Zeitschrift für Angewandte Mathematik und Mechanik

ZAMM

Applied Mathematics and Mechanics

Volume 67 1987 Number 3

ZAMM · Z. angew. Math. Mech. 67 (1987) 3, 145 – 152

ALEFELD, G.

Berechenbare Fehlerschranken für ein Eigenpaar unter Einschluß von Rundungsfehlern bei Verwendung des genauen Skalarprodukts

Herrn WOLFGANG WALTER, Karlsruhe, zum Sechzigsten Geburtstag am 2. 5. 1987 gewidmet.

In der vorliegenden Arbeit wird, ausgehend von einer Näherung für ein Matrizeneigenpaar mit einem algebraisch einfachen Eigenwert, zunächst eine Einschließung für den Eigenwert und für die Komponenten des dazugehörigen Eigenvektors berechnet. Daran anschließend wird ein Verfahren eingeführt, welches es gestattet, diese Schranken (bei rundungsfehlerfreier Rechnung) beliebig genau zu verbessern. Schließlich wird diskutiert, wie sich dieses Verfahren bei Berücksichtigung von Rundungsfehlern verhält. Dabei setzen wir voraus, daß das von U. Kulisch [3] eingeführte genaue Skalarprodukt zur Verfügung steht. Zwei numerische Beispiele schließen die Arbeit ab.

Let there be given a sufficiently accurate approximation to a real eigenpair of a real matrix, where the eigenvalue is a simple zero of the characteristic polynomial. Then we construct bounds for the eigenvalue and for the components of the eigenvector. Furthermore we introduce an iteration method by which these bounds can be made arbitrarily small. Finally we discuss how this method behaves if all rounding errors are taken into account. For this discussion we assume that the so-called precise scalar product which was first introduced by U. Kulisch [3] is available. The paper closes with two numerical examples.

Исходя из приближения для веществennозначной пары собственных решений вещественной матрицы с алгебраическо простым собственным значениям в настоящей статей сперва вычисляется включение для собственного значения и для компонентов соответствующего собственного вектора. Кроме того вводится метод итерации позволяющий улучшить эти грани на любую точность. На конец обсуждается, как этот метод работает, если принимаются в расчет все погрешности округления. При этом предполагается, что возможно располагать введенным У. Кулишом в [3] скалярным произведением. Статья заканчивает с двумя численным примером.

0. Einleitung

Wir betrachten die Aufgabe, ausgehend von einer reellen Näherung λ für einen einfachen reellen Eigenwert und einem reellen Näherungsvektor x für den dazugehörigen Eigenvektor einer reellen (n,n)-Matrix, Fehlerschranken zu berechnen. Diese Aufgabenstellung wurde in der Vergangenheit wiederholt betrachtet, zuletzt z. B. von S. RUMP [5], H. J. SYMM und J. H. WILKINSON [6] und T. YAMAMOTO [8]. In [5] und [6] wird auch auf die beim praktischen Rechnen entstehenden Rundungsfehler eingegangen. Während in [6] eine qualitative Diskussion der Rundungsfehler durchgeführt wird, werden in [5] Schranken unter Einschluß aller Rundungsfehler berechnet.

In dieser Arbeit wird zunächst in Abschnitt 1 wie in [6] ein nichtlineares Gleichungssystem mit n Gleichungen und n Unbekannten aufgestellt, dessen (exakte) Auflösung mit der Bestimmung eines Eigenpaares äquivalent ist. (Im Gegensatz dazu wurde in [5] ein nichtlineares Gleichungssystem mit $n+1$ Unbekannten und $n+1$ Gleichungen aufgelöst.)

Daran anschließend wird gezeigt, daß man mit hinreichend guten Näherungen λ und x den Eigenwert und die Komponenten des Eigenvektors in Schranken einschließen kann (Satz 2). Beginnend mit dieser Einschließung kann

man Folgen von Intervallvektoren berechnen, welche das Eigenpaar fortwährend einschließen und (theoretisch, d. h. bei rundungsfehlerfreier Rechnung) gegen dieses konvergieren.

Schließlich wird in Abschnitt 2 eine genaue Untersuchung des vorgeschlagenen Verfahrens angegeben, wenn es auf einer Rechneranlage unter Verwendung eines Gleitpunktzahlensystems durchgeführt wird. Dabei setzen wir voraus, daß — wie von U. KULISCH vorgeschlagen (siehe [3] und [4]) — Skalarprodukte mit maximaler Genauigkeit berechnet werden können, wie dies z. B. in der Programmiersprache PASCAL SC möglich ist. Diese Programmiersprache steht heute auf zahlreichen Kleinrechnern, das genaue Skalarprodukt auf allen IBM-Anlagen zur Verfügung.

Im letzten Abschnitt geben wir zwei numerische Beispiele an.

1. Berechnung von Schranken für ein Eigenpaar und ihre iterative Verbesserung

Gegeben sei die reelle Matrix A, λ und x seien reelle Näherungen für einen reellen einfachen Eigenwert $\lambda + \mu$ und einen dazugehörigen reellen Eigenvektor $x + \tilde{y}$, so daß also

$$A \cdot (x + \tilde{y}) = (\lambda + \mu) \cdot (x + \tilde{y}) \tag{1}$$

gilt. λ und x können z. B. mit einem der zahlreichen bekannten Näherungsverfahren berechnet worden sein. Siehe etwa [7]. Es sei mit $x = (x_i)$

$$||x||_\infty = |x_s| > 0\,, \tag{2}$$

wobei wir, falls es mehrere Indizes gibt, für welche die Unendlichnorm angenommen wird, z. B. s als den kleinsten solchen Index wählen können.

Da der Eigenvektor $x + \tilde{y}$ zunächst nicht eindeutig festgelegt ist, können wir die s-te Komponente von $\tilde{y} = (\tilde{y}_i)$ gleich Null setzen (Normierung von $x + \tilde{y}$):

$$\tilde{y}_s = 0\,. \tag{3}$$

Die Gleichung (1) kann geschrieben werden als

$$(A - \lambda I)\,\tilde{y} - \mu x = (\lambda I - A)\,x + \mu\tilde{y}\,. \tag{4}$$

Führen wir nun einen Vektor $y = (y_i) \in \mathbb{R}^n$ ein durch

$$y_i = \begin{cases} \tilde{y}_i\,, & i \neq s \\ \mu\,, & i = s \end{cases}, \tag{5}$$

so kann unter Beachtung von (3) die Gleichung (4) geschrieben werden als

$$By = r + y_s \cdot \tilde{y} \tag{6}$$

mit dem Residuenvektor

$$r = \lambda x - Ax \tag{7}$$

und der Matrix B

$$B = \left((A - \lambda I)_1, \ldots, (A - \lambda I)_{s-1}, -x, (A - \lambda I)_{s+1}, \ldots, (A - \lambda I)_n\right). \tag{8}$$

Die Matrix B entsteht aus $A - \lambda I$, in dem die s-te Spalte durch $-x$ ersetzt wird und alle anderen Spalten beibehalten werden.

Die Umformung von (1) in (6) mit der Normierung (3) wurde bereits in [6] durchgeführt.

Für hinreichend gute Näherungen λ und x ist die Matrix B nichtsingulär. Dies ist aus Stetigkeitsgründen eine Folge aus der in [6] bewiesenen Aussage:

Satz 1: *Es sei (λ, x) ein (exaktes) Eigenpaar von A. λ sei einfache Nullstelle des charakteristischen Polynoms. Dann ist die Matrix B, die aus $A - \lambda I$ dadurch entsteht, daß man die s-te Spalte (wobei s durch (3) definiert ist) durch den Vektor $-x$ ersetzt, nichtsingulär.*

Die Gleichung (6) ist der Ausgangspunkt für die Berechnung von gesicherten Schranken für $\tilde{y}$ und μ und damit für das Eigenpaar $(\lambda + \mu, x + \tilde{y})$.

Wir setzen voraus, daß λ und x so gute Näherungen sind, daß B nichtsingulär ist. L sei eine Näherung für die Inverse von B oder die exakte Inverse. Dann kann die Gleichung (6) geschrieben werden als

$$y = Lr + (I - LB)\,y + L(y_s \cdot \tilde{y})\,. \tag{9}$$

Wir bestimmen nun einen Intervallvektor $[y] = ([y]_i)$, für welchen

$$Lr + (I - LB)\,y + L(y_s \cdot \tilde{y}) \in [y] \tag{10}$$

für alle $y \in [y]$ gilt. Aufgrund des Brouwerschen Fixpunktsatzes besitzt die Gleichung (9) dann (mindestens) einen Fixpunkt y^* in $[y]$.

Zur Bestimmung eines Intervallvektors $[y]$, für welchen (10) gilt, machen wir den Ansatz

$$[y] = [-\beta, \beta] \cdot e\,, \tag{11}$$

wobei $0 < \beta \in \mathbb{R}$, $e = (1, \ldots, 1)^T \in \mathbb{R}^n$. Wir setzen außerdem

$$[\tilde{y}] = ([\tilde{y}]_i) \quad \text{mit} \quad [\tilde{y}]_i = \begin{cases} [y]_i\,, & i \neq s \\ 0\,, & i = s \end{cases}. \tag{12}$$

Aufgrund der Teilmengeneigenschaft (siehe [1], Chapter 1, Theorem 5) gilt für $y \in [y]$

$$Lr + (I - LB)\, y + L(y_s \cdot \tilde{y}) \in [w]$$

mit $[w] := Lr + (I - LB)\,[y] + L([y]_s \cdot [\tilde{y}])$. Daher gilt (10) sicherlich dann, wenn

$$[w] = Lr + (I - LB)\,[y] + L([y]_s \cdot [\tilde{y}]) \subseteq [y] \tag{13}$$

gilt.

Es gilt (13) genau dann, wenn für die Mittelpunkte $m[w]$ und $m[y]$ bzw. die Durchmesser $d[w]$ und $d[y]$ von $[w]$ und $[y]$ die Ungleichung

$$|m[w] - m[y]| + \tfrac{1}{2}\, d[w] \leqq \tfrac{1}{2}\, d[y] \tag{14}$$

gilt. Zur Definition von m und d sowie zu Rechenregeln dafür siehe [1], Chapter 10.

Für $[y]$ nach (11) gilt

$$m[y] = 0\,. \tag{15}$$

Außerdem ist

$$m[w] = Lr\,. \tag{16}$$

Mit $\tilde{e} = (\tilde{e}_i)$, $\tilde{e}_i = \begin{cases} 1, & i \neq s \\ 0, & i = s \end{cases}$, erhält man

$$d[w] = d\big(Lr + (I - LB)\,[y] + L([y]_s\,[\tilde{y}])\big) = 2\beta|I - LB|\, e + 2\beta^2|L|\, \tilde{e}\,. \tag{17}$$

Mit (15), (16) und (17) gilt (14) genau dann, wenn

$$|Lr| + \beta|I - LB|\, e + \beta^2|L|\, \tilde{e} \leqq \beta e$$

gilt. Diese Ungleichung ist sicherlich dann erfüllt, wenn β so gewählt wird, daß

$$|Lr| + \beta|I - LB|\, e + \beta^2|L|\, e \leqq \beta e \tag{18}$$

gilt. Wir setzen nun

$$\varrho = ||Lr||_\infty\,, \qquad \varkappa = ||I - LB||_\infty\,, \qquad l = ||L||_\infty\,, \tag{19), (20), (21)}$$

wobei der Index ∞ die Unendlichvektor- bzw. Unendlichmatrixnorm bezeichnet. Dann gilt (18) sicherlich dann, wenn β so gewählt wird, daß $\varrho + \beta\varkappa + \beta^2 l \leqq \beta$, d. h. wenn

$$\varrho + (\varkappa - 1)\,\beta + l\beta^2 \leqq 0 \tag{22}$$

gilt. Die quadratische Gleichung

$$l\beta^2 + (\varkappa - 1)\,\beta + \varrho = 0$$

hat unter den Voraussetzungen $\varkappa < 1$, $(\varkappa - 1)^2 - 4\varrho l \geqq 0$ die positiven Nullstellen

$$\beta_{1/2} = \frac{1 - \varkappa \mp \sqrt{(1 - \varkappa)^2 - 4\varrho l}}{2l}\,. \tag{23}$$

Somit haben wir den Beweis für den folgenden

Satz 2: *Es seien ϱ, $\varkappa$, l nach* (19), (20), (21) *definiert. Weiter sei $\varkappa < 1$ und $(1 - \varkappa)^2 - 4\varrho l \geqq 0$. Wird dann $\beta \in [\beta_1, \beta_2]$ gewählt, wobei β_1 und β_2, $\beta_1 \leqq \beta_2$, durch* (23) *definiert sind, so hat die Gleichung* (9) *(mindestens) einen Fixpunkt y^* im Intervallvektor $[y] = ([y]_i)$ mit $[y]_i = [-\beta, \beta]$.*

Ein Fixpunkt der Gleichung (9) ist sicherlich eine Lösung von (6), wenn L nichtsingulär ist. Unter der Voraussetzung $\varkappa < 1$ von Satz 2 ist dies stets der Fall, da wegen $||I - LB||_\infty = \varkappa < 1$ die Inverse von $I - (I - LB) = LB$ existiert.

Wir betrachten nun das *Iterationsverfahren*

$$[y]^0 = [-\beta, \beta]\, e\,, \qquad [y]^{k+1} = g([y]^k), \qquad k = 0, 1, 2, \ldots\,, \tag{24}$$

wobei

$$g([y]) = Lr + (I - LB)\,[y] + L \cdot ([y]_s \cdot [\tilde{y}])\,. \tag{25}$$

Es gelten die folgenden Aussagen.

Satz 3: *Es sei $\varkappa < 1$, $(\varkappa - 1)^2 - 4\varrho l > 0$, und β_1, β_2 seien durch* (23) *definiert. Genügt dann β (in* (24)*) der Ungleichung $\beta_1 \leqq \beta < \dfrac{\beta_1 + \beta_2}{2}$, so liefert* (24) *eine Folge von Intervallvektoren $\{[y]^k\}_{k=0}^\infty$ mit*

$$y^* \in [y]^k, \qquad k = 0, 1, 2, \ldots\,; \quad \textit{und} \quad \lim_{k \to \infty} [y]^k = y^*\,. \tag{26); (27)}$$

Dabei ist y^* der eindeutige Fixpunkt der Gleichung (9) *in $[y]^0$.*

Beweis: Nach Satz 2 gibt es (mindestens) einen Fixpunkt $y^* \in [y]^0$ der Gleichung (9). Ist $y^* \in [y]^k$, was für $k = 0$ der Fall ist, so folgt aufgrund der Teilmengeneigenschaft (siehe [1], Chapter 1, Theorem 5)

$$y^* = Lr + (I - LB)\, y^* + L \cdot (y_s^* \tilde{y}^*) \in Lr + (I - LB)\,[y]^k + L \cdot ([y]_s^k\,[\tilde{y}]^k) = [y]^{k+1}\,.$$

Also gilt (26).

Im Beweis von Satz 2 haben wir bereits gezeigt, daß $[y]^1 = g([y]^0) \subseteq [y]^0$ gilt. Es gelte $[y]^0 \supseteq [y]^1 \supseteq \ldots \supseteq [y]^{k-1} \supseteq [y]^k$, was für $k = 1$ gilt. Dann folgt aufgrund der Teilmengeneigenschaft (siehe [1], Chapter 1, Theorem 5)

$$[y]^k = g([y]^{k-1}) \supseteq g([y]^k) = [y]^{k+1} .$$

Die Folge $\{[y]^k\}_{k=0}^{\infty}$ bildet daher eine ineinander geschachtelte Folge von Intervallvektoren. Es gibt somit einen Intervallvektor $[y]^*$, für welchen $\lim_{k\to\infty} [y]^k = [y]^*$ gilt. Wir zeigen, daß $d[y]^* = 0$ gilt. Wegen $y^* \in [y]^k$ folgt dann die Behauptung (27) sowie die Eindeutigkeit von y^* in $[y]^0$.

Wir verwenden die Bezeichnung

$$d_k = ||d[y]^k||_\infty$$

und berücksichtigen, daß für zwei reelle Intervalle $[a]$, $[b]$ die Ungleichung

$$d([a] \cdot [b]) \leqq |[a]| \cdot d[b] + d[a] \cdot |[b]|$$

gilt. Siehe dazu [1], Chapter 2, Theorem 9. Dann erhält man aus (24) und (25)

$$\begin{aligned} d[y]^{k+1} &= |I - LB|\, d[y]^k + |L| \cdot d([y]_s^k \cdot [\tilde{y}]^k) \\ &\leqq |I - LB|\, d[y]^k + |L| \cdot \{d[y]_s^k \cdot |[\tilde{y}]^k| + |[y]_s^k| \cdot d[\tilde{y}]^k\} \\ &\leqq |I - LB|\, d[y]^k + |L| \cdot \{d_k \cdot \beta \cdot e + \beta \cdot d_k \cdot e\} \\ &\leqq d_k(|I - LB| + 2\beta|L|) \cdot e . \end{aligned}$$

Aus dieser Ungleichung folgt mit der Definition von $\varkappa$ und l die Ungleichung

$$d_{k+1} \leqq (\varkappa + 2\beta l)\, d_k . \tag{28}$$

Wegen $\beta < \dfrac{\beta_1 + \beta_2}{2} = \dfrac{1 - \varkappa}{2l}$ ist $\varkappa + 2\beta l < 1$. Damit folgt dann aus (28) $d[y]^* = \lim_{k\to\infty} d[y]^k = 0$.

2. Rundungsfehleranalysis

Um die nachfolgenden Ausführungen verständlicher zu machen, beginnen wir mit einigen allgemeinen Bemerkungen:

1. Für eine beliebige reelle nichtsinguläre Matrix B und einen Intervallvektor $[z]$ gilt aufgrund der Eigenschaften der Intervallrechnung

$$\{B^{-1} \cdot z \mid z \in [z]\} \subseteq B^{-1} \cdot [z] . \tag{29}$$

Es bezeichne IGA(B, $[z]$) *einen* Intervallvektor — auch bei rundungsfehlerfreier Rechnung ist IGA(B, $[z]$) nicht eindeutig, sondern hängt von der Art der Pivotisierung ab —, den man erhält, wenn man das Gaußsche Eliminationsverfahren mit B als Koeffizientenmatrix und $[z]$ als rechter Seite durchführt (siehe etwa [1], Chapter 15 für Einzelheiten). Dann gilt ebenfalls

$$\{B^{-1} \cdot z \mid z \in [z]\} \subseteq \mathrm{IGA}(B, [z]) . \tag{30}$$

In [2], Lemma 1, wurde gezeigt, daß stets

$$B^{-1} \cdot [z] \subseteq \mathrm{IGA}(B, [z]) \tag{31}$$

gilt, wobei i. allg. die echte Inklusion besteht. Die Berechnung von $B^{-1} \cdot [z]$ erfordert für großes n ungefähr den dreifachen Aufwand im Vergleich zur Berechnung von IGA(B, $[z]$). Andererseits zeigt (31), daß die Menge $\{B^{-1} \cdot z \mid z \in [z]\}$ durch $B^{-1} \cdot [z]$ gewöhnlich besser eingeschlossen wird als durch IGA(B, $[z]$).

2. Wenn wir nun $L := B^{-1}$ in (25) wählen, so erhalten wir, da B^{-1} eine Punktmatrix ist und somit das Distributivgesetz gilt

$$g([y]) = B^{-1}r + B^{-1}([y]_s\,[\tilde{y}]) = B^{-1}(r + [y]_s\,[\tilde{y}]) .$$

Wegen (31) erhalten wir mit $[z] := r + [y]_s\,[\tilde{y}]$ dann

$$g[y] = B^{-1} \cdot [z] \subseteq \mathrm{IGA}(B, [z]).$$

Anders ausgedrückt: Die Einschließung (13), welche die Existenz eines Fixpunktes garantiert, gilt sicherlich für $g[y] = B^{-1}(r + [y]_s\,[\tilde{y}])$, falls sie für IGA($B$, $r + [y]_s\,[\tilde{y}]$) gilt. Durch vollständige Induktion können wir außerdem mit (31) zeigen, daß für einen nach Satz 1.3 bestimmten Intervallvektor $[y]^0 = [-\beta, \beta]\,e$ das unter Verwendung des Gaußschen Algorithmus durchgeführte Verfahren

$$[y]^0 = [-\beta, \beta]\,e , \qquad [y]^{k+1} = \mathrm{IGA}(B, r + [y]_s^k\,[\tilde{y}]^k) \tag{32}$$

in jedem Schritt eine im Sinne der Inklusion gröbere Einschließung von y^* liefert, als Verfahren (24), (25). Da wir daran interessiert sind, y^* durch möglichst wenige Schritte möglichst gut einzuschließen, verwenden wir das Iterationsverfahren (24), (25) (anstelle von (32), welches vom Aufwand her günstiger ist, dafür aber aufgrund der vorangehenden Bemerkungen gewöhnlich gröbere Einschließungen liefert).

Wir untersuchen nun, wie sich das Verfahren (24), (25) verhält, wenn es auf einer Rechenanlage unter Verwendung eines Gleitpunktzahlensystems durchgeführt wird.

Es sei $b > 1$ die Basis des Zahlensystems, und t_1 sei die Mantissenlänge einer einfach genauen Gleitpunktzahl. Für die nachfolgende Diskussion machen wir die folgenden Annahmen (a) bis (c):

(a) Für zwei Maschinenintervalle $[a]$ und $[b]$ gilt

$$fl([a] * [b]) = [(1 - \varepsilon_1)\,([a] * [b])_1,\ (1 + \varepsilon_2)\,([a] * [b])_2] \tag{33}$$

wobei $* \in \{+, -, \times, /\}$,

$$[a] * [b] = [([a] * [b])_1,\ ([a] * [b])_2]\,, \qquad |\varepsilon_1|, |\varepsilon_2| \leqq \varepsilon = b^{1-t_1}\,.$$

$fl(.)$ bezeichnet das Ergebnis einer Maschinenintervalloperation. (33) besagt, daß die untere Schranke $([a] * [b])_1$ des exakten Ergebnisses nach unten zur nächsten Maschinenzahl gerundet wird (falls überhaupt gerundet werden muß). Das Analoge gilt für die obere Schranke. Man beachte, daß (33) und auch die nachfolgende Gleichung (34) im Unterlaufbereich nicht gelten können.

(b) Wir setzen voraus, daß das von U. KULISCH eingeführte sogenannte *genaue Skalarprodukt* zur Verfügung steht. (Siehe dazu [3], [4]):

Für zwei Intervallvektoren $[x] = ([x]_i)$ und $[y] = ([y]_i)$, die Maschinenintervalle als Komponenten haben, gilt

$$fl\,(\sum_{i=1}^{n} [x]_i \cdot [y]_i) = [(1 - \varepsilon_1)\,\sigma_1,\ (1 + \varepsilon_2)\,\sigma_2]\,, \tag{34}$$

wobei $\sum_{i=1}^{n} [x]_i \cdot [y]_i = [\sigma_1, \sigma_2]$ das exakte Skalarprodukt ist und $|\varepsilon_1|, |\varepsilon_2| \leqq b^{1-t_1}$ gilt. Siehe dazu die Bemerkungen im Anschluß an (33).

Im Normalfall ist die Genauigkeit des sogenannten genauen Skalarproduktes vergleichbar mit der doppelt genauen Akkumulation der doppelt langen Produkte einfach langer Gleitpunktintervalle und anschließender Rundung auf einfache Genauigkeit. Falls jedoch starke Auslöschung auftritt (was gewöhnliche bei der Berechnung von Residuen der Fall ist), so ist die Genauigkeit wesentlich höher. Aus (34) folgt

$$fl\,(\sum_{i=1}^{n} [x]_i \cdot [y]_i) \subseteq \sum_{i=1}^{n} [x]_i \cdot [y]_i + \varepsilon[-1, 1] \cdot |\sum_{i=1}^{n} [x]_i \cdot [y]_i| \tag{35}$$

mit $\varepsilon = b^{1-t_1}$.

(c) Wir setzen voraus, daß die gegebene Matrix A exakt auf der Maschine darstellbar ist. Das gleiche gelte für die Näherungen λ und x. Wir erhalten die Matrix B in (18) aus A, indem wir λ von den Diagonalelementen subtrahieren und die s-te Spalte durch $-x$ ersetzen. Die auf der Maschine – im Folgenden bedeuten überstrichene Größen immer auf der Maschine exakt dargestellte Vektoren, Matrizen, Intervallvektoren etc. – berechnete Matrix $\overline{B}$ unterscheidet sich daher von B höchstens in den Diagonalelementen. Wir setzen weiter voraus, daß eine auf der Maschine darstellbare Intervallmatrix $\overline{[BI]}$ bekannt ist, mit

$$B^{-1} \in \overline{[BI]} \subseteq B^{-1} + \varepsilon \cdot [-1, 1] \cdot |B^{-1}| + [-1, 1] \cdot \overline{E}\,, \tag{36}$$

wobei $\varepsilon = b^{1-t_1}$ ist. Wenn B^{-1} bezüglich kleiner Änderungen in der Diagonale von B gut konditioniert ist, so ist $\overline{E}$ eine (nichtnegative) Maschinenmatrix mit kleinen Elementen. Falls sogar $B = \overline{B}$ gilt, so ist $\overline{E} = 0$. (36) besagt dann, daß die Inverse von B auf der Maschine in eine Intervallmatrix $\overline{[BI]}$ eingeschlossen ist, deren Elemente Schranken besitzen, die sich um höchstens zwei benachbarte Maschinenzahlen unterscheiden. Unter Verwendung des genauen Skalarproduktes kann eine solch enge Einschließung der Inversen auch für bezüglich Inversion schlecht konditionierte Matrizen B berechnet werden. Siehe dazu [5]. Falls B nicht exakt darstellbar ist ($B \neq \overline{B}$), dann kann man unter Verwendung des genauen Skalarproduktes eine Intervallmatrix $\overline{[BI]}$ berechnen, für die (36) mit $||\overline{E}|| \approx \varepsilon \cdot ||B^{-1}|| \cdot ||D_B||$, $\varepsilon = b^{1-t_1}$, gilt. D_B bezeichnet dabei die Diagonale von B. Siehe [5].

Wir bezeichnen mit $\overline{[y]}^k$ die mit dem Verfahren (24), (25) tatsächlich berechneten Iterierten. Wir setzen außerdem

$$\overline{[\hat{z}]}^k := r + \overline{[y]}_s^k\,\overline{\widetilde{[y]}}^k\,,$$

wobei „~" die gleiche Bedeutung wie in Abschnitt 1 hat. $\overline{[\hat{z}]}^k$ wird komponentenweise unter Verwendung des genauen Skalarproduktes berechnet. Dabei werden wegen $r = \lambda x - Ax$ Skalarprodukte mit Intervallvektoren der Länge $n + 2$ berechnet. $\overline{[y]}^{k+1}$ wird durch

$$\overline{[y]}^{k+1} = fl(\overline{[BI]} \cdot \overline{[\hat{z}]}^k)$$

berechnet. Die rechte Seite wird wiederum durch Verwendung des genauen Skalarproduktes berechnet. Wegen (35) gilt

$$\overline{[\hat{z}]}^k \subseteq [\hat{z}]^k + \varepsilon[-1, 1] \cdot |[\hat{z}]^k|\,.$$

Mit (35) und (36) folgt daher für die tatsächlich berechneten Iterierten

$$\begin{aligned}
\overline{[y]}^{k+1} &= fl(\overline{[BI]} \cdot \overline{[\hat{z}]}^k) \subseteq \\
&\subseteq fl\big((B^{-1} + \varepsilon[-1, 1] \cdot |B^{-1}| + [-1, 1] \cdot \overline{E}) \cdot ([\hat{z}]^k + \varepsilon[-1, 1] \cdot |[\hat{z}]^k|)\big) \subseteq \\
&\subseteq (B^{-1} + \varepsilon[-1, 1] \cdot |B^{-1}| + [-1, 1] \cdot \overline{E})\,([\hat{z}]^k + \varepsilon[-1, 1] \cdot [\hat{z}]^k) + \\
&\qquad + \varepsilon \cdot [-1, 1] \cdot |B^{-1} + \varepsilon \cdot [-1, 1] \cdot |B^{-1}| + [-1, 1] \cdot \overline{E}|\,|[\hat{z}]^k + \varepsilon \cdot [-1, 1] \cdot |[\hat{z}]^k|| \subseteq \\
&\subseteq B^{-1} \cdot [\hat{z}]^k + [-1, 1] \cdot (3\varepsilon + 3\varepsilon^2 + \varepsilon^3) \cdot |B^{-1}| \cdot |[\hat{z}]^k| + \\
&\qquad + [-1, 1] \cdot \varepsilon \cdot (1 + \varepsilon) \cdot \overline{E} \cdot |[\hat{z}]^k| + [-1, 1] \cdot (1 + \varepsilon) \cdot \overline{E} \cdot |[\hat{z}]^k|\,.
\end{aligned}$$

Damit erhält man für die Durchmesser der tatsächlich berechneten Iterierten $\overline{[y]}^{k+1}$

$$d\overline{[y]}^{k+1} \leqq |B^{-1}| \cdot d[\hat{z}]^k + 2(3\varepsilon + 3\varepsilon^2 + \varepsilon^3) \cdot |B^{-1}| \cdot |[\hat{z}]^k| + 2(1+\varepsilon)^2 \cdot \overline{E} \cdot |[\hat{z}]^k| .$$

Wir setzen jetzt (ohne wesentliche Einschränkung) voraus, daß der nach Satz 3 berechnete Intervallvektor $[y]^0 = [-\beta_1, \beta_1]\, e$ auf der Maschine exakt darstellbar ist. Es gilt also dann $\overline{[y]}^0 = [y]^0$. (Falls dies nicht der Fall ist, so genügt es, bei der Berechnung von β_1 nach (23) alle Rechenoperationen „nach oben" zu runden.) Durch Durchschnittsbildung im Verfahren (24), (25) nach jedem Iterationsschritt kann man erreichen, daß alle Iterierten in $\overline{[y]}^0 = [y]^0$ liegen. Dann erhalten wir mit

$$\overline{d}_k = ||d\overline{[y]}^k||_\infty , \qquad \delta = 2\beta_1 l = 1 - \sqrt{1 - 4\varrho l} < 1 ,$$
$$\varepsilon = b^{1-t_1} , \qquad \tilde{\varepsilon} = ||\overline{E}||_\infty ,$$
$$s = 2(3 + 3\varepsilon + \varepsilon^2) \cdot ||B^{-1}||_\infty \cdot \big\||[\hat{z}]^0|\big\|_\infty , \qquad \tilde{s} = 2(1+\varepsilon)^2 \cdot \big\||[\hat{z}]^0|\big\|_\infty$$

die Ungleichung

$$\overline{d}_{k+1} \leqq \delta \overline{d}_k + \varepsilon \cdot s + \tilde{\varepsilon} \cdot \tilde{s} . \tag{37}$$

Für die Herleitung von (37) bemerken wir nur, daß vollständig analog wie beim Beweis von Satz 1.3

$$|B^{-1}|\, d[\hat{z}]^k \leqq \overline{d}_k \cdot 2 \cdot \beta_1 \cdot |B^{-1}| \cdot e$$

folgt.

Mit $\tau = \varepsilon s + \tilde{\varepsilon}\tilde{s}$ bekommen wir aus (37) durch vollständige Induktion

$$\overline{d}_{k+1} \leqq \delta^{k+1} \overline{d}_0 + (\delta^k + \delta^{k-1} + \ldots + \delta + 1)\, \tau$$

und daher

$$\overline{d}_{k+1} \leqq \delta^{k+1} \overline{d}_0 + \frac{\tau}{1-\delta} \quad \text{oder} \quad \overline{d}_{k+1} \leqq \delta^{k+1} \overline{d}_0 + \frac{\varepsilon s + \tilde{\varepsilon}\tilde{s}}{1-\delta} .$$

Diese Ungleichung kann man folgendermaßen interpretieren: Da $\delta < 1$ vorausgesetzt ist, geht der erste Summand für $k \to \infty$ gegen Null. Die auf der Maschine erzielbare Genauigkeit wird daher durch den zweiten Summanden bestimmt. Für hinreichend gute Näherungen λ und x und falls $||B^{-1}||_\infty$ nicht zu groß ist, ist $\delta \ll 1$ (siehe dazu die Definition von δ in Abhängigkeit von ϱ und l). Daher ist für solche Startwerte der Nenner des zweiten Summanden nicht wesentlich kleiner als Eins. Unter den gleichen Voraussetzungen besitzt auch $|[\hat{z}]^0|$ kleine Komponenten. Falls die Inverse von B gut konditioniert ist bezüglich Änderungen in der Diagonale von B, welche relativ die Größenordnung von $\varepsilon = b^{1-t_1}$ besitzen, so ist auch $\tilde{\varepsilon}$ klein. Unter den genannten Voraussetzungen erhält man daher kleine Durchmesser für den auf der Maschine berechneten Intervallvektor.

Diese Diskussion erfordert zwei Bemerkungen.

(I) Die Annahme, daß $||B^{-1}||_\infty$ klein, oder genauer, nicht zu groß ist, macht nur Sinn, wenn $||B||_\infty$ einen festen Wert besitzt, z. B. $||B||_\infty = 1$. Andernfalls kann man für $||B^{-1}||_\infty$ jeden positiven Wert dadurch erreichen, daß man B mit einem geeigneten Faktor multipliziert.

(II) Wegen

$$\overline{B} = B - D_B + \big(I + \operatorname{diag}(\varepsilon_i)\big) \cdot D_B , \qquad |\varepsilon_i| \leqq \varepsilon = b^{1-t_1} , \qquad 1 \leqq i \leqq n ,$$

gilt $\overline{B} = B + D_B \cdot \operatorname{diag}(\varepsilon_i)$ und daher

$$\overline{B}^{-1} = \big(I + B^{-1} \cdot D_B \cdot \operatorname{diag}(\varepsilon_i)\big)^{-1} \cdot B^{-1} .$$

Ist daher $||B^{-1}||_\infty$ nicht zu groß, so ist B^{-1} bezüglich kleiner Änderungen in der Diagonale von B gut konditioniert.

Zusammenfassend haben wir somit das folgende

Ergebnis: *Für hinreichend gute Näherungen λ und x für ein Eigenpaar und für $||B||_\infty = 1$ kann man bei nicht zu großem $||B^{-1}||_\infty$ mit Hilfe des Verfahrens* (24), (25) *einen Intervallvektor auf der Maschine berechnen, dessen Schranken den Fixpunkt y^* von* (9) *(und damit das Eigenpaar) eng einschließen.*

3. Numerische Beispiele

a) Wir betrachten die unsymmetrische (5,5)-Matrix

$$A = \begin{pmatrix} 15 & 11 & 6 & -9 & -15 \\ 1 & 3 & 9 & -3 & -8 \\ 7 & 6 & 6 & -3 & -11 \\ 7 & 7 & 5 & -3 & -11 \\ 17 & 12 & 5 & -10 & -16 \end{pmatrix} .$$

Siehe [9], Beispiel 5.11. Sie besitzt den (algebraisch) einfachen Eigenwert $\lambda = -1$ mit

$$(13, 22, 19, 16, 28)^T$$

als (einen) dazugehörigen Eigenvektor. Die anderen Eigenpaare sind in [9] angegeben. Als Näherung für λ wählen wir -0.99999999, als Näherung für den Eigenvektor

$$(13.000001, 21.999999, 18.999999, 16.000001, 27.999999)^T .$$

Nach dem Einlesen dieser Näherungen wird die Näherung für den Eigenvektor auf Unendlichnorm Eins normiert. Dies ergibt den Vektor

$$\begin{pmatrix} 0.464285766582 \\ 0.785714278061 \\ 0.678571417092 \\ 0.571428627551 \\ 1 \end{pmatrix},$$

der als Näherung für den Eigenvektor auf der Maschine exakt vorliegt. (Der verwendete Computer rechnet im Dezimalsystem. Siehe unten.) Mit diesem Vektor und der obigen Eigenwertnäherung ergibt sich für β_1 nach (23)

$$\beta_1 = 0.561224459410 \times 10^{-7} .$$

Nach einem Iterationsschritt mit dem Verfahren (24), (25) erhalten wir die folgenden Einschließungen für den Eigenwert:

$$[-1.00000000001; \quad -0.999999999999]$$

bzw. für die Komponenten des Eigenvektors:

$$\begin{aligned} &[0.464285714285; \quad 0.464285714286] \\ &[0.785714285714; \quad 0.785714285715] \\ &[0.678571428571; \quad 0.678571428572] \\ &[0.571428571428; \quad 0.571428571429] \\ &[1; \qquad\qquad 1 \qquad\qquad] . \end{aligned}$$

Zum Vergleich geben wir noch die Ergebnisse an, die man erhält, wenn man das Gleichungssystem (6) iterativ, d. h. durch „normale Gleitpunktrechnung" nach der Vorschrift

$$By^{k+1} = r + y_s^k \tilde{y}^k , \qquad k = 0, 1, 2, \ldots , \qquad y^0 \in \mathbb{R}^n ,$$

löst. Mit den gleichen Näherungen für den Eigenwert bzw. für die Komponenten des dazugehörigen Eigenvektors erhält man nach einem Iterationsschritt als Näherung für den Eigenwert

$$-0.100000004231 \times 10^1$$

und für die Komponenten des Eigenvektors

$$\begin{aligned} &0.464285714273 \\ &0.785714285723 \\ &0.678571428573 \\ &0.571428571406 \\ &1 \qquad\qquad . \end{aligned}$$

Diese Werte verändern sich in den nachfolgenden Schritten nicht mehr. Mit den exakten Werten (welche durch die oben berechneten Intervalle eingeschlossen sind) stimmen nur acht bis elf Ziffern in der Mantisse überein.

b) Wir betrachten die unsymmetrische (5,5)-Matrix

$$A = \begin{pmatrix} -4 & -9 & 6 & 4 & 2 \\ -9 & -4 & -3 & -2 & -1 \\ -2 & -2 & 0 & -1 & -1 \\ 3 & 3 & 3 & 5 & 3 \\ -9 & -9 & -9 & -9 & -4 \end{pmatrix}$$

aus [8]. Sie besitzt die Eigenwerte $\lambda_1 = 5$, $\lambda_2 = \lambda_3 = 2$, $\lambda_4 = 1 + i\sqrt{2}$, $\lambda_5 = 1 - i\sqrt{2}$. Ein auf Euklidische Länge Eins normierter zu $\lambda_1 = 5$ gehöriger Eigenvektor ist

$$x^T = \frac{1}{\sqrt{2}}(1, -1, 0, 0, 0) .$$

Wir wählen als Näherung für den Eigenwert λ_1 die Zahl

$$4.9999957$$

und als Näherung für den dazugehörigen Eigenvektor den Vektor

$$\begin{pmatrix} -9.999989817\,67 \times 10^{-1} \\ 1 \\ 5.475\,909\,243\,19 \times 10^{-7} \\ -4.706\,321\,091\,58 \times 10^{-7} \\ -7.185\,457\,850\,60 \times 10^{-7} \end{pmatrix} .$$

Damit erhalten wir für β_1 nach (23)

$$\beta_1 = 4.300045759\,67 \times 10^{-6}\,.$$

Nach 3 Iterationsschritten mit dem Verfahren (24), (25) erhalten wir die folgenden Einschließungen für den Eigenwert:

$$[4.999999999\,99;\quad 5.000000000\,01]$$

bzw. für die Komponenten des Eigenvektors:

$$\begin{aligned}
&[-1.000000000\,01;\quad -9.999999999\,99 \times 10^{-1}]\\
&[\ \ 1;\qquad\qquad\qquad\ \ 1]\\
&[-5.0 \times 10^{-18};\quad 6.0 \times 10^{-18}]\,.\\
&[-5.0 \times 10^{-18};\quad 6.0 \times 10^{-18}]\\
&[-1.3 \times 10^{-17};\quad 7.0 \times 10^{-18}]\,.
\end{aligned}$$

Im Gegensatz zu Beispiel a), in welchem sich die Mantissen der oberen und unteren Schranken jeweils nur um eine Einheit der letzten Stelle unterscheiden, ist dies für die drei letzten Komponenten der Eigenvektoreinschließung in Beispiel b) nicht der Fall. (Hier stimmen nicht einmal die Vorzeichen überein). Dies ist insofern nicht verwunderlich, als die exakten Werte dieser Komponenten gleich Null sind und im Unterlaufbereich die Voraussetzungen (33) und (34) im allgemeinen nicht gelten.

In den Fällen, in welchen wie in diesem Beispiel ein mit endlich vielen Schritten nach (24), (25) berechneter Intervallvektor $[y] = ([y]_i)$ in (mindestens) einer Komponente die Null enthält (Unterlaufbereich!) kann man häufig die Schranken noch durch folgenden einfachen Trick wesentlich verbessern: Man bildet den reellen Vektor $z = (z_i)$ mit

$$z_i = \begin{cases} 0 & \text{falls} \quad 0 \in [y]_i \\ m[y]_i & \text{sonst}\,. \end{cases}$$

Dabei bezeichnet $m[y]_i$ den Mittelpunkt des Intervalles $[y]_i$. Mit diesem Vektor z wird nun mit $[z]^0 := z$ anstelle von (24) das Verfahren

$$[z]^{k+1} = g([z]^k) \cup [z]^k\,, \qquad k = 0, 1, \ldots\,,$$

durchgeführt. Dabei bezeichnet „$\cup$" die konvexe Vereinigung der Intervallvektoren $g([z]^k)$ und $[z]^k$. Man erhält so eine Folge $\{[z]^k\}_{k=0}^{\infty}$ von Intervallvektoren mit

$$[z]^0 \subseteq [z]^1 \subseteq \ldots \subseteq [z]^k \subseteq [z]^{k+1} \subseteq \ldots$$

Aufgrund der Teilmengeneigenschaft liegen alle $[z]^k$ in $[y]^0 = [-\beta_1, \beta_1]\,e$. Daher muß es einen Index k geben mit $[z]^k = [z]^{k+1}$. Beginnend mit $[y]^0 = [z]^k$ wird dann (der zweite Teil von) (24) durchgeführt. Sehr häufig ist $k = 0$ oder $k = 1$.

In unserem Beispiel erhält man (auf der Maschine exakt) $z = (-1, 1, 0, 0, 0)^T$, und nach einem Schritt des beschriebenen Vorgehens erhält man eine Einschließung, für welche in allen Komponenten die Mantissen der Ober- und Unterschranken übereinstimmen, also das exakte Eigenpaar.

Die Beispiele wurden auf einem IBM-PC gerechnet. In PASCAL SC stehen dabei für Gleitpunktzahlen 12 Dezimalstellen in der Mantisse zur Verfügung.

Literatur

1 Alefeld, G.; Herzberger, J., Introduction to Interval Computations, Academic Press 1983.
2 Alefeld, G.; Platzöder, L.: A quadratically convergent Krawczyk-like algorithm, SIAM J. Numer. Anal. **20** (1983), 210–219.
3 Kulisch, U., Grundlagen des Numerischen Rechnens, Mathematische Begründung der Rechnerarithmetik, Bibliographisches Institut Mannheim 1976.
4 Kulisch, U.; Miranker, W. (Eds.), A New Approach to Scientific Computation, Academic Press 1983.
5 Rump, S., Solving algebraic problems with high accuracy, In [4], pp. 53–120.
6 Symm, H. J.; Wilkinson, J. H., Realistic error bounds for a simple eigenvalue and its associated eigenvector, Numer. Math. **35** (1980), 113–126.
7 Wilkinson, J. H.; Reinsch, C., Handbook for Automatic Computation. Volume 2: Linear Algebra, Springer Verlag, Berlin 1971.
8 Yamamoto, T., Error bounds for computed eigenvalues and eigenvectors, Numer. Math. **34** (1980), 189–199.
9 Gregory, R. T.; Karney, D. L., A Collection of Matrices for Testing Computational Algorithms, Wiley-Interscience, New York 1969.

Anschrift: Prof. Dr. G. Alefeld, Institut für Angewandte Mathematik, Universität Karlsruhe, Kaiserstraße 12, D-7500 Karlsruhe, BRD