Numerische Mathematik 9, 80–98 (1966)

Die numerische Berechnung der Wurzeln eines Polynoms*

KARL NICKEL

Eingegangen am 4. Februar 1966

Einen *Zahlenwert* X numerisch berechnen heißt, einen *Näherungswert* x und eine *Fehlerschranke* ξ anzugeben derart, daß $|X-x| \leqq \xi$ ist.

1. Einleitung

In der folgenden Arbeit wird eine für digitale Rechenanlagen geeignete Methode zur Berechnung aller Wurzeln Z_p $(p=1(1)n)$ eines Polynoms

$$P(Z) := \sum_{m=0}^{n} A_m Z^m, \qquad Z = X + iY \tag{1}$$

mitgeteilt. Es sei $n>0$, die Koeffizienten A_m seien gegebene komplexe Zahlen mit $A_n \neq 0$; man kann o. B. d. A. $A_n = 1$ setzen. Nach dem Fundamentalsatz der Algebra gibt es dann genau n komplexe Zahlen Z_p derart, daß

$$P(Z) = A_n \prod_{p=1}^{n} (Z - Z_p) \quad \text{ist.} \tag{2}$$

Es gibt bis heute schon eine sehr große Anzahl von Methoden zur Bestimmung der Wurzeln Z_p und es ist einem geübten menschlichen Rechner ohne weiteres möglich, die Wurzeln Z_p jedes numerisch gegebenen Polynoms mit beliebiger Genauigkeit zu bestimmen. Erstaunlicherweise scheint es aber noch fast keine Algorithmen zu geben, die uneingeschränkt zur Verwendung in Rechenautomaten geeignet wären[1]. Dafür scheinen zwei Gründe verantwortlich zu sein:

1. Während ein menschlicher Rechner je nach Bedarf in flexibler Weise mit 5, 10 oder gar 20 Ziffern rechnet, sind die heutigen Ziffernrechner mit ihrer im allgemeinen festen Zahlenlänge wenig anpassungsfähig an alle diejenigen Pro-

* Die folgende Arbeit entstand während eines Aufenthalts des Verfassers als visiting professor in den USA an der University of Notre Dame/Indiana. Die numerische Erprobung des Verfahrens wurde auf der UNIVAC 1107 des dortigen Computing Centers vorgenommen. Der Verfasser dankt dessen Direktor, Herrn DON MITTELMANN, für die Zurverfügungstellung der benötigten Rechenzeit.

[1] Die einzigen mir bekannten Methoden, die wenigstens theoretisch einigermaßen befriedigen, sind diejenigen von LEHMER [2] und NASITTA [3]. Allerdings werden dabei die Rundungsfehler nicht berücksichtigt, so daß die praktische Brauchbarkeit nicht unbedingt gewährleistet ist. Bezüglich weiterer Hinweise auf Schwierigkeiten und Methoden vgl. man WILKINSON [5].

bleme, die mit Rundungsfehlern zusammenhängen[2]. Die Gleichung $P(Z)=0$ kann (bis auf den trivialen Sonderfall $A_0=0$ mit $Z_1=0$) nur dadurch befriedigt werden, daß auf der linken Seite mindestens einmal die Differenz zweier Zahlen gebildet wird. Dies bedeutet numerisch den „Verlust führender Ziffern" und verursacht damit einen — zunächst unkontrollierbaren — Genauigkeitsverlust.

2. Die bisher bekannt gewordenen Lösungsmethoden sind im allgemeinen nicht „automatensicher" in dem Sinne, daß an jeder Stelle des Rechenablaufs klare Vorschriften über die weiteren Rechenschritte vorliegen. Fast stets ist die Beobachtung des Rechenablaufs und ein hohes Maß an Einsicht in dem Fortschritt der Rechnung unumgänglich notwendig; manche Methoden können überhaupt nur mit einem aus Erfahrung gewonnenen Fingerspitzengefühl in fruchtbarer Weise angewendet werden. So ist z.B. das Graeffe-Verfahren in der Erweiterung nach BRODETSKI-SMEAL (siehe z.B. ZURMÜHL [6], S. 70ff.) für die Handrechnung hervorragend geeignet, doch läßt es sich nicht in befriedigender Weise programmieren[3].

Das nachstehend beschriebene Verfahren beseitigt diese beiden Schwierigkeiten. Es ist — wie auch Tests an weit über 10000 Polynomen mit mehr als 100000 Wurzeln gezeigt haben — unbeschränkt anwendbar. Die erhaltene Genauigkeit ist jedoch — je nach dem gewählten Beispiel und der verwendeten Rechenanlage — unter Umständen nicht ausreichend. Diese Eigenschaft ist allerdings unvermeidlich, solange in den Rechenautomaten mit fester Zahlenlänge gearbeitet wird. Sie folgt aus der Aufgabenstellung und ist nicht durch das verwendete Rechenverfahren bedingt. Es ist also — worauf leider oft nicht, oder nicht intensiv genug hingewiesen wird — im allgemeinen unmöglich, die Polynomwurzeln Z_p maschinell mit vorgegebener Genauigkeit numerisch zu bestimmen. Das folgende Programm berechnet daher zu den exakten Nullstellen Z_p Näherungswerte z_p, die „so gut wie möglich" sind und bestimmt nachträglich reelle Fehlerschranken ζ_p^* derart, daß

$$|Z_p - z_p| \leqq \zeta_p^* \quad \text{für} \quad p = 1(1)n \tag{3}$$

gilt. Es ist ohne weiteres möglich, als Koeffizienten A_m fehlerbehaftete Zahlen zu betrachten. Damit läßt sich die Abhängigkeit der Wurzelwerte von den Fehlern der Ausgangsdaten numerisch verfolgen.

Da es kein „bestes" Programm zur Nullstellenbestimmung eines beliebigen Polynoms geben kann, wird in dem laufenden Text besonderer Wert auf die Begründung der einzelnen Schritte gelegt und auf mögliche Abänderungen hingewiesen.

---

[2] Daran ändert auch die Tatsache nichts, daß fast alle Ziffernrechenmaschinen mit „einfacher" und „doppelter" Zahlenlänge rechnen können.

[3] Wenn 3 verschiedene Methoden im statistischen Durchschnitt je in 10% aller Fälle versagen, so liefert ihre gemeinsame Verwendung durch einen menschlichen Rechner — der das Versagen frühzeitig erkennen kann — eine durchschnittliche Versagensrate von 1:1000, also einen für praktische Bedürfnisse beim menschlichen Rechnen völlig zufriedenstellenden Wert. Beim automatischen Rechnen dagegen wird im allgemeinen eine Fehlerquote von 1:10000 noch als untragbar schlecht angesehen; bei der Verwendung eines Verfahrens als Unterprogramm muß dieser Wert sogar exakt Null sein.

2. (Grobe) Skizze des Lösungsverfahrens

I. Beginne mit einem passenden Näherungswert z^*.

II. Verbessere z^* so, daß $|P(z^*)|$ kleiner wird.

III. Wiederhole den Schritt II so lange, bis $|P(z^*)|$ hinreichend klein ist. Dann ist $z_p := z^*$ ein Näherungswert für eine Wurzel Z_p von $P(Z)$.

IV. Bestimme zu z_p eine (reelle und positive) Fehlerschranke ζ_p^* derart, daß

$$|Z_p - z_p| \leqq \zeta_p^* \text{ ist.}$$

V. Reduziere $P(Z)$ durch Division durch den Linearfaktor $Z - Z_p$ zu einem Polynom vom Grade $n-1$ und wiederhole die Schritte I bis IV so lange, bis der Grad Eins erreicht und damit für $p = 1(1)n$ Näherungswerte z_p und Fehlerschranken ζ_p^* für die exakten Wurzeln Z_p bestimmt sind.

Bemerkung: Es ist mir nicht bekannt, ob das nachfolgend beschriebene, in den Schritten I und II verwendete Näherungsverfahren schon von anderen Autoren angegeben wurde, doch ist das bei dessen einfachem Bau sehr wahrscheinlich. (Ich selbst wurde durch die Lektüre der interessanten Arbeit von Herrn NASITTA [3] auf die benutzten Vorschriften geführt und war durch eine von mir zu haltende Vorlesung über Numerische Mathematik angeregt worden, mich mit diesem Fragenkomplex zu beschäftigen.) Die Mitteilung eines neuen Verbesserungsverfahrens erscheint mir als ein unwichtiges Detail, die Methode von I und II kann vom Leser durch eine beliebige andere passende Vorschrift ersetzt werden.

Wichtig und neuartig bei dem vorliegenden Programm scheinen mir vielmehr die beiden Tatsachen zu sein, daß hier wohl erstmals die logische Abfrage zur Iterationsbeendigung in Teil III als ein integrierter Bestandteil der verwendeten Arithmetik erscheint und daß wohl ebenfalls erstmalig die Berechnung einer Näherung (in Teil I und II) und die Angabe einer zugehörigen Fehlerschranke (in Teil IV) als eine Einheit aufgefaßt wird.

3. Beschreibung des Verfahrens

Bezeichnungen: Exakte (komplexe) *Zahlen* werden im folgenden durch große lateinische, (komplexe) *Näherungswerte* durch kleine lateinische Buchstaben dargestellt. (Komplexe) *Fehlerschranken* werden als kleine griechische Buchstaben geschrieben. *Ausnahmen:* Die Variabeln j, k, l, m, n, p, q bedeuten durchweg ganzzahlige reelle (exakte) Zahlen; i ist wie üblich die imaginäre Einheit. Ist zu einer exakten Zahl Z ein Näherungswert z und eine Fehlerschranke ζ bekannt, so wird die Schreibweise $Z = (z, \zeta)$ gebraucht, Z ist dann eine „*Schrankenzahl*" im Sinne von [4]. *Beispiel:* $Z = X + iY = (z, \zeta)$ mit $z = x + iy$, $\zeta = \xi + i\eta$ und $|X - x| \leqq \xi$, $|Y - y| \leqq \eta$. Für die Koeffizienten A_m des Polynoms (1) seien Näherungen a_m und Schranken α_m bekannt. Mit der soeben eingeführten Schreibweise gibt das $A_m = (a_m, \alpha_m)$ für $m = 0(1)n$. Entsprechend zu dem Vorschlag in dem Artikel [4] „Die Notwendigkeit einer Fehlerschrankenarithmetik" sollen Unterprogramme für die arithmetische Verknüpfung von Schrankenzahlen vorliegen[4]. Die Verknüpfung der Näherungswerte erfolgt dabei durch die übliche Gleitkommaarithmetik (Permanenzprinzip).

[4] Diese Voraussetzung ist wesentlich, das zu schildernde Verfahren kann *nicht* ohne eine Schrankenzahlenarithmetik formuliert werden!

Zählvariable: Die Variable p zähle die Nummer der augenblicklich zu bestimmenden Wurzel. Zu Beginn setzt man

$$p := 1.$$

Zu I. Algorithmus

Man setzt

$$z := -a_{n-1}/n \cdot a_n. \tag{4}$$

Begründung: a) Da in den Teilen I und II nur ein *Näherungs*wert z^* gefunden werden soll, dessen Fehler erst nachträglich in Teil IV abgeschätzt wird, ist ein Rechnen mit exakten Zahlen und eine Berücksichtigung der Eingangsfehler α_m vorläufig überflüssig. Die Rechnung erfolgt in (4) und in den nächsten Formeln daher mit der üblichen Gleitkommaarithmetik (Rechenzeitersparnis); dies wird durch das Kleinschreiben der auftretenden Variablen sichtbar.

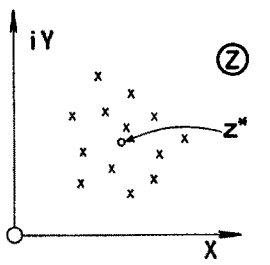

Fig. 1. Wurzelhaufen und Lage von z^* nach Formel (4) in der Z-Ebene

b) Nach den Vietaschen Wurzelsätzen ist z^* nach (4) das arithmetische Mittel aller Wurzeln (der Schwerpunkt, wenn man sich die Nullstellen in der komplexen Zahlenebene als gleichschwere Massenpunkte vorstellt). Im Fall eines Wurzelhaufens gibt daher (4) unmittelbar eine gute Näherung (s. Fig. 1).

c) Das unter II beschriebene Näherungsverfahren arbeitet im allgemeinen „von innen nach außen", d.h. bestimmt zuerst die betragskleinsten Wurzeln. Dies liefert besonders kleine Rundungsfehler. Auch aus diesem Grunde ist es nützlich, mit (4) zunächst eine Transformation derart durchzuführen, daß alle Wurzeln „gleichmäßig" um den Nullpunkt verteilt sind und sie dann, vom Ursprung aus beginnend, der Größe des absoluten Betrags nach „abzuarbeiten". Im folgenden Algorithmus wurde diese Nullpunktverschiebung allerdings nicht angewendet, um den Aufbau des Verfahrens übersichtlicher zu gestalten.

d) Für $n = 1$ ist z^* nach (4) die exakte Wurzel.

e) Ist $P(Z)$ ein Kreisteilungspolynom, dann liefert das folgende Verfahren im ersten Schritt eine exakte Wurzel. Durch eine Translation mit (4) kommt man — falls die geometrische Lage der Nullstellen es zuläßt — auf die Kreisteilungsform. Dies gelingt immer für $n = 2$, so daß in diesem Falle beide Wurzeln stets mit nur einem einzigen Verfahrensschritt [und zuzüglich zwei Anfangswerten nach (4)] bestimmt werden.

Andere Möglichkeiten: a) Das Näherungsverfahren von Teil II arbeitet unabhängig vom Anfangswert, so daß z.B. $z^* := 0$ gesetzt werden könnte.

b) Ist eine Nullstelle z_p schon (näherungsweise) bestimmt, so ist $z^* := z_p$ sinnvoll. Ist z_p mehrfach, so erhält man unmittelbar ohne Schritt II die nächsten Nullstellen.

c) Sind die Koeffizienten A_m alle reell und ist wieder eine Nullstelle z_p (näherungsweise) schon bekannt, so ist auch $z^* := \bar{z}_p$ (konjugiert komplexer Wert) empfehlenswert. Zu jeder bereits bestimmten echt komplexen Nullstelle wird

dann — ohne Iteration — sofort die zugehörige konjugiert komplexe angegeben. Eine mehrfache Nullstelle z_p wird ebenfalls wie bei b) ohne Iteration abgebaut, und zwar direkt, wenn sie reell ist und alternierend $z_p, \bar{z}_p, z_p, \bar{z}_p, \ldots$ wenn z_p komplex ist. Im Mittel wird — bei reellen Koeffizienten A_m — die Rechenzeit daher „etwa" auf die Hälfte reduziert.

Zu II. A. Normalschritt

Algorithmus: Zum gegebenen Näherungswert z^* werden die Koeffizienten $c_m := P^{(m)}(z^*)/m!$ des umentwickelten Polynoms (mit dem vollständigen Hornerschema) berechnet, so daß gilt

$$P(Z) = \sum_{m=0}^{n} c_m (Z - z^*)^m. \tag{5}$$

Es sei

$$\hat{z} := z^* + \sqrt[k]{-c_0/c_k}\ ^{5}, \tag{6}$$

wobei k so gewählt ist, daß

$$\sqrt[k]{|c_0/c_k|} = \operatorname*{Min}_{m=1\,(1)\,n} \sqrt[m]{|c_0/c_m|}\ ^{6} \tag{7}$$

gilt. Ist

$$\hat{z} = z^*, \tag{8}$$

so fahre man fort bei Teil IV. Ist (8) nicht erfüllt, aber

$$|P(\hat{z})| < |P(z^*)|, \tag{9}$$

so fahre man fort bei Teil III, andernfalls bei Teil II B.

Begründung: a) Wie oben wird nur die übliche Gleitkommaarithmetik verwendet, man erkennt dies daran, daß die auftretenden Variablen kleingeschrieben sind.

b) Ist z^* Näherungswert einer Polynomwurzel und ist $c_1 \neq 0$, dann können in der Entwicklung (5) im allgemeinen die Glieder $c_2(Z - z^*)^2 + \cdots$ vernachlässigt werden. Durch „Auflösen nach Z" erhält man so die wohlbekannte Verbesserungsvorschrift des Newtonschen Verfahrens.

$$z^* := z^* - c_0/c_1.$$

c) Ist $c_1 = 0$ oder $|c_1| \ll |c_0|$, so „löst man nach dem zweiten Glied von Gl. (5) auf", d.h. man vernachlässigt die Terme $c_1(Z - z^*)$ und $c_3(Z - z^*)^3 + \cdots$. Dies gibt

$$z^* := z^* + \sqrt{-c_0/c_2}.$$

d) Ist allgemein $c_1 = c_2 = \cdots = c_{k-1} = 0$, $c_k \neq 0$, so findet man die Vorschrift (6). Wählt man in (6) den Zahlenwert k so, daß noch (7) gilt, dann hat man auch noch den Fall „die Werte $|c_1|, |c_2|, \ldots, |c_{k-1}|, |c_{k+1}|, \ldots, |c_n|$ sind klein" in sinnvoller Weise erfaßt.

[5] Zur Berechnung des Werts der k-ten Wurzel ist ein beliebiger Zweig der Wurzelfunktion zu verwenden. Ist $-c_0/c_k = r^k e^{i\varphi}$ und φ wie üblich so normiert, daß $-\pi < \varphi \leqq \pi$ gilt, dann wird man etwa $\sqrt[k]{-c_0/c_k} := r\, e^{i\varphi/k}$ setzen.

[6] Indizes m mit $c_m/c_0 = 0$ sind zu streichen.

Die Formeln (6), (7) werden dem weiteren Vorgehen ohne weitergehende Begründung zugrunde gelegt. Sie sind eine sinnvolle Verallgemeinerung des Newtonschen Verfahrens und vermeiden insbesondere dessen beide Nachteile:

α) $|c_1| \ll |c_0|$ führt zu im allgemeinen unsinnigen Ergebnissen und häufig zur Bereichsüberschreitung.

β) Sind alle Koeffizienten A_m reell und ist z^* reell, so können keine komplexen Nullstellen gefunden werden. Die Forderung (6) findet sich z.B. schon bei NASITTA [3], allerdings ohne die Bedingung (7), durch die eine fruchtbare Anwendung erst gesichert erscheint.

e) Es kann geschehen, daß der Zusatzterm auf der rechten Seite der Gl. (6) im Rahmen der Rechengenauigkeit den Wert von z^* nicht mehr ändert. Dann ist (8) erfüllt. Dieser Fall tritt dann ein, wenn z^* schon eine sehr gute Näherung ist. Offensichtlich kann dann mit dem vorliegenden Verfahren keine weitere Verbesserung von z^* mehr geleistet werden: die Iteration „steht" und ist abzubrechen.

f) Es ist zwar fast selbstverständlich, jedoch soll ausdrücklich darauf hingewiesen werden, daß die Divisionen in den Formeln (6) und (7) die einzigen Divisionen dieses Schrittes sind (im Hornerschema treten nur $+$, $-$, $\times$ auf) und daß sie *nicht* zu Bereichsüberschreitungen führen können, wenn $P(Z)$ normiert ist. Bei dem Verfahren von NASITTA [3] ist diese Gewähr nicht gegeben.

B. Zusatzschritt (Halbierungsverfahren)

Es sei $\hat{z} \neq z^*$ und

$$|P(\hat{z})| \geqq |P(z^*)| > 0.$$

Man setzt für $m = 1, 2, \ldots$

$$\hat{z} := z^* + 2^{-m} r \sqrt[l_m]{-\frac{c_0}{c_{l_m}} \left| \frac{c_{l_m}}{c_0} \right|} \tag{10}$$

mit

$$r := \sqrt[k]{|c_0/c_k|}, \tag{11}$$

wobei l_m so gewählt ist, daß

$$|c_{l_m}| (2^{-m} r)^{l_m} = \max_{j=1\,(1)\,l_{m-1}} \{|c_j| (2^{-m} r)^j\} \tag{12}$$

gilt mit $l_0 = k$. Sobald für einen Wert m

$$\hat{z} = z^*$$

ist, fahre man fort beim Teil IV. Wenn für ein m

$$|P(\hat{z})| < |P(z^*)|$$

gilt, ist bei Teil III fortzufahren.

Begründung: a) Der Normalschritt A sichert keine monotone Konvergenz im Sinne von Ungleichung (9). *Beispiel:* $n = 3$, $P(Z) := 1 + Z + Z^2 + Z^3$, $z^* := 0$ gibt $|P(z^*)| = 1$ und mit $k = 1$ oder $k = 2$ oder $k = 3$ wird jeweils $|P(\hat{z})| = 2 > 1$. (Die praktische Erprobung zeigt allerdings, daß solche Betragserhöhungen (wenn auf Schritt B verzichtet wird) fast stets nur vorübergehend auftreten und schnell wieder abgebaut werden.)

b) Die Gefahr der gefürchteten Käfigbildung ist bei der Vorschrift (6), (7) ebenfalls nicht ausgeschlossen:

Beispiel: $n=2$, $P(Z):=3+Z^2$, $z^*:=1$ gibt $P(Z)=4+2(Z-1)+(Z-1)^2$ und damit $k=1$ oder $k=2$. Wählt man $k=1$, so findet man die nichtkonvergente Folge $\hat{z}=1, -1, 1, -1, ..$ mit $P(\hat{z})=4, 4, 4, 4, \ldots$. (In der Praxis werden solche Käfige jedoch fast immer durch Rundungseffekte vermieden.)

c) Der Sinn des Halbierungsverfahrens in Schritt B besteht darin, die beiden Fälle a) und b) unmöglich zu machen und damit eine stark monotone Konvergenz zu erzwingen. Da in einem Rechenautomaten nur endlich viele Zahlen zur Verfügung stehen, ist damit ein Abbrechen nach endlich vielen Schritten gesichert. Zwar ist die Konvergenzgeschwindigkeit bei B nur noch linear, doch zeigt die numerische Erfahrung (siehe Teil 4), daß der Sonderfall B im Durchschnitt in weniger als 10% aller Rechenschritte auftritt und daher die Rechengeschwindigkeit nicht wesentlich beeinträchtigt.

d) Nach dem Maximumprinzig gibt es wegen $|P(z^*)|>0$ in jeder Umgebung von z^* eine Stelle $\hat{z}$ mit $|P(\hat{z})|<|P(z^*)|$. Wie schon NASITTA [3] gezeigt hat, kann man solch eine Stelle $\hat{z}$ etwa durch systematische Verkleinerung der Schrittweite r finden. Daß durch die Formeln (10) bis (12) wenigstens theoretisch solch eine Stelle gefunden wird, soll nun begründet werden: Die Vorschrift (12) wählt $l=l_m$ für jedes $m=1, 2, \ldots$ so, daß in der Entwicklung

$$P(Z)=c_0+c_1(Z-z^*)+\cdots+c_l(Z-z^*)^l+\cdots+c_n(Z-z^*)^n$$

der l-te Term jeden einzelnen anderen im Absolutwert übertrifft, d.h., daß stets

$$|c_l||Z-z^*|^l \geqq |c_j||Z-z^*|^j$$

ist für $|Z-z^*|=r\cdot 2^{-m}$ und $j=1\,(1)\,n$. (Die Analogie dieses Vorgehens zu demjenigen von A ist offensichtlich.) Nach Konstruktion der Zahl k ist es dabei sogar schon ausreichend, allein den eingeschränkten Zahlbereich $j=1\,(1)\,k$ zugrundezulegen.

Nach Konstruktion strebt $\hat{z}-z^*$ für $m\to\infty$ in Zweierpotenzen gegen Null. Ist c_l der erste nichtverschwindende Koeffizient in der Folge der c_j für $j=1\,(1)\,n$, so gibt es also eine Zahl m^* derart, daß $l_m=l$ gilt für alle $m\geqq m^*$. Weiterhin gibt es eine Zahl $m^{**}\geqq m^*$ derart, daß sogar

$$\tfrac{1}{2}|c_l||\hat{z}-z^*|^l \geqq \Big|\sum_{j>l} c_j(\hat{z}-z^*)^j\Big|$$

gilt für alle $m>m^{**}$. Schließlich ist nach Definition (7) und wegen $c_l\neq 0$ noch $0<q:=r/\sqrt[l]{|c_0/c_l|}<1$. Damit ist für $m>m^{**}$

$$\begin{aligned}
|P(\hat{z})| &= \Big|\sum_{j=0}^{n} c_j(\hat{z}-z^*)^j\Big| \\
&\leqq |c_0+c_l(\hat{z}-z^*)^l| + \Big|\sum_{j>l} c_j(\hat{z}-z^*)^j\Big| \\
&\leqq |c_0+c_l(\hat{z}-z^*)^l| + \tfrac{1}{2}|c_l||\hat{z}-z^*|^l \\
&\leqq |c_0||1-q^l 2^{-ml}| + \tfrac{1}{2}|c_0|\, q^l 2^{-ml} \\
&= |c_0||1-q^l 2^{-ml-1}| < |c_0| = P(z^*)|,
\end{aligned}$$

d.h. die Konstruktion durch die Formeln (10) bis (12) liefert für hinreichend großes m sicherlich eine Stelle $\hat{z}$ derart, daß $|P(\hat{z})|<|P(z^*)|$ ist.

e) Allerdings gilt der vorstehende Beweis nur dann, wenn exakte Zahlenwerte betrachtet werden. Durch das Rechnen mit endlicher Zahlenlänge kann unter Umständen keine solche Stelle $\hat{z}$ gefunden werden. Bei fortgesetzter Schritthalbierung ergibt sich dann schnell der Fall (8), in dem im Rahmen der Rechengenauigkeit $\hat{z} = z^*$ ist. Eine weitere Verbesserung von z^* mit diesem Verfahren ist dann unmöglich, die Iteration ist abzubrechen.

Andere Möglichkeiten: Die beiden Schritte A und B von Teil II können durch *jedes* andere Verfahren ersetzt werden, das 1. theoretisch (d.h. beim Rechnen mit unendlich vielen Ziffern) konvergiert, 2. nach endlich vielen Schritten abbricht und 3. keine Bereichsüberschreitung erzeugt. Bis auf Punkt 3 genügen die Methoden von LEHMER [2] und NASITTA [3] diesen Forderungen. — Die „theoretische" Konvergenz des vorstehend beschriebenen Iterationsverfahrens läßt sich zwar beweisen, doch soll das — entsprechend der praktischen Zielsetzung dieser Note — hier nicht geschehen.

Zu III. Algorithmus

Man setze

$$Z^* := (\hat{z}, 0) \tag{13}$$

und berechne mit einer Schrankenarithmetik

$$B = (b, \beta) := P(Z^*). \tag{14}$$

Wenn gleichzeitig gilt

$$|Re\, b| \leqq Re\, \beta \quad \text{und} \quad |Im\, b| \leqq Im\, \beta, \tag{15}$$

so werde bei Teil IV fortgefahren, andernfalls soll zu Teil II A zurückgegangen werden.

Begründung: a) In den Teilen I und II wird der Wert von z^* bzw. $\hat{z}$ ohne Rücksicht auf Rundungsfehler berechnet. Da $\hat{z}$ nur als Näherungswert dient, der schrittweise verbessert wird, ist das zulässig. Dieser — möglicherweise noch recht ungenaue — Wert $\hat{z}$ wird nun als rundungsfehlerfrei angesehen [Schreibweise Z^* in Formel (13)] und es wird durch (15) getestet, ob die unvermeidlichen Rundungsfehler bei der Berechnung von $P(Z^*)$ nach (14) schon so groß sind, daß $P(Z^*) = 0$ sein könnte. Sobald das nach (15) der Fall ist, wird die Iteration beendet[7]. Es könnte sich dann zwar durchaus bei weiterer Iteration noch ein genauerer Wert $\hat{z}$ ergeben. Man könnte dementsprechend die Iteration erst dann beenden, wenn die nach Teil IV berechnete Fehlerschranke ζ^* zu $\hat{z}$ wieder zunimmt. Die numerische Erfahrung zeigt jedoch, daß der mögliche Gewinn im allgemeinen klein ist und in keinem günstigen Verhältnis zur Verlängerung der Rechenzeit steht.

Zu IV. Algorithmus

Man bestimme zu $Z^* := (\hat{z}, 0)$ mit einer Fehlerschrankenarithmetik (und dem großen Hornerschema) die Koeffizienten $C_m = (c_m, \gamma_m)$ so, daß

$$P(Z) = \sum_{m=0}^{n} C_m (Z - Z^*)^m$$

[7] Man vgl. dazu [4], wo auf die Nullstellenbestimmung einer ungenauen Funktion ausführlich eingegangen wird.

ist[8]. Man setzt

$$F=(f,\varphi):=\sqrt[k]{\binom{n}{k}\,|C_0/C_k|}$$

und

$$\zeta_p^*:=\operatorname*{Min}_{k=1\,(1)\,n}(f+\varphi).^{9}$$

Dann gilt für die p-te Wurzel Z_p des Polynoms die Abschätzung

$$|Z_p-Z^*|\leqq\zeta_p^*. \tag{16}$$

Man setzt weiter

$$Z_p=(z_p,\zeta_p):=(\hat{z},\zeta_p^*+i\zeta_p^*) \tag{17}$$

mit der Abschätzung

$$|Re(Z_p-Z^*)|\leqq\zeta_p,\ |Im(Z_p-Z^*)|\leqq\zeta_p.$$

Begründung: a) Nach allgemeinen Sätzen über die Nullstellen von Polynomen (vgl. etwa FEKETE [1], S. 302, Formel (8)) gibt es mindestens eine Wurzel $\hat{Z}$ des Polynoms $P(Z)$ derart, daß

$$|\hat{Z}-Z^*|\leqq\operatorname*{Min}_{k=1\,(1)\,n}\sqrt[k]{\binom{n}{k}\,|C_0/C_k|} \tag{18}$$

ist. $Z_p:=\hat{Z}$ gesetzt gibt die Abschätzung (16).

b) Für die weitere Verarbeitung der damit gefundenen p-ten Polynomwurzel Z_p ist eine Darstellung $Z_p=(z_p,\zeta_p)$ als Schrankenzahl erforderlich. Durch die Setzung (17) wird dies geleistet. Allerdings wird dabei vergröbernd eine komplexe Schranke ζ_p für den Fehler von Z_p eingeführt, während in (16) allein der absolute Betrag des Fehlers abgeschätzt wurde. (Übergang von einer Normabschätzung zu einer komponentenweisen Abschätzung).

Zu V. Algorithmus

Zu der durch (17) definierten Polynomwurzel $Z_p=(z_p,\zeta_p)$ bestimmt man mit einer Fehlerschrankenarithmetik und dem Hornerschema die Koeffizienten $B_m=(b_m,\beta_m)$ so, daß

$$\frac{P(Z)}{Z-Z_p}=\frac{B_0}{Z-Z_p}+\sum_{m=1}^{n}B_mZ^{m-1} \tag{19}$$

ist. Weiter wird gesetzt

$$n:=n-1;$$
$$A_m:=B_{m+1},\quad\text{für}\quad m=0\,(1)\,n;$$
$$p:=p+1.$$

Ist

$$n>0$$

so ist zum Schritt I zurückzukehren, andernfalls gilt $n=0$. In diesem Falle ist das Verfahren beendet, alle Nullstellen Z_p des Polynoms $P(Z)$ sind bestimmt, es ist zu stoppen.

[8] Wenn die Fehlerschrankenarithmetik, die in [4] empfohlene Permanenzeigenschaft besitzt, stimmen die Werte der Variablen c_m mit den in Gl. (5) zuletzt bestimmten überein.

[9] Das Zeichen $\dotplus$ bedeutet „Addition mit Aufrunden", s. [4].

Abschließende Bemerkungen: a) Wenn die Werte der gegebenen Koeffizienten A_m fehlerfrei sind, so besitzen nach der Abspaltung der ersten Nullstelle die neuen Koeffizienten (die wieder mit A_m bezeichnet werden) des reduzierten Polynoms vom Grade $n-1$ im allgemeinen unvermeidliche Rundungsfehler. Bei jeder weiteren Reduktion erhöhen sich diese Fehler, bis der Fall $n=1$ erreicht ist. Man wird daher erwarten, daß die zuerst bestimmten Nullstellen genauer sein werden als diejenigen, die erst nach vielen Reduktionen errechnet werden. Da die ständig ungenauer werdenden Werte der A_m auch in die Schrankenberechnung nach den Formeln (16), (17) eingehen, sind zusätzlich dazu noch für die später berechneten Wurzeln Z_p erheblich schlechtere Schranken ζ_p zu erwarten (s. dazu die numerischen Experimente von Teil 4).

b) Diese beiden Schwierigkeiten lassen sich scheinbar auf die folgende Weise umgehen: Nach Bestimmung eines genügend genauen Näherungswerts Z^* für die p-te Wurzel Z_p iteriert man mit dem Ausgangspolynom $P(Z)$ weiter. Ebenso bestimmt man die Fehlerschranke von Z^* mit den Koeffizienten des ursprünglichen Polynoms $P(Z)$. In der numerischen Mathematik wird dieses Verfahren der „Nachiteration" oft propagiert.

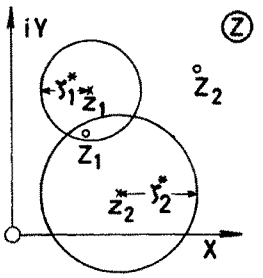

Fig. 2. Erklärung im Text

Leider ist dieser Weg jedoch nicht gangbar: Da es keine Möglichkeit gibt, die Polynomwurzeln zu „markieren", kann bei der Nachiteration eine andere Wurzel als die ursprünglich approximierte angesteuert werden. Bei dem Reduktionsverfahren dagegen wird sicherlich jedesmal eine neue Wurzel ausgewählt, weil die bereits bearbeiteten beseitigt worden sind.

Genau so ist es auch mit der Fehlerabschätzung. Wertet man die Formeln (16), (17) mit den Koeffizienten des ursprünglichen Polynoms aus, so kann man ein Minimum des Einflusses der Rundungsfehler erhoffen. Da jedoch bei (18) nur die Existenz einer Nullstelle $\hat{Z}$ von $P(Z)$ mit der Eigenschaft (18) behauptet werden kann, wäre es möglich, daß sich diese Fehlerabschätzung auf die „falsche" Nullstelle bezieht. Ein einfachstes Beispiel für $n=2$ zeigt die Fig. 2: Die Näherungswerte z_1 und z_2 der beiden exakten Nullstellen Z_1 und Z_2 liegen beide nahe bei Z_1. Eine Fehlerschrankenberechnung nach (18) gibt dann im allgemeinen zwei Schranken ζ_1^* und ζ_2^*, die sich beide auf Z_1 beziehen (in Fig. 2 durch zwei Kreise $|Z-z_j| \leqq \zeta_j^*$ $(j=1, 2)$ angedeutet)!

Diese Schwierigkeit tritt nicht auf, wenn alle Nullstellen des Polynoms einfach sind und wenn sich aus der Fehlerabschätzung Schrankenwerte ζ_p^* derart ergeben, daß die Fehlerkreisscheiben $|Z-z_p| \leqq \zeta_p^*$ $p=1(1)n$ keine gemeinsamen Punkte enthalten. Diese Eigenschaft läßt sich natürlich erst a posteriori feststellen. Man kann daher zweierlei Fehlerabschätzungen programmieren: 1. aus den Koeffizienten des ursprünglichen Polynoms und 2. aus dem Koeffizienten der jeweils reduzierten Polynome. Wenn die Fehlerkreise nach 1. alle punktfremd sind, ist die Abschätzung 1. gültig, die im allgemeinen erheblich ungünstigere Abschätzung 2. kann unberücksichtigt bleiben. Wenn k Kreisscheiben einen nichtleeren Durchschnitt besitzen, so liegt darin mindestens eine Nullstelle, jedoch sind sicher $n-k$ Nullstellen in den restlichen $n-k$ punktfremden Kreisscheiben enthalten.

Wenn man ein allgemeingültiges Programm aufstellen will, das auch mehrfache und engbenachbarte Nullstellen mit Sicherheit erfaßt, so gibt also allein die fortlaufende Reduktion von $P(Z)$ durch Division durch den Linearfaktor $Z-Z_p$ eine Gewähr dafür, sämtliche Wurzeln Z_p $(p=1\,(1)\,n)$ zu approximieren und zugehörige Fehlerschranken zu bekommen. Der dadurch entstehende Genauigkeitsverlust muß als unvermeidlich in Kauf genommen werden.

c) Es soll ausdrücklich betont werden:

Die Reduktion von $P(Z)$ durch die Formel (19) ist exakt gültig, obwohl die Nullstelle Z_p nur näherungsweise bekannt ist: Damit sind auch sämtliche Überlegungen, die sich auf das reduzierte Polynom $\sum_{m=0}^{n-1} B_{m+1} Z^m$ beziehen, wieder exakt richtig:

Der Grund dafür liegt darin, daß die (unbekannte) exakte Nullstelle $Z_p = (z_p, \zeta_p)$ in den Intervallen $\left\{\begin{matrix}\mathrm{Re}\\ \mathrm{Im}\end{matrix}\right\}(z_p-\zeta_p) \leqq \left\{\begin{matrix}\mathrm{Re}\\ \mathrm{Im}\end{matrix}\right\} Z_p \leqq \left\{\begin{matrix}\mathrm{Re}\\ \mathrm{Im}\end{matrix}\right\}(z_p+\zeta_p)$ eingefangen ist und daß die Fehlerschrankenarithmetik sich immer auf diese ganzen Intervalle bezieht. Die restlichen exakten Nullstellen von $P(Z)$ liegen daher in den Nullstellenbereichen von $\sum_{m=0}^{n-1} B_{m+1} Z^m$.

4. Numerische Ergebnisse

Es gibt leider sehr viele Veröffentlichungen über neue numerische Verfahren, in denen nur wenige oder gar keine Beispiele mitgeteilt werden. Häufig sind solche Beispiele noch nicht einmal typisch und verraten nichts über ein eventuelles Versagen, über Instabilität oder über einen Genauigkeitsverlust durch Rundungsfehler. Verfahren wie das vorliegende, durch das eines der Grundprobleme der numerischen Mathematik gelöst wird, sollten grundsätzlich an einer sehr großen Zahl von Beispielen kontrolliert werden, wobei die ungünstigsten Fälle selbstverständlich mit zu berücksichtigen sind.

Die Erprobung der vorstehend beschriebenen Methode geschah auf der elektronischen Rechenanlage UNIVAC 1107 an der University of Notre Dame/Indiana, USA[10]. (Rechnung mit einfacher Zahlenlänge, Gleitkommamantisse $= 27$ bit ≈ 8 Dezimalziffern). Es wurden dazu weit über 10000 Polynome mit mehr als 100000 Nullstellen verarbeitet. Für jeden Wurzelwert wurde dabei statistisch ausgewertet: Die Anzahl der notwendigen Iterationen, die erreichte Genauigkeit und die erhaltene Fehlerschranke.

In einem ersten Untersuchungsabschnitt wurden die Koeffizienten der betrachteten Polynome als reelle und als komplexe Zufallszahlen erzeugt. Zur Untersuchung der erreichbaren Genauigkeit wurden in einem zweiten Abschnitt die Nullstellen als komplexe Zufallszahlen vorgegeben, daraus die Koeffizienten des dazugehörigen Polynoms errechnet und schließlich allein aus diesen Polynomkoeffizienten wieder näherungsweise die Nullstellen bestimmt. Die Differenz der vorgegebenen und der erhaltenen Zahlenwerte gibt dann die exakten Fehler, die mit den Fehlerschranken des Verfahrens verglichen werden können. Die Wahrscheinlichkeit für das Auftreten mehrfacher Nullstellen ist natürlich in beiden Fällen gleich Null. Da die Behandlung von Polynomen mit mehrfachen Nullstellen

[10] Für die Herstellung des benötigten Unterprogramms für die Fehlerschranken-Arithmetik bin ich Herrn HANS HERMANS sehr dankbar.

oder mit Wurzelhaufen besonders schwierig ist und an jedes Verfahren die höchsten Anforderungen stellt, wurde in einem dritten Untersuchungsabschnitt dieser ungünstigste Fall gesondert behandelt. Es wurden dabei wieder die Polynomwurzeln als Zufallszahlen vorgegeben, doch wurde dafür gesorgt, daß künstlich mehrfache (q-fache) Nullstellen oder Wurzelhaufen (von q Nullstellen) auftraten, wobei q zwischen 1 und n variierte.

In der numerischen Erprobung wurde kein logisches Versagen des Programms beobachtet, ebenso traten (wie zu erwarten) keine Instabilitäten auf. Die erreichte Genauigkeit ist oft sehr schlecht, doch liegt das in der Natur des Problems und nicht an dem untersuchten Verfahren.

Um einen Begriff von der Geschwindigkeit des Programms zu geben sei erwähnt, daß auf der UNIVAC 1107 im Mittel alle Wurzeln eines Polynoms vom Grade $n = 10$ in weniger als 2 Sekunden berechnet werden. Die entsprechenden Zeiten für $n = 20$ und $n = 30$ lauten 12 Sekunden und 36 Sekunden. Der erforderliche Rechenaufwand geht etwa mit der Potenz n^3, ist also sehr hoch. Dies liegt sowohl an der Berechnung der Näherungswerte als auch an der Fehlerabschätzung. Sollten daher die Teile I bis III des vorliegenden Programms durch einen „schnelleren" Algorithmus ersetzt werden, so würde die Rechenzeit doch nicht wesentlich sinken, da die Fehlerabschätzung in jedem Falle einen Aufwand proportional zu n^3 erfordert (vollständiges Hornerschema).

Die in den folgenden Tabellen und Figuren niedergelegten Ergebnisse sind ein kleiner aber repräsentativer Ausschnitt aus dem Erprobungsprogramm.

Anzahl der benötigten Iterationen: Diese Anzahl erwies sich als erstaunlich klein, nach Tabelle 1 erste Zeile sind bis zum Polynomgrad $n = 30$ im Mittel pro Nullstelle weniger als 6 Iterationsschritte erforderlich: Z. B. werden beim Polynomgrad $n = 10$ im Mittel über 100 Polynome insgesamt 44,6 Iterationen benötigt, d.h. pro Wurzel 4,46 Iterationen. In manchen Fällen ist natürlich eine größere Anzahl von Iterationen erforderlich. Nach Tabelle 1 zweite Zeile war der ungünstigste Wert bis $n = 30$ jedoch nur 16 Iterationen, dieser ungünstigste Fall bezieht sich hier auf einen Vergleich von 1800 Polynomen mit zusammen 19500 Wurzeln!

Sollte die Anzahl der Zusatzschritte mit dem Halbierungsverfahren nach II B im Mittel sehr groß sein, so würde die dabei auftretende lineare Konvergenz das Verfahren sehr stark verlangsamen. Glücklicherweise ist das nicht der Fall, s. Tabelle 1, Zeile 3 und 4. Es treten zwar gelegentlich bis zu 7 Zusatzschritte auf, doch sind diese Fälle so selten, daß die mittlere Häufigkeit bis $n = 30$ stets unter 10% der gesamten Schrittzahl bleibt.

In Tabelle 2 wird an dem Beispiel von 100 Polynomen mit dem Grad $n = 20$ gezeigt, wie sich die in Tabelle 1 angegebenen Zahlenwerte auf die nacheinander ermittelten Wurzeln desselben Polynoms verteilen. Es ergibt sich, daß die Anzahl der benötigten Iterationen für die ersten gefundenen Wurzeln etwas kleiner ist und langsam ansteigt bis zu einem Maximum, wenn ungefähr die Hälfte aller Wurzeln schon ermittelt ist. Ein Grund dafür ist mir nicht bekannt. Anschließend sinken die Werte wieder ab, weil das jeweils zu bearbeitende reduzierte Polynom laufend kleineren Grad hat. Wenn dieser Grad auf Zwei bzw. Eins gesunken ist, werden natürlich nur noch eine bzw. keine Iterationen mehr benötigt.

Tabelle 3 gibt die mittlere Anzahl der benötigten Iterationsschritte für 100 Polynome vom Grad $n = 10$. Die Koeffizienten A_m wurden als komplexwertige

Tabelle 1. *Anzahl der benötigten Iterationen pro Wurzel. Mittelwert über alle Wurzeln eines Polynoms und Mittelwert über* 100 *Polynome. Die beteiligten Wurzeln* Z_ρ *wurden im Quadrat* $|\mathrm{Re}\, Z| \leqq 2{,}0$; $|\mathrm{Im}\, Z| \leqq 2{,}0$ *als Zufallszahlen erzeugt. Anfangswert jeweils* $z^* := -a_{n-1}/n a_n$ *nach Gl.* (4)

	Polynomgrad n																
	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	20	25	30
Anzahl der benötigten Iterationen	0	0,50	1,83	2,66	3,28	3,53	3,82	4,16	4,27	4,46	4,68	4,74	4,83	4,94	5,06	5,37	5,66	5,78
Größte beobachtete Anzahl der Iterationen	0	1	7	10	10	9	11	12	12	11	11	11	12	12	14	15	16	15
Anzahl der Zusatzschritte (Halbierungsverfahren)	0	0	0	0,06	0,10	0,10	0,13	0,18	0,20	0,22	0,26	0,27	0,28	0,30	0,31	0,41	0,50	0,57
Größte beobachtete Anzahl der Zusatzschritte	0	0	0	2	3	2	4	5	4	3	3	4	5	4	4	6	7	6

Tabelle 2. $n=20$. *Anzahl der Iterationen pro Wurzel. Mittelwert über* 100 *Polynome.* $p=$ *Nummer der Wurzel in der Reihenfolge der Berechnung* ($p=1$: *erste berechnete Wurzel, das reduzierte Ploynom ist gleich dem ursprünglichen Polynom.* $p=20$: *letzte berechnete Wurzel, das reduzierte Polynom hat den Grad Eins*). *Anfangswert jeweils* $z^* := -a_{n-1}/na_n$ *nach Gl.* (4). *Die beteiligten Wurzeln* Z_p *wurden im Quadrat* $|\mathrm{Re}\, Z| \leqq 2{,}0$; $|\mathrm{Im}\, Z| \leqq 2{,}0$ *als Zufallszahlen erzeugt*

	p																				
	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	Mittelwert
Anzahl der benötigten Iterationen	5,26	5,57	5,54	5,97	6,03	6,05	6,03	6,34	6,21	6,40	6,36	6,48	6,34	6,13	5,93	5,55	5,41	4,74	1,00	0,00	5,37
Größte beobachtete Anzahl der Iterationen	10	11	11	10	12	11	13	10	12	15	12	11	11	12	11	9	10	7	1	0	
Anzahl der Zusatzschritte (Halbierungsverfahren)	0,16	0,25	0,34	0,48	0,56	0,48	0,59	0,68	0,54	0,66	0,58	0,68	0,65	0,52	0,43	0,30	0,35	0,02	0	0	0,41
Größte beobachtete Anzahl der Zusatzschritte	2	3	4	3	4	3	5	3	5	6	3	3	4	4	3	2	3	1	0	0	

Zufallszahlen im Bereich $|ReA_m| \leqq 1$, $|ImA_m| \leqq 1$ (erste Zeile), bzw. als reellwertige Zufallszahlen im Intervall $|A_m| \leqq 1$ (zweite Zeile) erzeugt. Erster Anfangswert $z^* := 0$, nach Berechnung der Nullstellennäherung z_p wird $z^* := \bar{z}_p$ gesetzt. Der Vergleich der ersten Zeile mit der Tabelle 1 bzw. 2 zeigt keine wesentliche Erhöhung der Anzahl der Iterationen durch die veränderten Anfangswerte. (Man beachte jedoch die Tatsache, daß jetzt für $n = 1$ *stets eine* und für $n = 2$ *mehr als eine* Iteration erforderlich ist). Beim Übergang von komplexwertigen (Zeile 1) zu reellwertigen Koeffizienten (Zeile 2) fällt dagegen die Anzahl der benötigten Iterationen wie erwartet sehr stark, wenn auch nicht auf die Hälfte.

Tabelle 3. $n = 10$. *Mittlere Anzahl der Iterationen pro Wurzel für je* 100 *Polynome.* p = *Nummer der Wurzel in der Reihenfolge der Berechnung wie bei Tabelle 2. Anfangswert für* $p = 1$ *jeweils* $z^* := 0$, *anschließend* $z^* := \bar{z}_p$ *für* $p = 2(1)n$

	p										
	1	2	3	4	5	6	7	8	9	10	Mittelwert
A_m komplexwertig mit $\|\mathrm{Re}\, A_m\| \leqq 1$, $\|\mathrm{Im}\, A_m\| \leqq 1$	5,65	5,45	5,99	5,74	5,51	5,53	5,63	4,80	4,80	1,00	5,01
A_m reellwertig mit $\|A_m\| \leqq 1$	5,68	4,30	4,50	4,38	4,56	3,21	3,95	2,77	3,81	0,50	3,77

Erreichte Genauigkeit, Fehlerschranken

In Fig. 3 sind die auftretenden relativen Fehler $\frac{|Z_p - Z^*|}{|Z_p|}$ für die Fälle $n = 1(1)\,10, 15, 20, 25, 30$ als kleine Kreise, verbunden durch ausgezogene Linien, aufgetragen[11]. Sie stellen jeweils den Mittelwert über 100 Polynome und über sämtliche Wurzeln jedes dieser Polynome dar. Z. B. sind daher an dem Punkt für $n = 25$ insgesamt 2500 Polynomwurzeln beteiligt. Die beteiligten Wurzeln Z_p wurden im Quadrat $|Re\, Z| \leqq 2{,}0$; $|Im\, Z| \leqq 2{,}0$ als Zufallszahlen erzeugt.

Diese mittleren Fehler beginnen bei $n = 1$ mit 0 (d. h., die vorgegebenen Wurzeln werden wegen $A_n = 1$ selbstverständlich exakt wiedergefunden) und starten bei $n = 2$ mit $3{,}0 \cdot 10^{-8}$ ($\approx 10^{-8}$ bedeutet die Grenze der Maschinengenauigkeit bei einfacher Wortlänge). Zwischen $n = 2$ und $n = 12$ gehen für $\Delta n = 5$ je etwa 2 Zehnerpotenzen Genauigkeit verloren. Anschließend steigen die Fehler langsamer bis zu $1{,}0 \cdot 10^{-2}$ bei $n = 30$. Unter der Annahme, daß sich „Schutzstellen" additiv verhalten, kann man daraus die notwendige Ziffernlänge extrapolieren, wenn eine gewisse mittlere Genauigkeit erwartet wird. Sollen z. B. für $n = 15$ im Mittel 10 Ziffern des Ergebnisses richtig sein, so wäre eine Zahlenlänge von mindestens 15 Ziffern notwendig.

Die ungünstigsten Fälle sind durch kleine Kreuze markiert und ebenfalls durch ausgezogene Geraden verbunden. Sie liegen im Durchschnitt um etwa 2 Zehnerpotenzen über den mittleren Fehlern. Um also auch im (statistisch über 100 Poly-

[11] Da n ganzzahlig ist, gibt es nur diskrete Werte. Die hier und in den nachfolgenden Abbildungen gezeichneten Verbindungsgeraden zwischen den Rechenpunkten haben daher selbstverständlich keine selbständige Bedeutung; sie sollen nur zusammengehörige Punkte kennzeichnen und auf einen funktionellen Verlauf hinweisen.

nome ermittelten) ungünstigsten Falle für $n = 15$ noch 10 Ergebniszíffern sichern zu können, wären mindestens 17 Ziffern (mehr als doppelte Genauigkeit der UNIVAC 1107) pro Zahl erforderlich.

Die Fehlerschranken nach Formel (16) sind jeweils wieder durch Kreise bzw. Kreuze markiert, wenn es sich um mittlere bzw. ungünstigste relative Fehlerschranken handelt. Wird die Formel (16) so ausgewertet, wie es in dem vorliegenden Algorithmus angegeben wurde, so sind die Punkte durch eine gestrichelte

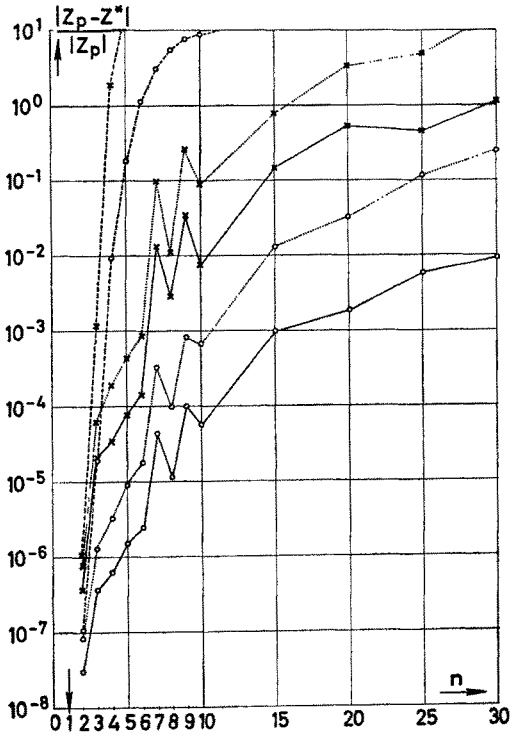

Fig. 3. Relativer Fehler $|Z_p - Z^*|/|Z_p|$ des Näherungswerts Z^* der p-ten Wurzel Z_p für verschiedene Werte des Polynomgrads n. Vergleich über alle Wurzeln von je 100 Polynomen, deren Nullstellen Z_p als Zufallszahlen im Quadrat $|\mathrm{Re}\, Z| \leqq 2$, $|\mathrm{Im}\, Z| \leqq 2$ erzeugt wurden. O——O Mittelwerte des relativen Fehlers. ×——× Größte beobachteten Werte des relativen Fehlers. O– – –O Mittelwerte der Fehlerschranken des relativen Fehlers nach Formel (16). × – – – × Größte beobachteten Werte der Fehlerschranken des relativen Fehlers nach Formel (16). O · · · · · · O Mittelwerte der Fehlerschranken des relativen Fehlers, aus dem nichtreduzierten Polynom bestimmt. × · · · · · · × Größte beobachteten Werte der Fehlerschranken des relativen Fehlers, aus dem nichtreduzierten Polynom bestimmt

Linie verbunden. Offenbar ist diese Vorschrift außerordentlich ungünstig, da die Schrankenwerte schon bei $n = 4$ bzw. $n = 6$ den Wert $1 = 10°$ übertreffen, d.h. einen Fehler von 100% und mehr zulassen. Zum Vergleich wurde daher noch derjenige Algorithmus ausgewertet, der die Fehlerschranke stets aus dem Koeffizienten des ursprünglichen (nicht reduzierten) Polynoms bestimmt. Die entsprechenden Punkte sind durch punktierte Linien verbunden.

Offensichtlich werden in diesem Falle die wirklichen Fehler durch die berechnete Schranke nur noch um rund eine Zehnerpotenz überschätzt. Dieser sehr günstigen Eigenschaft steht der Nachteil gegenüber, daß die ermittelten Fehler-

schranken nur im Falle einzelner Nullstellen (der bei den untersuchten Polynomen fast immer vorliegen dürfte) exakt sind.

In Fig. 4 sind die Werte des relativen Fehlers für alle Wurzeln von 100 Polynomen vom Grade $n=25$ ausgewertet. Aufgetragen sind wieder die mittleren Fehler (Kreise) und die ungünstigsten Fehler (Kreuze), auf die zugehörigen Fehlerschranken wurde verzichtet. Der Fehlerverlauf über der Nummer p der Reihenfolge der Bestimmung der Wurzeln ist typisch für alle untersuchten Polynome: Steigend bis zu einem Maximum zwischen $p=3$ bzw. $p=9$ und danach wieder fallend bis $p=25$. Die Deutung ist nicht schwer: Die Koeffizienten des Ausgangspolynoms sind noch fehlerfrei, während jeder Reduktionsschritt unvermeidliche Rundungsfehler akkumuliert. Dadurch steigen die Fehler zunächst mit wachsender Nummer p. Betrachtet man den absoluten Fehler der p-ten Wurzel, so hält diese monotone Zunahme sogar bis $n=25$ an. Da jedoch die Wurzeln bei dem vorliegenden Verfahren für wachsende Nummer p immer größere Beträge aufweisen, werden die relativen Fehler nach einem Maximum wieder kleiner. Bis $n=5$ wird dieses Maximum schon bei $p=2$ erreicht, mit wachsendem n verschiebt es sich zu größeren Werten p.

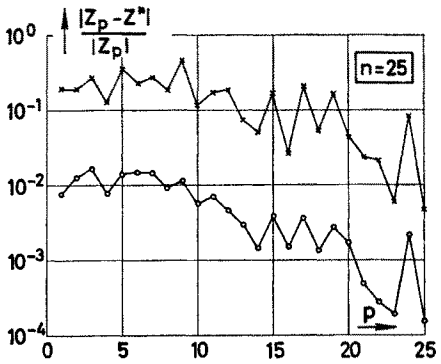

Fig. 4. Relativer Fehler $|Z_p-Z^*|/|Z_p|$ des Näherungswerts Z^* der p-ten Wurzel Z_p von 100 Polynomen vom Grade $n=25$, aufgetragen über p. Die Wurzeln Z_p waren als Zufallszahlen im Quadrat $|\mathrm{Re}Z|\leqq 2$, $|\mathrm{Im}Z|\leqq 2$ erzeugt worden. O——O Mittelwerte des relativen Fehlers. ×——× Größte beobachteten Werte des relativen Fehlers

Mehrfache Nullstellen, Wurzelhaufen

Während die meisten propagierten Verfahren bei mehrfachen Nullstellen und Wurzelhaufen versagen, erweist sich der vorliegende Algorithmus als weitgehend unempfindlich gegen solche Wurzelkonstellationen. Eine der vielen untersuchten

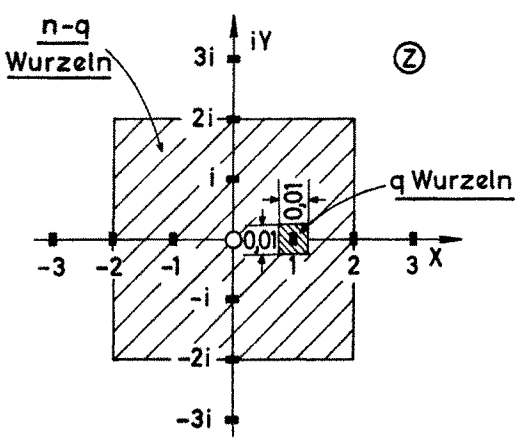

Fig. 5. Lage der Wurzeln in der Z-Ebene

Kombinationen war die folgende: Von den n Wurzeln des Polynoms liegen q in dem Quadrat $|Re\,Z-1| \leqq 0{,}01$, $|Im\,Z| \leqq 0{,}01$, während die übrigen $n-q$ als

Tabelle 4. $n=10$; *Wurzelhaufen mit q Nullstellen im Quadrat* $|\mathrm{Re}\,Z-1| \leqq 0{,}01$; $|\mathrm{Im}\,Z| \leqq 0{,}01$; *die restlichen* $n-q$ *Nullstellen liegen im Quadrat* $|\mathrm{Re}\,Z| \leqq 2$, $|\mathrm{Im}\,Z| \leqq 2$. *Anzahl der benötigten Iterationen pro Wurzel. Mittelwert über alle Wurzeln eines Polynoms und Mittelwert über* 100 *Polynome. Anfangswert jeweils* $z^* := -a_{n-1}/n a_n$ *nach Gl.* (4)

	q										
	0	1	2	3	4	5	6	7	8	9	10
Anzahl der benötigten Iterationen	4,48	4,49	4,78	5,14	5,23	5,39	5,14	5,02	4,10	3,40	1,49
Größte beobachtete Anzahl der Iterationen	11	11	13	14	15	15	15	16	15	14	9
Anzahl der Zusatzschritte (Halbierungsverfahren)	0,22	0,21	0,22	0,17	0,13	0,17	0,17	0,19	0,19	0,24	0,24
Größte beobachtete Anzahl der Zusatzschritte	4	4	5	4	2	3	3	3	4	6	4

Zufallszahlen in dem größeren Quadrat $|Re\,Z| \leqq 2$, $|Im\,Z| \leqq 2$ erzeugt werden, vgl. Fig. 5. Dabei ist $q=1\,(1)\,n$. In Tabelle 4 sind für $n=10$ analog zur Tabelle 2 und 3 die Anzahlen der benötigten Iterationen eingetragen. Die dritte und vierte Zeile (mittlere bzw. maximale Anzahl der Zusatzschritte) zeigt kaum einen Einfluß von q. Aber auch in den ersten beiden Zeilen (mittlere bzw. maximale Anzahl aller Iterationsschritte) ist der Einfluß von q nur wenig spürbar: Vielfache Nullstellen oder Wurzelhaufen beeinflussen die Konvergenzgeschwindigkeit des betrachteten Verfahrens nicht wesentlich.

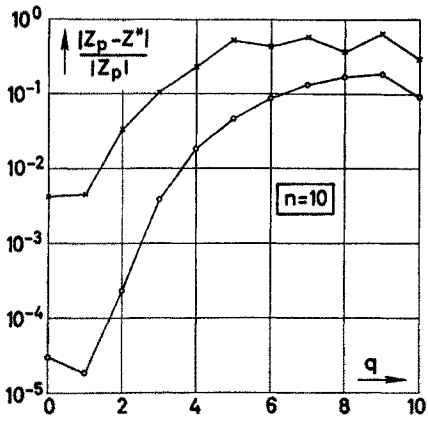

Fig. 6. Relativer Fehler $|Z_p-Z^*|/|Z_p|$ des Näherungswerts Z^* der p-ten Wurzel Z_p von 11mal je 100 Polynomen vom Grade $n=10$. Jeweils q Wurzeln liegen in dem kleinen Quadrat der Abb. 5, alle $n=10$ Wurzeln liegen in dem großen Quadrat. Vergleich über alle n Wurzeln und alle 100 Polynome. O——O Mittelwerte des relativen Fehlers. ×——× Größte beobachteten Werte des relativen Fehlers

Ungünstiger steht es natürlich mit der erzielbaren Genauigkeit. Daß jede Fehlerabschätzung zusammenbrechen muß, wenn nicht extrem lange Zahlen verarbeitet werden, zeigt das folgende Beispiel:

Es sei $n=p=10$ und der Näherungswert z zu der unbekannten, 10fachen Nullstelle Z so genau bestimmt, daß $|P(z)| \leqq 10^{-10}$ ist. Dann gilt $k=10$ in der Fehlerabschätzung (18) und man findet für ein normiertes Polynom $P(z)$ die Ungleichung

$$|Z-z| \leqq \sqrt[10]{|P(z)|} \leqq 10^{-1}.$$

Eine 10-ziffrige Genauigkeit für eine Nullstelle Z mit $|Z| \approx 1$ würde daher eine Berechnung von z und $P(z)$ bis auf $|P(z)| \leqq 10^{-100}$ verlangen: Es müßte daher mit mindestens 100-ziffrigen Zahlen gerechnet werden!!! In Fig. 6 ist das Anwachsen des relativen Fehlers der Wurzeln über p aufgetragen. Kreise bzw. Kreuze bedeuten wieder Mittelwerte über 100 Polynome und alle Polynomwurzeln bzw. die jeweils ungünstigsten Fälle. Man beobachtet — wie erwartet — ein starkes Anwachsen des mittleren Fehlers und zwar um 4 Zehnerpotenzen. Nach den bekannten Eigenschaften von Polynomen und den obigen Bemerkungen darf dieser starke Genauigkeitsverlust nicht dem verwendeten Verfahren zur Last gelegt, sondern muß als typisch für die Nullstellenbestimmung von Polynomen angesehen werden. Im Falle des Auftretens vieler benachbarter oder gar mehrfacher Nullstellen kann eine vorgeschriebene Genauigkeit bei Verfahren, die vom Polynomwert $P(z)$ ausgehen, im allgemeinen nur dadurch erreicht werden, daß mit extremer Ziffernlänge gearbeitet wird. Methoden, die mit Abzählkriterien analog zum Routh-Kriterium arbeiten — wie etwa das Lehmer-Verfahren — könnten im Prinzip dort günstigere Resultate erzielen. Das praktische Stabilitätsverhalten von solchen Algorithmen scheint jedoch noch nicht hinreichend untersucht zu sein.

Ausblick. Es ist mir wichtig, zu betonen, daß das vorstehend beschriebene Verfahren mannigfach abgeändert werden kann und werden sollte. Einige wenige Vorschläge wurden in Teil 3 gemacht, sehr viele weitere sind naheliegend. Es ist jedoch zu wünschen, daß Modifikationen und Konkurrenzmethoden, die nach anderen Prinzipien arbeiten, ebenfalls an einem möglichst großen Zahlenmaterial erprobt und die Ergebnisse mit den hier mitgeteilten verglichen werden. Es dürfte dann möglich sein, in Erweiterung der hier mitgeteilten Prinzipien endlich einen Algorithmus zu schaffen, der zu jedem beliebigen Polynom alle Nullstellen mit vorgeschriebener Genauigkeit liefert. Dies kann natürlich nur dadurch geschehen, daß der Rechenautomat in dynamischer Weise mit 10, 20, 100 oder noch mehr Ziffern rechnet, je nach der Kondition des eingegebenen Polynoms.

Literatur

[1] FEKETE, M.: Analoga zu den Sätzen von Rolle und Bolzano für komplexe Polynome und Potenzreihen mit Lücken. Jahresbericht der DMV **32**, 299—306 (1923).

[2] LEHMER, D. H.: A machine method for solving polynomial equations. Journal of the ACM **8**, 151—162 (1961).

[3] NASITTA, KH.: Ein immer konvergentes Nullstellenverfahren für analytische Funktionen. ZAMM **44**, 57—63 (1964).

[4] NICKEL, K.: Über die Notwendigkeit einer Fehlerschranken-Arithmetik für Rechenautomaten. Numer. Math. **6**, 69—79 (1966).

[5] WILKINSON, J. H.: Rounding errors in algebraic processes. Englewood Cliffs, (N. J.): Prentice-Hall Inc. 1963.

[6] ZURMÜHL, R.: Praktische Mathematik, 5. Aufl. Berlin-Heidelberg-New York: Springer 1965.

Institut für Angewandte Mathematik
der Technischen Hochschule Karlsruhe
7500 Karlsruhe