# EIN NEUES APPROXIMATIONSVERFAHREN ZUR SPEKTRALZERLEGUNG VON MATRIZEN

Von

T. FREY

Lehrstuhl für Mathematik, Fakultät für Elektrotechnik
Technische Universität, Budapest

(Eingegangen am 3. September 1971)

**1.** Die Dreiecksiteration [1] bzw. einige neuere Varianten dieses Verfahrens, wie z. B. der $L$—$R$- resp. der $Q$—$R$-Algorithmus [2, 3] sind die am meisten angewandten Verfahren für die Spektralzerlegung von reellen, quadratischen Matrizen. Alle diese Verfahren haben jedoch zwei wesentliche Nachteile: einerseits besteht Konvergenz (dann und) nur dann, wenn alle Eigenwerte reell sind und verschiedene Modulen haben, anderseits ist die Konvergenz langsam, nämlich linear. Wir geben eine solche Weiterentwicklung der Grundidee der Dreiecksiteration an, durch die alle genannten Nachteile behoben werden: die Konvergenz ist quadratisch, und besteht auch dann, wenn komplexe oder mehrfache Eigenwerte auftreten. Einen Nachteil, u. zw. von algorithmischem Charakter hat jedoch das hier angegebene Verfahren: es arbeitet durch eine Halbierung der Dimensionszahl, und deshalb müssen mehrere Iterationsprozesse einander nachgeschaltet werden. Die Grundidee könnte man auch so anwenden, daß nur ein Iterationsprozeß erforderlich sei, jedoch wären dann die Rechenformeln zu kompliziert; es werden einige Hinweise auch auf diese Möglichkeit angegeben.

**2.** Es ist nicht schwer einzusehen (s. Satz 1), daß man zu jeder reellen quadratischen Matrix $\mathbf{A}$, die in der Form

$$\begin{pmatrix} \mathbf{A}_{11} & \mathbf{A}_{12} \\ \mathbf{A}_{21} & \mathbf{A}_{22} \end{pmatrix}$$

partizioniert ist, reelle obere bzw. untere Blockdreiecksmatrizen $\mathbf{S}$, $\mathbf{T}$

$$\mathbf{S} = \begin{pmatrix} \mathbf{S}_{11} & \mathbf{S}_{12} \\ \mathbf{0} & \mathbf{S}_{22} \end{pmatrix}; \quad \mathbf{T} = \begin{pmatrix} \mathbf{E}_{11} & \mathbf{0} \\ \mathbf{H} & \mathbf{E}_{22} \end{pmatrix}$$

mit

$$\mathbf{A} = \mathbf{T}\mathbf{S}\mathbf{T}^{-1} \tag{1}$$

angeben kann, wo $\mathbf{A}_{11}$, $\mathbf{S}_{11}$ bzw. $\mathbf{A}_{22}$, $\mathbf{S}_{22}$ quadratische Blöcke sind, und mindestens eines dieser Paare eine gerade Dimensionszahl besitzt ($\mathbf{E}_{11}$ bzw. $\mathbf{E}_{22}$ sind

Einheitsmatrizen mit derselben Dimensionszahl wie $\mathbf{A}_{11}$, $\mathbf{S}_{11}$ bzw. wie $\mathbf{A}_{22}$, $\mathbf{S}_{22}$). Es ist nun bekannt, daß durch die Spektralzerlegung von $\mathbf{S}_{11}$ bzw. von $\mathbf{S}_{22}$ sogleich auch die Spektralzerlegung von $\mathbf{S}$ angegeben wird. Die Angabe von $\mathbf{S}$ und $\mathbf{T}$ bedeutet also, daß wir uns mit der Spektralzerlegung von $\mathbf{S}_{11}$ bzw. $\mathbf{S}_{22}$ beschäftigen, und somit die ursprüngliche Dimensionszahl ungefähr halbieren können. Ist nun $\mathbf{S}_{11}$ bzw. $\mathbf{S}_{22}$ ein- oder zweidimensional, so kann die Spektralzerlegung durch elementare Operationen erfolgen. Die Hauptfrage ist also, die Blockdreiecksmatrizen $\mathbf{S}$ und $\mathbf{T}$ auszusuchen. Dazu kann einerseits die Verallgemeinerung der Dreiecksiteration angewandt werden

$$\mathbf{A}\mathbf{T}_{n-1} = \mathbf{T}_n \mathbf{S}_n, \tag{2}$$

u. zw. bei beliebiger Wahl von $\mathbf{T}_0$. Es soll gezeigt werden (s. Satz 2), daß die verallgemeinerte Dreiecksiteration (2) konvergent ist, falls $\mathbf{A}$ mindestens zwei Eigenwerte mit verschiedenen Modulen besitzt und die Partizionierungen von $\mathbf{A}$, $\mathbf{T}$ und $\mathbf{S}$ der Regel entsprechen, und daß mindestens einer der Blöcke $\mathbf{A}_{11}$ und $\mathbf{A}_{22}$ eine gerade Dimensionszahl hat. Jedoch ist die Konvergenz des Prozesses (2) nur linear. Es läßt sich jedoch auch eine quadratische Konvergenz erreichen — u. zw. auch dann, wenn alle Eigenwerte von $\mathbf{A}$ einen gleichen Modulus haben —, falls eine hinreichend gute Approximation $\mathbf{T}_n$ von $\mathbf{T}$ bekannt ist (s. Satz 3). Ist nämlich $\mathbf{T}_n$ bekannt, so wird dadurch die Gleichung

$$\mathbf{A}\mathbf{T}_n = \mathbf{T}^{(n)} \mathbf{S}^{(n)} \tag{3}$$

befriedigende obere bzw. untere Blockdreiecksmatrix $\mathbf{S}^{(n)}$ bzw. $\mathbf{T}^{(n)}$ bestimmt. Wäre nun $\mathbf{T}_n = \mathbf{T}^{(n)}$, so hätten wir schon die Lösung von (1) in der Hand. Das ist natürlich im allgemeinen nicht der Fall, doch ist $\mathbf{T}_n$ eine gute Annäherung von $\mathbf{T}$, dann gibt eine geeignete geringe Variation von $\mathbf{T}_n$ — und dadurch von $\mathbf{T}^{(n)}$ und $\mathbf{S}^{(n)}$ — die gesuchte Lösung. Es soll diese Variation von $\mathbf{T}_n$, $\mathbf{S}^{(n)}$ und $\mathbf{T}^{(n)}$ in der Form

$$\Delta\mathbf{T}_n = \begin{pmatrix} \mathbf{0} & \mathbf{0} \\ \delta & \mathbf{0} \end{pmatrix}; \quad \Delta\mathbf{T}^{(n)} = \begin{pmatrix} \mathbf{0} & \mathbf{0} \\ \delta^{(n)} & \mathbf{0} \end{pmatrix}; \quad \Delta\mathbf{S}^{(n)} = \begin{pmatrix} \Delta_{11} & \Delta_{12} \\ \mathbf{0} & \Delta_{22} \end{pmatrix} \tag{4}$$

gesucht werden. Diese Variationen sollen einerseits die Gleichung

$$\mathbf{T}_n + \Delta\mathbf{T}_n = \mathbf{T}^{(n)} + \Delta\mathbf{T}^{(n)},$$

d. h.

$$\mathbf{H}_n + \delta = \mathbf{H}^{(n)} + \delta^{(n)}, \tag{5}$$

andererseits die Gleichung

$$\mathbf{A} \cdot (\mathbf{T}_n + \Delta\mathbf{T}_n) = (\mathbf{T}^{(n)} + \Delta\mathbf{T}^{(n)}) \cdot (\mathbf{S}^{(n)} + \Delta\mathbf{S}^{(n)}),$$

d. h. die Gleichungen

$$\mathbf{A}_{11}+\mathbf{A}_{12}\mathbf{H}_n+\mathbf{A}_{12}\delta=\mathbf{S}_{11}^{(n)}+\Delta_{11}, \tag{6}$$

$$\mathbf{A}_{12}=\mathbf{S}_{12}^{(n)}+\Delta_{12}, \tag{7}$$

$$\mathbf{A}_{21}+\mathbf{A}_{22}\mathbf{H}_n+\mathbf{A}_{22}\delta=\mathbf{H}^{(n)}\cdot\mathbf{S}_{11}^{(n)}+\mathbf{H}^{(n)}\Delta_{11}+\delta^{(n)}\cdot\mathbf{S}_{11}^{(n)}+\delta^{(n)}\cdot\Delta_{11}, \tag{8}$$

$$\mathbf{A}_{22}=\mathbf{H}^{(n)}\mathbf{S}_{12}^{(n)}+\mathbf{H}^{(n)}\cdot\Delta_{12}+\delta^{(n)}\mathbf{S}_{12}^{(n)}+\delta^{(n)}\cdot\Delta_{12}+\mathbf{S}_{22}+\Delta_{22} \tag{9}$$

befriedigen. Das Gleichungssystem (5)—(9) ist in geschlossener Form nicht auflösbar. Wird jedoch das Glied zweiter Ordnung $\delta^{(n)}\Delta_{11}$ in (8) vernachlässigt, so erhält man die folgende, näherungsweise Lösung des betrachteten Systems:

$$\Delta_{11}=\mathbf{A}_{12}-\mathbf{S}_{12}^{(n)} \tag{10}$$

$$\delta^{(n)}=\mathbf{H}_n-\mathbf{H}^{(n)}+\delta; \tag{11}$$

$$\Delta_{11}=\mathbf{A}_{11}+\mathbf{A}_{12}\mathbf{H}_n+\mathbf{A}_{12}\delta-\mathbf{S}_{11}^{(n)}; \tag{12}$$

$$\mathbf{A}_{22}\delta-\mathbf{H}^{(n)}\cdot\Delta_{11}-\delta^{(n)}\mathbf{S}_{11}^{(n)}\cong\mathbf{H}^{(n)}\cdot\mathbf{S}_{11}^{(n)}-\mathbf{A}_{21}-\mathbf{A}_{22}\mathbf{H}_n,$$

folglich, nach (11)—(12)

$$\begin{aligned}(\mathbf{A}_{22}-\mathbf{H}^{(n)}\mathbf{A}_{11})\,\delta-\delta\mathbf{S}_{11}^{(n)}\cong{}&\mathbf{H}^{(n)}\mathbf{S}_{11}^{(n)}-\mathbf{A}_{21}-\mathbf{A}_{22}\mathbf{H}_n+\\&+\mathbf{H}^{(n)}\mathbf{A}_{11}+\mathbf{H}^{(n)}\mathbf{A}_{12}\mathbf{H}_n-\mathbf{H}^{(n)}\mathbf{S}_{11}^{(n)}+\mathbf{H}_n\mathbf{S}_{11}^{(n)}-\mathbf{H}^{(n)}\mathbf{S}_{11}^{(n)},\end{aligned} \tag{13}$$

und schließlich

$$\Delta_{22}=-\mathbf{H}^{(n)}\mathbf{S}_{11}^{(n)}-\mathbf{H}^{(n)}\Delta_{12}+\mathbf{A}_{22}-\delta^{(n)}\mathbf{S}_{12}^{(n)}-\delta^{(n)}\cdot\Delta_{12}-\mathbf{S}_{22}. \tag{14}$$

Den Annäherungswert von $\delta$ — gewonnen aus der angenäherten Gleichung (13) — in (11) eingesetzt erhält man die Annäherung von $\delta^{(n)}$, und dadurch eine verbesserte Annäherung von $\mathbf{T}$. Das Newtonsche Verfahren (11)—(14) sichert nun eine quadratische Konvergenz. (Es sei hier bemerkt, daß mit Hilfe dieser ersten Annäherung in (8) auch das quadratische Glied abgeschätzt und somit eine zweite Verbesserung in den Unbekannten erreicht wird.)

**3.** In diesem Abschnitt sollen die oben benutzten Sätze in exakter Weise formuliert und bewiesen werden.

Satz **1.** *Genügt die Matrix* $\mathbf{A}$ *bzw. ihre Partitionierung den in* **2** *angegebenen Voraussetzungen, so bestehen die Relation* (1) *und in der angegebenen Darstellung partizionierbare Blockdreiecksmatrizen* $\mathbf{S}$ *und* $\mathbf{T}$.

Beweis: Es sei $\mathbf{A}$ in der Jordanschen Normalform

$$\mathbf{A}=[s_1, s_2, \ldots, s_n]\cdot\mathbf{J}\cdot\begin{bmatrix}z_1^*\\ z_2^*\\ \cdot\\ \cdot\\ z_n^*\end{bmatrix} \tag{15}$$

dargestellt, u. zw. in einer Reihenfolge der rechtsseitigen Hauptvektoren, daß, einerseits, in der Hauptdiagonalen von $[s_1, s_2, \ldots, s_n]$ alle Elemente von 0 verschieden seien, und daß anderseits, die zu konjugierten Eigenwerten gehörenden und somit konjugierten Hauptvektoren nebeneinander stehen. Im ersten Schritt werden aus der Darstellung (15) die komplexen Zahlen eliminiert. (Die Paritätsbedingungen, die für die Dimensionszahlen von $\mathbf{A}_{11}$ und $\mathbf{A}_{22}$ vorgeschrieben wurden, hängen mit diesem Schritt zusammen.) Sind z. B. $s_i$ und $s_{i+1}$ konjugiert, so sind

$$\hat{s}_i = \frac{1}{2}(s_i + s_{i+1}) \quad \text{und} \quad \hat{s}_{i+1} = \frac{1}{2j}(s_i - s_{i+1}) \tag{16}$$

schon reell und linear unabhängig. Betrachten wir nun einerseits das Produkt

$$[s_1, s_2, \ldots, s_i, s_{i+1}, \ldots s_n] \cdot \begin{bmatrix} e_1^* \\ e_2^* \\ \vdots \\ \frac{1}{2}(e_i^* + e_{i+1}^*) \\ \frac{1}{2j}(e_i^* - e_{i+1}^*) \\ \vdots \\ e_i^* \end{bmatrix}, \tag{17}$$

anderseits das Produkt

$$\begin{bmatrix} e_1^* \\ e_2^* \\ \vdots \\ \frac{1}{2}(e_i^* + e_{i+1}^*) \\ \frac{1}{2j}(e_i^* - e_{i+1}^*) \\ \vdots \\ e_n^* \end{bmatrix}^{-1} \cdot \mathbf{J} = \mathbf{H}_{i,i+1}^{-1} \cdot \mathbf{J}. \tag{18}$$

Es ist so gleich zu sehen, daß (17) die Matrix

$$[s_1, s_2, \ldots, s_{i-1}, \hat{s}_i, \hat{s}_{i+1}, \ldots, s_n]$$

angibt, und daß $\mathbf{H}_{i,i+1}^{-1} \cdot \mathbf{J}$ überall mit $\mathbf{J}$ übereinstimmt, abgesehen vom zweidimensionalen Block in den $i$- und $i+1$-ten Spalten und Zeilen, der von

$$\begin{bmatrix} \lambda & 0 \\ 0 & \bar{\lambda} \end{bmatrix} \tag{19}$$

in

$$\begin{bmatrix} \lambda & \bar{\lambda} \\ j\lambda & -j\bar{\lambda} \end{bmatrix}$$

übergeht. Es folgt dann, daß in der Darstellung

$$\mathbf{A} = \left\{[s_1, s_2, \ldots, s_n] \cdot \mathbf{H}_{i,i+1}\right\} \cdot \left\{\mathbf{H}_{i,i+1}^{-1} \mathbf{J} \mathbf{H}_{i,i+1}\right\} \cdot \left\{\mathbf{H}_{i,i-1}^{-1} \cdot \begin{bmatrix} z_1^* \\ z_2^* \\ \vdots \\ z_n^* \end{bmatrix}\right\} \tag{20}$$

der erste Faktor in der $i$-ten und in der $i+1$-ten Spalte nur reelle Zahlen und in der Hauptdiagonalen von 0 verschiedene Elemente besitzt, während der zweite Faktor mit $\mathbf{J}$ übereinstimmt, abgesehen vom in (19) genannten Block, der in

$$\begin{bmatrix} \dfrac{\lambda+\bar{\lambda}}{2} & \dfrac{\lambda-\bar{\lambda}}{2j} \\ -\dfrac{\lambda-\bar{\lambda}}{2j} & \dfrac{\lambda+\bar{\lambda}}{2} \end{bmatrix} \tag{21}$$

übergeht. Wird also für alle konjugierten Paare die angegebene Transformation benützt, erhält man schließlich die Darstellung

$$[\hat{s}_1, \hat{s}_2, \ldots, \hat{s}_n] \cdot \hat{\mathbf{J}} \cdot \begin{bmatrix} \hat{z}_1^* \\ \hat{z}_2^* \\ \vdots \\ \hat{z}_n^* \end{bmatrix}, \tag{22}$$

wo bereits alle Elemente reell sind, und in $[\hat{s}_1, \hat{s}_2, \ldots, \hat{s}_n]$ die Hauptdiagonale nur von 0 verschiedene Zahlen enthält. Um nun in der $i$-ten Zeile alle Elemente — bzw. vom $(k+1)$-ten zum $n$-ten Element (falls $i > k$ ist; $k$ ist die Dimensionszahl von $\mathbf{A}_{11}$) — zu annullieren (vom $i$-ten Element abgesehen), ist von links mit der Matrix

$$\mathbf{H}_i^{(1)} = \begin{bmatrix} & & \\ & & \\ -\dfrac{s_{1i}}{s_{ii}} e_i^* + e_1^* & & \\ \vdots & & \\ -\dfrac{s_{i-1,i}}{s_{ii}} e_i^* + e_{i-1}^* & & \\ e_i^* & & \\ \vdots & & \\ -\dfrac{s_{k,i}}{s_{ii}} e_i^* + e_k^* & & \\ \vdots & & \\ -\dfrac{s_{n,i}}{s_{ii}} e_i^* + e_n^* & & \end{bmatrix} \quad \text{bzw.} \quad \mathbf{H}_i^{(2)} = \begin{bmatrix} & & \\ e_1^* & & \\ \vdots & & \\ e_k^* & & \\ -\dfrac{s_{k-1,i}}{s_{ii}} e_i^* + e_{k+1}^* & & \\ \vdots & & \\ e_i^* & & \\ \vdots & & \\ -\dfrac{s_{n,i}}{s_{ii}} e_i^* + e_n^* & & \end{bmatrix}$$

zu multiplizieren; in der Darstellung

$$\mathbf{A} = \left\{[\hat{s}_1, \ldots, \hat{s}_n] \cdot \mathbf{H}_i\right\} \cdot \left\{\mathbf{H}_i^{-1} \cdot \hat{\mathbf{J}} \cdot \mathbf{H}_i\right\} \cdot \left\{\mathbf{H}_i^{-1} \cdot \begin{bmatrix} \hat{x}_1^* \\ \hat{x}_2^* \\ \vdots \\ \hat{x}_n^* \end{bmatrix}\right\}$$

hat also die $i$-te Zeile von $[\hat{s}_1, \ldots, \hat{s}_n] \cdot \mathbf{H}_i$ die Form, die der Form von $\mathbf{T}$ entspricht. Multipliziert man mit allen $\mathbf{H}_i$, so ergibt sich die gewünschte Darstellung für $\mathbf{T}$. Da aber der dritte Faktor die Reziproke des ersten ist, und $\mathbf{T}^{-1}$ notwendigerweise die angegebene Form hat, entspricht auch der dritte Faktor der angegebenen Form. Endlich gewährleistet die Struktur von $\hat{\mathbf{J}}$, daß der zweite Faktor

$$\left\{\left(\prod_{i=1}^{n} \mathbf{H}_i^{-1}\right) \cdot \hat{\mathbf{J}} \cdot \left(\prod_{i=1}^{n} \mathbf{H}_i\right)\right\}$$

eine obere Blockdreiecksmatrix sei; $\mathbf{H}_i^{-1}$ bedeutet ja (von links) eine Reihenoperation, bildet aber nur in der $i$-ten Reihe neue Elemente, und ist $i > k$, so bildet sie neue Elemente nur mit Hilfe solcher Reihen, die einen größeren Index als $k$ haben; daraus folgt, daß der zweite Faktor eine obere Blocksdreiecksmatrix ist, und damit ist unser Satz vollkommen bewiesen.

Um nun Satz 2 in exakter Weise zu formulieren, betrachten wir erst den Beweis von Satz 1. Wie wir gesehen haben, wird durch eine Jordansche Normalform dann, und nur dann, die Darstellung von $\mathbf{A}$ in der gewünschten Form ermöglicht, wenn die rechtsseitigen Hauptvektoren in $[x_1, x_2, \ldots, x_n]$ eine Reihenfolge haben, im welcher einerseits die Konjugierten nebeneinander stehen, anderseits wo in der Hauptdiagonalen keine $O$-Elemente vorkommen. Alle möglichen Reihenfolgen, die die obigen Anforderungen erfüllen, werden in der folgenden, eine Partitionierung-zulassenden Reihenfolge genannt.

**Satz 2.** *In der Blockdreiecksiteration* (2) *konvergieren — u. zw. linear — die Matrizen* $\mathbf{T}_n$ *und* $\mathbf{S}_n$ *gegen die Gleichung* (1) *genügende Grenzmatrizen* $\mathbf{T}$ *und* $\mathbf{S}$, *falls es für* $\mathbf{A}$ *eine solche, die Partitionierung-zulassende Reihenfolge der Hauptvektoren gibt, in welcher der größte Modulus der ersten* $k$*-Eigenwerte ungleich dem größten Modulus der letzten* $(n - k)$*-Eigenwerte ist.*

*Beweis:* Betrachtet man die Iteration (2), so erhält man nacheinander die Relationen:

$$\begin{aligned}
&\mathbf{A}\mathbf{T}_0 = \mathbf{T}_1 \mathbf{S}_1 ; \\
&\mathbf{A}^2 \mathbf{T}_0 = \mathbf{A}(\mathbf{A}\mathbf{T}_0) = \mathbf{A}(\mathbf{T}_1 \mathbf{S}_1) = \mathbf{A}\mathbf{T}_1 \mathbf{S}_1 = \mathbf{T}_2 \mathbf{S}_2 \mathbf{S}_1 ; \\
&\mathbf{A}^3 \mathbf{T}_0 = \mathbf{A}(\mathbf{A}^2 \mathbf{T}_0) = \mathbf{A}(\mathbf{T}_2 \mathbf{S}_2 \mathbf{S}_1) = \mathbf{T}_3 \mathbf{S}_3 \mathbf{S}_2 \mathbf{S}_1 ; \ldots . \\
&\mathbf{A}^n \mathbf{T}_0 = \mathbf{A}(\mathbf{A}^{n-1} \mathbf{T}_0) = \mathbf{A}(\mathbf{T}_{n-1} \mathbf{S}_{n-1} \mathbf{S}_{n-2} \ldots \mathbf{S}_1) = \\
&\qquad = \mathbf{T}_n \mathbf{S}_n \mathbf{S}_{n-1} \ldots \mathbf{S}_1 .
\end{aligned} \tag{23}$$

Da nun das Produkt aus diesen oberen Blockdreiecksmatrizen selbst eine obere Dreiecksmatrix ist, folgt aus (23), daß

$$\mathbf{A}^n = \mathbf{T}_n \mathbf{S}^{(n)} \mathbf{T}_0^{-1} \tag{24}$$

gilt. Die Partitionierung von $\mathbf{A}^n$ entsprechend (24) verläuft nun ebenso, wie es beim Beweis des Satzes 1 gezeigt wurde (hier ist im allgemeinen, $q_0^{-1} \neq q_n^{-1}$), jedoch ist $\mathbf{T}_0$ vorgeschrieben und $\mathbf{T}_0^{-1}$ hat eine gleiche Struktur wie $\mathbf{T}_0$; deshalb muß der Grundgedanke des gezeigten Algorithmus so modifiziert werden, daß rechts und links verschiedene Spaltenoperationen angewendet werden.

Betrachten wir also eine Jordansche Normalform von $\mathbf{A}_n$, wo einerseits, die Reihenfolge der Hauptvektoren zulässig, anderseits, der größte Modulus der Eigenwerte in der ersten $k$-Zeile ungleich dem der letzten $(n - k)$-Zeile ist. Wird die $n$-te Potenz des größten Modulus der Eigenwerte von $\mathbf{A}$ in $\mathbf{J}^n$ ausgeklammert, bleiben an einigen Stellen der Hauptdiagonalen von $\check{\mathbf{J}}^n$ komplexe Zahlen mit dem Modulus 1, an anderen Stellen aber Elemente, die eine Größenordnung $O(q^n)$ haben (mit $0 \leq q < 1$); u. zw. bleiben Elemente mit dem Modulus 1 nur entweder in der ersten $k$- oder aber in den letzten $n - k$-Zeilen. Wendet man nun die in Zusammenhang mit dem Beweis des Satzes 1 erklärten Transformationen an, um $[x_1, x_2, \ldots, x_n]$ in die Form $\mathbf{T}_n$, $\begin{vmatrix} z_1^* \\ \vdots \\ z_n^* \end{vmatrix}$ aber in die Form $\mathbf{T}_0^{-1}$ zu bringen, so transformiert sich $\check{\mathbf{J}}^n$ in eine Form — u. zw. in die $\mathbf{S}^n$ darstellende Form —, wo gewisse Elemente die Größenordnung $1 + O(q^n)$, andere aber die Größenordnung $O(q^n)$ haben. Von dem ausgeklammerten Faktor abgesehen, konvergiert $\mathbf{S}^{(n)}$ zu einer Grenzmatrix, und deswegen gilt dasselbe für $\mathbf{T}_n$. Wenn aber $\mathbf{T}_n$ konvergiert, so folgt aus (2), daß auch $\mathbf{S}_n$ konvergent ist, w. z. B. w.

Es sei hier bemerkt, daß falls $\mathbf{A}$ überhaupt Eigenwerte mit verschiedenen Modulen hat, aber bei der gewählten Partitionierung keine Reihenfolge existiert, die die Forderung des Satzes 2 erfüllt, so kann entweder die Partitionierung von $\mathbf{A}$ gewechselt oder eine Ähnlichkeitstransformation mit Hilfe von Permutationsmatrizenpaaren angewendet werden, um eine Vertauschung der Komponentenfolge der Hauptvektoren zu erzwingen, wodurch auch eine Reihenfolgenänderung der zulässigen Reihenfolgen dargestellt ist bzw. auch neue zulässige Reihenfolgen stattfinden.

**Satz 3.** *Der Newtonsche Iterationsprozeß* (3)—(4)—(10)—(11)—(12)—(13)—(14) *konvergiert quadratisch für eine beliebige Matrix* $\mathbf{A}$ *und für beliebige Partitionen mit entsprechenden Dimensionszahlen, falls* $\mathbf{T}_n$ *eine genügend gute Approximation von* $\mathbf{T}$ *ist. Nehmen wir die Korrektionen in* (8) *in Betracht, die sich so ergeben, erhält man einen Iterationsprozeß mit kubischer Konvergenz.*

Beweis: Im Satz 1 wurde bewiesen, daß bei geeigneten Dimensionszahlen die gewünschte Darstellung (1) von $\mathbf{A}$ immer existiert. Die Korrektion $\mathbf{T}^{(n)}$

bzw. die dadurch definierten Korrektionen $\mathbf{S}_n$ und $\mathbf{T}_n$ werden in (5)—(6)—(7) und in (9) ganz genau in Betracht gezogen und hier erhält man die gesuchte Lösung durch einfache Umordnung der Gleichungen. Nur in der Gleichung (8) haben wir ein quadratisches Glied, das in erster Approximation außer acht gelassen wird. Für die Unbekannte $\delta$ ergibt sich hier eine lineare Gleichung der Form

$$\mathbf{B}\delta + \delta\mathbf{C} = \mathbf{D}\,, \tag{25}$$

wo $\mathbf{B}$, $\mathbf{C}$ und $\mathbf{D}$ bekannte Matrizen sind. Hätten wir in (8) auch das quadratische Glied in Betracht gezogen, u. zw. mit seinem genauen (und hinreichend kleinen) Wert, so würde statt (25) die Gleichung

$$\mathbf{B}\delta + \delta\mathbf{C} = \mathbf{D} - \delta^{(n)}\,\mathbf{S}_{11}^{(n)} \tag{25'}$$

die genaue Lösung $\delta$ angegeben. $\mathbf{A}$ hat nun mehrere Darstellungen der Form (1), die sich in der Anordnung der rechtsseitigen Hauptvektoren unterscheiden (s. den Nachweis des Satzes 1); eben deshalb haben die möglichen $\mathbf{T}$- und $\mathbf{S}$-Matrizen, die die Darstellung (1) geben — einen Abstand (in einer gewissen Matrizennorm gemessen), der von 0 streng abgegrenzt ist. Ist also $\mathbf{T}^{(n)}$ eine Approximation eines möglichen $\mathbf{T}$ mit einem viel kleineren Abstand von diesem $\mathbf{T}$ als die untere Grenze der Abstände der möglichen $\mathbf{T}$- bzw. $\mathbf{S}$-Matrizen (u. zw. mit einem so geringen Abstand, daß durch (3) ein solches Paar $\mathbf{T}_n$, $\mathbf{S}_n$ definiert wird, dessen Glieder auch viel kleinere Abstände von $\mathbf{T}$ bzw. von $\mathbf{S}$ als diese untere Grenze aufweisen) so definiert (25′) eindeutig die Lösung $\delta$. Das kann aber nur dann der Fall sein, wenn die Gleichung (25) nicht entartet ist, also auch eine eindeutige Lösung besitzt. (Über die Gleichung (25) s. Teil 4.) Daraus folgt aber, daß die Lösung von (25) nur quadratisch von der genauen Lösung von (25′) abweichen kann, w. z. B. w. Nehmen wir noch die quadratische Approximation von $\delta^{(n)}$ und $\mathbf{S}_{11}^{(n)}$ in (25′) in Betracht, so ergibt sich eine kubische Approximation von $\delta$, wie es behauptet wurde.

4. Wir müssen uns noch mit der Lösung der linearen Gleichung (25) beschäftigen. Besitzen die quadratischen Matrizen $\mathbf{B}$, $\mathbf{C}$ und $\mathbf{D}$ die Dimensionszahl $n$, so ergibt sich das folgende Gleichungssystem, wenn wir die Spalten der Matrizen $\delta$ bzw. $\mathbf{D}$ in die (Spaltenvektor) Elemente der Hypervektoren $x$ bzw. $g$ umschreiben:

$$\mathbf{F} \cdot x = g\,, \tag{26}$$

wo

$$x = \begin{bmatrix} \delta_1 \\ \delta_2 \\ \vdots \\ \delta_n \end{bmatrix};\quad g = \begin{bmatrix} d_1 \\ d_2 \\ \vdots \\ d_n \end{bmatrix};$$

$$\mathbf{F} = \begin{bmatrix} \mathbf{B}+c_{11}\mathbf{E} & c_{21}\mathbf{E} & c_{31}\mathbf{E} & \dots & c_{n1}\mathbf{E} \\ c_{12}\mathbf{E} & \mathbf{B}+c_{22}\mathbf{E} & c_{32}\mathbf{E} & \dots & c_{n2}\mathbf{E} \\ \vdots & \vdots & \vdots & & \vdots \\ c_{1n}\mathbf{E} & c_{2n}\mathbf{E} & c_{3n}\mathbf{E} & \dots & \mathbf{B}+c_{nn}\mathbf{E} \end{bmatrix}. \tag{27}$$

Die Gleichung (26) hat nun — wie es aus (27) sichtbar ist — einen speziellen Charakter, der jedoch die direkte Lösung nicht erleichtert: bei einer Gauß- oder Gauß—Jordan-Elimination sind $O(n^6)$ Operationen durchzuführen. Es scheint also viel günstiger, die Lösung von (25) direkt, durch ein Iterationsverfahren zu suchen, da die Ersetzung einer Approximation von $\delta$ in (25) nur $O(n^3)$ Operationen erfordert.

Jedoch scheint es nicht leicht, die Gleichung (26) in eine Form

$$\delta = \mathrm{f}(\delta, \mathrm{B}, \mathrm{C}, \mathrm{D}) \tag{28}$$

so umzuschreiben, daß das Iterationsverfahren

$$\delta_{n-1} = \mathrm{f}(\delta_n, \mathrm{B}, \mathrm{C}, \mathrm{D}) \tag{29}$$

auch bei günstig gewähltem Anfangswert $\delta_0$ dem Fixpunkt von $f$ zustrebe, da die verallgemeinerte Fréchet-Derivierte von f im allgemeinen eine Norm über 1 hat. Die verallgemeinerte Regula-falsi-Methode ist natürlich anwendbar, jedoch ist für diese Methode eine gute Anfangsapproximation notwendig. Um eine solche Anfangsapproximation zu erhalten, wird das Schrödersche Verfahren verallgemeinert, das für lineare Gleichungen in der Regel bessere Konvergenz aufweist als das Iterationsverfahren (s. z. B. [4—6]). Betrachten wir also den Halbordnungs-Banachraum $R$, wo die Halbordnung eine Netzstruktur bildet, ferner den Operator $T : R \to R$, der in Form

$$T = T_1 + T_2 + T_3 + T_4$$

zerlegt werden kann, mit isotonen $T_1$ und $T_3$ bzw. mit antitonen $T_2$ und $T_4$. Es seien ferner vorausgesetzt:

a) Es gibt einen konvexen Bereich $D \subseteq R$, wo $S_1 = T_1 + T_2$ und $S_2 = T_3 + T_4$ jedes Intervall in einer kompakten Menge durchführt;

b) $E - S_1$ ist von monoton wachsender Art;

c) es gibt ein Intervall $\langle w_0, z_0 \rangle \subseteq D$, und zu jedem inneren Intervall $\langle w_i, z_i \rangle \subseteq \langle w_0, z_0 \rangle$ ein Intervallpaar $\langle u_{10}(i), v_{10}(i) \rangle \subseteq D$ bzw. $\langle u_{20}(i), v_{20}(i) \rangle \subseteq D$, mit

$$u_{11}(i) = T_1 u_{10}(i) + T_2 v_{10}(i) + T_3 w_i + T_4 z_i \geq u_{10}(i) ; \tag{30}$$

$$v_{11}(i) = T_1 v_{10}(i) + T_2 u_{10}(i) + T_3 w_i + T_4 z_i \leq v_{10}(i) ; \tag{31}$$

$$u_{21}(i) = T_1\, u_{20}(i) + T_2\, v_{20}(i) + T_3\, z_i + T_4\, w_i \geq u_{20}(i)\ ; \tag{32}$$

$$v_{21}(i) = T_1\, v_{20}(i) + T_2\, u_{20}(i) + T_3\, z_i + T_4\, w_i \leq v_{20}(i)\ . \tag{33}$$

d) Es gibt ein Indexpaar $n_1$ bzw. $n_2$ so, daß für die Folgen

$$u_{1,n+1} = T_1\, u_{1n} + T_2\, v_{1n} + T_3\, w_0 + T_4\, z_0\ ; \tag{34}$$

$$v_{1,n+1} = T_1\, v_{1n} + T_2\, u_{1n} + T_3\, w_0 + T_4\, z_0\ ; \tag{35}$$

$$u_{2,n+1} = T_1\, u_{2n} + T_2\, v_{2n} + T_3\, z_0 + T_4\, w_0\ ; \tag{36}$$

$$v_{2,n+1} = T_1\, v_{2n} + T_2\, u_{2n} + T_3\, z_0 + T_4\, w_0\ , \tag{37}$$

$$u_{1,n_1} \text{ und } v_{1,n_1} \in \langle w_0, z_0\rangle;\ u_{2,n_2} \text{ und } v_{2,n_2} \in \langle w_0, z_0\rangle \tag{38}$$

feststehen.

Satz 4. *Neben den obigen Voraussetzungen besitzt der Operator $T$ mindestens einen Fixpunkt in $<w_0, z_0>$. Ist $E - S_1$ eindeutig umkehrbar und von streng monotoner Art, und sind $T_3$ und $T_4$ paarweise streng monoton* (d. h. daß für $w < z$: $T_3\, w + T_4\, z < T_3\, z + T_4\, w$ gelte), *und kann für kein $w < z$, $w \in D$ $z \in D$ das Gleichungspaar*

$$w = (E - S_1)^{-1}\,(T_3\, w + T_4\, z)\ ;\ z = (E - S_1)^{-1}\,(T_3\, z + T_4\, w)$$

*feststehen, so besitzt $T$ einen und nur einen Fixpunkt in $<w_0, z_0>$, den man mit Hilfe eines Verfahrens des Typs* (34)—(37) *approximieren kann.*

*Beweis:* Es ist bekannt (s. z. B. [7]), daß unter den Voraussetzungen (30), (31) die Punkte, definiert durch (34)—(35), ineinander eingeschachtelte Intervallen bilden, d. h.

$$\langle u_{1,n+1}, v_{1,n+1}\rangle \subseteq \langle u_{1,n}, v_{1n}\rangle\,.$$

Laut des Schröderschen Fixpunktsatzes besitzt dann der Operator $S_1 + T_3\, w_0 + T_4\, z_0$ wegen a) mindestens einen Fixpunkt in $<u_{1n}, v_{1n}>$, und da (38) auch erfüllt ist, auch mindestens einen Fixpunkt in $<w_0, z_0>$. Ebenso kann gezeigt werden, daß auch der Operator $S_1 + T_3\, z_0 + T_4\, w_0$ mindestens einen Fixpunkt in $<w_0, z_0>$ besitzt. Betrachten wir einen Fixpunkt $w_1 \in <w_0, z_0>$ des ersten, und einen Fixpunkt $z_1 \in <w_0, z_0>$ des zweiten Operators.

Da weiterhin $E - S_1$ von monoton wachsender Art ist, und $T_3\, w_0 + T_4\, z_0 \leq T_3\, z_0 + T_4\, w_0$ gilt, steht auch $w_0 \leq w_1 \leq z_1 \leq z_0$ fest.

Laut (30)—(33) kann also der obige Gedankengang auch für die Operatoren $S_1 + T_3\, w_1 + T_4\, z_1$ bzw. $S_1 + T_3\, z_1 + T_4\, w_1$ verfolgt werden. Es bleibt nur die Frage offen, ob man auch jetzt ein Indexpaar $n_1(1)$, $n_2(1)$ mit der Eigenschaft

$$\langle u_{1,n_1(1)}(1), v_{1,n_1(1)}(1)\rangle \in \langle w_1, z_1\rangle\ ;\ \langle u_{2,n_2(1)}(1), v_{2,n_2(1)}(1)\rangle \subseteq \langle w_1, z_1\rangle$$

finden kann. Da aber beide Operatoren gewiß einen Fixpunkt in $<u_{1,n}(1), v_{1,n}(1)>$ bzw. in $<u_{2,n}(1), v_{2,n}(1)>$ haben, und $T_3 w_1 + T_4 z_1 \geq T_3 w_0 + T_4 z_0$ bzw. $T_3 z_1 + T_4 w_1 \leq T_3 z_0 + T_4 w_0$ auch feststeht, ferner der Operator $E - S_1$ von monotoner Art ist, so folgt daß sowohl $S_1 + T_3 w_1 + T_4 z_1$ einen Fixpunkt $w_2 \geq w_1$ als auch $S_1 + T_3 z_1 + T_4 w_1$ einen Fixpunkt $z_2 \leq z_1$ haben, ferner, daß auch $w_2 \leq z_2$ feststeht. Die Konstruktion kann also unbedingt fortgesetzt, und mit vollständiger Induktion bewiesen werden, daß sowohl der Operator $S_1 + T_3 w_n + T_4 z_n$ einen Fixpunkt $z_{n-1} \leq z_n$ besitzen, und $w_{n+1} \leq z_{n+1}$ auch feststeht. Es soll nun gezeigt werden, daß

$$T(\langle w_n, z_n \rangle) \subseteq \langle w_n, z_n \rangle \tag{38}$$

gültig ist. Es sei $x \in \langle w_n, z_n \rangle$. Da $T_3$ isoton, und $T_4$ antiton ist, folgt, daß $S_1 x + T_3 w_n + T_4 z_n \leq S_1 x + T_3 x + T_4 x = Tx \leq S_1 x + T_3 z_n + T_4 w_n$. Da ferner $S_1 + T_3 w_n + T_4 z_n$ einen Fixpunkt $w_{n+1} \in \langle w_n, z_n \rangle$ und $S_1 + T_3 z_n + T_4 w_n$ einen Fixpunkt $z_{n+1} \in \langle w_n, z_n \rangle$ besitzen, folgt, daß $w_{n+1} \leq T_1 w_n + T_2 z_n + T_3 w_n + T_4 z_n \leq S_1 x + T_3 w_n + T_4 z_n \leq Tx \leq S_1 x + T_3 z_n + T_4 w_n \leq T_1 z_n + T_2 w_n + T_3 z_n + T_4 w_n \leq z_{n+1}$, folglich ist

$$T(\langle w_n, z_n \rangle) \subseteq \langle w_{n+1}, z_{n+1} \rangle, \tag{39}$$

also steht auch (38) umso mehr fest. Damit haben wir den ersten Teil unseres Satzes bewiesen.

Der zweite Teil des Satzes folgt fast unmittelbar; da $(E - S_1)$ eindeutig umkehrbar und von streng monotoner Art ist, kann eine Relation der Form

$$\lim_{k \to \infty} u_{1k}^{(n)}(i) = u_1^{(n)}(i) < \lim_{k \to \infty} v_{1k}^{(n)}(i) = v_1^{(n)}(i) \tag{40}$$

bzw.

$$\lim_{k \to \infty} u_{2k}^{(n)}(i) = u_2^{(n)}(i) < \lim_{k \to \infty} v_{2k}^{(n)}(i) = v_2^{(n)}(i) \tag{41}$$

nicht richtig sein, da

$$u_1^{(n)}(i) = (E - S_1)^{-1} (T_3 z_n + T_4 z_n) = v_1^{(n)}(i) \tag{42}$$

bzw.

$$u_2^{(n)}(i) = (E - S_1)^{-1} (T_3 z_n + T_4 w_n) = v_2^{(n)}(i) \tag{43}$$

gelten. Die Punkte $w_n$, $z_n$ sind also eindeutig definiert. Es kann aber auch

$$\lim_{n \to \infty} w_n = w < \lim_{n \to \infty} z_n = z \tag{44}$$

nicht gelten, da dann

$$w = (E - S_1)^{-1} (T_3 w + T_4 z)$$

und

$$z = (E - S_1)^{-1}(T_3\, z + T_4\, w)$$

auch gültig wären in Gegensatz zu unseren Voraussetzungen. Es folgt also, daß in diesem Falle der eindeutig definierte Fixpunkt von $T$ folgendermaßen gefunden werden kann:

Die Folge

$$\begin{aligned} u^{(k)}_{1,n+1} &= T_1\, u^{(k)}_{1n} + T_2\, v^{(k)}_{1n} + T_3\, w_k + T_4\, z_k \\ v^{(k)}_{1,n+1} &= T_1\, v^{(k)}_{1n} + T_2\, u^{(k)}_{1n} + T_3\, w_k + T_4\, z_k \end{aligned} \tag{45}$$

bzw. die Folge

$$\begin{aligned} u^{(k)}_{2,n+1} &= T_1\, u^{(k)}_{2n} + T_2\, v^{(k)}_{2n} + T_3\, z_k + T_4\, w_k \\ v^{(k)}_{2,n+1} &= T_1\, v^{(k)}_{2n} + T_2\, u^{(k)}_{2n} + T_3\, z_k + T_4\, w_k \end{aligned} \tag{46}$$

sollen bis zu einem Index $n_1$ bzw. $n_2$ fortgesetzt werden, für die bereits

$$u^{(k)}_{1,n_1} > w_k \quad \text{bzw.} \quad v^{(k)}_{2,n_2} < z_k \tag{47}$$

feststehen: wir können dann

$$w_{k+1} = u^{(k)}_{1,n_1} \cap v^{(k)}_{2,n_2}; \quad z_{k+} = u^{(k)}_{1,n_1} \cup v^{(k)}_{2,n_2} \tag{48}$$

wählen, wo $\cap$ bzw. $\cup$ die Bildung der unteren bzw. der oberen Grenze bedeuten.

Nun stehen für die betrachtete lineare Gleichung $\mathfrak{B}\delta + \delta\mathfrak{C} - \mathfrak{D} = 0$ alle Voraussetzungen des Satzes 4 fest, falls wir sie in eine Form

$$\delta = \alpha(\delta + \varepsilon(\mathfrak{B}\delta + \delta\mathfrak{C} - \mathfrak{D})) + (1 - \alpha)\,(\delta + \varepsilon(\mathfrak{B}\delta + \delta\mathfrak{C} - \mathfrak{D}))$$

umschreiben, mit $\alpha \sim 3/4$ und genügend kleinem $|\varepsilon|$.

(Es sei erwähnt, daß für $\alpha \sim 1/2$ auch Elemente $w$ und $z < w$ angegeben werden können, die die Gleichungen

$$w = (E - S_1)^{-1}(T_3\, w + T_4\, z); \quad z = (E - S_1)^{-1}(T_3\, z + T_4\, w)$$

gleichzeitig befriedigen.)

Es sei noch bemerkt, daß das Verfahren (54)—(57) nicht leicht durchzuführen ist und nicht schnell konvergiert; eben deswegen wenden wir es nur für eine grobe Eingrenzung der Lösung von $\mathfrak{B}\delta + \delta\mathfrak{C} - \mathfrak{D} = 0$ an, und nur nachfolgend bedienen wir uns der verallgemeinerten Regula-falsi-Methode. Es soll aber auch über diese Methode eine Bemerkung hinzugefügt werden. Die Regula-falsi-Methode wird in der Regel in Banachräumen angewandt, die gleichzeitig einen kommutativen Ring bilden und in denen eine Invertier-

menge existiert, jedoch ist der Raum der $n$-dimensionalen quadratischen Matrizen kein kommutativer Ring, und deshalb wendet man die Methode nicht in der im Buche von Collatz angegebenen Form an, sondern in der Form

$$u_{n-1} = u_n - (u_{n-1} - u_n) \cdot (Tu_{n-1} - Tu_n)^{-1} Tu_n . \tag{49}$$

(Es sei bemerkt, daß hier der »Differenzenquotient« $(Tu_{n-1} - Tu_n) \cdot (u_{n-1} - u_n)^{-1}$ keine Approximation der Ableitung ist, da letztere ein »Array« mit vier Indexen ist; die Formel ist jedoch für die linearen Gleichungen $AX = B$ bzw. $X \cdot A = B$ noch genau; für den allgemeineren Fall $BX + XC = D$ ist das aber schon nicht der Fall.)

5. Wie schon erwähnt wurde, kann die hier angegebene Methode auch so formuliert werden, daß sie Formen der $L$—$R$-Methode näher steht. Man kann nämlich die Matrizen A, T und S »stärker« partitionieren, z. B. in eine Form

$$\begin{pmatrix} A_{11} & A_{12} & A_{13} \\ A_{21} & A_{22} & A_{23} \\ A_{31} & A_{32} & A_{33} \end{pmatrix} \cdot \begin{pmatrix} E & 0 & 0 \\ T_{21} & E & 0 \\ T_{31} & T_{32} & E \end{pmatrix} = \begin{pmatrix} E & 0 & 0 \\ T_{21} & E & 0 \\ T_{31} & T_{32} & TE \end{pmatrix} \cdot \begin{pmatrix} S_{11} & S_{12} & S_{13} \\ 0 & S_{22} & S_{23} \\ 0 & 0 & S_{33} \end{pmatrix} .$$

und um so mehr auch in höchstens zweidimensionale Blöcke; jedoch werden dann die Gleichungen viel komplizierter.

## Zusammenfassung

Eine Verallgemeinerung der L—R-Methode wird angegeben, die einerseits eine quadratische bzw. kubische Konvergenz sichert, anderseits auch dann konvergiert, wenn die Eigenwerte komplex oder mehrfach sind, oder einige den gleichen Modulus haben. Verfasser beschäftigt sich auch mit der iterativen Lösung der Matrizengleichung $BX + XC = D$, die ein Hilfsmittel der angegebenen Methode ist.

## Literatur

1. Bauer, F. L.: Beiträge zum Danilewski-Verfahren. Ber. Internat. Math. Koll. Nov. 1955. (1957)
2. Rutishauer, H.: Une méthode pour la détermination des valeurs propres d'une matrice. C.R. Ac. Sci. Paris 240 (1955).
3. Francis, J. G. F.: The QR Transformation. I—II. Computing J. 4. (1961)—4 (1962).
4. Schröder, J.: Anwendung von Fixpunktsätzen bei der numerischen Behandlung nichtlinearer Gleichungen in halbgeordneten Räumen. Arch. Rat. Mech. Anal. 4. (1960).
5. Albrecht, J.: Fehlerschranken und Konvergenzbeschleunigung bei einer monotonen oder alternierenden Iterationsfolge. Num. Math. (1962).
6. Albrecht, J.: Zur Fehlerabschätzung beim Gesamt- und Einzelschrittverfahren für lineare Gleichungssysteme. Z. angew. Math. Mech. 43. (1963).
7. Collatz, L.: Funktionalanalysis und numerische Mathematik. Springer, Berlin—Göttingen—Heidelberg, 1964.

Prof. Dr. Tamás Frey, Budapest XI., Egry József u. 18—20, Ungarn