ZAMM · Z. angew. Math. Mech. 69 (1989) 11, 375 – 391

CARSTENSEN, C.; STEIN, E.

Über die Falksche ECP-Transformation und Verallgemeinerungen

Die ECP-*Transformation von Falk ordnet einem Polynom f vom Grad n und n paarweise verschiedenen Knoten* $\alpha_1, \dots, \alpha_n$ *einen Vektor* $(d_1, \dots, d_n)$ *zu, so daß das charakteristische Polynom der dyadisch gestörten quadratischen Diagonalmatrix*

$$\operatorname{diag}(\alpha_1, \dots, \alpha_n) - \begin{pmatrix} 1 \\ \vdots \\ 1 \end{pmatrix} \cdot (d_1, \dots, d_n)$$

ein Vielfaches von f ist. Diese Transformation wird für mehrfache Knoten betrachtet und auf dyadisch gestörte Bidiagonalmatrizen verallgemeinert. Der Sonderfall einfacher Knoten wird an Beispielen numerisch getestet.

Falk's ECP *transformation associates to a polynomial f of degree n and to n pairwise different knots* $\alpha_1, \dots, \alpha_n$ *a vector* $(d_1, \dots, d_n)$ *such that the characteristic polynomial of the dyadically perturbed quadratic diagonal matrix*

$$\operatorname{diag}(\alpha_1, \dots, \alpha_n) - \begin{pmatrix} 1 \\ \vdots \\ 1 \end{pmatrix} \cdot (d_1, \dots, d_n)$$

becomes a multiple of f. This transformation is considered for multiple knots and is generalized to dyadically perturbed bidiagonal matrices. The special case of simple knots is numerically tested by means of examples.

РХМ-преобразование Фалька (РХМ — Разложение Характеристижеского Многочлена) соответствует многочлена f степени n и n попарно различным узлам $\alpha_1, \dots, \alpha_n$ вектор $(d_1, \dots, d_n)$ так, что характеристический многочлен диадично возмущенной квадратичной диагональной матрицы

$$\operatorname{diag}(\alpha_1, \dots, \alpha_n) - \begin{pmatrix} 1 \\ \vdots \\ 1 \end{pmatrix} \cdot (d_1, \dots, d_n)$$

является кратным от f. Это преобразование рассматривается для кратных узлов и обобщается на диадично возмущенные бидиагональные матрицы. Особенный случай простых узлов численно испытывается на примерах.

1. Einleitung

Die ECP-Transformation wurde für den Spezialfall einfacher Knoten (Stützstellen) von FALK 1986 erstmals in [1] beschrieben und von den Autoren bereits in [2] untersucht. Bei dieser Expansion des charakteristischen Polynomes (kurz ECP) wird einem Polynom f vom Grad n mit dem Hauptkoeffizienten $f_n \in \mathbb{K} \setminus \{0\}$ und frei wählbaren paarweise verschiedenen Knoten $\alpha_1, \dots, \alpha_n$ ein Vektor $(d_1, \dots, d_n)^T \in \mathbb{K}^n$ durch

$$\forall i \in \{1, 2, \dots, n\} : d_i := \frac{f(\alpha_i)}{f_n \cdot \prod_{\substack{\nu=1 \\ \nu \neq i}}^{n} (\alpha_i - \alpha_\nu)} \tag{1.1}$$

zugeordnet. Diese Zahlen werden nach FALK als *Defekte* bezeichnet. Mit dem Defektvektor und der Abkürzung $e := (1, \dots, 1)^T \in \mathbb{R}^n$ kann die Matrix

$$\operatorname{diag}(\alpha_1, \dots, \alpha_n) - e \cdot (d_1, \dots, d_n) \tag{1.2}$$

betrachtet werden, deren charakteristisches Polynom $(-1)^n \cdot f/f_n$ ist.

Bei nichtlinearen Matrizeneigenwertproblemen mit Polynommatrizen kann das charakteristische Polynom f dieses Problemes auf ein spezielles Eigenwertproblem mit einer dyadisch gestörten Diagonalmatrix transformiert werden, deren numerische Behandlung einfacher sein kann.

Ausgehend von dieser Eigenschaft hat FALK einen *Eigenwertalgorithmus* ECP als einen sehr leistungsfähigen Algorithmen konstruiert [1, p. 425f]. Dabei werden die Knoten als Näherungen für die Nullstellen von f gewählt und iterativ verbessert.

Tatsächlich sind die Defekte bereits in den 60-iger Jahren zur simultanen Verbesserung aller Näherungen für die einfachen Nullstellen von Polynomen [7], [9] und später für Fehlerabschätzungen von Nullstellennäherungen verwendet worden [8]. Dieses allerdings ohne auf die Matrix in (1.2) Bezug zu nehmen; Hilfsmittel war allein die Darstellung, die sich durch Lagrange-Interpolation von f an den Knoten ergibt.

Für mehrfache Nullstellen von f versagen die erwähnten Verbesserungsverfahren. Eine direkte Verallgemeinerung der Aussage über die Matrix (1.2) ist nur in Sonderfällen moglich. Bei der numerischen Behandlung von linearen Eigenwertproblemen in der Strukturmechanik mit weichen und starren Bindungen treten aber ‚Nullstellenhaufen auf, die erhebliche numerische Probleme bereiten. Deshalb ist es wünschenswert, die ECP-Transformation auf mehrfache Knoten zu verallgemeinern.

In dieser Arbeit wird im 3. Abschnitt zunächst eine Verallgemeinerung der Matrix (1.2) für mehrfache Knoten und eine nicht konstante dyadische Störung und schließlich eine solche Verallgemeinerung vorgestellt, die sich beim Übergang auf die Klasse der dyadisch gestörten Bidiagonalmatrizen ergibt. Durch Entkopplung verschiedener

Knoten in der Nebendiagonale entstehen dabei Bedingungsgleichungen, die rekursiv aufgelöst werden können. Abschließend werden mit dem Satz von Gerschgorin Fehlerabschätzungen gewonnen, die denen aus [8] ähneln.

Im 4. Abschnitt wird ein kurzer Überblick über den Spezialfall einfacher Knoten gegeben und ein Algorithmus formuliert, mit dem im 5. Abschnitt drei numerische Beispiele behandelt werden. Einige Bemerkungen zur Leistungsfähigkeit des Algorithmus in den betrachteten Beispielen im Abschnitt 5.4 beschließen die Arbeit.

2. Notationen und Vorbereitungen

Mit $\mathbb{K}$ werden der Körper der reellen oder komplexen Zahlen, mit $\mathbb{N}$ die Menge der natürlichen und mit $\mathbb{N}_0$ die der nicht negativen ganzen Zahlen bezeichnet.

Es seien $\alpha_1, \dots, \alpha_m \in \mathbb{K}$, $m \in \mathbb{N}$, paarweise verschiedene Knoten, die mit den Vielfachheiten $m_1, \dots, m_m \in \mathbb{N}$ in der Aufzählung aller

$$N := \sum_{\nu=1}^{m} m_\nu \in \mathbb{N} \tag{2.1}$$

Knoten

$$\underbrace{\alpha_1, \dots, \alpha_1}_{m_1}, \quad \underbrace{\alpha_2, \dots, \alpha_2}_{m_2}, \quad \dots, \quad \underbrace{\alpha_m, \dots, \alpha_m}_{m_m} \tag{2.2}$$

vorkommen. Mit diesen Knoten werden einige Funktionen definiert. Für $i \in \{1, 2, \dots, m\}$ sei

$$\Pi_i : \mathbb{K} \to \mathbb{K}, \quad x \mapsto \prod_{\substack{\nu=1 \\ \nu \neq i}}^{m} (x - \alpha_\nu)^{m_\nu} \quad \text{sowie} \quad \Pi : \mathbb{K} \to \mathbb{K}, \quad x \mapsto \prod_{\nu=1}^{m} (x - \alpha_\nu)^{m_\nu}. \tag{2.3), (2.4}$$

Für genügend oft differenzierbare Funktionen $h : \mathbb{K} \to \mathbb{K}$ und $i \in \mathbb{N}$ bezeichne $\partial^i h$ die i-te Ableitung von h, und $\partial^0 h$ bezeichne die Funktion h selbst. Leere Produkte seien durch 1, leere Summen und leere Positionen in Matrizen seien durch 0 zu ersetzen.

Interpolation

Für ein fixiertes Polynom f über $\mathbb{K}$ vom Grad $N = \sum_{\nu=1}^{m} m_\nu$ mit dem Hauptkoeffizienten $f_N \neq 0$ werde das Polynom

$$r : \mathbb{K} \to \mathbb{K}, \quad x \mapsto f(x) - f_N \cdot \Pi(x) \tag{2.5}$$

betrachtet. r ist vom Grad kleiner oder gleich $N - 1$, so daß r das Hermite-Interpolationsproblem zu den Knoten (2.2) und den Daten

$$\underbrace{\partial^0 r(\alpha_1), \dots, \partial^{m_1 - 1} r(\alpha_1)}_{m_1}, \quad \underbrace{\partial^0 r(\alpha_2), \dots, \partial^{m_2 - 1} r(\alpha_2)}_{m_2}, \quad \dots, \quad \underbrace{\partial^0 r(\alpha_m), \dots, \partial^{m_m - 1} r(\alpha_m)}_{m_m} \tag{2.6}$$

exakt löst. Folglich läßt sich

$$r = \sum_{i=1}^{m} \sum_{j=1}^{m_i} \partial^{j-1} r(\alpha_i) \cdot p_i^j \tag{2.7}$$

für solche Polynome p_i^j über $\mathbb{K}$ schreiben, die vom Grad kleiner oder gleich $N - 1$ sind und für die gilt:

$$\forall i, \nu \in \{1, 2, \dots, m\}\ \forall j \in \{1, 2, \dots, m_i\}\ \forall \mu \in \{1, 2, \dots, m_\nu\} : \partial^{\mu - 1} p_i^j(\alpha_\nu) = \delta_{\mu, j} \cdot \delta_{\nu, i}.$$

Hierbei sei δ das Kronecker-Delta. Da diese Polynome selbst gewisse Interpolationsaufgaben lösen, folgt leicht, daß für jedes $i \in \{1, 2, \dots, m\}$ und jedes $j \in \{1, 2, \dots, m_i\}$ das Polynom p_i^j in der Form

$$\forall x \in \mathbb{K} : p_i^j(x) = q_i^j(x) \cdot (x - \alpha_i)^{j-1} \cdot \Pi_i(x)$$

mit einem Polynom q_i^j vom Grad kleiner oder gleich $m_i - j$ notiert werden kann.

Wenn man abschließend diese Darstellung und die Definition von r in (2.5) in die Interpolationsgleichung (2.7) einsetzt, dann folgt für jedes $x \in \mathbb{K} \setminus \{\alpha_1, \alpha_2, \dots, \alpha_m\}$

$$f(x) = f_N \cdot \Pi(x) \cdot \left(1 + \sum_{i=1}^{m} \sum_{j=1}^{m_i} \partial^{j-1} f(\alpha_i) \cdot \frac{q_i^j(x)}{(x - \alpha_i)^{m_i + 1 - j}}\right). \tag{2.8}$$

Lineare Algebra

Für $n \in \mathbb{N}$ bezeichne I_n die n-dimensionale Einheitsmatrix in $\mathbb{K}^{n \times n}$. Mit $\operatorname{diag}(d_1, d_2, \dots, d_n)$ werde die n-dimensionale Diagonalmatrix mit den Diagonalelementen $d_1, d_2, \dots, d_n \in \mathbb{K}$ in der ersten bis n-ten Position notiert. Durchgehend werde e für den Spaltenvektor $(1, \dots, 1)^T \in \mathbb{R}^*$ mit geeigneter Dimension geschrieben.

Für Vektoren $a, b \in \mathbb{K}^n$ bezeichne b^* den komplex-konjugierten und transponierten Zeilenvektor und $a \cdot b^* \in \mathbb{K}^{n \times n}$ das dynadische Produkt dieser Vektoren a und b.

In dieser Arbeit wird die *dyadisch gestörte Diagonalmatrix*

$$\operatorname{diag}(d_1, d_2, \dots, d_n) - a \cdot b^* \tag{2.9}$$

betrachtet. Im Zusammenhang mit der Sherman-Morrison-Formel [5], nach der

$$(I_n - a \cdot b^*)^{-1} = I_n + \frac{1}{1 - b^* \cdot a} \cdot a \cdot b^* \tag{2.10}$$

gilt (man vgl. auch [1], Bd. 1, S. 310), kann auch die Determinante dieser Matrix,

$$\det (I_n - a \cdot b^*) = 1 - b^* \cdot a\,, \tag{2.11}$$

bestimmt werden [3]. Mit (2.11) gilt natürlich auch

$$\det (I_n - a \cdot b^{\mathsf{T}}) = 1 - b^{\mathsf{T}} \cdot a\,, \tag{2.12}$$

wobei b^{T} der zu b transponierte Vektor sei. Durch Multiplikation der dyadisch gestörten Einheitsmatrix in (2.12) mit einer regulären Diagonalmatrix kann man leicht die Gleichung

$$\det \left(\operatorname{diag}(d_1, \ldots, d_n) - a \cdot b^{\mathsf{T}}\right) = \left(\prod_{\nu=1}^{n} d_\nu\right) - \sum_{i=1}^{n} a_i \cdot b_i \cdot \left(\prod_{\substack{\nu=1 \\ \nu \neq i}}^{n} d_\nu\right) \tag{2.13}$$

verifizieren, die dann aus Stetigkeitsgründen auch für beliebige Diagonalmatrizen gilt.

3. Über Verallgemeinerte ECP-Transformationen

3.1 Über eine allgemeine ECP-Transformation für dyadisch gestörte Diagonalmatrizen

Es werden die im 2. Abschnitt eingeführten Größen betrachtet. Durch Kombination von (2.8) und (2.12) kann das Polynom f als Vielfaches einer Determinante dargestellt werden. In dem wichtigsten Spezialfall ergibt sich f dann als ein charakteristisches Polynom zu einer dyadisch gestörten Diagonalmatrix.

Um diese Kombination möglichst allgemein abhandeln zu können, werden die Summe und das Produkt in (2.13) zunächst beliebig zusammengefaßt.

Definition 1: Für $n \in \mathbb{N}$ seien

$$D_1, D_2, \ldots, D_n\colon \mathbb{K} \setminus \{\alpha_1, \ldots, \alpha_m\} \to \mathbb{K} \quad \text{und} \quad p_0, p_1, \ldots, p_n\colon \mathbb{K} \to \mathbb{K}$$

Funktionen mit den Eigenschaften, daß für jedes $x \in \mathbb{K} \setminus \{\alpha_1, \ldots, \alpha_m\}$ gilt:

$$\sum_{\nu=1}^{n} D_\nu(x) = \sum_{i=1}^{m} \sum_{j=1}^{m_i} \partial^{j-1} f(\alpha_i) \cdot \frac{q_i^j(x)}{(x - \alpha_i)^{m_i+1-j}} \quad \text{und} \quad \prod_{\nu=0}^{n} p_\nu(x) = \Pi(x)\,. \tag{3.1), (3.2}$$

Satz 1: *Unter obigen Voraussetzungen gilt für alle* $x \in \mathbb{K} \setminus \{\alpha_1, \ldots, \alpha_m\}$

$$f(x) = f_N \cdot p_0(x) \cdot \det\left(\operatorname{diag}\left(p_1(x), \ldots, p(x)\right) + \begin{pmatrix}1\\ \vdots \\ 1\end{pmatrix} \cdot \left(p_1(x) \cdot D_1(x), \ldots, p_n(x) \cdot D_n(x)\right)\right). \tag{3.3}$$

Beweis: Ausgehend von der Darstellung (2.8) wird die Doppelsumme durch (3.1) umgeschrieben. Dabei entsteht in der Klammer eine Summe, die mit (2.12) als Determinante einer dyadisch gestörten n-dimensionalen Einheitsmatrix geschrieben werden kann. Dabei werde als Spaltenvektor $(1, \ldots, 1)^{\mathsf{T}}$ verwendet. Wenn man das Produkt $\Pi(x)$ durch (3.2) ersetzt, kann jeder der Faktoren $p_1(x), \ldots, p_n(x)$ mit der ersten bis n-ten Spalte der Determinante multipliziert werden. Hieraus folgt (3.3). ∎

Beispiel 1: Zunächst werde die Doppelsumme in (3.1) mit $n = N$ als einfache Summe und das Produkt in (3.2) als Produkt von den entsprechenden Linearfaktoren geschrieben. Dazu sei $(i_1, j_1), (i_2, j_2), \ldots, (i_N, j_N)$ eine Abzählung der Menge

$$\{(i, j) \mid i \in \{1, \ldots, m\}, j \in \{1, \ldots, m_i\}\}\,.$$

Dann werde für jedes $\nu \in \{1, \ldots, N\}$ und jedes $x \in \mathbb{K} \setminus \{\alpha_1, \ldots, \alpha_m\}$

$$D_\nu(x) := \partial^{j_\nu - 1} f(\alpha_{i_\nu}) \cdot \frac{q_{i_\nu}^{j_\nu}(x)}{(x - \alpha_{i_\nu})^{m_{i_\nu}+1-j_\nu}}$$

und $p_0(x) := 1$ sowie $p_\nu(x) := x - \alpha_{i_\nu}$ definiert. Damit ergibt sich als typisches Element des Zeilenvektors in (3.3) für $\nu \in \{1, \ldots, N\}$ und $x \in \mathbb{K} \setminus \{\alpha_1, \ldots, \alpha_m\}$

$$p_\nu(x) \cdot D_\nu(x) = \partial^{j_\nu - 1} f(\alpha_{i_\nu}) \cdot q_{i_\nu}^{j_\nu}(x) \cdot (x - \alpha_{i_\nu})^{j_\nu - m_{i_\nu}}\,.$$

Dieses Element ist von x unabhängig, wenn $q_{i_\nu}^{j_\nu}$ ein Vielfaches von $(x - \alpha_{i_\nu})^{m_{i_\nu} - j_\nu}$ ist. Weil $q_{i_\nu}^{j_\nu}$ höchstens vom Grad $m_{i_\nu} - j_\nu$ ist, muß dann für alle $x \in \mathbb{K}$

$$q_{i_\nu}^{j_\nu}(x) = c_{i_\nu}^{j_\nu} \cdot (x - \alpha_{i_\nu})^{m_{i_\nu} - j_\nu} \tag{3.4}$$

für ein $c_{i_\nu}^{j_\nu} \in \mathbb{K}$ gelten.

Ein interessanter Fall ist nun der, daß der Zeilenvektor in (3.3) von x unabhängig wird und sich f in (3.3) als charakteristisches Polynom zu einem linearen speziellen Eigenwertproblem ergibt.

Wenn alle Elemente des Zeilenvektors in (3.3) von x unabhängig sein sollen, dann folgt aus (3.4) und (2.8), daß das Polynom f an den Stellen $\alpha_1, \ldots, \alpha_m$ jeweils eine $(m_1 - 1), \ldots, (m_m - 1)$-fache Nullstelle besitzt. Folglich sind

dann in dem Zeilenvektor aus (3.3) nur diejenigen Elemente von 0 verschieden, in denen von f die Ableitung $(m_i - 1)$-ter Ordnung an der Stelle α_i vorkommt. In diesem Fall ergibt sich aus (3.3) und einer genauen Bestimmung der c_i^j, $i \in \{1, \dots, m\}$, $j \in \{1, \dots, m_i\}$, der folgende

Satz 2: *Es sei f ein Polynom vom Grad $N \in \mathbb{N}$ mit dem Hauptkoeffizienten $f_N \in \mathbb{K} \setminus \{0\}$. f besitze bei den paarweise verschiedenen Knoten $\alpha_1, \alpha_2, \dots, \alpha_m \in \mathbb{K}$ Nullstellen mit den jeweiligen Vielfachheiten $m_1 - 1$, $m_2 - 1, \dots, m_m - 1 \in \mathbb{N}_0$. Für diese Vielfachheiten gelte $N = \sum_{i=1}^{m} m_i$. Für jedes $i \in \{1, \dots, m\}$ werde der „Defekt"*

$$d_i := \frac{\partial^{m_i - 1} f(\alpha_i)}{f_N \cdot (m_i - 1)! \cdot \Pi_i(\alpha_i)} \tag{3.5}$$

definiert. Dann gilt für jedes $x \in \mathbb{K}$

$$f(x) = (-1)^N \cdot f_N \cdot \{ \prod_{\nu=1}^{m} (\alpha_\nu - x)^{m_\nu - 1} \} \cdot \det \left(\operatorname{diag}(\alpha_1, \dots, \alpha_m) - e \cdot (d_1, \dots, d_m) - x \cdot I_m \right) \tag{3.6}$$

Alternativbeweis: Es werde für jedes $x \in \mathbb{K}$ der Rest

$$R(x) := f(x) - (-1)^N \cdot f_N \cdot \{ \prod_{\nu=1}^{m} (\alpha_\nu - x)^{m_\nu - 1} \} \cdot \det \left(\operatorname{diag}(\alpha_1, \dots, \alpha_m) - e \cdot (d_1, \dots, d_m) - x \cdot I_m \right) \tag{3.7}$$

betrachtet. R ist ein Polynom vom Grad kleiner oder gleich $N - 1$ und besitzt nach den Voraussetzungen an das Polynom f und die Knoten $\alpha_1, \dots, \alpha_m$ an diesen Stellen Nullstellen in der Vielfachheit $m_1 - 1, \dots, m_m - 1$. Es genügt daher zu zeigen, daß R in den Knoten sogar Nullstellen der Vielfachheit $m_1, \dots, m_m$ hat. Sei dazu $i \in \{1, 2, \dots, m\}$. Dann ist

$$\begin{aligned} \partial^{m_i - 1} R(\alpha_i) = {} & \partial^{m_i - 1} f(\alpha_i) - (-1)^N \cdot f_N \cdot (m_i - 1)! \times \\ & \times (-1)^{m_i - 1} \{ \prod_{\substack{\nu=1 \\ \nu \neq i}}^{m} (\alpha_\nu - \alpha_i)^{m_\nu - 1} \det \left(\operatorname{diag}(\alpha_1 - \alpha_i, \dots, \alpha_m - \alpha_i) - e \cdot (d_1, \dots, d_m) \right) . \end{aligned} \tag{3.8}$$

Nach (2.13) ergibt sich die Determinante in (3.8) zu

$$- d_i \cdot \prod_{\substack{\nu=1 \\ \nu \neq i}}^{m} (\alpha_\nu - \alpha_i) ,$$

so daß mit (3.5) schließlich $\partial^{m_i - 1} R(\alpha_i) = 0$ folgt. ∎

Bemerkung 1:

(i) Der Satz 2 verallgemeinert die Falksche ECP-Transformation in [1] auf mehrfache Knoten. Der Beweis verallgemeinert den aus [2]. — Damit sich die Determinante in (3.3) als charakteristisches Polynom zu einem linearen Eigenwertproblem schreiben läßt, müssen $p_1, \dots, p_n$ Polynome vom Grad 1, und die dyadische Störung muß konstant sein. Diese Forderung führte schließlich auf die Voraussetzungen bezüglich der Nullstelleneigenschaften von f. Andererseits ist klar, daß sich das Polynom f für mehrfache Knoten nur dann als charakteristisches Polynom eines speziellen Eigenwertproblemes mit dyadisch gestörter Diagonalmatrix ergeben kann, wenn f die im Satz 2 verlangten Voraussetzungen über die Lage der Nullstellen erfüllt. Aus (2.13) folgt nämlich für ein eben beschriebenes Polynom mit $d_\nu := \alpha_\nu - x$, daß dieses Polynom im Knoten α_ν eine Nullstelle der Ordnung m_ν hat.

(ii) Im Beweis zu Satz 1 wurde der Spaltenvektor $(1, \dots, 1)^T$ verwendet, um anschließend (2.12) anzuwenden. Dieses Vorgehen liefert dann im Satz 2 den Spaltenvektor in (3.6). Man könnte aber natürlich auch andere Darstellungen wählen, die dann im Satz 2 ein spezielles Eigenwertproblem mit einer durch $a \cdot b^T$ gestörten Diagonalmatrix erzeugen. Hierbei ist lediglich $d_i = a_i \cdot b_i$ für alle $i \in \{1, 2, \dots, m\}$ zu gewährleisten. Insbesondere kann dann $a_i := \operatorname{sign}(d_i) \cdot \sqrt[3]{|d_i|}$ und $b_i := \sqrt[3]{|d_i|}$ gewählt werden. Wegen des Vorzeichens ist die Begleitmatrix der ECP-Transformation i. a. nicht symmetrisch. Man kann das Vorzeichen im reellen Fall aber in einer Diagonalmatrix $J := \operatorname{diag}(\operatorname{sign}(d_1), \dots, \operatorname{sign}(d_m))$ zusammenfassen und erhält im Satz 2 ein allgemeines Eigenwertproblem mit einer sogenannten J-symmetrischen Matrix $\operatorname{diag}(\alpha_1, \dots, \alpha_m) - a_i \cdot b_i^T$. Zum Begriff einer J-symmetrischen Matrix vgl. man etwa [6].

Beispiele 2, 3: Es soll eine weitere kanonische Konkretisierung von Satz 1 betrachtet werden. Dazu werde $n = m$ und für jedes $x \in \mathbb{K} \setminus \{\alpha_1, \dots, \alpha_m\}$ und jedes $i \in \{1, 2, \dots, m\}$

$$D_i(x) = \sum_{j=1}^{m_i} \partial^{j-1} f(\alpha_i) \cdot \frac{q_i^j(x)}{(x - \alpha)^{m_i + 1 - j}} \quad \text{und} \quad p_i(x) = (x - \alpha_i)^{m_i} ,$$

sowie $p_0 = 1$ gesetzt. Dann gilt für jedes $i \in \{1, 2, \dots, m\}$ und jedes $x \in \mathbb{K}$

$$p_i(x) \cdot D_i(x) = \sum_{j=1}^{m_i} \partial^{j-1} f(\alpha_i) \cdot q_i^j(x) \cdot (x - \alpha_i)^{j-1} .$$

Damit ist f in (3.3) als charakteristisches Polynom eines nichtlinearen Eigenwertproblemes dargestellt. Wenn sich in dieser Darstellung ein lineares Eigenwertproblem ergeben soll, so ist die Störung entweder nicht konstant, oder es gilt eine ähnliche Beziehung wie in (3.4), die dann wieder die Nullstellenvoraussetzungen des Satzes 2 nach sich zieht.

Wenn man abweichend zu oben $p_i(x) = (x - \alpha_i)^{n_i}$ für $i \in \{1, 2, \dots, m\}$ für ein $n_i \in \{1, 2, \dots, m_i\}$ setzt, dann ist die Störung i. a. eine rationale Funktion mit bekanntem Pol α_i.

Numerisches Beispiel: Es werde das normierte Polynom f vom Grad 4 betrachtet, das durch einen Hauptkoeffizienten 1 und die Werte $f(0) = 0$, $f'(0) = a$, $f(1) = 1$ und $f'(1) = 1$ gegeben ist. Hierbei ist $a \in \mathbb{R}$ ein reeller Parameter. Die Nullstelle von f an der Stelle 0 sei bereits bekannt, von den anderen Nullstellen sei die ‚in der Nähe von 0' gesucht. Eine einfache Berechnung würde

$$\forall x \in \mathbb{K}: f(x) := a \cdot x + (1 - a) \cdot x^2 + (a - 1) \cdot x^2 \cdot (x - 1) + x^2 \cdot (x - 1)^2$$

ergeben, so daß man das Problem mit einem Newtonschen Verfahren lösen könnte. Hier soll jedoch demonstriert werden, daß die Informationen an den beiden zweifachen Knoten $\alpha_1 := 0$ und $\alpha_2 := 1$ ($m_1 := 2$, $m_2 := 2$, $N := 4$, $f_N := 1$) benutzt werden können die gesuchte Nullstelle zu finden. Für die Ausgangsdaten lassen sich die Polynome p_i^j aus Interpolationsbedingungen ermitteln. Danach ergeben sich die Polynome q_i^j zu

$$q_1^1(x) = 2x + 1\,, \qquad q_1^2(x) = 1\,, \qquad q_2^1(x) = 3 - 2x \quad \text{und} \quad q_2^2(x) = 1$$

für jedes $x \in \mathbb{K}$. Wenn man weiter $n = 2$ und die Funktionen D_1 und D_2 nach diesem Beispiel, aber abweichend $p_0(x) := x \cdot (x - 1)$, $p_1(x) := x$ und $p_2(x) := x - 1$ für $x \in \mathbb{K}$ als ‚gemischte Form' von Beispiel 1 und dem gerade betrachteten Beispiel wählt, dann ergibt sich nach Satz 1 für alle $x \in \mathbb{K}$

$$f(x) = x \cdot (x - 1) \cdot \det\left(\operatorname{diag}(x, x - 1) + \begin{pmatrix}1\\1\end{pmatrix} \cdot \left(a, \frac{2 - x}{x - 1}\right)\right).$$

Die Determinante in dieser Darstellung läßt sich als charakteristisches Polynom zu einem speziellen Eigenwertproblem in x mit der nichtkonstanten Matrix

$$\forall x \in \mathbb{K}: A_x := \begin{pmatrix} -a & \dfrac{x - 2}{x - 1} \\ -a & 1 + \dfrac{x - 2}{x - 1} \end{pmatrix}$$

auffassen. Die Werte von A_x werden sich für betragskleine $x \in \mathbb{K}$ nur wenig ändern, so daß man eine ‚Schaukeliteration' versuchen kann, bei der in jedem Schritt einer Näherung x_ν zunächst die Matrix A_{x_ν} und dieser Matrix einen Eigenwert $x_{\nu+1}$ (hier der betragskleinste, falls eindeutig gegeben) zuordnet. Dieser Eigenwert kann hier als Wurzel des charakteristischen Polynomes von A_{x_ν} direkt und sehr genau, sonst iterativ als Näherung ermittelt werden.

Als Zahlenbeispiel wurden verschiedene Parameter a für den Startwert $-a$ getestet und stets der kleinere Eigenwert gewählt. Dabei ergeben sich die in Tabelle 1 gezeigten Werte.

Tabelle 1. Zahlenwerte zum numerischen Beispiel nach der ersten Methode

a = 1.0000000000E+000	a = 1.0000000000E−001	a = 1.0000000000E−002
−1.000000000000E+000	−1.000000000000E−001	−1.000000000000E−002
−5.000000000000E−001	−3.515865876544E−002	−3.351823294266E−003
−4.683749459844E−001	−3.447666982972E−002	−3.344477297584E−003
−4.658034180507E−001	−3.446919028015E−002	−3.344469144586E−003
−4.655904865888E−001	−3.446910821311E−002	−3.344469135537E−003
−4.655728288118E−001	−3.446910731265E−002	−3.344469135527E−003
−4.655713643236E−001	−3.446910730277E−002	−3.344469135527E−003
−4.655712428617E−001	−3.446910730266E−002	−3.344469135527E−003
−4.655712327878E−001	−3.446910730266E−002	−3.344469135527E−003
−4.655712319523E−001	−3.446910730266E−002	−3.344469135527E−003
−4.655712318830E−001	−3.446910730266E−002	−3.344469135527E−003

Die Konvergenz kann hier durch die linearisierte Berücksichtigung der Änderungen von A_x verbessert werden. Für betragskleine x und x_ν ist

$$A_x \approx A_{x_\nu} + (x - x_\nu) \cdot \begin{pmatrix} 0 & 1/(x_\nu - 1)^2 \\ 0 & 1/(x_\nu - 1)^2 \end{pmatrix}.$$

Das Eigenwertproblem wird nun allgemein und lautet in jedem Schritt für bekanntes x_ν und gesuchtes $x_{\nu+1}$

$$0 \overset{!}{=} \det\left(A_{x_\nu} - x_{\nu+1} \cdot \begin{pmatrix} 0 & 1/(x_\nu - 1)^2 \\ 0 & 1/(x_\nu - 2)^2 \end{pmatrix} - x \cdot \begin{pmatrix} 1 & -1/(x_\nu - 1)^2 \\ 0 & 1 - 1/)x_\nu - 1)^2 \end{pmatrix}\right).$$

Durch Lösen dieser quadratischen Gleichung ergeben sich die in Tabelle 2 gezeigten Zahlenwerte.

Tabelle 2. Zahlenwerte zum numerischen Beispiel nach der zweiten Methode

a = 1.0000000000E+000	a = 1.0000000000E−001	a = 1.0000000000E−002
−1.000000000000E+000	−1.000000000000E−001	−1.000000000000E−002
−4.748096336327E−001	−3.450983281003E−002	−3.344517544264E−003
−4.655764270363E−001	−3.446910732045E−002	−3.344469135527E−003
−4.655712318784E−001	−3.446910730266E−002	−3.344469135527E−003
−4.655712318768E−001	−3.446910730266E−002	−3.344469135527E−003
−4.655712318768E−001	−3.446910730266E−002	−3.344469135527E−003
−4.655712318768E−001	−3.446910730266E−002	−3.344469135527E−003
−4.655712318768E−001	−3.446910730266E−002	−3.344469135527E−003
−4.655712318768E−001	−3.446910730266E−002	−3.344469135527E−003
−4.655712318768E−001	−3.446910730266E−002	−3.344469135527E−003
−4.655712318768E−001	−3.446910730266E−002	−3.344469135527E−003

3.2 Über dyadisch gestörte Bidiagonalmatrizen

In Satz 2 können auch mehrfache Knoten zur ECP-Transformation Verwendung finden. Wenn diese Transformation auf ein lineares Eigenwertproblem führen soll, so müssen jedoch nach Bemerkung 1(i) die im Satz 2 formulierten Nullstelleneigenschaften erfüllt sein. Wenn man diese Nullstelleneigenschaften umgehen will, dann muß man statt der Klasse der dyadisch gestörten Diagonalmatrizen eine andere Klasse von Matrizen betrachten. Wegen der Form der wohlbekannten Frobeniusschen oder Güntherschen Begleitmatrix werden hier dyadisch gestörte Bidiagonalmatrizen betrachtet. Als Spezialfälle ergeben sich dann tatsächlich Aussagen der gewünschten Form.

Lemma 1: *Für $n \in \mathbb{N}$ und $\lambda_1, \lambda_2, \dots, \lambda_n, \mu_1, \mu_2, \dots, \mu_{n-1} \in \mathbb{K}$ bezeichne*

$$\operatorname{bidiag}(\lambda_1, \dots, \lambda_n \mid \mu_1, \dots, \mu_{n-1}) := \begin{pmatrix} \lambda_1 & \mu_1 & & \\ & \lambda_2 & \ddots & \\ & & \ddots & \mu_{n-1} \\ & & & \lambda_n \end{pmatrix} \in \mathbb{K}^{n\times n} \tag{3.9}$$

die durch diese Daten gegebene Bidiagonalmatrix, wobei leere Positionen durch Nullen und für $n = 1$ diese Matrix durch (λ_1) zu ersetzen ist. Es sei weiter $a_1, \dots, a_n, b_1, \dots, b_n \in \mathbb{K}$.

Für $\lambda_1, \dots, \lambda_n \in \mathbb{K} \setminus \{0\}$ ist $\operatorname{bidiag}(\lambda_1, \dots, \lambda_n \mid \mu_1, \dots, \mu_{n-1})$ *regulär und die Inverse durch*

$$\forall i, j \in \{1, \dots, n\}: \operatorname{bidiag}(\lambda_1, \dots, \lambda_n \mid \mu_1, \dots, \mu_{n-1})^{-1}_{i,j} = \begin{cases} 0 & , \quad \text{falls} \quad i > j \\ (-1)^{i+j} \cdot \dfrac{\prod_{\nu=i}^{j-1} \mu_\nu}{\prod_{\varkappa=i}^{j} \lambda_\varkappa} & , \quad \text{falls} \quad i \leqq j \end{cases} \tag{3.10}$$

gegeben. Für alle $x \in \mathbb{K}$ ist

$$\det\left\{\operatorname{bidiag}(\lambda_1 - x, \dots, \lambda_n - x \mid \mu_1, \dots, \mu_{n-1}) - \begin{pmatrix} a_1 \\ \vdots \\ a_n \end{pmatrix} \cdot (b_1, \dots, b_n)\right\} =$$

$$= \prod_{\nu=1}^{n} (\lambda_\nu - x) - \sum_{1\leqq i\leqq j\leqq n} (-1)^{i+j} \cdot a_j \cdot b_i \cdot \Big(\prod_{\varkappa=1}^{i-1} (\lambda_\varkappa - x)\Big) \cdot \Big(\prod_{\varkappa=i}^{j-1} \mu_\varkappa\Big) \cdot \Big(\prod_{\varkappa=j+1}^{n} (\lambda_\varkappa - x)\Big), \tag{3.11}$$

wobei leere Produkte durch 1 zu ersetzen sind. ∎

Bemerkung 2: Wie in der Herleitung zu Satz 2 gibt es an dieser Stelle verschiedene Möglichkeiten, das Teilresultat in (3.11) in einen Satz umzuwandeln, der aussagt, wie die Größen in (3.11) zu wählen sind, damit ein gegebenes Polynom f sich als Vielfaches eines charakteristischen Polynomes zu einem speziellen Eigenwertproblem mit einer dyadisch gestörten Bidiagonalmatrix ergibt. Zum einen kann man (2.8) mit (3.11) identifizieren und Bedingungen für die Parameter ableiten. Zum anderen ist eine Identität zwischen Polynomen herzustellen, die durch Interpolationsbedingungen konstruiert werden kann. Dazu sei f ein Polynom vom Grad n mit dem Hauptkoeffizienten $f_n \in \mathbb{K} \setminus \{0\}$. Wenn man das Polynom in (3.11) mit $(-1)^N \cdot f_n$ multipliziert und von f subtrahiert, dann verbleibt ein Polynom vom Grade $n - 1$, das durch Interpolation an n Knoten identifiziert werden kann. Als Interpolationsstellen kann man etwa $\lambda_1, \dots, \lambda_n$ aus (3.11) verwenden. Um diese Interpolationsbedingungen zu entkoppeln, werden wie bei einer Jordanschen Normalenform die Nebendiagonalelemente gleich Null gesetzt, die zwei Blöcke von gleichen Hauptdiagonalelementen trennen.

Definition 2: Es seien die Bezeichnungen des Kapitels 2 verwendet. Für jedes $i \in \{1, 2, \dots, m\}$ sei

$$a_i := (a_i^{(1)}, a_i^{(2)}, \dots, a_i^{(m_i)})^\mathsf{T}, \qquad b_i := (b_i^{(1)}, b_i^{(2)}, \dots, b_i^{(m_i)})^\mathsf{T} \in \mathbb{K}^{m_i}$$

und

$$J_i := \operatorname{bidiag}(\alpha_i, \dots, \alpha_i \mid \beta_i^{(1)}, \beta_i^{(2)}, \dots, \beta_i^{(m_i-1)}) \in \mathbb{K}^{m_i \times m_i}.$$

Mit diesen Größen werde weiter

$$\operatorname{diag}(J_1, \dots, J_m) := \begin{pmatrix} J_1 & & \\ & \ddots & \\ & & J_m \end{pmatrix} \in \mathbb{K}^{N\times N}$$

sowie

$$(a_1, \dots, a_m)^\mathsf{T} := (a_1^{(1)}, a_1^{(2)}, \dots, a_1^{(m_1)}, a_2^{(1)}, \dots, a_m^{(m_m)})^\mathsf{T}$$

und

$$(b_1, \dots, b_m)^\mathsf{T} := (b_1^{(1)}, b_1^{(2)}, \dots, b_1^{(m_1)}, b_2^{(1)}, \dots, b_m^{(m_m)})^\mathsf{T} \in \mathbb{K}^N$$

gesetzt. Ferner sei χ das charakteristische Polynom, das für alle $x \in \mathbb{K}$ durch

$$\chi(x) := \det \operatorname{diag}\left((J_1, \dots, J_m) - \begin{pmatrix} a_1 \\ \vdots \\ a_m \end{pmatrix} \cdot (b_1^\mathsf{T}, \dots, b_m^\mathsf{T}) - x \cdot I_N\right) \tag{3.12}$$

gegeben ist.

Satz 3: *Es sei f ein Polynom in $\mathbb{K}$ vom Grad N mit dem Hauptkoeffizienten $f_N \in \mathbb{K} \setminus \{0\}$. Das Polynom χ sei durch* (3.12) *gegeben. Dann gilt*

$$f = (-1)^N \cdot f_N \cdot \chi$$

genau dann, wenn

$$\forall i \in \{1, 2, \dots, m\}\ \forall j \in \{1, 2, \dots, m_i\}: \quad \partial^{j-1} f(\alpha_i) = \sum_{1 \le k \le \nu \le j} f_N \cdot b_i^{(k)} \cdot a_i^{(m_i+k-\nu)} \cdot \frac{(j-1)!}{(j-\nu)!} \times \times \Big(\prod_{\varkappa=k}^{m_i+k-\nu-1} \beta_i^{(\varkappa)} \Big) \cdot \partial^{j-\nu} \Pi_i(\alpha_i) \,. \tag{3.13}$$

Beweis: f und $(-1)^N \cdot f_N \cdot \chi$ sind Polynome vom Grad N mit gleichem Hauptkoeffizienten. Folglich ist genau dann $f = (-1)^N \cdot f_N \cdot \chi$, wenn beide Polynome an den N Knoten aus (2.2) (in Vielfachheit) dieselben Werte annehmen. Damit genügt es zu zeigen, daß die rechte Seite in (3.13) gerade gleich $(-1)^N \cdot f^N \cdot \partial^{j-1}\chi(\alpha_i)$ ist. Um andere Darstellungen für χ zu erhalten, wird Lemma 1 angewendet.

Es sei $n = N$, und $(\lambda_1, \dots, \lambda_n)$ seien die Knoten aus (2.2) mit den Vielfachheiten $m_1, \dots, m_M \in \mathbb{N}$, wobei $N = \sum_{i=1}^{m} m_i$ gilt. Im Lemma 1 werde weiter $(a_1, \dots, a_n)^T$ durch $(a_1^{(1)}, \dots, a_1^{(m_1)}, a_2^{(1)}, \dots, a_m^{(m_m)})^T$, $(b_1, \dots, b_n)^T$ durch $(b_1^{(1)}, \dots, b_1^{(m_1)}, b_2^{(1)}, \dots, b_m^{(m_m)})^T$ und $(\mu_1, \dots, \mu_{n-1})$ durch

$$\underbrace{(\beta_1^{(1)}, \beta_1^{(2)}, \dots, \beta_1^{(m_1-1)}}_{m_1-1}, 0, \underbrace{\beta_2^{(1)}, \beta_2^{(2)}, \dots \beta, {}_2^{(m_2-1)}}_{m_2-1}, 0, \beta_3^{(1)}, \dots, 0, \underbrace{\beta_m^{(1)}, \beta_m^{(2)}, \dots, \beta_m^{(m_m-1)})}_{m_m-1} \tag{3.14}$$

ersetzt. Dann kann χ nach (3.11) berechnet werden. Wegen der speziellen Ersetzung in (3.14) kann man die Summe aus (3.11) schließlich so umordnen, daß für alle $x \in \mathbb{K}$

$$(-1)^N \cdot \chi(x) = \Pi(x) + \sum_{i=1}^{m} \sum_{k=1}^{m_i} \sum_{l=k}^{m_i} a_i^{(l)} b_i^{(k)} \cdot (x - a_i)^{m_i - l + k - 1} \cdot \Big(\prod_{\varkappa=k}^{l-1} \beta_i^{(\varkappa)} \Big) \cdot \Pi_i(x) \tag{3.15}$$

gilt.

Eine $(j-1)$-malige Differentiation von (3.15) zeigt dann leicht (etwa unter Verwendung der Leibnizschen Produktregel), daß die rechte Seite in (3.13) gleich $(-1)^N \cdot f_N \cdot \partial^{j-1}\chi(\alpha_i)$ ist. ∎

Folgerung 1: *Es gelten die Bezeichnungen aus Definition 2 und Satz 3. Für jedes $i \in \{1, \dots, m\}$ und $j \in \{1, \dots, m_i\}$ sei $\beta_i^{(j)} \in \mathbb{K} \setminus \{0\}$ fixiert. Dann ergeben sich die beiden folgenden Spezialfälle des letzten Satzes.*

(i) *Für jedes $i \in \{1, \dots, m\}$ werde*

$$c_i := (c_i^{(1)}, c_i^{(2)}, \dots, c_i^{(m_i)})^T \in \mathbb{K}^{m_i}$$

durch

$$c_i^{(1)} := f(\alpha_i) \Big(f_N \cdot \Pi_i(\alpha_i) \cdot \prod_{\varkappa=1}^{m_i-1} \beta_i^{(\varkappa)} \Big)$$

und für $j \in \{2, \dots, m_i\}$ durch

$$c_i^{(j)} := \Bigg[\frac{\partial^{j-1} f(\alpha_i)}{(j-1)! \cdot f_N} - \sum_{k=1}^{j-1} c_i^{(k)} \cdot \Bigg\{ \sum_{\nu=0}^{j-k} \frac{\partial^\nu \Pi_i(\alpha_i)}{\nu!} \cdot \prod_{\varkappa=0}^{m_i+\nu-j-1} \beta_i^{(k+\varkappa)} \Bigg\} \Bigg] \Bigg/ \Big(\Pi_i(\alpha_i) \cdot \prod_{\varkappa=j}^{m_i-1} \beta_i^{(\varkappa)} \Big) \tag{3.16}$$

rekursiv definiert. Dann gilt für alle $x \in \mathbb{K}$

$$f(x) = (-1)^N \cdot f_N \cdot \det \{ \operatorname{diag}(J_1, \dots, J_m) - e \cdot (c_1^T, \dots, c_m^T) - x \cdot I_N \} \,. \tag{3.17}$$

(ii) *Für jedes $i \in \{1, \dots, m\}$ werde*

$$e_i := (\underbrace{0, 0, \dots, 0}_{m_i-1}, 1)^T \in \mathbb{K}^{m_i}$$

sowie

$$d_i := (d_i^{(1)}, d_i^{(2)}, \dots, d_i^{(m_i)})^T \in \mathbb{K}^{m_i}$$

durch

$$d_i^{(1)} := f(\alpha_i) / \Big(f_N \cdot \Pi_i(\alpha_i) \cdot \prod_{\varkappa=1}^{m_i-1} \beta_i^{(\varkappa)} \Big)$$

und für $j \in \{2, \dots, m_i\}$ durch

$$d_i^{(j)} := \frac{\partial^{j-1} f(\alpha_i)}{(j-1)! \cdot f_N \cdot \Pi_i(\alpha_i) \cdot \prod_{\varkappa=j}^{m_i-1} \beta_i^{(\varkappa)}} - \sum_{k=1}^{j-1} d_i^{(k)} \cdot \frac{\partial^{j-k} \Pi_i(\alpha_i) \cdot \prod_{\varkappa=k}^{j-1} \beta_i^{(\varkappa)}}{(j-k)! \cdot \Pi_i(\alpha_i)} \tag{3.18}$$

rekursiv definiert. Dann gilt für alle $x \in \mathbb{K}$

$$f(x) = (-1)^N \cdot f_N \cdot \det \left\{ \operatorname{diag}(J_1, \dots, J_m) - \begin{pmatrix} e_1 \\ \vdots \\ e_m \end{pmatrix} \cdot (d_1^T, \dots, d_m^T) - x \cdot I_N \right\} . \tag{3.19}$$

Beweis: Im Satz 3 wird für $i \in \{1, \dots, m\}$ im ersten Fall $a_i = (1, \dots, 1)^T \in \mathbb{K}^{m_i}$, $b_i := c_i$ und im zweiten $a_i = e_i$, $b_i := d_i$ gesetzt. Dann ergeben sich spezielle Interpolationsbedingungen in (3.13), die leicht umgeformt und zur rekursiven Berechnung der beteiligten Größen in (3.16) bzw. (3.18) herangezogen werden können. ∎

Bemerkung 3:

(i) Weitere Spezialfälle lassen sich völlig analog für $b_i^{(1)} = 1, b_i^{(2)} = b_i^{(3)} = \ldots = b_i^{(m_i)} = 0$ oder $b_i^{(1)} = b_i^{(2)} = b_i^{(3)} = \ldots = b_i^{(m_i)} = 1$ gewinnen.

(ii) Die Aussage von Satz 2 ist als weiterer Spezialfall enthalten. Setzt man nämlich $\beta_i^{(j)} := 0$ (für jedes $i \in \{1, \ldots, m\}$ und jedes $j \in \{1, \ldots, m_i\}$), so lauten die Interpolationsbedingungen aus (3.13)

$$\forall i \in \{1, \ldots, m\}\ \forall j \in \{1, \ldots, m_i - 1\}: \partial^{j-1} f(\alpha_i) = 0$$

und

$$\partial^{m_i - 1} f(\alpha_i) = f_N \cdot \sum_{k=1}^{m_i} b_i^{(k)} \cdot a_i^{(k)} \cdot (m_i - 1)! \cdot H_i(\alpha_i)\,.$$

Hieraus folgt mit $b_i^{(j)} = a_i^{(j)} = 0$ für jedes $j \in \{1, \ldots, m_i - 1\}$ und $b_i^{(m_i)} \cdot a_i^{(m_i)} = d_i$ nach (3.5) die Aussage (3.6). Eine Grenzwertbetrachtung für $\beta_i^{(j)} \to 0$ liefert unter den Voraussetzungen des Satzes 2 für beide Folgerungen aus 3.2.5 ebenfalls (3.6).

(iii) Beide Teile von Folgerung 1 verallgemeinern die Falksche ECP-Transformation für einfache Knoten.

(iv) Die Folgerung 1 (ii) verallgemeinert die wohlbekannte Aussage über die Frobeniussche oder Günthersche Begleitmatrix. Ist nämlich $m = 1, \alpha_1 = 0, N = m_1, \beta_1^{(1)} = \beta_1^{(2)} = \beta_1^{(3)} = \ldots = \beta_1^{(N-1)} = 1$ und $f_N = 1$, so folgt aus (3.18) wegen $H_1(x) = 1$ (für alle $x \in \mathbb{K}$)

$$\forall j \in \{1, \ldots, N\}: d_1^{(j)} = \partial^{j-1} f(0)(j-1)!\,.$$

Für diese Daten liefert Folgerung 1(i)

$$c_1^{(1)} = f(0) \quad \text{und} \quad c_1^{(j)} = \frac{\partial^{j-1} f(0)}{(j-1)!} - \frac{\partial^{j-2} f(0)}{(j-2)!}, \quad \text{für} \quad j \in \{2, 3, \ldots, N\}$$

als Verallgemeinerung der Falkschen Begleitmatrix zur ECP-Transformation. Die Determinanten beider Begleitmatrizen lassen sich auch mit elementaren Spalten- und Zeilenumformungen ineinander überführen.

(v) Das hier am Beispiel der Bidiagonalmatrizen vorgestellte Vorgehen kann allgemein als Konstruktionsprinzip für Begleitmatrizen aufgefaßt werden:

„Zunächst wird eine Matrix (oder eine Klasse von Matrizen) $A \in \mathbb{K}^{N \times N}$ betrachtet, deren Inverse $(A - x \cdot I_N)^{-1}$ für fast alle $x \in \mathbb{K}$ explizit bekannt ist. Für $a, b \in \mathbb{K}^N$ wird $\det [A - x \cdot I_N - a \cdot b^T]$ analog (2.12) berechnet, indem zunächst $(A - x \cdot I_N)^{-1}$ ausgeklammert wird. Dieses liefert eine Summendarstellung ähnlich der aus Lemma 1. Durch geeignete Interpolationsbedingungen (etwa nach Satz 3) wird diese Summendarstellung in solche Gleichungen überführt, die verschiedene Rekursionsbeziehungen für die beteiligten Größen erzeugen. Wenn eine rekursive Auflösung und Berechnung dieser Größen möglich ist, dann gilt

$$f(x) = \det [A - x \cdot I_N - a \cdot b]^T \quad \text{für alle} \quad x \in \mathbb{K}\,."$$

Beispiel 4: Es seien die beiden Teile von Folgerung 1 für das oben gebrachte numerische Beispiel betrachtet. Wenn man $\beta_2^{(1)} := \beta \in \mathbb{K} \setminus \{0\}$ schreibt, dann folgt

$$c_1^{(1)} = 0\,, \qquad c_1^{(2)} = a\,, \qquad c_2^{(1)} = \frac{1}{\beta}\,, \qquad c_2^{(2)} = -\frac{1+\beta}{\beta}$$

sowie

$$d_1^{(1)} = 0\,, \qquad d_1^{(2)} = a\,, \qquad d_2^{(1)} = \frac{1}{\beta}\,, \qquad d_2^{(2)} = -1\,.$$

Damit ergeben sich die speziellen Eigenwertprobleme zu den Matrizen

$$\begin{pmatrix} 0 & & & \\ & 0 & & \\ & & 1 & \beta \\ & & & 1 \end{pmatrix} - \begin{pmatrix} 1 \\ 1 \\ 1 \\ 1 \end{pmatrix} \left(0, a, \frac{1}{\beta}, -\frac{1+\beta}{\beta}\right) \tag{i}$$

und

$$\begin{pmatrix} 0 & & & \\ & 0 & & \\ & & 1 & \beta \\ & & & 1 \end{pmatrix} - \begin{pmatrix} 0 \\ 1 \\ 0 \\ 1 \end{pmatrix} \left(0, a, \frac{1}{\beta}, -1\right). \tag{ii}$$

Da $\beta_1^{(1)} \in \mathbb{K} \setminus \{0\}$ nicht in einem Nenner vorkommt, kann auch $\beta_1^{(1)} = 0$ gesetzt werden.

Tabelle 3. Zahlenwerte zum Beispiel 4 für die Matrix (i)

$a = 1.0000000000\text{E}+000$	$a = 1.0000000000\text{E}-001$	$a = 1.0000000000\text{E}-002$
−1.000000000000E+000	−1.000000000000E−001	−1.000000000000E−002
−2.758620689655E−001	−2.609742544916E−002	−3.255945191264E−003
−4.658015330026E−001	−3.446941100681E−002	−3.344469135977E−003
−4.655712319159E−001	−3.446910730266E−002	−3.344469135527E−003
−4.655712318768E−001	−3.446910730266E−002	−3.344469135527E−003
−4.655712318768E−001	−3.446910730266E−002	−3.344469135527E−003
−4.655712318768E−001	−3.446910730266E−002	−3.344469135527E−003
−4.655712318768E−001	−3.446910730266E−002	−3.344469135527E−003
−4.655712318768E−001	−3.446910730266E−002	−3.344469135527E−003
−4.655712318768E−001	−3.446910730266E−002	−3.344469135527E−003
−4.655712318768E−001	−3.446910730266E−002	−3.344469135527E−003

Für die Matrizen aus (i) und (ii) soll jeweils der kleinste Eigenwert für verschiedene Werte von a berechnet werden. Bei der Berechnung der Eigenwerte einer dyadisch gestörten Bidiagonalmatrix

$$\left(\operatorname{diag}(J_1, \ldots, J_m) - \begin{pmatrix} a_1 \\ \vdots \\ a_m \end{pmatrix} \cdot (b_1^T, \ldots, b_m^T)\right) \tag{3.20}$$

Tabell 4. Zahlenwerte zum Beispiel 4 für die Matrix (ii)

$a = 1.0000000000\text{E}+000$	$a = 1.0000000000\text{E}-001$	$a = 1.0000000000\text{E}-002$
$-1.000000000000\text{E}+000$	$-1.000000000000\text{E}-001$	$-1.000000000000\text{E}-002$
$-3.076923076923\text{E}-001$	$-2.678675188844\text{E}-002$	$-3.256812750419\text{E}-003$
$-4.644571762325\text{E}-001$	$-3.446906050958\text{E}-002$	$-3.344469135470\text{E}-003$
$-4.655712318854\text{E}-001$	$-3.446910730266\text{E}-002$	$-3.344469135527\text{E}-003$
$-4.655712318768\text{E}-001$	$-3.446910730266\text{E}-002$	$-3.344469135527\text{E}-003$
$-4.655712318768\text{E}-001$	$-3.4469107930266\text{E}-002$	$-3.344469135527\text{E}-003$
$-4.655712318768\text{E}-001$	$-3.446910730266\text{E}-002$	$-3.344469135527\text{E}-003$
$-4.655712318768\text{E}-001$	$-3.446910730266\text{E}-002$	$-3.344469135527\text{E}-003$
$-4.655712318768\text{E}-001$	$-3.446910730266\text{E}-002$	$-3.344469135527\text{E}-003$
$-4.655712318768\text{E}-001$	$-3.446910730266\text{E}-002$	$-3.344469135527\text{E}-003$
$-4.655712318768\text{E}-001$	$-3.446910730266\text{E}-002$	$-3.344469135527\text{E}-003$

mit der inversen Iteration nach WIELANDT muß i. a. mit mehreren Links- und Rechtseigenvektornäherungen gearbeitet werden [1]. Dabei müssen mehrere Gleichungssysteme mit der Koeffizientenmatrix (3.20) gelöst werden, was bei Verwendung der Formel von SHERMAN-MORRISON [1, 5] leicht möglich ist. Zwar ist

$$\text{diag}\,(J_1, \ldots, J_m)^{-1} = \text{diag}\,(J_1^{-1}, \ldots, J_m^{-1})$$

und deshalb explizit durch (3.10) gegeben, für die Lösung der Gleichungssysteme ist aber die Berechnung durch Rückwärts- bzw. Vorwärtssubstitution mit nur linearem Aufwand möglich. Deshalb erscheint hier die Verwendung von progressiven Schifts (neuer Schift in jeder Iteration [1]) sinnvoll.

Im vorliegenden Beispiel wurden in den Matrizen (i) und (ii) jeweils die erste Zeile und Spalte gestrichen, $\beta = 1$ gesetzt und als Startvektoren für die Eigenvektornäherungen $(1, 1, 1)$ bzw. $(1, 1, 1)^T$ verwendet. Für verschiedene Parameter a sind die Ergebnisse in den Tabellen 3 und 4 aufgeführt. Der Startschift wurde wie im numerischen Beispiel zu $-a$ gewählt.

Die Ergebnisse sind mit denen der zweiten Methode aus Tabelle 2 vergleichbar.

3.3 Matrizenpolynome

Nichtlineare Eigenwertprobleme können durch sogenannte Matrizenpolynome approximiert werden. Das sind Polynome mit quadratischen Matrizen als Koeffizienten.

Definition 3: Für $n, k \in \mathbb{N}$ sei $A_0, \ldots, A_k \in \mathbb{K}^{n\times n}$ und

$$f: \mathbb{K} \to \mathbb{K},\ x \mapsto \det\,(A_0 + x \cdot A_1 + \ldots + x^k \cdot A_k)\,. \tag{3.21}$$

f ist ein (gewöhnliches) Polynom vom Grad kleiner oder gleich $N := n \cdot k \in \mathbb{N}$ und besitze eine Darstellung

$$\forall x \in \mathbb{K}: f(x) = f_0 + f_1 \cdot x + \ldots + f_N \cdot x^N \tag{3.22}$$

mit Koeffizienten $f_0, \ldots, f_N \in \mathbb{K}$.

Lemma 2: *Für $j \in \{0, 1, \ldots, N\}$ gilt*

$$f_j = \sum_{\substack{l_1, l_2, \ldots, l_n \in \{0,1,\ldots,k\} \\ \sum_{i=1}^{n} l_i = j}} \det\left(A_{l_1}\begin{pmatrix}1\\1 \ldots n\end{pmatrix}, A_{l_2}\begin{pmatrix}2\\1 \ldots n\end{pmatrix}, \ldots, A_{l_n}\begin{pmatrix}n\\1 \ldots n\end{pmatrix}\right). \tag{3.23}$$

Beweis: Mit P sei die Menge aller Permutationen der Menge $\{1, 2, \ldots, n\}$ bezeichnet. Eine Entwicklung der Determinante aus (3.21) liefert für jedes $x \in \mathbb{K}$

$$f(x) = \sum_{\sigma\in P} \text{sign}\,(\sigma) \cdot \prod_{\nu=1}^{n}\left(\sum_{\kappa=0}^{k} x^\kappa \cdot A_\kappa\begin{pmatrix}\sigma(\nu)\\ \nu\end{pmatrix}\right) = \sum_{\sigma\in P} \text{sign}\,(\sigma) \cdot \sum_{l_1, l_2, \ldots, l_n \in \{0,1,\ldots,k\}} \prod_{\nu=1}^{n} x^{l_\nu} \cdot A_{l_\nu}\begin{pmatrix}\sigma(\nu)\\ \nu\end{pmatrix} =$$

$$= \sum_{l_1, l_2, \ldots, l_n \in \{0,1,\ldots,k\}} x^{(l_1+l_2+\ldots+l_n)} \cdot \sum_{\sigma\in P} \text{sign}\,(\sigma) \cdot \prod_{\nu=1}^{n} A_{l_\nu}\begin{pmatrix}\sigma(\nu)\\ \nu\end{pmatrix}.$$

Durch Koeffizientenvergleich mit (3.22) und Rückentwicklung der Determinante folgt (3.23). ∎

Bemerkung 4: Mit Lemma 2 ist der Hauptkoeffizient $f_N = \det A_k$. Wenn die führende Koeffizientenmatrix singulär ist, dann können $f_{N-1}, f_{N-2}, \ldots$ nach Lemma 2 berechnet werden, bis der genaue Grad von f feststeht. Zusätzlich lassen sich Ableitungen von f berechnen. Danach kann die ECP-Transformation nach den Abschnitten 3.1 und 3.2 Anwendung finden. Die hierdurch festgelegte Transformation wurde von FALK in [1] für den Fall einfacher Knoten angegeben. Der Vorteil der ECP-Transformation, etwa im Vergleich zur Frobeniusschen oder Güntherschen Begleitmatrix ist der, daß nur Funktionswerte (bzw. wenige Ableitungen) von f benötigt werden, die lediglich die Berechnung einer (bzw. weniger) Determinanten erfordern. Die Berechnung aller Koeffizienten mit Lemma 2 wäre deutlich aufwendiger.

Aufwandsberechnung: Zur ECP-Transformation mit einfachen Knoten ist die Berechnung von $n \cdot k + 1$ n-dimensionalen Determinanten notwendig. Bei Verwendung des Gaußschen Algorithmus zu deren Berechnung sind etwa $\frac{1}{3} \cdot k \cdot n^4 + O(n^3)$ Operationen nötig. Bei einem speziellen Eigenwertproblem kann die Koeffizientenmatrix vorher einmalig auf obere Hessenberg-Form transformiert werden, so daß insgesamt $\frac{5}{6} \cdot n^3 + k \cdot n \cdot O(n^3) + O(n^3)$ Operationen benötigt werden. Für symmetrische Matrizen liegt der Aufwand bei $\frac{2}{3} \cdot n^3 + k \cdot n \cdot O(n^2) + O(n^2)$ Operationen, wenn vorher auf eine symmetrische Tridiagonalform transformiert wurde.

3.4 Fehlerbetrachtungen

Alle in dieser Arbeit formulierten Spezialfälle für die Transformation auf ein spezielles Eigenwertproblem mit einer dyadisch gestörten Bidiagonalmatrix haben die Eigenschaft, daß die Störung in der Vielfachheit verschwindet, in

der der dazugehörige Knoten Eigenwert ist. Wenn die gewählten Knoten gute Näherungen für die Eigenwerte sind, dann wird man kleine Störungen erwarten können. Zur Fehlerabschätzung für die Knoten (als Näherung für die Eigenwerte des Problemes) bietet sich im Falle kleiner dyadischer Störung der wohlbekannte Satz von Gerschgorin an. Dabei kann die ECP-transformierte Matrix (3.20) vor der Fehlerbetrachtung noch einer Ähnlichkeitstransformation unterzogen werden. Ein interessantes Beispiel dieser Vorgehensweise soll in diesem Abschnitt vorgestellt werden.

Definition 4: Es gelten die Notationen und Vereinbarungen der vorherigen Kapitel. Für alle $i \in \{1, 2, \dots, m\}$ bezeichne

$$a(i) := \sum_{j=1}^{m_j} |a_i^{(j)}|\,, \qquad b(i) := \max\{|b_i^{(j)}| \mid j \in \{1, 2, \dots, m_i\}\}\,,$$

$$\beta(i) := \max\{|\beta_i^{(j)}| \mid j \in \{1, 2, \dots, m_i - 1\}\}\,, \qquad D(i) := \min\{|\alpha_i - \alpha_j| \mid j \in \{1, 2, \dots, m\} \setminus \{i\}\}\,,$$

$$\varepsilon_i := \max\left\{\frac{\beta(i) + \beta(j)}{|\alpha_i - \alpha_j|} \,\middle|\, j \in \{1, 2, \dots, m\} \setminus \{i\}\right\}, \qquad \sigma_i := \sum_{\substack{j=1 \\ j \neq i}}^{m} \frac{a(j) \cdot b(j)}{|\alpha_i - \alpha_j|}\,, \qquad \delta_i := \frac{a(i) \cdot b(i)}{D(i)}\,.$$

Satz 4: *Für ein* $i \in \{1, 2, \dots, m\}$ *gelte*

$$\varepsilon_i + \sigma_i + \delta_i + 2 \cdot \sqrt{\sigma_i \cdot \delta_i} < 1\,. \tag{3.24}$$

Falls $\delta_i = 0$, *dann ist* α_i *ein* m_i*-facher Eigenwert der Matrix* (3.20). *Andernfalls ist*

$$t_i := \frac{1 - \delta_i - \sigma_i - \varepsilon_i + \sqrt{(1 - \delta_i - \sigma_i - \varepsilon_i)^2 - 4 \cdot \delta_i \cdot \sigma_i}}{2 \cdot \delta_i} > 0\,,$$

und in der abgeschlossenen Kreisscheibe um den Mittelpunkt α_i *mit dem Radius*

$$R_i := \beta(i) + a(i) \cdot b(i) + \sigma_i \cdot D(i)/t_i$$

liegen wenigstens m_i *Eigenwerte der Matrix* (3.20).

Beweis: Für $\delta_i = 0$ verschwindet die Störung für den Knoten α_i, und das Eigenwertproblem zerfällt (auch wenn (3.24) nicht gilt). Andernfalls ist δ_i positiv. Wegen (3.24) ist dann die Diskreminante zur Berechnung von t_i positiv, und es gilt

$$0 < \frac{1 - \varepsilon_i - \sigma_i - \delta_i}{2 \cdot \delta_i} < t_i\,.$$

Eine einfache Diskussion zeigt, daß das Polynom

$$h: \left]\frac{1 - \varepsilon_i - \sigma_i - \delta_i}{2 \cdot \delta_i}, t_i\right[\to \mathbb{R}\,, \qquad t \to \delta_i \cdot t^2 + t \cdot (\varepsilon_i + \sigma_i + \delta_i - 1) + \sigma_i$$

im betrachteten Intervall negativ ist. Für ein $t \in \left]\frac{1 - \varepsilon_i - \sigma_i - \delta_i}{2 \cdot \delta_i}, t_i\right[$ und jedes $j \in \{1, 2, \dots, m\}$ werde

$$\gamma_j(t) := \begin{cases} \dfrac{b(j)|}{|\alpha_i - \alpha_j|}\,, & \text{falls} \quad j \neq i \\ \dfrac{t \cdot b(i)}{D(i)}\,, & \text{falls} \quad j = i \end{cases}$$

gesetzt.

Im folgenden sei zusätzlich vorausgesetzt, daß $b(j) \neq 0$ für alle $j \in \{1, 2, \dots, m\} \setminus \{i\}$ gilt. Wenn $b(j) = 0$ für ein $j \in \{1, 2, \dots, m\} \setminus \{i\}$ ist, dann zerfällt das Eigenwertproblem, und man kann in der Matrix (3.20) die Zeilen und Spalten in den Positionen der $a_j^{(1)}, \dots, a_j^{(m_j)}$ und $b_j^{(1)}, \dots, b_j^{(m_j)}$ streichen. Für die so reduzierte Matrix liefert die nachfolgende Rechnung sogar etwas schärfere Abschätzungen als die direkte Anwendung des Satzes.

Dann hat die Matrix (3.20) dieselben Eigenwerte, wie die Matrix in (3.25),

$$\operatorname{diag}(J_1, \dots, J_m) - \begin{pmatrix} \gamma_1(t) \cdot a_1 \\ \vdots \\ \gamma_m(t) \cdot a_m \end{pmatrix} \cdot \left(\gamma_1(t)^{-1} \cdot b_1^{\mathrm{T}}, \dots, \gamma_m(t)^{-1} \cdot b_m^{\mathrm{T}}\right)\,, \tag{3.25}$$

die durch Ähnlichkeitstransformation mit der Matrix $\operatorname{diag}(\gamma_1(t) \cdot I_{m_1}, \dots, \gamma_m(t) \cdot I_{m_m})$ aus (3.20) hervorgeht. Für jedes $j \in \{1, 2, \dots, m\}$ liegen alle m_j Gerschgorinkreise um die Mittelpunkte $\alpha_j - a_j^{(1)} \cdot b_j^{(1)}, \dots, \alpha_j - a_j^{(m_j)} \cdot b_j^{(m_j)}$ zu den Spalten aus (3.25) in der abgeschlossenen Kreisscheibe mit dem Mittelpunkt α_j und dem Radius

$$r_j(t) := \beta(j) + \gamma_j(t)^{-1} \cdot b(j) \cdot \sum_{\nu=1}^{m} \gamma_\nu(t) \cdot a(\nu)\,.$$

Für jedes $j \in \{1, 2, \dots, m\} \setminus \{i\}$ gilt mit obigen Bezeichnungen wegen $D(i) \leqq |\alpha_i - \alpha_j|$ und der erwähnten Eigenschaft von h

$$r_i(t) + r_j(t) \leqq |\alpha_i - \alpha_j| \cdot \frac{t + h(t)}{t} < |\alpha_i - \alpha_j|\,.$$

Folglich liegen die großen Kreise um α_j mit den Radien $r_j(t)$ für $j \in \{1, 2, \dots, m\} \setminus \{i\}$ disjunkt zu dem um α_i mit dem Radius $r_i(t)$. Nach dem wohlbekannten Gerschgorinschen Kreisesatz liegen genau m_i Eigenwerte in der abgeschlossenen Kreisscheibe um α_i mit dem Radius $r_i(t)$. Diese Aussage gilt für jedes $t \in \left]\frac{1-\varepsilon_i-\sigma_i-\delta_i}{2\cdot\delta_i}, t_i\right[$ so daß im Kreis um α_i mit dem Radius $r_i(t_i) = R_i$ wenigstens m_i Eigenwerte der Matrix (3.25) liegen. ∎

Bemerkung 5:

(i) Eine ähnliche Abschätzung erhält man, wenn der Gerschgorinsche Kreisesatz auf die Zeilen der Matrix (3.25) angewendet wird.

(ii) Für den Fall einfacher Knoten reduzieren sich die Bedingungen des Satzes 3. In diesem Fall kann der Mittelpunkt zum i-ten Gerschgorinkreis genauer mit $\alpha_i - a_i \cdot b(i)$ abgeschätzt werden.

(iii) In [8] hat BÖRSCH-SUPAN Fehlerabschätzungen für Polynom-Nullstellen dadurch gewonnen, daß er die Funktionswerte der bekannten Näherungen berechnet, für die Daten eine Lagrange-Interpolation durchführt und damit aus dem Satz von Rouché Fehlerabschätzungen herleitet. Für den Fall einfacher Knoten reduziert sich der Satz 4 im wesentlichen auf den Satz 1 aus [8]. Für mehrfache Knoten besitzt der Satz 2 aus [8] eine ähnliche Struktur wie Satz 4. Dieser Satz 2 aus [8] kann auch zur Fehlerabschätzungen der Spezialfälle dieser Arbeit herangezogen werden.

4. Der Spezialfall einfacher Knoten

Für ein Polynom f in $\mathbb{K}$ vom Grad $n \in \mathbb{N}$ mit Hauptkoeffizienten $f_n \in \mathbb{K} \setminus \{0\}$ soll in diesem Abschnitt der Spezialfall einfacher Knoten

$$\alpha_1, \dots, \alpha_n \in \mathbb{K} \quad \text{paarweise verschieden} \tag{4.1}$$

genauer beschrieben werden. Nach Satz 2 folgt mit

$$d_i := f(\alpha_i)/(f_n \cdot \Pi_i(\alpha_i)), \qquad i \in \{1, \dots, n\}, \tag{4.2}$$

für alle $x \in \mathbb{K}$:

$$f(x) = (-1)^n \cdot f_n \cdot \det\left(\operatorname{diag}(\alpha_1, \dots, \alpha_n) - \alpha \cdot (d_1, \dots, d_n) - x \cdot I_n\right). \tag{4.3}$$

Wenn man die Knoten $\alpha_1, \dots, \alpha_n$ als Näherungen für die Nullstellen von f auffaßt, dann ist für jedes $i \in \{1, \dots, n\}$ d_i ein (gewichteter) Defekt $f(\alpha_i)$, also ein Maß für den Fehler. Diese Eigenschaft der Defekte $d_1, \dots, d_n$ soll zunächst konkretisiert und anschließend angewendet werden.

4.1 Verschiedene Deutungen der Defekte

Bei kleinen Defekten kann der Gerschgorinsche Kreisesatz auf die Matrix in (4.3) angewendet werden. Dann liegt (für $i \in \{1, \dots, n\}$) der Kreis um $\alpha_i - d_i$ in der (abgeschlossenen) Kreisscheibe mit dem Mittelpunkt α_i und dem Radius $n \cdot |d_i|$. Wenn die Defekte betragsmäßig klein sind, dann erhält man Fehlerabschätzungen für einen oder mehrere Knoten. Dieses wurde bereits im Beweis zu Satz 4 benutzt, der wie die Sätze in [8] ebenfalls Fehlerschranken liefert. Im Spezialfall einfacher Knoten läßt sich die Abschätzung auf $\alpha_i - d_i$ anwenden und liefert eine strengere Fehlerabschätzung für die Näherung $\alpha_i - d_i$. Die Defekte stellen tatsächlich ein linearisiertes Fehlermaß dar. Ist nämlich μ eine Nullstelle von f in der Nähe von α_i, $i \in \{1, \dots, n\}$, so gilt (sofern der Nenner nicht verschwindet)

$$\mu = \alpha_i - \frac{d_i}{1 + \sum\limits_{\substack{j=1 \\ j \neq 1}}^{n} \frac{d_i + d_j}{\alpha_i - \alpha_j}} + O(|\mu - \alpha_i|^2), \qquad |\mu - \alpha_i| \to 0. \tag{4.4}$$

Diese Darstellung folgt aus einer Entwicklung von f an der Stelle μ, wie sie vom klassischen Newton-Verfahren bekannt ist, wenn man (3.7) ausnutzt, um $f(\alpha_i)$ und $f'(\alpha_i)$ zu berechnen; vgl. auch [2, (11)]. Für kleine Defekte folgt

$$\mu \approx \alpha_i - d_i. \tag{4.5}$$

An dieser Stelle könnte man definieren, wann die Defekte $d_1, \dots, d_n$ als klein zu bezeichnen wären; nämlich dann, wenn die Näherung α_i so gut ist, daß der quadratische Anteil in (4.4) und die Summe im Nenner gegenüber 1 vernachlässigbar sind. Die Anwendung von (4.5) ist bei einem reellen Polynom mit hinreichend verschiedenen einfachen reellen Nullstellen schon in [7] und (auch für den komplexen Fall) in [9] beschrieben worden. (Für diesen Hinweis danken wir Herrn W. BÖRSCH-SUPAN.) Dort wird ein Verfahren vorgestellt, das in unserer Notation (bei Vernachlässigung der Indizes für den Iterationsschritt) simultan alle Knoten durch den Übergang

$$(\alpha_1, \dots, \alpha_n) \mapsto (\alpha_1 - d_1, \dots, \alpha_n - d_n)$$

verbessert. Für dieses Verfahren wird quadratische Konvergenz garantiert, sofern die Knoten genügend dicht an den einfachen Nullstellen liegen.

Diese Eigenschaft der Defekte motiviert die Deutung der Zahlen $d_1, \dots, d_n$ als einen Anhalt für den tatsächlichen Fehler der Knoten (4.1) und läßt eine Steuerung im Programmablauf zu, wie sie etwa in [1], [2] und in 4.4 beschrieben ist.

4.2 Verbesserung der Näherungen

Mit der ECP-Transformation, d. h. mit der Berechnung der Defekte in (4.2) und Betrachtung des linearen, speziellen Eigenwertproblemes in (4.3) mit der Koeffizientenmatrix

$$\operatorname{diag}(\alpha_1, \dots, \alpha_n) - e \cdot (d_1, \dots, d_n), \tag{4.6}$$

können Eigenwertalgorithmen zur Bestimmung der Nullstellen von f eingesetzt werden. FALK schlägt in [1] insbesondere seinen Algorithmus BONAVENTURA vor, vgl. auch [2]. In dieser Arbeit wurde bei Verwendung von Links- und Rechtseigenvektornäherungen und dem Rayleigh-Quotienten nach WIELANDT invers iteriert. Bei kleinen Defekten ist (4.6) diagonaldominant, so daß das Verfahren bei einfachen Nullstellen und hinreichend guten Startwerten rasch konvergiert. Dabei kann die Matrix (4.6) mit der SHERMAN-MORRISON-Formel [5] leicht invertiert und sogar ein progressiver Schift verwendet werden.

Man kann leicht zeigen, daß für einen Eigenwert $\mu \in \mathbb{K} \setminus \{\alpha_1, \dots, \alpha_n\}$ von (4.6)

$$\left(\frac{d_1}{\alpha_1 - \mu}, \dots, \frac{d_n}{\alpha_n - \mu}\right) \quad \text{bzw.} \quad \left(\frac{1}{\alpha_1 - \mu}, \dots, \frac{1}{\alpha_n - \mu}\right)^T$$

ein Links- bzw. Rechtseigenvektor zu μ ist. Diese Informationen können bei der Wahl der Startwerte einfließen.

4.3 Aktualisierung und Deflation

Es sei das Polynom f fixiert und

$$\Delta: \{(\alpha_1, \dots, \alpha_n) \in \mathbb{K}^n \mid \alpha_1, \dots, \alpha_n \text{ paarweise verschieden}\} \to \mathbb{K}^n$$

die Abbildung, die den Knoten $\alpha_1, \dots, \alpha_n$ die Defekte $d_1, \dots, d_n$ gemäß (4.2) zuordnet. Wenn $\Delta(\alpha_1, \dots, \alpha_i, \dots, \alpha_n) =: (d_1, \dots, d_i, \dots, d_n)$ bekannt ist und der i-te Knoten α_i gegen einen Wert $\lambda \in \mathbb{K} \setminus \{\alpha_1, \dots, \alpha_n\}$ ausgetauscht werden soll ($i \in \{1, \dots, n\}$), dann gilt für $\Delta(\alpha_1, \dots, \lambda, \dots, \alpha_n) =: (\hat{d}_1, \dots, \hat{d}_n)$

$$\forall j \in \{1, \dots, n\} \setminus \{i\}: \hat{d}_j = d_j \cdot \frac{\alpha_j - \alpha_i}{\alpha_j - \lambda} \tag{4.7}$$

und

$$\hat{d}_i = \lambda - \alpha_i + d_i + \sum_{\substack{k=1 \\ k \neq i}}^{n} d_k \cdot \frac{\alpha_i - \lambda}{\alpha_k - \lambda}. \tag{4.8}$$

Die Gleichungen (4.7) folgen unmittelbar aus (4.2); (4.8) kann aus (4.3) und (2.13) zur Berechnung von $f(\lambda)$ oder aus der Invarianz der Spur der Matrix (4.6) gefolgert werden.

Wenn λ eine exakte Nullstelle ist, dann folgt aus $f(\lambda) = 0$ auch $\hat{d}_i = 0$, und das Eigenwertproblem mit der neuen Matrix (4.6) zerfällt. Das Problem reduziert sich auf die weitere Betrachtung der Matrix

$$\operatorname{diag}(\alpha_1, \dots, \alpha_{i-1}, \alpha_{i+1}, \dots, \alpha_n) - \varepsilon \cdot (d_1, \dots, d_{i-1}, d_{i+1}, \dots, d_n) \in \mathbb{K}^{(n-1)\times(n-1)}.$$

Dieser Deflationsalgorithmus ist genau der Spezialfall von FALKS VELOCITAS in [1], [2].

Bei der Anwendung dieser Deflation ist meistens ein guter Schift $\lambda \in \mathbb{K} \setminus \{\alpha_1, \dots, \alpha_n\}$ gegeben, mit dem das Problem verkleinert werden soll. Wenn man zum Austausch gegen λ einen Knoten α_i mit

$$|\alpha_i - \lambda| = \min\{|\alpha_j - \lambda| : j \in \{1, \dots, n\}\} \tag{4.9}$$

wählt, dann vergrößern sich die Beträge der übrigen Knoten höchstens um den Faktor 2.

Diese Aussage gilt i. a. nicht, wenn die Deflation mehrfach ausgeführt wird, in (4.9) einige Defekte bereits vernachlässigt wurden, und die dazugehörigen Näherungen nicht mehr betrachtet werden.

4.4 Algorithmus

1. Wähle n paarweise verschiedene Startwerte $\alpha_1, \alpha_2, \dots, \alpha_n$.
2. Berechne $d_1, d_2, \dots, d_n$ nach (4.2).
3. Entscheide, ob Abbruch, wähle sonst $0 < \varepsilon_1 < \varepsilon_2 < \varepsilon_3$ für die gewünschte Genauigkeit in diesem Iterationszykel, wähle **Maxstep** und **Maxsteps**. Setze **Steps** $:= 0$, $\hat{\mathrm{N}} :=$ n.
4. Bestimme $\mathrm{N} \in \{1, 2, \dots, n\}$ und $\mathrm{K} \in \{1, 2, \dots, \mathrm{N}\}$ und vertausche die Indizes von $(\alpha_1, d_1), (\alpha_2, d_2), \dots, (\alpha_n, d_n)$, so daß gilt $\forall i \in \{1, 2, \dots, \mathrm{K}\}: \varepsilon_1 < |d_i| \leqq \varepsilon_2$ und $\forall i \in \{\mathrm{N}+1, \dots, n\}: |d_i| \leqq \varepsilon_1$.

5a. Wenn $\mathrm{N} \leqq 2$ berechne die Eigenwerte des N-dimensionalen Eigenwertproblemes und setze $\alpha_1, \dots, \alpha^{\mathrm{N}}$ gleich diesen Werten. Weiter mit 2.

5b. Wenn $\mathrm{N} < \mathrm{K}$, berechne $M := \max\{|d_i| : i \in \{1, 2, \dots, \mathrm{N}\}\}$.

5b1. Wenn $M \leqq \varepsilon_3$, weiter mit 2.

5b2. Andernfalls

wenn **Steps** $\leqq$ **Maxsteps**

wenn $\mathrm{N} < \hat{\mathrm{N}}$ **Steps** $:= 0$, sonst **Steps** $:=$ **Steps** $+ 1$

$\hat{\mathrm{N}} := \mathrm{N}$, $\mathrm{K} := 0$. Weiter mit 6.

sonst weiter mit 2.

5c. Wenn $\mathrm{N} \geqq \mathrm{K}$, weiter mit 6.

6. Wähle Startwerte

Tausche Indizes, so daß $|d_{\mathrm{N}}| = \min\{|d_i| : i \in \{\mathrm{K}+1, \dots, \mathrm{N}\}\}$.

Wähle Startwert ‚in der Nähe von' α_{N}. Wähle Startvektoren, weiter mit 7.

7. Verbessere den Startwert λ als Näherung für einen Eigenwert des speziellen N-dimensionalen Eigenwertproblems mit der Matrix

$$\operatorname{diag}(\alpha_1, \dots, \alpha_{\mathrm{N}}) - e \cdot (d_1, \dots, d_{\mathrm{N}}) \tag{4.6'}$$

höchstens **Maxstep** mal. Höre auf, falls die Iteration zum Stehen gekommen ist, oder ähnliche Gründe einen Abbruch motivieren. Die letzte Näherung heiße wieder λ.

8. Wähle $j \in \{1, \dots, N\}$ mit $|\lambda - \alpha_j| = \min\{|\lambda - \alpha_i| : i \in \{1, \dots, N\}\}$.
Berechne

$$d := \lambda - \alpha_j + d_j + \sum_{\substack{i=1 \\ i \neq j}}^{N} d_i \cdot \frac{\alpha_j - \lambda}{\alpha_i - \lambda}. \tag{4.8'}$$

Vertausche die Indizes von j und N. Weiter mit 9.

9. Wenn $|d| < \frac{|d_N|}{2 \cdot N}$, dann setze $\forall i \in \{1, \dots, N-1\}: d_i := d_i \cdot \frac{\alpha_i - \alpha_N}{\alpha_i - \lambda}$.
$\alpha_N := \lambda$, $d_N := d$. Weiter mit 10.

10. Wenn $|d| \leqq \varepsilon_1$, dann setze $N := N - 1$.
Andernfalls setze $K := K + 1$ und vertausche die Indizes von (α_j, d_j) und (α_K, d_K). Weiter mit 5.

Bemerkungen zu den Beispielrechnungen: In den Beispielen des folgenden Abschnittes wurde eine Genauigkeitsschranke δ gewählt und $\varepsilon_1 := \delta * 10^{-10}$, $\varepsilon_2 := \delta * 10^{-5}$ sowie $\varepsilon_3 := \delta * 10^{0}$ gesetzt.

Bemerkung 6: Im Algorithmus 4.4 werden die Defekte durchgehend zur Steuerung eingesetzt, wie bereits in [1, 2] beschrieben. Dieses Vorgehen wird für kleine Defekte durch die verschiedenen Deutungen der Zahlen $d_1, \dots, d_n$ als Fehler motiviert. Bei mehrfachen Knoten und den im 3. Abschnitt vorgestellten Verallgemeinerungen ist diese einfache Deutung der Defekte i. a. nicht möglich. Eine andere Verallgemeinerung mit dieser Eigenschaft wäre daher wünschenswert.

Tabelle 5. Ausgangsmatrix, Startwerte und einige Näherungen zu Beispiel 5.1

Ausgangsmatrix

J	$a(J, J-1)$	$a(J, J)$	$a(J, J+1)$
1		−.100E2	−.400E1
2	.100E2	−.140E2	−.100E2
3	.180E2	−.140E2	−.180E2
4	.240E2	−.100E2	−.280E2
5	.280E2	−.200E1	−.400E2
6	.300E2	.100E2	−.540E2
7	.300E2	.260E2	−.700E2
8	.280E2	.460E2	−.880E2
9	.240E2	.700E2	−.108E3
10	.180E2	.980E2	−.130E3
11	.100E2	.130E3	

Startwerte und Anfangsdefekte

J	alpha (J)	$d(J)$
1	−.3333E2	−.7456911E−1
2	−.6666E1	.6824949E−3
3	.2000E2	−.1081577E−6
4	.4666E2	−.2219965E−6
5	.7333E2	.2097852E+0
6	.1000E3	−.6371504E+2
7	.1266E3	.2001035E+4
8	.1533E3	−.1725665E+5
9	.1800E3	.5674437E+5
10	.2066E3	−.7709314E+5
11	.2333E3	.3643796E+5

Näherungen und Defekte nach dem 1. Iterationszykel

J	alpha (J)	$d(J)$
1	0.5120446767153210600E−03	0.9843401E−03
2	0.1198863043762975700E+02	−0.1113593E+00
3	0.1439079781477693840E+02	−0.2724452E+02
4	0.1799503270296267490E+02	0.1407968E+02
5	0.2199999628968029340E+02	−0.2215336E−05
6	0.3783387911927682980E+02	−0.6000055E−02
7	0.3955215736625745390E+02	0.1287583E−02
8	0.4227894028698453610E+02	−0.2763739E−02
9	0.4569825685135989350E+02	0.1576909E+02
10	0.4757574997674394980E+02	−0.1768211E+03
11	0.4812296973905189650E+02	0.1717716E+03

Näherungen und Defekte nach dem 4. Iterationszykel

J	alpha (J)	$d(J)$
1	0.0000000000000000000E+00	0.0000000E+00
2	0.1200000000022502900E+02	0.3378796E−10
3	0.2200000000269699230E+02	0.2164158E−08
4	0.2999435003669922390E+02	−0.2860058E−02
5	0.3000560578166220080E+02	0.2809391E−02
6	0.3599917391716888910E+02	0.2336440E−02
7	0.3600128582148372660E+02	0.2826304E−02
8	0.3999850032619394820E+02	0.5005916E−02
9	0.4000245108611891000E+02	0.6832247E−03
10	0.4199850984506257840E+02	0.1438125E−02
11	0.4200165108466759990E+02	0.1599958E−03

Näherungen und Defekte nach dem 10. Iterationszykel

J	alpha (J)	$d(J)$
1	0.0000000000000000000E+00	0.0000000E+00
2	0.1199999999981579200E+02	−0.1053346E−10
3	0.2200000000040866150E+02	0.5776760E−09
4	0.2999961843958879730E+02	−0.4305049E−03
5	0.3000058001377944800E+02	0.5830280E−03
6	0.3599787123896872030E+02	−0.3050326E−02
7	0.3600091655310365720E+02	0.1113840E−02
8	0.3998803555296563420E+02	−0.2531836E−01
9	0.3999379576428548690E+02	0.6139636E−02
10	0.4199985562599810150E+02	−0.2031306E−03
11	0.4200806333408552010E+02	0.8238626E−02

5. Numerische Beispiele

Um einen Überblick zur Effektivität des in 4.4 vorgestellten Programmes zu erhalten, wurden verschiedene Eigenwertaufgaben betrachtet. Dabei wurden die Problemkreise von mehrfachen Nullstellen, ungünstige Startwerte und große Defekte, gute Startwerte und günstig verteilte Startwerte in reeller Rechnung untersucht.

5.1 Beispiel zu doppelten Eigenwerten kleiner Dimension

An der Testmatrix A aus [4],

$$a_{i,j} := \begin{cases} -[(2i+1)N + is - 2i^2] & \text{für } i = j, \\ (i+1)(N+s-i) & \text{für } j = i+1, \\ i(N-i+1) & \text{für } j = i-1, \\ 0 & \text{sonst}, \end{cases} \tag{5.1}$$

einer $(N+1)$-dimensionalen Tridiagonalmatrix mit den reellen Eigenwerten

$$(\lambda_j := -j(s+j+1) : j \in \{0, 1, \ldots, N\})$$

wurden für $N = 10$, $s = -14$ die Eigenwerte 0, 12, 22, 30, 30, 36, 36, 40, 40, 42, 42 berechnet. (In [2] wurde mit $s = 0$ die symmetrische Matrix untersucht.) Als Startwerte wurden für A die Gerschgorinkreise ausgewertet. Danach liegen alle Eigenwerte im Intervall $[-60, 260]$. Dieses Intervall wurde in $N + 2 = 12$ äquidistante Teile unterteilt, deren 11 Mittelpunkte die Startwerte definierten.

Bei dieser linearen Verteilung entstehen große Defekte, die wegen der verhältnismäßig kleinen Dimension hier mit $\delta = 10^{-10}$ in wenigen Schritten verkleinert werden können; man vgl. Tabelle 5. Nach 4 Iterationszyklen stagniert das Verfahren und liefert für die doppelten Nullstellen nur etwa 4 signifikante Ziffern. Die Werte sind aber von den übrigen getrennt, so daß Fehlerabschätzungen möglich sind, die eine Genauigkeit der letzten Näherungen von etwa $\pm$ 0.05 garantieren.

5.2 Beispiel zu einfachen Eigenwerten und schlechten Startwerten

Für die Testmatrix (5.1) wurde $N = 99$, $s = -10$ gewählt, so daß nur einfache reelle Eigenwerte zu erwarten sind. Die Startwerte wurden analog zum ersten Beispiel linear generiert und ergeben große Defekte.

Tabelle 6. Startwerte und einige Näherungen zu Beispiel 5.2

Startwerte und Anfangsdefekte

J	alpha (J)	$d(J)$
1	−0.10904950495049504800E+05	−0.7106094E+17
2	−0.10795900990099009500E+05	−0.1699797E+19
3	−0.10686851485148514300E+05	0.4407267E+18
4	−0.10577801980198019000E+05	0.1638131E+20
5	−0.10468752475247523800E+05	0.4188522E+19
40	−0.66520198019801982800E+04	−0.6130638E+16
60	−0.44710297029702969700E+04	0.4385815E+06
96	−0.54524752475247510100E+03	−0.1153700E−28
97	−0.43619801980198008100E+03	0.1466895E−30
98	−0.32714851485148506100E+03	0.1957897E−31
99	−0.21809900990098998400E+03	0.6936967E−33
100	−0.10904950495049493500E+03	−0.5726634E−34

Näherungen und Defekte nach dem 1. Iterationszykel

J	alpha (J)	$d(J)$
1	−0.10890016232125108200E+05	−0.4933365E−04
2	−0.10766986527660077600E+05	−0.1724472E+01
3	−0.10681825820495929300E+05	−0.1546535E−01
4	−0.10577801980198019000E+05	0.1830898E+02
5	−0.10485775685911901700E+05	−0.3733613E+02
40	−0.64500000000004583800E+04	0.1072595E−13
60	−0.41382128047131600400E+04	0.7364015E−14
96	−0.32714851485148506100E+03	0.2150251E−46
97	−0.21809900990098998400E+03	0.1146030E−47
98	−0.10904950495049493500E+03	−0.1899615E−48
99	0.00000000000000000000E+00	0.0000000E+00
100	0.42574851887699158600E+21	0.4257485E+21

Näherungen und Defekte nach dem 4. Iterationszykel

J	alpha (J)	$d(J)$
1	−0.14892324333223828600E+06	−0.4137411E+05
2	−0.29757998013649514200E+05	0.1126431E+04
3	−0.17818477628613480200E+05	−0.5192245E+03
4	−0.11621666287619917100E+05	0.1439273E+04
5	−0.10890000000000000000E+05	−0.1350283E−11
40	−0.55199999999999863600E+04	−0.1511968E−15
60	−0.30499999889460791600E+04	0.6577561E−15
96	0.26065164053366920600E+04	0.1011460E−06
97	0.10907469059950533800E+05	−0.1166136E−01
98	0.34663387098999519400E+05	0.1628782E+03
99	0.52050756544072639400E+05	−0.1322509E+04
100	0.45444122214252472600E+06	0.3475316E+06

Näherungen und Defekte nach dem 9. Iterationszykel

J	alpha (J)	$d(J)$
1	−0.10890000000000000000E+05	−0.1489479E−12
2	−0.10682000000000000000E+05	0.3506786E−13
3	−0.10476000000000000000E+05	−0.6495309E−13
4	−0.10272000000000000000E+05	0.9228253E−13
5	−0.10070000000000000000E+05	0.2460964E−13
40	−0.43920000000000000000E+04	0.7069183E−13
60	−0.25848598139236564700E+04	0.7728020E+05
96	0.22288279419647760700E+04	0.3805973E+05
97	0.23602188874312996600E+04	−0.6450222E+05
98	0.30227038593151414700E+04	0.1250035E+06
99	0.36763760356358902800E+04	−0.1593341E+07
100	0.37200772943496704100E+04	0.1500109E+07

Näherungen und Defekte nach dem 13. Iterationszykel

J	alpha (J)	$d(J)$
1	−0.10890000000000000000E+05	−0.7270695E−13
2	−0.10682000000000000000E+05	0.1655082E−13
3	−0.10476000000000000000E+05	−0.2957412E−13
4	−0.10272000000000000000E+05	0.4043669E−13
5	−0.10070000000000000000E+05	0.1035013E−13
40	−0.42600000000000000000E+04	−0.9179555E−15
60	−0.20400000000000000000E+04	−0.4113616E−14
96	−0.59999999999999730000E+02	0.1516629E−12
97	−0.41999999999999836600E+02	0.1652905E−13
98	−0.26000000000000071100E+02	−0.1390253E−13
99	−0.11999999999999964500E+02	0.1997535E−13
100	0.00000000000000000000E+00	0.0000000E+00

Bild 1. Verteilung der Defekte in logarithmischer Einteilung zum Beispiel 5.2

	Iterationszykel														
	0	1	2	3	4	5	6	7	8	9	10	11	12	13	14
Exp.															
21	1														
20	16	1													
19	9														
18	4														
17	4														
16	4														
15	3														
14	3														
13	2														
12	2														
11	1														
10	3														
9	2							2	1						
8	1							4							
7	1							2							
6	3		2					2		2					
5	1			3	1	1	1	1		5	2				
4	1				1	2	2	3		3	2				
3	2			2	5	2	3	3		2	2				
2	2	5	3	4	2	3	2	1		1	2	7	1		
1	1	6	2	1	1	2	3	5	1			4	1		
0	1	10	3	4	1	1	1	5	6			5	2		
−1	1	6	8	2	1	1		3	3		4	1	1		
−2	1	6	1	1	1	1	1	2	1				1		
−3	1	3	5					2	1			4			
−4	1	1		1				2	1	2		1	1		
−5		2	3		1	1	1		2	1	2	3			
−6	3		2	1				6	1	2	2	1			
−7	1	1	2	1	2			5		1	3	2			
−8			3	2	1	2	1	5		6	3		1		
−9	3		3	4	5	3	2	6		4	3	1			
−10	1		4	12	3	1	2	8		4	1			2	
−11	1		10	13	2	3	4	5		11	8	2	3	2	2
−12	1	1	15	11	17	11	9	6	4	9	11	6	11	4	3
−13	1	1	3	5	5	11	12	4		22	30		4	3	1
−14	1	11	5	8	9	8	9	13	10	18	16	33	43	52	55
−15	1	6	2	3	7	12	11		5	2	5	26	27	32	34
−16	1	5	2	2	7	5	8	1	5	1	3	3	3	4	4
−17	1	1	1		4	6	4	1	8	1					
−18		1	1	1	2	1	2		8	2					
−19	2	1	1	2	1	2	1	1	16						
−20		1	1	1	2	1	2	1	20						
−21	1	1	1		1	2	2		2						
−22	1	5	1	2	4	2	4		1						
−23					4	6	3		1						
−24	2	1	2	2	1		1		1						
−25	1	1	1	2		1	1		1						
−26		2	2	1	1	1									
−27	1	1	1	1	1	1	1								
−28	1	1	1				1								
−29	1		1	1	1	1	1								
−30		1		1	1	1									
−31	1	1	1				1								
−32	1	1	1	1	2	2	1								
−33		1					1								
−34	1		1	2	2	1	1								
−35	1	2	1			1									
−36			1	1											
−37		2		1											
−38			2												
−39															
−40		2													
−41		1													
−42		1													
−43		1													
−44															
−45		1													
−46		1													
−47		1													
−49		1													
−49		1													
<−50		1	1	1	1	1	1	1	1	1	1	1	1	1	1

Mit $\delta = 1.0$ können die meisten Näherungen in wenigen Schritten sehr gut verbessert werden, aber wenige sehr große Defekte verhindern eine rasche und gleichmäßige Konvergenz. Diese Konvergenzverzögerung ist weniger ein lokales Problem, denn in jedem Durchlauf für $K := 0$ (vgl. 4.4) konnte wenigstens 1 Defekt betragsmäßig kleiner ε_1 gemacht werden. Vielmehr handelt es sich hier um ein Verteilungsproblem. Dazu sind für jeden Iterationszyklus die Defekte d_i in Klassen nach dem Exponenten von $|d_i|$, d. h. nach dem ganzzahligen Anteil von $\log_{10}|d_i|$, eingeteilt worden. Die Mächtigkeit dieser Klassen ist in Tabelle 6 für die Startwerte und die ersten 14 Iterationen aufgetragen worden. Nach 13 Iterationsschritten wird für alle Defekte die gewünschte Genauigkeit von $\delta = 1$ erreicht und liegt dann in der Größenordnung von $\varepsilon_3 = 10^{-10}$. Durch eine andere Wahl der Größen $\varepsilon_1, \varepsilon_2, \varepsilon_3$ kann dieser Effekt variiert werden. Interessant ist nun, daß nicht nur die betragsmäßig großen Defekte verkleinert werden, sondern auch, daß dabei viele der betragsmäßig kleinen Defekte vergrößert werden. Insgesamt läßt sich aus Bild 1 eine Angleichung der Defekte mit zunehmender Zahl der Iterationszyklen erkennen. Die Rechenzeiten betrugen im Beispiel auf einem Rechner der PC AT-Klasse für den ersten Iterationszykel allein etwa 1/2 und insgesamt etwa 1 Stunde.

5.3 Schwingungsproblem mit guten Näherungen

Als einfaches mechanisches Anwendungsbeispiel wurden die ersten 20 Eigenkreisfrequenzen eines eingespannten Kragträgers mit konstanter Biegesteifigkeit $EI = 10.8 \cdot 10^6$ MNm^2, $l = 120$ m und $m = 4 \cdot 10^5$ kg/m gesucht.

Tabelle 7. Ausgangsmatrix, Startwerte und einige Näherungen zu Beispiel 5.3

Ausgangsmatrix nach Lanczos-Tridiagonalisierung

J	$a(J, J)$	$a(J, J-1)$
1	0.1947819666703E+00	
2	0.4284866888423E+00	0.2865996878470E+00
3	0.1381099251160E−01	0.9403898524912E−02
4	0.1947464616815E−02	0.1161295644235E−02
5	0.5260074443683E−03	0.3090040887172E−03
6	0.2123976317939E−03	0.1280430399253E−03
7	0.9105471964440E−04	0.5609584629695E−04
8	0.5196878393178E−04	0.2927704353959E−04
9	0.2871443908344E−04	0.1780970479430E−04
10	0.1749623464267E−04	0.9836835779652E−05
11	0.1231941582170E−04	0.7062542213208E−05
12	0.7468194514568E−05	0.4475039693928E−05
13	0.5850468713203E−05	0.3085308830802E−05
14	0.4026641693350E−05	0.2428024960342E−05
15	0.2895713807837E−05	0.1569043363179E−05
16	0.2443731230988E−05	0.1334153248608E−05
17	0.1635450741661E−05	0.9694452463032E−06
18	0.1442899974949E−05	0.7273628170362E−06
19	0.1104309324171E−05	0.6515397435529E−06
20	0.8355595913391E−06	0.4470440038979E−06
21	0.7832677506142E−06	0.4098387455828E−06
22	0.5479315425151E−06	0.3251402134248E−06
23	0.5095132425376E−06	0.2494290229391E−06
24	0.4220290632711E−06	0.2436312757505E−06
25	0.3216211297637E−06	0.1727269164611E−06
26	0.3245473826458E−06	0.1632089404116E−06
27	0.2340271362398E−06	0.1395858432089E−06
28	0.2210031780073E−06	0.1064379840480E−06
29	0.1968121882374E−06	0.1106807105734E−06
30	0.1480893431309E−06	0.8071307971351E−07
31	0.1575267708337E−06	0.7630488547880E−07
32	0.1172442698921E−06	0.7017022621379E−07
33	0.1095355697424E−06	0.5251243941612E−07
34	0.1046930776722E−06	0.5700051142147E−07
35	0.7723828239963E−07	0.4299440428043E−07
36	0.8495711410749E−07	0.3979581222062E−07
37	0.6577256426621E−07	0.3929540513196E−07
38	0.5962234996752E−07	0.2879475164665E−07
39	0.6099037884999E−07	0.3200854517199E−07
40	0.4417627911051E−07	0.2523962604847E−07
41	0.4926333621023E−07	0.2252318079351E−07
42	0.4009411776817E−07	0.2380810838897E−07
43	0.3473764332207E−07	0.1717960094693E−07
44	0.3789843940522E−07	0.1915298781204E−07
45	0.2718259738474E−07	0.1600898143058E−07
46	0.3010547348941E−07	0.1358092693668E−07
47	0.2617538534907E−07	0.1525503562776E−07
48	0.2150284863498E−07	0.1093239800000E−07
49	0.2482631752377E−07	0.1193633050975E−07
50	0.1792581518204E−07	0.1067436216522E−07

Startwerte und Anfangsdefekte

J	alpha (J)	$d(J)$
1	−0.00000028205899547519	−0.2512539E−01
10	−0.00000009916633500960	0.2275102E−04
20	0.00000012098226522808	0.3673378E−17
30	0.00000040886899768049	−0.7903168E−08
40	0.00000662725437642330	0.1094736E−06
41	0.00001008194129065527	0.3383037E−06
42	0.00001560948252521929	0.4871724E−06
43	0.00002528280883854636	0.3675689E−06
44	0.00004462997809035598	0.4664325E−06
45	0.00008678162038134156	0.6297211E−06
46	0.00019252166013218505	0.2541985E−06
47	0.00052579053495968348	0.3861170E−06
48	0.00201819475201914667	0.6185602E−06
49	0.01581864261347947640	0.5123723E−06
50	0.62124079328168946300	0.3576578E−06

Näherungen und Defekte nach dem 1. Iterationszykel

J	alpha (J)	$d(J)$
1	−0.00000957555798993897	−0.3158263E−05
10	−0.00000005852352157280	0.3318838E−13
20	0.00000015060582022084	−0.4869325E−16
30	0.00000054517942079068	−0.8753555E−17
40	0.00000940501911757852	0.5808476E−07
41	0.00001495471938413877	0.8913246E−07
42	0.00002486225053637150	0.3036564E−06
43	0.00002782823654260782	0.1852304E−04
44	0.00004413039202228053	−0.7875945E−07
45	0.00008613611751370346	−0.3202671E−07
46	0.00019226543541960256	−0.4884542E−08
47	0.00052540320519054091	−0.2555577E−08
48	0.00201757580863514948	−0.1041314E−08
49	0.01581813021965138010	−0.1093049E−09
50	0.62124043562370301300	−0.1941119E−11

Näherungen und Defekte nach dem 3. Iterationszykel

J	alpha (J)	$d(J)$
1	0.00000000001743515809	−0.1482247E−10
10	0.00000002443248587366	−0.8878070E−12
20	0.00000009100563533200	0.1970661E−15
30	0.00000044631748726407	0.9332816E−23
40	0.00000648630578979319	−0.1599753E−19
41	0.00000967971427645904	−0.6405190E−20
42	0.00001510368667280007	−0.3077400E−19
43	0.00002491809404839887	−0.6114067E−19
44	0.00004416795131301283	−0.4936402E−19
45	0.00008616094242517801	−0.1231643E−18
46	**0.00019226988001142512**	**−0.1238079E−18**
47	**0.00052540568144820684**	**0.7160643E−19**
48	0.00201757684175333249	−0.3356124E−18
49	0.01581813032884746280	0.2107954E−17
50	0.62124043562564412700	0.4436891E−16

Nach einer gleichmäßigen Diskretisierung mit 100 Finiten-Elementen (kubischer Verschiebungsansatz für die Steifigkeitsmatrix und die konsistente Massenmatrix) wurde die Steifigkeitsmatrix K nach Cholesky in $L^T L = K$ zerlegt und mit $L^{-T} M L^{-1}$ eine Lanczos-Tridiagonalisierung mit vollständiger Nachorthogonalisierung nach [6, 4.6.1] zur Dimension 50 durchgeführt. Mit einem Bisektionsalgorithmus wurden die Startwerte bis zu einer Genauigkeit von 10^{-6} eingeschlossen bzw. stückweise linear generiert, wenn bei dieser Genauigkeit keine Trennung möglich war.

In Tabelle 7 sind die symmetrische Tridiagonalmatrix, die Startwerte und einige Zahlenwerte (hier die reziproken Quadrate der Eigenkreisfrequenzen in s^2) der ersten Iterationen gezeigt. Mit dem unter 4.4 angegebenen Verfahren mit $\delta = 10^{-12}$ konnten diese Startwerte in wenigen Iterationsschritten verbessert werden.

5.4 Abschließende Bemerkungen

Obwohl der Algorithmus 4.4 in den ersten beiden numerischen Beispielen trotz schlechter Startwerte brauchbare Ergebnisse liefert, erscheint ein solches Vorgehen hinsichtlich einer wirtschaftlichen Berechnung nicht zweckmäßig. Im ersten Beispiel bleibt die Konvergenz nach wenigen Iterationszykeln stehen. Hier ist ein Übergang zu einer ECP-Transformation für mehrfache Knoten empfehlenswert. Im zweiten Beispiel ist der Aufwand für den ersten Iterationszykel zu groß. Eine wirtschaftliche Entscheidung ist es, wenn man nach der Defektberechnung nicht mit dem Algorithmus (4.4) fortfährt, sondern zunächst einen anderen Globallöser verwendet. In diesen Beispielen können insbesondere Zerlegungsalgorithmen verwendet werden. Für betragsmäßig kleine Defekte, wie im Beispiel 5.3, ist rasche Konvergenz zu beobachten. Ähnliches wurde von den Autoren bereits in [2] formuliert. Die ECP-Transformation kann daher zur Nachbehandlung von guten Näherungen der einfachen Polynomnullstellen fast durchgehend empfohlen werden, wenn die Defektberechnung nicht zu aufwendig ist. Dann erhält man ggf. Fehlerabschätzungen und mit kurzen Algorithmen bessere Näherungen oder die negative Information, daß die Qualität der Näherungen noch nicht ausreichend war und weitere Iterationen mit einem anderen Algorithmus sinnvoll erscheinen.

Für die linearen Eigenwertprobleme der Strukturmechanik ergibt sich damit das prinzipielle Vorgehen aus Beispiel 5.3: Nach einer Tridiagonalisierung werden mit einer QR-Zerlegung und/oder einem Bisektionsalgorithmus gute Näherungen aller einfachen Eigenwerte berechnet. Dabei wird die Qualität dieser Näherungen durch gelegentliche Defektberechnung überwacht. Sobald die Defekte hinreichend klein sind, so daß Fehlerabschätzungen möglich werden, werden die Näherungen je nach Problemstellung defektgesteuert oder simultan unter Ausnutzung von (4.5) mit quadratischer Konvergenz verbessert.

Die wesentlichen Nachteile bei diesem Vorgehen sind der hohe Aufwand bei der Defektberechnung, die Nichtnormalität der ECP-transformierten Matrix, mangelnde Informationen über die Eigenvektoren des Originalproblemes sowie die Tatsache, daß für eine effektive Defektsteuerung alle Defekte klein, also alle Näherungen der Polynomnullstellen gut sein müssen. Folglich ist dieses Vorgehen nur dann besonders geeignet, wenn alle Eigenwerte des transformierten Problemes gesucht sind.

Literatur

1 ZURMÜHL, R.; FALK, S.: Matrizen und ihre Anwendungen. Band 1 und 2, 5. Auflage, Springer Verlag Berlin, Heidelberg, New York, Tokyo 1984 und 1986.

2 CARSTENSEN, C.; STEIN, E.: Analysis und Berechnung der Falkschen ECP-Transformation und verwandte Probleme. In COLLATZ et al. (Hrsg.): ISNM 83 Tagungsband ‚Numerische Behandlung von Eigenwertaufgaben', Band 4. Oberwolfach-Tagung Dezember 1986; Birkhäuser Verlag, Basel 1986.

3 HOUSEHOLDER, A. S.: The theory of matrices in numerical analysis. Dover Publications 1964.

4 EBERLEIN, P. J.: A two parameter test matrix. Math. Comput. **18** (1964), 296—298.

5 SHERMAN, J.; MORRISON, W. J.: Adjustment of an inverse matrix corresponding to changes in a given column or a given row of the original matrix. Ann. Math. Statist. **21** (1950), 124—127.

6 BUNSE, W.; BUNSE-GERSTNER, A.: Numerische lineare Algebra. B. G. Teubner, Stuttgart 1985.

7 ILIEF, L.; DOVCEV, K.: Über Newtonsche Iterationen. Wiss. Z. TU Dresden **12** (1963), 117—118.

8 BÖRSCH-SUPAN, W.: Residuenabschätzung für Polynom-Nullstellen mittels Lagrange-Interpolation. Numer. Math. **14** (1970), 287—296.

9 KERNER, I. O.: Ein Gesamtschrittverfahren zur Berechnung der Nullstellen von Polynomen. Numer. Math. 8 (1966), 290—294.

Eingegangen: 5. Mai 1988, revidiert: 29. November 1988

Anschrift: C. CARSTENSEN, Prof. Dr.-Ing. ERWIN STEIN, Universität Hannover, Institut für Baumechanik und Numerische Mechanik, Appelstraße 9a, D-3000 Hannover, BRD