Ein Inverses Eigenwertproblem

K. P. HADELER
Institut für Angewandte Mathematik
der Universität Hamburg, Germany

Communicated by Lothar Collatz

SUMMARY

The paper is concerned with the problem of finding a real, diagonal matrix M such that $A + M$ has prescribed eigenvalues, where A is any given symmetric matrix. This problem represents a discrete analog of the inverse eigenvalue problem in which we seek to determine a "potential" $g(x)$ such that the operator in Hilbert space, $l(y) = -y'' + g(x)y$ with appropriate boundary conditions, possesses a prescribed spectrum.

Section 2 defines the "inverse problem" and its generalization.

In Section 3, upper and lower bounds are given for the eigenvalues of a symmetric matrix, obtained by the iterative application of Temple's theorem. Also, an estimate is given for the changes in the eigenvalues produced by perturbing a given symmetric matrix by a diagonal matrix.

Section 4 contains a complete discussion of the case of 2×2 matrices.

In Section 5, from simple considerations concerning rotations in an $n(n+1)/2$-dimensional space, necessary conditions are given for the inverse problem to have a solution.

In Section 6, the Brouwer fixed point theorem is applied to an operator equation which is equivalent to the "inverse problem." A sufficient condition is given for the problem to have a solution: let k denote a certain seminorm on the set of all symmetric matrices and let d be the minimal mutual distance of the given eigenvalues. Then the inverse problem has a solution provided $k(A)/d \leqslant 0.288$.

Under only slightly stronger conditions ($k(A)/d \leqslant 0.25$), it is shown in Section 7, using the Banach contraction theorem, that the iteration converges and the solution is locally unique.

Section 8 contains a numerical example.

1. EINLEITUNG

Seit mehreren Jahrzehnten werden neben den direkten Spektral- (Eigenwert-) Aufgaben sogenannte inverse Aufgaben betrachtet. Während

Linear Algebra and Its Applications 1, 83–101 (1968)

bei ersteren das Spektrum eines linearen Operators in einem Hilbertraum (oder in einem allgemeineren linearen Raum) beschrieben oder bestimmt werden soll, ist bei den inversen Aufgaben ein Element aus einer Klasse von Operatoren zu bestimmen, das ein vorgegebenes Spektrum besitzt und eventuell noch weitere Eigenschaften aufweist.

Die ersten Arbeiten über inverse Aufgaben waren dem Sturm-Liouville-Problem gewidmet. Auf einem (endlichen oder unendlichen) Intervall definiert eine Klasse von Differentialausdrücken

$$ly = -y'' + q(x)y \tag{1}$$

mit Randbedingungen eine Klasse von linearen Operatoren in einem Hilbertraum. Das Potential q (und eventuell freie Parameter in den Randbedingungen) ist so zu bestimmen, daß der zugehörige Operator ein vorgeschriebenes Spektrum besitzt. Die wichtigsten Tatsachen über dieses Problem finden sich im Anhang des Buches von M. A. Neumark [5]. Es zeigt sich, daß man die Eindeutigkeit der Lösung nicht durch Vorgabe des Spektrums allein erzwingen kann. Auch die Spektralverteilungsfunktion muß gegeben sein. Neben analogen Problemen für den Dirac-Operator wurden neuerdings auch Aufgaben der Art

$$ly = -(py')' \tag{2}$$

untersucht, wo eine Funktion p durch das Spektrum bestimmt werden soll.

Es scheint, daß man sich im Gegensatz zur direkten Aufgabe beim inversen Problem bisher nur wenig mit endlichdimensionalen Analoga sowie mit den (1), (2) entsprechenden Differenzengleichungen beschäftigt hat.

Ohne Vollständigkeit anzustreben, erwähnen wir nur ein zu (2) analoges Problem, das Anwendungen in der theoretischen Chemie besitzt. Zu einer Matrix A ist eine Diagonalmatrix D gesucht, so daß

$$A \cdot D \tag{3}$$

vorgeschriebene Eigenwerte λ_j erhält. Diese Aufgabe wurde von J. Uhlig [7], H. Heinrich [3], A. Fadini [2] behandelt. Zumeist konstruieren die Autoren Iterationsverfahren zur Lösung der aus den Beziehungen $\det(A \cdot D - \lambda_j E) = 0$ folgenden algebraischen Gleichungen.

In der vorliegenden Arbeit wird das endlichdimensionale Analogon zu (1) betrachtet. Zu einer fest gewählten symmetrischen Matrix A soll eine reelle Diagonalmatrix M gefunden werden, so daß

$$A + M \tag{4}$$

vorgegebene Eigenwerte d_j, $j = 1, \ldots, n$, besitzt. Dieses Problem wurde u. a. von A. C. Downing und A. S. Householder [1] gestellt. Man erkennt leicht, daß es ein Spezialfall einer allgemeineren Aufgabe ist, die auch die direkte Eigenwertaufgabe einer symmetrischen Matrix als Sonderfall enthält.

Wir geben eine Inhaltsübersicht: In Nr. 2 werden das inverse Problem und seine Verallgemeinerung definiert. In Nr. 3 geben wir eine simultane Einschließung der Eigenwerte einer symmetrischen Matrix, die man durch iterierte Anwendung des Templeschen Satzes erhält, sowie eine Abschätzung für die Änderungen der Eigenwerte bei Störung einer symmetrischen Matrix durch eine Diagonalmatrix. Diese Ungleichungen sind auch an sich von Interesse. In Nr. 4 wird der Fall, daß die Ordnung der Matrizen in (4) gleich 2 ist, vollständig erledigt. In Nr. 5 erhalten wir durch einfache Überlegungen über Drehungen im Raum der Dimension $n(n+1)/2$ notwendige Bedingungen für die Lösbarkeit der inversen Aufgabe. Nr. 4 und Nr. 5 sind von Nr. 3 unabhängig. In Nr. 6 wird auf eine der inversen Aufgabe äquivalente Operatorgleichung der Brouwersche Fixpunktsatz angewandt. Es ergibt sich eine für die Lösbarkeit hinreichende Bedingung: Bezeichne $\mathfrak{h}$ eine gewisse Halbnorm auf der Menge der symmetrischen Matrizen und d den minimalen Abstand der Zahlen d_j, $j = 1, \ldots, n$. Dann ist die inverse Aufgabe für $\mathfrak{h}(A)/d \leqslant 0.288$ lösbar. Eine andere Abschätzung ermöglicht die Anwendung des Banachschen Kontraktionssatzes, der unter einer nur wenig stärkeren Voraussetzung ($\mathfrak{h}(A)/d \leqslant 0.25$) lokale Eindeutigkeit und Konvergenz eines Iterationsverfahrens sichert (Nr. 7). In Nr. 8 wird das Verfahren an einem Beispiel demonstriert.

Herrn Roland Wais danke ich für die Durchführung der Rechnung auf der Rechenanlage TR4 der Universität Hamburg.

2. PROBLEMSTELLUNG

Sei H der reelle R^n, $n \geqslant 2$, mit der gewöhnlichen euklidischen Metrik. Das innere Produkt sei $(\ ,\ )$. Für $x = \{x_j\} \in H$ sei

$$\|x\|_2 = \left(\sum_{j=1}^{n} x_j^2\right)^{1/2}, \qquad \|x\| = \max_j |x_j|.$$

$\mathfrak{A}$ sei die Algebra aller reellen quadratischen Matrizen der Ordnung n. $\mathfrak{A}$ ist ein reeller Vektorraum der Dimension n^2. Wir versehen auch $\mathfrak{A}$

mit der euklidischen Metrik. Für Matrizen $B = (b_{jk})$, $C = (c_{jk}) \in \mathfrak{A}$ ist das Skalarprodukt

$$\{B, C\} = \sum_{j,k=1}^{n} b_{jk} c_{jk} = \operatorname{tr}(BC^*). \tag{5}$$

(Das Symbol $\operatorname{tr}(B)$ bedeutet die Spur von B.) Die Norm von B ist $|B|$,

$$|B|^2 = \operatorname{tr}(BB^*). \tag{6}$$

In der Theorie der Matrixnormen (vgl. [4]) bezeichnet man (6) als E. Schmidt- bzw. Hilbert-Schmidt-Norm.

$\mathfrak{H}$ sei der Teilraum aller symmetrischen Matrizen aus $\mathfrak{A}$, $\mathfrak{D}$ der Teilraum aller Diagonalmatrizen. Die Dimensionen von $\mathfrak{H}$ und $\mathfrak{D}$ sind $n(n+1)/2$ und n. $\mathfrak{U}$ sei die multiplikative Gruppe aller reellen orthogonalen Matrizen der Ordnung n.

Wir betrachten das folgende Problem. Seien eine Matrix $A \in \mathfrak{H}$ und n reelle Zahlen $d_1, \ldots, d_n$ gegeben. Gesucht ist $M \in \mathfrak{D}$, so daß $A + M$ die Eigenwerte $d_1, \ldots, d_n$ besitzt. Diese Aufgabe ist ein Spezialfall des folgenden allgemeineren Problems.

PROBLEM I. *Seien $A, S \in \mathfrak{H}$ vorgegeben. Gesucht ist $U \in \mathfrak{U}$, so daß $U^*SU - A \in \mathfrak{D}$ ist.*

Der oben formulierte Spezialfall ergibt sich als

PROBLEM II. *Wie* I, *aber $S \in \mathfrak{D}$.*

Auch die gewöhnliche Eigenwertaufgabe läßt sich hier als elementarer Sonderfall einordnen:

PROBLEM III. *Wie* I, *aber $A = 0$.*

Die Probleme I und II sind im wesentlichen äquivalent, da es nur auf die Eigenwerte von S ankommt. Problem III ist immer lösbar, Problem II im allgemeinen nicht.

Wir machen im folgenden die nicht einschränkende Voraussetzung, daß die Diagonale der Matrix A in Problem I nur Nullen enthält.

3. HILFSSÄTZE

Sei $B \in \mathfrak{H}$ und $x \in H$, $x \neq 0$. Die Momente (oder Schwarzschen Konstanten) seien

$$b_0 = (x, x), \qquad b_1 = (Bx, x), \qquad b_2 = (Bx, Bx).$$

Bezeichnen wir das Spektrum von B mit $\sigma(B)$ und die leere Menge mit $\varnothing$, so gelten

LEMMA 1 (*Krylov-Bogoljubov-Weinstein*).

$$\{z: |b_0 z - b_1|^2 \leqslant b_2 b_0 - b_1^2\} \cap \sigma(B) \neq \varnothing.$$

LEMMA 2 (*Temple*). *Sei* $\lambda_1 \in \sigma(B)$, $r > 0$, *und*

$$\left(\lambda_1, \frac{b_1}{b_0} + r\right) \cap \sigma(B) = \varnothing.$$

Dann ist

$$\frac{b_1}{b_0} - \frac{b_2 b_0 - b_1^2}{b_0^2 r} \leqslant \lambda_1.$$

Wir werden Lemma 2 auch in der hierzu symmetrischen Form benutzen, um obere Schranken für Eigenwerte zu erhalten. Für die Beweise verweisen wir auf [4].

Sei nun $x_{(j)}$ $(j = 1, \ldots, n)$ ein System normierter Vektoren und

$$\begin{aligned} b_0^{(j)} &= (x_{(j)}, x_{(j)}) = 1, \\ b_1^{(j)} &= (Bx_{(j)}, x_{(j)}), \\ b_2^{(j)} &= (Bx_{(j)}, Bx_{(j)}), \\ \mathfrak{k}_j^2 &\doteq b_2^{(j)} b_0^{(j)} - b_1^{(j)2}, \qquad \mathfrak{k} \geqslant 0. \end{aligned} \tag{7}$$

Ohne Beschränkung der Allgemeinheit können wir annehmen, daß für die Rayleigh-Quotienten $b_1^{(j)}$ und die Eigenwerte λ_j von B die Numerierung

$$\begin{gathered} b_1^{(1)} \leqslant b_1^{(2)} \leqslant \cdots \leqslant b_1^{(n)}, \\ \lambda_1 \leqslant \lambda_2 \leqslant \cdots \leqslant \lambda_n \end{gathered} \tag{8}$$

gilt. Wir zeigen

LEMMA 3. *Sei* $\mathfrak{k} > 0$, $d > 0$, $d \geqslant 2\mathfrak{k}$ *und*

$$\mathfrak{k}_j \leqslant \mathfrak{k} \qquad (j = 1, \ldots, n),$$

$$|b_1^{(j)} - b_1^{(k)}| \geqslant d(1 - \delta_{jk}) \qquad (j, k = 1, \ldots, n).$$

Dann ist

$$|b_1^{(j)} - \lambda_j| \leqslant \tfrac{1}{2}[d - (d^2 - 4\mathfrak{k}^2)^{1/2}] \qquad (j = 1, \ldots, n).$$

Beweis: Sei $d > 2\mathfrak{k}$. Nach Lemma 1 enthalten die Intervalle

$$[b_1^{(j)} - \mathfrak{k}, b_1^{(j)} + \mathfrak{k}] \qquad (j = 1, \ldots, n)$$

je mindestens einen Eigenwert. Sie sind paarweise disjunktiv, also enthält jedes genau einen Eigenwert. Es gilt

$$b_1^{(1)} - \mathfrak{k} \leqslant \lambda_1 \leqslant b_1^{(1)}, \tag{9}$$

$$b_1^{(j)} - \mathfrak{k} \leqslant \lambda_j \leqslant b_1^{(j)} + \mathfrak{k} \qquad (j = 2, \ldots, n - 1),$$

$$b_1^{(n)} \leqslant \lambda_n \leqslant b_1^{(n)} + \mathfrak{k}. \tag{10}$$

Die Ungleichungen (9), (10) folgen aus den Extremaleigenschaften der Eigenwerte. Wir wenden Lemma 2 an: Für $j = 2, \ldots, n$ liefert die untere Schranke von λ_j eine verbesserte untere Schranke von λ_{j-1}, für $j = 1, \ldots, n - 1$ liefert die obere Schranke von λ_j eine verbesserte obere Schranke von λ_{j+1}. Es folgt

$$b_1^{(1)} - \frac{\mathfrak{k}^2}{(d - \mathfrak{k})} \leqslant \lambda_1 \leqslant b_1^{(1)},$$

$$b_1^{(j)} - \frac{\mathfrak{k}^2}{(d - \mathfrak{k})} \leqslant \lambda_j \leqslant b_1^{(j)} + \frac{\mathfrak{k}^2}{(d - \mathfrak{k})} \qquad (j = 2, \ldots, n - 1),$$

$$b_1^{(n)} \leqslant \lambda_n \leqslant b_1^{(n)} + \frac{\mathfrak{k}^2}{(d - \mathfrak{k})}.$$

Wir führen das Verfahren fort und erhalten eine Folge von Fehlern

$$r_1 \doteq \mathfrak{k}, \qquad r_2 \doteq \frac{\mathfrak{k}^2}{d - \mathfrak{k}}, \qquad r_3 \doteq \frac{\mathfrak{k}^2}{d - \mathfrak{k}^2/(d - \mathfrak{k})}, \ldots$$

oder

$$r_1 = \mathfrak{k}, \qquad r_{l+1} = \frac{\mathfrak{k}^2}{(d - r_l)} \qquad (l = 1, 2, \ldots). \tag{11}$$

Die Folge nimmt monoton ab und strebt gegen den Grenzwert

$$\tilde{\mathfrak{k}} \doteq \tfrac{1}{2}[d - (d^2 - 4\mathfrak{k}^2)^{1/2}] < \mathfrak{k}. \tag{12}$$

Da für die beiden Endeigenwerte nach (9), (10) die Abschätzungen auf einer Seite trivial sind, ist das Lemma für $d > 2\mathfrak{k}$ gezeigt. Für $d = 2\mathfrak{k}$ folgt die Behauptung aus der stetigen Abhängigkeit der Eigenwerte von der Matrix. Q.E.D.

Die Voraussetzungen von Lemma 3 seien erfüllt. Sei

$$y_{(j)} = \{y_{j1}, \ldots, y_{jn}\}^* \qquad (j = 1, \ldots, n), \tag{13}$$

normierter Eigenvektor der Matrix B zum Eigenwert λ_j. Wir geben eine Abschätzung für die Projektion von $y_{(j)}$ auf den von $x_{(j)}$ erzeugten Raum. Dazu setzen wir mit festem j

$$B_1 = \lambda_j y_{(j)} y_{(j)}^*, \qquad B_2 = \sum_{k \neq j} \lambda_k y_{(k)} y_{(k)}^*, \qquad B_1 + B_2 = B, \tag{14}$$

$$z_1 = (x_{(j)}, y_{(j)}) y_{(j)}, \qquad z_2 = x_{(j)} - z_1. \tag{15}$$

Mit $c \in R^1$ ist

$$\begin{aligned} b_2(c) \doteq \|(B - cE)x_{(j)}\|_2^2 &= \|(B_1 - cE)z_1\|_2^2 + \|(B_2 - cE)z_2\|_2^2 \\ &\geqslant r^2 \cdot \|z_2\|_2^2 \end{aligned}$$

für jedes r mit

$$r \leqslant \min_{k \neq j} |c - \lambda_k|. \tag{16}$$

Speziell erhalten wir für $c = b_1^{(j)}$, also

$$b_2(c) = b_2^{(j)} - b_1^{(j)2} \leqslant \mathfrak{k}^2,$$

und (vgl. (12)) $r = d - \tilde{\mathfrak{k}}$ die Ungleichung

$$\|z_2\|_2^2 \leqslant \frac{\mathfrak{k}^2}{(d - \tilde{\mathfrak{k}})^2} = \frac{\mathfrak{k}^2}{(d - \tilde{\mathfrak{k}})} \cdot \frac{1}{(d - \tilde{\mathfrak{k}})} = \frac{\tilde{\mathfrak{k}}}{(d - \tilde{\mathfrak{k}})}, \tag{17}$$

dabei ergibt sich die letzte Umformung aus (11), (12). Nun seien die $x_{(j)}$ speziell die Einheitsvektoren,

$$x_{(j)} = e_j = (0, \ldots, \underset{j}{1}, \ldots, 0) \qquad (j = 1, \ldots, n). \tag{18}$$

Bei dieser Wahl der $x_{(j)}$ werden die Diagonalelemente von B Rayleigh-Quotienten. Es gilt

$$||z_2||_2{}^2 = \sum_{k \neq j} (e_j, y_{(k)})^2 = 1 - (e_j, y_{(j)})^2 = \sum_{k \neq j} (y_{(j)}, e_k)^2$$
$$= \sum_{k \neq j} y_{jk}^2. \tag{19}$$

Mit (17) erhalten wir für den Eigenvektor $y_{(j)}$ von B die Abschätzung

$$\sum_{k \neq j} y_{jk}^2 \leqslant \frac{d - (d^2 - 4\mathfrak{z}^2)^{1/2}}{d + (d^2 - 4\mathfrak{z}^2)^{1/2}}. \tag{20}$$

Wir stören B durch eine Diagonalmatrix

$$P = (p_j \, \delta_{jk}) \in \mathfrak{D}. \tag{21}$$

Mit $p = \{p_j\} \in H$ setzen wir

$$||P|| \doteq ||p|| = \max_j |p_j|. \tag{22}$$

Sei $\sigma(B + P) = \{\mu_1, \ldots, \mu_n\}$, $\mu_1 \leqslant \mu_2 \leqslant \cdots \leqslant \mu_n$. Nach einem Satz von Weyl (vgl. z. B. [6]) gilt

$$|\lambda_j - \mu_j| \leqslant ||P|| \qquad (j = 1, \ldots, n). \tag{23}$$

Wir nehmen $y_{(j)}$ als Probevektor für die gestörte Matrix $B + P$ und erhalten die Momente

$$\begin{aligned} a_0{}^{(j)} &\doteq (y_{(j)}, y_{(j)}) = 1, \\ a_1{}^{(j)} &\doteq ([B + P]y_{(j)}, y_{(j)}) = \lambda_j + (Py_{(j)}, y_{(j)}), \\ a_2{}^{(j)} &\doteq ([B + P]y_{(j)}, [B + P]y_{(j)}) \\ &= \lambda_j{}^2 + 2\lambda_j(Py_{(j)}, y_{(j)}) + (Py_{(j)}, Py_{(j)}), \\ a_2{}^{(j)}a_0{}^{(j)} - a_1{}^{(j)2} &\leqslant (Py_{(j)}, Py_{(j)}) \leqslant ||P||^2. \end{aligned} \tag{24}$$

Wir fordern

$$d^2 > 4(\mathfrak{z}^2 + ||P||^2). \tag{25}$$

Wegen

$$\begin{aligned} |\lambda_j - \lambda_k| &= |(\lambda_j - b_{jj}) + (b_{jj} - b_{kk}) + (b_{kk} - \lambda_k)| \\ &\geqslant d - 2\tilde{\mathfrak{z}} = (d^2 - 4\mathfrak{z}^2)^{1/2} \qquad \text{für} \quad j \neq k \end{aligned} \tag{26}$$

gilt

$$\begin{aligned}|a_1^{(j)} - \mu_k| &\geqslant |\lambda_j - \lambda_k| - |a_1^{(j)} - \lambda_j| - |\lambda_k - \mu_k| \\ &\geqslant (d^2 - 4\mathfrak{k}^2)^{1/2} - 2\|P\| \qquad \text{für} \quad j \neq k. \end{aligned} \tag{27}$$

Nun können wir Lemma 2 anwenden. An die Stelle von B tritt $B + P$, statt b_l ist $a_l^{(j)}$, $l = 0, 1, 2$, zu nehmen. Zähler und Nenner des in Lemma 2 auftretenden Quotienten werden nach (24), (27) abgeschätzt. Es ergibt sich

$$|\lambda_j + (Py_{(j)}, y_{(j)}) - \mu_j| \leqslant \frac{\|P\|^2}{(d^2 - 4\mathfrak{k}^2)^{1/2} - 2\|P\|} \qquad (j = 1, \ldots, n). \tag{28}$$

Wegen

$$p_j - (Py_{(j)}, y_{(j)}) = \sum_{k=1}^{n} (p_j - p_k) y_{jk}^2$$

gilt

$$|p_j - (Py_{(j)}, y_{(j)})| \leqslant \max_k |p_j - p_k| \cdot \sum_{k \neq j} y_{jk}^2. \tag{29}$$

Mit (20), (28), (29) folgt nun

Lemma 4. *Sei $d^2 > 4(\mathfrak{k}^2 + \|P\|^2)$. Dann gilt für die Eigenwerte μ_j der gestörten Matrix $B + P$ die Abschätzung*

$$|\lambda_j + p_j - \mu_j| \leqslant \rho(d, \mathfrak{k}, \|P\|) \cdot \|P\| \qquad (j = 1, \ldots, n), \tag{30}$$

wobei die Funktion ρ durch

$$\rho(\alpha, \beta, \gamma) \doteq \left\{ \frac{\gamma}{(\alpha^2 - 4\beta^2)^{1/2} - 2\gamma} + 2\,\frac{\alpha - (\alpha^2 - 4\beta^2)^{1/2}}{\alpha + (\alpha^2 - 4\beta^2)^{1/2}} \right\} \tag{31}$$

definiert ist.

4. Der Fall $n = 2$

Wenn $n = 2$ ist, läßt sich Problem I elementar behandeln. Sei

$$S = \begin{bmatrix} s_1 & s_2 \\ s_2 & s_4 \end{bmatrix}, \qquad A = \begin{bmatrix} 0 & a \\ a & 0 \end{bmatrix}, \qquad M = \begin{bmatrix} m_1 & 0 \\ 0 & m_2 \end{bmatrix}.$$

Für die Eigenwerte λ_1, λ_2 von S gilt

$$\lambda_1 + \lambda_2 = s_1 + s_4, \qquad \lambda_1 \cdot \lambda_2 = s_1 s_4 - s_2^2.$$

Andererseits sollen λ_1, λ_2 Eigenwerte von $A + M$ sein,

$$\lambda_1 + \lambda_2 = m_1 + m_2, \qquad \lambda_1 \cdot \lambda_2 = m_1 m_2 - a^2.$$

Durch Elimination von λ_1, λ_2 folgt

$$m_{1,2} = \tfrac{1}{2}\{s_1 + s_4 \pm [(s_1 - s_4)^2 + 4s_2{}^2 - 4a^2]^{1/2}\}. \tag{32}$$

Die Forderung, daß m_1, m_2 reell seien, führt auf

$$(s_1 - s_4)^2 + 4s_2{}^2 \geqslant 4a^2. \tag{33}$$

Äquivalent hierzu ist

$$|\lambda_1 - \lambda_2| \geqslant 2|a|. \tag{34}$$

Nach leichter Umformung von (33) folgt

SATZ 1. *Notwendig und hinreichend für die Lösbarkeit des Problems* I *im Falle $n = 2$ ist das Bestehen der Ungleichung*

$$\mathrm{tr}(S^2) - \mathrm{tr}(A^2) \geqslant \tfrac{1}{2}[\mathrm{tr}(S)]^2. \tag{35}$$

5. FÜR DIE LÖSBARKEIT NOTWENDIGE BEDINGUNGEN

Jedes $U \in \mathfrak{U}$ definiert eine orthogonale Transformation

$$\varphi_U \colon C \to U^*CU, \qquad C \in \mathfrak{H} \tag{36}$$

von $\mathfrak{H}$ in sich. Die Gesamtheit der φ_U, $U \in \mathfrak{U}$, bildet eine zur n-dimensionalen Drehgruppe isomorphe Gruppe Φ, die eine Untergruppe der Gruppe aller orthogonalen Transformationen von $\mathfrak{H}$ ist.

Sei E die Einheitsmatrix. Es gilt

$$\varphi_U E = E \qquad \text{für alle } U \in \mathfrak{U}. \tag{37}$$

Andererseits gilt $(\varphi_U)^* = \varphi_{U^*}$ und damit

$$\mathrm{tr}(C) = 0 \Leftrightarrow \{C, E\} = 0 \Leftrightarrow \{C, \varphi_{U^*}E\} = 0 \Leftrightarrow \{\varphi_U C, E\} = 0. \tag{38}$$

Folglich sind die Teilräume

$$\mathfrak{H}_1 \doteq \{C : \{C, E\} = 0\}, \qquad \mathfrak{H}_2 = \mathfrak{H} \ominus \mathfrak{H}_1$$

von $\mathfrak{H}$ invariant bei Φ.

Das Element S aus Problem I werde in

$$S = S_1 + S_2, \qquad S_j \in \mathfrak{H}_j \quad (j = 1, 2)$$

zerlegt. Ein normiertes Basiselement in $\mathfrak{H}_2$ ist

$$\tilde{E} = n^{-1/2}E \Rightarrow S_2 = \{S, \tilde{E}\}\tilde{E} = n^{-1}\operatorname{tr}(S)\cdot E,$$
$$|S_2|^2 = n^{-1}[\operatorname{tr}(S)]^2. \tag{39}$$

Problem I sei lösbar. Dann gibt es $U \in \mathfrak{U}$ mit

$$\varphi_U S = \varphi_U S_1 + S_2 = A + M, \qquad M \in \mathfrak{D},$$

das ist

$$\varphi_U S_1 = A + M - S_2 \doteq A + M_1, \quad M_1 \in \mathfrak{D}.$$

Für die Normen gilt

$$|S_1|^2 = |\varphi_U S_1|^2 = |A|^2 + |M_1|^2 \Rightarrow |S_1|^2 \geqslant |A|^2.$$

Nach (39) gilt

$$|S_1|^2 = |S|^2 - |S_2|^2 = \operatorname{tr}(S^2) - n^{-1}[\operatorname{tr}(S)]^2,$$

damit haben wir den

SATZ 2. *Notwendig für die Lösbarkeit von Problem* I *ist das Bestehen der Ungleichung*

$$\operatorname{tr}(S^2) - \operatorname{tr}(A^2) \geqslant n^{-1}[\operatorname{tr}(S)]^2. \tag{40}$$

Nach Satz 1 ist diese Bedingung im Falle $n = 2$ auch hinreichend.

Beim Jacobi-Verfahren zur Lösung von Problem III wird eine Lösung U als (im allgemeinen unendliches) Produkt von Elementarrotationen U_j aufgebaut. Die U_j sind Drehungen, die $n - 2$ Einheitsvektoren fest lassen, sie lösen jeweils ein zweidimensionales Teilproblem. Der folgende Satz gibt einen ähnlichen Zusammenhang zwischen Problem I und seinen Teilproblemen.

SATZ 3. *Sei* $n \geqslant 3$. *Gilt*

$$s_{ii}^2 + s_{ij}^2 + 2s_{ij}^2 - 2a_{ij}^2 \leqslant \tfrac{1}{2}(s_{ii} + s_{jj})^2 \tag{41}$$

für alle Indexpaare $i, j = 1, \ldots, n$, $i \neq j$, *und gilt für mindestens ein Indexpaar starke Ungleichheit, so ist Problem* I *unlösbar.*

Beweis: Aus (41) folgt durch Summation über alle Paare $i \neq j$ mit der zusätzlichen Voraussetzung die Ungleichung

$$2(n-1)\sum_i s_{ii}^2 + 2\sum_{\substack{i,j\\ i\neq j}} s_{ij}^2 - 2\sum_{i,j} a_{ij}^2$$

$$< (n-1)\sum_i s_{ii}^2 + \sum_{\substack{i,j\\ i\neq j}} s_{ii}s_{jj}, \tag{42}$$

das ist

$$2|S|^2 - 2|A|^2 < \sum_{\substack{i,j\\ i\neq j}} s_{ii}s_{jj} - (n-3)\sum_i s_{ii}^2. \tag{43}$$

Für beliebige reelle c_i $(i = 1, \ldots, n)$, gilt

$$0 \leqslant \sum_{i,j} c_i c_j \leqslant n \sum_i c_i^2,$$

$$n\left(\sum_{i,j} c_i c_j - (n-2)\sum_i c_i^2\right) \leqslant 2\left(\sum_i c_i\right)^2,$$

so daß aus (43) die Beziehung

$$|S|^2 - |A|^2 < \tfrac{1}{2}\left[\sum_{i,j} s_{ii}s_{jj} - (n-2)\sum_i s_{ii}^2\right] \leqslant \frac{1}{n}\left(\sum_i s_{ii}\right)^2$$

folgt. Die notwendige Bedingung (40) ist nicht erfüllt.

6. EIN HINREICHENDES KRITERIUM

Sei $A \in \mathfrak{H}$ mit $a_{jj} = 0$ für $j = 1, \ldots, n$ fest gewählt. Sei $\mathfrak{k} > 0$ mit

$$\sum_{k=1}^{n} a_{jk}^2 \leqslant \mathfrak{k}^2 \qquad \text{für} \quad j = 1, \ldots, n. \tag{44}$$

Jedem Vektor $v = \{v_k\} \in H$ werde die Matrix

$$D(v) = (v_j \cdot \delta_{jk}) \tag{45}$$

eineindeutig zugeordnet. Die Matrix $A + D(v)$ besitze die Eigenwerte

$$\lambda_1(v) \leqslant \lambda_2(v) \leqslant \cdots \leqslant \lambda_n(v), \tag{46}$$

die wir zu einem Vektor

$$\lambda(v) = \{\lambda_1(v), \ldots, \lambda_n(v)\} \tag{47}$$

zusammenfassen. Durch

$$F(v) \doteq \lambda(v), \qquad v \in H, \tag{48}$$

erklären wir eine Abbildung von H in den Bereich

$$G \doteq \{v : v_1 \leqslant v_2 \leqslant \cdots \leqslant v_n\} \subset H.$$

Wir versehen H mit der durch

$$\|v\| \doteq \max_j |v_j| \tag{49}$$

definierten Norm.

Zu $v \in H$ definieren wir

$$d(v) \doteq \min |v_j - v_k|, \qquad j \neq k, \quad j, k = 1, \ldots, n. \tag{50}$$

Wir wenden Lemma 3 auf die Matrix $A + D(v)$ an. Dazu wählen wir die $x_{(j)}$ wie in (18) als Einheitsvektoren. Die Diagonale von $A + D(v)$ ist $D(v)$, $\mathfrak{k}$ hängt nur von A ab. Unter der Voraussetzung

$$d(v) \geqslant 2\mathfrak{k} \tag{51}$$

erhalten wir

$$\|F(v) - v\| \leqslant \tfrac{1}{2}[d(v) - (d(v)^2 - 4\mathfrak{k}^2)^{1/2}]. \tag{52}$$

Sei $s \in G$ fest gewählt, sei $\varepsilon > 0$ und $d(s) \geqslant 2\varepsilon$. Dann ist

$$K_\varepsilon \doteq \{v : \|v - s\| \leqslant \varepsilon\} \subset G, \tag{53}$$

und es gilt

$$d(v) \geqslant d(s) - 2\varepsilon \qquad \text{für alle } v \in K_\varepsilon. \tag{54}$$

Wir definieren eine neue Abbildung

$$T(v) \doteq v + s - F(v), \tag{55}$$

$$T(v) - s = v - F(v). \tag{56}$$

Fordern wir nun

$$\delta \doteq d(s) \geqslant 2(\mathfrak{k} + \varepsilon), \tag{57}$$

so gilt in K_ε die Ungleichung (51) und damit auch (52), also nach (56), (52), (54)

$$\|T(v) - s\| \leqslant \tfrac{1}{2}\{\delta - 2\varepsilon - [(\delta - 2\varepsilon)^2 - 4\mathfrak{k}^2]^{1/2}\}. \tag{58}$$

Hinreichend dafür, daß T die Kugel (bzgl. der Norm (49)) K_ε in sich abbildet, ist also

$$\tfrac{1}{2}\{\delta - 2\varepsilon - [(\delta - 2\varepsilon)^2 - 4\mathfrak{k}^2]^{1/2}\} \leqslant \varepsilon \tag{59}$$

zusammen mit (57).

Wir normieren durch

$$\kappa \doteq \mathfrak{k}/\delta, \qquad \eta \doteq \varepsilon/\delta, \qquad \kappa > 0, \qquad \eta > 0, \tag{60}$$

und erhalten aus (57), (59) die Ungleichungen

$$1 \geqslant 2(\kappa + \eta), \tag{61}$$

$$1 - 4\eta \leqslant [(1 - 2\eta)^2 - 4\kappa^2]^{1/2}. \tag{62}$$

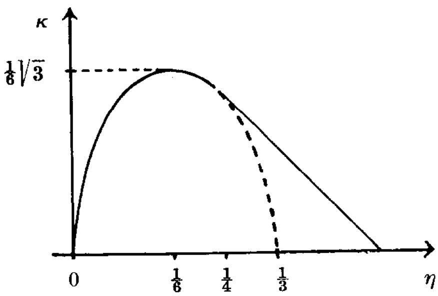
Fall 1: $4\eta \leqslant 1$. Dann gibt (62)

$$3\eta^2 + \kappa^2 \leqslant \eta, \tag{63}$$

und daraus folgt (61).

Fall 2: $4\eta \geqslant 1$. Dann folgt (62) aus (61).

Der durch (61), (62) und $\eta \geqslant 0$, $\kappa \geqslant 0$ beschriebene Bereich wird also von einem Ellipsenbogen und zwei Strecken begrenzt. Aus der Skizze



ist zu ersehen, daß der maximale zulässige Wert von κ sich für $\eta = \frac{1}{6}$ ergibt und $\kappa = \sqrt{3}/6$ beträgt. Da K_ε konvex und kompakt ist, liefert der Brouwersche Fixpunktsatz die Existenz eines Fixpunktes von T in K_ε, also eines Punktes v_0 mit $F(v_0) = s$. Damit gilt

Satz 4. *Das Problem* II *sei vorgelegt. Für die Matrix* $A = (a_{jk})$ *gelte*

$$a_{jj} = 0, \qquad \sum_{k=1}^{n} a_{jk}^2 \leqslant \mathfrak{k}^2 \qquad (j, k = 1, \ldots, n), \quad \mathfrak{k} \geqslant 0.$$

Für die Matrix $S = (s_j \cdot \delta_{jk})$ *gelte*

$$|s_j - s_k| \geqslant \delta(1 - \delta_{jk}), \qquad \delta > 0 \quad (j, k = 1, \ldots, n).$$

Sei

$$\kappa = \frac{k}{\delta} \leqslant \frac{1}{6}\sqrt{3} = 0.288\ldots . \tag{64}$$

Dann ist Problem II *lösbar.*

Nach Satz 1 gibt es für $\kappa > \frac{1}{2}$ unlösbare Probleme. Zu vorgegebenem $\kappa \leqslant \sqrt{3}/6$ ist wegen (63) das minimale zulässige η durch

$$\eta_0(\kappa) = \tfrac{1}{6}[1 - (1 - 12\kappa^2)^{1/2}]. \tag{65}$$

gegeben, daher gilt das

Corollar. *Wenn die Voraussetzungen von Satz* 4 *erfüllt sind, gibt es eine Lösung* $v_0 \in G$ *bzw.* $V_0 \in \mathfrak{D}$ *von Problem* II *mit*

$$\|S - V_0\| = \|s - v_0\| \leqslant \tfrac{1}{6}[\delta - (\delta^2 - 12k^2)^{1/2}]. \tag{66}$$

7. KONTRAKTION BEI DER TRANSFORMATION T

T sei bezüglich der Norm (49) kontrahierend in einer Menge $\mathscr{F} \subset H$, d. h. es gelte

$$\|T(u) - T(v)\| \leqslant K\|u - v\|, \qquad K < 1, \tag{67}$$

für alle $u, v \in \mathscr{F}$.

Setzen wir $u - v = p$, so ist dies wegen

$$T(v + p) - T(v) = F(v) + p - F(v + p)$$

äquivalent zu:

$$\|F(v + p) - F(v) - p\| \leqslant K\|p\|, \qquad K < 1, \tag{68}$$

für alle v, p mit $v, v + p \in \mathscr{F}$.

Wir wenden Lemma 4 auf die Matrix $A + D(v)$ an, dabei setzen wir

$$B = A + D(v), \qquad P = D(p), \qquad \|P\| = \|p\|,$$
$$F(v) = \{\lambda_j\}, \qquad F(v + p) = \{\mu_j\}.$$

Es ist zu beachten, daß die Größe $\mathfrak{k}$ nur von A, die Größe $d = d(v)$ (definiert durch (50)) nur von v abhängt. Unter der Voraussetzung

$$d(v)^2 > 4(\mathfrak{k}^2 + ||P||^2) \tag{69}$$

gilt

$$||F(v+p) - F(v) - p|| \leqslant \rho(d(v), \mathfrak{k}, ||p||) \cdot ||p||. \tag{70}$$

Sei

$$\varepsilon > 0, \qquad \delta = d(s) > 2\varepsilon, \tag{71}$$

$$(\delta - 2\varepsilon)^2 > 4(\mathfrak{k}^2 + 4\varepsilon^2). \tag{72}$$

In der Kugel K_ε gilt

$$d(v) \geqslant \delta - 2\varepsilon, \tag{73}$$

und für $v_1, v_2 \in K_\varepsilon$, $p = v_1 - v_2$ gilt

$$||p|| = ||v_1 - v_2|| \leqslant 2\varepsilon. \tag{74}$$

Da $\rho(\alpha, \beta, \gamma)$ in α monoton fällt und in γ wächst, gilt in K_ε die Ungleichung

$$\rho(d(v), \mathfrak{k}, ||p||) \leqslant \rho(\delta - 2\varepsilon, \mathfrak{k}, 2\varepsilon). \tag{75}$$

Damit erhalten wir

LEMMA 5. *Wenn die Größen* $\delta, \mathfrak{k}, \varepsilon$ *die Ungleichungen* (71), (72) *und*

$$\rho(\delta - 2\varepsilon, \mathfrak{k}, 2\varepsilon) < 1 \tag{76}$$

erfüllen, so ist T *in der Kugel* K_ε *kontrahierend.*

Wir normieren wieder nach (60), kombinieren (59) und Lemma 5 und erhalten als hinreichende Bedingung dafür, daß T auf K_ε kontrahierend ist und K_ε in sich abbildet, das System von Ungleichungen

$$\begin{gathered} \eta > 0, \qquad 1 \geqslant 2(\eta + \kappa), \\ 1 - 4\eta \leqslant [(1 - 2\eta)^2 - 4\kappa^2]^{1/2}, \\ (1 - 2\eta)^2 > 4(\kappa^2 + 4\eta^2), \\ \rho(1, \kappa, \eta) < 1. \end{gathered} \tag{77}$$

Nach leichter Umformung folgt

SATZ 5. *Seien* $\kappa = \mathfrak{k}/\delta$ *und* $\eta = \varepsilon/\delta$ *derart, daß mit reellem*

$$w \doteq +[(1-2\eta)^2 - 4\kappa^2]^{1/2}$$

die Ungleichungen

$$\eta > 0, \qquad 1 > 2\eta, \qquad 4\eta < w, \qquad 1 - 4\eta \leqslant w,$$

$$\rho = \frac{2\eta}{w - 4\eta} + 2\,\frac{1 - 2\eta - w}{1 - 2\eta + w} < 1$$

gelten. Dann ist $TK_\varepsilon \subset K_\varepsilon$, *und* T *ist auf* K_ε *kontrahierend mit der Lipschitz-Konstanten* ρ.

Für $\kappa \leqslant \frac{1}{4}$ lassen sich die Voraussetzungen mit $\eta = \frac{1}{12}$ erfüllen. Es ergibt sich $\rho \approx \frac{13}{18}$. Mit dem Banachschen Fixpunktsatz erhalten wir das

COROLLAR. *Für* $\kappa \leqslant \frac{1}{4}$ *konvergiert das Iterationsverfahren*

$$w_0 = s, \qquad w_{k+1} = Tw_k, \qquad k = 0, 1, 2, \ldots \tag{78}$$

gegen eine Lösung v_0 *von*

$$F(v) = s.$$

v_0 *ist die einzige Lösung in jeder Kugel* K_ε, *für die* ε *die Voraussetzung in Satz* 5 *erfüllt.*

8. EIN BEISPIEL

Wegen (48), (55) sind bei jedem Schritt der Iteration (78) sämtliche Eigenwerte einer symmetrischen Matrix zu bestimmen. Dies kann mit einem Jacobi-Verfahren geschehen. Man wird das Jacobi-Verfahren für ein festes k in (78) nicht beliebig genau durchführen, sondern nur soweit, bis der Fehler der Jacobi-Näherung für $F(w_k)$ wesentlich kleiner (z. B. um eine Zehnerpotenz) als der Fehler von w_k selbst ist. Daher hat man im Prinzip ein zweistufiges Verfahren.

Beispiel:

$$A = \begin{pmatrix} 0 & 4 & 3 & 2 & 1 \\ 4 & 0 & 3 & 2 & 1 \\ 3 & 3 & 0 & 2 & 1 \\ 2 & 2 & 2 & 0 & 1 \\ 1 & 1 & 1 & 1 & 0 \end{pmatrix}, \qquad s = s_\nu = 0.8^\nu \cdot \begin{pmatrix} 0 \\ 32 \\ 60 \\ 90 \\ 115 \end{pmatrix}.$$

Diese Aufgabe wurde für $\nu = 0, \ldots, 8$ durchgerechnet. Die folgende Tabelle gibt die Anzahl k_ν der Iterationsschritte an, die nötig sind, um $\|F(w_k) - s\| \leqslant 10^{-5}$ zu erreichen.

ν	0	1	2	3	4	5	6	7	8
k_ν	4	5	5	6	8	11	17	27	Divergenz

Die Bedingung von Satz 5 ist wegen $k = \sqrt{30}$ und $d(s_0) = 25$ nur für $\nu = 0$ erfüllt. Diese Bedingung ist also durchaus nicht notwendig für die Konvergenz.

$\nu = 3$

$s = w_0$	0.000	16.384	30.720	46.080	58.880
w_1	1.135166	16.474473	30.167754	45.606163	58.680444
w_2	1.192392	16.474391	30.119162	45.595653	58.682402
w_3	1.195747	16.475060	30.114607	45.596010	58.682577
w_4	1.195920	16.475259	30.114138	45.596100	58.682583
w_5	1.195922	16.475299	30.114084	45.596111	58.682584
w_6	1.195921	16.475316	30.114078	45.596112	58.682583

$\nu = 7$

$s = w_0$	0.000	6.710886	12.582912	18.874368	24.117248
w_1	2.068051	7.362790	11.840076	17.671844	23.342653
w_2	2.327299	7.635126	11.648909	17.361719	23.312361
w_{10}	2.272374	7.947337	11.576137	17.145172	23.344394
w_{20}	2.272646	7.944887	11.579515	17.143887	23.344480
w_{26}	2.272689	7.944720	11.579663	17.143863	23.344480
w_{27}	2.272691	7.944711	11.579671	17.143861	23.344780

Schon die ersten Schritte der Iteration bringen häufig eine recht gute Näherung.

LITERATUR

1 A. C. Downing und A. S. Householder, Some Inverse Characteristic Value Problems, *Journ. Assoc. for Computing Machinery* **3**(1956), 203–207.

2 A. Fadini, Beitrag zur Lösung eines inversen Eigenwertproblems, *ZAMM* **44**(1964), 502–508.

3 H. Heinrich, Ein inverses Eigenwertproblem für endliche Matrizen und seine graphische Lösung für $n = 3$, *ZAMM* **40**(1960), 62–64.

4 A. S. Householder, *The Theory of Matrices in Numerical Analysis*, Ginn (Blaisdell), New York, **1964**.

5 M. A. Neumark, *Lineare Differentialoperatoren*, Akademie-Verlag, Berlin, 1960.
6 F. Riesz und B. Sz.-Nagy, *Vorlesungen über Funktional-Analysis*, Deutscher Verlag der Wissenschaften, Berlin, 1956.
7 J. Uhlig, Ein Iterationsverfahren für ein inverses Eigenwertproblem endlicher Matrizen, *ZAMM* **40**(1960), 123–125.

Received February 28, 1967