Beiträge zur Numerischen Mathematik 8

Beiträge zur Numerischen Mathematik 8

Herausgegeben von
Frieder Kuhnert und Jochen W. Schmidt

R. Oldenbourg Verlag München Wien 1980

CIP-Kurztitelaufnahme der Deutschen Bibliothek

Beiträge zur Numerischen Mathematik / hrsg.
von Frieder Kuhnert u. Jochen W.
Schmidt. — München, Wien : Oldenbourg.

NE: Kuhnert, Frieder [Hrsg.]

8. — 1980.
ISBN 3-486-22941-9

Printed in the German Democratic Republic
Gesamtherstellung: VEB Druckhaus „Maxim Gorki", Altenburg
ISBN: 3-486-22941-9

Inhalt

Beiträge zur Numerischen Mathematik
8 (1980), 7–13

Über die Lösungen der Gleichung $\varrho^{(P)} = \|P\|_z$

HANS-JÜRGEN ALBRAND

1. Einleitung

Es bezeichne $P = (p_{ij})$ eine reelle quadratische $n \times n$ Matrix, $\varrho(P)$ den Spektralradius von P und $\|P\|_z$ die Zeilensummennorm von P, d. h.

$$\|P\|_z := \max_{1 \leq i \leq n} \sum_{j=1}^{n} |p_{ij}|.$$

Für jede Matrixnorm gilt bekanntlich die Ungleichung

$$\varrho(P) \leqq \|P\|. \tag{1}$$

Für symmetrisches und positiv definites P gilt in Verbindung mit der Spektralnorm in (1) das Gleichheitszeichen.

Das Anliegen der Arbeit ist, eine notwendige und hinreichende Bedingung für die Lösungen P der Gleichung

$$\varrho(P) = \|P\|_z \tag{2}$$

zu geben. Es bezeichne L die Lösungsmenge von (2), d. h.

$$L := \bigl(P : \varrho(P) = \|P\|_z\bigr).$$

Die notwendige und hinreichende Bedingung wird in Form einer Streichungsregel gegeben. Sie entscheidet darüber, ob eine vorgegebene Matrix P der Menge L angehört oder nicht. Die Streichungsregel ist insofern von einfacher Struktur, als sie nur mit den Vorzeichen der Elemente von P operiert. Die gefundene Charakterisierung beantwortet damit auch vollständig die Frage nach der (theoretischen) Konvergenz der Iterationsvorschrift

$$x^{s+1} := Px^s + v, \qquad s = 0, 1, 2, \ldots,$$

bei beliebig vorgegebenem Startvektor x^0 und $\|P\|_z = 1$ oder $\|P^{\mathsf{T}}\|_z = 1$. Insbesondere enthält die zu gebende notwendige und hinreichende Bedingung eine weitere Abschwächung der bekannten schwachen Zeilen- und Spaltensummenkriterien, und zwar so, daß eine weitere Abschwächung nicht mehr möglich ist.

2. Eine notwendige und hinreichende Charakterisierung der Lösungen

In diesem Abschnitt wird die zu gebende notwendige und hinreichende Bedingung hergeleitet, die sämtliche Matrizen P aus L charakterisiert. O. B. d. A. sei $\|P\|_z = 1$ vorausgesetzt. Es bezeichnen $z_1^{(k)}, \ldots, z_n^{(k)}$ die Zeilenvektoren von P^k $(:= PP^{k-1})$, $k = 1, 2, \ldots$ $N^{(m)}$, $m = 0, 1, 2, \ldots$ seien Indexmengen. $V_0(0) := V_0(P)$ sei die Vorzeichenmatrix von P, d. h., beispielsweise mit $P = \begin{pmatrix} -2 & 3 \\ 0 & 4 \end{pmatrix}$ wird $V_0(0) = \begin{pmatrix} - & + \\ 0 & + \end{pmatrix}$. $V_{m+1}(k) := \left(v_{ij}^{(m+1)}(k)\right)$ sei die 2^k-te Potenz der Matrix, die aus $V_m(k_m)$ durch Streichung der i-ten Zeilen und Spalten, i aus $N^{(m)}$, hervorgeht, wobei die k_m natürliche Zahlen sind, $k_0 = 0$ gesetzt wird und $m, k = 0, 1, 2, \ldots$ ist. Die Indizierung der Zeilen und Spalten soll erhalten bleiben. Es sei bemerkt, daß die Potenzen der Vorzeichenmatrizen nur dann gebildet werden, wenn die entsprechende Multiplikation zweier Matrizen in folgender Weise einen Sinn hat: Die Skalarprodukte, die in der üblichen Weise gebildet werden und die die Elemente der Produktmatrix ergeben, lassen sich eindeutig mit $-$, $+$ oder 0 identifizieren. Bei $V_0(0)$ im obigen Beispiel wäre dieser Sachverhalt nicht gegeben. Aber z. B. bei

$$V := \begin{pmatrix} 0 & - & + \\ - & 0 & + \\ + & + & - \end{pmatrix}.$$

Hier ergibt sich V^2 zu $V^2 = \begin{pmatrix} + & + & - \\ + & + & - \\ - & - & + \end{pmatrix}$. Mit $V_0(0) := \begin{pmatrix} - & + & 0 \\ 0 & - & 0 \\ - & 0 & - \end{pmatrix}$ und $N^{(0)} = (3)$ wird $V_1(k) = \begin{pmatrix} + & - \\ 0 & + \end{pmatrix}$ für $k = 1, 2, 3, \ldots$ Nach diesen Vorbereitungen kommen wir zur Formulierung der Streichungsregel. Diese Regel arbeitet bis auf die Bestimmung von $N_0^{(0)}$, was trivial ist, ausschließlich mit den Vorzeichen der Elemente von P.

Streichungsregel

I. $$N_0^{(0)} := \left(i : \sum_{j=1}^{n} |p_{ij}| < \|P\|_z\right),$$

$$N_{g+1}^{(0)} := \left(i : i \notin \bigcup_{j=0}^{g} N_j^{(0)}, \quad s \in \bigcup_{j=0}^{g} N_j^{(0)}, \quad v_{is}^{(0)}(0) \neq 0\right),$$

$$g = 0, 1, \ldots, g_0 + 1, \quad \text{wobei} \quad N_{g_0+1}^{(0)} = \emptyset, \quad N_g^{(0)} \neq \emptyset \text{ für } g \leqq g_0.$$

$$N^{(0)} := \bigcup_{j=0}^{g_0} N_j^{(0)}, \quad \text{falls} \quad g_0 \geqq 0; \quad N^{(0)} = \emptyset, \quad \text{falls } N_0^{(0)} = \emptyset.$$

II. Falls $N^{(0)} \neq (1, 2, \ldots, n)$ ist, werden für $m = 1, 2, \ldots, m_0$, $m_0 \leqq n$, Indexmengen $N^{(m)}$ auf folgende Weise konstruiert: k_m sei die kleinste natürliche

Zahl mit

$$N_0^{(m)}(k_m) := \Big(i : i \notin \bigcup_{j=0}^{m-1} N^{(j)}, \quad \exists\, r(i), s(i), h(i) :$$
$$v_{i,r(i)}^{(m)}(k_m)\, v_{r(i),h(i)}^{(m)}(k_m) = +,$$
$$v_{i,s(i)}^{(m)}(k_m)\, v_{s(i),h(i)}^{(m)}(k_m) = -\Big) \neq \emptyset.$$

Wir setzen $\overline{N}_g^{(m)} := \left(\bigcup_{j=0}^{m-1} N^{(j)}\right) \cup \left(\bigcup_{j=0}^{g} N_j^{(m)}\right)$,

$$N_{g+1}^{(m)} := \left(i : i \notin \overline{N}_g^{(m)},\ s \in \overline{N}_g^{(m)},\ v_{is}^{(m)}(k_m) \neq 0\right),$$

$g = 0, 1, \ldots, g_m + 1$, wobei $N_{g_m+1}^{(m)} = \emptyset$ und $N_g^{(0)} \neq \emptyset$ für $g \leqq g_m$;

$$N^{(m)} := \bigcup_{j=0}^{g_m} N_j^{(m)}, \text{ falls } g_m \geqq 0;\ N^{(m)} := \emptyset \text{ sonst}.$$

Bemerkung 1. Nach dieser Regel werden nacheinander die i-ten Zeilen und Spalten von $V_0(P)$ mit i aus $N^{(0)}, N^{(1)}, \ldots, N^{(m_0)}$ gestrichen. Das Verfahren bricht bei einem gewissen $m_0 \leqq n$ ab, wenn alle Zeilen und Spalten gestrichen sind oder $N^{(m_0)} = \emptyset$ ist für ein $m_0 \geqq 1$.

Bemerkung 2. Der Fall $N^{(m_0)} = \emptyset$ tritt genau dann ein, wenn $N_0^{(m_0)}(k) = \emptyset$ ist für $k = 1, 2, \ldots$ Da es nur endlich viele verschiedene Vorzeichenmatrizen $V_{m_0}(k) := V_{m_0}^{2^k}(0)$ gibt, existieren natürliche Zahlen $t(m_0)$, $\bar{t}(m_0)$ mit $t(m_0) < \bar{t}(m_0)$ und

$$v_{ij}^{(m_0)}\big(t(m_0)\big) = v_{ij}^{(m_0)}\big(\bar{t}(m_0)\big) \tag{3}$$

für i, j aus $(1, 2, \ldots, n) \setminus \bigcup_{j=0}^{m_0-1} N^{(j)}$, so daß aus (3) und $N_0^{(m_0)}(k) = \emptyset$, $k = 1, 2, \ldots, \bar{t}(m_0)$, die Beziehung $N_0^{(m_0)}(k) = \emptyset$ für $k > \bar{t}(m_0)$ folgt. Denn für $0 \leqq k \leqq \bar{t}(m_0) - t(m_0)$ gilt wegen

$$V_{m_0}^{2^{\bar{t}(m_0)+k}}(0) = V_{m_0}^{2^{t(m_0)+k}}(0)$$

die Beziehung $N_0^{(m_0)}\big(\bar{t}(m_0) + k\big) = \emptyset$, woraus in analoger Weise $N_0^{(m_0)}\big(2\bar{t}(m_0) + k\big) = \emptyset$ folgt für $0 \leqq k \leqq \bar{t}(m_0) - t(m_0)$, usw.

In Verbindung mit diesen Bemerkungen ergibt sich die

Folgerung 1. *Die Streichungsregel bricht nach endlich vielen Schritten ab. Für $P \geqq 0$ bzw. $P \leqq 0$ bricht sie bereits mit ihrem Teil I ab.*

$S(P)$ bezeichne die Menge der Indizes der nicht gestrichenen Zeilen, d. h.

$$S(P) := (1, 2, \ldots, n) \setminus \bigcup_{j=0}^{m_0} N^{(j)}.$$

Satz 1. *Eine reelle $n \times n$-Matrix P ist genau dann eine Lösung der Gleichung (2), wenn $S(P) \neq \emptyset$ ist.*

Beweis. Die Gleichung (2) ist äquivalent mit $\varrho(cP) = \|cP\|_z$, falls die Zahl $c \neq 0$ ist. Deshalb können wir o. B. d. A. $\|P\|_z = 1$ voraussetzen.

1. Es sei $\varrho(P) = \|P\|_z = 1$. Dann ist $S(P) \neq \emptyset$.

Für natürliche Zahlen k und l ist

$$\|z_i^{(k+l)}\|_1 := \sum_{j=1}^{n} |p_{ij}^{(k+l)}| \leqq \sum_{j=1}^{n} |p_{ij}^{(l)}| \, \|z_j^{(k)}\|_1 \leqq \|z_i^{(l)}\|_1 \leqq 1,$$

wobei $\sum_{j=1}^{n} |p_{ij}^{(l)}| \, \|z_j^{(k)}\|_1 < \|z_i^{(l)}\|_1$ genau dann gilt, wenn ein j_0 existiert mit $\|z_{j_0}^{(k)}\|_1 < 1$ und $p_{ij_0}^{(l)} \neq 0$ ist. Die Ungleichung

$$\|z_i^{(k+l)}\|_1 < \sum_{j=1}^{n} |p_{ij}^{(l)}| \, \|z_j^{(k)}\|_1$$

gilt genau dann, wenn Indizes $r(i)$, $s(i)$, $h(i)$ existieren mit

$$p_{i,r(i)}^{(l)} p_{r(i),h(i)}^{(k)} > 0, \qquad p_{i,s(i)}^{(l)} p_{s(i),h(i)}^{(k)} < 0.$$

Auf Grund dieser Ungleichungen wird $\|z_i^{(k)}\|_1 < 1$ für $k = g_0 + 1$ und i aus $N^{(0)}$, für $k = (2 + g_r)\, 2^{k_1 + \cdots + k_r}$ und i aus $N^{(r)}$, $r = 1, 2, \ldots, m_0$. Für jeden Index i aus $\bigcup_{j=0}^{m_0} N^{(j)}$ existiert also ein $k(i)$ mit $\|z_i^{(k(i))}\|_1 < 1$. Wegen $\varrho(P) = \lim \|P^k\|_z^{1/k} = 1$ (siehe z. B. [1]) muß es mindestens einen Index i_0 geben, für den

$$\|z_{i_0}^{(k)}\|_1 = 1, \qquad k = 1, 2, \ldots,$$

gilt, d. h., i_0 ist nicht aus $\bigcup_{j=0}^{m_0} N^{(j)}$. Folglich ist i_0 aus $S(P)$ und $S(P) \neq \emptyset$.

2. Es sei $S(P) = (i_1, \ldots, i_s) \neq \emptyset$. Dann gilt (2).

Wegen $N_0^{(m_0)}(k) = \emptyset, k = 1, 2, \ldots$, folgt $\|z_{i_t}^{(q)}\|_1 = 1$ für $q = 2^{k_1 + \cdots + k_{m_0} + k}, k = 1, 2, \ldots$, $t = 1, 2, \ldots, s$, und mit $\varrho(P) = \lim \|P^k\|_z^{1/k}$ ergibt sich $\varrho(P) = \|P\|_z = 1$. Damit ist der Satz 1 bewiesen.

Satz 2. *Es sei P eine reelle $n \times n$-Matrix, $S(P) = (i_1, \ldots, i_s) \neq \emptyset$ und $s < n$. Dann ist P zerlegbar.*

Beweis. Aus der Streichungsregel folgt $p_{i_r j} = 0$ für $j \neq i_1, \ldots, i_s$; $r = 1, 2, \ldots, s$. Dann gibt es aber eine Permutationsmatrix E, so daß

$$EPE^{\mathsf{T}} = \begin{pmatrix} P_{11} & P_{12} \\ 0 & P_{22} \end{pmatrix}$$

wird, wobei P_{11} und P_{22} quadratische Untermatrizen von P sind.

Die Umkehrung dieses Satzes gilt natürlich nicht, wovon man sich leicht anhand von Beispielen überzeugt.

Bemerkung 3. Ist P nicht zerlegbar und ist das schwache Zeilensummenkriterium erfüllt (d. h., $\sum_{j=1}^{n} |p_{ij}| < 1$ für mindestens ein i), so folgt $N^{(0)} = (1, 2, \ldots, n)$, d. h., $S(P) = \emptyset$ ergibt sich bereits aus dem Teil I der Streichungsregel.

3. Einige Folgerungen

Im folgenden werden weitere Aussagen über Matrizen P gemacht, die Lösung der Gleichung (2) sind. Zunächst wird eine wichtige Klasse von Matrizen P angegeben, für die $S(P) \neq \emptyset$ ist.

Wir betrachten einen Vorzeichenvektor $v := (v_1, \ldots, v_n)^{\mathsf{T}}$ mit v_r gleich $+$ oder $-$. Es bezeichne $\operatorname{sgn} s_r$ den Vektor, dessen Komponenten die Vorzeichen der Spalte s_r der zu konstruierenden Matrix P sind. Wir setzen $\operatorname{sgn} s_r := v_1 v_r v$ für $r = 1, 2, \ldots, n$. In der so erhaltenen Vorzeichenmatrix, z. B.

$$\begin{pmatrix} + & - & - & + \\ - & + & + & - \\ - & + & + & - \\ + & - & - & + \end{pmatrix}$$

(hier ist $v = (+, -, -, +)^{\mathsf{T}}$), kann man an die Stelle der Plus- und Minuszeichen (bis auf jeweils ein Zeichen in jeder Zeile) auch eine Null setzen, z. B.

$$\begin{pmatrix} 0 & - & - & + \\ - & + & 0 & - \\ 0 & 0 & + & - \\ + & - & - & + \end{pmatrix}.$$

In der so konstruierten Vorzeichenmatrix können nun an die Stellen der Pluszeichen beliebige positive und an die Stellen der Minuszeichen beliebige negative Zahlen eingesetzt werden, so daß $\|z_i\|_1 = \|P\|_z$ wird für $i = 1, 2, \ldots, n$. Für die so konstruierte Matrix P wird dann $N^{(0)} = \emptyset$, $N_0^{(1)}(k) = \emptyset$ für $k = 0, 1, 2, \ldots$ und folglich $S(P) \neq \emptyset$. Der Vektor v heißt erzeugender Vorzeichenvektor von P.

Für die folgende, etwas allgemeinere Struktur von P, gilt ebenfalls $S(P) \neq \emptyset$. Es sei $\|z_j\|_1 \leqq \|P\|_z$, $\|z_{r_i}\|_1 = \|P\|_z$, $i = 1, 2, \ldots, s$, die Matrix $(p_{r_i r_j})_{i,j=1,2,\ldots,s}$ habe einen erzeugenden Vorzeichenvektor v und $p_{r_i j} = 0$ für $j \neq r_1, \ldots, r_s$; $i = 1, 2, \ldots, s$. Unter diesen Voraussetzungen gilt $(r_1, \ldots, r_s) \subseteqq S(P)$.

Die hier angegebene Struktur von P ist hinreichend für $S(P) \neq \emptyset$, sie ist aber nicht notwendig, wie man an dem folgenden Beispiel sieht:

$$P = \begin{pmatrix} 0 & -1 & 0 \\ 1 & 0 & 0 \\ 1/2 & 1/2 & 0 \end{pmatrix}, \qquad V_0 = \begin{pmatrix} 0 & - & 0 \\ + & 0 & 0 \\ + & + & 0 \end{pmatrix}.$$

Für V_0 existiert kein erzeugender Vorzeichenvektor. Es ist aber $N^{(0)} = \emptyset$,

$$V_0^2 = \begin{pmatrix} - & 0 & 0 \\ 0 & - & 0 \\ + & - & 0 \end{pmatrix},$$

woraus $N_0^{(1)}(k) = \emptyset$ für $k = 0, 1, 2, \ldots$ folgt, da V_0^2 mit $v = (-, +, +)^{\mathsf{T}}$ einen erzeugenden Vorzeichenvektor besitzt. Damit wird $S(P) = (1, 2, 3)$.

Die nächsten Sätze sind eine Folgerung aus Satz 1.

Mit natürlichen Zahlen s und r sei $P^{s/r}$ eine r-te Wurzel der s-ten Potenz von P (siehe z. B. [2]).

Satz 3. *Es seien s und r natürliche Zahlen. Dann ist $S(P) \neq \emptyset$ genau dann, wenn*

$$\varrho(P^{s/r}) = \|P\|^{s/r}_z$$

gilt.

Beweis. Es sei $S(P) \neq \emptyset$. Dann wird $\varrho(P^{1/r}) = \|P\|_z^{1/r}$ und $\varrho(P^{s/r}) = \|P\|_z^{s/r}$. Gilt umgekehrt $\varrho(P^{s/r}) = \|P\|_z^{s/r}$, so folgt $\varrho(P^s) = \|P\|_z^s$, d. h., $\varrho(P) = \|P\|_z$ und also nach Satz 1 $S(P) \neq \emptyset$, w. z. z. w.

Bemerkung 4. Aus $S(P) \neq \emptyset$ folgt im allgemeinen nicht $S(P^{s/r}) \neq \emptyset$. Man kann aber $\varrho(P^{s/r})$ nach Satz 3 sofort angeben.

Satz 4. *Es seien $S(P) \neq \emptyset$, $S(Q) \neq \emptyset$. Dann gelten die Ungleichungen*

$$\varrho(P + Q) \leqq \varrho(P) + \varrho(Q), \qquad \varrho(PQ) \leqq \varrho(P)\,\varrho(Q).$$

Beweis. $\varrho(P + Q) \leqq \|P + Q\|_z \leqq \|P\|_z + \|Q\|_z = \varrho(P) + \varrho(Q)$, $\varrho(PQ) \leqq \|PQ\|_z \leqq \|P\|_z\,\|Q\|_z = \varrho(P)\,\varrho(Q)$, w. z. z. w.

Satz 5. *Es sei P eine reelle $n \times n$-Matrix. Das Iterationsverfahren*

$$x^{s+1} := Px^s + v, \qquad s = 0, 1, 2, \ldots,$$

konvergiert bei beliebigem Startvektor $x^{(0)}$ und $\|P\|_z \leqq 1$ (bzw. $\|P^{\mathsf{T}}\|_z \leqq 1$) *genau dann gegen die Lösung des Gleichungssystems* $(I - P)\,x = v$, *wenn* $S(P) = \emptyset$ (bzw. $S(P^{\mathsf{T}}) = \emptyset$) *ist.*

Beweis. Es ist $\varrho(P) = \varrho(P^{\mathsf{T}})$, und aus Satz 1 folgt in Verbindung mit $\varrho(P) \leqq \|P\|_z$ und $S(P) = \emptyset$ (bzw. $S(P^{\mathsf{T}}) = \emptyset$) die Ungleichung $\varrho(P) < 1$. Umgekehrt folgt aus $\varrho(P) < 1$ und $\|P\|_z = 1$ die Relation $S(P) = \emptyset$ (und Entsprechendes für P^{T}), w. z. z. w.

Mit dem Satz 5 ist das Konvergenzverhalten des in Frage stehenden Iterationsverfahrens im Fall des schwachen Zeilen- bzw. des schwachen Spaltensummenkriteriums und im Fall der weiteren Abschwächung (alle Betrags-Zeilen- bzw. alle Betrags-Spaltensummen gleich 1) theoretisch geklärt.

4. Einfache Beispiele zur Streichungsregel

a) Es sei $P = (p_{ij})$ eine 7×7-Matrix mit $p_{31} = p_{32} = p_{41} = p_{42} = p_{62} = 0$, $p_{12} > 0$, $p_{21} < 0$, $p_{22} > 0$; $p_{35}, p_{47}, p_{52}, p_{61}, p_{72} \neq 0$; $\|z_i\|_1 = \|P\|_z$ für $i = 1, 2, \ldots, 7$, d. h. $N_0^{(0)} = \emptyset$. Sonst sei p_{ij} beliebig:

$$\begin{pmatrix} . & + & . & . & . & . & . \\ - & + & . & . & . & . & . \\ 0 & 0 & . & . & \neq 0 & . & . \\ 0 & 0 & . & . & . & . & \neq 0 \\ . & \neq 0 & . & . & . & . & . \\ \neq 0 & 0 & . & . & . & . & . \\ . & \neq 0 & . & . & . & . & . \end{pmatrix}.$$

Dann ist $N_0^{(1)}(0) = (2)$, $N_1^{(1)} = (1, 5, 7)$, $N_2^{(1)} = (6, 3, 4)$. Nach der Streichungsregel werden also alle Zeilen von P gestrichen, d. h. $\varrho(P) < \|P\|_z$.

b) Es sei $\|z_i\|_1 = \|P\|_z$ für $i = 1, 2, \ldots, n$, $P = (p_{ij})$ vom Format $n \times n$ und $p_{ij} > 0$. Dann kann keine Zeile gestrichen werden, d. h. $\varrho(P) = \|P\|_z$. Wählt man dagegen ein $p_{i_0 j_0} < 0$, so wird $N_0^{(1)}(0) = (i_0)$ und $N_1^{(1)} = (1, \ldots, i_0 - 1, i_0 + 1, \ldots, n)$, d. h. $\varrho(P) < \|P\|_z$.

c) Für die Koeffizientenmatrix

$$P := \begin{pmatrix} 0 & -0.25 & 0.5 & -0.25 \\ 0.5 & 0 & -0.425 & 0.075 \\ 0.225 & 0.357 & 0 & -0.4 \\ -0.4 & 0.5 & -0.1 & 0 \end{pmatrix}$$

wird $N_0^{(0)} = \emptyset$, $N_0^{(1)}(0) = (1, 2, 3, 4)$, d. h. $\varrho(P) < 1$.

Literatur

[1] YOSIDA, K., Functional analysis, Springer-Verlag, Berlin—Göttingen—Heidelberg 1965.
[2] GANTMACHER, F. R., Matrizenrechnung I, VEB Deutscher Verlag der Wissenschaften, Berlin 1970.

Manuskripteingang: 6. 2. 1978

VERFASSER:
Dr. HANS-JÜRGEN ALBRAND, Sektion Mathematik der Wilhelm-Pieck-Universität Rostock

Beiträge zur Numerischen Mathematik
8 (1980), 15 – 30

Monotone Regula-Falsi-ähnliche Verfahren bei nichtkonvexen Operatorgleichungen

Götz Alefeld

1. Einleitung

Bei der numerischen Auflösung von nichtlinearen Gleichungen ist die sogenannte *Globalisierung* heute ein wesentlicher Bestandteil der neueren Entwicklungen. Nach J. W. Schmidt [15] versteht man unter Globalisierung das Bestreben, ursprünglich nur lokal konvergente Verfahren so zu erweitern, daß alle oder möglichst viele Vektoren konvergente Folgen liefern, sofern sie als Startvektoren genommen werden.

Neben Abstiegsverfahren und Einbettungsverfahren können verschiedene Verfahren mit Monotonieeigenschaft in gewissem Sinne als globalisierte Verfahren angesehen werden, da es bei ihnen gewöhnlich nicht erforderlich ist, daß die Startvektoren hinreichend nahe an einer Lösung liegen. In der Vergangenheit wurde beispielsweise wiederholt das Problem betrachtet, in halbgeordneten Räumen Aussagen über die Monotonie von Folgen zu gewinnen, welche mit dem Newton-Verfahren oder mit den entsprechenden Übertragungen der Regula falsi berechnet werden. Siehe dazu insbesondere Baluev [4], Collatz [5], Hofmann [7], Ortega-Rheinboldt [11], Schmidt [14, 16] und Voss [20]. Zur Erzielung solcher Aussagen muß unter anderem gewöhnlich vorausgesetzt werden, daß die zu lösende Operatorgleichung konvex ist, oder es werden Voraussetzungen gemacht, aus denen die Konvexität folgt.

In [1] wurde für Gleichungssysteme im R^n gezeigt, daß durch intervallmäßige Auswertung der Fréchet-Ableitung das Newton-Verfahren so modifiert werden kann, daß man auch ohne Konvexität quadratisch konvergente monotone Folgen erhält. Diese Verfahren wurden kürzlich von Mönch [10] weiter modifiziert und bezüglich ihres Wirkungsgrades untersucht. Weitgehend ohne Konvexität kommt auch J. W. Schmidt in [16] aus.

In der hier vorliegenden Arbeit geben wir nun zur Berechnung einer Lösung der Gleichung $F(x) = 0$ durch fortwährende monotone Einschließung ein quadratisch konvergentes Verfahren an, bei welchem in jedem Schritt ausschließlich Steigungen erster Ordnung, d. h. also nur Funktionswerte berechnet werden müssen. Sonst werden zur Durchführung des Iterationsverfahrens nur *feste* Schranken für die Steigungen zweiter Ordnung verwendet. Die gewöhnlich notwendige Konvexitätsvoraussetzung für den Operator F kann dann wieder entfallen.

In Abschnitt 2 werden zunächst die benötigten Hilfsmittel bereitgestellt. Anschließend werden einige Sätze über das betrachtete Verfahren bewiesen. In Abschnitt 4 befinden sich numerische Beispiele. Die in dieser Arbeit durchgeführten Überlegungen ermöglichen eine Verallgemeinerung der von Voss [20] angegebenen Verfahren.

2. Hilfsmittel

Wir setzen im folgenden voraus, daß B ein reeller Banachraum ist, der durch einen regulären Kegel K halbgeordnet ist. K ist dabei eine abgeschlossene Teilmenge von B mit den Eigenschaften: $x, y \in K \Rightarrow x + y \in K$; $x, -x \in K \Rightarrow x = 0$; $\lambda \geqq 0$, $x \in K \Rightarrow \lambda x \in K$. Durch „$x \leqq y \Leftrightarrow y - x \in K$" wird dann eine mit der linearen Struktur von B verträgliche Halbordnung in B erklärt. K heißt regulär, wenn jede monoton wachsende (oder fallende) und in der Halbordnung beschränkte Folge konvergent ist. Ein regulärer Kegel ist normal, d. h., die Norm ist in dem Sinne monoton, daß aus $0 \leqq x \leqq y$ folgt $\|x\| \leqq \alpha \|y\|$ mit einer von x und y unabhängigen Konstanten α. Mit $[y, z]$ bezeichnen wir die Menge $\{x \mid y \leqq x \leqq z\} \subset B$. Ein Operator F von B nach B heißt isoton, wenn aus $x \leqq y$ folgt $Fx \leqq Fy$. In der Menge der stetigen linearen Operatoren von B nach B führen wir eine Halbordnung ein durch „$A_1 \leqq A_2 \Leftrightarrow (A_2 - A_1)\, x \geqq 0$ für alle $x \geqq 0$". Ein linearer Operator G von B nach B heißt positiv, falls $G \geqq 0$ gilt. In der Menge der stetigen bilinearen Operatoren von $B \times B$ nach B führen wir eine Halbordnung ein durch „$M \leqq N \Leftrightarrow (N - M)\,(x, y) \geqq 0$ für alle $x, y \geqq 0$". Ein bilinearer Operator H heißt positiv, wenn $H \geqq 0$ gilt. Ein Operator F von B nach B heißt invers isoton, wenn aus $Fx \leqq Fy$ folgt $x \leqq y$. Ein invers isotoner Operator ist injektiv.

Die hier aufgezählten Begriffe und Eigenschaften halbgeordneter linearer Räume findet man im wesentlichen bei KRASNOSELSKI [8]. Ein stetiger linearer Operator $\delta F(u, v)$ von B nach B heißt Steigung erster Ordnung des Operators F von $D \subset B$ nach B, wenn

$$F(u) - F(v) = \delta F(u, v)\,(u - v), \qquad u, v \in D, \tag{1}$$

gilt. Siehe SCHMIDT [13].

Wir fordern darüber hinaus das Bestehen der Gleichungen

$$\begin{aligned} \delta F(u, v) - \delta F(u, w) &= \delta^2 F(u, v, w)\,(v - w), \\ \delta F(u, w) - \delta F(v, w) &= \delta^2 F(u, v, w)^*\,(u - v). \end{aligned} \tag{2}$$

Dabei seien die Steigungen zweiter Ordnung $\delta^2 F(u, v, w)$ und $\delta^2 F(u, v, w)^*$ bei festem u, v und w stetige bilineare Operatoren von $B \times B$ nach B.

Ein Operator F von $D \subset B$ nach B heißt konvex, wenn mit der Steigung $\delta F(u, v)$ für jedes $x \geqq 0$ auch $\delta F(u, v)\, x$ bezüglich u und v isoton ist. F heißt konkav, wenn $-F$ konvex ist. Siehe HOFMANN [7].

Wir benötigen die folgende bekannte Aussage von KANTOROWITSCH [9].

Lemma. *B sei durch einen regulären Kegel K halbgeordnet. H sei ein stetiger isotoner Operator von $[y, z] \subset B$ nach B. Es gelte $y \leqq H(y)$, $z \geqq H(z)$. Dann besitzt H einen Fixpunkt $x \in [y, z]$.* □

3. Verfahren und Sätze

Zur Auflösung der Gleichung $F(x) = 0$ betrachten wir die folgende Iterationsvorschrift:

$$\begin{aligned} &F(x_k) + \{\delta F(x_k, y_k) + R(y_k - x_k)\}\,(x_{k+1} - x_k) = 0, \\ &F(y_k) + \{\delta F(x_k, y_k) + R^*(y_k - x_k)\}\,(y_{k+1} - y_k) = 0, \end{aligned} \quad k = 0, 1, \ldots 2, \qquad \text{(V)}$$

Dabei seien R und R^* stetige bilineare Operatoren von $B \times B$ nach B.

Wir beweisen über das Verfahren (V) zunächst die folgenden Aussagen.

Satz 1. *Der Banachraum B besitze die im zweiten Abschnitt genannten Eigenschaften. F sei eine stetige Abbildung von $D = [x_0, y_0] \subset B$ nach B, die in D Steigungen besitzt, welche den Forderungen* (1) *und* (2) *genügen. Es gelte $F(x_0) \leqq 0$ und $F(y_0) \geqq 0$. Weiter seien zwei positive stetige bilineare Operatoren R und R^* von $B \times B$ nach B bekannt mit den Eigenschaften*

$$\begin{aligned} &-R \leqq \delta^2 F(u, v, w), \\ &\delta^2 F(u, v, w)^* \leqq R^* \end{aligned} \qquad (3)$$

für $u, v, w \in D$. Schließlich gebe es zwei stetige lineare Operatoren G_1 und G_2, die beide stetige positive Inverse G_1^{-1} und G_2^{-1} besitzen, und es gelte

$$\begin{aligned} &\delta F(u, v) + R(v - u) \leqq G_1, \\ &\delta F(u, v) + R^*(v - u) \leqq G_2 \end{aligned} \qquad (4)$$

für $x_0 \leqq u \leqq v \leqq y_0$.

Dann gelten die folgenden Aussagen für das Verfahren (V):

1. (V) *ist durchführbar, d. h., die beiden Gleichungen besitzen für jedes $k \geqq 0$ Lösungen x_{k+1} und y_{k+1}.*

2. *Es existieren die Grenzwerte $\lim_{k\to\infty} x_k = x^*$ und $\lim_{k\to\infty} y_k = y^*$. x^* und y^* sind Nullstellen von F.*

3. *Es besteht die monotone Einschließung*

$$x_0 \leqq \cdots \leqq x_k \leqq x_{k+1} \leqq \cdots \leqq x^* \leqq z^* \leqq y^* \leqq \cdots \leqq y_{k+1} \leqq y_k \leqq \cdots \leqq y_0$$

für alle Lösungen z^* von $F(x) = 0$ im Intervall $[x_0, y_0]$.

Beweis. Wir zeigen zunächst, daß F in $[x_0, y_0]$ eine Nullstelle z^* besitzt. Dazu betrachten wir den Operator T mit

$$T(z) = z - G_1^{-1} F(z)$$

auf dem Intervall $[x_0, y_0]$. T ist stetig. Es gilt

$$\begin{aligned} &T(x_0) = x_0 - G_1^{-1} F(x_0) \geqq x_0, \\ &T(y_0) = y_0 - G_1^{-1} F(y_0) \leqq y_0, \end{aligned}$$

sowie wegen (4) für $v \geqq u$

$$\begin{aligned} T(v) - T(u) &= G_1^{-1}\{F(u) - F(v) - G_1(u - v)\} \\ &= G_1^{-1}\{\delta F(u, v) - G_1\}(u - v) \geqq G_1^{-1}R(v - u)^2 \geqq 0 \end{aligned}$$

d. h., T ist isoton. Auf Grund des Lemmas aus Abschnitt 2 gibt es daher mindestens einen Fixpunkt z^* von T, d. h. mindestens eine Lösung z^* von $F(x) = 0$ im Intervall $[x_0, y_0]$.

Durch vollständige Induktion zeigen wir jetzt das Bestehen der Ungleichungen

$$x_0 \leqq \cdots \leqq x_{k-1} \leqq x_k \leqq z^* \leqq y_k \leqq y_{k-1} \leqq \cdots \leqq y_0,$$
$$F(x_k) \leqq 0 \leqq F(y_k)$$

für $k = 0, 1, 2, \ldots$ und jede Lösung z^* von $F(x) = 0$ im Intervall $[x_0, y_0]$. Für $k = 0$ gelten diese Ungleichungen nach Voraussetzung. Wir betrachten nun den Operator

$$H_1(z) = z - G_1^{-1}\Big\{F(x_k) + \big(\delta F(x_k, y_k) + R(y_k - x_k)\big)(z - x_k)\Big\}$$

auf dem Intervall $[x_k, z^*]$. H_1 ist stetig. Wegen

$$H_1(v) - H_1(u) = G_1^{-1}\{G_1 - \delta F(x_k, y_k) - R(y_k - x_k)\}(v - u) \geqq 0$$

für $v \geqq u$ ist H_1 isoton. Außerdem ist $H_1(x_k) = x_k - G_1^{-1}F(x_k) \geqq x_k$. Schließlich ergibt sich wegen (3)

$$-\delta^2 F(x_k, z^*, y_k)(y_k - z^*) \leqq R(y_k - z^*) \leqq R(y_k - x_k),$$

also $\delta^2 F(x_k, z^*, y_k)(z^* - y_k) - R(y_k - x_k) \leqq 0$ und damit

$$\begin{aligned} H_1(z^*) &= z^* - G_1^{-1}\Big\{F(x_k) - F(z^*) - \big(\delta F(x_k, y_k) + R(y_k - x_k)\big)(x_k - z^*)\Big\} \\ &= z^* - G_1^{-1}\{\delta F(x_k, z^*) - \delta F(x_k, y_k) - R(y_k - x_k)\}(x_k - z^*) \\ &= z^* - G_1^{-1}\{\delta^2 F(x_k, z^*, y_k)(z^* - y_k) - R(y_k - x_k)\}(x_k - z^*) \\ &\leqq z^*. \end{aligned}$$

Auf Grund des Lemmas aus Abschnitt 1 besitzt daher H_1 einen Fixpunkt $x_{k+1} = H_1(x_{k+1})$ im Intervall $[x_k, z^*]$, d. h., es gilt

$$F(x_k) + \{\delta F(x_k, y_k) + R(y_k - x_k)\}(x_{k+1} - x_k) = 0.$$

Wir zeigen, daß $F(x_{k+1}) \leqq 0$ gilt. Wegen (3) gilt zunächst

$$\begin{aligned} 0 &\leqq \delta^2 F(x_k, y_k, x_{k+1})(y_k - x_{k+1}) + R(y_k - x_{k+1}) \\ &\leqq \delta^2 F(x_k, y_k, x_{k+1})(y_k - x_{k+1}) + R(y_k - x_k). \end{aligned}$$

Damit folgt durch Einsetzen der Iterationsvorschrift in

$$F(x_{k+1}) = F(x_k) - \delta F(x_k, x_{k+1})(x_k - x_{k+1})$$

die Ungleichung

$$F(x_{k+1}) = \{\delta F(x_k, y_k) - \delta F(x_k, x_{k+1}) + R(y_k - x_k)\} (x_k - x_{k+1})$$
$$= \{\delta^2 F(x_k, y_k, x_{k+1}) (y_k - x_{k+1}) + R(y_k - x_k)\} (x_k - x_{k+1}) \leqq 0 .$$

Die Aussagen für die Folge $\{y_k\}$ werden weitgehend analog bewiesen. Dazu betrachtet man den Operator

$$H_2(z) = z - G_2^{-1}\big\{F(y_k) + \big(\delta F(x_k, y_k) + R^*(y_k - x_k)\big) (z - y_k)\big\}$$

auf dem Intervall $[z^*, y_k]$ und zeigt die Existenz eines Fixpunktes y_{k+1} im Intervall $[z^*, y_k]$. $F(y_{k+1}) \geqq 0$ folgt ähnlich wie $F(x_{k+1}) \leqq 0$.

Durch Grenzübergang $k \to \infty$ folgt auf Grund der Voraussetzungen über die Halbordnung

$$\lim_{k\to\infty} x_k = x^* \leqq z^* \leqq y^* = \lim_{k\to\infty} y_k$$

für jede Lösung z^* von $F(x) = 0$ im Intervall $[x_0, y_0]$. Es gilt nach (4)

$$0 \leqq -F(x_k) = \{\delta F(x_k, y_k) + R(y_k - x_k)\} (x_{k+1} - x_k) \leqq G_1(x_{k+1} - x_k)$$

und

$$0 \leqq F(y_k) = \{\delta F(x_k, y_k) + R^*(y_k - x_k)\} (y_k - y_{k+1}) \leqq G_2(y_k - y_{k+1}) .$$

Aus der Konvergenz der beiden Folgen $\{x_k\}$ und $\{y_k\}$ folgt mit der Stetigkeit von F und den Voraussetzungen über die Halbordnung für $k \to \infty$ die Gleichheit $F(x^*) = F(y^*) = 0$. Damit sind alle Behauptungen des Satzes nachgewiesen. □

Der nächste Satz gibt eine Bedingung dafür an, daß $x^* = y^*$ gilt.

Satz 2. *Neben den Voraussetzungen von Satz* 1 *sei mit einer linearen Abbildung S, die invers isoton ist, die Ungleichung*

$$S \leqq \delta F(u, v) \tag{5}$$

für $x_0 \leqq u \leqq v \leqq y_0$ erfüllt. Dann ist $x^* = y^*$ die einzige Nullstelle von F in $[x_0, y_0]$.

Beweis. Wegen $x^* \leqq y^*$ gilt

$$0 = F(x^*) - F(y^*) = \delta F(x^*, y^*) (x^* - y^*) \leqq S(x^* - y^*) ,$$

somit $y^* \leqq x^*$, also $x^* = y^*$. □

Die folgenden Aussagen zeigen, daß unter geeigneten Voraussetzungen die Konvergenz des Verfahrens (V) asymptotisch quadratisch erfolgt.

Satz 3. *Es seien die Voraussetzungen der Sätze* 1 *und* 2 *erfüllt. Darüber hinaus sei S^{-1} stetig. Neben den Ungleichungen* (3) *mögen die Abschätzungen*

$$\begin{aligned} &\delta^2 F(u, v, w) \leqq R_1 , \\ &-R_1^* \leqq \delta^2 F(u, v, w)^* \end{aligned} \tag{6}$$

mit stetigen bilinearen Operatoren R_1 und R_1^* für $u, v, w \in [x_0, y_0]$ bestehen. Dann konvergieren die beiden Folgen $\{x_k\}$ und $\{y_k\}$ quadratisch gegen $x^* = y^*$.

Beweis. Mit (5) folgt aus dem ersten Teil von (V) mit (3)

$$
\begin{aligned}
&S(x^* - x_{k+1}) \leqq \{\delta F(x_k, y_k) + R(y_k - x_k)\}\,(x^* - x_{k+1}) \\
&= F(x_k) - F(x^*) + \{\delta F(x_k, y_k) + R(y_k - x_k)\}\,(x^* - x_k) \\
&= \delta F(x_k, x^*)\,(x_k - x^*) + \{\delta F(x_k, y_k) + R(y_k - x_k)\}\,(x^* - x_k) \\
&= \{\delta F(x_k, y_k) - \delta F(x_k, x^*) + R(y_k - x_k)\}\,(x^* - x_k) \\
&= \{\delta F(x_k, y_k, x^*)\,(y_k - x^*) + R(y_k - x_k)\}\,(x^* - x_k) \\
&\leqq \{R_1(y_k - x^*) + R(y_k - x^*) + R(x^* - x_k)\}\,(x^* - x_k).
\end{aligned}
$$

Setzt man nun

$$r_k := \max\{\|x^* - x_k\|, \|x^* - y_k\|\},$$

so folgt mit der Monotonie der Norm

$$
\begin{aligned}
\|x^* - x_{k+1}\| &\leqq \alpha\, \|S^{-1}\| \cdot \{\|R_1\| \cdot \|y_k - x^*\| + \|R\| \cdot \|y_k - x^*\| + \|R\| \cdot \|x^* - x_k\|\} \\
&\qquad \times \|x^* - x_k\| \\
&\leqq \alpha\, \|S^{-1}\|\, \{\|R_1\| + 2\,\|R\|\}\, r_k^2.
\end{aligned}
$$

Vollständig analog folgt $\|y_{k+1} - x^*\| \leqq \alpha\, \|S^{-1}\|\, \{\|R_1{}^*\| + 2\,\|R^*\|\}\, r_k^2$, insgesamt also

$$r_{k+1} \leqq \gamma r_k^2, \qquad k = 0, 1, 2, \ldots,$$

mit

$$\gamma = \alpha\, \|S^{-1}\| \cdot \max\{\|R_1\| + 2\,\|R\|, \|R_1{}^*\| + 2\,\|R^*\|\}. \ \square$$

Auf Grund der geforderten Regularität des Kegels K wird in Satz 1 z. B. der Banachraum $C[0, 1]$ der stetigen Funktionen mit der natürlichen Halbordnung ausgeschlossen. Verzichtet man auf die Forderung der Regularität des Kegels, so erhält man den folgenden Satz. Die Existenz einer Lösung kann man dabei allerdings mit den hier zur Verfügung stehenden Mitteln nicht beweisen. Sie muß auf anderem Wege nachgewiesen werden.

Satz 1*. *Der Banachraum B sei durch einen Kegel K halbgeordnet. F sei eine Abbildung von $D = [x_0, y_0] \subset B$ nach B, die in D Steigungen besitzt, welche den Forderungen* (1) *und* (2) *genügen. Es gelte $F(x_0) \leqq 0$ und $F(y_0) \geqq 0$. Weiter seien zwei positive stetige bilineare Operatoren R und R^* von $B \times B$ nach B bekannt mit den Eigenschaften* (3). *$\delta F(u, v) + R(v - u)$, $\delta F(u, v) + R^*(v - u)$ sowie $\delta F(u, v)$ seien für $x_0 \leqq u \leqq v \leqq y_0$ invers isoton. Außerdem seien $\delta F(u, v) + R(v - u)$ und $\delta F(u, v) + R^*(v - u)$ surjektiv. Dann gilt für die nach* (V) *berechneten Iterierten*

$$x_0 \leqq x_1 \leqq \cdots \leqq x_k \leqq x_{k+1} \leqq y_{k+1} \leqq y_k \leqq \cdots \leqq y_1 \leqq y_0. \ \square$$

Zum Beweis dieses Satzes bemerken wir nur, daß die Ungleichungen $x_k \leqq x_{k+1}$ und $y_{k+1} \leqq y_k$ durch vollständige Induktion unmittelbar aus den Voraussetzungen folgen unter Verwendung von $F(x_k) \leqq 0 \leqq F(y_k)$, was man ebenfalls durch vollständige Induktion zeigt. Die Ungleichung $x_{k+1} \leqq y_{k+1}$ sieht man folgendermaßen