BAYERISCHE AKADMIE DER WISSENSCHAFTEN
MATHEMATISCH-NATURWISSENSCHAFTLICHE KLASSE

ABHANDLUNGEN · NEUE FOLGE, HEFT 95

KLAUS SAMELSON

Faktorisierung von Polynomen durch funktionale Iteration

Vorgelegt von Herrn Robert Sauer
am 7. März 1958

MÜNCHEN 1959
VERLAG DER BAYERISCHEN AKADEMIE DER WISSENSCHAFTEN
IN KOMMISSION BEI DER C. H. BECK'SCHEN VERLAGSBUCHHANDLUNG MÜNCHEN

Gedruckt mit Unterstützung der Deutschen Forschungsgemeinschaft

Druck der C. H. Beck'schen Buchdruckerei Nördlingen
Printed in Germany

INHALT

I. EINLEITUNG

Die rasche Entwicklung der programmgesteuerten elektronischen Rechenautomaten in den letzten Jahren hat der numerischen Analysis mannigfache Anregungen gebracht. Sie gibt insbesondere dazu Anlaß, die klassischen Rechenverfahren kritisch zu sichten und in eine systematische Ordnung nach übergeordneten Gesichtspunkten zu bringen, sowie an Hand der dabei gewonnenen Erkenntnisse neue, für die Verwendung an Rechenautomaten geeignete Verfahren zu entwickeln. Einen Beitrag in diesen Richtungen auf dem Gebiet der Auflösung algebraischer Gleichungen soll die vorliegende Arbeit leisten.

Die Verfahren zur Auflösung der Gleichung

$$P(z) = \sum_{i=0}^{n} p_i z^{n-i} = 0$$

sind grundsätzlich infinit und daher von iterativer Natur. Hinsichtlich ihres Konvergenzverhaltens lassen sie sich in zwei Klassen einordnen.

Die Verfahren der ersten Klasse, deren wichtigster Vertreter das Verfahren von BERNOULLI ist, konvergieren in gewissen Gebieten der z-Ebene, die je eine Nullstelle von $P(z)$ enthalten, und konvergieren nicht in dem mehrfach zusammenhängenden Komplementärbereich. Sie sollen wegen ihres klaren Konvergenzverhaltens weiterhin als definite Verfahren bezeichnet werden. Die Verfahren der zweiten Klasse, zu der z. B. das Newtonsche Verfahren gehört, beruhen auf Abbildungen der z-Ebene, die die Wurzel von $P(z)$ zu Fixpunkten haben. Die Bestimmung der Wurzeln geschieht durch Iteration der Abbildung geeignet gewählter Ausgangspunkte. Hinsichtlich der Konvergenz gelten die allgemeinen Fixpunktsätze, es gibt jedoch keine Zerlegung der Ebene in endlich viele Konvergenz- und Nichtkonvergenz-Gebiete wie im Falle der definiten Verfahren. Die Verfahren dieser zweiten Klasse sollen als Verfahren der funktionalen Iteration im engeren Sinne bezeichnet werden.

In Abschnitt II wird eine im Sinne des Graeffeverfahrens quadratisch konvergente Abkürzung eines von BAUER angegebenen Faktorisierungsverfahrens beschrieben. Abschnitt III ist den Methoden der funktionalen Iteration gewidmet, die sämtlich, wie sich zeigt, aus dem Prinzip der Nullpunktverschiebung an definiten Verfahren ableitbar sind. In Abschnitt IV wird aus diesem Prinzip ein allgemeines Verfahren der funktionalen Iteration zur Faktorisierung abgeleitet, dessen einfachster Fall als quadratisch konvergent im Sinne des Newtonverfahrens nachgewiesen wird. Abschnitt V enthält die Diskussion des Spezialfalls der Abspaltung eines Linearfaktors, für den sich etwas weitergehende Konvergenzaussagen machen lassen.

II. DEFINITE VERFAHREN DER NULLSTELLENBESTIMMUNG

II.1. Das Verfahren von BERNOULLI

Fast alle praktisch wichtigen definiten Verfahren zur Bestimmung der Nullstellen eines Polynoms $P(z) = \sum_{i=0}^{n} p_i z^{n-i}$ lassen sich auf das klassische Verfahren von BERNOULLI zurückführen. Dieses Verfahren liefert bekanntlich mit Hilfe der Differenzengleichung

$$a_i = -p_1 a_{i-1} - p_2 a_{i-2} - \ldots\ldots - p_n a_{i-n} \tag{II. 1. 1.}$$

eine Folge gewichteter Potenzsummen

$$a_i = \sum_{j=1}^{n} e_j \zeta_j^i \qquad (i = 1, 2, \ldots)$$

der Wurzeln ζ_j von $P(z)$ bei willkürlicher Vorgabe von n Ausgangsgrößen $a_0, a_1, \ldots a_{n-1}$, die die Gewichte e_j eindeutig festlegen. Besitzt $P(z)$ eine betragsgrößte Wurzel ζ_1, so gilt unabhängig von der Wahl der Ausgangsgrößen[1]

$$\lim_{i \to \infty} \frac{a_{i+1}}{a_i} = \zeta_1.$$

Das Konvergenzverhalten steht also von vornherein fest. Die Konvergenz ist linear, d. h. der Fehler niedrigster Ordnung multipliziert sich in der Grenze bei jedem Schritt mit dem festen Konvergenzfaktor ζ_1/ζ_2. Das Verfahren ist selbstkorrigierend, d. h. ein einmaliger Fehler im Laufe der Rechnung führt zwar zu einer Veränderung der Folge der a_i (ein Fehler bedeutet ja eine Neufestlegung der Gewichte e_j), jedoch hat die neue Folge den gleichen Grenzwert wie die ursprüngliche. Besonders bemerkenswert ist, daß diese Aussage sogar für beliebig große Fehler gilt.

An das Verfahren von BERNOULLI schließt sich (historisch und systematisch) eine beträchtliche Zahl von weiteren Verfahren als Spezialfälle, Verallgemeinerungen und Abkürzungen an. Als wichtigste Verfahren sind dabei zu nennen:

Der Quotienten-Differenzen-(*Q-D*-)Algorithmus von RUTISHAUSER (21) und die im folgenden beschriebene Treppeniteration von BAUER (6, 7) als linear konvergente Prozesse sowie das GRAEFFE-Verfahren, die *A-P*-Transformation von RUTISHAUSER und BAUER (23) in Anwendung auf die zu $P(z)$ gehörige Frobeniusmatrix

---

[1] (Abgesehen von dem Ausnahmefall $e_1 = 0$).

$$\mathfrak{F} = \begin{bmatrix} -p_1 & 1 & & \\ -p_2 & & 1 & \\ \cdot & & & \cdot \\ \cdot & & & \cdot \\ -p_{n-1} & & & 1 \\ -p_n & & & \end{bmatrix}$$

und die abgekürzte Iteration von BAUER (3), die alle quadratisch konvergieren.

II.2. Die Treppeniteration

Eine echte, weitreichende Verallgemeinerung des Bernoullischen Verfahrens mit den gleichen Konvergenz- und Selbstkorrektureigenschaften wie das letztere ist dagegen die bereits erwähnte Treppeniteration von BAUER (7) sowie die damit eng verwandte direkte Faktorisierung des gleichen Autors (6). Die Treppeniteration läßt sich in bequemer Weise als Iterationsverfahren an der Frobeniusmatrix $\mathfrak{F}$ von $P(z)$ darstellen, was besonders auch für die anschließend gegebene Entwicklung der quadratischen Abkürzung von Vorteil ist und außerdem in einfacher Weise zu der direkten Faktorisierung führt. Das Verfahren ist dann folgendermaßen zu beschreiben:

Es sei Γ_0 eine beliebige n-zeilige und r-spaltige ($1 < r < n$) Matrix von Trapezgestalt, d. h. Γ_0 sei oberhalb der von der linken oberen Ecke ausgehenden Diagonale mit Nullen und in der Diagonale selbst mit Einsen besetzt. Für $i = 0, 1, 2, \ldots$ werde nun aus Γ_i durch Abspaltung einer $r \times r$-reihigen oberen Dreiecksmatrix R_{i+1} nach der Beziehung

$$\mathfrak{F} \cdot \Gamma_i = \Gamma_{i+1} R_{i+1}$$

eine neue Trapezmatrix Γ_{i+1} hergestellt. In R_{i+1} sind dabei stets nur die Diagonale und die erste obere Nachbardiagonale besetzt, die letztere mit Einsen. Die Diagonalelemente streben mit wachsendem i gegen die r betragsgrößten Wurzeln ζ_k von $P(z)$, wenn nur $|\zeta_1| > |\zeta_2| > \ldots > |\zeta_r| > |\zeta_{r+1}|$. Gleichzeitig streben die Elemente der Spalten von Γ_i gegen die Koeffizienten der Polynome $P_1(z) = P(z)/(z - \zeta_1)$, $P_{12}(z) = P_1(z)/(z - \zeta_2), \ldots$ $P_{12\ldots r} = P_{12\cdot r-1}(z)/(z - \zeta_r)$.

Den sogenannten Fürstenauschen Fall erhält man, wenn man für Γ_0 die ersten r Spalten der n-dimensionalen Einheitsmatrix wählt.

Die Faktorisierung, d. h. die iterative Zerlegung von $P(x)$ in zwei Faktoren $Z(z)$ und $V(z)$, läßt sich nun aus der Treppeniteration gewinnen, wenn man an Stelle der trapezförmigen Iterationsmatrizen Γ_i Parallelogramm-Matrizen $\bar{\Gamma}_i$ einführt, die auch unterhalb der von der unteren rechten Ecke ausgehenden Diagonalen mit Nullen besetzt sind. An Stelle der bei jedem Iterationsschritt rechts abgespaltenen oberen Dreiecksmatrix tritt jetzt eine $r \times r$-reihige Matrix $\bar{\mathfrak{F}}_i$ von Frobeniusgestalt. Dabei stellen die jeweils letzte Spalte von $\bar{\Gamma}_i$ die neueste Approximation $U_i(z)$ an $U(z)$, die Frobeniusmatrix $\bar{\mathfrak{F}}_i$ die Approximation $V_i(z)$ an $V(z)$ dar. Die Beziehungen zwischen Treppeniteration und Faktorisierung sind also so eng, daß man durch einmalige Abspaltung des unteren Dreiecksanteils von der Iterationsmatrix $\bar{\Gamma}_i$ der Treppeniteration zur Faktorisierung übergehen kann. Es ist im übrigen auch möglich, diese Abspaltung des untersten Dreiecksanteils auf $s < r$ nebeneinanderliegende Spalten von $\bar{\Gamma}_i$ zu beschränken, was zum Beispiel ($s = 2$) von Nutzen sein kann, wenn die Iterationsfolgen für zwei benachbarte Spalten auf das Vorhandensein eines komplexen Paares hindeuten.

II. 3. Eine quadratische Abkürzung der Treppeniteration

Zu der Bauerschen Treppeniteration läßt sich eine quadratische Abkürzung konstruieren. Sie beruht auf folgender Eigenschaft der Frobeniusmatrix $\mathfrak{F}$:

Es sei e der Spaltenvektor $(0; 0; \ldots; 0; 1)^T$. Dann gilt für alle $i \geqslant 0$

$$\mathfrak{F}^i = (\mathfrak{F}^{i+n-1} \cdot e;\ \mathfrak{F}^{i+n-2} \cdot e;\ \ldots;\ \mathfrak{F}^i \cdot e). \tag{II. 3. 1}$$

Daher ist offensichtlich, daß man im Fürstenauschen Spezialfall mit Hilfe von $n - r + 1$ aufeinanderfolgenden Rechteckmatrizen $\mathfrak{F}^{i+r-n} \cdot \Gamma_0$, $\mathfrak{F}^{i+r-n+1} \cdot \Gamma_0$, $\ldots\ldots \mathfrak{F}^{i-1} \cdot \Gamma_0$, $\mathfrak{F}^i \cdot \Gamma_0$ durch Aneinanderreihung der passenden Spalten gerade die Matrix $\mathfrak{F}^i$ aufbauen kann, deren Anwendung auf $\mathfrak{F}^i \cdot \Gamma_0$ sofort $\mathfrak{F}^{2i} \cdot \Gamma_0$ ergibt.

Das wesentliche Problem dabei ist nun wie im Falle der Matrixpotenzierung, daß bei direkter Iteration ohne besondere Stabilisierungsmaßnahmen die Spalten der höheren Matrixpotenzen sehr schnell einander proportional werden, was bei der abschließend notwendigen Dreieckszerlegung zu progressiver Auslöschung führt.

Man muß daher ähnlich wie bei der AP-Transformation von vornherein die für den quadratischen Schnitt benötigte Matrix $\mathfrak{F}^i$ aus den bereits zerlegten Teilmatrizen $\mathfrak{F}^{i-j} \cdot \Gamma_0 = \Gamma_{i-j} \cdot P_{i-j}$ derart in Links- und Rechtsbestandteil A und B getrennt aufbauen, daß im Ausdruck $\mathfrak{F}^i \cdot \Gamma_i P_i = A B \Gamma_i P_i$ nur das innere Produkt $B\Gamma_i$ eine Trapezzerlegung und Abspaltung eines oberen Dreiecks P_i', $B\Gamma_i = \Gamma_i' P_i'$, erfordert und die abschließenden Multiplikationen $A\Gamma_i'$ bzw. $P_i' P_i$ die Trapezgestalt des ersten bzw. die Dreiecksgestalt des zweiten Produkts erhalten.

Daraus ergibt sich folgende Minimalforderung an die Gestalt von A: In den ersten $r - 1$ Spalten muß es von der gleichen Trapezgestalt sein wie die Γ-Matrizen. Das Spiegelbild dieses Trapezes an der Diagonalen, d. h. die ersten $r - 1$ Zeilen von der Diagonale (ausschließlich) an, muß mit Nullen besetzt sein. Der Rest, also das rechts von der $r - 1$-ten Spalte und unter der $r - 1$-ten Zeile liegende Quadrat, darf voll besetzt sein. Die Matrix A hat also folgendes Aussehen:

$$\begin{bmatrix} 1 & 0 & 0 & 0 & 0 & 0 & 0 \\ * & 1 & 0 & 0 & 0 & 0 & 0 \\ * & * & 1 & 0 & 0 & 0 & 0 \\ * & * & * & 1 & 1 & 1 & 1 \\ * & * & * & * & * & * & * \\ * & * & * & * & * & * & * \\ * & * & * & * & * & * & * \end{bmatrix}$$
$$\underbrace{\qquad\qquad}_{r}$$

Damit ist aber der tatsächliche Aufbau von A eine Selbstverständlichkeit: Als das führende Trapez ist Γ_i zu verwenden und die übrigen $n - r$ Spalten sind mit den jeweils letzten, wie verlangt gerade $n - r + 1$-komponentigen, Spalten von Γ_{i-1}, Γ_{i-2}, $\ldots$, Γ_{i-n+r} zu besetzen.

B ist eine obere Dreiecksmatrix, deren erste r Spalten aus den entsprechenden Spalten von P_i gebildet sind. Die weiteren Spalten von B sind aus den P_{i-j} so zu bestimmen, daß das Produkt AB mit $\mathfrak{F}^i$ übereinstimmt. Wie dies spaltenweise geschehen kann, soll am Beispiel der $r + 1$-ten Spalte von B gezeigt werden. Diese Spalte gibt diejenige Linearkombination der Spalten von Γ_i sowie der letzten Spalte von Γ_{i-1} an, die die $r + 1$-te Spalte von $\mathfrak{F}^i$ liefert.

Es sei nun S_{i-1} die obere Dreiecksmatrix, die aus P_{i-1} dadurch entsteht, daß die letzte Spalte durch die Einheitsspalte $(0, 0 \ldots, 1)^T$ ersetzt wird. Bildet man nun $W_{i-1} = S_{i-1}^{-1} P_{i-1}$, so ist darin nur die letzte Spalte voll besetzt, im übrigen enthält sie Einsen in der Diagonale und Nullen sonst. Dementsprechend sind die ersten $r-1$ Spalten von $T_{i-1} = \Gamma_{i-1} S_{i-1}$ direkt die zweite bis r-te Spalte von $\mathfrak{F}^i$, da $T_{i-1} W_{i-1} = \Gamma_{i-1} P_{i-1} = \mathfrak{F}^{i-1} \cdot \Gamma_0$ ist. Die letzte Spalte von W_{i-1} gibt an, wie sich die $r+1$-te Spalte von $\mathfrak{F}^i$ aus den $r-1$ vorangehenden Spalten von $\mathfrak{F}^i$ selbst und aus der letzten Spalte von Γ_{i-1} linear kombiniert. Die $r-1$ vorangehenden Spalten von $\mathfrak{F}^i$ sind aber die zweite bis r-te Spalte von $\Gamma_i P_i$, und man erhält sie bereits durch Linearkombination der zweiten bis r-ten Spalte von P_i gemäß den ersten $r-1$ Komponenten der letzten Spalte von W_{i-1}.

Ergänzt man den entstehenden r-komponentigen Spaltenvektor unten durch das rechte untere Eckelement von W_{i-1} (und weitere $n-r-1$ Nullen), so ist damit die $r+1$-te Spalte des Rechtsfaktors B von $\mathfrak{F}^i$ fertig aufgebaut.

Auf die gleiche Weise sind nun auch die folgenden Spalten aufzubauen: Man bestimmt die letzte Spalte von W_{i-j}, kombiniert entsprechend den ersten $r-1$ Komponenten die letztvorangegangenen $r-1$ Spalten von B und ergänzt nach unten durch das Eckelement von W_{i-j} und Nullen. Die Matrix B ergibt sich damit als eine volle obere Dreiecksmatrix.

Man erhält somit $\mathfrak{F}^i$ in einer insofern unorthodoxen Zerlegung, als der Linksbestandteil A keine vollständige untere Dreiecksmatrix ist. Durch die Trapezgestalt der führenden r Spalten wird aber ein Proportionalwerden dieser entscheidenden Spalten mit Sicherheit verhindert. Die restlichen Spalten werden einander (und der r-ten Spalte) naturgemäß schnell proportional, doch ist dies belanglos, da die gesamte relevante Information bereits in den ersten r Spalten enthalten ist. Es ist aus Normierungsgründen vielleicht sogar vorteilhaft, die $r+1$-te bis n-te Spalte von A durch die jeweilige Differenz gegen die r-te Spalte zu ersetzen. Gleichzeitig müssen dann in jeder Spalte des Rechtsbestandteils B alle unter dem Element der r-ten Zeile stehenden Komponenten zu diesem addiert werden.

Durch die sukzessiven Abkürzungsschritte wird eine Folge von Matrizenpaaren $(\Gamma_{i_j}; P_{i_j})$ erzeugt. Sie ist offensichtlich eine Teilfolge der aus der BAUERschen Treppeniteration im FÜRSTENAUschen Falle resultierenden Folge $(\Gamma_i; P_i)$, die durch $\mathfrak{F}^i \Gamma_0 = \Gamma_i P_i$ definiert ist. Für die Indizes der Teilfolge gilt

$$i_j = (2^{j+1} - 2) \cdot n - (2^j - 2) \cdot r \qquad (j = 1, 2, \ldots) \tag{II. 3. 2}$$

Asymptotisch verhält sich also i_j wie $(2n - r) \cdot 2^j$.

Das Verfahren konvergiert daher unter den gleichen Voraussetzungen wie die Treppeniteration, und zwar quadratisch im Sinne des Graeffe-Verfahrens.

Der Fall $r = 1$ ist die abgekürzte Iteration von BAUER, der Fall $r = n$ wird mit der A-P-Transformation identisch.

III. NULLSTELLENBESTIMMUNG DURCH FUNKTIONALE ITERATION

III. 1. Das Prinzip der funktionalen Iteration

Als erster hat wohl SCHRÖDER (24) die funktionale Iteration in Allgemeinheit als Mittel zur Nullstellenbestimmung vorgeschlagen, wenn auch die ältesten und bekanntesten Verfahren, nämlich die Regula falsi und das Newtonsche Verfahren, zu dieser Klasse gehören. SCHRÖDER beschreibt die Methode in folgender Weise:

„Es handelt sich dann darum, eine solche Funktion F aufzufinden, daß die Gleichung

$$z^+ = F(z)$$

stets einen Punkt z^+ liefert, welcher der Wurzel z_1 näher liegt als der anfänglich angenommene Punkt z."

Und weiter:

„Diese Gleichung lehrt, daß die Funktion F erstens in dem Punkt z_1 stetig sein, und zweitens die Bedingung: $F(z_1) = z_1$ erfüllen muß."

In diesen Sätzen ist das Prinzip der funktionalen Iteration klar ausgesprochen: Man konstruiere sich eine Abbildung des Lösungsraumes (hier der Zahlenebene), die die Lösungen zu Fixpunkten hat, und suche nun die Fixpunkte durch Iteration der Abbildung eines willkürlichen Ausgangselementes zu bestimmen.

Daher kann man sich hinsichtlich der Konvergenz der Iterationen auf den Fixpunktsatz stützen, während die Existenz der Lösungen schon durch den Fundamentalsatz der Algebra gesichert ist. Jedoch garantiert der Fixpunktsatz, der Abbildungen eines vollständigen metrischen Raumes voraussetzt, die Konvergenz nur in gewissen durch Lipschitzbedingungen definierten Umgebungen der Nullstellen und macht keinerlei Aussage über das Konvergenzverhalten im Komplementärbereich dieser Umgebungen. Daher gibt er auch im allgemeinen keinen Hinweis auf die Wahl des Ausgangspunktes der Iteration.

In einer für den vorliegenden Zweck ausreichenden Allgemeinheit lautet der Fixpunktsatz:

Genügt die Funktion $F(z)$ in einer gewissen Umgebung $U_r(\zeta) : |z - \zeta| < r$ der Lösung ζ (die Existenz ist vorausgesetzt) einer Lipschitzbedingung

$$|F(z') - F(z'')| < k \cdot |z' - z''| \qquad (z' \text{ und } z'' \text{ in } U_r(\zeta))$$

mit $0 < k < 1$, und liegt Punkt z_0 in $U_r(\zeta)$, so liegen alle Punkte der Folge $z_i = F(z_{i-1})$ $(i = 1, 2, \ldots)$ in $U_r(\zeta)$, und die Folge konvergiert gegen den Lösungspunkt ζ.

Ist nun $F(z)$ in einer gewissen Umgebung von ζ analytisch (oder auch nur stetig differenzierbar), und ist $|F'(\zeta)| < 1$, so gibt es sicher wenigstens ein Wertepaar k, r so, daß die Bedingungen des Satzes erfüllt sind.

Bei Ersetzung des Betrages durch eine geeignete Distanz gilt der Satz auch, was im Hinblick auf Abschnitt IV vermerkt sei, wenn $z = (z_1, z_2, \ldots, z_n)$ und $F = (F_1, F_2, \ldots, F_n)$ Punkte eines n-dimensionalen R_n sind. An die Stelle der Ableitung $F'(z)$ tritt dann die Funktionalmatrix $\partial F_i / \partial z_k$.

Die Definition der Ordnung der Konvergenz einer funktionalen Iteration schließt sich (ebenfalls nach SCHRÖDER) an das Verhalten der Ableitungen von F im Lösungspunkte ζ an. Ist die erste Ableitung von Null verschieden (aber natürlich dem Betrage nach kleiner als 1), so nennt man die Konvergenz linear. Im übrigen ist die Ordnung der ersten nicht verschwindenden Ableitung von F die Ordnung der Konvergenz. Eine Taylorentwicklung von F um den Lösungspunkt zeigt im übrigen, daß die Definition in der Grenze mit der für die Bernoullischen Verfahren mit Hilfe des Konvergenzfaktors abgeleiteten Definition übereinstimmt.

Die für die Verfahren der funktionellen Iteration typische Tatsache, daß sich die Konvergenz nur im kleinen nachweisen läßt, ist naturgemäß ein Nachteil gegenüber den Methoden vom Bernoullischen Typ, die eine globale Konvergenzaussage erlauben. Andererseits lassen sich bequem Iterationen zweiter und höherer Ordnung angeben, die, wie sich aus dem Konvergenzsatz ablesen läßt, wenigstens im kleinen selbstkorrigierend sind. Im

übrigen hat die Erfahrung gezeigt, daß die Konvergenzeigenschaften der wesentlichen Verfahren weit besser sind, als sich aus den Bedingungen des Konvergenzsatzes ablesen läßt. Man kann sogar sagen, daß sie praktisch fast immer konvergieren, da schon die Rundungsfehler dafür sorgen, daß theoretisch mögliche nichtkonvergente Folgen nicht stabil sind.

Was nun die das Verfahren bestimmende Wahl der Abbildung $F(z)$ anbetrifft, so hat schon SCHRÖDER bemerkt, daß praktisch wegen der Stationaritätsbedingung nur Funktionen in Frage kommen, die die Form

$$F(z) = z - W(z) = z - P(z) \cdot R(z)$$

mit weitgehend willkürlichem $R(z)$ haben, das nur der Bedingung

$$|1 - P'(\zeta) \cdot R(\zeta)| < 1$$

genügen muß.

Diese Bedingung läßt natürlich noch eine ungeheure Freiheit in der Wahl von W. Es zeigt sich aber, daß man Verfahren jeder Ordnung nach folgendem Prinzip erhält: Die Funktion $W(z)$ wird aus einer endlichen Zahl von Schritten eines von dem Ausgangspunkt z_0 gestarteten definiten Verfahrens abgeleitet. Es werden also in einem beliebigen Ausgangspunkt je nach der Ordnung der gewünschten Approximation ein oder mehrere Schritte eines definiten Verfahrens ausgeführt und damit ein neuer Näherungspunkt im Sinne der funktionalen Iteration bestimmt, von dem aus das Verfahren wiederholt wird.

Tatsächlich liefert ja ein in einem beliebigen Punkte z_0 des Konvergenzgebietes angesetztes definites Verfahren einen unendlichen Ausdruck $\zeta(z_0)$ für eine Wurzel ζ von $P(z)$. Faßt man diese Abbildung der Punkte z_0 des Konvergenzgebietes (ungeachtet der Tatsache, daß sie in ihren Einzelschritten eine Iteration darstellt) als einen Iterationsschritt auf, so liefern die definiten Verfahren solche ausgearteten Abbildungen in der z-Ebene, die ihre Definitionsbereiche auf die n Nullstellen von $P(z)$ abbilden. Die funktionale Iteration macht davon Gebrauch, indem sie den unendlichen Ausdruck durch einen endlichen Abschnitt ersetzt, in der begründeten Hoffnung, dadurch günstige Lipschitzkonstanten zu erzielen.

III. 2. Verfahren der funktionalen Iteration

Nach diesen grundsätzlichen Bemerkungen soll nun kurz auf die verschiedenen Klassen gebräuchlicher Iterationen eingegangen werden.

Die Klasse S_1

Eine erste Klasse läßt sich aus dem eigentlichen Bernoulliverfahren ableiten. Sie ist vollständig von SCHRÖDER behandelt worden und hat mit $G(z) = Q(z)/P(z)$ die Form

$$z^+ = z + i \left(\frac{d}{dz}\right)^{i-1} G(z) \Big/ \left(\frac{d}{dz}\right)^{i} G(z) \qquad (i = 1, 2, \ldots; \text{fest}). \tag{III. 2. 1}$$

Hinsichtlich der Wahl von $Q(z)$ kommen praktisch nur der Fürstenausche und der sogenannte Lagrangesche Spezialfall, $Q(z) = 1$ und $Q(z) = P'(z)$, in Frage (SCHRÖDER selbst hat allerdings auch die Fälle $Q(z) = z^k$, $k = 1, 2, \ldots, n-1$, diskutiert). Die Ordnung der Iteration ist $i + 1$. Für $i = 1$ ergibt der Fürstenaufall

$$z^+ = z + (1/P(z))/(1/P(z))' = z - P(z)/P'(z),$$

also das Newtonsche Verfahren. Der Lagrangesche Fall liefert für $i = 1$ die Iteration

$$z^+ = z - \frac{P}{P'} \cdot \frac{1}{1 - \frac{PP''}{P'^2}},$$

von der schon SCHRÖDER bemerkt, daß sie im Gegensatz zum Newtonschen Verfahren auch für mehrfache Nullstellen quadratisch konvergiert. Dies ist zwar richtig und aus der Tatsache, daß P'/P nur einfache Pole hat, sofort verständlich. Praktisch hat es aber wenig zu bedeuten, da die Aussage nur im Sinne eines numerisch kaum faßbaren Grenzüberganges für den unbestimmt werdenden Bruch im Nenner gilt.

Alle Verfahren dieser Klasse haben den Nachteil, daß die reelle Achse eine Ausnahmelinie ist, die freiwillig nicht verlassen wird. Für komplexe Rechnung muß ausdrücklich ein komplexer Anfangswert vorgegeben werden.

Die Klasse B

Diesen Mangel behebt eine Klasse von Verfahren, bei der zur Bestimmung der Korrektur $k = z^+ - z$ eine quadratische Hilfsgleichung, und zwar die Jacobische Determinantengleichung

$$\text{(III. 2. 2)} \qquad \begin{vmatrix} a_{i-2} & a_{i-1} & a_i \\ a_{i-1} & a_i & a_{i+1} \\ 1 & k^{-1} & k^{-2} \end{vmatrix} = 0 \qquad \begin{matrix} (i = 2, 3, \ldots; \text{fest}) \\ \left(a_i = \left(\frac{d}{dz}\right)^i G(z)\right) \end{matrix}$$

benützt wird. Als Korrektur ist der Wert k mit dem kleineren Betrage zu wählen, wobei die Annahme zugrunde gelegt wird, daß dies zu der näher gelegenen Nullstelle von $P(z)$ hinführt.

MAEHLY (18) hat kürzlich den Fall $i = 2$ diskutiert. Für $i = 1$ ist das Element a_{i-2} aus der Taylorentwicklung nicht definiert. Bei Zurückgehen zur Potenzsummendefinition ergibt sich $a_{-1} = n$. Auch diesen Fall hat MAEHLY behandelt und dabei gezeigt, daß sich aus der zusätzlichen Annahme einer $n - 1$-fachen Wurzel das Laguerresche Verfahren ergibt.

Die Verfahrensklasse in Allgemeinheit hat wohl zuerst BAUER (3) angegeben. Alle genannten Diskussionen legen übrigens stets den Lagrangeschen Fall zugrunde, der ja den Vorzug hat, mehrfache Nullstellen zu unterdrücken. Der Fürstenausche Fall dagegen, der einfachere Ausdrücke liefert, scheint keinerlei Beachtung gefunden zu haben.

Die Ordnung der Konvergenz ist in allen Fällen $i + 2$.

Die Klasse S_2

Als nächstes ist die zweite Schrödersche Klasse zu nennen, die die aus der Schröder-Thereminschen Reihe, d. h. der formalen Lagrange-Buermann-Entwicklung der Umkehrfunktion $z = z(P)$ von $P(z)$ für den Wert $P = 0$, abgeleiteten Verfahren umfaßt. Diese Verfahren haben die Form

$$\text{(III. 2. 3)} \qquad z^+ = z - \sum_{k=1}^{i} (-1)^{k-1} \frac{P^k}{k!} \left(\frac{1}{P'}\frac{d}{dz}\right)^{k-1} \frac{1}{P'}.$$

Für $i = 1$ ergibt sich wieder das Newtonsche Verfahren, für $i = 2$ die z. B. von OLVER (20) genannte Iteration

$$z^+ = z - \frac{P}{P'} - \frac{P^2 P''}{2P'^3} = z - \frac{P}{P'}\left(1 + \frac{PP''}{2P'^2}\right).$$

Die Konvergenz ist auch hier wieder von der Ordnung $i + 1$.

Die Klasse I

Als letztes schließlich ist die Klasse der inversen Interpolation mit der Regula falsi als dem wohl einzig gebräuchlichen Verfahren anzuführen. Man erhält sie analog der Ge-

winnung der Schröder-Thereminschen Reihe aus der Lagrange-Buermannschen Potenzreihe: Die Umkehrfunktion $z = z(P)$ von $P(z)$ wird mit Hilfe einer Serie fester Stützwerte $P_1, z_1; P_2, z_2; \ldots$ durch eine „Fakultätenreihe"

$$z = z_1 + a_1(P - P_1) + a_2(P - P_2)(P - P_1) + \ldots \tag{III. 2. 4}$$

dargestellt. Der Wert $P = 0$ muß auch hier, wenn die Reihe konvergiert, eine Wurzel ζ liefern. Der erste Abschnitt der Reihe liefert die Regula falsi mit einem festen Stützwert z_2. Denn zunächst ergibt sich a_1 durch Einsetzen von $P = P_2$ zu

$$a_1 = (z_2 - z_1)/(P_2 - P_1).$$

Damit aber ergibt der erste Abschnitt mit $P = 0$, wenn noch z^+ für z und z für z_1 gesetzt wird, die Formel der Regula falsi

$$z^+ = z - P(z) \cdot \frac{z - z_2}{P(z) - P(z_2)}.$$

Aus der Reihe (III. 2. 4) lassen sich naturgemäß durch Hinzunahme neuer Stützwerte Mehrpunktformeln der inversen Interpolation gewinnen, doch scheinen solche Formeln bisher nicht verwendet worden zu sein. Die Konvergenz ist für die Regula falsi linear, falls z_2 der Nullstelle ζ nahe genug liegt.

Wie man sieht, lassen sich aus den definiten Verfahren beliebig viele Iterationen beliebiger Ordnung ableiten. Trotzdem wird für die praktische Anwendung über die dritte Ordnung, die selbst schon selten genug benützt wird, kaum hinausgegangen, und die Norm ist fast immer das quadratisch konvergente Verfahren. Der Grund ist, daß die zu berechnenden Ausdrücke mit höherer Ordnung immer komplizierter werden, so daß es bequemer ist, mehrere kleine Schritte an Stelle eines großen zu machen.

III. 3. Das Verfahren von Bairstow

Grundsätzlich anders als auf dem Gebiet der Nullstellenbestimmung liegen die Dinge hinsichtlich der funktionalen Iteration auf dem Gebiet der Faktorisierung. Hier gibt es, abgesehen von einer direkten reellen Zusammenfassung des Newtonschen Verfahrens im Komplexen für konjugiert komplexe Paare, nur ein einziges Verfahren zur Bestimmung reeller quadratischer Faktoren, das von Bairstow (2) angegeben worden und von Collatz und Zurmühl (9, 28) in eine rechentechnisch übersichtliche Form gebracht worden ist. Das Verfahren ist eine Übertragung des Newtonschen Verfahrens. Ist $d(z) = z^2 + p_0 z + q_0$ eine Näherung für einen quadratischen Faktor und ergibt sich durch Division $P(z) = Q(z)d(z) + sz + t$, so sind damit s und t als Funktion von p und q definiert, die gleichzeitig verschwinden, wenn $d(z)$ Teiler von $P(z)$ ist. An die Stelle von $P(z)$ im Newtonschen Verfahren tritt damit das Funktionspaar s und t, an die Stelle der Ableitung $P'(z)$ die Funktionalmatrix, und die Formeln für die neuen Näherungen p und q lauten dementsprechend

$$\begin{vmatrix} p \\ q \end{vmatrix} = \begin{vmatrix} p_0 \\ q_0 \end{vmatrix} - \begin{vmatrix} s_p & s_q \\ t_p & t_q \end{vmatrix}^{-1} \cdot \begin{vmatrix} s \\ t \end{vmatrix}. \tag{III. 2. 5}$$

Berechnet werden die Werte der Ableitungen dabei durch zweimalige Division von $P(z)$ durch $d(z)$. Als Newtonsches Verfahren ist das Verfahren selbstverständlich quadratisch konvergent und selbstkorrigierend.

IV. EIN NEUES FAKTORISIERUNGSVERFAHREN

Nach dem in Abschnitt IV über die funktionale Iteration Gesagten stellt eine funktionale Iteration zur vollständigen Faktorisierung von $P(z)$ in zwei Faktoren $U(z)$ und $V(z)$ der Grade k und $m = n - k$ eine Abbildung des n-dimensionalen Koeffizientenraumes dar in folgender Form

$$u'_\mu = f_\mu(u_r, v_s; p_r) \qquad \mu = 1, 2, \ldots, k$$
$$v'_\nu = g_\nu(u_r, v_s; p_r) \qquad \nu = k+1, \; k+2, \ldots, \; n,$$

die die Lösungen, d. h. die Koeffizienten der tatsächlichen Faktorpolynome, zu Fixpunkten hat. Der Konvergenzsatz von Abschnitt IV besagt, daß das Konvergenzverhalten in der Umgebung der Lösung durch das Verhalten der Funktionalmatrix bestimmt wird. Ein quadratisch konvergentes Faktorisierungsverfahren z. B. ist also durch eine Abbildung der obengenannten Art gegeben, deren Funktionalmatrix in dem oder den Lösungspunkten verschwindet.

IV. 1. Die Nullpunktverschiebung

Bei der Behandlung der Nullstellenbestimmung war gesagt worden, daß funktionale Iterationen aus definiten Verfahren durch Abbrechen und Nullpunktverschiebung gewonnen werden können. Es liegt nahe, die gleiche Methode für die Faktorisierung zu verwenden. Bei der Übertragung des Prinzips macht sich jedoch die Tatsache bemerkbar, daß die Aufgabe hier in bestimmter Hinsicht grundsätzlich von der Nullstellenbestimmung verschieden ist. Bei der Nullstellenbestimmung wird eine Folge von Punkten in der Ebene der analytischen Veränderlichen z durch Auswertung der analytischen Funktion $P(z)$ und ihrer Ableitungen bestimmt. Bei der Faktorisierung dagegen ist eine Folge von Punkten in dem $m + k = n$-dimensionalen direkten Produktraum der Koeffizientenräume der Polynome m-ten und k-ten Grades in der algebraischen Unbestimmten z zu bestimmen. Von den Werten der analytischen Funktion $P(z)$ ist dabei keine Rede mehr.

Daher ist zunächst einmal die Frage zu klären, wie die „Nullpunktverschiebung an einem definiten Faktorisierungsverfahren"zweckmäßig zu definieren ist. Dies erfordert eine Analyse des einzigen in Frage kommenden Verfahrens, nämlich der Faktorisierung von BAUER. Für den hier vorliegenden Zweck ist es bequem, an Stelle der in Abschnitt II beschriebenen die „inverse" Iteration zu benützen, die durch die Gleichung

$$P(z) - \sum_{j=1}^{m} v_{i,\,m-j}\, z^{m-j}\, U_{i-j}(z) = z^m U_i(z) \tag{IV. 1. 1}$$

dargestellt wird und sich hinsichtlich ihrer Konvergenzeigenschaften von der früher angegebenen Vorschrift nur dadurch unterscheidet, daß die Polynome V_i hier gegen den Faktor mit den kleinsten Nullstellen streben.

Das für die augenblickliche Fragestellung wesentliche Merkmal dieser Iteration ist, daß zur Bildung der Linearkombination, die $P(z)$ ergibt, die Produkte U_{i-j} mit den $m + 1$ festen Polynomen z^{m-j} herangezogen werden. Von diesen repräsentiert insbesondere das erste, also z^m, den Nullpunkt des Raumes Π_m der normierten Polynome $V(z)$ vom genauen Grade m, während die übrigen Polynome z^{m-j} die Nullpunkte der Räume Π_{m-j}

darstellen. Es liegt nun nahe, als „Nullpunktverschiebung hinsichtlich der Faktorisierung" eine Abbildung eines Teiles oder aller dieser Polynome auf die laufend anfallenden Näherungspolynome $V_i(z)$ zu verlangen.

Eine Vorschrift über die Zuordnung der festen Polynome z^{m-j} zu den Näherungspolynomen soll nun aus den einzelnen Schritten des Faktorisierungsalgorithmus abgeleitet werden. Dabei soll hier als Anfangszustand der Faktorisierung nur ein solcher Zustand betrachtet werden, in dem alle bereits vorliegenden Näherungspolynome U gleich dem einen festen Ausgangspolynom $U_0(z)$ sind. Die ersten Faktorisierungsschritte lauten dann

$$\begin{aligned} &P(z) + (z^m - V_1(z))\, U_0(z) = z^m U_1(z) \\ &P(z) + (z^m - v_{2,1}\, z^{m-1} - V_2(z))\, U_0(z) - v_{2,1}\, z^{m-1} U_1(z) = z^m U_2(z) \\ &P(z) + (z^m + v_{3,1}\, z^{m-1} + v_{3,2}\, z^{m-2} - V_3(z))\, U_0(z) \\ &\quad - v_{3,2}\, z^{m-2} U_1(z) - v_{3,1}\, z^{m-1} U_2(z) = z^m U_3(z) \qquad \text{usw.} \end{aligned} \tag{IV. 1. 2}$$

An diesen Gleichungen ist nun eine Nullpunktverschiebung im oben umrissenen Sinne vorzunehmen.

Völlig unproblematisch ist dabei die Nullpunktverschiebung im eigentlichen Sinne: der „Nullpunkt" z^m ist durch das letztvorangegangene Näherungspolynom $V_0(z)$ zu ersetzen. Nicht so eindeutig ist dagegen die Vorschrift für die Abbildung der weiteren Polynome $z^{m-j}\,(j = 1, 2, \ldots, m)$. Bei strengster Übertragung des Begriffes der Nullpunktverschiebung wäre zu verlangen, daß jedes dieser Polynome auf ein Polynom aus dem gleichen Raum, d. h. ein Polynom mit dem gleichen Grade $m - j$, abgebildet wird. Dies führt jedoch zwangsläufig dazu, daß die Näherungen der beiden Faktoren U_i und V_i in die Iterationsgleichungen in verschiedener Weise eingehen. Der Zweck der Nullpunktverschiebung ist aber, eine einmal erreichte Näherung vollständig zur Fortführung der Iteration auszunützen. Damit wird die Forderung nach Symmetrie der Gleichungen hinsichtlich der Näherungen U_i und V_i, die ja völlig gleichberechtigt sind, zu einer Selbstverständlichkeit.

Die obengenannte Forderung nach Erhaltung der Grade soll daher ersetzt werden durch die schwächere Forderung, daß das für z^{m-j} eintretende Polynom den Höchstgrad $m - 1$ haben soll, sowie durch die zusätzliche Symmetrieforderung hinsichtlich der U_i und der V_i.

Beide Forderungen sind erfüllt, wenn man in der s-ten der Gleichungen den Ausdruck $\sum_{r=0}^{j} v_{s,r}\, z^{m-r}$ durch V_j ersetzt. Dies entspricht der Zuordnung

$$\begin{aligned} z^m &\longrightarrow V_0 \\ z^{m-j} &\longrightarrow \frac{1}{v_{s,j}}\,(V_j - V_{j-1}) \qquad (j = 1, 2, \ldots, s-1). \end{aligned}$$

IV. 2. Die Iterationsvorschrift

Damit geht die Formelgruppe (IV. 1. 2) über in

$$\begin{aligned} &P(z) + (V_0 - V_1)\, U_0 = V_0 U_1 \\ &P(z) + (V_1 - V_2)\, U_0 - (V_1 - V_0)\, U_1 = V_0 U_2 \\ &P(z) + (V_{i-1} - V_i)\, U_0 - (V_{i-1} - V_{i-2}) U_1 - \ldots = V_0 U_i \qquad (i \leqslant m) \\ &\qquad\qquad = V_{i-m} U_i \qquad (i > m) \end{aligned} \tag{IV. 2. 1}$$

Jeweils die ersten s-Gleichungen dieser Gruppe definieren nun ein unabhängiges Iterationsverfahren, wobei man sich im übrigen mittels der ersten $s - 1$-Gleichungen die Größen U_j

und $V_j (j = 1, \ldots, s-1)$ aus der s-ten Gleichung eliminiert zu denken hat, so daß eine Gleichung zwischen U_0, V_0 einerseits und U_s, V_s andererseits entsteht. Mit ihrer Hilfe läßt sich nun die Polynomfolge U_{ks}, V_{ks} $(k = 1, 2, \ldots)$ iterativ berechnen.

Daß die exakte Zerlegung $U(z), V(z)$ stationäre Lösung des Gleichungssystems ist, ist unmittelbar abzulesen, da alle Differenzen verschwinden. Daß bei Teilerfremdheit von U und V aus $U_0 = U, V_0 = V$ auch $U_s = U, V_s = V$ folgt, also die Fixpunkteigenschaft der durch die Gleichungen definierten Abbildung, ergibt sich folgendermaßen:

Die erste Gleichung geht über in $V(U_1 - U) + U(V_1 - V) = 0$. Wegen der vorausgesetzten Teilerfremdheit von U und V führt die Annahme, das Polynom $k-1$-ten Grades $U_1 - U$ sei nicht das Nullpolynom, auf einen Widerspruch. Also ist $U_1 = U, V_1 = V$. Damit geht die zweite und infolgedessen auch alle weiteren Gleichungen, abgesehen von der Ersetzung des Index 1 durch 2, 3, ... s, in die erste Gleichung über, und der Beweisgang wiederholt sich. Von dem Verhalten der Iteration bei nicht teilerfremden U, V wird noch später die Rede sein.

Was nun die Konvergenzeigenschaften anbetrifft, so ließen die Eigenschaften der Iterationen zur Nullstellenbestimmung eine Konvergenz der aus s Gleichungen aufgebauten Iteration von der Ordnung $s + 1$ erwarten. Die durch die erste Gleichung allein definierte Iteration konvergiert tatsächlich, wie sich zeigen wird, quadratisch. Was die übrigen anbetrifft, so läßt sich auf Grund der dabei auftretenden Eliminationsschwierigkeiten nicht mehr sagen, als daß sie mindestens quadratisch konvergieren. Da es darüber hinaus ebenso wie bei der Nullstellenbestimmung mindestens fraglich erscheint, ob die größere Kompliziertheit der Gleichungen durch die höhere Konvergenzzahl aufgewogen wird, soll die weitere Diskussion auf den einfachsten Fall der durch die erste Gleichung

$$P(z) + (V_0(z) - V_1(z))\, U_0(z) = V_0(z)\, U_1(z) \tag{IV. 2. 2}$$

definierten Iteration beschränkt bleiben.

IV. 3. Diskussion des Gleichungssystems der Iteration

Die Gleichung (IV. 2. 2) repräsentiert ein lineares Gleichungssystem für die Koeffizienten $u_{1,j}$ von U_1 und $v_{1,k}$ von V_1, das man am besten der für die in IV. 5 zu behandelnde numerische Durchrechnung bequemsten Umformung

$$V_0(U_1 - z^k) + U_0(V_0 - V_1) = P - z^k V_0$$

entnimmt. Zur Vermeidung allzu vieler Indizierungen sollen nun U und V für U_0 und V_0, U^+ und V^+ für U_1 und V_1 sowie ΔV für $V^+ - V$ gesetzt werden. Mit diesen Bezeichnungen erhält man das Gleichungssystem

$$\underbrace{\begin{bmatrix} 1 & & & & 1 & & \\ v_1 & 1 & & & u_1 & \cdot & \\ \cdot & v_1 & \cdot & & \cdot & \cdot & 1 \\ \cdot & \cdot & 1 & & & u_1 & 1 \\ v_m & & v_1 & 1 & & \cdot & u_1 \\ & v_m & & v_1 & u_k & & \cdot \\ & & \cdot & \cdot & & \cdot & \\ & & & v_m & \cdot & & u_k \\ & & & & v_m & & u_k \end{bmatrix}}_{\substack{k\text{-Spalten} \\ m\text{-Spalten}}} \cdot \begin{bmatrix} u_1^+ \\ u_2^+ \\ \cdot \\ \\ u_k^+ \\ \Delta v_1 \\ \cdot \\ \\ \Delta v_m \end{bmatrix} = \begin{bmatrix} p_1 - v_1 \\ p_2 - v_2 \\ \cdot \\ p_m - v_m \\ p_{m+1} \\ \\ \\ \\ p_n \end{bmatrix}. \tag{IV. 3. 1}$$

Die Matrix des Systems ist die bekannte Matrix $\mathfrak{R}(U, V)$ der Resultante $R(U, V)$ der Polynome U und V. Sie ist invertierbar, wenn die Resultante von Null verschieden ist, also bei Teilerfremdheit von U und V. Diese Bedingung ist aber praktisch ohne Bedeutung, wenn nur die exakten Faktoren $\overline{U}$ und $\overline{V}$ teilerfremd sind; denn das zufällige Auftreten eines exakten gemeinsamen Teilers der beiden Iterationspolynome U und V hat für die numerische Rechnung schon wegen der beschränkten Stellenzahl praktisch verschwindende Wahrscheinlichkeit, und ein approximativer Teiler führt zwar zu einer schlechten Kondition des Gleichungssystems und damit zu einer Konvergenzverzögerung, aber nicht zu einem zwangsläufigen Abbrechen der Iteration. Man kann eine solche Situation etwa vergleichen dem Verhalten des Newtonschen Verfahrens in der Umgebung einer Nullstelle der Ableitung P'.

Anders ist die Lage, wenn $\overline{U}$ und $\overline{V}$ selbst einen gemeinsamen Teiler haben. Dies übersieht man am schnellsten, wenn man das Gleichungssystem im Lösungspunkt selbst betrachtet, also $U = \overline{U}$ und $V = \overline{V}$ setzt. Dann ist $U^+ = U$, $\Delta V = 0$ eine Lösung. Für $R(U, V) \neq 0$ ergibt dies trivialerweise die Fixpunkteigenschaften. Haben aber U und V den gemeinsamen Teiler S vom Grade s, also $U = SX$, $V = SY$, so ergeben sich zusätzliche Lösungen der Form

$$U = H(z)\,X(z) \qquad V = -H(z)\,Y(z)$$

mit willkürlichem $H(z)$ vom Grade $s-1$. Dies deutet darauf hin, daß in der Umgebung der gewünschten Faktorisierung Fremdlösungen eingeschleppt werden, die den gemeinsamen Teiler in unkontrollierbarer Weise verfälschen, während die Konvergenz der teilerfremden Anteile nicht gestört wird.

IV. 4. Auflösung des Gleichungssystems

Schließt man diesen Fall aus, so läßt sich die Inverse der Resultantenmatrix $\mathfrak{R}(U, V)$ explizit angeben. Zur Konstruktion der Inversen geht man am einfachsten auf die Eigenschaften der bereits früher erwähnten Frobeniusmatrix $\mathfrak{F}$ des Polynoms $Q(z) = U(z) \cdot V(z)$ zurück. $\mathfrak{F}$ hat einerseits die Polynome $Q_\mu(z) = Q(z)/(z - \zeta_\mu)$ zu Eigenvektoren (lineare Elementarteiler, d. h. einfache Nullstellen von $Q(z)$ vorausgesetzt), es ist also

$$\mathfrak{F} \cdot (Q_\mu) = (Q_\mu) \cdot (\Lambda), \tag{IV. 4. 1}$$

wo Λ die Diagonalmatrix der Eigenwerte von $\mathfrak{F}$, d. h. der Nullstellen ζ_μ von $Q(z)$ darstellt.

Andererseits bringt die Resultantenmatrix $\mathfrak{R}$ die Frobeniusmatrix zum Zerfall in die beiden Frobeniusmatrizen $\mathfrak{F}_u$ und $\mathfrak{F}_v$ der Faktoren U und V:

$$\mathfrak{F} \cdot \mathfrak{R} = \mathfrak{R} \cdot \begin{pmatrix} \mathfrak{F}_u & 0 \\ 0 & \mathfrak{F}_v \end{pmatrix} = \mathfrak{R} \cdot \mathfrak{F}_{uv}.$$

$\mathfrak{F}_u$ und $\mathfrak{F}_v$ haben nun natürlich wieder die Polynome $U_\mu = U(z)/(z - \zeta_\mu)$ und $V_\varrho = V(z)/(z - \zeta_\varrho)$ zu Eigenvektoren, und die Gesamtheit ihrer Eigenwerte stimmt mit denen von $\mathfrak{F}$ überein. Mit der Eigenvektormatrix $[U,V]_\mu = \begin{pmatrix} U_\mu & 0 \\ 0 & V_\varrho \end{pmatrix}$ von $\mathfrak{F}_{uv}$ ergibt sich daher

$$\mathfrak{F}\mathfrak{R} \cdot [U, V]_\mu = \mathfrak{R} \cdot \mathfrak{F}_{uv}[U, V]_\mu = \mathfrak{R}[U, V]_\mu \cdot \Lambda. \tag{IV. 4. 2}$$

Aus den beiden Gleichungen (IV.4.1) und (IV.4.2) folgt zunächst nur, daß $\mathfrak{R} \cdot [U,V]_\mu \times$ $\times (Q_\mu)^{-1}$ eine Diagonalmatrix ist. Tatsächlich sieht man aber sofort, daß $\mathfrak{R} \cdot [U, V]_\mu$ gleich (Q_μ) ist, da die Spalten die Polynome $V \cdot U_\mu$ und $U \cdot V_\varrho$ ergeben, was bei richtiger Anordnung mit Q_μ identisch ist. Daher ist $\mathfrak{R}^{-1} = [U, V]_\mu \cdot (Q_\mu)^{-1} \cdot (Q_\mu)^{-1}$ ist schließlich die Linkseigenvektormatrix von $\mathfrak{F}: L_{\mu\nu} = Q_\mu(\zeta_\mu)^{-1} \cdot (\zeta_\mu^{n-\nu})$. Damit ergibt sich endgültig

$$\text{(IV. 4. 3)} \qquad \mathfrak{R}(U, V)^{-1} = [U, V]_\mu \cdot L = \begin{pmatrix} \sum_{\mu=1}^{k} \dfrac{U_{\mu,i} \cdot \zeta_\mu^{n-j}}{U_\mu(\zeta_\mu) \cdot V(\zeta_\mu)} \\ \sum_{\varrho=k+1}^{n} \dfrac{V_{\varrho i} \cdot \zeta_\varrho^{n-j}}{V_\varrho(\zeta_\varrho) \cdot U(\zeta_\varrho)} \end{pmatrix}.$$

Die Herleitung der Inversen bleibt im übrigen gültig, wenn U oder V mehrfache (aber nicht gemeinsame) Nullstellen besitzen, wobei nur an Stelle gewisser Eigenvektoren die entsprechenden Hauptvektoren treten.

Damit ist die Inverse der Resultantenmatrix als Funktion der Nullstellen von $U(z)$ und $V(z)$ dargestellt. Daß sie sich als symmetrische Funktion in den beiden Nullstellensätzen auch durch die Koeffizienten von U und V ausdrücken läßt, folgt trivialerweise aus der Determinantendarstellung der Inversen. Eine rekursive Vorschrift zu ihrer Berechnung läßt sich aber anscheinend ebensowenig wie für die Resultante selbst geben, was wohl im wesentlichen darauf zurückzuführen ist, daß die Ausdrücke in den Variablen u_i, v_k homogen sind. Für den hier verfolgten Zweck des Konvergenzbeweises der Iteration genügt jedoch die Tatsache, daß die Inverse sich als rationale Funktion der Koeffizienten darstellen läßt, zusammen mit der expliziten Darstellung in den Nullstellen.

Mit Hilfe dieser Darstellung, wobei nun die Nullstellen von $U(z)$ mit $\alpha_\varrho (\varrho = 1, 2 \ldots, k)$ und die Nullstellen von $V(z)$ mit $\beta_\sigma (\sigma = 1, 2 \ldots m)$ bezeichnet seien, ergibt sich die Lösung des Gleichungssystems (IV. 3. 1) zu

$$\text{(IV. 4. 4)} \qquad \begin{aligned} u_i^+ &= \sum_{\varrho=1}^{k} u_{\varrho, i-1} \frac{P(\alpha_\varrho) - \alpha_\varrho^k V(\alpha_\varrho)}{U_\varrho(\alpha_\varrho) V(\alpha_\varrho)} \\ \Delta v_j &= \sum_{\sigma=1}^{m} v_{\sigma, j-1} \frac{P(\beta_\sigma)}{V_\sigma(\beta_\sigma) U(\beta_\sigma)}. \end{aligned}$$

Nun ist $u_{\varrho, i-1} = \sum_{s=0}^{i-1} \alpha_\varrho^s u_{i-1-s}$ und infolgedessen

$$\sum_{\varrho=1}^{k} u_{\varrho, i-1} \cdot \frac{\alpha_\varrho^k}{U_\varrho(\alpha_\varrho)} = \sum_{s=0}^{i-1} u_{i-1-s} \cdot \sum_{\varrho=1}^{k} \frac{\alpha_\varrho^{k+s}}{U_\varrho(\alpha_\varrho)} = -u_i,$$

da $\sum_{\varrho=1}^{k} \frac{\alpha_\varrho^{k+s}}{U_\varrho(\alpha_\varrho)}$ die Alephfunktion h_{s+1} der α_ϱ ergibt, die wiederum der Gleichung $\sum_{i=0}^{i} h_s u_{i-s} = 0$ genügt.

Damit ergibt sich schließlich

$$\begin{aligned} u_i^+ &= u_i + \sum_{\varrho=1}^{k} u_{\varrho, i-1} \frac{P(\alpha_\varrho)}{U_\varrho(\alpha_\varrho) V(\alpha_\varrho)} \\ \Delta v_j &= \sum_{\sigma=1}^{m} v_{\sigma, j-1} \frac{P(\beta_\sigma)}{V_\sigma(\beta_\sigma) U(\beta_\sigma)}. \end{aligned}$$

IV. 5. Beweis der quadratischen Konvergenz

Wegen $v_j = v_j^+ - v_j$ sind die Koeffizienten von U^+ und V^+

$$u_i^+ = u_i + \sum_{\varrho=1}^{k} u_{\varrho, i-1} \cdot \frac{P(\alpha_\varrho)}{U_\varrho(\alpha_\varrho)\, V(\alpha_\varrho)}$$

$$v_j^+ = v_j + \sum_{\sigma=1}^{m} v_{\sigma, j-} \cdot \frac{P(\beta_\sigma)}{V_\sigma(\beta_\sigma)\, U(\beta_\sigma)},$$

und dies ist die der funktionalen Iteration zugrunde liegende Abbildung des Koeffizientenraumes, deren Verhalten im Lösungspunkte zu untersuchen ist.

Zufolge dem Konvergenzsatz von Abschnitt III ist die Konvergenz quadratisch, wenn die Matrix der Ableitungen der gestrichenen nach den ungestrichenen Größen

$$\frac{D(u_i^+, v_j^+)}{D(u_r, v_s)} = \begin{pmatrix} \partial u_i^+/\partial u_r & \partial u_i^+/\partial v_s \\ \partial v_j^+/\partial u_r & \partial v_j^+/\partial v_s \end{pmatrix}$$

bei Einsetzen der exakten Lösung verschwindet.

Nun stehen die u_i^+, v_j^+ nur als Funktionen der Wurzeln α_ϱ, β_σ zur Verfügung, wobei die u_i, v_j als ebenfalls durch die Wurzeln dargestellt zu betrachten sind. Solange aber alle Nullstellen einfach sind, lassen sich die Ableitungen nach den Koeffizienten u_r, v_s linear in den Ableitungen nach den Wurzeln ausdrücken. Daher genügt es für den Konvergenzbeweis, das Verschwinden aller Ableitungen der gestrichenen Größen nach den Wurzeln zu zeigen, wenn in die Ableitungen die Nullstellen von $P(z)$ eingesetzt werden.

Das Verschwinden der Ableitungen nach den Wurzeln aber ist sofort einzusehen. Es ist

$$\partial u_i^+/\partial \alpha_\varrho = -u_{\varrho, i-1} + u_{\varrho, i-1} \frac{P'(\alpha_\varrho)}{U_\varrho(\alpha_\varrho)\, V(\alpha_\varrho)} + \sum_{\tau=1}^{k} P(\alpha_\tau) \frac{\partial}{\partial \alpha_\varrho} \cdot \frac{u_{\tau, i-1}}{U_\tau(\alpha_\tau)\, V(\alpha_\tau)} \Bigg| \begin{matrix} \alpha_\varrho = \zeta_\varrho \\ \beta_\sigma = \zeta_\sigma \end{matrix}$$

(IV. 5. 2) $$= 0$$

$$\partial u_i^+/\partial \beta_\sigma = \sum_{\varrho=1}^{k} u_{\varrho, i-1} \frac{P(\alpha_\varrho)}{U_\varrho(\alpha_\varrho)} \cdot \frac{\partial}{\partial \beta_\sigma} \cdot \frac{1}{V(\alpha_\varrho)} \Bigg| \begin{matrix} \alpha_\varrho = \zeta_\varrho \\ \beta_\sigma = \zeta_\sigma \end{matrix}$$

wegen $P(\alpha_\varrho) = 0$ und $P'(\alpha_\varrho) / (U_\varrho(\alpha_\varrho) V(\alpha_\varrho)) = 1$ für $\alpha_\varrho = \zeta_\varrho$, und entsprechend verschwinden die Ableitungen der v_j'.

Damit ist die quadratische Konvergenz des Verfahrens nachgewiesen. Da das Verfahren auch als spezielles Newtonsches Verfahren für nichtlineare Systeme aufgefaßt werden kann, unterliegt es im übrigen den dafür geltenden allgemeinen Konvergenzsätzen. Wir glauben jedoch, daß der direkte Beweis der Einsicht in die Zusammenhänge besser dient. Im übrigen ist das Konvergenzverhalten wie stets bei funktionalen Iterationen abhängig von der Ausgangsnäherung, und eine Aussage darüber, gegen welche der möglichen Zerlegungen der vorgeschriebenen Grade die Iteration konvergiert, läßt sich a priori nicht machen.

IV. 6. Darstellung der Näherungspolynome

Aus den Gleichungen (IV. 4. 1) ergeben sich die neuen Näherungspolynome $U^+(z)$ und $V^+(z)$ zu

$$U^+(z) = U(z) + \sum_{\varrho=1}^{k} \frac{U_\varrho(z)}{U_\varrho(a_\varrho)} \cdot \frac{P(a_\varrho)}{V(a_\varrho)}$$

(IV. 6. 1)

$$V^+(z) = V(z) + \sum_{\sigma=1}^{m} \frac{V_\sigma(z)}{V_\sigma(\beta_\sigma)} \cdot \frac{P(\beta_\sigma)}{U(\beta_\sigma)}.$$

$U^+(z)$ wird also mit Hilfe der normierten Interpolationspolynome $\frac{U_\varrho(z)}{U_\varrho(a_\varrho)}$ so bestimmt, daß es an den Nullstellen der letzten Näherung gerade die Werte $P(a_\varrho)/V(a_\varrho)$ annimmt, und Entsprechendes gilt für $V(z)$.

Für die numerische Berechnung ist diese Darstellung der Näherungspolynome leider nicht verwendbar, da natürlich die Nullstellen der Näherungen nicht bekannt sind. Es muß daher tatsächlich das Gleichungssystem (IV. 3. 1) gelöst werden. Dieses läßt sich allerdings beträchtlich reduzieren, so daß der Grad des tatsächlich zu lösenden Gleichungssystems gleich dem Grad des kleineren der beiden Faktoren U und V wird. Über die Durchführung der Rechnung wird nun anschließend zu sprechen sein.

IV. 7. Zur numerischen Durchführung der Iteration

Die Matrix des zu lösenden Gleichungssystems

$$\underbrace{\left[\begin{array}{ccc:ccc}
1 & & & 1 & & \\
v_1 & 1 & & u_1 & \cdot & \\
\cdot & v_1 & \cdot & \cdot & \cdot & 1 \\
\cdot & \cdot & 1 & & u_1 & 1 \\
v_m & & v_1\ 1 & & \cdot & u_1 \\
\hdashline
 & v_m & v_1 & u_k & & \cdot \\
 & \cdot\ \cdot & & & \cdot & \\
 & v_m\ \cdot & & & u_k\ \cdot & \\
 & v_m & & & u_k &
\end{array}\right]}_{k\text{-Spalten}}\ \underbrace{\phantom{\left[\begin{array}{ccc:ccc} 1 & & & 1 & & \\ v_1 & 1 & & u_1 & \cdot & \\ \cdot & v_1 & \cdot & \cdot & \cdot & 1 \\ \cdot & \cdot & 1 & & u_1 & 1 \\ v_m & & v_1\ 1 & & \cdot & u_1 \\ \hdashline & v_m & v_1 & u_k & & \cdot \\ & \cdot\ \cdot & & & \cdot & \\ & v_m\ \cdot & & & u_k\ \cdot & \\ & v_m & & & u_k & \end{array}\right]}}_{m\text{-Spalten}}
\left[\begin{array}{c} u_1^+ \\ u_2^+ \\ \cdot \\ \\ \\ u_k^+ \\ \hdashline \Delta v_1 \\ \cdot \\ \\ \Delta v_m \end{array}\right]
=
\left[\begin{array}{c} p_1 - v_1 \\ p_2 - v_2 \\ \cdot \\ p_m - v_m \\ p_{m+1} \\ \hdashline \\ \\ \\ p_n \end{array}\right]$$

setzt sich in der durch Strichelung angedeuteten Weise aus Untermatrizen zusammen in der Form

(IV. 7. 1)

$$\begin{pmatrix} u^+ \\ \Delta v \end{pmatrix} = \begin{pmatrix} L & T \\ S & R \end{pmatrix} \begin{pmatrix} u^+ \\ \Delta v \end{pmatrix} = \begin{pmatrix} p \\ q \end{pmatrix}.$$

Darin ist L eine $k \cdot k$-reihige untere Dreiecksmatrix, R eine $m \cdot m$-reihige obere Dreiecksmatriz, T und S sind $k \cdot m$- bzw. $m \cdot k$-reihige Trapeze. Unter der oben über die Grade gemachten Annahme ist L der Bestandteil größerer, R derjenige kleinerer Dimension.

Daher ist es zweckmäßig, die Invertierung so zu führen, daß eine tatsächliche Gleichungsauflösung nur im Raume von R notwendig wird. Da sich auch die Inverse der unteren Dreiecksmatrix L unmittelbar angeben läßt, erscheint die Gaußsche Elimination an den Matrizen als das gegebene Verfahren.

Es wird also zunächst die Matrix

$$G = \begin{pmatrix} L^{-1} & 0 \\ -SL^{-1} & E_m \end{pmatrix} \quad (E_m \text{ sei die } m\text{-dimensionale Einheit})$$

von links auf das Gleichungssystem (IV. 7. 1) angewendet. Damit ergibt sich

$$\text{(IV. 7. 2)} \qquad \begin{pmatrix} E_k & L^{-1}T \\ 0 & R - SL^{-1}T \end{pmatrix} \begin{pmatrix} u^+ \\ \Delta v \end{pmatrix} = \begin{pmatrix} L^{-1}p \\ q - SL^{-1}p \end{pmatrix}.$$

Anschließend muß das untere, $m \cdot m$-reihige Gleichungssystem für Δv numerisch aufgelöst werden. Mit $\bar{R} = R - SL^{-1}T$ ergibt dies

$$\text{(IV. 7. 3)} \qquad \begin{aligned} \Delta v &= \bar{R}^{-1}(q - SL^{-1}p) \\ u^+ &= L^{-1}(p - T\Delta v). \end{aligned}$$

Die Matrizen L^{-1} und SL^{-1} stellen sich in den bereits erwähnten Alephfunktionen h_i der Nullstellen von V und den zugeordneten $h_{i,i+j}$ dar in der Form

$$\text{(IV. 7. 4)} \qquad \begin{aligned} L^{-1} &= h_{j-i} && (i,j = 1, 2, \ldots k) \\ SL^{-1} &= h_{k+1-j,\,k+i-j} && \begin{matrix}(i = 1, 2, \ldots m) \\ (j = 1, 2, \ldots k).\end{matrix} \end{aligned}$$

Mit Hilfe dieser Größen lassen sich nun alle übrigen Elemente der Gleichung (IV. 7. 3) berechnen und in den rekursiv definierten Größen

$$\text{(IV. 7. 5)} \qquad M_i = \sum_{r=0}^{i} u_r h_{i-r} = u_i - \sum_{r=1}^{i} v_r M_{i-r} \qquad (i = 1, 2, \ldots k)$$

$$\bar{M}_{i,j} = u_{k+i-j} + \sum_{r=j}^{k} u_{r-j} h_{k+1-r,\,k+i-r} = \sum_{r=0}^{i-1} v_r M_{k+i-j-r}$$

$$= u_{k+i-j} - \sum_{r=i}^{k+i-j} v_r M_{k+i-j-r} \qquad (i,j = 1, 2, \ldots m)$$

$$K_i = \sum_{r=0}^{i} p_r h_{i-r} = p_i - \sum_{r=1}^{i} v_r K_{i-r} \qquad (i = 1, 2, \ldots k)$$

$$\bar{K}_i = p_{k+i} + \sum_{r=0}^{k} p_r h_{k+1-r,\,k+i-r} = \sum_{r=0}^{i-1} v_r K_{k+i-r} \qquad (i = 1, 2, \ldots m)$$

$$= p_{k+i} - \sum_{r=i}^{m} v_r K_{k+i-r}$$

ausdrücken, wobei selbstverständlich $v_r = 0$ für $r > m$ und $u_r = 0$ für $r > k$ zu setzen ist.

Dabei ergibt sich

$$
\begin{aligned}
(L^{-1}T)_{i,j} &= M_{i-j} && (i = 1, 2, \ldots k)\ (j = 1, 2, \ldots m)\\
(R - SL^{-1}T)_{i,j} &= \bar{M}_{i,j} && (i, j = 1, 2, \ldots m)\\
(L^{-1}p)_i &= K_i && (i = 1, 2, \ldots k)\\
(q - SL^{-1}p)_i &= \bar{K}_i && (i = 1, 2, \ldots m)
\end{aligned}
\tag{IV. 7. 6}
$$

Die endgültigen Lösungen sind also

$$
\begin{aligned}
u_1^+ &= K_i - M_{i-j}(v_j^+ - v_j)\\
v_i^+ &= v_i + \bar{M}_{i,j}^{-1} \cdot \bar{K}_j.
\end{aligned}
\tag{IV. 7. 7}
$$

Die Rekursionen für die K_i, $\bar{K}_i$ und M_i, $\bar{M}_{i,j}$ sind nun nichts anderes als Divisionsalgorithmen in der Form eines verallgemeinerten Hornerschemas. Die tatsächliche Rechnung ist also verwandt mit der des BAIRSTOW-COLLATZ-Verfahrens zur Bestimmung eines quadratischen Faktors, jedoch werden hier beide Faktoren iterativ bestimmt und gehen in die Berechnung der jeweiligen Korrekturen ein. Dadurch wird einmal die Konvergenzprüfung, die gewöhnlich durch Vergleich zweier aufeinanderfolgender Näherungen vorgenommen wird, zu einer automatischen vollständigen Verprobung beider Faktoren. Außerdem gibt es die Möglichkeit, ähnlich wie bei der Treppeniteration mehrere Faktoren beliebiger Grade kaskadenartig nebeneinander zu berechnen nach dem Schema

$$
\begin{aligned}
U_i^{(s-1)} + (V_i^{(s)} - V_{i+1}^{(s)})\, U_i^{(s)} &= V_i^{(s)}\, U_{i+1}^{(s)} && (s = 1, 2, \ldots),\\
U^{(0)} &= P(z).
\end{aligned}
\tag{IV. 7. 8}
$$

V. DER SPEZIALFALL DER BESTIMMUNG VON LINEARFAKTOREN

Für den Spezialfall $m = 1$, d. h. Abspaltung reeller Linearfaktoren V, ist das Verfahren einschließlich der Vorschrift (IV. 6. 8) zur simultanen Berechnung mehrerer Nullstellen vor kurzem von BAUER und dem Verfasser (8), aus anderen Gesichtspunkten als hier hergeleitet, vorgeschlagen worden. Dabei ist zu erwarten (und hat sich auch bei numerischen Versuchen schon bestätigt), daß die Konvergenz der Iteration für ein bestimmtes s nicht notwendig an die Konvergenz aller darüberliegenden Iterationen mit $s' < s$ geknüpft ist, wie dies auch bei der Treppeniteration der Fall ist, da ein Paar komplexer Linearfaktoren $V^{(s-1)}$, $V^{(s)}$ nur zu Oszillationen der entsprechenden Faktoren $U_i^{(s-1)}$, $U_i^{(s)}$ im Raume von Realteil und Imaginärteil von $U^{(s-1)}$ führt, die aber beide den Faktor $U^{(s)}$ enthalten.

V. 1. Ein Konvergenzsatz

Für den Fall eines linearen Faktors $V(z)$ und nur einfacher Nullstellen von $P(z)$ läßt sich noch eine über den Konvergenzbeweis aus Abschnitt V hinausgehende Konvergenzaussage machen, die nicht unerwähnt bleiben soll, obwohl sie zunächst nur von theoretischem Interesse ist. Für den Fall eines solchen Polynoms konvergiert ja bekanntlich das Newtonsche Verfahren vom positiv Unendlichen her monoton gegen die größte Nullstelle. Dies legt die Frage nahe, ob nicht etwas Ähnliches für die hier beschriebene Iteration gilt, wobei noch der Einfachheit halber nur positive Nullstellen vorausgesetzt seien.

Die Iteration lautet jetzt

$$P(z) - \frac{P(\beta_i)}{U_i(\beta_i)} \cdot U_i(z) = (z - \beta_i) U_{i+1}(z)$$

(V. 1. 1)

$$\beta_{i+1} = \beta_i - \frac{P(\beta_i)}{U_i(\beta_i)}.$$

$U_i(z)$ läßt sich jetzt durch die Polynome $P_\mu(z)$, denen die μ-te Nullstelle von $P(z)$ fehlt, linear ausdrücken. Es sei also (für $i = 0$) $U_0(z) = \sum_{\mu=1}^{n} e_\mu P_\mu(z)$ eine Ausgangsnäherung für $U(z)$ und $\beta_0 > \zeta_1 > \zeta_2 > \cdots > \zeta_n$ eine Ausgangsnäherung für ζ_1. Dann sind alle $P_\mu(\beta_0)$ positiv und größenmäßig wie die Nullstellen selbst geordnet.

Sind nun alle e_μ positiv, so folgt wegen $\sum e_\mu = 1$ (alle $U_i(z)$ sind wie die $P_\mu(z)$ auf den führenden Koeffizienten Eins normiert) aus $U_0(\beta_0) = \sum_{\mu=1}^{n} e_\mu P_\mu(\beta_0)$ sofort

$$P_1(\beta_0) > U_0(\beta_0) > P_n(\beta_0) > 0,$$

und damit

$$\beta_1 = \beta_0 - \frac{P(\beta_0)}{U_0(\beta_0)} = \beta_0 - (\beta_0 - \zeta_1) \frac{P_1(\beta_0)}{U_0(\beta_0)} < \zeta_1.$$

Wird also $U_0(z)$ so gewählt, daß alle Gewichte positiv sind, insbesondere für $U_0(z) = \frac{1}{n} P'(z)$, so ist eine monotone Approximation von ζ_1 nicht möglich.

Ist dagegen $U_0(\beta_0) > P_1(\beta_0)$, so folgt sofort $\beta_1 > \zeta_1$, und die Permanenz dieser Bedingung liefert monotone Approximation. Die Frage ist nun, ob sich spezielle Ausgangsnäherungen U_0 angeben lassen, für die aus $U_i(\beta_i) > P_1(\beta_i)$ dasselbe für $i + 1$ an Stelle von i folgt.

Aus Gleichung (V. 1. 1) folgt durch Einsetzen von ζ_μ

$$U_1(\zeta_\mu) = \frac{P_\mu(\beta_0)}{U_0(\beta_0)} U_0(\zeta_\mu).$$

Aus $U_0(\beta_0) > P_1(\beta_0)$ folgt also, daß für alle μ $|U_1(\zeta_\mu)| < |U_0(\zeta_\mu)|$ und daß die Vorzeichen von $U_0(\zeta_\mu)$ und $U_1(\zeta_\mu)$ übereinstimmen. Wird U_0 so gewählt, daß es lauter reelle Nullstellen $\alpha_\mu (\mu = 2, \ldots n)$ hat, so folgt durch Einsetzen in (V. 1. 1)

$$U_1(\alpha_\mu) = \frac{\alpha_\mu - \zeta_\mu}{\alpha_\mu - \beta_0} \cdot P_\mu(\alpha_\mu),$$

für $\alpha_\mu < \zeta_\mu$ stimmen daher die Vorzeichen von $U_1(\alpha_\mu)$ und $P_\mu(\alpha_\mu)$ überein.

Ist nun $U_0(z)$ speziell so gewählt, daß seine Nullstellen mit den Nullstellen von $P_1(z)$ alternieren in der Folge

$$\zeta_2 > \alpha_2 > \zeta_3 > \alpha_3 > \ldots > \zeta_n > \alpha_n,$$

so ist in allen Nullstellen ζ_μ von $P(z)$ mit Ausnahme von ζ_1 das Vorzeichen von $U_0(\zeta_\mu)$ dem von $P_\mu(\zeta_\mu)$ entgegengesetzt. Daher tritt zwischen je zwei gleichnumerierten Nullstellen von P_1 und U_0 ein Vorzeichenwechsel von $U_1(z)$ auf, d. h. $U_1(z)$ besitzt in diesen Intervallen je genau eine Nullstelle. Die Nullstellen von U_1 liegen also den zugehörigen Nullstellen von $P(z)$ näher als die entsprechenden Nullstellen von $U_0(z)$ und erfüllen alle an die letzteren gestellten Bedingungen.

Damit ist gezeigt, daß die Folge der β_i sich ζ_1 monoton nähert und daß ebenso die Folgen der $\alpha_{\mu,i}$ die Nullstellen ζ_μ monoton approximieren. Daß die Folgen überhaupt konvergieren, folgt aus der Monotonität und der Beschränktheit. Zur Bestimmung der Grenzwerte genügt die Betrachtung der Folge der β_i.

Aus der Tatsache der Konvergenz folgt aber insbesondere, daß der Betrag von $\beta_{i+1} - \beta_i$ und damit wegen der Beschränktheit aller $U_i(\beta_i)$ auch $P(\beta_i)$ gegen Null strebt. Die Folge der β konvergiert demnach gegen eine Nullstelle von $P(z)$, also notwendig gegen die untere Schranke ζ_1. Damit ergibt sich wegen der Eindeutigkeit der Faktorzerlegung aus der Gleichung (V. 1. 1) sofort, daß $U_i(z)$ gegen $P_1(z)$ konvergiert.

V. 2. Zweckmäßige Ausgangsnäherungen

Die angegebene Wahl von $U_0(z)$ garantiert also die monotone Konvergenz der Iteration. Jedoch ist diese Tatsache, wie bereits oben bemerkt, zunächst nur, als Beispiel für einen Konvergenzsatz im großen, von theoretischem Interesse, da sie eine Kenntnis der Verteilung der Nullstellen voraussetzt, die zu Beginn einer Rechnung nicht vorhanden ist.

Immerhin ist die Überlegung insofern nicht ganz ohne Wert, als sie gewisse Fingerzeige hinsichtlich einer vernünftigen Wahl der Ausgangsnäherung gibt. In dem behandelten Spezialfall bilden nämlich die Gewichte $e_{1,i}$, mit denen P_1 in U_i eingeht, eine von einem Anfangswert größer Eins monoton gegen Eins fallende Folge. Dies deutet darauf hin, daß U_0 grundsätzlich so gewählt werden sollte, daß das entsprechende Gewicht $e_{1,0}$ größer Eins ist, eine Vorschrift, die leicht zu realisieren ist, indem man eine unter Abschätzung für alle Nullstellen von $P(z)$ als einzige, $n-1$-fache Nullstelle von U_0 vorgibt. Bei allen bisher gerechneten numerischen Beispielen hatte diese Vorschrift tatsächlich monotone Konvergenz zur Folge.

LITERATURVERZEICHNIS

1. A. C. Aitken: Further numerical Studies in Algebraic Equations and Matrices XII. Proc. Roy. Soc. Edinburgh, Sec. A 51, 80–90 (1931).
2. L. Bairstow: Investigation relating to the Stability of the Aeroplane. Rep. Memor. Ado. Comm. Aero, London 154, 51–63 (1914).
3. F. L. Bauer: Quadratisch konvergente Durchführung der Bernoulli-Jacobischen Methode zur Nullstellenbestimmung von Polynomen. Sitz.-Ber. Bayer. Akad. Wiss. 1954, 275–303 (1954).
4. F. L. Bauer: Das Verfahren der abgekürzten Iteration für algebraische Eigenwertprobleme, insbesondere zur Nullstellenbestimmung eines Polynoms. Z. angew. Math. Phys. 7, 17–32 (1956).
5. F. L. Bauer: Ein direktes Iterationsverfahren zur Hurwitz-Zerlegung eines Polynoms. Arch. Elektr. Übertragung 9, 285–290 (1955).
6. F. L. Bauer: Direkte Faktorisierung von Polynomen. Sitz.-Ber. Bayer. Akad. Wiss. 1956, 163–309 (1956).
7. F. L. Bauer: Das Verfahren der Treppeniteration und verwandte Verfahren zur Lösung algebraischer Eigenwertprobleme. Z. Angew. Math. Phys. 8, 214–235 (1957).
8. F. L. Bauer und Klaus Samelson: Polynomkerne und Iterationsverfahren. Math. 67, 93–98 (1957).
9. L. Collatz: Das Hornersche Schema bei komplexen Wurzeln algebraischer Gleichungen. Z. angew. Math. Mech. 20, 235–236 (1940).
10. E. Frank: On the zeros of polynomials with complex coefficients. Bull. Amer. Math. Soc. 52, 144–157 (1946).
11. E. Frank: The location of the zeros of polynomials with complex coefficients. Bull. Amer. Math. Soc. 52, 890–898 (1946).
12. B. Friedmann: Note on Approximating Complex zeros of a Polynomial. Comms. Pure Appl. Math. 2, 195–208 (1949).
13. E. Fürstenau: Darstellung der Wurzeln algebraischer Gleichungen durch Determinanten der Koeffizienten. Marburg 1860.
14. J. König: Über eine Eigenschaft der Potenzreihen. Math. Ann. 23, 447–449 (1884).
15. J. L. Lagrange: Résolution des équations numériques. Note 6.
16. Shi-Nye Lin: A Method of Successive Approximation of Evaluating the Real and Complex Roots of Cubic and Higher Order Equations. J. Math. Phys. 20, 231–242 (1941).
17. Y. L. Luke und D. Ufford: On the roots of algebraic equations. J. Math. Phys. 30, 94–101 (1951).
18. H. J. Maehly: Zur iterativen Auflösung algebraischer Gleichungen. Z. angew. Math. Phys. 5, 260 (1954).
19. L. Occhini: Beitrag zu Walls Verfahren der getrennten Berechnung von Real- und Imaginärteilen der Nullstellen eines Polynoms. Z. angew. Math. Phys. 36, 139–145 (1956).
20. F. W. J. Olver: The Evaluation of Zeros of High-Degree Polynomials. Phil. Trans. Roy. Soc. A 244, 385–415 (1952).
21. H. Rutishauser: Der Quotienten-Differenzen-Algorithmus. Mitt. Inst. angew. Math. ETH Zürich Nr. 7 (Birkhäuser, Basel-Stuttgart 1957).
22. H. Rutishauser: Report on the Solution of Eigenvalue-Problems with the LR-Transformation. Appl. Math. Series vol. 49 (NBS, Washington).

23. H. Rutishauser und F. L. Bauer: Détermination des vecteurs propres d'une matrice par une méthode itérative avec convergence quadratique. (CR. Acad. Sci. Paris 240, 34 (1955).

24. E. Schröder: Über unendlich viele Algorithmen zur Auflösung der Gleichungen. Math. Ann. 2, 317 bis 365 (1870).

25. Theremin: Récherches sur la résolution des équations de tous les degrés. Crellé's J. 49, 187–243.

26. H. S. Wall: Polynomials whose zeros have negative real parts. Bull. Amer. Math. Soc. 52, 308–322 (1945).

27. E. T. Whittaker and G. Robinson: The calculus of observations. Blackie & Son Ltd., Glasgow 1940.

28. R. Zurmühl: Zum Graeffe-Verfahren und Horner-Schema bei komplexen Wurzeln. Z. angew. Math. Mech. 30, 283–285 (1940).