ABSCHÄTZUNG DER FORTPFLANZUNG DER UNGENAUIGKEIT DER DATEN IN DIE LÖSUNG BEI LINEAREN GLEICHUNGSSYSTEMEN UND MATRIZENGLEICHUNGEN

von

WERNER DÜCK

Bei den in der Praxis auftretenden linearen Gleichungssystemen sind die Daten der Aufgabe meist mit Ungenauigkeiten behaftet. Die Frage nach der Abschätzung der Fortpflanzung der Ungenauigkeit der Daten in die Lösung des Systems ist sehr wichtig aber bis zum heutigen Tage noch nicht befriedigend beantwortet worden. Die folgende Arbeit soll ein kleiner Beitrag zu diesem Fragenkreis sein. Es werden einige Abschätzungsformeln zusammengestellt und lineare Gleichungssysteme und Matrizengleichungen behandelt. Für Gleichungssysteme mit stark überwiegender Hauptdiagonale lassen sich die Abschätzungsformeln sehr leicht anwenden. Für allgemeine Gleichungssysteme wird die Abschätzung mit Hilfe des Defekts durchgeführt, wozu außerdem noch eine Abschätzung für die Inverse benötigt wird, die in verschiedener Weise erfolgen kann. Parallel zu den Betrachtungen über lineare Gleichungssysteme laufen die Untersuchungen über Matrizengleichungen. Die ermittelten Abschätzungsformeln werden auf Beispiele angewandt. Auf die Ergebnisse der Literatur wird am Ende der Arbeit kurz eingegangen.

Mein verehrter Lehrer, Herr Professor Dr.-Ing. habil E. WEINEL, Jena, hat in seinen Vorlesungen wiederholt auf diesen Problemkreis hingewiesen. Für die vielen wertvollen Anregungen möchte ich Herrn Professor WEINEL an dieser Stelle herzlichst danken.

§ 1. Abschätzung der Fehlerfortpflanzung bei linearen Gleichungssystemen mit überwiegender Hauptdiagonale

1.1. *Problemstellung.* In der Praxis wird man immer wieder vor die Aufgabe gestellt, ein lineares Gleichungssystem der Form

$$(\mathbf{C} - \delta\mathbf{C})(\boldsymbol{x} + \delta\boldsymbol{x}) - (\boldsymbol{c} + \delta\boldsymbol{c}) = 0 \tag{1}$$

zu lösen, bei dem die Matrix $\mathbf{C}$ und der Spaltenvektor $\boldsymbol{c}$ als zahlenmäßig gegeben anzusehen sind, während für die Elemente δc_{ij} bzw. δc_i der Fehlermatrix $\delta\mathbf{C}$ bzw. des Fehlervektors $\delta\boldsymbol{c}$ nur Schranken für ihre Beträge bekannt sind. Bezeichnen wir diese Schranken mit Δc_{ij} bzw. Δc_i, so gilt

$$|\delta c_{ij}| \leqq \Delta c_{ij}, \quad |\delta c_i| \leqq \Delta c_i . \tag{2}$$

Numerisch kann man nur das System

$$\mathbf{C}\boldsymbol{x} - \boldsymbol{c} = 0 \tag{3}$$

lösen, dessen Matrix $\mathbf{C}$ als nicht singulär vorausgesetzt wird. Mit $\boldsymbol{x}$ wird die exakte Lösung dieser Gleichung bezeichnet. Es soll die Fortpflanzung der Ungenauigkeit der Daten in die Lösung abgeschätzt werden, d. h. es ist der Einfluß der Fehlerglieder $\delta\,\mathbf{C}$ und $\delta\,\boldsymbol{c}$, also $\delta\,\boldsymbol{x}$ abzuschätzen.

Wir wollen voraussetzen, daß die Störung $\delta\,\mathbf{C}$ des Systems (1) so klein ist, daß wenn die Matrix $\mathbf{C}$ stark überwiegende Hauptdiagonalelemente besitzt, dieses auch von der Matrix $\mathbf{C} + \delta\,\mathbf{C}$ behauptet werden kann. Dann erscheint es angebracht, das System (1) auf iterierfähige Form

$$\boldsymbol{x} + \delta\,x = (\mathbf{A} + \delta\,\mathbf{A})\,(\boldsymbol{x} + \delta\,\boldsymbol{x}) + (\boldsymbol{a} + \delta\,\boldsymbol{a}) \tag{4}$$

zu bringen, wobei numerisch wieder nur das System

$$\boldsymbol{x} = \mathbf{A}\boldsymbol{x} + \boldsymbol{a} \tag{5}$$

gelöst werden kann. Wir haben die Gewinnung der Gleichung (4) jetzt näher zu verfolgen, um Abschätzungen für die Fehlerglieder $\delta\,\mathbf{A}$, $\delta\,\boldsymbol{a}$ des iterierfähigen Systems zu erhalten.

1.2. *Abschätzung der Fehlerglieder des iterierfähigen Systems.* Zur Überführung des Systems (1) in iterierfähige Form zerlegen wir die Matrix $\mathbf{C}$ entsprechend der Gleichung

$$\mathbf{C} = \mathbf{D} - \mathbf{B} \tag{6}$$

in zwei Matrixen $\mathbf{D}$ und $\mathbf{B}$. Dabei ist $\mathbf{D}$ eine Diagonalmatrix, deren Diagonalelemente mit denen der Matrix $\mathbf{C}$ übereinstimmen. Die Matrix $\mathbf{B}$ enthält in der Hauptdiagonale Nullen, während sonst die Elemente von $\mathbf{B}$ entgegengesetzt gleich den entsprechenden Elementen von $\mathbf{C}$ sind. Mit (6) können wir dann für das System (1) schreiben

$$\mathbf{D}(\boldsymbol{x} + \delta\,\boldsymbol{x}) = (\mathbf{B} + \delta\,\mathbf{C})\,(\boldsymbol{x} + \delta\,\boldsymbol{x}) + (\boldsymbol{c} + \delta\,\boldsymbol{c})\,,$$

$$\boldsymbol{x} + \delta\,\boldsymbol{x} = (\mathbf{D}^{-1}\,\mathbf{B} + \mathbf{D}^{-1}\,\delta\,\mathbf{C})\,(\boldsymbol{x} + \delta\,\boldsymbol{x}) + (\mathbf{D}^{-1}\,\boldsymbol{c} + \mathbf{D}^{-1}\,\delta\,c)\,.$$

Setzen wir

$$\mathbf{A} = \mathbf{D}^{-1}\,\mathbf{B}\,, \qquad \delta\,\mathbf{A} = \mathbf{D}^{-1}\,\delta\,\mathbf{C}\,, \tag{7}$$

$$\boldsymbol{a} = \mathbf{D}^{-1}\,\boldsymbol{c}\,, \qquad \delta\,\boldsymbol{a} = \mathbf{D}^{-1}\,\delta\,\boldsymbol{c}\,, \tag{8}$$

so stoßen wir auf das System (4).

Unter der Norm[2] einer n-reihigen, quadratischen Matrix $\mathbf{C}$ mit den Elementen c_{ij} wollen wir in dieser Arbeit die maximalen Zeilensumme der Beträge der Elemente verstehen:

$$\|\mathbf{C}\| = \max_i \sum_{k=1}^{n} |c_{ik}|\,.$$

Die Norm eines Spaltenvektors $\boldsymbol{x}$ mit den Komponenten x^i ist durch

$$\|\boldsymbol{x}\| = \max_i |x^i|$$

definiert.

[2] Auf andere Normdefinitionen soll hier nicht eingegangen werden.

Für die Elemente der Fehlerglieder $\delta\,\mathbf{C}$, $\delta\,\boldsymbol{c}$ sind Schranken entsprechend (2) bekannt. Für die Normen der Fehlerglieder gilt somit die Abschätzung

$$\|\delta\,\mathbf{C}\| \leqq \max_i \sum_{k=1}^{n} |\Delta\, c_{ik}|\,, \qquad \|\delta\,\boldsymbol{c}\| \leqq \max_i |\Delta\, c_i|\,.$$

Dann können wir auch die Normen der Fehlerglieder $\delta\,\mathbf{A}$, $\delta\,\boldsymbol{a}$ des Systems (4) wegen der für zwei Matrizen $\mathbf{A}$, $\mathbf{B}$ gültigen Beziehung

$$\|\mathbf{AB}\| \leqq \|\mathbf{A}\| \cdot \|\mathbf{B}\| \tag{9}$$

nach (7), (8) leicht abschätzen

$$\|\delta\,\mathbf{A}\| \leqq \|\mathbf{D}^{-1}\| \cdot \|\delta\,\mathbf{C}\|,$$
$$\|\delta\,\boldsymbol{a}\| \leqq \|\mathbf{D}^{-1}\| \cdot \|\delta\,\boldsymbol{c}\|\,.$$

Damit haben wir Abschätzungen für die Normen der Fehlerglieder des iterierfähigen Systems erhalten, so daß wir für die weiteren Betrachtungen von den Gleichungen (4) und (5) ausgehen können.

1.3. *Abschätzung von* $\delta\,\boldsymbol{x}$. Für $\delta\,\boldsymbol{x}$ läßt sich leicht eine Abschätzung angeben. Ziehen wir Gleichung (5) von (4) ab, so finden wir

$$\delta\,\boldsymbol{x} = \delta\,\mathbf{A} \cdot \boldsymbol{x} + \mathbf{A} \cdot \delta\,\boldsymbol{x} + \delta\,\mathbf{A} \cdot \delta\,\boldsymbol{x} + \delta\,\boldsymbol{a},$$
$$\delta\,\boldsymbol{x} = (\mathbf{E} - \mathbf{A} - \delta\,\mathbf{A})^{-1}\,(\delta\,\mathbf{A} \cdot \boldsymbol{x} + \delta\,\boldsymbol{a})\,^{3},$$
$$\|\delta\,\boldsymbol{x}\| \leqq \|(\mathbf{E} - \mathbf{A} - \delta\,\mathbf{A})^{-1}\|\,\{\|\boldsymbol{x}\| \cdot \|\delta\,\mathbf{A}\| + \|\delta\,\boldsymbol{a}\|\}\,. \tag{10}$$

Zur Abschätzung der Norm von $(\mathbf{E} - \mathbf{A} - \delta\,\mathbf{A})^{-1}$ erinnern wir uns daran, daß bekanntlich $(\mathbf{E} - \mathbf{A})^{-1}$ analog zur geometrischen Reihe gewöhnlicher Zahlen in eine Matrizenreihe, die sog. Neumannsche Reihe, entwickelt werden kann

$$(\mathbf{E} - \mathbf{A})^{-1} = \mathbf{E} + \mathbf{A} + \mathbf{A}^2 + \ldots\,,$$

die für $\|\mathbf{A}\| < 1$ konvergiert. Daher gilt unter Berücksichtigung von (9)

$$\|(\mathbf{E} - \mathbf{A})^{-1}\| \leqq \sum_{i=0}^{\infty} \|\mathbf{A}^i\| \leqq \sum_{i=0}^{\infty} \|\mathbf{A}\|^i = \frac{1}{1 - \|\mathbf{A}\|}\,. \tag{11}$$

Wir setzten voraus, daß das vorgelegte Gleichungssystem überwiegende Hauptdiagonalelemente besitzt, also $\|\mathbf{A}\| < 1$ ist. Wir setzen weiterhin voraus, daß die Störung $\delta\,\mathbf{A}$ so klein ist, daß auch noch

$$\|\mathbf{A}\| + \|\delta\,\mathbf{A}\| < 1 \tag{12}$$

gilt. Dann erhalten wir aus (10) unter Berücksichtigung von (11) die gesuchte Abschätzungsformel

$$\boxed{\|\delta\,\boldsymbol{x}\| \leqq \frac{\|\boldsymbol{x}\| \cdot \|\delta\,\mathbf{A}\| + \|\delta\,\boldsymbol{a}\|}{1 - \|\mathbf{A}\| - \|\delta\,\mathbf{A}\|}}\,. \tag{13}$$

[3] Mit $\mathbf{E}$ wird die Einheitsmatrix bezeichnet.

In diese Abschätzung geht außer den Normen von $\mathbf{A}$, $\delta\,\mathbf{A}$, $\delta\,\boldsymbol{a}$, für die wir obere Schranken angeben können, noch die Norm der Lösung $\boldsymbol{x}$ des Systems (5) ein, die wir jetzt geeignet abschätzen wollen.

1.4. *Abschätzung der Norm von* $\boldsymbol{x}$. Die Norm der Lösung $\boldsymbol{x}$ können wir abschätzen, ohne daß wir eine Näherungslösung des Systems (5) kennen. Das gestattet uns, die Fortpflanzung der Ungenauigkeit der Daten bereits vor der Lösung des numerisch zu behandelnden Systems (5) zu beurteilen. Aus (5) finden wir nämlich

$$(\mathbf{E}-\mathbf{A})\,\boldsymbol{x}=\boldsymbol{a}\,,$$

und mit (11) ergibt sich unter der gemachten Voraussetzung $\|\mathbf{A}\|<1$

$$\|\boldsymbol{x}\| \leqq \frac{\|\boldsymbol{a}\|}{1-\|\mathbf{A}\|}\,. \tag{14}$$

Ist jedoch eine Näherungslösung $\overline{\boldsymbol{x}}$ des Systems (5) bekannt, so können wir das zur Abschätzung der Norm von $\boldsymbol{x}$ ausnutzen. $\overline{\boldsymbol{x}}$ steht mit dem Defekt $\boldsymbol{r}$ in der Beziehung

$$\overline{\boldsymbol{x}}=\mathbf{A}\overline{\boldsymbol{x}}+\boldsymbol{a}-\boldsymbol{r}\,. \tag{15}$$

Ziehen wir (15) von Gleichung (5) ab, so finden wir

$$\begin{aligned}\boldsymbol{x}-\overline{\boldsymbol{x}}&=\mathbf{A}(\boldsymbol{x}-\overline{\boldsymbol{x}})+\boldsymbol{r}\,,\\ \boldsymbol{x}&=\overline{\boldsymbol{x}}+(\mathbf{E}-\mathbf{A})^{-1}\boldsymbol{r}\,,\\ \|\boldsymbol{x}\|&\leqq\|\overline{\boldsymbol{x}}\|+\|(\mathbf{E}-\mathbf{A})^{-1}\|\cdot\|\boldsymbol{r}\|\,.\end{aligned}$$

Wegen (11) erhalten wir schließlich unter der Voraussetzung $\|\mathbf{A}\|<1$

$$\|\boldsymbol{x}\| \leqq \|\overline{\boldsymbol{x}}\|+\frac{\|\boldsymbol{r}\|}{1-\|\mathbf{A}\|}\,. \tag{16}$$

1.5. *Beispiel.* Die bisherigen Betrachtungen sollen auf ein Beispiel angewandt werden. Wir wählen

$$\mathbf{C}=\begin{bmatrix}200 & 40 & 20\\ 45 & 150 & 15\\ 10 & 10 & 100\end{bmatrix},\qquad \boldsymbol{c}=\begin{bmatrix}340\\ 390\\ 330\end{bmatrix}$$

und nehmen an, daß die Daten des linearen Gleichungssystems mit der Ungenauigkeit

$$|\delta\,c_{ij}| \leqq 1=\Delta\,c_{ij},\quad |\delta\,c_i| \leqq 1=\Delta\,c_i$$

behaftet sind. Dann gilt

$$\|\delta\,\mathbf{C}\| \leqq 3,\quad \|\delta\,\boldsymbol{c}\| \leqq 1.$$

Es ist

$$\mathbf{D} = \begin{bmatrix} 200 & 0 & 0 \\ 0 & 150 & 0 \\ 0 & 0 & 100 \end{bmatrix}, \quad \mathbf{A} = \begin{bmatrix} 0 & -\frac{1}{5} & -\frac{1}{10} \\ -\frac{3}{10} & 0 & -\frac{1}{10} \\ -\frac{1}{10} & -\frac{1}{10} & 0 \end{bmatrix}, \quad \boldsymbol{a} = \begin{bmatrix} \frac{17}{10} \\ \frac{13}{5} \\ \frac{33}{10} \end{bmatrix}$$

und

$$\|\mathbf{D}^{-1}\| = \frac{1}{100}, \quad \|\mathbf{A}\| = \frac{2}{5}, \quad \|\boldsymbol{a}\| = \frac{33}{10}.$$

Weiter ergibt sich

$$\|\delta\,\mathbf{A}\| \leqq \frac{3}{100}, \quad \|\delta\,\boldsymbol{a}\| \leqq \frac{1}{100}.$$

Bei der Berechnung der Matrix $\mathbf{A}$ tritt im Beispiel kein zusätzlicher Rechnungsfehler durch Divisionen auf, so daß wir auch bei der Abschätzung der Normen der Fehlerglieder $\delta\,\mathbf{A}$ und $\delta\,\boldsymbol{a}$ einen zusätzlichen Ungenauigkeitsfaktor nicht zu berücksichtigen brauchen.

Für die Abschätzung der Norm des Lösungsvektors $\boldsymbol{x}$ nach (14) finden wir

$$\|\boldsymbol{x}\| \leqq \frac{3{,}3}{1 - 0{,}4} = 5{,}5. \tag{17}$$

Verwenden wir für die Abschätzung nach (16) als Näherungsvektor $\overline{\boldsymbol{x}}$ mit dem zugehörigen Defekt $\boldsymbol{r}$

$$\overline{\boldsymbol{x}} = \begin{bmatrix} 0{,}99 \\ 2{,}02 \\ 3{,}01 \end{bmatrix}, \quad \boldsymbol{r} = \begin{bmatrix} 0{,}005 \\ -0{,}018 \\ -0{,}011 \end{bmatrix},$$

so ist

$$\|\overline{\boldsymbol{x}}\| = 3{,}01, \quad \|\boldsymbol{r}\| = 0{,}018,$$

und wir finden für (16)

$$\|\boldsymbol{x}\| \leqq 3{,}01 + \frac{0{,}018}{1 - 0{,}4} = 3{,}04. \tag{18}$$

Für die exakte Lösung würde gelten

$$\|\boldsymbol{x}\| = 3. \tag{19}$$

Die Abschätzung (13) ergibt bei Verwendung der Normen (17) bis (19) für $\boldsymbol{x}$:

$$\|\delta\,\boldsymbol{x}\| \leqq \frac{5{,}5 \cdot 0{,}03 + 0{,}01}{1 - 0{,}4 - 0{,}03} < 0{,}31 \qquad \text{mit (17)},$$

$$\|\delta\,\boldsymbol{x}\| \leqq 0{,}178 \qquad \text{mit (18)},$$

$$\|\delta\,\boldsymbol{x}\| \leqq 0{,}176 \qquad \text{mit (19)}.$$

Alle drei Abschätzungen für δx lassen erwarten daß der Einfluß der Ungenauigkeit der Daten der Aufgabe nicht zu vernachlässigen ist. Daher ist es auch nicht sinnvoll, die Lösung x mit größerer Genauigkeit zu berechnen.

§ 2. Abschätzung der Fehlerfortpflanzung bei allgemeinen linearen Gleichungssystemen mit Hilfe des Defekts

2.1. *Problemstellung.* Während sich die Abschätzungen in § 1 auf lineare Gleichungssysteme mit stark überwiegender Hauptdiagonale beziehen, ist die in diesem Abschnitt anzugebende Abschätzung allgemein für lineare Gleichungssysteme gültig, wenn nur die Koeffizientenmatrix nicht fast singulär ist und die Störung ihrer Elemente nicht zu groß ist. Nach den Ausführungen in § 1 kann der Einfluß der Ungenauigkeit der Daten der Aufgabe selbst dann abgeschätzt werden, wenn keine Näherungslösung des Systems bekannt ist. Jetzt werden wir grundsätzlich voraussetzen müssen, daß wir eine solche Näherungslösung kennen.

Die Gleichung (1) schreiben wir in dem Folgenden in der Form

$$(\mathbf{C} - \delta\mathbf{C})\,\tilde{x} - (c + \delta c) = 0\,. \tag{20}$$

Mit $\bar{x}$ bezeichnen wir eine Näherungslösung der Gleichung (3) und mit r den zugehörigen Defekt

$$\mathbf{C}\,\bar{x} - c = -r\,. \tag{21}$$

Gesucht wird eine Abschätzung der Norm von $\tilde{x} - \bar{x}$.

2.2. *Fehlerabschätzung.* Aus den Gleichungen (20), (21) finden wir

$$(\mathbf{C} - \delta\mathbf{C})\;\tilde{x} - \mathbf{C}\,\bar{x} = \delta c + r\,,$$

$$(\mathbf{C} - \delta\mathbf{C})\,(\tilde{x} - \bar{x}) = \delta\mathbf{C}\,\bar{x} + \delta c + r\,,$$

$$\|\tilde{x} - \bar{x}\| \leqq \|(\mathbf{C} - \delta\mathbf{C})^{-1}\|\,\{\|\delta\mathbf{C}\| \cdot \|\bar{x}\| + \|\delta c\| + \|r\|\}\,. \tag{22}$$

Es muß jetzt noch die Norm von $(\mathbf{C} - \delta\mathbf{C})$, abgeschätzt werden. Es gilt

$$\mathbf{C} - \delta\mathbf{C} = \mathbf{C}(\mathbf{E} - \mathbf{C}^{-1}\,\delta\mathbf{C}),$$

$$\|(\mathbf{C} - \delta\mathbf{C})^{-1}\| \leqq \|\mathbf{C}^{-1}\| \cdot \|(\mathbf{E} - \mathbf{C}^{-1}\,\delta\mathbf{C})^{-1}\|\,.$$

Setzen wir

$$\|\mathbf{C}^{-1}\| \cdot \|\delta\mathbf{C}\| < 1 \tag{23}$$

voraus, was erfüllt ist, wenn die Koeffizienten des linearen Gleichungssystems nicht mit zu großen Störungen behaftet sind, und die Matrix $\mathbf{C}$ nicht fast singulär ist, so erhalten wir bei Beachtung der Formel (11) von § 1

$$\|(\mathbf{C} - \delta\mathbf{C})^{-1}\| \leqq \|\mathbf{C}^{-1}\|\,\frac{1}{1 - \|\mathbf{C}^{-1}\,\delta\mathbf{C}\|} \leqq \frac{\|\mathbf{C}^{-1}\|}{1 - \|\mathbf{C}^{-1}\| \cdot \|\delta\mathbf{C}\|}\,.$$

(22) führt dann zu der gesuchten Abschätzung

$$\boxed{\|\tilde{x} - \bar{x}\| \leqq \|\mathbf{C}^{-1}\|\,\frac{\|\delta\mathbf{C}\| \cdot \|\bar{x}\| + \|\delta c\| + \|r\|}{1 - \|\mathbf{C}^{-1}\| \cdot \|\delta\mathbf{C}\|}}\,. \tag{24}$$

In der Abschätzung (24) treten neben der Norm der Näherungslösung $\bar{\boldsymbol{x}}$ und des Defektenvektors $\boldsymbol{r}$ die Normen der Fehlerglieder $\delta\,\mathbf{C}$ und $\delta\,\boldsymbol{c}$ auf, die wir im Sinne von 1.2. als bekannt ansehen können. In (24) geht aber weiterhin die Norm der Matrix $\mathbf{C}^{-1}$ ein, was uns vor eine neue Aufgabe stellt, nämlich Abschätzungen für die Norm der Inversen der Matrix $\mathbf{C}$ anzugeben. Mit dieser Aufgabe werden wir uns in § 4 näher befassen.

Die Abschätzung (24) beantwortet zugleich die Frage, wie bei der Lösung eines linearen Gleichungssystems der Einfluß der Abrundungsfehler und ihre Fortpflanzung abgeschätzt werden kann. Dazu ist die Norm der Differenz der exakten Lösung $\tilde{\boldsymbol{x}}=\boldsymbol{x}$ der Gleichung (3) und der Näherungslösung $\bar{\boldsymbol{x}}$ von (21) abzuschätzen. Eine solche Abschätzungsformel erhalten wir aber aus (24), wenn wir dort $\| \delta\,\mathbf{C} \| = \| \delta\,\boldsymbol{c} \| = 0$ setzen.

§ 3. Abschätzungen bei Matrizengleichungen

3.1. *Problemstellung.* Die Bestimmung der Inversen einer Matrix ist eine Aufgabe von großer praktischer Wichtigkeit. Sind die Koeffizienten der zu invertierenden Matrix mit Ungenauigkeiten behaftet, so wird sich diese Ungenauigkeit in die inverse Matrix fortpflanzen. Wieder besteht die Frage, wie sich diese Fortpflanzung der Ungenauigkeit der Daten der Aufgabe abschätzen läßt.

Um die Inverse einer Matrix zu bestimmen, haben wir die Matrizengleichung

$$(\mathbf{C} - \delta\,\mathbf{C})\,(\mathbf{X} + \delta\,\mathbf{X}) = \mathbf{E} \tag{25}$$

zu lösen. Die Matrix $\mathbf{C}$ ist als zahlenmäßig gegeben anzusehen, während für die Elemente $\delta\,c_{ij}$ der Fehlermatrix $\delta\,\mathbf{C}$ nur Schranken für ihre Beträge im Sinne von § 1 bekannt sind. Numerisch können wir nur die Matrizengleichung

$$\mathbf{CX} = \mathbf{E} \tag{26}$$

lösen, wobei mit $\mathbf{X}$ die exakte Lösung dieser Gleichung bezeichnet wird. Die für lineare Gleichungssysteme angegebenen Abschätzungen in § 1 und § 2 lassen sich sofort übertragen.

3.2. *Übertragung der Abschätzungen von* § 1. Die Matrix $\mathbf{C}$, welche stark überwiegende Hauptdiagonalelemente besitzen möge, wird wieder entsprechend (6) in die Matrizen $\mathbf{D}$ und $\mathbf{B}$ zerlegt, und es werden die Gleichungen (25), (26) auf iterierfähige Form übergeführt

$$\mathbf{X} + \delta\,\mathbf{X} = (\mathbf{A} + \delta\,\mathbf{A})\,(\mathbf{X} + \delta\,\mathbf{X}) + \mathbf{D}^{-1}, \tag{27}$$

$$\mathbf{X} = \mathbf{AX} + \mathbf{D}^{-1}. \tag{28}$$

Die Matrizen $\mathbf{A}$ und $\delta\,\mathbf{A}$ bestimmen sich dabei aus $\mathbf{D}$, $\mathbf{B}$ und $\delta\,\mathbf{C}$ wie in (7) angegeben. Für die Matrix $\delta\,\mathbf{X}$ finden wir dann unter der Voraussetzung (12) die zu (13) analoge Abschätzung

$$\boxed{\| \delta\,\mathbf{X} \| \leq \frac{\| \mathbf{X} \| \cdot \| \delta\,\mathbf{A} \|}{1 - \| \mathbf{A} \| - \| \delta\,\mathbf{A} \|}}. \tag{29}$$

In (29) tritt wieder die Norm der Lösung $\mathbf{X}$ von Gleichung (28) auf, die wir analog zu (14) in der Form

$$\|\mathbf{X}\| \leqq \frac{\|\mathbf{D}^{-1}\|}{1 - \|\mathbf{A}\|} \tag{30}$$

abschätzen können.

Bezeichnen wir mit $\overline{\mathbf{X}}$ eine Näherungslösung von (28) und mit $\mathbf{R}$ die zugehörige Defektenmatrix

$$\overline{\mathbf{X}} = \mathbf{A}\overline{\mathbf{X}} + \mathbf{D}^{-1} - \mathbf{R}\,, \tag{31}$$

so finden wir die (16) entsprechende Abschätzung

$$\|\mathbf{X}\| \leqq \|\overline{\mathbf{X}}\| + \frac{\|\mathbf{R}\|}{1 - \|\mathbf{A}\|}\,. \tag{32}$$

3.3. *Übertragung der Abschätzung von* § 2. Die jetzt anzugebende Abschätzung wird wieder von der Voraussetzung frei sein, daß die zu invertierende Matrix stark überwiegende Hauptdiagonalelemente besitzt, dafür aber voraussetzen, daß die Matrix $\mathbf{C}$ nicht fast singulär ist und die Störung ihrer Elemente nicht zu groß ist. An Stelle von (20), (21) betrachten wir die Matrizengleichungen

$$(\mathbf{C} - \delta\,\mathbf{C})\,\widetilde{\mathbf{X}} = \mathbf{E}\,, \tag{33}$$

$$\mathbf{C}\overline{\mathbf{X}} = \mathbf{E} - \mathbf{R}\,. \tag{34}$$

Analog zu (24) finden wir unter der Voraussetzung (23) die Abschätzung

$$\boxed{\|\widetilde{\mathbf{X}} - \overline{\mathbf{X}}\| \leqq \|\mathbf{C}^{-1}\|\,\frac{\|\delta\,\mathbf{C}\| \cdot \|\overline{\mathbf{X}}\| + \|\mathbf{R}\|}{1 - \|\mathbf{C}^{-1}\| \cdot \|\delta\,\mathbf{C}\|}}\,. \tag{35}$$

3.4. *Angabe einer weiteren Abschätzung.* Bei Matrizengleichungen läßt sich noch eine weitere Abschätzung für die Norm von $\widetilde{\mathbf{X}} - \overline{\mathbf{X}}$ angeben. Ausgehend von den Gleichungen (33), (34) finden wir

$$(\mathbf{C} - \delta\,\mathbf{C})\,(\widetilde{\mathbf{X}} - \overline{\mathbf{X}}) = \delta\,\mathbf{C}\overline{\mathbf{X}} + \mathbf{R}\,. \tag{36}$$

Berücksichtigen wir Gleichung (33), so können wir dafür schreiben

$$\mathbf{E} - (\mathbf{C} - \delta\,\mathbf{C})\,\overline{\mathbf{X}} = \delta\,\mathbf{C}\overline{\mathbf{X}} + \mathbf{R}\,,$$

$$(\mathbf{C} - \delta\,\mathbf{C}) = (\mathbf{E} - \delta\,\mathbf{C}\overline{\mathbf{X}} - \mathbf{R})\,\overline{\mathbf{X}}^{-1}\,.$$

Setzen wir voraus, daß

$$\|\mathbf{R}\| + \|\overline{\mathbf{X}}\| \cdot \|\delta\,\mathbf{C}\| < 1 \tag{37}$$

ist, was wir als erfüllt ansehen können, wenn die Matrix $\mathbf{C}$ nicht fast singulär ist, die Störung $\delta\,\mathbf{C}$ genügend klein ist und $\overline{\mathbf{X}}$ eine ausreichend gute Näherungs-

matrix für die Lösung $\mathbf{X}$ von Gleichung (26) ist, so können wir nach der bekannten Formel (11) schreiben

$$\|(\mathbf{C}-\delta\mathbf{C})^{-1}\| \leqq \frac{\|\overline{\mathbf{X}}\|}{1-\|\delta\mathbf{C}\overline{\mathbf{X}}+\mathbf{R}\|} \leqq \frac{\|\overline{\mathbf{X}}\|}{1-\|\delta\mathbf{C}\|\cdot\|\overline{\mathbf{X}}\|-\|\mathbf{R}\|}.$$

Gleichung (36) liefert uns dann die Abschätzung

$$\|\widetilde{\mathbf{X}}-\overline{\mathbf{X}}\| \leqq \|\overline{\mathbf{X}}\| \frac{\|\mathbf{R}\|+\|\overline{\mathbf{X}}\|\cdot\|\delta\mathbf{C}\|}{1-\|\mathbf{R}\|-\|\overline{\mathbf{X}}\|\cdot\|\delta\mathbf{C}\|}. \tag{38}$$

Zur Anwendung von (38) ist nur die Kenntnis einer brauchbaren Näherungsmatrix $\overline{\mathbf{X}}$ für $\mathbf{C}^{-1}$ und der zugehörigen Defektenmatrix $\mathbf{R}$ bei als gegeben anzusehendem $\|\,\delta\,\mathbf{C}\,\|$ notwendig.

3.5. *Vergleich der Fehlerabschätzungen.* Die Voraussetzung (23) für die Anwendbarkeit der Abschätzung (35) wird bei praktischen Aufgaben in nicht besonders ungünstig gelagerten Fälle wohl meist erfüllt sein. Dafür muß zur Anwendung von (35) eine Abschätzung für die Norm von $\mathbf{C}^{-1}$ bekannt sein. Häufiger sind die Fälle, in denen die Voraussetzungen für die Anwendung der Abschätzungen (29) und (38) nicht oder nur mit numerisch unbrauchbaren Zahlwerten erfüllt sind. Die Abschätzung (29), (30) nimmt insofern eine Sonderstellung gegenüber den anderen Abschätzungen ein, weil durch sie der Einfluß der Ungenauigkeiten der Daten der Aufgabe abgeschätzt werden kann, ohne daß eine Näherungslösung für die Inverse benötigt wird. Die Abschätzung (38) hat gegenüber (35) den Vorteil, daß nicht die Kenntnis der Norm der Inversen notwendig ist. Sie hat gegenüber Formel (29) den Vorteil, daß die Matrix $\mathbf{C}$ nicht überwiegende Hauptdiagonalelemente besitzen muß.

3.6. *Beispiel.* Die Fehlerabschätzungen sollen auf ein Beispiel angewandt werden. Wir legen den Berechnungen die in 1.5. angegebene Matrix $\mathbf{C}$ zugrunde, die mit den dort angenommenen Ungenauigkeiten behaftet sein möge. Für (30) finden wir sofort mit den in § 1 angeführten Zahlwerten

$$\|\mathbf{X}\| \leqq \frac{\frac{1}{100}}{1-\frac{2}{5}} = \frac{1}{60} < 0{,}0167\,.$$

Mit diesem Wert für $\|\,\mathbf{X}\,\|$ ergibt (29)

$$\|\delta\,\mathbf{X}\| \leqq \frac{\frac{1}{60}\cdot\frac{3}{100}}{1-\frac{2}{5}-\frac{3}{100}} = \frac{1}{1140} < 0{,}00088\,. \tag{39}$$

Zur Anwendung von (32) benötigen wir eine Näherungsmatrix $\overline{\mathbf{X}}$ für die Inverse $\mathbf{C}^{-1}$; wir wählen

$$\overline{\mathbf{X}} = \begin{bmatrix} \frac{1}{200} & -\frac{1}{1000} & -\frac{1}{1000} \\ -\frac{1}{500} & \frac{7}{1000} & -\frac{1}{1000} \\ -\frac{1}{1000} & -\frac{1}{1000} & \frac{11}{1000} \end{bmatrix}. \tag{40}$$

Es ist dann

$$\|\overline{\mathbf{X}}\| = \frac{13}{1000},$$

und für die zugehörige Defektenmatrix $\mathbf{R}$ finden wir

$$\|\mathbf{R}\| = \frac{19}{10000},$$

so daß die Anwendung von (32) ergibt

$$\|\mathbf{X}\| \leqq \frac{13}{1000} + \frac{\frac{19}{10000}}{1 - \frac{2}{5}} = \frac{97}{6000} < 0{,}0162\,.$$

(29) liefert somit

$$\|\delta\mathbf{X}\| \leqq \frac{\frac{97}{6000} \cdot \frac{3}{100}}{1 - \frac{2}{5} - \frac{3}{100}} = \frac{97}{114000} < 0{,}00085\,. \tag{41}$$

Formel (38) führt mit der Näherungsmatrix (40) und

$$\|\mathbf{R}\| = \frac{19}{100}$$ [4]

zu der Abschätzung

$$\|\tilde{\mathbf{X}} - \overline{\mathbf{X}}\| \leqq \frac{13}{1000} \cdot \frac{\frac{19}{100} + \frac{13}{1000} \cdot 3}{1 - \frac{19}{100} - \frac{13}{1000} \cdot 3} = \frac{2977}{771000} < 0{,}0039\,. \tag{42}$$

[4] Man beachte, daß die durch (34) definierte Defektenmatrix $\mathbf{R}$ sich von der in (32) verwendeten Matrix $\mathbf{R}$ unterscheidet, welche durch (31) erklärt wird. Man erhält die Matrix $\mathbf{R}$ aus (31) aus der durch (34) definierten Defektenmatrix, indem man diese linksseitig mit $\mathbf{D}^{-1}$ multipliziert.

Wenden wir zum Vergleich Abschätzung (35) für die Näherungsmatrix (40) mit dem exakten Wert für die Norm von $\mathbf{C}^{-1}$

$$\|\mathbf{C}^{-1}\| = \frac{617}{55500} < 0{,}0112 \tag{43}$$

an, so finden wir

$$\|\tilde{\mathbf{X}} - \bar{\mathbf{X}}\| \leqq 0{,}0112 \cdot \frac{3 \cdot \frac{13}{1000} + \frac{19}{100}}{1 - 0{,}0112 \cdot 3} < 0{,}0027\,. \tag{44}$$

Im Beispiel haben die Abschätzungen (39), (41) zu numerisch gleichwertigen Resultaten geführt. Die Kleinheit der erhaltenen Zahlwerte darf aber nicht darüber hinwegtäuschen, daß die Abschätzungen (39), (41) sehr ungünstig sind; denn die Zahlwerte der Elemente der Matrix $\mathbf{C}^{-1}$ sind, wie (43) zeigt, ebenfalls sehr klein. Entsprechende Resultate liefern die Abschätzungen (42) und (44), die auch den starken Einfluß der Ungenauigkeit der Daten zeigen.

§ 4. Abschätzungen für die Norm der Inversen

4.1. *Einleitende Bemerkungen.* In den Abschätzungen (24) und (35) von § 2 und § 3 tritt die Norm der Inversen der Matrix $\mathbf{C}$ auf. Da die Inverse nicht bekannt ist, müssen Abschätzungen für ihre Norm gesucht werden. Besitzt die Matrix $\mathbf{C}$ stark überwiegende Hauptdiagonalelemente, so wird die Abschätzung sehr einfach zu führen sein. Weitere Abschätzungen werden unter der Voraussetzung angegeben, daß eine Näherungsmatrix für die Inverse bekannt ist. Schließlich wird ein Vorgehen erläutert, das unmittelbar die Entnahme der Größenordnung der Norm der Inversen aus einem Eliminationsverfahren gestattet.

4.2. *Eine Abschätzung für Matrizen mit überwiegender Hauptdiagonale.* Besitzt die Matrix $\mathbf{C}$ stark überwiegende Hauptdiagonalelemente, so zerlegen wir sie entsprechend (6) in die Matrizen $\mathbf{D}$ und $\mathbf{B}$ und erhalten

$$\mathbf{C} = \mathbf{D}(\mathbf{E} - \mathbf{D}^{-1}\mathbf{B})\,.$$

Dafür können wir wegen (7)

$$\mathbf{C} = \mathbf{D}(\mathbf{E} - \mathbf{A})$$

schreiben. Es ergibt sich somit

$$\|\mathbf{C}^{-1}\| \leqq \|\mathbf{D}^{-1}\| \cdot \|(\mathbf{E} - \mathbf{A})^{-1}\|\,,$$

und setzen wir voraus, daß $\|\mathbf{A}\| < 1$ ist, was der Annahme einer stark überwiegenden Hauptdiagonale entspricht, so finden wir bei Berücksichtigung der bekannten Formel (11):

$$\boxed{\|\mathbf{C}^{-1}\| \leqq \frac{\|\mathbf{D}^{-1}\|}{1 - \|\mathbf{A}\|}}\,. \tag{45}$$

4.3. *Abschätzungen mit Hilfe des Iterationsverfahrens von Schulz.* Die folgenden Abschätzungen setzen die Kenntnis einer Näherungsmatrix für die Inverse $\mathbf{C}^{-1}$ voraus. Haben wir nach Aufgabenstellung die Inversion einer Matrix durchzuführen, so wird uns auch meist eine Näherungsmatrix für die Inverse bekannt sein. Anders ist es jedoch, wenn wir die Lösung eines linearen Gleichungssystems zu bestimmen haben. Hier bedeutet diese Voraussetzung eine Erschwerung des Problems, da wir dann faktisch nicht die Lösung eines Gleichungssystems sondern eine Matrizeninversion durchzuführen haben. Das müssen wir aber in Kauf nehmen, da bei linearen Gleichungssystemen die Abschätzung (24) faktisch die einzige Abschätzungsformel ist, mit deren Anwendbarkeit wir im allgemeinen immer rechnen können.

Für die Bestimmung der Inversen einer Matrix, d. h. für die Lösung der Matrizengleichung

$$\mathbf{CX} = \mathbf{E}\,, \tag{46}$$

hat Schulz [1] ein Iterationsverfahren angegeben

$$\mathbf{X}_{n+1} = \mathbf{X}_n(2\mathbf{E} - \mathbf{CX}_n)\,, \tag{47}$$

für das der Verfasser [2] Fehlerabschätzungen bewiesen hat. Ist eine Näherungsmatrix $\overline{\mathbf{X}}$ für $\mathbf{C}^{-1}$ bekannt, so können wir sie als Ausgangsmatrix für das Iterationsverfahren von Schulz wählen: $\mathbf{X}_0 = \overline{\mathbf{X}}$. Nach (47) berechnen wir die Matrix $\mathbf{X}_1$. In [2] sind Abschätzungen für die Norm von $\mathbf{X} - \mathbf{X}_1$ angegeben:

$$\|\mathbf{X} - \mathbf{X}_1\| \leqq \frac{q}{1-q}\|\mathbf{X}_1 - \mathbf{X}_0\|, \tag{48}$$

$$\|\mathbf{X} - \mathbf{X}_1\| \leqq \frac{q^2}{1-q}\|\mathbf{X}_0\|\,. \tag{49}$$

Dabei gilt

$$q = \|\mathbf{E} - \mathbf{CX}_0\|\,, \tag{50}$$

und es wird der Zahlwert von q unmittelbar durch das Verfahren geliefert. Die Abschätzungen (48) und (49) sind unter der Voraussetzung $q<1$ gültig. Zum Vergleich der Abschätzungen (48), (49) sei auf [2] verwiesen.

Wegen (46) ist $\mathbf{X} = \mathbf{C}^{-1}$, und wir erhalten für (48), (49):

$$\boxed{\|\mathbf{C}^{-1}\| \leqq \|\mathbf{X}_1\| + \frac{q}{1-q}\|\mathbf{X}_1 - \mathbf{X}_0\|} \tag{51}$$

$$\boxed{\|\mathbf{C}^{-1}\| \leqq \|\mathbf{X}_1\| + \frac{q^2}{1-q}\|\mathbf{X}_0\|}\,. \tag{52}$$

Zu einer noch einfacheren Abschätzung kommen wir, wenn wir die Norm von $\mathbf{X}_1$ in (52) abschätzen. Wegen (47) ist

$$\mathbf{X}_1 = \mathbf{X}_0 + \mathbf{X}_0(\mathbf{E} - \mathbf{CX}_0)\,.$$

Damit ergibt sich unter Berücksichtigung von (50)

$$\|\mathbf{X}_1\| \leqq \|\mathbf{X}_0\| + \|\mathbf{X}_0\| \cdot \|\mathbf{E} - \mathbf{C}\mathbf{X}_0\| = \|\mathbf{X}_0\| (1 + q) .$$

Tragen wir dieses Ergebnis in (52) ein, so finden wir

(53)
$$\boxed{\|\mathbf{C}^{-1}\| \leqq \frac{\|\mathbf{X}_0\|}{1 - q}} .$$

Zur Anwendung der Abschätzung (53) ist es damit gar nicht mehr notwendig, daß ein Schritt des Iterationsverfahrens durchgeführt wird. Die zur Abschätzung nach (53) bei Kenntnis einer Näherungsmatrix $\mathbf{X}_0$ wesentlich zu leistende Rechenarbeit besteht in der Ermittlung von q, wozu faktisch eine Matrizenmultiplikation auszuführen ist. Bei der Abschätzung nach (51) und (52) haben wir die erste Iterierte $\mathbf{X}_1$ zu berechnen, wozu wesentlich zwei Matrizenmultiplikationen erforderlich sind.

Die Abschätzung (53) ist zweifellos numerisch am einfachsten, dafür aber auch schlechter als (52) und (51). Formel (51) führt zu den numerisch besten Resultaten. Welche der Abschätzungen (51) bis (53) in der Praxis Verwendung finden wird, hängt wesentlich von der Größenordnung von q ab.

Zur Abschätzung (53) können wir auch leicht auf einem anderen Wege gelangen. Nach (38) läßt sich ja die Norm der Differenz zwischen der Lösung $\tilde{\mathbf{X}} = \mathbf{X} = \mathbf{C}^{-1}$ der Gleichung (46) und der Näherungslösung $\overline{\mathbf{X}}$ dieser Gleichung abschätzen, indem in (38) $\| \delta \mathbf{C} \| = 0$ gesetzt wird. Wir finden dann

$$\|\mathbf{C}^{-1}\| \leqq \|\overline{\mathbf{X}}\| + \frac{\|\mathbf{R}\| \cdot \|\overline{\mathbf{X}}\|}{1 - \|\mathbf{R}\|} = \frac{\|\overline{\mathbf{X}}\|}{1 - \|\mathbf{R}\|} .$$

Beachten wir, daß oben $\overline{\mathbf{X}} = \mathbf{X}_0$ gesetzt wurde und daß wegen (50) und der Definitionsgleichung (34) für die Defektenmatrix $q = \| \mathbf{R} \|$ ist, so erkennen wir, daß die erhaltene Abschätzung mit (53) identisch ist. In derselben Weise kann ausgehend von (35) Abschätzung (53) erhalten werden.

4.4. *Beispiel.* Wir wollen die erhaltenen Fehlerabschätzungen auf die Mustermatrix $\mathbf{C}$ von § 1 anwenden. Die Abschätzung nach (45) ist trivial. Zur Anwendung der Fehlerformeln von 4.3. benötigen wir eine Ausgangsmatrix für das Iterationsverfahren. Als Ausgangsmatrix $\mathbf{X}_0$ wählen wir die Matrix (40) von § 3. Dann ist

$$\|\mathbf{X}_0\| = \frac{13}{1000}, \quad q = \|\mathbf{E} - \mathbf{C}\mathbf{X}_0\| = \frac{19}{100},$$

und durch Berechnung von $\mathbf{X}_1$ nach (47) lassen sich leicht die Werte

$$\|\mathbf{X}_1\| = \frac{11}{1000}, \qquad \|\mathbf{X}_1 - \mathbf{X}_0\| = \frac{11}{10000}$$

bestätigen.

Die Ergebnisse der Fehlerabschätzungen (45) und (51) bis (53) sind in der folgenden Tabelle zusammengestellt und mit dem exakten Wert von $\| \mathbf{C}^{-1} \|$ verglichen worden. In der Tabelle stehen außerdem neben diesen Werten die Resultate der Abschätzungen von $\| \tilde{\boldsymbol{x}} - \bar{\boldsymbol{x}} \|$ nach (24) für das Beispiel von § 1 und von $\| \tilde{\mathbf{X}} - \bar{\mathbf{X}} \|$ nach (35) mit den jeweiligen Schranken für die Norm von $\mathbf{C}^{-1}$. Zum Vergleich sind in der Tabelle schließlich noch die bereits früher in § 1 und § 3 ermittelten Abschätzungen von $\| \delta \boldsymbol{x} \|$ und $\| \delta \mathbf{X} \|$ sowie die Abschätzung von $\| \tilde{\mathbf{X}} - \bar{\mathbf{X}} \|$ nach (38) angeführt worden.

	Abschätzung von $\|\mathbf{C}^{-1}\|$	Abschätzung von $\|\tilde{\boldsymbol{x}} - \bar{\boldsymbol{x}}\|$ bzw. $\|\delta \boldsymbol{x}\|$	Abschätzung von $\|\tilde{\mathbf{X}} - \bar{\mathbf{X}}\|$ bzw. $\|\delta \mathbf{X}\|$
exakt	0,0112	0,117	0,0027
nach (45)	0,0167	0,177	0,0041
nach (51)	0,0113	0,118	0,0027
nach (52)	0,0116	0,121	0,0028
nach (53)	0,0161	0,170	0,0039
nach (38)		—	0,0039
nach (13), (14) bzw. (29), (30)		0,31	0,00088
nach (13), (16) bzw. (29), (32)		0,178	0,00085
nach (13) bzw. (29); $\| \boldsymbol{x} \|$, $\| \mathbf{X} \|$ exakt		0,176	0,00059

4.5. *Vorgehen bei Verwendung eines Eliminationsverfahrens.* So bestechend auch die Abschätzungen in 4.3. sind, so bedeutet doch die Voraussetzung, daß eine Näherungsmatrix für die Inverse bekannt sein muß, oft eine wesentliche Erschwerung der Aufgabe. Dieser Nachteil haftet der Abschätzung (45) nicht an, dafür setzt sie aber eine stark überwiegende Hauptdiagonale voraus. In den Fällen, in denen die Hauptdiagonale nicht stark überwiegt, werden wir jedoch häufig das Gleichungssystem nach einem Eliminationsverfahren lösen oder die Matrizeninversion mit Hilfe eines Eliminationsverfahrens durchführen. Besonders günstig gestalten sich die Verhältnisse beim GAUSS—JORDANschen Verfahren, bei dem die Ausgangsmatrix auf Diagonalform transformiert wird. In diesem Falle kann die Größenordnung der Norm von $\mathbf{C}^{-1}$ mit praktisch meist ausreichender Genauigkeit unmittelbar aus dem Verfahren entnommen werden. Zur Erläuterung dieses Vorgehens soll das Verfahren von GAUSS—JORDAN kurz skizziert werden. Das Verfahren konstruiert eine Matrix $\mathbf{C}^{(n)}$ [5] nach der Vorschrift

$$\mathbf{C}^{(n)} = \mathbf{Q}^{(n)} \mathbf{Q}^{(n-1)} \cdot \ldots \cdot \mathbf{Q}^{(1)} \mathbf{C} . \tag{54}$$

Dabei sind die Matrizen $\mathbf{Q}^{(j)}$ durch die Quotienten $q_i^{(j)}$ des Verfahrens bestimmt, welche bekanntlich so gewählt werden, daß die Matrix $\mathbf{C}$ auf Diagonalform

[5] Mit n wird die Ordnung der Matrix $\mathbf{C}$ bezeichnet.

transformiert wird, also $\mathbf{C}^{(n)}$ eine Diagonalmatrix ist. Die Matrizen $\mathbf{Q}^{(i)}$ besitzen folgende Gestalt:

$$\underline{Q}^{(1)} = \begin{bmatrix} 1 & & & \\ q_2^{(1)} & 1 & & 0 \\ \vdots & & \ddots & \\ q_n^{(1)} & 0 & & 1 \end{bmatrix}, \quad \underline{Q}^{(2)} = \begin{bmatrix} 1 & q_1^{(2)} & & \\ 0 & 1 & & 0 \\ 0 & q_3^{(2)} & \ddots & \\ \vdots & & & 1 \\ 0 & q_n^{(2)} & 0 & 1 \end{bmatrix}, \ldots, \underline{Q}^{(n)} = \begin{bmatrix} 1 & & & q_1^{(n)} \\ & \ddots & & \vdots \\ & 0 & 1 & q_{n-1}^{(n)} \\ 0 & & & 1 \end{bmatrix}$$

Aus (54) finden wir nun sofort

$$\|\mathbf{C}^{-1}\| \leqq \|\mathbf{Q}^{(n)}\| \cdot \|\mathbf{Q}^{(n-1)}\| \cdot \ldots \cdot \|\mathbf{Q}^{(1)}\| \cdot \|(\mathbf{C}^{(n)})^{-1}\| \,. \tag{55}$$

$\mathbf{C}^{(n)}$ ist die durch das Verfahren gelieferte Diagonalmatrix, die Norm ihrer Inversen also leicht zu berechnen. Auch die Norm der Quotientenmatrizen $\mathbf{Q}^{(i)}$ kann unmittelbar aus dem Verfahren entnommen werden. Schreiben wir

$$\bar{q}_j = \max_i |q_i^{(j)}| \,,$$

so können wir für (55), wenn wir unter der Norm einer Matrix wie verabredet die Zeilensummennorm verstehen, wegen der angegebenen Gestalt der $\mathbf{Q}^{(i)}$ die Darstellung

$$\boxed{\|\mathbf{C}^{-1}\| \leqq (1 + \bar{q}_n)(1 + \bar{q}_{n-1}) \cdot \ldots \cdot (1 + \bar{q}_1) \cdot \|(\mathbf{C}^{(n)})^{-1}\|} \tag{56}$$

geben. Das Verfahren liefert also unmittelbar eine Abschätzung für die Norm der Inversen $\mathbf{C}^{-1}$, allerdings ohne Berücksichtigung der bei der Durchführung des Verfahrens auftretenden Rundungsfehler und ihrer Fortpflanzung, was zunächst paradox erscheinen mag. Aber häufig wird für Überschlagsrechnungen die durch (56) gelieferte Abschätzung ausreichen. Auch können die Rechnungen mit einigen Schutzstellen durchgeführt werden, um eine größere Sicherheit gegen das Häufen der Rundungsfehler zu haben.

Wenden wir auf die in § 1 angegebene Matrix $\mathbf{C}$ das Verfahren von GAUSS—JORDAN an und führen wir die Rechnungen vollkommen mit rationalen Zahlen durch, so daß keine Rundungsfehler auftreten, so erhalten wir für die Abschätzung nach (56)

$$\|\mathbf{C}^{-1}\| < 0{,}0188 \,.$$

Das geschilderte Vorgehen läßt sich auch beim Verfahren von GAUSS anwenden, bei dem die Matrix $\mathbf{C}$ auf Dreiecksform transformiert wird. Die Matrizen $\mathbf{Q}^{(1)}, \ldots, \mathbf{Q}^{(n-1)}$ sind hier untere Dreiecksmatrizen. Unmittelbar gelangen wir so zu einer Abschätzung für $\|\mathbf{C}^{-1}\|$, die zu (56) analog ist. Erschwerend ist hier nur, daß die Norm der Inversen der durch das Verfahren gelieferten Dreiecksmatrix $\mathbf{C}^{(n-1)}$ bestimmt werden muß.

§ 5. Vergleich mit den Ergebnissen der Literatur

Dem in dieser Arbeit behandelten Problemkreis ist in der numerischen Mathematik stets große Aufmerksamkeit geschenkt worden. Die Ermittlung nicht zu mühsamer und nicht zu grober Abschätzungen für diesen Fragenkomplex wird auch weiterhin im Interesse der Anwendung liegen. So werden z. B. in [3] Angaben gemacht, wie ein Anhaltspunkt für die bei der Lösung

eines linearen Gleichungssystems auftretenden Abrundungsfehler und ihre Fortpflanzung aus den Defekten erhalten werden kann. Das kann für viele praktische Fälle oft ausreichen, auch wenn es sich dabei nicht um eine exakte Fehlerabschätzung handelt. Eine exakte Fehlerabschätzung hat WITTMEYER [4] bereits 1936 angegeben. Die Abschätzung wurde von COLLATZ [5] modifiziert und verbessert. Es ist wichtig, die Ergebnisse der Überlegungen von COLLATZ kurz zu skizzieren. In der bis jetzt benutzten Bezeichnungsweise schätzt COLLATZ die Differenz der Lösungen der Gleichungen (20), (21) ab und beweist die Formel

$$|\tilde{\boldsymbol{x}} - \bar{\boldsymbol{x}}| \leqq \frac{|\overline{\delta \mathbf{C}}| \cdot |\bar{\boldsymbol{x}}| + |\delta \boldsymbol{c}| + |\boldsymbol{r}|}{|\underline{\mathbf{C}}| - |\overline{\delta \mathbf{C}}|}\,, \tag{57}$$

die gültig ist, sobald der Nenner positiv ist. Dabei wird für einen Vektor $\boldsymbol{z}$ mit den Komponenten z^i der Betrag $|\boldsymbol{z}|$ durch die Gleichung

$$|\boldsymbol{z}| = \sqrt{\sum_{i=1}^{n} |z^i|^2}$$

definiert. Das entspricht aber gerade der Quadratsummennorm[6] für den Vektor $\boldsymbol{z}$. In (57) tritt weiterhin der obere Betrag $|\bar{\mathbf{C}}|$ und der untere Betrag $|\underline{\mathbf{C}}|$ einer Matrix $\mathbf{C}$ auf, wobei für eine beliebige Matrix $\mathbf{C}$ und einen beliebigen Vektor $\boldsymbol{z}$ die Ungleichung

$$|\underline{\mathbf{C}}| \cdot |\boldsymbol{z}| \leqq |\mathbf{C}\boldsymbol{z}| \leqq |\bar{\mathbf{C}}| \cdot |\boldsymbol{z}|$$

besteht. Der obere Betrag der Matrix $\mathbf{C}$ kann durch die im Sinne der Quadratsummennorm gebildete Norm der Matrix $\mathbf{C}$ nach oben abgeschätzt werden. Das Hauptproblem bei der Formel (57) ist also die Bestimmung des unteren Betrags der Matrix $\mathbf{C}$. Mit dieser Frage befaßt sich die Arbeit von BARTSCH [7], in welcher der Sonderfall einer positiv definiten hermiteschen Matrix untersucht wird; in diesem Falle ist $|\underline{\mathbf{C}}|$ die kleinste und $|\overline{\mathbf{C}}|$ die größte charakteristische Zahl der Matrix $\mathbf{C}$. Die Arbeit von BARTSCH erscheint in unserem Zusammenhang als eine Ergänzung zu den angeführten Überlegungen. Ein unterer Betrag einer Matrix $\mathbf{C}$ kann aber auch leicht mit Hilfe der Norm der inversen Matrix angegeben werden. Es ist ja

$$\|\boldsymbol{z}\| = \|\mathbf{C}^{-1}\mathbf{C}\boldsymbol{z}\| \leqq \|\mathbf{C}^{-1}\| \cdot \|\mathbf{C}\boldsymbol{z}\|$$

und damit

$$\frac{\|\boldsymbol{z}\|}{\|\mathbf{C}^{-1}\|} \leqq \|\mathbf{C}\boldsymbol{z}\|\,. \tag{58}$$

Verstehen wir (58) im Sinne der Quadratsummennorm, so haben wir für die Abschätzung (57) von COLLATZ in $1/\|\mathbf{C}^{-1}\|$ einen unteren Betrag für die Matrix $\mathbf{C}$ gefunden. Setzen wir diesen Wert für $|\underline{\mathbf{C}}|$ in (57) ein, so erhalten wir sofort die Abschätzung (24), welche nur auf Quadratsummennorm zu beziehen ist. Damit ist ein Übergang von (57) zu (24) hergestellt.

In diesem Zusammenhang sei auch auf eine Arbeit des Verfassers [8] verwiesen, in der bereits über einige der erörterten Probleme kurz berichtet wurde.

(Eingegangen: 15. März, 1960.)

[6] Zur Erklärung dieser Begriffe vergleiche man etwa [6].

LITERATURVERZEICHNIS

[1] Schulz, G.: »Iterative Berechnung der reziproken Matrix.« *Zeitschrift für angewandte Mathematik und Mechanik* **13** (1933) 57—59.
[2] Dück, W.: »Fehlerabschätzungen für das Iterationsverfahren von Schulz zur Bestimmung der Inversen einer Matrix.« *Zeitschrift für angewandte Mathematik und Mechanik* **40** (1960) 192—194.
[3] Collatz, L.: *Numerische und graphische Methoden. Handbuch der Physik, Bd. II.* Springer-Verlag, 1955, S. 388.
[4] Wittmeyer, H.: »Einfluß der Änderung einer Matrix auf die Lösung des zugehörigen Gleichungssystems, sowie auf die charakteristischen Zahlen und die Eigenvektoren.« *Zeitschrift für angewandte Mathematik und Mechanik* **16** (1936) 287—300.
[5] Collatz, L.: »Zur Fehlerabschätzung bei linearen Gleichungssystemen.« *Zeitschrift für angewandte Mathematik und Mechanik* **34** (1954) 71—72.
[6] Collatz, L.: »Einige Anwendungen funktionalanalytischer Methoden in der praktischen Analysis.« *Zeitschrift für angewandte Mathematik und Physik* **4** (1953) 327—357.
[7] Bartsch, H.: »Abschätzungen für die kleinste charakteristische Zahl einer positivdefiniten hermiteschen Matrix.« *Zeitschrift für angewandte Mathematik und Mechanik* **34** (1954) 72—74.
[8] Dück, W.: »Zur Abschätzung der Fortpflanzung der Datenfehler bei linearen Gleichungssystemen nach der Formel von Wittmeyer—Collatz.« *Zeitschrift für angewandte Mathematik und Mechanik* **41** (1961).

ОЦЕНКА НЕУСТРАНИМОЙ ПОГРЕШНОСТИ РЕШЕНИЙ СИСТЕМ ЛИНЕЙНЫХ УРАВНЕНИЙ И МАТРИЧНЫХ УРАВНЕНИЙ

W. DÜCK

Резюме

Пусть дана система линейных уравнений вида

$$\boldsymbol{x} = \mathbf{A}\,\boldsymbol{x} + \boldsymbol{a}$$

Обозначим через $\delta\mathbf{A}$ погрешность матрицы $\mathbf{A}$, а через $\delta\,\boldsymbol{a}$ погрешность вектора $\boldsymbol{a}$. Если

$$\|\mathbf{A}\| = \max_i \sum_{k=1}^{n} |a_{ik}| < 1\,,\ \text{более того}\ \|\mathbf{A}\| + \|\delta\,\mathbf{A}\| < 1\,,$$

то для неустранимой погрешности $\delta\,\boldsymbol{x}$ решения $\boldsymbol{x}$ имеет место оценка

$$\|\delta\,\boldsymbol{x}\| = \max_i |\delta x_i| \leqq \frac{\|\boldsymbol{x}\|\,\|\delta\,\mathbf{A}\| + \|\delta\,\boldsymbol{a}\|}{1 - \|\mathbf{A}\| - \|\delta\,\mathbf{A}\|}\,.$$

Если $\boldsymbol{x}$ не известно, то его можно оценить с помощью неравенства

$$\|\boldsymbol{x}\| \leqq \frac{\|\boldsymbol{a}\|}{1 - \|\mathbf{A}\|}$$

а если известно приближенное решение $\overline{\boldsymbol{x}}$, то с помощью неравенства

$$\|\boldsymbol{x}\| \leqq \|\overline{\boldsymbol{x}}\| + \frac{\|\boldsymbol{r}\|}{1 - \|\mathbf{A}\|}\,,$$

где

$$\boldsymbol{r} = \mathbf{A}\,\overline{\boldsymbol{x}} + \boldsymbol{a} - \overline{\boldsymbol{x}}\,.$$

Пусть далее дана линейная система уравнений вида

$$\mathbf{C}\,\boldsymbol{x} = \boldsymbol{c}$$

погрешность матрицы $\mathbf{C}$ есть $\delta\mathbf{C}$, а вектора a $\delta\boldsymbol{a}$. Пусть $\tilde{\boldsymbol{x}}$ есть решение уравнения

$$(\mathbf{C} - \delta\mathbf{C})\,\tilde{\boldsymbol{x}} = (\boldsymbol{c} + \delta\boldsymbol{c})$$

$\boldsymbol{r}$ обозначает разность

$$\boldsymbol{c} - \mathbf{C}\,\bar{\boldsymbol{x}}$$

где $\bar{\boldsymbol{x}}$ известное приближенное решение уравнения $\mathbf{C}\,\boldsymbol{x} = \boldsymbol{c}$. Для оценки неустранимой погрешности решения получается уравнение

$$\|\tilde{\boldsymbol{x}} - \bar{\boldsymbol{x}}\| \leqq \|\mathbf{C}^{-1}\| \frac{\|\delta\mathbf{C}\|\,\|\bar{\boldsymbol{x}}\| + \|\delta\boldsymbol{c}\| + \|\boldsymbol{r}\|}{1 - \|\mathbf{C}^{-1}\|\,\|\delta\mathbf{C}\|},$$

если

$$\|\mathbf{C}^{-1}\|\,\|\delta\mathbf{C}\| < 1\,.$$

Если известно приближенное значение $\mathbf{X}_0$ матрицы, обратной матрице $\mathbf{C}$, то преобразуя итерационный метод SCHULZ-а [1] и оценку погрешности автора [2], при обозначениях

$$\mathbf{X}_1 = \mathbf{X}_0\,(2\,\mathbf{E} - \mathbf{C}\mathbf{X}_0)$$

и

$$q = \|\mathbf{E} - \mathbf{C}\mathbf{X}_0\|\,,$$

получаем

$$\|\mathbf{C}^{-1}\| \leqq \|\mathbf{X}_1\| + \frac{q}{1-q}\,\|\mathbf{X}_1 - \mathbf{X}_0\|\,,$$

$$\|\mathbf{C}^{-1}\| \leqq \|\mathbf{X}_1\| + \frac{q^2}{1-q}\,\|\mathbf{X}_0\|$$

и

$$\|\mathbf{C}^{-1}\| \leqq \frac{\|\mathbf{X}_0\|}{1-q}\,,$$

так что приведенное выше неравенство для оценки $\|\tilde{\boldsymbol{x}} - \bar{\boldsymbol{x}}\|$ можно практически исзпользовать.

Аналогичные оценки получаются в § 3 для матричного уравнения

$$\mathbf{C}\,\mathbf{X} = \mathbf{E}\,.$$

§ 4 занимается оценкой нормы обратной матрицы.

§ 5 сравнивает полученные результаты с полученными ранее.