boldsymbol{V}_{k+1}(\boldsymbol{R}_k \boldsymbol{R}_{k-1} \cdots \boldsymbol{R}_2 \boldsymbol{R}_1) = \boldsymbol{A}^k.$$

Ü 13.7.2. Gegeben sei die Matrix $\boldsymbol{P} = \begin{pmatrix} \alpha & \beta \\ \beta & \gamma \end{pmatrix}$ mit $\beta \neq 0$, und es sei $\delta := \dfrac{\gamma - \alpha}{2\beta}$. Man zeige:

(i) Ein Schritt des Jacobi-Verfahrens aus 13.2 liefert bei Anwendung auf $\boldsymbol{P}$ die Eigenwerte $\mu' = \gamma + \beta t$, $\mu'' = \gamma - \beta/t$, wobei

$$t := \operatorname{sgn}(\delta)/(|\delta| + \sqrt{1 + \delta^2}) \quad \text{im Fall} \quad \delta \neq 0,$$

$$t := 1 \quad \text{im Fall} \quad \delta = 0$$

ist.

(ii) Es gilt

$$|\mu' - \gamma| < |\mu'' - \gamma| \quad \text{im Fall} \quad \delta \neq 0,$$

$$|\mu' - \gamma| = |\mu'' - \gamma| \quad \text{im Fall} \quad \delta = 0,$$

so daß

$$\mu := \begin{cases} \gamma + \operatorname{sgn}(\delta)\, \beta/(|\delta| + \sqrt{1 + \delta^2}) & \text{für} \quad \delta \neq 0, \\ \gamma - \operatorname{sgn}(\gamma)\, |\beta| & \text{für} \quad \delta = 0, \gamma \neq 0, \\ \beta & \text{für} \quad \delta = 0, \gamma = 0 \end{cases}$$

die zu $\boldsymbol{P}$ gehörende Wilkinson-Verschiebung gemäß 13.7.11 ist.

Ü 13.7.3. Es sei $\boldsymbol{T} - \mu \boldsymbol{I} = \boldsymbol{Q}\boldsymbol{R}$ eine $\boldsymbol{QR}$-Faktorisierung von $\boldsymbol{T} - \mu \boldsymbol{I}$, und $\boldsymbol{T}$ sei tridiagonal und nichtzerfallend. Man zeige, daß dann $r_{jj} \neq 0$ $(j = 1, \ldots, n-1)$ gilt und daß $\boldsymbol{T} - \mu \boldsymbol{I}$ genau dann singulär ist, wenn $r_{nn} = 0$ gilt. Die letzte Bedingung ist gleichwertig dazu, daß die n-te Zeile von $\boldsymbol{R}\boldsymbol{Q}$ verschwindet, also $\boldsymbol{R}\boldsymbol{Q} + \mu \boldsymbol{I}$ die n-te Zeile $(0, \ldots, 0, \mu)$ hat.

Bemerkungen zum Kapitel 13

B 13.1. Die im Abschnitt 13.1 zusammengestellten Ergebnisse gehören zum Standardwissen der Eigenwerttheorie, siehe etwa WILKINSON [65] und PARLETT [80a], wo auch historische Kommentare zu finden sind. Der Rayleigh-Quotient wurde durch Lord RAYLEIGH [1899] eingeführt.

B 13.2. Das klassische Jacobi-Verfahren geht auf JACOBI [1846] zurück. Es wurde 100 Jahre später zu Beginn des Computer-Zeitalters durch BARGMANN/MONTGOMERY/VON NEUMANN [1946] wiederentdeckt; vgl. auch GOLDSTINE/MURRAY/VON NEUMANN [59]. Schwellwertstrategien sind von POPE/TOMPKINS [57], RUTISHAUSER [66] u. a. vorgeschlagen und analysiert worden. Zum Nachweis der asymptotisch quadratischen Konvergenz des zyklischen Jacobi-Verfahrens sei auf SCHÖNHAGE [64] und WILKINSON [62] verwiesen. Eine ausgefeilte

Implementierung hat RUTISHAUSER [66/71] publiziert; dort sind alle im Text beschriebenen Tricks zur Verringerung des Rundungsfehlereinflusses zu finden. Die zweite Formel (13.2.23), mit der sich eine Multiplikation einsparen läßt, scheint nicht allgemein bekannt zu sein. Zur Fehleranalyse siehe WILKINSON [65]; die dort angegebene Schranke (13.2.36) für F bezieht sich auf eine zyklische Realisierung nach den Formeln (13.2.10) und (13.2.11) und ist daher i. allg. zu pessimistisch.

B 13.3. Die einfache Vektoriteration wird seit Jahrzehnten in jedem Lehrbuch der Numerischen Mathematik beschrieben; der Name von-Mises-Iteration geht auf die Arbeit von MISES/POLLACZEK-GEIRINGER [29] zurück. Die bei der Vektoriteration erzeugte Vektorfolge $\{\boldsymbol{v}, \boldsymbol{A}\boldsymbol{v}, \boldsymbol{A}^2\boldsymbol{v}, \ldots\}$ bildet die Grundlage zur Betrachtung der sog. Krylov-Teilräume $\mathcal{K}_p := \operatorname{span}\{\boldsymbol{v}, \boldsymbol{A}\boldsymbol{v}, \ldots, \boldsymbol{A}^{p-1}\boldsymbol{v}\}$, mit deren Hilfe ebenfalls die dominanten Eigenwerte und zugehörigen Eigenvektoren von $\boldsymbol{A}$ approximiert werden können, z. B. nach den verschiedenen Varianten des sog. Lanczos-Algorithmus, siehe PARLETT [80a] für Details. Diese Technik ist allerdings nur für große schwach besetzte Matrizen zweckmäßig und wird deshalb hier nicht untersucht.

B 13.4. Wesentliche Arbeiten zur Teilraumiteration gehen auf BAUER [57] — dort simultane Iteration genannt — und RUTISHAUSER [69, 70/71] zurück, siehe auch STEWART [69, 75]. Auf Grund der bei der Teilraumiteration vorgenommenen Orthogonalisierung der Basis besteht ein enger Zusammenhang zum QR-Algorithmus, siehe 13.7.2. Die zur Teilraumapproximation benötigten Begriffe sind bei DAVIS/KAHAN [70], STEWART [73a] und PARLETT [80a] zu finden. Der Rayleigh-Ritz-Algorithmus zur Beschaffung von Eigenwert/Eigenvektor-Näherungen aus approximierenden Teilräumen ist eine numerische Standardtechnik; die Ursprünge liegen bei RAYLEIGH [1899] und RITZ [1909]. Eine sehr ausgefeilte Prozedur ritzit zur Kombination von Teilraumiteration und RR-Algorithmus, in die noch weitere Verbesserungen aufgenommen worden sind, hat RUTISHAUSER [70] angegeben. Die dabei benutzte Version des RR-Algorithmus nach 13.3.17(vi) ist von REINSCH vorgeschlagen worden; zur Analyse siehe wieder PARLETT [80a]. Alternative Varianten sind für allgemeinere Aufgabenklassen bei MCCORMICK/NOE [77] zu finden.

B 13.5. Spezielle Verfahren der Teilraumiteration beruhen auf einer Maximierung des Funktionals $\operatorname{tr}(\boldsymbol{P}) = \operatorname{tr}(\boldsymbol{Q}^T\boldsymbol{A}\boldsymbol{Q})$ über alle $\boldsymbol{Q} \in \mathbf{R}^{n,p}$ mit $\boldsymbol{Q}^T\boldsymbol{Q} = \boldsymbol{I}$. Hierzu gehören die Verfahren der simultanen Koordinatenrelaxation — siehe etwa SCHWARZ [77] —, simultane Gradientenverfahren und simultane Verfahren der konjugierten Gradienten — siehe GERADIN [71], LONGSINE/MCCORMICK [80] und DÖHLER [82]. Diese Verfahren sind in der Regel für das allgemeine symmetrische Eigenwertproblem $\boldsymbol{A}\boldsymbol{x} = \lambda\boldsymbol{B}\boldsymbol{x}$ entwickelt und formuliert worden, dazu siehe Kapitel 14. Eine zusammenfassende Darstellung gibt MEYER [87].

B 13.6. Die inverse Iteration geht auf WIELANDT [44] zurück. Modifikationen, bei denen neben $\boldsymbol{v}$ auch μ verbessert wird, haben UNGER [50] und viele andere vorgeschlagen; diese Verfahren hängen eng mit dem Newton-Verfahren zusammen, siehe etwa PETERS/WILKINSON [79]. Die sogar kubisch konvergente Rayleigh-Quotienten-Iteration geht nicht etwa auf Lord RAYLEIGH, sondern auf OSTROWSKI [58, 59] zurück. Das monotone Fallen der Residuen bemerkte KAHAN, siehe PARLETT [74, 80a]. Der Beweis der globalen Konvergenz ist von PARLETT/KAHAN [69] geführt worden, siehe wieder PARLETT [80a]. Zwischen der RQ-Iteration und dem QR-Algorithmus mit speziellen Verschiebungen besteht ein enger Zusammenhang, siehe 13.7.9. Während noch vor 25 Jahren angenommen wurde, daß die Fastsingularität von $\boldsymbol{A} - \mu\boldsymbol{I}$ sich nachteilig auf die inverse Iteration auswirkt, hat später WILKINSON in verschiedenen Arbeiten gezeigt, daß sie sich sogar sehr vorteilhaft auf die berechneten Eigenvektornäherungen auswirkt. Eine detaillierte Rundungsfehleranalyse ist bei WILKINSON [77] zu finden.

B 13.7. Die Transformation einer symmetrischen Matrix mittels Givens-Drehungen auf Tridiagonalform wurde erstmals von GIVENS [54] beschrieben. HOUSEHOLDER/BAUER [59] verwenden später für dieselbe Aufgabe Householder-Spiegelungen. Daß $\boldsymbol{Q}$ und $\boldsymbol{T}$ empfindlich von $\boldsymbol{A}$ und der zur Berechnung genutzten Arithmetik abhängen, zeigen instruktive Bei-

spiele bei WILKINSON [65] und PARLETT [80a]. Zur Anwendung impliziter Givens-Drehungen sei auf RATH [82] verwiesen. Eine alternative, besonders für große schwach besetzte Matrizen geeignete Tridiagonalisierungsmethode stellt der sog. Lanczos-Algorithmus dar, siehe Ü 13.5.6, der auf LANCZOS [50] zurückgeht und detailliert von PAIGE [76], PARLETT [80a], GOLUB/VAN LOAN [83] und CULLUM/WILLOUGHBY [85] untersucht wird.

B 13.8. Das Bisektionsverfahren mit Berechnung von $s(\mu)$ über die Folge $\{\tau_0, \ldots, \tau_n\}$ gemäß Ü 13.6.2 wurde erstmals von GIVENS [53, 54] vorgeschlagen. Da die Hauptabschnittsdeterminanten $\tau_k = \tau_k(\mu)$ eine sog. Sturmsche Kette bilden, spricht man auch von der *Methode der Sturmschen Ketten*. Wesentliche Verbesserungen und eine Fehleranalyse gehen auf WILKINSON [65] zurück; dort ist auch die Matrix (13.6.5) zu finden. Bereits bei der Implementierung in WILKINSON/REINSCH [71] wird Über- und Unterlauf durch Übergang zu $d_k = \tau_k/\tau_{k-1}$ vermieden, vgl. Ü 13.6.2. Die natürlichere Motivierung ohne Sturmsche Ketten mittels des Trägheitsgesetzes für quadratische Formen kongruenter Matrizen wird von verschiedenen Autoren bevorzugt, z. B. von PARLETT [80a].

B 13.9. Der $\boldsymbol{QR}$-Algorithmus ist das Analogon des auf RUTISHAUSER [58] zurückgehenden sog. $\boldsymbol{LR}$-Algorithmus und wurde unabhängig und etwa gleichzeitig von FRANCIS [61] und KUBLANOVSKAJA [61] gefunden. Ausführungen zur historischen Entwicklung können bei PARLETT [64] nachgelesen werden. Bei der Motivierung der Grundform des $\boldsymbol{QR}$-Algorithmus über die Teilraumiteration folgen wir methodisch der Darstellung bei WATKINS [82]. Der Zusammenhang zwischen dem $\boldsymbol{QR}$-Algorithmus mit den Verschiebungen $\mu_k = a_{nn}^{(k)}$ und der RQ-Iteration wurde von WILKINSON und PARLETT bemerkt, siehe PARLETT [80a]. Die globale Konvergenz des tridiagonalen $\boldsymbol{QR}$-Algorithmus mit den nach ihm benannten Verschiebungen hat WILKINSON [68] bewiesen, siehe auch HOFFMAN/PARLETT [78] und PARLETT [80a], wo auch Aussagen über die in der Regel schnellere als kubische Konvergenz zu finden sind. Alternative Verschiebungen haben z. B. REINSCH/BAUER [71] verwendet. Bei PARLETT [80a] wird auch auf die Vermeidung von Quadratwurzeln eingegangen, siehe dazu ebenfalls RATH [82], wo implizite Givens-Drehungen herangezogen werden. Das Kriterium (13.7.56) für das Nullsetzen von Subdiagonalelementen ist eine leichte Verbesserung eines auf KAHAN zurückgehenden Kriteriums, das bei PARLETT [80a] angegeben und diskutiert wird.

14. Das allgemeine symmetrische Eigenwertproblem

14.1. Grundlegende Eigenschaften

Das allgemeine symmetrische Eigenwertproblem lautet wie folgt: Für gegebene Matrizen $\boldsymbol{A}, \boldsymbol{B} \in \mathbf{S}^{n,n}$ sind Zahlen $\lambda_j = \lambda_j[\boldsymbol{A}, \boldsymbol{B}]$, für die

$$\boldsymbol{A}\boldsymbol{z}^j = \lambda_j \boldsymbol{B}\boldsymbol{z}^j \quad \text{mit} \quad \boldsymbol{z}^j \neq \boldsymbol{o} \tag{1}$$

gilt, und gegebenenfalls die Vektoren $\boldsymbol{z}^j$ zu bestimmen. Wie beim speziellen symmetrischen Eigenwertproblem, das sich aus (1) für $\boldsymbol{B} = \boldsymbol{I}$ ergibt, werden die λ_j *Eigenwerte*, die $\boldsymbol{z}^j$ zugehörige *Eigenvektoren* und die Paare $\{\lambda_j, \boldsymbol{z}^j\}$ *Eigenpaare* von $\{\boldsymbol{A}, \boldsymbol{B}\}$ genannt. Die Matrizenschar $\boldsymbol{A} - \lambda \boldsymbol{B}$ heißt das durch $\{\boldsymbol{A}, \boldsymbol{B}\}$ erzeugte *Büschel* (engl. „pencil", russ. „пучок"); häufig wird das Paar $\{\boldsymbol{A}, \boldsymbol{B}\}$ selbst Büschel genannt. Man beachte, daß $\{\lambda, \boldsymbol{z}\}$ genau dann ein Eigenpaar von $\{\boldsymbol{A}, \boldsymbol{B}\}$ ist, wenn

$$(\boldsymbol{A} - \lambda \boldsymbol{B})\,\boldsymbol{z} = \boldsymbol{o} \tag{2}$$

gilt, also

$$p(\lambda) := \det(\boldsymbol{A} - \lambda \boldsymbol{B}) = 0 \quad \text{und} \quad \boldsymbol{z} \in \mathcal{N}(\boldsymbol{A} - \lambda \boldsymbol{B}), \qquad \boldsymbol{z} \neq \boldsymbol{o}, \tag{3}$$

ist. Die Funktion $p(\lambda)$ ist ein Polynom vom Höchstgrad n, das *charakteristische Polynom* von $\{\boldsymbol{A}, \boldsymbol{B}\}$. Im Unterschied zum speziellen Eigenwertproblem kann jedoch der Grad von $p(\lambda)$ kleiner als n sein. In diesem Fall hat das aus (1) durch die Substitution $\lambda = 1/\omega$ entstehende Eigenwertproblem

$$\omega \boldsymbol{A}\boldsymbol{z} = \boldsymbol{B}\boldsymbol{z} \tag{4}$$

den Eigenwert $\omega = 0$, siehe Ü 14.1.1. Wir sagen dann, daß (1) den Eigenwert $= 1/0 = \infty$ hat. Es kann insbesondere eintreten, daß $p(\lambda)$ identisch verschwindet, d. h., daß jedes λ Eigenwert ist. Außerdem können auch komplexe Eigenwerte auftreten. In Ü 14.1.2 sieht man, daß die beschriebenen ungewöhnlichen Fälle bereits für $n = 2$ möglich sind.

Ein wesentliches theoretisches wie praktisches Hilfsmittel bei der Behandlung des speziellen symmetrischen Eigenwertproblems $\boldsymbol{A}\boldsymbol{x} = \lambda \boldsymbol{x}$ war der Übergang von $\boldsymbol{A} - \lambda \boldsymbol{I}$ zu

$$\hat{\boldsymbol{A}} - \lambda \hat{\boldsymbol{I}} = \boldsymbol{E}(\boldsymbol{A} - \lambda \boldsymbol{I})\,\boldsymbol{F} \tag{5}$$

mittels regulärer Transformationsmatrizen $\boldsymbol{E}, \boldsymbol{F}$. Damit das transformierte Problem wieder ein spezielles Eigenwertproblem wird, muß $\hat{\boldsymbol{I}} = \boldsymbol{E}\boldsymbol{I}\boldsymbol{F} = \boldsymbol{I}$, also $\boldsymbol{E} = \boldsymbol{F}^{-1}$ gefordert werden, und $\hat{\boldsymbol{A}}$ ergibt sich dann zu $\hat{\boldsymbol{A}} = \boldsymbol{F}^{-1}\boldsymbol{A}\boldsymbol{F}$, d. h., $\hat{\boldsymbol{A}}$ entsteht durch Ähnlichkeitstransformation mit $\boldsymbol{F}$ aus $\boldsymbol{A}$. Damit $\hat{\boldsymbol{A}}$ mit $\boldsymbol{A}$ symmetrisch ist, haben wir $\boldsymbol{F}^{-1} = \boldsymbol{F}^{\mathsf{T}}$, also die Orthogonalität der Transformationsmatrix $\boldsymbol{F}$ gefordert, vgl. 1.2.B und 13.1.A.

Für das allgemeine Eigenwertproblem $\boldsymbol{A}\boldsymbol{z} = \lambda \boldsymbol{B}\boldsymbol{z}$ ist die Situation etwas anders. Die zu (5) analoge Transformation führt auf

$$\hat{\boldsymbol{A}} - \lambda \hat{\boldsymbol{B}} = \boldsymbol{E}(\boldsymbol{A} - \lambda \boldsymbol{B})\,\boldsymbol{F} \quad \text{mit} \quad \hat{\boldsymbol{A}} = \boldsymbol{E}\boldsymbol{A}\boldsymbol{F}, \qquad \hat{\boldsymbol{B}} = \boldsymbol{E}\boldsymbol{B}\boldsymbol{F}. \tag{6}$$

Damit die Klasse der allgemeinen Eigenwertprobleme beim Übergang von $\{\boldsymbol{A}, \boldsymbol{B}\}$ zu $\{\hat{\boldsymbol{A}}, \hat{\boldsymbol{B}}\}$ nicht verlassen wird, braucht außer der Regularität keine weitere Forderung an $\boldsymbol{E}$ und $\boldsymbol{F}$ gestellt zu werden. Büschel, die durch die simultanen Äquivalenztransformationen (6) auseinander hervorgehen, heißen *äquivalent*. Äquivalente Büschel haben dieselben Eigenwerte, und $\hat{\boldsymbol{z}}$ ist Eigenvektor von $\{\hat{\boldsymbol{A}}, \hat{\boldsymbol{B}}\}$ genau dann, wenn $\boldsymbol{z}$ mit

$$\boldsymbol{z} = \boldsymbol{F}\hat{\boldsymbol{z}} \tag{7}$$

Eigenvektor von $\{\boldsymbol{A}, \boldsymbol{B}\}$ zum selben Eigenwert ist. Man beachte jedoch, daß bei der Transformation (6) die Symmetrie von $\{\boldsymbol{A}, \boldsymbol{B}\}$ i. allg. verlorengeht. Damit sich die Symmetrie von $\{\boldsymbol{A}, \boldsymbol{B}\}$ auf $\{\hat{\boldsymbol{A}}, \hat{\boldsymbol{B}}\}$ überträgt, muß $\boldsymbol{E} = \boldsymbol{F}^{\mathsf{T}}$ gefordert werden, d. h., die Äquivalenztransformationen (6) müssen auf die symmetrieerhaltenden simultanen Kongruenztransformationen

$$\hat{\boldsymbol{A}} - \lambda \hat{\boldsymbol{B}} = \boldsymbol{F}^{\mathsf{T}}(\boldsymbol{A} - \lambda \boldsymbol{B})\,\boldsymbol{F} \quad \text{mit} \quad \hat{\boldsymbol{A}} = \boldsymbol{F}^{\mathsf{T}}\boldsymbol{A}\boldsymbol{F}, \qquad \hat{\boldsymbol{B}} = \boldsymbol{F}^{\mathsf{T}}\boldsymbol{B}\boldsymbol{F} \tag{8}$$

eingeschränkt werden; die Relation (7) zwischen den Eigenvektoren ändert sich dabei nicht. Büschel, die gemäß (8) auseinander hervorgehen, heißen *kongruent. Der Übergang zu geeigneten einfacheren kongruenten Büscheln, bei dem die Symmetrie erhalten bleibt, ist die für das allgemeine symmetrische Eigenwertproblem angemessene Transformation.* Wir weisen darauf hin, daß in (8) nicht die Orthogonalität von $\boldsymbol{F}$ gefordert wird.

Es zeigt sich, daß die eingangs beschriebenen pathologischen Situationen bei einem symmetrischen Büschel nicht auftreten können, wenn $\boldsymbol{B}$ zusätzlich positiv definit ist.

14.1.1. *Satz.* Gegeben sei das symmetrische Büschel $\{\boldsymbol{A}, \boldsymbol{B}\}$, und $\boldsymbol{B}$ sei positiv definit. Dann gibt es eine Matrix $\boldsymbol{Z}$ mit

$$\boldsymbol{Z}^{\mathsf{T}}\boldsymbol{A}\boldsymbol{Z} = \boldsymbol{\Lambda} = \operatorname{diag}(\lambda_j), \qquad \lambda_j \in \mathbf{R}, \tag{9}$$

und

$$\boldsymbol{Z}^{\mathsf{T}}\boldsymbol{B}\boldsymbol{Z} = \boldsymbol{I}, \tag{10}$$

d. h., $\{\boldsymbol{A}, \boldsymbol{B}\}$ ist kongruent zum diagonalen Büschel $\{\boldsymbol{\Lambda}, \boldsymbol{I}\}$, und die Transformationsmatrix ist $\boldsymbol{Z}$.

Beweis. Wir konstruieren die gesuchte Transformation in zwei Schritten. Es sei zunächst

$$\boldsymbol{Q}^{\mathsf{T}}\boldsymbol{B}\boldsymbol{Q} = \boldsymbol{\Delta} = \operatorname{diag}(\delta_j), \qquad \delta_1 \geqq \cdots \geqq \delta_n > 0, \qquad \boldsymbol{Q} \text{ orthogonal}, \tag{11}$$

die Eigenwertzerlegung von $\boldsymbol{B}$; wegen der positiven Definitheit von $\boldsymbol{B}$ sind alle Eigenwerte δ_j positiv. Mit

$$\boldsymbol{\Delta}^{1/2} := \operatorname{diag}(\sqrt{\delta_j}), \qquad \boldsymbol{F} := \boldsymbol{Q}\boldsymbol{\Delta}^{-1/2}$$

gilt dann

$$\boldsymbol{F}^{\mathsf{T}}\boldsymbol{A}\boldsymbol{F} = \boldsymbol{\Delta}^{-1/2}\boldsymbol{Q}^{\mathsf{T}}\boldsymbol{A}\boldsymbol{Q}\boldsymbol{\Delta}^{-1/2} =: \boldsymbol{C} \quad \text{und} \quad \boldsymbol{F}^{\mathsf{T}}\boldsymbol{B}\boldsymbol{F} = \boldsymbol{\Delta}^{-1/2}\boldsymbol{Q}^{\mathsf{T}}(\boldsymbol{Q}\boldsymbol{\Delta}\boldsymbol{Q}^{\mathsf{T}})\,\boldsymbol{Q}\boldsymbol{\Delta}^{-1/2} = \boldsymbol{I},$$

d. h., $\{\boldsymbol{A}, \boldsymbol{B}\}$ ist kongruent zu $\{\boldsymbol{C}, \boldsymbol{I}\}$ mit $\boldsymbol{F}$, wobei $\boldsymbol{\Delta}^{-1/2} := (\boldsymbol{\Delta}^{1/2})^{-1}$ gesetzt ist. Die Matrix $\boldsymbol{C}$ ist symmetrisch und besitzt ihrerseits die Eigenwertzerlegung

$$\boldsymbol{U}^{\mathsf{T}}\boldsymbol{C}\boldsymbol{U} = \boldsymbol{\Lambda} = \operatorname{diag}(\lambda_j), \qquad \lambda_j \in \mathbf{R}, \quad \boldsymbol{U} \text{ orthogonal}, \quad \text{d. h.} \quad \boldsymbol{U}^{\mathsf{T}}\boldsymbol{I}\boldsymbol{U} = \boldsymbol{I}. \tag{12}$$

Die Gleichungen (12) besagen, daß $\{\boldsymbol{C}, \boldsymbol{I}\}$ mit $\boldsymbol{U}$ kongruent zu $\{\boldsymbol{\Lambda}, \boldsymbol{I}\}$ ist, d. h., $\{\boldsymbol{A}, \boldsymbol{B}\}$ ist mit dem Produkt $\boldsymbol{Z} := \boldsymbol{F}\boldsymbol{U}$ der einzelnen Transformationsmatrizen kongruent zu $\{\boldsymbol{\Lambda}, \boldsymbol{I}\}$. □

Die $\{\lambda_j, \boldsymbol{e}^j\}$ sind offensichtlich die Eigenpaare von $\{\boldsymbol{\Lambda}, \boldsymbol{I}\}$, so daß die $\{\lambda_j, \boldsymbol{z}^j\}$, wobei die $\boldsymbol{z}^j = \boldsymbol{Z}\boldsymbol{e}^j$ die Spalten von $\boldsymbol{Z} = (\boldsymbol{z}^1, \ldots, \boldsymbol{z}^n)$ bezeichnen, die Eigenpaare von $\{\boldsymbol{A}, \boldsymbol{B}\}$ sind; man lese (7) mit $\boldsymbol{Z}$ statt $\boldsymbol{F}$ oder schreibe (9) unter Beachtung von (10) in der Form $\boldsymbol{A}\boldsymbol{Z} = \boldsymbol{B}\boldsymbol{Z}\boldsymbol{\Lambda}$, was zu (1) äquivalent ist. Die Eigenschaft (10) von $\boldsymbol{Z}$ wird *$\boldsymbol{B}$-Orthogonalität* genannt; in elementweiser Lesart bedeutet sie

$$(\boldsymbol{z}^i)^{\mathsf{T}}\boldsymbol{B}\boldsymbol{z}^j = \begin{cases} 1 & \text{für} \quad i = j, \\ 0 & \text{für} \quad i \neq j \end{cases} \qquad (i, j = 1, \ldots, n). \tag{13}$$

Die Eigenvektoren $\{\boldsymbol{z}^j\}$ heißen deswegen auch *$\boldsymbol{B}$-orthonormal*, vgl. Ü 14.1.3. Offenbar sind $\boldsymbol{B}$-orthonormale Vektoren stets linear unabhängig. 14.1.1 bedeutet daher: *Jedes*

symmetrische Büschel $\{\boldsymbol{A}, \boldsymbol{B}\}$ *mit positiv definitem* $\boldsymbol{B}$ *hat* n *reelle Eigenwerte* $\{\lambda_j\}$, *und es gibt* n *zugehörige* $\boldsymbol{B}$*-orthonormale Eigenvektoren* $\{\boldsymbol{z}^j\}$. Wie im Fall des speziellen symmetrischen Eigenwertproblems kann diese Aussage noch verschärft werden: Zu verschiedenen Eigenwerten $\lambda_i \neq \lambda_j$ gehörende Eigenvektoren $\boldsymbol{z}^i$ und $\boldsymbol{z}^j$ sind notwendig $\boldsymbol{B}$*-orthogonal* im Sinne von $[\boldsymbol{z}^i, \boldsymbol{z}^j]_{\boldsymbol{B}} = (\boldsymbol{z}^i)^{\mathsf{T}} \boldsymbol{B}\boldsymbol{z}^j = 0$. Falls λ eine α-fache Nullstelle von $p(\lambda)$ ist, gibt es α linear unabhängige Eigenvektoren, die den zugehörigen Eigenraum $\mathcal{N}(\boldsymbol{A} - \lambda\boldsymbol{B})$ aufspannen, und diese Eigenvektoren können $\boldsymbol{B}$-orthonormal gewählt werden. Im folgenden setzen wir stets voraus, daß $\{\boldsymbol{A}, \boldsymbol{B}\}$ symmetrisch und $\boldsymbol{B}$ positiv definit ist.

Wir bemerken an dieser Stelle, daß auch für das allgemeine symmetrische Eigenwertproblem eine Reihe von nützlichen Residualkriterien und Störungsaussagen existiert, siehe B 14.1. Aus Platzgründen verzichten wir jedoch auf eine Wiedergabe.

Übungsaufgaben

Ü 14.1.1. Man überlege sich, daß $p(\lambda) = \det(\boldsymbol{A} - \lambda\boldsymbol{B})$ und $q(\omega) = \det(\omega\boldsymbol{A} - \boldsymbol{B})$ den Beziehungen

$$q(\omega) = p(1/\omega)\,\omega^n \quad \text{und} \quad p(\lambda) = q(1/\lambda)\,\lambda^n$$

genügen, und folgere hieraus, daß $\omega = 0$ Eigenwert von $\{\boldsymbol{B}, \boldsymbol{A}\}$, also $\lambda = 1/0 = \infty$ Eigenwert von $\{\boldsymbol{A}, \boldsymbol{B}\}$ genau dann ist, wenn der Grad von $p(\lambda)$ kleiner als n ist.

Ü 14.1.2. Man zeige (PARLETT [80a]):

(i) Für $\boldsymbol{A} = \begin{pmatrix} 1 & 0 \\ 0 & 0 \end{pmatrix}$, $\boldsymbol{B} = \begin{pmatrix} 1 & 0 \\ 0 & 0 \end{pmatrix}$ sind $\{1, \boldsymbol{e}^1\}$ und $\{\lambda, \boldsymbol{e}^2\}$, λ beliebig, Eigenpaare.

(ii) Für $\boldsymbol{A} = \begin{pmatrix} 1 & 0 \\ 0 & 0 \end{pmatrix}$, $\boldsymbol{B} = \begin{pmatrix} 0 & 0 \\ 0 & 1 \end{pmatrix}$ sind $\{\infty, \boldsymbol{e}^1\}$ und $\{0, \boldsymbol{e}^2\}$ Eigenpaare.

(iii) Für $\boldsymbol{A} = \begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix}$, $\boldsymbol{B} = \begin{pmatrix} 1 & 0 \\ 0 & -1 \end{pmatrix}$ sind $\{\mathrm{i}, (\mathrm{i}, -1)^{\mathsf{T}}\}$ und $\{-\mathrm{i}, (\mathrm{i}, 1)^{\mathsf{T}}\}$ Eigenpaare, wobei i mit $\mathrm{i}^2 = -1$ die imaginäre Einheit bedeutet.

Ü 14.1.3. Mit der positiv definiten Matrix $\boldsymbol{B} \in \mathbf{S}^{n,n}$ werde

$$[\boldsymbol{x}, \boldsymbol{y}] := [\boldsymbol{x}, \boldsymbol{y}]_{\boldsymbol{B}} := \boldsymbol{x}^{\mathsf{T}}\boldsymbol{B}\boldsymbol{y} \tag{14}$$

definiert. Man zeige, daß die derart erklärte Abbildung $(\boldsymbol{x}, \boldsymbol{y}) \to [\boldsymbol{x}, \boldsymbol{y}]_{\boldsymbol{B}}$ von $\mathbf{R}^n \times \mathbf{R}^n$ in $\mathbf{R}$ die Eigenschaften

(S1) $[\boldsymbol{x}, \boldsymbol{y}] = [\boldsymbol{y}, \boldsymbol{x}]$,
(S2) $[\lambda_1\boldsymbol{x}^1 + \lambda_2\boldsymbol{x}^2, \boldsymbol{y}] = \lambda_1[\boldsymbol{x}^1, \boldsymbol{y}] + \lambda_2[\boldsymbol{x}^2, \boldsymbol{y}]$,
(S3) $[\boldsymbol{x}, \boldsymbol{x}] \geqq 0$, $[\boldsymbol{x}, \boldsymbol{x}] = 0$ genau für $\boldsymbol{x} = \boldsymbol{o}$

hat, d. h., $[\boldsymbol{x}, \boldsymbol{y}]$ hat dieselben Eigenschaften wie das gewöhnliche Skalarprodukt $[\boldsymbol{x}, \boldsymbol{y}]_I = \boldsymbol{x}^{\mathsf{T}}\boldsymbol{y}$ und wird daher das *durch* $\boldsymbol{B}$ *erzeugte Skalarprodukt* oder kurz $\boldsymbol{B}$*-Skalarprodukt* genannt. Durch die Vorschrift

$$\|\boldsymbol{x}\|_{\boldsymbol{B}} := \sqrt{[\boldsymbol{x}, \boldsymbol{x}]_{\boldsymbol{B}}} = \sqrt{\boldsymbol{x}^{\mathsf{T}}\boldsymbol{B}\boldsymbol{x}} \tag{15}$$

wird auf $\mathbf{R}^n$ eine Norm definiert, die sog. *durch* $\boldsymbol{B}$ *erzeugte* (*energetische*) *Norm* oder kurz $\boldsymbol{B}$*-Norm*. Für $\boldsymbol{B}$-Skalarprodukt und $\boldsymbol{B}$-Norm gilt die (verallgemeinerte) Schwarzsche Ungleichung

$$|[\boldsymbol{x}, \boldsymbol{y}]_{\boldsymbol{B}}| \leqq \|\boldsymbol{x}\|_{\boldsymbol{B}}\,\|\boldsymbol{y}\|_{\boldsymbol{B}}. \tag{16}$$

Ü 14.1.4. Betrachtet wird das quadratische Eigenwertproblem

$$(\boldsymbol{A} - \lambda \boldsymbol{B} - \lambda^2 \boldsymbol{C})\, \boldsymbol{x} = 0 \tag{17}$$

mit $\boldsymbol{A}, \boldsymbol{B}, \boldsymbol{C} \in \mathbf{S}^{n,n}$ und positiv definiten Matrizen $\boldsymbol{A}, \boldsymbol{C}$. Man zeige: Wenn $\{\lambda, \boldsymbol{x}\}$ ein Eigenpaar von (17) ist, ist $\{\lambda, \boldsymbol{z}\}$ mit $\boldsymbol{z} := \begin{pmatrix} \boldsymbol{x} \\ \lambda \boldsymbol{x} \end{pmatrix}$ ein Eigenpaar von

$$\boldsymbol{N}\boldsymbol{z} = \lambda \boldsymbol{M}\boldsymbol{z} \tag{18}$$

mit

$$\boldsymbol{N} := \left(\begin{array}{c|c} \boldsymbol{O} & \boldsymbol{A} \\ \hline \boldsymbol{A} & -\boldsymbol{B} \end{array}\right), \qquad \boldsymbol{M} := \left(\begin{array}{c|c} \boldsymbol{A} & \boldsymbol{O} \\ \hline \boldsymbol{O} & \boldsymbol{C} \end{array}\right) \in \mathbf{S}^{2n,2n}. \tag{19}$$

Ist umgekehrt $\{\lambda, \boldsymbol{z}\}$ mit $\boldsymbol{z} = \begin{pmatrix} \boldsymbol{x} \\ \boldsymbol{y} \end{pmatrix}$ ein Eigenpaar von (18), so gilt $\boldsymbol{y} = \lambda \boldsymbol{x}$, und $\{\lambda, \boldsymbol{x}\}$ ist ein Eigenpaar von (17). Das quadratische symmetrische Eigenwertproblem (17) kann also auf das lineare allgemeine symmetrische Eigenwertproblem (18) mit positiv definitem $\boldsymbol{M}$ zurückgeführt werden, wobei sich die Dimension verdoppelt.

14.2. Explizite Reduktion auf ein spezielles Eigenwertproblem

Wir betrachten in diesem Abschnitt Methoden zur Lösung des allgemeinen symmetrischen Eigenwertproblems

$$\boldsymbol{A}\boldsymbol{z} = \lambda \boldsymbol{B}\boldsymbol{z}, \tag{1}$$

die auf der Transformation von $\{\boldsymbol{A}, \boldsymbol{B}\}$ in ein äquivalentes bzw. kongruentes Büschel $\{\hat{\boldsymbol{A}}, \hat{\boldsymbol{B}}\} = \{\boldsymbol{C}, \boldsymbol{I}\}$ beruhen, d. h., die (1) auf das spezielle Eigenwertproblem

$$\boldsymbol{C}\boldsymbol{u} = \lambda \boldsymbol{u} \tag{2}$$

zurückführen. Werden äquivalente Büschel mit $\hat{\boldsymbol{A}} = \boldsymbol{E}\boldsymbol{A}\boldsymbol{F}$, $\hat{\boldsymbol{B}} = \boldsymbol{E}\boldsymbol{B}\boldsymbol{F}$ zugelassen, so muß also

$$\hat{\boldsymbol{B}} = \boldsymbol{E}\boldsymbol{B}\boldsymbol{F} = \boldsymbol{I} \tag{3}$$

gelten, was am einfachsten durch die Festlegung

$$\boldsymbol{E} := \boldsymbol{E}_1 := \boldsymbol{B}^{-1}, \qquad \boldsymbol{F} := \boldsymbol{F}_1 := \boldsymbol{I}, \quad \text{also} \quad \boldsymbol{C} := \boldsymbol{C}_1 := \boldsymbol{B}^{-1}\boldsymbol{A} \tag{4}$$

erreicht werden kann. Dies entspricht dem Übergang von (1) zu

$$\boldsymbol{B}^{-1}\boldsymbol{A}\boldsymbol{z} = \lambda \boldsymbol{z}. \tag{5}$$

Da die entstehende Matrix $\boldsymbol{C}_1 = \boldsymbol{B}^{-1}\boldsymbol{A}$ nur in trivialen Ausnahmefällen symmetrisch ist, kann dieses naheliegende Vorgehen jedoch nicht empfohlen werden. Wir beschränken uns daher auf kongruente Büschel, bei denen die Symmetrie erhalten bleibt. Die Bedingung (3) geht dabei in

$$\boldsymbol{F}^{\mathsf{T}}\boldsymbol{B}\boldsymbol{F} = \boldsymbol{I}, \quad \text{also} \quad \boldsymbol{B} = \boldsymbol{F}^{-\mathsf{T}}\boldsymbol{F}^{-1} \tag{6}$$

über, d. h., es wird eine symmetrische Faktorisierung von $\boldsymbol{B}$ benötigt. Eine Möglichkeit zur Festlegung von $\boldsymbol{F}$ haben wir bereits im ersten Schritt des Beweises zu 14.1.1 kennengelernt:

14.2.1. *Reduktion mittels der Eigenwertzerlegung von* $\boldsymbol{B}$. Es sei

$$\boldsymbol{B} = \boldsymbol{Q}\boldsymbol{\Delta}\boldsymbol{Q}^{\mathsf{T}} \quad \text{mit} \quad \boldsymbol{\Delta} = \operatorname{diag}(\delta_j), \quad \delta_1 \geqq \cdots \geqq \delta_n > 0, \quad \boldsymbol{Q} \text{ orthogonal}, \tag{7}$$

die Eigenwertzerlegung von $\boldsymbol{B}$, und mit $\boldsymbol{\Delta}^{1/2} := \operatorname{diag}\left(\sqrt{\delta_j}\right)$ werde

$$\boldsymbol{F}_2 := \boldsymbol{Q}\boldsymbol{\Delta}^{-1/2} \tag{8}$$

gesetzt. Dann gilt

$$\boldsymbol{F}_2^{\mathsf{T}}(\boldsymbol{A} - \lambda\boldsymbol{B})\,\boldsymbol{F}_2 = \boldsymbol{C}_2 - \lambda\boldsymbol{I} \tag{9}$$

mit

$$\boldsymbol{C}_2 := \boldsymbol{F}_2^{\mathsf{T}}\boldsymbol{A}\boldsymbol{F}_2 = \boldsymbol{\Delta}^{-1/2}\boldsymbol{Q}^{\mathsf{T}}\boldsymbol{A}\boldsymbol{Q}\boldsymbol{\Delta}^{-1/2}. \tag{10}$$

14.2.2. *Bemerkung.* (i) Für voll besetztes $\boldsymbol{B}$ sollte der $\boldsymbol{QR}$-Algorithmus — d. h. die Kombination von 13.5.1 und 13.7.17 — zur Berechnung der Eigenwertzerlegung verwendet werden. Man beachte, daß bei der Reduktion die Matrix $\boldsymbol{Q}$ als Produkt der Householder-Spiegelungen aus 13.5.1 und der Givens-Drehungen aus 13.7.17 überhaupt nicht explizit gebildet zu werden braucht: Von $\{\boldsymbol{A}, \boldsymbol{B}\}$ wird direkt zu $\{\tilde{\boldsymbol{A}}, \tilde{\boldsymbol{B}}\}$ mit $\tilde{\boldsymbol{A}} := \boldsymbol{Q}^{\mathsf{T}}\boldsymbol{A}\boldsymbol{Q}$ und $\tilde{\boldsymbol{B}} := \boldsymbol{Q}^{\mathsf{T}}\boldsymbol{B}\boldsymbol{Q} = \boldsymbol{\Delta}$ übergegangen, indem die einzelnen Faktoren von $\boldsymbol{Q}$ nicht nur auf $\boldsymbol{B}$, sondern sofort bei ihrem Auftreten auch auf $\boldsymbol{A}$ angewendet werden. Das kostet insgesamt $\sim Kn^3$ opms mit $K \lessapprox 4$, vgl. 13.7.18(ii), denn der Aufwand zur Berechnung von $\tilde{\boldsymbol{A}}$ ist derselbe wie zur expliziten Berechnung von $\boldsymbol{Q}$, und die Berechnung von $\boldsymbol{\Delta}$ aus der Tridiagonalform kostet nur $O(n^2)$ opms, also relativ fast nichts. Zum Schluß wird $\boldsymbol{C}_2 := \boldsymbol{\Delta}^{-1/2}\tilde{\boldsymbol{A}}\boldsymbol{\Delta}^{-1/2}$ mit $\sim n^2$ opm gebildet. Für die so berechnete Matrix $\boldsymbol{C}_2$ gilt dann

$$\boldsymbol{C}_2 = \boldsymbol{\Delta}^{-1/2}\boldsymbol{Q}^{\mathsf{T}}(\boldsymbol{A} + \delta\boldsymbol{A})\,\boldsymbol{Q}\boldsymbol{\Delta}^{-1/2} \quad \text{mit kleinem } \delta\boldsymbol{A} \in \mathbf{S}^{n,n},$$

und $\boldsymbol{C}_2$ enthält fast die volle Information über das Originalproblem. Falls die Eigenvektoren auch gesucht sind, muß $\boldsymbol{Q}$ für die Rücktransformation $\boldsymbol{z} = \boldsymbol{F}_2\boldsymbol{u} = \boldsymbol{Q}\boldsymbol{\Delta}^{-1/2}\boldsymbol{u}$ allerdings bereitgestellt werden, was den Aufwand verdoppelt.

(ii) Im Fall $\delta_1 \gg \delta_n$ ist $\boldsymbol{C}_2$ ansteigend gestuft. Es sind dann die in 13.5.4 und 13.7.18 beschriebenen Vorsichtsmaßnahmen durchzuführen, damit eventuell auch die betragskleinen Eigenwerte ausreichend genau berechnet werden. Für die Behandlung des semidefiniten Falles $\delta_n = 0$ siehe B 14.2.

(iii) Falls $\boldsymbol{A}, \boldsymbol{B}$ Bandmatrizen sind, kann das Büschel $\{\boldsymbol{A}, \boldsymbol{B}\}$ in Analogie zu 13.5.B ohne Vergrößerung der Bandbreite kongruent in ein Büschel $\{\boldsymbol{T}, \boldsymbol{I}\}$ mit tridiagonalem $\boldsymbol{T}$ transformiert werden, siehe B 14.3. □

Eine billigere Reduktion als in 14.2.1 ist möglich, wenn (6) mittels der $\boldsymbol{LDL}^{\mathsf{T}}$-Faktorisierung von $\boldsymbol{B}$ realisiert wird. Zur Verbesserung der Stabilität verwenden wir dazu die mit Diagonalpivotisierung gemäß 6.1.5(iii) berechnete Faktorisierung, bei der die Diagonalelemente monoton fallen.

14.2.3. *Reduktion mittels* $\boldsymbol{LDL}^{\mathsf{T}}$*-Faktorisierung von* $\boldsymbol{B}$. Es sei

$$\boldsymbol{P}\boldsymbol{B}\boldsymbol{P}^{\mathsf{T}} = \boldsymbol{L}\boldsymbol{D}\boldsymbol{L}^{\mathsf{T}} \quad \text{mit} \quad \boldsymbol{D} = \operatorname{diag}(d_j), \qquad d_1 \geqq \cdots \geqq d_n > 0, \tag{11}$$

einer Permutationsmatrix $\boldsymbol{P}$ und einer unteren Einsdreiecksmatrix $\boldsymbol{L}$ die mittels Diagonalpivotisierung gemäß 6.1.1/6.1.5 berechnete $\boldsymbol{LDL}^{\mathsf{T}}$-Faktorisierung von $\boldsymbol{B}$, und mit $\boldsymbol{D}^{1/2} := \operatorname{diag}\left(\sqrt{d_j}\right)$ werde

$$\boldsymbol{F}_3 := \boldsymbol{P}^{\mathsf{T}}\boldsymbol{L}^{-\mathsf{T}}\boldsymbol{D}^{-1/2} \tag{12}$$

gesetzt. Dann gilt

$$\boldsymbol{F}_3^{\mathsf{T}}(\boldsymbol{A} - \lambda\boldsymbol{B})\,\boldsymbol{F}_3 = \boldsymbol{C}_3 - \lambda\boldsymbol{I} \tag{13}$$

mit

$$\boldsymbol{C}_3 := \boldsymbol{F}_3^{\mathsf{T}}\boldsymbol{A}\boldsymbol{F}_3 = \boldsymbol{D}^{-1/2}\boldsymbol{L}^{-1}\boldsymbol{P}\boldsymbol{A}\boldsymbol{P}^{\mathsf{T}}\boldsymbol{L}^{-\mathsf{T}}\boldsymbol{D}^{-1/2}. \tag{14}$$

Beweis. Der Vergleich von $\boldsymbol{B} = \boldsymbol{P}^{\mathsf{T}}\boldsymbol{L}\boldsymbol{D}\boldsymbol{L}^{\mathsf{T}}\boldsymbol{P}$ mit (6) zeigt, daß $\boldsymbol{F}^{-1} = \boldsymbol{D}^{1/2}\boldsymbol{L}^{\mathsf{T}}\boldsymbol{P}$ gesetzt werden muß. □

14.2.4. *Bemerkung.* (i) Bei der Realisierung von 14.2.3 sollte die $\boldsymbol{LDL}^{\mathsf{T}}$-Faktorisierung in der $(n-1)$-stufigen Version analog zu 6.1.B durchgeführt werden, was der Produktdarstellung

$$\tilde{\boldsymbol{B}} := \boldsymbol{L}^{-1}\boldsymbol{P}\boldsymbol{B}\boldsymbol{P}^{\mathsf{T}}\boldsymbol{L}^{-\mathsf{T}} = \boldsymbol{D} = \boldsymbol{L}_{n-1}\cdots\boldsymbol{L}_1\boldsymbol{T}_{n-1}\cdots\boldsymbol{T}_1\boldsymbol{B}\boldsymbol{T}_1\cdots\boldsymbol{T}_{n-1}\boldsymbol{L}_1^{\mathsf{T}}\cdots\boldsymbol{L}_{n-1}^{\mathsf{T}} \tag{15}$$

mit

$$\boldsymbol{L}^{-\mathsf{T}} = \boldsymbol{L}_1^{\mathsf{T}}\cdots\boldsymbol{L}_{n-1}^{\mathsf{T}} \tag{16}$$

entspricht, wobei die Matrizen $\boldsymbol{L}_k = \boldsymbol{L}_k(-\boldsymbol{l}^k)$ die zur Elimination verwendeten nichtorthogonalen elementaren Transformationsmatrizen bezeichnen. Die Matrix $\boldsymbol{L}^{-1}$ wird nicht explizit gebildet, sondern $\tilde{\boldsymbol{A}} := \boldsymbol{L}^{-1}\boldsymbol{P}\boldsymbol{A}\boldsymbol{P}^{\mathsf{T}}\boldsymbol{L}^{-\mathsf{T}}$ wird analog zu (15) aus $\boldsymbol{A}$ erzeugt, indem dieselben Transformationen wie mit $\boldsymbol{B}$ auch mit $\boldsymbol{A}$ ausgeführt werden. In dieser Form kostet die Transformation von $\{\boldsymbol{A}, \boldsymbol{B}\}$ in $\{\tilde{\boldsymbol{A}}, \tilde{\boldsymbol{B}}\}$ insgesamt $\sim 2n^3/3$ opms, nämlich $\sim n^3/2$ opms für $\tilde{\boldsymbol{A}}$ und $\sim n^3/6$ für $\tilde{\boldsymbol{B}} = \boldsymbol{D}$, ist also um den Faktor 6 billiger als die entsprechende Transformation in 14.2.1. Aus $\tilde{\boldsymbol{A}}$ wird $\boldsymbol{C}_3 = \boldsymbol{D}^{-1/2}\tilde{\boldsymbol{A}}\boldsymbol{D}^{-1/2}$ mit $\sim n^2$ opm berechnet. Es gilt dann

$$\boldsymbol{C}_3 = \boldsymbol{D}^{-1/2}\boldsymbol{L}^{-1}\boldsymbol{P}(\boldsymbol{A} + \delta\boldsymbol{A})\,\boldsymbol{P}^{\mathsf{T}}\boldsymbol{L}^{-\mathsf{T}}\boldsymbol{D}^{-1/2},$$

aber die Störung $\delta\boldsymbol{A} \in \mathbf{S}^{n,n}$ ist a priori nur gemäß

$$\|\delta\boldsymbol{A}\|_F \leqq \nu F\,[\operatorname{cond}(\boldsymbol{L})]^2\,\|\boldsymbol{A}\|$$

mit akzeptablem F beschränkt. Auf Grund der Diagonalpivotisierung wird cond $(\boldsymbol{L})$ meist nicht sehr groß sein, so daß $\delta\boldsymbol{A}$ klein wird und $\boldsymbol{C}_3$ ausreichende Information über $\{\boldsymbol{A}, \boldsymbol{B}\}$ enthält. Falls cond $(\boldsymbol{L})$ jedoch groß ist — dies wird selten eintreten, theoretisch kann cond $(\boldsymbol{L})$ aber exponentiell mit n wachsen, vgl. Ü 11.2.2 —, entsteht beim Übergang zu $\boldsymbol{C}_3$ ein großer Informationsverlust. Dann wird i. allg. auch $\tilde{\boldsymbol{A}}$ groß sein, so daß $\|\tilde{\boldsymbol{A}}\|/\|\boldsymbol{A}\|$ als Kriterium für die Zulässigkeit genommen werden kann. Ist dieser Wert groß, so sollte besser die teurere, aber stabilere Reduktion 14.2.1 verwendet werden. Wegen (16) gilt außerdem $\boldsymbol{z} = \boldsymbol{F}_3\boldsymbol{u} = \boldsymbol{P}^{\mathsf{T}}\boldsymbol{L}_1^{\mathsf{T}}\cdots\boldsymbol{L}_{n-1}^{\mathsf{T}}\boldsymbol{D}^{-1/2}\boldsymbol{u}$, d. h., $\boldsymbol{L}^{-1}$ wird auch bei der Rücktransformation der Eigenvektoren von $\boldsymbol{C}_3$ in solche von $\{\boldsymbol{A}, \boldsymbol{B}\}$ nicht explizit benötigt. Diese Transformation kostet $\sim n^2/2$ opms pro Vektor, ist also viel billiger als in 14.2.1.

(ii) Wenn auf die Pivotisierung verzichtet wird, vergrößert sich das Risiko, ein großes cond $(\boldsymbol{L})$ zu erhalten, wesentlich.

(iii) Im Fall $d_1 \gg d_n$ ist $\boldsymbol{C}_3$ ansteigend gestuft, so daß wie in 14.2.2(ii) vorzugehen ist. □

14.2.5. *Bemerkung.* (i) Die explizite Reduktion nach 14.2.1 oder 14.2.3 ist nur für kleine bzw. mittlere Dimensionen n zweckmäßig; eine eventuelle schwache Besetztheit von $\boldsymbol{A}, \boldsymbol{B}$ wird beim Übergang zu $\boldsymbol{C}$ zerstört.

(ii) Wenn mehr als $n/4$ Eigenwerte bzw. Eigenpaare gesucht sind, sollten alle Eigenwerte bzw. Eigenpaare von $\boldsymbol{C}$ nach dem $\boldsymbol{QR}$-Algorithmus bestimmt werden. Interessieren nur wenige Eigenwerte bzw. Eigenpaare, so empfiehlt sich das Bisektionsverfahren aus 13.6 bzw. die inverse Iteration aus 13.4. In beiden Fällen müssen die Eigenvektoren $\boldsymbol{u}$ von $\boldsymbol{C}$ gemäß $\boldsymbol{z} = \boldsymbol{F}\boldsymbol{u}$ in solche von $\{\boldsymbol{A}, \boldsymbol{B}\}$ transformiert werden.

(iii) Falls $\boldsymbol{A}, \boldsymbol{B}$ Bandgestalt haben, kann das Bisektionsverfahren direkt mit $\boldsymbol{A} - \mu\boldsymbol{B}$ durchgeführt werden: Existiert die Faktorisierung $\boldsymbol{A} - \mu\boldsymbol{B} = \boldsymbol{L}_\mu \boldsymbol{D}_\mu \boldsymbol{L}_\mu^\mathsf{T}$, so ist die Zahl der negativen Diagonalelemente von $\boldsymbol{D}_\mu$ gleich der Zahl $s(\mu)$ der Eigenwerte von $\{\boldsymbol{A}, \boldsymbol{B}\}$, die kleiner als μ sind, vgl. Ü 14.2.1. Für weiterführende Hinweise zur Behandlung von Bandmatrizen siehe B 14.3, vgl. auch 13.6.11 (vi). □

Übungsaufgabe

Ü 14.2.1. Man beweise die Aussage von 14.2.5(iii).
Hinweis: Man zeige mit 14.1.1, daß $\boldsymbol{A} - \mu\boldsymbol{I}$ und $\boldsymbol{D}_\mu$ kongruent sind, und beachte 1.2.13.

14.3. Vektor- und Teilraumiteration

Wenn die Dimension n groß ist, sind die expliziten Reduktionsverfahren aus dem vorangegangenen Abschnitt aus Aufwandsgründen nicht mehr anwendbar. Meist sind dann jedoch nur einige Eigenwerte bzw. Eigenpaare von Interesse — etwa die p betragsgrößten oder betragskleinsten —, so daß sich die Verfahren der direkten bzw. inversen Vektor- und Teilraumiteration — siehe 13.3 bzw. 13.4 — als Alternative anbieten. Da wir diese Verfahren für das spezielle symmetrische Eigenwertproblem formuliert haben, setzen wir zunächst voraus, daß das Büschel $\{\boldsymbol{A}, \boldsymbol{B}\}$ bereits mittels einer regulären Matrix $\boldsymbol{F}$ kongruent in das Büschel $\{\boldsymbol{C}, \boldsymbol{I}\}$ transformiert worden ist, d. h., es gelte

$$\boldsymbol{C} = \boldsymbol{F}^\mathsf{T}\boldsymbol{A}\boldsymbol{F} \quad \text{und} \quad \boldsymbol{I} = \boldsymbol{F}^\mathsf{T}\boldsymbol{B}\boldsymbol{F}, \quad \text{also} \quad \boldsymbol{B} = \boldsymbol{F}^{-\mathsf{T}}\boldsymbol{F}^{-1}, \tag{1}$$

mithin

$$\boldsymbol{C} - \lambda\boldsymbol{I} = \boldsymbol{F}^\mathsf{T}(\boldsymbol{A} - \lambda\boldsymbol{B})\,\boldsymbol{F}. \tag{2}$$

Die Vektoriteration 13.3.1 für $\boldsymbol{C}$ lautet dann

$$\hat{\boldsymbol{v}}^{k+1} := \boldsymbol{C}\hat{\boldsymbol{v}}^k/\hat{\omega}_{k+1} \qquad (k = 1, 2, \ldots), \tag{3}$$

wobei $\|\hat{\boldsymbol{v}}^1\| = 1$ und $\hat{\omega}_{k+1}$ so gewählt wird, daß $\|\hat{\boldsymbol{v}}^{k+1}\| = 1$ gilt. Falls λ_1 ein dominanter Eigenwert von $\boldsymbol{C}$ bzw. $\{\boldsymbol{A}, \boldsymbol{B}\}$ ist, konvergiert $\{\pm\hat{\boldsymbol{v}}^k\}$ gegen den zu λ_1 gehörenden

Eigenvektor $\boldsymbol{u}^1$ von $\boldsymbol{C}$, vgl. 13.3.A. Da die Eigenvektoren $\boldsymbol{u}$ von $\boldsymbol{C}$ und $\boldsymbol{z}$ von $\{\boldsymbol{A}, \boldsymbol{B}\}$ durch

$$\boldsymbol{z} = \boldsymbol{F}\boldsymbol{u} \tag{4}$$

verknüpft sind, siehe (2), konvergieren die i. allg. nicht normierten transformierten Iterierten

$$\boldsymbol{v}^k := \boldsymbol{F}\hat{\boldsymbol{v}}^k \tag{5}$$

bei geeigneter Vorzeichenfestlegung gegen den $\boldsymbol{u}^1$ entsprechenden Eigenvektor $\boldsymbol{z}^1$ von $\{\boldsymbol{A}, \boldsymbol{B}\}$. Aus (3), (5) und (1) folgt nun

$$\boldsymbol{v}^{k+1} = \boldsymbol{F}\hat{\boldsymbol{v}}^{k+1} = \boldsymbol{F}\boldsymbol{C}\hat{\boldsymbol{v}}^k/\hat{\omega}_{k+1} = (\boldsymbol{F}\boldsymbol{F}^{\mathsf{T}})\,\boldsymbol{A}(\boldsymbol{F}\hat{\boldsymbol{v}}^k)/\hat{\omega}_{k+1} = \boldsymbol{B}^{-1}\boldsymbol{A}\boldsymbol{v}^k/\hat{\omega}_{k+1}, \tag{6}$$

d. h., die eigentlich interessierenden, rücktransformierten Iterierten $\{\boldsymbol{v}^k\}$ sind — abgesehen von der für die Approximation unwesentlichen Normierung — gerade die Iterierten der Vektoriteration für die i. allg. nichtsymmetrische Matrix $\boldsymbol{B}^{-1}\boldsymbol{A}$, und die Konvergenzaussagen für $\{\hat{\boldsymbol{v}}^k\}$ übertragen sich sinngemäß auf $\{\boldsymbol{v}^k\}$. Die oben vorausgesetzte Reduktion auf $\{\boldsymbol{C}, \boldsymbol{I}\}$ braucht also überhaupt nicht ausgeführt zu werden; das Verfahren läßt sich allein mit den Originaldaten $\{\boldsymbol{A}, \boldsymbol{B}\}$ formulieren. Zur Berechnung von $\boldsymbol{w}^{k+1} := \boldsymbol{B}^{-1}\boldsymbol{A}\boldsymbol{v}^k$ braucht $\boldsymbol{B}^{-1}\boldsymbol{A}$ — die Matrix $\boldsymbol{C}_1$ aus 14.2 — auch nicht explizit gebildet zu werden, sondern $\boldsymbol{w}^{k+1}$ wird als Lösung des Gleichungssystems $\boldsymbol{B}\boldsymbol{w}^{k+1} = \boldsymbol{A}\boldsymbol{v}^k$ etwa unter Verwendung einer symmetrischen Faktorisierung von $\boldsymbol{B}$ bestimmt. Bei Normierung von $\boldsymbol{v}^k$ auf $\|\boldsymbol{v}^k\|_{\boldsymbol{B}} = 1$ — siehe Ü 14.1.3 — ergibt sich damit die folgende implizite Realisierung der Vektoriteration für $\boldsymbol{C}$:

14.3.1. *Implizite Vektoriteration.*

S0: Berechne Faktorisierung $\boldsymbol{B} = \boldsymbol{L}\boldsymbol{D}\boldsymbol{L}^{\mathsf{T}}$, wähle $\boldsymbol{v}^1$ mit $\|\boldsymbol{v}^1\|_{\boldsymbol{B}} = 1$, setze $k := 1$

S1: Berechne $\boldsymbol{w}^{k+1}$ unter Verwendung der $\boldsymbol{L}\boldsymbol{D}\boldsymbol{L}^{\mathsf{T}}$-Faktorisierung von $\boldsymbol{B}$ aus dem Gleichungssystem

$$\boldsymbol{B}\boldsymbol{w}^{k+1} = \boldsymbol{A}\boldsymbol{v}^k \tag{7}$$

S2: Bestimme

$$\omega_{k+1} := \|\boldsymbol{w}^{k+1}\|_{\boldsymbol{B}} = \sqrt{(\boldsymbol{w}^{k+1})^{\mathsf{T}}\,\boldsymbol{B}\boldsymbol{w}^{k+1}} = \sqrt{(\boldsymbol{w}^{k+1})^{\mathsf{T}}\,(\boldsymbol{A}\boldsymbol{v}^k)}, \tag{8}$$

setze

$$\boldsymbol{v}^{k+1} := \boldsymbol{w}^{k+1}/\omega_{k+1}$$

S3: Setze $k := k + 1$, **goto** S1

Aus den $\boldsymbol{v}^k$ können analog zu 13.3.A Näherungen für λ_1 ermittelt werden, insbesondere kann dazu der Rayleigh-Quotient benutzt werden. Der Vektor $\hat{\boldsymbol{v}}^k = \boldsymbol{F}^{-1}\boldsymbol{v}^k$ ist eine i. allg. nicht normierte Eigenvektorapproximation für $\boldsymbol{C}$, und der zugehörige Rayleigh-Quotient ergibt sich wegen (1) zu

$$\varrho_k = \varrho_{\boldsymbol{C}}(\hat{\boldsymbol{v}}^k) = \frac{(\hat{\boldsymbol{v}}^k)^{\mathsf{T}}\,\boldsymbol{C}\hat{\boldsymbol{v}}^k}{(\hat{\boldsymbol{v}}^k)^{\mathsf{T}}\,\hat{\boldsymbol{v}}^k} = \frac{(\boldsymbol{v}^k)^{\mathsf{T}}\,\boldsymbol{F}^{-\mathsf{T}}(\boldsymbol{F}^{\mathsf{T}}\boldsymbol{A}\boldsymbol{F})\,\boldsymbol{F}^{-1}\boldsymbol{v}^k}{(\boldsymbol{v}^k)^{\mathsf{T}}\,\boldsymbol{F}^{-\mathsf{T}}\boldsymbol{F}^{-1}\boldsymbol{v}^k} = \frac{(\boldsymbol{v}^k)^{\mathsf{T}}\,\boldsymbol{A}\boldsymbol{v}^k}{(\boldsymbol{v}^k)^{\mathsf{T}}\,\boldsymbol{B}\boldsymbol{v}^k}, \tag{9}$$

vgl. (13.1.44).

In den Anwendungen ist meist nicht der betragsgrößte, sondern der betragskleinste oder der am dichtesten an einer vorgegebenen Zahl μ gelegene Eigenwert gesucht. Dann muß statt der Vektoriteration die inverse Iteration aus 13.4 verwendet werden, d. h., statt $\boldsymbol{C}$ wird die Matrix

$$\hat{\boldsymbol{C}} := (\boldsymbol{C} - \mu \boldsymbol{I})^{-1} \quad \text{mit den Eigenwerten } \hat{\lambda}_j = 1/(\lambda_j - \mu) \tag{10}$$

zur Iteration verwendet. Wir numerieren die Eigenwerte λ_j von $\boldsymbol{C}$ bzw. $\{\boldsymbol{A}, \boldsymbol{B}\}$ wie in 13.4 gemäß

$$0 \leqq |\lambda_1 - \mu| \leqq |\lambda_2 - \mu| \leqq \cdots \leqq |\lambda_n - \mu| \tag{11}$$

und setzen $0 < |\lambda_1 - \mu| < |\lambda_2 - \mu|$ voraus, d. h., $\hat{\boldsymbol{C}}$ existiere, und $\hat{\lambda}_1$ sei ein dominanter Eigenwert von $\hat{\boldsymbol{C}}$. Die Iterierten $\{\pm \hat{\boldsymbol{v}}^k\}$ der inversen Iteration

$$\hat{\boldsymbol{v}}^{k+1} := \hat{\boldsymbol{C}} \hat{\boldsymbol{v}}^k / \hat{\omega}_{k+1} = (\boldsymbol{C} - \mu \boldsymbol{I})^{-1} \hat{\boldsymbol{v}}^k / \hat{\omega}_{k+1} \tag{12}$$

konvergieren dann gegen den zu λ_1 gehörenden Eigenvektor $\boldsymbol{u}^1$ von $\boldsymbol{C}$. Nun folgt aus (2) mit μ statt λ unter Beachtung von (1)

$$(\boldsymbol{C} - \mu \boldsymbol{I})^{-1} = \boldsymbol{F}^{-1} (\boldsymbol{A} - \mu \boldsymbol{B})^{-1} \boldsymbol{B} \boldsymbol{F},$$

so daß (12) mit den rücktransformierten Größen $\boldsymbol{v}^k$ in der Form

$$\boldsymbol{v}^{k+1} = \boldsymbol{F} \hat{\boldsymbol{v}}^{k+1} = (\boldsymbol{A} - \mu \boldsymbol{B})^{-1} \boldsymbol{B} \boldsymbol{v}^k / \hat{\omega}_{k+1} \tag{13}$$

geschrieben werden kann. Man beachte, daß die i. allg. nichtsymmetrische Matrix $(\boldsymbol{A} - \mu \boldsymbol{B})^{-1} \boldsymbol{B}$ ähnlich zu $(\boldsymbol{C} - \mu \boldsymbol{I})^{-1}$ ist und folglich dieselben Eigenwerte $1/(\lambda_j - \mu)$ wie die letztere hat. Wie die direkte Vektoriteration kann daher auch die inverse Iteration allein mit den Originaldaten $\{\boldsymbol{A}, \boldsymbol{B}\}$ durchgeführt werden:

14.3.2. *Implizite inverse Iteration.*

S0: Wähle $\mu \in \mathbf{R}$, berechne numerisch gutartige Faktorisierung von $\boldsymbol{A} - \mu \boldsymbol{B}$, wähle $\boldsymbol{v}^1$ mit $\|\boldsymbol{v}^1\|_{\boldsymbol{B}} = 1$, setze $k := 1$

S1: Berechne $\boldsymbol{w}^{k+1}$ unter Verwendung der in S0 berechneten Faktorisierung aus dem Gleichungssystem

$$(\boldsymbol{A} - \mu \boldsymbol{B}) \boldsymbol{w}^{k+1} = \boldsymbol{B} \boldsymbol{v}^k \tag{14}$$

S2: Setze $\boldsymbol{v}^{k-1} := \boldsymbol{w}^{k+1} / \|\boldsymbol{w}^{k+1}\|_{\boldsymbol{B}}$

S3: Setze $k := k + 1$, **goto** S1

14.3.3. *Bemerkung.* (i) Im Unterschied zur Situation beim speziellen Eigenwertproblem ist sowohl in 14.3.1 als auch in 14.3.2 pro Schritt ein lineares Gleichungssystem zu lösen. Bei der direkten Vektoriteration ist die Koeffizientenmatrix $\boldsymbol{B}$ positiv definit, so daß die $\boldsymbol{LDL}^T$-Faktorisierung mit positiven Diagonalelementen — und gegebenenfalls mit Diagonalpivotisierung — berechnet werden kann, vgl. 6.1.A. Die bei der inversen Iteration auftretende Matrix $\boldsymbol{A} - \mu \boldsymbol{B}$ ist jedoch i. allg. indefinit, so daß eine auch dann numerisch gutartige Faktorisierung — etwa die spaltenpivotisierte Gauß-Faktorisierung oder die BKP-Faktorisierung — verwendet werden sollte, vgl. 13.4.2(i).

(ii) Bei der Faktorisierung von $\boldsymbol{B}$ bzw. $\boldsymbol{A} - \mu\boldsymbol{B}$ kann eine für großes n fast immer vorliegende schwache Besetztheit von $\boldsymbol{A}$ und $\boldsymbol{B}$ nach den in 6.4 beschriebenen bzw. erwähnten Methoden ausgenutzt werden. Sollte n so groß sein, daß auch dann die Faktorisierung wegen zu großer Speicherplatzforderungen nicht möglich ist, müssen die Systeme (7) bzw. (14) durch ein inneres Iterationsverfahren gelöst werden.

(iii) Wenn in S1 die feste Verschiebung μ durch den zu $\boldsymbol{v}^k$ gehörenden Rayleigh-Quotienten $\mu = \varrho_k$ gemäß (9) ersetzt wird, geht 14.3.2 in die implizit realisierte Rayleigh-Quotienten-Iteration 13.4.6 für $\boldsymbol{C}$ über. Allerdings muß dann als Preis für die kubische Konvergenz — siehe 13.4.7 — pro Schritt eine neue Faktorisierung von $\boldsymbol{A} - \varrho_k\boldsymbol{B}$ berechnet werden. □

Wenn nicht nur der erste, sondern p dominante Eigenvektoren von $\{\boldsymbol{A}, \boldsymbol{B}\}$ gesucht sind, wird zur Teilraumiteration übergegangen. Wir diskutieren zunächst, wie die Rayleigh-Ritz-Prozedur 13.3.13 für das allgemeine Eigenwertproblem zu realisieren ist. Ausgangspunkt ist wieder die durch (1), (2) beschriebene Reduktion. Der zu $\{\boldsymbol{A}, \boldsymbol{B}\}$ gehörende approximierende m-dimensionale Teilraum $\mathcal{R}(\boldsymbol{W})$ mit $p \leqq m \leqq n$ werde durch die spaltenreguläre Matrix $\boldsymbol{W} \in \mathbf{R}^{n,m}$ repräsentiert. Wegen (2), (4) ist dann

$$\mathcal{R}(\hat{\boldsymbol{W}}) \quad \text{mit} \quad \hat{\boldsymbol{W}} := \boldsymbol{F}^{-1}\boldsymbol{W} \tag{15}$$

der entsprechende Teilraum für $\boldsymbol{C}$. Um 13.3.12 mit $\boldsymbol{C}$ statt $\boldsymbol{A}$ anwenden zu können, benötigen wir eine durch die Spalten einer Matrix $\hat{\boldsymbol{Q}}$ repräsentierte orthonormale Basis von $\mathcal{R}(\hat{\boldsymbol{W}})$; wie bisher werden die zu $\{\boldsymbol{C}, \boldsymbol{I}\}$ gehörenden Größen durch ein „$\wedge$" gekennzeichnet. Die Forderung $\mathcal{R}(\hat{\boldsymbol{Q}}) = \mathcal{R}(\hat{\boldsymbol{W}})$ ist gleichbedeutend mit

$$\hat{\boldsymbol{Q}} = \hat{\boldsymbol{W}}\boldsymbol{G} \quad \text{mit regulärem} \quad \boldsymbol{G} \in \mathbf{R}^{m,m}, \tag{16}$$

und die Bedingung der Spaltenorthonormalität führt auf $\hat{\boldsymbol{Q}}^\mathsf{T}\hat{\boldsymbol{Q}} = \boldsymbol{G}^\mathsf{T}\hat{\boldsymbol{W}}^\mathsf{T}\hat{\boldsymbol{W}}\boldsymbol{G} = \boldsymbol{I}$, also

$$\hat{\boldsymbol{W}}^\mathsf{T}\hat{\boldsymbol{W}} = \boldsymbol{W}^\mathsf{T}\boldsymbol{F}^{-\mathsf{T}}\boldsymbol{F}^{-1}\boldsymbol{W} = \boldsymbol{W}^\mathsf{T}\boldsymbol{B}\boldsymbol{W} = \boldsymbol{G}^{-\mathsf{T}}\boldsymbol{G}^{-1}, \tag{17}$$

man beachte (15) und (1). Da $\boldsymbol{W}^\mathsf{T}\boldsymbol{B}\boldsymbol{W}$ positiv definit ist, existiert die Cholesky-Faktorisierung

$$\boldsymbol{W}^\mathsf{T}\boldsymbol{B}\boldsymbol{W} = \boldsymbol{L}\boldsymbol{L}^\mathsf{T}, \tag{18}$$

und (17) ist mit $\boldsymbol{G} := \boldsymbol{L}^{-\mathsf{T}}$ erfüllt, was auf

$$\hat{\boldsymbol{Q}} = \hat{\boldsymbol{W}}\boldsymbol{L}^{-\mathsf{T}} = \boldsymbol{F}^{-1}\boldsymbol{W}\boldsymbol{L}^{-\mathsf{T}} \tag{19}$$

führt. Die in 13.3.12 benötigte Matrix $\hat{\boldsymbol{P}}$ ist dann wegen (1)

$$\hat{\boldsymbol{P}} := \hat{\boldsymbol{Q}}^\mathsf{T}\boldsymbol{C}\hat{\boldsymbol{Q}} = \boldsymbol{L}^{-1}\boldsymbol{W}^\mathsf{T}\boldsymbol{F}^{-\mathsf{T}}(\boldsymbol{F}^\mathsf{T}\boldsymbol{A}\boldsymbol{F})\,\boldsymbol{F}^{-1}\boldsymbol{W}\boldsymbol{L}^{-\mathsf{T}} = \boldsymbol{L}^{-1}\boldsymbol{W}^\mathsf{T}\boldsymbol{A}\boldsymbol{W}\boldsymbol{L}^{-\mathsf{T}}, \tag{20}$$

und die zugehörige Eigenwertzerlegung sei

$$\hat{\boldsymbol{P}} = \hat{\boldsymbol{X}}\boldsymbol{M}\hat{\boldsymbol{X}}^\mathsf{T} \quad \text{mit} \quad \boldsymbol{M} = \operatorname{diag}(\mu_j) \text{ und orthogonalem } \hat{\boldsymbol{X}}. \tag{21}$$

Die mit den Spalten von $\hat{\boldsymbol{V}} := (\hat{\boldsymbol{v}}^1, \ldots, \hat{\boldsymbol{v}}^m) := \hat{\boldsymbol{Q}}\hat{\boldsymbol{X}}$ gebildeten Ritzschen Näherungen $\{\mu_j, \hat{\boldsymbol{v}}^j\}$ für $\boldsymbol{C}$ bezüglich $\mathcal{R}(\hat{\boldsymbol{Q}}) = \mathcal{R}(\hat{\boldsymbol{W}})$ minimieren

$$\begin{aligned} &\|\hat{\boldsymbol{R}}\|_F^2 = \operatorname{tr}(\hat{\boldsymbol{R}}^\top\hat{\boldsymbol{R}}) = \|\boldsymbol{C}\hat{\boldsymbol{V}} - \hat{\boldsymbol{V}}\boldsymbol{M}\|_F^2 = \sum_{j=1}^{m} \|\boldsymbol{C}\hat{\boldsymbol{v}}^j - \mu_j\hat{\boldsymbol{v}}^j\|^2, \\ &\hat{\boldsymbol{R}} := \boldsymbol{C}\hat{\boldsymbol{V}} - \hat{\boldsymbol{V}}\boldsymbol{M}, \end{aligned} \tag{22}$$

über alle $\boldsymbol{M} = \operatorname{diag}(\mu_j)$ und alle spaltenorthonormalen $\hat{\boldsymbol{V}}$ mit $\mathcal{R}(\hat{\boldsymbol{V}}) = \mathcal{R}(\hat{\boldsymbol{Q}})$. Die entsprechenden Paare $\{\mu_j, \boldsymbol{v}^j\}$ bezüglich $\{\boldsymbol{A}, \boldsymbol{B}\}$ entstehen dann wegen (2), (15) durch die Rücktransformation

$$\boldsymbol{v}^j := \boldsymbol{F}\hat{\boldsymbol{v}}^j \qquad (j = 1, \ldots, m), \tag{23}$$

und für die zugehörige Matrix $\boldsymbol{V} := (\boldsymbol{v}^1, \ldots, \boldsymbol{v}^m)$ gilt wegen (19) und (15)

$$\begin{aligned} &\boldsymbol{V} = \boldsymbol{F}\hat{\boldsymbol{V}} = \boldsymbol{F}\hat{\boldsymbol{Q}}\hat{\boldsymbol{X}} = \boldsymbol{F}\hat{\boldsymbol{W}}\boldsymbol{L}^{-\top}\hat{\boldsymbol{X}} = \boldsymbol{W}\boldsymbol{L}^{-\top}\hat{\boldsymbol{X}} =: \boldsymbol{W}\boldsymbol{X} \\ &\text{mit} \quad \boldsymbol{X} := (\boldsymbol{x}^1, \ldots, \boldsymbol{x}^m) := \boldsymbol{L}^{-\top}\hat{\boldsymbol{X}}. \end{aligned} \tag{24}$$

Zufolge (18) gilt dabei

$$\boldsymbol{X}^\top(\boldsymbol{W}^\top\boldsymbol{B}\boldsymbol{W})\,\boldsymbol{X} = \boldsymbol{V}^\top\boldsymbol{B}\boldsymbol{V} = \hat{\boldsymbol{X}}^\top\boldsymbol{L}^{-1}\boldsymbol{W}^\top\boldsymbol{B}\boldsymbol{W}\boldsymbol{L}^{-\top}\hat{\boldsymbol{X}} = \hat{\boldsymbol{X}}^\top\hat{\boldsymbol{X}} = \boldsymbol{I}, \tag{25}$$

d. h., die $\{\boldsymbol{x}^j\}$ sind $\boldsymbol{W}^\top\boldsymbol{B}\boldsymbol{W}$-orthonormal, die $\{\boldsymbol{v}^j\}$ dagegen $\boldsymbol{B}$-orthonormal. Aus (20), (21) ergibt sich außerdem

$$\boldsymbol{X}^\top(\boldsymbol{W}^\top\boldsymbol{A}\boldsymbol{W})\,\boldsymbol{X} = \hat{\boldsymbol{X}}^\top\boldsymbol{L}^{-1}\boldsymbol{W}^\top\boldsymbol{B}\boldsymbol{W}\boldsymbol{L}^{-\top}\hat{\boldsymbol{X}} = \hat{\boldsymbol{X}}^\top\hat{\boldsymbol{P}}\hat{\boldsymbol{X}} = \boldsymbol{M}, \tag{26}$$

d. h., die Matrix $\boldsymbol{X}$ transformiert $\{\boldsymbol{W}^\top\boldsymbol{A}\boldsymbol{W}, \boldsymbol{W}^\top\boldsymbol{B}\boldsymbol{W}\}$ kongruent in $\{\boldsymbol{M}, \boldsymbol{I}\}$. Schließlich kann $\hat{\boldsymbol{R}}$ in (22) mit (1) und (24) gemäß

$$\hat{\boldsymbol{R}} = (\boldsymbol{F}^\top\boldsymbol{A}\boldsymbol{F})\,\boldsymbol{F}^{-1}\boldsymbol{V} - \boldsymbol{F}^{-1}\boldsymbol{V}\boldsymbol{M} = \boldsymbol{F}^\top(\boldsymbol{A}\boldsymbol{V} - \boldsymbol{B}\boldsymbol{V}\boldsymbol{M}) =: \boldsymbol{F}^\top\boldsymbol{R} \tag{27}$$

mit $\boldsymbol{R} := (\boldsymbol{r}^1, \ldots, \boldsymbol{r}^m) := \boldsymbol{A}\boldsymbol{V} - \boldsymbol{B}\boldsymbol{V}\boldsymbol{M}$ dargestellt werden. Wieder mit (1) folgt $\|\boldsymbol{F}^\top\boldsymbol{r}\|^2 = \boldsymbol{r}^\top\boldsymbol{F}\boldsymbol{F}^\top\boldsymbol{r} = \boldsymbol{r}^\top\boldsymbol{B}^{-1}\boldsymbol{r} = \|\boldsymbol{r}\|_{\boldsymbol{B}^{-1}}^2$, so daß (22) in

$$\|\hat{\boldsymbol{R}}\|_F^2 = \operatorname{tr}(\boldsymbol{R}^\top\boldsymbol{B}^{-1}\boldsymbol{R}) = \sum_{j=1}^{m} \|\boldsymbol{F}^\top\boldsymbol{r}^j\|^2 = \sum_{j=1}^{m} \|\boldsymbol{r}^j\|_{\boldsymbol{B}^{-1}}^2 = \sum_{j=1}^{m} \|\boldsymbol{A}\boldsymbol{v}^j - \mu_j\boldsymbol{B}\boldsymbol{v}^j\|_{\boldsymbol{B}^{-1}}^2 \tag{28}$$

übergeht. Die Paare $\{\mu_j, \boldsymbol{v}^j\}$ minimieren also die mit $\boldsymbol{B}^{-1}$ gewichteten Residuumsnormen. Als Zusammenfassung erhalten wir das folgende Analogon zu 13.3.12:

14.3.4. *Satz.* Gegeben seien das symmetrische Büschel $\{\boldsymbol{A}, \boldsymbol{B}\}$ mit positiv definitem $\boldsymbol{B}$ und die reguläre Matrix $\boldsymbol{W} \in \mathbf{R}^{n,m}$, $m \leqq n$. Es sei $\boldsymbol{X} = (\boldsymbol{x}^1, \ldots, \boldsymbol{x}^m) \in \mathbf{R}^{m,m}$ diejenige Matrix, die $\{\bar{\boldsymbol{A}}, \bar{\boldsymbol{B}}\}$ mit $\bar{\boldsymbol{A}} := \boldsymbol{W}^\top\boldsymbol{A}\boldsymbol{W}$, $\bar{\boldsymbol{B}} := \boldsymbol{W}^\top\boldsymbol{B}\boldsymbol{W}$ kongruent in $\{\boldsymbol{M}, \boldsymbol{I}\}$ mit $\boldsymbol{M} = \operatorname{diag}(\mu_j)$ transformiert, d. h., es gelte

$$\bar{\boldsymbol{A}}\boldsymbol{x}^j = \mu_j\bar{\boldsymbol{B}}\boldsymbol{x}^j \qquad (j = 1, \ldots, m)$$

mit $\bar{\boldsymbol{B}}$-orthonormalen Eigenvektoren $\{\boldsymbol{x}^j\}$. Aus $\boldsymbol{W}$ und $\boldsymbol{X}$ werde

$$\boldsymbol{V} := (\boldsymbol{v}^1, \ldots, \boldsymbol{v}^m) := \boldsymbol{W}\boldsymbol{X}$$

gebildet. Dann sind die Paare $\{\mu_j, \boldsymbol{v}^j\}$ die Ritzschen Eigenpaare für das durch $\{\boldsymbol{A}, \boldsymbol{B}\}$ definierte allgemeine Eigenwertproblem bezüglich des Teilraumes $\mathcal{R}(\boldsymbol{W})$, d. h., sie minimieren (28) über alle diagonalen $\boldsymbol{M} \in \mathbf{R}^{m,m}$ und alle $\boldsymbol{V} \in \mathbf{R}^{n,m}$ mit $\boldsymbol{V}^T\boldsymbol{B}\boldsymbol{V} = \boldsymbol{I}$ und $\mathcal{R}(\boldsymbol{V}) = \mathcal{R}(\boldsymbol{W})$.

Wir geben abschließend eine zu 13.3.17(i) analoge Kombination der Teilraumiteration und des RR-Algorithmus aus 14.3.4 an. Aus den bereits diskutierten Gründen beziehen wir uns dabei nicht auf $\boldsymbol{C}$, sondern auf $(\boldsymbol{C} - \mu\boldsymbol{I})^{-1}$, wobei die Realisierung wie in 14.3.2 implizit mit den Originaldaten und rücktransformierten Teilräumen erfolgt.

14.3.5. *Implizite inverse Teilraumiteration mit RR-Algorithmus.*

S0: Wähle $\mu \in \mathbf{R}$ und berechne numerisch gutartige Faktorisierung von $\boldsymbol{A} - \mu\boldsymbol{B}$, wähle m mit $1 \leqq m \leqq n$ und $\boldsymbol{V}_1 \in \mathbf{R}^{n,m}$ mit $\boldsymbol{V}_1^T\boldsymbol{B}\boldsymbol{V}_1 = \boldsymbol{I}$, setze $k := 1$

S1: Berechne $\boldsymbol{W}_{k+1} = (\boldsymbol{w}^{k+1,1}, \ldots, \boldsymbol{w}^{k+1,m}) \in \mathbf{R}^{n,m}$ unter Verwendung der in S0 berechneten Faktorisierung von $\boldsymbol{A} - \mu\boldsymbol{B}$ als Lösung von

$$(\boldsymbol{A} - \mu\boldsymbol{B})\,\boldsymbol{W}_{k+1} = \boldsymbol{B}\boldsymbol{V}_k \tag{29}$$

S2: Bilde $\bar{\boldsymbol{A}}_{k+1} := \boldsymbol{W}_{k-1}^T\boldsymbol{A}\boldsymbol{W}_{k+1}$, $\bar{\boldsymbol{B}}_{k+1} := \boldsymbol{W}_{k-1}^T\boldsymbol{B}\boldsymbol{W}_{k+1}$. Bestimme die Eigenwerte $\{\mu_j\}$ und die zugehörigen $\bar{\boldsymbol{B}}_{k+1}$-orthonormalen Eigenvektoren $\{\boldsymbol{x}^j\}$ des allgemeinen symmetrischen Eigenwertproblems

$$\bar{\boldsymbol{A}}_{k+1}\boldsymbol{x}^j = \mu_j\bar{\boldsymbol{B}}_{k+1}\boldsymbol{x}^j \qquad (j = 1, \ldots, m) \tag{30}$$

S3: Bilde $\boldsymbol{V}_{k+1} := (\boldsymbol{v}^{k+1,1}, \ldots, \boldsymbol{v}^{k+1,m}) := \boldsymbol{W}_{k+1}\boldsymbol{X}_{k+1}$ mit $\boldsymbol{X}_{k+1} := (\boldsymbol{x}^1, \ldots, \boldsymbol{x}^m)$

S4: Setze $k := k + 1$, **goto** S1

14.3.6. *Bemerkung.* (i) Unter geeigneten Voraussetzungen approximieren die Ritzschen Eigenpaare $\{\mu_j, \boldsymbol{v}^j\}$ die in der Numerierung (11) zugehörigen Eigenpaare $\{\lambda_j, \boldsymbol{z}^j\}$ von $\{\boldsymbol{A}, \boldsymbol{B}\}$, siehe 13.3.B,C.

(ii) Die Matrixgleichung (29) entspricht den m linearen Gleichungen

$$(\boldsymbol{A} - \mu\boldsymbol{B})\,\boldsymbol{w}^{k+1,j} = \boldsymbol{B}\boldsymbol{v}^{k,j} \qquad (j = 1, \ldots, m)$$

für die Spalten von $\boldsymbol{W}_{k+1}$; die Koeffizientenmatrix ist für alle j dieselbe. Für ihre Lösung gilt 14.3.3 sinngemäß.

(iii) In der Regel ist $m \ll n$, d. h., das Eigenwertproblem (30) ist von kleiner Dimension und i. allg. voll besetzt. Zur Lösung empfiehlt sich daher die explizite Reduktion auf ein spezielles Eigenwertproblem nach den Methoden aus 14.2. □

14.3.7. *Bemerkung.* Alternative Verfahren zur Lösung des allgemeinen symmetrischen Eigenwertproblems sind einmal die Relaxations- und Gradientenverfahren, die auf einer Minimierung des Rayleigh-Quotienten $\varrho = \boldsymbol{v}^T\boldsymbol{A}\boldsymbol{v}/\boldsymbol{v}^T\boldsymbol{B}\boldsymbol{v}$ bzw. einer geeigneten, auf Teilräume definierten Verallgemeinerung von diesem beruhen, siehe B 14.5 für Hinweise. Zum anderen können die bereits in 13.3 erwähnten Verfahren

vom Lanczos-Typ mit $\boldsymbol{C}$ bzw. $(\boldsymbol{C} - \mu \boldsymbol{I})^{-1}$, aber auch direkt mit $\boldsymbol{B}^{-1}\boldsymbol{A}$ bzw. $(\boldsymbol{A} - \mu \boldsymbol{B})^{-1}\boldsymbol{B}$ angewendet werden; man beachte, daß die beiden letztgenannten Matrizen bezüglich des durch $\boldsymbol{B}$ definierten Skalarproduktes selbstadjungiert sind, siehe Ü 14.3.1. Die Lanczos-Algorithmen, die in den letzten Jahren sehr intensiv untersucht worden sind, konvergieren i. allg. deutlich besser als die Teilraumiteration. Wir weisen auch hier auf die Spezialliteratur hin, siehe wieder B 14.5. □

Übungsaufgabe

Ü 14.3.1. Es sei $\boldsymbol{B}$ symmetrisch und positiv definit. Dann heißt die Matrix $\boldsymbol{M}$ *selbstadjungiert bezüglich* $[\boldsymbol{x}, \boldsymbol{y}]_{\boldsymbol{B}}$ oder kurz $\boldsymbol{B}$*-selbstadjungiert*, wenn

$$[\boldsymbol{x}, \boldsymbol{My}]_{\boldsymbol{B}} = [\boldsymbol{Mx}, \boldsymbol{y}]_{\boldsymbol{B}}$$

für alle $\boldsymbol{x}, \boldsymbol{y}$ gilt. Man zeige:

(i) $\boldsymbol{M}$ ist $\boldsymbol{I}$-selbstadjungiert genau dann, wenn $\boldsymbol{M}$ symmetrisch ist.

(ii) Es sei $\boldsymbol{A}$ symmetrisch und μ sei kein Eigenwert von $\{\boldsymbol{A}, \boldsymbol{B}\}$. Dann sind die Matrizen $\boldsymbol{B}^{-1}\boldsymbol{A}$ und $(\boldsymbol{A} - \mu \boldsymbol{B})^{-1}\boldsymbol{B}$ beide $\boldsymbol{B}$-selbstadjungiert.

Bemerkungen zum Kapitel 14

B 14.1. Eine detaillierte Diskussion des allgemeinen symmetrischen Eigenwertproblems ist bei PARLETT [80a] zu finden, wo viele nützliche Fakten und Hinweise gegeben werden. Zur Störungstheorie sei auf STEWART [72, 78, 79a], ELSNER/SUN [82], SUN [83] verwiesen. Neuere Resultate und weiterführende Literaturhinweise über Probleme, die mit Matrixbüscheln zusammenhängen, sind in dem von KÅGSTRÖM/RUHE [83] herausgegebenen Sammelband enthalten.

B 14.2. Die in 14.2 angegebenen Methoden zur expliziten Reduktion des allgemeinen Eigenwertproblems auf ein spezielles gehören zu den Standardtechniken der numerischen linearen Algebra, siehe WILKINSON [65], PETERS/WILKINSON [70b] und PARLETT [80a]. Sie sind an die positive Definitheit von $\boldsymbol{B}$ gebunden. Falls $\boldsymbol{B}$ nur positiv semidefinit ist, also $\delta_n = 0$ Eigenwert von $\boldsymbol{B}$ ist, kann die nach FIX/HEIBERGER [72] benannte Reduktion vorgenommen werden; sie wird auch bei PARLETT [80a] beschrieben. Numerisch unterscheidet sich der Fall einer semidefiniten Matrix $\boldsymbol{B}$ nicht von dem einer positiv definiten mit schlechter Kondition.

B 14.3. Die in 14.2.2(iii) erwähnte Transformation eines Büschels $\{\boldsymbol{A}, \boldsymbol{B}\}$ aus Bandmatrizen in ein kongruentes Büschel $\{\boldsymbol{T}, \boldsymbol{I}\}$ mit tridiagolem $\boldsymbol{T}$ geht auf CRAWFORD [73] zurück. Andere Verfahren für Bandmatrizen werden von PETERS/WILKINSON [69], SCHWARZ [77, 80] und WALDVOGEL [82] beschrieben.

B 14.4. Ein alternatives Verfahren zur Lösung des nicht notwendig symmetrischen allgemeinen Eigenwertproblems stellt der $\boldsymbol{QZ}$-Algorithmus von MOLER/STEWART [73] dar, bei dem $\{\boldsymbol{A}, \boldsymbol{B}\}$ mittels einer Folge orthogonaler Äquivalenztransformationen simultan in ein äquivalentes Büschel $\{\boldsymbol{R}_{\boldsymbol{A}}, \boldsymbol{R}_{\boldsymbol{B}}\}$ mit oberen Dreiecksmatrizen $\boldsymbol{R}_{\boldsymbol{A}}, \boldsymbol{R}_{\boldsymbol{B}}$ transformiert wird. Die Quotienten $(\boldsymbol{R}_{\boldsymbol{A}})_{jj}/(\boldsymbol{R}_{\boldsymbol{B}})_{jj}$ sind dann die Eigenwerte von $\{\boldsymbol{A}, \boldsymbol{B}\}$. Falls $\boldsymbol{B}$ regulär ist, ist der $\boldsymbol{QZ}$-Algorithmus identisch mit einer impliziten Realisierung des $\boldsymbol{QR}$-Algorithmus für $\boldsymbol{AB}^{-1}$. Wir gehen auf diesen Algorithmus nicht weiter ein, da bei ihm die Symmetrie von $\{\boldsymbol{A}, \boldsymbol{B}\}$ zerstört wird; der Aufwand für symmetrische Büschel ist derselbe wie für nichtsymmetrische.

B 14.5. Die Verfahren der direkten und inversen Vektor- und Teilraumiteration für das allgemeine Eigenwertproblem werden ausführlich bei PARLETT [80a] behandelt. Stärker auf die Belange der Ingenieurwissenschaften eingegangen wird in den Büchern von BATHÉ/WILSON [76] und JENNINGS [77]. Algorithmus 14.3.5 ist z. B. bei PARLETT [80b] zu finden. Eine Variante der RQ-Iteration für das quadratische Eigenwertproblem (14.1.17) ohne Reduktion auf ein lineares der doppelten Dimension beschreiben SCOTT/WARD [82]. Zur Literatur über Relaxations- und Gradientenverfahren siehe B 13.5. Zur Klasse der Lanczos-Algorithmen sei wieder auf PARLETT [80a] und CULLUM/WILLOUGHBY [85] verwiesen. Hier sind auch in nächster Zeit noch interessante neue Ergebnisse zu erwarten.

15. Zusammenfassung zum Teil IV

Das Vorgehen zur Lösung des speziellen symmetrischen Eigenwertproblems $\boldsymbol{A}\boldsymbol{x} = \lambda\boldsymbol{x}$ hängt wesentlich von der Struktur der Matrix $\boldsymbol{A}$ und der Anzahl der gesuchten Eigenwerte und gegebenenfalls Eigenvektoren ab.

Wenn $\boldsymbol{A}$ voll besetzt ist und mehr als ca. 25% der Eigenwerte bzw. Eigenvektoren berechnet werden sollen, ist die Tridiagonalisierung von $\boldsymbol{A}$ nach 13.5.A und die vollständige Lösung des entstehenden tridiagonalen Eigenwertproblems nach dem $\boldsymbol{QR}$-Algorithmus aus 13.7.C — etwa in der impliziten $\boldsymbol{QL}$-Version — der effektivste Weg. Von der Genauigkeit vergleichbar, aber vom Aufwand her ungünstiger ist das Jacobi-Verfahren aus 13.2. Es sollte nur dann angewendet werden, wenn $\boldsymbol{A}$ wenig von einer Diagonalmatrix abweicht, eine geringe Genauigkeit gefordert ist oder Wert auf ein kurzes, kompaktes Programm gelegt wird. Sind nur wenige Eigenwerte gesucht, sollten diese durch Schneiden des Spektrums der Tridiagonalmatrix nach dem Bisektionsverfahren aus 13.6.B ermittelt werden. Zugehörige Eigenvektoren können mittels inverser Iteration nach 13.4 bestimmt werden; bei mehrfachen Eigenwerten bzw. Eigenwerthaufen sind dann Zusatzmaßnahmen erforderlich, um die Orthogonalität der berechneten Eigenvektoren zu sichern.

In der Regel werden die betragskleinen Eigenwerte von $\boldsymbol{A}$ mit geringerer relativer Genauigkeit als die betragsgroßen berechnet, siehe 13.1.B. Für gestuftes $\boldsymbol{A}$ können jedoch auch die betragskleinen Eigenwerte bei sachgemäßem Vorgehen ausreichend genau berechnet werden. Für ansteigend gestuftes $\boldsymbol{A}$ muß dazu die Tridiagonalisierung von rechts nach links erfolgen, und der $\boldsymbol{QR}$-Algorithmus ist in der $\boldsymbol{QL}$-Variante zu verwenden, siehe 13.5.A und 13.7.C.

Wenn $\boldsymbol{A}$ eine Bandmatrix ist, kann die Tridiagonalisierung mittels Givens-Drehungen bei Erhalt der Bandbreite vorgenommen werden, siehe 13.5.B. Da dann i. allg. nur wenige Eigenwerte bzw. Eigenvektoren gesucht sind, sollten diese durch Schneiden des Spektrums bzw. inverse Iteration bestimmt werden, wobei unter Umständen günstiger direkt mit $\boldsymbol{A} - \mu\boldsymbol{I}$ statt mit der Tridiagonalmatrix $\boldsymbol{T} - \mu\boldsymbol{I}$ gearbeitet wird.

Wenn $\boldsymbol{A}$ schwach besetzt und von hoher Dimension ist, sind die Verfahren der Vektor- und Teilraumiteration zu empfehlen. Falls der betragsgrößte oder die p betragsgrößten Eigenwerte und zugehörigen Eigenvektoren gesucht sind, kann die direkte Vektor- oder Teilraumiteration aus 13.3 verwendet werden, bei der nur

Ausdrücke der Form $\boldsymbol{w} = \boldsymbol{A}\boldsymbol{v}$ für gegebenes $\boldsymbol{v}$ ausgewertet werden müssen. Sind dagegen die betragskleinsten oder allgemeiner die nahe an einer vorgegebenen Zahl μ gelegenen Eigenwerte gesucht, so kommt die inverse Vektor- oder Teilraumiteration in Frage, siehe 13.4. Pro Schritt sind dann ein oder mehrere Gleichungssysteme $(\boldsymbol{A} - \mu\boldsymbol{I})\,\boldsymbol{w} = \boldsymbol{v}$ für gegebenes $\boldsymbol{v}$ zu lösen, wobei die schwache Besetztheit nach den in 6.4 angegebenen Methoden ausgenutzt werden kann. Bei der Kopplung mit dem RR-Algorithmus kommt pro Schritt die Lösung eines speziellen Eigenwertproblems niedriger Dimension hinzu. Alternative Methoden sind Relaxations- und Gradientenverfahren und die Verfahren vom Lanczos-Typ, siehe dazu B 13.5 und B 13.7.

Die berechneten Eigenwerte und Eigenvektoren können nach Dongarra/Moler/Wilkinson [83] verbessert werden.

Für das allgemeine symmetrische Eigenwertproblem $\boldsymbol{A}\boldsymbol{z} = \lambda\boldsymbol{B}\boldsymbol{z}$ mit positiv definitem $\boldsymbol{B}$ ist im Fall voll besetzter Matrizen $\boldsymbol{A}, \boldsymbol{B}$ die explizite Reduktion auf ein spezielles symmetrisches Eigenwertproblem nach 14.2 der Standardweg. In den meisten Fällen kann dazu die möglichst mit Diagonalpivotisierung berechnete $\boldsymbol{L}\boldsymbol{D}\boldsymbol{L}^{\mathrm{T}}$-Faktorisierung von $\boldsymbol{B}$ verwendet werden; andernfalls sollte mit der stabileren Eigenwertzerlegung gearbeitet werden. Falls $\boldsymbol{B}$ semidefinit oder sehr schlecht konditioniert ist, ist die Reduktion nach Fix/Heiberger — siehe B 14.2 — zu empfehlen.

Wenn $\boldsymbol{A}$, $\boldsymbol{B}$ Bandmatrizen hoher Dimension sind, kann das Büschel $\{\boldsymbol{A}, \boldsymbol{B}\}$ unter Erhalt der Bandbreite kongruent in $\{\boldsymbol{T}, \boldsymbol{I}\}$ mit tridiagonalem $\boldsymbol{T}$ transformiert werden, siehe B 14.3. Da dann i. allg. nur wenige Eigenwerte gesucht sind, sollten diese durch Schneiden des Spektrums bestimmt werden, wobei auch direkt auf $\boldsymbol{A} - \mu\boldsymbol{B}$ zurückgegriffen werden kann. Letzteres trifft auch auf die inverse Iteration zur Ermittlung zugehöriger Eigenvektoren zu.

Für schwach besetzte Matrizen $\boldsymbol{A}$, $\boldsymbol{B}$ hoher Dimension gelten die Ausführungen für das spezielle Eigenwertproblem sinngemäß. Die direkte bzw. inverse Vektor- und Teilraumiteration kann direkt mit den Originaldaten durchgeführt werden, wobei bei der direkten Iteration Gleichungssysteme des Typs $\boldsymbol{B}\boldsymbol{w} = \boldsymbol{A}\boldsymbol{v}$, bei der inversen solche des Typs $(\boldsymbol{A} - \mu\boldsymbol{B})\,\boldsymbol{w} = \boldsymbol{A}\boldsymbol{v}$ für gegebenes $\boldsymbol{v}$ zu lösen sind. Bei Kombination mit dem RR-Algorithmus tritt pro Schritt zusätzlich ein allgemeines symmetrisches Eigenwertproblem niedriger Dimension auf, siehe 14.3. Auch hier stellen Relaxations- und Gradientenverfahren sowie die Verfahren vom Lanczos-Typ alternative Möglichkeiten dar, siehe B 13.5, B 13.7 und B 14.5.

Den Problemkreis der Rang-1-Modifikation symmetrischer Eigenwertaufgaben haben wir aus Platzgründen nicht diskutiert; siehe dazu Golub [73] und Bunch/Nielsen/Sorensen [78].

In manchen Anwendungen spielen sog. inverse Eigenwertprobleme eine Rolle, bei denen Matrizen $\boldsymbol{A}$ bzw. $\{\boldsymbol{A}, \boldsymbol{B}\}$ innerhalb gewisser Klassen so festzulegen sind, daß sie vorgegebene Eigenwerte besitzen. Wir weisen dazu auf Golub [73], Friedland [75], de Boor/Golub [78], Wang/Garbow [83], aber auch auf die entsprechenden Abschnitte bei Parlett [80a] und Golub/van Loan [83] hin.

Eigenwertzerlegungen können auch zur Berechnung von Matrixfunktionen $f(\boldsymbol{A})$ — etwa $e^{\boldsymbol{A}}$ — herangezogen werden; siehe Golub/van Loan [83] und speziell Moler/van Loan [78].

Eigenwertprobleme für nichtsymmetrische Matrizen sind von der Theorie wie von der Praxis wesentlich komplizierter als solche für symmetrische; wir verweisen auf WILKINSON [65] und GOLUB/VAN LOAN [83].

16. Software für Aufgabenklassen der numerischen linearen Algebra

Wir beginnen mit der — möglicherweise überspitzten — Feststellung, daß das Hauptziel der numerischen linearen Algebra die Schaffung von qualitativ hochwertiger numerischer Software ist. Unter *numerischer Software* soll dabei die Gesamtheit der Computerprogramme verstanden werden, in denen Algorithmen zur Lösung numerischer Probleme aus gewissen Problemklassen dem fortgeschrittenen Entwicklungsstand entsprechend realisiert worden sind. In diesem Sinne impliziert die Bezeichnung „numerische Software" einen Qualitätsanspruch, der natürlich relativ ist: Er hängt vom jeweiligen theoretischen Erkenntnisstand über die Problemklasse und die zu ihrer Lösung bekannten Algorithmen, von den Möglichkeiten der Computer, den verwendeten Programmiersprachen und Betriebssystemen, vom Stand der Programmiertechnologie und schließlich auch vom ins Auge gefaßten Nutzerkreis ab.

Die auf die Software wirkenden Faktoren weisen eine große Variationsbreite auf: Die zu einer Klasse gehörenden Probleme können in „kleine" und „große" eingeteilt werden. Dabei versteht man unter kleinen Problemen in der Regel solche, bei denen alle Eingangsdaten und Zwischenergebnisse im Hauptspeicher mit schnellem Zugriff Platz finden und wo eine mögliche schwache Besetztheit nicht systematisch ausgenutzt wird; die Einteilung hängt also vom Computer ab. Das Spektrum der Computer selbst reicht vom Ein-Chip-Rechner bis zu den imposanten Supercomputern mit mehreren 100 Millionen Gleitpunktoperationen pro Sekunde. Die Operationsgeschwindigkeiten können sich um Faktoren in der Größenordnung 10^4 bis 10^5 unterscheiden; dasselbe trifft in etwa auch für die Preise zu.

Die Mehrzahl der Nutzer numerischer Software sind Ingenieure, Naturwissenschaftler u. a., die bei der Behandlung ihrer fachspezifischen Aufgaben auf numerische Probleme treffen und diese möglichst einfach gelöst haben wollen. Sie verfügen in der Regel nicht über Spezialkenntnisse in Numerischer Mathematik und sehen die Software als „black box" an, die aus den Eingangsdaten in hoffentlich vernünftiger, sie im einzelnen jedoch nicht interessierender Weise die gewünschten Ausgangsdaten erzeugen soll. Der Kreis dieser Nutzer vergrößert sich in dem Maße, wie leistungsfähige Rechentechnik am Arbeitsplatz verfügbar gemacht und die Mathematik in den Fachdisziplinen mehr und mehr angewendet wird. Es ist klar, daß dieser große Nutzerkreis auf entsprechend konzipierte Software angewiesen ist. Den Gegenpol bildet die kleine Schar der hochspezialisierten Nutzer, die wesentlich komplexere Aufgaben bearbeiten und z. T. die derzeit verfügbaren Erkenntnisse der Theorie wie auch die Möglichkeiten der Computer bis zur Grenze ausschöpfen.

Für das Gebiet der numerischen linearen Algebra muß außerdem beachtet werden, daß die einzelnen Probleme i. allg. nicht eigenständig, sondern als Bestandteil umfangreicher technisch-wissenschaftlicher u. ä. Berechnungen auftreten. Die Eingangsdaten sind das Ergebnis vorhergehender Rechnungen, und die Ausgangsdaten werden im nachfolgenden Schritt als Eingangsdaten benutzt, ohne daß sie etwa über einen Drucker ausgegeben oder über einen Bildschirm angezeigt werden. Wegen dieses sog. verdeckten Datenflusses ist es zweckmäßig — wenn auch nicht immer nötig —, von den Ausgangsdaten die höchste erreichbare Genauigkeit zu verlangen.

Die bisherigen Ausführungen motivieren die folgenden Forderungen an numerische Software:

- *Zuverlässigkeit:* Alle Probleme, die zur Problemklasse gehören, für die der implementierte Algorithmus geeignet ist und die im Rahmen der Computerarithmetik auch gelöst werden können, sollen mit maximal möglicher Genauigkeit gelöst werden. Wünschenswert ist eine Information über die erreichte Genauigkeit. Die verwendeten Algorithmen sollten numerisch stabil und möglichst sogar numerisch gutartig sein; gegebenenfalls sollte ein Stabilitätsverlust durch Beobachtung der dafür charakteristischen Größen signalisiert werden. Sonderfälle sollten sachgemäß behandelt und eine erfolglose Bearbeitung mit Angabe von Gründen mitgeteilt werden.
- *Effektivität:* Die numerische Lösung soll mit möglichst geringer Rechenzeit und möglichst wenig zusätzlichem Speicherplatz berechnet werden. Fast immer sind Zuverlässigkeit und Effektivität gegensätzliche Forderungen: Die zuverlässigsten Algorithmen sind meist nicht die billigsten, und die korrekte Behandlung aller denkbaren Sonderfälle erfordert zusätzlichen Aufwand. Für die große Klasse der Nutzer, die nur wenige nicht zu große Probleme lösen wollen, spielt die Effektivität gegenüber der Zuverlässigkeit i. allg. eine untergeordnete Rolle. Sind dagegen sehr viele und/oder sehr große Probleme zu lösen, so wird die Forderung nach Effektivität dominieren.
- *Portabilität:* Die Software soll möglichst ohne bzw. mit wenigen Änderungen auf unterschiedlichen Computern und unter unterschiedlichen Betriebssystemen verwendbar sein. Da numerische Software heute fast ausschließlich in der Programmiersprache FORTRAN geschrieben wird, sollte eine solche Teilmenge dieser Sprache verwendet werden, die auf der Mehrzahl und insbesondere auch auf kleineren Computern verfügbar ist. Die Software sollte keine Ein- und Ausgabeoperationen enthalten; wenn solche nicht vermieden werden können, sollten sie über Unterprogramme realisiert werden, die dann i. allg. vom Computer und der genutzten Peripherie abhängen. Die Datenübermittlung sollte möglichst über Parameterlisten und nicht über COMMON-Bereiche erfolgen. Computerspezifische Größen wie ν, MAX, MIN usw. sollten nicht explizit verwendet werden. Der Kleinheitstest

  $$\textbf{if } (|\alpha| \leqq \nu * |\beta|) \tag{1}$$

  kann z. B. etwa gleichwertig in der Form

  $$\textbf{if } (|\alpha| + |\beta| = |\beta|) \tag{2}$$

ohne explizite Verwendung von v realisiert werden. Wenn doch gewisse Computerkonstanten benutzt werden müssen, sollten sie am Programmanfang durch eine DATA-Anweisung und damit leicht erkennbar gemacht werden, oder sie sollten durch Unterprogramme vermittelt werden, vgl. auch B 2.5.

— *Nutzerfreundlichkeit:* Dazu gehört eine verständliche und übersichtliche Dokumentation, und zwar möglichst auf zwei Niveaus: eins für den black-box-Nutzer, ein zweites für den erfahrenen Nutzer, der Anpassungen, Erweiterungen o. ä. vornehmen und daher im Detail informiert sein will. Letzteres wird durch eine ausreichende Erläuterung der Quelltexte durch Kommentare gefördert. Sofern die Auswahl der Verfahren nicht automatisch erfolgt, sollten Hinweise über die Leistungsfähigkeit und Anwendungsgebiete der implementierten Algorithmen gegeben werden. Programme, die für eine oder mehrere benachbarte Problemklassen gedacht sind, sollten nach einheitlichen Gesichtspunkten implementiert und dokumentiert werden, in verschiedenen Programmen auftretende Größen mit derselben oder ähnlicher Bedeutung sollten dieselben Bezeichnungen haben, die Parameterlisten sollten vereinheitlicht und nicht zu lang sein, die Bezeichnung der Teilprogramme und Parameter sollte nach einem durchschaubaren, einheitlichen Prinzip erfolgen.

Eine nach den letztgenannten Gesichtspunkten erarbeitete Zusammenstellung von aufeinander abgestimmten Programmen für die Lösung von Aufgaben aus gewissen Aufgabenklassen wird *Programmpaket* genannt. Programmpakete stellen die heute übliche Form dar, in der numerische Software ausgearbeitet, angeboten und genutzt wird. Wenn der Automatisierungsgrad durch Ermöglichen der anwendernahen Formulierung des Problems in einer angepaßten Spezialsprache, automatische Algorithmenauswahl, Übernahme der Datenorganisation und der Ein- und Ausgabe u. ä. weiter erhöht wird, spricht man von *Programmsystemen*. Die Erarbeitung solcher Programmsysteme ist allerdings außerordentlich aufwendig, so daß sie sich nur für komplexere und häufig auftretende Problemklassen wie etwa Schaltkreisentwurf, mechanische Berechnungen u. ä. lohnt.

Als erstes maßstabsetzendes Programmpaket zur numerischen linearen Algebra müssen die ALGOL-Prozeduren aus dem von WILKINSON/REINSCH [71] herausgegebenen und wesentlich mitgetragenen Band „Linear Algebra" des „Handbook for Automatic Computation" angesehen werden. Mit diesen Programmen wurde ein bis zu diesem Zeitpunkt nicht erreichter Stand demonstriert, der die weitere Entwicklung im Bereich der numerischen Software wesentlich beeinflußt hat. Auf der Grundlage der Eigenwertprogramme des „Handbooks" entstanden in den darauffolgenden Jahren die weiterentwickelten FORTRAN-Pakete EISPACK von SMITH et al. [74] und GARBOW et al. [77]; eine überarbeitete dritte, allerdings nicht in Buchform veröffentlichte Version liegt inzwischen auch vor. Die EISPACK-Routinen sind für das symmetrische und nichtsymmetrische Eigenwertproblem konzipiert. Den Problemklassen der linearen Gleichungssysteme und Quadratmittelprobleme ist das von DONGARRA et al. [79] erarbeitete Programmpaket LINPACK gewidmet. Hinsichtlich Sorgfalt der Implementierung, Berücksichtigung theoretischer Erkenntnisse und Programmierstil können die in den genannten Paketen enthaltenen

FORTRAN-Routinen als Musterbeispiel hochwertiger numerischer Software dienen und jedem, der sich mit der Implementierung numerischer Algorithmen befaßt, zum intensiven Studium empfohlen werden. Auch in der Ausbildung der Studenten ist es viel vernünftiger, sie etwa ein LINPACK-Programm verstehen und testen zu lassen als zu fordern, daß sie ein — dann in der Regel schlechtes — Programm zur Lösung des entsprechenden Problems selbst schreiben.

Eine Besonderheit des letztgenannten Paketes LINPACK ist die durchgängige Verwendung der sog. „Basic Linear Algebra Subprograms for FORTRAN Usage" — kurz: BLAS —, siehe B 16.2. Diese Unterprogramme realisieren die in der numerischen linearen Algebra häufigsten Grundoperationen wie Addition des Vielfachen eines Vektors zu einem anderen, Skalarproduktberechnung usw., wobei die Vektoren sowohl Spalten als auch Zeilen von Matrizen sein können, die in zweidimensionalen Feldern gespeichert sind. Der Ersatz der sonst benötigten FORTRAN-Schleifen durch die BLAS-Aufrufe führt zu einer Verkürzung der Programme. Außerdem besteht die Möglichkeit, die BLAS in der Maschinensprache zu kodieren und dabei Besonderheiten des Computers — etwa vorhandene Vektorprozessoren — auszunutzen, womit die Effektivität erhöht werden kann. Beim Übergang zu anderen Computern brauchen dann nur die BLAS neu geschrieben zu werden. Als Nachteil steht dem der durch den Organisationsaufwand bei den BLAS-Aufrufen bedingte Effektivitätsverlust bei der Behandlung von Problemen niedriger Dimension gegenüber, vgl. die entsprechenden Angaben in der LINPACK-Dokumentation.

Besondere Aufmerksamkeit muß bei Verwendung von FORTRAN den in der linearen Algebra typischen zweidimensionalen Feldern gewidmet werden. Da solche Felder spaltenweise abgespeichert werden, sollten die Algorithmen stets spaltenweise orientiert werden, d. h., der innerste Schleifenindex sollte in einer Spalte laufen. Ist die (M,N)-Matrix A z. B. in dem gleichnamigen Feld A der Dimension (M,N) gespeichert worden, so entspricht dem zweifach indizierten Element A(I,J) in der im Speicher vorhandenen linearen Anordnung das einfach indizierte Element A(L) mit

$$L = (J-1)*M + I, \tag{3}$$

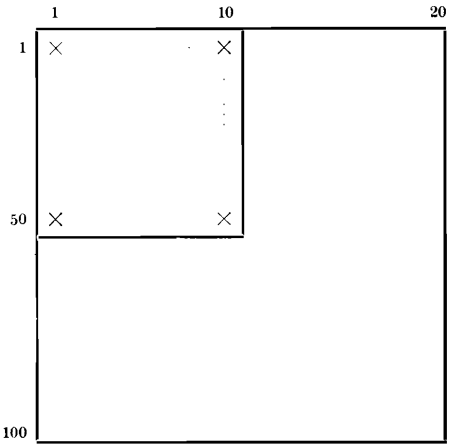
vgl. 1.1.C. Bei Zugriff auf die J-te Spalte werden die aufeinanderfolgenden Elemente A(L) mit L zwischen (J−1)*M + 1 und J*M benötigt; das Inkrement von L ist 1. Bei Zugriff auf die I-te Zeile wächst L dagegen von I bis (N−1)*M + I um das Inkrement M. Bei Betriebssystemen mit virtuellem Speicher kann dies zu häufigem Seitenwechsel und daher zu großem Effektivitätsverlust führen.

Das zweite Problem entsteht, wenn mit einem Programm mehrere Probleme mit Matrizen unterschiedlicher Dimensionen bearbeitet werden sollen, was in den Anwendungen meist der Fall ist. Da die Vereinbarung von Feldern variabler Dimension in FORTRAN nicht zulässig ist, muß am Anfang des Hauptprogramms ein Feld mit maximaler Dimension vereinbart werden, das alle aktuellen Matrizen aufnimmt, etwa durch die Vereinbarung

```
DIMENSION A(100,20).
```
(4)

In diesem Feld A können (M,N)-Matrizen — sie mögen ebenfalls mit A bezeichnet sein — mit $M \leqq 100$, $N \leqq 20$ gespeichert werden. Als Beispiel betrachten wir eine (50, 10)-Matrix $\boldsymbol{A}$, die im Hauptprogramm durch die Anweisungen

```
      M=50
      N=10
      DO  20   I=1,M
          DO  10  J=1,N
              A(I,J)=...
10            CONTINUE
20        CONTINUE
```



erzeugt werden möge. Im danebenstehenden Schema sind die dadurch belegten Teile des Feldes A wie üblich gekennzeichnet worden. In linearer Reihenfolge stehen also am Anfang die 50 Elemente der ersten Spalte, dann kommen 50 nicht belegte Plätze, dann die 50 Elemente der zweiten Spalte, wieder 50 „Lücken" usw. Einem Unterprogramm — etwa einer SUBROUTINE QRFACT zur $\boldsymbol{QR}$-Faktorisierung der Matrix $\boldsymbol{A}$ nach HOUSEHOLDER — muß dann neben der aktuellen Zeilenzahl M=50 und Spaltenzahl N=10 noch die erste Dimension (engl. „leading dimension") LDA=100 des tatsächlich gemäß (4) vereinbarten Feldes mitgeteilt werden. Das Unterprogramm müßte also mit

```
SUBROUTINE QRFACT(A,LDA,M,N,...)          (5)
DIMENSION A(LDA,1)                        (6)
```

beginnen, und im Hauptprogramm muß dem aktuellen Parameter LDA vor Aufruf von QRFACT der Wert 100 durch die Anweisung

```
LDA=100   bzw.   DATA LDA/100/            (7)
```

zugewiesen werden. Man beachte, daß durch (6) kein Feld vereinbart wird, sondern dem Compiler wird lediglich mitgeteilt, daß A ein zweidimensionales Feld der Zeilenzahl LDA ist, so daß im Unterprogramm der Index von A(I,J) gemäß (3) mit LDA statt M richtig ausgewertet wird. Da die zweite Dimension — die Spaltenzahl des vereinbarten Feldes — in (3) überhaupt nicht vorkommt und daher für die Indexauswertung nicht benötigt wird, kann in (6) für sie 1 gesetzt werden.

Bei IBM- und ESER-Computern wird statt (6) häufig

```
REAL A(LDA,1)
```

geschrieben; bei doppelter Genauigkeit ist

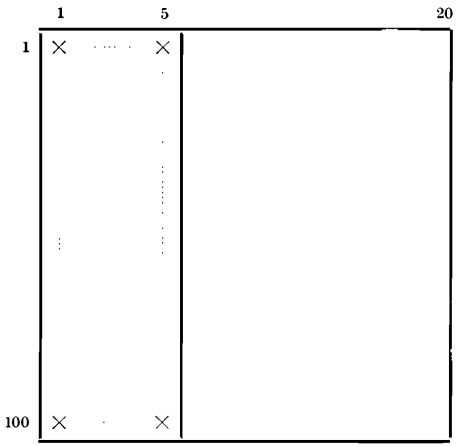
DOUBLE PRECISION A(LDA,1) bzw. REAL*8 A(LDA,1)

zu verwenden.

Würde das Unterprogramm in naiver Weise in der Form

SUBROUTINE QRFACT(A,M,N,...)
DIMENSION A(M,1)

programmiert werden, so müßten die Elemente der (50,10)-Matrix A spaltenweise nacheinander auf den ersten 500 Plätzen des Feldes A stehen, d. h., das Feld müßte nach dem Schema



belegt werden. Im Hauptprogramm könnte dann nicht mit zweifacher Indizierung gearbeitet werden, da die dort gültige Auswertungsformel (3) mit LDA = 100 statt M nicht der oben angegebenen Belegung entspricht. Zum Beispiel wird das im Unterprogramm mit A(1,2) erreichbare Element im Hauptprogramm als A(51,1) geführt. Die Indexrechnung im Hauptprogramm müßte also vom Bearbeiter selbst in eindimensionaler Weise organisiert werden, was umständlich ist, zu unübersichtlichen Programmen führt und besonders für Anfänger eine beliebte Fehlerquelle darstellt.

Wir bemerken abschließend, daß die oben zitierten Programmpakete für Probleme mit voll besetzten Matrizen — bei Gleichungssystemen und Eigenwertaufgaben auch für solche mit Bandmatrizen — konzipiert wurden. Zur Software für Probleme mit großen und schwach besetzten Matrizen sei auf B 16.4 verwiesen.

Bemerkungen zum Kapitel 16

B 16.1. Es gibt inzwischen eine umfangreiche Literatur über numerische Software und speziell über Software zur numerischen linearen Algebra. Wir zitieren lediglich die Sammelbände von Rice [77], Jacobs [78], Fosdick [79] und Cowell [84] sowie das informative, auch für

Anfänger geeignete Buch von RICE [81]. Zum Problem der Portabilität — und andere interessante Fragen — sei auf COWELL [76] verwiesen. Auf die in den letzten Jahren sehr stark in den Blickpunkt des Interesses getretenen Vektor- und Parallelcomputer gehen u. a. HELLER [78], RODRIGUE [82], DONGARRA et al. [84] und KOTOV/MIKLOŠKO [84] ein. In jüngster Zeit werden auch sog. „array processors" und „systolic arrays" untersucht; als Beispiele seien etwa die Arbeiten von HELLER/IPSEN [83] und BOJANCZYK/BRENT/KUNG [84] genannt.

B 16.2. FORTRAN-Programme für Quadratmittelprobleme mit voll besetzten Matrizen sind außer im LINPACK auch bei LAWSON/HANSON [74] zu finden. Für Eigenwertaufgaben mit voll besetzten und Bandmatrizen wurden von LANG et al. [79] auf der Grundlage von EISPACK das weiterentwickelte Programmsystem MEIWEP erarbeitet; siehe auch LANG [82]. Die BLAS gehen auf LAWSON et al. [79] zurück; sie sind auch als Anhang im LINPACK aufgelistet.

B 16.3. Eine Vorstellung über die Zentralprozessorzeit, die zur Lösung eines Gleichungssystems der Dimension 100 mittels der LINPACK-Implementierung des Gaußschen Algorithmus erforderlich ist, vermitteln die von DONGARRA [84] zusammengestellten Tabellen, wo eine Vielzahl von Computern vom Apple II bis zur CRAY-XM erfaßt worden ist. Instruktive Zeitvergleiche sind auch in den LINPACK- und EISPACK-Dokumentationen enthalten.

B 16.4. Eine Übersicht über Software für Probleme mit schwach besetzten Matrizen gibt DUFF [84]. Zu den am meisten verwendeten Paketen zur Lösung von Gleichungssystemen mit schwach besetzten, symmetrischen und positiv definiten Matrizen gehören das von GEORGE/LIU [81] entwickelte SPARSPAK und das auf EISENSTAT et al. [77] zurückgehende Yale Sparse Matrix Package. Falls die Matrix indefinit ist und aus Stabilitätsgründen auch (2,2)-Pivots zugelassen werden müssen — vgl. 6.1.B —, sind die Programme von DUFF/REID [82] zu empfehlen. Für nichtsymmetrische Systeme erwähnen wir wieder das Yale Sparse Matrix Package, die von DUFF/REID [77] erarbeiteten Routinen und die darauf aufbauenden Entwicklungen von ZLATEV/WASNIEWSKI/SCHAUMBURG [81] und ØSTERBY/ZLATEV [83]. Zur Software für große Eigenwertaufgaben sei außer auf DUFF [84] noch auf die Übersichten von STEWART [76a] und PARLETT [84] sowie auf CULLUM/WILLOUGHBY [85] und die dort zitierte Literatur verwiesen, für große Quadratmittelprobleme auf die am Schluß des Kapitels 12 erwähnten Arbeiten.

Literatur

Aasen, J. O. (71). On the reduction of a symmetric matrix to tridiagonal form. BIT **11**, 233 to 242.

Babuška, I. (69). Numerical stability in numerical analysis, in: Information processing 68. North-Holland, Amsterdam, 11–23.

Babuška, I. (72). Numerical stability in problems of linear algebra. SIAM J. Numer. Anal. **9**, 53–77.

Bard, Y. (73). Nonlinear parameter estimation. Academic Press, New York. Russ. Übers.: Nelineĭnoe ocenivanie parametrov. Statistika, Moskva 1979.

Bargmann, V.; Montgomery, C.; von Neumann, J. (46). Solution of linear systems of high order. Institute for Advanced Study, Princeton.

Bartels, R. H.; Stoer, J.; Zenger, C. (71). A realization of the simplex method based on triangular decompositions, in: Wilkinson, J. H.; Reinsch, C. (Hrsg.), 152–190.

Barwell, V.; George, A. (76). A comparison of algorithms for solving symmetric indefinite systems of linear equations. ACM Trans. Math. Software **2**, 242–251.

Bathé, K. J.; Wilson, E. (76). Numerical methods in finite element analysis. Prentice-Hall, Englewood Cliffs, N. J.

Bauer, F. L. (57). Das Verfahren der Treppeniteration und verwandte Verfahren zur Lösung algebraischer Eigenwertprobleme. Z. Angew. Math. Phys. **8**, 214–235.

Bauer, F. L. (63). Optimally scaled matrices. Numer. Math. **5**, 73–87.

Bauer, F. L. (66). Genauigkeitsfragen bei der Lösung linearer Gleichungssysteme. Z. Angew. Math. Mech. **46**, 409–421.

Bauer, F. L. (69). Remarks on optimally scaled matrices. Numer. Math. **13**, 1–3.

Bauer, F. L. (74). Computational graphs and rounding errors. SIAM J. Numer. Anal. **11**, 87–96.

Bennett, J. M. (65). Triangular factors of modified matrices. Numer. Math. **7**, 217 to 221.

Bersenev, S. M. (79). O peresčete faktorizacii Holeckogo. Ž. Vyčisl. Mat. i Mat. Fiz. **19**, 1318–1319.

Björck, A. (67a). Solving least squares problems by Gram-Schmidt orthogonalization. BIT **7**, 1–21.

Björck, A. (67b, 68). Iterative refinement of linear least squares solutions I, II. BIT **7**, 257–278; **8**, 8–30.

Björck, A. (76). Methods for sparse least squares problems, in: Bunch, J. R.; Rose, D. J. (Hrsg.), 177–199.

Björck, A. (78). Comment on the iterative refinement of least squares solutions. J. Amer. Statist. Assoc. **73**, 161–166.

Björck, A. (81). Least squares methods in physics and engineering. Report CERN 81–16, Geneva.

Björck, A.; Golub, G. H. (67). Iterative refinement of linear least squares solutions by Householder transformations. BIT **7**, 322–337.

Blue, J. L. (78). A portable FORTRAN program to find the Euclidean norm of a vector. ACM Trans. Math. Software **4**, 15–23.

Bojanczyk, A.; Brent, R. P.; Kung, H. T. (84). Numerically stable solution of dense systems of linear equations using mesh-connected processors. SIAM J. Sci. Statist. Comput. **5**, 95–104.

Brown, W. S. (81). A simple but realistic model of floating-point computation. ACM Trans. Math. Software **7**, 445–480.

Bunch, J. R. (71). Analysis of the diagonal pivoting method. SIAM J. Numer. Anal. **8**, 656–680.

Bunch, J. R.; Kaufman, L. (77). Some stable methods for calculating inertia and solving symmetric linear systems. Math. Comp. **31**, 162–179.

Bunch, J. R.; Kaufman, L.; Parlett, B. N. (76). Decomposition of a symmetric matrix. Numer. Math. **27**, 95–109.

Bunch, J. R.; Nielsen, C. P.; Sorensen, D. C. (78). Rank-one modification of the symmetric eigenproblem. Numer. Math. **31**, 31–48.

Bunch, J. R.; Parlett, B. N. (71). Direct methods for solving symmetric indefinite systems of linear equations. SIAM J. Numer. Anal. **8**, 639–655.

Bunch, J. R.; Rose, D. J. (Hrsg.) (76). Sparse matrix computations. Academic Press, New York.

Bunse, W.; Bunse-Gerstner, A. (85). Numerische lineare Algebra. Teubner, Stuttgart.

Burmeister, W. (76). Ein Verfahren zur Modifizierung von *LU*-Zerlegungen bei Spaltenersetzung. Unveröffentl. Manuskript.

Businger, P. (68). Matrices which can be optimally scaled. Numer. Math. **12**, 346–348.

Businger, P.; Golub, G. H. (65). Linear least squares solutions by Householder transformations. Numer. Math. **7**, 269–276. Auch in: Wilkinson, J. H.; Reinsch, C. (Hrsg.), 111–118.

Chan, T. F. (82). An improved algorithm for computing the singular value decomposition. ACM Trans. Math. Software **8**, 72–83.

Cline, A. K.; Conn, A. R.; van Loan, C. (82). Generalizing the LINPACK condition estimator, in: Hennart, J. P. (Hrsg.), Numerical analysis. Lecture Notes in Mathematics, Vol. 909. Springer, Berlin, 73–83.

Cline, A. K.; Moler, C. B.; Stewart, G. W.; Wilkinson, J. H. (79). An estimate for the condition number of a matrix. SIAM J. Numer. Anal. **16**, 368–375.

Cline, A. K.; Rew, R. K. (83). A set of counter-examples to three condition number estimators. SIAM J. Sci. Statist. Comput **4**, 602–611.

Cowell, W. R. (Hrsg.) (76). Portability of numerical software. Lecture Notes in Computer Science, Vol. 57. Springer, Berlin.

Cowell, W. R. (Hrsg.) (84). Sources and development of mathematical software. Prentice-Hall, Englewood Cliffs, N. J.

Crawford, C. R. (73). Reduction of a band-symmetric generalized eigenvalue problem. Comm. ACM **16**, 41–44.

Cullum, J. K.; Willoughby, R. A. (85). Lanczos algorithms for large symmetric eigenvalue computations. Vol. I: Theory, Vol. II: Programs. Progress in Scientific Computing, Vol. 3, 4. Birkhäuser, Basel.

Daniel, J.; Gragg, W. B.; Kaufman, L.; Stewart, G. W. (76). Reorthogonalization and stable algorithms for updating the Gram-Schmidt *QR* factorization, Math. Comp. **30**, 772–795.

Davis, C.; Kahan, W. M. (70). The rotation of eigenvectors by a perturbation. III. SIAM J. Numer. Anal. **7**, 1–46.

Dax, A. (83). A diagonal modification for the downdating algorithm. SIAM J. Sci. Statist. Comput. **4**, 85–93.

DE BOOR, C.; GOLUB, G. H. (78). The numerically stable reconstruction of a Jacobi matrix from spectral data. Linear Algebra Appl. **21**, 245—260.

DÖHLER, B. (82). Ein neues Gradientenverfahren zur simultanen Berechnung der kleinsten und größten Eigenwerte des allgemeinen Eigenwertproblems. Numer. Math. **40**, 79—91.

DONGARRA, J. J. (84). Performance of various computers using standard linear equations software in a FORTRAN environment. Technical Memorandum MCS-TM-23, Argonne National Laboratory, Argonne.

DONGARRA, J. J.; BUNCH, J. R.; MOLER, C. B.; STEWART, G. W. (79). LINPACK. Users guide. SIAM Publications, Philadelphia.

DONGARRA, J. J.; GUSTAVSON, F. G.; KARP, A. (84). Implementing linear algebra algorithms for dense matrices on a vector pipeline machine. SIAM Rev. **26**, 91—112.

DONGARRA, J. J.; MOLER, C. B.; WILKINSON, J. H. (83). Improving the accuracy of computed eigenvalues and eigenvectors. SIAM J. Numer. Anal. **20**, 23—45.

DRAPER, N. R.; SMITH, H. (66). Applied regression analysis. J. Wiley, New York.

DUFF, I. S. (Hrsg.) (81). Sparse matrices and their use. Academic Press, New York.

DUFF, I. S. (84). A survey of sparse matrix software, in: COWELL, W. R. (Hrsg.).

DUFF, I. S.; REID, J. K. (76). A comparison of some methods for the solution of sparse overdetermined systems of linear equations. J. Inst. Math. Appl. **17**, 267—280.

DUFF, I. S.; REID, J. K. (77). MA 28 — A set of FORTRAN-subroutines for sparse unsymmetric matrices. Report R-8730, Computer Science and Systems Division, AERE Harwell.

DUFF, I. S.; REID, J. K. (82). MA 27 — A set of FORTRAN-subroutines for solving sparse symmetric sets of linear equations. Report R-10533, Computer Science and Systems Division, AERE Harwell.

DUFF, I. S.; STEWART, G. W. (Hrsg.) (79). Sparse matrix proceedings 1978. SIAM Publishers, Philadelphia.

EFROYMSON, M. A. (60). Multiple regression analysis, in: RALSTON, A.; WILF, H. (Hrsg.), Mathematical methods for digital computers, Vol. 1, J. Wiley, New York. Dtsch. Übers.: Mathematische Methoden für Digitalrechner, Bd. 1. Oldenbourg, München 1963.

EISENSTAT, S. C.; GURSKY, M. C.; SCHULTZ, M. H.; SHERMAN, A. H. (77). Yale sparse matrix package. I. The symmetric codes. Research Report no. 112, II. The nonsymmetric codes. Research Report no. 114. Department of Computer Science, Yale University, New Haven.

ELDÉN, L. (72). Stepwise regression analysis with orthogonal transformations. Report 2-1972. Department of Applied Mathematics, Linköping Institute of Technology, Linköping.

ELDÉN, L. (77). Algorithms for the regularization of illconditioned least squares problems. BIT **17**, 134—145.

ELSNER, L.; SUN, J. (82). Perturbation theorems for the generalized eigenvalue problem. Linear Algebra Appl. **48**, 341—357.

FADDEEV, D. K.; FADDEEVA, V. N. (63). Vyčislitel'nye metody lineĭnoĭ algebry. Fizmatgiz, Moskva. Dtsch. Übers.: Numerische Methoden der linearen Algebra. 5. Auflage, VEB Deutscher Verlag der Wissenschaften, Berlin; Oldenbourg, München—Wien 1978.

FADDEEV, D. K.; FADDEEVA, V. N. (69). Stability in linear algebra problems, in: Information processing 68. North-Holland, Amsterdam, 33—39.

FADDEEV, D. K.; FADDEEVA, V. N. (70). Natural norms in algebraic processes. SIAM J. Numer. Anal. **7**, 520—531.

FADDEEV, D. K.; KUBLANOVSKAJA, V. N.; FADDEEVA, V. N. (68). Sur les Systèmes Linéaires Algébriques de Matrices Rectangularies et Mal-Conditionnées, in: Programmation en Mathématiques Numériques VII. Editions Centre Nat. Recherche Sci., Paris, 161—170.

FADDEEVA, V. N.; IKRAMOV, H. D. et al. (82). Vyčislitel'nye metody lineĭnoĭ algebry. Bibliografičeskiĭ ukasatel' 1975—1980. Leningradskoe Otdelenie Mat. Inst. Steklova, Leningrad.

FADDEEVA, V. N.; KUZNECOV, JU. A. et al. (76). Vyčislitel'nye metody lineĭnoĭ algebry. Bibliografičeskiĭ ukasatel' 1828—1974. Leningradskoe Otdelenie Mat. Inst. Steklova, Vyčisl. Centr Sibirskogo Otdelenija Akad. Nauk SSSR, Novosibirsk.

Fenner, T. I.; Loizou, G. (77). Optimally scalable matrices. Philos. Trans. Roy. Soc. London Ser. A **287**, no. 1345, 307–349.

Fix, G.; Heiberger, R. (72). An algorithm for the ill-conditioned generalized eigenvalue problem. SIAM J. Numer. Anal. **9**, 78–88.

Fletcher, R.; Matthews, S. P. J. (83). A stable algorithm for updating triangular factors under a rank one change. Report NA/69. Department of Mathematical Sciences, University of Dundee, Dundee.

Fletcher, R.; Powell, M. J. D. (74). On the modification of LDL^T factorizations. Math. Comp. **29**, 1067–1087.

Forsythe, G. E.; Moler, C. B. (67). Computer solutions of linear algebraic systems. Prentice Hall, Englewood Cliffs, N. J. Dtsch. Übers.: Computer-Verfahren für lineare algebraische Systeme. Oldenbourg, München 1971. Russ. Übers.: Čislennoe rešenie sistem lineĭnyh algebraičeskyh uravnenii. Mir, Moskva, 1969.

Forsythe, G. E.; Straus, E. G. (55). On best conditioned matrices. Proc. Amer. Math. Soc. **6**, 340–345.

Fosdick, L. (Hrsg.) (79). Performance evaluation of numerical software. North-Holland, Amsterdam.

Francis, J. G. F. (61). The QR transformation – a unitary analogue to the LR transformation. Comput. J. **4**, 265–271, 332–345.

Friedland, S. (75). On inverse multiplicative eigenvalue problems for matrices. Linear Algebra Appl. **12**, 127–138.

Friedrich, V.; Hofmann, B.; Tautenhahn, U. (79). Möglichkeiten der Regularisierung bei der Auswertung von Meßdaten. Schriftenreihe der TH Karl-Marx-Stadt **10**, 1–42.

Gantmaher, F. R. (66). Teorija matric. Izd. vtoroe. Nauka, Moskva. Dtsch. Übers.: Matrizentheorie. VEB Deutscher Verlag der Wissenschaften, Berlin 1986.

Garbow, B. S.; Boyle, J. M.; Dongarra, J. J.; Moler, C. B. (77). Matrix eigensystem routines – EISPACK guide extension. Lecture Notes in Computer Science, Vol. 51. Springer, Berlin.

Gentleman, W. M. (73). Least squares computations by Givens transformations without square roots. J. Inst. Math. Appl. **12**, 329–336.

Gentleman, W. M. (75). Error analysis of QR decompositions by Givens transformations. Linear Algebra Appl. **10**, 189–197.

Gentleman, W. M.; Marovich, S. B. (74). More on algorithms that reveal properties of floating point arithmetic units. Comm. ACM **17**, 276–277.

George, A.; Liu, J. W. H. (81). Computer solution of large positive definite systems. Prentice Hall, Englewood Cliffs, N. J. Russ. Übers.: Čislennoe rešenie bol'ših razrežennyh sistem uravnenii. Mir, Moskva 1984.

George, A.; Heath, M. T.; Ng, E. (83). A comparison of some methods for solving sparse linear least squares problems. SIAM J. Sci. Statist. Comput. **4**, 177–187.

George, A.; Heath, M. T.; Plemmons, R. J. (81). Solution of large-scale sparse least squares problems using auxiliary storage. SIAM J. Sci. Statist. Comput. **2**, 416–429.

Geradin, M. (71). The computational efficiency of a new minimization algorithm for eigenvalue analysis. J. Sound Vibration **19**, 319–331.

Gill, P. E.; Golub, G. H.; Murray, W.; Saunders, M. A. (74). Methods for modifying matrix factorizations. Math. Comp. **28**, 505–535.

Gill, P. E.; Murray, W. (77). Modification of matrix factorizations after a rank-one change, in: Jacobs, D. E. (Hrsg.), 55–83.

Gill, P. E.; Murray, W.; Saunders, M. A. (75). Methods for computing and modifying the LDV factors of a matrix. Math. Comp. **29**, 1051–1077.

Gill, P. E.; Murray, W.; Wright, M. H. (82). Practical optimization. Academic Press, New York.

Givens, W. (53). A method for computing eigenvalues and eigenvectors suggested by classical results on symmetric matrices. Nat. Bur. Standards Appl. Math. Ser. **29**, 117–122.

Givens, W. (54). Numerical computation of the characteristic values of a real symmetric matrix. Report ORNL-1574, Oak Ridge National Laboratory, Oak Ridge.

GIVENS, W. (58). Computation of plane unitary rotations transforming a general matrix to triangular form. SIAM J. Appl. Math. **6**, 26—50.

GOLDSTINE, H. H.; MURRAY, F. J.; VON NEUMANN, J. (59). The Jacobi method for real symmetric matrices. J. Assoc. Comput. Mach. **6**, 59—96.

GOLUB, G. H. (65). Numerical methods for solving linear least squares problems. Numer. Math. **7**, 206—216.

GOLUB, G. H. (69). Matrix decompositions and statistical computation, in: MILTON, R. C.; NELDER, J. A. (Hrsg.), Statistical computation. Academic Press, New York.

GOLUB, G. H. (73). Some modified matrix eigenvalue problems. SIAM Rev. **15**, 318—334.

GOLUB, G. H.; HEATH, M.; WAHBA, G. (79). Generalized cross-validation as a method for choosing a good ridge parameter. Technometrics **21**, 215—223.

GOLUB, G. H.; KAHAN, W. (65). Calculating the singular values and pseudoinverse of a matrix. SIAM J. Numer. Anal. **2**, 205—224.

GOLUB, G. H.; KAHAN, W. (68). Calculating the singular values and pseudoinverse of a matrix. Appl. Mat. **13**, 44—51.

GOLUB, G. H.; REINSCH, C. (70). Singular value decomposition and least squares solutions. Numer. Math. **14**, 403—420. Auch in: WILKINSON, J. H.; REINSCH, C. (Hrsg.), 134—151.

GOLUB, G. H.; STYAN, G. P. (73). Numerical computations for univariate linear models. J. Statist. Comput. Simulation **2**, 253—274.

GOLUB, G. H.; VAN LOAN, C. F. (80). An analysis of the total least squares problem. SIAM J. Numer. Anal. **17**, 883—893.

GOLUB, G. H.; VAN LOAN, C. F. (83). Matrix computations. North Oxford Academic, Oxford.

GOLUB, G. H.; WILKINSON, J. H. (66). Note on the iterative refinement of least squares solution. Numer. Math. **9**, 139—148.

GREGORY, R. T.; KRISHNAMURTHY, E. V. (84). Methods and applications of error-free computation. Springer, Berlin.

GRIMES, R. G.; LEWIS, J. G. (81). Condition number estimation for sparse matrices. SIAM J. Sci. Statist Comput. **2**, 384—388.

HACKBUSCH, W.; TROTTENBERG, U. (Hrsg.) (82). Multigrid methods. Lecture Notes in Mathematics, Vol. 960. Springer, Berlin.

HAGEMAN, L. A.; YOUNG, D. M. (81). Applied iterative methods. Academic Press, New York. Russ. Übers.: Prikladnye iteracionnye metody. Mir, Moskva 1986.

HAMMARLING, S. (74). A note on modifications of Givens plane rotation. J. Inst. Math. Appl. **13**, 215—218.

HASKELL, K. H.; HANSON, R. J. (81). An algorithm for linear least squares problems with equality and nonnegativity constraints. Math. Programming **21**, 98—118.

HEATH, M. T. (83). Numerical methods for large sparse linear least squares problems. Report ORNL/CSD-114, Oak Ridge National Laboratory, Oak Ridge.

HEINRICH, H. (63). Bemerkungen zu einem Konditionsmaß für lineare Gleichungssysteme. Z. Angew. Math. Mech. **43**, 568.

HELLER, D. E.; IPSEN, I. C. F. (83). Systolic networks for orthogonal decompositions. SIAM J. Sci. Statist. Comput. **4**, 261—269.

HELLER, P. (78). A survey of parallel algorithms in numerical linear algebra. SIAM Rev. **20**, 740—777.

HENRICI, P. (58). On the speed of convergence of cyclic and quasicyclic Jacobi methods for computing eigenvalues of Hermitian matrices. J. Soc. Industr. Appl. Math. **6**, 144—162.

HOFFMANN, W.; PARLETT, B. N. (78). A new proof of global convergence for the tridiagonal *QL* algorithm. SIAM J. Numer. Anal. **15**, 929—937.

HOFMANN, B. (86). Regularization for applied inverse and ill-posed problems. Teubner, Leipzig.

HOUSEHOLDER, A. S. (58). Unitary triangularization of a nonsymmetric matrix. J. Assoc. Comput. Mach. **5**, 339—342.

HOUSEHOLDER, A. S. (64). The theory of matrices in numerical analysis. Blaisdell, New York.

HOUSEHOLDER, A. S.; BAUER, F. L. (59). On certain methods for expanding the characteristic polynomial. Numer. Math. **1**, 29—37.

Humak, K. M. S. (77/83). Statistische Methoden der Modellbildung. Bd. I/II. Akademie-Verlag, Berlin.

Ikramov, H. D. (82). Razrežennye matricy. Itogi Nauki i Tehniki. VINITI, Ser. Mat. Analiz, 179–257.

Ikramov, H. D. (84). Čislennoe rešenie matričnyh uravnenií. Ortogonal'nye metody. Nauka, Moskva.

Jacobi, C. G. J. (1846). Über ein leichtes Verfahren, die in der Theorie der Säcularstörungen vorkommenden Gleichungen numerisch aufzulösen. J. Reine Angew. Math. **30**, 51–95.

Jacobs, D. E. (Hrsg.) (77). The state of the art in numerical analysis. Academic Press, London.

Jacobs, D. A. H. (Hrsg.) (78). Numerical software – needs and availability. Academic Press, London.

Jankowski, M.; Woźniakowski, H. (77). Iterative refinement implies numerical stability. BIT **17**, 303–311.

Jennings, A. (77). Matrix computations for engineers and scientists. J. Wiley, London.

Kågström, B.; Ruhe, A. (Hrsg.) (83). Matrix pencils. Lecture Notes in Mathematics, Vol. 973. Springer, Berlin.

Kahan, W. (66). Numerical linear algebra. Canad. Math. Bull. **9**, 757–801.

Kahan, W. (71). A survey of error analysis, in: Information processing 71. North-Holland, Amsterdam, 1214–1239.

Kiełbasiński, A. (78). Basic concepts in numerical error analysis, in: Mathematical models and numerical methods. Banach Center Publications. Vol. 3. PWN, Warszawa.

Kotov, V. J.; Mikloško, J. (Hrsg.) (84). Algorithms, software, and hardware of parallel computers, VEDA, Bratislava; Springer, Berlin.

Kublanovskaja, V. N. (61). O nekotoryh algoritmah dlja rešenija polnoï problemy sobstvennyh značeniï, Ž. Vyčisl. Mat. i Mat. Fiz. **1**, 555–570.

Kublanovskaja, V. N. (78a). K rešeniju spektral'noï zadači dlja singuljarnogo pučka matric. Ž. Vyčisl. Mat. i Mat. Fiz. **18**, 1056–1060.

Kublanovskaja, V. N. (78b). Primenenie normalizovannogo processa k postroeniju algoritmov rešenija spektral'nyh zadač dlja pučkov matric, in: Proceedings of the Fourth Sympos. on Basic Problems of Numerical Mathematics. Prague.

Kuznecov, Ju. A. (83). Metod soprjažennyh gradientov, ego obobščenija i primenenija, in: Marčuk, G. I. (Hrsg.), 267–301.

Lanczos, C. (50). An iteration method for the solution of the eigenvalue problem of linear differential and integral operators. J. Res. Nat. Bur. Standards, Sect. B **45**, 255–282.

Lang, W. et al. (79). Dokumentation zum Programmsystem MEIWEP. Teil 1: Dokumentation des Gesamtsystems. Teil 2: Dokumentation der Moduln (Heft 1: Symmetrische Matrizen, Heft 2: Reell nichtsymmetrische Matrizen, Heft 3: Komplexe Matrizen). TH Karl-Marx-Stadt, Sektion Mathematik, Karl-Marx-Stadt.

Lang, W. (82). Ein Programmsystem für Matrizeneigenwertaufgaben. Akademie-Verlag, Berlin.

Lawson, C. L.; Hanson, R. J. (74). Solving least squares problems. Prentice-Hall, Englewood Cliffs, N. J. Russ. Übers.: Čislennoe rešenie zadač metoda naimen'ših kvadratov. Mir, Moskva 1986.

Lawson, C. L.; Hanson, R. J.; Kincaid, D. R.; Krogh, F. T. (79). Basic linear algebra subprograms for FORTRAN usage. ACM Trans. Math. Software **5**, 308–323.

Lemeire, F. (75). Bounds for condition numbers of triangular and trapezoid matrices. BIT **15**, 58–64.

Levenberg, K. (44). A method for the solution of certain nonlinear problems in least squares. Quart. Appl. Math. **2**, 164–168.

Linnik, I. V. (62). Metod naimen'ših kvadratov i osnovy matematiko-statističeskoï teorii

obrabotki nabljudenií. Fizmatgiz, Moskva. Dtsch. Übers. d. 1. Auflage: Die Methode der kleinsten Quadrate in moderner Darstellung. Deutscher Verlag der Wissenschaften, Berlin 1961.

LÖTSTEDT, P. (84). Solving the minimal least squares problem subject to bounds on the variables. BIT **24**, 206–224.

LONGSINE, D. E.; MCCORMICK, S. F. (80). Simultaneous Rayleigh-quotient minimization methods for $Ax = \lambda Bx$. Linear Algebra Appl. **34**, 195–234.

MAESS, G. (84). Vorlesungen über Numerische Mathematik. Bd. 1. Akademie-Verlag, Berlin.

MALCOLM, M. (72). Algorithms to reveal properties of floating-point arithmetic. Comm. ACM **15**, 949–951.

MARČUK, G. I. (Hrsg.) (83). Vyčislitel'nye processy i sistemy. Vyp. 1. Nauka, Moskva.

MARKOWITZ, H. M. (57). The elimination form of the inverse and its application to linear programming. Management Sci. **3**, 255–269.

MCCORMICK, S. F.; NOE, T. (77). Simultaneous iteration for the matrix eigenvalue problem. Linear Algebra Appl. **16**, 43–56.

MEYER, A. (87). Modern algorithms for large sparse eigenvalues problems. Akademie-Verlag, Berlin.

MILLER, W. (75). Computational complexity and numerical stability. SIAM J. Comput. **4**, 97–107.

MILLER, W.; WRATHALL, C. (80). Software for roundoff analysis of matrix algorithms. Academic Press, New York.

MOLER, C. B. (67). Iterative refinement in floating point. J. Assoc. Comput. Mach. **14**, 316–371.

MOLER, C. B.; STEWART, G. W. (73). An algorithm for generalized matrix eigenvalue problems. SIAM J. Numer. Anal. **10**, 241–256.

MOLER, C. B.; VAN LOAN, C. F. (78). Nineteen dubious ways to compute the exponential of a matrix. SIAM Rev. **20**, 801–836.

MOORE, R. E. (79). Methods and applications of interval analysis. SIAM Publications, Philadelphia.

MOROZOV, V. A. (66). O reguljarizacii nekorrektno postavlennyh zadač i vybore parametra reguljarizacii. Ž. Vyčisl. Mat. i Mat. Fiz. **6**, 170–175.

MOROZOV, V. A. (73). O principe nevjaski pri rešenii nesovmestnyh uravnenií metodom reguljarizacii A. N. Tyhonova. Ž. Vyčisl. Mat. i Mat. Fiz. **13**, 1099–1111.

MOROZOV, V. A. (84). Methods for solving incorrectly posed problems. Springer, Berlin.

NICKEL, K. (Hrsg.) (80). Interval mathematics 1980. Academic Press, New York.

O'LEARY, D. P. (80). Estimating matrix condition numbers. SIAM J. Sci. Statist. Comput. **1**, 205–209.

OSBORNE, M. R. (76). On the computation of stepwise regression. Austral. Comput. J. **8**, 61–68.

ØSTERBY, O.; ZLATEV, Z. (83). Direct methods for sparse matrices. Lecture Notes in Computer Science, Vol. 157. Springer, Berlin. Russ. Übers.: Prjamye metody dlja razrežennyh matric. Mir, Moskva 1987.

OSTROWSKI, A. M. (58/59). On the convergence of the Rayleigh quotient iteration for the computation of characteristic roots and vectors. I–VI. Arch. Rat. Mech. Anal. **1** (1958), 233–241; **2** (1959), 423–428; **3** (1959), 325–340, 341–347, 472–481; **4** (1959), 153 to 165.

PAIGE, C. C. (76). Error analysis of the Lanczos algorithm for tridiagonalizing a symmetric matrix. J. Inst. Math. Appl. **18**, 341–349.

PAIGE, C. C. (79a). Computer solution and perturbation analysis of generalized linear least squares problems. Math. Comp. **33**, 171–184.

PAIGE, C. C. (79b). Fast numerically stable computations for generalized linear least squares problems. SIAM J. Numer. Anal. **16**, 165–171.

Paige, C. C. (80). Error analysis of some techniques for updating orthogonal decompositions. Math. Comp. **34**, 465–471.

Parlett, B. N. (64). The origin and development of methods of *LR* type. SIAM Rev. **6**, 275 to 295.

Parlett, B. N. (74). The Rayleigh quotient iteration and some generalizations for nonnormal matrices. Math. Comp. **28**, 679–693.

Parlett, B. N. (80a). The symmetric eigenvalue problem. Prentice-Hall, Englewood Cliffs, N. J. Russ. Übers.: Simmetričnaja problema sobstvennyh značeniĭ. Čislennye metody. Mir, Moskva 1983.

Parlett, B. N. (80b). How to solve $(K - \lambda M)\, z = 0$ for large K and M. Numer. Methods Engrg. **1**, 97–106.

Parlett, B. N. (84). The software scene in the extraction of eigenvalues from sparse matrices. SIAM J. Sci. Statist. Comput. **5**, 590–604.

Parlett, B. N.; Kahan, W. (69). On the convergence of a practical *QR* algorithm, in: Information processing 68. North-Holland, Amsterdam, 114–118.

Peters, G.; Wilkinson, J. H. (69). Eigenvalues of $Ax = \lambda Bx$ with band symmetric A and B. Comput. J. **12**, 398–404.

Peters, G.; Wilkinson, J. H. (70a). The least squares problem and pseudoinverses. Comput. J. **13**, 309–316.

Peters, G.; Wilkinson, J. H. (70b). $Ax = \lambda Bx$ and the generalized eigenproblem. SIAM J. Numer. Anal. **7**, 479–492.

Peters, G.; Wilkinson, J. H. (75). On the stability of Gauss-Jordan elimination with pivoting. Comm. ACM **18**, 20–24.

Peters, G.; Wilkinson, J. H. (79). Inverse iteration, ill-conditioned equations, and Newton's method. SIAM Rev. **21**, 339–360.

Pissanetzky, S. (84). Sparse matrix technology. Academic Press, New York.

Pope, D. A.; Tompkins, C. (57). Maximizing functions of rotations — experiments concerning speed of diagonalization of symmetric matrices using Jacobi's method. J. Assoc. Comput. Mach. **4**, 459–466.

Rath, W. (82). Fast Givens rotations for orthogonal similarity transformations. Numer. Math. **40**, 47–56.

Rice, J. R. (66). Experiments on Gram-Schmidt orthogonalization. Math. Comp. **20**, 325–328.

Rice, J. R. (Hrsg.) (77). Mathematical software III. Academic Press, New York.

Rice, J. R. (81). Matrix computations and mathematical software. McGraw-Hill, New York. Russ. Übers.: Matričnye vyčislenija i matematičeskoe obespečenie. Mir, Moskva 1984.

Rodrigue, G. (Hrsg.) (82). Parallel computations. Academic Press, New York.

Rutishauser, H. (58). Solution of eigenvalue problems with the *LR* transformation. Nat. Bur. Standards Appl. Math. Ser. **49**, 47–81.

Rutishauser, H. (66). The Jacobi method for real symmetric matrices. Numer. Math. **9**, 1–10. Auch in: Wilkinson, J. H.; Reinsch, C. (Hrsg.), 202–211.

Rutishauser, H. (69). Computational aspects of F. L. Bauer's simultaneous iteration method. Numer. Math. **13**, 4–13.

Rutishauser, H. (70). Simultaneous iteration method for symmetric matrices. Numer. Math. **16**, 205–223. Auch in: Wilkinson, J. H.; Reinsch, C. (Hrsg.), 202–211.

Samarskiĭ, A. A.; Nikolaev, E. S. (78). Metody rešenija setočnyh uravneniĭ. Nauka, Moskva.

Schittkowski, K.; Stoer, J. (79). A factorization method for the solution of constrained linear least squares problems allowing for subsequent data changes. Numer. Math. **31**, 431 to 463.

Schönhage, A. (64). Zur quadratischen Konvergenz des Jacobi-Verfahrens. Numer. Math. **6**, 410–412.

Schwarz, H. R. (77). Two algorithms for treating $Ax = \lambda Bx$. Comput. Methods Appl. Mech. Engrg. **12**, 181–199.

Schwarz, H. R. (79). Zur Eigenwertaufgabe $Ax = \lambda Bx$, in: Numerische Behandlung von Eigenwertaufgaben. ISNM 43. Birkhäuser, Basel, 161–175.
Schwarz, H. R. (80). Methode der finiten Elemente. Teubner, Stuttgart.
Schwetlick, H. (79). Numerische Lösung nichtlinearer Gleichungen. VEB Deutscher Verlag der Wissenschaften, Berlin; Oldenbourg, München.
Schwetlick, H.; Tiller, V. (85). Numerical methods for estimating parameters in nonlinear models with errors in the variables. Technometrics **27**, 17–24.
Scott, D. S.; Ward, R. C. (82). Solving symmetric-definite quadratic λ-matrix problems without factorization. SIAM J. Sci. Statist. Comput. **3**, 58–67.
Skeel, R. D. (79). Scaling for numerical stability in Gaussian elimination. J. Assoc. Comput. Mach. **26**, 494–526.
Skeel, R. D. (80). Iterative refinement implies numerical stability for Gaussian elimination. Math. Comp. **35**, 817–832.
Skeel, R. D. (81). Effect of equilibration on residual size for partial pivoting. SIAM J. Numer. Anal. **18**, 449–454.
Smith, B. T.; Boyle, J. M.; Dongarra, J. J.; Garbow, B. S.; Ikebe, Y.; Klema, V. C.; Moler, C. B. (74). Matrix eigensystem routines — EISPACK guide. Lecture Notes in Computer Science, Vol. 6. Springer, Berlin. 2nd ed. 1976.
Sorensen, D. C. (77). Updating the symmetric indefinite factorization with applications in a modified Newton's method. Report ANL-77-49, Argonne National Laboratory, Argonne.
Sterbenz, P. H. (74). Floating-point computation. Prentice-Hall, Englewood Cliffs, N. J.
Stewart, G. W. (69). Accelerating the orthogonal iteration for the eigenvalues of a Hermitian matrix. Numer. Math. **13**, 362–376.
Stewart, G. W. (72). On the sensitivity of the eigenvalue problem $Ax = \lambda Bx$. SIAM J. Numer. Anal. **9**, 669–686.
Stewart, G. W. (73a). Error and perturbation bounds for subspaces associated with certain eigenvalue problems. SIAM Rev. **15**, 727–764.
Stewart, G. W. (73b). Introduction to matrix computations. Academic Press, New York.
Stewart, G. W. (75). Methods of simultaneous iteration for calculating eigenvectors of matrices, in: Miller, J. H. (Hrsg.), Topics in numerical analysis II. Academic Press, New York, 185–196.
Stewart, G. W. (76a). A bibliographic tour of the large, sparse generalized eigenvalue problem, in: Bunch, J. R.; Rose, D. J. (Hrsg.), 113–130.
Stewart, G. W. (76b). The economical storage of plane rotations. Numer. Math. **25**, 137 to 138.
Stewart, G. W. (77). On the perturbation of pseudo-inverses, projections, and linear least squares problems. SIAM Rev. **19**, 634–662.
Stewart, G. W. (78). Perturbation theory for the generalized eigenvalue problem, in: de Boor, C.; Golub, G. H. (Hrsg.), Recent advances in numerical analysis. Academic Press, New York, 69–86.
Stewart, G. W. (79a). Perturbation bounds for the definite generalized eigenvalue problem. Linear Algebra Appl. **23**, 69–86.
Stewart, G. W. (79b). The effects of rounding error on an algorithm for downdating a Cholesky factorization. J. Inst. Math. Appl. **23**, 203–213.
Stewart, G. W. (80). The efficient generation of random orthogonal matrices with an application to condition estimators. SIAM J. Numer. Anal. **17**, 403–409.
Stiefel, E. (61). Einführung in die Numerische Mathematik. Teubner, Stuttgart.
Stoer, J. (71). On the numerical solution of constrained least squares problems. SIAM J. Numer. Anal. **8**, 382–411.
Stoer, J. (83). Einführung in die Numerische Mathematik. Bd. 1. 4., verb. Auflage. Springer, Berlin.
Strassen, V. (69). Gaussian elimination is not optimal. Numer. Math. **13**, 354–356.
Sun, J. (83). Perturbation analysis for the generalized eigenvalue and the generalized singular value problem, in: Kågström, B.; Ruhe, A. (Hrsg.), 221–224.

TURING, A. M. (48). Rounding-off errors in matrix processes. Quart. J. Mech. Appl. Math. **1**, 287–308.

TYHONOV, A. N. (63). O reguljarizacii nekorrektno postavlennyh zadač. Dokl. Akad. Nauk SSSR **153**, 49–52.

TYHONOV, A. N.; ARSENIN, V. JA. (79). Metody rešenija nekorrektnyh zadač. Izd. vtoroe. Nauka, Moskva.

TYHONOV, A. N.; GONČARSKIĬ, A. V.; STEPANOV, V. V.; JAGOLA, A. G. (83). Reguljarizirujuščie algoritmy i apriornaja informacija. Nauka, Moskva.

UNGER, H. (50). Nichtlineare Behandlung von Eigenwertaufgaben. Z. Angew. Math. Mech. **30**, 281–282.

VAN DER SLUIS, A. (69). Condition numbers and equilibration of matrices. Numer. Math. **14**, 14–23.

VAN DER SLUIS, A. (70). Condition, equilibration, and pivoting in linear algebraic systems. Numer. Math. **15**, 74–86.

VARGA, R. S. (62). Matrix iterative analysis. Prentice-Hall, Englewood Cliffs, N. J.

VOEVODIN, V. V. (69a). O metode reguljarizacii. Ž. Vyčisl. Mat. i Mat. Fiz. **9**, 671–673.

VOEVODIN, V. V. (69b). Ošibki okruglenija i ustoĭčivost' v prjamyh metodah lineĭnoĭ algebry. Izd. MGU, Moskva.

VOEVODIN, V. V. (77). Vyčislitel'nye osnovy lineĭnoĭ algebry. Nauka, Moskva.

VOEVODIN, V. V. (80). Lineĭnaja algebra. Nauka, Moskva. Engl. Übers.: Linear algebra. Mir, Moscow 1983.

VOEVODIN, V. V.; KUZNECOV, JU. A. (84). Matricy i vyčislenija. Nauka, Moskva.

VOEVODIN, V. V.; TYRTYŠNIKOV, E. E. (83). Vyčislenija s teplicevymi matricami, in: MARČUK, G. I. (Hrsg.), 124–266.

VON MISES, R.; POLLACZEK-GEIRINGER, H. (29). Praktische Verfahren der Gleichungsauflösung. Z. Angew. Math. Mech. **9**, 152–164.

VON NEUMANN, J.; GOLDSTINE, H. H. (47). Numerical inverting of matrices of high order. Bull. Amer. Math. Soc. **53**, 1021–1099.

WALDVOGEL, P. (82). Bisection for $Ax = \lambda Bx$ with matrices of variable band width. Computing **28**, 171–180.

WANG, J. Y.; GARBOW, B. S. (83). A numerical method for solving inverse real symmetric eigenvalue problems. SIAM J. Sci. Statist. Comput. **4**, 45–51.

WATKINS, D. S. (82). Understanding the QR algorithm. SIAM Rev. **24**, 427–440.

WEDIN, P. A. (73). Perturbation theory for pseudo-inverses. BIT **13**, 217–232.

WILKINSON, J. H. (61). Error analysis of direct methods of matrix inversion. J. Assoc. Comput. Mach. **8**, 281–330.

WILKINSON, J. H. (62). Note on the quadratic convergence of the cyclic Jacobi process. Numer. Math. **4**, 296–300.

WILKINSON, J. H. (63). Rounding errors in algebraic processes. Prentice-Hall, Englewood Cliffs, N. J. Dtsch. Übers.: Rundungsfehler. Springer, Berlin 1969. Russ. Übers.: Algebraičeskaja problema sobstvennyh značeniĭ. Nauka, Moskva 1970.

WILKINSON, J. H. (65). The algebraic eigenvalue problem. Clarendon Press, Oxford.

WILKINSON, J. H. (68). Global convergence of tridiagonal QR algorithm with origin shifts. Linear Algebra Appl. **1**, 409–420.

WILKINSON, J. H. (71). Modern error analysis. SIAM Rev. **13**, 548–568.

WILKINSON, J. H. (77). Some recent advances in numerical linear algebra, in: JACOBS, D. E. (Hrsg.), 3–53.

WILKINSON, J. H.; REINSCH, C. (Hrsg.) (71). Linear algebra. Springer, Berlin. Russ. Übers.: Lineĭnaja algebra. Spravočnik algoritmov na jazyke ALGOL. Izd. Mašinostroenie, Moskva 1976.

YOUNG, D. M. (71). Iterative solution of large linear systems. Academic Press, New York.

ZIELKE, G. (70). Numerische Berechnung von benachbarten inversen Matrizen und linearen Gleichungssystemen. Vieweg, Braunschweig.

ZIELKE, G. (78). Zur historischen Entwicklung von verallgemeinerten inversen Matrizen. Wiss. Z. Martin-Luther-Univ. Halle-Wittenberg, Math.-Natur. Reihe 4, 109—118.

ZIELKE, G. (86). Report on test matrices for generalized inverses. Computing **36**, 105—162.

ZLATEV, Z.; WASNIEWSKI, J.; SCHAUMBURG, K. (81). Y 12 M. Solution of large and sparse systems of linear algebraic equations. Documentation of subroutines. Lecture Notes in Computer Science, Vol. 121. Springer, Berlin.

Sachverzeichnis

Abbildung 20
–, identische 22
–, inverse 21
–, lineare 21
absolute Norm 30
Ähnlichkeitstransformation 35, 354
Akkumulation in höherer Genauigkeit 84
Äquilibriertheit 134
Äquivalenz numerischer Probleme 96
Äquivalenztransformation 41
Aufdatierung einer LR-Faktorisierung 214
– einer LL^T-Faktorisierung 217f.
– einer inversen Matrix 235f.
– einer QR-Faktorisierung 301ff.
Auffüllung 227
Aufwand 74
Ausgangsdaten 50
Auslöschung 70

Bandbreite 15, 221
Bandmatrix 15, 221
Basis einer Computerzahl 64
– eines Teilraumes 14
Bauer-Fike-Theorem 356
Betragsmatrix 29
Betragsvektor 29
Bidiagonalisierung 316ff.
–, modifizierte, für $m \gg n$ 324
Bidiagonalmatrix 15
Bisektionsverfahren 411ff.
Bit 65
BKP-Faktorisierung → Bunch-Kaufman-Parlett-Faktorisierung
BLAS – Basic Linear Algebra Subprograms 453
Blockelimination 191ff.
Blockmatrix 12
B-Norm 437
B-Orthogonalität 436
B-Selbstadjungiertheit 447
B-Skalarprodukt 437
Bunch-Kaufman-Parlett-Faktorisierung 197
Büschel 434
–, äquivalente 435
–, kongruente 436
Byte 65

Charakteristisches Polynom eines Büschels 435
– – einer Matrix 32, 353
Cholesky-Faktorisierung → LL^T-Faktorisierung
Computerarithmetik 69ff.
Computerzahlen 64ff.

Darstellungsfehler einer Matrix 85
– eines Vektors 85
– einer Zahl 67
Deflation bei QR-Algorithmus 420
– bei SVD-Berechnung 320
– bei Vektoriteration 381
Determinanten 22f.
–; Berechnung 47, 160, 163
–; Entwicklungssatz 23
Diagonalmatrix 15
Diagonalpivotisierung 168, 189
Dimension einer linearen Mannigfaltigkeit 242
– eines Teilraumes 14
direkte Verfahren 96, 142
dominanter Eigenwert 374, 441, 444
– Teilraum 374
doppelte Genauigkeit 65
Drehungsmatrix 112
Dreiecksfaktorisierung → LR-Faktorisierung
Dreiecksmatrix 15
Dreiecksungleichung 25
dyadisches Produkt 19

Eigenwerte 31, 434

Eigenwerte, konjugiert-komplexe 33
Eigenpaar 32, 434
Eigenvektor 32, 434
Eigenwerthaufen 358, 372, 387, 394f.
Eigenwertproblem, allgemeines symmetrisches 49, 434
—, inverses 449
—, spezielles nichtsymmetrisches 49, 54f.
—, — symmetrisches 49, 353
Eigenwertzerlegung 36
einfache Genauigkeit 65
Eingangsdaten 50
Einhüllende einer Matrix 226
Einsdreiecksmatrix 145
Eliminationsschritt mittels Givens-Drehungen 114
— mittels impliziter Givens-Drehungen 121
— mittels Householder-Spiegelung 110
— mittels LNT-Matrix 103
— mittels stabilisierter LNT-Matrix 105
Euklidische Norm 25
Exponent 63ff.
Exponentenüberlauf 66
Exponentenunterlauf 66

Fehler, absoluter 51
—, relativer 56
—, unvermeidlicher 53, 91
Fehlerabschätzung bei linearen Gleichungssystemen 174ff.
Fehleranalyse 79, 90ff.
Fehlerkumulationskonstante 79, 90ff.
Fehlerniveau 51ff.
—, optimales 52, 91
Frobeniusnorm 27

Gauß-Jordan-Verfahren 234
Gaußsche Dreiecksfaktorisierung 158ff.
→ ***LR***-Faktorisierung
—r Algorithmus 143, 153ff.
—r —, verketteter 203
Givens-Drehungen 111ff.
—, implizite 116ff.
Givens-Orthogonalisierung 290ff.
—, implizite 291
Givens-Spiegelungen 125
Gleitpunktzahlen 64
—, normalisierte 64
Gradientenverfahren 446
Gram-Schmidt-Orthogonalisierung 277ff.
—; Instabilität 279
—, modifizierte 281ff.

Hauptachsentransformation 40
Hessenbergmatrix 16
Hölderstetigkeit 54
Householder-Orthogonalisierung 288ff.
— mit Pivotisierung 292
Householder-Spiegelungen 106ff.

Inkorrekt gestelltes Problem 59
inverse Iteration 389ff., 443ff.
— Matrix 21
— —; Berechnung 46f., 229ff.
iterative Verbesserung bei Gleichungssystemen 176ff.
— — bei Quadratmittelproblemen 275ff., 299ff., 301, 315
— Verfahren 96
— — für Eigenwertprobleme 373ff., 441ff.
— — für lineare Gleichungssysteme 239

Jacobi-Verfahren 365ff.
Jordanblock 37
Jordansche Normalform 37

Kaskadensummation 93
Kompaktspeicherung 228
Konditionsschätzer 178ff.
Konditionszahl, absolute 56
— einer Matrix 130
—, partielle 58
—, relative 57
Kongruenztransformation 40
Konsistenz 241
Koordinatenvektor 13
korrekt gestelltes Problem 53
Kreisesatz von Geršgorin 39

Lanczos-Algorithmus 405f., 434, 447
LDL^T-Faktorisierung, direkte 205
— indefiniter Matrizen 190ff.
— positiv definiter Matrizen 185
— zum Schneiden des Spektrums 407ff.
lineare Abhängigkeit 14
— Gleichungssysteme mit Dreiecksmatrizen 145ff.
— —, inkonsistente 241
— —, konsistente 241
— — mit orthogonalen Matrizen 148ff.
— —, quadratische reguläre 45, 142ff.
— —, rechteckige 240ff.
— —, schwach besetzte 46
— —, überbestimmte 47, 240ff.
— —, unterbestimmte 301
— Mannigfaltigkeit 242
— Quadratmittelprobleme 47f., 243ff.
— —, gewichtete 266f., 268, 274
— — mit Nebenbedingungen 269
— —, rangdefiziente 253, 264, 283, 315ff.

lineare Quadratmittelprobleme, spaltenreguläre 264, 270ff., 277ff.
— —, steife 267, 268, 274, 292
— —, totale 269
— Unabhängigkeit 14
—s Ausgleichsproblem 243
Linearkombination 12
Lipschitzkonstante 54
—, lokale 53
—, relative 56
Lipschitzstetigkeit 54
—, lokale 53
LL^T-Faktorisierung 188ff.
—; Aufdatierung 217ff.
— von Bandmatrizen 224, 226ff.
—, direkte 206f.
LNT-Matrix 101
Lösbarkeitsbedingung 241
Lösungsmenge eines Gleichungssystems 242
— eines Quadratmittelproblems 243, 244
LR-Faktorisierung 143, 158, 162
— von Bandmatrizen 221ff.
—, direkte 201ff.
— von diagonaldominanten Matrizen 165ff.
— von positiv definiten Matrizen 166ff., 204ff.
— von schwach besetzten Matrizen 220, 226ff.
— von symmetrischen Matrizen 226
— von Tridiagonalmatrizen 225

Mantisse 63, 64
Matrix 14
—, defektive 34
—, diagonalähnliche 37
—, dominante 165
—, gestufte 359
—, indefinite 22
—, inverse 21
—, orthogonale 24
—, positiv definite 22
—, — semidefinite 22
—, quadratische 18
—, rangdefiziente 21
—, reguläre 21
—, schiefsymmetrische 30
—, schwach besetzte 46
—, singuläre 21
—, spaltenorthonormale 24
—, spaltenreguläre 21
—, symmetrische 17
—, transponierte 16
—, zeilenorthonormale 24
Matrixnormen 27ff.
Matrizen 90
Matrizen, ähnliche 35
—, äquivalente 41
—, kongruente 40
—, vertauschbare 19
Meßfehler 137
Methode der kleinsten Quadrate 243
MGS-Orthogonalisierung 281ff.
— mit Pivotisierung 282f.
— mit Re-Orthogonalisierung 284
modifizierte Gram-Schmidt-Orthogonalisierung → MGS-Orthogonalisierung
monotone Norm 30
Moore-Penrose-Inverse → Pseudoinverse

Nichtorthogonale elementare Transformationsmatrix 101
— — —, stabilisierte 105
NNE = Nichtnullelemente 220ff.
Normalgleichungen 247
Normalgleichungsverfahren 270ff.
— mit höherer Genauigkeit 274
—; Instabilität 273
— mit iterativer Verbesserung 275
Normallösung 248ff., 301, 315ff.
NT-Matrix 101
Nullmatrix 16
Nullraum einer Matrix 21
numerische Gutartigkeit 79, 90
— Probleme 50
— Problemklassen 50
— Stabilität 79, 91
—r Algorithmus 73
—r —; Computerrealisierung 73
—s Verfahren → numerischer Algorithmus

Orthogonale Faktorisierung → QR-Faktorisierung
— Projektion 244
—r Projektor 245, 250
—s Komplement 246
Orthogonalität 24
Orthogonalisierungsverfahren 143, 277ff., 316ff.
Orthonormalität 24

Penrose-Bedingungen 249
Permutation 97
Permutationsmatrix 98
Pivotelement 153
Pivotisierung beim Gaußschen Algorithmus 153ff., 161ff.
— bei impliziten Givens-Drehungen 123, 292
— beim Jacobi-Verfahren 369ff.
— bei der QR-Faktorisierung 282f., 292
Potenzmethode → Vektoriteration

Profil einer Matrix 226
Programmpaket 452
Programmsystem 452
Projektion → orthogonale Projektion
Projektor → orthogonaler Projektor
Pseudoinverse 48, 249
Pseudorang 327, 332

QL-Algorithmus 431f.
QR-Algorithmus 415ff.
—, expliziter 426
—; Grundform 416
—, impliziter 428
—, tridiagonaler 424ff., 430
QR-Faktorisierung 143
—; Eindeutigkeit 279
— nach Givens 290ff.
— nach Gram-Schmidt 277ff.
— nach Householder 285ff.
quadratische Form 22
Quadratmittellösung 243
QZ-Algorithmus 447

Rang einer Matrix 21
Rang-1-Modifikation 208, 234f., 275, 301ff., 449
rangerhöhende Störungen 253f., 326, 340
Rayleigh-Quotient 362, 385, 442
Rayleigh-Quotienten-Iteration 395, 444
Rayleigh-Quotienten-Verschiebung 423
Rayleigh-Ritz-Algorithmus 384ff., 444ff.
Reduktion eines allgemeinen Eigenwertproblems auf ein spezielles, explizite 438ff.
— — — — — — —, implizite 442ff.
Regularisierung 59
—, diskrete 325ff.
—, kontinuierliche 338ff.
Regularisierungsparameter 327, 338
relative Maschinengenauigkeit 67
—s Rundungsfehlerniveau 67
Re-Orthogonalisierung bei Aufdatierung von *QR*-Faktorisierungen 313
— bei Gram-Schmidt-Orthogonalisierung 284
— bei inverser Iteration 394
Residualkriterien bei Eigenwertproblemen 360ff.
— bei Gleichungssystemen 139ff.
— bei Quadratmittelproblemen 261ff.
Residuum bezüglich eines allgemeinen Eigenwertproblems 445
— bezüglich eines speziellen Eigenwertproblems 360
— bezüglich eines Gleichungssystems 139
— bezüglich eines Quadratmittelproblems 261, 299
Reskalierung 292
Ritzsche Eigenwerte und Eigenvektoren 385
RQ-Iteration → Rayleigh-Quotienten-Iteration
RR-Algorithmus → Rayleigh-Ritz-Algorithmus
Rücksubstitution 159
Rückwärtsanalyse 79
Rundung 66
—, symmetrische 66
—, unsymmetrische 66
Rundungsfehler 67, 91

Satz von Schur 38
Schneiden des Spektrums 408ff.
schrittweise Regression 266, 269
Schwarzsche Ungleichung 25
Sherman-Morrison-Formel 235
simultane Iteration 387
Singulärvektoren 42
Singulärwerte 41
Singulärwertzerlegung 41
—; Berechnung 316ff., 321
Skalarprodukt 19
Skalierung bei linearen Gleichungssystemen 135
— — — Quadratmittelproblemen 261, 266, 291
SLNT-Matrix 105
Software 450ff.
Spaltenpivotisierung nach Bunch-Kaufman-Parlett 197
— beim Gaußschen Algorithmus 161, 203
— bei Givens-Drehungen 123
— bei *QR*-Faktorisierung 292
Spaltenraum einer Matrix 21
Spektralnorm 27
Spektralverschiebung 355
— bei inverser Iteration 389ff.
— bei *QR*-Algorithmus 420ff.
— bei SVD 319f.
Spektrumslücke 358
Spiegelungsmatrizen 107ff.
Störungslemma 128
Störungstheorie von Eigenwertproblemen 355ff.
— von Gleichungssystemen 129
— von iterativen Matrizen 128
— von Pseudoinversen 251ff.
— von Quadratmittelproblemen 257ff.
Sturmsche Ketten 434
SVD → Singulärwertzerlegung
symmetrische Blockfaktorisierung 193

Teilraum 13
—, invarianter 374

Teilraumiteration 381ff., 441ff.
Trägheitsgesetz für quadratische Formen 40
Transformation auf Tridiagonalform 397ff.
Trapezmatrix 15
Tridiagonalmatrix 224, 398
—; Dreiecksfaktorisierung 225
—, nichtzerfallende 406
—; *QR*-Algorithmus 424ff.

Vektoren 11
—, linear abhängige/unabhängige 14
—, orthogonale/orthonormale 24
Vektornormen 25, 26
Vektoriteration 375ff., 441ff.
verallgemeinerte Inverse → Pseudoinverse
Verschiebung → Spektralverschiebung
Vertauschungsmatrix 99
Vielfachheit, algebraische 33
—, geometrische 34
Vollrang 21
Vollrangmatrix 21
vollständige Pivotisierung 165, 197
Vorwärtsanalyse 79
Vorwärtselimination 159

Wertebereich einer Matrix 21
Wielandt-Hoffman-Theorem 356
Wilkinson-Verschiebung 424, 432
Winkel zweier Vektoren 26
— zweier Teilräume 381

Zeilenpivotisierung beim Gaußschen Algorithmus 164