Beiträge zur Numerischen Mathematik 2

Beiträge zur Numerischen Mathematik 2

Herausgegeben von
Frieder Kuhnert und Jochen W. Schmidt

R. Oldenbourg Verlag München Wien 1974

Lizenz-Nr. 206 · 435/138/74
Gesamtherstellung: VEB Druckhaus „Maxim Gorki", 74 Altenburg
ISBN: 3-486-34411-0

Inhalt

Beiträge zur Numerischen Mathematik
2 (1974), 7–11

Über ein modifiziertes Gesamtschrittverfahren

Hans-Jürgen Albrand

Es sei $Ax = b$ ein lineares Gleichungssystem. Ist A eine symmetrische und positiv definite Matrix, dann konvergiert das Einzelschrittverfahren. Für das Gesamtschrittverfahren (Jacobisches Verfahren) gibt es in diesem Fall keine analoge und ebenso allgemein gehaltene Aussage zur Konvergenz. Bekanntlich konvergiert das Gesamtschrittverfahren nicht für jedes Gleichungssystem mit symmetrischer und positiv definiter Koeffizientenmatrix (man vgl. mit [1, 2]). In dieser Arbeit wird ein modifiziertes Gesamtschrittverfahren vorgestellt, das für jedes Gleichungssystem mit symmetrischer und positiv definiter Koeffizientenmatrix konvergent ist. Das Zeilen- und Spaltensummenkriterium gilt wie beim Jacobi-Verfahren. Der im Vergleich zum Jacobi-Verfahren zusätzliche Rechenaufwand ist nur unerheblich größer.

Es seien H ein Hilbertraum und $U_1, \ldots, U_r$ Unterräume von $L_n = L(x_1, \ldots, x_n) \subset H$ mit

$$L\left(\bigcup_{s=1}^{r} U_s\right) = L_n. \tag{1}$$

$L(x_1, \ldots, x_n)$ bezeichnet die Menge aller Linearkombinationen von $x_1, \ldots, x_n$. Wir setzen für festes k

$$\min_{y \in U_s} \|x^{(k)} - y\| = \|x^{(k)} - y_s^{(k+1)}\| \quad (s = 1, \ldots, r) \tag{2}$$

$$x^{(k+1)} := x^{(k)} - \frac{1}{r} \sum_{s=1}^{r} y_s^{(k+1)} \tag{3}$$

und lassen k die Zahlenfolge 0, 1, 2, ... durchlaufen. Es sei $x^{(0)} = x_0$ ein Element aus H.

Satz 1. *$x_0 - x^{(k)}$ konvergiert gegen die beste Approximation von x_0 in L_n bezüglich der verwendeten Hilbertraumnorm.*

Beweis. Es ist

$$\|x^{(k+1)}\| \leqq \|x^{(k)}\| \tag{4}$$

für jedes $k = 1, 2, \ldots$ Folglich existiert

$$\lim_{k \to \infty} \|x^{(k)}\| =: m. \tag{5}$$

Mit

$$\|x^{(k)} - y_s^{(k+1)}\| = \|x^{(k)}\| - c_{k,s} \quad (s = 1, \ldots, r), \tag{6}$$

$$c_{k,s} \geqq 0$$

wird

$$\|x^{(k+1)}\| \leqq \frac{1}{r} \sum_{s=1}^{r} \|x^{(k)} - y_s^{(k+1)}\| = \|x^{(k)}\| - \frac{1}{r} \sum_{s=1}^{r} c_{k,s},$$

d. h.

$$\sum_{s=1}^{r} c_{k,s} \leqq r(\|x^{(k)}\| - \|x^{(k+1)}\|),$$

und wegen (6)

$$\lim_{k\to\infty} c_{k,s} = 0 \quad (s = 1, \ldots, r). \tag{7}$$

Es sei $x_1^{(s)}, \ldots, x_{l(s)}^{(s)}$ eine orthonormierte Basis von U_s $(s = 1, \ldots, r)$. Dann gilt

$$\|x^{(k)} - y_s^{(k+1)}\|^2 = \|x^{(k)}\|^2 - \sum_{i=1}^{l(s)} (x^{(k)}, x_i^{(s)})^2 \tag{8}$$

$(s = 1, \ldots, r)$. Wegen (6) und (7) folgt aus (8)

$$\lim_{k\to\infty} (x^{(k)}, y) = 0 \tag{9}$$

für alle y aus U_s $(s = 1, \ldots, r)$. Mit (1) ergibt sich aus (9)

$$\lim_{k\to\infty} (x^{(k)}, y) = 0$$

für alle y aus L_n. Ist $e_1, \ldots, e_n$ eine orthonormierte Basis von L_n und $x^{(k)} = x_0 - y^{(k)}$ mit $y^{(k)} = a_1^{(k)} e_1 + \cdots + a_n^{(k)} e_n$, dann folgt

$$\lim_{k\to\infty} a_i^{(k)} = (x_0, e_i)$$

und damit

$$\lim_{k\to\infty} y^{(k)} = \sum_{i=1}^{n} (x_0, e_i)\, e_i,$$

w.z.b.w.

Bemerkung. In (3) kann man anstelle des Faktors $\frac{1}{r}$ auch den Faktor $\frac{1}{r+a}$ mit reellem $a > 0$ schreiben. Der Satz 1 gilt unverändert.

Im folgenden betrachten wir den Spezialfall $U_i = L(x_i)$ $(i = 1, 2, \ldots, n)$. In

$$x^{(k)} = x_0 - d_1^{(k)} x_1 - \cdots - d_n^{(k)} x_n$$

haben die Koeffizienten $d_i^{(k)}$ die Darstellung

$$\begin{aligned} d_i^{(k)} &= \frac{1}{n} \sum_{s=0}^{k-1} \frac{(x^{(s)}, x_i)}{(x_i, x_i)} \\ &= \frac{k}{n} \frac{(x_0, x_i)}{(x_i, x_i)} - \frac{1}{n} \frac{1}{(x_i, x_i)} \sum_{s=1}^{k-1} \sum_{r=1}^{n} d_r^{(s)} \frac{(x_i, x_r)}{(x_i, x_i)} \end{aligned} \tag{10}$$

$(i = 1, 2, \ldots, n)$. Mit

$$d^{(s)\mathsf{T}} = (d_1{}^{(s)}, \ldots, d_n{}^{(s)}) \quad (s = 1, 2, \ldots, n)$$

und

$$z_i{}^{\mathsf{T}} = ((x_i, x_1), \ldots, (x_i, x_k))$$

ergibt sich aus (10)

$$d^{(k)} = \frac{1}{n}\begin{pmatrix} k\dfrac{(x_0, x_1)}{(x_1, x_1)} - \dfrac{1}{(x_1, x_1)}\sum\limits_{s=1}^{k-1} d^{(s)\mathsf{T}} z_1 \\ \cdots\cdots\cdots\cdots \\ k\dfrac{(x_0, x_n)}{(x_n, x_n)} - \dfrac{1}{(x_n, x_n)}\sum\limits_{s=1}^{k-1} d^{(s)\mathsf{T}} z_n \end{pmatrix},$$

d. h.

$$d^{(k+1)} = d^{(k)} + \frac{1}{n}\begin{pmatrix} \dfrac{(x_0, x_1)}{(x_1, x_1)} \\ \vdots \\ \dfrac{(x_0, x_n)}{(x_n, x_n)} \end{pmatrix} - \frac{1}{n}\begin{pmatrix} \dfrac{d^{(k)\mathsf{T}} z_1}{(x_1, x_1)} \\ \vdots \\ \dfrac{d^{(k)\mathsf{T}} z_n}{(x_n, x_n)} \end{pmatrix}$$

oder

$$d^{(k+1)} = \frac{1}{n}(-L - R)\, d^{(k)} + \frac{n-1}{n}\, d^{(k)} + \frac{1}{n} D^{-1} b. \tag{11}$$

In (11) bedeuten

$$L = \begin{pmatrix} 0 & 0 & \ldots & 0 & 0 \\ \dfrac{(x_2, x_1)}{(x_2, x_2)} & 0 & \ldots & 0 & 0 \\ \vdots & \vdots & & \vdots & \vdots \\ \dfrac{(x_n, x_1)}{(x_n, x_n)} & \dfrac{(x_n, x_2)}{(x_n, x_n)} & \ldots & \dfrac{(x_n, x_{n-1})}{(x_n, x_n)} & 0 \end{pmatrix},$$

$$R = \begin{pmatrix} 0 & \dfrac{(x_1, x_2)}{(x_1, x_1)} & \ldots & \dfrac{(x_i, x_n)}{(x_i, x_1)} \\ \vdots & \vdots & & \vdots \\ 0 & 0 & \ldots & \dfrac{(x_{n-1}, x_n)}{(x_{n-1}, x_{n-1})} \\ 0 & 0 & \ldots & 0 \end{pmatrix},$$

$$D^{-1} = \begin{pmatrix} (x_1, x_1)^{-1} & & 0 \\ \vdots & \ddots & \vdots \\ 0 & \ldots & (x_n, x_n)^{-1} \end{pmatrix}, \qquad b = \begin{pmatrix} (x_0, x_1) \\ \vdots \\ (x_0, x_n) \end{pmatrix}.$$

Die Iterationsvorschrift (11) stimmt also fast mit der Jacobischen Vorschrift

$$d^{(k+1)} = (-L - R)\, d^{(k)} + D^{-1}b \tag{12}$$

für das Gleichungssystem

$$(x_i, x_1)\, d_1 + \cdots + (x_i, x_n)\, d_n = (x_0, x_i) \quad (i = 1, 2, \ldots, n) \tag{13}$$

überein. Nun können wir jedes Gleichungssystem $Ad = b$ mit symmetrischer und positiv definiter Koeffizientenmatrix $A = (a_{ij})$ in der Form (13) darstellen, denn

$$x^{\mathrm{T}}Ay = (x, y) \tag{14}$$

ist ein Skalarprodukt. Setzen wir $x_i = e_i \quad (i = 1, \ldots, n)$, wobei e_i in der i-ten Komponente eine 1 hat und in den übrigen Komponenten 0 ist, so wird

$$(e_i, e_j) = a_{ij},$$

und mit $x_0 = b$ geht (13) bei Verwendung des Skalarproduktes (14) über in $Ad = b$. Aus dem Satz 1 folgt also der

Satz 2. *Das modifizierte Gesamtschrittverfahren* (11) *konvergiert für jede symmetrische und positiv definite Koeffizientenmatrix gegen die Lösung des Gleichungssystems.*

Es sei nun $Ad = b$ ein beliebiges Gleichungssystem mit der Koeffizientenmatrix $A = (a_{ij})$ vom Format $n \times n$. Dann gilt der

Satz 3. *Für die Konvergenz der Iterationsvorschrift* (11) *ist hinreichend*

$$\text{a)} \quad \sum_{l \neq i} \left| \frac{a_{il}}{a_{ii}} \right| < 1 \text{ für } i = 1, 2, \ldots, n$$

bzw.

$$\text{b)} \quad \sum_{l \neq i} \left| \frac{a_{li}}{a_{ii}} \right| < 1 \text{ für } i = 1, 2, \ldots, n.$$

Beweis. Aus

$$\sum_{l \neq i} \left| \frac{1}{n} \frac{a_{il}}{a_{ii}} \right| + \frac{n-1}{n} < 1 \quad (i = 1, 2, \ldots, n)$$

folgt a). Entsprechend folgt b), w.z.b.w.

Die Iterationsvorschrift (11) stellt also ein modifiziertes Gesamtschrittverfahren dar, für das ebenso wie beim Jacobi-Verfahren die bekannten Zeilen- und Spaltensummenkriterien gelten. Gegenüber dem Jacobi-Verfahren aber konvergiert das modifizierte Gesamtschrittverfahren für jedes Gleichungssystem mit symmetrischer und positiv definiter Koeffizientenmatrix. Der zusätzliche Rechenaufwand bei (11) gegenüber (12) ist offenbar gering.

Literatur

[1] FORSYTHE, G. E., and W. R. WASOW, Finite-difference methods for partial differential equations, J. Wiley & Sons, New York 1960.
[2] SCHWARZ, H. R., H. RUTISHAUSER und E. STIEFEL, Numerik symmetrischer Matrizen, B. G. Teubner, Stuttgart 1968/ BSB B. G. Teubner Verlagsgesellschaft, Leipzig 1969.

Manuskripteingang: 6. 7. 1972

VERFASSER:

Dr. rer. nat. HANS-JÜRGEN ALBRAND, Sektion Mathematik der Universität Rostock

Beiträge zur Numerischen Mathematik
2 (1974), 13–18

Automatische Lösung algebraischer Gleichungen

Karl-Heinz Bachmann

Zur Lösung algebraischer Gleichungen wird, falls alle Wurzeln gesucht sind, häufig die Methode benutzt, sukzessive einzelne Wurzeln zu bestimmen und durch Division den Grad des Polynoms zu erniedrigen. Die Wurzeln werden dann jeweils als Nullstellen der einzelnen Polynome der so erhaltenen Polynomfolge bestimmt. Durch dabei auftretende Akkumulation von Fehlern können die Nullstellen eventuell stärker von den eigentlich gesuchten Wurzeln abweichen. In [6] wurde auf die Notwendigkeit hingewiesen, Fehlerbetrachtungen in die Rechnung einzubeziehen, z. B. durch Verwendung einer Fehlerschrankenarithmetik.

Die hier vorgeschlagene Methode zur automatischen Lösung algebraischer Gleichungen sei im folgenden erläutert. Vom jeweils betrachteten Polynom wird eine Näherung für eine Nullstelle mit einem modifizierten Abstiegsverfahren bestimmt. Das Abstiegsverfahren wurde häufig untersucht, von Lipschitz wurde es bereits zum Beweis des Fundamentalsatzes der Algebra benutzt [8, 3, 1, 5, 7]. Für die gefundene Näherung wird eine Fehlerabschätzung unter Benutzung des Ausgangspolynoms ausgeführt und geprüft, ob die erhaltenen Fehlerschranken kleiner als eine vorgeschriebene Toleranzgröße sind. Dabei wird gleichzeitig die Vielfachheit der betreffenden Nullstelle bestimmt, wobei innerhalb der Toleranzgrenzen nicht mehr trennbare einfache Nullstellen zu einer mehrfachen zusammengefaßt werden. Sind die erhaltenen Fehlerschranken zu groß, so wird die Rechengenauigkeit erhöht. Ist auf diese Weise eine Nullstelle genügend genau bestimmt, so wird ein entsprechender Linearfaktor vom betrachteten Polynom abgespaltet und das Verfahren wiederholt. Die Eigenschaften der verwendeten Rechenanlage begrenzen dabei die erreichbare Rechengenauigkeit. Ist infolge der Fehlerfortpflanzung ein durch mehrfache Abspaltung von Linearfaktoren erhaltenes Polynom zu ungenau, so wird das daran erkannt, daß eine genügend gute Fehlerschranke trotz Genauigkeitssteigerung nicht erreichbar ist. In diesem Fall werden, sofern die verfügbare Stellenzahl es zuläßt, alle vorher berechneten Nullstellen mit höherer Genauigkeit verbessert, um den Einfluß der Fehlerfortpflanzung zu reduzieren. Es ist zu empfehlen, zwei Genauigkeitsstufen zu benutzen, eine globale Genauigkeitsstufe zur Abspaltung der Linearfaktoren und eine lokale Genauigkeitsstufe, die bei der Durchführung des Abstiegsverfahrens vorübergehend erhöht werden kann.

Zur Fehlerabschätzung werden folgende auf dem Satz von Rouché basierende Sätze herangezogen [2]:

Satz 1. *Das Polynom* $g(w) = \sum_{i=0}^{n} c_i w^i$ *hat in* $|w| \leqq \tilde{r}$ *genau* k *Nullstellen, wenn* $\tilde{r}$ *die kleinste einfache und positive Nullstelle des Polynoms* $h_k(x) = \sum_{i \neq k} |c_i|\, x^i - |c_k|\, x^k$ *ist, sofern diese existiert. Dabei sei* $c_0 \neq 0$ *vorausgesetzt.*

Satz 2. *Mit* $m_i \leqq |c_i| \leqq M_i$ *gilt, daß eine solche Nullstelle* $\tilde{r}$ *von* $h_k(x)$ *existiert, wenn* $h_k^*(x) = \sum_{i \neq k} M_i x^i - m_k x^k$ *eine ebensolche Nullstelle* r^* *hat, und es gilt* $\tilde{r} \leqq r^*$.

Satz 3. *Eine notwendige Bedingung für die Existenz einer einfachen positiven Nullstelle von* $h_k^*(x)$ *ist für* $k < n$ *das Bestehen der Ungleichung*

$$s_k = \max_{i<k} \sqrt[k-i]{\frac{M_i}{m_k}} < \min_{i>k} \sqrt[i-k]{\frac{m_k}{M_i}} = S_k.$$

Im Fall $k = n$ *existiert stets eine solche Nullstelle.*

Satz 4. *Eine hinreichende Bedingung für die Existenz einer einfachen positiven Nullstelle von* $h_k^*(x)$ *ist für* $k < n$ *das Bestehen der Ungleichung*

$$2t_k = 2 \cdot \max_{i<k} \sqrt[k-i]{\frac{2M_i}{m_k}} \leqq \frac{1}{2} \cdot \min_{i>k} \sqrt[i-k]{\frac{m_k}{2M_i}} = \frac{1}{2}\, T_k,$$

und $r = 2t_k$ *ist obere Schranke für diese Nullstelle. Für* $k = 1$ *ist bereits* $t_1 < \frac{1}{2}\, T_1$, *hinreichend, und* t_1 *ist obere Schranke für die betreffende Nullstelle. Für* $k = n$ *existiert stets eine solche Nullstelle, und* $2 \cdot \max_{i<n} \sqrt[n-i]{\frac{M_i}{m_n}}$ *ist obere Schranke für diese.*

Der Spezialfall $k = n$ in Satz 4 ergibt die Schranke von FUJIWARA für alle Nullstellen eines Polynoms [4].

Wird nun das Ausgangspolynom $f(z)$ für eine Nullstellennäherung z_0 nach Potenzen von $w = z - z_0$ entwickelt, so sind, falls z_0 eine Näherung für k dicht beieinanderliegende Nullstellen ist, die ersten k Entwicklungskoeffizienten absolut klein gegenüber dem Entwicklungskoeffizienten c_k. Daher wird die Ungleichung aus Satz 4 erfüllbar sein, wenn z_0 eine genügend gute Näherung ist und mit genügend hoher Stellenzahl gerechnet wird. Um die Vielfachheit k eines solchen Nullstellenbüschels zu bestimmen, wird man das kleinste k suchen, für das die Ungleichung in Satz 3 erfüllt ist. Für $k > 1$ kann es zweckmäßig sein, dann noch eine Näherung für eine Nullstelle der $(k-1)$-ten Ableitung $f^{(k-1)}(z)$ in der Umgebung von z_0 zu bestimmen, da diese angenähert der Schwerpunkt des Nullstellenbüschels ist.

Die Ungleichungen aus Satz 3 und Satz 4 lassen sich mit Hilfe der im folgenden beschriebenen ALGOL-60-Prozedur prüfen:

```
procedure pruefe (n, gm, km, faktor, k, bed, r);
    value n, faktor, k; integer n, k; real faktor, r;
    array gm, km; Boolean bed;
begin real c, c1, tau, tau1, gt; integer j;
        bed := false; c := (if n = k then 1 else faktor) * km[k];
        if c ≠ 0 then
```

```
begin tau := gm[k − 1]/c; for j := 2 step 1 until k do
      begin tau1 := (gm[k − j]/c)↑(1/j);
            if tau1 > tau then tau := tau1
      end; gt := schr * 2;
      comment schr ist eine globale Groeße und bezeichnet eine obere
            Schranke fuer alle Nullstellenbetraege;
      for j := 1 step 1 until n − k do
      begin c1 := gm[k + j]; if c1 > 0 then
            begin tau1 := (c/c1)↑(1/j);
                  if tau1 < gt then gt := tau1
            end
      end;
      r := if k = 1 then tau else tau/faktor;
      bed := r ≦ faktor * gt
end
end;
```

Dabei bezeichnet n den Grad des Polynoms, *gm* den Vektor der M_i, *km* den Vektor der m_i $(i = 0, \ldots, n)$, k die zu prüfende Vielfachheit. Zur Prüfung der Ungleichung in Satz 3 wird *faktor* mit 1 belegt, zur Prüfung der Ungleichung in Satz 4 mit 0.5. In *bed* wird die Erfüllung der Ungleichung festgehalten, in r wird gegebenenfalls die berechnete Schranke vermittelt.

Als Hilfsmittel für die Berechnung der Entwicklungskoeffizienten und von Fehlerschranken für die Entwicklungskoeffizienten dient ein modifiziertes Hornersches Schema. Die im Hornerschen Schema verwendete Rekursionsformel lautet $a_k' = a_k + a_{k+1}' \cdot x$. Nimmt man an, daß a_k und a_{k+1}' mit Fehlern d_k und d_{k+1}' behaftet sind, so entsteht unter Berücksichtigung von Rundungsfehlern e_m für die Multiplikation und e_a für die Addition das Resultat

$$a_k' + d_k' = a_k + d_k + (a_{k+1}' + d_{k+1}')\, x + e_m + e_a.$$

Es gilt also

$$d_k' = d_k + d_{k+1}' \cdot x + e_m + e_a.$$

Dabei wird angenommen, daß das Hornersche Schema für ein genau bekanntes Argument x auszuführen ist. Kennt man obere Schranken für die Absolutbeträge der Fehler, und zwar

$$|d_{k+1}'| \leqq D_{k+1}', \quad |d_k| \leqq D_k, \quad |e_m| \leqq E_m, \quad |e_a| \leqq E_a,$$

so erhält man

$$|d_k'| \leqq D_k' = D_k + D_{k+1}' \cdot x + E_m + E_a.$$

Der übliche Ablauf der Rechnung ist nun so, daß zunächst das in der Rekursionsformel enthaltene Produkt gebildet und dessen Rundungsfehler abgeschätzt wird, der im allgemeinen eine halbe Einheit der letzten Stelle des berechneten Produktes nicht übersteigt. Darauf folgt die in der Rekursionsformel enthaltene Addition, bei der ebenfalls ein Rundungsfehler auftreten kann. Dazu wird angenommen, daß die Summanden Gleitkommazahlen gleicher Mantissenlänge sind. Von beiden Summanden wird eine Einheit der letzten Stelle gebildet. Die größere dieser beiden Einheiten

legt fest, welcher Fehler beim Angleichen der Exponenten höchstens gemacht werden kann, und zwar eine halbe Einheit dieser Stelle, falls korrekt gerundet wird. Sind beide Einheiten gleich, so brauchen die Exponenten nicht angeglichen zu werden, es tritt dann auch kein Fehler auf. Bei der Addition kann eine um eine Stelle längere Mantisse entstehen, durch Verkürzen dieser Mantisse um eine Stelle tritt dann ein zusätzlicher Rundungsfehler in Größe einer halben Einheit der letzten Resultatstelle auf. Die Verhältnisse seien an einem Beispiel in vierstelliger dezimaler Gleitkommarechnung erläutert.

Zu addieren seien $a_1 = 0{,}9900 \cdot 10^2$ und $a_2 = 0{,}6045 \cdot 10^1$. Die Einheiten der letzten Stellen sind $e_1 = 10^{-2}$ und $e_2 = 10^{-3}$. Die Addition erfolgt in der Form $\hat{a}_3 = 0{,}9900 \cdot 10^2 + 0{,}0605 \cdot 10^2 = 1{,}0505 \cdot 10^2$. Dieses Resultat wird mit vierstelliger Mantisse als $a_3 = 0{,}1051 \cdot 10^3$ dargestellt. Die Einheit der letzten Resultatstelle ist $e_3 = 10^{-1}$. Der Fehler gegenüber dem korrekten Resultat 105,045 ist 0,055 $= \frac{1}{2}(e_3 + \max(e_1, e_2))$.

Bei Rechnung im Dualsystem oder in einem anderen Zahlensystem sind die Verhältnisse entsprechend. Bei dezimaler Rechnung mit achtstelliger Mantisse kann durch die ALGOL-60-Prozedur

```
real procedure deltax(x); value x; real x;
deltax := if x = 0 then 0 else 10↑(entier(.43429448 ln(abs(x)) + 10—7)—7);
```

eine Einheit der letzten Stelle bestimmt werden. Es empfiehlt sich allerdings, hierfür eine Prozedur im Maschinencode zu verwenden. Der Rundungsfehler f bei Addition läßt sich in der beschriebenen Form durch die folgende ALGOL-60-Prozedur berechnen:

```
procedure rund (e1, e2, e3, f); value e1, e2, e3; real e1, e2, e3, f;
begin real e,ee; e  := if e1 > e2 then e1 else e2;
                 ee := if e1 = e2 then 0 else e;
                 f  := 0.5 * (if e3 > e then ee + e3 else ee) end;
```

Mit Hilfe dieser Prozeduren läßt sich ein vollständiges Hornersches Schema zur Berechnung der Entwicklungskoeffizienten mit zugehörigen Fehlerschranken aufbauen, für komplexe Rechnungen lassen sich diese Betrachtungen leicht erweitern. Auf die Darstellung von Einzelheiten sei hier verzichtet.

Ein mit diesen Hilfsmitteln arbeitendes Verfahren zur Nullstellenbestimmung sei im folgenden skizziert: Zunächst wird die erwähnte Schranke von FUJIWARA für die Nullstellenbeträge berechnet. Ausgehend von einem Anfangswert, der beliebig gewählt werden kann und hier entweder zu 0 oder auf eine vorher bekannte Nullstellennäherung festgelegt wird, wird ein Abstiegsverfahren durchgeführt. Dieses führt entweder zu einer Nullstelle oder zu einem Sattelpunkt der Betragsfläche. Eine Nullstellennäherung wird als erreicht angesehen, wenn entweder der Funktionswert exakt 0 ist oder der Korrekturbetrag bei einem Newtonschritt kleiner als eine relativ zum Betrag des Näherungswertes gewählte Schranke ist. Ist es nicht möglich, den Funktionsbetrag bei einem Abstiegsschritt auf weniger als das 0,8fache zu verkleinern, so wird angenommen, daß die Näherung sich entweder in der Nähe eines Sattelpunktes der Betragsfläche oder in der Umgebung einer Nullstelle be-

findet. Dann werden Entwicklungskoeffizienten mit Fehlerschranken berechnet, und an Hand der Fehlerschranke für den Funktionswert wird entschieden, ob der betreffende Punkt als Nullstellennäherung angesehen wird. Ist nämlich der Funktionsbetrag kleiner als das Doppelte der berechneten Fehlerschranke, so kann die übliche Newtonkorrektur mit einer solchen Unsicherheit behaftet sein, daß eine Weiterführung der Rechnung ohne Genauigkeitserhöhung nicht mehr sinnvoll ist. Andernfalls wird versucht, den Funktionsbetrag unter Verwendung der zweiten Ableitung wie in [1] beschrieben zu verkleinern, wobei gegebenenfalls die Richtung des stärksten Abstiegs verlassen wird. Dabei zeigt es sich, daß die Umgebung eines Sattelpunktes relativ schnell verlassen werden kann, wenn auf einem vom Sattelpunkt in Richtung auf kleinere Funktionsbeträge wegführenden Strahl die Korrekturbeträge sukzessive verdoppelt werden.

Sollte sich ein nach dem erwähnten Verfahren theoretisch auffindbarer Punkt mit kleinerem Funktionsbetrag als im Sattelpunkt durch den Einfluß von Rundungsfehlern nicht finden lassen, so ist die Genauigkeit zu erhöhen. Ist das nicht mehr möglich, so ist die Umgebung des Sattelpunktes systematisch nach einem solchen Punkt (z. B. auf einer Spirale) zu durchsuchen.

Ist auf diese Weise ein als Nullstellennäherung anzusehender Punkt gefunden, so wird wie oben beschrieben eine Fehlerabschätzung unter Einschluß der Vielfachheitsbestimmung ausgeführt, und nach Abspaltung des entsprechenden Linearfaktors wird das Verfahren wiederholt. Bei Polynomen mit reellen Koeffizienten und komplexen Näherungen wird man auch versuchen, den konjugiert komplexen Linearfaktor abzuspalten. Dabei darf jedoch der Rest der Division nicht zu groß werden. Als Kriterium dafür kann wieder die für den Funktionswert bestimmte Fehlerschranke benutzt werden. Gegebenenfalls muß zur komplexen Rechnung übergegangen werden. Bei Polynomen mit reellen Koeffizienten können auch kleine Imaginärteile von Nullstellennäherungen vorgetäuscht werden, daher ist in einem solchen Fall erst zu prüfen, ob eine reelle Näherung in Frage kommt. Eine versuchte Abspaltung des konjugiert komplexen Linearfaktors führt dann im allgemeinen zu einer starken Nullabweichung des Divisionsrestes.

Durch Wiederholung dieses Verfahrens werden für ein Polynom n-ten Grades insgesamt n Näherungen — entsprechend ihrer Vielfachheit gezählt — für Nullstellen bestimmt und ihre Fehler abgeschätzt. Sind die durch die Fehlerabschätzung festgelegten Fehlerkreise mit den Nullstellennäherungen als Mittelpunkten disjunkt, so hat man mit Sicherheit Kreise gefunden, in denen genau die durch die berechnete Vielfachheit angegebene Zahl von Nullstellen des Ausgangspolynoms liegt. Ist das nicht der Fall, so war die Rechengenauigkeit noch nicht ausreichend.

Literatur

[1] Bachmann, K.-H., Lösung algebraischer Gleichungen nach der Methode des stärksten Abstieges, ZAMM **40** (1960), 132—135.

[2] Bachmann, K.-H., Fehlerabschätzung für Nullstellen von Polynomen, Vorträge zur Numerischen Verfahrenstechnik, TU Dresden, Heft 1/71 (1973), 82—88.

[3] Brooker, R. A., The solution of algebraic equations on the EDSAC, Proc. Cambr. philos. Soc. **48** (1952), 255—270.

[4] FUJIWARA, M., Über die obere Schranke des absoluten Betrags der Wurzeln einer algebraischen Gleichung, Tohoku math. J., Ser. **10** (1916), 167—170.
[5] NASITTA, K., Ein immer konvergentes Nullstellenverfahren für analytische Funktionen, ZAMM **44** (1964), 57—63.
[6] NICKEL, K. Über die Notwendigkeit einer Fehlerschrankenarithmetik für Rechenautomaten, Numer. Math. **9** (1966), 69—79.
[7] NICKEL K., Die numerische Berechnung der Wurzeln eines Polynoms, Numer. Math. **9** (1966), 80—98.
[8] WEBER, H., Lehrbuch der Algebra, Bd. 1, Braunschweig 1898, S. 143—147.

Manuskripteingang: 25. 7. 1972

VERFASSER:

Dr. rer. nat. habil. KARL-HEINZ BACHMANN, Zentralinstitut für Mathematik und Mechanik der Akademie der Wissenschaften der DDR

Beiträge zur Numerischen Mathematik
2 (1974), 19–23

Stabilitätsuntersuchung eines verallgemeinerten Iterationsprozesses zur Lösung linearer Gleichungssysteme

Michael Fröhner

1. Einleitung

Löst man das lineare Gleichungssystem $Lx = g$ nach Umformung auf die Form $x = Bx + b$ mit Hilfe eines dreigliedrigen Iterationsprozesses der Gestalt $x_k = f_k(B, b, x_{k-1}, x_{k-2})$, so ist eine Rückführung dieser Vorschrift auf das Kontraktionsprinzip im allgemeinen nicht mehr möglich. Es besteht deshalb das Problem, die Fortpflanzung eingeschleppter Fehler, d. h. die Stabilität des Lösungsverfahrens der als Differenzengleichung aufzufassenden Iteration, zu untersuchen. Es wird gezeigt, daß das Iterationsverfahren unter den Bedingungen, unter denen es konvergiert, auch stabil ist, und weiterhin wird eine Fehlerschranke angegeben.

Von der Matrix L sei bekannt, daß alle Eigenwerte im Intervall (m, M), $m > 0$, $M > 0$, liegen. Erzeugt man die iterierfähige Gestalt durch die Umformungen $B = E - 2 \cdot L/(M + m)$, $b = 2 \cdot g/(M + m)$, so liegen die Eigenwerte von B symmetrisch zum Nullpunkt im Intervall $\left(-\frac{M-m}{M+m}; +\frac{M-m}{M+m}\right)$, d. h., für den Spektralradius gilt $\varrho(B) < 1$. Wir betrachten die Iterationsvorschrift

$$\left.\begin{aligned} x_k &= \alpha_k(B \cdot x_{k-1} + b) - \beta_k x_{k-2}, \\ x_1 &= Bx_0 + b \end{aligned}\right\} \beta_k = \alpha_k - 1, \quad k = 2, 3, \ldots \tag{1}$$

Das Verfahren (1) konvergiert für $\varrho(B) < 1$ und $0 < \alpha_k < 2$ (vgl. [1]). Für $\alpha_k = 1$ erhält man das Gesamtschrittverfahren, und für $\alpha_k = 1 + T_{k-2}\left(\frac{M+m}{M-m}\right) \Big/ T_k\left(\frac{M+m}{M-m}\right)$ erhält man einen für Gleichungssysteme beliebiger Ordnung bezüglich der Anzahl der Iterationsschritte optimal konvergenten Prozeß [1]. Dabei ist $T_k(t)$ das k-te Tschebyscheff-Polynom. Wie man leicht sieht, konvergiert die Folge der α_k sehr rasch gegen einen Grenzwert $\alpha \in (1; 2)$.

Bei der Vorschrift (1) handelt es sich um eine Differenzengleichung zweiter Ordnung mit (im allgemeinen) variablen Koeffizienten, für die der Einfluß von Störungen (Rundungsfehler) untersucht werden soll. Wir wollen zusätzlich zu (1) die gestörte Gleichung betrachten:

$$\left.\begin{aligned} \tilde{x}_k &= \alpha_k(B \cdot \tilde{x}_{k-1} + b) - \beta_k \tilde{x}_{k-2} + \tilde{\varrho}_k, \\ \tilde{x}_1 &= B \cdot \tilde{x}_0 + b + \varrho_1 \end{aligned}\right\} \beta_k = \alpha_k - 1, \quad k = 2, 3, \ldots \tag{1'}$$

Bezeichnen wir den Fehler mit $\tilde{z}_i = \tilde{x}_i - x_i$ und nehmen wir an, daß B ein vollständiges System von Eigenvektoren besitze, so daß eine Transformationsmatrix U der Gestalt

$$U^{-1}BU = \Lambda = \operatorname{diag}(\lambda_1, \lambda_2, \ldots, \lambda_m)$$

existiert, so erhält man mit $z_k = U^{-1}\tilde{z}_k$, $\varrho_k = U^{-1}\tilde{\varrho}_k$ die Gleichungen

$$\left.\begin{aligned} z_k &= \alpha_k \Lambda z_{k-1} - \beta_k z_{k-2} + \varrho_k, \\ z_1 &= \varrho_1, \\ z_0 &= 0 \end{aligned}\right\} \tag{2}$$

oder in Komponenten

$$\left.\begin{aligned} z_k^{(i)} &= \alpha_k \lambda_i z_{k-1}^{(i)} - \beta_k z_{k-2}^{(i)} + \varrho_k^{(i)}, \\ z_1^{(i)} &= \varrho_1^{(i)}, \\ z_0^{(i)} &= 0 \end{aligned}\right\} \qquad \begin{aligned} i &= 1(1)m, \\ k &= 2, 3, \ldots, \\ \beta_k &= \alpha_k - 1. \end{aligned} \tag{2'}$$

Die Transformation von (2′) auf ein System von Differenzengleichungen erster Ordnung liefert

$$\left.\begin{aligned} Z_k &= A_k \cdot Z_{k-1} + R_k, \\ Z_1 &= R_1 \end{aligned}\right\} \qquad k = 2, 3, \ldots, \tag{3}$$

wobei zur Abkürzung

$$Z_k = \begin{pmatrix} z_k^{(i)} \\ z_{k-1}^{(i)} \end{pmatrix}, \qquad A_k = \begin{pmatrix} \alpha_k \lambda_i & -\beta_k \\ 1 & 0 \end{pmatrix}, \qquad R_k = \begin{pmatrix} \varrho_k^{(i)} \\ 0 \end{pmatrix}$$

gesetzt wurde. Die allgemeine Lösung lautet

$$Z_k = \prod_{p=k}^{2} A_p \cdot Z_1 + \sum_{t=2}^{k-1} \left(\prod_{p=k}^{t+1} A_p \right) R_t + R_k. \tag{4}$$

2. Der stationäre Prozeß

Ist $\alpha_k = \alpha = \text{const}$, so erhalten wir aus (4) die Lösung

$$Z_k = A^{k-1} Z_1 + \sum_{p=2}^{k-2} A^p \cdot R_{k-p}, \quad A = \begin{pmatrix} \alpha\lambda_i & -\beta \\ 1 & 0 \end{pmatrix}.$$

Die Matrix A hat die Eigenwerte $t_{1/2} = (\alpha\lambda_i)/2 \pm \sqrt{(\alpha\lambda_i/2)^2 - \beta}$, die unter den Voraussetzungen $\lambda_i \in (-1; 1)$, $0 < \alpha < 2$, $\beta = \alpha - 1$ betragsmäßig kleiner als Eins sind. Für diesen Fall wächst aber die Lösung des homogenen Systems von Differenzengleichungen bei auftretenden Störungen $\tilde{\varrho}_k^{(i)}$ nicht an (d. h., für $k \to \infty$ verschwindet der erste Term in (4), und somit ist der stationäre Iterationsprozeß stabil.

3. Ein nichtstationärer Prozeß

Wir betrachten den Fall $\alpha_k \to \alpha$, der z. B. den praktisch bedeutenden und eingangs erwähnten Prozeß der optimalen Konvergenz mit umfaßt. Für die Matrix A mit dem Grenzparameter α und $\lambda_i \in (-1; 1)$ sind die Eigenwerte kleiner als Eins, folglich existiert immer eine Transformationsmatrix H, so daß für $D = H^{-1}AH$ die Beziehung $\|D\| \leqq q < 1$ gilt, falls nur $0 < \underline{\alpha} \leqq \alpha \leqq \bar{\alpha} < 2$ und $|\lambda_i| \leqq \bar{\lambda} < 1$ ist.

Werden die A_k ebenfalls mit H transformiert, so läßt sich die Darstellung

$$\tilde{D}_k = H^{-1}A_kH = D + (\alpha - \alpha_k) \cdot K(\alpha, \lambda_i)$$

finden, wobei die Matrix K nur von λ_i und $\alpha = \lim\limits_{k\to\infty} \alpha_k$ abhängt. $\tilde{D}_k$ ist also eine „Fastdiagonalmatrix", für deren Eigenwerte man die Beziehung $t_i + \delta t_i$ angeben kann, wenn t_i Eigenwerte von A und wobei

$$\delta t_i = (\alpha - \alpha_k) \cdot \text{const}$$

ist. Damit verschwindet auch in diesem Fall die Lösung des homogenen Systems (4), und für den untersuchten Prozeß liegt ebenfalls Stabilität vor, falls $k \to \infty$ gilt.

4. Fehlerabschätzung

Wir wenden uns sofort dem nichtstationären Fall zu, der den stationären Prozeß als Spezialfall mit enthält. Wegen $\|D\| \leqq q < 1$ kann immer ein k_0 gefunden werden, so daß für alle $k > k_0$ auch $\|\tilde{D}_k\| < 1$ gilt. Hier soll für k_0 diejenige Zahl gewählt werden, für die sich α_{k_0} und α auf Grund der Maschinengenauigkeit der verwendeten Rechenanlage nicht mehr unterscheiden, d. h., für alle $k > k_0$ gilt $\tilde{D}_k = D$. Durch Transformation von (4), Einführung einer Norm und Rücksubstitution aller Hilfsgrößen erhält man für den Grenzwert des Fehlers beim betrachteten Iterationsprozeß

$$\limsup_{k\to\infty} \|\tilde{x}_k - x^*\| \leqq \|U^{-1}\| \cdot \|\tilde{\varrho}\| \cdot \|H\| \max_i 1/(1 - \|D^{(i)}\|),$$

wobei $D^{(i)}$ aus der zum Eigenwert λ_i gehörenden Matrix A_k (bzw. A) hervorgegangen ist. Die Norm von H kann im einfachsten Fall durch 2 abgeschätzt werden.

Weiterhin ist

$$\|D^{(i)}\| = \max\{|t_1|, |t_2|\} \quad \text{mit} \quad t_{1/2} = (\alpha\lambda_i)/2 \pm \sqrt{(\alpha\lambda_i/2)^2 - \beta}.$$

x^* ist die exakte Lösung des vorgegebenen Gleichungssystems. Im Fall einer symmetrischen Koeffizientenmatrix B vereinfacht sich die Abschätzung wegen $U^{-1} = U^{\mathsf{T}}$.

Bemerkung. Die Abschätzung für das gewöhnliche Gesamtschrittverfahren ($\alpha = 1$) stimmt mit einer Abschätzung überein, die auf anderem Wege über das Fixpunktprinzip erhalten wurde. Es gilt dann $\|D^{(i)}\| = |\lambda_i|$.

5. Numerische Ergebnisse

Nach dem angegebenen Iterationsprozeß wurden lineare Gleichungssysteme der Dimension 3 bis 6 mit unterschiedlichen Eigenwertverteilungen der Koeffizientenmatrix und Lösungen in verschiedenen Größenordnungen untersucht, wobei die Fälle $\alpha = \text{const}$, $\alpha_k \to \alpha$ und α_k zufälligen Änderungen im Intervall $(0, 2)$ unterworfen, betrachtet wurden. Die Rechnungen wurden im Gleitkomma mit einer Mantissenlänge von 8 Ziffern ausgeführt.[1]) Für die Abschätzung des Rundungsfehlers $\|\tilde{\varrho}\|$ wurden die von WILKINSON [2] angegebenen Schranken für Skalarprodukte ohne akkumulierende Multiplikation und für Additionen im Gleitkomma unter Berücksichtigung der Matrixstruktur und der Rechenvorschrift des ALGOL-Programmes verwendet.

Die folgenden beiden Tabellen enthalten für zwei symmetrische Gleichungssysteme der Dimension 4 und für ausgewählte Werte von α in den einzelnen Spalten den maximalen Rundungsfehler je Schritt $\|\tilde{\varrho}\|$, den nach obiger Abschätzung berechneten Fehler des Prozesses $\Delta_{\max} = \limsup_{k\to\infty} \|\tilde{x}_k - x^*\|$ und das praktisch erhaltene Fehlerintervall $[\delta]$, das nach N Iterationsschritten erstmalig erreicht und innerhalb weiterer 50 Iterationen nicht mehr verlassen wurde. Der Fehler variierte in diesem Intervall völlig unregelmäßig.

Beispiel 1. Eigenwerte $\lambda_i^B \in (-0{,}75; 0{,}75)$ (Kondition von $L: \varkappa = 6{,}85$)

α	$\|\tilde{\varrho}\| \cdot 10^7$	$\Delta_{\max} \cdot 10^7$	$[\delta] \cdot 10^7$	N
1,0	24	192	(13 ··· 52)	64
1,2	30	108	(8 ··· 51)	30
1,2*	30	108	(17 ··· 63)	26
1,4	36	201	(11 ··· 118)	42

Beispiel 2. Eigenwerte $\lambda_i^B \in (-0{,}8112; 0{,}8112)$ (Kondition von $L: \varkappa = 9{,}6$)

α	$\|\tilde{\varrho}\| \cdot 10^7$	$\Delta_{\max} \cdot 10^7$	$[\delta] \cdot 10^7$	N
1,0	0,5	5	(1 ··· 2)	74
1,1	0,6	4,6	(0,8 ··· 1,5)	57
1,262*	0,6	2,7	(0,5 ··· 1)	26
1,5	0,8	10,4	(2 ··· 6)	44

Anmerkung. Die Werte für (*) wurden aus dem optimal konvergenten Prozeß erhalten.

Bemerkung. Der beschriebene Iterationsprozeß bildet im Sinne von BABUŠKA eine α_0-L-Folge (vgl. hierzu Artikel von FRIEDRICH und MÜLLER in Heft 3 dieser Schriftenreihe).

[1]) Rechenanlage R 300 des Rechenzentrums der TH Karl-Marx-Stadt.

Literatur

[1] FADDEJEW, D. K., und W. N. FADDEJEWA, Numerische Methoden der linearen Algebra, 3. Aufl., VEB Deutscher Verlag der Wissenschaften, Berlin/R. Oldenbourg Verlag, München—Wien 1973 (Übersetzung aus dem Russischen).

[2] WILKINSON, J. H., Rundungsfehler, Springer-Verlag, Berlin—Heidelberg—New York 1969 (Übersetzung aus dem Englischen).

Manuskripteingang: 12. 7. 1972

VERFASSER:

Dipl.-Math. MICHAEL FRÖHNER, Sektion Mathematik der Technischen Hochschule Karl-Marx-Stadt