

Lineare positive Operatoren und Fehlerabschätzungen bei Operatorgleichungen¹

Von

H. Schwetlick², Dresden

(Eingegangen am 6. Januar 1969)

Zusammenfassung — Summary

Lineare positive Operatoren und Fehlerabschätzungen bei Operatorgleichungen. In der vorliegenden Arbeit werden implizite und zugeordnete explizite Iterationsverfahren zur näherungsweisen Lösung von Fixpunktgleichungen untersucht, um die für die Konvergenz und die Berechnung von Fehlerschranken bekannten hinreichenden Bedingungen zu vergleichen. Es wird gezeigt, daß die genannten Bedingungen entweder für beide Verfahren gleichzeitig oder aber für keines der Verfahren erfüllt werden können. Dabei lassen sich für das implizite Verfahren stets Fehlerschranken angeben, die nicht schlechter als die für das explizite Verfahren sind. Grundlage dieser Aussagen ist eine Verallgemeinerung des Satzes von STEIN-ROSENBERG auf den Fall linearer positiver Operatoren, wobei auf Vollstetigkeit oder ähnliche Voraussetzungen verzichtet werden konnte. Weiter wird ein iteratives Verfahren angegeben, das die Berechnung von Fehlerschranken in endlich vielen Schritten genau dann gestattet, wenn die zitierten Bedingungen erfüllt sind.

Linear Positive Operators and Error Estimates to Operator-Equations. In the note presented here implicit and associated explicit iterations for approximate solution of fixed-point equations are considered in order to compare the known conditions sufficient for convergence and for calculation of error bounds. The mentioned conditions are shown to be satisfied either for both methods or for none of them. Moreover, it is possible to give error bounds for the implicit iteration which are not worse than the bounds for the explicit one. The basic idea of these statements is a generalization of the STEIN-ROSENBERG-Theorem to linear positive operators without requiring the assumption of compactness. At last an iterative method is given which allows under additional assumptions the calculation of error bounds after a finite number of steps if and only if the conditions cited above are satisfied.

1. Einleitung

Zur genäherten Lösung der Fixpunktgleichung $z = Fz$ in einem N -metrischen vollständigen Raum X kann das implizite Iterationsverfahren

$$y_n = G(y_n, y_{n-1}), \quad n = 1, 2, \dots, \quad (I)$$

¹ Erweiterter Auszug aus der Dissertation des Verfassers, Technische Universität Dresden, 1967. Es sei an dieser Stelle den Referenten Prof. Dr. H. HEINRICH und Prof. Dr. P. H. MÜLLER, Dresden, für ihr anhaltendes Interesse an dieser Arbeit gedankt. Ebenso ist der Verfasser Prof. Dr. J. W. SCHMIDT, Dresden, für zahlreiche Diskussionen und Anregungen zu Dank verpflichtet.

² Über die vorliegende Arbeit hat der Verfasser auf der V. Jahrestagung der Mathematischen Gesellschaft der DDR im Februar 1968 in Rostock vorgetragen.

herangezogen werden, wobei die Operatoren G und F über die Beziehung $G(z, z) = Fz$, $z \in D \subset X$, verknüpft sein sollen. Ebenso kann das i. a. einfachere zugeordnete explizite Verfahren

$$x_n = G(x_{n-1}, x_{n-1}), \quad n = 1, 2, \dots, \quad (E)$$

benutzt werden. Falls G bezüglich der beiden Argumente durch positive LIPSCHITZ-Operatoren mit genügend kleinem Spektralradius beschränkt ist und gewisse Kugelbedingungen erfüllt sind, konvergieren die Verfahren (I) bzw. (E), und man kann Fehlerschranken für die Näherungen x_n bzw. y_n angeben (s. [11], [13]).

In der vorliegenden Arbeit wird gezeigt, daß die genannten hinreichenden Konvergenzbedingungen für die beiden Verfahren (I) und (E) genau gleichzeitig erfüllbar sind. Außerdem können für das implizite Verfahren stets Fehlerschranken angegeben werden, die nicht größer als die für das explizite Verfahren sind, wenn beide Iterationen von einem gemeinsamen Startelement $x_0 = y_0$ ausgehen. Der Beweis dieser im Abschnitt 4 gebrachten Aussagen beruht wesentlich auf einer Verallgemeinerung des Satzes von STEIN-ROSENBERG auf den Fall linearer positiver Operatoren in einem durch einen normalen erzeugenden Kegel halbgeordneten BANACH-Raum. Diese im 3. Abschnitt vorgenommene Verallgemeinerung liefert eine etwas schwächere Aussage als die von MAREK [10] angegebene Übertragung, kommt dafür jedoch ohne die dort geforderten zusätzlichen Bedingungen an die Positivität und an das Spektrum der Operatoren aus.

Im 2. Abschnitt werden die dafür benötigten Begriffe und Sätze aus der Theorie der linearen positiven Operatoren zum Teil ohne Beweis bereitgestellt; es sei an dieser Stelle auf die Monographie [9] verwiesen. Im abschließenden 5. Abschnitt wird auf die numerische Auswertung der Fehlerschranken für den Fall eingegangen, daß der Kegel innere Punkte besitzt. Eine von SCHRÖDER [13] vorgeschlagene Methode ist hier stets anwendbar. Falls die darin benötigten Größen nach einer auch von BOHL [4] betrachteten Hilfsiteration bestimmt werden, können die SCHRÖDERSchen vergrößerten Schranken den im Konvergenzsatz angegebenen Fehlerschranken beliebig nahe gebracht werden.

2. Kegel und lineare positive Operatoren

Es seien N ein reeller oder komplexer BANACH-Raum und $K \neq \{0\}$ ein Kegel in N , d. h. K sei eine abgeschlossene Teilmenge von N mit den Eigenschaften $K + K \subset K$, $\lambda K \subset K$ für $\lambda \geq 0$, $\lambda \in \mathbb{R}^3$ und $K \cap -K = \{0\}$. Durch die Festlegung „ $x \leq y$, wenn $y - x \in K$ “ wird auf N eine mit der linearen Struktur verträgliche Halbordnung eingeführt. Ein Kegel K heißt *erzeugend*, wenn jedes $x \in N$ in der Form $x = u - v$ mit $u, v \in K$ darstellbar ist. Die Elemente u, v können dann sogar so gewählt werden, daß $\|u\|, \|v\| \leq M \|x\|$ mit einer von x un-

³ \mathbb{R} bezeichne den Körper der reellen Zahlen.

abhängigen Konstanten M gilt (s. etwa [3]). K heißt *räumlich*, wenn das Innere K^i von K nicht leer ist; für $y - x \in K^i$ wird auch $x < y$ geschrieben.

Lemma 2.1. *Es seien K räumlich und $e \in K^i$. Dann gelten folgende Aussagen:*

- (1) *Es gibt eine Zahl $\kappa > 0$, so daß $\frac{\|x\|}{\kappa} e \pm x \in K$ für alle $x \in N$ ist.*
- (2) *Die Zahl $\text{MAX} \frac{x}{e} := \min \{\mu \in R : \mu e - x \in K\}$ existiert für jedes $x \in N$.*
- (3) *$\text{MAX} \frac{x}{e}$ ist bei festem $x \in N$ eine stetige Funktion von $e \in K^i$.*

Beweis. (1) ist eine einfache Folgerung aus $e \in K^i$; (2) folgt aus (1) und der Abgeschlossenheit von K (siehe [9]).

Zu (3): Es sei $2\delta > 0$ der Radius einer Kugel

$$U(e, 2\delta) := \{z \in N : \|e - z\| \leq 2\delta\},$$

die mit e in K^i liegt. Wir wählen $f \in U(e, \delta) \subset K^i$. Dann gelten

$$\pm f \leq \left(\pm 1 + \frac{\|e - f\|}{\delta} \right) e \quad (2.1)$$

und

$$\pm e \leq \left(\pm 1 + \frac{\|e - f\|}{\delta} \right) f, \quad (2.2)$$

wobei entweder die oberen oder die unteren Vorzeichen zu nehmen sind.

1. Fall: Es sei $\text{MAX} \frac{x}{e} \geq 0$. Dann ist auch $\text{MAX} \frac{x}{f} \geq 0$, und mit (2.2) gilt $x \leq \text{MAX} \frac{x}{e} e \leq \text{MAX} \frac{x}{e} \left(1 + \frac{\|e - f\|}{\delta} \right) f$, also

$$\text{MAX} \frac{x}{f} \leq \text{MAX} \frac{x}{e} \left(1 + \frac{\|e - f\|}{\delta} \right). \quad (2.3)$$

Analog beweist man mit (2.1) die Ungleichung

$$\text{MAX} \frac{x}{e} \leq \text{MAX} \frac{x}{f} \left(1 + \frac{\|e - f\|}{\delta} \right). \quad (2.4)$$

Aus (2.3) folgt

$$\text{MAX} \frac{x}{f} - \text{MAX} \frac{x}{e} \leq \text{MAX} \frac{x}{e} \frac{\|e - f\|}{\delta}, \quad (2.5)$$

aus (2.4) ergibt sich mit (2.3)

$$\begin{aligned} -\text{MAX} \frac{x}{f} + \text{MAX} \frac{x}{e} &\leq \text{MAX} \frac{x}{f} \frac{\|e - f\|}{\delta} \leq \\ &\leq \text{MAX} \frac{x}{e} \left(1 + \frac{\|e - f\|}{\delta} \right) \frac{\|e - f\|}{\delta}. \end{aligned} \quad (2.6)$$

Wegen $\|e - f\| \leq \delta$ erhält man aus (2.5) und (2.6)

$$\left| \text{MAX} \frac{x}{f} - \text{MAX} \frac{x}{e} \right| \leq \left| \text{MAX} \frac{x}{e} \right| \frac{2}{\delta} \|e - f\|. \quad (2.7)$$

2. Im Fall $\text{MAX} \frac{x}{e} < 0$, $\text{MAX} \frac{x}{f} < 0$ ergibt sich die Abschätzung (2.7) in analoger Weise unter Benutzung der unteren Vorzeichen in (2.1) und (2.2); *q. e. d.*

Aus Lemma 2.1. folgt, daß jeder räumliche Kegel erzeugend ist.

Beispiel 2.2. Im Fall $N = R^n$, $x = (x_k) \in R^n$, erzeugt der Kegel $K := \{x \in R^n : x_k \geq 0, k = 1, \dots, n\}$ die natürliche (komponentenweise) Halbordnung. Mit $K^i = \{e = (e_k) \in R^n : e_k > 0, k = 1, \dots, n\}$ ergibt sich $\text{MAX} \frac{x}{e} = \max_k \frac{x_k}{e_k}$.

Ein reeller BANACH-Raum N mit einem Kegel K heißt *Rieszscher Banach-Raum*, wenn zu jedem $x \in N$ das Element $\sup(x, 0)$ in N existiert. In einem RIESZSchen BANACH-Raum sind zu zwei Elementen stets Supremum und Infimum vorhanden. Mit $|x| := \sup(x, -x)$, $x^+ := \sup(x, 0)$, $x^- := \sup(-x, 0)$ gelten $x = x^+ - x^-$ und $|x| = x^+ + x^-$. Ersichtlich ist der Kegel eines RIESZSchen BANACH-Raumes erzeugend.

Schließlich heißt ein Kegel *normal*, wenn eine Zahl $\alpha > 0$ existiert, so daß aus $0 \leq x \leq y$ folgt $\|x\| \leq \alpha \|y\|$.

Lemma 2.3. *Es sei N ein Rieszscher Banach-Raum mit einem normalen Kegel K . Dann gibt es eine Zahl $\gamma > 0$, so daß*

$$(1) \|x\| \leq 2\alpha \| |x| \| \leq \gamma \|x\| \quad \text{und}$$

$$(2) \| |x| - |y| \| \leq \gamma \alpha \|x - y\|$$

für alle $x, y \in N$ gelten.

Beweis. Zu (1): Wegen der Normalität folgt aus $0 \leq x^+ \leq |x|$ und $0 \leq x^- \leq |x|$ sofort $\|x\| = \|x^+ - x^-\| \leq \|x^+\| + \|x^-\| \leq 2\alpha \| |x| \|$.

Es sei jetzt $x = u - v$ mit $u, v \in K$ und $\|u\|, \|v\| \leq M \|x\|$. Dann gilt $0 \leq |x| \leq u + v$, also $\| |x| \| \leq \alpha \|u + v\| \leq 2\alpha M \|x\| =: \frac{\gamma}{2\alpha} \|x\|$.

Zu (2): Da $| \cdot |$ die Dreiecksungleichung erfüllt, gilt

$$\| |x| - |y| \| \leq \|x - y\|, \text{ also mit (1) }$$

$\frac{1}{2\alpha} \| |x| - |y| \| \leq \| |x| - |y| \| \leq \alpha \| |x - y| \| \leq \frac{\gamma}{2} \|x - y\|$ wie behauptet.

Es sei T ein auf N definierter linearer beschränkter Operator. Falls N reell ist, wird T auf die komplexe Erweiterung $\tilde{N} = N + iN$ ($\|z\| := \max_{0 \leq \varphi \leq 2\pi} \|x \sin \varphi + y \cos \varphi\|$ für $z = x + iy \in \tilde{N}$, $x, y \in N$) zu einem Operator \tilde{T} nach der Vorschrift $\tilde{T}z := Tx + iTy$ fortgesetzt. Da \tilde{N} mit N ein BANACH-Raum und $\tilde{K} := K + iK$ mit K ein normaler (erzeugender, räumlicher) Kegel in \tilde{N} ist (siehe [12]), werden im folgenden für \tilde{N} und \tilde{T} wieder die Bezeichnungen N und T verwendet.

Mit $\sigma(T)$ werde das *Spektrum* von T bezeichnet; ferner sei $r(T) := \sup \{ \|\lambda\| : \lambda \in \sigma(T) \}$ der *Spektralradius* von T . Es gilt dann

$$r(T) = \lim_{n \rightarrow \infty} \|T^n\|^{1/n}. \quad (2.8)$$

Ein auf N erklärter Operator A heißt *positiv*, wenn $AK \subset K$ gilt.

Satz 2.4. *Es seien K ein normaler erzeugender Kegel⁴ und A ein linearer positiver Operator. Dann gelten folgende Aussagen:*

- (1) A ist beschränkt.
- (2) $r(A) \in \sigma(A)$.
- (3) *Es gibt eine Folge $(z_n) \subset N$ mit $\|z_n\| = 1$, $n = 1, 2, \dots$, und $\lim_{n \rightarrow \infty} \|(r(A)I - A)z_n\| = 0$.*

Beweis. Aussage (1) wurde in [3] bewiesen; zu (2) siehe [12].

Zu (3): Sei $r := r(A) \in \sigma_P(A)$. Dann gibt es ein $z_0 \in N$, $\|z_0\| = 1$, mit $(rI - A)z_0 = 0$. Man setze $z_n = z_0$, $n = 1, 2, \dots$. Falls $r \in \sigma_C(A)$, folgt nach [7, S. 583] die Behauptung. Da r als Randpunkt des Spektrums nicht zum reinen Residualspektrum gehören kann ([8], S. 70), ist $(rI - A)^{-1}$ im Fall $r \in \sigma_R(A)$ auf $(rI - A)N$ nicht beschränkt. Es gibt dann stets eine Folge (z_n) mit den in (3) geforderten Eigenschaften.

Satz 2.5. *Es seien K ein normaler erzeugender Kegel, A ein linearer positiver Operator und μ eine reelle Zahl. Dann sind die folgenden Aussagen äquivalent:*

- (1) $r(A) < \mu$.
- (2) $(\mu I - A)^{-1}$ existiert und ist positiv.
- (3) *Es ist $\mu > 0$, und $Rx := \frac{1}{\mu} \sum_{k=0}^{\infty} \left(\frac{1}{\mu} A\right)^k x$ konvergiert für jedes $x \in N$.*

Falls K überdies räumlich ist, sind (1), (2) und (3) noch äquivalent zu

- (4) *Es gibt ein $e \in K^i$ mit $(\mu I - A)e \in K^i$.*

Beweis. Die Äquivalenz von (1) und (2) wurde unter schwächeren Voraussetzungen an den Kegel in [12] bewiesen. (3) folgt aus (1) wegen der Konvergenz der Reihe $\frac{1}{\mu} \sum_{k=0}^{\infty} \left(\frac{1}{\mu} A\right)^k$ in der gleichmäßigen Operatorentopologie, während der Teil (3) \Rightarrow (2) in [6] gezeigt wurde.

(2) \Rightarrow (4): Es sei K räumlich. Dann besitzt $e := (\mu I - A)^{-1} x_0$ die in (4) geforderten Eigenschaften, wenn $x_0 \in K^i$ gewählt wurde.

⁴ Die Kegel der bezüglich der natürlichen (punktweisen) Halbordnungen nicht-negativen Elemente in den Räumen R^n , $C[\Omega]$ und L_p , $p \geq 1$, sind normal und erzeugend.

(4) \Rightarrow (1): Aus $e \in K^i$ und $(\mu I - A)e \in K^i$ folgt $\mu > 0$. Wegen Lemma 2.1. gibt es dann ein $\tilde{\mu}$, $0 \leq \tilde{\mu} < \mu$, so daß $(\tilde{\mu} I - A)e \in K$, also $Ae \leq \tilde{\mu}e$ gilt. Nach einer von STECENKO [15] bewiesenen Verallgemeinerung des COLLATZschen Quotientensatzes folgt hieraus $r(A) \leq \tilde{\mu} < \mu$; *q. e. d.*

Eine mit Satz 2.5. im wesentlichen identische Aussage wurde in [5] und teilweise in [11] für positive Operatoren in topologischen halbgeordneten Räumen angegeben, jedoch wird dort der Kegel zusätzlich als regulär⁵ vorausgesetzt.

Satz 2.6. *Es seien K ein normaler erzeugender Kegel und T, P lineare Operatoren, und es gelte*

$$(P \pm T)x \in K \quad \text{für} \quad x \in K. \quad (2.9)$$

Dann ist T beschränkt, und es ist $r(T) \leq r(P)$.

Zum Beweis sei auf [10] und [14] verwiesen. Wir bemerken, daß die Bedingung (2.9) in einem RIESZschen BANACH-Raum äquivalent zu

$$|Tx| \leq P|x| \quad \text{für alle } x \in N$$

ist; P ist in diesem Sinne ein Schrankenoperator für T .

Folgerung 2.7. *Es seien K normal und erzeugend und A, B lineare positive Operatoren. Falls $(B - A)x \in K$ für $x \in K$ gilt, ist $r(A) \leq r(B)$.*

3. Der Satz von Stein-Rosenberg

Es seien Q, R zwei auf einem BANACH-Raum N mit einem normalen erzeugenden Kegel K definierte lineare positive Operatoren. Aus ihnen werden die Operatoren $T := Q + R$ und, falls $r(Q) < 1$ ist⁶, $P := (I - Q)^{-1}R$ gebildet. Wir beweisen zunächst

Lemma 3.1. *Es sei $r(Q) < 1$. Dann ist $r(P) \leq r(PQ + R)$.*

Beweis. Nach Satz 2.4 gibt es eine Folge $(z_n) \subset N$ mit $\|z_n\| = 1$ und

$$\lim_{n \rightarrow \infty} \|(pI - P)z_n\| = 0, \quad p := r(P). \quad (3.1)$$

Nun gilt

$$p^k I - (pQ + R)^k = \left(\sum_{i=0}^{k-1} p^{k-1-i} (pQ + R)^i \right) (I - Q) (pI - P),$$

woraus mit (3.1) $\lim_{n \rightarrow \infty} \|(p^k I - (pQ + R)^k)z_n\| = 0$ und wegen $\|z_n\| = 1$

$$p^k = \lim_{n \rightarrow \infty} \|(pQ + R)^k z_n\|, \quad k = 1, 2, \dots,$$

⁵ K heißt *regulär*, wenn jede monotone ordnungsbeschränkte Folge aus K einen Limes besitzt. Der Kegel aller punktweise nichtnegativen Funktionen in $C[\Omega]$ ist nicht regulär.

⁶ Nach Satz 2.5. existiert in diesem Fall $(I - Q)^{-1}$ und ist positiv.

folgen. Damit ergibt sich

$$p = \lim_{n \rightarrow \infty} \| (pQ + R)^k z_n \|^{1/k} \leq \| (pQ + R)^k \|^{1/k}, \quad k = 1, 2, \dots$$

Grenzübergang $k \rightarrow \infty$ liefert die Behauptung.

Für die Ausführungen des nächsten Abschnitts ist der folgende Satz von wesentlicher Bedeutung.

Satz 3.2. *Es seien K ein normaler erzeugender Kegel und Q, R zwei lineare positive Operatoren. Dann sind die Aussagen*

$$(1. a) \quad r(T) < 1 \quad \text{mit} \quad T = Q + R \quad \text{und}$$

$$(1. b) \quad r(Q) < 1 \quad \text{und} \quad r(P) < 1 \quad \text{mit} \quad P = (I - Q)^{-1} R$$

äquivalent. Ist eine der beiden Aussagen erfüllt, so gilt

$$(1) \quad r(P) \leq r(T) < 1.$$

Beweis. (1. a) \Rightarrow (1. b): Für $x \in K$ gilt $(T - Q)x = Rx \in K$, nach Folgerung 2.7. ist daher $r(Q) \leq r(T) < 1$. Wir nehmen $p := r(P) \geq 1$ an. Dann gilt $(pT - (pQ + R))x = (p - 1)Rx \in K$ für $x \in K$. Nach derselben Folgerung ergibt sich daraus $r(pQ + R) \leq r(pT) = pr(T)$. Mit Lemma 3.1. folgt $p \leq r(pQ + R) \leq pr(T)$, also $1 \leq r(T)$ im Widerspruch zu (1. a).

(1. b) \Rightarrow (1. a): Wegen $r(Q) < 1$ und $r(P) < 1$ existieren $(I - Q)^{-1}$ und $(I - P)^{-1}$ als positive Operatoren. Aus der Identität $(I - Q)(I - P) = I - T$ folgt daher, daß $(I - T)^{-1} = (I - P)^{-1}(I - Q)^{-1}$ existiert und positiv ist. Nach Satz 2.5. ist dies mit $r(T) < 1$ gleichbedeutend.

Zu (1): Es seien $r(T) < 1$, $r(Q) < 1$ und $p := r(P) < 1$. Für $x \in K$ ist dann $(T - (pQ + R))x = (1 - p)Qx \in K$, d. h. es gilt $r(pQ + R) \leq r(T)$ nach Folgerung 2.7. Mit Lemma 3.1. folgt $p = r(P) \leq r(T)$; q. e. d.

Die Aussage des eben bewiesenen Satzes läßt sich erweitern zu einer Verallgemeinerung des Satzes von STEIN-ROSENBERG (s. [16]) über Matrizen mit nichtnegativen Elementen.

Satz 3.3. *Es seien N ein Rieszscher Banach-Raum mit einem normalen Kegel K und Q, R lineare positive Operatoren, wobei $r(Q) < 1$ gelte. Dann genügen die Spektralradien der Operatoren $T = Q + R$ und $P = (I - Q)^{-1} R$ genau einer der beiden folgenden Ungleichungen*

$$(1) \quad r(P) \leq r(T) < 1,$$

$$(2) \quad r(P) \geq r(T) \geq 1.$$

Beweis. Wegen Satz 3.2. genügt es zu zeigen, daß im Fall $t := r(T) \geq 1$ und $p := r(P) \geq 1$ die Ungleichung (2) gilt.

Sei $(z_n) \subset N$ eine nach Satz 2.4 existierende Folge mit $\|z_n\| = 1$ und $\lim_{n \rightarrow \infty} \|(tI - T)z_n\| = 0$. Nach Lemma 2.3. ist

$$\| |tz_n| - |Tz_n| \| \leq \gamma \alpha \|(tI - T)z_n\|,$$

wegen $|t z_n| = t |z_n|$ gilt daher

$$\lim_{n \rightarrow \infty} \|\delta_n\| = 0, \quad \delta_n := |t z_n| - |T z_n|. \quad (3.2)$$

Unter Beachtung von $|T z_n| \leq T |z_n|$ und $t \geq 1$ erhält man weiter

$t |z_n| = |T z_n| + \delta_n \leq T |z_n| + \delta_n \leq (t Q + R) |z_n| + \delta_n$, d. h.
 $t (I - Q) |z_n| \leq R |z_n| + \delta_n$. Multiplikation mit $(I - Q)^{-1}$ liefert
 $t |z_n| \leq P |z_n| + (I - Q)^{-1} \delta_n$. Durch Induktion beweist man, daß

$$t^k |z_n| \leq P^k |z_n| + S_k (I - Q)^{-1} \delta_n, \quad k = 1, 2, \dots, \quad (3.3)$$

gilt, wobei $S_k := \sum_{i=0}^{k-1} t^i P^{k-1-i}$ gesetzt wurde. Aus (3.3) folgt

$$t^k \| |z_n| \| \leq \alpha \|P^k\| \| |z_n| \| + \alpha \|S_k\| \|(I - Q)^{-1}\| \|\delta_n\|,$$

mit Lemma 2.3. und wegen $\|z_n\| = 1$ somit

$$\frac{1}{2\alpha} t^k \leq \frac{\gamma}{2} \|P^k\| + \alpha \|S_k\| \|(I - Q)^{-1}\| \|\delta_n\| \text{ für jedes } k \text{ und } n.$$

Für $n \rightarrow \infty$ ergibt sich mit (3.2) $t^k \leq \alpha \gamma \|P^k\|$, $k = 1, 2, \dots$, woraus über $t \leq \|P^k\|^{1/k} (\alpha \gamma)^{1/k}$ schließlich $t = r(T) \leq r(P)$ folgt; *q. e. d.*

Unter Voraussetzungen, die die Existenz von Eigenelementen der Operatoren T bzw. P zu den Eigenwerten $r(T)$ bzw. $r(P)$ im Kegel K sichern, und zusätzlichen Positivitätsforderungen an die Operatoren wie etwa u_0 -positiv wurde der Satz von STEIN-ROSENBERG kürzlich von MAREK [10] in der ursprünglichen Form mit strengen Ungleichheitszeichen bewiesen. Für den im folgenden Abschnitt vorzunehmenden Vergleich von Iterationsverfahren genügt jedoch die Aussage von Satz 3.2., die nur mit den für Schrankenoperatoren Q und R sowieso geforderten Eigenschaften „linear und positiv“ auskommt.

4. Vergleich von expliziten und impliziten Iterationsverfahren

Im folgenden bezeichne X einen N -metrischen vollständigen Raum und D eine Teilmenge von X , wobei N wie bisher als BANACH-Raum mit einem normalen erzeugenden Kegel K vorausgesetzt wird (zur Definition „ N -metrischer Raum“ siehe etwa [6]; in den Bezeichnungen folgen wir [11]).

Es seien Abbildungen

$$F|D \rightarrow X \text{ und } G|D \times D \rightarrow X \text{ mit } G(z, z) = Fz \text{ für } z \in D$$

vorgegeben, und es gebe auf N erklärte lineare positive Operatoren Q und R , so daß

$$\left. \begin{aligned} Q(G(u, w), G(v, w)) &\leq Q Q(u, v) \\ Q(G(w, u), G(w, v)) &\leq R Q(u, v) \end{aligned} \right\} u, v, w \in D$$

gilt.

Wir betrachten das implizite Iterationsverfahren

$$y_n = G(y_n, y_{n-1}), \quad n = 1, 2, \dots, \quad (I)$$

und das zugehörige explizite Verfahren

$$x_n = G(x_{n-1}, x_{n-1}) = F x_{n-1}, \quad n = 1, 2, \dots, \quad (E)$$

zur Bestimmung eines Fixpunktes $z^* = F z^* \in D$.

Über die Durchführbarkeit und Konvergenz dieser Verfahren sind die folgenden Aussagen bekannt [11]⁷, [13]:

Voraussetzung (VE):

Es sei $r(T) < 1$, $T = Q + R$, und es gebe ein $x_0 \in D$, so daß mit

$$x_1 = F x_0 = G(x_0, x_0) \text{ gilt}$$

$$U := \{z \in X : \varrho(z, x_1) \leq (I - T)^{-1} T \varrho(x_0, x_1)\} \subset D.$$

Aussage 1: *Die Voraussetzung (VE) sei erfüllt. Dann besitzt F genau einen Fixpunkt $z^* \in D$, die Iteration (E) ist mit x_0 als Startelement unbeschränkt durchführbar, und es gelten $\lim_{n \rightarrow \infty} x_n = z^*$ sowie*

$$\varrho(z^*, x_n) \leq (I - T)^{-1} T^n \varrho(x_0, x_1) =: \xi_n.$$

Voraussetzung (VI):

Es seien $r(Q) < 1$ und $r(P) < 1$, $P = (I - Q)^{-1} R$, und es gebe ein $y_0 \in D$, so daß die Gleichung $y = G(y, y_0)$ eine Lösung $y_1 \in D$ besitzt, für die

$$S := \{z \in X : \varrho(z, y_1) \leq (I - P)^{-1} P \varrho(y_0, y_1)\} \subset D \text{ gilt.}$$

Aussage 2: *Die Voraussetzung (VI) sei erfüllt. Dann besitzt F genau einen Fixpunkt $z^* \in D$, die Iteration (I) ist mit y_0 als Startelement unbeschränkt durchführbar (d. h. die Gleichungen $y = G(y, y_{n-1})$ besitzen für $n = 1, 2, \dots$ eine Lösung $y_n \in D$), und es gelten $\lim_{n \rightarrow \infty} y_n = z^*$ sowie*

$$\varrho(z^*, y_n) \leq (I - P)^{-1} P^n \varrho(y_0, y_1) =: \eta_n.$$

Es soll untersucht werden, in welcher Beziehung die für die Anwendung der Verfahren (E) bzw. (I) hinreichenden Bedingungen (VE) bzw. (VI) stehen.

Satz 4.1. *Die Bedingungen (VE) bzw. (VI) sind höchstens beide gleichzeitig erfüllbar. Es gilt genauer: Falls (VE) mit x_0 erfüllt ist, so gilt (VI) für $y_0 := x_0$. Sind umgekehrt die Bedingungen in (VI) mit y_0 erfüllt, so genügt $x_0 := y_1$ der Voraussetzung (VE).*

Beweis. (VE) \Rightarrow (VI): Mit $r(T) < 1$ gelten nach Satz 3.2. auch die Ungleichungen $r(Q) < 1$ und $r(P) < 1$. Wir setzen $y_0 := x_0$, $x_1 = G(x_0, x_0)$ und $S' := \{z \in X : \varrho(z, x_1) \leq (I - Q)^{-1} Q \varrho(x_0, x_1) =: \eta'\}$.

⁷ Man beachte, daß nach Satz 2.5. ein linearer positiver Operator T genau dann konvergent ist (d. h. seine geometrische Reihe $\sum_{k=0}^{\infty} T^k x$ konvergiert für jedes $x \in N$), wenn $r(T) < 1$ gilt. Die Überlegungen von [11] lassen sich damit sofort auf den hier betrachteten Fall übertragen.

Wegen $\eta' \leq \xi_1$ gilt $S' \subset U \subset D$, und für $z \in S'$ folgt

$$\begin{aligned} \varrho(G(z, x_0), x_1) &= \varrho(G(z, x_0), G(x_0, x_0)) \leq Q \varrho(z, x_0) \leq \\ &\leq Q[\varrho(z, x_1) + \varrho(x_1, x_0)] \leq Q[(I - Q)^{-1} Q \varrho(x_0, x_1) + \varrho(x_0, x_1)] = \eta', \end{aligned}$$

also $G(\cdot, x_0) S' \subset S'$. Zufolge des Kontraktionsprinzips gibt es daher ein $y_1 \in S' \subset U \subset D$ mit $y_1 = G(y_1, y_0)$ und $\varrho(y_1, x_1) \leq \eta'$.

Sei jetzt $z \in S$, d. h. $\varrho(z, y_1) \leq (I - P)^{-1} P \varrho(x_0, y_1)$. Dann gilt

$$\begin{aligned} \varrho(z, x_1) &\leq \varrho(z, y_1) + \varrho(y_1, x_1) \\ &\leq (I - P)^{-1} P \varrho(x_0, y_1) + \varrho(y_1, x_1) = \eta_1 + \varrho(y_1, x_1) \\ &\leq (I - P)^{-1} P \varrho(x_0, x_1) + (I - P)^{-1} P \varrho(x_1, y_1) + \varrho(y_1, x_1) \\ &= (I - P)^{-1} P \varrho(x_0, x_1) + (I - P)^{-1} \varrho(x_1, y_1) \\ &\leq (I - P)^{-1} [P \varrho(x_0, x_1) + (I - Q)^{-1} Q \varrho(x_0, x_1)] \\ &= (I - P)^{-1} (I - Q)^{-1} (R + Q) \varrho(x_0, x_1) \\ &= (I - T)^{-1} T \varrho(x_0, x_1) = \xi_1, \end{aligned}$$

also $z \in U$ und daher $S \subset U \subset D$.

(VI) \Rightarrow (VE): Nach Satz 3.2. ist $r(T) < 1$. Wir setzen $x_0 := y_1$, dann ist $x_1 = G(x_0, x_0) = G(y_1, y_1)$. Es sei nun $z \in U$, d. h. es gelte $\varrho(z, x_1) \leq (I - T)^{-1} T \varrho(x_0, x_1)$. Dann gilt

$$\begin{aligned} \varrho(z, y_1) &= \varrho(z, x_0) \leq \varrho(z, x_1) + \varrho(x_1, x_0) \\ &\leq (I - T)^{-1} T \varrho(x_0, x_1) + \varrho(x_0, x_1) = (I - T)^{-1} \varrho(x_0, x_1) \\ &= (I - T)^{-1} \varrho(G(y_1, y_0), G(y_1, y_1)) \leq (I - T)^{-1} R \varrho(y_0, y_1) \\ &= (I - P)^{-1} (I - Q)^{-1} R \varrho(y_0, y_1) = (I - P)^{-1} P \varrho(y_0, y_1) = \eta_1, \end{aligned}$$

also $z \in S$ und daher $U \subset S \subset D$; q. e. d.

Wir zeigen als nächstes, daß die Fehlerschranken für das explizite Verfahren stets nicht kleiner sind als diejenigen des impliziten Verfahrens, also (I) in diesem Sinne „besser“ als (E) ist.

Satz 4.2. *Es seien die Voraussetzungen (VE) und (VI) für ein gemeinsames Startelement $x_0 = y_0$ erfüllt⁸. Dann gelten für die Fehlerschranken ξ_n bzw. η_n der aus $x_0 = y_0$ gemäß (E) bzw. (I) berechneten Iterierten x_n bzw. y_n die Ungleichungen*

$$\eta_n \leq \xi_n - T^{n-1} \varrho(x_1, y_1) \leq \xi_n$$

und

$$\eta_n \leq \xi_n - P^{n-1} \varrho(x_1, y_1) \leq \xi_n, \quad n = 1, 2, \dots$$

Beweis. Für $n = 1$ wurden die Ungleichungen bereits im ersten Teil des Beweises zu Satz 4.1. bewiesen. Wir nehmen an, daß die erste Ungleichung für $n = k$ richtig ist, also $\xi_k - \eta_k \geq T^{k-1} \varrho(x_1, y_1)$ gilt. Wegen der Monotonie von T folgt daraus

$$T(\xi_k - \eta_k) \geq T^k \varrho(x_1, y_1). \quad (4.1)$$

⁸ Nach Satz 4.1. ist dies stets möglich, wenn auch nur eine der genannten Voraussetzungen erfüllt ist.

Weiter gilt

$$(T - P) \eta_k = (T - P) (I - P)^{-1} P^k \varrho(y_0, y_1) = Q P^k \varrho(y_0, y_1) \geq 0. \quad (4.2)$$

Mit (4.1) und (4.2) ergibt sich

$$\xi_{k+1} - \eta_{k+1} = T \xi_k - P \eta_k = T (\xi_k - \eta_k) + (T - P) \eta_k \geq T^k \varrho(x_1, y_1).$$

Unter Beachtung von

$$\begin{aligned} (T - P) \xi_k &= (T - P) (I - T)^{-1} T^k \varrho(x_0, x_1) = \\ &= Q (I - Q)^{-1} T^k \varrho(x_0, x_1) \geq 0 \end{aligned}$$

beweist man in analoger Weise die noch ausstehende Ungleichung.

Im Spezialfall eines linearen Gleichungssystems ($D = X = N = R^n$) kann durch geeignete Festlegung von F und G erreicht werden, daß (E) bzw. (I) in das Gesamt- bzw. Einzelschrittverfahren übergehen. Für diesen Fall wurde eine mit den Sätzen 4.1 und 4.2. vergleichbare Aussage von ALBRECHT [1] bewiesen. Die dort betrachtete „Iteration mit monotonen Folgen“ läßt sich durch eine Transformation auf die hier betrachtete Form bringen. Eine Übertragung der ALBRECHTSchen Ergebnisse auf den Fall sog. monoton-zerlegbarer linearer vollstetiger Operatoren in BANACH-Räumen mit einem Kegel ist in [14] zu finden.

5. Auswertung der Fehlerschranken⁹

Zur Berechnung der Fehlerschranken ξ_n ist die Invertierung des Operators $(I - T)$ bzw. die Lösung der Gleichung $(I - T) \xi_n = T^n \varrho(x_0, x_1)$ erforderlich¹⁰. Falls der Kegel K räumlich ist und ein Element

$$e \in K^i \quad \text{mit} \quad (I - T) e \in K^i \quad (5.1)$$

existiert, kann dies nach SCHRÖDER [13] bei Vergrößerung der Schranke ξ_n durch eine einfache Maximumbildung umgangen werden. Es gilt dann nämlich (für den endlichdimensionalen Fall siehe [2]):

$$\varrho(z^*, x_n) \leq \xi_n \leq \text{MAX} \left\{ \frac{\varrho(x_0, x_1)}{(I - T)^i e} \right\} T^n e =: \delta_n(e). \quad (5.2)$$

Zum Beweis setzen wir $M := \text{MAX} \{ \varrho(x_0, x_1) / (I - T) e \}$. Über

$$\varrho(x_0, x_1) \leq M (I - T) e$$

folgt wegen der Monotonie von $(I - T)^{-1} T^n$ sofort

$$\xi_n = (I - T)^{-1} T^n \varrho(x_0, x_1) \leq M T^n e = \delta_n(e).$$

⁹ Wie der Verfasser nach Einreichen dieser Arbeit von Herrn Prof. Dr. L. COLLATZ erfuhr, werden ähnliche bzw. äquivalente Aussagen wie die in diesem Abschnitt angegebenen in zwei Arbeiten von E. BOHL bewiesen: 1. Operator Equations on a Partially Ordered Vector Space. Erscheint demnächst in *aequationes mathematicae*. 2. Über Fehlerabschätzungen bei nichtlinearen Operatorgleichungen. Zur Veröffentlichung eingereicht.

¹⁰ Wir beschränken uns im folgenden auf das explizite Verfahren (E). Sämtliche Ausführungen gelten auch für die Iteration (I), falls T durch P und ξ_n durch η_n ersetzt werden.

Wenn die Voraussetzung (VE) für das explizite Verfahren erfüllt und der Kegel K räumlich ist, gibt es nach Satz 2.5. stets ein Element e mit den in (5.1) geforderten Eigenschaften. Zum Aufsuchen eines solchen Elementes schlägt ALBRECHT [2] (im R^n) die Iteration

$$e_{k+1} = T e_k, \quad k = 0, 1, \dots, \quad e_0 \in K^i \quad (5.3)$$

vor. Existiert ein Index m , so daß $e_{m+1} = T e_m < e_m$ gilt, kann e_m als Hilfselement e zur Berechnung von $\delta_n(e_m)$ herangezogen werden. Im Fall $N = R^n$, Kegel wie im Beispiel 2.2., gibt es zu jedem $e_0 \in K^i$ einen Index m mit dieser Eigenschaft, wenn nur die nichtnegative Matrix T nichtzerfallend und primitiv ist [16]. Es können jedoch nichtzerfallende (notwendig zyklische) Matrizen T angegeben werden, für welche die Ungleichung $T e_k < e_k$ für keinen Index $k = 0, 1, \dots$, erfüllt ist, obwohl $e_0 \in K^i$ gewählt wurde und $r(T) < 1$ gilt (Beispiel siehe [14]). Aus dem Beweis zum Satz 2.5. ist ersichtlich, wie die Iteration (5.3) abgeändert werden muß, damit sie in endlich vielen Schritten ein Element mit den Eigenschaften (5.1) liefert.

Satz 5.1. (Siehe auch Fußnote 9). *Es seien N ein Banach-Raum mit einem normalen räumlichen Kegel und T ein linearer positiver Operator. Die Folge (e_k) sei durch*

$$e_{k+1} = T e_k + e_0, \quad k = 0, 1, \dots, \quad e_0 \in K^i \quad (5.4)$$

definiert. Dann sind die folgenden Aussagen äquivalent:

- (1) $r(T) < 1$.
- (2) Zu jedem $e_0 \in K^i$ gibt es einen Index m , so daß

$$e_k \in K^i \quad \text{und} \quad (I - T) e_k \in K^i$$

für $k \geq m$ gelten.

Beweis. (1) \Rightarrow (2): Wegen $r(T) < 1$ gilt $\lim_{k \rightarrow \infty} (I - T) e_k = e_0$. Da $e_0 \in K^i$ vorausgesetzt wurde, ist auch $(I - T) e_k \in K^i$ für $k \geq m$, m hinreichend groß. Durch Induktion folgt weiter $e_k \geq 0$, $k = 0, 1, \dots$, also $e_k = (I - T) e_k + T e_k \in K^i$. Die Umkehrung ist im Satz 2.5. bewiesen.

Im Fall eines vollstetigen Operators T ist die Iteration (5.4) als Sonderfall in einem auch für gewisse nichtlineare Aufgaben anwendbaren Algorithmus von BOHL [4] enthalten.

Wir setzen jetzt zusätzlich voraus, daß $\varrho(x_0, x_1) \in K^i$ ist. Dann kann in (5.4) $e_0 = \varrho(x_0, x_1)$ gewählt werden. Die mit diesem Startelement berechneten iterierten e_k können für $k \geq m$ gemäß (5.2) zur Bestimmung der Schranken $\delta_n(e_k)$ benutzt werden. Im folgenden Satz wird gezeigt, daß diese speziellen Schranken „asymptotisch optimal“ sind.

Satz 5.2. *Es sei $r(T) < 1$, und die Folge (e_k) sei nach (5.4) mit $e_0 = \varrho(x_0, x_1) \in K^i$ bestimmt. Dann gilt*

$$\lim_{k \rightarrow \infty} \text{MAX} \left\{ \frac{\varrho(x_0, x_1)}{(I - T) e_k} \right\} T^n e_k = (I - T)^{-1} T^n \varrho(x_0, x_1)$$

für $n = 1, 2, \dots$

Beweis. Wegen $\lim_{k \rightarrow \infty} (I - T) e_k = \varrho(x_0, x_1) \in K^i$ und der im Lemma 2.1. nachgewiesenen Stetigkeit von $\text{MAX} \frac{x}{f}$ für $f \in K^i$ gilt

$$\lim_{k \rightarrow \infty} \text{MAX} \left\{ \frac{\varrho(x_0, x_1)}{(I - T) e_k} \right\} = \text{MAX} \left\{ \frac{\varrho(x_0, x_1)}{\varrho(x_0, x_1)} \right\} = 1.$$

Mit $\lim_{k \rightarrow \infty} T^n e_k = T^n (I - T)^{-1} \varrho(x_0, x_1)$ folgt daraus die Behauptung.

Unter zusätzlichen Voraussetzungen konvergiert die Iteration (5.3) — falls die e_k nach jedem Schritt normiert werden — gegen das zum Eigenwert $r(T)$ gehörende Eigenelement von T . Satz 5.2. zeigt, daß bei der Fehlerabschätzung nach (5.2) nicht dieses Eigenelement eine ausgezeichnete Rolle als Hilfselement spielt, sondern mit den gemäß (5.4) berechneten e_k im Fall $e_0 = \varrho(x_0, x_1) \in K^i$ zumindest für große Werte von k bessere Schranken zu erwarten sind. Dies ist zum Beispiel im Spezialfall linearer Gleichungssysteme auch numerisch zu beobachten.

Beispiel 5.3. Das [1] entnommene lineare Gleichungssystem $Ax = b$,

$$A = \begin{pmatrix} 4,33 & -1,12 & -1,08 & 1,14 \\ -1,12 & 4,33 & 0,24 & -1,22 \\ -1,08 & -0,24 & 7,21 & -3,22 \\ 1,14 & -1,22 & -3,22 & 5,43 \end{pmatrix}, \quad b = \begin{pmatrix} 3,52 \\ 1,57 \\ 0,54 \\ -1,09 \end{pmatrix},$$

wird auf die durchdividierte Form $x = Mx + s$ gebracht und nach dem Gesamtschrittverfahren iterativ behandelt. Mit $N = R^n$, $\varrho(x, y) = (|x_i - y_i|)$, $T = |M| = (|m_{ij}|)$ ergeben sich für $x_1 = (1,046293; 0,562746; 0,110962; -0,228068)$ die folgenden Fehlerschranken:

Tabelle 1

k	Fehlerschranken $\delta_1(e_k)$			
	mit $e_0 = (1, 1, 1, 1)$		mit $e_0 = \varrho(x_0, x_1)$	
	e_k nach (5.3)	e_k nach (5.4)	e_k nach (5.3)	e_k nach (5.4)
0	— ¹¹	— ¹¹	0,000 353	0,000 353
	—	—	0,000 299	0,000 299
	—	—	0,000 332	0,000 332
	—	—	0,000 434	0,000 434
1	0,001 531	0,000 503	0,000 353	0,000 347
	0,001 379	0,000 417	0,000 298	0,000 294
	0,001 564	0,000 455	0,000 330	0,000 326
	0,001 761	0,000 631	0,000 435	0,000 427
2	0,000 735	0,000 547	0,000 349	0,000 348
	0,000 595	0,000 450	0,000 296	0,000 295
	0,000 645	0,000 491	0,000 328	0,000 326
	0,000 947	0,000 691	0,000 429	0,000 428

¹¹ Hier ist eine Abschätzung nach (5.2) nicht möglich, da die Bedingung $T e_0 < e_0$ nicht erfüllt ist.

Zum Vergleich sei noch $\varrho(z^*, x_1) = (0,000\ 044; 0,000\ 037; 0,000\ 041; 0,000\ 054)$ angegeben.

Literatur

- [1] ALBRECHT, J.: Monotone Iterationsfolgen und ihre Verwendung zur Lösung linearer Gleichungssysteme. Numer. Math. **3**, 345–358 (1961).
- [2] ALBRECHT, J.: Zur Fehlerabschätzung beim Gesamt- und Einzelschrittverfahren für lineare Gleichungssysteme. Z. angew. Math. Mech. **43**, 83–85 (1963).
- [3] BAHȚIN, I. A., M. A. KRASNOSEL'SKIȚ und V. Ja. STEČENKO: Über die Stetigkeit linearer positiver Operatoren (Russ.). Sibirsk. Mat. Z. **3**, 156–160 (1962).
- [4] BOHL, E.: Nichtlineare Aufgaben in halbgeordneten Räumen. Numer. Math. **10**, 220–231 (1967).
- [5] BOHL, E.: An Iteration Method and Operators of Monotone Type. Arch. Rational Mech. Anal. **29**, 395–400 (1968).
- [6] COLLATZ, L.: Funktionalanalysis und Numerische Mathematik. Berlin-Göttingen-Heidelberg: Springer-Verlag. 1964.
- [7] DUNFORD, N., and J. T. SCHWARTZ: Linear Operators. I. New York: Interscience Publishers. 1958.
- [8] HADELER, K. P.: Einschließungssätze bei normalen und bei positiven Operatoren. Arch. Rational Mech. Anal. **21**, 58–88 (1966).
- [9] KRASNOSEL'SKIȚ, M. A.: Positive Lösungen von Operatorgleichungen. (Russ.). Moskau. 1962.
- [10] MAREK, I.: u_0 -Positive Operators and Some of Their Applications. SIAM J. Appl. Math. **15**, 484–493 (1967).
- [11] ORTEGA, J. M., and W. C. RHEINOLDT: On a Class of Approximate Iterative Processes. Arch. Rational Mech. Anal. **23**, 352–365 (1967).
- [12] SCHAEFFER, H.: Some Spectral Properties of Positive Linear Operators. Pacific J. Math. **10**, 1009–1019 (1960).
- [13] SCHRÖDER, J.: Das Iterationsverfahren bei verallgemeinertem Abstandsbegriff. Math. Z. **66**, 111–116 (1956).
- [14] SCHWETLICK, H.: Spektraleigenschaften linearer positiver Operatoren und Fehlerabschätzungen bei Operatorgleichungen. Diss. Techn. Univ. Dresden. 1967.
- [15] STEČENKO, V. Ja.: Über eine Abschätzung des Spektrums gewisser Klassen linearer Operatoren (Russ.). Dokl. Akad. Nauk SSSR **157**, 1054–1057 (1964).
- [16] VARGA, R. S.: Matrix Iterative Analysis. Englewood Cliffs, N. J.: Prentice Hall. 1962.

Dr. Hubert Schwetlick
Sektion Mathematik
Technische Universität Dresden
DX-8027 Dresden, Zellescher Weg 12–14
Deutsche Demokratische Republik