A priori Fehlerschranken für sukzessiv abgespaltene Polynomnullstellen

Von Martin Gutknecht, Seminar für angewandte Mathematik, Eidgenössische Technische Hochschule Zürich[1])

1. Einleitung

Wir gehen aus von einem Polynom

$$p_N(z) := z^N + a_1 z^{N-1} + \cdots + a_N$$

mit komplexen Koeffizienten und von einer Zahl $\varepsilon > 0$. Zudem sei ein numerisches Verfahren gegeben, das gestattet eine komplexe Zahl ω_N zu finden, für die gilt

$$|p_N(\omega_N)| \leqslant \varepsilon .$$

ω_N kann als Approximation einer Nullstelle von $p_N(z)$ aufgefasst werden. An Stelle der unbekannten Nullstelle spalten wir die Approximation ω_N ab, d.h. wir dividieren $p_N(z)$ durch den Linearfaktor $z - \omega_N$ und vernachlässigen den Rest (Deflation). Es resultiert ein Polynom $p_{N-1}(z)$ vom Grade $N-1$, von welchem i.a. wieder eine Nullstellenapproximation bestimmt werden kann. Iteration dieses Prozesses ergibt in bekannter Weise N Approximationen $\omega_1, \ldots, \omega_N$ der N Nullstellen $\zeta_1, \ldots, \zeta_N$ von $p_N(z)$. Gesucht sind nun Schranken $T_1, \ldots, T_N$, so dass (bei geeigneter Numerierung der ζ_k) gilt

$$|\omega_k - \zeta_k| \leqslant T_k, \quad k = 1, \ldots, N .$$

2. A priori Schranke

Wird das nach $N-k$ Deflationen erreichte Polynom vom Grade k mit $p_k(z)$ bezeichnet, so gilt

$$p_{k-1}(z) = \frac{p_k(z) - p_k(\omega_k)}{z - \omega_k}, \quad k = 1, \ldots, N . \tag{1}$$

Nur im konstanten Koeffizienten verschieden von $p_k(z)$ ist das Polynom

$$q_k(z) := p_k(z) - p_k(\omega_k) = (z - \omega_k)\, p_{k-1}(z), \quad k = 1, \ldots, N . \tag{2}$$

Bezeichnen wir die Nullstellen von $p_k(z)$ mit $\zeta_{k1}, \ldots, \zeta_{kk}$, so hat $q_k(z)$ gerade die Nullstellen $\zeta_{k-11}, \ldots, \zeta_{k-1k-1}, \omega_k$. Um den Abstand dieser Nullstellen abzuschätzen,

[1]) Für zahlreiche Diskussionen und Verbesserungen bin ich den Herren Prof. Dr. P. Henrici und dipl. Math. Rolf Jeltsch zu Dank verpflichtet.

verwenden wir folgende, von Ostrowski [1] bewiesene, quantitative Aussage über die Stetigkeit von Polynomnullstellen.

SATZ A. *Seien* $\zeta_1, \ldots, \zeta_N$ *die Nullstellen von* $p(z) := z^N + a_1 z^{N-1} + \cdots + a_N$ *und* $\omega_1, \ldots, \omega_N$ *die Nullstellen von* $q(z) := z^N + b_1 z^{N-1} + \cdots + b_N$, *wobei gelte* $|\zeta_i| \leqslant R$, $|\omega_i| \leqslant R$, $i = 1, \ldots, N$. *Weiter sei*

$$\Delta := \left(\sum_{j=1}^{N} |b_j - a_j| \, R^{N-j} \right)^{1/N} . \tag{3}$$

Dann liegen in jeder Komponente des abgeschlossenen Bereichs

$$S := \bigcup_{j=1}^{N} \{z \mid |z - \zeta_j| \leqslant \Delta\} \tag{4}$$

gleich viele ω_i *wie* ζ_i. *Zudem existiert zu jedem* ω_i *sogar ein* $\zeta_{s(i)}$ *mit*

$$|\omega_i - \zeta_{s(i)}| \leqslant \Delta , \tag{5}$$

doch ist s i.a. nicht als Permutation wählbar.

Da der Durchmesser einer Komponente des Bereichs S die obere Schranke $2N\Delta$ hat und da jedes ζ_i mindestens um Δ vom Rand der Komponente entfernt ist, gibt es sicher eine Permutation σ, so dass $|\omega_i - \zeta_{\sigma(i)}| \leqslant (2N - 1)\Delta$ ist. Dadurch dass man σ günstig wählt, kann man aber sogar

$$|\omega_i - \zeta_{\sigma(i)}| \leqslant (2i - 1)\Delta , \quad i = 1, \ldots, N , \tag{6}$$

erreichen, z.B. indem man $\sigma(i)$ in der Reihenfolge wachsender i wie folgt festlegt: ω_i wird von den noch nicht reservierten, in derselben Komponente von S liegenden Nullstellen ζ_j diejenige zugeordnet, deren Abstand von ω_i am kleinsten ist.

Weiter ist zu beachten, dass in (6) die Numerierung der ω_i willkürlich ist. Dementsprechend erhalten wir allgemeiner: *Zu einer vorgegebenen Permutation* τ *von* $(1, \ldots, N)$ *existiert eine Permutation* σ, *so dass unter den Voraussetzungen von Satz A gilt*

$$|\omega_i - \zeta_{\sigma(i)}| \leqslant (2\tau(i) - 1)\Delta , \quad i = 1, \ldots, N . \tag{7}$$

Im übrigen treten in den Voraussetzungen von Satz A ω_i und ζ_i symmetrisch auf. In den Formeln (4) bis (7) dürfen also die Buchstaben ω und ζ vertauscht werden.

Die Anwendung von Satz A und (6) auf die eingangs definierten Polynome $p_k(z)$ und $q_k(z)$ ergibt nun unter der Voraussetzung, dass die Nullstellen ζ_{ki} günstig numeriert sind (d.h. so, dass $\sigma(i) = i$ gilt)

$$|\zeta_{k-1,i} - \zeta_{ki}| \leqslant (2i - 1)\Delta_k , \quad i = 1, \ldots, k - 1 ,$$

$$|\omega_k - \zeta_{kk}| \leqslant (2k - 1)\Delta_k ,$$

wobei

$$\Delta_k := |p_k(\omega_k)|^{1/k} , \quad k = 1, \ldots, N . \tag{8}$$

Nach der Dreiecksungleichung folgt nun mit $\zeta_{Nk} = \zeta_k$

$$|\omega_k - \zeta_k| \leqslant |\omega_k - \zeta_{kk}| + |\zeta_{kk} - \zeta_{k+1,k}| + \cdots + |\zeta_{N-1\,k} - \zeta_{Nk}|$$

$$\leqslant (2k-1)\sum_{s=k}^{N} \Delta_s .$$

Wir erhalten damit

SATZ B. *Bei günstiger Numerierung der Nullstellen ζ_k von $p_N(z)$ gilt*

$$|\omega_k - \zeta_k| \leqslant (2k-1)\sum_{s=k}^{N} \Delta_s , \quad k = 1, \ldots, N , \tag{9}$$

mit den in (8) *definierten Δ_s. Insbesondere folgt unter der Voraussetzung*

$$|p_k(\omega_k)| \leqslant \varepsilon , \quad k = 1, \ldots, N , \tag{10}$$

die a priori Abschätzung

$$|\omega_k - \zeta_k| \leqslant T_k , \quad wo \quad T_k := (2k-1)\sum_{s=k}^{N} \varepsilon^{1/s} . \tag{11}$$

Es ist klar, dass man bei Verwendung von (7) zu weiteren ähnlichen Abschätzungen gelangen kann.

Beispiele von Werten der a priori Schranken T_k zu $\varepsilon = 10^{-10}$ und verschiedenen N sind in Tabelle 1 zusammengestellt. Es fällt auf, dass die Werte sehr gross sind im Vergleich zur recht kleinen Toleranz ε. In der Tat existiert für $N > 1$ kein Beispiel, in dem die Schranken wirklich angenommen werden. Doch ist zu beachten, dass beim Übergang von $p_k(z)$ zu $q_k(z)$ Nullstellenverschiebungen um $\varepsilon^{1/k}$ oder mehr durchaus auftreten können (vgl. [1], Appendix A, Abschnitt 10). Dementsprechend ist auch für $\varepsilon \to 0$ die Ordnung $0(\varepsilon^{1/N})$ der Schranken richtig.

Tabelle 1

A priori Schranken T_k aus (11) zu $\varepsilon = 10^{-10}$. Es bezeichnet m denjenigen Index, für den T_k maximal ist.

N	T_1	m	T_m	T_N
1	$1.00\ 10^{-10}$	1	$1.00\ 10^{-10}$	$1.00\ 10^{-10}$
2	$1.00\ 10^{-5}$	2	$3.00\ 10^{-5}$	$3.00\ 10^{-5}$
5	$1.36\ 10^{-2}$	4	$9.21\ 10^{-2}$	$9.00\ 10^{-2}$
10	$3.06\ 10^{-1}$	7	3.52	1.90
20	2.54	13	$4.91\ 10^{1}$	$1.23\ 10^{1}$
50	$1.78\ 10^{1}$	29	$6.92\ 10^{2}$	$6.25\ 10^{1}$
100	$5.43\ 10^{1}$	54	$3.70\ 10^{3}$	$1.58\ 10^{2}$

3. Ergänzungen

Verschärfung von Satz B. Die Herleitung von Satz B beruht auf der ersten Aussage von Satz A, wogegen die Abschätzung (5) nicht benützt wird. Bei Berücksichti-

gung von (5) ist es möglich, folgende zwei schärfere Sätze herzuleiten, deren Beweise einfach, aber recht lang sind, so dass auf die Wiedergabe verzichtet werden soll.

SATZ C. *Zu einer vorgegebenen Permutation* τ *von* $(1, \ldots, N)$ *existiert eine Permutation* σ, *so dass gilt*

$$|\omega_k - \zeta_{\sigma(k)}| \leqslant 2(\tau(k) - 1) \max_{s=k+1, \ldots, N} \Delta_s + \sum_{s=k}^{N} \Delta_s . \tag{12}$$

Insbesondere folgt unter der Voraussetzung (10) *bei günstiger Numerierung der* ζ_k *die a priori Abschätzung*

$$|\omega_k - \zeta_k| \leqslant T'_k , \quad wo \quad T'_k := (2k - 1)\, \varepsilon^{1/N} + \sum_{s=k}^{N-1} \varepsilon^{1/s} . \tag{13}$$

Im Vergleich zu (11) tritt bei (13) der Faktor $(2k - 1)$ nur beim grössten Summanden $\varepsilon^{1/N}$ auf.

SATZ D. *Es sei für* $k = 1, \ldots, N$

$$R_k := \sum_{s=k}^{N} \Delta_s , \; D_k := \{z \mid |z - \omega_k| \leqslant R_k\} , \; D'_k := \{z \mid |z - \zeta_k| \leqslant R_1\}$$

und zudem

$$S := \bigcup_{k=1}^{N} D_k , \qquad S' := \bigcup_{k=1}^{N} D'_k .$$

Dann gilt:

(i) *Jede Kreisscheibe* D_k *enthält mindestens eine Nullstelle* ζ_i *von* $p_N(z)$.
(ii) *Jede Kreisscheibe* D'_k *enthält mindestens eine Approximation* ω_i.
(iii) *Jede Komponente von* S *enthält gleich viele* ζ_i *wie* ω_i.
(iv) *Jede Komponente von* S' *enthält gleich viele* ζ_i *wie* ω_i.

Sind die Nullstellen ζ_i bekannt, so ergeben sich unter der Voraussetzung (10) aus (ii) und (iv) a priori Aussagen über die Approximationen ω_i.

Da die Schranken T'_k für $\varepsilon \leqslant 1$ mit k monoton zunehmen und da $T'_1 = T_1$ und $T'_N = T_N$ gilt, kann man ihre Grössenordnung auch aus Tabelle 1 entnehmen. Ebenso ist das dort aufgeführte T_1 die obere Grenze der Kreisradien R_k, da diese monoton fallen und $R_1 = T_1$ ist.

Relative Fehler. Ostrowski [1] hat auch einen Satz über den relativen Abstand der Nullstellen zweier Polynome bereitgestellt. Mit diesem lassen sich in ähnlicher Weise Abschätzungen für den relativen Fehler $(\zeta_i - \omega_i)/\omega_i$ herleiten. Dabei muss allerdings (10) durch eine andere, einschränkendere Genauigkeitsforderung ersetzt werden.

Berücksichtigung der Rundungsfehler. Infolge der Rundungsfehler erhält man bei der Deflation nicht die durch (1) charakterisierten Polynome $p_k(z)$, sondern Polynome $\tilde{p}_k(z)$ mit leicht veränderten Koeffizienten. Sofern wir statt der in (2) definierten Polynome $q_k(z)$ nun die Polynome $\tilde{q}_k(z) := (z - \omega_k)\, \tilde{p}_{k-1}(z)$, $k = 1, \ldots, N$, verwenden, so können wir aber unsere Überlegungen übertragen, und zwar sowohl bei Festkomma- als auch bei Gleitkomma-Arithmetik. Allerdings werden nun die Schranken Δ_k

komplizierter als in (8), denn bei der Anwendung des Satzes von Ostrowski auf die Polynome $\tilde{p}_k(z)$ und $\tilde{q}_k(z)$ fällt in (3) die Zahl R nicht mehr heraus. Bekanntlich (vgl. z.B. Polya und Szegö [2], Kap. 3, Aufgaben 16, 17, 21) lässt sich aber eine Zahl R mit den gewünschten Eigenschaften aus den Koeffizienten (oder oberen Schranken für diese) von $\tilde{p}_k(z)$ und $\tilde{q}_k(z)$ berechnen. Auf diese Weise gelangen wir sogar wieder zu a priori gültigen Δ_k, mit welchen dann (9), (12) und Satz D unverändert gelten.

Zur praktischen Berechnung von *a posteriori* Schranken sind aber andere Methoden [3–5] angezeigt, die den Satz von Rouché direkt statt in der Form von Satz A benützen.

Wahl der Rechengenauigkeit. Als Anwendung lässt sich folgendes Problem lösen: Gegeben sei das Polynom $p_N(z)$ und eine Zahl $T > 0$; man bestimme ε und die Mantissenlänge, so dass das gegebene Nullstellen-Suchverfahren und der zur Deflation verwendete Hornersche Algorithmus Approximationen ω_i der Nullstellen ζ_i von $p_N(z)$ liefern, für die gilt $|\omega_i - \zeta_i| \leqslant T$, $i = 1, \ldots, N$. Ohne Beweis sei erwähnt, dass (bei entsprechender Mantissenlänge) $\varepsilon = (T/(2N))^N/2$ gewählt werden kann.

Literatur

[1] A. M. Ostrowski, *Solution of Equations and Systems of Equations*, 2nd ed. (Academic Press, New York 1966) (Appendices A, B).

[2] G. Polya und G. Szegö, *Aufgaben und Lehrsätze der Analysis*, Band 1, 3. Aufl. (Springer-Verlag, Berlin 1964).

[3] W. Börsch-Supan, *Residuenabschätzung für Polynom-Nullstellen mittels Lagrange-Interpolation*, Numer. Math. *14*, 287 (1970).

[4] B. T. Smith, *Zero Sets for Polynomials and Computable Regions Containing These Sets*, to appear.

[5] M. Gutknecht and B. T. Smith, *A Posteriori Error Bounds for the Zeros of a Polynomial*, in preparation.

Summary

The errors of the approximations to the zeros of a polynomial are analyzed, supposing these approximations have been found successively using factorization of the polynomial. We deduce an error bound depending only of the degree of the polynomial and the values of the reduced polynomials at the approximation being factored. The same method may be used to calculate error bounds in the case where round-off is involved.

(Eingegangen: 26. Januar 1971)